Repaso De Estadística

0.1. Medidas de tendencia central

0.2. Medias

0.2.1. Media aritmética

La media aritmética de una muestra es la suma de cada uno de los valores posibles multiplicado por su frecuencia, es decir.

Si la siguiente tabla representa la tabla de frecuencia de la muestra

M	n_i	f_i
x_1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k

La media es el valor :

$$\_x = \frac{k\sum_{i=1} x_i n_i}{n} \tag{1}$$

y si los datos no están ordenados entonces

$$\_x = \frac{k\sum_{i=1} x_i}{n} \tag{2}$$

Observación 1 *En la definición de media se consideró que la variable de interés X es discreta, pero si la variable X no es discreta sino continua. En la fórmula se reemplaza cada valor x_i por la marca de clase correspondiente es decir*

$$\_x = \frac{k\sum_{i} = 1 m_i n_i}{n} \tag{3}$$

Este proceso hace que la media aritmética difiera de la media obtenida según (2.1), es decir habrá una perdida de precisión que será mayor en cuanto mayor sea la diferencia entre las marcas de clase y los valores reales, o sea entre mayor sea la longitud a_i de los intervalos

Desventajas de la media La media es una medida muy usada en estadística, pero a pesar de eso posee ciertas desventajas

- La media aritmética es muy sensible a los valores extremos, es decir si una medida se aleja mucho de las otras hará que la media se aproxime mucho a ella
- No se recomienda usar cuando los datos se desplazan hacia los extremos
- En el caso de variables continuas depende de los intervalos de clase
- en le caso de variables discretas el valor puede no ser un valor de la muestra.

0.2.2.

Otra media es la llamada media cuadrática $\bar{x}_c$ la cual es la raíz cuadrada de la media aritmética de los cuadrados

$$\bar{x}_c = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

0.3. La mediana

Sea X una variable discreta cuyas observaciones han sido ordenadas de mayor de mayor a menor, entonces se le llama mediana $\widetilde{x}$ al primer valor de la variable que deja por debajo de si el 50 % de las observaciones es decir si n es el número de observaciones, la mediana será la observación $\left[\left|\frac{n}{2}\right|\right] + 1$

Definición 1 *Sea $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n)}$ las observaciones de una muestra para una variable X donde $x_{(1)}$ representa la observación más pequeña, $x_{(2)}$ la observación que le sigue en valor y así sucesivamente $x_{(n)}$ denota la observación de mayor valor, entonces la mediana se define*

$$\widetilde{x} = \begin{cases} x_{([n+1]/2)} & \text{si } n \text{ impar} \\ \frac{x_{(n/2)} + x_{([n/2]+1)}}{2} & \text{si } n \text{ par} \end{cases}$$

En el caso de variables continuas, las clases vienen dadas por intervalos como se indicó en el capítulo anterior por tal razón para determinar la mediana se escoge el intervalo donde se encuentra el valor para el cual están debajo de él la mitad de los datos. Entonces a partir de ese

intervalo se observan las frecuencias absolutas acumuladas y se aplica la siguiente fórmula

$$\widetilde{x} = x_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i$$

de aquí se puede deducir que $\widetilde{x}$ el "punto" que divide al histograma en dos partes de áreas iguales

0.3.1. Propiedades y desventajas de la mediana

1. Tiene la ventaja de no ser afectada por los valores extremos y por eso se aconseja para distribuciones para las cuales los datos no se concentran en el centro
2. Es fácil de calcular
3. En el caso de variables discretas el valor de la mediana es un valor de la variable
4. El mayor defecto es que las propiedades matemáticas son muy complicadas y esto hace que muy poco se use para realizar inferencias
5. Es función de los intervalos escogidos en el caso de variables continuas

0.3.2. La moda

Llamaremos moda $\widehat{x}$ a cualquier máximo relativo de la distribución de frecuencias, es decir, cualquier valor de la variable que posea una frecuencia mayor que su anterior y posterior valor.

En el caso de variables continuas es más correcto hablar de intervalos modales. Luego de determinar el intervalo de clase o intervalo modal,que es aquel para el cual la distribución de frecuencia posee un máximo relativo, se determina la moda utilizando la siguiente fórmula

$$\widehat{x} = x_{i-1} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} a_i$$

Propiedades de la moda

La moda posee la siguientes propiedades

- Es muy fácil de calcular
- Puede no ser única
- Es función de los intervalos de su amplitud,número y límites

0.3.3. Relación entre la media, la moda y la mediana

En el caso de distribuciones unimodales, la mediana está con frecuencia comprendida entre la media y la moda (incluso más cerca de la media).

En distribuciones que presentan cierta inclinación, es más aconsejable el uso de la mediana. Sin embargo en estudios relacionados con propósitos estadísticos y de inferencia suele ser más apta la media.

Veamos un ejemplo de cálculo de estas tres magnitudes.

1. Medidas de posición

A veces es importante obtener los valores de la variable que dividen la población en cuatro,diez o cien partes iguales,usualmente llamados cuartiles deciles y percentiles respectivamente.

El procedimiento es similar al utilizado para determinar la mediana, como lo indicaremos ahora.

1.0.4. Percentil

Para una variable discreta, se define el percentil de orden k, como la observación que deja por debajo de si el $k\,\%$ de la población es decir $\mathrm{N}_k = n\frac{k}{100}, sinesimpar$ es decir $\mathrm{p}_k = x_{[|(n+1)\frac{k}{100}|]}$ donde en sub índice $[|(n+1)\frac{k}{100}|]$ indica que es la posición k a la que le corresponde ese valor de la frecuencia absoluta acumulada. Si n es par

$$p_k = \frac{x_{[|n\frac{k}{100}|]} + x_{[|n\frac{k}{100}|]+1}}{2}$$

En el caso de variables continuas se busca el intervalo donde se encuentra p_k, es decir se busca el valor que deja por debajo de si el $k\,\%$ de las observaciones y se determina el intervalo $(x_{i-1}, x_i]$ donde se encuentra y se utiliza la relación

$$p_k = x_{i-1} + \frac{n\frac{k}{100} - N_{i-1}}{n_i} \cdot a_i$$

1.0.5. Quartiles

Los cuartiles son tres y se definen:

- $Q_1 = p_{25}$
- $Q_2 = p_{50}$
- $Q_3 = p_{75}$

1.1. Deciles

De manera análoga se definen los deciles

los deciles son los valores que dividen las observaciones en 10 grupos de igual tamaño es decir son el conjunto

$D_1, D_2, D_3, \cdots, D_{10}$ y se definen

$$D_i = p_{10 \cdot i} \qquad i = 1, 2, 3, \cdots, 10$$

2. Medidas de variabilidad o dispersión

2.1. Varianza y desviación t´ı pica

Como forma de medir la dispersión de los datos hemos descartado:

La varianza, S_n^2, se define como la media de las diferencias cuadráticas de n puntuaciones con respecto a su media aritmética, es decir

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

Para datos agrupados en tablas, usando las notaciones establecidas en el capítulo anterior, la varianza se puede escribir como

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i$$

Una fórmula equivalente para el cálculo de la varianza está basada en lo siguiente:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 n_i - \overline{x}^2$$

La varianza no tiene la misma magnitud que las observaciones (ej. si las observaciones se miden en metros, la varianza lo hace en $metros^2$). Si queremos que la medida de dispersión sea de la misma dimensionalidad que las observaciones bastará con tomar su raíz cuadrada. Por ello se define la desviación típica, S_n, como

$$S_n = \sqrt{S_n^2}$$