UNIVERSIDAD NACIONAL DE CHIMBORAZO

FACULTAD DE INGENIERÍA

Carrera de Arquitectura

Probabilidad y estadística

Byron Obregón

Fredy Barahona

INDICE DE TABLAS

INDICE DE ILUSTRACIONES

ÍNDICE GENERAL

1. UNIDAD 1: ESTADÍSTICA DESCRIPTIVA

Es la parte de la estadística que permite analizar todo un conjunto de datos, de los cuales se extraen conclusiones valederas, únicamente para ese conjunto. Para realizar este análisis se procede a la recolección y representación de la información obtenida. Como ejemplo de estas estadísticas podemos citar a aquellas que se obtienen generalmente en los deportes, en los rendimientos académicos de los estudiantes de determinada materia, en los negocios al determinar las ventas obtenidas mensualmente en un determinado año por una empresa en particular. (Del Castillo Galarza & Salazar Pinto, 2018)

1.1. Elementos básicos de la estadística, Variables: Colectivo estadístico, muestra, variables cuantitativas, variables cualitativas

1.1.1. Población, Muestra y Variables

La disciplina de estadística proporciona métodos de organizar, resumir datos y de sacar conclusiones basadas en la información contenida en los mismos. Por tanto, en una investigación el enfoque se basa en una colección bien definida de objetos que constituyen una "población de interés". Cuando la información deseada está disponible para todos los objetos de la población, se tiene lo que se llama "censo". Pero se dificulta de alguna manera, por restricciones de tiempo, dinero y otros recursos, que lo vuelven un proceso impráctico. Para evitar esta situación, se utiliza un subconjunto de la población, "una muestra". Por lo general, si se quiere conocer ciertas características de objetos de una población, se hace uso de la variable. Una variable es cualquier característica cuyo valor puede cambiar de un objeto a otro en la población. Inicialmente las letras minúsculas del alfabeto denotaran las variables (x,y,z). (Devore, 2011)

1.1.2. Tipos de Variables

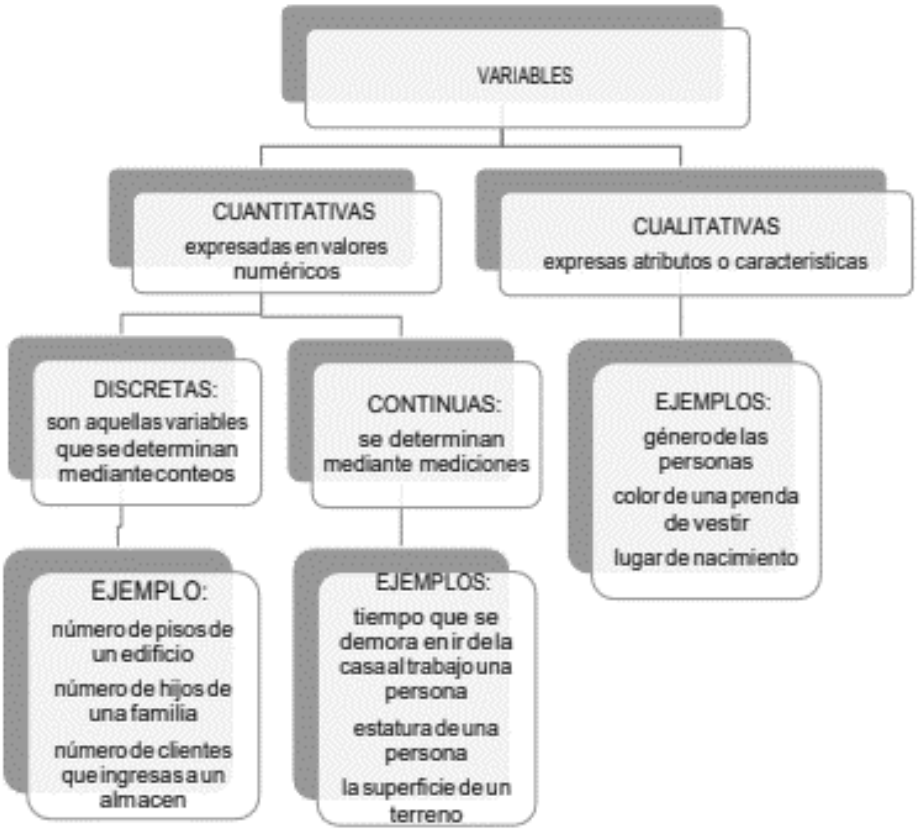
Se pueden clasificar variables en una de 2 categorías: cualitativas y cuantitativas.

Las variables cualitativas miden una cualidad o característica en cada unidad experimental. Este tipo de variables producen datos que se pueden clasificar de acuerdo a similitudes o diferencias en clase, por tanto, con frecuencia se denominan datos categóricos. Por ejemplo, las variables como género, año y especialidad.

En cambio, las variables cuantitativas miden una cantidad numérica en cada unidad experimental. Con frecuencia son representadas por la letra x, que producen datos numéricos. Por ejemplo, x = peso de un paquete listo para ser enviado. (Mendenhall, Beaver, & Beaver, 2010)

Ilustración 1.

Variables



Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

Tabla 1.

Ejemplo de Variables

Variables		
CUALITATIVAS	CUANTITATIVAS DICRETAS	CUANTITATIVAS CONTINUAS
Barrio donde se halla ubicada	Número de dormitorios	Área de construcción
Tipo de estructura	Número de baños	Área del terreno sobre el que está ubicada
Disponibilidad e garaje	Número de pisos	Área de patios y jardines exterior
Tipo de cubierta	Antigüedad de la construciión	Longitud del terreno de la casa
Material utilizado en las paredes		Altura de la casa
Disponibilidad de los servicios básicos		

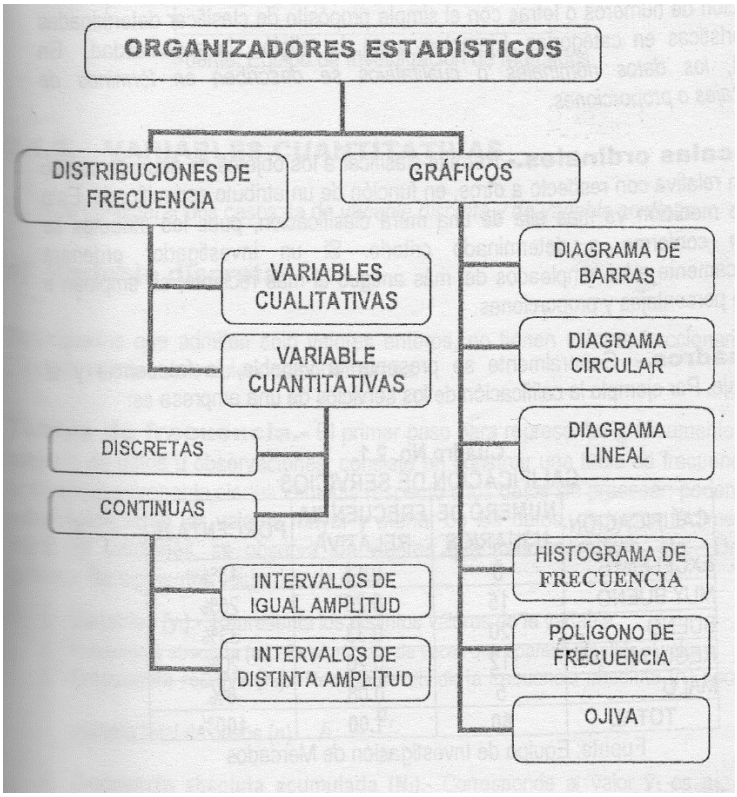
Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

1.2. Distribuciones de frecuencia

Se trata de una representación tabular de datos estadísticos que ofrece una síntesis de la información, permitiendo una rápida comprensión de su comportamiento. Este comportamiento incluye la identificación aproximada de los valores centrales, la variabilidad y la simetría en relación con un valor central. Las tablas de frecuencias pueden resumir diferentes tipos de datos, ya sean categóricos (nominales), ordinales, discretos o continuos. Para los datos nominales, ordinales y discretos, la distribución de frecuencias generalmente consta de dos columnas: una para las categorías o valores observados y otra para las frecuencias correspondientes a cada categoría, (Del Castillo Galarza & Salazar Pinto, 2018).

Ilustración 2.

Organizadores Estadísticos



Fuente: (Alvarez Roman, 2004)

Notación

Es importante primero adquirir familiaridad con ciertos símbolos que serán empleados tanto para variables discretas como para variables continuas.

$\overline{n}$	=	n	= Tamaño de la muestra.
N	=	N	= Tamaño de la población o universo.
X_i	=	x_i	= Identificación para cada valor observado. (minúscula en la muestra)
f_i	=	n_i	= Frecuencias absolutas.
f_i/n	=	h_i	= Frecuencias relativas.
F_i	=	N_i	= Frecuencias absolutas acumuladas.
H_i	=	H_i	= Frecuencias relativas acumuladas.
X_i	=	y_i	= Identifica la variable discreta o las marcas de clase en la continua.
X'_{i-1}-X'_i	=	y'_{i-1}- y'_i	= Identifica a la variable continua con sus intervalos.
i	=	c	= Amplitud del intervalo
$\underline{m}$		m	= Número de valores de la variable o de intervalos.

Se ha decidido mostrar las dos notaciones más comunes en los textos de Estadística para que los estudiantes puedan consultar o utilizar cualquiera de ellas sin distinción.

(Martínez Bencardino, 2012)

1.2.1. Distribución estadística unitaria

Variables Cualitativas: Se distinguen por registros que no poseen valores numéricos, aunque es factible asignarles códigos numéricos, sin seguir un orden específico en la asignación de dichos códigos. (Alvarez Roman, 2004)

a) Escalas nominales: Es el nivel más bajo de medición, consiste en la asignación de números o letras con el simple propósito de clasificar determinadas características en categorías. Ejemplos: sexo, la religión, la especialidad. En general, los datos nominales o cualitativos se describen en términos de porcentajes o proporciones. (Alvarez Roman, 2004)
b) Escalas ordinales: Facilitan la clasificación de objetos según su posición relativa respecto a otros, según un atributo específico. Este nivel de medición va más allá de una simple clasificación, ya que los atributos se disponen en función de un criterio determinado. Por ejemplo, si un investigador ordenara jerárquicamente a los empleados desde el más antiguo hasta el más reciente, se suelen emplear porcentajes y proporciones. (Alvarez Roman, 2004)
c) Cuadros: Generalmente se presenta la variable, la frecuencia y el porcentaje. Por ejemplo, la calificación de los servicios de una empresa es:

Tabla 2.

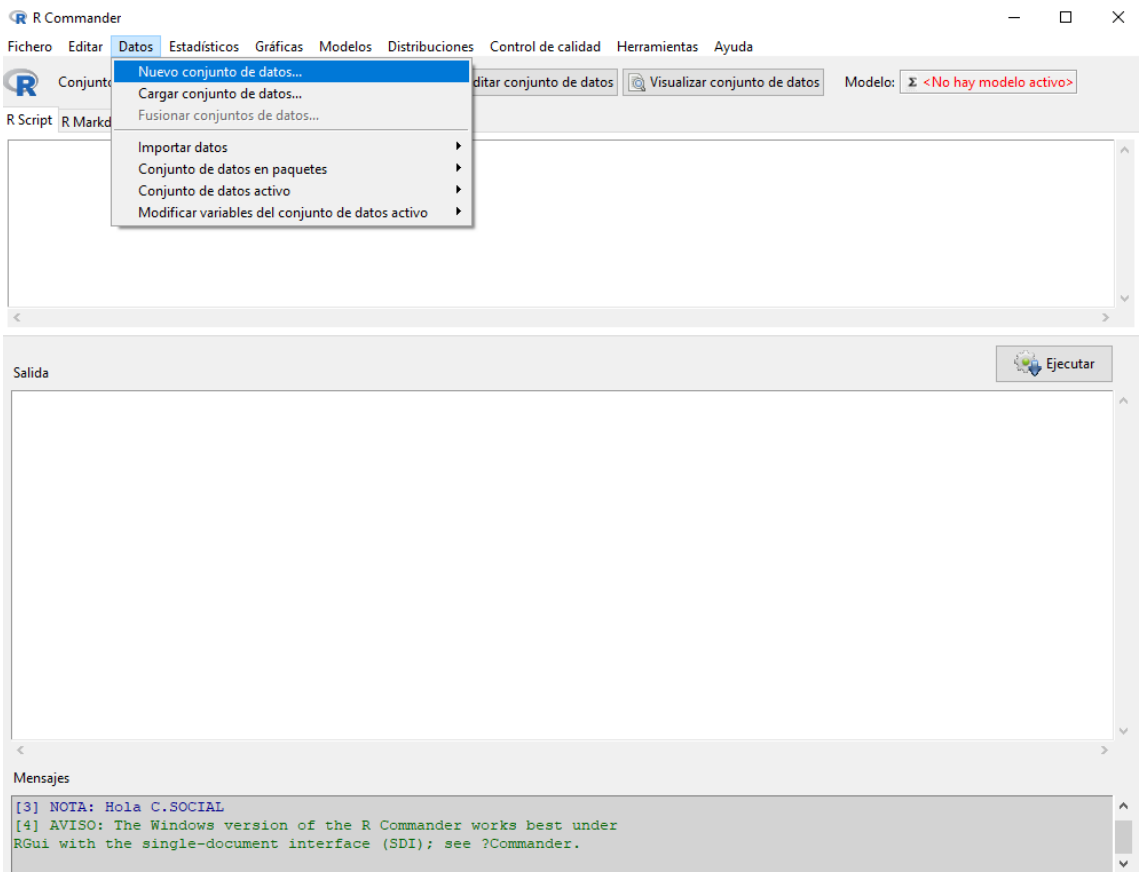
Calificación de Servicios

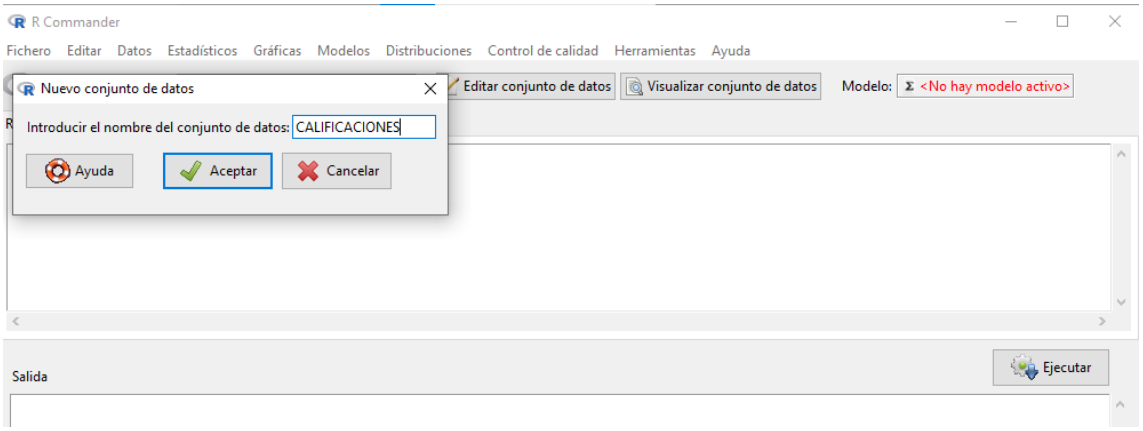
CALIFICACIÓN	NÚMERO DE USUARIOS	FRECUENCIA RELATIVA	PORCENTAJE
EXCELENTE	8	0.13	13,33… %
MUY BUENO	15	0.25	25%
BUENO	20	0.33	33,33…%
REGULAR	12	0.20	20%
MALO	5	0.09	8,33…%
TOTAL	60	1	100%

Fuente: (Alvarez Roman, 2004)

Ilustración 3.

Comprobación de Calificación de Servicios en el software R Commander



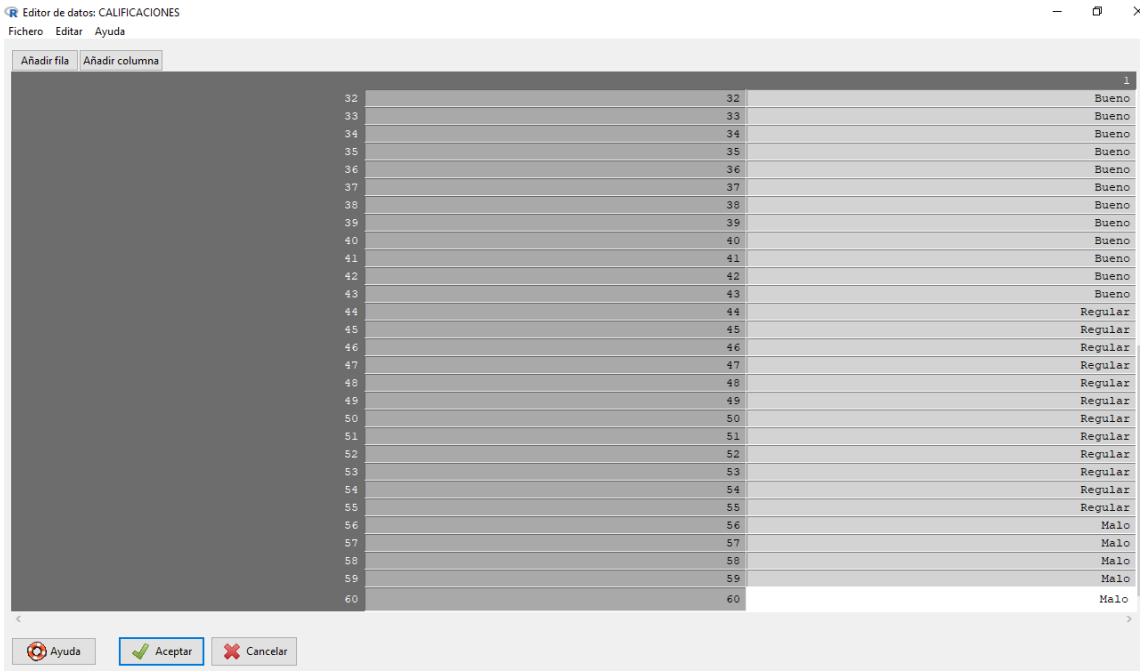


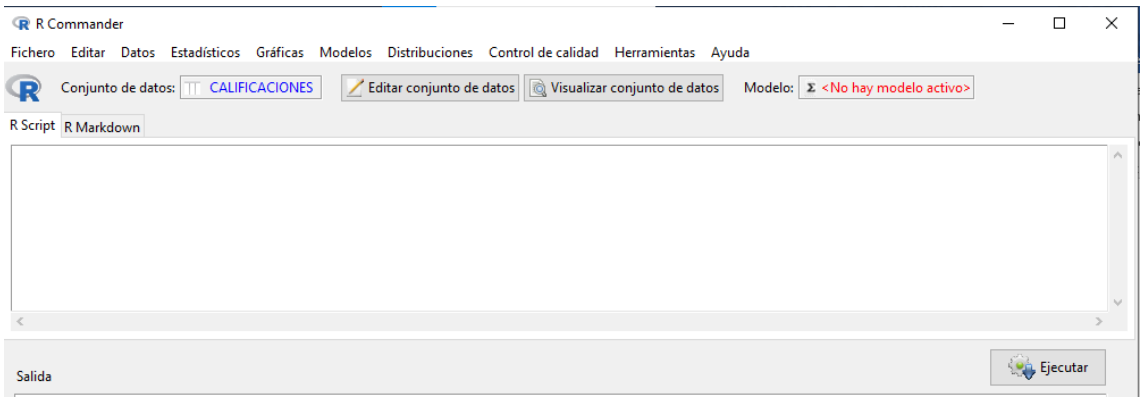
Editor de datos: CALIFICACIONES

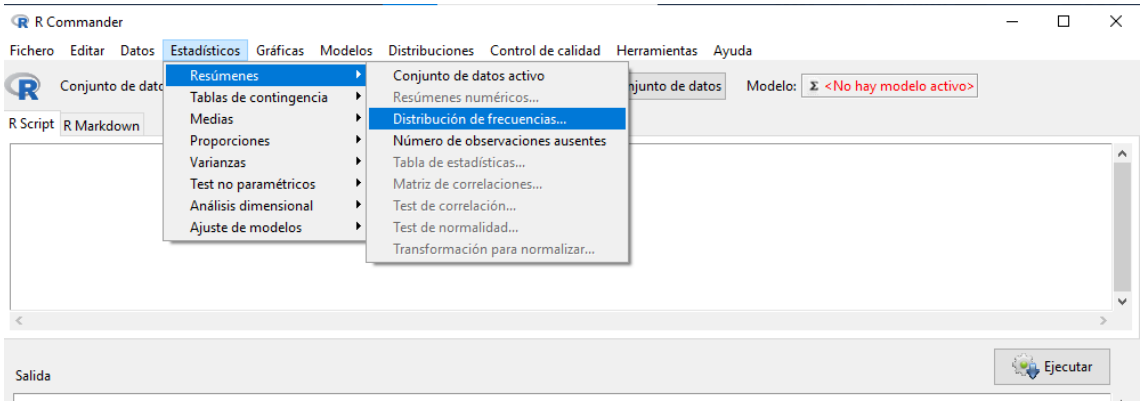
Fichero Editar Ayuda

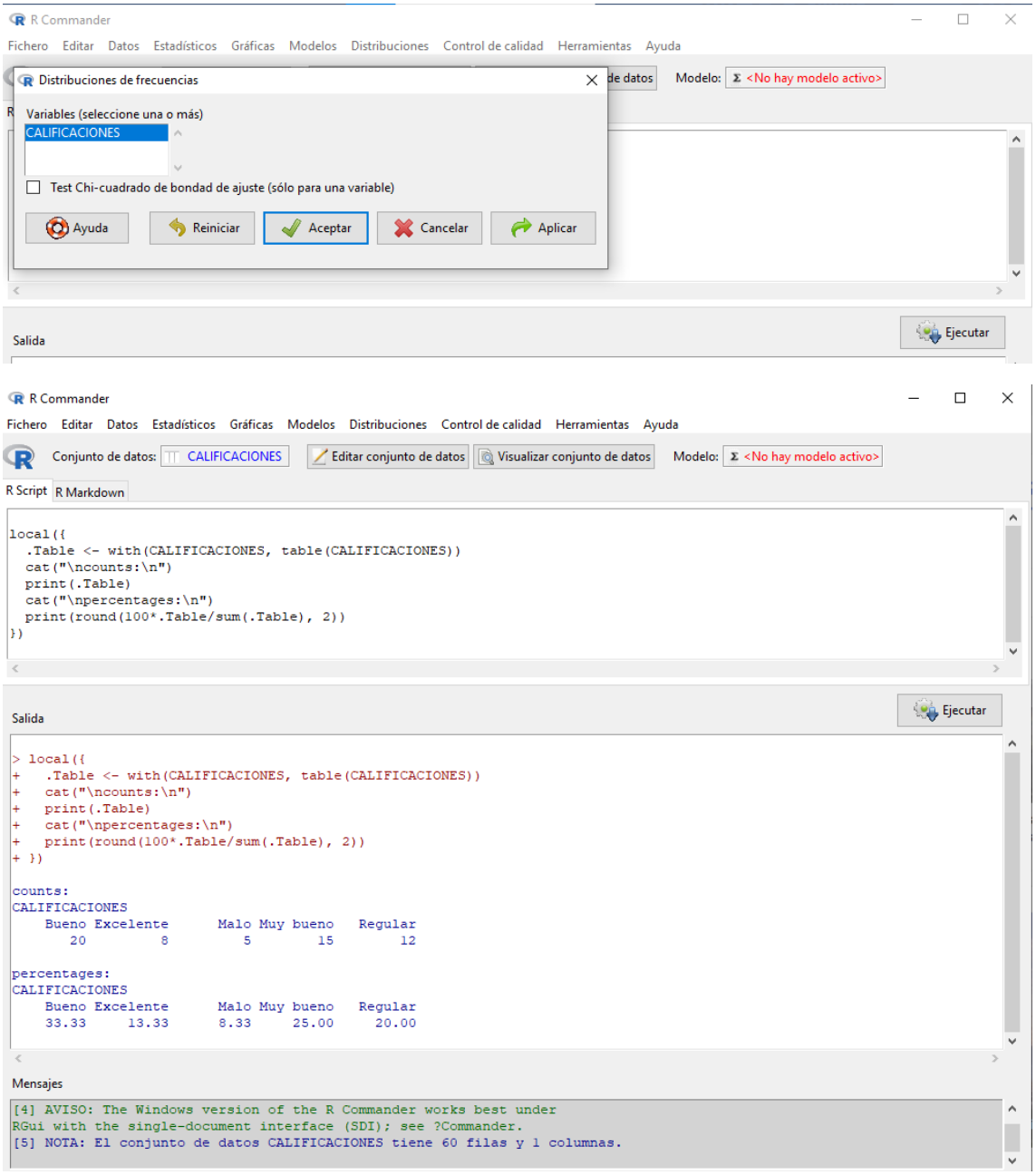
Añadir fila | Añadir columna

	1
rowname	CALIFICACIONES
1	Excelente
2	Excelente
3	Excelente
4	Excelente
5	Excelente
6	Excelente
7	Excelente
8	Excelente
9	Muy bueno
10	Muy bueno
11	Muy bueno
12	Muy bueno
13	Muy bueno
14	Muy bueno
15	Muy bueno
16	Muy bueno
17	Muy bueno
18	Muy bueno
19	Muy bueno
20	Muy bueno
21	Muy bueno
22	Muy bueno
23	Muy bueno
24	Bueno
25	Bueno
26	Bueno
27	Bueno
28	Bueno


Editor de datos: CALIFICACIONES
Fichero Editar Ayuda
Añadir fila
Añadir columna
1
32 32 Bueno
33 33 Bueno
34 34 Bueno
35 35 Bueno
36 36 Bueno
37 37 Bueno
38 38 Bueno
39 39 Bueno
40 40 Bueno
41 41 Bueno
42 42 Bueno
43 43 Bueno
44 44 Regular
45 45 Regular
46 46 Regular
47 47 Regular
48 48 Regular
49 49 Regular
50 50 Regular
51 51 Regular
52 52 Regular
53 53 Regular
54 54 Regular
55 55 Regular
56 56 Malo
57 57 Malo
58 58 Malo
59 59 Malo
60 60 Malo
Ayuda
Aceptar
Cancelar


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Conjunto de datos: CALIFICACIONES
Editar conjunto de datos
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
R Script
R Markdown
Salida
Ejecutar


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Resúmenes
Tablas de contingencia
Medias
Proporciones
Varianzas
Test no paramétricos
Análisis dimensional
Ajuste de modelos
Conjunto de datos activo
Resúmenes numéricos...
Distribución de frecuencias...
Número de observaciones ausentes
Tabla de estadísticas...
Matriz de correlaciones...
Test de correlación...
Test de normalidad...
Transformación para normalizar...
Modelo: <No hay modelo activo>
R Script
R Markdown
Salida
Ejecutar



Fuente: Elaboración Propia

Observaciones: Para calcular la frecuencia relativa, es necesario dividir el número total entre la cantidad específica en cuestión. Como se observa en la tabla, el total es de 60, y al dividirlo entre los 8 usuarios excelentes, obtenemos 0.133… como resultado. Para obtener el porcentaje, multiplicamos la frecuencia relativa por 100.**Distribución estadística de frecuencias sin clases**

Variables Cuantitativas: Se debe considerar dos casos, la de variable discreta y de variable continua.

a. Variable discreta: son aquellas que admiten solo valores enteros, no tienen valores fraccionarios. Ejemplo: número de empleados por empresa.

Tablas de frecuencia: El inicio del proceso de representación gráfica de un conjunto de datos u observaciones implica la creación de una tabla de frecuencias. Esta tabla ofrece ventajas sobre los datos originales: facilita la identificación rápida de los valores máximos y mínimos, permite una división sencilla en secciones y revela los valores que se repiten con mayor frecuencia. Se recomienda que la tabla contenga las siguientes columnas de datos:Variables (Y_i): Representa los distintos valores de la variable.

i. Frecuencia absoluta (n_i): Es el número de veces que aparece repetido el valor Y_i.
ii. Frecuencia relativa (h_i): Resulta de dividir la frecuencia absoluta (n_i) para el número total de casos (n). $hi = \frac{ni}{n}$
iii. Frecuencia absoluta acumulada (N_i): Corresponde al valor $Y_i$es n_i, por definición será igual a: $N_i = n_1 + n_2 + n_3 + \cdots + n_i$
iv. Frecuencia relativa acumulada (H_i): Correspondiente al valor $Y_i$que por definición será igual a: $H_i = h_1 + h_2 + h_3 + \cdots + h_i$
v. . (Alvarez Roman, 2004)

Ejemplo: En el país existen 2252 establecimientos de alojamiento y se desea realizar una investigación sobre el numero de turistas que atienden en promedio por día. Por razones económicas y de tiempo, se desea que la investigación no sea exhaustiva (no examinar la población total), por lo que se selecciono una muestra de 20 establecimientos de alojamiento. El resultado es el siguiente:

23	22	20	23	23	23	21	21	20	21
23	23	24	24	23	22	24	22	24	22

Tabla 3.

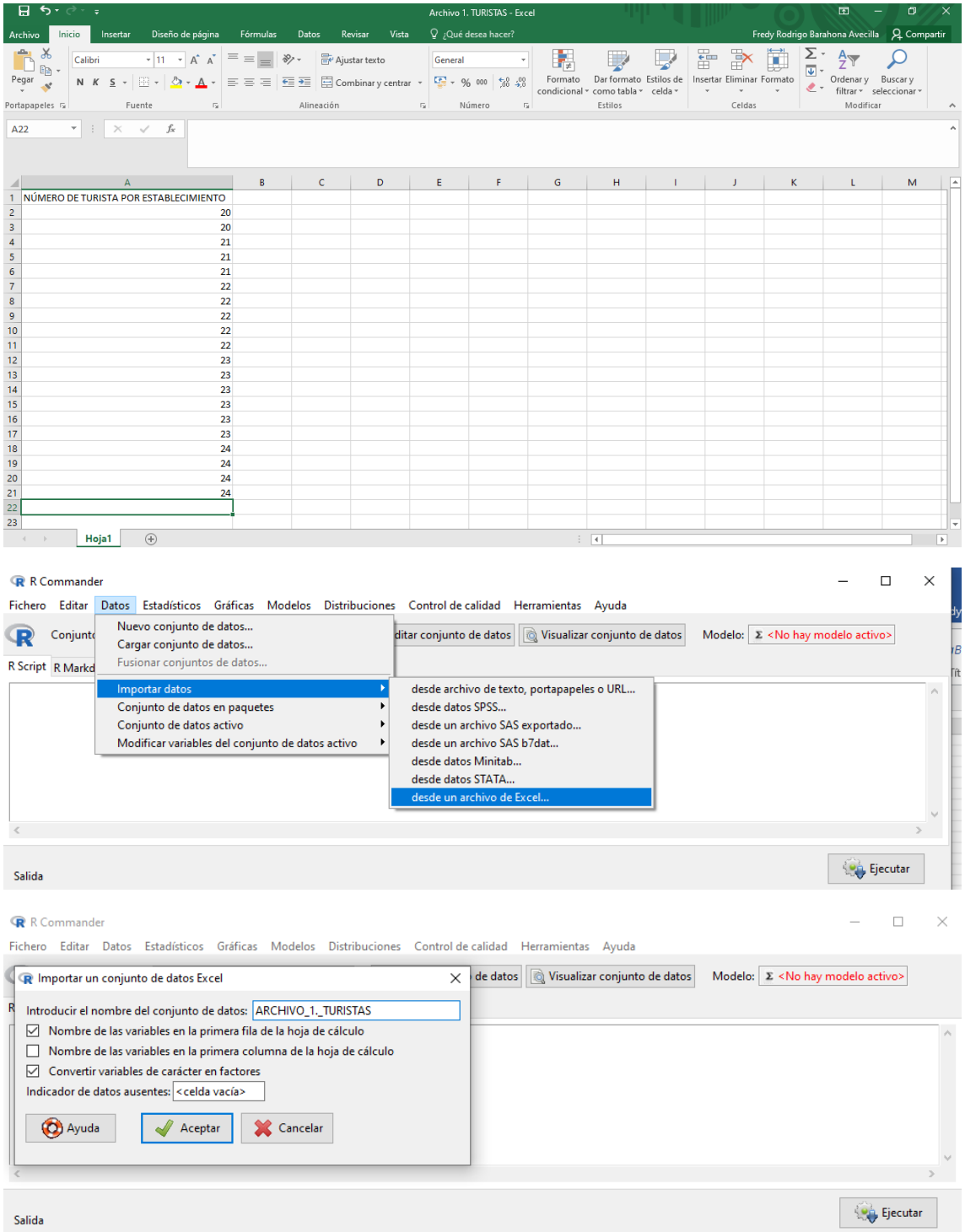
Número de Turistas por Establecimiento

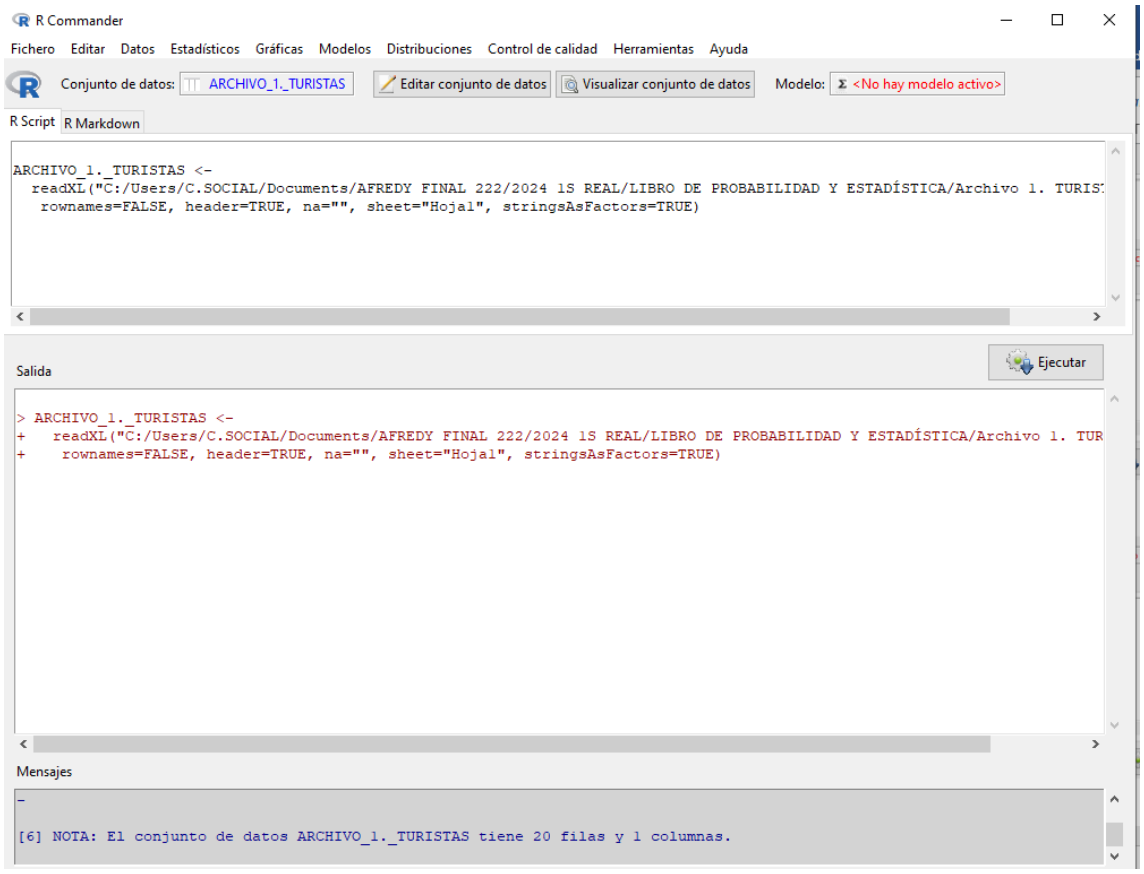
Valores e la variable (y_i)	Frecuencia absoluta n_i	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada N_i	Frec. Relativa acumulada (H_i)
$y_1 = 20$	$n_1 = 2$	$h_1 = 0.10$	$N_1 = 2$	$H_1 = 0.10$
$y_2 = 21$	$n_2 = 3$	$h_2 = 0.15$	$N_2 = 5$	$H_2 = 0.25$
$y_3 = 22$	$n_3 = 5$	$h_3 = 0.25$	$N_3 = 10$	$H_3 = 0.50$
$y_4 = 23$	$n_4 = 6$	$h_4 = 30$	$N_4 = 16$	$H_4 = 0.80$
$y_5 = 24$	$n_5 = 4$	$h_5 = 20$	$N_5 = 20$	$H_5 = 1.00$
	$n = 20$	1.00		

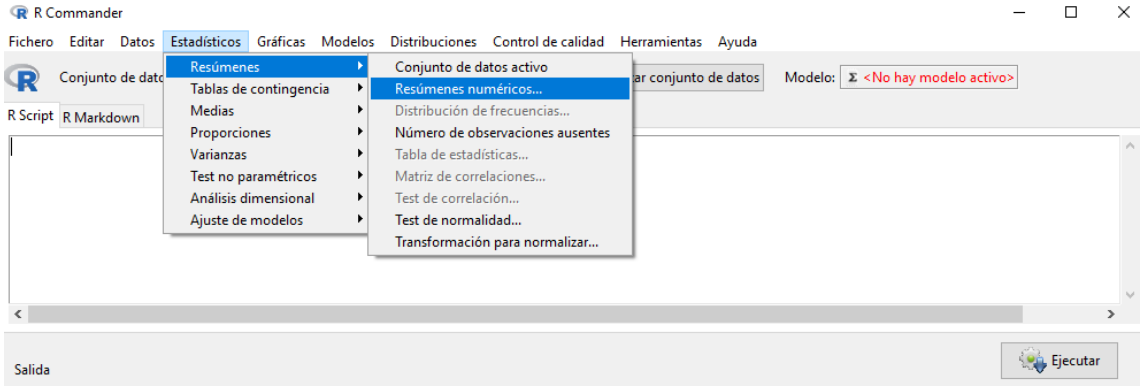
Fuente: (Alvarez Roman, 2004)

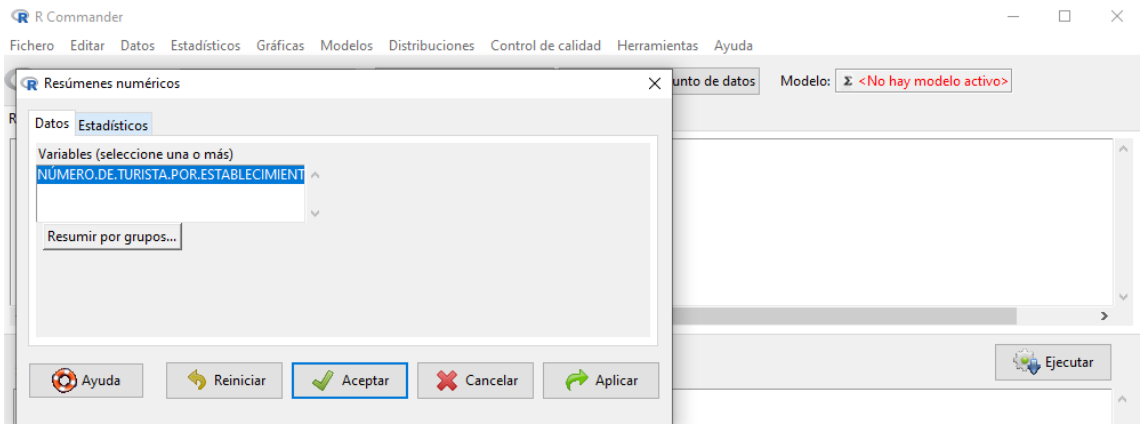
Ilustración 4.

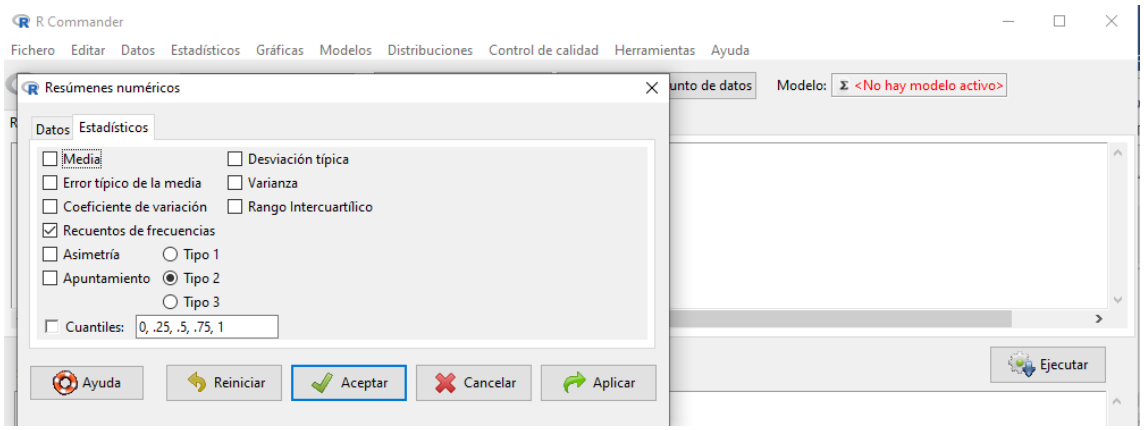
Comprobación de Número de turistas en el software R Commander

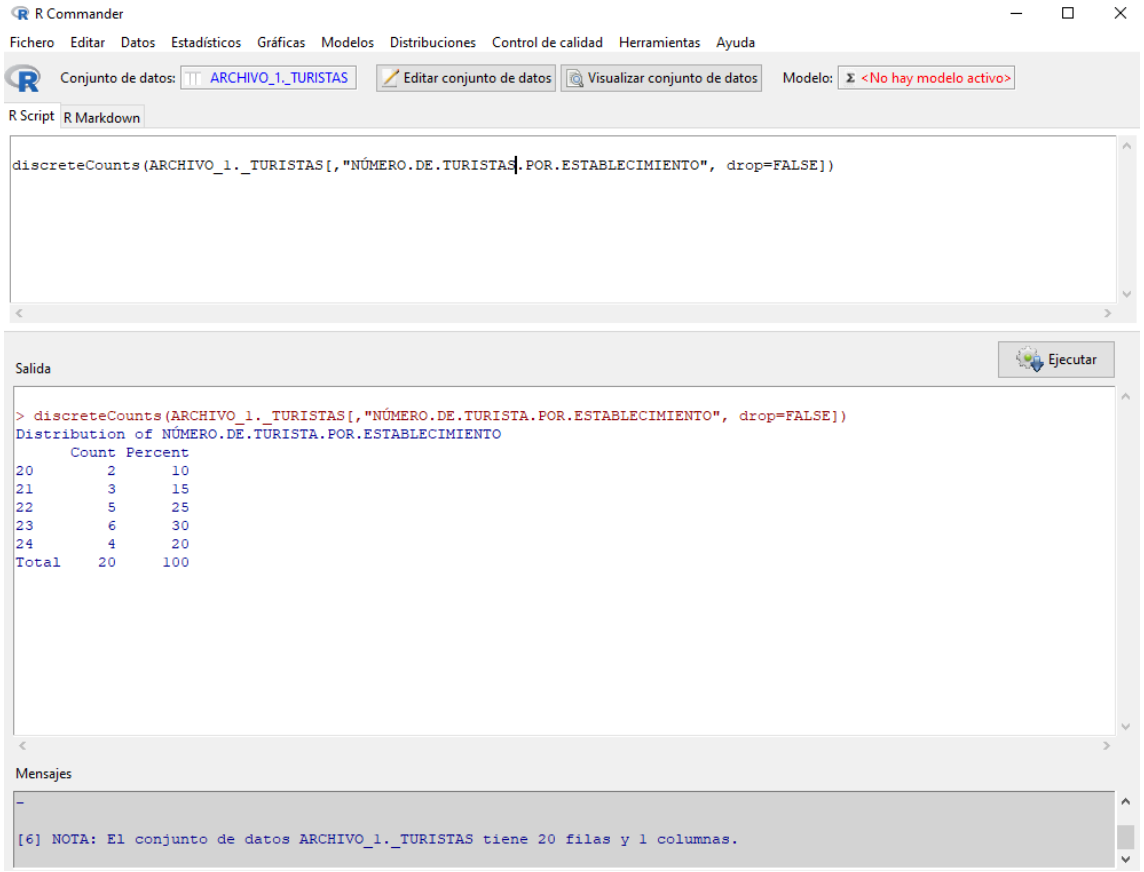



R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Conjunto de datos: ARCHIVO_1._TURISTAS
Editar conjunto de datos
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
R Script R Markdown
ARCHIVO_1._TURISTAS <-
readXL("C:/Users/C.SOCIAL/Documents/AFREDY FINAL 222/2024 1S REAL/LIBRO DE PROBABILIDAD Y ESTADÍSTICA/Archivo 1. TURIS
rownames=FALSE, header=TRUE, na="", sheet="Hoja1", stringsAsFactors=TRUE)
Salida
Ejecutar
> ARCHIVO_1._TURISTAS <-
+ readXL("C:/Users/C.SOCIAL/Documents/AFREDY FINAL 222/2024 1S REAL/LIBRO DE PROBABILIDAD Y ESTADÍSTICA/Archivo 1. TUR
+ rownames=FALSE, header=TRUE, na="", sheet="Hoja1", stringsAsFactors=TRUE)
Mensajes
[6] NOTA: El conjunto de datos ARCHIVO_1._TURISTAS tiene 20 filas y 1 columnas.


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Resúmenes
Tablas de contingencia
Medias
Proporciones
Varianzas
Test no paramétricos
Análisis dimensional
Ajuste de modelos
Conjunto de datos activo
Resúmenes numéricos...
Distribución de frecuencias...
Número de observaciones ausentes
Tabla de estadísticas...
Matriz de correlaciones...
Test de correlación...
Test de normalidad...
Transformación para normalizar...
Modelo: <No hay modelo activo>
R Script R Markdown
Salida
Ejecutar


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Resúmenes numéricos
Datos Estadísticos
Variables (seleccione una o más)
NÚMERO.DE.TURISTA.POR.ESTABLECIMIENT
Resumir por grupos...
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar
Modelo: <No hay modelo activo>
Ejecutar






```
> discreteCounts(turistas[,"Numero.de.turistas.por.establecimiento", drop=FALSE])
Distribution of Numero.de.turistas.por.establecimiento
      Count Percent
20        2      10
21        3      15
22        5      25
23        6      30
24        4      20
Total    20     100
```

Fuente: Elaboración Propia

1.2.2. Distribución estadística de frecuencias con clases

Variable continua: son aquellas que admiten que el fenómeno se exprese en forma fraccionaria o decimal. Ejemplo: La estatura de una persona es 1,75m, el sueldo de trabajador es de $520,35, etc. (Alvarez Roman, 2004)

Tablas de intervalos: debe contener las siguientes columnas:

i. Intervalos o clases $(y_{i-1} - y_i)$: Para la construcción de los intervalos debemos considerar los siguientes pasos:
   a. El recorrido de la variable o rango (L): $L = X_{max} - X_{min}$
   b. Numero de Intervalos (k): El número de clases o intervalos se establece en forma convencional, es decir, dependerá del problema en cuestión (a más datos más intervalos y menos datos menos intervalos). En general, se recomienda que el número de clases este entre cinco y dieciséis. Sugerimos tres métodos:
      1. Método practico: $k = \sqrt{n}$
      2. Método de Sturges: $k = 1 + 3.322\ \log(n)$ (recomendado)
      3. El investigador sugiere: De acuerdo a la necesidad cada investigador (el sugiere el número de intervalos)
   c. La amplitud del intervalo (c): $c = \frac{L}{k}$ o $c = \frac{Xmax - Xmin}{k}$
      i. Si c=número entero debe construir directamente los intervalos
      ii. Si c=número decimal, solo por facilitar los cálculos se recomienda llevar de inmediato superior y continuar con el siguiente proceso:
         1. Amplitud del intervalo modificado (Cm): cuando llevamos al inmediato superior el valor de c.
         2. Recorrido de la variable modificada (LM) es: $Cm = \frac{Lm}{k}$
         3. La diferencia (d) entre Lm y L es: $d = Lm - L$
         4. Esta diferencia debe repartirse entre los valores de Xmax y Xmin, obteniéndose los valores de Xmax y Xmin modificadas.

ii. Punto medio o marca o marca de clase (Yi): $MC = Yi = \frac{y`j-1+y`j}{2}$

iii. Frecuencia absoluta (ni): Es el número de veces que aparece repetido el valor Y1 en cada intervalo. Para facilitar su conteo se recomienda utilizar el diagrama troco hoja.

iv. Frecuencia relativa (hi): Resulta de dividir la frecuencia absoluta (ni) para el número total de casos (n). $hi = \frac{ni}{n}$

vi. Frecuencia absoluta acumulada (Ni): Corresponde al valor Y1 es n1, por definición será igual a: $N_i = n_1 + n_2 + n_3 + \cdots + n_i$

v. Frecuencia relativa acumulada (Hi): Correspondiente al valor Yi que por definición será igual a: $H_i = h_i + h_2 + h_3 \ldots + h_n$.

Ejemplo: Se conoce que visitan una playa alrededor de 1200 turistas a la semana y se desea realizar una investigación sobre el gasto diario en dólares que tienen los turistas.

Por razones económicas y de tiempo, se desea que la investigación no sea exhaustiva (no examinar la población total), por lo que se seleccionó una muestra de 20 turistas. El resultado es el siguiente:

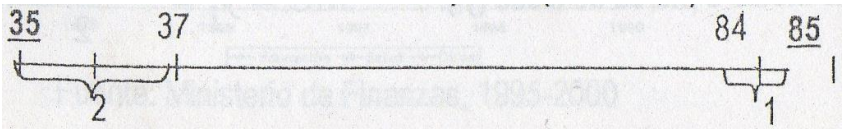
72	45	48	66	84	57	60	59	42	61
78	37	55	75	51	67	64	57	47	69

Solución:

1. $L = X_{max} - X_{min} = 84 - 37 = 47$
2. $k = \sqrt{n} = \sqrt{20} = 4{,}5 = 5$
3. $c = \frac{X_{max} - X_{min}}{k} = \frac{47}{5} = 9{,}4 = 10 \ (C_m = 10)$

Nota: Cuando la amplitud del intervalo (c) es un numero entero se facilita el trabajo, razón por la que se aproxima al entero inmediato superior (10) y le denominamos Amplitud del intervalo modificada (C_m).

1. $C_m = \frac{Lm}{k}$, $10 = \frac{Lm}{5}$, $L_m = 50$
2. $D = L_m - L$, $d = 50 - 47$, $d = 3$ (Esta diferencia debe repartirse entre los valores de X_{max} y X_{min}: (2,1 o 1,2) (Alvarez Roman, 2004)



Método 1:

Tabla 4.

Gastos en Dólares de los Turistas (Método 1)

GASTOS EN DÓLARES DE LOS TURISTAS					
$(y'_{i-1} - i)$	Y_i	n_i	h_i	N_i	H_i
$(35 - 45)$	40	3	0,15	3	0,15
$(45 - 55)$	50	4	0,20	7	0,35
$(55 - 65)$	60	6	0,30	13	0,65
$(65 - 75)$	70	5	0,25	18	0,90
$(75 - 85)$	80	2	0.10	20	1
10		20	1		
C		n	$\sum h_i$		

Fuente: (Alvarez Roman, 2004)

Método 2:

Tabla 5

Gastos en Dólares de los Turistas (Método 2)

GASTOS EN DÓLARES DE LOS TURISTAS					
$(y'_{i-1} - i)$	Y_i	n_i	h_i	N_i	H_i
$(35{,}1 - 45)$	$Y_1 = 40$	$n_1 = 3$	$h_1 = 0{,}15$	$N_1 = 3$	$H_1 = 0{,}15$
$(45{,}1 - 55)$	$Y_2 = 50$	$n_2 = 4$	$h_2 = 0{,}20$	$N_2 = 7$	$H_2 = 0{,}35$
$(55{,}1 - 65)$	$Y_3 = 60$	$n_3 = 6$	$h_3 = 0{,}30$	$N_3 = 13$	$H_3 = 0{,}65$
$(65{,}1 - 75)$	$Y_4 = 70$	$n_4 = 5$	$h_4 = 0{,}25$	$N_4 = 18$	$H_4 = 0{,}90$
$(75{,}1 - 85)$	$Y_5 = 80$	$n_5 = 2$	$h_5 = 0.10$	$N_5 = 20$	1
10		20	1		
C		n	$\sum h_i$		

Fuente: (Alvarez Roman, 2004)

Método 3:

Tabla 6.

Gastos en Dólares de los Turistas (Método 3)

GASTOS EN DÓLARES DE LOS TURISTAS					
$(y'_{i-1} - i)$	Y_i	n_i	h_i	N_i	H_i
$(36 - 45)$	40,5	3	0,15	3	0,15
$(46 - 55)$	50,5	4	0,20	7	0,35
$(56 - 65)$	60,5	6	0,30	13	0,65
$(66 - 75)$	70,5	5	0,25	18	0,90
$(76 - 85)$	80,5	2	0.10	20	1
10		20	1		
C		n	$\sum h_i$		

Fuente: (Alvarez Roman, 2004)

1.3. Representaciones Graficas

1.3.1. Histograma

Básicamente, este tipo de representación visual se compone de una serie de rectángulos contiguos, cada uno correspondiente a una categoría, asegurando que el área de cada rectángulo sea igual o proporcional a la frecuencia de la categoría respectiva. La variable de interés se coloca en el eje horizontal, mientras que la frecuencia de la clase (absoluta, relativa o porcentual) se muestra en el eje vertical. Cuando los intervalos de clase son uniformes para todas las categorías, la altura de cada rectángulo coincide con la frecuencia de la clase. Este tipo de gráfico es común en las distribuciones de frecuencia donde la variable estudiada es de naturaleza cuantitativa continua, con las clases o categorías definidas por intervalos.

Si el análisis abarca todo el proceso desde la creación de la distribución de frecuencias, se puede utilizar la herramienta "Histograma" disponible en la opción Insertar "Gráficos recomendados", Histograma y luego en + se puede ubicar elementos del gráfico como ejes, títulos de ejes…

. (Del Castillo Galarza & Salazar Pinto, 2018)

Ejemplo: Se realizó un estudio sobre una muestra de 100 familias de cuatro integrantes, para determinar cuál es el gasto semanal en alimentación, obteniéndose la siguiente distribución:

20	51	60	95	115	135	160	190	220	255
25	51	78	95	115	135	160	190	240	280
30	57	78	95	120	135	170	190	240	280
35	57	78	100	120	140	170	190	240	290
35	60	78	100	120	140	170	200	240	290
40	60	90	100	125	145	170	200	245	300
42	60	90	110	125	145	180	210	245	300
43	60	90	110	125	145	180	210	245	340
48	60	90	110	130	150	185	220	245	350
50	60	95	115	130	150	190	220	245	255

Tabla 7.

Nº de familias

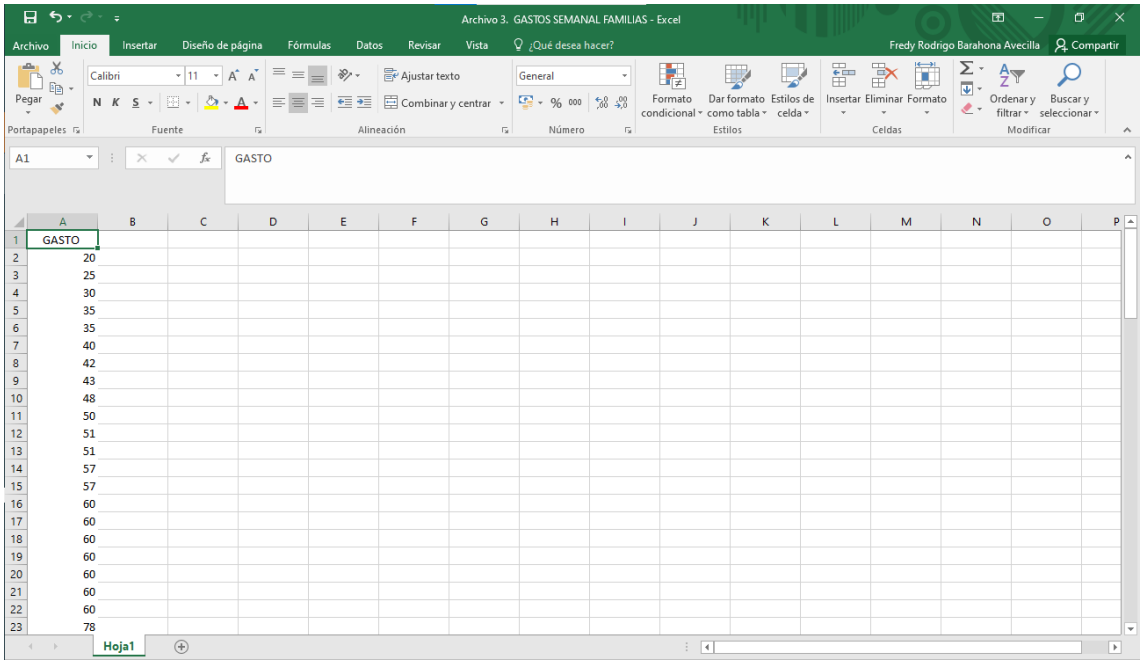
Archivo 3. GASTOS SEMANAL FAMILIAS

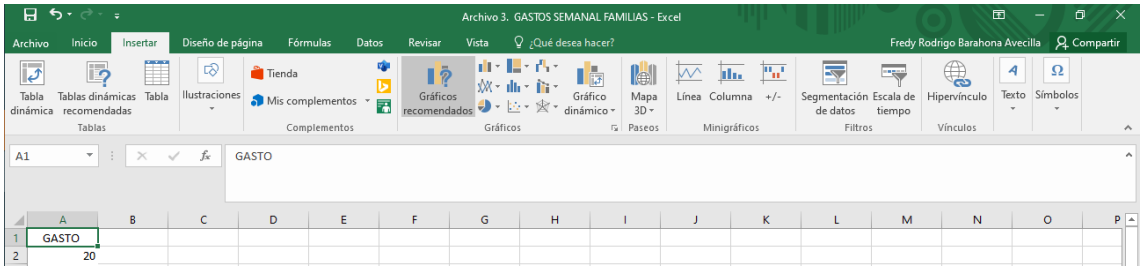
GASTO EN $	N° FAMILIAS
0-50	10
50-100	26
100-150	24
150-200	17
200-250	13
250-300	8
300-350	2
TOTAL	100

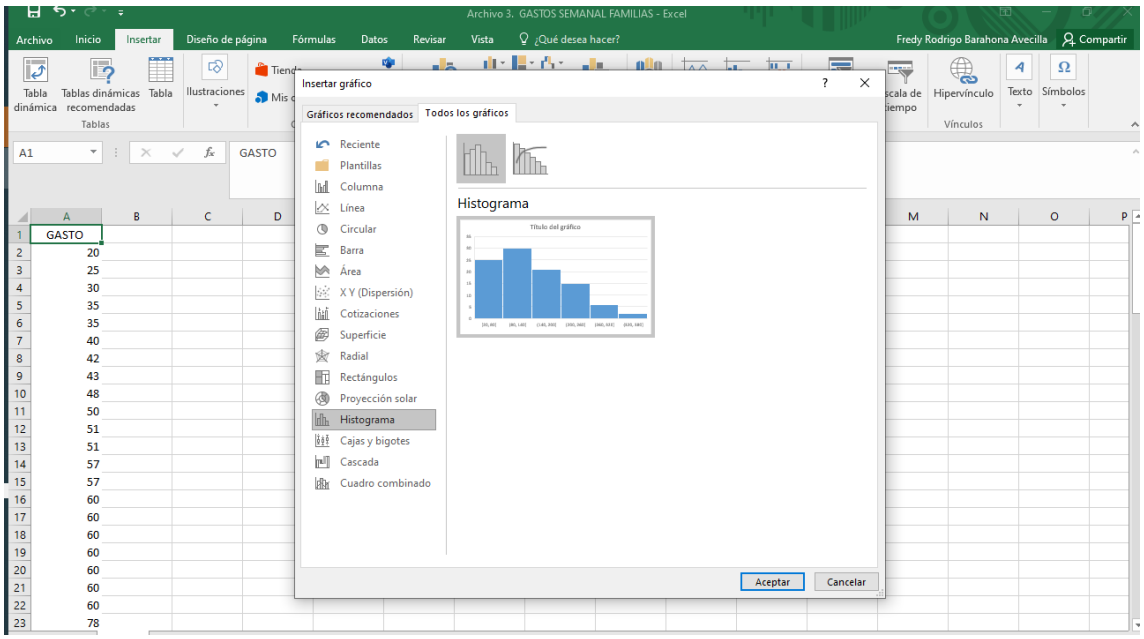
Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

Ilustración 5.

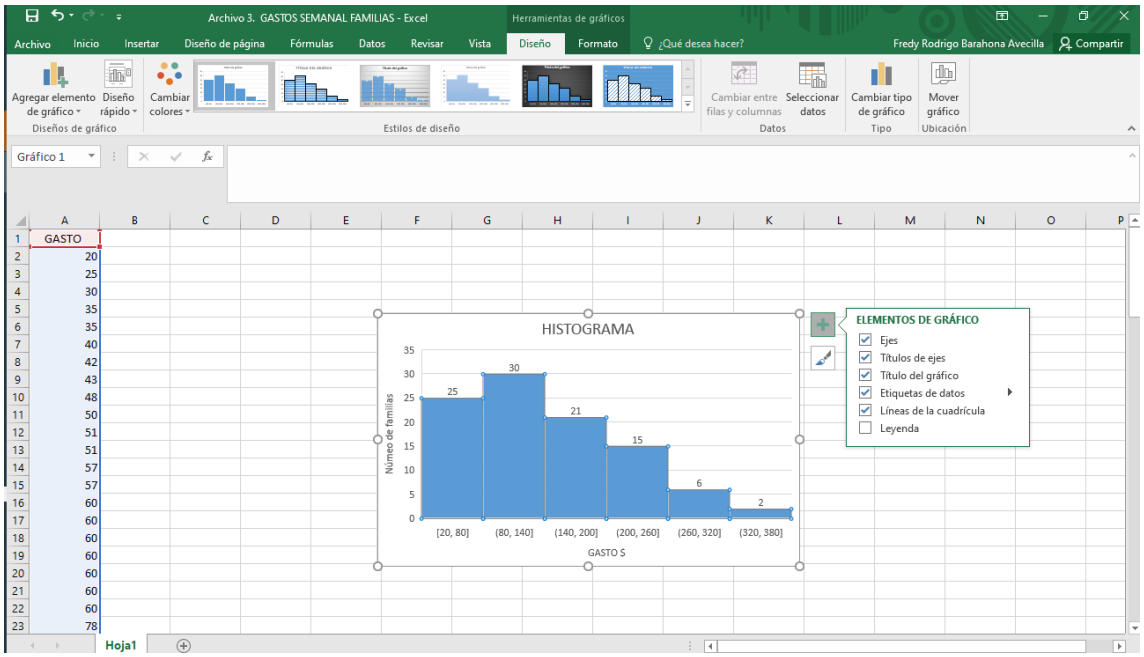
Histograma de Nª de familias (Excel)

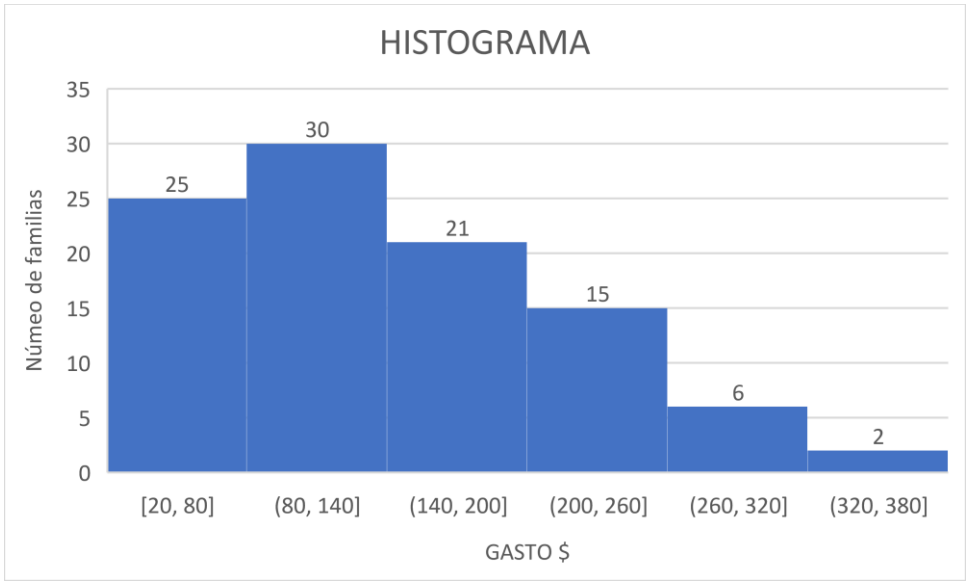




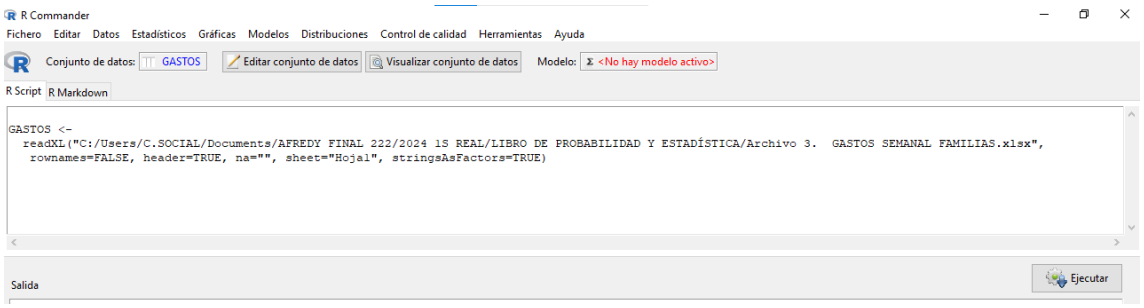


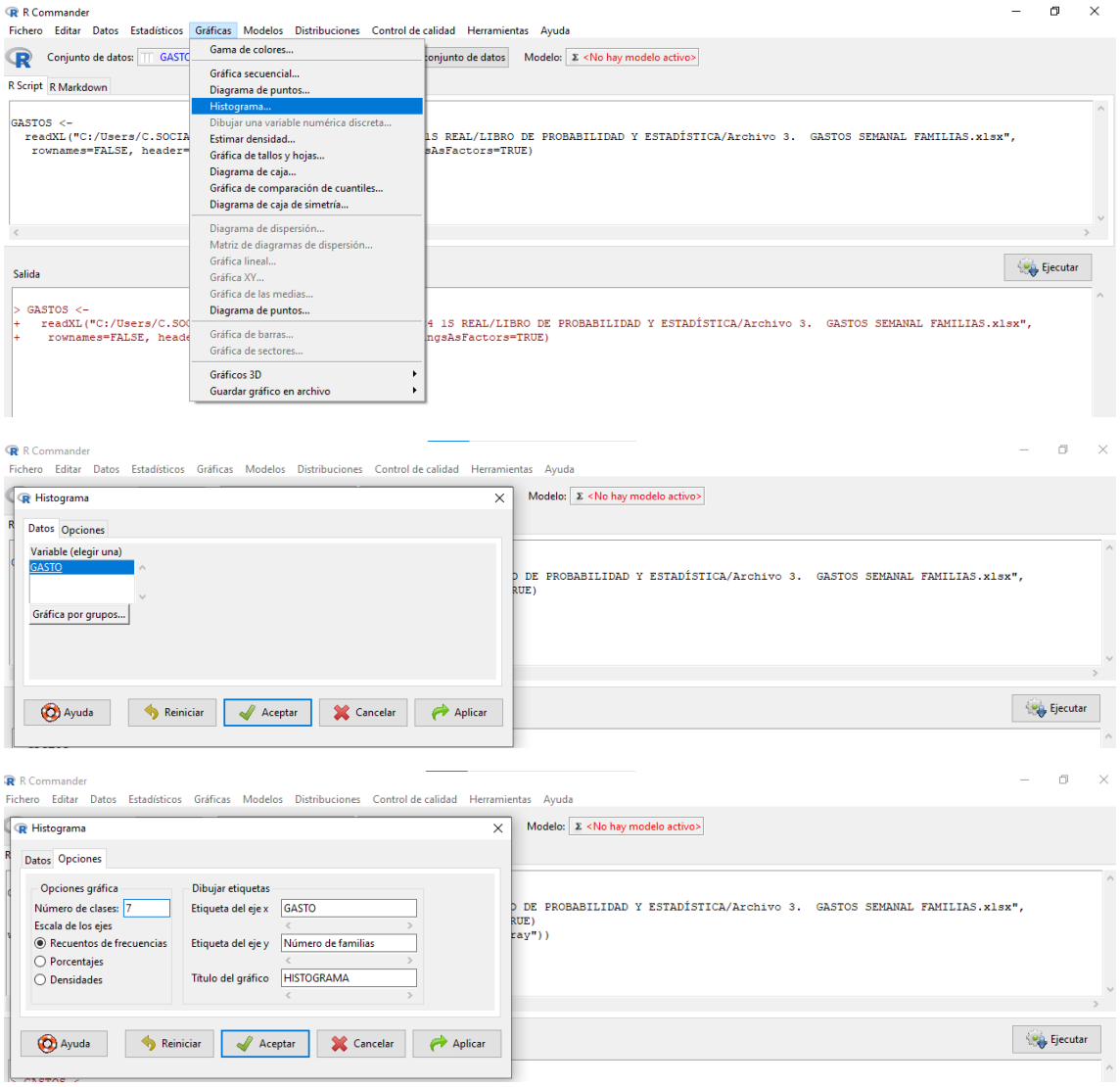
Excel por defecto realizó el histograma con 6 intervalos de clase

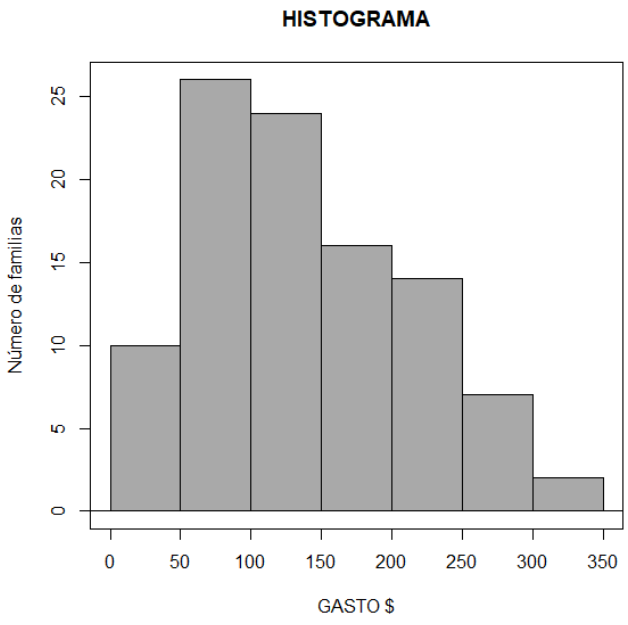




Histograma de Nª de familias (R Commander)







Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

1.3.2. Diagrama de caja

En este gráfico, una caja ocupa el centro, con sus límites marcados por el primer y tercer cuartil, y la mediana representada como una línea divisoria en el medio de la caja. Los "bigotes", que son extensiones de la caja, están conectados por un segmento que atraviesa la caja y proporciona una estimación visual del rango de los datos.

Ejemplo: A continuación, se relacionan las edades de una muestra de usuarios de un centro de rehabilitación fisioterapéutica:

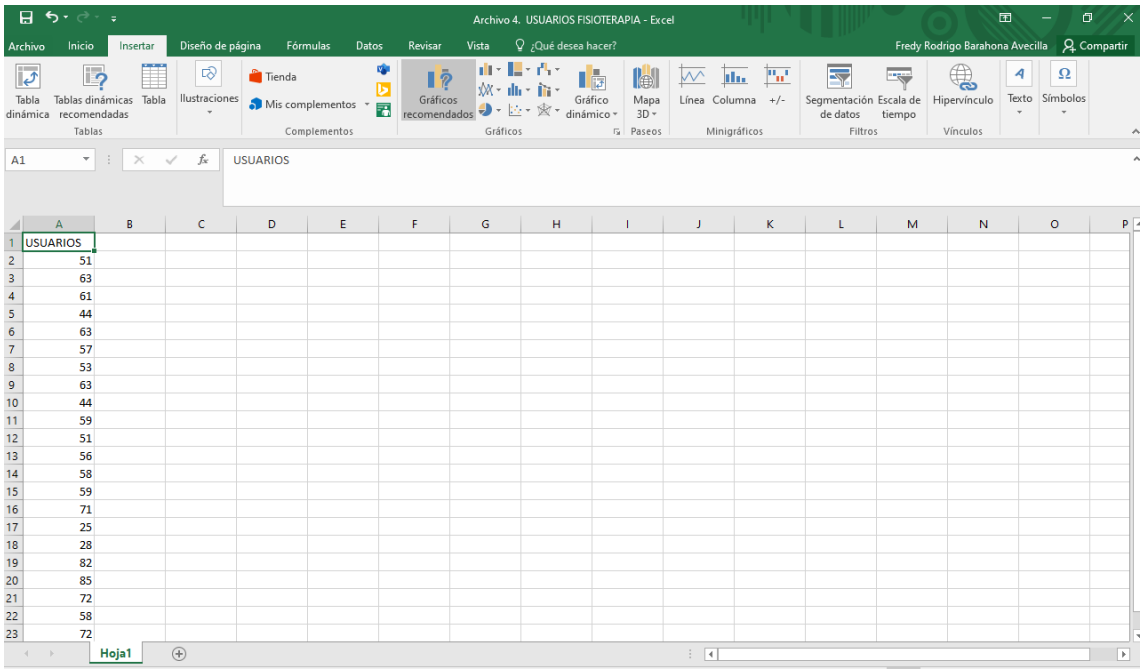
51	63	61	44	63	57
53	63	44	59	51	56
58	59	71	25	28	82
85	72	58	72	58	

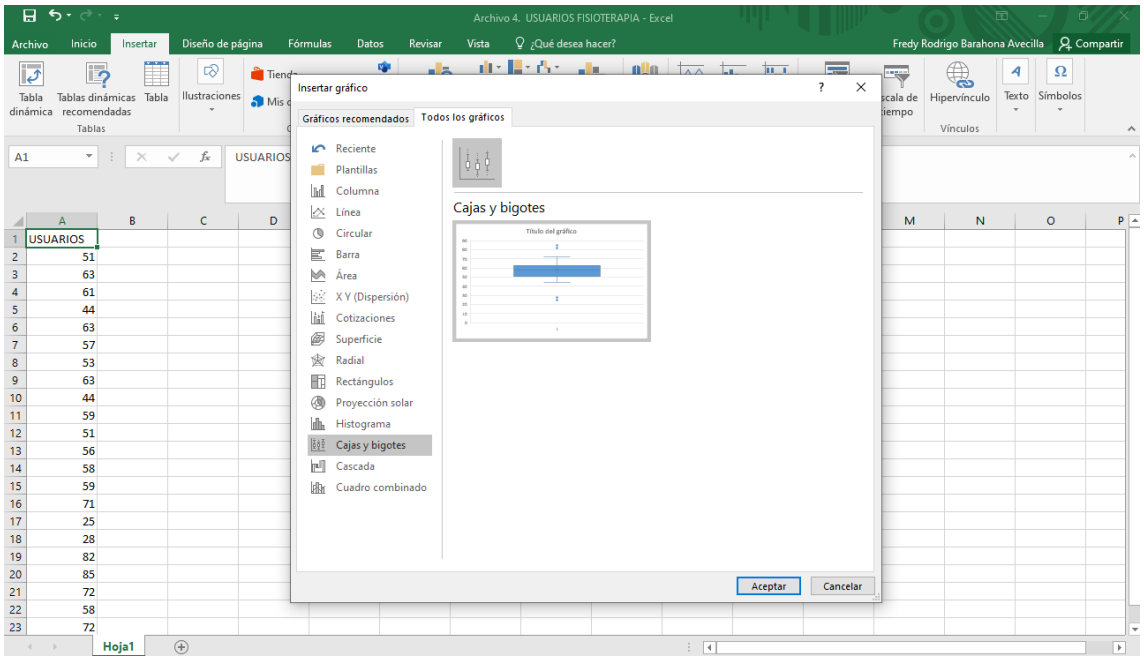
Representa mediante un diagrama de cajas estos datos.

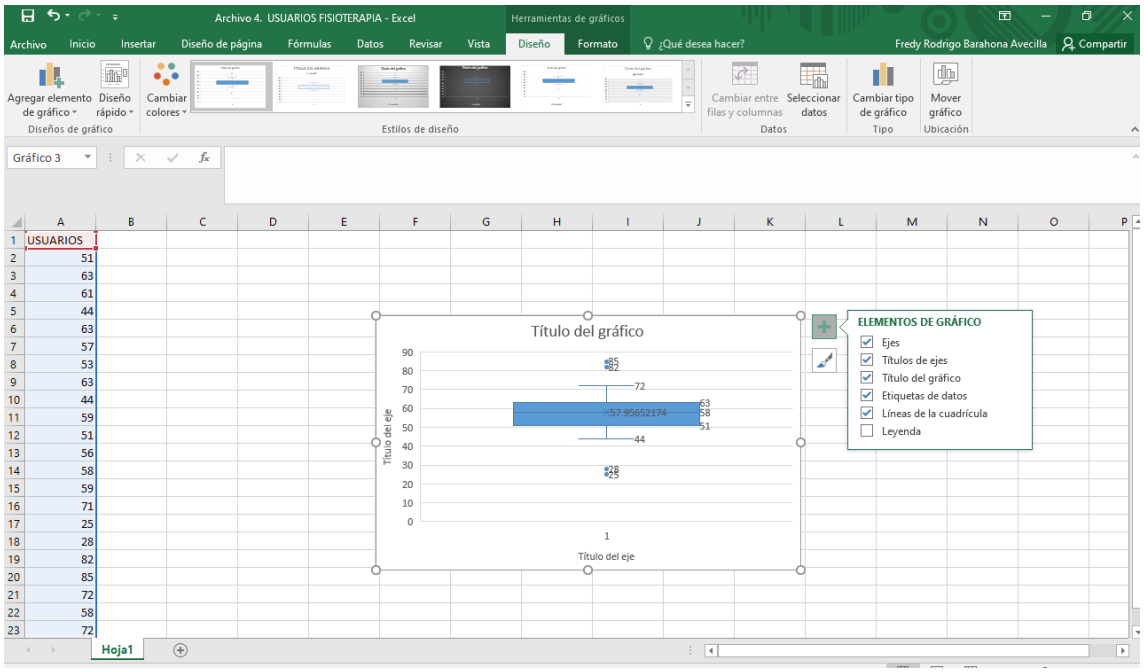
Tabla 8

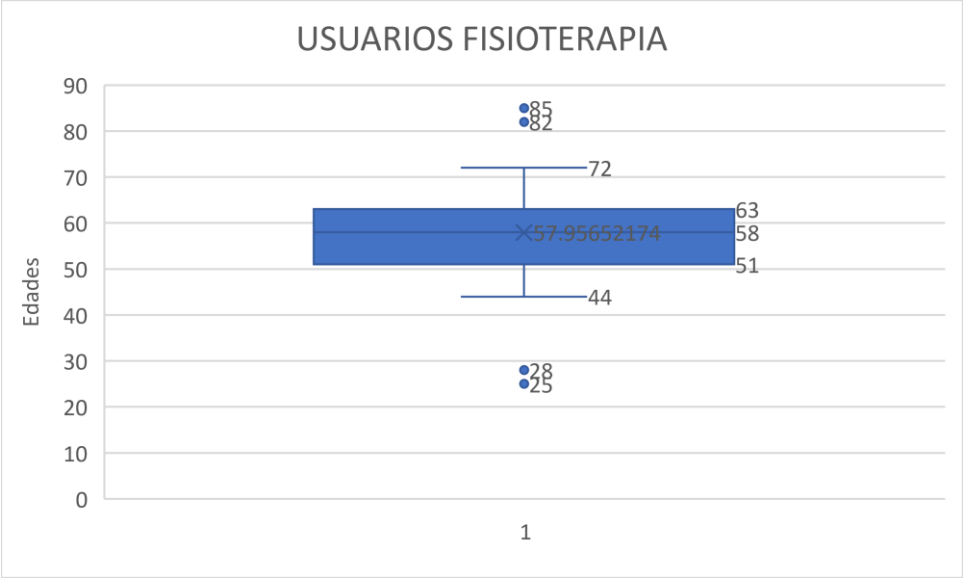
Diagrama de Cajas (Excel)

Archivo 4. USUARIOS FISIOTERAPIA




Archivo 4. USUARIOS FISIOTERAPIA - Excel
Insertar gráfico
Gráficos recomendados
Todos los gráficos
Reciente
Plantillas
Columna
Línea
Circular
Barra
Área
X Y (Dispersión)
Cotizaciones
Superficie
Radial
Rectángulos
Proyección solar
Histograma
Cajas y bigotes
Cascada
Cuadro combinado
Cajas y bigotes
Aceptar
Cancelar
USUARIOS
Hoja1


Archivo 4. USUARIOS FISIOTERAPIA - Excel
Herramientas de gráficos
Título del gráfico
Título del eje
Título del eje
ELEMENTOS DE GRÁFICO
Ejes
Títulos de ejes
Título del gráfico
Etiquetas de datos
Líneas de la cuadrícula
Leyenda
USUARIOS
Hoja1



Fuente: Elaboración propia

En primer lugar, ordenaremos la muestra y calcularemos los estadísticos que necesitamos:

25 28 44 44 51 51 53 56 57 58 58 58 59 59 61 63 63 63 71 72 72 82 85

$$Mínimo = 25$$

$$P25 = Q1 = 51$$

$$Mediana = P50 = Q2 = 58$$

$$P75 = Q3 = 63$$

$$Máximo = 85$$

$$RIC = 63 - 51 = 12$$

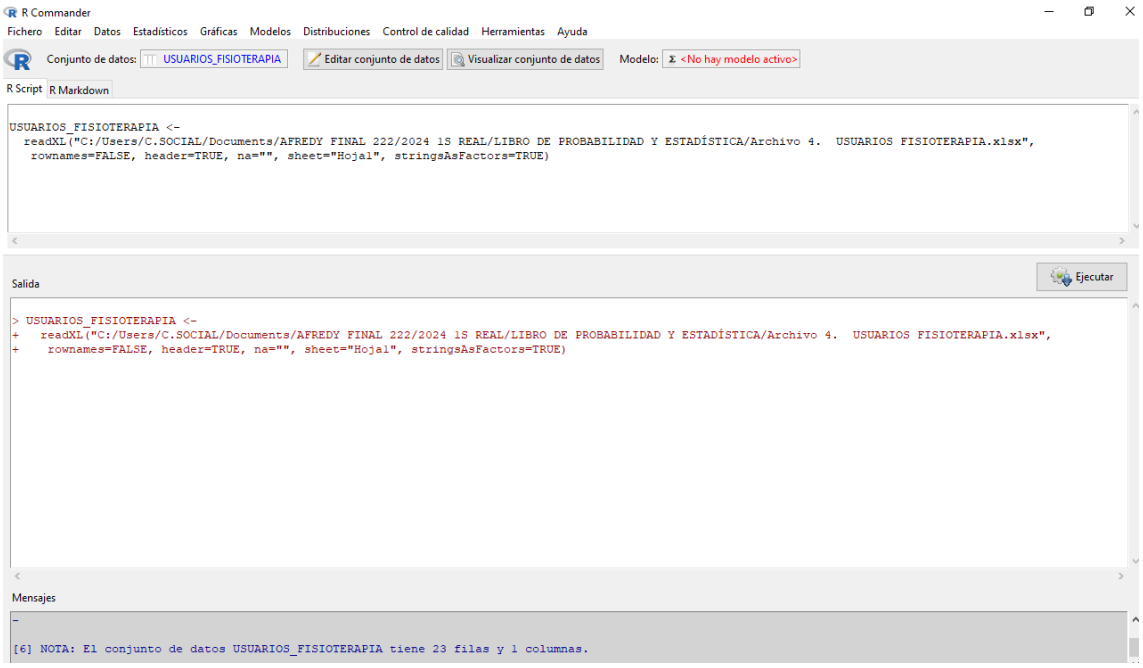
Intervalo que determinara los valores atípicos:

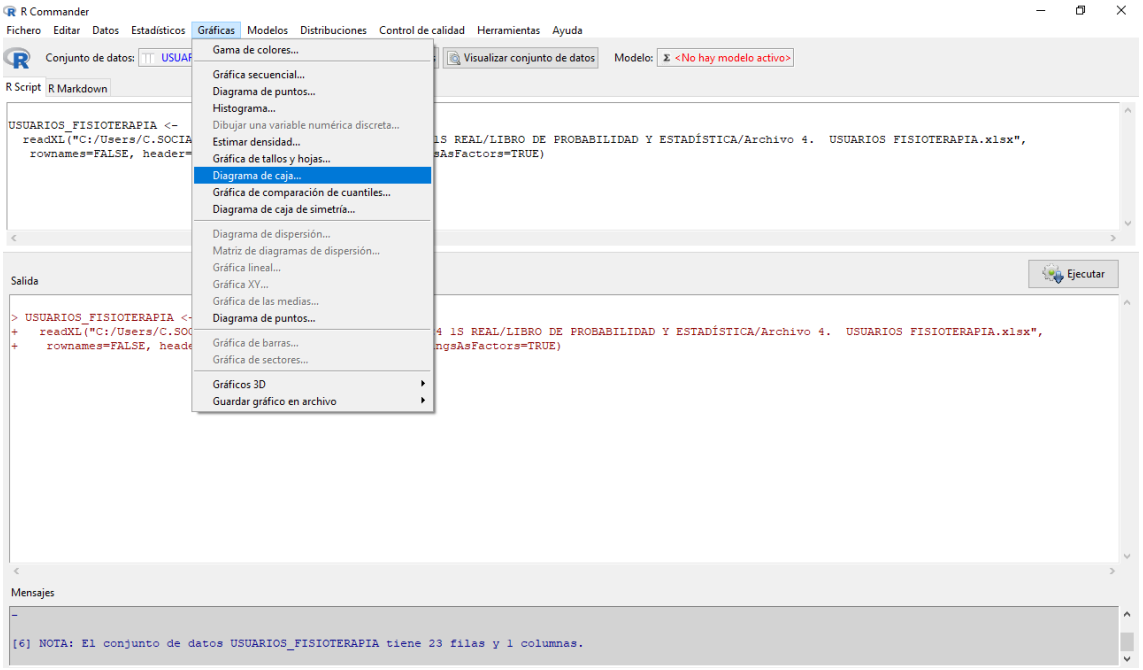
(51 – 1,5 * 12,63 + 1,5 * 12) = (33,81)

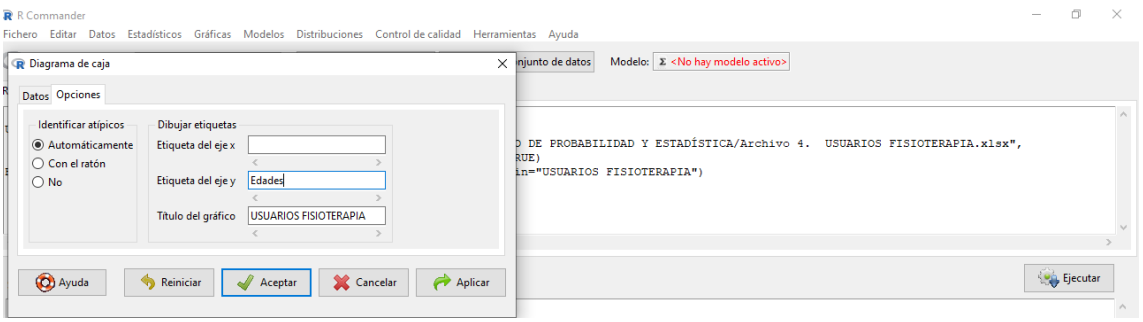
Hay dos valores en la muestra que son atípicos por defecto (el 25 y el 28) y otros dos valores que son atípicos por exceso (el 82 y el 85). Por tanto, los bigotes los representarán el valor 33 y el 81 y los valores atípicos aparecerán en la representación gráfica como puntos aislados. (Botella Rocmora, Alacreu Garci, & Martinez Beneito , 2014)

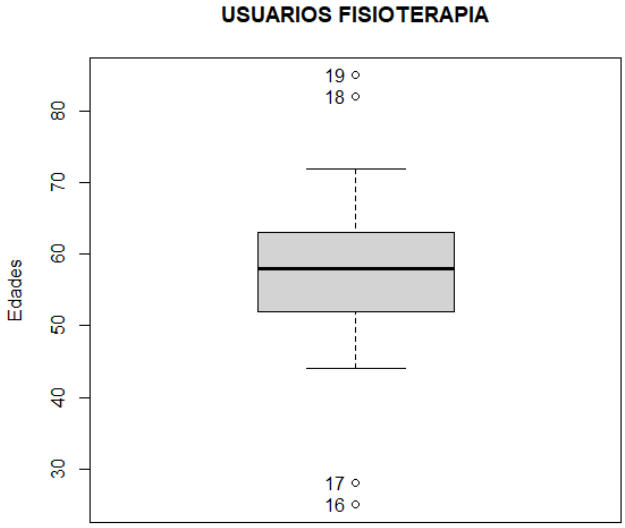
Diagrama de Cajas (R Commander)

Archivo 4. USUARIOS FISIOTERAPIA









Fuente: Elaboración propia

1.3.3. Diagrama de sectores

Es un tipo de representación gráfica que muestra información mediante un círculo dividido en secciones, cada una representando una categoría de distribución. Es útil para variables cualitativas o discretas. En Excel, hay varias opciones para crear este gráfico, como presentarlo en dos dimensiones o en perspectiva, mostrándose como un disco. También puede representarse como anillos o coronas circulares, aunque básicamente son equivalentes. Cuando hay muchas categorías, es preferible utilizar un gráfico circular con subgráficos de barras. (Del Castillo Galarza & Salazar Pinto, 2018)

Tabla 9.

Relación entre del Tipo de Servicio y Nº Personas

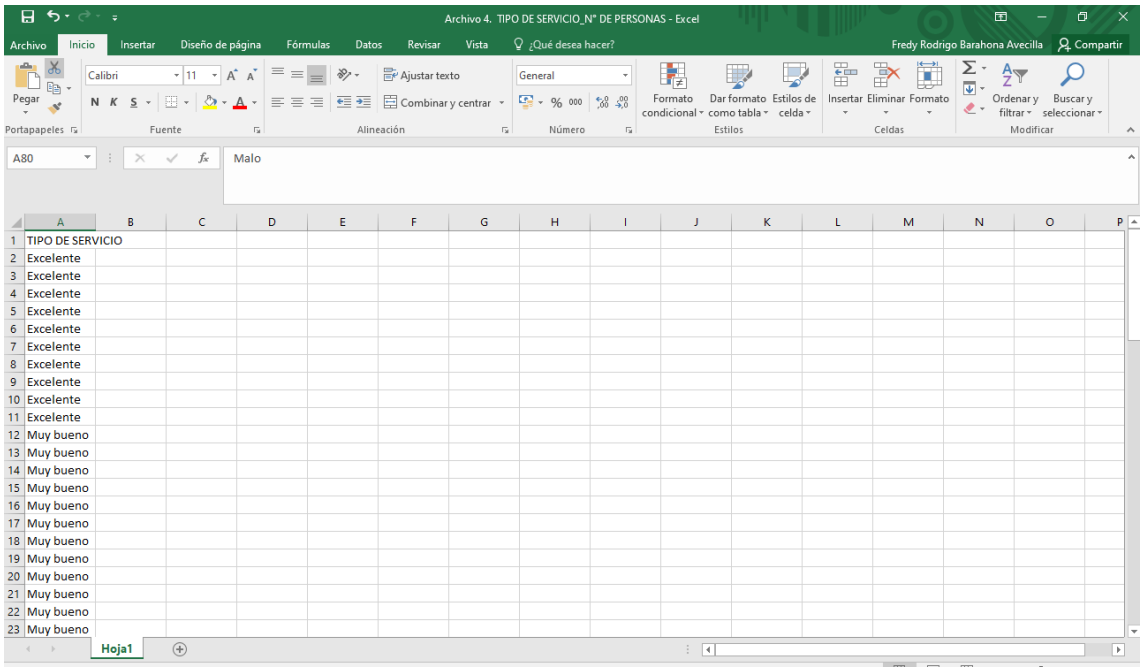
TIPO DE SERVICIO	N° DE PERSONAS
Excelente	10
Muy bueno	15
Bueno	20
Regular	18
Malo	17
Total	80

Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

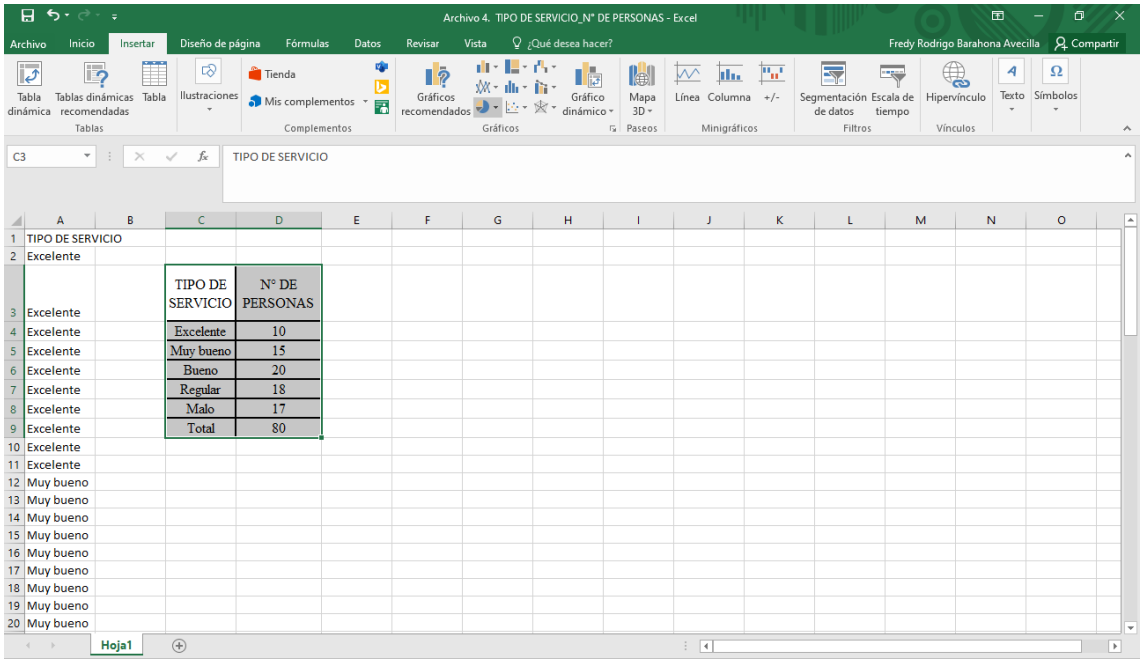
Ilustración 6.

Diagrama circular o de sectores (Excel)

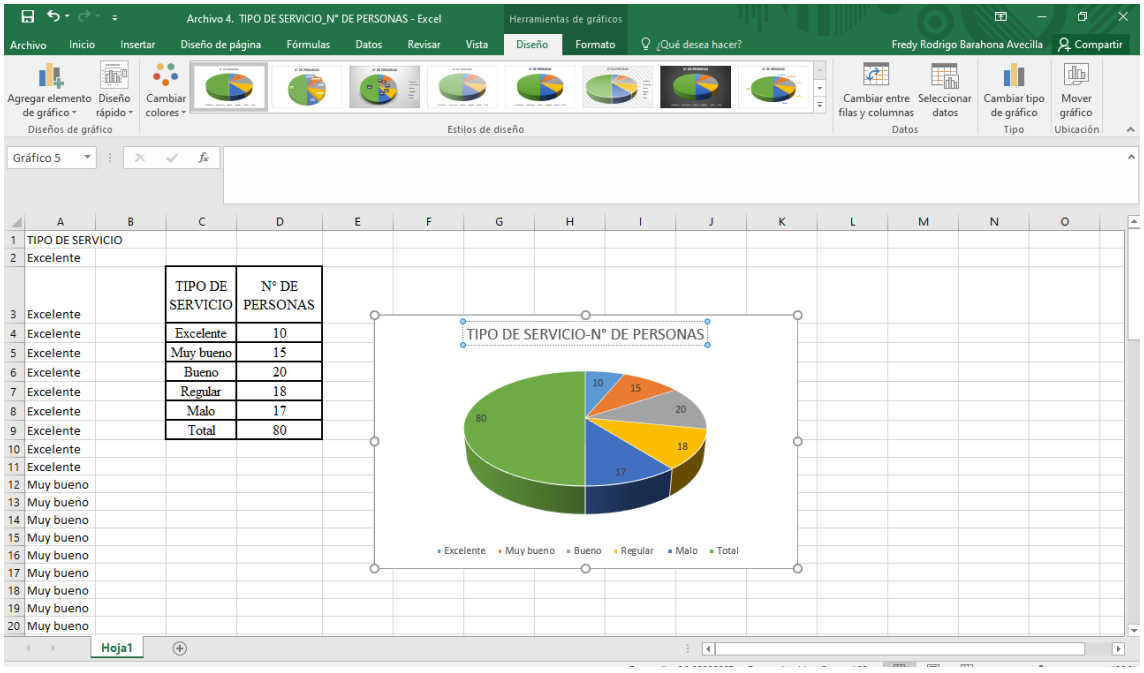
Archivo 5. TIPO DE SERVICIO N° DE PERSONAS

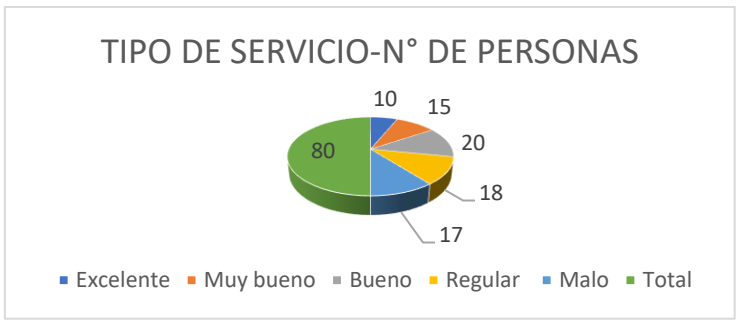


Utilizar la función Insertar-gráfico circular o de anillos



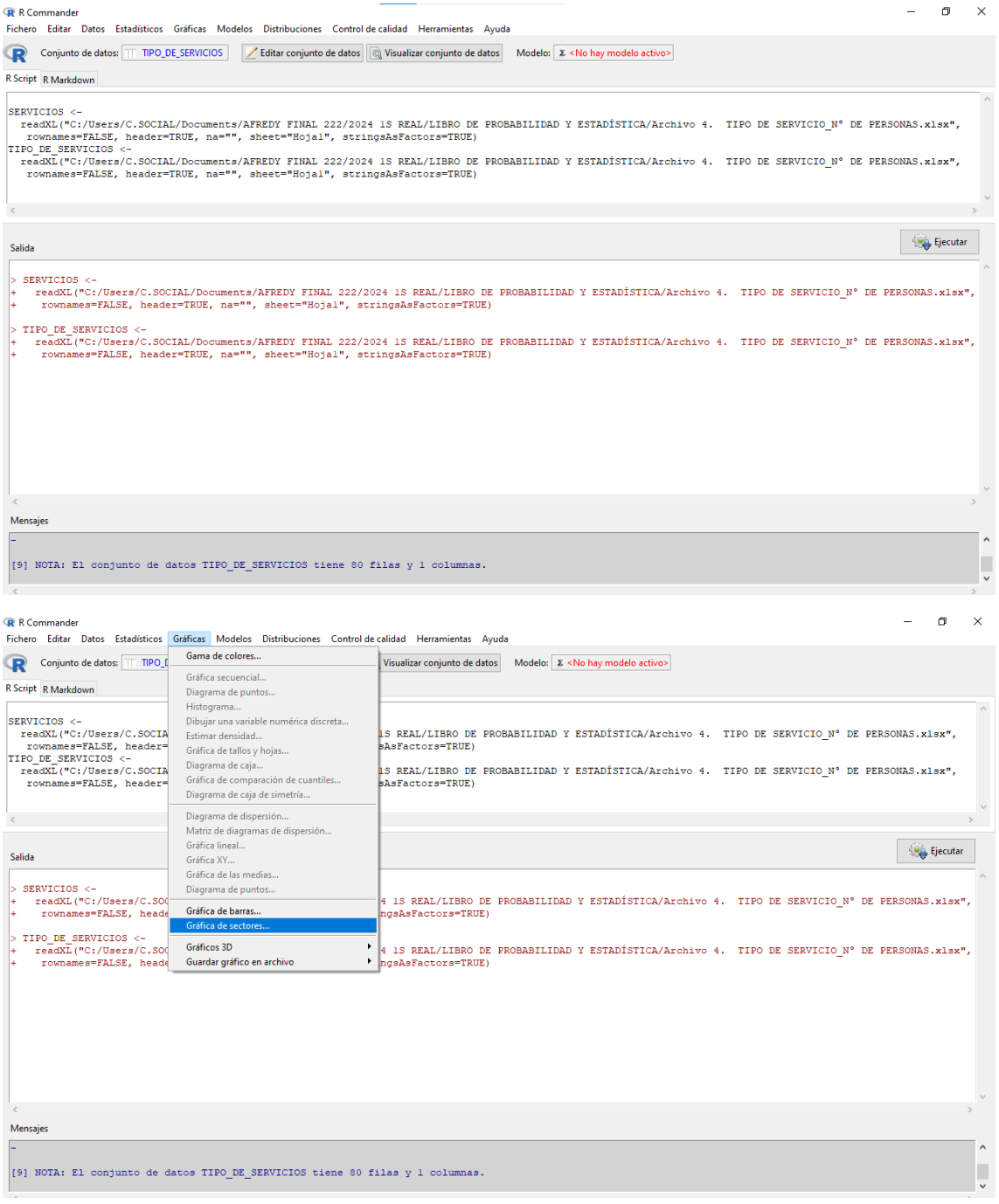
TIPO DE SERVICIO	N° DE PERSONAS
Excelente	10
Muy bueno	15
Bueno	20
Regular	18
Malo	17
Total	80

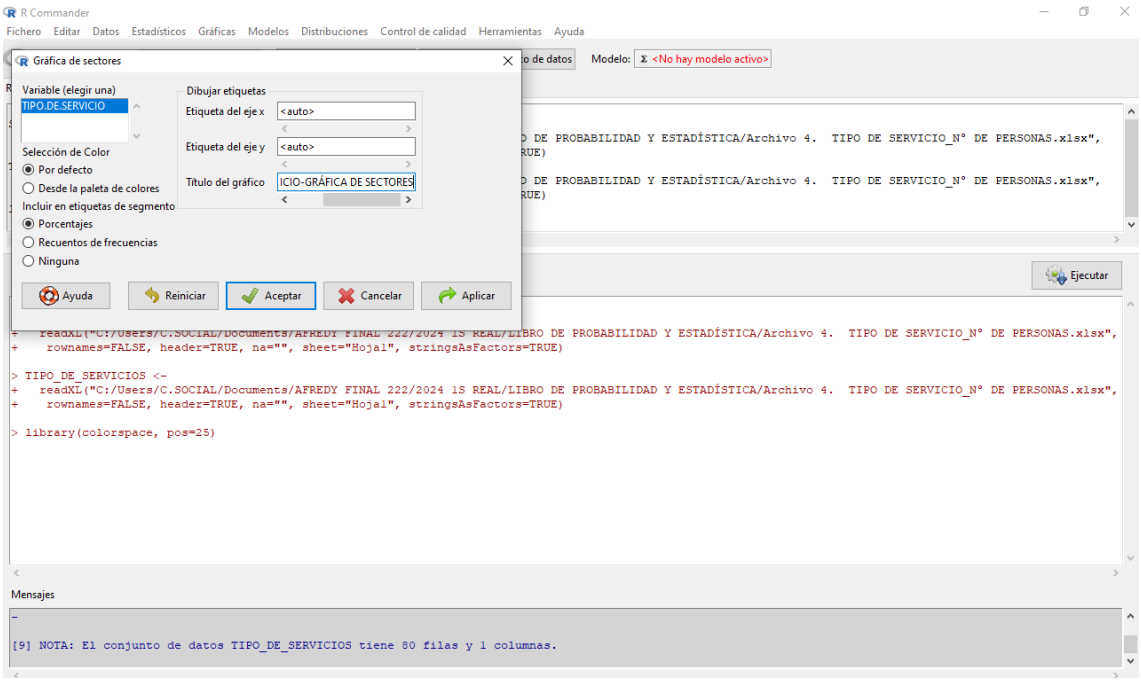


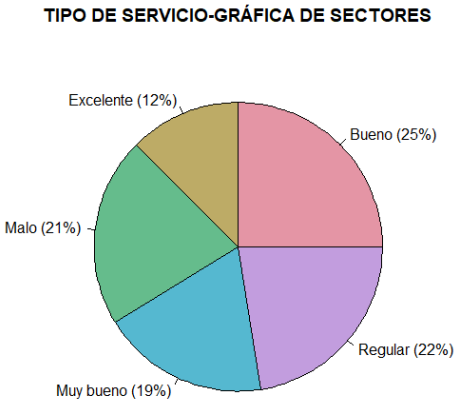


Fuente: Elaboración propia

Diagrama circular o de sectores (R Commander)







Fuente: Elaboración propia

1.3.4. Diagrama de barras

Se trata de un tipo de gráfico que consta de dos ejes, donde uno se elige para representar la variable de estudio según la distribución de frecuencias generada, mientras que el otro muestra la frecuencia de cada categoría. Si se utiliza el eje vertical para las frecuencias, el gráfico será de columnas; si se usa el eje horizontal, será de barras. En este tipo de gráfico, la altura de las columnas o la longitud de las barras reflejan la frecuencia de cada categoría. Además de estas opciones, también existen variantes como las columnas apiladas y las columnas apiladas al 100%, cuyo uso específico recomendamos al lector investigar. (Del Castillo Galarza & Salazar Pinto, 2018)

Ejemplo: Se ha realizado un estudio sobre la provincia de nacimiento de una muestra de 60 personas, obteniéndose la siguiente distribución:

Tabla 10

Estudio sobre la provincia de nacimiento de una muestra de 60 personas

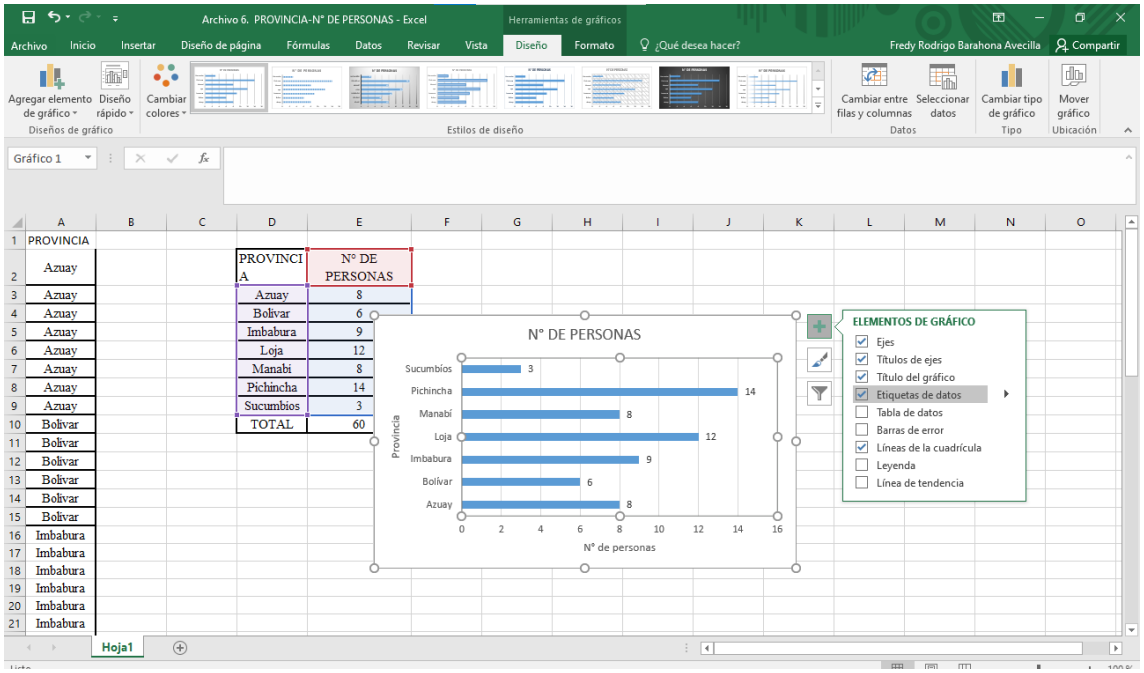
PROVINCIA	N° DE PERSONAS
Azuay	8
Bolívar	6
Imbabura	9
Loja	12
Manabí	8
Pichincha	14
Sucumbíos	3
TOTAL	60

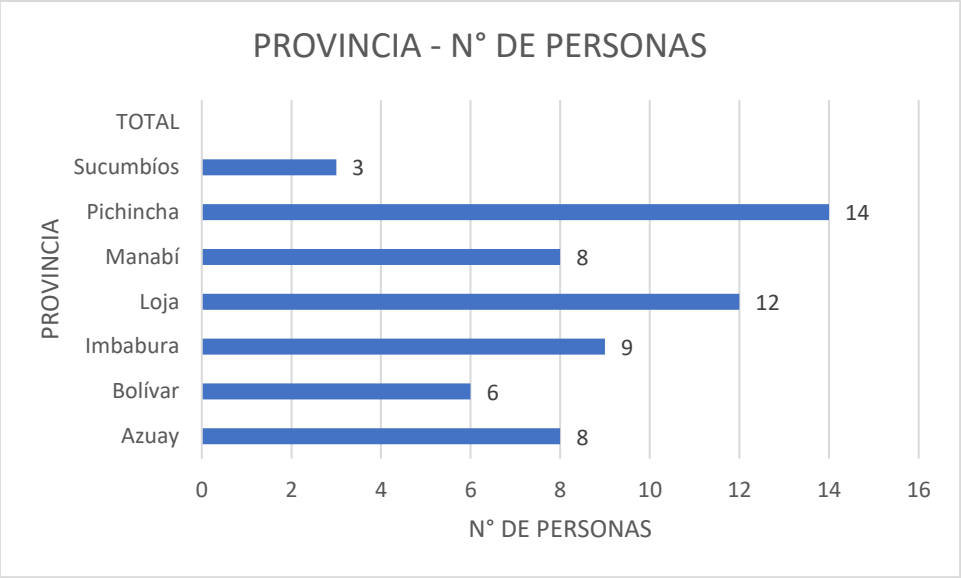
Fuente: (Del Castillo Galarza & Salazar Pinto, 2018)

Dado que la variable de estudio (provincia de nacimiento) es de naturaleza cualitativa, se optará por utilizar un diagrama de columnas para representar este estudio y su correspondiente gráfico. Sin embargo, si se representan los mismos datos utilizando barras, se obtendrá el mismo gráfico.

Estudio sobre la provincia de nacimiento de una muestra de 60 personas (Excel)

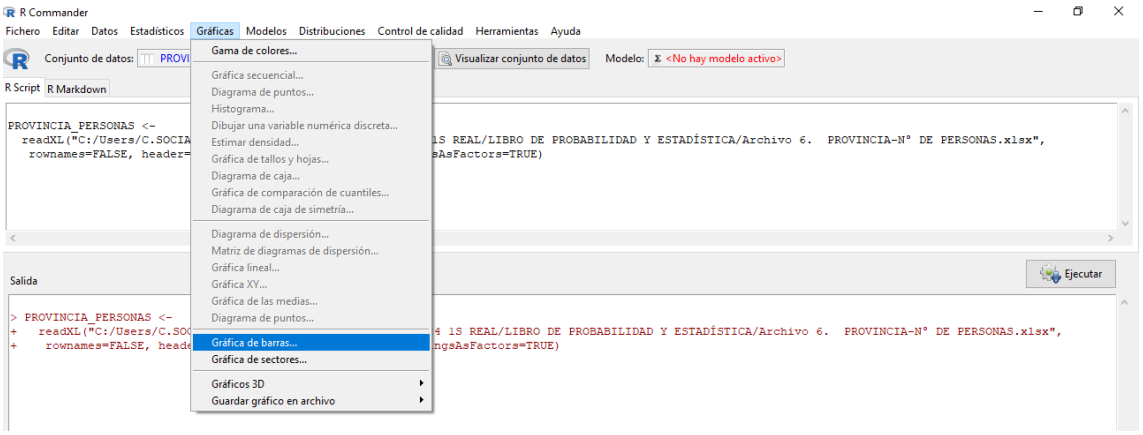
Archivo 6. PROVINCIA-N° DE PERSONAS

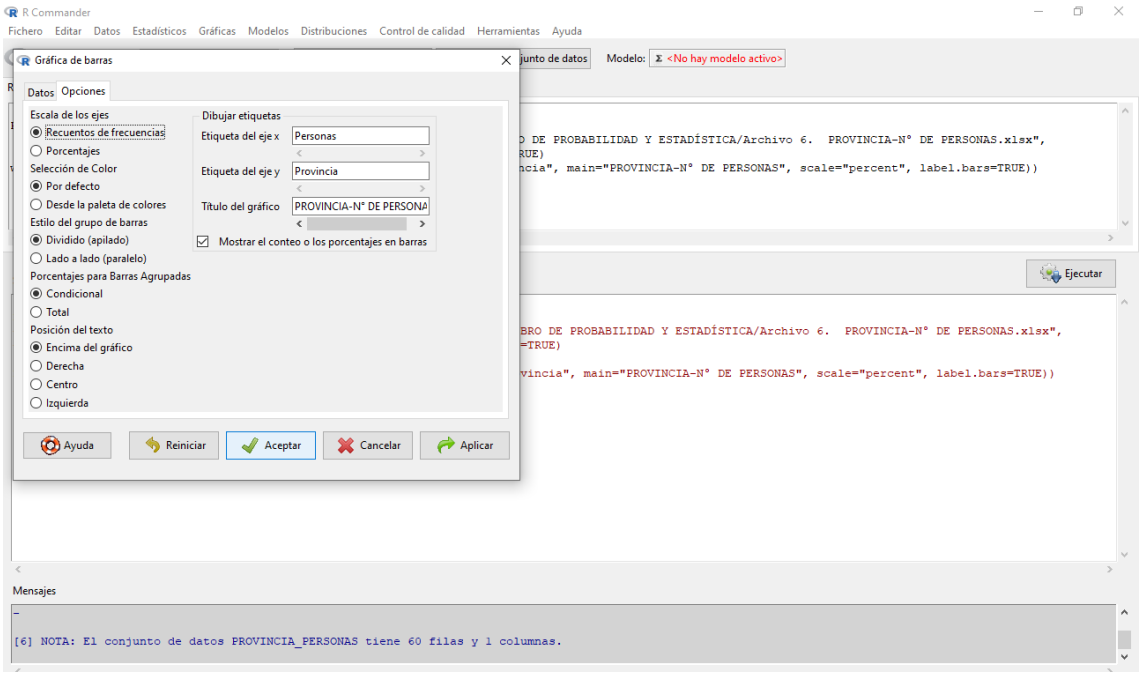


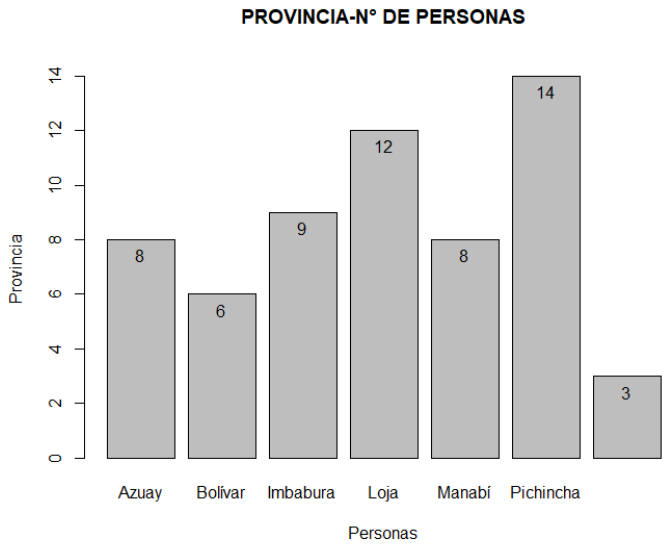


Fuente: Elaboración propia

Estudio sobre la provincia de nacimiento de una muestra de 60 personas (R Commander)


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Conjunto de datos:
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
R Script R Markdown
Gama de colores...
Gráfica secuencial...
Diagrama de puntos...
Histograma...
Dibujar una variable numérica discreta...
Estimar densidad...
Gráfica de tallos y hojas...
Diagrama de caja...
Gráfica de comparación de cuantiles...
Diagrama de caja de simetría...
Diagrama de dispersión...
Matriz de diagramas de dispersión...
Gráfica lineal...
Gráfica XY...
Gráfica de las medias...
Diagrama de puntos...
Gráfica de barras...
Gráfica de sectores...
Gráficos 3D
Guardar gráfico en archivo
Salida
Ejecutar


R Commander
Gráfica de barras
Datos Opciones
Escala de los ejes
Recuentos de frecuencias
Porcentajes
Selección de Color
Por defecto
Desde la paleta de colores
Estilo del grupo de barras
Dividido (apilado)
Lado a lado (paralelo)
Porcentajes para Barras Agrupadas
Condicional
Total
Posición del texto
Encima del gráfico
Derecha
Centro
Izquierda
Dibujar etiquetas
Etiqueta del eje x Personas
Etiqueta del eje y Provincia
Título del gráfico PROVINCIA-N° DE PERSONA
Mostrar el conteo o los porcentajes en barras
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar
Ejecutar
Mensajes
[6] NOTA: El conjunto de datos PROVINCIA_PERSONAS tiene 60 filas y 1 columnas.


PROVINCIA-N° DE PERSONAS
Provincia
0
2
4
6
8
10
12
14
8
6
9
12
8
14
3
Azuay
Bolívar
Imbabura
Loja
Manabí
Pichincha
Personas

Fuente: Elaboración propia

1.4. Medidas de Tendencia Central. Medidas de dispersión. Medidas de posición no central. Medidas de forma.

En los análisis estadísticos, es fundamental examinar la información relacionada con variables cualitativas y cuantitativas mediante la tabulación y la representación gráfica de los datos. Además, implica analizar los datos mediante cálculos matemáticos que resuman el comportamiento de las características del objeto de estudio.

En la mayoría de los casos, los conjuntos de datos obtenidos, ya sea de una muestra o de una población, tienden a agruparse alrededor de un valor central. De este modo, es posible obtener un valor típico o representativo de todo el conjunto de datos, conocido como medida de tendencia central. Entre las medidas de tendencia central más relevantes se encuentran la media aritmética, la mediana y la moda.

1.4.1. Media aritmética, mediana y moda

a) La media aritmética

La media aritmética es la medida de tendencia central más comúnmente empleada y la más representativa en los análisis estadísticos. Esta medida representa el promedio del conjunto de datos de la muestra. Se calcula sumando todos los valores de los datos y dividiendo el resultado entre el número total de datos que conforman la muestra. Si la variable de estudio se denota como X, la media aritmética se representa como $\overline{X}$. (Posada Hernandez, 2016)

Cuando se dispone de una cantidad limitada de datos y estos no han sido organizados en clases o intervalos, se calcularía la media aritmética como:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde

$\overline{X}$: media aritmética de la muestra

n : total de datos de la muestra

x_i : dato de la variable

$\sum_{i=1}^{n} x_i$: suma de todos los valores de la muestra

Por ejemplo, sea X el tiempo que tarda en horas un grupo de 4 estudiantes en realizar una actividad, cuyos valores son: 2,4,3 y 5.

La media aritmética es $X = \frac{2+4+3+5}{4} = \frac{14}{4} = 3{,}5\ horas$

En este caso, el tiempo promedio que tardo el grupo de estudiantes en realizar la actividad fue 3,5 horas. (Posada Hernandez, 2016)

Imaginemos que hemos escogido una muestra de 30 estudiantes para determinar su peso en kilogramos; para simplificar, hemos redondeado las cifras.

Tabla 11.

PESO-ALUMNOS				
74	67	94	70	69
61	71	79	47	85
82	55	65	88	52
58	76	57	72	66
48	56	63	71	60
64	68	83	74	92

Fuente: (Martínez Bencardino, 2012)

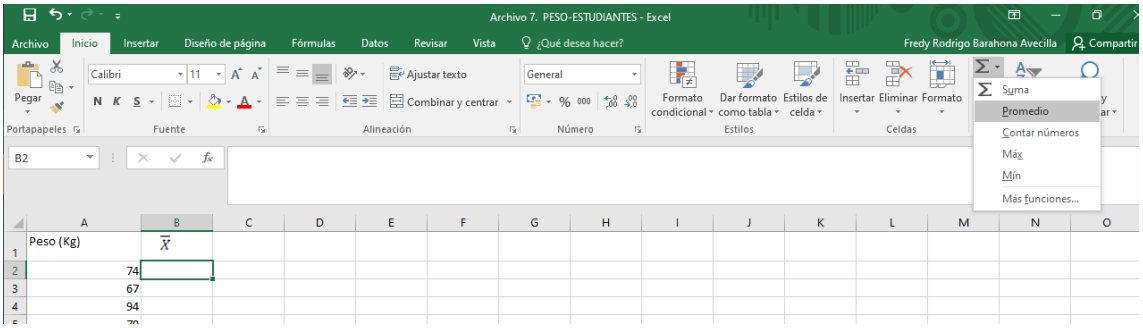
Estudio sobre el peso una muestra de 30 estudiantes (Mediante la fórmula)

Archivo 7. PESO-ESTUDIANTES

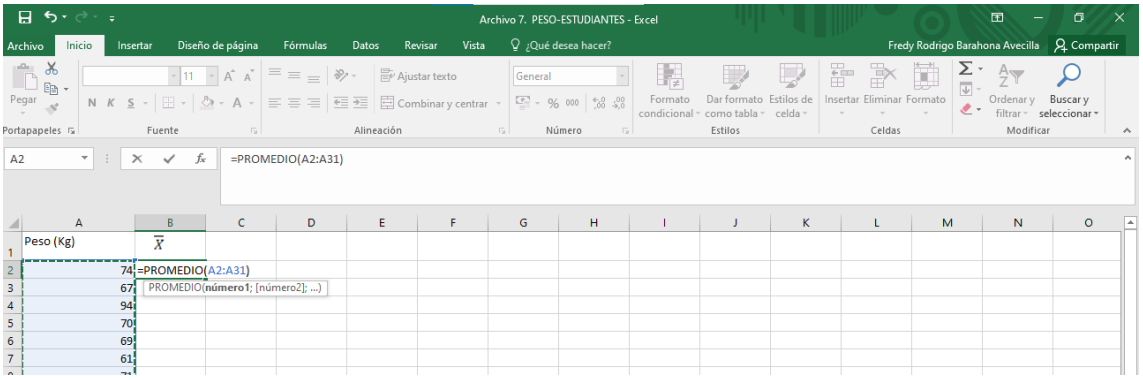
x_i	$\sum_{i=1}^{n} x_i$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
74	2067	69,7
61		
82		
58		
48		
64		
67		
71		
55		
76		
56		
68		
94		
79		
65		
57		
63		
83		
70		
47		
88		
72		
71		
74		
69		
85		
52		
66		
60		
92		

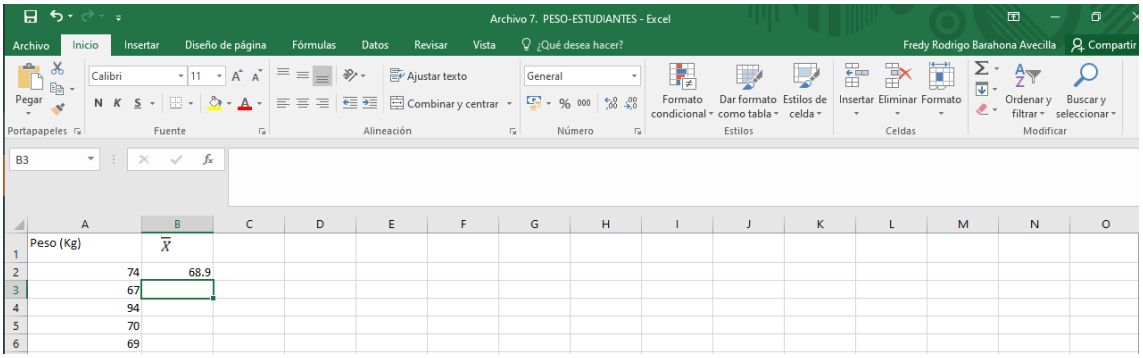
Estudio sobre el peso una muestra de 60 estudiantes (Excel)

Ubicar la celda en la cual se quiere el valor de la media y click en promedio

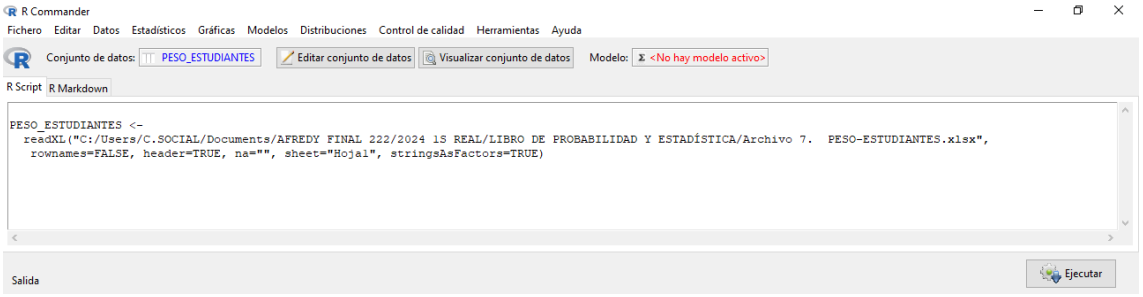


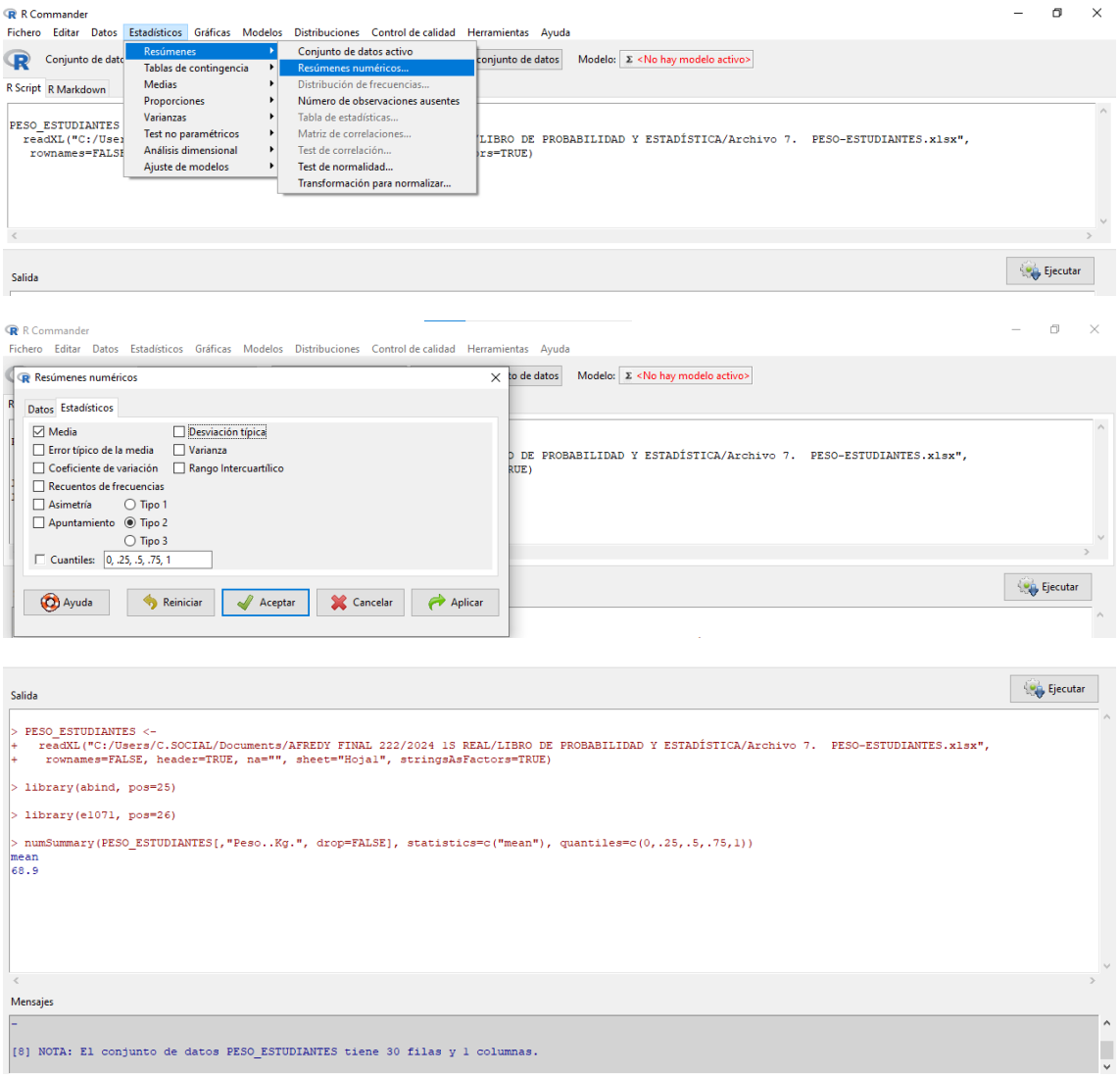
Marcar el conjunto de datos y Enter.





Estudio sobre el peso una muestra de 30 estudiantes (R Commander)





La media es 68,9 Kg

Cuando los datos se organizan en una tabla de frecuencias sin crear intervalos, se determina la media aritmética utilizando la siguiente fórmula:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i \,.\, n_i}{n}$$

Donde n_i es frecuencia absoluta para cada valor de la variable.

Por ejemplo, sea X el número de hijos de los empleados de una organización, los cuales se representan en la tabla 13.

Tabla 13.

Número de hijos de los empleados

Número de hijos (x_i)	Frecuencia (n_i)	$x_i.n_i$
0	1	0
1	2	2
2	4	8
3	2	6
4	1	4
	$n = \sum n_i = 10$	$\sum x_i \, . n_i = 20$

Fuente: (Posada Hernandez, 2016)

$$= \frac{20}{10} = 2 \, hijos$$

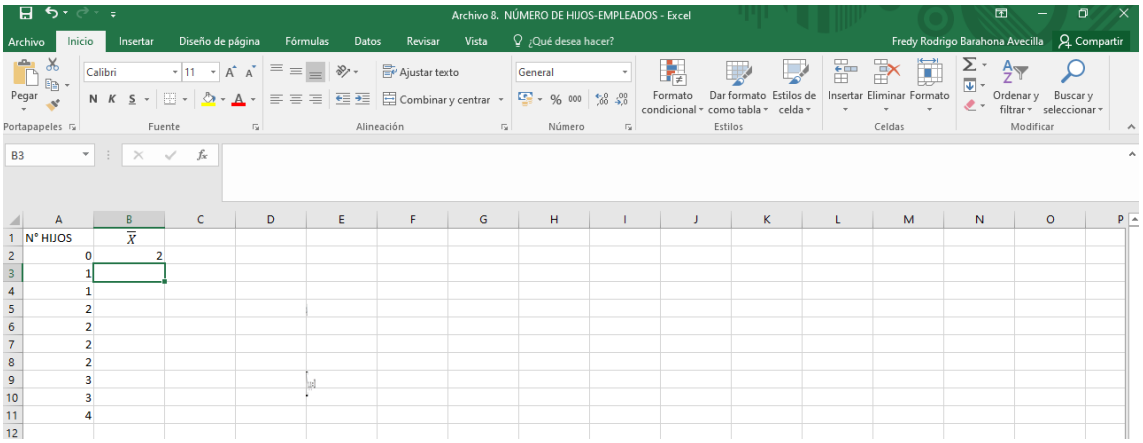
Lo que significa que el promedio de hijos para el grupo de empleados es 2.

Si el conjunto de datos se ha agrupado en intervalos, el cálculo de la media aritmética se realiza mediante la siguiente formula:

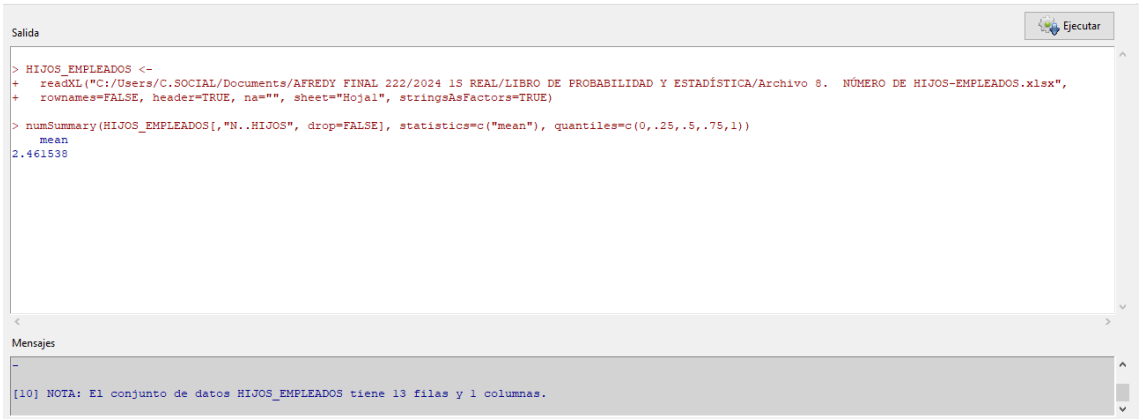
Donde x_i es la marca de clase de cada intervalo.

Estudio del número de hijos de los empleados (Excel)

Archivo 8. NÚMERO DE HIJOS-EMPLEADOS



Estudio del número de hijos de los empleados (R Commander)



En la tabla que sigue, se examina el lapso que un conjunto de individuos emplea en llevar a cabo una tarea; la media aritmética se calculará como:

Tabla 12

Intervalos y frecuencias para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad.

N° de intervalo	Intervalo (Tiempo en minutos)	n_i	h_i (%)	N_i	H_i (%)	X_i
1	$[45-50]$	2	4	2	4	47,5
2	$[50-55]$	9	18	11	22	52,5
3	$[55-60]$	12	24	23	46	57,5
4	$[60-65]$	11	22	34	68	62,5
5	$[65-70]$	9	18	43	86	67,5
6	$[70-75]$	7	14	50	100	72,5

Tabla 13

Media aritmética para el ejemplo de tiempo (en minutos) que tarda un grupo de personas en realizar una actividad.

N° de intervalo	Intervalo (Tiempo en minutos)	x_i	n_i	$x_i n_i$
1	$[45-50]$	47,5	2	95
2	$[50-55]$	52,5	9	472,5
3	$[55-60]$	57,5	12	690
4	$[60-65]$	62,5	11	687.5
5	$[65-70]$	67,5	9	607.5
6	$[70-75]$	72,5	7	507.

$$n = \sum n_i = 50$$

$$\sum x_i n_i = 3060$$

Fuente: (Posada Hernandez, 2016)

Lo que significa que el tiempo promedio que tarda el grupo de personas en realizar la actividad es 61,2 minutos.

b. Mediana

La mediana de un conjunto de datos se define como el valor que se sitúa en el centro, de modo que divide el conjunto en dos partes iguales: el 50% de las observaciones quedan por debajo de este valor y el otro 50% por encima. Para encontrar la posición de la mediana, es necesario ordenar los datos en orden ascendente. La mediana se denota comúnmente como Me. (Posada Hernandez, 2016)

Si el conjunto de datos no se ha agrupado, la posición i de la mediana se ubica según los siguientes criterios:

Cuando el total de datos (n) es impar, la posición de la mediana estará determinada por la fórmula:

$$Me = x_{\frac{n+1}{2}}$$

Mientras que si el total de datos (n) es par, la posición de la mediana estaría determinada por:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Por ejemplo, sea X el número de errores por página cometidos por un grupo de digitadores.

Tabla 14.

Numero de errores por página cometidos por un grupo de digitadores

Digitador	A	B	C	D	E
N° de errores	3	6	4	5	8

Fuente: (Posada Hernandez, 2016)

En primer lugar, es necesario clasificar los datos de manera ascendente, es decir: $3, 4, 5, 6, 8$. Dado que se trata de una muestra, se considera Me como un estimador de la mediana para la población. Esto implica que el tamaño total de la muestra es n = 5 y la posición estimada para la mediana será:

$$Me = x_{\frac{n+1}{2}} = x_{\frac{5+1}{2}} = x_3 = 5$$

Tabla 15.

Nº de errores

N° de errores	3	4	5	6	8
x_i	x_1	x_2	x_3	x_4	x_5

Fuente: (Posada Hernandez, 2016)

Por lo tanto, el estimador de la mediana, denotado como Me, es 5. Esto significa que el 50% de los digitadores registran 5 errores o menos por página, mientras que el otro 50% comete 5 errores o más por página.

Considerando el escenario previo, si se examina otro conjunto de digitadores y se organizan los resultados de manera ascendente, se obtienen los siguientes datos: 5, 5, 7, 9, 11, 13, 13, 15.

(Posada Hernandez, 2016)

Tabla 16

N^a de errores 2

N° de errores	5	5	7	9	11	13	13	13
Posición	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

En este caso, el total de datos es $n = 8$. Al calcular la posición para la mediana se tendrá:

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{\frac{8}{2}} + x_{\frac{8}{2}+1}}{2} = \left(\frac{x_4 + x_5}{2}\right) = \frac{9 + 11}{2} = 10$$

Los valores que corresponden a las posiciones x_4 y x_5 en el conjunto de datos, luego de ser ordenados, son 9 y 11, respectivamente. (Posada Hernandez, 2016)

Es decir, el 50% de los digitadores cometen menos de 10 errores por página, y el otro 50% cometen 10 o más errores por página.

En caso de que el conjunto de datos este agrupado en intervalos, el cálculo de la mediana se realiza mediante el siguiente procedimiento:

a) Hallar $N/2$
b) Ubicar el intervalo cuya frecuencia absoluta acumulada Ni contiene a $N/2$
c) Calcular la mediana mediante la formula

$$Me = l_i + \left(\frac{\frac{N}{2} - N_{i-1}}{n_i}\right) * c$$

Donde:

Li: límite inferior del intervalo que contiene a N/2

N: numero total de datos de la población

Ni-1: frecuencia absoluta acumulada anterior al intervalo que con tiene a N/2

Ni: frecuencia absoluta del intervalo que contiene a N/2

C: amplitud del intervalo que contiene a N/2

Es importante aclarar que cuando el conjunto de datos corresponde a una muestra, la fórmula para la mediana se asume como un estimador y, en consecuencia, el total de datos se representa por n.

Para el ejemplo mencionado anteriormente en la tabla 13, sobre el tiempo que tarda un grupo de personas en realizar una actividad, se tiene la información de la tabla 15. (Posada Hernandez, 2016)

Tabla 17

Cálculo de la mediana para el ejemplo del tiempo (minutos) requerido por un grupo de personas para realizar una actividad

N°. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Fuente: (Posada Hernandez, 2016)

Para calcular el estimador de la mediana, se utilizan los pasos descritos en el enunciado anterior, esto es:

a. El total de personas que realizaron la actividad es 50, por lo tanto, N/2=50/2=25 personas.
b. Al analizar la frecuencia absoluta acumulada, se encuentra que 25 se ubica en el 4º intervalo (no es posible ubicar el valor de 25 en el tercer intervalo, debido a que solo acumula 23 personas).
c. Los datos para el calculo de la mediana serán:

$$l_i = 60$$
$$n/2 = 25$$
$$N_{i-1} = 23$$
$$n_i = 11$$
$$c = 65 - 60 = 5$$

Luego,

$$Me = l_i + \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right) * c = 60 + \left(\frac{25-23}{11} \right) * 5 = 60 + \left(\frac{2}{11} \right) * 5 = 60 + 0.9 = 60,9 \text{ minutos}$$

Significa que el 50% de las personas realizaron la actividad en 60,9 minutos o menos y el otro 50% tardaron más de 60,9 minutos.

En los análisis estadísticos, la medida de tendencia central mas representativa es la media aritmética. Sin embargo, en aquellos casos en los cuales se presentan valores extremos, es preferible usar la mediana en vez de la media, debido a que esta no es afectada por valores extremos y por lo tanto, no es tan sensible como la media aritmética.

Por ejemplo: sea X la edad (en años) de un grupo de personas pertenecientes a un club de actividades lúdicas, estas son: 17,16,17,18,17,16,17,18,35.

Al calcular la media aritmética X se tendría un promedio de 19 años y la mediana Me de 17 años. Sin embargo, al analizar el comportamiento de las edades de los deportistas, se observa que estas tienden a agruparse más alrededor de 17 que a 19 años. Además, la media aritmética se afecta directamente con la presencia del valor extremo de 35 años, mientras que la mediana se mantiene en su valor, independiente de los valores extremos que se presenten en el conjunto de datos. En estos casos, es decir cuando se presentan valores extremos que afectan visiblemente el promedio en el conjunto de datos, se prefiere como medida de tendencia central a la mediana y no a la media aritmética. (Posada Hernandez, 2016)

c. Moda

En la vida diaria, es común oír la frase "está de moda" cuando algo se ve o se presenta con frecuencia. En términos estadísticos, este concepto se refiere a la moda de un conjunto de datos, que es el valor que aparece con mayor frecuencia, ya sea un atributo o un valor específico. La moda, denotada por Mo, se aplica tanto a variables cualitativas como cuantitativas, ya sean discretas o continuas.

Para calcular la moda de un conjunto de datos no agrupados, primero se determinan las frecuencias de cada valor y luego se identifica aquel con la frecuencia más alta. Por ejemplo, si preguntamos a varias personas sobre su color preferido y obtenemos respuestas como blanco, azul, rosa, azul, negro, azul, morado, azul, negro y blanco, podemos construir las frecuencias de cada color para determinar la moda.. (Posada Hernandez, 2016)

Tabla 18

Moda para la variable cualitativa

Color	n_i
Blanco	2
Azul	4

Rozado	1
Negro	2
Morado	1

Fuente: (Posada Hernandez, 2016)

En la tabla 19 se puede observar que el color más frecuente es el azul, por lo tanto, la preferencia de color más común entre las personas es el azul.

En el ejemplo anterior, se presenta una única moda, lo que hace que este conjunto de datos se clasifique como una distribución unimodal. Cuando hay múltiples modas, se denomina distribución multimodal, y si no hay ninguna moda, se llama distribución amodal. Por ejemplo, consideremos el número de cursos matriculados por varios estudiantes en un semestre: $6, 7, 6, 6, 7, 8, 9, 7$. En este caso, hay dos modas: 6 y 7, ya que ambos tienen la misma frecuencia máxima. Por lo tanto, este conjunto de datos se clasifica como una distribución multimodal, específicamente bimodal.

Cuando los datos han sido agrupados en clases o intervalos, la moda se calcula utilizando la ponderación en el intervalo, con el siguiente procedimiento:

a. Ubicar el intervalo (o los intervalos) con mayor frecuencia absoluta ni.
b. Calcular la moda (o las modas) con la fórmula:

$$Mo = l_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta} \right) * C$$

Donde:

l_i : límite inferior del intervalo con mayor frecuencia absoluta

Δ_1: diferencia entre la mayor frecuencia absoluta y la anterior

Δ_2: diferencia entre la mayor frecuencia absoluta y la siguiente

c : amplitud del intervalo con mayor frecuencia absoluta

Al retomar el ejemplo mencionado anteriormente en la tabla 18, sobre el tiempo que tarda un grupo de personas en realizar una actividad, se toma la siguiente información para el cálculo de la moda (ver tabla 20):

Tabla 19.

Moda para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Fuente: (Posada Hernandez, 2016)

i. El intervalo de mayor frecuencia absoluta es el 3.
ii. Los valores para el cálculo de la moda son:

$$\frac{1}{f_i} 55$$

$$\Delta_1 12 - 9 = 3$$

$$\Delta_2 12 - 11 = 1$$

$$c = 60 - 55 = 5$$

Por lo tanto, la moda seria:

$$Mol = {}_i + \left(\frac{\Delta_1}{\Delta_{12} + \Delta}\right) * C = 55 + \left(\frac{3}{3+1}\right) * 5 = 55 + \frac{15}{4} = 58,75 \text{ minutos}$$

Es decir, el tiempo que la mayoría de personas invierten para realizar la actividad es 58,75 minutos.

1.4.2. Varianza, desviación estándar, rango, coeficiente de variación

- **Varianza**

La varianza es una medida de dispersión basada en la diferencia de cada dato con la media aritmética. Posada y Buitrago (2008) plantean que "la diferencia entre cada Xi y el promedio (X para una muestra y μ para una población) se llama desviación respecto al promedio. Para una muestra, la desviación respecto a la media se expresa como (xi – x); para una población es (xi - μ)". Al sumar el total de la desviación respecto al promedio, este tiende a cero por la compensación de las desviaciones positivas (cuando los datos están por encima del promedio), con las desviaciones negativas (cuando los datos están por debajo del promedio). De esta manera, no es posible obtener efectivamente la desviación de los datos respecto del promedio, por lo cual se hace necesario elevar cada

desviación al cuadrado, garantizando así que tolas las desviaciones obtenidas presenten cantidades positivas; el resultado entonces quedara en unidades cuadradas. La varianza es una medida de dispersión basada en la diferencia de cada dato con la media aritmética. Posada y Buitrago (2008) plantean que "la diferencia entre cada X_i y el promedio $\overline{x}$para una muestra y μ para una población) se llama desviación respecto al promedio. Para una muestra, la desviación respecto a la media se expresa como (xi – x); para una población es (xi - μ)". Al sumar el total de la desviación respecto al promedio, este tiende a cero por la compensación de las desviaciones positivas (cuando los datos están por encima del promedio), con las desviaciones negativas (cuando los datos están por debajo del promedio). De esta manera, no es posible obtener efectivamente la desviación de los datos respecto del promedio, por lo cual se hace necesario elevar cada desviación al cuadrado, garantizando así que tolas las desviaciones obtenidas presenten cantidades positivas; el resultado entonces quedara en unidades cuadradas. (Posada Hernandez, 2016)

Cuando se tiene la totalidad de los datos de la población, el promedio de las desviaciones elevadas al cuadrado se denomina varianza poblacional y se representa con la letra del alfabeto griego sigma ((σ^2). Para una población con total de datos N y promedio μ, el parámetro para la varianza se calcula mediante la siguiente ecuación:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

La varianza de la muestra ((s^2) tiene como objetivo convertirse en un estimador de la variación para la población; por tal razón, se define como la suma de las desviaciones elevadas al cuadrado, distribuidas entre el tamaño de la muestra, menos uno. El estimador para la varianza muestral se calcula mediante la siguiente ecuación:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Donde:

$\overline{x}$:media aritmética de la muestra

n :total de los datos de la muestra

x_i :cada dato u observación de la variable X

En relación al denominador (n-1), Posada y Buitrago manifiestan que:

Si el denominador fuera n en lugar de (n-1), se obtendría el promedio de los cuadrados de las diferencias con respecto a la media. Sim embargo, se utiliza (n-1) debido a ciertas propiedades matemáticas deseadas que tiene el estadístico s^2, las cuales lo hacen apropiado para hacer inferencias estadísticas. Al aumentar el tamaño de la muestra, la diferencia entre n y (n-1) disminuye cada vez más.

Al calcular la varianza, los datos se elevan al cuadrado, por tanto, las unidades con las cuales se midieron también se elevan al cuadrado, imposibilitando la interpretación. En consecuencia, en la mayoría de los análisis estadísticos se emplea la varianza como una

medida que permite comprar la dispersión entre dos o más variables, identificando la de mayor varianza como aquella que posee mayor dispersión o variabilidad. La importancia de la varianza esta en que es una medida transitoria para el cálculo de la desviación típica o estándar de un conjunto de datos.

Por ejemplo, en la tabla 21 se presenta la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa. La puntuación es valorada en la escala de 1 a 5. Se requiere conocer la varianza de la calificación de los empleados. (Posada Hernandez, 2016)

Tabla 20.

Varianza para la evaluación de desempeño de siete empleados del área de mercadeo de una empresa

Empleado	Calificación (x_i)	Media de la muestra $\bar{x}$	Desviación $(x_i-\bar{x})$	Desviación al cuadrado $(x_i-\bar{x})^2$
1	3,5	3,6	-0,1	0,01
2	4,5	3,6	0,9	0,81
3	4,2	3,6	0,6	0,36
4	3,0	3,6	-0,6	0,36
5	2,7	3,6	-0,9	0,81
6	3,3	3,6	-0,3	0,09
7	4,0	3,6	0,4	0,16
			$\sum(x_i-\bar{x})=0$	$\sum(x_i-\bar{x})^2=2,6$

Fuente: (Posada Hernandez, 2016)

Luego, la varianza será:

$$s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{2,6}{6}=0,43$$

Nótese que si se interpreta la varianza se estaría diciendo que la variación en la calificación de desempeño de los empleados es de 0,43 puntos cuadrados, lo cual no es lógico. En este sentido, cobra importancia la varianza como medida de transición para la desviación típica o estándar.

Si los datos se agruparon en frecuencias o en intervalos, la varianza puede ser calculada mediante las siguientes formulas:

$$\sigma^2=\frac{\sum x_i^2*n_i}{N}-\mu^2$$ Como parámetro para la población.

$$s^2=\frac{\sum x_i^2*n_i}{n-1}-\bar{x}^2$$ Como estimador para la muestra.

Donde:

$\bar{x}$: media aritmética

n : total de datos de la muestra

N: total de datos de la población

x_i : cada dato de la variable o marca de clase si es intervalo

n_i : frecuencia absoluta

Para los datos del ejemplo de la estatura en centímetros de un grupo de mujeres que asisten al gimnasio, presentados en la tabla 22, la varianza seria (ver tabla 23):

Tabla 21.

Cuartiles para la estatura en centímetros de un grupo de mujeres que asisten al gimnasio

Nº. de Intervalo	Intervalo (Estatura en cm)	n_i	f_i	N_i	F_i	$\dot{x}$
1	[150 – 155]	1	3%	1	3%	152.5
2	(155 – 160]	11	31%	12	34%	157.5
3	(160 – 165]	13	37%	25	71%	162.5
4	(165 – 170]	6	17%	31	89%	167.5
5	(170 – 175]	4	11%	35	100%	172.5

Fuente: (Posada Hernandez, 2016)

Tabla 22.

Varianza para la estura en centímetros de un grupo de mujeres que asisten al gimnasio

Nº. de Intervalo	Intervalo (Estatura en cm)	x	x^2	n_i	$x_i^2 * n_i$
1	[150 – 155]	152,5	23.256,25	1	23.256,25
2	(155 – 160]	157,5	24.806,25	11	272.868,75
3	(160 – 165]	162,5	26.406,25	13	343.281,25
4	(165 – 170]	167,5	28.056,25	6	168.337,50
5	(170 – 175]	172,5	29.756,25	4	119.025,00
					$\sum x_i^2 * n_i = 926.768,75$

Fuente: (Posada Hernandez, 2016)

Al calcular el promedio de la estatura para las 34 mujeres se obtiene:

$$x = 162{,}6\ cm$$

Luego, la varianza será:

$$s^2 = \frac{\sum x_i^2 * n_i}{n-1} - \bar{x}^2 = \frac{926.768,75}{35-1} - (162,6)^2 = 27.257,9 - 26.438,8 = 819,1$$

- **Desviación estándar**

La desviación estándar es considerada la medida de dispersión con mayor representatividad para un conjunto de datos. Matemáticamente se calcula como la raíz cuadrada positiva de la varianza, y se denota por (s) cuando se estima para la muestra y por (σ) si se calcula para la población:

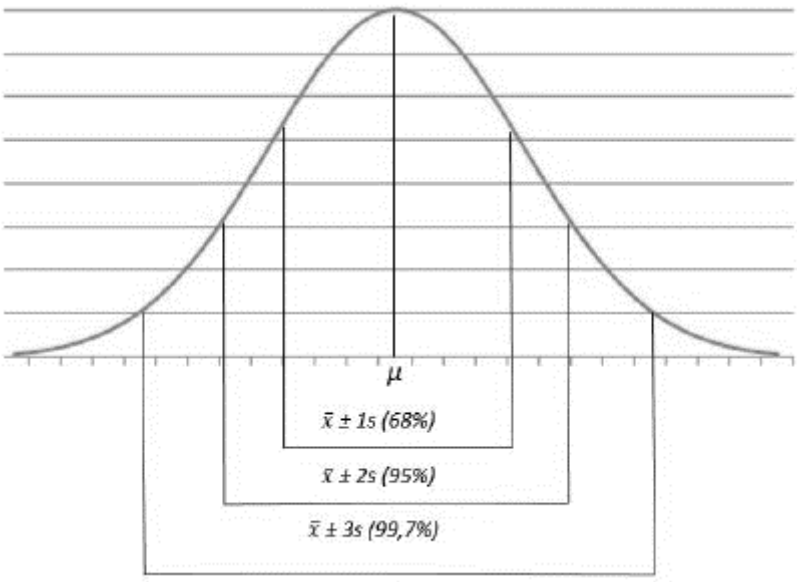
$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\sigma^2}$$

La desviación estándar indica la distribución de los datos alrededor de la media aritmética o promedio. Cuando la distribución de los datos se aproxima a una forma de campana o es simétrica, como se ilustra en la ilustración, la desviación estándar puede interpretarse mediante la regla empírica, esta es: el 68% de los datos se agrupan alrededor de media, entre el intervalo (x-1s) y (x+1s), el 95% entre (x-2s) y (x+2s), el 99,7% entre (x-3s) y (x+3s). Para los análisis estadísticos solo se analiza la dispersión de los datos a partir de una variación de la desviación alrededor de la media aritmética, es decir, el intervalo que cubre aproximadamente el 68% de los datos, teniendo en cuenta que la distribución de estos debe ser simétrica. (Posada Hernandez, 2016)

Ilustración 7

Variación de la desviación estándar alrededor de la media aritmética



Fuente: (Posada Hernandez, 2016)

Retomando la información de la tabla 23 sobre la estatura en centímetros de un grupo de mujeres que asisten al gimnasio, la desviación estándar seria:

$$s = \sqrt{s^2} = \sqrt{819{,}1} = 28{,}6 \text{ cm.}$$

Al interpretar la desviación estándar, significa que la estatura varia 28,6 cm alrededor de la media (162,6 cm). Por la regla empírica, podría decirse que el 68% de las estaturas está dentro de una desviación estándar de la media, se estima que el 95% de las estaturas estará entre $(x \pm 2s)$ y el 99,7% estará entre $(x \pm 3s)$.

Es importante resaltar que las medidas del rango, rango intercuartil, varianza y desviación estándar nunca asumen valores negativos. La relación de estas medidas con la dispersión es directa, es decir, si los valores de las medidas son altos, la dispersión también será alta y viceversa. (Posada Hernandez, 2016)

- **Rango**

El rango es considerado como la medida de dispersión más simple para el análisis de los datos. No ofrece mucha información sobre la variabilidad de los datos por estar basada solo en los valores extremos, razón por la cual debe ser usada como complemento de otras medidas de dispersión. Para el cálculo del rango se utiliza la siguiente ecuación:

Rango = valor máximo – valor mínimo

Por ejemplo, para los datos de la tabla 22, sobre la talla (en centímetros) de un grupo de mujeres que asisten al gimnasio, el rango serio:

Rango = 175 -150 = 25cm

Al interpretar el rango se deben relacionar los valores mínimo y máximo; es decir, resaltar las cantidades entre las cuales se encuentra el rango. Para el ejemplo mencionado anteriormente, la variación de la talla de las mujeres que asisten al gimnasio es de 25cm, la cual oscila entre 150 y 175cm. Si no se hace claridad que el rango esta entre los valores 150 y 175 cm. Si no se hace claridad que el rango esta entre los valores 150 y 175cm, puede generar confusión debido a que pueden existir muchos valores extremos con rango de 25cm. (Posada Hernandez, 2016)

- **Coeficiente de variación**

El coeficiente de variación (CV) es una medida que relaciona la desviación estándar con la media aritmética para determinar qué tan homogénea o dispersa es la información. Expresa el porcentaje que representa la desviación con relación a la media aritmética y se calcula por medio de la siguiente ecuación:

$$CV = \frac{S}{\overline{X}} * 100$$

Cuando se tiene una muestra, el coeficiente de variación puede ser utilizado para calificar estadísticamente la calidad de las estimaciones. Para ello se consideran los siguientes criterios:

CV menor o igual al 7%, las estimaciones se consideran precisas.

CV entre el 8% y el 14%, las estimaciones tienen precisión aceptable.

CV entre el 15% y 20%, la precisión es regular.

CV mayor del 20% indica que la estimación es poco precisa.

Para el ejemplo de la tabla 23 sobre la talla de un grupo de mujeres que asisten al gimnasio, la media aritmética fue 162,6cm y la desviación estándar 28,6 cm. Al calcular el coeficiente de variación, se obtiene:

$$CV = \frac{S}{\overline{X}} * 100 = \frac{28,6}{162,6} * 100 = 17,6\%$$

Al interpretar los datos, es posible establecer que la desviación representa el 17,6% de la media. En términos del ejercicio, podría interpretarse que los datos varían 17,6% alrededor de la media, lo cual intuye que la precisión de estimación de los parámetros para esta población es regular.

El coeficiente de variación, por ser una medida de dispersión relativa, se utiliza para comparar la variabilidad de distintas muestras o poblaciones, aunque tengan unidades de medida diferentes. En el siguiente ejemplo se muestra esta situación:

Una persona desea realizar una inversión en un negocio que tenga buena rentabilidad, para ello se le presentan dos proyectos con posibilidades diferentes. El primer proyecto ha presentado utilidades promedio en el último año de $150 millones y desviación de $50 millones. ¿Cuál proyecto presenta más estabilidad para generar confianza al inversionista?

Al analizar la desviación estándar, el primer proyecto es más variable que el segundo proyecto. Si embargo, como el promedio de las utilidades de los proyectos es diferente, se recomienda considerar la variación de la utilidad con respecto al promedio, para observar la estabilidad de ambos proyectos. (Posada Hernandez, 2016)

Los coeficientes de variación para los proyectos serían:

Primer proyecto: $$CV_1 = \frac{S}{\overline{X}} * 100 = \frac{\$50}{\$150} * 100 = 33,3\%$$

Segundo proyecto: $$CV_2 = \frac{S}{\overline{X}} * 100 = \frac{\$12}{\$120} * 100 = 10\%$$

En consecuencia, en relación con la media, la utilidad del primer proyecto es más variable que la del segundo. Por tanto, a pesar de presentar el segundo proyecto menor utilidad promedio, es mas estable que el primero, lo cual puede generar mayor confianza para el inversionista. (Posada Hernandez, 2016)

1.4.3. Cuartiles, quintiles y deciles

- **Cuartiles**

Los cuartiles (Q1) son valores que fraccionan la distribución de los datos en cuatro partes iguales. Existen 3 cuartiles y cada una de las partes representa un 25% de los datos.

El primer cuartil Q1 deja por debajo el 25% de la distribución de los datos o el 75% por encima de él. El segundo cuartil (Q2) acumula el 50% de los datos por debajo y el otro

50% por encima del (por tal razón es igual a la mediana); y el tercer cuartil (Q3) deja por debajo el 75% de los datos y por encima el 25%.

El cálculo de los cuartiles se realiza mediante el siguiente procedimiento:

a) Ordenar los datos de forma ascendente.
b) Calcular la posición i con la ecuación: $i = (\frac{k}{4})^n$. Donde K es el numero del cuartil (K = 1,2,3) y n el número total de datos.
c) Si i no es un numero entero, se debe redondear al entero siguiente y el valor que ocupa esta posición será el cuartil requerido. Si i es un numero entero, el cuartil es el promedio de los valores e . ii + 1.

Ejemplo: se le consulto a un grupo de siete estudiantes el número de horas semanal que dedican para el repaso de los temas vistos en clase, obteniendo los siguientes resultados: 3,5,2,7,6,4,9 horas. Para el cálculo de los cuartiles, se empleará el procedimiento descrito anteriormente.

a) Ordenar los datos en forma ascendente: 2,3,4,5,6,7,9.
b) Para el cuartil Q1 la posición i seria: $i = (\frac{1}{4})\ 7 = 1{,}75$
c) Dado que i no es un entero, se redondea al entero siguiente, es decir a 2. En este caso, el cuartil Q1 corresponde al valor ubicado en la posición 2, el cual es 3 horas. Su interpretación significa que el 25% de los estudiantes dedican máximo 3 horas semanales para el repaso a los temas vistos en clase.

De forma similar, para el cuartil Q2 la posición i seria: $i = (\frac{2}{4})7 = 3{,}5$

Como i no es un entero, se redondea al entero siguiente, es decir a 4. Por tanto, el cuartil Q2 será el valor correspondiente a la posición 4, el cual es 5 horas. Esto es, el 50% de los estudiantes dedican máximo 5 horas semanales para el repaso a los temas vistos en clase. Nótese que este valor corresponde a la mediana.

En este caso, para el cuartil Q3 la posición i seria:

$$i = (\frac{3}{4})7 = 5{,}25$$

Al redondearla quedaría en 6, y el valor del cuartil Q3 es 7 horas. Indica que el 75% de los estudiantes dedican máximo 7 horas semanales para el repaso a los temas vistos en clase.

Ejemplo: la talla de los neonatos prematuros nacidos en los partos durante una noche en un hospital fueron: 40,37,29,31,32,38,38,38cm; para el cálculo de los cuartiles se empleará el procedimiento del ejemplo anterior, teniendo en cuenta el resultado obtenido al calcular la posición i.

Primer paso: ordenar los datos en forma ascendente: 29,31,32,37,38,38,38,40.

Segundo paso: para el cuartil Q1 la posición i seria: $i = (\frac{1}{4})8 = 2$

Tercer paso: dado que i es un entero, el cuartil Q1 corresponde al promedio entre los valores ubicados en las posiciones 2 y 3. Esto es, $Q1 = \frac{31+32}{2} = 31{,}5\,cm$. Su interpretación significa que el 25% de los neonatos prematuros presentaron una talla máxima de 31,5 cm.

El cuartil Q2 tendría la posición $i = (\frac{2}{4})8 = 4$ y sería el promedio entre los valores ubicados en las posiciones 4 y 5. Esto es,

$$Q_2 = \frac{37+38}{2} = 37{,}5 \text{ cm.}$$

Su interpretación significa que el 50% de los neonatos prematuros presentaron una talla máxima de 37,5 cm, igual a la mediana.

Para el Q3, la posición será: $i = (\frac{3}{4})8 = 6$ y sería el promedio entre los valores ubicados en las posiciones 6 y 7. Esto es, $Q2 = \frac{37+38}{2} = 37cm$. Su interpretación significa que el 75% de los neonatos prematuros presentaron una talla máxima de 38 cm.

Si los datos se han agrupado en clases o intervalos, los cuartiles se calculan mediante la siguiente ecuación. (Posada Hernandez, 2016)

$$Q_k = l_i + \left[\frac{k(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C$$

Donde:

k : número del cuartil, k= 1, 2, 3.

n : número total de datos.

l_i : límite inferior del intervalo que contiene a k(n/4).

N_{i-1} : frecuencia absoluta acumulada anterior al intervalo que con tiene a k(n /4).

n_i : frecuencia absoluta del intervalo que contiene a k(n /4).

C : amplitud del intervalo.

Ejemplo: en la tabla 22 se presentan los datos ordenados de la estatura, en centímetros, de un grupo de mujeres que asisten al gimnasio.

El cuartil uno se calcula mediante el siguiente procedimiento:

a) Se halla k(n/4). (1*35/4 = 8,75)
b) Se ubica el intervalo que contiene a k(n/4) en la frecuencia absoluta acumulada N1. (El segundo intervalo contiene a 8,75 en la frecuencia absoluta acumulada).
c) El primer cuartil se obtiene mediante la fórmula:

$$Q_1 = l_i + \left[\frac{1(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C$$

Nota: la descripción de los componentes de la fórmula es la misma que se realizó en la medicina.

$$Q_1 = 155 + \left[\frac{1(\frac{35}{4}) - 1}{11} \right] * 5 = 159,4 \text{ centímetros}$$

Se estima que el 25% de las mujeres que asisten al gimnasio presentan una estatura máxima de 159,4 cm.

De forma similar se obtienen los cuartiles dos y tres.

$$Q_2 = l_i + \left[\frac{2(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C \qquad Q_2 = 160 + \left[\frac{2(\frac{35}{4}) - 12}{13} \right] * 5 = 162,1 \text{ cm}$$

$$Q_3 = l_i + \left[\frac{3(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C \qquad Q_3 = 165 + \left[\frac{3(\frac{35}{4}) - 25}{6} \right] * 5 = 166 \text{ cm}$$

El 50% de las mujeres presentan una estatura máxima de 162,1 cm (cuartil dos) y el 75% tienen una estatura máxima de 166 cm (cuartil tres). (Posada Hernandez, 2016)

- **Quintiles**

Es un fractil se obtienen dividiendo al conjunto de datos en cinco partes iguales cada parte representa el 20% del total. Se pueden calcular 4 quintiles. (Berrocal de Montestruque, Asurza Olaechea, & Billon, 2016)

- **Deciles**

Los deciles (D1) son valores que fraccionan la distribución de los datos en diez partes iguales. En la distribución se presentan nueve deciles: el D1 acumula el 10% del conjunto de datos, el D2 deja el 20%, y así sucesivamente hasta el D9, que acumula el 90% de los datos. Para el cálculo de los deciles se usa un procedimiento similar al de los cuartiles:

a) Ordenar los datos de forma ascendente.
b) Calcular la posición i con la ecuación: $i = (\frac{k}{10})^n$. Donde K es el numero de decil (k = 1,2,3,4,5,6,7,8,9) y n el número total de datos.
c) Si la posición i no es un numero entero, se debe redondear al entero siguiente y el valor que ocupa esta posición sera el cuartil requerido. Si la posición es un numero entero, el decil es el promedio de los valores i e i + 1. (Berrocal de Montestruque, Asurza Olaechea, & Billon, 2016)

Para datos agrupados en intervalos:

$$D_k = l_i + \left[\frac{k(\frac{n}{10}) - N_{i-1}}{n_i} \right] * C$$

El cálculo de los deciles uno y nueve para el ejemplo de la estatura de las mujeres, presentado en la tabla 22, se detalla a continuación:

$$D_1 = l_i + \left[\frac{1(\frac{n}{10}) - N_{i-1}}{n_i} \right] * C = 155 + \left[\frac{3,5-1}{11} \right] * 5 = 155 + 1,1 = 156,1 \text{ centímetros}$$

$$D_9 = l_i + \left[\frac{9(\frac{n}{10}) - N_{i-1}}{n_i} \right] * C = 170 + \left[\frac{31,5-31}{4} \right] * 5 = 170 + 0,6 = 170,6 \text{ centímetros}$$

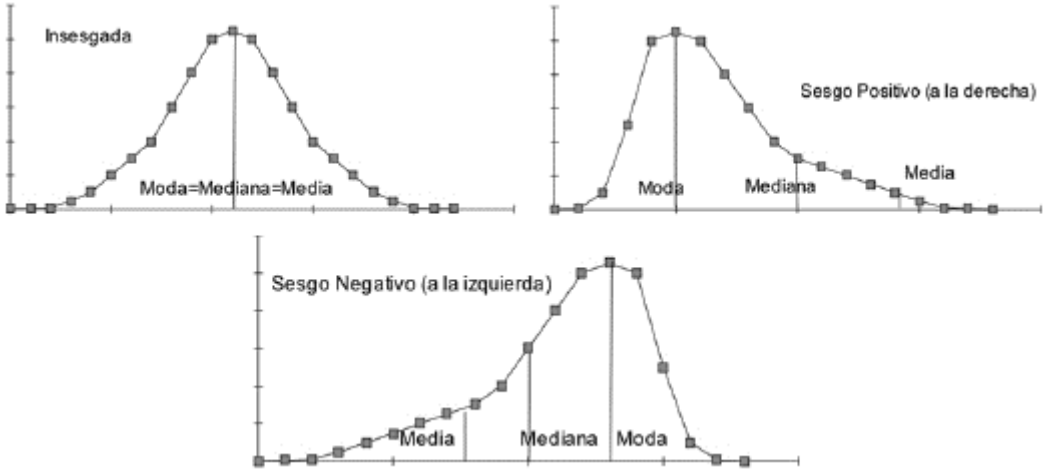
1.4.4. Coeficientes de Asimetría

Al analizar la distribución de los datos es posible que exista una tendencia de estos hacia uno de los extremos (derecho o izquierdo). Esta tendencia se denomina sesgo y permite mostrar la inclinación de los datos hacia los extremos.

Para describir el sesgo o la forma de la distribución de los datos, se comparan la media aritmética, la mediana y la moda. Si estas medidas son exactamente iguales, se considera que la distribución de los datos es insesgada o simétrica (con sesgo cero). En otro caso, cuando la media aritmética es superior a la mediana, la distribución de los datos estará sesgada a la derecha (o con sesgo positivo), tal como se muestra en la ilustración 8. (Posada Hernandez, 2016)

Ilustración 8.

Forma o sesgo de la distribución de los datos



Fuente: (Posada Hernandez, 2016)

El sesgo mantiene relación directa con la media aritmética, es decir, si la media se afecta por valores extremos, esto se verá reflejado en el sesgo. Si no hay valores extremos (muy pequeños o muy grandes) la distribución se comporta de forma simétrica, en tal forma existe una compensación entre los valores grandes y los pequeños.

La asimetría de un conjunto de datos se puede calcular mediante varios coeficientes, entre ellos están:

a) Coeficiente de asimetría de Pearson: Relaciona la diferencia entre media aritmética y la moda con la desviación. Pese a que este eficiente es fácil de calcular, no se utiliza con frecuencia en la práctica, ya que la distribución de los datos debe ser unimodal y moderada o ligeramente asimétrica, condiciones que no se observan de forma directa en la distribución, por lo que resultan muy exigentes. El coeficiente de Pearson varia entre -3 y 3 y la fórmula es:

$$Ap = \frac{\bar{x} - M_o}{s}$$

b) Coeficiente de asimetría de Bowley: Este coeficiente es el menos usado por sus altas exigencias. Para emplearlo se requiere que tanto al extremo izquierdo como al derecho de la distribución de los datos, se presente un comportamiento similar, de lo contrario no es imposible estimar la asimetría. El calculo se basa en la posición que presentan los cuartiles y la mediana. La medida de Bowley varia entre -1 y 1 y se calcula de acuerdo con la siguiente expresión:

$$Ab = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

c) Coeficiente de asimetría de Fisher: está basado en las desviaciones que presentan los datos con respecto a la media. Es el coeficiente más usado para determinar la asimetría de un conjunto de datos, debido a que no es afectado por valores extremos y solo vincula la media aritmética y la desviación. La fórmula para su cálculo es la siguiente:

$$g_1 = \frac{\sum (x_i - \bar{x})^3}{n * s^3} \text{ ; para datos sin agrupar.}$$

$$g_1 = \frac{\sum n_i (x_i - \bar{x})^3}{n * s^3} \text{ para datos agrupados en frecuencias.}$$

Es importante resaltar que para determinar el sesgo se sugiere utilizar el coeficiente de asimetría de Fisher, el cual es más confiable para analizar la similitud de la distribución de los datos con la Distribución Normal; además el valor obtenido es muy similar con el estimado por el Excel.

Debe tenerse en cuenta que el análisis del sesgo se realiza a partir del signo que arroja cualquiera de los coeficientes mencionados y, particularmente para el coeficiente de Fisher, mientras más se aleje de cero, mayor es el sesgo de la distribución de los datos, tanto a la derecha como a la izquierda. En síntesis:

Sesgo > 0: sesgo positivo o a la derecha.

Sesgo = 0: simetría en la distribución de los datos.

Sesgo < 0: sesgo negativo o a la izquierda. (Posada Hernandez, 2016)

Por ejemplo, el sesgo para la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa (tabla 21), con media aritmética de 3,6 y desviación de 0,66, calculado mediante el coeficiente de asimetría de Fisher, será:

Tabla 23

Sesgo para la puntuación de la evaluación de desempeño de siete empleados del área de mercadero de una empresa

Empleado	Calificación (X_i)	Media de la muestra ($\bar{x}$)	Diferencia ($x_i - \bar{x}$)	$(x_i - \bar{x})^3$
1	3,5	3,6	-0,1	-0,001
2	4,5	3,6	0,9	0,729
3	4,2	3,6	0,6	0,216
4	3,0	3,6	-0,6	-0,216
5	2,7	3,6	-0,9	-0,729
6	3,3	3,6	-0,3	-0,027
7	4,0	3,6	0,4	0,064
			TOTAL	0,036

Fuente: (Posada Hernandez, 2016)

Aplicando la formula del coeficiente de asimetría de Fisher se tiene:

$$g_1 = \frac{\sum (x_i - \bar{x})^3}{N * s^3} = \frac{0,036}{7 * 0,66^3} = \frac{0,036}{2,01} = 0,0179$$

Dado que le valor es positivo, el sesgo es la a la derecha o positivo, es decir, la calificación de la mayoría de los empleados tiende a estar por debajo del promedio de 3,6. (Posada Hernandez, 2016)

1.4.5. Coeficiente de Curtosis

La curtosis es una medida que permite analizar la concentración de los datos alrededor de los valores medios de la muestra. Se calcula con el coeficiente de Fisher para la curtosis; la ecuación es la siguiente:

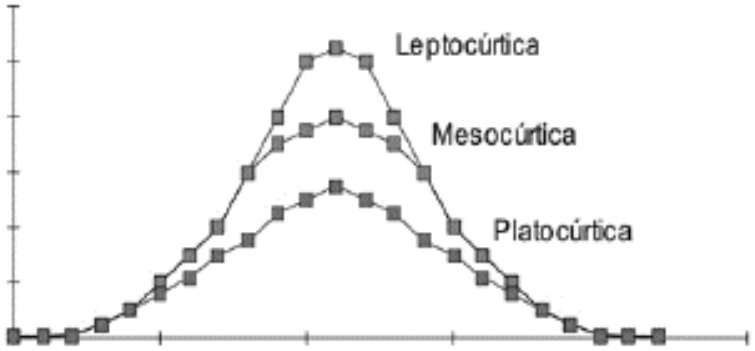
$$g_2 = \frac{\sum (x_i - \bar{X})^4}{n * s^4} - 3 \quad \text{para datos sin agrupar.}$$

$$g_2 = \frac{\sum n_{ii} (x - \bar{X})^4}{n * s^4} - 3 \quad \text{para datos agrupados en frecuencias.}$$

El coeficiente de curtosis diferencia tres clases de distribuciones, que se ilustran en la ilustración:

Ilustración 9

Tipos de distribución según el coeficiente de curtosis



Fuente: (Posada Hernandez, 2016)

El grado de concentración alrededor de los valores centrales de la variable es moderado. Este mismo comportamiento se presenta en una distribución simétrica o normal. El coeficiente presenta valor igual a cero.

a) Distribución mesocúrtica: El grado de concentración alrededor de los valores centrales de la variable es moderado. Este mismo comportamiento se presenta en una distribución simétrica o normal. El coeficiente presenta valor igual a cero.
b) Distribución leptocúrtica: El grado de concentración alrededor de los valores centrales de la variable es elevado, lo que la hace ver de forma puntiaguda, dado que las frecuencias altas están alrededor de la media. El valor del coeficiente es mayor a cero.
c) Distribución platicurtica: El grado de concentración alrededor de los valores centrales de la variable es reducido, mostrándose de forma aplanada dado que las frecuencias bajas están alrededor de la media. El valor del coeficiente es menor a cero.

Para el ejemplo sobre la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa con media aritmética de 3,6 y desviación estándar de 0,66, la curtosis será (ver tabla 25):

Tabla 24.

Curtosis para la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa

Empleado	Calificación (x_i)	Media de la muestra $(\bar{x})$	Diferencia $(x_i-\bar{x})$	$(x_i-\bar{x})^4$
1	3,5	3,6	-0,1	-0,001
2	4,5	3,6	0,9	0,6561
3	4,2	3,6	0,6	0,1296
4	3,0	3,6	-0,6	0,1296
5	2,7	3,6	-0,9	0,6561
6	3,3	3,6	-0,3	0,0081
7	4,0	3,6	0,4	0,0256
			Total	1,6052

Fuente: (Posada Hernandez, 2016)

Aplicando la formula del coeficiente de asimetría de Fisher se tiene:

$$g_2 = \frac{\sum (x_i - \overline{X})^4}{N * s^4} - 3 = \frac{1,6052}{7 * 0,66^4} - 3 = \frac{1,6052}{1,3282} - 3 = 1,208 - 3 = -1,792$$

Dado que el coeficiente de Fisher para la curtosis es -1,792, se considera una distribución platicurtica, lo cual significa que existe reducida concentración de los datos alrededor de los valores centrales de la distribución. (Posada Hernandez, 2016)

2. UNIDAD II: PROBABILIDAD Y DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

2.1. Aspectos básicos de la probabilidad

En general, la probabilidad es la posibilidad de que algo suceda. Las probabilidades se expresan como fracciones (1/6) o como decimales (0,167) que están entre cero y uno. Tener la probabilidad de cero (0), significa que nunca va a suceder. Una probabilidad de uno (1), sucederá siempre.

a) EVENTO: Es uno o más posibles resultados de hacer algo. Es un subconjunto del Espacio Muestral. Un evento se indica con letras mayúsculas del alfabeto. Ejemplo: Al lanzar una moneda, si cae cruz es un evento y si cae cara es otro.
b) EXPERIMENTO: Es la actividad que origina uno de dichos eventos. Ejemplo:
    a. Lanzamiento de un dado
    b. En un hotel se desea detectar personas que prefieren un lugar turístico de un grupo de 100.

c. Un día se decide tomar una muestra 10 empleados del hotel, eligiendo al azar.

c) ESPACIO MUESTRAL (Ω): Conjunto de todos los resultados posibles de un experimento. Ejemplo:
   a. Para el experimento 1: **Ω** = {1,2,3,4,5,6}
   b. Para el experimento 3: **Ω** = {0,1,2,…,100}
   c. Para el experimento 4: **Ω** = {empleados del hotel). (Alvarez Roman, 2004)

2.2.1. Experimentos aleatorios y deterministas

Un fenómeno o experimento aleatorio, es el que satisface las siguientes características:

a) Se conocen todos sus resultados
b) Cuando se lleva a cabo no se sabe con certeza el resultado que se va a obtener y
c) Puede ser repetido bajo idénticas condiciones

Un experimento no aleatorio se denomina fenómeno o experimento determinista y por ello, cuando se realiza bajo las mismas condiciones se sabe con certeza su resultado.

Un ensayo es una realización de un experimento aleatorio.

De acuerdo a lo anterior, el experimento es aleatorio si para un ensayo particular su resultado no es previsible con certidumbre, mientras que en un fenómeno determinista en cualquier ensayo haya certeza del resultado que se obtendrá y si se realiza bajo las mismas condiciones se obtiene siempre el mismo resultado.

Ejemplo 1: El experimento E que consiste en lanzar una moneda honesta, no cargada, es aleatorio, ya que

a) Se saben los resultados que se pueden obtener: águila o sol,
b) En cada lanzamiento o ensayo, no se conoce con seguridad cual de las dos caras va a salir y
c) Puede considerarse que los lanzamientos se realizan bajo idénticas condiciones

Ejemplo 2: El experimento de medir la velocidad a la que debe viajar Fulanito para recorrer 270 kms en 1 hora 30 minutos, es un experimento determinista, puesto que la velocidad (v) es igual a la distancia (d) entre el tiempo (t), $v = \frac{d}{t}$, de forma que se sabe con certeza que la velocidad a la que debe manejar Fulanito es de 3 kms por minuto o bien, 180 kms por hora, si quiere llegar a su destino en 90 minutos.

Si Zutanito también desea hacer el recorrido en el mismo tiempo, la velocidad a la que viajará será la misma que la de Fulanito, puesto que las condiciones bajo las cuales se realiza ambos ensayos son las mismas, recorrer 270 kms en 1 hora 30 minutos. Cualquier individuo que quiera hacer ese recorrido en el tiempo propuesto, debe viajar a la velocidad obtenida.

Ejemplo 3: ¿Cuáles de los siguientes experimentos son aleatorios?

E1: Lanzar un dado no cargado y anotar el número que se muestra

E2: Contar las páginas de un libro de 87 hojas

E3: Medir el tiempo de vida de un foco

E4: Seleccionar un hombre de un equipo de basquetbol varonil

Los experimentos E1 y E3 son aleatorios. En el de lanzar el dado se sabe que resultados se deben obtener: 1,2,3,4,5 o 6, pero al lanzarlo no se conoce el número que se va a mostrar, aun cuando sea lanzado en las mismas condiciones. Por su parte, en el experimento de medir el tiempo de vida de un foco, puede ser que cuando se encienda por primera vez ya no funcione o que dure unos segundos, minutos, horas, días, etc., es decir, se saben los posibles resultados, pero para un foco particular no se conoce con seguridad su tiempo de vida aun cuando se coloque en el mismo lugar.

En el experimento E2 que consiste en contar el número de páginas de un libro de 87 hojas, se sabe con certeza que el resultado es 174 páginas, hecho que lo hace ser no aleatorio, es un experimento determinista. De igual forma, en el E4, al seleccionar una persona del equipo de básquetbol varonil con toda certeza se va a elegir a un varón, por lo que el experimento tampoco es aleatorio. (Garcia Salazar & Ruiz Galindo, 2013)

2.1.2. Variables aleatorias

a) Las variables aleatorias unidimensionales:

Una variable aleatoria se define como una función real de una partición € del espacio muestral (Ω) asociado a un fenómeno de comportamiento no determinístico, formada por los eventos incompatibles E = {E1, E2, E3, …} que representan el conjunto exhaustivo de resultados posibles de dicho fenómeno: X(w) = x1|E1| + x2|E2| + x3|E3| + … (weΩ), donde los |Ei| denotan indicadores que pueden asumir los valores 1 o 0 según que el evento Ei ocurra o no y los Xi son números reales. De esta forma, la definición de variable aleatoria traslada el ámbito del análisis del espacio Ω de los eventos al "dominio" (o "soporte") Rn al cual pertenecen los valores que puede asumir (este traslado implica una importante simplificación conceptual, en la medida que Ω es un espacio abstracto en tanto que Rn es un espacio Euclidiano sobre el cual las operaciones son más simples).

La teoría de las variables aleatorias está dirigida fundamentalmente a la asignación de sus probabilidades asociadas. Es decir, dada una variable aleatoria X, su análisis esta relacionado, en principio, con la asignación de las probabilidades de realización de cada uno de los valores de su dominio (xEΩ(X)) o, en términos mas generales, de las probabilidades de asumir valores pertenecientes a un conjunto BEΩ(X):

$$p(x\epsilon B) = p[X(w)\epsilon B] = p\{w/[X(w)\epsilon B]\} = p[X^{-1}(B)]$$

(donde B está definido por un intervalo o por un numero finito o una infinidad numerable de operaciones y $x^{-1}(B)$ denota la imagen inversa de B).

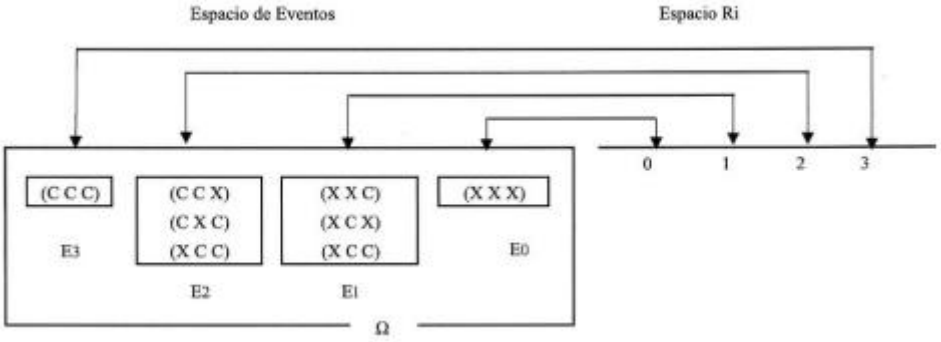
La partición E que define a una variable aleatoria puede ser numerable o poseer la cardinalidad del continuo (es decir, ser biunívocamente correspondiente con el conjunto de los números reales, card (E) = card (R)). En el primer caso se dice que la variable es discreta, en el segundo que es continua.

Como una extensión de la definición anterior, se dice que Z(w) = X(w) + iY(w) es una variable aleatoria compleja si X(w) e Y(w) son variables aleatorias reales.

Ejemplo 1: Sea una prueba consistente en arrojar tres monedas "clásicas". La variable aleatoria que representa el número de veces que se puede obtener el resultado "cara" esta definida de la siguiente forma:

Ilustración 10

Arrojar tres monedas y el número de veces que se puede obtener resultado "cara"



Fuente: (Landro & Gonzalez, 2018)

De acuerdo con el supuesto de "perfección" de las monedas, la probabilidad a asignar al evento "que la variable asuma el valor =", es decir la probabilidad de que ocurra el evento:

Eo: no obtener ninguna vez el resultado "cara"

Sera $p(Eo) = \frac{1}{8}$. De la misma forma, las probabilidades a asignar a la ocurrencia de los eventos:

Ei: obtener i (= 1,2,3) veces "cara"

Serán, respectivamente: $p(E1) = \frac{3}{8}$, $p(E2) = \frac{3}{8}$ y $p(E3) = \frac{1}{8}$.

b) Las variables aleatorias unidimensionales discretas:

Sea X(w) una variable aleatoria discreta y sea **Ω**(X) = {Xi, i = 1,2,…} su dominio y sea la sucesión:

$$p_i = p[X(w) = x_i] = p\{w/[X(w) = x_i]\} = p[X^{-1}(x_i)](i = 1,2,3, \dots)$$

La asignación de la probabilidad de que la variable asuma cada uno de estos valores, la cual define la "función de probabilidades" de la variable.

Teniendo en cuenta que los eventos X(Xi) son incompatibles, de acuerdo con los axiomas de no-negatividad de la probabilidad y de aditividad numerable, se obtiene que la función de probabilidades posee las siguientes propiedades:

i) $p_i \geq 0 \qquad (i = 1,2,3, \dots)$

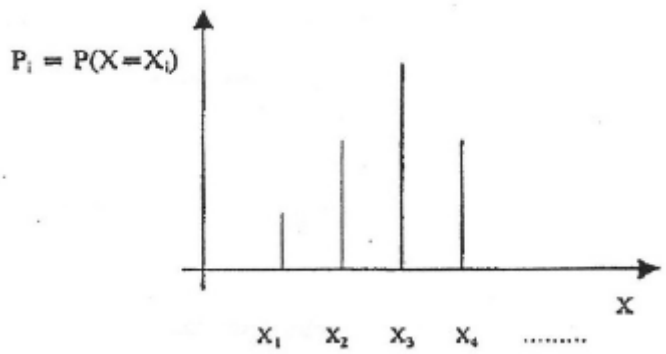
ii) $p[X(w)\epsilon\Omega(X)] = p\{X^{-1}[\Omega(X)]\} = p[(X_1 = x_1) \cup (X_2 = x_2) \cup \dots] =$

$= \sum_{x_i \epsilon \Omega(X)} p_i = 1$

Esta propiedad de " aditividad numerable" o "completa" o "monótona" o " $\sigma - aditividad$" fue introducida por Borel (1896)(1897)(1898), quien la convirtió en el tema central de la teoría de la probabilidad al demostrar su ley fuerte de los grandes números. Deben reconocerse, además, los aportes de Lesbesque (1901)(1904), Radon (1913)(1915), Frechet (1915 a)(1915 b) (1930 a) (1930 b), Daniell (1918)(1919 a)(1919 b)(1920)(1921), Wiener (1920)(1921 a)(1921 b)(1923)(1924) y Steinhaus (1923)(1930 a)(1930 b). Si bien Kolmogorov (1933), a partir del teorema de continuidad, incorporo a su axiomática la condición de aditividad numerable, manifiesto ciertas reservas sobre su validez y la justifico exclusivamente en virtud de su utilidad en ciertos ámbitos de la investigación con respecto a la interpretación frecuencista de la probabilidad (la aditividad numerable constituye un ejemplo del riesgo que traen aparejadas las conclusiones que se obtienen de un sistema axiomático basado exclusivamente en conveniencias matemáticas. (Landro & Gonzalez, 2018)

Ilustración 11.

Plano cartesiano en el que los valores de la variable figuran en el eje de abscisas y las probabilidades respectivas están representas por segmentos de longitud proporcional a cada probabilidad



Fuente: (Landro & Gonzalez, 2018)

El dominio Ω(X) = {Xi, i = 1,2, …} de la variable X y la función de probabilidades {Xi, Pi,i = 1,2, …} componen su " distribución de probabilidades".

En este caso particular de las variables unidimensionales, cualquier conjunto de valores de su dominio puede ser definido a partir de intervalos, utilizando operaciones de unión y negación. Se puede concluir, entonces, que los intervalos de la recta constituyen la base para la definición de la distribución de probabilidades de una variable unidimensional: el conjunto de posibles valores de la variable comprendidos en un intervalo [Xh, Xk] (para Xh < Xk) puede ser expresado por la diferencia [Xh, Xk] =]-∞, Xk] -] -∞, Xh] a partir de la cual se obtiene que:

$$p(x_h \leq X \leq x_k) = p(X \leq x_k) - p(X \leq x_h)$$

En resumen, para calcular la probabilidad de un evento (XEB) basta conocer, para todo xEΩ(X), la probabilidad del intervalo] -∞, x]; es decir, basta conocer la función:

$$F_X(x) = p(X \leq x) = p\left[X^{-1} \in]-\infty, x]\right] = \sum_{x_i \leq x} p(X = x_i) = \sum_{x_i \leq x} p_i$$

$$p(x_h \leq x \leq x_k) = F_X(x_k) - F_X(x_h) = \Delta_{x_h}^{x_k} F_X(x)$$

Denominada "función de distribución" o "de probabilidades acumuladas" de la variable X, la cual posee las siguientes propiedades:

a) Es una función no-negativa: Fx(Xi) ≥ 0 (i = 1,2,3,…) y continua por la izquierda: Fx (Xi -) = Fx(Xi) (los valores Xi se denominan "puntos de continuidad" de X).
b) Dados dos eventos: A = {X ≤ Xj}, donde Xi ≤ Xj (es decir, tales que A ⊂ B, de acuerdo con el axioma de aditividad, se verificara que p (A) ≤ p (B). Es decir que Fx (-) es una función no-decreciente, Fx(Xi) ≤ Fx(Xj)(∨Xi ≤ Xj) [5]
c) Se verifica que:

$$\lim_{x \to -\infty} F_X(x) = F_X(-\infty) = 0 \qquad \lim_{x \to +\infty} F_X(x) = F_X(+\infty) = 1$$

Estas condiciones son necesarias y suficientes para que una función pueda ser considerada una función de distribución. Es decir, si una función real Fx(X) satisface las condiciones enunciadas, se puede asegurar que existe una única función de distribución de probabilidades correspondiente a la variable X que es igual a Fx (x), para el intervalo] - ∞,x].

d) Sean dos variables aleatorias, X e Y, con funciones de distribución Fx (x) y Fy (y), respectivamente y tales que $p(|X - Y| \leq \eta) > 1 - \varepsilon$ (donde n y e denotan dos constantes positivas). De acuerdo con el teorema de la probabilidad total, se puede escribir:

$$p[(|X - Y| \leq \eta) \cup (Y \leq y)]$$
$$= p(|X - Y| \leq \eta) + p(Y \leq y) - p[(|X - Y| \leq \eta) \cap (Y \leq y)]$$

De las relaciones anteriores se obtiene que:

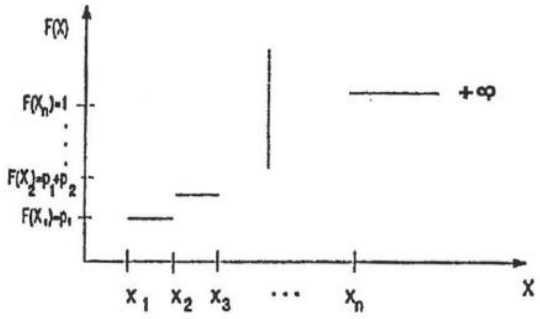
$$p[(|X - Y| \leq \eta) \cap (Y \leq y)] = p(X \leq Y + \eta) \geq p(Y \leq y) - \varepsilon$$

Y, por lo tanto, que $p(X > y - \eta) \geq p(Y > y) - \varepsilon$. Es decir, resulta que:

$$F_X(y - \eta) - \varepsilon \leq F_Y(y) \leq F_X(y + \eta) + \varepsilon$$
$$F_X(y - \eta) - F_X(y + \eta) - \varepsilon \leq F_X(y) - F_Y(y) \leq F_X(y + \eta) - F_Y(y - \eta) + \varepsilon$$
$$|F_X(y) - F_Y(y)| \leq [F_X(y + \eta) - F_X(y - \eta)] + \varepsilon$$

Ilustración 12.

La función de distribución



Fuente: (Landro & Gonzalez, 2018)

La función Fx (x) asume el valor pi para todos los valores de X comprendidos en el intervalo $[x_1, x_2[$. En x2 y para todos los puntos incluidos en el intervalo $[x_2, x_3[$, será Fx (X2) = p1 + p2, y así sucesivamente.

Ejemplo nº 2:

Sea X una variable aleatoria definida como la diferencia de los puntos a obtener entre la primera y la segunda tirada de un dado "clásico". Su dominio esta definido de la siguiente:

Tabla 25

Variable aleatoria de un dado clásico

(1 - 1)	(2 - 1)	(3 - 1)	(4 - 1)	(5 - 1)	(6 - 1)
(1 - 2)	(2 - 2)	(3 - 2)	(4 - 2)	(5 - 2)	(6 - 2)
(1 - 3)	(2 - 3)	(3 - 3)	(4 - 3)	(5 - 3)	(6 - 3)
(1 - 4)	(2 - 4)	(3 - 4)	(4 - 4)	(5 - 4)	(6 - 4)
(1 - 5)	(2 - 5)	(3 - 5)	(4 - 5)	(5 - 5)	(6 - 5)
(1 - 6)	(2 - 6)	(3 - 6)	(4 - 6)	(5 - 6)	(6 - 6)

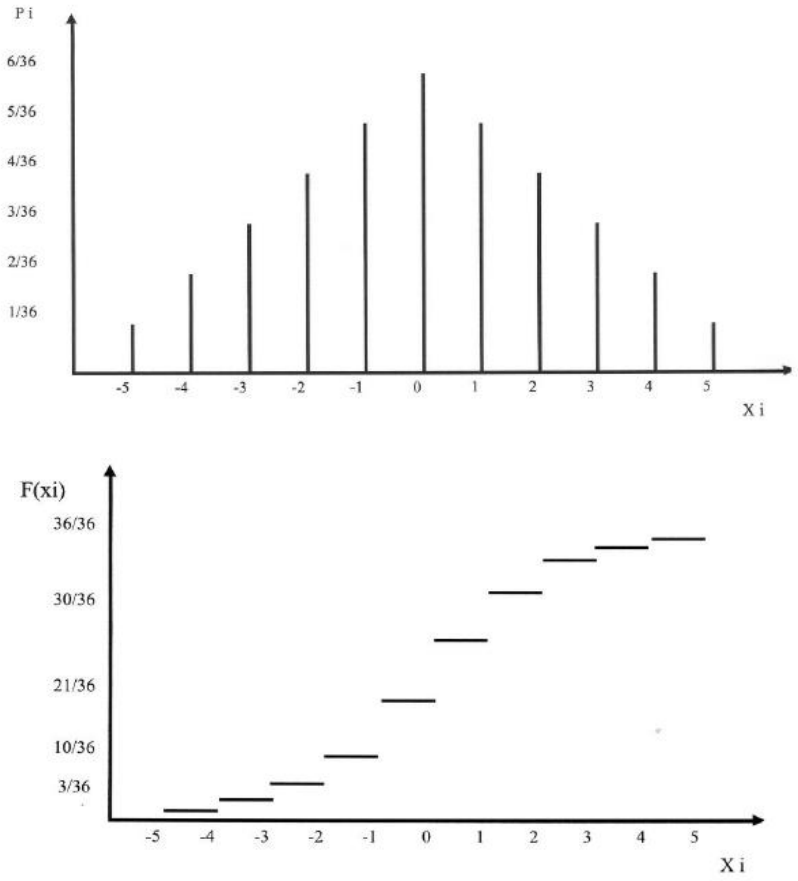
Ω

Fuente: (Landro & Gonzalez, 2018)

Tabla 26

Las funciones de probabilidades y de distribución

X	P (X=xi)	F(xi)	X	P (X=xi)	F(xi)
X1 = -5	1/36	1/36	X7 = 1	5/36	26/36
X2 = -4	2/26	3/26	X8 = 2	4/36	30/36
X3 = -3	3/36	6/36	X9 = 3	3/36	33/36
X4 = -2	4/36	10/36	X10 = 4	2/36	35/36
X5 = -1	5/36	15/36	X11 = 5	1/36	36/36
X6 = 0	6/36	21/36			



Fuente: (Landro & Gonzalez, 2018)

Ejemplo Nº 3: La variable tiempo de espera: Supónganse una serie de lanzamientos independientes de una moneda tal que la probabilidad de que en una prueba dad se produzca el resultado "cara" © es igual a p y la probabilidad de que se produzca el resultado "ceca" (X) es igual a q = 1 – p.

Sean los eventos:

$$C_n: que\ se\ produzca\ el\ resultado\ "cara"\ en\ el\ n-ésimo\ lanzamiento$$

$$X_n: que\ se\ produzca\ el\ resultado\ "ceca"\ en\ el\ n-ésimo\ lanzamiento$$

De modo que p(Cn) = p y p(Xn) = q. La probabilidad de que en n lanzamientos de la moneda se produzca una sucesión dada de k resultados "cara" y n – k resultados "ceca" (por ejemplo, la sucesión C1, C2, … , Ck, Xk + 1, Xk + 2, … , Xn) será, entonces:

$$p(C_1 \cap C_2 \cap \ldots \cap C_k \cap X_{k+1} \cap X_{k+2} \cap \ldots \cap X_n) = p^k q^{n-k}$$

Sea $Z^{(n)}$ la variable aleatoria que representa el numero de resultados "cara" a obtener en una serie de n lanzamientos. Su distribución de probabilidades será de la forma:

$$p\left(Z^{(n)} = k\right) = p[(C_1 \cap C_2 \cap \ldots \cap X_k \cap C_{k+1} \cap \ldots \cap X_n) \cup$$

$$\cup (C_1 \cap C_2 \cap \ldots \cap X_{k-1} \cap C_k \cap C_{k+1} \cap \ldots \cap X_n) \cup$$

$$\cup \ldots \cup (X_1 \cap X_2 \cap \ldots \cap X_{n-k} \cap C_{n-k+1} \cap \ldots \cap C_n)] =$$

$$= p[\cup(C_1 \cap C_2 \cap \ldots \cap C_k \cap X_{k+1} \cap \ldots \cap X_n)] =$$

$$= p^k q^{n-k} + p^k q^{n-k} + \cdots + p^k q^{n-k} = \binom{n}{k} p^k q^{n-k} \qquad (k = 0,1,2,\ldots,n)$$

Sea T la variable aleatoria que representa el número de tiradas hasta la aparición por primera vez del resultado "cara":

$$(T = n) = (X_1 \cap X_2 \cap \ldots \cap X_{n-1} \cap C_n) \qquad (n = 1,2,3,\ldots)$$

Su distribución de probabilidades será de la forma:

$$p_T(n) = \begin{cases} pq^{n-1} & si \quad n = 1,2,\ldots. \\ 0 & \qquad\quad n = 0 \end{cases}$$

Se comprueba fácilmente que:

$$\sum_{n=1}^{\infty} p_T(n) = \sum_{n=1}^{\infty} pq^{n-1} = p\sum_{n=1}^{\infty} q^{n-1} = \frac{p}{1-q} = 1$$

La función de distribución queda, entonces, definida de la siguiente forma:

$$F_T(n) = p(T \leq n) = 1 - p(T > n) = 1 - p(T \geq n+1)$$

$$= 1 - \sum_{j=n+1}^{\infty} p(T = j) =$$

$$= 1 - \sum_{j=n+1}^{\infty} pq^{j-1} = 1 - pq^n \sum_{j=0}^{\infty} \frac{pq^n}{1-q} = 1 - q^n$$

Un problema interesante relacionado con la variable T, es el referido a la distribución de la denominada "espera residual", es decir, a la probabilidad de que el resultado "cara" se produzca al cabo de n2 lanzamientos de la moneda, si ya sufre un retraso de n1 tiradas (en otros términos, la probabilidad de que el tiempo de espera del resultado "cara" sea n1 + n2, sabiendo que debe ser mayor que n1). De acuerdo con el análisis precedente, se tiene que:

$$p[T = n_1 + n_2/(T > n_1)] = \frac{p[(T = n_1 + n_2) \cap (T > n_1)]}{p(T > n_1)} =$$

$$= \frac{p(T = n_1 + n_2)}{p(T > n_1)} =$$

$$= \frac{pq^{n_1+n_2-1}}{q^{n_1}} = pq^{n_2-1} = p(T > n_2)$$

De lo que se concluye que la variable espera residual tiene la misma distribución de probabilidades que la variable "espera desde el inicio". Un resultado que contradice la idea intuitiva que el tiempo de espera ya transcurrido debería de alguna forma reducir el tiempo de espera no transcurrido. En realidad, debido a la independencia postulada entre estos eventos, el tiempo de espera se reproduce con las mismas características que poseía al inicio, sin que se vea afectado por la extensión del tiempo de espera ya transcurrido.

Asociada a una variable aleatoria X es posible definir también una "función de supervivencia":

$$S_X(x_i) = 1 - F_X(x_i) = p(X > x_i) = \sum_{x>x_i} p_j \qquad (x_i \epsilon \Omega(X))$$

Se denomina "espectro" de una variable aleatoria al conjunto de puntos con probabilidad no-nula; es decir, al conjunto de puntos de discontinuidad de Fx(X). Se puede concluir en forma inmediata que, para una variable aleatoria discreta, el espectro coincide con su dominio (Ω(X)) y, por lo tanto, tiene probabilidad igual a uno.

Como se verá en la sección siguiente, para una variable aleatoria continua el espectro es vacío y con probabilidad nula. Por lo que se puede concluir que como máximo el espectro puede ser un conjunto numerable. Sea una variable X tal que:

$$F_X(x) = I_{(x>a)=} \begin{cases} 0 & si \quad x \leq a \\ 1 & si \quad x > a \end{cases}$$

(donde l(x>a) denota una función indicadora del conjunto (x>a)). Su espectro está definido por el punto a, de modo que Fx (X) representa la función de distribución del caso límite de una variable aleatoria cuto dominio es Ω(X) = {a} y su función de probabilidades es p(X = a) = 1. Estas variables se denominan "constantes" o "quasi constantes" o "quasi constantes con certeza" o "degeneradas" o "concentradas en un punto". (Landro & Gonzalez, 2018)

c) Las variables aleatorias unidimensionales continuas

Se dice que X es una variable aleatoria continua si su función de distribución, Fx (x) (no-decreciente), es continua para todo xeR1 y si existe, además, una "función de densidad de probabilidad" $(f_X(\theta, x))$ no-negativa tal que:

$$F_X(x) = p(X \leq x) = \int_{-\infty}^{x} f_X(\theta, y)dy \qquad (\forall x \epsilon \mathbb{R})$$

(donde ɵ denota un vector de parámetros).

De modo que, si Fx (x) es absolutamente continua, será:

$$f_X(\theta, x) = \frac{d}{dx} F_X(x) =$$

$$= \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = \lim_{\Delta x \to 0} p(x < X < x + \Delta x) \quad (\Delta x > 0)$$

Si bien desempeña el mismo rol que la función pi (i = 1,2, …) para las variables aleatorias discretas, estrictamente hablando, $(f_X(\theta, x)$ no es una función de probabilidades, ya que, para las variables continuas la probabilidad en cada punto es nula:

$$p(X = a) = \int_{a}^{a} f_X(x)dx = 0$$

Se puede concluir entonces que, en este caso, por razones puramente matemáticas, el concepto de probabilidad resulta en si mismo insuficiente para atribuir distintos grados

de confiabilidad a los diferentes eventos de la partición que define a la variable aleatoria, ya que los grados de creencia estan todos identificados con la probabilidad del evento imposible. En consecuencia, se conviene que, si fx (x) ≥ fx (y), debe interpretarse que el grado de confiabilidad en la ocurrencia del evento (X = x) es mayor o igual que el correspondiente al evento (X 0 y) aun sabiendo que a los dos eventos les corresponde la misma probabilidad, p(X = x) = p(X = y) = 0)8.

De acuerdo con una interpretación de la probabilidad en términos de teoría de la medida, una probabilidad nula es asimilable a la probabilidad de acertar en un blanco definido por un punto con una flecha cuya punta tiene espesor cero. Se genera entonces una paradoja que podría ser expresada de la siguiente forma: si es imposible acertar en cada punto, ¿Cómo se puede explicar que la probabilidad en certeza, de ignorar que entre un evento "posible", aunque de probabilidad nula y un evento "imposible" existe una diferencia de carácter cualitativo?

Según la definición Boreliana, la probabilidad de que la variable X asuma valores incluidos en un conjunto cualquiera de su dominio queda definida por la siguiente integral de Lebesgue-Stieljes:

$$p(X \epsilon B) = \int_B dF_X(x) = \int_B f_X(\theta, x)dx \qquad (B \epsilon \mathbb{R}_1)$$

De acuerdo con el concepto de densidad y teniendo en cuenta la propiedad de aditividad compleja, será:

$$f_X(\theta, x)dx = \int_x^{x+dx} f_X(\theta, u)du = p(x \leq X \leq x + dx)$$

La función de densidad posee las siguientes propiedades:

1) $f_X(\theta, x) \geq 0 \qquad (\forall x \epsilon \mathbb{R}_1)$
2) $\int_{-\infty}^{\infty} f_X(\theta, x)\, dx = 1$

Estas condiciones son suficientes para poder asegurar que $f_X(\theta, x)$ es una función de densidad. Es decir, si una función $f_X(\theta, x)$ en R1 cumple estas condiciones, entonces se puede asegurar que la función $f_X(\theta, x)$ definida en la página precedente, es una función de distribución de la variable X y, por lo tanto, que $f_X(\theta, x)$ define una función de densidad. Observece que $f_X(\theta, x)$ dx es el valor del área de un rectángulo de base dx y altura proporcional a $f_X(\theta, x)$. En gneral, dada una variable aleatoria X que existe en el intervalo Ω (X) = (c,d), la probabilidad de que X asuma valores comprendidos en un ntervalo (a,b) (c≤a≤b≤d), puede ser representada como la suma de las áreas de n rectángulos (n - ∞) de base dx:

$$p(a \leq X \leq b) = \lim_{n \to \infty} \sum_{i=1}^{n} f_X(\xi_i)\, dx = p\left(\bigcup_{x \epsilon [a,b]} E_X \right) = \int_a^b f_X(x)dx =$$

$$= \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)\,dx = F_X(b) - F_X(a) =$$

$$= \Delta_a^b F_X(x) \qquad (-\infty < a < b < \infty)$$

Sea una variable aleatoria X con función de distribución Fx (x) (xeΩ(X)), entonces toda potencia positiva (a > 0) $F_X^a(x)$ es una función de distribución. Una propiedad similar se cumple para las funciones de supervivencia. (Landro & Gonzalez, 2018)

2.1.3. Eventos y espacio muestral

Dado el experimento aleatorio E. El espacio muestral omega, Ω, de E, es el conjunto de todos los posibles resultados de E.

Ω es un espacio muestral discreto si es un conjunto contable, es decir, si es un conjunto finito o infinito contable.

Ω es un espacio muestral continuo si no es contable.

Un evento es cualquier subconjunto de Ω.

Un evento elemental es un evento constituido por un solo elemento de Ω.

Observe que:

Ejemplo 1: En el experimento E que consiste en lanzar una moneda honesta, no cargada, es aleatorio, ya que:

a) Se saben los resultados que se pueden obtener: águila o sol,
b) En cada lanzamiento o ensayo, no se conoce con seguridad cual de las dos caras va a salir y
c) Puede considerarse que los lanzamientos se realizan bajo idénticas condiciones.

El espacio muestral es:

$$\Omega = \{a, s\},$$

Donde a denota que la cara mostrada al lanzar la moneda es águila y s que es sol. Como los elementos de Ω se pueden contar, Ω es un espacio discreto. (a) y (s) son eventos de Ω: $\{a\} \subseteq \Omega$, $\{s\} \subseteq \Omega$, y son elementales, su unión es Ω y su intersección es el ɵ.

Ejemplo 2: ¿Cuáles de los siguientes experimentos son aleatorios?

E!: Lanzar un dado no cargado y anotar el número que se muestra.

E2: Contar las páginas de un libro de 87 hojas.

E3: Medir el tiempo de vida de un foco.

E4: Seleccionar un hombre de un equipo de basquetbol varonil.

El espacio muestral es:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Y es discreto ya que sus elementos se pueden contar. El evento A constituido por los números pares es A = {2,4,6}, el B integrado por los números menores que 3 es B = {1,2}, el Ci formado por el numero i, (i = 1,2, …,6) es Ci = {ili = 1,2,…,6}, o bien, C1 = {1}, C2 = {2}, … , C6 = {6}, y el D, el de los números primos es D = {2,3,5}. De todos esos eventos solo los Ci son elementales, su unión es todo **Ω** y su interacción es el conjunto vacío, es decir,

$$\cup_{i=1}^{6} C_i = \Omega \quad \text{y} \quad \cap_{i=1}^{6} C_i = \emptyset.$$

Además, como $A^c = \{1, 3, 5\},\ A \cup D = \{2, 3, 4, 5, 6\},\ A \cap B = \{2\}$ son subconjuntos de **Ω**, entonces también son eventos. Por otro lado, el conjunto E formado por los números mayores que 6, también es un evento de **Ω** (¿Por qué?), pero F = {7} no lo es (¿Por qué?). (Landro & Gonzalez, 2018)

2.1.4. Enfoques de probabilidad

Concepción Clásica: Se define la probabilidad "a priori", de una manera teórica:

$$P(A) = \frac{m}{n} = \frac{No\ de\ eventos\ simples\ favorables}{No\ de\ eventos\ simples\ posibles}$$

La probabilidad clásica supone una especie de simetría en el mundo, posición que ocasiona muchos problemas. Por las situaciones de la vida real, desordenadas y poco probables como son a menudo, hace que definamos la probabilidad de otras maneras.

Concepción Subjetiva: La probabilidad es el grado de confianza que cada persona atribuye a un evento aleatorio. Tiene sentido intuitivo, no proporciona una definición estricta de probabilidad.

Concepción Estadística: Se define la probabilidad "a posteriori", después de haber hecho muchos experimentos. Parte del concepto de frecuencia relativa, utiliza el concepto empírico que resulta al contar m eventos simples favorables producidos en n pruebas.

$$hi = P(A) = \frac{m}{n} = \frac{No\ de\ eventos\ que\ han\ resultado\ triunfales}{No\ total\ de\ eventos\ simples\ que\ han\ ocurrido}$$

La definición estadística, aunque útil en la práctica, tiene dificultades desde el punto de vista matemático, puesto que puede no existir un numero limite. Por esta razón, la moderna teoría de probabilidad ha sido desarrollada AXIOMATICAMENTE. (Landro & Gonzalez, 2018)

Concepción Axiomática: Se fundamenta en el Algebra Abstracta.

P(A) ≥ 0	La probabilidad P(A) es un numero positivo
P(**Ω**) = 1	La probabilidad de un suceso seguro, **Ω**, es la unidad
P(ɵ) = 1	La probabilidad del suceso imposible ɵ, es nula
0 < P(A) < 1	La probabilidad de cualquier otro suceso elemental A
P(A)+P(cA) = 1	P(A) = 1 – P(cA) o (p + q = 1)

2.2. Propiedades y teoremas de la probabilidad

2.2.1. Propiedades de la probabilidad

De la definición de probabilidad se deducen algunas propiedades muy útiles.

a) La probabilidad del vicio es cero: $\Omega = \Omega \cup \emptyset \cup \emptyset \cup \ldots$ y por la aditividad numerable, $P(\Omega) = P(\Omega) + \sum_{k>1} P(\emptyset)$, de modo que $P(\emptyset) = 0$.

b) Adividad finita: Si A1, …, An son elementos disjuntos de A, aplicando la σ-aditividad, la propiedad anterior y haciendo $A_i = \emptyset,\ i > n$ tendremos.

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Se deduce de aquí fácilmente que $\forall A \in \mathcal{A},\ P(A^c) = 1 - P(A)$.

c) Monotonía: Si $\text{Si } A, B \in \mathcal{A},\ A \subset B,$ entonces de P(B) = P(A) + P(B – A) se deduce que P(A) ≤ P(B).

d) Probabilidad de una unión cualquiera de sucesos (formula de inclusión-exclusión): Si $A_1, \ldots, A_n \in \mathcal{A},$ entonces

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n).$$

Para su obtención observemos que si n = 2 es cierta, pues

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2 - A_1) = P(A_1) + P(A_2 - A_1) + P(A_1 \cap A_2) - P(A_1 \cap A_2) \\ &= P(A_1) + P(A_2) - P(A_1 \cap A_2). \end{aligned}$$

El resto se sigue por inducción.

e) Subactividad: Dados los sucesos A1, …, An, la relación existente entre la probabilidad de la unión de los Ai y la probabilidad de cada uno de ellos es la siguiente:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

En efecto, sean $B_1 = A_1$ y $B_i = A_i - \bigcup_{j=1}^{i-1} A_j$ para i = 2, …, n. Los Bi son disjuntos y $\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$. Por la aditividad finita y la monotonía de P se tiene

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i) \leq \sum_{i=1}^{n} P(A_i).$$

Si se trata de una sucesión de sucesos, $\{A_n\}_{n\geq 1},$ se comprueba, análogamente, que

$$P\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} P(A_n).$$

f) Continuidad de la probabilidad: Sea $\{A_n\}_{n\geq 1}$, una sucesión monótona creciente de sucesos y sea A su límite. Es decir, $A_n \subset A_{n+1}$, $\forall n$ y $\bigcup_{n>1} A_n = A$ (que en lo que sigue denotaremos mediante $A_n \uparrow A$). Si a partir de la sucesión inicial definimos $B_n = A_n - \bigcup_{i=1}^{n-1} A_i = A_n - A_{n-1}$, para $n > 1$, y $B_1 = A_1$, se tiene (Montes Suay, 2007)

$$P(A) = P\left(\bigcup_{n\geq 1} A_n\right) = P\left(\bigcup_{n\geq 1} B_n\right) = \sum_{n\geq 1} P(B_n) =$$

$$\lim_{n\to+\infty} \sum_{j=1}^{n} P(B_j) = \lim_{n\to+\infty} P\left(\bigcup_{j=1}^{n} B_j\right) = \lim_{n\to+\infty} P(A_n),$$

2.2.2. Independencia y condicional

- **Probabilidad Condicionada. Teorema de Bayes:**

Si compramos un numero para una rifa que se celebra anualmente durante las fiestas de verano en nuestro pueblo y que está compuesta por 100 boletos numerados del 1 al 100, sabemos que nuestra probabilidad ganar el premio, suceso que designaremos por A, vale

$$P(A) = \frac{1}{100}$$

Supongamos que a la mañana siguiente de celebrarse el sorteo alguien nos informa que el boleto premiado termina en 5. Con esta información, ¿continuaremos pensando que nuestra probabilidad de ganar vale 10^{-2}? Desde luego seria absurdo continuar pensando lo si nuestro número termina en 7, porque evidentemente la nueva probabilidad valdría P(A) = 0, pero, aunque terminara en 5 también nuestra probabilidad de ganar habría cambiado, porque los números que terminan en 5 entre los 100 son 10 y entonces

$$P'(A) = \frac{1}{10},$$

10 veces mayor que la inicial.

Supongamos que nuestro número es el 35 y repasemos los elementos que han intervenido en la nueva situación. De una parte, un suceso original $A = \{ganar\ el\ premio\ con\ el\ número\ 35\}$, de otra, un suceso $B = \{el\ boleto\ premiado\ termina\ en\ 5\}$ de cuya ocurrencia se nos informa a priori. Observemos que $A \cap B = \{el\ número\ 35\}$ y que la nueva probabilidad encontrada verifica,

$$P'(A) = \frac{1}{10} = \frac{1/100}{10/100} = \frac{P(A \cap B)}{P(B)},$$

Poniendo en evidencia algo que cabía esperar, que la nueva probabilidad a depender de P(B). Estas propiedades observadas justifican la definición que damos a continuación.

Definición 1. Sea (Ω, A, P) un espacio de probabilidad y sean A y B dos sucesos, con P(B) > 0, se define la probabilidad de A condicionada a B mediante la expresión,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

A la anterior expresión se la denomina probabilidad con toda justicia, porque verifica los tres axiomas que definen el concepto de probabilidad, como fácilmente puede comprobarse. De entre los resultados y propiedades que se derivan de este nuevo concepto, tres son especialmente relevantes: el teorema de factorización, el teorema de la probabilidad total y el teorema de Bayes. (Montes Suay, 2007)

a) Teorema de Factorización:

A partir de la definición de probabilidad condicionada, la probabilidad de la intersección de dos sucesos puede expresarse de la forma $P(A \cap B) = P(A|B)P(B)$. El teorema de factorización extiende este resultado para cualquier intersección finita de sucesos.

Consideremos los sucesos A1, A2, …, An, tales que $P(\cap_{i=1}^{n} A_i) > 0$, por inducción se comprueba fácilmente que

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P\left(A_n | \cap_{i=1}^{n-1} A_i\right) P\left(A_{n-1} | \cap_{i=1}^{n-2} A_i\right) \ldots P(A_2|A_1)P(A_1).$$

Ejemplo 1: En una urna que contiene 5 bolas blancas y 4 negras, llevamos a cabo 3 extracciones consecutivas sin reemplazamiento. ¿Cuál es la probabilidad de las 2 primeras sean blancas y la tercera negra?

Cada extracción altera la composición de la urna y el total de bolas que contiene. De acuerdo con ello tendremos (la notación es obvia)

$$P(B_1 \cap B_2 \cap N_3) = \\ = P(N_3|B_1 \cap B_2)P(B_2|B_1)P(B_1) = \tfrac{4}{7} \cdot \tfrac{4}{8} \cdot \tfrac{5}{9}$$

b) Teorema de la probabilidad total:

Si los sucesos A1, A2, …, An constituyen una partición del Ω, tal que P(Ai) > 0, $\forall i$, tendremos que cualquier suceso B podrá particionarse de la forma, $B = \cup_{i=1}^{n} B \cap A_i$ y tratándose de una unión disjunta podremos escribir

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Este resultado se conoce con el nombre de teorema de la probabilidad total.

Encuesta sobre cuestiones delicadas: una aplicación del teorema de la probabilidad total

Es bien conocida la reticencia de la gente a contestar cualquier encuesta, reticencia que se convierte en clara desconfianza y rechazo si el cuestionario aborda lo que podríamos denominar temas delicados: situación económica, creencias religiosas, afinidades políticas, costumbres sexuales, consumo de estupefacientes, … El rechazo y la desconfianza están casi siempre basados en la creencia de una insuficiente garantía de anonimato. Es comprensible, por tanto, el afán de los especialistas en convencer a los encuestados de que el anonimato es absoluto. El teorema de la probabilidad total puede ayudar a ello.

Supongamos que un sociólogo está interesado en conocer el consumo de drogas entre los estudiantes de un Instituto de Bachillerato. Elige 100 estudiantes al azar y para garantizar la confidencialidad de las respuestas, que sin duda recundara en un resultado mas fiable,

diseña una estrategia consistente en que cada estudiante extrae al azar una bola de un saco o urna que contiene 100 bolas numeradas del 1 al 100, conservándola sin que nadie la vea,

- Si el numero de la bola elegida esta entre el 1 y el 70, contesta a la pregunta ¿has consumido drogas alguna vez?,
- Si el numero de la bola elegida esta entre el 71 y el 100, contesta a la pregunta ¿es par la última cifra de tu DNI?

En ambos casos la repuesta se escribe sobre un trozo de papel sin indicar, lógicamente, a cuál de las dos preguntas se está contestando.

Realizado el proceso, las respuestas afirmativas han sido 25 y para estimar la proporción de los que alguna vez han consumido droga aplicamos (1.2), (Montes Suay, 2007)

$$\begin{aligned} P(sí) &= P(sí|pregunta\ delicada)P(pregunta\ delicada)+ \\ & P(sí|pregunta\ intrascendente)P(pregunta\ intrascendente) \end{aligned}$$

Sustituyendo,

$$0{,}25 = P(sí|pregunta\ delicada) \times 0{,}7 + 0{,}5 \times 0{,}3,$$

Y despejando,

$$P(sí|pregunta\ delicada) = \frac{0{,}25 - 0{,}15}{0{,}7} \approx 0{,}14$$

c) Teorema de Bayes:

Puede tener interés, y de hecho así ocurre en muchas ocasiones, conocer la probabilidad asociada a cada elemento de la partición dado que ha ocurrido B, es decir, P(Ai/B). Para ello, recordemos la definición de probabilidad condicionada y apliquemos el resultado anterior.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}.$$

Este resultado, conocido como el teorema de Bayes, permite conocer el cambio que experimenta la probabilidad de Ai como consecuencia de haber ocurrido B. En el lenguaje habitual del Cálculo de Probabilidades a P(Ai) se la denomina probabilidad a priori y a P(Ai/B) probabilidad a posteriori, siendo la ocurrencia de B la que establece la frontera entre el antes y el después. ¿Cuál es, a efectos prácticos, el interés de este resultado? Veámoslo con un ejemplo.

Ejemplo 1: Tres urnas contienen bolas blancas y negras. La composición de cada una de ellas es la siguiente: $U_1 = \{3B, 1N\}$, $U_2 = \{2B, 2N\}$, $U_3 = \{1B, 3N\}$. Se elige al azar una de las urnas, se extrae de ella una bola al azar y resulta ser blanca. ¿Cuál es la urna con mayor probabilidad de haber sido elegida?

Mediante U1, U2 y U3, representaremos también la urna elegida. Estos sucesos constituyen una partición de **Ω** y se verifica, puesto que la elección de la urna es al azar,

$$P(U_1) = P(U_2) = P(U_3) = \frac{1}{3}.$$

Si $B=\{\text{la bola extraída es blanca}\}$, tendremos

$$P(B|U_1)=\frac{3}{4},\quad P(B|U_2)=\frac{2}{4},\quad P(B|U_3)=\frac{1}{4}.$$

Lo que nos piden es obtener P(Ui/B) para conocer cuál de las urnas ha originado, más probablemente, la extracción de la bola blanca. Aplicando el teorema de Bayes a la primera de las urnas,

$$P(U_1|B)=\frac{\frac{1}{3}\cdot\frac{3}{4}}{\frac{1}{3}\cdot\frac{3}{4}+\frac{1}{3}\cdot\frac{2}{4}+\frac{1}{3}\cdot\frac{1}{4}}=\frac{3}{6},$$

Y para las otras dos, P(U2/B) = 2/6 y P(U3/B) = 1/6. Luego la primera de las urnas es la que con mayor probabilidad dio lugar a una extracción de bola blanca.

El teorema de Bayes es uno de aquellos resultados que inducen a pensar que la cosa no era para tanto. Se tiene ante el la sensación que produce lo trivial, hasta el punto de atrevernos a pensar que lo hubiéramos podido deducir nosotros mismos de haberlo necesitado, aunque afortunadamente el Reverendo Thomas Bayes se ocupo de ello en un trabajo titulado An Essay towards solving a Problem in the Doctrine of Chances, publicado en 1763. Conviene precisar que Bayes no planteo el teorema en su forma actual, que es debida a Laplace. (Montes Suay, 2007)

- **Independencia:**

La información previa que se nos proporcionó sobre el resultado del experimento modifico la probabilidad inicial del suceso. ¿Ocurre esto siempre? Veámoslo.

Supongamos que, en lugar de comprar un único boleto, el que lleva el número 35, hubiéramos comprado todos aquellos que terminan en 5. Ahora P(A) = 1/10 puesto que hemos comprado 10 boletos, pero al calcular la probabilidad condicionada a la información que se nos ha facilitado, $B=\{el\ boleto\ premiado\ termina\ en\ 5\}$, observaremos que $P(A\cap B)=1/100$ porque a la intersección de ambos sucesos es justamente el boleto que esta premiado, en definitiva

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{1/100}{10/100}=\frac{1}{10},$$

La misma que originalmente tenia A. Parecen existir situaciones en las que la información previa no modifica la probabilidad inicial del suceso. Observemos que este resultado tiene una consecuencia inmediata,

$$P(A\cap C)=P(A|C)P(C)=P(A)P(C).$$

Esta es una situación de gran importancia en probabilidad que recibe el nombre de independencia de sucesos y que generalizamos mediante la siguiente definición.

a. Sucesos independientes: Sean A y B dos sucesos. Decimos que A y B son independientes si $P(A\cap B)=P(A)P(B).$

De esta definición se obtiene como propiedad,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A),$$

Y su simetría P(B/A) = P(B).

En ocasiones se define la independencia de dos sucesos a partir de este resultado, obteniéndose entonces como propiedad la que nosotros hemos dado como definición. Existe equivalencia entre ambas definiciones, aunque a fuerza de ser rigurosos, hay que matizar que definir el concepto a partir de la probabilidad condicional exige añadir la condición de que el suceso condicionante tenga probabilidad distinta de cero. Hay además otra ventaja a favor de la definición basada en la factorización de $P(A \cap B)$, pone de inmediato en evidencia la simetría del concepto. El concepto de independencia puede extenderse a una familia finita de sucesos de la siguiente forma.

b. Independencia mutua: Se dice que los sucesos de la familia (A1, …, An) son mutuamente independientes cuando

$$P(A_{k_1} \cap \ldots \cap A_{k_m}) = \prod_{i=1}^{m} P(A_{k_i}) \quad (1.6)$$

Siendo $\{k_1, \ldots, k_m\} \subset \{1, \ldots, n\}$ *y los* k_i *distintos.*

Conviene señalar que la independencia mutua de los n sucesos supone que han de verificarse $\binom{n}{n} + \binom{n}{n-1} + \ldots \binom{n}{2} = 2^n - n - 1$ ecuaciones del tipo dado en (1.6).

Si solamente se verificasen aquellas igualdades que implican a dos elementos diríamos que los sucesos son independientes dos a dos, que es un tipo de independencia menos restrictivo que el anterior como pone de manifiesto el siguiente ejemplo. Solo cuando n = 2 ambos conceptos son equivalentes.

Ejemplo 1: Tenemos un tetraedro con una cara roja, una cara negra, una cara blanca y la cuarta cara pintada con los tres colores. Admitimos que el tetraedro esta bien construido, de manera que al lanzarlo sobre una mesa tenemos la misma probabilidad de que se apoye sobre una cualquiera de las cuatro caras, a saber, $p = \frac{1}{4}$. El experimentó consiste en lanzar el tetraedro y ver en qué posición ha caído. Si

$$R = \{el\ tetraedro\ se\ apoya\ en\ una\ cara\ con\ color\ \text{rojo}\}$$
$$N = \{el\ tetraedro\ se\ apoya\ en\ una\ cara\ con\ color\ \text{negro}\}$$
$$B = \{el\ tetraedro\ se\ apoya\ en\ una\ cara\ con\ color\ \text{blanco}\},$$

Se comprueba fácilmente que son independientes dos a dos, pero no son mutuamente independientes.

El tipo de independencia habitualmente exigida es la mutua, a la que nos referiremos simplemente como independencia.

Digamos, por último, que si la colección de sucesos es infinita diremos que son independientes cuando cualquier su colección finita lo sea. (Montes Suay, 2007)

- **Independencia de clases de sucesos:**

Si A y B son sucesos independientes, ni A ni B nos proporcionan información sobre la ocurrencia del otro. Parece lógico que tampoco nos digan mucho sobre los complementarios respectivos. La pregunta es ¿son A y B independientes? La respuesta afirmativa la deducimos a continuación.

$$P(A \cap B^c) = \\ = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c).$$

Del mismo modo se comprueba que A^c es independiente tanto de B como de B^c. Resulta así que los conjuntos de sucesos $\{A, A^c\}$ y $\{B, B^c\}$ son independientes en el sentido que al tomar un suceso de cada una de ellos, los sucesos son independientes. De forma mas general podemos hablar de clases independientes de sucesos.

Definición 1: Clases independientes de sucesos: Las clases de sucesos A1, …, An c A se dicen independientes, si al tomar Ai en cada Ai, i = 1, …, n, los sucesos de la familia $\{A_1, \ldots, A_n\}$ son independientes.

Notemos que en la definición no se exige que los elementos de cada clase Ai sean independientes entre sí. De hecho, A Y A^c solo lo son si P(A) = 0 o P(A) = 1.

Para una colección infinita de clases de sucesos la anterior definición se extiende con facilidad. Diremos que $\{\mathcal{A}_n\}_{n\geq 1} \subset \mathcal{A}$ son independientes si cualquier subcolección finita lo es. (Montes Suay, 2007)

2.2.3. Teorema del límite central

Una aplicación inmediata es el Teorema de De Moivre-Laplace, una versión temprana del TCL, que estudia el comportamiento asintótico de una B(n,p).

a) Teorema de Moivre-Laplace: Sea $X_n \sim B(n,p)$ y definamos $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$. Entonces

$$Z_n \xrightarrow{L} N(0,1).$$

Demostración: Aplicando los resultados anteriores, se obtiene

$$\phi_{Z_n}(t) = \left((1-p)e^{-it\sqrt{\frac{p}{n(1-p)}}} + pe^{it\sqrt{\frac{(1-p)}{np}}}\right)^n,$$

Que admite un desarrollo en serie de potencias de la forma

$$\phi_{Z_n}(t) = \left[1 - \frac{t^2}{2n}(1 + R_n)\right]^n,$$

Con $R_n \to 0$, si $n \to \infty$. En consecuencia,

$$\lim_{n\to\infty} \phi_{Z_n}(t) = e^{-\frac{t^2}{2}}.$$

La unicidad y el teorema de continuidad hacen el resto.

b) Lo que el teorema afirma es que si $X \sim B(n,p),$ para n suficientemente grande, tenemos

$$P\left(\frac{X-np}{\sqrt{np(1-p)}} \leq x\right) \simeq \Phi(x),$$

Donde $\Phi(x)$ es la función de distribución de la N(0,1).

¿De que forma puede generalizarse este resultado? Como ya sabemos $X_n \sim B(n,p)$ es la suma de n v.a. i.i.d., todas ellas Bernoulli $(Y_k \sim B(1,p))$, cuya varianza común, var(Y1) = p(1-p), es finita. En esta dirección tiene lugar la generalización: variables independientes, con igual distribución y con varianza finita.

c) Lindeberg: Sean X1, X2, …, Xn, v.a. i.i.d. con media y varianza finitas, μ y σ^2, respectivamente. Sea $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ su media muestral, entonces

$$Y_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{var(\overline{X}_n)}} \xrightarrow{L} N(0,1).$$

Teniendo en cuenta la definición de Xn podemos escribir

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k,$$

Con Zk = (Xk - μ) / σ variables aleatorias i.i.d. con E(Z1) = 0 y var(Z1) = 1. Aplicando $\mathrm{P}\phi 4$ y

$$\phi_Y(t) = E\left(e^{it(X_1+X_2+\cdots+X_n)}\right) = E\left(\prod_{k=1}^{n} e^{itX_k}\right) = \prod_{k=1}^{n} E\left(e^{itX_k}\right) = \prod_{k=1}^{n} \phi_{X_k}(t),$$

Tendremos

$$\phi_{Y_n}(t) = \left[\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Pero existiendo los dos primeros momentos de Z1 y teniendo en cuenta

$$\phi_X(t) = \sum_{k\geq 0} \frac{i^k E(X^k)}{k!} t^k.$$

$\phi_{Z_1}(t)$ puede también expresarse de la forma

$$\phi_{Z_1}(t) = 1 - \frac{t^2}{2n}(1+R_n),$$

Con $R_n \to 0$, si $n \to \infty$. En consecuencia

$$\phi_{Y_n}(t) = \left[1 - \frac{t^2}{2n}(1 + R_n)\right]^n.$$

Así pues,

$$\lim_{n\to\infty} \phi_{Y_n}(t) = e^{-\frac{t^2}{2}},$$

Que es la función característica de una N(0,1).

Observemos que el Teorema de De Moivre-Laplace es un caso particular del Teorema de Lindeberg, acerca de cuya importancia se invita al lector a reflexionar porque lo que en el se afirma es, ni más ni menos, que sea cual sea la distribución común a las Xi, su media muestral Xn, adecuadamente tipificada, converge a una N (0,1) cuando n - ∞.

El teorema de Linderberg, que puede considerarse el teorema central del límite básico, admite una generalización en la dirección de relajar la condición de equidistribución exigida a las variables. Las llamadas condiciones de Lindeberg y Lyapunov muestran sendos resultados que permiten eliminar aquella condición.

Ejemplo 1: La fórmula de Stirling para aproximar n!: Consideremos una sucesión de variables aleatorias X1, X2, …, independientes e idénticamente distribuidas, Poisson de parámetro $\lambda = 1.$ La variable $S_n = \sum_{i=1}^{n} X_i$ es también Poisson con parámetro $\lambda_n = n.$ Si $Z \sim N(0,1),$ para n suficientemente grande el TCL nos permite escribir,

$$\begin{aligned} P(S_n = n) &= P(n-1 < S_n \leq n) \\ &= P\left(-\frac{1}{\sqrt{n}} < \frac{S_n - n}{\sqrt{n}} \leq 0\right) \\ &\approx P\left(-\frac{1}{\sqrt{n}} < Z \leq 0\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-1/\sqrt{n}}^{0} e^{-x^2/2} dx \\ &\approx \frac{1}{\sqrt{2\pi n}}, \end{aligned}$$

En donde la última expresión surge de aproximar la integral entre $[-1/\sqrt{n}, 0]$ de $f(x) = e^{-x^2/2}$ mediante el área del rectángulo que tiene por base el intervalo de integración y por altura el $f(0) = 1.$

Por otra parte,

$$P(S_n = n) = e^{-n}\frac{n^n}{n!}.$$

¡Igualando ambos resultados y despejando n! se obtiene la llamada formula de Stirling.

$$n! \approx n^{n+1/2} e^{-n} \sqrt{2\pi}.$$

d) Una curiosa aplicación del TCL: estimación del valor de π:

De Moivre y Laplace dieron en primer lugar una versión local del TCL al demostrar que si $X \sim B(n,p)$,

$$P(X=m)\sqrt{np(1-p)} \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad (4.12)$$

Para n suficientemente grande y $x = \frac{m-np}{\sqrt{np(1-p)}}$. Esta aproximación nos va a servir para estudiar la credibilidad de algunas aproximaciones al número π obtenidas a partir del problema de la aguja de Buffon. (Montes Suay, 2007)

Recordemos que en el problema planteado por Buffon se pretende calcular la probabilidad de que una aguja de longitud l, lanzada al azar sobre una trama de paralelas separadas entre si una distancia a, con a > 1, corte a alguna de las paralelas. Puestos de acuerdo sobre el significado de lanzada al azar, la respuesta es

$$P(corte) = \frac{2l}{a\pi},$$

Resultado que permite obtener una aproximación de π si, conocidos a y l, sustituimos en $\pi = \frac{2l}{aP(corte)}$ la probabilidad de corte por su estimador natural la frecuencia relativa de corte, p, a lo largo de n lanzamientos. Podremos escribir, si en lugar de trabajar con π lo hacemos con su inverso,

$$\frac{1}{\pi} = \frac{am}{2ln},$$

Donde m es el número de cortes en los n lanzamientos.

En el año 1901 Lazzarini realizo 3408 lanzamientos obteniendo para π el valor 3,1415929 con ¡6 cifras decimales exactas! La aproximación es tan buena que merece como mínimo alguna pequeña reflexión. Para empezar, supongamos que el número de cortes aumenta en una unidad, las aproximaciones de los inversos de π correspondientes a los m y m + 1 cortes diferirían en

$$\frac{a(m+1)}{2ln} - \frac{am}{2ln} = \frac{a}{2ln} \geq \frac{1}{2n},$$

Que si $n \approx 5000$, da lugar a $\frac{1}{2n} \approx 10^{-4}$. Es decir, un corte más produce una diferencia mayor que la precisión de 10^{-6} alcanzada. No queda más alternativa que reconocer que Lazzarini tuvo la suerte de obtener exactamente el número de cortes, m, que conducía a tan excelente aproximación. La pregunta inmediata es, ¿Cuál es la probabilidad de que ello ocurriera?, y para responderla podemos recurrir a (4.12) de la siguiente forma,

$$P(X=m) \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(m-np)^2}{2np(1-p)}} \leq \frac{1}{\sqrt{2\pi np(1-p)}}.$$

Por ejemplo, si a = 2l entonces p = 1/π y para P(X = m) obtendríamos la siguiente cota

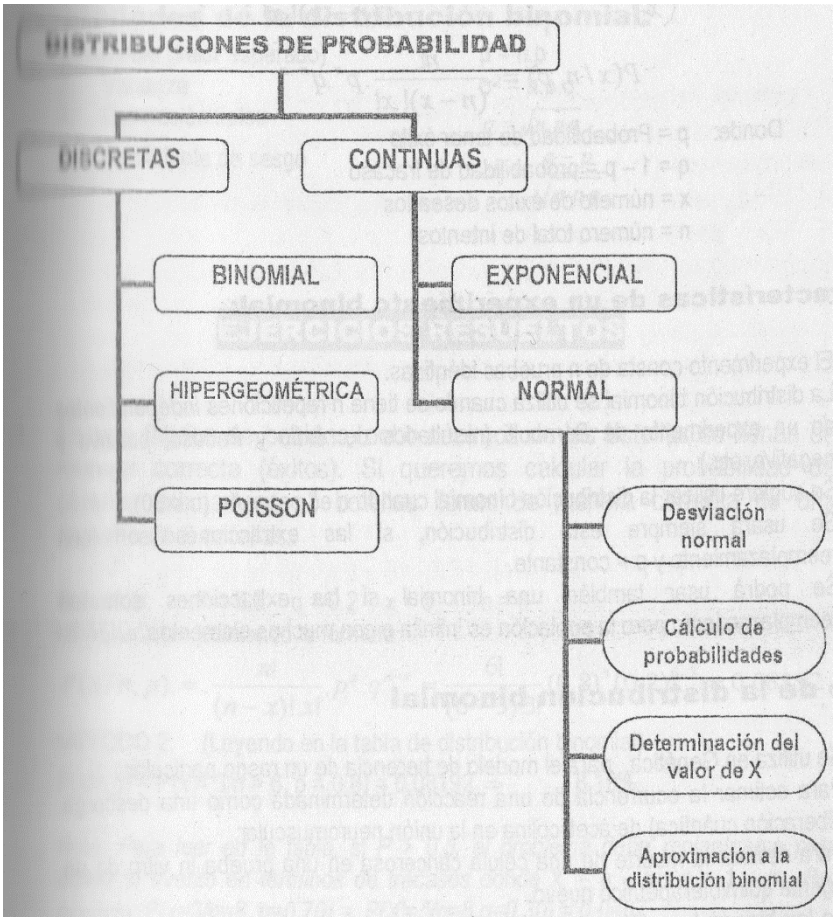
$$P(X = m) \leq \sqrt{\frac{\pi}{2n(\pi - 1)}}.$$

Para el caso de Lazzarini n = 3408 y P(X = m) ≤ 0,0146, $\forall m.$ Parece ser que Lazzarini era un hombre de suerte, quizás demasiada. (Montes Suay, 2007)

2.3. Distribuciones discretas

Ilustración 13.

Distribuciones de Probabilidad



Fuente: (Alvarez Roman, 2004)

Las distribuciones de probabilidad se utilizan para solucionar muchos problemas de los negocios, utilizando variables discretas y continuas. (Alvarez Roman, 2004)

2.3.1. Distribución Binomial

Sea X una v.a. que representa el número de éxitos de n pruebas y p la probabilidad de éxito. Se dice entonces que X tiene una distribución binominal con función de probabilidad. Se aplica como modelo en la toma de decisiones en condiciones de incertidumbre. (Alvarez Roman, 2004)

$$P(x) = \binom{n}{x} p^x q^{n-x}; P(x) = \frac{n!}{(n-x)!\,x!} p^x q^{n-x};$$

$$P(x/n,p) = \frac{n!}{(n-x)!\,x!} p^x q^{n-x};$$

Donde:

p = Probabilidad de tener éxito

q = 1-p = probabilidad de fracaso

x = número de éxitos deseados

n = número total de intentos

a) Características de un experimento binominal:
    a. El experimento consta de n pruebas idénticas.
    b. La distribución binomial se utiliza cuando se tiene n repeticiones independientes de un experimento de Bernoulli (resultados de: éxito y fracaso, positivo y negativo, etc.)
    c. Se sugiere utilizar la distribución binomial cuando n es pequeño (n < 30).
    d. Se usará siempre esta distribución, si las extracciones son con reemplazamiento y p = constante.
    e. Se podrá usar también una binomial si las extracciones son sin reemplazamiento, pero la población es infinita o con muchos elementos. (Alvarez Roman, 2004)

b) Uso de la distribución binomial:
    a. Se utiliza en Genética, para el modelo de herencia de un rasgo particular.
    b. Para estimar la ocurrencia de una reacción determinada como una descarga (liberación cuántica) de acetilcolina en la unión neuromuscular.
    c. Para estimar la muerte de una célula cancerosa en una prueba in vitro de un agente quimioterapéutico nuevo.
    d. La ley binomial es útil en el control de calidad. Si se desea determinar entre análisis defectuosos y no defectuosos. (Alvarez Roman, 2004)

El término "binomial" proviene de las probabilidades p(x;n,p) son términos del desarrollo del binomio. (Alvarez Roman, 2004)

$$(q+p)^n = (\frac{n}{0})q^n + (\frac{n}{0})pq^{n-1} + (\frac{n}{2})p^2q^{n-2} + \ldots + (\frac{n}{n})p^n = \sum_{x=0}^{n} p\,(x;n,p)$$

$$(q+p)^n = q^n + \frac{n}{1!}q^{n-1}p + \frac{n(n-1)}{2!}q^{n-2}p^2 + \frac{n(n-1)(n-2)}{3!}q^{n-3}p^3 + \ldots + p^n$$

$$(p+q)^n = \sum_{x=0}^{n} (\frac{n}{x})p^x q^{n-x} = \sum_{x=0}^{n} \frac{n!}{(n-x)!\,x!}\, p^x q^{n-x}$$

c) Propiedades de la distribución binomial:

a. Media (valor esperado) μ=n.p
b. Varianza $\sigma^2 = n.p.q$
c. Desviación típica $\sigma = \sqrt{n.p.q}$
d. Coeficiente de sesgo $\propto 3 = \frac{q-p}{\sqrt{n.p.q}}$

Ejemplo 1: En una fábrica de bebidas, ocho decimos de las botellas se llenan de manera correcta (éxitos). Si queremos calcular la probabilidad de obtener exactamente 3 botellas llenas de manera correcta de una muestra de 6 botellas. (Alvarez Roman, 2004)

P = 0.8 q = 0.2 x = 3 n = 6

Método 1. Utilizando la formula

$$P(x/n,p) = \frac{n!}{(n-x)!\,x!}p^x q^{n-x} = \frac{6!}{(6-3)!\,3!}(0.8)^3(0.2)^{6-3} = 0.08192$$

Método 2. (Leyendo en la tabla de distribución binomial)

P(x/n,p) = P(x=3/n=6,p=0.8) = 0.08192 = 8.19%

Para leer en la tabla, si P > 0.5, el problema debe replantearse para definir el evento en términos de fracasos donde X = n – x; q = 1 – p. Por ejemplo: P(x=3/n=8,p=0.70) = P(X=5/n=8,q=0.30) = 0.0467

Ejemplo 2: Supongamos que el 30% de turistas estudiados ingieren licor. Hallar la probabilidad de que, en una muestra aleatoria de 15 turistas, 5 hayan ingerido licor: (Alvarez Roman, 2004)

Exactamente 5 P(X = 5)

Método 1: Utilizando la formula

$$P(x/n,p) = \frac{n!}{(n-x)!\,x!}p^x q^{n-x} = \frac{15!}{(15-5)!\,5!}(0.3)^5(0.7)^{15-5} = 0.2061$$

Método 2. (Leyendo en la tabla de distribución binomial)

P(x/n,p) = P(x=5/n=15,p=0.3) = 0.2061 = 20.61%

Ejemplo 3: En general, la probabilidad de que prefieran un lugar turístico es del 20%. De una muestra aleatoria de 6 turistas. Cuál es la probabilidad de que prefieran: (Alvarez Roman, 2004)

a. Exactamente 4 P(X=4)
b. Mayor que 4 P(X>4)
c. Menor que 4 P(X<4)
d. Cuatro o menos P(X≤4)
e. Cuatro o más P(X≥4)
f. Un numero comprendido entre 3 y 5. P(3 ≤ x ≤ 5)

Soluciones:

1) Exactamente 4 P (X = 4)

P(x/n,p) = P(x=4/n=6,p=0.2) = 0.0154 = 1.54%

2) Mayor que 4 P(X>4)

P(x/n,p) = P(x=4/n=6,p=0.2) = P(x=5) + P(x=6) = 0.0015 + 0.0001 = 0.0016

3) Menor que 4 P(X<4)

P(x/n,p) = P(x=4/n=6,p=0.2) = P(x=0) + P(x=1) + P(x=2) + P(x=3) = 0.2621 + 0.3932 + 0.2458 + 0.0819 = 0.983

4) Cuatro o menos P(X≤4)

P(x/n,p) = P(x=4/n=6,p=0.2) = P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4) = 0.2621 + 0.3932 + 0.2458 + 0.0819 + 0.0154 = 0.983

5) Cuatro o más P(X≥4)

P(x/n,p) = P(x=4/n=6,p=0.2) = P(x=4) + P(x=5) + P(x=6) = 0.0154 + 0.0015 + 0.0001 = 0.017

6) Un numero comprendido entre 3 y 5. P(3≤x≤5)

P(3≤x≤5/n=6,p=0.2) = P(x=3) + P(x=4) + P(x=5) = 0.0819 + 0.0154 + 0.0015 = 0.0983

2.3.2. Distribución de Poisson

Es muy útil donde la variable aleatoria representa el número de eventos independientes que ocurren a una velocidad constante. Desde la teoría de los limites la distribución binomial se aproxima a la distribución de Poisson. Cuando el número de repeticiones de un experimento es muy grande (n ≥ 30) y se hace muy laboriosa la aplicación de la formula binomial. Cuando esto ocurre existen dos distribuciones teóricas que se aproximan a la distribución binomial, una de ellas es cuando p es muy pequeña, la cual se conoce como distribución de Poisson y se define así: (Alvarez Roman, 2004)

$$P(X/\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \;\; o \;\; P(x) = \frac{e^{-(n.p)}(n.p)^{x}}{x!}$$

(En poblaciones grandes $\lambda = \mu = n.p = \sigma^2$)

Donde:

P(x) = Probabilidad de tener exactamente x presentaciones.

$\lambda = n.p = El\ numero\ medio\ de\ presentaciones\ por\ intervalo\ de\ tiempo.$

$e =$ 2.71828.. (Base de los logaritmos naturales o neperianos).

$x! = x\ factorial.$

a) Casos donde se utiliza:
   a. El número de artículos defectuosos en una hora de producción.
   b. El número de automóviles que llegan a una caseta de cobro en 1 hora.

c. El número de llamadas telefónicas en una central, durante ciertas horas.
d. El número de accidentes registrados en la intersección de dos calles.
e. Atención medica que requieren los pacientes en un hospital en una determinada, etc. (Alvarez Roman, 2004)

b) Características:
a. El experimento consiste en contar el número de veces que ocurre un evento, en particular durante una unidad de tiempo dada, o en un área o volumen (o peso, distancia o cualquier otra medida) dada.
b. La probabilidad de que un evento ocurra en una unidad dada de tiempo, área o volumen es la misma para todas.
c. El número de eventos que ocurren en una unidad de tiempo, área o volumen es independiente del número de los que ocurren en otras unidades.
d. El número medio (o esperado) de eventos en cada unidad se denota por la letra griega LAMBDA (λ). (Alvarez Roman, 2004)

Ejemplo 1: En una fábrica se realizan ciertos análisis y la probabilidad de que una pieza sea defectuosa es del 0.02. Calcular la probabilidad de encontrar en un lote de 100 piezas que 2 sean defectuosas. (Alvarez Roman, 2004)

$$n = 100, \quad p = 0.02. x = 2$$

$$\lambda = np = (100)(0.02) = 2 \quad (2 < 5, por\ tanto\ la\ aproximacion\ es\ adecuada)$$

Método 1: Utilizando la formula

$$P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad P(x = 2) = \frac{e^{-2}2^2}{2!} = \frac{4e^{-2}}{2} = 0.2707$$

Método 2: Utilizando la tabla de Poisson: $P(x = 2/\lambda = 2) = 0.2707$

Método 3: Utilizando tablas acumulativas de distribución de Poisson

$$P(X, \lambda) = F(x, \lambda) - F(x - 1, \lambda)$$

Para $\lambda = 2, P(x = 2) = P(x \leq 2) - P(x \leq 1) = 0.677 - 0.406 = 0.271$

Ejemplo 2: La probabilidad de que un equipo que se utiliza en un hotel se descomponga al cabo de 900 horas de trabajo es 0.004. si se seleccionan al azar 1000 equipos con 900 horas o mas de trabajo, calcular la probabilidad de que se descompongan 4 equipos. (Alvarez Roman, 2004)

$$n = 1000, \quad p = 0.004, \quad x = 4, \qquad \lambda = n.p = 1000.0{,}004 = 4 < 5$$

$$P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{(2.71828)^{-4}.4^4}{4!} = \frac{(2.71828)^{-4}.256}{4!} = 0.1954$$

Método 2: Utilizando las tablas acumulativas de la distribución de Poisson

$$P(X, \lambda) = F(x, \lambda) - F(x - 1, \lambda)$$

Para $\lambda = 4, P(x = 4) = P(x \leq 4) - P(x \leq 3) = 0.629 - 0.433 = 0.196$

Ejemplo 3: Se sabe en un banco, que dos clientes en promedio por mes, dan información incorrecta ¿Cuál es la probabilidad de que en un mes: (Alvarez Roman, 2004)

i) Ningún cliente de información incorrecta (x=0).
ii) Un cliente de información incorrecta (x=1).
iii) Dos clientes den información incorrecta (x=2).
iv) Tres clientes den información incorrecta (x=3).

Datos: $\lambda = 2$

i) $P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{(2.71828)^{-2}.2^0}{0!} = 0.1353$

ii) $P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{(2.71828)^{-2}.2^1}{1!} = 0.2706$

iii) $P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{(2.71828)^{-2}.2^2}{2!} = 0.2706$

iv) $P(x/\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{(2.71828)^{-2}.2^3}{3!} = 0.1804$

3. UNIDAD 3: DISTRIBUCIONES DE PROBABILIDAD CONTINUAS MUESTREO

3.1. Distribución Normal

Llamada también como Distribución Laplace-Gauss, Gaussiana, Laplaciana, curva normal, curva de error, curva de campana o curva de Moivre. Aparentemente descubierta por Moivre (1756) como forma límite de la Distribución Binomial. Todo ejercicio de Binomial se puede resolver mediante la Distribución Normal, conocida como método aproximado. Existen dos razones básicas para que ocupe un lugar importantísimo en la estadística. Primero, tiene algunas propiedades que la hacen aplicable a un gran número de situaciones en las que es necesario hacer inferencias mediante la toma de muestras. Segundo, la distribución normal casi se ajusta a la distribución de frecuencias reales observadas en muchos fenómenos naturales. (Alvarez Roman, 2004)

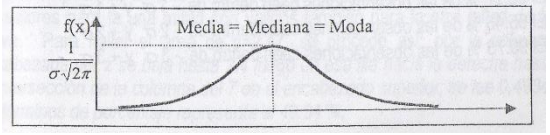
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad \text{o} \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(Z)^2}$$

Donde: e = 2.71828…; π = 3,14159265…; $\sigma = \sqrt{n.p.q}$; $\mu = n.p$

a) Características de la distribución normal:

Ilustración 14

El diagrama nos ayudara a determinar ciertas características importantes



Fuente: (Alvarez Roman, 2004)

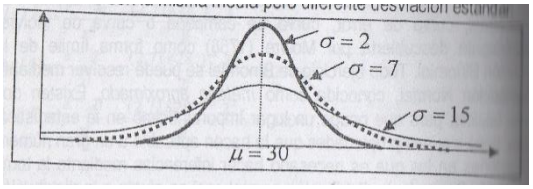
1) Según el valor de σ, habrá mayor o menor concentración de los datos respecto al valor central que es μ.

2) La función está definida en todo para el eje x.
3) Para todos los valores de x, la función toma valores positivos, es decir, la curva normal está situada sobre el eje x.
4) Es asíntota respecto al eje x, es decir, los dos extremos de la curva se extienden indefinidamente y nunca tocan el eje horizontal.
5) La curva tiene un solo pico; por tanto, es unimodal. Tiene la forma de campana.
6) La media, mediana y moda tienen el mismo valor, coinciden en el punto medio y dividen en dos partes iguales la curva.
7) Es simétrica respecto al eje y.
8) El área bajo la curva vale 1.

No existe una sola distribución normal, sino una familia de distribuciones normales. (Alvarez Roman, 2004)

Ilustración 15

Tres curvas con la misma media, pero diferente desviación estándar



Fuente: (Alvarez Roman, 2004)

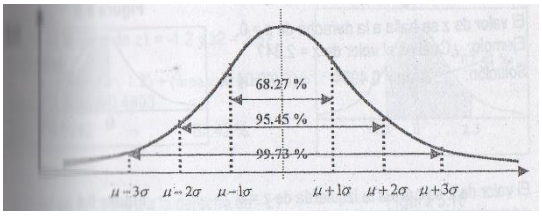
b) Áreas bajo la curva normal:

El área bajo la curva normal vale 1, sin importar los valores de $\mu\ y\ \sigma$. En términos matemáticos podemos pensar en áreas bajo las curvas como si fueran probabilidades. La relación más importante entre la desviación típica o estándar y la curva normal, lo observamos en el siguiente gráfico. En la distribución normal, la desviación estándar es usada como unidad para determinar el porcentaje de población. (Alvarez Roman, 2004)

1) El 68,27% de las observaciones caen dentro de $-1\sigma\ y\ +1\sigma$
2) El 95,45% de las observaciones caen dentro de $-2\sigma\ y\ +2\sigma$
3) El 99,73% de las observaciones caen dentro de $-3\sigma\ y\ +3\sigma$

Ilustración 16

Áreas bajo la curva normal



Fuente: (Alvarez Roman, 2004)

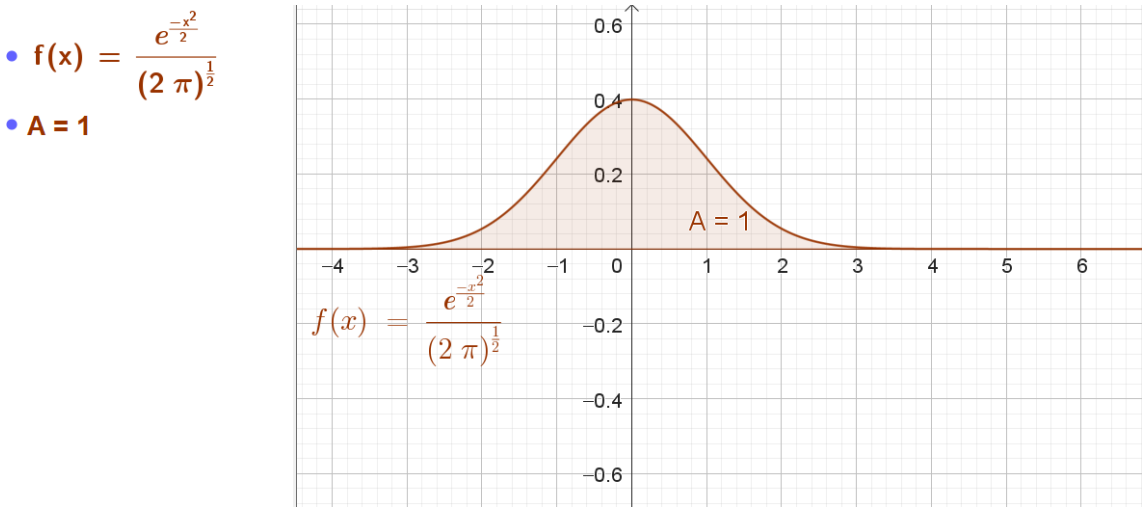
c) Medida estándar o valor tipificado z

El valor tipificado z sirve para convertir observaciones individuales en unidades disponibles en función de la desviación típica. La fórmula para medir las distancias bajo la curva normal es:

$$z = \frac{xi - \mu}{\sigma} \quad o \quad z = \frac{xi - x}{s}$$

¿Por qué utilizamos z en lugar del "numero de desviaciones estándar"? Las variables aleatorias normalmente distribuidas tienen muchas unidades diferentes de medición: dólares, pulgadas, kilogramos, segundos. Para determinar el área bajo la curva normal, trabajaremos en términos de puntuaciones estándares o típicas (z) y emplearemos. El uso de z permite solamente cambiar la escala de medición del eje horizontal. Algunos textos trabajan con la tabla de distribución acumulada, cuya lectura también es fácil.

¿Cómo utilizar la tabla de distribución normal? La tabla, da los valores de la mitad del área bajo la curva normal, empezando en 0.00 en la media. Como la curva normal es simétrica, los valores para la una mitad son válidos también para la otra mitad de la curva. Para hallar el área de z=2,47, se procede así: por la columna encabezada por z se baja hasta 2,4 luego en esa fila hacia la derecha hasta la intersección de la columna del 7 en el encabezado superior, se lee 0.4934, en términos de porcentaje representa el 49.34%. (Alvarez Roman, 2004)



d) Casos que se presentan en el cálculo de áreas

1) El valor de z se halla a la derecha de z = 0.

Ejemplo 1: ¿Cuál es el valor de z = 2,34?

Área = 0,4904 = 49,04%

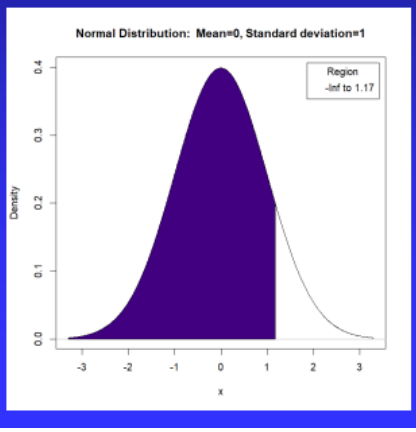
Probabilidad Normal Estándar En Rcmdr

Probabilidad de que Z obtenga los siguientes valores:

- **P(Z ≤ 1.17)**

pnorm(c(1.17), mean=0, sd=1, lower.tail=TRUE)

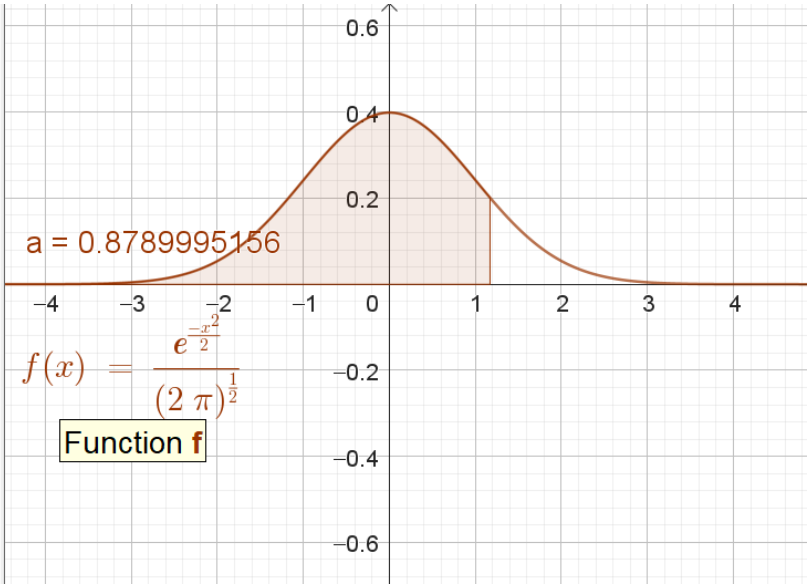
 # 0.8789995

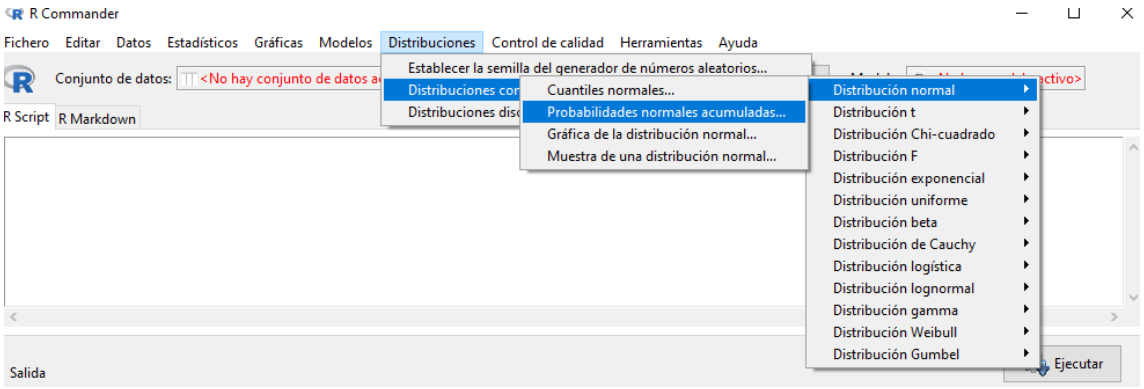


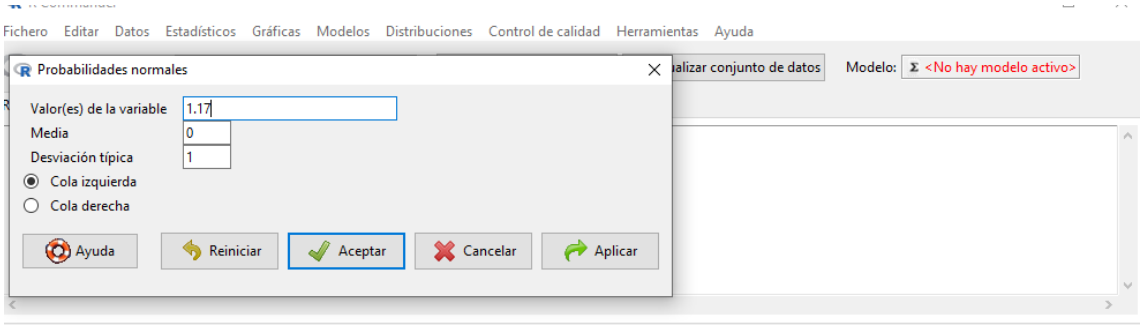
GEOGEBRA

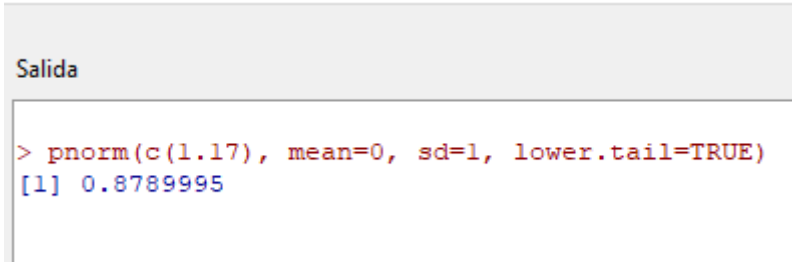
Cola izquierda

- $f(x) = \frac{e^{\frac{-x^2}{2}}}{(2\,\pi)^{\frac{1}{2}}}$
- **a = 0.8789995156**

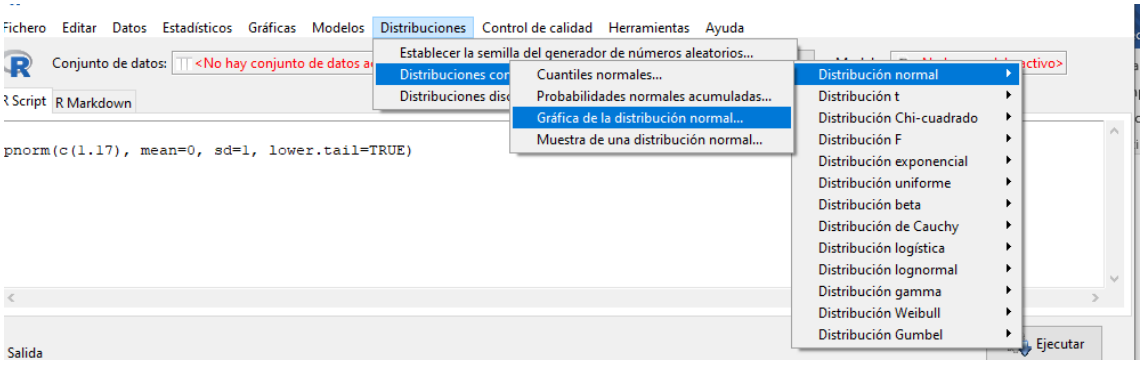


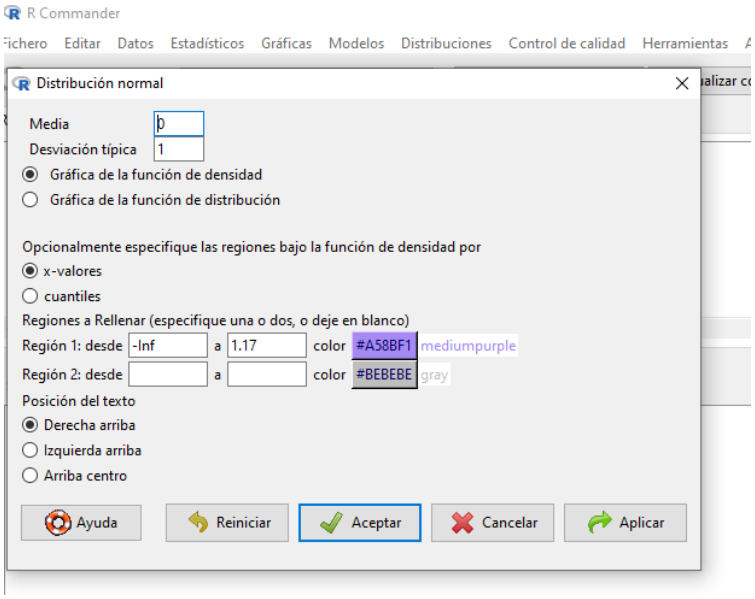

R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Establecer la semilla del generador de números aleatorios...
Cuantiles normales...
Probabilidades normales acumuladas...
Gráfica de la distribución normal...
Muestra de una distribución normal...
Distribución normal
Distribución t
Distribución Chi-cuadrado
Distribución F
Distribución exponencial
Distribución uniforme
Distribución beta
Distribución de Cauchy
Distribución logística
Distribución lognormal
Distribución gamma
Distribución Weibull
Distribución Gumbel
Conjunto de datos:
R Script
R Markdown
Salida
Ejecutar

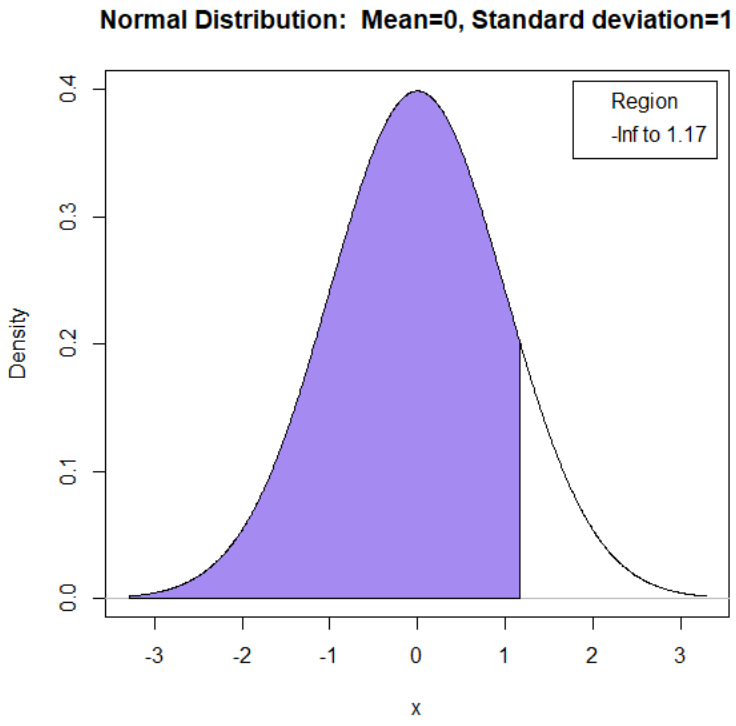

Probabilidades normales
Valor(es) de la variable
1.17
Media
0
Desviación típica
1
Cola izquierda
Cola derecha
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar
Modelo:
<No hay modelo activo>


Salida
> pnorm(c(1.17), mean=0, sd=1, lower.tail=TRUE)
[1] 0.8789995

Gráfica


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ayuda
Establecer la semilla del generador de números aleatorios...
Cuantiles normales...
Probabilidades normales acumuladas...
Gráfica de la distribución normal...
Muestra de una distribución normal...
Distribución normal
Distribución t
Distribución Chi-cuadrado
Distribución F
Distribución exponencial
Distribución uniforme
Distribución beta
Distribución de Cauchy
Distribución logística
Distribución lognormal
Distribución gamma
Distribución Weibull
Distribución Gumbel
Conjunto de datos:
R Script
R Markdown
pnorm(c(1.17), mean=0, sd=1, lower.tail=TRUE)
Salida
Ejecutar


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas
Distribución normal
Media
0
Desviación típica
1
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde -Inf a 1.17 color #A58BF1 mediumpurple
Región 2: desde a color #BEBEBE gray
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar

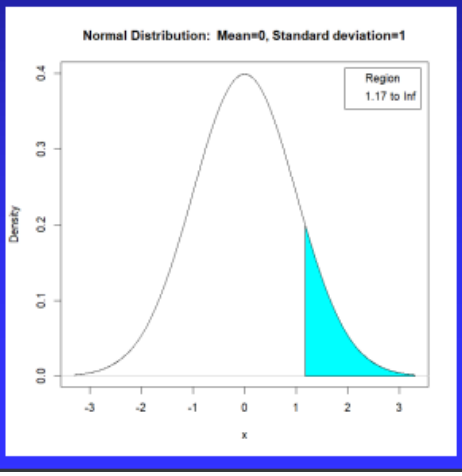

Normal Distribution: Mean=0, Standard deviation=1
Region
-Inf to 1.17
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

Probabilidad Normal Estándar En Rcmdr

Probabilidad de que Z obtenga los siguientes valores:

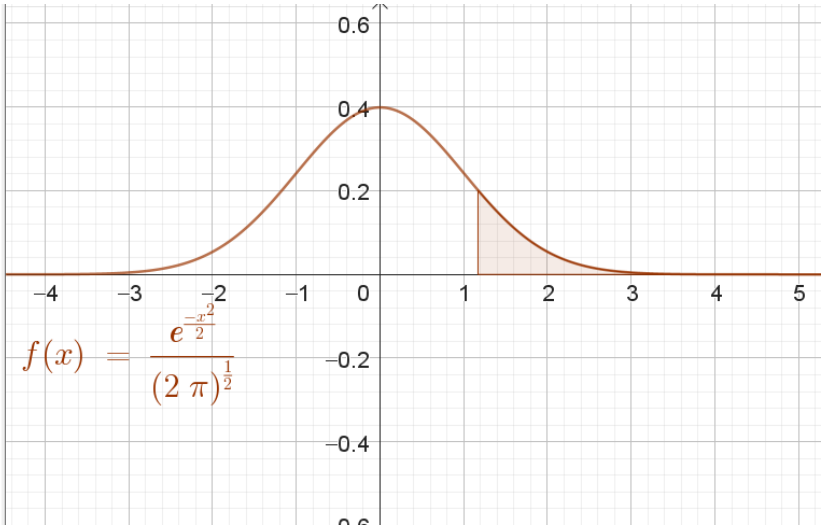
- **P(Z ≥ 1.17)**

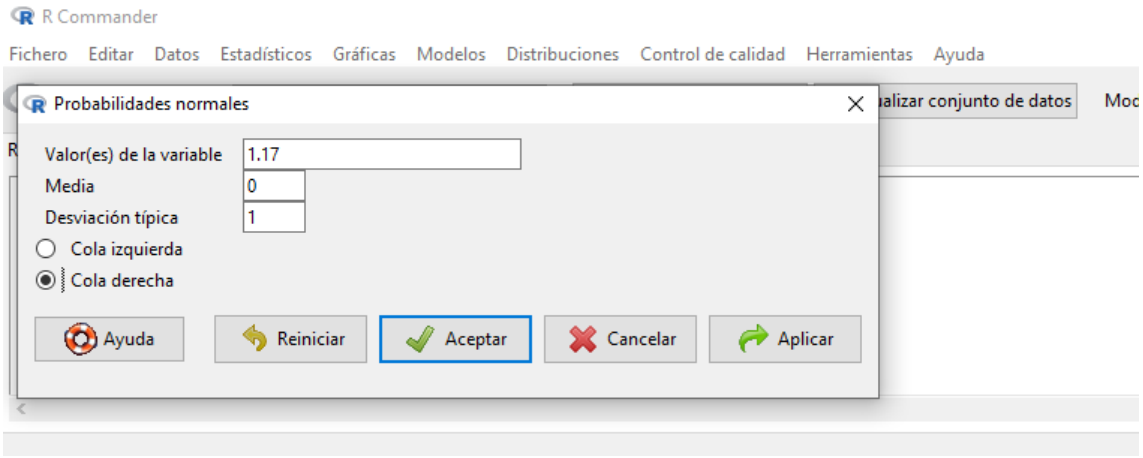
```
pnorm(c(1.17), mean=0, sd=1, lower.tail=FALSE)
# 0.1210005
```



GEOGEBRA

- $f(x) = \frac{e^{\frac{-x^2}{2}}}{(2\,\pi)^{\frac{1}{2}}}$
- **a = 0.1210004844**

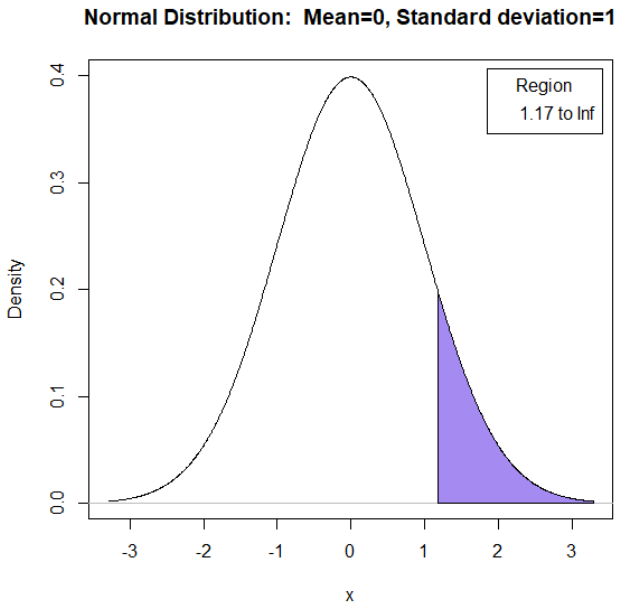




Salida

```
> pnorm(c(1.17), mean=0, sd=1, lower.tail=FALSE)
[1] 0.1210005
```

GRÄFICA

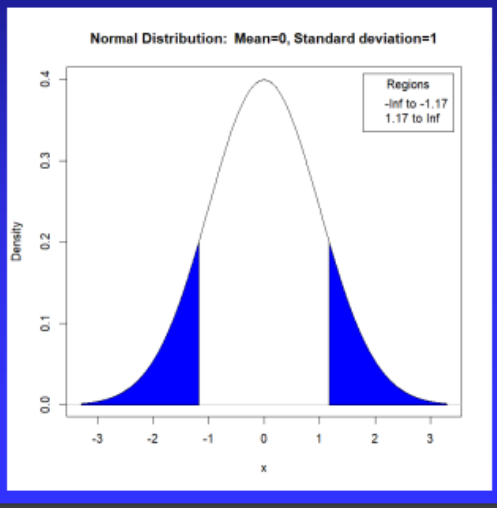


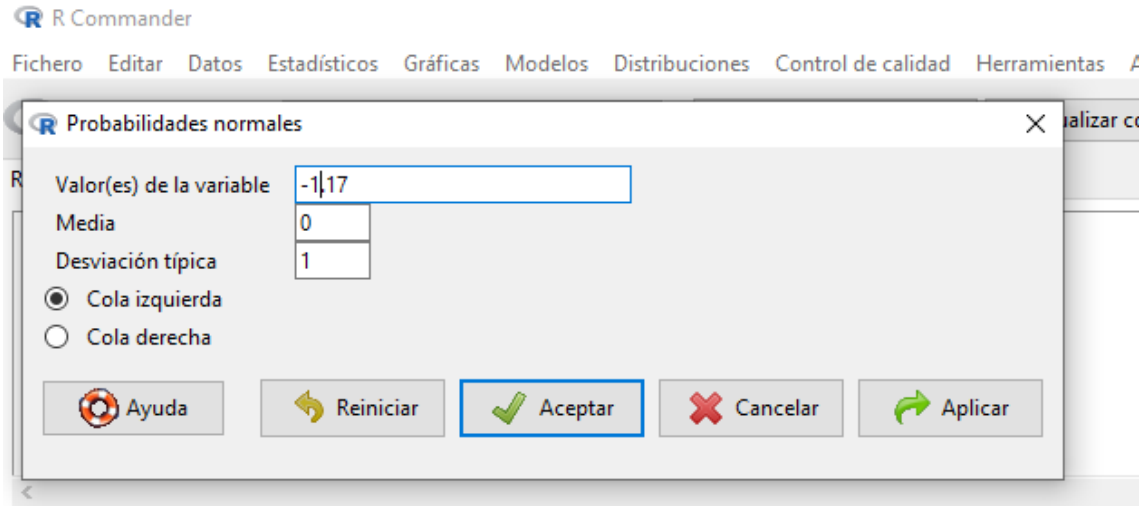
Probabilidad Normal Estándar En Rcmdr

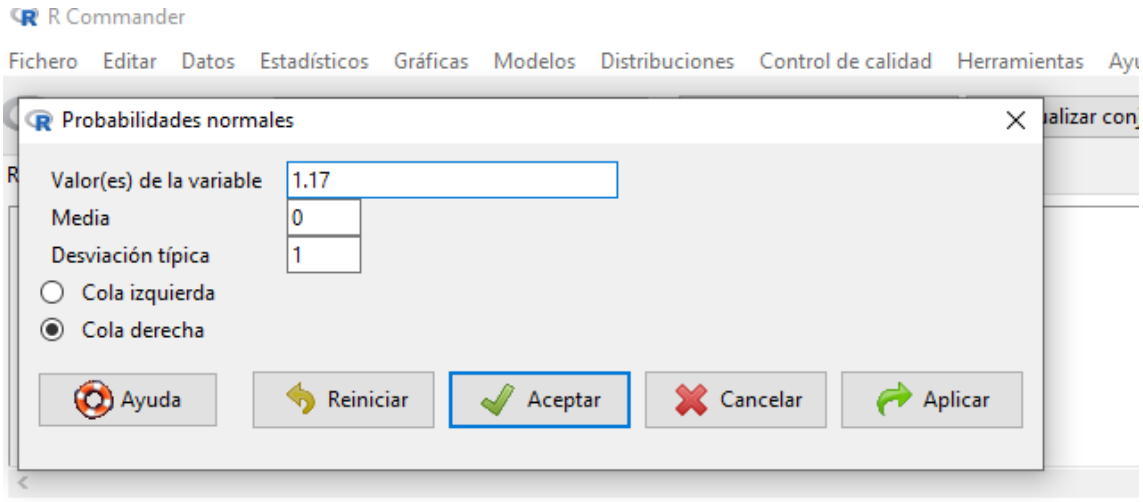
h) **P(|Z| ≥ 1.17)**

- Determinar el área de -∞ a -1.17 y de 1.17 a +∞. Como la curva es simétrica, simplemente multiplicamos el valor de P(Z ≥ 1.17) del ejemplo anterior por 2:

P(|Z| ≥ 1.17) = 2 x P(Z ≥ 1.17) = 2 x 0.121 = 0.242







Salida

```
> pnorm(c(-1.17), mean=0, sd=1, lower.tail=TRUE)
[1] 0.1210005

> pnorm(c(1.17), mean=0, sd=1, lower.tail=FALSE)
[1] 0.1210005
```

GRAFICA

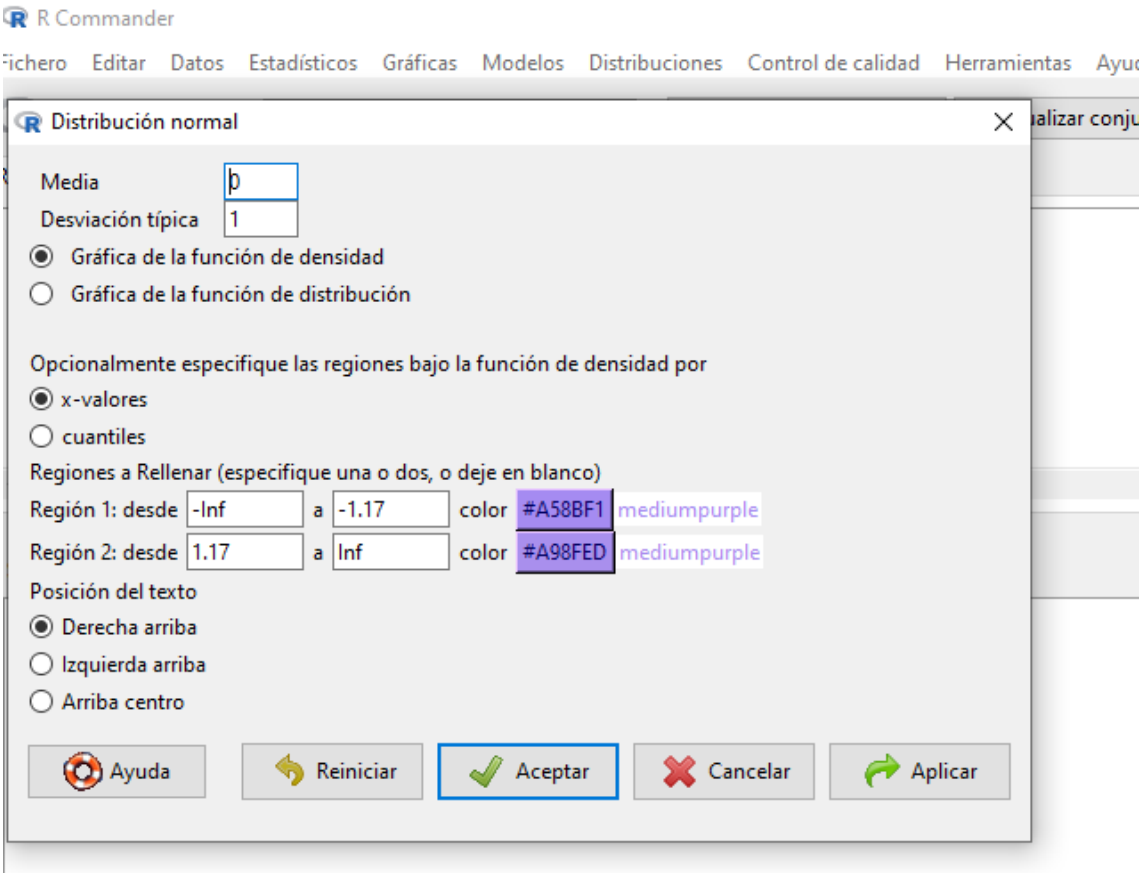
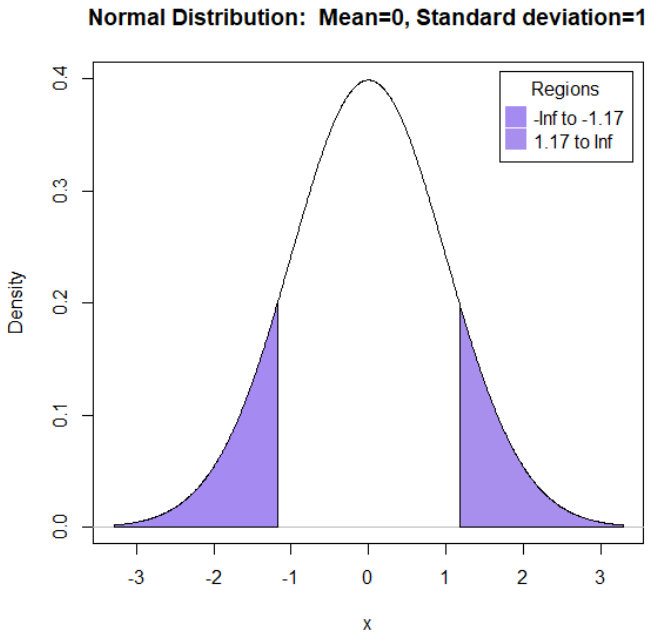
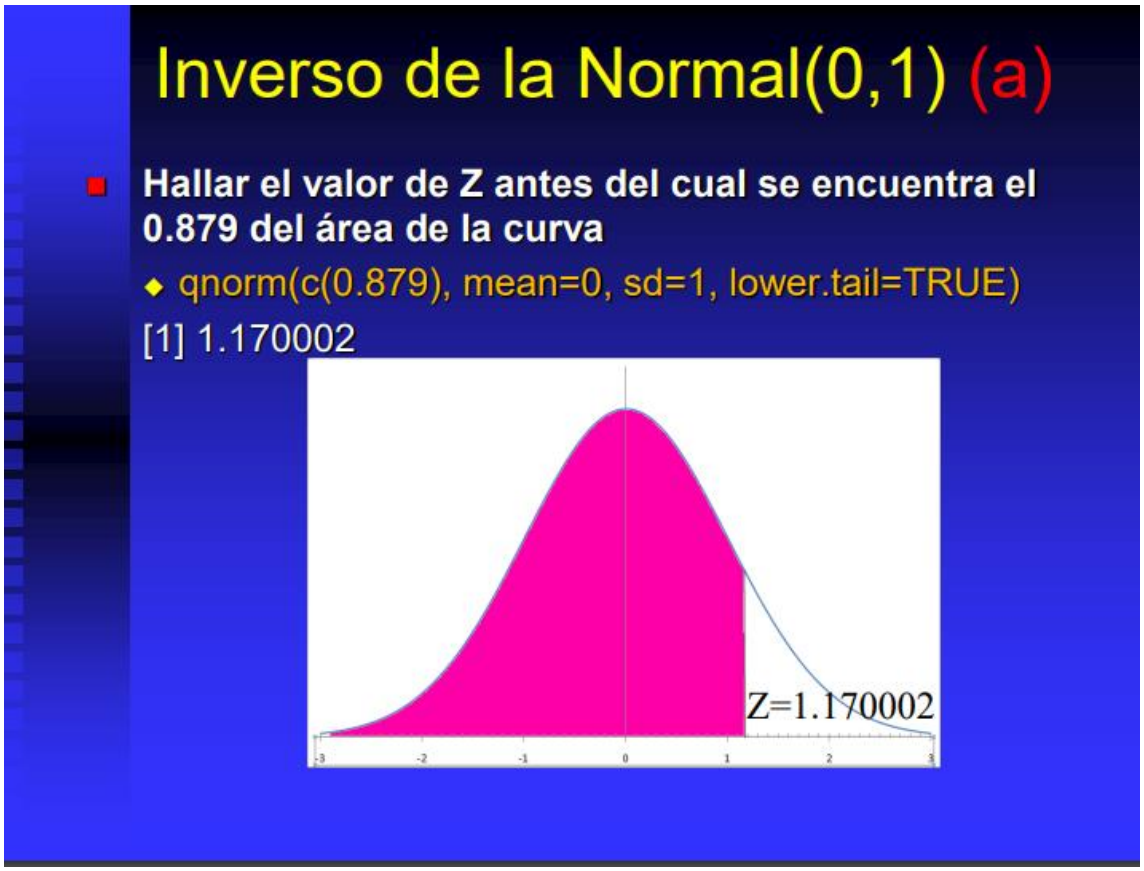
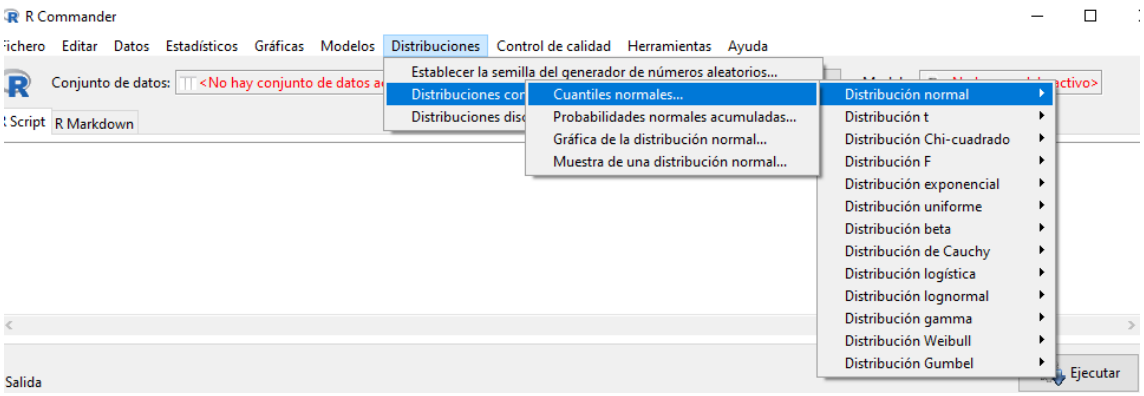


Ilustración 17



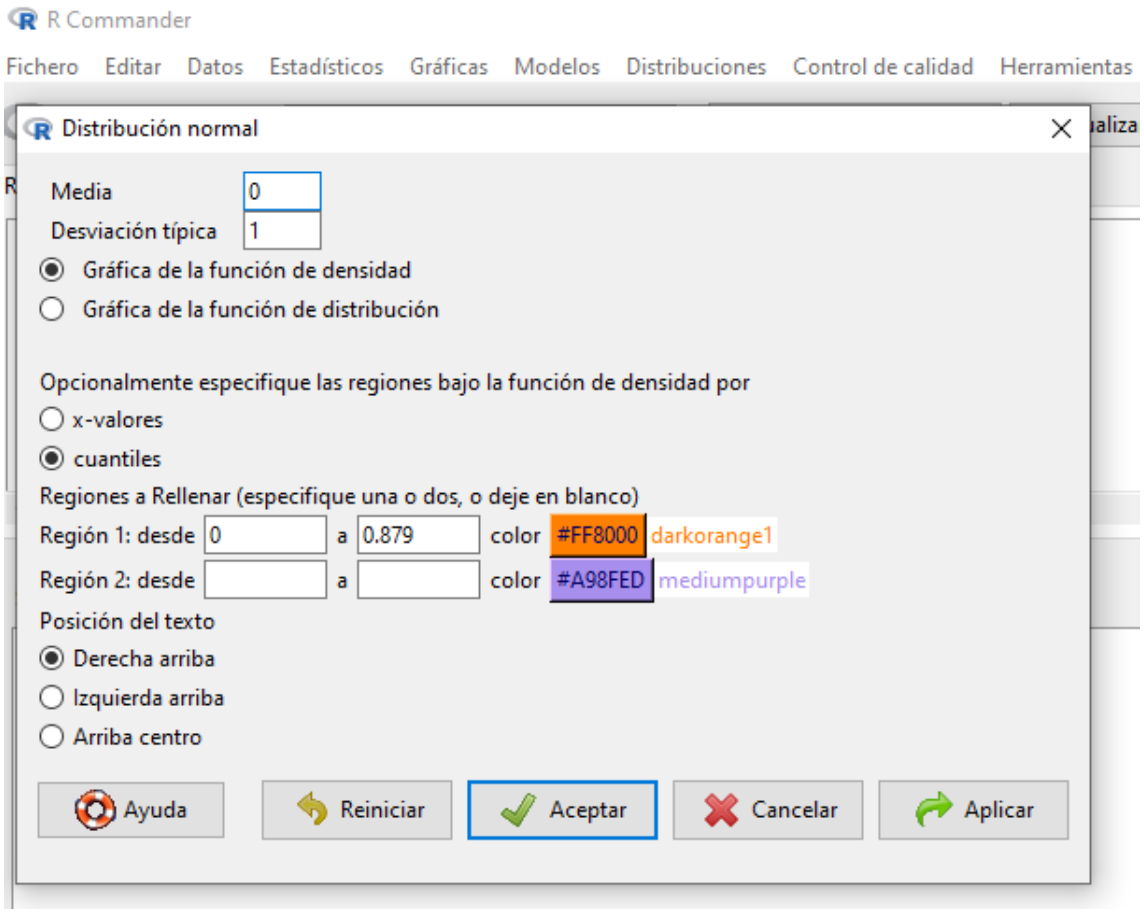


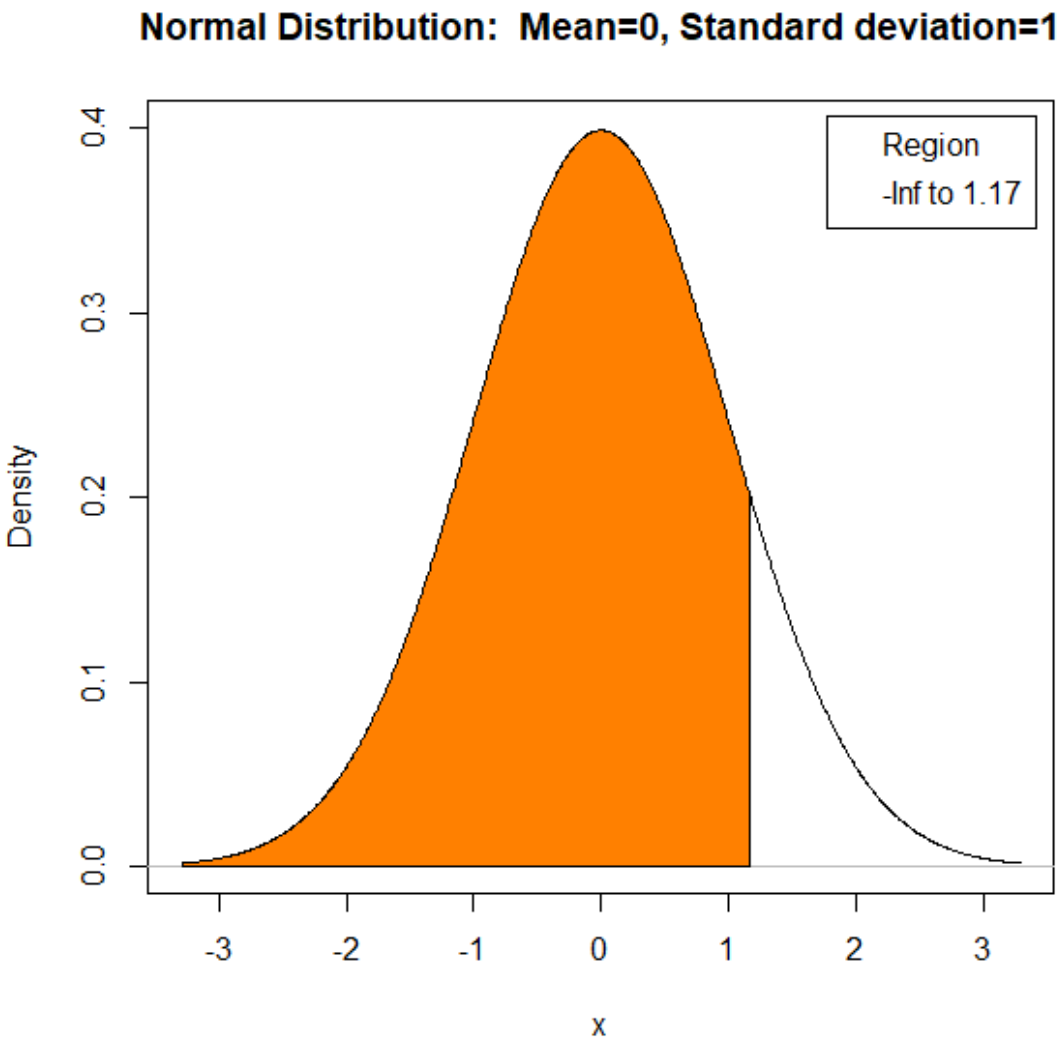


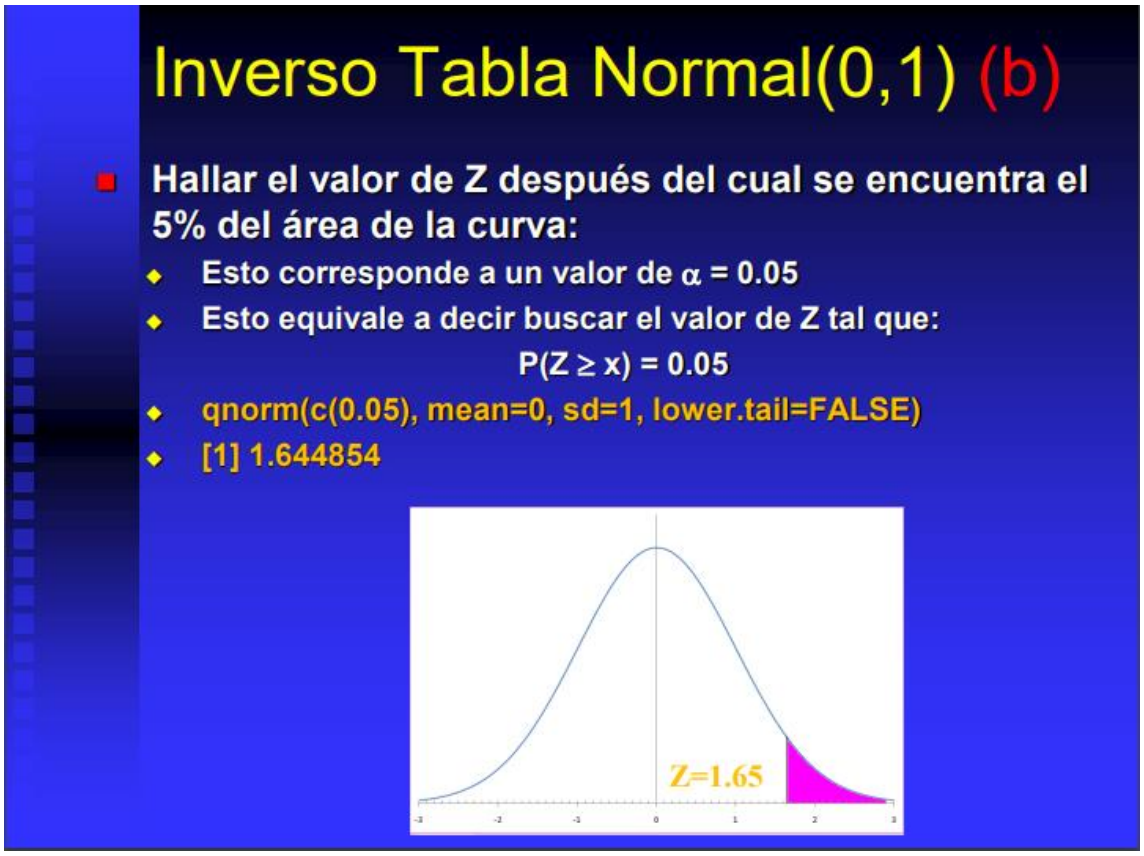
Salida

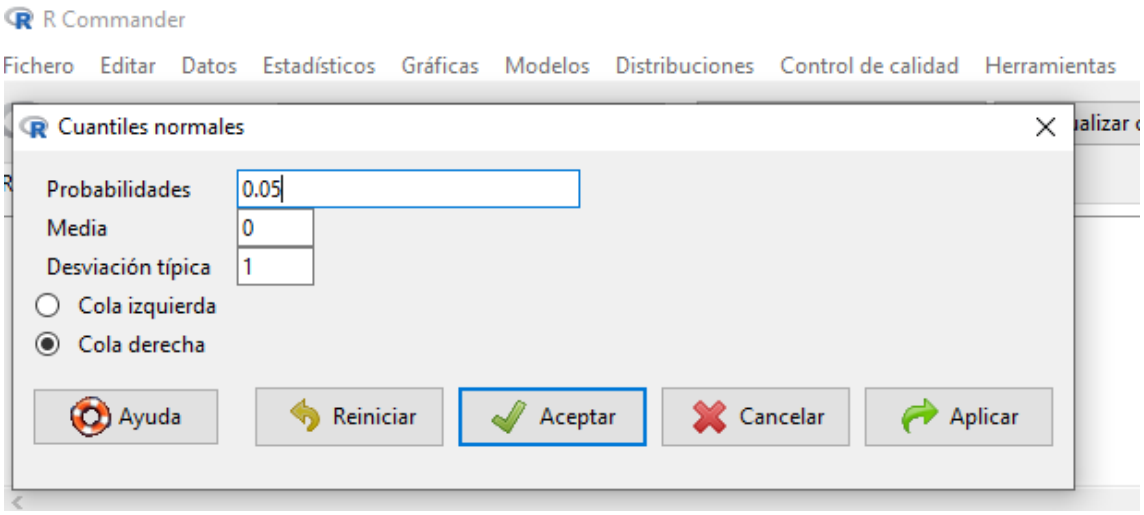
```
> qnorm(c(0.879), mean=0, sd=1, lower.tail=TRUE)
[1] 1.170002
```

GRÁFICA




Normal Distribution: Mean=0, Standard deviation=1
Region
-Inf to 1.17
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

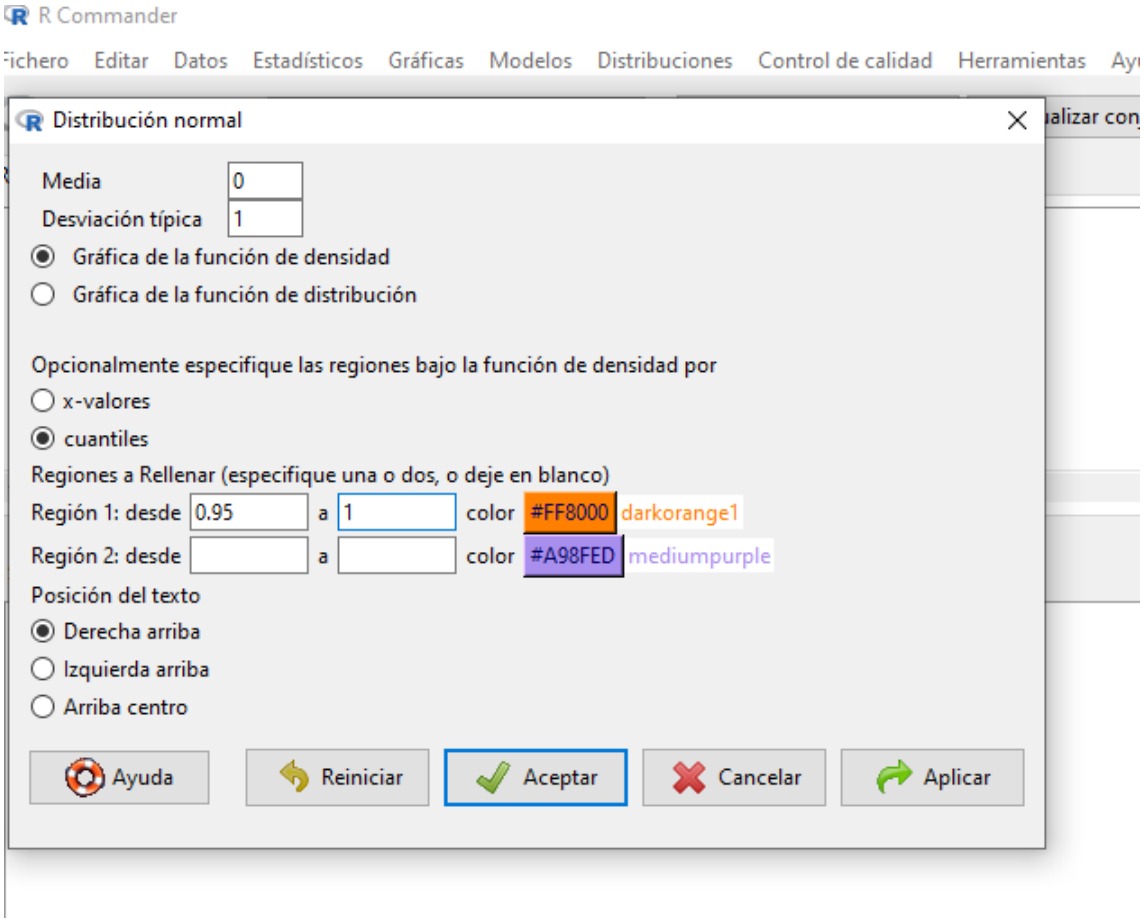


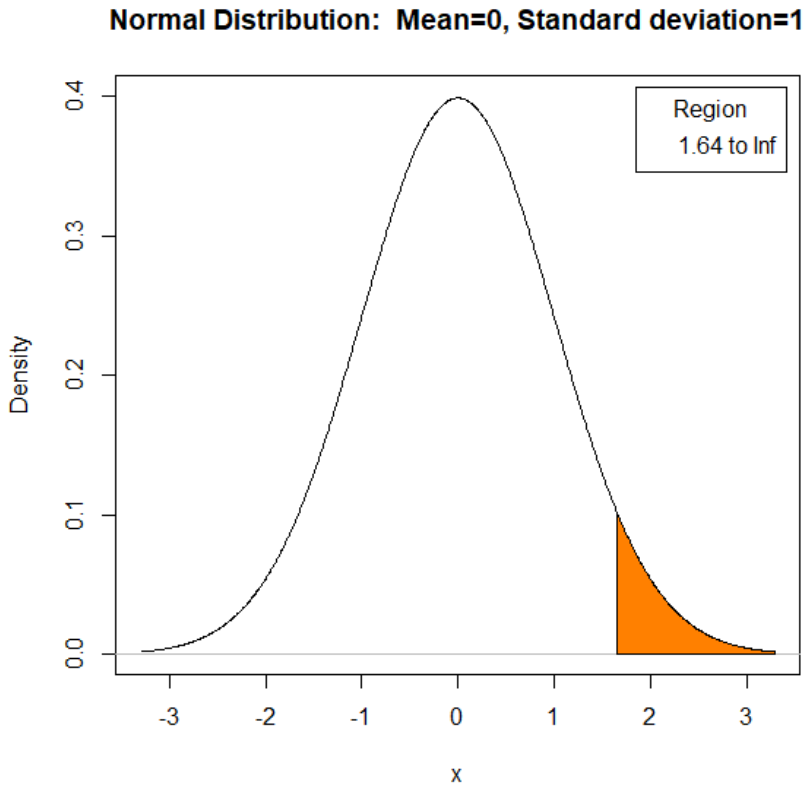


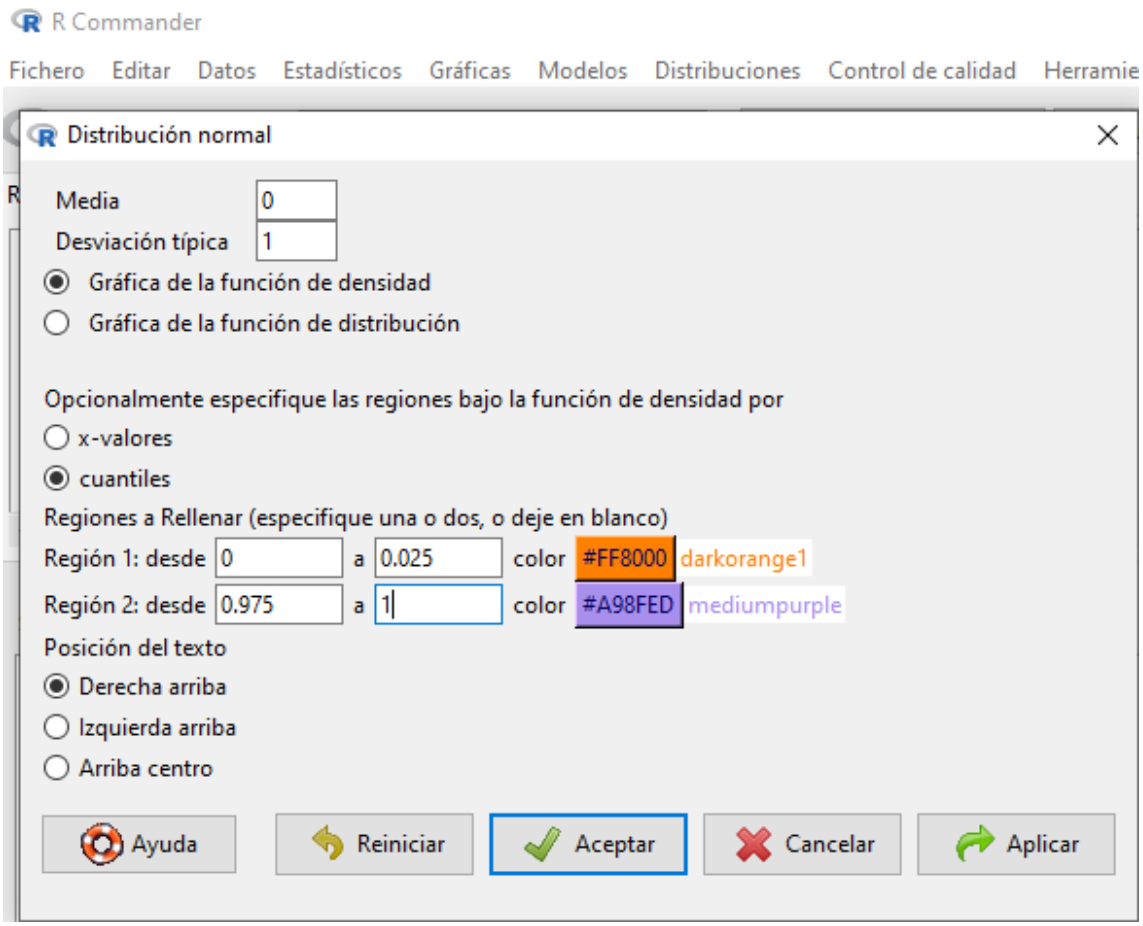
Salida

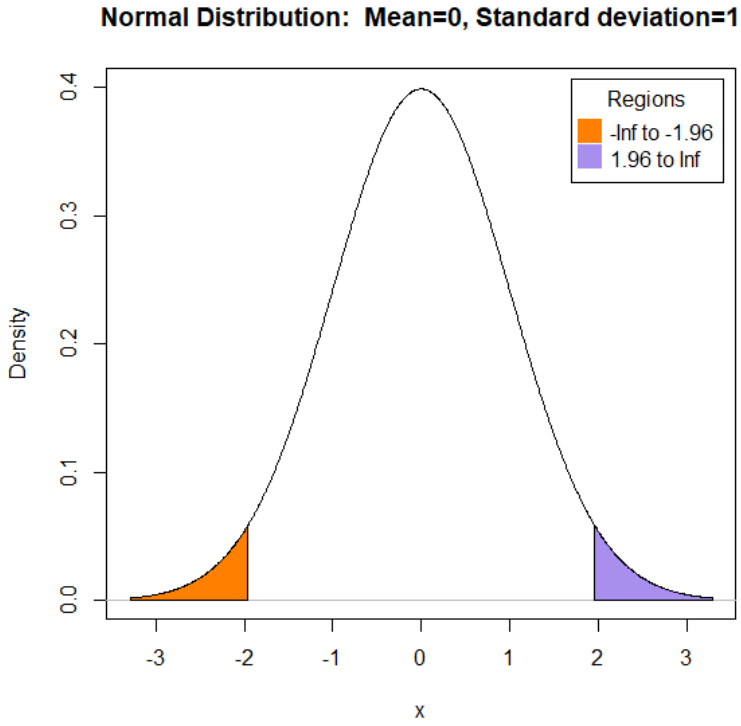
```
> qnorm(c(0.05), mean=0, sd=1, lower.tail=FALSE)
[1] 1.644854
```

GRÁFICA


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas Ay
Distribución normal
Media 0
Desviación típica 1
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0.95 a 1 color #FF8000 darkorange1
Región 2: desde a color #A98FED mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar


Normal Distribution: Mean=0, Standard deviation=1
Region
1.64 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramie
Distribución normal
Media 0
Desviación típica 1
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0 a 0.025 color #FF8000 darkorange1
Región 2: desde 0.975 a 1 color #A98FED mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar


Normal Distribution: Mean=0, Standard deviation=1
Regions
-Inf to -1.96
1.96 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

Inverso Tabla Normal(0,1)

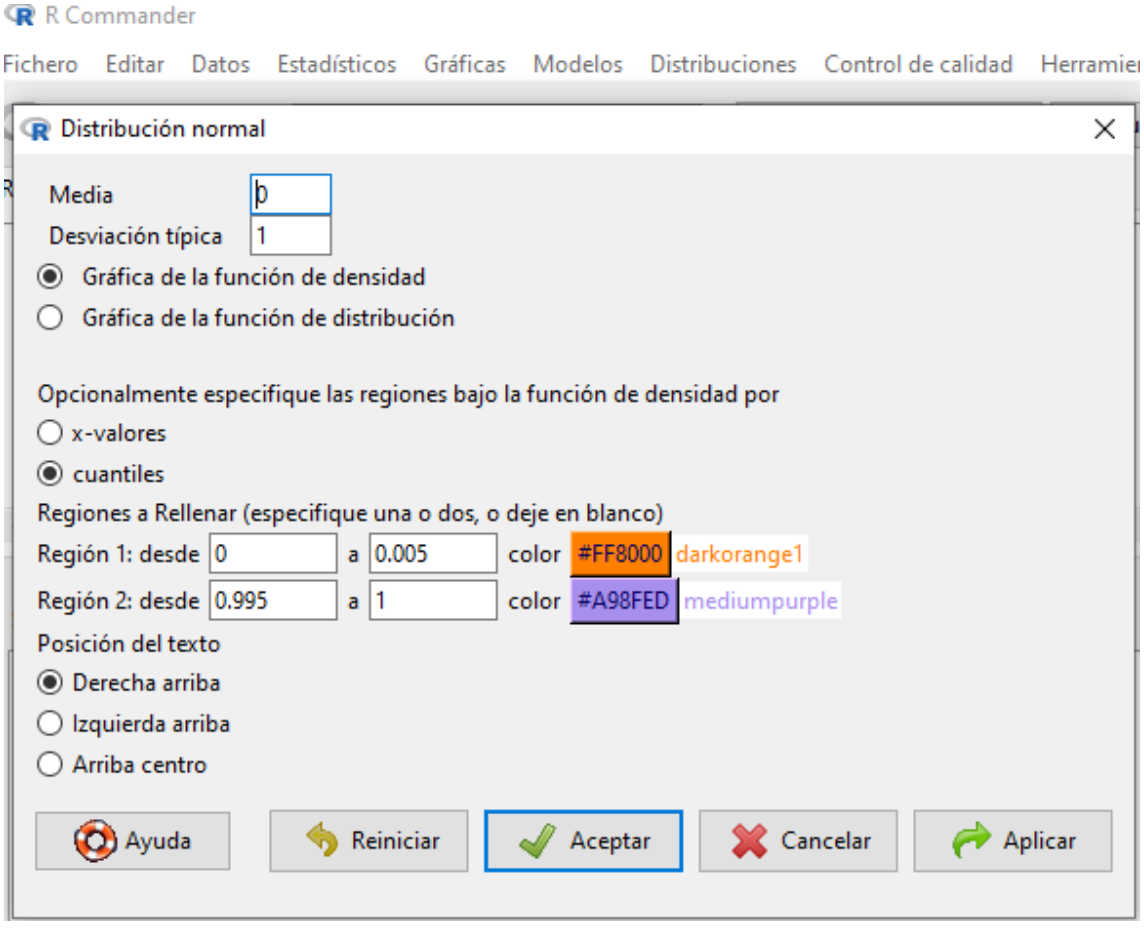
- **(d) Hallar el valor de Z después del cual se encuentra el 1% del área de la curva:**
  - Esto corresponde a un valor de $\alpha = 0.01$

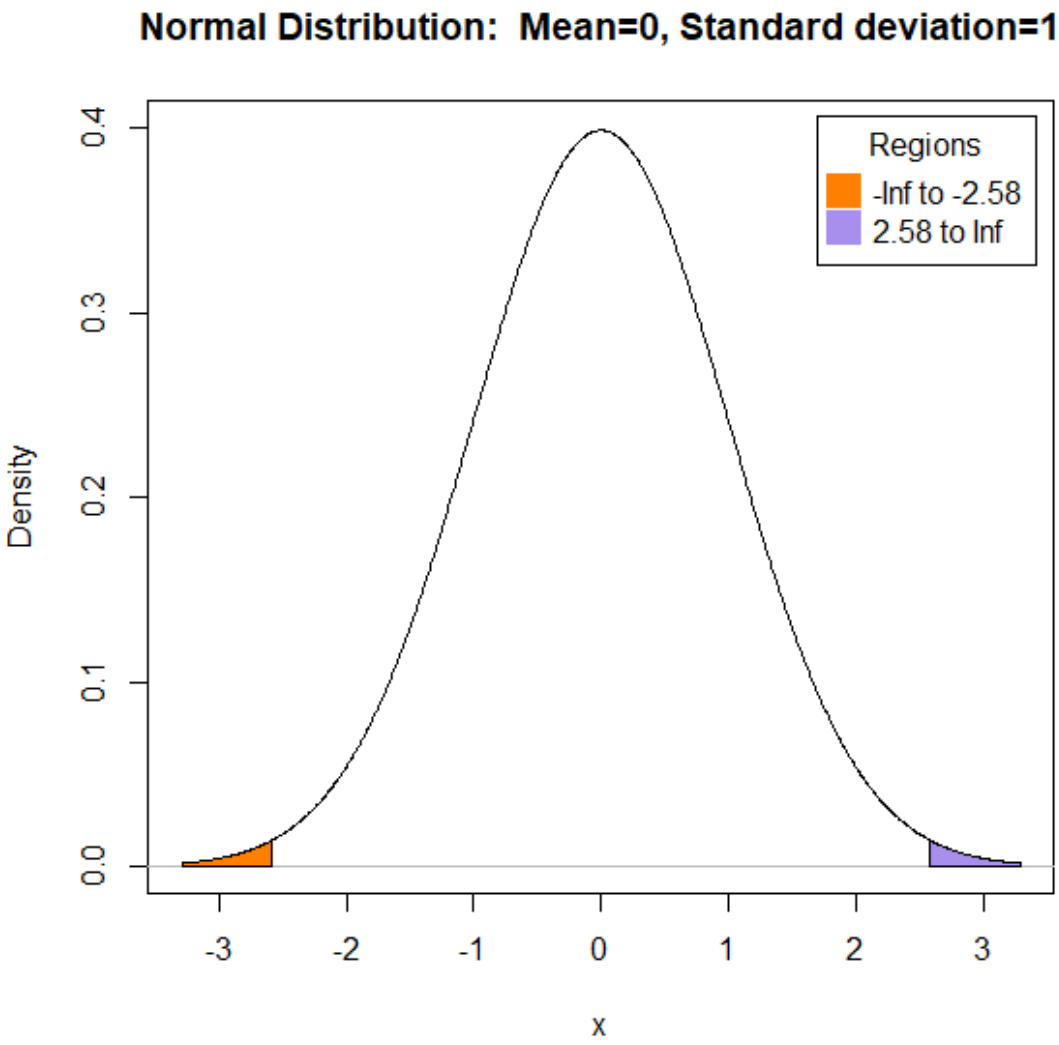
    $Z_{(0.01)} = 2.326$

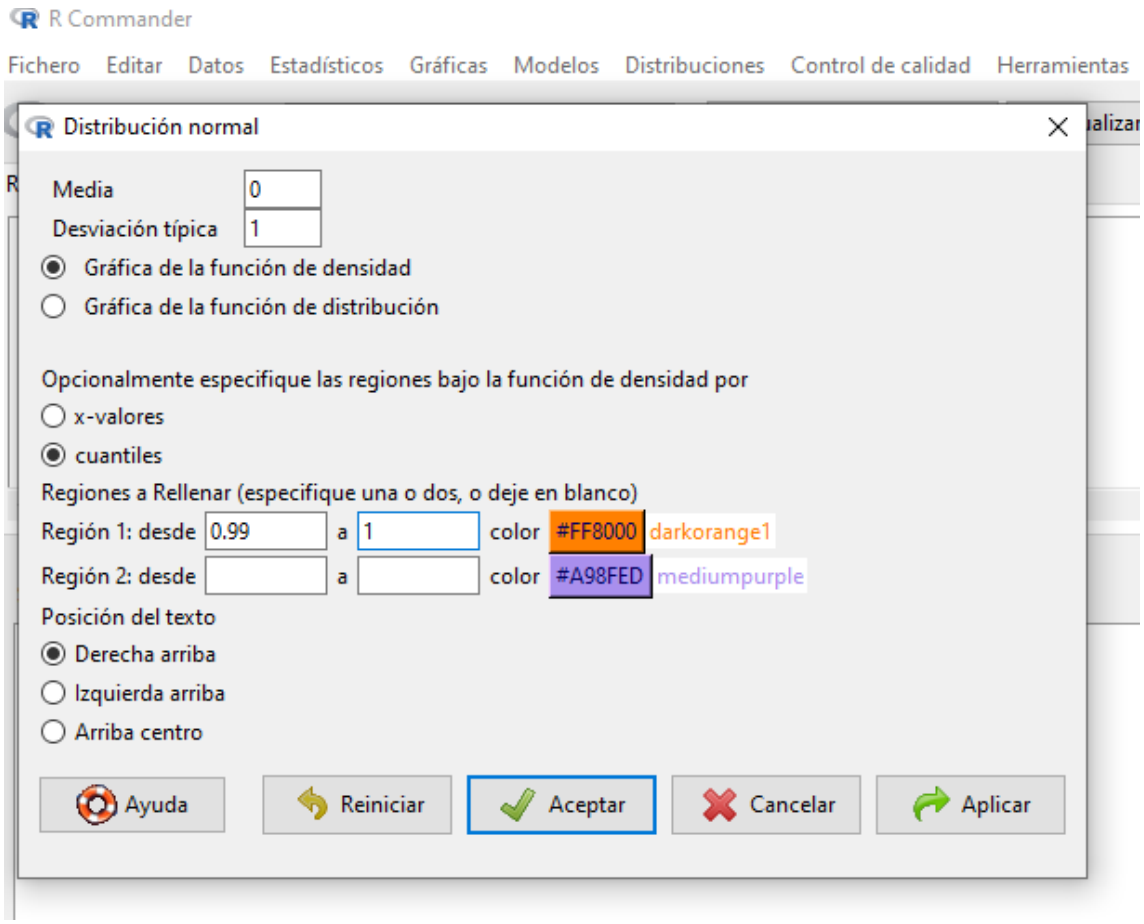
- **(e) Hallar el valor de Z tal que el área fuera del intervalo de –Z a Z es igual a 0.01:**
  - Como es una curva simétrica: $\alpha/2 = 0.01/2=0.005$

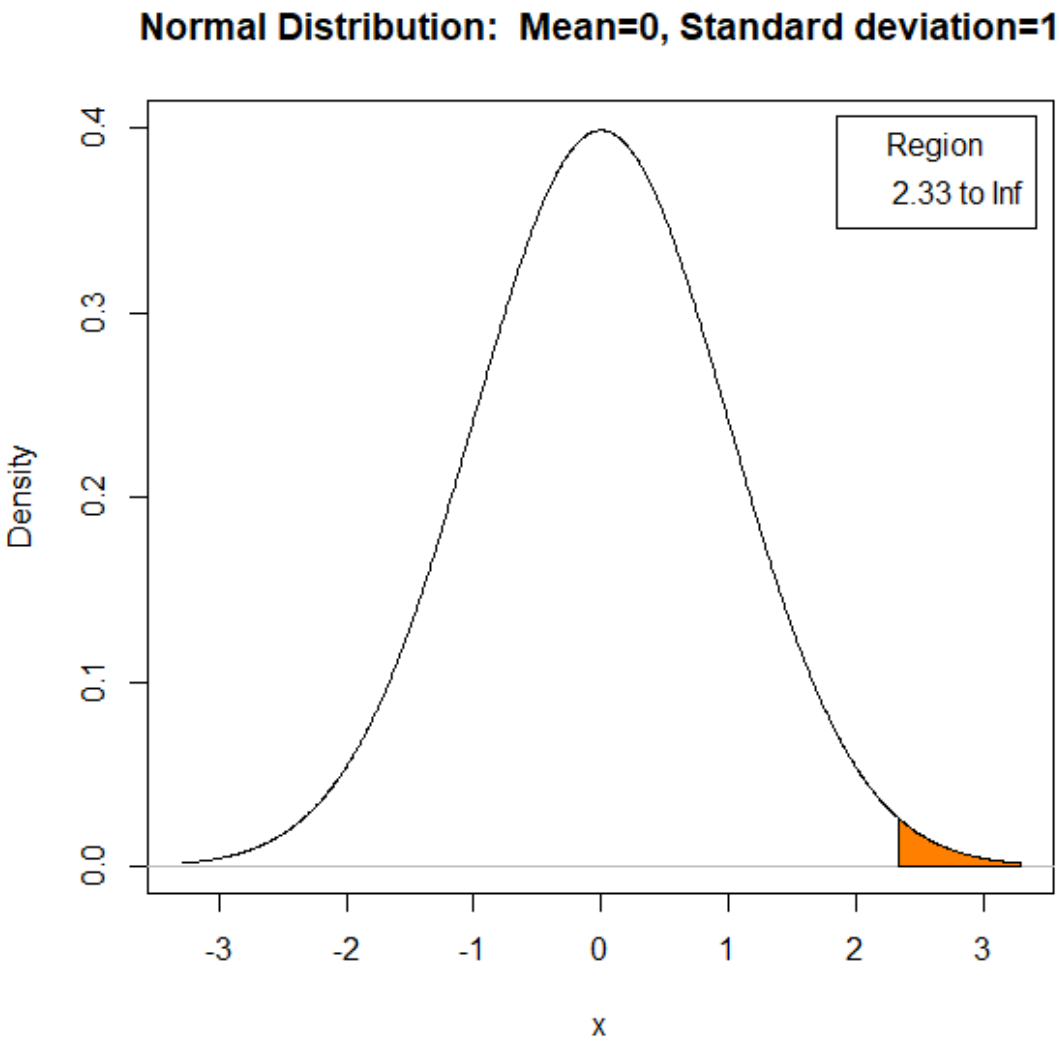
    $Z_{(0.005)} = 2.576$

Calcular en Rcmdr

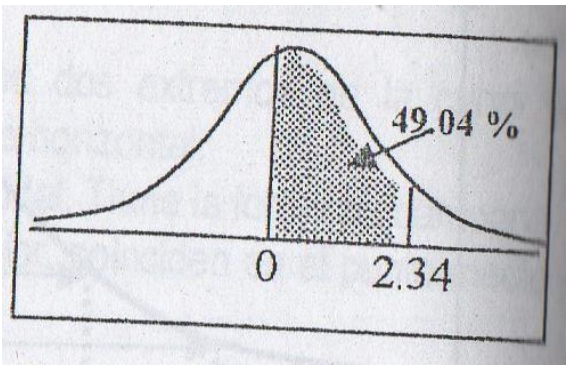



Normal Distribution: Mean=0, Standard deviation=1
Regions
-Inf to -2.58
2.58 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x


R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas
Distribución normal
Media 0
Desviación típica 1
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0.99 a 1 color #FF8000 darkorange1
Región 2: desde a color #A98FED mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar



Grafica del Ejemplo 1



Fuente: (Alvarez Roman, 2004)

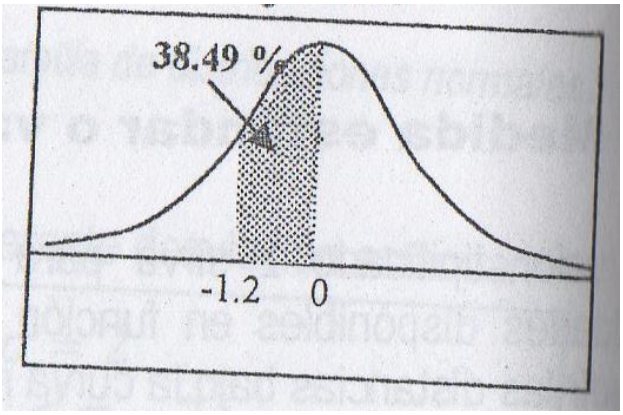
2) El valor de z se halla a la izquierda de z = 0

Ejemplo 2: ¿Cuál es el valor de z = -1,2?

Área = 0,3849 = 38,49%

Ilustración 18

Grafica del Ejemplo 2



Fuente: (Alvarez Roman, 2004)

3) Z1 y Z2 se encuentra a la derecha de z = 0

Ejemplo 3: ¿Cuál es el área entre z1 = 1,2 y z2 = 2,7?

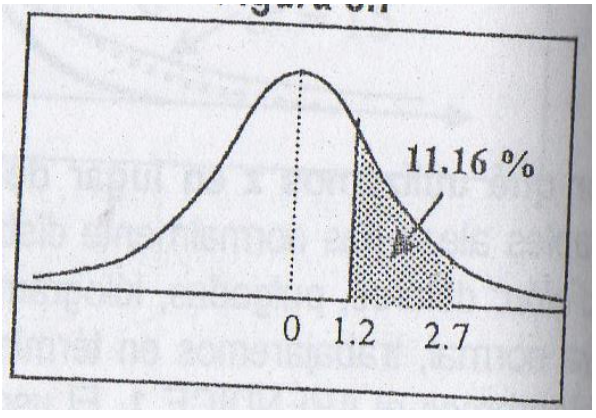
A = (área z=0 y z2=2,7) – (área z=0 y z1=1,2)

Área = 0,4965 – 0,3849

Área = 0,1116 = 11,16%

Ilustración 19

Grafica del Ejemplo 3



Fuente: (Alvarez Roman, 2004)

4) Z1 y Z2 se encuentra a la izquierda de z = 0

Ejemplo 4: ¿Cuál es el valor de z1 = -0,6 y z2 = -1,9?

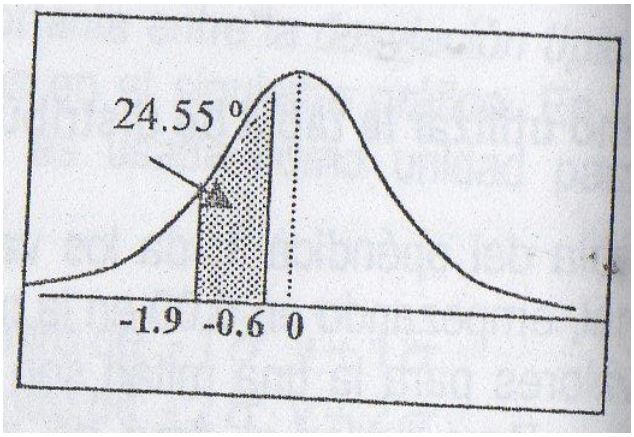
A = (área z=0 y z2=-1,9) – (área z=0 y z1=-0,6)

Área = 0,4713 – 0,2258

Área = 0,2455 = 24,55%

Ilustración 20

Grafica del Ejemplo 4



Fuente: (Alvarez Roman, 2004)

5) Z1 se encuentra a la izquierda y z2 se encuentra a la derecha de z=0

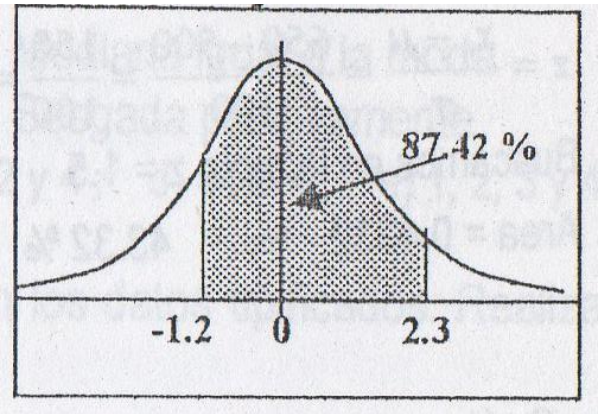
Ejemplo 5: ¿Cuál es el valor de z1 = -1,2 y z2 = 2,3?

A = (área z=0 y z=-1,2) + (área z=0 y z=2,3)

Área = 0,8742 = 87,42%

Ilustración 21

Grafica del Ejemplo 5



Fuente: (Alvarez Roman, 2004)

3.2. Distribución "t" de student

Cuando se muestrea una distribución normal con desviación estándar conocida σ, la distribución de $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ es N(0,1). Desde un punto de vista práctico, la necesidad de conocer σ impide formular inferencias con respecto a μ debido a que generalmente no se conoce el valor de la desviación estándar de la población. Dada la disponibilidad de una muestra aleatoria, el camino lógico que se sigue en este caso es reemplazar σ con una estimación s, que es el valor de la desviación estándar muestral S. Desafortunadamente, cuando lo anterior se lleva a cabo, la distribución de $(x - \mu)/(S/\sqrt{n})$ no es $N(0,1)$, aun cuando la muestra provenga de una distribución normal. Sin embargo, es posible determinar la distribución de muestreo exacta de $(x - \mu)/(S/\sqrt{n})$ cuando se muestrea N(μ,σ), con μ y σ^2 desconocidos. Para finalizar esta sección se examinarán los aspectos teóricos de lo que se conoce como la distribución t de Student.

Supóngase que se realiza un experimento en que se observan dos variables aleatorias X y Z; X tiene una distribución normal con media cero y desviación estándar uno. Sea T otra variable aleatoria que es función de X y Z, de manera tal que

$$T = \frac{Z}{\sqrt{X/v}}$$

Es decir, T se define como el coeficiente entre una variable aleatoria normal estándar y la raíz cuadrada de una variable aleatoria chi-cuadrada dividida por sus grados de libertad. El conjunto de todos los posibles valores de la variable aleatoria T es el intervalo $(-\infty, -\infty)$ puesto que los valores de Z se encuentran en este y los valores de X son positivos. El valor

$$t = \frac{z}{\sqrt{x/v}}$$

recibe el nombre de valor de la variable aleatoria de t de student. Lo anterior lleva al siguiente teorema.

Sea Z una variable aleatoria normal estándar y X una variable aleatoria chi cuadrada con v grados de libertad. Si Z y X son independientes, entonces la variable aleatoria

$$T = \frac{Z}{\sqrt{X/v}}$$

Tiene una distribución t de Student con v grados de libertad y una función de densidad de probabilidad dada por

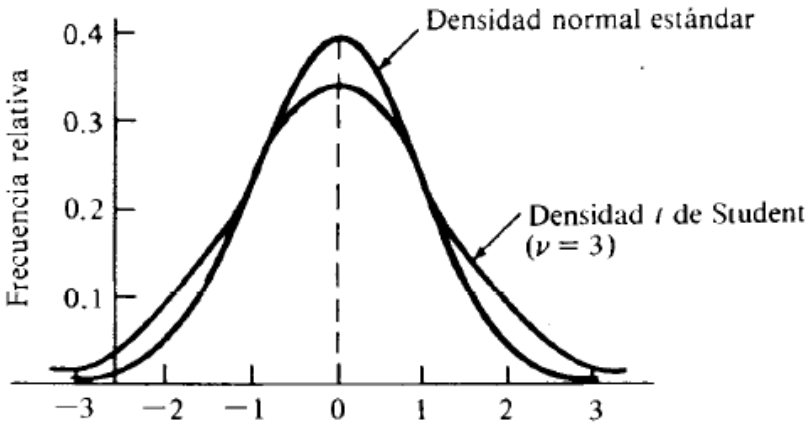
$$f(t;v) = \frac{T((v+1)/2)}{\sqrt{\pi v}\,T(v/2)}\,(1 + (t^2/v))^{-(v+1)/2}, \quad -\infty < 1 < \infty, \quad v > 0.$$

En la formula anterior se observa que el parámetro de distribución t es v, que, al igual que para la distribución chi cuadrada, recibe el nombre de grados de libertad. Para cualquier

v > 0, la distribución t es simétrica con respecto al origen y la función de densidad tiene su valor máximo cuando t = 0. De la ilustración 22 es evidente que la forma de la función de densidad t de Student es muy similar a la de la densidad normal estándar y con los extremos de la distribución t menos pronunciados que los de la distribución normal. De hecho, conforme se tiene un número mayor de grados de libertad, la distribución t de Student tiende hacia la normal estándar. (Canavos, 1988)

Ilustración 22

Comparación entre las densidades normal estándar y t de Student



Fuente: (Canavos, 1988)

Puede demostrarse que el valor esperado de T es

$$E(T) = 0 \qquad v > 1,$$

Y la varianza está dada por

$$Var(T) = v/(v-2) \qquad v > 2.$$

Los valores cuantiles t1-$(-\propto, v)$ tales que:

$$P(T \leq t1, \propto, v) = \int f(t; v)dt = 1 - \propto, \quad 0 \leq \propto \leq 1.$$

Para los distintos valores de v y de las proporciones acumulativas seleccionadas $1 - \propto$. Por ejemplo, si v = 15.

$$P(T \leq t0.90{,}15) = P(T \leq 1{,}341) = 0{,}90$$

$$P(T \leq t0.95{,}15) = P(T \leq 1{,}753) = 0{,}95$$

$$P(T \leq t0.95{,}15) = P(T \leq 1{,}753) = 0{,}95$$

Dado que la distribución t es simétrica con respecto al cero, para $\propto > 0{,}5$ los valores cuantiles serán negativos, pero sus magnitudes serán las mismas que las de los correspondientes valores que se encuentran en el lado derecho. De esta forma, para v = 15, (Canavos, 1988)

$$P(T \leq t0.10,15) = P(T \leq -1,341) = 0,10$$

$$P(T \leq t0.05,15) = P(T \leq -1,753) = 0,05$$

$$P(T \leq t0.01,15) = P(T \leq -2,602) = 0,01$$

A fin de ilustrar la similitud que existe entre la distribución t de Student y la normal estándar para valores relativamente grandes de v, en la tabla 28 se encuentra una comparación entre los valores cuantiles t y los correspondientes valores normales estándar para valores crecientes de v. Para $\propto = 0,1\ o\ 0,05$, la concordancia se encuentra en aproximadamente 0,05 unidades, aun para valores tan bajos de v como 30. De hecho, muchos autores sugieren que, desde un punto de vista práctico, es muy poca la ganancia que se tiene al emplear la distribución t de Student en lugar de la norma estándar cuando $v \geq 30$.

Recuérdese que para formular inferencias con respecto a μ cuando el muestreo se lleva a cabo sobre una distribución normal con media y varianza desconocidas, se necesita determinar la distribución de (X - μ)(S/$\sqrt{n}$). Cuando se muestra una distribución $N(\mu, \sigma)$, la distribucion de $(X - \mu)/(\sigma/\sqrt{n})$ es N(0,1). Para la misma condición, se sabe que, de $(\frac{(n-1)S^2}{\sigma^2} + (\frac{X-\mu}{\sigma/\sqrt{n}})^2 = \frac{\sum_{i=1}^{n}(Xi-\mu)^2}{\sigma^2}$ y del teorema (Si X1 y X2 son dos variables aleatorias independientes y cada una tiene distribución chi-cuadrada con v1 y v2 grados de libertad respectivamente), la distribución de $(n - 1)S^2/\sigma^2$ es chi-cuadrada con n – 1 grados de libertad. Dado que puede demostrarse que X y S^2 son independientes, del (Sea Z una variable aleatoria normal estándar y X una variable aleatoria chi-cuadrada con v grados de libertad. Si Z y X son independientes, entonces la variable aleatoria), se desprende que la distribución de (Canavos, 1988)

$$\frac{\frac{X - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n - 1)S^2/\sigma^2}{(n - 1)}}} = \frac{X - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{\sqrt{S^2}},$$

$$T = \frac{X - \mu}{S/\sqrt{n}},$$

Es la t Student con n – 1 grados de libertad.

Tabla 27

Comparación entre los valores cuantiles de las distribuciones t de Student y normal estándar

α	$t_{1-\alpha, 20}$	$t_{1-\alpha, 30}$	$t_{1-\alpha, 40}$	$t_{1-\alpha, 50}$	$z_{1-\alpha}$
0.10	1.325	1.310	1.303	1.299	1.282
0.05	1.725	1.697	1.684	1.676	1.645
0.01	2.528	2.457	2.423	2.403	2.326

Fuente: (Canavos, 1988)

Ejemplo 1: El Departamento de Protección a Medio Ambiente asegura que, para un automóvil compacto en particular, el consumo de gasolina en carretera es de un galón por cada 45 millas. Una organización independiente de consumidores adquiere uno de estos automóviles y lo somete a prueba con el propósito de verificar la cifra proporcionada por el DPMA. El automóvil recorrió una distancia de 100 millas en 25 ocasiones. En cada recorrido se anotó el número de galones necesarios para realizar el viaje. Los 25 ensayos, el valor promedio y la desviación estándar, tuvieron un valor de 43,5 y 2,5 millas por galón, respectivamente. Si se supone que el número de millas que se recorre por galón es una variable aleatoria distribuida normalmente, con base en esta prueba ¿existe alguna razón para dudar de la veracidad del dato proporcionado por el DONA?

Este problema ilustra algunas de las dificultades prácticas que pueden encontrarse al ponerse en práctica la noción de muestra aleatoria. En forma ideal, se debieron seleccionar 25 carros de la misma marca, modelo y configuración de motor, de manera aleatoria, del mismo proceso de armado, de manera que fuese posible considerar el consumo de combustible como una variable aleatoria. Sin embargo, en este y otros, lo anterior representa un costo prohibitivo. A pesar de lo anterior, debe determinarse la veracidad de la información proporcionada por el DPMA con base en la probabilidad. Esto es, si μ fuese realmente igual a 45 millas por galón, ¿Cuál es la probabilidad de que se observe un valor X no mayor de 43,5 millas por galón, con base en una muestra de tamaño 25 y una estimación de σ igual a 2,5?

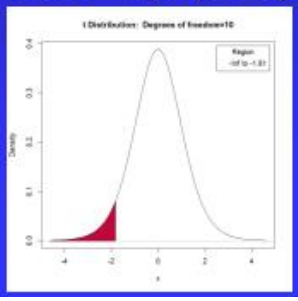
$$t = \frac{x - \mu}{s/\sqrt{n}} = \frac{43{,}5 - 45}{2{,}5/\sqrt{25}} = -3$$

es un valor de la distribución t de student con 24 grados de libertad. $P(T \leq -3) < 0{,}005$. Es decir, si el valor verdadero de la media es 45, la probabilidad de observar un valor de T no mayor de -3 unidades, es menor de 0,005. En cualquier caso, se ha observado algo que tiene una posibilidad de ocurrir menos de 5 en 1000, o μ tiene un valor real menor de 45. Para esta situación es preferible elegir la segunda explicación. (Canavos, 1988)

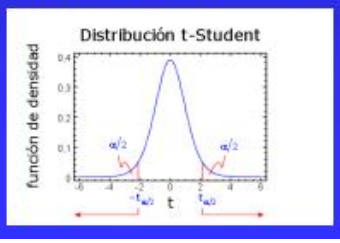
Probabilidad “t” en Rcmdr

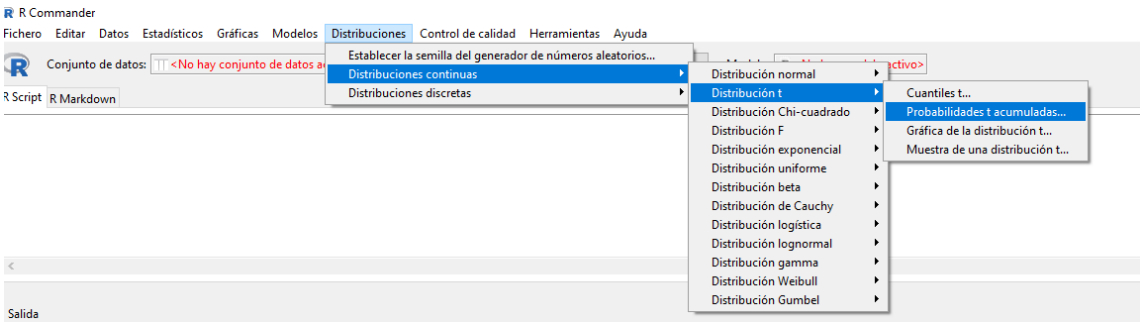
- pt(c(-1.8), df=10, lower.tail=TRUE)
- [1] 0.05102612
  - Devuelve el área a la izquierda de x (α)
  - x= valor de t
  - ν= grados de libertad
  - Colas = 1 o 2 colas

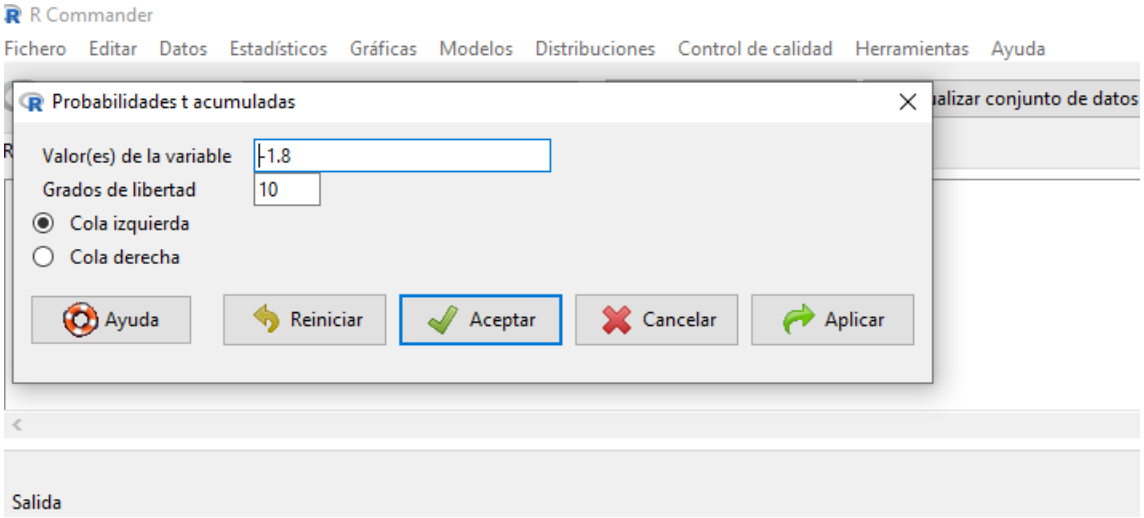
colas= 1, P(X<t)



colas = 2, P(|X| > t); P(X > t o X < -t).

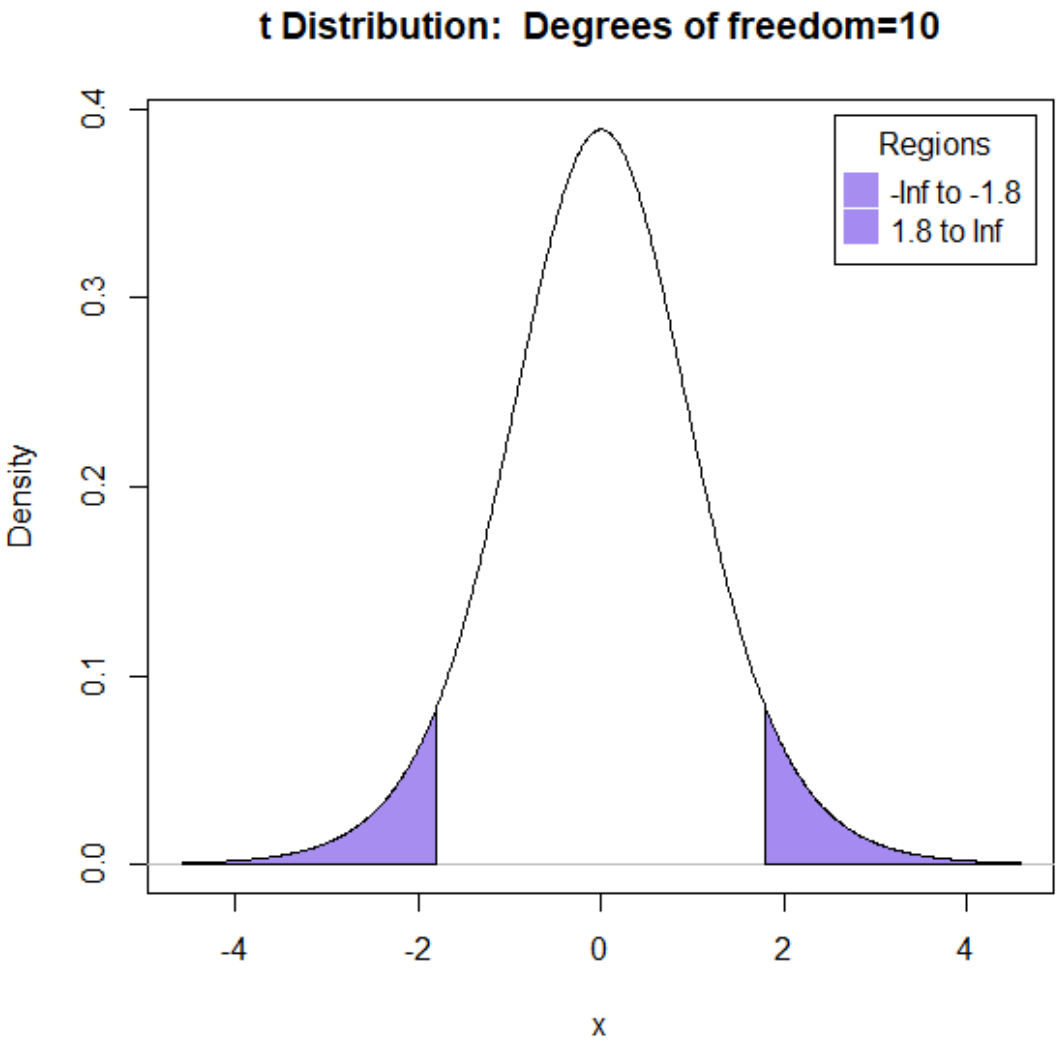


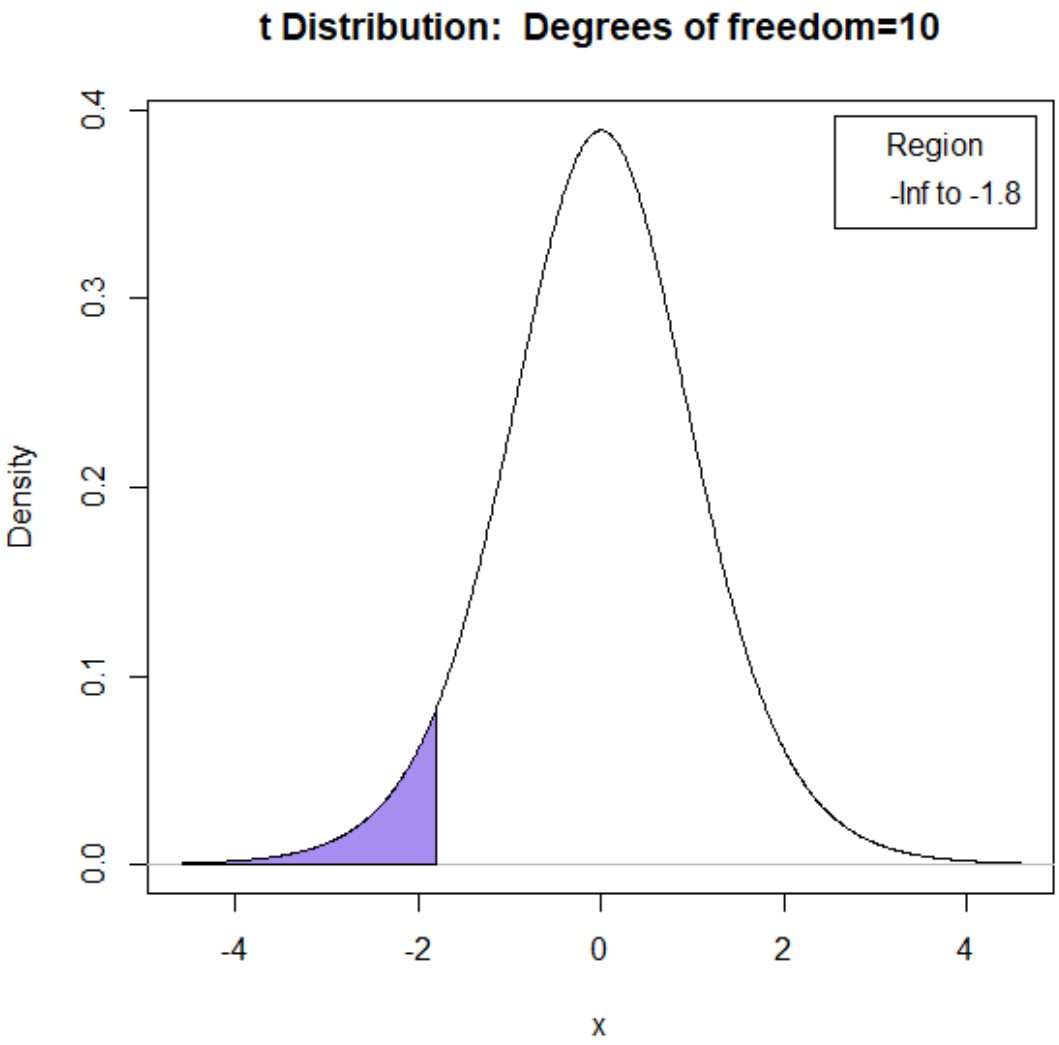




Salida

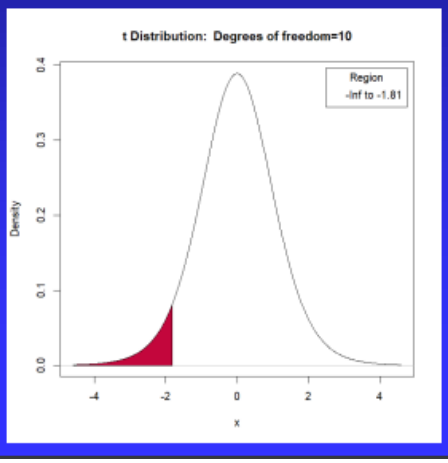
```
> pt(c(-1.8), df=10, lower.tail=TRUE)
[1] 0.05102612
```

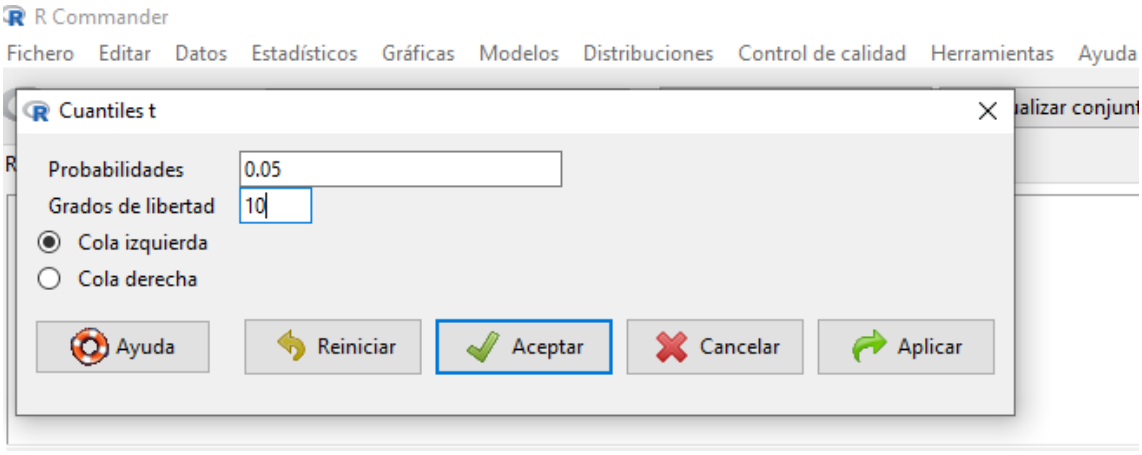



t Distribution: Degrees of freedom=10
Region
-Inf to -1.8
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Probabilidad “t” Inversa en Rcmdr

- qt(c(0.05), df=10, lower.tail=TRUE)
- [1] -1.812461
- Devuelve el valor de t de una cola a la izquierda, antes del cual se encuentra el α x 100% del área de la curva.
  - P(X < -t)





Salida

```
> qt(c(0.05), df=10, lower.tail=TRUE)
[1] -1.812461
```

Ejercicio en Rcmdr
Realizar un gráfico en cada uno

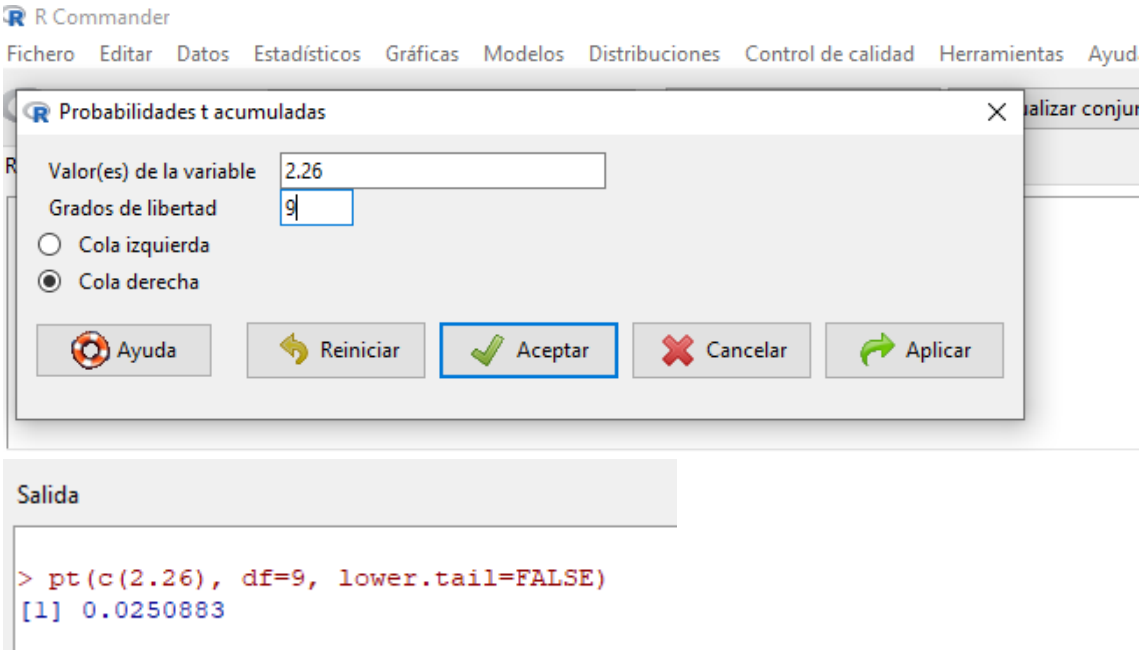
a) Calcular la probabilidad de obtener un valor mayor que 2.26 en una distribución t con 9 gdl

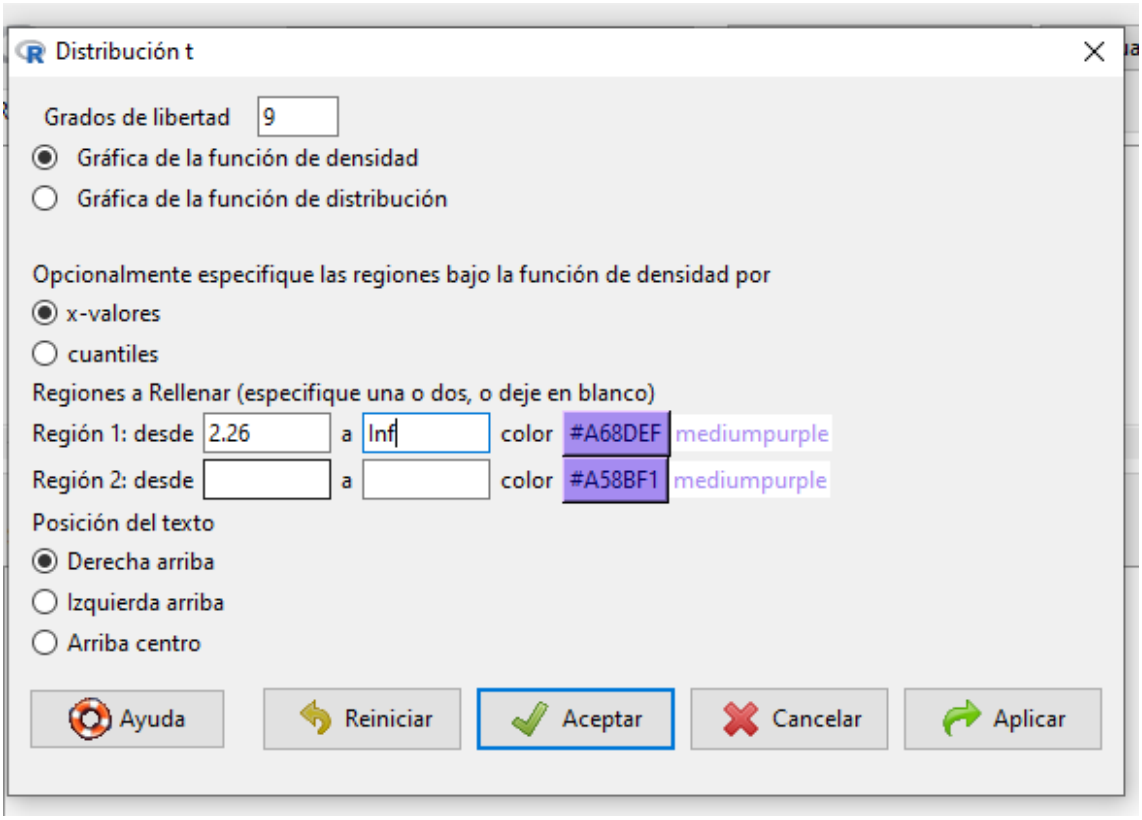
b) Calcular la probabilidad de obtener un valor mayor que 2.26 o menor que -2.26 en una distribución t con 9 gdl

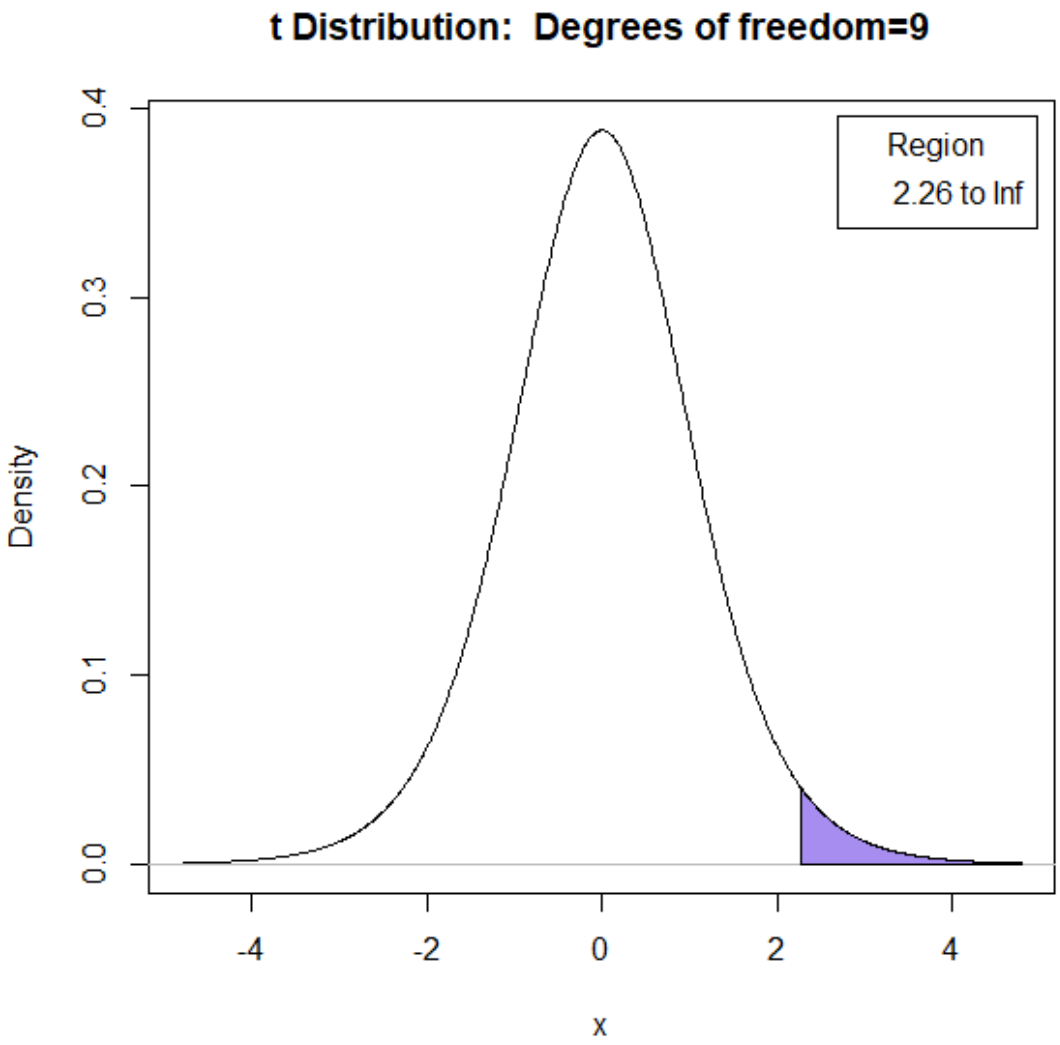
c) Calcular el valor de t después del cual se encuentre el 5% del área dela curva con 9 gdl

d) Calcular el valor de t para alfa= 0.05 con 9 gdl y dos colas

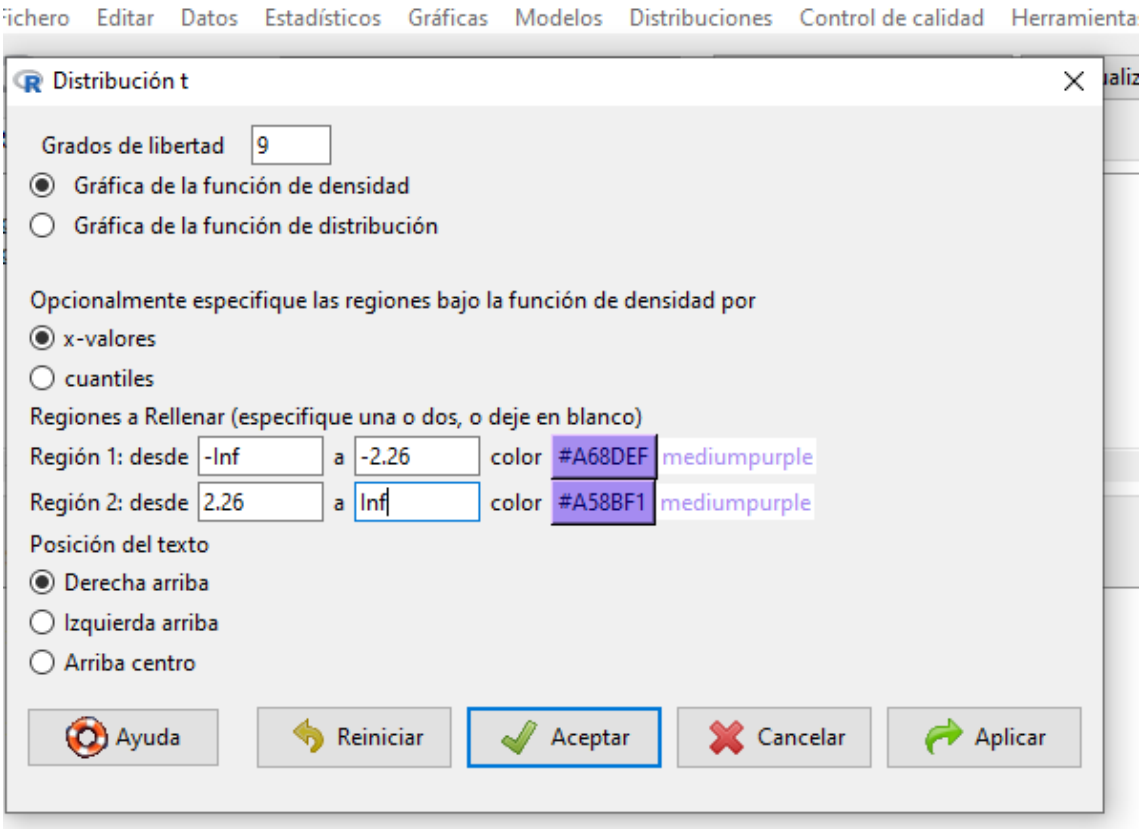
a) Solución




Distribución t
Grados de libertad
9
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde
2.26
a
Inf
color
#A68DEF
mediumpurple
Región 2: desde
a
color
#A58BF1
mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar


t Distribution: Degrees of freedom=9
Region
2.26 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

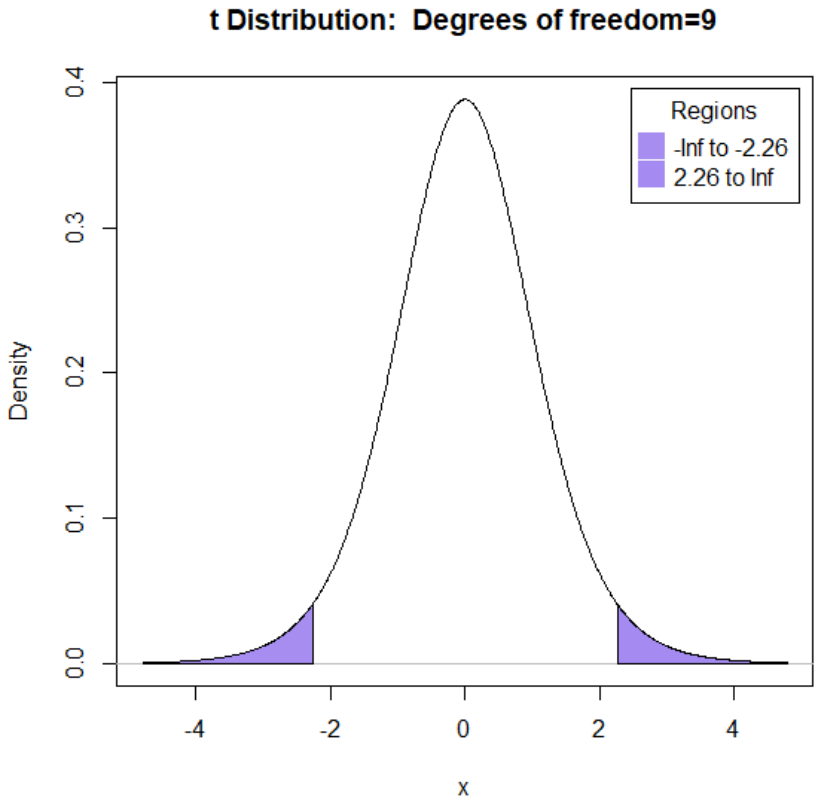
b) Solución



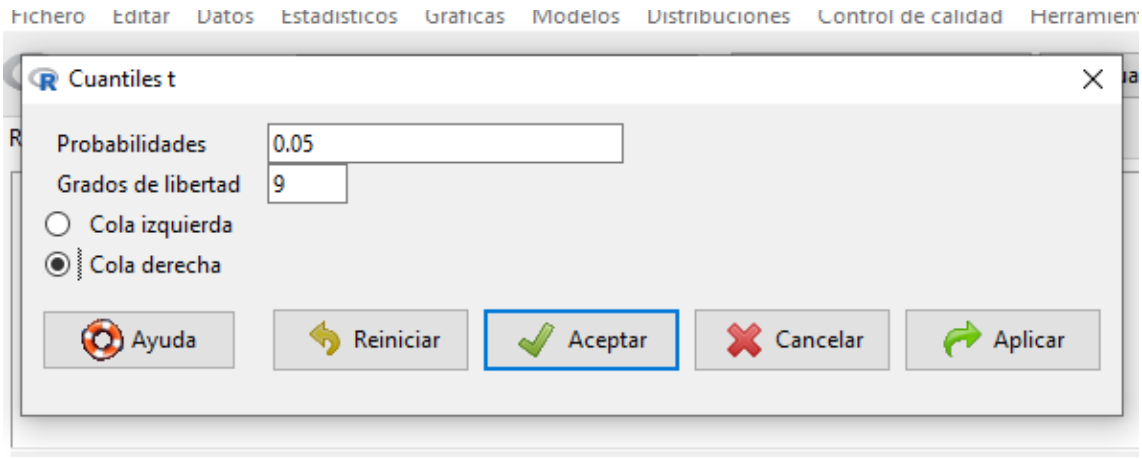
Salida

```
> pt(c(2.26), df=9, lower.tail=FALSE)
[1] 0.0250883

> pt(c(-2.26), df=9, lower.tail=TRUE)
[1] 0.0250883
```

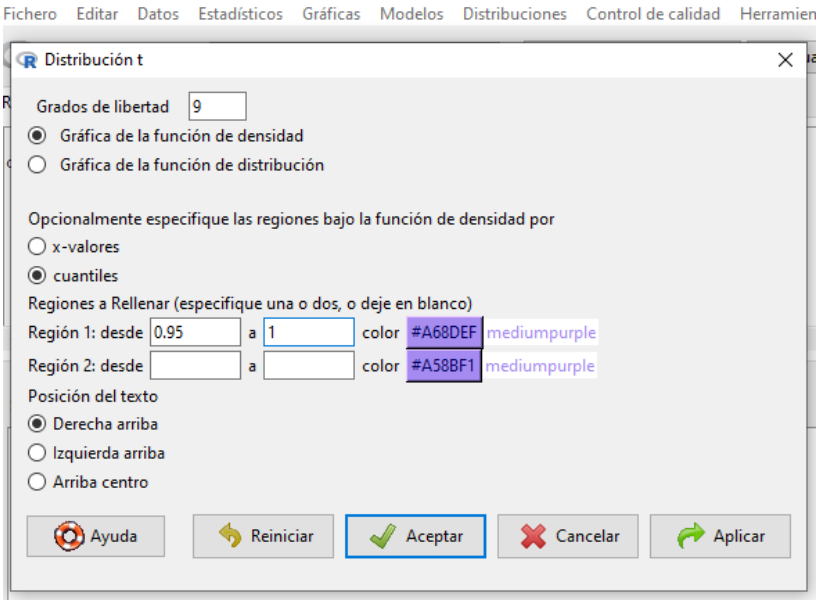


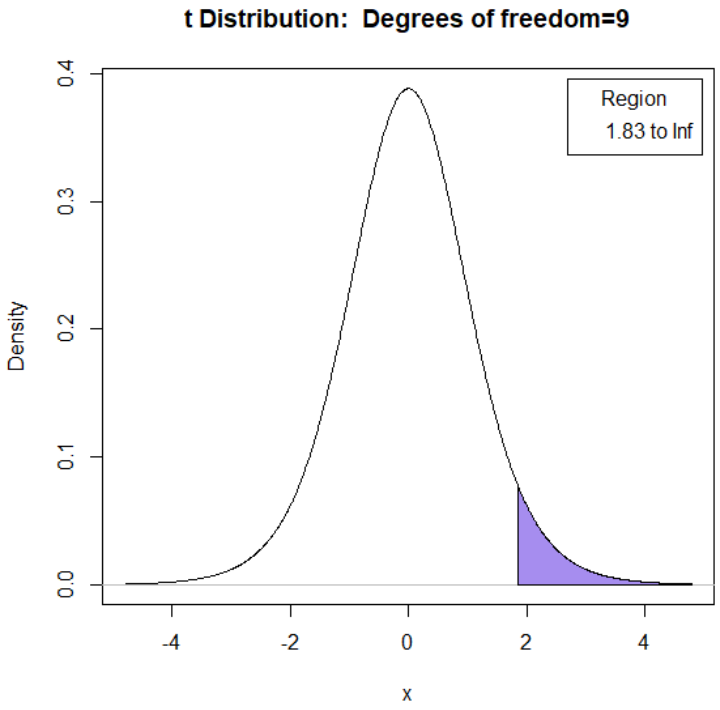
c) Solución



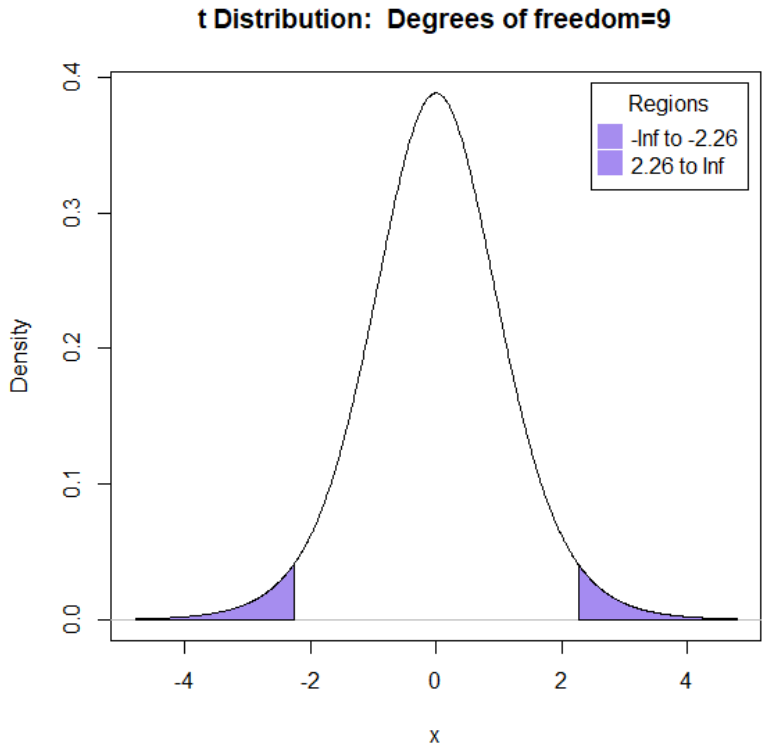
Salida

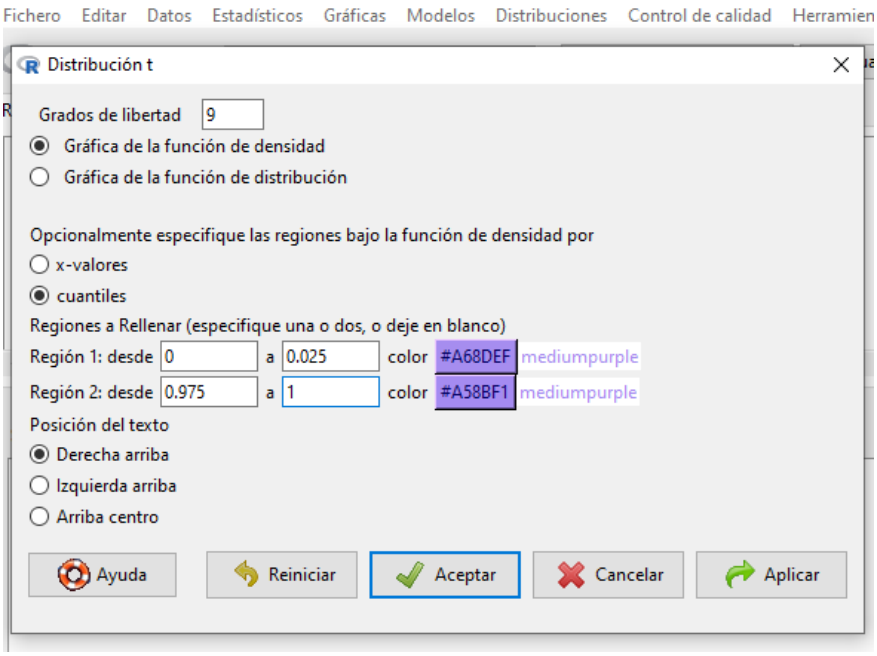
```
> qt(c(0.05), df=9, lower.tail=FALSE)
[1] 1.833113
```

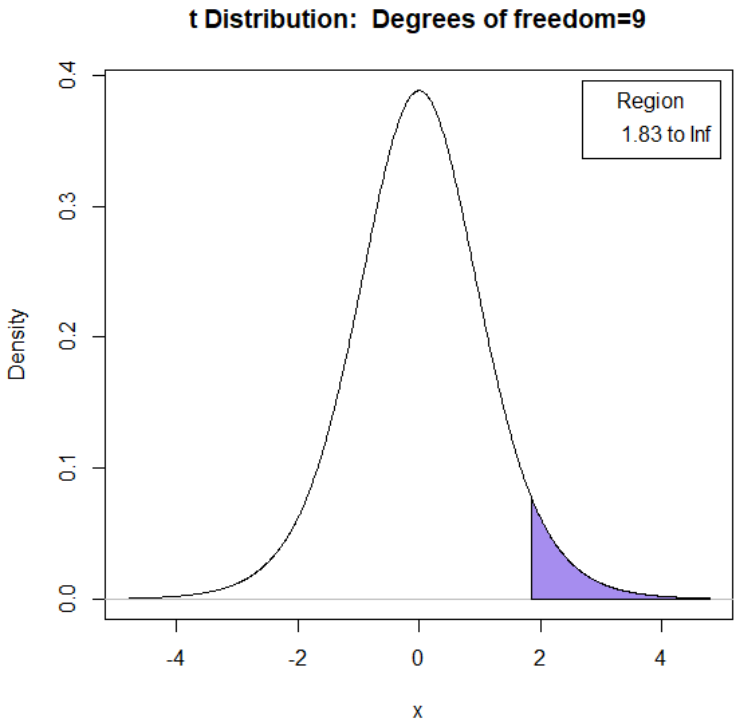

Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramien
Distribución t
Grados de libertad 9
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0.95 a 1 color #A68DEF mediumpurple
Región 2: desde a color #A58BF1 mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar


t Distribution: Degrees of freedom=9
Region
1.83 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

d) Solución


t Distribution: Degrees of freedom=9
Regions
-Inf to -2.26
2.26 to Inf
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramien
Distribución t
Grados de libertad 9
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0 a 0.025 color #A68DEF mediumpurple
Región 2: desde 0.975 a 1 color #A58BF1 mediumpurple
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar



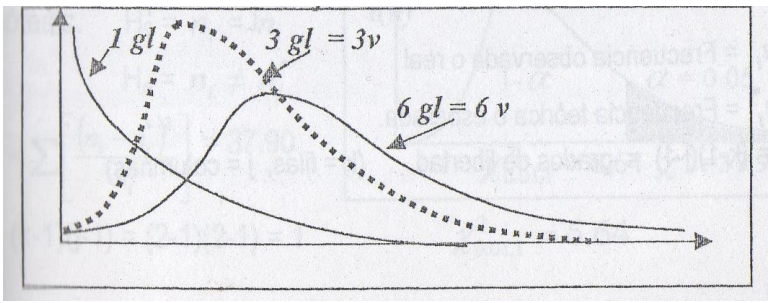
3.3. Distribución Chi cuadrado y Fisher

- **Distribución Chi cuadrado:**

Esta distribución fue introducida por F.R Helmert en 1876 y redescubierta en 1900 por Kart Pearson. Tiene muchos usos importantes, incluyendo ensayos de hipótesis acerca de proporciones y cálculo de intervalos de confianza para varianzas. Hay una distribución ji cuadrada diferente según el valor de n-1, lo cual representa los grados de libertad (gl). Así:

Ilustración 23

Las distribuciones "ji cuadrada" no son simetricas



Fuente: (Alvarez Roman, 2004)

Cuando gl es grande (v > 30), la distribución ji cuadrada se aproxima a la distribución normal. La variable $\sqrt{2x^2}$ es asintóticamente normal con media $\sqrt{2v-1}$ y varianza 1.

La curva está dada por: $Y = C(X^2)^{\frac{V-2}{2}} e^{\frac{x^2}{2}}$

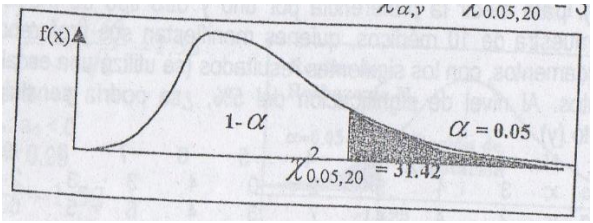
Donde v=n-1=grados de libertad c=constante que depende de v, para que el área bajo la curva sea 1. (Alvarez Roman, 2004)

a) ¿Cómo leer en la tabla?

Se busca en la primera fila Xv^2Y en la primera columna gl, en la intersección de la fila y la columna correspondiente se encuentra el valor de x^2 correspondiente.

Ejemplo 1: Si se tiene una variable aleatoria que sigue una distribución x^2 con 20 grados de libertad, obtener $x^2 \propto, v$ para:

$$\propto = 5\% \qquad x^2 \propto, v = x^2 0.05,20 = 31,42$$



b) Proceso para la prueba Ji-cuadrada:

1) Formular la hipótesis.
2) Establecer las diferencias entre las frecuencias observadas y las esperadas, se eleva a cuadrado y se divide cada una de ellas para la frecuencia teórica esperada.
3) Se suma y se obtiene Ji-cuadrada.

c) Ecuación sin corregir:

$$x^2 = \sum \left(\frac{(ni - ni*)^2}{ni*}\right)$$

d) Ecuación con corrección de Yates:

$$x^2 = \sum \left(\frac{((ni - ni*) - 0{,}5)^2}{ni*}\right)$$

Donde: ni = Frecuencia; ni* = Frecuencia teórica o esperada; $v = (k-1)(j-1) = grados\ de\ libertad.$ (k = filas, j = columnas)

La corrección de yates se utiliza cuando la tabla es de 2x2, es decir, v = 1 y las variables son discretas. En muestras grandes se obtienen prácticamente los mismos resultados. La corrección de Yates hoy es muy poco utilizada por cuanto se ha demostrado que, en la mayoría de casos la hipótesis nula no se rechaza. (Alvarez Roman, 2004)

1) Durante una epidemia se obtuvieron los siguientes datos sobre la efectividad de una vacuna como medida preventiva para los médicos. Estos datos, ¿indican la efectividad de la vacunación con base en el nivel de significación del 1%?

Tabla 28

Ejercicio 1

TRATAMIENTO	ENFERMOS	NO ENFERMOS	TOTAL
Vacunados	192	4	196
No vacunados	113	34	147
TOTAL	305	38	343

Fuente: (Alvarez Roman, 2004)

Calculamos: $ni *= n.p$

$$n1 *= 305(\frac{196}{343}) = 174{,}28;$$

$$n2 *= 305(\frac{147}{343}) = 130{,}71;$$

$$n3 *= 38(\frac{196}{343}) = 21{,}71;$$

$$n4 *= 38(\frac{147}{343}) = 16{,}29;$$

Calculamos x^2

Tabla 29

Cálculo de x2

n_i	n_i^*	$n_i - n_i^*$	$(n_i - n_i^*)^2$	$\frac{(n_i - n_i^*)^2}{n_i^*}$
192	174.28	17.72	313.99	1.802
113	130.71	-17.71	313.64	2.399
4	21.71	-17.71	313.64	14.45
34	16.29	17.71	313.64	19.25
343				37.90
n				$\chi^2 = \sum\left[\frac{(n_i - n_i^*)^2}{n_i^*}\right]$

Fuente: (Alvarez Roman, 2004)

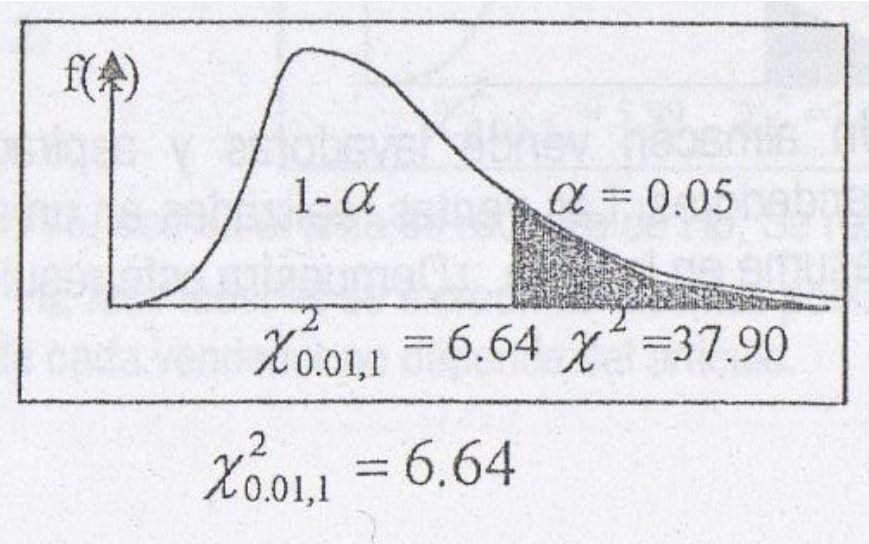
Hipótesis:

$$Ho = ni = ni *$$

$$Ha = ni \neq ni *$$

$$x^2 = \sum (\frac{(ni - ni*)^2}{ni*}) = 37{,}90$$

$$v = (k-1)(j-1) = (2-1)(2-1) = 1$$



Decisión: Como $x^2 = 37{,}90$, cae en el área de rechazo de Ho. Se rechaza la hipótesis nula y se acepta $Ha = ni \neq ni*$. Es decir, la diferencia es significativa. (Alvarez Roman, 2004)

Tabla 30

Aplicación de la corrección de Yates

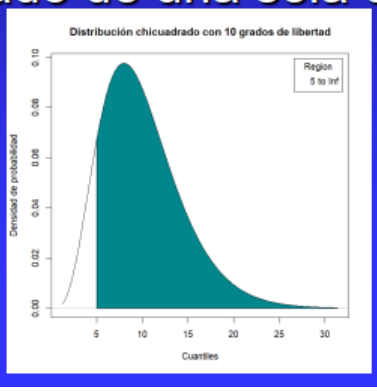
n_i	n_i^*	$\lvert n_i - n_i^* \rvert$	$\lvert n_i - n_i^* \rvert - 0.5$	$(\lvert n_i - n_i^* \rvert - 0.5)^2$	$\frac{(\lvert n_i - n_i^* \rvert - 0.5)^2}{n_i^*}$
192	174.28	17.72	17.21	296.18	1.699
113	130.71	17.71	17.21	296.18	2.266
4	21.71	17.71	17.21	296.18	13.64
34	16.29	17.71	17.21	296.18	18.18
					35.785
n					$\chi^2 = \sum \left[\frac{(\lvert n_i - n_i^* \rvert - 0.5)^2}{n_i^*} \right]$

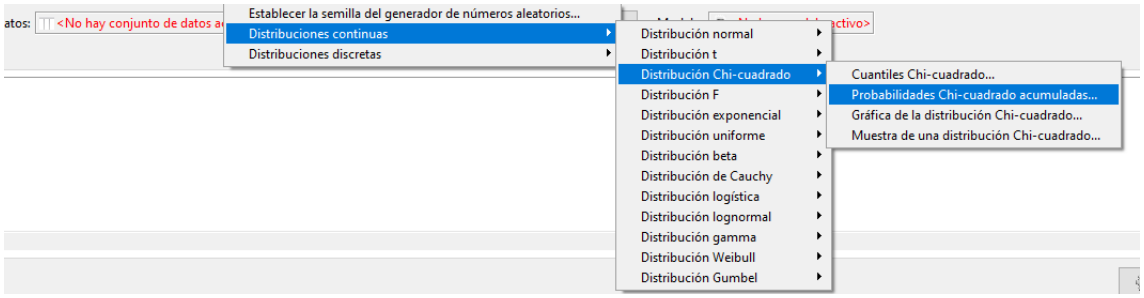
$\chi^2 = \sum \left[\frac{(\lvert n_i - n_i^* \rvert - 0.5)^2}{n_i^*} \right] = 35.785$ (se llega a la misma conclusión)

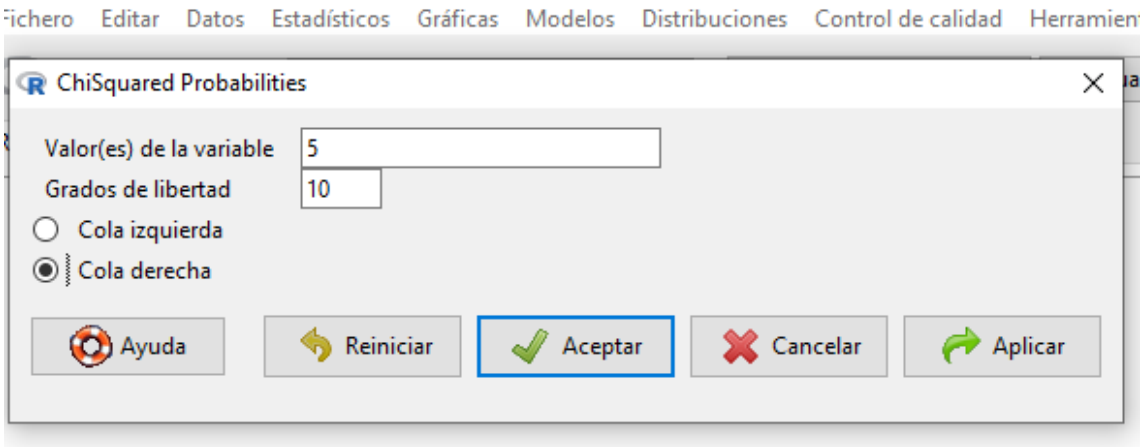
Fuente: (Alvarez Roman, 2004)

Probabilidad χ^2 Rcmdr

- pchisq(c(5), df=10, lower.tail=FALSE)
- [1] 0.891178
- Devuelve la probabilidad de una variable aleatoria continua siguiendo una distribución chi cuadrado de una sola cola con ν g.d.l.
- $P(X>\chi^2)$

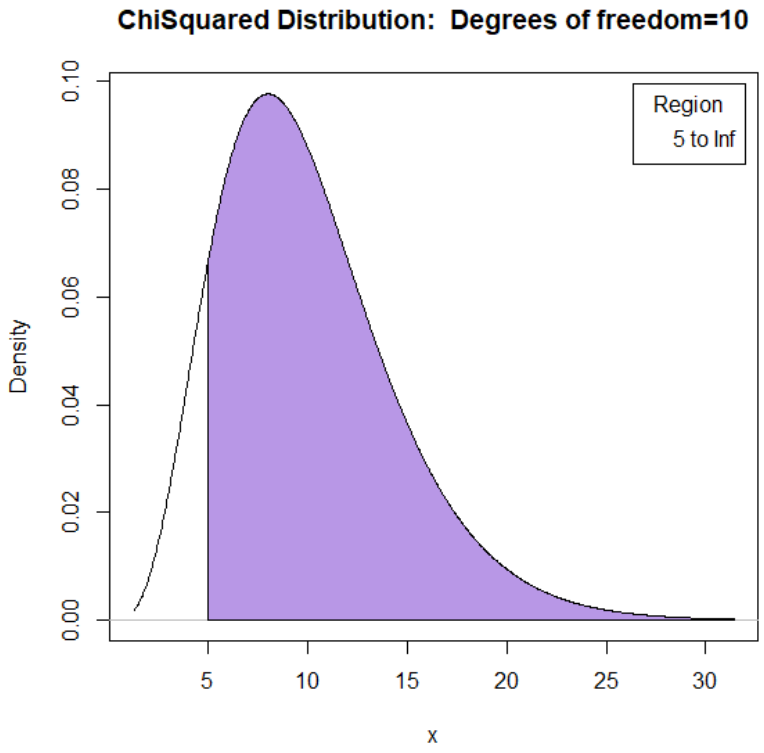






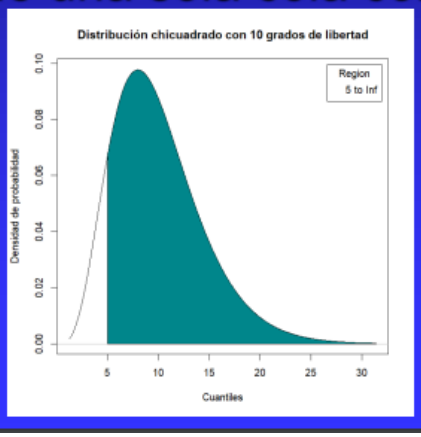
Salida

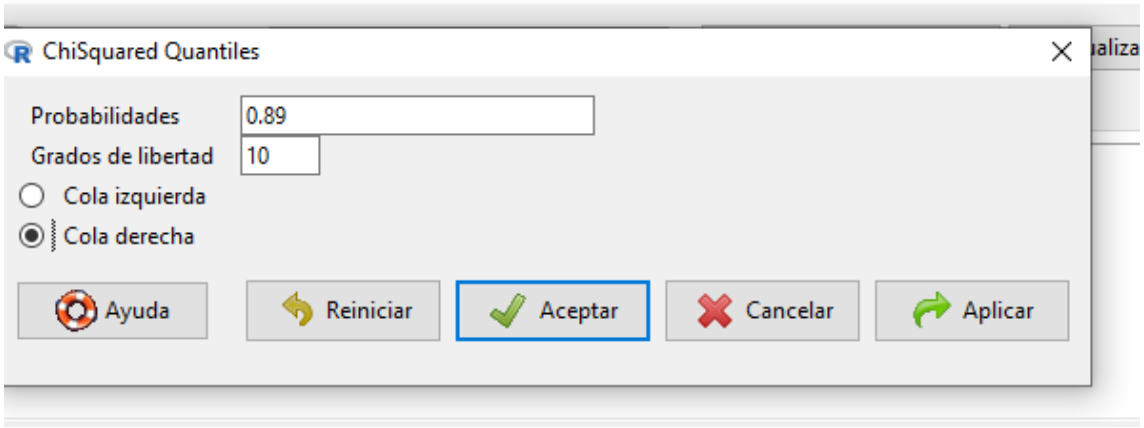
```
> pchisq(c(5), df=10, lower.tail=FALSE)
[1] 0.891178
```



Probabilidad χ^2 Inversa en Rcmdr

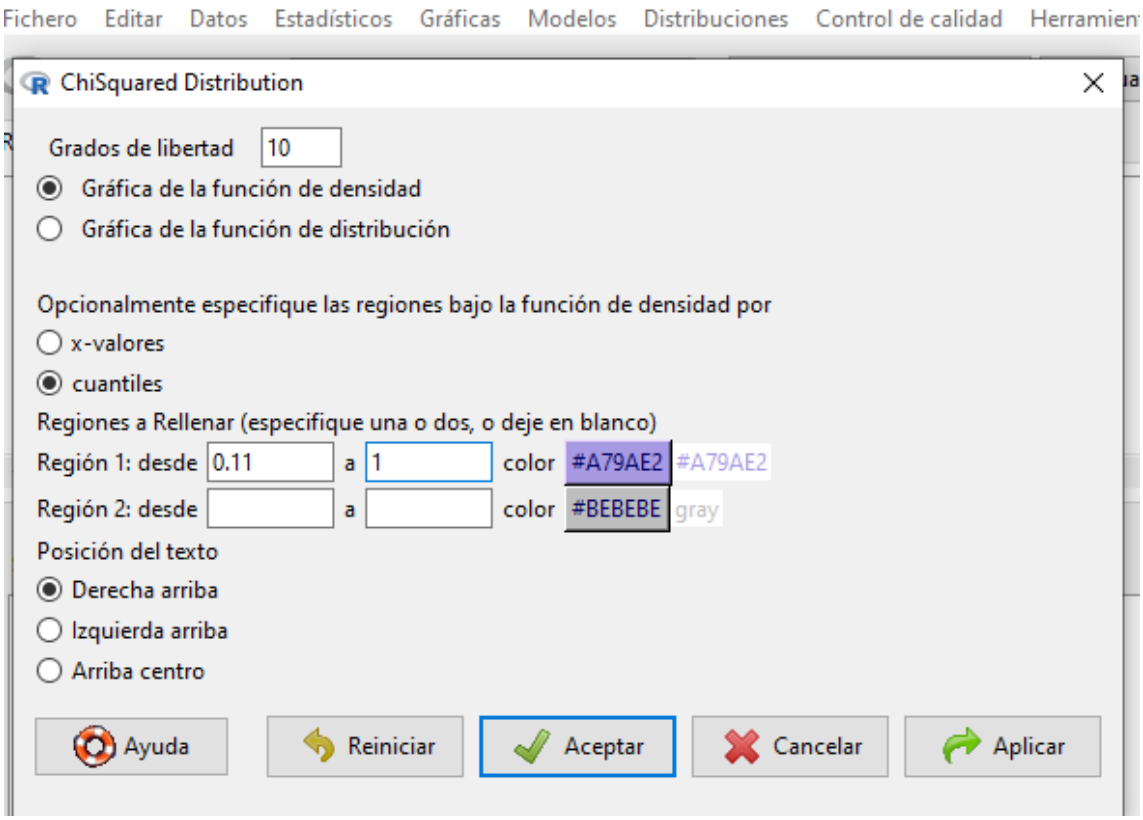
- qchisq(c(0.89), df=10, lower.tail=FALSE)
- [1] 5.017588
- Devuelve el valor de χ^2 para una probabilidad dada, de una distribución Ji-cuadrado de una sola cola con ν g.d.l.

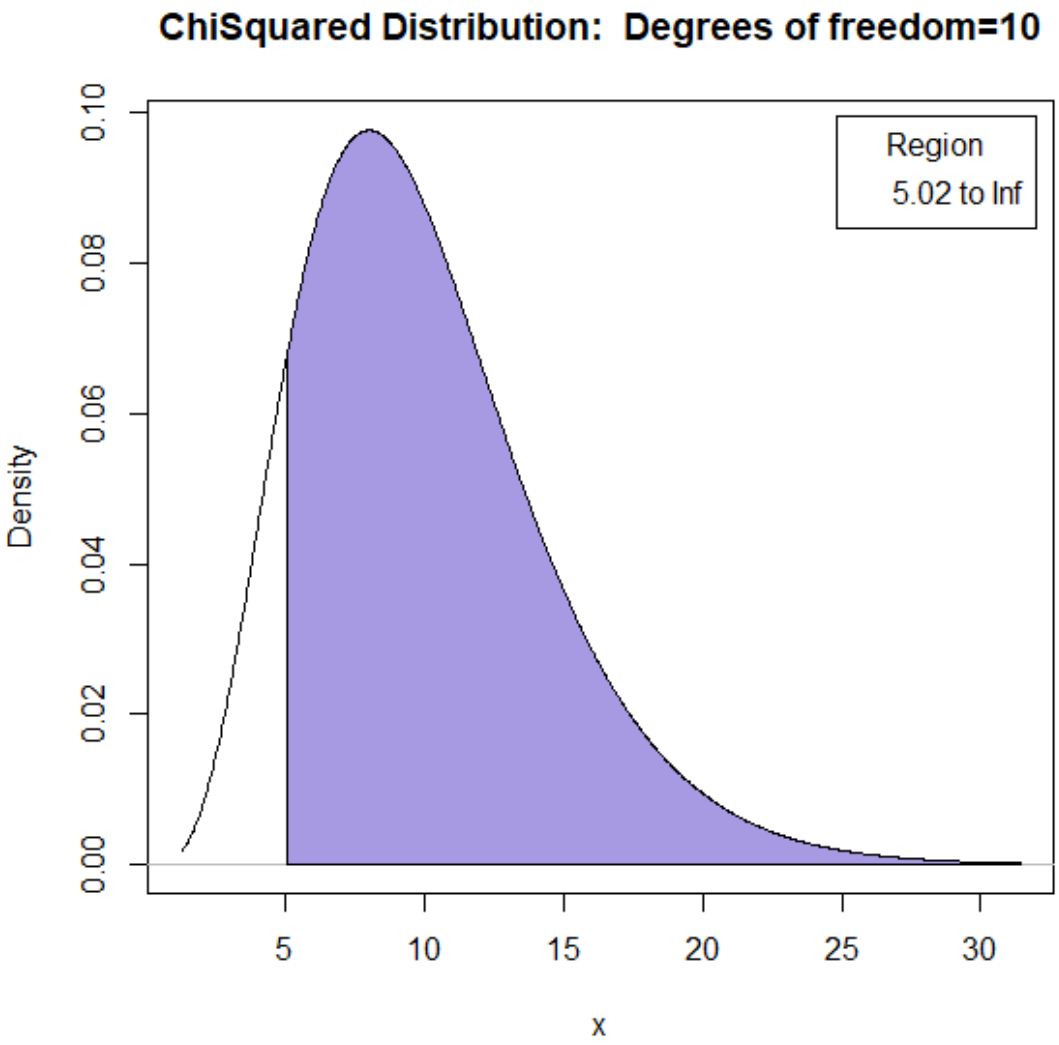




Salida

```
> qchisq(c(0.89), df=10, lower.tail=FALSE)
[1] 5.017588
```


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramien
ChiSquared Distribution
Grados de libertad 10
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0.11 a 1 color #A79AE2 #A79AE2
Región 2: desde a color #BEBEBE gray
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar

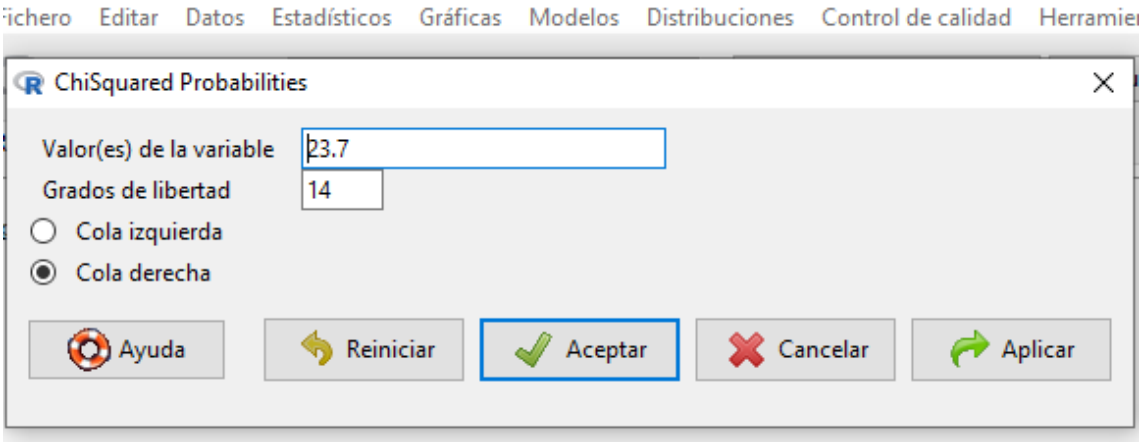


Ejercicios de la DistribucionJi-cuadrado con gráficos

a) Calcular la probabilidad de obtener un valor mayor de 23.7 en una distribución χ^2 con $\nu = 14$ g.d.l.

b) Calcular el valor de χ^2 despues del cual se encuentre el 5% del área en una distribución Ji-cuadrado con 4 g.d.l.

a) Solución

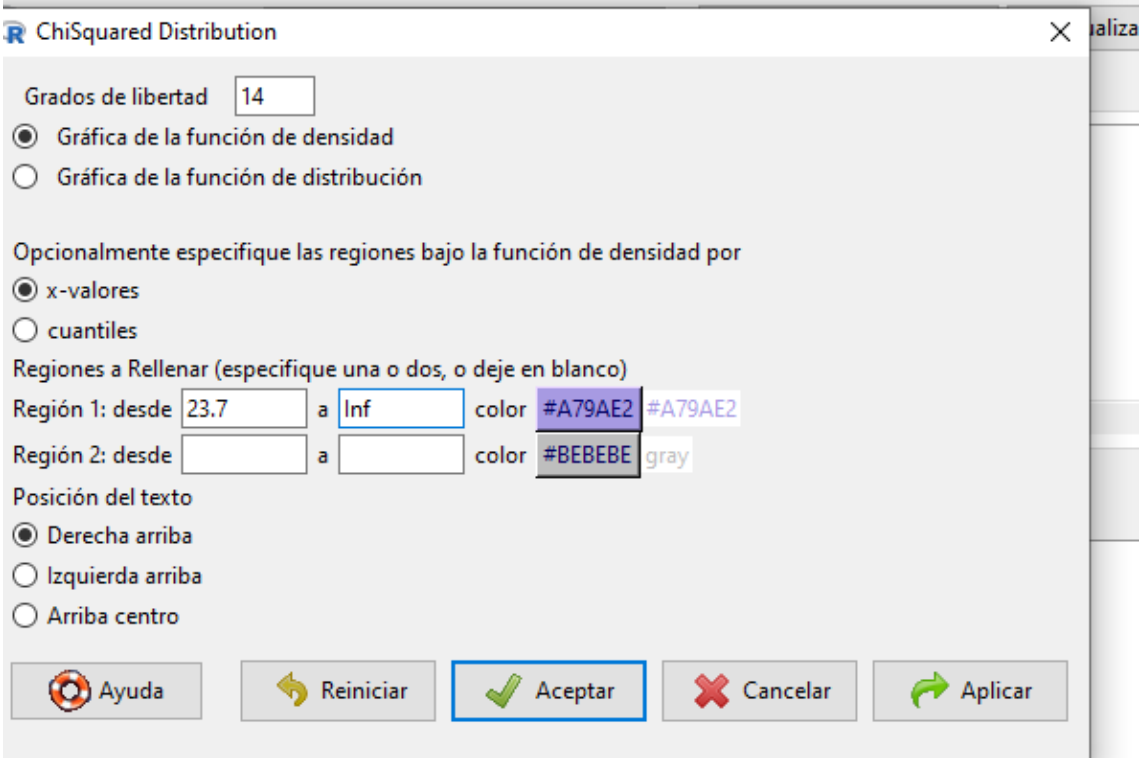


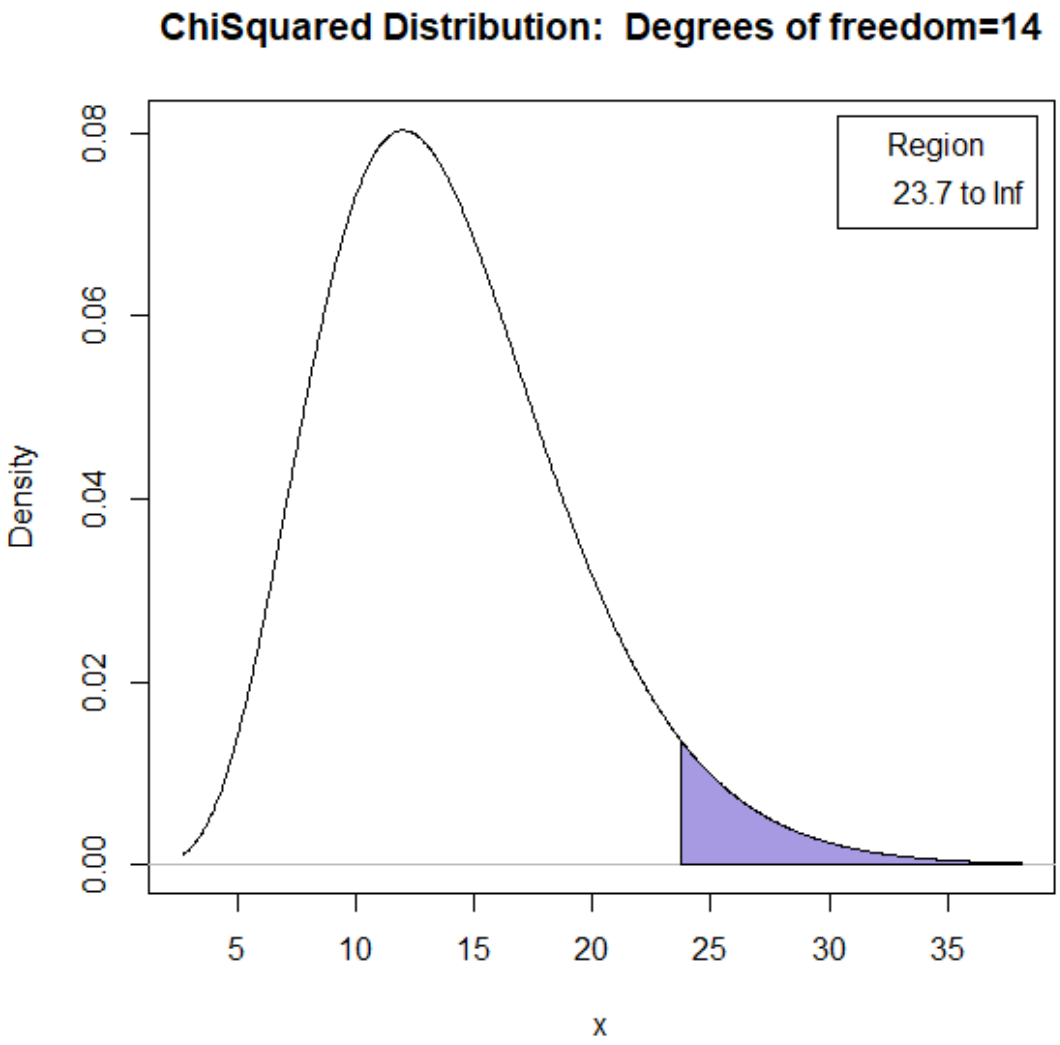
Salida

```
> pchisq(c(23.7), df=14, lower.tail=FALSE)
[1] 0.04979085
```

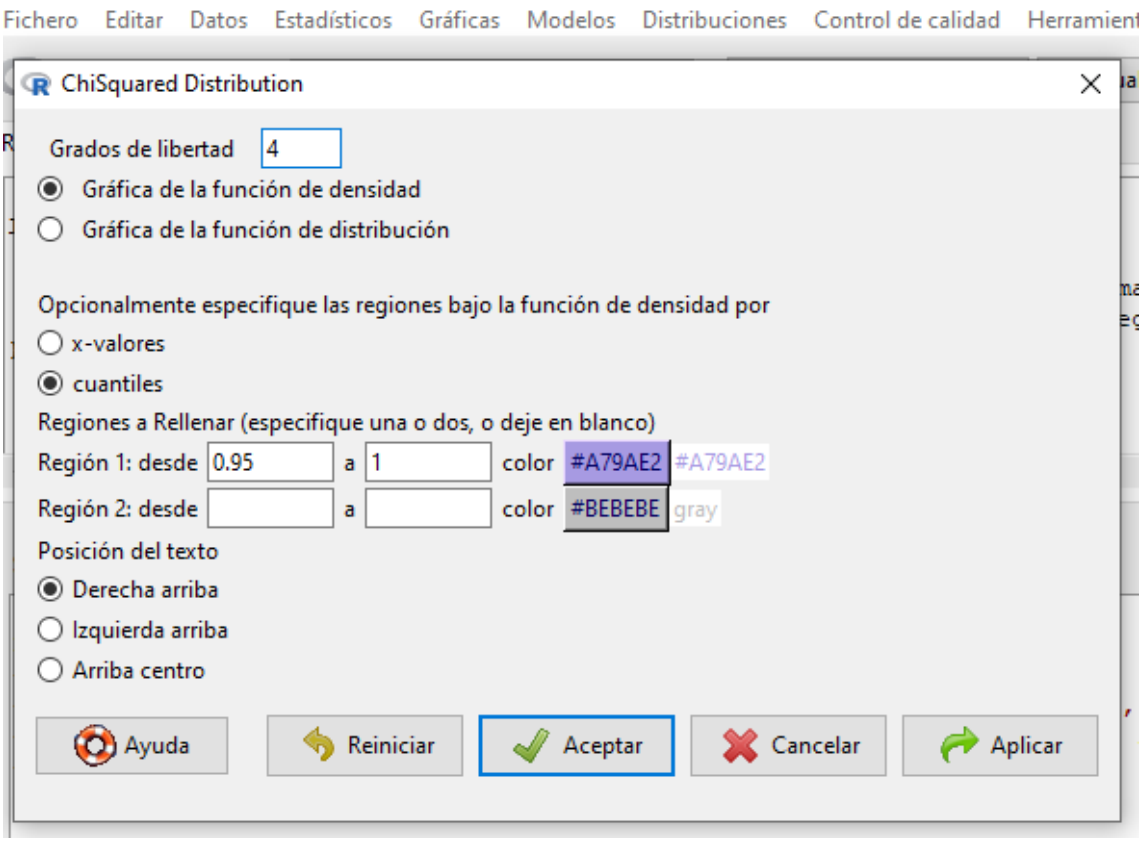
Salida

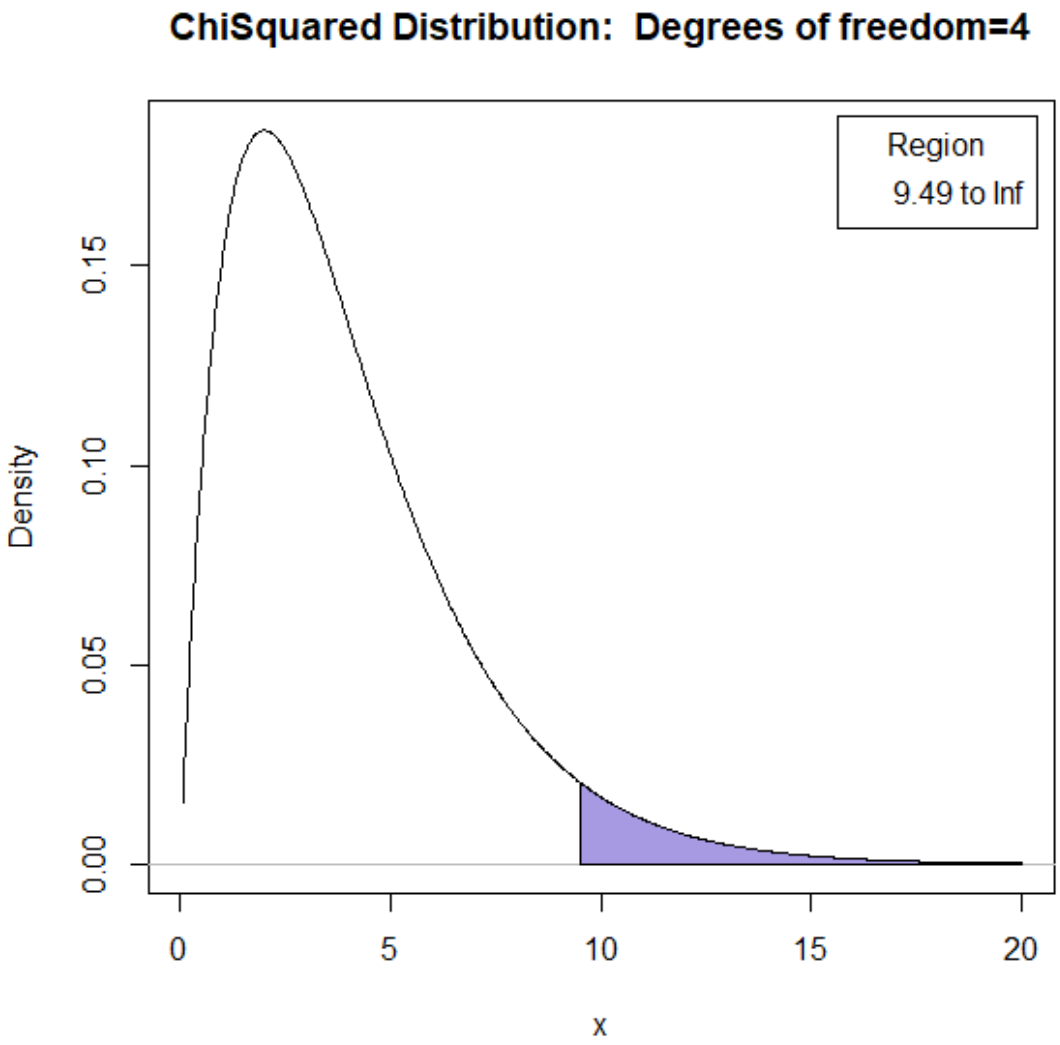
```
> pchisq(c(23.7), df=10, lower.tail=FALSE)
[1] 0.008437766
```




ChiSquared Distribution: Degrees of freedom=14
Region
23.7 to Inf
Density
0.00
0.02
0.04
0.06
0.08
5
10
15
20
25
30
35
x

b) Solución


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramient
ChiSquared Distribution
Grados de libertad 4
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0.95 a 1 color #A79AE2 #A79AE2
Región 2: desde a color #BEBEBE gray
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar



- **Fisher:**

Las inferencias con respecto a σ^2 cuando se muestrea una distribución normal, se formulan con base en la estadística $(n-1)S^2/\sigma^2$, la que tiene una distribución chi-cuadrada con n – 1 grados de libertad. En esta sección se desarrollará la estadística apropiada para emplearse en la formación de inferencias con respecto a las varianzas de dos distribuciones normales independientes con base en las muestras aleatorias de cada una. Por último, se analizará la teoría de una distribución muy útil, la cual se conoce como distribución F.

Supóngase un experimento en que se observan dos variables aleatoria independientes X y Y, cada una con una distribución chi cuadrada con v1 y v2 grados de libertad respectivamente. Sea F una variable aleatoria que es función de X y Y, de manera tal que

$$F = \frac{X/v1}{Y/v2}$$

Esto es, la variable aleatoria F es el cociente de dos variables aleatorias chi-cuadrada, cada una dividida por sus grados de libertad. Lo anterior lleva al siguiente teorema.

Teorema 1: Sean X y Y dos variables aleatorias independientes chi cuadrada con v1 y v2 grados de libertad, respectivamente. La variable aleatoria

$$F = \frac{X/v1}{Y/v2}$$

Tiene una distribución F con una función de densidad de probabilidad dada por

$$g(f; v1, v2) *= \begin{cases} \dfrac{T((v1+v2)/2)v1^{v1/2}v2^{v2/2}}{T(v1/2)T(v2/2)}\, f^{(v1-2)/2}(v2+v1f)^{-(v1+v2)/2} & f > 0. \\ 0 & para\ cualquier\ otro\ valor \end{cases}$$

(Le deducción de la función de densidad de probabilidad de F es similar a la de t de Student y se deja como ejercicio para el lector.)

La distribución F se caracteriza completamente por los grados de libertad v1 y v2. Puede demostrarse que el valor esperado es

$$E(F) = V2/(V2\ -\ 2) \qquad\qquad V2 > 2$$

Y la varianza está dada por

$$Var(F) = \frac{V2^2(2V2 + 2V1 - 4)}{V1(V2\ -\ 2)^2(V2\ -\ 4)} \qquad V2 > 4.$$

La distribución F tiene asimetría positiva para cualesquiera valores de v1 y v2, pero esta va disminuyendo conforme v1 y v2 toman valores cada vez más grandes.

Los valores cuantiles f1 $(-\propto, v1, v2)$ tales que

$$P(F \leq f1{-}\propto, v1, v2) = \int_0^{f1-\propto,v1,v2} g(f; v1, v2)df = 1 - \propto, \qquad 0 \leq \propto \leq 1$$

Se emplea g para denotar la función de densidad y de esta forma evitar cualquier confusión con respecto al argumento f.

Para las proporciones acumulativas seleccionadas $1 - \propto$ y distintas combinaciones de los grados de libertad del numerador v1, y del denominador v2 del cociente. Por ejemplo, si v1 = 5 y v2 = 10, entonces:

$$P(F \leq f0.90{,}5.10) = P(F \leq 2.52) = 0.90$$

$$P(F \leq f0.95{,}5.10) = P(F \leq 3.33) = 0.95$$

$$P(F \leq f0.99{,}5.10) = P(F \leq 5.64) = 0.99$$

Los valores cuantiles f1 $-\propto, v1, v2$ únicamente para $\propto <\ 0.5$. Si se desean los cuantiles del lado izquierdo (es decir, para $\propto >\ 0.5$) estos pueden encontrarse mediante el siguiente procedimiento: si la variable aleatoria F tiene una distribución F con v1 y v2 grados de libertad, entonces la variable $F` = 1/F$ también tiene una distribución F pero con v2 y v1 grados de libertad. Puede verse que lo anterior es cierto, a partir de:

$$F` = \frac{1}{\frac{X/v1}{Y/v2}} = \frac{Y/v2}{X/v1}.$$

Si se desean los valores cuantiles f1 $-\propto, v1, v2$ únicamente para $\propto >$ 0,5.

$$P(F \leq f1{-}\propto, v1, v2) = P(\frac{1}{F} > \frac{1}{f1{-}\propto, v1, v2}) = 1 - \propto,$$

O

$$P(\frac{1}{F} \leq \frac{1}{f1{-}\propto, v1, v2}) = \propto$$

Pero $1/F = F` \sim F$ se encuentra distribuida con v2 y v1 grados de libertad. Entonces el $\propto -esimo$ valor cuantil de F` es tal que

$$P(F` \leq f1{-}\propto, v1, v2) = \propto$$

Dado que $P(\frac{1}{F} \leq \frac{1}{f1-\propto,v1,v2}) = \propto$ y $P(F` \leq f1{-}\propto, v1, v2) = \propto$ son idénticas, se sigue que

$$f1{-}\propto, v1, v2) = 1/\, f1{-}\propto, v1, v2 \qquad \text{for } \propto > 0.5$$

Como ejemplo, sea v1 = 8 y v2 = 12. Entonces

$$P(F \leq f0.05,8,12) = P(F \leq 1/f``0,95,12,8) = P(F \leq 1/3,28) = P(F \leq 0,305) = 0,05$$

$$P(F \leq f0.04,8,12) = P(F \leq 1/f``0,99,12,8) = P(F \leq 1/5,67) = P(F \leq 0,176) = 0,01$$

Regresando al problema de desarrollar una estadística apropiada para usarse en la formulación de inferencias con respecto a las varianzas de dos varianzas de dos distribuciones normales independientes, sea X1, X2, …, Xnx una muestra aleatoria de variables aleatorias independientes y normalmente distribuidas cada una con media μxy varianza $\sigma^2 x$. También se Y1, Y2, …, Yny un conjunto de ny variables aleatorias independientes normalmente distribuidas, cada una con media μy y varianza $\sigma^2 y$. Si se supone que las X y las Y son independientes, las estadísticas

$$(nx - 1)S^2 x/\sigma^2 x$$

y

$$(ny - 1)S^2 y/\sigma^2 y$$

Son dos variables aleatorias chi-cuadrada i independientes con $nx - 1$ y $ny - 1$ grados de libertad, respectivamente. Entonces, se desprende que la variable aleatoria

$$\frac{\frac{(nx-1)S^2 x}{\sigma^2 x}/(nx - 1)}{\frac{(ny-1)S^2 x}{\sigma^2 y}/(ny - 1)} = \frac{S^2 x/\sigma^2 x}{S^2 y/\sigma^2 y}$$

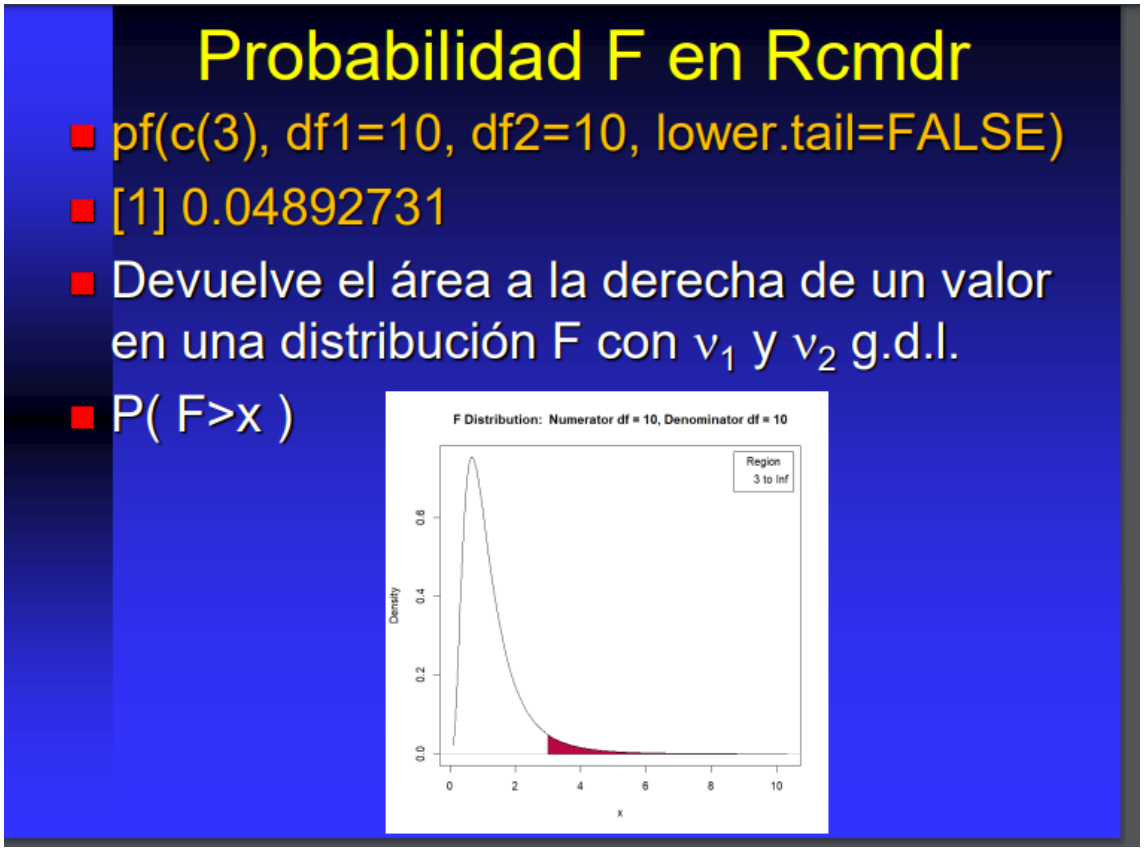
Tiene una distribución F con $nx - 1$ y $ny - 1$ grados de libertad.

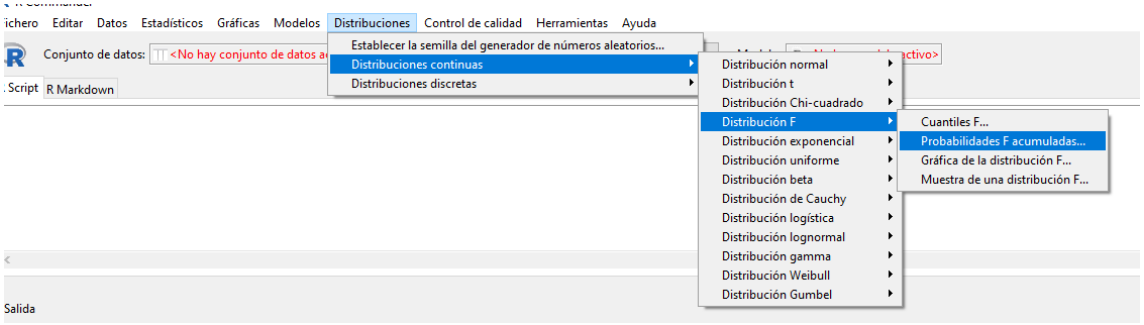
Una aplicación de la formula anterior es inmediata si se recuerda el problema general. Esto es, el formular una inferencia con respecto a la diferencia entre dos medias poblacionales ya sea cuando se conocen las varianzas de las poblaciones o cuando se supone que se conoce, al menos, el cociente de estas. Una forma factible de verificar la validez de esta suposición es mediante el empleo de la formula. Si la suposición de que $\sigma^2 x = \sigma^2 y$ es correcta, la estadística F dada en la formula anterior, se reduce a

$$F = s^2 x / s^2 y,$$

Cuando se obtienen los valores de $S^2 x$ y $S^2 y$ a partir de las muestras y se calcula el cociente de la anterior formula, puede concluirse que la hipótesis de varianza iguales es falsa si el valor de este cociente es, de manera suficiente, distinto de I. En otras palabras, si las dos varianzas son iguales, la probabilidad de observar un valor de F distinto, de manera suficiente, es pequeña.

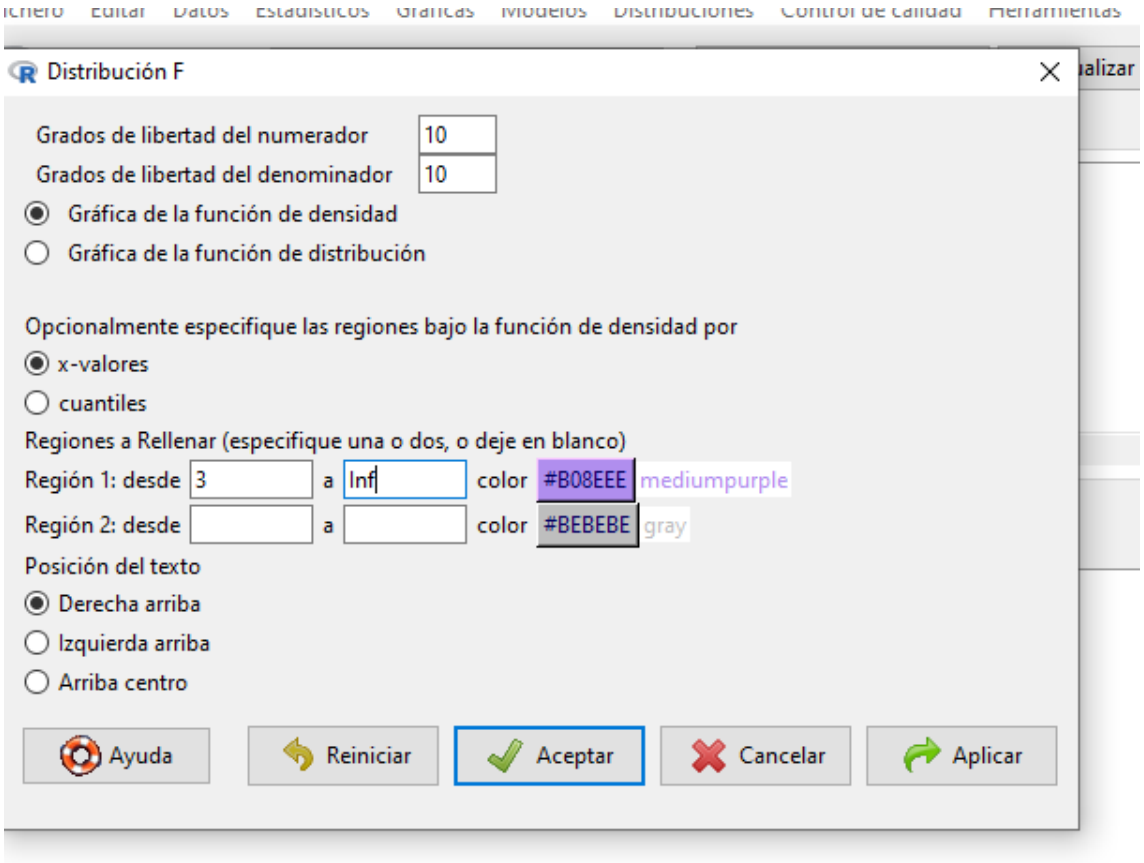
Para finalizar, debe notarse que, en esta sección, se desarrolló el material que se presentó bajo la hipótesis de realizar un muestreo aleatorio sobre poblaciones que tienen una distribución normal. En la realidad, la hipótesis de normalidad puede o no ser justificable. Sin embargo, desde un punto de vista práctico, el lector debe darse cuenta que la diferencia entre la distribución normal y el modelo de probabilidad de la población de interés es inversamente proporcional a las técnicas delineadas para formular inferencias. La afirmación anterior es particularmente cierta cuando se formulan inferencias con respecto a las varianzas cuando se emplean la distribución chi-cuadrada o la F.

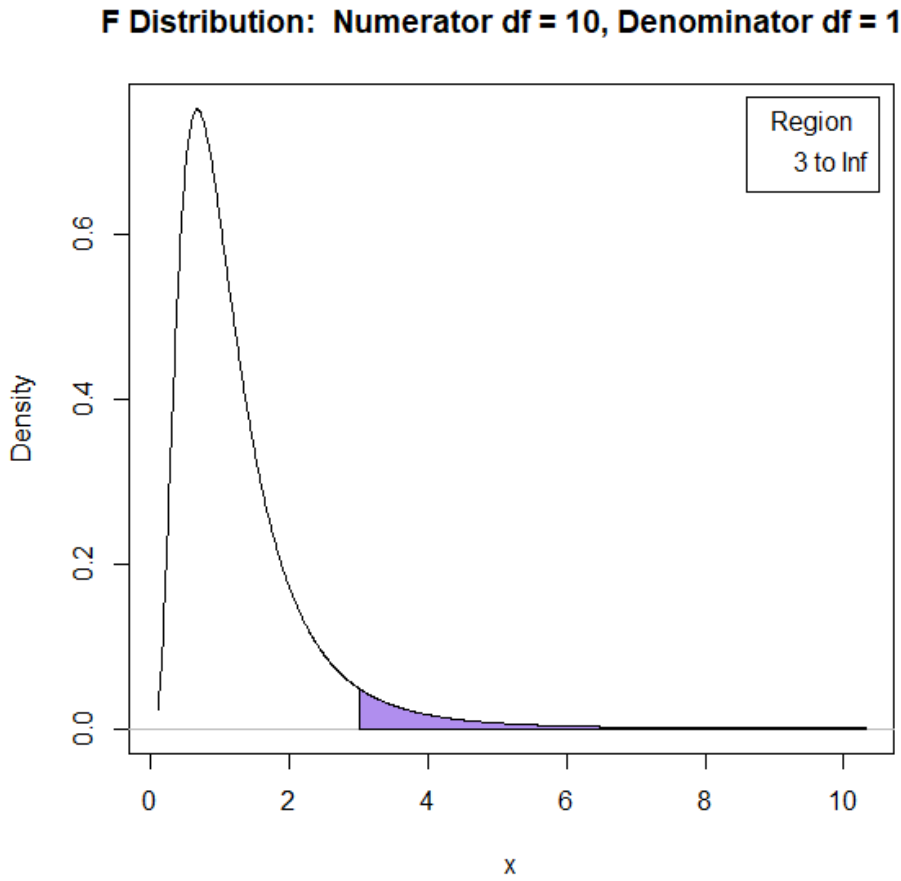




Salida

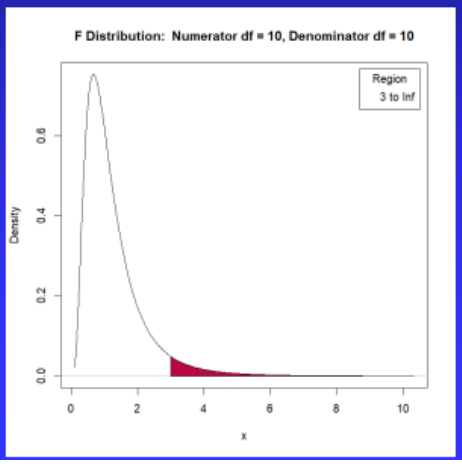
```
> pf(c(3), df1=10, df2=10, lower.tail=FALSE)
[1] 0.04892731
```

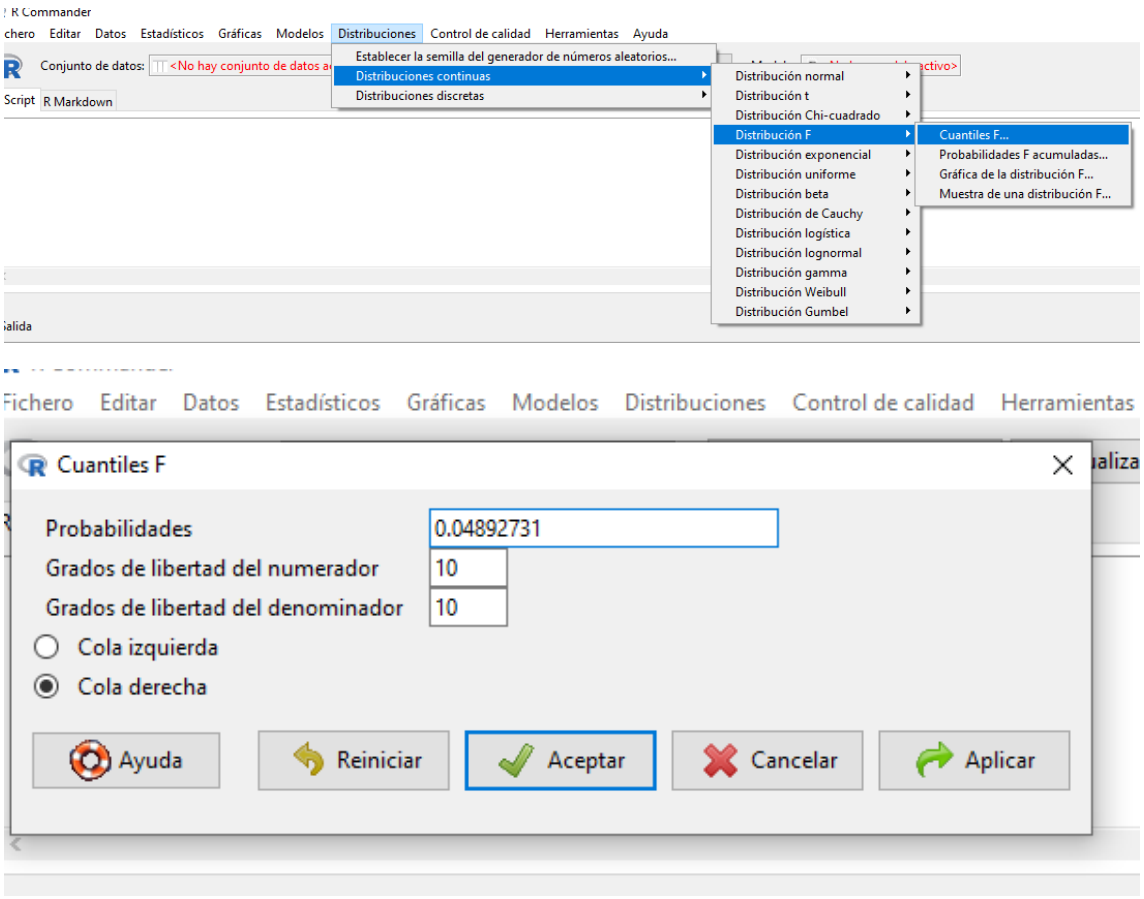




Probabilidad F Inversa

- qf(c(0.05), df1=10, df2=10, lower.tail=FALSE)
- [1] 2.978237
- Devuelve el valor crítico de F(α) para una distribución F con ν_1, ν_2 g.d.l.





Salida

```
> qf(c(0.04892731), df1=10, df2=10, lower.tail=FALSE)
[1] 3
```

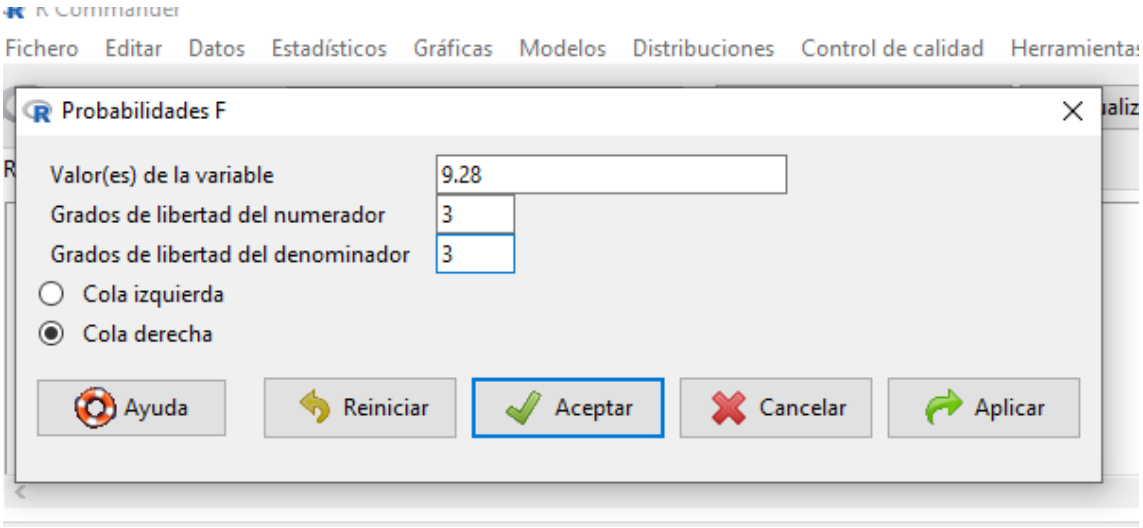
Ejercicios con gráficos

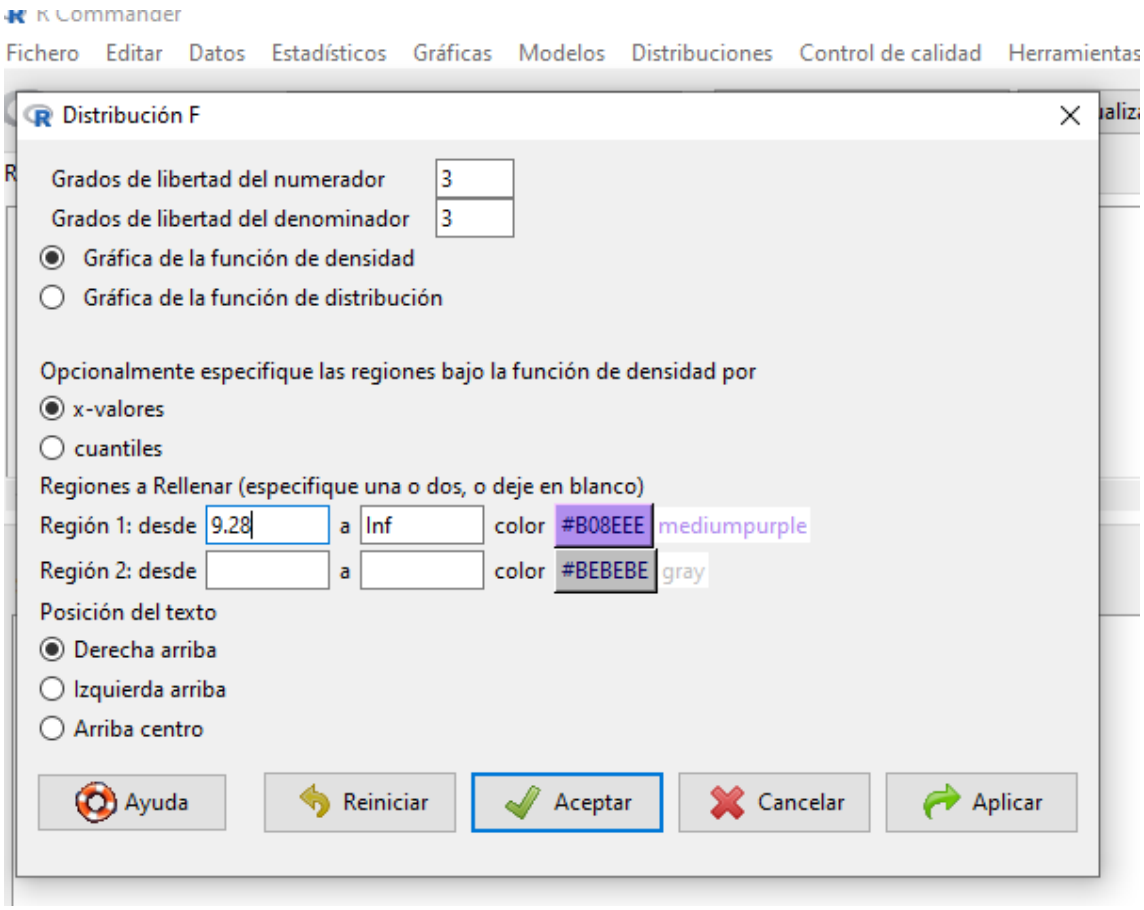
- Ejercicio – Distribucion F

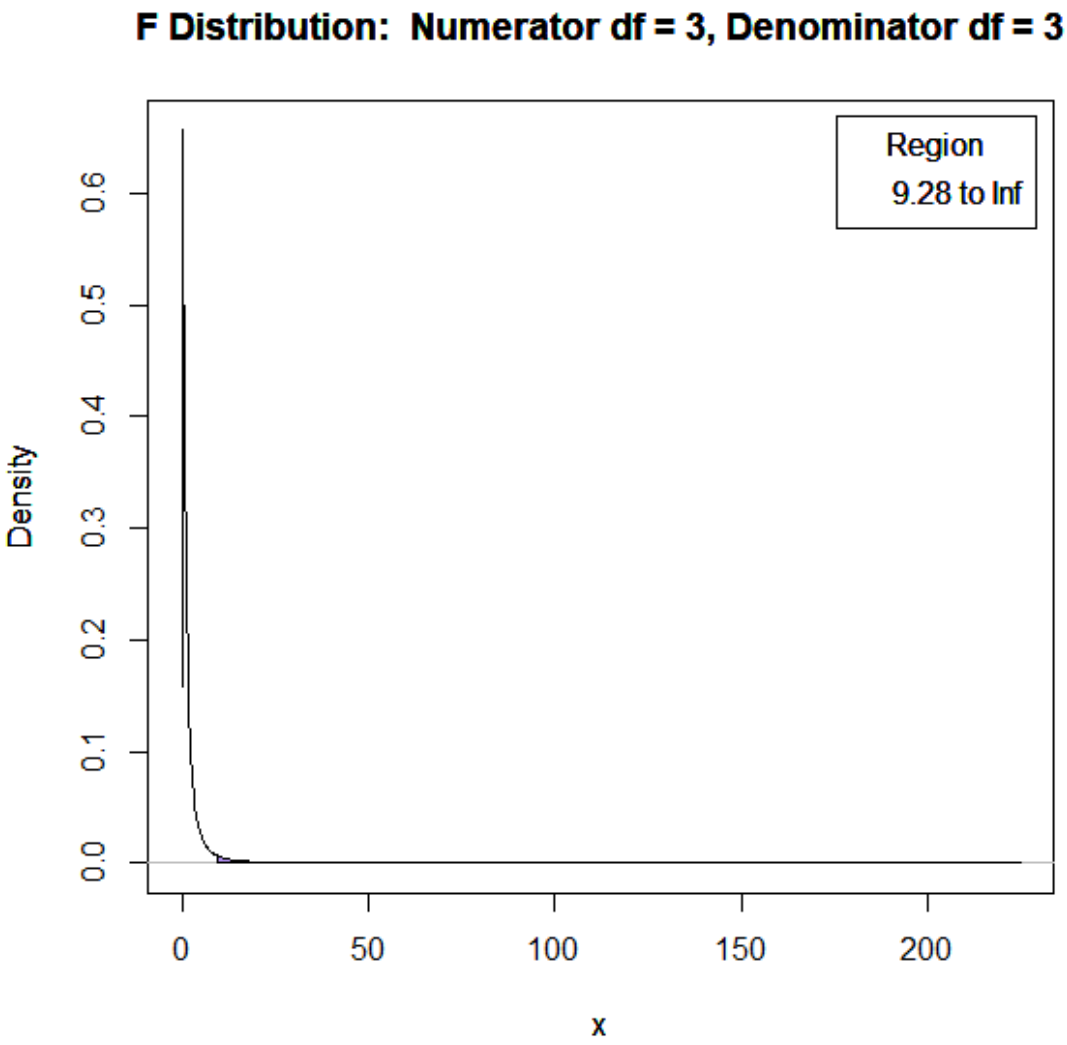
a) Determine la probabilidad de tener un valor de F mayor que 9.28 en una distribución F con ν_1=3 y ν_2=3 g.d.l.

b) Halle el valor crítico de $F_{(0.05)}$ para ν1=3 y ν2=15 g.d.l.

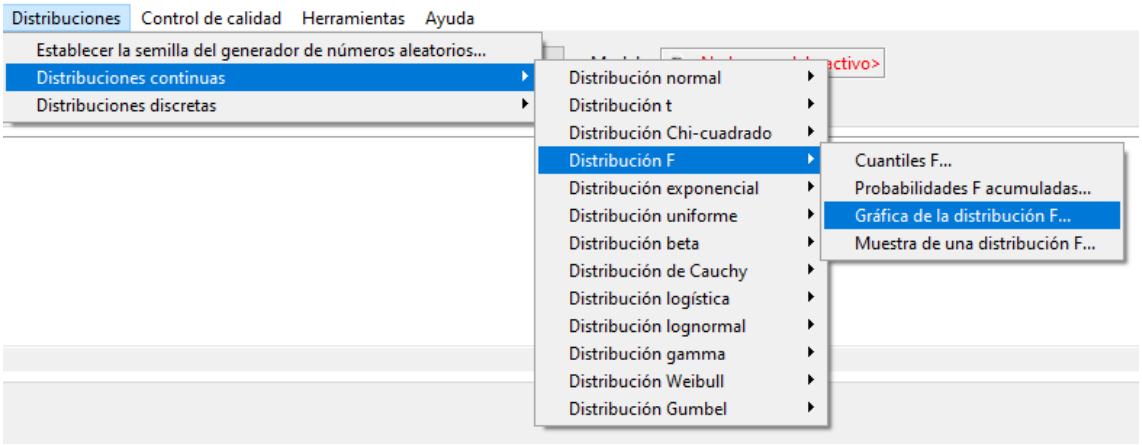
a) Solución

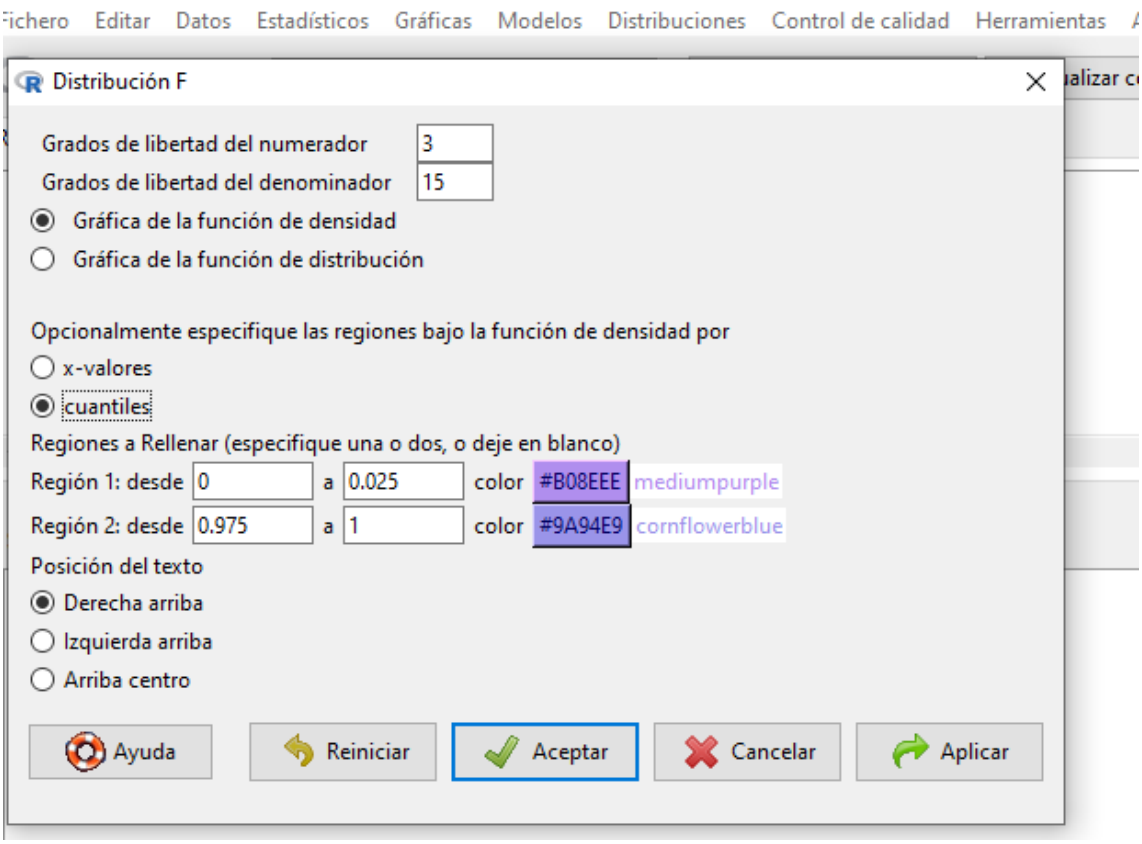


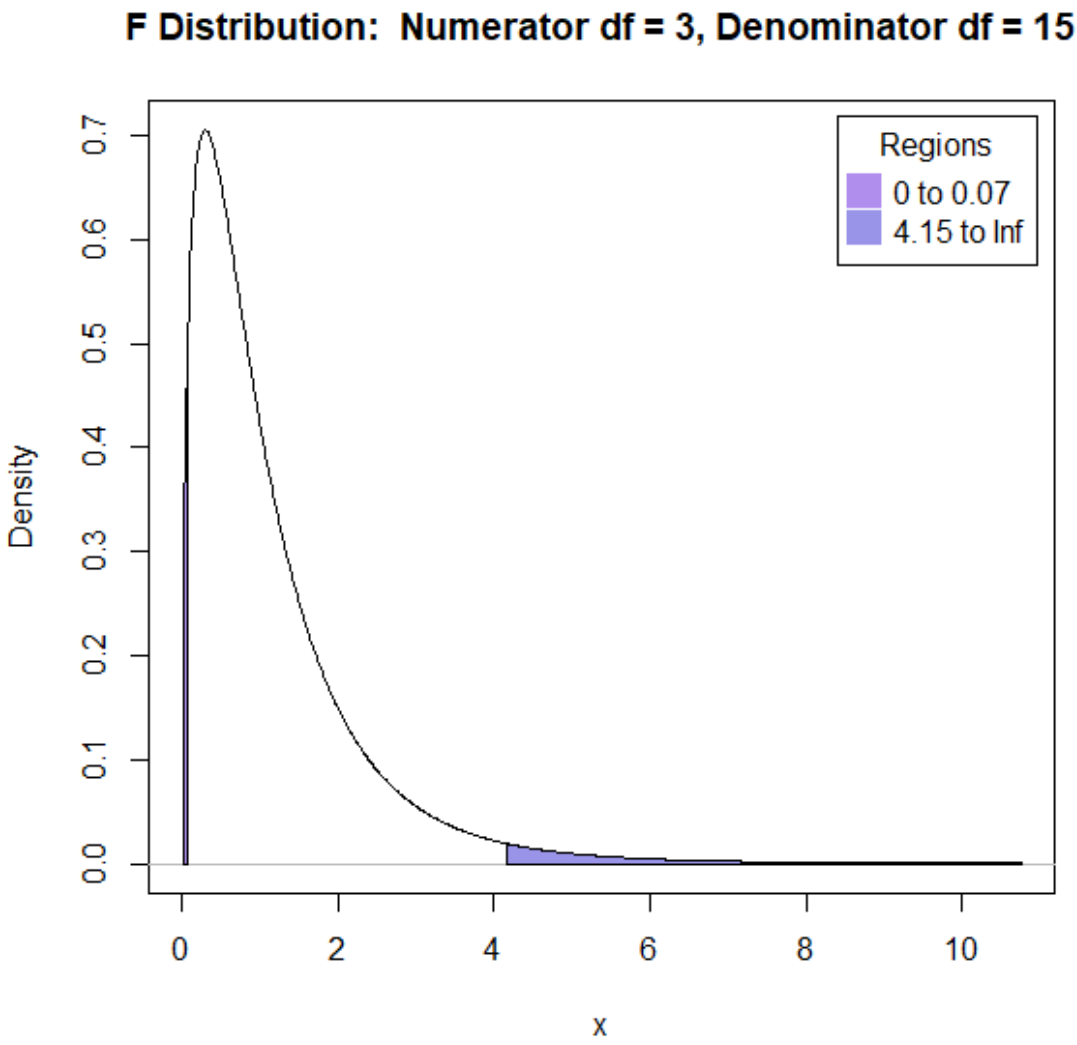

R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas
Distribución F
Grados de libertad del numerador 3
Grados de libertad del denominador 3
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 9.28 a Inf color #B08EEE mediumpurple
Región 2: desde a color #BEBEBE gray
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar


F Distribution: Numerator df = 3, Denominator df = 3
Region
9.28 to Inf
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
x

b) Solución


Distribuciones
Control de calidad
Herramientas
Ayuda
Establecer la semilla del generador de números aleatorios...
Distribuciones continuas
Distribuciones discretas
Distribución normal
Distribución t
Distribución Chi-cuadrado
Distribución F
Distribución exponencial
Distribución uniforme
Distribución beta
Distribución de Cauchy
Distribución logística
Distribución lognormal
Distribución gamma
Distribución Weibull
Distribución Gumbel
Cuantiles F...
Probabilidades F acumuladas...
Gráfica de la distribución F...
Muestra de una distribución F...

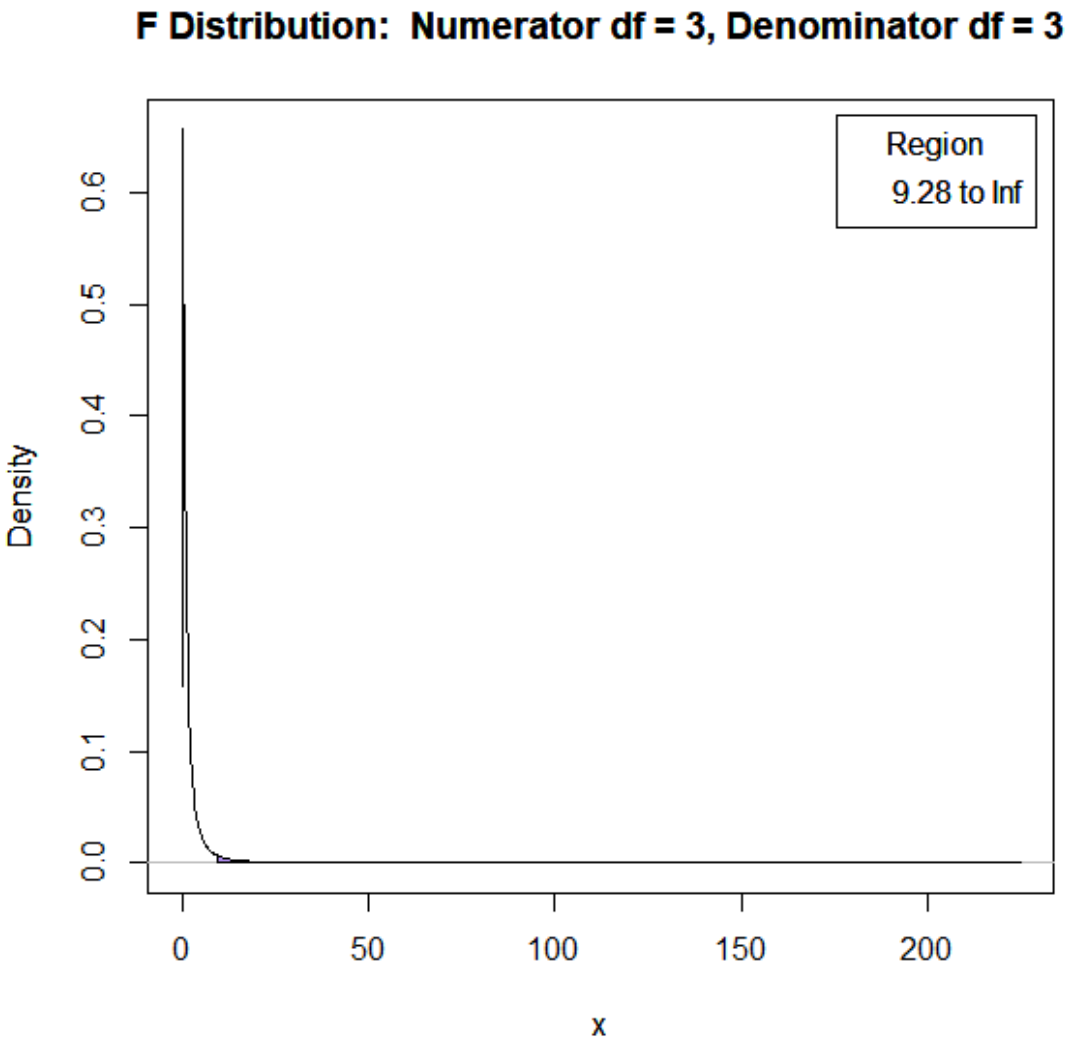

Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Control de calidad Herramientas
Distribución F
Grados de libertad del numerador 3
Grados de libertad del denominador 15
Gráfica de la función de densidad
Gráfica de la función de distribución
Opcionalmente especifique las regiones bajo la función de densidad por
x-valores
cuantiles
Regiones a Rellenar (especifique una o dos, o deje en blanco)
Región 1: desde 0 a 0.025 color #B08EEE mediumpurple
Región 2: desde 0.975 a 1 color #9A94E9 cornflowerblue
Posición del texto
Derecha arriba
Izquierda arriba
Arriba centro
Ayuda
Reiniciar
Aceptar
Cancelar
Aplicar



Salida

```
> pf(c(9.28), df1=3, df2=3, lower.tail=FALSE)
[1] 0.04997591
```

a



3.4. Teoría del muestreo. Muestreo no Probabilístico

El requerimiento básico de una muestra es que sea representativa de la población. La forma de seleccionar los individuos que han de constituir la muestra tiene, como es lógico, una importancia capital para garantizar que esta permita obtener conclusiones que puedan extrapolarse a la población. No hay que olvidar nunca que el objeto final del estudio siempre es la población y que la muestra solo es un medio para obtener información sobre esta. El propósito de trabajar con muestras es disminuir el tiempo, los costos, cuando no podemos contar o medir todos los elementos de la población. Con el fin de permitir inferir conclusiones validas sobre una población la muestra debe ser "representativa" de esta. En teoría la única forma de garantizar representatividad de una muestra es seleccionar al azar los individuos. Aunque esta forma de proceder rara vez sea aplicable de forma estricta en la práctica, siempre hay que extremar las precauciones para que la forma real de obtener la muestra sea la mas parecida posible a la ideal. (Alvarez Roman, 2004)

a) Conceptos Básicos:

Población: Son todos los individuos o elementos de un conjunto, de las mismas o similares características de donde se toman las muestras para ser observadas. Ejemplo: El conjunto de estudiantes de la UNACH.

Muestra: Es un subconjunto de la población o del conjunto universo, solo se toman en cuenta unas pocas unidades para ser observadas; por cuanto resulta posible, fácil y económico en una investigación. Ejemplo: 100 alumnos de la UNACH.

Parámetros: Es cualquier característica de la población que sea medible. El valor verdadero del parámetro no se conoce, es lo que tratamos de describir mediante el muestreo.

Estimación: Llamada también "Estadístico", es la medición que resulta de la muestra escogida.

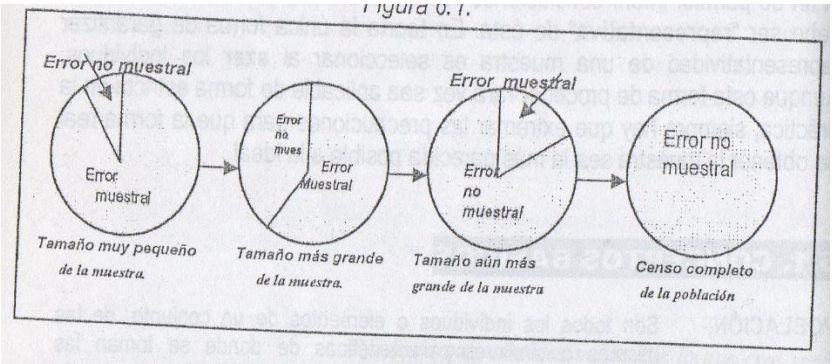
Confianza: Es el grado de certidumbre que tenemos sobre la exactitud de la estimación de la muestra. Existe un nexo estrecho entre el nivel de confianza y el grado de exactitud.

Error Muestral: Es ocasionado por el muestreo; este error es inevitable en el proceso.

Error No Muestral: Denominado también "sesgo "o tendencia a un error direccional. Puede presentarse incluso cuando hayamos hecho un censo de toda la población. (Alvarez Roman, 2004)

Ilustración 24

Error no muestral y muestral



Fuente: (Alvarez Roman, 2004)

b) Procedimiento Muestral:

Podemos resumir los pasos que intervienen en el proceso de muestreo:

1) Determinación de la población y los parámetros pertinentes
2) Seleccionar e marco apropiado del muestreo
3) Escoger entre el muestreo probabilístico y el no probabilístico
4) Escoger el método de muestreo que se utilizara
    (1) Método Probabilístico:
        (a) Muestreo Aleatorio Simple
        (b) Muestreo Estratificado
        (c) Muestreo por Conglomerados

(2) Método No Probabilístico:
   (a) Muestreo por Cuotas
   (b) Muestreo por Conveniencia (Alvarez Roman, 2004)

3.4.1. Muestreo Probabilístico

En el muestreo probabilístico, además, se puede considerar una distinción cuando los diferentes elementos de la muestra o bien son idénticos o similares y tienen la misma probabilidad de ser elegidas, o bien esta probabilidad es diferente. Así se diferencia entre muestreo con probabilidades desiguales. También el muestreo tiene implicaciones en función de si se trata de muestreo con reposición o sin reposición (o reemplazamiento).

3.4.2. Muestreo Aleatorio Simple

El muestreo aleatorio simple (MAS) es el tipo de muestreo más sencillo, pero fundamental pues constituye la técnica muestral básica de la estadística inferencial de donde se derivan las demás y con la que se comparan los demás métodos.

Una muestra aleatoria simple se define como aquella donde las unidades se seleccionan o extraen aleatoriamente cumpliendo estas condiciones:

1) Cada unidad de la población tiene idéntica probabilidad de ser incluida en la muestra: $Prob(Ui) = \frac{1}{N}$
2) Cada combinación de unidades, es decir, cada muestra posible de tamaño n que se puede seleccionar tiene igual probabilidad de constituirse (condición de equiprobabilidad). En total existen múltiples combinaciones de posibles muestras donde se seleccionan n casos entre una población de N^7.

De estas dos condiciones se deriva que la probabilidad de una unidad cualquiera Ui de pertenecer a una muestra será n/N. En este caso hemos supuesto una extracción de unidades sin reemplazamiento. Cuando existe reemplazamiento una misma unidad puede salir más de una vez en la selección y, en cada caso, todas las unidades tienen una probabilidad 1/N de aparecer en cada extracción.

Una muestra aleatoria, extraída al azar, no significa una muestra extraída casualmente o de forma azarosa, sino que se establecen determinadas condiciones para considerarla extracción al azar. Garantizar estas condiciones es de suma importancia. (Lopez Roldan & Fachelli, 2015)

Para extraer una muestra aleatoria tradicionalmente se disponía de tablas en papel con una relación de números dispuestos al azar y que determinaban las unidades a extraer. Hoy en día el software estadístico genera los números aleatorios. El procedimiento de extracción consiste en cuatro pasos simples:

1) Se numeran de 1 a N correlativamente todos los individuos de la población en la base de datos de la muestra
2) Se determina el tamaño de la muestra n.
3) Se generan los n números aleatorios.
4) Los n números se extraen del listado de la base de datos y constituyen la muestra.

En el contexto del MAS podemos realizar las estimaciones de los estadísticos muestrales. Los fundamentos de la utilización de la estadística descriptiva e inferencial en el contexto del muestreo. Aquí recuperamos algunas ideas generales que nos servirán para plantear dos tipos de tareas implicadas en el análisis de la información cuantitativa obtenida por muestreo: la determinación del tamaño de la muestra y la determinación del error asociado a una estimación una vez se ha obtenido la muestra. La primera tarea es el aspecto mas importante del diseño de muestras, la segunda, aplicando razonamientos similares, es una tarea propia del análisis posterior de los datos que veremos reiteradamente a lo lardo del texto; aquí simplemente se apuntaran algunas ideas generales de introducción. (Lopez Roldan & Fachelli, 2015)

a) Determinación del tamaño de la muestra

El objetivo de una muestra está en alcanzar la mayor representatividad o precisión posible en la estimación de los parámetros poblacionales. En otras palabras, y refiriéndonos en general al diseño de encuestas por muestreo, se trata de reducir el error muestral dadas unas limitaciones de tiempo, dinero y trabajo, lo que se traduce siempre en la decisión de fijar el numero de unidades de la muestra: cuantas encuestas se realizaran.

Por la ley de regularidad estadística se sabe que, a partir de un determinado número de unidades, los valores tienden a estabilizarse, nuevos elementos en la muestra aumentan cada vez en menor cuantía la fiabilidad y disminuyen poco el error, hecho que hace innecesario el aumento del tamaño a partir de un determinado momento. Se trata por tanto de encontrar el equilibrio entre la fiabilidad y disminuyen poco el error, hecho que hace innesario el aumento del tamaño a partir de un determinado momento. Se trata por tanto de encontrar el equilibrio entre la fiabilidad deseada, los objetivos de la investigación y los costes en tiempo, dinero y trabajo. (Lopez Roldan & Fachelli, 2015)

En la determinación del tamaño muestral se conjugan estos cuatro elementos que intervendrán en la fórmula de cálculo:

1) La amplitud del Universo, diferenciando dos situaciones: si la población es finita (si tienes menos de 100,000 individuos) o es infinita (a partir de 100,000 individuos). Con poblaciones finitas el tamaño muestral tiende a ser cada vez mas sensible al tamaño poblacional por lo que es necesario introducir un factor de corrección por finitud e igual a $\sqrt{\frac{N-n}{N-1}}$. Cuando la población es muy numerosa la diferencia con respecto al numero de muestra es muy pequeña y el numerador de ese factor tiene a igualarse al denominador, por lo que el factor tiende a ser 1; y tiende a ser un valor cada vez mayor a medida que población y muestra se aproximan en número de casos.
2) El nivel de confianza adoptado. Como comentamos en el primer apartado el nivel de confianza establece la probabilidad o confiablidad de nuestros resultados, los cuales se elaboran y razonan en términos probabilísticos. El criterio habitual que se sigue es considerar un nivel de la confianza del 95,5%, lo que implica considerar 2σ, es decir, un valor de dos unidades de desviación a partir de la media en la distribución normal, que se puede expresar también diciendo que z = 2 (valor tipificado de la distribución normal). Este es el criterio que fija el valor de la fórmula de determinación de la muestra que veremos seguidamente, donde

simplemente substituiremos el valor de z por un 2. También se puede elegir el 95%, en ese caso el número de unidades sigma de desviación seria de $1{,}96\sigma$, es decir z = 1,96. Si quisiéramos mayor confianza, por ejemplo, del 99,7%, el número de unidades de desviación seria de $3\sigma, o\ z = 3$.

3) El error muestral e asociado al estadístico elegido de estimación. Veremos sobre todo estimaciones de medias y proporciones (o porcentajes) y en cada caso los cálculos tendrán sus especificidades.
4) La varianza (o desviación típica) de la población: σ^2. Cuando se estiman medias es la formula habitual de calculo de la varianza de una variable, y puede ser un dato disponible o estimado. Cuando se estiman proporciones el valor de la varianza de una variable, y puede ser un dato disponible o estimado. Cuando se estiman proporciones el valor de la varianza es igual a $PX(1-P) = PxQ$, es decir, la proporción (o porcentaje) P asociado a una estimación (por ejemplo, la proporción de desempleados) multiplicado por Q que es el complementario de P (la proporción de los que no están desempleados), que es $1-P$ $(100-P)$ si fueran porcentajes). (Lopez Roldan & Fachelli, 2015)

En el contexto del muestreo aleatorio simple las fórmulas de determinación del tamaño de la muestra n, teniendo en cuenta si se estima una media o una proporción, y teniendo en cuenta si se estima una media o una proporción, y teniendo en cuenta si se estudia una población finita o infinita, se presentan en la tabla 32. En las fórmulas aparecen los símbolos siguientes:

x^2: el numero de unidades de desviación que indica el nivel de confianza adoptado, elevado al cuadrado.

σ^2: la varianza de la variable cuantitativa sobre la que se calcula la media.

e^2: el error muestral considerado, elevado al cuadrado.

N: el tamaño de la población.

P: la proporción (o porcentaje) de individuos que tienen una característica.

Q: la proporción (o porcentaje) de individuos que no tienen la característica. (Lopez Roldan & Fachelli, 2015)

Tabla 31

Formulas de determinacion del tamaño de la muestra segun el tipo de poblacion y el parametro estimado, dado un error muestral

Tamaño en función del error		**Población**	
		Infinita	Finita
Parámetro	Media	$n = \frac{z^2 \times \sigma^2}{e^2}$	$n = \frac{z^2 \times \sigma^2 \times N}{(N-1) \times e^2 + z^2 \times \sigma^2}$
	Proporción	$n = \frac{z^2 \times P \times Q}{e^2}$	$n = \frac{z^2 \times P \times Q \times N}{(N-1) \times e^2 + z^2 \times P \times Q}$

Fuente: (Lopez Roldan & Fachelli, 2015)

Como se puede observar en las fórmulas, los valores del tamaño y del error están en relación inversa (en el numerador y el denominador) y son dos valores de la formula con los que se puede "jugar" hasta encontrar el equilibrio entre precisión y costes: podemos fijar el tamaño asociado. El valor de las unidades de desviación z lo fijamos por convención en niveles del 95,5%, es decir, con z=2 (o bien z=1,96).

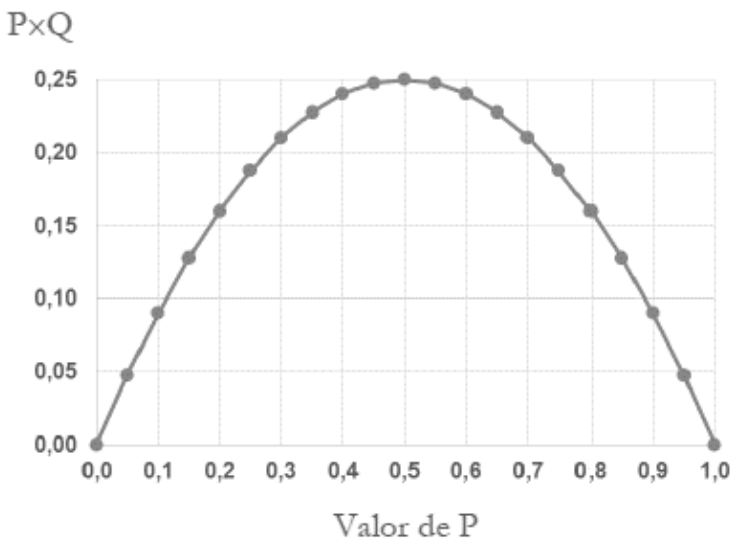
En todos los casos, finalmente, debemos especificar el valor de la varianza, ya sea σ^2 en el caso de una media o PxQ en el caso de una proporción. Pero estos valores hacen referencia a la varianza poblacional, lo que nos plantea el siguiente dilema: por un lado, estoy planteando es un estudio para conocer las características relacionadas con una determinada variable y, por otro lado, para determinar el tamaño de la muestra necesito conocer primero los datos poblacionales de la variable que quiero estudiar. En consecuencia, si disponemos de esa información para que necesitamos realizar el estudio. Bien, la cuestión merece dos comentarios. Por una parte, efectivamente, no tendría sentido; de hecho, desconocemos casi siempre el dato poblacional, lo que nos obliga a buscarlo o a estimarlo, aproximándonos a él a través de información externa como un censo, a través de investigaciones anteriores o adboc, considerando variables parecidas muy correlacionadas con la de interés, realizando conjeturas o mediante pruebas piloto. Por otra parte, aun si conociéramos el dato de la varianza de la variable, los estudios no se limitan simplemente a conocer el comportamiento de una variable, sino que se recoge información múltiple con un número elevado de variables distintas que justifican el estudio. En esos casos la variable de la que se precisa conocer la varianza actúa de variable central o sintética de la naturaleza del estudio realizado y permite diseñar la muestra teniendo en cuenta sus características.

Este dilema se plantea tanto para la estimación de medias como de proporciones. En el caso de las medias no queda más remedio que proceder como hemos indicado. Pero en el caso de las proporciones existe una solución muy practica que resuelve el problema del conocimiento previo de la varianza. La solución surge del comportamiento de los valores posibles de la varianza de una proporción, del producto de P por Q. Sabemos que este producto alcanza un valor máximo, por tanto, proporciona un error máximo. Para cualquier par P y Q, siempre se cumple que: $P\ x\ Q\ \leq$ ¼, es decir, siempre es menor o igual como máximo a 0,25, situación que se produce cuando $P = 0{,}5\ y\ Q = 0{,}5$. Este resultado se puede ilustrar gráficamente a partir del calculo de todos los posibles valores del producto de P por Q (Ilustración 25). El máximo valor de todas las combinaciones posibles de P y Q se alcanza cuando P=0,5 y Q=0,5, cuyo producto PxQ=0,25, todos los demas valores son inferiores, es decir, la varianza, y el error que comporta, siempre son inferiores. Si en lugar de proporciones utilizamos los porcentajes el valor máximo seria de 2,500, es decir, con $P = Q = 50\%$. (Lopez Roldan & Fachelli, 2015)

Ilustración 25

Determinación del valor máximo de PxQ

P	Q	P×Q
0,00	1,00	0,0000
0,05	0,95	0,0475
0,10	0,90	0,0900
0,15	0,85	0,1275
0,20	0,80	0,1600
0,25	0,75	0,1875
0,30	0,70	0,2100
0,35	0,65	0,2275
0,40	0,60	0,2400
0,45	0,55	0,2475
0,50	0,50	0,2500
0,55	0,45	0,2475
0,60	0,40	0,2400
0,65	0,35	0,2275
0,70	0,30	0,2100
0,75	0,25	0,1875
0,80	0,20	0,1600
0,85	0,15	0,1275
0,90	0,10	0,0900
0,95	0,05	0,0475
1,00	0,00	0,0000



Fuente: (Lopez Roldan & Fachelli, 2015)

Cuando se considera esta situación extrema se dice que estamos en la situación de máxima incertidumbre o de máxima indeterminación. Este supuesto es el que se asume también habitualmente en los sondeos electorales y encuestas sociológicas en el momento de determinar el tamaño de la muestra.

En las fórmulas de la Tabla 33

Tabla 32

Fórmulas de determinación del tamaño de la muestra según el tipo de población y el parámetro estimado, dado un error muestral

Tamaño en función del error		Población	
		Infinita	Finita
Parámetro	Media	$n=\frac{z^2\times\sigma^2}{e^2}$	$n=\frac{z^2\times\sigma^2\times N}{(N-1)\times e^2+z^2\times\sigma^2}$
	Proporción	$n=\frac{z^2\times P\times Q}{e^2}$	$n=\frac{z^2\times P\times Q\times N}{(N-1)\times e^2+z^2\times P\times Q}$

Fuente: (Lopez Roldan & Fachelli, 2015)

Se ha considerado el tamaño de las muestras en función del resto de parámetros de la formula. Alternativamente, y más practico, se puede calcular el error que se comete dados el resto de parámetros de la formula, fijando en particular el tamaño de la muestra y viendo el error asociado. Las fórmulas son, entonces, las de la tabla 34

Tabla 33

Fórmulas de determinación del error de la muestra según el tipo de población y el parámetro estimado, dado un tamaño muestral

Error en función del tamaño		Población	
		Infinita	Finita
Parámetro	Media	$e = z \times \sqrt{\frac{\sigma^2}{n}}$	$e = z \times \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}}$
	Proporción	$e = z \times \sqrt{\frac{P \times Q}{n}}$	$e = z \times \sqrt{\frac{P \times Q}{n} \times \frac{N-n}{N-1}}$

Fuente: (Lopez Roldan & Fachelli, 2015)

Aplicaremos lo visto para calcular el tamaño de la muestra que habitualmente se plantea en los sondeos de opinión y electorales. Un sondeo elaborado por el Instituto Opina con motivo de las elecciones al Parlamento de Cataluña de 2003, publicado en el diario El País, donde se adjuntaba la ficha técnica siguiente:

Ilustración 26

Articulo del el diario EL Pais

ELECCIONES EN CATALUÑA *Sondeo de Opina para EL PAÍS*

FICHA TÉCNICA

La encuesta ha sido realizada por el INSTITUTO OPINA. **Realización del trabajo de campo:** 31 de octubre, 1 y 2 de noviembre de 2003. **Ámbito geográfico:** Cataluña. **Recogida de la información:** mediante entrevista telefónica. **Universo de análisis:** población mayor de 18 años residente en hogares con teléfono. **Error muestral:** El margen de error para el total de la muestra es de ±2,14% para un margen de confianza del 95% y bajo el supuesto de máxima indeterminación (p=q=50%). **Procedimiento de muestreo:** selección polietápica del entrevistado: unidades primarias de muestreo (MUNICIPIOS) seleccionadas de forma aleatoria proporcional para cada provincia. unidades secundarias (HOGARES) mediante la selección aleatoria de números de teléfono. Unidades últimas (INDIVIDUOS) según cuotas cruzadas de SEXO, EDAD y RECUERDO DE VOTO GENERALES 2000.

Tamaño de la muestra: 2.100 entrevistas. 900 en la provincia de Barcelona (290 en Barcelona ciudad, 256 en el área metropolitana, 354 en el resto de la provincia). **400 en la provincia de Girona** (53 en Girona ciudad, 347 en el resto de la provincia). **400 en la provincia de Lleida** (126 en Lleida ciudad, 274 en el resto de la provincia). **400 en la provincia de Tarragona** (78 en Tarragona ciudad, 322 en el resto de la provincia).

Ponderación: el total de la muestra se ha ponderado para otorgar a cada una de las provincias su peso real dentro del conjunto de la población de Cataluña. **De 900 en la provincia de Barcelona a 1.594** (514 en Barcelona ciudad, 453 en el área metropolitana, 627 en el resto de la provincia). **De 400 en la provincia de Girona a 185** (25 en Girona ciudad, 160 en el resto de la provincia). **De 400 en la provincia de Lleida a 122** (38 en Lleida ciudad, 84 en el resto de la provincia). **De 400 en la provincia de Tarragona a 199** (39 en Tarragona ciudad, 160 en el resto de la provincia).

Fuente: (Lopez Roldan & Fachelli, 2015)

Se puede leer que se ha realizado una encuesta con una muestra de 2,100 personas mayores de 18 años en Cataluña. El margen de error ha sido del 2,14% para un nivel de confianza del 95% (z = 1,96) y bajo el supuesto de máxima indeterminación (P=Q=50%): Con estos datos, el tamaño de la muestra es el resultado de aplicar la formula del cálculo del tamaño muestral de una población infinita pues los electores superan los 100,000 individuos:

$$n = \frac{z^2 x\ P\ x\ Q}{e^2} = \frac{1{,}96^2 x 0{,}5 x 0{,}5}{0{,}0214^2} \cong 2100$$

El mismo resultado se obtiene si hacemos los cálculos en porcentajes:

$$n = \frac{z^2 x\ P\ x\ Q}{e^2} = \frac{1{,}96^2 x 50 x 50}{2{,}14^2} \cong 2100$$

Los valores obtenidos con decimales siempre se redondean al entero superior.

b) Ejercicios de cálculo del tamaño de la muestra

Ejercicio 1: En un estudio sobre el sindicato CC.OO. se requiere obtener una muestra aleatoria de la población afiliada a la que administrar un cuestionario para estudiar su perfil y la valoración de la acción sindical. ¿Cuál debe ser el tamaño de la muestra? Sabemos que a diciembre de 1998 el número de afiliados/ as era de 123,440 personas.

Tipo de población: Infinita

Nivel de error que queremos asumir: 3%

Varianza: Supuesto de máxima indeterminación, P=Q=50%

Nivel de confianza: 95,5%

Cálculo de n: $n = \frac{z^2 xPQ}{e^2} = \frac{2^2.50.50}{3^2} = \frac{2^2}{4\ .0{,}03^2} = 1111{,}11 = 1112$

Ejercicio 2: En un estudio sobre mercado de trabajo y empresa en la Región Metropolitana de Barcelona en el contexto de los Juegos Olímpicos se considera el análisis de diversas características de los centros de trabajo y su comportamiento frente al mercado de trabajo. Se requiere recoger información por encuesta en relación a una población de 24.141 empresas de más de 10 trabajadores/ as. ¿Cuánto hay que seleccionar para obtener una muestra representativa?

Tipo de población: Finita

Nivel de error que queremos asumir: 3,9%

Varianza: Supuesto de máxima incertidumbre, P=Q=50%

Nivel de confianza: 95,5%

Cálculo de n: $n = \frac{2^2.50.50.24141}{(24141-1).3{,}9^2+2^2.50.50} = 640$

Ejercicio 3: En un segmento de mercado con una población identificada de 1000 hogares y hábitos de consumo similares se quiere realizar un muestreo aleatorio simple para conocer el número medio de unidades de consumo anual de un cierto producto de consumo cultural. Decidir un nivel de error muestral y calcular el tamaño de la muestra necesaria para realizar este estudio sabiendo que según las estimaciones de otros estudios anteriores la varianza poblacional del número de unidades de consumo se sitúa sobre un valor de 100. A continuación, determinar el tamaño de la muestra si la población fuera de 200.000 hogares.

Tipo de población: Finita

Nivel de error que queremos asumir: por ejemplo, 0.5 o 1 unidad de consumo

Varianza: 100

Nivel de confianza: 95,5%

Cálculo de n: $n = \frac{z^2\sigma^2 N}{(N-1)e^2+z^2\sigma^2} = \frac{2^2 100.1000}{(1000-1)0{,}5^2+2^2 100} = 615{,}6 = 616$

$$n = \frac{z^2\sigma^2 N}{(N-1)e^2+z^2\sigma^2} = \frac{2^2 100.1000}{(1000-1)1^2+2^2 100} = 258{,}91 = 286$$

Si N=200.000, cálculo de n: $n = \frac{z^2\sigma^2}{e^2} = \frac{2^2 100}{0{,}5^2} = 1{,}600$

$$n = \frac{z^2\sigma^2}{e^2} = \frac{2^2 100}{1^2} = 400$$

Ejercicio 4: Un sondeo electoral realizo las estimaciones de intención de voto a partir de una muestra de la población de 2100 entrevistas. ¿Qué error muestral se esta considerando?

Tipo de población: Infinita

Tamaño de la muestra: 2,100

Varianza: Supuesto de máxima incertidumbre, P=Q=50%

Nivel de confianza: 95%

Cálculo de e: $e = z\sqrt{\frac{PQ}{n}} = 1{,}96\sqrt{\frac{50.50}{2100}} = 2{,}14\%$

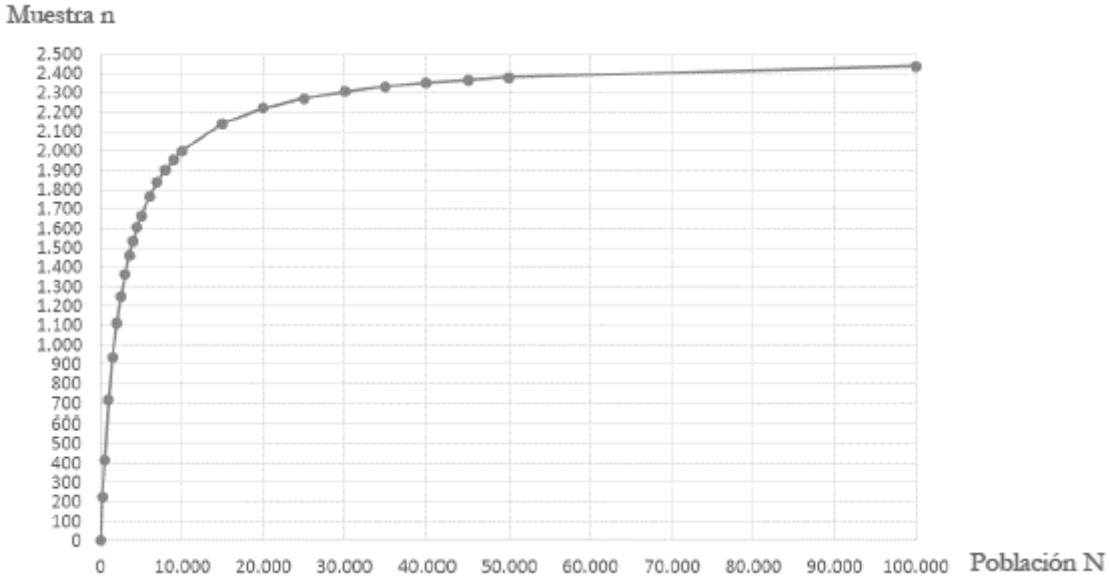
c) Relación entre el tamaño de la muestra, de la población y del error

Finalmente incluimos un apartado con un análisis ilustrativo de las relaciones que dan entre el tamaño de la muestra y el tamaño de la población, si como entre el tamaño de la muestra y el error muestral.

En el primer caso una representación gráfica del tamaño de la muestra en función del tamaño muestral (Ilustración 27) dibuja una curva que pasa por el origen de coordenadas con un comportamiento asintótico que nos pone de manifiesto una relación directa y un comportamiento creciente (a medida que aumenta el tamaño de la población necesitamos más muestras para un determinado nivel de error muestral) pero no es un comportamiento de proporcionalidad, de crecimiento lineal. Al inicio, con poblaciones pequeñas, se requiere un número importante de elementos de la muestra, y a medida que crece el tamaño poblacional los elementos muestrales requeridos crecen notablemente, pero a partir de un determinado N los incrementos son cada vez más reducidos, hasta llegar a un momento en que los aumentos de N no producen aumentos en el tamaño de la muestra. A partir de entonces la población es infinita. (Lopez Roldan & Fachelli, 2015)

Ilustración 27

Relación entre el tamaño de la muestra y el tamaño de la población Poblaciones finitas, e=2%, z=2, P=Q=50%



Fuente: (Lopez Roldan & Fachelli, 2015)

En particular, al pasar de una población 1 a 1000 la muestra necesaria pasa de 1 a 714, mientras que de 1000 a 2000 tan solo necesitamos aumentar la muestra en 397, y de 2000 a 3000, el aumento es de 253, progresivamente la necesidad de nuevos elementos en la muestra disminuye hasta estabilizarse a partir de 100.000 individuos.

Un comportamiento similar se da entre el error muestral y el tamaño de la muestra en el sentido de que tampoco se trata de una relación proporcional lineal (Ilustración 28). Error y tamaño tienen una relación inversa, a medida que aumenta el tamaño muestral se reduce el error. Pero el tamaño n es inversamente proporcional al cuadrado del error de muestreo, lo que implica que cualquier pequeño incremento de unidades en la muestra, cuando estas son pequeñas, producen una reducción muy importante del error cometido, pero llega un momento en que las reducciones son cada vez más pequeñas de modo que reducciones mínimas del error suponen incrementos muy elevados del tamaño de la muestra. (Lopez Roldan & Fachelli, 2015)

Ilustración 28

Relación entre el tamaño de la muestra y el margen de error. Población infinita, z=2, P=Q=50%

Fuente: (Lopez Roldan & Fachelli, 2015)

En particular, al pasar de un tamaño de 1 a 100 se reduce el error del 100% al 10%, una reducción de 90 puntos porcentuales, pero al pasar de 100 a 200, la reducción es tan solo del 3%.

d) El error asociado a una estimación

Para finalizar este apartado apuntamos someramente uno de los aspectos que permanentemente se plantea en el análisis de los datos derivados de una muestra estadística: el error que se deriva de una estimación puntual en la muestra obtenida. Como hemos destacado anteriormente, toda afirmación estadística expresada, por ejemplo, en términos de un porcentaje o de una media responde al aspecto descriptivo de los resultados de un estudio cuantitativo por muestreo. Este dato se acompaña de un elemento adicional en términos inferenciales para dar cuenta del error asociado al estadístico en cuestión, así como para plantear distintas cuestiones sobre su significación. Seria el caso, por ejemplo, de estar interesados en conocer la media de los ingresos de una población y de plantearse si esa media es igual entre varones o mujeres, o bien difiere significativamente entre ambos sexos. Una prueba de hipótesis estadística establecerá, basándose en el error asociado a esas estimaciones, si podemos concluir la igualdad o la significación de la diferencia.

Pero consideremos el caso sencillo de estimar una sola media o un solo porcentaje, y planteemos la cuestión del error asociado a estas estimaciones. Por ejemplo, supongamos desconocida la media de ingresos μ de la población de un municipio. Una estimación de este valor se puede obtener tomando una muestra aleatoria de sus habitantes. Pongamos que obtenemos una muestra de n=1.000 personas a las que les preguntamos por sus ingresos, y con las 1000 respuestas u observaciones calculamos la media de los ingresos, y con las 1000 respuestas u observaciones calculamos la media de los ingresos. Supongamos que los ingresos medios obtenidos fueran x = 1.500ϵ, esta sería la

estimación puntual de la media poblacional. Tengamos presente que en otra posible muestra de 1000 personas hubiera dado unos ingresos medios diferentes, por ejemplo, de 1600 ϵ. De hecho, cada posible muestra generaría puntualmente un resultado distinto, lo que nos esta indicado la existencia de una variabilidad y de que estamos cometiendo un error de estimación. Este es un aspecto crucial que nos conducirá a hablar de estimaciones por intervalos teniendo en cuenta el error que se comete. Por el hecho de disponer de una parte de la población, aunque sea representativa e incluso numerosa, cada estimación del valor verdadero (y desconocido) del parámetro poblacional a partir del estadístico muestral tiene un error respecto del valor real y desconocido del parámetro. Nuestro resultado que puede estar más cercano (si tenemos buena suerte con la muestra) o menos cercano (si tenemos mala suerte), pero difícilmente coincidirá con el verdadero valor de la media. El margen de error e es la diferencia máxima entre la estimación puntual obtenida en la muestra y el valor verdadero del parámetro poblacional y nos mide el grado de exactitud o de precisión con el que inferimos de la muestra a la población. Este valor vendrá determinado por la variabilidad del estadístico, es decir, que el error se cuantifica mediante las varianzas del estadístico considerado en cada caso. Como veremos esta variabilidad se denomina error típico del estadístico, y se corresponde con un cálculo que dependen de la varianza y del tamaño de la muestra, además de otra característica como el nivel de confianza, de forma similar a como hemos visto en el cálculo del tamaño de la muestra. (Lopez Roldan & Fachelli, 2015)

Pero el error muestral se define simplemente como la divergencia entre los valores de los estadísticos obtenidos en la muestra de una variable y los existentes en la población. En el caso del estadístico de la media, el error muestral seria la diferencia entre el valor muestral de la media, X, y el valor poblacional de la media, μ, es decir, x – μ. Esta es una definición de lo real pero no conocido, ya que el parámetro poblacional es desconocido y es lo que se quiere saber. Para solucionar este interrogante, razonaremos en términos probabilísticos a partir del conocimiento de la forma de la distribución del estadístico. En este sentido, cualquier afirmación para dar cuenta del valor de un parámetro poblacional desconocido a partir del valor del estadístico que lo estima, tendrá un grado de desviación, de variación al alza o de variación a la baja, que es el que mide el error de muestreo. De aquí se deriva una relación básica del muestreo y de las estimaciones estadísticas:

$$\frac{Valor\ poblacional}{(Parametro)} = \frac{Valor\ estimado\ en\ la\ muestra}{(Estadistico)} \pm Error\ de\ muestreo$$

De esta forma, cada estadístico (x, p, etc.) debemos entenderlo como si fuera una variable, en el sentido de que puede tomar varios valores según la muestra que consideramos (de todas las muestras posibles que podemos tomar de la población). Así, todos los posibles valores que estiman el único valor verdadero del parámetro forman lo que se llama una distribución muestral del estadístico. En el caso de una media y una proporción es distribución muestral teórica es la normal, la cual nos ayudara a determinar, con una certeza suficiente, el margen de error que cometemos al hacer estimaciones.

Si, como ejemplo, obtenemos, tras preguntar a una muestra de 1000 personas, n = 1000, que la intención de voto a un determinado partido en las próximas elecciones es del 30% de los votos, p=30, siendo q=100-p=70, el error asociado se calcula a partir del producto

del valor del a distribución normal z y el error típico de la estimación de la proporción Sp, es decir:

$$e = zxSp$$

Asumiendo, como es habitual, un nivel de confianza del 95,5%, es decir, Z=2, y siendo el error típico de la estimación es igual a $\sqrt{pxq/n}$, por tanto, igual a $\sqrt{30x70/1000} = 1{,}45$, se obtiene que el error asociado a la estimación es e = z x Sp = 2 x 1,45 = 2,9 , es decir, del 2,9%.

Cuando a la estimación puntual le añadimos el margen de error, sumando y restando el error de muestreo, realizamos la estimación por intervalos de confianza. Es decir, que el valor poblacional se situara en un intervalo definido por:

$$30\% \pm 2{,}9\%$$

Es decir, en un intervalo entre el

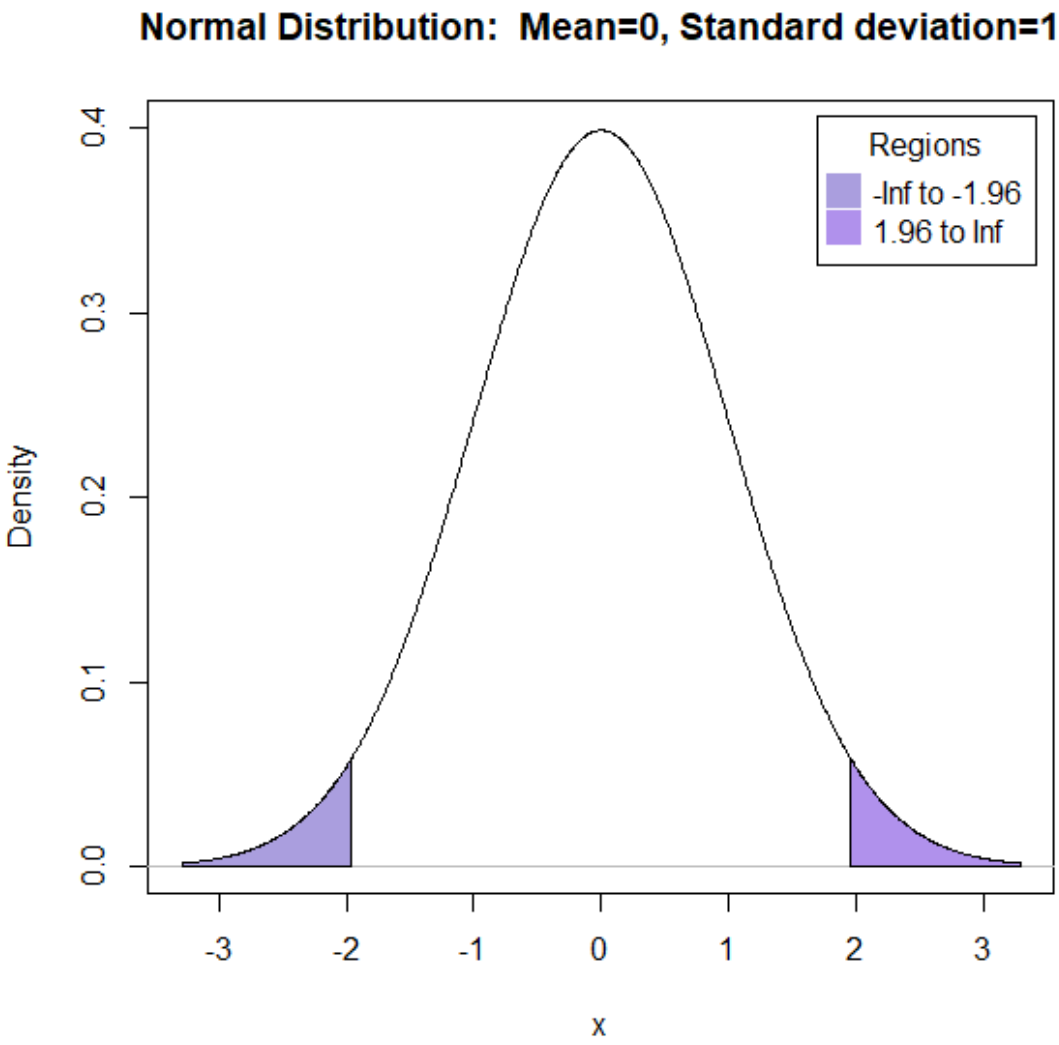
$$27{,}1\% \; y \; el \; 32{,}9\%$$

Y que escribimos así (27,1%,32,9%)

Cualquier valor puntual estimado es una aproximación al valor poblacional desconocido afectado por el error muestral, por lo tanto, esto no significa que el porcentaje sea estrictamente del 30% en nuestro ejemplo, sino que solo podemos establecer un intervalo de valores, entre un mínimo y un máximo, en cuyo interior se situara, probablemente el parámetro poblacional. Y decimos probablemente porque el intervalo se establece con un grado de probabilidad o nivel de confianza, del 95%. De esta forma funciona el razonamiento estadístico en todos los casos a partir de muestras aleatorias representativas de una población. (Lopez Roldan & Fachelli, 2015)

Hallar el tamaño de la muestra para una población N=120000, con un nivel de confianza del 95%, error de estimación de la muestra de el 5%.

Nivel de significancia es igual 0.05 (5%)



Hallar el tamaño de la muestra para una población N=180000, con un nivel de confianza del 98%, error de estimación de la muestra del 6%.

3.4.3. Muestreo Estratificado

El muestreo aleatorio estratificado (MAE) es un método de muestreo que tiene la gran ventaja de permitir mejorar la precisión de las estimaciones en relación al muestreo aleatorio simple, es decir, disminuye el error muestral y de las estimaciones para un mismo tamaño de muestra, o bien reduce el tamaño para un mismo margen de error.

En el MAE se parte de la consideración de que la población no es homogénea según los objetivos de la investigación, y se trata de dividir a la población en categorías o grupos que tienen un interés analítico en función de una o más características o variables criterio de la población que definen la heterogeneidad. Estos grupos son las subpoblaciones que definen los estratos. Desde el punto de vista de la investigación resulta fundamental que estas características se respeten de forma mas o menos estricta en la muestra y asi garantizar su representatividad en términos muestrales, a la vez que, con esta forma de

proceder, como hemos dicho, se consigue una ganancia en la reducción de la variabilidad y del error de las estimaciones que se hacen para toda la población.

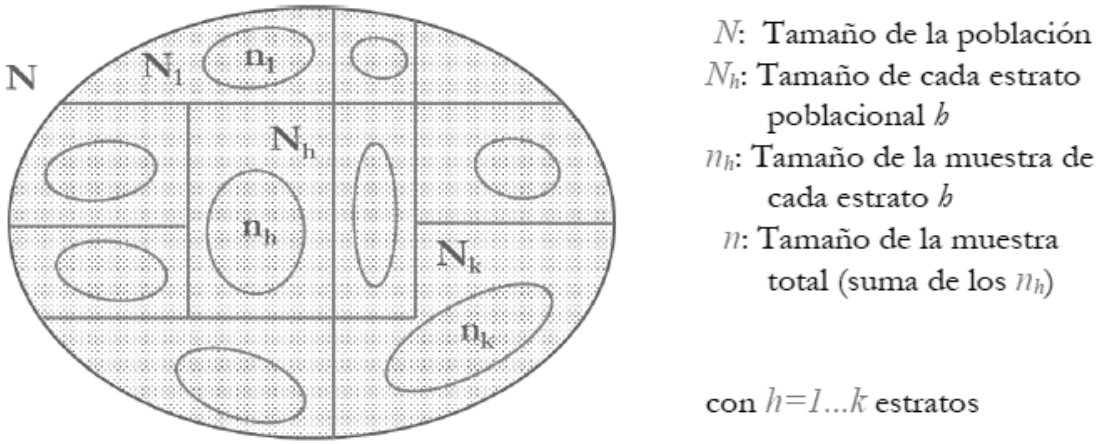
Este tipo de muestreo, por otro lado, nos permite plantear y nos facilita la necesidad de efectuar estudios o estimaciones de características poblacionales a partir de la definición de subpoblaciones que se hacen corresponder con los estratos. Estas subpoblaciones objeto de estudio específico resultan representativas en sus respectivas submuestras con un grado de error aceptable que el MAE favorece, siempre que el tamaño de la muestra sea suficientemente grande. En todo caso, el número de muestra necesario siempre será menor, para el mismo nivel de error muestral, que en el muestreo aleatorio simple. En consecuencia, se trata de un método que reporta un menor coste y es más eficiente también en este sentido.

Pero, la cuestión crucial en este tipo de muestreo, que limita en la práctica su aplicación generalizada, es la exigencia de disponer de la información de las variables criterio para todas las unidades de las unidades de la población para clasificarlas en estratos poblacionales. El método estratificado implica, por un lado, definir y construir los estratos poblacionales h donde se agrupan todas las unidades (un total de k estratos) de forma que todas sean asignadas a un estrato (criterio de exhaustividad, de forma que la unión de todos los estratos es igual a al número de unidades de la población) y cada una pertenezca a un solo estrato (criterio de exclusividad, la intersección entre dos estratos cualesquiera es nula). A continuación, se extrae una muestra aleatoria independiente en cada uno de estos estratos (Nh). El número de unidades que se deben extraer de cada estrato viene determinado por el criterio de reparto que se aplique en función de un tamaño de muestra que es fijada, operación llamada de afijación de la muestra. Posteriormente, reunimos toda la información para obtener estimaciones globales de toda la población.

En la Ilustración 29 adjunto se representa la forma de proceder de este método de muestreo. (Lopez Roldan & Fachelli, 2015)

Ilustración 29

Representación del muestreo aleatorio estratificado



Fuente: (Lopez Roldan & Fachelli, 2015)

El tipo de muestreo que se realiza en cada estrato habitualmente es el mismo, y es característico proceder a obtener una muestra aleatoria simple dentro de cada estrato. Pero también puede ser diferente, utilizando diferentes estrategias en cada estrato para optimizar los recursos económicos, por razones derivadas de las características particulares que definen el estrato o simplemente por la imposibilidad de poder obtener un muestreo aleatorio en alguno de ellos.

La estratificación implica pues conocer unas características poblacionales para cada unidad que son objeto de interés central para la investigación. Pero suele suceder que estas características de interés directo no son conocidas o no están disponibles. Y si se conocieran la pregunta es inmediata, ¿Por qué hacer una muestra? Como señalamos en un momento anterior, habitualmente los estudios no se centran en la necesidad de obtener observaciones de una sola característica, a pesar de que haya una que sea de interés central de la investigación. En estos casos en los no disponemos de la variable principal de estratificación la solución es recurrir a otro tipo de información para estratificar siempre que sean determinantes del fenómeno estudiado y consecuentemente tengan una elevada correlación con los criterios de estratificación deseados.

Las variables que actúan como criterio clasificatorio se pueden reducir a una única variable o pueden considerarse varias de ellas simultáneamente. Veamos algunos ejemplos de variables de estratificación. Es habitual estratificar unidades que son personas mediante el sexo, la edad, la categoría socioeconómica, el número de habitantes del municipio donde residen (tamaño del hábitat). Cuando se hacen estudios de familias es posible estratificar según el tamaño de los hogares. Si estudiamos la población universitaria un criterio de estratificación es considerar la facultad en la que estudian. Si analizamos el comportamiento de los consumidores es recomendable dividir la población en segmentos de mercado. Cuando se consideran estudios de empresas es frecuente emplear como criterio de estratificación el tamaño de la empresa, utilizando como indicador o bien el número de trabajadores o bien el nivel de facturación, de la misma forma que se podría utilizar el sector de actividad, o una combinación de sector y tamaño de la empresa. Si se quiere hacer un estudio de la sanidad, a partir de los hospitales, se estratifica por el número de pacientes. Si la investigación es educativa y consideran las escuelas un criterio es considerar el número de alumnos del centro. (Lopez Roldan & Fachelli, 2015)

El método de estratificación comprende dos etapas principales: la determinación de los estratos y la afijación de la muestra.

1) La determinación de los estratos de la población a partir de una o varias características conocidas a priori de esta (corte de la muestra) implica considerar finalmente una variable clasificatoria categórica, de tipo nominal u ordinal. La obtención de esta clasificación cuando se hace a partir de un único criterio clasificatorio es inmediata, hay que considerar directamente sus categorías (distritos del municipio, comunidad autónoma, sexo,…) o proceder a una agrupación de sus valores (números de trabajadores en intervalos, grupos de edad, niveles de estudios, etc.). Si la clasificación tiene en cuenta varios criterios clasificatorios simultáneamente (usamos por tanto una tipología con diferentes tipos que definen los estratos) se puede generar por varios procedimientos. Lo más

sencillo es considerar el cruce de las diversas variables y considerar tantos estratos como combinaciones de valores resulten. Alternativamente, cuando el número de variables es elevado, se puede utilizar alguna técnica de análisis multivariable clasificatoria que nos ayude a generar una tipología de la población en estratos.

El número idóneo de estratos no se puede determinar apriorísticamente, depende de varios factores donde hay que considerar las particularidades de cada diseño y los objetivos de investigación. Si consideramos el procedimiento de combinación de valores de diferentes variables, el número de estratos se multiplica rápidamente cuando el número de variables aumenta y sus valores son numerosos. No obstante, por estudios empíricos y teóricos se aconseja multiplicar el número de estratos, hasta cierto límite en que nuevos estratos no reportan una ganancia significativa de precisión. La consideración de una diversidad de estratos tiene que ver y se justifica con el comportamiento de la variabilidad de los estratos.

Cuando el objetivo de la estratificación es considerar grupos de unidades homogéneas para dar cuenta de la heterogeneidad del fenómeno estudiado, lo que estamos diciendo es que se consideran grupos en el interior de los cuales la variabilidad es reducida, y la fuente de variación se da entre los diferentes grupos. Una variabilidad pequeña en el interior de un estrato se traduce inmediatamente en una reducción del tamaño de muestra necesaria; al límite, si todas las unidades fueran idénticas, la variabilidad interna seria nula y solo nos haría falta una unidad para dar cuenta de las características del estrato. Llevando este razonamiento al extremo se trataría de considerar tantos estratos como unidades tiene la población, pero evidentemente ya no estaríamos hablando de muestreo sino de un censo. Pero el razonamiento nos sugiere la conveniencia de considerar una diversidad de estratos ya que de esta forma conseguimos reducir la variabilidad interna.

Este razonamiento es bien conocido en estadística, y se expresa en la descomposición de la varianza total de una variable (cantidad constante) en dos partes, la varianza interna o intra grupos y la variabilidad externa o entre grupos. El análisis de varianza se basa en esta distinción y algunos métodos de clasificación automática hacen uso igualmente de estos conceptos. Cuando se consigue que la variable o variables de estratificación hagan mínima la variabilidad intra grupos (se formen estratos homogéneos), lo que significa al mismo que tiempo hacer máxima la variabilidad entre grupos (los grupos son heterogéneos entre ellos), el resultado es una ganancia en eficiencia del muestreo.

2) La afijación de la muestra consiste en repartir un tamaño de muestra determinada previamente entre los diferentes estratos. Con la cuota de muestra que corresponde a cada estrato se procede a efectuar la extracción aleatoria de la muestra del estrato, a través de un muestreo aleatorio simple, por ejemplo. La afijación de la muestra se puede operar de tres formas diferentes:
    a. Afijación igual. Simplemente asignamos la misma muestra a cada uno de los estratos: $nh = \frac{n}{k}$. Pero suele ser un criterio poco eficiente.
    b. Afijación proporcional. En este caso la afijación distribuye el tamaño de muestra total n proporcionalmente en función del tamaño poblacional de cada estrato: $nh = \frac{Nh}{N}.n$, de forma que cuando mayor sea el extracto más cuota de muestra le toca y logrando así que cada unidad de la muestra represente el mismo número de unidades de la población. Cuando cada

unidad de la muestra tiene el mismo peso y representa al mismo número de unidades de la población la fracción de muestreo es la misma en todos los estratos, $\frac{nh}{N} = \frac{n}{N}$, y de la muestra se dice que es auto ponderada; de esta manera el cálculo, por ejemplo, de una media se puede hacer a partir del conjunto de unidades de todos los estratos.

c. Afijación no proporcional u optima. La distribución de la muestra tiene en cuenta además del tamaño del estrato, la variabilidad de este. La afijación llamada optima de Neyman, considera diferentes fracciones de muestreo según este doble criterio que se expresa mediante las fórmulas:

- Para la estimación de medias: $nh = \frac{Nh\ x\ \sigma h}{\sum_{h=1}^{k} Nh\ x\ \sigma h}\ x\ n$
- Para la estimación de proporciones $nh = \frac{Nh\ x\ \sqrt{Ph\ x\ Qh}}{\sum_{h=1}^{k} Nh\ x\ \sqrt{Ph\ x\ Qh}}\ x\ n$

Así cuanto mayor sea el estrato mayor cuota de muestra le corresponde, y cuando más variable (heterogéneo) internamente sea el estrato mayor cuota de muestra necesita también. Este doble criterio plantea alterar la estricta proporcionalidad, lo que lleva a considerar la ponderación posterior para restablecerla. De este modo se consigue, primero, la presencia en la muestra de fenómenos menos frecuentes, más atípicos y variables, y, segundo, devolver su peso específico en términos de proporción poblacional. La afijación optima puede basarse también en un criterio adicional de coste, minimizando este para una varianza dada o minimizando la varianza para un coste dado. El coste, por otro lado, puede ser igual para cada estrato o con costes desiguales para cada estrato. En el caso de la afijación optima de Neyman se supone que los costes de los estratos son aproximadamente iguales.

El muestreo estratificado siempre supone una ganancia de precisión sobre el MAS, mejor cuanto mayor sea la diferencia entre los estratos, más correlación se dé entre las variables estudiadas y la variable de estratificación y mejor sea la información o la medida de los estratos. Por otra parte, se demuestra que el muestreo estratificado con afijación optima es mas eficiente que el muestreo estratificado con afijación proporcional, las estimaciones están afectadas de un menor error, y su eficiencia será mayor cuando las diferencias de las estimaciones de los estadísticos entre los estratos sean más grandes.

En esta exposición hemos considerado siempre el caso de la estratificación a priori, pero también existe la posibilidad de operar estratificaciones a posteriori, por ejemplo, a partir de la extracción de una muestra aleatoria simple y estratificando según determinantes criterios poblacionales aplicados a la muestra. Aunque de hecho se trata más bien de un ajuste o equilibrio de la muestra hecha después de su obtención. (Lopez Roldan & Fachelli, 2015)

3.4.4. Muestreo por conglomerados

El muestreo por conglomerados (MC), llamado también muestreo de cluster, de conjuntos, de racimos o de grapas. A diferencia del muestreo aleatorio simple o del estratificado, en este tipo de muestreo las unidades no son simples sino compuestas o complejas (con un tamaño mayor de 1), es decir, se clasifica a la población en grupos, llamados conglomerados, cada uno de los cuales incluye a la vez otras unidades más simples o desagregados. Por ejemplo, las unidades agregadas o conglomerados (unidades

de muestreo primarias) podrían ser las escuelas de educación primaria de un territorio, y las unidades mas simples de cada conglomerado (unidades de muestreo secundarias) el profesorado de estos centros. La estrategia de muestreo consiste en hacer una extracción aleatoria inicial de algunas de las unidades primarias o conglomerados (mediante un muestreo aleatorio simple u otro tipo de diseño) y luego tomar todas las unidades secundarias de cada conglomerado escogido. De esta forma, los elementos individuales de la población solo pueden participar en la muestra si pertenecen a un conglomerado incluido en la muestra. La unidad de muestreo primaria no es la unidad de observación, sino la secundaria, y hay que tener presente el tamaño de ambos tipos de unidades. (Lopez Roldan & Fachelli, 2015)

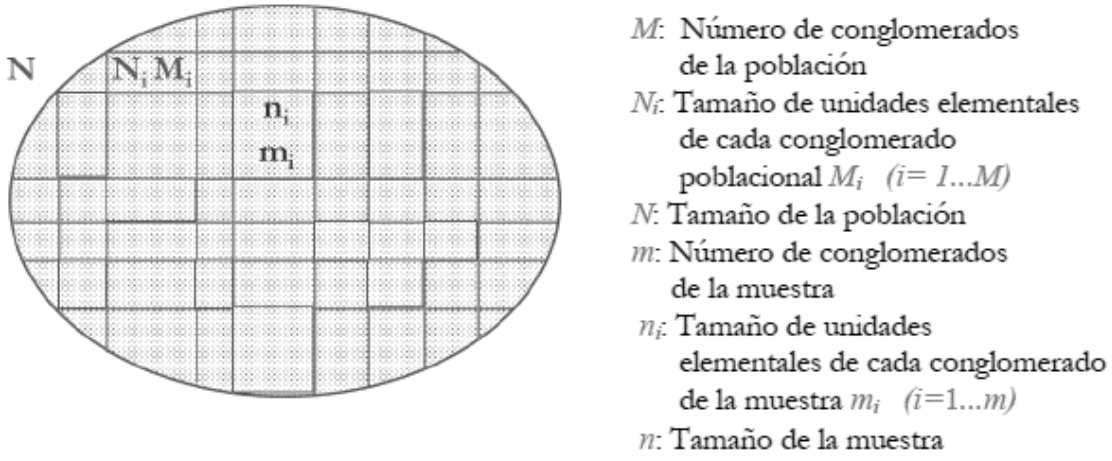
Si esquematizamos el procedimiento del muestreo por conglomerados, se trata de:

1) Dividir a la población N en M conglomerados, con Ni elementos cada uno. Cada uno de los Mi conglomerados pueden ser iguales o de diferente tamaño.
2) Se escogen aleatoriamente m conglomerados de M. La selección se puede hacer con idéntica probabilidad para cada conglomerado o con probabilidad proporcional a su tamaño.
3) De cada conglomerado elegido Mi se toman todos los elementos que lo forman, ni.

Gráficamente se podría representar de la forma siguiente:

Ilustración 30

Representación del muestreo por conglomerados



Fuente: (Lopez Roldan & Fachelli, 2015)

Frente al muestreo aleatorio estratificado que pretende establecer también grupos poblacionales (estratos) homogéneos, en el muestreo por conglomerados lo que se busca es precisamente lo contrario, constituir grupos o unidades compuestas que sean lo mas heterogéneas internamente y lo más homogéneas entre ellas. Porque lo que se pretende idealmente es que cada conglomerado sea una representación, a escala reducida, de todo el universo considerado; cada conglomerado actúa como un pequeño universo poblacional con toda la diversidad de lo que se estudia, y con la elección de unos cuantos de ellos representamos toda la población con la gran ventaja de concentrar todas las

unidades más simples de la muestra en estos conglomerados. Cuando esta concentración se realiza sobre la base de conglomerados que corresponden a divisiones (administrativas) del territorio es evidente que la elección de unos cuantos conglomerados reduce el número de desplazamiento en el espacio, reduciendo así los costes del muestreo.

De hecho, esta es la clave principal del recurso a esta estrategia de muestreo. La dispersión geográfica de las unidades en el territorio eleva notablemente los gastos de la investigación en tiempo y recursos, de forma que la distribución de la población en áreas geográficas de las que se escogen unas cuantas disminuye la necesidad de desplazamientos. Por este motivo también se identifica y designa al muestreo por conglomerados como muestreo por áreas.

Pero esta no es la única razón que justifica el uso de este método. La necesidad de disponer de un marco de muestreo de todas las unidades poblacionales se encuentra a menudo con las dificultades prácticas para su obtención o confección, lo que impide hacer viable el planteamiento de un muestreo estrictamente aleatorio. Con el muestreo por conglomerados lo que se consigue es reducir el problema de elaborar un listado de las unidades poblacionales para solamente aquellos conglomerados que se han seleccionado inicialmente. Por tanto, no se necesita un marco muy específico de las unidades secundarias, y la constitución de un marco inicial de conglomerados o unidades primarias es económico y fácil de conseguir.

Sin embargo, la estrategia del muestreo por conglomerados resulta de las más ineficientes o imprecisas. En general para el mismo tamaño de muestra es menos eficiente que el MAS. Siempre será más representativa y precisa una muestra aleatoria simple del profesorado de educación primaria extraída del conjunto de docentes que la muestra del mismo tamaño extraída a partir de una selección previa de colegios a partir de los cuales se toma en todo su profesorado. En este sentido se debería compensar el mayor error en las estimaciones con un aumento del tamaño muestral.

La eficiencia del muestreo por conglomerados aumenta cuando aumenta la heterogeneidad interna de los conglomerados y, en consecuencia, aumenta la homogeneidad entre ellos. Pero conseguir este grado de semejanza entre los conglomerados para representar fielmente las características de la población no es fácil ni habitual. También aumenta la eficiencia cuando el tamaño de los conglomerados disminuye, por lo que es conveniente escoger conglomerados de tamaño no muy grande. Es recomendable igualmente que el tamaño de los conglomerados sea similar ya que la precisión de las estimaciones aumenta.

Por otra parte, como inconvenientes adicionales, hay que señalar que exige tratamientos estadísticos complejos y que pierde parte de su carácter probabilístico dado que los conglomerados no seleccionados determinan una probabilidad 0 de las unidades finales de pertenecer a la muestra y de 1 de los seleccionados.

La situación que hemos presentado aquí hasta ahora corresponde a lo que se denomina muestreo monoetapico de conglomerados o de conglomerados en una etapa. Pero también se puede plantear un muestreo en sucesivas etapas (escalones) en la que se consideran unidades terciarias, cuartas, etc. En la primera etapa se realiza una selección aleatoria de los conglomerados considerados como unidades primarias, después cada conglomerado

seleccionado se divide también en conglomerados como unidades secundarias. Si tomamos todas las unidades terciarias de estos conglomerados tendremos un muestreo por conglomerados bietapico, también llamado muestreo por conglomerados con submuestreo. Si consideramos más conglomerados el muestreo será progresivamente de tres o más etapas, denominándose muestreo polietápico por conglomerados. En cada extracción aleatoria se pueden seguir varias estrategias de muestreo.

Como ejemplos de conglomerados podríamos considerar cualquier división territorial que, partiendo de la unidad elemental final, el individuo, considerara sucesivamente unidades agregadas o conglomerados en términos de hogares, manzanas, secciones censales, barrios, distritos, municipios, comarcas, provincias, regiones. Las urnas electorales representan igualmente conglomerados del comportamiento del voto. Los departamentos universitarios son unidades agregadas de su profesorado que se pueden considerar en un muestreo por conglomerados. En un estudio para medir la satisfacción de los usuarios / as del transporte público se pueden considerar los como conglomerados. Los centros penitenciarios, los asilos, las escuelas, los centros hospitalarios, los museos, etc. Suelen ser consideradas como unidades primarias naturales, es decir, un tipo de unidad agregada propia de la organización social, política, económica o cultural de una sociedad.

En relación a la determinación del número de conglomerados y su tamaño se pueden hacer las siguientes observaciones como condiciones de eficiencia. Por un lado, el número de conglomerados debe ser suficientemente grande para que actúe la ley de los grandes números y se produzcan compensaciones entre los conglomerados extraídos. Por otro, los conglomerados deben tener un tamaño mareado pero suficiente, es decir, deben garantizar la suficiente variabilidad (o la máxima heterogeneidad) en su interior sin constituir conglomerados excesivamente grandes. Además, el tamaño de los conglomerados debe ser lo mas parecido posible. (Lopez Roldan & Fachelli, 2015)

3.4.5. Muestreo no probabilístico

Hay dos tipos básicos de muestras, o de selecciones muestrales: el muestreo probabilístico, lo que hemos tenido ocasión de ver hasta ahora, y el muestreo no probabilístico o no aleatorio.

En todos los métodos destinados a obtener una muestra aleatoria, los elementos de la población se seleccionan de tal manera que cada uno tiene una probabilidad no nula y conocida de ser incluido en la muestra. La muestra de la población pues se debe seleccionar por métodos aleatorios que aseguren la extracción independientemente de cualquier juicio subjetivo, o no más allá de los criterios definidos en la investigación que orientan el diseño de la muestra. Esto significa que podemos elaborar conclusiones de inferencia estadística desde los datos de la muestra en los datos de la población. Siempre que se planteen estudios extensivos, en particular por encuesta, la muestra aleatoria es la más conveniente y la más representativa de la población y en consecuencia es la que menos sesgos comportara. En las muestras perfectamente aleatorias las diferencias entre los datos de la muestra y los atribuibles a la población son únicamente debidos a los errores muestrales, siempre que los errores de medida y otros errores sistemáticos no introduzcan un grado de invalidez y pese a que una muestra perfectamente aleatoria es un ideal difícilmente alcanzable.

Por su parte, las muestras no probabilísticas se seleccionan en base a la apreciación de los investigadores/as en función de determinados objetivos analíticos propios y particulares. En ellas algunas unidades de la base de sondeo tienen una probabilidad diferente y desconocida de salir a la muestra en relación a otras unidades. Por tanto, las muestras no probabilísticas se fundamentan en el criterio de selección del propio investigador/a según los objetivos de la investigación y con un juicio y decisiones objetivadas que juega una función clave para determinar que unidades han formar parte de la muestra. Este criterio puede estar fundamentado sobre la conveniencia y la facilidad, aunque justificado, o sobre la base de una norma sistemáticamente empleada.

De hecho, las muestras no probabilísticas se utilizan en muchas investigaciones, ya sean extensivas y para producir información de tipo cuantitativo como en estudios de orientación cualitativa. La estrategia no probabilística se plantea en particular:

- Cuando un muestreo probabilístico es prohibitivo económicamente.
- Cuando es difícil de implementar, en particular, si no se dispone de un marco de muestreo adecuado por varias razones: la población a analizar esta muy dispersa, o tiene (se le ha impuesto) algún estigma social que no permite su libre identificación: drogadictos, enfermos de SIDA, prostitución, alcohólicos, criminales, inmigrantes, elites económicas y sociales, etc.
- Cuando no son necesarias representaciones precisas: en los primeros estadios de la investigación para explotar determinados conceptos, testear el cuestionario, etc., donde no se está interesado en extender la representatividad a toda la población, como por ejemplo en ensayos, para la construcción de escalas, para generar hipótesis, en análisis exploratorios, para construir modelos, etc.
- Cuando corresponde a un diseño de análisis cualitativo en el que la selección de las unidades informantes responde a un criterio intensivo y calificado o analítico para dar cuenta en profundidad de un rasgo, vivencia o dinámica social. (Lopez Roldan & Fachelli, 2015)

3.4.6. Muestreo por cuotas

Es un tipo de muestreo muy generalizado en las investigaciones aplicadas por encuentra porque ha demostrado su eficiencia en relación a los diseños más cuidadosos de las muestras probabilísticas, y de hecho se utiliza a menudo como si se tratara de un método probabilísticos. Pero en la medida en que no parte de una extracción aleatoria sobre la base de un marco de muestreo no puede equiparse. En su lugar, esta estrategia de muestreo deja finalmente la elección de la unidad muestra al criterio de los encuestadores/as a partir de determinadas indicadores o perfiles que debe respetar, siguiendo, en particular, estrategias pseudoaleatorias con las llamadas rutas aleatorias.

El procedimiento es sencillo y se puede resumir en las siguientes tareas básicas:

1) Elección de las variables criterio y de sus valores correspondientes para establecer las cuotas poblacionales y de muestreo.
2) Cruce de las variables con sus valores y determinación, a partir de la tabla correspondiente, de las cuotas que corresponde a cada casilla.

3) Atribución a cada casilla o cuota del peso que le corresponde en términos de la proporción que estas características tienen en la población total.
4) Atribución de la cuota o números de elementos de la muestra correspondiente a cada perfil.
5) Orientaciones e instrucciones a los encuestadores/as para elegir convenientemente a los entrevistados/as.

En primer lugar, pues, se divide la población en subgrupos según los valores de las variables que se consideran son más influyentes en el modelo de análisis que se lleva a cabo. Estas variables suelen ser las denominadas estructurales, variables independientes básicas (sociodemográficas) relacionadas con los objetivos de la investigación como, por ejemplo, el sexo para definir cuotas de hombres y mujeres, la edad para determinar subgrupos de personas más jóvenes o mayores, el nivel educativo alcanzado para configurar cuotas educativo-culturales, etc. Evidentemente el número de variables a retener y el de los valores de cada una de ellas no debe ser elevado ya que la recogida de datos se complicaría enormemente. Si consideramos varias de estas variables se configura una matriz con forma de tabla de contingencia donde cada casilla, resultando del cruce de los valores de estas variables, configura una cuota de características de la población (conocidas habitualmente a partir de datos censales) cuya proporción se debe respetar en los elementos seleccionados finalmente en la muestra seleccionada. Si todos los elementos están bien ponderados la muestra debería ser una razonable representación de la población.

Imaginemos que queremos realizar un estudio utilizando el muestreo por cuotas teniendo en cuenta el nivel de estudios, la edad y el sexo de la población ocupada española. En la Tabla 35 aparecen las cuotas poblacionales que deberán respetarse en la extracción de la muestra. Contiene los porcentajes sobre el total de la población ocupada de 16 y mas año, 17,514,455 personas según los datos del Censo de Población de 2011, de cada combinación de categorías de las tres variables. (Lopez Roldan & Fachelli, 2015)

Tabla 34

Cuotas poblacionales según Estudios, Edad y Sexo. Porcentaje sobre el total de la población de 15 y mas año

Cuotas poblacionales (%)		Estudios		
		Primarios	Secundarios	Terciarios
Varones	16-29	0,93	2,79	2,66
	31-49	5,18	20,98	8,33
	50 o más	1,54	8,87	3,40
Mujeres	16-29	0,43	1,61	2,20
	31-49	4,25	14,60	6,01
	50 o más	2,65	10,63	2,93

Fuente: (Lopez Roldan & Fachelli, 2015)

La tabla 36, aplica esas cuotas a un tamaño de muestra, en este caso sobre 1000 individuos.

Tabla 35

Cuotas muestrales según Estudios, Edad y Sexo

Cuotas muestrales (encuestas)		Estudios			Total
		Primarios	Secundarios	Terciarios	
Varones	16-29	9	28	27	**64**
	31-49	52	210	83	**345**
	50 o más	15	89	34	**138**
Mujeres	16-29	4	16	22	**42**
	31-49	42	146	60	**249**
	50 o más	26	106	29	**162**
Total		**142**	**521**	**201**	**1.000**

Fuente: (Lopez Roldan & Fachelli, 2015)

A partir de estas cuotas los encuestadores reciben instrucciones especificas para recoger el numero de cuestionarios que aseguran las proporciones poblacionales en términos de cuotas de muestra. Así, por ejemplo, un encuestador en concreto puede tener el encargo de encuestar a 50 personas en un determinado barrio, 23 deben ser mujeres y 26 varones; los entrevistados de cada sexo se distribuirán en tres grupos de edades y tres grupos de estudios de la forma que recoge la tabla 37. (Lopez Roldan & Fachelli, 2015)

Tabla 36

Asignación de encuestas al encuestador/a

		Estudios			Total
		Primarios	Secundarios	Terciarios	**Total**
Varones	16-29	1	1	1	**4**
	31-49	3	9	4	**17**
	50 o más	1	4	2	**7**
Mujeres	16-29	1	1	1	**3**
	31-49	2	7	3	**12**
	50 o más	1	5	1	**8**
Total		**7**	**26**	**10**	**50**

Fuente: (Lopez Roldan & Fachelli, 2015)

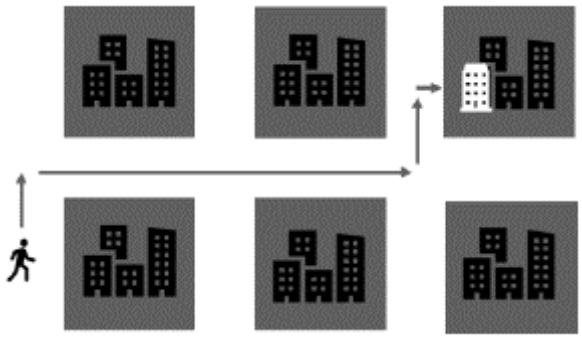
Los encuestadores/as tienen que identificar los individuos de las cuotas simplemente preguntando al tomar con la persona que se seleccione. Las personas se pueden seleccionar a partir de una ruta aleatoria, a la salida de algún tipo de establecimiento o edificio (centro comercial, colegio electoral, etc.), entre los peatones en lugares determinados de la ciudad, etc.

En el caso de las rutas aleatorias los elementos de la muestra se seleccionan empezando la ruta en un punto elegido al azar dentro de un área geográfica. Desde el encuestador/a,

a partir de un criterio de aleatorización predefinido, va seleccionado viviendas y a los individuos que en ellas habitan. Por ejemplo, desde el punto de arranque el encuestador/a camina dos manzanas a la derecha, gira a la izquierda y en la acera de enfrente selecciona de forma aleatoria una vivienda de un edificio y una persona que se corresponda a las cuotas asignadas.

Ilustración 31

Con las distintas rutas elegidas se trata de alcanzar la mayor cobertura geográfica



Fuente: (Lopez Roldan & Fachelli, 2015)

Con esta estrategia se busca implementar un procedimiento de selección semi aleatorio o semi probabilístico pues no garantiza las condiciones exigibles a una muestra estrictamente aleatoria, en particular, se desconoce la probabilidad de ser elegido un individuo y no se garantiza que todos tengan la misma probabilidad de ser seleccionados. De ello se deriva que los razonamientos inferenciales o probabilísticos no se puedan plantear en sentido estricto. Aun así, se suelen aplicar los mismos conceptos, formulaciones y conclusiones estadísticas, si bien contendrán un grado de incertidumbre mayor que en muchas ocasiones genera resultados carentes de la suficiente fiabilidad.

El método tiene bastantes similitudes con el muestreo aleatorio estratificado, de tipo proporcional. También se trata de dividir la población en subpoblaciones, y aquí las cuotas actúan a modo de estratos de características poblacionales que son conocidas, pero con una diferencia importante: la muestra no se extrae de forma probabilística, sino que la selección de los individuos depende finalmente del juicio de los entrevistadores/as. El muestreo por cuotas representa una buena alternativa cuando el criterio de economía de costes prima a la investigación, pero, al no tratarse de una muestra aleatoria nunca podemos saber los errores que se cometen, y el nivel de precisión supuesto depende de la idoneidad del modelo de que justifica los criterios de selección de las unidades.

Pero además se pueden introducir sesgo específicos: a veces no se dispone de la información suficiente para atribuir la cuota en cada casilla; pese a que las cuotas sean correctas se pueden dar sesgos en la selección de las unidades de la muestra tendiendo a escoger personas fácilmente accesibles o condicionadas por criterios subjetivos del entrevistador (dejar las personas que viven en pisos altos sin ascensor, pisos de más difícil acceso o apariencia de abandono, escoger personas afines al entrevistador/a, buscar a las personas en determinados momentos del día, etc.) que el procedimiento de rutas aleatorias con una hoja de ruta bien detallado intenta corregir; puede resultar difícil completar determinados perfiles muy especificados definidos por una cuota (por ejemplo, “mujer de

categoría profesional alta de 30 a 45 años residente en un barrio determinado"). Finalmente puede suceder que la muestra realmente mantenga las proporciones poblacionales definidas por las variables que configuran las cuotas, pero en relación o otras variables estudiadas no se dé una verdadera representatividad.

Para mejor la eficiencia en la utilización de este tipo de muestreo habitualmente se combina con otras estrategias en un diseño polietápico, por ejemplo, estratificando geográficamente en una primera etapa (o más etapas) y recogiendo las unidades finales con las cuotas definidas en la segunda etapa (o ulterior etapa). (Lopez Roldan & Fachelli, 2015)

3.4.7. Muestreo por conveniencia

Es un tipo de muestreo en el que las unidades están disponibles y son fáciles de localizar, tienen un carácter de representatividad de la población que se quiere analizar, pero se hace una selección conveniente de varias unidades con el objetivo de constituir grupos de control y experimentales a partir de la elección aleatoria de unos cuantos individuos que los forman. (Lopez Roldan & Fachelli, 2015)

4. UNIDAD 4: ESTADÍSTICA INFERENCIAL

La inferencia estadística comprende una serie de técnicas de uso imprescindible para tomar decisiones con respecto a la cuestión planteada por el investigador al comienzo de su tarea de análisis de datos.

Es necesario destacar aquí que las decisiones que debe tomar el investigador ante la situación de incertidumbre que implica inferir de casos particulares a la generalidad, deben estar respaldadas por la objetividad que garantiza la aplicación del método científico.

De este modo, los resultados obtenidos en situaciones experimentales, serán idealizados de acuerdo a un modelo probabilístico conveniente, permitiendo al investigador medir en términos de probabilidad la incertidumbre que trae aparejada la generalización de sus resultados. En otras palabras, podrá medir y comunicar el "error" que puede cometer o la confianza que deposita en sus decisiones.

Si la distribución de frecuencias de las observaciones puede asimilarse a la distribución de probabilidad teórica en la cual está basada la aplicación de la metodología inferencial elegida, entonces el investigador podrá estar seguro de que el error que informa en el proceso de prueba de hipótesis estadísticas planteadas o la confianza con que realiza sus estimaciones son correctos.

Aplicar cualquier metodología estadística inferencial sin estudiar a fondo el cumplimiento de los supuestos en los cuales ella esta basada, lleva irremediablemente a conclusiones erróneas. (Sacco , 2011)

a) Ramas de la Estadística Inferencial:

Los dos principales procedimientos de la estadística inferencial son la estimación (puntual o por intervalos) y las pruebas de hipótesis. Comúnmente, los parámetros de una población son desconocidos, siendo necesario estimar el valor de estos o, si no, efectuar indagaciones (pruebas de hipótesis) para comprobar si los valores a ellos atribuidos pueden ser considerados como verdaderos. (Sacco , 2011)

Ilustración 32

Ramas de la Estadística Inferencial

Fuente: (Sacco , 2011)

b) Población y muestra:

Una población se define como la totalidad de elementos sobre los cuales se desea estudiar un tema en particular. Por ejemplo, si se desea estudiar el ingreso promedio de las familias en la ciudad de San Nicolas, la población estará constituida por todas las familias que habitan en esta ciudad.

Es evidente que, por razones de costo y tiempo, sería casi imposible encuestadas a todas. Generalmente se encuesta a una pocas seleccionadas de tal manera que representen al total de familias que componen la población en cuestión.

A ese conjunto de familias seleccionadas a partir de una cierta población, se lo denomina muestra.

El procedimiento que generalmente se sigue en cualquier investigación consiste en obtener resultados a partir de una muestra y luego generalizarlos a la población objetivo.

Una población cualquiera queda perfectamente especificado por ciertas medidas denominadas parámetros poblacionales.

Un parámetro poblacional es una medida que se calcula teniendo en cuenta todos los elementos que componen una cierta población.

Por ejemplo, si el ingreso promedio de las familias de la ciudad de San Nicolas se calcula considerando el ingreso de todas las familias que habitan en la ciudad, este ingreso promedio es un parámetro poblacional.

Es evidente que los parámetros poblacionales son generalmente imposibles de calcular. En la práctica, casi siempre se trabaja con muestras.

Las medidas calculadas a partir de las observaciones muestrales, se conocen con el nombre de estadísticos muestrales.

Un estadístico muestral es una medida que se calcula teniendo en cuenta solamente los elementos que integran una muestra determinada.

Así, si se toma una muestra de 100 familias de la ciudad de San Nicolas, se les realiza una entrevista en la que se pregunta el ingreso familiar y, en base a la información recogida se calcula un ingreso promedio, este promedio es un estadístico muestral.

Ahora bien, ¿para qué nos sirve calcular el ingreso promedio de las 100 familias que salieron seleccionadas en la muestra?

Nos sirve pues ella es la única información con la que contaremos para decir algo acerca de todas las familias de la ciudad de San Nicolas.

Debido a esto, la utilizaremos para estimar al parámetro poblacional (ingreso promedio de todas las familias de la ciudad de San Nicolas) que prácticamente nunca llegaremos a conocer.

El proceso por medio del cual se establecen relaciones entre los estadísticos muestrales y los parámetros poblacionales es el objeto de la inferencia estadística.

4.1. Estimación de Parámetros

Como se dijo anteriormente, el objetivo de la estadística consiste en hacer inferencias acerca de los parámetros de una población teniendo en cuenta la información contenida en la muestra.

Ahora bien, como en general los parámetros poblacionales son desconocidos, existe una amplia gama de técnicas estadísticas que tienen como objetivo la estimación de estos parámetros a través de estadísticos muestrales adecuados a cada caso en particular.

La base teórica que sustenta la metodología que aplicamos a los resultados de una muestra, se fundamenta en la distribución de probabilidad del estimador calculando en cada una de las muestras posibles.

Existen dos tipos de estimaciones para parámetros, puntuales y por intervalo.

Una estimación puntual es un único valor estadístico y se usa para estimar un parámetro. El estadístico usado se denomina estimador.

Una estimación por intervalo es un rango, generalmente de ancho finito, que se espera que contenga el parámetro.

a) Estimadores:

Un estimador es una función que permite calcular valores aproximados al del parámetro, además, se lo considera una variable aleatoria ya que puede tomar valores que pertenecen a un intervalo de números reales.

Es posible definir muchos estadísticos para estimar un parámetro desconocido. Por ejemplo, puede elegirse la media muestral para estimar el valor de la media poblacional, o también la mediana muestral.

Por ejemplo, la media muestral se obtiene sumando todas las observaciones de la muestra y dividiendo esta suma por el tamaño de la muestra.

Cualquier persona podría definir otra combinación de las observaciones muestrales como estimador del parámetro μ y entonces cabria la pregunta:

¿Cuál es el "mejor" estimador de μ?

Un problema importante que debió resolver la teoría estadística, fue el de determinar el mejor estimador de cada parámetro en particular.

b) Propiedades de los estimadores

Cuando se analizan conceptos generales y métodos de inferencia es conveniente tener un símbolo genérico para el parámetro de interés. Se utilizará la letra griega θ para nombrar al parámetro y con θ al estimador.

Entonces, ¿Cómo seleccionar un buen estimador de θ?, ¿Cuáles son los criterios para juzgar cuando un estimador es "bueno" o "malo"?

Si se piensa en términos de estimadores humanos como se encuentran en las grandes compañías, entonces, quizá un buen estimador es aquella persona cuyas estimaciones siempre se encuentran muy cercanas a la realidad.

De aquí surgen dos propiedades deseables de un estimador:

1) La distribución muestral de θ debe tener una media igual al parámetro estimado θ, en este caso se dice que el estimador es insesgado.

Si se usa la media muestral X para estimar la media poblacional μ, se sabe que la $\mu x = \mu$, por lo tanto, la media es un estimador insesgado.

2) La varianza del estimador debe ser la menor posible, en este caso se dice que el estimador es eficiente o con varianza mínima.

Suponiendo que $\theta 1$ y $\theta 2$ son dos estimadores insesgados de θ. Entonces, aun cuando la distribución de cada estimador este centrada en el valor verdadero de θ, las dispersiones de las distribuciones alrededor del valor verdadero pueden ser diferentes.

Entre todos los estimadores θ de θ que son insesgados, es necesario seleccionar al que tenga varianza mínima. En otras palabras, la eficiencia se refiere al tamaño de error estándar de la estadística.

Si comparamos dos estadísticas de una muestra del mismo tamaño y tratamos de decidir cuál de ellas es un estimador más eficiente, escogeríamos la estadística que tuviera el menor error estándar, o la menor desviación estándar de la distribución de muestreo.

Tiene sentido pensar que un estimador con un error estándar menor tendrá una mayor oportunidad de producir una estimación más cercana al parámetro de población que se está considerando. Como se puede observar las dos distribuciones tienen un mismo valor en el parámetro solo que la distribución muestral de medias tiene una menor varianza, por lo que la media se convierte en un estimador eficiente e insesgado.

Es posible sumar a estas dos propiedades:

- Coherencia: Una estadística es un estimador coherente de un parámetro de población, si al aumentar el tamaño de la muestra se tiene casi la certeza de que

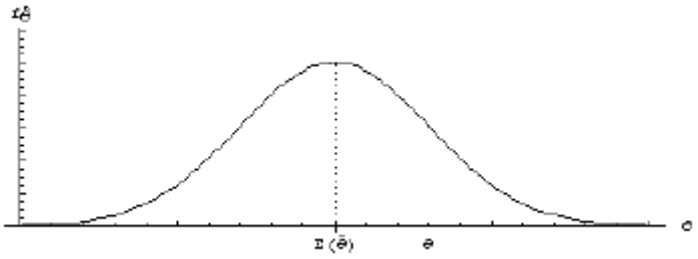
el valor de la estadística se aproxima bastante al valor del parámetro de la población. Si un estimador es coherente se vuelve mas confiable si tenemos tamaños de muestras más grandes.

- Suficiencia: Un estimador es suficiente si utiliza una cantidad de la información contenida de la muestra que ningún otro estimador podría extraer información adicional de la muestra sobre el parámetro de la población que se está estimando. Es decir, se pretende que al extraer la muestra el estadístico calculado contenga toda la información de esa muestra. Por ejemplo, cuando se calcula la media de la muestra solo se utiliza a un dato o a dos. Esto es solo el dato o los datos del centro son los que van a representar la muestra. Con esto se deduce que si utilizamos a todos los datos de la muestra como es en el caso de la media, la varianza, desviación estándar, etc.; se tendrá un estimador suficiente.

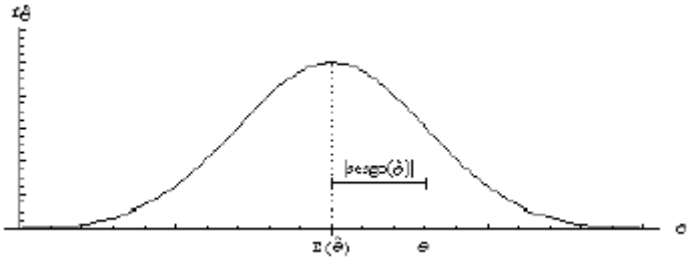
c) Interpretación grafica de estas propiedades:

Supongamos que estamos estimando la media de una población normal. Es decir, la media de una población que sabemos que es normal, aunque no separamos su media. Si como estimador de la media usamos, por ejemplo, alguna combinación lineal de los valores de una muestra tomada de esa población, entonces como el valor de cada valor de la muestra es una variable normal en si misma, y una combinación lineal de variables normales es una variable normal, nuestro estimador también es una variable aleatoria normal.

Si calcularemos como vimos antes el valor esperado del estimador y lo graficaremos, podríamos llegar a un gráfico como este:

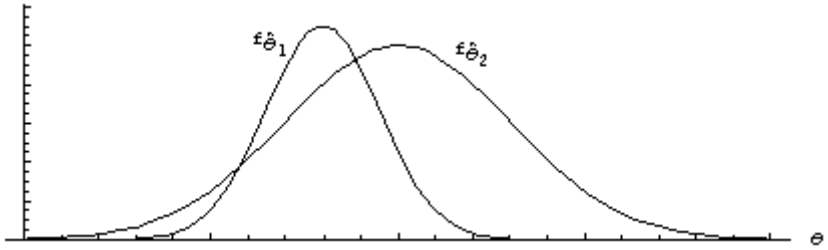


En este grafico podemos apreciar que es deseable que el valor esperado del estimador coincida con el parámetro estimado. Denominamos sesgo a la diferencia $E(\theta) - \theta$. Por eso cuando el sesgo de un estimador es cero, se lo denomina “insesgado”.



Como podemos observar, el estimador graficado no es insesgado. Por lo que dijimos antes, es deseable que el sesgo de un estimador sea pequeño.

Otra característica importante que analizamos fue la varianza. Es deseable que la varianza de un estimador sea pequeña, para que la variabilidad respecto de su valor esperado sea pequeña.



En el ejemplo graficado, la varianza de $\theta 1$ es mas pequeña que la de $\theta 2$. Vemos que su variabilidad respecto de su valor esperado es menor.

Ejemplo 1: Consideremos una población compuesta por 5 escuelas rurales en las que se ha registrado el número de maestros obtenidos: 2,3,6,8,11.

Analizar como la estimadora media muestral cumple con todas las propiedades de un buen estimador.

(i) Insesgabilidad

En primer lugar, calculamos la media aritmética y la varianza correspondiente a la variable X = cantidad de maestros por escuela. Se obtiene $\mu = 6$ maestros por escuela y $\sigma^2 = 10{,}8$ con una desviación típica de $\sigma = 3{,}29$

Estos resultados nos indican que el promedio de maestros en escuelas rurales en la población es de 6 maestros por escuela con una dispersión de 3,29 maestros por escuela.

Supongamos, ahora, que seleccionamos todas las muestras posibles de tamaño n = 2 por medio de un muestreo con reemplazo.

Como podemos observar en la Tabla 38, cada una de estas muestras es el resultado de un experimento aleatorio y todas tienen la misma probabilidad de ser seleccionada $(1/N^n = 1/25)$. Luego, el muestreo es aleatorio y cada muestra es una muestra aleatoria simple.

En cada muestra podemos obtener la media aritmética, como vemos en la columna 4 de la tabla 38.

Tabla 37

Muestras posibles y Promedios de cada muestra

Muestra	Observaciones en la muestra (x_i, x_i)	Probabilidad de cada muestra $p(x_i)$	Media muestra $\bar{x}_i$
1	(2,2)	1/25	2,0
2	(2,3)	1/25	2,5
3	(2,6)	1/25	4,0
4	(2,8)	1/25	5,0
5	(2,11)	1/25	6,5
6	(3,2)	1/25	2,5
7	(3,3)	1/25	3,0
8	(3,6)	1/25	4,5
9	(3,8)	1/25	5,5
10	(3,11)	1/25	7,0
11	(6,2)	1/25	4,0
12	(6,3)	1/25	4,5
13	(6,6)	1/25	6,0
14	(6,8)	1/25	7,0
15	(6,11)	1/25	8,5
16	(8,2)	1/25	5,0
17	(8,3)	1/25	5,5
18	(8,6)	1/25	7,0
19	(8,8)	1/25	8,0
20	(8,11)	1/25	9,5
21	(11,2)	1/25	6,5
22	(11,3)	1/25	7,0
23	(11,6)	1/25	8,5
24	(11,8)	1/25	9,5
25	(11,11)	1/25	11,0

Fuente: (Sacco , 2011)

Vemos que, de esta forma, nos ha quedado definida una nueva variable aleatoria: la variable aleatoria media muestral (última columna de la tabla). El valor que ella toma, depende de la muestra a la que corresponda.

Como cada media aritmética está calculada con las observaciones muestrales, el valor obtenido en cada muestra será un estadístico muestral.

Ahora bien, por ser la media aritmética una variable aleatoria, podemos establecer su correspondiente distribución de probabilidad y calcular la esperanza matemática y varianza. Para eso, construiremos primero la tabla de frecuencias, computando los diferentes valores de xi y sus repeticiones.

Esto se presenta en las dos primeras columnas de la Tabla 39.

Las columnas restantes, resultan de asimilar las frecuencias relativas a las probabilidades (Teoría frecuencial de la probabilidad) y utilizarlas para obtener la Esperanza y la Varianza. (Sacco , 2011)

Tabla 38

Distribución de probabilidad de la variable Aleatoria media muestral

$\bar{x}_i$	n_i	$p(\bar{x}_i)$	$x_i\ p(\bar{x}_i)$	$(\bar{x}_i - \bar{x})^2\ p(\bar{x}_i)$
2,0	1	1/25	2/25	16/25
2,5	2	2/25	5/25	24,5/25
3,0	1	1/25	3/25	9/25
3,5	0	0/25	0	0
4,0	2	2/25	8/25	8/25
4,5	2	2/25	9/25	4,5/25
5,0	2	2/25	10/25	P-125
5,5	2	2/25	11/25	0,5/25
6,0	1	1/25	6/25	0
6,5	2	2/25	13/25	0,5/25
7,0	4	4/25	28/25	4/25
7,5	0	0/25	0	0
8,0	1	1/25	8/25	4/25
8,5	2	2/25	17/25	12,5/25
9,0	0	0/25	0	0
9,5	2	2/25	19/25	24,5/25
10,0	0	0/25	0	0
10,5	0	0/25	0	0
11,0	1	1/25	11/25	25/25
	25	**1**	**150/25**	**135/25**

Fuente: (Sacco , 2011)

Se obtiene $E(x) = x = \sum_{i=1}^{N^n} Xi.p(Xi) = 6$ La primera conclusión importante que hemos obtenido es $E(x) = \mu$ y esto equivale a decir que x es un estimador insesgado de μ.

Sin embargo, es oportuno aclarar aquí que la propiedad de insesgamiento es un concepto teórico. Únicamente se da en términos de valores esperados, puesto que si nos fijamos en la tabla 38 encontraremos que, de las 25 muestras posibles, solo una media muestral coincide con el valor del parámetro μ.

Calculemos ahora la varianza de la variable aleatoria media muestral utilizando también la información proporcionada por la tabla.

Obtenemos $V(x) = \sum_{i=1}^{N^n} Xi.p(xi) = 6$. La primera conclusión importante que hemos obtenido es $E(x) = \mu$ y esto equivale a decir que x e un estimador insesgado de μ.

Sin embargo, es oportuno aclarar aquí que la propiedad de insesgamiento es un concepto teórico. Únicamente se da en términos de valores esperados, puesto que si nos fijamos en la tabla 38 encontraremos que, de las 25 muestras posibles, solo una media muestral coincide con el valor del parámetro μ.

Calculemos ahora la varianza de la variable aleatoria media muestral utilizando también la información proporcionada por la tabla.

Obtenemos $V(x) = \sum_{i=1}^{N^n}(X1 - X)^2.p(Xi) = 5{,}40$. Vemos que este valor es exactamente igual al de la varianza poblacional divido por el tamaño de la muestra, es decir:

$$V(x) = 5{,}40 = \frac{10{,}80}{2} = \frac{\sigma^2}{n}$$

La segunda conclusión importante a la que llegamos es que la varianza de la variable aleatoria media muestral es directamente proporcional a la varianza de la variable aleatoria media muestral es directamente proporcional a la varianza poblacional e inversamente proporcional al tamaño de la muestra. Esto quiere decir que a medida que se incrementó el tamaño de la muestra menor es la variabilidad de la variable media muestral, mientras que, cuanto más variable es la característica en estudio en la población (expresada por una mayor varianza σ^2), mayor será también la varianza de la variable aleatoria media muestral. (Sacco , 2011)

(ii) Insesgabillidad de mínima varianza (eficiencia)

Veamos ahora si la media muestral es un estimador eficiente. Para verificar esta propiedad debemos comparar la varianza de esta variable aleatoria con la de algún otro estimador insesgado de μ.

Cuando la distribución de la característica en estudio en la población es perfectamente simétrica, la mediana también es un estimador insesgado de μ.

Ahora bien, la varianza de la estimadora mediana muestral es $V(Me) = \frac{\sigma^2}{n}.\frac{4}{\pi}$ (verificar), lo que nos permite verificar que siempre $V(Me) > V(x)$

Con ello verificamos que la media muestral es un estimador más eficiente que la mediana para estimar al parámetro poblacional μ. En símbolos:

$$\frac{V(Me)}{V(x)} = \frac{\frac{\sigma^2}{n}\cdot\frac{4}{\pi}}{\frac{\sigma^2}{n}} = \frac{4}{\pi} > 1$$

En general, podemos verificar que la media muestral es un estimador que tiene varianza mínima cuando se lo compara con cualquier otro estimador del parámetro media poblacional. (Sacco , 2011)

(iii) Consistencia

En cuanto a la propiedad de consistencia, ella no necesita verificación. Es evidente que si la media muestral es un estimador insesgado también será consistente.

(iv) Distribución asintóticamente normal

Analizaremos ahorra si la estimadora media muestral cumple con la propiedad 4, es decir, tiene distribución asintóticamente normal.

Evidentemente hay dos situaciones posibles:

- Cuando la muestra se selecciona aleatoriamente de una población con distribución normal.
- Cuando la muestra se selecciona aleatoriamente de una población sin distribución normal.

En el primer caso de la teoría estadística establece que la distribución de la variable aleatoria media muestral, calculada en base a una muestra seleccionada aleatoriamente de una población con distribución normal, responde a las características de la distribución normal de la población de origen.

En el segundo caso de la teoría estadística establece que, aun cuando la distribución poblacional de la característica en estudio se aleja bastante de la forma normal, la distribución de la variable aleatoria media muestral se aproxima a la distribución normal a medida que se incrementó el tamaño de la muestra.

Esta última afirmación está sustentada en un importante teorema de la teoría estadística conocido con el nombre de teorema central del límite.

Luego es fácil concluir que la estimadora media muestral cumple con la propiedad 4. (Sacco , 2011)

4.1.1. Estimación puntual

La estimación puntual es un proceso mediante el cual se estima el parámetro en un punto, dando un valor especifico como estimación.

Los dos métodos tradicionales de estimación puntual de parámetros son el de mínimos cuadrados y el de máxima verosimilitud.

El objetivo de la estimación puntual es seleccionar solo un número, basados en datos de la muestra, que represente el valor más razonable de θ

Una estimación puntual de un parámetro θ es un solo numero que se puede considerar como el valor mas razonable de θ. La estimación puntual se obtiene al seleccionar un estadístico apropiado y calcular su valor a partir de datos de la muestra dada. El estadístico seleccionado se llama estimador puntual de θ. (Sacco , 2011)

Tabla 39

Estimadores más frecuentes de los parámetros

Parámetro	Estimador más probable
μ	$\bar{x}$, la media muestral
σ^2 o σ	s^2 o s, la varianza muestral o desviación estándar muestral
ρ	$\hat{\rho}=X/n$, la proporción muestral, donde x es el número de objetos en una muestra aleatoria de tamaño n que pertenece a la clase de interés
$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$, la diferencia de medias muestrales de dos muestras aleatoria independientes
$\rho_1-\rho_2$	$\hat{\rho}_1-\hat{\rho}_2$, la diferencia entre las proporciones de dos muestras aleatorias independientes

Fuente: (Sacco , 2011)

En el mejor de los casos, se encontrará un estimador θ para el cual $\theta = \theta$ siempre. Sin embargo, θ es una función de los Xi muestrales, por lo que es en si misma, una variable aleatoria. Entonces, $\theta = \theta + error\ de\ estimación$ lo cual nos permite deducir que el estimador preciso seria uno que produzca solo pequeñas diferencias de estimación, de modo que los valores estimados se acerquen al valor verdadero.

La distancia entre una estimación y el parámetro estimado recibe el nombre de error de estimación.

Para cualquier estimador puntual con una distribución normal, la regla empírica dice que aproximadamente el 95% de todas las estimaciones puntuales estarán a no mas de dos (exactamente 1,96) desviaciones estándar de la media de esa distribución.

Ejemplo 1: La longitud de los tornillos que produce una determinada maquina es una variable normal, pero no sabemos cuánto vale el parámetro (media poblacional) μ de esa distribución normal.

Podemos hacer el experimento de tomar 10 tornillos, calcular el promedio de sus longitudes, y usar ese promedio como estimación de μ. En este caso el estimador es:

$$\mu = x = \frac{\sum_{i=1}^{n} Xi}{n}$$

El margen de error se estima como

$$\pm 1{,}96 \sqrt{\frac{\sum_{i=1}^{n} (Xi - X)^2}{n}}$$

Ejemplo 2: En el futuro habrá cada vez más interés en desarrollar aleaciones de Mg de bajo costo, para varios procesos de fundición. En consecuencia, es importante contar con métodos prácticos para determinar varias propiedades mecánicas de esas aleaciones. Examine la siguiente muestra de mediciones del módulo de elasticidad obtenidas de un proceso de fundición a presión 44.2 43.9 44.7 44.2 44.0 43.8 44.6 43.1

Suponga que esas observaciones son el resultado de una muestra aleatoria. Se desea estimar la varianza poblacional σ^2

Un estimador natural es la varianza muestral

$$\sigma^2 = \frac{\sum_{i=1}^{n} (xi - x)^2}{n - 1}$$
$$= \frac{(44{,}2 - 44{,}0625)^2 + (43{,}9 - 44{,}0625)^2 + \ldots + (43{,}1 - 44.0625)^2}{8 - 1}$$
$$= 0{,}251$$

Este valor nos permite encontrar el margen de error que se comete al estimar el valor de la media muestral. (Sacco , 2011)

4.1.2. Estimación por intervalos

Un estimado puntual, por ser un solo número, no proporciona por sí mismo información alguna sobre la precisión y confiablidad de la estimación.

Por ejemplo, si se tiene interés en estimar la resistencia promedio a la ruptura de cierto elemento estructural, es probable que un solo número no sea tan significativo como un intervalo, dentro del cual se espera encontrar el valor de este parámetro.

Una alternativa para reportar un solo valor del parámetro que se esté estimando es calcular e informar todo un intervalo de valores factibles, un estimado de intervalo o intervalo de confianza (IC).

Una estimación por intervalos de un parámetro desconocido θ es un intervalo de la forma $I \leq \theta \leq u$, donde los puntos extremo I y u dependen del valor numérico de la estadística θ para una muestra en particular y de la distribución de muestreo de θ.

De la distribución de muestreo de θ es posible determinar los valores de I y u tales que la siguiente proposición sea verdadera:

$$P(I \leq \theta \leq u) = 1 - \propto \qquad ; \qquad 0 <\propto< 1$$

Por tanto, se tiene una probabilidad de $1 - \propto$ de seleccionar una muestra que produzca un intervalo que contiene el valor verdadero de θ. El intervalo resultante $1 \leq \theta \leq u$ se conoce como Intervalo de Confianza del $100(1 - \propto)$ por ciento.

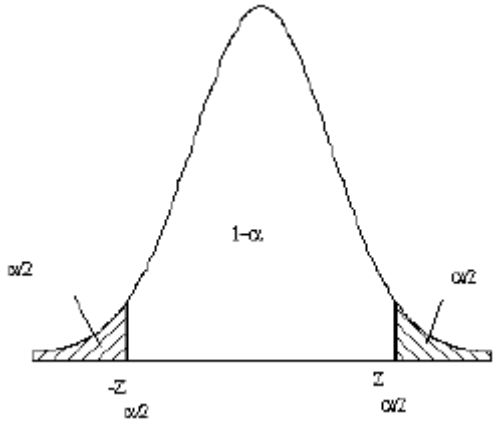
Las cantidades I y u se denominan límites de confianza inferior y superior y $1 - \propto$ es el coeficiente de confianza (o nivel de confianza, que es una medida del grado de fiabilidad en el intervalo).

De tal forma, cuando $\propto = 0{,}05$, se tiene un intervalo de confianza del 95% y cuando $\propto = 0{,}01$ se tiene uno del 99%. Entre mayor es el intervalo de confianza se tiene mas seguridad de que el mismo contenga el parámetro desconocido.

Sobre 100 muestras aleatorias de un cierto tamaño n de una población, si en cada una se calcula la media muestral x y, a partir de ellas, se construyen 100 intervalos de confianza para el parámetro que se desea estimar 95 contendrán al verdadero valor del parámetro poblacional, mientras que 5 no lo abarcaran. (Sacco , 2011)

Un intervalo del tipo $I \leq \theta \leq u$ recibe el nombre mas apropiado de Intervalo de Confianza Bilateral. También existen intervalos de confianza Unilaterales $I \geq \theta \; y \; \theta \leq u$ donde los limites de confianza se eleigen de modo que $P(\theta \geq I) = 1 - \propto$ y $P(\theta \geq I) = 1 - \propto$

i. Intervalos de confianza para las medias



Para estimar la media μ de una característica de la población, es necesario, primero saber si la varianza de la población es conocida o no lo es.

Para estimar la media μ de una característica de la población, cuando se considera conocida la varianza de esa población, se toma una muestra de tamaño "n" y se le calcula la media muestral X.

Por el Teorema del Limite Central, se sabe que de la distribución de la media muestral se obtiene que $Z = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}$ tenga una distribución como una normal estándar, con media E(X) = μ y varianza $\frac{\sigma^2}{n}$.

Teniendo en cuenta que la distribución de la media muestral sigue o tiende a una Distribución Normal, y considerando la varianza como conocida, el intervalo de confianza debe abarcar el área de $(1-\propto)\%$ ente sus limites superior e inferior en dicha distribución. Cada limite es expresado en unidades de desviación típica y esas unidades esta expresadas por $Za/2$. En consecuencia, el intervalo de confianza bilateral del $100(1-\propto)\%$ para μ este dado por $I = (X - Za/2\frac{\sigma}{\sqrt{n}}; X - Za/2\frac{\sigma}{\sqrt{n}})$

Tener en cuenta que esta expresión obedece al concepto:

Valor del parámetro=estimación puntual $\pm$ una función de la confianza y la dispersión (directamente proporcionales) y del tamaño de la muestra (inversamente proporcionales).

Ejemplo 1: Supongamos que le director de investigaciones de mercado de una fábrica automotriz necesita hacer una estimación de la vida promedio de las bacterias que su compañía produce. Selecciona aleatoriamente 200 usuarios y resulta tener una vida promedio de sus bacterias de 36 años.

Encontrar el intervalo dentro del cual es probable que este la media de población desconocida.

Para ello, es necesario encontrar el error estándar de la media $\sigma x = \frac{\sigma}{\sqrt{n}}$, conociendo que la desviación de la población es 10 meses, obtenemos $\sigma x = 0{,}707$

Ahora podemos decir que la vida útil de las bacterias está dentro del intervalo $I = (X - \sigma x; X + \sigma x)$,es decir, $I = (35{,}293; 36{,}707)$

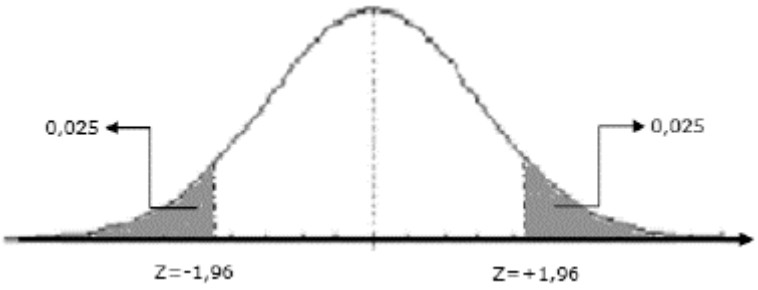
Aunque estos datos son útiles no son suficientes, pues no tiene así un nivel de confianza significativo. Para esto debemos recordar que cuando trabajamos con la distribución normal de probabilidad, hemos visto que posiciones especificas del área bajo la curva normal están localizadas a una distancia de cierto número dado de desviaciones estándares por debajo y por arriba de la media. Esto se puede aplicar al error estándar de la media.

Así podemos decir que, por ejemplo, el 95,5% de las medias de muestra están dentro de $\pm 2\sigma$ de μ, y en consecuencia, μ esta dentro de $\pm 2\sigma$ de la media de cada una de tales muestras. De manera parecida, también podemos decir que la probabilidad de que la media de la muestra este dentro de $\pm\sigma$ de la media de la población μ es de 0,683. (Sacco , 2011)

Ejemplo 2: Un vendedor de partes del automotor, mayorista, necesita una estimación de la vida media que puede esperar de los limpiadores de parabrisas en condiciones normales de manejo. La administradora ya ha determinado que la desviación estándar de la vida útil de la población es de 6 meses. Se selecciona una sola muestra de 100 limpiadores y se obtienen que x = 21 meses, encontrar el intervalo de confianza de un 95%.

Como la muestra es mayor de 30, debemos calcular el error estándar de la media $\sigma x = 0{,}6$. Ahora, considerando un nivel de confianza del 95%, podemos obtener Z de la siguiente manera $\propto = 100 - 95 = 5 = \frac{\propto}{2} = 2{,}5 = \frac{2{,}5}{100} = 0{,}025$

Con este valor planteamos que $\frac{F(Z)=1-F(-Z)}{F(Z)=1-0{,}025=0{,}975}$ este valor lo buscamos en tabla y obtenemos Z=1,96. (Sacco , 2011)



Con este valor calculamos el intervalo de confianza en el que puede estar la media poblacional

$$I = (X - Za/2\frac{\sigma}{\sqrt{n}}; X - Za/2\frac{\sigma}{\sqrt{n}}) = (19{,}824; 22{,}176)$$

Ejemplo 3: Se quiere estimar el ingreso medio anual de 700 familias que vive en una sección de 4 manzanas. Si se toma una muestra de 50 familias y se hallan los siguientes resultados x = $11800 y s = $950. Encontrar un intervalo con un nivel de confianza del 90% en el que pueda encontrarse la media poblacional.

Calculo el error producido en la desviación estándar de la media $\sigma x = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = 129.56$.

Con el valor del 90% de nivel de confianza obtenemos que $\propto = 100 - 90 = 10; \frac{\propto}{2} = 5; \frac{5}{100} = 0{,}05$. De igual manera que lo hicimos antes, obtenemos z = 1,64.

Entonces $I = (X - Za/2\sigma x; X - Za/2\sigma x) = (11587{,}52; 12012{,}48)$

Como síntesis, podemos decir que para reducir la amplitud de un intervalo de confianza y en consecuencia aumentar su precisión, debemos reducir el error estándar de la media muestral x que es σ/n. Esto puede lograrse solamente disminuyendo la variabilidad de los datos ya sea homogeneizando el material experimental o, si esto no puede llevarse a cabo, aumentando el tamaño de la muestra.

Es costumbre utilizar coeficientes de confianza del 90%, 95% y 99%. Por este motivo es posible considerar la siguiente tabla que resume los valores de probabilidad de la distribución normal estandarizada para estos niveles de confianza.

Coeficiente de confianza	z
0,90	1,645
0,95	1,960
0,99	2,576

¿Por qué "intervalos de confianza" y no "de probabilidad"?

Si observamos las expresiones de los intervalos obtenidos para la media poblacional, de un 95% de confianza o de un 99% de confianza, se puede apreciar que en ellos no está implicada ninguna variable aleatoria, ya que en el centro del intervalo esta un parámetro μ y en los extremos se tienen números obtenidos sumando y restando $Z\sigma/\sqrt{n}$ a la media muestral), el 95% o 99% (según el valor de Z) de los posibles intervalos contendrán al verdadero valor de μ. De allí la expresión "existe entre un 95 (o 99) por ciento de confianza de que el intervalo contenga al parámetro". (Sacco , 2011)

4.2. Pruebas de Hipótesis

Uno de los principales propósitos del análisis estadístico es hacer inferencias acerca de la población mediante la comparación de una o más muestras de la población, para esto hay que formularse hipótesis estadísticas. La primera hipótesis se denomina hipótesis nula, abreviada Ho, esta supone un efecto dado: por ejemplo, podríamos tratar de probar la hipótesis de Ho. El poder aglutinante del suero es el mismo en los cirróticos que en los sujetos normales, o bien Ho. La densidad de pájaro ג es la misma en los bosques tropicales que el los bosques de coníferas. Esta hipótesis expresa el concepto de no diferencia. Estas hipótesis difieren de la hipótesis alternativa o contraria (H1 o Ha (y supone lo contrario de Ho. Por ejemplo, H1: La densidad del pájaro "ג" es diferente en los bosques tropicales que los bosques de coníferas.

Una vez definidas las hipótesis estadísticas, estas deben ser sometidas a una prueba de verdad, es decir definir si los resultados obtenidos por la muestra son conforme a los resultados supuestos, de esta manera definir si las diferencias registradas entre las observaciones y las hipótesis provienen del azar. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

a) Riesgo de Error en un Test Estadístico:

Cada test estadístico comporta dos riesgos de error.

(i) Si se rechaza una hipótesis que debería ser aceptada, se comete un error tipo $\propto$, también llamado de primera especie (Tipo I), y, por el contrario.

(ii) Si se acepta una hipótesis que debería ser rechazada se comete un error de tipo β o de segunda especie (Tipo II). En el ejemplo de los cirróticos, $\propto$ representa el riesgo de declarar diferente la media del poder aglutinante del suero de los cirróticos y los sujetos normales, mientras que son idénticas a nivel de poblaciones, y β es el riesgo de declarar idénticas las medias de ambos grupos mientas que son diferentes.

El cálculo del riesgo de error, depende de la manera de la manera como ha sido construido el test estadístico; así por ejemplo en el caso de formular un test para dimorfismo sexual en camarones a partir de la longitud total, se tiene una muestra de 121 individuos de los cuales se determinó el sexo por observación del telico. La longitud total media de 65 machos es 61.16 mm y la desviación estándar de 1.11mm. Si se admite que la distribución de la variable es normal, y que la media de la muestra es una buena estimación de la media poblacional, se puede decir que:

$$P(X \geq X + Z \propto Sx) = 0.05$$

$$P(X \geq 61.16 + 1.64 * 1.11)$$
$$= 0.05;\ 1.64\ corresponde\ al\ valor\ de\ z\ para\ un\ test\ unilateral\ donde\ \propto = 0.05$$

$$P(X \geq 62.98) = 0.05$$

Hipótesis establecidas:

H0: el camarón extraído aleatoriamente de la población que posee una longitud total Xi es un macho.

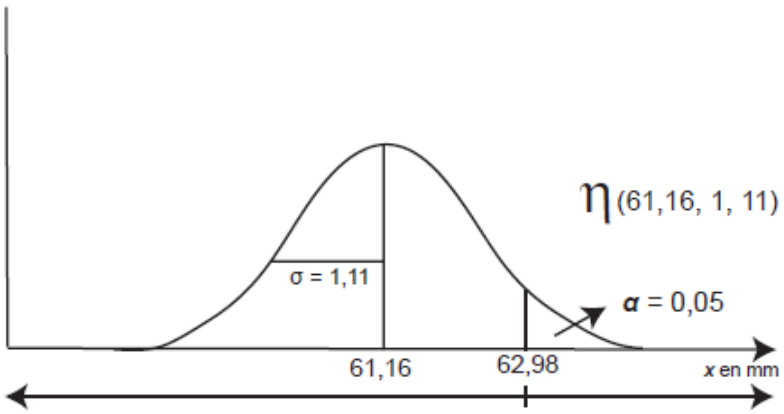
H1: el camarón extraído aleatoriamente de la población que posee una longitud total Xi es una hembra.

Las reglas de decisión pueden establecerse de la siguiente manera:

- Si Xi es superior al valor critico 62.98 mm la hipótesis es rechazada y el camarón es considerado como una hembra.
- Si Xi es inferior a 62.98 mm la hipótesis principal es aceptada y el camarón es un macho (Ilustración 33). (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Ilustración 33

Distribución de la longitud de camarón (mm) y nivel de significatividad.



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Con esta regla, el riesgo de tomar a una hembra por un macho es $\propto = 0{,}05$, en cuanto al riesgo β de tomar un macho por una hembra, este puede ser calculado a partir de las características de la población de machos inmaduros.

Si se considera que la longitud media de los camarones hembras, calculada sobre 56 individuos es 63.74mm con una desviación estándar de 1.20 mm y se admite que la distribución de la variable obedece a una distribución normal, es posible calcular la probabilidad de tener un macho con longitud inferior a 62.98 mm, nivel $\propto = 0{,}05$ (Ilustración 34).

$$P(X < 62.98) = \beta$$

$$P(Z < \frac{62.98 - 63.74}{1.2}) = \beta$$

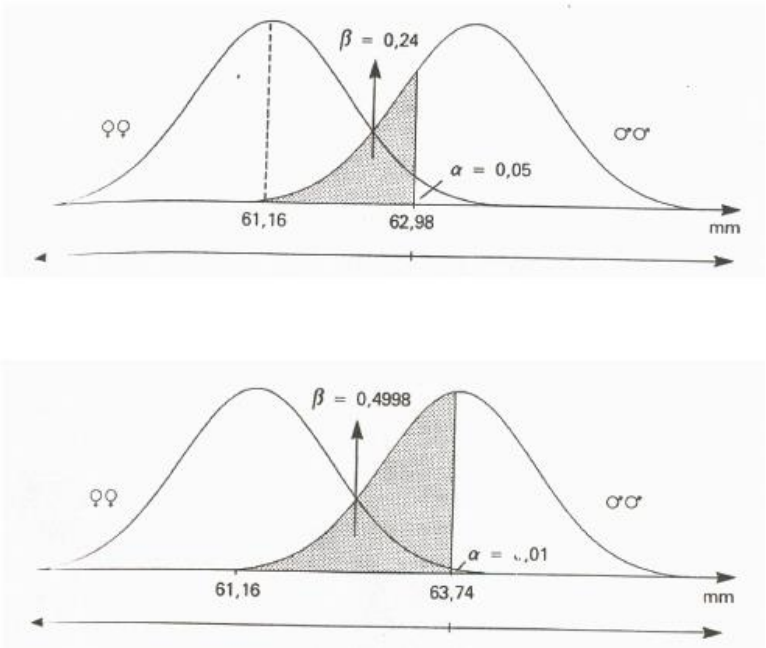
$$P(X < -0.716) = \beta$$

Ahora se lee 0.716 en la tabla de áreas bajo la curva de Z, para este valor Z = 0.7611 para obtener $\beta = 0.239$ es decir que la probabilidad de tomar a una hembra por un macho es 23.9%.

Para que un test estadístico sea eficaz, debe de estar construido de manera que los errores en la decisión sean mínimos. Lo que no es simple, puesto que, para una muestra de un tamaño dado, la disminución de un tipo de error, es normalmente acompañada por el incremento del otro tipo. En la práctica, un tipo de error puede ser mas importante que otro, el investigador tiene que definir el nivel de error que arriesga, con el fin de limitar el error más importante. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Ilustración 34

Distribución de la longitud de camarón (mm) y nivel de significatividad y error B.



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

b) Umbral de Probabilidad o Nivel de Significatividad

Cuando se prueba una hipótesis se arriesga un error $\propto$ hasta un cierto nivel, a este se le llama umbral de probabilidad o nivel de significatividad. Se reconoce como significativo al nivel de probabilidad igual al 0.05, muy significativo en nivel 0.01 y altamente significativo el nivel 0.001. Así si por ejemplo si se escoge el nivel de 0.05 entonces se dice que hay 5 oportunidades sobre 100 de rechazar una hipótesis Ho cuando esta debe ser aceptada. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

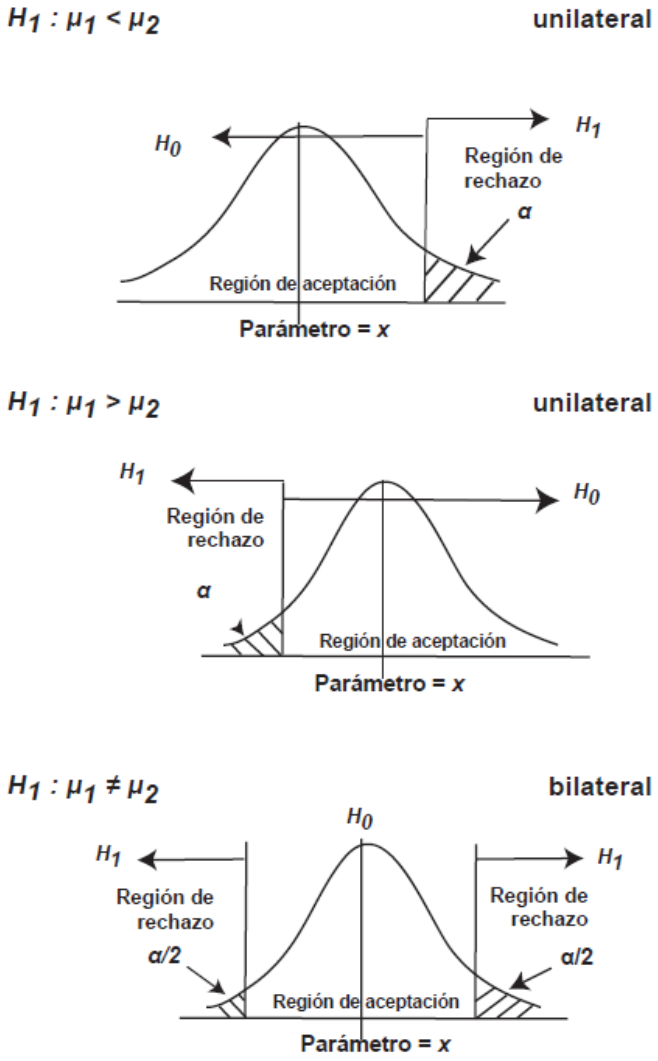
c) Pruebas de Hipótesis de una Cola o Dos Colas, llamado también Test Unilateral o Bilateral

Antes de aplicar todo test estadístico hay que definir el problema propuesto, así según las hipótesis formuladas, se habla de un test bilateral o unilateral. El test bilateral se utiliza cuando se tiene que definir entre dos estimaciones, o entre una estimación y un valor dado sin tener en cuenta el signo o sentido de la diferencia, es decir solo se utiliza si la Ho (hipótesis nula) señala la igualdad. El test unilateral se aplica cuando se necesita saber si una estimación H1 (hipótesis alternativa) es superior o inferior a otra. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Las determinaciones de las zonas de aceptación o de rechazo en un test unilateral y bilateral se muestran en la Ilustración 35:

Ilustración 35

Regiones de aceptación y rechazo de la hipótesis nula para test unilateral y bilateral



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

4.2.1. Pruebas de hipótesis para una media

- **Pruebas de hipótesis acerca de una media con varianza poblacional supuesta conocida:**

Sea X la media de una muestra aleatoria de tamaño n seleccionada de una población con media μ y varianza σ^2 supuestamente conocida.

Si la población es normal $N(\mu, \sigma^2)$, entonces, la distribución de la estadística X es exactamente normal $N(\mu, \sigma^2/n)$ para cualquier valor de n $(n \geq 2)$.Si la población no es normal, pero el tamaño de la muestra es suficientemente grande $(n \geq 30)$, entonces, la distribución de X es aproximadamente normal $N(\mu, \sigma^2/n)$. Entonces,

La estadística para la prueba acerca de μ con varianza σ^2 conocida es

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

Cuya distribución es exacta o aproximadamente normal estándar N(0,1), según sea la población normal o no.

Si se supone verdadera la hipótesis nula: $Ho: \mu = \mu o$, la estadística especificada por esta hipótesis es entonces:

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

a) Prueba bilateral o de dos colas

Si se prueba $Ho: \mu = \mu o$ contra $Ho: \mu \neq \mu o$, dado el nivel de significación $\propto$, en la distribución de $Z = (X - \mu o)/(\sigma/\sqrt{n})$, que es normal N(0,1), se determina el valor $Z1 - a/2$ tal que la probabilidad de rechazar Ho cuando se supone verdadera sea (Ilustración 36)

$$P(Z < -Z1{-}\propto/2) = \propto/2 \quad o \quad P(Z > Z1{-}\propto/2) = \propto/2$$

En consecuencia, la región critica en el rango de variación de Z es:

$$R.C. = (Z < -z1{-}\propto/2 \quad o \quad Z > z1{-}\propto/2)$$

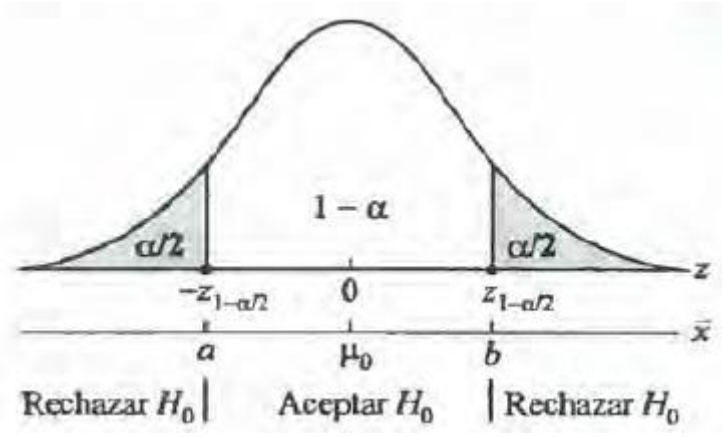
Por otro lado, la probabilidad de aceptar Ho cuando se supone verdadera es:

$$P(-z1{-}\propto/2 \leq Z \leq z1{-}\propto/2) = 1{-}\propto$$

Resultando la región de aceptación: R.A.=$(-z1{-}\propto/2 \leq Z \leq z1{-}\propto/2)$

Ilustración 36

Región critica bilateral en escalas z y X



Fuente: (Cordova Zamora, 1995)

Regla de decisión es: Si $Zk = (x - \mu o)/(\sigma/\sqrt{n})$ es un valor de Z obtenido de la muestra, entonces, se rechazará Ho con riesgo igual a $\propto$, si $x \in R.C.$ (o si $x \notin R.C.$ (o si $x \in R.A.$

No se rechazará Ho en caso contrario (Ilustración 36).

Si se rechaza Ho se dice que el valor Zk es significativo con un riesgo cuyo valor es $\propto$.

(Región critica en X)

Si se sustituye $Z = (X - \mu o)/(\sigma/\sqrt{n})$ en RC resulta la región critica en el rango de variación de X:

$$R.A = (a \leq X \leq b)$$

La regla de decisión es: Si x es el valor de x obtenido a partir de una muestra aleatoria, se rechazará Ho con un riesgo $\propto$, si $x \in R.C.$ ($o\ si\ x \notin R.A.$).

No se rechazará Ho en caso contrario (Ilustración 37).

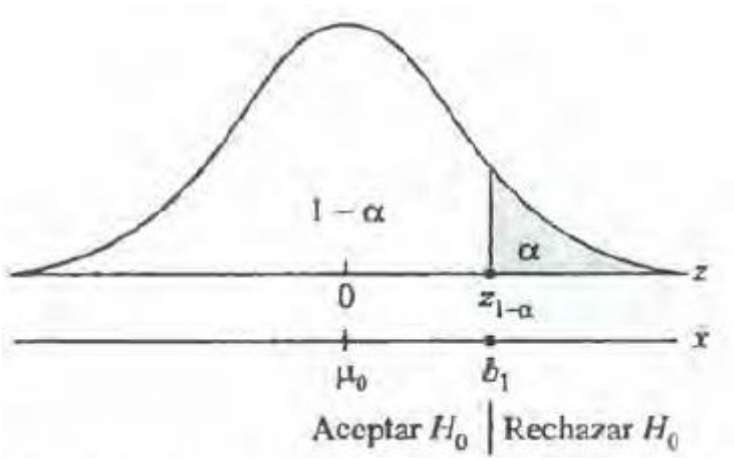
b) Prueba unilateral de cola a la derecha

Si se prueba $Ho: \mu = \mu o$, dado el nivel de significación $\propto$, en la distribución de $Z = (X - \mu o)/(\sigma/\sqrt{n})$ que es normal N(0,1), se determina el valor $Z1-\propto$ tal que (Ilustración 37), (Cordova Zamora, 1995)

$$P(Z > z1-\propto / H: \mu = \mu o\ verdadera) = \propto$$

Ilustración 37

Región critica cola a la derecha en escalas z y x



Fuente: (Cordova Zamora, 1995)

En consecuencia, la región critica en el rango de variación de Z es:

$$R.C = (Z > z1-\propto)$$

La región de aceptación es: $R.A = (Z \leq z1-\propto)$,

La regla de decisión es: Si $Zk = (x - \mu o)/(\sigma/\sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará Ho si $zk \in R.C.$ (o si $Zk \notin R.A.$).

No se rechazará Ho en caso contrario (Ilustración 37).

(Región critica en X)

Si se sustituye $Z = (X - \mu o)/(\sigma/\sqrt{n})$ en RC resulta la región critica en el rango de variación de X:

$$R.C = (X > bi)$$

Donde, $bi = \mu o + z1{-}\propto (\sigma/\sqrt{n})$

La región de aceptación es el intervalo:

$$R.A. = (X \leq bi)$$

La regla de decisión es: Siendo x el valor de x obtenido a partir de una muestra aleatoria de tamaño n, se rechazará Ho con un riesgo $\propto$, si $x \epsilon R.C.$ (o si $x \notin R.A$). No se rechazara Ho en caso contrario. (Ilustración 37). (Cordova Zamora, 1995)

c) Prueba unilateral de cola a la izquierda

Si se prueba $Ho: \mu = \mu o$ contra $Hi: \mu < \mu o$, dado el nivel de significación $\propto$, en la distribución de $Z = (X - \mu o)/(\sigma/\sqrt{n})$, se puede determinar el valor $z1{-}\propto$ tal que (Ilustración 38)

$$P(Z < -z1{-}\propto / H: \mu = \mu o\ verdadera) = \propto$$

En consecuencia, la región critica en el rango de variación de Z es:

$$R.C = (Z < -z1{-}\propto)$$

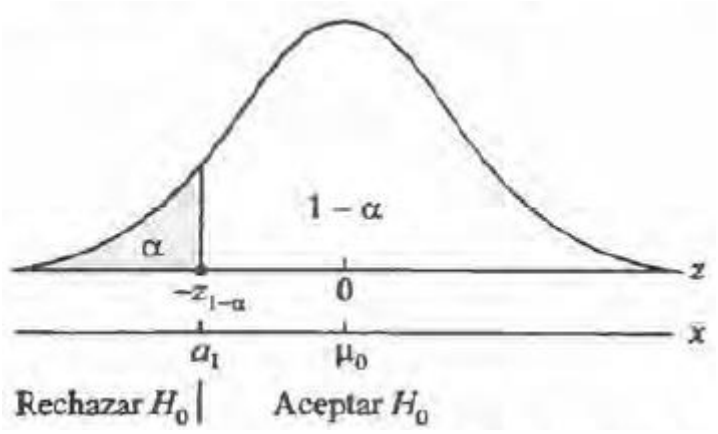
Región de aceptación es: $R.A = (Z \geq .z1{-}\propto)$.

Regla de decisión: Si $Zk = (x - \mu o)/(\sigma/\sqrt{n})$ es un valor de Z obtenido a partir de una muestra, se rechazará Ho con un riesgo $\propto$, si $x \epsilon R.C. (o\ si\ x\ \notin R.A.)$.

No se rechazará Ho en caso contrario (Ilustración 38). (Cordova Zamora, 1995)

Ilustración 38

Región critica cola a la izquierda en escalas z y x



Fuente: (Cordova Zamora, 1995)

(Región critica en X)

Si se sustituye $Z = (X - \mu o)/(\sigma/\sqrt{n})$ en RC se obtiene la región critica en el rango de variación de X:

$$RC = (X < a1)$$

Donde, $$a1 = \mu o - z1{-}\propto (\sigma/\sqrt{n})$$

Región de aceptación: $RA = (X \geq a1)$

Regla de decisión es: Si x es un valor de X obtenido a partir de una muestra aleatoria de tamaño n, se rechazará Ho con un riesgo $\propto$ si $x \in R.C.$ (o si $x \notin R.A.$).

No se rechazará Ho en caso contrario (Ilustración 38).

Ejemplo 1: Un determinado proceso de empaquetar un producto está controlado, si el peso medio del producto empaquetado es 400 gramos. Si en una muestra aleatoria de 100 paquetes del producto se ha encontrado que el peso medio es de 395 gramos. ¿Se podría concluir que el proceso está fuera de control al nivel de significación 5%?. Suponga que el peso de los productos empaquetados se distribuye normalmente con desviación estándar de 20 gramos.

Sea X la variable aleatoria definida como el peso de los paquetes del producto. Se supone que la distribución de X es $N(\mu, (20)^2)$.

1) Hipótesis: Ho ; $\mu = 400$ $(el\ proceso\ esta\ controlado)$.

   H1 $\mu \neq 400$ $(el\ proceso\ esta\ fuera\ de\ control)$.

2) Nivel de significación: $\propto = 0.05$.
3) Estadística: Población normal con varianza conocida, la estadística es

$$Z = \frac{X - \mu}{\sigma/\sqrt{n}}$$

Cuya distribución es normal N(0,1).

4) Región critica: Si se supone verdadera la hipótesis nula Ho, para $\propto = 0{,}5$ y la alternativa bilateral, en la distribución de $Z = (X - 400)/(20/\sqrt{100})$, se encuentra el valor critico:

$$Z1{-}\propto/2 = Z0{,}975 = 1{,}96$$

Luego, la región critica en la variable Z está dada por:

$$RC = (Z < -1{,}96\ o\ Z > 1{,}96)$$

5) Cálculos: De los datos se tiene:

$$n = 100, x = 395, \sigma = 20,$$

$$Zk = \frac{x - \mu o}{\sigma/\sqrt{n}} = \frac{395 - 400}{2} = -2.5.$$

6) Decisión: Puesto que Zk = $-2.5 \in R.C$, debemos rechazar Ho y concluir con un riesgo de 5%, que el proceso de empaquetar no está controlado.

En el rango de variación de X, la región critica es:

$$R.C = (X < 400 - 1.96x2 \quad o \quad X > 400 + 1.96x2)$$

$$R.C = (X < 396.08 \quad o \quad X > 403.92)$$

El hecho de $X = 395\ \epsilon\ R.C$, se debe rechazar Ho y concluir con un riesgo de 5%, que el proceso de empaquetar no está controlado.

(Regla de decisión en Intervalo de confianza)

La prueba bilateral de la hipótesis nula $Ho: \mu = \mu 0$ contra $H1: \mu \neq \mu 0$ a un nivel de significación dado $\propto$, equivale a calcular el intervalo de confianza (I.C.) de $(1-\propto)x100\%$ para el parámetro μ y luego rechazar la hipótesis nula $Ho: \mu = \mu 0$ si es que $\mu o \notin I.C$.

En efecto, si x es un valor de X, no se rechazará $Ho: \mu = \mu 0$ si el valor

$Zk \in R.A. = (-z1-\propto/2, z1-\propto/2)$, donde $Zk = (x - \mu o)/(\sigma/\sqrt{n})$

O, si $$-z1-\propto/2 \leq \frac{x-\mu o}{\frac{\sigma}{\sqrt{n}}} \leq z1-\propto/2$$

Esto es, no se rechazará $Ho: \mu = \mu 0$ si

$$x \in R.A. = (\mu o\ -\ z1-\propto/2\frac{\sigma}{\sqrt{n}}, \mu o + z1-\propto/2\frac{\sigma}{\sqrt{n}})$$

o equivalentemente si μo se encuentra dentro del intervalo de confianza (I.C.) del $(1-\propto)x100\%$ para μ:

$$x \in I.C. = (x\ -\ z1-\propto/2\frac{\sigma}{\sqrt{n}}, \mu o + z1-\propto/2\frac{\sigma}{\sqrt{n}})$$

Por tanto, se rechazará Ho con riesgo $\propto$ si,

$$x \notin R.A. \quad o\ si \quad \mu 0 \notin I.C$$

Por ejemplo, en el ejemplo 1, para $\propto = 0.05$ se tiene:

$$I.C. = (x\ -\ z1-\propto/2\frac{\sigma}{\sqrt{n}}, x + z1-\propto/2\frac{\sigma}{\sqrt{n}}) = (391.08.398.92)$$

$$R.A. = (\mu o\ -\ z1-\propto/2\frac{\sigma}{\sqrt{n}}, \mu o + z1-\propto/2\frac{\sigma}{\sqrt{n}}) = (396.08.403.92)$$

Dado que $\mu o = 400 \notin I.C.$ (o que $x = 395 \notin R.A.$) se debe rechazar Ho con un riesgo del 5%.

(Método del valor P en la prueba)

Otra forma de establecer la regia de decisión, en estadística aplicada, es calculando el valor P, a partir del valor absoluto de $Zk = (x - \mu o)/(\sigma/\sqrt{n})$, (que se obtiene de la muestra), de manera que

a) $P = P(Z < -zk) + P(Z > zk) = 2P(Z > zk)$ $\quad (para\ dos\ colas)$.
b) $P = P(Z < zk)$ $\quad (cola\ a\ la\ derecha)$.
c) $P = P(Z < -zk)(colaalaizquierda)$.

Si el valor de $P <\propto$, entonces se rechazara Ho. No se rechazará Ho, en caso contrario.

Las pruebas de hipótesis con el paquete estadístico MCEST contienen el método del valor P.

En el ejemplo 1, el valor absoluto de Zk es igual a 2.5, entonces,

$$P = 2P(Z > zk) = 2P(Z > 2.5) = 2(0.0062) = 0.0124$$

Dado que $P = 0.0124 < \propto = 0.05$, se debe rechazar Ho, con un riesgo $\propto = 0.05$. Observar que Zk es significativo en un valor mucho menor que $\propto = 0.05$ y que este valor de Zk solo ocurrirá en 124 casos de 10.000 experimentos.

Una región critica de tamaño 0.0124 es muy pequeña y, por lo tanto, es poco probable que se cometa error tipo I. (Cordova Zamora, 1995)

- **Pruebas de hipótesis acerca de una media con varianza poblacional supuesta desconocida:**

a) Población no normal: Si la población no tiene distribución normal y si la varianza es desconocida, para probar hipótesis acerca de la media μ, solo si, el tamaño de la muestra es grande $(n \geq 30)$, se suele utilizar la estadística:

$$Z = \frac{X - \mu o}{\sigma/\sqrt{n}}$$

Cuya distribución es aproximadamente N(0.1). La desviación estándar σ se estima puntualmente por s.

Luego, las regiones críticas de las pruebas de $Ho: \mu = \mu o$ contra cualquiera de las tres alternativas $H1: \mu > \mu o$ o $H1: \mu < \mu o$ o $H1: \mu \neq \mu o$ son las mismas (aproximadamente) de la (Prueba de hipótesis acerca de una media con varianza poblacional supuesta conocida).

b) Población normal: Si la población tiene distribución normal $N(\mu, \sigma^2)$, donde $\mu \; y \; \sigma^2$ son parámetros desconocidos, para $n \geq 2$ la estadística de la prueba acerca de la media μ es:

$$T = \frac{X - \mu}{S/\sqrt{n}}$$

Cuya distribución es t-Student con n-1 grados de libertad.

Si se supone verdadera la hipótesis nula: $Ho: \mu = \mu o$, la estadística especificada por esta hipótesis es entonces, ahora:

$$T = \frac{X - \mu o}{S/\sqrt{n}}$$

La estructura de la prueba es idéntica que en el caso de σ conocida, salvo que el valor de σ se estima por s y la distribución normal estándar se sustituye por la distribución t de Student con n-1 grados de libertad.

1) Prueba bilateral o de dos colas

Si se prueba $Ho: \mu = \mu o$ contra $H1: \mu \neq \mu o$, dado el nivel de significación $\propto$ en la distribución de $T = (X - \mu o)/(s/\sqrt{n}) - t(n-1)$, se determinan los valores $\pm t1{-}\propto /2, n-1$, tales que la probabilidad de rechazar Ho cuando se supone verdadera sea (Ilustración 39).

$$P(T < t1{-}\propto/2, n-1) = \propto/2 \qquad o \qquad P(T < t1{-}\propto/2, n-1) = \propto/2.$$

Luego, la región critica en el rango de variación de T es:

$$P(T < -t1{-}\propto/2, n-1) = \propto/2 \qquad o \qquad P(T > t1{-}\propto/2, n-1)$$

La región de aceptación es el intervalo

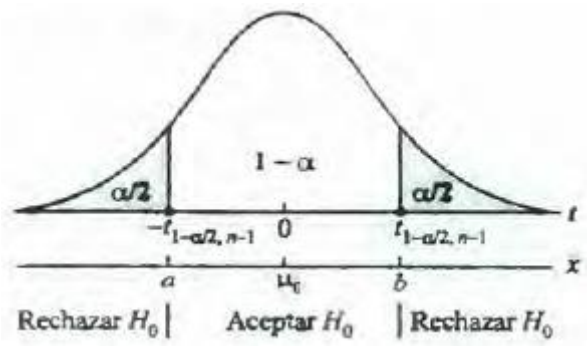
$$R.A = (-t1{-}\propto/2, n-1 \leq T \leq t1{-}\propto/2, n-1)$$

Regla de decisión: Se rechazará Ho con riesgo $\propto$, si $tk \in R.C.$ (o, si $tk \in R.A.$).

No se rechazará Ho en caso contrario. (Cordova Zamora, 1995)

Ilustración 39

Región critica bilateral en escalas t y x



Fuente: (Cordova Zamora, 1995)

Si se sustituye $T = (X - \mu o)/(s/\sqrt{n})$ en R.C se obtiene:

La región critica en el rango de variación de X:

$$R.C. = (X < a \quad o \quad X > b)$$

Donde $\propto = \mu o - t1{-}\propto/2, n-1(s/\sqrt{n}), \quad y \quad b = \mu o + t1 - \alpha/2.n-1(s/\sqrt{n})$

Regla de decisión: Siendo x el valor de X obtenido a partir de una muestra aleatoria de tamaño n, se rechazará Ho con un riesgo α. Si $x \in R.C.$ (o si $x \notin R.A. = (R.C.)^c$). No se rechazara Ho en caso contrario (Ilustración 39)

2) Prueba unilateral de cola a la derecha

Si se prueba $Ho: \mu = \mu o$ contra $H1: \mu > \mu o$, dado el nivel de significación $\propto$, en la distribución de $T = (X - \mu o)/(s/\sqrt{n}) - t(n-1)$, se determina el valor t1-$\alpha.n-1$ tal que: (Ilustración 40)

$$P(T > t1 - \alpha.n - 1/Ho: \mu = \mu o\ verdadera) = \alpha$$

Luego, la región critica en el rango de variación de T es:

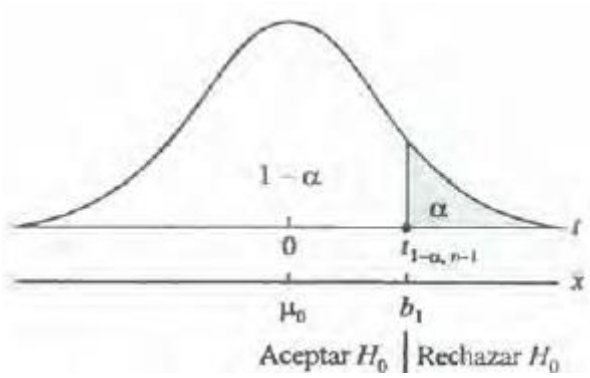
$$RC = (T > t1 - \alpha.n - 1)$$

La región de aceptación es el intervalo:

$$RA = (T > t1 - \alpha.n - 1)$$

Ilustración 40

Región critica cola a la derecha en escalas t y X



Fuente: (Cordova Zamora, 1995)

Regla de decisión: Se rechazará Ho si $tk \in R.C.$ (o si $tk \notin R.A.$). No se rechazará Ho en caso contrario.

La región critica en X (Ilustración 40) es: $RC = (X > b1)$, donde $b1 = X < \mu o + t1 - \alpha.n - 1(s/\sqrt{n})$

3) Prueba unilateral de cola a la izquierda

Si se prueba $Ho: \mu = \mu o$ contra $H1: \mu < \mu o$, dado el nivel de significación α, en la distribución de $T = (X - \mu o)/(s/\sqrt{n}) - t(n-1)$ se determina el valor $t1 - \alpha.n - 1$, tal que; (Ilustración 41)

$$P(T < -t1 - \alpha.n - 1/Ho: \mu = \mu o\ verdadera) = \alpha$$

Luego, la región critica en el rango de variación de T es:

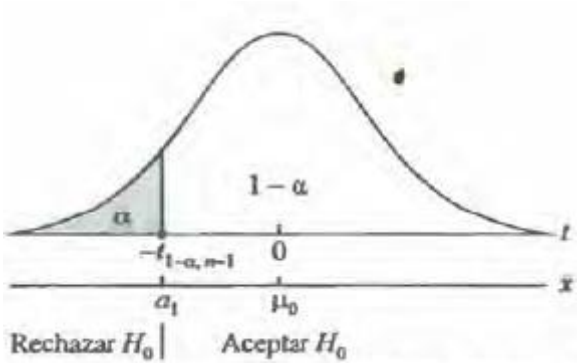
$$RC = (T < -t1 - \alpha.n - 1)$$

La región de aceptación es el intervalo:

$$RA = (T < -t1 - \alpha.n - 1)$$

Ilustración 41

Región critica cola a la izquierda en escalas t y X



Fuente: (Cordova Zamora, 1995)

Regla de decisión: Se rechazará Ho si $tk \in R.C.$ (o si $tk \notin R.A.$). No se rechazará Ho en caso contrario (Ilustración 41).

La región critica en X (Ilustración 41) es $RC = (X < \alpha 1)$, donde $\alpha 1 = X < \mu o - t1 - \alpha.n - 1(s/\sqrt{n})$

Ejemplo 2: Las cajas de cierto tipo de cereal procesadas por una fábrica deben tener un contenido promedio de 160 gramos. Por una queja ante el defensor del consumidor de que tales cajas de cereal tienen menos contenido, un inspector tomo una muestra aleatoria de 10 cajas encontrando los siguientes pesos de cereal en gramos:

$$157, 157, 163, 158, 161, 159, 162, 159, 158, 156$$

¿Es razonable que el inspector multe al fabricante? Utilice un nivel de significación del 5% y suponga que los contenidos tienen distribución normal.

Sea X la variable aleatoria que representa los pesos de las cajas del cereal. Se supone que la distribución X es normal con media μ y varianza σ^2 desconocidas.

i) Hipótesis. $Ho: \mu = 160$ $(No\ multa\ al\ fabricante)$

$$H1: \mu = 160\ (Multa\ al\ fabricante)$$

ii) Nivel de significación $\alpha = 0.05$

iii) Estadística: Población normal, con varianza desconocida y n = 10. Si $Ho: \mu = 160$ es verdadera, la estadística es

$$T = \frac{X - 160}{s/\sqrt{n}}$$

Que se distribuye según una t-Student con 9 grados de libertad.

iv) Región critica: Con el nivel de significación $\alpha = 0.05$ y para una prueba de hipótesis unilateral cola a la izquierda, en la tabla de probabilidades de t-Student se encuentra: $t0.95.9 = 1.833$.

Consecuentemente, la región critica es: $RC = (T < -1.833)$

v) Cálculos: De los datos de la muestra se obtiene:

$$n = 10, x = 159, s = 2.309, error\ estandar: s/\sqrt{n} = 0.73$$

$$tk = \frac{x - 160}{s/\sqrt{n}} = \frac{159 - 160}{0.73} = -1.37.$$

vi) Decisión: Dado que $tk = -1.37 \notin R.C$, debemos aceptar Ho y concluir que la media de los ingresos quincenales no ha variado.

Utilizando el paquete de cómputo MCEST, se encuentra la probabilidad $P = P(T < -1.37) = 0.1012$, por lo que debemos aceptar Ho. (Cordova Zamora, 1995)

4.2.2. Pruebas de hipótesis para dos medias

- **Pruebas de hipótesis acerca de dos medias con varianza poblacional supuesta conocida:**

Sean X1 y X2 las medias de dos muestras aleatorias independientes de tamaños n1 y n2 seleccionadas respectivamente de dos poblaciones independientes, con medias $\mu 1$ y $\mu 2$ y varianzas $\sigma 1^2$ y $\sigma 2^2$ respectivas supuestas conocidas.

Si las dos poblaciones son normales, entonces, las estadísticas X1 y X2 tienen respectivamente distribución normal $N(\mu 1. \sigma 1^2/n1)$ y $N(\mu 2. \sigma 2^2/n2)$ para $n1 \geq 2$, y $n2 \geq 2$. Luego, la estadística X1 – X2 tiene distribución exactamente normal $N(\mu 1 - \mu 2. \sigma 1^2/n1 + \sigma 2^2/n2)$.

Luego, según sean las dos poblaciones normales o no, la estadística

$$Z = \frac{X1 - X2 - (\mu 1 - \mu 2)}{\sqrt{\sigma 1^2/n1 + \sigma 2^2/n2}}$$

Tiene distribución exactamente o aproximadamente normal N(0,1).

Si suponemos verdadera la hipótesis nula $Ho: \mu 1 = \mu 2 \quad o \quad \mu 1 - \mu 2 = 0$, la estadística es:

$$Z = \frac{X1 - X2}{\sqrt{\sigma 1^2/n1 + \sigma 2^2/n2}} \sim N(0,1)$$

Su valor $Z = \frac{X1 - X2}{\sqrt{\sigma 1^2/n1 + \sigma 2^2/n2}}$ que resulta de dos muestras, se utiliza para probar Ho contra cualquiera de las hipótesis alternativas $H1: \mu 1 \neq \mu 2 \quad o \quad H1: \mu 1 - \mu 2 \quad o \quad H1: \mu 1 < \mu 2$.

La estructura de la prueba es similar a los casos descritos en la (una media), usando la distribución de Z.

1) Prueba bilateral o de dos colas

Si se prueba $Ho: \mu 1 = \mu 2$ contra $H1: \mu 1 \neq \mu 2$, la región critica en el rango de variación de Z es:

$$R.C = (Z < z1 - \alpha/2 \quad o \quad Z > 1 - \alpha/2)$$

2) Prueba unilateral de cola a la derecha

Si se prueba $Ho: \mu1 = \mu2$ contra $H1: \mu1 > \mu2$, la región critica en la variación de Z es,

$$R.C = (Z > z1 - \alpha)$$

3) Prueba unilateral de cola a la derecha

Si se prueba $Ho: \mu1 = \mu2$ contra $H1: \mu1 > \mu2$, la región critica en los valores de Z es:

$$R.C = (Z > -z1 - \alpha)$$

Cuando las hipótesis son de la forma

a) $Ho: \mu1 - \mu2 = do\ contra\ h1: \mu1 - \mu2 \neq do$
b) $Ho: \mu1 - \mu2 = do\ contra\ h1: \mu1 - \mu2 > do$
c) $Ho: \mu1 - \mu2 = do\ contra\ h1: \mu1 - \mu2 < do$

La estadística de la prueba es,

$$Z = \frac{(X1 - X2) - do}{\sqrt{(\sigma1^2/n1) + (\sigma2^2/n2)}}$$

Cuya distribución es exactamente o aproximadamente normal N(0,1), según sean las dos poblaciones normales o no.

Ejemplo 1: Un fabricante quiere comparar dos marcas de máquinas, A y B: para fabricar un tipo de articulo. Observa dos muestras aleatorias de 60 artículos procesados por A y B respectivamente y encuentra que las medias respectivas son 1,230 y 1,190 segundos.

Suponga $\sigma1 = 120 \quad y \quad \sigma2 = 90\ segundos.$

a) Al nivel de significación del 5%, ¿se puede inferir que la maquina B es más rápida que la maquina A?
b) Al nivel de significación del 5%, ¿se puede inferir que la media de B es menor que la media de A en menos de 7 segundos?
c) ¿En cuánto deberían incrementarse los tamaños de las muestras de cada proceso para que la diferencia observada de 40 segundos en los tiempos promedios muestrales de A menos B sea significativa al nivel $\alpha = 1\%$?

Sean X1 y X2 los tiempos de proceso con las maquinas A y B respectivamente y $\mu1\ y\ \mu2$ sus medias respectivas.

Se desconocen las distribuciones de probabilidades de X1 y X2, pero las muestras son grandes.

1) Hipótesis: $Ho: \mu1 = \mu2$ contra $H1: \mu1 > \mu2$
2) Nivel de significación: $\alpha = 0.05$
3) Estadística: Si se supone verdadera la hipótesis Ho y para muestras grandes, la estadística apropiada es:

$$Z = \frac{X1 \ - \ X2}{\sqrt{(\sigma 1^2/n1) + (\sigma 2^2/n2)}}$$

Cuya distribución es aproximadamente normal estándar N(0,1).

4) Región critica: Para $\alpha = 0.05$ y una prueba unilateral de cola a la derecha, en la distribución de Z se encuentra el valor $z0.9500 = 1{,}645$. Luego, la región critica es
5) Cálculos: De los datos se tiene

$$n1 = n2 = 60. \qquad x1 = 1{,}230. x2 = 1190 \qquad \sigma 1 = 120 \ y \ \sigma 2 = 90$$

$$E.S = Error\ estandar = \sqrt{(\sigma 1^2/n1) + (\sigma 2^2/n2)} = 19.365$$

$$Zk = \frac{x1 - x2}{ES} = \frac{1{,}230 - 1{,}190}{19{,}365} = 2{,}07$$

6) Decisión: Ya que $Zk = 2.07 \in R.C.$, debemos rechazar Ho y concluir que el equipo B utiliza menor tiempo en el proceso de fabricación.

Se debe probar $Ho: \mu 1 - \mu 2 = 7$ contra $H1: \mu 1 - \mu 2 > 7$.

Si Ho es verdadera, la estadística de la prueba es

$$Z = \frac{(X1 \ - \ X2) - 7}{\sqrt{(\sigma 1^2/n1) + (\sigma 2^2/n2)}} - N(0{,}1)$$

La región critica de la prueba unilateral cola derecha al nivel $\alpha = 0.05$ es: la misma del caso (hipótesis)

$$R.C = (Z > 1.645)$$

$$Zk = \frac{(X1 \ - \ X2) - 7}{ES} = \frac{(1230 - 1190) - 7}{19.365} = 1.7$$

Ya que $Zk = 1.7 \ \epsilon \ R.C.$, debemos rechazar Ho y concluir que la maquina B utiliza un tiempo promedio menos de 7 segundos debajo del tiempo promedio de A.

Sea n el tamaño de cada una de las dos muestras tomadas de los artículos procesados por las maquinas A y B.

Si la hipótesis nula Ho es verdadera, y si además.

$$x1 - x2 = 1230 - 1190 = 40$$

se tiene $Zk = \frac{40}{\sqrt{\frac{120^2+90^2}{n}}} = 0.27\sqrt{n}.$

Para $\alpha = 0.01$ y una prueba unilateral de cola a la derecha, en la distribución de Z de halla:

$$Z1 - \alpha = Z0.99 = 2.33$$

La región critica es: $R.C = (Z > 2.33)$

La diferencia observada de 40 en las medias de las muestras será significativa al nivel de 1%, si

$$0.27\sqrt{n} \in R.C.$$

Esto es, si $0.27\sqrt{n} > 2.33, \quad \sqrt{n} > 8.63, \quad n \geq 75.$

De aquí que se debe incrementar cada muestra en al menos: 75-60 = 15 casos. (Cordova Zamora, 1995)

- **Pruebas de hipótesis acerca de dos medias con varianza poblacional supuesta desconocida:**

a) Poblaciones no normales:

Si las dos muestras aleatorias independientes de tamaños n1 y n2 se seleccionan respectivamente de dos poblaciones cuyas distribuciones son no normales con varianzas $\sigma 1^2$ y $\sigma 2^2$ supuestas desconocidas, entonces, siempre que los tamaños de las muestras sean grandes; $n1 \geq 30$ y $n2 \geq 30$ los parámetros $\sigma 1\ y\ \sigma 2$ se estiman respectivamente por $s1\ y\ s2$.

Para probar la hipótesis nula $Ho: \mu 1 - \mu 2 = 0$ contra una alternativa bilateral o unilateral, se utiliza la estadística:

$$Z = \frac{(X1\ -\ X2) - (\mu 1 - \mu 2)}{\sqrt{(s1^2/n1) + (s2^2/n2)}}$$

Que se distribuye aproximadamente normal N(0,1).

Las regiones críticas y las reglas de decisión para las pruebas de la hipótesis nula $Ho: \mu 1 - \mu 2 = 0$ (o $Ho: \mu 1 - \mu 2 = do$) contra una alternativa unilateral o bilateral son las mismas del método con varianzas conocidas.

b) Poblaciones normales

Sean X1 y X2 las medias y $S1^2\ y\ S2^2$ las varianzas de dos muestras aleatorias independientes de tamaños n1 y n2 respectivamente seleccionadas de dos poblaciones normales con medias $\mu 1\ y\ \mu 2$ y varianzas $\sigma 1^2\ y\ \sigma 2^2$ desconocidas.

i) Varianzas desconocidas supuesta iguales: $\sigma 1^2 = \sigma 2^2 = \alpha^2$

Si las poblaciones son normales, independientes, y con varianzas desconocidas supuestas iguales, entonces, la estadística

$$T = \frac{(X1 - X2) - (\mu 1 - \mu 2)}{\sqrt{\frac{Sc^2}{n1} + \frac{Sc^2}{n2}}}$$

Tienen distribución t-student con $n1 + n2 - 2$ grados de libertad, en donde la varianza común:

$$Sc^2 = \frac{(n1 - 1)S1^2 + (n2 - 1)S2^2}{n1 + n2 - 2}$$

Es un estimador insesgado de la varianza común σ^2.

Si la hipótesis nula $Ho: \mu 1 = \mu 2$ es verdadera, entonces, la estadística

$$T = \frac{X1 - X2}{\sqrt{\frac{Sc^2}{n1} + \frac{Sc^2}{n2}}} - t(n1 + n2 - 2)$$

Su valor $tk = \frac{x1-x2}{\sqrt{\frac{(n1-1)S1^2+(n2-1)S2^2}{n1+n2-2}(\frac{1}{n1}+\frac{1}{n2})}}$

Que resulta de dos muestras aleatorias, se utiliza para probar Ho contra una alternativa unilateral o bilateral.

La estructura de la prueba es similar a los casos descritos (una media) usando la distribución de t.

1) Prueba bilateral o de dos colas

Si se prueba $Ho; \mu 1 = \mu 2$ contra $H1; \mu 1 \neq \mu 2$, la región critica es el intervalo:

$$R.C. = (T < -t1 - \alpha/2.n1 + n2 - 2 \quad o \quad T > t1 - \alpha/2.n1 + n2 - 2)$$

2) Prueba unilateral de cola a la derecha

Si se prueba $Ho; \mu 1 = \mu 2$ contra $H1; \mu 1 > \mu 2$, la región critica es el intervalo:

$$R.C. = (T < -t1 - \alpha/2.n1 + n2 - 2)$$

3) Prueba unilateral de cola a la izquierda

Si se prueba $Ho; \mu 1 = \mu 2$ contra $H1; \mu 1 < \mu 2$, la región critica es el intervalo:

$$R.C. = (T < -t1 - \alpha/2.n1 + n2 - 2)$$

ii) Varianzas desconocidas supuestas distintas $\sigma 1^2 \neq \sigma 2^2$

Si las varianzas de las dos poblaciones normales independientes son desconocidas supuestas diferentes, entonces, la estadística

$$T = \frac{(X1 - X2) - (\mu 1 - \mu 2)}{\sqrt{S1^2/n1 + S2^2/n2}}$$

Tienen distribución t-student con r grados de libertad, siendo

$$r = \frac{(s1^2/n1 + s2^2/n2)^2}{\frac{(S1^2/n1)^2}{n1 - 1} + \frac{(S2^2/n2)^2}{n2 - 1}}$$

Dado que r rara vez es un entero, se redondea al entero más cercano.

Si la hipótesis nula $Ho: \mu 1 = \mu 2$ se supone verdadera. Entonces

$$T = \frac{X1 - X2}{\sqrt{S1^2/n1 + S2^2/n2}} \sim t(r)$$

Su valor $tk = \frac{x1-x2}{\sqrt{S1^2/n1+S2^2/n2}}$ que resulta de dos muestras aleatorias independientes, se utiliza para probar Ho contra una alternativa unilateral o bilateral.

Las regiones críticas y las reglas de decisión son similares a los del caso de (Varianzas desconocidas supuestas iguales), pero con r grados de libertad.

Ejemplo 2: Una medicina A es aplicada a 10 pacientes aquejados de cierta enfermedad. Otra medicina B es aplicada a otros 9 pacientes aquejados de la misma enfermedad. Los tiempos de recuperación de los pacientes, en días, fueron los siguientes:

Medicina A: 6, 5, 6, 7, 4, 7, 6, 4, 3, 6.

Medicina B: 7, 6, 7, 9, 5, 8, 7, 6, 8.

Utilizando un nivel de significación del 5% y suponiendo poblaciones normales.

i) ¿Se podría concluir que las varianzas poblacionales son iguales?
ii) ¿Se puede aceptar la hipótesis nula que son iguales las medias de los tiempos de tratamiento de las dos medicinas?
iii) ¿Cuál de las medicinas es más eficaz?

Sean X1 y X2 las variables aleatorias que representan los tiempos en días de tratamiento de las medicinas A y B respectivamente. Se supone que $X1 \sim N(\mu 1.\sigma 1^2)\ y\ X2 \sim N(\mu 2.\sigma 2^2)$.

a) Prueba de la homogeneidad de varianzas
1) Hipótesis: $Ho: \sigma 1^2 = \sigma 2^2 \quad contra \quad H1: \sigma 1^2 \neq \sigma 2^2$
2) Nivel de significación: $\alpha = 0.05$
3) Estadística: Poblaciones normales. Suponiendo verdadera la hipótesis nula Ho, para $n1 = 10\ y\ n2 = 9$, la estadística de la prueba es:

$$F = \frac{S1^2}{S2^2}$$

Que se distribuye como F(9,8).

4) Región critica. Para $\alpha = 0.05$ y una prueba bilateral, en la distribución de F(9,8) se encuentran:

$$f0.975.9.8 = 4.36 \quad y \quad f0.025.9.8 = 1/f0.975.8.9 = 1/4.10 = 0.244$$

Luego, la región critica está dada por:

$$R.C = (F < 0.244\ \ o\ \ F > 4.36)$$

5) Cálculos. De los datos de la muestra se obtiene:

$$S1^2 = 1.822 \quad S2^2 = 1.5 \quad y \quad fk = S1^2/S2^2 = 1.215$$

6) Decisión. Como $fk = 1.215 \notin R.C$. Se debería aceptar Ho y concluir que las varianzas de los tiempos de A y B son iguales.

b) Prueba de la diferencia de las dos medias
1) Hipótesis: $Ho: \mu 1^2 = \mu 2^2 \quad contra \quad H1: \mu 1^2 \neq \mu 2^2$

2) Nivel de significación: $\alpha = 0.05$.
3) Estadística de la prueba: Si se supone Ho verdadero y dado que hay prueba de que las varianzas poblacionales son iguales, la estadística apropiada es:

$$T = \frac{X1 - X2}{\sqrt{\frac{Sc^2}{n1} + \frac{Sc^2}{n2}}}$$

Que se distribuye según un t-Student con n1+n2-2-17 grados de libertad.

4) Región critica: Para $\alpha = 0.05$ y una prueba de hipótesis bilateral, en la distribución t(17) se encuentra $t1 - \alpha/2.n1 + n2{\sim}2 = t0.975.17 = 2.110$.

La región critica en la variación de T es

$$R.C = (T < -2.110 \; o \; T > 2.110)$$

5) Cálculos. De los datos se tiene:

$$n1 = 10. \quad x1 = 5.4 \quad s1^2 = 1.822, \qquad n2 = 9. x2 = 7.0 S2^2 = 1.5$$

$$Sc = \frac{(n1 - 1)S1^2 + (n2 - 1)S2^2}{n1 + n2 - 2} = \frac{9(1.822) + 8(1.5)}{10 + 9 - 2} = 1.67$$

$$Error\ Estandar = \sqrt{\frac{Sc^2}{n1} + \frac{Sc^2}{n2}} = 0.594$$

$$tk = \frac{x1 - x2}{\sqrt{\frac{Sc^2}{n1} + \frac{Sc^2}{n2}}} = \frac{5.4 - 7.0}{0.594} = -2.694$$

6) Decisión: Ya que $tk = -2.694 \in R.C.$, debemos rechazar Ho y concluir que los promedios de los tiempos de tratamientos con las medicinas A y B son diferentes.

c) Como las medias de las dos poblaciones son diferentes, planteamos las hipótesis:

$$Ho: \mu 1 = \mu 2 \; (Ambas\ medicinas\ son\ iguales).$$

$$H1: \mu 1 < \mu 2 \; (Medicina\ A\ es\ mejor\ que\ B).$$

Con $\alpha = 0.05$ y 17 grados de libertad, para la prueba unilateral de cola a la izquierda se encuentra el valor critico: $t0.95.17 = 1.740$. Luego, la región critica es:

$$R.C. = (T < -1.740)$$

Como $tk = -2.694 \in R.C.$, debemos rechazar Ho y concluir que la medicina A mas eficaz que la medicina B.

Con el paquete MCEST para la prueba de varianzas se obtiene: $P(F > 1.215) = 0.397$. Dado que 0.397>0.05 se infiere que las varianzas poblacionales son iguales al nivel 5%.

También para la prueba de dos medias se obtiene P(T>2.694)=0.007. Luego, 0.007 es la significación para una prueba unilateral o bilateral al 5%. (Cordova Zamora, 1995)

4.2.3. Pruebas de hipótesis para tres o más medias

Sea X1,X2,….,Xn una muestra aleatoria de tamaño n, seleccionada de una población normal con media μ y varianza σ^2, parámetros desconocidos, y sea la varianza muestral,

$$S^2 = \frac{\sum_{i=1}^{n}(Xi\ -\ X)^2}{n-1}$$

Entonces, la variable aleatoria,

$$X = \frac{(n-1)S^2}{\sigma^2}$$

Tiene distribución chi-cuadrado con n-1 grados de libertad. Esta estadística se utiliza para probar hipótesis acerca de una varianza.

Si se supone verdadera la hipótesis nula Ho:$\sigma^2 = \sigma o^2$, la estadística es:

$$X = \frac{(n-1)S^2}{\sigma o^2} \sim X^2(n-1)$$

Su valor $Xk = \frac{(n-1)S^2}{\sigma o^2}$ que resulta de la muestra aleatoria, se utiliza para la prueba de Ho, contra una alternativa unilateral o bilateral.

1) Prueba bilateral o de dos colas

Si se prueba $Ho: \sigma^2 = \sigma o^2$ contra $H1: \sigma^2 \neq \sigma o^2$, dado un nivel de significación α, en la distribución $x^2(n-1)$ se determinan los valores $x\alpha^2/2, n-1$ y $x1^2 - \alpha/2, n-1$ (Ilustración 42) tales que la probabilidad de rechazar la hipótesis nula Ho cuando realmente es verdadera es igual a:

$$P(X < X\alpha^2/2.n-1) = \alpha/2 \quad o \quad P(X < X1^2 - \alpha/2.n-1) = \alpha/2.$$

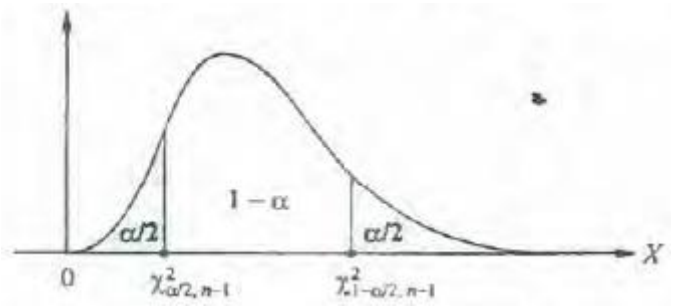
La región critica de la prueba, es entonces.

$$R.C. = (X < X\alpha^2/2.n-1 \quad o \quad X > X1^2 - \alpha/2.n-1).$$

La regla de decisión es rechazar Ho con un riesgo α, si $Xk \in R.C.$ (o si $Xk \notin R.A. = (X\alpha^2/2.n-1.X1^2 - \alpha/2.n-1)$). No rechazar Ho en caso contrario.

Ilustración 42

Región critica para la prueba de Ho:$\sigma^2 = \sigma o^2$contra $H1: \sigma^2 \neq \sigma o^2$



Fuente: (Cordova Zamora, 1995)

2) Contraste unilateral de cola a la derecha

Si se prueba Ho: $\sigma^2 = \sigma o^2$ contra H1: $\sigma^2 > \sigma o^2$, dado un nivel de significación α, en la distribución $x^2(n-1)$ se determina el valor $x1^2 - \alpha.n-1$ (Ilustración 43) tal que la probabilidad de rechazar la hipótesis nula Ho cuando realmente es verdadera es igual a:

$$P(X > x1^2 - \alpha.n-1) = \alpha$$

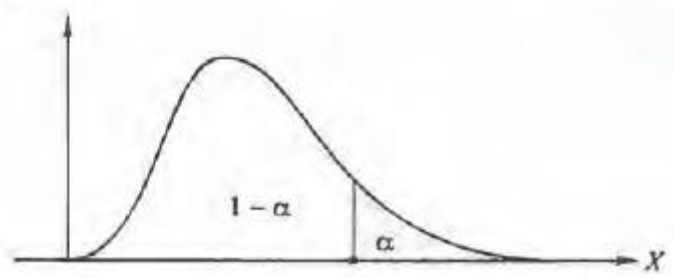
Luego, la región critica es:

$$R.C. = (X > x1^2 - \alpha.n-1).$$

La regla de decisión es: rechazar Ho al nivel α si $Xk \in R.C.$ (o si $Xk \notin R.A. = (X \leq x1^2 - \alpha.n-1)$). No rechazar Ho, en caso contrario. (Cordova Zamora, 1995)

Ilustración 43

Región critica para la prueba de Ho:$\sigma^2 = \sigma o^2$contra $H1: \sigma^2 > \sigma o^2$



Fuente: (Cordova Zamora, 1995)

3) Contraste unilateral cola o la izquierda

Si la prueba es de Ho: $\sigma^2 = \sigma o^2$ contra H1: $\sigma^2 < \sigma o^2$, dado un nivel de significación α, en la distribución $x^2(n-1)$ se determina el valor $x1^2 - \alpha.n-1$ (Ilustración 44) tal que la probabilidad de rechazar la hipótesis nula Ho cuando realmente es verdadera es igual a:

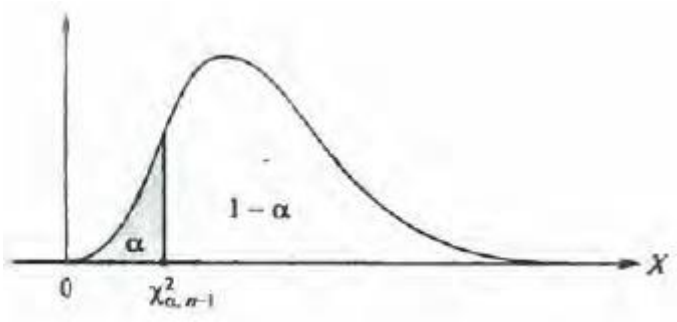
$$P(X > x\alpha^2.n-1) = \alpha$$

Luego, la región critica es:

$$R.C. = (X > x\alpha^2.n - 1).$$

Ilustración 44

Región critica para la prueba de Ho:$\sigma^2 = \sigma o^2$*contra* $H1: \sigma^2 < \sigma o^2$



Fuente: (Cordova Zamora, 1995)

Regla de decisión: rechazar Ho si $Xk \in R.C.(o\ si\ Xk \notin R.A. = (X \geq X\alpha^2.n - 1))$. No rechazar Ho en caso contrario.

Ejemplo 1: En un proceso de fabricación, se plantea la hipótesis que la desviación estándar de las longitudes de cierto tipo de tornillo es 2.0 mm. En una muestra de diez tornillos elegidos al azar del proceso de producción se han encontrado las siguientes longitudes en milímetros:

$$71,66,64,72,69,67,70,68,65,69.$$

Con estos datos, ¿se justifica la suposición que la desviación estándar verdadera es 2.00 mm?

Use el nivel de significación $\alpha = 0.05$, y suponga que la distribución de las longitudes es normal.

1) Hipótesis: $Ho: \sigma^2 = 4$ contra $H1: \sigma^2 \neq 4$
2) Nivel de significación: $\alpha = 0.05$.
3) Estadística: Población normal, con n = 10, y suponiendo verdadera la hipótesis $Ho: \sigma^2 = 4$, la estadística

$$X = \frac{(n-1)S^2}{4}$$

Se distribuye como chi-cuadrado con 9 grados de libertad.

4) Región critica: Para $\alpha = 0.05$ y para contraste bilateral, en la tabla chi cuadrado se encuentran los siguientes valores críticos:

$$X\alpha^2/2.n - 1 = x^2 0.025.9 = 2.70$$

$$X1^2 - \alpha/2.n - 1 = x^2 0.975.9 = 19.02$$

Luego, la región critica es: $R.C. = (X < 2.70 o X > 19.02)$.

5) Cálculos: De los datos de la muestra resulta $S^2 = 6.77$, entonces,

$$Xk = \frac{(n-1)s^2}{4} = \frac{9s^2}{4} = \frac{9(6.77)}{4} = 15.23,$$

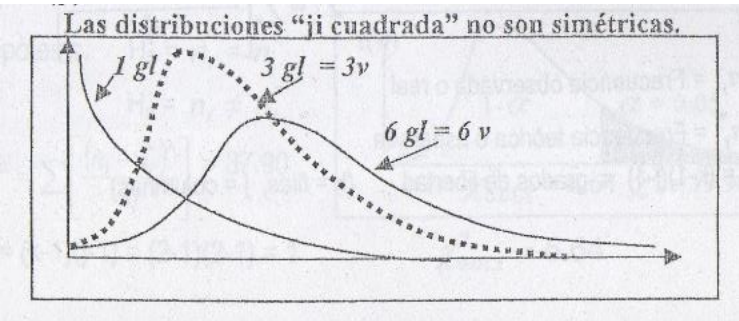
6) Decisión: Como $Xk = 15.23 \notin R.C.$ no se debe rechazar Ho y concluimos que la varianza de la población es igual a 4 gr^2. (Cordova Zamora, 1995)

4.2.4. Pruebas de hipótesis para variables categóricas

Esta distribución fue introducida por F.R.Helmert en 1876 y redescubierta en 1900 por Kart Pearson. Tiene muchos usos importantes, incluyendo ensayos de hipótesis acerca de proporciones y cálculo de intervalos de confianza para varianzas. Hay una distribución ji cuadrada diferente según el valor de n-1, lo cual representa los grados de libertad (gl), Así:

Ilustración 45

Las distribuciones "ji cuadrada" no son simétricas



Fuente: (Alvarez Roman, 2004)

Cuando gl es grande (v > 30), la distribución ji cuadrada se aproxima a la distribución normal. La variable $\sqrt{2x^2}$ es asintóticamente normal con media $\sqrt{2v-1}$ y varianza 1.

La curva está dada por: $Y = C(x^2)^{\frac{v-2}{2}}.e^{-\frac{x^2}{2}}$

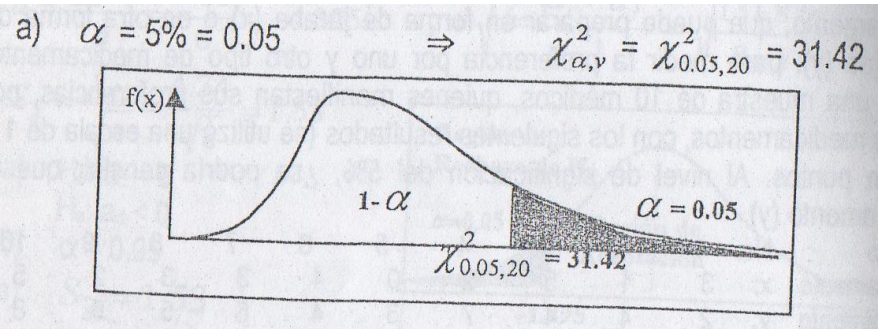
Donde $v = n - 1 = grados\ de\ libertad$

$C = constante\ que\ depende\ de\ v, para\ que\ el\ area\ bajo\ la\ curva\ sea$ 1.

a) Como leer en la tabla

Se busca en la primera fila Xv^2 y en la primera columna gl, en la intersección de la fila y la columna correspondiente se encuentra el valor de x^2 correspondiente. (Alvarez Roman, 2004)

Ejemplo 1: Si se tiene una variable aleatoria que sigue una distribución x^2 con grados de libertad, obtener $x^2\alpha, v$ para:



Proceso para la prueba Ji-cuadrada

1) Formular la hipótesis
2) Establecer las diferencias entre las frecuencias observadas y las esperadas, se eleva la cuadrado y se divide cada una de ellas para la frecuencia teórica esperada.
3) Se suma y se obtiene Ji-cuadrada

Ecuación sin corregir:

$$x^2 = \sum (\frac{(n1 - n1*)^2}{n1*})$$

Ecuación con corrección de Yates:

$$x^2 = \sum \left(\frac{\left((n1 - n1*) - 0{,}5\right)^2}{n1*} \right)$$

Donde: ni = frecuencia observada o real

ni*=Frecuencia teórica o esperada

$$v = (k-1)(j-1) = grados\ de\ libertad \qquad (k = filas, j = columnas)$$

La corrección de yates se utiliza cuando la tabla es de 2x2, es decir, v=1 y la variable es discreta. En muestra grandes se obtienen prácticamente los mismos resultados. La corrección de Yates hoy es muy poco utilizada por cuanto se ha demostrado que, en la mayoría de casos la hipótesis nula no se rechaza. (Alvarez Roman, 2004)

1) Durante una epidemia se obtuvieron los siguientes datos sobre la efectividad de una vacuna como medida preventiva para los médicos. Estos datos, ¿indican la efectividad de la vacunación con base en el nivel de significación del 1%?

nivel de significación del 1%?

TRATAMIENTO	ENFERMOS	NO ENFERMOS	TOTAL
Vacunados	192	4	196
No vacunados	113	34	147
TOTAL	305	38	343

a) Calculamos: $ni* = n.p$

$$n1 *= 305(\frac{196}{343}) = 174.28$$

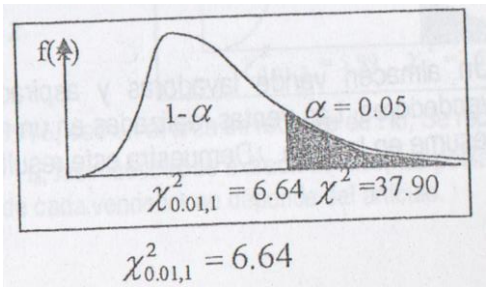
$$n2 *= 305(\frac{147}{343}) = 130.71$$

$$n3 *= 38(\frac{196}{343}) = 21.71$$

$$n4 *= 38(\frac{147}{343}) = 16.29$$

b) Calculamos x^2

n_i	n_i^*	$n_i - n_i^*$	$(n_i - n_i^*)^2$	$\frac{(n_i - n_i^*)^2}{n_i^*}$
192	174.28	17.72	313.99	1.802
113	130.71	-17.71	313.64	2.399
4	21.71	-17.71	313.64	14.45
34	16.29	17.71	313.64	19.25
343				37.90
n				$\chi^2 = \sum\left[\frac{(n_i - n_i^*)^2}{n_i^*}\right]$



$\chi^2_{0.01,1} = 6.64$

c) Hipótesis.

$$Ho = ni = ni *$$

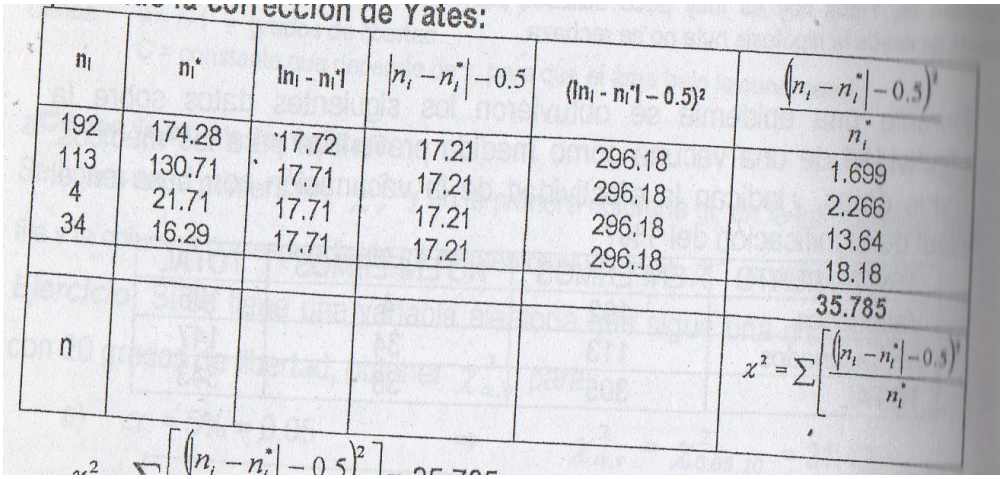
$$Ha = ni \neq ni *$$

d) $x^2 = \sum\left(\frac{(ni-ni)^2}{ni*}\right) = 37.90$

e) v = $(k-1)(j-1) = (2-1)(2-1) = 1$

f) Decisión: Como $x^2 = 37.90$, cae en el área de rechazo de Ho. Se rechaza la hipótesis nula y se acepta $Ha = ni \neq ni *$. Es decir, la diferencia es significativa.

Aplicando la corrección de Yates:



corrección de Yates:

n_i	n_i^*	$\lvert n_i - n_i^* \rvert$	$\lvert n_i - n_i^* \rvert - 0.5$	$(\lvert n_i - n_i^* \rvert - 0.5)^2$	$\frac{(\lvert n_i - n_i^* \rvert - 0.5)^2}{n_i^*}$
192	174.28	17.72	17.21	296.18	1.699
113	130.71	17.71	17.21	296.18	2.266
4	21.71	17.71	17.21	296.18	13.64
34	16.29	17.71	17.21	296.18	18.18
					35.785
n					$\chi^2 = \sum \left[\frac{(\lvert n_i - n_i^* \rvert - 0.5)^2}{n_i^*} \right.$

$$x^2 = \sum \left(\frac{((ni - ni) - 0.5)^2}{ni *} \right) = 35.785 \qquad (se\ llega\ a\ la\ misma\ conclusion)$$

4.3. Regresión y Correlación

Al analiza los datos para las ciencias, con frecuencia resulta que es conveniente saber la relación entre 2 variables. Por ejemplo, la relación entre la presión de la sangre y la edad, o la estatura y el peso, o entre la concentración de un medicamento inyectado y la rapidez de los latidos del corazón. El consumo de un nutriente y su ganancia en peso; la intensidad de un estímulo y el tiempo de reacción o hasta el ingreso total familiar y los gastos médicos. La naturaleza y la intensidad de las relaciones entre variables como estas pueden examinarse por medio del análisis de regresión y correlación.

La regresión es útil para averiguar la forma probable de la relación entre 2 variables y el objetivo es predecir o estimar el valor de una variable correspondiente a un valor dado de otra variable. El análisis de correlación se refiere a la medición de la intensidad de la relación entre las variables. Cuando se calculan medidas de correlación a partir de un conjunto de datos, el interés se centra en el grado de correlación entre las variables. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

a) La correlación:

El concepto fue utilizado por primera vez en 1888 por el Sir Francis Galton.

Esta noción mide el grado de unión que hay entre varias variables, según la naturaleza y el numero de variables implicadas se le asigna un nombre.

- La unión entre dos variables cuantitativas distribuidas normalmente, se denomina correlación lineal simple.
- La unión entre una variable dependiente y varias variables cuantitativas independientes, se denomina correlación múltiple.
- La unión entre dos conjuntos de variables cuantitativas, se denomina correlación canónica.

- La relación entre dos variables semicuantitativas, se denomina correlación de rango.
- La relación entre dos variables cualitativas, se denomina asociación; y se trata además de variables cualitativas binarias, se denomina coeficiente de correlación de punto. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

4.3.1. Correlación de Pearson

La correlación de Pearson es una medida de unión lineal existente entre dos variables cuantitativas aleatorias.

Esta dada por la siguiente expresión: $\rho x, y = \frac{\sigma xy}{\sigma x \sigma y}$

Donde σxy es la covarianza entre x e y.

$$\sigma xy = \frac{\sum_{i=1}^{N}(x-\mu x)(y-\mu y)}{N}$$

$\sigma x\ y\ \sigma y$ corresponden a las desviaciones estándar, estos términos se refieren a la población estadística.

Para una muestra aleatoria simple de talla "n" el estimador de $\rho x, y$ es:

$$rxy = \frac{Sx,y}{SxSy} = \frac{\sum(x-x)(y-y)}{(\sum(x-x)^2 \sum(y-y)^2)^{1/2}}$$

Donde Sx,y es la covarianza estimada:

$$rxy = \frac{\sum_{i=1}^{n}(x-x)(y-y)}{n-1} = Sxy = \frac{n\sum(xy-(\sum x)(\sum y)}{n(n-1)}$$

$$Sx^2 = \frac{\sum x2 - \frac{(\sum x)^2}{n}}{(n-1)} = \ varianza\ de\ x;\ \sqrt{Sx^2} = \ desviacion\ estandar\ de\ x$$

La correlación lineal es la covarianza de dos variables centradas y reducidas.

Las correlaciones pueden ser integradas en una tabla de doble entrada:

	x	y
x	$r_{x,x}=1$	$r_{x,y}$
y	$r_{x,y}$	$r_{x,x}=1$

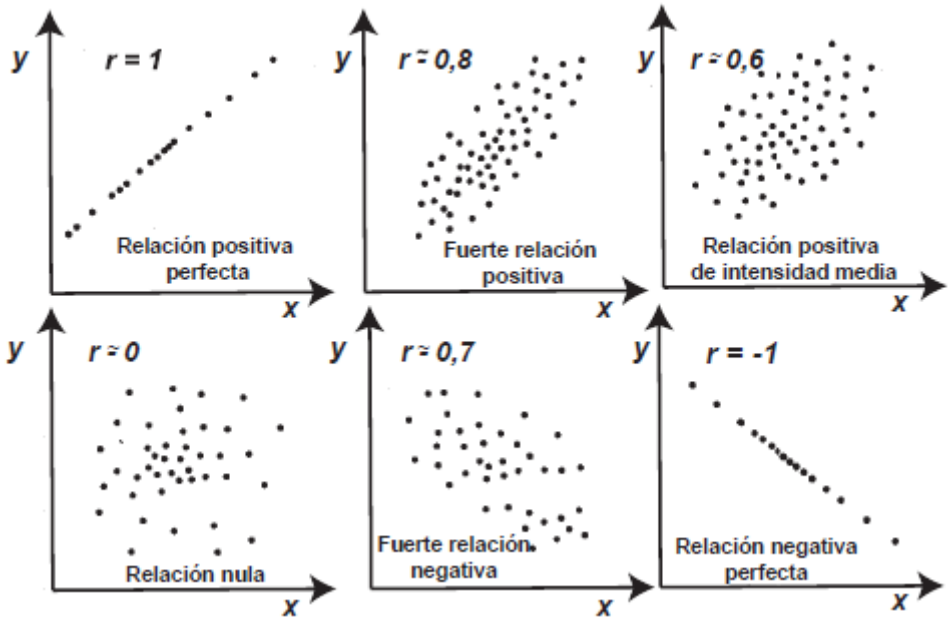
Esta tabla presenta las siguientes propiedades (Ilustración 42)

- Tiene solo 1 en la diagonal, ya que por definición la varianza de una variable centrada y reducida es 1; además de tener simetría respecto a la diagonal puesto que $rx, y = rx, y$
- Varía entre -1 y +1, si $r = -1o + 1$ todos los puntos están situados en una línea; si la nube de puntos no muestra ninguna tendencia.

- Y la última propiedad, es referente al signo. Si es (+) indica que la variable dependiente aumenta al mismo tiempo que la independiente. Si el signo es (-) significa que una variable aumenta cuando la otra disminuye. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Ilustración 46

Coeficientes de correlación relacionados a diferentes nubes de dispersión



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

El coeficiente de correlación de Pearson, mide la intensidad de la unión y la eficacia de ajuste de los datos a un modelo lineal o linealizado; sin embargo, no indica necesariamente una dependencia directa de las variables o una relación causa-efecto.

Ejemplo 1:En un estudio de la capacidad reproductiva del insecto "A" del pino. Se busca determinar la intensidad de la relación entre la longitud del nido con el número de ovocitos.

El numero de ovocitos por nido (y) y la longitud del nido (x) son dos variables aleatorias cuantitativas. Entonces su relación se mide por r de Pearson.

$$Sxy = 5.51 \qquad rxy = \frac{Sxy}{SxSy} = \frac{\sum(x-x)(y-y)}{(\sum(x-x)^2 \sum(y-y)^2)^{1/2}}$$

$$S^2x = 0.3039 \qquad rxy = \frac{\sum_{i=1}^{n}(x-x)(y-y)}{n-1} = Sxy = \frac{n\sum(xy-(Ex)(Ey)}{n(n-1)}$$

$$S^2x = 344.13 \qquad r = \frac{5.51}{\sqrt{0.3039}\,.\sqrt{344.13}} = 0.544$$

La relación que une estas variables es débil. En otras palabras, hay una fuerte dispersión de puntos. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

a) Calculo de Significatividad de r

Tomando como ejemplo el ejercicio anterior, se quiere probar si la correlación entre la longitud del nido y el numero de ovocitos es altamente significativa. Para probar la significatividad de r, se utiliza la prueba de t con n-2 grados de libertad.

-Condiciones de aplicación del test:

-Variables aleatorias cuantitativas,

-con distribución normal,

-Para que t1 sea válido, x e y deben tener una distribución binomial.

$$Ho; \rho = 0$$

$$Ho; \rho > 0$$

Prueba: $tr = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

$$r = 0.544 \quad n = 69; \quad \alpha = 0.01$$

$$tr = \frac{0.544\sqrt{69-2}}{\sqrt{1-0.544^2}} = 5.24$$

Como n = 69 y no hay un valor preciso en la tabla y para test unilateral, el valor critico se obtiene interpolando los valores de gl = 65 y gl = 70, entonces:

$t0.01(1),85 = 2.381 = Ca$ $\qquad Ca = Valor\ critico\ de\ a$

$t0.01(1),70 = 2.386 = Cb$ $\qquad Cb = Valor\ critico\ de\ b$

Se tiene entonces que g/a<g/b;

Ahora se estima la proporción utilizando n-2 gl; así se tiene: EQ

$$p = \frac{gl - a}{b - a} = \frac{67 - 65}{70 - 65} = \frac{2}{5} = 0.4$$

Se calcula el valor critico: $Cgl = 67 = Ca + p(Ca - Cb) = 2.381 + 0.4 * 0.005 = 2.383$

$$t0.01(1),67 = 2.383$$

Decisión estadística:

$tr = 5.24 > t\alpha = 2.383$ se rechaza Ho, al 99% de confiabilidad, lo que indica que el número de ovocitos aumenta de manera muy significativa con la talla del nido. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

4.3.2. Regresión Lineal Simple

a) La regresión lineal

La regresión lineal tiene un triple objetivo

- Permite resumir o sintetizar la relación existente entre una variable aleatoria dependiente "y" y una o varias variables aleatorias (modelo II) o controladas (modelo I) x1 llamada variables explicativas (independientes).
- Describe la forma de relación que une a las variables. Puede ser una relación lineal; puede ser una asociación de 2º, 3º o n grado, entonces se habla de una regresión polinomial, que pueden ser funciones hiperbólicas, logísticas u otras.
- Predecir; los valores de y1 en función de x1. Es decir, estimar con un mínimo de error el desconocido de "y" de un elemento a partir de los resultados obtenidos de las variables predictivas x1.

Cuando la estimación se realiza sobre varias variables descriptivas cuantitativas se trata de regresión múltiple. Si se agregan varias variables cualitativas, se trata de regresión múltiple con variables mudas.

Si solo hay dos variables cuantitativas se trata de regresión simple. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

b) Regresión Lineal Simple

Es una función de primer grado que une las variables x e y: $y = ax + b$

Se aplica a los modelos I y II. Corresponde a la recta que atraviesa de la mejor manera la nube de puntos cuando se relacionan dos variables.

c) Cálculo de la Recta de Regresión y Función de X: Método de Mínimos Cuadrados (MC)

Consiste en escoger una pendiente (a) y una ordenada al origen (b) de la recta que minimiza la suma de los cuadrados de los errores (SCE).

Error = Separación entre el valor observado y1y el valor predicho por la recta y1, es el ei residuo (Ilustración 43).

El principio de los mínimos cuadrados es minimizar los errores:

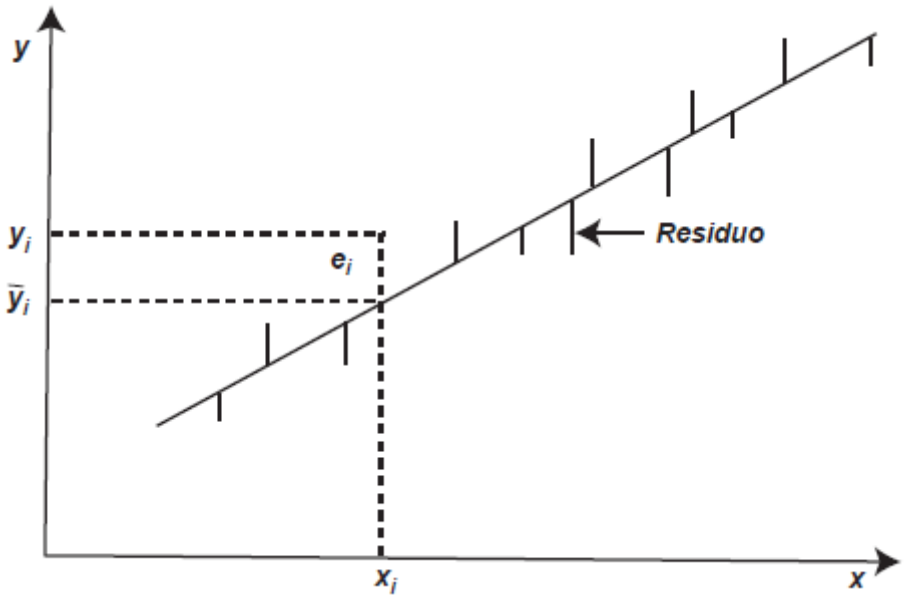
$$\sum_{i=1}^{n} e1^2 = \sum_{i=1}^{n} (y1 - y1)^2$$

En calculo diferencial se vio que el mínimo de una función se encuentra igualando a cero su primera derivada. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

$$Como: \quad y1 = axi + b \quad entonces: \quad \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y1 - y1)^2 = \sum_{i=1}^{n} (y1 - (ax1 + b))^2$$

Ilustración 47

Grafica de los residuos de una recta de regresión de y en x



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Se calculan las derivadas parciales en función de a y b y se seleccionan los parámetros o valores que satisfacen simultáneamente la anulación de las dos derivadas parciales. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Las dos ecuaciones con 2 incógnitas se llaman ecuaciones normales:

Sea: $S = \sum^{n} e1^2$; $\quad \frac{\partial S}{\partial b} = -2\sum(y1 - b - axi) = 0$

$$\frac{\partial S}{\partial b} = -2\sum^{n} X1(y1 - b - axi) = 0$$

Del desarrollo se obtiene:

$$b = y - ax; \qquad a = \frac{Sxy}{Sx^2}$$

Ejemplo 1: Se quiere obtener el modelo de regresión lineal de la concentración de DDt, DDE, DDD contra la edad de algunos peces:

n=45	$\bar{x}$=3.44 años
$\sum x$=155	S_x= 1.235
$\sum x^2$= 601	S^2_x=1.525

Modelo I	
$\sum y$=20.09	Sy= 0.276
$\sum y^2$= 12.33	$\bar{y}$=0.446µg/g•año
$\sum xy$=83.325	S^2_x=0.076

$$Sxy = \frac{n\sum xy - (\sum x)(\sum y)}{n(n-1)}$$

$$Sxy = \frac{45x83.32 - 155x20.09}{45(44)} = 0.321$$

$$r = \frac{Sxy}{SxSy} = \frac{0.321}{1.235x0.276} = 0.94; \quad a = \frac{Sxy}{Sx^2} = \frac{0.321}{1.525} = 0.210\mu g/g.\text{año}$$

$$b = y - ax = 0.466 - 0.210x3.44 = -0.278\mu/g; \quad y = 0.210x + (-0.278)\ o\ bien$$

$$y = -0.278 + 0.210x$$

4.3.3. Correlación de Spearman

Esta prueba estadística permite medir la correlación o asociación de dos variables y es aplicable cuando las mediciones se realizan en una escala ordinal, aprovechando la clasificación por rangos. El coeficiente de correlación de Spearman $\rho(rho)$, es una prueba no paramétrica que mide la asociación entre dos variables discretas. Para calcular ρ, los datos son ordenados y reemplazados por su respectivo orden:

El estadístico ρ viene dado por la expresión:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

Donde D es la diferencia entre los correspondientes valores de x-y. N es el número de parejas. Se tiene que considerar la existencia de datos idénticos a la hora de ordenarlos, aunque si estos son pocos, se puede ignorar tal circunstancia.

Para muestras mayores de 20 observaciones, podemos utilizar la siguiente aproximación a la distribución de t de Student.

$$t = \frac{\rho}{\sqrt{(1-\rho^2)/(n-2)}}$$

El coeficiente de correlación de Spearman se rige por las reglas de la correlación simple de Pearson, y las mediciones de este índice corresponden de +1ª-1, pasando por el cero, donde este último significa no correlación entre las variables estudiadas, mientras que los dos primeros denotan la correlación máxima.

La ecuación utilizada en este procedimiento, cuando en el ordenamiento de los rangos de las observaciones no hay datos empatados o ligados, es la siguiente:

$$rs = 1 - \frac{6\sum d^2}{N^3 - N}$$

Donde:

rs = coeficiente de correlación de Spearman.

d^2= diferencias existentes entre los rangos de las dos variables, elevadas al cuadrado.

N = tamaño de la muestra expresada en parejas de rangos de las variables.

Pasos.

1) Clasificar en rangos cada medición de las observaciones.
2) Obtener las diferencias de las parejas de rangos de las variables estudiadas y elevadas al cuadrado.
3) Efectuar la sumatoria de todas las diferencias al cuadrado.
4) Aplicar la ecuación.
5) Calcular los grados de libertad (gl).gl = número de parejas – 1. Solo se utilizará cuando la muestra sea mayor a 10.
6) Compara el valor r calculado con respecto a los valores críticos de la tabla de valores críticos de coeficiente de correlación por rangos de Spearman en función de probabilidad.
7) Decidir si se acepta o rechaza la hipótesis.

Ejemplo 1: Un investigador está interesado en conocer si el desarrollo mental de un niño está asociado a la educación formal de su madre. De esta manera, obtiene la calificación de desarrollo mental en la escala de Gesell de ocho niños elegidos aleatoriamente y se informa del grado de escolaridad de las madres.

a) Elección de la prueba estadística

Se desea medir asociación o correlación. Las calificaciones de la educación formal de las madres están dadas en una medición cualitativa, pero tienen una escala ordinal, por lo cual es posible ordenarlas en rangos.

b) Planteamiento de la hipótesis

- Hipótesis alternativa (H1). El desarrollo mental de los hijos es una variable dependiente de la educación formal de la madre; por lo tanto, existe una correlación significativa.
- Hipótesis nula (Ho). La asociación entre las variables de educación formal de la madre y el desarrollo mental de los hijos no es significativa, ni hay correlación.

c) Nivel de significación. Para todo valor de probabilidad igual o menor que 0.05, se acepta H1 y se rechaza Ho.

d) Zona de rechazo. Para todo valor de probabilidad mayor que 0.05, se acepta Ho y se rechaza H1.

e) Aplicación de la prueba estadística. Las observaciones de cada variable se deben ordenar en rangos, así como obtener las diferencias entre los rangos, efectuar la sumatoria y elevar esta al cuadrado.

Cálculo de rs de Spearman.

$$rs = \frac{1 - 6\sum d^2}{N^3 - N} = \frac{1 - 6\ x\ 26}{8^3 - 8} = \frac{1 - 156}{504} = 0.69$$

Cálculo de los grados de libertad (gl).

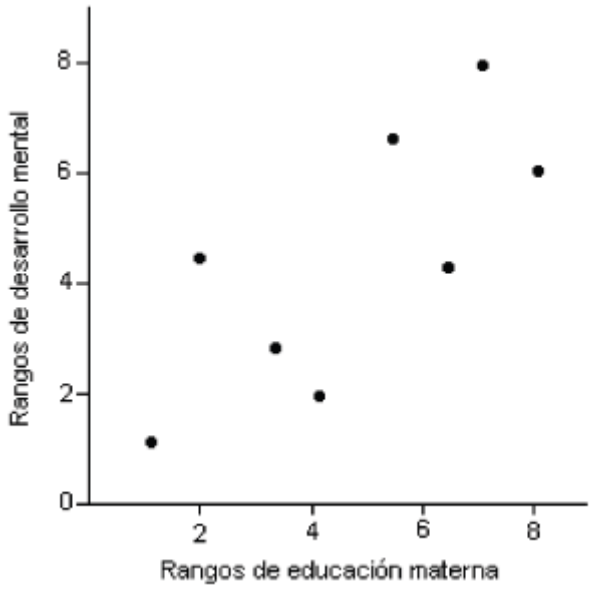
$$gl = numero\ de\ parejas\ -\ 1 = 8 - 1 = 7$$

f) El valor rs calculado se compara con los valores críticos de rs del coeficiente de correlación por rangos de Spearman.

g) El valor critico de rs con 7 grados de libertad, para una probabilidad de 0.05 del nivel de significatividad es 0.714, o sea, mayor que el calculado. Por lo tanto, este tiene una probabilidad mayor que 0.05.
h) Decisión. Como el valor de probabilidad de rs de 0.69 es mayor que 0.05, se acepta Ho y se rechaza H1.
i) Interpretación. El coeficiente de correlación de Spearman de 0.69 es menor que los valores críticos de la tabla, pues a estos corresponde la probabilidad de obtener esa magnitud, al nivel de confianza de 0.05 y 0.01, para 0.714 y 0.893. Esto significa que para aceptar H1, se requiere tener un valor igual o más lato que 0.714. Por lo tanto, se acepta Ho y se rechaza H1, aun cuando, como se observa en la (Ilustración 44), existe una asociación relativa entre la educación formal de la madre y el desarrollo mental de sus hijos; sin embargo, esta no es significativa. (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

Ilustración 48

Asociación relativa entre la educación formal de la madre y el desarrollo mental de sus hijos



Fuente: (Flores Hernandez, Ramos Miranda, & Sosa Lopez, 2007)

4.4. Aplicaciones en la Arquitectura

4.4.1. Software estadístico R Commander

Ejemplo 1: Si decidís utilizar el código de R para hacer los ejercicios, hacedlo usando un script. En el caso de que decidáis utilizar R Commander, procura guardar, al menos, el fichero de instrucciones.

El fichero *JaenIndicadores.xls* se refiere a los municipios de la provincia de Jaen en el 2011, e incluye las siguientes variables:

- Código INE del municipio.
- Nombre del municipio.
- Consumo de energía eléctrica en megavatios por hora.

- Consumo medio de agua en invierno, en metros cúbicos por día.
- Consumo medio de agua en verano, en metros cúbicos por día.
- Destino de los residuos sólidos urbanos: las posibilidades son vertedero controlado, vertedero incontrolado, compostaje.
- Cantidad de residuos sólidos urbanos, en toneladas.

1) Importar el fichero y llamar a la hoja de datos Jaen.
2) Recodificar la variable "Población" en una variable cualitativa tipo factor llamada "Tamaño" con tres categorías:

- Si la población es inferior a 2000 habitantes, Tamaño será "Pequeño".
- Si la población esta entre 2000 y 4500 habitantes, Tamaño será "Mediano".
- Si la población es superior a 4500 habitantes, Tamaño será "Grande".

3) Calcular los siguientes promedios a partir de las variables existentes:

- Consumo de energía eléctrica por habitante, elec.hab, obtenido como:

$$\frac{Consumo.de.energia,electrica}{Poblacion},$$

- Consumo medio de agua por habitante y dia, agua.hab, obtenido como:

$$\frac{Consumo.de.agua.Invierno + Consumo.de.agua..Verano}{Poblacion},$$

- Residuos sólidos urbanos por habitante, res.hab, obtenido como:

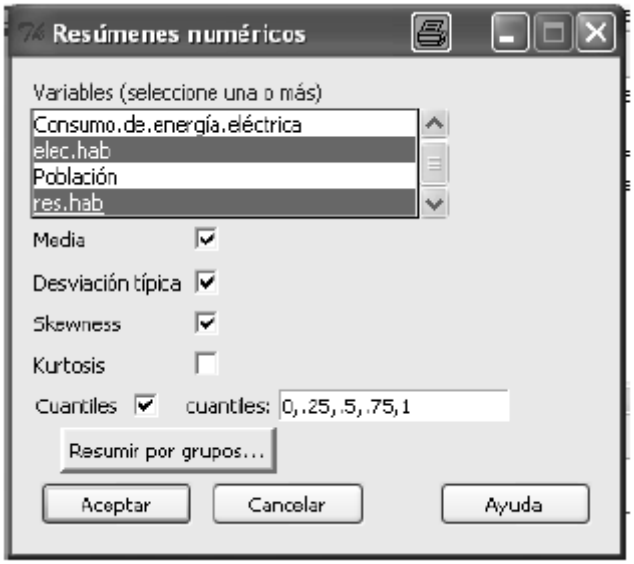
$$\frac{Residuos.solidos.urbanos..Cantidad}{Poblacion}$$

4) Guardar esta hoja de datos (Jaen) en un archivo de datos de R llamado *JaenIndicadores.RData* o *JaenIndicadores.rda.*
5) Definir una nueva hoja de datos con todas las variables que contienen los datos originales, pero referida solo a los municipios de tamaño mediano.
6) Cálculo de medidas de posición, dispersión y forma:

Las medidas de posición, dispersión y forma más comunes, media, mediana, percentiles, desviación típica y coeficiente de asimetría, se hallan en la opción del menú Estadisticos → Resumenes → IP − SUR − Numerical Summaries. A modo de ejemplo, vamos a obtener estas medidas para los promedios elec.hab, agua.hab y res.hab. En la Ilustración 45, aparece las entradas de esa opción del menú. En ella es posible elegir varias variables a la vez, pulsando a la vez, pulsando la tecla Control mientras se clica en las variables deseadas. (Saez Castillo, 2010)

Ilustración 49

Opción Resúmenes numéricos del menú



Fuente: (Saez Castillo, 2010)

En la Tabla se muestra los resultados que aparece. En esta, mean se refiere a la media, sd a la raíz de la cuasi-varianza, skewness es el coeficiente de asimetría, el percentil 0 es el valor mínimo de la variable, el percentil 50, como ya sabemos, es la mediana y el percentil 100 es el valor máximo de la variable. (Saez Castillo, 2010)

Tabla 40

Descriptivos básicos de elec.hab, agua.hab y res.hab

	mean	sd	skewness	0 %	25 %	50 %	75 %	100 %
agua.hab	0.53	0.14	1.22	0.14	0.46	0.51	0.56	1.09
elec.hab	2.66	1.40	2.05	0.94	1.75	2.20	3.10	9.19
res.hab	0.23	0.04	1.39	0.18	0.21	0.23	0.24	0.35

Fuente: (Saez Castillo, 2010)

7) Mediante grupo: La funciones mean ( ), sd ( ) y quantile ( ) proporcionan la media, la desviación típica y los cuantiles de cualquier muestra. Todas estas órdenes responden al mismo tipo de formato. Por ejemplo, si queremos calcular la media de las anteriores variables escribimos

mean(Datos$agua.hab,na.rm=TRUE)

mean(Datos$elec.hab,na.rm=TRUE)

mean(Datos$res.hab,na.rm=TRUE)

O bien, las tres medias juntas mediante

mean(Datos[,c(“agua.hab”,”elec.hab”,”res.hab”)], na.rm=TRUE)

El argumento na.rm-TRUE indica que los valores faltantes NA se eliminan para realizar los cálculos. Si no se incluyen esta opción entonces la media resultante será N.A.

De igual forma, la desviación típica se lograría mediante

sd(Datos$agua.hab,na.rm=TRUE)

sd(Datos$elec.hab,na.rm=TRUE)

sd(Datos$res.hab,na.rm=TRUE)

O bien,

sd(Datos[,c("agua.hab","elec.hab", "res.hab")],na.rm=TRUE)

Finalmente, para obtener los cuantiles necesitamos especificar las variables y las probabilidades de los cuantiles que deseamos mediante el argumento probs. Por ejemplo, para obtener los percentiles 5 y 95.

quantile(Datos$agua.hab,na.rm=TRUE,probs=c(0.05,0.95))

quantile(Datos$elec.hab,na.rm=TRUE,probs=c(0.05,0.95))

quantile(Datos$res.hab,na.rm=TRUE,probs=c(0.05,0.95))

El coeficiente de asimetría no viene definido en el paquete base que se carga por defecto con R, de manera que debemos buscarlo. Si ponemos ??skewness podremos comprobar como hay un paquete llamado e1071 que tiene una función con ese nombre. Si lo cargamos y vemos su ayuda (?skewness), comprobaremos que, en efecto, lo que hace esa función es calcular el coeficiente de asimetría. Por tanto,

library(e1071)

skewness(Datos$agua.hab,na.rm=TRUE)

skewness(Datos$elec.hab,na.rm=TRUE)

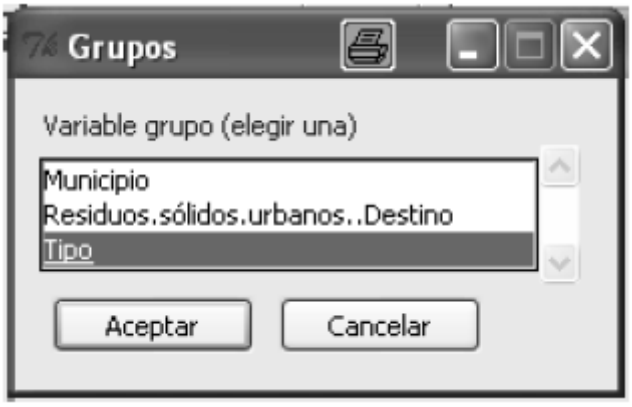
skewness(Datos$res.hab,na.rm=TRUE)

8) Resúmenes por grupos: ¿Y si deseamos hacer el mismo análisis, pero en cada uno de los grupos que determina la variable tipo? Tengamos en cuenta que esta variable es un factor, asi que podemos utilizarla para ello.

Veámoslo en primer lugar mediante R Commander. Utilizamos la misma opción $Estadisticos \rightarrow Resumenes \rightarrow IPSUR - Numerical\ summaries,$ pero ahora clicamos en la pestaña Resumir por grupos. Esta opción abre una ventana (Ilustración 46) donde aparecen como posibilidades todas aquellas variables que pueden dividir al conjunto de datos en grupo. En nuestro caso elegimos la variable Tipo. La ventana de resultados mostrara los mismos estadísticos, pero separado cada uno de los tres grupos, para cada variable. Por ejemplo, para la variable agua.hab tenemos. (Saez Castillo, 2010)

Ilustración 50

Resúmenes por grupos



Fuente: (Saez Castillo, 2010)

Tabla 41

Descriptivos básicos de agua.hab en función de Tipo

	mean	sd	skewness	0 %	25 %	50 %	75 %	100 %
Grande	0.51	0.12	1.45	0.18	0.47	0.51	0.54	1.01
Mediano	0.51	0.07	0.23	0.41	0.45	0.51	0.57	0.65
Pequeño	0.55	0.19	0.70	0.14	0.46	0.49	0.66	1.09

Fuente: (Saez Castillo, 2010)

Mediante código, los anteriores resultados se pueden calcular a través de la orden topply ( ). Por ejemplo, la media, la desviación típica y los percentiles 5 y 95 de la variable agua.hab en función de los grupos de la variable. Tipo se obtendría de la forma siguiente:

Tapply(Datos$agua.hab,Datos$Tipo,mean,na.rm=TRUE)

Tapply(Datos$agua.hab,Datos$Tipo,sd,na.rm=TRUE)

Tapply(Datos$agua.hab,Datos$Tipo,quantile,probs=c(0.05,0.95),na.rm=TRUE)

Obsérvese que en la última línea hemos tenido que especificar las probabilidades requeridas a la función quantile.

9) Distribuciones de frecuencias: Las variables Tipo y Residuos.solidos.urbanos. Destino son de tipo cualitativo, por lo que no pueden ser resumidas mediante medidas numéricas. Para este tipo de variables el resumen más conveniente es, simplemente, su distribución de frecuencias.

Para obtener la distribución de frecuencias de una o varias variables de un conjunto de datos mediante R Commander elegimos la opción $Estadisticos \rightarrow Resumenes \rightarrow Dsitribucion\ de\ frecuencias$. En la ventana emergente elegimos las variables que queremos analizar y la tabla aparece en la ventana de resultados, incluyendo las frecuencias absolutas y relativas.

10) Mediante código: Las funciones de código correspondientes son table ( ) y prop.table ( ). Asi, para la obtencion de las distribuciones de frecuencias (absolutas y relativas) de la variable Tipo escribimos

Tabla < - table(Datos$Tipo)

Tabla # frecuencias absolutas

prop.table(Tabla)#frecuencias relativas

11) Diagrama de barras y diagrama de sectores: No obstante, asumiendo el dicho una imagen vale más que mil palabras, abemos que existen dos formas de plasmar en un gráfico la distribución de frecuencias de una variable cualitativa o discreta con pocos valores: el diagrama de barras y el diagrama de sectores.

- Diagrama de barras para variables cualitativas: En R Commander este tipo de graficos están en la opción $Graficas \rightarrow IPSUR - Bar\ Graph\ldots$.. La ventana de entradas aparece en la Ilustración 47. En esta ventana hemos solicitado un análisis de la variable Tipo. Es muy importante tener en cuenta que solo pueden representarse variables cualitativas de tipo factor.

La función barplot ( ) nos permite obtener la distribución de barras mediante código. Aquí tenemos posibilidad de controlar más cosas. Por ejemplo, es interesante añadir las frecuencias exactas al grafico o retocar los títulos de los ejes y del gráfico. La sintaxis para obtener la Ilustración 48 es la siguiente:

diagrama<-barplot(Tabla,col=rainbow(3)

,xlab="Municipios según su tipo", ylab="Frecuencias absolutas")

Text (diagrama, Tabla+1, labels=Tabla,xpd=TRUE)

Title(main="Distribución de frecuencias de la variable Tipo", Font.main = 4)

Existen otras muchísimas posibilidades en la construcción del grafico mediante código: aquí solo mostramos aspectos de su sintaxis mas básica. (Saez Castillo, 2010)

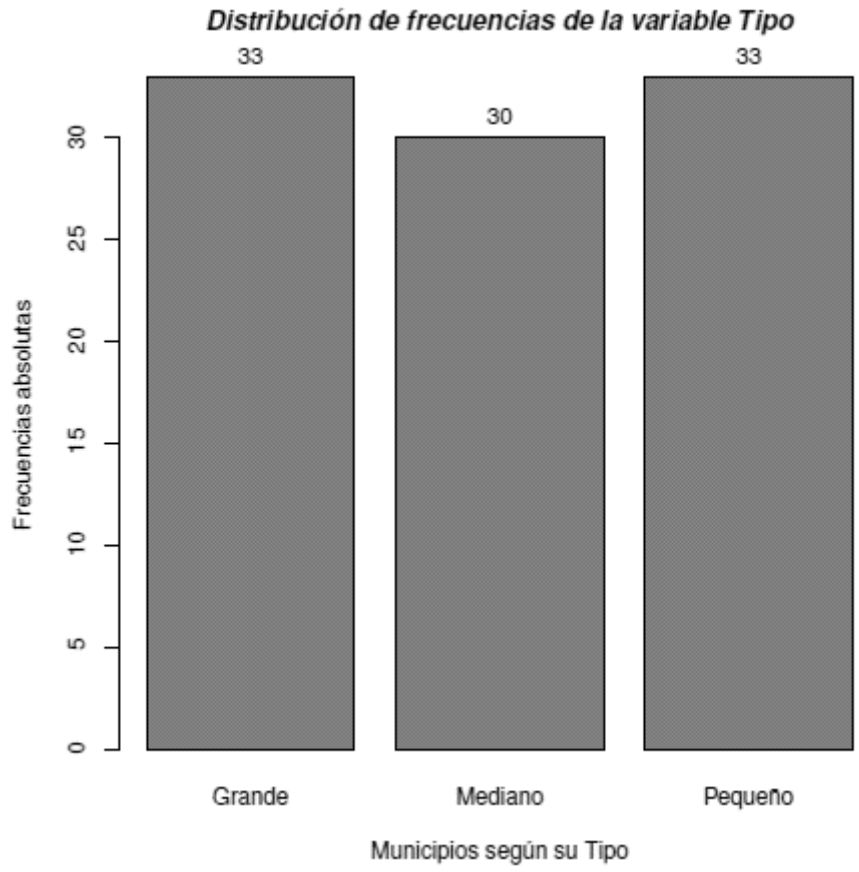
Ilustración 51

Entradas para la construcción de un diagrama de barras

Fuente: (Saez Castillo, 2010)

Ilustración 52

Diagrama de barras según el tipo de municipio del destino de los residuos solidos



Fuente: (Saez Castillo, 2010)

Ilustración 53

Diagrama de sectores del destino de los residuos sólidos urbanos en los municipios de la provincia de Jaen

Fuente: (Saez Castillo, 2010)

12) Diagrama de sectores para variables cualitativas: Para realizar un diagrama de sectores mediante R Commander elegiremos la opción $Graficas \rightarrow Diagrama\ de\ sectores$. La ventana emergente solo permite elegir una variable cualitativa. De nuevo es muy importante tener en cuenta que solo pueden representarse variables cualitativas de tipo factor. El diagrama correspondiente al destino de los residuos sólidos de los municipios aparece en la Ilustración 49.

La opción mediante código en su versión más básica es

Tabla.destino<-prop.table(table(Datos$Residuos.solidos.urbanos..Destino))

Tabla.destino<-round(100*tabla.destino,2)

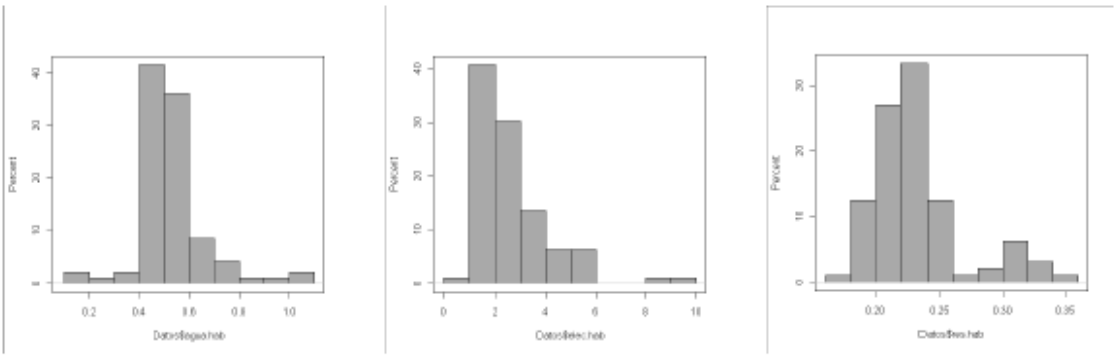
Sectores.destino<-pie(tbla.destino,
labels=paste(name(tabla.destino),tabla.destino,"1/0"),

Main=" Distribución de porcentajes de la variable Destino de los residuos sólidos urbanos")

Observemos que la función paste ( ) nos ha ayudado a pegar las etiquetas de la variable con el porcentaje y el símbolo "%", que luego hemos añadido a los sectores del gráfico. (Saez Castillo, 2010)

Ilustración 54

De izquierda a derecha, histogramas de agua.hab, elec.hab y res.hab



Fuente: (Saez Castillo, 2010)

13) Histograma para variables continuas y discretas

- Histograma para variables continuas: Como ya sabemos, los diagramas de barras o sectores no son adecuados para datos de variable continuas. Frente a estas representaciones, el histograma aparece como la alternativa válida, ya que obliga a agrupar los valores en intervalos cuya frecuencia si es relevante.

Para realizar un histograma con R Commander elegimos $Graficas \rightarrow Histograma$. La ventana de entrada permite elegir solo una variable para cada análisis, el número de intervalos del histograma y la escala de este (frecuencias absolutas, porcentajes y densidades).

En el caso de las variables agua.hab, elec.hab, res.hab hemos seleccionado histogramas con escala en porcentajes y 10 intervalos. Los resultados aparecen en la Ilustración 50.

La función hist ( ) nos permite representar el histograma mediante código. La sintaxis básica de esta función es la siguiente:

Hist(x,breaks=”Sturges”,freq=NULL,

Main=paste(“Histogramof”,xname),labels=FALSE)

- X es el vector de datos.
- Breaks puede especificar el número de intervalos que deseamos o los extremos de los intervalos que deseamos considerar, mediante un vector. Por defecto, asigna el número de intervalos por el conocido como método de Sturges.
- Freg especifica si la escala del histograma es tal que el área de las barras es igual a la proporción de datos en cada intervalo (escala de densidad, con valor freq=FALSE) o su altura es simokemente el recuerdo de las frecuencias (escalas de frecuencias, con valor freq=TRUE).
- Main especifica el título del gráfico, mientras que xlab e ylab especifican el de los ejes.
- Labels añade una etiqueta a cada barra con el valor de las frecuencias.

Por ejemplo, para calcular el histograma de frecuencias de la variable agua.hab debemos escribir

Hist(Datos$agua.hab,breaks=10,freq=true,

Main=" Histograma del consumo de agua por habitante",

Xlab="",ylab="Frecuencias")

- Histograma para variables discretas: Parece contradictorio hablar de un histograma para variables discretas, ya que una variable discreta debe representarse con un diagrama de barras. Sin embargo, para los ordenadores y los programas estadísticos, la división entre variables discretas y continuas suele no ser habitual. Se entiende que ambas son variables numéricas, al contrario de las variables cualitativas, que son caracteres, y para variables numéricas se ofrece el histograma como representación gráfica.

Lo que podemos hacer es utilizar la función que realiza el histograma para realizar un diagrama de barras de una variable discreta. La idea es muy simple: le pediremos a R a través de Commander o de la consola el histograma de la variable discreta de manera que las barras del histograma representen las frecuencias de los valores de la variable discreta.

14) Detección de valores atípicos. Diagrama de caja

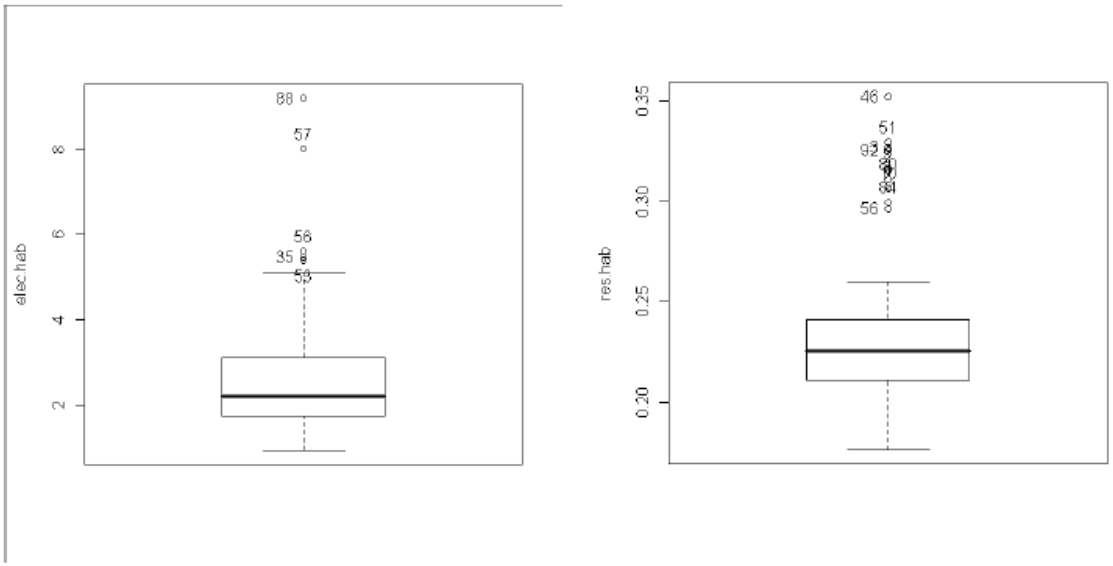
Un método común para la detección de valores atípicos es el llamado de caja o boxplot. Este método es válido para cualquier conjunto de datos, independientemente de la forma de su distribución de frecuencias.

Vamos a obtener el diagrama de caja de las variables elec.hab y res.hab e identificar los municipios atípicos en cuanto al consumo de energía eléctrica y de los residuos generados por habitante. Para ello, elegimos la opción $Graficas \rightarrow Diagrama\ de\ caja\ o\ la\ opcion\ Graficas \rightarrow IPSUR - Boxplot$. Las dos ventanas de entradas son muy parecidas: en ellas tenemos que elegir la variable que queremos analizar y existen dos opciones muy interesantes:

a) Podemos elegir un análisis por grupos, sin más que clicar en la pestaña Resúmenes por grupos. En esta ocasión no es necesario, pero podríamos hacerlo con los grupos generados por la variable Tipo.
b) Podemos elegir la opción Identificar atípicos con el ratón, que es la forma más fácil de señalar los municipios atípicos. Recordemos que en un diagrama de caja los municipios atípicos se señalan con círculos, y los atípicos extremos con asteriscos. (Saez Castillo, 2010)

Ilustración 55

Diagramas de caja de las variables elec.hab (izquierda) y res.hab (derecha)



Fuente: (Saez Castillo, 2010)

En la Ilustración 51, a la izquierda aparece el diagrama de caja de la variable elec.hab. En ella podemos ver gráficamente la asimetría a la derecha de la distribución, ya que el lado a la derecha de la mediana, que separa la caja, es más grande. Por su parte, también podemos ver que hemos detectado como municipios atípicos a la derecha de la distribución, es decir, que destacan por su fuerte consumo por habitante, a los municipios 35,53, 56, 57 y 88 (estos dos últimos de forma muy clara), que corresponden con Guarroman, Lupion, Martos, Mengibar y Villanueva de la Reina. Una forma eficiente de ver sus nombres es introducir el siguiente código en la ventana de instrucciones:

Datos$Municipio[c(35,53,56,57,88)].

15) La función bloxplot ( )

La sintaxis básica de la función bloxplot ( ) obliga simplemente a especificar el conjunto de datos.

Por ejemplo,

Boxplot(Datos$elec.hab)

También es posible añadir un título al gráfico y a los ejes, como en el caso de hist ( ).

Por su parte, si no cerramos el grafico, también podemos especificar los atípicos mediante la función.

Identify( ). Vamos a verlo en el ejemplo:

Identify(rep(1,length(Datos$elec.hab)),

Datos$elec.hab, rownames(Datos))

El primer argumento, rep(1,length(Datos$elec.hab)), esta especificando que la coordenada de x de todos los puntos es 1 (lugar donde se sitúan todos los valores). Datos$elec.hab especifica la coordenada y de los puntos (ya que el diagrama es vertical), mientras que rownames (Datos) especifica como identificar los puntos, en este caso con el numero de la fila. Podríamos cambiar esta opción por Datos$Municipio y al clicar nos daría directamente el nombre del municipio (aunque si señalamos varios, probablemente se solaparan). (Saez Castillo, 2010)

5. BIBLIOGRAFÍA

Alvarez Roman, J. (2004). *Estadistica Aplicada a Proyectos.* Riobamba: Diego Basantes.

Barrios Zamudio, E., Garcia Perez, J., & Matuk Villazon, J. (2016). *Tablas de Probabilidades.* Ciudad de Mexico: Departamento Academico de Estadistica.

Berrocal de Montestruque, L., Asurza Olaechea, H., & Billon, S. A. (2016). *Glosario basico de terminos estadisticos.* Lima: Centro de ediciones del INEI.

Botella Rocmora, P., Alacreu Garci, M., & Martinez Beneito , M. A. (2014). *Estadistica en Ciencias de la Salud.* Univ. CEU-Cardenal Herrera.

Canavos, G. C. (1988). *Probabilidad y Estadistica Aplicaciones y metodos.* Ciudad de Mexico: Camara Nacional de la Industria .

Cordova Zamora, M. (1995). *Estadistica Descriptiva e Inferencial.* Lima: Moshera S.R.I.

Del Castillo Galarza, R. S., & Salazar Pinto, R. C. (2018). *Fundamentos Basicos de Estadistica.* Del Castillo Galarza, Raul Santiago.

Devore, J. L. (2011). *Probabilidad y Estadistica para Ingenieria y Ciencias.* Cengage Learning Latin America.

Flores Hernandez, D., Ramos Miranda, J., & Sosa Lopez, A. (2007). *Estadistica Descriptiva Probabilidad y Pruebas de Hipotesis.* Universidad Autonoma de Campeche.

Garcia Salazar, M. G., & Ruiz Galindo, L. A. (2013). *Introduccion a la Probabilidad.* Universidad Autonoma Metropolitana - Azcapotzalco.

Landro, A. H., & Gonzalez, M. L. (2018). *Teoria General de las Variables Aleatorias.* Centro de Investigaciones en Econometria.

Lopez Roldan, P., & Fachelli, S. (2015). *Metodologia de la Investigacion Social Cuantitativa.* Barcelona: Universidad Autonoma de Barcelona.

Martínez Bencardino, C. (2012). *Estadística y muestreo.* Bogotá: ECOE EDICIONES.

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2010). *Introduccion a la probabilidad y estadistica.* Cengage Learning.

Montes Suay, F. (2007). *Introduccion a la Probabilidad.* Valencia: Departamento de Estadistica e Invesstigacion Operativa.

Posada Hernandez, G. J. (2016). *Elementos Basicos de Estadistica Descriptiva para el Analisis de Datos.* Colombia: Luis Amigo.

Quezada, N. (2010). *Estadística para ingenieros.* Lima, Perú: Macro E.I.R.L.

Sacco , L. C. (2011). *Probabilidad y Estadistica II .* Buenos Aires: Universidad Tecnologica Nacional Facultad Regional San Nicolas.

Saez Castillo, A. J. (2010). *Metodos Estadisticos con R y R Commander.* Jaen: Departamento de Estadistica e Investigacion Operativa.