Capítulo 1: Introducción y Estadística Descriptiva

Fernando A. Crespo R.

3 de Agosto de 2023

Índice

1.1.1 Definición y términos básicos de la estadística

- La estadística es la ciencia de los datos.
- La estadística es aplicada comúnmente a dos tipos de problemas:
  - Resumir, describir y explorar datos.
  - Ejemplo: Resultados del censo.
  - Usando muestras de datos para inferir la naturaleza del conjunto de datos desde los cuales la muestra fue seleccionada.
  - Ejemplo: Cuando se estudia la sobrevida de las personas para calcular el valor de la prima de un seguro de vida.

Definición (Ramas de Estudio de la Estadística)

Las áreas que estudian las diferentes problemáticas de la estadística se reconocen como:

- *La rama que se dedica a resumir, describir y explorar datos se denomina estadística descriptiva.*
- *La rama que se decida a usar muestras de datos para inferir la naturaleza del conjunto de datos desde los cuales la muestra fue seleccionada, se denomina estadística inferencial.*

1.1.1 Definición y términos básicos de la estadística

- Todo ello con el fin de comprender la variablidad.
- Se desarrolla el pensamiento estadístico con el fin de poder enfrentar la variabilidad.
- Esa variabilidad, puede provenir de:
  - desde los distintos factores que influyen en un fenómeno: por ejemplo, pensemos en la medición del rendimiento de [km/l] del automóvil.
  - Puede ser implícita, porque las variables no se pueden medir de manera precisa, o el fenómeno tiene variabilidad. Por ejemplo: Genes, medidas atómicas.
- La estadística en conjunto con el método científico permite crear modelos coherentes capaces de soportar la variabilidad de los fenómenos.

1.2 Elementos Fundamentales de Estadística

- Una población estadística es un conjunto de datos (usualmente grande, otras veces conceptual) que es el objetivo de interés.
- Una muestra (sample) es un subconjunto de datos seleccionados de la población objetivo.
- El objeto (ie persona, cosa, transacción, espécimen, evento, u otra construcción) sobre el cual se observan las medidas se denomina unidad experimental. Una población puede considerarse como datos recolectados sobre muchas unidades experimentales.
- Una variable es una característica o propiedad de una unidad experimental individual.
- Ejemplo: Un estudio que desea observar las esquinas que tienen más accidentes de la comuna de Santiago.
- Una inferencia es una afirmación sustentada a partir de los datos.
- En un problema de inferencia estadística se puede indentificar cuatro puntos: una población, una o más variables, una muestra, y una inferencia. Hay que añadir la confiabilidad de la inferencia, es decir, una medida que nos diga cuan verdadera es la inferencia.
- Una medida de confiabilidad es una declaración (cuantificada) acerca del grado de incerteza asociada a una inferencia estadística.

1.2 Elementos Fundamentales de Estadística

- Cuatro elementos de Problemas de Estadística Descriptiva:
  - La población o muestra de interés.
  - Una o más variables que son investigadas.
  - Tablas, gráficos, o herramientas de resumen númerico.
- Cuatro elementos de Problemas de Inferencia Estadística:
  - La población de interés.
  - Una o más variables que son investigadas.
  - La muestra de unidades experimentales.
  - La inferencia acerca de la población basada en la información contenida en la muestra.
  - Una medida de confiabilidad para la inferencia.

1.3 Tipos de Datos

- Datos Cuantitativos son los que representan cantidades de algo, medidos en una escala númerica.
- Datos Cualitativos no poseen interpretación cuantitativa. Sólo pueden ser clasificados. Ejemplo: Los n trabajos que realizan n graduados de ingeniería después de un año. La clasificación de los estratos económicos, es cualitativa pero ordinal, sabemos que a mayor número mayor ingreso.
- La herramienta estadística propiamente tal, usada para describir y analizar datos, dependerá del tipo de dato. De ahí la importancia de si es cuantitativo o cualitativo.

1.4 Estadística Descriptiva

- El objetivo es presentar métodos gráficos y númericos para explorar, resumir, y describir datos.

1.4.1 Métodos Gráficos y Númericos para Describir Datos Cualitativos

- Asumiendo que tenemos un conjunto de datos reunidos de interés para uno, ¿Como podemos darle sentido? ¿Cómo podemos organizarlos de tal forma que sean más comprensibles y significativos?
- La respuesta depende de los datos.
- Cuando es cualitativo se grupa en categorías.
- La frecuencia de categoría (o clases) para una categoría dada es el número de observaciones que cuentan en esa categoría.
- La frecuencia relativa de la categoría (o clase) para una categoría dada es la proporción de el número total de observaciones que cuentan en esa categoría.
- Ejemplo 1: Investigación de seguridad de reactores nucleares y el riesgo de uso de distintas fuentes de energía. Accidentes observados desde 1977, publicado en "Safety of nuclear power reactors". Nuclear Issues Briefing Paper 14, November 2004.

1.4.1 Métodos Gráficos y Númericos para Describir Datos Cualitativos

- El gráfico de barras da la frecuencia (o frecuencia relativa) para cada categoría donde el largo de la barra es proporcional a la frecuencia (o frecuencia relativa) de la categoría.
- El gráfico de tortas divide un círculo completo en trozos, uno para cada categoría, donde el ángulo es proporcional la frecuencia (o frecuencia relativa) para cada categoría donde el largo de la barra es proporcional a la frecuencia (o frecuencia relativa) de la categoría.
- El diagrama de Pareto (en honor a Vilfredo Pareto un economista italiano) es un gráfico de barras de frecuencias, desplegadas en orden descendente. Es muy usado en control de procesos y calidad, con la primera categoría indicando la mayor falla, etc. La acumulación (denominada línea de acumulación) es graficada con una línea impuesta sobre las barras.

1.4.2 Métodos Gráficos para Describir Datos Cuantitativos

- Los datos cuantitativos son grabados en escalas númericas significativas.
- Hay métodos gráficos de punto, despliegue stem-and-leaf (tallos y hojas), e histogramas. Los primeros dos ya no se usan, por razones de potencia gráfica y de cálculo computacional.
- Ejemplo 2: Datos de rendimiento de los nuevos vehículos medidos en millas por galón, recolectados por la Environmental Protection Agency (EPA).
- El histograma es un gráfico que se construye de partir de generar intervalos de clases para los cuales contamos la frecuencia de datos observados que caen en los distintos intervalos de clases.
- Desventaja, no muestra el valor de las medidas individuales, por ejemplo, el hecho que se repita un punto.

1.4.2 Métodos Gráficos para Describir Datos Cuantitativos

- Pasos a seguir para construir un histograma:
  1. Cálculo del rango de los datos:

     $$\text{rango} = \text{máximo dato observado - mínimo dato observado.}$$

  2. Divida el rango entre 5 a 20 clases de igual ancho. El valor más bajo va primero.
  3. Por cada clase, se cuenta el número de observaciones en esa clase. Ello es denominado la frecuencia de la clase.
  4. Calcular cada frecuencia relativa de clase:

     $$\text{Frecuencia relativa de clase} = \frac{\text{Frecuencia de clase}}{\text{Número total de medidas}}.$$

  5. El histograma es un gráfico de barras en el cual las categorías son conjunto. Si es un histograma de frecuencia, las alturas son determinadas por la frecuencia de clases. Y en un histograma de frecuencia relativa de clase, las alturas de las barras son determinadas por la frecuencia relativa de clase.

1.4.2 Métodos Gráficos para Describir Datos Cuantitativos

- Otro gráfico, es el gráfico de densidad, donde se gráfica la distribución de probabilidad de los datos.

1.4.3 Métodos Númericos para Describir Datos Cuantitativos

- Las medidas descriptivas númericas son valores calculados desde los datos, y nos ayuda a crear una imagen mental de su histograma de frecuencias relativas.
- Las medidas a presentar están en tres categorías:
  1. Las que ayudan a localizar el centro de la distribución de las frecuencias relativas.
  2. Las que miden la dispersión alrededor del centro.
  3. Las que miden la posición relativa de una observación dentro del conjunto de datos.
- Una estadística es una medida numérica calculada desde la muestra de datos.
- Un parámetro es una medida numérica descriptiva de una población, generalmente notada con símbolos griegos.

1.4.4 Medidas de Tendencia Central

- La media aritmética de un conjunto de n medidas, $x_1, \ldots, x_n$, es el promedio de las medidas:

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{1}$$

también se denomina media muestral, como la media de una muestra de n medidas.

- La mediana de un conjunto de n medidas, $x_1, \ldots, x_n$, es el número medio cuando las medidas son arregladas en orden ascendente (o descendente), i.e, el valor de x localizado a la mitad del área bajo el histograma de frecuencia relativa que tiene lugar a su izquierda y la mitad del área que tiene lugar a su derecha. Se usa el símbolo m para representar la mediana de la muestra, y τ para representar la mediana de la población.
  Si $x_{(i)}$ denota el i-ésimo valor de la muestra cuando esta ordenada en orden ascendente. La mediana de la muestra es calculada como sigue:

$$m = \begin{cases} x_{\lfloor \frac{(n+1)}{2} \rfloor} & \text{si } n \text{ es impar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ es par} \end{cases}. \tag{2}$$

1.4.4 Medidas de Tendencia Central

- La moda de un conjunto de n medidas, $y_1, \ldots, y_n$, es el valor de x que ocurre con mayor frecuencia.
- La media es la medida preferida de tendencia central, pero no dice nada respecto de la asímetria (skewness) (la cola de la distribución).
- La mediana es denominada una medida de resistencia de la tendencia central, ya que la media, es resistente a las influencias de observaciones extremas.
- La moda sólo es importante si la frecuencia relativa de x es de interés.

1.4.5 Medidas de Variación

- Las medidas de variación más usadas son el rango, la varianza y la desviación estándar.
- El rango es igual a la diferencia entre la mayor y la menor medida en un conjunto de datos:

$$\text{rango} \;=\; \text{máximo dato observado - mínimo dato observado.} \tag{3}$$

- La varianza de una muestra de n medidas, $x_1, \ldots, x_n$, es definida como:

$$s_n^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2}{n-1}, \tag{4}$$

La varianza de una población finita con n medidas es definida como:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}. \tag{5}$$

1.4.5 Medidas de Variación

- La desviación estándar de una muestra de n medidas es igual a la raíz de la varianza:

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}{n-1}}, \tag{6}$$

  La desviación estándar de una población es:

$$\sigma = \sqrt{\sigma^2}. \tag{7}$$

- Ver Ejemplo 2, varianza.

1.4.6 Medidas de Posición Relativa

- Las medidas de Posición Relativa de una observación son los percentiles y los z-scores, e indican la localización de una observación relativa respecto a otros puntos en la distribución.
- El percentil $100p^\circ$ de un conjunto es un valor de x localizado tal que el $100p\%$ de el área bajo la distribución de frecuencia relativa para los datos está contenida a la izquierda de el $100p^\circ$ percentil y $100(1-p)\%$ del área está contenida a su derecha. (Notar que $0 \leq p \leq 1$.)
- El cuartil más bajo, Q_L, para un conjunto de datos es el percentil 25°.
- El cuartil medio o mediana, m, para un conjunto de datos es el percentil 50°.
- El cuartil superior, Q_U, para un conjunto de datos es el percentil 75°.

1.4.6 Medidas de Posición Relativa

- Pasos a seguir para construir cuartiles:
  1. Ordene los datos de menor a mayor. Sean $x_{(1)}, \ldots, x_{(n)}$ los datos ordenados.
  2. Cálcule la cantidad $l = \frac{1}{4}(n+1)$ y redondeé al entero más cercano. La medida con este rango, $x_{(l)}$, representa el cuartil más bajo o percentil 25°.
  3. Cálcule la cantidad $u = \frac{3}{4}(n+1)$ y redondeé al entero más cercano. La medida con este rango, $x_{(u)}$, representa el cuartil más alto o percentil 75°.
- Ver ejemplo 2.

1.4.6 Medidas de Posición Relativa

- El z-scores para un valor x del conjunto de datos es la distancia de x sobre o bajo la media, en unidades de la desviación estándar:

$$\text{z-cores muestra } = \frac{x - \bar{x}_n}{s_n}, \tag{8}$$

$$\text{z-cores población } = \frac{x - \mu}{\sigma}. \tag{9}$$

1.4.7 Medidas de Asimetría

- Skewness o Coeficiente de asimetría de Fisher:

$$\gamma_1 = IE\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}. \tag{10}$$

- Si $\gamma_1 > 0$, la distribución es asimétrica positiva o a la derecha.
- Si $\gamma_1 < 0$, la distribución es asimétrica negativa o a la izquierda.

1.4.8 Medidas de Concentración de Datos

- Curtosis o Kurtosis:

$$\beta_2 = IE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}. \tag{11}$$

- Si $\beta_2 > 3$, la distribución es más apuntada y con colas más gruesas que la normal.
- Si $\beta_2 < 3$, la distribución es menos apuntadas y con colas menos gruesas que la normal.
- Si $\beta_2 = 3$, la distribución es normal.