4.1 Conceptos básicos de Estadística I

Importancia de la Estadística en la Investigación en Salud. Análisis descriptivo de los individuos de la muestra. Métodos para obtener conclusiones sobre la Media de una variable cuantitativa en la población, a partir de la información proporcionada por los datos de la muestra.

Autora: Inma Jarrín Vera

Coordinadora de Estadística del Máster en Salud Pública

Se recomienda imprimir 2 páginas por hoja

Citación recomendada:

Jarrín Vera I. Conceptos básicos de Estadistica I [Internet]. Madrid: Escuela Nacional de Sanidad; 2012 [consultado día mes año]. Tema 4.1 Disponible en: direccion url del pdf.

Resumen:

Este tema comienza enfatizando la importancia que los Métodos Estadísticos tienen en la Investigación en Salud, tanto en las etapas de diseño como en la selección de muestras y análisis de datos.

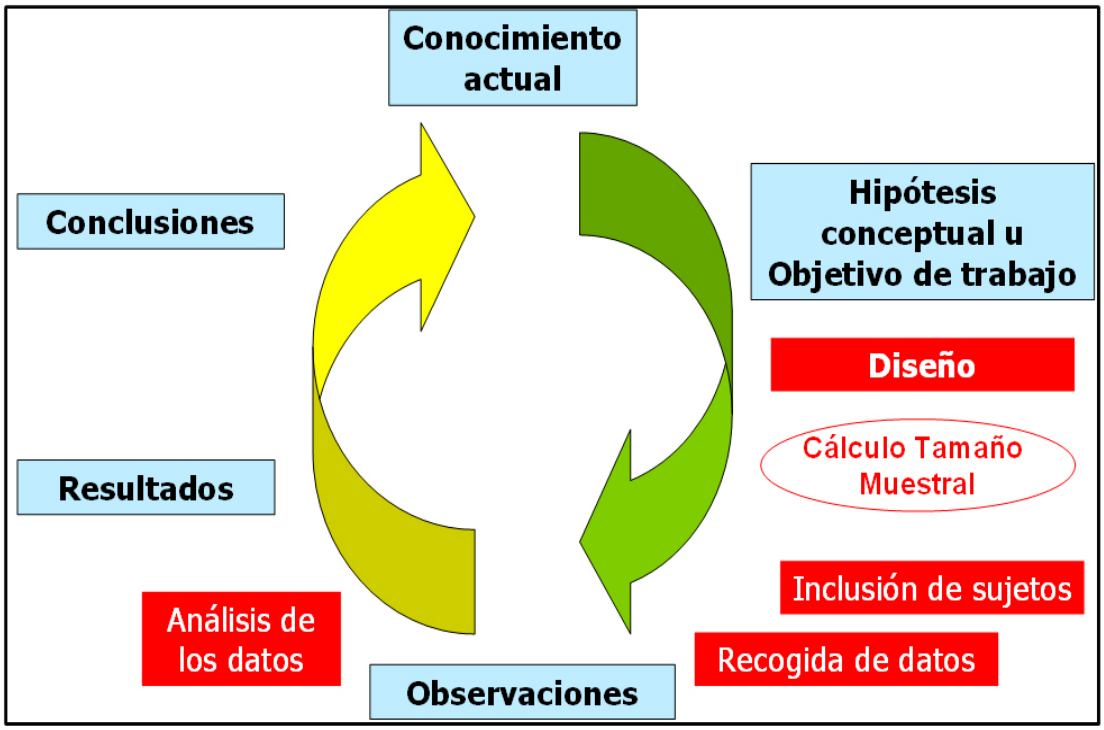
Se distingue entre población, muestra y muestra aleatoria. Se describen los diferentes tipos de variables, en función de su escala de medida y de su papel en el estudio. Se presentan los métodos,

numéricos y gráficos, para realizar el análisis descriptivo de los individuos de la muestra. Se presentan los métodos analíticos utilizados en el análisis de variables de interés cuantitativas. En este sentido, se presentan los métodos que nos permitirán obtener conclusiones sobre la media de una variable cuantitativa en la población a estudio, a partir de la información proporcionada por los datos de la muestra. Como parte de este proceso de Inferencia Estadística, se presenta la Distribución Normal. Se describen los métodos para comparar la media de una variable cuantitativa en 2 ó más grupos de individuos. Se introducen los conceptos de correlación y regresión lineal para el estudio de la relación entre dos variables cuantitativas. Y, finalmente, se presentan los métodos no paramétricos, como alternativa a los métodos paramétricos, para el análisis de variables de interés cuantitativas que no siguen una distribución aproximadamente Normal.

Introducción

Los Métodos Estadísticos son una herramienta fundamental en la Investigación en Salud, tanto en las etapas de diseño como en los procesos de selección de muestras y análisis de datos.

Los métodos estadísticos proporcionan herramientas básicas para la descripción y cuantificación de los procesos de salud y enfermedad, convirtiéndose en una disciplina imprescindible para los estudios en salud.



Proporcionan herramientas básicas para la descripción y cuantificación de los procesos de salud y enfermedad, convirtiéndose en una disciplina imprescindible para los estudios en salud.

1.- Poblaciones, muestras y muestras aleatorias

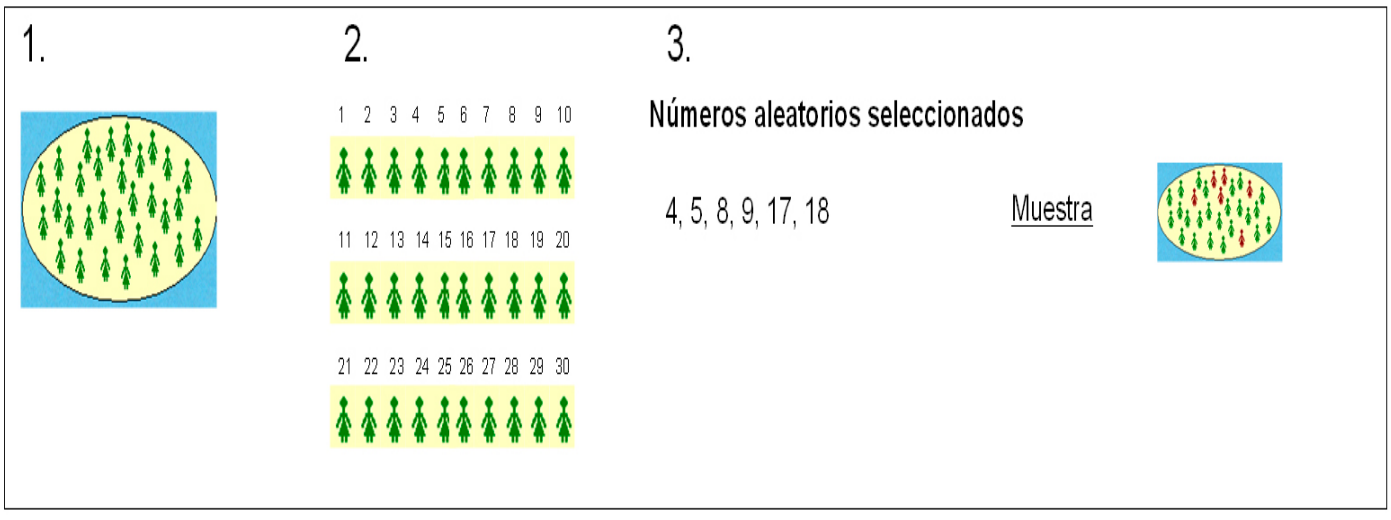
Se ha diseñado un estudio para describir el peso y determinar las variables asociadas al mismo en los niños entre 5 y 36 meses residentes en Bolivia.

Al total de niños entre 5 y 36 meses residentes en el país se le denomina **población**. Por razones prácticas y financieras, no podemos acceder a todos los niños de la población a estudio. En su defecto, podemos estudiar un subconjunto del total de niños del país. A este subconjunto de la población se le denomina **muestra**. En adelante, asumimos que los métodos para calcular el tamaño de la muestra han mostrado que debemos seleccionar un total de 160 niños de la población. El número total de individuos de la muestra se conoce como **tamaño muestral (n)**.

¿Cómo seleccionamos a los individuos de la muestra? Podríamos pensar en seleccionar 160 niños de la capital, La Paz. Sin embargo, las características de los niños que viven en La Paz pueden ser diferentes a las de los niños que viven en el resto del país, y por tanto, no ser representativos de la población a estudio.

Para evitar sesgos en la selección de la muestra, seleccionamos una **muestra aleatoria**: una muestra en la que cada miembro de la población tiene las mismas posibilidades de ser seleccionado, con independencia de los miembros seleccionados previamente, y la elección de los diferentes miembros de la muestra está basada en el azar. Los pasos a seguir para seleccionar una muestra aleatoria son:

1. Disponer de un listado de todos los miembros de la población mediante censos poblacionales o registros electorales.
2. A cada miembro de la población se le asigna un número de identificación.
3. Se seleccionan tantos números aleatorios como sujetos queramos incluir en la muestra.



*Al conjunto total de sujetos que estamos interesados en estudiar se le denomina **población**. Al subconjunto de sujetos de la población que observamos se le denomina **muestra**. El número de sujetos de la muestra se conoce como **tamaño muestral**.*

*Una **muestra aleatoria** es una muestra en la que cada miembro de la población tiene las mismas posibilidades de ser seleccionado, con independencia de los miembros seleccionados previamente, y la elección de los diferentes miembros de la muestra está basada en el azar.*

La selección de muestras aleatorias no siempre es posible. En ese caso, la interpretación de los resultados debe realizarse con cautela. Por ejemplo, si se ha seleccionado una muestra de pacientes VIH positivos atendidos en un Centro de Información y Prevención del SIDA (CIPS), los resultados obtenidos son extrapolables a la población de pacientes VIH positivos que acuden a los CIPS, y no al total de pacientes VIH positivos.

2. Tipos de variables

Una vez seleccionados los sujetos de la muestra, se recoge información sobre las características a estudio. Generalmente, los **sujetos** bajo observación son individuos, aunque no siempre (ejemplo: hogares familiares o hospitales). Las características medidas en los sujetos se denominan **variables.** Los valores que toman cada una de las variables en los diferentes sujetos se denominan **datos.**

En la siguiente tabla se presenta la información de los primeros 10 niños de la muestra:

	Variables					
	Id	**Sexo**	**Edad (meses)**	**Altura (cm)**	**Clase social**	**Peso (kg.)**
Sujetos	1	Masculino	8	68.3	Baja	7.0
	2	Masculino	9	68.8	Baja	6.9
	3	Masculino	6	68.6	Media	8.7
	4	Masculino	25	85.4	Alta	12.3
	5	Femenino			Media	9.3
	6	Masculino		Datos	Alta	15.7
	7	Femenino	17	70.7	Media	9.6
	8	Femenino	19	76.3	Media	9.5
	9	Masculino	10	70.6	Media	9.0
	10	Femenino	6	66.5	Media	8.4

El primer paso, antes de elegir el método más apropiado para analizar los datos, consiste en clasificar las variables en función de su escala de medida y de su papel en el estudio.

2.1. Clasificación de las variables en función de su escala de medida

Variables categóricas (o cualitativas): variables que no son susceptibles de ser medidas numéricamente. Se dividen en:

- Ordinales: Las categorías son susceptibles de ser ordenadas de un modo lógico, siguiendo un orden ascendente o descendente. Ejemplo: clase social (baja, media, alta).
- Nominales: Las categorías no siguen ningún orden natural. Ejemplo: estado civil (soltero, casado, viudo, divorciado) o grupo sanguíneo (A, B, AB, O).

Las variables categóricas que sólo toman dos valores se denominan dicotómicas (o binarias). Ejemplo: sexo (hombre, mujer) o estado vital (vivo, muerto).

Variables cuantitativas (o numéricas): Variables susceptibles de ser medidas numéricamente. Resultan de realizar mediciones o conteos. Se clasifican en:

- Discretas: Variables que pueden tomar un número limitado de valores, normalmente números enteros. Ejemplo: número de partos o número de hijos.
- Continuas: Variables medidas en escala continua, que pueden tomar cualquier valor dentro del eje real. Ejemplo: peso o altura.

*Una vez seleccionados los **sujetos** de la muestra, se recoge información sobre las características a estudio. Las características medidas en los sujetos se denominan **variables**, y los valores que toman las variables en los diferentes sujetos constituyen los **datos***

2.2. Clasificación de las variables en función de su papel en el estudio

La información sobre una variable se recoge por una de las siguientes razones:

Variable de interés (variable respuesta o dependiente): variable que es el centro de nuestra atención, cuya ocurrencia estamos interesados en comprender. En nuestro ejemplo, la variable de interés sería el peso.

Variable de exposición (variable explicativa o independiente): variable que puede influir en la ocurrencia de la variable de interés. En nuestro ejemplo, las variables de exposición serían el sexo, la edad, la altura y la clase social.

*Las variables, en función de su escala de medida, se clasifican en **Variables categóricas** (ordinales o nominales) y **Variables cuantitativas** (discretas o continuas)*

*Las variables, en función de su papel en el estudio, se clasifican en **Variable de interés** (variable respuesta o dependiente) y **Variable/s de exposición** (variable/s explicativa/s o independientes/s*

3.- Estadística descriptiva

Conjunto de procedimientos que permiten resumir, numérica y gráficamente, un conjunto de datos. Tiene un doble objetivo: controlar la calidad de los datos y describir las características de los individuos de la muestra. El análisis descriptivo a realizar depende de la escala de medida de las variables.

3.1. Análisis descriptivo de variables categóricas

Se realiza, numéricamente, mediante tablas de frecuencias, y gráficamente, mediante diagramas de barras o diagramas de sectores.

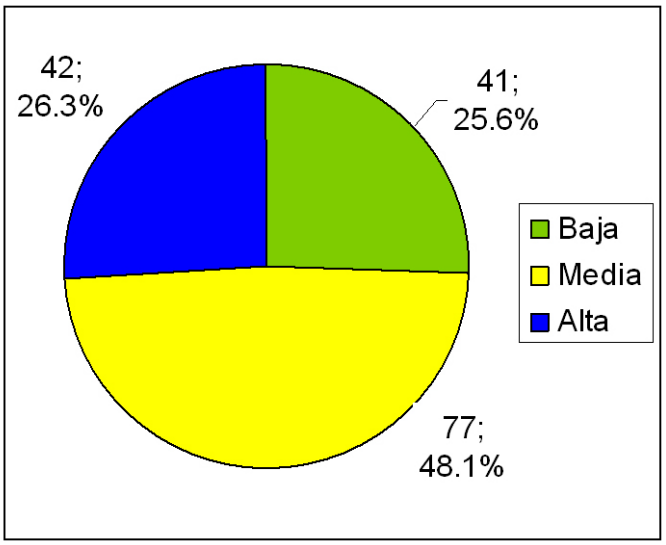
➢ Tabla de frecuencias

Tabla en la que se muestra el número (frecuencia absoluta) y el porcentaje (frecuencia relativa) de individuos que hay en cada una de las categorías de la variable. La Tabla de frecuencias de la Clase social de los 160 niños de la muestra sería:

Clase social	Número	Porcentaje
Baja	41	25.6 (= 41/160)
Media	77	48.1 (= 77/160)
Alta	42	26.3 (= 42/160)
Total	160	100.0

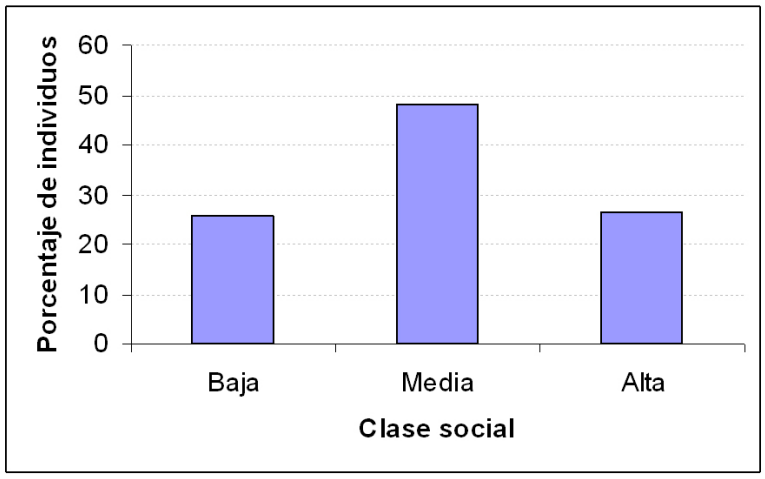
➢ Diagrama de sectores

Es un círculo en el que a cada categoría de la variable se le asigna un sector de área proporcional a su frecuencia absoluta o relativa. El diagrama de sectores de la Clase social de los 160 niños de la muestra sería:



➢ Diagrama de barras

Es un gráfico en el que las categorías de la variable se representan sobre el eje horizontal y sus frecuencias absolutas o relativas sobre el eje vertical. El diagrama de barras de la Clase social de los 160 niños de la muestra sería:



*El análisis descriptivo de **variables categóricas** se realiza, numéricamente, mediante **tablas de frecuencias** y, gráficamente, mediante **diagramas de barras** y **diagramas de sectores**.*

3.2. Análisis descriptivo de variables cuantitativas

La descripción numérica de una variable cuantitativa se realiza mediante una medida de tendencia central y una medida de dispersión. La elección de ambas medidas depende de la

*El análisis descriptivo de **variables cuantitativas** se realiza, numéricamente, mediante una **medida de tendencia central** y una **medida de dispersión** y, gráficamente, mediante **histogramas** y **diagramas de cajas**.*

distribución de la variable. La descripción gráfica de variables cuantitativas se realiza mediante Histogramas y Diagramas de cajas.

➢ Medida de tendencia central

Es un valor alrededor del cual se concentran los datos. Las medidas de tendencia central son:

Media

Es el promedio de las observaciones, es decir, la suma de las observaciones dividida por el número de observaciones:

$$Media\ (\bar{x}) = \frac{\sum x_i}{n}$$

Si los pesos de los primeros 10 niños de la muestra fueran 7.0, 6.9, 8.7, 12.3, 9.3, 15.7, 9.6, 9.5, 9.0, 8.4 kg, la media del peso sería:

$$Media\ (\bar{x}) = \frac{7.0+6.9+8.7+12.3+9.3+15.7+9.6+9.5+9.0+8.4}{10} = \frac{96.4}{10} = 9.64\ kg$$

Es la medida de tendencia central más utilizada. Su principal desventaja es que está fuertemente afectada por los valores extremos. Si los días de estancia en un hospital de 8 pacientes fueran 1, 3, 4, 2, 3, 4, 2 y 1 días, la medía sería 2.5 días. Sin embargo, el hecho de cambiar un 1 por un 25 en el último paciente, cambiaría drásticamente la media pasando a ser 5.5 días, un valor que no es representativo de los días de estancia de la mayoría de los pacientes. Por eso, en presencia de valores extremos, es preferible utilizar la Mediana como medida de tendencia central.

Mediana

Es el valor que divide la distribución de los datos en dos partes iguales, de forma que hay el mismo número de observaciones por debajo que por encima de la mediana.

El cálculo de la Mediana varía en función de si el número de observaciones es par o impar:

Número de observaciones impar

Si los pesos de los primeros 9 niños de la muestra fueran 7.0, 6.9, 8.7, 12.3, 9.3, 15.7, 9.6, 9.5, 9.0 kg, ordenamos los pesos de menor a mayor:

6.9, 7.0, 8.7, 9.0, 9.3, 9.5, 9.6, 12.3, 15.7

y seleccionamos el valor central como la Mediana.
Mediana = 9.3 kg.

Número de observaciones par

Si los pesos de los primeros 10 niños de la muestra fueran 7.0, 6.9, 8.7, 12.3, 9.3, 15.7, 9.6, 9.5, 9.0, 8.4, ordenamos los pesos de menor a mayor:

6.9, 7.0, 8.4, 8.7, 9.0, 9.3, 9.5, 9.6, 12.3, 15.7

y calculamos la mediana como la media de los dos valores centrales:

$$Mediana = \frac{9.0 + 9.3}{2} = 9.15\ kg$$

Moda

Es el valor que ocurre con más frecuencia. Si las edades de los primeros 10 niños de la muestra fueran 8, 9, 6, 25, 16, 29, 17, 19, 10 y 6 años, la Moda sería 6.

- **Medida de dispersión**

Es un valor que nos indica lo dispersos que se encuentran los datos alrededor de la medida de tendencia central. Las principales medidas de dispersión son:

Rango

Es la medida de dispersión más simple e intuitiva. Se calcula como la diferencia entre el mayor y el menor valor. Si los pesos de los primeros 10 niños de la muestra fueran 7.0, 6.9, 8.7, 12.3, 9.3, 15.7, 9.6, 9.5, 9.0 y 8.4 kg., el rango sería:

Rango = 15.7 – 6.9 = 8.8 kg

Se expresa en las mismas unidades que los datos originales. Sin embargo, está basado únicamente en los dos valores extremos y, por tanto, su valor aumenta conforme aumenta el tamaño muestral ya que aumentan las posibilidades de que aparezcan valores extremos.

Varianza

Para ilustrar el proceso intuitivo que da lugar a la definición de Varianza, utilizamos los pesos de los primeros 5 niños de la muestra: 7.0, 6.9, 8.7, 12.3 y 9.3 kg.

El rango es 12.3 – 6.9 = 5.4 kg. La principal limitación del Rango es que se basa únicamente en dos valores. Nos planteamos encontrar una medida de dispersión en la que participen todos los datos de la muestra. Parece conveniente calcular la diferencia entre cada peso y el peso medio (8.83 kg.):

Individuo	Peso (kg)	Peso – Peso medio (kg)
1	7.0	7.0 – 8.84 = - 1.84
2	6.9	6.9 – 8.84 = - 1.94
3	8.7	8.7 – 8.84 = - 0.14
4	12.3	12.3 – 8.84 = +3.46
5	9.3	9.3 – 8.84 = +0.46

A continuación, calculamos la media de las diferencias entre cada peso y el peso medio:

$$Media\ (diferencias) = \frac{-1.84 - 1.94 - 0.14 + 3.46 + 0.46}{5} = \frac{0}{5} = 0$$

La media de las diferencias es 0; sin embargo, los diferentes pesos si que presentan dispersión respecto al peso medio (8.84 kg).

El error del razonamiento es permitir que las "diferencias" puedan ser negativas. Para evitar valores negativos de las diferencias, elevamos al cuadrado las diferencias entre cada peso y el peso medio:

Individuo	Peso (kg)	Peso – Peso medio (kg)	(Peso – Peso medio)^2 (kg^2)
1	7.0	7.0 – 8.84 = - 1.84	$(-1.84)^2 = 3.39$
2	6.9	6.9 – 8.84 = - 1.94	$(-1.94)^2 = 3.76$
3	8.7	8.7 – 8.84 = - 0.14	$(-0.14)^2 = 0.02$
4	12.3	12.3 – 8.84 = +3.46	$(+3.46)^2 = 11.97$
5	9.3	9.3 – 8.84 = +0.46	$(+0.46)^2 = 0.21$

Finalmente, calculamos la media de los cuadrados de las diferencias, dando lugar a lo que se conoce como Varianza:

$$Varianza\,(s^2) = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{3.39 + 3.76 + 0.02 + 11.97 + 0.21}{5} = 3.87\ kg^2$$

El principal inconveniente de la Varianza es que se expresa en unidades que son el cuadrado de las unidades de las observaciones originales. Para evitar esto, se hace la raíz cuadrada obteniendo lo que se conoce como Desviación estándar.

Desviación estándar (o Desviación típica)

Es la medida de dispersión más utilizada. Es la raíz cuadrada de la varianza. Expresa la dispersión de los datos respecto a la media y se expresa en las mismas unidades de medida que la variable original.

En nuestro ejemplo, la desviación estándar sería:

$$Desviación\ estándar\ (s) = \sqrt{Varianza} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{3.87} = 1.97\ kg$$

En el proceso de Inferencia Estadística, es habitual utilizar el tamaño muestral menos uno (n – 1), en lugar del tamaño muestral (n), como denominador para el cálculo de la Varianza y Desviación Estándar. El motivo principal es que la nueva expresión tiene mejores propiedades para hacer Inferencia Estadística, uno de los principales objetivos de la estadística.

Percentiles

Si ordenamos los datos de menor a mayor, se define el percentil i como el valor que deja a su izquierda el i% del total de observaciones. Los percentiles 25, 50 y 75 se denominan primer, segundo (o Mediana) y tercer cuartil. A continuación, se presenta el procedimiento para llevar a cabo el cálculo de los percentiles 25 y 75 de los pesos de los primeros 12 niños de la muestra: 7.0, 6.9, 8.7, 12.3, 9.3, 15.7, 9.6, 9.5, 9.0, 8.4, 8.9, 11.3

Percentil 25. Si ordenamos los datos de menor a mayor, el percentil 25 será aquél valor que deja a su izquierda el 25% del total de observaciones y a su derecha el 75% de las mismas.

Ordenamos los datos de menor a mayor:
6.9, 7.0, 8.4, 8.7, 8.9, 9.0, 9.3, 9.5, 9.6, 11.3, 12.3, 15.7
Identificamos el valor que deja a su izquierda el 25% de las observaciones, esto es, 3 observaciones: 8.7. Y el valor que deja a su derecha el 75% de las observaciones, esto es, 9 observaciones: 8.4. El Percentil 25 es el promedio de ambas observaciones:

$$Percentil\,25\ (P_{25}) = \frac{8.4+8.7}{2} = 8.55$$

Percentil 75 (o Tercer cuartil). Si ordenamos los datos de menor a mayor, el percentil 75 será aquél valor que deja a su izquierda el 75% de las observaciones y a su derecha el 25% de las mismas.

Ordenamos los datos de menor a mayor:
6.9, 7.0, 8.4, 8.7, 8.9, 9.0, 9.3, 9.5, 9.6, 11.3, 12.3, 15.7

Identificamos el valor que deja a su izquierda el 75% de las observaciones, esto es, 9 observaciones: 11.3. Y el valor que deja a su derecha el 25% de las observaciones, esto es, 3 observaciones: 9.6. El Percentil 75 se calcula como el promedio de ambas observaciones:

$$Percentil\,75\ (P_{75}) = \frac{9.6+11.3}{2} = 10.45$$

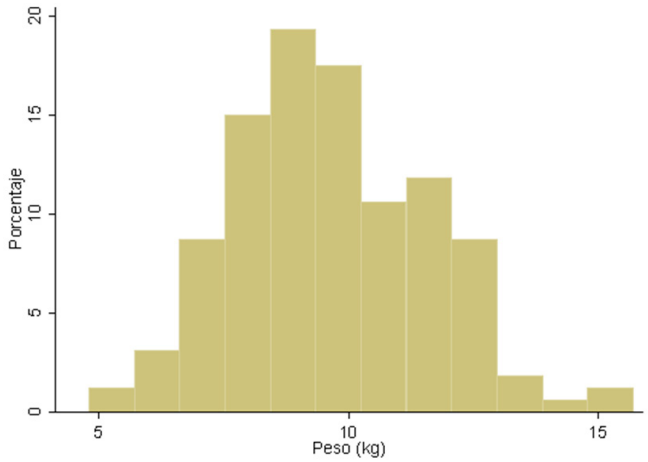
Rango Intercuartílico

Se define como la diferencia entre los percentiles 25 y 75. En nuestro ejemplo, el Rango Intercuartílico sería RI = (8.55; 10.45)

Histograma

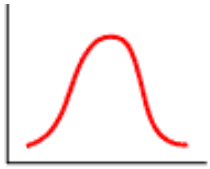
Es la representación gráfica más utilizada en investigación.

Los valores de la variable cuantitativa se dividen en intervalos, que se representan sobre el eje horizontal. En el eje vertical, se representan las frecuencias absolutas o relativas de cada intervalo en forma de rectángulo. Su forma es similar a la del diagrama de barras con la diferencia de que no hay espacio entre las barras. El Histograma del Peso de los 160 niños de la muestra sería:



El Histograma proporciona información sobre la distribución de la variable. Las tres formas más comunes que puede presentar la distribución de una variable cuantitativa son:

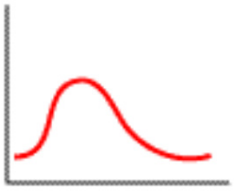
Simétrica



Tiene forma de campana; la cola derecha de la distribución es igual que la cola izquierda.

Ejemplo. Altura o Peso.

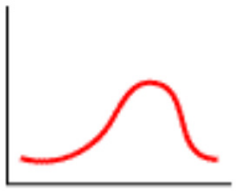
Asimétrica a la derecha



La cola derecha de la distribución es más larga que la cola izquierda.

Ejemplo. Puntuación GHQ

Asimétrica a la izquierda



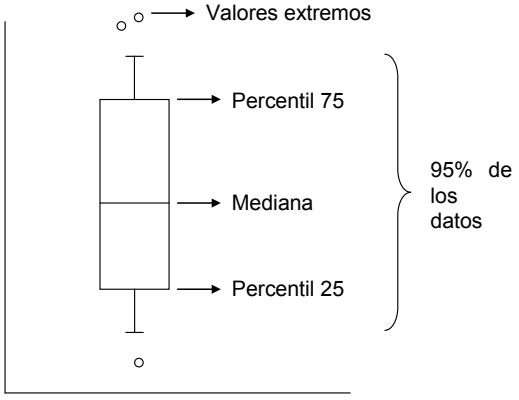
La cola izquierda de la distribución es más larga que la cola derecha.

Ejemplo. Período de gestación

Si la distribución de los datos es simétrica, la media, mediana y moda son iguales. Si la distribución es asimétrica a la derecha, la media es mayor que la mediana. Si la distribución es asimétrica a la izquierda, la media es menor que la mediana.

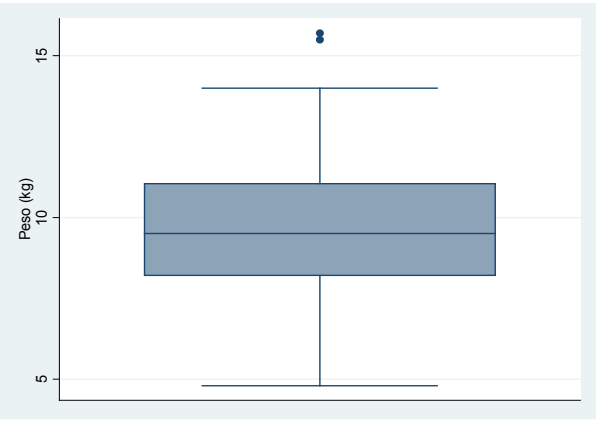
Diagrama de cajas

Es un gráfico en el que se representa la mediana, los percentiles 25 y 75, los valores atípicos y los valores extremos:



Si la distribución de la variable es simétrica, la distancia entre el percentil 25 y la mediana será similar a la distancia entre el percentil 75 y la mediana, y la distancia entre el bigote superior y la mediana será similar a la distancia entre el bigote inferior y la mediana.

El diagrama de cajas del peso de los 160 niños de la muestra es:



El Histograma y el Diagrama de cajas del Peso de los 160

*Si la distribución de una **variable cuantitativa** es **aproximadamente simétrica**, se utiliza la **media** como medida de tendencia central y la **desviación estándar** como medida de dispersión para describirla. Si, por el contrario, la distribución de la variable es **marcadamente asimétrica**, se utiliza la **mediana y el rango intercuartílico** como medidas de tendencia central y de dispersión, respectivamente*

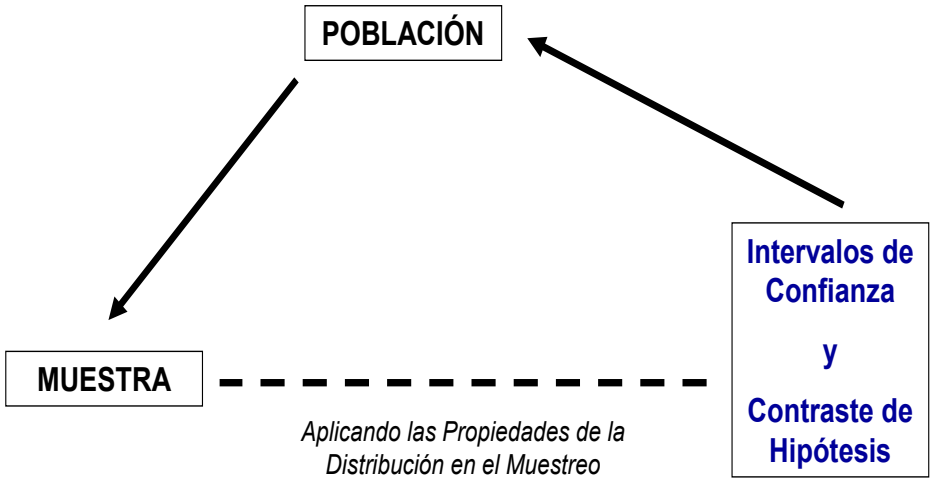
niños de la muestra reflejan que la distribución del Peso es aproximadamente simétrica.

La descripción numérica de una variable cuantitativa se realiza indicando una medida de tendencia central y una medida de dispersión. Si la distribución de la variable es aproximadamente simétrica, se utiliza la media, como medida de tendencia central, y la desviación estándar, como medida de dispersión. Si la distribución de la variable es marcadamente asimétrica, se utiliza la mediana y el rango intercuartílico para describirla.

4.- Inferencia Estadística

*El principal objetivo del Análisis Estadístico es utilizar la información de la muestra para sacar conclusiones acerca de la población a estudio. Las dos herramientas que permiten obtener conclusiones sobre la población a partir de la información proporcionada por los datos de la muestra son: los **Intervalos de Confianza** y los **Contrastes de Hipótesis**.*

El principal objetivo del Análisis Estadístico es utilizar la información de la muestra para sacar conclusiones acerca de la población a estudio. Hay dos herramientas que nos permiten obtener conclusiones sobre la población a estudio a partir de la información proporcionada por los datos de la muestra: los Intervalos de Confianza y los Contrastes de Hipótesis.



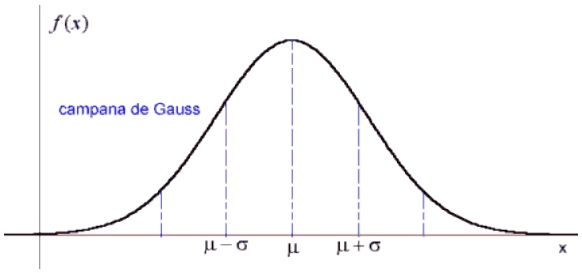
Un Intervalo de Confianza es un rango de valores dentro de los cuales podemos estar seguros que se encuentra un valor poblacional, denominado parámetro, que queremos estudiar. Un Contraste de Hipótesis es un procedimiento que nos permite decidir sobre la veracidad de una hipótesis planteada sobre un valor poblacional.

El cálculo de Intervalos de Confianza y Contrastes de Hipótesis

requiere del uso de la Probabilidad, una herramienta que permite medir el grado de incertidumbre con el que ocurren los fenómenos aleatorios. La mayoría de los fenómenos de la Naturaleza siguen exacta o aproximadamente una serie de distribuciones de probabilidad teóricas bien definidas; la identificación de la distribución que mejor se ajusta al comportamiento de un fenómeno nos permitirá calcular cualquier probabilidad asociada a la ocurrencia del mismo.

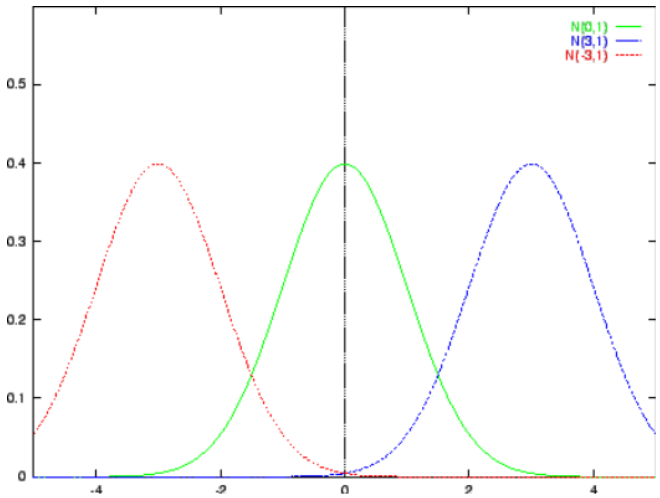
4.1. Distribución Normal

Es la distribución más importante en Estadística ya que numerosas variables asociadas a fenómenos naturales siguen, aproximadamente, una distribución Normal (ejemplo: estatura o peso). Está determinada por dos parámetros: la Media (μ) y la Desviación Estándar (σ), tal y como se muestra a continuación:



Tiene forma de campana y es simétrica respecto a su media.

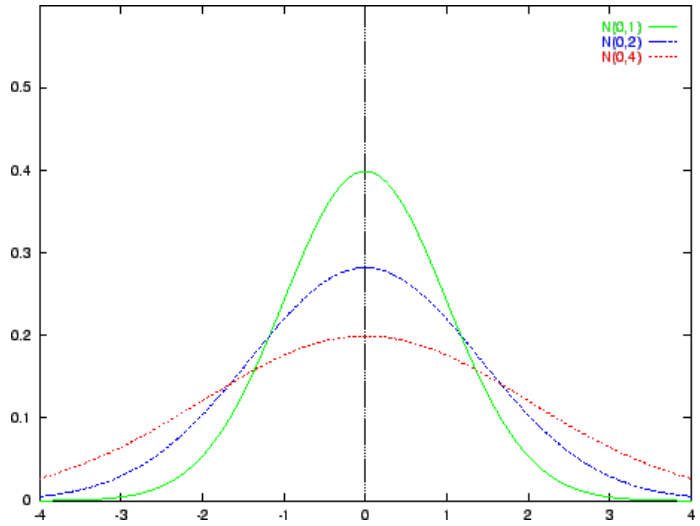
Si la media aumenta, la distribución se desplaza a la derecha; si la media disminuye, la distribución se desplaza hacia la izquierda.



*Un **Intervalo de Confianza** es un rango de valores dentro de los cuales podemos estar seguros que se encuentra un valor poblacional, denominado parámetro, que queremos estudiar. Un **Contraste de Hipótesis** es un procedimiento que nos permite decidir sobre la veracidad de una hipótesis planteada sobre un valor poblacional.*

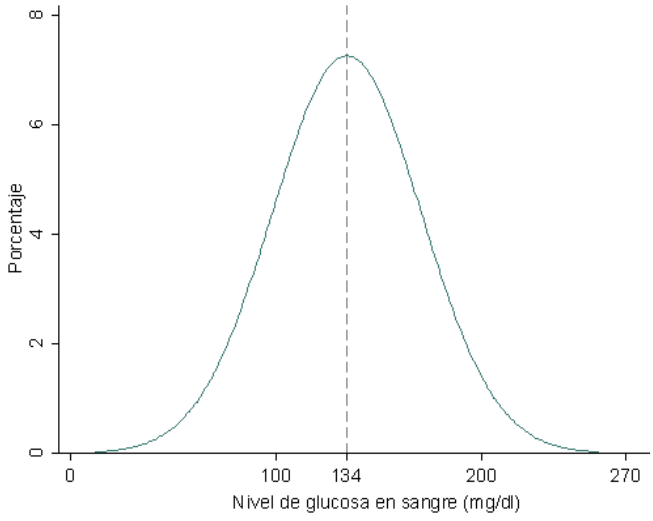
*La **distribución Normal** es la distribución más importante en Estadística. Está determinada por dos parámetros: la Media (μ) y la Desviación Estándar (σ). Tiene forma de campana y es simétrica respecto a su media.*

Si la desviación estándar aumenta, la altura disminuye y la curva se ensancha; si la desviación estándar disminuye, la altura aumenta y la curva se estrecha.



Una vez caracterizada la distribución de una variable, podemos calcular probabilidades asociadas a la ocurrencia de la variable.

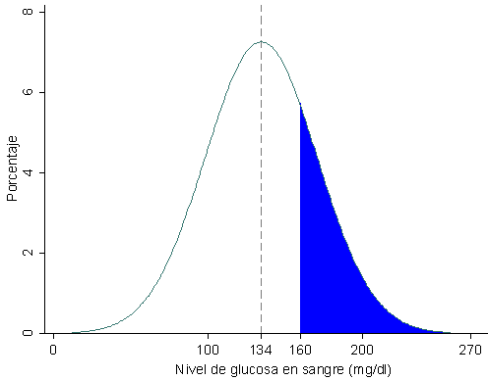
La siguiente figura muestra la distribución del nivel de glucosa en sangre en una población de pacientes diabéticos tipo II. La distribución del nivel de glucosa en sangre es Normal con media (μ) 134 mg/dl, y desviación estándar (σ) 36 mg/dl.



Una vez caracterizada la distribución de una variable, podemos calcular probabilidades asociadas a la ocurrencia de la variable.

Supongamos que estamos interesados en determinar el porcentaje de individuos de esta población que tienen un nivel de glucosa en sangre superior a 160 mg/dl (o, equivalentemente, la probabilidad de que un individuo de esta población tenga un nivel de glucosa en sangre superior a 160 mg/dl). Para calcular

este porcentaje basta con determinar el área bajo la curva de una Normal (μ = 134, σ = 36) que está por encima del valor 160, tal y como se muestra a continuación:



La distribución Normal estándar es una distribución con media 0 y desviación estándar 1.

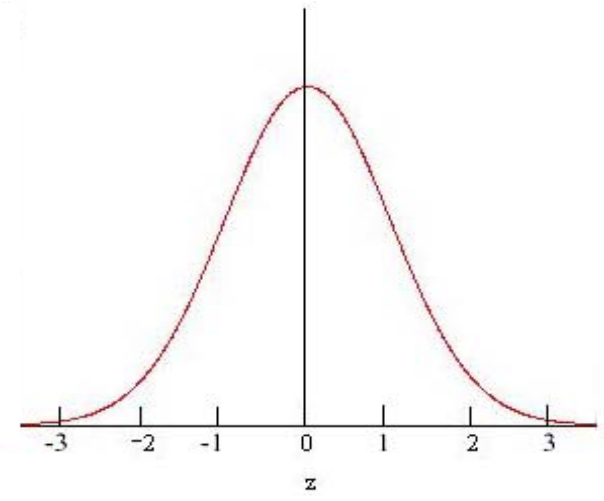
El cálculo de esta área se haría integrando entre 160 e infinito la función que define la curva de una Normal (μ = 134, σ = 36):

$$\Pr(X \geq 160) = \int_{160}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx = \int_{160}^{\infty} \frac{1}{\sqrt{2\pi 36^2}} \exp\left(\frac{-(x-134)^2}{2\cdot 36^2}\right) dx$$

Afortunadamente, podemos utilizar un procedimiento alternativo, basado en la distribución Normal estándar para llevar a cabo este cálculo.

La Distribución Normal Estándar

Es una distribución Normal con media 0 y desviación estándar 1.



Los valores de la distribución Normal estándar se denominan puntuaciones z. Es posible transformar cualquier variable con distribución Normal de media μ y desviación estándar σ en una distribución Normal estándar, mediante la siguiente fórmula:

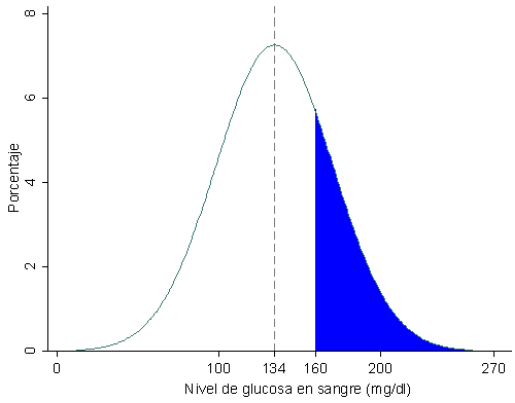
$$z = \frac{x - \mu}{\sigma}$$

Continuamos con el ejemplo anterior en el que estábamos interesados en determinar el porcentaje de individuos de la población con nivel de glucosa en sangre superior a 160 mg/dl. Para calcular este porcentaje basta con determinar el área bajo la curva de una Normal (μ = 134, σ = 36) que está por encima del valor 160. Para determinar esta área, calculamos la puntuación z correspondiente al valor 160 de una distribución Normal (μ = 134, σ = 36). Para ello, al valor 160 le restamos la media y dividimos por la desviación estándar:

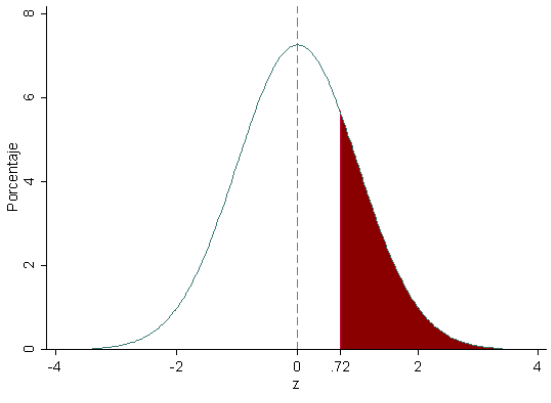
$$z = \frac{160 - 134}{36} = 0.72$$

El área bajo la curva de una Normal (μ = 134, σ = 36) que está por encima del valor 160, (a), es exactamente igual que el área bajo la curva de una Normal estándar que está por encima del valor 0.72, (b).

(a)

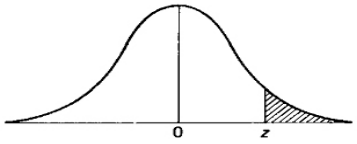


(b)



El área bajo la curva de una Normal estándar que está por encima del valor 0.72 puede calcularse utilizando la Tabla de la distribución Normal Estándar:

Distribución Normal Estándar
Z~ Normal (0, 1)
Pr (Z ≥ z) = p



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0-0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
3.3	0.00048	0.00047	0.00045	0.00043	0.000 42	0.00040	0.00039	0.00038	0.00036	0.00035
3.4	0.00034	0.00032	0.00031	0.0030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003

Las probabilidades aparecen en el interior de la tabla. El área bajo la curva de una Normal estándar que está por encima del valor 0.72 es la probabilidad correspondiente al valor 0.7 de las filas y al valor 0.02 de las columnas:

z	0.00	0.01	**0.02**	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	**0.2358**	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

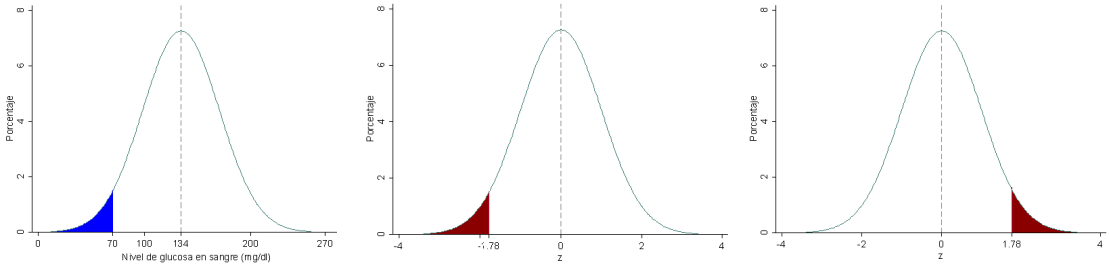
La probabilidad buscada es 0.2358. En esta población, el 23.58% de los individuos tienen un nivel de glucosa en sangre superior a 160 mg/dl, o equivalentemente, la probabilidad de que un individuo elegido al azar de esta población tenga un nivel de glucosa en sangre superior a 160 mg/dl es 0.2358.

Supongamos que estamos interesados en determinar qué porcentaje de individuos de esta población tienen un nivel de glucosa en sangre inferior a 70 mg/dl. Para ello, determinamos el área bajo la curva de una Normal (μ = 134, σ = 36) que está por debajo del valor 70. Para determinar esta área calculamos la puntuación z correspondiente al valor 70 de una distribución Normal (μ = 134, σ = 36):

$$z = \frac{70 - 134}{36} = -1.78$$

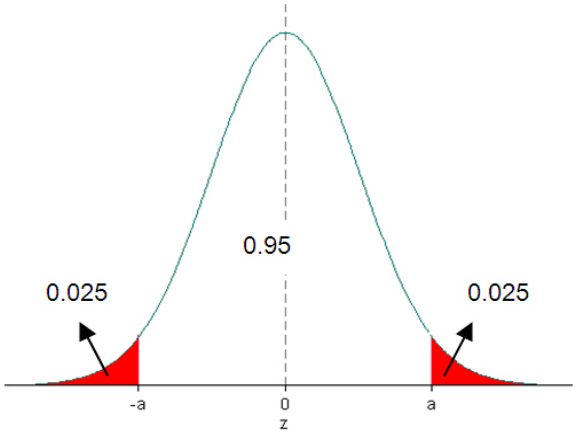
El área bajo la curva de una Normal (μ = 134, σ = 36) que está por debajo del valor 70, (a), es exactamente igual que el área bajo la curva de una Normal estándar que está por debajo del valor -1.78, (b).

(a) (b) (c)



Como la distribución Normal Estándar es simétrica respecto al 0, el área bajo la curva que está por debajo del valor -1.78 es igual que el área bajo la curva que está por encima del valor +1.78, (c). Buscando en la Tabla de la distribución Normal Estándar, obtenemos que el área bajo la curva Normal estándar por encima del valor +1.78 es 0.0375; es decir, el 3.75% de los individuos de esta población tienen un nivel de glucosa en sangre por debajo de 70 mg/dl.

A partir de la Tabla de la distribución Normal Estándar podemos determinar entre qué dos puntuaciones z hay una determinada probabilidad. Por ejemplo, para determinar entre qué dos valores de la Distribución Normal estándar hay una probabilidad del 95%, basta con determinar el valor "a" que deja a su derecha una probabilidad del 2.5%:



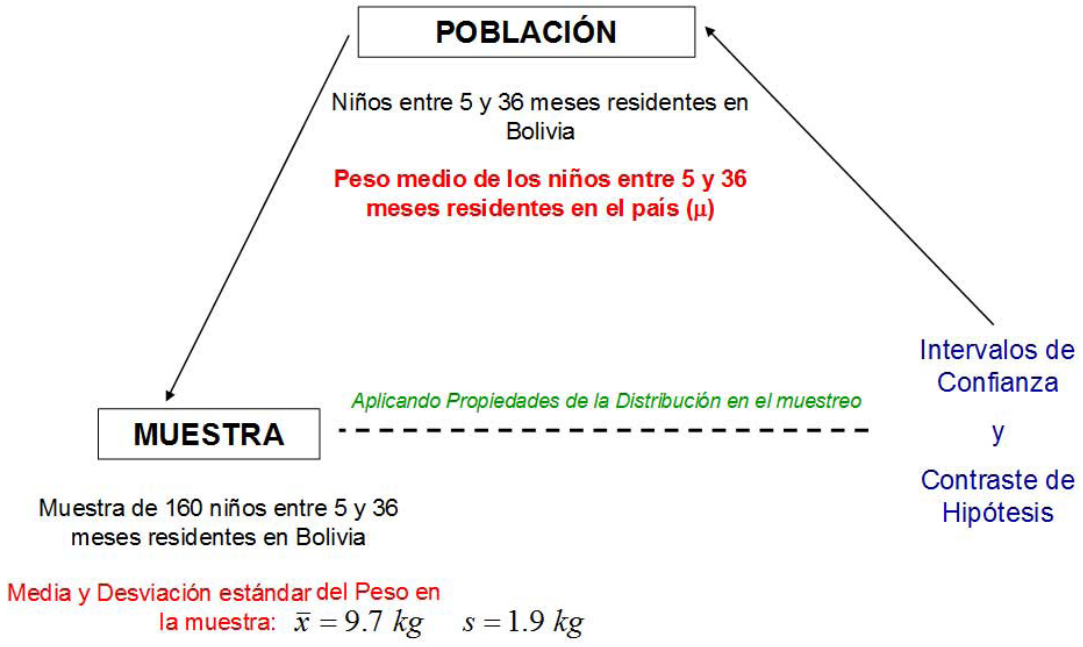
La puntuación z que deja a su derecha una probabilidad de

0.025 es +1.96. Dado que la distribución Normal Estándar es simétrica respecto al 0, la puntuación -1.96 deja a su izquierda una probabilidad de 0.025. Por lo tanto, entre -1.96 y +1.96 hay una probabilidad del 95%.

Siguiendo el mismo procedimiento, podemos determinar que entre -1.64 y +1.64 hay una probabilidad del 90%, y entre -2.58 y +2.58 hay una probabilidad del 99%.

4.2. Inferencia sobre una media

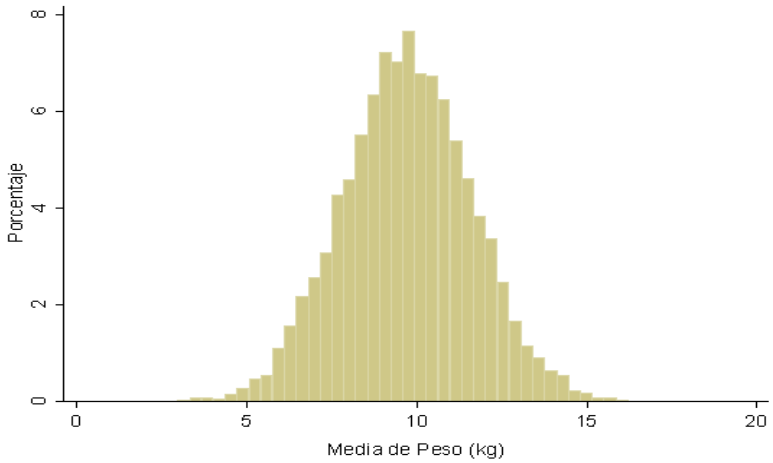
Supongamos que estamos interesados en determinar el peso medio de los niños entre 5 y 36 meses residentes en Bolivia. La población a estudio es el total de niños entre 5 y 36 meses del país. El parámetro poblacional que estamos interesados en conocer es el Peso medio de los niños. En lo que sigue, a la media de una variable cuantitativa en la población la denotaremos como μ. Dado que por razones prácticas y financieras, no podemos acceder a todos los niños de la población, hemos seleccionado una muestra de 160 niños, a los que les hemos medimos el peso. En los 160 niños de la muestra, el peso medio ($\bar{x}$) es 9.7 kg y la desviación estándar (σ) es 1.9 kg.



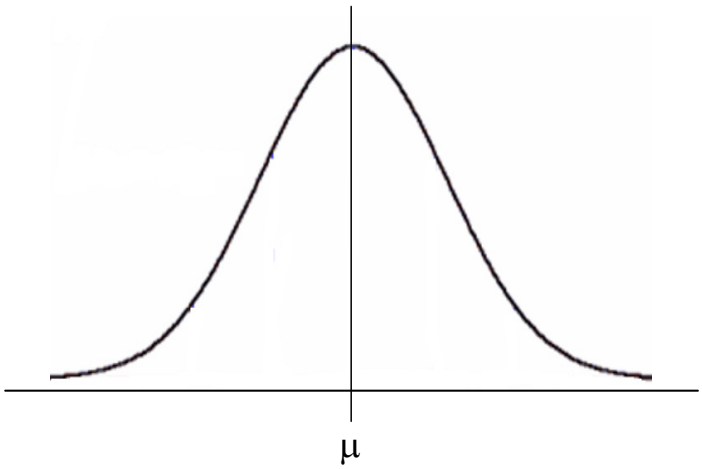
Si hubiéramos seleccionado una muestra diferente, habríamos obtenido una media y una desviación estándar diferente, debido a lo que se conoce como **Variación en el muestreo,** es decir, debido a la variabilidad que se observa al estudiar muestras en lugar de poblaciones.

*La **variación en el muestreo** es la variabilidad que se observa al estudiar muestras en lugar de poblaciones.*

Si en lugar de una única muestra, seleccionáramos 1000 muestras de tamaño 160 de la población, calculáramos la media de peso en cada una de las 1000 muestras y representáramos las 1000 medias de peso en un histograma, tendríamos lo que se conoce como la **Distribución en el muestreo de la Media Muestral.**



Si seleccionáramos 10.000 muestras de tamaño 160 de la población, la distribución en el muestreo de las 10.000 medias tendría la siguiente forma:



Si el tamaño muestral (n) es lo suficientemente grande:

- La distribución en el muestreo de las medias es aproximadamente Normal

- La Media de la distribución en el muestreo de las medias es la Media poblacional (μ)
- La desviación estándar de la distribución en el muestreo de las medias, conocida como Error estándar, es la Desviación estándar poblacional (σ) dividida por la raíz cuadrada del tamaño muestral:

$$EE(\bar{x}) = \sigma / \sqrt{n}$$

Intervalo de Confianza al 95% para la Media

Un Intervalo de Confianza para la Media de una variable cuantitativa en la población de la que se extrajo la muestra es un rango de valores, obtenidos a partir de los datos de la muestra, dentro de los cuales podemos estar seguros que se encuentra la Media de la variable cuantitativa en la población.

La distribución en el muestreo de las medias, $\overline{X}$, es aproximadamente $Normal(\mu, \sigma/\sqrt{n})$.

Por lo tanto, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ *sigue una distribución* $Normal(0,1)$

Sabemos que el 95% de las puntuaciones z están entre -1.96 y +1.96.

$$z = -1.96 \qquad -1.96 = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \overline{X} = \mu - (1.96 \times \sigma/\sqrt{n})$$

$$z = +1.96 \qquad +1.96 = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \overline{X} = \mu + (1.96 \times \sigma/\sqrt{n})$$

Por lo tanto, el 95% de las medias muestrales está en el rango

$$[\mu - (1.96 \times \sigma/\sqrt{n}),\ \mu + (1.96 \times \sigma/\sqrt{n})]$$

Y esto es equivalente a decir, que:

μ está en el rango: $[\bar{x} - (1.96 \times \sigma/\sqrt{n}),\ \bar{x} + (1.96 \times \sigma/\sqrt{n})]$ con una confianza del 95%.

El intervalo $[\bar{x} - (1.96 \times \sigma/\sqrt{n}), \bar{x} + (1.96 \times \sigma/\sqrt{n})$ se conoce como Intervalo

de Confianza al 95% de la Media poblacional.

En la mayoría de las situaciones no conocemos la desviación estándar poblacional (σ). En su lugar, utilizaremos la desviación estándar muestral (s), y calcularemos un Intervalo de Confianza al 95% para una Media con la siguiente fórmula:

$$IC_{95\%}(\mu) = \bar{x} \pm (1.96 \times s / \sqrt{n})$$

En nuestro ejemplo, un Intervalo de Confianza al 95% para el Peso medio de los niños entre 5 y 36 meses residentes en Bolivia se calcularía como:

$$IC_{95\%}(\mu) = \bar{x} \pm (1.96 \times s / \sqrt{n}) = 9.7 \pm (1.96 \times 1.9 / \sqrt{160}) = 9.7 \pm (1.96 \times 0.15) = (9.41;\ 9.99)$$

Estamos seguros al 95% de que el peso medio en la población de niños entre 5 y 36 meses residentes en Bolivia está entre 9.41 y 9.99 kg.

La fórmula general para calcular un Intervalo de Confianza al 95% de un parámetro poblacional es:

$$IC_{95\%}(parámetro) = estimador \pm 1.96 \times EE(estimador)$$

Podemos calcular un Intervalo de Confianza de una Media a un nivel de confianza distinto del 95%, por ejemplo al 90 o 99%. Para ello, basta con cambiar el valor 1.96 por el punto de la distribución Normal estándar que deja en las colas una probabilidad del 10% o 1%, respectivamente:

$$IC_{90\%}(\mu) = \bar{x} \pm (1.64 \times s / \sqrt{n})$$

$$IC_{99\%}(\mu) = \bar{x} \pm (2.58 \times s / \sqrt{n})$$

Un Intervalo de Confianza será tanto más preciso cuanto más estrecho sea, es decir, cuanto menor sea la distancia entre el límite superior y el límite inferior.

Existen dos alternativas para disminuir la amplitud de un Intervalo de Confianza: aumentar el tamaño de la muestra o disminuir el nivel de confianza. La primera, opción aconsejable, responde a una regla general de la Estadística: "Cuanto más grande es una muestra, más información proporciona y más precisas son las conclusiones obtenidas a partir de ella".

Un Intervalo de Confianza será más preciso cuanto más estrecho sea. Existen dos alternativas para disminuir la amplitud de un Intervalo de Confianza: aumentar el tamaño de la muestra, opción aconsejable, o disminuir el nivel de confianza.

Muestras pequeñas

Si el tamaño muestral es pequeño (n<60), la desviación estándar muestral, s, puede no ser un buen estimador de la desviación estándar poblacional, σ. Por este motivo, utilizaremos la distribución t de Student, en lugar de la distribución Normal, para calcular un Intervalo de Confianza para la Media.

Distribución t de Student

Es una distribución continua determinada por un parámetro conocido como grados de libertad: t_n es una distribución t de Student con n grados de libertad.

Su rango es todo el eje real: (-∞, +∞).

Tiene propiedades similares a la distribución Normal Estándar:

- Tiene Media 0 y es simétrica respecto a la Media
- Es más dispersa que la distribución Normal estándar, pero la desviación estándar decrece hasta 1 conforme aumentan los grados de libertad
- Conforme aumentan los grados de libertad, la distribución t de Student se aproxima a la distribución Normal Estándar

Figura. Función de densidad de una distribución t de Student con diferentes grados de libertad

Un Intervalo de Confianza al 95% para una Media viene dado por:

$$IC_{95\%}(\mu) = \bar{x} \pm (t_{n-1} \times s / \sqrt{n})$$

donde t_{n-1} es el punto de la distribución t de Student con n-1 grados de libertad que deja en las colas una probabilidad del 5%.

Para el cálculo de este punto puede utilizarse la Tabla de la distribución t de Student o las calculadoras de la mayoría de los programas estadísticos.

Supongamos que estamos interesados en determinar el Número medio de horas de sueño sin dolor en pacientes artríticos tras recibir un nuevo tratamiento. Se han seleccionado 6 pacientes y se ha observado el número de horas sin dolor tras recibir el tratamiento: 2.2, 2.4, 4.9, 2.5, 3.7 y 4.3 horas.

En los 6 pacientes de la muestra, la media y la desviación estándar del número de horas sin dolor son $\bar{x} = 3.3$ y $s = 1.13\ horas$, respectivamente.

Un Intervalo de Confianza al 95% para la Media del Número de horas sin dolor tras recibir el nuevo tratamiento en la población de pacientes artríticos de la que se extrajo la muestra, se calcularía como:

$$IC_{95\%}(\mu) = \bar{x} \pm (t_{n-1} \times s/\sqrt{n}) = 3.3 \pm (2.57 \times 1.13/\sqrt{6} = 3.3 \pm (2.57 \times 0.46) = (2.1;\ 4.5)\ horas$$

Estamos seguros al 95% de que en la población de pacientes artríticos de la que se extrajo la muestra, el número medio de horas sin dolor tras recibir el nuevo tratamiento está entre 2.1 y 4.5 horas.

La mayoría de los programas estadísticos utilizan, de forma general, la distribución t de Student para el cálculo de Intervalos de Confianza para una Media. La razón es que la distribución t de Student es la apropiada si el tamaño muestral es pequeño, y se aproxima a la Normal Estándar si el tamaño muestral es grande.

4.3. Comparación de dos medias

Son numerosas las ocasiones en las que el interés se centra en determinar si la media de una variable cuantitativa es igual en dos grupos diferentes de individuos.

Supongamos que estamos interesados en estudiar si hay diferencias por sexo en el peso medio de los niños entre 5 y 36 meses residentes en Bolivia.

Primero, calculamos la media y desviación estándar del peso en los niños y niñas de la muestra:

Niños	$n_1 = 68$	$\bar{x}_1 = 10.1$	$s_1 = 2.1$
Niñas	$n_2 = 92$	$\bar{x}_2 = 9.3$	$s_2 = 1.7$

El peso medio de los 68 niños de la muestra es 10.1 kg y el peso medio de las 92 niñas es 9.3 kg. A la vista del análisis descriptivo, parece que los niños pesan más que las niñas. Pero, ¿la diferencia observada puede ser explicada por azar? Para poder responder a esta pregunta, necesitamos realizar un Contraste de Hipótesis

que nos permita determinar si, en la población de la que se extrajeron las muestras, el peso medio de los niños es igual al peso medio de las niñas. Antes de esto, necesitamos conocer la distribución en el muestreo de la diferencia de medias.

*Un **Contraste de Hipótesis** es un procedimiento cuyo objetivo es comprobar si una determinada hipótesis enunciada acerca de la población es compatible o no con los datos de la muestra.*

Distribución en el muestreo de la Diferencia de Medias

Conforme n_1 y n_2 aumentan:

1. La distribución en el muestreo de la diferencia de medias se aproxima a la Normal
2. La media de la distribución en el muestreo es la diferencia de medias poblacionales, $(\mu_1 - \mu_2)$
3. La desviación estándar de la distribución en el muestreo, esto es, el error estándar de $(\bar{x}_1 - \bar{x}_2)$ es una combinación de los errores estándar de las medias individuales:

$$EE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Como en la mayoría de las situaciones no conocemos las desviaciones estándar poblacionales (σ_1, σ_2) , utilizaremos las desviaciones estándar muestrales (s_1, s_2) para estimar el error estándar de la diferencia de medias muestrales. Por lo tanto,

$$EE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

*El **primer paso** de un **Contraste de Hipótesis** consiste en definir la **Hipótesis Nula** (generalmente, ausencia de asociación entre dos variables) y la **Hipótesis Alternativa** (existencia de asociación).*

Contraste de Hipótesis para la Diferencia de Medias

Un Contraste de Hipótesis es un procedimiento cuyo objetivo es comprobar si una determinada hipótesis enunciada acerca de la población es compatible o no con los datos de la muestra. Permite decidir entre dos hipótesis, la Hipótesis Nula (generalmente, ausencia de asociación entre dos variables) y la Hipótesis Alternativa (existencia de asociación).

El primer paso de un Contraste de Hipótesis consiste en definir las Hipótesis Nula y Alternativa. En nuestro ejemplo, la definición de ambas hipótesis sería:

Hipótesis Nula: En la población de niños entre 5 y 36 meses residentes en Bolivia, no hay una asociación estadísticamente significativa entre el Sexo y el Peso; es decir, el peso medio de los niños es igual que el peso medio de las niñas

$$H_0: \mu_1 = \mu_2$$

Hipótesis Alternativa: En la población de niños entre 5 y 36 meses residentes en Bolivia, existe una asociación estadísticamente significativa entre el Sexo y el Peso; es decir, el peso medio de los niños es diferente al peso medio de las niñas.

$$H_1: \mu_1 \neq \mu_2$$

donde μ_1 y μ_2 son el peso medio de los niños y las niñas en la población, respectivamente

De forma descriptiva, la evaluación de la asociación entre el sexo y el peso se obtiene de la comparación del peso medio de los niños, 10.1 kg, y el peso medio de las niñas, 9.3 kg. Si no existiera asociación, ambas medias serían iguales aunque no necesariamente iguales, debido a la variación en el muestreo, al hecho de estudiar muestras en lugar de poblaciones.

El objetivo del Contraste de Hipótesis es determinar si la diferencia en los pesos medios que hemos observado puede explicarse por azar o es una diferencia que existe en la población. El procedimiento se basa en poner a prueba la Hipótesis Nula de no asociación, calculando cómo de probable sería encontrar una diferencia entre los pesos medios como la que hemos observado (10.1-8.3 = 0.8 kg) o más extrema, es decir más a favor de la Hipótesis Alternativa, si la Hipótesis Nula fuera cierta.

Para calcular esta probabilidad, utilizamos las propiedades de la distribución en el muestreo de la diferencia de medias. La distribución en el muestreo de la diferencia de medias, $\overline{X}_1 - \overline{X}_2$, es aproximadamente $Normal\ (\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$.

Bajo la Hipótesis Nula, $\overline{X}_1 - \overline{X}_2$, se distribuye $Normal\ (0, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$.

El primer paso para calcular la Probabilidad de que la diferencia de medias, $\overline{X}_1 - \overline{X}$, sea mayor o igual que 0.8, consiste en calcular la puntuación z correspondiente. Para ello, a la diferencia de

medias, $\overline{X}_1 - \overline{X}_2$, le restamos la diferencia de medias poblacionales, que bajo la Hipótesis Nula es 0, y lo dividimos por su error estándar:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.8}{\sqrt{\frac{2.1^2}{68} + \frac{1.7^2}{92}}} = \frac{0.8}{0.31} = 2.58$$

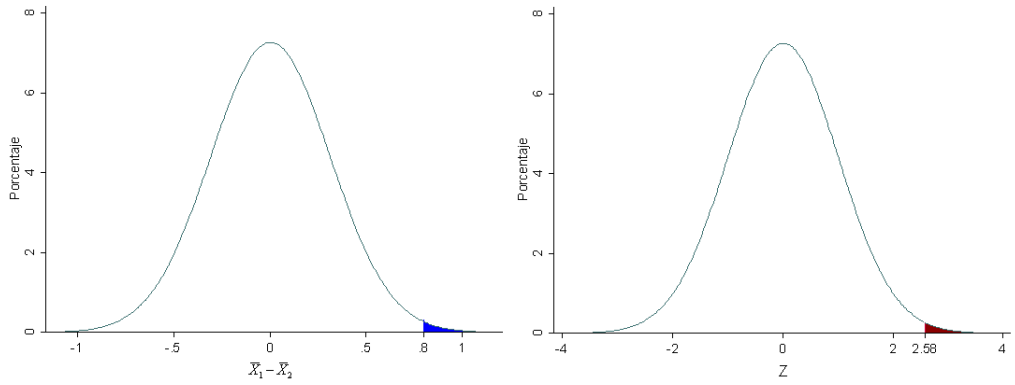
El valor z que acabamos de calcular se conoce con el nombre de test estadístico.

La forma general de un test estadístico es:

$$z = \frac{estimador}{EE(estimador)}$$

El siguiente paso consiste en calcular la Probabilidad de que la puntuación z de una distribución Normal estándar sea mayor o igual que 2.58, (b), lo que es equivalente a calcular la Probabilidad de que la diferencia de medias sea mayor o igual que 0.8, (a):

(a) (b)



*El **segundo paso** de un **Contraste de Hipótesis** se basa en el cálculo del **test estadístico**, cuya forma general es:*

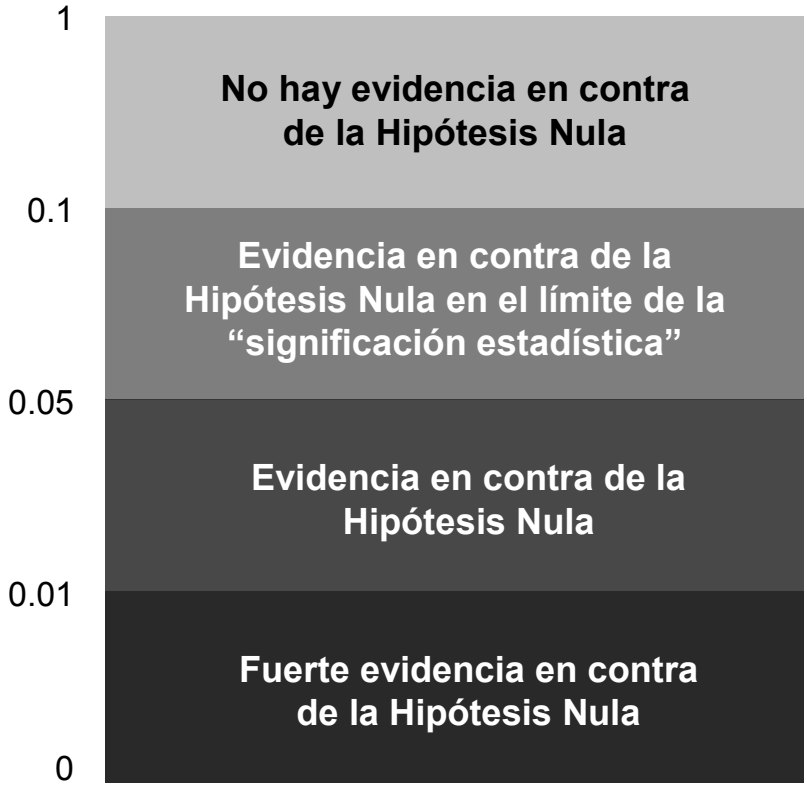
$$z = \frac{estimador}{EE(estimador)}$$

A partir de la Tabla de la distribución Normal Estándar, obtenemos que el área bajo la curva de la Normal estándar que está por encima del valor 2.58 es 0.00494. Este valor, conocido como p-valor unilateral, nos indica cómo de probable sería encontrar una diferencia entre los pesos medios mayor o igual que 0.8 kg. Dado que una diferencia en el otro extremo de la curva, diferencia entre los pesos medios menor o igual que -0.8, también iría a favor de la Hipótesis Alternativa, el procedimiento general consiste en multiplicar por 2 el p-valor unilateral, obteniendo lo que se conoce como p-valor bilateral, o simplemente, p-valor.

En nuestro ejemplo, el p-valor sería 2x0.00494 = 0.01. Si la Hipótesis Nula fuera cierta, es decir, si no hubiera diferencias en

el Peso medio de los niños y las niñas, habría una probabilidad de 0.01 de obtener una diferencia de medias como la que hemos observado o más extrema. Es poco probable que los datos provengan de una población en la que no hay diferencias entre el Peso medio de los niños y el Peso medio de las niñas; por lo tanto, los datos proporcionan evidencia estadística suficiente para rechazar la Hipótesis Nula, es decir, para afirmar que en la población de la que se extrajeron las muestras, el peso medio de los niños NO es el mismo que el peso medio de las niñas.

Conforme el p-valor es más pequeño, mayor es la evidencia en contra de la Hipótesis Nula:



Habitualmente, aunque es algo arbitrario y no puede dársele una consideración estricta, se adopta el valor p = 0.05 como punto de corte por debajo del cual se considera que se dispone de suficientes evidencias para rechazar la Hipótesis Nula, concluyendo que la asociación es estadísticamente significativa. Si el valor de p es superior a 0.05, se considera que es muy probable que las diferencias observadas se deban únicamente al azar, concluyendo que la asociación no es estadísticamente significativa.

En la siguiente tabla se muestra los dos tipos de error que se pueden cometer al realizar un Contraste de Hipótesis:

*El **tercer paso** de un **Contraste de Hipótesis** consiste en calcular el **p-valor**, es decir, la probabilidad de observar una diferencia como la observada en la muestra o más extrema (más a favor de la Hipótesis Alternativa), si la Hipótesis Nula fuera cierta*

Conforme el p-valor es más pequeño, mayor es la evidencia en contra de la Hipótesis Nula. Habitualmente, se considera que si el p-valor es menor de 0.05 se dispone de suficientes evidencias para rechazar la Hipótesis Nula, concluyendo que la asociación es estadísticamente significativa. Si el p-valor es mayor de 0.05, se considera que es muy probable que las diferencias observadas se deban únicamente al azar, concluyendo que la asociación no es estadísticamente significativa.

		VERDAD (REALIDAD)	
		H_0 (Hipótesis nula) *Ausencia de asociación: El azar puede explicar la asociación encontrada en la muestra*	**H_1: Hipótesis alternativa** *La asociación encontrada en la muestra es debida a que existeuna asociación entre el hábito de fumar y el riesgo de cardiopatía coronaria*
DECIDE	**H_0**	Decisión correcta	Error tipo II (β)
	H_1	Error tipo I (α)	Decisión correcta Potencia (1- β)

Error tipo I: Afirmar que la asociación es estadísticamente significativas cuando no lo es.
Error tipo II: Afirmar que no hay asociación significativa cuandorealmente la hay. Este error se puede producir bien porque el efecto sea pequeño (asociación real pero de poca magnitud), bien porque el número de sujetos estudiados sea escaso (tamaño muestral pequeño) o por ambas cosas a la vez.
Potencia (1-β): Capacidad del test para detectar una asociación cuando realmente existe. Un tamaño de muestra muy pequeño reduce la potencia, impidiendo encontrar asociaciones significativas. Un test será tanto mejor cuanta mayor potencia tenga siendo el incremento del tamaño muestral el único modo de aumentar la potencia.

El p-valor indica si la asociación encontrada es estadísticamente significativa pero no mide su magnitud o relevancia, ya que su valor depende tanto de la magnitud de la asociación como del tamaño muestral. Tamaños suficientemente grandes permiten El p-valor indica si la asociación encontrada es estadísticamente significativa pero no mide su magnitud o relevancia, ya que su valor depende tanto de la magnitud de la asociación como del tamaño muestral. Tamaños suficientemente grandes permiten encontrar resultados con altísima significación estadística pero de escasa magnitud, algo que puede carecer de relevancia desde el punto de vista clínico. En el otro extremo, tamaños suficientemente pequeños podrían llevarnos a concluir que una asociación no es estadísticamente significativa por problemas de potencia estadística del test para detectar asociaciones realmente existentes.

Por lo tanto, es fundamental acompañar el p-valor de una medida que cuantifique la magnitud de la asociación en la muestra y un Intervalo de Confianza al 95% para la medida de asociación utilizada. En la situación en la que la variable de exposición es dicotómica (ejemplo: sexo) y la variable de interés es cuantitativa (ej. peso), la medida que nos permite cuantificar la magnitud de la asociación entre ambas variables es la Diferencia de Medias.

En nuestro ejemplo, la diferencia entre el peso medio de los niños y el peso medio de las niñas en los 160 individuos de la muestra es:

$$Diferencia\ Medias = \bar{x}_1 - \bar{x}_2 = 10.1 - 9.3 = 0.8$$

En los 160 individuos de la muestra, la diferencia entre el peso medio de los niños y el peso medio de las niñas es 0.8 kg; los niños pesan, en media, 0.8 kg más que las niñas.

Una vez cuantificada la magnitud de la asociación en los individuos de la muestra, el siguiente paso es cuantificar la magnitud de la asociación en la población. Para ello, calculamos un Intervalo de Confianza al 95% para la Diferencia de Medias.

El p-valor indica si la asociación encontrada es estadísticamente significativa pero no mide su magnitud o relevancia, ya que su valor depende tanto de la magnitud de la asociación como del tamaño muestral.

Intervalo de Confianza al 95% para la Diferencia de Medias

La fórmula general para calcular un Intervalo de Confianza al 95% para un parámetro poblacional es:

$$IC_{95\%}(parámetro) = estimador \pm 1.96 \times EE(estimador)$$

Un Intervalo de Confianza al 95% para la diferencia de medias viene dado por:

$$IC_{95\%}(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm 1.96 \times EE(\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2) \pm 1.96 \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

En nuestro ejemplo, un Intervalo de Confianza al 95% para la Diferencia entre la Media de Peso de los niños y la Media de Peso de las niñas vendría dado por:

$$IC_{95\%}(\mu_1 - \mu_2) = (10.1 - 9.3) \pm 1.96 \times \sqrt{\frac{2.1^2}{68} + \frac{1.7^2}{92}} = 0.8 \pm 1.96 \times 0.31 = (0.19; 1.41)$$

Es fundamental acompañar el p-valor de una medida que cuantifique la magnitud de la asociación en la muestra y un Intervalo de Confianza al 95% para la medida de asociación utilizada.

En la población de la que se extrajeron las muestras, estamos seguros al 95% de que la diferencia entre el peso medio de los niños y el peso medio de las niñas está entre 0.19 y 1.41 kg.

Como el Intervalo de Confianza al 95% no incluye el 0, estamos seguros al 95% de que el peso medio de los niños es diferente al peso medio de las niñas; de hecho, los niños pesan, en media, entre 0.19 y 1.41 kg. más que las niñas.

Muestras pequeñas

Al comparar la media de una variable cuantitativa en dos grupos diferentes de individuos procedentes de muestras de tamaño pequeño, usaremos la distribución t de Student, en lugar de la distribución Normal, para calcular Intervalos de Confianza y realizar Contrastes de Hipótesis.

El número de grados de libertad de la distribución t de Student es: $n_1 + n_2 - 2$ donde n_1 y n_2 son el número de individuos en las muestras 1 y 2, respectivamente.

El procedimiento es similar al utilizado con una muestra, con la

excepción del cálculo del error estándar. En el caso de muestras pequeñas, estimamos una única varianza basada en los datos de las dos muestras.

La varianza común es una media de las varianzas muestrales:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^3}{(n_1 - 1) + (n_2 - 1)}$$

El error estándar de la diferencia de medias es:

$$EE(\bar{x}_1 - \bar{x}_2) = s \times \sqrt{(1/n_1 + 1/n_2)}$$

Se ha diseñado un estudio para determinar la influencia de la hipertensión de los padres en la presión arterial sistólica de los hijos. Se seleccionó un grupo de 12 niños con uno de sus padres hipertenso (grupo 1), y un grupo de 10 niños con ambos padres normotensos (grupo 2), obteniéndose los siguientes datos:

Niños con uno de sus padres hipertenso (grupo 1)	Niños con ambos padres normotensos (grupo 2)
100 102 96 106 110 110 120 112 112 90 111 108	104 88 100 98 102 92 96 100 96 96
$n_2 = 12$ $\bar{x}_2 = 106.4$ $s_2 = 8.1$	$n_1 = 10$ $\bar{x}_1 = 97.2$ $s_1 = 4.7$

En primer lugar, planteamos un Contraste de Hipótesis para determinar si existen diferencias en la Presión arterial sistólica media de los niños en función de que sus dos padres sean normotensos o alguno de ellos sea hipertenso.

Definimos las Hipótesis Nula y Alternativa:

Hipótesis Nula: En la población de la que se extrajeron las muestras, no existe una asociación estadísticamente significativa

entre la hipertensión de los padres y la presión arterial sistólica de los niños; la presión arterial sistólica media de los niños es la misma en aquéllos que tienen padres normotensos que en aquéllos en los que uno de los padres es hipertenso

$$H_0 : \mu_1 = \mu_2$$

Hipótesis Alternativa: En la población de la que se extrajeron las muestras, existe una asociación estadísticamente significativa entre la hipertensión de los padres y la presión arterial sistólica de los niños; la presión arterial sistólica media de los niños NO es la misma en aquéllos niños que tienen padres normotensos que en aquéllos en los que uno de los padres es hipertenso

$$H_1 : \mu_1 \neq \mu_2$$

A continuación, calculamos el test estadístico y el p-valor. Para ello, seguimos los siguientes pasos:

Diferencia de medias en la muestra: $\bar{x}_1 - \bar{x}_2 = 106.4 - 97.2 = 9.2\ mmHg$

Varianza común:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(12-1)\times 8.1^2 + (10-1)\times 4.7^2}{(12-1)+(10-1)} = \frac{920.52}{20} = 46.03$$

$$s = \sqrt{s^2} = \sqrt{46.03} = 6.78$$

Error estándar de la diferencia entre las medias muestrales:

$$EE(\bar{x}_1 - \bar{x}_2) = s \times \sqrt{(1/n_1 + 1/n_2)} = 6.78 \times \sqrt{(1/12 + 1/10)} = 2.90$$

El valor del estadístico t se calcularía como:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{EE(\bar{x}_1 - \bar{x}_2)} = \frac{9.2}{2.90} = 3.17 \qquad gl = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$$

$$p - valor\ (unilateral) = \Pr(t_{20} \geq 3.17) = 0.0024$$
$$p - valor\ (bilateral) = 2 \times 0.0024 = 0.0048$$

El p-valor del Contraste de Hipótesis es 0.0048. Los datos presentan evidencia estadística suficiente para rechazar la Hipótesis Nula. Existe una asociación estadísticamente significativa entre la hipertensión de los padres y la presión arterial sistólica de los niños; los niños que tienen algún padre hipertenso tienen una

presión arterial sistólica media diferente a la de los niños con padres normotensos.

A continuación, cuantificamos la magnitud de la asociación entre la hipertensión de los padres y la presión arterial sistólica de los niños mediante el cálculo de un Intervalo de Confianza al 95% para la Diferencia entre la Media de presión arterial sistólica de niños con algún padre hipertenso y la Media de presión arterial sistólica de los niños con padres normotensos:

$$IC_{95\%}(\mu_1-\mu_2)=(\bar{x}_1-\bar{x}_2)\pm t_{n_1+n_2-2}\times EE(\bar{x}_1-\bar{x}_2)=9.2\pm(2.09\times 2.90)=(3.14;\ 15.26)\ mmHg$$

La diferencia en las medias de la Presión arterial sistólica entre los niños con un padre hipertenso y aquéllos con padres normotensos es 9.2 mmHg. En la población, estamos seguros al 95% de que la diferencia entre las medias está entre 3.14 y 15.26 mmHg.

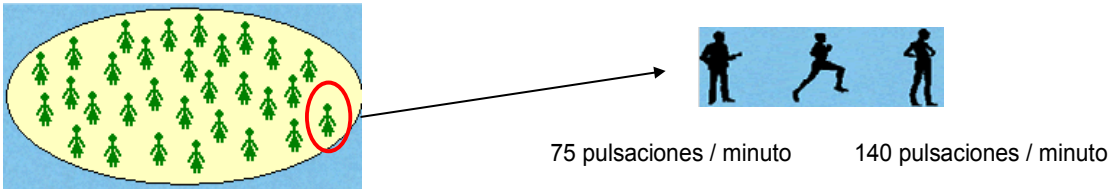
Como el Intervalo de Confianza al 95% excluye al 0 y ambos límites son positivos, estamos seguros al 95% de que los niños con algún padre hipertenso tienen una presión arterial sistólica entre 3.14 y 15.26 mmHg más alta que los niños con ambos padres normotensos.

*El **Contraste de Hipótesis t de Student** permite comparar la media de una variable cuantitativa en dos grupos diferentes de individuos.*

La mayoría de los programas estadísticos utilizan el procedimiento para muestras pequeñas para calcular Intervalos de Confianza y realizar Contrastes de Hipótesis sobre la Diferencia de Medias. Es por ésto por lo que el Contraste sobre la Diferencia de Medias se suele conocer como el Test t de Student para la comparación de medias.

Muestras dependientes

En algunas situaciones, nuestros datos son pares de mediciones realizadas sobre el mismo individuo, en diferentes circunstancias. Por ejemplo, para determinar el efecto del ejercicio físico en las pulsaciones por minuto, se selecciona a un grupo de estudiantes a los que se les pide que corran durante 2 minutos. A cada uno de los estudiantes, se les mide el pulso antes y después de llevar a cabo el ejercicio físico, de forma que cada estudiante tiene dos medidas del pulso, una antes de realizar la carrera y otra después.



Al realizar el análisis, debemos tener en cuenta que nuestros datos están "apareados". Esto se hace calculando las diferencias entre cada par de observaciones apareadas, y aplicando posteriormente los métodos para realizar Inferencias sobre una Media, presentados previamente.

*El análisis de **datos "apareados"** se hace calculando las diferencias entre cada par de observaciones apareadas, y aplicando posteriormente los métodos para realizar Inferencias sobre una Media.*

Se ha diseñado un estudio para determinar si existe una diferencia en el ángulo de torsión conseguido en el brazo derecho e izquierdo en individuos con parálisis en las extremidades superiores. Se han seleccionado 80 individuos a los que se les ha medido el ángulo de torsión conseguido con el brazo derecho e izquierdo, respectivamente.

En la siguiente tabla se muestra la información de los primeros 5 individuos de la muestra:

	Ángulo de torsión (grados)	
Individuo	Brazo derecho	Brazo izquierdo
1	35	30
2	34	28
3	30	29
4	26	26
5	27	28

El primer paso consiste en calcular la diferencia entre cada par de observaciones apareadas; calculamos la diferencia en el ángulo de torsión entre el brazo derecho y el izquierdo de cada individuo. Estas diferencias las denotamos como d_i:

	Ángulo de torsión (grados)		
Individuo	Brazo derecho	Brazo izquierdo	d_i
1	35	30	5
2	34	28	6
3	30	29	1
4	26	26	0
5	27	28	-1

La mayoría de las diferencias observadas son positivas, lo que sugiere que los individuos tienen un mayor ángulo de torsión con el brazo derecho que con el izquierdo.

En los 80 individuos de la muestra, la media y la desviación estándar de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo son: $\bar{d} = 3.9 \quad s_d = 3.3 \ \ grados$

A continuación, planteamos un Contraste de Hipótesis para determinar si en la población de la que se extrajeron las muestras,

existe una asociación estadísticamente significativa entre el brazo derecho e izquierdo y el ángulo de torsión conseguido, es decir, si la media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo es 0.

Definimos las Hipótesis Nula y Alternativa:

$H_0: \delta \neq 0$	En la población de la que se extrajo la muestra, la Media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo es 0
$H_0: \delta \neq 0$	En la población de la que se extrajo la muestra, la media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo es distinta de 0

A continuación, calculamos el valor del test estadístico y del p-valor:

$$z = \frac{\bar{d}}{EE(\bar{d})} = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{3.9}{3.3 / \sqrt{80}} = 10.58$$

El valor z = 10.58 no aparece en la Tabla de la Distribución Normal Estándar, lo que nos indica que la probabilidad de observar una puntuación z mayor o igual que 10.58 es muy pequeña. En estos casos, se suele indicar que el p-valor es < 0.001.

Los datos muestran evidencia estadística suficiente para rechazar la Hipótesis Nula; es decir, en la población de la que se extrajo la muestra, la media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo, respectivamente, es distinto de 0.

A continuación, calculamos un Intervalo de Confianza al 95% para la Media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo en la población de la que se extrajeron las muestras:

$$IC_{95\%}(\delta) = \bar{d} \pm 1.96 \times (s_d / \sqrt{n}) = 3.9 \pm 1.96 \times (3.3 / \sqrt{80}) = (3.2; 4.6)\ grados$$

En los 80 individuos de la muestra, la media de las diferencias

en los ángulos de torsión conseguidos con el brazo derecho e izquierdo, respectivamente, es 3.9 grados.

En la población de la que se extrajo la muestra, estamos seguros al 95% de que la media de las diferencias en los ángulos de torsión conseguidos con el brazo derecho e izquierdo, respectivamente, está entre 3.2 y 4.6 grados.

Como el Intervalo de Confianza al 95% no incluye al 0 y ambos límites son positivos, estamos seguros al 95% de que los individuos con parálisis en las extremidades superiores consiguen entre 3.2 y 4.6 grados más en la torsión con el brazo derecho que con el izquierdo.

Muestras pequeñas

Si el tamaño muestral es pequeño (n<60), utilizaremos la distribución t de Student, en lugar de la distribución Normal, para calcular Intervalos de Confianza y realizar Contrastes de Hipótesis; el número de grados de libertad de la distribución t de Student viene dado por n-1, donde n es el número de individuos.

4.4. Comparación de más de dos medias

Son numerosas las ocasiones en las que el interés se centra en determinar si la media de una variable cuantitativa es igual en más de dos grupos diferentes de individuos.

Supongamos que estamos interesados en determinar si existen diferencias en el Peso de los niños entre 5 y 36 meses residentes en Bolivia en función de su clase social.

Primero, calculamos la media y la desviación estándar del Peso de los niños en función de la clase social (baja, media, alta) en la muestra:

Baja	$n_1 = 41$	$\bar{x}_1 = 7.4$	$s_1 = 0.8$
Media	$n_2 = 77$	$\bar{x}_2 = 9.5$	$s_2 = 0.7$
Alta	$n_3 = 42$	$\bar{x}_3 = 12.2$	$s_3 = 1.0$
Total	$n = 160$	$\bar{x} = 9.7$	$s = 1.9$

En la muestra, el peso medio es: 7.4 kg en los 41 niños de clase social baja, 9.5 kg en los 77 de clase social media, y 12.2 en los

42 de clase social alta. Parece que el peso de los niños es mayor conforme la clase social es mayor.

Pero, ¿las diferencias observadas pueden explicarse por azar? Para responder a esta pregunta, utilizaremos un Contraste de Hipótesis, denominado ANOVA.

*El **Contraste de Hipótesis ANOVA** permite comparar la media de una variable cuantitativa en los grupos definidos por una variable de exposición con 2 ó más categorías.*

El ANOVA permite comparar la media de una variable cuantitativa en los grupos definidos por una variable de exposición con 2 ó más categorías. Se basa en la descomposición de la varianza total de la variable cuantitativa en:

- Variabilidad atribuida a las diferencias entre las medias de los grupos definidos por la variable de exposición
- Variabilidad debida a las diferencias entre las observaciones dentro de cada grupo

Primero, calculamos la varianza del Peso de los 160 niños, ignorando la subdivisión de los niños en función de la clase social:

$$Varianza = s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{586.31}{159} = 3.69$$

Segundo, llevamos a cabo la partición de la varianza.

El numerador de la varianza, denominado **suma de cuadrados** $(SC_{Total} = \sum (x-\bar{x})^2)$, se divide en:

- La suma de cuadrados debida a diferencias entre las medias de los grupos $(SC_{Entre-Grupos} = \sum n_i (\bar{x}_i - \bar{x})^2)$
- La suma de cuadrados debida a diferencias entre las observaciones dentro de cada grupo $(SC_{Dentro-Grupos} = \sum (n_i - 1) \times s_i^2)$.

  Esta suma de cuadrados se conoce como **suma de cuadrados residual.**

El denominador de la varianza, denominado **grados de libertad** (gl = n-1), se divide en:

- k – 1 grados de libertad asociados a la suma de cuadrados entre grupos, siendo k el número de grupos
- n – k grados de libertad asociados a la suma de cuadrados residual

En nuestro ejemplo,

$$SC_{Total} = \sum (x - \bar{x})^2 = 586.31$$
$$SC_{Entre-Grupos} = \sum n_i \times (\bar{x}_i - \bar{x})^2 = [41 \times (7.4 - 9.7)^2] + [77 \times (9.5 - 9.7)^2] + [42 \times (12.2 - 9.7)^2] = 482.47$$
$$SC_{Dentro-Grupos} = \sum (n_i - 1) \times s_i^2 = [(41-1) \times 0.8^2)] + [(77-1) \times 0.7^2)] + [(42-1) \times 1.0^2)] = 103.84$$
$$gl_{Total} = n - 1 = 160 - 1 = 159$$
$$gl_{Entre-Grupos} = k - 1 = 3 - 1 = 2$$
$$gl_{Dentro-Grupos} = n - k = 160 - 3 = 157$$

Construimos la Tabla del Análisis de Varianza:

Fuente de variación	Suma Cuadrados (SC)	Grados libertad (gl)	Media Cuadrática (MC = SC/gl)
Entre Grupos	482.47	2	241.23
Dentro Grupos	103.84	157	0.66
Total	586.31	157	3.69

A continuación, definimos las Hipótesis Nula y Alternativa, calculamos el test estadístico y el p-valor asociado.

Hipótesis Nula (H_0): En la población de la que se extrajeron las muestras, no hay una asociación estadísticamente significativa entre la clase social y el peso; el peso medio de los niños es el mismo en los de clase social baja, media y alta

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

Hipótesis Alternativa (H_1): En la población de la que se extrajeron las muestras, existe una asociación estadísticamente significativa entre la clase social y el peso; el peso medio de los niños varía en función de la clase social

$$H_1 : Existe\ i, j\ tal\ que\ \mu_i \neq \mu_j \qquad (i, j = 1,2,3)$$

El test estadístico utilizado para resolver el Contraste de Hipótesis ANOVA es el cociente entre la Media cuadrática Entre-Grupos y la Media cuadrática Dentro-Grupos:

$$F = \frac{MC_{Entre-Grupos}}{MC_{Dentro-Grupos}} \qquad gl = gl_{Dentro-Grupos} - gl_{Entre-Grupos} = k - 1, n - k$$

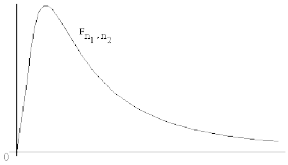
El valor del test estadístico F tomará el valor 1, si no hay diferencias reales entre los grupos, y tomará un valor mayor que 1, si hay diferencias entre los grupos.

Bajo la Hipótesis Nula, el estadístico F sigue una distribución F

de Snedecor con k-1 grados de libertad en el numerador y n-k grados de libertad en el denominador.

Distribución F de Snedecor
Distribución continua fundamentalmente asociada al análisis de la varianza (ANOVA) y a la comparación de varianzas.
Está determinada por dos parámetros: grados de libertad del numerador (n_1) y grados de libertad del denominador (n_2):
$F_{n1,\ n2}$ es una distribución F de Snedecor con n_1 grados de libertad en el numerador y n_2 grados de libertad en el denominador.
Su rango es el eje real positivo: (0, +∞)

Figura. Función de densidad de una distribución F de Snedecor con n_1 grados de libertad en el numerador y n_2 grados de libertad en el denominador



En nuestro ejemplo,

$$F = \frac{MC_{Entre-Grupos}}{MC_{Dentro-Grupos}} = \frac{240.10}{0.68} = 353.09 \qquad con\ (2, 157)\ grados\ de\ libertad$$

$$p - valor = \Pr(F_{2,157} \geq 353.09) < 0.001$$

El p-valor asociado al contraste de hipótesis es <0.001. Los datos muestran evidencia estadística suficiente para rechazar la Hipótesis Nula. Existe una asociación estadísticamente significativa entre la Clase social y el Peso; el peso medio de los niños entre 5 y 36 meses residentes en Bolivia varía en función de la clase social.

El contraste de hipótesis ANOVA se basa en dos asunciones: (1) La distribución de la variable de interés es aproximadamente Normal, y (2) La desviación estándar poblacional de la variable de interés es la misma en los diferentes grupos definidos por la variable de exposición.

Desviaciones moderadas de la Normalidad pueden ignorarse, pero el efecto de desviaciones estándar desiguales puede ser serio.

Cuando sólo hay 2 grupos, el Análisis de la Varianza da exactamente el mismo resultado que el contraste t de Student para la comparación de la media en 2 grupos diferentes de individuos.

Cuando sólo hay 2 grupos, el Análisis de la Varianza da exactamente el mismo resultado que el contraste t de Student para la comparación de la media en 2 grupos diferentes de individuos.

4.5. Correlación y Regresión Lineal

En la investigación en salud surge frecuentemente la necesidad de estudiar la relación entre dos variables cuantitativas. Abordamos el caso en el que la relación entre las variables es lineal.

Supongamos que estamos interesados en estudiar si existe una relación entre la Edad y el Peso de los niños entre 5 y 36 meses de Bolivia.

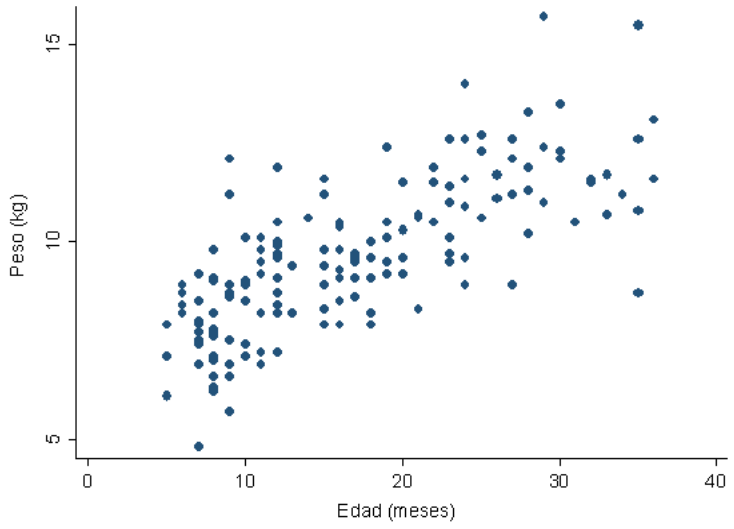
Correlación

La relación entre dos variables cuantitativas puede explorarse gráficamente mediante un Diagrama de dispersión y numéricamente, mediante el Coeficiente de correlación lineal de Pearson.

*La relación entre dos variables cuantitativas puede explorarse gráficamente mediante un **Diagrama de dispersión** y numéricamente, mediante el **Coeficiente de correlación lineal de Pearson**.*

Diagrama de dispersión

Es un gráfico que permite representar conjuntamente dos variables cuantitativas para examinar la posible relación entre ellas. El Diagrama de Dispersión de la Edad y el Peso de los 160 niños de la muestra es:



Cada par de valores de Edad y Peso se representan por un símbolo donde la posición horizontal se determina por el valor de la primera variable (Edad) y la posición vertical viene determinada por el valor de la segunda variable (Peso). Por convención, la variable de interés se representa en el eje vertical y la variable de exposición en el eje horizontal.

*En el **Diagrama de Dispersión**, el sentido de la asociación viene determinado por la inclinación de la nube de puntos: positiva, si valores altos de una variable se asocian con valores altos de la otra; y negativa, si valores altos de una variable se asocian con valores bajos de la otra. La fuerza de la asociación viene determinada por lo aplastado de la nube de puntos; asociación más fuerte conforme los puntos estén más cerca unos de otros.*

El Diagrama de Dispersión nos proporciona información sobre el sentido y la fuerza de la asociación entre las dos variables. El sentido de la asociación viene determinado por la inclinación de la nube de puntos: positiva, si valores altos de una variable se asocian con valores altos de la otra; y negativa, si valores altos de una variable se asocian con valores bajos de la otra. La fuerza de la asociación viene determinada por lo aplastado de la nube de puntos; asociación más fuerte conforme los puntos estén más cerca unos de otros.

El Diagrama de dispersión de nuestro ejemplo muestra la existencia de una relación positiva fuerte entre la Edad de los niños y su Peso; valores altos de edad se asocian con valores altos de peso.

Coeficiente de correlación lineal de Pearson (r)

Mide el grado de relación lineal entre dos variables cuantitativas. Se calcula como:

$$r = \frac{Covarianza(x,y)}{Desviación\ estándar(x) \cdot Desviación\ estándar(y)}$$, siendo

$$Covarianza(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

En nuestro ejemplo, el valor del coeficiente de correlación lineal entre edad y peso es 0.74.

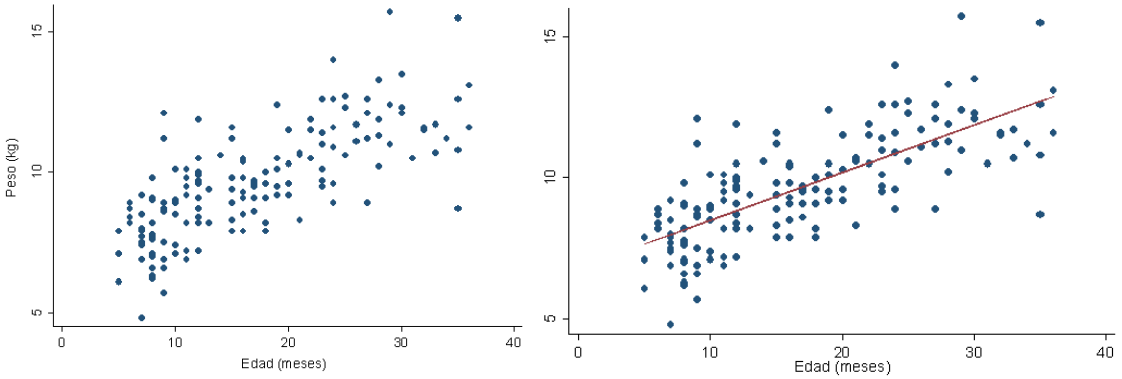
El coeficiente de correlación lineal es un número comprendido entre -1 y 1. El signo indica el sentido de la asociación: positiva si r > 0, negativa si r < 0 y ausencia de correlación lineal si r = 0. La magnitud absoluta indica la fuerza de la asociación.

En nuestro ejemplo, r = 0.74, muestra una relación lineal positiva entre la Edad y el Peso de los niños; valores altos de Edad se asocian con valores altos de Peso.

Regresión Lineal

La Correlación cuantifica la fuerza de la asociación entre dos variables cuantitativas, tratándolas de modo simétrico. La Regresión Lineal permite estudiar la relación entre dos variables cuantitativas, describiendo el comportamiento de una variable en función de la otra.

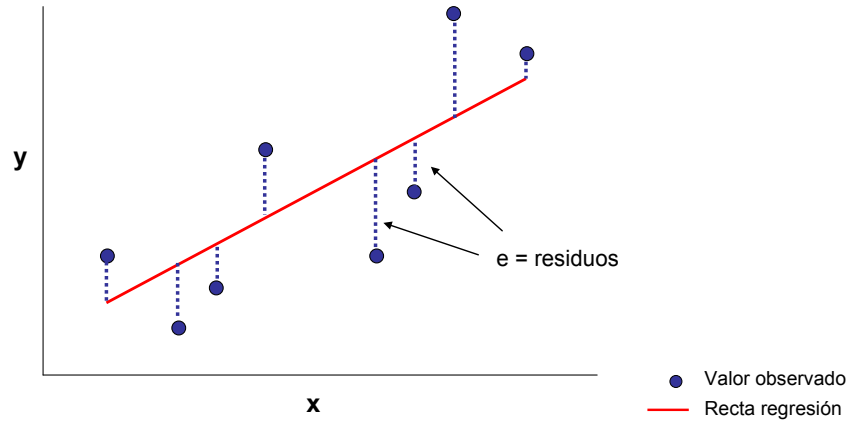
La idea intuitiva de la Regresión Lineal consiste en intentar resumir la información del Diagrama de Dispersión mediante una Recta que se ajuste a la nube de puntos.



*La **Regresión Lineal** permite estudiar la relación entre dos variables cuantitativas, describiendo el comportamiento de una variable en función de la otra*

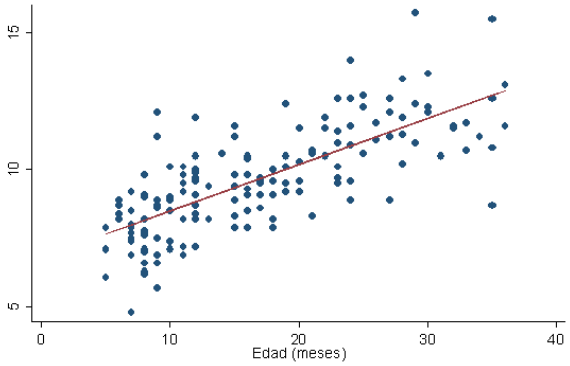
El objetivo es determinar la línea recta que mejor describa la relación entre la variable de exposición y la variable de interés, es decir, entre la Edad y el Peso. Intuitivamente, la Recta de Regresión será aquélla que esté más cerca de todos los puntos. Para determinar esta recta se utiliza el Método de los Mínimos Cuadrados, que elige como recta de regresión aquélla que minimiza las distancias verticales de las observaciones a la recta, tal y como se refleja a continuación:

*El objetivo de la Regresión Lineal consiste en determinar la línea recta, denominada **Recta de Regresión**, que mejor describa la relación entre la variable de exposición y la variable de interés.*



La distancia vertical entre el valor observado y el valor ajustado por la recta, se denomina residuo (e_i). Los residuos pueden ser positivos o negativos y al sumarlos podrían cancelarse. Por tanto, el Método de los Mínimos Cuadrados se basa en la Minimización de la Suma de los Residuos al cuadrado (e_i^2).

A la recta que minimiza la suma de los residuos al cuadrado se le denomina **Recta de Regresión.** En nuestro ejemplo, la Recta de Regresión es:



*El **intercepto (a)** se interpreta como la media de **y** cuando **x** toma el valor 0; la media de la variable de interés cuando la variable de exposición vale 0. Si la variable de exposición no puede tomar el valor 0, el intercepto no es interpretable. La **pendiente (b)** se interpreta como el cambio por término medio en **y** por cada aumento de una unidad en **x**; el cambio por término medio en la variable de interés por cada aumento de una unidad en la variable de exposición.*

La expresión matemática de la Recta de Regresión es:

$$y = a + bx$$

donde ***a*** es el intercepto y ***b*** es la pendiente de la recta.

El intercepto (a) se interpreta como la media de ***y*** cuando ***x*** toma el valor 0; la media de la variable de interés cuando la variable de exposición vale 0. Si la variable de exposición no puede tomar el valor 0, el intercepto no es interpretable.

La pendiente (b) se interpreta como el cambio por término medio en ***y*** por cada aumento de una unidad en ***x***; el cambio por término medio en la variable de interés por cada aumento de una unidad en la variable de exposición.

En nuestro ejemplo, la Recta de regresión es:

peso = 6.81 + 0.169 edad

Como la edad no puede tomar el valor 0, el intercepto (a = 6.81) no es interpretable.

La pendiente (b = 0.169) se interpretaría como: *El peso de los niños se incrementa, en media, 169 gramos por cada aumento de 1 mes en su edad.*

En los individuos de la muestra, se observa una relación positiva entre la edad y el peso. Pero, ¿esta relación puede ser explicada por azar o existe en la población?

Para responder a esta pregunta, hacemos lo siguiente: Primero, realizamos un Contraste de Hipótesis sobre la Pendiente de la Recta de Regresión en la población, β, para determinar si existe una relación lineal entre la Edad y el Peso de los niños en la población. Segundo, calculamos un Intervalo de Confianza al 95% para la Pendiente de la Recta de Regresión para cuantificar

la magnitud de la asociación entre la Edad y el Peso en la población.

Contraste de Hipótesis para la Pendiente (ß)

Primero, definimos las Hipótesis Nula y Alternativa:

Hipótesis Nula: En la población de la que se extrajo la muestra, no hay una relación lineal entre la Edad y el Peso de los niños

$$H_0 : \beta = 0$$

Hipótesis Alternativa: En la población de la que se extrajo la muestra, hay una relación lineal entre la Edad y el Peso de los niños

$$H_0 : \beta \neq 0$$

Segundo, calculamos el valor del test estadístico. Asumiendo que EE(b) es 0.012, el valor del test estadístico sería:

$$t = \frac{b}{EE(b)} = \frac{0.169}{0.012} = 13.83$$

Bajo la Hipótesis Nula, el estadístico t sigue una distribución t de Student con n-2 grados de libertad. Por lo tanto, el p-valor se calcularía como:

$$p - valor = \Pr(t_{n-2} \geq t) = \Pr(t_{158} \geq 13.83) < 0.001$$

El p-valor del contraste es <0.001. Los datos muestran evidencia estadística suficiente para rechazar la Hipótesis Nula; es decir, para afirmar que existe una relación lineal entre la Edad y el Peso de los niños entre 5 y 36 meses residentes en Bolivia.

Intervalo de Confianza al 95% para la Pendiente (ß)

Un Intervalo de Confianza al 95% para β se calcula como:

$$IC_{95\%}(\beta) = b \pm t_{n-2} \times EE(b)$$

donde t_{n-2} es el punto de la distribución t de Student con n-2 grados de libertad que deja en las colas una probabilidad del 5%.

En nuestro ejemplo,

$$IC_{95\%}(\beta) = b \pm t_{n-2} \times EE(b) = 0.169 \pm 1.96 \times 0.012 = (0.145;\ 0.193)$$

En la población de niños entre 5 y 36 meses residentes en Bolivia,

el peso se incrementa, en media, entre 145 y 193 gramos por cada aumento de 1 mes en su edad.

Asunciones de la Regresión Lineal

1. La relación entre la variable de exposición y la variable de interés es lineal

Además, para que los Intervalos de Confianza y los p-valores sean correctos, debe cumplirse:

2. Los residuos siguen una distribución Normal
3. Los residuos tienen varianza constante

5. Métodos no paramétricos

***Los métodos no paramétricos** se utilizan para analizar variables de interés que no siguen una distribución Normal. Se basan en el análisis de los rangos, reemplazando cada valor de la variable de interés por su rango correspondiente.*

Los métodos presentados hasta ahora, conocidos como métodos paramétricos, asumen que la variable de interés sigue una distribución aproximadamente Normal. Pero, ¿qué método utilizar si esta asunción no se cumple?

Métodos no paramétricos: Se utilizan para analizar variables de interés que no siguen una distribución Normal. Se basan en el análisis de los rangos, reemplazando cada valor de la variable de interés por su rango correspondiente.

Bootstrapping: Técnica que permite calcular Intervalos de Confianza haciendo muy pocas asunciones sobre la distribución de la variable de interés.

Errores estándar robustos: Técnica que permite calcular Intervalos de Confianza y Errores Estándar a partir de la distribución observada, y no asumida, de la variable de interés.

A continuación, se presentan las principales ventajas y desventajas de los Métodos no Paramétricos.

Ventajas

Son más robustos que los métodos paramétricos, en el sentido de que están menos afectados por observaciones extremas.

Limitaciones

Los métodos no paramétricos se han utilizado tradicionalmente para realizar Contrastes de Hipótesis; el desarrollo de métodos no paramétricos para el cálculo de Intervalos de Confianza es

muy reciente. Esto representa una limitación en la estadística moderna en la que se presta mucha atención a la estimación de la magnitud de las asociaciones, y a la interpretación de los p-valores en el contexto de los Intervalos de Confianza.

P-valores grandes resultantes de comparar dos muestras pequeñas mediante métodos no paramétricos, se han mal interpretado, en ausencia de intervalos de confianza, como ausencia de diferencias entre los grupos, cuando en realidad los datos podrían ser compatibles tanto con la ausencia de diferencias como con la existencia de diferencias.

Además, los métodos no paramétricos presentan mayores dificultades que los métodos paramétricos, para generalizarlos a situaciones en las que se desea tener en cuenta el efecto que más de una variable de exposición tiene en la variable de interés.

A continuación, se muestra una Tabla que resume los principales Contrastes de Hipótesis, y en la que se incluye la alternativa no paramétrica de los métodos paramétricos presentados.

Tabla. Principales Contrastes de Hipótesis

	Variable de Interés	Variable de Exposición		Contraste de Hipótesis	Ejemplo
MUESTRAS INDEPENDIENTES	***Cuantitativa***	***Categórica***			
		2 grupos			
			Paramétrico	t-Student	*Comparar niveles medios de hematocrito en un grupo estudio y uno control*
			No paramétrico	U Mann-Whitney	*Comparar si el grado de dolor (de 0 a 10) es igual en dos grupos tratados con dos técnicas diferentes*
		> 2 grupos			
			Paramétrico	ANOVA	*Comparar si nivel medio de colesterol es el mismo en 3 grupos con dietas diferentes*
			No paramétrico	Kruskal-Wallis	*Comprobar si el grado de ansiedad (de 0 a 10) es el mismo en 3 enfermedades distintas*
		Cuantitativa			
		Paramétrico		Pearson/ Regresión lineal	*Relación entre tensión arterial sistólica y edad*
		No paramétrico		Spearman/Kendall	*Analizar asociación entre grado ansiedad (0 a 100) y sensación de dolor en parto (0 a 20) de embarazadas*
	Dicotómica	***Categórica***		Chi-cuadrado Exacto de Fisher	*Analizar si existe asociación entre el uso de medicación tiroidea y la aparición de cáncer de mama*
MUESTRAS DEPENDIENTES	***Cuantitativa***	***Categórica***			
		2 grupos			
			Paramétrico	t-Student apareado	*Analizar si tensión sistólica media de grupo de pacientes es la misma antes y después de tratamiento*
			No paramétrico	Signo-rango de Wilcoxon	*Comparar si grado de ansiedad (de 0 a 10) de grupo de pacientes es igual antes y después de una terapia*
		> 2 grupos			
			Paramétrico	ANOVA de Medidas Repetidas	*Comparar la tensión sistólica media en un momento base, a los 3 meses y a los 6 meses de tratamiento*
			No paramétrico	Friedman	*Comparar si grado de dolor (de 0 a 20) de grupo de pacientes es el mismo antes y después de tratamiento*
	Categórica	***Categórica***			
		2 grupos		Test de McNemar	
		> 2 grupos		Test Cochran	

Conclusiones

La Estadística juega un papel fundamental en la Investigación en Salud, tanto en las etapas de diseño como en la selección de muestras y análisis de datos. En este tema se han descrito los métodos para realizar el análisis descriptivo de los individuos de la muestra. Y, se han desarrollado los métodos que nos permiten estimar la media de una variable cuantitativa en la población a estudio, determinar qué variables de exposición se asocian con la variable de interés cuantitativa, y cuantificar la magnitud de la asociación entre la variable de exposición y la variable de interés, en la población a estudio.

Referencias bibliográficas

1. *Peña D, Romo J. Introducción a La Estadística para las Ciencias Sociales. Editorial McGraw Hill, 2003*

2. *Martínez M.Bioestadística amigable. Editorial Díaz de Santos, 2006*

3. *Hernández-Aguado I, Gil A, Delgado M, Bolumar F. Manual de Epidemiología y Salud Pública. Editorial Médica Panamericana, 2005*

4. *Kirkwood B, Sterne J. Essential Medical Statistics. Blackwell Science Ltd, 2001.*