Manual de inferencia estadística

Marianela Luzardo Briceño
Manuel Arturo Jiménez Ramírez

Escuela de Ingeniería Escuela de Ingeniería

Universidad Pontificia Bolivariana

Marianela Luzardo Briceño

Nacida en Mérida, Venezuela, el 24 de abril de 1965. Licenciada en Estadística (julio de 1989), Magister Scientiarum en Estadística Aplicada (Mayo de 1996) y Doctora en Estadística (noviembre de 2008) de la Universidad de Los Andes, Mérida-Venezuela. Ha participado en decenas de eventos científicos nacionales e internacionales; autora y coautora de diversos libros, capítulos de libros y artículos en revistas científicas indexadas como Scientometrics, Comunicaciones Estadísticas, Economía, Actualidad Contable, Revista Virtual de la Universidad Católica del Norte, Revista Pensando Psicología, y Psicogente, entre otras. Directora del gupo de investigación GeeTIC, Empresa, Educación y TIC. Investigadora Asociada de Colciencias –Colombia (2014-2015), Investigador B, PEI-ONTIC, Venezuela (2014-2015). Docente de la Facultad de Ingeniería Industrial de la Universidad Pontificia Bolivariana, Seccional Bucaramanga, Colombia.

Contacto: marianela.luzardo@upb.edu.co

Manuel Arturo Jiménez Ramírez

Es graduado de Ingeniería Industrial de la Universidad Central, Bogotá-Colombia y tiene una maestría en Ingeniería Industrial de la Universidad de los Andes, Bogotá Colombia. En su experiencia profesional ha trabajado como investigador en diferentes proyectos, asistido a congresos nacionales e internacionales, publicado libros de investigación y como docente en áreas organizacionales y métodos cuantitativos.

Contacto: manuel.jimenezr@upb.edu.co

Manual de inferencia estadística

Marianela Luzardo Briceño

Manuel Arturo Jiménez Ramírez

519.54
L979

Luzardo Briceño, Marianela, autor
Manual de inferencia estadística / Marianela Luzardo Briceño, Manuel Arturo Jiménez Ramírez -- Medellín: UPB, Seccional Bucaramanga, 2018.
236 páginas ; 17 x 24 cm.
ISBN: 978-958-764-531-6

1. Inferencia (Estadística) -- 2. Probabilidades (Estadística) – I. Jiménez Ramírez, Manuel Arturo, autor – II. Título

CO-MdUPB / spa / rda
SCDD 21 / Cutter-Sanborn

Manual de inferencia estadística
ISBN: 978-958-764-531-6
Primera edición 2018
Escuela de Ingeniería
Facultad de Ingeniería Industrial
Dirección de Investigaciones y Transferencia - DIT
Seccional Bucaramanga

Arzobispo de Medellín y Gran Canciller UPB: Mons. Ricardo Tobón Restrepo
Rector General: Pbro. Julio Jairo Ceballos Sepúlveda
Rector Seccional Bucaramanga: Presbítero Gustavo Méndez Paredes
Vicerrectora Académica Seccional Bucaramanga: Ana Fernanda Uribe Rodríguez
Decano de la Escuela de Ingeniería: Edwin Dugarte Peña
Director de la Facultad de Ingeniería Industrial: María Teresa Castañeda Galvis
Gestora Editorial Seccional Bucaramanga: Ginette Rocío Moreno Cañas
Editor: Juan Carlos Rodas Montoya
Coordinación de Producción: Ana Milena Gómez Correa
Diagramación: María Isabel Arango Franco
Corrección de Estilo: Juana Manuela Montoya Velásquez

Dirección Editorial:
Editorial Universidad Pontificia Bolivariana, 2018
E-mail: editorial@upb.edu.co
www.upb.edu.co
Telefax: (57)(4) 354 4565
A.A. 56006 - Medellín - Colombia

Radicado: 1645-17-10-17

Contenido

Introducción

Introducción

Una de las situaciones más comunes cuando se hace un análisis de tipo estadístico es realizar inferencias acerca de la media μ de una población. Es por esto que se debe investigar sobre la distribución muestral de este estadístico, incluyendo la forma o tipo de su distribución de probabilidad y algunas de sus características, como son la media y la varianza.

Otras veces se puede estar interesado en investigar distintas características de una población que no sean necesariamente la media de la misma. En estos casos, la distribución muestral de proporciones es la más adecuada para dar respuesta a este par de situaciones.

Por lo anterior, se pueden definir indicadores numéricos asociados a una muestra, los cuales van a reflejar propiedades de ella y a depender exclusivamente de los valores o elementos de esa muestra. Estos indicadores se denominan *estadísticos* y corresponden a diferentes técnicas, como la estimación de parámetros y la contrastación de hipótesis a la hora de aplicar inferencia estadística.

Como es sabido, la estadística se divide en dos ramas fundamentales: estadística descriptiva y estadística inferencial. La estadística descriptiva se encarga de la recolección, organización y presentación de los datos en cuadros o gráficos, así como también del cálculo de medidas numéricas que permiten destacar los aspectos más importantes de los mismos. La estadística inferencial, por su parte, se apoya en los resultados obtenidos de una muestra para sacar conclusiones y tomar decisiones sobre la población en estudio.

Por consiguiente, el presente manual aborda el contenido programático de la asignatura Inferencia Estadística de una manera clara y sencilla, para que el estudiante pueda vencer las dificultades que con frecuencia se suelen presentar.

El trabajo está constituido por seis capítulos, siguiendo por supuesto el contenido programático de la materia en cuestión. Se comienza con un recorrido por aquellas distribuciones continuas necesarias para la estimación de los parámetros, tales como la media poblacional, la varianza poblacional, la proporción poblacional, entre otros.

El capítulo dos versa sobre la distribución en el muestreo y el teorema fundamental en estadística: el teorema central del límite. El capítulo tres despliega los conceptos necesarios en la inferencia estadística, como parámetros, estadísticas, estimadores y propiedades de los estimadores.

En el capítulo cuatro se desarrollan los procedimientos de inferencia estadística para una sola media poblacional cuando el tamaño de muestra es grande o pequeño, y cuando se conoce la varianza poblacional o no se conoce. Además, se hace referencia a los procesos de inferencia para dos medias poblacionales independientes y dependientes, utilizando las distribuciones normales y *t-student*, según el caso.

En el capítulo cinco se analizan los procedimientos de inferencia estadística para una y dos proporciones poblacionales usando la distribución normal. Por otra parte, el capítulo seis se relaciona con los procesos de inferencia estadística para una varianza y dos varianzas poblacionales utilizando las distribuciones ji-cuadrado o chi-cuadrado y la F de Snedecor, respectivamente.

El capítulo siete trata sobre situaciones en las cuales intervienen dos variables cuantitativas con el fin de observar las relaciones existentes entre ellas a través de dos técnicas: la regresión y la correlación; la primera para fines de predicción y la segunda para medir la fuerza o asociación entre las variables objeto de estudio. Para finalizar, el capítulo ocho hace referencia a los elementos de muestreo y los tipos de muestreo probabilístico y no probabilístico.

En cada uno de estos capítulos se presentan, además de la teoría básica, ejemplos adecuados para reforzar lo expuesto en el mismo.

1. Distribuciones continuas

DISTRIBUCIONES CONTINUAS

En este capítulo se describen brevemente las distribuciones continuas necesarias para los procedimientos de inferencia en relación con los diferentes parámetros de interés, a saber, distribución normal, chi-cuadrado, *t-student* y distribución F-Snedecor.

1.1 Distribución normal[1]

Esta distribución debe su origen al matemático francés Abraham de Moivre, en 1933. Sin embargo, Pierre Laplace (1944) y Carl Gauss (1809 y 1816) fueron figuras importantes en el desarrollo de la misma. Gracias a Gauss, esta distribución alcanzó mayor popularidad, y fue difundida como la "ley normal de los errores de mediciones", particularmente relacionada con observaciones astronómicas. La curva de la distribución normal se conoce también como la curva de Gauss o como campana de Gauss.

Se dice que una variable aleatoria continua X se distribuye como una normal, y se escribe $X \sim N(\mu; \sigma)$ si su función de densidad viene dada por:

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}}.\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \qquad -\infty < x < \infty \qquad (1.1)$$

Donde μ es la media poblacional y $\sigma > 0$ es la desviación estándar.

[1] Entre las aplicaciones que tiene la distribución normal se encuentra la inferencia para la media, la diferencia de medias, la proporción y la diferencia de proporciones poblacionales. Estas aplicaciones se verán con mayor profundidad a partir del capítulo cuatro.

1.1.1 Características de la distribución

- La distribución normal es una variable aleatoria continua
- Su rango de variación es $(-\infty; \infty)$
- Es simétrica con respecto a la media μ
- Su punto máximo lo obtiene en la media μ
- Está definida por sus dos parámetros μ y σ
- La media de la distribución $E(X) = \mu$
- La varianza de la distribución es σ^2
- El área total bajo la curva es igual a uno
- Tiene forma de campana

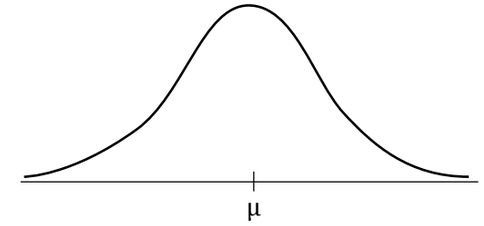


Figura 1.1 Forma de la distribución normal

1.1.2 Distribución normal estándar

Se dice que una variable aleatoria normal sigue una distribución estándar si su media es cero y su varianza es uno.

Cuando una variable aleatoria está estandarizada se le denota con la letra Z, se expresa como y su función de densidad viene dada por: $Z \sim N(0;1)$

$$f(z) = \frac{1}{\sqrt{2\pi}}.\exp\left[\frac{-z^2}{2}\right] \qquad -\infty < z < \infty \qquad (1.2)$$

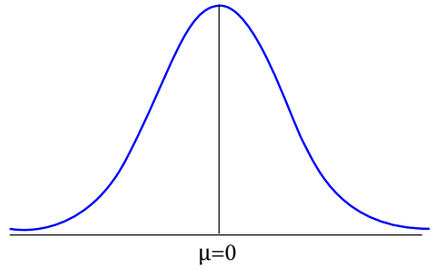


Figura 1.2 Forma de la distribución normal estándar
Fuente: MLB

1.1.3 Uso de la tabla normal estándar

En la tabla 1.1, se presenta un extracto de la tabla estadística de la distribución normal estándar acumulada, tanto para valores negativos como para valores positivos.

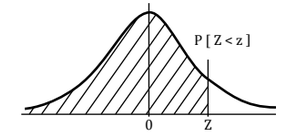
A partir de ejemplos, se van a distinguir los tres posibles casos: cuando el número es negativo, positivo o está entre dos valores.

1.1.4 Busca la probabilidad acumulada de un número negativo

Ejemplo 1.1:

Suponga que se quiere calcular Pr(Z ≤ -2,63). Dicha probabilidad está representada por el área sombreada en la figura 1.3.

Tabla 1.1 Valores de la distribución normal estándar



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549

Fuente: MLB

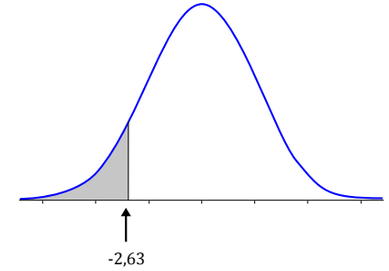


Figura 1.3 Pr(Z ≤ -2,63)
Fuente: MLB

En la tabla de la distribución normal se ubica el entero y la primera cifra decimal, mientras que la segunda cifra decimal se sitúa en la primera fila.

Tabla 1.2 Muestra de valores de la distribución normal estándar

Z	0.00	0.01	0.02	0.03
-3	0,0013	0,0013	0,0013	0,0012
-2,9	0,0019	0,0018	0,0018	0,0017
-2,8	0,0026	0,0025	0,0024	0,0023
-2,7	0,0035	0,0034	0,0033	0,0032
-2,6	0,0047	0,0045	0,0044	0,0043
-2,5	0,0062	0,0060	0,0059	0,0057
-2,4	0,0082	0,0080	0,0078	0,0075

Fuente: MLB

Por lo tanto, la Pr(Z ≤ -2,63)=0.0043. Esto quiere decir que el área sombreada es de 0.0043.

1.1.5 Busca la probabilidad acumulada de un número positivo

Ejemplo 1.2:

Suponga que se quiere calcular Pr(Z ≤ 0,32). Dicha probabilidad está representada por el área sombreada en la figura 1.4.

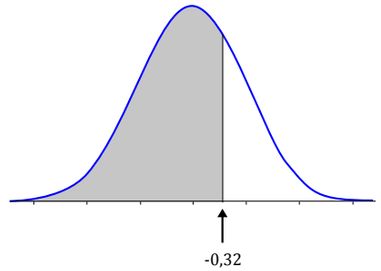


Figura 1.4 Pr(Z ≤ 0,32)
Fuente: MLB

Tabla 1.3 Muestra de valores de la distribución normal estándar

Z	0.00	0.01	0.02	0.03	0.04
0	0,5000	0,5040	0,5080	0,5120	0,5160
0,1	0,5398	0,5438	0,5478	0,5517	0,5557
0,2	0,5793	0,5832	0,5871	0,5910	0,5948
0,3	0,6179	0,6217	0,6255	0,6293	0,6331

Fuente: MLB

De igual manera como se procedió en el ejemplo 1.1, se busca el área correspondiente para este valor. El resultado es que Pr(Z ≤ 0,32)=0.6255.

1.1.6 Busca la probabilidad entre dos valores

Ejemplo 1.3:

Suponga que se quiere calcular Pr(-2,43 < Z ≤ 0,24). Esta probabilidad está representada por el área sombreada en la figura 1.5.

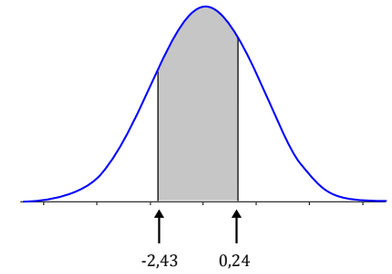


Figura 1.5 Pr(-2,43 < Z ≤ 0,24)
Fuente: MLB

En este caso se debe proceder de la siguiente manera:

- Buscar las probabilidades asociadas a cada uno de los valores que forman la desigualdad, tal como se representan en las áreas sombreadas de las figuras 1.6.a y 1.6.b:

$$\Pr[Z \leq 0{,}24] = 0{,}5948$$
$$\Pr[Z \leq -2{,}43] = 0{,}0075$$

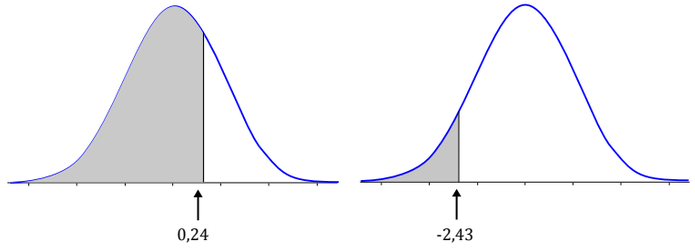


Figura 1. 6.a. Pr [Z ≤ 0,24]
Fuente: MLB

Figura 1.6.b. Pr[Z ≤ -2,43]
Fuente: MLB

- Restar las dos probabilidades:

$$\Pr[Z \leq 0{,}24] - \Pr[Z \leq -2{,}43] = 0{,}5948 - 0{,}0075 = 0.5873$$

Por lo tanto, la Pr(-2,43 < Z ≤ 0,24) es 0.5873.

1.2 Distribución chi-cuadrado o ji-cuadrado[2]

Esta distribución debe su origen a Karl Pearson hacia 1900. Su uso fundamental está basado en la inferencia para una varianza poblacional y las pruebas de la bondad del ajuste y de independencia.

[2] Entre las aplicaciones de la distribución chi-cuadrado se encuentran hacer inferencias para una sola varianza, una sola desviación estándar, pruebas de la bondad del ajuste y pruebas de independencia. Estas aplicaciones se verán con mayor detenimiento en el capítulo seis.

Esta variable aleatoria surge como la suma de variables aleatorias con distribución normal estándar, independientes elevadas al cuadrado, es decir,

$$\chi^2_\upsilon = Z_1^2 + Z_2^2 + \cdots + Z_\upsilon^2 = \sum_{i=1}^{\nu} Z_i^2 \quad 1.3$$

Donde:

$$Z_i \sim N(0;1) \qquad i = 1,2,\ldots \nu$$

Por otro lado, se dice que una variable aleatoria sigue una distribución chi-cuadrado o ji-cuadrado con υ grados de libertad si su función de densidad viene dada por:

$$f_x(x) = \frac{1}{2^{\nu/2}.\Gamma\left(\frac{\upsilon}{2}\right)}.x^{(\upsilon/2)-1} \qquad x > 0 \qquad (1.4)$$

1.2.1 Características de la distribución

- La distribución chi-cuadrado es una variable aleatoria continua
- Su rango de variación es (0; ∞)
- Es asimétrica positiva y unimodal
- Está definida por un solo parámetro: los grados de libertad υ
- El área total bajo la curva es igual a uno
- La media de la distribución coincide con los grados de libertad: $E(\chi^2) = \upsilon$
- La varianza de la distribución está dada por dos veces los grados de libertad: $Var(\chi^2) = 2\upsilon$

Su forma se presenta en la figura 1.7.

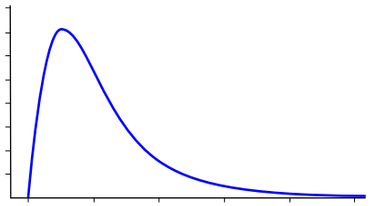


Figura 1.7 Forma de la distribución chi-cuadrado
Fuente: MLB

A medida que se incrementan los grados de libertad, la distribución se aproxima a la distribución normal, tal como se muestra en la figura 1.8.

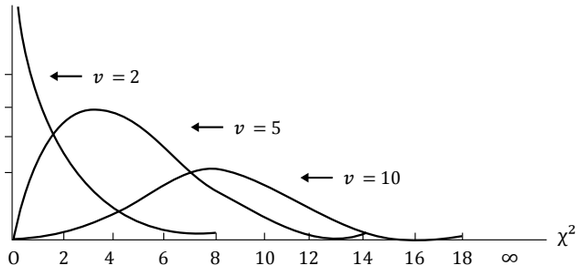


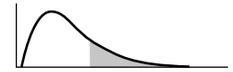
Figura 1.8 Diferentes distribuciones chi-cuadrado
Fuente: Farfán, J., p.8

α

1.2.2 Uso de la tabla chi-cuadrado

En la tabla 1.4, se presenta una parte de los valores críticos de una distribución chi-cuadrado. En la primera fila se muestran las diferentes probabilidades que se encuentran en la cola derecha de la distribución, en la primera columna los valores de los grados de libertad, y los valores en el cuerpo de la tabla representan el valor de la variable chi-cuadrado que deja a su derecha un área específica, tal como se muestra en la figura 1.9.

Tabla 1.4 Valores de la distribución chi-cuadrado



Grados de libertad	0,001	0,005	0,01	0,02	0,025	0,03	0,04	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4
1	10,83	7,879	6,635	5,412	5,024	4,709	4,218	3,841	2,706	2,072	1,642	1,323	1,074	0,873	0,708
2	13,82	10,6	9,21	7,824	7,378	7,013	6,438	5,991	4,605	3,794	3,219	2,773	2,408	2,1	1,833
3	16,27	12,84	11,34	9,837	9,348	8,947	8,311	7,815	6,251	5,317	4,642	4,108	3,665	3,283	2,946
4	18,47	14,86	13,28	11,67	11,14	10,71	10,03	9,488	7,779	6,745	5,989	5,385	4,878	4,438	4,045
5	20,52	16,75	15,09	13,39	12,83	12,37	11,64	11,07	9,236	8,115	7,289	6,626	6,064	5,573	5,132
6	22,46	18,55	16,81	15,03	14,45	13,97	13,2	12,59	10,64	9,446	8,558	7,841	7,231	6,695	6,211
7	24,32	20,28	18,48	16,62	16,01	15,51	14,7	14,07	12,02	10,75	9,803	9,037	8,383	7,806	7,283
8	26,12	21,95	20,09	18,17	17,53	17,01	16,17	15,51	13,36	12,03	11,03	10,22	9,524	8,909	8,351
9	27,88	23,59	21,67	19,68	19,02	18,48	17,61	16,92	14,68	13,29	12,24	11,39	10,66	10,01	9,414
10	29,59	25,19	23,21	21,16	20,48	19,92	19,02	18,31	15,99	14,53	13,44	12,55	11,78	11,1	10,47

Grados de libertad	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,98	0,99	0,995
1	0,571	0,455	0,357	0,275	0,206	0,148	0,102	0,064	0,036	0,016	0,004	1E-03	6E-04	2E-04	4E-05
2	1,597	1,386	1,196	1,022	0,862	0,713	0,575	0,446	0,325	0,211	0,103	0,051	0,04	0,02	0,01
3	2,643	2,366	2,109	1,869	1,642	1,424	1,213	1,005	0,798	0,584	0,352	0,216	0,185	0,115	0,072
4	3,687	3,357	3,047	2,753	2,47	2,195	1,923	1,649	1,366	1,064	0,711	0,484	0,429	0,297	0,207
5	4,728	4,351	3,996	3,655	3,325	3	2,675	2,343	1,994	1,61	1,145	0,831	0,752	0,554	0,412
6	5,765	5,348	4,952	4,57	4,197	3,828	3,455	3,07	2,661	2,204	1,635	1,237	1,134	0,872	0,676
7	6,8	6,346	5,913	5,493	5,082	4,671	4,255	3,822	3,358	2,833	2,167	1,69	1,564	1,239	0,989
8	7,833	7,344	6,877	6,423	5,975	5,527	5,071	4,594	4,078	3,49	2,733	2,18	2,032	1,646	1,344
9	8,863	8,343	7,843	7,357	6,876	6,393	5,899	5,38	4,817	4,168	3,325	2,7	2,532	2,088	1,735
10	9,892	9,342	8,812	8,295	7,783	7,267	6,737	6,179	5,57	4,865	3,94	3,247	3,059	2,558	2,156

Fuente: MLB

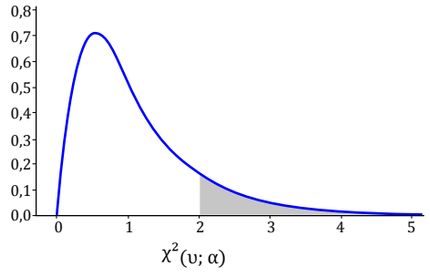


Figura 1.9 Distribución chi-cuadrado cola derecha
Fuente: MLB

Ejemplo 2.1:

Dada una variable aleatoria que sigue una distribución chi-cuadrado con 8 grados de libertad, hallar el valor que dicha variable deja a su derecha un área de α=0.01.

Solución:

En la tabla 1.4 se localiza en la primera fila el valor del α (0,01 ó 1%), y en la primera columna el valor de los grados de libertad, en este caso 8. Luego la intersección de ellos, 20.090, representa el valor de la distribución chi que deja a su derecha un área de 0.01 (ver figura 1.10).

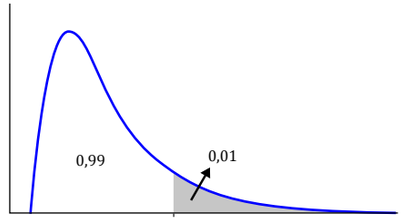


Figura 1.10 Distribución chi-cuadrado con 8 g.l. y cola derecha de α=0.01
Fuente: MLB

Tabla 1.5 Muestra de valores de la distribución chi-cuadrado

Gdos de libertad	0,001	0,005	0,01	0,02
1	10,83	7,879	6,635	5,412
2	13,82	10,6	9,21	7,824
3	16,27	12,84	11,34	9,837
4	18,47	14,86	13,28	11,67
5	20,52	16,75	15,09	13,39
6	22,46	18,55	16,81	15,03
7	24,32	20,28	18,48	16,62
8	26,12	21,95	20,09	18,17
9	27,88	23,59	21,67	19,68

Fuente: MLB

Ejemplo 2.2:

Dada una variable aleatoria chi-cuadrado con 13 grados de libertad, hallar el valor que dicha variable deja a su izquierda un área de α=0.05.

Solución:
En vista de que la tabla 1.4 solo tiene valores para cola derecha y además esta distribución es asimétrica, entonces se debe localizar en la primera fila el valor del complemento que se está pidiendo, 1-α (0,95 o 95 %), y en la primera columna el valor de los grados de libertad, que serían 13. Luego la intersección entre ellos, 5.892, representa el valor de la distribución chi que deja a su derecha un área de 0.05 (ver figura 1.11).

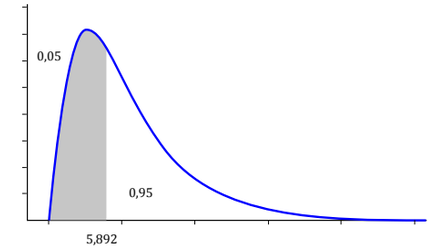


Figura 1.11 Distribución chi-cuadrado con 13 g.l.y cola izquierda de α=0.05
Fuente: MLB

Ejemplo 2.3:

Dada una variable aleatoria chi-cuadrado con 10 grados de libertad, hallar los valores de dicha variable que deja un área total del 5 % a ambos lados de la curva.

Solución:
En primer lugar, se debe dividir el área en cuestión entre dos, eso daría un área a cada lado del gráfico de 0,025 o de 2,5 %.

El valor correspondiente al lado derecho de la gráfica se busca directamente en la tabla, tal como en el ejemplo 2.1, con 10 grados de libertad y un área de 0,025. El valor de la distribución sería de 20,483.

Para el lado izquierdo de la gráfica se procede como en el ejemplo 2.2, con un área acumulada de 0,975 (0,95+0,025), que con 10 grados de libertad da un valor para la distribución de 3,94, tal como se muestra en la figura 1.12.

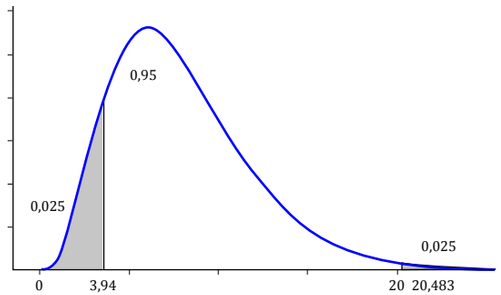


Figura 1.12 Distribución chi-cuadrado con 10 g.l. y dos colas de α=0.05
Fuente: MLB

1.3 Distribución *t-student*[3]

El desarrollo de la distribución de probabilidad *t-student* se debe al químico inglés William Sealy Gosset en el año 1899.

Este químico publicó los hallazgos que hizo mientras trabajaba en el Departamento de Control de Calidad de las destilerías *Guiness* en Dublín bajo el seudónimo de *student* y de ahí el nombre de la distribución.

Para obtener la función de esta distribución, Gosset supuso que las muestras eran tomadas de una población normal. Sin embargo, pudo demostrarse que aun cuando la población no es normal, si la distribución tiene forma acampanada, sigue proporcionando valores que se aproximan bastante a la *t-student*.

La distribución *t-student* surge como el cociente de variables aleatorias independientes: en el numerador una normal estándar (Z), y en el denominador la raíz cuadrada de una chi-cuadrado dividida entre sus grados de libertad (υ). Esto es:

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{\upsilon}}} \sim t_\upsilon \qquad (1.5)$$

De donde:

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1) \qquad (1.6)$$

Y

$$U = \sum_{i=1}^{\upsilon} \frac{(x_i-\bar{x})^2}{\sigma} \sim \chi^2_\upsilon \qquad (1.7)$$

[3] La principal aplicación de la distribución t-student radica en hacer inferencia para una media y diferencia de medias cuando el tamaño de la muestra es pequeño, se desconoce la varianza poblacional y se supone que la población de donde proviene la muestra sigue una distribución normal o aproximadamente normal. En el capítulo cuatro se verán con mayor detenimiento las aplicaciones de esta distribución.

Además, *Z* y *U* son variables aleatorias independientes.

Por otro lado, se dice que una variable aleatoria sigue una distribución *t-student* con u grados de libertad si su función de densidad viene dada por:

$$f_x(x) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)} \cdot \left(1 + \frac{x^2}{\upsilon}\right)^{-\left(\frac{\upsilon+1}{2}\right)} \quad -\infty < x < \infty \qquad (1.8)$$

1.3.1 Características de la distribución

- La *t-student* es una variable aleatoria continua
- Su rango de variación es $(-\infty;\ \infty)$
- Es simétrica y unimodal
- Está definida por su único parámetro: los grados de libertad υ
- El área total bajo la curva es igual a uno
- La media de la distribución es cero: $E\,(t) = 0$
- La varianza de la distribución cuando $\upsilon > 3$ viene dada por: $Var(t) = \frac{\upsilon}{\upsilon - 2}$
- Su forma es la que se presenta en la figura 1.13.

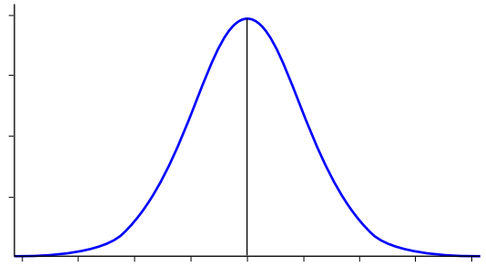


Figura 1.13 Forma de la distribución *t-student*
Fuente: MLB

- Existe una distribución t diferente para cada grado de libertad (figura 1.14)

- A medida que se incrementan los grados de libertad, la distribución se aproxima a la normal (figura 1.14)

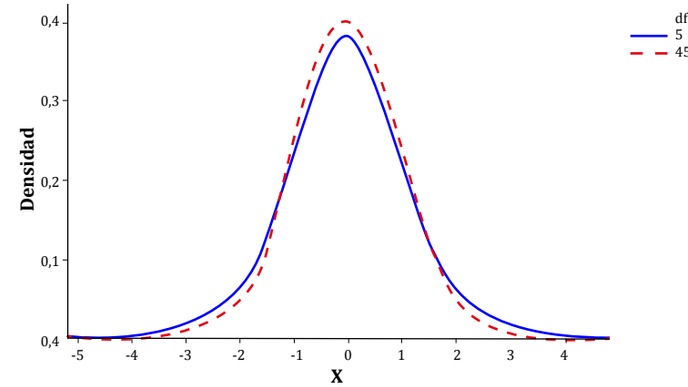


Figura 1.14 Aproximación de la distribución *t-student* a la normal
Fuente: MLB

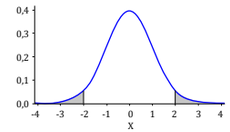
1.3.2 Uso de la tabla *t-student*

En la siguiente tabla se presenta una parte de los valores críticos de una distribución *t-student*. Tal como puede apreciarse, solo existen valores positivos de la distribución, los valores negativos quedan implícitos por la simetría de la misma con respecto al valor promedio de la distribución.

En este sentido, en la primera columna de la tabla 1.6 se encuentran los valores que representan los grados de libertad inherentes a las diferentes distribuciones de probabilidades según la *t-student*.

Con respecto a las filas, en la primera fila de la tabla (α) se especifican las probabilidades ubicadas en la cola superior (figura 1.15.a) o en la cola inferior (figura 1.15.b) de la distribución, mientras que en la segunda fila de la tabla (α) se especifican las probabilidades ubicadas en los dos extremos de la distribución (figura 1.15.c). Cada una de las probabilidades presentadas en la fila 2α es la suma de ambas áreas: de la cola superior y de la cola inferior de la distribución.

Tabla 1.6 Valores de la distribución *t-student*



Gdos de libertad	a	0,4	0,25	0,1	0,05	0,025	0,0125	0,0063	0,0025	0,0013	0,0005
	2α	0,8	0,5	0,2	0,1	0,05	0,025	0,0125	0,005	0,0025	0,001
1		0,3249	1	3,0777	6,3138	12,706	25,452	50,923	127,32	254,65	636,62
2		0,2887	0,8165	1,8856	2,92	4,3027	6,2053	8,8602	14,089	19,962	31,599
3		0,2767	0,7649	1,6377	2,3534	3,1824	4,1765	5,3919	7,4533	9,4649	12,924
4		0,2707	0,7407	1,5332	2,1318	2,7764	3,4954	4,3147	5,5976	6,7583	8,6103
5		0,2672	0,7267	1,4759	2,015	2,5706	3,1634	3,81	4,7733	5,6042	6,8688
6		0,2648	0,7176	1,4398	1,9432	2,4469	2,9687	3,5212	4,3168	4,9807	5,9588
7		0,2632	0,7111	1,4149	1,8946	2,3646	2,8412	3,3353	4,0293	4,5946	5,4079
8		0,2619	0,7064	1,3968	1,8595	2,306	2,7515	3,206	3,8325	4,3335	5,0413
9		0,261	0,7027	1,383	1,8331	2,2622	2,685	3,1109	3,6897	4,1458	4,7809
10		0,2602	0,6998	1,3722	1,8125	2,2281	2,6338	3,0382	3,5814	4,0045	4,5869

Fuente: MLB

Y, por último, el cuerpo interno de la tabla indica el valor de la distribución *t*, a partir del cual a su derecha, a su izquierda o a ambos extremos hay un área determinada, bien sea α 2α.

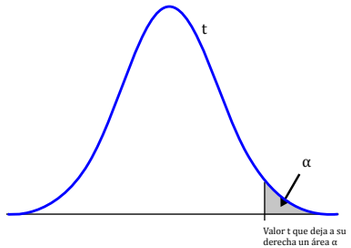


Figura 1.15.a Valores de la distribución *t* con probabilidades ubicadas a la derecha

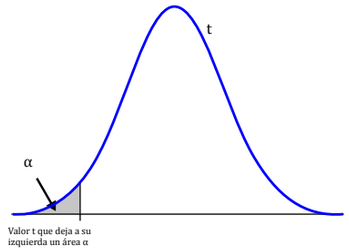


Figura 1.15.a Valores de la distribución *t* con probabilidades ubicadas a la derecha

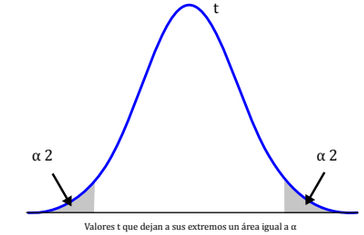


Figura 1.15 Valores de la distribución *t* con probabilidades ubicadas ambos extremos

Fuente: MLB

Ejemplo 3.1:

Dada una variable aleatoria con 8 grados de libertad, hallar el valor *t* que deja a su izquierda un área de 0,05.

Solución:
En la tabla 1.6 se localiza en la primera fila el valor del α (0,05 o 5 %), y en la primera columna el valor de los grados de libertad, en este caso 8. Luego la intersección de ellos, - 1,8595, representa el valor *t* de la distribución, que deja a su izquierda un área de 0,05 (ver figura 1.16).

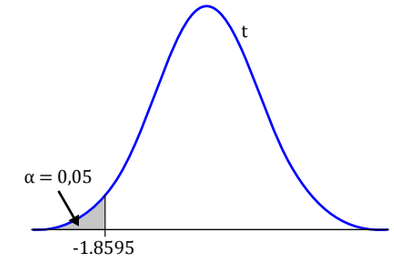


Figura 1.16 Distribución *t* con 8 g.l. y área izquierda de 0,05

Fuente: MLB

Tabla 1.7 Muestra de valores de la distribución *t-student*

	α	0,4	0,25	0,1	0,05
Gdos de libertad	**2α**	**0,8**	**0,5**	**0,2**	**0,1**
1		0,3249	1	3,0777	6,3138
2		0,2887	0,8165	1,8856	2,92
3		0,2767	0,7649	1,6377	2,3534
4		0,2707	0,7407	1,5332	2,1318
5		0,2672	0,7267	1,4759	2,015
6		0,2648	0,7176	1,4398	1,9432
7		0,2632	0,7111	1,4149	1,8946
8		0,2619	0,7064	1,3968	1,8595
9		0,261	0,7027	1,383	1,8331

Fuente: MLB

Ejemplo 3.2:

Dada una variable aleatoria con 8 grados de libertad, hallar el valor *t* que deja a su derecha un área de 0,05.

Solución:
En la tabla 1.6 se localiza en la primera fila el valor del α (0,05 o 5 %), y en la primera columna el valor de los grados de libertad, en este caso 9. Luego la intersección de ellos, 1,8595, representa el valor *t* de la distribución, que deja a su derecha un área de 0,05 (ver figura 1.17).

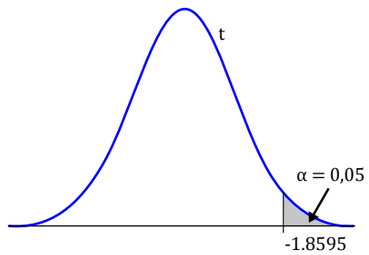


Figura 1.17 Distribución *t* con 8 g.l. y área derecha de 0,05
Fuente: MLB

Ejemplo 3.3:

Dada una variable aleatoria con 9 grados de libertad, hallar el valor *t* que deja a ambos lados un área de 0,05.

Solución:
En la tabla 1.6 se localiza en la primera fila el valor del 2a (0,05 o 5 %), y en la primera columna el valor de los grados de libertad, en este caso 9. Luego la intersección de ellos, 2,2622, representa el valor *t* de la distribución, que deja a ambos lados un área de 0,05.

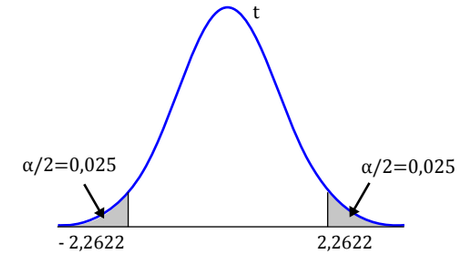


Figura 1.18 Distribución *t* con 9 g.l. y área derecha e izquierda de 0,05
Fuente: MLB

1.4 Distribución *F*[4]

Esta distribución recibe su nombre en honor a R. A. Fisher (1890-1962). Fisher era matemático y biólogo, y fue la primera persona que utilizó los métodos estadísticos para el diseño de experimentos en el año 1919.

Fisher desarrolló técnicas que permiten obtener mayor información significativa a partir de muestras más pequeñas e inició el principio de la aleatoriedad y la técnica del análisis de la varianza.

[4] Son varias las aplicaciones que tiene la distribución F, entre las que destacan la inferencia para el cociente de varianzas y la comparación de dos medias poblacionales simultáneamente a través del análisis de la varianza. En el capítulo seis se verán con mayor detenimiento.

La distribución F surge como el cociente de variables aleatorias independientes chi-cuadrado divididas entre sus respectivos grados de libertad. Esto es:

$$F = \frac{\left(\frac{\chi_1^2}{\upsilon_1}\right)}{\left(\frac{\chi_2^2}{\upsilon_2}\right)} \sim F_{(\upsilon_1;\upsilon_2)} \qquad (1.9)$$

De donde:

$$\chi_1^2 \sim \chi^2(\upsilon_1) \text{ y } \chi_2^2 \sim \chi^2(\upsilon_2)$$

Además χ_1^2 y χ_2^2 son independientes.

La función de densidad de una variable aleatoria que sigue una distribución F con υ_1 y υ_2 grados de libertad viene dada por:

$$f_x(x) = \frac{\Gamma\left(\frac{\upsilon_1+\upsilon_2}{2}\right)}{\Gamma\left(\frac{\upsilon_1}{2}\right)\Gamma\left(\frac{\upsilon_2}{2}\right)} \cdot \left(\frac{\upsilon_1}{\upsilon_2}\right)^{\upsilon_1/2} \cdot x^{(\upsilon_1/2)-1} \left(1+\frac{\upsilon_1}{\upsilon_2}x\right)^{-\left(\frac{\upsilon_1+\upsilon_2}{2}\right)} \quad \text{si} \quad x > 0 \qquad (1.10)$$

En caso contrario, $f_x(x) = 0$

1.4.1 Características de la distribución

- La distribución F es una variable aleatoria continua
- Su rango de variación es $(0;\ \infty)$
- Es asimétrica positiva cuando $\upsilon_1 + \upsilon_2$ son pequeños
- Está definida por s únicos parámetros: los grados de libertad del numerador y del denominador ($\upsilon_1 y\ \upsilon_2$), respectivamente
- El área total bajo la curva es igual a uno
- La media de la distribución está definida cuando el número de grados de libertad del denominador υ_2 es mayor a 2 cero:

$$E(t) = \frac{\upsilon_2}{\upsilon_2 - 2} \qquad (1.11)$$

- La varianza de la distribución cuando $\upsilon_2 > 4$ viene dada por:

$$Var(t) = \frac{2\,\upsilon_2^2(\upsilon_1 + \upsilon_2 - 2)}{\upsilon_1(\upsilon_2 - 2)(\upsilon_2 - 4)} \qquad (1.12)$$

- Su forma se presenta en la figura 1.19.

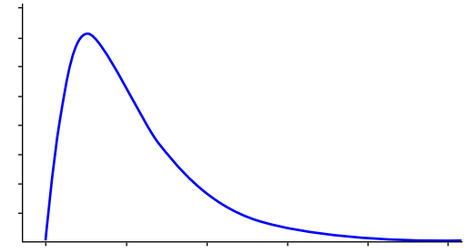


Figura 1.19 Forma de la distribución F

Fuente: MLB

- Existe una distribución F diferente para cada grado de libertad
- A medida que se incrementan los grados de libertad, la distribución se aproxima a la distribución normal, tal como se muestra en la figura 1.20.

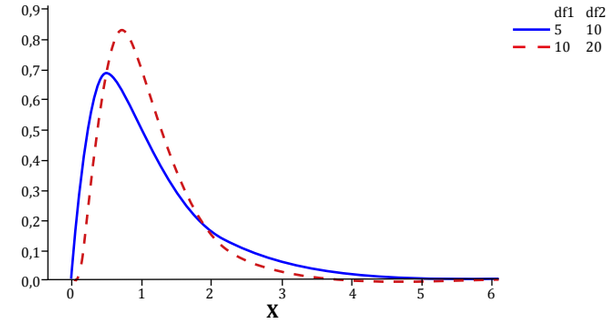


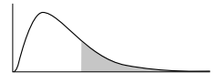
Figura 1.20 Distribución F para distintos grados de libertad

Fuente: MLB

1.4.2 Uso de la tabla *F*

En la tabla 1.8 se presenta una parte de los valores críticos de una distribución *F*. Se aprecia que solo existen valores positivos de la distribución, ya que el rango es entre cero e infinito.

Tabla 1.8 Valores de la distribución F (α = 0,05)



	1	2	3	4	5	6	7	8	9	10
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494

Fuente: MLB

Hay diferentes tablas *F* según el nivel de significación que se esté estudiando. Para ilustrar se tomó en cuenta la probabilidad de rechazar la hipótesis nula cierta más común entre todos, que es 005.

En este sentido, se tienen en la primera fila los grados de libertad del numerador (υ_1) y en la primera columna los grados de libertad del denominador (υ_2).

La tabla 1.8, contiene los valores de una variable que se distribuye según una F, tal que a su derecha se encuentra un área igual a alfa (α); estos valores se denotan por . El subíndice consta del valor de α que se refiere al

porcentaje dejado en la cola derecha de la distribución, (υ_1) a los grados de libertad del numerador y (υ_2)a los grados de libertad del denominador (ver figura 1.21).

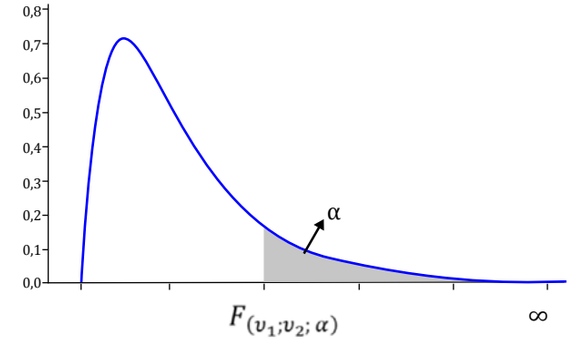


Figura 1.21 Distribución *F* con (υ_1) y (υ_2) grados de libertad
Fuente: MLB

Ejemplo 4.1:

Dada una variable aleatoria que sigue un comportamiento *F*, con 6 y 15 grados de libertad para el numerador y denominador respectivamente, hallar el valor *F* que deja un área de 5 % a su derecha.

Solución:
En la tabla 1.8, que es la correspondiente a un nivel de significación de 0,05, se localiza en la primera fila el valor del grado de libertad del numerador, que en este caso es 6, y en la primera columna los grados de libertad del denominador, 15. Luego la intersección de estos dos valores, 2,79, representa el valor *F* de la distribución, que a la derecha un área de 0,05 (ver figura 1.22).

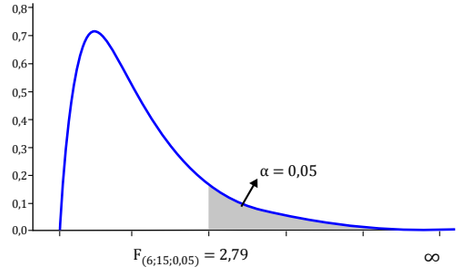


Figura 1.22 Distribución *F* con 6 y 15 grados de libertad cola derecha
Fuente: MLB

Tabla 1.9 Muestra de valores de la distribución *F* (α = 0,05)

	1	2	3	4	5	6
1	161,448	199,500	215,707	224,583	230,162	233,986
2	18,513	19,000	19,164	19,247	19,296	19,330
3	10,128	9,552	9,277	9,117	9,013	8,941
4	7,709	6,944	6,591	6,388	6,256	6,163
5	6,608	5,786	5,409	5,192	5,050	4,950
6	5,987	5,143	4,757	4,534	4,387	4,284
7	5,591	4,737	4,347	4,120	3,972	3,866
8	5,318	4,459	4,066	3,838	3,687	3,581
9	5,117	4,256	3,863	3,633	3,482	3,374
10	4,965	4,103	3,708	3,478	3,326	3,217
11	4,844	3,982	3,587	3,357	3,204	3,095
12	4,747	3,885	3,490	3,259	3,106	2,996
13	4,667	3,806	3,411	3,179	3,025	2,915
14	4,600	3,739	3,344	3,112	2,958	2,848
15	4,543	3,682	3,287	3,056	2,901	2,790

Ejemplo 4.2:

Dada una variable aleatoria que sigue un comportamiento *F*, con 3 y 8 grados de libertad para el numerador y denominador respectivamente, hallar el valor *F* que deja a su izquierda un área de 5 %.

Solución:
En este caso no se puede usar directamente la tabla 1.8 con υ_1=3 y υ_2=8, pues se trata de una tabla que toma en cuenta solamente valores que dejan a su derecha un área determinada y cuyos valores posibles son: 0,01, 0,05, 0,025 y 0,10. Por otro lado, tampoco es posible ir a la tabla con la probabilidad de 1−α que deje a su derecha, porque no se tiene la tabla con dicha probabilidad.

En este sentido, para obtener el valor *F* correspondiente a la cola izquierda se utiliza la propiedad recíproca de la distribución *F*, que consiste en invertir los grados de libertad del numerador por el del denominador y viceversa, y calcular el recíproco del mismo. Esto es:

$$F_{cola\ izquierda\ (\upsilon_1;\upsilon_2;\alpha)} = \frac{1}{F_{cola\ derecha\ (\upsilon_2;\upsilon_1;\alpha)}} \qquad (1.13)$$

Por lo tanto, para el caso que se requiere una $F_{cola\ izquierda\ (3;8;0.05)}$, se intercambian los grados de libertad, $F_{cola\ derecha\ (8;3;0.05)}$ y se saca el recíproco:

$$F_{cola\ izquierda\ (3;8;0.05)} = \frac{1}{F_{cola\ derecha\ (8;3;0.05)}} = \frac{1}{8.845} = 0.1125$$

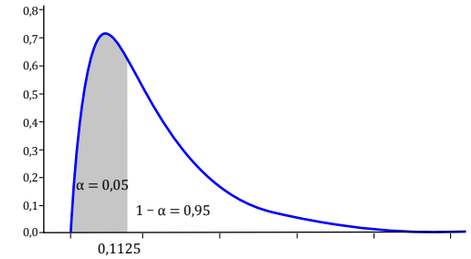


Figura 1.23 Distribución *F* con 3 y 8 grados de libertad cola izquierda
Fuente: MLB

Ejemplo 4.3:

Dada una variable aleatoria que sigue un comportamiento *F*, con 3 y 8 grados de libertad, hallar el valor *F* que deja un área total del 10 % en ambos extremos.

Solución:
Tal como se dijo en el ejemplo 4.2, las tablas utilizadas no proporcionan los valores de la distribución *F* para cola izquierda; en ese sentido, se deben intercambiar los grados de libertad y sacar el inverso. Por otro lado, se debe dividir el área que se está utilizando entre dos (ver figura 1.24).

$$F_{cola\ derecha\ (3;8;005)} = 4.066$$

$$F_{cola\ izquierda\ (3;8;0.05)} = \frac{1}{F_{cola\ derecha\ (8;3;0.05)}} = \frac{1}{8.845} = 0.1125$$

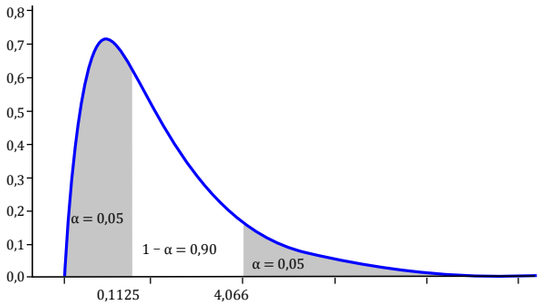


Figura 1.24 Distribución *F* con 3 y 8 grados de libertad dos colas
Fuente: MLB

Ejercicios

1. Dada una variable aleatoria *t-student* con 15 grados de libertad, hallar el valor *t* que deja a la izquierda un área de 0,10.
2. Dada una variable aleatoria chi-cuadrado con 22 grados de libertad, hallar el valor chi que deja a ambos lados un área de 0,05.
3. Dada una variable aleatoria que sigue una distribución *F*, con 6 y 21 grados de libertad para el numerador y denominador respectivamente, hallar el valor *F* que deja a la derecha un área de 0,01.
4. Dada una variable aleatoria *t-student* con 18 grados de libertad, hallar el valor *t* que deja a la derecha un área de 0,01.
5. Dada una variable aleatoria con distribución chi-cuadrado de 6 grados de libertad, hallar el valor chi que deja al lado izquierdo un área de 0,125.
6. Dada una variable aleatoria normal estándar, cuál será la probabilidad de:
   i. Pr(-0,59 < Z ≤ -0,24)
   ii. Pr(Z ≤ 2,35)
   iii. Pr(Z > -1,34)
7. Investigue cómo obtener los valores de los ejercicios 1 al 6 utilizando Excel. Verifique los resultados.

2. Distribución en el muestreo

DISTRIBUCIÓN EN EL MUESTREO

2.1 Población y muestra

El desarrollo de la inferencia estadística se basa en tres conceptos fundamentales: universo, población y muestra.

Universo: un conjunto, finito o infinito, de seres vivos, elementos o cosas sobre las cuales están definidas características o variables que interesan analizar. Se puede hablar del universo de todos los jefes de hogar de un departamento de Colombia, el universo de todos los grupos familiares de una región, el universo de los estudiantes universitarios de un país, etc.

Población: la totalidad de mediciones que se pueden hacer de una característica en estudio en un lugar y momento determinado. La población es finita o infinita. Se puede hablar de la población constituida por las estaturas de los jefes de hogar de una región, la población de los ingresos mensuales de los grupos familiares de una región, la población de las edades de los estudiantes universitarios del país, etc.

Es importante resaltar que el universo está compuesto por el conjunto de elementos sobre los cuales se mide al menos una característica, mientras que la población la constituyen los valores de las mediciones de esas variables en los elementos en cuestión.

Un mismo universo puede dar origen a diferentes poblaciones. Por ejemplo, el universo de los jefes de hogar de un departamento de Colombia puede generar la población de estatura, la población de ingresos, entre otras.

Según el número de unidades que tenga una población, se puede considerar finita o infinita. Una población es finita cuando está compuesta de una cantidad limitada de elementos. Por ejemplo, el número de estudiantes, el número de obreros, etc. Una población infinita es la que tiene un número extremadamente grande de componentes, como el conjunto de especies que tiene el reino animal, la totalidad de los potenciales rendimientos de plantas de maíz a las cuales se les aplica cierta cantidad de un fertilizante, etc.

Muestra: un subconjunto del conjunto de elementos que constituyen la población; en otras palabras, es una parte o porción de la población debidamente seleccionada con la finalidad de analizar y sacar conclusiones sobre ciertas propiedades de la población.

Siempre es deseable que la muestra represente en pequeña escala a la población de la cual se extrajo la misma en cuanto a sus propiedades, pero dado que de una misma población se pueden seleccionar diferentes muestras, es posible que no ocurra en la seleccionada. Este es uno de los riesgos que se asume cuando se utilizan muestras.

2.2 Censo, muestreo y tipos de muestreo

a. Censo: al realizar cualquier estudio sobre una población, puede utilizarse una muestra o analizar todos los elementos de ella. Este último procedimiento se conoce como censo.
b. Muestreo: proceso mediante el cual se selecciona una muestra de la población. El muestreo puede ser probabilístico o no probabilístico.

Se dice que el muestreo (y la muestra) es probabilístico cuando cada uno de los elementos de la población tiene una determinada probabilidad de formar parte de la muestra o, alternativamente, cuando cada una de las muestras de tamaño n que es posible seleccionar de esa población tiene una probabilidad específica de ser elegida.

En el caso particular de que cada uno de los elementos de la población tengan la misma probabilidad de integrar la muestra, y que todas las muestras de tamaño n tengan la misma probabilidad de ser elegidas, se

dice que el muestreo es aleatorio simple y que las muestras son muestras aleatorias (simples). El muestreo probabilístico y específicamente el aleatorio simple constituyen el soporte fundamental de la inferencia estadística. Por otro lado, en la teoría de probabilidades se utilizan muestras aleatorias en el desarrollo e ilustración de su formulación.

El muestreo probabilístico comprende, además del muestreo aleatorio simple, otros métodos, tales como el muestreo sistemático, el muestreo estratificado, el muestreo por conglomerados, entre otros. Cada uno de estos métodos de muestreo tienen un desarrollo teórico particular y la aplicación de cualquiera de ellos es conveniente realizarla con la asesoría de profesionales calificados de la estadística. Para un tratamiento detallado de los métodos de muestreo, se recomienda al lector consultar bibliografía especializada.

El muestro no probabilístico, también llamado muestreo subjetivo o por conveniencia, se caracteriza porque en la elección de los elementos de la muestra interviene el conocimiento y la opinión de la persona que realiza la selección. Usualmente, la muestra se selecciona de acuerdo con la conveniencia y comodidad de la persona. En este tipo de muestreo se ubican: muestreo por cuotas, muestreo bola de nieve, muestreo por juicio (opinión) y muestreo sin norma (por conveniencia).

El muestreo no probabilístico es el que intuitiva o lógicamente se emplea en la vida diaria, por ejemplo, cuando se juzga la calidad o el sabor de una torta analizando o probando un pedacito de ella. Este tipo de muestreo no requiere ningún conocimiento técnico especial y no tiene ninguna utilidad para realizar inferencias estadísticas, tal como se explica a continuación.

2.3 Parámetros y estadísticos

a. **Parámetros:** cualquier característica numérica de una población que describa alguna propiedad particular de esa población.

   Usualmente los parámetros se denotan con letras griegas: la letra μ para la medida poblacional, σ^2 para la varianza, π para la desviación estándar poblacional y π o P para la proporción poblacional.

Si se considera la población constituida por los sueldos mensuales de los empleados de la industria petrolera del país, son ejemplos de parámetros de esa población: el sueldo promedio mensual de la industria, la varianza poblacional de esos sueldos, la proporción de empleados cuyos sueldos están por encima de cierto valor, etc. Y en general, cualesquiera de las medidas descriptivas numéricas de localización o variabilidad de los datos constituyen parámetros poblacionales.

Se debe tener en cuenta que los parámetros poblacionales son valores únicos para una población determinada; es decir, no cambian al menos que cambie la población. Además, para determinar el valor de un parámetro es necesario conocer toda la población. Sin embargo, dado que generalmente no se trata con toda la población, sino con una muestra de ella, entonces los parámetros poblacionales son usualmente valores desconocidos. La inferencia estadística pretende justamente realizar estimaciones de los parámetros poblacionales sobre la base de la información que proporciona una muestra.

b. Estadísticos: de manera similar a como se hace con una población, se pueden definir indicadores numéricos asociados a una muestra, los cuales van a reflejar propiedades de ella y a depender exclusivamente de los valores o elementos de esa muestra. Estos indicadores se denominan estadísticos. Por ejemplo, la media aritmética muestral, la mediana, la moda, los percentiles, el rango, la varianza, la desviación estándar, las proporciones y porcentajes muestrales, entre otras. En lenguaje matemático, un estadístico es una función (matemática) de los valores muestrales que no depende de ningún valor desconocido. Si denotamos por $x_1, x_2, \dots, x_n$ una muestra aleatoria de tamaño n proveniente de una población determinada, entonces un estadístico será una función $f(x_1, x_2, \dots, x_n)$ que no contenga ningún término desconocido. Las funciones que se muestran a continuación constituyen ejemplos de estadísticos:

- **i. La medida de dispersión denominada rango:**

$$Rgo = valor\ max(x_i) - valor\ min(x_i)$$

- **ii. La media muestral:**

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

- **iii. La mediana de la distribución:**

$$Med = X_{\left(\frac{x+1}{2}\right)}$$

- **iv. El cuartil *h***

$$Q_h = L_i + \left(\frac{\frac{h.n}{4} - F_i^{\uparrow}}{f_i}\right) C_i \quad [5]$$

Note que las expresiones dadas en los numerales a, b, c y d solo dependen de los valores de la muestra. Si se considera que la media poblacional, μ, es desconocida, entonces la expresión $A = \overline{X} - \mu$ no es un estadístico, ya que depende del parámetro desconocido μ.

Es interesante observar que cualquier medida descriptiva numérica (media, mediana, moda, varianza, proporción, etc.), cuando se maneja en el contexto de una población, constituye un parámetro, mientras que si está referida a una muestra, representa un estadístico.

Los estadísticos cumplen un papel protagónico en el proceso de inferencia estadística. Por ejemplo, cuando se quiere estimar, sobre la base de una muestra, el valor de algún parámetro poblacional desconocido o tomar una decisión sobre cualquier hipótesis o teoría que se plantee acerca del valor de ese parámetro, habrá que utilizar procedimientos y metodologías basados en estadísticos.

[5] Ecuación del cuartil cuando los datos están agrupados en distribución de frecuencias: h indica el cuartil solicitado, $F_i^{\uparrow}$ es la frecuencia acumulada hasta la clase anterior al cuartil, f_i es la frecuencia absoluta de la clase donde se encuentra el cuartil y C_i es la amplitud de dicha clase.

2.4 Distribución muestral de un estadístico

Observe que un estadístico es una función que hace corresponder a cada muestra de tamaño n, que sea posible seleccionar de la población, un único número real, el cual representa el valor que toma el estadístico en esa muestra. En consecuencia, un estadístico puede tomar diferentes valores dependiendo de cuál sea la muestra que le corresponda.

Ejemplo 2.1:

Considere una población ficticia constituida por los años de servicio de tres empleados de una compañía A {2; 4; 6} y seleccione una muestra aleatoria de tamaño n=2 con reposición y en la cual el orden en que aparecen los elementos se utiliza como criterio de diferenciación de las muestras, esto es, la muestra {2,4} es diferente de la muestra {4,2}.

Si se denota la muestra por {x1, x2} y se considera la media aritmética $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ de los años de servicio de cada muestra, se tiene que esa función $(\bar{X})$ es un estadístico, ya que solo depende de los valores muestrales, tal como se muestra en la tabla 2.1.

Tabla 2.1 Muestras de tamaño 2 y sus respectivos promedios muestrales

Muestras posibles	Media muestral $(\bar{X})$
{2;2}	2
{2;4}	3
{2;6}	4
{4;2}	3
{4;4}	4
{4;6}	5
{6;2}	4
{6;4}	5
{6;6}	6

Fuente: MLB

En virtud de que un estadístico asigna un único número real a cada muestra y que cada muestra es un posible resultado del experimento aleatorio que consiste en seleccionar una muestra aleatoria de una población,

entonces se concluye que todo estadístico es una variable aleatoria y, en consecuencia, tendrá una determinada distribución de probabilidad.

Se denomina *distribución muestral de un estadístico* a la distribución de probabilidad que sigue ese estadístico.

La distribución de un estadístico puede ser de tipo discreto o continuo. Además contará con todos los elementos que caracterizan a cualquier distribución de probabilidad, como, por ejemplo, media o esperanza, varianza, desviación estándar, forma de la distribución, asimetría, etc. En consecuencia, se puede hablar de la distribución de probabilidad de la media muestral $(\bar{X})$, de la distribución de probabilidad de la varianza muestral S^2, de la distribución de probabilidad de la proporción muestral p, etc., así como también de la media, varianza y desviación estándar de esas distribuciones o de esos estadísticos. A la desviación estándar de un estadístico también se acostumbra denominarla *error estándar del estadístico*.

El valor esperado o promedio de la media muestral $(\bar{X})$ se denota como $E(\bar{X})$ o $\mu_{\bar{X}}$; su varianza, $Var(\bar{X})$ o $\sigma^2_{(\bar{X})}$, y su error estándar, $DE(\bar{X})$ o $\sigma_{\bar{X}}$, y están representados respectivamente por:

a. $$E(\bar{X}) = \mu_{\bar{X}} = \sum \bar{X}\, P(\bar{X} = \bar{X}) \qquad (2.1)$$

b. $$V(\bar{X}) = \sigma^2_{\bar{x}} = \sum_{i=1}^{n} (\bar{X} - \mu_{\bar{x}})^2 \times \Pr\,(\bar{X} = \bar{x}) \qquad (2.2)$$

c. $$DE(\bar{X}) = \sigma_{(\bar{X})} = \sqrt{\sum_{i=1}^{n} (\bar{X} - \mu_{\bar{x}})^2 \times \Pr\,(\bar{X} = \bar{x})} \qquad (2.3)$$

A continuación, se ilustrará mediante un ejemplo un caso sencillo de la distribución muestral de la media aritmética.

Ejemplo 2.2:

Hallar la distribución de probabilidad muestral de la media aritmética para el ejemplo 1, es decir, obtener todos los posibles valores que asume

la variable aleatoria $(\bar{X})$, acompañada por sus respectivas probabilidades: $\Pr(\bar{X} = \bar{x})$.

En este caso, $(\bar{X})$ solo puede tomar un número finito de valores y, en consecuencia, su distribución de probabilidad será de tipo discreta.

En cuanto a los valores de todas las muestras posibles de tamaño n = 2 de la tabla 2.1, se observa que el estadístico $(\bar{X})$ puede tomar los valores 2, 3, 4, 5 y 6. Tomando en cuenta que las muestras son aleatorias y cada una con la misma probabilidad de ser seleccionada, se tiene que las correspondientes probabilidades para los valores de $(\bar{X})$ son:

- $\Pr(\bar{X} = 2) = \Pr(\{2;2\}) = 1/9$
- $\Pr(\bar{X} = 3) = \Pr(\{2;4\};\{4;2\}) = 2/9$
- $\Pr(\bar{X} = 4) = \Pr(\{2;6\};\{4;4\};\{6;2\}) = 3/9$
- $\Pr(\bar{X} = 5) = \Pr(\{4;6\};\{6;4\}) = 2/9$
- $\Pr(\bar{X} = 6) = \Pr(\{6;6\}) = 1/9$

Por lo que la distribución muestral de $(\bar{X})$ para este ejemplo viene dado por:

$\bar{X}$	2	3	4	5	6
$\Pr(\bar{X} = \bar{X})$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$

De esta manera, utilizando las ecuaciones (2.1), (2.2) y (2.3), la esperanza matemática $(E(\bar{X}))$ la varianza $(V(\bar{X}))$ y la desviación estándar $(DE(\bar{X}))$ de esta distribución son:

a. $E(\bar{X}) = \mu_{\bar{X}} = \sum \bar{X}\, P(\bar{X} = \bar{X}) = 2\left(\frac{1}{9}\right) + 3\left(\frac{2}{9}\right) + 4\left(\frac{3}{9}\right) + 5\left(\frac{2}{9}\right) + 6\left(\frac{1}{9}\right) = 4$

b. $V(\bar{X}) = \sigma_{\bar{x}}^2 = \sum_{i=1}^{n} (\bar{X} - \mu_{\bar{x}})^2 \times \Pr(\bar{X} = \bar{x})$ $(2-4)^2 \times \left(\frac{1}{9}\right) + (3-4)^2 \times \left(\frac{2}{9}\right) +$

$(4-4)^2 \times \left(\frac{3}{9}\right) + (5-4)^2 \times \left(\frac{2}{9}\right) + (6-4)^2 \times \left(\frac{1}{9}\right) = \frac{8}{9} =$

c. $DE(\bar{X}) = \sigma_{(\bar{X})} = \sqrt{\sum_{i=1}^{n} (\bar{X} - \mu_{\bar{x}})^2 \times \Pr(\bar{X} = \bar{x})} = \sqrt{\frac{8}{9}} = 0{,}1047$

2.5 Distribución de la media muestral

Una de las acciones más comunes cuando se hace un análisis de tipo estadístico es realizar inferencias acerca de la media μ de una población. Estas inferencias se basan en el estadístico media muestral $\bar{X}$, lo cual debe resultar un hecho natural y lógico. Es por esto que se debe investigar sobre la distribución muestral de este estadístico, incluyendo forma o tipo de su distribución de probabilidad y algunas de sus características, como la media y la varianza.

Antes de considerar la forma que debe seguir la distribución de probabilidad de la media muestral $\bar{X}$, se debe destacar un resultado sustancial acerca de la esperanza y la varianza de este estadístico.

Considérese el estadístico $\bar{X}$, basado en una muestra aleatoria de tamaño n, la cual se denota como $x_1, x_2, \ldots, x_n$, donde x_i tiene la misma distribución de probabilidad de la población. Entonces se cumple que:

$$E(x_1) = E(x_2) = \cdots = E(x_n) = \mu$$
$$Var(x_1) = Var(x_2) = \cdots = Var(x_n) = \sigma^2$$

Tomando en cuenta que:

$$\bar{X} = \frac{x_1, x_2, \ldots, x_n}{n}$$

Se pueden usar las propiedades del valor esperado y de la varianza para obtener:

a. $E(\bar{X}) = \frac{1}{n} E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)]$
$= \frac{1}{n}(\mu + \mu + \cdots + \mu)\frac{n\mu}{n} = \mu$

b. $Var(\bar{X}) = \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n) = \frac{1}{n^2}[Var(X_1) + Var(X_2) + \cdots +$
$Var(X_n)] = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots \sigma^2) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

Dada una población cualquiera con media μ y varianza σ^2, la media, la varianza y el error estándar del estadístico media muestral ($\bar{X}$) son $E(\bar{X}) = \mu, Var = \frac{\sigma^2}{n}$ y $DE = \frac{\sigma}{\sqrt{n}}$ respectivamente.

Ejemplo 2.3:

Volviendo al ejemplo 2.1 (una población ficticia constituida por los años de servicio de tres empleados de una compañía A {2; 4; 6}), la media y la varianza de esa población son:

$$\mu = \frac{\sum_{i=1}^{3} X_i}{N} = \frac{(2+4+6)}{3} = 4$$

$$\sigma^2 = \sum_{i=1}^{3} \frac{(X_i - \mu)^2}{N} = \frac{(2-4)^2}{3} + \frac{(4-4)^2}{3} + \frac{(6-4)^2}{3} = \frac{8}{3}$$

Ahora bien, de este último resultado se pueden conocer directamente los valores de la media y de la varianza de la media de la muestra. Esto es,

$$E(\bar{X}) = \mu = 4 \text{ y } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{\left(\frac{8}{3}\right)}{3} = \frac{8}{9}$$

que fueron los valores obtenidos en el ejemplo 2.2.

2.6 Tipos de distribución de la media muestral ($\bar{X}$)

Una vez conocidas la media y la varianza de $\bar{X}$, se puede determinar el tipo o la forma de la correspondiente distribución de probabilidad. En ese sentido, se van a presentar dos situaciones:

a. Cuando la población de la cual se selecciona la muestra sigue una distribución normal
b. Cuando la población de la cual se selecciona la muestra no tiene una distribución normal

Con respecto al punto a, se obtiene el siguiente resultado:

Cuando la población de origen sigue una distribución normal con media μ y varianza σ^2 , el estadístico $\bar{X}$, basado en una muestra de tamaño n, sigue una distribución normal con media μ y varianza $\frac{\sigma^2}{n}$ [6]

Esta afirmación permite concluir automáticamente que la forma de la distribución de $\bar{X}$ es normal cuando la población de la cual se muestrea es normal.

Ejemplo 2.4:

Calcular la probabilidad de que la edad promedio de una muestra aleatoria de 15 trabajadores de una compañía esté comprendida entre 40 y 45 años, sabiendo que la edad de todos los trabajadores sigue una distribución aproximadamente normal con media 38 años y desviación estándar de 3,4 años.

Solución:
En primer lugar, se debe definir la variable aleatoria:
$\bar{X}$: edad promedio de los trabajadores

Se sabe que:

a. E ($\bar{X}$) = 38
b. DE ($\bar{X}$) = 3,4/ $\sqrt{15}$ = 0,88

Por otra parte, dado que la población es normal, entonces $\bar{X}$ también sigue una distribución normal con parámetros N (38;0.88).

En consecuencia, usando la tabla I del apéndice A tenemos:

$$Pr\left(\frac{40-38}{0{,}88} < \frac{X-38}{0{,}88} < \frac{45-38}{0{,}88}\right) = Pr\,(2{,}24 < Z < 7{,}95)$$

$$Pr(40 < \bar{X} < 45) = Pr(Z < 7{,}95) - Pr(Z < 2{,}27) = 1 - 0{,}9884 = 0{,}116$$

[6] Meyer, 1973:259.

Respecto al segundo caso, cuando la población no es normal, la distribución de $\bar{X}$ dependerá del tipo de distribución de la población y habrá que obtenerla utilizando cálculos estadísticos apropiados, los cuales pueden resultar muy sencillos o con cierto grado de complejidad.

Adicionalmente, existe un caso muy particular en el cual se puede obtener de manera aproximada el tipo de distribución de probabilidad de $\bar{X}$. Este caso no depende de la forma que tenga la distribución de la población, sino más bien de la magnitud del tamaño de la muestra. Y es tratado a través de uno de los teoremas más importantes que existe en el campo de la estadística, conocido como el teorema central del límite.

2.7 Teorema del límite central (TCL)

La distribución de probabilidad del estadístico $\bar{X}$, basada en una muestra aleatoria de tamaño n, de una población cualquiera con media μ y varianza σ^2 es aproximadamente normal con media μ y varianza $\frac{\sigma^2}{n}$ cuando el tamaño de n es suficientemente grande[7].

En relación con este teorema, es razonable comentar que:

a. Su importancia radica en que proporciona la distribución de probabilidad aproximada de $\bar{X}$ independientemente de cual sea la distribución de la población. Ya sea una distribución discreta o continua, simétrica o asimétrica, la distribución de $\bar{X}$ será aproximadamente normal para un valor grande de la muestra n.
b. Evidentemente el TCL tiene aplicación cuando no se conoce la distribución poblacional o cuando se sabe que no es normal. Cuando la población no es normal, sea n grande o pequeño, la distribución de $\bar{X}$ es normal y exacta.
c. Observe que la distribución de la media muestral siempre presentará menos variabilidad que la distribución poblacional, debido a que la varianza poblacional va dividida entre el tamaño de la muestra.
d. El TCL es de gran utilidad en problemas de inferencia estadística cuando se desconocen los valores de μ o σ y solo se dispone de una muestra en la cual se pueden calcular $\bar{X}$ y otros estadísticos, con los que se harán estimaciones apropiadas de los parámetros desconocidos.
e. En cuanto a la aplicación del TCL en casos reales, es necesario dilucidar qué se entiende por un tamaño de muestra "suficientemente grande". Es claro que mientras menos se parezca la distribución poblacional a la distribución normal, mayor tendrá que ser el valor de n para que la aproximación sea buena. Se ha aceptado satisfactoriamente un n > 30 para la aplicación del teorema.

[7] Moore, 1995:304.

Ejemplo 2.5:

Se sabe que la temperatura de enfriamiento de las neveras de cierta marca de una compañía determinada es de -3 °C, con una varianza de 5,2 °C^2. Si se tiene una muestra aleatoria de 40 neveras, cuál es la probabilidad de que:

a. ¿La nevera tenga una temperatura de enfriamiento promedio superior a -2 °C?
b. ¿La nevera tenga una temperatura de enfriamiento promedio menor a -1,5 °C?

Solución:
En este caso, la media de la población es -3 °C y la varianza es 5,2 °C^2. Por lo tanto, la desviación estándar, que es la raíz de la varianza, es 2,28 °C. En vista de que el tamaño de la muestra es grande, se puede emplear el TCL. Así, la temperatura de enfriamiento promedio de las neveras de la muestra se distribuye aproximadamente normal, con media y desviación estándar dadas por:

$$\mu = -3^oC \text{ y } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2{,}28}{6{,}32} = 0{,}361$$

Por lo tanto,

a. $Pr(\bar{X} > -2) = 1 - Pr\ (\bar{X} < -2) = 1 - Pr\left(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{-2-(-3)}{0{,}361}\right)$
$= 1 - Pr(Z < 2{,}77) = 1 - 0{,}9972 = 0{,}0028$

b. $Pr(\bar{X} < -1{,}5) = Pr\left(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{-1{,}5-(-3)}{0{,}361}\right) = Pr(Z < 4{,}15) = 0{,}99998$

Se utilizó la fórmula del error estándar de la media muestral, $\sigma_{\bar{x}}$ como el cociente entre la desviación estándar de la muestra y la raíz del tamaño de la misma, $\left(\frac{\sigma}{\sqrt{n}}\right)$ Cabe destacar que esta ecuación se usa cuando se considera que la población es infinita o cuando, a pesar de que la población es finita, el muestreo se realiza con reposición. Ahora bien, si el muestreo que se está realizando es sin reposición, la fórmula para el error estándar de la media viene dado por:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{n-1}}\right) \quad (2.4)$$

Donde:

σ = desviación estándar poblacional
N= tamaño de la población
n = tamaño de la muestra

El término $\sqrt{\frac{N-n}{n-1}}$ se conoce como factor de corrección para población finita (fcf) y se utilizará para calcular la desviación estándar de la media muestral cuando la población sea finita y el tamaño de la muestra no sea pequeño en relación con el tamaño de la población $\frac{n}{N} > 0{,}05)$ (. En los demás casos se usará $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

2.8 Distribución de la proporción muestral ($\bar{X}$)

Muchas veces se puede estar interesado en investigar otras características de una población que no sea necesariamente la media de la misma. Por ejemplo, se quisiera conocer la proporción de piezas defectuosas de un proceso de producción o la proporción de personas que están a favor de cierto candidato en las próximas elecciones; en estos casos, la distribución muestral de proporciones es la más adecuada.

La proporción no es más que un promedio resultante de una variable que toma 2 posibles valores: 0 (ausencia de la característica de interés) y

1 (presencia de la característica de interés). Es decir que $\bar{X}$ representa la proporción de éxitos en una muestra y constituye un estadístico denominado proporción muestral, representado por *p*.

Para un tamaño de muestra grande, se puede aplicar el TCL y obtener la distribución de probabilidad del estadístico proporción muestral (*p*). En este caso, *p* sigue una distribución aproximadamente normal con media P y varianza $\frac{\boldsymbol{P(1-P)}}{\boldsymbol{n}}$; es decir,

$$p \sim N\left(\mathrm{E(p)} = P\,;\mathrm{V(p)} = \frac{P(1-P)}{n}\right)$$

Y, por lo tanto,

$$Z = \frac{p-P}{\sqrt{P\,(1-P)/\,n}} \sim \mathrm{N}(0;1)$$

Se debe tener presente que además de la condición de que *n* sea grande, es necesario que P no esté muy cercano a 0 ni a 1 para que la aproximación sea buena.

Ejemplo 2.6:

Si se sabe que el 25 % de los automóviles de una región no está en condiciones adecuadas para circular públicamente, calcular la probabilidad de que en una muestra aleatoria de 200 automóviles, la proporción muestral de automóviles en malas condiciones sea superior a 0,30.

Solución:
Se tiene:

p = la proporción de automóviles en la muestra que están en malas condiciones
n = 200
P = 0,25
Pr(p > 0,30) = ?

En virtud de que n = 200 es lo suficientemente grande, al aplicar el teorema del límite central se obtiene que:

$$p \sim N(\, P = 0{,}25\, ; V(p) = 0{,}0009375)$$

Luego,

$$Pr(\mathrm{p} > 0{,}30) = Pr\left(Z > \frac{0{,}30 - 0{,}25}{0{,}0306}\right) = Pr(Z > 1{,}63) = 0{,}0516$$

Ejercicios:

1. Dada una población de N = 110 elementos con una media igual a 24 y una varianza igual a 36, se seleccionan 3 muestras aleatorias sin reposición de tamaño n = 9, n = 36 y n = 64, respectivamente.
   - Calcule el error estándar de la media muestral para cada una de esas muestras. Compare el valor del error estándar muestral para cada una de las muestras. ¿Qué concluye?
   - Responda nuevamente la pregunta anterior, pero asumiendo que las muestras se seleccionan con reposición.
   - Sobre la base de los resultados obtenidos en (a) y (b), ¿qué conclusiones saca?

2. Considere una población de tamaño N = 980 y desviación estándar igual a 8,2. Calcule el error estándar de la media para una muestra de tamaño n = 12.
   - Sin utilizar el factor de corrección para población finita (fcf).
   - Utilizando el factor de corrección para la población finita.
   - Comparando los valores obtenidos, ¿qué concluye acerca del error estándar de la media en relación con los tamaños n y N?

3. El número de accidentes de trabajo que ocurre en un día en una industria metalmecánica es una variable aleatoria X cuya distribución de probabilidad es:

X	0	1	2	3
P(X=x)	0,5	0,3	0,15	0,05

   - Calcule el número promedio de accidentes diarios y su varianza.
   - Obtenga la distribución de probabilidad, con su correspondiente media y varianza, del número promedio de accidentes $\bar{X}$ en una muestra aleatoria de 64 días.
   - Calcule la probabilidad de que el promedio de accidentes diarios en los 64 días sea mayor que 1.

4. Se ha establecido que el consumo por persona de los clientes de un bar es una variable aleatoria que sigue una distribución normal con media igual a 1.800 unidades y desviación estándar igual a 400

unidades. Sea $\bar{X}$ el consumo promedio por persona en una muestra aleatoria de 12 clientes.

- Determine la distribución de probabilidad de $\bar{X}$.
- Esta distribución de $\bar{X}$, ¿es exacta o aproximada?
- Calcule la probabilidad de que el consumo promedio de los clientes sea superior a 2.000 unidades.

5. El tiempo promedio de duración de las llamadas telefónicas de los empleados de un hotel es de 3,5 minutos, con una desviación estándar de 1,2 minutos. Con la finalidad de establecer algunos mecanismos de control, el gerente decide medir la duración de 40 llamadas elegidas al azar.
   - Obtener la probabilidad de que la duración promedio de las llamadas de la muestra sobrepase los 3,6 minutos.
   - Si efectivamente ocurre que el promedio de duración de las llamadas es de 3,62 minutos, ¿debe preocuparse el gerente por cuanto los empleados están hablando más tiempo promedio por teléfono?

6. Sabiendo que la varianza de las estaturas (en centímetros) de los alumnos del último año de un liceo es $\sigma^2 = 70{,}56$, calcule la probabilidad de que la estatura promedio de una muestra aleatoria de 36 alumnos difiera en más de 3 centímetros de la estatura promedio de todos los estudiantes del último año.

3. Inferencia estadística

INFERENCIA ESTADÍSTICA

Una vez adquiridos los conocimientos sobre las distribuciones continuas y las distribuciones muestrales en los capítulos anteriores, se comienza con la inferencia estadística y las diferentes técnicas que abarca, como la estimación de parámetros y la contrastación de hipótesis.

3.1 Inferencia estadística

Tal como es sabido, la estadística se divide en dos ramas fundamentales: estadística descriptiva y estadística inferencial. La *estadística descriptiva* se encarga de la recolección, organización y presentación de los datos en cuadros o gráficos, así como también del cálculo de medidas numéricas que permiten destacar los aspectos más importantes de los mismos. La *estadística inferencial,* por su parte, se apoya en los resultados obtenidos de una muestra para sacar conclusiones y tomar decisiones sobre la población en estudio.

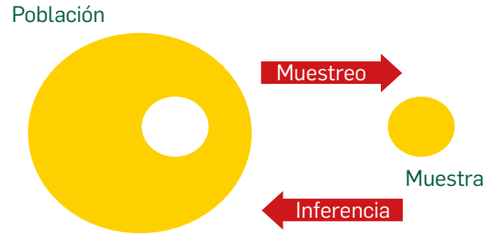


Figura 3.1 Representación gráfica de la estadística
Fuente: MLB

La estadística inferencial se divide en dos partes principales:

a. Estimación de parámetros poblacionales y

b. Contrastación de hipótesis

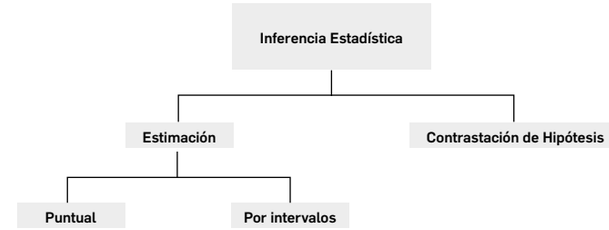


Figura 3.2 Representación gráfica de la estadística inferencial
Fuente: MLB

3.1.1 La estimación de parámetros poblacionales o teoría de estimación

Es aquella división de la inferencia estadística que proporciona las herramientas necesarias para determinar las mejores aproximaciones a aquellos valores (parámetros) desconocidos de la población. Por lo tanto, el objetivo de la estimación es obtener una aproximación al verdadero valor del parámetro poblacional.

La estimación puede ser: puntual (consiste en asignar un valor único como estimación del parámetro que esté lo más próximo posible al verdadero valor, utilizando la información proporcionada por la muestra aleatoria que se ha seleccionado de la población) o por intervalos (usa dos valores, entre los que pudiera estar el parámetro).

3.1.2 Estimación puntual

La estimación puntual de un parámetro poblacional θ es el valor numérico particular $\hat{\theta}$ de la muestra aleatoria $x_1, x_2 \ldots x_n$, que es obtenido de un estadístico (estimador) y se denota por:

$$\hat{\theta} = g(x_1, x_2, \ldots x_n)$$

La estimación puntual se utiliza principalmente cuando se quiere conocer un valor determinado de un parámetro poblacional que no se dispone. Ahora bien, para obtener ese valor único es necesario definir o construir una función que dependa solamente de los valores muestrales y que al sustituirlos en ella produzca el valor que estimará puntualmente al parámetro.

Estadístico

Un estadístico es una función matemática de los valores muestrales que no depende de ningún parámetro desconocido. Se debe tener en cuenta que todo estadístico es a su vez una variable aleatoria y, en consecuencia, posee una distribución de probabilidad específica y unos elementos que la caracterizan, tales como la media y varianza de la distribución, entre otras. Los estadísticos constituyen la base para realizar las estimaciones puntuales de los parámetros.

Estimador

Un estimador de un parámetro poblacional θ es una función $\hat{\theta}$ de las variables aleatorias $x_1, x_2 \ldots x_n$ que se aproxima a θ.

$$\hat{\theta} = g(x_1, x_2, \ldots x_n)$$

Se puede decir que el estimador es un estadístico cuyos valores son utilizados para estimar un parámetro cualquiera (ver tabla 3.1).

Tabla 3.1 Parámetros poblacionales y sus estimadores

Parámetro poblacional	Símbolo	Estimador
Media	μ	$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Varianza	σ^2	$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i\ 1}^{n}(X_i - \bar{X})^2$
Proporción	P	$p = \hat{p} = \frac{X}{n} = \frac{No.éxitos}{No.pruebas}$

Fuente: MLB

Ejemplo 3.1:

Un fabricante de bebidas refrescantes afirma que, en promedio, el volumen de llenado de los envases de las bebidas que produce es de al menos 300 ml. Para corroborarlo, mide el volumen de una muestra aleatoria y, a partir de ella, infiere el resultado.

Ejemplo 3.2:

En el proceso de aceptación de un lote de producción, una compañía recibe el lote si como máximo el porcentaje de piezas defectuosas es del 0,1 %. Para tomar la decisión de rechazar el lote se basa en una muestra aleatoria de piezas.

La calidad de la estimación obtenida depende de la adecuada elección del estimador puntual. En vista de que existe una gran variedad de estimadores posibles en cada situación particular, se deben tener algunos criterios de selección.

En este sentido, para seleccionar un buen estimador entre un conjunto de posibles estimadores, los estadísticos propuestos son estudiados teniendo en cuenta ciertas propiedades deseables.

Propiedades deseables de un estimador

Para que sea un "buen estimador" debe cumplir, por lo menos, con las siguientes cuatro propiedades:

a. Insesgado / insesgabilidad
b. Eficiente / eficiencia
c. Consistente / consistencia
d. Suficiente / suficiencia

a. Insesgado / insesgabilidad: se dice que un estimador $\hat{\theta}$es insesgado del parámetro θ si el promedio o valor esperado del mismo coincide con el valor del parámetro a estimar (Mood, Graybill y Boes, 1974).

Esto es:

$$E\left(\hat{\theta}(x)\right) = \theta$$

En la figura 3.3 se puede apreciar claramente que es un estimador insesgado de ya que su esperanza coincide con el valor del parámetro, mientras que no lo es, puesto que no cumple esta condición.

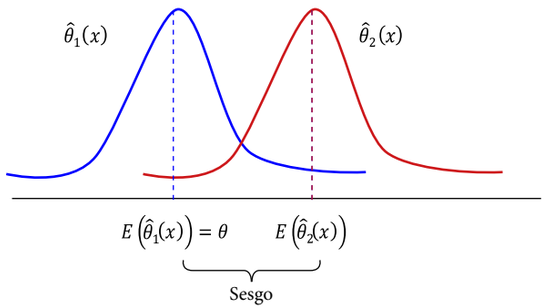


Figura 3.3 Estimador insesgado y estimador sesgado
Fuente: MLB

En otras palabras, si se tiene un gran número de muestras de tamaño n y se obtiene el valor del estimador $\hat{\theta}(x)$ en cada una de ellas, sería deseable que el promedio de estas estimaciones coincidiera con el valor de $\hat{\theta}$.

A continuación, se demuestra que la media muestral $\bar{X}$ es un estimador insesgado de la media poblacional (μ).

Si se considera a la muestra de n observaciones como una colección de n variables aleatorias, todas idénticamente distribuidas con $E(X_i) = \mu, \quad \forall i,$ entonces:

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

$$= \frac{1}{n} n\mu$$
$$E(\bar{X}) = \mu$$

b. Eficiente / eficiencia: una medida de la calidad de un estimador para θ no debe ser solo que su media sea el parámetro poblacional, sino que además haya una alta probabilidad de que los valores observados de $\hat{\theta}(x)$ sean próximos a θ (varianza lo más pequeña posible). Por lo tanto, se dice que un estimador es eficiente cuando presenta menor varianza.

En este sentido, tal como se muestra en la figura 3.4, si se tienen dos estimadores insesgados $\hat{\theta}_1(x)$ y $\hat{\theta}_2(x)$ de un parámetro θ dado, se dice que $\hat{\theta}_1(x)$ es un estimador más eficiente que $\hat{\theta}_2(x)$ si se cumple que $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.

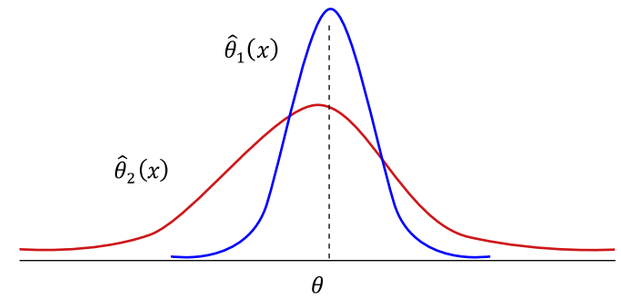


Figura 3.4 Estimador insesgado y eficiente
Fuente: MLB

c. Consistente / consistencia: esta propiedad tiene que ver con el tamaño de la muestra, ya que muchas veces el comportamiento de un estimador varía cuando se aumenta el tamaño de la muestra.

Se dice que un estimador es consistente cuando a medida que aumenta el tamaño de la muestra, el valor del estimador se aproxima al verdadero valor del parámetro que trata de estimar. En otras palabras, se dice que $\hat{\theta}$ es estimador consistente de θ si $\hat{\theta} \to \theta$ cuando $n \to \infty$.

Se debe notar que la consistencia es una propiedad que está referida a cualquier estimador, sea insesgado o no. Realmente, que un estimador sea consistente quiere decir que al aumentar el tamaño de la muestra suficientemente ($n \to \infty$), el estimador se vuelve insesgado y su varianza tiende a cero.

Por ejemplo, la media muestral ($\bar{X}$), la varianza muestral (S^2) y la proporción muestral (p)son estimadores consistentes de sus respectivos parámetros poblacionales: μ, σ^2 y P.

d. Suficiente / suficiencia: se dice que un estimador es suficiente cuando puede aportar toda la información que la muestra proporciona sobre el parámetro poblacional.

Por ejemplo, la media muestral ($\bar{X}$) es un estimador suficiente de μ por cuanto toma en cuenta cada uno de los valores muestrales en sí mismos, mientras que la mediana muestral no es un estimador suficiente, ya que solo toma en cuenta la magnitud de los valores centrales y el rango o jerarquía del resto de los valores.

Una vez obtenidas las propiedades que debe tener todo estimador, se deben tener en cuenta los siguientes comentarios al respecto:

- En la práctica, no siempre un estimador va a cumplir las propiedades vistas anteriormente de manera simultánea, ya que hay factores (costos, simplicidad, tiempo, etc.) que intervienen en el proceso de selección y que deben ser considerados; en esta situación, un estimador teóricamente mejor que otro puede ser rechazado. A veces, por ejemplo, es preferible un estimador menos eficiente en comparación con otro, pero que resulte menos costoso de obtener; en otros casos es preferible un estimador suficiente y consistente a uno que sea insesgado, y hay casos en que un estimador sesgado tiene menor varianza que otro insesgado y le resulta más conveniente al investigador.
- Los estimadores puntuales que se utilizan en este curso son insesgados y además son los mejores estimadores de sus respectivos parámetros: la media muestral es el mejor estimador de la media poblacional, la proporción muestral es el mejor estimador de la proporción poblacional y la varianza muestral (dividida entre n-1) es el mejor estimador de la varianza poblacional.

Ejemplo 3.3:

Los siguientes números representan el tiempo (minutos) que tardaron 15 operarios de una determinada empresa en realizar una tarea.

3,4	4,8	2,8	4,4	2,5	4	3,3	4,8
2,9	5,6	5,2	3,7	3	3,6	2,8	

Obtener la estimación puntual del tiempo promedio que tardaron todos los operarios de la empresa en realizar la tarea y la varianza de los tiempos.

Solución:
El mejor estimador puntual para la media poblacional μ es la media muestral $\bar{X} = \frac{\sum X_i}{n}$

El mejor estimador para la varianza poblacional σ^2 es $S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{\sum X_i^2 - n\bar{X}^2}{n-1}$

Luego:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{(3{,}4 + 4{,}8 + \cdots + 3{,}6 + 2{,}8)}{15-1} = 3{,}8\ min$$

El verdadero tiempo promedio que tardaron todos los operarios de la empresa en realizar la tarea es de 3,8 minutos.

Por otro lado,

$$S^2 = \frac{\sum X_i^2 - n\bar{X}^2}{n-1} = \frac{228{,}3 - 15(3{,}8)^2}{15-1} = 0{,}834\ min^2$$

Por lo tanto, $\bar{X} = 3{,}8\ min$ y $S^2 = 0{,}834\ min^2$son estimaciones puntuales de μ y σ^2 respectivamente.

Tal como se acaba de apreciar, cuando se usa la estimación puntual es prácticamente imposible que el valor de la estimación coincida con el verdadero valor del parámetro. Por ello, se debe buscar una manera que permita tener información sobre los parámetros que se están estimando con un cierto grado de confiabilidad; en este sentido, se hace uso de la estimación por intervalos.

3.1.3 Estimación por intervalos

La estimación de un parámetro poblacional θ mediante un intervalo consiste en obtener un par de valores sobre los que se espera que se encuentre el valor de θ con un determinado nivel de certidumbre. Este intervalo, conocido como intervalo de confianza para θ, se obtiene sobre la base de la información conocida de una muestra.

Dada una muestra aleatoria simple $(X_1; X_2; \ldots; X_n)$, proveniente de una población caracterizada por un parámetro θ desconocido y el cual se quiere estimar, un intervalo de confianza para θ estará formado por dos valores, digamos (L_i y L_s), límites inferior y superior respectivamente, que se obtienen de la muestra, asegurando de esta manera que exista una alta probabilidad (1 - α) de que la misma contenga a θ.

A esta probabilidad (1 - α) se le conoce también como nivel de certeza, certidumbre o confianza del intervalo, y es fijada previamente por el investigador.

Formalmente, se tiene que:

Dada una muestra aleatoria simple $(X_1; X_2; \ldots; X_n)$, es un intervalo de confianza para el parámetro θ de un (1 - α) si:

$$\Pr[L_i; L_s] = \Pr[L_i \leq \theta \leq L_s] = 1 - \alpha \qquad (3.1)$$

Se debe tener claro que L_i y L_s son variables aleatorias, por lo tanto, se obtendrán diferentes valores de L_i y L_s para diversas muestras. Es decir, hay probabilidad (1 - α) de que al seleccionar la muestra, esta produzca un intervalo que contenga el verdadero valor de θ.

En relación con el nivel de confianza 1 - α , es conveniente aclarar que su interpretación como probabilidad solo tiene sentido antes de obtener la muestra y de calcular los valores de L_i y L_s, ya que una vez obtenidos, no se puede hablar de la probabilidad de que un intervalo conocido (L_i; L_s) contenga al parámetro θ debido a que este no es una variable aleatoria, sino un valor constante, pero desconocido. Calculados los valores de (L_i; L_s), el parámetro estará dentro o fuera de ese intervalo, y en consecuencia la probabilidad de que L_i y L_s contenga a θ será cero o uno, pero nunca podrá ser un valor entre ellos.

Se debe observar con detenimiento que si (1 - α) representa la probabilidad de que (L_i; L_s) contenga a θ, entonces α representa la probabilidad de que el intervalo no contenga a θ. Este valor α juega un papel muy importante en la docimasia de hipótesis y será estudiado más adelante.

Para calcular un intervalo de confianza para un parámetro cualquiera se parte del estimador puntual de ese parámetro y se toma en cuenta su distribución de probabilidad y el nivel de confianza que queremos para la estimación. Es decir que los límites L_i y L_s dependen de dos factores fundamentales:

a. El estimador puntual del parámetro y su correspondiente distribución de probabilidad, y
b. El nivel de certeza 1 - α.

Al interpretar un intervalo de confianza se debe tener en cuenta que si se obtiene un número muy grande de muestras aleatorias del mismo tamaño y para cada una de ellas se calcula su correspondiente intervalo de confianza (100(1 − α) %) para θ, entonces el 100(1 − α) % de estos intervalos va a contener el verdadero valor del parámetro θ.

Para proporcionar más detalles sobre la manera de construir un intervalo de confianza, en las próximas secciones se van a tratar casos particulares.

Tal como se dijo anteriormente, el objetivo de la estimación es obtener una aproximación al valor de cierto parámetro de la población θ, bien sea de manera puntual (estimación puntual) o con un grado de incertidumbre a partir de un intervalo (estimación por intervalos de

confianza). Sin embargo, muchos problemas de las diferentes áreas del saber requieren que se tome alguna decisión entre aceptar o rechazar una proposición sobre algún parámetro θ en particular.

En esta parte del capítulo se hace referencia a la contrastación de hipótesis, y su objetivo fundamental es decidir si una afirmación acerca de una característica de la población es o no verdadera.

Por ejemplo, suponga que se tiene interés en la resistencia a la ruptura de la fibra textil usada en la fabricación de material para cortinas. Entonces la resistencia a la ruptura es una variable aleatoria que puede describirse con una distribución de probabilidad. Considere que el interés se enfoca en la media de la resistencia a la ruptura; específicamente quiere decidirse si la media de la resistencia a la ruptura de la fibra textil es de 100 psi o no. Si bien se hace referencia al tema en la sección 3.2.2., formalmente la conjetura anterior se puede expresar como:

$$H_0: \mu = 100\ psi$$
$$H_0: \mu \neq 100\ psi$$

3.1.4 Contrastación de hipótesis

La contrastación de hipótesis, también conocida como docimasia de hipótesis, prueba de hipótesis, test de hipótesis, entre otras, es un procedimiento estadístico basado en la evidencia de la muestra y la teoría de la probabilidad, que permite determinar si una conjetura o afirmación es cierta o no.

A continuación, se van a dar algunos conceptos básicos de un contraste de hipótesis.

3.1.4.1 Hipótesis estadística

Es cualquier enunciado, afirmación o conjetura sobre una o más características desconocidas de la población. Estas pueden ser un parámetro, la distribución de probabilidad de una población, etc.

Debe tenerse claro que de ningún modo se sabe con absoluta certeza la verdad o falsedad de una hipótesis estadística, a menos que se examine toda la población. Esto, por supuesto, no sería práctico en la mayoría de las situaciones. Por lo anterior se toma una muestra aleatoria de la población bajo estudio y se usa la información disponible en la misma para proporcionar evidencias que confirmen o no la hipótesis en cuestión.

Cuando la evidencia de la muestra no es consistente con la hipótesis planteada, entonces se produce un rechazo de la misma; en caso contrario, se conduce a su aceptación. Así, el objetivo primordial de la prueba de hipótesis es determinar si la diferencia entre un valor propuesto de un parámetro poblacional y el valor del estadístico de la muestra se debe o no a la variabilidad del muestreo.

La hipótesis estadística puede ser simple (exacta) o compuesta (inexacta). Una hipótesis simple es aquella en la cual se especifica un solo valor para el parámetro. Por ejemplo, siendo θ un parámetro cualquiera, la hipótesis $\theta = 15$ es una hipótesis simple, de manera general $H_0: \theta = \theta_0$. Una hipótesis compuesta es aquella en la que se especifica un conjunto de valores o un rango de valores para un parámetro poblacional desconocido, por ejemplo $\mu \leq 15$, de manera general, $H_0: \theta \leq \theta_0$.

En toda prueba de hipótesis siempre se va a encontrar dos hipótesis que se enfrentan: la hipótesis nula y la hipótesis alternativa.

3.1.4.2 Hipótesis nula

Se denota por H_0. Es la hipótesis que se plantea para considerar si puede ser o no rechazada; es decir, es aquella que se formula con la intención de determinar si puede ser rechazada.

Hay que subrayar que si la hipótesis nula no se rechaza con base en los datos muestrales, no es posible afirmar que sea verdadera. Para probar sin duda alguna que la hipótesis nula es verdadera tendría que conocerse el parámetro poblacional y, por lo general, no es posible.

3.1.4.3 Hipótesis alternativa

Se denota por H_1 y se le conoce como hipótesis de investigación. Representa la hipótesis de interés para el investigador debido a que es, generalmente, la proposición que él desea probar. En otras palabras, es aquella que se enfrenta o se opone a la hipótesis nula. Aquí se expresa lo que se presume que está sucediendo (actualmente) y que ha cambiado con respecto a lo que se suponía como verdadero (anteriormente).

En el contexto del contraste de hipótesis clásico, la hipótesis nula se considera cierta inicialmente. La tarea de persuadir de lo contrario corresponde a los datos de la muestra. La "aceptación" de una hipótesis nula implica tan solo que los datos de la muestra no proporcionan evidencia suficiente para rechazarla. Por otro lado, el rechazo implica que la evidencia muestral la refuta.

La selección de hipótesis nula y alternativa no es arbitraria; por el contrario, resulta crucial que sean elegidas adecuadamente. Ahora bien, ¿cómo elegir una y otra? El criterio más conveniente es adoptar como hipótesis nula a aquello que se tiene como cierto o que se ha venido aceptando hasta el momento, y tomar como hipótesis alternativa a lo que se plantea como nuevo, aquello que se enfrenta a lo establecido, es decir, lo que se quiere investigar y proponer como un cambio y además se espera que la muestra proporcione información que lo respalde. Se podría decir que la hipótesis alternativa constituye la hipótesis de trabajo o de investigación.

Tal como se muestra en la figura 3.5, en la hipótesis nula se encuentra la posibilidad de que el parámetro poblacional θ sea igual a un valor específico, θ_0; esto se debe a que siempre la hipótesis nula se considera inicialmente cierta.

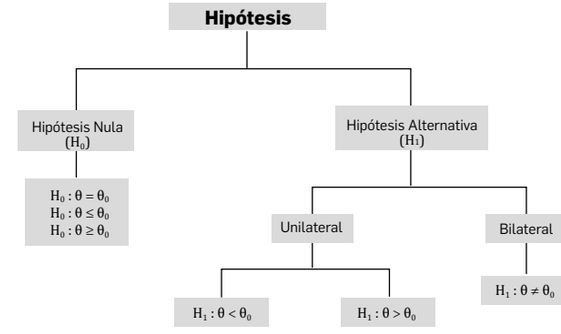


Figura 3.5 Tipos de hipótesis
Fuente: MLB

3.1.4.4 Tipos de contraste

La forma de la hipótesis alternativa es determinante para conducir la prueba de hipótesis de una u otra manera. En este sentido, los contrastes pueden ser unilaterales o bilaterales.

El contraste es unilateral o de una cola, bien sea por la derecha o por la izquierda, cuando en la hipótesis alternativa (H_1) se establece que el parámetro es mayor o menor que el valor especificado en la hipótesis nula (H_0).

El contraste es bilateral o de dos colas cuando en la hipótesis alternativa (H_1) se especifica que el parámetro es diferente del valor indicado en la hipótesis nula (H_0).

A continuación, se muestran algunos casos que pueden presentarse:

a. $H_0: \theta = \theta_0$ *vs.* $H_1: \theta > \theta_0$ → Contraste unilateral cola derecha
b. $H_0: \theta = \theta_0$ *vs.* $H_1: \theta < \theta_0$ → Contraste unilateral cola derecha
c. $H_0: \theta = \theta_0$ *vs.* $H_1: \theta \neq \theta_0$ → Contraste bilateral

Al finalizar un procedimiento de contrastación de hipótesis, hay que inclinarse por una de las dos hipótesis planteadas. Rechazar la hipótesis

nula (H_0) implica automáticamente aceptar la alternativa (H_1); esto quiere decir que la muestra ha proporcionado suficiente evidencia para rechazar lo que se plantea en H_0 y para respaldar lo expresado en H_1.

Por otro lado, aceptar (no rechazar) la hipótesis nula (H_0) no implica que se rechaza lo establecido en la alternativa (H_1), sino que la evidencia proporcionada por la muestra no es lo suficientemente fuerte como para rechazar la nula y aceptar la alternativa. Por ello, en un contraste de hipótesis se habla de rechazar H_0 (y aceptar H_1) o de no rechazar H_0, pero no se debe aceptar H_0.

Es importante dejar claro que con referencia a la contrastación de hipótesis, no es lo mismo no rechazar H_0 que aceptar H_0. Cuando se hace un contraste de hipótesis, las hipótesis nula y alternativa se especifican con la intención de que la información de la muestra sea concluyente en contra de H_0 y apuntale vigorosamente lo establecido en H_1. Cuando la muestra da evidencias concluyentes en contra de lo que dice H_0, no se tiene duda de rechazar H_0 y aceptar H_1. Ahora bien, si la evidencia de la muestra no es suficiente o convincente en contra de H_0, entonces no es que se acepta H_0, sino que simplemente *no hay evidencia suficiente en contra* y, por lo tanto, no se rechaza. Si se acepta H_0, se dice que estamos convencidos de que H_0 es verdad, y en realidad no se trata de eso en un contraste de hipótesis; lo que persigue un contraste de hipótesis *son evidencias razonables en contra de H_0,* y si eso no se logra con la muestra, quiere decir que o bien esas evidencias realmente no existen o, en caso de existir, no se encuentran reflejadas en la muestra que se eligió. La similitud entre un juicio que se le sigue a una persona y un contraste de hipótesis permite aclarar esta situación.

Imagine que una persona es acusada de asesinato y se le lleva a juicio. Se sabe que *toda persona se presume inocente hasta tanto se demuestre que es culpable*. Las dos hipótesis a manejar en el juicio, guardando la similitud con el contraste de hipótesis, serían:

H_0: La persona es inocente (no culpable)
H_1: La persona es culpable

Luego, sobre la base de las evidencias presentadas por los organismos indicados para el caso, la policía y el acusador privado, el juez debe tomar la decisión de declarar a la persona culpable o inocente (no culpable). En el contraste de hipótesis las evidencias están representadas por la muestra y sobre ellas se decide rechazar o no a la hipótesis nula. Si las evidencias que se presentan en contra de la persona son definitivas (por ejemplo, se dispone del cadáver, del arma disparada con las huellas digitales, de testigos que declaran haber presenciado el hecho, de las pruebas técnicas de la policía que señalan al acusado, etc.), entonces el juez rechaza la inocencia de la persona y la condena como culpable del asesinato (rechaza H_0 y acepta H_1). Por el contario, si las pruebas presentadas no son suficientes en contra de la persona (por ejemplo, no se presentan los testigos, no se encuentra el arma del homicida, hay duda con las pruebas técnicas, etc.), entonces el juez no puede declarar culpable al acusado y se le sigue considerando inocente del crimen (se acepta H_0). Sin embargo, hay que puntualizar que aun cuando en el lenguaje cotidiano se dice que la persona es inocente, en realidad el juez no declara su inocencia, sino que dice que, sobre la base de las pruebas presentadas, él no puede condenarla. Afirmar que alguien es inocente de un asesinato es estar seguro de que esa persona no lo hizo. Es posible que el acusado haya cometido el delito, pero si las pruebas presentadas en el juicio no son lo suficientemente contundentes, no se le puede declarar culpable. Se dice que la persona es no culpable (inocente) de lo que se acusa. Este es exactamente el principio que se aplica en un contraste de hipótesis: se rechaza o no se rechaza H_0.

3.1.4.5 Tipos de error

Al tomar una decisión en un contraste de hipótesis se puede presentar las siguientes posibilidades:

Si H_0 es cierto, se puede aceptar o rechazar
Si H_0 es falso, se puede aceptar o rechazar

De estas dos situaciones surgen las cuatro decisiones posibles en un contraste de hipótesis, las cuales se presentan en la tabla 3.2, donde se puede apreciar que dos decisiones son correctas y dos son incorrectas.

Tabla 3.2 Posibles resultados de un contraste de hipótesis

Decisión	Situación de H_0	
	H_0 es cierta	**H_0 es falsa**
Aceptar H_0	Decisión correcta	Decisión incorrecta (error tipo II)
Rechazar H_0	Decisión incorrecta (error tipo I)	Decisión correcta

Fuente: MLB

Dado que se trabaja con una muestra y nunca se sabe si es cierta o falsa, al tomar una decisión existe la posibilidad de que esta sea incorrecta. Es necesario distinguir dos tipos de errores posibles:

a. Error tipo I, que consiste en rechazar H_0, dado que es cierta
b. Error tipo II, que consiste en aceptar H_0, cuando es falsa

Se denota por α y se llama nivel de significación a la probabilidad de cometer el error tipo I, y por β a la probabilidad de cometer el error tipo II. Observe que ambas probabilidades son condicionales:

$$\alpha = Pr(\text{error tipo I}) = Pr(\text{rechazar } H_0 \,/\, H_0 \text{ es cierta})$$
$$\beta = Pr(\text{error tipo II}) = Pr(\text{aceptar } H_0 \,/\, H_0 \text{ es falsa})$$

Observe que $(1 - \alpha)$ es la probabilidad de tomar la decisión correcta, de aceptar H_0 cuando es cierta, y $(1 - \beta)$ la probabilidad de decidir correctamente rechazar H_0 cuando es falsa. A la probabilidad $(1 - \beta)$ se le denomina potencia del contraste y representa el poder o la fortaleza que tiene el contraste para reconocer correctamente que H_0 es falsa cuando en verdad H_1 es cierta. Lo ideal en cualquier contraste de hipótesis es que tenga una potencia grande (próxima a uno), o equivalentemente, que β sea muy pequeño cuando H_0 es falsa.

Es de hacer notar que en un contraste de hipótesis, los valores de β y de $(1 - \beta)$ solo pueden calcularse para un valor determinado del parámetro dentro del rango de valores especificados para este en la hipótesis alternativa. Por ejemplo, en el contraste H_0:μ < 15, se puede calcular la probabilidad β asumiendo que H_1 es cierta y que en verdad μ = 16.

$$\beta = Pr\,(\text{aceptar } H_0 / \mu = 16)$$

También puede calcularse β para otros valores de μ bajo H_1. En el caso de μ = 18, $\beta = Pr(\text{aceptar } H_0 / \mu = 18)$.

Si se calcula la probabilidad $Pr(\text{rechazar } H_0/\mu)$ para todos los posibles valores del parámetro μ especificados, tanto en H_0 como en H_1, y se grafican esos valores contra los correspondientes valores de μ, se obtiene una curva continua denominada curva de potencia, tal como se muestra en la figura 3.6.

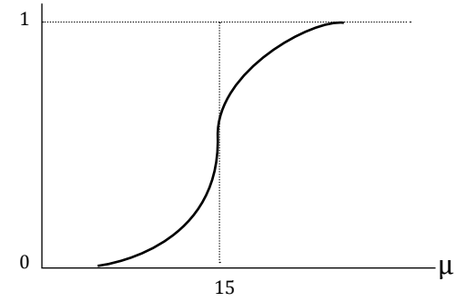


Figura 3.6 Curva de potencia para un contraste sobre μ

Fuente: MLB

Para los valores de μ en los cuales H_0 es cierta, la ordenada de la curva es igual a la probabilidad de α de cometer error tipo I, mientras que para aquellos valores de μ donde H_1 es cierta, la ordenada es igual a $1 - \beta$.

Dado que α y β representan las probabilidades de incurrir en decisiones incorrectas, lo deseable es que cada una de ellas sea lo más pequeña posible; sin embargo, α y β no son independientes entre sí y tampoco lo son del tamaño de muestra n. Se ha demostrado que para un tamaño fijo de muestra, al disminuir el valor de α aumenta el valor de β y viceversa. También se cumple que al aumentar el tamaño de muestra n es posible reducir los valores de α y β. Aunque lo indicado parece ser incrementar n para reducir α y β no olvidemos que esto se traduciría en mayores costos, que el investigador muchas veces no puede enfrentar. Lo usual en un contraste de hipótesis es que se busque un equilibrio entre los valores de α y β y el tamaño de muestra n. Asumiendo que el tamaño de muestra es dado y considerando que incurrir en el error tipo I es más perjudicial que incurrir en el error tipo II, se acostumbra fijar o contro-

lar α bajo un cierto nivel razonable y luego se usa aquel contraste que producen los valores de β más pequeños.

Fijar el nivel de α es decir, la probabilidad de rechazar H_o cuando es cierta, es a criterio del investigador; para ello se toma en cuenta la magnitud o la fuerza que se desea que tenga la evidencia en contra de H_o para poder rechazarla. Si se quiere rechazar H_o solamente cuando la evidencia en contra de ella sea muy fuerte, entonces quiere decir que se está dispuesto a correr un riesgo muy pequeño de rechazar H_o cuando sea cierta y, en consecuencia, el valor de la probabilidad α debe fijarse muy próxima a cero. Si, por el contrario, se está dispuesto a rechazar H_o ante la más pequeña evidencia en contra de ella, entonces quiere decir que se desea correr un riesgo grande de cometer el error de rechazar H_o aun siendo cierta, y por lo tanto α va a ser relativamente grande. Sin embargo, no hay que olvidar que α es la probabilidad de tomar una decisión incorrecta y se debe tratar de que sea razonablemente pequeña.

Los valores más utilizados para α son 0,01, 0,05 y 0,10. Un valor de 0,01 significa que se está dispuesto a correr el riesgo de rechazar (equivocadamente) H_o cuando sea cierta en 1 de cada 100 contrastes independientes que se realicen.

De manera general, con un valor de *p* menor de:

a. 0,10, se tiene regular evidencia de que H_0 no es verdadera
b. 0,05, se tiene fuerte evidencia de que H_0 no es verdadera
c. 0,01, se tiene muy fuerte evidencia de que H_0 no es verdadera
d. 0,001, se tiene evidencia extremadamente fuerte de que H_0 no es verdadera

3.1.4.7 Selección del estadístico de contraste

Una vez establecidas las hipótesis nulas y alternativas, y fijado el nivel de significación α, es necesario seleccionar un estadístico sobre el cual se va a tomar la decisión de aceptar o rechazar la hipótesis nula. La decisión sobre el parámetro poblacional se basa en la información que proporcione la muestra. El estadístico a seleccionar se denomina estadístico del contraste y debe tener relación estrecha y lógica con el parámetro bajo estudio.

Por lo tanto, el estadístico de contraste es aquella función de las observaciones muestrales que se usa para determinar si la hipótesis nula debe ser rechazada o no.

3.1.4.7 Región de rechazo y región de aceptación

El conjunto de valores del estadístico del contraste que sugiere el rechazo de H_0 constituye la denominada región de rechazo o región crítica, mientras que el conjunto de valores del estadístico que lleva a aceptar H_0 constituye la región de aceptación del contraste. El valor o los valores que separan la región crítica de la región de aceptación se denominan valores críticos, y estos siempre forman parte de la región crítica. Dependiendo de si la hipótesis alternativa es unilateral o bilateral, se tendrán diferentes regiones y valores críticos. Gráficamente, la figura 3.7 muestra las posibles zonas de aceptación y de rechazo de la hipótesis nula para los diferentes tipos de contrastes.

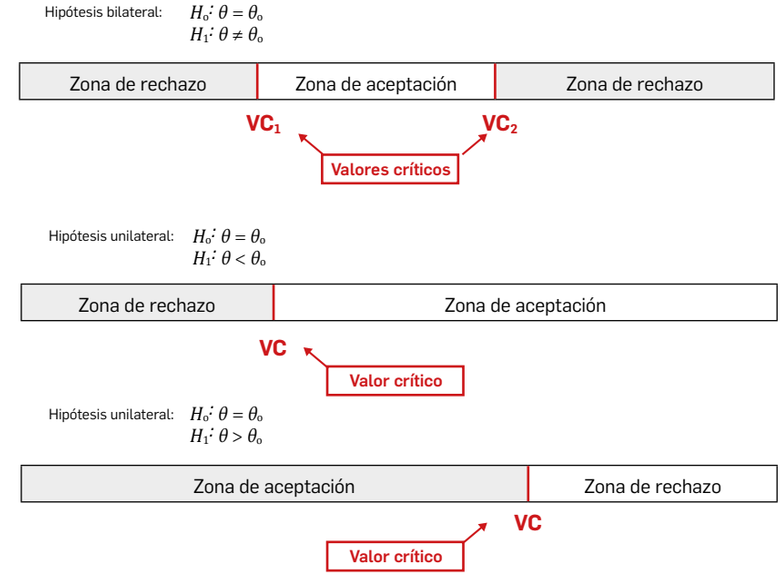


Figura 3.7 Regiones de aceptación y de rechazo
Fuente: MLB

3.1.4.8 Pasos a seguir en la contrastación de hipótesis

Al realizar una contrastación de hipótesis se deben seguir una serie de pasos que conlleva la toma de decisión sobre una hipótesis en particular. Los pasos a seguir son los siguientes:

a. Identificar el parámetro o los parámetros poblacionales sobres los cuales se va a realizar el contraste.
b. Establecer las hipótesis nula H_0 y alternativa H_1 : se precisa el tipo de hipótesis a contrastar y si el contraste es unilateral o bilateral.
c. Fijar el nivel de significación α.
d. Especificar las condiciones y establecer los supuestos necesarios bajo los cuales se realiza el contraste.
e. Seleccionar el estadístico del contraste y determinar su distribución de probabilidad bajo H_0.
f. Determinar la región de rechazo y la región de aceptación del contraste. Lo anterior es equivalente a determinar los valores críticos del contraste, los cuales, como ya se dijo, dependen de tres factores: la distribución de probabilidad del estadístico, el nivel de significación α y el tipo de contraste (unilateral o bilateral).
g. Establecer la regla de decisión del contraste.
h. Calcular el valor estadístico del contraste sobre la base de los datos muestrales, asumiendo que H_0 es cierta.
i. Tomar la decisión estadística acerca del contraste mediante la aplicación de la regla de decisión.
j. Expresar la decisión estadística en términos no estadísticos.

Es muy importante traducir la decisión de tipo estadístico que se haya tomado en términos del problema analizado, de tal manera que sea comprensible para las personas desconocedoras del lenguaje estadístico.

Ejercicios

1. Si X es una variable aleatoria binomial[8], muestre que:

   a. $\hat{B} = \frac{X}{n}$ es un estimado insesgado de b

   b. $B' = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$ es un estimador sesgado de p

2. Si X es una variable aleatoria binomial, muestre que:

   a. $\hat{P} = X/n$ es un estimador de p

   b. $P' = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$ es un estimador sesgado de p

3. Muestra que el estimador P' del ejercicio 2(b) se vuelve insesgado conforme $n \to \infty$.

[8] Una variable aleatoria se distribuye binomial si $\Pr(X = x) = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k \in \{0; 1; 2; \ldots; n\}$.

4. Inferencia acerca de la medida poblacional

INFERENCIA ACERCA DE LA MEDIA POBLACIONAL

En el capítulo anterior se abordaron las dos maneras de realizar inferencia estadística, la estimación y el contraste de hipótesis, concernientes a los parámetros de una población. En este capítulo se considera la manera de hacer inferencia sobre una sola media poblacional y diferencia de medias de dos poblaciones independientes y dependientes, así como también la realización de contraste de hipótesis acerca de las mismas.

4.1 Inferencia para una sola media poblacional (μ)

Para hacer inferencia sobre una sola media poblacional se pueden tener diferentes escenarios dependiendo de las circunstancias que se presenten, tales como:

a. La población sea normal o no
b. La desviación estándar poblacional σ sea o no sea conocida y
c. El tamaño *n* de la muestra sea grande o no

En este sentido, se realizará la inferencia estadística (estimación por intervalos y contraste de hipótesis) para las diferentes combinaciones de los puntos (a), (b) y (c) nombrados anteriormente.

Se debe tener en cuenta la distribución (normal o *t-student*) que sigue la media muestral dependiendo del escenario presentado.

Escenario 1:

Cuando la población se distribuye normal con media μ, la desviación estándar poblacional σ es conocida y el tamaño de la muestra n es grande.

En primer lugar, se debe resaltar que el mejor estimador puntual para la media poblacional es la media muestral $\bar{X}$

Tal como se dijo en el capítulo dos, la media muestral, $\bar{X}$ sigue una distribución normal o aproximadamente normal con parámetros media (μ) y desviación estándar (σ), es decir:

$$\bar{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \quad (4.1)$$

En este caso, se sabe que:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1) \quad (4.2)$$

1. Estimación por intervalos

Recordando que la distribución normal es simétrica con respecto a la media(μ), y fijando el nivel de confianza en (1 - α) %, se pueden encontrar dos valores, $-k$ y k, que satisfagan la relación:

$$P\left(-k \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq k\right) = (1-\alpha)\% \quad (4.3)$$

Ahora bien, dado que el cociente representado por Z sigue una distribución normal estándar y que entre $-k$ y k hay un área de probabilidad igual a (1 - α), entonces k es un valor de Z tal que a su derecha hay un área igual a α / 2 y $-k$, es un valor de Z tal que a su izquierda hay un área del mismo tamaño.

En la figura 4.1 se aprecia que si se denota a k como $Z_{\alpha/2}$ y a $-k$ como $-Z_{\alpha/2}$, se trata del mismo valor de Z, solo que de un lado tiene un valor positivo y del otro lado es negativo.

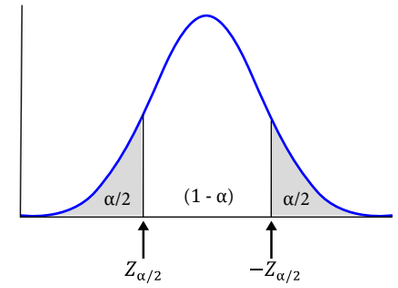


Figura 4.1 Valores de $Z_{\alpha/2}$ y $-Z_{\alpha/2}$ en la distribución N(0;1)
Fuente: MLB

En este caso, la expresión probabilística dada en la ecuación (4.3) quedaría expresada en términos de Z como:

$$P\left(-Z\frac{\alpha}{2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z\alpha/2\right) = 1 - \alpha \quad (4.4)$$

Al realizar operaciones algebraicas dentro del paréntesis se puede llegar a la expresión (4.5.):

$$P\left(-Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) \quad (4.5)$$

Ahora se puede identificar:

$$L_i = \bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \text{ y } L_s = \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad (4.6)$$

Cuando ya se ha tomado la muestra y se ha definido el nivel de significación α, dado que σ es conocida, se puede calcular el intervalo $(L_i; L_s)$, el cual constituye un intervalo de (1 - α) 100 % de confianza para la media poblacional μ.

Resumiendo, se tiene que:

Dada una población normal N(μ; σ) o una población no normal, pero fijando un tamaño de muestra n grande y siendo σ conocida, un intervalo de confianza de (1 - α) 100 % para μ viene dada por:

$$\bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \qquad (4.7)$$

siendo $Z_{\alpha/2}$ el valor correspondiente de normal estandarizada tal que a su derecha hay un área bajo la curva igual α/2.

Ejemplo 4.1:

Se sabe que el peso promedio de los paquetes enviados a través de una empresa de encomiendas sigue una distribución normal con una desviación estándar de 2 libras. Si el peso promedio de 30 paquetes enviados por medio de esta empresa fue de 51,3 libras, estime el peso medio poblacional mediante un intervalo de confianza del 95 %.

Solución:
Datos:

- $\bar{X} = 51{,}3$
- $\sigma = 2$
- $n = 30$

Además, dado que 1 - α = 0,95, entonces α = 1 − 0,95 = 0,05.

$$\frac{\alpha}{2} = 0{,}025 \quad \text{y} \quad Z_{0{,}025} = 1{,}96^{9}$$

Y usando la expresión (4.7), un intervalo de confianza del 95 % para μ viene dado por:

$$51{,}3 - 1{,}96\left(\frac{2}{\sqrt{30}}\right) \leq \mu \leq 51{,}3 + 1{,}96\left(\frac{2}{\sqrt{30}}\right)$$

[9] Ver tabla distribución normal anexo A.

$$51{,}3 - 1{,}96\,(0{,}365) \leq \mu \leq 51{,}3 + 1{,}96\,(0{,}365)$$

$$51{,}3 - 0{,}716 \leq \mu \leq 51{,}3 + 0{,}716$$

$$50{,}584 \leq \mu \leq 52{,}016$$

Luego, la estimación por intervalo para μ con un nivel de confianza del 95 % viene dado por (50,584; 52,016). En este sentido, se tiene una confianza del 95 % en que el verdadero peso medio poblacional estará contenido en el intervalo (50,584; 52,016) libras.

Con estos resultados se puede puntualizar lo siguiente en relación con los intervalos de confianza:

(i) En primer lugar, es necesario preguntarse: ¿Qué significa tener una confianza del 95 % de que el intervalo (50,584; 52,016) contiene a μ?

Se sabe que al tomar diferentes valores para la media muestral ($\bar{X}$), se producirán diferentes intervalos de confianza para la media poblacional μ, y, como este parámetro es fijo, entonces algunos intervalos contendrán a μ y otros no, tal como se muestra en la figura 4.2.

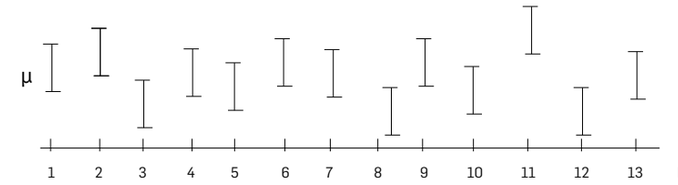


Figura 4.2 Intervalos de confianza para diferentes muestras del mismo tamaño
Fuente: MLB

Por lo tanto, si se repite muchas veces el experimento de seleccionar muestras de tamaño 30 (n = 30) y de calcular valores para intervalos de confianza del 95 % para μ, entonces se espera que el 95 % de esos intervalos incluyan a μ y el 5 % de ellos no la incluyan. Más sencillamente, es como si se tuviera que seleccionar aleatoriamente

un intervalo de un recipiente donde hay 100 intervalos y se sabe que 95 de ellos contienen a μ y 5 no la contienen. Ahora bien, el intervalo favorecido es el (50,584; 52,016). Este intervalo puede ser de los que contienen o no a μ, pero eso no lo sabremos, ya que μ es desconocida. Lo único que podemos afirmar es que tenemos un 95 % de seguridad en que el intervalo (50,584; 52,016) contiene a μ.

Hay que recordar que cuando se hace inferencia, se selecciona y trabaja con una sola muestra, y que cuando se habla de seleccionar varias muestras y construir intervalos de confianza, solo se hace hipotéticamente y con la finalidad de entender el significado de la seguridad o confianza que se tiene en la estimación por intervalo.

(ii) El nivel de confianza (1 – α) es una probabilidad que solo tiene sentido interpretar como tal antes de seleccionar la muestra y calcular el intervalo de confianza correspondiente.

Es correcto escribir:

$$Pr\left(\bar{X} - 1{,}96\left(\frac{\sigma}{\sqrt{30}}\right) \leq \mu \leq \bar{X} + 1{,}96\left(\frac{\sigma}{\sqrt{30}}\right)\right) = 0{,}95$$

Por cuanto los límites del intervalo dependen de los valores muestrales y, por lo tanto, son aleatorios.

Una vez seleccionada la muestra y obtenido el intervalo (50,584; 52,016), es incorrecto escribir:

$$Pr\,(50{,}584 \leq \mu \leq 52{,}016) = 0{,}95$$

ya que los límites son valores fijos y μ es también una constante, aunque desconocida. En consecuencia, aún sin saber el valor de μ y dado el intervalo (50,584; 52,016), solo caben dos posibilidades: (a) que μ esté dentro del intervalo, en cuyo caso la probabilidad anterior es igual a 1, (b) que μ esté afuera del intervalo y, en consecuencia, la probabilidad es igual a 0. Es decir:

$$Pr(50{,}584 \leq \mu \leq 52{,}016) = \left\{\begin{matrix} 1\ si\ \ \mu\ está\ en\ (50{,}584 \leq \mu \leq 52{,}016) \\ 0\ si\ \mu\ no\ está\ en\ (\ 50{,}584 \leq \mu \leq 52{,}016\ ) \end{matrix}\right\}$$

(iii) Existe una relación directa entre la longitud de un intervalo de confianza y su nivel de seguridad. Al aumentar el nivel de confianza (1 – α) también aumenta la longitud del intervalo. En consecuencia, siempre se puede tener un intervalo con un nivel de confianza lo más grande posible, pero va a tener una longitud tan extensa que lo hará poco útil a los fines de estimación.

El intervalo $(50{,}584 \leq \mu \leq 52{,}016)$ tiene una longitud de 1,426 y una confianza del 95 %. Si en este ejemplo aumentamos el nivel de confianza al 99 %, se obtendría el intervalo de confianza $(50{,}451 \leq \mu \leq 52{,}149)$, el cual tiene una longitud de 1,698.

Es ideal encontrar intervalos cortos con un nivel de confianza alto. La manera de lograrlo es aumentar el tamaño de la muestra, ya que esto disminuiría la longitud del intervalo sin reducir el nivel de confianza.

(iv) El intervalo de confianza $\left(\bar{X} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right); \bar{X} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$ está centrado en $\bar{X}$ y $\bar{X}$ es el mejor estimador puntual de μ; sin embargo, esto ocurre así porque la distribución de $\bar{X}$ es normal, la cual es simétrica. Cuando se estima un parámetro y la distribución del estimador puntual no es simétrica, se encuentra que el correspondiente intervalo de confianza no está centrado en el estimador. Lo importante es que cuando se hace estimación por intervalo, no importa el punto medio del mismo ni ningún otro punto intermedio, sino todo el intervalo.

2. Contraste de hipótesis

Para contrastar hipótesis cuando la población es normal o aproximadamente normal, la desviación estándar poblacional σ es conocida y el tamaño de la muestra es grande (n > 30), se utiliza la distribución normal para probar hipótesis respecto a la media de una población utilizando el siguiente procedimiento:

a. Plantear las hipótesis a contrastar:

$$\bullet\ H_0: \mu = \mu_0 \quad H_1: \mu < \mu_0 \qquad \bullet\ H_0: \mu = \mu_0 \quad H_1: \mu > \mu_0 \qquad \bullet\ H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0$$

b. Fijar el nivel de significación: μ

c. Establecer el estadístico a utilizar:
En virtud de que la población se distribuye normal, la varianza es conocida, el tamaño de la muestra es grande y se quiere hacer inferencia para una sola media, se tiene que el estadístico apropiado es:

$$Z_c = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1) \qquad (4.8)$$

d. Determinar el valor y la región críticos:
Cuando la hipótesis alternativa es unilateral, la región crítica está localizada en la dirección correspondiente al signo de desigualdad. Esto es, para hipótesis unilateral izquierda, la región crítica se sitúa en la cola inferior de la distribución; si la hipótesis es unilateral derecha, se ubica en la parte superior de la distribución; mientras que si es bilateral, las mismas se encontrarán en ambos lados de la gráfica, tal como se muestra en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución normal.

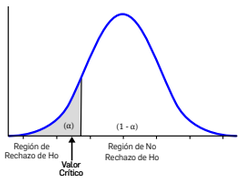


Figura 4.3. a $H_1: \mu < \mu_0$
Unilateral izquierda

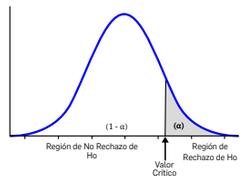


Figura 4.3. b $H_1: \mu > \mu_0$
Unilateral derecha

Figura 4.3. c $H_1: \mu \neq \mu_0$
Bilateral

Figura 4.3 Regiones y valores críticos para el contraste de hipótesis
Fuente: MLB

e. Calcular el estadístico:
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$Z_c = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

f. Decidir e interpretar:
Al comparar el valor del estadístico debe cerciorarse si cae en la región de rechazo de la hipótesis nula. En caso contrario, se expresa que no se rechaza la hipótesis nula o que el contraste no es estadísticamente significativo.

Ejemplo 4.2:

Con respecto al ejemplo 4.1, suponga que las especificaciones de las empresas de encomiendas requieren que el peso promedio de los paquetes sea de 50 libras. Además el analista decide especificar un nivel de significación de 0,05. ¿A qué conclusiones debe llegarse?

Solución:

a. Plantear las hipótesis a contrastar
$H_0: \mu = 50\ libras$
$H_1: \mu \neq 50\ libras$

b. Fijar el nivel de significación: = α = 0,05

c. Establecer el estadístico a utilizar
Debido a que la población se distribuye normal, la varianza es conocida, el tamaño de la muestra es grande y se quiere hacer inferencia para una sola media, se tiene que el estadístico apropiado es:

$$Z_c = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos

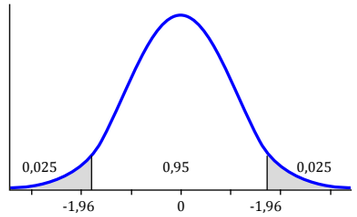


Figura 4.4 Valor y región crítica
Fuente: MLB

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$Z_c = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{51{,}3 - 50}{\frac{2}{\sqrt{30}}} = 3{,}56$$

f. Decidir e interpretar

Como 3,56 es mayor que 1,96, que es el valor crítico, entonces se rechaza la hipótesis nula, es decir, que el peso promedio de los paquetes difiere de 50 libras.

Escenario 2:

Cuando la población se distribuye normal con media μ, la desviación estándar poblacional σ es desconocida y el tamaño de la muestra n es grande.

1. Estimación por intervalos

Considere una situación más real en la cual se tiene una población con una media μ y desviación estándar poblacional σ desconocida. Suponga además que el tamaño de la muestra n es grande.

Bajo estas condiciones, el intervalo de confianza obtenido para μ en el capítulo 3 es válido, solo que no es posible calcularlo por cuanto el valor de σ es desconocido. Sin embargo, dado que n es grande, se puede sustituir σ por su estimador S sin afectar notablemente el proceso y obtener una estimación por intervalo para μ en el caso de una muestra grande.

Cuando el tamaño de una muestra es grande y la desviación estándar poblacional σ es desconocida, un intervalo de confianza $(1 - \alpha)$ 100 % para μ viene dado por:

$$\bar{X} - Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \quad (4.9)$$

Donde S^2 es la varianza de la muestra:

$$S^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \quad (4.10)$$

y S la desviación estándar muestral, dada por la raíz de la varianza:

$$S = \sqrt[2]{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \quad (4.11)$$

En este caso no importa el tipo de distribución de probabilidad que tenga la población, por cuanto n es grande.

Ejemplo 4.3:

Cierta compañía automovilística hace una prueba entre cien propietarios de automóviles comprados en su concesionario. En un municipio en particular, se tiene que un carro se maneja en promedio 23.500 km/año, con una desviación estándar de 3.900 km/año. Construir un intervalo de confianza del 90 % para el verdadero número de kilómetros promedio que se maneja un automóvil.

Solución:
Si μ es el número de kilómetros promedio que se maneja un automóvil y dado que n es grande, podemos aproximar σ^2 por S^2, y aplicando el teorema del límite central se tiene que $\bar{X}$ sigue una distribución aproximadamente normal con media μ y varianza S^2/n. Luego, el intervalo de confianza para μ es:

$$\left(\bar{X} - Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) ; \bar{X} + Z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)\right) \quad (4.12)$$

Debido a que el nivel de significación es $(1 - \alpha) = 0{,}90$, entonces $\alpha/2 = 0{,}05$ y $Z_{0{,}05} = 1{,}64$. Siendo S = 3.900.

Sustituyendo en el intervalo:

$$\left(23.500 - 1{,}64\left(\frac{3.900}{\sqrt{100}}\right); 23.500 + 1{,}64\left(\frac{3.900}{\sqrt{100}}\right)\right)$$

$$(22.860{,}4;\ 24.139{,}6)$$

Hay una confianza del 90 % de que el verdadero número de kilómetros promedio que se maneja un automóvil en cierta ciudad está entre 22.860,4 y 24.139,6 km.

2. Contraste de hipótesis

Para contrastar hipótesis cuando la población es normal o aproximadamente normal, la desviación estándar poblacional σ es desconocida y el tamaño de la muestra es grande (n > 30), se utiliza la distribución normal para probar hipótesis respecto a la media de una población utilizando el siguiente procedimiento:

a. Plantear las hipótesis a contrastar

1.	2.	3.
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
En este caso, como la varianza de la población σ^2 es desconocida, se utiliza la varianza muestral (S^2); por otro lado, como la población se distribuye normal, el tamaño de la muestra es grande y se quiere hacer inferencia para una sola media poblacional, se tiene que el estadístico apropiado es:

$$Z_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos
De igual manera que en el caso anterior, cuando la hipótesis alternativa es unilateral, la región crítica se localiza en la dirección correspondiente al signo de desigualdad. Si es bilateral, se encontrará en ambos lados de la gráfica, tal como se mostró en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución normal.

e. Calcular el estadístico

$$Z_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

f. Decidir e interpretar
Se compara el valor del estadístico calculado con el valor crítico, y si cae en la región de rechazo de la hipótesis nula, se concluye que el contraste es estadísticamente significativo.

Ejemplo 4.4:

Con los datos del ejemplo 4.3, verifique si los automóviles se manejan en promedio menos de 24.700 km/año con un nivel de significación del 5 %.

Solución:

a. Plantear las hipótesis a contrastar

$$H_0: \mu = 24.700$$
$$H_1: \mu < 24.700$$

b. Fijar el nivel de significación: $\alpha = 0{,}05$

c. Establecer el estadístico a utilizar

$$Z_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos

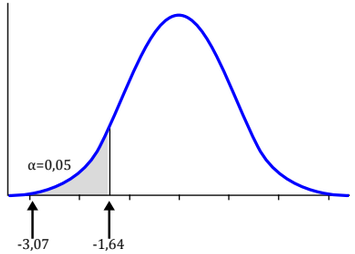


Figura 4.5 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$Z_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{23.500 - 24.700}{\frac{3.900}{\sqrt{100}}} = -3{,}07$$

f. Decidir e interpretar
Como el estadístico -3,07 < -1,64, se rechaza la hipótesis nula; por lo tanto, los automóviles se manejan en promedio menos de 24.700 km/año, con un nivel de significación del 5 %.

Escenario 3:

Cuando la población se distribuye normal con media μ, la desviación estándar poblacional σ es desconocida y el tamaño de la muestra n es pequeño.

La distribución normal no es adecuada en aquellos casos que no se conoce la desviación estándar de la población (σ), el tamaño de la muestra es pequeño (n < 30) y la población de la cual se selecciona la muestra es normal o aproximadamente normal. Cuando se presentan estas características, la distribución apropiada para hacer inferencia sobre la media poblacional es la distribución *t-student.*

1. Estimación por intervalos

Bajo las condiciones que se están considerando en este caso, un intervalo de (1 – α) 100 % de confianza para una sola media poblacional (μ) viene dado por:

$$\bar{X} - t_{(n-1;\frac{\alpha}{2})}\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + t_{(n-1;\frac{\alpha}{2})}\left(\frac{S}{\sqrt{n}}\right) \quad (4.13)$$

donde S representa la desviación estándar de la muestra de estudio.

Ejemplo 4.5:

Los datos de las fallas en las pruebas por tracción para la adhesión en 22 muestras de cierta aleación mostraron un promedio de 13,71 fallas, con una desviación estándar de 3,55 fallas. Calcule e interprete un intervalo de confianza del 95 % para el verdadero número de fallas promedio.

Solución:
Se trata del promedio en el número de fallas en las pruebas de tracción, y dado que n es pequeño, que el nivel de significación es (1 – α) = 0,95, entonces $t_{0,005;n-1=21} = 2{,}080$.

Sustituyendo en la ecuación (4.12.), se tiene que:

$$\left(13{,}71 - 2{,}080\left(\frac{3{,}55}{\sqrt{22}}\right) \leq \mu \leq 13{,}71 + 2{,}080\left(\frac{3{,}55}{\sqrt{22}}\right)\right)$$

$$(13{,}71 - 1{,}57 \leq \mu \leq 13{,}71 + 1{,}57)$$

$$(12{,}14 \leq \mu \leq 15{,}28)$$

Hay una confianza del 95 % de que el verdadero número de fallas promedio en las pruebas de tracción está entre 12,14 y 15,28 fallas.

2. Contraste de hipótesis

Para contrastar las hipótesis cuando la población es normal o aproximadamente normal, la desviación estándar poblacional σ es desconocida y el tamaño de la muestra es pequeño (n < 30), se utiliza la distribución *t-student* para probar hipótesis respecto a la media de una población utilizando el siguiente procedimiento:

a. Plantear las hipótesis a contrastar

• $H_0: \mu = \mu_0$	• $H_0: \mu = \mu_0$	• $H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
Al cumplirse que la desviación estándar de la población (σ) es desconocidala población se distribuye normal, el tamaño de la muestra es pequeño y se quiere hacer inferencia sobre una sola media poblacional, el estadístico apropiado es:

$$t_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos
De igual manera que en los casos anteriores, la región crítica se localiza en la dirección correspondiente al signo de desigualdad en la hipótesis alternativa. Si es bilateral, se encontrará en ambos lados de la gráfica, tal como se muestra en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución *t-student*.

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$t_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

f. Decidir e interpretar
Se compara el valor del estadístico calculado con el valor crítico, y si cae en la región de rechazo de la hipótesis nula, se concluye que el contraste es estadísticamente significativo.

Ejemplo 4.6:

Se afirma que una máquina de aspirar gasta un promedio de 46 kilowatt-hora al año. Si una muestra aleatoria de 12 hogares que se incluye en un estudio planeado indica que las aspiradoras gastan un promedio de 42 kilowatt-hora al año, con una desviación estándar de 11,9 kilowatt-hora, ¿sugiere esto que las aspiradoras gastan en promedio menos de 46 kilowatt-hora anualmente? Use un nivel de significación de 0,05.

Solución:

a. Plantear las hipótesis a contrastar

$$H_0: \mu = 46\ kw/hora$$
$$H_0 < \mu = 46\ kw/hora$$

b. Fijar el nivel de significación: α= 0,05

c. Establecer el estadístico a utilizar

$$t_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos

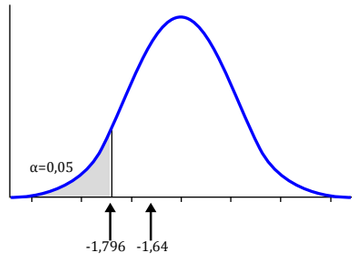


Figura 4.6 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$t_c = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{42 - 46}{11{,}9/\sqrt{12}} = -1.16$$

f. Decidir e interpretar
Como el estadístico -1,16 > -1,796, no se rechaza la hipótesis nula, por lo tanto, el número promedio de kilowatt-hora que gastan al año las aspiradoras domésticas no es significativamente menor que 46.

4.2 Inferencia para la diferencia de dos medias poblacionales (μ_1-μ_2)

En cuanto a la metodología a implementar para realizar inferencias acerca de $\mu_1 - \mu_2$, se debe decir que es similar a la utilizada en el caso de una población, y que, al igual que allí, se presentarán diferentes escenarios o condiciones para hacer esas inferencias.

La notación a utilizar para los datos muestrales es la siguiente:

$x_1, x_2, \ldots, x_{n_j}$ Muestras aleatorias de tamaño n_j de la población j, para j = 1,2

$\bar{x}_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j}$ Media de la muestra j, para j = 1,2

$S_j^{\,2} = \frac{\sum_{i=1}^{n_1} (x_{ij} - \bar{x}_j)^2}{n_j - 1}$ Varianza de muestra j, para j = 1,2

El estadístico razonable para hacer inferencias sobre $\mu_1 - \mu_2$ es $\bar{X}_1 - \bar{X}_2$ y, efectivamente, resulta ser el más apropiado para desarrollar los procedimientos de estimación y contrastación de hipótesis. De tal manera que es imprescindible conocer la distribución de probabilidad del estadístico $\bar{X}_1 - \bar{X}_2$ bajo las diferentes condiciones en las cuales se desea hacer inferencias acerca de $\mu_1 - \mu_2$.

Ahora bien, antes de determinar el tipo de distribución de probabilidad que tiene $\bar{X}_1 - \bar{X}_2$, es posible adelantar cuál será la media o valor esperado y la varianza de esa distribución, independientemente de las condiciones que se presenten.

Tal como es conocido, la media y la varianza de la diferencia de variables aleatorias viene dada por:

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) - 2COV(\bar{X}_1; \bar{X}_2)$$

$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2COV(\bar{X}_1; \bar{X}_2)$$

Si las muestras de las dos poblaciones son independientes, entonces también $\bar{X}_1$ y $\bar{X}_2$ son independientes y, por lo tanto, $\mathrm{COV}(\bar{X}_1; \bar{X}_2) = 0$, de donde:

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Estos resultados muestran que la media y la varianza del estadístico $\bar{X}_1 - \bar{X}_2$, en caso de muestras independientes, son $\mu_1 - \mu_2$ y $\bar{X}_1$ y $\bar{X}_2$ respectivamente, con independencia de cuál sea el tipo de distribución que siga $\bar{X}_1$ y $\bar{X}_2$. Se observa, por otro lado, que el estadístico $\bar{X}_1$ y $\bar{X}_2$ es un estimador insesgado de $\mu_1 - \mu_2$.

Se presentan en esta sección diferentes escenarios para comparar dos medias poblacionales $\mu_1 - \mu_2$ a partir de la información contenida en dos muestras extraídas al azar de dichas poblaciones. Cada escenario dependerá, en primer lugar, de si las muestras se asumen independientes; luego, de si se conocen las varianzas poblacionales, y, por último, del tamaño de la muestra.

4.2.1 Muestras independientes

Escenario 4:

Cuando las poblaciones se distribuyen normal con medias μ_1 y μ_2, las desviaciones estándar poblacionales σ_1 y σ_2 son conocidas y los tamaños de las muestras son grandes ($n_1 > 30$ y $n_2 > 30$).

Dado que las poblaciones son normales y que las muestras son independientes, se cumple que el estadístico $\bar{X}_1 - \bar{X}_2$ sigue una distribución normal $N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ y, por lo tanto:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0;1) \qquad (4.14)$$

En cuanto a las inferencias acerca de $\mu_1 - \mu_2$, se comienza por establecer que el mejor estimador puntual de esa diferencia de medias poblacionales es precisamente la correspondiente diferencia de medias muestrales $(\bar{X}_1 - \bar{X}_2)$.

El estimador $(\bar{X}_1 - \bar{X}_2)$ satisface todas las condiciones deseables en un buen estimador puntual.

1. Estimación por intervalos

Para estimar $\mu_1 - \mu_2$ mediante un intervalo de confianza se sigue un procedimiento similar al descrito en el caso de una media, llegando al siguiente resultado. Bajo las condiciones de poblaciones normales, varianza conocidas y muestras independientes, un intervalo de confianza del $(1 - \alpha)$ 100 % para $\mu_1 - \mu_2$ viene dado por:

$$(\bar{X}_1 - \bar{X}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (4.15)$$

siendo $Z_{\alpha/2}$ el valor de la normal estandarizada tal que a su derecha hay un área bajo la curva igual a $\alpha/2$.

2. Contraste de hipótesis

a. Plantear las hipótesis a contrastar

- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 < \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 \neq \mu_2$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
Al cumplirse que las desviaciones estándar poblacionales (σ_1^2 y σ_2^2) son conocidas, las poblaciones son normales e independientes, el tamaño de las muestras es grande ($n_1 > 30$ y $n_2 > 30$) y se quiere hacer inferencia sobre la diferencia de medias poblacionales, el estadístico apropiado es:

$$Z_c = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0;1) \qquad (4.16)$$

d. Determinar el valor y la región críticos
De igual manera que en los casos anteriores, la región crítica se localiza en la dirección correspondiente al signo de desigualdad en la hipótesis alternativa. Si es bilateral, se encontrará en ambos lados de la gráfica, tal como se muestra en la figura 4.3. Los valores críticos se deben ubicar en la tabla de la distribución normal estandarizada.

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$Z_c = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

f. Decidir e interpretar
Se compara el valor del estadístico calculado con el valor crítico, y si cae en la región de rechazo de la hipótesis nula, se concluye que el contraste es estadísticamente significativo.

Ejemplo 4.7:

Para medir el rendimiento en millas/galón de gasolina, se comparan dos tipos de motores: T_1 y T_2. Para esto se realizan 50 experimentos con el motor T_1 y 75 con el motor T_2, con rendimientos promedio de 36 y 42 millas respectivamente. Se conoce además que las desviaciones estándar poblacionales son de 6 para los motores T_1 y de 8 para los motores T_2. Se pide que:

i. Encuentre un intervalo del 96 % de confianza para la diferencia en el rendimiento promedio de gasolina entre ambos motores.
ii. ¿Habrá diferencia significativa en el rendimiento promedio de los dos motores? Use un nivel de significación del 4 %.

Solución:

Tabla 4.1 Datos

Parámetro	Motores T_1	Motores T_2
Media	36	42
Desviación estándar poblacional	6	8
Tamaño muestra	75	50

Fuente: MLB

i. Sustituyendo los datos en la ecuación (4.14), se tiene que:

$$\left((36-42) - 2{,}05\sqrt{\frac{36}{75} + \frac{64}{50}} \leq \mu_1 - \mu_2 \leq (36-42) + 2{,}05\sqrt{\frac{36}{75} + \frac{64}{50}}\right)$$

$$(-8{,}73 \leq \mu_1 - \mu_2 \leq -3{,}27) \text{ millas/galón}$$

Se concluye que hay una confianza del 96 % y que la diferencia en el rendimiento promedio de gasolina entre ambos motores se encuentra entre -8,73 y -3,27 millas/galón.

ii. Para saber si hay diferencia significativa entre los dos tipos de motores se debe proceder tal como lo establecen los pasos para la contrastación vistos anteriormente, a saber:

a. Plantear las hipótesis a contrastar

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

b. Fijar el nivel de significación: $\alpha = 0{,}04$

c. Establecer el estadístico a utilizar

$$Z_c = \frac{((\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0;1)$$

d. Determinar el valor y la región críticos

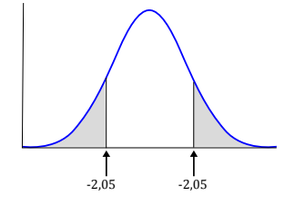


Figura 4.7 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$Z_c = \frac{(36-42)-(0)}{\sqrt{\frac{36}{75}+\frac{64}{50}}} = \frac{-6}{1{,}33} = -4{,}51$$

f. Decidir e interpretar
Como el estadístico $-4{,}51 < -2{,}05$, se rechaza la hipótesis nula. Por lo tanto, existe una diferencia significativa en el rendimiento promedio de gasolina entre ambos motores.

Escenario 5:

Cuando las poblaciones se distribuyen normal con medias μ_1 y μ_2, las desviaciones estándar poblacionales σ_1 y σ_2 son desconocidas, pero se suponen iguales, y las muestras son independientes, de tamaño pequeño ($n_1 < 30$ y $n_2 < 30$).

Bajo estas condiciones, el estadístico $\bar{X}_1 - \bar{X}_2$ sigue una *distribución t-student* con $(n_1 + n_2 - 2)$ grados de libertad y, por lo tanto:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

Siendo:

$$S_c = \sqrt[2]{\frac{(n_1-1)S_1^2 + (n_2-1)S_1^2}{n_1+n_2-2}}$$

Debe tenerse en cuenta que el mejor estimador puntual de esa diferencia de medias poblacionales es la diferencia de medias muestrales $(\bar{X}_1 - \bar{X}_2)$. Además, este satisface todas las condiciones deseables en un buen estimador puntual.

1. Estimación por intervalos

La estimación de $\mu_1 - \mu_2$ mediante un intervalo de confianza viene dada por:

$$(\bar{X}_1 - \bar{X}_2) - t_{(n_1+n_2-2;\,\alpha/2)}S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X}_1 - \bar{X}_2) + t_{(n_1+n_2-2;\,\alpha/2)}S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (4.18)$$

donde $t_{(n_1+n_2-2;\,\alpha/2)}$ el valor en la tabla de la *t-student,* tal que a su derecha hay un área bajo la curva igual a $\alpha/2$.

2. Contraste de hipótesis

a. Plantear las hipótesis a contrastar

- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 < \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 \neq \mu_2$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
Al cumplirse que las desviaciones estándar poblacionales (σ_1 y σ_2) son desconocidas, pero se asumen iguales, las poblaciones son normales e independientes, el tamaño de las muestras es pequeño (y se quiere hacer inferencia sobre la diferencia de promedios poblacionales, el estadístico apropiado es:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2-2)} \qquad (4.19)$$

d. Determinar el valor y la región críticos
De igual manera que en los casos anteriores, la región crítica se localiza en la dirección correspondiente al signo de desigualdad en la hipótesis alternativa. Si es bilateral, se encontrará en ambos lados de la gráfica, tal como se muestra en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución *t-student.*

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

f. Decidir e interpretar
Se compara el valor del estadístico calculado con el valor crítico, y si cae en la región de rechazo de la hipótesis nula, se concluye que el contraste es estadísticamente significativo.

Ejemplo 4.8:

Con el objetivo de determinar los efectos del ejercicio por un tiempo prolongado en los ingenieros de una compañía inscritos en un programa supervisado de acondicionamiento físico, se registraron los datos de 13 ingenieros que voluntariamente se inscribieron en el programa (activos) y de 17 que decidieron no inscribirse (sedentarios). Se les pidió a los 30 individuos realizar sentadillas por 30 segundos, obteniéndose la siguiente información:

Tabla 4.2 Datos

	Activos	Sedentarios
Media	21	12,1
Desviación estándar	4,9	5,6
Tamaño muestra	13	17

Fuente: MLB

Asumiendo que las varianzas poblacionales desconocidas son iguales, se pide que:

i. Elabore un intervalo de confianza del 95 % para la diferencia de medias poblacionales.
ii. ¿Habrá diferencia significativa en el número de sentadillas promedio entre los ingenieros que hacen ejercicio y los sedentarios? Use un nivel de significación del 5 %.

Solución:

i. En este caso, se observa que no se conocen las varianzas poblacionales, pero que se están asumiendo iguales; en este sentido, el intervalo de confianza viene dado por la ecuación (4.18.):

$$\left((\bar{X}_1-\bar{X}_2)-\ t_{(n_1+n_2-2;\,\alpha/2)}S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le \mu_1-\ \mu_2\le (\bar{X}_1-\bar{X}_2)+\ t_{(n_1+n_2-2;\,\alpha/2)}S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

Donde el valor de la $t_{(n_1+n_2-2;\,\alpha/2)}=t_{(13+17-2;0{,}025)=}2{,}0484$ se obtiene de la tabla de la distribución *t-student* (anexo B).

Además,

$$S_c=\sqrt[2]{\frac{(n_1-1)S_1^2+(n_2-1)S_1^2}{n_1+n_2-2}}=\sqrt[2]{\frac{(13-1)\times 24{,}01+(17-1)\times 31{,}36}{13+17-2}}=5{,}31$$

Sustituyendo, se tiene:

$$\left((21-12{,}1)-\ 2{,}0484\times 5{,}31\sqrt{\frac{1}{13}+\frac{1}{17}}\le \mu_1-\ \mu_2\le (21-12{,}1)+\ 2{,}0484\times 5{,}31\sqrt{\frac{1}{13}+\frac{1}{17}}\right)$$

$$(8{,}09-\ 4{,}0085\le \mu_1-\ \mu_2\le 8{,}09-\ 4{,}0085)$$

$$(4{,}9\le \mu_1-\ \mu_2\le 12{,}9)$$

Hay una confianza del 95 % y la diferencia en el número de sentadillas promedio entre los ingenieros que hacen ejercicio y los sedentarios se encuentra entre 4,9 y 12,9.

ii. Los pasos a seguir para la contrastación se presentan a continuación:

a. Plantear las hipótesis a contrastar

$$H_0{:}\mu_1=\mu_2$$
$$H_1{:}\mu_1\neq\mu_2$$

b. Fijar el nivel de significación: α =0,05

c. Establecer el estadístico a utilizar

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\ \mu_2)}{S_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{(n_1+n_2-2;\alpha/2)}$$

d. Determinar el valor y la región críticos

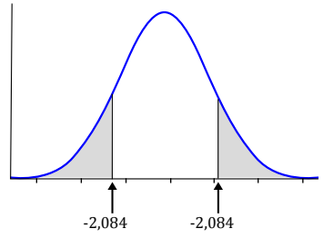


Figura 4.8 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico
f. Decidir e interpretar
Como 12,39 > 2,084, se rechaza la hipótesis nula, evidenciando una diferencia significativa en el número de sentadillas promedio que realizan los ingenieros activos en comparación con los sedentarios.

Escenario 6:

Cuando las poblaciones se distribuyen normal con medias μ_1 y μ_2, las desviaciones estándar poblacionales σ_1 y σ_2 son desconocidas, pero se suponen diferentes, y las muestras son independientes, de tamaño pequeño ($n_1 < 30$ y $n_2 < 30$).

Bajo estas condiciones, el estadístico $\bar{X}_1 - \bar{X}_2$ sigue una *distribución t-student* con (υ) grados de libertad y, por lo tanto:

$$t = \frac{8{,}9 - 0}{5{,}31\sqrt{\frac{1}{13} + \frac{1}{17}}} = \frac{8{,}9}{0{,}721} = 12{,}39$$

Donde:

$$\upsilon = \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad (4.20)$$

Al igual que en los casos anteriores, el mejor estimador puntual de esa diferencia de medias poblacionales es la diferencia de medias muestrales $\bar{X}_1 - \bar{X}_2$.

1. Estimación por intervalos

La estimación de $\mu_1 - \mu_2$ mediante un intervalo de confianza viene dada por:

$$(\bar{X}_1 - \bar{X}_2) - t_{(\upsilon;\, \alpha/2)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{(\upsilon;\, \alpha/2)}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \qquad (4.21)$$

donde $t_{(\upsilon;\, \alpha/2)}$ el valor en la tabla de la *t-student* tal que a su derecha hay un área bajo la curva igual a $\alpha/2$.

2. Contraste de hipótesis

a. Plantear las hipótesis a contrastar

- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 < \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 > \mu_2$
- $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 \neq \mu_2$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
Al ser las varianzas poblacionales(σ_1^2 y σ_2^2) desconocidas, pero diferentes, las poblaciones normales e independientes, el tamaño de las muestras pequeño ($n_1 < 30$ y $n_2 < 30$)y se quiere hacer inferencia sobre la diferencia de medias poblacionales, el estadístico apropiado es:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \upsilon \qquad (4.22)$$

d. Determinar el valor y la región críticos
De igual manera que en los casos anteriores, la región crítica se localiza en la dirección correspondiente al signo de desigualdad en la

hipótesis alternativa. Si es bilateral, se encontrará en ambos lados de la gráfica, tal como se muestra en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución *t-student*.

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$t_c = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

f. Decidir e interpretar
Se compara el valor del estadístico calculado con el valor crítico, y si cae en la región de rechazo de la hipótesis nula, se concluye que el contraste es estadísticamente significativo.

Ejemplo 4.9:

Un estudio para determinar la cantidad de fósforo químico (miligramos por litro) medido en dos estaciones diferentes (A y B) de un río produjo los siguientes resultados:

Tabla 4.3 Datos

	Estación A	Estación B
Media	3,84	1,49
Desviación estándar	3,07	0,80
Tamaño muestra	15	12

Fuente: MLB

Suponga que las mediciones provienen de poblaciones normales con varianzas diferentes. Se pide que:

i. Elabore un intervalo de confianza del 95 % para la diferencia en la cantidad de fósforo químico promedio entre las dos estaciones.
ii. A un nivel de significación del 5 %, ¿será mayor la cantidad de fósforo en la estación A del río?

Solución:
Para el intervalo de confianza se deben conocer los grados de libertad asociados. Usando la ecuación (4.19.), se tiene que:

$$\upsilon = \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} =$$

$$\frac{\frac{9,43}{15} + \frac{0,64}{12}}{\frac{\left(\frac{9,43}{15}\right)^2}{15 - 1} + \frac{\left(\frac{0,64}{12}\right)^2}{12 - 1}} =$$

$$\upsilon = 16,3 = 16$$

En la tabla de la distribución *t* se observa que con 16 grados de libertad y un nivel de significación de 0,05, el valor correspondiente es 2,120 (anexo B).

Luego, sustituyendo en la ecuación (4.21), se obtiene:

$$\left((3,84 - 1,49) - 2,120\sqrt{\frac{9,43}{15} + \frac{0,64}{12}} \leq \mu_1 - \mu_2 \leq (3,84 - 1,49) + 2,120\sqrt{\frac{9,43}{15} + \frac{0,64}{12}}\right)$$

$$(0,60 \leq \mu_1 - \mu_2 \leq 4,10)$$

Por lo que con una confianza del 95 %, la verdadera diferencia en la cantidad de fósforo químico promedio entre las dos estaciones se encuentra entre 0,60 y 4,10 miligramos por litro.
ii. Para la contrastación se sigue el siguiente esquema:

a. Plantear las hipótesis a contrastar

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 > \mu_2$$

b. Fijar el nivel de significación: $\alpha = 10$

c. Establecer el estadístico a utilizar

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt[2]{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(\upsilon; \alpha/2)$$

d. Determinar el valor y la región críticos

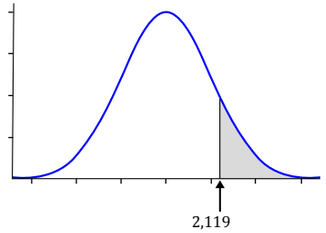


Figura 4.9 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$t = \frac{2{,}35 - 0}{0{,}8256} = 2{,}84$$

f. Decidir e interpretar
Como el valor del estadístico 2,84 es mayor al valor de la tabla 2,119, se rechaza la hipótesis nula, dejando en evidencia que la cantidad de ortofósforo en la estación A es mayor.

Muestras dependientes

En todos los casos anteriores de inferencias acerca de dos medias poblacionales se ha exigido que las muestras sean independientes. Sin embargo, existen situaciones muy comunes en las cuales las muestras están relacionadas, es decir, los valores de una muestra no son independientes de los valores de la otra muestra. En estos casos, se habla de muestras dependientes o muestras apareadas.

Básicamente, la idea de formar pares o de recabar información de parejas es que haya homogeneidad en cada dupla y a la vez heterogeneidad de un par a otro.

Los casos de muestras apareadas se presentan en numerosas situaciones, entre las cuales se pueden mencionar:

a. Cuando se desea comparar el rendimiento de dos variedades A y B de maíz. Para ello se someten a prueba en diferentes estaciones agronómicas o experimentales, las cuales presentan entre sí diferentes condiciones de suelo, luz, humedad, etc. En cada estación experimental se utilizan dos parcelas similares, una para el maíz A y otra para el maíz B, de tal manera que ambas variedades sean cosechadas bajo las mismas condiciones en cada estación y en diferentes condiciones de estación a estación. Se forman pares para tratar de eliminar el efecto que tenga cada estación en el rendimiento del maíz.
b. Cuando se desea investigar si un nuevo método de entrenamiento es efectivo para mejorar la capacidad de memorización de las personas por medio de mediciones de antes y después de aplicación del método. En general, en los casos de situaciones de antes y después surgen los datos apareados.
c. En ciertas ocasiones existen limitaciones ambientales o de tiempo que impiden que un experimento se realice totalmente en una sola jornada y se hace necesario completarlo en varios días, pero las condiciones climáticas pueden influenciar el resultado del experimento. Si se desea comparar dos “tratamientos” particulares, se debe planificar el experimento de forma tal que en cada uno de los días se sometan a prueba los dos tratamientos y de esa manera se generan datos apareados diariamente.
d. Hay investigaciones biológicas en las cuales los elementos de experimentación lo constituyen gemelos o camadas de animales. En estas indagaciones resulta obligado planificar el experimento para generar datos apareados por cada camada o gemelos.

Se debe puntualizar que la formación de pares persigue el objetivo de eliminar el efecto que tiene cada pareja en el rendimiento de las variables que se comparan.

En términos formales, la estructura de los datos y del problema es la siguiente:

a. Se dispone de n pares de datos $(x_1; x_2), (x_2, y_2), \dots, (x_n, y_n)$; siendo los mismos independientes entre sí, pero dependientes entre x_i y y_i.

b. Interesa trabajar con las diferencias $d_i = x_i - y_i (I = 1,2, \dots, n)$ para eliminar el efecto de cada pareja.

c. Se va a denotar por μ_d y σ_d^2 a la media y varianza de la población constituida por todas las posibles diferencias d_i, es decir:

$$\mu_d = E(d_i) = E(x_i - y_i) = \mu_x - \mu_y$$
$$\sigma_d^2 = V(d_i) = V(x_i - y_i)$$

En consecuencia, se puede considerar a $d_1, d_2, \dots, d_n$ como una muestra aleatoria proveniente de una población con media μ_d y varianza σ_d^2

Si $\mu_d = 0$ quiere decir que las medias μ_x y μ_y son iguales o que los dos "tratamientos" tienen el mismo efecto. Si μ_d es mayor que cero, entonces μ_x es mayor que μ_y y si μ_d es menor que cero, entonces μ_x es menor que μ_y.

Si $d_1, d_2, \dots, d_n$ es muestra aleatoria de tamaño n proveniente de una población con media μ_d y varianza σ_d^2, se pueden realizar inferencias acerca de μ_d exactamente como se hizo acerca de una media poblacional μ anteriormente.

A continuación, se presentan los resultados que pueden deducirse bajo dos escenarios diferentes. El primero establece lo siguiente:
Si $d_1, d_2, \dots, d_n$ es una muestra aleatoria proveniente de una población $N(\mu_d; \sigma^2{}_d)$ entonces un intervalo de confianza del $(1 - \alpha)$ para μ_d viene dado por:

$$\bar{d} - t_{(n-1;\frac{\alpha}{2})}\left(\frac{S_d}{\sqrt{n}}\right) \leq \mu_d \leq \bar{d} + t_{(n-1;\frac{\alpha}{2})}\left(\frac{S_d}{\sqrt{n}}\right) \quad (4.23)$$

Siendo:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \; ; \; S_d = \frac{\sum (d_i - \bar{d})^2}{n-1}$$

$t_{(n-1;\frac{\alpha}{2})}$ el valor de la variable t con $(n - 1)$ grados de libertad, que deja a su derecha un área bajo la curva de $\alpha/2$

Para realizar cualquier contraste de hipótesis de μ_d se utiliza el estadístico:

$$t_c = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} \quad (4.24)$$

el cual bajo H_0 sigue una distribución *t-student* con $(n - 1)$ grados de libertad.

El otro escenario que interesa es aquel en el cual el tamaño de la muestra es grande y no es necesario exigir que la población siga una distribución normal.

Si el tamaño n de la muestra es suficientemente grande, se puede realizar cualquier contraste de hipótesis acerca de μ_d basándose en el estadístico:

$$Z_c = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}$$

el cual, bajo la hipótesis nula, sigue una distribución aproximadamente $N(0;1)$. Así mismo, un intervalo de confianza del $(1 - \alpha)$ para μ_d viene dado por:

$$\bar{d} - Z_{\alpha/2}\left(\frac{S_d}{\sqrt{n}}\right) \leq \mu_d \leq \bar{d} - Z_{\alpha/2}\left(\frac{S_d}{\sqrt{n}}\right) \quad (4.25)$$

Ejemplo 4.10:

En una empresa vendedora de electrodomésticos se somete a prueba un nuevo método de entrenamiento para mejorar la eficiencia de sus vendedores. Este nuevo método ha sido desarrollado por el departamento de sicología de la universidad de la ciudad. La eficiencia en ventas se mide en una escala de 0 a 100 a través de un índice especial diseñado para tal efecto. Se seleccionan aleatoriamente 10 vendedores de la empresa, los cuales tienen una experiencia variada en ventas, y se les mide

su eficiencia actual. Se someten al nuevo entrenamiento y luego de un tiempo razonable de trabajo se les mide nuevamente su eficiencia en ventas. La información obtenida es la siguiente:

Tabla 4.4 Datos

Vendedor	1	2	3	4	5	6	7	8	9	10
Antes	80	70	77	63	66	68	80	62	70	55
Después	84	75	74	70	69	66	86	65	74	60

Fuente: MLB

¿Hay razón para creer que el nuevo método de entrenamiento es realmente útil para mejorar el rendimiento de los vendedores? Usar α = 0,02

Solución:
A cada vendedor se la ha medido su efectividad antes y después del entrenamiento. Se denota con x_1 a los datos después del entrenamiento y con y_i a los datos antes del entrenamiento. Esta designación es arbitraria y puede hacerse a conveniencia del investigador, pero una vez elegidos los x_1 y y_i, debe desarrollarse la solución del problema de acuerdo con esta selección, especialmente en los referentes al planteamiento de la hipótesis del contraste.

Para plantear la hipótesis se debe tener en cuenta que se quiere mejorar la eficiencia. En este sentido, la variable en el momento dos debe ser mayor que en el momento uno para que esto se pueda cumplir. Las hipótesis se plantean como:

a. Plantear las hipótesis a contrastar

$$H_0: \mu_1 = \mu_2 \quad \text{ó} \quad H_0: \mu_1 - \mu_2 = \mu_d$$
$$H_1: \mu_1 < \mu_2 \quad \text{ó} \quad H_0: \mu_1 - \mu_2 < \mu_d$$

b. Fijar el nivel de significación: α = 0,02

c. Establecer el estadístico a utilizar

$$t_c = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}$$

d. Determinar el valor y la región críticos

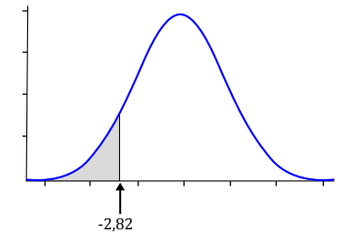


Figura 4.10 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico
Se debe conocer el promedio y la varianza de la nueva variable aleatoria, para lo cual sirve de apoyo la tabla a continuación:

Tabla 4.5 Datos

Vendedor	1	2	3	4	5	6	7	8	9	10	Suma
Antes	80	70	77	63	66	68	80	62	70	55	
Después	84	75	74	70	69	66	86	65	74	60	
d_i	-4	-5	3	-7	-3	2	-6	-3	-4	-5	-32
d_i^2	16	25	9	49	9	4	36	9	16	25	198

Fuente: MLB

Se tiene entonces que la media de la variable aleatoria *d* viene dada por:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{-32}{10} = -3.2$$

Y la desviación estándar de la variable aleatoria d viene dada por:

$$S_d = \frac{\sum(d_i - \bar{d})^2}{n-1} = \frac{198 - 10\times(-3.2)^2}{10-1} = 10,$$

Luego, sustituyendo en la ecuación (4.23.) se tiene que:

$$t = \frac{-3,2 - 0}{3,26} = -0,98$$

f. Decidir e interpretar
Como el valor del estadístico -0,98 es mayor al valor de la tabla -2,82, no se rechaza la hipótesis nula. En conclusión, los vendedores no han mejorado la eficiencia.

Ejercicios

1. Una empresa proporciona agua embotellada en contenedores de 15 galones a las casas de un sector de 3 parroquias. El gerente desea estimar el número promedio de contenedores que una casa utiliza cada mes. Se toma una muestra de 25 casas y se registra el número de contenedores, y se obtiene una media de 3,2 y una desviación estándar de 0,78.

   i. ¿Qué revelaría un intervalo de 92 % de confianza?
   ii. Pruebe la hipótesis de que el verdadero número medio de contenedores que usa una casa cada mes no es superior a 3. Use un nivel de significación de 5 %.
   iii. Se selecciona una muestra más pequeña de 10 casas para estimar el número promedio de miembros de la familia por casa. Los resultados son: 1, 3, 4, 7, 2, 2, 3, 5, 6 y 6 personas en cada casa. ¿Cuáles son los resultados de un intervalo de confianza de 99 % para el número promedio de miembros de la familia?

2. La estatura de una muestra aleatoria de 50 estudiantes universitarios arroja una media de 174,5 centímetros y una desviación estándar de 6,9 centímetros.
   i. Construya un intervalo de confianza de 98 % para la estatura media de todos los estudiantes de la universidad.
   ii. ¿Qué podemos afirmar con 98 % de confianza sobre el tamaño posible de nuestro error si estimamos que la estatura media de todos los estudiantes de la universidad es de 174,5 centímetros?

3. Una empresa eléctrica fabrica focos que tienen una duración aproximadamente distribuida de forma normal, con una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una duración promedio de 780 horas, encuentre un intervalo de confianza de 96 % para la media de la población de todos los focos que produce esta empresa.

4. Las siguientes mediciones se registraron para el tiempo de secado, en horas, de cierta marca de puntura de látex:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones representan una muestra aleatoria de una población normal. Encuentre los límites de tolerancia de 99 % que contendrán 95 % de los tiempos de secado.

5. Una muestra aleatoria de tamaño $n_1 = 25$ que se toma de una población normal con una desviación estándar de $\sigma_1 = 5$, tiene una media $\overline{x_1} = 80$. Una segunda muestra aleatoria de tamaño $n_2 = 36$, que se toma de una población normal diferente con una desviación estándar $\sigma_2 = 3$, tiene una media $\overline{x_2} = 75$. Encuentre un intervalo de confianza de 95 % para $\mu_1 - \mu_2$.

6. Una muestra aleatoria de 10 barras de chocolate energético de cierta marca tiene, en promedio, 230 calorías, con una desviación estándar de 15 calorías. Construya un intervalo de confianza de 99 % para el contenido medio real de calorías de esta marca de barras de chocolate energético. Suponga que la distribución de las calorías es aproximadamente normal.

7. En un proceso químico por lotes se comparan los efectos de 2 catalizadores sobre la potencia de la reacción del proceso. Se preparó una muestra de 12 lotes con el uso del catalizador 1 y se obtuvo una muestra de 10 lotes con el catalizador 2. Los 12 lotes para los que se utilizó el catalizador 1 dan un rendimiento promedio de 85, con una desviación estándar muestral de 4, y para la segunda muestra el promedio es 81, con una desviación estándar muestral de 5. Encuentre un intervalo de confianza de 90 % para la diferencia entre las medias poblacionales. Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales.

8. Un experimento reportado en *Popular Science* compara las economías en combustible para dos tipos de camiones compactos a diésel equipados de forma similar. Suponga que se utilizaron 12 camiones de la marca A y 10 de la marca B en pruebas de velocidad constante de 90 kilómetros por litro. Si los 12 de la marca A promedian 16 km/litro con una desviación estándar de 1.0 km/litro, y los 10 de la marca B promedian 11 km/litro, con una desviación estándar de 0,8 km/litro. Construya un intervalo de confianza de 90 % para la diferencia entre los kilómetros promedio por litro de estos 2 camiones compactos. Suponga que las distancias por litro para cada modelo de camión están distribuidas de forma aproximadamente normal con varianzas iguales.

9. El gobierno otorga fondos para los departamentos de agricultura de 9 universidades para probar las capacidades de rendimiento de 2 nuevas variedades de trigo. Cada variedad se planta en parcelas de área igual en cada universidad y el rendimiento, en kilogramos por parcela, se registra como sigue:

Tabla 1. Rendimientos

	Universidad								
Variedad	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

   i. Encuentre un intervalo de confianza de 95 % para la diferencia media entre los rendimientos de las dos variedades. Suponga que las diferencias de rendimiento se distribuyen de forma aproximadamente normal.
   ii. Responda: ¿por qué se necesita el apareamiento en este problema?

10. Una compañía de taxis trata de decidir si comprar neumáticos de la marca A o de la B para su flotilla. Para estimar la diferencia de las dos marcas, se lleva a cabo un experimento con 12 neumáticos de cada marca. Los neumáticos se utilizan hasta que se gastan. Los resultados son:

Tabla 2. Duración de los neumáticos

Marca A:	Marca B:
$\bar{x}_A = 36.300\ km$	$\bar{x}_B = 38.100\ km$
$\hat{S}_A = 5.000\ km$	$\hat{S}_B = 6.100\ km$

Calcule un intervalo de confianza de 95 % para $\mu_1 - \mu_2$. Suponga que las poblaciones se distribuyen aproximadamente normal, y que las varianzas son iguales.

11. Para el ejercicio 10 pruebe la hipótesis de que no hay diferencia en las dos marcas de llantas con un nivel de significancia de 0,05. Suponga que las poblaciones se distribuyen de forma aproximadamente normal con varianzas iguales

12. Una muestra aleatoria de 64 bolsas de palomitas de maíz con queso cheddar pesa, en promedio, 5,23 onzas, con una desviación estándar de 0,24 onzas. Pruebe la hipótesis de que $\mu = 5.5$ onzas contra la hipótesis alternativa $\mu < 5.5$ onzas en el nivel de significancia de 0,05.

13. Una empresa eléctrica fabrica focos con una duración que se distribuye de forma aproximadamente normal con una media de 800 horas y una desviación estándar de 40 horas. Pruebe la hipótesis de que $\mu = 800$ horas contra la alternativa $\mu \neq 800$ horas si una muestra aleatoria de 30 focos tiene una duración promedio de 788 horas. Utilice un nivel de significancia de 0,04.

14. Pruebe la hipótesis de que el contenido promedio de los envases de un lubricante particular es de 10 litros si los contenidos de una muestra aleatoria de 10 envases son 10,2, 9,7, 10,1, 10,3, 10,1, 9,8, 9,9, 10,4, 10,3 y 9,8 litros. Utilice un nivel de significancia de 0,01 y suponga que la distribución del contenido es normal.

15. Un fabricante afirma que la resistencia a la tracción promedio del hilo A excede la resistencia a la tracción promedio del hilo B en al menos 12 kilogramos. Para comprobar esta afirmación, se prueban 50 piezas de cada tipo de hilo bajo condiciones similares. El hilo tipo A tiene una resistencia a la tracción promedio de 86,7 kilogramos, con una desviación estándar de 6,28 kilogramos, mientras que el hilo tipo B tiene una resistencia promedio a la tracción de 77,8 kilogramos, con una desviación estándar de 5,61 kilogramos. Pruebe la afirmación del fabricante con el uso de un nivel de significancia de 0,05.

5. Inferencia acerca de la proporción poblacional

INFERENCIA ACERCA DE LA PROPORCIÓN POBLACIONAL

En este capítulo se trata lo concerniente a la estimación por intervalos y el contraste de hipótesis para la proporción poblacional de manera individual y para la diferencia.

5.1 Inferencia para una sola proporción poblacional (p)

Se debe recordar que $p = \frac{X}{n}$ es un estimador puntual de la proporción P de la población a la que se esté haciendo referencia. Además, la proporción muestral se distribuye normalmente con media P y varianza $\frac{P(1-P)}{n} = \frac{PQ}{n}$, siempre y cuando n sea grande y p no esté muy cerca ni de cero ni de uno. Típicamente para aplicar esta aproximación se necesita que np y $n(1-p)$ sean mayores o iguales a cinco.

1. Estimación por intervalos

Si p es la proporción de observaciones en una muestra aleatoria de tamaño n que pertenece a una clase de interés, entonces un intervalo del $100(1-\alpha)$ % aproximado para la proporción P de la población a la que pertenece esta clase es:

$$p - Z_{\alpha/2}\left(\sqrt{\frac{P(1-P)}{n}}\right) \leq P \leq p + Z_{\alpha/2}\left(\sqrt{\frac{P(1-P)}{n}}\right) \quad (5.1)$$

Obsérvese que en la ecuación (5.1) se presenta el problema de que el parámetro P que se desea estimar aparece en los límites del intervalo de confianza y, en consecuencia, no se podría hacer la estimación. Una solución satisfactoria es reemplazar p por P en el error estándar. Luego, para un tamaño de muestra n grande, un intervalo de confianza del $(1 - \alpha)100$ % para la proporción poblacional P viene dado por:

$$\left(p - Z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right) \leq P \leq p + Z_{\alpha/2}\left(\sqrt{\frac{p(1-p)}{n}}\right)\right) \qquad (5.2)$$

siendo p la proporción muestral y $Z_{\alpha/2}$ el valor de la variable normal estandarizada, que deja a su derecha un área de $\alpha/2$.

2. Contrastación de hipótesis

Para contrastar hipótesis sobre una proporción poblacional se deben seguir los pasos que se presentan a continuación:

a. Plantear las hipótesis a contrastar

- $H_0{:}P = P_0$
 $H_1{:}P < P_0$
- $H_0{:}P = P_0$
 $H_1{:}P > P_0$
- $H_0{:}P = P_0$
 $H_1{:}P \neq P_0$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar

$$Z_c = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0;1) \qquad (5.3)$$

d. Determinar el valor y la región críticos

Tal como se vio en el capítulo 4, cuando la hipótesis alternativa es unilateral, la región crítica está localizada en la dirección correspondiente al signo de desigualdad, esto es, para la hipótesis unilateral izquierda, la región crítica se sitúa en la cola inferior de la distribución; si la hipótesis es unilateral derecha, se ubica en la parte superior de la distribución; mientras que si es bilateral, se encontrarán en ambos lados de la gráfica, tal como se mostró en la figura 4.3. Los valores críticos correspondientes se deben ubicar en la tabla de la distribución normal.

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

$$Z_c = \frac{p - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

f. Decidir e interpretar

Al comparar el valor del estadístico debe cerciorarse si cae en la región de rechazo de la hipótesis nula; en caso contrario, se expresa que no se rechaza la hipótesis nula o que el contraste no es estadísticamente significativo.

Ejemplo 5.1:

Formaron parte de una muestra aleatoria 85 rodamientos para el cigüeñal del motor de un automóvil, de los cuales 10 tenían un acabado muy rugoso sobre su superficie, de hecho, mayor al que dicen las especificaciones. Por lo tanto, una estimación puntual de la proporción de rodamientos de la población que excede la especificación de aspereza es $p = x/n = 10/85 = 0{,}12$. Se pide que:

i. Construya e interprete un intervalo de 95 % de confianza para la verdadera proporción de rodamientos que excede las especificaciones de aspereza.
ii. Responda: ¿será la proporción de rodamientos que excede las especificaciones de aspereza menor a 10 %? Use alfa de 0,025.

Solución:

i. La información necesaria para la resolución de este ejemplo es:
   a. $p = x/n = 10/85 = 0{,}12$
   b. $1 - \alpha = 0{,}95 \ \therefore \alpha = 0{,}05 \therefore \alpha/2 = 0{,}025$
   c. $Z_{\alpha/2} = Z_{0{,}025} = 1{,}96$

Sustituyendo en la ecuación (5.2.):

$$\left(0{,}12 - 1{,}96\sqrt{\frac{0{,}12(1-0{,}12)}{85}} \leq P \leq 0{,}12 + \ 1{,}96\sqrt{\frac{0{,}12(1-0{,}12)}{85}}\right)$$

$$\left(0{,}12 - 1{,}96\sqrt{\frac{0{,}12(0{,}88)}{85}} \leq P \leq 0{,}12 + \ 1{,}96\sqrt{\frac{0{,}12(0{,}88)}{85}}\right)$$

$$(0{,}05 \leq P \leq 0{,}19)$$

Luego, se tiene una confianza de 95 % de que la verdadera proporción de rodamientos que excede las especificaciones de aspereza estará contenida en el intervalo (0,05; 0,19) o (5; 19) %.

ii. Para conocer si la proporción de rodamientos que excede las especificaciones de aspereza es menor a 10 % se debe plantear, en primer lugar, la hipótesis a contrastar:

   a. Plantear las hipótesis a contrastar

$$H_0: P = 0{,}10$$

$$H_1: P < 0{,}10$$

   b. Fijar el nivel de significación: α = 0,025

   c. Establecer el estadístico a utilizar

$$Z_c = \frac{p - \ P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0;1$$

d. Determinar el valor y la región críticos

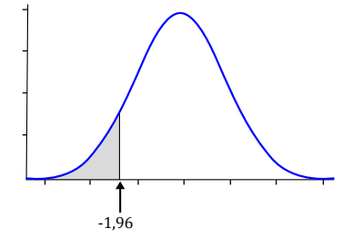


Figura 5.1 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$Z_c = \frac{p - \ P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0{,}12 - \ 0{,}10}{\sqrt{\frac{0{,}10(1-0{,}10)}{85}}} = 0{,}03$$

f. Decidir e interpretar
   Como el resultado obtenido en el punto anterior (0,03) es mayor que -1,96, que es el valor crítico, entonces se no rechaza la hipótesis nula; es decir, la proporción de rodamientos que excede las especificaciones de aspereza no es menor a 10 %.

5.2 Inferencia para dos proporciones poblacionales ($P_1 - P_2$)

Son muy frecuentes las comparaciones de dos proporciones poblacionales. Por ejemplo, las proporciones de productos defectuosos en dos procesos de producción; los porcentajes de efectividad de dos medicamentos contra cierta enfermedad; la proporción de muertes por el accidente de tránsito de una ciudad A en comparación con la de una ciudad B, etc.

Para realizar pruebas de hipótesis y estimaciones de dos proporciones poblacionales P_1 y P_2 sobre la base de la información suministrada por dos muestras independientes de las respectivas poblacionales, nos basamos en las correspondientes proporciones muestrales p_1 y p_2.

Tal como en el caso de dos medias poblacionales, cualquier inferencia acerca de P_1 y P_2 se puede realizar en términos de las diferencias $(p_1 - p_2)$. El estadístico $(p_1 - p_2)$ constituye el mejor estimador puntual de $(P_1 - P_2)$ y en él nos apoyaremos para realizar las inferencias correspondientes. En consecuencia, necesitaremos conocer la distribución de probabilidad de ese estadístico.

En el capítulo dos de este texto se estableció que la esperanza y la varianza de la proporción de una población, digamos 1, son respectivamente:

$$E(p_1) = P_1$$

$$V(p_1) = \frac{P_1(1-P_1)}{n_1} = \frac{P_1Q_1}{n_1}$$

donde: $Q_1 = 1 - P_1$ y n_1 es el tamaño de la muestra 1.

Para una población 2, la esperanza y la varianza para la proporción de esa población son:

a. Esperanza: $E(p_2) = P_2$

b. Varianza: $V(p_1) = \frac{P_2(1-P_2)}{n_2} = \frac{P_2Q_2}{n_2}$

siendo $Q_2 = 1 - P_2$ y n_2 el tamaño de la muestra 2.

Por otro lado, la media, la varianza y la desviación estándar de la diferencia de dos variables aleatorias independientes son:

c. Media: $E(p_1 - p_2) = E(p_1) - E(p_2) = P_1 - P_2$

d. Varianza: $V(p_1 - p_2) = V(p_1) + p_1(p_2) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}$

e. Desviación estándar: $DE(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

En el caso de que n_1 y n_2 sean suficientemente grandes (ambas mayores que 30), se tiene que:

$$(p_1 - p_2) \sim N\left(P_1 - P_{2;}\ \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)$$

Por lo tanto:

$$\frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}} \sim N(0;1)$$

Si en el denominador sustituimos $(P_1 - P_2)$ por sus estimadores puntuales $(p_1 - p_2)$ respectivamente, se obtiene una buena estimación de la desviación estándar, y la relación anterior se mantiene. Es decir:

$$\frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}} \sim N(0;1) \qquad (5.4)$$

Siendo $q_1 = 1 - p_1$, $q_2 = 1 - p_2$

A partir de esta última expresión se puede deducir una estimación por intervalo para $(P_1 - P_2)$.

1. Estimación por intervalos

Si los tamaños de las muestras n_1 y n_2 y son suficientemente grandes, un intervalo de confianza del $1(-\alpha)$ 100 % para $(P_1 - P_2)$ viene dado por:

$$\left[(p_1 - p_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} < (P_1 - P_2) < (p_1 - p_2) + Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right] \qquad (5.5)$$

2. Contrastación de hipótesis

Para contrastar la hipótesis nula H_0:$P_1 = P_2$ contra cualquier hipótesis alternativa H_1 vamos a denotar por P a la proporción poblacional común, es decir, $P_1 = P_2 = P$. Recordemos que al realizar un contraste de hipótesis interesa conocer la distribución de probabilidad del estadístico bajo el supuesto de que H_0 es cierta. En este sentido, el estadístico $(p_1 - p_2)$, bajo H_0 y en el caso de muestras grandes, tiene una distribución aproximadamente normal, con media y varianza dadas por:

Media: $E(p_1 - p_2) = P - P = 0$

Varianza: $V(p_1 - p_2) = \frac{P(1-P)}{n_1} + \frac{p(1-P)}{n_2} = P(1-p)\frac{1}{n_1} + \frac{1}{n_2}$

Es decir que:

$$\frac{(p_1 - p_2) - 0}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0;1) \qquad (5.6)$$

Dado que es grande, si sustituimos P por su estimador, que denotamos por p, el resultado anterior se mantiene.

Ahora bien, como p es el estimador de la proporción común P de dos poblaciones, lo razonable es que el valor de p se obtenga combinado la información proveniente de las dos muestras.

Si tenemos presente que $p_1 = x_1/n_1$ y $p_2 = x_2/n_2$ siendo $\bar{X}_1 - \bar{X}_2$ variables binomiales que representan el número de éxitos en la muestra n_1 y n_2 respectivamente, entonces la forma más adecuada de definir p es la siguiente:

$$p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

Luego, asumiendo que H_0 es cierta, el estadístico

$$\frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

sigue una distribución aproximadamente N(0;1).

En resumen, para contrastar la hipótesis nula H_0:$P_1 = P_2$ contra cualquier hipótesis alternativa H_1 en el caso en que los tamaños de muestras sean grandes, se utiliza como estadístico de contraste:

$$Z_c = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

el cual, bajo H_0 sigue una distribución aproximadamente N(0;1).

Finalizamos señalando que la desviación estándar estimada del estadístico $p_1 - p_2$ que utiliza en el intervalo de confianza para $P_1 - P_2$ es diferente de la que se usa en el contraste de hipótesis, por cuanto en el caso de la estimación por intervalo no se sabe si $P_1 = P_2$

Resumiendo, los pasos a seguir en la contrastación de hipótesis son:

a. Plantear las hipótesis a contrastar

- H_0:$P_1 = P_2$
 H_1:$P < P_0$
- H_0:$P_1 = P_2$
 H_1:$P_1 > P_2$
- H_0:$P_1 = P_2$
 H_1:$P_1 \neq P_2$

b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar

$$Z_c = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0;1)$$

d. Determinar el valor y la región críticos

Los valores críticos correspondientes se deben ubicar en la tabla de la distribución normal (figura 4.3) del capítulo 4.

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser cotejado con el valor crítico.

f. Decidir e interpretar
Al comparar el valor del estadístico debe cerciorarse si cae en la región de rechazo de la hipótesis nula; en caso contrario, se expresa que no se rechaza la hipótesis nula o que el contraste no es estadísticamente significativo.

Ejemplo 5.2:

Se desea averiguar si un programa de televisión tiene la misma preferencia entre hombres y mujeres de una región. Se selecciona una muestra aleatoria de tamaño 100 de la población de hombres y se encuentra que a 30 de ellos les gusta el programa; igualmente se seleccionan aleatoriamente 125 mujeres y se obtiene que a 40 de ellas les agrada el programa. Realizar el contraste de hipótesis adecuado utilizando $\alpha = 0{,}05$.

Solución:
Denotemos por P_1 y P_2 a la proporción de hombres y mujeres en sus respectivas poblaciones que les gusta el programa de TV. Si lo que se quiere averiguar es si P_1 y P_2 son iguales o no, las hipótesis a contrastar son:

a. Plantear las hipótesis a contrastar

$$H_0{:}P_1 = P_2 \text{ lo que es equivalente a } (P_1 - P_2 = 0)$$
$$H_1{:}P_1 \neq P_2 \text{ lo que es equivalente a } (P_1 - P_2 \neq 0)$$

b. Fijar el nivel de significación: $\alpha = 0{,}05$

c. Establecer el estadístico a utilizar

$$Z_c = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0;1)$$

d. Determinar el valor y la región críticos
Se trata de un contraste bilateral (dos colas) y, por lo tanto, para $\alpha = 0{,}05$ los valores se obtienen de la tabla normal $N(0;1)$.

$$Z_{\alpha/2} = Z_{0{,}025} = 1{,}96 \text{ y } Z_{0{,}025} = -1{,}96$$

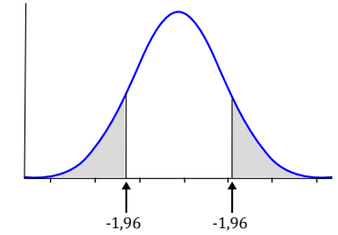


Figura 5.2 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso *c* para ser cotejado con el valor crítico.

Se tiene que:

$$p_1 = \frac{30}{100} = 0{,}30 \qquad p_2 = \frac{40}{125} = 0{,}32$$
$$p = \frac{100(0{,}3) + 125(0{,}32)}{100 + 125} = 0{,}31$$
$$Z_c = \frac{0{,}30 - 0{,}32}{\sqrt{0{,}31(0{,}69)\left(\frac{1}{100} + \frac{1}{125}\right)}} = -0{,}32$$

f. Decidir e interpretar
El criterio de decisión es rechazar H_0 si $Z_c \leq -1{,}96$o o $Z_c \geq -1{,}96$.

Dado que Z_C está comprendido entre -1,96 y 1,96, no se rechaza H_0 con $\alpha = 0{,}05$. No hay evidencia suficiente en las muestras en contra de la presunción de que el programa de televisión es igualmente preferido por los hombres y las mujeres de la región.

Ejercicios

1. Se selecciona una muestra aleatoria de 200 votantes y se encuentra que 114 apoyan un convenio de anexión.
   i. Encuentre el intervalo de confianza de 96 % para la fracción de la población votante que favorece el convenio.
   ii. ¿Qué podemos asegurar con 96 % de confianza acerca de la posible magnitud de nuestro error si estimamos que la fracción de votantes que favorece la anexión es de 0,57?

2. En una muestra aleatoria de 1.000 casas en cierta ciudad, 228 se calientan con petróleo. Encuentre el intervalo de confianza de 99 % para la proporción de casas que se calientan con petróleo.

3. Calcule un intervalo de confianza de 98 % para la proporción de artículos defectuosos en un proceso cuando una muestra de tamaño 100 arroja 8 defectuosos.

4. En el estudio *Germination and Emergence of Broccoli*, que lleva a cabo el departamento de horticultura de un instituto universitario, un investigador encuentra que a 5 °C, 10 semillas de 20 germinaron, mientras que a 15 °C, 15 de 20 semillas germinaron. Calcule el intervalo de confianza de 952 % para la diferencia entre la proporción de germinación en las dos diferentes temperaturas y decida si hay una diferencia significativa.

5. Un fabricante de reproductores de música y video utiliza un conjunto de pruebas amplias para evaluar la función eléctrica de su producto. La totalidad de los reproductores debe pasar todas las pruebas antes de venderse. Una muestra aleatoria de 500 reproductores tiene como resultado 15 que fallan en una o más pruebas. Encuentre un intervalo de confianza de 90 % para la proporción de los reproductores de música y video que pasan todas las pruebas.

6. Un experto en mercadotecnia de una compañía fabricante de pasta considera que 40 % de los amantes de la pasta prefieren la lasaña. Si 9 de 20 amantes de la pasta eligen lasaña sobre otras pastas, ¿qué se puede concluir acerca de la afirmación del experto? Utilice un nivel de significancia de 0,05.

7. Se cree que al menos 60 % de los residentes de cierta área favorece una demanda de anexión de una ciudad vecina. ¿Qué conclusión extraería si solo 110 en una muestra de 200 votantes está a favor de la demanda? Utilice un nivel de significancia de 0,05.

8. Una compañía petrolera afirma que 1/5 de las casas en cierta ciudad se calienta con petróleo. ¿Tenemos razón en dudar de esta afirmación si en una muestra aleatoria de 1.000 casas en esta ciudad, 136 se calientan con petróleo? Utilice un nivel de significancia de 0,01.

9. En una investigación se estima que 25 % de los estudiantes van en bicicleta a la escuela. ¿Esta parece ser una estimación válida si en una muestra aleatoria de 90 estudiantes universitarios, 28 van en bicicleta a la escuela? Utilice un nivel de significancia de 0,05.

10. En un estudio para estimar la proporción de residentes de cierta ciudad y sus suburbios que están a favor de la construcción de una planta de energía nuclear, 63 de 100 residentes urbanos están a favor de la construcción, mientras que solo 59 de 125 residentes suburbanos la favorecen. ¿Hay una diferencia significativa entre la proporción de residentes urbanos y de suburbanos que favorecen la construcción de la planta nuclear? Use un valor de significación de 5 %.

11. Se considera un nuevo dispositivo de radar para cierto sistema de misiles de defensa. El sistema se verifica mediante la experimentación con aeronaves reales en las que se simula una situación de *baja* o *no baja.* Si en 300 pruebas ocurren 250 muertes, no rechace o rechace, en un nivel de significancia de 0,04, la afirmación de que la probabilidad de una *baja* con el sistema nuevo no excede la probabilidad de 0,8 del sistema existente.

6. Inferencia acerca de una y dos varianzas poblacionales

INFERENCIAS ACERCA DE UNA Y DOS VARIANZAS POBLACIONALES

En este capítulo se considera la estimación puntual, por intervalos y contrastación de hipótesis en aquellos casos en los cuales se desea estimar una sola varianza / desviación estándar y el cociente entre dos varianzas poblacionales. Para los primeros casos se hace uso de la distribución chi-cuadrado y para los segundos, la distribución F de Snedecor.

6.1 Inferencia estadística con respecto a una varianza (σ^2) o desviación estándar poblacional (σ)

Se han analizado los procedimientos a seguir para construir intervalos de confianza y prueba de hipótesis referentes a una o dos medias poblacionales. Sin embargo, existen situaciones prácticas donde se está interesado en hacer inferencia sobre la variabilidad de una población. Por lo tanto, en este capítulo se estudia cómo hacer inferencia (estimación y prueba de hipótesis) acerca de la varianza o desviación estándar de una población, usando la distribución chi-cuadrado.

6.2 Estimación para la varianza (o desviación estándar) poblacional

Se ha visto que la varianza de la población (σ^2) es un número que cuantifica la cantidad de variabilidad de una población. Muchas veces se desconoce el valor real de σ^2 y, por esta razón, se requiere estimarlo utilizando una estimación puntual o una estimación por intervalo.

1. Estimación puntual

El estadístico muestral $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ es un estimador insesgado para σ^2. Por ello se utiliza S^2 como un estimador puntual del parámetro poblacional σ^2 o S como un estimador puntual del parámetro poblacional σ.

2. Estimación por intervalo

Dada una muestra aleatoria de tamaño pequeño seleccionada de una población normal o aproximadamente normal, con una media μ y una varianza σ^2, ambas desconocidas, el estadístico a partir del cual se obtiene[2] un intervalo de confianza para σ^2 es:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \qquad (6.1)$$

que tiene distribución χ^2 con (n - 1) grados de libertad.

El intervalo de confianza para σ^2 se deduce a partir de la figura 6.1.

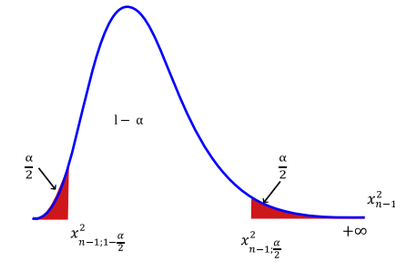


Figura 6.1 Gráfico de una distribución chi-cuadrado
Fuente: MLB

Con base en la figura 6.1 se puede deducir que un valor cualquiera de la distribución, digamos χ^2_{n-1}, se encuentra en el intervalo $\left(\chi^2_{n-1,1-\frac{\alpha}{2}};\ \chi^2_{n-1;\frac{\alpha}{2}}\right)$. Con un $(1-\alpha)$ % se tiene que:

$$P_r\left(\chi^2_{n-1,1-\frac{\alpha}{2}} < \chi^2_{n-1} < \chi^2_{n-1;\frac{\alpha}{2}}\right) = 1-\alpha \qquad (6.2)$$

donde:
$\chi^2_{n-1,1-\alpha/2}$ es un valor de la distribución chi-cuadrado que deja un área de $(1-\alpha)$ a la izquierda de la curva.

$\chi^2_{n-1,\alpha/2}$ es un valor de la distribución chi-cuadrado que deja un área de $\alpha/2$ a la derecha de la curva.

Sustituyendo χ^2_{n-1} por $\frac{(n-1)S^2}{\sigma^2}$ en la ecuación (6.2.) nos queda:

$$P_r\left(\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right) = 1-\alpha$$

Esta afirmación de probabilidad puede expresarse como:

$$P_r\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1,\ \alpha/2}}\right) = 1-\alpha$$

Multiplicando por $(n-1)S^2$ cada término de la desigualdad se obtiene:

$$P_r\left[\frac{(n-1)S^2}{\chi^2_{n-1,\ \alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right] = 1-\alpha$$

Para una muestra aleatoria de tamaño pequeño (n < 100), se calcula la varianza de la muestra S^2, y el intervalo de confianza de $1 - \alpha/2$ para σ^2 está dado por:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\ \alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right) \qquad (6.3)$$

Si calculamos la raíz cuadrada de cada uno de los tres términos de esta doble desigualdad se obtiene un intervalo de confianza para la desviación típica de la población σ^2.

A continuación, se presenta un ejemplo para estimar la varianza y la desviación estándar poblacional.

Ejemplo 6.1:

En el semestre 2014-1 se selecciona una muestra aleatoria de 10 estudiantes de la asignatura Estadística II en la Facultad de Economía de una prestigiosa universidad y se mide el tiempo que tardan en terminar el primer examen parcial. Los resultados en minutos son los siguientes:

120, 100, 90, 120, 80, 70, 98, 60, 90, 105

Se pide que:

i. Obtenga la varianza muestral.
ii. Calcule e interprete un intervalo de confianza de 90 % para la verdadera varianza del tiempo que tardan los estudiantes en terminar el examen.
iii. Calcule e interprete un intervalo de 90 % para la desviación estándar poblacional.

Solución:

i. Varianza muestral (S^2)
Para obtener la varianza muestral, se tiene que calcular, en primer lugar, la media muestral:

$$\bar{X} = \frac{\sum X_i}{n} = 93{,}3 \text{ minutos}$$

Por lo tanto, la varianza muestral de los valores dados es:

$$S^2 = \frac{\sum X_i^2 - n\,\bar{X}^2}{n-1} = \frac{90.529 - 10\,(93{,}3)^2}{9} = 386{,}68 \text{ minutos}^2$$

ii. Como la muestra es aleatoria de tamaño pequeño (n < 100) y se considera que el tiempo que tardan los alumnos en terminar el examen se distribuye normalmente, el intervalo de confianza para σ^2 sería:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\ \alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$$

Como $\upsilon = 10 - 1 = 9$ grados de libertad y $\alpha = 0{,}10$ se localiza en la tabla del anexo C, el valor de:

$$\chi^2_{n-1,\ \frac{\alpha}{2}} = \chi^2_{9,\frac{0{,}10}{2}} = \chi^2_{9,0{,}05} = 16{,}9190 \ y$$

$$\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{9,\ 1-0.05} = \chi^2_{9,0{,}95} = 3{,}32511$$

Al sustituir estos valores junto con n = 10 y $S^2 = 386{,}68$ en la fórmula del intervalo de confianza para σ^2 se obtiene:

$$\left(\frac{9\,(386{,}68)}{16{,}9190} < \sigma^2 < \frac{9\,(386{,}68)}{3{,}32511}\right)$$

$$(205{,}69 < \sigma^2 < 1.046{,}62)\,minutos^2$$

Interpretación:

Se espera con un 90 % de confianza que la verdadera varianza del tiempo que tardan los estudiantes en terminar el examen se encuentre entre 205,69 y 1.046,62 minutos^2.

iii. El intervalo de confianza de 90 % para σ sería:

$$\left(\sqrt{205{,}69} < \sigma < \sqrt{1.046{,}62}\right)$$

$$(14{,}34 < \sigma < 32{,}25)\ minutos$$

Interpretación:

La desviación estándar poblacional(σ) del tiempo que tardan los estudiantes en resolver el examen se encuentra entre 14,34 y 32,35 minutos, con una confianza de 90 %.

3. Contrastación de hipótesis

Cuando el tamaño de la muestra es pequeño y la población de la que se selecciona la muestra es normal o aproximadamente normal, se utiliza la distribución chi-cuadrado para probar hipótesis con respecto a una varianza o desviación estándar poblacional. Para tal fin se aplica el siguiente procedimiento:

a. Plantear las hipótesis a contrastar α
Se plantea la hipótesis nula y la alternativa. Se debe escoger una de las siguientes opciones:

$$H_0: \sigma^2 = \sigma_0^2 \quad H_0: \sigma^2 = \sigma_0^2 \quad H_0: \sigma^2 = \sigma_0^2$$
$$H_1: \sigma^2 > \sigma_0^2 \quad H_1: \sigma^2 < \sigma_0^2 \quad H_1: \sigma^2 \neq \sigma_0^2$$

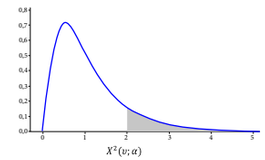
b. Fijar el nivel de significación:

c. Establecer el estadístico a utilizar
La distribución a utilizar en este caso es chi-cuadrado y el estadístico apropiado bajo la hipótesis nula $\sigma^2 = \sigma_0^2$ es el siguiente:

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2} \quad (6.4)$$

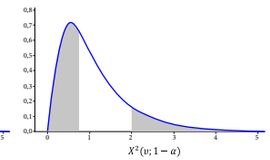
que tiene distribución chi-cuadrado con n – 1 grados de libertad.

d. Determinar el valor y la región críticos



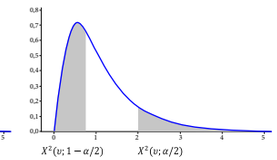
$H_1: \sigma^2 > \sigma_0^2$

Figura 6.2. *a*



$H_1: \sigma^2 < \sigma_0^2$

Figura 6.2. *b*



$H_1: \sigma^2 \neq \sigma_0^2$

Figura 6.2. *c*

Figura 6.2 Regiones y valores críticos para el contraste de hipótesis
Fuente: MLB

e. Calcular el estadístico
En este paso se calcula el estadístico fijado en el paso (iii) para ser comparado con el valor crítico.

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

f. Decidir e interpretar
Si el valor calculado del estadístico de prueba cae en la región crítica, entonces se rechaza la hipótesis nula y se acepta lo planteado en la hipótesis alternativa al nivel de significación α. Cuando el valor calculado cae en la región de aceptación, se decide no rechazar H_0, al nivel de significación α . Se concluye de acuerdo con lo planteado en el problema.

Ejemplo 6.2:

Suponga que a nivel nacional se considera que la variable gasto mensual estudiantil se distribuye normalmente X~N ($\mu = 8.000\$, \sigma = 50\$$). Se selecciona aleatoriamente un grupo de 30 estudiantes de la UPB y la varianza mensual resultante es de $S^2 = 4.000\$^2$ ¿Tienen los gastos mensuales de los estudiantes de esta universidad una varianza significativamente mayor que la de todos los estudiantes universitarios del país? Haga la prueba con un nivel de significación de 5 % (Nota: $\sigma^2 = 50^2 = 2.500$).

Solución:

a. Plantear las hipótesis a contrastar

$H_0 : \sigma^2 = 2.500$ (La varianza del gasto mensual de los estudiantes de la UPB es significativamente igual a la varianza del gasto mensual de todos los universitarios del país).

$H_1 : \sigma^2 > 2.500$ (La varianza del gasto mensual de los estudiantes de la UPB es significativamente mayor a la varianza del gasto mensual de todos los universitarios del país).

b. Fijar el nivel de significación: $\alpha = 0{,}05$

c. Establecer el estadístico a utilizar
Como el tamaño de la muestra es pequeño, seleccionada de una población normal, el estadístico apropiado sería:

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

d. Determinar el valor y la región críticos (ver figura 6.3)

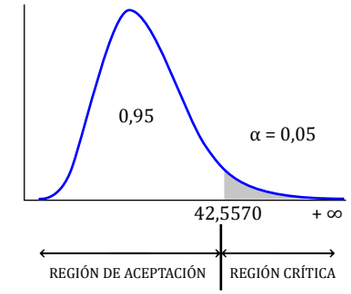


Figura 6.3 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(29)(4.000)}{2.500} = \frac{116.000}{2.500} = 46{,}40$$

f. Decidir e interpretar
Como el valor calculado del estadístico de prueba cae dentro de la región crítica, se toma la decisión de rechazar la hipótesis nula. Se comprueba que los gastos mensuales de los estudiantes de la UPB tienen una varianza significativamente mayor que la varianza presentada a nivel nacional. Esta prueba se ha realizado con un nivel de significación de 5 %.

Ejemplo 6.3:

Por investigaciones anteriores, se conoce que el aumento de peso en kilos para las razas comunes de ovejas durante los primeros 6 meses de vida se distribuye según una normal X~N ($\mu = 20, \sigma = 6$). En un estudio actual se sugiere que para cierta raza de ovejas, el aumento de peso en ese periodo de tiempo es más uniforme (es decir, su varianza es menor de 36). Se toma una muestra aleatoria de 10 ovejas; sus aumentos de peso durante los primeros 6 meses de vida fueron: 15, 10, 20, 30, 25, 17,

23, 19, 21, 20 kilos. Utilizando $\alpha = 0{,}10$ ¿puede llegarse a la conclusión de que la varianza en el aumento de peso en esta raza de ovejas es menor que la de las razas comunes?

Solución:

a. Plantear las hipótesis a contrastar

$H_0 : \sigma^2 = 36$ (La varianza del aumento de peso de esa raza de ovejas durante los primeros seis meses de vida es significativamente igual al de las razas comunes).

$H_1 : \sigma^2 < 36$ (La varianza del aumento de peso de esa raza de ovejas durante los primeros seis meses de vida es significativamente menor al de las razas comunes).

b. Fijar el nivel de significación: $\alpha = 0{,}10$

c. Establecer el estadístico a utilizar
Como el tamaño de la muestra es pequeño, seleccionada de una población normal, el estadístico apropiado sería:

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

d. Determinar el valor y la región críticos

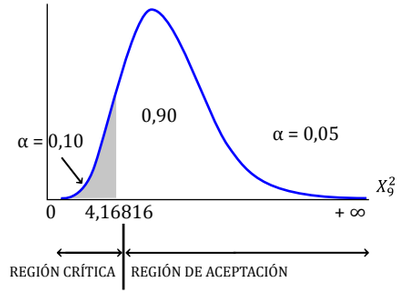


Figura 6.4 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico
En primer lugar, se debe conocer la media y la varianza para sustituirlo en la ecuación (6.4.).

Media:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{200}{10} = 20\ kilos$$

Varianza:

$$S^2 = \frac{\sum X_i^2 - n\,\bar{X}^2}{n-1} = \frac{4.270 - (10)(20)^2}{9} = \frac{270}{9} = 30$$

Sustituyendo en la ecuación (6.4) se tiene:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(9)(30)}{36} = \frac{270}{36} = 7{,}50$$

f. Decidir e interpretar
Como el valor calculado del estadístico de prueba cae en la región de aceptación, se decide no rechazar H_0. La muestra no ofrece suficiente evidencia para concluir que la varianza en el aumento de peso de esa raza de ovejas es menor que la de las razas comunes. Esta prueba se ha realizado con un nivel de significación de 10 %.

Ejemplo 6.4:

El gerente de control de calidad de una fábrica de bombillas necesita conocer si la desviación estándar de la vida de las bombillas de un gran embarque ha cambiado de 100 horas. Si una muestra aleatoria de 15 bombillas señala una desviación estándar de 110 horas, al nivel de significación de 0,01, ¿existe evidencia en la muestra de que la desviación estándar ha cambiado? Suponga que la vida de las bombillas se distribuye normalmente.

Solución:

a. Plantear las hipótesis a contrastar

$H_0: \sigma = 100$ ó $\sigma^2 = 10.000$ (La desviación estándar de la vida de las bombillas no ha cambiado, es decir, es igual a 100).

$H_1: \sigma \neq 100$ ó $\sigma^2 \neq 10.000$ (La desviación estándar de la vida de las bombillas ha cambiado, es decir, es diferente a 100).

b. Fijar el nivel de significación: $\alpha = 0{,}01$

c. Establecer el estadístico a utilizar

$$\chi_c^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

d. Determinar el valor y la región críticos

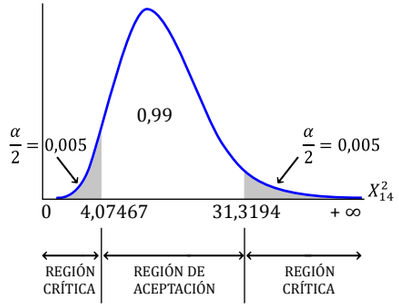


Figura 6.5 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico

$$\chi_c^2 = \frac{(15-1)(110)^2}{(100)^2} = 16{,}94$$

f. Decidir e interpretar
Como el valor del estadístico de prueba se encuentra en la zona de aceptación, se decide no rechazar H_0. La muestra no ofrece suficiente evidencia para que el gerente pueda llegar a la conclusión de que la desviación estándar ha cambiado, con un nivel de significación de 0,01.

1. Inferencia estadística con respecto a dos varianzas poblacionales

Muchas veces surge la necesidad de obtener información sobre la variabilidad de dos poblaciones. Esta información se puede adquirir por medio de la estimación o la prueba de hipótesis. En este apartado se presenta el procedimiento a seguir para la estimación y la prueba de hipótesis de dos varianzas poblacionales cuando las muestras son aleatorias e independientes, de tamaño pequeño, seleccionadas de poblaciones normales o aproximadamente normales.

Para obtener información sobre el valor desconocido de las varianzas de dos poblaciones se ha ideado un procedimiento estadístico basado en la razón de las dos varianzas muestrales, utilizando una estimación puntual o una estimación por intervalo.

a. Estimación puntual

La estimación puntual de la razón de dos varianzas poblacionales, digamos $(\frac{\sigma_1^2}{\sigma_2^2})$, está dada por el cociente de sus respectivas varianzas muestrales $(\frac{S_1^2}{S_2^2})$.

b. Estimación por intervalo

Para dos muestras aleatorias independientes de tamaño pequeño, provenientes de poblaciones chi-cuadrado, el estadístico a partir del cual se obtiene un intervalo de confianza para $(\frac{\sigma_1^2}{\sigma_2^2})$ es $(\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2})$.

Suponga que:

$$U \sim \chi^2_{(n_1-1)} \quad \text{y} \quad V \sim \chi^2_{(n_2-1)}$$

Por lo tanto:

$$\chi_1^2 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \quad \text{y} \quad \chi_2^2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \qquad (6.5)$$

Donde U y V son independientes, se tiene que el cociente:

$$\frac{\left(\frac{\chi_1^2}{(n_1-1)}\right)}{\left(\frac{\chi_2^2}{(n_2-1)}\right)} \qquad (6.6)$$

Si sustituimos χ_1^2 y χ_2^2 dada en (6.5) en (6.6):

$$\frac{\left(\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}}{(n_1-1)}\right)}{\left(\frac{\frac{(n_2-1)S_2^2}{\sigma_2^2}}{(n_2-1)}\right)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

que tiene una distribución F con $(n-1)$ y $(n-2)$ grados de libertad.

El intervalo de confianza se puede deducir a partir de la figura 6.6.

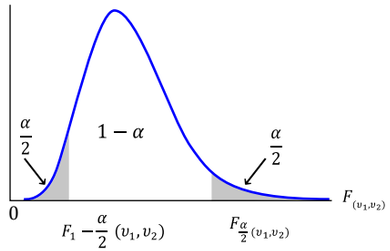


Figura 6.6 Distribución F
Fuente: MLB

Con base en la figura se puede describir:

$$P_r\left[F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)} < F_{(\upsilon_1,\upsilon_2)} < F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}\right] = 1-\alpha$$

Sustituyendo $F_{(\upsilon_1,\upsilon_2)}$ por $\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$, nos queda:

$$P_r\left[F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)} < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}\right] = 1-\alpha$$

Al multiplicar cada término de la desigualdad por $\frac{S_2^2}{S_1^2}$, se obtiene que:

$$P_r\left[\frac{S_2^2}{S_1^2} F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}\right] = 1-\alpha$$

Esta afirmación probabilística puede expresarse como:

$$P_r\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}}\right] = 1-\alpha$$

Invirtiendo los términos, nos queda:

$$P_r\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}}\right] = 1-\alpha$$

Como $F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}$ no se encuentra directamente en la tabla, hay que aplicar la propiedad recíproca:

$$F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)} \frac{1}{F_{\frac{\alpha}{2}(\upsilon_2,\upsilon_1)}}$$

Remplazando $F_{1-\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}$ por $\frac{1}{F_{\frac{\alpha}{2}(\upsilon_2,\upsilon_1)}}$ se tiene:

$$P_r\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}(\upsilon_1,\upsilon_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2}(\upsilon_2,\upsilon_1)}\right] = 1-\alpha$$

Cuando se seleccionan muestras aleatorias independientes de tamaño pequeño, provenientes de poblaciones normales, y calculamos las varianzas muestrales, el intervalo de confianza de $(1-\alpha)$ % para $\frac{\sigma_1^2}{\sigma_2^2}$ está dado por:

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}(v_1,v_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\ F_{\frac{\alpha}{2}(v_2,v_1)}\right) \qquad (6.7)$$

Ejemplo 6.5:

Un grupo de 25 varones y 16 mujeres presentan el mismo examen de estadística. Los varones obtienen una calificación promedio de 16 puntos y una desviación estándar de 8 puntos, mientras que las mujeres obtienen una calificación promedio de 12 puntos, con una desviación estándar de 6 puntos. Encuentre un intervalo de confianza de 98 % para el cociente de las varianzas poblacionales ($\frac{\sigma_1^2}{\sigma_2^2}$), donde σ_1^2 y σ_2^2 representan las varianzas de las poblaciones de calificaciones de varones y mujeres, respectivamente.

Solución:
Como las muestras son independientes, de tamaño pequeño y seleccionadas de poblaciones normales, el intervalo de confianza apropiado para $\frac{\sigma_1^2}{\sigma_2^2}$ es:

$$\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\frac{\alpha}{2}(v_1,v_2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\ F_{\frac{\alpha}{2}(v_2,v_1)}\right)$$

Donde:

$$\frac{S_1^2}{S_2^2} = \frac{64}{36} = 1{,}78$$
$$1-\alpha = 0{,}93$$
$$\alpha = 0{,}02$$
$$\frac{\alpha}{2} = 0{,}01$$
$$v_1 = n_1 - 1 = 25 - 1 = 24$$

$$v_2 = n_2 - 1 = 16 - 1 = 15$$
$$F_{\frac{\alpha}{2}(v_1,v_2)} = F_{0{,}01\,(24,15)} = 3{,}29$$
$$F_{\frac{\alpha}{2}(v_2,v_1)} = F_{0{,}01\,(15,24)} = 2{,}89$$

Sustituyendo estos valores en la fórmula del intervalo de confianza se obtiene:

$$\left(1{,}78\left(\frac{1}{3{,}29}\right) < \frac{\sigma_1^2}{\sigma_2^2} < 1{,}78\ (2{,}89)\right)$$
$$\left(0{,}54 < \frac{\sigma_1^2}{\sigma_2^2} < 5{,}14\right)$$

Interpretación:
Se tiene un 98 % de confianza de que la razón entre las variabilidades de las calificaciones de las poblaciones de varones y mujeres se encuentre en el intervalo hallado. Como el intervalo abarca el valor 1,0, esto indica que se tiene un 98 % de confianza de que σ_1^2 no es significativamente diferente de σ_2^2.

c. Contrastación de hipótesis

A veces es necesario demostrar si dos poblaciones tienen la misma varianza, bien para probar la suposición de igualdad de varianzas (cuando se hace inferencia acerca de la diferencia entre dos medias poblacionales, con muestras aleatorias independientes de tamaño pequeño), o bien para obtener información sobre las varianzas de dos poblaciones.

Para probar la hipótesis $\sigma_1^2 = \sigma_2^2$ mediante la distribución F se ha creado un procedimiento estadístico basado en la razón de las varianzas muestrales S_1^2/S_2^2, y se considera el grado en que la razón S_1^2/S_2^2 difiere de 1,0. Si se cumple que $\sigma_1^2 = \sigma_2^2$, cabría esperar que la razón S_1^2/S_2^2 tuviese un valor cercano a 1,0. Así, mientras mayor sea la discrepancia entre S_1^2/S_2^2 y 1,0, menor confianza se tendrá de que σ_1^2 sea significativamente igual a σ_2^2.

El procedimiento general para probar hipótesis con respecto a dos varianzas poblacionales se presenta a continuación:

a. Plantear las hipótesis a contrastar
Se plantea la hipótesis nula y la alternativa. Se debe escoger una de las siguientes opciones:

$$H_0 : {\sigma_1}^2 = \sigma_2^2 \quad H_0 : {\sigma_1}^2 = \sigma_2^2 \quad H_0 : {\sigma_1}^2 = \sigma_2^2$$
$$H_1 : {\sigma_1}^2 > \sigma_2^2 \quad H_1 : {\sigma_1}^2 < \sigma_2^2 \quad H_1 : {\sigma_1}^2 \neq \sigma_2^2$$

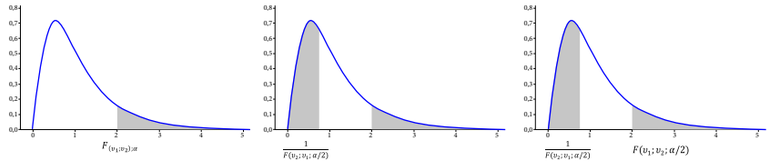
b. Fijar el nivel de significación: α

c. Establecer el estadístico a utilizar
Se determina el estadístico de prueba. La distribución a utilizar en este caso es la F de Snedecor y el estadístico apropiado bajo la hipótesis nula $\sigma_1^2 = \sigma_2^2$ es:

$$F_c = \frac{S_1^2}{S_2^2} \qquad (6.8)$$

que tiene una distribución F con $\upsilon_1 = n_1 - 1$ y $\upsilon_2 = n_2 - 1$.

d. Derminar el valor y región críticos (figura 6.7)



$H_1 : {\sigma_1}^2 > \sigma_2^2$ $\quad$ $H_1 : {\sigma_1}^2 < \sigma_2^2$ $\quad$ $H_1 : {\sigma_1}^2 \neq \sigma_2^2$

Figura 6.2. *a* **Figura 6.2.** *b* **Figura 6.2.** *c*

Figura 6.7 Regiones y valores críticos para el contraste de hipótesis
Fuente: MLB

e. Calcular el estadístico

$$F_c = \frac{S_1^2}{S_2^2}$$

f. Decidir e interpretar
Si el valor calculado del estadístico de prueba cae en la región de rechazo, entonces se rechaza la hipótesis nula y se acepta lo planteado en la hipótesis alternativa, al nivel de significación α. Cuando el valor calculado cae en la región de aceptación, se decide no rechazar H_0 al nivel de significación α. Se interpreta o se concluye de acuerdo con lo planteado en el problema.

Ejemplo 6.6:

Un gerente de producción piensa que la productividad promedio de los empleados expertos es mayor que la de los nuevos, pero no espera que la varianza de ella difiera en los 2 grupos. En un grupo de 16 empleados nuevos, la producción unitaria promedio es de 20 unidades por hora, con una varianza de 56 unidades al cuadrado. En un grupo de 21 empleados con 5 años de experiencia, la producción promedio en el mismo tipo de trabajo es de 30 unidades por hora, con una varianza muestral de 28 unidades al cuadrado. ¿Hay evidencia suficiente en los datos para concluir que existe una diferencia en la varianza de la producción en los 2 niveles de experiencia? Use un nivel de significación de 0,05.

Solución:

a. Plantear las hipótesis a contrastar

$H_0 : {\sigma_1}^2 = \sigma_2^2$ (La varianza de la producción en el primer nivel de experiencia es significativamente igual a la varianza del segundo nivel).
$H_1 : {\sigma_1}^2 \neq \sigma_2^2$ (La varianza de la producción en el primer nivel de experiencia difiere significativamente de la varianza del segundo nivel).

b. Fijar el nivel de significación: $\alpha = 0{,}05$

c. Establecer el estadístico a utilizar
Como las muestras son aleatorias, independientes, de tamaño pequeño y fueron seleccionadas de poblaciones normales, el estadístico apropiado sería:

$$F_c = \frac{S_1^2}{S_2^2}$$

d. Determinar el valor y la región críticos

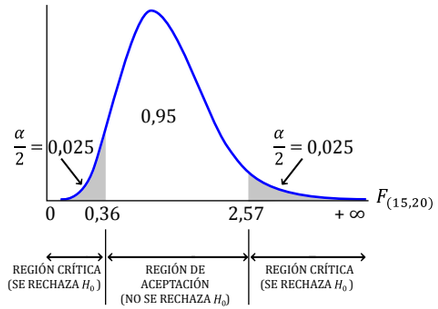


Figura 6.8 Valor y región críticos
Fuente: MLB

$$F_{1-\frac{\alpha}{2};(v_1,v_2)} = \frac{1}{F_{\frac{\alpha}{2};(v_2,v_1)}}$$

$$F_{0,975;\,(15,20)} = \frac{1}{F_{0,25;\,(20,15)}} = \frac{1}{2,76} = 0,36$$

$$F_{\frac{\alpha}{2}(v_1,v_2)} = F_{0,025\,(15,20)} = 2,57$$

e. Calcular el estadístico

$$F_c = \frac{56}{28} = 2$$

f. Decidir e interpretar
Como el valor calculado cae en la región de aceptación, entonces no se rechaza H_0. La varianza de la producción del primer nivel no es significativamente diferente a la del segundo nivel con un $\alpha = 0,05$.

Ejemplo 6.7:

El supervisor de control de calidad de una compañía está interesado en determinar si la varianza en la longitud de los tornillos producidos por la máquina 1 es mayor a la varianza de la longitud de los tornillos producidos por la máquina 2. Estas máquinas están diseñadas para producir tornillos idénticos, los cuales deben medir 3 cm de longitud. Sin embargo, debido a ciertos factores implicados en el proceso de producción, las longitudes de los tornillos varían ligeramente de la longitud establecida. Se seleccionan 2 muestras aleatorias de 13 y 16 tornillos de las máquinas 1 y 2, obteniéndose las siguientes varianzas: $S_1^2 = 0,6$ *y* $S_2^2 = 0,4$. Supóngase que las 2 poblaciones tienen distribución normal. Con un nivel de significación de 0,05, realice la prueba de hipótesis de interés para el supervisor.

Solución:

a. Plantear las hipótesis a contrastar

$H_0 : {\sigma_1}^2 = \sigma_2^2$ (La varianza en la longitud de los tornillos producidos por la máquina 1 es significativamente igual a la de los tornillos producidos por la máquina 2).
$H_1 : {\sigma_1}^2 > \sigma_2^2$ (La varianza en la longitud de los tornillos producidos por la máquina 1 es significativamente mayor a la de los tornillos producidos por la máquina 2).

b. Fijar el nivel de significación: $\alpha = 0,05$

c. Establecer el estadístico a utilizar
Como las muestras son aleatorias, independientes, de tamaño pequeño y fueron seleccionadas de poblaciones normales, el estadístico apropiado sería:

$$F_c = \frac{S_1^2}{S_2^2}$$

d. Determinar el valor y la región críticos

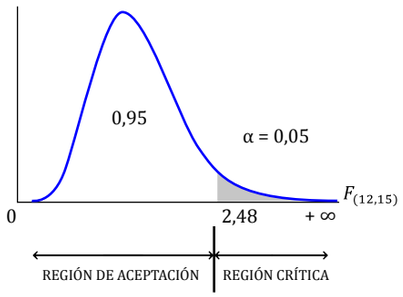


Figura 6.9 Valor y región críticos
Fuente: MLB

$$v_1 = n_1 - 1 = 13 - 1 = 12$$
$$v_2 = n_2 - 1 = 16 - 1 = 15$$
$$F_{\alpha(v_1,v_2)} = F_{0,05\ (12,15)} = 2,48$$

e. Calcular el estadístico

$$F_c = \frac{S_1^2}{S_2^2} = \frac{0,6}{0,4} = 1,5$$

f. Decidir e interpretar
Como el valor calculado cae dentro de la región de aceptación, entonces no se rechaza H_0. No existe evidencia en la muestra para concluir que la varianza en la longitud de los tornillos producidos por la máquina 1 es significativamente mayor a la varianza en la longitud de los tornillos producidos por la máquina 2 con α = 0,05.

Ejercicios

1. Un fabricante de baterías para automóvil afirma que sus baterías durarán, en promedio, 3 años, con una varianza de 1 año. Si 5 de estas baterías tienen duraciones de 1,9, 2,4, 3,0, 3,5 y 4,2 años, construya un intervalo de confianza de 95 % para σ^2 y decida si la afirmación del fabricante de que $\sigma^2 = 1$ es válida. Suponga que la población de duraciones de las baterías se distribuye de forma aproximadamente normal.

2. Una muestra aleatoria de 20 estudiantes obtiene una media de $\bar{x}_1 = 72$ y una varianza de $s^2 = 16$ en un examen de colocación en matemáticas. Suponga que las calificaciones se distribuyen normalmente y construya un intervalo de confianza de 98 % para σ^2.

3. La experiencia anterior indica que el tiempo que se requiere para que los estudiantes de último año de preparatoria completen una prueba estandarizada es una variable aleatoria normal con una desviación estándar de 6 minutos. Pruebe la hipótesis de que $\sigma = 6$ contra la alternativa de que $\sigma < 6$ si una muestra aleatoria de 20 estudiantes de último año de preparatoria tiene una desviación estándar $s = 4.51$. Utilice un nivel de significancia de 0,05.

4. Se deben supervisar las aflatoxinas producidas por moho en cosechas de cacahuate en cierta región. Una muestra de 64 lotes de cacahuate revela niveles de 24,17 ppm, en promedio, con una varianza de 4,25 ppm. Pruebe la hipótesis de que $\sigma^2 = 4,2$ ppm con la alternativa de que $\sigma^2 \neq 4,2$ ppm. Utilice un nivel de significancia de 0,05.

5. Algunos datos históricos indican que la cantidad de dinero con que contribuyeron los residentes trabajadores de una ciudad grande a un escuadrón de rescate voluntario es una variable aleatoria normal con una desviación estándar de $1,40, es decir, 4.200 pesos aproximadamente. Se sugiere que las contribuciones al escuadrón de rescate de los empleados del departamento de sanidad son mucho más variables. Si las contribuciones de una muestra aleatoria de 12 empleados del departamento de sanidad tienen una desviación estándar de $1,75, ¿podemos concluir con un nivel de significancia

de 0,01 que la desviación estándar de las contribuciones de los trabajadores de sanidad es mayor que la de todos los trabajadores que viven en esta ciudad?

6. Se dice que una máquina despachadora de refrescos está fuera de control si la varianza de los contenidos excede 1,15 decilitros. Si una muestra aleatoria de 25 bebidas de esta máquina tiene una varianza de 2,03 decilitros, ¿lo anterior indica con un nivel de significancia de 0,05 que la máquina está fuera de control? Suponga que los contenidos se distribuyen aproximadamente normal.

7. Se comparan dos tipos de instrumentos para medir la cantidad de monóxido de azufre en la atmósfera en un experimento de contaminación del aire. Se desea determinar si los dos tipos de instrumentos dan mediciones con la misma variabilidad. Se registran las siguientes lecturas para los dos instrumentos:

Tabla 1. Mediciones

Monóxido de azufre	
Instrumento A	**Instrumento B**
0.86	0.87
0.82	0.74
0.75	0.63
0.61	0.55
0.89	0.76
0.64	0.7
0.81	0.69
0.68	0.57
0.65	0.53

Suponga que las poblaciones de mediciones se distribuyen de forma aproximadamente normal y pruebe la hipótesis de que $\sigma_A = \sigma_B$ contra la alternativa de que $\sigma_A \neq \sigma_B$. Use un valor $\alpha = 0{,}10$

8. Construya un intervalo de confianza de 99 % para la desviación estándar usando la información del ejercicio 6 del capítulo 4.

9. Construya un intervalo de confianza de 98 % para $\left(\frac{\sigma_1^2}{\sigma_2^2}\right)$ usando la información del ejercicio 7 del capítulo 4, donde σ_1^2 y σ_2^2son, respectivamente, las desviaciones estándar para las distancias que se obtienen por litro de combustible en los camiones compactos Volkswagen y Toyota.

7. Análisis regresión lineal simple

ANÁLISIS REGRESIÓN LINEAL SIMPLE

El objetivo principal de algunas investigaciones estadísticas consiste en establecer la relación entre dos o más variables. Para estudiar la relación existente entre ellas se utilizan dos técnicas: el análisis de regresión y el análisis de correlación.

El análisis de regresión se utiliza para fines de predicción o descripción. Su objetivo principal es desarrollar una ecuación de predicción, es decir, una fórmula matemática que se pueda usar para predecir los valores de una variable dependiente o de respuesta, basada en los valores de otra u otras variables independientes o explicatorias. Por ejemplo, se podría estar interesado en predecir las ventas de un nuevo producto para el próximo mes en términos de su precio. En este ejemplo, las ventas representan la variable dependiente y el precio sería la variable independiente.

En otros casos, la regresión puede ser utilizada para describir la relación entre algunos valores conocidos de dos o más variables. Por ejemplo, se puede estudiar la relación entre el consumo y los niveles pasado y presente de ingreso.

Sería ideal si se pudiera predecir o describir los valores exactos de una variable dependiente en términos de cualquier otra u otras variables independientes, pero rara vez es posible, ya que pueden existir muchos elementos que causen variaciones en la variable dependiente para un conjunto dado de valores de la variable o las variables independientes. A causa de estas posibles variaciones, nuestro interés consistirá en predecir o describir el valor promedio de una variable en términos del valor conocido de otra u otras variables. Por ejemplo, no se puede predecir con exactitud los valores de las ventas de un nuevo producto para el próximo mes en términos de su precio, pero con datos apropiados se

puede predecir las ventas promedio del nuevo producto para el próximo mes en términos de su precio. De igual forma, se puede predecir la calificación promedio de los alumnos que inician un programa de maestría en Administración en términos de su calificación promedio en pregrado.

El término *regresión* surgió de estudios de la herencia biológica realizados por Francis Galton (1822 -1911) a fines del siglo XIX. En su conocida experiencia, Galton notó que los padres altos tenían hijos cuya altura era mayor a la altura promedio, pero no eran más altos que sus padres. También, padres bajos tenían hijos con altura menor a la altura promedio, pero eran más altos que sus padres. Esta tendencia de las características de los grupos de moverse, en la siguiente generación, hacia el promedio de la población o de regresión hacia la media fue descubierta por Galton. El término no tiene hoy el mismo significado que le dio Galton, pero se usa extensamente para referirse al estudio de relaciones funcionales entre variables cuando hay un componente aleatorio involucrado.

El análisis de regresión puede incluir una o más variables independientes. Si se estima el valor de la variable dependiente con base en una variable independiente, el análisis de regresión se denomina simple, mientras que si se estima el valor de la variable dependiente con base en dos o más variables independientes, el análisis de regresión se denomina múltiple. En este capítulo nos referimos al modelo de regresión simple.

La naturaleza de la relación existente entre las variables puede adoptar muchas formas, desde funciones matemáticas muy sencillas hasta otras complicadas. La más sencilla consiste en una relación en línea recta o relación lineal. En este capítulo se estudia el modelo de regresión lineal simple. El análisis de correlación se utiliza para medir la fuerza de la relación entre las variables. Su objetivo no es usar una o más variables para predecir otra, sino más bien medir la fuerza de la asociación entre las variables de interés. En este capítulo solo se describe el análisis de correlación lineal simple.

7.1 El diagrama de dispersión

Los gráficos de dispersión son útiles en la etapa exploratoria, tanto en el análisis de regresión como de correlación. La representación gráfica de los datos es frecuentemente el punto de partida de cualquier análisis que involucra más de una variable.

Los gráficos de dispersión muestran una nube de puntos, y cada punto representa una observación. Por eso, el primer paso para determinar si existe una relación entre dos variables consiste en elaborar y examinar el diagrama de dispersión de los datos, que es una gráfica en la que se traza cada uno de los puntos que representa un par de valores observados, para las variables independiente y dependiente.

Para elaborar el diagrama de dispersión se procede de la siguiente manera:

a. Se determina cuál es la variable independiente y cuál es la variable dependiente.
b. En un sistema de coordenadas cartesianas, en el eje de las abscisas se ubican los valores de la variable independiente (X) y en el eje de las ordenadas se colocan los valores de la variable dependiente o variable respuesta (Y).
c. Se coloca un punto en el plano por cada par de valores (X_i, Y_i).
d. El patrón de puntos obtenidos se denomina diagrama de dispersión.

Una vez elaborado el diagrama de dispersión se puede observar visualmente si las variables están relacionadas. De ser cierto, se puede tener una idea aproximada de qué clase de línea o ecuación de estimación describe mejor dicha relación. A continuación, se presenta un ejemplo que ilustra el procedimiento a seguir para elaborar el diagrama de dispersión.

Ejemplo 7.1:

Una cadena de almacenes ha tenido grandes fluctuaciones en sus ingresos durante los últimos años. Muchas ofertas, nuevos productos y técnicas de publicidad se han utilizado durante este tiempo, por lo cual es difícil determinar cuáles de estos factores han tenido mayor influencia en las ventas. El departamento de mercadotecnia ha estudiado varias

relaciones y piensa que los gastos mensuales destinados a la publicidad por la televisión pueden ser significativos. Se seleccionó una muestra aleatoria de siete meses con los resultados que se muestran en la tabla 7.1. Elabore el diagrama de dispersión y determine si un análisis de regresión lineal es apropiado.

Tabla 7.1 Gastos publicitarios e ingreso por ventas de una cadena de almacenes

Gastos mensuales en publicidad por televisión ($ 1.000)	Ingreso mensual por ventas ($ 100.000)
X_i	Y_i
25	34
16	14
42	48
34	32
10	26
21	29
19	20

Fuente: MLB

Solución:
Para elaborar el diagrama de dispersión se debe, en primer lugar, determinar la variable independiente y la dependiente. Puesto que se desea emplear los gastos mensuales en publicidad por televisión para predecir los ingresos mensuales por ventas, la variable independiente serían los gastos mensuales y la dependiente, los ingresos mensuales por ventas. En segundo lugar, en un sistema de coordenadas cartesianas en el eje horizontal o eje *X* se localizan los valores de la variable independiente (gastos mensuales) y en el eje vertical o eje *Y*, los valores de la variable dependiente (ingreso mensual por ventas), ubicando un punto por cada par (X_i, Y_i). En la figura 7.1 se observa el diagrama de dispersión terminado.

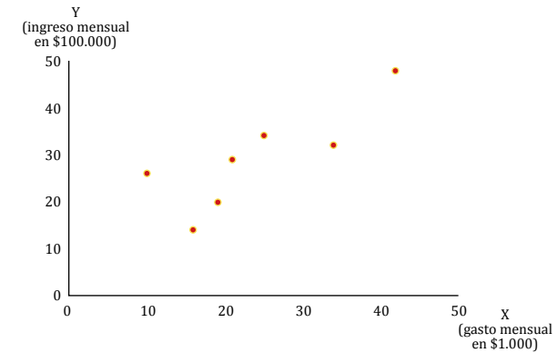


Figura 7.1 Diagrama de dispersión
Fuente: MLB

En términos generales, los puntos trazados siguen una relación lineal, por lo tanto, podemos trazar o ajustar una recta a través del diagrama para representar la relación entre las variables ingreso mensual y gasto mensual. Se observa, además, que a medida que aumenta el gasto mensual, aumenta el ingreso mensual por venta. Por tal motivo, se espera una relación lineal directa entre dichas variables.

En la figura 7.2 se ilustran algunos tipos de patrones que se podrían encontrar en un diagrama de dispersión.

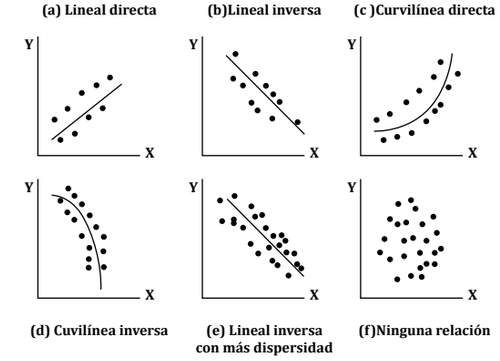


Figura 7.2 Posibles relaciones entre X y Y en los diagramas de dispersión
Fuente: Cari, 2011

Las gráficas *a* y *b* muestran relaciones lineales de tipo directo e inverso. Las gráficas *c* y *d* son ejemplos de relaciones curvilíneas que evidencian asociaciones directas e inversas entre variables, respectivamente. La gráfica *e* ilustra una relación lineal inversa con un patrón ampliamente disperso de puntos. Esta mayor dispersión revela que hay menor grado de asociación entre la variable independiente y la dependiente que en la gráfica *b*. El patrón de puntos en la gráfica *f* parece indicar que no existe relación entre las dos variables; por tanto, el conocimiento del pasado de una variable no permite predecir las ocurrencias futuras de la otra.

7.2 Modelo de regresión lineal simple poblacional

El modelo de regresión lineal simple se emplea en aquellas situaciones en las cuales nos interesa estudiar la relación que existe entre dos variables, una de las cuales es la variable dependiente (*Y*) y la otra es la independiente (*X*), admitiendo que la relación poblacional promedio entre la variable dependiente y la variable independiente pueda ser representada por una línea recta.

Como se desea determinar el valor medio de *Y* para un valor dado de *X*, nuestro interés es obtener μ_{y/x_i}, que representa el valor medio de *Y* para un valor dado X_i. La ecuación que representa la relación poblacional lineal entre X_i y la media de *Y* se llama recta de regresión poblacional, la cual puede escribirse de la siguiente manera:

$$\mu_{y/x_i} = \beta_0 + \beta_1 X_i \qquad (7.1)$$

Donde:

β_0 y β_1: Son constantes, llamados coeficientes de regresión.

β_0: Es la ordenada en el origen. Representa el valor medio de *Y* cuando *X* es igual a cero.

β_0: Es la pendiente de la recta de regresión poblacional. Representa el cambio medio en *Y* (aumento o disminución) por un incremento unitario particular en *X*.

Además del valor medio μ_{y/x_i}, puede que se necesite obtener información acerca de un valor particular de *Y* para un valor dado de *X*. Este valor particular se denota por Y_i.

Un valor observado de Y_i para el valor dado de X_i, por lo general, no es igual al valor medio de μ_{y/x_i}. La diferencia entre μ_{y/x_i} y es el elemento imprevisible en el análisis de regresión, conocido también como perturbación aleatoria; este puede tomar valores positivos (Y_i se pone por encima del valor medio), valores negativos (Y_i se pone por debajo del valor medio) y nulos (si son iguales). Es decir:

$$\varepsilon_i = Y_i - \mu_{y/x_i}$$

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) \qquad (7.2)$$

Despejando Y_i en la expresión (7.2) se obtiene la ecuación del modelo de regresión lineal simple poblacional.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (7.3)$$

7.3 Supuestos del modelo de regresión lineal simple

En el modelo de regresión lineal simple deben cumplirse una serie de supuestos sobre la distribución de las variables. Mediante ellos se obtienen estimadores que poseen propiedades deseables. A continuación, se describen estos supuestos.

Supuesto 1

La variable independiente *X* no es una variable aleatoria, sino una variable fija en el muestreo, porque sus niveles de observación son seleccionados de antemano por el investigador y para cada valor de *X*, la variable

que se va a predecir, Y tiene distribución normal con media $\beta_0 + \beta_1 X_i$ y varianza σ^2.

Supuesto 2

La variable aleatoria ε es estadísticamente independiente de los valores de la variable dependiente X. Cuando X es una variable fija, tiene que ser independiente de la variable aleatoria ε, puesto que la covarianza de una variable aleatoria y una constante es igual a cero.

Supuesto 3

La variable aleatoria ε_i representa un conjunto de factores extraños (errores en la especificación del modelo, errores de observación o medida, etc.) no relacionados entre sí. Mediante el teorema central del límite se garantiza que su efecto conjunto tenga distribución normal.

Supuesto 4

El valor esperado de la variable aleatoria ε_i para cualquier X_i dado es igual a cero; es decir,

$$\mathrm{E}(\varepsilon_i/X_i) = 0$$

Este supuesto expresa que para un X_i dado, las diferencias entre Y y μ_{y/x_i} son a veces positivas y a veces negativas, pero su valor esperado es igual a cero. Por lo tanto, la distribución de ε_i respecto a la recta de regresión μ_{y/x_i} tiene siempre su centro en el valor μ_{y/x_i} para cualquierX_i dado.

Supuesto 5

Cualquier par de errores ε_k y ε_j son estadísticamente independientes entre sí; es decir, su covarianza es igual a cero.

$\varepsilon_k \ \varepsilon_j$

$$Cov\left(\varepsilon_k \ \varepsilon_j\right) = 0$$

Este supuesto implica que el error de un punto de la población no puede ser relacionado sistemáticamente con el error de cualquier otro punto de la población.

Supuesto 6

Las variables aleatorias ε_i tienen una varianza finita que es constante para todos los valores de X_i.

$$Var(\varepsilon_i/X_i) = \sigma^2$$

Este supuesto expresa que cada una de las perturbaciones asociadas a los diferentes valores de X_i posee la misma varianza σ^2. Se conoce como homocedasticidad (igual dispersión). Las poblaciones que no tienen igual dispersión se llaman heterocedásticas.

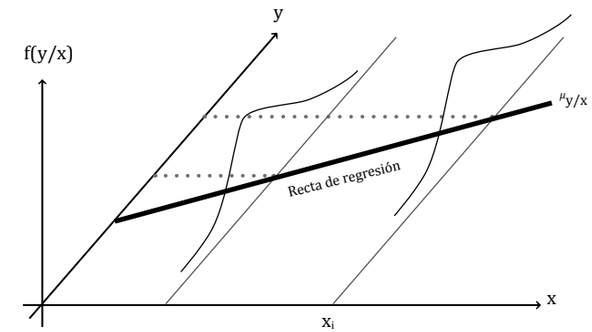


Figura 7.3 Errores ε_i que se distribuyen normalmente con centro en μ_{y/x_i} para cualquier valor X (considerando que también se cumplen los supuestos 4 y 6)
Fuente: MLB

7.4 Modelo de regresión lineal simple muestral

Es preciso tener en cuenta que para estimar los parámetros poblacionales se deben utilizar los datos muestrales. En el análisis de regresión lineal simple, los dos parámetros a estimar son β_0 y β_0, debido a que después se puede obtener una estimación de μ_{y/x_i}. La recta de regresión muestral sería:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (7.4)$$

Se observa que esta recta de regresión muestral está relacionada con la regresión poblacional $\mu_{y/x_i} = \beta_0 + \beta_1 X_i$ debido a que β_0 es el estimador puntual de β_0 y $\hat{\beta}_1$ es el estimador puntual de β_1. Con los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ y con un valor dado X_i se predice un valor de Y, designado por $\hat{Y}$, el cual representa el estimador de μ_{y/x_i}.

Para un valor dado X_i, el valor estimado $\hat{Y}$ está ubicado sobre la recta de regresión muestral $\hat{Y}_i = \beta_0 + \beta_1 X_1$, como se observa en la figura 7.4.

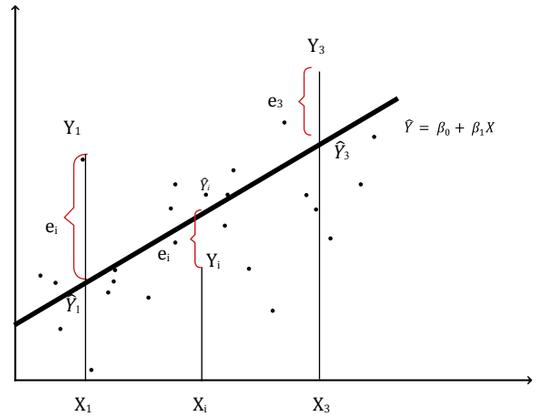


Figura 7.4 Representación gráfica de los valores estimados ($\hat{Y}_i$) y los residuos (e_i) en el modelo de regresión lineal simple muestral
Fuente: MLB

La diferencia entre el valor observado Y_i y el estimado $\hat{Y}_i$ se denomina error y se representa por e_i, es decir:

$$e_i = Y_i - \hat{Y}_i \quad ; \quad i = 1,2,\ldots n \quad (7.5)$$

Usualmente, los e_i se llaman residuos, ya que representan lo que no queda explicado después de usar el valor $\hat{Y}_i$ para estimar Y_i. Los residuos se pueden obtener por medio de gráficas. En la figura 7.5, la distancia vertical entre el valor observado Y_i y el estimado $\hat{Y}_i$, para un X_i dado, representan los residuos e_i. Se observa que los residuos pueden ser positivos o negativos. Son positivos si el valor observado de Y_i es menor que el estimado $\hat{Y}_i$.

El residuo e_i puede ser considerado como el estimador de e_i. Despejando Y_i en la expresión (7.5), obtendremos la ecuación del modelo de regresión lineal simple muestral.

$$Y_i = \hat{Y}_i + e_i \text{ ó}$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \quad (7.6)$$

Una vez que se ha especificado el modelo de regresión literal simple poblacional y muestral, se necesita un procedimiento para determinar los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ de forma tal que representen las mejores estimaciones de los parámetros desconocidos $\hat{\beta}_0$ y $\hat{\beta}_1$. El procedimiento para hallar tales estimaciones se denomina método de los mínimos cuadrados.

7.5 Estimación de los parámetros β_0 y β_1 por el método de los mínimos cuadrados ordinarios

Este método consiste en obtener los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ de modo que los valores resultantes de $\hat{Y}_i$ en la ecuación $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ sean los más cercanos posible a los valores observados Y_i; es decir, determinaremos los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la suma de los cuadrados de los residuos ($\sum e_i^2$), para hallar la recta que mejor se ajusta a los datos. Dicho método se define de la manera siguiente:

$$\text{Minimizar} \sum_{i=1}^{n} e_i^2 = Min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = Min \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$

Al diferenciar parcialmente $\hat{\beta}_0$ y $\hat{\beta}_1$ e igualar estas derivadas parciales a cero, se obtiene:

$$\frac{\partial e_i^2}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{\partial e_i^2}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0$$

Luego de realizar las operaciones algebraicas, se consiguen las siguientes ecuaciones, llamadas *ecuaciones normales.*

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \qquad (7.7)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad (7.8)$$

Al resolver este sistema de ecuaciones, la estimación de mínimos cuadrados de $\hat{\beta}_0$ y $\hat{\beta}_1$ son, respectivamente:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \qquad (7.9)$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} \qquad (7.10)$$

Ejemplo 7.2:

Obtenga la recta de regresión de mínimos cuadrados para los datos del ejemplo 7.1.

Solución:
Tal como se acaba de ver, la recta de regresión se obtiene mediante la siguiente ecuación:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Para lo anterior se deben calcular las estimaciones de $\hat{\beta}_0$ y $\hat{\beta}_1$. En tal sentido, se calculan en una tabla (7.2) todas aquellas expresiones que serán necesarias al momento de estimar los parámetros en cuestión.

Tabla 7.2 Cálculos para obtener la recta de mínimos cuadrados

	X_i	Y_i	X_iY_i	X_i^2
	25	34	850	625
	16	14	224	256
	42	48	2016	1764
	34	32	1088	1156
	10	26	260	100
	21	29	609	441
	19	20	380	361
Suma	167	203	5427	4703
Media	23,8571	29	-	-

Fuente: Ejemplo 7.1.

Una vez calculados los valores que se encuentran en la tabla 7.2, se pueden estimar los parámetros de la ecuación de regresión de la siguiente manera:

Para $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$\hat{\beta}_1 = \frac{7\times(5427) - 167\times203}{7\times4703 - 167^2}$$

$$\hat{\beta}_1 = \frac{37989 - 33901}{32921 - 27889} = 0{,}8124$$

Interpretación:
0,8124x100.000 = 81.240 $ indica el incremento en el ingreso promedio mensual por ventas por cada 1.000 $ en el gasto mensual de publicidad por televisión.

Para $\hat{\beta}_0$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_0 = 29 - 0{,}8124 \times 23{,}8571 = 9{,}6185$$

Interpretación:
9,6185x100.000 = 961850 $indica el ingreso promedio mensual por ventas cuando no se gasta en publicidad por televisión.

Por lo tanto, la recta de regresión por mínimos cuadrados ordinarios es:

$$\hat{Y}_i = 9{,}6185 + 0{,}8124 X_i$$

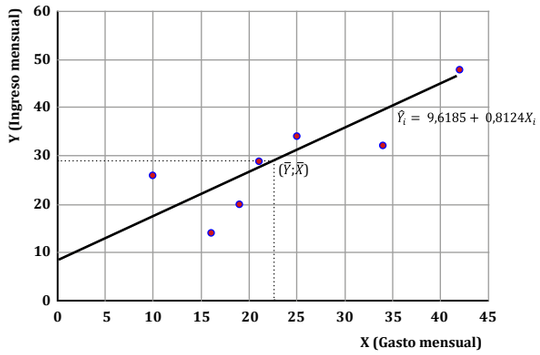


Figura 7.5 Recta de regresión por mínimos cuadrados ordinarios
Fuente:MLB

Una vez obtenida la recta de regresión muestral, se pueden calcular los valores estimados y los residuos para los puntos muestrales.

Ejemplo 7.3:

Obtenga los valores de la recta ajustada y de los errores para cada uno de los puntos muestrales del ejemplo 7.1.

Solución:
En la tabla 7.3 se presentan los valores correspondientes a:

$$\hat{Y}_i = 9{,}6185 + 0{,}8124 X_i$$

Tabla 7.3 Valores observados, estimados y residuos

Valor observado X_i	**Valor observado** Y_i	**Valor estimado** $\hat{Y}_i = 9{,}6185 + 0{,}8124X_i$	**Residuos** $e_i = Y_i - \hat{Y}_i$
25	34	29,9285	4,0715
16	14	22,6169	-8,6169
42	48	43,7393	4,2607
34	32	37,2401	-5,2401
10	26	17,7425	8,2575
21	29	26,6789	2,3211
19	20	25,0541	-5,0541

Fuente: MLB

La recta que se acaba de obtener sirve para diversos propósitos:

Los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ constituyen estimaciones puntuales de los parámetros poblacionales $\hat{\beta}_0$ y $\hat{\beta}_1$ respectivamente.

Mediante la recta de regresión se pueden predecir los valores promedio de la variable dependiente, dado un valor de la variable independiente. Suponga por ejemplo que se está interesado en predecir el ingreso medio mensual por ventas cuando el gasto por mes en publicidad por

televisión es igual a 20. En este caso, se sustituye el valor de $X_i = 20$ en la recta de regresión obtenida, es decir:

$$\hat{Y}_i = 9{,}6185 + 0{,}8124 \times 20 = 25{,}8665$$

Este resultado indica que el ingreso medio mensual por ventas es de 25,8665x100.000 = 258.665 $ cuando el gasto mensual por publicidad es de 20.000 dólares.

Cuando se usa la recta de regresión con fines de predicción es importante tener en cuenta solo el rango concerniente a la variable independiente. Este rango incluye todos los valores de *X*, desde el más pequeño hasta el más grande, usados para obtener la recta de regresión. Por tanto, cuando se predice el valor medio de *Y* para un determinado valor de *X*, se puede interpolar dentro del rango de los valores de *X*, pero no se puede extrapolar fuera del rango de los valores de la misma. Por ejemplo, cuando se usa el gasto mensual de publicidad para predecir el ingreso promedio mensual por ventas, se observa en el ejemplo 7.1 que la variable independiente (<) varía entre 10 y 42; en consecuencia, solo se debe hacer predicciones del ingreso promedio mensual por ventas cuando los gastos mensuales en publicidad por televisión estén entre 10 y 42.

7.6 Medidas de la bondad del ajuste

Aunque hemos visto que el método de mínimos cuadrados da como resultado una recta que se ajusta a los datos con el mínimo de variación, la recta de regresión no es un indicador perfecto de la predicción, a menos que todos los puntos de los datos observados se encuentren sobre la recta de regresión, lo cual es muy difícil que ocurra en la realidad. Por lo tanto, la recta de regresión solo se puede utilizar para hacer predicciones en forma aproximada. Necesitamos obtener medidas que nos indiquen la confiabilidad de la recta de regresión; para ello se utilizan las medidas de la bondad del ajuste, debido a que miden la bondad con la cual la línea de regresión se ajusta a las observaciones.

A continuación, se presenta un esquema de las medidas de bondad del ajuste de la recta de regresión en términos absolutos y relativos:

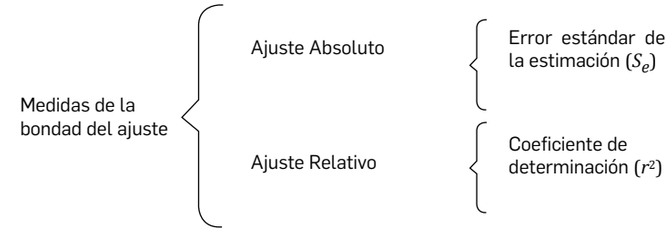


Figura 7.6 Medidas de la bondad del ajuste
Fuente: MLB

Para obtener las fórmulas de las medidas de la bondad del ajuste (S_e y r^2), en primer lugar presentaremos algunas medidas de la variabilidad en el análisis de regresión.

La diferencia entre Y_i y la media de estos valores ($\bar{Y}$) se llama desviación total de *Y*, y representa cuanto se desvía la i-ésima observación respecto a la media de todos los valores de *Y*.

$$\underset{\text{Desviación total}}{(Y_i - \bar{Y})} = \underset{\text{Desviación no explicada}}{(Y_i - \hat{Y}_i)} + \underset{\text{Desviación explicada}}{(\hat{Y}_i - \bar{Y}_i)}$$

Si elevamos al cuadrado cada una de las desviaciones anteriores y sumamos todos los valores correspondientes a la *n* observaciones, obtenemos las medidas de variación, es decir:

$$\underset{\text{Variación total}}{} = \underset{\text{Variación no explicada}}{} + \underset{\text{Variación explicada}}{}$$

$$\underset{\text{Suma de cuadrados del total (SCT)}}{\sum (Y_i - \bar{Y})^2} = \underset{\text{Suma de cuadrados del error (SCE)}}{\sum (Y_i - \hat{Y}_i)^2} + \underset{\text{Suma de cuadrados de regresión (SCR)}}{\sum (\hat{Y}_i - \bar{Y})^2}$$

$$\text{SCT} = \text{SCE} + \text{SCR} \qquad (7.12)$$

Las fórmulas abreviadas para obtener las sumas de cuadrados son:

$$SCT = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 \qquad (7.13)$$

$$SCR = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 (\sum X_i^2 - n\bar{X}^2) \qquad (7.14)$$

$$SCE = \sum (Y_i - \hat{Y}_i)^2 = SCT - SCR \qquad (7.15)$$

Una vez que se ha descompuesto la variación total en estos dos componentes, se pueden obtener las medidas de la bondad del ajuste.

El error estándar de la estimación se basa en el valor de la SCE y el coeficiente de determinación se basa en la magnitud relativa SCR respecto a SCT. A continuación, se describen estas medidas.

7.6.1 Error estándar de la estimación (S_e)

El error estándar de la estimación mide la variabilidad o dispersión de los valores observados alrededor de la línea de regresión. Viene expresado por S_e y se define como:

$$S_e = \sqrt{\frac{\sum (Y_i - Y_i)^2}{n-2}} = \sqrt{\frac{SCE}{n-2}} = \sqrt{CME} \qquad (7.16)$$

Se observa en la fórmula que la SCE se divide entre n – 2, debido a que los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ son obtenidos de una muestra de puntos de datos, perdiéndose dos grados de libertad cuando se emplean estos valores para estimar la recta de regresión. Cuanto más grande sea S_e , mayor será la dispersión de los puntos alrededor de la línea de regresión. En el caso extremo en que $S_e = 0$, se dice que la ecuación de predicción es un estimador perfecto, porque todos los puntos observados caen en la línea de regresión.

El error estándar de estimación (S_e) se utiliza, fundamentalmente, con fines comparativos (al seleccionar dos o más modelos que utilizan las mismas variables).

Ejemplo 7.4:

Calcular el error estándar de estimación (S_e) para la recta de regresión muestral que se ajusta a los datos del ejemplo 7.1.

Solución:
Como en la fórmula (7.14) se necesita calcular $\sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2$ y los valores de los e_i ya se han calculado en la columna 4 de la tabla 7.3, entonces solo falta elevar al cuadrado los e_i y obtener la sumatoria. A continuación, se presentan estos cálculos:

$e_i^2 = (Y_i - \hat{Y}_i)^2$
16,5771
74,2510
18,1536
27,4586
68,1863
5,3875
25,5439
235,5580

Por lo tanto,

$$S_e = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{235,5580}{7-2}} = 6,86378;$$

$\therefore S_e = 686.378$ \$

Existe un método abreviado para calcular el error estándar de la estimación, que no necesita realizar la suma de cuadrados del error y que se expresa de la siguiente manera:

$$S_e = \sqrt{\frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum X_i Y_i}{n-2}} \qquad (7.17)$$

Ahora se puede calcular S_e mediante el método abreviado. En la tabla 7.2 se obtuvo que $\sum Y_i = 203$ y $\sum X_i Y_i = 5.427$ y luego, los valores de $\hat{\beta}_0 = 9{,}6185$ y $\hat{\beta}_1 = 0{,}8124$. Por lo tanto, solo faltaría $\sum Y_i^2$, la cual se presenta a continuación:

Y_i^2
1.156
196
2.304
1.024
676
841
400
6.597

Al sustituir estos valores en la fórmula (7.15), se obtiene que el valor de S_e por el método abreviado es:

$$S_e = \sqrt{\frac{6.597 - 9{,}6185\,(203) - 0{,}8124\,(5.427)}{7-2}} = \sqrt{\frac{235{,}5497}{5}} = 6{,}86367;$$

$\therefore\ S_e = 686.367\ \$$

Este error estándar de la estimación igual a 6,86367 representa una medida de la variación alrededor de la línea de regresión ajustada.

7.6.2 Coeficiente de determinación (r^2)

El coeficiente de determinación es la medida de la bondad del ajuste, que sirve para obtener la cantidad relativa de la variación de la variable dependiente *Y* explicada por la variable independiente *X*. Se denota por r^2 y se obtiene por medio de SCT, SCR y SCE.

Al dividir la relación SCT = SCR + SCE por SCT se obtiene que:

$$\frac{\text{SCT}}{\text{SCT}} = \frac{\text{SCR}}{\text{SCT}} + \frac{\text{SCE}}{\text{SCT}} \qquad (7.18)$$

$$1 = \frac{\text{SCR}}{\text{SCT}} + \frac{\text{SCE}}{\text{SCT}} \qquad (7.19)$$

Se observa que el cociente $\left(\frac{\text{SCR}}{\text{SCT}}\right)$ es la parte de la variación total explicada por la regresión, y que $\left(\frac{\text{SCE}}{\text{SCT}}\right)$ es la parte de la variación total no explicada por la regresión.

El cociente $\left(\frac{\text{SCR}}{\text{SCT}}\right)$ es la medida relativa de la bondad del ajuste y se denomina coeficiente de determinación (muestral). Es decir:

$$r^2 = \frac{\text{SCR}}{\text{SCT}} = \frac{\text{Variación explicada por la regresión}}{\text{Variación total}} \qquad (7.20)$$

El coeficiente de determinación mide la proporción o el porcentaje de la variación de la variable dependiente *Y* explicada por la variable independiente *X*, en el modelo de regresión lineal simple.

En forma equivalente, se puede obtener otra fórmula para calcular r^2, si se despeja de la expresión (7.19) $\left(\frac{\text{SCR}}{\text{SCT}}\right)$, es decir:

$$\frac{\text{SCR}}{\text{SCT}} = 1 - \frac{\text{SCE}}{\text{SCT}} \qquad (7.21)$$

Al observar las fórmulas dadas para obtener r^2, se deducen las siguientes características:

a. Como el coeficiente de determinación es un cociente de dos sumas de cuadrados, entonces r^2 es una medida no negativa.

b. Como SCR es menor o igual a SCT y por ser un valor no negativo, el coeficiente de determinación va a estar comprendido entre 0 y 1, es decir:

$$0 \leq r^2 \leq 1$$

c. Un $r^2 = 0$ indica que no existe relación lineal entre las variables X en Y, lo cual significa que ninguna parte de la variación de Y está explicada por X. El valor de r^2 va a ser igual a 0 cuando SCR = 0 y SCE = SCT.

d. Un $r^2 = 1$ indica que existe una relación lineal perfecta entre las variables X y Y, lo cual significa que la variación existente en Y es explicada totalmente por X y todos los puntos observados están sobre la recta de regresión muestral. Por lo tanto, SCE = 0 y SCR = SCT son el ajuste perfecto.

Ejemplo 7.5:

Calcular e interpretar el coeficiente de determinación muestral para los datos del ejemplo 7.1.

Solución:
Si se utiliza la fórmula (7.18), se necesita conocer los valores de SCR y SCT.

$$SCR = \hat{\beta}_1^2 \left(\sum X_i^2 - n\bar{X}^2 \right)$$

$$SCR = (0{,}8124)^2 [4.703 - 7\,(23{,}8571)^2]$$

$$= 0{,}65999\,(718{,}8715)$$

$$= 474{,}4480$$

$$SCT = \sum Y_i^2 - n\,\bar{Y}^2 = 6.597 - 7\,(29)^2 = 710$$

Al sustituir estos valores en la fórmula (7.18), se obtiene que el valor del coeficiente de determinación muestral es:

$$r^2 = \frac{\text{SCR}}{\text{SCT}} = \frac{474{,}4480}{710} = 0{,}6682$$

Este resultado indica que el 66,82 % de la variación muestral total en el ingreso mensual por ventas (Y) se explica por el gasto mensual en publicidad (X). El 33,18 % restante de la variación en el ingreso mensual por ventas no ha sido explicado por el modelo.

7.7 Inferencia estadística con respecto a los parámetros β_0 y β_1

Aun cuando la relación entre las dos variables en estudio de una población es muy poca o nula, es posible obtener valores muestrales que hagan parecer que las variables X y Y están relacionadas linealmente, debido a que los factores aleatorios en el muestreo han dado lugar a una relación de tipo lineal, cuando en realidad no existe este tipo de relación. Por ejemplo, en la figura 7.7.a se representa una población en la que X y Y no tienen ninguna relación. En la figura 7.7.b se muestran algunas posibles observaciones muestrales seleccionadas de la misma población que hacen parecer que existe una relación lineal, cuando en realidad no existe este tipo de relación. Por esta razón, una vez que se ha calculado la recta de regresión muestral, es importante conocer si esta misma puede ser utilizada para fines predictivos. Es decir, nos interesa determinar si el conocimiento de la variable independiente X resulta útil para predecir los valores de la variable dependiente Y.

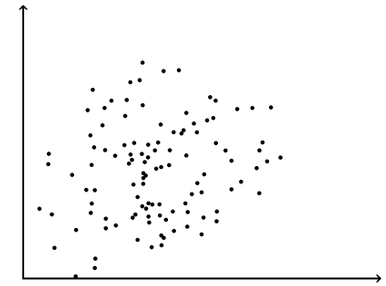


Figura 7.7.a Población en la cual no existe relación entre las dos variables

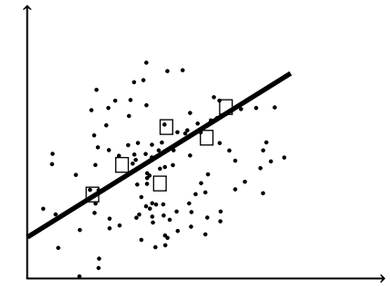


Figura 7.7.b Observaciones hipotéticas muestrales obtenidas de la población que hacen parecer que exista una relación lineal

Fuente: elaboración propia

7.8 Prueba de hipótesis y estimación por intervalo para β_1

Se centra nuestra atención en la prueba de hipótesis y en la estimación por intervalo para β_1 ya que mediante estos métodos podemos determinar si existe o no una relación lineal entre las variables X y Y.

Prueba de hipótesis para β_1

Suponga que en la población de variables, X y Y no están relacionadas linealmente, lo cual indica que $\beta_1 = 0$. Es decir, la recta de regresión poblacional es una línea horizontal y, por lo tanto, $Y_i = \bar{Y}$, puesto que Y_i es constante cuando $\beta_1 = 0$. Los valores de X no sirven para predecir Y.

Si las variables X y Y en la población están relacionadas linealmente, significa que $\beta_1 \neq 0$ y que los valores de X se pueden utilizar para predecir los valores de Y. Así, para determinar si la recta de regresión se puede usar con fines predictivos, debemos probar la hipótesis nula $H_0 : \beta_1 = 0$. La formulación de la hipótesis alternativa depende del conocimiento previo que se tenga de β_1. Si se tiene información a priori sobre β_1, en el sentido de que no puede ser negativo, la hipótesis alternativa (H_1) es unilateral derecha ($H_1 : \beta_1 > 0$); si la información es que β_1 no puede ser positivo, la hipótesis alternativa es unilateral izquierda ($H_1 : \beta_1 < 0$) y si no se tiene conocimiento sobre los posibles valores del parámetro de β_1, la hipótesis alternativa es bilateral ($H_1 : \beta_1 \neq 0$).

Para probar la hipótesis ($H_0 : \beta_1 = 0$) se usa la distribución t.

La prueba de hipótesis para β_1 es muy parecida a la media poblacional (μ), debido a que también en este caso la prueba se refiere a una media (β_1), el estimador es $\hat{\beta}_1$ (en vez de $\bar{X}$), el valor hipotético es cero (en vez de μ_0) y el error estándar del estimador es $S_{\hat{\beta}_1}$ (en vez de $S_{\bar{x}}$), donde $S_{\hat{\beta}_1}$ se define de la siguiente forma:

$$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{\sum (X_i - X)^2}} = \frac{S_e}{\sqrt{\sum X_i^2 - n\,\bar{X}^2}} \qquad (7.22)$$

Por lo tanto, el estadístico apropiado sería:

$$t = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \qquad (7.23)$$

que tiene una distribución t con (n – 2) grados de libertad.

También es posible probar la hipótesis nula $H_0 : \beta_1 = \beta_1^*$, es decir, la hipótesis de que la pendiente es un cierto valor especificado distinto de cero. Este valor suele denotarse por β_1^*, de modo que las hipótesis se formulan de la siguiente manera:

$$H_0 : \beta_1 = \beta_1^*$$
$$H_1 : \beta_1 > \beta_1^*$$
$$>$$
$$\neq$$

y el estadístico sería:

$$t_c = \frac{\hat{\beta}_1 - \beta_1^*}{S_{\hat{\beta}_1}}$$

que tiene una distribución *t-student* con n – 2 grados de libertad.

Ejemplo 7.6:

Usando los datos del ejemplo 7.1, pruebe la hipótesis $H_0 : \beta_1 = 0$ contra la alternativa $\beta_1 > 0$, a un nivel de significación del 0,05 (se usa una hipótesis alternativa unilateral derecha, porque no es razonable esperar que exista una relación inversa entre el ingreso mensual por ventas y el gasto mensual en publicidad).

Solución:

a. Plantear las hipótesis a contrastar:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

o lo que es lo mismo:

H_0 no existe relación lineal entre el ingreso mensual por ventas y el gasto mensual en publicidad
H_1 existe relación lineal directa entre el ingreso mensual por ventas y el gasto mensual en publicidad

b. Fijar el nivel de significación: $\alpha = 0{,}05$

c. Establecer el estadístico a utilizar:

$$t_c = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

d. Determinar el valor y la región críticos:

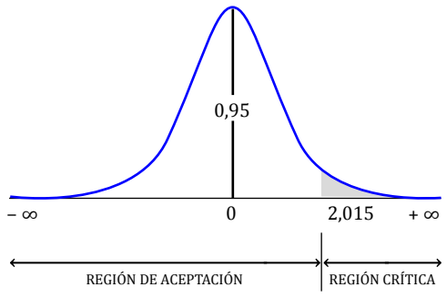


Figura 7.8 Valor y región críticos
Fuente: MLB

e. Calcular el estadístico:
Para calcular el estadístico de prueba, primero se debe calcular $S_{\hat{\beta}_1}$, sustituyendo en la fórmula (7.22) los valores previamente calculados de $S_e = 6,86367$ y $\sum X_i^2 - n\bar{X}^2 = 718{,}8715$. Así se obtiene que:

$$S_{\hat{\beta}_1} = \frac{6{,}86367}{\sqrt{718{,}8715}} = 0{,}2560$$

Por lo tanto, el valor calculado del estadístico es:

$$t_c = \frac{0{,}8124}{0{,}2560} = 3{,}1734$$

f. Decidir e interpretar:
Como el valor calculado t = 3,1734 cae en la región de rechazo, se decide rechazar H_0 al nivel de significación de 0,05. Se concluye que con base en la información proporcionada por la muestra, entre las variables ingreso mensual y gasto mensual existe una relación lineal directa.

7.9 Intervalo de confianza para β_1

Un método equivalente a la prueba de hipótesis que se utiliza para probar la existencia de una relación lineal entre dos variables es obtener un intervalo de confianza para β_1 y determinar si el valor hipotético ($\beta_1 = 0$) está incluido en el intervalo. Los límites del intervalo de confianza para la pendiente de la población β_1 se obtienen de la siguiente manera:

$$\hat{\beta}_1 \pm t_{n-2;\,\frac{a}{2}} S_{\hat{\beta}_1} \qquad (7.24)$$

Este intervalo de confianza proporciona la siguiente información:

a. Indica el intervalo probable en el que puede estar el valor verdadero de β_1.

b. Si el intervalo de confianza incluye el valor cero, sería equivalente a una prueba en la que $H_0 : \beta_1 = 0$ no se puede rechazar y las variables X y Y no están relacionadas linealmente.

c. Si el intervalo de confianza no incluye el valor cero, sería equivalente a una prueba en la que $H_0 : \beta_1 = 0$ se rechaza y las variables X y Y están relacionadas linealmente.

Ejemplo 7.7:

Encuentre un intervalo de confianza de 95 % para β_1, basándose en los datos del ejemplo 7.1.

Solución:
Siguiendo con los datos del ejemplo 7.1, se puede obtener un intervalo de 95 % para β_1 de la siguiente manera:

Dado que $\hat{\beta}_1 = 0{,}8124$, $\quad t_{n-2;\frac{a}{2}} = t_{5;\ 0{,}025} = 2{,}571$ y $S_{\hat{\beta}_1} = 0{,}2560$

por lo tanto:

$$\hat{\beta}_1 \pm t_{n-2;\frac{a}{2}} S_{\hat{\beta}_1}$$

$$0{,}8124 \pm 2{,}571\ (0{,}2560)$$

$$0{,}8124 \pm 0{,}6582$$

$$0{,}1542 < \beta_1 < 1{,}4706$$

Interpretación:
Se espera, con una confianza de 95 %, que la pendiente de la recta de regresión poblacional se encuentre entre 0,1542 y 1,4706. Puesto que estos valores son positivos, se puede concluir que existe una relación lineal directa entre el ingreso mensual por ventas y el gasto mensual en publicidad, con una confianza de 95 %.

7.10 Prueba de hipótesis y estimación por intervalo para β_0

Con un procedimiento similar al anterior se puede realizar pruebas de hipótesis o estimación por intervalo referentes al parámetro poblacional β_0. En este caso, el estadístico apropiado sería:

$$t_c = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \qquad (7.25)$$

que tiene una distribución t con n – 2 grados de libertad.

$S_{\hat{\beta}_0}$ se obtiene de la siguiente manera:

$$S_{\hat{\beta}_0} = \frac{S_e \sqrt{\sum X_i^2}}{\sqrt{n \sum (X_i - \bar{X})^2}} = \frac{S_e \sqrt{\sum X_i^2}}{\sqrt{n\ (\sum X_i^2 - n\ \bar{X}^2)}} \qquad (7.26)$$

Los límites del intervalo de confianza para β_0 se calculan así:

$$\hat{\beta}_0 \pm t_{n-2;\ \frac{a}{2}} S_{\hat{\beta}_0} \qquad (7.27)$$

7.11 Análisis de varianza en la regresión lineal simple

Existe una forma equivalente de probar la hipótesis de no existencia de una relación lineal entre X y Y ($H_0 : \beta_1 = 0$): por medio del método del análisis de varianza.

En la sección 7 se descompuso la variación muestral total en variación no explicada (variación debida al error) y variación explicada (variación debida a la regresión); es decir:

$$\text{SCT} = \text{SCE} + \text{SCR}$$

Ahora se necesita conocer los grados de libertad para calcular los cuadrados medios. Por lo tanto, los grados de libertad asociados a:

a. La suma de cuadrados total (SCT) son n – 1 (ya que solo es necesario determinar $\bar{Y}$ para poder calcular SCT).

b. La suma de cuadrados del error (SCE) son n – 2 (debido a que para calcular la SCE han de determinarse previamente $\hat{\beta}_0$ y $\hat{\beta}_1$).

a. La suma de cuadrados de la regresión (SCR) tiene 1 grado de libertad (ya que los grados de libertad de SCT son iguales a la suma de los grados de libertad de SCE y SCR; es decir, n – 1 = ((n – 2) + 1).

7.11.1 Cuadrados medios

Si se dividen las sumas de los cuadrados entre sus respectivos grados de libertad, se obtienen los cuadrados medios. Los dos cuadrados medios que se necesitan son el cuadrado medio de la regresión (CMR) y el cuadrado medio del error (CME).

$$CMR = \frac{SCR}{1} = SCR$$

$$CMR = \frac{SCR}{1} = SCR$$

Se acostumbra presentar la información obtenida en una tabla de análisis de varianza, como se indica en la tabla 7.4.

Tabla 7.4 Tabla de análisis de varianza para la regresión lineal simple

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Regresión	SCR	1	SCR / 1
Error (o residuo)	SCE	n – 2	SCE / (n – 2)
Total	SCT	n – 1	

Fuente: MLB

Si el valor de CMR es alto en relación con el de CME, una buena parte de la variabilidad en *Y* está siendo explicada por la recta de regresión, lo cual implica que se debe rechazar la hipótesis nula. Sin embargo, si CMR es pequeña en relación con CME, la recta de regresión no explica la variabilidad existente en los valores de *Y*, razón por la cual no se

puede rechazar la hipótesis nula. Así, puede ser comprobado $H_0 : \beta_1 = 0$ utilizando el cociente de los cuadrados medios CMR/CME. Este cociente tiene una distribución *F* con 1 y (n – 2) grados de libertad.

$$F_c = \frac{CMR}{CME} \qquad (7.28)$$

Como se sabe, en el análisis de varianza la región crítica se localiza en el extremo superior de la distribución *F* y el valor crítico se localiza en la tabla *F* con 1 grado de libertad en el numerador y n – 2 grados de libertad en el denominador.

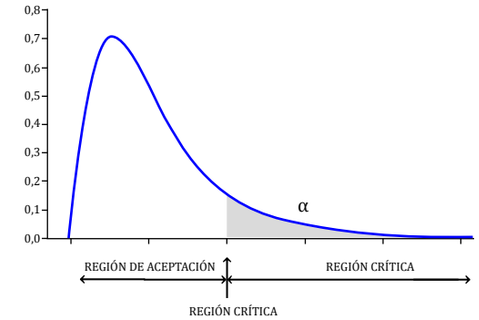


Figura 7.9 Valor y región críticos
Fuente: MLB

Cuando el valor del estadístico de prueba se encuentre en la región crítica, se rechaza H_0 y se prueba la existencia de la relación lineal entre la variable dependiente y la independiente, para un nivel de significación dado (α).

Ejemplo 7.8:

Con los datos del ejemplo 7.1, pruebe la hipótesis $H_0 : \beta_1 = 0$, utilizando el método de análisis de varianza.

Solución:
Para ilustrar la aplicación del análisis de varianza en la regresión lineal simple, nos basamos en los cálculos de las sumas de los cuadrados presentados en la sección 7.2, obteniéndose que:

SCT = 710; SCR = 474,4480 y por diferencia se obtiene que SCE = 235,5520

En la tabla 7.5 se presenta el análisis de varianza para este ejemplo.

Tabla 7.5 Tabla de análisis de varianza para el ejemplo dado

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Regresión	474,4480	1	474,4480
Error	235,5520	5	47,1104
Total	710,0000	6	

Fuente: MLB

Con base en la tabla anterior se calcula el valor del estadístico de prueba:

La región crítica y el valor crítico serían:

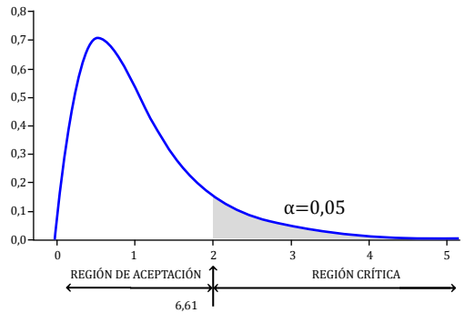


Figura 7.10 Valor y región críticos
Fuente: MLB

Como el valor del estadístico cae en la región crítica, se rechaza la hipótesis nula y se concluye que existe una relación lineal entre el ingreso mensual por ventas y el gasto mensual en publicidad, con una α = 0,05.

Puede demostrarse que t y F son pruebas equivalentes al determinar la significación de la relación lineal entre las variables X y Y. El valor del estadístico F es igual al cuadrado del valor del estadístico t, en nuestro ejemplo, t = 3,1734. De modo que:

La ventaja de la prueba t es que puede utilizarse en los casos en que β_1 toma valores diferentes de cero, mientras que F, en un análisis de varianza, solo es apropiada para probar la hipótesis $H_0 : \beta_1 = 0$.

7.12 Predicción en el análisis de regresión lineal simple

Se ha visto que una de las aplicaciones más importantes de la regresión lineal simple es la predicción de los valores de la variable dependiente Y para valores dados de la variable independiente X. La predicción puede ser puntual o por intervalos, y se pueden realizar dos tipos de predicciones: predicción para el valor medio de Y dado un valor de X denominado $X_0(\mu_y/X_0)$ y predicción para un valor real individual de Y denominado Y_0 dado $X_0(\mu_y/X_0)$.

7.13 Estimaciones puntuales de las predicciones

Para obtener la mejor estimación puntual de las predicciones de valor medio y del valor real de Y dado el valor de la variable independiente (X_0), se sustituye este valor en la ecuación de la recta de regresión muestral. De esta manera, resulta que el valor de predicción es:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Así, $\hat{Y}_0$ es una estimación puntual tanto de μ_{y/x_0} como de Y_0/X_0.

Cuando el gasto mensual en publicidad por televisión es igual a 15.000 $, se obtiene que el ingreso mensual promedio por ventas es 21,8045

(100.0000) = 2.180.450 $. Análogamente, la mejor estimación puntual del ingreso mensual por ventas es de 2.180.450 $ cuando el gasto mensual individual es 15.000 $.

7.14 Intervalo de confianza para la predicción media μ_{y/x_0}

Los límites del intervalo de confianza para μ_{y/x_0} se obtienen de la siguiente manera:

$$\hat{Y}_0 \pm t_{n-2\,;\frac{a}{2}}\, S_{\hat{Y}_0} \qquad (7.29)$$

donde:

$$S_{\hat{Y}_0} = S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \qquad (7.30)$$

De manera equivalente, $S_{\hat{Y}_0}$ se puede obtener como:

$$S_{\hat{Y}_0} = S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i^2 - n\,\bar{X}^2}} \qquad (7.31)$$

Ejemplo 7.9:

Obtenga un intervalo de 95 % para estimar el ingreso medio mensual por ventas cuando el gasto en publicidad por televisión es de 15.000 $, con base en los datos del ejercicio 7.1.

Solución:
Se tiene $X_0 = 15$ $\hat{Y}_0 = 21{,}8045$, $t_{n-2\,;\frac{a}{2}} = t_{5;\,0{,}025} = 2{,}57$ y el error estándar de Y_0 sería:

$$S_{\hat{Y}_0} = S_e \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum X_i^2 - n\,\bar{X}^2}}$$

$$= 6{,}86367 \sqrt{\frac{1}{7} + \frac{(15 - 23{,}8571)^2}{718{,}8715}} = 3{,}4456$$

Al sustituir en los límites del intervalo nos queda:

$$21{,}8045 \pm 2{,}571\ (3{,}4456)$$

$$21{,}8045 \pm 8{,}8586$$

$$(12{,}9459 < \mu_{Y/X_{0=15}} < 30{,}6631)$$

Al multiplicar por 100.000 ambos extremos del intervalo nos resultan:

$$(1.294.590 < \mu_{Y/X_{0=15.000}} < 3.066.310)\ \$$$

Interpretación:

Se espera con un 95 % de confianza que el ingreso medio mensual por ventas se encuentre entre 1.294.590 $ y 3.066.310, cuando el gasto mensual en publicidad es igual a 15.000 $.

7.15 Intervalo de confianza para la predicción individual Y_0/X_0

Si queremos conocer la predicción media, nos puede interesar el intervalo de confianza para la predicción individual. En este caso, lo único que cambia en el intervalo para la predicción individual en relación con la predicción media es el error estándar, que se denota por ($S_{\hat{Y}_0^*}$) y se expresa como:

$$S_{\hat{Y}_0^*} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \qquad (7.32)$$

o equivalentemente:

$$S_{\hat{Y}_0^*} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{\sum X_i^2 - n\,\bar{X}^2}} \qquad (7.33)$$

Los límites del intervalo de confianza para Y_0/X_0 se obtienen de la siguiente manera:

$$\hat{Y}_0 \pm t_{n-2\,;\frac{a}{2}} S_{\hat{Y}_0^*} \qquad (7.34)$$

Ejemplo 7.10:

Con los datos del ejemplo 7.1, calcular un intervalo de confianza de 95 % para el ingreso por venta de un mes individual, cuando el gasto en publicidad sea de 15.000 $, es decir .

Solución:
Se ha calculado que $\hat{Y}_0 = 21{,}8045$; $t_{5;\,0{,}025} = 2{,}571$ y el error estándar de Y_0/X_0 sería:

$$S_{\hat{Y}_0^*} = 6{,}8637\sqrt{1 + \frac{1}{7} + \frac{(15 - 23{,}8571)^2}{718{,}8715}}$$

$$S_{\hat{Y}_0^*} = 6{,}8637\sqrt{1{,}2520} = 7{,}6798$$

Y al sustituir en los límites del intervalo nos quedaría:

$$21{,}8045 \pm 2{,}571\ (7{,}6798)$$

$$21{,}8045 \pm 19{,}7448$$

$$(2{,}0597 < Y_0/X_0 = 15 < 41{,}5493)$$

Al multiplicar por 100.000 ambos extremos resulta:

$$(205.970 < Y_0/X_0 = 15.000 < 4.154.930)\ \$.$$

Interpretación:
Se espera, con un 95 % de confianza, que el ingreso para un mes individual por venta se encuentre en el intervalo hallado, cuando el gasto mensual en publicidad sea de 15.000 $.

7.16 Análisis de correlación lineal simple

Frecuentemente, este análisis se utiliza junto con el análisis de regresión lineal simple para medir la eficacia con que la recta de regresión explica la variación de la variable dependiente *Y*. También puede usarse la correlación para medir el grado de asociación o relación lineal entre las variables *X* y *Y*.

Los supuestos sobre la población en que se basa el análisis de correlación lineal simple son:

Supuesto 1:
La relación entre las variables *X* y *Y* es lineal.

Supuesto 2:
Las variables *X* y *Y* son aleatorias, puesto que ninguno de sus valores es predeterminado.

Supuesto 3:
Para cada una de las variables, las varianzas condicionales para los diferentes valores de la otra variable son iguales $\left(\sigma^2_{Y/X_1} = \sigma^2_{X/Y_1} = \sigma^2\right)$

Supuesto 4:
Para cada variable, las distribuciones condicionales, dados diferentes valores de la otra variable, son distribuciones normales, es decir, se supone una distribución normal bivariada.

Con base en estos supuestos, se pasa ahora a definir cómo se calcula el coeficiente de correlación poblacional o coeficiente de correlación momento-producto o correlación de Pearson, que se denota por la letra griega *p* (rho) y se utiliza para medir el grado de asociación lineal entre las variables *X* y *Y*. Se expresa de la siguiente forma:

$$\rho = \frac{Covarianza\ de\ X\ y\ Y}{(Desv.est.de\ X)(Des.est.de\ )} = \frac{Cov\,(X,Y)}{\sigma_x\,\sigma_Y} \qquad (7.35)$$

Si en la población todos los pares de valores de *X* y *Y* están sobre una recta de pendiente positiva, se dice que hay una correlación lineal positiva

perfecta entre X y Y. En este caso, el valor de Cov (X, Y) será exactamente igual al producto $\sigma_X\ \sigma_Y$ de modo que ρ será igual a +1. Es decir, $\rho = +1$.

Si en la población todos los pares de valores de X y Y están sobre una recta de pendiente negativa, se dice que hay correlación lineal negativa perfecta entre X y Y. En este caso, el valor de Cov (X, Y) será exactamente igual al producto $-\sigma_X\ \sigma_Y$ de modo que ρ será igual a -1. Es decir, $\rho = -1$.

Si X y Y no están relacionadas linealmente (si son variables aleatorias independientes), entonces la Cov (X, Y) será igual a cero y el coeficiente de correlación será también igual a cero. Es decir, $\rho = 0$.

De las observaciones anteriores se deduce que:

a. $-1 \leq \rho \leq +1$

b. Los valores de ρ cercanos a cero indican una débil correlación lineal entre X y Y.

c. Los valores de ρ cercanos a +1, 0 indican una fuerte correlación lineal positiva entre X y Y.

d. Los valores de ρ cercanos a − 1, 0 indican una fuerte correlación lineal negativa entre X y Y.

La figura 7.7 ilustra los valores de ρ correspondientes a algunos diagramas de dispersión seleccionados.

Nótese en las figuras 7.11.b y 7.11.c que dos poblaciones bastante diferentes pueden tener el mismo coeficiente de correlación, pero diferente pendiente en la recta poblacional. Las figuras 7.11.c y 7.11.d muestran la diferencia entre la correlación positiva y la negativa. Los dos últimos diagramas son ejemplos de poblaciones con correlación cero. En la figura 7.11.f, X y Y están relacionadas, pero en forma no lineal, a pesar de que existe cierto tipo de relación $\rho = 0$, lo cual remarca el hecho de que ρ mide la fuerza de la relación no lineal.

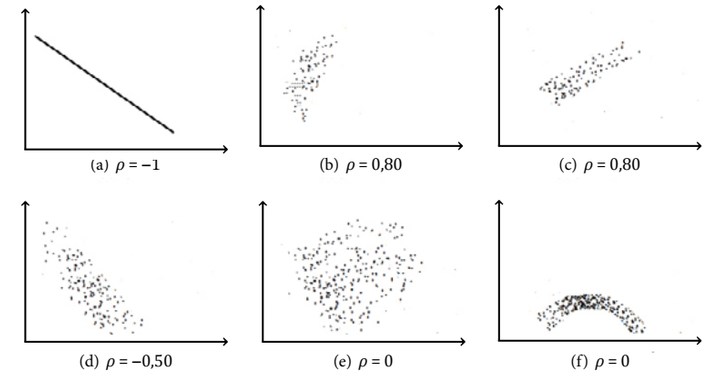


Figura 7.11 Coeficiente de correlación poblacional
Fuente: MLB

7.17 Coeficiente de correlación muestral (r)

Como en todos los problemas de estimación, para estimar el parámetro poblacional ρ se usarán los datos muestrales (o sea, algunos valores de dos variables aleatorias X y Y). En este caso, el estadístico muestral recibe el nombre de coeficiente de correlación muestral y se denota por la letra r . El valor de r se define del mismo modo que ρ, pero en términos muestrales.

$$r = \frac{Covarianza\ de\ los\ valores\ muestrales\ X\ y\ Y}{\left(\begin{matrix}desviación\ estándar\\ muestral\ de\ X\end{matrix}\right)\left(\begin{matrix}desviación\ estándar\\ muestral\ de\ Y\end{matrix}\right)} = \frac{S_{xy}}{S_x\ S_y}$$

$$= \frac{\frac{1}{n-1}\ \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\ \sum_{i=1}^{n}(X_i^2 - \bar{X})^2}\sqrt{\frac{1}{n-1}\ \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (7.36)$$

La siguiente fórmula simplifica el cálculo de r:

$$r = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{n \sum X_i^2 - (\sum X_i)^2}\sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} \qquad (7.37)$$

El coeficiente de correlación muestral se interpreta de la misma forma que ρ, excepto porque se refiere a los valores muestrales y no a los poblacionales. Por ejemplo, cuando $r = \pm 1$ hay un ajuste lineal perfecto entre los valores muestrales de X y Y, por lo que se dice que tienen una correlación lineal perfecta. Si entre los valores muestrales de X y Y no hay ninguna relación, el valor de r es igual a cero.

Para que r sea un estimador insesgado de ρ, la distribución conjunta de X y Y debe ser normal.

Ejemplo 7.11:

Obtenga el coeficiente de correlación muestral (r) basándose en los datos del ejemplo 7.1.

Solución:
Para calcular el coeficiente de correlación muestral (r) se usan los datos del ejemplo 7.1 y los cálculos de la recta de regresión muestral, ya que se observa que el numerador de β_1 coincide con el de r y que $n \sum X_i^2 - (\sum X_i)^2 = 5.032$. Por lo tanto, se obtiene que:

$$r = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{n \sum X_i^2 - (\sum X_i)^2}\sqrt{n \sum Y_i^2 - (\sum Y_i)^2}} = \frac{4.088}{\sqrt{5.032}\sqrt{4.970}} = 0{,}8175$$

Este resultado nos indica que existe una relación lineal positiva fuerte entre las variables ingreso mensual por ventas y gasto mensual en publicidad.

Se recomienda precaución a quienes intenten inferir una relación causa-efecto del análisis de correlación o de regresión, ya que una correlación alta o un buen ajuste de una recta de regresión no implica que X sea causa de Y. Ni siquiera implica que X proporcionará una buena estimación de Y en el futuro o con otro conjunto de observaciones muestrales.

Muchos estudiantes recién ingresados a la universidad consideran, por ejemplo, que la nota obtenida en la prueba de aptitud académica es un buen indicador de su futuro éxito académico, y es de suponer que basan su opinión en el hecho de que cualquiera de las aptitudes que mide esta prueba (memoria, inteligencia, vocabulario) influirá en sus notas futuras y no en la consideración de que las notas altas en el examen de ingreso influyan en la obtención de notas altas en la universidad. Otro ejemplo es el de los economistas, que usan con frecuencia las técnicas de la regresión para determinar posibles relaciones causa-efecto que puedan ser útiles para predecir el futuro de la economía o el efecto de alguna política económica; en estos casos se recomienda establecer primero la base teórica del problema en cuestión y luego explicar la asociación que existe entre las dos variables. Sin una base teórica que sustente la relación entre ambas variables, no tiene sentido el cálculo de t.

7.18 Prueba de hipótesis acerca del coeficiente de correlación poblacional (ρ)

Para conocer cuándo un valor dado de r es significativamente diferente de cero, se realiza la prueba de hipótesis para $\rho = 0$. Cuando $\rho = 0$ la distribución de r es simétrica, mientras que cuando ρ es diferente de cero, la distribución r es asimétrica. La propiedad de que la distribución de r es simétrica cuando $\rho = 0$ hace posible probar la hipótesis $H_0 : \rho = 0$. Para probar esta hipótesis se utiliza la siguiente variable aleatoria:

$$t_c = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad (7.38)$$

que tiene distribución t con n – 2 grados de libertad.

Ejemplo 7.12:

Pruebe la hipótesis $H_0 : \rho = 0$ con un nivel de significación de 5 % basándose en los datos del ejemplo 7.1.

Solución:
Para probar la hipótesis $H_0 : \rho = 0$ se usa, nuevamente, el ejemplo 7.1, en el cual ya se calculó que $r = 0{,}8115$ Por la misma razón que antes se utilizó la hipótesis alternativa $H_1 : \beta_1 > 0$ se usa ahora $H_1 : \rho > 0$, y el valor del estadístico de prueba sería:

$$t_c = \frac{0{,}8175}{\sqrt{\frac{1-(0{,}8571)^2}{5}}} = 3{,}173$$

El valor crítico es el mismo que se obtuvo para probar $H_1 : \beta_1 > 0$, es decir, $t_{5,\ 0{,}05} = 2{,}015$. Por lo tanto, se rechaza H_0 y se concluye que existe una correlación lineal positiva entre las variables de estudio.

El valor de t en esta prueba, $t = 3{,}173$, es exactamente el mismo que se obtuvo al probar $H_0 : \beta_1 = 0$. Esta concordancia se debe a que son pruebas equivalentes. Además, $\hat{\beta}$ y r están estrechamente relacionados, pero producen diferentes interpretaciones, ya que r mide la asociación lineal entre X y Y, mientras que $\hat{\beta}_1$ mide el tamaño del cambio medio en Y cuando X cambia en una unidad. Así, se puede obtener $\hat{\beta}_1$ en función de r y viceversa.

$$\hat{\beta}_1 = r\frac{S_Y}{S_X} \quad y \quad r = \hat{\beta}_1 \frac{S_x}{S_Y} \qquad (7.39)$$

Ejercicios

1. Considere la estimación de la función de producción que expresa la relación entre el nivel de salida y el nivel de entrada de mercancía de un factor que tienen los datos que se indica en la siguiente tabla:

Tabla 1. Datos

Entrada x	**Salida y**
1,00	0,58
2,00	1,10
3,00	1,20
4,00	1,30
5,00	1,95
6,00	2,55
7,00	2,60
8,00	2,90
9,00	3,45
10,00	3,50
11,00	3,60
12,00	4,10
13,00	4,35
14,00	4,40
15,00	4,50

Fuente: MLB

i. Asuma que los datos pueden ser descritos por el modelo estadístico $y_t = \beta_0 + x_t\beta_1 + e_t$, donde la variable aleatoria $e_t \sim (0, \sigma^2)$. Utilice la regla de los cuadramos mínimos para estimar β_0 y β_1.

ii. Dé una interpretación económica de los parámetros estimados.
iii. Utilice los resultados del literal (a) para trazar la función de producción.
iv. Si el costo de alimentación es de 6 centavos por libra, obtenga las funciones del costo total y el costo marginal.

v. Muestre cómo utilizar la función del costo total (la cual relaciona el costo total con la salida) para obtener estimados de los parámetros de la función de producción conectando entrada de alimentación y salidas de carne de aves de corral.

2. Considere las siguientes 5 observaciones sobre $y_t = \{5,2,3,2,-2\}$ y $x_t = \{3,2,1,-1,0\}$.

   i. Encuentre $\sum_{t=1}^{5} x_t y_t$, $\sum_{t=1}^{5} x_t y_t$, $\sum_{t=1}^{5} y_t$, y $\sum_{t=1}^{5} y_t$.

   ii. Encuentre $\beta_1 = \frac{5\sum x_t y_t - \sum x_t \sum y_t}{5\sum x_t^2 - (\sum x_t)^2}$ y $b_0 = \bar{y} - b_1\bar{x}$.

   iii. Dé una interpretación de las cantidades que usted ha calculado.

   iv. Utilice la fórmula de derivación $\beta_1 = \frac{\sum(y_t - \bar{y})(x_t - \bar{x})}{\sum(x_t - \bar{x})^2}$ para estimar .

3. Pruebe la linealidad de la regresión en los ejercicios 1 y 2.

4. Se realiza un estudio sobre la cantidad de azúcar transformada en cierto proceso a varias temperaturas. Los datos se recolectan y se registran como sigue:

Temperatura, x	Azúcar transformada, y
1	8,1
1,1	7,8
1,2	8,5
1,3	9,8
1,4	9,5
1,5	8,9
1,6	8,6
1,7	1,02
1,8	9,3
1,9	9,2
2	10,5

Fuente: MLB

a. Estime la línea de regresión lineal.

b. Estime la cantidad media de azúcar transformada que se produce cuando la temperatura codificada es de 1,75.
c. Utilice un procedimiento de análisis de varianza para probar la hipótesis de que $\beta_0 = 0$ contra la alternativa de que $\beta_0 \neq 0$ a un nivel de significancia de 0,05.

5. A los estudiantes de primer año en un pequeño colegio se les aplica un examen de clasificación en matemáticas. Al estudiante que obtiene una calificación por debajo de 35 se le niega la admisión al curso regular de matemáticas y se le matricula en un grupo de regularización. Las calificaciones del examen de clasificación y las calificaciones finales de 20 estudiantes que toman el curso regular se registran a continuación:

Tabla 3. Calificaciones del examen de clasificación

Examen de clasificación	Calificación del curso	Examen de clasificación	Calificación del curso
50	53	90	54
35	41	80	91
35	61	60	48
40	56	60	71
55	68	60	71
65	36	40	47
35	11	55	53
60	70	50	68
90	79	65	57
35	59	50	79

Fuente: MLB

Con referencia a esta información, construya:

a. Un intervalo de confianza de 95 % para las calificaciones promedio del curso de estudiantes que obtienen 35 en el examen de colocación.
b. Un intervalo de predicción de 95 % para las calificaciones del curso de un estudiante que obtiene 35 en el examen de colocación.

8. Elementos de muestreo

ELEMENTOS DE MUESTREO

Se debe tener en cuenta que el muestreo parte de las definiciones básicas de universo de estudio, variables de interés y parámetros a investigar, y utiliza conceptos propios de la estadística. Algunos se han visto en capítulos anteriores y otros se presentan a continuación.

a. **Marco muestral**: es la lista o registro de las unidades de muestreo, es decir, el material o dispositivo usado para tener acceso a los elementos de la población de interés. Es la base sobre la cual deben diseñarse los procesos de selección.

El muestreo es una metodología que, apoyándose en la teoría estadística y de acuerdo con las características del estudio, indica cómo seleccionar y medir una parte de los elementos de la población (muestra) para hacer inferencia válida sobre el comportamiento global de la población. El muestreo puede ser probabilístico o no probabilístico.

En la teoría de muestreo se parte de un universo finito. Una muestra es el conjunto de elementos extraídos del universo, ya sea mediante un método sin reposición, en el cual las muestras son de tamaño menor o igual al universo, o mediante métodos con reposición, en los que es posible que las muestras sean mayores al universo.

b. **Representatividad de la muestra:** grado en el cual la muestra reproduce las características de la población.

c. **Diseño muestral**: todo diseño muestral comprende las siguientes partes:

i. Método de selección de la muestra
ii. Estimadores a utilizar y propiedades
iii. Determinación del tamaño de muestra
iv. Modificaciones al diseño básico

Una muestra sin reposición puede seleccionarse:

De manera aleatoria: garantiza inferencias estadísticas válidas y mejoramientos acumulativos a través de la separación y evaluación objetiva de sus fuentes de error
Por tablas de números aleatorios
Por computadora

8.1 Muestreo probabilístico

Una muestra es probabilística si:

a. Dispone de un marco muestral para los elementos del universo a ser seleccionados, denominados unidades de muestreo.
b. Todas las unidades de muestreo tienen una probabilidad conocida de antemano y una probabilidad mayor a cero de ser incluidas en una muestra.
c. El mecanismo de selección de la muestra corresponde a las probabilidades asignadas con anterioridad a cada objeto.

El muestreo probabilístico comprende el muestreo aleatorio simple, el muestreo sistemático, el muestreo estratificado, el muestro por conglomerados, entre otros. Cada uno de estos métodos de muestreo tiene un desarrollo teórico particular y la aplicación de cualquiera de ellos es conveniente realizarla con la asesoría de profesionales calificados de la estadística. Para un tratamiento detallado de los métodos de muestreo, se recomienda al lector consultar bibliografía especializada.

8.1.1 Muestreo aleatorio simple

En el caso particular en que cada uno de los elementos de la población tenga la misma probabilidad de integrar la muestra, y que todas las muestras de tamaño n tengan la misma probabilidad de ser elegidas, se dice que el muestreo es aleatorio simple y que las muestras son muestras aleatorias simples. El muestreo probabilístico y específicamente el aleatorio simple constituyen el soporte fundamental de la inferencia estadística. Por otro lado, en la teoría de probabilidades se utilizan muestras aleatorias en el desarrollo e ilustración de su formulación.

Para obtener una muestra aleatoria simple se enumeran las unidades de la población de 1 a N y posteriormente se extrae una serie de n números aleatorios entre 1 y N (tarea que se puede realizar usando una tabla de números aleatorios o mediante un programa de computación que produce una tabla semejante). Las unidades cuya enumeración coincide con la serie de números seleccionados conformarán la muestra aleatoria. En este esquema muestral, si una unidad muestral fue previamente seleccionada, entonces no puede ser seleccionada nuevamente (muestreo sin reposición). En cada extracción el proceso debe garantizar la misma oportunidad de selección a todos y cada uno de los elementos que no hayan sido seleccionados aún.

El muestreo aleatorio sin reemplazo de poblaciones finitas consiste en la selección de n elementos tomados de una población con N unidades, de modo que todas las posibles $\binom{N}{n}$ muestras de tamaño n tengan la misma probabilidad de ser seleccionadas.

La probabilidad de elegir cualquier muestra individual, digamos S, de n unidades viene dada por:

$$P(S) = \frac{1}{\binom{N}{n}} = \frac{1}{\frac{N!}{n!\,(N-n)!}} = \frac{n!\,(N-n)!}{N!} \qquad (8.1)$$

Este método de muestreo se usa en poblaciones suficientemente homogéneas, es decir, cuya varianza poblacional tienda a cero. Exige disponer una lista enumerada de 1 a N y mediante un experimento aleatorio seleccionar a cada uno de los n elementos de la muestra.

Dos factores afectan la cantidad de información contenida en la muestra y, por tanto, la precisión: tamaño de la muestra y cantidad de variación que se controla por el tipo de muestreo.

8.1.1.1 Estimación de la media

Suponga que $y_1, y_2, \ldots, y_n$ es una muestra aleatoria sin reposición de una población con media μ y varianza σ^2.

La estimación de la media es: $\bar{y} = \frac{\sum y_i}{n}$

La varianza para la muestra viene dada por:

$$S^2 = \frac{\sum y_i^2 - n\bar{y}^2}{n-1} \qquad (8.2)$$

Y la covarianza poblacional es diferente de cero, tal como se aprecia en la ecuación 8.4.

$$Cov(y_i, y_j) = -\frac{1}{N-1}\sigma^2 \qquad (8.3)$$

Para determinar el tamaño de la muestra cuando se quiere estimar la media poblacional, recordando lo visto en el capítulo cuatro sobre el intervalo de confianza para la estimación de la media poblacional, se tiene que:

$$\bar{Y} \mp t_{(\alpha/2)}\sqrt{\hat{V}(\bar{Y})} \qquad (8.4)$$

En este caso, $t_{(\alpha/2)}\sqrt{\hat{V}(\bar{Y})}$ es el error de estimación, el cual se representa como B. Luego, para estimar el tamaño de la muestra se despeja n de B, bien sea que se conozca la varianza poblacional a partir de la ecuación (8.5) o que no se conozca, en cuyo caso se usaría la ecuación (8.6).

$$e = B = t_{(\alpha/2)}\sqrt{\hat{V}(\bar{Y})} = t_\alpha . \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)} \qquad (8.5)$$

$$e = B = t_{(\alpha/2)}\sqrt{\hat{V}(\bar{Y})} = t_\alpha . \sqrt{\frac{\hat{S}^2}{n}\left(\frac{N-n}{N-1}\right)} \qquad (8.6)$$

Despejando n de la ecuación (8.5), el tamaño de la muestra se obtiene mediante:

$$n = \frac{N\sigma^2}{\left(\frac{e^2}{t_\alpha^2}\right)(N-1) + \sigma^2} \Longrightarrow n = \frac{N\sigma^2}{D\,(N-1) + \sigma^2} \qquad (8.7)$$

Para obtener el tamaño de la muestra cuando se está haciendo inferencia para el total poblacional ($\tau = N\mu$), se utiliza la expresión dada en (8.8).

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad (8.8)$$

Donde: $D = \frac{e^2}{t_\alpha^2\, N^2}$

Tabla 8.1 Estimador de la media, la varianza de la media, el total y la varianza del total bajo un muestreo aleatorio simple

	Media	Total	Proporción
Estimador	$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_i$	$\hat{\tau} = N\bar{y} = N.\frac{1}{n}\sum_{j=1}^{n} y_i$	$\hat{p} = \frac{1}{n}.\sum_{i=1}^{n} y_i$
Varianza	$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$	$\hat{V}(\hat{\tau}) = N^2\frac{s^2}{n}\left(\frac{N-n}{N}\right)$	$\hat{V}(\hat{p}) = \frac{\hat{p}.\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$
	$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$		
Tamaño de muestra	$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$ $D = \frac{B^2}{t_\alpha^2}$	$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}$ $D = \frac{B^2}{t_\alpha^2 N^2}$	$n = \frac{Npq}{(N-1)D + pq}$ $D = \frac{B^2}{t_\alpha^2}$

Fuente: MLB

El muestreo aleatorio simple tiene como ventajas:

a. El solo hecho de realizar muestreo.
b. La simplicidad en determinar la precisión de las estimaciones a partir de las observaciones muestrales.

c. La tendencia a reflejar las características del universo; esto es, cuando el tamaño de la muestra crece, esta se hace cada vez más representativa del universo o población.

Entre las desventajas del muestreo aleatorio simple se encuentran:

a. Se supone un listado completo.
b. Para poblaciones muy grandes, la enumeración de los elementos demanda tiempo y trabajo que se pudieran utilizar en otro diseño muestral.
c. Implica costos mayores con la dispersión espacial de las unidades muestreadas.

8.1.2 Muestreo sistemático

En el muestreo sistemático los elementos de la población se enumeran o se ordenan.

Elección de una muestra sistemática

Una muestra sistemática de 1 en k es la que se extrae de la siguiente forma:

a. Se selecciona aleatoriamente un elemento (llamado punto de inicio) de los primeros k elementos de la población.
b. Después se selecciona cada k-ésimo elemento hasta conseguir una muestra de tamaño n.
c. En general, k se toma como el número entero menor o igual que el cociente $\frac{n}{N}$:

$$k \leq \frac{n}{N}$$

De esta manera, nos podemos encontrar con las siguientes situaciones:

a. **k=n/N** es entero**.** Entonces se obtienen exactamente n observaciones.

Por ejemplo, si N = 100 y n = 5, entonces k = 20. Y aun tomando la última observación del primer intervalo (20º), obtenemos 5 observaciones: 20º, 40º,…, 100º.

b. **k=n/N** no es entero.

Por ejemplo, si N = 103 y n = 5, entonces n/N = 20,6 y tomamos k = 20.

Según el punto inicial nos podemos encontrar con estas situaciones:

a. Si elegimos, por ejemplo, el 2º como punto inicial, obtendríamos:

2º, 22º, 42º, 62º, 82º...

Al dividir la población en 5 intervalos de 20 elementos, sobran 3. Si no hay problema de costo podríamos elegir también el 102º y la muestra sería de tamaño 6.

b. Si se elige, por ejemplo, la observación 18º como la inicial, obtendríamos una muestra de tamaño 5:

18º, 38º, 58º, 78º, 98º

8.1.2.1 *N* es desconocido

En este caso, la decisión sobre el valor de k se tomará de forma que se asegure el número mínimo deseado de elementos de la muestra. Se deber tener en cuenta que N se estima por defecto, así k será menor de lo necesario y, por tanto, el tamaño muestral será mayor o igual de lo requerido.

Ventajas del muestreo sistemático frente al muestreo aleatorio simple (m.a.s.)

a. En la práctica, el muestreo sistemático es más fácil de llevar a cabo y está expuesto a menos errores del encuestador (en el m.a.s. se nos juntaría el trabajo si dos números aleatorios fueran consecutivos o muy próximos).

b. Con igual tamaño de muestra el muestreo sistemático proporciona más información que el muestreo aleatorio simple. Esto se debe a que la muestra sistemática se extiende uniformemente a lo largo de toda la población, mientras que en el muestreo aleatorio simple puede ocurrir que un gran número de observaciones se concentre en una zona y descuide otras.

Usos del muestreo sistemático:
Este tipo de muestreo es utilizado en los planes de muestreo para el control de calidad en el proceso de fabricación, por los auditores cuando se enfrentan a largas listas de apuntes para comprobar y por los investigadores de mercados cuando analizan personas en movimiento.

Tabla 8.2 Estimador de la media, la varianza de la media, el total y la varianza del total bajo un muestreo aleatorio sistemático

	Media	Total	Proporción
Estimador	$\hat{\mu} = \bar{y}_{sy} = \frac{1}{n}\sum_{j=1}^{n} y_i$	$\hat{\tau} = N.\bar{y}_{sy} = N.\frac{1}{n}\sum_{j=1}^{n} y_i$	$\hat{p}_{sy} = \frac{1}{n}.\sum_{i=1}^{n} y_i$
Varianza	$\hat{V}(\bar{y}_{sy}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$	$\hat{V}(\bar{y}_{sy}) = N^2\frac{s^2}{n}\left(\frac{N-n}{N}\right)$	$\hat{V}(\hat{p}_{sy}) = \frac{\hat{p}_{sy}.\hat{q}_{sy}}{n-1}\left(\frac{N-n}{N}\right)$
	$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{sy})^2}{n-1}$		
Tamaño de muestra	$n = \frac{N\sigma^2}{(N-1)D+\sigma^2}$ $D = \frac{B^2}{t_\alpha^2}$	$n = \frac{N\sigma^2}{(N-1)D+\sigma^2}$ $D = \frac{B^2}{t_\alpha^2 N^2}$	$n = \frac{Npq}{(N-1)D+pq}$ $D = \frac{B^2}{t_\alpha^2}$

Fuente: MLB

8.1.3 Muestreo estratificado

Es el procedimiento mediante el cual se escoge una muestra aleatoria estratificada. Ahora bien, una muestra aleatoria estratificada es la obtenida mediante la división de la población en subpoblaciones denominadas estratos. De cada estrato se selecciona en forma independiente una muestra aleatoria simple, tal como se muestra en la figura 8.1.

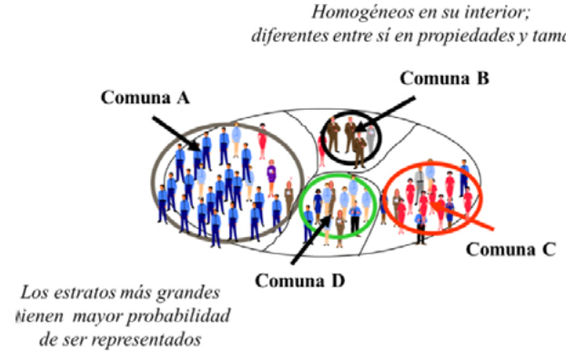


Figura 8.1 Muestreo estratificado
Fuente: adaptado de Universidad de Alicante (s.f.)

Las razones que tiene el investigador para estratificar son, entre otras:

a. Aumentar la precisión de las estimaciones al disminuir la variación de los estratos.
b. Disminuir los costos al estratificar y variar las fracciones de muestreo de los estratos.
c. Definir los estratos como dominios de estudio y obtener estimaciones con precisión conocida para los estratos.

Para seleccionar una muestra aleatoria estratificada, se debe dividir la población en estratos de acuerdo con las razones que se tengan para la segmentación. Luego hay que ubicar cada una de las unidades muestrales en sus respectivos estratos. Una vez realizado lo anterior, se debe asignar el tamaño muestral de cada estrato, digamos n_i, y por último seleccionar muestras aleatorias simples en cada estrato de manera independiente.

Tabla 8.3 Estimador de la media, la varianza de la media, el total y la varianza del total en un muestreo aleatorio estratificado

	Media	Total	Proporción
Estimador	$\hat{\mu} = \bar{y}_{est} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$ donde $\bar{y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i}$	$\hat{t}_{est} = N.\bar{y}_{est} = \sum_{i=1}^{L} N_i \bar{y}_i$	$\hat{p}_{est} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i$
Varianza	$\hat{V}(\bar{y}_{est}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{N_i - n_i}{N_i}.\frac{\hat{S}_i^2}{n_i}$	$\hat{V}(\hat{t}_{est}) = \sum_{i=1}^{L} N_i^2 \frac{N_i - n_i}{N_i}.\frac{\hat{S}_i^2}{n_i}$	$\hat{V}(\hat{p}_{est}) = \sum_{i=1}^{L} N_i^2 \frac{N_i - n_i}{N_i}.\frac{\hat{p}_i\hat{q}_i}{n_i - 1}$
	$\hat{S}_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n_i - 1}$		
Tamaño de muestra	$n = \frac{\sum W_i \frac{\sigma_i^2}{w_i}}{D + \frac{1}{N}\sum W_i \sigma_i^2}$ $W_i = \frac{N_i}{N}$ $D = \frac{B^2}{t_\alpha^2}$	$n = \frac{\sum N_i^2 \frac{\sigma_i^2}{w_i}}{D + \frac{1}{N}\sum N_i \sigma_i^2}$ $w_i = \frac{n_i}{n}$ $D = \frac{B^2}{t_\alpha^2.N^2}$	$n = \frac{\sum N_i^2 \frac{\hat{p}_i\hat{q}_i}{w_i}}{D + \frac{1}{N}\sum N_i \hat{p}_i\hat{q}_i}$ $D = \frac{B^2}{t_\alpha^2}$

Fuente: MLB

Se debe recalcar que según sea la afijación utilizada en el muestreo sistemático, a saber, proporcional, de Neyman u óptima, el tamaño de la muestra puede verse afectado.

8.1.4 Muestreo por conglomerado

Este tipo de muestreo se le conoce también como en etapas o multietápico. Se emplea cuando se desea estudiar una población grande y dispersa, y no se dispone de ningún listado para aplicar las técnicas anteriores. En lugar de seleccionar sujetos, se seleccionan subgrupos o conglomerados a los que se les da el nombre de unidades de primera etapa o unidades primarias. La diferencia con los estratos del tipo de muestreo anterior es que los conglomerados ya están agrupados de forma natural (hospitales, escuelas, etc.). Posteriormente, se seleccionan las unidades de la segunda etapa de manera aleatoria a partir de las unidades primarias.

Así sucesivamente, hasta llegar a las unidades de análisis, que serán los individuos que compongan la muestra de estudio.

Este muestreo, en numerosas ocasiones, es efectivo para obtener la información deseada a un menor costo, aunque el uso de los conglomerados conlleve, en algunos casos, a una varianza mayor de los estimadores.

Este diseño se justifica cuando:

a. Existe un alto costo por la movilización o traslado entre las unidades primarias, ya que permite disminuir las distancias. Por lo general, los conglomerados son áreas físicas o geográficas donde las unidades primarias están contiguas.
b. No existe una lista de las unidades primarias (o últimas) sobre las cuales hay que tomar las observaciones, y el costo de levantar un marco muestral de estas unidades es alto, en comparación con el costo de muestrear sobre conglomerados, los cuales sí pueden disponer de un marco o directorio.
c. Hay pequeñas unidades donde puede ser difícil fijar con precisión sus límites; sin embargo, puede ser posible y fácil dividir en unidades mayores y luego muestrear y medir aquellas unidades mayores seleccionadas. Ejemplo: animales.

La diferencia de objetivos entre estratificados y conglomerados conduce a diferentes criterios para establecer los conglomerados o los estratos. En contraste, con el estratificado, la varianza del estimador se hace pequeña al hacer el conglomerado, tanto como sea posible, representativo de la diversidad de toda población, y los conglomerados deben ser en lo posible construidos lo más semejantes entre sí. A diferencia del muestreo estratificado, donde los estratos deben ser homogéneos dentro de sí y heterogéneos entre sí.

8.2 Muestreo no probabilístico

El muestreo no probabilístico, también llamado muestreo subjetivo o por conveniencia, se caracteriza porque en la elección de los elementos

de la muestra interviene el conocimiento y la opinión de la persona que realiza la selección. Usualmente, la muestra se selecciona de acuerdo con la conveniencia y comodidad de la persona. En este tipo de muestreo se ubican muestreo por cuotas, muestreo bola de nieve, muestreo por juicio (opinión) y muestreo sin norma (por conveniencia).

Ejercicios

1. Defina las poblaciones adecuadas de las que se seleccionaron las siguientes muestras:
   i. Se llamó por teléfono a 200 personas en una ciudad y se les pidió nombrar al candidato por el que votarían en la elección para que sea el representante del consejo comunal.
   ii. Se lanzó 200 veces una moneda y se registraron 34 cruces.
   iii. Se probaron 200 pares de un nuevo tipo de tenis en un torneo profesional y, en promedio, duraron cuatro meses.
   iv. Una abogada tardó 21, 26, 24, 22 y 21 minutos en manejar de su casa en los suburbios a su oficina en el centro de la ciudad..

2. Una empresa eléctrica fabrica focos que tienen una duración aproximadamente distribuida de forma normal con una desviación estándar de 40 horas. Si una muestra de 30 focos tiene una duración promedio de 780 horas, ¿de qué tamaño se necesita una muestra si deseamos tener 96 % de confianza de que nuestra media muestra esté dentro de 10 horas de la media real?

3. Se fabrica cierto tipo de hilo con una resistencia a la tracción media de 78,3 kilogramos y una desviación estándar de 5,6 kilogramos. ¿Cómo cambia la varianza de la media muestra cuando el tamaño de la muestra:
   i. Aumenta de 64 a 196?
   ii. Disminuye de 784 a 49?

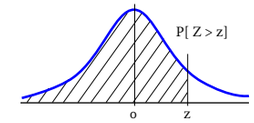
4. Si la desviación estándar de la media para la distribución muestral de muestras aleatorias de tamaño 36 de una población grande o infinita es 2, ¿qué tan grande debe ser el tamaño de la muestra si la desviación estándar se reduce a 1,2?

ANEXOS

ANEXOS

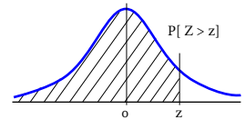
Anexo A

(Tabla de distribución normal)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Anexo A (Tabla de distribución normal)



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Anexo B

(Tabla de distribución chi-cuadrado)

Gdos de libertad	0,001	0,005	0,01	0,02	0,025	0,03	0,04	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4
1	10,83	7,879	6,635	5,412	5,024	4,709	4,218	3,841	2,706	2,072	1,642	1,323	1,074	0,873	0,708
2	13,82	10,6	9,21	7,824	7,378	7,013	6,438	5,991	4,605	3,794	3,219	2,773	2,408	2,1	1,833
3	16,27	12,84	11,34	9,837	9,348	8,947	8,311	7,815	6,251	5,317	4,642	4,108	3,665	3,283	2,946
4	18,47	14,86	13,28	11,67	11,14	10,71	10,03	9,488	7,779	6,745	5,989	5,385	4,878	4,438	4,045
5	20,52	16,75	15,09	13,39	12,83	12,37	11,64	11,07	9,236	8,115	7,289	6,626	6,064	5,573	5,132
6	22,46	18,55	16,81	15,03	14,45	13,97	13,2	12,59	10,64	9,446	8,558	7,841	7,231	6,695	6,211
7	24,32	20,28	18,48	16,62	16,01	15,51	14,7	14,07	12,02	10,75	9,803	9,037	8,383	7,806	7,283
8	26,12	21,95	20,09	18,17	17,53	17,01	16,17	15,51	13,36	12,03	11,03	10,22	9,524	8,909	8,351
9	27,88	23,59	21,67	19,68	19,02	18,48	17,61	16,92	14,68	13,29	12,24	11,39	10,66	10,01	9,414
10	29,59	25,19	23,21	21,16	20,48	19,92	19,02	18,31	15,99	14,53	13,44	12,55	11,78	11,1	10,47
11	31,26	26,76	24,72	22,62	21,92	21,34	20,41	19,68	17,28	15,77	14,63	13,7	12,9	12,18	11,53
12	32,91	28,3	26,22	24,05	23,34	22,74	21,79	21,03	18,55	16,99	15,81	14,85	14,01	13,27	12,58
13	34,53	29,82	27,69	25,47	24,74	24,12	23,14	22,36	19,81	18,2	16,98	15,98	15,12	14,35	13,64
14	36,12	31,32	29,14	26,87	26,12	25,49	24,49	23,68	21,06	19,41	18,15	17,12	16,22	15,42	14,69
15	37,7	32,8	30,58	28,26	27,49	26,85	25,82	25	22,31	20,6	19,31	18,25	17,32	16,49	15,73

Gdos de libertad	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,98	0,99	0,995
1	0,571	0,455	0,357	0,275	0,206	0,148	0,102	0,064	0,036	0,016	0,004	1E-03	6E-04	2E-04	4E-05
2	1,597	1,386	1,196	1,022	0,862	0,713	0,575	0,446	0,325	0,211	0,103	0,051	0,04	0,02	0,01
3	2,643	2,366	2,109	1,869	1,642	1,424	1,213	1,005	0,798	0,584	0,352	0,216	0,185	0,115	0,072
4	3,687	3,357	3,047	2,753	2,47	2,195	1,923	1,649	1,366	1,064	0,711	0,484	0,429	0,297	0,207
5	4,728	4,351	3,996	3,655	3,325	3	2,675	2,343	1,994	1,61	1,145	0,831	0,752	0,554	0,412
6	5,765	5,348	4,952	4,57	4,197	3,828	3,455	3,07	2,661	2,204	1,635	1,237	1,134	0,872	0,676
7	6,8	6,346	5,913	5,493	5,082	4,671	4,255	3,822	3,358	2,833	2,167	1,69	1,564	1,239	0,989
8	7,833	7,344	6,877	6,423	5,975	5,527	5,071	4,594	4,078	3,49	2,733	2,18	2,032	1,646	1,344
9	8,863	8,343	7,843	7,357	6,876	6,393	5,899	5,38	4,817	4,168	3,325	2,7	2,532	2,088	1,735
10	9,892	9,342	8,812	8,295	7,783	7,267	6,737	6,179	5,57	4,865	3,94	3,247	3,059	2,558	2,156
11	10,92	10,34	9,783	9,237	8,695	8,148	7,584	6,989	6,336	5,578	4,575	3,816	3,609	3,053	2,603
12	11,95	11,34	10,76	10,18	9,612	9,034	8,438	7,807	7,114	6,304	5,226	4,404	4,178	3,571	3,074
13	12,97	12,34	11,73	11,13	10,53	9,926	9,299	8,634	7,901	7,042	5,892	5,009	4,765	4,107	3,565
14	14	13,34	12,7	12,08	11,45	10,82	10,17	9,467	8,696	7,79	6,571	5,629	5,368	4,66	4,075
15	15,02	14,34	13,68	13,03	12,38	11,72	11,04	10,31	9,499	8,547	7,261	6,262	5,985	5,229	4,601

Anexo C

(Tabla de distribución t-student)

	Q	0,4	0,25	0,1	0,05	0,025	0,013	0,006	0,003	0,001	0,0005
Gdos de libertad	2Q	0,8	0,5	0,2	0,1	0,05	0,025	0,013	0,005	0,003	0,001
1		0,325	1,000	3,078	6,314	12,706	25,452	50,92	127,32	254,65	636,62
2		0,289	0,816	1,886	2,920	4,303	6,205	8,860	14,089	19,962	31,599
3		0,277	0,765	1,638	2,353	3,182	4,177	5,392	7,453	9,465	12,924
4		0,271	0,741	1,533	2,132	2,776	3,495	4,315	5,598	6,758	8,610
5		0,267	0,727	1,476	2,015	2,571	3,163	3,810	4,773	5,604	6,869
6		0,265	0,718	1,440	1,943	2,447	2,969	3,521	4,317	4,981	5,959
7		0,263	0,711	1,415	1,895	2,365	2,841	3,335	4,029	4,595	5,408
8		0,262	0,706	1,397	1,860	2,306	2,752	3,206	3,833	4,334	5,041
9		0,261	0,703	1,383	1,833	2,262	2,685	3,111	3,690	4,146	4,781
10		0,260	0,700	1,372	1,812	2,228	2,634	3,038	3,581	4,005	4,587
11		0,260	0,697	1,363	1,796	2,201	2,593	2,981	3,497	3,895	4,437
12		0,259	0,695	1,356	1,782	2,179	2,560	2,934	3,428	3,807	4,318
13		0,259	0,695	1,356	1,782	2,179	2,560	2,934	3,428	3,807	4,318
14		0,258	0,692	1,345	1,761	2,145	2,510	2,864	3,326	3,675	4,140
15		0,258	0,691	1,341	1,753	2,131	2,490	2,837	3,286	3,624	4,073
16		0,258	0,690	1,337	1,746	2,120	2,473	2,813	3,252	3,581	4,015
17		0,257	0,689	1,333	1,740	2,110	2,458	2,793	3,222	3,543	3,965
18		0,257	0,688	1,330	1,734	2,101	2,445	2,775	3,197	3,510	3,922
19		0,257	0,688	1,328	1,729	2,093	2,433	2,759	3,174	3,481	3,883
20		0,257	0,687	1,325	1,725	2,086	2,423	2,744	3,153	3,455	3,850
21		0,257	0,686	1,323	1,721	2,080	2,414	2,732	3,135	3,432	3,819
22		0,256	0,686	1,321	1,717	2,074	2,405	2,720	3,119	3,412	3,792
23		0,256	0,685	1,319	1,714	2,069	2,398	2,710	3,104	3,393	3,768
24		0,256	0,685	1,318	1,711	2,064	2,391	2,700	3,091	3,376	3,745
25		0,256	0,684	1,316	1,708	2,060	2,385	2,692	3,078	3,361	3,725
30		0,256	0,683	1,310	1,697	2,042	2,360	2,657	3,030	3,300	3,646
40		0,255	0,681	1,303	1,684	2,021	2,329	2,616	2,971	3,227	3,551
60		0,254	0,679	1,296	1,671	2,000	2,299	2,575	2,915	3,156	3,460
120		0,254	0,677	1,289	1,658	1,980	2,270	2,536	2,860	3,088	3,373

Anexo D

(Tabla de distribución F con α = 0,01)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	60,473	60,705	60,903	61,073	61,220	61,350
2	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,222	5,216	5,210	5,205	5,200	5,196
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,907	3,896	3,886	3,878	3,870	3,864
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,282	3,268	3,257	3,247	3,238	3,230
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,920	2,905	2,892	2,881	2,871	2,863
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,684	2,668	2,654	2,643	2,632	2,623
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,519	2,502	2,488	2,475	2,464	2,455
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,396	2,379	2,364	2,351	2,340	2,329
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,302	2,284	2,269	2,255	2,244	2,233
11	3,225	2,860	2,660	2,536	2,451	2,389	2,342	2,304	2,274	2,248	2,227	2,209	2,193	2,179	2,167	2,156
12	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,166	2,147	2,131	2,117	2,105	2,094
13	3,136	2,763	2,560	2,434	2,347	2,283	2,234	2,195	2,164	2,138	2,116	2,097	2,080	2,066	2,053	2,042
14	3,102	2,726	2,522	2,395	2,307	2,243	2,193	2,154	2,122	2,095	2,073	2,054	2,037	2,022	2,010	1,998
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,037	2,017	2,000	1,985	1,972	1,961
16	3,048	2,668	2,462	2,333	2,244	2,178	2,128	2,088	2,055	2,028	2,005	1,985	1,968	1,953	1,940	1,928
17	3,026	2,645	2,437	2,308	2,218	2,152	2,102	2,061	2,028	2,001	1,978	1,958	1,940	1,925	1,912	1,900
18	3,007	2,624	2,416	2,286	2,196	2,130	2,079	2,038	2,005	1,977	1,954	1,933	1,916	1,900	1,887	1,875
19	2,397	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,990	2,990	2,990	2,990	2,990	2,990	2,990
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,913	1,892	1,875	1,859	1,845	1,833
21	2,961	2,575	2,365	2,233	2,142	2,075	2,023	1,982	1,948	1,920	1,896	1,875	1,857	1,841	1,827	1,815
22	2,949	2,561	2,351	2,219	2,128	2,060	2,008	1,967	1,933	1,904	1,880	1,859	1,841	1,825	1,811	1,798
23	2,937	2,549	2,339	2,207	2,115	2,047	1,995	1,953	1,919	1,890	1,866	1,845	1,827	1,811	1,796	1,784
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,680	1,657	1,637	1,619	1,603	1,589
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,625	1,601	1,580	1,562	1,545	1,530

Anexo D
(Tabla de distribución F con α = 0,01)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882	242,983	243,906	244,690	245,364	245,950	246,464
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396	19,405	19,413	19,419	19,424	19,429	19,433
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,763	8,745	8,729	8,715	8,703	8,692
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912	5,891	5,873	5,858	5,844
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678	4,655	4,636	4,619	4,604
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000	3,976	3,956	3,938	3,922
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575	3,550	3,529	3,511	3,494
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,313	3,284	3,259	3,237	3,218	3,202
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073	3,048	3,025	3,006	2,989
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913	2,887	2,865	2,845	2,828
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788	2,761	2,739	2,719	2,701
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687	2,660	2,637	2,617	2,599
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604	2,577	2,554	2,533	2,515
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534	2,507	2,484	2,463	2,445
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475	2,448	2,424	2,403	2,385
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,456	2,425	2,397	2,373	2,352	2,333
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,413	2,381	2,353	2,329	2,308	2,289
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,374	2,342	2,314	2,290	2,269	2,250
19	2,895	9,277	9,277	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552	9,552
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,310	2,278	2,250	2,225	2,203	2,184
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,366	2,321	2,283	2,250	2,222	2,197	2,176	2,156
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342	2,297	2,259	2,226	2,198	2,173	2,151	2,131
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320	2,275	2,236	2,204	2,175	2,150	2,128	2,109
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,952	1,917	1,887	1,860	1,836	1,815
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,869	1,834	1,803	1,775	1,750	1,728

Anexo E
(Tabla de distribución F con α = 0,10)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	60,473	60,705	60,903	61,073	61,220	61,350
2	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526	8,526
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,222	5,216	5,210	5,205	5,200	5,196
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,907	3,896	3,886	3,878	3,870	3,864
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,282	3,268	3,257	3,247	3,238	3,230
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,920	2,905	2,892	2,881	2,871	2,863
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,684	2,668	2,654	2,643	2,632	2,623
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,519	2,502	2,488	2,475	2,464	2,455
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,396	2,379	2,364	2,351	2,340	2,329
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,302	2,284	2,269	2,255	2,244	2,233
11	3,225	2,860	2,660	2,536	2,451	2,389	2,342	2,304	2,274	2,248	2,227	2,209	2,193	2,179	2,167	2,156
12	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,166	2,147	2,131	2,117	2,105	2,094
13	3,136	2,763	2,560	2,434	2,347	2,283	2,234	2,195	2,164	2,138	2,116	2,097	2,080	2,066	2,053	2,042
14	3,102	2,726	2,522	2,395	2,307	2,243	2,193	2,154	2,122	2,095	2,073	2,054	2,037	2,022	2,010	1,998
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	2,037	2,017	2,000	1,985	1,972	1,961
16	3,048	2,668	2,462	2,333	2,244	2,178	2,128	2,088	2,055	2,028	2,005	1,985	1,968	1,953	1,940	1,928
17	3,026	2,645	2,437	2,308	2,218	2,152	2,102	2,061	2,028	2,001	1,978	1,958	1,940	1,925	1,912	1,900
18	3,007	2,624	2,416	2,286	2,196	2,130	2,079	2,038	2,005	1,977	1,954	1,933	1,916	1,900	1,887	1,875
19	2,397	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,606	2,990	2,990	2,990	2,990	2,990	2,990	2,990
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,913	1,892	1,875	1,859	1,845	1,833
21	2,961	2,575	2,365	2,233	2,142	2,075	2,023	1,982	1,948	1,920	1,896	1,875	1,857	1,841	1,827	1,815
22	2,949	2,561	2,351	2,219	2,128	2,060	2,008	1,967	1,933	1,904	1,880	1,859	1,841	1,825	1,811	1,798
23	2,937	2,549	2,339	2,207	2,115	2,047	1,995	1,953	1,919	1,890	1,866	1,845	1,827	1,811	1,796	1,784
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,680	1,657	1,637	1,619	1,603	1,589
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,625	1,601	1,580	1,562	1,545	1,530

REFERENCIAS

Armas, José M. (2000) *Estadística Sencilla: Inferencia*, Consejo de Publicaciones, Universidad de Los Andes, Mérida, Venezuela

Cari Mogrovejo, L. H. (2011). Estadística para la investigación (sesión 5). Recuperado de https://es.slideshare.net/zarlenin/estadistica-para-la-investigacin-sesin5-version-mejorable.

Farfán, Johnny, Distribución chi-cuadrada. Recuperado de https://www.google.com.co/search?q=distribucion+f+grados+de+libertad&espv=2&biw=1440&bih=794&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjnx4nxveDRAhUD6SYKHVDSABIQ_AUIBigB#imgrc=7VgpzBk2EUzN7M%3A

Gujarati, D. N., & Porter, D. C. (2011). *Econometría básica*-5. AMGH Editora.

Levine, D. M., Berenson, M. L. y Krehbiel, T. C. (2006). *Estadística para administración*. EUA. Pearson Educación.

Lohr, S. (2000). *Muestreo: diseño y análisis*. México: International Thomson Editores.

Meyer, P. (1992). *Probabilidad y aplicaciones estadísticas*. México: Addison-Wesley.

Montgomery, D. C., Runger, G. C. y Medal, E. G. U. (1996). *Probabilidad y estadística aplicadas a la ingeniería*. EUA, McGraw Hill.

Mood, A. M., Graybill, F. A. y Boes, D. C. (1974). *Introduction to the theory of statistics*. Sigapore: McGraw-Hill.

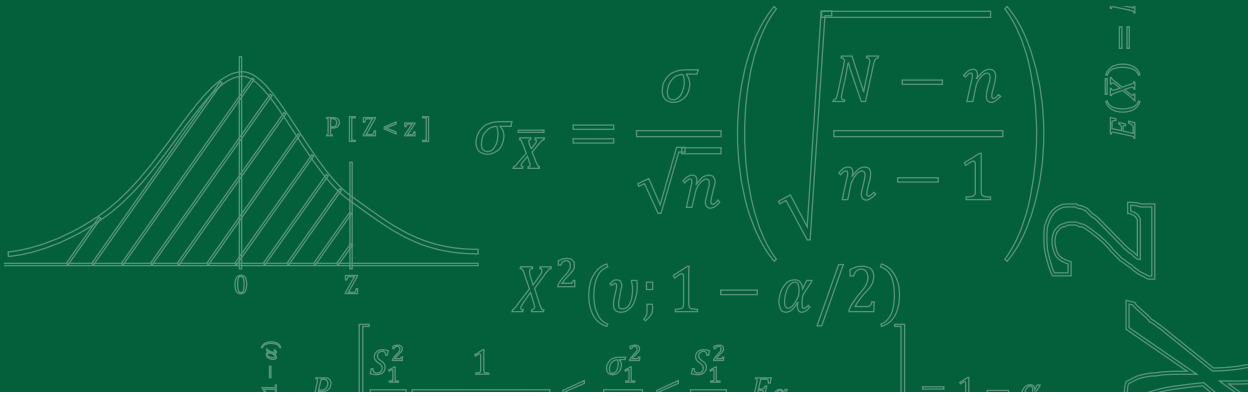
Moore, D. S. (1995). *Estadística aplicada básica*. Barcelona: Bosch.

Scheaffe, R., Mendenhall, W. y Ott, L. (1991) Elementos de muestreo. Boston: Duxbury Press.

Universidad de Alicante (s.f.). Técnicas de investigación social. Muestreo aleatorio estratificado. Recuperado de https://sites.google.com/site/tecninvestigacionsocial/temas-y-contenidos/tema-3-las-tecnicas-distributivas-la-investigacion-cuantitativa-y-la-encuesta/seleccion-de-los-casos-muestreos-probabilisticos/tipos-de-muestreo-probabilistico/muestreo-aleatorio-estratificado.

Walpole, R. E., Myers, R. H. y Myers, S. L. (1999). *Probabilidad y estadística para ingenieros*. EUA, Pearson Educación.

Esta obra se publicó en archivo digital
en el mes de mayo de 2018.



El presente manual aborda el contenido programático de la asignatura Inferencia Estadística de una manera clara y sencilla permitiendo al estudiante vencer las dificultades que con frecuencia se les suelen presentar.

El trabajo está constituido por ocho capítulos, siguiendo por supuesto, el contenido programático de la materia en cuestión. Se comienza haciendo un recorrido por aquellas distribuciones continuas necesarias para la estimación de los parámetros tales como: la media poblacional, la varianza poblacional, la proporción poblacional, entre otros.

El capítulo dos hace mención a la distribución en el muestreo y al teorema fundamental en estadística como lo es el Teorema Central del Límite. El capítulo tres aborda los conceptos necesarios en la inferencia estadística como lo son: los parámetros, estadísticas, estimadores, propiedades de los estimadores.

En el capítulo cuatro se desarrollan los procedimientos de inferencia estadística para una sola media poblacional cuando el tamaño de muestra es grande o pequeño. En el capítulo cinco se analizan los procedimientos de inferencia estadística para una y dos proporciones poblacionales utilizando la distribución normal. El capítulo seis se relaciona con los procesos de inferencia estadística para una varianza y dos varianzas poblacionales. En el capítulo siete se analizan situaciones donde intervienen dos variables cuantitativas con el fin de observar las relaciones existentes entre ellas, a través de dos técnicas: la regresión y la correlación y para finalizar el capítulo ocho, hace referencia a los elementos de muestreo, tipos de muestreo probabilístico y no probabilístico.

En cada uno de estos capítulos se presentan, además de la teoría básica, ejemplos adecuados para reforzar lo expuesto en el mismo.

ISBN: 978-958-764-531-6