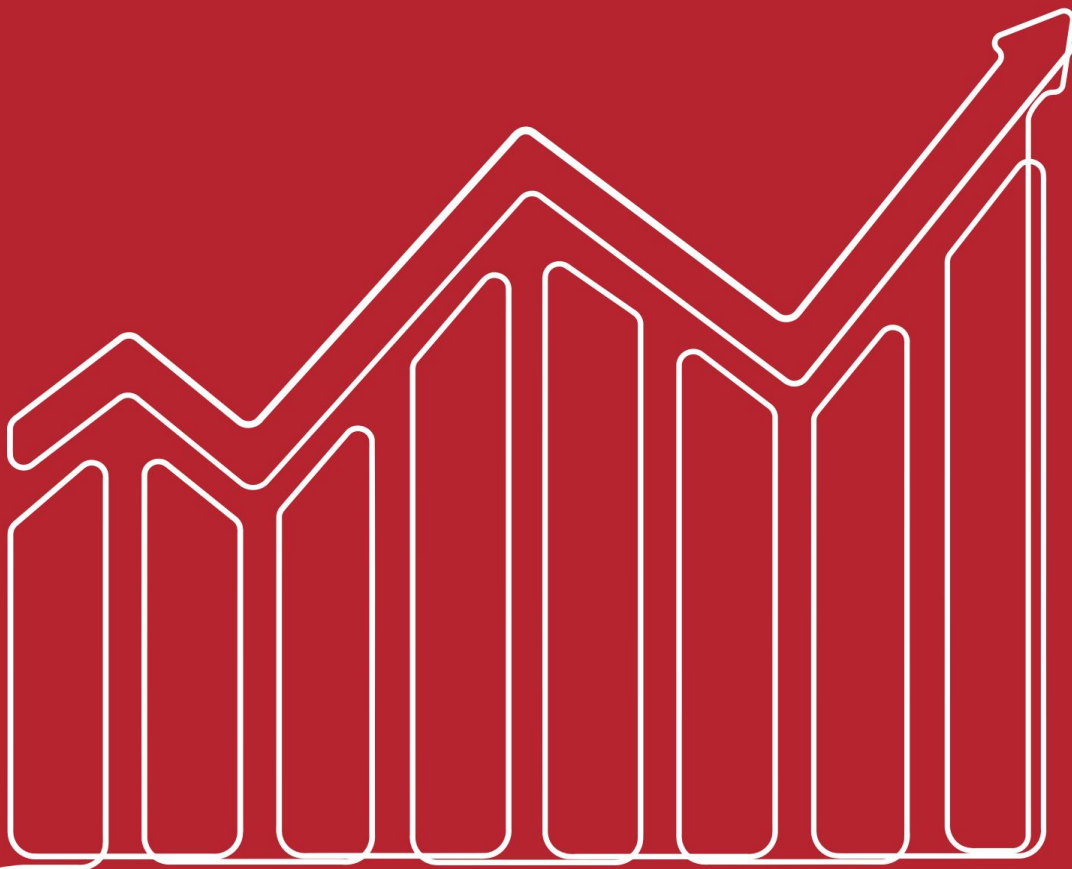


ESTADÍSTICA

— Christian Ochoa —



ÍNDICE

01

pág. 11

Conceptos y Estadística descriptiva

- Introducción
- Bioestadística
- Variables
- Estadística Descriptiva – Gráficos
- Gráficos Exploratorios

02

pág. 19

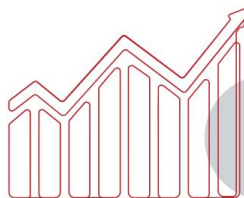
Estadística Inferencial

- Introducción
- Teoría de Probabilidades
- Distribución de Probabilidades
- Muestreo
- Estadística Inferencial

03

pág. 26

Cuestionario



01. CONCEPTOS Y ESTADÍSTICA DESCRIPTIVA

Introducción

En el Manual de Epidemiología aprendimos los fundamentos de la investigación clínica. Aprendimos la importancia de la pregunta de investigación, definir la población de estudio, diseñar el estudio, entre otras cosas. Sin embargo, ¿Qué vamos a hacer con los datos que generamos u obtendremos en el estudio? ¿Cómo analizaremos los datos y cómo vamos a interpretar los resultados? En este Manual nos introduciremos dentro de la bioestadística, donde les proporcionaremos conocimientos y herramientas para formular un plan de análisis de datos que nos ayude a responder estas preguntas en la parte clínica y a la hora del examen.

Bioestadística



Recuerda

La Estadística puede clasificarse en 2 categorías: Estadística Descriptiva y la Estadística Inferencial.

- La bioestadística proporciona métodos para describir y resumir los datos que hemos recogido de una muestra y nos permite extrapolar los resultados para hacer inferencias sobre la población de la que se ha extraído la muestra. La Estadística puede clasificarse en 2 categorías: Estadística Descriptiva y la Estadística Inferencial.
- El término Estadística Descriptiva se refiere a las medidas que resumen y caracterizan un conjunto de datos que nos permiten comprender mejor los atributos de un grupo o población. Cabe mencionar que, estas medidas pueden ser gráficas o numéricas. Mientras que la Estadística Inferencial y la prueba de hipótesis tienen como objetivo utilizar los datos de la muestra para aprender sobre la población de la que se ha extraído esta, basándose en la teoría de la probabilidad.

Variables



Recuerda

Se conoce como variable a toda característica o atributo susceptible de tomar un valor y ser clasificado o medido.

- Se conoce como variable a toda característica o atributo susceptible de tomar un valor y ser clasificado o medido. Debe tener dos o más valores. Cuando hablamos del valor operacional de la variable, nos referimos al lugar donde nosotros exponemos cómo vamos a medir nuestra variable. Ahora, esto depende de una herramienta, del instrumento de medición. Con este se le asignará un tipo (cualitativa o cuantitativa) y una escala (nominal, ordinal, razón, intervalo). Cuando nos referimos al tipo, nos referimos a si la variable será cualitativa o cuantitativa. Por ejemplo, imaginémonos que queremos medir la luz de nuestra habitación. Cualitativamente podemos medirla mencionando si hay o no hay luz. Mientras que cuantitativamente la podemos medir a través de la candela (unidad de medida de luz).
- Las variables cualitativas se clasifican en dicotómicas o politómicas dependiendo de

la cantidad de valores que serán asignadas. Por ejemplo, al querer medir la luz de la habitación, una variable dicotómica solo tendría 2 opciones (hay luz o no hay luz). Mientras que la variable politómica puede tener más de 1 valor (no hay luz, poca luz, regular luz, mucha luz)([ENAM 2019-A](#)).

- Las variables cuantitativas pueden dividirse en discretas (cuando los valores sólo pueden asignarse como número enteros) o en continuas (cuando se dan en número racionales).

Por ejemplo:

— Variables Cuantitativas Discretas:

- ◆ Los datos discretos son valores numéricos que representan cantidades medibles. Se limitan a valores enteros y a menudo se denominan datos de recuento. Ejemplos de datos discretos son el número de muertes en Perú en 2012 y el número de años que un grupo de individuos ha recibido educación, el número de hijos.

— Variables Cuantitativas Continuas:

- ◆ Los datos continuos también representan cantidades medibles, pero no se limitan a valores enteros (números racionales) y pueden incluir valores fraccionarios y decimales. Por lo tanto, la diferencia entre dos valores puede ser arbitrariamente pequeña en función de la precisión de nuestro instrumento de medición. Algunos ejemplos de datos continuos son la temperatura, el peso y el nivel de colesterol en sangre, entre otros.

CUALITATIVAS	CUANTITATIVAS
DICOTÓMICAS (Ejemplo: Masculino o Femenino)	DISCRETAS (Ejemplo: 1,2,3,4,5)
POLITÓMICAS (Ejemplo: lavado de manos, uso de guantes, uso de mascarilla, etc)	CONTINUAS (Ejemplo, 1.1, 1.2, 1.3, etc)

Tabla 1. Resumen de la medición de Variables según el Tipo

“

Recuerda

La Escala de Medición es el grado de exactitud con que se va a expresar la medida de una variable

La razón asigna distancias constantes y numéricas entre categorías, pero el 0 es REAL.

Por otro lado, la Escala de Medición es el grado de exactitud con que se va a expresar la medida de una variable. Según el instrumento que usamos para acercarnos a la verdad, tendremos una escala de medición Nominal, Ordinal, Intervalo y Razón.

- **Nominal:** representan categorías o clases de datos no ordenadas. Por ejemplo, una de las posibles formas de categorizar la etnia en los seres humanos es "Blanco", "Negro" y "Otras etnias"
- **Ordinal:** Cuando existe un orden natural entre las categorías, los datos se denominan ordinales. La clasificación de la New York Heart Association (NYHA) describe cuatro categorías de insuficiencia cardíaca según la gravedad de los síntomas y el grado de limitación para realizar las actividades diarias actividades diarias o el score de APGAR. (Ver tabla 2).
- **Intervalo:** se le conoce así porque asigna distancias constantes y numéricas entre las categorías. Por ejemplo, medir la temperatura con el termómetro en grados Celsius, por cada grado de temperatura se dilata el mercurio, los valores que aumentan son constantes. En esta escala el cero no significa ausencia real del fenómeno. Es arbitrario. Por ejemplo, 0° Celsius no significa ausencia de temperatura.
- **Razón:** asigna distancias constantes y numéricas entre categorías, pero el 0 es REAL. En otras palabras, el 0 es ausencia total del fenómeno. Por ejemplo, medir la temperatura, pero en grados Kelvin, donde el 0 representa lo absoluto. Otros ejemplos son las unidades de los sistemas internacionales de medidas (metro, milímetros, kilo, litro, newton, etc) .

TIPO	CARACTERÍSTICAS	EJEMPLOS
NOMINAL	Categorías o clases no ordenadas, nomenclatura especial: dicotómica o binario- 2 valores distintos	Raza: Blanco, Negra, Asiática, otros. Género: Femenino y masculino
ORDINAL	Categorías ordenadas, la magnitud no es importante (ENAM 2012-B)	New York Heart Association (NYHA) Clasificación de la IC: I, II, III, IV
INTERVALO	El espacio entre los valores es significativo	Temperatura en grados C°
RAZÓN	Posibilidad de valores fraccionarios y decimales, se pueden aplicar operaciones aritméticas	Temperatura en grados K° donde el 0 es absoluto

Tabla 2. Escalas de medición, características y ejemplos.

Estadística descriptiva – Gráficos

El primer paso en el análisis de datos es describir o resumir los datos que has recogido mediante tablas, gráficos y/o valores numéricos. Este es un paso importante, porque nos permitirá evaluar cómo se distribuyen los datos y cómo deben ser analizados.

Los gráficos proporcionan una evaluación general de los datos y pueden permitirle comprender rápidamente cómo se distribuyen los datos o encontrar patrones y relaciones entre las variables. Las estadísticas de resumen numérico son números que representan los datos y resumen cuantitativamente lo que puede verse a través de los gráficos. Tanto las medidas de resumen gráfico como las numéricas constituyen estadísticas descriptivas.



Recuerda

Gráfico de Dispersión: sirve para ver asociación de 2 variables.

Si estás contando variables cualitativas o cuantitativas discretas se usa principalmente las Barras Simples, gráfico donde las bases no se tocan.

A. Gráficos exploratorios

- **Gráfico de Tallo y hojas (stem plot):** sirve para explorar las unidades de todos los datos con el objetivo de no perder el dato individual (**ENAM 2006-A, ENAM 2014-B, ENAM 2006-A**).
- **Gráfico de Dispersión:** sirve para ver asociación de 2 variables, también es útil para evaluar la correlación (R^2) (**ENAM 2020**).
- **Gráfico de Cajas (boxplot):** ayuda a comparar grupos de acorde a sus medianas y su rango intercuartilar. De tal manera que nos permita tener una mejor lectura a la hora de comparar ambos grupos.

B. Gráficos de distribuciones de frecuencia

Son aquellos gráficos que se utilizan para contar la frecuencia de ocurrencia de una variable.

- Si estás contando variables cualitativas o cuantitativas discretas se usa:
 - ♦ Barras simples → gráfico donde las bases no se tocan (**ENAM 2004-B**)
 - ♦ Circular o "pie" → expresan porcentajes

“

Recuerda

En el histograma las bases de las columnas se tocan porque los números son infinitos.

- ◆ Pictograma → se le añade un dibujo explicativo
- ◆ Dot chart → compara 2 o más mediciones por categoría.
- Si estás contando variables cuantitativas continuas:
 - ◆ Histograma → las bases de las columnas se tocan porque los números son infinitos (**ENAM 2006-A**).
 - ◆ Polígonos de frecuencias → se utiliza para comparar varios histogramas a la vez. Se trazan líneas a partir de los puntos medios de las columnas del histograma.

TASAS	CÁLCULO	TASAS	CÁLCULO
Gráfico de Tallo y Hojas	Tallo 2 3 4 5 1 3 6 0 6 7 Hoja	Pictograma	
Gráfico de Dispersión		Dot Chart	
Gráfico de Cajas (boxplot)		Histograma	
Barras Simples		Polígono de Frecuencias	
Circular o Pie			

B. Gráficos de tendencia

Gráficos que sirven para predecir estadísticamente el futuro y permiten evaluar las variables y/o medidas en el tiempo. Solo aplica con variables cuantitativas. Tenemos las curvas que responde a una función matemáticas y de estas, un tipo especial son las curvas lineales. Se les conoce como lineal porque la función representa: $X + Y$ (ambos elevados a la 1) y son las más usadas en la bioestadística.

Medidas de Resumen Numérico

Nos permite resumir las medidas en números para permitirnos hacer cálculos con las variables. Para ello tenemos que evaluar primero que tipo es nuestra variable. Al igual que con la representación gráfica, la elección de la representación numérica dependerá del tipo de variable que se esté considerando.

“

Recuerda

Razón: X/Y nos permite ver la relación numérica entre 2 variables cualitativas.

“

Recuerda

- Tasa de mortalidad infantil: (muertos < 1 año/nacidos vivos) $\times 1000$
- Razón de mortalidad materna (muertesmaternas/nacidos vivos) $\times 100000$

“

Recuerda

La media es sensible a los valores extremos y es única.

Para variables Cualitativas:

- **Razón:** X/Y , nos permite ver la relación numérica entre 2 variables cualitativas (**ENAM 2003-A**). Por ejemplo, cuántos cardiólogos hay por cada neurólogo. Cabe resaltar que X e Y son totalmente excluyentes, no tienen nada que ver una con la otra (**ENAM EXTRA 2021-II**).
- **Proporción:** $X/X+Y$, son las que permiten ver el porcentaje (%) o la fracción que representa del total. En otras palabras, yo puedo ver cuántos cardiólogos hay de entre el total de médicos.
- **Tasa:** ocurre cuando expresas la Razón o la Proporción en función de las variables de la estadística descriptiva. Entre estas tenemos tiempo, persona y lugar. A continuación, algunas tasas epidemiológicas fundamentales (ver tabla 3).

TASAS	CÁLCULO
TASA DE LETALIDAD	$\left(\frac{\text{MUERTES}}{\text{ENFERMOS}} \right) \times 100$
TASA DE MORTALIDAD INFANTIL	$\left(\frac{\text{MUERTOS} < 1 \text{ AÑO}}{\text{NACIDOS VIVOS}} \right) \times 1000$
TASA DE MORTALIDAD MATERNA	$\left(\frac{\text{MUERTES MATERNAS}}{\text{MUJERES EN EDAD FÉRTIL}} \right) \times 1000$
RAZÓN DE MORTALIDAD MATERNA	$\left(\frac{\text{MUERTES MATERNAS}}{\text{NACIDOS VIVOS}} \right) \times 100000$

Tabla 3. Principales Tasas a nivel nacional

Para variables Cuantitativas:

Medidas de Tendencia Central

A. Moda:

- El valor más frecuente de un determinado conjunto de datos se denomina moda. La moda puede ser un resumen estadístico útil para los datos categóricos u ordinales, pero normalmente no es informativa para los datos discretos o continuos, ya que los valores únicos pueden ocurrir con baja frecuencia

110mmHg	134mmHg	126mmHg	154mmHg	168mmHg
128mmHg	168mmHg	158mmHg	170mmHg	188mmHg

- Partamos de este ejemplo: la moda sería el valor que más se repite, en este caso 168 mmHg (2 veces)

168 mmHg (2 veces)



Recuerda

Si la distribución de los datos es simétrica y unimodal, la media, la mediana y la moda deben ser las mismas. Cuando los datos no son simétricos, la mediana es la mejor medida de la tendencia central.

B. Promedio o media:

La medida de tendencia central más común para los datos discretos y continuos es la media, también denominada promedio. La media de una variable se calcula sumando todas las observaciones y dividiéndolas por el número total de observaciones. Por lo tanto, la media es sensible a los valores extremos y es única. **(ENAM 2005-B)** Tomando como ejemplo la figura anterior, en este grupo de presiones el cálculo del promedio sería:

$$(110 + 134 + 126 + 154 + 168 + 128 + 168 + 158 + 170 + 188) / 10 = 150.4 \text{ mmHg.}$$

C. Mediana:

— La mediana se define como el número medio de una lista de valores ordenados de menor a mayor **(ENAM EXTRA 2021-II)**. (Si no hay un número medio, la mediana es la media de los dos valores medios). La mediana es una medida de tendencia central que no es tan sensible a los valores atípicos extremos en comparación con la media. Puede utilizarse para resumir datos discretos o continuos. Tomando el ejemplo anterior, ordenamos de menor a mayor:

110, 126, 128, 134, 154, 158, 168, 168, 170, 188

— Dado que tenemos 10 valores, la mediana es el promedio entre los 2 valores medios, en este caso:

$$154 \text{ y } 158 \rightarrow 156 \text{ mmHg.}$$

— OJO: La medida de tendencia central más adecuada depende de la distribución de los valores. Si la distribución de los datos es simétrica y unimodal, la media, la mediana y la moda deben ser las mismas. En este caso, se suele preferir la media. Cuando los datos no son simétricos, la mediana es la mejor medida de la tendencia central.

Medidas de Dispersión

Aunque dos distribuciones diferentes pueden tener la misma media, mediana y moda, el conjunto total de datos puede ser muy diferentes. Por ello, las medidas de dispersión son necesarias para describir mejor los datos y complementar la información proporcionada por las medidas de tendencia central.

A. Rango

El rango de un grupo de observaciones se define como la diferencia entre la observación más grande y la más pequeña. El rango es fácil de calcular y nos da una idea aproximada de la dispersión de los datos; sin embargo, su utilidad es limitada. El rango es muy sensible a los valores atípicos, ya que sólo tiene en cuenta los dos valores más extremos de un conjunto de datos, el mínimo y el máximo.

B. Rango Intercuartílico

El rango intercuartílico (IQR) representa el medio 50% de todos los datos. Para calcular el rango intercuartílico, primero hay que encontrar los percentiles 25 y 75. El percentil 25, también llamado primer cuartil y denotado Q1, es el valor por debajo del cual cae el 25% de los de los datos, cuando los datos están ordenados de menor a mayor. Del mismo modo, el percentil 75, también llamado tercer cuartil y denominado Q3, es el valor por debajo del cual cae en el 75% de los datos. El rango intercuartílico se encuentra tomando la diferencia entre los percentiles 75 y 25. Este se suele comunicar junto con la mediana, ya que no se ve afectado por los valores extremos.

C. Varianza y Desviación Estándar

La medida de dispersión más común es la desviación estándar. La varianza de la muestra se define como la desviación estándar de la muestra al cuadrado. Ambas describen la cantidad de variabilidad en torno a la media. La desviación estándar puede considerarse como la distancia media de una observación individual con respecto a X. Normalmente,

“

Recuerda

Medidas de posición: en el aspecto clínico, los más usados son los percentiles, se utilizan en el diagnóstico clínico.

la media y la desviación estándar se utilizan para describir las características de toda la distribución de valores. Si aumenta el número de individuos la varianza disminuye.

Medidas de Posición

- Son aquellas medidas que nos permiten saber cuál es el lugar de un valor dentro de la curva. En primer lugar, la mediana es el punto medio de todo conjunto de datos. Nos permite determinar el 50% de los datos inferiores a la mediana y a los 50% restantes superiores a la misma.
- Los cuartiles (Q1, Q2 o mediana y Q3) son los que dividen a la curva en 4 partes (25% en 4 partes), y por último los percentiles, que dividen a la data hasta en 100 partes. En el aspecto clínico, los más usados son los percentiles, se utilizan en el diagnóstico clínico.

Medidas de Forma

- Son medidas que nos permiten estudiar la forma de la curva representada por los datos obtenidos del estudio. Para ello debemos tener en cuenta dos aspectos fundamentales, la Simetría y la Kurtosis. La primera, es la relación que tiene el centro con sus colas. Por ende, se dice que una curva es simétrica cuando en un mismo punto coinciden la media, la moda y la mediana (ver figura 1) (**ENAM 2008-B**).
- En la otra mano, la Kurtosis se define como el grado de apuntalamiento, en otras palabras, la distancia que hay entre la punta de la curva respecto de sus colas (ver figura y). Por ende, si en el gráfico observamos una curva bien parada hacia arriba se le conoce de Kurtosis positiva (curva picuda). Mientras que, si el apuntalamiento es más corto de la distancia de las colas, estamos frente a una Kurtosis negativa (curva plana).

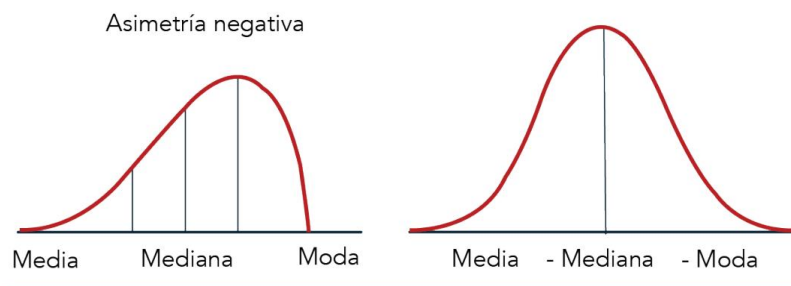


Figura 1. Curvas simétricas y asimétricas

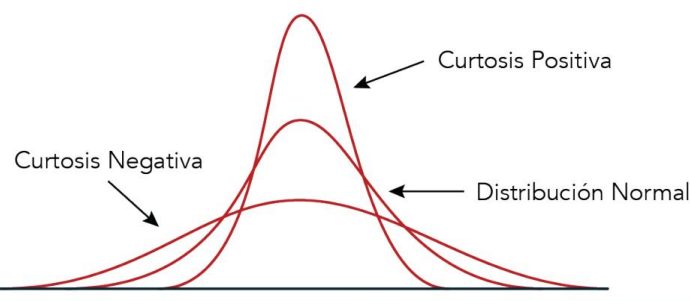
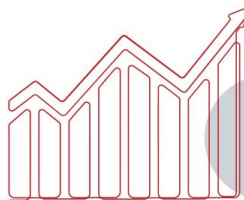


Figura 2. Kurtosis positiva, normal y negativa



02. ESTADÍSTICA INFERENCIAL

Introducción

La estadística descriptiva nos enseñó su utilidad en resumir y evaluar un conjunto de datos, lo que representa el primer paso en el análisis estadístico de un conjunto de datos. No obstante, cuando realizamos un experimento u observamos un fenómeno en una muestra, nuestros intereses van hacia poder generalizar nuestros resultados a la población de la que se ha extraído la muestra. Esto último es posible gracias a la inferencia estadística. En la presente unidad, se utilizará mucho este concepto para explicar la base de las diferentes pruebas estadísticas y para hablar sobre el muestreo. El antecedente más relevante para comenzar con el tema de la estadística inferencial es la teoría de las probabilidades.

Teoría de probabilidades

La teoría de las probabilidades cumple con los siguientes principios:

- Experimento aleatorio (e): el resultado es producto del azar, en otras palabras, no se sabe a ciencia cierta lo que va a suceder.
- Espacio muestral (omega): representa todos los resultados posibles. Estos se pueden colocar en una curva lo que forma una curva de probabilidades (distribución de probabilidades u Omega)
- Evento: es lo que ocurre. Por ejemplo, que nos salga cara en el lanzamiento de la moneda, que salga 2 en el lanzamiento de un dado, etc.
- Eventos mutuamente excluyentes: se define cuando la ocurrencia del evento A no tiene nada que ver con la ocurrencia del evento B. Por ejemplo, el hecho de sacar cara en el lanzamiento de la moneda no guarda ningún tipo de relación con sacar un número en un lanzamiento de dados, etc.

Con estos principios vamos a poder calcular lo ejemplificado en la siguiente tabla (ver tabla).

PRINCIPIOS	CÁLCULO
FÓRMULA PROBABILIDAD	Probabilidad de A = Evento de A / Omega
PROBABILIDAD CONJUNTA	Probabilidad A x Probabilidad B
PROBABILIDAD DE 2 EVENTOS CUALESQUIERA	Probabilidad A + Probabilidad B
PROBABILIDAD CONDICIONADA	Probabilidad A / Probabilidad B

Tabla 4. Principios de las Probabilidades y su cálculo

Distribución de probabilidades

Si queremos colocar todos los resultados posibles de un ensayo aleatorio en una tabla, gráfico o función matemática, entonces estamos creando una DISTRIBUCIÓN de probabilidades. Esta distribución variará de acuerdo con la variable en cuestión del experimento aleatorio.



Recuerda

Distribución Binomial

- Deriva del ensayo de Bernoulli
- 2 eventos mutuamente excluyentes



Recuerda

La función de densidad de probabilidad de una variable aleatoria continua se dice que es normal si cumple con las siguientes proporcionalidades → 1.96 Desviaciones estándar (DS) = 95% de la curva, 1 DS = 68.2% de la curva.



Recuerda

La curva normal de Gauss se estandariza o es estándar cuando a la media se le asigna un valor de 0 y a la DS un valor de 1.

A. Distribución de Probabilidades de Variables Discretas

Distribución Binomial

- Deriva del ensayo de Bernoulli (**ENAM 2005-A**)
- 2 eventos mutuamente excluyentes

Distribución Poisson o Politémica

- Más de 2 eventos excluyentes
- No relacionados

B. Distribución de Probabilidades de Variables Continuas - Curva Normal de Gauss

- La función de densidad de probabilidad de una variable aleatoria continua se dice que es normal si cumple con las siguientes proporcionalidades → 1.96 Desviaciones estándar (DS) = 95% de la curva, 1 DS = 68.2% de la curva.
- La Curva de Gauss es una curva simétrica que respeta un porcentaje de proporción constante que dice que, si partimos de la media, 1DS para cada lado, se está representando al 68.2% de la curva. Luego, 1.96 DS para cada lado, representa al 95% de la curva (**ENAM 2004-A, ENAM 2009-B**). Por ello, Gauss demostró que, si la variable en cuestión es un fenómeno biológico, probablemente la distribución de probabilidades sea normal. En otras palabras, si nos encontramos estudiando una variable biológica, esta debería ser normal, y si esta es así, nosotros podemos calcular sus probabilidades de ocurrencia según la teoría de Gauss, lo que nos permite conocer el 95% que está al centro (en el aspecto clínico, lo correcto) y al 5% que está afuera (en el aspecto clínico, lo malo). A este punto de intersección se le conoce como el punto de Significancia o también conocido como el error Alfa. Se le conoce así porque todo lo que esté fuera de ese punto lo hemos considerado como anormal. Sin embargo, debemos recordar que esta distinción es arbitraria y se hizo por consenso científico. Finalmente, la curva normal de Gauss se estandariza o es estándar cuando a la media se le asigna un valor de 0 y a la DS un valor de 1.

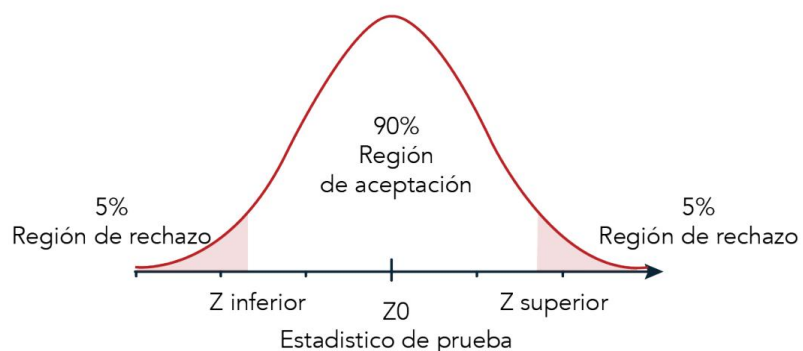


Figura 3. Curva de Gauss

Muestreo

El muestreo es importante en la estadística porque representa la herramienta principal que nos permite ahorrar tiempo y dinero utilizando la matemática, asumiendo que la variable es normal, para así poder extrapolar nuestros resultados en poblaciones grandes.

El Muestreo debe ser representativo (calidad) y significativo (cantidad) y partir de este estudio infiere (deduce) lo que sucede en la población. Por ejemplo, si queremos vacunar a todos los peruanos contra el COVID-19, partimos desde el ejemplo que los laboratorios han

“

Recuerda

El Muestreo debe ser representativo (calidad) y significativo (cantidad) y partir de este estudio infiere (deduce) lo que sucede en la población.

vacunado a miles de pacientes, y gracias a estos resultados podemos extrapolar y vacunar a millones de personas. Y todo esto es gracias a un buen muestreo con buena estadística probabilística de fondo.

Para recordar un poco más el concepto veamos las siguientes definiciones para tener en cuenta.

- **Población Objetivo:** Es aquella población a la cual queremos extrapolar los resultados de nuestro estudio. Por ejemplo, todos los millones de peruanos que deben ser vacunados contra el COVID-19. Estos necesitan Criterios de Inclusión y Criterios de Exclusión.
 - ◆ Criterios de Inclusión: son generales y van primero. Ejem. Adultos mayores de 18 años que vivan en el territorio peruano (**ENAM 2005-A**).
 - ◆ Criterios de Exclusión: son específicos y van al final. Ejem. Se excluyen adultos peruanos con prueba positiva actual de COVID-19.
- **Población Accesible - Marco Muestral:** Es la población donde se obtiene la muestra. Se diferencia de la población objetivo ya que ésta última es un concepto o una idea, mientras que la población accesible o marco muestral es algo concreto y representa la fuente de datos real.
 - ◆ Ejem. Si mi población objetivo son peruanos mayores de 18 años, mi población accesible serán los peruanos mayores de 18 años registrados en el Instituto Nacional de Estadística Informática (INEI).
- **Muestra:** es el conjunto de unidades muestrales tomados de la población accesible. Todo lo que se calcule en la muestra se llamará Estadístico Muestral (e) (**ENAM 2005-A**), se podrá calcular la media, desviación estándar, R de Pearson, entre otros. Sin embargo, cuando nosotros deseamos inferir a nuestra población, estos datos serán transformados en un Parámetro (P).

“

Recuerda

Criterios de Inclusión: son generales y van primero.
Muestra: todo lo que se calcule en la muestra se llamará Estadístico Muestral (e),

ORDEN	PASO
1	Determinar la población (criterios de inclusión y exclusión) y parámetros a medir
2	Escoger un marco muestral apropiado
3	Seleccionar tipo de muestreo (disminuir error alfa) y seleccionar método de muestreo
4	Definir tamaño de muestra (minimizar error beta)
5	Seleccionar muestra y tomar información (validar la muestra)

Tabla 5. Pasos para la realización de un correcto muestreo

“

Recuerda

El muestreo es representativo cuando comete poco error alfa.

Muestreo Representativo

El muestreo es representativo cuando comete poco error alfa. Por ende, si queremos un muestreo “representativo” tenemos que optar por un muestreo de tipo probabilístico.

- Muestreo No Probabilístico: no son representativos, ya que no es posible reducir el error alfa. Se utiliza en estudios cualitativos, donde la variable principal es de tipo subjetiva (una opinión, estudios de calidad) (Ver tabla 6)
- Muestreo Probabilístico: utiliza la probabilidad de Gauss, por ende, comete un error alfa controlado, hasta un 5% como máximo, según consenso científico. Se utiliza generalmente en estudios cuantitativos objetivos. (Ver tabla 7)

“

Recuerda

- El muestreo estratificado se utiliza para controlar una variable confusora al momento de tomar la muestra. Es el muestreo más representativo en el ámbito de la medicina
- El muestreo por conglomerados se utiliza cuando trabajamos con una población muy heterogénea.

MUESTREO NO PROBABILÍSTICO	USO/APLICACIÓN
POR CONVENIENCIA	El investigador elige a quién va y quién no va a su muestra. Cuando al investigador le interesa a quién muestrear usamos este tipo de muestreo
CASOS CONSECUTIVOS	Una muestra llega en función del orden de llegada de los sujetos.
POR CUOTAS	Se da cuando cerramos un número exacto de cuotas requerido para la muestra.
BOLA DE NIEVE	Donde cada sujeto muestreado nos recomienda a otro sujeto similar para la muestra y así consecutivamente (se da cuando las muestras son difíciles de conseguir, por ejemplo, trabajadoras sexuales, drogadicción, etc).
POR SATURACIÓN	Se usa en la investigación cualitativa y ocurre cuando ya se ha escuchado una cierta cantidad de ideas y con cada entrevista u observación adicional no aparecen nuevos elementos de análisis. (ENAM 2012-A)

Tabla 6. Resumen de los principales tipos de muestreo no probabilístico

MUESTREO PROBABILÍSTICO	USO/APLICACIÓN
ALEATORIO SIMPLE	Se basa en el azar para la selección de sujetos
SISTEMÁTICO	Trata de abarcar todo el marco muestral
ESTRATIFICADO	Se utiliza para controlar una variable confusora al momento de tomar la muestra. Es el muestreo más representativo en el ámbito de la medicina (ENAM 2004-A)
POR CONGLOMERADOS	Se utiliza cuando trabajamos con una población muy heterogénea (ENAM 2017-A)
MULTIETÁPICO	Cuando se combina 1 o más de los anteriores tipos de muestreo probabilístico

Tabla 7. Resumen de los principales tipos de muestreo probabilístico

Muestreo Significativo

El muestreo es significativo cuando tiene suficiente cantidad de sujetos para no cometer el error beta. Es importante tener el tamaño suficiente de muestra para no cometer este error, pero no lo suficientemente grande como para incurrir en costos adicionales innecesarios.

Fórmula para calcular el tamaño de muestra infinita (ENAM 2010-B)

$$n = \frac{Z_{\alpha}^2 \times p \times q}{e^2}$$

- n: Tamaño de muestra buscado
 N: Tamaño de la población o Universo
 z: Parámetro estadístico que depende el Nivel de Confianza (NC)
 e: Error de estimación máximo aceptado
 p: Probabilidad de que ocurra el evento estudiado (éxito)
 q: (1-p)= Probabilidad de que no ocurra el evento estudiado

Cálculo muestral para media

$$N = Z_{\alpha}^2 \frac{S^2}{\delta^2}$$

- N: Tamaño muestral
 Z_α: Nivel de confianza
 S²: varianza poblacional
 δ: precisión de la estimación

Ajuste en poblaciones finitas

$$N_c = \frac{N}{1 + \frac{N}{PT}}$$

- N_c: tamaño muestral corregido;
 N: tamaño muestral calculado;
 PT: tamaño de la población

Estadística inferencial

Es la parte de la estadística que utiliza la teoría de probabilidades y el muestreo para extrapolar los resultados o estadísticos muestrales en la población objetivo, determinando así la estimación del parámetro poblacional.

“

Recuerda

- Hipótesis de una Cola: se utiliza cuando el operador matemático en bioestadística es mayor $>$ o menor $<$
- Hipótesis de dos Colas: se utiliza cuando hablamos de hipótesis iguales o diferentes.

“

Recuerda

El error alfa o también conocido como nivel de significancia nos permite establecer un nivel de confianza (1-alfa o 95%).

“

Recuerda

Podemos crear un intervalo alrededor de la media de la muestra con un margen de error que es 2 veces el error estándar de la media (SEM), lo que se llama un intervalo de confianza del 95% para la verdadera media de la población.

Antes de adentrarnos en la estadística inferencial debemos definir los tipos de las hipótesis de investigación.

Tipos hipótesis de Investigación

- **Descriptivas:** se utiliza cuando solo queremos conocer y describir a profundidad una variable
- **Correlacionales:** se utiliza cuando queremos asociar a una variable con la otra (ejem. tabaco se asocia con cáncer) ($A \leftrightarrow B$)
- **De Diferencia:** se utiliza cuando queremos probar que el grupo A es diferente al grupo B. ($A \neq B$)
- **De Causalidad:** se utiliza cuando queremos demostrar que A genera o causa B ($A \rightarrow B$)

Tipos de hipótesis Estadísticas

Cuando hablamos de estadística tenemos otro tipo de hipótesis a las cuáles nos referimos desde el punto de vista matemático, estas son:

- **Hipótesis de una Cola:** se utiliza cuando el operador matemático en bioestadística es mayor $>$ o menor $<$
 - **Hipótesis de dos Colas:** se utiliza cuando hablamos de hipótesis iguales o diferentes.
- Una vez revisado esto, para simplificar el estudio clasificaremos a la estadística inferencial en intervalo de confianza y prueba de hipótesis.

Intervalo de Confianza (IC)

Es el conjunto de valores que estiman el parámetro poblacional indicando el error: alfa y beta. Se usa para estudios de una sola variable donde se busca extrapolar el estadístico con el parámetro poblacional sin compararlo con otro grupo de estudio o variable en cuestión.

- Condiciones que debe cumplir: que la población provenga de una variable normal (para poder usar la curva Z de Gauss) y que se conozca la varianza para poder hablar del error beta de la muestra.
- ¿Cómo se interpreta? Según un 95% de probabilidad podemos encontrar el parámetro poblacional en este intervalo.

La precisión del IC será la amplitud del intervalo. Cabe resaltar que todo lo anteriormente mencionado se da gracias al Teorema del Límite Central (TLC), el cuál es la explicación matemática del cálculo del IC. Dentro de este teorema se explica el error estándar de la media, el cual se calcula a través de la varianza dividida entre la muestra, lo que representa el valor del error beta. El error alfa o también conocido como nivel de significancia nos permite establecer un nivel de confianza (1-alfa o 95%).

Como se ha mencionado anteriormente, la media de una muestra es sólo una estimación de la media real, de la que se han extraído los datos. Se puede concebir que hay algún error en la estimación de la población con la estimación mediante la media de una sola muestra. Podemos crear un intervalo alrededor de la media de la muestra con un margen de error que es 2 veces el error estándar de la media (SEM), lo que se llama un intervalo de confianza del 95% para la verdadera media de la población. Decimos que "tenemos un 95% de confianza en que la verdadera media de la población cae en este intervalo". Lo que realmente significa es lo siguiente: imaginemos que se extraen muchas muestras del mismo tamaño de una población; entonces el 95% de estas muestras tendrán intervalos de confianza que capturan la verdadera media de la población.

Prueba de Hipótesis

Generalmente se utiliza cuando se desea comprobar si el parámetro de una población A es igual, diferente, mayor o menor al parámetro de una población B. Es decir, la prueba de hipótesis supone una comparación de grupos. Para formularla se requiere seguir una serie de pasos de forma ordenada para no cometer error sistemático:

ORDEN	PASO
1	Ver las variables y los parámetros. Verificar la normalidad de la variable en estudio.
2	Formular las hipótesis alterna y nula. H_0 y H_1
3	Fijar el nivel de significación (error alfa)
4	Seleccionar la prueba estadística dependiendo del tipo de variable y la normalidad.
5	Formular la regla de decisión, es decir elegir la zona de rechazo de la H_0
6	Calcular el estadístico de prueba, se somete a cálculo estadístico la H_0 solamente.
7	Formular la decisión estadística - No rechazar la hipótesis nula H_0 - Rechazar la hipótesis nula H_0 y aceptar hipótesis alterna H_1
8	Conclusión en términos del problema de investigación

Tabla 8. Pasos para hacer una correcta prueba de hipótesis

Condiciones:

- La conclusión deseada es H_1 (hipótesis alterna o lo que busca demostrar el investigador)
- Crear una hipótesis nula (H_0) para demostrar lo contrario
- La H_0 debe contener la igualdad, NO puede tener la desigualdad.
- La H_0 es la que debe ser comprobada matemáticamente.
- La H_1 y H_0 son complementarias, es decir suman 100% de todas las probabilidades posibles.
- La conclusión no es verdadera, es probablemente verdadera desde el punto de vista estadístico.

Ahora veamos los 4 escenarios posibles de una toma de decisión estadística, asumiendo que NO tenemos certeza de que la H_0 sea falsa o verdadera.



Recuerda

Condiciones:
La H_1 y H_0 son complementarias, es decir suman 100% de todas las probabilidades posibles.

DECISIÓN ESTADÍSTICA	REALIDAD DEL FENÓMENO	
	H_0 : 1 (cierta)	H_0 = 0 (falsa)
NO RECHAZAR H_0	Nivel de confianza	Error II o beta
RECHAZAR H_0	Error I o alfa	Potencia estadística

Tabla 9. Decisión estadística

“

Recuerda

- Error tipo 1 (falso positivo): Rechazar la hipótesis nula, aunque esta es verdadera (en otras palabras, afirmar que hay una diferencia significativa cuando en realidad no hay diferencia).
- Error tipo 2 (falso negativo): No rechazar la hipótesis nula cuando es falsa (en otras palabras, afirmar que no hay diferencias significativas cuando en realidad sí las hay. Además, se puede acortar aumentando el tamaño de la muestra

De aquí pueden verse 2 escenarios:

- **Escenario 1:** la H_0 es verdadera (supongamos que es posible saberlo). Se realiza el experimento y, efectivamente, se encuentra un valor p superior a 0,05, por lo que no se rechaza la H_0 . Por tanto, el resultado del experimento coincide con la verdad (de nuevo, si fuera posible conocer la verdad)
- **Escenario 2:** la H_0 es falsa (de nuevo, supongamos que es posible saberlo). Usted realiza el experimento y, efectivamente, encuentra un valor p inferior a 0,05, con lo que rechaza la H_0 . Por tanto, el resultado del experimento coincide con la verdad (de nuevo, si fuera posible conocer la verdad)

En estos dos escenarios el experimento coincide con la verdad. Pero ¿y si no lo hace?

La tabla anterior muestra que también pueden producirse dos tipos de errores al realizar una prueba estadística:

- **Error tipo 1 (falso positivo):** Rechazar la hipótesis nula, aunque esta es verdadera (en otras palabras, afirmar que hay una diferencia significativa cuando en realidad no hay diferencia). Este error NO se puede acortar aumentando el tamaño de la muestra, ya que es un error de diseño. (ENAM 2004-A, ENAM EXTRA 2020).
- **Error tipo 2 (falso negativo):** No rechazar la hipótesis nula cuando es falsa (en otras palabras, afirmar que no hay diferencias significativas cuando en realidad sí las hay; esto ocurre cuando el experimento tiene poca potencia). Directamente relacionado con la potencia (Potencia = $1-\beta$) (ENAM 2005-A). La mayoría de los estudios establecen un β de 0,2. Esto significa que su potencia será de 0,8 (80%), y que aceptaremos un 20% de posibilidades de no rechazar la hipótesis nula (H_0) cuando esta es realmente cierta. Este error tipo 2 se puede acortar aumentando el tamaño de la muestra.

Test Estadísticos

OBJETIVO	CONTINUOS	ORDINALES	BINARIOS	SUPERVIVENCIA
DESCRIPCIÓN 1 GRUPO	Media, DS	Media, DS	Proporción	Curva Kaplan-Meier
COMPARAR CON 1 VARIABLE	T Student – Z Gauss (ENAM 2015-B)	Wilcoxon	Chi2, prueba binomial	
COMPARAR 2 GRUPOS INDEPENDIENTE	T Student muestras independientes	Mann-Whitney (ENAM 2019-A)	Fisher, Chi 2 (ENAM 2003-B)	Long-rank test Mantel-Haenszel
COMPARAR +2 GRUPOS INDEPENDIENTE	ANOVA 1 vía	Kruskal-Wallis Jockheere-Terpstra	Chi2 (ENAM 2006-A)	Regresión Cox
COMPARAR 2 GRUPOS DEPENDIENTE	T Student muestras apareadas (ENAM 2009-A)	Wilcoxon (ENAM 2003-B)	McNemar	Ref. Riesgos proporcionales
COMPARAR +2 GRUPOS DEPENDIENTE	ANOVA 2 vías	Friedman	Q Cochran	¿?
ASOCIACIÓN 2 VARIABLE	Correlación Pearson	Correlación Spearman	Coeficiente contingencia	¿?
ASOCIACIÓN +2 VARIABLE	Correlación canónica	¿?	An. De fc multivía	¿?
PREDECIR 1 VARIABLE	Regresiones (lineal, no lineal, simple, múltiple) (ENAM 2014-A)	Reg. No paramétrica	Regresión logística	Regresión Cox

Tabla 10. Test estadísticos