Probabilidad y Estadística Fundamental

Estadística descriptiva - Medidas descriptivas de centro, dispersión, localización y forma

Profesor: Nicolás López

Universidad Nacional de Colombia

Contenido

Contenido

Motivación

¿Para qué usar números?

Hasta ahora, ¿qué hemos logrado?

- Los datos han sido recolectados y dispuestos en una base de manera adecuada.
- Las variables han sido clasificadas según su escala, dimensión y clase.
- Dada la caracterización de las variables univariadas, estas han sido resumidas de manera apropiada a través de gráficos y/o tablas.
- Diferentes características tales como ubicación, dispersión y simetría han sido determinadas a partir de gráficas.

Motivación

¿Para qué usar números?

Es importante resaltar que

- La descripción gráfica y tabular es muy útil en la **descripción general** de un conjunto de datos.
- **Cuantificar** las características de las variables de interés permitiría una caracterización **más objetiva y particular** de las mismas.

Motivación

¿Para qué usar números?

Se requiere una forma concisa resumir las características importantes de los datos:

- La ubicación en la escala de medida de un punto central o *centro*, aquel con preponderancia de observaciones. ¿Cómo lo mido?.
- La *dispersión* de las observaciones respecto a este punto central. ¿hay una alta/baja concentración de valores cerca a este punto?. ¿Cómo lo mido?.
- La ubicación en la escala de medida del valor máximo, mínimo ó alguna otra *localización* de interés de la variable. ¿Cómo lo mido?.
- La *forma* de la distribución de la variable. ¿Cómo lo mido?.

Motivación

¿Para qué usar números?

La **cuantificación** de estas propiedades permite una posterior elaboración de **procedimientos estadísticos** en el análisis de datos.

Motivación

¿Para qué usar números?

Ya sacrifiqué la información individual de los datos a través de gráficos con el objetivo de ganar interpretabilidad en la distribución de los datos. Ahora, a través de estos números, o *medidas representativas* ¿perderé conocimiento de la estructura distribucional de los datos?

Contenido

Estadísticas y parámetros

El conjunto de números que proveen una buena "imagen mental" de la distribución de frecuencias son llamados **parámetros**. Cuando estos números provienen de una muestra, son llamados **estadísticas**:

Parámetros

Las mediciones descriptivas numéricas asociadas con una **población** de mediciones se llaman parámetros.

Estadísticas

Las mediciones descriptivas numéricas asociadas con una **muestra** de mediciones se llaman estadísticas.

Estadísticas y parámetros

Así como un gráfico/tabla pueden variar dependiendo de la muestra tomada, una estadística puede variar de muestra a muestra. Por el contrario, un parámetro poblacional es constante.

En general, para un parámetro dado:

1. El parámetro es desconocido.
2. Se obtiene **una** muestra de la población.
3. A partir de la información de la muestra se calcula **un** estadístico para **estimar** el parámetro.

Contenido

Medidas de centro

¿Dónde preponderan los datos?

Hay diferentes opciones para determinar un número que defina la ubicación central de los datos, un *punto de equilibrio* de la distribución. Las **medidas de tendencia central** más conocidas son:

- Media.
- Mediana.
- Moda.

Algunas otras son

- Media geométrica.
- Media armónica.
- Punto rango medio.

Medidas de centro

¿Dónde preponderan los datos? - Media

Para un conjunto de n mediciones $x_1,...,x_n$ se define la media aritmética muestral, o media muestral $\bar{x}$, como

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

La media poblacional es usualmente denotada por la letra μ.

Medidas de centro

¿Dónde preponderan los datos? - Media

Notas

- $\bar{x}$ considera todos los datos con la misma importancia. Les da un **peso** de $1/n$. De manera similar para μ.
- $\bar{x}$ es **sensible** ante datos atípicos[1].
- Aplicable a datos de intervalo y de razón.

[1] Cuando un conjunto de datos tiene valores muy pequeños u observaciones muy grandes, la media muestral se traza hacia la dirección de las mediciones extremas.

Medidas de centro

¿Dónde preponderan mis datos? - Mediana

Para un conjunto de n mediciones $x_1,\ldots,x_n$ se define la mediana m como es el valor de la variable que cae en la posición media cuando las mediciones son ordenadas de menor a mayor.

Medidas de centro

¿Dónde preponderan los datos? - Mediana

Notas

- Una vez ordenadas las mediciones, el valor $0{,}5(n+1)$ indica la posición de la mediana.
- A diferencia de $\bar{x}$, m es una estadística **robusta**.
- Aplicable a datos ordinales, de intervalo y de razón.

Medidas de centro

¿Dónde preponderan los datos? - Moda

Para un conjunto de n mediciones $x_1,...,x_n$ se define la moda como es el valor de la variable que se presenta con más frecuencia.

Medidas de centro

¿Dónde preponderan los datos? - Moda

Notas

- Es posible que una distribución de mediciones tenga más de una moda, o que no tenga ninguna.
- Si hay múltiples modas y son adyacentes, la distribución se considera unimodal, con moda igual a la media de las modas.
- Cuando las mediciones en una variable continua se han agrupado como histograma de frecuencia, la clase con mayor frecuencia se llama **clase modal**, y el punto medio de esa clase se toma como la moda.
- Aplicable a datos nominales, ordinales, de intervalo y de razón.

Contenido

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos?

Hay diferentes opciones para determinar un número que defina la dispersión de los datos respecto al centro de la distribución, que cuándo las observaciones se encuentren concentradas alrededor del centro de la distribución provean una medida de baja variabilidad, y cuándo se encuentren dispersas, una alta variabilidad. Las **medidas de variabilidad** más conocidas son:

- Rango.
- Varianza y desviación estándar.
- Coeficiente de variación.

Algunas otras son

- Rango intercuartílico.
- Promedio de desvíos absolutos.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Rango

Para un conjunto de n mediciones $x_1,\ldots,x_n$ se define el rango muestral R, como la diferencia entre la medición más grande y la más pequeña.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Rango

Notas

- R tiene **las mismas** unidades que los datos observados.
- Si todos los datos de la muestra son iguales, $R = 0$.
- R aumenta a medida que exista mayor variabilidad en los datos.
- El rango muestral usualmente subestima el rango poblacional.
- Al estar definido en función del máximo y el mínimo, es **altamente** afectado por outliers.
- Aplicable a datos ordinales, de intervalo y de razón.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Varianza

Para un conjunto de n mediciones $x_1,...,x_n$ se define la varianza muestral s^2, como

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

La varianza poblacional es usualmente denotada por la letra σ^2.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Varianza

Notas

- s^2 provee el grado de representatividad de **la media** de los datos y tiene como unidades las unidades **cuadradas** de los datos observados.
- Si todos los datos de la muestra son iguales, $s^2 = 0$.
- s^2 aumenta a medida que exista mayor variabilidad en los datos.
- Aplicable a datos de intervalo y de razón.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Varianza

La varianza poblacional está definida como:

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

El dividir por $n - 1$ en lugar de n en la fórmula de s^2 permite obtener un mejor estimador para σ^2 (¿qué hará a un estimador mejor que otro?).

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Desviación estándar

Para un conjunto de n mediciones $x_1,\ldots,x_n$ se define la desviación estándar muestral s, como la raíz cuadrada positiva de la varianza muestral, es decir

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

La desviación estándar poblacional es usualmente denotada por la letra σ.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Desviación estándar

Notas

- s provee el grado de representatividad de la media de los datos y tiene como unidades **las mismas** de los datos observados.
- Si todos los datos de la muestra son iguales, $s = 0$.
- s aumenta a medida que exista mayor variabilidad en los datos.
- Aplicable a datos de intervalo y de razón.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Coeficiente de variación

Para un conjunto de n mediciones $x_1,\ldots,x_n$ se define la el coeficiente de variación muestral cv, como el cociente entre la desviación estándar muestral y la media muestral, es decir

$$cv = \frac{s}{\bar{x}}$$

Al dividir s por $\bar{x}$, cv **no tiene unidades**.

Medidas de dispersión

Respecto al punto medio, ¿qué tan dispersos se encuentran los datos? - Coeficiente de variación

Notas

- Al no tener unidades, cv permite cuantificar la variabilidad respecto al valor medio independiente de las unidades de medida.
- Empíricamente, se ha observado, que cuando $|cv| < 0{,}1$ la varianza de los datos, por lo general, no es grande.

Ejemplo de aplicación

Cálculo manual de medidas resumen

Calcular manualmente (papel y lápiz) para las variables número de hijos = c(1,2,3,4,5) y salario = (999.998,999.999,1.000.000,1.000.001,1.000.002) las medidas de centro y dispersión detalladas en las clase.

Conclusión

- Los estadísticos (muestrales) y parámetros (poblacionales) formalizan las características de la distribución de datos: además de **cualificar** mediante gráficos, podemos **cuantificar** las características distribucionales de las variables.
- Existen múltiples formas de cuantificar la tendencia central y la dispersión de un conjunto de datos. Cada una tiene supuestos para la variable medida.

Contenido

Medidas de localización

¿Dónde están ubicados los datos?

Antes de continuar. ¿Cómo se diferencian las variables salario y número de hijos del ejercicio anterior en términos de su dispersión? ¿a qué conclusión puede llegar respecto a este par de variables?

Medidas de localización

¿Dónde están ubicados los datos?

Cuándo interesa la posición de una observación respecto a otras de un conjunto de datos, se puede determinar su **posición relativa** mediante

- Puntaje muestral z.
- Cua**n**tíles.

Dónde los cuantiles presentan diferentes variaciones, como por ejemplo:

- Centíles ó percentíles.
- Cua**r**tiles.
- Quintiles.
- Deciles.

Medidas de localización

¿Dónde están ubicados los datos? - Puntaje z

Para un conjunto de n mediciones $x_1,\ldots,x_n$ con media y desviación estándar muestral $\bar{x}$ y s respectivamente, se define el puntaje muestral z_i para la i-ésima observación x_i como

$$z_i = \frac{x_i - \bar{x}}{s}$$

Medidas de localización

¿Dónde están ubicados los datos? - Puntaje z

Notas

- El puntaje z mide la distancia entre una observación x y la media $\bar{x}$, medidas en unidades de desviación estándar s.

$$x = \bar{x} + sz$$

- El puntaje z es una herramienta para determinar la **atipicidad** de una observación dada. Los puntajes z *grandes* en valor absoluto tienden a ser poco comunes.
- Aplicable a datos de intervalo y de razón.

Medidas de localización

¿Dónde están ubicados los datos? - Cuantíles

Para un conjunto de n mediciones $x_1,...,x_n$, se definen los cuantíles como valores que separan proporciones determinadas de los n datos.

Medidas de localización

¿Dónde están ubicados los datos? - Cuantíles

Notas

- La mediana es el cua**n**tíl medio. Separa en 2 partes iguales el conjunto ordenado de datos.
- Si los datos ordenados son divididos en 4 partes iguales se obtienen 3 cuantíles denominados **cuartíles**:
  1. $1/4$=25 % de las observaciones ordenadas son menores que el primer cuartíl $\mathbf{Q_1}$.
  2. $2/4$=50 % de las observaciones ordenadas son menores que el segundo cuartíl $\mathbf{Q_2}$.
  3. $3/4$=75 % de las observaciones ordenadas son menores que el tercer cuartíl $\mathbf{Q_3}$.

Medidas de localización

¿Dónde están ubicados los datos? - Cuantíles

Notas

- Si los datos ordenados son divididos en 100 partes iguales se obtienen 99 cuantíles denominados **percentíles**:
  1. $1/100{=}1\,\%$ de las observaciones ordenadas son menores que el primer percentíl p_1.
  2. $2/100{=}2\,\%$ de las observaciones ordenadas son menores que el segundo percentíl p_2.
  3. ...

Medidas de localización

¿Dónde están ubicados los datos? - Cuantíles

Notas

- Los cuantiles son **valores de la variable** en posiciones determinadas, **no** las posiciones.
- La mediana, que es le cuantíl medio, también es igual al segundo cuartíl, al quinto decíl, el 50vo percentíl, ...
- Aplicable a datos ordinales, de intervalo y de razón.

Medidas de localización

¿Dónde están ubicados los datos? - Medidas de localización para medir dispersión

Notas

- El rango R es una medida de dispersión definida en términos de medidas de localización (*máx* - *mín*).
- Los cuantiles permiten definir medidas de dispersión similares, *pero mejores*, al rango. Por ejemplo, a partir de cuartíles se obtiene la **amplitud intercuartílica** como $Q_3 - Q_1$.

Contenido

Medidas de forma

¿Qué forma tienen los datos?

Además de la magnitud de la dispersión respecto al punto medio, interesa la **manera** en la que los datos se distribuyen en torno a este en términos de la **asimetría** y la **curtosis** de los datos.

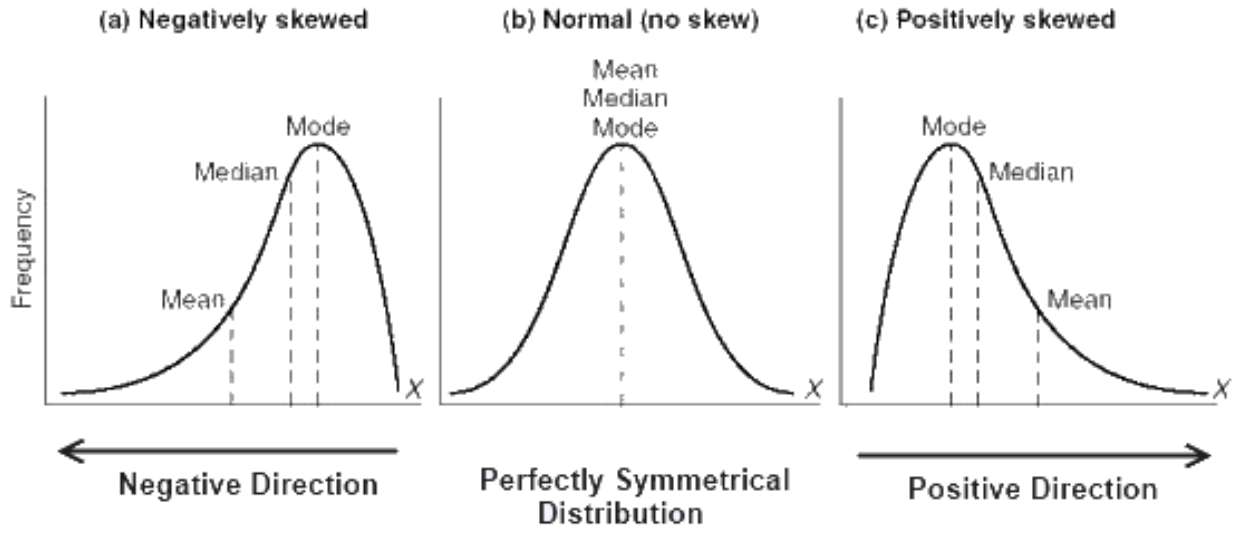
El objetivo en este punto es cuantificar la **forma** de la representación gráfica de la distribución.

Medidas de forma

¿Qué forma tienen los datos? - Simetría

¿Cómo cuantificar la simetría?

Figura 1: Simetría de distribuciones. Variable continua



En orden: distribución con asimetría a izquierda, insesgada y con asimetría a derecha.

Medidas de forma

¿Qué forma tienen los datos? - Simetría

Si definimos

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

Nótese que

- μ_3 tiene como unidades las unidades cúbicas de los datos observados.
- $\mu_3 < 0$ cuándo la distribución tiene sesgo negativo respecto a $\bar{x}$.
- $\mu_3 = 0$ cuándo la distribución es simétrica respecto a $\bar{x}$.
- $\mu_3 > 0$ cuándo la distribución tiene sesgo positivo respecto a $\bar{x}$.

Medidas de forma

¿Qué forma tienen los datos? - Simetría

Por lo cual, al definir

$$g_1 = \frac{\mu_3}{s^3}$$

Se tiene

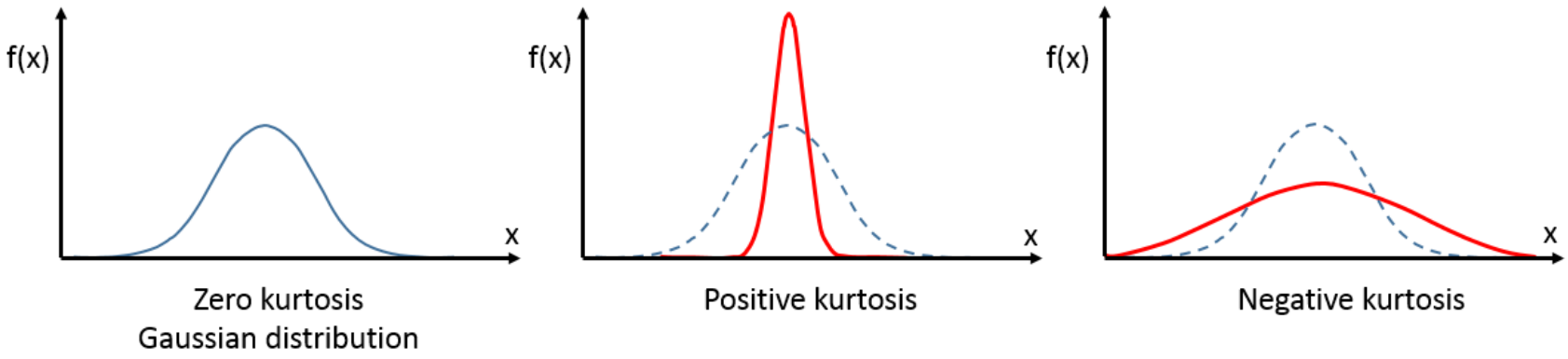
- g_1 no tiene unidades.
- $g_1 < 0$ cuando la distribución tiene sesgo negativo respecto a $\bar{x}$.
- $g_1 = 0$ cuando la distribución es simétrica respecto a $\bar{x}$.
- $g_1 > 0$ cuando la distribución tiene sesgo positivo respecto a $\bar{x}$.

Medidas de forma

¿Qué forma tienen los datos? - Curtosis

¿Cómo cuantificar la curtosis?

Figura 2: Curtosis de distribuciones. Variable continua



En orden: distribución mesocúrtica, leptocúrtica y platicúrtica.

Medidas de forma

¿Qué forma tienen los datos? - Curtosis

Nótese que

- ¿cuándo existe curtosis? $\rightarrow$ es necesaria una **referencia** de cuándo no existe curtosis.
- Tomando como referencia la **distribución normal**, también llamada distribución **gaussiana** (ya llegaremos a este tema), se puede definir si una distribución observada es mas o menos apuntada que esta.

Medidas de forma

¿Qué forma tienen los datos? - Curtosis

Cuando la distribución es normal, $\mu_4/s^4=3$, por lo cual, al definir

$$g_2 = \frac{\mu_4}{s^4} - 3$$

Se tiene

- g_2 no tiene unidades.
- $g_2 < 0$ cuando la distribución es platicúrtica.
- $g_2 = 0$ cuando la distribución es mesocúrtica (curtosis gausiana).
- $g_2 > 0$ cuando la distribución es leptocúrtica.

Ejemplo de aplicación

Tercer laboratorio de programación en R

Realice el tercer laboratorio de programación en R.

Conclusión

- Los estadísticos (muestrales) y parámetros (poblacionales) formalizan las características de la distribución de datos: además de **cualificar** mediante gráficos, podemos **cuantificar** las características distribucionales de las variables.
- Las características de ubicación y forma complementan la caracterización de centro y dispersión, y viceversa. Son todas requeridas para comprender la distribución de un conjunto de datos.