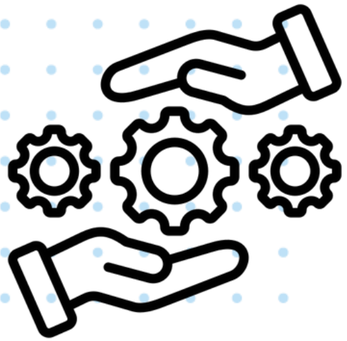
COLOMBIA POTENCIA DE LA VIDA

TIC

# Lección 1
# Análisis estadístico de los datos



TALENTO TECH | AZ PROYECTOS EDUCATIVOS | UTP Universidad Tecnológica de Pereira

Tiempo de ejecución: 3 horas

| PLANTEAMIENTO DE LA SESIÓN | MATERIALES |
| --- | --- |
| El análisis a través de las técnicas estadísticas permite modelar y entender el comportamiento de una variable aleatoria. Usualmente en un conjunto de datos, cada una de las características o columnas se pueden entender empleando el concepto de variable aleatoria para determinar qué comportamiento está modelado por los datos y saber las mejores técnicas matemáticas y de ciencia de datos que se pueden aplicar tanto para el tratamiento de los datos como para el análisis. | |

Cualquier proceso de adquisición de datos requiere de un elemento que mida los datos, un elemento que procese los datos medidos y un elemento que permita a un observador ver los datos que se están midiendo. En general, estos procesos se incorporan en instrumentos de medidas, los cuales, permiten censar información y visualizarla de alguna forma entendible.

Si bien este proceso se crea de tal forma que la precisión sea la mejor posible bajo algunas restricciones. Siempre habrá errores. Es decir, cualquier proceso que implique la observación de un fenómeno directamente altera el fenómeno y hace que las mediciones vengan contaminadas con variaciones a las que se les conoce como ruido.

Existen diferentes tipos de errores que alteran las medidas, como por ejemplo el error aleatorio, que no se puede eliminar e implica que todo dato medido u observado viene con cierto grado de distorsión. Al valor no distorsionado se le denomina valor verdadero. Por mas preciso que sea un proceso de adquisición de datos siempre existirá error causando que el valor verdadero no pueda ser medido. Por eso se requieren técnicas de estimación que ayuden a modelar cual es el valor más probable dadas algunas observaciones de la variable de interés y en qué rango de confianza se encuentra el valor verdadero.

También existen fuentes de error eliminables, como es el caso de errores de equipos, de instrumentos de medidas, de ajuste, de aproximación, entre muchos otros corregibles.

El error aleatorio se debe analizar con la repetición de las medidas. Para tal fin se aplican métodos estadísticos que nos permitan conocer el valor más probable.

Al medir repetidas veces un evento, se puede conocer la distribución de probabilidad y modelar el ruido y el comportamiento de una variable aleatoria.

Es común que las distribuciones de probabilidad tengan algunos datos con más repeticiones que los demás. Estos comportamientos se describen mediante una función que nos ayuda a conocer cuáles son las características de esa variable aleatoria. Para establecer dicha función es necesario conocer la posición y la dispersión de la variable.

Conociendo la distribución de frecuencias podemos tener una idea del comportamiento para modelarlo con la medida de posición o de tendencia central y la medida de dispersión. De esta forma es posible comparar y operar variables aleatorias para realizar cálculos, propagación de las distribuciones, y validación de experimentos.

En el caso de las variables cuantitativas se tienen las siguientes medidas de tendencia central:

**Moda**: Es el valor que tiene la máxima frecuencia (preguntar la moda de la edad en el titanic.). En ocasiones la moda no difiere de otras medidas, es decir, los valores de varias frecuencias son iguales o similares. En este caso se puede decir que la cariable puede ser bimodal (dos modas) o multimodal (varias modas). Cuando todos los valores tienen la misma frecuencia no existe la moda (tal es el caso de una distribución uniforme).
Cuando la escala de una variable se divide en clase, se puede definir la clase modal, como aquella que reúne la máxima frecuencia.

**Cuantiles:** Un cuantil es un valor de una variable a la cual corresponde una determinada frecuencia relativa acumulada.

El cuantil α de una distribución de frecuencias es un valor de la variable al cual corresponde la frecuencia relativa acumulada fra = α. Por ejemplo, el cuantil 0,15 es un valor de la variable al cual corresponde la frecuencia relativa acumulada fra = 0,15. Es común referirse a percentiles que no son otra cosa que los cuantiles identificados por el valor de α expresado en porcentaje. Por ejemplo, en lugar de cuantil 0,15 podemos decir percentil 15. Los cuantiles 0,25, 0,50 y 0,75 se denominan respectivamente primer cuartil, segundo cuartil o mediana y tercer cuartil. La mediana es el valor que corresponde a la mitad de la distribución de frecuencias. Por eso decimos que la mediana es una medida de posición central.

Los gráficos de cajas o de bigotes (box plots) resumen la distribución de frecuencias a partir de unos cuantiles, como se muestra en la figura 1. En estos gráficos, los bordes de la caja indican el primer y tercer cuartil, la línea horizontal que corta la caja indica la mediana y los extremos muestran el valor mínimo y máximo de la variable. Los valores de las medidas de posición se aprecian en el eje vertical de la caja.

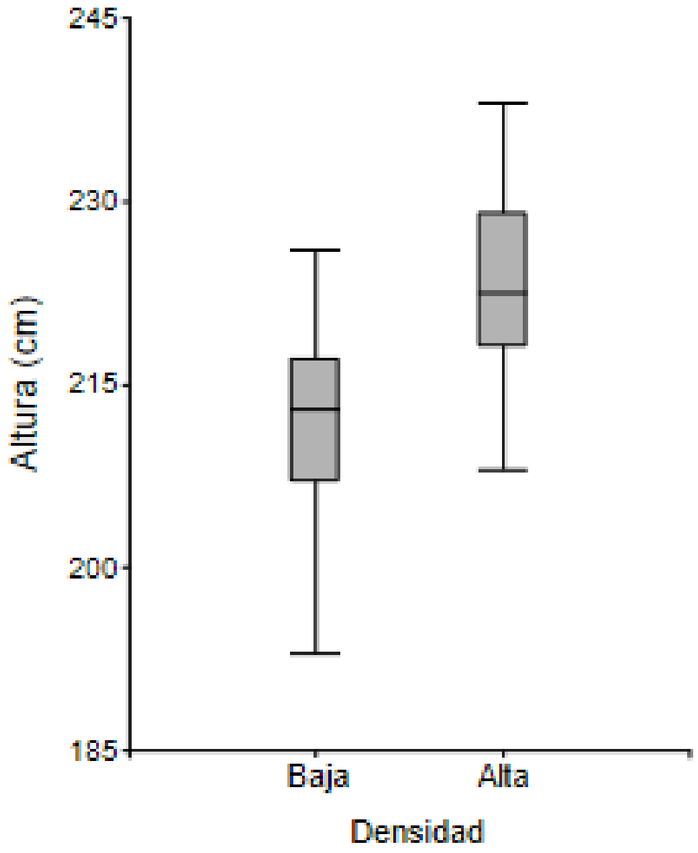



**Figura 1. Ejemplo de diagrama de cajas o de bigotes.**

**Media aritmética**

La media aritmética (promedio) es una de las medidas de tendencia central más conocidas de la estadística. Esta se calcula como el cociente entre la suma de los valores de la variable aleatoria y el número de mediciones. Generalmente se denota como una variable con una barra superior en la escritura como se muestra en la ecuación 1.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{N} x_i$$

**Ecuación 1: cálculo de la media aritmética.**

La media aritmética brinda información sobre la posición central de una distribución de frecuencias. Sin embargo, la media no nos informa acerca de cómo se distribuyen los datos. Con lo que no es posible aun comparar muchas distribuciones de frecuencia solo con la media aritmética. Por eso se necesitan conocer algunas medidas de dispersión (se verán en la siguiente lección). Si se conoce una medida de tendencia central y una medida de dispersión, es posible comparar variables aleatorias y modelarlas dados estos dos comportamientos.

Si bien las medidas descritas (de tendencia central) ayudan a conocer una parte de cómo están distribuidos los datos, también es necesario entender qué tan dispersos están los datos alrededor de las medidas de tendencia central. Es allí donde toman utilidad las medidas de dispersión.

Las medidas de dispersión son herramientas estadísticas que complementan las medidas de tendencia central al proporcionar información sobre la variabilidad o dispersión de un conjunto de datos. Mientras que las medidas de tendencia central como la media, la mediana y la moda nos dan una idea de dónde se centran los datos, las medidas de dispersión nos indican cuán dispersos están las observaciones alrededor de este centro.

Las medidas de dispersión más comunes son la varianza y la desviación estándar, que son especialmente útiles para describir la dispersión de los datos en torno a la media. La varianza se calcula como la media de los cuadrados de las desviaciones de cada observación respecto a la media. Como se muestra en la ecuación 1. La desviación estándar es la raíz cuadrada positiva de la varianza (ecuación 2). Estas medidas nos dan una idea de qué tan "juntos" o "dispersos" están los datos alrededor de la media.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right) \; ecuación\ 1: varianza$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}\left(x_i - \bar{x}\right)} \quad Ecuación\ 2: Desviación\ estándar$$

En general en la literatura se representa a la desviación estándar con la letra griega minúscula sigma (σ) y a la varianza como sigma al cuadrado. Otra medida de dispersión es el rango, se calcula como la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Aunque el rango es fácil de calcular, puede no ser tan robusto como la varianza o la desviación estándar, ya que solo se basa en dos observaciones extremas. Si las observaciones corresponden a casos atípicos, podría ser sesgada la visibilidad de la dispersión, es decir, se podría pensar que los datos están más dispersos de lo que realmente están.

Rango inter-cuartil: Consiste en la diferencia entre el tercer y el primer cuartil de una distribución de probabilidad. A diferencia del rango, presenta una estabilidad numérica mayor.

El rango inter-cuartil se define como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1). Es decir: Rq = Q3 – Q1.

A la mitad del rango intercuartil se le llama desviación cuartil y permite conocer la dispersión en distribuciones sesgadas. También se usa para dibujar los diagramas de cajas de tales distribuciones.

Conociendo las medidas de tendencia central de los datos y la dispersión, se puede asignar una distribución de probabilidad a los datos. Dentro de las funciones más comúnes se encuentran:

Su importancia radica en que gran cantidad de fenómenos naturales y artificiales se ajustan a este patrón. Por ejemplo, la altura de las personas, los puntajes en pruebas estandarizadas, el error en cualquier medición y muchas variables asociadas con fenómenos naturales. Por esto es muy conocida y se suelen explicar muchos conceptos estadísticos basados en las características de una distribución gaussiana.

Esta distribución se puede modelar con dos parámetros que son la media y la desviación estándar. De la distribución gaussiana se deriva la regla empírica, que significa que aproximadamente el 68% de los datos se encuentran contenidos dentro de una desviación estándar del valor medio y el 95% de los datos se encuentra en dos desviaciones estándar. También implica que en tres desviaciones estándar alrededor de la media se encuentra el 99.7% de los datos.

La función que modela la forma de una distribución gaussiana se modela con la ecuación:

$$f(x) = a \times e^{-(x-b)^2/2c^2}$$

En esta ecuación a, b y c son constantes reales mayores que -1. El valor a simboliza el punto más alto de la campana, b es la posición del centro de la campana y c es la deviación estándar que modela el ancho de la campana. Usualmente al valor de a se le asigna:

$$a = \frac{1}{c\sqrt{2\pi}}$$

También se suele encontrar en la literatura que los valores de b se le llama también con la letra griega μ (valor de la media) y a c se le asigna la letra griega  (valor de la desviación estándar). En la figura 2 se muestran algunas curvas para diferentes valores de media y de desviación estándar.

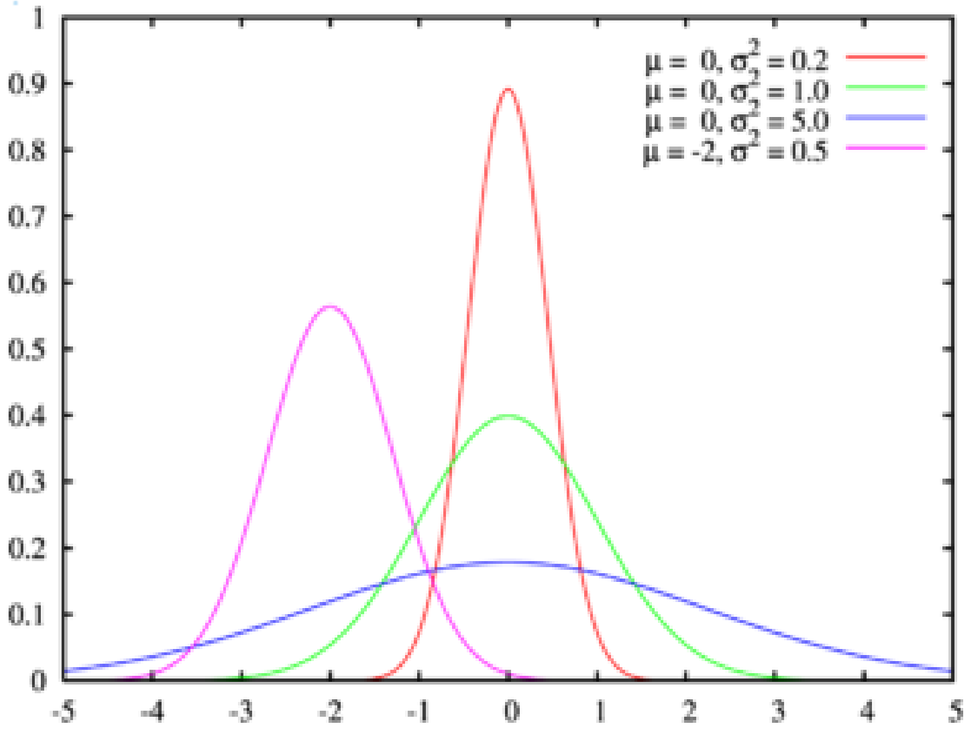



**Figura 2: Curvas gaussianas**

Distribución rectangular

Es una familia de distribuciones de probabilidad en donde todos los valores de un rango determinado tienen una probabilidad uniforme de aparición. El dominio de estas funciones está definido por los parámetros a y b que son su valor mínimo y su máximo respectivamente. La función de densidad está definida por:

$$f(x) = \frac{1}{b-a}; para\ x \in [a,\ b]$$

Es importante resaltar que solo está definida para el intervalo entre los valores a y b (figura 3).

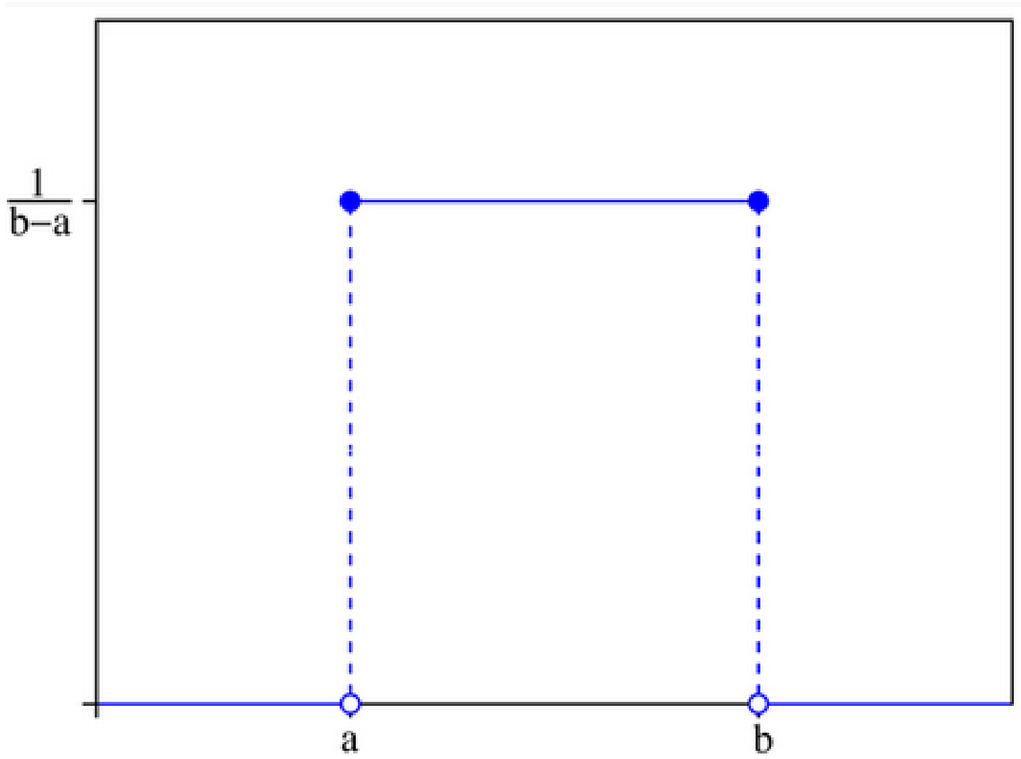



**Figura 3: representación de la función uniforme de probabilidad**

Para este tipo de distribuciones la esperanza (valor medio de la variable aleatoria) está dado por:

$$E[x] = \frac{a+b}{2}$$

Y la varianza está dada por:

$$\frac{(b-a)^2}{12}$$

## Distribución Pearson ($x^2$)

La distribución de Pearson o ($x^2$) chi cuadrado es la suma del cuadrado de k variables aleatorias independientes con distribución normal estándar. Esta función se emplea en inferencia estadística para las pruebas de hipótesis y la construcción de los intervalos de confianza. En la figura 4 se muestra la densidad de probabilidad de esta función. La media está dada por el valor k y la varianza es 2k.

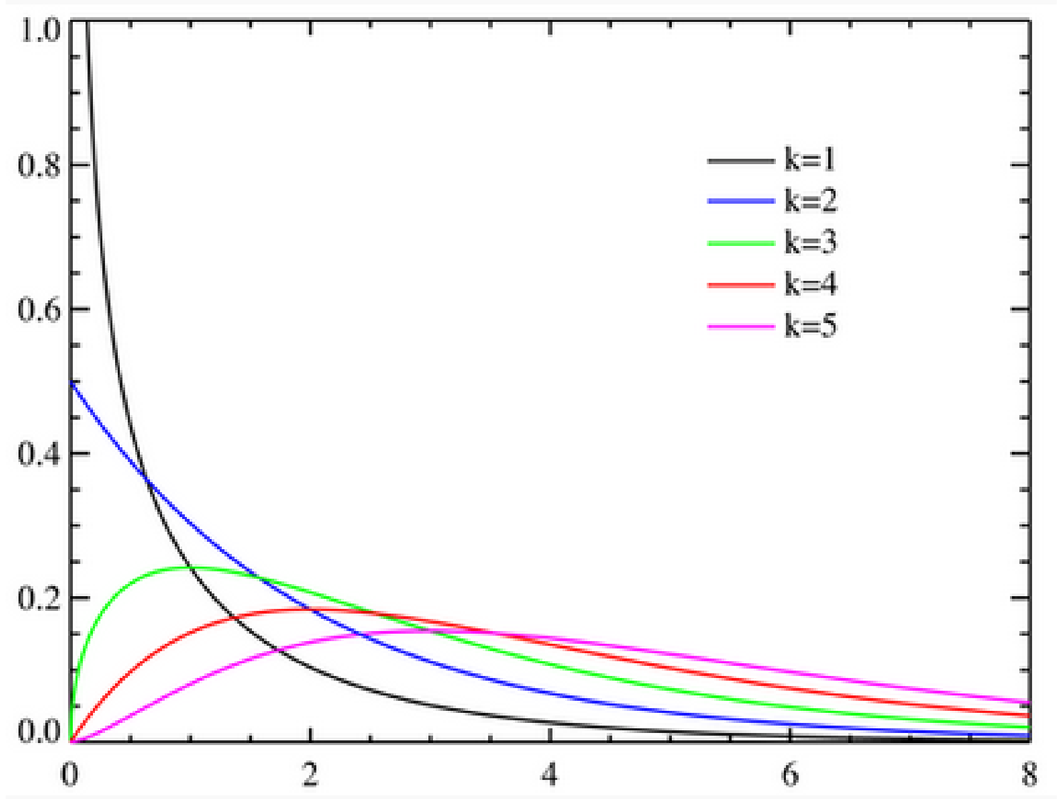



**Figura 4: Densidad de probabilidad $x^2$**

## Distribución T Student

Representa la distribución de probabilidad cuando se quiere estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño y la desviación estándar poblacional es desconocida. Su media es cero, y su varianza es:

$$\frac{v}{v-2}$$

Para v > 2 y es indefinida para otros valores.

En la figura 5 se muestra la curva de la densidad de probabilidad

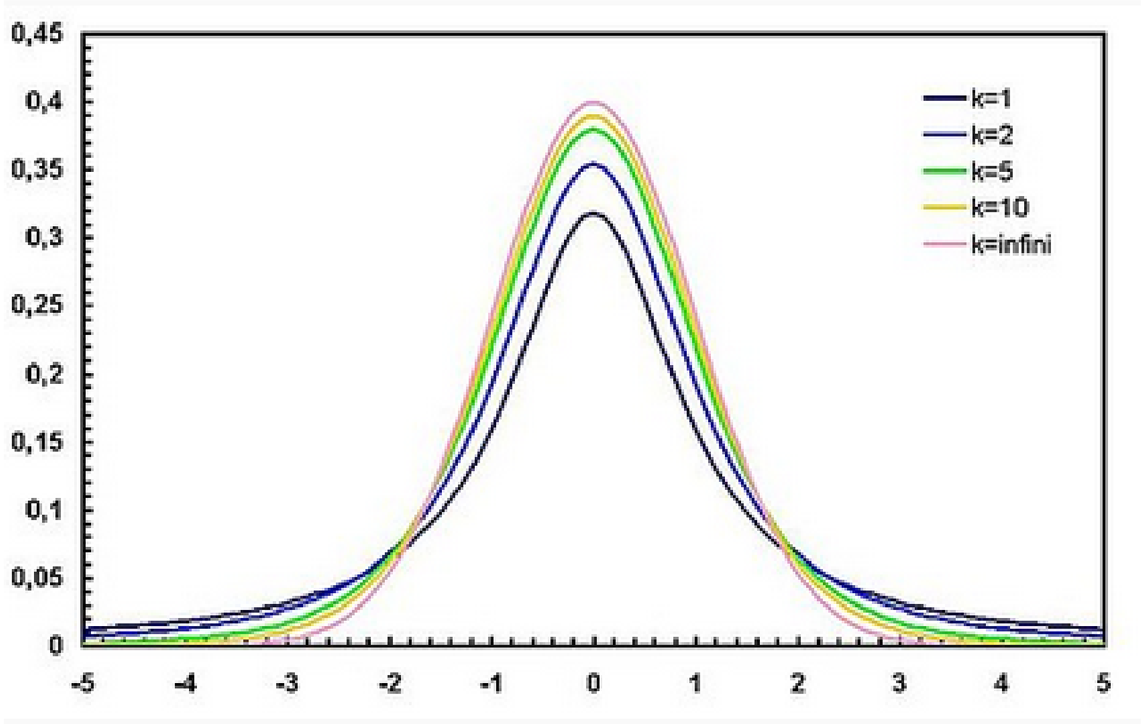



**Figura 5: densidad de probabilidad T student.**

Es común que en cualquier análisis de datos se suela iniciar con el conteo de cuantas variables existen (generalmente se ubican en las columnas de un conjunto de datos), luego se determina de qué tipo de variables aleatorias se trata cada columna (cualitativa o cuantitativa) y para las variables cuantitativas se suelen extraer las medidas de tendencia central y las medidas de dispersión, modelando la distribución de probabilidad de cada variable aleatoria.

La información de la distribución de probabilidad de las variables aleatorias ayuda a aplicar las técnicas de procesamiento de datos correctas, ya que no todas las variables se pueden tratar de la misma forma. Tampoco se puede considerar que absolutamente todos los eventos medidos tienen una distribución gaussiana y se pueden modelar tan solo con la media y la varianza. Siempre es bueno crear un histograma de la información y revisar qué tipo de técnicas de análisis y de tratamiento de los datos aplican para esa distribución de probabilidad en específico.