ESTADÍSTICA CON EXCEL

Contenido

La información aquí obtenida se conseguió del diplomado **Estadística con Excel** por parte de SciData.

INTRODUCCION A LA ESTADISTICA

- **Bases de la Estadística:**

  Antes que otra cosa, veamos algunas nociones conceptuales sobre la Estadística. A saber:

- **¿Qué es la Estadística?**

  La Estadística es una rama de la Matemática, tal como la Geometría o el Álgebra. Sin embargo, a diferencia de la gran mayoría de las matemáticas, está enfocada en los datos numéricos.

  Los procesos estadísticos comienzan con el **levantamiento de los datos;** continúan con la **transformación de los datos recolectados** para tenerlos de manera manejable y tales que **podamos analizarlos** para que, finalmente, a través de los resultados podamos **tomar decisiones.**

- **¿Cómo se aplica la Estadística a los negocios e investigaciones?**
- Cada vez que aprendemos algo nuevo nos surge la misma pregunta: ¿Esto para qué me puede ser útil? Describe los datos, analiza y presenta los datos, realiza predicciones, mejora procesos.

- **Tipos de Datos:**

  a) **Según su tamaño:**
     - Datos poblacionales: todo el universo de datos
     - Datos muestrales**:** Una pequeña muestra de ese gran universo de datos. Poner criterios que deben cumplir las muestras

  Las encuestas son cuando ya les metes un factor probabilístico. Las encuestas de Facebook NO SON ENCUESTAS, son sondeos.

  b) **Según lo que miden:**
     - **Datos Categóricos:** Sirven para clasificar. Ej. Número de tu INE, número de tu matricula, "¿Te gusta ir a la escuela?", géneros. Sirve para hacer conteos. Se usan en las Regresiones Logísticas
     - **Datos Numéricos:** Sirven para hacer medidas. Se usan en las Regresiones Lineales.

     Para diferenciarlos, hay que fijarnos si la suma tiene sentido. Por ejemplo, ¿Tiene sentido que se sumen 2 matriculas? No, es categórico. ¿Tiene sentido sumar años? Sí, es numérico

  c) **Por tipo de mediciones:**
     - **Cualitativos (Categóricos):**
       i. Nominales: Sirven para clasificar. Un ejemplo es las matrículas de empleados o estudiantes.
       ii. Ordinales: Sirven para clasificar, pero tienen un orden. Un ejemplo son las encuestas de tipo ¿Cuánto te gusta el refresco? Mucho, poco, nada. Un ejemplo es la Escala de Likert: Escalas pares de preferencia, para evitar sesgos centrales

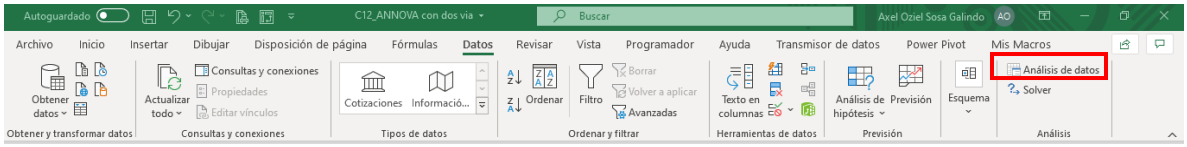
- **Cuantitativos:**
    i. Intervalos: Las restas tienen sentido (la suma no tiene sentido). El ejemplo clásico son las temperaturas en C°. No tienen un cero absoluto (significa ausencia) (En ventas, un 0 representa que no hubo ventas, un 0 absoluto)
    ii. Ratio: Sirven para medir el orden y el valor exacto. Tienen un cero absoluto. Algunos ejemplos son la altura, el peso, la duración. Aquí sí tienen sentido las proporciones.

- **Proceso del Análisis Estadístico**
    1. Definir el objetivo
    2. Recolectar y clasificar los datos
    3. Definir un método de estudio
    4. Analizar los resultados
    5. Reportar y presentar las conclusiones

- **Análisis de Datos:**
- Para configurar la zona de Análisis de datos, sigue esta ruta

    Archivo -> Opciones -> Complementos -> Herramientas para análisis

    Selecciona esta opción y da click en **Aceptar.** De inmediato, en la zona de Datos tendrás activada una subzona de Análisis de Datos como en la siguiente imagen.



- Si el paso anterior no funcionó, una alternativa es realizar la misma ruta:

    Archivo -> Opciones -> Complementos -> Administrar

    Selecciona ahora la opción **Complementos de Excel** y da click en **Ir.** En el cuadro de diálogo que se abrirá, selecciona la casilla de **Herramientas para análisis.** Da click en **Aceptar.**

- **Conocimientos necesarios de Excel:**
    - Funciones Básicas: SUMAR y SUMAR. SI, CONTAR y CONTAR.SI, MIN y MAX
    - Gráficos:
        - Gráficos de líneas: Suelen usarse para visualizar cambios de los datos en el tiempo. Pueden identificar tendencias para realizar predicciones.
        - Gráficos de pastel: Suelen usarse para visualizar proporciones de los datos. Pueden identificar categorías dónde enfocar la atención.
        - Nubes de Puntos: Suelen usarse para establecer relaciones entre dos medidas diferentes. Pueden identificar valores atípicos
        - Histogramas: indican la frecuencia de un hecho mediante una distribución de los datos.

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/nxn3trpWivo

MEDIDAS DE TENDENCIA CENTRAL

- En Estadística, todos los datos están distribuidos a través de varios puntos, que es lo que se conoce como la distribución de los datos.
- A partir de la distribución de los datos es muy complicado llegar a conclusiones, pues muchas veces corresponden a distribuciones demasiado complejas. Sin embargo, muchas veces existe una tendencia de los datos a juntarse alrededor de un cierto valor. (por ejemplo, en la distribución gaussiana, tienden a juntarse en el centro (los datos se quieren parecer a un cierto dato))
  (en el eje x hay datos, y en el eje Y hay frecuencias)
- Este valor es lo que se conoce como una tendencia central. De esta manera, la tendencia central es un representante global de los datos. En particular, representa las características generales del conjunto de datos.
- Las medidas de tendencia central se clasifican en tres formas: valores promedios, valores de partición y valores repetidos.

- **Medidas de Promedio**:
  - **a) Media Aritmética:**
    - Valor promedio de las observaciones.
    - Se trata de **sumar** todos los valores y dividir entre el numero de obs.
    - La suma de todas las desviaciones de cada valor respecto de su media aritmética siempre es 0.
      ($dato_1$, $dato_2$, …, $dato_n$ -> ($dato_1$-promedio) + ($dato_2$-promedio) ($dato_n$-promedio) = 0)
    - Se ve afectada por cambios en las escalas, así como si se suma un valor constante a cada uno de los datos:
      constante*$dato_1$, c*$dato_2$, …, c*$dato_n$ -> c * promedio
      esta observación sirve por sí es necesario convertir a alguna otra unidad de medida el promedio
    - En Excel, la media aritmética se calcula con la función **PROMEDIO**
    - **Descripción Matemática**
      - Datos No Agrupados: Viene dada por $\frac{1}{n}\sum x_i$ donde x_i es el valor de la observación i-esima y n es el número de observaciones
      - Datos Agrupados: $\frac{1}{n}\sum f_i x_i = \frac{\sum f_i x_i}{\sum f_i}$ donde fi es la frecuencia en que el valor del grupo i, xi, se repite.

  - **b) Media Geométrica:**
    - Se trata de multiplicar todos los valores y calcular a ese productor la raíz n-esima donde n es la cantidad de observaciones.
    - Suele aplicarse en Finanzas ya que está relacionada con la tasa de crecimiento anual compuesto
    - La media geométrica es más pequeña que la media aritmética (demostración con el Teorema de Pitágoras)

- La Media Geométrica es equivalente a una media aritmética, sólo que con logaritmos. El logaritmo de la media geométrica es la aritmética de los logaritmos de los datos
- En Excel, se usa la función **MEDIA.GEOM**
- **Descripción Matemática:**
  - Datos No Agrupados: $\sqrt[n]{x_1 * x_2 * \dots * x_n}$
  - Datos Agrupados: $\sqrt[n]{{x_1}^{f1} * {x_2}^{f2} * \dots * {x_n}^{fn}}$

c) **Media Armónica**:

- Se trata de **dividir** el total de elementos entre la suma de los inversos multiplicativos de cada dato.
- Se suele utilizar cuando los datos son de tipo fracción (proporciones) (se calcularon a través de razones (velocidad promedio, goles por partido))
- La media siempre es más pequeña que la media geométrica
- En Excel, se usa la función **MEDIA.ARMO**
- **Descripción Matemática:**
  - Datos No Agrupados: $\frac{n}{\sum \frac{f_i}{x_i}}$
  - Datos Agrupados: $\frac{N}{\sum \frac{f_i}{x_i}}$

Para más información, echa un vistazo a la siguiente lectura complementaria:
https://towardsdatascience.com/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0

- **Medidas de partición**:
- Dividen la colección de datos en varios subconjuntos.
- Hay 4 medidas de partición:

a) Mediana:

- Se utiliza para dividir las observaciones, en el 50% más alto y el 50% más bajo. Se le considera también una medida de posición
- Se debe ordenar de menor a mayor
- Importante en los llamados gráficos de bigote
- En Excel, utilizamos la función **MEDIANA**
- **Descripción Matemática:**
  - Si n es impar, se tiene mediana = $\frac{x_{n+1}}{2}$
  - Si n es par, se tiene mediana = $\frac{\frac{x_n}{2} \; y \; \frac{x_n}{2}+1}{2}$

b) Cuartil:

- Se utiliza para dividir las observaciones en cuatro partes iguales. Por lo tanto, existen tres ellos: Q_1, Q_2 y Q_3, donde Q_1 es la partición hasta el 25% de los datos; Q_2 es la partición hasta el 50% de los datos, y por lo tanto Q_2 = mediana; y Q_3 es la partición hasta el 75% de los datos.
- En Excel utilizamos las funciones **CUARTIL.INC** y **CUARTIL.EXC**. La diferencia es que INC incluye los extremos, y EXC no.
- **Descripción Matemática:** Si tenemos los valores x_1, x_2, …, x_n entonces

$$Q_1 \sim x_{(n+1)/4},\ Q_2 \sim x_{(n+1)/2},\ Q_3 \sim 3(n+1)/4$$

c) Percentiles:

- Se utiliza para dividir las observaciones en cien partes iguales. Por lo tanto, existen 99 de ellos. P_1, P_2, …, P_{99}, donde P_1 es la partición hasta el 1% de los datos; P_2 es la partición hasta el 2% de los datos, …, y P_{99} es la partición hasta el 99% de los datos.
- En Excel utilizamos las funciones **PERCENTIL.INC** y **PERCENTIL.EXC** para calcular los cuartiles. La sintaxis es idéntica a las de CUARTIL
- **Descripción Matemática:** $P_i \sim (n+1) * \frac{i}{100}$

d) Deciles:

- Se utiliza para dividir las observaciones en 10 partes iguales. Por lo tanto existen 9 de ellos: D_1, D_2, …, D_9, donde D_1 es la partición hasta el 10% de los datos; D_2 es la partición hasta el 20% de los datos, …, y D_9 es la partición hasta el 90% de los datos.
- No hay una función para deciles en Excel; sin embargo, utilizamos las funciones **PERCENTIL.INC** y **PERCENTIL.EXC** como Di = PERCENTIL.EXC(rango,10*i).
- **Descripción Matemática:** $D_i \sim (n+1) * \frac{i}{10}$

Revisar el siguiente video a partir del minuto 33:47:
https://youtu.be/7i6prOTvQis

- **Medidas de Repetición:**

a) Moda:

- Esta se refiere al valor que más veces se repite. En la práctica es la menos usada de todas las medidas de tendencia central, ya que no provee suficiente claridad acerca de las características de un conjunto de datos, aunque suele usarse para los valores nominales.
- En Excel utilizamos la función **MODA.UNO** en el caso en que hay valores repetidos. Su sintaxis es igual a la de la función SUMA. La función **MODA.VARIOS** devuelve varias modas si es que existe más de una.

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/qS4BqqOrLI8

MEDIDAS DE DISPERSION

- Las dispersiones o desviaciones son una manera de medir qué tan alejados están los puntos de alguna medida de tendencia central.
- Se pueden calcular para cualquiera de las medidas de tendencia central.
- Son medidas para conocer que tan dispersos se encuentran nuestros datos de algún valor central. Es decir, si hay una desviación alta, entonces nuestros datos se alejan mucho de la medida de tendencia central y por lo tanto está última no es un buen representante de nuestros datos.
- La medida de tendencia central es el representante de los datos, y la medida de dispersión es qué tan buen representante es la medida de tendencia central.

- Existen dos tipos de medida de dispersión: absolutas y relativas:
  - **a) Absoluta:**
    - Rango:
      - Se define como la diferencia entre el valor más alto y el más pequeño. Por ende, se ve fuertemente afectado por los valores atípicos.
      - Nos indica cuando mide la distancia entre los datos. El tamaño del conjunto que abarcan los datos.
      - El rango sí se ve afectado drásticamente con los outliers
      - No se trata de la mejor medida de dispersión
      - Su cálculo no involucra ninguna medida de tendencia central
      - En Excel se involucra las funciones **MAX** Y **MIN**
      - **Descripción Matemática:** Como la misma definición de rango lo indica, simplemente tenemos

$$\text{Rango} = \text{val}_{max} - \text{val}_{min}$$

    - Desviación Media Absoluta
      - Considera directamente todos los valores de los datos uno por uno. Se trata de medir como se aleja, en promedio, la "distancia" entre cada dato y alguna medida de tendencia central.
      - Generalmente se toma como la media aritmética o la mediana.
      - En Excel, cuando la medida de tendencia central es la media aritmética, se utiliza la función **DESVPROM**
      - **Descripción Matemática:** Supongamos que tenemos los n datos x1, x2, …, xn (no necesariamente ordenados). Sea A alguna medida de tendencia central. Entonces la desviación media absoluta se calcula como $MD = \frac{1}{n}\sum|x_i - A|$
        (Que tan lejos está xi de la medida de tendencia central) (el valor absoluto en la formula evita que dé 0, como en la media aritmética)

(mientras más pequeña sea la desviación, más se parecen los datos)

- Desviación Inter cuartil
  - Representa qué tanto se desvía el 50% de los datos "de en medio" respecto de la mediana, por lo que se calcula utilizando Q1 y Q3.
  - Como no toma en cuenta valores debajo de Q1 ni arriba de Q3, no se ve afectado por los valores extremos. Tampoco se ve afectado si se suma una constante a todos los datos, pero sí se afecta si se multiplican todos los datos por cualquier número.
  - No existe en Excel una función directa, pero recordemos que tenemos la función **CUARTIL.EXC**
  - **Descripción Matemática:** $\boldsymbol{DQ = \frac{Q3 - Q1}{2}}$
    Al numerador de la formula se le conoce como **rango Inter cuartil** y se representa como RQ
    (indica qué tan lejos estás de la MEDIANA)

- Desviación estándar:
  - Considerada como la mejor medida de dispersión.
  - Se utiliza, junto con la media aritmética, para dar reglas empíricas acerca de los valores poco comunes (como la regla de las desviaciones para distribuciones gaussianas).
  - No se ve afectada por si sumamos un valor a todos los datos; sin embargo, se afecta (y en gran medida) si todos los datos se multiplican por un mismo número
  - En Excel existen dos funciones para calcularla: **DESVEST.M** y **DESVEST.P** . La diferencia entre ellas es que la primera calcula la desviación estándar MUESTRAL (y por lo tanto aplica la fórmula de la desviación estándar) y la segunda calcula la desviación estándar POBLACIONAL (que en lugar de dividir por n-1, divide por el tamaño de la población)
  - **Descripción Matemática:** Si consideramos una muestra x1, x2, …, xn (los datos no están necesariamente ordenados), se calcula como
    $SD = \sqrt{\frac{1}{n-1}\sum(x_i - MA)^2}$
    (La fórmula es para la desviación estándar Muestral)
    (Se divide entre n-1 (estimadores sin sesgados))
    (Que tan lejos queda cada puntito de la media aritmética. A diferencia de la media absoluta, es que al elevar al cuadrado se implica el uso del Teorema de Pitágoras) (la desviación estándar se usa para regresiones lineales (predicciones de variables continuas, mientras que la media absoluta cuando se quieran hacer clasifs.)

- Varianza:
  - Se trata de la desviación estándar elevada al cuadrado.
  - En Excel existen dos funciones para calcular la varianza: **VAR.S** y **VARS.P.** La diferencia entre ellas es que la primera calcula la varianza MUESTRAL (sample en ingles) y la segunda calcula la varianza POBLACIONAL
  - **Descripción Matemática:** Si consideramos una muestra x1, x2, …, xn (los datos no están necesariamente ordenados), se calcula como $Var = \frac{1}{n-1}\sum(x_i - MA)^2$

b) Relativa: Estas medidas nos ayudan a comparar con respecto a otros datos. Las fórmulas de las distintas medidas de dispersión relativas se encuentran en el archivo de Excel C03_Medidas de Dispersión:

- Coeficiente de Desviación Media: $\frac{Desviacion\ Media}{Media\ Aritmetica}$
- Coeficiente de Desviación Inter Cuartil: $\frac{Q3 - Q1}{Q3 + Q1}$
- Coeficiente de Varianza: $\frac{Desviacion\ Estandar}{Media\ Aritmetica}$

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/i96l0JrtWK0

CORRELACIÓN

- Hasta ahora ya sabemos que existen números que pueden ser usados para representar un conjunto de datos de una característica (medidas de tendencia central) y una manera de ver qué tan buena esa es esa representación (medidas de dispersión absolutas). Más aun, sabemos comparar, dadas dos características en una muestra o más muestras, cuál de ellas tiene menor dispersión (medidas de dispersión relativas).
- El siguiente problema al que nos enfrentamos es: **cómo sabemos si, dada una muestra con dos o más características (=columnas), hay alguna de ellas que dependa de alguna manera, de alguna otra (¿existirá al menos 2 columnas donde 1 dependa de la otra?).**
- Por ejemplo, en una fábrica se hacen 10 productos diferentes. Para cada uno, se anota el número de unidades producidas, el número de unidades vendidas y el precio de venta durante un periodo de tiempo. ¿Habrá alguna relación matemática entre algunas DOS de estas mediciones?

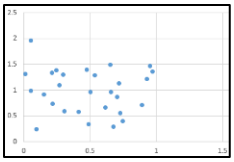
- Sean X_1 y X_2 dos características que medimos de una muestra. Lo que nos hemos preguntado anteriormente es: ¿Existe una función f tal que $f(X_1) = X_2$? Más aún: en caso de existir, ¿Cómo es esa función?
- Si existe una función f que relaciones X_1 y X_2, entonces existe correlación.
- La correlación es una medida de qué tan relacionadas están dos características dentro de una misma muestra.
- Existe un tipo de correlación, la llamada lineal, que trata de establecer una medida bajo la cual, bajo ciertos criterios subjetivos, podamos decir que dicha relación entre las dos características existe y además es lineal. Es decir, $X_2 = aX_1 + b$ para algunos números a y b.
- En otras palabras, la correlación lineal es una medida de qué tan relacionadas están, linealmente, dos características dentro de una misma muestra. Además, una propiedad importante que tiene es que carece de unidades y no se altera al sumar el mismo número a las dos características o multiplicas ambas por un mismo número.

- **Tipos de Correlación**

  Existen 6 tipos de correlación. Estos se pueden identificar haciendo un gráfico de puntos tomando como abscisas y ordenadas las dos características que se quieren analizar.

  1) **Perfectamente Positiva**: Los puntos forman una recta ascendente (si una característica aumenta de valor, la otra también). (X1 y X3 (C04 Tabla_datos))
  2) **Perfectamente Negativa**: Los puntos forman una recta descendente (si una característica aumenta de valor, la otra disminuye). (X1 y X4 (C04 Tabla_datos))
  3) **Parcialmente positiva**: Los puntos no forman una recta, pero si una característica aumenta de valor, la otra también. (X1 y X5 (C04 Tabla_datos))
  4) **Parcialmente negativa**: Los puntos no forman una recta, pero si una característica aumenta de valor, la otra disminuye. (X1 y X6 (C04 Tabla_datos))
  5) **Uncorrelacion**: No se ve una relación funcional de ningún tipo. (X1 y X2 (“”))
  6) **Correlación Curva**: Hay una relación funcional no lineal. (X1 y X7 (“”))

- Recordemos que una nube de puntos es una gráfica de la forma



- **Coeficiente de Correlación de Pearson:**
- Este es un número que mide el nivel de CORRELACIÓN LINEAL (no mide la curva) entre dos características de una misma muestra. Se trata de un número entre -1 y 1. Si su valor absoluto es 1, entonces hay correlación perfecta. Si el valor es 1, la correlación es además positiva; si es -1, es correlación negativa. Finalmente, si es 0, **no hay correlación lineal,** pudiendo haber otro tipo de correlación.
- En Excel es simplemente **COEF.DE.CORREL(Característica1, Característica2)**
- **Descripción Matemática:** Sean X1 y X2 dos características de la misma muestra. Digamos $X_1 = (x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)})$ y $X_2 = (x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)})$. Si $\overline{X_1}$ y $\overline{X_2}$ son las medias aritméticas de las características, respectivamente, entonces se define el coeficiente de correlación de Pearson de las características como

$$\rho(X_1, X_2) = \frac{\sum(X_i^{(1)} - \overline{X_1})(X_i^{(2)} - \overline{X_2})}{\sqrt{\sum(X_i^{(1)} - \overline{X_1})^2}\sqrt{\sum(X_i^{(1)} - \overline{X_2})^2}}$$

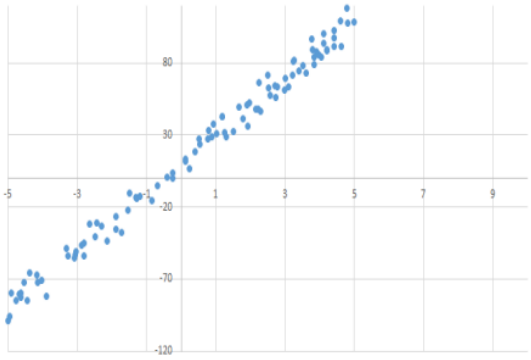
- **Desventajas de la correlación**
    - Los coeficientes de correlación más utilizadas sólo miden una relación lineal. Por lo tanto, es perfectamente posible que, si bien existe una fuerte relación no lineal entre las variables, r está cerca de 0 o igual a 0. En tal caso, una nube de puntos puede indicar aproximadamente la existencia o no de una relación no lineal.
    - Hay que tener cuidado al interpretar el valor de ρ. Por ejemplo, se podría calcular ρ entre el número de calzado y la inteligencia de las personas, la altura y los ingresos. Cualquiera sea el valor de ρ, no tiene sentido y por lo tanto es llamado correlación de oportunidad o sin sentido.
    - El ρ no debe ser utilizado para decir algo sobre la relación entre causa y efecto. Dicho de otra manera, al examinar el valor de ρ podríamos concluir que las variables X_1 y X_2 están relacionadas. Sin embargo, el mismo valor de ρ no nos dice si X_1 causa a X_2 o al revés. La correlación estadística no debe ser la herramienta principal para estudiar la causalidad, por el problema con las terceras variables.

Leer coeficiente de rangos de Spearman (verificar juicios) https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-spearman/
Para ver la clase de este tema, vaya al siguiente link: https://youtu.be/Bycf-rVgmYs

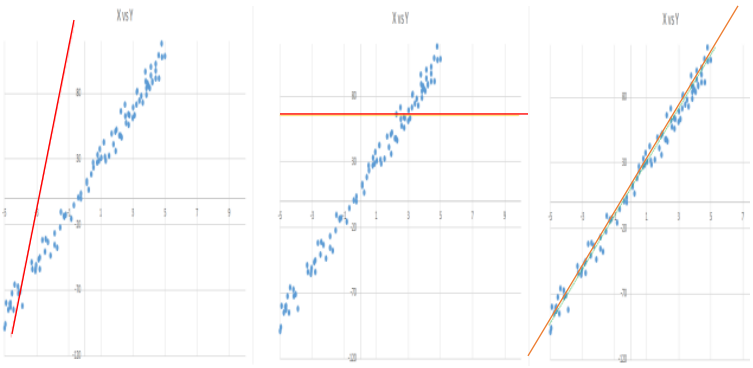
REGRESIONES LINEALES

- Las regresiones son maneras de MODELAR una característica medida utilizando otras, también medidas, de la misma muestra con el objetivo de crear predicciones. Esto es: si X_1, X_2, …, X_n, X_{n+1} son algunas de las columnas de las tablas, encontrar una función *f* tal que $X_{n+1} = f(X_1, X_2, \ldots, X_n)$. En cristiano: **¿será posible explicar el comportamiento de una de las características a través del conocimiento de otras?** (a diferencia de la correlación, aquí no se busca un número (el coeficiente), sino una función *f*)
- Bajo la idea anterior, decimos que las características X_1, X_2, …, X_n son **explicativas** y la característica X_{n+1} es la variable objetivo (o variable explicada).

- En este capítulo platicaremos de un problema de regresión muy sencillo conocido como **regresión lineal**. Observemos la siguiente nube de puntos:



- Debido a su forma, vale preguntarse cuál será la recta que mejor se aproxime, en algún sentido, a todos los puntos al mismo tiempo.
- Observemos varias rectas graficadas con la nube de puntos. ¿Cuál dirías que es la que más se ajusta a todos los puntos al mismo tiempo?



- Antes de avanzar, recordemos nuestros cursos de Geometría Analítica: toda recta en el plano en el plano es una ecuación de la forma

$$y = mx + b$$

- Donde
  - b se conoce como ordenada al origen y es el valor sobre el eje Y en que en la recta lo atraviesa. En regresión lineal se llama **intercepto.**
  - m se conoce como **pendiente** de la recta y se identifica como la tangente inversa del ángulo que hace la recta con el eje X. En cristiano: m mide la inclinación de la

recta. También en regresión lineal se llama pendiente. Si m>0 la recta va hacia arriba; si m<0 la recta va hacia abajo; si m=0, la recta es horizontal.

- Por lo tanto, **hallar la ecuación de una recta equivale a hallar los valores de b y m.**

- **Terminología:**

- Recordemos nuestro objetivo: tenemos una lista de parejas de puntos (x_1, y_1), (x_2, y_2), …, (x_d, y_d). Los graficamos como nube de puntos y buscamos la recta que mejor aproxime, en cierto sentido, a todos los puntos al mismo tiempo. Digamos que y=mx+b es una recta que tomamos como aproximación. Por lo tanto, a cada xi se le asignan dos números: el y_i (que es un valor que conocemos) y el $\widehat{y_i}$ = mx+b (y_i y $\widehat{y_i}$ no son lo mismo, aunque pueden ser iguales: uno es la "pareja natural" y el otro es la imagen con la recta respectivamente), que es el valor que nos da la recta para ese número x_i.
    - **Predicciones:** Son los valores $\widehat{y_1}, \widehat{y_2}, \ldots, \widehat{y_d}$. Es decir, los valores que la recta predice.
    - **Residuos:** "¿Qué tanto se equivocó la recta?". Recordemos: la recta le asigna a xi el valor $\widehat{y_i}$. $Pero\ el\ valor\ verdadero\ que\ acompaña\ a\ xi\ es\ yi$. Los residuos son los errores que la recta cometió: $\varepsilon_i = y_i - y_i$ (¿Qué tanto me equivoque en ese punto i?)(si el residuo es positivo, el punto está por arriba de la recta; si es negativo, el punto está por debajo de la recta; mientras más se acerque a 0, mejor será la recta graficada para ese conjunto de puntos).
    - **Error estándar:** Es la desviación estándar de los residuos. También se conoce como **error cuadrático medio** y se interpreta como el error cometido conjuntamente por la recta en todos los puntos (¿Cuánto me equivoque globalmente (se basa en el residuo))
    - **Coeficiente de determinación:** Es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, determinado por R^2, refleja la bondad del ajuste de un modelo a la variable que pretende explicar.

- Como dijimos anteriormente, se trata de encontrar la recta que mejor aproxime a los puntos en cierto sentido. Esto con el objetivo de poder realizar predicciones.
- Es decir, supongamos que tenemos una tabla que contiene al menos dos columnas y d renglones. Además, hay dos columnas, llamadas X y Y, de las cuales obtenemos la regresión lineal, lo que significa que encontramos números m y b tales que la recta y=mx+b es la que mejor aproxima a la nube de puntos.
- Si tenemos evidencia para afirmar que el modelo es bueno (lo que significa que los errores cometidos fueron suficientemente pequeños), entonces al recibir un nuevo renglón d+1 del cual desconocemos su valor en la característica Y pero conocemos su valor en la característica X, digamos x_{d+1}, entonces el valor en Y deberá ser aproximadamente

$$mx_{d+1} + b$$

- Ahora bien… ¿qué significa que la recta sea la que mejor aproxime a todos los puntos al mismo tiempo? Tiene muchas rectas de donde elegir!! Bueno, cada recta que das te genera una lista de residuos. La suma de todos los residuos te da el error cometido globalmente. Pero eso es precisamente el error estándar. O sea que buscas la recta que te consiga el menor error estándar posible.

- Analíticamente, el problema anterior se escribe como: Hallar números m y b tales que la función $\mathrm{C(m,b)} = \sqrt{\frac{1}{d}\sum(mx_i + b - y_i)^2}$ sea mínima. Esto es lo que se conoce como **Problema de los mínimos cuadrados.** Para el caso de la regresión lineal simple, se sabe que si tomamos $m = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2}$ $y\ b = \bar{Y} - m\bar{X}$ entonces la recta y=mx+b es la que mejor aproxima.

- **En Excel**, para realizar la regresión lineal se usa la aplicación **Regresión** del panel especial de **Análisis de datos** dentro del menú **Datos.**

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/3sX0TFeiuZg

INTRODUCCIÓN A LA INVESTIGACION

- El análisis y la investigación nos permiten buscar información pertinente acerca de un tema en específico. Básicamente nos permite:
  - Entender mejor
  - Determinar la frecuencia de un evento
  - Comparar relaciones entre las variables
  - Describir varias características de nuestros datos
  - Reportar y presentar conclusiones

- En Investigación, podemos aplicar la Estadística a cualquier problema enfocado con procesos sistemáticos y científicos mediante:
  - Recolección de datos
  - Recopilación de datos
  - Análisis de los datos: Ya se hacen operaciones matemáticas. Se entienden a profundidad los datos recolectados
  - Interpretación

- **Diferencia entre investigación básica y aplicada**

  Existen, a grandes rasgos, dos tipos de investigación: básica y aplicada

  a) Básica:
     - Es usada únicamente para la expansión de conocimientos (por ejemplo, es usada por los matemáticos puros o los físicos teóricos).
     - Viene impulsado por la simple curiosidad
     - Busca responder preguntas sobre fundamentos

  b) Aplicada:
     - Busca un uso comercial o de desarrollo (ya no se hace la investigación por la expansión del conocimiento)
     - Se busca una explotación de investigación.

Es impulsada por la solución de un problema real
- Responde preguntas especificas

- **Toma de decisiones**

El proceso en la toma de decisiones es el siguiente:

1. Planteamiento del problema u oportunidades (en qué área tienes todavía oportunidad de crecer; en qué áreas tienes debilidades)
2. Diagnóstico del problema
3. Pasos de Investigación
4. Tomar una decisión
5. Implementación de las decisiones

- **Tipos de Investigación**

Existen varios tipos de investigación. A saber:

A. Exploratoria: Se usa primordialmente para explorar ideas sobre el fenómeno de estudio. Se generan ideas generales.
Se aplican diferentes fuentes de información como:
   1. Análisis de datos de fuentes secundarias
   2. Análisis de expertos
   3. Entrevistas a grupos focalizados
   4. Entrevistas en profundidad
   5. Análisis por casos
   6. Técnicas proyectivas

B. Descriptiva: Se aplica para describir las características del fenómeno de estudio basado en los datos. Sólo describe, no hace inferencia.
Por ejemplo: "El año pasado, la temperatura promedio en la Ciudad fue de…"
Existen dos tipos:
   I. Secciones cruzadas:
      - ❖ Recolecta la información únicamente para un punto en el tiempo
      - ❖ La muestra puede no ser la misma en cada estudio
      - ❖ Busca un enfoque más cuantitativo
   II. Longitudinal:
      - ❖ Recolecta la información para un periodo de tiempo
      - ❖ La muestra es la misma en todos los estudios
      - ❖ Trabaja enfoque cuantitativo y cualitativo

C. Causal: Identifica las causas y efectos entre las variables dentro del fenómeno de estudio. Aquí si se hacen inferencias
   - Busca identificar las causas y los efectos.
   - Utiliza la experimentación como herramienta.
   - Es mucho más estructurada en la teoría y la práctica, por lo que su uso es aplicable cuando el problema o fenómeno de estudio está claramente definido.

Debido a esto, únicamente se utiliza en la investigación aplicada, ya que siempre es necesario tener al menos dos variables medidas para poder definir sus relaciones.

- **Trabajo con datos**

El proceso del trabajo con datos consiste en los siguientes pasos:

1. Recolección de datos y trabajo de campo
2. Preparación de los datos
3. Análisis de los datos
4. Interpretación de resultados
5. Implementación de las decisiones

- **Tipos de Fuentes:**

Se suelen dividir de dos maneras

a) Primarias:
   - El investigador recolecta los datos con un propósito en específico.
   - Hay más control y calidad en los datos
   - Más costoso y consume más tiempo consumido
   - Hay 3 básicas:
     - Encuestas
     - Observaciones
     - Experimentos

b) Secundarias:
   - El investigador utiliza los datos recolectados por alguien más
   - Puede haber menor calidad en los datos
   - Puede ser o no ser costoso, pero en definitiva se consume menos tiempo.
     - Bases de datos internas
     - Libros y artículos
     - Reportes y publicaciones
     - Fuentes públicas
     - Fuentes privadas
     - Redes sociales

- **Prueba de Hipótesis:**

- Una **HIPOTESIS** es una suposición que se hace sobre un parámetro poblacional (TODAS las personas de tu universo están incluidas), la cual puede ser verdadera o falsa.
- Por ejemplo, basado en la información que el investigador tiene sobre los últimos 10 años acerca de un fenómeno, puede medir el comportamiento del mismo fenómeno en el año siguiente.
- Luego, las pruebas de hipótesis son procedimientos estadísticos utilizados para verificar las suposiciones del investigador.
- Tipos de hipótesis:
  a) Nula:

- Suposición que queremos probar
- Se denota por H0

b) Alternativa:

- Conclusión que obtenemos si la H0 es rechazada.
- Se denota por Ha

- **Interpretación de los Resultados:**
- Para poder interpretar correctamente los resultados, el investigador debe definir un **valor crítico** usualmente basado en su experiencia u objetivos.

Valor calculado < Valor crítico En esta condición, aceptamos la hipótesis nula

Valor calculado >= Valor critico En esta condición, rechazamos la hipótesis nula y se acepta la hipótesis alternativa

- **Tipos de errores cometidos:**
- Cuando aceptamos o rechazamos una hipótesis, podemos cometer algunos errores. Estos se dividen en dos tipos: **Tipo I** y **Tipo II:**

	Decisión tomada	
	Aceptar hipótesis nula	**Rechazar hipótesis nula**
H0 es verdadera	No hay error	Error de tipo I
H0 es falsa	Error de tipo II	No hay error

- ¿cómo pueden surgir estos errores?
  - Pueden deberse a la muestra (estaba contaminada o sesgada)

- Para ver la clase de este tema, vaya al siguiente link: https://youtu.be/zvtmSU92fvg

- **Recolección de Datos**
- Por lo general, buscamos que nuestros datos cumplan las siguientes características:
  - Temporalidad: Es decir, que sean relativamente recientes. Si los datos no son relativamente recientes, difícilmente se podrán usar para realizar predicciones.
  - Adecuados: También llamada **representatividad.** Si eliges muy poca cantidad de muestra o una cantidad demasiado grande, puedes terminar provocando un análisis inadecuado de los datos.
  - Validos: Tus datos deben ser válidos.
  - Medibles: Tus datos deben medirse.

- **Métodos de recolección de datos:**
    - **a)** Encuestas:
        - Se ocupan sobre todo para el análisis cuantitativo.
        - Son mucho menos caros y ahorran tiempo.
        - Las respuestas se pueden ver afectadas por el entrevistados
    - **b)** Censos:
        - Se ocupan para tener un conocimiento general del ambiente poblacional.
        - Son costosos tanto económicos como temporalmente.
        - Las respuestas no representan gran problema.
    - **c)** Sondeos
    - **d)** Casos particulares de estudio
    - **e)** Agencias recolectoras de datos
    - **f)** Fuentes públicas
    - **g)** Experimentos
    - **h)** Redes sociales e internet

- **Cuestionarios**
- Debe tener un conjunto de preguntas formales y específicas. Estos es, preguntas que se alineen únicamente con el objeto de nuestra investigación y que formulados con un formato correcto.
- Las preguntas pueden ser o no estructuradas. Esto significa que pueden seguir una estructura de opciones múltiples o bien, preguntas cuyas respuestas pueden ser de cualquier tipo.
- Algo que siempre debes considerar es que tus preguntas deben incluir situaciones u objetivos específicos de tu interés de estudio. Recuerda, enfócate en los objetivos.

- **Proceso del diseño de un cuestionario**
    1. Entiende bien tu objetivo de investigación
    2. Toma en cuenta las características del entrevistado
    3. Selecciona correctamente el formato de cada pregunta
    4. Secuencia de las preguntas: fácil -> Difícil
    5. Borrador y visualización

- **Tipos de Preguntas I**

  Existen tres tipos básicos:
    I. Final abierto
    II. Preguntas dicotómicas
    III. Opción Múltiple
- **Tipos de Preguntas II**

  Dependiendo del formato y la seriación de las preguntas, también tenemos:
    I. Detección: Preguntas que sirven para saber si el entrevistado está o no en condiciones de responder todo o parcialmente el cuestionario.
    II. Apertura: Son preguntas iniciales que sirven para dirigir al entrevistado. Deben ser atrevidas, pero sencillas y fáciles de responder.

III. Declaraciones de transición: Sirven para hacer que el entrevistado se mueva de una a otra sección del cuestionario.
IV. Difíciles: Se trata de preguntas sensibles o difíciles de responder y deben dejarse para el final del cuestionario.

- **Niveles de mediciones de datos**
  Existen cuatro niveles para clasificar nuestros datos. Ya los hemos comentado antes: Nominales, Ordinales, Intervalo, Razón

- **Escalas con un solo Ítem**
  - Multiple Choice Scale: Multiple Choice are provides to collect nominal data
  - Forced Choice Ranking: Respondents rank different objects from a list
  - Constant Sum Scale: Respondents allocates points to objects, but total remains same.
  - Direct Quantification: Directly ask question, to collect ratio scaled data
- **Escalas con varios ítems**
  - Likert Scarle: 5 categories from strongly disagree to strongly agree
  - Semantic Differential: Scale of 1 to 5 OR 1 to 7. Extreme points hihglights characteristics.
  - Staple Scales: Presented vertical scale with positive and negative ratings both sides
  - Numerical Scales: Numbers as scale points are used
- **Escalas Continuas:**
  - Continuos Rating: Two extreme measuring ends are given.

- **Muestreo:**
- Se trata de una herramienta (proceso) para obtener la muestra de nuestra población de interés. Es bastante útil ya que disminuye el tiempo de recolección de datos y los factores económicos. Además, si la población de interés de estudio es muy grande, ofrece ventajas muy evidentes y notorias.

- **Diseño del muestreo:**
  1. Definir la población de estudio
  2. Determinar el marco del muestreo
  3. Seleccionar una técnica de muestreo
  4. Seleccionar el tamaño de la muestra
  5. Seleccionar la muestra

- **Métodos de muestreo:**
  Se dividen en dos categorías. Lo que los diferencia es la oportunidad que tiene una unidad para pertenecer o no a la muestra:
  - Aleatorios:
    - Todos tienen la misma oportunidad de ser elegidos
    - No se ocupan criterios de selección

- No Aleatorios:
  - No todos tienen la misma oportunidad de selección
  - Se basan en criterios de selección

- **Ejemplos de muestreo aleatorio:**
  - Simple: Todas las muestras posibles tienen la misma oportunidad de ser elegidas (si el tamaño del universo es N, hay 2N -1 muestras)
  - Estratificado: Antes de seleccionar la muestra, se divide a la población en estratos considerando características homogéneas al interior. Cada grupo se llama estrato.
  - Conglomerado: También se divide la población antes de seleccionar la muestra. Sin embargo, se hace mediante marcas o áreas geográficas. Cada grupo se llama conglomerado.
  - Sistemático: Es un tipo de muestreo donde la parte aleatoria viene dada al escoger al primer individuo de la muestra.
  - Multietapico: Se aplican varios de los muestreos anteriores y se arma una muestra final.

- **Ejemplos de muestreo no aleatorio**
  - Por cuotas: Se seleccionan los elementos con el objetivo de cumplir con una cuota, la cual es establecida por el investigador, quien debe tener en cuenta la proporción de cada grupo. Si su población objetivo tiene un 40% de mujeres, el muestreo por cuotas también debería incluir elementos en la misma proporción.
  - Por conveniencia: El investigador elige a los miembros solo por su proximidad y no considera si realmente estos representan muestra representativa de toda la población o no.
  - Por juicio: El investigador elige, a su juicio, la muestra
  - Bola de nieve: El investigador elige la muestra a partir de las sugerencias de alguna observación ya captada.

- **Errores:**

  Los errores más comunes al seleccionar una muestra son:
  - ✓ Muestreo: No hay representatividad
  - ✓ Respuesta: El entrevistado no respondió correctamente.
  - ✓ Sesgo por no respuesta: El entrevistado no quiso responder
  - ✓ Errores del investigador: Errores propios cometidos por el equipo de investigación.

- **Codificación:**
- Se trata de la codificación en números de cada una de las variables a estudiar. Se asignan valores y códigos a cada respuesta. Por ejemplo, en la escala de Likert para una pregunta en particular, puede ser codificada con los valores 1, 2 , 3, etc.
- Con esto, cumplimos el criterio de medibilidad de nuestros datos

- **Entrada y edición de datos:**
- En la entrada de datos trataremos de separar nuestra información. Se eligen, por ejemplo, características demográficas para hacerlo. La elección de las características con las que haremos la separación es muy útil dependiendo del tipo de investigación que hagamos. Luego, se incluyen todos los valores validos disponibles y se vacían los datos en Excel.
- Esto creará lo que se conoce como una **base de datos.**
- Además, probablemente tengamos que editar algunos datos. Por ejemplo, el trato de datos faltantes o las contradicciones entre los datos debidos a errores del investigador.

- **Limpieza:**

  En general no se tienen los datos en forma pura; esto es, se pudieron cometer errores durante la recolección. Para solucionar esto, tenemos técnicas de limpieza de datos. En particular, algunos de los problemas que se buscan resolver son:

  1. Datos faltantes:
  2. Datos no coincidentes
  3. Gestión de metadatos (descripción de cada una de las columnas de la tabla)
  4. Filtrado de datos

- **Operaciones entre múltiples fuentes:**

  Cuando nuestra información esta diseminada en múltiples fuentes, algunas operaciones para concentrar la información son las siguientes:

    - Concatenación de tablas
    - Combinación de tablas
    - Crear pegados horizontales (joins)
    - Crear pegados verticales (unions)
    - Quitar duplicados

- **Herramientas para preparar datos:**
- Algunos programas y lenguajes especializados en la preparación de datos son:
    - Tablean
    - Excel
    - SQL
    - R
    - Python

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/qrZbR5sWmgQ

ESTADISTICA DESCRIPTIVA

- Básicamente, la Estadística descriptiva se enfoca, como su nombre lo dice, en describir los atributos y características de un fenómeno visto como un conjunto de datos. Su propósito principal es hacer resúmenes de los datos y hacerlos presentables para los interesados.
- Se utilizan tanto las medidas de tendencia central como las medidas de dispersión.
- Dado que se enfoca en hacer resúmenes de la información, sirve para darnos una idea general acerca de la distribución de los datos.

- **Usos de la Estadística Descriptiva:**
  De lo anterior, en síntesis, podemos deducir algunos de sus usos
  1. Realizar resúmenes para presentar la información
  2. Conocer sobre la distribución de los datos
  3. Encontrar características "ocultas" del fenómeno.
  4. Prepararnos para establecer alguna hipótesis.

- **Agregaciones Básicas:**
  Excel provee varias agregaciones básicas que ya hemos estudiado, las cuales son:
  - Mínimo: MIN
  - Máximo: MAX
  - Suma: SUM
  - Conteo: CONTA

- **Medidas de Tendencia Central**
  Ya hemos aprendido que las medidas de tendencia central son números que usamos para representar nuestros datos. Recordemos que las más usuales son
  - Media aritmética: PROMEDIO
  - Mediana: MEDIANA
  - Moda: MODA

- **Medidas de dispersión:**
  A su vez, ya que hemos elegido el número con el que se representa una característica (es decir, una medida de tendencia central), tenemos mediciones para saber qué tan bien ese número representa a la característica. Estas son las medidas de dispersión:
  - Rango: MAX – MIN
  - Desviación Estandas: DESVEST.M
  - Varianza muestral: VAR.S o DESVEST.M^2

- **Medidas de Forma:**
  Veamos ahora las llamadas medidas de forma:
  - Sesgo: Mide la falta de simetría en nuestros datos. En general, es de tipo izquierdo o derecho. Se refleja como una cola larga, así que se enfoca en la forma horizontal de nuestros datos.
  - Curtosis: Mide la cantidad de datos distribuidos cerca de un pico, que generalmente es la medida de tendencia central elegida.

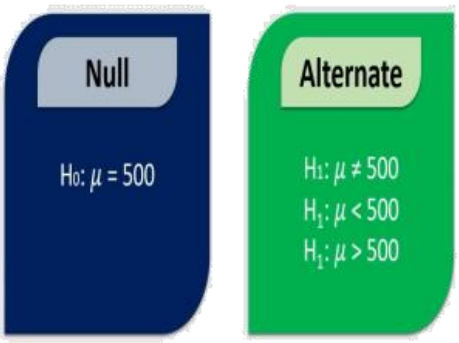
- Excel nos provee una herramienta muy sencilla para realizar los cálculos de la Estadística descriptiva; el panel de Análisis de datos que ya hemos usado anteriormente. En ella, se encuentra la opción **Estadística descriptiva.** Se abrirá un cuadro de dialogo

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/_RFCEJV2SGY

PRUEBA T

- Se trata de una prueba estadística para demostrar una hipótesis sobre la media poblacional. A saber:
    - Se utiliza la prueba T de una muestra para averiguar si la media poblacional toma o no cierto valor. Es decir, se toma como hipótesis nula el enunciado "'H0: la media poblacional real vale A"'.
    - Se puede utilizar cuando la muestra tiene menos de 30 elementos.
    - Más general: se ocupa cuando la desviación estándar poblacional es desconocida.

- **Hipótesis**



- La media poblacional se denota por μ.
- La media poblacional no tiene que ver con el tamaño de la muestra ni con el número de elementos. La media poblacional representa una característica general de una población menor a 30 elementos.
- Se quiere comprobar que μ es real.
- Las demás alternativas son que la media poblacional μ es distinta (<>, <, >) de A.

- **Proceso de Investigación:**
- Los pasos para la aplicación de la prueba T son:
    1. **Seleccionar el nivel de significación:** Este es denotado por α. Generalmente α = 0.01, α = 0.05, α = 0.10. Esto representa la probabilidad de rechazar la hipótesis nula cuando es verdadera (Error Tipo I). Por ejemplo, un nivel de significación de 0.05 indica un riesgo del 5% de concluir que existe una diferencia entre los resultados del estudio y la hipótesis nula cuando en realidad no hay ninguna diferencia. Al número 1 – α se le llama **nivel de confianza.** ("Tengo una confianza del 90% de que cada mexicano mayor de 18 años fue víctima de, en promedio, 1.4 delitos (se toma a α del 10%" (es más común agarrar 95% de confianza) (la interpretación de esta frase (por ejemplo): de 100 pruebas, el 95 de ellas serán verdad, y fallará en 5).

        Mientras menos confianza haya, menos probabilidad hay de equivocarse; y mientras más confianza, más probabilidad hay de error.

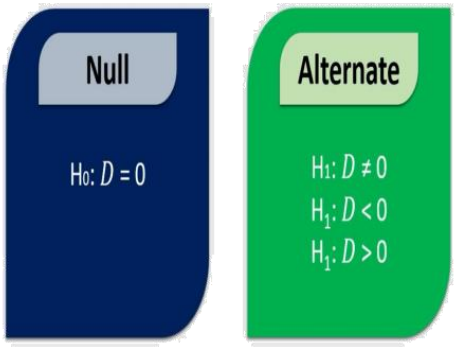
2. **Encontrar el valor crítico:** Este es denotado por $t_{n-1,\ \alpha'}$ (es α' si se toma una hipótesis alternativa). Antiguamente se usaba una tabla de T. Se seleccionaba la columna basada en α y la fila basado en los **grados de libertad**, que en este caso en n – 1 (siendo n el tamaño de la muestra).
3. **Calcular un parámetro:** Se refiere a calcular el número $T = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ (S/ es la desviación muestral) con los datos de la muestra.
4. **Comparar y decidir:** Si $|T| < t_{n-1,\alpha'}$, aceptamos H0. En caso contrario, rechazamos la hipótesis nula.

- **Proceso en Excel**
  En la opción **Análisis de Datos,** elegimos la opción **PRUEBA t para dos muestras suponiendo varianzas desiguales**, para luego llenar los campos requeridos.

PRUEBA T APAREADA

- Se trata de una prueba estadística para demostrar una hipótesis sobre las medias poblacionales de dos poblaciones. A saber:
  - Se utiliza la prueba T apareada para averiguar si las medias poblacionales de dos poblaciones diferentes son iguales (prueba independiente)
  - En este sentido, también, sirve para comparar las medias de dos características de una misma población (prueba dependiente).
  - Es decir, se toma como hipótesis nula el enunciado "'H0; las medias poblacionales de P1 y P2 son iguales"'.

- **Hipótesis:**



- D representa "diferencia entre las medias". Es decir, que los promedios de las dos muestras (columnas) hayan sido iguales.
- Las hipótesis es que la Diferencia sea <>0, es decir, los promedios son diferentes.
- En la Hipótesis Alternativa, nos enfocaremos en la de dos colas (<>)

- **Proceso de Investigación:**
- Los pasos para la aplicación de la prueba T apareada son:
  1. **Seleccionar el nivel de significación:** Este es denotado por α. Generalmente α = 0.01, α = 0.05, α = 0.10. Esto representa la probabilidad de rechazar la hipótesis nula cuando es verdadera.
  2. **Encontrar el valor crítico:** Este es denotado por $t_{n-1,\ \alpha'}$.

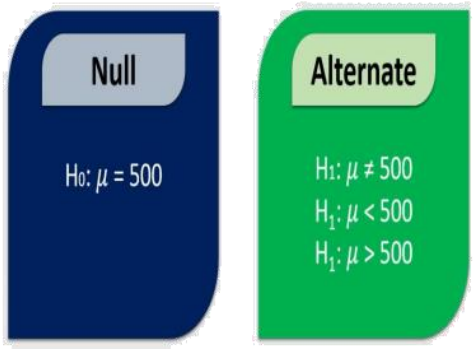
3. **Calcular un parámetro:** Se refiere a calcular el número $T = \frac{\bar{d} - D}{S_{diff}/\sqrt{n}}$ con los datos de la muestra. $\bar{d}$ representa el promedio de las diferencias (se hará una nueva columna), D supondremos que es 0, S_{diff} representa la desviación estándar de la diferencia)
4. **Comparar y decidir:** Si $|T| < t_{n-1,\alpha}$, aceptamos H0. En caso contrario, rechazamos la hipótesis nula.
   Si rechazamos la H0, entonces para saber qué muestra es más grande sólo basta con ver los promedios y ver cuál es el más grande.

- **Proceso en Excel:**
- En **Análisis de Datos**, elegimos **Prueba T para medias de dos muestras emparejadas.**
- Si se trabajará con datos temporales (años, por ejemplo), el rango para la variable 1 será el año más grande

PRUEBA Z

- Se trata de una prueba estadística para demostrar una hipótesis sobre la media poblacional. A saber:
  - Se utiliza la prueba Z de una muestra para averiguar si la media poblacional toma o no cierto valor. Es decir, se toma como hipótesis nula el enunciado "'H0: la media poblacional real vale A"'.
  - Se utiliza cuando la muestra tiene más de 30 elementos.
  - Más general: Se ocupa cuando la desviación estándar poblacional es conocida.

- **Hipótesis:**



- **Proceso de Investigación:**
- Los pasos para la aplicación de la prueba Z son:
  1. **Seleccionar el nivel de significación:** Este es denotado por α. Generalmente α = 0.01, α = 0.05 o α = 0.10. Esto representa la probabilidad de rechazar la hipótesis nula cuando es verdadera. Por ejemplo, un nivel de significación de 0.05 indica un riesgo del 5% de concluir que existe una diferencia entre los resultados del estudio y la hipótesis nula cuando en realidad no hay ninguna diferencia. Al número 1 – α se le llama **nivel de confianza.**

2. **Encontrar el valor critico:** Este es denotado por $Z_{n-1,\alpha}$. Antiguamente se usaba una tabla de T. Se seleccionaba la Columba basado en α y la fila basado en los **grados de libertad**, que en este caso es n – 1 (siendo n el tamaño de la muestra).
3. **Calcular un parámetro.** Se refiere a calcular el número $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ con los datos de la muestra.
4. **Comparar y decidir:** Si $|Z| < t_{n-1,\alpha}$, aceptamos H0. En caso contrario, rechazamos la hipótesis nula.

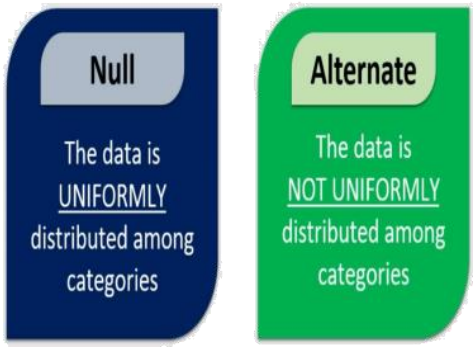
- **Proceso en Excel:**
- En **Análisis de Datos**, elegimos la opción **Prueba Z para medias de dos muestras.**

Para ver las clases de estos temas, vaya a los siguientes links:
https://youtu.be/n3NdbhDEmiY
https://youtu.be/FC7__aq9gTA

Prueba CHI CUADRADA

- Se trata de una prueba estadística utilizada cuando tenemos datos categóricos
  - Nos ayuda cuando buscamos la dependencia entre características de nuestro fenómeno. Por ejemplo, determine si las ventas de diferentes colores de automóviles dependen de la ciudad donde se venden.
  - Es una prueba de hipótesis que compara la distribución observada de los datos con una distribución esperada de los datos. Por ejemplo, puedes comprobar si un dado es justo lanzando el dado muchas veces y utilizando una prueba chi-cuadrada para determinar si los resultados siguen una distribución uniforme (cuando todas las categorías tienen más igual la misma cantidad de observaciones).

- **Hipótesis**



- **H0:** Los datos siguen una misma categoría.
- La hipótesis alternativa es que una categoría jala más observaciones que otra.

- **Proceso de Investigación:**

Los pasos para la aplicación de la prueba chi cuadrada son:

1. **Seleccionar el nivel de significación:** Este es denotado por α. Generalmente α = 0.01, α = 0.05 o α = 0.10.

2. **Encontrar el valor critico:** Este es denotado por $\chi_{n-1,\alpha}$. Antiguamente se usaba una tabla de χ^2. Se seleccionaba la columna basado en α y la fila basado en los **grados de libertad**, que en este caso es n – 1 (siendo n el tamaño de la muestra).
3. **Calcular un parámetro.** Se refiere a calcular el número $\chi^2 = \sum \frac{(o-e)^2}{e}$ con los datos de la muestra (valores originales y valores esperados).
4. **Comparar y decidir:** Si $\chi^2 < x_{n-1,\alpha}$, aceptamos H0. En caso contrario, rechazamos la hipótesis nula.

- **Proceso en Excel:**

No hay una fórmula directa en Excel para hacer pruebas de χ^2. Sin embargo, se pueden utilizar las funciones de agregación (MIN, MAX, CONTAR) que ya hemos estudiado junto con la función **INV.CHICUAD.CD** que devuelve la función inversa de la probabilidad de una cola de la distribución chi cuadrado. La sintaxis de esta función es INV.CHICUAD.CD(∝,n-1). Esto significa que $x^2{}_{n-1,\alpha} = INV.CHICUAD.CD(\propto, n-1)$

- Para hacerlo con puras formulas, se usa **PRUEBA.CHICUAD,** y el número que arroje, lo usamos en parámetro "Probabilidad" de **INV.CHICUAD.CD**

Para ver las clases de estos temas, vaya a los siguientes links:
https://youtu.be/Y156jlQMS40

ANOVA DE UNA VÍA (PRUEBA F)

- Se trata del Análisis de la Varianza (de ahí su nombre). En particular, para ANOVA de una vía tenemos lo siguiente:
  - Se utiliza para comparar la varianza entre diferentes muestras elegidas de una misma población.
  - La idea es comparar dos o más muestras basados en la diferencia de las varianzas y se trata de una prueba de hipótesis sobre las medias poblacionales.
  - "Si tengo varios grupos en una población, ver qué tanto se parecen entre sí".
  - Puede ser usado cuando tenemos al menos dos variables, donde una es categórica y la otra es continua. A la categórica la llamamos **variable independiente** y a la continua la llamamos **variable de respuesta.**

- **Condiciones para ser aplicable**

- Supongamos que tenemos la variable independiente X y la variable de respuesta Y. Para poder aplicar correctamente el ANOVA de una vía se debe cumplir:
  - Los grupos son independientes.
  - Y debe ser aproximadamente normal en cada grupo (siendo menos estricta esta condición cuanto mayor sea el tamaño de cada grupo. (debe parecer montaña).

- Todos los grupos tienen la misma varianza (esta condición es más importante cuanto menor es el tamaño de los grupos).
- No tener datos atípicos

- **Hipótesis**

- Supongamos que la muestra se divide, según la variable independiente, en los grupos G_1, G_2, …, G_k (grupos de X). Sea μ_m la media poblacional del grupo m. Entonces la prueba de hipótesis se establece de la siguiente manera:

$$\begin{cases} H_0 : & \mu_1 = \mu_2 = \ldots = \mu_K \ (¿tendran\ los\ mismos\ valores?) \\ H_a : & hay\ al\ menos\ dos\ medias\ diferentes \end{cases}$$

- Por lo tanto, en caso de rechazar H0, se tiene que proceder a lo que se conoce como **pruebas post hoc**, siendo la prueba por parejas las más conocida. Ver https://github.com/scidatmath2020/Inferencia-Estadistica/blob/master/C08.%20ANOVA.ipynb

- **Proceso de Investigación**

Los pasos para la aplicación del ANOVA de una vía son:

1. **Seleccionar el nivel de significación:** Este es denotado por α. Generalmente α = 0.01. α = 0.05 o α = 0.10
2. **Encontrar el valor crítico:** Este es denotado por *f*. Antiguamente se usaba una tabla de F. Se seleccionaba la columna basado en α y la fila basado en los **grados de libertad.**
3. **Calcular un parámetro:** Se refiere a calcular el número F con los datos de la muestra como la media de la suma de los cuadrados.
4. **Comparar y decidir:** Si F < *f*, aceptamos H0. En caso contrario, rechazamos la hipótesis nula.

- **Proceso en Excel**

Nos vamos a la pestaña Datos, a Análisis de Datos, y elegimos “Análisis de Varianza de un Factor”.

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/kmrg-ghgT1s

ANNOVA DE DOS VÍAS

- Es un diseño de ANNOVA que permite estudiar simultáneamente los efectos de dos fuentes de variación:
  - En un ANNOVA de dos vías se clasifica a los individuos de acuerdo a dos factores (o vías) para estudiar simultáneamente sus efectos.
  - Sirve para estudiar la relación entre una variable dependiente cuantitativa y dos variables independientes cualitativas (factores) cada uno con varios niveles.
- Es como hacer 2 ANOVAS de una vía.

- **Ejemplo:**

- Tenemos interés en conocer los niveles de estrés de un grupo de jóvenes y cómo estos niveles se relacionan con el sexo y la percepción de salud. Entonces, en este caso tenemos tres variables: a) Estrés que es medida con un cuestionario y tiene puntajes y por ende será

nuestra variable de intervalo (variable a); b) Sexo, que es una variable categórica que puede ser hombre o mujer (variable b); c) Percepción de salud, que es otra variable categórica que tiene cinco opciones. Mala, Regular, Buena, Muy Buena y Excelente (variable c).

- **ANOVA de dos vías sin replicación:**
- Sin entrar en detalles, se puede decir que aplicamos un ANOVA de dos vías con replicación cuando una de las vías está a su vez dividida en niveles
- Se realiza un ANOVA de dos vías sin replicación para comparar, por ejemplo, las puntuaciones de los estudiantes en una batería de pruebas donde se tiene más de un grupo (digamos, de dos escuelas diferentes).

	Matemáticas	Español
Escuela 1	2	1
Escuela 2	5	3
Escuela 3	6	2

- **Proceso en Excel:**
- En Datos, nos vamos a "Análisis de varianza de dos factores con una sola muestra por grupo".

Para ver la clase de este tema, vaya al siguiente link:
https://youtu.be/ssrE1KkSVGQ