**ESTADISTICA DESCRIPTIVA**

**Parte I: Análisis Univariados**

**Contenidos**:

* 1. Vocabulario básico
  2. Medidas descriptivas
     1. De tendencia central
     2. De dispersión
     3. De posición
     4. De forma y apuntamiento
     5. De asociación
  3. Gráficos estadísticos
  4. Representaciones Tabulares
  5. Práctica con Software Excel, estat, R
  6. **Vocabulario básico**

**Población** es un grupo de unidades que tienen alguna característica cuantificable en común. Las unidades pueden ser personas, árboles, bacterias, compuestos químicos, etc.. Pueden ser finitas o infinitas en número.

Población Objetivo (blanco): la población que realmente se quiere estudiar

Población en estudio: bajo algunas limitaciones, es esta la población a la cual realmente se tendrá acceso en el estudio.

**Muestra** un subconjunto de la población, seleccionada de manera tal que representa a una población mayor

* 1. **Medidas descriptivas**

El Objetivoperseguido con estas medidas esresumir y sintetizar un conjunto de datos mediante **un único** valor numérico.

En adelante utilizaremos la siguiente notación:

N: tamaño de la población

n: tamaño de la muestra

Muestra de datos para la variable X: 

* + 1. **De tendencia central:**

El objetivo es resumir un conjunto de datos a través de un solo valor numérico, estas medidas pretender lograr este difícil propósito, en general se pueden ver como valores alrededor de los cuales se agrupan los datos.

**Media:** punto medio en la distribución de los datos. También se le conoce

como media aritmética o promedio.

Notación: si se tiene una muestra  esta es la media muestral

Fórmula:

Si se tiene toda la población  esta es la verdadera media o media poblacional.

Ventajas:

Fácil de interpretar.

Es única.

Se puede comparar

El cálculo se realiza con todos los valores de la variable.

Tiene un cálculo sencillo

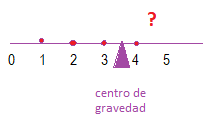
Desventajas: Es susceptible a la presencia de valores extremos.

Tenga en cuenta que la media se ubica al interior del intervalo definido entre el mínimo y el máximo de la muestra pero no necesariamente es un valor observado o acorde a la escala de medición, ejemplo, el número promedio de hijos por familia puede ser 1.2.

Problemas con la interpretación

Ejemplo 1: ¿Es posible inventar tres conjuntos con 4 notas (de 0 a 5) con características muy diferentes pero que su media sea de 3.0?

Ejemplo 2: Usando notas (de 0 a 5), es posible adicionar un dato al conjunto de datos {2.0, 4.0,3.0, 1.0, \_\_ } tenga un media de 3.5?



Ejemplo 3: Si estas son las notas de tres estudiantes, ¿qué podemos concluir al comparar su rendimiento académico a través del promedio?

|  |  |
| --- | --- |
| Notas |  |
| 3.0 3.0 3.0 3.0 | 3.0 |
| 2.5 3.0 3.0 3.5 | 3.0 |
| 0.0 4.0 4.0 4.0 | 3.0 |

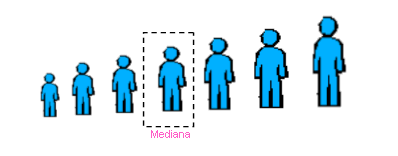
Ahora considere, ¿qué diría si compara el comportamiento examinando en detalle las notas de cada estudiante?

Ejemplo 4: Para cierta institución educativa en la cual sólo la tercera parte son mujeres, los promedios en el exámen del ICFES discriminados por género son de 268 para hombres y 255 para las mujeres. Las directivas de la institución explican que este comportamiento se debe a que hubo más hombres presentando la prueba. ¿Es válida esta interpretación?

**Mediana**: la observación central una vez **ordenados** los datos.

Notación: no es estándar por ejemplo: , Me, …

Fórmula:



Si n es impar, es el dato en la posición .

Si n es par, la mediana es el promedio de los datos que están en la posición  y .

Ventajas:

No se ve afectada por la presencia de valores extremos.

Es fácil de calcular y de entender.

Se recomienda reportarla cuando la distribución de los datos muestra valores extremos ( distribución asimétrica).

|  |  |
| --- | --- |
|  |  |

Ejemplo: Observe los siguientes conjuntos de datos:

|  |  |  |
| --- | --- | --- |
| Datos | Media | Mediana |
| 2.5 **3.0 3.0** 3.5 | 3 |  |
| 0.0 **3.0 3.0** 3.5 | 2.37 |  |
| 1.0 2.0 3.0 4.0 5.0 | 3 |  |

**Moda:** es el dato que más se repite.

Notación: no es estándar Mo. Se debe utilizar cuando la variable es categórica.

Fórmula: simplemente observar los conteos o frecuencias de cada categoría de la variable. Puede darse el caso de que haya una, dos o más de dos modas, entonces se dice que la distribución es unimodal, bimodal o multimodal respectivamente.

**Actividad**: Explore el comportamiento de la media y la mediana a través del este recurso <https://www.geogebra.org/m/K8zZ8eXg>

¿Qué características tienen los datos en los cuales la media es igual a la mediana?

¿Qué características tienen los datos en los cuales la media y la mediana difieren considerablemente?

¿Qué cambios en los datos no afectan a la media?, y a la mediana?

¿Qué características tienen los datos en los cuales la media, la mediana y la moda difieren considerablemente?

**Nota**: Existen otras medidas de tendencia central como son: Media geométrica, Media Recortada y Trimedia. Es conveniente consultar en qué casos es apropiada cada una de ellas.

* + 1. **De dispersión:**

Se calculan para complementar la descripción adicionando información del comportamiento de los datos alrededor de la medida de tendencia central reportada.

**Rango** : intervalo donde se encuentran distribuidos los datos. También se le conoce como amplitud o recorrido.

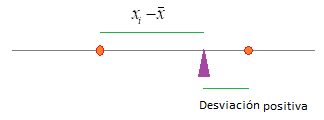
Notación: no es estándar

Fórmula: Dato mayor – Dato menor

**Varianza:**  Observe cuidadosamente la deducción de la fórmula esto le ayudará a entender qué información lleva esta medida.

La idea es reportar un número que de cuenta del grado de dispersión de los datos **alrededor de una medida de tendencia central, lo usual es usar la media**. Para ello empezaremos definiendo una desviación.

Desviación:  (mide la distancia del datos i-ésimo a la media)



Suma de desviaciones:  (Suma de todas las distancias)

Pero hay problemas, veamos que pasa con estos dos conjuntos de datos:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datos | Desviación |  | Datos | Desviación |
| 2.5 | -0.5 |  | 0.0 | -3.0 |
| 3.0 | 0.0 |  | 4.0 | 1.0 |
| 3.0 | 0.0 |  | 4.0 | 1.0 |
| 3.5 | 0.5 |  | 4.0 | 1.0 |
| 3.0 | Suma= 0.0 |  | 3.0 | Suma =0.0 |

Solución: elimínenos el signo! Esto puede hacerse de dos formas, elevando al cuadrado o sacando valor absoluto.

Sumatoria de las desviaciones al cuadrado , veamos que pasa con los datos anteriores:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datos | Desviación |  | Datos | Desviación |
| 2.5 | 0.25 |  | 0.0 | 9.0 |
| 3.0 | 0.0 |  | 4.0 | 1.0 |
| 3.0 | 0.0 |  | 4.0 | 1.0 |
| 3.5 | 0.25 |  | 4.0 | 1.0 |
| 3.0 | Suma= 0.5 |  | 3.0 | Suma =12 |

Observe que la suma de desviaciones da información sobre cuál conjunto de datos tiene más dispersión complementando la información que nos aporta la sola consideración del promedio.

Para que la medida sea representativa podemos pensar en promediarla (dividir entre n) y así se involucra también el tamaño de muestra.

Varianza= esta es la varianza muestral

 esta es la verdadera varianza o varianza poblacional

Más problemas ….

Note que las desviaciones están en las mismas unidades de la variable en estudio pero al elevar al cuadrado cambiarían. Así por ejemplo una estatura esta en metros, pero al hacer las desviaciones al cuadrado nos darán en metros cuadrados lo cual carece de interpretación. Para solucionar este nuevo problema, la solución es muy evidente: saque raíz cuadrada a la varianza.

**Desviación estándar:**

Fórmula:  esta es la desviación estándar muestral.

La desviación estándar de un conjunto de datos es una medida de cuánto se desvían los datos de su media. Esta medida es más estable que el rango y toma en consideración el valor de cada dato. Observe que por la forma cómo se calcula la varianza y la desviación estándar nos asegura que **siempre van a ser mayores que cero**.

Por ejemplo, las tres muestras

Datos 1: 0 0 14 14 Datos 2: 0 6 8 14 Datos 3: 6 6 8 8

cada una tiene una media de 7. Sus desviaciones estándar son 7, 5 y 1, respectivamente. La tercera muestra tiene una desviación mucho menor que las otras dos porque sus valores están más cerca de 7.

Usted encontrará fórmulas para la varianza (o la desviación estándar) divididas por n o por n-1, cuál utilizar?

Cuándo usted divide por n-1, la varianza muestral provee un estimador mucho más cercano a la varianza poblacional que cuando lo hace por n. Note que para tamaños de muestra grandes son muy similares los dos valores por lo que la diferencia sólo es relevante para tamaños de muestras pequeños.

Ejemplo: Comparación de los cálculos para la desviación estándar.

|  |  |  |  |
| --- | --- | --- | --- |
| Datos |  | Sn-1 | Sn |
| 3.0 3.0 3.0 3.0 | 3.0 | 0.0 | 0.0 |
| 2.5 3.0 3.0 3.5 | 3.0 | 0.4 | 0.35 |
| 0.0 4.0 4.0 4.0 | 3.0 | 2.0 | 1.73 |

Si estas son notas de 3 estudiantes, ¿cuál estudiante es mejor?

Ejercicios:

1. Para el grupo M de estudiantes la desviación estándar de las notas en una prueba fué de 2.4 mientras que para el grupo Q fué de 1.2. ¿Qué se puede decir sobre las clases?

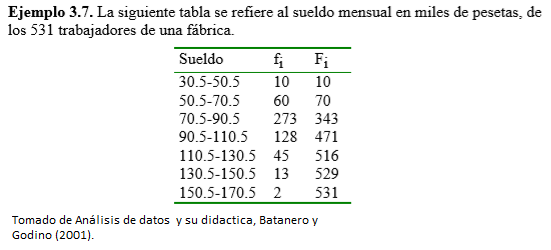
a. El grupo M es más homogéneo que el grupo Q.

b. El grupo Q es menos hetereogéneo que el grupo M.

c. El grupo Q tuvo un rendimiento inferior al grupo M.

d. El grupo M tuvo un desempeño doblemente superior con relación al grupo Q.

1. Trabajo con datos agrupados.



¿Cómo calcular las medidas de tendencia central para este tipo de datos?, ¿Cómo analizar la variabilidad en este caso?