



Guía para el análisis de datos del Proyecto terminal



División de Ciencias de la Salud, Biológicas y Ambientales



Guía para el análisis de datos del Proyecto terminal

Proyecto terminal II





Contenido

Introducción	3
Tu Proyecto y la investigación.....	3
Elección de prueba estadística	6
Variables, muestreo y tamaño de muestra	9
Análisis de datos para variables cualitativas	16
Medidas de tendencia central, de dispersión y distribución.....	24
Prueba U Mann-Whitney	33
Prueba Z (muestras pareadas)	37
Prueba t (muestras pareadas).....	39
Prueba Wilcoxon (prueba de rango con signo).....	42
Prueba Z (muestras independientes)	44
Prueba t (muestras independientes)	47
Regresión lineal	49
Comparación de dos rectas	54
Análisis de varianza de un factor o de una vía.....	60
Símbolos y operadores matemáticos.....	63
Referencias.....	64



Introducción

Has comenzado la realización de la segunda parte de tu proyecto terminal, en la primera parte: Proyecto terminal I, construiste tu protocolo, e investigaste tu marco teórico, planteaste los objetivos que persiguen tu proyecto, estableciste la metodología con la cual obtendrás resultados que satisfacen a los objetivos que planteaste. Ahora en el proyecto terminal II analizarás los resultados, escribirás la versión final de tu documento y harás la presentación que expondrás sobre tu Proyecto terminal.

En la presentación que hagas de tu proyecto terminal deberás mostrar el análisis de resultados, las conclusiones y propuestas a las que llegarás con base a tales resultados, por lo cual este documento tiene como objetivo de ayudarte a que hagas un análisis de los datos de la forma más satisfactoria sobre todo tu trabajo del proyecto terminal. Esto se verá reflejado en tu documento final, así como en la presentación audio visual. Es importante señalar que cada uno de los pasos a seguir deben ser revisados por tu docente en línea y tu asesor externo, tales pasos deben llevarse acabo de acuerdo a un orden coherente y con el cronograma establecido al inicio del Proyecto terminal I y II.

Tu Proyecto y la investigación

En el semestre anterior elegiste la empresa o institución de salud donde se aprobó la estancia para la realización de tu proyecto terminal, además presentaste una propuesta del título de tu proyecto terminal, el planteamiento del problema, el marco teórico (junto con antecedentes), la justificación, los objetivos (general y específicos) y la metodología, es decir el contenido teórico de tu Proyecto terminal. Ahora en el Proyecto Terminal II deberás presentar los resultados con su correspondiente análisis y con base en ellos dar una conclusión junto con las propuestas pertinentes de acuerdo con los resultados y a lo que la unidad de atención a la salud demanda. El primer paso que diste en el Proyecto terminal I fue ubicar el tipo de investigación que estas llevando a cabo, para que en función de esto sepas los alcances y limitaciones que tendrás al analizar los resultados de tu proyecto terminal, esto puedes revisarlo con más detalle en la tabla 1, y constituye un repaso del contenido que viste en la asignatura de Fundamentos de Investigación, si quieres ahondar más en el tema se te sugiere revises el material de la asignatura antes mencionada.



Tabla 1. Tipos de Investigación de acuerdo con ciertos criterios

CRITERIO UTILIZADO	TIPO	SUBTIPO
Por el propósito o finalidades perseguidas	Básica	
	Aplicada	Exploratoria: Se realiza para conocer el tema que se abordará, lo que nos permita “familiarizarnos” con algo que hasta el momento desconocíamos
		Descriptiva: no va mucho más allá del nivel descriptivo; ya que consiste en plantear lo más relevante de un hecho o situación concreta.
		Confirmatoria: Este tipo de investigación proporciona principios generales de explicación.
Desde el punto de vista del método frente al objeto de estudio	Experimental	Laboratorio: Se ejerce el máximo control en escenarios no naturales, dificulta la validez externa.
		Campo: Se efectúan en escenarios naturales. disminuye la artificialidad, facilita la validez externa
		Naturales: Se produce un evento que se convierte en la variable independiente (v.i.).
	No experimental	Estudios de campo: Intensivos no interesa el muestreo probabilístico.
		Correlacionales: Proporciona predicciones
		Encuesta: Por lo general es un estudio extensivo representativo
		Observaciones naturales:
	Intermedia	Ex.post-facto: No proporciona explicaciones funcionales
		Cuasi-experimentales: No permiten el control propiamente dicho de la varianza externa.
Por la clase de medios utilizados en la		Documental: Estudia problemas con el propósito de ampliar y profundizar el conocimiento de su naturaleza, con apoyo, principalmente, en trabajos previos, información y datos divulgados por medios impresos, audiovisuales o electrónicos, y el investigador



obtención de datos:		desarrolla la capacidad reflexiva y crítica a través del análisis, interpretación y confrontación de la información regida.
		Investigación de Campo: Se efectúan en escenarios naturales. disminuye la artificialidad, facilita la validez externa
		Cuantitativa: Unifica y analiza datos numéricos de variables determinadas con antelación, estudia la relación entre las variables que han sido cuantificados.
		Cualitativa: Estudia la calidad de los elementos en un determinado problema, trata de analizar profundamente el problema en particular.
		Cualicuantitativa: es un tipo de investigación que mezcla características de las investigaciones cuantitativa y cualitativa.
		Explicativa: se orienta a establecer las causas que originan un fenómeno determinado. Se trata de un tipo de investigación cuantitativa que descubre el por qué y el para qué de un fenómeno
		Inferencia o predictiva: Predice la dirección futura de los eventos investigados, la predicción de situaciones se hacen a partir de estudios detallados del progreso de los eventos y la relación con el contexto.
Según el tiempo en que se efectúan.		Sincrónicas o Transversal: son aquellas que estudian fenómenos que se dan en un corto periodo
		Diacrónicas: Son aquellas que estudian fenómenos en un período largo con el objetivo de verificar los cambios que se pueden producir.
		Histórica: Se realiza cuando se desea estudiar desde una perspectiva histórica de la realidad, recurriendo a las fuentes primarias y secundarias para la reconstrucción de la misma
		Longitudinal: Consiste en estudiar y evaluar a los mismos elementos de la población por un período prolongado de tiempo, es la examinación de cambios producidos en el tiempo en una misma muestra.
		Dinámica: Es donde se controla la información, todo sobre un hecho desde cómo surgió hasta como acabo
		Estática: En ella no se admite ninguna variación



Una vez que determinaste que tipo de investigación va a ser la que realizarás en tu proyecto terminal, debiste establecer cuál sería tu población, de que tamaño sería tu muestra, cómo sería tu muestreo y con qué tipo de variables irías a trabajar. A manera de repaso se presentan en este documento estos aspectos. Si embargo, es importante revises el material que se te fue proporcionado en las asignaturas de Estadística básica y Seminario de Investigación.

Elección de prueba estadística

Una de las primeras decisiones que se deben de tomar en el análisis de datos y obtención de resultados es ¿Qué prueba estadística puedo realizar con mis datos? La respuesta depende principalmente al tipo de variables que tenemos, en este documento se ejemplificarán solo algunas de estas pruebas. Es importante e indispensable que revises los materiales de las asignaturas Estadística básica y Bioestadística.

A continuación, se presenta una guía sencilla con las pruebas estadísticas más frecuentemente empleadas en el manejo de datos, revísala para que elijas la prueba estadística que has de emplear de acuerdo con el tipo de variables que está utilizando en tu proyecto terminal (figura 1).

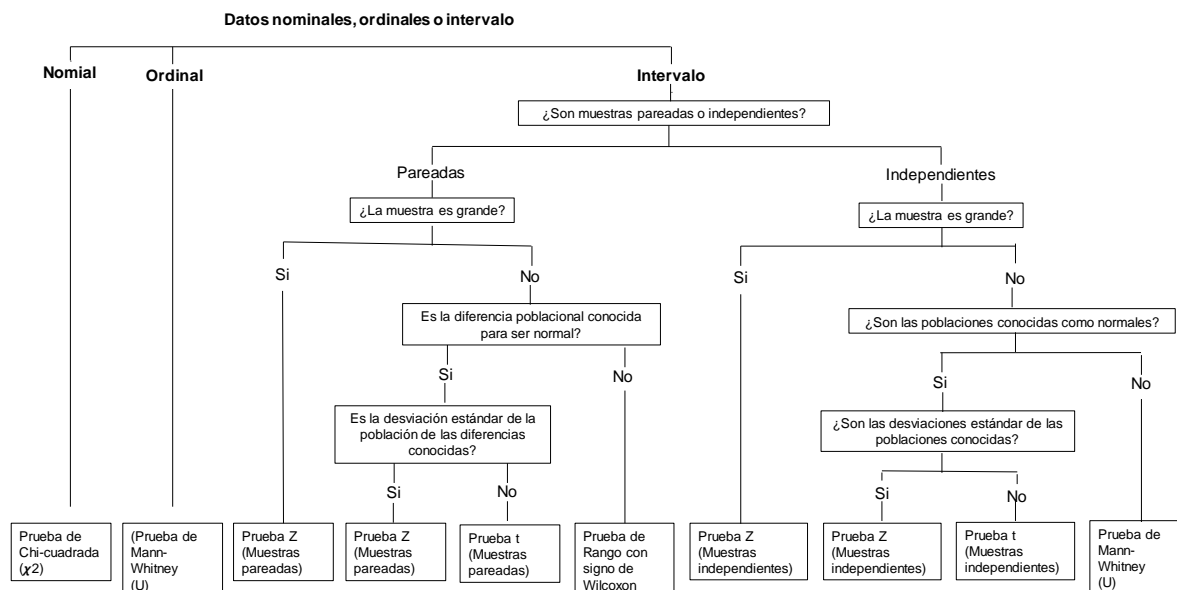


Figura 1. Diagrama de flujo para la elección de pruebas estadísticas correctas. Tomado de Edmondson y Druce, 1996.



Con el diagrama de flujo de la figura 1, como se mencionó previamente, elegirás qué prueba estadística es la más apropiada para analizar tus datos y obtener resultados, obviamente esto debe ser establecido previamente en la sección de metodología, aun cuando en algunas ocasiones debido a que no siempre se obtienen la cantidad de datos o surgen problemas, se modifica la metodología, en específico la elección de la prueba.

En la bioestadística para tener una muestra que realmente sea representativa de la población de deben cumplir con algunos supuestos, uno de los supuestos que son más frecuentemente mencionados para varias pruebas estadísticas es el que se refiere a la normalidad, la cual se define como los valores de determinada medición en un grupo de individuos normales de una población definida. Se ajusta a una distribución teórica conocida como: Distribución normal o Gaussiana.

El teorema del límite central (Soporte de Minitab, 2017)

El teorema del límite central es un teorema fundamental de probabilidad y estadística. El teorema describe la distribución de la media de una muestra aleatoria proveniente de una población con varianza finita. Cuando el tamaño de la muestra es lo suficientemente grande, la distribución de las medias sigue aproximadamente una distribución normal. El teorema se aplica independientemente de la forma de la distribución de la población. Muchos procedimientos estadísticos comunes requieren que los datos sean aproximadamente normales. El teorema de límite central le permite aplicar estos procedimientos útiles a poblaciones que son considerablemente no normales. El tamaño que debe tener la muestra depende de la forma de la distribución original. Si la distribución de la población es simétrica, un tamaño de muestra de 5 podría producir una aproximación adecuada. Si la distribución de la población es considerablemente asimétrica, es necesario un tamaño de muestra más grande. Por ejemplo, la distribución de la media puede ser aproximadamente normal si el tamaño de la muestra es mayor que 50. Las siguientes gráficas muestran ejemplos de cómo la distribución afecta el tamaño de la muestra que se necesita.

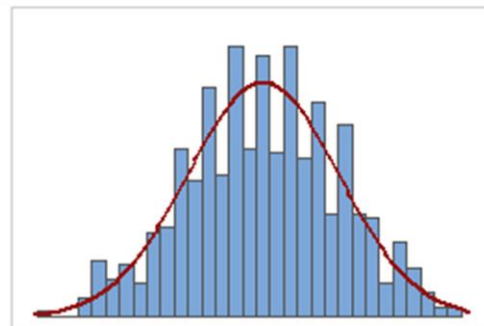


El teorema del límite central (Soporte de Minitab, 2017)

Una población que sigue una distribución uniforme es simétrica, pero marcadamente no normal, como lo demuestra el primer histograma. Sin embargo, la distribución de las medias de 1000 muestras de tamaño 5 de esta población es aproximadamente normal debido al teorema del límite central, como lo demuestra el segundo histograma. Este histograma de las medias de las muestras incluye una curva normal superpuesta para ilustrar esta normalidad.

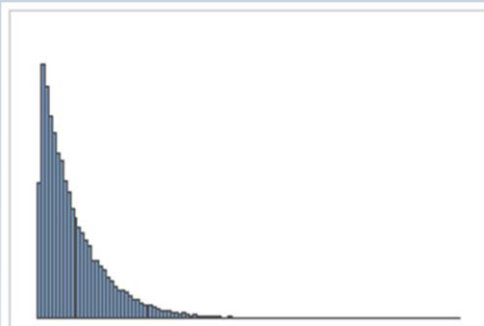


Distribución uniforme

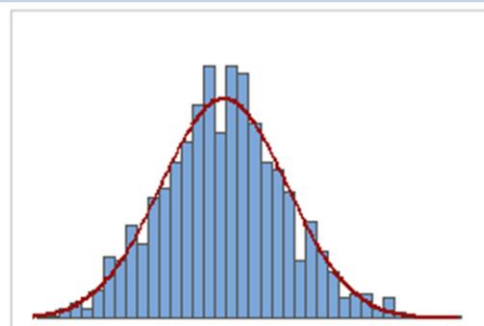


Medias de las muestras

Una población que sigue una distribución exponencial es asimétrica y no normal, como lo demuestra el primer histograma. Sin embargo, la distribución de las medias de 1000 muestras de tamaño 50 de esta población es aproximadamente normal debido al teorema del límite central, como lo demuestra el segundo histograma. Este histograma de las medias de las muestras incluye una curva normal superpuesta para ilustrar esta normalidad.



Distribución exponencial



Medias de las muestras



La distribución normal (figura 2) está determinada por dos parámetros: Media y desviación estándar; es simétrica en torno a la media; la media, la mediana y la moda son iguales; y el área total bajo la curva arriba del eje X es igual a la unidad (50% a la derecha y 50% a la izquierda) (Moreno, s.f.).

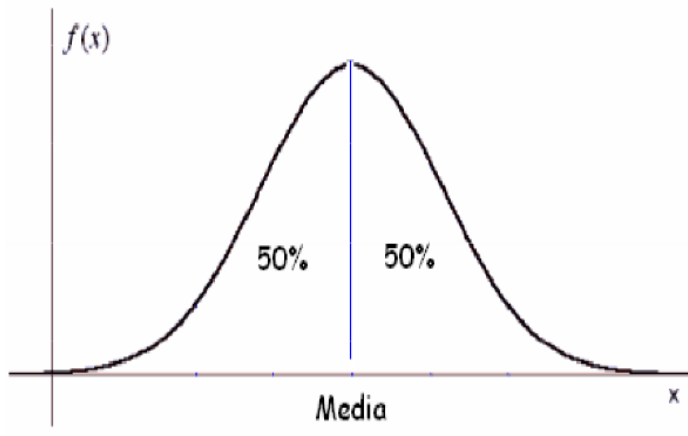


Figura 2. Distribución normal o Gausiana Tomada de Moreno (s. f.)

Ahora repasemos los conceptos y el significado de variable, muestreo y tamaño de muestra, aspectos que serán fundamentales para la validación de tu análisis estadístico y con ello para la discusión y conclusiones de tu Proyecto terminal.

Variables, muestreo y tamaño de muestra

En el Proyecto terminal I estableciste tu metodología, en este sentido seleccionaste y aplicaste correctamente los métodos de muestreo y su análisis, esto es fundamental para cumplir con los objetivos de tu proyecto y con base en ellos darás conclusiones sólidas y verificables. Esto es de suma importancia para las funciones y toma de decisiones de un TSU en Gestión de Servicios de Salud, tales decisiones se verán reflejadas en la calidad, eficacia y satisfacción de los servicios de salud que se ofrecen a la población.

A continuación, iniciaremos revisando los conceptos de población y tamaño de muestra con las que empezaste a trabajar en tu Proyecto terminal I.

La población estadística es un conjunto de elementos que comparten ciertas características, existen dos tipos de poblaciones:

Población finita: se refiere al conjunto de elementos que son cuantificables y que por ende se sabe exactamente cuántos elementos hay de una determinada población, por ejemplo, en un hospital tenemos un número finito de pacientes que se encuentra hospitalizados por eso es una población finita.



Población infinita: esto se refiere cuando la población es muy grande y prácticamente incuantificable, en este caso por cuestiones prácticas resulta frecuentemente imposible estudiar a todos los elementos de esa población, por ejemplo, la población de personas de un país.

Cuando se trabaja con poblaciones denominadas infinitas con la que se realizará un proyecto, se requiere obtener una muestra (figura 3); una muestra es una proporción de la población en su totalidad, esta muestra debe ser representativa, es decir que sea una representación de esa población infinita, es decir si el tamaño de la población es desconocida o infinita >10,000 elementos del universo.



Figura 3. Población y muestra. Tomado de CanStockPhoto, 2017.

De este modo tenemos los siguientes modelos de estimación para el tamaño de muestra para datos cuantitativos:

Modelo de la estimación del tamaño de la muestra para la población infinita o desconocida:

$$n = \frac{Z_{\alpha}^2 \cdot p \cdot q}{i^2}$$

Modelo de la estimación del tamaño de la muestra para la población finita y conocida:

$$n = \frac{Z_{\alpha}^2 \cdot N \cdot p \cdot q}{i^2(N - 1) + Z_{\alpha}^2 \cdot p \cdot q}$$

Donde:

n: tamaño de la muestra.

N: tamaño de la población

Z: valor correspondiente a la distribución de gauss, $\alpha = 0.05 = 1.96$ y $\alpha = 0.01 = 2.58$



p: prevalencia esperada del parámetro a evaluar, en caso de desconocerse ($p = 0.5$), que hace mayor el tamaño de la muestra.

q: $1 - p$ (si $p = 70\%$, $q = 30\%$).

i: error que se prevé cometer si es del 10% , $i = 0.1$

(Murray y Stephens, 2008).

Para su mejor comprensión se ejemplifica cada caso:

Para población infinita, es decir que desconoce el tamaño de la población. Se necesita estimar el tamaño de muestra de adultos mayores en una colonia popular de la Cd. de México, de este modo se tiene:

$$n = \frac{Z_{\alpha}^2 \cdot p \cdot q}{i^2}$$

Sustituyendo el modelo con datos, donde:

$Z_{\alpha}^2 = 1.96$, es decir para tener el 95% de nivel de confianza

p: prevalencia esperada del parámetro a evaluar, en caso de desconocerse ($p = 0.5$), que hace mayor el tamaño de la muestra.

q: $1 - p$ (si $p = 0.5$, $q = 0.5$).

$i^2 = 0.1^2$

Entonces:

$$n = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.1^2}$$

(Murray y Stephens, 2008).

$n = 96.04$

Como no existen 96.04 personas se redondea y serían 96 personas

Para población finita, es decir conocido el tamaño de la población, se presenta este ejemplo: Se desea estimar la prevalencia de apnea del sueño en la población femenina de la consulta de neumología de un hospital de segundo nivel, el tamaño de la población es de 249 personas.

Entonces se tiene:

$$n = \frac{Z_{\alpha}^2 \cdot N \cdot p \cdot q}{i^2(v - 1) + Z_{\alpha}^2 \cdot p \cdot q}$$



$$Z^2_{\alpha}=1.96^2$$

$$N=249$$

p: prevalencia esperada del parámetro a evaluar, en caso de desconocerse ($p = 0.5$), que hace mayor el tamaño de la muestra.

q: $1 - p$ (si $p = 0.5$, $q = 0.5$).

$$i^2=0.1^2$$

(Murray y Stephens, 2008)).

Substituyendo os datos en el modelo:

$$n = \frac{1.96^2 \cdot 249 \cdot 0.5 \cdot 0.5}{0.1^2(249 - 1) + 1.96^2 \cdot 0.5 \cdot 0.5}$$

$n = 80.52$, es decir 81 personas.

Con respecto al muestreo para datos cualitativo, para estimar el tamaño de muestra para población finita (cuando se emplean escalas nominales, como, por ejemplo: ausencia o presencia del fenómeno a investigar, se utiliza el modelo:

$$n = \frac{n'}{1 + n'/N}$$

Donde

$$n' = \frac{s^2}{\sigma^2}$$

$$s^2 = p(1 - p) \text{ y } \sigma^2 = (se)^2$$

Donde:

n = tamaño muestral

N = tamaño de la población

s^2 = varianza muestral

σ^2 = varianza poblacional



se= error estándar

p= porcentaje de confianza

(Murray y Stephens, 2008).

Para comprender esta modelo, se ejemplifica con un estudio sobre el conocimiento que se tiene en la población del 3er grado de 5 escuelas secundarias de la delegación Gustavo A. Madero en la Cd. De México sobre las formas de transmisión de VIH/SIDA.

La población está formada por 1098 estudiantes, los datos con los que se cuentan es un error estándar de 1.2% y confianza del 95%.

N=1098

se=1.2% =0.012

$s^2 = p(1-p) = 0.95(1-0.95) = 0.0475$

$\sigma^2 = (se)^2 = (0.012)^2 = 0.000144$

$$n' = \frac{0.0475}{0.000144}$$

$$n' = 329.86$$

$$n = \frac{329.86}{1 + \frac{329.86}{1098}}$$

$$n = \frac{329.86}{1 + \left(\frac{329.86}{1098}\right)}$$

n= 253.65, es decir, 254 estudiantes

Ahora se revisará que es una variable, la variable es una característica de la población o de la muestra cuya medida puede cambiar de valor. Según su naturaleza puede ser cualitativa y cuantitativa (Asuarza, 2006), de este modo existen:

- Variable cualitativa, categórica (o alfanumérica): Pueden tomar valores no cuantificables numéricamente. Se denomina categoría a cada uno de los valores que toma la variable, y pueden ser:

- Nominales: si no existe ningún orden entre las categorías de la variable. Ejemplos: el grupo sanguíneo (A, B, AB, O); el color del cabello (negro, café, rubio, rojo, blanco). Se debe distinguir las variables binarias, que son las que únicamente pueden tener dos valores, por ejemplo, el sexo; o la presencia o ausencia de alguna característica.

- Ordinales: cuando existe un cierto orden entre las categorías de la variable. Ejemplo: el nivel de estudios (sin estudios, básicos, medios, superiores), el grado de miopía (ausencia, bajo, medio, alto) (Grané, s.f.).



- Por intervalos: Pueden tratarse como ordinales y se pueden calcular distancias numéricas entre dos niveles. (Ejemplo: El número de años de educación recibidos (0, 1, 2...) es una variable cuantitativa que puede ser agrupada por intervalos).
- Variable cuantitativa (o numérica): Pueden tomar valores cuantificables numéricamente, y pueden ser:
 - Discretas: si solamente toman valores aislados (generalmente enteros). Suelen corresponder a contajes. Ejemplos: el número de hermanos, el número de tazas de café al día, el número de veces que se toma un medicamento en un día.
 - Continuas: potencialmente puede tomar cualquier valor numérico dentro de un intervalo o de una unión de intervalos. Ejemplos: el tiempo de reacción a un cierto medicamento, el peso de un individuo, la cantidad de ácido úrico en la sangre (Grané, s.f.).

De esta forma la definición de las variables es de particular importancia en proyectos experimentales; por su parte, los experimentos son una de las actividades a la que se recurre frecuentemente para comprender el comportamiento de un fenómeno, para comprobar o demostrar lo que teóricamente se deduce. Para ello es necesario comenzar por la identificación de las variables que caracterizan el fenómeno, experimento, características y cuyo cambio puede ser observado cualitativa y cuantitativamente. El conocimiento de las variables y la forma en que dependen unas de otras, es decir, saber cuál es la relación que existe entre ellas, es lo que permite obtener información del proceso y posteriormente elaborar un modelo de él.

La selección de variables de acuerdo con el tipo de información que se busca es el primer paso para efectuar un experimento. Cuando un fenómeno puede ser descrito por dos variables, se dice que una es la variable independiente y la otra dependiente. La variable independiente es la variable controlada, es decir, aquella a la que el experimentador le asigna valores determinados, y la variable dependiente es la variable que resulta afectada por los valores asignados a la variable independiente (Oda, 2005).

Retomando el tema de la muestra, debemos considerar que para la obtención de ésta existen diversos tipos de muestreo, el muestreo es un conjunto de métodos y procedimientos estadísticos destinados a la selección de una o más muestras; el objetivo principal de un diseño de muestreo es proporcionar procedimientos para la selección de muestras que sean representativas de la población en estudio (Asurza, 2006), así pues, existen dos tipos principales de muestreo (figura 4):

Muestreo probabilístico son aquellos que se basan en el principio de equiprobabilidad. Es decir, aquellos en los que todos los individuos tienen la misma probabilidad de ser elegidos para formar parte de una muestra y, consiguientemente, todas las posibles muestras de tamaño tienen la misma probabilidad de ser elegidas. Sólo estos métodos de muestreo probabilísticos nos aseguran la representatividad de la muestra extraída y son, por tanto, los más recomendables (Cuesta y Herrero, s.f.).

No probabilísticos, son estudios exploratorios, se acude a este método cuando el muestreo probabilístico resulta muy costoso, aun siendo conscientes de que no sirven para realizar



generalizaciones, pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos. En general se seleccionan a los sujetos siguiendo determinados criterios procurando que la muestra sea representativa (Cuesta y Herrero, s.f.). A continuación, se muestran los tipos de muestreo y subtipos de muestreo.

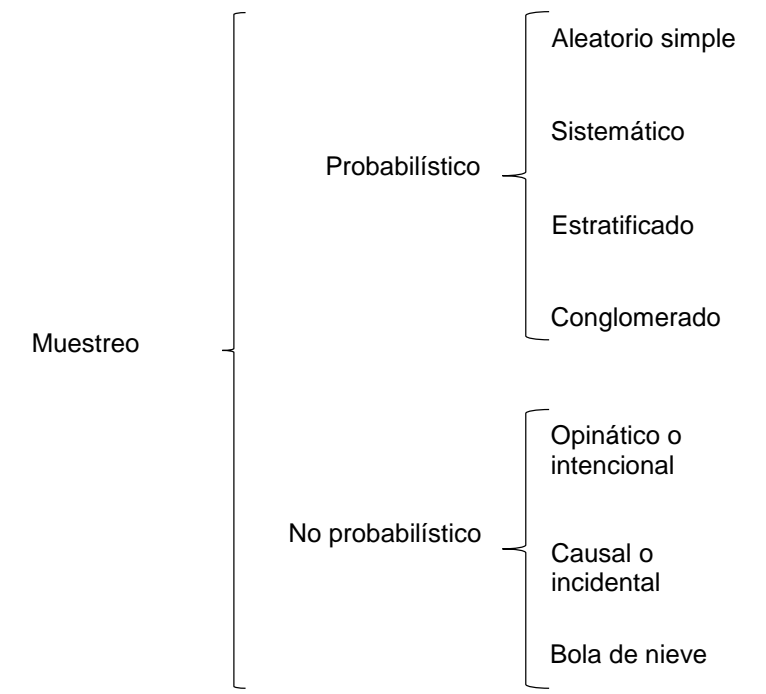


Figura 4. Tipos y subtipos de muestreo

El tamaño de muestra es fundamental para dar soporte y verosimilitud a los resultados estadísticos de la investigación. Para determinar el tamaño de muestra que debemos de obtener es necesario considerar:

El tipo de muestreo y los siguientes elementos:

- **Parámetro a estimar:** Son las medidas o datos que se obtendrán con respecto a la población.
- **Error muestral aceptable: (error estándar)** Es una medida de variabilidad de las estimaciones de muestras repetidas en torno al valor de la población, nos da una idea de la probabilidad en que la estimación basada en una muestra se aleja del valor que hubiera obtenido por medio de un censo completo.
- **La varianza poblacional:** Cuando una población es más homogénea la varianza es menor y el número de entrevistas necesarias para construir un modelo reducido de la población será más pequeño.
- **El nivel de confianza:** Es la probabilidad de que la estimación efectuada se ajuste a



la realidad (Cuesta y Herrero, s.f.)

Después de describir a la población se describe el método de obtención de muestra y de estimación del tamaño de muestra; y después se deben describir los y las técnicas y modelos de análisis de datos, todo esto con base al tipo de investigación, tipo de variables y revisión exhaustiva en la literatura, recordemos que frecuentemente existe herramientas y métodos variados para analizar los datos, es ahí cuando el investigador debe tomar la decisión de cuál de todos ellos elegirá y porqué, es importante que el investigador no presuponga que el lector conoce las técnicas, métodos o herramientas empleadas en la metodología; recuerda que las personas que lean tu proyecto no tienen que ser expertos en el tema. Es importante recomendar que la descripción del método debe llevar el mismo orden en que fueron planteados los objetivos específicos. Por otra parte, si hay metodologías de explicación amplia, es recomendable incluirlas en la sección de “Apéndice”, con el objetivo de que el lector no se pierda o distraiga en la lectura del documento, si bien es importante esta información eventualmente se describen con términos técnicos que no son siempre sencillos de comprender; todo lo anterior debiste haberlo incluido en la sección de Metodología.

Análisis de datos para variables cualitativas

Una vez descrita la metodología, el siguiente paso es la presentación de resultados. Si los datos de tu proyecto terminal incluyen variables cuantitativas, el primer acercamiento para analizar los datos es realizar gráficas, pongamos un ejemplo.

En tu proyecto se obtuvieron datos (números), lo cual es recomendable como un primer acercamiento a su análisis mostrarlos en figuras, específicamente en gráfica de distribución de frecuencias; por ejemplo, se tiene una muestra de 233 personas (hombres y mujeres) que fuman o no fuman, en este caso se tienen:

1. Frecuencias relativas marginales:

$$P(\text{ser hombre}) = 108 / 233 = 46.4\%$$

$$P(\text{ser mujer}) = 125 / 233 = 53.6\%$$

$$P(\text{fumar}) = 123 / 233 = 52.8\%$$

$$P(\text{no fumar}) = 110 / 233 = 47.2\%$$

2. Frecuencias relativas conjuntas:

$$P(\text{hombre y fumar}) = 65 / 233 = 27.9\%$$

$$P(\text{hombre y no fumar}) = 43 / 233 = 18.5\%$$

$$P(\text{mujer y fumar}) = 58 / 233 = 24.9\%$$



$$P(\text{mujer y no fumar}) = 67 / 233 = 28.8\%$$

3. Frecuencias relativas teóricas esperadas en caso de independencia:

$$E(\text{hombre y fumar}) = 46.4\% \times 52.8\% = 24.5\%$$

$$E(\text{hombre y no fumar}) = 46.4\% \times 47.2\% = 21.9\%$$

$$E(\text{mujer y fumar}) = 53.6\% \times 52.8\% = 28.3\%$$

$$E(\text{mujer y no fumar}) = 53.6\% \times 47.2\% = 25.3\%$$

4. Frecuencias absolutas teóricas esperadas en caso de independencia:

$$E(\text{hombre y fumar}) = 123 \times 108 / 233 = 57$$

$$E(\text{hombre y no fumar}) = 108 \times 110 / 233 = 51$$

$$E(\text{mujer y fumar}) = 123 \times 125 / 233 = 66$$

$$E(\text{mujer y no fumar}) = 125 \times 110 / 233 = 59$$

Estos datos pueden verse mejor explicados realizando figuras (1 y 2) como se observa a continuación:

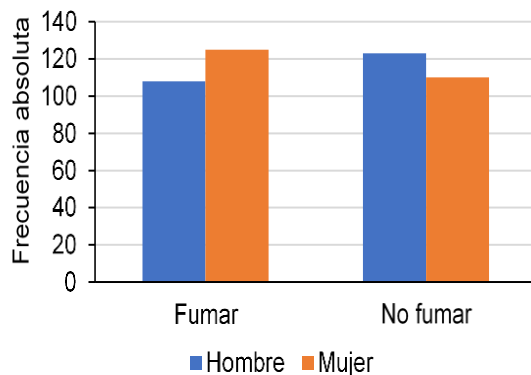


Figura 5. Frecuencia absoluta de personas que fuman y no fuman, n=233

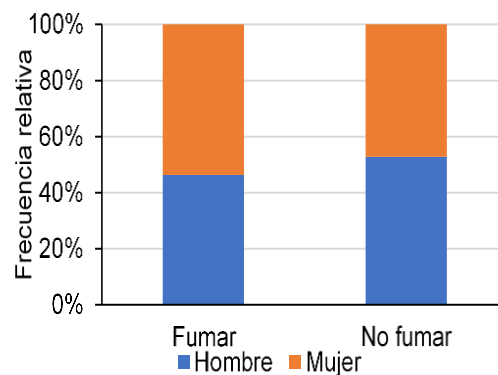


Figura 6. Frecuencia relativa de personas que fuman y no fuman, n=233

Aquí es a criterio del autor, es decir tú deberás tomar la decisión de qué es más útil o claro, presentar frecuencia absoluta, es decir los datos numéricos tal cual; o la frecuencia relativa, es decir el porcentaje que cada conjunto de dato representa con respecto al total de los casos.

Otro tipo de gráfica son las áreas; por ejemplo, según el canal Endémico General de Infecciones Respiratorias Agudas, el municipio de San Mateo del mar durante el año 2010 se encuentra en zona de seguridad, con tendencia del mismo comportamiento en la zona de San Pedro Pochutla, Santiago Niltepec y Santa María Petapaque están en la zona de alarma, se grafican los datos del promedio de niños (figura 7), menores de 5 años que son



diagnosticados a causa de neumococo cada mes.

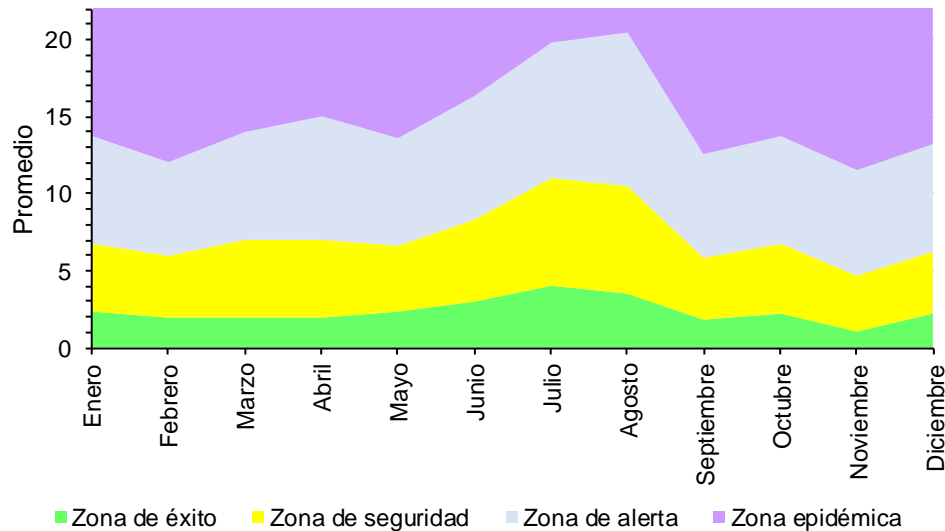


Figura 7. Promedio de niños con diagnóstico de neumococo.

Si bien este tipo de figuras ayudan a explicar los datos, son de forma muy simple, por otra parte, existen medidas de asociación entre variables cualitativas, medidas de asociación más empleadas en la práctica.

La primera que se revisará es la Chi cuadrada (χ^2), que es una prueba que compara los valores (n_{ij}) observados en la tabla de contingencia con los que teóricamente se obtendrían (t_{ij}) bajo la hipótesis nula (Universidad de Granada).

Ejemplo: Un médico muestrea a una población de mujeres para saber cuántas fuman y cuántas no, en dos colonias de la Ciudad de México, una considerada popular de nivel socioeconómico bajo, y otra de clase media

Se aplicará una Chi cuadrada, pero lo cual se realiza una tabla de contingencia para calcular la frecuencia esperada en cada caso, y se plantean dos hipótesis nulas:

H_{01} : La razón de la frecuencia en ambas columnas no muestran diferencias significativas.

H_{02} : La razón de la frecuencia en ambas filas no muestran diferencias significativas.

Es decir: frecuencia de no fumadoras es la misma que la frecuencia de las fumadoras, y la frecuencia de clase baja es la misma que la frecuencia en la clase media.

	No Fumadoras	Fumadoras	TOTAL
Clase baja	a	b	a+b



	14	6	20
Clase media	c 22	d 46	c+d 68
TOTAL	a+c 36	b+d 52	a+b+c+d 88

De este modo se calcula la frecuencia esperada de los cuatro grupos:

Por ejemplo, para estimar la frecuencia esperada de mujeres no fumadoras de clase baja es $((a+c) \times (a+b) \div (a+b+c+d) = 36 \times 20 \div 88 = 8.182$

Es decir, la suma de la columna de no fumadoras por la suma de la fila de clase baja, dividido entre la suma total, esto se hace para todos los grupos obteniéndose lo siguiente:

	No Fumadoras	Fumadoras
Clase baja	$36 \times 20 \div 88 = 8.182$	$52 \times 20 \div 88 = 11.818$
Clase media	$36 \times 68 \div 88 = 27.818$	$52 \times 68 \div 88 = 40.182$

Para corroborar que los cálculos sean correctos se suman las frecuencias esperadas: $8.182 + 11.818 + 27.818 + 40.182 = 88$

Con los valores de las cuatro frecuencias observadas y sus respectivas frecuencias esperadas, los componentes individuales de la prueba estadística pueden ahora ser estimados.

Pero antes se debe conocer los grados de libertad, en esta prueba de tablas de contingencia se estima como:

$$gl = (\text{número de columnas} - 1) \times (\text{número de filas} - 1)$$

(Edmonson y Druce, 2006).

Substituyendo los valores:

$$gl = (2 - 1) \times (2 - 1) = (1 \times 1) = 1$$



	No Fumadoras	Fumadoras
Clase baja	$\frac{(14 + 8.182 - 0.5)^2}{8.182}$ $=3.457$	$\frac{(6 + 11.818 - 0.5)^2}{11.818}$ $=2.393$
Clase media	$\frac{(22 + 27.818 - 0.5)^2}{27.818}$ $=1.017$	$\frac{(46 + 40.182 - 0.5)^2}{40.182}$ $=0.704$

$$\chi^2 = 3.457 + 2.393 + 1.017 + 0.704$$

Estimación de $\chi^2=7.571$

Consultando la tabla de distribución de Chi cuadrada se busca el valor en 1 grado de libertad, y con $p=0.01$ y 0.05 , es decir, 6.6349 y 3.8415, respectivamente; como se aprecia el valor estimado fue 7.571 que es mayor a 6.6349 y 3.8415 (valor de tablas), por lo tanto se acepta que las mujeres no fumadores se relacionan significativamente con la clase baja, y que las mujeres fumadoras se relacionan significativamente con la clase media.

Coefficiente de contingencia o C de Pearson: Es una medida del grado de intensidad de la relación que existe entre dos o más variables cualitativas; se sustenta en la comparación de las frecuencias calculadas de dos características con respecto a las frecuencias que se esperan con independencia de estas características.

$$C = \sqrt{\frac{\chi_{exp}^2}{(\chi_{exp}^2 + n)}}$$

El valor máximo es:

$$Max\{C\} = \sqrt{\frac{Min\{r-1, c-1\}}{1 + Min\{r-1, c-1\}}}$$

Y toma valores del intervalo:



$$0 < C < \sqrt{\frac{\text{Min}\{r-1, c-1\}}{1 + \text{Min}\{r-1, c-1\}}}$$

Si $C=0$ indica dependencia absoluta

$C=$ Máxima (C) indica dependencia perfecta

(Eomondson y Druce, 1996)

Ejemplo: Para analizar si el estado civil no es una variable relevante para explicar la aceptación de métodos anticonceptivos, se encuestó a 500 personas, y se obtuvieron los resultados siguientes:

	Actitud a Favor de Anticonceptivos	Actitud en Contra Anticonceptivos	Total
Solteros	120	30	150
Casados	50	200	250
Divorciados	30	70	100
Total	200	300	500

Calculamos las frecuencias esperadas

Calculamos el valor Chi-cuadrado

$$\chi^2_{exp} = \sum_i \sum_j \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = 145.83$$

Calculamos el valor C :

$$C = \sqrt{\frac{\chi^2_{exp}}{(\chi^2_{exp} + n)}} = 0.475$$

Calculamos el valor máximo de C :



$$\begin{aligned} \text{Max}\{C\} &= \sqrt{\frac{\text{Min}\{r-1, c-1\}}{1 + \text{Min}\{r-1, c-1\}}} = \\ \text{Max}\{C\} &= \sqrt{\frac{\text{Min}\{3-1, 2-1\}}{1 + \text{Min}\{3-1, 2-1\}}} = 0.7071 \end{aligned}$$

Por tanto, la intensidad de la asociación es considerable, ya que 0.475 es más de la mitad del valor máximo que en este caso fue 0.7071

Coefficiente Φ de Cramer: es otro coeficiente usado para ver la asociación de las variables nominales cuando sus categorías son de dos o tres clases. El coeficiente varía entre cero y uno.

Si la tabla de contingencia es de dos filas por dos columnas, o es de tres filas por tres columnas, es válido este coeficiente. Cuanto más próximo a cero se encuentre, más independientes serán las variables; cuanto más próximo a uno sea el número, más asociadas estarán las variables que se estudien. Es importante señalar que, para el cálculo del coeficiente de Cramer, se necesita previamente tener calculado el estadístico Chi cuadrada. El coeficiente de Cramer se estima como:

$$\Phi_1 = \sqrt{\frac{\chi^2}{n}}$$

Oscila entre 0 o 1 (al igual que Φ_2^1 , que también se puede usar como medida de asociación). Esto se basa en χ^2 y se puede describir como:

$$\Phi_2 = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{C_1 C_2 R_1 R_2}}$$

Y su intervalo de valores va de 0 hasta 1, donde:

$\Phi = 0$, no hay relación entre X e Y

$\Phi = 1$, hay una relación perfecta entre X e Y

$\Phi = 0,6$, hay una correlación relativamente intensa entre X e Y

(Zar, 1984).



Paciente	Presencia de Sintomatología	Presencia del parásito
1	+	+
2	+	+
3	-	-
4	-	+
5	+	+
6	-	+
7	-	-
8	+	+
9	-	+
10	-	-
11	+	+
12	-	-
13	+	+
14	-	+

Los datos pueden ser tabulados en una tabla de contingencia 2 x 2, quedando como la siguiente tabla de contingencia:

		Sintomatología		TOTAL
		Presencia	Ausencia	
Presencia de parásito	Presencia	6	4	10
	Ausencia	0	4	4
TOTAL		6	8	14



$$\Phi_2 = \frac{(6x4)-(4x0)}{\sqrt{(6x8x10x4)}} = \Phi_2 = \frac{24}{\sqrt{1920}} = 0.55$$

$$\Phi_2 = \frac{(6x4) - (4x0)}{\sqrt{(6x8x10x4)}}$$

Es decir, la relación es relativamente intensa, según lo previamente señalado ($\Phi = 0.6$).

Por otra parte, en tu Proyecto terminal posiblemente utilices variables que sean numéricas, es decir cuantitativas, por lo cual también debes tener nociones de pruebas estadísticas que puedes utilizar para analizar tus datos y obtener resultados, y que a continuación se revisarán. Se te sugiere que revises los materiales que utilizaste en las asignaturas de Estadística básica y Bioestadística.

Medidas de tendencia central, de dispersión y distribución

Para variables numéricas, en las que puede haber un gran número de valores observados distintos, se ha de optar por un método de análisis distinto, respondiendo a las siguientes preguntas:

1. ¿Alrededor de qué valor se agrupan los datos?
2. Supuesto que se agrupan alrededor de un número, ¿Cómo lo hacen? ¿Muy concentrados? ¿Muy dispersos?

Para dar respuesta a estas preguntas se utilizan las medidas de tendencia central y que a continuación se explican.

a. Medidas de tendencia central

Las medidas de centralización son las más evidentes que podemos calcular para describir un conjunto de observaciones numéricas es su valor medio. La media no es más que la suma de todos los valores de una variable dividida entre el número total de datos de los que se dispone.

Las medidas de tendencia central son:

- **Mediana.** Es el valor que divide al conjunto de datos ordenados, en aproximadamente dos partes: 50% de valores son inferiores y otro 50% son superiores (Asurza, 2006), ver ejemplo en la tabla X.
- **Media aritmética para datos simples.** Es una medida de tendencia central que denota el promedio de un conjunto de datos (ver ejemplo en la tabla X), y se estima como:



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

Ver ejemplo en la tabla 4

- **Media aritmética para datos agrupados:** Se calcula multiplicando cada valor de los elementos por el número de veces que se repite. La suma de todos estos elementos se divide entre el total de datos, ver ejemplo en la tabla X.

$$\bar{X} = \frac{(x_1 \cdot n_1) + (x_2 \cdot n_2) \dots + (x_m \cdot n_m)}{N}$$

x_i representa el valor de la marca de clase o punto medio del intervalo.

n_i representa la frecuencia absoluta

N representa el total de datos

Ver ejemplo en la tabla 4

- **Media armónica:** Es un valor que se obtiene como la inversa de la media de las inversas de las observaciones (ver ejemplo en la tabla X). Se le denota por H .

$$H = \frac{1}{\sum_{i=1}^n \frac{1}{x_i} \cdot n_i}$$

c_i representa el valor de la variable o en su caso la marca de clase.

n_i representa la frecuencia absoluta

Media geométrica, dados dos números y_1 e y_2 , llamaremos media geométrica (G) de estos números a la raíz cuadrada del producto de los mismos. Cuando se tiene N observaciones (más de dos datos): x_1, x_2, \dots, x_p y cada uno de ellos se repite n_1, n_2, \dots, n_p veces entonces, generalizando la primera expresión se tiene:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_p^{n_p}}$$

La media armónica solo se puede calcular si no hay observaciones negativas o valores cero. Es menos sensible que la media aritmética a los valores extremos. Su valor es siempre menor o igual que la media aritmética. Su uso más frecuente es el de promediar porcentajes, tasas, números índices, entre otros, es decir en los



casos que se supone que la variable presenta variaciones acumulativas, ver ejemplo en la tabla 4.

- **Moda:** Es el valor de la variable que tiene mayor frecuencia absoluta, la que más se repite es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo. Por su propia definición, la moda no es única, pues puede haber dos o más valores de la variable que tengan la misma frecuencia siendo esta máxima. Entonces tendremos una distribución bimodal o polimodal según el caso.

$$Mo = Ll + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i + n_{i+1})} \cdot C_i$$

LI Es el límite inferior de la clase modal.

$n_i - n_{i-1}$ Es la diferencia de la frecuencia absoluta de la clase modal menos la frecuencia del intervalo anterior.

$n_i - n_{i+1}$ Es la diferencia de la frecuencia absoluta de la clase modal menos la frecuencia del intervalo posterior

C_i Es la amplitud del intervalo.

Clase modal es el intervalo que tiene mayor frecuencia o frecuencia relativa.

Para la mejor comprensión de estos conceptos, se presentan datos y los cálculos de las medidas de tendencia central para el tamaño de la población de los 18 municipios de Sinaloa:

Municipio	Población en miles de personas
Cosalá	16
San Ignacio	21
Concordia	27
Badiraguato	31
Choix	33
Mocorito	45
Angostura	47



Rosario	53
Elota	53
Escuinapa	59
Salvador Alvarado	81
Sinaloa de Leyva	87
El Fuerte	100
Navolato	154
Guasave	295
Ahome	450
Mazatlán	502
Culiacán	905
N	18
Sumatoria	2959
Mediana 18 datos, divididos a la mitad, son 9; el valor del dato 9 es 53	
Media aritmética para datos simples , suma de todos los datos y divididos entre el número de datos =164.38	
Media aritmética para datos agrupados , cada dato multiplicado por el número de veces que aparece, se suman todos estos datos y dividen entre el número total de datos= 164.38	
Moda , es el dato que más se repite, en este ejemplo, es bimodal= 53, con 2 veces.	

b. Medidas de dispersión

Tal y como se adelantaba antes, otro aspecto para tener en cuenta al describir datos (cuantitativos) continuos es la dispersión de estos. Existen distintas formas de cuantificar esa variabilidad. De todas ellas, la varianza (S^2) de los datos es la más utilizada. Es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.



Por su parte, las medidas de dispersión son las que a continuación se explican (Fernández y Pértega, 2001); y que para que se comprendan mejor se realizan los cálculos utilizando los datos siguientes:

	Peso (kg)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	52	-14.5	210.25
	57	-9.5	90.25
	58	-8.5	72.25
	60	-6.5	42.25
	65	-1.5	2.25
	66	-0.5	0.25
	66	-0.5	0.25
	66	-0.5	0.25
	66	-0.5	0.25
	67	0.5	0.25
	68	1.5	2.25
	69	2.5	6.25
	70	3.5	12.25
	70	3.5	12.25
	71	4.5	20.25
	74	7.5	56.25
	75	8.5	72.25
	77	10.5	110.25
Media	66.500	S^2	35.53
		S	5.96



- **Rango:** es el intervalo en que se distribuyen los datos en observaciones de una muestra y se determina restandole al mayor valor el menor valor. La definición matemática del rango es:

$$R = X_n - X_1$$

Donde:

X_n = valor mayor

X_1 = valor menor

(García y Matus, 2010)

$$R = 77 - 52$$

$$R = 25$$

- **Varianza.** Es la media de los cuadrados de las diferencias entre cada valor de la variable y la media aritmética de la distribución.

$$S_x^2 = \frac{\sum_{j=1}^n (X_j - (\bar{X}))^2}{n}$$

Donde:

S_x^2 = varianza

X_j = cada dato

\bar{X} = Media de los datos

n = número de datos

$$S_x^2 = \frac{\sum (X_j - (66.5))^2}{20}$$

$$S_x^2 = 35.525$$

- **Desviación típica (S):** Es la raíz cuadrada de la varianza. Dice cuánto se alejan de ésta los valores en general. Si todos los valores están más cerca de la media aritmética, la desviación estándar también se hace más pequeña (e indica una distribución más homogénea). En distribuciones con alta desviación estándar (distribuciones más heterogéneas), usualmente la media aritmética deja de ser



representativa. Expresa la dispersión de la distribución y se expresa en las mismas unidades de medida de la variable. La desviación típica es la medida de dispersión más utilizada en estadística.

$$S_x = \sqrt{\frac{\sum_{j=1}^n (X_j - (\bar{X}))^2}{n - 1}}$$

Dado que la desviación estándar es la raíz cuadrada de la varianza, entonces,

$$S_x = \sqrt{735.525}$$

$$S_x = 5.96$$

- **Coeficiente de variación (CV).** Es una medida de dispersión relativa de los datos y se calcula dividiendo la desviación típica muestral por la media y multiplicando el cociente por 100.

$$CV = \left(\frac{\sigma}{\mu} \right) \cdot 100$$

$$CV = \left(\frac{5.96}{66.5} \right) \cdot 100$$

$$CV = 8.96$$

- **Los cuartiles y percentiles** son medidas de posición: El percentil es el valor de la variable que indica el porcentaje de una distribución que es igual o menor a esa cifra. Por su parte, los cuartiles son valores que dividen una muestra de datos en cuatro partes iguales. Utilizando cuartiles puede evaluar rápidamente la dispersión y la tendencia central de un conjunto de datos, que son los pasos iniciales importantes para comprender sus datos. Los cuartiles son valores calculados, no observaciones en los datos (tabla 2). A menudo es necesario interpolar entre dos observaciones para calcular un cuartil con exactitud (Minitab, 2017).



Tabla 2. Descripción de cuartiles.

Cuartil	Descripción
1er cuartil (Q1)	25% de los datos es menor que o igual a este valor.
2do cuartil (Q2)	La mediana. 50% de los datos es menor que o igual a este valor.
3er cuartil (Q3)	75% de los datos es menor que o igual a este valor.
Rango intercuartil	La distancia entre el primer 1er cuartil y el 3er cuartil (Q3-Q1); de esta manera, abarca el 50% central de los datos.

Para comprender mejor la distribución de los datos pongamos un ejemplo: Se obtuvieron los datos de la ingesta de carbohidratos totales (CHO gr) por día de 45 niños de 8 años:

CHOS (g)	CHOS (g)	CHOS (g)	CHOS (g)	CHOS (g)
85.5	92.0	89.9	82.6	77.9
89.2	82.1	87.6	71.1	92.1
84.3	93.5	93.5	88.6	77.3
79.3	70.0	50.5	87.3	70.3
94.7	82.2	90.2	85.9	93.6
63.4	83.4	88.3	87.2	93.6
88.9	56.1	88.6	72.6	73.4
91.9	62.1	88.0	88.2	92.9
76.4	76.8	83.2	83.8	92.5

Al revisar la tabla se observa que el valor de CHOS más bajo es 50.5g, y el más alto es 94.7 g; se puede entonces crear un intervalo utilizando el modelo de Sturges:

$$K = 1 + (3.3 \times \log n)$$

Con nuestro conjunto de datos sustituimos los valores en el modelo:

$$K = 1 + (3.3 \times \log 45) = 6.5$$



De acuerdo con la regla de Sturges, deberíamos tener 6 o 7 clases, dado que el resultado fue 6.5 el valor de K se aproxima al entero más próximo, en este caso 7.

Ahora se calcula la amplitud del intervalo, de este modo:

$$AI = \frac{\text{Valor del dato más alto} - \text{Valor del dato menor}}{K}$$

Substituimos los valores:

$$AI = \frac{94.7 - 50.5}{7}$$

$$AI = \frac{44.2}{7} = 6.31$$

La amplitud del intervalo se aproxima al entero más cercano, es decir 6.

Obteniéndose la figura 8. Esta gráfica es lo que se denomina Distribución de frecuencias, que significa que una vez que se obtienen el rango (valor máximo menos el valor mínimo de la variable x), este intervalo de valores se divide en intervalos más pequeños que se ponen en el eje horizontal; y después se cuantifica el número de datos de la variable y que caen dentro de cada intervalo de la variable x.

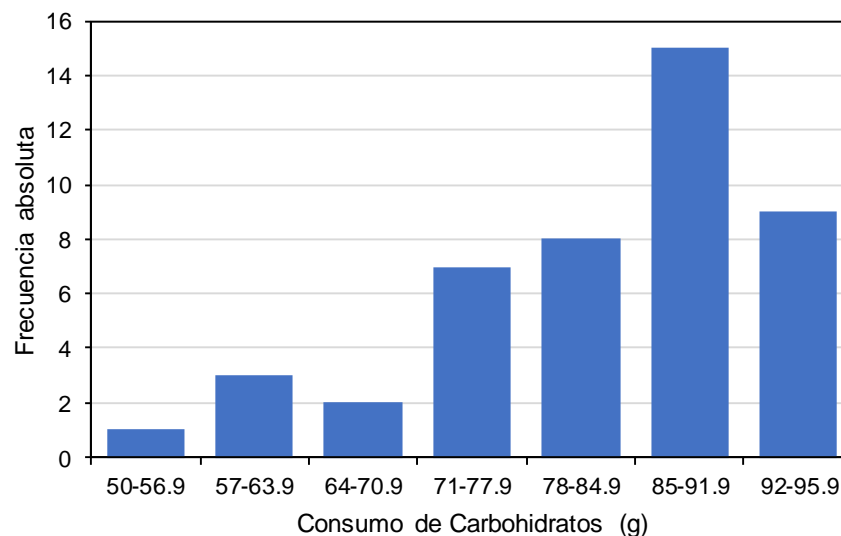


Figura 8. Frecuencia de distribución del consumo de carbohidratos en niños de 8 años.

Es importante que conozcas como obtener el número de intervalos, dado que estos influyen en la forma que tendrá la curva de distribución (normal o no), y que a su vez es un requisito para varias pruebas estadísticas.



Para conocer más consulta el material de Medidas de tendencia central y dispersión



A continuación, se describirán de manera específica cada una de las pruebas indicadas en la figura 1. Se realiza cada una de las pruebas paso a paso; sin embargo, es importante señalar que, por cuestiones prácticas, es altamente recomendable que utilices algún programa de estadística, pues en ocasiones el número de datos es muy grande, y hacerlo paso a paso representará una inversión grande de tiempo.

Prueba U Mann-Whitney

La prueba de U Mann-Whitney se realiza cuando se cumplen las siguientes condiciones estadísticas:

Las muestras deberán ser independientes, y al menos una de las muestras debe tener más de cinco valores.

Para variables con datos en intervalos deben ser variables independientes, al menos una de las muestras debe tener más de cinco valores; las muestras se extraen de poblaciones con distribución no normal, pero con formas similares, o las formas de las poblaciones son desconocidas (Edmondson y Druce, 1996).

Ejemplo: Se ha demostrado que la cantidad de manipulaciones que reciben los animales en los experimentos de laboratorio afecta a su comportamiento.

Una investigación sobre el efecto del personal de enfermería registró la actividad de 10 enfermeras que tuvieron preparación sobre el tema de tanatología que duró los primeros 25 días de su ingreso como personal de un hospital de alta especialidad y ocho resultados se muestran en la tabla siguiente:

	Marcador de conducta realizada por un experto del tema									
Enfermeras con preparación sobre el tema	215	220	249	254	260	265	290	300	306	320



Enfermeras sin preparación sobre el tema	140	170	192	205	215	240	245	305		
--	-----	-----	-----	-----	-----	-----	-----	-----	--	--

Los observadores predijeron que las enfermeras que tuvieron preparación sobre el tema de tanatología serían más activas y eficientes en brindar atención humanitaria a los pacientes con respecto a aquellas que no tuvieron preparación, y eligieron una prueba U de Mann Whitney de una cola.

Los datos se miden en el nivel ordinal, ya que la calificación de los observadores solo se puede utilizar para colocar a las enfermeras en orden, desde las más activas y eficientes hasta las menos activas.

1. 18 enfermeras recién contratadas por un hospital de alta especialidad y se asignaron a las dos muestras, un grupo tuvo preparación del tema de tanatología y otro no. Por lo tanto, las muestras son independientes.
2. Ambas muestras tienen más de cinco valores.

Se plantean dos hipótesis:

HN: Las enfermeras con preparación sobre el tema de tanatología no influyen en el nivel de actividad y eficiencia hacia los pacientes.

HA: Las enfermeras con preparación sobre el tema de tanatología influyen en el nivel de actividad y eficiencia hacia los pacientes.

Nivel de significancia $P=2.5\%$ (0.025).

Es importante hacer notar que se elegirá una prueba de una cola porque los investigadores que observaron la actividad de las enfermeras predijeron la tendencia del resultado. Por otra parte, la hipótesis nula H_0 e H_A (hipótesis alterna) son frases redactadas en términos de causa y efecto porque esto fue un experimento controlado; alguna diferencia en el nivel de actividad y eficiencia serán debidas a la cantidad de preparación que tuvieron.

Sea n_1 el tamaño de la muestra más pequeño, y n_2 el tamaño de la muestra más grande, en este ejemplo $n_1 = 8$, $n_2=10$.

Se ordenan todos los valores, desde el más pequeño (rango número 1) al más grande.



	MARCADOR DE ACTIVIDAD Y EFICIENCIA																	
Enfermeras con preparación						215	220			249	254	260	265	290	300		306	320
Rango de enfermeras con preparación						5.5	7			10	11	12	13	14	15		17	18
Rango de enfermeras sin preparación	1	2	3	4	55			8	9							16		
Enfermeras sin preparación	140	170	192	2005	215			240	245							305		

Si dos o más valores son los mismo, se da a cada valor el promedio de los rangos que ellos ocupan. En nuestro ejemplo el valor de 215 que ocupan en quinto y sexto lugar, el promedio de estos es 5.5 ($\frac{5+6}{2} = 5.5$).

Se calcula la sumatoria de los rangos de a muestra más pequeña, representada algebraicamente por **R**

De este modo se estima **R** para la muestra n_1 :

$$R = 1+2+3+4+5.5+8+9+16=48.5$$

Se calcula el valor de U utilizando la fórmula:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R$$

(Edmondson y Druce, 1996).

$$U = 8 \times 10 + \frac{8(8 + 1)}{2} - 48.5$$

$$U = 67.5$$

Se estima el valor de U'

$$U' = n_1 n_2 - U$$



$$U' = 8 \times 10 - 67.5$$

$$U' = 12.5$$

Utilizando la tabla de valores críticos de U de Mann-Whitney, para $n_1=8$ y $n_2=10$, con nivel de significancia de 2.5% (una cola).

Rechazamos la hipótesis nula y aceptamos la hipótesis alterna si el valor más pequeño de U y U' es menor o igual al valor crítico de U.

Ahora $U=67.5$ y $U'=12.5$ y la más pequeña de los dos valores es $U' = 12.5$

Entonces $U'=12.5$ es menor que el valor crítico de $U=17$ (valor de tablas), por ende, debemos rechazar la hipótesis nula y aceptar la hipótesis alterna que indica que la preparación de las enfermeras en el tema de tanatología incrementa la actividad y eficiencia hacia los pacientes.

¿Cuál es la diferencia entre una prueba de dos colas (bilateral) y de una cola (unilateral)?
(XLStat, 2017)

Una prueba estadística se basa en dos hipótesis competitivas: la hipótesis nula H_0 y la hipótesis alternativa H_a .

El tipo de hipótesis alternativa H_a define si una prueba es de una cola (unilateral) o de dos colas (bilateral).

Pruebas bilaterales o de dos colas

Una prueba de dos colas se asocia a una hipótesis alternativa para la cual se desconoce el signo de la potencial diferencia. Por ejemplo, supongamos que deseamos comparar las medias de dos muestras A y B. Antes de diseñar el experimento y ejecutar la prueba, esperamos que, si se resalta una diferencia entre las dos medias, realmente no sabemos si A debería ser superior a B o a la inversa. Esto nos lleva a elegir una prueba de dos colas, asociada a la siguiente hipótesis alternativa: H_a : $\text{media}(A) \neq \text{media}(B)$. Las pruebas de dos colas son con diferencia las más utilizadas.

Pruebas unilaterales o de una cola

Una prueba de una cola normalmente está asociada a una hipótesis alternativa para la cual se conoce el signo de la potencial diferencia antes de ejecutar el experimento y la prueba. En el ejemplo descrito más arriba, la hipótesis alternativa referida a una prueba de una cola podría redactarse así: $\text{media}(A) < \text{media}(B)$ o $\text{media}(A) > \text{media}(B)$, dependiendo de la dirección esperada de la diferencia.



Prueba Z (muestras pareadas)

Puedes utilizar esta prueba en el caso de tengas muestras pareadas. Las muestras pareadas se obtienen usualmente de distintas observaciones realizadas sobre los mismos individuos. Por ejemplo:

Medir la glucosa en 10 personas en la mañana = Muestra A

Medir la glucosa de las mismas personas en la tarde= Muestra B

En otro ejemplo, se tienen ocho muestras tomadas del mismo individuo, 4 con los ojos abiertos y 4 con los ojos cerrados. Por lo tanto, se tienen muestras pareadas.

En contraparte, se toman datos entre diferentes individuos, por ejemplo, ancianos y jóvenes, entonces se tendrán muestras NO pareadas (Random Notes, 2010).

Esta prueba también se emplea cuando la distribución es normal en forma de campana (Gausiana) y la desviación estándar es conocida, además si las muestras son mayores a 30 valores (Edmondson y Druce, 1996).

Veamos el siguiente ejemplo:

Se realiza una investigación sobre el efecto que tienen el anuncio del aumento del 80% en el bono de puntualidad y su eficiencia en el área de quirófanos de un hospital, se tienen datos de la hora de llegada y eficiencia del personal (60 personas) antes y después de la notificación del aumento en el bono de puntualidad.

La diferencia, d , en la hora de llegada antes y después de la noticia de incremento en el bono fue calculada para cada persona, la tabla siguiente:

Persona	Tiempo de llegada con respecto a la hora establecida de entrada (segundos)		Diferencia en hora de llegada (segundos)
	Antes de la notificación del aumento en el bono de puntualidad	Después de la notificación del aumento en el bono de puntualidad	
A	285	270	$285-270=15$
B	315	285	$315-285=30$
C	255	245	$255-245=10$
⋮	⋮	⋮	⋮



La diferencia media, \bar{d} , y la desviación estándar de las diferencias. S. fue calculada como:

$$n = 60$$

Diferencia media de la hora de llegada = $\bar{d} = 18.2$ seg.

Desviación estándar de diferencias = $s = 25.6$ seg.

Se quiere determinar si el anuncio de aumento en el bono de puntualidad afecta en la hora de llegada y en su eficiencia en el área de trabajo.

Es importante tomar en cuenta en la hora de llegada antes y después de la notificación en cada persona (muestra pareada), el tiempo de reacción es medida en intervalos, la muestra es de tamaño grande.

Las hipótesis que se plantean son:

HN: El anuncio de aumento en el bono de puntualidad y eficiencia no tiene relación con la puntualidad y eficiencia.

HA: El anuncio de aumento en el bono de puntualidad y eficiencia no tiene relación con la puntualidad y eficiencia.

Se utiliza dos colas por ser muestras pareadas, y porque no se predice la dirección o tendencia del resultado.

Nivel de significancia de $P = 5\%$ (0-05)

La HN e HA son frases en términos de causa y efecto porque este es un ejemplo controlado, algunas diferencias en el tiempo de llegada será debido por el anuncio del aumento del bono de puntualidad y eficiencia.

Ahora se calcula el valor de Z, usando la siguiente fórmula:

$$Z = \frac{\bar{d} \sqrt{n}}{s}$$

(Edmondson y Druce, 1996).

Para nuestro caso:

$$Z = \frac{18.2 \times \sqrt{60}}{25.6}$$

$$Z = \frac{18.2 \times 7.75}{25.6}$$

$$Z = 5.51$$

Para la prueba Z de dos colas, con en el nivel de significancia de 5%, tenemos $\frac{5\%}{2} = 2.5\%$

Así que la probabilidad es $100\% - 2.5\% = 97.5\%$, que equivale a 0.975



Se utilizará el apéndice X para encontrar el valor crítico de Z correspondiendo a la probabilidad de 0.975, de este modo el valor crítico de $Z=1.96$

En la primera columna se busca el valor de 1, y de punto nueve, y se recorre el renglón hasta encontrar el valor de z de 0.06, de este modo la intersección señala 0.9750.

Si el valor de la prueba es negativo, se convierte en positivo.

Dado que el valor de prueba $Z = 5.51$ es un número positivo, no hay necesidad de convertirlo.

De este modo se rechaza la hipótesis nula y se acepta la hipótesis alterna ya que el valor de la prueba es positivo de Z y es más grande que el valor crítico de Z. Ya que el valor de prueba $z = 5.51$ es mayor que el valor crítico $Z = 1.96$ se debe rechazar la hipótesis nula y aceptar la hipótesis alternativa de que los tiempos de respuesta son diferentes antes y después de anunciar los estímulos por puntualidad y eficiencia.

Prueba t (muestras pareadas)

La prueba t para muestras pareas se utiliza en los siguientes casos:

Muestras pareadas, la población debe tener distribución normal, la desviación estándar es conocida y el tamaño de la muestra es menor a 30 valores (Edmondson y Druce, 1996).

Veamos un ejemplo, se han registrado el tiempo que 21 empleados de farmacia tardan en surtir una receta en dos turnos diferentes (son los mismos empleados en dos turnos diferentes); en el día y en la noche.

Con base de hallazgos de investigaciones anteriores, el equipo de investigación predijo que el tiempo que los empleados de la noche tarden en surtir una receta es mayor durante la noche. Los resultados de la investigación actual se muestran a continuación.

		Enfermeras																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Pacientes graves	Día	4.7	2.5	3.2	1.6	4.9	4.2	5.5	3.8	7.5	6.7	6.3	4.6	6.8	3.4	1.9	1.8	4.3	7.1	5.3	2.9	4.5
	Noche	4.8	2.8	3.3	1.6	4.9	4.4	5.4	4.0	7.4	6.8	6.3	4.8	6.6	3.5	1.9	1.7	4.4	7.4	5.3	3.1	4.7
Diferencia $d=\text{día-noche}$		0.1	0.3	0.1	0	0	0.2	-0.1	0.2	-0.1	0.1	0	0.2	-0.2	0.1	0	-0.1	0.1	0.3	0	0.2	0.1

Se utilizará una prueba de t de una cola porque el tiempo en que tarda en surtir una



receta son un par para cada empleado, es decir es una muestra pareada; además de que el equipo que realiza esta investigación ha predicho la tendencia del resultado.

La tabla del conteo a continuación muestra las diferencias y se distribuyen normalmente

Diferencia d	Número de empleados
-0.2	1
-0.1	3
0	5
0.1	6
0.2	4
0.3	3

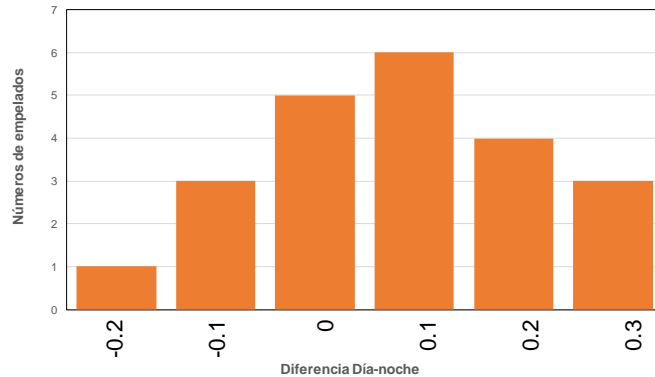


Figura 9. Distribución de datos Gaussiana o Distribución normal

Es importante señalar que cada muestra tiene menos de 30 valores.

HN: El tiempo en que cada empleado tarda en surtir una receta no es diferentes entre el turno de día con respecto al de la noche.

HA: El tiempo en que cada empleado tarda en surtir una receta es mayor en la noche que durante el día.

Nivel de significancia: $P=1\%$ (0.01)

Aunque este es un experimento controlado, la hipótesis nula y la hipótesis alterna se han eliminado en términos de causa y efecto. No tiene sentido decir que la hora del día provoca un cambio en el tiempo en que se tarda en surtir una receta; otros factores asociados con la hora del día pueden ser la causa inmediata, por ejemplo, cantidad de luz.

Se calcula la diferencia media, representada algebraicamente por \bar{d}

$$\bar{d} = \frac{\sum d}{n} = \frac{0.1 + 0.3 + 0.1 \dots + 0.1}{21} = \frac{1.5}{21} = 0.071 \text{ minutos}$$

Si la media de la diferencia, \bar{d} , resulta ser negativa, esto contradeciría inmediatamente la hipótesis alterna, ya que mostraría que el tiempo en que se surte la receta en la noche no es menor que el tiempo que toma surtir una receta en el día (en promedio); entonces no habría necesidad de seguir adelante con una prueba de una cola.



		Enfermeras																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pacientes graves	Día	4.7	2.5	3.2	1.6	4.9	4.2	5.5	3.8	7.5	6.7	6.3	4.6	6.8	3.4	1.9	1.8	4.3	7.1	5.3	2.9	4.5
	Noche	4.8	2.8	3.3	1.6	4.9	4.4	5.4	4.0	7.4	6.8	6.3	4.8	6.6	3.5	1.9	1.7	4.4	7.4	5.3	3.1	4.7
Diferencia d=día-noche		0.1	0.3	0.1	0	0	0.2	-0.1	0.2	-0.1	0.1	0	0.2	-0.2	0.1	0	-0.1	0.1	0.3	0	0.2	0.1
Cuadrado de la diferencia d ²		0.01	0.09	0.01	0	0	0.04	0.01	0.04	0.01	0.01	0	0.04	0.04	0.01	0	0.01	0.01	0.09	0	0.04	0.01

Ahora se realiza la suma de los cuadrados de las diferencias

$$= \sum d^2 = 0.01 + 0.09 + \dots 0.01 = 0.47$$

$$\begin{aligned} \text{Varianza de las diferencia de la muestra} &= s_d^2 = \frac{\sum d^2}{n} - \bar{d}^2 = \frac{0.47}{21} - 0.714^2 \\ &= 0.0224 - 0.0051 = 0.017 \text{ minutos} \end{aligned}$$

$$\text{Desviación estandar de diferencia} = s_d = \sqrt{\text{Varianza}} = \sqrt{0.017} = 0.131 \text{ minutos}$$

Se calcula el valor de la prueba de t, usando la fórmula:

$$t = \frac{\bar{d}\sqrt{n-1}}{s_d}$$

(Edmondson y Druce, 1996).

En nuestro caso:

$$t = \frac{\bar{d}\sqrt{n-1}}{s_d} = t = \frac{0.0714 \times \sqrt{21-1}}{0.131} = t = \frac{0.0714 \times 4.472}{0.131} = 2.44$$

Ahora se calcula los grados de libertad (gl):

$$gl = \text{Número de pares de valores} - 1$$

$$gl = 21 - 1 = 20$$

Se busca el valor crítico de t de una cola y con 1% de nivel de significancia en la tabla de la página X.

$$\text{Valor crítico} = t_{\text{crit}} = 2.528$$

Se rechaza la hipótesis nula y se acepta la hipótesis alterna si el valor positivo de la



prueba es más grande que el valor crítico.

En nuestro ejemplo, el valor positivo de 2.44 no es más grande que el valor crítico de $t=2.528$; por tal razón no se rechaza la hipótesis nula, es decir no hay diferencia entre el tiempo que les toma a los empleados surtir una receta en el turno de la noche con respecto a surtirla en el día; es decir no existen diferencias significativas.

Prueba Wilcoxon (prueba de rango con signo)

La prueba de Wilcoxon (prueba de rango con signo), se realiza cuando se cumplen las siguientes condiciones: Distribución de la población no normal o desconocida, cinco o más diferencias (no cero) (Edmondson y Druce, 1996).

La prueba de rangos con signo de Wilcoxon también puede ser empleada para datos ordinales previstos en ciertas condiciones.

Para comprenderlo mejor se presenta un ejemplo. El director de un hospital ha tenido quejas recurrentes de los jefes de los servicios médicos con respecto a la labor que han desempeñado dos subdirectores médicos, por desgracia solo existen dos doctores que cumplen con los requisitos para asumir el cargo, por lo cual se pide a cada jefe de servicio se evalúe a cada uno de los dos subdirectores que se han encargado del puesto.

HN: No hay diferencias entre subdirectores médicos con respecto a la calificación que obtienen de los jefes de servicios.

HA: Hay diferencias entre subdirectores médicos con respecto a la calificación que obtienen de los jefes de servicios.

Nivel de significancia $P=5\%$ (0.05).

Prueba de Wilcoxon (prueba de rango con signos) de dos colas, por qué no se está prediciendo que subdirector médico tendrá mejor evaluación por parte de los jefes de servicios.

Se calculan las diferencias de la evaluación que cada jefe de servicio ha hecho de cada uno de los subdirectores médicos.

En caso de que la diferencia tenga un valor de signo negativo se convierte a valor absoluto y se asigna el rango, iniciando con el valor más pequeño, si se repiten valores se obtiene el correspondiente promedio, por ejemplo, en el caso del valor 3, $(3+4)/2=3.5$



	Jefes de servicios											
	A	B	C	D	E	F	G	H	I	J	K	L
Director A	44	44	45	47	48	52	57	52	52	57	62	67
Director B	36	39	39	43	49	49	51	54	58	60	62	72
Rango de diferencia	44-36=8	44-39=5	45-39=5	47-43=3	48-49=-1	52-49=3	57-51=6	52-54=-2	52-58=-6	57-60=-3	62-62=0	67-72=-5
Diferencia Absoluta	8	5	6	4	1	3	6	2	6	3	0	5
Rango de diferencia	11	6.5	9	5	1 *	3.5	9	2*	9*	3.5*	-	6.5*

* valores pertenecientes a rango con signo negativo.

El siguiente paso es hacer la suma de los valores del rango R^+ , es decir los valores del rango en signo positivo:

$$R^+ = 11 + 6.5 + 9 + 5 + 9 = 44$$

De forma similar se calcula la sumatoria de los valores del rango R^- , es decir los valores del rango con signo negativo:

$$R^- = 1 + 2 + 9 + 3.5 + 6.5 = 22$$

Usando el apéndice de valores críticos de R para la prueba de rangos con signo para Wilcoxon con un nivel de significancia de 5% (0.05):

Valor crítico $R=11$

Se rechaza la hipótesis nula y se acepta la hipótesis alterna si el valor más pequeño de R^+ y R^- es menor o igual que el valor crítico de R .

Puesto que $R^+=44$ y $R^-=22$, el valor más pequeño es 22 el cual no es menor o igual que el valor crítico de $R=11$.

De este modo no rechazamos la hipótesis nula, es decir que no hay diferencias entre subdirectores médicos con respecto a la calificación que obtienen de los jefes de servicios. Entonces se dice que no hay diferencias significativas entre los subdirectores médicos con base a las evaluaciones de los jefes de servicios.



Prueba Z (muestras independientes)

Se utiliza la prueba Z (muestras independientes) cuando se satisfacen las condiciones de datos en nivel de intervalo, muestras independientes, la población se distribuye normalmente y la desviación estándar es conocida, y las muestras son grandes, más de 30 valores (Edmondson y Druce, 1996).

Se predijo que el número de pacientes atendidos en urgencias por médicos que trabajaron solos serían diferentes al número de pacientes atendidos por parejas de médicos. Se registró el promedio de pacientes atendidos por 50 médicos que trabajaron solos y de 40 que lo hicieron trabajando en pareja durante 40 minutos.

Se obtuvieron los siguientes resultados:

	Media	Desviación estándar
Médicos trabajando en pareja	30.3	1.8
Médicos trabajando solos	31.2	2.9

Se decidió usar la prueba Z de dos colas (muestras independientes) ya que el número de pacientes fue medido en niveles de intervalo, los médicos en cada muestra fueron seleccionados al azar, y ambas muestras son grandes y mayores a 30 datos.

HN: No hay diferencias significativas entre el número de pacientes atendidos por médicos que trabajan solos y médicos que trabajan en pareja.

HA: Hay diferencias significativas entre el número de pacientes atendidos por médicos que trabajan solos y médicos que trabajan en pareja.

Nivel de significancia $P=10\%$ (0.10).

Una prueba de dos colas fue seleccionada porque no se predice cuál de las muestras tendría el número de pacientes más alto, y la hipótesis alterna no tienen una tendencia.

La hipótesis nula y la hipótesis alterna no son frases que estén redactadas en términos de causa-efecto ya que esto no es controlado.

Se dejará a X represente el número de pacientes atendidos en urgencias por médicos que trabajan solos y en parejas. Ahora se calcula la diferencia entre la media de las muestras, representadas algebraicamente por $\bar{x} - \bar{y}$

Así $s_x = 30.3$ pacientes y $\bar{y} = 31.2$ pacientes, entonces:

Diferencias entre medias $= \bar{x} - \bar{y} = 30.3 - 31.2 = -0.9$

El cuadrado de la desviación estándar de las muestras se obtienen sus varianzas



representadas algebraicamente como s_x^2 y s_y^2

Así que $s_x = 1.8$ y $s_y = 2.9$ entonces:

$$s_x^2 = 1.8^2 = 3.24 \quad y \quad s_y^2 = 2.9^2 = 8.41$$

Divide cada varianza por el tamaño de la muestra correspondiente

Si n_x y n_y representan el tamaño de dos muestras entonces:

$$\frac{s_x^2}{n_x} = \frac{3.24}{40} = 0.081 \quad y \quad \frac{s_y^2}{n_y} = \frac{8.41}{50} = 0.1682$$

Ahora se obtiene la estimación de la varianza combinada representada algebraicamente como:

$$\text{Varianza combinada} = \hat{\sigma}^2 = \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} = 0.081 + 0.1682 = 0.2492$$

Se calcula la raíz cuadrada de la varianza combinada

$\hat{\sigma}^2$, representada algebraicamente como $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = \sqrt{0.2492} = 0.4992$$

Se calcula el valor de la prueba Z dividido en la diferencia entre las medias, $\bar{x} - \bar{y}$, por el valor de $\hat{\sigma}$

$$\text{valor de la prueba} = Z = \frac{\bar{x} - \bar{y}}{\hat{\sigma}} = \frac{-0.9}{0.4992} = -1.80$$

Si el valor de la prueba Z es negativo, se convierte al valor positivo, en este caso $Z = -1.80$ es negativo, se convierte a su valor en positivo, es decir 1.80

Para una prueba de dos colas, se reduce a la mitad el nivel de significación, se resta el nivel de significancia, a esto se le llama resultado de la probabilidad.

Así, el nivel de significancia es 10%, entonces,

$$\frac{10\%}{2} = 5\%, \text{ y así:}$$

Probabilidad = $100\% - 5\% = 95\%$ (0.95)

Para comprender mejor veamos, utilizando la tabla de valor crítico de Z, correspondiendo a la probabilidad 0.95

La columna de la izquierda y la fila superior de la tabla de Z dan el valor de Z_{crit} (la fila superior da el segundo lugar decimal de Z_{crit})

Las probabilidades de la mayor parte de la tabla son correctas hasta los cuatro decimales.

La probabilidad más cercana a 0.95 en las tablas es 0.9495 correspondiente a un valor de $Z_{\text{crit}} = 1.64$



Z	0.00	0.01	0.02	0.03	0.04	0.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
.1	.5398	.5438	.5478	.5517	.5557	.5596
.2	.5793	.5832	.5871	.5910	.5948	.5987
.3	.6179	.6217	.6255	.6293	.6331	.6368
.4	.6554	.6591	.6628	.6664	.6700	.6736
.5	.6915	.6950	.6985	.7019	.7054	.7088
.6	.7257	.7291	.7324	.7357	.7389	.7422
.7	.7580	.7611	.7642	.7673	.7704	.7734
.8	.7881	.7910	.7939	.7967	.7995	.8023
.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
.1	.8643	.8665	.8686	.8708	.8729	.8749
.2	.8849	.8869	.8888	.8907	.8925	.8944
.3	.9032	.9049	.9066	.9082	.9099	.9115
.4	.9192	.9207	.9222	.9236	.9251	.9265
.5	.9332	.9345	.9357	.9370	.9382	.9394
.6	.9452	.9463	.9474	.9484	.9495	.9505
.7	.9554	.9564	.9573	.9582	.9591	.9599
.8	.9641	.9649	.9656	.9664	.9671	.9678
.9	.9713	.9719	.9726	.9732	.9738	.9744
2.0	.9772	.9778	.9783	.9788	.9793	.9798
.1	.9821	.9826	.9830	.9834	.9838	.9842

Un valor más preciso de Z_{crit} es 1.645 la mitad entre 1.64 y 1.65 (ver tabla anterior).

Se rechaza la hipótesis nula y se acepta la hipótesis alterna si el valor positivo de Z es más grande que el valor crítico Z_{crit} .

Entonces el valor positivo de prueba de 1.80 es más grande que el valor crítico $Z_{crit}=1.64$, se debe rechazar la hipótesis nula y aceptar la hipótesis alterna. Por lo tanto, hay diferencias en el número de pacientes atendidos en urgencias por parte de médicos que trabajan solos y quienes trabajan en pareja.

Así que el valor positivo de la prueba de 1.8 es más grande que el valor crítico de $Z_{crit}=1.64$, la hipótesis nula debería haber sido rechazada si el valor crítico, Z_{crit} , fue ligeramente menor que 1.8. En la tabla se puede ver que el valor crítico de 1.80 corresponde a .9641, o 96.41%

De este modo se puede encontrar en el nivel de significancia P, correspondiendo a este valor crítico:

$$100\% - 96.41\% = 3.59\%$$

$$2 \times 3.59\% = 7.18\% \approx 7.5\% \text{ (redondeado al 0.5\% más cercano)}$$

Así, la hipótesis nula debería ser rechazada si el nivel de significancia P se había fijado al menos en 7.5% (0.075), en lugar de 10% (0.10).

Establecemos este nivel mínimo de significación P como:

“Rechazamos la hipótesis nula para $P < 0.075$ ” o simplemente “ $P < 0.075$ ”



Prueba t (muestras independientes)

Se usa la prueba de t para muestras independientes, la población debe presentar una distribución más o menos normal, la población debe tener aproximadamente la misma desviación estándar o la desviación estándar es desconocida.

El modelo es definido como:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\bar{x}_1 - \bar{s}_2}$$

Ahora revisemos un ejemplo en el que se puede aplicar esta prueba. Se presentan los datos de millones de pesos del coste de detección, diagnóstico y tratamiento para un mes para el mismo número de pacientes con trastorno metabólico empelado por 13 hospitales de la Ciudad de México.

Las hipótesis para la prueba de t para muestras independientes (dos colas) son:

$$HN = \mu_1 - \mu_2 = 0$$

$$HA = \mu_1 - \mu_2 \neq 0$$

Es decir, que el coste de la detección, diagnóstico y tratamiento para un mes del número de pacientes (mismo número) cobrado por el distribuidor 1 tiene en promedio el mismo coste del que es cobrado por el distribuidor 2, para los 13 hospitales a los que se les proporciona.

Esta hipótesis es comúnmente expresada como:

$$HN = \mu_1 = \mu_2$$

$$HA = \mu_1 \neq \mu_2$$

Se presentan los datos

	Millones de pesos por el distribuidor 1	Millones de pesos por el distribuidor 2	$\bar{x}_1 - x_i$ Distribuid or 1	$\bar{\mu}_2 - x_i$ Distribuid or 2	$(\bar{x}_1 - x_1)^2$	$(\bar{x}_2 - x_2)^2$
	8.8	9.9	0.05	0.157	0.003	0.025
	8.4	9	-0.35	-0.743	0.123	0.552
	7.9	11.1	-0.85	1.357	0.722	1.842
	8.7	9.6	-0.05	-0.143	0.003	0.020
	9.1	8.7	0.35	-1.043	0.123	1.088



	9.6	10.4	0.85	0.657	0.722	0.432
		9.5		-0.243		0.059
n	6	7				
v	5	6				
media	8.75	9.742857143				
SS					1.695	4.017

$$s_p^2 = \frac{SS_1 + SS_2}{v_1 + v_2} = s_p^2 = \frac{1.695 + 4.017}{5 + 6} = \frac{5.712}{11} = 0.519$$

$$S_{\bar{x}_1} - \bar{x}_2 = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$S_{\bar{x}_1} - \bar{x}_2 = \sqrt{\frac{0.519}{6} + \frac{0.519}{7}}$$

$$S_{\bar{x}_1} - \bar{x}_2 = \sqrt{0.087 + 0.074}$$

$$S_{\bar{x}_1} - \bar{x}_2 = \sqrt{0.161}$$

$$= 0.40$$

$$t = \frac{8.75 - 9.74}{0.40}$$

$$t = \frac{-0.90}{0.40}$$

$$= -2.475$$

$$t_{0.05(2)v} = t_{0.05(2)11} = 2.201$$



∴ Rechazamos la hipótesis nula

Esto significa que el coste de la detección, diagnóstico y tratamiento para un mes del número de pacientes (mismo número en todos los hospitales) cobrado por el distribuidor 1 no tiene en promedio el mismo coste del que es cobrado por el distribuidor 2. Es decir, existen diferencias estadísticamente significativas ambos distribuidores.

Regresión lineal

Frecuentemente dos variables (una dependiente y otra independiente) mantienen una relación que describe un comportamiento de una recta, la cual se describe por el modelo:

$$y = mx + b$$

Donde:

La Constante b (ordenada al origen)

m, pendiente

x, valor que toma la variable independiente

b, ordenada al origen

y, valor que toma la variable dependiente y que es el resultado con respecto a las variaciones de la variable x.

Y, por ende, el valor que se pretende hallar es la variable y, es decir la variable dependiente.

El método más efectivo y sencillo para determinar los parámetros m y b se conoce método de mínimos cuadrados que mejor se ajuste a los datos.

Los modelos para estimar la pendiente y la ordenada al origen son:

$$m = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{(\sum x_i^2) - (\sum x_i)^2}$$

$$b = \frac{(\sum y_i) - a(\sum x_i)}{n}$$

Donde:

n es el número de datos.



Σ representa la suma de todos los datos.

Las variables pueden presentar diferentes tendencias y grado de correlación, como se observa en la figura 10, graficar los datos permite saber a priori el signo que tendrá la pendiente, y con los cálculos que se describen más abajo se obtendrán los valores para b, m y R^2 de la recta que mejor se ajuste.

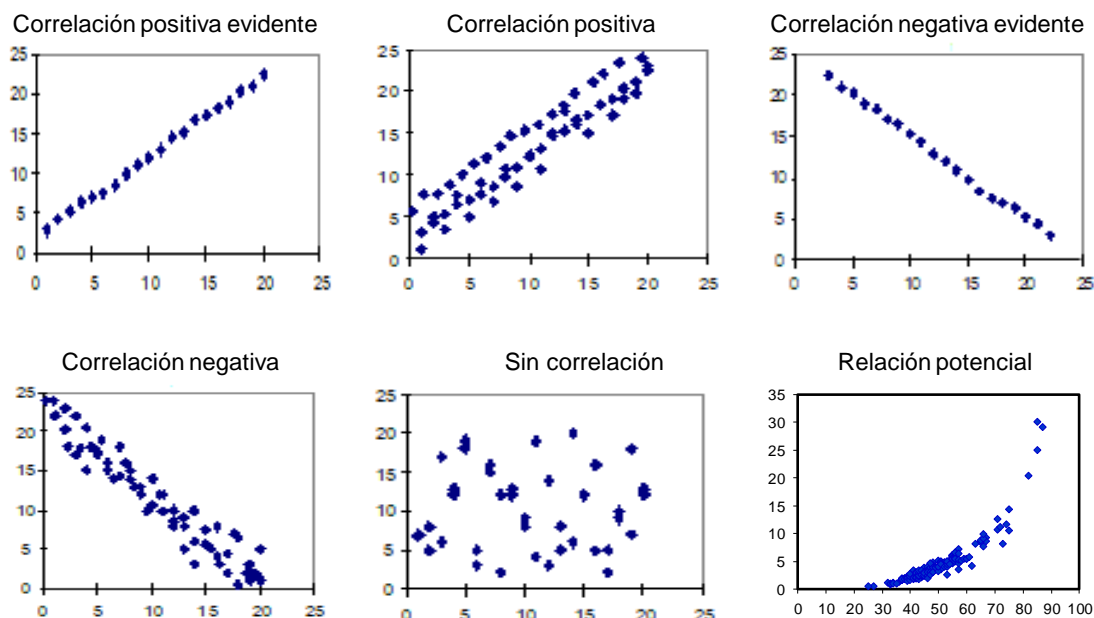
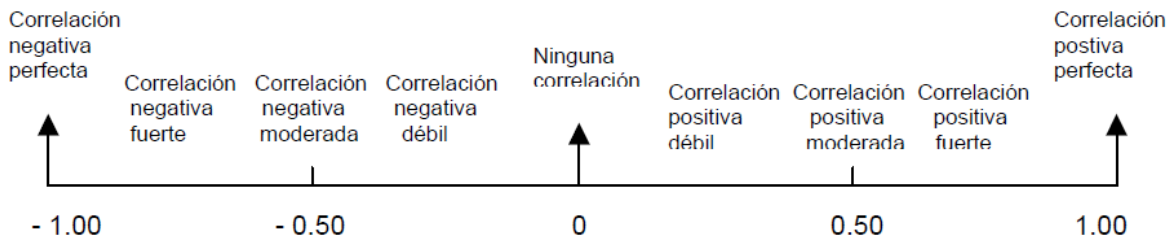


Figura 10. Tipos de correlación entre dos variables (correlación positiva, correlación negativa sin correlación, y

Por otra parte, el coeficiente de correlación es otro parámetro para el estudio de la relación entre variables que describen una relación de la recta e indica el grado de dependencia entre las variables x e y. El coeficiente de correlación r es un número que se obtiene mediante el modelo:

$$r = \frac{n (\sum x_i y_i) - (\sum x_i) (\sum y_i)}{\sqrt{[n (\sum x_i^2) - (\sum x_i)^2] [n (\sum y_i^2) - (\sum y_i)^2]}}$$

El valor del coeficiente de correlación oscila entre 1 y -1, cada valor del coeficiente de correlación representa el grado de dependencia entre las variables, como se observa a continuación.



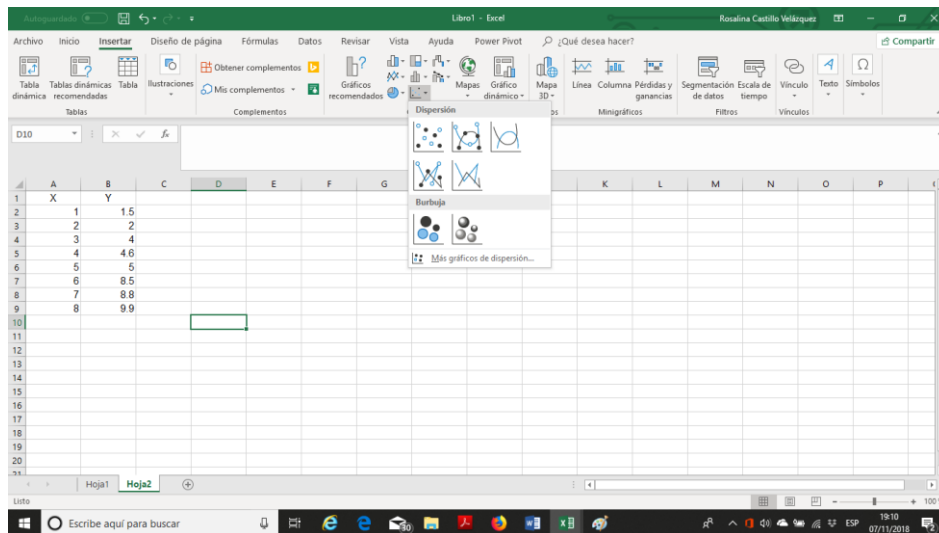
Afortunadamente en la actualidad utilizando programas como Excel se puede obtener la relación de la recta y el coeficiente de correlación de forma sencilla, como se explica a continuación ejemplificándolo con un conjunto de datos:

x	y
1	1.5
2	2
3	4
4	4.6
5	5
6	8.5
7	8.8
8	9.9

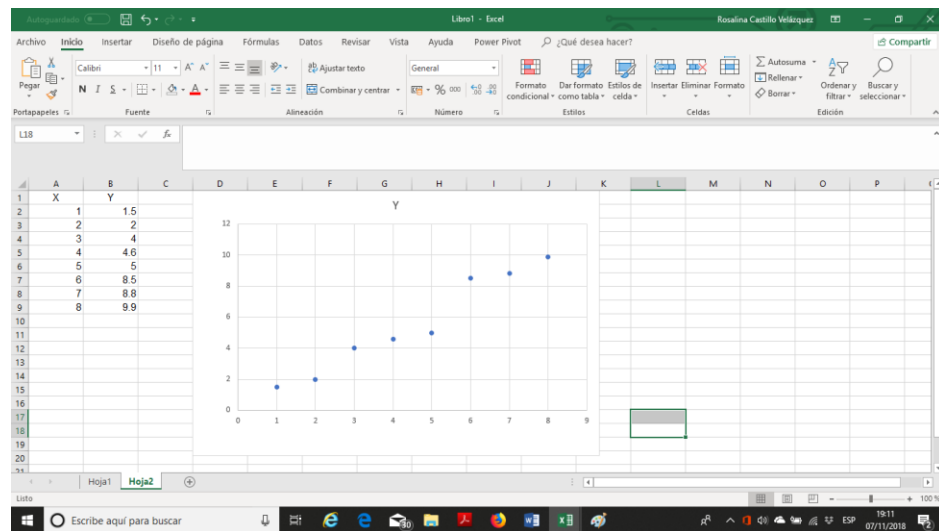
Se seleccionan los datos, se elige la opción insertar y se elige dispersión.

Gestión de servicios de salud

Guía para el análisis de datos del Proyecto terminal



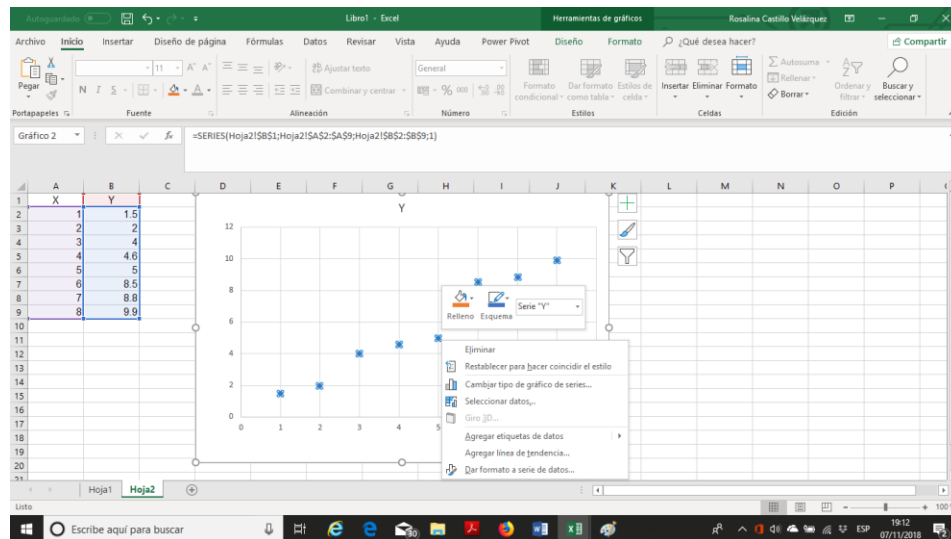
El resultado es:



Se coloca el curso sobre los puntos de la gráfica y se oprime el botón derecho, seleccionando agregar línea de tendencia, eligiendo la opción presentar ecuación en el gráfico, y presentar el valor de R cuadrado en el gráfico.

Gestión de servicios de salud

Guía para el análisis de datos del Proyecto terminal



Obteniéndose la figura 11 que evidentemente tiene el comportamiento de una recta con signo positivo.

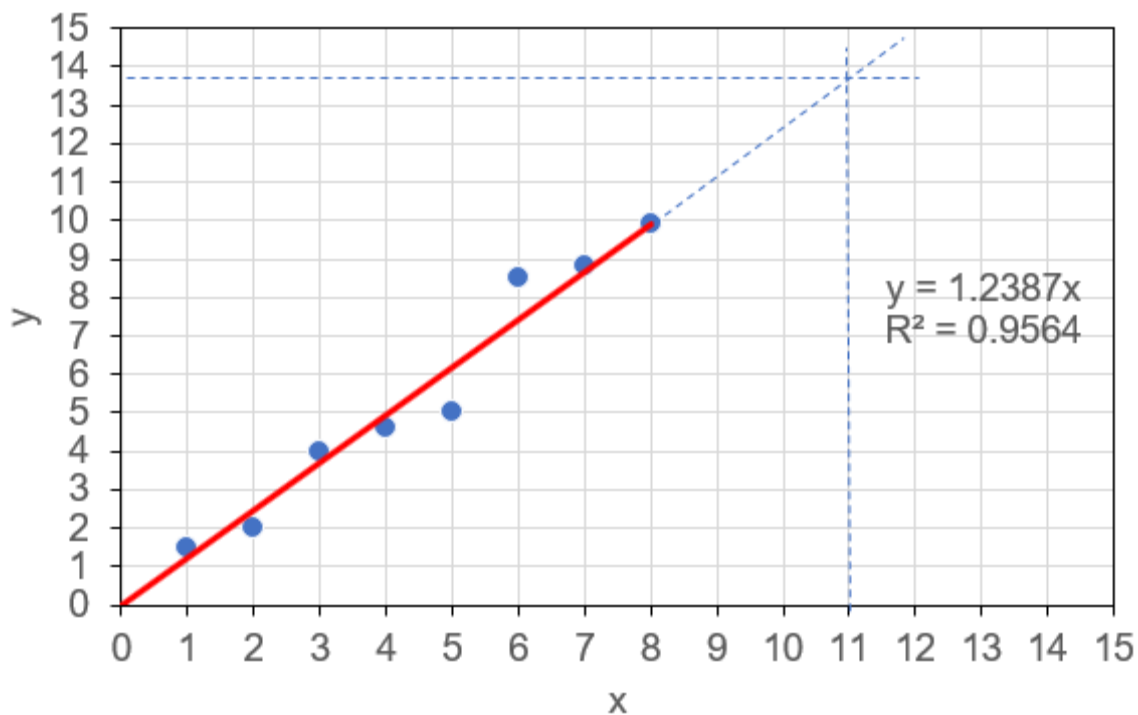


Figura 11. Relación entre variables que describe el comportamiento de una recta.

De este modo al tener la ecuación de la recta que mejor se ajusta a los datos podemos estimar valores para y con base a cualquier valor de x. Podemos estimar el valor de y para cualquier x de dos formas, la primera de forma matemática y que es la más



recomendable y de forma gráfica, revisemos ambos métodos:

En este caso para el conjunto de datos se obtuvo con ayuda de Excel la ecuación de la recta $y = 1.2387x$

Para este ejemplo el valor de m (la pendiente) fue 1,2387; el valor de b (ordenada al origen) es cero, por eso no aparece en la ecuación, si obligatoriamente se incluyera quedaría la ecuación como:

$$y = 1.2387x + 0$$

Entonces, al hacer la figura en Excel e incluir la línea de tendencia ésta aparece desde cero hasta 8 (línea sólida en rojo), que es el valor de x más grande que el conjunto de datos tiene, pero si se quisiera estimar el valor de y para $x=11$, entonces sustituimos en la fórmula:

$$y = 1.2387(11)$$

$$y = 13.6257$$

Si se desea estimar el valor de y para $x=11$ de forma gráfica, entonces prolongamos la recta más allá del valor que toma en $x=8$, en la figura 11 se hizo en línea azul punteada, y en el valor deseado de x para el cual se quiere conocer y se sube la línea de forma paralela al eje y, donde corta con la recta se lleva al eje y, al hacerlo podemos observar que efectivamente, el corte de esta última recta en el eje y es muy cercano al valor estimado de forma matemática, es decir a 13.6.

Comparación de dos rectas

En algunos casos los datos que se tienen sobre un fenómeno describen un comportamiento lineal cuando son graficados, en el caso de que se quiera comparar tal comportamiento entre dos poblaciones, no es suficiente observar la gráfica y determinar a priori si tienen o no la misma pendiente, ¿cómo se puede saber a ciencia cierta que la pendiente de cada una de estas rectas es la misma? Para ello existe un método simple para probar la hipótesis sobre la igualdad de dos coeficientes de regresión de dos poblaciones, e involucra el uso de t de Student, de forma análoga a la prueba de diferencia entre dos medias poblacionales.

$$t = \frac{b_1 - b_2}{S_{b_1} - S_{b_2}}$$

Donde el error estándar de la diferencia entre los coeficientes de regresión está dado por:

$$S_{b_1} - S_{b_2} = \sqrt{\frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_1} + \frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_2}}$$



(Zar, 1984).

Veamos un ejemplo, para su mejor comprensión. Se tiene el registro del cómo se va haciendo uso de los recursos destinados a insumos de papelería durante el año 2014 en una clínica de atención primaria, sin embargo, se decidió implementar para el gasto de 2015 medidas de control estricto en el proceso de compra de tales insumos, en el año 2016 se analizarán los datos de ambas muestras para determinar si los nuevos procedimientos muestran diferencias significativas con respecto año 2014 cuando no existía tal control.

	X	Y ₁	Y ₂	X ₂	Y ₁ ²	Y ₂ ²	XY ₁	XY ₂
	1	7.5	6.3	1	56.25	39.69	7.5	6.3
	2	7.2	6.2	4	51.84	38.44	14.4	12.4
	3	6.8	6	9	46.24	36	20.4	18
	4	6.5	5.75	16	42.25	33.0625	26	23
	5	5.8	5.7	25	33.64	32.49	29	28.5
	6	5.3	5.5	36	28.09	30.25	31.8	33
	7	5.2	5.3	49	27.04	28.09	36.4	37.1
	8	4.6	5	64	21.16	25	36.8	40
	9	4.3	4.8	81	18.49	23.04	38.7	43.2
	10	4.2	4.5	100	17.64	20.25	42	45
	11	4	4.4	121	16	19.36	44	48.4
	12	3.8	4	144	14.44	16	45.6	48
SUMA	78	65.2	63.45	650	373.08	341.6725	372.6	382.9
(Sum X) ²	6084						138830.76	146612.41

Se obtuvo con los datos de ambas muestras comportamiento de rectas que se observan en la figura 12, ambas con signo negativo.

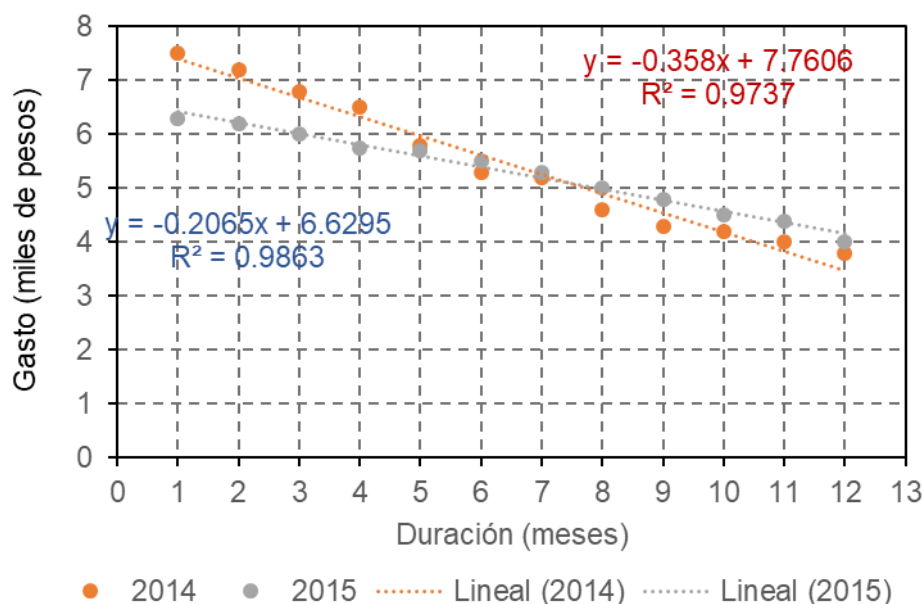


Figura 12. Datos con comportamiento de rectas para el gasto en 12 meses consecutivos aplicando un protocolo de compras para el insumo de papelería de una clínica de atención primaria.

Se plantean las hipótesis:

HN: $B_1=B_2$

HA: $B_1 \neq B_2$

Disminución del presupuesto asignado a compra de insumos de papelería durante 2014

$$\sum x^2 = 650$$

$$\sum xy_1 = 372.6$$

$$\sum y_1^2 = 373.08$$

$$(\sum y_1)^2 = 138830.76$$

$$n = 12$$

$$b = \frac{\sum xy}{\sum x^2} \rightarrow b = \frac{372.6}{650} = 0.57$$

$$Residual SS_1 = \sum y_1^2 - \frac{(\sum xy_1)^2}{\sum x^2}$$

$$Residual SS_1 = 373.08 - \frac{372.6^2}{650} = 372.51$$

Disminución del presupuesto asignado a compra de insumos de papelería durante 2015

$$\sum x^2 = 650$$

$$\sum xy = 382.9$$

$$\sum y^2 = 341.67$$

$$(\sum y_2)^2 = 14661.41$$

$$n = 12$$

$$b = \frac{\sum xy}{\sum x^2} \rightarrow b = \frac{382.9}{650} = 0.60$$

$$Residual SS_1 = \sum y_2^2 - \frac{(\sum xy_2)^2}{\sum x^2}$$

$$Residual SS_1 = 341.67 - \frac{341.67^2}{650} = 341.14$$



$$\text{Residual } GL_1 = n - 2 \rightarrow 12 - 2 = 10$$

$$\text{Residual } GL_1 = n - 2 \rightarrow 12 - 2 = 10$$

$$(S_{y \cdot x}^2)_p = \frac{\text{Residual } SS_1 + \text{Residual } SS_2}{\text{Residual } GL_1 + \text{Residual } GL_2}$$

$$(S_{y \cdot x}^2)_p = \frac{372.51 + 341.14}{10 + 10} = 35.68$$

$$S_{b_1} - S_{b_2} = \sqrt{\frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_1} + \frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_2}}$$

$$S_{b_1} - S_{b_2} = \sqrt{\frac{35.58}{650} + \frac{35.58}{650}}$$

$$S_{b_1} - S_{b_2} = 0.1095$$

$$t = \frac{b_1 - b_2}{S_{b_1} - S_{b_2}}$$

$$t = \frac{0.57 - 0.60}{0.1095} = -0.2739, \text{ valor absoluto} = 0.2739$$

$$GL = 20$$

$$\text{Rechazo la } H_0 \text{ si } |t| \geq t_{\alpha(2), GL}$$

$$t_{0.05(2), 22} = 2.086$$

∴ Acepto la H_0 , dado que la t estimada (0.2739) es menor que la t de tablas (2.086); es decir $H_0: \beta_1 = \beta_2$

En conclusión, el gasto en consumibles de papelería no varió en 2015 con respecto a cómo se gastó en 2014.

Se muestra un ejemplo para su mejor comprensión. La secretaria de salud está revisando la asignación presupuestaria para el año 2019 de dos clínicas de primer nivel en la zona Raramuri de Chihuahua, y se ha pedido a los responsables de las clínicas que se envíe información del presupuesto que han recibido durante los 15 años previos únicamente

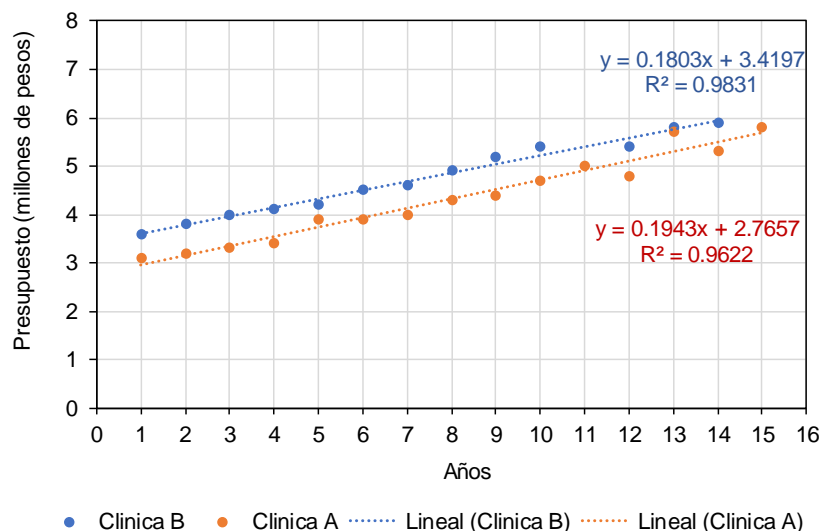


para equipamiento, insumos, y mantenimiento, las dos clínicas tienen los mismos servicios, equipo, personal administrativo, número de médicos, número de enfermeras, y atienden aproximadamente a un número similar de pacientes, la única diferencia es que la clínica B no tiene los datos de dos años debido a que extraviaron documentos probatorios de la asignación de los recursos que recibieron, a continuación se muestran los datos:

X(año)		Y ₁ (millones de pesos)	Y ₂ (millones de pesos)	X ²	Y ₁ ²	Y ₂ ²	XY ₁	XY ₂
2004	1	3.6	3.1	1	12.96	9.61	3.6	3.1
2005	2	3.8	3.2	4	14.44	10.24	7.6	6.4
2006	3	4	3.3	9	16	10.89	12	9.9
2007	4	4.1	3.4	16	16.81	11.56	16.4	13.6
2008	5	4.2	3.9	25	17.64	15.21	21	19.5
2009	6	4.5	3.9	36	20.25	15.21	27	23.4
2010	7	4.6	4	49	21.16	16	32.2	28
2011	8	4.9	4.3	64	24.01	18.49	39.2	34.4
2012	9	5.2	4.4	81	27.04	19.36	46.8	39.6
2013	10	5.4	4.7	100	29.16	22.09	54	47
2014	11		5	121	0	25	0	55
2015	12	5.4	4.8	144	29.16	23.04	64.8	57.6
2016	13	5.8	5.7	169	33.64	32.49	75.4	74.1
2017	14	5.9	5.3	196	34.81	28.09	82.6	74.2
2018	15		5.8	225	0	33.64	0	87
Sumatoria	120	61.4	64.8	1240	297.08	290.92	482.6	572.8
Media	8	4.72	4.32					



Al graficar los datos se obtuvo la siguiente figura, donde ambas rectas son similares con respecto a sus pendientes, sin embargo, para corroborarlo se debe aplicar una prueba estadística que permite afirmar si hay o no diferencias estadísticamente significativas.



Presupuesto de la clínica B en los últimos años

$$n = 13$$

$$\bar{X} = 8$$

$$\bar{Y} = 4.72$$

$$\sum x^2 = 1240$$

$$\sum xy_1 = 482.6$$

$$\sum y_1^2 = 297.1$$

$$b = 0.1813$$

$$a = 3.4197$$

$$Residual SS_1 = \sum y_1^2 - \frac{\sum xy_1^2}{\sum x^2}$$

$$Residual SS_1 = 297 - \frac{482.6^2}{1240} = 109.275$$

$$Residual GL_1 = 13 - 2 = 11$$

Presupuesto de la clínica A en los últimos años

$$n = 15$$

$$\bar{X} =$$

$$\bar{Y} = 4.32$$

$$\sum x^2 = 1240$$

$$\sum xy = 572.8$$

$$\sum y^2 = 290.9$$

$$b = 0.1943$$

$$a = 2.7657$$

$$Residual SS_1 = \sum y_2^2 - \frac{\sum xy_2^2}{\sum x^2}$$

$$Residual SS_1 = 290.8 - \frac{3572.8^2}{612400} = 26.303$$

$$Residual GL_1 = 15 - 2 = 13$$



$$(S_{y \cdot x}^2)_p = \frac{\text{Residual } SS_1 + \text{Residual } SS_2}{\text{Residual } GL_1 + \text{Residual } GL_2}$$

$$(S_{y \cdot x}^2)_p = \frac{109275 + 26.303}{11 + 15} = 5.649$$

$$S_{b_1} - S_{b_2} = \sqrt{\frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_1} + \frac{(S_{y \cdot x}^2)_p}{(\sum x^2)_2}}$$

$$S_{b_1} - S_{b_2} = \sqrt{\frac{35.58}{650} + \frac{35.58}{650}}$$

$$S_{b_1} - S_{b_2} = 0.1095$$

$$t = \frac{b_1 - b_2}{S_{b_1} - S_{b_2}}$$

$$t = \frac{0.57 - 0.60}{0.1095} = -0.2739, \text{ valor absoluto} = 0.2739$$

$$GL = 24$$

$$\text{Rechazo la } H_0 \text{ si } |t| \geq t_{\alpha(2), GL}$$

$$t_{0.05(2), 24} = 2.064$$

\therefore No rechazo la H_0 , dado que la t estimada $|0.1361|$ es menor que la t de tablas (2.064); es decir $H_0: \beta_1 = \beta_2$

En conclusión, el presupuesto asignado en los últimos 15 años a ambas clínicas se ha incrementado de forma equitativa, no mostrando diferencias estadísticamente significativas.

Análisis de varianza de un factor o de una vía

El análisis de varianza (ANOVA) de una vía se utiliza para estimar si existen diferencias



estadísticamente significativas entre las medias de tres o más grupos.

Para saber si los grupos tienen comportamiento distinto, es decir conocer entre qué grupos hay o no diferencias estadísticamente significativas se realizan pruebas *post hoc*.

Los supuestos que deben cumplirse para realizar un ANOVA de una vía son: La distribución de probabilidad de la variable dependiente correspondiente a cada factor debe ser normal; las muestras de cada tratamiento son independientes; las poblaciones tienen varianza igual, es decir existe homoscedasticidad (Zar, 1984).

Ahora hagamos un ejercicio para utilizar ANOVA de una vía.

Se muestran los datos obtenidos en un experimento donde a 19 empleados del archivo de un hospital de tercer nivel se les pide que utilicen cuatro diferentes protocolos de apertura de expediente, revisión de vigencia de derechos y entrega del expediente en la consulta externa de oncología, los datos que se obtienen se refieren al tiempo en minutos y se muestran en la tabla de abajo. En este caso la hipótesis nula sería $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$.

	Protocolo 1	Protocolo 2	Protocolo 3	Protocolo 4
	60.8	68.7	102.6	87.9
	57.0	67.7	102.1	84.2
	65.0	74.0	100.2	83.1
	58.6	66.3	96.5	85.7
	61.7	69.8		90.3
n_i	5	5	4	5
$\sum_{j=1}^{n_i} X_{ij}$	303.10	346.50	401.40	431.20
\bar{X}_i	60.62	69.30	100.35	86.24
$\frac{(\sum_{j=1}^{n_i} X_{ij})^2}{n_i}$	18373.922	24012.450	40280.490	37186.688

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

(Zar, 1984).

H_A : La media del tiempo que los empleados de archivo emplean en cada tipo de protocolo



no son todos iguales

$$\sum_{i=1}^k \frac{(\sum_{j=1}^{n_i} X_{ij})^2}{n_i} = 119853.550$$

$$\sum_i \sum_j X_{ij} = 1482.200$$

$$\sum_i \sum_j X_{ij}^2 = 119981.900$$

Total DF= N-1=19-1=18

Grupos DF= k-1=4-1=3

Error DF= N-k= 19-4=15

DF, son los grados de libertad, esta abreviatura es por cómo se escribe en el idioma inglés degree of freedom, es según Levin 1996 “Se definen como el número de valores que podemos escoger libremente”.

$$C = \frac{(\sum_i \sum_j X_{ij})^2}{N} = 115627.202$$

$$C = \frac{(1482.2)^2}{19} = 115627.201$$

Suma total de cuadrados (SS):

$$\sum_i \sum_j X_{ij}^2 - C = 1482.200$$

Suma de grupos al cuadrado:

$$= \sum_i \frac{(\sum_j X_{ij})^2}{n_i} - C = 119853.550 - 115627.202 = 4226.348$$

Suma de errores al cuadrado:

$$\text{Total SS- grupos SS} = 4354.698 - 4226.348 = 128.350$$

En resumen, tenemos:

Fuente de varianza	SS	DF	MS
Total	4354.698	18	



Grupos	4226.348	3	1408.783
Error	128.350	15	8.557

$$F = \frac{\text{Grupos } MS}{\text{error } MS}$$

$$F = \frac{1408.783}{8.557} = 165$$

$$F_{0.05, (1), 3, 15} = 3.29$$

Se rechaza la H_0 .

$$P < 0.0005$$

Es decir, el tiempo promedio en minutos de los cuatro diferentes protocolos de apertura de expediente, revisión de vigencia de derechos y entrega del expediente en la consulta externa de oncología son diferentes.

Sin embargo, no sabemos entre que pares de protocolo existen las diferencias es decir existen cuatro posibilidades de diferencias:

$$\mu_1 \neq \mu_2$$

$$\mu_1 \neq \mu_3$$

$$\mu_1 \neq \mu_4$$

$$\mu_2 \neq \mu_3$$

$$\mu_2 \neq \mu_4$$

$$\mu_3 \neq \mu_4$$

Para determinar entre cuales pares de medias existen diferencias estadísticamente significativas se aplican otras pruebas estadísticas

Símbolos y operadores matemáticos

$|x|$ Barras paralelas significa valor absoluto



- ≥ Mayor o igual que
- ≤ Menor o igual que
- ≠ Diferente de
- ∴ Por lo tanto
- Σ Sumatoria

Referencias

- Asurza O., H. (mayo de 2006). *Glosario básico de términos estadísticos*. Obtenido de Instituto Nacional de estadística e Informática:
https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib0900/Libro.pdf
- CanStockPhoto, 2017. Selección, Estadística, Investigación, Muestra, Metodología, Concepto, Encuesta, Población. Recuperado de
<https://www.canstockphoto.es/selecci%C3%B3n-estad%C3%ADstica-investigaci%C3%B3n-49453612.html>
- Cuesta, M., y Herrero, F. (s.f.). Universidad de Oviedo. *Introducción al muestreo*. Obtenido de <http://mey.cl/apuntes/muestrasunab.pdf>
- Edmondson, A. y Druce, D. (1996). *Advanced Biology Statistics*. College of North Press. Oxford. 176pp.
- Fernández, P., y Pértiga Díaz, S. (6 de marzo de 2001). *Estadística descriptiva de los datos*. Obtenido de
<https://www.fisterra.com/mbe/investiga/10descriptiva/10descriptiva2.pdf>
- Grané Chávez, A. (s.f.). Departamento de estadística. *Introducción*. Obtenido de Universidad carlos III de Madrid:
http://www.est.uc3m.es/agrane/ficheros_docencia/EDAD/introduccion_tema1_reducido.pdf
- Levin R, R UBIND. *Estadística para administradores*. México: Prentice-Hall. 1996.
- Moreno, L. (s. f.) Normalidad, Departamento de Salud pública. Facultad de medicina UNAM. Recuperado de
http://paginas.facmed.unam.mx/deptos/sp/wp-content/uploads/2015/10/U6_matcompl_morenoalta_epiclin.pdf
- Murray S. y L. Stephens. (2008). *Estadística*. McGraw-Hill Companies. México. 577p.
- Oda, N. B., (2005). *Introducción al análisis gráfico de datos experimentales*.



México. Facultad de Ciencias, UNAM.

Random Notes (2010). Muestras pareadas y muestras no pareadas. Recuperado de <http://aleatorynotes.blogspot.com/2010/11/muestras-pareadas-y-muestras-no.html>

Soporte Minitab 18. (2017). Recuperado de <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/calculations-data-generation-and-matrices/calculator/calculator-functions/statistics-calculator-functions/percentile-function/>

XLStat. (2017). ¿Cuál es la diferencia entre una prueba de dos colas (bilateral) y de una cola (unilateral)?

Zar, J. (1984). Biostatistical Analysis. Prentice Hall. New Jersey. 718pp