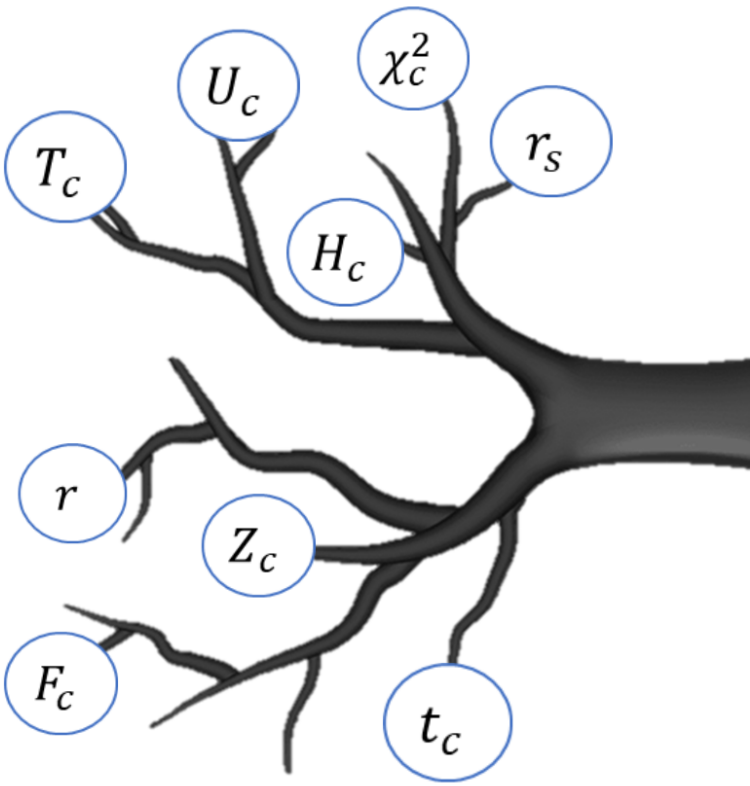
Métodos estadísticos

Guía básica para el uso de la estadística inferencial en investigación



Jorge Sucasaire Pilco

Rodolfo Ticona Vilcapaza

Jorge Sucasaire Pilco

Docente egresado de la Facultad de Educación de la UNMSM con especialidad en Matemática y Física. Maestro y doctor por la Universidad César Vallejo. Ha dictado cursos de Matemática y Estadística en el Instituto Superior Tecnológico Paul Müller, Instituto de Educación Superior Avansys (Cibertec), Instituto de Formación Bancaria IFB Certus, Instituto Superior IDAT, Centro Preuniversitario de la Universidad Autónoma del Perú, Universidad Privada César Vallejo, Universidad Tecnológica del Perú, Universidad Privada del Norte y la Universidad Le Cordon Bleu.

Rodolfo Ticona Vilcapaza

Docente Licenciado de la Facultad de Educación de la UNMSM con especialidad en Matemática y Física. Egresado de la Facultad de Ingeniería de Sistemas de la UNTELS. Ha dictado cursos de Matemática y Estadística a nivel superior en el Instituto ADEX, Centro Preuniversitario de la Universidad de Ingeniería y Tecnología (UTEC), Universidad Tecnológica del Perú (UTP), Universidad Peruana de Ciencias Aplicadas (UPC), Universidad Privada del Norte (UPN). Ha coordinado el curso de Matemática en la Universidad Peruana Cayetano Heredia.

Jorge Sucasaire Pilco

Rodolfo Ticona Vilcapaza

MÉTODOS ESTADÍSTICOS

Guía básica para el uso de la estadística inferencial en investigación

Jorge Sucasaire Pilco, editor

©Jorge Sucasaire Pilco
©Rodolfo Ticona Vilcapaza

Primera edición digital – febrero 2023

Editado por:
©Jorge Sucasaire Pilco
Jr. Los Lirios 380, Urb. La Primavera – El Agustino, Lima - Perú.
E-mail: jorgesuca1982@gmail.com

Libro electrónico disponible en:
https://repositorio.concytec.gob.pe/

Hecho el Depósito Legal en la Biblioteca Nacional del Perú N° 2023-01973

ISBN 978-612-00-8393-2

MÉTODOS ESTADÍSTICOS
Guía básica para el uso de la estadística inferencial en investigación

Diseño de portada: Jorge Sucasaire Pilco

No se permite reproducir, almacenar en sistemas de recuperación de la información ni transmitir parte o el total de esta publicación, cualquiera que sea el medio empleado, sin el permiso previo de los autores o del editor

Lima-Perú

Presentación

Generalmente, en la literatura estadística, se trata de presentar una ruta completa que muestre al investigador como seleccionar el método adecuado para su trabajo de investigación. Estos libros muestran una gran cantidad de fórmulas, pruebas estadísticas y las condiciones que deben verificarse antes de la aplicación del método seleccionado. Sin embargo, los ejemplos que se muestran, en la mayoría de textos, no están orientados a ningún esquema de presentación y análisis de resultados, de modo que, no queda clara la conexión entre los datos que ha recolectado el investigador y la forma de utilizarlos en la estimación de parámetros o la prueba de hipótesis de la investigación.

El aporte de este texto se centra en brindar ejemplos que no solo permiten comprender la aplicación de los métodos estadísticos, sino también se acompaña cada situación con su respectiva estructura metodológica. Es decir, a cada ejemplo de aplicación se le agrega de manera tentativa el título del trabajo de investigación, la pregunta de investigación, el objetivo, la hipótesis, el tipo de investigación y las variables respectivas.

Los ejemplos se explican de manera sencilla y en muchos casos se utiliza información real de fuentes verificables para hacer el contraste de las hipótesis. Los ejemplos se acompañan de una pequeña matriz de datos que servirá para que el lector pueda verificar por sí mismo, los resultados que se muestran en cada capítulo.

Los procedimientos realizados permiten ver, en la mayoría de los casos, el desarrollo completo del método estadístico a través del uso de diferentes expresiones matemáticas. Esto se complementa con resultados obtenidos por medio de programas estadísticos como Excel, SPSS o Minitab. En cada caso, se muestra la ruta para el acceso al método estadístico según el programa seleccionado.

Los resultados que se muestran en los diferentes ejemplos corresponden a la estimación de intervalos de confianza o a la prueba de hipótesis, en algunos casos se muestra la estadística descriptiva pero solo como un aspecto complementario, es decir, no se aborda al análisis descriptivo con detalle. Quienes deseen complementar el desarrollo de la parte descriptiva pueden revisar el libro "*Estadística descriptiva para trabajos de investigación: presentación e interpretación de los resultados*".

Los temas del texto se desarrollan tratando de usar de forma mínima las expresiones matemáticas y la simbología no es estrictamente formal. Este último aspecto no reduce la rigurosidad de los métodos estadísticos, sino que permite que los resultados se muestren de manera más cercana al lenguaje y entendimiento del lector.

En este texto no se abordan todos los métodos estadísticos, sino que, como se indica en el título, se exponen los métodos básicos de la estadística paramétrica y no paramétrica. De modo que, en los diferentes capítulos se explicará sobre la estimación por intervalos de confianza, la prueba de hipótesis para la media, la prueba de hipótesis para la proporción, la prueba t, el ANOVA para muestras independientes, el coeficiente de correlación lineal de Pearson, la regresión lineal simple, la prueba de Mann-Whitney, la prueba de Wilcoxon, el coeficiente de correlación de Spearman, la prueba de Chi cuadrado, la prueba exacta de Fisher y la prueba de Kruskall-Wallis.

Jorge Sucasaire Pilco
Rodolfo Ticona Vilcapaza

Contenido

Capítulo 1: Aspectos metodológicos 11

1.1. ¿En qué etapa de la investigación nos ubicamos? 13

1.2. Diseño metodológico o metodología 14

1.3. Los métodos que utilizaremos ¿son útiles para todo tipo de investigaciones? 15

1.4. ¿En todas las investigaciones se deben formular hipótesis? 16

1.5. ¿Las investigaciones descriptivas pueden tener hipótesis? 17

1.6. Muestreo probabilístico 17

Capítulo 2: Estimación estadística 20

2.1. Estadígrafos y parámetros 20

2.2. Estimación puntual 22

2.3. Estimadores y sesgo 22

2.4. Estimación por intervalo 24

2.4.1. Nivel de confianza $(1 - \alpha)$ 25

2.4.2. Distribución de probabilidad 25

2.4.3. Distribución muestral 29

2.4.4. Estandarización de valores 32

2.5. Intervalo de confianza bilateral para la media μ 34

2.5.1. Intervalo de confianza para la media poblacional con σ conocida 35

2.5.2. Intervalo de confianza para la media poblacional con σ desconocida 38

2.5.3. Intervalo de confianza para la proporción (π) 45

Capítulo 3: Prueba de hipótesis con respecto a la media 50

3.1. Hipótesis estadística 50

3.2. Hipótesis nula $(H0)$ 50

3.3. Hipótesis alterna $(H1)$ 51

3.4. Prueba de hipótesis 52

3.5. Nivel de significancia (α) 53

3.6. Región de rechazo 54

3.7. El p valor 56
3.8. Prueba de hipótesis para la media 57
3.8.1. Prueba unilateral de cola inferior 58
3.8.2. Prueba unilateral de cola superior 64
3.8.3. Prueba bilateral o de dos colas 70
3.8.4. Prueba bilateral para muestras pequeñas 75
3.9. Prueba de hipótesis para la diferencia de medias 82
Capítulo 4: Prueba de hipótesis con respecto a la proporción 87
4.1. Prueba de la cola inferior 88
4.2. Prueba de la cola superior 94
4.3. Prueba bilateral o de dos colas 100
4.4. Prueba de hipótesis para la diferencia de proporciones 106
Capítulo 5: Prueba t 110
5.1. Prueba de normalidad 110
5.2. Prueba de homocedasticidad 112
5.3. Prueba t para muestras independientes 113
5.3.1. Si no se cumplen los supuestos ¿Qué podemos hacer? 125
5.4. Prueba t para muestras relacionadas 126
5.4.1. Si no se cumplen los supuestos ¿Qué se puede hacer? 135
Capítulo 6: Análisis de la varianza 136
6.1. ANOVA de un solo factor 136
6.2. ANOVA de un solo factor para muestras independientes 137
6.2.1. Prueba F 138
6.2.2. Comparaciones post hoc a la prueba F 140
6.2.3. ¿Qué sucede cuando no se cumplen los supuestos? 154
Capítulo 7: Coeficiente de correlación lineal de Pearson 155
7.1. Diagrama de dispersión 156
7.2. Coeficiente de correlación lineal de Pearson (r) 159
7.2.1. Interpretación del coeficiente de correlación lineal 160
7.2.2. Supuestos o condiciones 162

7.2.3. Prueba de hipótesis para el coeficiente de correlación de Pearson 162
7.2.4. ¿Qué sucede cuando no se cumplen los supuestos? 171
Capítulo 8: Regresión Lineal 172
8.1. Regresión lineal simple 172
8.2. Recta de regresión lineal 174
8.3. Error estándar de estimación (Syx) 176
8.4. Significancia de los coeficientes de la ecuación de la recta de regresión lineal 176
8.5. Condiciones para validar el modelo de regresión lineal 177
Capítulo 9: Prueba U de Mann-Whitney 187
9.1. Pruebas no paramétricas 187
9.2. Prueba U de Mann-Whitney 188
Capítulo 10: Prueba de Wilcoxon 198
Capítulo 11: Coeficiente de correlación lineal de Spearman 209
11.1. Interpretación del coeficiente rs 210
Capítulo 12: Prueba Chi cuadrado 218
12.1. Chi cuadrado de Bondad de ajuste 218
12.2. Chi cuadrado de independencia 219
12.2.1. Tabla de contingencia 219
12.2.2. Frecuencia Observada 220
12.2.3. Frecuencia Esperada 220
12.2.4. El nivel de significancia para la prueba de hipótesis con el uso de software 224
12.2.5. Medidas del grado asociación 224
Capítulo 13: Prueba de Fisher 244
Capítulo 14: Prueba de Kruskal-Wallis 254
APÉNDICE 260
Referencias bibliográficas 262

Capítulo 1: Aspectos metodológicos

Luego del proceso de recolección de datos, a través de un instrumento válido y con base en una muestra representativa, se procede a describir los resultados. Por ejemplo, si se ha recolectado información sobre la estatura entre los estudiantes varones y mujeres de un colegio, los datos recolectados se pueden presentar de la siguiente manera:

Tabla 1.1 Matriz de datos recolectados.

Estatura en metros	
Mujeres	Varones
1.66	1.70
1.58	1.75
1.60	1.72
1.60	1.72
1.56	1.65
1.60	1.68
1.65	1.70
1.70	1.65
1.62	1.68
1.65	1.75

De la tabla 1.1 no se puede establecer una afirmación concluyente respecto a la variable estatura. Pero, podemos utilizar la estadística descriptiva para determinar el promedio y la desviación estándar.

Tabla 1.2 Estadísticos descriptivos.

Grupo	Promedio	Desviación Estándar
Mujeres	1.622	0.042
Varones	1.700	0.036

En la tabla 1.2 se observa que la estatura promedio de los varones (1.7 m) es superior a la estatura promedio de las mujeres (1.622 m). También se puede observar que las estaturas de los varones presentan menor

variabilidad (0.036 m) en comparación con la variabilidad de las estaturas de las mujeres (0.042 m).

Pero, la descripción anterior, es válida solo para la muestra (20 estudiantes analizados). Obviamente, el colegio tiene más de 20 alumnos, entonces ¿Qué se puede afirmar sobre la estatura de varones y mujeres de la población de donde proviene la muestra? Para responder a esta pregunta debemos introducirnos el campo de la estadística inferencial.

Para acercarnos a la estatura de la población de varones y mujeres podríamos calcular intervalos de confianza con los cuales limitar el valor de la media poblacional de estaturas. Si queremos comparar los promedios de estatura de la población de varones y la población de mujeres, se puede utilizar la prueba t. Sin embrago, el uso de esta prueba depende de ciertas condiciones. De no cumplirse estas condiciones se puede usar la prueba de Mann-Whitney, en ese caso no se comparan los promedios, sino las medianas.

También se pueden clasificar la estatura en las categorías: alta y baja; de modo que podemos agrupar los datos en una tabla de doble entrada.

Tabla 1.3 Tabla de contingencia para la estatura y el género.

		Género	
		Varones	Mujeres
Estatura	Baja		
	Alta		

Con los datos agrupados en categorías (tabla 1.3), y con la finalidad de determinar si existe asociación entre el género y la estatura, se puede utilizar una prueba de chi cuadrado, de cumplirse ciertas condiciones. Otra opción, dada la cantidad de filas y columnas, es utilizar la prueba exacta de Fisher.

Con lo expuesto, el lector ya debe tener claro que la elección del método estadístico no es una tarea sencilla. Se puede preguntar, entonces, ¿Qué método utilizar? Por ello, el los capítulos siguientes, se brinda información importante que todo investigador debe tener en cuenta para seleccionar la prueba o método estadístico pertinente.

1.1. ¿En qué etapa de la investigación nos ubicamos?

La investigación científica se desarrolla a través de una serie de etapas secuenciales enlazadas coherentemente. Esto implica que no podemos saltarnos ningún paso dentro del proceso de investigación. El uso de los métodos estadísticos aparece en la etapa del análisis de resultados, pero el método específico a utilizar se menciona dentro del diseño metodológico como la herramienta que permitirá extraer conclusiones a partir de los datos recolectados.

Tamayo (2003) muestra un esquema general sobe los pasos a seguir para elaborar una investigación.

Tabla 1.4 Estructura de la investigación científica según Tamayo (2003).

Pasos	Descripción
Elección del tema	Planteamiento
Delimitación del tema	Revisión del conocimiento, alcances, límites y recursos.
Problema	Identificación, descripción y recursos.
Objetivos	Generales y específicos.
Marco teórico	Antecedentes, definición conceptual, hipótesis y variables.
Metodología	Población, muestra, recolección y procesamiento de datos.
Informe	Resultados discusión, conclusiones y recomendaciones.

En la tabla 1.4 podemos ubicar el desarrollo de los métodos estadísticos en la parte de los resultados, aquí se desarrolla el análisis descriptivo e inferencial.

1.2. Diseño metodológico o metodología

Cuando mencionamos a la metodología en una investigación nos referimos a una etapa muy importante en el proceso de investigación. Morán, G. y Alvarado, D. (2010) consideran que el diseño de investigación hace referencia a las estrategias concebidas para obtener información pertinente.

Quezada et al. (2018) señalan que existen etapas bien definidas en la investigación. A continuación, se muestra una parte del esquema que elaboraron para explicar el proceso de la investigación.

Tabla 1.5 Algunas etapas de la investigación.

Etapa	Descripción	
Metodología	Establecimiento del diseño de la investigación.	Métodos. Técnicas. Modalidad de la investigación. Nivel o tipo de investigación.
	Determinación de la población y muestra. Operacionalización de los indicadores de las variables. Plan de recolección y procesamiento de datos.	
Análisis de datos	Recopilación de datos. Codificación y tabulación de datos. Explicación e interpretación de resultados. Comprobación de la hipótesis.	

Existen diferentes protocolos parta la presentación de la metodología de investigación. Sin embargo, podemos generalizar esta etapa de la investigación mediante los siguientes procesos:

- Selección del tipo de investigación.
- Selección de la población y la muestra.

- Método de recolección de la información.
- Método de procesamiento y análisis de datos.

La selección del método estadístico como parte del diseño de investigación se relaciona con el problema, los objetivos y las hipótesis de investigación. Estos elementos están íntimamente relacionados y deben definirse adecuadamente.

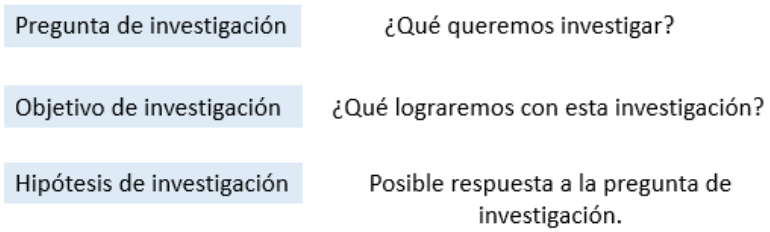


Figura 1.1 Elementos básicos para el desarrollo metodológico.

Usando el criterio de la figura 1.1 podemos elaborar el siguiente ejemplo:

Pregunta de investigación: ¿Qué relación existe entre el nivel de estrés y el rendimiento académico en los estudiantes del colegio X?

Objetivo de investigación: conocer la relación entre el nivel de estrés y el rendimiento académico en los estudiantes del colegio X.

Hipótesis de investigación: existe una relación inversa entre el nivel de estrés y el rendimiento académico en los estudiantes del colegio X.

1.3. Los métodos que utilizaremos ¿son útiles para todo tipo de investigaciones?

Hernández et al. (2010) señala que existen tres enfoques de investigación: cuantitativo, cualitativo y mixto. El contenido de este texto se orienta al desarrollo de investigaciones con enfoque cuantitativo.

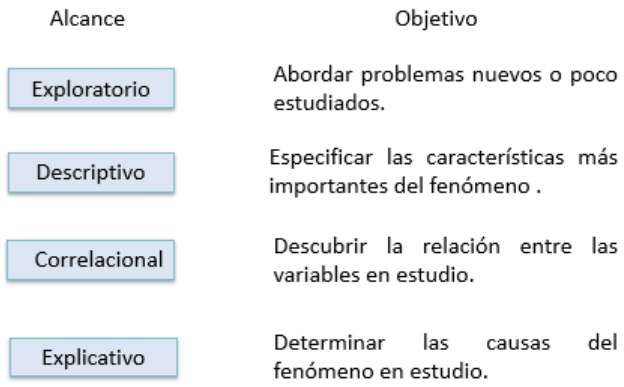


Figura 1.2 Alcances de la investigación según Hernández et al. (2010).

En la figura 1.2 se observa la clasificación de los alcances de una investigación. En este libro desarrollaremos ejemplos para casi todos los alcances mostrados.

También es importante recordar que, cuando los datos se recolectan en un determinado momento, la investigación se denomina transversal. Es decir, la muestra se analiza una sola vez. Asimismo, cuando la recolección de datos se realiza en diferentes momentos el tipo de investigación se denomina longitudinal. Es este último caso la muestra se puede analizar en diferentes etapas.

1.4. ¿En todas las investigaciones se deben formular hipótesis?

No. Existen investigaciones en las que no es necesario hacer uso de las hipótesis. Esto se debe a que dichas investigaciones se han planificado de esa forma, es decir, en ellas no se realiza alguna afirmación tentativa acerca de los resultados debido a que no se cuenta con mucha información al respecto o porque es un estudio con objetivos básicos.

En una investigación sencilla de alcance descriptivo se puede trabajar con una pregunta de investigación y su respectivo objetivo. Sin embargo, para una pregunta de investigación se pueden formular más de un objetivo de investigación.

- Pregunta de investigación.
- Objetivo de investigación.

Cuando contamos con información adecuada en nuestros antecedentes podemos establecer una hipótesis para nuestra investigación. En ese caso usamos los siguientes elementos:

- Pregunta de investigación.
- Objetivo de investigación.
- Hipótesis de investigación.

De manera general, en muchos protocolos de investigación elaborados por las universidades se pueden encontrar los siguientes elementos:

- Pregunta de investigación (general y específicos).
- Objetivo de investigación (general y específicos).
- Hipótesis de investigación (general y específicos).

1.5. ¿Las investigaciones descriptivas pueden tener hipótesis?

Las investigaciones descriptivas pueden contar con hipótesis, así como también se puede prescindir de ellas.

El uso de hipótesis en la investigación descriptiva se justifica cuando se quiere contrastar los resultados de nuestra investigación con los resultados de nuestros antecedentes. También utilizamos hipótesis cuando queremos comparar grupos.

En investigaciones cuyo objetivo consiste solo en describir las características de un hecho o fenómeno, se puede prescindir del uso de las hipótesis.

Sin embargo, no plantear hipótesis en una investigación no implica que ya no se pueda utilizar la estadística inferencial. Para complementar el análisis de nuestros resultados si una investigación carece de hipótesis, se pueden hacer estimaciones para la población a partir de los resultados muestrales. En estos casos se hace uso de los intervalos de confianza.

1.6. Muestreo probabilístico

Existen dos tipos de muestreo. El muestreo probabilístico y el no probabilístico. El muestreo probabilístico se fundamenta en el azar. Es decir que cada elemento de la población tiene la misma probabilidad de

ser seleccionado para formar parte de la muestra y por ese motivo se pueden generar inferencias para la población a partir de dicha muestra. El muestreo no probabilístico no se basa en el azar, la selección de la muestra se rige por criterios de conveniencia, la experiencia o por la sugerencia de los expertos, por ese motivo con los resultados del muestreo no probabilístico no se puede hacer inferencia hacia la población.

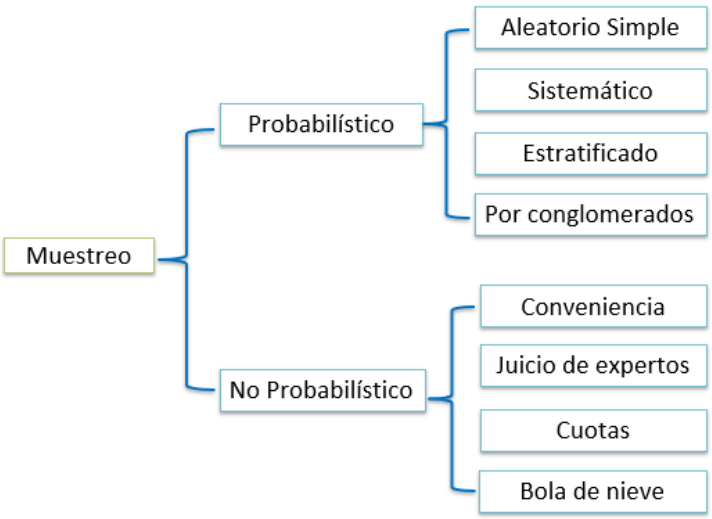


Figura 1.3 Tipos de muestreo.

Quintero y Duran (2004) resaltan que es muy común que los investigadores no elaboren un proceso adecuado de muestreo y utilicen la prueba estadística pensando que los resultados serán los mismos como cuando si se realiza un muestreo aleatorio.

Del párrafo anterior se desprende el criterio necesario para obtener buenos resultados cuando se utilizan los métodos estadísticos, se debe trabajar con muestras representativas obtenidas a través de un muestreo probabilístico. La omisión de un adecuado proceso de muestreo genera resultados poco favorables para la investigación. En muchos casos, las hipótesis, con adecuado fundamento teórico, no se comprueban estadísticamente y se generan resultados contradictorios.

Para un mejor entendimiento de los tipos de muestreo puede revisar el libro "*Orientaciones para la selección y el cálculo del tamaño de la muestra en investigación*".

Hopkins et al. (1997) ponen énfasis en que los libros de métodos estadísticos presuponen que cuando se habla de una muestra es porque se ha realizado un muestreo aleatorio.

Capítulo 2: Estimación estadística

En muchas actividades del quehacer humano es importante conocer ciertos valores o cantidades, sin embargo, no siempre es posible conocer de manera exacta estos valores. Entonces, surge la necesidad de estimar dichos valores y entendemos por estimación al proceso que permite aproximarnos al valor real que queremos conocer según una determinada variable.

La estimación estadística se compone de un conjunto de procesos que nos permiten aproximarnos a ciertos valores de una población a partir de muestras que se extraen de ella. Los valores a los que nos referimos pueden ser, por ejemplo, el promedio o la proporción. Así:

- Para un centro educativo de alto rendimiento, es necesario conocer la calificación promedio de sus estudiantes.

- Para una empresa pesquera es indispensable conocer la proporción de productos marinos que llegan en mal estado a sus instalaciones.

- En una fábrica es muy importante comparar la eficiencia de dos tipos de proceso.

Podemos establecer diferentes situaciones en las que se requiere conocer cierta información de una población. Pero, en muchos casos, por temas de factibilidad no se puede acceder a toda la población, sino que se recurre al uso de muestras representativas.

2.1. Estadígrafos y parámetros

Los parámetros se refieren a los valores o medidas que caracterizan a una población. Los parámetros se infieren a partir de los valores de la

muestra. Blair y Taylor (2008) definen los parámetros como valores que resumen las características de interés en una población.

El estadígrafo o estadístico es un valor que caracteriza a la muestra. Estos valores se obtienen directamente con los datos de la muestra. Lind et al. (2016) indican que cualquier medición que se basa en una muestra recibe el nombre de estadístico o estadígrafo

Tabla 2.1 Estadígrafos y parámetros.

Denominación	Estadígrafo	Parámetro
Media	$\bar{X}$	μ
Desviación estándar	S	σ
Varianza	S^2	σ^2
Proporción	p	π

En investigación se analizan las características de la población. Sin embargo, podemos aproximarnos, de forma muy cercana, a estas características a partir de una muestra.

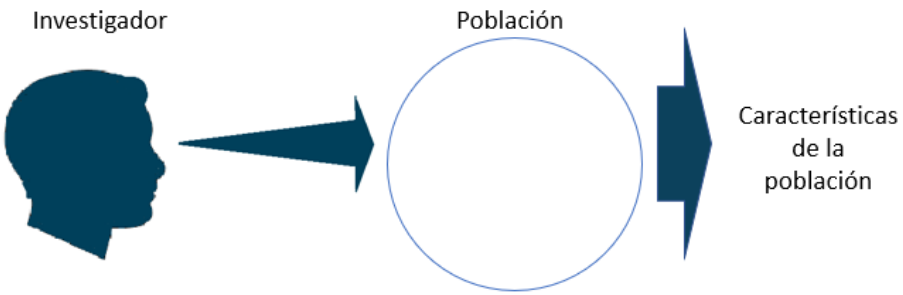


Figura 2.1 La investigación tiene como objeto de estudio a la población.

Las características de la muestra sirven de base para conocer las características de la población a través de un proceso llamado inferencia. Los resultados obtenidos luego de la inferencia son determinados con cierto grado de probabilidad. Es decir, no se llega a conocer con exactitud las características de la población, pero se pueden delimitar sus valores.

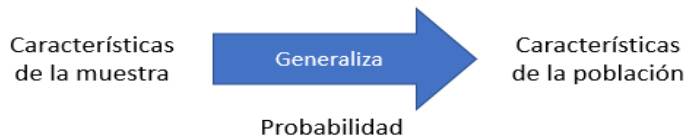


Figura 2.2 Con la muestra se estiman las características de la población.

2.2. Estimación puntual

La estimación puntual consiste en realizar la estimación de un parámetro a partir de un solo valor. Pliego y Maya (1999) señalan al estimador como una función de los valores muestrales, en ese sentido, al valor que asume el estimador para una determinada muestra se le define como estimación puntual.

Ejemplo:

Si calculamos una media muestral $\bar{X} = 50.2$ a partir de cierta muestra de una población.

La media $\bar{X} = 50.2$ es un estimador puntual del parámetro μ.

2.3. Estimadores y sesgo

Según Wackerly et al. (2010) un estimador nos indica como calcular el valor de una estimación con base en los datos de la muestra.

Si la media de la distribución muestral de un estadístico es igual al parámetro poblacional correspondiente se dice que el estadístico es un estimador insesgado del parámetro. De lo contrario, se dice que el estadístico es un estimador sesgado del parámetro.

Para tener una mejor comprensión de este tema podemos utilizar un ejemplo sencillo. Si tenemos una pequeña población P que se compone de los siguientes elementos:

$$P = \{2; 4; 5; 6\}$$

La media poblacional es:

$$\mu = \frac{2 + 4 + 5 + 6}{4} = 4.25$$

Si tomamos muestras diferentes de tamaño 2 obtenemos:

$$M_1 = \{2; 4\}$$
$$M_2 = \{2; 5\}$$
$$M_3 = \{2; 6\}$$

$$M_4 = \{4; 5\}$$
$$M_5 = \{4; 6\}$$
$$M_6 = \{5; 6\}$$

Las medias muestrales respectivas son:

$$\bar{X}_1 = \frac{2+4}{2} = 3$$
$$\bar{X}_2 = \frac{2+5}{2} = 3.5$$
$$\bar{X}_3 = \frac{2+6}{2} = 4$$
$$\bar{X}_4 = \frac{4+5}{2} = 4.5$$
$$\bar{X}_5 = \frac{4+6}{2} = 5$$
$$\bar{X}_6 = \frac{5+6}{2} = 5.5$$

Se puede apreciar que cada una de las medias muestrales son diferentes a la media poblacional. Esto quiere decir que, al tomar una sola muestra para estimar la media poblacional el resultado tiene un margen de error.

Podemos calcular la media de los 3 primeros promedios muestrales $\mu_{\bar{X}}$:

$$\frac{3 + 3.5 + 4}{3} = 3.5$$

Podemos calcular la media de los 4 primeros promedios muestrales $\mu_{\bar{X}}$:

$$\frac{3 + 3.5 + 4 + 4.5}{4} = 3.75$$

Podemos calcular la media de los 5 primeros promedios muestrales $\mu_{\bar{X}}$:

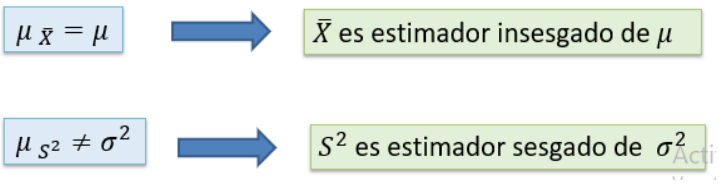
$$\frac{3 + 3.5 + 4 + 4.5 + 5}{5} = 4$$

Se puede observar que, a medida que se utilizan más muestras, la media de los promedios muestrales, el estimador, se acerca más al valor de la media poblacional. En los tres primeros casos el estimador es sesgado

porque difiere de valor de la media poblacional. Pero cuando utilizamos la media de todos los promedios muestrales obtenemos:

$$\mu_{\bar{X}} = \frac{3 + 3.5 + 4 + 4.5 + 5 + 5.5}{6} = 4.25$$

Entonces, podemos decir que $\mu_{\bar{X}}$ es un estimador insesgado de μ.



.

Figura 2.3 Sesgo en la estimación.

En el gráfico 2.3 se muestra que el estimador promedio de las varianzas muestrales μ_{S^2} es un estimador sesgado de la varianza poblacional σ^2.

2.4. Estimación por intervalo

La estimación por intervalo considera dos valores entre los cuales se debe encontrar el parámetro. La estimación por intervalo es la más utilizada porque nos señala el grado de precisión.

Ejemplo:

El parámetro μ se encuentra en el intervalo 50.2 ± 0.8.

Es decir:

$$50.2 - 0.8 \leq \mu \leq 50.2 + 0.8$$

$$49.4 \leq \mu \leq 51.0$$

Entonces:

$$\mu \in [49.4,\ 51.0]$$

Este intervalo se denomina intervalo de confianza y los extremos de dichos intervalos se obtienen a partir de un cálculo.

2.4.1. Nivel de confianza ($1 - \alpha$)

Para entender el nivel de confianza partimos imaginando una población de la cual se extraen muchas muestras, algunas muestras pueden generar intervalos que contienen al parámetro y el resto de muestras genera intervalos que no contienen al parámetro. El tanto por ciento que representa la cantidad de muestras, que generan intervalos que contienen al parámetro, con respecto al total de muestras se denomina nivel de confianza.

Por ejemplo, si se calcula un intervalo de confianza para la media poblacional con un nivel de confianza de 90% esto significa que, si de la población se extraen 100 muestras, 90 de estas muestras generan intervalos de confianza que contienen a la media poblacional.

Para el cálculo de los intervalos de confianza se utilizarán valores Z de la distribución normal. Estos valores Z están asociados al nivel de confianza. Los niveles de confianza más utilizados son el 90%, 95% y 99%. Para intervalos de confianza bilaterales tenemos los siguientes valores:

Tabla 2.2 Valor z según el nivel de confianza.

Nivel de confianza	Valor z asociado
90%	1.645
95%	1.96
99%	2.58

2.4.2. Distribución de probabilidad

En todos los procesos de estimación y las pruebas de hipótesis que se desarrollarán más adelante se hace uso de las distribuciones muestrales y las distribuciones de probabilidad.

Para entender el significado de una distribución de probabilidad podemos usar la información de la siguiente tabla:

Tabla 2.3 Tabla de distribución de la frecuencia absoluta.

Cantidad de hijos	Frecuencia
0	4
1	6
2	10
3	15
4	8
5	7
Total	50

Ahora elaboramos una tabla con las respectivas frecuencias relativas.

Tabla 2.4 Tabla de distribución de la frecuencia relativa.

Cantidad de hijos	F. Relativa
0	$\frac{4}{50} = 0.08$
1	$\frac{6}{50} = 0.12$
2	$\frac{10}{50} = 0.20$
3	$\frac{15}{50} = 0.30$
4	$\frac{8}{50} = 0.16$
5	$\frac{7}{50} = 0.14$
Total	1

Las frecuencias relativas de la tabla 2.4 se pueden reemplazar por probabilidades de modo que se obtiene el siguiente resultado:

Tabla 2.5 Tabla de distribución de probabilidades.

Cantidad de hijos	Probabilidad
0	$P[X = 0] = 0.08$
1	$P[X = 1] = 0.12$
2	$P[X = 2] = 0.20$
3	$P[X = 3] = 0.30$
4	$P[X = 4] = 0.16$
5	$P[X = 5] = 0.14$
Total	1

La tabla 2.5 representa una distribución de probabilidad para la variable aleatoria cantidad de hijos. Pero, como señalan Murray y Stephens (2009) la distribución de probabilidad hace referencia a la distribución de la población, en cambio las frecuencias relativas son distribuciones de las muestras. En la figura 2.4 se muestra la distribución de frecuencias relativas de forma gráfica.

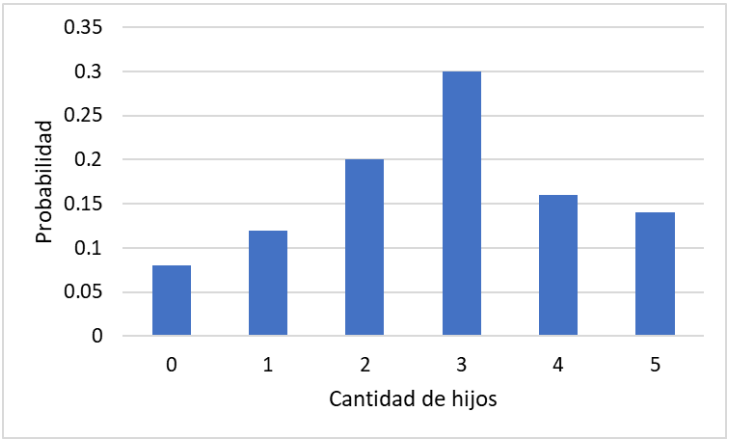


Figura 2.4

A continuación, se muestran las distribuciones que se utilizarán en este libro:

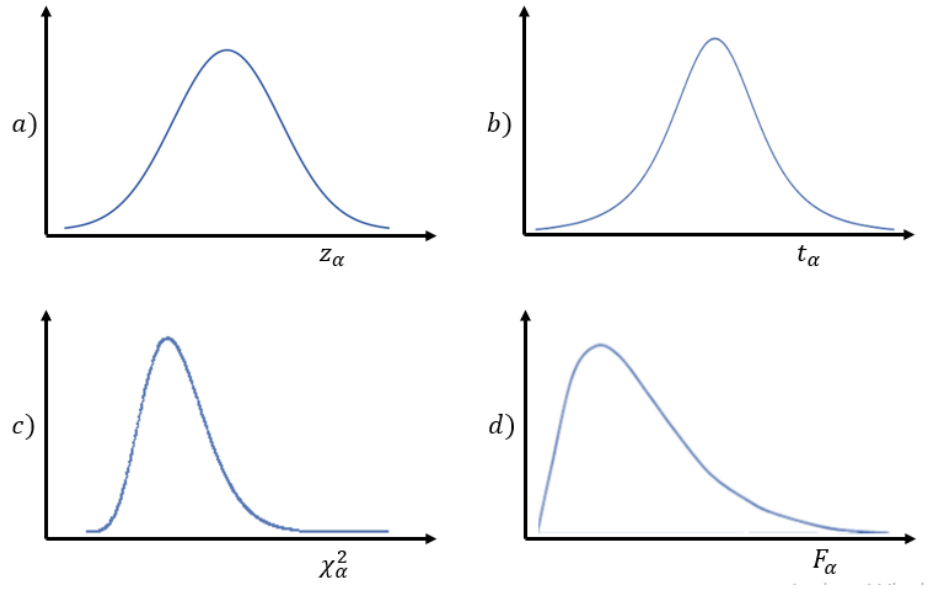


Figura 2.5 Algunas distribuciones de probabilidad.

En la figura 2.5 se observa la gráfica de la distribución normal (a), la distribución t de Student (b), la distribución Chi cuadrada (c) y la distribución F de Snedecor (d). Estas no son todas las distribuciones existentes. Hay otras como la distribución binomial y la distribución de Poisson. Además, las distribuciones de probabilidad pueden ser discretas o continuas. Las distribuciones que usaremos en el desarrollo de este libro son distribuciones continuas.

En la figura 2.5 también se observa los valores z, t, χ^2 y F que corresponden a las distribuciones normal, t, Chi cuadrado y F de Snedecor respectivamente. En las pruebas de hipótesis se compara los valores calculados con los valores teóricos, también llamados críticos, según el tipo de distribución.

Observe la notación z_α, en la figura 2.5 a), hace referencia a un valor z para determinado nivel de significancia. En otros textos se utilizan también $z_{\alpha/2}$, $z_{1-\alpha}$ o $z_{1-\alpha/2}$ dependiendo del tipo de prueba unilateral o bilateral. En este texto usaremos, de manera general, z con la finalidad de simplificar la notación.

Podemos calcular los valores críticos según la significancia utilizando el programa Excel.

En la figura 2.6 se muestra la fórmula para obtener el valor crítico z para una prueba unilateral con 95% de confianza.

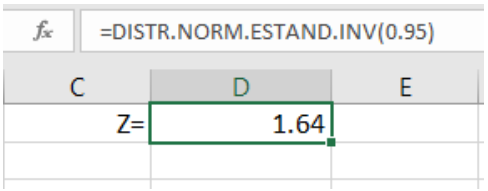


Figura 2.6 Cálculo del valor z unilateral.

Para una prueba bilateral con un 95% de confianza, el valor crítico z se obtiene con la siguiente fórmula:

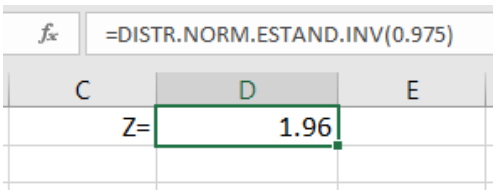


Figura 2.7 Cálculo del valor z bilateral.

Para una prueba bilateral con un 5% de significancia, el valor crítico t se obtiene con la siguiente fórmula:

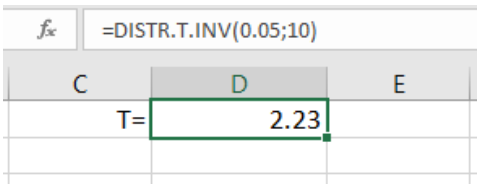


Figura 2.8 Cálculo del valor t bilateral.

2.4.3. Distribución muestral

De forma sencilla podemos definir a una distribución muestral como la distribución de las probabilidades de cierto estadístico que se puede obtener con las muestras de igual tamaño extraídas de una misma población. Por ello, se puede establecer una distribución muestral para la media, para la varianza, para la proporción, para la diferencia de medias, etc.

Para una mejor explicación consideremos el siguiente conjunto de datos que representa el gasto en transporte público por día:

$$P = \{2; 3; 4; 5; 6\}$$

Para elaborar la distribución muestral de las medias, podemos utilizar muestras con 3 elementos:

$$M_1 = \{2; 3; 4\}$$
$$M_2 = \{2; 3; 5\}$$
$$M_3 = \{2; 3; 6\}$$
$$M_4 = \{2; 4; 5\}$$
$$M_5 = \{2; 4; 6\}$$
$$M_6 = \{2; 5; 6\}$$
$$M_7 = \{3; 4;\ 5\}$$
$$M_8 = \{3; 4; 6\}$$
$$M_9 = \{3; 5; 6\}$$
$$M_{10} = \{4; 5; 6\}$$

Ahora calculamos las medias de cada muestra:

$$\bar{X}_1 = \frac{2+3+4}{3} = 3$$

$$\bar{X}_2 = \frac{2+3+5}{3} = 3.33$$

$$\bar{X}_3 = \frac{2+3+6}{3} = 3.67$$

$$\bar{X}_4 = \frac{2+4+5}{3} = 3.67$$

$$\bar{X}_5 = \frac{2+4+6}{3} = 4$$

$$\bar{X}_6 = \frac{2+5+6}{3} = 4.33$$

$$\bar{X}_7 = \frac{3+4+5}{3} = 4$$

$$\bar{X}_8 = \frac{3+4+6}{3} = 4.33$$

$$\bar{X}_9 = \frac{3+5+6}{3} = 4.67$$

$$\bar{X}_{10} = \frac{4+5+6}{3} = 5$$

Luego construimos una tabla de distribución de frecuencias para las medias muestrales:

Tabla 2.6 Tabla de distribución de probabilidades.

Media	Frecuencia	Probabilidad
3	1	0.1
3.33	1	0.1
3.67	2	0.2
4	2	0.2
4.33	2	0.2
4.67	1	0.1
5	1	0.1
Total	10	1

Luego realizamos la respectiva gráfica:

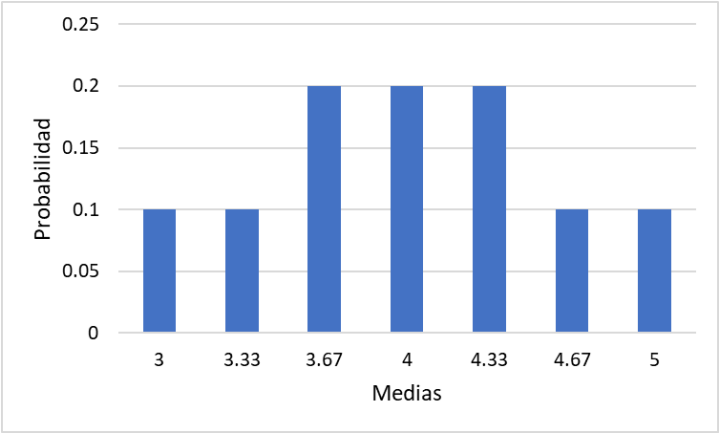


Figura 2.9 Distribución de probabilidad.

A medida que aumenta el tamaño de la población y el tamaño de la muestra, la distribución adopta la forma de una curva especial.

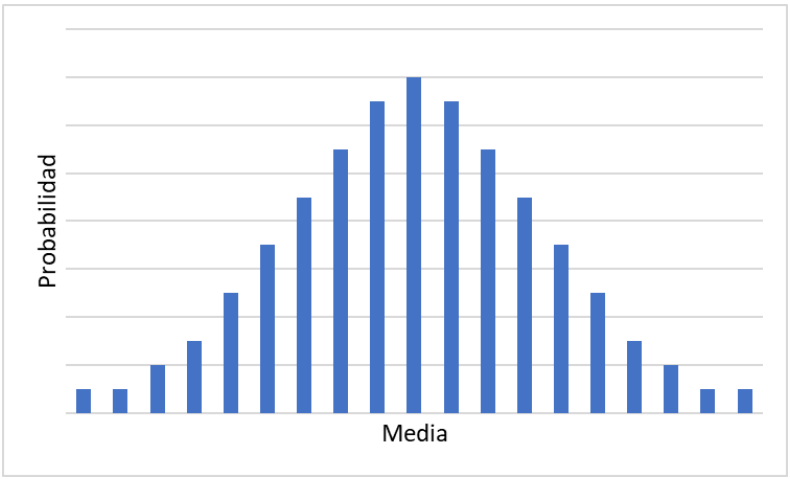


Figura 2.10 Aproximación a la curva de la distribución normal.

En la figura 2.10 se verifica el teorema del límite central. Este teorema nos indica que cuando el tamaño de la muestra se hace más grande, la distribución muestral de las medias se aproxima a la distribución normal (Hopkins et al.,1997).

2.4.4. Estandarización de valores

En estadística inferencial se dice que una variable aleatoria X se distribuye normalmente con media μ y desviación estándar σ. Simbólicamente:

$$X \rightarrow N(\mu, \sigma)$$

La distribución normal estándar tiene la forma:

$$X \rightarrow N(0, 1)$$

Esto quiere decir que la distribución normal se define por su media y su desviación estándar. Moya y Saravia (2004) afirman que las variables aleatorias con distribución normal se pueden transformar en variables aleatorias estandarizadas usando la expresión:

$$z = \frac{x - \mu}{\sigma}$$

Entonces, cuando tenemos variables numéricas como la masa, el costo o el volumen, cuyas unidades son el kilogramo, soles o litros; todas ellas se pueden expresar en unidades estándar. Note que el valor z no tiene una unidad especial, pero hace referencia a la cantidad de desviaciones estándar que contiene la diferencia $x - \mu$.

Ejemplo: si tenemos una población de personas cuya masa corporal promedio es de 68.4 kg con una desviación estándar de 4.3 kg ¿Cuál será el valor estandarizado para x=70 kg?

Solución: extraemos los datos de manera simplificada.

$\mu = 68.4\ kg$ $\sigma = 4.3\ kg$ $x = 70\ kg$ $z =?$

Reemplazamos en la expresión:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{70\ kg - 68.4\ kg}{4.3\ kg}$$

$$z = \frac{1.6\ kg}{4.3\ kg} = \frac{1.6}{4.3}$$

$$z = 0.37$$

Observe que el valor z no tiene unidades. Este resultado quiere decir que los 70 kg superan a la media 68.4 kg en 0.37 desviaciones estándar. La finalidad del proceso de estandarización es el acceso a las probabilidades de la distribución normal.

Los promedios de las muestras de una población grande también se distribuyen normalmente. Sabemos que para calcular un promedio se divide la suma de n elementos con la cantidad n. Por ello, si la varianza de la distribución de una variable es σ^2, entonces la varianza de una distribución de medias será σ^2/n. Extrayendo la raíz cuadrada tenemos la desviación estándar $\sigma/\sqrt{n}$. Entonces, la distribución de las medias muestrales se puede representar como:

$$X \rightarrow N(\mu, {}^{\sigma}/_{\sqrt{n}})$$

Esto implica que para estandarizar los valores de la media usamos la expresión:

$$z = \frac{\bar{X} - \mu}{{}^{\sigma}/_{\sqrt{n}}}$$

Observe que en el numerador usamos el promedio $\bar{X}$ en lugar de un valor determinado de la variable X. Pero el desarrollo de este caso se verá en las siguientes secciones.

Los valores z están asociados a una determinada probabilidad. Sin embargo, esta probabilidad no necesita ser calculada manualmente porque ya existen tablas y programas con las probabilidades para cada valor z.

2.5. Intervalo de confianza bilateral para la media μ

El intervalo de confianza para la media poblacional μ, para cierto nivel de confianza $1 - \alpha$, es $[a; b]$.

Esto quiere decir que la probabilidad de que la media poblacional pertenezca al intervalo $[a; b]$ es igual al nivel de confianza $1 - \alpha$ (Moya y Saravia, 2004).

$$P[a \leq \mu \leq b] = 1 - \alpha$$

α es la significancia y su explicación se abordará durante el desarrollo del tema de la prueba de hipótesis.

En investigación, para $1 - \alpha$ se suele utilizar 90%, 95% o 99%, en estos casos, la significancia será 10%, 5% o 1% respectivamente.

Por ejemplo, cuando tenemos que $\mu \epsilon [a; b]$ con un nivel de confianza del 95%, esto significa que de las muestras que se extraen de la población el 95% de ellas producirán intervalos $[a; b]$ que contienen a μ.

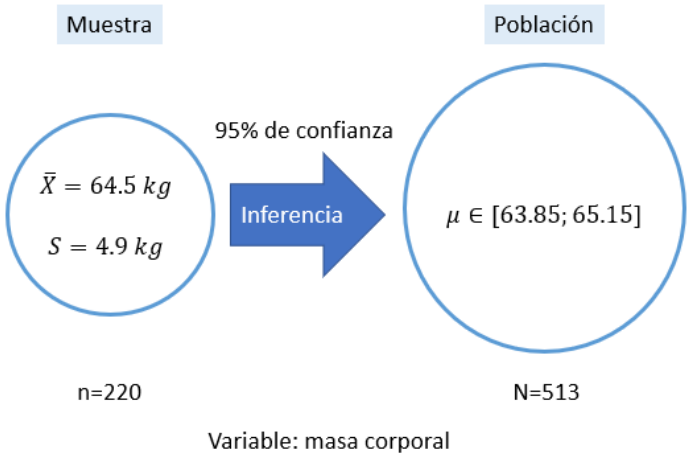


Figura 2.11 Estimación de la media poblacional para una muestra grande.

En la figura 2.11 se ilustra que la media poblacional se puede estimar a partir de la media muestral. Es decir, se conoce que el promedio de las masas corporales de los 220 individuos es 64.5 kg, pero no tenemos la certeza de que este promedio sea el mismo para la población compuesta de 513 individuos. Lo que se puede afirmar, con un nivel de confianza de 95% y luego de realizar los respectivos cálculos, es que el promedio de las masas corporales de todos los individuos de la población se encuentra en el intervalo de 63.85 kg a 65.15 kg.

2.5.1. Intervalo de confianza para la media poblacional con σ conocida

Recordemos el intervalo de confianza definido en la sección anterior:

$$P[a \leq \mu \leq b] = 1 - \alpha$$

Si usamos valores estandarizados, basándonos en la expresión mostrada por Moya y Saravia (2004), tenemos:

$$P[-z \leq Z_c \leq z] = 1 - \alpha$$

Pero el Z_c representa la expresión estandarizada de la distribución muestral. Es decir:

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

Reemplazando en la expresión anterior:

$$P\left[-z \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z\right] = 1-\alpha$$

$$P\left[-z \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq z \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$P\left[-\bar{X}-z \cdot \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+z \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

$$P\left[\bar{X}-z \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+z \cdot \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

Finalmente, el intervalo de confianza para la media de una población que se distribuye normalmente queda determinado por:

$$\mu \in [\bar{X}-z\frac{\sigma}{\sqrt{n}}; \bar{X}+z\frac{\sigma}{\sqrt{n}}]$$

z: valor asociado al nivel de confianza.
$\bar{X}$: media muestral.
σ: desviación estándar de la población (se puede obtener de los antecedentes de la investigación).
n: tamaño de la muestra.

Ejemplo: en una investigación realizada en diferentes hospitales de Ciudad de México en el año 2022 se trató de conocer el desarrollo antropométrico de los recién nacidos prematuros sanos. Una de las medidas antropométricas de interés en este estudio fue el perímetro cefálico. Por ello, luego de analizar una muestra representativa de 155 recién nacidos prematuros se obtuvo un promedio de crecimiento del perímetro cefálico de 0.82 cm por semana. En un antecedente encontrado, Cárdenas et al. (2005) indican que los recién nacidos prematuros sanos del Instituto Nacional Perinatal de México tienen aumentos promedio de 0.86 ± 0.39 cm a la semana en su perímetro cefálico.

Título: desarrollo antropométrico en recién nacidos prematuros sanos en Ciudad de México.

Pregunta: ¿Cuál es crecimiento del perímetro cefálico en recién nacidos prematuros sanos en Ciudad de México?

Objetivo: conocer el crecimiento del perímetro cefálico en recién nacidos prematuros sanos en Ciudad de México.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: perímetro cefálico.

Si se utiliza un nivel de confianza de 95% ¿Cuál será el intervalo de confianza para la media de la población de recién nacidos prematuros?

Solución: tenemos que resumir la información.

$\bar{X} = 0.82 \qquad \sigma = 0.39 \qquad n = 155 \qquad z = 1.96$

Reemplazamos los valores en la expresión adecuada:

$$\mu \in [\bar{X} - z\frac{\sigma}{\sqrt{n}}; \bar{X} + z\frac{\sigma}{\sqrt{n}}]$$

$$\mu \in [0.82 - 1.96 \times \frac{0.39}{\sqrt{155}}; 0.82 + 1.96 \times \frac{0.39}{\sqrt{155}}]$$

$$\mu \in [0.82 - 0.06; 0.82 + 0.06]$$

$$\mu \in [0.76; 0.88]$$

Se puede afirmar, con un nivel de confianza de 95%, que la media poblacional pertenece al intervalo de 0.76 cm a 0.88 cm de crecimiento del perímetro cefálico por semana.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas – Z de una muestra

Tabla 2.7 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
155	0.82	0.0313	[0.7586; 0.8814]

Como se puede apreciar en la tabla 2.7, el resultado obtenido por medio del procedimiento de desarrollo completo es igual al obtenido con Minitab.

2.5.2. Intervalo de confianza para la media poblacional con σ desconocida

Cuando el tamaño de muestra es grande $n \geq 30$ (Spiegel y Stephens, 2009) la variable aleatoria tiene una distribución aproximadamente normal. Por lo tanto, se hará uso de los niveles de confianza que se han estudiado anteriormente (90%, 95%, 99%).

Casi siempre se desconoce σ, por ello se utiliza la desviación estándar de la muestra S.

Muestras grandes

Para muestras grandes se seguirá utilizando la distribución normal.

$$\mu \in [\bar{X} - z\frac{S}{\sqrt{n}}; \bar{X} + z\frac{S}{\sqrt{n}}]$$

Ejemplo 1: Un estudiante necesita conocer, para su investigación, el promedio de la cantidad de horas de uso del celular por día. Por ello, toma una muestra de 120 estudiantes y obtiene una media de 6.5 horas con una desviación estándar de 1.4 horas.

Título: uso del celular en estudiantes del colegio "X".

Pregunta: ¿Cuál es la cantidad de horas de uso del celular en los estudiantes del colegio "X"?

Objetivo: conocer la cantidad de horas de uso del celular en los estudiantes del colegio "X".

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: cantidad de horas de uso del celular.

Si se utiliza un nivel de confianza de 99% ¿Cuál será el intervalo de confianza para la media de la población de estudiantes?

Solución: tenemos la siguiente información.

$\bar{X} = 6.5 \qquad S = 1.4 \qquad n = 120 \qquad z = 2.58$

Reemplazamos los valores en la expresión adecuada:

$$\mu \in [6.5 - 2.58 \times \frac{1.4}{\sqrt{120}}; 6.5 + 2.58 \times \frac{1.4}{\sqrt{120}}]$$

$$\mu \in [6.5 - 0.33; 6.5 + 0.33]$$

$$\mu \in [6.17; 6.83]$$

Se puede afirmar, con un nivel de confianza de 99%, que la media poblacional de horas de uso del celular pertenece al intervalo de 6.17 horas a 6.83 horas.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas – Estadísticas básicas – Z de una muestra

Tabla 2.8 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
120	6.5	0.125	[6.171; 6.829]

En la tabla 2.8 se verifica que el resultado obtenido por medio del programa Minitab es igual al resultado obtenido por medio del desarrollo completo.

Ejemplo 2: un estudiante de enfermería realiza un estudio para conocer el ritmo cardiaco de los adultos mayores del hospital X luego de ser sometidos al tratamiento de salud M. Dichas personas presentan características similares y han sido sometidas al mismo tratamiento de salud. Para conocer el ritmo cardiaco promedio, el estudiante, toma una muestra de 72 adultos mayores y realiza la medición del ritmo cardiaco bajo condiciones similares. Los resultados indican que el ritmo cardiaco

promedio, luego del tratamiento, es de 75 latidos por minuto con una desviación estándar de 11 latidos por minuto.

Título: Tratamiento M y ritmo cardiaco en adultos mayores del Hospital "X".

Pregunta: ¿Cuál es ritmo cardiaco en adultos mayores del Hospital "X" luego de someterse al tratamiento M?

Objetivo: conocer ritmo cardiaco en adultos mayores del Hospital "X" luego de someterse al tratamiento M.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: ritmo cardiaco luego del tratamiento M.

Si se utiliza un nivel de confianza de 95% ¿Cuál será el intervalo de confianza para la media de la población de adultos mayores?

Solución: tenemos la siguiente información.

$\bar{X} = 75 \qquad S = 11 \qquad n = 72 \qquad z = 1.96$

Reemplazamos los valores en la expresión adecuada:

$$\mu \in [75 - 1.96 \times \frac{11}{\sqrt{72}}; 75 + 1.96 \times \frac{11}{\sqrt{72}}]$$

$$\mu \in [75 - 2.54; 75 + 2.54]$$

$$\mu \in [72.46; 77.54]$$

Se puede afirmar, con un nivel de confianza de 95%, que la media poblacional pertenece al intervalo de 72.46 a 77.54 latidos por minuto.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas – Estadísticas básicas – Z de una muestra

Tabla 2.9 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
72	75	1.30	[72.46; 77.54]

En la tabla 2.9, se verifica el resultado obtenido por medio del procedimiento de desarrollo completo.

Muestras pequeñas

Para muestras pequeñas ($n < 30$) se hará uso de la distribución t. Wackerly et al. (2010) justifican este procedimiento señalando que "la distribución t tiene una función de densidad muy semejante a la función de densidad de la distribución normal estándar".

$$\mu \in [\bar{X} - t\frac{S}{\sqrt{n}}; \bar{X} + t\frac{S}{\sqrt{n}}]$$

S: desviación estándar de la muestra.

Para utilizar la distribución t, se introduce el término grados de libertad cuyo valor se obtiene usando $n - 1$. Es decir, si tenemos una muestra de tamaño 15, entonces en dicho análisis se usará $15 - 1 = 14$ grados de libertad.

Ejemplo 1: en un estudio, realizado en una empresa panificadora, se debe determinar el peso promedio de los queques de tamaño mediano que son parte de la producción. Por ello, se toma una muestra aleatoria de 10 queques y se mide los pesos, en gramos, con una balanza electrónica. Los resultados se muestran a continuación:

Pesos:	750	748	746	754	758	750	748	744	750	748

Título: características de los queques medianos elaborados en la panificadora "X".

Pregunta: ¿Cuál es el peso de los queques medianos elaborados en la panificadora "X"?

Objetivo: conocer el peso de los queques medianos elaborados en la panificadora "X".

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: peso de los queques medianos.

Calcula e interpreta el intervalo de confianza para el peso promedio poblacional de los queques utilizando un nivel de confianza de 95%.

Solución: primero calculamos la media ($\bar{X}$) y la desviación estándar de la muestra (S).

Usando Excel y las fórmulas predeterminadas =DESVEST.M y =PROMEDIO, obtenemos los valores requeridos.

$$\bar{X} = 749.60$$

$$S = 3.98$$

Además, el tamaño de la muestra es 10. Es decir, $n = 10$

El valor de t crítico para una prueba bilateral con 5%=0.05 de significancia y $n - 1 = 9$ grados de libertad es 2.26 (de la tabla del apéndice).

$$t = 2.26$$

Reemplazando los valores en el intervalo de confianza para la media poblacional para muestras pequeñas:

$$\mu \in [\bar{X} - t\frac{S}{\sqrt{n}}; \bar{X} + t\frac{S}{\sqrt{n}}]$$

$$\mu \in [749.60 - 2.26 \times \frac{3.98}{\sqrt{10}}; 749.60 + 2.26 \times \frac{3.98}{\sqrt{10}}]$$

$$\mu \in [749.60 - 2.84; 749.60 + 2.84]$$

$$\mu \in [746.76; 752.44]$$

Se puede afirmar, con un nivel de confianza de 95%, que la media poblacional pertenece al intervalo de 746.76 g a 752.44 g.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas – t de una muestra

Tabla 2.10 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
10	749.6	1.26	[746.75; 752.45]

En la tabla 2.10 se verifica el resultado obtenido por medio del procedimiento anterior. Los límites del intervalo de confianza en ambos procedimientos no son totalmente iguales, esto se debe a que el programa Minitab y otros programas, realizan los cálculos usando todos los decimales. En nuestro primer procedimiento realizamos una aproximación a dos decimales.

Ejemplo 2: en una investigación, realizada en el laboratorio, se quiere conocer el peso que ganan los conejos cuando reciben una estricta dieta "X". Por ello, se toma una muestra aleatoria de 11 conejos, se les aplica la dieta durante tres semanas y se mide los pesos ganados, en gramos, con una balanza electrónica. Los resultados se muestran a continuación:

78 86 76 84 80 70 82 90 88 84 84

Título: influencia de la dieta "X" en el peso de los conejos.

Pregunta: ¿Cuál es el peso ganado por los conejos a los que se les aplicó la dieta "X"?

Objetivo: conocer el peso ganado por los conejos a los que se les aplicó la dieta "X".

Tipo de investigación: enfoque cuantitativo, cuasi experimental, alcance explicativo y de corte longitudinal.

Variable x: dieta.

Variable y: peso ganado por conejo.

Calcula e interpreta el intervalo de confianza para el peso promedio poblacional de los conejos utilizando un nivel de confianza de 95%.

Solución: primero calculamos la media ($\bar{X}$) y la desviación estándar de la muestra (S).

Usando Excel y las fórmulas predeterminadas =DESVEST.M y =PROMEDIO, obtenemos los valores requeridos.

$$\bar{X} = 82$$
$$S = 5.73$$

Además, el tamaño de la muestra es 11. Es decir, $n = 11$

El valor de t teórico para una prueba bilateral con 5%=0.05 de significancia y $n - 1 = 10$ grados de libertad es 2.23 (de la tabla del apéndice).

Reemplazando los valores en el intervalo de confianza para la media poblacional para muestras pequeñas:

$$\mu \in [\bar{X} - t\frac{S}{\sqrt{n}}; \bar{X} + t\frac{S}{\sqrt{n}}]$$

$$\mu \in [82 - 2.23 \times \frac{5.73}{\sqrt{11}}; 82 + 2.23 \times \frac{5.73}{\sqrt{11}}]$$

$$\mu \in [82 - 3.85; 82 + 3.85]$$

$$\mu \in [78.15; 85.85]$$

Se puede afirmar, con un nivel de confianza de 95%, que la media poblacional pertenece al intervalo de 78.15 g a 85.85 g.

Usando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas – t de una muestra

Obtenemos:

Tabla 2.11 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
11	82	1.73	[78.15; 85.85]

En la tabla 2.11 se verifica el resultado obtenido por medio del procedimiento de desarrollo completo.

2.5.3. Intervalo de confianza para la proporción (π)

El intervalo de confianza para la proporción poblacional π es $[a; b]$. Es decir:

$$P[a \leq \pi \leq b] = 1 - \alpha$$

Para muestras grandes ($n \geq 30$) se considera que la distribución es aproximadamente normal. Por lo tanto, se hará uso de los niveles de confianza que se han utilizado anteriormente (90%, 95%, 99%).

$$\pi \in [p - z\sqrt{\frac{p(1-p)}{n}}; p + z\sqrt{\frac{p(1-p)}{n}}]$$

Dado que, p es la proporción de éxitos y la distribución es binomial, entonces $q = 1 - p$ es la proporción de fracasos.

Para entender mejor el significado de la proporción p vamos a considerar a la cantidad de éxitos como x y n como el total de casos o ensayos. Si queremos conocer la proporción de empleados que tienen auto propio en una empresa, encuestamos a una muestra de 300 empleados y resulta que 180 de ellos tienen auto propio, la proporción de esta muestra se calcula de la siguiente manera:

$$p = \frac{x}{n}$$

Tenemos que:

x: cantidad de personas que tienen auto propio (éxito).
n: tamaño de la muestra.

Reemplazando los datos podemos calcular la proporción muestral:

$$p = \frac{180}{300} = 0.6 = 60\%$$

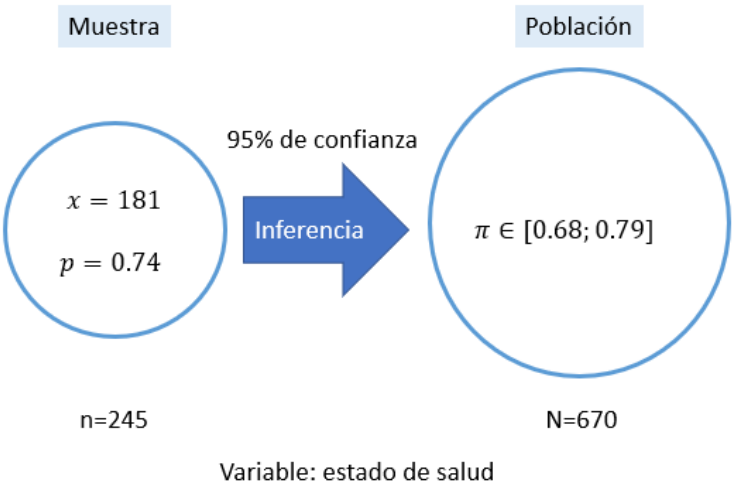


Figura 2.12 Estimación de la proporción poblacional.

En la figura 2.12 se muestra que la proporción poblacional se puede estimar a partir de la proporción muestral. Es decir, se conoce que la proporción de individuos con buen estado de salud representa el 74% de los 245 individuos que componen la muestra, pero no tenemos la certeza de que esta proporción sea la misma para la población compuesta de 670 individuos. Lo que se puede afirmar, con un nivel de confianza de 95% y luego de realizar los respectivos cálculos, es que la proporción poblacional de individuos con buena salud se encuentra en el intervalo de 68% a 79%.

Ejemplo 1: mediante un estudio, en una empresa avícola, se quiere conocer la proporción poblacional de clientes que paga en efectivo. Por ello, se toma una muestra de 120 clientes y observa que 75 de ellos pagan en efectivo.

Título: características de consumo en la empresa avícola "X".

Pregunta: ¿Cuál es la proporción de clientes que paga en efectivo en la empresa avícola "X"?

Objetivo: conocer la proporción de clientes que paga en efectivo en la empresa avícola "X".

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: tipo de pago.

Calcula el intervalo de confianza para la proporción poblacional de clientes que paga en efectivo. Utilice un nivel de confianza de 95%.

Solución: Primero calculamos la proporción p, es decir, la proporción muestral de clientes que pagan en efectivo.

$$p = \frac{75}{120} = 0.625$$

Además, el tamaño de la muestra es 120. Es decir, $n = 120$

El valor de z para una prueba bilateral con 95% de confianza es 1.96.

Reemplazando los valores en el intervalo de confianza para la proporción poblacional para muestras grandes:

$$\pi \in [p - z\sqrt{\frac{p(1-p)}{n}}; p + z\sqrt{\frac{p(1-p)}{n}}]$$

$$\pi \in [0.625 - 1.96 \times \sqrt{\frac{0.625 \times 0.375}{120}}; 0.625 + 1.96 \times \sqrt{\frac{0.62 \times 0.375}{120}}]$$

$$\pi \in [0.625 - 0.087; 0.625 + 0.087]$$

$$\pi \in [0.538; 0.712]$$

Se puede afirmar, con un nivel de confianza de 95%, que la proporción poblacional pertenece al intervalo de 53.8% a 71.2%.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas – 1 proporción

Obtenemos el resultado:

Tabla 2.12 Estadísticas descriptivas

N	Evento	Muestra p	IC de 99% para p
120	75	0.625	[0.538381; 0.711619]

En la tabla 2.12 se verifica el resultado obtenido por medio del procedimiento de desarrollo completo.

Ejemplo 2: Un investigador requiere conocer la proporción de estudiantes que usan tutoriales para mejorar su rendimiento académico en el colegio "X". Por ello, toma una muestra de 420 estudiantes y observa que 300 de ellos utiliza tutoriales para mejorar su rendimiento académico.

Título: uso de tutoriales académicos en estudiantes del colegio "X".

Pregunta: ¿Cuál es la proporción de estudiantes que usan tutoriales para mejorar su rendimiento académico en el colegio "X"?

Objetivo: conocer la proporción de estudiantes que usan tutoriales para mejorar su rendimiento académico en el colegio "X".

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: uso de tutoriales para mejorar el rendimiento académico.

Calcula el intervalo de confianza para la proporción poblacional de estudiantes que usa tutoriales para mejorar su rendimiento académico. Utilice un nivel de confianza de 99%.

Solución: Primero calculamos la proporción p, es decir, la proporción muestral de estudiantes que usan tutoriales para mejorar su rendimiento académico.

$$p = \frac{300}{420} = 0.71$$

Además, el tamaño de la muestra es 420. Es decir, $n = 420$

El valor de z para una prueba bilateral con 99% de confianza es 2.58.

Reemplazando los valores en el intervalo de confianza para la proporción poblacional para muestras grandes:

$$\pi \in [p - z\sqrt{\frac{p(1-p)}{n}}; p + z\sqrt{\frac{p(1-p)}{n}}]$$

$$\pi \in [0.71 - 2.58 \times \sqrt{\frac{0.71 \times 0.29}{420}}; 0.71 + 2.58 \times \sqrt{\frac{0.71 \times 0.29}{420}}]$$

$$\pi \in [0.71 - 0.06; 0.71 + 0.06]$$

$$\pi \in [0.65; 0.77]$$

Se puede afirmar, con un nivel de confianza de 99%, que la proporción poblacional de estudiantes que usan tutoriales para mejorar su rendimiento académico pertenece al intervalo de 65% a 77%.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas – 1 proporción

Tabla 2.13 Estadísticas descriptivas

N	Evento	Muestra p	IC de 99% para p
420	300	0.714286	[0.657506; 0.771066]

En la tabla 2.13 se verifica el resultado obtenido por medio del procedimiento de desarrollo completo. Sin embargo, los resultados no son idénticos. Esto se debe a que en el primer procedimiento se han realizado aproximaciones para trabajar con dos decimales.

Capítulo 3: Prueba de hipótesis con respecto a la media

3.1. Hipótesis estadística

Es una afirmación que se hace acerca de la distribución de una o más variables aleatorias (Moya y Saravia, 2004). También se puede decir que son afirmaciones con respecto a alguna característica de la población. Además, se puede decir que la hipótesis estadística es una conjetura acerca de algún valor determinado que un parámetro puede tomar (Pliego y maya, 1999).

3.2. Hipótesis nula (H_0)

Es la hipótesis que se pone a prueba estadísticamente, supone la ausencia de efectos o diferencias en las variables que se investiga.

A continuación, se muestran ejemplos de hipótesis nulas.

H_0: las medias del grupo 1 y el grupo 2 son iguales.

H_0: la proporción poblacional de individuos con la característica X es igual en el grupo 1 y el grupo 2.

H_0: las medias de todos los grupos son iguales.

H_0: no existe asociación entre la variable 1 y la variable 2.

H_0: no existe correlación significativa entre la variable 1 y la variable 2.

H_0: el factor 1 no tiene efecto significativo sobre la variable respuesta.

H_0: el promedio de las edades de los empleados es igual a 41 años.

H_0: la proporción de estudiantes aprobados es igual a 89%.

En investigación siempre se trata de rechazar esta hipótesis para poder generalizar nuestros supuestos. Por ejemplo, consideremos dos grupos de estudiantes, a un grupo se le somete a un reforzamiento académico y al otro grupo no, se realiza una prueba única a ambos grupos y se observan los resultados. La hipótesis nula es la siguiente.

H_0: Los puntajes promedio del grupo 1 y del grupo 2 son iguales.

Esa afirmación, la hipótesis nula, no tiene mucho sentido, dado que uno de los grupos debe tener ventaja por el reforzamiento académico que se le brindó. Por ello, lo que se quiere es rechazar la hipótesis nula, así llegaremos a la conclusión de que el puntaje promedio no es igual en ambos grupos y lo más lógico sería verificar que el puntaje promedio del grupo que recibió el reforzamiento académico es mayor.

El razonamiento anterior es importante al momento de plantear las hipótesis. Sin embargo, esto último se desarrollará luego con mayor detalle siguiendo una serie de pasos predeterminados.

La hipótesis nula se rechaza cuando se encuentra suficiente evidencia estadística de que dicha afirmación no es cierta. Es decir, cuando se verifica numéricamente que existe una probabilidad muy pequeña de que la afirmación contenida en H_0 sea cierta.

3.3. Hipótesis alterna (H_1)

Es la conclusión a la que se desea llegar. Se expresa en forma contraria a la hipótesis nula, pero de forma menos precisa.

La hipótesis alterna es la hipótesis que se toma como conclusión cuando se rechaza la hipótesis nula. Anderson et al. (2012) consideran que la hipótesis alternativa es la hipótesis de investigación.

A continuación, se muestran algunos ejemplos de hipótesis alternativas:

H_1**:** las medias del grupo 1 y el grupo 2 no son iguales.

H_1: la proporción poblacional de individuos con la característica X no es igual en el grupo 1 y el grupo 2.

H_1: las medias de todos los grupos no son iguales.

H_1: existe asociación entre la variable 1 y la variable 2.

H_1: existe correlación significativa entre la variable 1 y la variable 2.

H_1: el factor 1 tiene efecto significativo sobre la variable respuesta.

H_1: el promedio de las edades de los empleados es menor a 41 años.

H_1: la proporción de estudiantes aprobados es mayor a 89%.

3.4. Prueba de hipótesis

Es un proceso que permite tomar la decisión de rechazar o no rechazar la hipótesis nula del estudio con base en los datos obtenidos a través de la observación o experimentación.

Se debe resaltar que la hipótesis que se debe rechazar o no rechazar es la hipótesis nula no la hipótesis alterna.

Si se rechaza la hipótesis nula, entonces queda como opción aceptar la hipótesis alterna.

Tabla 3.1 Tipos de error.

	H_0 es verdadera	H_1 es verdadera
Rechazar H_0	**Error de tipo I**	Correcto
Rechazar H_1	Correcto	**Error de tipo II**

Error de tipo I

Consiste en rechazar H_0 cuando H_0 es verdadera.

Error de tipo II

Consiste en aceptar H_0 cuando H_0 es falsa.

Como señalan Moya y Saravia (2004) no es fácil reducir la probabilidad de cometer estos errores, sin embargo, se puede lograr aumentando el tamaño de la muestra.

3.5. Nivel de significancia (α)

Se define como la probabilidad de cometer el error de tipo I.

$$\alpha = P[rechazar\ H_0\ /\ H_0\ es\ verdadera]$$

$$\alpha = P[aceptar\ H_1\ /\ H_1\ es\ falsa]$$

En el capítulo anterior se estableció el nivel de confianza como $1 - \alpha$. Por lo tanto, se entiende que el nivel de confianza y la significancia se complementan para sumar 1 o 100%.

Es decir, si tenemos un nivel de confianza de 95%=0.95. Entonces α se calcula de la siguiente manera:

$$0.95 = 1 - \alpha$$

$$\alpha = 1 - 0.95$$

$$\alpha = 0.05$$

Si tenemos un nivel de confianza de 99%=0.99. Entonces α se calcula de la siguiente manera:

$$0.99 = 1 - \alpha$$

$$\alpha = 1 - 0.99$$

$$\alpha = 0.01$$

Hernández et al. (2010) nos explican de forma sencilla que el nivel de significancia es la probabilidad de equivocarse con respecto a una afirmación. Además, esta significancia se debe fijar o elegir desde un inicio para poder realizar la estimación o la prueba de hipótesis. El nivel de significancia es elegido por el investigador.

También se puede entender que el propósito de fijar un nivel de significancia es definir un límite para que H_0 sea considerado como un evento raro (Siegel y Castellan,1995). Así, cuando se realizan los cálculos y la probabilidad de que se rechace H_0. cuando esta es verdadera, es menor que la significancia fijada, se toma la decisión de rechazar H_0. Es decir, al ser H_0 un evento poco probable, también será poco probable que rechacemos H_0 cuando esta sea verdadera.

De todo lo anterior se concluye que la significancia debe ser un valor pequeño, dado que es la probabilidad de cometer error. Generalmente se utiliza el 5% de significancia en diferentes ramas del conocimiento. Pero algunas investigaciones requieren una significancia mucho menor debido a la importancia que constituye la toma de una decisión adecuada. En investigaciones en el campo de las ciencias de la salud, por ejemplo, se debe estar seguro de que el error es tan pequeño que permita rechazar la hipótesis nula. Por ello, en esos casos, se suele utilizar una significancia de 0.01.

3.6. Región de rechazo

Es la región de una distribución muestral que representa la probabilidad de que ocurra H_0. También se le denomina región crítica.

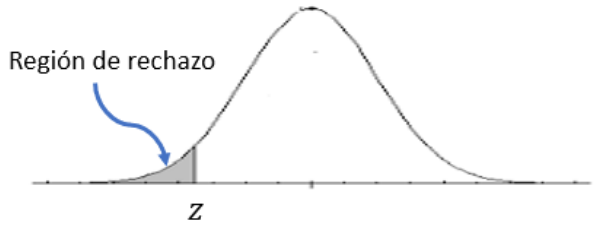


Figura 3.1 Región de rechazo en la distribución normal.

Por ejemplo, en el gráfico 3.2 se tiene una región de rechazo que tiene como límite inferior $z = 2.9$. Si calculamos un valor estadístico de prueba z_c=3.08. Este valor está dentro de la región de rechazo. Por lo tanto, la decisión a tomar es rechazar H_0.

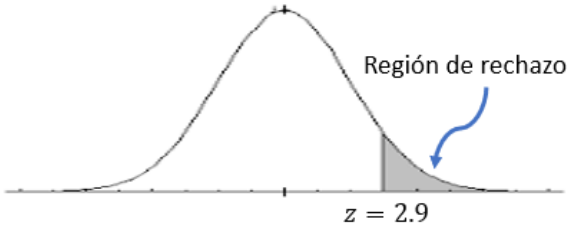


Figura 3.2 Región de rechazo en la distribución normal.

En el mismo gráfico 3.2, si calculamos un valor estadístico de prueba z_c=1.38. Este valor no está dentro de la región de rechazo. Por lo tanto, la decisión a tomar es no rechazar H_0.

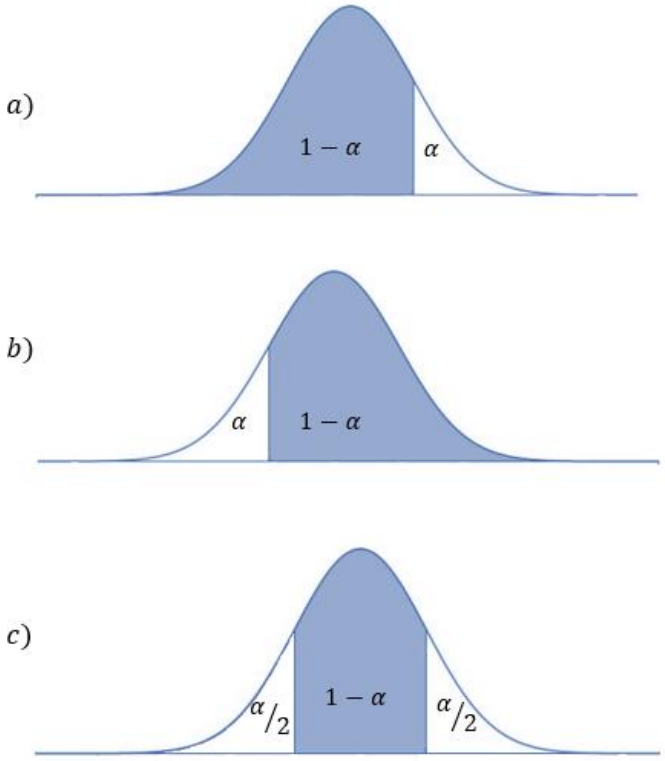


Figura 3.3 Región de rechazo según el tipo de prueba.

En la figura 3.3 se muestra la región de rechazo, de color blanco representada por la significancia α para las pruebas unilaterales a y b. En el caso c, también tenemos la significancia α, pero como la prueba es bilateral este valor se divide en dos partes iguales para cada una de las colas de modo que tenemos $\alpha/2$ y $\alpha/2$.

3.7. El p valor

En el apartado anterior se muestra el proceso para decidir si se rechaza o no se rechaza H_0. Para ello se utiliza un estadístico de prueba y se compara con un valor crítico que generalmente se obtiene a partir de una tabla de valores como z, t o F. Pero este no es el único método para tomar una decisión estadística.

Desde hace ya bastante tiempo, las pruebas de hipótesis se realizan usando programas estadísticos. Estos programas nos brindan mucha información y entre toda esa información se encuentra el p valor. Este p valor hace referencia a lo que denominaremos significancia calculada. Generalmente consideramos un nivel de confianza de 95%, esto implica una significancia de $\alpha = 0.05$ para nuestros resultados. La regla para tomar la decisión, de rechazar o no rechazar H_0, consiste en comparar la significancia calculada p con la significancia asumida α:

Si $p < \alpha$, se rechaza H_0.
Si $p \geq \alpha$, no se rechaza H_0.

Molina (2017) explica que el valor p es la probabilidad de que las diferencias encontradas, entre los parámetros a comparar, sean producto del azar. Por ejemplo, si analizamos el IMC en una muestra y planteamos la siguiente hipótesis nula:

H_0: el promedio de los IMC es igual a 23.5.

Obtener un valor $p = 0.002$ significa que hay una probabilidad de 0.002 para la afirmación de que la media 23.5, para la población de donde se extrajo la muestra, se haya obtenido por azar. Entonces esa probabilidad pequeña menor que la significancia $\alpha = 0.05$, que se ha considerado para la investigación, nos permite tomar la decisión de rechazar H_0. En otras palabras, es poco probable que la afirmación H_0 sea válida para la población.

Si el p valor hubiese sido mayor, $p = 0.09$ por ejemplo, la lógica nos dirige a pensar que esa probabilidad es muy grande para no tenerla en consideración. El $p = 0.09$ es mayor que la significancia $\alpha = 0.05$ definida para la investigación. Entonces, no se rechaza H_0.

3.8. Prueba de hipótesis para la media

En esta prueba de hipótesis se compara una media poblacional con la media poblacional estimada en base a la media muestral.

En este tipo de prueba se hace uso de una media referencial para establecer la comparación con la media que se quiere conocer. Esta media referencial se obtiene a partir de información recolectada en los antecedentes y que el investigador considera relevante porque le permite hacer una comparación con sus resultados obtenidos.

A continuación, se muestran algunos ejemplos de hipótesis para la media:

H_1: el consumo promedio de agua en los habitantes de la ciudad X excede los 18 m^3 por mes.

H_1: la edad promedio de los docentes de la universidad X es menor a 45 años.

H_1: el gasto mensual promedio por la compra de diarios en los padres de familia del colegio X es diferente a S/ 45.

También es importante señalar que el investigador decide si debe incluir o no incluir hipótesis en su investigación descriptiva. En este texto mostramos ejemplos de investigaciones descriptivas en los cuales se incluye hipótesis, pero no es una regla estricta el hacer uso de hipótesis en estudios de alcance descriptivo.

Las pruebas de hipótesis pueden ser unilaterales o bilaterales, también se les conoce como prueba de cola superior, prueba de cola inferior y prueba de dos colas.

Pruebas unilaterales

$H_0: \mu = \mu_0.$
$H_1: \mu < \mu_0.$

$H_0: \mu = \mu_0.$
$H_1: \mu > \mu_0$

Prueba bilateral

$H_0: \mu = \mu_0.$
$H_1: \mu \neq \mu_0$

3.8.1. Prueba unilateral de cola inferior

Se utiliza cuando queremos verificar si nuestra media poblacional μ es menor que una media referencial μ_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \mu = \mu_0.$
$H_1: \mu < \mu_0.$

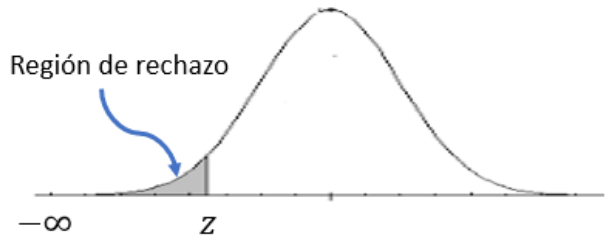


Figura 3.4 Región de rechazo para una cola inferior.

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle -\infty; z \rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Cuando tenemos muestras grandes ($n \geq 30$) y desconocemos σ, podemos utilizar S.

$$Z_c = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

4) Si $Z_c$pertenece a la R.R. se rechaza H_0, en caso contrario no se rechaza.

Usaremos los siguientes valores z:

Tabla 3.2 Valor z según el nivel de confianza

Nivel de confianza	Valor z asociado
90%	-1.28
95%	-1.64
99%	-2.33

Ejemplo 1: un investigador del sector salud**,** mediante un estudio, quiere verificar si el consumo anual de granos andinos en la ciudad de Arequipa en el año 2021 es menor con respecto al consumo de granos andinos a nivel nacional en el año 2019. El Ministerio de agricultura y Riego del Perú (2019) informa que el consumo anual de granos andinos (quinua, kiwicha, cañihua, y tarwi) es de 2.3 kilogramos por persona. Se toma una muestra de 330 personas adultas y se obtiene un promedio anual de consumo de 2.28 kg por persona con una desviación estándar de 0.28 kg.

Título: consumo de cereales andinos en habitantes de la ciudad de Arequipa.

Pregunta: ¿Cuál es el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa?

Objetivo: conocer el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa.

Hipótesis: el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa es menor a 2.3 kg por año.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo anual de cereales andinos por persona.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola inferior.

H_0: el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa no es menor a 2.3 kg por año.

H_1: el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa es menor a 2.3 kg por año.

Extraemos los datos necesarios de forma resumida.

$\bar{X} = 2.28 \qquad \mu_0 = 2.3 \qquad S = 0.28 \qquad n = 330$

Para una prueba unilateral con cola inferior y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle -\infty;\ -1.64 \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{2.28 - 2.3}{0.28/\sqrt{330}} = -1.30$$

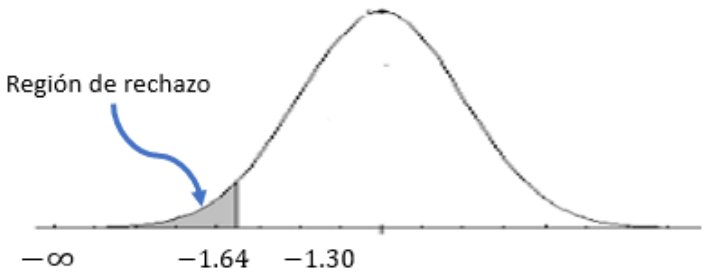


Figura 3.5 Región de rechazo para una cola inferior.

-1.30 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que el consumo promedio de cereales andinos en los habitantes de la ciudad de Arequipa sea menor a 2.3 kg por año.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Los resultados se muestran a continuación:

Tabla 3.3 Estadísticas descriptivas

N	Media	Error estándar de la media	Límite superior de 95% para μ
330	2.28	0.0154	2.3054

En la tabla 3.3 se muestran los elementos básicos como el tamaño y la media muestral además del límite superior del intervalo de confianza para la media.

Tabla 3.4 Prueba de hipótesis en Minitab.

Hipótesis nula	H_0: μ= 2.3
Hipótesis alterna	H_1: μ < 2.3
Valor z	-1.30
Valor p	0.097

En la tabla 3.4 se muestra la prueba de hipótesis completa. Para tomar la decisión de rechazar o no la hipótesis nula se utiliza el valor p. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.097 no es menor que 0.05, no se rechaza la hipótesis nula.

Ejemplo 2: En una investigación se desea conocer las características del consumo de carne vacuna en los habitantes de la ciudad de Lima. Según el Ministerio de Agricultura y Riego del Perú (2019) los limeños consumen en promedio 8.31 kg de carne vacuna al año. Otros antecedentes indican que debido al problema de la pandemia por COVID 19 el promedio de consumo de carne vacuna debe haber disminuido. Por ello, se toma una muestra de 385 habitantes adultos de la ciudad de lima y se obtiene una media de consumo de 7.92 kg de carne vacuna por año con una desviación estándar de 0.82 kg.

Título: consumo de carne vacuna en habitantes adultos de la ciudad de Lima.

Pregunta: ¿Cuántos kilogramos de carne vacuna consumen los habitantes adultos de la ciudad de Lima?

Objetivo: conocer la cantidad de kilogramos de carne vacuna que consumen los habitantes adultos de la ciudad de Lima.

Hipótesis: el consumo promedio de carne vacuna en los habitantes de la ciudad de Lima es menor a 8.31 kg por año.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo anual de carne vacuna.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para la prueba unilateral de cola inferior.

H_0: el consumo promedio de carne vacuna en los habitantes de la ciudad de Lima no es menor a 8.31 kg por año.

H_1: el consumo promedio de carne vacuna en los habitantes de la ciudad de Lima es menor a 8.31 kg por año.

Extraemos los datos necesarios de forma resumida.

$\bar{X} = 7.92 \qquad \mu_0 = 8.31 \qquad S = 0.82 \qquad n = 385$

Para una prueba unilateral con cola inferior y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle -\infty;\ -1.64 \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{7.92 - 8.31}{0.82/\sqrt{385}} = -9.33$$

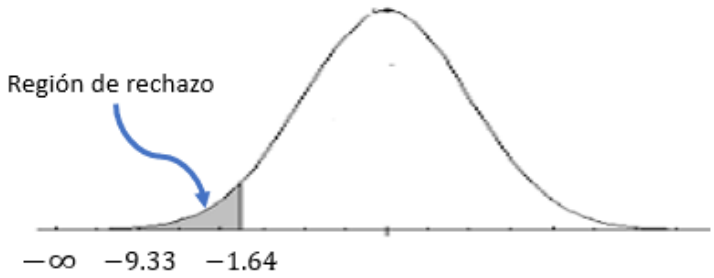


Figura 3.6 Región de rechazo para una cola inferior.

-9.33 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el consumo promedio de carne vacuna en los habitantes de la ciudad de Lima es menor a 8.31 kg por año.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Los resultados se muestran a continuación:

Tabla 3.5 Estadísticas descriptivas

N	Media	Error estándar de la media	Límite superior de 95% para μ
385	7.92	0.0418	7.9887

En la tabla 3.5 se muestra los elementos básicos como el tamaño y la media muestral además del límite superior del intervalo de confianza para la media.

Tabla 3.6 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 8.31
Hipótesis alterna	H_1: μ < 8.31
Valor z	-9.33
Valor p	0.000

En la tabla 3.6 se muestra la prueba de hipótesis completa. Para tomar la decisión de rechazar o no la hipótesis nula se utiliza el valor p. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.000 es menor que 0.05, se rechaza la hipótesis nula.

3.8.2. Prueba unilateral de cola superior

Se utiliza cuando queremos verificar si nuestra media poblacional μ es mayor que una media referencial μ_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \mu = \mu_0.$
$H_1: \mu > \mu_0.$

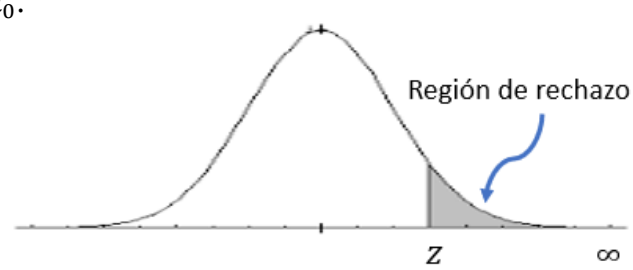


Figura 3.7 Región de rechazo para una cola superior.

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle z;\ \infty \rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Cuando tenemos muestras grandes ($n \geq 30$) y desconocemos σ, podemos utilizar S.

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

4) Si Z_c pertenece a la R.R. se rechaza H_0, en caso contrario no se rechaza.

Usaremos los siguientes valores z:

Tabla 3.7 Valor z según el nivel de confianza.

Nivel de confianza	Valor z asociado
90%	1.28
95%	1.64
99%	2.33

Ejemplo 1: en un artículo del año 2021, el diario La República informa que un reporte hecho por la empresa App Annie, indica que el promedio global de tiempo de uso apps es de 4,2 horas al día. Un investigador ha encontrado antecedentes que le indican que este promedio es mayor en algunas regiones del mundo sobre todo cuando los usuarios son jóvenes. Por ello, toma una muestra representativa de 375 usuarios jóvenes de la ciudad de Lima y realiza una encuesta. Los resultados señalan que la media de uso de las apps es de 5.4 horas al día con una desviación estándar de 1.5 horas al día.

Título: tiempo de uso de las apps en jóvenes de la ciudad de Lima.

Pregunta: ¿Cuántas horas al día usan las apps los jóvenes de la ciudad de Lima?

Objetivo: conocer la cantidad de horas al día que usan las apps los jóvenes de la ciudad de Lima.

Hipótesis: los jóvenes de la ciudad de Lima usan las apps por más de 4.2 horas al día.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: horas de uso de las apps.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola superior.

H_0: el promedio de horas de uso de las apps es igual a 4.2 horas por día en los jóvenes de la ciudad de Lima.

H_1: el promedio de horas de uso de las apps es mayor a 4.2 horas por día en los jóvenes de la ciudad de Lima.

Extraemos los datos necesarios de forma resumida.

$\bar{X} = 5.4 \qquad \mu_0 = 4.2 \qquad S = 1.5 \qquad n = 375$

Para una prueba unilateral con cola superior y con 95% de confianza, la región de rechazo será:

$$R.R. = \langle 1.64; \infty \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{5.4 - 4.2}{1.5/\sqrt{375}} = 15.49$$

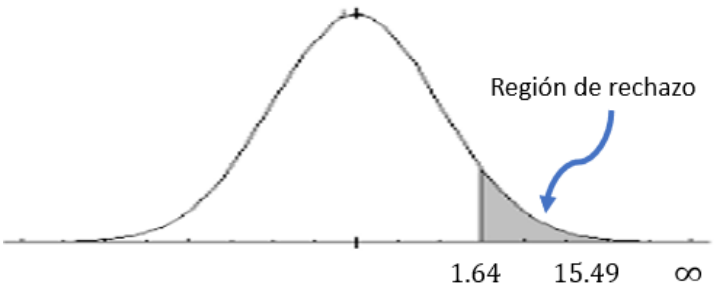


Figura 3.8 Región de rechazo para una cola superior.

15.49 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el promedio de horas de uso de las apps es mayor a 4.2 horas por día en los jóvenes de la ciudad de Lima.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Obtenemos:

Tabla 3.8 Estadísticas descriptivas

N	Media	Error estándar de la media	Límite inferior de 95% para μ
375	5.4	0.0775	5.2726

En la tabla 3.8 se muestra los elementos básicos como el tamaño y la media muestral además del límite inferior del intervalo de confianza para la media.

Tabla 3.9 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 4.2
Hipótesis alterna	H_1: μ > 4.2
Valor z	15.49
Valor p	0.000

En la tabla 3.9 se muestra la prueba de hipótesis completa. Observamos el valor p=0.000 y lo comparamos con la significancia 0.05. Dado que el valor p=0.000 es menor que 0.05, se rechaza la hipótesis nula.

Ejemplo 2: según el portal Infobae (2022), en promedio, los peruanos consumen 46 litros de cerveza al año. Un investigador quiere realizar una comparación de este promedio de consumo con el promedio de consumo de los habitantes de la ciudad de Trujillo. Observa los antecedentes y considera que el consumo promedio debe ser mayor en la ciudad de Trujillo. Por ello, toma una muestra representativa de 248 adultos de la ciudad de Trujillo y obtiene una media de consumo de cerveza de 47.2 litros por año con una desviación estándar de 12 litros.

Título: consumo de cerveza en adultos de la ciudad de Trujillo.

Pregunta: ¿Cuántos litros de cerveza por año consume cada habitante adulto de la ciudad de Trujillo?

Objetivo: conocer la cantidad de litros de cerveza por año que consume cada habitante adulto de la ciudad de Trujillo.

Hipótesis: cada habitante adulto de la ciudad de Trujillo consume más de 46 litros de cerveza por año.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo de cerveza.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola superior.

H_0: el consumo promedio de cerveza en los habitantes adultos de la ciudad de Trujillo no es mayor a 46 litros por año.

H_1: el consumo promedio de cerveza en los habitantes adultos de la ciudad de Trujillo es mayor a 46 litros por año.

Extraemos los datos necesarios de forma resumida.

$$\bar{X} = 47.2 \qquad \mu_0 = 46 \qquad S = 12 \qquad n = 248$$

Para una prueba unilateral con cola superior y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle 1.64; \infty \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{47.2 - 46}{12/\sqrt{248}} = 1.57$$

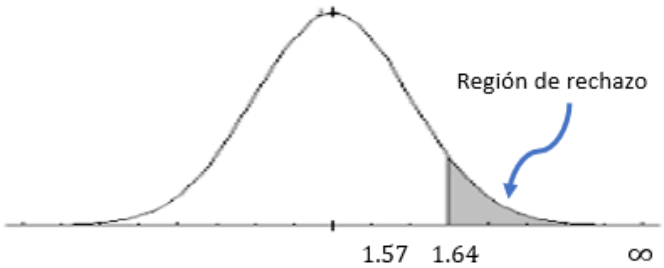


Figura 3.9 Región de rechazo para una cola superior.

1.57 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que el consumo promedio de cerveza en los habitantes adultos de la ciudad de Trujillo sea mayor a 46 litros por año.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Obtenemos:

Tabla 3.10 Estadísticas descriptivas

N	Media	Error estándar de la media	Límite inferior de 95% para μ
248	47.2	0.762	45.947

En la tabla 3.10 se muestra los elementos básicos como el tamaño y la media muestral además del límite inferior del intervalo de confianza para la media.

Tabla 3.11 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 46
Hipótesis alterna	H_1: μ > 46
Valor z	1.57
Valor p	0.058

En la tabla 3.11 se muestra la prueba de hipótesis completa y se comprueba el resultado anterior. Dado que el valor p=0.058 no es menor que 0.05, no se rechaza la hipótesis nula.

3.8.3. Prueba bilateral o de dos colas

Se utiliza cuando queremos verificar si nuestra media poblacional μ es diferente a una media referencial μ_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \mu = \mu_0.$
$H_1: \mu \neq \mu_0.$

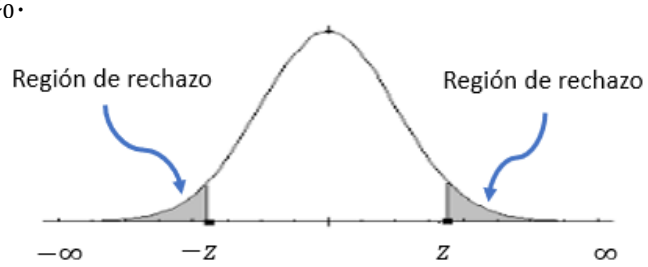


Figura 3.10 Región de rechazo para dos colas.

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle -\infty; z\rangle \cup \langle z, \infty\rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Cuando tenemos muestras grandes ($n \geq 30$) y desconocemos σ, podemos utilizar S.

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

4) Si Z_c pertenece a la R.R. se rechaza H_0, en caso contrario no la rechazamos.

Usaremos los siguientes valores:

Tabla 3.12 Valor z según el nivel de confianza

Nivel de confianza	Valor z asociado
90%	1.645
95%	1.96
99%	2.58

Ejemplo 1: la Universidad Católica de Santa María (2020) señala que en el Perú el consumo de papa en el 2020 alcanzó los 90 kg por persona. Un investigador quiere verificar si este promedio todavía se mantiene en el 2021 en la ciudad de Lima. Por ello, realiza una encuesta a un grupo de 380 amas de casa y obtiene un consumo promedio de 88.6 kg por persona con una desviación estándar de 11.9 kg. Además, se debe precisar que no se cuenta con información que determine si el consumo de papa ha disminuido o aumentado luego del 2020.

Título: consumo anual de papa en habitantes de la ciudad de Lima.

Pregunta: ¿Cuál es el consumo anual de papa en los habitantes de la ciudad de Lima?

Objetivo: conocer el consumo anual de papa en los habitantes de la ciudad de Lima.

Hipótesis: el consumo promedio anual de papa en los habitantes de la ciudad de Lima es diferente a 90 kg por persona.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo anual de papa por persona.

Usando un nivel de confianza de 99% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: el consumo promedio anual de papa en los habitantes de la ciudad de Lima es igual a 90 kg por persona.

H_1: el consumo promedio anual de papa en los habitantes de la ciudad de Lima es diferente a 90 kg por persona.

Extraemos los datos necesarios de forma resumida.

$\bar{X} = 88.6$ $\mu_0 = 90$ $S = 11.9$ $n = 380$

Para una prueba bilateral y con 99% de confianza, la región de rechazo es:

$$R.R. = \langle -\infty; -2.58\rangle \cup \langle 2.58;\ \infty\rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{87.8 - 90}{7.9/\sqrt{380}} = -2.29$$

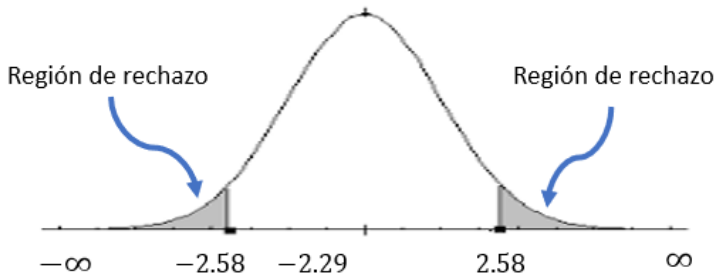


Figura 3.11 Región de rechazo para dos colas.

-2.29 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que el consumo promedio anual de papa en los habitantes de la ciudad de Lima sea diferente a 90 kg por persona en el año 2021.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Obtenemos:

Tabla 3.13 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
380	88.6	0.610	[87.028; 90.172]

En la tabla 3.13 se muestra los elementos básicos como el tamaño y la media muestral, además del intervalo de confianza para la media.

Tabla 3.14 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 90
Hipótesis alterna	H_1: $\mu \neq 90$
Valor z	-2.29
Valor p	0.022

En la tabla 3.14 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 99%, entonces la significancia es de 0.01. Dado que el valor p=0.022 no es menor que 0.01, no se rechaza la hipótesis nula.

Ejemplo 2: el Ministerio de Salud del Perú (2019) informó que muchos peruanos no cumplen con los 150 minutos de actividad física semanal recomendados para una vida saludable. Un investigador desea verificar si el tiempo dedicado a la actividad física semanal en el año 2022 es similar a 150 minutos señalados por el MINSA. Para este fin, toma una muestra representativa de 661 habitantes de la ciudad de Arequipa y obtiene un promedio de 118 minutos de actividad física por semana con una deviación estándar de 32 minutos.

Título: Actividad física en los habitantes de la ciudad de Arequipa.

Pregunta: ¿Cuánto tiempo de actividad física semanal acumulan los habitantes de la ciudad de Arequipa?

Objetivo: conocer la cantidad de tiempo de actividad física semanal que acumulan los habitantes de la ciudad de Arequipa.

Hipótesis: la cantidad de tiempo de actividad física semanal que acumulan los habitantes de la ciudad de Arequipa es diferente a 150 minutos.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: minutos de actividad física semanal.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: el tiempo promedio de actividad física semanal que acumulan los habitantes de la ciudad de Arequipa es igual a 150 minutos.

H_1: el tiempo promedio de actividad física semanal que acumulan los habitantes de la ciudad de Arequipa es diferente a 150 minutos.

Extraemos los datos necesarios de forma resumida.

$$\bar{X} = 118 \qquad \mu_0 = 150 \qquad S = 32 \qquad n = 661$$

Para una prueba bilateral y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle -\infty; -1.96\rangle \cup \langle 1.96;\ \infty\rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$Z_c = \frac{118 - 150}{32/\sqrt{661}} = -25.71$$

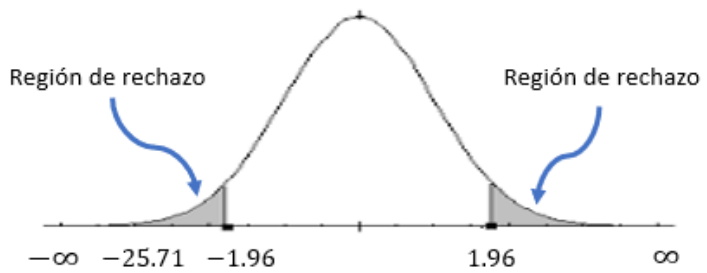


Figura 3.12 Región de rechazo para dos colas.

-25.71 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el tiempo promedio de actividad física semanal que acumulan los habitantes de la ciudad de Arequipa es diferente a 150 minutos.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - Z de una muestra

Tabla 3.15 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 99% para μ
661	118	1.24	[115.56; 120.44]

En la tabla 3.15 se muestra los elementos básicos como el tamaño y la media muestral además del intervalo de confianza para la media.

Tabla 3.16 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 150
Hipótesis alterna	H_1: $\mu \neq 150$
Valor z	-25.71
Valor p	0.000

En la tabla 3.16 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.000 es menor que 0.05, se rechaza la hipótesis nula.

3.8.4. Prueba bilateral para muestras pequeñas

Cuando la muestra es pequeña ($n < 30$) se utiliza la distribución t. Aunque, la distribución t también se puede utilizar para analizar muestras grandes y los resultados serán similares. Spiegel y Stephens (2009) señalan que la distribución t es parte de la teoría del muestreo exacto y sus resultados son válidos tanto para muestras pequeñas como para muestras grandes.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \mu = \mu_0.$
$H_1: \mu \neq \mu_0$

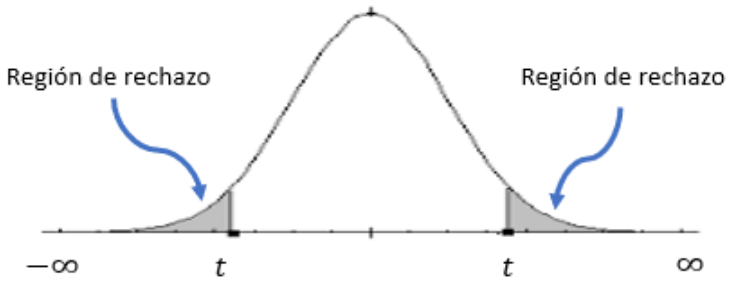


Figura 3.13 Región de rechazo para dos colas.

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor t con $n-1$ grados de libertad.
2) Elabora la Región de rechazo: $R.R. = \langle -\infty; -t \rangle \cup \langle t; \infty \rangle$
3) Determina el valor de t_c :

$$t_c = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

4) Si t_c pertenece a la R.R. se rechaza H_0, en caso contrario no se rechaza.

Ejemplo 1: Un docente quiere verificar si las calificaciones de los estudiantes del aula X han variado significativamente en el segundo bimestre con respecto al primer bimestre. Por ello toma una muestra de 25 estudiantes y obtiene una media de 17.1 puntos con una desviación estándar de 2.1 puntos. El primer bimestre la calificación promedio fue de 18 puntos.

Título: calificaciones de los estudiantes del aula X en el segundo bimestre.

Pregunta: ¿Cuál es el promedio de calificaciones en los estudiantes del aula X en el segundo bimestre?

Objetivo: conocer la calificación promedio de los estudiantes del aula X en el segundo bimestre.

Hipótesis: la calificación promedio de los estudiantes del aula X en el segundo bimestre es diferente a 18 puntos.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: calificación de los estudiantes.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: el promedio de calificaciones en los estudiantes del aula X en el segundo bimestre es igual a 18 puntos.

H_1: el promedio de calificaciones en los estudiantes del aula X en el segundo bimestre es diferente a 18 puntos.

Extraemos los datos necesarios de forma resumida.

$\bar{X} = 17.1 \qquad \mu_0 = 18 \qquad S = 2.1 \qquad n = 25$

Para una prueba bilateral, con 0.05 de significancia y 25-1=24 grados de libertad, la región de rechazo es:

$$R.R. = \langle -\infty; -2.06\rangle \cup \langle 2.06; \infty\rangle$$

El valor 2.06 se obtiene a partir de la tabla de la distribución t de student (buscar en las tablas del apéndice). También se puede usar Excel.

Determinamos el valor de t_c usando la expresión:

$$t_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$t_c = \frac{17.1 - 18}{2.1/\sqrt{25}} = -2.14$$

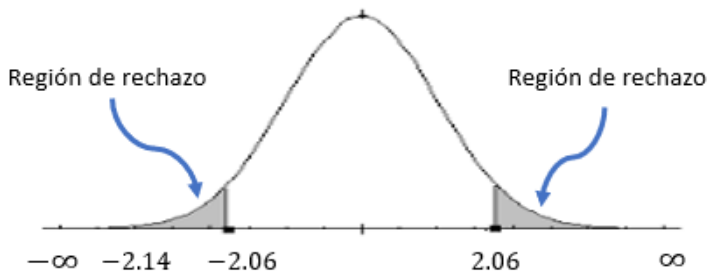


Figura 3.14 Región de rechazo para dos colas.

-2.14 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el promedio de calificaciones en los estudiantes del aula X en el segundo bimestre es diferente a 18 puntos.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - t de una muestra

Obtenemos:

Tabla 3.17 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 95% para μ
25	17.1	0.420	[16.233; 17.967]

En la tabla 3.17 se muestra los elementos básicos como el tamaño y la media muestral además del intervalo de confianza para la media.

Tabla 3.18 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 18
Hipótesis alterna	H_1: $\mu \neq 18$
Valor z	-2.14
Valor p	0.042

En la tabla 3.18 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.042 es menor que 0.05, se rechaza la hipótesis nula.

Ejemplo 2: En el año 2021 el promedio de consumo de galletas en el colegio X fue de 65 galletas por estudiante. Para conocer si ese promedio de consumo de galletas se mantiene en el año 2022 se tomó una muestra de 20 estudiantes y se analizó el consumo anual aproximado de galletas a través de una encuesta. Los resultados son los siguientes:

Tabla 3.19 Matriz de datos.

Estudiante	Cantidad de galletas
1	60
2	66
3	65
4	62
5	64
6	66
7	65
8	65
9	64
10	67
11	62
12	65
13	64
14	65
15	66
16	62
17	65
18	68
19	61
20	70

Título: Consumo de galletas en los estudiantes del colegio X en el año 2022.

Pregunta: ¿Cuántas galletas consumieron los estudiantes del colegio X en el año 2022?

Objetivo: conocer la cantidad de galletas que consumieron los estudiantes del colegio X en el año 2022.

Hipótesis: la cantidad de galletas que consumieron los estudiantes del colegio X en el año 2022 es diferente a 65 unidades.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo de galletas por año.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: la cantidad promedio de galletas que consumieron los estudiantes del colegio X en el año 2022 es igual a 65 unidades.

H_1: la cantidad promedio de galletas que consumieron los estudiantes del colegio X en el año 2022 es diferente a 65 unidades.

Usando Excel y las fórmulas predeterminadas =DESVEST.M y =PROMEDIO, obtenemos los valores requeridos.

$$\bar{X} = 64.6$$

$$S = 2.39$$

Además, el tamaño de la muestra es 20. Es decir, $n = 20$

Para una prueba bilateral, con 0.05 de significancia y 20-1=19 grados de libertad, la región de rechazo es:

$$R.R. = \langle -\infty; -2.09\rangle \cup \langle 2.09; \infty\rangle$$

El valor 2.09 se obtiene a partir de la tabla de la distribución t de student (buscar en las tablas del apéndice). También se puede usar Excel.

Determinamos el valor de t_c usando la expresión:

$$t_c = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$t_c = \frac{64.6 - 65}{2.39/\sqrt{20}} = -0.75$$

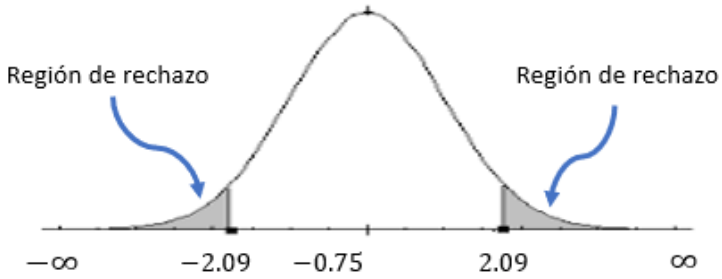


Figura 3.15 Región de rechazo para dos colas.

-0.75 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que la cantidad promedio de galletas que consumieron los estudiantes del colegio X en el año 2022 sea diferente a 65 unidades.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - t de una muestra

Obtenemos:

Tabla 3.20 Estadísticas descriptivas

N	Media	Error estándar de la media	IC de 95% para μ
20	64.6	0.534	[63.553; 65.647]

En la tabla 3.20 se muestra los elementos básicos como el tamaño y la media muestral además del intervalo de confianza para la media.

Tabla 3.21 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: μ= 65
Hipótesis alterna	H_1: $\mu \neq 65$
Valor z	-0.75
Valor p	0.454

En la tabla 3.21 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.454 no es menor que 0.05, se no rechaza la hipótesis nula.

3.9. Prueba de hipótesis para la diferencia de medias

Se utiliza cuando dos poblaciones están distribuidas normalmente y de cada una de ellas se toma una muestra para comparar sus medias. Los tamaños de muestra son n y m respectivamente, y lo que se quiere es verificar si existen diferencias significativas entre las medias de ambas poblaciones.

Prueba bilateral para muestras grandes ($n \geq 30$)

$$Z_c = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

Donde:

$\bar{X}$: es la media de la población 1.
$\bar{Y}$: es la media de la población 2.
σ_x^2: es la varianza de la población 1.
σ_y^2: es la varianza de la población 2.
n: es el tamaño de la muestra 1.
m: es el tamaño de la muestra 1.

Moya y Saravia (2004) sostienen que este procedimiento se puede utilizar aun cuando la población no se distribuya normalmente, pero la muestra debe ser grande. Si no se conoce σ también se puede usar este procedimiento, con la misma condición de trabajar con una muestra grande. Entonces, se usa la desviación estándar muestral S en lugar de σ.

$$Z_c = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

Donde:

S_x^2: es la varianza de la muestra 1.
S_y^2: es la varianza de la muestra 2.

Para muestras pequeñas se utiliza la distribución t con $n + m - 2$ grados de libertad:

$$t_c = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

El desarrollo de la prueba t se verá más adelante con un mayor detalle en el capítulo 5. Pero, ya se puede comprender que es una prueba que permite comparar las medias de dos poblaciones.

Hipótesis bilateral

H_0: $\mu_x = \mu_y$.
H_1: $\mu_x \neq \mu_y$.

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle -\infty; -z\rangle \cup \langle z;\ \infty\rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

4) Si Z_c pertenece a la R.R. rechazamos H_0, en caso contrario no se rechaza.

Ejemplo: El gerente de una fábrica X quiere comparar el desempeño de sus empleados en dos de sus sedes. Por ello toma una muestra de 65 empleados de la sede A y 70 empleados de la sede B. el desempeño se mide a través de una lista de cotejo que brinda un puntaje máximo de 40 puntos al mejor desempeño. La media y la desviación estándar obtenidas en la sede A fueron 32 y 6.5 respectivamente. La media y la desviación estándar obtenidas en la sede B fueron 34 y 3.5 respectivamente.

Título: desempeño laboral en empleados de las sedes A y B de la fábrica X.

Pregunta: ¿Existe diferencia en el desempeño laboral de los empleados de la sede A y los empleados de la sede B?

Objetivo: conocer la diferencia en el desempeño laboral de los empleados de la sede A y los empleados de la sede B.

Hipótesis: el desempeño laboral de los empleados de la sede A es diferente al desempeño laboral de los empleados de la sede B.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable 1: sede.

Variable 2: desempeño laboral.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: el promedio de los puntajes de desempeño laboral de los empleados de la sede A es igual al promedio de los puntajes de desempeño laboral de los empleados de la sede B.

H_1: el promedio de los puntajes de desempeño laboral de los empleados de la sede A es diferente al promedio de los puntajes de desempeño laboral de los empleados de la sede B.

Para un nivel de confianza de 95% en una prueba bilateral tenemos la siguiente región de rechazo:

$$R.R. = \langle -\infty; -1.96\rangle \cup \langle 1.96;\ \infty\rangle$$

Extraemos los datos más importantes del enunciado inicial.

$$\bar{X} = 32 \qquad \bar{Y} = 34$$
$$n = 65 \qquad m = 70$$
$$S_x = 6.5 \qquad S_y = 3.5$$

Como no se conoce σ, utilizamos el estadístico de prueba:

$$Z_c = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

$$Z_c = \frac{32 - 34}{\sqrt{\frac{6.5^2}{65} + \frac{3.5^2}{70}}} = -2.20$$

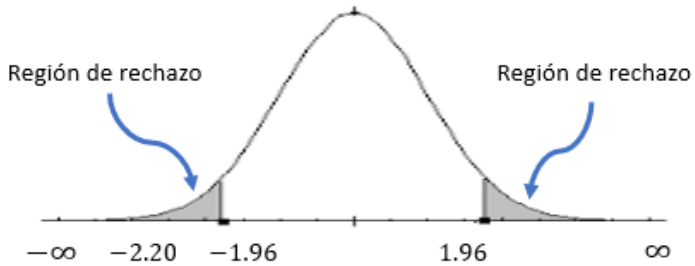


Figura 3.16 Región de rechazo para dos colas.

-2.20 pertenece a la región de rechazo. Entonces, se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que el promedio de los puntajes de desempeño laboral de los empleados de la sede A es diferente al promedio de los puntajes de desempeño laboral de los empleados de la sede B.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - t de 2 muestras

Los resultados son los siguientes:

Tabla 3.22 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\mu_1 - \mu_2 = 0$
Hipótesis alterna	H_1: $\mu_1 - \mu_2 \neq 0$

Valor T	GL	Valor p
-2.20	96	0.030

En la tabla 3.22 se observa la significancia igual a 0.03 y es menor a 0.05. Entonces, se rechaza H_0.

Una vez que se ha verificado la diferencia de medias. Se puede establecer una nueva prueba de hipótesis para verificar que la media de la sede B es mayor a la media de la sede A.

Capítulo 4: Prueba de hipótesis con respecto a la proporción

Estas pruebas de hipótesis son parecidas a las pruebas relativas a las medias. En el proceso de prueba de hipótesis para la proporción también se utiliza la distribución normal.

En este caso el análisis esta referido a eventos con distribución binomial. Eventos donde solo se consideran dos resultados posibles. Como ya hemos visto, en el desarrollo del intervalo de confianza para la proporción, estos posibles resultados hacen referencia al éxito o al fracaso. Consideremos que p representa el tanto por ciento de éxitos.

Por ejemplo, si queremos conocer la proporción muestral de personas adultas en un centro comercial, tomamos una muestra representativa de 450 personas y de estas 325 son adultas. Entonces, la proporción muestral p se obtiene de la siguiente manera:

$$p = \frac{eventos\ favorables}{total\ de\ ensayos} = \frac{325}{450} = 0.72 = 72\%$$

No confundir la proporción muestral p con la significancia calculada p.

A continuación, se muestran algunos ejemplos de hipótesis para la proporción:

H_1: la proporción de estudiantes puntuales en la universidad X excede a 90%.

H_1: la proporción de electores a favor del candidato X es menor a 20%.

H_1: la proporción de familias limeñas que consumen pan en su desayuno es diferente a 86%.

En esta prueba de hipótesis se compara la proporción poblacional con una proporción poblacional referencial.

Considerando π_0 como la proporción poblacional referencial. Las posibles hipótesis a probar son las siguientes:

Pruebas unilaterales

$H_0: \pi = \pi_0.$
$H_1: \pi < \pi_0.$

$H_0: \pi = \pi_0.$
$H_1: \pi > \pi_0$

Prueba bilateral

$H_0: \pi = \pi_0.$
$H_1: \pi \neq \pi_0$

4.1. Prueba de la cola inferior

Se utiliza cuando queremos verificar si nuestra proporción poblacional π es menor que una proporción referencial π_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \pi = \pi_0.$
$H_1: \pi < \pi_0.$

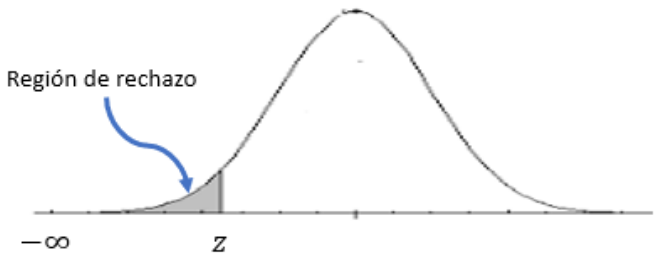


Figura 4.1 Región de rechazo para una cola inferior.

Para muestras grandes n≥30

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle -\infty; z \rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

x es el número de casos a favor

4) Si Z_c pertenece a la R.R. rechazamos H_0 , en caso contrario no se rechaza.

Usaremos los siguientes valores z:

Tabla 4.1 Valor z según el nivel de confianza.

Nivel de confianza	Valor z asociado
90%	-1.28
95%	-1.64
99%	-2.33

Ejemplo 1: según el Organismo Supervisor de Inversión Privada en Telecomunicaciones – OSIPTEL (2022) a finales del 2021, el 87,7 % de hogares peruanos, cuentan con acceso a internet sea este fijo o móvil. Un investigador, luego de un trabajo decampo previo, considera que la proporción de hogares que cuenta con internet en el distrito de Ancón debe ser menor a la proporción que nos brinda OSIPTEL. Por ello, selecciona una muestra representativa de 350 hogares. Los resultados muestran que solo 243 hogares analizados cuentan con servicio de internet.

Título: conexión a internet en hogares del distrito de Ancón.

Pregunta: ¿Cuál es la proporción de hogares del distrito de Ancón que cuentan con conexión a internet?

Objetivo: conocer la proporción de hogares del distrito de Ancón que cuentan con conexión a internet.

Hipótesis: la proporción de hogares del distrito de Ancón que cuentan con conexión a internet es menor a 87.7%.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: conexión a internet.

Elabora la prueba de hipótesis para verificar si la proporción de hogares que cuentan con conexión a internet es menor al 87.7% señalado por OSIPTEL. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola inferior.

H_0: la proporción de hogares del distrito de Ancón que cuentan con conexión a internet es igual a 87.7%.

H_1: la proporción de hogares del distrito de Ancón que cuentan con conexión a internet es menor a 87.7%.

Extraemos los datos necesarios de forma resumida.

$$x = 243 \qquad n = 350 \qquad \pi_0 = 87.7\% = 0.877$$

Para una prueba unilateral con cola inferior y con 95% de confianza, la región de rechazo será:

$$R.R. = \langle -\infty; -1.64 \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{243 - 350(0.877)}{\sqrt{350(0.877)(1-0.877)}} = -10.41$$

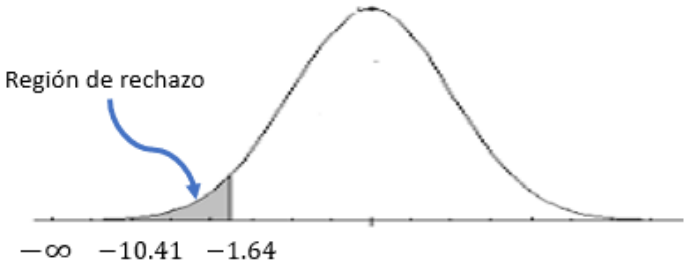


Figura 4.2 Región de rechazo para una cola inferior.

-10.41 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que la proporción de hogares del distrito de Ancón que cuentan con conexión a internet es menor a 87.7%.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Los resultados se muestran a continuación:

Tabla 4.2 Estadísticas descriptivas

N	Evento	Muestra p	Límite superior de 95% para π
350	243	0.694286	0.734792

En la tabla 4.2 se muestra los elementos básicos como la proporción muestral p y el límite superior del intervalo de confianza para π.

Tabla 4.3 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: π = 0.877
Hipótesis alterna	H_1: π < 0.877
Valor z	-10.41
Valor p	0.000

En la tabla 4.3 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la

significancia es de 0.05. Dado que el valor p=0.000 es menor que 0.05, se rechaza la hipótesis nula.

Ejemplo 2: Zavaleta (2017) afirma que, en el Perú, la anemia afecta al 43,6% de los niños menores de 3 años. Un investigador especialista considera que, dadas las políticas en salud para combatir la anemia infantil, la proporción de niños menores de 3 años que padecen de anemia debe ser mucho menor en la ciudad de Abancay en el año 2022. Por ello, toma una muestra representativa de 455 niños menores de 3 años. De este grupo de niños, 202 padecen de anemia.

Título: prevalencia de anemia infantil en la ciudad de Abancay.

Pregunta: ¿Cuál es la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay?

Objetivo: conocer la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay.

Hipótesis: la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay es menor a 43.6%.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: estado nutricional.

Elabora la prueba de hipótesis para verificar si la proporción de niños menores de 3 años que padecen anemia es menor al 43.6% señalado por Zavaleta. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola inferior.

H_0: la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay es igual a 43.6%.

H_1: la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay es menor a 43.6%.

Extraemos los datos necesarios de forma resumida.

$$x = 202 \qquad n = 455 \qquad \pi_0 = 43.6\% = 0.436$$

Para una prueba unilateral con cola inferior y con 95% de confianza, la región de rechazo será:

$$R.R. = \langle -\infty; -1.64 \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{202 - 455(0.436)}{\sqrt{455(0.436)(1-0.436)}} = 0.34$$

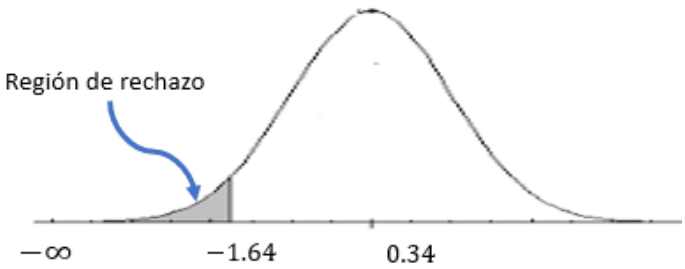


Figura 4.3 Región de rechazo para una cola inferior.

0.34 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que la proporción de niños menores de 3 años que padecen anemia en la ciudad de Abancay sea menor a 43.6%.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Los resultados se muestran a continuación:

Tabla 4.4 Estadísticas descriptivas

N	Evento	Muestra p	Límite superior de 95% para π
455	202	0.443956	0.482269

En la tabla 4.4 se muestra los elementos básicos como la proporción muestral p y el límite superior del intervalo de confianza para π.

Tabla 4.5 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi = 0.436$
Hipótesis alterna	H_1: $\pi < 0.436$
Valor z	0.34
Valor p	0.634

En la tabla 4.5 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.634 no es menor que 0.05, no se rechaza la hipótesis nula.

4.2. Prueba de la cola superior

Se utiliza cuando queremos verificar si nuestra proporción poblacional π es mayor que una proporción referencial π_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \pi = \pi_0.$
$H_1: \pi > \pi_0.$

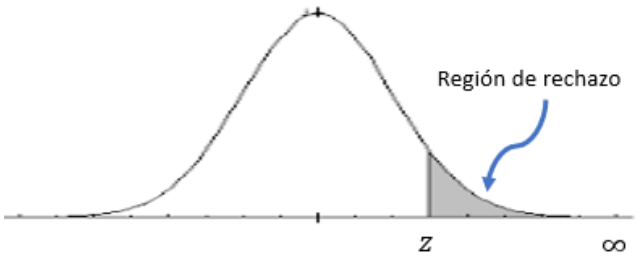


Figura 4.4 Región de rechazo para una cola superior.

Para muestras grandes n≥30

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle z; \infty \rangle$
3) Determina el valor de Z_c :

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

4) Si Z_c pertenece a la R.R. se rechaza H_0, en caso contrario no se rechaza.

Usaremos los siguientes valores z:

Tabla 4.6 Valor z según el nivel de confianza.

Nivel de confianza	Valor z asociado
90%	1.28
95%	1.64
99%	2.33

Ejemplo 1: En una investigación se quiere conocer la proporción de estudiantes de secundaria de la ciudad de Lima que consume tabaco. Por ello, se realiza una encuesta a 660 estudiantes y se observa que 75 de ellos son consumidores de tabaco.

Además, se ha encontrado información del Ministerio de Salud del Perú que indica que 10% de los adolescentes en el Perú son consumidores de tabaco (MINSA, 2018). Sin embargo, hay antecedentes más recientes que indican que dicha proporción puede ser mayor al 10% señalado.

Título: consumo de tabaco estudiantes de la ciudad de Lima.

Pregunta: ¿Cuál es la proporción de estudiantes que consume tabaco en la ciudad de Lima?

Objetivo: conocer la proporción de estudiantes que consume tabaco en la ciudad de Lima.

Hipótesis: la proporción de estudiantes que consume tabaco en la ciudad de Lima es mayor al 10% señalado por el MINSA.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: consumo de tabaco.

Elabora la prueba de hipótesis para verificar si la proporción de estudiantes que consume tabaco es mayor al 10% señalado por el MINSA. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola superior.

H_0: la proporción poblacional de estudiantes que consume tabaco en la ciudad de Lima no es mayor al 10% señalado por el MINSA.

H_1: la proporción poblacional de estudiantes que consume tabaco en la ciudad de Lima es mayor al 10% señalado por el MINSA.

Extraemos los datos necesarios de forma resumida.

$$x = 75 \qquad n = 660 \qquad \pi_0 = 10\% = 0.1$$

Para una prueba unilateral con cola superior y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle 1.64;\ \infty \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{75 - 660(0.1)}{\sqrt{660(0.1)(1-0.1)}} = 1.17$$

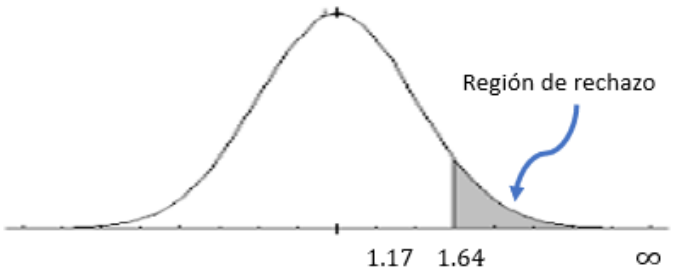


Figura 4.5 Región de rechazo para una cola superior.

1.17 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que la proporción poblacional de estudiantes que consume tabaco en la ciudad de Lima sea mayor al 10% señalado por el MINSA.

También se puede utilizar el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Los resultados se muestran a continuación:

Tabla 4.7 Estadísticas descriptivas

N	Evento	Muestra p	Límite inferior de 95% para π
660	75	0.113636	0.093317

En la tabla 4.7 se muestra los elementos básicos como la proporción muestral p y el límite inferior del intervalo de confianza para π.

Tabla 4.8 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi = 0.1$
Hipótesis alterna	H_1: $\pi > 0.1$
Valor z	1.17
Valor p	0.121

En la tabla 4.8 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.121 no es menor que 0.05, no se rechaza la hipótesis nula.

Ejemplo 2: el diario El Comercio (2021), en su página web informa que, "en la encuesta realizada por Trabajando.com, más del 50% de los peruanos sufre de estrés laboral crónico como consecuencia del trabajo remoto y el desempleo". Un grupo de investigadores considera que, dada la coyuntura política y social del 2022 en el Perú, este tanto por ciento debe ser mayor. Por ello, se toma una muestra representativa de 589 trabajadores y, luego de ser evaluados con un test, se observa que 345 de ellos padece de estrés laboral crónico.

Título: estrés laboral crónico en trabajadores de la ciudad de Lima.

Pregunta: ¿Cuál es la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima?

Objetivo: determinar la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima.

Hipótesis: la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima es mayor a 50%.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: nivel de estrés.

Elabora la prueba de hipótesis para verificar si la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima es mayor a 50%. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba unilateral de cola superior.

H_0: la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima es igual a 50%.

H_1: la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima es mayor a 50%.

Extraemos los datos necesarios de forma resumida.

$$x = 345 \qquad n = 589 \qquad \pi_0 = 50\% = 0.5$$

Para una prueba unilateral con cola superior y con 95% de confianza, la región de rechazo es:

$$R.R. = \langle 1.64;\ \infty \rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{345 - 589(0.5)}{\sqrt{589(0.5)(0.5)}} = 4.16$$

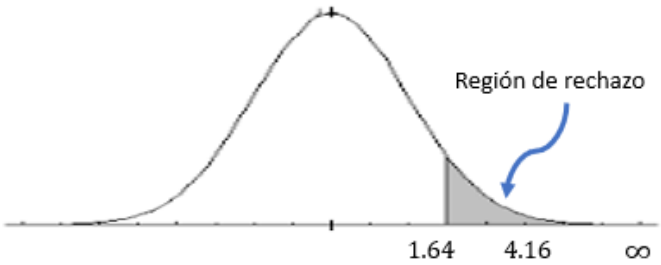


Figura 4.6 Región de rechazo para una cola superior.

4.16 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que la proporción de trabajadores que padece estrés laboral crónico en la ciudad de Lima es mayor a 50%.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Obtenemos:

Tabla 4.9 Estadísticas descriptivas

N	Evento	Muestra p	Límite inferior de 95% para π
589	345	0.585739	0.552353

En la tabla 4.9 se muestra los elementos básicos como la proporción muestral p y el límite inferior del intervalo de confianza para π.

Tabla 4.10 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi = 0.5$
Hipótesis alterna	H_1: $\pi > 0.5$
Valor z	4.16
Valor p	0.000

En la tabla 4.10 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.000 es menor que 0.05, se rechaza la hipótesis nula.

4.3. Prueba bilateral o de dos colas

Se utiliza cuando queremos verificar que nuestra proporción poblacional π es diferente a una proporción referencial π_0 obtenida a partir de los antecedentes de la investigación.

Simbólicamente las hipótesis se redactan de la siguiente manera:

$H_0: \pi = \pi_0.$
$H_1: \pi \neq \pi_0.$

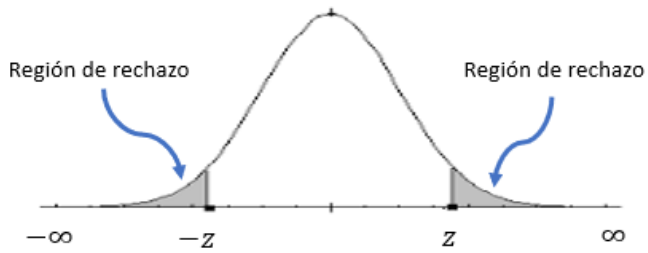


Figura 4.7 Región de rechazo para dos colas.

Para muestras grandes n≥30

Procedimiento:

1) De acuerdo al nivel de confianza, se determina el valor z
2) Elabora la Región de Rechazo: $R.R. = \langle -\infty; -z \rangle \cup \langle z; \infty \rangle$

3) Determina el valor de Z_c :

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

4) Si Z_c pertenece a la R.R. rechazamos H_0, en caso contrario no se rechaza.

Usaremos los siguientes valores:

Tabla 4.11 Valor z según el nivel de confianza.

Nivel de confianza	Valor z asociado
90%	1.645
95%	1.96
99%	2.58

Ejemplo 1: los resultados de la Encuesta Demográfica y de Salud Familiar – ENDES 2015, revelaron que el 35.5% de personas de 15 y más años de edad presenta sobrepeso (INEI, 2015). El responsable de un centro de salud considera que, dada la coyuntura actual del año 2022, la proporción de personas con sobrepeso debe haber cambiado con respecto a los resultados del 2015. Por ello, toma una muestra representativa de 375 personas de la urbanización X y verifica que 152 de estas personas presentan sobrepeso.

Título: prevalencia de sobrepeso en personas de 15 años a más en la urbanización X.

Pregunta: ¿Cuál es la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X?

Objetivo: determinar la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X.

Hipótesis: la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X es diferente al 35.5% señalado por el INEI.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: Índice de masa corporal.

Elabora la prueba de hipótesis para verificar si la proporción de personas con sobrepeso es diferente al 35.5% señalado por el INEI. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X es igual al 35.5% señalado por el INEI.

H_1: la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X es diferente al 35.5% señalado por el INEI.

Extraemos los datos necesarios de forma resumida.

$$x = 152 \qquad n = 375 \qquad \pi_0 = 35.5\% = 0.355$$

Para una prueba bilateral y con 95% de confianza, la región de rechazo será:

$$R.R. = \langle -\infty; -1.96\rangle \cup \langle 1.96; \infty\rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{152 - 375(0.355)}{\sqrt{375(0.355)(1-0.355)}} = 2.04$$

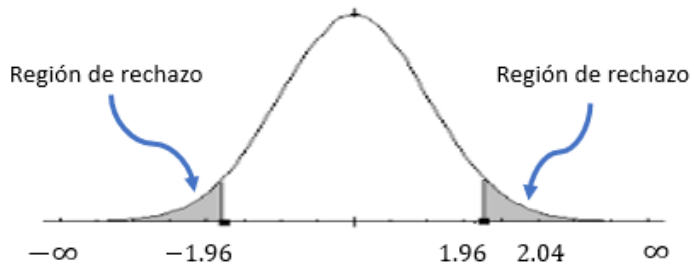


Figura 4.8 Región de rechazo para dos colas.

2.04 pertenece a la región de rechazo. Por lo tanto, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que la prevalencia del sobrepeso en personas de 15 años a más en la urbanización X es diferente al 35.5% señalado por el INEI.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Obtenemos:

Tabla 4.12 Estadísticas descriptivas

N	Evento	Muestra p	IC de 95% para π
375	152	0.40533	$\langle 0.355643;\ 0.455204 \rangle$

En la tabla 4.12 se muestra los elementos básicos como la proporción muestral p y el intervalo de confianza para π.

Tabla 4.13 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi = 0.355$
Hipótesis alterna	H_1: $\pi \neq 0.355$
Valor z	2.04
Valor p	0.042

En la tabla 4.13 se muestra la prueba de hipótesis completa. El nivel de confianza que utilizamos en este ejemplo es de 95%, entonces la significancia es de 0.05. Dado que el valor p=0.042 es menor que 0.05, se rechaza la hipótesis nula.

Ejemplo 2: según el diario el comercio (2019), en un informe de Concortv, las plataformas Netflix o HBO son utilizadas por el 49% de los limeños. En una investigación sobre el uso de diferentes plataformas digitales en la ciudad de Lima se quiere contrastar la proporción de usuarios de las plataformas Netflix o HBO del 2019 con la proporción respectiva para el año 2022. Por ello se toma una muestra representativa de 540 limeños y se observa que 283 de ellos son usuarios de las plataformas mencionadas.

Título: uso de las plataformas digitales para entretenimiento en los habitantes de la ciudad de Lima.

Pregunta: ¿Cuál es la proporción de limeños que utilizan las plataformas Netflix o HBO?

Objetivo: determinar la proporción de limeños que utilizan las plataformas Netflix o HBO.

Hipótesis: la proporción de limeños que utilizan las plataformas Netflix o HBO es diferente a 49%.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable: uso de las plataformas Netflix o HBO.

Elabora la prueba de hipótesis para verificar si la proporción de limeños que utilizan las plataformas Netflix o HBO es diferente al 49% señalado por la empresa Concortv. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: la proporción de limeños que utilizan las plataformas Netflix o HBO es igual a 49%.

H_1: la proporción de limeños que utilizan las plataformas Netflix o HBO es diferente a 49%.

Extraemos los datos necesarios de forma resumida.

$$x = 540 \qquad n = 283 \qquad \pi_0 = 49\% = 0.49$$

Para una prueba bilateral y con 95% de confianza, la región de rechazo será:

$$R.R. = \langle -\infty; -1.96\rangle \cup \langle 1.96; \infty\rangle$$

Determinamos el valor de Z_c usando la expresión:

$$Z_c = \frac{x - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}}$$

$$Z_c = \frac{283 - 540(0.49)}{\sqrt{540(0.49)(1-0.49)}} = 1.58$$

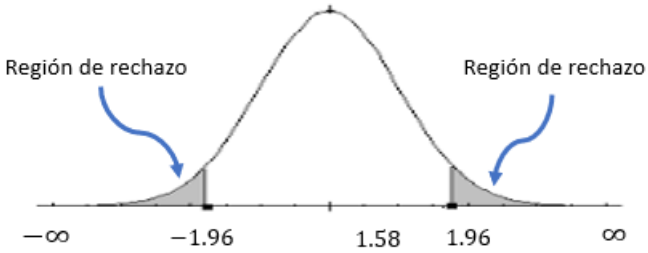


Figura 4.9 Región de rechazo para dos colas.

1.58 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que la proporción de limeños que utilizan las plataformas Netflix o HBO sea diferente a 49%.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 1 proporción

Obtenemos:

Tabla 4.14 Estadísticas descriptivas

N	Evento	Muestra p	IC de 95% para π
540	283	0.524074	$\langle 0.481951;\ 0.566197 \rangle$

En la tabla 4.14 se muestra los elementos básicos como la proporción muestral p y el intervalo de confianza para π.

Tabla 4.15 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi = 0.49$
Hipótesis alterna	H_1: $\pi \neq 0.49$
Valor z	1.58
Valor p	0.113

En la tabla 4.15 se muestra la prueba de hipótesis completa. La significancia para este ejemplo es de 0.05. Dado que el valor p=0.113 no es menor que 0.05, no se rechaza la hipótesis nula.

4.4. Prueba de hipótesis para la diferencia de proporciones

Para muestras grandes n≥30.

Procedimiento:

Es igual al procedimiento de las pruebas unilaterales o bilaterales, para muestras grandes, cuando se comparan las medias. Para muestras grandes, la distribución de la diferencia de las proporciones poblaciones $\pi_1 - \pi_2$ es aproximadamente normal.

El valor del estadístico de prueba es Z_c (Anderson et al., 2012):

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

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

Donde:

p_1 y p_2 son las proporciones muestrales.
n_1 y n_2 son los tamaños de las muestras.
x_1 y x_2 son los casos o eventos a favor.

Ejemplo: mediante un estudio de mercado se quiere determinar si existen diferencias entre el consumo de cerveza en lata en los distritos del sur y los distritos del norte de la ciudad de Trujillo. Para ello, se toma una muestra representativa de 195 consumidores de los distritos del sur y 235 consumidores de los distritos del norte. Entre los encuestados de los distritos del sur, 101 consumen cerveza en lata y entre los encuestados del distrito del norte 102 consumen cerveza en lata.

Título: Consumo de cerveza enlatada en los habitantes de la ciudad de Trujillo.

Pregunta: ¿Existe diferencia en el consumo de cerveza en los distritos del sur y los distritos del norte de la ciudad de Trujillo?

Objetivo: determinar si existe diferencia en el consumo de cerveza enlatada en los distritos del sur y los distritos del norte de la ciudad de Trujillo.

Hipótesis: en la ciudad de Trujillo, la proporción de habitantes de los distritos del sur que consume cerveza enlatada es diferente a la proporción de habitantes de los distritos del norte que consumen cerveza enlatada.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable 1: distrito de residencia.

Variable 2: tipo de envase de cerveza preferido.

Elabora la prueba de hipótesis para verificar si existe diferencia significativa entre la proporción de los grupos analizados. Utiliza un nivel de confianza de 95%.

Solución: primero se plantean las hipótesis para una prueba bilateral.

H_0: la proporción de habitantes de los distritos del sur, que consume cerveza enlatada, es igual a la proporción de habitantes de los distritos del norte, que consume cerveza enlatada.

H_1: la proporción de habitantes de los distritos del sur, que consume cerveza enlatada, es diferente a la proporción de habitantes de los distritos del norte, que consume cerveza enlatada.

Extraemos los datos necesarios de forma resumida.

$$x_1 = 101 \qquad x_2 = 102$$

$$n_1 = 195 \qquad n_2 = 235$$

$$p_1 = \frac{101}{195} = 0.52 \qquad p_2 = \frac{102}{235} = 0.43$$

$$p = \frac{101 + 102}{195 + 235} = 0.47$$

Para una prueba bilateral con 95% de confianza, la región de rechazo será:

$$R.R. = \langle -\infty; -1.96\rangle \cup \langle 1.96; \infty\rangle$$

Determinamos el valor de Z_c usando la expresión:

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

$$Z_c = \frac{0.52 - 0.43}{\sqrt{0.47(1-0.47)\,(\frac{1}{195} + \frac{1}{235})}} = 1.86$$

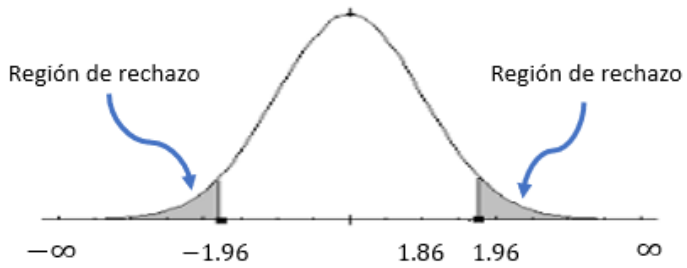


Figura 4.10 Región de rechazo para dos colas.

1.86 no pertenece a la región de rechazo. Por lo tanto, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que la proporción de habitantes de los distritos del sur que consume cerveza enlatada sea diferente a la proporción de habitantes de los distritos del norte que consumen cerveza enlatada.

Utilizando el programa Minitab con la siguiente ruta:

Estadísticas - Estadísticas básicas - 2 proporciones

Obtenemos:

Tabla 4.16 Estimación de la diferencia

Diferencia	IC de 95% para la diferencia
0.0839062	$\langle -0.010614;\ 0.178427 \rangle$

IC basado en la aproximación a la normal

En la tabla 4.16 se muestra que la diferencia de proporciones muestrales es mayor a 8%, pero el intervalo de confianza es muy amplio (note que el límite superior del intervalo llega hasta casi 18% de diferencia). Sin embargo, esa diferencia de proporciones, aparentemente grande, no es suficiente para afirmar que exista diferencia significativa entre las proporciones poblacionales. Esto se evidencia en la tabla 4.17 donde el p=0.082 es mayor que 0.05 no permite rechazar la hipótesis nula.

Tabla 4.17 Prueba de hipótesis en Minitab

Hipótesis nula	H_0: $\pi_1 - \pi_2 = 0$	
Hipótesis alterna	H_1: $\pi_1 - \pi_2 \neq 0$	
Método	Valor Z	Valor p
Aproximación normal	1.74	0.082
Exacta de Fisher		0.099

Capítulo 5: Prueba t

Es una prueba de comparación de dos poblaciones. El objetivo de esta prueba es comparar las medias de los grupos y determinar que la diferencia no se deba al azar (Juárez, Villatoro y López, 2002).

Para utilizar esta prueba se deben verificar ciertas condiciones o supuestos, entre estos supuestos tenemos la normalidad, la homocedasticidad y el nivel de medición escalar. El nivel de medición escalar hace referencia a las variables cuantitativas.

5.1. Prueba de normalidad

Esta prueba permite conocer si un conjunto de datos proviene de una población con distribución normal, para ello se debe realizar una prueba de hipótesis y podemos apoyarnos en el programa SPSS.

SPSS brinda dos pruebas de normalidad Kolmorogov - Smirnof y Shapiro – Wilk. El criterio para decidir que prueba utilizar depende del tamaño de la muestra. En la tabla 5.1 se observa las condiciones para utilizar estas pruebas:

Tabla 5.1 Pruebas de normalidad

Kolmorogov - Smirnof	Shapiro - Wilk
$n \geq 50$	$n < 50$

Para realizar la prueba seguimos el proceso:

1) Planteamos las hipótesis:
   H_0: Los datos provienen de una población con distribución normal.
   H_1: Los datos no provienen de una población con distribución normal.
2) Asignamos la significancia: $\alpha = 0.05$

3) Identificamos el p-valor: Sig. Bilateral según SPSS
4) Decisión: Si $p < \alpha$ se rechaza H_0, en caso contrario, no se rechaza.

La ruta para realizar la prueba de hipótesis en SPSS es la siguiente:

Analizar-Estadísticos descriptivos-Explorar-Gráficos-Gráficos de normalidad

Ejemplo: determina si los siguientes datos provienen de una población con distribución normal.

Tabla 5.2 Matriz de datos.

Individuo	Edad
1	45
2	40
3	35
4	40
5	40
6	40
7	35
8	42
9	42
10	35
11	40
12	35
13	30
14	42
15	42

Solución: primero planteamos las hipótesis.

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

Usando SPSS obtenemos la siguiente tabla:

Tabla 5.3 Prueba de normalidad en SPSS

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Edad	0.890	15	0.067

En la tabla 5.3 observamos que la significancia calculada es 0.067 y es mayor que 0.05. por lo tanto, no se rechaza H_0. En conclusión, Los datos provienen de una población con distribución normal.

5.2. Prueba de homocedasticidad

La homocedasticidad también es conocida como la igualdad de varianzas. Esta prueba permite conocer si los conjuntos de datos provienen de poblaciones con varianzas iguales, para ello se debe realizar una prueba de hipótesis y podemos utilizar en el programa SPSS.

Para realizar la prueba seguimos el proceso:

1) Planteamos las hipótesis:
   H_0: Los datos provienen de poblaciones con varianzas iguales.
   H_1: Los datos no provienen de poblaciones con varianzas iguales.
2) Asignamos la significancia: $\alpha = 0.05$
3) Identificamos el p-valor: Sig. Bilateral según SPSS
4) Decisión: Si $p < \alpha$ se rechaza H_0, en caso contrario, no se rechaza.

La ruta para realizar la prueba de hipótesis en SPSS es la siguiente:

Analizar-Comparar medias-ANOVA de un factor-Opciones

Ejemplo: determina si los siguientes datos provienen de poblaciones con varianzas iguales.

Tabla 5.4 Matriz de datos.

Grupo 1	Grupo 2
12	16
10	13
9	13
16	13
14	14
12	10
14	12
14	13
14	14
15	14
14	14

Solución: planteamos las hipótesis.

H_0: Los datos provienen de poblaciones con varianzas iguales.
H_1: Los datos no provienen de poblaciones con varianzas iguales.

Los resultados que nos brinda SPSS son los siguientes:

Tabla 5.5 Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
Se basa en la media	2.136	1	20	0.159
Se basa en la mediana	0.532	1	20	0.474
Se basa en la mediana y con gl ajustado	0.532	1	16.777	0.476
Se basa en la media recortada	1.953	1	20	0.178

Luego de definir la significancia 0.05, observamos la tabla 5.5. En dicha tabla p=0.159 no es menor que 0.05, entonces, no se rechaza H_0. Es decir, Los datos provienen de poblaciones con varianzas iguales.

5.3. Prueba t para muestras independientes

En este caso se trabaja con poblaciones distintas y se desea comparar sus medias. Los tamaños de las muestras no necesariamente tienen que ser iguales.

Para utilizar esta prueba se deben verificar las siguientes condiciones o supuestos:

- Nivel de medición de Intervalo o de razón (Escalar).
- Selección aleatoria de los grupos.
- Distribución normal de la VD en cada grupo (Normalidad).
- Homogeneidad de las varianzas de la VD (Homocedasticidad).

El estadístico de prueba está determinado por:

$$t_c = \frac{\bar{X}_1 - \bar{X}_2}{S_{D\bar{X}}}$$

Donde $\bar{X}_1$ y $\bar{X}_2$ son las medias muestrales y $S_{D\bar{X}}$ es el error estándar de la diferencia entre medias.

$$S_{D\bar{X}} = \sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

n_1 y n_2 son los tamaños de muestra, S^2 es la varianza del grupo

$$S^2 = \frac{\sum(x_i - \bar{X}_1)^2 + \sum(x_j - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$\sum(x_i - \bar{X}_1)^2 = \sum x_i^2 \; - \frac{(\sum x_i)^2}{n_1}$$

$$\sum(x_j - \bar{X}_2)^2 = \sum x_j^2 \; - \frac{(\sum x_j)^2}{n_2}$$

Una vez que se ha calculado el estadístico de prueba t_c, se procede de la siguiente manera para tomar la decisión estadística de rechazar o no rechazar la hipótesis nula.

Regla de decisión

Si $|t_c| \geq t$ entonces, se rechaza H_o
Si $|t_c| < t$ entonces, no se rechaza H_o

Recuerda que en la distribución de probabilidad utilizaremos $n_1 + n_2 - 2$ grados de libertad. Por ello, ubicamos en la tabla el valor t asociado al nivel de significancia elegido.

Ejemplo 1: En una investigación se desea comparar las calificaciones de los estudiantes de las secciones A y B del quinto grado de secundaria del colegio X. Por ello, se toma una muestra de estudiantes de cada sección y se observan las calificaciones. Los resultados se muestran a continuación:

Tabla 5.6 Matriz de datos.

Calificaciones	
Sección A	Sección B
16	15
12	19
14	17
14	16
13	12
15	11
15	15
14	15
15	16
14	16
14	15
17	16
15	

Título: rendimiento académico en estudiantes de quinto grado de secundaria del colegio X.

Pregunta: ¿Qué rendimiento académico tienen los estudiantes de las secciones A y B del quinto grado de secundaria del colegio X?

Objetivo: comparar el rendimiento académico en los estudiantes de las secciones A y B del quinto grado de secundaria del colegio X.

Hipótesis: el rendimiento académico en los estudiantes de las secciones A y B del quinto grado de secundaria del colegio X es diferente.

Variable x: sección.

Variable y: calificación de los estudiantes.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas entre las calificaciones de ambas secciones del quinto grado de secundaria. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias podemos trabajar con el promedio de calificaciones de cada sección.

Los grupos son diferentes. Por lo tanto, se puede utilizar la prueba t para muestras independientes.

Para usar la prueba t se debe verificar si se cumple el supuesto de normalidad. La muestra es pequeña, por ello, se utiliza la prueba de Shapiro-Wilk en el programa SPSS.

Tabla 5.7 Prueba de normalidad en SPSS

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Sección A	0,941	13	0,464
Sección B	0,891	12	0,123

Planteamos las hipótesis para cada uno de los grupos:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

En la tabla 5.7 se observa que la significancia para la sección A y la sección B es mayor a 0.05. Entonces, no se rechaza H_0; se cumple el supuesto de normalidad para ambos grupos.

Tabla 5.8 Prueba de homogeneidad de varianzas

		F	Significancia
Consumo	Se asumen varianzas iguales	0.958	0,338

Para la prueba de homocedasticidad planteamos las hipótesis:

H_0: Los datos provienen de poblaciones con varianzas iguales.
H_1: Los datos no provienen de poblaciones con varianzas iguales.

En la tabla 5.8 se observa el resultado de la prueba de homocedasticidad en SPSS y se verifica que la significancia 0.338 es mayor a 0.05. Entonces, se cumple el supuesto de la igualdad de varianzas.

Puesto que se verifican los supuestos de normalidad y homocedasticidad, podemos utilizar la prueba t para muestras independientes. Luego, planteamos las hipótesis para una prueba bilateral:

H_0: el promedio de calificaciones en los estudiantes de la sección A es igual al promedio de calificaciones en los estudiantes de la sección B.

H_1: el promedio de calificaciones en los estudiantes de la sección A es diferente al promedio de calificaciones en los estudiantes de la sección B.

Solo en este caso utilizaremos el procedimiento con el desarrollo completo de los cálculos, en ejemplos posteriores se utilizará programas estadísticos.

Tabla 5.9 Sumatorias de los datos y los cuadrados de los datos.

x_1	x_2	x_1^2	x_2^2
16	15	256	225
12	19	144	361
14	17	196	289
14	16	196	256
13	12	169	144
15	11	225	121
15	15	225	225
14	15	196	225
15	16	225	256
14	16	196	256
14	15	196	225
17	16	289	256
15		225	
$\sum x_1$=188	$\sum x_2$=183	$\sum x_1^2$=2738	$\sum x_2^2$=2839

Definimos los tamaños de muestra:

$$n_1 = 13$$

$$n_2 = 12$$

Calculamos los promedios:

$$\bar{X}_1 = \frac{188}{13} = 14.46$$

$$\bar{X}_2 = \frac{183}{12} = 15.25$$

Calculamos la suma de cuadrados para la sección A:

$$\sum (x_i - \bar{X}_1)^2 = \sum x_i^2 \ - \frac{(\sum x_i)^2}{n_1}$$

$$\sum (x_i - \bar{X}_1)^2 = 2738 - \frac{188^2}{13} = 19.23$$

Calculamos la suma de cuadrados para la sección B:

$$\sum (x_j - \bar{X}_2)^2 = \sum x_j^2 \ - \frac{(\sum x_j)^2}{n_2}$$

$$\sum (x_j - \bar{X}_2)^2 = 2839 - \frac{183^2}{12} = 48.25$$

Calculamos la varianza del grupo:

$$S^2 = \frac{\sum (x_i - \bar{X}_1)^2 + \sum (x_j - \bar{X}_2)^2}{n_1 + n_2 - 2}$$

$$S^2 = \frac{19.23 + 48.25}{13 + 12 - 2} = 2.934$$

Calculamos el error estándar de la diferencia entre medias:

$$S_{D\bar{X}} = \sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{2.934 \left(\frac{1}{13} + \frac{1}{12}\right)} = 0.686$$

El estadístico de prueba está determinado por:

$$t_c = \frac{\bar{X}_1 - \bar{X}_2}{S_{D\bar{X}}} = \frac{14.46 - 15.25}{0.686} = -1.15$$

Para una prueba bilateral, con una significancia de 0.05 y $13 + 12 - 2 = 23$ grados de libertad, la tabla de distribución de probabilidad t nos indica el siguiente valor:

$$t = 2.07$$

La tabla para ubicar el valor crítico de t se puede ubicar en el Apéndice de este libro. Aunque también se puede usar Excel, como ya se explicó en el capítulo 2.

Utilizando la regla de decisión verificamos que $|t_c| = 1.15$ no es mayor que $t = 2.07$. Por lo tanto, no se rechaza H_0.

En conclusión, no existe evidencia estadísticamente significativa para afirmar que el promedio de calificaciones en los estudiantes de la sección A sea diferente al promedio de calificaciones de los estudiantes de la sección B.

Probablemente este resultado nos haga reflexionar sobre la diferencia de las medias muestrales. Las medias $\bar{X}_1 = 14.46$ y $\bar{X}_2 = 15.25$ son diferentes, pero esta diferencia no basta para inferir que exista diferencia significativa entre las medias poblacionales.

Para realizar la prueba t también se puede usar Excel. La ruta es la siguiente:

Datos-Análisis de datos-prueba t para dos muestras suponiendo varianzas iguales.

Tabla 5.10 Prueba t para dos muestras suponiendo varianzas iguales

	Variable 1	Variable 2
Media	14.4615385	15.25
Varianza	1.6025641	4.38636364
Observaciones	13	12
Varianza agrupada	2.93394649	
Diferencia hipotética de las medias	0	
Grados de libertad	23	
Estadístico t	-1.14986463	
P(T<=t) una cola	0.13100799	
Valor crítico de t (una cola)	1.71387153	
P(T<=t) dos colas	0.26201598	
Valor crítico de t (dos colas)	2.06865761	

El objetivo de la investigación consiste en determinar si los promedios son iguales o diferentes. Por ello, se utilizará la probabilidad de dos colas.

Regla de decisión:

Si la significancia calculada es menor que 0.05, se rechaza la hipótesis nula.

En la Tabla 5.10 podemos verificar que la significancia para una prueba bilateral (dos colas) es 0.262 y este valor no es menor a 0.05. Entonces, no se rechaza la hipótesis nula.

Para realizar la prueba t también se puede usar el programa Minitab. La ruta es la siguiente:

Estadísticas-Estadísticas básicas-t de 2 muestras

Tabla 5.11 Prueba de hipótesis

Hipótesis nula	$H_0: \mu_1 - \mu_2 = 0$
Hipótesis alterna	$H_1: \mu_1 - \mu_2 \neq 0$

Valor T	GL	Valor p
-1.15	23	0.262

En la tabla 5.11 se puede observar que la significancia calculada p=0.262 no es menor que 0.05. Por lo tanto, no se rechaza H_0.

Ejemplo 2: En una investigación, un grupo de madres recibió una serie de charlas sobre la importancia de la preparación de alimentos ricos en hierro para combatir la anemia en los niños. El objetivo de la investigación es determinar si existen diferencias significativas entre el consumo mensual de hígado vacuno en el grupo de madres que recibió la charla (grupo experimental) y un grupo de madres que no recibió dichas charlas (grupo de control). Los resultados se muestran en la tabla 5.12.

Tabla 5.12 Matriz de datos.

Consumo mensual de hígado vacuno (kg)	
Con charla	Sin charla
3	1.5
2.5	2
2	1
3	2
3.5	1.5
2	1
2	1.5
2	2
3	2.5
3	1.5
1.5	1
2.5	1
1.5	1.5
3	2

Título: charlas sobre alimentación saludable y consumo mensual de hígado vacuno en madres de familia del distrito de Comas.

Pregunta: ¿Las charlas sobre alimentación saludable generan algún cambio en el consumo mensual de hígado vacuno?

Objetivo: conocer si las charlas sobre alimentación saludable generan algún cambio en el consumo mensual de hígado vacuno.

Hipótesis: Las charlas sobre alimentación saludable generan un incremento en el consumo mensual de hígado vacuno.

Tipo de investigación: enfoque cuantitativo, experimental, alcance explicativo y de corte transversal.

Variable x: participación en las charlas sobre la importancia de la preparación de alimentos ricos en hierro para combatir la anemia en los niños.

Variable y: consumo mensual de hígado vacuno.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas en el consumo de hígado vacuno entre ambos grupos en comparación. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias podemos trabajar con el consumo mensual promedio.

Los grupos son diferentes. Por lo tanto, se puede utilizar la prueba t para muestras independientes.

Para usar la prueba t se debe verificar si se cumple el supuesto de normalidad. La muestra es pequeña, por ello, se utiliza la prueba de Shapiro-Wilk en el programa SPSS.

Tabla 5.13 Prueba de normalidad en SPSS.

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Con charla	0.901	14	0.116
Sin charla	0.882	14	0.063

Planteamos las hipótesis para cada uno de los grupos:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

En la tabla 5.13 se observa que la significancia para el consumo de hígado en el grupo de madres que si recibió las charlas y las que no recibieron las charlas es mayor a 0.05. Entonces, no se rechaza H_0; se cumple el supuesto de normalidad.

Luego, verificamos el supuesto de homocedasticidad utilizando SPSS.

Tabla 5.14 Prueba de homogeneidad de varianzas

		F	Significancia
Consumo	Se asumen varianzas iguales	2.343	0.138

Planteamos las hipótesis:

H_0: los datos provienen de poblaciones con varianzas similares.
H_1: los datos provienen de poblaciones con varianzas similares.

En la tabla 5.14 se observa el resultado de la prueba de homocedasticidad en SPSS y se verifica que la significancia 0.138 es mayor a 0.05. Entonces, no se rechaza H_0; se cumple el supuesto de la igualdad de varianzas.

Puesto que se verifican los supuestos de normalidad y homocedasticidad, podemos utilizar la prueba t para muestras independientes. Luego, planteamos las hipótesis para una prueba bilateral:

H_0: el promedio de consumo de hígado vacuno es igual entre el grupo de madres que recibieron las charlas y el grupo de madres que no recibieron la charla.

H_1: el promedio de consumo de hígado vacuno es diferente entre el grupo de madres que recibieron las charlas y el grupo de madres que no recibieron la charla.

Regla de decisión:

Si la significancia calculada es menor que 0.05, se rechaza H_0.

Para realizar la prueba t utilizaremos Excel. La ruta es la siguiente:

Datos-Análisis de datos-prueba t para dos muestras suponiendo varianzas iguales.

Tabla 5.15 Prueba t para dos muestras suponiendo varianzas iguales

	Variable 1	Variable 2
Media	2.46428571	1.57142857
Varianza	0.40247253	0.22527473
Observaciones	14	14
Varianza agrupada	0.31387363	
Diferencia hipotética de las medias	0	
Grados de libertad	26	
Estadístico t	4.21651437	
P(T<=t) una cola	0.00013265	
Valor crítico de t (una cola)	1.70561792	
P(T<=t) dos colas	0.00026529	
Valor crítico de t (dos colas)	2.05552944	

El objetivo de la investigación consiste en determinar si los promedios son iguales o diferentes. Por ello, se utilizará la probabilidad de dos colas.

En la Tabla 5.15 podemos verificar que la significancia para una prueba bilateral (dos colas) es 0.000265 y este valor es menor a 0.05. Entonces, se rechaza la hipótesis nula.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el promedio de consumo de hígado vacuno es diferente entre el grupo de madres que recibieron las charlas y el grupo de madres que no recibieron la charla.

¿Qué más se puede inferir?

La prueba utilizada nos indica que el consumo es diferente entre los grupos analizados. Si queremos ir más allá en el análisis, nos podríamos preguntar ¿Qué grupo presenta el mayor consumo de hígado vacuno?

Si observamos la tabla 5.15, el consumo promedio de hígado vacuno en el grupo de madres que recibió las charlas es de 2.46 kg y el consumo promedio de hígado vacuno en el grupo de madres que no recibió las charlas es de 1.57 kg. Eso quiere decir que la participación en las charlas podría generar un incremento en el consumo de hígado vacuno. Pero esto solo es válido para la muestra, aunque lógicamente se puede inferir que también es válido para la población ya que se ha demostrado con la prueba anterior que los promedios de consumo no son iguales. Si queremos extender esta afirmación, de manera formal para la población, se debe realizar una nueva prueba de hipótesis, pero esta vez una prueba unilateral.

H_0: el promedio de consumo de hígado vacuno en el grupo de madres que recibieron las charlas es igual al promedio de consumo de hígado vacuno en el grupo de madres que no recibieron la charla.

H_1: el promedio de consumo de hígado vacuno en el grupo de madres que recibieron las charlas es mayor al promedio de consumo de hígado vacuno en el grupo de madres que no recibieron la charla.

Para realizar la prueba t también podemos utilizar el programa Minitab. La ruta es la siguiente:

Estadísticas-Estadísticas básicas-t de 2 muestras

Tabla 5.16 Prueba de hipótesis.

Hipótesis nula	$H_0: \mu_1 - \mu_2 = 0$
Hipótesis alterna	$H_1: \mu_1 - \mu_2 > 0$

Valor T	GL	Valor p
4.22	26	0.000

En la tabla 5.16 se puede observar que la significancia calculada p=0.000 es menor que 0.05. Por lo tanto, se rechaza H_0.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el promedio de consumo de hígado vacuno en el grupo de madres que recibieron las charlas es mayor al promedio de consumo de hígado vacuno en el grupo de madres que no recibieron la charla.

También podemos usar los resultados de la tabla 5.15, donde se observa que P(T<=t) de una cola es igual a 0.00013265 y este valor menor a 0.05 permite rechazar H_0.

5.3.1. Si no se cumplen los supuestos ¿Qué podemos hacer?

Pruebas de Normalidad: la prueba t es robusta. Es decir, que así no se cumpla con el supuesto de normalidad, la aplicación de esta prueba nos aporta resultados válidos.

Wackerly et al. (2012) señalan que "en la mayor parte de las aplicaciones con muestras iguales o casi del mismo tamaño, y de manera que el tamaño total de la muestra, $n1 + n2$, sea por lo menos 20, se esperan muy buenos resultados aun cuando las poblaciones no sean normales".

También se pueden eliminar valores extremos o transformar los datos. Por ejemplo, se puede extraer la raíz cuadrada, el logaritmo, etc. De los datos recolectados.

Homogeneidad de varianzas: si no se cumple este supuesto, se pueden transformar los datos o eliminar los valores extremos (atípicos).

Finalmente, se pueden usar pruebas no paramétricas. En este caso, la prueba que corresponde es la prueba U de Mann-Whitney.

5.4. Prueba t para muestras relacionadas

Este tipo de prueba se realiza en investigaciones de tipo longitudinal, es decir en investigaciones donde se mide la variable en un mismo grupo, pero en diferentes momentos.

La prueba t para muestras relacionadas, al ser una prueba paramétrica cuenta con ciertas condiciones para su aplicación. Así, las variables deben estar en el nivel de medición escalar, debe existir la selección aleatoria de los elementos de la muestra y las diferencias emparejadas deben cumplir con el supuesto de normalidad.

Esta prueba se suele utilizar en diseños cuasi experimentales o en diseños experimentales, con pruebas pre test y post test. También se le conoce como prueba t para datos apareados, datos emparejados o medidas repetidas.

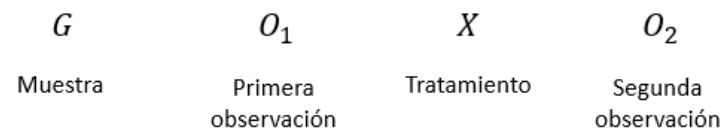


Figura 5.1 Diseño pre test-post test con un solo grupo según Hernández, Fernández y Baptista (2014).

Blair y Taylor (2008) señalan que una prueba *t* de muestras apareadas "no es más que una prueba *t* de una media realizada con las diferencias de puntuaciones". Es decir, en esta prueba no se compran las medias de los grupos, sino que se analiza la media de las diferencias.

El estadístico de prueba está determinado por:

$$t_c = \frac{\overline{D}}{S_{\overline{D}}}$$

Donde $\overline{D}$ es la diferencia media y $\boldsymbol{S_{\overline{D}}}$ es el error estándar de la diferencia media.

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n-1}}$$

S_D es la desviación estándar de las diferencias.

$$S_D = \frac{\sqrt{n\sum D^2 - (\sum D)^2}}{n}$$

D: representa las diferencias entre pares
n: es el tamaño de la muestra

Una vez que se ha calculado el estadístico de prueba t_c, se procede de la siguiente manera para tomar la decisión estadística de rechazar o no rechazar la hipótesis nula.

Regla de decisión

Si $|t_c| \geq t$ entonces, se rechaza H_o
Si $|t_c| < t$ entonces, no se rechaza H_o

Recuerda que en esta distribución de probabilidad utilizaremos $n-1$ grados de libertad. Por ello, ubicamos en la tabla el valor t asociado al nivel de significancia elegido. Recuerda que la muestra se compone de los mismos elementos por ello se considera un solo valor para n en el cálculo de los grados de libertad.

Ejemplo 1: un investigador quiere conocer si las nuevas promociones por la compra de bebidas gaseosas generan un impacto en el consumo de este producto en los habitantes de la urbanización X. Por ello, realiza una encuesta a un grupo de habitantes antes de lanzar las promociones y vuelve a realizar la encuesta al mismo grupo luego de lanzar las promociones. La finalidad de la encuesta es conocer los cambios en el consumo familiar mensual de bebidas gaseosas durante el verano. Los resultados se muestran en la siguiente tabla:

Tabla 5.17 Matriz de datos.

Consumo mensual de bebidas gaseosas (en litros)	
Antes	Después
2	2
2	4
3	4
3	7
3	2
5	6
2	4
5	6
3	6
3	5
2	5
2	3
2	2
3	4
2	3
3	5

Título: promociones de verano y consumo mensual de bebidas gaseosas en los habitantes de la urbanización X.

Pregunta: ¿Cuál es el cambio que las promociones de verano generan en el consumo mensual de bebidas gaseosas en los habitantes de la urbanización X?

Objetivo: conocer el cambio que las promociones de verano generan en el consumo mensual de bebidas gaseosas en los habitantes de la urbanización X.

Hipótesis: las promociones de verano generan un cambio en el consumo mensual de bebidas gaseosas en los habitantes de la urbanización X.

Tipo de investigación: enfoque cuantitativo, cuasi experimental, alcance explicativo y de corte longitudinal.

Variable x: promociones de verano.

Variable y: consumo mensual de bebidas gaseosas.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas en el consumo mensual de bebidas gaseosas antes y después del lanzamiento de las promociones de verano. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias entre el consumo mensual de bebidas gaseosas podemos trabajar con el promedio de las diferencias de consumo.

El grupo en estudio es el mismo, solo ha sido analizado en dos momentos distintos. Por lo tanto, se puede utilizar la prueba t para muestras relacionadas.

Planteamos las hipótesis para una prueba bilateral:

H_0: el promedio mensual de consumo de gaseosas antes del lanzamiento de las promociones de verano es igual al promedio mensual de consumo de gaseosas después del lanzamiento de las promociones de verano.

H_1: el promedio mensual de consumo de gaseosas antes del lanzamiento de las promociones de verano es diferente al promedio mensual de consumo de gaseosas después del lanzamiento de las promociones de verano.

Para verificar el supuesto de normalidad usamos los datos de la columna de diferencias D de la tabla 5.19 y planteamos las hipótesis:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

Debido al tamaño de la muestra usamos la prueba de Shapiro-Wilk en el programa SPSS.

Tabla 5.18 Prueba de normalidad en SPSS.

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Sin charla	0.950	16	0.494

En la tabla 5.18 se observa que la significancia calculada 0.494 es mayor a 0.05, entonces, no se rechaza H_0. Es decir, se cumple el supuesto de normalidad.

Tabla 5.19 Cálculo de las diferencias y los cuadrados respectivos.

x_1	x_2	D	D^2
2	2	0	0
2	4	-2	4
3	4	-1	1
3	7	-4	16
3	2	1	1
5	6	-1	1
2	4	-2	4
5	6	-1	1
3	6	-3	9
3	5	-2	4
2	5	-3	9
2	3	-1	1
2	2	0	0
3	4	-1	1
2	3	-1	1
3	5	-2	4
		$\sum D$=-23	$\sum D^2$=57

Calculamos la diferencia media:

$$\overline{D} = \frac{-23}{16} = -1.438$$

Calculamos la desviación estándar:

$$S_D = \frac{\sqrt{n\sum D^2 - (\sum D)^2}}{n} = \frac{\sqrt{16(57) - (-23)^2}}{16} = 1.223$$

Calculamos el error estándar de la diferencia media:

$$S_{\overline{D}} = \frac{S_D}{\sqrt{n-1}} = \frac{1.223}{\sqrt{16-1}} = 0.316$$

Calculamos el estadístico de prueba:

$$t_c = \frac{\overline{D}}{S_{\overline{D}}} = \frac{-1.438}{0.316} = -4.551$$

Con una significancia de 0.05 y 16-1=15 grados de libertad para una prueba bilateral tenemos:

$$t = 2.13$$

$|-4.551| = 4.551$ es mayor que 2.13 entonces, se rechaza H_o.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el promedio mensual de consumo de gaseosas antes del lanzamiento de las promociones de verano es diferente al promedio mensual de consumo de gaseosas después del lanzamiento de las promociones de verano.

Podemos utilizar Excel para realizar esta prueba t. La ruta es la siguiente:

Datos-Análisis de datos-prueba t para medias de dos muestras emparejadas.

Tabla 5.20 Prueba t para medias de dos muestras emparejadas

	Variable 1	Variable 2
Media	2.8125	4.25
Varianza	0.9625	2.46666667
Observaciones	16	16
Coeficiente de correlación de Pearson	0.59491626	
Diferencia hipotética de las medias	0	
Grados de libertad	15	
Estadístico t	-4.5517047	
P(T<=t) una cola	0.00019095	
Valor crítico de t (una cola)	1.75305036	
P(T<=t) dos colas	0.0003819	
Valor crítico de t (dos colas)	2.13144955	

En la Tabla 5.20 podemos verificar que la significancia para una prueba bilateral (dos colas) es 0.00019 y este valor es menor a 0.05. Entonces, se rechaza la hipótesis nula.

También podemos utilizar SPSS para realizar esta prueba t. La ruta es la siguiente:

Analizar-Compara medias-prueba t para muestras relacionadas.

El programa nos arroja los siguientes resultados:

Tabla 5.21 Prueba de muestras emparejadas

	t	gl	Sig. Bilateral
antes - después	-4,552	15	0.000

Como se puede ver en la tabla 5.21 la significancia bilateral 0.000 es menor a 0.05. Por lo tanto, se rechaza H_0.

Un paso complementario, una vez que se han verificado las diferencias significativas, es utilizar una prueba unilateral para verificar si el consumo de gaseosas aumentó luego de las promociones.

Ejemplo 2: en la empresa de productos alimenticios M se implementa un sistema de supervisión X para las buenas prácticas de manufactura (BPM). Mediante un estudio se desea conocer si este nuevo sistema de supervisión mejora el desempeño en las BPM. Por ello, se utiliza una lista de cotejo para comparar el desempeño antes y después de implementado el sistema de supervisión. El desempeño es evaluado en una muestra de trabajadores con la lista de cotejo que brinda puntajes con un mínimo de 0 puntos y un máximo de 20 puntos. Los resultados fueron los siguientes:

Tabla 5.22 Matriz de datos.

Puntaje para el desempeño en BPM	
Antes	Después
12	14
15	17
13	15
13	15
15	15
15	18
16	18
15	17
12	16
17	17

16	16
17	18
15	17
13	17
14	17
16	18
14	17
15	15
16	15
15	17
15	16
16	16

Título: sistema de supervisión X y buenas prácticas de manufactura en trabajadores de la empresa M.

Pregunta: ¿Qué efecto genera el sistema de supervisión X en las buenas prácticas de manufactura de los trabajadores de la empresa M?

Objetivo: conocer el efecto que el sistema de supervisión X genera en las buenas prácticas de manufactura de los trabajadores de la empresa M.

Hipótesis: el sistema de supervisión X mejora el desempeño de los trabajadores de la empresa M con respecto a las buenas prácticas de manufactura.

Tipo de investigación: enfoque cuantitativo, cuasi experimental, alcance explicativo y de corte longitudinal.

Variable x: tipo de supervisión.

Variable y: desempeño en buenas prácticas de manufactura.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas en el desempeño en buenas prácticas de manufactura antes y después de la implementación del sistema de supervisión. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias podemos trabajar con el promedio de las diferencias de los puntajes.

El grupo en estudio es el mismo, solo ha sido analizado en dos momentos distintos. Por lo tanto, se puede utilizar la prueba t para muestras relacionadas. Hay que resaltar que usaremos una prueba unilateral ya que los antecedentes del estudio señalan que este nuevo sistema genera cambios positivos luego de su implementación en diferentes campos de la industria.

H_0: el puntaje promedio de desempeño en BPM después de la implementación del sistema de supervisión X es igual al puntaje promedio de desempeño en BPM antes de la implementación del sistema de supervisión X.

H_1: el puntaje promedio de desempeño en BPM después de la implementación del sistema de supervisión X es mayor al puntaje promedio de desempeño en BPM antes de la implementación del sistema de supervisión X.

Para realizar la prueba t usaremos el programa Minitab. La ruta es la siguiente:

Estadísticas-Estadísticas básicas-t pareada

Tabla 5.23 Prueba de hipótesis

Hipótesis nula	H_0: diferencia $\mu = 0$
Hipótesis alterna	H_1: diferencia $\mu > 0$

Valor T	Valor p
5.63	0.000

Debemos considerar que el programa trabaja con la diferencia de medias. Para este caso la diferencia se ha realizado de modo que la muestra 1 se refiere al después y la muestra 2 se refiere al antes. Por lo tanto, lo que se quiere comprobar es que la diferencia de medias entre al después y el antes de la implementación del sistema de supervisión X sea mayor que cero. Es decir:

$$\mu_{después} - \mu_{antes} > 0$$

En la tabla 5.23 se puede observar que la significancia calculada p=0.000 es menor que 0.05. Por lo tanto, se rechaza H_0.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el puntaje promedio de desempeño en BPM después de la

implementación del sistema de supervisión X es mayor al puntaje promedio de desempeño en BPM antes de la implementación del sistema de supervisión X.

También podemos usar los resultados de la tabla 5.24, donde se observa que P(T<=t) de una cola es igual a 0.00000695 y este valor menor a 0.05 permite rechazar H_0.

Tabla 5.24 Prueba t para medias de dos muestras emparejadas

	Variable 1	Variable 2
Media	16.4090909	14.7727273
Varianza	1.3961039	2.08874459
Observaciones	22	22
Coeficiente de correlación de Pearson	0.47532182	
Diferencia hipotética de las medias	0	
Grados de libertad	21	
Estadístico t	5.62552323	
P(T<=t) una cola	6.9547E-06	
Valor crítico de t (una cola)	1.7207429	
P(T<=t) dos colas	1.3909E-05	
Valor crítico de t (dos colas)	2.07961384	

5.4.1. Si no se cumplen los supuestos ¿Qué se puede hacer?

Pruebas de Normalidad: La prueba t es robusta. Es decir, que así no se verifique el supuesto de normalidad la aplicación de esta prueba nos puede generar resultados válidos.

Existe la opción de eliminar valores extremos o transformar los datos. Por ejemplo, se les puede extraer la raíz cuadrada, el logaritmo, etc.

Otra opción es usar la prueba no paramétrica de Wilcoxon.

Capítulo 6: Análisis de la varianza

6.1. ANOVA de un solo factor

Este método se utiliza cuando se desea comparar más de dos grupos. Dado que es una prueba paramétrica requiere cumplir con los supuestos de normalidad, homocedasticidad, selección aleatoria de grupos y el nivel de medición escalar (Anderson et al., 2012).

El ANOVA (Analysis of variance) se considera un método muy común que se utiliza cuando se prueban medias poblacionales (Walpole et al., 2007).

Las hipótesis se plantean de la siguiente manera:

H_0: Las medias de los grupos son iguales.
H_1: Al menos dos grupos tienen medias diferentes.

También se puede escribir:

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
H_1: Al menos dos grupos tienen medias diferentes.

El objetivo del ANOVA es determinar el efecto de una variable nominal u ordinal sobre una escalar. También se puede decir, que el ANOVA permite determinar si es razonable o no concluir que no todas las medias del muestreo provienen de la misma población (Hopkins, Hopkins y Glass, 1997).

Figura 6.1 Esquema de una relación causa-efecto.

En la figura 6.1 se muestra un esquema para una investigación experimental. Sin embargo, el ANOVA también se puede usar en investigaciones descriptivas. Por ejemplo, cuando queremos conocer si existen diferencias significativas entre el rendimiento académico de 4 grupos de estudiantes. En ese ejemplo, las calificaciones para el rendimiento académico corresponden a una variable cuantitativa, se tienen más de dos grupos en comparación y verificando los demás supuestos se puede utilizar el ANOVA para hacer una comparación entre el rendimiento académico de los 4 grupos.

Existen dos tipos de prueba ANOVA de un solo factor: el ANOVA para muestras independientes y el ANOVA para muestras relacionadas.

6.2. ANOVA de un solo factor para muestras independientes

Se aplica en situaciones donde existen diferentes grupos a los cuales se les ha aplicado un determinado tratamiento (variable independiente).

Tabla 6.1 Datos de muestras distintas.

Grupos en comparación		
A	B	C
a_1	b_1	c_1
a_2	b_2	c_2
a_3	b_3	c_3
.	.	.
.	.	.
.	.	.
a_n	b_n	c_n

Con la información de la tabla 6.1 se muestra que cada grupo ha recibido un tratamiento diferente y por ello se quiere conocer si esto genera valores muy diferenciados en la variable respuesta. En un estudio descriptivo podemos simplificar la explicación para la prueba de

hipótesis, señalando que los elementos de los grupos A, B y C corresponden a poblaciones distintas y presentan diferencias significativas.

Se pueden presentar dos casos:

- Cuando hay K muestras de tamaño n.
- Cuando hay k muestras de tamaño n_1, n_2, n_3, …

En cada caso el procedimiento estadístico es diferente. Solo veremos el primer caso.

6.2.1. Prueba F

El ANOVA utiliza la prueba F, de modo que se recurre a la distribución F, desarrollada por el estadístico R. Fisher.

Usaremos el estadístico de prueba:

$$F_c = \frac{CM_{inter}}{CM_{intra}}$$

Donde CM_{inter} es el cuadrado medio intergrupos y CM_{intra} es el cuadrado medio intragrupos.

$$CM_{inter} = \frac{SC_{inter}}{gl_{inter}}$$

$$CM_{intra} = \frac{SC_{intra}}{gl_{intra}}$$

Donde SC_{inter} es la suma de cuadrados intergrupos y SC_{intra} es la suma de cuadrados intragrupos.

gl_{inter}: número de grupos menos 1, $k - 1$

gl_{intra}: suma de casos en cada grupo menos 1, $(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + \cdots + (n_k - 1)$

gl_T: número total de casos menos 1, $N - 1$

Suma de cuadrados total:

$$SC_T = \sum x^2 - \frac{(\sum x)^2}{N}$$

Suma de cuadrados intergrupal:

$$SC_{inter} = \sum \frac{(\sum x_i)^2}{n_i} - \frac{(\sum x)^2}{N}$$

Suma de cuadrados intragrupal:

$$SC_{intra} = SC_T - SC_{inter}$$

Regla de decisión para la prueba ANOVA

Una vez que se ha calculado el estadístico de prueba F_c, se procede de la siguiente manera para tomar la decisión estadística de rechazar o no rechazar la hipótesis nula.

Para identificar el F crítico en la tabla de distribución de probabilidad, utilizaremos los grados de libertad. Por ello, en la tabla ubicamos el valor F asociado al nivel de significancia elegido.

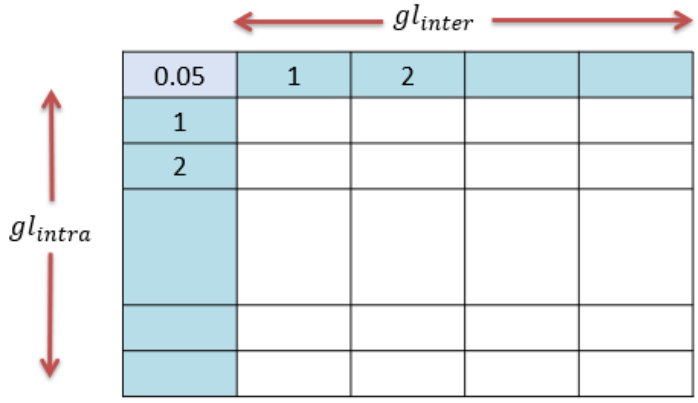


Figura 6.2 Forma de identificar el valor F crítico de la tabla.

Si $F_c \geq F$ entonces, se rechaza H_o
Si $F_c < F$ entonces, no se rechaza H_o

6.2.2. Comparaciones post hoc a la prueba F

Una vez que se ha verificado que existe al menos dos grupos con medias diferentes por medio de la prueba F. Es importante identificar las diferencias entre pares de medias.

En otras palabras, podemos identificar entre todos los grupos a quienes presentan medias significativamente diferentes.

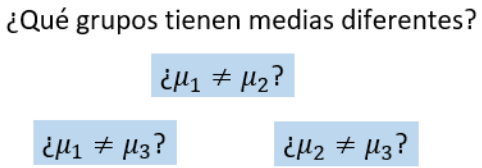


Figura 6.3 Posibles casos a verificar en una diferencia de medias.

Pruebas post hoc

Solo cuando se han detectado diferencias significativas entre los grupos se procede a comparar los grupos por pares. Existen diferentes pruebas para hacer estas comparaciones.

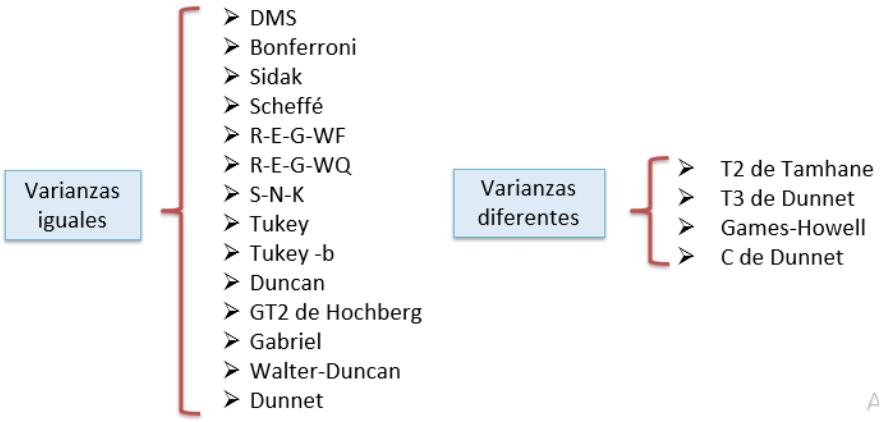


Figura 6.4 Pruebas posteriores según la igualdad de varianzas.

Ejemplo 1: mediante un estudio se quiere conocer si existen diferencias entre las masas corporales de las gallinas de 3 gallineros. Por ello, se toma una muestra representativa de 10 gallinas de cada gallinero y se obtienen los siguientes resultados:

Tabla 6.2 Matriz de datos.

Masa corporal en Kg		
A	B	C
1.80	1.78	1.80
1.85	1.80	1.72
1.86	1.74	1.76
1.84	1.77	1.76
1.86	1.77	1.76
1.85	1.74	1.77
1.86	1.80	1.77
1.92	1.77	1.77
1.85	1.77	1.73
1.88	1.77	1.79

Título: masa corporal de las gallinas en la empresa avícola X.

Pregunta: ¿Existen diferencias entre las masas corporales de las gallinas de los gallineros A, B y C?

Objetivo: conocer si existen diferencias entre las masas corporales de las gallinas de los gallineros A, B y C.

Hipótesis: las masas corporales de las gallinas de los gallineros A, B y C son diferentes.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable 1: gallinero.

Variable 2: masa corporal.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas en la masa corporal de las gallinas de los gallineros A, B y C. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias podemos trabajar con el promedio de las masas corporales de las gallinas.

Los grupos a comparar son grupos distintos y la masa corporal es una variable cuantitativa. Por lo tanto, se puede utilizar el ANOVA para muestras independientes.

H_0: el promedio de las masas corporales de las gallinas de los gallineros A, B y C son iguales.

H_1: el promedio de las masas corporales de las gallinas de los gallineros A, B y C son diferentes.

.

Para realizar la prueba ANOVA, en este ejemplo, usaremos el desarrollo completo y luego verificaremos los resultados con los programas Excel, SPSS y Minitab.

Para usar el ANOVA se debe verificar si se cumple el supuesto de normalidad para cada uno de los grupos de catos. Las muestras son pequeñas, por ello, se utiliza la prueba de Shapiro-Wilk en el programa SPSS.

Analizar-Estadísticos descriptivos-Explorar-Gráficos-Gráficos de normalidad

Tabla 6.3 Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Gallinero A	0.896	10	0.199
Gallinero B	0.859	10	0.075
Gallinero C	0.925	10	0.397

Planteamos las hipótesis para cada uno de los grupos:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

Para verificar el supuesto de normalidad utilizamos la significancia calculada. Dado que el nivel de confianza es de 95% la significancia calculada es 0.05. En la tabla 6.3 se observa que las significancias calculadas para los diferentes gallineros A, B y C son 0.199, 0.075 y 0.397 respectivamente. Todas las significancias calculadas son mayores que 0.05. Entonces, no se rechaza H_0; se verifica el supuesto de normalidad para cada grupo.

También se debe verificar si se cumple el supuesto de homocedasticidad. Por ello, usaremos el estadístico de Levene que se obtiene con el programa SPSS.

Planteamos las hipótesis para verificar la homogeneidad de varianzas:

H_0: Los tres conjuntos de datos provienen de poblaciones con varianzas iguales.
H_1: Los tres conjuntos de datos no provienen de poblaciones con varianzas iguales.

Analizar-Comparar medias-ANOVA de un factor-Opciones-Prueba de homogeneidad de varianzas.

Tabla 6.4 Prueba de homogeneidad de varianzas

		Estadístico de Levene	gl1	gl2	Sig.
Masa	Se basa en la media	0.247	2	27	0.783
	Se basa en la mediana	0.280	2	27	0.758
	Se basa en la mediana y con gl ajustado	0.280	2	23.46	0.759
	Se basa en la media recortada	0.243	2	27	0.786

Para este ejemplo la significancia es 0.05 puesto que el nivel de confianza es de 95%. En la tabla 6.4 se observa que la significancia basada en la media es 0.738 y es mayor que 0.05. Entonces no se rechaza H_0. Es decir, se verifica el supuesto de homocedasticidad.

En el siguiente apartado veremos el proceso completo para calcular el estadístico de prueba F_c.

Calculamos los cuadrados de los datos recolectados y sus respectivas sumatorias:

Tabla 6.5 Datos y sus respectivos cuadrados.

x_1	x_2	x_3	x_1^2	x_2^2	x_3^2
1.8	1.78	1.8	3.24	3.1684	3.24
1.85	1.8	1.72	3.4225	3.24	2.9584
1.86	1.74	1.76	3.4596	3.0276	3.0976
1.84	1.77	1.76	3.3856	3.1329	3.0976
1.86	1.77	1.76	3.4596	3.1329	3.0976
1.85	1.74	1.77	3.4225	3.0276	3.1329
1.86	1.8	1.77	3.4596	3.24	3.1329
1.92	1.77	1.77	3.6864	3.1329	3.1329
1.85	1.77	1.73	3.4225	3.1329	2.9929
1.88	1.77	1.79	3.5344	3.1329	3.2041
$\sum x_1$ =18.57	$\sum x_1$ =17.71	$\sum x_1$ =17.63	$\sum x_1^2$ =34.4927	$\sum x_2^2$ =31.3681	$\sum x_3^2$ =31.0869
$\sum x$ =53.91			$\sum x^2$ =96.9477		
$n_1 = n_2 = n_3 = 10$			$N = 30$		

Calculamos la suma de cuadrados total:

$$SC_T = \sum x^2 - \frac{(\sum x)^2}{N} = 96.9477 - \frac{53.91^2}{30} = 0.07143$$

Calculamos la suma de cuadrados intergrupal:

$$SC_{inter} = \sum \frac{(\sum x_i)^2}{n_i} - \frac{(\sum x)^2}{N}$$

$$SC_{inter} = \frac{18.57^2}{10} + \frac{17.71^2}{10} + \frac{17.63^2}{10} - \frac{53.91^2}{30} = 0.05432$$

Calculamos la suma de cuadrados intragrupal:

$$SC_{intra} = SC_T - SC_{inter} = 0.07143 - 0.05432 = 0.01711$$

Definimos los grados de libertad:

$$gl_{inter} = k - 1 = 3 - 1 = 2$$

$$gl_{intra} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) = 9 + 9 + 9 = 27$$

$$gl_T = N - 1 = 30 - 1 = 29$$

Calculamos los cuadrados medios:

$$CM_{inter} = \frac{SC_{inter}}{gl_{inter}} = \frac{0.05432}{2} = 0.02716$$

$$CM_{intra} = \frac{SC_{intra}}{gl_{intra}} = \frac{0.01711}{27} = 0.000634$$

Ahora calculamos el estadístico de prueba:

$$F_c = \frac{CM_{inter}}{CM_{intra}} = \frac{0.02716}{0.000634} = 42.84$$

Considerando un nivel de confianza de 95%, $gl_{inter} = 2$ y $gl_{intra} = 27$. El valor f en la tabla de distribución F es:

$$F = 3.35$$

También podemos usar Excel para determinar el valor de F crítico:

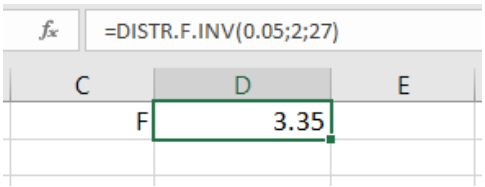


Figura 6.5 Cálculo del valor F bilateral.

42.84 es mayor que 3.35. Entonces, se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que el promedio de las masas corporales de las gallinas de los gallineros A, B y C son diferentes.

Pruebas posteriores

Luego de establecer que los promedios de las masas corporales de las gallinas en los gallineros A, B y C son significativamente diferentes, nos

podemos preguntar ¿En qué gallinero se encuentran las gallinas con mayor peso? O ¿En qué gallinero se encuentran las gallinas con mayor peso?

Para responder a las preguntas planteadas, podemos formular afirmaciones como las siguientes.

- La masa corporal de las gallinas del gallinero A es mayor a la masa corporal de las gallinas del gallinero B.

- La masa corporal de las gallinas del gallinero A es mayor a la masa corporal de las gallinas del gallinero C.

- La masa corporal de las gallinas del gallinero B es mayor a la masa corporal de las gallinas del gallinero C.

Existen pruebas de hipótesis para comparar las medias obtenidas, esta comparación se realiza por pares de grupos. Sin embargo, en nuestro ejemplo nos podemos orientar observando las medias de cada grupo.

Tabla 6.6 Masa promedio por gallinero.

Gallinero	Promedio (kg)
A	1.857
B	1.771
C	1.763

En la tabla 6.6 se observa que la mayor masa promedio está en el gallinero A y la menor masa promedio está en el gallinero C. Sin embargo, entre el promedio de masas de los gallineros B y C no hay mucha diferencia. En cambio, la masa promedio del gallinero A si se diferencia mucho de las masas promedio de los demás gallineros. Más adelante se hará uso de una prueba post hoc que permite establecer exactamente qué grupo o grupos presentan las diferencias más significativas.

Si utilizamos el programa Excel para realizar el ANOVA. La ruta es la siguiente:

Datos-Análisis de datos-Análisis de varianza de un factor.

Los resultados se muestran a continuación:

Tabla 6.7 Resumen descriptivo.

Grupos	Cuenta	Suma	Promedio	Varianza
Columna 1	10	18.57	1.857	0.00091222
Columna 2	10	17.71	1.771	0.00041
Columna 3	10	17.63	1.763	0.00057889

En la tabla 6.7 se muestra la parte descriptiva de los resultados como la media y varianza de cada grupo.

Tabla 6.8 Análisis de varianza.

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Prob.*	*Valor crítico para F*
Entre grupos	0.05432	2	0.02716	42.8591467	4.18E-09	3.35413083
Dentro de los grupos	0.01711	27	0.0006337			
Total	0.07143	29				

En la tabla 6.8 se muestran los valores necesarios para la prueba de hipótesis. Utilizando el método de la significancia, para un nivel de confianza de 95% la significancia es 0.05. la significancia calculada o probabilidad es 0.00000000418 y es menor que 0.05. Entonces se rechaza H_0.

Usando el programa SPSS podemos realizar el ANOVA. La ruta es la siguiente:

Analizar-Comparar medias-ANOVA de un factor.

Los resultados se muestran a continuación:

Tabla 6.9 ANOVA en SPSS.

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	0.054	2	0.027	42.859	0.000
Dentro de grupos	0.017	27	0.001		
Total	0.071	29			

En la tabla 6.9 se verifica que el valor de la significancia calculada es 0.000. Esto permite rechazar H_0.

El programa SPSS también nos permite utilizar las pruebas posteriores. En este caso usaremos la prueba de Tukey.

Tabla 6.10 Comparaciones múltiples HSD Tukey

(I) Gallinero	(J) Gallinero	Diferencia de medias (I-J)	Desv. Error	Sig.
A	B	0.08600*	0.01126	0.000
	C	0.09400*	0.01126	0.000
B	A	-0.08600*	0.01126	0.000
	C	0.00800	0.01126	0.759
C	A	-0.09400*	0.01126	0.000
	B	-0.00800	0.01126	0.759

*. La diferencia de medias es significativa en el nivel 0.05.

Para interpretar la tabla 6.10 se debe hacer uso de la significancia. Si la significancia calculada es menor que 0.05 se interpreta que existen diferencias significativas entre las medias de la pareja de grupos comparados. Entonces, como la significancia calculada 0.759 no es menor que 0.05 podemos afirmar que las medias de los gallineros B y C no presentan una diferencia significativa. La siguiente tabla nos ayudará a entender mejor la interpretación de prueba de Tukey.

Tabla 6.11 HSD Tukey

		Subconjunto para alfa = 0.05	
Gallinero	N	1	2
C	10	1.7630	
B	10	1.7710	
A	10		1.8570
Sig.		1.759	1.000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.

En la tabla 6.11 se observa que los promedios de las masas corporales en los gallineros B y C no presentan diferencias significativas y estos promedios son menores con respecto al promedio de masas corporales del gallinero A. También se puede decir que A presenta el mayor

promedio de masa corporales y su diferencia con los promedios de los demás grupos es significativa.

También podemos utilizar el programa Minitab para realizar el ANOVA. La ruta es la siguiente:

Estadísticas-ANOVA-un solo factor.

Los resultados son los siguientes:

Tabla 6.12 Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Factor	2	0.05432	0.027160	42.86	0.000
Error	27	0.01711	0.000634		
Total	29	0.07143			

En la tabla 6.12 se observa el valor p=0.000 y es menor que 0.05. entonces se rechaza H_0. En este ejemplo, el término factor hace referencia a los gallineros A, B y C. Rechazar H_0 implica que el factor gallinero tiene un efecto significativo sobre las masas corporales de las gallinas. Es decir, que existen diferencias significativas entre los promedios de las masas corporales.

Debemos resaltar que el término factor es más adecuado para investigaciones experimentales donde se trata de determinar la relación de causa-efecto.

Tabal 6.13 Medias por grupo.

Factor	N	Media	Desv.Est.	IC de 95%
A	10	1.85700	0.03020	(1.84067; 1.87333)
B	10	1.77100	0.02025	(1.75467; 1.78733)
C	10	1.76300	0.02406	(1.74667; 1.77933)

Desv.Est. agrupada = 0.0251735

En la tabla 6.13 se muestran las medias muestrales. Esta información nos ayuda a determinar la relación de orden entre los promedios muestrales. Es decir, nos permite diferenciar la menor y mayor media.

Tabla 6.14 Comparaciones en parejas de Tukey.
Agrupar información utilizando el método de Tukey y una confianza de 95%.

Factor	N	Media	Agrupación	
A	10	1.85700	A	
B	10	1.77100		B
C	10	1.76300		B

Las medias que no comparten una letra son significativamente diferentes.

La nota de la tabla 6.14 permite identificar que el gallinero A posee la mayor masa corporal promedio y este promedio se diferencia significativamente con los promedios de los gallineros B y C. En cambio, los promedios de las masas corporales de los gallineros B y C no presentan diferencias significativas entre sí.

Ejemplo 2: En una investigación se desea conocer si existen diferencias entre la densidad del aceite producido por tres máquinas de tipo X. Por ello, se toma una muestra de 12 botellas de aceite producidas por cada máquina. Los resultados son los siguientes.

Tabla 6.15 Matriz de datos.

Densidad del aceite en Kg		
Máquina A	Máquina B	Máquina C
0.91	0.92	0.89
0.93	0.92	0.89
0.91	0.92	0.90
0.90	0.92	0.90
0.93	0.95	0.90
0.93	0.95	0.92
0.93	0.88	0.92
0.93	0.88	0.92
0.96	0.86	0.92
0.96	0.92	0.94
0.97	0.95	0.94
0.93	0.96	0.95

Título: densidad del aceite producido por las máquinas de Tipo X.

Pregunta: ¿Existen diferencias entre la densidad del aceite producido por las máquinas A, B y C?

Objetivo: conocer si existen diferencias entre la densidad del aceite producido por las máquinas A, B y C.

Hipótesis: la densidad del aceite producido es diferente en las máquinas A, B y C.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo y de corte transversal.

Variable 1: máquina.

Variable 2: densidad del aceite.

Realiza la prueba de hipótesis para determinar si existen diferencias significativas en la densidad del aceite producido por las máquinas A, B y C. Utilice un nivel de confianza de 95%.

Solución: Para establecer si existen diferencias podemos trabajar con el promedio de las densidades del aceite producido por cada máquina.

Los grupos a comparar son grupos distintos y la densidad es una variable cuantitativa. Por lo tanto, se puede utilizar el ANOVA para muestras independientes.

Planteamos las hipótesis para una prueba bilateral:

H_0: los promedios de las densidades de aceite producidos por las máquinas A, B y C son iguales.

H_1: los promedios de las densidades de aceite producidos por las máquinas A, B y C son diferentes.

.

Para realizar la prueba ANOVA, en este ejemplo, usaremos Excel y el programa SPSS.

Para usar el ANOVA se debe verificar si se cumple el supuesto de normalidad para cada uno de los grupos de datos. Las muestras son pequeñas, por ello, se utiliza la prueba de Shapiro-Wilk en el programa SPSS.

Analizar-Estadísticos descriptivos-Explorar-Gráficos-Gráficos de normalidad

Tabla 6.16 Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Máquina A	0.885	12	0.101
Máquina B	0.887	12	0.109
Máquina C	0.909	12	0.204

Planteamos las hipótesis para cada uno de los grupos:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

En la tabla 6.16 se observa que todas las significancias calculadas son mayores que 0.05. Entonces, se no se rechaza H_0; se verifica el supuesto de normalidad para cada grupo.

También se debe verificar si se cumple el supuesto de homocedasticidad. Por ello, usaremos el estadístico de Levene que se obtiene con el programa SPSS.

Planteamos las hipótesis para verificar la homogeneidad de varianzas:

H_0: Los tres conjuntos de datos provienen de poblaciones con varianzas iguales.
H_1: Los tres conjuntos de datos no provienen de poblaciones con varianzas iguales.

Analizar-Comparar medias-ANOVA de un factor-Opciones-Prueba de homogeneidad de varianzas.

Tabla 6.17 Prueba de homogeneidad de varianzas

	Estadístico de Levene	gl1	gl2	Sig.
Se basa en la media	0.794	2	33	0.461
Se basa en la mediana	0.818	2	33	0.450
Se basa en la mediana y con gl ajustado	0.818	2	27,538	0.452
Se basa en la media recortada	0.709	2	33	0.499

Para este ejemplo la significancia es 0.05 puesto que el nivel de confianza es de 95%. En la tabla 6.17 se observa que la significancia basada en la media es 0.461 y es mayor que 0.05. Entonces no se rechaza H_0. Es decir, se verifica el supuesto de homocedasticidad.

Ahora usaremos Excel para realizar el ANOVA. La ruta es la siguiente:

Datos-Análisis de datos-Análisis de la varianza de un factor

Los resultados son los siguientes:

Tabla 6.18 Resumen descriptivo.

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Máquina A	12	11.19	0.9325	0.00045682
Máquina B	12	11.03	0.91916667	0.00100833
Máquina C	12	10.99	0.91583333	0.00040833

En la tabla 6.18 se muestran los promedios de las densidades de cada grupo. Estos promedios son muy parecidos, por ello, es muy probable que no se verifiquen diferencias significativas en la prueba de hipótesis.

Tabla 6.19 Análisis de la varianza

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Prob.*	*Valor crítico para F*
Entre grupos	0.00186667	2	0.00093333	1.494541	0.2391	3.28491765
Dentro de los grupos	0.02060833	33	0.00062449			
Total	0.022475	35				

En la tabla 6.19 se observa que la probabilidad 0.2391 no es menor que 0.05. Entonces, no se rechaza H_0.

En conclusión, no existe evidencia estadística para afirmar que los promedios de las densidades de aceite producidos por las máquinas A, B y C sean diferentes.

Finalmente, podemos agregar que no es necesario elaborar las pruebas posteriores. Debido a que no se ha verificado que existan diferencias

significativas entre los promedios de densidades de aceite producidas por las máquinas A, B y C.

6.2.3. ¿Qué sucede cuando no se cumplen los supuestos?

El ANOVA es una prueba robusta y los resultados que nos brinda cuando no se cumplen los supuestos aún pueden ser válidos. También está la opción transformar los datos o eliminar valores atípicos como se sugirió en la prueba t.

Finalmente, se puede optar por el uso de la prueba no paramétrica de Kruskal-Wallis.

Capítulo 7: Coeficiente de correlación lineal de Pearson

La correlación se entiende como el grado de relación entre dos variables o más de dos variables. En este apartado solo analizaremos la correlación entre dos variables.

Generalmente se utiliza el término correlación cuando analizamos la relación entre dos variables cuantitativas o en la escala de medición de intervalo o de proporción. También se puede afirmar que, para utilizar esta prueba de correlación lineal, el nivel de medición de las variables debe ser, por lo menos, de escala ordinal.

En una investigación relacional no se analiza la relación causa-efecto. Por ello no es adecuado definir a las variables como independiente o dependiente, sino que, dichas variables deben definirse como variable 1 y variable 2.

La relación entre dos variables puede ser directa, inversa o nula. Consideramos una relación directa entre dos variables x_1 y x_2 cuando se observa que el incremento de los valores de la variable x_1 se asocia con el incremento de los valores de la variable x_2 y viceversa.

Consideramos una relación inversa entre dos variables x_1 y x_2 cuando se observa que el incremento de los valores de la variable x_1 se asocia con la reducción en los valores de la variable x_2 y viceversa.

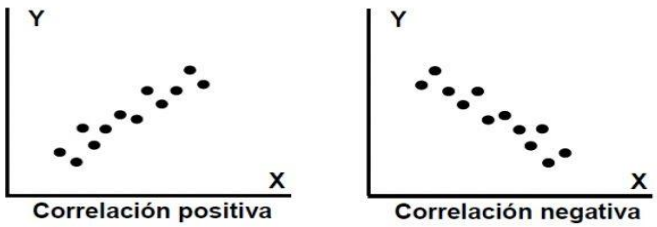


Figura 7.1 Tipos de correlación.

Cuando no se observa una relación directa ni una relación inversa entre las variables, se puede decir que no existe relación entre las variables, o en todo caso, se dice que la relación entre las variables es nula.

7.1. Diagrama de dispersión

Es un gráfico que permite tener un conocimiento inicial acerca de la relación entre dos variables.

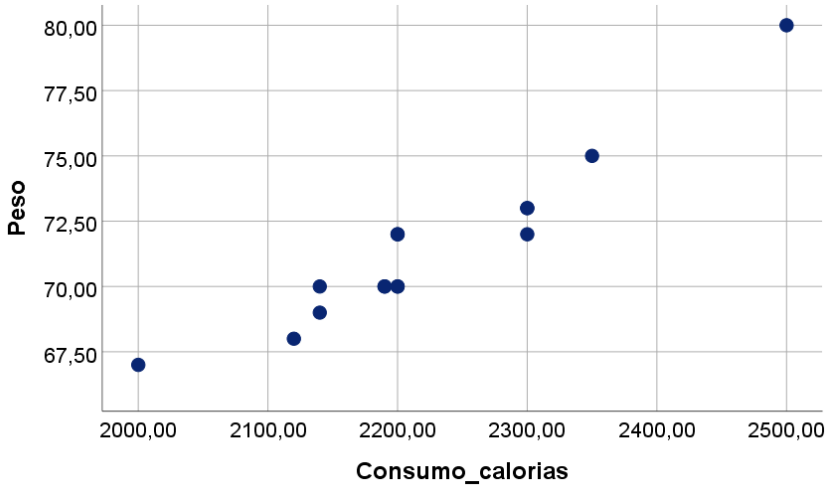


Figura 7.2 Diagrama de dispersión Peso-Consumo de calorías.

Este tipo de gráfico es útil para establecer visualmente el tipo de correlación entre las variables. Cada uno de los puntos representa dos observaciones que se han realizado en una unidad de análisis. Por ejemplo, en la figura 7.2 el punto más alto del lado derecho del diagrama de dispersión representa a un individuo que consume 2500 calorías y tiene un peso (masa corporal) de 80 kg.

En la figura 7.2 también se puede resaltar la forma en la que están ubicados los puntos. Esta ubicación nos permite identificar si la relación entre las variables es lineal o no lineal. Es decir, cuando el conjunto o nube de puntos se ordena de tal forma que se puede asociar a una línea recta, se dice que la tendencia es lineal. Cuando el conjunto o nube de puntos se ordena de tal forma que se puede asociar a una curva, se dice que la tendencia es no lineal. Finalmente, cuando no se observa una tendencia lineal o no lineal, se puede concluir que no existe relación entre las variables.

A continuación, se muestran algunos diagramas de dispersión con sus respectivas interpretaciones. Recordemos que el diagrama de dispersión nos muestra de manera inicial si existe o no una relación lineal entre las variables. Además, cuando los puntos están cerca a la recta de referencia se dice que existe buen ajuste, en caso contrario se dice que existe mal ajuste.

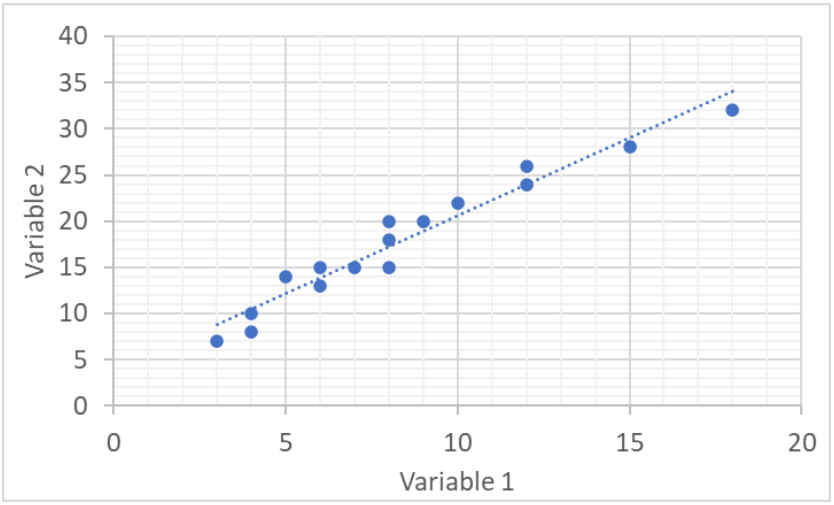


Figura 7.3 Diagrama de dispersión Variable 1-Variable 2.

En la figura 7.3 se observa una tendencia lineal entre las variables y se observa un buen ajuste. Además, se puede ver que la relación entre las variables es directa.

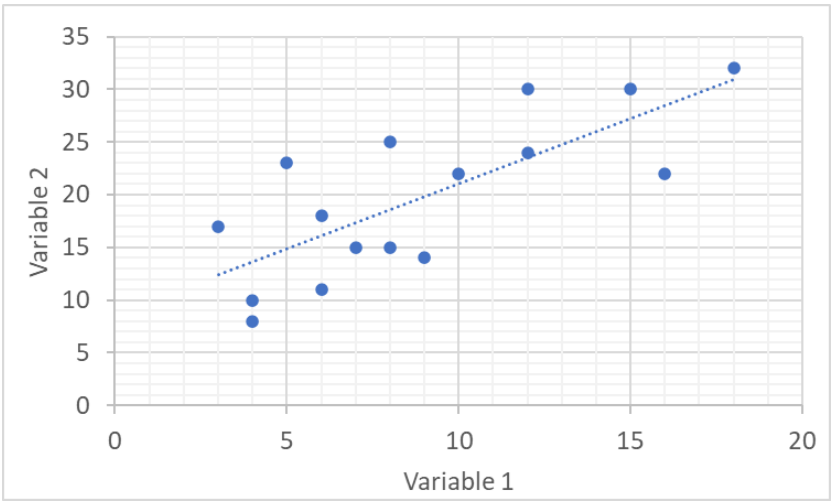


Figura 7.4 Diagrama de dispersión Variable 1-Variable 2.

En la figura 7.4 se observa una tendencia lineal entre las variables, pero no se observa un buen ajuste. Además, se puede ver que la relación entre las variables es directa.

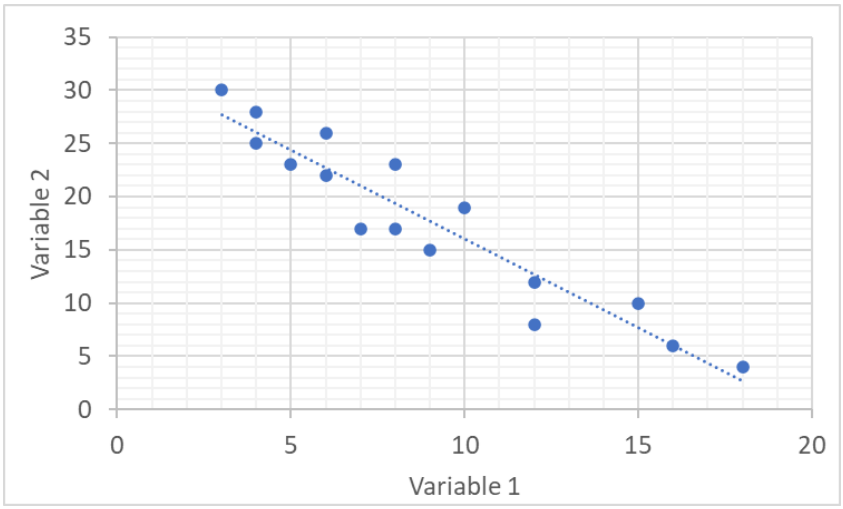


Figura 7.5 Diagrama de dispersión Variable 1-Variable 2.

En la figura 7.5 se observa una tendencia lineal entre las variables y se observa un buen ajuste. Además, se puede ver que la relación entre las variables es inversa.

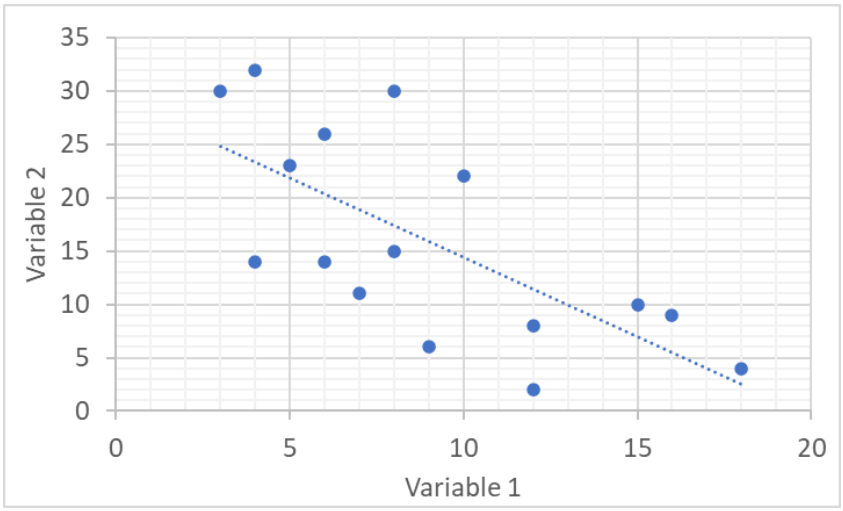


Figura 7.6 Diagrama de dispersión Variable 1-Variable 2.

En la figura 7.6 se observa una tendencia lineal entre las variables, pero no se observa un buen ajuste. Además, se puede ver que la relación entre las variables es inversa.

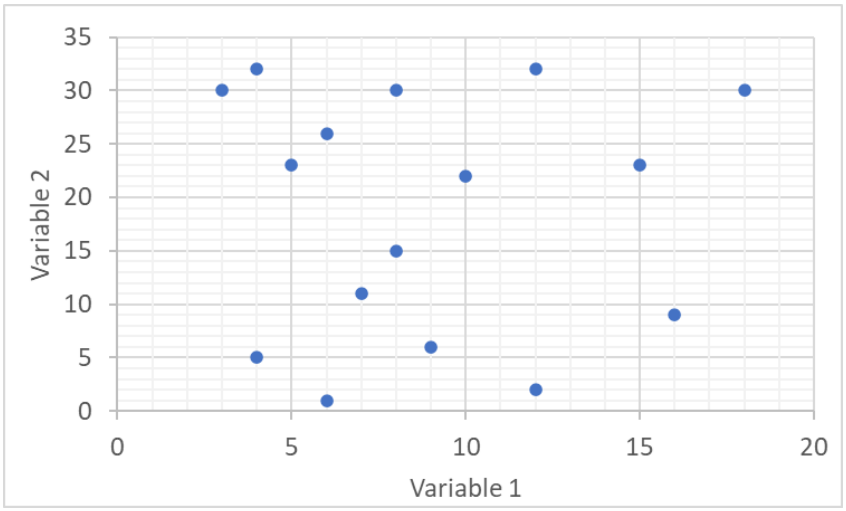


Figura 7.7 Diagrama de dispersión Variable 1-Variable 2.

En la figura 7.7 no se observa una tendencia lineal entre las variables. Existen casos en los cuales existe una tendencia no lineal. Pero en el caso mostrado no se observa tendencia alguna.

7.2. Coeficiente de correlación lineal de Pearson (r)

Este coeficiente es un valor que nos indica el tipo de relación y el grado de relación entre las variables. Maya y Pliego (1999) utilizan la siguiente expresión para el cálculo del coeficiente de correlación lineal:

$$r = \frac{\sum(x - \bar{X})(y - \bar{Y})}{\sqrt{\sum(x - \bar{X})^2 \sum(y - \bar{Y})^2}}$$

Para evitar el uso de los promedios, r también se calcula mediante la siguiente expresión:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Donde:

n: representa la cantidad de parejas de datos.

$\sum x$: es la suma de los valores de x.

$\sum y$: es la suma de los valores de y.

$\sum xy$: es la suma de los productos de los valores de x e y.

$\sum x^2$: es la suma de los cuadrados de los valores de x.

$\sum y^2$: es la suma de los cuadrados de los valores de y.

El coeficiente de correlación lineal no puede tomar cualquier valor, sino que está acotado en el siguiente intervalo:

$$-1 \leq r \leq 1$$

7.2.1. Interpretación del coeficiente de correlación lineal

Existen muchas formas de interpretar el valor del coeficiente de correlación. Por ejemplo, Wathen, Lind y Marchal (2012) presentan la siguiente forma de interpretación:

Tabla 7.1 interpretación del coeficiente r.

Coeficiente	Interpretación
$r = -1$	Existe una relación negativa perfecta.
$-1 < r < -0{,}5$	Existe relación negativa fuerte.
$r = -0{,}5$	Existe una relación negativa moderada.
$-0{,}5 < r < 0$	Existe relación negativa débil.
$r = 0$	No existe relación lineal.
$0 < r < 0{,}5$	Existe relación positiva débil.
$r = 0{,}5$	Existe una relación positiva moderada.
$0{,}5 < r < 1$	Existe relación positiva fuerte.
$r = 1$	Existe una relación positiva perfecta.

Pero no es una imposición el hacer una interpretación basándonos estrictamente en la tabla 7.1. En situaciones particulares, es necesario usar el sentido lógico. Si tenemos r=0.52, este valor está más cercano a 0.5 que a 1, por ello no podemos decir que la relación es positiva fuerte, sino que la relación es positiva moderada.

En general, para establecer una forma de interpretación práctica del coeficiente de correlación podemos usar el siguiente criterio:

Cuando r está cerca de -1 o de 1, la relación es fuerte.
Cuando r está cerca de -0.5 o de 0.5, la relación es moderada.
Cuando r está cerca de 0, la relación es débil.

Una correlación positiva también se denomina correlación directa, una correlación negativa también se denomina correlación inversa y una correlación media también se denomina correlación moderada.

Por ejemplo:

Si $r = 0.78$, la correlación es directa y fuerte.

Si $r = -0.18$, la correlación es inversa pero débil.

Si $r = 0.91$, la correlación es directa y muy fuerte.

Si $r = -0.48$, la correlación es inversa y moderada.

Si $r = 0.084$, la correlación es directa pero muy débil.

Si $r = -0.99$, la correlación es inversa y casi perfecta.

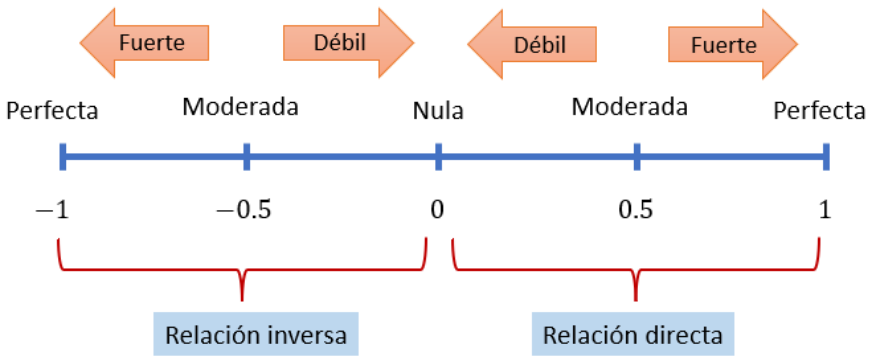


Figura 7.8 Interpretación del coeficiente de correlación.

7.2.2. Supuestos o condiciones

Las variables deben pertenecer a un nivel de medición de intervalo o de razón.

Los datos deben poseer una relación lineal y esto se puede determinar a través de una gráfica de dispersión.

Los datos provienen de una población con distribución normal (distribución normal bivariada), esto es muy importante cuando realicemos pruebas de hipótesis. Sin embargo, cuando cada grupo de datos tiene distribución normal de manera independiente, también se puede utilizar el coeficiente de correlación lineal de Pearson (Blair y Taylor, 2008).

Las observaciones utilizadas (datos) para el análisis deberían recolectarse de forma aleatoria de la población de referencia. Cuando esto no ocurre, el coeficiente de correlación podría no ser el adecuado para la población de interés.

7.2.3. Prueba de hipótesis para el coeficiente de correlación de Pearson

Cuando ya se ha determinado el valor del coeficiente de correlación, se realiza la prueba de hipótesis para determinar si este coeficiente de correlación calculado es significativo.

Planteamos las hipótesis:

H_0: No existe correlación significativa entre las variables.

H_1: Existe correlación significativa entre las variables.

En esta prueba de hipótesis se puede usar la distribución t. Por ello, el estadístico de prueba es el siguiente (Hopkins et al., 1997):

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Donde:

n: es la cantidad de parejas de datos.

r: es el coeficiente de correlación calculado.

Con una significancia α, y $n-2$ grados de libertad para una prueba bilateral se utiliza la tabla de valores críticos de la distribución t (ver el apéndice) y se obtiene el valor critico t asociado.

Regla de decisión

Dado que estamos usando la distribución t, entonces la regla de decisión es la misma que utilizamos anteriormente:

Si $|t_c| \geq t$ entonces, se rechaza H_o
Si $|t_c| < t$ entonces, no se rechaza H_o

Ejemplo 1: mediante una investigación se desea conocer la relación entre el conocimiento sobre nutrición en las madres de familia y la cantidad de panes que consumen por semana.

La primera variable se evalúa a través de un cuestionario de conocimientos compuesto de 15 preguntas donde a cada respuesta correcta se le asigna un punto y a cada respuesta incorrecta se le asigna 0 puntos. La segunda variable se mide a través de una pregunta directa sobre el consumo de pan. Los resultados se muestran en la siguiente tabla:

Tabla 7.2 Matriz de datos.

Individuo	Variable 1 Conocimiento	Variable 2 Consumo
1	12	8
2	12	7
3	10	7
4	13	7
5	8	10
6	8	10
7	13	10
8	13	5
9	10	7
10	10	8
11	10	10
12	12	10

13	7	14
14	8	14
15	10	14
16	10	14

Título: relación entre el conocimiento sobre nutrición y el consumo de pan en madres de familia del distrito de San Luis.

Pregunta: ¿Existe relación entre el conocimiento sobre nutrición y el consumo de pan en madres de familia del distrito de San Luis?

Objetivo: conocer la relación entre el conocimiento sobre nutrición y el consumo de pan en madres de familia del distrito de San Luis.

Hipótesis: Existe una relación inversa entre el conocimiento sobre nutrición y el consumo de pan en las madres de familia del distrito de San Luis.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo correlacional y de corte transversal.

Variable 1: conocimientos sobre nutrición.

Variable 2: consumo semanal de pan.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar si existe correlación significativa entre las variables en estudio.

Solución: Para tener un acercamiento al comportamiento conjunto de las variables se elabora el diagrama de dispersión para verificar si existe una tendencia lineal.

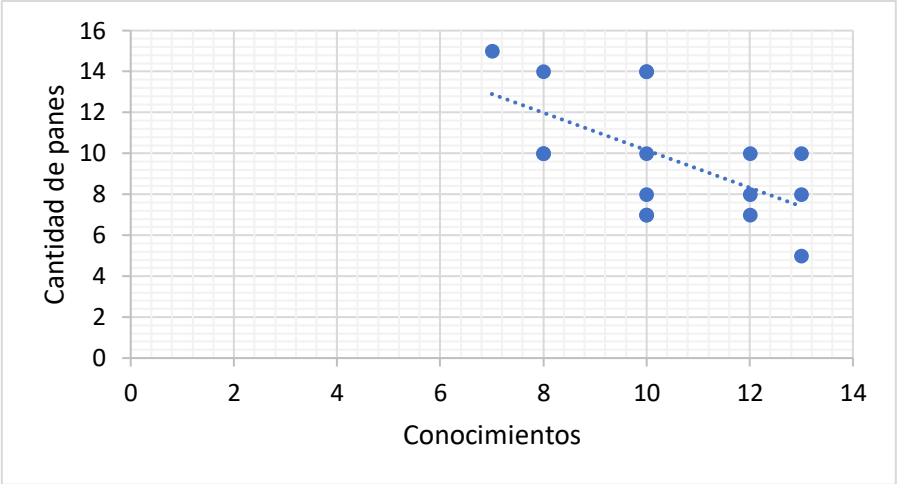


Figura 7.9 Diagrama de dispersión Conocimientos-cantidad de panes.

En la figura 7.9 se observa una tendencia lineal, pero con poco ajuste. Es decir, la correlación que se pueda establecer no será alta.

Para utilizar el coeficiente de correlación de Pearson se debe cumplir con el supuesto de normalidad. Por ello, se realizó la prueba de normalidad de Shapiro-Wilk porque el tamaño de las muestras es menor a 50. Los resultados fueron los siguientes:

Tabla 7.3 Prueba de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Significancia
Conocimiento	0.899	16	0.077
Consumo	0.901	16	0.084

El supuesto de normalidad se cumple cuando la significancia calculada es mayor que 0.05. En la tabla 7.3 se verifica la significancia para ambos grupos es mayor que 0.05, Por lo tanto, se verifica el supuesto de normalidad para ambos grupos de datos.

Luego, se calcula el coeficiente de correlación lineal de Pearson usando Excel. La ruta es la siguiente:

Datos-Análisis de datos-Coeficiente de correlación.

Los resultados se muestran a continuación:

Tabla 7.4 Coeficiente de correlación.

	Columna 1	Columna 2
Columna 1	1	
Columna 2	-0.59567244	1

En la tabla 7.4 se observa que el coeficiente de correlación es aproximadamente:

$$r = -0.596$$

Con base en el valor del coeficiente, realizamos la prueba de hipótesis para determinar si este coeficiente de correlación calculado es significativo.

Planteamos las hipótesis:

H_0: No existe correlación significativa entre los conocimientos sobre nutrición y el consumo semanal de panes.

H_1: Existe correlación significativa entre los conocimientos sobre nutrición y el consumo semanal de panes.

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.596\sqrt{16-2}}{\sqrt{1-0.596^2}} = -2.78$$

Con una significancia de 0.05, y 16-2=14 grados de libertad para una prueba bilateral:

$$t = 2.14$$

$|-2.78| = 2.78$ es mayor que 2.14. Entonces, se rechaza H_0.

En conclusión, existe correlación significativa entre los conocimientos sobre nutrición y el consumo semanal de panes. Además, dicha correlación es inversa, pero moderada.

También podemos utilizar SPSS para realizar esta prueba. La ruta es la siguiente:

Analizar-Correlacionar-Bivariadas-Pearson

El programa nos arroja los siguientes resultados:

Tabla 7.5 Coeficiente de correlación lineal.

		Variable 1	Variable 2
Variable 1	Correlación de Pearson	1	-0.598
	Sig. (bilateral)		0.014
	N	16	16
Variable 2	Correlación de Pearson	-0.598	1
	Sig. (bilateral)	0.014	
	N	16	16

En la tabla 7.5 se puede verificar el valor del coeficiente de correlación de Pearson $r = -0.598$. Además, la significancia calculada 0.014 es menor que 0.05. por lo tanto, se rechaza H_0.

Ejemplo 2: en una investigación se desea conocer la relación entre el nivel de estrés y el consumo de bebidas alcohólicas en los trabajadores jóvenes residentes de la urbanización X.

Para medir el nivel de estrés se utilizó una escala del 1 al 10 de tal manera que 1 representa un nivel bajo de estrés y 10 representa un nivel alto de estrés. Para medir el consumo de alcohol se les preguntó acerca de su consumo mensual en botellas y luego se calculó su equivalente en litros. Los resultados se muestran en la siguiente tabla:

Tabla 7.6 Matriz de datos.

Individuo	Variable 1 Estrés	Variable 2 Consumo
1	3	2
2	5	2
3	6	2
4	5	2.5
5	5	2.5
6	5	1
7	5	2.5
8	5	4
9	5	2.5
10	4	2
11	4	3
12	6	3

13	7	3
14	3	1
15	4	1
16	4	2.5
17	6	2.5
18	6	3

Título: relación entre el nivel de estrés y el consumo de bebidas alcohólicas en los trabajadores jóvenes residentes de la urbanización X.

Pregunta: ¿Existe relación entre el nivel de estrés y el consumo de bebidas alcohólicas en los trabajadores jóvenes residentes de la urbanización X?

Objetivo: conocer la relación entre el nivel de estrés y el consumo de bebidas alcohólicas en los trabajadores jóvenes residentes en la urbanización X.

Hipótesis: Existe una relación directa entre el nivel de estrés y el consumo de bebidas alcohólicas en los trabajadores jóvenes residentes en la urbanización X.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo correlacional y de corte transversal.

Variable 1: nivel de estrés.

Variable 2: consumo mensual de bebidas alcohólicas.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar si existe correlación significativa entre las variables en estudio.

Solución: Para tener un acercamiento al comportamiento conjunto de las variables se elabora el diagrama de dispersión para verificar si existe una tendencia lineal.

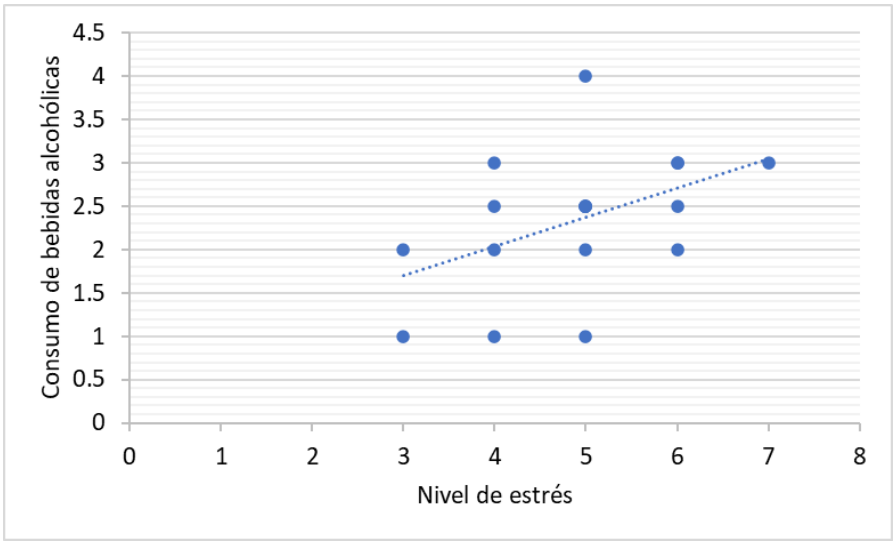


Figura 7.10 Diagrama de dispersión Nivel de estrés-Consumo de bebidas A.

En la figura 7.10 se observa una tendencia lineal, pero con poco ajuste. Es decir, la correlación que se pueda establecer no será alta.

Para utilizar el coeficiente de correlación de Pearson se debe cumplir con el supuesto de normalidad. Por ello, se realizó la prueba de normalidad de Shapiro-Wilk dado que el tamaño de muestra es menor a 50. Los resultados fueron los siguientes:

Tabla 7.7 Prueba de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Significancia
Nivel de estrés	0.928	18	0.177
Consumo de alcohol	0.909	18	0.083

El supuesto de normalidad se cumple cuando la significancia calculada es mayor que 0.05. En la tabla 7.7 se verifica que la significancia para ambos grupos es mayor que 0.05. Por lo tanto, se cumple el supuesto de normalidad y es adecuado realizar el análisis de correlación lineal de Pearson.

Luego, se calcula el coeficiente de correlación lineal de Pearson usando Excel. La ruta es la siguiente:

Datos-Análisis de datos-Coeficiente de correlación.

Tabla 7.8 Coeficiente de correlación.

	Columna 1	Columna 2
Columna 1	1	
Columna 2	0.462621	1

En la tabla 7.8 se observa que el coeficiente de correlación es aproximadamente:

$$r = 0.463$$

Con base en el valor del coeficiente, realizamos la prueba de hipótesis para determinar si este coeficiente de correlación calculado es significativo.

H_0: No existe correlación significativa entre el nivel de estrés y el consumo de bebidas alcohólicas.

H_1: Existe correlación significativa entre el nivel de estrés y el consumo de bebidas alcohólicas.

$$t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.463\sqrt{18-2}}{\sqrt{1-0.463^2}} = 2.09$$

Con una significancia de 0.05, y 18-2=16 grados de libertad para una prueba bilateral:

$$t = 2.12$$

2.09 no es mayor que 2.12. Entonces, no se rechaza H_0.

En conclusión, no existe correlación significativa entre los el nivel de estrés y el consumo de bebidas alcohólicas.

Se debe tener en consideración que la correlación calculada r=0.463 es la correlación para la muestra, pero este coeficiente no es válido para ser generalizado hacia la población. Por ello, no es necesario identificar si la correlación es fuerte, moderada o débil.

Utilizando SPSS para realizar esta prueba con la siguiente ruta:

Analizar-Correlacionar-Bivariadas-Pearson

El programa nos arroja los siguientes resultados:

Tabla 7.9 Coeficiente de correlación.

		Variable 1	Variable 2
Variable 1	Correlación de Pearson	1	0.463
	Sig. (bilateral)		0.053
	N	18	18
Variable 2	Correlación de Pearson	0.463	1
	Sig. (bilateral)	0.053	
	N	18	18

En la tabla 7.9 se puede verificar el valor del coeficiente de correlación de Pearson $r = 0.463$. Además, la significancia calculada 0.053 no es menor que 0.05. por lo tanto, no se rechaza H_0.

7.2.4. ¿Qué sucede cuando no se cumplen los supuestos?

En este caso se pueden transformar los datos como se sugirió en la prueba t. Sin embargo, es mejor optar por el uso de la prueba no paramétrica del coeficiente de correlación de Spearman.

Capítulo 8: Regresión Lineal

8.1. Regresión lineal simple

La regresión lineal simple permite identificar la relación lineal numérica que existe entre dos variables. Es decir, no solo permite identificar si la relación es directa o inversa, sino también estimar el valor de una variable a partir de la otra variable. La regresión lineal permite obtener la relación funcional entre las variables.

A este nivel de investigación ya se puede establecer la denominación de variable independiente y variable dependiente. A través del modelo de regresión lineal se puede cuantificar aproximadamente la variación que producen los cambios de la variable independiente sobre la variable dependiente.

El modelo matemático o modelo de regresión lineal es el siguiente:

$$y = a + bx + E$$

x: es la variable independiente.
y: es la variable dependiente.
a y b: son constantes. Es la pendiente de la recta asociada a la ecuación.
E: es el error de estimación.

Motgomery et al. (2006) consideran que el error E puede estar formado por el efecto de otras variable o factores. En los diseños experimentales dichos factores son denominados variables de control y variables intervinientes. Los mismos autores señalan que es más adecuado denominar a las variables x e y como variable predictora y variable de respuesta respectivamente.

Wackerly et al. (2010) también afirman que cuando se usa el modelo lineal para predecir valores de y, la predicción no es exacta. Es decir, siempre existe la posibilidad de que la predicción este sujeta a un error y este se puede determinar usando técnicas estadísticas.

La gran mayoría de fenómenos están determinados por diferentes variables o factores, por eso el modelo de regresión lineal adecuado es el que se ha señalado. Pero en algunos casos, cuando la variable respuesta (y) depende únicamente de una sola variable predictora (x), el modelo de regresión se puede simplificar:

$$y = a + bx$$

Por ejemplo, consideremos las variables minutos acumulados por llamadas telefónicas y costo mensual por el servicio de telefonía; obviamente estas variables están relacionadas y el modelo adecuado es el modelo lineal. Con una tarifa básica de S/ 20 y S/ 0.4 por cada minuto de llamadas telefónicas en un determinado mes, el costo mensual por el servicio de telefonía esta dado por la ecuación:

$$y = 20 + 0.4x$$

Donde x representa los minutos acumulados por llamadas telefónicas e y representa el costo mensual por el servicio de telefonía. En este modelo no aparece E puesto que la relación es exacta, el costo mensual por el servicio depende únicamente de los minutos de llamas telefónicas.

Así, cuando en un determinado mes se acumulan 45 minutos de llamadas telefónicas. El costo mensual por el servicio de telefonía será de S/ 38, como se verifica en la siguiente expresión.

$$y = 20 + 0.4(45) = 38$$

Cuando en un mes se acumulan 60 minutos de llamadas telefónicas. El costo mensual por el servicio de telefonía será de S/ 44, como se verifica en la siguiente expresión.

$$y = 20 + 0.4(60) = 44$$

Esta ecuación no solo permite estimar y a partir de x, también permite estimar x a partir de y. Por ejemplo, si en un determinado mes el costo por el servicio de telefonía es de S/ 46.8 ¿Cuántos minutos en llamadas

telefónicas se han acumulado dicho mes? Para responder a esta pregunta resolvemos la ecuación:

$$46.8 = 20 + 0.4x$$

$$46.8 - 20 = 0.4x$$

$$26.8 = 0.4x$$

$$\frac{26.8}{0.4} = x$$

$$67 = x$$

Por lo tanto, la cantidad de minutos acumulado en llamadas telefónicas en dicho mes es igual a 67 minutos.

Es necesario precisar que la regresión lineal no define automáticamente una relación de causa-efecto. La relación de causa-efecto debe tener un fundamento teórico. Por ello, en investigación se analiza adecuadamente la relación entre las variables y se sustenta el uso de la regresión lineal como ayuda para confirmar la relación de causa-efecto (Montgomery et al., 2006).

La regresión lineal se utiliza con más frecuencia y tiene mejores resultados cuando se aplica al estudio de fenómenos de las ciencias físicas, las ciencias naturales, gestión o las ingenierías porque en este tipo de fenómenos se puede reducir con mayor facilidad el efecto de otras variables que pueden afectar a la variable respuesta, pero que no son de interés para la investigación. En cambio, el uso de la regresión lineal en las ciencias sociales o las ciencias de la conducta presenta ciertas dificultades ya que las variables a analizar son muy complejas y dependen de muchos factores que no son fáciles de controlar.

8.2. Recta de regresión lineal

Es la recta teórica a la cual se ajustan la nube de puntos del diagrama de dispersión. Mientras más cerca estén los puntos a esta recta de regresión lineal, el modelo será más adecuado y se pueden hacer estimaciones más precisas.

Mediante las técnicas de regresión de una variable y sobre una variable x, buscamos una función que sea una buena aproximación a la nube de puntos (diagrama de dispersión).

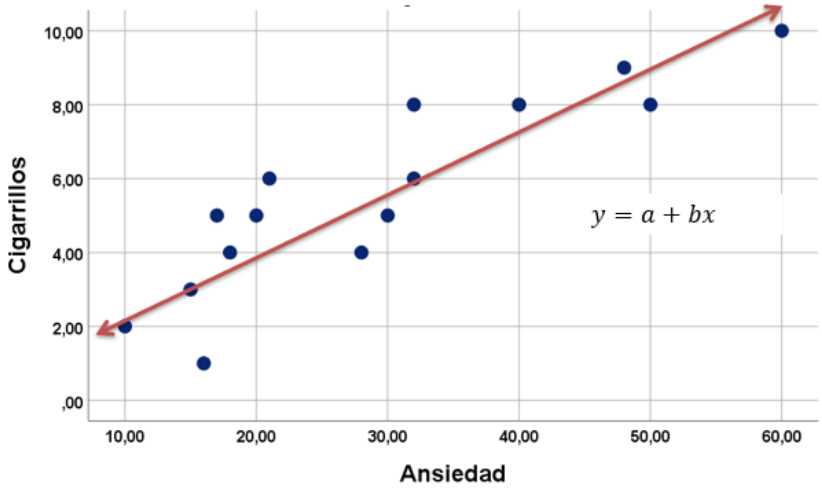


Figura 8.1 Diagrama de dispersión Ansiedad-Consumo de cigarrillos.

$y = a + bx + E$ representa el modelo lineal. Si hacemos una simplificación a la ecuación de la recta, presentada como modelo lineal, de manera que no se considere el error de estimación. Podemos usar la siguiente expresión como ecuación de la recta de regresión lineal:

$$\hat{y} = a + bx$$

Donde a y b son constantes o coeficientes que se deben calcular.

Los coeficientes de la ecuación de la recta de regresión lineal se calculan utilizando el método de mínimos cuadrados. Este método hace referencia a las distancias que existen entre los puntos del diagrama de dispersión y la recta de regresión. Se trata de establecer que el cuadrado de dichas distancias sea mínimo. Con este método se pueden calcular los valores de a y b utilizando sumatorias de los valores de x, y, del producto xy y de los cuadrados de x.

$$a = \frac{\sum y - b \sum x}{n}$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \sum x \sum x}$$

8.3. Error estándar de estimación (S_{yx})

Para estimar la capacidad de predicción de una ecuación de regresión se puede utilizar el error estándar de estimación que es la medida de la dispersión de los valores observados respecto de la recta de regresión para un valor dado de x (Wathen, Lind y Marchal, 2012).

$$S_{yx} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

Cuando el error estándar es pequeño, entonces, los valores $(x;\ y)$ que se calculan son cercanos a la recta de regresión. Esto quiere decir que mientras más pequeño es el error, la ecuación de la recta de regresión brinda mejores predicciones de y con base en los valores de x.

8.4. Significancia de los coeficientes de la ecuación de la recta de regresión lineal

Los coeficientes de la ecuación calculada hacen referencia solo a la muestra. Para establecer si estos coeficientes son significativos para la población, se establece una prueba de hipótesis para ambos coeficientes.

Por ejemplo, para determinar la significancia de la pendiente b se establecen las siguientes hipótesis.

H_0: la pendiente calculada no es significativa para la población.
H_1: la pendiente calculada es significativa para la población.

Simbólicamente:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

8.5. Condiciones para validar el modelo de regresión lineal

El desarrollo de la regresión lineal se sustenta con métodos paramétricos. Por ello, se debe verificar el cumplimiento de ciertos supuestos. Anderson et al. (2012) explican que estos supuestos están relacionados con el error E. Este valor E se origina de las diferencias entre los valores observados y los valores estimados por el modelo lineal.

Normalidad: se debe verificar que el error *E*, considerado como variable aleatoria, se distribuya normalmente.

Independencia de los Residuos: los valores de E deben ser independientes. Es decir, que un valor E solo está relacionado con un solo valor de x.

Homocedasticidad: la varianza de *E* con respecto a la recta de regresión es la misma para todos los valores de *x*.

Para el cálculo de la ecuación de la recta de regresión y la prueba de significancia de los coeficientes utilizaremos EXCEL.

Ejemplo 1: en una investigación se desea determinar la relación numérica lineal entre el consumo diario de azúcar y el índice de masa corporal (IMC) de los pacientes con problemas de sobrepeso en el hospital X. por ello, se toma una muestra representativa y se realiza la medición de las variables. Los resultados fueron los siguientes:

Tabla 8.1 Matriz de datos.

Consumo diario de azúcar en gramos	IMC
90	31
120	33.2
115	33
100	32
125	33.4
110	31.2
150	36.8
180	37.1
120	34.2
145	35.1
140	32.8

120	31.4
120	31.1
142	34.3
152	35.1
150	37.2

Título: Consumo de azúcar e IMC en pacientes con sobrepeso del Hospital X.

Problema: ¿Cómo influye el consumo de azúcar en el IMC de los pacientes con sobrepeso del Hospital X?

Objetivo: conocer la influencia del consumo de azúcar en el IMC de los pacientes con sobrepeso del Hospital X.

Hipótesis: el consumo de azúcar influye significativamente en el IMC de los pacientes con sobrepeso del Hospital X.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo correlacional causal y de corte transversal.

Variable x: Consumo diario de azúcar en gramos.

Variable y: índice de masa corporal.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la significancia del modelo lineal entre las variables en estudio.

Solución: Para tener un acercamiento al comportamiento conjunto de las variables se elabora el diagrama de dispersión para verificar si existe una tendencia lineal.

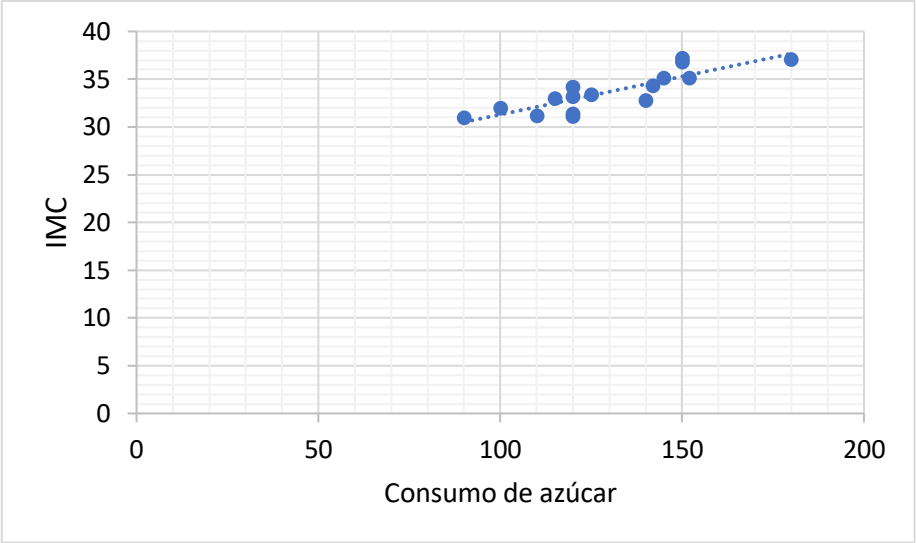


Figura 8.2 Diagrama de dispersión Consumo de azúcar-IMC.

En la figura 8.2 se observa una tendencia lineal con buen ajuste. Además, se puede decir que la relación entre las variables es directa.

Utilizamos Excel para el análisis de regresión lineal simple. La ruta es la siguiente:

Datos-Análisis de datos-Regresión

Tabla 8.2 Estadísticas de la regresión

Coeficiente de correlación múltiple	0.85347008
Coeficiente de determinación R^2	0.72841117
R^2 ajustado	0.70901197
Error típico	1.14908423
Observaciones	16

En la tabla 8.2 se observa que $r = 0.853$ lo que indica una correlación directa fuerte entre el consumo de azúcar y el IMC de los pacientes, el coeficiente de determinación $r^2 = 72.8\%$ nos indica que el 72.8% de la variación en el IMC está determinada por la variación en el consumo de azúcar.

Tabla 8.3 Análisis de varianza.

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados	F	Valor crítico de F
Regresión	1	49.578851	49.578851	37.5485	2.61E-05
Residuos	14	18.485524	1.32039457		
Total	15	68.064375			

En la tabla 8.3 el ANOVA nos permite establecer si el modelo lineal es significativo. Por ello, planteamos las hipótesis:

H_0: el modelo lineal no es significativo.
H_1: el modelo lineal es significativo.

La regla de decisión es la siguiente:

Si el F calculado es mayor que el F crítico, se rechaza H_0, en caso contrario no se rechaza.

En la tabla 8.3 se puede verificar que $F_C = 37.5485$ es mayor que $F = 0.0000261$. Por lo tanto, se rechaza H_0.

Esto quiere decir que el modelo lineal es significativo para explicar la relación entre las variables en estudio.

Tabla 8.4 Coeficientes de la ecuación de la recta de regresión lineal.

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Intercepción	23.3237803	1.7145126	13.6037381	1.8404E-09
Variable X 1	0.07971117	0.01300837	6.12768403	2.6162E-05

En Excel la intercepción corresponde al coeficiente a y la variable X1 corresponde al coeficiente b. por lo tanto, la ecuación de regresión lineal, a partir de los datos mostrados en la tabla 8.4, es la siguiente:

$$\hat{y} = 23.3238 + 0.0797x$$

En la tabla 8.4 también encontramos información para establecer la significancia de los coeficientes a y b. Para ello observamos los valores de probabilidad en cada caso.

Cuando realizamos la prueba de hipótesis con base en la probabilidad calculada o significancia calculada, se debe rechazar cuando la probabilidad es menor a la significancia de la prueba. En este ejemplo usamos un nivel de confianza de 95% lo que implica una significancia de 0.05.

H_0: el coeficiente a no es significativo.
H_1: el coeficiente a es significativo.

En la tabla 8.4 se observa que la significancia o probabilidad calculada es 0.0000000018 y es menor que 0.05. Entonces, se rechaza H_0. En conclusión, existe evidencia estadística para afirmar que el coeficiente a es significativo.

H_0: el coeficiente b no es significativo.
H_1: el coeficiente b es significativo.

En la tabla 8.4 se observa que la significancia o probabilidad calculada es 0.000026 y es menor que 0.05. Entonces, se rechaza H_0. En conclusión, existe evidencia estadística para afirmar que el coeficiente b es significativo.

Para realizar el análisis de regresión lineal usando SPSS debemos seguir la siguiente ruta:

Analizar-Regresión- Lineales

Los resultados son los siguientes:

Tabla 8.5 Resumen del modelo[b].

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	0.853[a]	0.728	0.709	1.14908	1.479

a. Predictores: (Constante), Consumo

b. Variable dependiente: IMC

En las tablas 8.5 y 8.6 se puede verificar los resultados obtenidos con Excel.

Tabla 8.6 ANOVA[a].

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	49.579	1	49.579	37.549	0.000[b]
	Residuo	18.486	14	1.320		
	Total	68.064	15			

a. Variable dependiente: IMC

b. Predictores: (Constante), Consumo

En la tabla 8.7 también se pueden verificar los coeficientes que componen la ecuación de la recta regresión lineal.

Tabla 8.7 Coeficientes[a].

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	Constante	23.324	1.715		13.604	0.000
	Consumo	0.080	0.013	0.853	6.128	0.000

a. Variable dependiente: IMC

Condiciones para validar el modelo de regresión lineal

Normalidad: Para esto debemos calcular los residuos, es decir las diferencias entre los valores y observados con valores estimados con la ecuación de regresión.

Tabla 8.8 Diferencia $y - \hat{y}$

x	y	$\hat{y} = a + bx$	Residuos
90	31	30.4977853	0.50221471
120	33.2	32.8891203	0.31087972
115	33	32.4905645	0.50943555
100	32	31.294897	0.70510305
125	33.4	33.2876761	0.11232388
110	31.2	32.0920086	-0.89200862
150	36.8	35.2804553	1.51954472
180	37.1	37.6717903	-0.57179027
120	34.2	32.8891203	1.31087972
145	35.1	34.8818994	0.21810056

140	32.8	34.4833436	-1.68334361
120	31.4	32.8891203	-1.48912028
120	31.1	32.8891203	-1.78912028
142	34.3	34.6427659	-0.34276595
152	35.1	35.4398776	-0.33987761
150	37.2	35.2804553	1.91954472

Planteamos las hipótesis para el conjunto de residuos:

H_0: Los datos provienen de una población con distribución normal.
H_1: Los datos no provienen de una población con distribución normal.

Analizar-Estadísticos descriptivos-Explorar-Gráficos-Gráficos de normalidad

Tabla 8.9 Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
Residuos	0.966	16	0.775

En la tabla 8.9 se observa que la significancia calculada 0.775 es mayor que 0.05. Entonces, no se rechaza H_0. Es decir, se verifica el supuesto de normalidad para los residuos (E).

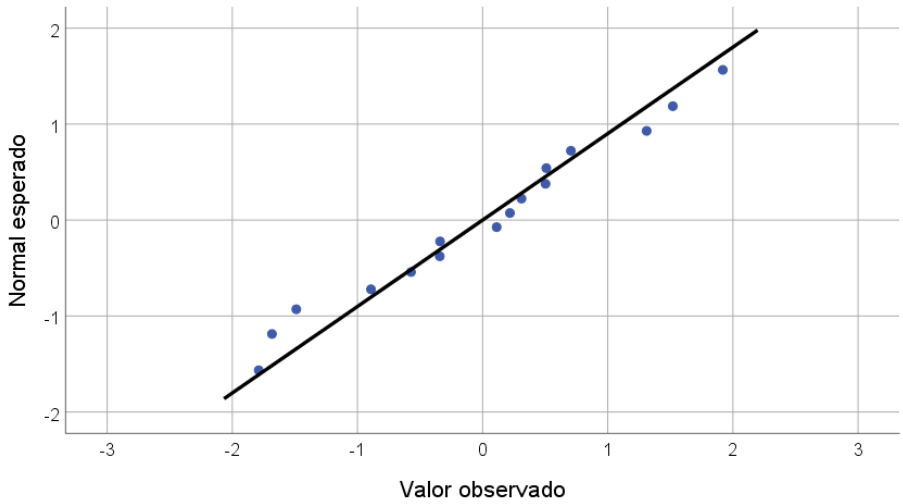


Figura 8.3 Gráfico Q-Q normal de residuos

La normalidad también se puede verificar de forma gráfica. SPSS nos brinda el gráfico Q-Q de normalidad. En la figura 8.3 lo que se debe tener en cuenta para verificar la normalidad es que los puntos estén muy cerca a la recta. Excel también nos brinda el gráfico de normalidad (figura 8.4). la opción para llegar a este gráfico se muestra en el mismo menú de la regresión lineal.

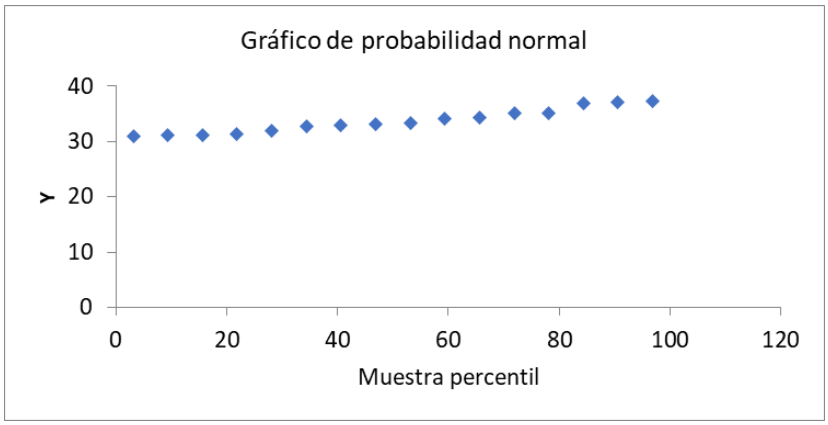


Figura 8.4 Gráfico de normalidad según Excel.

Homocedasticidad: planteamos las hipótesis para verificar la homogeneidad de varianzas:

H_0: Los conjuntos de datos provienen de poblaciones con varianzas iguales.
H_1: Los conjuntos de datos no provienen de poblaciones con varianzas iguales.

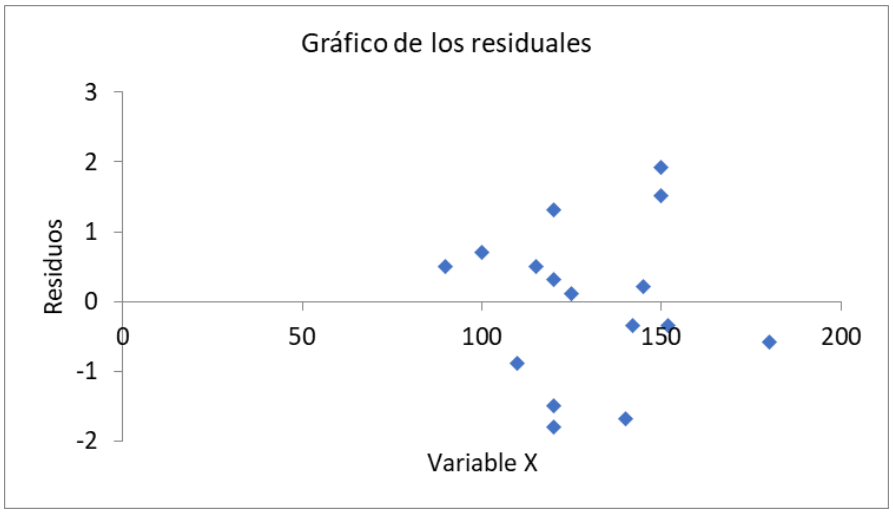


Figura 8.5 Gráfico de residuales según Excel.

Para verificar la homocedasticidad, Excel nos muestra la gráfica de los residuales. Para verificar la igualdad de varianzas en la nube de puntos no se debe observar ninguna tendencia. En la figura 8.5 no se observa ningún patrón o tendencia en la nube de puntos, entonces se verifica el supuesto de homocedasticidad.

Independencia de los Residuos: para verificar este supuesto usaremos el estadístico Durbin-Watson que se encuentra en SPSS.

Planteamos las hipótesis:

H_0: no existe autocorrelación.
H_1: existe autocorrelación.

La tabla 8.10 ya se ha utilizado al inicio del análisis con el uso de SPSS.

Tabla 8.10 Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	0.853[a]	0.728	0.709	1.14908	1.479

a. Predictores: (Constante), Consumo
b. Variable dependiente: IMC

El estadístico de Durbin-Watson (d) nos permite establecer la autocorrelación. Para usar este estadístico se establece un límite superior (d_U) y un límite inferior (d_L) que permiten tomar decisiones acerca de la presencia o ausencia de autocorrelación (Arranz y Zamora, 2002).

Si $d_U < d < 4 - d_U$, no se rechaza H_0, no existe autocorrelación.

De la tabla del apéndice de los anexos obtenemos los valores para 0.05 de significancia y $n = 16$: $d_L = 1.106$ y $d_U = 1.371$.

Sabemos que $d = 1.479$ (tabla 8.10). Entonces:

$$1.371 < 1.479 < 4 - 1.371$$

$$1.371 < 1.479 < 2.629$$

Dado que *d* pertenece al intervalo señalado, no se rechaza H_0. Se infiere que no existe autocorrelación.

Se han verificado los supuestos. Entonces, el modelo de regresión lineal permite hacer estimaciones muy significativas.

Capítulo 9: Prueba U de Mann-Whitney

9.1. Pruebas no paramétricas

Cuando los supuestos de las pruebas paramétricas no se cumplen, existen otros procedimientos alternativos, llamados no paramétricos o métodos de distribución libre, ya que no se asocian con las distribuciones estudiadas anteriormente (Walpole, Myers y Ye, 2007).

Se debe resaltar que las pruebas paramétricas tienen ciertas ventajas como:

- Mayor exactitud en las estimaciones.
- Mayor poder estadístico.
- Menor posibilidad de errores.
- Mayor sensibilidad a los rasgos de los datos.

Sin embargo, estas pruebas presentan algunas dificultades como el desarrollo de cálculos más complejos y limitaciones en los tipos de datos que se pueden evaluar. Triola (2004) afirma que las pruebas no paramétricas generalmente implican cálculos más simples.

Las pruebas no paramétricas están libres de los supuestos de las pruebas paramétricas, pero son pruebas con un menor poder estadístico. Siegel y Castellan (1998) señalan que las pruebas más potentes son aquellas que tienen las suposiciones más fuertes o extensas.

9.2. Prueba U de Mann-Whitney

Se utiliza para comparar dos grupos independientes. Esta prueba permite comparar dos grupos de rangos (medianas) y determinar que la diferencia sea estadísticamente significativa (Juárez, Villatoro y López, 2002).

Esta prueba es la prueba no paramétrica equivalente a la prueba t para muestras independientes y se utiliza cuando no se cumple el supuesto de normalidad o cuando la escala de medición es, por lo menos, ordinal.

El estadístico de prueba viene dado por:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

Donde:

n_1 y n_2 son los tamaños de muestra de cada grupo.
$\sum R_1$ y $\sum R_2$ son las sumas de rangos de cada grupo.

Regla de decisión para la U de Mann-Whitney

Una vez que se ha calculado los estadísticos de prueba U_1 y U_2, se procede de la siguiente manera: Se elige el menor valor de U_c.

Luego, ubicamos U en la tabla de valores críticos para U de Mann-Whitney.

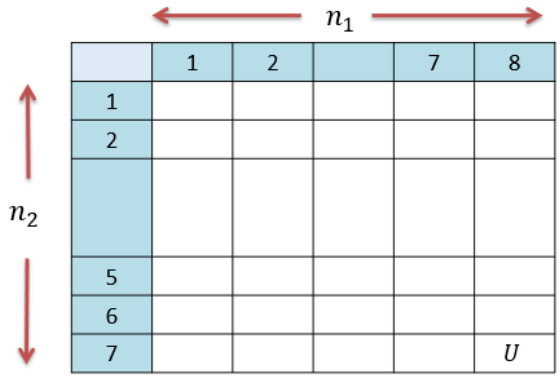


Figura 9.1 Forma de ubicar el valor de U crítico.

Si $U_c \leq U$ entonces, se rechaza H_o
Si $U_c > U$ entonces, no se rechaza H_o

Ejemplo 1: en una investigación se desea conocer el nivel de satisfacción de los clientes con respecto al tipo de servicio de desayuno que brinda la cadena de hoteles X. Por ello, se realiza una encuesta sobre la satisfacción del servicio de desayuno A y el servicio de desayuno B utilizando puntajes del 1 al 7. El 1 representa el menor grado de satisfacción y el 7 representa el mayor grado de satisfacción. Los resultados fueron los siguientes:

Tabla 9.1 Matriz de datos.

Satisfacción	
Servicio A	Servicio B
5	5
6	4
5	4
7	5
6	4
5	4
6	6
5	3
7	4
5	4
6	5

Título: Nivel de satisfacción por los servicios de desayuno en los clientes de la cadena de hoteles X.

Pregunta: ¿Existe diferencia en el nivel de satisfacción de los clientes por el tipo de servicio de desayuno brindado en la cadena de hoteles X?

Objetivo: conocer si existe diferencia en el nivel de satisfacción de los clientes por el tipo de servicio de desayuno brindado en la cadena de hoteles X.

Hipótesis: existe diferencia en el nivel de satisfacción de los clientes por el tipo de servicio de desayuno brindado en la cadena de hoteles X.

Tipo de investigación: enfoque cuantitativo, descriptivo y de corte transversal.

Variable 1: tipo de servicio.

Variable 2: nivel de satisfacción por el servicio de desayuno.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis para determinar si existen diferencias significativas en la satisfacción por los servicios de desayuno brindados por la cadena de hoteles.

Solución: planteamos las hipótesis para una prueba bilateral.

H_0: el nivel de satisfacción por el servicio A es igual al nivel de satisfacción por el servicio B.

H_1: el nivel de satisfacción por el servicio A es diferente al nivel de satisfacción por el servicio B.

Todos los datos, de ambos grupos, se deben ordenar en forma ascendente y luego se les asignan rangos:

Tabla 9.2 Cálculo de los rangos.

Posición	Valor	Rango
1	3	1
2	4	$\frac{2+3+4+5+6+7}{6} = 4.5$
3	4	
4	4	
5	4	
6	4	
7	4	
8	5	$\frac{8+9+10+11+12+13+14+15}{8} = 11.5$
9	5	
10	5	
11	5	
12	5	
13	5	
14	5	
15	5	
16	6	

17	6	$\frac{16+17+18+19+20}{5}=18$
18	6	
19	6	
20	6	
21	7	$\frac{21+22}{2}=21.5$
22	7	

En la tabla 9.2 se puede ver que cuando a una posición le corresponde un único valor, sin que este se repita, la posición es el rango de dicho valor. Pero si a un valor le corresponden varias posiciones, el rango de dicho valor es el promedio las posiciones correspondientes.

Luego se asignan los rangos correspondientes a cada valor y se calcula la suma de rangos.

Tabla 9.3 Asignación de los rangos.

Servicio A	Servicio B	Rango 1	Rango 2
5	5	11.5	11.5
6	4	18	4.5
5	4	11.5	4.5
7	5	21.5	11.5
6	4	18	4.5
5	4	11.5	4.5
6	6	18	18
5	3	11.5	1
7	4	21.5	4.5
5	4	11.5	4.5
6	5	18	11.5
		$\sum R_1$=172.5	$\sum R_2$=80.5

n_1 y n_2 son los tamaños de muestra de cada grupo y en este caso son iguales.

$$n_1 = n_2 = 11$$

Reemplazamos los datos en las siguientes expresiones:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U_1 = 11(11) + \frac{11(11+1)}{2} - 172.5 = 14.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

$$U_2 = 11(11) + \frac{11(11+1)}{2} - 80.5 = 106.5$$

Se elige el menor valor de U:

$$U_C = 14.5$$

Ubicamos U en la tabla de valores críticos para U de Mann-Whitney con $n_1 = n_2 = 11$.

$$U = 30$$

Como $U_C = 14.5$ es menor que $U = 30$, se rechaza H_0.

En conclusión, existe evidencia estadísticamente significativa para afirmar que el nivel de satisfacción por el servicio A es diferente al nivel de satisfacción por el servicio B.

Si usamos el programa SPSS, debemos seguir la ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-2 muestras independientes

Los datos se ingresan de la siguiente manera:

Tabla 9.4 Matriz de datos.

Tipo de servicio	Nivel de satisfacción
Servicio A	5.00
Servicio A	6.00
Servicio A	5.00
Servicio A	7.00
Servicio A	6.00
Servicio A	5.00
Servicio A	6.00

Servicio A	5.00
Servicio A	7.00
Servicio A	5.00
Servicio A	6.00
Servicio B	5.00
Servicio B	4.00
Servicio B	4.00
Servicio B	5.00
Servicio B	4.00
Servicio B	4.00
Servicio B	6.00
Servicio B	3.00
Servicio B	4.00
Servicio B	4.00
Servicio B	5.00

Obtenemos los siguientes resultados:

Tabla 9.5 Rangos

	Tipo de Servicio	N	Rango promedio	Suma de rangos
Satisfacción	Servicio A	11	15.68	172.50
	Servicio B	11	7.32	80.50
	Total	22		

En la tabla 9.5 se pueden observar los valores descriptivos para ambos tipos de servicio.

Tabla 9.6 Estadísticos para la prueba de hipótesis.

Estadísticos de prueba	Satisfacción
U de Mann-Whitney	14.500
W de Wilcoxon	80.500
Z	-3.148
Sig. asintótica(bilateral)	0.002
Significación exacta [2*(sig. unilateral)]	0.001

En la tabla 9.6 se puede ver que la significancia bilateral calculada es 0.002 y es menor que 0.05. Por lo tanto, se rechaza H_0.

¿Qué más se puede inferir?

Como ya se verificó que existen diferencias significativas en el nivel de satisfacción, sería necesario conocer ¿Con qué tipo de servicio se encuentran más satisfechos los clientes del hotel X?

Si calculamos los promedios de los puntajes obtenidos por grupo tenemos:

Tabla 9.7 Promedios por grupo.

Tipo de servicio	Promedio
Servicio A	5.73
Servicio B	4.36

En la tabla 9.7 se verifica que el mayor puntaje promedio corresponde el servicio de Tipo A. Pero, cuando usamos pruebas no paramétricas, porque no se verifica el supuesto de normalidad, lo que se compara son las medianas, no las medias. Además, las medias calculadas corresponden a la muestra no a la población.

Lo más adecuado es utilizar una prueba de hipótesis adicional. Por ello planteamos las hipótesis para una prueba unilateral:

H_0: el nivel de satisfacción por el servicio A no es superior al nivel de satisfacción por el servicio B.

H_1: el nivel de satisfacción por el servicio A es superior al nivel de satisfacción por el servicio B.

Usando el programa Minitab seguimos la ruta:

Estadísticas-No paramétricas-Mann-Whitney

Los resultados son los siguientes:

Tabla 9.8 Estadísticas descriptivas

Muestra	N	Mediana
Servicio A	11	6
Servicio B	11	4

Como se había comentado antes, en la prueba de Mann-Whitney se comparan las medianas (Tabla 9.8).

Tabla 9.9 Prueba de hipótesis.

Hipótesis nula	$H_0: \eta_1 - \eta_2 = 0$	
Hipótesis alterna	$H_1: \eta_1 - \eta_2 > 0$	
Método	Valor W	Valor p
No ajustado para empates	172.50	0.001
Ajustado para empates	172.50	0.001

En la tabla 9.9 se verifica que la significancia p=0.001 es menor que 0.05. Entonces se rechaza la hipótesis nula.

En conclusión, existe evidencia estadística para afirmar que el nivel de satisfacción por el servicio A es superior al nivel de satisfacción por el servicio B.

Ejemplo 2: en un estudio de mercado se quiere determinar si existen diferencias entre el consumo de yogurt natural en los distritos de San Miguel y Magdalena. Por ello, se toma una muestra representativa consumidores de cada distrito y se analiza su consumo mensual de yogurt natural. Los resultados son los siguientes:

Tabla 9.10 Matriz de datos.

Consumo mensual de yogurt (en litros)	
San Miguel	Magdalena
4	6
4	5
5	4
4.5	5
6	5
5	6
4	6
5	6
4	4

4	4
6	5
4	5
5	6
4.5	6
5.5	4.5
6	6
4	6

Título: consumo de yogurt natural en habitantes de los distritos de San Miguel y Magdalena.

Pregunta: ¿Existe diferencia en el consumo de yogurt natural en los habitantes de los distritos de San Miguel y Magdalena?

Objetivo: conocer si existe diferencia en el consumo de yogurt natural en los habitantes de los distritos de San Miguel y Magdalena.

Hipótesis: existe diferencia en el consumo de yogurt natural en los habitantes de los distritos de San Miguel y Magdalena.

Tipo de investigación: enfoque cuantitativo, descriptivo y de corte transversal.

Variable 1: distrito de residencia.

Variable 2: consumo mensual de yogurt natural.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis para determinar si existe diferencia significativa entre el consumo de los grupos analizados.

Solución: planteamos las hipótesis para una prueba bilateral.

H_0: el promedio de consumo mensual de yogurt natural en los habitantes de San Miguel es igual al promedio de consumo mensual de yogurt natural en los habitantes de Magdalena.

H_1: el promedio de consumo mensual de yogurt natural en los habitantes de San Miguel es diferente al promedio de consumo mensual de yogurt natural en los habitantes de Magdalena.

Para usar el programa SPSS, debemos seguir la ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-2 muestras independientes

Los resultados son los siguientes:

Tabla 9.11 Rangos

	Distrito	N	Rango promedio	Suma de rangos
Consumo	San Miguel	17	14.71	250.00
	Magdalena	17	20.29	345.00
	Total	34		

En la tabla 9.12 se observa la significancia bilateral p=0.088 y este valor no es menor que 0.05. Entonces, no se rechaza la hipótesis nula.

Tabla 9.12 Estadísticos de prueba[a].

	Consumo
U de Mann-Whitney	97.000
W de Wilcoxon	250.000
Z	-1.703
Sig. asintótica(bilateral)	0.088
Significación exacta [2*(sig. unilateral)]	0.106[b]

a. Variable de agrupación: Distrito

b. No corregido para empates.

En conclusión, no existe evidencia estadística para afirmar que el promedio de consumo mensual de yogurt natural en los habitantes de San Miguel sea diferente al promedio de consumo mensual de yogurt natural en los habitantes de Magdalena.

Capítulo 10: Prueba de Wilcoxon

Es una prueba de comparación de dos muestras relacionadas. Para utilizar esta prueba no se requiere verificar el supuesto de normalidad.

La variable dependiente debe pertenecer a una escala de medición ordinal o escalar y la variable independiente debe pertenecer a la escala de medición nominal.

La prueba de Wilcoxon se utiliza para comparar dos grupos de datos emparejados y determinar que la diferencia sea estadísticamente significativa (Juárez, Villatoro y López, 2002). Los datos corresponden a dos mediciones realizadas a una misma muestra en diferentes momentos.

Tabla 10.1 Datos de muestras relacionadas.

	Prueba 1 (Antes)	Prueba 2 (después)
Nivel de aceptación	4	2
	2	3
	5	3
	2	5
	3	3
	3	3
	4	2
	4	3
	5	4

En la tabla 10.1 se observa un ejemplo de cuando se puede utilizar la prueba de Wilcoxon. Ahí se observa la medición de la variable nivel de aceptación en dos momentos: antes de un determinado tratamiento y después de dicho tratamiento. La prueba en mención es adecuada cuando la escala de medición es ordinal, como se observa en este caso, o cuando

las diferencias entre los datos de cada grupo no provienen de una población con distribución normal.

La prueba de Wilcoxon es la prueba no paramétrica equivalente a la prueba t para muestras relacionadas. Por lo tanto, la prueba es utilizada para investigación cuasi experimental o investigación experimental ya que se desea conocer el efecto de una variable sobre otra variable.

Los datos recolectados se deben ubicar en una tabla de la siguiente manera:

Tabla 10.2 Matriz para la asignación de rangos.

X	Y	D	Rangos positivos	Rangos negativos
x_1	y_1	$x_1 - y_1$		
x_2	y_2	$x_2 - y_2$		
x_3	y_3	$x_3 - y_3$		
x_n	y_n	$x_n - y_n$		
n			T^+	T^-

Donde $\boldsymbol{T^+}$ y $\boldsymbol{T^-}$ son las sumas de rangos positivos y negativos.

Si alguna diferencia $x - y$ es igual a cero, se debe ajustar el número de sujetos a $\boldsymbol{n - 1}$.

El estadístico de prueba se determina seleccionando la menor suma de rangos.

$$T_c = mín\{T^+;\ T^-\}$$

Una vez que se ha determinado T_c y ajustado n, se debe ubicar en la tabla de valores críticos el valor T asociado.

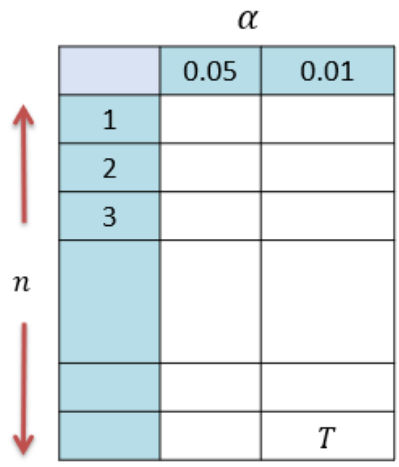


Figura 10.1 Forma de ubicar el valor de T crítico.

Regla de decisión

Para una prueba bilateral.

Si $T_c \leq T_{tabla}$ entonces, se rechaza H_o. En caso contrario, no se rechaza.

Ejemplo 1: en una investigación se quiere conocer el efecto que produce el consumo del edulcorante X en el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M". Por ello, se realiza una primera medición de la glucosa en el grupo experimental. Luego, al grupo experimental se le endulza los alimentos solo con el edulcorante y se les suministra una dieta regular. Además, se trata de controlar los demás factores que pueden afectar el nivel de glucosa. Los resultados obtenidos fueron los siguientes:

Tabla 10.3 Matriz de datos.

Nivel de glucosa (mg/dL)	
Antes	Después
160	150
190	170
175	150
180	160
165	170
165	160
180	170
150	150
140	160

120 110

Título: consumo de edulcorante X y el nivel de glucosa en los habitantes de la casa de reposo para el adulto mayor "M".

Pregunta: ¿Qué efecto tiene el consumo de edulcorante X sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M"?

Objetivo: conocer el efecto que tiene el consumo de edulcorante X sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

Hipótesis: el consumo de edulcorante X tiene efecto sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

Tipo de investigación: enfoque cuantitativo, experimental y de corte longitudinal.

Variable 1: Consumo de edulcorante X.

Variable 2: nivel de glucosa en la sangre.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis para determinar si el consumo del edulcorante X presenta un efecto significativo sobre el nivel de glucosa en la sangre.

Solución: planteamos las hipótesis para una prueba bilateral.

H_0: el consumo de edulcorante X no tiene efecto significativo sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

H_1: el consumo de edulcorante X tiene efecto significativo sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

Elaboramos una primera tabla en la cual se consideran las diferencias entre los valores del antes y el después.

Tabla 10.4 Diferencias $X - Y$.

X	Y	Diferencia (D)
160	150	10
190	170	20
175	150	25
180	160	20
165	170	-5
165	160	5
180	170	10
150	150	0
140	160	-20
120	110	10

Luego, sin considerar la diferencia igual a 0, ordenamos las diferencias en forma ascendente para determinar los rangos:

Tabla 10.5 Cálculo de los rangos.

Posición	Diferencia (D)	Rangos
1	-5	$\frac{1+2}{2} = 1.5$
2	5	
3	10	$\frac{3+4+5}{3} = 4$
4	10	
5	10	
6	-20	$\frac{6+7+8}{3} = 7$
7	20	
8	20	
9	25	9

Luego de quitar la diferencia igual a 0, $n = 9$.

Tabla 10.6 Asignación de rangos.

X	Y	D	Rangos positivos	Rangos negativos
160	150	10	4	
190	170	20	7	
175	150	25	9	
180	160	20	7	
165	170	-5		1.5
165	160	5	1.5	
180	170	10	4	
140	160	-20		7
120	110	10	4	
			$T^+ = 36.5$	$T^- = 8.5$

Entonces:

$$T_c = mín\{36.5;\ 8.5\} = 8.5$$

Para una prueba bilateral con $n = 9$ con 0.05 de significancia se debe ubicar en la tabla de valores críticos el T asociado.

$$T = 5$$

$T_c = 8.5$ no es menor que $T = 5$. Entonces, no se rechaza H_0.

En conclusión, no existe evidencia estadística para afirmar que el consumo de edulcorante X tiene efecto significativo sobre el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

Si usamos el programa SPSS, debemos seguir la ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-2 muestras relacionadas

Obtenemos los siguientes resultados:

Tabla 10.7 Descripción de los rangos.

		N	Rango promedio	Suma de rangos
Antes - Después	Rangos negativos	2[a]	4.25	8.50
	Rangos positivos	7[b]	5.21	36.50
	Empates	1[c]		
	Total	10		

a. Antes < Después
b. Antes > Después
c. Antes = Después

En la tabla 10.7 se muestra que en dos casos el nivel de glucosa se incrementó después del tratamiento, en 7 casos el nivel de glucosa se redujo y en 1 caso el nivel de glucosa se mantuvo igual. En la mayoría de casos el nivel de glucosa en la sangre se redujo, pero esto no es suficiente, estadísticamente, para afirmar que exista un efecto significativo del consumo de edulcorante X sobre el nivel de glucosa en

la sangre. Esto último se verifica con la prueba de hipótesis de la tabla 10.8.

Tabla 10.8 Estadísticos de prueba.

Estadísticos de prueba[a]	Antes - Después
Z	-1.672[b]
Sig. asintótica(bilateral)	0.095

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos negativos.

Cuando usamos SPSS verificamos que la significancia calculada sea menor que 0.05 para poder rechazar H_0. En la tabla 10.8 se puede ver que la significancia calculada es 0.095 y no es menor que 0.05, por lo tanto, no se rechaza H_0.

Observación: Cuando se cuenta con antecedentes o evidencias anteriores de que el edulcorante X es efectivo en la reducción del nivel de glucosa, se puede utilizar una prueba unilateral.

Para este caso planteamos las hipótesis para una prueba unilateral.

H_0: el consumo de edulcorante X no reduce el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

H_1: el consumo de edulcorante X reduce el nivel de glucosa de los habitantes de la casa de reposo para el adulto mayor "M".

En el ejemplo mostrado no es necesario utilizar una prueba unilateral, puesto que se ha verificado que no existe diferencia significativa entre los niveles de glucosa antes y después del consumo del edulcorante.

Ejemplo 2: En la cadena de restaurantes P se ha implementado el sistema digital X para que los clientes puedan seleccionar y solicitar su pedido. Mediante un estudio se desea verificar si esta implementación del sistema digital genera algún cambio en la satisfacción de los clientes con respecto al servicio de atención. Por ello, se toma una muestra representativa de clientes y se compara su satisfacción antes y después implementado el sistema digital. La satisfacción se mide con un instrumento que maneja una escala del 1 al 8, de menor a mayor satisfacción. Los resultados fueron los siguientes:

Tabla 10.9 Matriz de datos.

Satisfacción	
Antes	Después
7	7
6	7
5	7
5	6
5	8
5	8
6	8
4	8
6	7
6	6
8	7
5	7
5	7
6	6
5	7
4	7
8	8
6	8

Título: sistema digital X y satisfacción por el servicio en los clientes de la cadena de restaurantes P.

Pregunta: ¿Qué efecto tiene la implementación del sistema digital X sobre la satisfacción de los clientes de la cadena de restaurantes P?

Objetivo: conocer el efecto que tiene la implementación del sistema digital X sobre la satisfacción de los clientes de la cadena de restaurantes P.

Hipótesis: la implementación del sistema digital X genera cambios en la satisfacción de los clientes de la cadena de restaurantes P.

Tipo de investigación: enfoque cuantitativo, cuasi experimental y de corte longitudinal.

Variable 1: Tipo de sistema de pedido.

Variable 2: satisfacción por el servicio.

Usando un nivel de confianza de 95% realiza la prueba de hipótesis para determinar si la implementación del sistema digital X tiene efecto sobre la satisfacción de los clientes.

Solución: planteamos las hipótesis para una prueba bilateral.

H_0: la satisfacción de los clientes de la cadena de restaurantes P antes de la implementación del sistema digital X es igual a la satisfacción de los clientes después de la implementación de dicho sistema digital.

H_1: la satisfacción de los clientes de la cadena de restaurantes P antes de la implementación del sistema digital X es diferente a la satisfacción de los clientes después de la implementación de dicho sistema digital.

Usamos el programa SPSS, siguiendo la ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-2 muestras relacionadas

Obtenemos los siguientes resultados:

Tabla 10.10 Rangos.

		N	Rango promedio	Suma de rangos
Después - Antes	Rangos negativos	1[a]	2.50	2.50
	Rangos positivos	13[b]	7.88	102.50
	Empates	4[c]		
	Total	18		

a. Después < Antes
b. Después > Antes
c. Después = Antes

En la tabla 10.10 se observa que la satisfacción aumentó en 13 de los casos analizados, la satisfacción se redujo en 1 caso analizado y se mantuvo igual en 4 casos.

Tabla 10.11 Estadísticos de prueba[a]

	Después - Antes
Z	-3.177[b]
Sig. asintótica(bilateral)	0.001

a. Prueba de rangos con signo de Wilcoxon
b. Se basa en rangos negativos.

En la tabla 10.11 se observa que la significancia calculada 0.001 es menor que 0.05. Entonces, se rechaza H_0.

En conclusión, existe evidencia estadísticamente significativa para afirmar que la satisfacción de los clientes de la cadena de restaurante P antes de la implementación del sistema digital X es diferente a la satisfacción de los clientes después de la implementación de dicho sistema digital.

¿Qué más se puede concluir?

En el ejemplo desarrollado se verifica que existen diferencias significativas en los niveles de satisfacción antes y después de la implementación del sistema digital X, pero no se ha establecido con precisión si el nivel de satisfacción aumentó o disminuyó en la segunda medición de la satisfacción.

En la tabla 10.10 se verifica que, en la mayoría de los casos, la satisfacción aumentó luego de la implementación del sistema digital X (13 casos de un total de 18). Esto nos lleva a pensar que la implementación del sistema digital X aumenta la satisfacción de los clientes. Pero, el proceso adecuado, ahora, es realizar una prueba de hipótesis unilateral para saber si el nivel de satisfacción es mayor después de la implementación del sistema digital.

Planteamos las hipótesis para la prueba unilateral (Después>Antes).

H_0: la satisfacción de los clientes de la cadena de restaurante P después de la implementación del sistema digital X no es mayor a la satisfacción de los clientes antes de la implementación de dicho sistema digital.

H_1: la satisfacción de los clientes de la cadena de restaurante P después de la implementación del sistema digital X es mayor a la satisfacción de los clientes antes de la implementación de dicho sistema digital.

Usamos el programa Minitab, siguiendo la ruta:

Estadísticas-No paramétricos-Wilcoxon de 1 muestra

Ingresando las diferencias de los datos Después-Antes obtenemos los siguientes resultados:

Tabla 10.12 Prueba de hipótesis.

Hipótesis nula		H_0: $\eta = 0$	
Hipótesis alterna		H_1: $\eta > 0$	
Muestra	Número de prueba	Estadística de Wilcoxon	Valor p
Diferencia	14	102.50	0.001

En la tabla 10.12 verificamos que el p valor 0.001 es menor que 0.05. Entonces se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que la satisfacción de los clientes de la cadena de restaurante P después de la implementación del sistema digital X es mayor a la satisfacción de los clientes antes de la implementación de dicho sistema digital.

Capítulo 11: Coeficiente de correlación lineal de Spearman

Es una prueba no paramétrica que permite establecer la relación entre dos variables. El nivel de medición de las variables no necesariamente debe ser escalar. Esta prueba es la prueba alternativa a la prueba del coeficiente de correlación lineal de Pearson y se usa cuando no se verifica el supuesto de normalidad. Es una medida de asociación entre dos variables que requiere que ambas estén medidas, al menos, en una escala ordinal (Siegel y Castellan, 1998).

Esto último concuerda con los señalado por Juárez, Villatoro y López, (2002) quienes afirman que esta prueba se puede utilizar, aunque una de sus variables presente datos con nivel de medición escalar. Con frecuencia se utiliza esta prueba cuando ambas variables presentan un nivel de medición escalar. Cuando decimos nivel escalar, nos referimos al nivel de medición de intervalo y al nivel de medición de razón. En términos más sencillos, el nivel escalar hace referencia a las variables cuya medición nos arroja datos numéricos con los cuales tiene sentido realizar operaciones matemáticas.

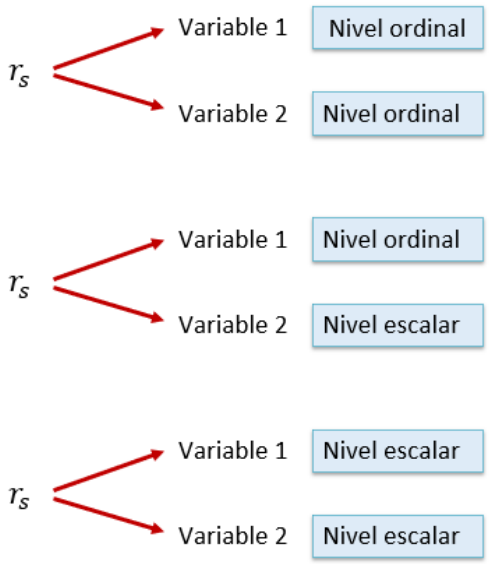


Figura 11.1 Formas de seleccionar variables para aplicar el r_s.

El cálculo del coeficiente de correlación de Spearman es más complejo que el cálculo del coeficiente de correlación de Pearson. Por ello, usaremos el programa SPSS para determinar el valor del coeficiente y la significancia calculada.

11.1. Interpretación del coeficiente r_s

Para el cálculo del coeficiente de correlación de Spearman los datos se traducen en rangos y la asociación se establece entre estos rangos. Por ello, la interpretación de este coeficiente es igual a la interpretación del coeficiente de correlación de Pearson. Este coeficiente de correlación por rangos adopta cualquier valor en el intervalo de -1 a 1 (Lind, Marchal y Wathen, 2012).

$$-1 \leq r_s \leq 1$$

Un valor de -1 indica una correlación negativa perfecta, y un valor de 1, una correlación positiva perfecta entre los rangos. Una correlación de rangos de 0 indica que no hay asociación entre los rangos. Los valores negativos del coeficiente indican una correlación inversa y los valores positivos una correlación directa.

Regla de decisión

Dado que para esta prueba utilizaremos el programa SPSS, entonces, si la significancia calculada por el programa es menor a 0.05, se rechaza H_0. En caso contrario, no se rechaza.

Ejemplo 1: en una investigación se desea conocer la relación entre el nivel de estrés y el IMC en estudiantes de preuniversitarios de la ciudad de Cusco. Por ello se toma una muestra de estudiantes y se les pide valorar su nivel de estrés usando una escala del 1 al 7. La variable IMC se obtiene a partir de la medición de la masa y la estatura de los estudiantes que son parte de la muestra. Los resultados fueron los siguientes:

Tabla 11.1 Matroz de datos.

Individuo	Variable 1 Estrés	Variable 2 IMC
1	3	21
2	5	23
3	6	22.8
4	5	22
5	5	22.5
6	7	24.5
7	7	23.5
8	5	24
9	6	24.5
10	4	22
11	4	23
12	6	24.1
13	7	24.5
14	3	22
15	4	22
16	4	22
17	7	24
18	6	23
19	6	23.5
20	4	22.5

Título: relación entre el nivel de estrés y el IMC en estudiantes de preuniversitarios de la ciudad de Cusco.

Pregunta: ¿Qué relación existe entre el nivel de estrés y el IMC en estudiantes de preuniversitarios de la ciudad de Cusco?

Objetivo: conocer relación que existe entre el nivel de estrés y el IMC en estudiantes de preuniversitarios de la ciudad de Cusco.

Hipótesis: existe una relación directa entre el nivel de estrés y el IMC en estudiantes de preuniversitarios de la ciudad de Cusco.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo correlacional y de corte transversal.

Variable 1: nivel de estrés.

Variable 2: IMC.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar si existe correlación significativa entre las variables en estudio.

Solución: Para tener un acercamiento al comportamiento conjunto de las variables se elabora el diagrama de dispersión para verificar si existe una tendencia lineal.

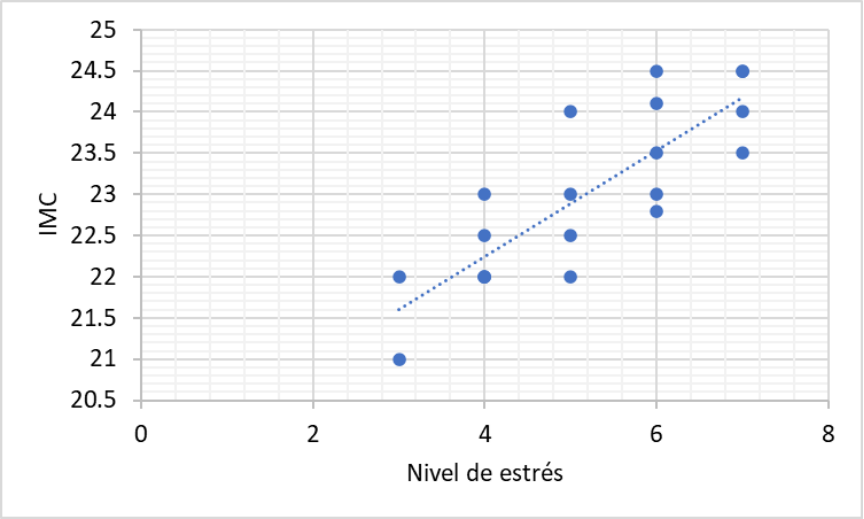


Figura 11.2 Diagrama de dispersión Nivel de estrés-IMC.

En la figura 11.2 se observa una tendencia lineal con buen ajuste. Es decir, la correlación que se pueda establecer será alta.

No es necesario verificar el supuesto de normalidad. Sin embargo, realizaremos la prueba de normalidad de Shapiro-Wilk. Los resultados son los siguientes:

Tabla 11.2 Prueba de normalidad.

	Shapiro-Wilk		
	Estadístico	gl	Significancia
Nivel de estrés	0.896	20	0.035
Consumo de alcohol	0.935	20	0.192

El supuesto de normalidad se cumple cuando la significancia calculada es mayor que 0.05. En la tabla 11.2 se verifica la normalidad para el consumo de alcohol, pero no se verifica la normalidad para el nivel de estrés puesto que 0.035 no es mayor que 0.05. Por lo tanto, no es adecuado realizar el análisis de correlación lineal de Pearson, en su lugar usaremos la prueba del coeficiente de correlación de Spearman.

Planteamos las hipótesis:

H_0: No existe correlación significativa entre el nivel de estrés y el IMC en estudiantes de preuniversitarios.

H_1: Existe correlación significativa entre el nivel de estrés y el IMC en estudiantes de preuniversitarios.

Utilizando SPSS para realizar esta prueba con la siguiente ruta:

Analizar-Correlacionar-Bivariadas-Spearman

El programa nos arroja los siguientes resultados:

Tabla 11.3 Estadísticos de prueba Rho de Spearman.

		Variable 1	Variable 2
Variable 1	Coeficiente de correlación	1	0.823
	Sig. (bilateral)		0.000
	N	20	20
Variable 2	Coeficiente de correlación	0.823	1
	Sig. (bilateral)	0.000	
	N	20	20

En la tabla 11.3 se puede verificar que la significancia calculada 0.000 es menor que 0.05. Por lo tanto, se rechaza H_0.

Es decir, existe evidencia estadística para afirmar que existe correlación significativa entre el nivel de estrés y el IMC en estudiantes de preuniversitarios.

Además, el valor del coeficiente de correlación de Spearman $r = 0.823$ indica que la correlación es directa y fuerte.

Ejemplo 2: en una investigación se desea conocer la relación entre la satisfacción laboral y los minutos de tardanza acumulados en el último año por los trabajadores de la empresa X.

Para la satisfacción laboral se elaboró un cuestionario con ítems que acumulan un puntaje máximo igual a 20. Para la variable minutos de tardanza acumulados, se solicitó la información al área de recursos humanos. Los resultados fueron los siguientes:

Tabla 11.4 Matriz de datos.

Individuo	Variable 1 Satisfacción	Variable 2 Minutos
1	12	15
2	15	10
3	10	20
4	18	15
5	18	20
6	16	15
7	12	20
8	15	10
9	15	12
10	16	20
11	13	22
12	16	16
13	10	30
14	18	18
15	13	20
16	15	15
17	15	10
18	18	20
19	8	15
20	16	10

21	18	5
22	17	10

Título: relación entre la satisfacción laboral y los minutos de tardanza acumulados en trabajadores de la empresa X.

Pregunta: ¿Qué relación existe entre la relación entre la satisfacción laboral y los minutos de tardanza acumulados en el último año por los trabajadores de la empresa X?

Objetivo: conocer la relación que existe entre la satisfacción laboral y los minutos de tardanza acumulados en el último año por los trabajadores de la empresa X.

Hipótesis: existe una relación inversa entre la satisfacción laboral y los minutos de tardanza acumulados en el último año por los trabajadores de la empresa X.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo correlacional y de corte transversal.

Variable 1: satisfacción laboral.

Variable 2: minutos de tardanza acumulados en el último año.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar si existe correlación significativa entre las variables en estudio.

Solución: Para tener un acercamiento al comportamiento conjunto de las variables se elabora el diagrama de dispersión para verificar si existe una tendencia lineal.

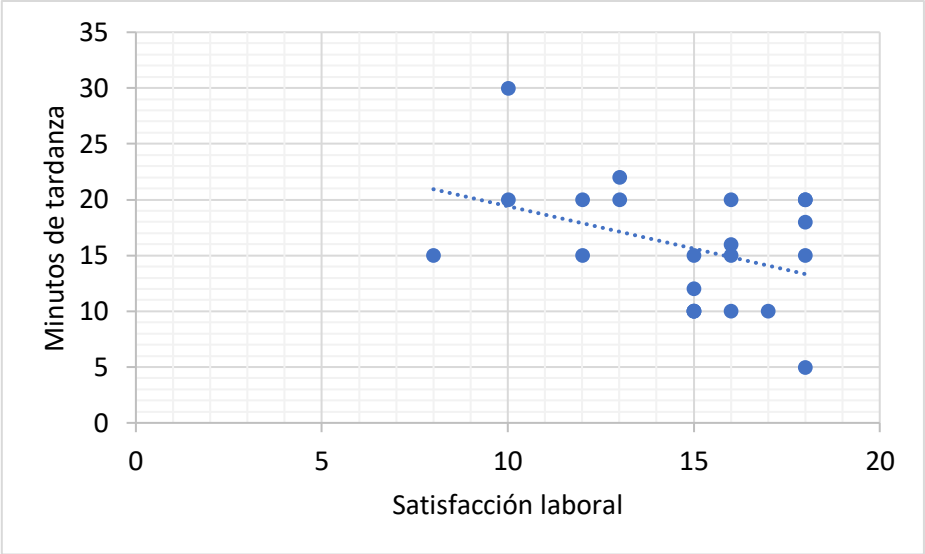


Figura 11.3 Diagrama de dispersión Satisfacción L.-Minutos de tardanza.

En la figura 11.3 se observa una tendencia lineal, pero con poco ajuste. Es decir, la correlación que se pueda establecer no será alta.

Planteamos las hipótesis para una prueba bilateral:

H_0: No existe correlación significativa entre la satisfacción laboral y los minutos de tardanza acumulados.

H_1: Existe correlación significativa entre la satisfacción laboral y los minutos de tardanza acumulados.

Utilizando SPSS para realizar esta prueba con la siguiente ruta:

Analizar-Correlacionar-Bivariadas-Spearman

El programa nos arroja los siguientes resultados:

Tabla 11.5 Estadísticos de prueba Rho de Spearman.

		Variable 1	Variable 2
Variable 1	Coeficiente de correlación	1	-0.262
	Sig. (bilateral)		0.239
	N	22	22
Variable 2	Coeficiente de correlación	-0.262	1
	Sig. (bilateral)	0.239	
	N	22	22

En la tabla 11.5 se puede verificar que la significancia calculada 0.239 no es menor que 0.05. Por lo tanto, no se rechaza H_0.

Es decir, no existe evidencia estadística para afirmar que la correlación entre la satisfacción laboral y los minutos de tardanza acumulados por los en el último año por los trabajadores de la empresa X sea significativa.

Sin embargo, para la muestra, el valor del coeficiente de correlación de Spearman $r = -0.262$ indica que la correlación es inversa y débil.

Capítulo 12: Prueba Chi cuadrado

Esta prueba estadística es muy útil para analizar la asociación entre variables categóricas. Además, Gorgas et al. (2011) afirman que este método estadístico es aplicable a datos que provienen de una variable tanto cualitativa como para las variables discretas o continuas.

Wackerly et al. (2008) consideran variables categóricas a las variables cualitativas las que se encuentran en el nivel de medición nominal y ordinal.

12.1. Chi cuadrado de Bondad de ajuste

Es una prueba que permite observar si existe diferencia entre las frecuencias observadas y alguna distribución teórica.

Se usa para determinar que tan bien se ajusta una distribución teórica como la distribución normal o la distribución t, a una distribución empírica, es decir, una distribución obtenida de datos muestrales (Spiegel y Stephens, 2009).

Tabla 12.1 Ajuste de los datos a una distribución teórica.

Eventos	Frecuencia observada	Frecuencia esperada
E_1	o_1	e_1
E_2	o_2	e_2
E_3	o_3	e_3
E_k	o_k	e_k

12.2. Chi cuadrado de independencia

Es una prueba que permite observar si existe asociación estadísticamente significativa entre dos variables (generalmente se usa para analizar variables categóricas).

En ambos casos, la prueba de chi cuadrado o ji cuadrado, se utiliza para determinar si las frecuencias observadas en las categorías difieren significativamente de las frecuencias esperadas (Hopkins, Hopkins y Glass, 1997).

12.2.1. Tabla de contingencia

La presentación de los datos para iniciar el análisis de la relación entre variables categóricas es a través de un cuadro de doble entrada como el que se muestra a continuación.

Tabla 12.2 Tabla de contingencia.

		Variable 1	
		Categoría 1	Categoría 2
Variable 2	Categoría 1		
	Categoría 2		
	Categoría 3		

La tabla de doble entrada o tabla de contingencia es una tabla en cuyo margen superior y margen izquierdo se sitúan las variables con sus respectivas categorías. También se le denomina tabla cruzada. En las casillas interiores de la tabla se ubica la frecuencia o número de casos observados.

Tabla 12.3 Tabla de contingencia.

	Comida rápida	Comida casera
Estudiante	18	15
Técnico	12	11
Profesional	10	12

En la tabla 12.3 se cruza la información acerca del nivel educativo y el tipo de comida que se consume. Los números que aparecen en el interior de la tabla se denominan frecuencias observadas.

12.2.2. Frecuencia Observada

Número real de casos en nuestro estudio. Es decir, la cantidad de casos por cada casilla en la tabla de contingencia.

12.2.3. Frecuencia Esperada

Número de casos que debería haber en las casillas si las variables fuesen independientes. En el ejemplo1, esto se explica con más detalle.

Las frecuencias esperadas se obtienen a partir de las frecuencias marginales y el total de casos observados.

Tabla 12.4 Frecuencias marginales.

	Comida rápida	Comida casera	Total
Estudiante	18	15	33
Técnico	12	11	23
Profesional	10	12	22
Total	40	38	78

40, 38, 33, 23, y 22 son las frecuencias marginales. El cálculo de las frecuencias esperadas se realiza de la siguiente manera:

Tabla 12.5 Cálculo de frecuencias esperadas.

	Comida rápida	Comida casera	Total
Estudiante	$\frac{40 \times 33}{78}$	$\frac{38 \times 33}{78}$	33
Técnico	$\frac{40 \times 23}{78}$	$\frac{38 \times 23}{78}$	23
Profesional	$\frac{40 \times 22}{78}$	$\frac{38 \times 22}{78}$	22
Total	40	38	78

Realizando las operaciones tenemos las frecuencias esperadas:

Tabla 12.6 Frecuencias esperadas.

	Comida rápida	Comida casera	Total
Estudiante	16.9	16.1	33
Técnico	11.8	11.2	23
Profesional	11.3	10.7	22
Total	40	38	78

Wackerly et al. (2010) resaltan que se necesita que todas las cantidades esperadas por celda sean al menos cinco para tener una aproximación adecuada a la distribución χ^2.

Para que se entienda el razonamiento de la prueba chi cuadrado podemos usar una situación en la que se quiere determinar si existe relación entre el estado civil y el hábito de fumar en un grupo de varones.

Tabla 12.7 Tabla de contingencia.

		Estado civil		
		Casado	Soltero	Total
Fumador	Si	6	10	16
	No	12	14	26
Total		18	24	42

Ahora, establecemos las frecuencias esperadas m, n, p y q, considerando que estas se distribuyen proporcionalmente a sus frecuencias marginales y al total de datos:

Tabla 12.7 Cálculo de frecuencias esperadas.

		Estado civil		
		Casado	Soltero	Total
Fumador	Si	m	n	16
	No	p	q	26
Total		18	24	42

Entonces, la razón entre la cantidad de fumadores casados y el total de casados debe ser igual a la razón entre el total de fumadores y el total de personas.

$$\frac{m}{18} = \frac{16}{42}$$

Resolviendo la ecuación obtenemos la primera frecuencia esperada:

$$m = \frac{18 \times 16}{42} = 6.86$$

Realizamos el mismo proceso para las demás frecuencias:

$$\frac{n}{24} = \frac{16}{42} \rightarrow m = 9.14$$

$$\frac{p}{18} = \frac{26}{42} \rightarrow p = 11.14$$

$$\frac{q}{24} = \frac{26}{42} \rightarrow q = 14.86$$

Luego completamos la tabla con las frecuencias esperadas:

Tabla 12.8 Frecuencias esperadas.

		Estado civil		
		Casado	Soltero	Total
Fumador	Si	6.86	9.14	16
	No	11.14	14.86	26
Total		18	24	42

Como ya se expresó anteriormente, estas frecuencias esperadas son proporcionales a las frecuencias marginales. Esta proporcionalidad implica la falta de relación entre las variables.

En el caso de los casados, la razón entre los que fuman y no fuman es igual a la razón entre los solteros que fuman y no fuman.

$$\frac{casados\ que\ fuman}{casados\ que\ no\ fuman} = \frac{6.86}{11.14} = 0.62$$

$$\frac{solteros\ que\ fuman}{solteros\ que\ no\ fuman} = \frac{9.14}{14.86} = 0.62$$

Esta situación nos lleva a concluir que el estado civil no determina ni se relaciona con el hábito de fumar.

Por lo tanto, en una investigación en la que se intenta demostrar una relación entre variables, se debe verificar que las frecuencias observadas sean diferentes a las frecuencias esperadas. Con más precisión, se debe verificar que la diferencia entre las frecuencias observadas y esperadas sea significativa.

Para realizar la prueba de hipótesis utilizaremos el siguiente estadístico:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Donde f_o es la frecuencia observada y f_e es la frecuencia esperada.

Los grados de libertad se calculan de la siguiente manera:

$$gl = (r-1)(c-1)$$

r y c son el número de filas y columnas respectivamente.

Ubicamos el valor crítico χ^2_{tabla} en la tabla Chi Cuadrada de acuerdo a los grados de libertad y al nivel de significancia.

	Significancia	
gl	0.05	0.01
1		
2		
3		
4		

Grados de libertad

Figura 12.1 Forma de ubicar el χ^2 en la tabla.

Decisión estadística:

Si $\chi^2 \geq \chi^2_{tabla}$, entonces se rechaza H_0.
Si $\chi^2 < \chi^2_{tabla}$, entonces no se rechaza H_0.

12.2.4. El nivel de significancia para la prueba de hipótesis con el uso de software

La significancia se puede calcular mediante tres métodos cuya elección responde a las características de la tabla de contingencia y al tamaño de la muestra:

- **Asintótico:** Se usa cuando la muestra tiene más de 20 elementos, menos del 20% de casillas tienen frecuencias esperadas menores a 5 y no hay casillas con frecuencia observada menor a 1.

- **Monte Carlo:** se puede usar cuando no se cumplen los anteriores supuestos, pero $n > 30$.

- **Prueba Exacta de Fisher:** se utiliza cuando no se cumplen los supuestos anteriores.

12.2.5. Medidas del grado asociación

Chi cuadrado no indica la magnitud de la asociación entre las variables, hay que usar estadísticos adicionales para medir el grado de asociación entre las variables. La elección del estadístico depende del tipo de tabla que presente nuestro estudio.

- Para tablas 2x2 se puede utilizar el Coeficiente Phi r(φ), la V de Cramer y el Riesgo de Odd.
- Para tablas Jxk: se puede utilizar el Coeficiente de contingencia o la V de Cramer.
- Cuando se tienen muestras relacionadas con pre y post test se utiliza el estadístico de McNemar.

Coeficiente Phi (φ)

- Este coeficiente puede tomar valores entre 0 y $\sqrt{q} - 1$
- «q» representa a la menor cantidad de categorías entre las variables.

V de Cramer

- Este coeficiente puede tomar valores entre 0 y 1
- «0» es independencia absoluta y «1» dependencia perfecta.
- La V de Cramer es una medida del tamaño del efecto para la prueba chi-cuadrado de independencia y se puede interpretar de la siguiente manera (IBM):

Tabla 12.9 Interpretación de la V de Cramer.

V de Cramer	Interpretación
$V \leq 0.2$	La asociación es débil. Aunque el resultado es estadísticamente significativo.
$0.2 < V \leq 0.6$	La asociación es moderada.
$V > 0.6$	La asociación es fuerte.

Coeficiente de contingencia (C)

- «0» representa la independencia absoluta
- $C = \sqrt{\chi^2/(\chi^2 + n)}$.
- Valor máximo de C: $Máx(c) = \sqrt{\frac{k-1}{k}}$
- K: es el menor valor de filas o columnas
- Comparando C con el valor máximo de C se identifica si la asociación es baja, moderada o alta.

Ejemplo 1: mediante una investigación se quiere conocer la relación que existe entre el consumo de productos veganos y el género. Por ello, se toma una muestra representativa de pobladores jóvenes de la urbanización X. Los resultados son los siguientes:

Tabla 12.10 Matriz de datos.

Poblador	Consumo de productos veganos	Genero
1	Si	Masculino
2	Si	Femenino
3	No	Masculino
4	No	Masculino

5	Si	Masculino
6	No	Femenino
7	Si	Femenino
8	Si	Femenino
9	Si	Femenino
10	No	Masculino
11	No	Masculino
12	No	Femenino
13	No	Masculino
14	Si	Femenino
15	Si	Femenino
16	No	Masculino
17	No	Masculino
18	Si	Masculino
19	Si	Femenino
20	Si	Femenino
21	Si	Femenino
22	No	Femenino
23	No	Femenino
24	No	Masculino
25	No	Masculino
26	No	Masculino
27	No	Masculino
28	No	Femenino

Título: Consumo de productos veganos y género en pobladores jóvenes de la urbanización X.

Pregunta: ¿Cómo se relaciona el consumo de productos veganos y el género en los pobladores jóvenes de la urbanización X?

Objetivo: conocer la relación entre el consumo de productos veganos y el género en los pobladores jóvenes de la urbanización X.

Hipótesis: el consumo de productos veganos se relaciona con el género de pobladores jóvenes de la urbanización X.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo relacional y de corte transversal.

Variable 1: género.

Variable 2: consumo de productos veganos.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la relación entre las variables en estudio.

Solución:

Planteamos las hipótesis:

H_0: el consumo de productos veganos no se relaciona significativamente con el género de los pobladores jóvenes de la urbanización X.

H_1: el consumo de productos veganos se relaciona significativamente con el género de los pobladores jóvenes de la urbanización X.

Como primer paso se debe elaborar la tabla de contingencia. Para ello, clasificamos las respuestas de los pobladores.

Tabla 12.11 Clasificación de respuestas.

Poblador	Consumo de productos veganos	Genero
1	Si	Masculino
2	Si	Femenino
3	No	Masculino
4	No	Masculino
5	Si	Masculino
6	No	Femenino
7	Si	Femenino
8	Si	Femenino
9	Si	Femenino
10	No	Masculino
11	No	Masculino
12	No	Femenino
13	No	Masculino
14	Si	Femenino
15	Si	Femenino
16	No	Masculino
17	No	Masculino
18	Si	Masculino
19	Si	Femenino
20	Si	Femenino

21	Si	Femenino
22	No	Femenino
23	No	Femenino
24	No	Masculino
25	No	Masculino
26	No	Masculino
27	No	Masculino
28	No	Femenino

Realizamos el conteo respectivo y se genera la siguiente tabla:

Tabla 12.12 Frecuencias observadas.

	Si	No	Total
Masculino	3	11	14
Femenino	9	5	14
Total	12	16	78

Calculamos las frecuencias esperadas:

Tabla 12.12 Cálculo de frecuencias esperadas.

	Si	No	Total
Masculino	$\frac{12 \times 14}{28}$	$\frac{16 \times 14}{28}$	14
Femenino	$\frac{12 \times 14}{28}$	$\frac{16 \times 14}{28}$	14
Total	12	16	28

Finalmente tenemos los valores de las frecuencias esperadas:

Tabla 12.13 Frecuencias esperadas.

	Si	No	Total
Masculino	6	8	14
Femenino	6	8	14
Total	12	16	28

Con los resultados de la tabla 12.13 se puede entender mejor el significado de la frecuencia esperada. Como se puede ver hay 12 personas que si consumen productos veganos, pero 6 son varones y 6 son mujeres, hay 16 personas que no consumen productos veganos, pero 8 son varones y 8 son mujeres; esto quiere decir que el género no es un factor que determina el consumo. También se puede interpretar que el consumo de productos veganos no depende del género. En otras palabras, las variables son independientes. Recordemos que la definición de frecuencia esperada hace referencia al número de casos que debería haber en las casillas si las variables fueran independientes.

Para calcular el estadístico de prueba reemplazamos las frecuencias observadas y esperadas en la siguiente expresión:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(11-8)^2}{8} + \frac{(9-6)^2}{6} + \frac{(5-8)^2}{8}$$

$$\chi^2 = \frac{(-3)^2}{6} + \frac{(3)^2}{8} + \frac{(3)^2}{6} + \frac{(-3)^2}{8}$$

$$\chi^2 = \frac{9}{6} + \frac{9}{8} + \frac{9}{6} + \frac{9}{8}$$

$$\chi^2 = 5.25$$

Sabemos que la cantidad de filas y columnas de la tabla de contingencia es la misma $r = c = 2$. Entonces, los grados de libertad son:

$$gl = (2-1)(2-1) = 1$$

Luego, con 1 grado de libertad para una significancia de 0.05 ubicamos el valor del chi cuadrado de la tabla (ver apéndice):

$$\chi^2_{tabla} = 3.84$$

5.25 es mayor que 3.94. Entonces, se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que el consumo de productos veganos se relaciona significativamente con el género de los pobladores jóvenes de la urbanización X.

Para realizar la prueba Chi cuadrado podemos usar el programa SPSS. La ruta es la siguiente:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos

Los resultados son los siguientes:

Tabla 12.14 Pruebas de chi-cuadrado.

	Valor	df	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	5.250[a]	1	0.022		
Corrección de continuidad[b]	3.646	1	0.056		
Razón de verosimilitud	5.445	1	0.020		
Prueba exacta de Fisher				0.054	0.027
Asociación lineal por lineal	5.062	1	0.024		
N de casos válidos	28				

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 6.00.

b. Sólo se ha calculado para una tabla 2x2

Usaremos la significancia asintótica bilateral porque la muestra tiene más de 20 elementos (n=28), los resultados de la tabla 12.14 nos indican que, menos del 20% de casillas tienen frecuencias esperadas menores a 5 (0.0%) y no hay casillas con frecuencia observada menor a 1.

Entonces, como la significancia asintótica bilateral es 0.022 y es menor que 0.05, se rechaza H_0.

¿Qué tipo de relación existe?

El valor calculado Chi cuadrado nos permite verificar que existe relación significativa, pero no nos brinda mayor información acerca de esta relación. Por ello debemos analizar la tabla de contingencia para interpretar el tipo de relación existente entre las variables.

Tabla 12.15 Frecuencias observadas.

	Si	No	Total
Masculino	3	11	14
Femenino	9	5	14
Total	12	16	78

En la tabla 12.15 se observa que de las personas que consumen productos veganos la mayoría son mujeres y del grupo de personas que no consumen productos veganos la mayoría son varones. Por ello, podemos identificar que el consumo de productos veganos se asocia con el género femenino. En una investigación experimental se podría decir que el género es un factor que determina el consumo de productos veganos.

¿Cuál es el grado de asociación?

El valor calculado Chi cuadrado no indica el grado de asociación entre las variables en estudio. Es decir, no establece si la asociación es fuerte o débil. Por lo tanto, se debe recurrir a otros estadísticos que miden el grado de asociación. Las condiciones para el uso de estos estadísticos ya se detallaron durante la exposición teórica del tema.

Por las características de este ejemplo (tabla $k \times k$) utilizaremos la V de Cramer. El programa SPSS brinda este estadístico siguiendo la ruta.

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos-Nominal

Tabla 12.16 Medidas simétricas.

		Valor	Significación aproximada
Nominal por	Phi	-0.433	0.022
Nominal	V de Cramer	0.433	0.022
N de casos válidos		28	

En la tabla 12.16 se observa que le valor de la V de Cramer es 0.433. este valor se encuentra en el intervalo:

$$0.2 < V \leq 0.6$$

Por lo tanto, el grado de asociación entre las variables es moderado.

Ejemplos 2: en una investigación se quiere conocer si existe relación entre el grupo etario y el consumo de animes en los integrantes del colegio X. Por ello, se toma una muestra representativa de 32 personas entre niños, adolescentes y adultos. Los resultados fueron los siguientes:

Tabla 12.17 Matriz de datos.

Integrante	Grupo etario	Consumo de animes
1	Adulto	Si
2	Adulto	No
3	Niño	Si
4	Niño	No
5	Niño	Si
6	Adulto	No
7	Niño	No
8	Niño	No
9	Adulto	No
10	Adolescente	Si
11	Niño	Si
12	Niño	Si
13	Adolescente	No
14	Adolescente	No
15	Niño	No
16	Niño	Si
17	Niño	Si
18	Adolescente	Si
19	Adolescente	Si
20	Adolescente	No
21	Adolescente	Si
22	Adulto	Si
23	Adolescente	Si
24	Adulto	No
25	Adulto	Si
26	Adulto	No
27	Niño	Si
28	Adolescente	Si
29	Adolescente	Si
30	Adolescente	No
31	Adolescente	Si
32	Adolescente	Si
33	Niño	No
34	Niño	Si

Título: Consumo de animes y grupo etario en integrantes del colegio X.

Pregunta: ¿Cómo se relaciona el consumo de animes y el grupo etario en integrantes del colegio X?

Objetivo: conocer la relación entre el consumo de animes y el grupo etario en los integrantes del colegio X.

Hipótesis: el consumo de animes se relaciona con el grupo etario en los integrantes del colegio X.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo relacional y de corte transversal.

Variable 1: grupo etario.

Variable 2: consumo de animes.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la relación significativa entre las variables en estudio.

Solución:

Planteamos las hipótesis:

H_0: el consumo de animes no se relaciona significativamente con el grupo etario en los integrantes del colegio X.

H_1: el consumo de animes se relaciona significativamente con el grupo etario en los integrantes del colegio X.

Como primer paso se debe elaborar la tabla de contingencia. Para ello, clasificamos las respuestas de los integrantes del colegio.

Tabla 12.18 Clasificación de las respuestas.

Integrante	Grupo etario	Consumo de animes
1	Adulto	Si
2	Adulto	No
3	Niño	Si
4	Niño	No
5	Niño	Si
6	Adulto	No
7	Niño	No
8	Niño	No
9	Adulto	No
10	Adolescente	Si
11	Niño	Si
12	Niño	Si
13	Adolescente	No
14	Adolescente	No
15	Niño	No
16	Niño	Si
17	Niño	Si
18	Adolescente	Si
19	Adolescente	Si
20	Adolescente	No
21	Adolescente	Si
22	Adulto	Si
23	Adolescente	Si
24	Adulto	No
25	Adulto	Si
26	Adulto	No
27	Niño	Si
28	Adolescente	Si
29	Adolescente	Si
30	Adolescente	No
31	Adolescente	Si
32	Adolescente	Si
33	Niño	No
34	Niño	Si

Realizamos el conteo respectivo y se genera la siguiente tabla:

Tabla 12.19 Frecuencias observadas.

	Si	No	Total
Niño	8	5	13
Adolescente	9	4	13
Adulto	3	5	8
Total	20	14	34

Calculamos las frecuencias esperadas:

Tabla 12.20 Cálculo de las frecuencias esperadas.

	Si	No	Total
Niño	$\frac{20 \times 13}{34}$	$\frac{14 \times 13}{34}$	13
Adolescente	$\frac{20 \times 13}{34}$	$\frac{14 \times 13}{34}$	13
Adulto	$\frac{20 \times 8}{34}$	$\frac{14 \times 8}{34}$	8
Total	20	14	34

Finalmente tenemos los valores aproximados de las frecuencias esperadas:

Tabla 12.21 Frecuencias esperadas.

	Si	No	Total
Niño	7.6	5.4	13
Adolescente	7.6	5.4	13
Adulto	4.7	3.3	8
Total	20	14	34

Para calcular el estadístico de prueba reemplazamos las frecuencias observadas y esperadas en la siguiente expresión:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(8-7.6)^2}{7.6} + \frac{(5-5.4)^2}{5.4} + \frac{(9-7.6)^2}{7.6} + \frac{(4-5.4)^2}{5.4} + \frac{(3-4.7)^2}{4.7} + \frac{(5-3.3)^2}{3.3}$$

$$\chi^2 = \frac{(0.4)^2}{7.6} + \frac{(-0.4)^2}{5.4} + \frac{(1.4)^2}{7.6} + \frac{(-1.4)^2}{5.4} + \frac{(-1.7)^2}{4.7} + \frac{(1.7)^2}{3.3}$$

$$\chi^2 = \frac{2}{95} + \frac{4}{135} + \frac{49}{190} + \frac{49}{135} + \frac{289}{470} + \frac{289}{330}$$

$$\chi^2 = 2.16$$

Sabemos que la cantidad de filas y columnas de la tabla de contingencia son $r = 3$ y $c = 2$. Entonces, calculamos los grados de libertad:

$$gl = (3-1)(2-1) = 2$$

Luego, con 2 grado de libertad para una significancia de 0.05 ubicamos el valor del chi cuadrado de la tabla:

$$\chi^2_{tabla} = 5.99$$

2 no es mayor que 5.99. Entonces, no se rechaza H_0.

En conclusión, no existe evidencia estadística para afirmar que el consumo de animes se relacione significativamente con el grupo etario en los integrantes del colegio X.

Para realizar la prueba Chi cuadrado podemos usar el programa SPSS. La ruta es la siguiente:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos

Los resultados son los siguientes:

Tabla 12.22 Pruebas de chi-cuadrado.

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2.123[a]	2	0.346
Razón de verosimilitud	2.113	2	0.348
Asoc. lineal por lineal	0.838	1	0.360
N de casos válidos	34		

a. 2 casillas (33,3%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,29.

En la tabla 12.22 la nota de la parte inferior nos alerta que no se cumple una condición para usar la significancia asintótica en la prueba de hipótesis, se supera el 20% de casillas con frecuencia esperada menor a 5. Por lo tanto, podemos usar el método de Montecarlo que es una prueba de significancia no asintótica.

Para usar el programa SPSS, la ruta es la siguiente:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Exacta

Los resultados se muestran a continuación:

Tabla 12.23 Pruebas de chi-cuadrado.

	Valor	df	Significación asintótica (bilateral)	Sig. Monte Carlo (bilateral)
Chi-cuadrado de Pearson	2.123[a]	2	0.346	0.368[b]
Razón de verosimilitud	2.113	2	0.348	0.368[b]
Prueba exacta de Fisher	2.070			0.368[b]
Asoc. lineal por lineal	0.838[c]	1	0.360	0.378[b]
N de casos válidos	34			

a. 2 casillas (33.3%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3.29.
b. Se basa en 10000 tablas de muestras con una semilla de inicio 2000000.
c. El estadístico estandarizado es 0.916.

En la tabla 12.23 se muestra que tanto la significancia asintótica como la significancia de Montecarlo, 0.346 y 0.368 respectivamente (aunque solo nos interesa la significancia de Monte Carlo), son mayores que 0.05. entonces, no se rechaza H_0.

¿Qué tipo de relación existe?

La prueba de hipótesis nos revela que no existe relación significativa. Por lo tanto, no tiene mucho sentido determinar el tipo de relación entre las variables. Sin embargo, podemos establecer algunas conclusiones para la muestra.

Tabla 12.24 Frecuencias observadas.

	Si	No	Total
Niño	8	5	13
Adolescente	9	4	13
Adulto	3	5	8
Total	20	14	34

En la tabla 12.24 se observa que en el grupo de integrantes del colegio que si consumen animes, la mayoría son niños y adolescentes; aunque no existe una diferencia significativa entre ambos grupos. En el grupo de las personas que no consumen animes vemos que la cantidad es similar para cada grupo. Esta falta de diferencias marcadas genera la ausencia de una relación significativa entre las variables.

¿Cuál es el grado de asociación?

Como se explicó en los párrafos anteriores, si la prueba de hipótesis revela que no existe relación significativa, no tiene mucho sentido determinar el grado de relación entre las variables. Sin embargo, si habría que calcular el grado se puede usar el coeficiente de contingencia ya que la matriz de contingencia es de 3×2.

Reemplazamos $\chi^2 = 2.123$ (el valor 2.16, que habíamos calculado, no es exacto porque realizamos aproximaciones con los decimales) y $n = 34$ en la expresión:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{2.123}{2.123 + 34}} = 0.242$$

Luego calculamos el valor máximo de C. k es el menor número de filas o columnas. Como la tabla de contingencia tiene 3 filas y 2 columnas, $k = 2$.

$$Máx(c) = \sqrt{\frac{k-1}{k}} = \sqrt{\frac{2-1}{2}} = 0.707$$

Considerando que 0 es el mínimo valor y 0.707 el máximo valor. El coeficiente de contingencia calculado 0.242 está más cerca de 0. Por ello, se puede concluir que la asociación entre las variables es débil.

El programa SPSS brinda este estadístico siguiendo la ruta.

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos-Nominal

Tabla 12.25 Medidas simétricas.

		Valor	Significación aproximada
Nominal por Nominal	Coeficiente de contingencia	0.242	0.346
N de casos válidos		34	

El programa Minitab nos brinda la opción de prueba chi cuadrado siguiendo la ruta.

Estadísticas-Tablas-Prueba chi cuadrada para asociación-Estadísticas

Los resultados son los siguientes:

Tabla 12.26 Prueba de chi-cuadrada.

	Chi-cuadrada	GL	Valor p
Pearson	2.123	2	0.346
Relación de verosimilitud	2.113	2	0.348

2 celda(s) con conteos esperados menores que 5.

En la tabla 12.26 se verifican los resultados obtenidos con el desarrollo completo y con el uso de SPSS.

Ejemplo 3: en una investigación se desea determinar si existe relación entre el consumo de cereales andinos y el nivel de educación de los pobladores del distrito de Lince. Por ello, se toma una muestra representativa compuesta de 90 personas. Los resultados se observan en la siguiente tabla:

Tabla 12.27 Matriz de datos.

	Nivel Ed.	Consumo		Nivel Ed.	Consumo		Nivel Ed.	Consumo
1	Básico	Bajo	31	Básico	Alto	61	Superior	Alto
2	Básico	Bajo	32	Básico	Bajo	62	Superior	Bajo
3	Básico	Bajo	33	Básico	Bajo	63	Superior	Bajo
4	Básico	Alto	34	Básico	Bajo	64	Superior	Bajo
5	Básico	Alto	35	Básico	Bajo	65	Superior	Alto
6	Básico	Alto	36	Básico	Alto	66	Superior	Bajo
7	Básico	Alto	37	Superior	Alto	67	Superior	Bajo
8	Básico	Bajo	38	Superior	Alto	68	Superior	Alto
9	Básico	Alto	39	Superior	Alto	69	Superior	Bajo
10	Básico	Alto	40	Superior	Alto	70	Superior	Alto
11	Básico	Alto	41	Superior	Bajo	71	Superior	Alto
12	Básico	Bajo	42	Superior	Bajo	72	Superior	Alto
13	Básico	Bajo	43	Superior	Bajo	73	Superior	Alto
14	Básico	Bajo	44	Superior	Alto	74	Superior	Alto
15	Básico	Bajo	45	Superior	Alto	75	Superior	Bajo
16	Básico	Alto	46	Superior	Alto	76	Superior	Alto
17	Básico	Bajo	47	Superior	Alto	77	Superior	Bajo
18	Básico	Alto	48	Superior	Alto	78	Superior	Alto
19	Básico	Bajo	49	Superior	Bajo	79	Superior	Alto
20	Básico	Bajo	50	Superior	Alto	80	Superior	Alto
21	Básico	Bajo	51	Superior	Alto	81	Superior	Alto
22	Básico	Alto	52	Superior	Bajo	82	Superior	Bajo
23	Básico	Bajo	53	Superior	Alto	83	Superior	Alto
24	Básico	Bajo	54	Superior	Alto	84	Superior	Alto
25	Básico	Bajo	55	Superior	Alto	85	Superior	Alto
26	Básico	Alto	56	Superior	Alto	86	Superior	Bajo
27	Básico	Alto	57	Superior	Alto	87	Superior	Bajo
28	Básico	Bajo	58	Superior	Alto	88	Superior	Alto
29	Básico	Bajo	59	Superior	Alto	89	Superior	Bajo
30	Básico	Bajo	60	Superior	Alto	90	Superior	Alto

Título: Nivel educativo y consumo de cereales andinos en pobladores del distrito de Lince.

Pregunta: ¿Cómo se relaciona el nivel educativo y el consumo de cereales andinos en los pobladores del distrito de Lince?

Objetivo: conocer la relación entre el nivel educativo y el consumo de cereales andinos en los pobladores del distrito de Lince.

Hipótesis: el nivel educativo se relaciona con el consumo de cereales andinos en los pobladores del distrito de Lince.

Tipo de investigación: enfoque cuantitativo, alcance descriptivo relacional y de corte transversal.

Variable 1: nivel educativo.

Variable 2: consumo de cereales andinos.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar si relación entre las variables en estudio.

Solución:

Planteamos las hipótesis:

H_0: el nivel educativo no se relaciona significativamente con el consumo de cereales andinos en los pobladores del distrito de Lince.

H_1: el nivel educativo se relaciona significativamente con el consumo de cereales andinos en los pobladores del distrito de Lince.

Para este ejemplo usaremos solo el Programa SPSS. La ruta ya se mostró en los ejercicios anteriores:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos

Los resultados se muestran a continuación:

Tabla 12.28 Tabla cruzada Nivel educativo*Consumo de cereal.

		Consumo de cereal		
		Bajo	Alto	Total
Nivel educativo	Básico	22	14	36
	Superior	17	37	54
Total		39	51	90

En la tabla 12.28 se puede apreciar que del grupo de pobladores que tienen un consumo bajo, 22 tienen un nivel educativo básico y 17 tienen un nivel educativo superior, esta diferencia no es muy grande. Pero, cuando observamos el grupo de las personas que tienen un consumo alto de cereales andinos, la diferencia es muy grande entre las personas que tienen nivel educativo básico y las que tienen un nivel educativo superior.

Tabla 12.29 Pruebas de chi-cuadrado.

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	7.722[a]	1	0.005
Corrección de continuidad[b]	6.563	1	0.010
Razón de verosimilitud	7.775	1	0.005
Prueba exacta de Fisher			
Asociación lineal por lineal	7.637	1	0.006
N de casos válidos	90		

a. 0 casillas (0.0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 15.60.

b. Sólo se ha calculado para una tabla 2x2

En las notas de tabla 12.29 se observa que se verifican las condiciones para utilizar la significancia asintótica. Estamos trabajando con un nivel de confianza de 95% lo que implica que la significancia es 0.05. Observamos que la significancia asintótica bilateral es igual a 0.005 y este valor es menor a 0.05. Por lo tanto, se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que el nivel educativo se relaciona significativamente con el consumo de cereales andinos en los pobladores del distrito de Lince

¿Qué tipo de relación existe?

Para interpretar la relación entre las variables analizadas podemos observar que, en el grupo de personas con nivel educativo básico la mayoría tiene un consumo bajo de cereales andinos y en el grupo de personas con nivel educativo alto la mayoría de personas tiene un consumo alto de cereales andinos. Este análisis nos permite establecer que a un mayor nivel educativo se le asocia un alto consumo de cereales andinos.

¿Cuál es el grado de asociación?

Tenemos una tabla de contingencia 2×2 (2 filas y 2 columnas) por ello, podemos utilizar la V de Cramer. Según SPSS:

Tabla 12.30 Medidas simétricas.

		Valor	Significación aproximada
Nominal por	Phi	0.293	0.005
Nominal	V de Cramer	0.293	0.005
N de casos válidos		90	

En la tabla 12.30 se observa que le valor de la V de Cramer es 0.293. este valor se encuentra en el intervalo:

$$0.2 < V \leq 0.6$$

Por lo tanto, el grado de asociación entre las variables es moderado.

Capítulo 13: Prueba de Fisher

También llamada prueba exacta de Fisher. Al igual que la prueba de hipótesis de chi cuadrado, la prueba de Fisher permite establecer si existe asociación entre variables categóricas, cuando las tablas de contingencia tienen 2 filas y dos columnas como en la figura 13.1. Esto quiere decir que las dos variables categóricas deben tener solo dos niveles cada una.

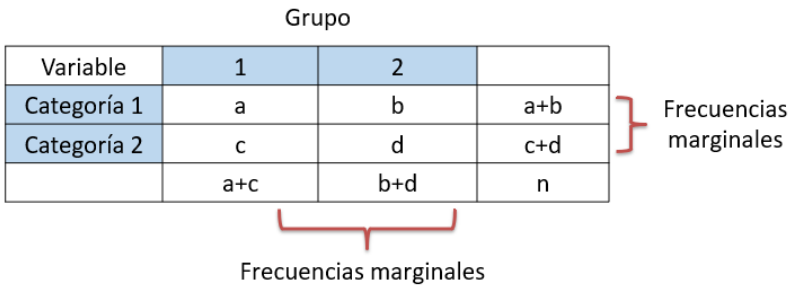


Figura 13.1 Frecuencias en una tabla 2 × 2.

Esta prueba se utiliza cuando no se verifican las condiciones para poder utilizar la prueba chi cuadrado. Se utiliza, sobre todo cuando el tamaño de muestra es pequeño. También es útil para comparar variables de tipo ordinal.

Esta prueba se basa en las probabilidades que se obtienen al asignar diferentes frecuencias en la tabla, pero manteniendo las frecuencias marginales constantes.

No usaremos un estadístico de prueba sino la siguiente probabilidad:

$$p = \frac{(a+c)! \times (b+d)! \times (a+b)! \times (c+d)!}{n! \times a! \times b! \times c! \times d!}$$

Esta probabilidad p es la probabilidad de que los valores de las frecuencias observadas hayan sido asignados al azar considerando las frecuencias marginales constantes (Siegel y Castellan, 1995)

Para la prueba de hipótesis bilateral planteamos:

H_0: No existe asociación entre las variables.
H_1: Existe asociación entre las variables.

Decisión estadística:

Si $p < 0.05$, entonces se rechaza H_0.
Si $p \geq 0.05$, entonces no se rechaza H_0.

Ejemplo 1: mediante un experimento se desea conocer si determinado tipo de entrenamiento de un gimnasio tiene efecto sobre el consumo de suplementos deportivos. Para ello, se toma dos muestras de deportistas del gimnasio X con similares características, de modo que, un grupo recibe el entrenamiento tipo Regular y otro grupo recibe el entrenamiento tipo Competición. Luego se les toma una encuesta para determinar su nivel de consumo de suplementos deportivos después del entrenamiento. Los resultados son los siguientes.

Tabla 13.1 Matriz de datos.

Tipo de entrenamiento	Nivel de consumo de suplementos deportivos
Regular	Bajo
Competición	Alto
Competición	Alto
Regular	Bajo
Regular	Bajo
Competición	Alto
Regular	Bajo
Competición	Alto
Regular	Bajo
Regular	Alto
Competición	Alto
Competición	Bajo
Competición	Alto
Regular	Bajo
Regular	Bajo

Competición	Bajo
Competición	Bajo
Regular	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Regular	Bajo

Título: tipo de entrenamiento y nivel de consumo de suplementos deportivos en deportistas del gimnasio X.

Pregunta: ¿Qué efecto tiene el tipo de entrenamiento sobre el nivel de consumo de suplementos deportivos en deportistas del gimnasio X?

Objetivo: conocer el efecto que tiene el tipo de entrenamiento sobre el nivel de consumo de suplementos deportivos en deportistas del gimnasio X.

Hipótesis: los deportistas del gimnasio X que practican el entrenamiento regular tienen un nivel bajo de consumo de suplementos deportivos.

Tipo de investigación: enfoque cuantitativo, experimental y de corte transversal.

Variable x: tipo de entrenamiento.

Variable y: nivel consumo de suplementos deportivos.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la relación significativa entre las variables en estudio.

Solución: Solo se cuenta con 22 datos. Por lo tanto, no se puede usar la prueba chi cuadrado con su significancia asintótica. Usaremos la prueba exacta de Fisher.

Planteamos las hipótesis para una prueba unilateral:

H_0: los deportistas del gimnasio X que practican el entrenamiento regular no tienen un nivel bajo de consumo de suplementos deportivos.

H_1: los deportistas del gimnasio X que practican el entrenamiento regular tienen un nivel bajo de consumo de suplementos deportivos.

Ordenamos los datos para generar la tabla de contingencia.

Tabla 13.2 Clasificación de las respuestas.

Tipo de entrenamiento	Nivel de consumo de suplementos deportivos
Regular	Alto
Regular	Alto
Regular	Bajo
Regular	Bajo
Regular	Bajo
Regular	Bajo
Regular	Bajo
Regular	Bajo
Regular	Bajo
Regular	Bajo
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Alto
Competición	Bajo
Competición	Bajo
Competición	Bajo
Competición	Bajo

Luego construimos la tabla de contingencia o tabla cruzada.

Tabla 13.3 Frecuencias observadas.

		Tipo de entrenamiento		
		Regular	Competición	Total
Nivel de consumo	Bajo	8	4	12
	Alto	2	8	10
	Total	10	12	22

Calculamos la probabilidad respectiva:

$$p_1 = \frac{10! \times 12! \times 12! \times 10!}{22! \times 8! \times 4! \times 2! \times 8!} = 0.034447$$

Luego buscamos las diferencias extremas. Para esto aumentamos el valor de la frecuencia de deportistas que tienen un entrenamiento regular y un nivel de consumo bajo en una unidad, pero mantenemos constantes las frecuencias marginales.

Tabla 13.4 Primera variación en las frecuencias observadas.

	Regular	Competición	Total
Bajo	9	3	12
Alto	1	9	10
Total	10	12	22

Calculamos la probabilidad respectiva:

$$p_2 = \frac{10! \times 12! \times 12! \times 10!}{22! \times 9! \times 3! \times 1! \times 9!} = 0.003402$$

Nuevamente aumentamos el valor de la frecuencia de deportistas que tienen un entrenamiento regular y un nivel de consumo bajo en una unidad, pero mantenemos constantes las frecuencias marginales.

Tabla 13.5 Segunda variación en las frecuencias observadas.

	Regular	Competición	Total
Bajo	10	2	12
Alto	0	10	10
Total	10	12	22

Calculamos la probabilidad respectiva:

$$p_3 = \frac{10! \times 12! \times 12! \times 10!}{22! \times 10! \times 2! \times 0! \times 10!} = 0.000102$$

Luego sumamos las probabilidades:

$$p_1 + p_2 + p_3 = 0.034447 + 0.003402 + 0.000102 = 0.037951$$

Entonces, el valor de p es aproximadamente:

$$p = 0.038$$

0.038 es menor que 0.05. Entonces se rechaza H_0.

En conclusión, existe evidencia estadística para afirmar que los deportistas del gimnasio X que practican el entrenamiento regular tienen un nivel bajo de consumo de suplementos deportivos.

Lo que se ha demostrado con esta prueba también se explica señalando que el bajo nivel de consumo de suplementos deportivos en los deportistas que tienen un entrenamiento regular no se debe al azar.

Para realizar la prueba de Fisher utilizando SPSS, seguimos la siguiente ruta:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos

Los resultados se aprecian en la siguiente tabla:

Tabla 13.6 Pruebas de chi-cuadrado.

	Valor	df	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	4.791[a]	1	0.029		
Corrección de continuidad[b]	3.094	1	0.079		
Razón de verosimilitud	5.032	1	0.025		
Prueba exacta de Fisher				0.043	0.038
Asociación lineal por lineal	4.573	1	0.032		
N de casos válidos	22				

a. 1 casillas (25.0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 4.55.

b. Sólo se ha calculado para una tabla 2x2

En la tabla 13.6 se observa la significancia exacta unilateral $p = 0.038$. Como este valor p es menor que 0.05, se rechaza H_0.

Ejemplo 2: un investigador quiere determinar si existe asociación entre el trabajo bajo presión y la productividad de los empleados de la empresa X. Por ello, se toma una muestra de 20 empleados, se analiza su ritmo de trabajo y su desempeño. Finalmente se obtienen los siguientes resultados:

Tabla 13.7 Matriz de datos.

Trabajo bajo presión	Productividad
No	Baja
Si	Alta
No	Alta
No	Baja
No	Baja
No	Alta
Si	Alta
Si	Alta
Si	Alta
No	Alta
Si	Alta
Si	Baja
No	Alta
No	Baja
Si	Alta
Si	Baja
Si	Baja
No	Alta
Si	Alta
Si	Alta

Título: trabajo bajo presión y productividad en los empleados de la empresa X.

Pregunta: ¿El trabajo bajo presión se relaciona con la productividad de los empleados de la empresa X?

Objetivo: conocer si existe relación entre el trabajo bajo presión y la productividad en los empleados de la empresa X.

Hipótesis: el trabajo bajo presión se relaciona con la productividad de los empleados de la empresa X.

Tipo de investigación: enfoque cuantitativo, descriptivo relacional y de corte transversal.

Variable 1: ritmo de trabajo.

Variable 2: productividad del empleado.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la relación significativa entre las variables en estudio.

Solución: Solo se cuenta con 20 datos. Por lo tanto, no se puede usar la prueba chi cuadrado con su significancia asintótica. Usaremos la prueba exacta de Fisher.

Planteamos las hipótesis para una prueba bilateral:

H_0: el trabajo bajo presión no se relaciona significativamente con la productividad de los empleados de la empresa X.

H_1: el trabajo bajo presión se relaciona significativamente con la productividad de los empleados de la empresa X.

Para realizar la prueba de Fisher utilizando SPSS, seguimos la siguiente ruta:

Analizar-Estadísticos descriptivos-Tablas cruzadas-Estadísticos

Los resultados son los siguientes:

Tabla 13.8 Tabla cruzada Trabajo bajo presión*Productividad.

		Productividad		
		Baja	Alta	Total
Trabajo bajo presión	Si	3	8	11
	No	4	5	9
Total		7	13	20

En la tabla 13.8 se puede observar que es frecuente encontrar trabajadores que laboran bajo presión y tienen una productividad alta. Pero dicha frecuencia no es muy diferente a la que corresponde a los trabajadores que no trabajan bajo presión, pero tienen una productividad alta. Del mismo modo se puede observar que en grupo de trabajadores que tienen baja productividad, la cantidad de trabajadores que laboran bajo presión y los que no laboran bajo presión es casi la misma.

Tabla 13.9 Pruebas de chi-cuadrado.

	Valor	df	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	,642[a]	1	,423		
Corrección de continuidad[b]	,109	1	,742		
Razón de verosimilitud	,642	1	,423		
Prueba exacta de Fisher				,642	,370
Asociación lineal por lineal	,610	1	,435		
N de casos válidos	20				

a. 2 casillas (50,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,15.

b. Sólo se ha calculado para una tabla 2x2

El nivel de confianza para este ejemplo es de 95%, esto implica que la significancia es de 0.05. En la tabla 13.9 se observa que la significancia exacta bilateral es 0.642 y este valor es mayor que 0.05. Entonces no se rechaza H_0.

En conclusión, no existe evidencia estadística para afirmar que el trabajo bajo presión se relaciona significativamente con la productividad de los empleados de la empresa X.

Si se hubiese rechazado la hipótesis nula nos podríamos preguntar ¿y cómo es la relación entre el trabajo bajo presión y la productividad? Pero, en este caso, no se estableció una relación significativa, así que no es necesario que busquemos una respuesta a la pregunta planteada. Sin

embargo, es necesario identificar la relación, especialmente cuando se utiliza una prueba bilateral.

Para realizar la prueba de Fisher también podemos utilizar el programa Minitab, siguiendo la ruta:

Estadísticas-Tablas-Tabulación cruzada y Chi cuadrada-Otras estadísticas

Los resultados son los siguientes:

Tabla 13.10 Prueba de chi-cuadrada.

	Chi-cuadrada	GL	Valor p
Pearson	0.642	1	0.423
Relación de verosimilitud	0.642	1	0.423

2 celda(s) con conteos esperados menores que 5.

La inclusión de la prueba de chi cuadrado, como se muestra en la tabla 13.10, es opcional. En este ejemplo solo se ha considerado para que el lector verifique los resultados obtenidos en SPSS.

Tabla 13.11 Prueba exacta de Fisher.

Valor p
0.642415

Minitab nos brinda una tabla muy simple para mostrar el valor de p en la prueba exacta de Fisher. La tabla 13.11 muestra que p=0.642. Este valor mayor a 0.05 implica que no se rechaza H_0.

Capítulo 14: Prueba de Kruskal-Wallis

Es una prueba no paramétrica equivalente al ANOVA para muestras independientes. Triola (2004) afirma que esta prueba consiste en verificar si tres o más muestras independientes provienen de poblaciones iguales.

Además, esta prueba se puede utilizar cuando los datos de la variable dependiente no provienen de poblaciones con distribución normal. En general, se puede utilizar cuando la variable dependente es cuantitativa.

Así como las pruebas no paramétricas de Mann-Whitney y Wilcoxon, la prueba de Kruskal-Wallis se fundamenta en la comparación de las medianas, no de las medias.

Ejemplo 1: en una investigación se quiere conocer la satisfacción laboral de los docentes del colegio X en las sedes A, B y C. Por ello, se toma una muestra representativa de docentes y se les toma una encuesta. Esta encuesta se compone de 10 preguntas y el máximo puntaje que se puede obtener es 20. Los resultados fueron los siguientes:

Satisfacción laboral		
Sede A	Sede B	Sede C
14	14	13
14	16	14
13	16	14
12	16	13
12	19	16
15	15	16
15	16	14
16	17	14
15	17	13

14	15	14

Título: satisfacción laboral en docentes del colegio X.

Pregunta: ¿Existe diferencia entre la satisfacción laboral de los docentes de las sedes A, B y C del colegio X?

Objetivo: conocer si existe diferencia entre la satisfacción laboral de los docentes de las sedes A, B y C del colegio X.

Hipótesis: existe diferencia entre la satisfacción laboral de los docentes de las sedes A, B y C del colegio X.

Tipo de investigación: enfoque cuantitativo, descriptivo y de corte transversal.

Variable 1: sede.

Variable 2: satisfacción laboral.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar la existencia de diferencias significativas entre los grupos de docentes.

Solución:

Planteamos las hipótesis para una prueba bilateral:

H_0: la satisfacción laboral es igual entre docentes de las sedes A, B y C del colegio X.

H_1: la satisfacción laboral es diferente entre docentes de las sedes A, B y C del colegio X.

Para realizar la prueba de Kruskal-Wallis utilizando SPSS, seguimos la siguiente ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-k muestras independientes

Los resultados son los siguientes.

Tabla 14.1 Rangos

	Sede	N	Rango promedio
Satisfacción	Sede A	10	11.85
	Sede B	10	23.00
	Sede C	10	11.65
	Total	30	

En la tabla 14.1 se observa que el mayor rango promedio le corresponde a la sede B. De otro lado, el rango promedio de la sede A es muy similar al rango promedio de la sede C. La gran diferencia entre los rangos promedio indica que la satisfacción laboral no es la misma en todas las sedes.

Tabla 14.2 Estadísticos de prueba[a,b]

H de Kruskal-Wallis	11.418
gl	2
Sig. asintótica	0.003

a. Prueba de Kruskal Wallis

b. Variable de agrupación: Sede

Para completar la prueba de hipótesis observamos la tabla 14.2, en ella se observa la significancia calculada p=0.003. El valor de p es menor a 0.05, entonces se rechaza la hipótesis nula.

En conclusión, existe evidencia estadística para afirmar que la satisfacción laboral es diferente entre docentes de las sedes A, B y C del colegio X.

Luego de encontrar diferencias significativas, es necesario precisar entre que sedes se encuentran dichas diferencias de satisfacción laboral. Por ello, se elaboran pruebas posteriores para hacer una comparación por parejas para la satisfacción laboral entre las sedes Ay B, las sedes B y C, y las sedes A y C.

Ejemplo 2: En una fábrica panificadora se quiere conocer las diferencias entre la firmeza de 3 tipos de pan complementados con harina de Tipo A, Tipo B y Tipo C. Para ello se realiza una prueba de compresión con una muestra de panes de cada tipo. Los resultados son los siguientes:

Firmeza del pan (N)		
Tipo A	Tipo B	Tipo C
3.8	4.2	4.8
3.5	3.6	4.2
4.1	3.8	4.2
4.6	3.9	3.8
4.8	4.5	3.8
4,0	3.8	3.8
3.8	4.9	3.9
3.9	4.2	4.3
4.5	4.0	4.3

Título: Características del pan elaborado con harinas complementarias.

Pregunta: ¿Existe diferencia entre la firmeza del pan elaborado con harina de Tipo A, harina de Tipo B y harina de Tipo C?

Objetivo: conocer si existe diferencia entre la firmeza del pan elaborado con harina de Tipo A, harina de Tipo B y harina de Tipo C.

Hipótesis: existe diferencia entre la firmeza del pan elaborado con harina de Tipo A, harina de Tipo B y harina de Tipo C.

Tipo de investigación: enfoque cuantitativo, experimental y de corte transversal.

Variable 1: tipo de harina.

Variable 2: firmeza del pan.

Utilizando un nivel de confianza de 95% realiza la prueba de hipótesis para verificar diferencias significativas entre los grupos a comparar.

Solución:

Planteamos las hipótesis para una prueba bilateral:

H_0: la firmeza es igual entre los panes complementados con harina de Tipo A, los panes complementados con harina de Tipo B, y los panes complementados con harina de Tipo C.

H_1: la firmeza es diferente entre los panes complementados con harina de Tipo A, los panes complementados con harina de Tipo B, y los panes complementados con harina de Tipo C.

Para realizar la prueba de Kruskal-Wallis utilizando SPSS, seguimos la siguiente ruta:

Analizar-Pruebas no paramétricas-Cuadro de diálogos antiguos-k muestras independientes

Los resultados son los siguientes.

Tabla 14.3 Rangos

	Tipo	N	Rango promedio
Firmeza	Tipo A	9	13.83
	Tipo B	9	13.67
	Tipo C	9	14.50
	Total	27	

En la tabla 14.3 se observa que los rangos para los tres tipos de productos son muy cercanos. Es decir, no se evidencia una diferencia significativa entre la firmeza del pan elaborado con los diferentes tipos de harina.

Tabla 14.4 Estadísticos de prueba[a,b]

H de Kruskal-Wallis	0.057
gl	2
Sig. asintótica	0.972

a. Prueba de Kruskal Wallis

b. Variable de agrupación: Tipo

En la tabla 14.4 se observa la significancia calculada p=0.972. El valor de p no es menor a 0.05, entonces, no se rechaza la hipótesis nula.

En conclusión, no existe evidencia estadística para afirmar que la firmeza sea diferente entre los panes complementados con harina de Tipo A, los panes complementados con harina de Tipo B, y los panes complementados con harina de Tipo C.

También se puede decir que El tipo de harina con que se complementa los panes no tiene efecto sobre la firmeza del pan elaborado.

Además, se debe resaltar que no es necesario elaborar las pruebas posteriores dado que no se han encontrado diferencias significativas luego de realizar la prueba de hipótesis.

APÉNDICE

Tabla de valores críticos de la distribución t de Student

La tabla de valores críticos de esta distribución se encuentra en el libro *Estadística matemática con aplicaciones* de Wackerly, Mendenhall y Scheaffer (2010).

También se puede encontrar en:
https://es.scribd.com/document/379718283/Distribucion-T-Student#

Tabla de valores críticos de la U de Mann-Whitney

Disponible en la página web de libretext (2022, 3 de octubre):
https://espanol.libretexts.org/Ciencias_Sociales/Ciencias_Sociales/Antropologia/Ling%C3%BC%C3%ADstica/Ling%C3%BC%C3%ADstica_del_corpus%3A_una_gu%C3%ADa_para_la_metodolog%C3%ADa_(Stefanowitsch)/14%3A_Tablas_estad%C3%ADsticas/14.3%3A_Valores_cr%C3%ADticos_para_el_texto_Mann-Whitney-Text

Tabla de valores críticos para la prueba de Wilcoxon

Disponible en la página web de statologos:
https://statologos.com/prueba-de-rango-con-signo-de-wilcoxon/

Tabla de valores críticos de la distribución chi cuadrado

La tabla de valores críticos de esta distribución se encuentra en el libro *Estadística matemática con aplicaciones* de Wackerly, Mendenhall y Scheaffer (2010).

También se puede encontrar en:
https://www.medwave.cl/series/MBE04/5266.html

Tabla de valores críticos de la distribución F

Se puede encontrar en el libro *Probabilidad y estadística para ingeniería y ciencias* de Walpole, Myers, Myers y Ye (2007).

También se puede encontrar en:
https://www.coursehero.com/file/38051518/tabla-f-005pdf/

Puntos críticos para el estadístico de Durbin-Watson

Estadístico de Durbin-Watson - Puntos críticos de d_L y d_u al nivel de significación del 5 %
k^* corresponde al número de regresores del modelo excluido el término independiente (es decir, $k^* = k - 1$)

	$k^*=1$		$k^*=2$		$k^*=3$		$k^*=4$		$k^*=5$		$k^*=6$	
n	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
6	0.610	1.400										
7	0.700	1.356	0.467	1.896								
8	0.763	1.332	0.559	1.777	0.368	2.287						
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588				
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822		
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645	0.203	3.005
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506	0.268	2.832
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.445	2.390	0.328	2.692
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.472
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.257
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.692	2.162
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.732	2.124
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.012
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.958
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900

Tomado de Salmeron, R. (s.f.). *Tabla del Estadístico de Durbin-Watson*. Universidad de Granada. Disponible en: https://www.ugr.es/~romansg/material/WebEco/01-comunes/dw.pdf

Referencias bibliográficas

Anderson, D., Sweeney, D. y Williams, T. (2012). *Estadística para negocios y economía*.11° ed. México D.F. Cengage Learning Editores.

Arranz, J.M. y Zamora, M.M. (2002): "*Autocorrelación I*" 5campus.com. Sistemas Informativos Contables. https://www.ciberconta.unizar.es/leccion/autocorrelacion/analisis%20de%20autocorrelacion.PDF. 8 de febrero de 2023.

Blair, R.C. y Taylor, R. A. (2008). *Bioestadística*. México. Pearson Educación de México.

Cárdenas, C., Haua, K., Suverza, A., y Perichart, O. (2005). *Mediciones antropométricas en el neonato*. *Boletín médico del Hospital Infantil de México*, *62*(3), 214-224. Recuperado en 14 de enero de 2023, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1665-11462005000300009&lng=es&tlng=es

Gorgas, J., Cardiel, N. y Zamorano, J. (2011). *Estadística básica para estudiantes de ciencias*. Facultad de Ciencias Físicas de la Universidad Complutense de Madrid, España.

Hernández, R., Fernández, C. y Baptista, M. (2010). *Metodología de la Investigación*. 5° ed. México D.F. Editorial McGraw Hill.

Hopkins, K., Hopkins, B. y Glass, G. (1997). *Estadística básica para las ciencias sociales y del comportamiento*. México. Prentice-Hall Hispanoamericana.

IBM Corp. (2017). *IBM SPSS Statistics for Windows*, Version 25.0. Armonk, NY: IBM Corp.

IBM (2023, 3 de enero) V de Cramer. Cognos analytics. https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=terms-cramrs-v

INEI (2016, 20 de junio). *El 35,5% de la población peruana de 15 y más años de edad padece de sobrepeso*. Encuesta Demográfica y de Salud Familiar – ENDES 2015. https://m.inei.gob.pe/prensa/noticias/el-355-de-la-poblacion-peruana-de-15-y-mas-anos-de-edad-padece-de-sobrepeso-9161/#:~:text=A%20nivel%20nacional%2C%20el%2035,35%2C2%25%20de%20hombres.

Infobae (2022, 4 de agosto). *Peruanos beben 46 litros de cerveza al año: Lima, Arequipa y Trujillo registran más compras de esta bebida.* https://www.infobae.com/america/peru/2022/08/04/peruanos-beben-46-litros-de-cerveza-al-ano-lima-arequipa-y-trujillo-registran-mas-compras-de-esta-bebida/

Juárez, F., Villatoro, J., López, E. (2002), *Apuntes de Estadística Inferencial*. México, D.F.: Instituto Nacional de Psiquiatría Ramón de la Fuente.

La República (2021, 11 de abril). *Smartphone: uso de apps supera las cinco horas diarias en algunas regiones.* https://larepublica.pe/tecnologia/2021/04/11/smartphone-uso-de-apps-supera-las-cinco-horas-diarias-en-algunas-regiones

Lind, D., Marchal, W. y Wathen, S. (2012). *Estadística aplicada a los negocios y la economía*. 15° ed. México D.F. Editorial McGraw Hill.

Maya, L. y Pliego, J. (1999). *Fundamentos de inferencia estadística*. España Editorial Alfa Centauro.

Mendoza, M. (2020, 23 de enero). *Plataformas como Netflix son usadas por el 33% de peruanos, según Concortv*. El Comercio. https://elcomercio.pe/economia/dia-1/plataformas-como-netflix-son-usadas-por-el-33-de-peruanos-segun-concortv-internet-noticia/#:~:text=No%20obstante%2C%20destaca%20el%20incremento,%2C%20resalta%20Trujillo%20con%2027%25.

MINAGRI (2019, 18 de octubre). *Minagri impulsa producción y consumo de carne para luchar contra la anemia infantil*. https://www.gob.pe/institucion/midagri/noticias/52971-minagri-impulsa-produccion-y-consumo-de-carne-para-luchar-contra-la-anemia-infantil

Minitab (2019) Versión 19.1.1. Disponible en: https://www.minitab.com

MINSA (2019. 3 de junio). *Solo uno de tres peruanos realiza actividad física*. https://www.gob.pe/institucion/minsa/noticias/29074-solo-uno-de-tres-peruanos-realiza-actividad-fisica

Molina, M. (2017) ¿Qué significa realmente el valor de p? Revista Pediatría Atención Primaria. 2017;19:377-81. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322017000500014#:~:text=El%20valor%20de%20p%20nos%20indica%20la%20importancia%20del%20resultado,cl%C3%ADnico%20la%20establece%20el%20investigador.

Montgomery. D. Peck, E. y Vining, G. (2006). *Introducción al análisis de regresión lineal*. 4° ed. México. Compañía Editorial Continental.

Morán, G. y Alvarado, D. (2010). *Métodos de investigación*. México. Pearson educación.

Moya, R. y Saravia, G. (2004). *Probabilidades e inferencia estadística*. Lima. 2da Edición. Editorial San Marcos.

OSIPTEL (2022, 12 de julio). *OSIPTEL: más de 8 millones de hogares peruanos tienen acceso a internet*. https://www.osiptel.gob.pe/portal-del-usuario/noticias/osiptel-mas-de-8-millones-de-hogares-peruanos-tienen-acceso-a-internet/#:~:text=Al%20cierre%20de%202021%2C%20el,Privada%20en%20Telecomunicaciones%20(OSIPTEL).

Quezada, C., Apolo, N. y Delgado K. (2018). *Investigación científica*. En Neill, D. y Cortez, L. (Ed). *Procesos y Fundamentos de la*

Investigación Científica (pp. 12-37). Ecuador. Editorial UTMACH.

Salmeron, R. (s.f.). *Tabla del Estadístico de Durbin-Watson*. Métodos cuantitativos para la economía y la empresa. Universidad de Granada. https://www.ugr.es/~romansg/material/WebEco/01-comunes/dw.pdf

Siegel, S. y Castellan, J. (1995). *Estadística no paramétrica; Aplicada a las ciencias de la conducta*. 4° ed. México D.F. Editorial Trillas.

Spiegel, M. y Stephens, L. (2009). *Estadística*. 4° ed. México D.F. Editorial McGraw Hill.

Tamayo, M. (2003). *El proceso de la investigación científica*. *Incluye evaluación y administración de proyectos de investigación.* 4° ed. México D.F. Limusa Editores.

Triola, M. (2004). *Estadística*. 9° ed. México. Pearson Educación.

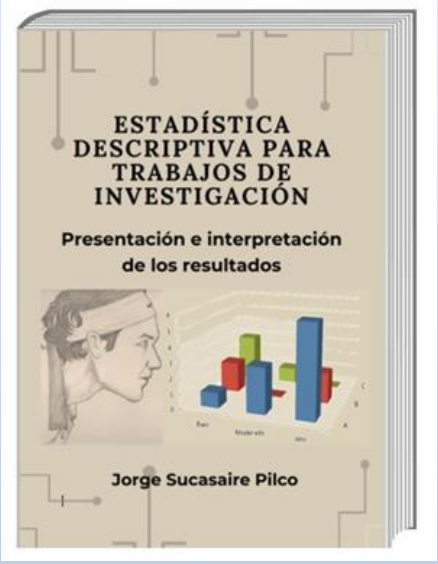
UCSM (2021, mayo). *Al año cada peruano consume 90 kilos de papa como parte de su dieta alimenticia*. https://www.ucsm.edu.pe/al-ano-cada-peruano-consume-90-kilos-de-papa-como-parte-de-su-dieta-alimenticia/

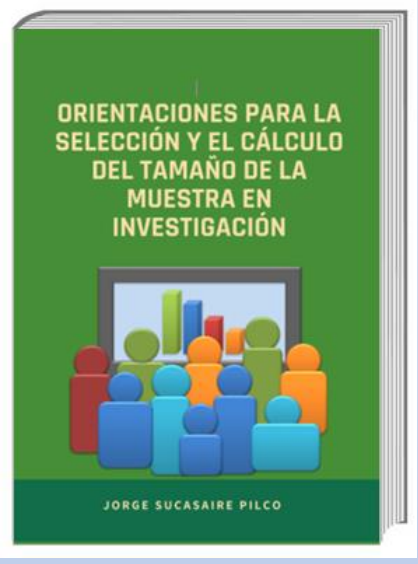
Walpole, R., Myers, R., Myers, S. y Ye, K. (2007). *Probabilidad y estadística para ingeniería y ciencias*. 4° edición. México. Pearson Educación.

Wackerly, D., Mendenhall, W. y Scheaffer, R. (2010). *Estadística matemática con aplicaciones*. 7° ed. México D.F. Cengage Learning Editores.

Zavaleta, N. (2017, 7 de diciembre) *Anemia infantil: retos y oportunidades al 2021*. Centro Nacional de Alimentación y Nutrición, Instituto Nacional de Salud. https://rpmesp.ins.gob.pe/rpmesp/article/view/3281/2906#:~:text=En%20el%20Per%C3%BA%2C%20la%20anemia,esfuerzos%20del%20gobierno%20por%20reducirla.

Otras obras del autor





ISBN: 978-612-00-8393-2