PROBABILIDADES Y ESTADÍSTICAS

RESUMEN FINAL

Unidad 1
Datos Estadísticos

- **Estadística:** Métodos para recoger, organizar, resumir y analizar de datos con el fin de tomar decisiones frente a la incertidumbre. Su objetivo es facilitar la comprensión, hacer más sencillo el conocimiento y lograr un mejor análisis de los hechos que se quieren interpretar.
  Puede ser:
  - **Estadística descriptiva:** Técnica que se encarga de la recopilación, tratamiento, presentación y análisis de los datos con el objetivo de resumirlos para su interpretación.
  - **Estadística Inferencial:** Es la técnica que permite a partir de la información provista por una muestra sacar conclusiones que puedan generalizarse para el resto de la población de la cual provienen.

- **Población o Universo:** Es la totalidad de individuos u objetos de interés, acerca de los cuales se desea información. Tienen características en común. Tiene que quedar bien definida en tiempo y espacio. Puede ser finita o infinita.

- **Muestra:** Es un subconjunto seleccionado de la población. Una muestra representativa es aquella que puede proporcionar una visión útil de la naturaleza de la población.

- **Censo:** Es una medición de todos los elementos de interés.

- **Muestreo:** Es la medición de todos los elementos de una muestra.

- **Datos estadísticos:** Son los valores observados de las variables que pueden ser comparados, analizados e interpretados.

- **Variable:** Es toda característica o dimensión de un individuo y objeto que puede adoptar distintos valores. Pueden ser:

  - **Variables Cuantitativas:** Son aquellas que adoptan valores numéricos. Pueden ser:
    - **Discretas:** Son aquellas que surgen de un proceso de conteo. Adoptan sólo valores enteros o fraccionarios determinados.
    - **Continuas:** Son aquellas que surgen de un proceso de medición. Pueden asumir cualquier valor numérico real, de modo que haya un flujo continuo de valores con graduaciones infinitamente pequeñas.

    Puedo discretizar una variable continua pero su naturaleza sigue siendo continua.

  - **Variables Cualitativas:** Son aquellas que arrojan respuestas que se describen por palabras, los individuos son poseedores o no poseedores de cierta propiedad. Solo

pueden clasificarse, no medirse
.

- **Unidad estadística:** Es el elemento del conjunto o universo poseedor de la característica sobre la cual queremos hacer el análisis.

- **Unidad de relevamiento:** Objeto o lugar del que se obtienen los datos.

- **Escalas de medida:** Se puede descender de escala pero no ascender.
    - **Escala nominal:** Consiste en categorías cualitativas mutuamente excluyentes que no tienen ningún orden lógico. Es la escala de medición más baja. Ej: país de origen.
    - **Escala Ordinal:** Consiste en categorías cualitativas en las que hay una progresión en el orden que pueden clasificarse por grados de acuerdo a algún criterio. No se puede medir las distancias entre categorías. Ej: Excelente/Bueno/Malo.
    - **Escala de intervalos:** Es un conjunto de intervalos para los que la distancia entre ellos es medible. Tienen un punto cero arbitrario y la razón entre los intervalos no tiene significado.
    - **Escala de razón:** Consiste en medidas numéricas para las cuales las distancias entre números es conocido, y donde la razón entre los números tienen algún significado. Existe un punto cero fijo.

- **Variables según el papel que cumplen:**

    - **Independientes:** Aquellas que toman valores que influyen en otras variables.
    - **Dependientes:** El valor de estas dependen del valor que tienen las variables independientes.
    - **De Control:** Sirven para comprender mejor la relación de dependencia entre las variables.

- **Etapas del método científico en el análisis de datos:**

1. Definición del problema (¿Qué?): Consiste en definir el objetivo de la investigación y precisar el universo o población.

2. Recogida de la información: Consiste en recolectar los datos necesarios relacionados al problema de investigación.

3. Análisis descriptivo: Consiste en resumir los datos disponibles para extraer la información relevante en el estudio.

4. Inferencia estadística: Consiste en suponer un modelo para toda la población partiendo de los datos analizados para obtener conclusiones generales. Solo se hace si trabaje con una muestra.

5. Diagnóstico: Consiste en verificar la validez de los supuestos del modelo que nos han permitido interpretar los datos y llegar a conclusiones sobre la población.

- **2. Recopilación de datos estadísticos:** Extracción y recolección de datos a partir de las fuentes que lo suministran. Los datos pueden ser primarios o secundarios según la fuente.

    - **Datos Secundarios:** Son aquellos que ya se han compilado y están disponibles para el análisis estadístico.

- **Datos Primarios:** Son aquellos que se recogen específicamente para el análisis deseado.

- **Relevamiento:**
  - **Estático:** Los datos son obtenidos en una fecha determinada.
  - **Dinámico:** Los datos corresponden a aquellas operaciones que se realizan en forma continuada a través del tiempo.

- **Técnicas para la recolección de datos primarios:**

  - **Grupos de interés:** Están integradas por un reducido número de personas que se reúnan para debatir qué datos son importantes. Se utilizan como guía para la investigación.
  - **Teléfono:** Entrevistas por teléfono. Rápidas, bajo costo, preguntas acotadas.
  - **Cuestionarios por correo:** Poca tasa de respuesta.
  - **Abordaje en centros comerciales:** Se utilizan para obtener opiniones de clientes.
  - **Registros.**
  - **Observaciones:** Difícil de analizar y comprobar.
  - **Entrevistas:** Se usa cuanto se necesita determinar en forma profunda las opiniones y actitudes (datos de calidad, costo y tiempo de alta).
  - **Experimentos.**

- **3. Clasificación, tabulación y presentación de datos:** Se refiere a la organización, presentación y descripción de los datos recopilados a los fines de facilitar la interpretación y el análisis de los mismos.
  Cuando los datos son pocos se puede presentar la información obtenida en forma de explicación literal, pero cuando los datos son muchos se hace preciso presentarlos en cuadros o tablas (distribuciones de frecuencias) o a través de gráficos y diagramas.

Formas de presentar los datos:

- **Escrita:** Se utiliza cuando la serie de datos es poca.
- **Tabla o Gráficos:** Cuando la serie de datos es mayor. Los gráficos se dividen en los que son para variables cualitativas y los que son para variables cuantitativas.

Unidad 2
Presentación de Datos Estadísticos

- **Partes de una tabla estadística:**
  - **Título.**
  - **Encabezamiento** (Cabeza de las columnas).
  - **Conceptos o columna matriz** (Columna de la izquierda).
  - **Cuerpo** (Datos).
  - **Notas del encabezamiento** (Aclaran la tabla).
  - **Notas al pie.**
  - **Fuente.**

- **Formas de agrupar variables cuantitativas:**
  - **Series simples:** Se utilizan cuando la serie de datos con la que contamos es pequeña por lo que no es necesario agruparlos. Cada valor de la serie representa una observación (X_1, X_2, …. X_n) según el orden en el que se presentan.
  - **Datos agrupados:** Cuando el número de datos es muy grande se utiliza una distribución de frecuencia para simplificar la información sin perder muchos detalles, agrupando los datos en forma de lista o intervalos:

Tablas de frecuencias: Son arreglos ordenados de los datos y en la columna de al lado las veces o frecuencia con la que se repite cada dato. Si los datos son pocos y tienen poca variedad, alcanza con ordenarlos en una lista. Si son muchos y con poca variedad se pueden agrupar en rangos. Si los datos son muchos y con mucha variedad, entonces conviene agruparlos con rangos de agrupación de los datos.

- **Distribución de frecuencias en lista:** Se construye una tabla de dos columnas. En la primera se listan los diferentes valores que asumió la variable ordenadas. En la segunda columna se indica el número de veces que cada valor distinto aparece, o bien la proporción que esa cantidad representa en el total.
  Cabe destacar que la variable que se analiza por tabla es única.

  - **Frecuencias Absolutas (ni):** Es la cantidad de veces que se repite un determinado valor de la variable. La suma de las frecuencias absolutas es igual al tamaño de la muestra.

$$n_i, \sum n_i = n$$

  - **Frecuencias Relativas (hi):** Proporción de los distintos valores de la variable que se obtiene dividiendo la frecuencia absoluta sobre el tamaño de la muestra. La suma de las frecuencias relativas es igual a 1.

$$h_i = \frac{n_i}{n}$$

- ➢ **Frecuencias Acumuladas:** Las frecuencias anteriores pueden ir sumándose desde el primer valor hasta la variable que nos interesa, de esa forma pueden ser:
  - **Frecuencias Absolutas Acumuladas (Ni):** Cantidad de datos que están por encima o por debajo de un valor determinado del intervalo i.

    $$N_i = \sum_{j=1}^{i} n_j = n_1 + n_2 + \ldots + n_i$$

  - **Frecuencias Relativas Acumuladas (Hi):** Proporción de datos que están por encima o por debajo de un valor determinado del intervalo i.

    $$H_i = \sum_{j=1}^{i} h_j = h_1 + h_2 + \ldots + h_j$$

❖ **Distribución de frecuencias en intervalos:** Se construye una tabla de dos columnas. En la primera se listan los valores de las variables distribuidas en intervalos. En la segunda columna se indica la cantidad de veces que alguno de los valores del intervalo aparece o la proporción que esa cantidad representa en el total.
Los intervalos son siempre cerrados en el extremo izquierdo y abiertos en el derecho [).

- ➢ **Clases:** Intervalos en los cuales se agrupan convenientemente los valores de las variables $[y , y_i]$.
- ➢ **Marcas de clases:** Son los puntos medios del intervalo. El subíndice de la marca coincidirá con el del límite superior.

  $$y_i = \frac{y_{i-1} + y_i}{2}$$

- ➢ **Rango:** Diferencia entre el mayor valor y el menor valor de la serie.

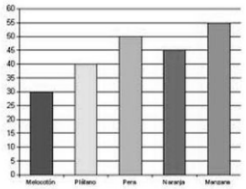
  $$R = y_{max} - y_{min} \quad C_i = \frac{R}{Nº\ de\ intervalos}$$ **Amplitud del intervalo**

- *Las frecuencias son las mismas que la distribución de frecuencias en lista.*

1. *Definir el rango. Ej: R = 88 - 61 = 27*
2. *Definir la amplitud de cada intervalo. Ej: Si quiero 6 intervalos entonces Ci = 27/6 = 4.5 mas o menos 5.*
3. *Definir un nuevo recorrido R' que se llama recorrido ampliado que se calcula como número de intervalos x amplitud. Ej: R' = 6 x 5 = 30.*
4. *Distribuir el R' - R en los extremos de la serie (puede ser de cualquier forma). De esta forma amplio la serie y cualquier valor que caiga en el primer y último intervalo será también contemplado.*

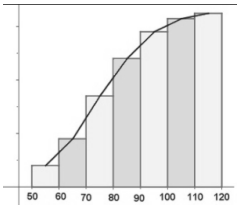
- **Formas de agrupar variables <u>cualitativas</u>:**
  - **Tablas de Contingencia:** Muestra el número o las proporciones de observaciones para cada una de las clases cualitativas.
  - **Representaciones Gráficas:** Tienen que tener título, fuente o nota y número.
    - **Gráficos Lineales:** Se utilizan para los agrupamientos en **lista**.

- **Gráfico de Bastones:** Consiste en un eje de coordenadas donde el eje “x” muestra los valores de la variable y el eje “y” los valores de las frecuencias. Se trazan líneas verticales desde el valor de la variable hasta el de la frecuencia.



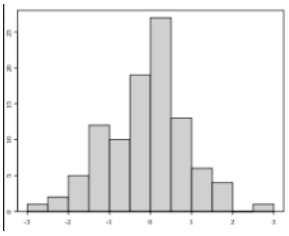
$$Eje\ "x":\ n_i\ -\ h_i\ \mid\ Eje\ "y":\ y_i$$

- **Gráfico Acumulativo de Frecuencias:** Se utiliza para representar las frecuencias absolutas o relativas acumuladas. Es igual al gráfico de bastones pero se trazan líneas horizontales formando una escalera.



$$Eje\ "x":\ N_i\ \mid\ Eje\ "y":\ y_i$$

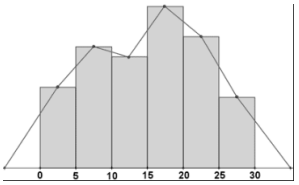
- **Gráficos de Superficies:** Se utiliza para los agrupamientos en intervalos o clases.
- **Histograma:** Se deja un espacio en cada extremo del eje “x”, formando columnas para cada clase con su respectivo valor de **frecuencia**. Los límites de las columnas o barras deben tocarse unas con otras.



$$Eje\ "x":\ \frac{h_i}{n_i}\ \mid\ Eje\ "y":\ y_i$$

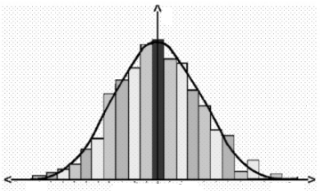
$$Dónde\ comienza\ cada\ barra:\ y_i\ |\ Dónde\ termina\ cada\ barra:\ y_{i+1}$$

- **Polígono de frecuencias:** Se utiliza para comparar y superponer gráficos. Es igual al histograma pero se utilizan como coordenadas las marcas de clase y las frecuencias, habiendo luego una línea que atraviesa los puntos. Sólo es útil para variables cuantitativas. El eje x es la variable y el y es la frecuencia. El área es igual a la del histograma.

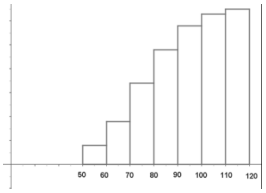


$$Eje\ "x":\ \frac{n_i}{h_i}\ \ |\ \ Eje\ "y":\ y_i$$

- **Curva Suave o Modelo de Población:** Cuando la muestra es muy grande, los intervalos de clase son muy estrechos y el histograma formará prácticamente una curva suave.
  Es importante porque representa la verdadera distribución de la población de la que se extrae la muestra.
  Debe presentarse siempre con el histograma.

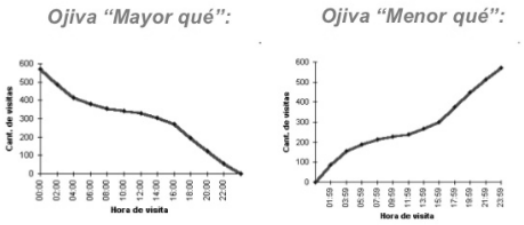


- **Diagrama Escalonado:** Representa a una distribución de frecuencias acumuladas mediante una serie de líneas horizontales trazadas en las correspondientes clases a la altura de las respectivas frecuencias.



$$Eje\ "x":\ N_i\ \ |\ \ Eje\ "y":\ y_i$$

- **Ojiva:** Es un polígono que representa una distribución acumulada en forma de diagrama de líneas. Puede ser:
  - Ojiva Mayor que: y_{i-1}
  - Ojiva Menor que: y_i

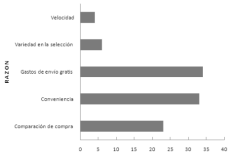


- **Curva acumulativa suavizada:** Con una muestra lo suficientemente grande se puede suavizar el diagrama escalonado o la ojiva para representar una distribución de población.

- **Gráficos especiales:**

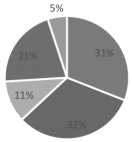
➔ **Distribuciones categóricas:** Para variables categóricas.

- **Diagrama de Barras Horizontales:**



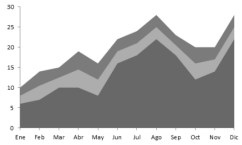
- **Diagrama de barras de Componentes de porcentajes:** El rectángulo es el 100% del fenómeno.

- **Diagrama de pastel:** Representa porcentajes.

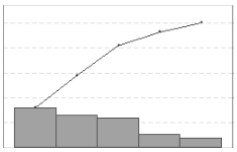


➔ **Otros diagramas:**

- **Gráfico de zonas:** Sirve para comparar en un mismo gráfico varios fenómenos.

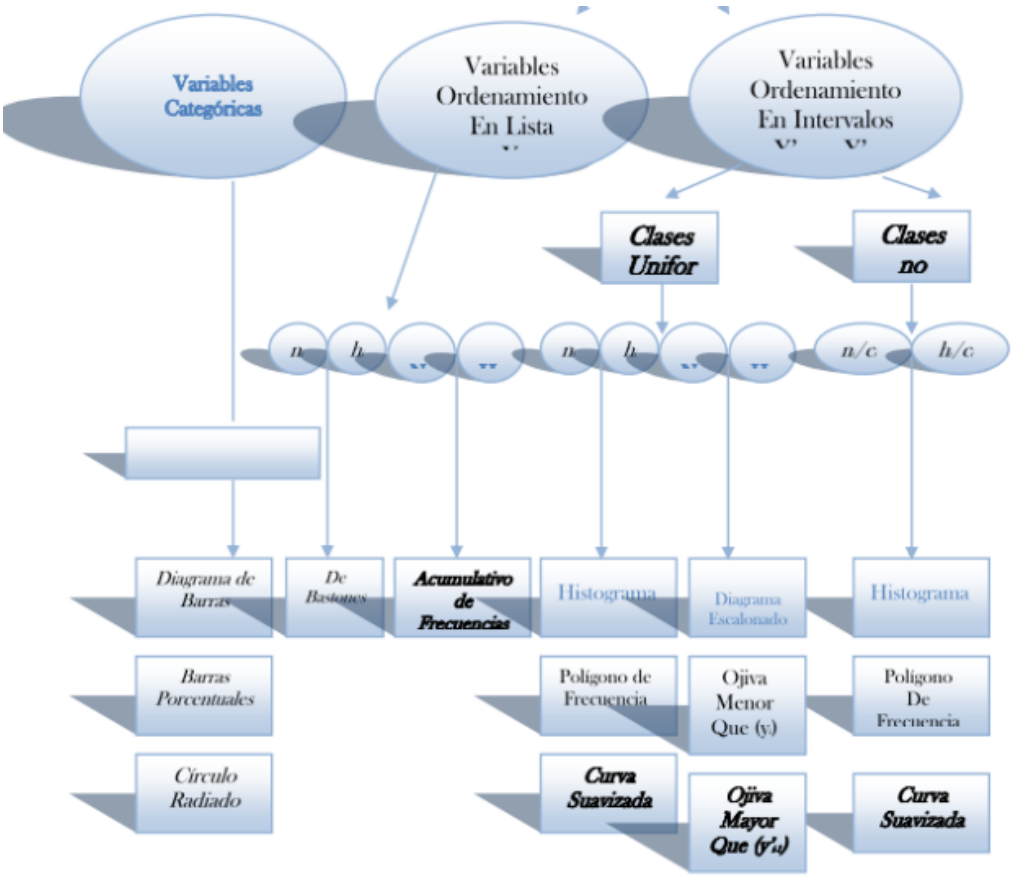


- **Diagrama de pareto:**



- **Diagrama de tallo y hoja:**

Tallo	Hoja
4	4 5 9
5	0 2 3 3 4 4 6 7 7 7 8
6	1 2 2 3 4 7 8 9
7	0 1 1 2 3 4 4 5 6 6 8 9
8	0 1 3 5



Unidad 3
Descripción de Datos Estadísticos

Las **medidas descriptivas** son aquellas que nos ayudan a resumir aún más el comportamiento de los datos bajo estudio para obtener un conocimiento más preciso de los datos que el que se obtiene a partir de tablas y gráficas. Se usan cuando se quiere proceder al análisis de datos previamente recolectados. No basta con organizar los datos a través de arreglos de distribuciones de frecuencias, ni tampoco el presentarlos a través de gráficas que permiten su comportamiento.
Toda medida que se calcula de una muestra se llama **estadístico** y de una población **parámetro.**

- **Medidas de posición o Medidas de Centralización:** Nos da un centro de distribución de frecuencias. Es un valor que se puede tomar como representativo de todos los datos, sea una muestra o una población.

    - **Media Aritmética M(x):** Indica la localización del centro de la distribución, la tendencia central se define y calcula dividiendo la suma de los valores de la variable por el número de observaciones (un promedio).
      La media siempre se puede calcular para un conjunto de números. Existe una media única para un conjunto dado de números. La media es sensible (o afectada) o cada valor del conjunto, por lo que, si cambia algún valor la media también cambiará.

- **Series simples:** $M(x) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **Datos Agrupados:** $M(y) = \bar{y} = \frac{\sum_{i=1}^{m} Y_i . n_i}{n}$

  sí $\frac{n_i}{n} = h_i \Rightarrow \bar{y} = \sum_{i=1}^{m} Y_i . h_i$
- **Distribución de frecuencia de variables continuas:**

  $M(y) = \bar{y} = \frac{\sum_{i=1}^{m} y_i . n_i}{n} = \sum_{i=1}^{m} y_i . h_i$ siendo Yi la marca de clase
- **Propiedades:**
  - **1.** La suma de las desviaciones de las variables respecto a la media aritmética es igual a cero.

    $\sum_{i=1}^{m} (y_i - \bar{y}) . n_i = 0 \quad | \quad \sum_{i=1}^{n} (x_i - \bar{x}) = 0$
  - **2.** $\sum_{i=1}^{m} (y_i - \bar{y})^2 n_i < \sum_{i=1}^{m} (y_i - a)^2 . n_i$
  - **3.** La Media de una cte es la misma cte: $M(k) = k$
  - **4.** La Media de un producto de una cte por una variable, es igual, al producto de la cte por la media de la variable:

    $M(y.k) = k . \bar{y}$
  - **5.** La Media de una suma de una variable más una cte es igual a la media de la variable más la cte:

    $M(x + k) = M(x) + k$
  - **6.** $M(\sum_{i=1}^{k} x_i) = \sum_{i=1}^{k} M(x_i)$

Mediana Me(x): Es el valor central de la variable cuando los datos están ordenados por su magnitud. Divide a un conjunto de datos ordenados en dos grupos iguales, la mitad serán menores a la mediana y los demás mayores.

- **Series Simples:** Cuando la posición es par se promedian los dos valores centrales y se dice que es una mediana no real.

  Posición: $(\frac{n+1}{2})^{o}$

  1. Ordenar los datos.
  2. Contar para saber si existe un número de datos par o impar.

3. En el caso de que sea impar la mediana vendrá a ser el valor intermedio. Por el contrario, el promedio de los valores intermedios.

- **Moda Md(x):** Es el valor más frecuente. Si existe más de un valor con la misma frecuencia, siendo esta la más alta, se dice que es una distribución multimodal pero se vuelve difícil interpretarlos y compararlos. Si ningún valor se repite más de una vez se dice que no se puede determinar la moda.
  En la distribución por lista: $md = y_j \ \ si \ \ n_{j-1} < n_j < n_{j+1}$
  En la distribución de intervalos: Se utiliza el histograma.
  Los valores de los extremos no afectan a la moda.

- Las distribuciones diamétricas que sólo contienen una moda siempre tienen el mismo valor para la media, la mediana y la moda.
- Cuando la población está sesgada negativa o positivamente, la mediana suele ser la mejor medida de posición, debido a que siempre está entre la media y la moda.
- En cualquier otro caso, no existen guías universales para la aplicación de la media, la mediana o la moda como medidas de tendencia central para distintas poblaciones. Cada vaso deberá considerarse de manera independiente, de acuerdo con las líneas generales que se analizaron.

- **Fractiles:** En una distribución de frecuencia cierta cantidad de los datos cae en un fractil o por debajo de este. Estos subdividen una distribución de mediciones de acuerdo con la proporción de frecuencias observadas. La mediana divide la distribución en dos, los cuadriles en cuatro y así, todos estos son fractiles.
  Fractiles: 4 ; Deciles: 10; Percentiles: 100

$$Q_k = \left[k\,.\,(\tfrac{n}{4})\right]^{o} ; \left[k\,.\,(\tfrac{n+1}{4})\right]^{o}$$

- **Medidas de Dispersión:** Indica el grado de variación entre los valores de la serie de los datos recopilados en relación a alguna medida de posición (que tan cerca o que tan lejos están los datos de las medidas de posición). Es utilizada como complemento a las medidas de posición. Cuanto mayor sea este valor mayor será la variabilidad y cuando menor sea más homogénea sea la media.

  - **Recorrido o Amplitud:** Es una <u>medida de dispersión absoluta</u> que indica la diferencia entre el valor máximo y el valor mínimo de los valores observados:

$$R = x_{max} - x_{min} = y_{max} - y_{min}$$

    Si dos valores tienen el mismo rango la media va a dar igual aunque la forma en la que están distribuidos sea distinta, por esto es poco significativa.

  - **Desviación media:** Es la media aritmética o promedio de los valores absolutos de las desviaciones de los valores de la variable con respecto a una medida de tendencia central (media, moda, mediana). Es una <u>medida de dispersión absoluta.</u>

$$DM_{\bar{x}} = \frac{\sum_{i=1}^{m} \left|y_i - \bar{x}\right| . n_i}{n} \quad \text{ó} \quad DM_{\bar{x}} = \frac{\sum_{i=1}^{n} \left|x_i - \bar{x}\right|}{n}$$

- **Varianza:** Es la media aritmética de los cuadrados de las desviaciones con respecto a la media aritmética de la distribución. Es una medida de dispersión absoluta.
  No da mucha precisión con respecto a la dispersión porque los valores están elevados al cuadrado.

  $V(X) = \sigma_x^2$ ⇒ Población

  $V(x) = s_x^2$ ⇒ Muestra

  - **Series Simples:**

    $$V(x) = M\left[(x_i - \bar{x})^2\right] = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

  - **Datos Agrupados:**

    $$V(y) = M\left[(y_i - \bar{y})^2\right] = \frac{\sum_{i=1}^{m} (y_i - \bar{y}) \cdot n_i}{n}$$

    ni es la frecuencia de cada valor (o cantidad de veces que se repite).

  - **Propiedades:**
    - $V(y) \geq 0$ (la varianza es siempre positiva).
    - $V(k) = 0$ (la varianza de una cte es siempre 0 porque no hay variabilidad en los datos).
    - $V(k \cdot y) = k^2 \cdot V(y)$ (si los valores de la variable se modifican por una cte la varianza queda multiplicada por el cuadrado de esa cte).
    - $V(k \pm y) = V(y)$ (si a los valores se le suma una cte la varianza no se modifica).
    - $V(x \pm y) = U(x) + V(y)$

- **Desviación Estándar:** Es el valor positivo de la raíz cuadrada de la varianza. Sirve para poder comparar el valor de la desviación con los valores de las variables. Es una medida de dispersión absoluta. Representa la medida del grado de dispersión de los datos con respecto al valor promedio o media aritmética.

  $DS(X) = \sigma_X$ ⇒ Población

  $DS(x) = s_x$ ⇒ Muestra

  $$DS(x) = \sqrt{V(x)}$$

  Propiedades:

  - $D(Y) \geq 0$
  - $D(k) = 0$
  - $D(k \pm y) = D(y)$
  - $D(k \cdot y) = k \cdot D(y)$
  - $D(x \pm y) = \sqrt{V(x) + V(y)}$

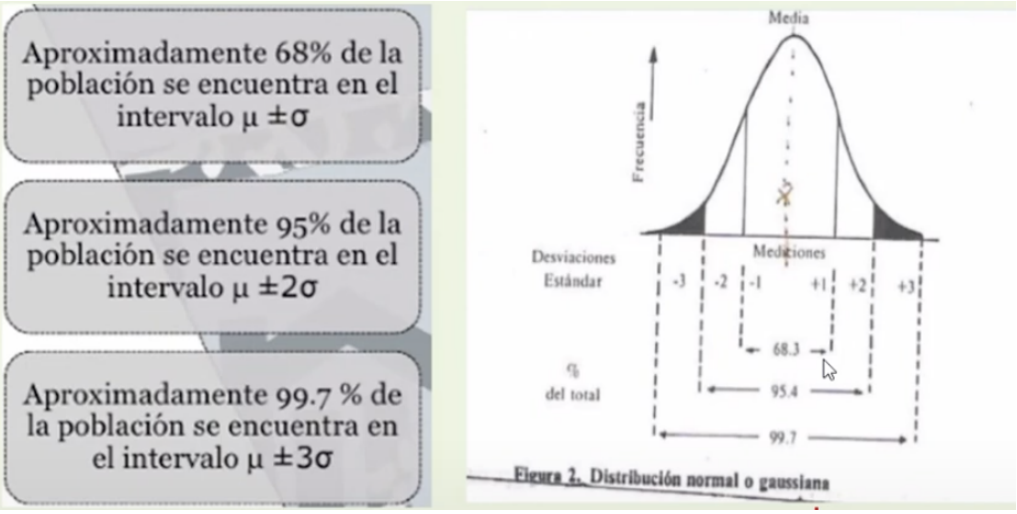
Es el parámetro de dispersión más utilizado.

Es afectada por el valor de cada observación.

Como consecuencia de considerar desviaciones cuadráticas pone mayor énfasis en las desviaciones extremas que en las demás desviaciones.

Al construir la tabla de frecuencias de una variables discreta y calcular a partir de ella la desviación típica, no hay pérdida de información por lo que la desviación para los datos observados es igual que para los datos tabulados.

Tanto la varianza como la desviación requieren de otro conjunto de datos para poder compararlos y hacer un análisis de la concentración de los datos. Si no tengo otro conjunto de datos uso la **regla empírica**:



Si tomamos como parámetro la media la regla empírica me dice que el 68% de los datos están agrupados entre +-1 la desviación. El 95.4% entre +-2 la desviación y el 99.7% entre +-3 la desviación.

- **Coeficiente de Variación:** Es una <u>medida de dispersión relativa</u>, nos da una relación entre el tamaño de la media y la variabilidad de la variable. Nos permite comparar la variabilidad de dos o más conjuntos de datos, ya que elimina la dimensionalidad de las variables. Es el cociente entre la desviación tipica y la media aritmética de una distribución.

  - **Series Simples y Datos Agrupados:** $CV(x) = \frac{DS(x)}{M(x)}(100\%)$

    Si el CV es menos o igual al 20% (0.2), entonces el promedio es representativo, y los datos son homogéneos.
    Es importante que todos los valores sean positivos y la media también.
    Mientras mayor es más heterogéneo (hay más dispersión y la curva de gauss va a estar más aplanada) y mientras menor es más homogéneo.

- **Variable desvío estandarizado:** Se utiliza para comparar dos individuos del mismo conjunto de datos.

$$Z_i = \frac{y_i - \bar{y}}{DS(Y)}$$

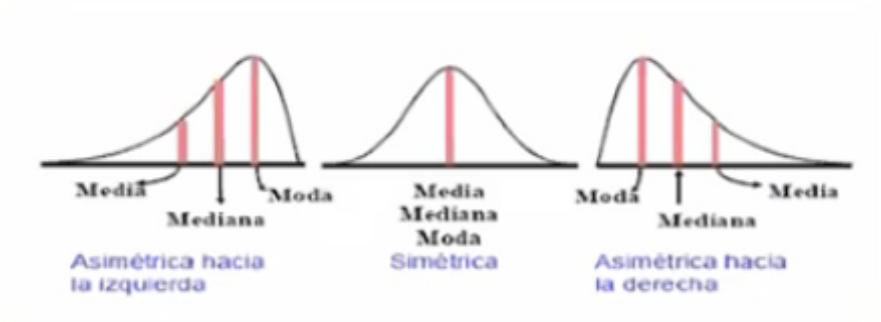
- **Varianza corregida:**

$$\hat{s}^2 = s^2 \frac{n}{n-1} \quad \Rightarrow \text{Muestral}$$

$$\hat{s}^2 = \sigma^2 \frac{N}{N-1} \quad \Rightarrow \text{Poblacional}$$

Medidas de Forma de la distribución:

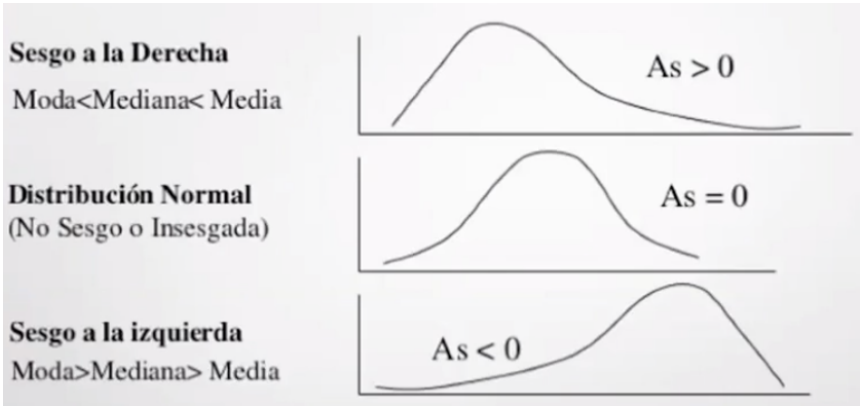
- **Medidas de Asimetría:** Permiten identificar y describir la manera en que los datos tienden a reunirse en función de su frecuencia. Pueden ser:
  - **Simétrica:** Se da cuando en una distribución se reúne aproximadamente la misma cantidad de los datos a ambos lados de la media aritmética, en forma de campana, llamada Campana de Gauss.
    Es simétrica cuando la media, la moda y la mediana son iguales.
    $$\bar{x} = Me = Md$$
  - **Asimetría Negativa:** Se da cuando en una distribución la minoría de los datos está en la parte izquierda de la media, es decir, cuando la media es menor que la mediana y esta menor que la moda.
    $$\bar{x} < Me < Md$$
  - **Asimetría Positiva:** Se da cuando en una distribución la minoría de los datos está en la parte derecha de la media, es decir, cuando la media es mayor que la mediana y esta mayor que la moda.
    $$\bar{x} > Me > Md$$



  - **Coeficiente de Asimetría:** Para estandarizar la medición de la asimetría, quitándole las unidades de medida y el grado de variabilidad, se utiliza el Coeficiente de Pearson:
    $$K_p = \frac{3\,(\bar{x} - Md)}{\hat{s}}$$
    Donde $\hat{s}$ es la desviación estándar muestral corregida.

Cómo vimos: (Método empírico) $Md = \bar{x} - 3(\bar{x} - Me)$
Podemos sacar la moda de la función ya que: $\bar{x} - Md = 3(\bar{x} - Me)$
Quedando: $K_p = \frac{3(\bar{x} - Me)}{\hat{s}}$
El cual:

- Sí $K_p < 0 \Rightarrow$ Asimetría negativa
- Sí $K_p = 0 \Rightarrow$ Simetría
- Sí $K_p > 0 \Rightarrow$ Asimetría positiva

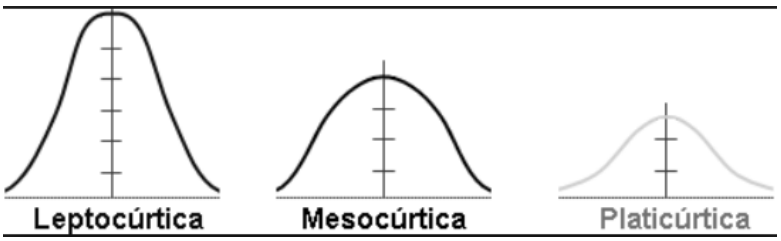


- **Medida de Puntiagudez: Coeficiente de Curtosis de Fisher:**
  La curtosis determina el grado de concentración que presentan los valores alrededor de la media de la distribución de datos (se fija en cómo sube o baja la curva).
  Partiendo de que en una distribución normal se verifica:

$$\mu = \frac{\sum_{i=1}^{m} (xi - \bar{x})^4}{N \cdot s4} - 3$$

N es el número de datos; s la desviación estándar; $\bar{x}$ la media

  - **Leptocúrtica:** Un elevado grado de concentración alrededor de los valores centrales. $k_4 > 0$.
  - **Mesocúrtica:** Un grado de concentración medio alrededor de los valores centrales. $k_4 = 0$.
  - **Platicúrtica:** Un reducido grado de concentración alrededor de los valores centrales. $k_4 < 0$.



Unidad 4
Teoría de Probabilidades

- **Teoría de las Probabilidades:** Es la base para la estimación estadística y la toma de decisiones a través de la docimasia de hipótesis. Además, se utiliza para todos aquellos problemas donde interviene la incertidumbre.
  En la estadística se saca un puñado de la caja negra. En la probabilidad se predice que se va a extraer de la caja transparente.
  Un suceso aleatorio es aquel fenómeno que ejecutado bajo las mismas condiciones puede tener varios resultados y en cada ocurrencia del suceso es imposible predecir el resultado, atribuyéndose el mismo al azar. Si lanzamos una moneda puedo tener cara o cruz como resultados pero no sabemos qué puede salir.

  Un **experimento** es una observación de un fenómeno que ocurre en la naturaleza. Pueden ser:
  - Determinísticos: Son aquellos en donde no hay incertidumbre acerca del resultado que ocurrirá cuando éstos son repetidos varias veces.
  - Aleatorios: Son aquellos en donde no se puede anticipar el resultado que ocurrirá, pero si tiene una completa idea acerca de todos los resultados posibles del experimento cuando este es ejecutado.

- **Espacio Probabilístico o Muestral (Ω):** Es el conjunto de todos los resultados posibles de un experimento aleatorio. Equivale al conjunto universal. Los elementos de este conjunto (llamados puntos muestrales) deben ser mutuamente excluyentes y colectivamente exhaustivos.
  - **Espacios Muestrales Discretos:** Son espacios muestrales cuyos elementos resultan de hacer conteos, y por lo general son subconjuntos de los números enteros.
  - **Espacios Muestrales Continuos:** Son espacios muestrales cuyos elementos resultan de hacer mediciones.

- **Eventos:** Es un resultado particular de un experimento aleatorio (es un subconjunto del espacio muestral), se representa cómo E_0, E_1, … , E_n.

    - **Evento Simple:** Contiene exactamente un elemento del espacio probabilístico. Ej: Que salga el número 1 al lanzar un dado.
    - **Evento Compuesto:** Contiene más de un elemento del espacio probabilístico. Se presenta si algún elemento ocurre. Ej: Que salga un número par al lanzar un dado.
    - **Evento Imposible:** Un conjunto que no contiene elementos del espacio probabilístico. Ej: Que salga el número 7 al lanzar un dado.
    - **Evento Cierto:** Contiene todos los eventos elementales, o sea, es el espacio muestral. Su probabilidad es igual a 1. Ej: Que salga un número al lanzar un dado.

- **Unión de eventos:** Es el evento que contiene los elementos que están en A o en B o en ambos. Ocurre si al menos uno de los dos eventos ocurren.
- **Intersección de eventos:** Es el evento que contiene los elementos que están en A y en B al mismo tiempo. Ocurre si los dos eventos ocurren simultáneamente.

❖ **Evento no mutuamente excluyentes:** Aquella que tienen elementos en común, su intersección es distinta al vacío $E_1 \cap E_2 \neq \emptyset$. Estos pueden ser:

- **Dependientes:** La ocurrencia o no de un m evento afecta a la probabilidad del otro.
- **Independientes:** La ocurrencia o no de un m evento no afecta a la probabilidad del otro.

❖ **Evento mutuamente excluyentes:** Aquella que NO tienen elementos en común, su intersección es igual al vacío $E_1 \cap E_2 = \emptyset$.

❖ **Eventos colectivamente exhaustivos:** Cuando la unión de dos eventos es igual al espacio probabilístico $E_1 \cup E_2 = \Omega$.

- **Teorías Probabilísticas:**

❖ **Teoría Clásica:** La define Jakob Bernoulli (1713): “Una fracción en la que el numerador es igual al número de apariciones del suceso y el denominador es igual al número total de casos en los que ese suceso pueda o no ocurrir. Tal fracción expresa la probabilidad de que ocurra ese suceso”.

$$P(E) = \frac{Números\ de\ casos\ favorables\ a\ E}{Números\ de\ casos\ posibles\ en\ \Omega} = \frac{n(E)}{n(\Omega)}$$

❖ **Teoría Frecuencial:** Bernoulli introdujo el concepto de **probabilidad frecuentista o estadística**: Asignar como probabilidad de un suceso el resultado que se obtendría si el proceso se repitiera en condiciones similares un gran número de veces, ideando la Ley de los grandes Números, en base a que era consciente de que las frecuencias observadas se acercaban a un cálculo previo de su probabilidad al aumentar el número de repeticiones del experimento.

Si un experimento es ejecutado “N” veces en las mismas condiciones y hay “n” resultados en que ocurrió el hecho ($n \leq N$) entonces hay una probabilidad de ese hecho que es: $\frac{ni}{n} = hi.$

❖ **Teoría Subjetivista o personalista:** En este caso la probabilidad mide el grado de creencia de un individuo en la verdad de una preposición variando entre 0 y 1. Basada en la confianza personal del investigador. Es un método subjetivo.

- **Propiedades para la probabilidad de eventos:**
  - A es un evento entonces existe un número para $P(A)$.
  - $P(A) \geq 0$para todo evento A.
  - $P(\Omega) = 1$.
  - Si $A \cap B = \emptyset$entonces $P(A \cup B) = P(A) + P(B)$.
  - Si $A \subset B$ entonces $P(B - A) = P(B) - P(A)$.
  - Ley de complementación: $P(\overline{A}) = 1 - P(A)$.
  - $0 \leq P(A) \leq 1$.
  - Si $A \subset B \Rightarrow P(A) < P(B)$.
  - Desigualdad de Boole: Si $A \cap B \neq = \emptyset \Rightarrow P(A \cup B) \leq P(A) + P(B)$

  - **Propiedad Total:** La probabilidad de la unión de dos eventos.
    - Si los eventos son <u>no mutuamente excluyentes</u> es igual a la suma de sus probabilidades menos la probabilidad de la intersección de dichos eventos.
      Si $A \cap B \neq \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - Si los eventos son <u>mutuamente excluyentes</u> es igual a la suma de sus probabilidades de dichos eventos.
      Si $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

- **Tipos de Probabilidad:**

  - **Probabilidades Condicionales:** Se utiliza cuando tengo cierta información a priori, ya que, es posible disponer de información que reduce el espacio probabilístico original a un subconjunto. Se simboliza$P(A/B)$, lease probabilidad de que ocurra “A” dado “B”.

    $$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

    Se puede ver también como una probabilidad conjunta sobre la marginal.

  - **Probabilidad Compuesta o Conjunta:** Son los valores del cuerpo de una tabla de doble entrada de probabilidades de eventos. Son los eventos de intersección
    - Si los eventos son <u>dependientes:</u> La ocurrencia de A afecta a la de B $P(A \cap B) = P(B) \,.\, P(A/B) = P(A) \,.\, P(B/A)$. Se calcula despejando la probabilidad condicional.
      Para más de dos hechos: $P(ABC) = P(A) \,.\, P(B/A) \,.\, P(C/AB)$.
    - SI los eventos son <u>independientes:</u> La ocurrencia de A no afecta a la de B entonces: $P(A \cap B) = P(A) \,.\, P(B)$

- **Probabilidad Marginal:** Se obtiene sumando una fila o una columna de la tabla: $P(A) = P(AC) + P(AD) = P(A \cap C) + P(A \cap D)$

	A	B	
C	P(A∩C)	P(B∩C)	
D	P(A∩D)	P(B∩D)	
	P(A)	P(B)	1

- **Dependencia e Independencia estadística:** Dos eventos son dependientes si la probabilidad de ocurrencia es afectada por la ocurrencia del otro. En general, este tipo de eventos son generados por el muestreo sin reposición.
  Se determina: Por hechos **independientes** entre sí, si la probabilidad de la ocurrencia conjunta de A y B es igual al producto de sus respectivas probabilidades independientes: $P(A \cap B) = P(A) \,.\, P(B)$.
  Si esta igualdad no se obtiene los eventos son **dependientes.**
  Esto surge de: $P(A/B)$debe ser igual a $P(A)$:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = P(A)$$
$$P(A \cap B) = P(A) \,.\, P(B)$$

- **Teorema de Bayes:** Establece una relación entre las probabilidades condicionales con el resto de las probabilidades para cuando no se cuenta con todos los datos en la tabla:

$$P(C/A) = \frac{P(C \cap A)}{P(A)} = \frac{P(A/C) \,.\, P(C)}{P(A/C) \,.\, P(C) + P(A/D) \,.\, P(D) + \ldots}$$

Utilizando una probabilidad condicional y aplicando la regla del producto en el numerador y probabilidad total en el denominador se obtiene la regla de Bayes que permite calcular fácilmente probabilidades condicionales, llamadas a posteriori siempre cuando se conozca la probabilidad a priori $P(C)$y las condicionales.

Unidad 5
Variables aleatorias y distribuciones de probabilidad

- **Variable Aleatoria (x):** Es una función que asigna un número real, a cada resultado del espacio muestral de un experimento aleatorio. Es una función X definida $X: \Omega \rightarrow IR$, o sea, una función cuyo dominio es el espacio muestral (como si trabajamos con una población) y el rango es el conjunto de los números reales.
  Representará al conjunto de resultados posibles (van a estar todos porque trabajo con la población), es decir, los valores que puede asumir una nueva variable llamada variable aleatoria y cada valor de dicha variable tendrá asociado una probabilidad de presentación.
  El espacio muestral en muchas ocasiones, no está constituido por números (variables categóricas), pero a través de la variable aleatoria se puede expresar en forma numérica todo tipo de espacio muestral lo que facilita el análisis. A estos valores de la variable le voy

a poder calcular la probabilidad de que se presente ese valor, esto es la distribución de probabilidad de una variable aleatoria.
Ejemplo: Secuencia del sexo de los dos primeros bebes que nacen en un Hospital. El dominio sería $\Omega = \{MM, MF, FM, FF\}$(son los resultados posibles) y $x = \{número\ de\ masculinos\}$ entonces si $x = MM \Rightarrow X(w) = 2$; si $x = MF \Rightarrow X(w) = 1; x = FF \Rightarrow X(w) = 0$ entonces mi rango es el conjunto {0, 1, 2}.

Distribuciones de Probabilidad: La probabilidad de los eventos se traslada a los valores de las variables aleatorias. Entonces una distribución de probabilidad muestra a través de una tabla, gráfico o fórmula todos los valores posibles que puede asumir una variable aleatoria (o sea su comportamiento) y su correspondiente probabilidad de presentación (valor entre 0 y 1).

Como ya tengo una variable aleatoria numerica la misma puede ser:

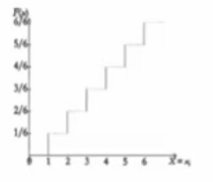
- **Variable Aleatorias Discretas:** Cuando tienen una cantidad numerable de valores o infinitos pero de valores reales.
  - **Puntual**: Se calcula con la Función de Probabilidad (o de Cuantía).
  - **Acumulada**: Se calcula con la Función de Distribución (o de Acumulación).
- **Variable Aleatorias Continuas:** Cuando pueden asumir cualquier valor real o cuando la diferencia entre un valor y otro es de un infinitésimo. No se pueden expresar para un valor puntual, sino que se trabaja con intervalos de valores, entonces cuando se calcula una probabilidad, se hace en función de un área de valores y se trabaja integrando. Acá es donde aparece:
  - **Función de densidad**: Sirven para trabajar con variables aleatorias con probabilidades por intervalos, y su integral coincide con esa probabilidad.

- **Función de Cuantía:** Permite obtener la probabilidad para un valor exacto de la variable aleatoria discreta. $P(x_1 = x_i)$ indica la probabilidad de que la variable aleatoria x tome un valor real x_i lo cual implica que se presente un determinado evento representado por x_i .
$\sum_{i}^{n} P_i = 1$ para $P_i \leq 0$.
Se puede expresar a través de tablas, gráficos o una expresión simbólica.

- **Función de distribución o de acumulación:** $P(x \leq x_i)$es la probabilidad de que la variable aleatoria discreta x asuma valores menores o iguales a x_i.
Esta probabilidad es la suma de las probabilidades puntuales de todos los valores menores o iguales a x_i. Es la probabilidad del intervalo definido por $x \leq x_i$.

$$F(x_i) = P(x \leq x_i) = \sum_{x_j \leq x_i} P_{x_j}$$

Se puede expresar a través de tablas, gráficos o una expresión simbólica.



Se trata de una función creciente donde:

$$F(0) = Pr(x \leq 0) = 0$$
$$F(n) = Pr(x \leq n) = 1$$

Y donde:

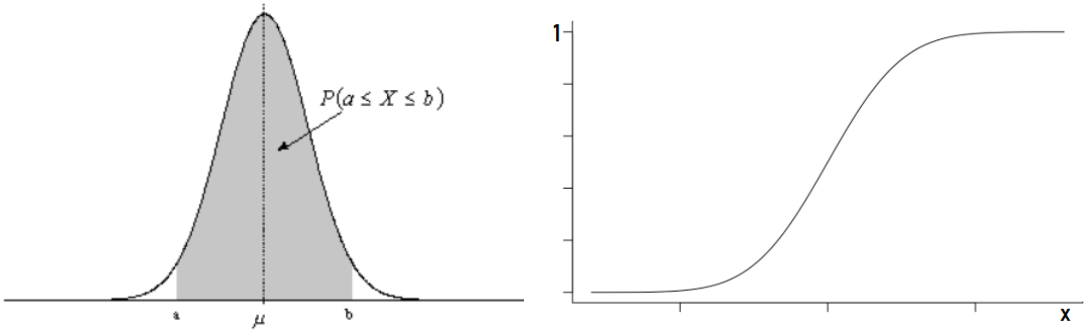
$$P(x \geq a) = 1 - P(x \leq a)$$
$$P(a \leq x \leq b) = P(x \leq b) - P(x \leq a)$$

- **Función de densidad:** Una variable continua puede tomar un valor fraccionario en un determinado rango de valores, entonces cómo va a existir un número infinito de mediciones no puedo calcular una única probabilidad para esos valores, así que se define la función de densidad. Por esto se habla de un área debajo de una curva que denota la densidad de probabilidad.
  Existe una función no negativa $f(x)$ llamada función de densidad, que satisface la siguiente relación para todo valor real de x:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx \quad \Rightarrow \frac{dF(x)}{dx} = F'(x) = f(x)$$

  $F(x)$es la función de acumulación. Si calculo la probabilidad en un punto me daría 0.
  Cuando hablamos de una variable aleatoria continua se define para un intervalo:

$$P(a \leq x \leq b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$$



Parámetros en las distribuciones de Probabilidad: Son medidas que se calculan con los valores posibles que una variable puede asumir en las distribuciones de probabilidad y se utilizan para caracterizar el fenómeno. Me ayudan a analizar mejor el comportamiento de la variable aleatoria. Son parámetros porque trabajo con todos los valores posibles de la variable aleatoria.

- **Esperanza matemática o valor esperado:** Es el promedio de un fenómeno aleatorio. Describe la tendencia central de la variable.

    - **Variables Aleatorias Discretas:** Es la suma de los productos de todos los posibles valores de la variable aleatoria por sus respectivas probabilidades.

      $$E(x) = \mu_x = \sum_{i=0}^{N} x_i . P_i$$

    - **Variables Aleatorias Continuas:** Cuando se trata de una variable aleatoria continua la suma se transforma en integral.

      $$E(x) = \mu_x = \int_{-\infty}^{\infty} x . f(x)\, dx$$

    - **Propiedades:**
        - $E(cte) = cte$
        - $E(c.x) = c . E(x)$
        - $E(c \pm x) = E(x) \pm c$
        - $E(x \pm x) = E(x) \pm E(x)$
        - $E(x . y) = E(x) . E(y)$

- **Varianza:** Es una medida de dispersión que permite conocer la concentración de los datos alrededor de la esperanza correspondiente a dicha distribución.

    - **Variable Aleatorias Discretas:** Es la sumatoria de la desviación cuadrada de los valores de dicha variable con respecto a su media multiplicada por la probabilidad de ocurrencia del evento:

      $$V(x) = \sigma^2 = \sum_i \left[x_i - E(x)\right]^2 . P_i = E(x^2) - [E(x)]^2$$

      Tiene como valor mínimo el 0.

    - **Variable Aleatorias Continuas:** Es la integral de los valores de la variable al cuadrado, multiplicados por la función valuada en el punto, menos la esperanza al cuadrado:

      $$V(x) = \sigma^2 = \int_{-\infty}^{\infty} [x - E(x)]^2 . f(x)\, dx = \int_{-\infty}^{\infty} x^2 . f(x)\, dx - E(x)^2$$

    - **Propiedades:**
        - $V(x) \geq 0$
        - $V(c) = 0$
        - $V(c . x) = c^2 . V(x)$
        - $V(c \pm x) = V(x)$
        - $V(x \pm y) = V(x) + V(y)$

- **Desviación Estándar:** Es la raíz cuadrada positiva de la varianza $\sigma = \sqrt{\sigma^2}$.

- **Momentos en las distribuciones de Probabilidad:** Son la expectativa o esperanza de diferentes potencias de la variable aleatoria. Se divide en:
    - **Momentos naturales de orden K:** Tienen como objetivo calcular la esperanza de $x,\ x^2,\ x^3,\ ....,\ x^k$.

$$a_k = E(x^k) = \sum_i x_i^k \,.\, P(x_i)$$

    - **Momentos centrados de orden K:**

$$m_k = \mu^k = \left[E(x_i - E(x))\right]^2$$

        - **Cuando tienen orden par miden dispersión:** $k^3 = \frac{\mu^3}{\sigma^3}$
            - $k^3 = 0 \Rightarrow$ Simétrica
            - $k^3 < 0 \Rightarrow$ Asimetría Negativa
            - $k^3 > 0 \Rightarrow$Asimetría Positiva
        - **Cuando tienen orden impar miden asimetrías:** $k^4 = \frac{\mu^4}{\sigma^4}$
            - $k^4 = 0 \Rightarrow$ Mesocúrtica
            - $k^4 < 0 \Rightarrow$ Platicúrtica
            - $k^4 > 0 \Rightarrow$Leptocúrtica

Unidad 6
Modelos especiales de Probabilidad para variables aleatorias discretas

- **Modelo Probabilístico:** Nos permite decir que, dadas ciertas condiciones iniciales, ocurrieron ciertos eventos con determinadas probabilidades, ósea, son modelos matemáticos apropiados para situaciones del mundo real en condiciones específicas. Son importantes porque ayudan a predecir la conducta de futuras repeticiones de un experimento.

    - **Modelo de Bernoulli o Bipuntual:** Se aplica a una variable que puede asumir solo dos valores, éxito o fracaso, por ello, se habla de población dicotómica. Generamos la variable aleatoria, dando el valor “1” a la característica estudiable (éxito) y “0” a la característica no estudiada (fracaso). Asignando “P” como la probabilidad que la variable tome el valor “1” y “Q” a la probabilidad que tome el valor “0”, siendo Q = 1-P.

y	P(y)

0	Q =1-P
1	P
	1

Función de Cuantía: $f(x) = p^x q^{1-x}$ calculo una probabilidad puntual
Función de Acumulación: No tengo porque realizo la prueba una sola vez.

- **Parámetros:**
  - $E(y) = P$ porque $E(x) = Q * 0 + P * 1$
  - $V(y) = P . Q$ porque$V(x) = p - p^2 = p(1 - p) = p * q$
  - $\sigma = \sqrt{P.Q}$

  Características:
  - Solo puede haber dos resultados mutuamente excluyentes (éxito o fracaso).
  - Las pruebas en las que se obtiene éxito o fracaso son independientes (realizo una sola prueba y los resultados posibles son dos).
  - Las probabilidades de éxito o fracaso son constantes.

- **Modelo Binomial:** Se repite una prueba simple un número "n" de veces bajo las mismas condiciones (son "n" pruebas Bernoulli). En cada prueba sólo pueden presentarse dos alternativas mutuamente excluyentes, éxito y fracaso con probabilidades P y Q=1-P.
  Los "n" ensayos Bernoulli son independientes entre sí, o sea que el resultado de un ensayo no afecta al resultado de los demás (muestreo con reposición). Es importante que todas las pruebas estén realizadas bajo las mismas condiciones. Tanto P y Q se mantienen constantes en cada una de las pruebas. Para asegurarme de que sean constantes debo ver si la población es finita o infinita y si el muestreo es con o sin reemplazo:
  - Si la muestra se toma con reemplazo me da lo mismo que sea una población finita o infinita porque se vuelve a dejar en su lugar.
  - Si la muestra se toma sin reemplazo y es una población finita, el número de unidades va disminuyendo de a uno, lo que hace que la probabilidad aumente. Va a permanecer constante siempre que n<5% s/N.
  - Si la muestra se toma sin reemplazo y es una población infinita es despreciable y no afecta a las probabilidades.

  $$x \sim B(n; P)$$

  - **Función de Cuantía:** Por tratarse de una variable discreta la Función de Probabilidad se llama Función de Cuantía.

    $P(x = x_i; n; P) = C_n^x . P^x . Q^{n-x}$ donde C es el combinatorio.

- **Función de Distribución ó Acumulación:** Al trabajar con “n” pruebas se pueden acumular haciendo la sumatoria.

$$F(x \leq x_i ; n ; P) = \sum_{x=0}^{x_i} C_n^x . P^x . Q^{n-x}$$

- **Parámetros:**
  - **Esperanza:** $E(x) = n . P$ donde n es la cantidad de pruebas.
  - **Varianza:** $V(x) = n . P . Q$
  - **Desviación Estándar:** $\sigma = \sqrt{n . P . Q}$

- **Modelo de Poisson:** Se utilizan para procesos en los que hay una observación por intervalo de tiempo, espacio o volumen, que se caracterizan por el número de éxitos esperados en una unidad específica.
  Es parte de la distribución binomial porque se realiza el experimento un número “n” muy elevado de veces, por lo que, la probabilidad de éxito “p” en cada ensayo es reducida.
  Se la intenta transformar en una Binomial dividiendo en intervalos tan pequeños que la probabilidad de una tercera posibilidad además de éxito y fracaso sea casi nula.
  Se hace $\lambda = n P \Rightarrow$ Intervalo
  Será $P(x = x_i ; \lambda) \Rightarrow$Cantidad de “$x_i$” de éxitos en el intervalo “λ”.

  - **Función de Cuantía:** $P(x = x_i ; \lambda) = \frac{e^{-\lambda} . \lambda^x}{x!}$
  - **Función de Acumulación:** $P(x \leq x_i ; \lambda) = \sum_{x=0}^{x_i} \frac{e^{-\lambda} . \lambda^x}{x!}$
  - **Parámetros:**
    - **Esperanza:** $E(x) = n . P = \lambda$
    - **Varianza:** $V(x) = n . P = \lambda$ es igual a la E(x) porque hay poca variabilidad
    - **Desviación Estándar:** $\sigma = \sqrt{n . P} = \sqrt{\lambda}$

- **Modelo Hipergeométrico:** Se aplica cuando la población es finita de N elementos y la muestra aleatoria se toma <u>sin reposición</u> o sin reemplazo, por lo que la probabilidad cambiará para cada nueva observación.
  Aclaración: Si la muestra es con reposición la probabilidad de obtener un éxito en cada una de las extracciones en *K/N*, es constante y las extracciones son independientes. Si la muestra es sin reposición la probabilidad de éxito es variable y las extracciones dependientes.

  - **Función de Probabilidad (Cuantía):** Tenemos una población finita de “N” elementos dentro de los cuales “K” tiene cierta característica, y “N-K” no la tiene.
    K: Número de elementos de la población que posee la característica.

n: Muestra aleatoria de tamaño n obtenida de la población anterior (MSR).

Basada en: $Pr = \frac{casos\ favorables}{casos\ posibles} = \frac{K}{N}$

Obtenemos $Pr(x = x_i ; N ; K ; n) = \sum_{x=0}^{x_i} \frac{C_k^x . C_{N-K}^{n-x}}{C_N^n}$

Básicamente si tenemos una población, voy a tener *K* éxitos y *N-K* fracasos. Si yo de esa población extraigo una muestra *n* voy a tener *x* éxitos y *n-x* fracasos.

Para obtener una determinada cantidad de éxitos dentro de la muestra necesitamos algunos fracasos.

- **Función de Acumulación:** $P(x \leq x_i ; N ; K ; N) = \sum_{x=0}^{x_i} \frac{C_K^x . C_{N-K}^{n-x}}{C_N^n}$
- **Parámetros:**
  - **Esperanza:** La propiedad aditiva de la esperanza no requiere independencia entre las variables.

    $E(x) = n P = n . \frac{K}{N}$
  - **Varianza:** No es aditiva para las variables dependientes por lo que cuando se trata de un muestreo sin reposición se introduce un factor de corrección de poblaciones finitas "$\frac{N-n}{N-1}$", que cuando "$N \to \infty$" tiende a "1".

    $V(x) = (\frac{N-n}{N-1}) . n. P. Q = (\frac{N-n}{N-1}). n . \frac{K}{N}. (1 - \frac{K}{N})$ Se le aplica un factor de corrección por ser un MSR.
  - **Desviación estándar:** $\sigma = \sqrt{(\frac{N-n}{N-1}) n P Q}$

- **Modelo Uniforme Discreto:** Es una distribución en la cual la probabilidad asociada con los resultados es una constante, todos tienen la misma probabilidad de presentarse. Son "N" resultados mutuamente excluyentes ya igualmente probables.

  $P(x = x_i) = \frac{1}{N}$

- **Modelo de Proporción de Éxitos:** A veces no nos interesa el número de éxitos sino la proporción de los mismos. Si es un MCR voy a ir por un Modelo Binomial y si es MSR por el Modelo Hipergeométrico.
  - $P = \frac{K}{N}$ Es la Proporción Poblacional donde K es el número de éxitos y N el tamaño de la población
  - $p = \widehat{P} = \frac{x}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$ Es la Proporción Muestral donde x es el número de éxitos en pruebas, n el tamaño muestral e y la variable aleatoria Bernolli.

Cómo podemos ver las proporciones se ven afectadas con el número de éxitos, por ende tenemos dos caminos:

- Si el muestreo es <u>con reposición</u> y *x* es Binomial, la probabilidad de que se presente una proporción de éxitos dados en la muestra, es igual a la probabilidad de que se presente ese número de éxitos en la muestra. Entonces, cálculo la probabilidad puntual y acumulada a través de una distribución **Binomial.**
- Si el muestreo es <u>sin reposición</u> y *x* es Hipergeométrica cálculo la probabilidad puntual y acumulada a través de una distribución **Hipergeométrica.**

  - **Parámetros para Binomial:**
    - $E(y) = P$ porque $E(\widehat{P}) = E(\frac{x}{n}) = \frac{1}{n} * E(x) = \frac{1}{n} * n * P = P$
    - $V(y) = \frac{P.Q}{n}$ porque
      $V(\widehat{P}) = V(\frac{x}{n}) = \frac{1}{n^2} * V(x) = \frac{1}{n^2} * n * P * (1 - P) = \frac{P*(1-P)}{n}$
    - $\sigma = \sqrt{(P.Q)}/n$

  - **Parámetros para Hipergeométrica:** Son iguales a la de la Binomial pero a la varianza y a la desviación estándar le tengo que aplicar el factor de corrección porque es un MSR.
    - $V(y) = \frac{P.Q}{n} * \frac{N-n}{N-1}$
    - $\sigma = \sqrt{P * Q/n} * \frac{N-n}{N-1}$

<u>Unidad 7</u>

<u>Modelos especiales de Probabilidad para variables aleatorias continuas</u>

- **Modelos probabilísticos:**
  - **Modelo Uniforme Continuo:** Una variable aleatoria cuyo valor se encuentra solo en un intervalo infinitamente divisible “(a; b)” tiene una distribución uniforme si su función de densidad (probabilidad puntual no existente) es constante en dicho intervalo.
    - **Función de densidad:** Como es uniforme $f(x) = k$

      Y como la probabilidad del universo probabilístico es 1: $\int_{-\infty}^{\infty} f(x)\,dx = 1$

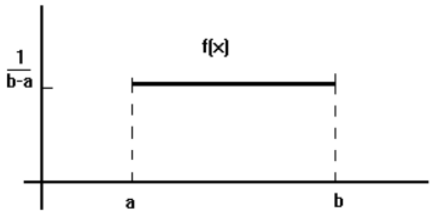
      Si integramos en el intervalo:

$$\int_a^b k\,dk \;=\; [k\,.\,x]\;{}_a^b \;=\; kb \;-\; ka \;=\; k(b \;-\; a\,)$$

Para que como resultado de 1: $\int_a^b \frac{k}{k\,(b-a)}dx \;=\; \frac{k\,(b-a)}{k\,(b-a)}$

Finalmente: $\int_a^b \frac{1}{(b-a)}dx \;=\; 1$

Por lo que la función de densidad f(x) será: $f(x) \;=\; \frac{1}{(b-a)}$



- **Función de distribución o acumulación:**

$$F(x) \;=\int_a^x f(x)\,dx \;=\int_a^x \frac{1}{(b-a)}dx \;=\; \left[\frac{x}{(b-a)}\right]\;{}_a^x \;=\; \frac{x}{(b-a)} \;-\; \frac{a}{(b-a)}$$

$$F(x) \;=\; \frac{x-a}{b-a} \quad 0 \;\Rightarrow\; x \;\leq\; a \quad 1 \;\Rightarrow\; x \;\geq\; b$$

- **Parámetros:**
  - **Esperanza:** Partiendo de que $E(x) \;=\; \int_{-\infty}^{\infty} x\,.\,f(x)\,dx$ para las variables continuas obtenemos: $E(x) \;=\; \frac{a+b}{2}$

    Diferencias de cuadras $b^2 \;-\; a^2 \;=\; (b \;-\; a)\,(b \;+\; a)$
  - **Varianza:** Partiendo de $V(x) \;=\; E(x^2) \;-\; [E(x)]^{\,2}$

    $$E(x^2) \;=\; \frac{b^3 - a^3}{3\,.(b-a)}$$

    $$[E(x)]^{\,2} = \left(\frac{a+b}{2}\right)^{\,2}$$

    $$V(x) \;=\; \frac{b^3 - a^3}{3\,.(b-a)} \;-\; \left(\frac{a+b}{2}\right)^{\,2} \;=\frac{(b-a)^{\,2}}{12}$$
  - **Desviación Estándar:** $\sigma \;=\; \sqrt{\frac{(b-a)^{\,2}}{12}} = \;\frac{(b-a)}{\sqrt{12}}$

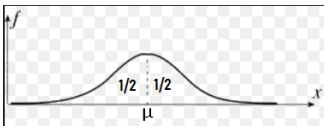
- **Modelo Exponencial:** Se utiliza como modelo para la distribución de tiempos entre la presentación de eventos sucesivos. Existe un tipo de variable aleatoria continua que obedece a una distribución exponencial la cuál se define como el tiempo que ocurre desde un instante dado hasta que ocurre el primer suceso.

  - **Función de Acumulación:** $F(x) = P(x \leq x_i) = 1 - e^{-\lambda x}$
  - **Función de Densidad:** Se dice que una variable aleatoria continua tiene una distribución exponencial con parámetro $\lambda > 0$ y su función de densidad es: $t(x) = \lambda . e^{-\lambda x}$ para $x \geq 0$.
  - **Parámetros:**
    - $E(x) = \frac{1}{\lambda}$
    - $V(x) = \frac{1}{\lambda^2}$
    - $DS(x) = E(x)$

- **Modelo Normal o de Gauss:** Al incrementar el tamaño de la muestra de los modelos Binomial, Poisson e Hipergeométrico se aproximan a la normal.
  La función de densidad es simétrica, mesocúrtica y unimodal (tiene un solo pico), entonces, la media, la mediana y la moda son iguales.
  La regla empírica se basa en el modelo Normal (pág 13).
  Pueden tratarse del "Modelo Normal General" o el "Modelo Normal Estándar".

❖ **Modelo Normal General:** Dada una variable aleatoria continua x que va de menos infinito a más infinito, diremos que tiene distribución aproximadamente normal, con media μ y desviación σ, en símbolos, $(x \sim N(\mu; \sigma))$, si y sólo si su **función de densidad** es:

$$f(x) = \frac{1}{\sigma . \sqrt{2 . \pi}} . e^{-\frac{(x-\mu)^2}{2 . \sigma^2}}$$

Gráficamente: 

- El área de probabilidad bajo la curva es 1 y cada mitad tiene un área de probabilidad de ½ = 0,5.
- El área bajo la curva acotada por un intervalo es la probabilidad de este intervalo.
- Es simétrica con respecto a μ y unimodal. μ=*Md*=*Mo*
- Los extremos de la curva son asintóticas con respecto al eje x.
- μ traslada la curva hacia la derecha si aumenta o hacia la izquierda si disminuye porque es un factor de traslación.

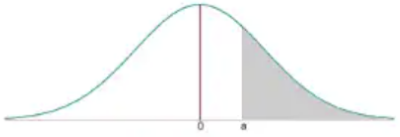
- σ achata la curva porque si aumenta se acumulan los datos en el centro de la distribución (se hace más empinada), si decrece, se extienden a lo largo del eje x (se achata).
- El dominio de $f(x)$ es infinito.
- La probabilidad de que x tome algún valor exacto x_i es 0.
- La curtosis y asimetría es igual a 0.
- Es simétrica con respecto al valor medio:
  $\int_{-\infty}^{\mu} f(x)dx = \int_{\mu}^{\infty} f(x)\,dx = \frac{1}{2}$
- **Parámetros:**
  - **Esperanza:** $E(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx$
  - **Varianza:** $V(x) = \int_{-\infty}^{\infty} [x - E(x)]^2 . f(x)\,dx$
  - **Desviación Estándar:** $DS(x) = \sqrt{V(X)}$

❖ **Modelo Normal Estándar:** A los fines de trabajar con una sola tabla de probabilidades se busca simplificar la función de densidad a través de una variable tipificada $z = \frac{x-\mu}{\sigma}$.

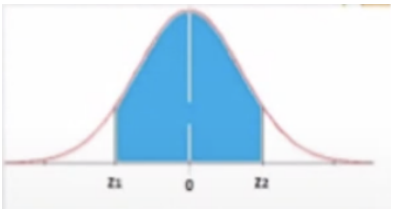
μ siempre va a ser 0 (siempre va a estar centrada) y σ siempre va a ser 1 (altura) entonces el único valor que va a ir variando en la función va a ser z que se va a comportar como $z \sim N(0;\ 1)$ que me deja el Modelo Normal mucho más simplificado.

- **La Función de densidad:** $f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$
- Para calcular probabilidades asociadas a una variable normal estandarizada en la tabla las propiedades tipo $P(Z \leq z)$ se buscan directamente en la tabla:

- Las probabilidades de tipo $P(Z \geq z) = 1 - P(Z \leq z)$



- Las probabilidades de la forma
  $P(z1 \leq Z \leq z2) = P(Z \leq z2) - P(Z \leq z1)$
  A todas las áreas menores a *z2* le resto las menores a *z1*.



- **Parámetros:**
  - **Esperanza:** $E(z) = \int_{-\infty}^{\infty} z\, f(z)\, dz = 0$
  - **Varianza:** $\int_{-\infty}^{\infty} z^2.\, f(z)\, dx - [E(z)]^2$

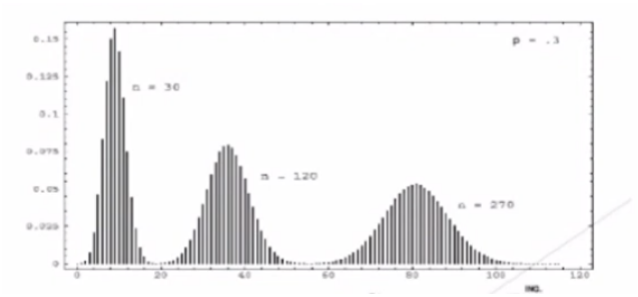
- **Función de Acumulación:** Nos proporciona la probabilidad de que la variable aleatoria asuma un valor que de como resultado Zel cual se encuentra en el intervalo $(-\infty; z_i)$.

$$F(z_i) = P(z \leq z_i) = \int_{-\infty}^{z_i} f(z)\, dz$$

Si queremos encontrar la distribución de un intervalo $[a - b]$ debemos buscar los $z$de a y b.

$$P(a \leq x \leq b) = P(\frac{a-\mu}{\sigma} \leq z \leq \frac{b-\mu}{\sigma})$$

- **Relación entre modelos discretos y modelos normales:** En este caso, se resuelve usando el modelo normal situaciones que normalmente responden a una variable que responde a un comportamiento binomial. Al aumentar el tamaño muestral las distribuciones de modelos discretos se aproximan al modelo normal:



  - **Del Modelo Binomial al Modelo Normal:** Se debe definir $z = \frac{x-\mu}{\sigma}$ reemplazando a μ y σpor su igual en la binomial (n, P, Q) entonces: $z = \frac{x-nP}{\sqrt{nPQ}}$ y a partir de ahí lo calculo a partir del modelo normal.
  Esto se puede hacer siempre que n sea grande y P no esté muy próxima a 0 o a 1 para que sea una campana simétrica.
  Para hacerse la transformación, primeramente debemos aplicar la siguiente corrección:

$$P(a \leq x \leq b) \approx P(a - 0.5 \leq x \leq b + 0.5)$$

Luego, se transforma a Normal: $P(\frac{(a-0.5)-nP}{\sqrt{nPQ}} \leq \frac{x-nP}{\sqrt{nPQ}} \leq \frac{(b+0.5)-nP}{\sqrt{nPQ}})$

- **Del Modelo de Poisson al Modelo Normal:** La distribución normal también es el límite de la distribución de Poisson. Dado el valor de probabilidad P, entonces nP aumenta al aumentar n. A medida que μo nP aumenta la distribución de Poisson se acercará cada vez más a la curva continua acampanada. Se utiliza cuando $\lambda > 20$.
  Si $x \sim P(\lambda) \Rightarrow E(x) = nP \mid V(x) = nP \mid \sigma = \sqrt{nP}$
  Entonces tipificamos x a z: $z = \frac{x-nP}{\sqrt{nP}}$.

- **Del Modelo Hipergeométrico al Modelo Normal:** Si $x \sim H(N, n, X)$

  $$E(x) = n.\frac{X}{N} \quad \text{y} \quad \sigma = \sqrt{\frac{N-n}{N-1}.\, n\frac{X}{N}.(1-\frac{X}{N})}$$

  Por lo que el reemplazar en z será: $z = \frac{x-(n.\frac{X}{N})}{\sqrt{\frac{N-n}{N-1}.n\frac{X}{N}.(1-\frac{X}{N})}}$.

- **Del Modelo de Proporción al Modelo Normal:**
  - Si la proporción muestral proviene de una Binomial y el tamaño de muestra es igual o mayor a 30 si P=Q, tenderá a normal.
    Si $\frac{x}{n} \sim \widehat{P}$ con $E(\widehat{P}) = P \; y \; \sigma = \sqrt{\frac{PQ}{n}}$ entonces $\frac{\widehat{P}-P}{\sqrt{\frac{PQ}{n}}} \sim N(0,1)$.
  - Si la proporción muestral proviene de una Hipergeométrica, y el tamaño de la muestra es igual o mayor a 100 tenderá a normal.
    Si $E(\widehat{P}) = P \; y \; \sigma = \sqrt{\frac{PQ}{n}\frac{N-n}{N-1}}$ entonces $\frac{\widehat{P}-P}{\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1)$.

- **Distribuciones de las pequeñas muestras:** T de Student y Chi Cuadrado trabajan con un tamaño de muestra inferior a 30 elementos n<30. Estas distribuciones se utilizan en la inferencia estadística porque trabajan con muestras y los generalizan a aspectos poblacionales.

  - **Grados de Libertad:** Es el número de observaciones linealmente independientes que ocurren en una suma de cuadrados. Cantidad de valores que dan como media el mismo valor.
    $\bar{x} = 5 = \frac{1}{2}(x_i + x_2) \Rightarrow x_1 + x_2 = 10 \Rightarrow (8;2); (6;4); (10;0); (5;5); (7;3); (9;1)$

    Para n números $x_1, x_2, \ldots, x_n$ y dada la condición de que el promedio debe ser $\bar{x} = k$, podremos elegir los valores de x_n pero uno de ellos quedará determinado por la igualdad, por lo que diremos que tenemos $n-1$ **grados de libertad** y se representa por: $\varphi = \delta = n - 1$.Si no hay ningún parámetro: $\varphi = \delta = n$.

- **Distribución de Chi-Cuadrado:** Esta distribución tiene un solo parámetro denominado grados de libertad. Al no tener parámetros ya estimado trabajamos con $\varphi = n$. Se utiliza mucho para hacer la inferencia de la varianza ya que nunca puede asumirse para valores negativos.
  - Con $n = 1$: Elevamos z^2: $z^2 = \frac{(x-\mu)^2}{\sigma^2} = \chi^2_{(1)}$
  - Con $\varphi = n$: $\chi^2_{(n)} = \sum_{i=0}^{n} \frac{(x_i-\mu)^2}{\sigma^2}$
  - **Función de Densidad:** Es asimétrica positiva porque solo tienen densidad los valores positivos. Se va haciendo mas simétrica (casi gaussiana) cuando aumenta el numero de grados de libertad.

    $f(\chi^2) = \frac{(\chi^2)^{(\varphi/2)-1} \cdot e^{-(\chi^2/2)}}{2^{\varphi/2} \cdot \tau(\varphi/2)}$
  - **Función de Acumulacion:** $F(x) = \int_0^x f(x)\,dx$
  - **Parámetro:**
    - $E(\chi^2) = \varphi$
    - $V(\chi^2) = 2\,\varphi$
    - $\sigma = \sqrt{2\,\varphi}$
    - Grados de libertad

- **Distribución t de Student:** Cuando tenemos una muestra $x_1, x_2, \ldots, x_n$ proveniente de una población normal cuya media es μ y varianza σ^2. La variable t es una razón entre la variable normal estandarizada sobre la raíz cuadrada de división entre una variable Chi Cuadrado por su número de Grados de Libertad.

$$t_\varphi = \frac{z}{\sqrt{\frac{\chi^2}{\varphi}}}$$

t de Student se utiliza cuando queremos estimar la media de una población normalmente distribuida a partir de una muestra pequeña. A partir de 30 observaciones la distribución t se parece mucho a una distribución normal, por lo tanto, utilizaremos el Modelo Normal aunque la curva va a ser más aplanada. Podemos observar también que mientras mas aumentan los grados de libertad más se acerca a N(0,1). En este caso, no se conoce la desviación estándar de una población y tiene que ser estimada a partir de las observaciones de la muestra.
Puede ser utilizada para trabajar con valores de la variable que va de menos infinito a infinito para hacer inferencia de la media o la proporción.
Es simétrica con respecto al cero.

  - **Función de Densidad:**

$$f(t) = \left(\frac{1}{\sqrt{\varphi\pi}}\right) \cdot \left[\frac{\tau\left(\frac{\varphi+1}{2}\right)}{\tau\left(\frac{\varphi}{2}\right)}\right] \left(1 + \frac{t^2}{\varphi}\right)^{-(\varphi+1)/2}$$

- **Función de Distribución:** $\int_{-\infty}^{t} f(t)\,dt$ $\qquad t = \frac{x-\mu}{\hat{\sigma}}$
- **Parámetros:**
  - $E(x) = 0$
  - $V(t) = \frac{\varphi}{\varphi-2}$
  - Grados de libertad

Unidad 8
Teoría del Muestreo

- **Teoría del Muestreo:** Es un estudio de las relaciones existentes entre una población y muestra extraídas de la misma. Fija pautas, técnicas y procedimientos necesarios para una correcta selección y uso de las muestras, a los finales de que las mismas sean representativas de la población para la inferencia estadística.

- **Inferencia Estadística:** Tiene dos aspectos importantes:
  - **Estimación de parámetros de población:** Partiendo de estimadores muestrales y calculando la precisión de esta.
  - **Docimasia o prueba de hipótesis:** Consiste en la utilización de las muestras para verificar la veracidad de un supuesto sobre la población.

Clave: En la unidad 1 hablamos de los pasos de un método científico y ahí estaba la inferencia estadística.

- **Parámetro:** Medida calculada en base a datos poblacionales.

- **Estadístico:** Medida calculada en base a datos muestrales.

- **Razones para el muestreo:**
  - Las poblaciones pueden ser infinitas.
  - La captación del dato requiere la destrucción del elemento, no se puede destruir toda la población.
  - Hay veces en las cuales la población no es accesible.
  - Mayor Exactitud: Se evitan los errores de encuesta, reemplazandolas por errores (en casos, menores) de muestreo, los cuales pueden ser calculados haciendo posible el conocimiento del tamaño probable del error, logrando un mayor grado de confianza.
  - Costo.
  - Tiempo: Extraer una muestra requiere menor tiempo que el censo. Además, la corrección, codificación y tabulación de las muestras consumen menos tiempo.

- **Consideraciones sobre muestras y muestreo:**
  - Cualquier subconjunto de elementos de una población es una muestra de ella.
  - Cuando se utiliza la muestra se pretende conocer las características de la población.
  - La muestra para estudiar, por lo tanto, debe ser representativa de la población.

- Muestra representativa es aquella que reúne en sí las características principales de la población y guarda relación con las condición particular que se estudia.
- Los aspectos fundamentales que se deben considerar en la extracción de la muestra representativa son: el sistema de muestreo utilizado y el tamaño de la muestra.

- **Base Teórica del muestreo:**
  - **Diversidad:** Cualquier población tiene propiedades características y la variación en sus elementos son limitados porque lo que si se toman algunas muestras al azar serán similares pero nunca idénticas
  - **Uniformidad:** Es la tendencia de las características mensurables a agruparse o concentrarse alrededor de una medida de tendencia central.

- **Una muestra representativa debería tener:**
  - **Fiabilidad:** Se mide la fiabilidad o precisión del muestreo, es una medida inversamente proporcional o la varianza. A mayor varianza menor fiabilidad en el resultado.
  - **Efectividad:** Un diseño de muestreo se considera efectivo si se obtienen el mismo grado de fiabilidad al menor costo posible.

  ❖ **Procedimientos para la selección de muestras:**

- **Muestreo no Probabilístico:** Las unidades de la población que integrarán la muestra se eligen según el criterio del investigador, ya sea por facilidad, o por algún objetivo, por lo cual no dependen de la probabilidad.
  No se puede conocer:
    - La probabilidad que tiene la muestra de ser seleccionada.
    - El error del muestreo, la confianza y el riesgo.
    - Precisión del estimador.

  - **Muestreo discrecional o por juicio:** Aquí desempeña un papel importante el juicio del investigador en la selección de los elementos que van a componer la muestra. Es aconsejable cuando el responsable del estudio conoce estudios anteriores similares o idénticos y sabe con precisión que la muestra que utilizó; también es conveniente en poblaciones reducidas y conocidas por el investigador.
  - **Muestreo de la muestra disponible o por conveniencia:** La muestra la conforman elementos de la población que se encuentren convenientemente disponibles. Obtener una muestra de esta manera es rápido y económico, pero la gente que contactan no es representativa de toda la población.

- **Muestreo Probabilístico:** Todos los elementos de la población tienen la misma posibilidad de ser escogidos. Se obtienen usando una herramienta estadística en base a la teoría de probabilidad que da como resultado una cantidad representativa de la población. Se puede conocer:
    - La probabilidad que tiene la muestra de ser seleccionada.
    - El error del muestreo, la confianza y el riesgo.
    - Precisión del estimador.

- **Muestreo Aleatorio Simple:** Se extraen de manera tal que cada una de las muestras posibles del mismo tamaño tengan la misma probabilidad de ser seleccionadas, o sea, de manera imparcial (el sesgo muestral se suprime). Se utiliza la tabla de números aleatorios u otra técnica de selección al azar. Las probabilidades de cada elemento será:
  - **Con reposición:** $Pr = \frac{1}{N}$
  - **Sin reposición:** $Pr = \frac{1}{C_N^n}$

- **Muestreo Aleatorio Estratificado:** Separa la población en grupos que no se traslapan llamados estratos y se elige después una muestra aleatoria simple de cada estrato. Se utiliza cuando la población no es homogénea o se desea información más precisa en ciertos subconjuntos de ésta. Cada estrato va a ser homogéneo en sí mismo pero heterogéneo en relación con los otros estratos.
  - **Procedimiento:** La población se divide en cierto número de grupos llamados estratos, mutuamente excluyentes y colectivamente exhaustivos.
    La población será: $N = N_1 + N_2 + \ldots + N_r$ y la muestra
    $n = n_1 + n_2 + \ldots + n_r$.
    El muestreo estratificado puede ser más eficiente que el aleatorio simple, sobre todo cuando existen características distintas entre los elementos.

  - **Planeación del Tamaño de las Muestras (Afijación):** Formas de determinar el tamaño de las submuestras en función del tamaño de la muestra n para minimizar costos y maximizar ganancias.
    - **Afijación Igual:** Donde todos los n_i son iguales. Voy a sacar la misma cantidad de elementos en cada muestra sin importar lo que pase dentro de cada estrato. Al tamaño de la muestra por la cantidad de estratos y me da cuantos elementos voy a sacar de los estratos. No tengo en cuenta cuántos elementos hay en cada estrato.
      n_i: $submuestra$ | n: $tamaño\ de\ muestra$ | r: $cantidad\ de\ elementos$
      $$n_i = \frac{n}{r}$$
    - **Afijación Proporcional:** Tengo en cuenta cuántos elementos hay en cada estrato y según esto voy a sacar una proporción de cada estrato.
      $n_i = (\frac{N_i}{N})\, n$
    - **Afijación Óptima:** Tengo mejor información porque se como se distribuyen los datos dentro del estrato en relación al promedio. Si los datos están más concentrados puedo sacar menos elementos del estrato.
      σ_i: $desviación\ estándar\ del\ estrato$

$$n_i = \left(\frac{N_i \cdot \sigma_i}{\sum_{i=1}^{\tau} N_i \cdot \sigma_i}\right) \cdot n \qquad \sum N.\sigma = N_A\sigma_A + N_B\sigma_B + \ldots + N_i\sigma_i$$

- **Muestreo Sistemático:** Requiere de una selección aleatoria inicial de observaciones seguida de otra selección de observaciones obtenida por un sistema o regla.
  1.Se deben ordenar y enumeran los elementos de la población con algún criterio.
  2. Se selecciona uno cada k elementos, donde k es la razón o patrón del muestreo.
  $k = \frac{N}{n}$ El primer elemento es seleccionado al azar entre los primeros k.
  Ej: Cada 10 elementos tomo 1, entonces k es 1.
  3. Se determina la unidad muestral por la que se iniciará la selección de la muestra utilizando un número aleatorio comprendido entre 1 y k. Sirve como número aleatorio raíz.
  4. De acuerdo con ese número obtenido y como la población está ordenada en una lista, el elemento que figura en la posición determinada por el número aleatorio será el primero que conformará la muestra.
  5. Se le va adicionando $k$hasta complementar la cantidad de elementos que integrarán la muestra.
  Ej: Si mi muestra es de 5 yo tomo un número aleatorio entre 1 y 5. Me sale 3. Entonces, 3 es mi primer elemento de la muestra, de ahí sumo 5. 8 es mi segundo elemento de la muestra y así.

- **Muestreo por conglomerado:** Son muestras aleatorias de unidades heterogéneas entre sí. Es lo opuesto al Muestreo Estratificado, cada conglomerado es heterogéneo en sí mismo y homogéneo con respecto a los otros conglomerados (se hacen conglomerados lo más parecido posibles entre sí, mientras que las diferencias entre los elementos de cada conglomerado se hacen lo más grandes posibles). Lo ideal sería que cada conglomerado sea una miniatura de toda la población y así un solo conglomerado se llama unidad de Muestreo Primaria.
  - **Parámetros:**
    - **Media Poblacional:** $\mu = \frac{\sum_{i=1}^{N} x_i^2}{N}$
    - **Varianza Poblacional:** $\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2$
    - **Desviación Poblacional:** $\sigma = \sqrt{\sigma^2}$
    - **Varianza Poblacional Corregida:** $s^2 = \frac{\sum_{i=1}^{N} x_i^2}{N-1} - \frac{\mu^2 . N}{N-1}$
    - **Desviación Poblacional Corregida:** $s = \sqrt{s^2}$

- **Corrección:** $s^2 = \sigma^2 \frac{N}{N-1}$ | $s = \sigma\sqrt{\frac{N}{N-1}}$

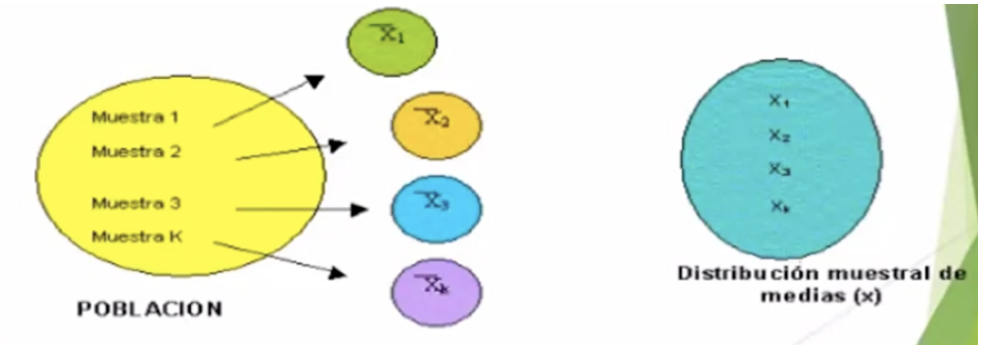
- **Comparaciones entre métodos:**
  El muestreo **aleatorio** es el punto de referencia para todos los demás métodos. Sin embargo, pocas encuestas a gran escala usan solamente el muestreo irrestricto aleatorio, debido a que otros métodos proporcionan mayor precisión o eficiencia o ambas cosas.
  El muestreo **estratificado,** usualmente produce un estimador que posee una varianza menor a la que puede ser obtenida por muestreo irrestricto aleatorio: por lo tanto, el costo de una encuesta puede reducirse seleccionando pocos elementos. La situación ideal para este método es tener mediciones iguales dentro de cualquier estrato pero que ocurran diferencias conforme se pasa de un estrato a otro.
  El muestreo **sistemático** suele realizarse cuando recolectar una muestra irrestricta aleatoria o una muestra aleatoria estratificada es extremadamente costosa o imposible. Se lo utiliza como una conveniencia. Por la periodicidad o sistematicidad de sus elementos, podría arrojar deficientes resultados.
  El muestreo por **conglomerados** puede reducir el costo porque cada unidad de muestreo es una colección de elementos usualmente seleccionados con el fin de que estén juntos físicamente. Suele realizarse cuando no se dispone de un marco que liste todo los elementos de la población o cuando los costos de transporte de un elemento a otro son considerables. El muestreo por conglomerados reduce el costo de la encuesta, reduciendo el costo de la recolección de datos. La situación ideal para este muestreo es tener conglomerados con mediciones tan diferentes como sea posible, por tener medidas iguales.

Distribución de Muestreo:

- Las muestras aleatorias obtenidas de una población son impredecibles.
- Si yo hablo de un parámetro es un conjunto de datos que se obtiene de una población. Si hablo de estadísticos hablo de valores calculados a través de datos obtenidos a través de una muestra.
- No se espera que dos muestras aleatorias del mismo tamaño y tomadas de la misma población tenga la misma media muestral o que sean completamente parecidas, si no que se espera que cualquier estadístico cambie su valor de una muestra a otra.
- Son muy importantes porque las inferencias sobre las poblaciones se harán usando estadísticos muestrales. Con el análisis de las distribuciones asociadas con los estadísticos muestrales, se puede juzgar la confiabilidad de un estadístico muestral como un instrumento para hacer inferencias sobre un parámetro poblacional desconocido.

- **Distribución de Probabilidad por muestreo de la media muestral:** La distribución muestral de un estadístico (en este caso la media muestral) es la de todos sus valores posibles calculados a partir de muestras del mismo tamaño. Como los valores de un estadístico varían de una muestra aleatoria a otra se la puede considerar como una variable

aleatoria con su correspondiente distribución de frecuencias.



Para elaborar las distribuciones muestrales se procede de la siguiente manera:

Ejemplo

- Se eligen muestras ordenadas de tamaño 2, con reemplazo, de la población de valores 0, 2, 4 y 6.
- Encuentre:

μ la media poblacional.

σ , la desviación estándar poblacional.

μ_x, la media de la distribución muestral de medias.

σ_x, la desviación estándar de la distribución muestral de medias.

0 . Calcular la media y la desviación estándar poblacional:

a) **La media poblacional es:**

$$\mu = \frac{0+2+4+6}{4} = 3$$

b. La desviación estándar de la población es:

$$\sigma = \sqrt{\frac{(0-3)^2+(2-3)^2+(4-3)^2+(6-3)^2}{4}} = 2.236$$

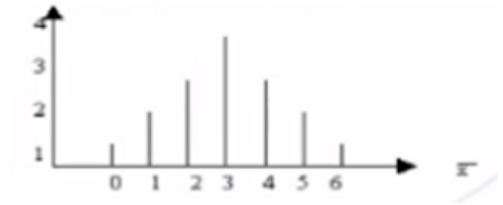
1. De una población finita de tamaño N, se extraen de manera aleatoria todas las muestras posibles de tamaño n. Si es un MCR la cantidad de elementos que puede tener la muestra es N^n.

2. Se calcula la estadística de interés para cada muestra. Si quiero calcular la media, la cálculo para cada valor de la muestra.

Muestra	x
(0,0)	0
(0,2)	1
(0,4)	2
(0,6)	3
(2,0)	1
(2,2)	2
(2,4)	3
(2,6)	4
(4,0)	2
(4,2)	3
(4,4)	4
(4,6)	5
(6,0)	3
(6,2)	4
(6,4)	5
(6,6)	6

3. Se ordenan en una columna los distintos valores observados de la estadística y, en otra columna las respectivas probabilidades o frecuencias de cada valor posible de esa variable.

Distribución de frecuencias de x	
$\bar{x}$	f
0	1
1	2
2	3
3	4
4	3
5	2
6	1



IMPORTANTE: Como para cualquier variable aleatoria, la distribución muestral de medias tiene una media o valor esperado, una varianza y una desviación estándar, se puede demostrar que tiene una media igual a la media poblacional:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu = 3$$

La desviación estándar de la distribución muestral de medias es: $\sigma_{\bar{x}} = \sqrt{\frac{\sum (\bar{x}-\mu_{\bar{x}})^2 f}{\sum f}}$

$$\sigma_{\bar{x}} = \sqrt{\frac{(0-3)^2 1 + (1-3)^2 2 + (2-3)^2 3 + (3-3)^2 4 + (4-3)^2 3 + (5-3)^2 2 + (6-3)^2 1}{16}} = 1.58$$

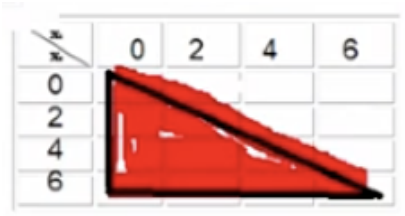
De aquí podemos deducir que la desviación de las medias muestrales va a ser igual a la desviación poblacional sobre la raíz de n: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

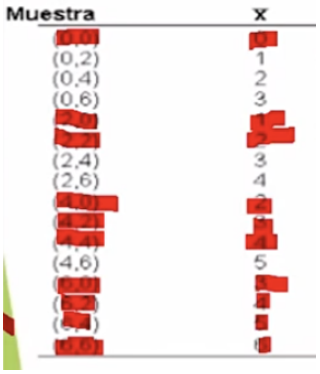
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.236}{2} = 1.58$$ (al 2 le falta la raíz)

Entonces: Del ejercicio anterior se puede ver que una distribución muestral se general extrayendo todas las posibles muestras del mismo tamaño de la población y calculándolas a estas sus estadísticas.

- Si la población de la que se extraen las muestras es normal, la distribución muestral de medidas será normal sin importar el tamaño de la muestra.
- Si la población de donde se extraen las muestras no es normal, entonces el tamaño de la muestra debe ser mayor o igual a 30, para que la distribución muestral tenga una forma acampanada.
- Mientras mayor sea el tamaño de la muestra, más cerca estará la distribución muestral de ser normal. La aproximación normal se considera buena si se cumple n=30.
    - **La media muestral es:** $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ al variable aleatoria por lo que también puede calcularse:
        - Función de probabilidad.
        - Su esperanza.
        - Su varianza y desviación.
    - La cantidad de muestras posibles son:

- Con reposición: N^n
- Sin reposición: C_N^n (Voy a ir sacando elementos de la población y no los voy a reponer así que si armo un cuadro de dos entradas la diagonal principal y algún extremo, el de arriba o el de abajo, no se pueden dar nunca, así que me quedan menos muestras) y te lo muestro con esta explicación bien fiera:





Todo lo que es fórmula queda igual excepto la varianza y la desviación estándar que hay que agregarle el factor de corrección.

- La media muestral es una
- **La esperanza es:** $E(\bar{x}) = \bar{x} = \sum_{i=1}^{m} \bar{x}_i \,.\, P(\bar{x}_i) = \frac{\sum_{i=1}^{m} \bar{x}_i \cdot n_i}{m}$

  m será la cantidad de muestras que será N^nsi es MCR ó C_N^nsi es MSR y $P(\bar{x}_i) = \frac{n_i}{m}$.
- **Varianza y desviación:**

$$\sigma^2 = E(\bar{x}^2) - [E(\bar{x})]^2$$

$$\sigma = \sqrt{E(\bar{x}^2) - [E(\bar{x})]^2}$$

$$\sigma^2 = \sum_{i=1}^{m} \bar{x}_i^{\,2} \,.\, P(\bar{x}_i) - \left[E(\bar{x})\right]^2 = \frac{\sum_{i=1}^{m} \bar{x}_i^{\,2}\, n_i}{m} - \left[E(\bar{x})\right]^2$$

$$\sigma = \sqrt{\sum_{i=1}^{m} \bar{x}_i^{\,2} \,.\, P(\bar{x}_i) - \left[E(\bar{x})\right]^2} = \sqrt{\frac{\sum_{i=1}^{m} \bar{x}_i^{\,2}\, n_i}{m} - \left[E(\bar{x})\right]^2}$$

- Relaciones: Sólo en el muestreo sin reemplazo.

$$\sigma_{\bar{x}}^2 = \frac{\sigma.\chi^2}{n} \cdot \frac{N-n}{N-1} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- **Distribución muestral de proporciones:** Esta distribución se genera de igual manera que la distribución muestral de medias, a excepción de que al extraer las muestras de la población se calcula el estadístico proporción $\widehat{P} = \frac{x}{n}$ en donde $x$es el número de éxitos u observaciones de interés y n el tamaño de la muestra. Entonces, tomo a la proporción como una variable aleatoria tomando varias muestras y de cada muestra una proporción para luego sacar sus probabilidades o frecuencias.

Muestra	Proporción ($\hat{p}$)
A,A	0
A,B	0
A,C	0.5
A,D	0.5
B,A	0
B,B	0
B,C	0.5
B,D	0.5
C,A	0.5
C,B	0.5
C,C	1
C,D	1
D,A	0.5
D,B	0.5
D,C	1
D,D	1

$\hat{p}$	$P(\hat{p})$	$\hat{p}\,P(\hat{p})$	$\hat{p}^2\,P(\hat{p})$
0	4/16	0	0
0,5	8/16	4/16	2/16
1	4/16	4/16	4/16
	1	0,5	6/16

$E(\hat{p}) = \sum \hat{p}\ P(\hat{p}) = 0{,}5$

$V(\hat{p}) = \sum \hat{p}^2\ P(\hat{p}) - [\,E(\hat{p})\,]^2 = 0{,}125$

$DE(\hat{p}) = \sqrt{V(\hat{p})} = 0{,}3535$

Entonces las relaciones entre estadísticas y parámetros se cumplen:

$E(\hat{p}) = P$	$v(\hat{p}) = \frac{P(1-P)}{n}$	$DE(\hat{p}) = \sqrt{\frac{P(1-P)}{n}}$
$0{,}5 = 0{,}5$	$0{,}125 = 0{,}125$	$0{,}3535 = 0{,}3535$

RELACIONES

M.C.R.	$E(\bar{x}) = \mu$	$E(\hat{P}) = P$	$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \rightarrow \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$	$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} \rightarrow \sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}}$
M.S.R.	$E(\bar{x}) = \mu$	$E(\hat{P}) = P$	$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{n} \cdot \frac{N-n}{N-1} \rightarrow$ $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$	$\sigma_{\hat{P}}^2 = \frac{P(1-P)}{n} \cdot \frac{N-n}{N-1} \rightarrow$ $\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

- **Distribución por muestreo de la varianza muestral corregida:**
  - La varianza muestral para series simples es: $s^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$
  - La varianza muestral corregida para series simples es: $s^2 = \frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x^2 - n.\bar{x}^2}{n-1}$
  - La relación entre ambos es: $\hat{s}^2 = s^2 \frac{n}{n-1}$
  - Las desviaciones serán: $s = \sqrt{s^2}$; $\hat{s} = \sqrt{\hat{s}^2}$
  - La esperanza de la distribución de la varianza muestral corregida será:

    $E(\hat{s}^2) = \sum_{i=1}^{m} \hat{s}_i^2 . P(\hat{s}_i^2) = \frac{\sum_{i=1}^{m} \hat{s}^2}{m}$ siendo m igual a N^n o C_N^n según corresponda.

❖ Para cualquier estadística muestral basada en un muestreo aleatorio, existe una distribución de muestreo de esa estadística que indica los valores que puede tomar esta estadística con todas las muestras posibles y la probabilidad que esos valores se presentan.

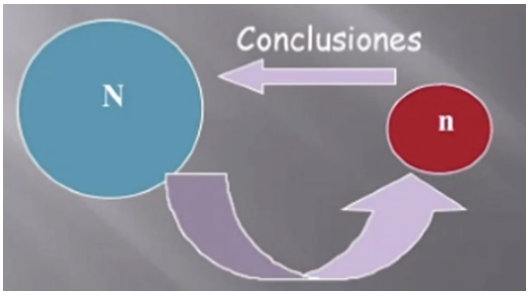
- **Ley de Grandes Números:** La probabilidad de que la diferencia entre el estadístico y el parámetro sea superior a un número d, arbitrariamente elegido, tiende a 0 a medida que $n \to \infty$.

$$\lim_{n \to \infty} Pr(|\hat{\theta} - \theta| > d) = 0$$

- **Teorema Central del Límite:** Es una teoría estadística que establece que, dada una muestra suficientemente grande de la población, la distribución de las medias muestrales seguirá una distribución normal. Asegura que a medida que el tamaño de la muestra se incrementa, la media muestral se acercará a la media de la población. Por ende, mediante TCL podemos definir la distribución de la media muestral de una determinada población con una varianza conocida. De manera que la distribución seguirá una distribución normal si el tamaño de la muestra es lo suficientemente grande.
  Propiedades:
  - Si el tamaño de la muestra es suficientemente grande (n>30), la distribución de las medias muestrales seguirá una distribución normal independientemente de la forma de la distribución con la que estemos trabajando.
  - La media poblacional y la media muestral serán iguales, o sea, la media de la distribución de todas las medias muestrales será igual a la media del total de la población.
  - La varianza de la distribución de las medias muestrales será $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$(la varianza de la población dividido el tamaño de la muestra).

Unidad 9
Estimación Estadística

- **Estadística Inferencial:** Se define como el conjunto de procedimientos empleados para llegar a mayores generalizaciones o inferencias acerca de poblaciones, basándose en datos muestrales. Quiero obtener información de una población basándose en una muestra:



Sabemos que un **parámetro** es una medida numérica que describe las características de la población; que un **estadístico** es una medida cuantitativa, derivada de un conjunto de datos para una muestra, con el objetivo de estimar características de una población. Y el

estimador es un estadístico que es “bueno” como para representar la muestra y se usa para estimar el parámetro que nos interesa de la población.

Los parámetros se estiman utilizando estadísticas muestrales.

- **Estimador:** Es el método que se utiliza para hacer la estimación, puede ser:
  - **Estimador Puntual:** Proporciona un solo número como estimación.
  - **Estimador por Intervalos:** El parámetro se estima situado entre dos límites para una confianza dada.

- **Propiedades de los buenos estimadores:**
  - **Insesgabilidad:** El estimador es insesgado cuando su esperanza es igual al parámetro de la población que se estima. (Cumple con el TCL).

    $$E(\widehat{\theta}) = \theta \Rightarrow TCL: E(\overline{x}) = \mu \mid E(\widehat{P}) = P$$

    $$E(\widehat{\theta}): \ estimador\ del\ parámetro \mid \theta: parámetro\ a\ estimar$$

    Para un estimador $\widehat{\theta}$del parámetro θ.
    - $E(\widehat{\theta}) = \theta$ ⇒Estimador insesgado.
    - $E(\widehat{\theta}) > \theta$ ⇒Estimador con sesgo positivo.
    - $E(\widehat{\theta}) < \theta$ ⇒Estimador con sesgo negativo.

  - **Eficiencia:** La varianza de un estimador proporciona una idea del grado de confianza que puede tener en el mismo, entonces, dados dos estimadores insesgados se dice que es más eficiente el que tenga menor varianza.

  - **Consistencia:** Un estimador es consistente si para muestras grandes hay una probabilidad cercana a 1, de que la estimación esté cerca del parámetro estimar. Es decir, el estimador $\widehat{\theta}$ del parámetro θ es consistente si para $n \to N$ se verifica que $\widehat{\theta} \to \theta$.

    $$\lim_{n \to \infty} Pr(|\widehat{\theta} - \theta| < \delta)) = 1$$

    Entonces, se dice que es más consistente cuanto mayor sea la muestra porque el valor del estimador se aproxima al del parámetro conforme aumenta la misma.
    A medida que n crece mas se va acercar a N, entonces $\overline{x}$ más se va a acercar a μ y $\widehat{P}$a P.

  - **Suficiencia:** Un estimador suficiente es un estimador que utiliza toda la información contenida en la muestra. Por ejemplo, con la media uso todos los valores de la muestra, pero con la moda y la mediana no, entonces la media es suficiente.

IMPORTANTE: La media y la proporción son los estimadores que cumplen con todas estas condiciones.

- **Estimación Puntual:** Proporciona un solo número como estimación. Tiene la limitación de que no proporciona información acerca de la precisión de la estimación obtenida, o sea, de

la magnitud del error debido al muestreo. Por ello, se utiliza demasiado poco.

- **Error:** Es la diferencia entre el estimador puntual y el parámetro $e = \hat{\theta}_o - \theta$. El error máximo aceptable es $|e|$ por lo tanto: $|\hat{\theta} - \theta| \leq e$.

- **Riesgo:** Es la probabilidad de que se produzca un error mayor al aceptable.

  $Pr\{(\hat{\theta} - \theta) > e\}$ =Riesgo

  Para poder calcularla se la estandariza: $z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ $\sigma = \frac{\sigma_\theta}{\sqrt{n}}$ $\sigma\ es\ la\ dirección\ típica$

  Relación:

  $$z = \frac{e\sqrt{n}}{\sigma} \quad \sigma = \frac{e\sqrt{n}}{z} \quad n = \left(\frac{z\,\sigma}{e}\right)^2 \quad e = \frac{\sigma\, z}{\sqrt{n}}$$

- **Estimación por Intervalos:** Consiste en obtener una cierta intervalo aleatorio $[L_i;\ L_s]$a partir de la estimación puntual considerando un cierto error de estimación, y un determinado grado de confianza de que el intervalo construido contiene al parámetro que queremos estimar. Para determinar los límites del intervalo se hace:

  $\hat{\theta} \pm 3\,\sigma_{\hat{\theta}} \Rightarrow \left[\hat{\theta} - 3\sigma_{\hat{\theta}};\ \hat{\theta} + 3\sigma_{\hat{\theta}}\right]$

  Este intervalo recibe el nombre de “Intervalo de confianza para el parámetro de la población” y la probabilidad de una aseveración correcta es de 0,997.

  Coeficiente de confianza: $1 - \alpha$ $Pr\left\{\theta \in \left[\hat{\theta} - 3\sigma_{\hat{\theta}};\ \hat{\theta} + 3\sigma_{\hat{\theta}}\right]\right\}$

  Nivel de confianza: $(1 - \alpha).100$

  Y se calcula:

  $$\left(1 - \tfrac{\alpha}{2}\right) - \tfrac{\alpha}{2} \Rightarrow 1 - \tfrac{\alpha}{2} = Pr\left(z \leq z_{1-\frac{\alpha}{2}}\right);\ \tfrac{\alpha}{2} = Pr\left(z \leq z_{\frac{\alpha}{2}}\right)$$

  Entonces:

  $$1 - \alpha = Pr\left(z \leq z_{1-\frac{\alpha}{2}}\right) - Pr\left(z \leq z_{\frac{\alpha}{2}}\right)$$

  Los límites de confianza se calculan: $L_i = \hat{\theta} - e\ ;\ L_s = \hat{\theta} + e$

  También: $1 - \alpha = Pr\{k_1 \leq \theta \leq k_2\}$

  $$k_1 = \hat{\theta} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad k_2 = \hat{\theta} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

  Y en ese caso de muestreo sin reposición: $k_{12} = \hat{\theta} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

- **Determinación del Tamaño de la Muestra:**

Ya sabemos que: $z = \frac{\hat{\theta} - \theta}{\frac{\sigma}{\sqrt{n}}}$ y $e = \hat{\theta} - \theta$

Entonces: $z = \frac{e}{\frac{\sigma}{\sqrt{n}}} = \frac{e\sqrt{n}}{\sigma} \Rightarrow z\sigma = e\sqrt{n} \Rightarrow \sqrt{n} = \frac{z\,\sigma}{e}$

Por lo que n debe ser: $n \geq \left(\frac{z\,\sigma}{e}\right)^2$

Entonces para determinar el tamaño de la muestra se debe conocer:

- El nivel confianza (z).
- El error permitido (e).
- La desviación estándar.

Unidad 10
Dócima y Verificación de Hipótesis

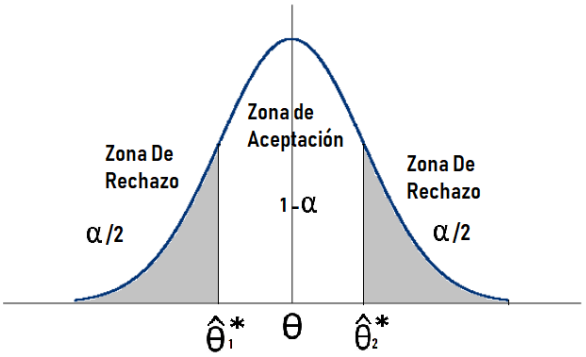
Es un procedimiento de la inferencia estadística para la toma de decisiones.

- **Decisión Estadística:** Una decisión que se tome con respecto a algún aspecto de la población, en base a evidencias proporcionadas por las muestras.

- **Hipótesis Estadística:** Suposiciones o conjeturas que se establecen acerca del valor de un parámetro, puede tratarse de:
  - **Hipótesis Nula:** $H_o = \hat{\theta} = \theta$ Es un supuesto acerca de uno o más parámetros y lo que se propone es que no existe diferencia entre el verdadero valor del parámetro de la población y el valor obtenido de la muestra.
  - **Hipótesis Alternativa:** $H_1 = \hat{\theta} >< \theta \Rightarrow \hat{\theta} \neq \theta$Si la hipótesis nula es falsa se propone que existe algún cambio con respecto a la población.

- **Docimasia de Hipótesis (Dosis):** Es un experimento aleatorio que se realiza para decidir sobre la veracidad o falsedad de una hipótesis.
  Pueden darse dos tipos de Errores:
  - Aceptar una hipótesis falsa (Tipo 1)
  - Rechazar una hipótesis cierta (Tipo 2)

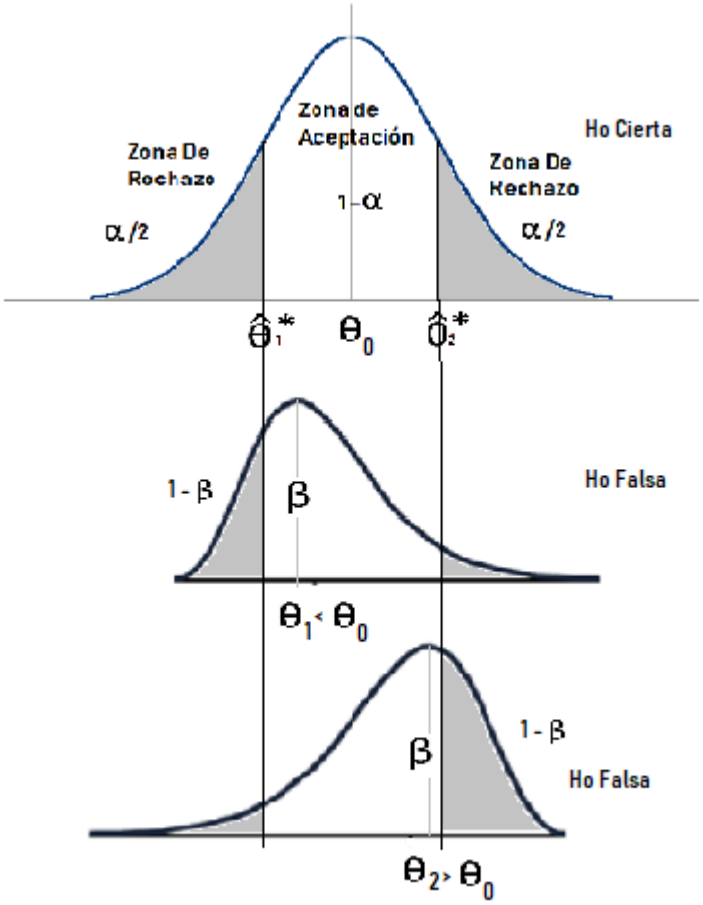
  Luego se puede determinar la probabilidad de que una estimación puntual se aleje del parámetro θ o más allá de ciertos límites llamados puntos críticos $(\hat{\theta}_1^*; \hat{\theta}_2^*)$ y asuma valores en una zona llamada “Zona de Rechazo”. Existe una probabilidad α, de que la hipótesis nula sea cierta y se la rechace como falsa. A esta probabilidad la fija el investigador a niveles muy bajos y se llama nivel de significación:



  - **Nivel de Significación:** $\alpha = Pr\{Rechazar\ H_o / H_o\ Cierta\}$

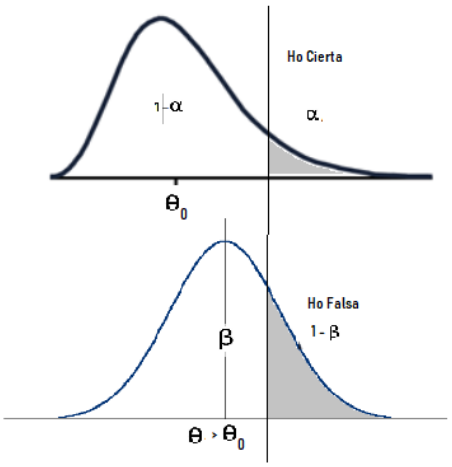
- **Errores:** Pueden ocurrir dos tipos de errores:
  - **Error Tipo 1:** H_o cae en zona de rechazo siento verdadero. Nivel de significación:
    $\alpha = Pr\{Error\ Tipo\ 1\} = Pr\left\{Rechazar\ H_0 / H_o\ Cierta\right\}$
  - **Error Tipo 2:** H_o es falsa pero cae en zona de aceptación.
    $\beta = Pr\{Error\ Tipo\ 2\} = Pr\left\{Aceptar\ H_o / H_o\ Falsa\right\}$

La única forma de activar? $\alpha\ y\ \beta$(simultáneamente) es aumentar el tamaño de la muestra. Si H_o es cierta se tomará como parámetro θ_o, pero sí H_0 es falsa habrá un conjunto de valores para el estimador, supondremos valores mayores y menores a θ_o.

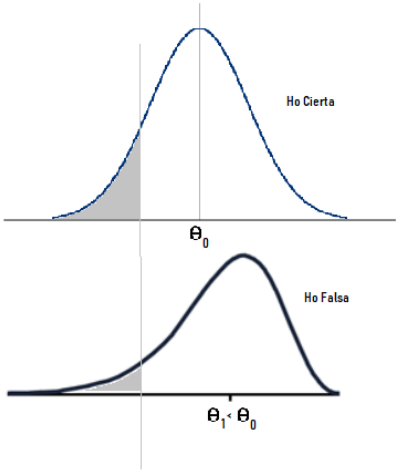


- **Tipos de Dócima:** Dado un parámetro θ, objeto de la dócima y un valor particular de este parámetro θ_o para el cual se verifica la hipótesis nula se puede plantear:
  - **Dócimas de Hipótesis Compuestas:**
    - **Dócimas bilaterales:** La hipótesis nula dice que el parámetro es igual a θ_o, mientras que la hipótesis alternativa dice que es distinto.
      $H_o: \theta = \theta_o \qquad H_1: \theta \neq \theta_o$

- **Dócimas laterales derecha:** La H_o se rechazará cuando la evidencia proporcionada por la muestra nos haga pensar que el parámetro es superior a θ_o. La zona de rechazo está ubicada en el extremo derecho de la distribución.



- **Dócimas laterales izquierdas:** La hipótesis nula plantea que el parámetro es igual a θ_0 y la alternativa, que es menor. La H_o se rechazará si la evidencia proporcionada por la muestra nos hace suponer que el parámetro es inferior a θ_0. La zona de rechazo está en el extremo izquierdo.
  $H_o: \theta = \theta_o \quad H_1 = \theta < \theta_o$



- **Dócimas de Hipótesis Simples:**
  Cuando se contraste un valor para H_o contra un valor para H_1 donde $H_o: \theta = \theta_o \ y \ H_1: \theta = \theta_1$. Las décimas serán laterales izquierdas o derechas según θ_1 sea menor que θ_0.
  - **Dócimas laterales derechas:** Mismos gráficos anteriores.
  - **Dócimas laterales izquierdas:** Mismos gráficos anteriores.

- **Esquema para la Dócima:**
    - 1. Identificación del parámetro a dominar θ.
    - 2. Selección del estimador $\widehat{\theta}$.
    - 3. Determinación del estadístico $k = (\widehat{\theta}\,; \theta) = \frac{\widehat{\theta} - \theta}{\sigma/\sqrt{n}}$
    - 4. Cálculo de los puntos críticos $\widehat{\theta}^{*}$.
    - 5. Regla de decisión.
    - 6. Cálculo de probabilidades.

- **Dócima para P. Uso de la Distribución Normal:**

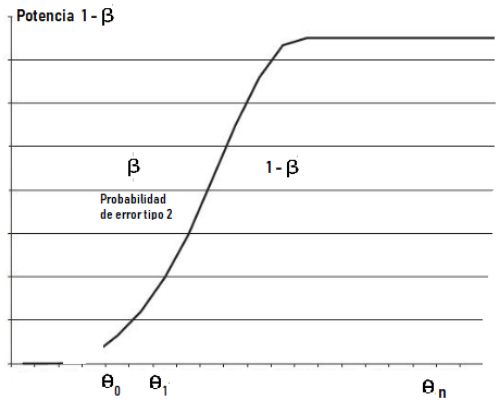
$$\theta = P \qquad \widehat{\theta} = P \sim N(P; \sqrt{\tfrac{PQ}{n}}) \qquad Z = \frac{\widehat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

$$k(\widehat{\theta}\,;\, \theta) = \frac{\widehat{P} - P}{\sqrt{\frac{PQ}{n}}} \sim N(0;1)$$

- **Potencia de la Dócima:** Mide la potencia que tiene la dócima para rechazar la hipótesis nula cuando es falsa.

$$1 - \beta = Pr\{Rechazar\ H_o\ /\ H_o\ Falsa\} = Potencia\ de\ la\ Dócima$$

- **Curva de Potencia:** Relaciona todos los valores posibles del parámetro con la probabilidad de rechazar H_o para un determinado α. Sí la H_o es cierta $1 - \beta = \alpha$. Si H_o es falsa, $1 - \beta \neq \alpha$.



Hay una asíntota en 1.

- **Curva Operatoria Característica:** Relaciona los posibles valores del parámetro con la probabilidad H_o para un determinado α.Es el complemento de la curva de potencia. Cuando H_o es cierta $\beta = 1 - \alpha$.


Rechazo β
1
β
1 - β
θ0
θ1
θn