Introducción a los métodos e investigación estadística

Isabel Maniega

Constructor

Definición:

1. Es cualquier cosa que es difícil de medir porque se puede definir y medir de muchas maneras diferentes.
2. Es la unidad de medida que estamos usando para el constructor. una vez que definimos algo operativamente, deja de ser una contructor.

Ejemplos:

a) El volumen es un constructor. Sabemos que el volumen es el espacio que ocupa algo, pero no hemos definido cómo estamos midiendo ese espacio, es decir en litros, galones, etc.

Si definimos el volumen en litros, entonces ya no sería un constructor porque estaría operacionalmente definido.

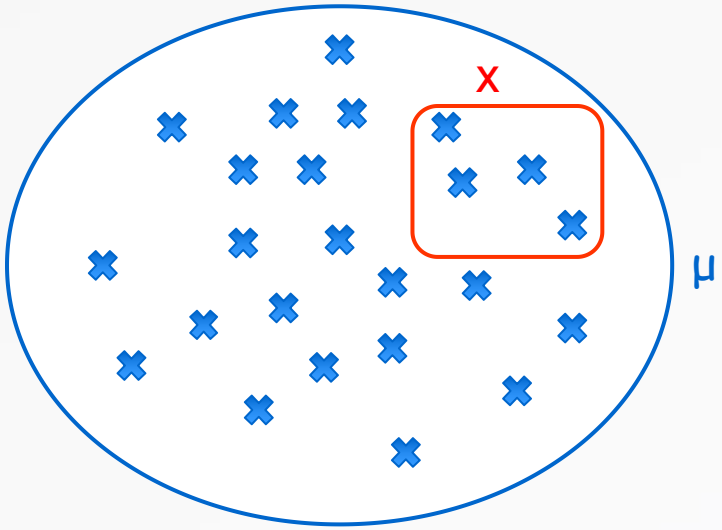
Población Vs Muestra

Definición:

1. ***Población***: es todos los individuos de un grupo.
2. ***Muestra:*** es alguno de los individuos de un grupo.
3. Parámetro VS estadística: ***parámetro*** es una característica de la población, mientras que ***estadística*** es una característica de la muestra.

Ejemplos:

a) La medida de una población se define con el símbolo μ, mientras que la medida de una muestra se define como x



Experimentación

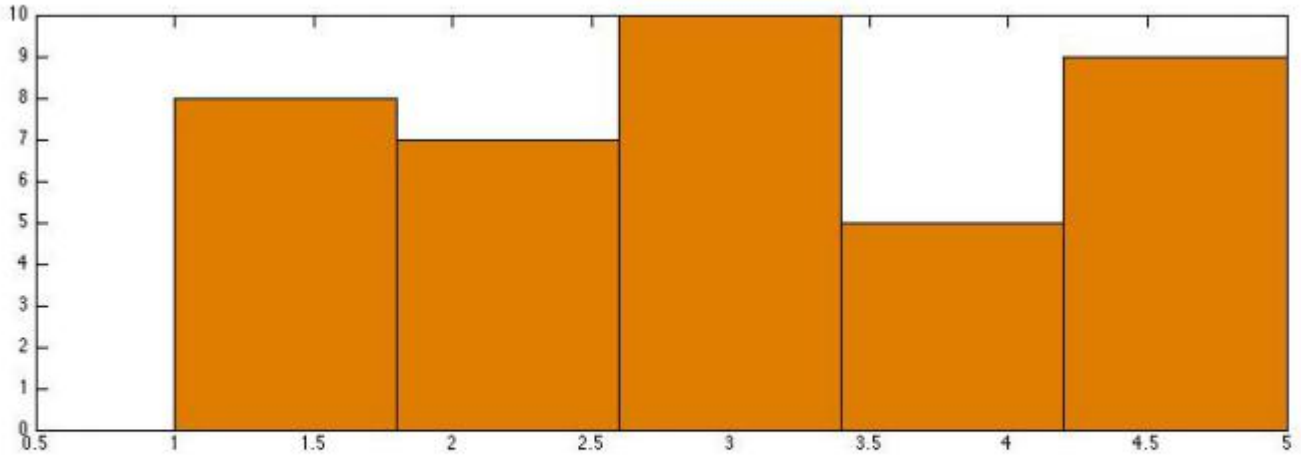
Definición:

1. ***Tratamiento***: en un experimento, la forma en que los investigadores manejan a los sujetos se denomina tratamiento. Los investigadores están específicamente interesados en cómo los diferentes tratamientos pueden producir resultados diferentes.
2. ***Estudio observacional:*** es cuando un experimentador observa a un grupo de sujetos y no introduce un tratamiento. Ej. las encuentas.
3. ***Variable independiente de un estudio:*** es la variable que los experimentadores eligen manipular; generalmente se traza a lo largo del eje x de un gráfico.
4. ***Variable dependiente de un estudio:*** es la variable que los experimentadores eligen medir durante un experimento; generalmente se traza a lo largo del eje y de un gráfico.
5. ***Grupo de tratamiento:*** el grupo de un estudio que recibe niveles de la variable independiente. Estos grupos se utilizan para medir el efecto de un tratamiento.
6. ***Grupo control:*** el grupo de un estudio que no recibe tratamiento. este grupo se utiliza como referencia cuando se comparan grupos tratamiento.
7. ***Placebo:*** Algo que se les da a los sujetos en el grupo de control para que piensen que están recibiendo el tratamiento, cuando en realidad están recibiendo algo que no les causa ningún efecto. (por ejemplo, una pastilla de azúcar)
8. ***Cegamiento:*** El cegamiento es una técnica utilizada para reducir el sesgo. El doble ciego garantiza que tanto los que administran los tratamientos como los que los reciben no saben quién recibe qué tratamiento.

Visualización

Definición:

1. ***Frecuencia***: La frecuencia de un conjunto de datos es el número de veces que ocurre un determinado resultado.



Este histograma muestra los puntajes en las pruebas de los estudiantes de 0 a 5. Podemos ver que ningún estudiante obtuvo una puntuación de 0, 8 estudiantes obtuvieron una puntuación de 1. Estos recuentos son lo que llamamos la frecuencia de las puntuaciones de los estudiantes.

Visualización

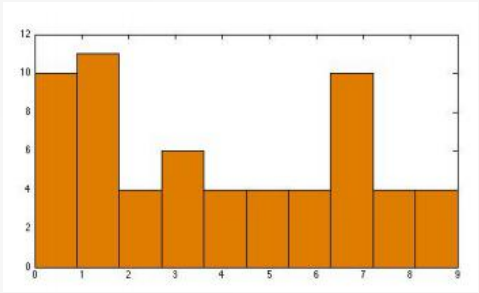
Definición:

1. ***Porporción***: Una proporción es la fracción de conteos sobre la muestra total. Una proporción se puede convertir en un porcentaje al multiplicar la proporción por 100.

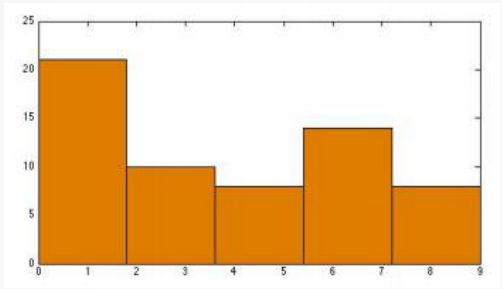
Ejemplo:

Usando nuestro histograma de arriba podemos ver la proporción de estudiantes que obtuvo un 1 en la prueba es igual a 8 de 39 alumnos es 0.2051 o 20.51%

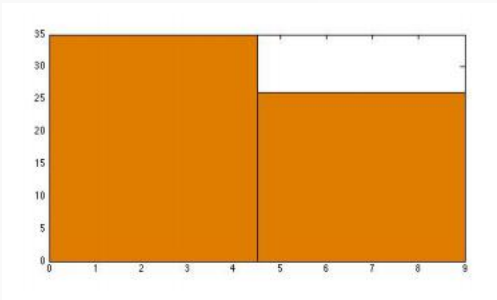
2. ***Histograma***: .es una representación gráfica de la distribución de datos, se deciden intervalos discretos (bins) para formar anchos para nuestras cajas.



Tamaño de contenedor 1



Tamaño de contenedor 2



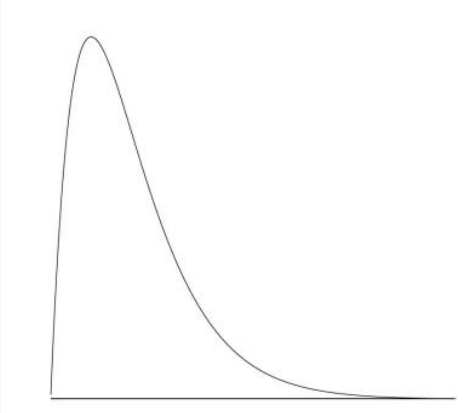
Tamaño de contenedor 5

Visualización

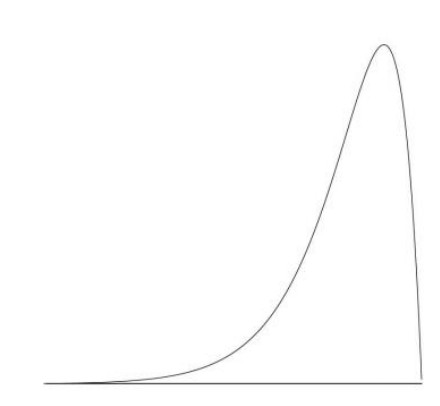
Definición Distribución sesgada:

1. ***Sesgo positivo***: es cuando hay valores atípicos a lo largo de la distribución en el extremo más a la derecha de la distribución.

2. ***Sesgo negativo***: Un sesgo negativo es cuando hay valores atípicos a lo largo del extremo izquierdo de la distribución.



Sesgo positivo



Sesgo negativo

Visualización

Ejemplos de como realizarlo Python (26/27):

Histograma

Scatter plot (nube de puntos (2D))

Pie chart (Diagrama de sectores)

Gráficos en 3 dimensiones (3D)

Tendencia central

Definición:

1. ***Media***: La media de un conjunto de datos es el promedio numérico y se puede calcular dividiendo la suma de todos los puntos de datos por el número de puntos de datos.

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$

⚠ La media se ve muy afectada por los valores atípicos, por lo que decimos que la media NO es una medida robusta.

2. ***Mediana***: .La mediana de un conjunto de datos es el punto de datos que está directamente en el medio del conjunto de datos. Si dos números están en el medio, la mediana es el promedio de los dos.

   a) Si n es impar, la mediana es el valor que ocupa la posición (n+1)/2 una vez que los datos han sido ordenados (en orden creciente o decreciente), porque este es el valor central.
   b) Si n es par, la mediana es la media aritmética de los dos valores centrales: $\frac{x_k + x_{k+1}}{n}$

⚠ La mediana es robusta a los valores atípicos, por lo tanto, un valor atípico no afectará el valor de la mediana.

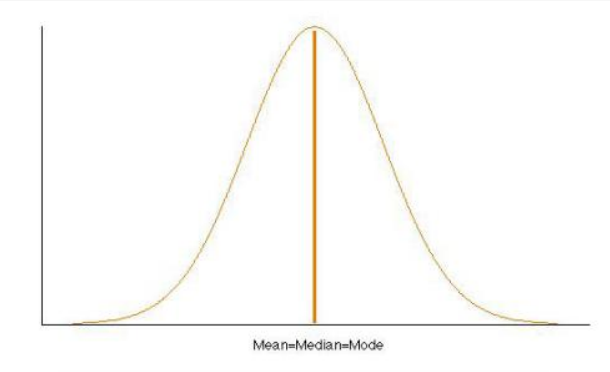
Tendencia central

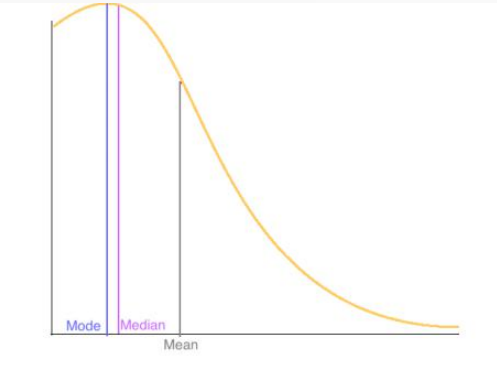
Definición:

3. ***Moda***: La moda de un conjunto de datos es el punto de datos que ocurre con más frecuencia en el conjunto de datos.

⚠ La moda también es robusta para los valores atípicos.

⚠ En la distribución normal la media = mediana = moda.





Ejemplos de como realizarlo Python (28)

Variabilidad

Definición:

2. ***Encontrar valores atípicos (Finding Outliers):*** *Puede utilizar el IQR para identificar valores atípico*
   a) *Valores atípicos superiores:* $Q3 + 1{,}5 \cdot IQR$
   b) *Valores atípicos más bajos:* $Q1 - 1{,}5 \cdot IQR$

3. ***Varianza:*** *La varianza es el promedio de las diferencias al cuadrado de la media. La fórmula para calcular la varianza es:*

$$\sigma^2 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}$$

4. ***Desviación estándar:*** La desviación estándar es la raíz cuadrada de la varianza y se usa para medir la distancia desde la media.

⚠ En una distribución normal, el 65 % de los datos se encuentra dentro de 1 desviación estándar de la media, el 95 % dentro de 2 desviaciones estándar y el 99,7 % dentro de 3 desviaciones estándar

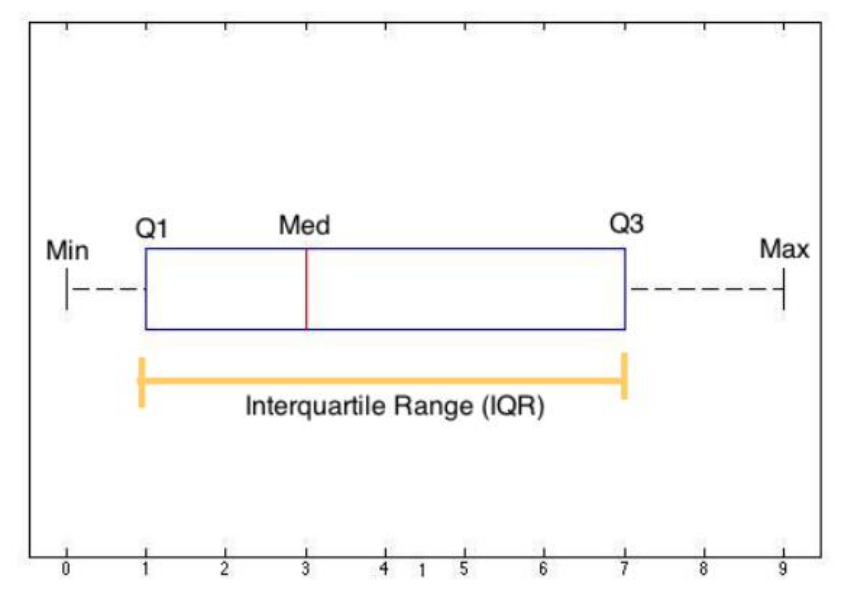
Variabilidad

Diagramas de caja y el IQR

Un diagrama de caja es una excelente manera de mostrar el resumen de 5 números de un conjunto de datos de una manera visualmente atractiva. El resumen de 5 números consiste en el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo.

Definición:

1. ***Rango intercuartílico:*** El rango intercuartil (IQR) es la distancia entre el primer *cuartil y el tercer cuartil y nos da el rango del 50% medio de nuestros datos. El IQR se encuentra fácilmente calculando:* $Q3 - Q1$



Variabilidad

Definición:

2. ***Corrector de Bessel:*** *Corrige el sesgo en la estimación de la varianza de la población y parte (pero no todo) del sesgo en la estimación de la desviación estándar de la población. Para aplicar la corrección de Bessel multiplicamos la varianza por* $\frac{n}{n-1}$

Utilice la corrección de Bessel principalmente para estimar la desviación estándar de la población

Ejemplos de como realizarlo Python (29)

Estandarización

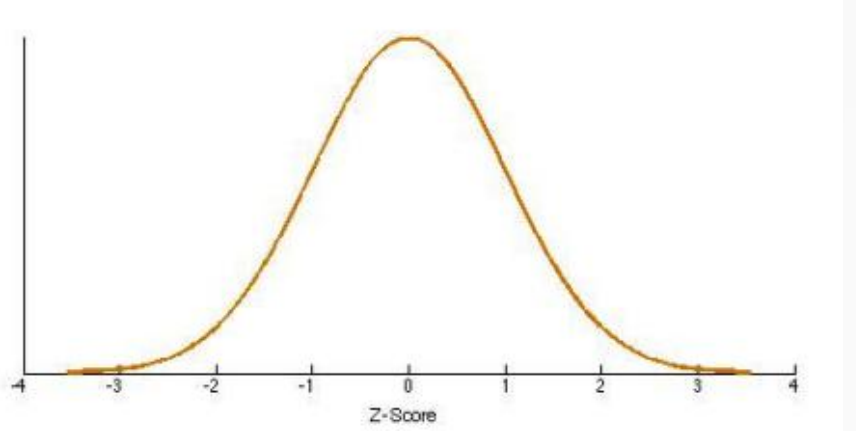
Z score

Dado un valor observado X, la Z score encuentra el número de desviaciones estándar X está lejos de la media.

$$Z = \frac{x - \mu}{\sigma}$$

Curva normal estándar

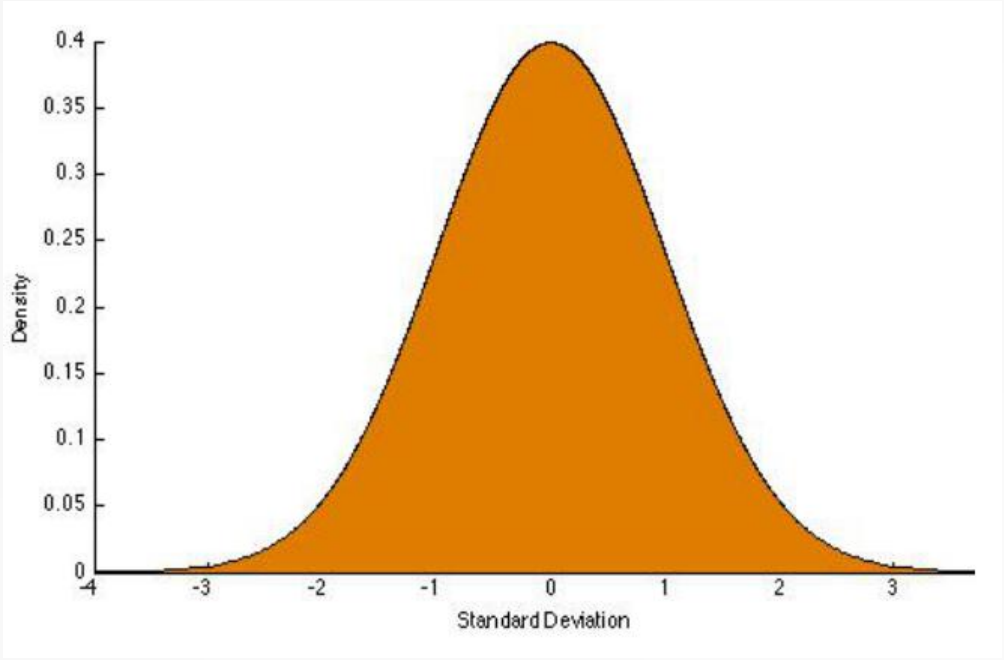
La curva normal estándar es la curva que usaremos para la mayoría de los problemas en esta sección. Esta curva es la distribución resultante que obtenemos cuando estandarizamos nuestros puntajes. Usaremos esta distribución junto con la tabla Z para calcular los porcentajes por encima, por debajo o entre observaciones en secciones posteriores.



Distribución normal

Función de distribución de probabilidad

La función de distribución de probabilidad es una curva normal con un área de 1 debajo, para representar la frecuencia acumulada de los valores.



Ejemplos de como realizarlo Python (31)

Distribuciones de muestreo

Teorema del límite central

El teorema del límite central se usa para ayudarnos a comprender los siguientes hechos, independientemente de si la distribución de la población es normal o no:

1. la media de las medias muestrales es igual a la media poblacional.

2. la desviación estándar de las medias muestrales siempre es igual al error estándar (es decir, SE = $\frac{\sigma}{\sqrt{n}}$)

3. la distribución de las medias muestrales será cada vez más normal a medida que aumente el tamaño de la muestra, n, aumenta.

Distribuciones de muestreo

Distribución de muestreo

La distribución muestral de una estadística es la distribución de esa estadística. Puede considerarse como la distribución del estadístico para todas las muestras posibles de una misma población de un tamaño dado

Ejemplos de como realizarlo Python (32)