Apuntes Probabilidad y Estadística

Felipe Colli

2024

Índice

1. 14/03

1.1. Materia

Población: Conjunto de todos los elementos que se quieren estudiar. Cuando la información deseada está disponible para todos los objetos de la población, lo llamamos **censo**. En la práctica es muy difícil o casi imposible realizar un censo.

Muestra: Subconjunto de la población que se mide u observa.

Parámetro: Es una medición numérica que describe algunas características de una población.

Estadístico (o estadígrafo): Es una medición numérica que describe algunas características de la muestra.

Variables cualitativas:

- Se describen mediante palabras o categorías.
- Se usan para categorizar a los individuos o para identificar.
- Sirven para comprender aspectos subjetivos y complejos.
- Se pueden clasificar en nominales y ordinales.
- Ejemplos: el color del cabello, el deporte favorito, la comida favorita, el lugar de nacimiento.

Variables cuantitativas:

- Se expresan mediante números, es decir, se pueden contar o medir.
- Permiten más operaciones matemáticas.
- Se pueden usar para conocer fenómenos o situaciones a través de la recolección y generación de números y datos.
- Ejemplos: la edad, los ingresos, el peso, la altura, la presión, la humedad o cantidad de hermanos.

1.2. Ejercicios:

Para cada una de las siguientes situaciones, identifica la población de interés, la variable estadística, clasifícala, y entrega un ejemplo de cuál podría ser una posible muestra.

1. Un investigador universitario desea estimar la proporción de ciudadanos chilenos de la *GEN X* que están interesados en iniciar sus propios negocios.

   a) **Población:** Chilenos de Gen X

   b) **Muestra (ejemplo):** Santiaguinos de Gen X (seleccionados aleatoriamente)

   c) **Variable:** Interés en iniciar un negocio (Sí/No) (*cualitativa nominal*)

2. Durante más de un siglo, la temperatura corporal normal en seres humanos ha sido aceptada como 37°C. ¿Es así realmente? Los investigadores desean estimar el promedio de temperatura de adultos sanos en Chile.

   a) **Población:** Adultos sanos en Chile

   b) **Muestra (ejemplo):** Adultos sanos de Santiago (seleccionados de diversos centros de salud)

   c) **Variable:** Temperatura corporal (*cuantitativa continua*)

3. Un ingeniero municipal desea estimar el promedio de consumo semanal de agua para unidades habitacionales unifamiliares en la ciudad.

   a) **Población:** Unidades habitacionales unifamiliares de la ciudad

   b) **Muestra (ejemplo):** Unidades habitacionales unifamiliares de un sector de la ciudad (seleccionadas aleatoriamente)

   c) **Variable:** Consumo semanal de agua (*cuantitativa continua*)

4. El National Highway Safety Council desea estimar la proporción de llantas para automóvil con dibujo o superficie de rodadura insegura, entre todas las llantas manufacturadas por una empresa específica durante el presente año de producción.

   a) **Población:** Todas las llantas para automóvil manufacturadas por la empresa específica durante el presente año de producción
   b) **Muestra (ejemplo):** Una selección aleatoria de llantas producidas en diferentes lotes o días del año de producción
   c) **Variable:** Estado de la superficie de rodadura (segura/insegura) (*cualitativa nominal*)

5. Un politólogo desea estimar si la mayoría de los residentes adultos de una región están a favor de una legislatura unicameral.
   a) **Población:** Residentes adultos de la región
   b) **Muestra (ejemplo):** Residentes adultos de una comuna (o varias comunas seleccionadas aleatoriamente) de la región
   c) **Variable:** Opinión sobre la legislatura unicameral (a favor/en contra/indeciso) (*cualitativa nominal*)

6. Un científico del área médica desea determinar el tiempo promedio para que se vuelva a presentar cierta enfermedad infecciosa, una vez que las personas se recuperan de ella por primera vez.
   a) **Población:** Personas que se han recuperado de la enfermedad infecciosa por primera vez
   b) **Muestra (ejemplo):** Pacientes recuperados seleccionados de registros médicos de diversos hospitales o regiones
   c) **Variable:** Tiempo hasta la recurrencia de la enfermedad (*cuantitativa continua*)

7. Un ingeniero electricista desea determinar si el promedio de vida útil de transistores de cierto tipo es mayor que 500 horas.
   a) **Población:** Todos los transistores de cierto tipo
   b) **Muestra (ejemplo):** Una muestra de 100 transistores de ese tipo, seleccionados aleatoriamente de la producción
   c) **Variable:** Vida útil del transistor (en horas) (*cuantitativa continua*). (Alternativamente, si se define como "vida útil $>$500 horas (Sí/No)", sería *cualitativa nominal*).

2. 17/03

2.1. Materia

Medidas de Tendencia Central
Medidas de Posición
Medidas de Dispersión

Tablas de Frecuencia: Conceptos Básicos

- **Dato o Intervalo:** Información (variable) que se estudia en estadística.
- **Marca de Clase (c_i):** Promedio entre los extremos de un intervalo.
- **Amplitud de un intervalo:** Es la diferencia entre el límite superior y el límite inferior del intervalo.

Tipos de Frecuencia:

- **Frecuencia Absoluta (f_i):** Cantidad de veces que se repite un dato o que los datos caen en un intervalo.
- **Frecuencia Absoluta Acumulada (F_i):** Suma de las frecuencias absolutas hasta determinado dato o intervalo. $F_i = \sum_{j=1}^{i} f_j$.
- **Frecuencia Relativa (h_i o f_{ri}):** Es la proporción (fracción, decimal o porcentaje) de observaciones que corresponden a cierto valor o intervalo. ($h_i = \frac{f_i}{n}$), donde n es el número total de datos.
- **Frecuencia Relativa Acumulada (H_i):** Es la proporción (fracción, decimal o porcentaje) de la frecuencia acumulada hasta cierto dato o intervalo. ($H_i = \frac{F_i}{n} = \sum_{j=1}^{i} h_j$).

Medidas de Tendencia Central:

- **Media Aritmética ($\bar{x}$):** Es el cociente entre la suma de todos los datos y el número total de datos (n). Si se tienen n datos $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Para datos agrupados en una tabla de frecuencia con k clases:

Marca de clase (c_i)	**Frecuencia (f_i)**
c_1	f_1
c_2	f_2
$\vdots$	$\vdots$
c_k	f_k

$$\bar{x} = \frac{c_1 f_1 + c_2 f_2 + \cdots + c_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} c_i f_i}{n}$$

(donde $n = \sum_{i=1}^{k} f_i$)

- **Mediana (M_e):** Es el valor que ocupa la posición central de la muestra cuando los datos se encuentran ordenados en forma creciente o decreciente. **Si la muestra tiene un número par de datos, la mediana es la media aritmética de los dos términos centrales.** Se aplica principalmente a variables cuantitativas y ordinales.
  - Si n es impar, la posición del dato mediano es $\frac{n+1}{2}$. $M_e = x_{(\frac{n+1}{2})}$.
  - Si n es par, las posiciones de los datos centrales son $\frac{n}{2}$ y $\frac{n}{2}+1$. $M_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$.
- **Moda (M_o):** Es el dato, valor o intervalo que presenta la mayor frecuencia absoluta. La muestra puede ser:
  - **Amodal:** No presenta moda (todos los datos tienen la misma frecuencia, o no hay una frecuencia que destaque).
  - **Unimodal:** Un solo dato (o intervalo) tiene la frecuencia máxima.
  - **Bimodal:** Dos datos (o intervalos) no adyacentes tienen la misma frecuencia máxima.
  - **Polimodal (o Multimodal):** Más de dos datos (o intervalos) no adyacentes tienen la misma frecuencia máxima.
  - **Intervalo Modal:** (Para datos agrupados) El intervalo que presenta la mayor frecuencia absoluta.

3. 26/03

3.1. Ejercicios:

Si las notas de Esteban en una asignatura son $3, 4, 6, 3, 5, 5, 6, 3, 4$ y de estas notas se cambian un 6 por un 7. ¿Cuál(es) de las siguientes medidas de tendencia central cambia(n)?

1. La moda
2. La mediana
3. La media aritmética

Solución: Notas originales (ordenadas): $3, 3, 3, 4, 4, 5, 5, 6, 6$. $n = 9$.

- Moda original: 3 (frecuencia 3).
- Mediana original: El dato en la posición $\frac{9+1}{2} = 5$. Mediana $= 4$.
- Media original: $\frac{3 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 + 2 \cdot 6}{9} = \frac{9+8+10+12}{9} = \frac{39}{9} \approx 4{,}33$.

Notas nuevas (se cambia un 6 por un 7): $3, 3, 3, 4, 4, 5, 5, 6, 7$. $n = 9$.

- Moda nueva: 3 (sigue siendo la frecuencia más alta, con 3 ocurrencias). $\rightarrow$ No cambia.
- Mediana nueva: El dato en la posición 5. Mediana $= 4$. $\rightarrow$ No cambia.
- Media nueva: $\frac{3 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 + 1 \cdot 6 + 1 \cdot 7}{9} = \frac{9+8+10+6+7}{9} = \frac{40}{9} \approx 4{,}44$. $\rightarrow$ Cambia.

Respuesta: Solo la media aritmética cambia (opción 3).

La siguiente tabla muestra los valores de una variable X y sus respectivas frecuencias. ¿Cuál es el valor de la mediana?

X_i	**Frecuencia** (f_i)	**Frecuencia Acumulada** (F_i)
4	4	4
5	8	12
6	10	22
7	20	42
8	8	50
Total	**n=50**	

Solución: Total de datos $n = 50$ (par). La mediana es el promedio de los valores de los datos en las posiciones $\frac{n}{2} = \frac{50}{2} = 25$ y $\frac{n}{2} + 1 = 26$. Buscamos en la columna de Frecuencia Acumulada (F_i) dónde caen estas posiciones:

- Hasta $X = 6$, se han acumulado 22 datos.
- Para $X = 7$, la frecuencia acumulada llega a 42. Esto significa que los datos desde la posición 23 hasta la 42 (inclusive) corresponden al valor $X = 7$.

Por lo tanto, tanto el dato en la posición 25 ($x_{(25)}$) como el dato en la posición 26 ($x_{(26)}$) son 7. La mediana es $M_e = \frac{x_{(25)}+x_{(26)}}{2} = \frac{7+7}{2} = 7$.
Respuesta: La mediana es 7.

De acuerdo a la siguiente muestra $a+2, a+4, a+6, a+6, a+6, a+4, a+2$, la suma de la mediana y la moda es:
Solución: Muestra original: $a+2, a+4, a+6, a+6, a+6, a+4, a+2$. Muestra ordenada: $a+2, a+2, a+4, a+4, a+6, a+6, a+6$. $n = 7$.

- **Moda (M_o):** El dato más frecuente es $a + 6$ (aparece 3 veces). $M_o = a + 6$.
- **Mediana (M_e):** Como $n = 7$ (impar), la mediana es el dato en la posición $\frac{7+1}{2} = 4$. El cuarto dato en la muestra ordenada es $a + 4$. $M_e = a + 4$.

Suma $= M_o + M_e = (a + 6) + (a + 4) = 2a + 10$
Respuesta: $2a + 10$.

Los datos de una muestra son todos números naturales consecutivos, si no hay ningún dato repetido y la mediana de la muestra es 11.5, entonces ¿Qué cantidad de datos no puede ser? *Solución:* La mediana es 11.5. Dado que los datos son números naturales (enteros y positivos), una mediana que no es un número natural (sino un decimal .5) implica que el número de datos (n) debe ser par. Si n es par, la mediana es el promedio de los dos datos centrales. Sean estos dos datos centrales x_k y x_{k+1}, donde $k = n/2$. Como los datos son naturales consecutivos y no repetidos, $x_{k+1} = x_k + 1$. La mediana es $M_e = \frac{x_k + x_{k+1}}{2} = \frac{x_k + (x_k+1)}{2} = \frac{2x_k+1}{2} = x_k + 0{,}5$. Se nos dice que la mediana es 11.5, entonces $x_k + 0{,}5 = 11{,}5$, lo que implica $x_k = 11$. El siguiente dato consecutivo es $x_{k+1} = 11 + 1 = 12$. Los dos datos centrales son 11 y 12. Para que la mediana sea el promedio de dos datos centrales, la cantidad de datos n debe ser par. Si n fuera impar, la mediana sería uno de los datos de la muestra (un número natural), lo cual contradice que la mediana es 11.5. Por lo tanto, la cantidad de datos n no puede ser un número impar.
Respuesta: La cantidad de datos no puede ser un número impar.

4. 04/04

4.1. Población y Muestra

¿Qué inconvenientes puede implicar realizar un censo?

- **Cardinalidad (tamaño) de la población:** Puede ser demasiado grande para estudiar todos sus elementos (incluso infinita).
- **Destrucción de los objetos de estudio:** En algunos casos, el proceso de medición destruye el elemento (ej. pruebas de vida útil de bombillas, control de calidad destructivo de alimentos).
- **Costos asociados:** Implica altos costos en términos de tiempo, dinero y recursos humanos.
- **Dificultad de acceso:** Puede ser logísticamente imposible acceder a todos los miembros de la población.
- **Tiempo requerido:** Un censo puede tomar tanto tiempo que la información obtenida ya no sea relevante cuando esté disponible.

4.2. Muestreo

Proceso de diseñar e implementar mecanismos para escoger los elementos que conformarán la muestra.
Es fundamental que la muestra esté bien escogida (sea representativa) para realizar una inferencia estadística válida sobre la población.

4.3. Muestra

4.3.1. Representatividad

Para que una muestra sea representativa, debe reflejar las características relevantes de la población en la misma proporción en que se encuentran en ella. Claves para lograrlo:

- El **tamaño** de la muestra (n): Debe ser suficientemente grande. Se abordará más adelante cómo determinarlo (criterios probabilísticos).

- **Aleatoriedad:** El mecanismo de selección debe asegurar que todos los elementos (o grupos de elementos) de la población tengan una probabilidad conocida (y a menudo igual, aunque no siempre) de ser seleccionados para la muestra. Esto ayuda a minimizar el sesgo de selección.

Por lo general designaremos con la letra **N** la cardinalidad (tamaño) de la población y con **n** la cardinalidad de la muestra.

4.4. Tipos de Muestreo Probabilístico

En el muestreo probabilístico, cada unidad de la población tiene una probabilidad conocida y no nula de ser seleccionada.

4.4.1. Muestreo Aleatorio Simple (M.A.S)

Una **M.A.S** de tamaño **n** se selecciona de tal modo que cada posible muestra del mismo tamaño n tiene la misma probabilidad de ser elegida. Requiere un listado completo y actualizado de todas las unidades de la población (marco muestral).

Ejemplos:

- Seleccionar 200 pacientes al azar de una lista completa de registros médicos de un hospital.
- Usar un generador de números aleatorios para seleccionar 500 estudiantes de una lista nacional de todos los estudiantes de educación media, sin agruparlos por colegio.
- Elegir 100 tornillos de una gran producción diaria para control de calidad, asumiendo que la producción es homogénea y se puede numerar cada tornillo o seleccionar al azar en momentos aleatorios.

4.4.2. Muestreo Estratificado

Se utiliza cuando la población no es homogénea con respecto a la variable de estudio, pero puede dividirse en subgrupos o estratos que son internamente más homogéneos.

1. Se divide la población (N) en L estratos ($N_1, N_2, \ldots, N_L$) mutuamente excluyentes y colectivamente exhaustivos ($N = \sum N_h$).

2. Se selecciona una muestra aleatoria simple (u otro método probabilístico) dentro de cada estrato, de tamaño n_h. La muestra total es $n = \sum n_h$.

3. Es más eficiente (produce estimaciones más precisas para un tamaño de muestra dado) si la variabilidad dentro de los estratos es baja (homogeneidad intra-estrato) y la variabilidad entre estratos es alta (heterogeneidad inter-estrato).

Muestreo Estratificado Proporcional (o Afijación Proporcional)

- El número de elementos extraído de cada estrato (n_h) es proporcional al tamaño relativo del estrato en la población ($W_h = N_h/N$).
- $n_h = n \cdot \frac{N_h}{N} = n \cdot W_h$.
- Se utiliza cuando el propósito principal es obtener una buena representatividad global de la población y estimar parámetros poblacionales generales. Cada elemento de la población tiene la misma probabilidad de ser seleccionado.

Muestreo Estratificado No Proporcional (ej. Afijación Óptima o de Neyman)

- El tamaño de la muestra en cada estrato (n_h) no es directamente proporcional a N_h/N.
- En la **afijación óptima**, n_h se elige para minimizar la varianza del estimador para un costo fijo, o minimizar el costo para una varianza fija. Generalmente, se asigna un tamaño muestral mayor a estratos más grandes, con mayor variabilidad interna (σ_h), y/o menor costo de muestreo por unidad.
- Fórmula (Neyman, sin costos): $n_h = n \cdot \frac{N_h \sigma_h}{\sum_{j=1}^{L} N_j \sigma_j}$.
- Los elementos de la población no necesariamente tienen la misma probabilidad global de ser incluidos en la muestra, a menos que se usen ponderaciones en el análisis.

4.4.3. Muestreo por Conglomerados (Clusters)

- Se utiliza cuando la población está dividida naturalmente en grupos (conglomerados), como ciudades, escuelas, manzanas de viviendas, etc. Es útil cuando es difícil o costoso obtener un marco muestral de unidades individuales.
- **Proceso (una etapa):**
  1. Se selecciona una muestra aleatoria de conglomerados.
  2. Se incluyen en la muestra **todos** los individuos dentro de los conglomerados seleccionados.
- **Muestreo polietápico (o multietápico):** Se realizan varias etapas de muestreo. Ej: seleccionar conglomerados, luego submuestrear unidades dentro de esos conglomerados.
- **Idealmente, cada conglomerado debe ser internamente heterogéneo (una mini-representación de la población)** y los conglomerados deben ser similares entre sí. Es más eficiente (en términos de costo, no necesariamente de precisión para un n dado) si la variabilidad *dentro* de los conglomerados es alta y *entre* conglomerados es baja (opuesto al estratificado en términos de varianza).

4.4.4. Muestreo Aleatorio Sistemático

1. Se utiliza cuando se dispone de una lista ordenada de los N elementos de la población.
2. Se calcula un intervalo de muestreo $k = N/n$ (aproximado a un entero si no lo es).
3. Se elige un punto de partida aleatorio (a) entre 1 y k.
4. Se seleccionan los elementos $a, a+k, a+2k, \ldots, a+(n-1)k$.
5. Si la lista está ordenada según alguna característica relacionada con la variable de estudio, puede ser más preciso que un M.A.S. Sin embargo, puede ser problemático si hay alguna periodicidad en la lista que coincida con el intervalo k.

4.5. Muestreos No Aleatorios (No Probabilísticos)

La selección de la muestra se basa en criterios subjetivos, conveniencia o juicio, y no se conoce la probabilidad de selección de cada unidad. No permiten realizar inferencias estadísticas formales sobre la población.

4.5.1. Muestreo por Cuotas

- Técnica común en estudios de mercado y sondeos de opinión.
- La población se divide en grupos según características demográficas (sexo, edad, región, etc.).
- Se fija una cuota (número de individuos a entrevistar) para cada grupo, a menudo proporcional a su tamaño en la población.
- La selección de los individuos dentro de cada grupo queda a criterio del entrevistador (no es aleatoria), quien busca personas que cumplan con las características hasta llenar la cuota.

4.5.2. Muestreo Bola de Nieve (Snowball Sampling)

1. Indicado para estudiar poblaciones difíciles de localizar o contactar (minoritarias, ocultas, estigmatizadas, o muy dispersas pero conectadas en red).
2. Se contacta a unos pocos individuos iniciales que cumplen los criterios del estudio.
3. Estos individuos iniciales ayudan a localizar y contactar a otros miembros de la población, y así sucesivamente, como una bola de nieve que crece.

4.5.3. Muestreo por Juicio (o Intencional o de Conveniencia)

- **Por Juicio/Intencional:** La selección de la muestra se basa en el juicio o criterio del investigador, quien elige a los individuos que considera más representativos, típicos o informativos para los propósitos del estudio, basándose en su experiencia o conocimiento previo de la población.

- **De Conveniencia:** Se seleccionan los individuos que son más fáciles de acceder o que están disponibles en un momento dado (ej. entrevistar a estudiantes en un campus, usar pacientes de una clínica específica).

4.6. Preguntas

1. ¿Cuándo ocupar un muestreo estratificado en vez de uno por conglomerados? *Respuesta:* Usar **muestreo estratificado** cuando:

   - La población es heterogénea globalmente respecto a la variable de interés.
   - Se pueden identificar subgrupos (estratos) que son internamente homogéneos (baja varianza intra-estrato).
   - Hay alta varianza entre los estratos (los estratos son diferentes entre sí).
   - El objetivo principal es aumentar la precisión de las estimaciones y asegurar la representación de todos los subgrupos importantes.
   - Se dispone de un marco muestral para cada estrato.

   Usar **muestreo por conglomerados** cuando:

   - La población está naturalmente agrupada en conglomerados (ej. geográficamente).
   - Es costoso o difícil obtener un marco muestral de unidades individuales para toda la población, pero es más fácil obtener un marco de conglomerados.
   - Idealmente, los conglomerados son internamente heterogéneos (representan la variabilidad de la población, alta varianza intra-conglomerado).
   - Hay baja varianza entre conglomerados (los conglomerados son similares entre sí).
   - El objetivo principal es la eficiencia operativa y la reducción de costos, aunque puede ser menos preciso que el M.A.S. o estratificado para el mismo número de unidades finales.

2. ¿En qué se diferencia un muestreo por cuotas de un muestreo estratificado? *Respuesta:* Ambos métodos dividen la población en grupos o

estratos. La diferencia fundamental radica en el método de selección de los elementos *dentro* de esos grupos:

- **Muestreo Estratificado:** Es un método *probabilístico*. Una vez definidos los estratos, se selecciona una muestra aleatoria (generalmente M.A.S.) *dentro de cada estrato*. Todos los elementos de un estrato tienen una probabilidad conocida de ser seleccionados. Permite realizar inferencias estadísticas formales sobre la población.
- **Muestreo por Cuotas:** Es un método *no probabilístico*. Aunque se definen cuotas para los grupos (similares a los estratos), la selección de los individuos para cumplir esas cuotas queda a *criterio del entrevistador o por conveniencia*. No hay aleatoriedad en la selección final de los participantes dentro de cada cuota. No permite generalizar los resultados a la población con un nivel de confianza medible.

5. 10/04

Objetivo: Aplicar y comprender propiedades de las medidas de dispersión

5.1. Medidas de Dispersión

Las medidas de tendencia central (como la media) no son suficientes por sí solas para describir un conjunto de datos, ya que no indican cuán dispersos o concentrados están los datos alrededor de ese centro. Consideremos dos conjuntos con la misma media $\bar{x} = 0$:

$$A = \{-4, 4, -4, 4\} \quad (\text{Media } \bar{x}_A = 0)$$

$$B = \{7, 1, -6, -2\} \quad (\text{Media } \bar{x}_B = 0)$$

Ambos tienen $\bar{x} = 0$, pero los datos en el conjunto A están menos dispersos (más concentrados alrededor de la media) que en el conjunto B. Las medidas de dispersión cuantifican esta variabilidad o .^{es}parcimiento"de los datos.

5.1.1. Rango (o Amplitud Total):

Se define como la diferencia entre el valor máximo y el valor mínimo de los datos.

$$Rango = x_{max} - x_{min}$$

Es una medida simple pero muy sensible a valores extremos y no considera la distribución de los datos intermedios.

5.1.2. Desviación Media (DM):

Dada una variable X, con n datos $x_1, x_2, \ldots, x_n$ y media aritmética $\bar{x}$. Se define la desviación media como el promedio de las desviaciones absolutas de cada dato respecto a la media:

$$DM = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Mide el promedio de cuánto se desvían los datos de la media, en valor absoluto.

5.1.3. Varianza (σ^2 para población, s^2 para muestra):

Es el promedio de las desviaciones al cuadrado de cada dato respecto a la media. Es la medida de dispersión más utilizada junto con su raíz cuadrada (la desviación estándar). Para una **población** de N datos:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Donde μ es la media poblacional. Si los datos $x_1, \ldots, x_n$ constituyen toda la población (y $\bar{x}$ es su media):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Para una **muestra** de n datos, la varianza muestral *insesgada* (estimador de σ^2) es:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

(En este curso, si no se especifica, σ^2 con denominador n se refiere a la varianza de un conjunto de datos específico, sea este una población o una muestra descripta como tal).

5.1.4. Desviación Estándar (o Típica) (σ para población, s para muestra):

Es la raíz cuadrada positiva de la varianza. Tiene la ventaja de estar expresada en las mismas unidades que los datos originales.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

5.1.5. Propiedades de σ y σ^2 (usando la definición con denominador n)

1. $\sigma \geq 0$ y $\sigma^2 \geq 0$. Son siempre no negativas.

2. $\sigma = 0 \iff \sigma^2 = 0 \iff x_i = \bar{x}$ para todo $i \iff x_i = x_j$ para todo $i, j \in \{1, \ldots, n\}$. La desviación estándar (y varianza) es cero si y sólo si todos los datos son iguales.

3. Si a todos los datos de un conjunto se les suma (o resta) una constante k (transformación $y_i = x_i + k$), la nueva media es $\bar{y} = \bar{x} + k$, pero la varianza y la desviación estándar no cambian: $\sigma_y^2 = \sigma_x^2$ y $\sigma_y = \sigma_x$.

4. Si todos los datos de un conjunto se multiplican (o dividen) por una constante k (transformación $y_i = k \cdot x_i$), la nueva media es $\bar{y} = k\bar{x}$, la nueva varianza es $\sigma_y^2 = k^2\sigma_x^2$, y la nueva desviación estándar es $\sigma_y = |k|\sigma_x$.

5. Fórmula computacional (o abreviada) para la varianza: $\sigma^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2$. Es decir, la varianza es la media de los cuadrados de los datos menos el cuadrado de la media de los datos.

6. $\sigma^2 = \sigma \iff \sigma = 0 \vee \sigma = 1$. (Asumiendo que σ es el valor numérico de la desviación estándar).

7. $\sigma^2 < \sigma \iff 0 < \sigma < 1$.

8. $\sigma^2 > \sigma \iff \sigma > 1$.

6. 16/04

6.1. Demostración Propiedad 4 (Multiplicación por una constante)

Sea la variable X con datos $x_1, \ldots, x_n$, media $\bar{x}$ y varianza σ_x^2. Sea Y una nueva variable tal que $y_i = k \cdot x_i$ para cada i. Sabemos que la media de Y es $\bar{y} = k \cdot \bar{x}$. La varianza de Y, σ_y^2, se define como:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Sustituyendo $y_i = kx_i$ y $\bar{y} = k\bar{x}$:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(kx_i - k\bar{x})^2}{n}$$

Factorizando k dentro del paréntesis al cuadrado:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}[k(x_i - \bar{x})]^2}{n}$$

Aplicando la potencia al producto:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n} k^2(x_i - \bar{x})^2}{n}$$

Como k^2 es una constante para la sumatoria, puede salir fuera:

$$\sigma_y^2 = k^2 \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Reconociendo que $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$ es la definición de σ_x^2:

$$\sigma_y^2 = k^2 \cdot \sigma_x^2$$

Tomando la raíz cuadrada positiva para obtener la desviación estándar (ya que $\sigma_x \geq 0$):

$$\sigma_y = \sqrt{k^2 \cdot \sigma_x^2} = \sqrt{k^2} \cdot \sqrt{\sigma_x^2} = |k| \cdot \sigma_x$$

L.Q.Q.D. (Lo Que Queríamos Demostrar)

6.2. Ejercicio

Dados los datos: -2, 0, 2, 4, 6. ($n = 5$). Determinar:

1. $\bar{x}$
   Solución: $\bar{x} = \frac{-2+0+2+4+6}{5} = \frac{10}{5} = 2$.

2. σ (desviación estándar)
   Solución: Primero calculamos la varianza σ^2:

$$\begin{aligned}\sigma^2 &= \frac{\sum(x_i - \bar{x})^2}{n} \\ &= \frac{(-2-2)^2 + (0-2)^2 + (2-2)^2 + (4-2)^2 + (6-2)^2}{5} \\ &= \frac{(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2}{5} \\ &= \frac{16+4+0+4+16}{5} = \frac{40}{5} = 8\end{aligned}$$

   Ahora la desviación estándar: $\sigma = \sqrt{\sigma^2} = \sqrt{8} = \sqrt{4 \cdot 2} = 2\sqrt{2} \approx 2{,}828$.

3. $\overline{x^2}$ (el promedio de los cuadrados de los datos)
   Solución: Los cuadrados de los datos son: $(-2)^2 = 4, 0^2 = 0, 2^2 = 4, 4^2 = 16, 6^2 = 36$.

$$\overline{x^2} = \frac{4+0+4+16+36}{5} = \frac{60}{5} = 12$$

4. Calcular $\overline{x^2} - (\bar{x})^2$ y comparar con σ^2.
   Solución: $\overline{x^2} - (\bar{x})^2 = 12 - (2)^2 = 12 - 4 = 8$.
   Este resultado (8) es igual a la varianza σ^2 calculada en el punto 2, lo cual verifica la propiedad 5 (fórmula computacional de la varianza).

6.3. Demostración Propiedad 5 (Fórmula Computacional de la Varianza)

Partimos de la definición de varianza (usando denominador n):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Expandimos el binomio al cuadrado $(a-b)^2 = a^2 - 2ab + b^2$:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + (\bar{x})^2)}{n}$$

Distribuimos la sumatoria y el denominador n:

$$\sigma^2 = \frac{\sum x_i^2}{n} - \frac{\sum 2x_i\bar{x}}{n} + \frac{\sum(\bar{x})^2}{n}$$

En el segundo término, $2\bar{x}$ es una constante respecto a la suma $\sum x_i$. En el tercer término, $(\bar{x})^2$ es una constante, y $\sum_{i=1}^{n}(\bar{x})^2 = n(\bar{x})^2$.

$$\sigma^2 = \frac{\sum x_i^2}{n} - 2\bar{x}\frac{\sum x_i}{n} + \frac{n(\bar{x})^2}{n}$$

Reconocemos que $\frac{\sum x_i^2}{n} = \overline{x^2}$ (la media de los cuadrados) y $\frac{\sum x_i}{n} = \bar{x}$ (la media):

$$\sigma^2 = \overline{x^2} - 2\bar{x}(\bar{x}) + (\bar{x})^2$$
$$\sigma^2 = \overline{x^2} - 2(\bar{x})^2 + (\bar{x})^2$$
$$\sigma^2 = \overline{x^2} - (\bar{x})^2$$

L.Q.Q.D.

7. 23/04

7.1. Demostraciones de Propiedades Relacionadas con el Valor de σ

Recordar que $\sigma \geq 0$ por definición (es una raíz cuadrada positiva o cero).

7.1.1. Propiedad 6: $\sigma^2 = \sigma \iff \sigma = 0 \vee \sigma = 1$

Partimos de la ecuación:

$$\sigma^2 = \sigma$$

Reordenamos para formar una ecuación cuadrática en σ:

$$\sigma^2 - \sigma = 0$$

Factorizamos σ:

$$\sigma(\sigma - 1) = 0$$

Esto implica que uno de los factores debe ser cero:

$$\sigma = 0 \quad \text{o} \quad \sigma - 1 = 0$$

Por lo tanto:

$$\sigma = 0 \vee \sigma = 1$$

L.Q.Q.D.

7.1.2. Propiedad 7: $\sigma^2 < \sigma \iff 0 < \sigma < 1$

Partimos de la desigualdad:

$$\sigma^2 < \sigma$$

Reordenamos:

$$\sigma^2 - \sigma < 0$$

Factorizamos:

$$\sigma(\sigma - 1) < 0$$

Para que el producto de dos factores sea negativo, uno debe ser positivo y el otro negativo. Analizamos los signos de σ y $(\sigma - 1)$:

Intervalo	$(-\infty, 0)$	0	$(0, 1)$	1	$(1, +\infty)$
Signo de σ	$-$	0	$+$	$+$	$+$
Signo de $(\sigma - 1)$	$-$	$-$	$-$	0	$+$
Signo de $\sigma(\sigma - 1)$	$+$	0	$-$	0	$+$

La desigualdad $\sigma(\sigma - 1) < 0$ se cumple cuando $\sigma \in (0, 1)$. Dado que $\sigma \geq 0$ por definición, el intervalo $(-\infty, 0)$ no es relevante para la desviación estándar. Por lo tanto:

$$0 < \sigma < 1$$

L.Q.Q.D.

7.1.3. Propiedad 8: $\sigma^2 > \sigma \iff \sigma > 1$

Partimos de la desigualdad:

$$\sigma^2 > \sigma$$

Reordenamos:

$$\sigma^2 - \sigma > 0$$

Factorizamos:

$$\sigma(\sigma - 1) > 0$$

Para que el producto de dos factores sea positivo, ambos deben ser positivos o ambos deben ser negativos. Usando la tabla de signos anterior:

- Ambos negativos: $\sigma < 0$ y $\sigma - 1 < 0$ (es decir, $\sigma < 0$). No es posible para σ.
- Ambos positivos: $\sigma > 0$ y $\sigma - 1 > 0$ (es decir, $\sigma > 1$).

La desigualdad $\sigma(\sigma - 1) > 0$ se cumple cuando $\sigma \in (-\infty, 0) \cup (1, +\infty)$. Considerando la restricción $\sigma \geq 0$:

- Si $\sigma = 0$, entonces $\sigma(\sigma - 1) = 0$, lo cual no satisface $0 > 0$.
- El intervalo $(-\infty, 0)$ no es válido para σ.
- Nos queda el intervalo $(1, +\infty)$.

Por lo tanto:

$$\sigma > 1$$

L.Q.Q.D.

8.