**Resumen de estadística 2**

# **Repaso de algunos conceptos**

**Probabilidad:**

Es la parte de las matemáticas que se encarga del estudio de los fenómenos o experimentos aleatorios

**Experimento aleatorio:**

Experimento que cuando se le repite bajo las mismas condiciones iniciales, el resultado que se obtiene no siempre es el mismo

**Espacio Muestral:**

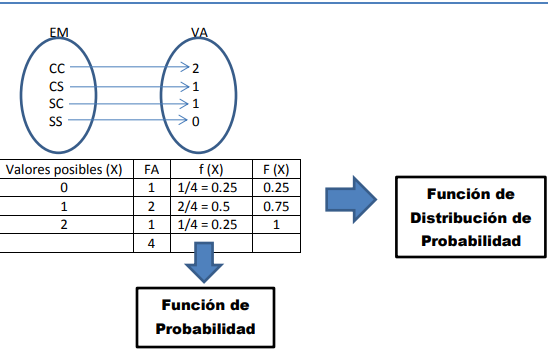
Asociado a un experimento aleatorio, el conjunto de todos los resultados posibles de dicho experimento.

**Evento:**

Todo subconjunto de un espacio muestral.

## **1.1) Distribuciones de probabilidad – Función de probabilidad para variables discretas**

**Variable aleatoria:** Tirar simultáneamente dos monedas al aire y observar el número de caras del experimento



**Esperanza Matemática**: E(X) = ∑ X \* p(X)

Propiedades de la Esperanza Matemática

1. E (K \* X) = K \* E (X)
2. E (K) = K
3. E ( X + K) = E (X) + K
4. E (K \* X + b) = K \* E (X) + b

**Varianza**: V (X) = E (X2) – (E (X))2

Propiedades de la Varianza

1. V (K) = 0
2. V (X + K) = V (X), siendo k una constante
3. V ( K\*X) = K2 \* V(X), siendo k una constante
4. V (aX + b) = a2 \* V (X), siendo X una variable y a, b constantes
5. V (X +Y) = V (X) + V (Y), siendo X e Y dos variables independientes

# **¿Qué es un modelo?**

Es una simplificación de la realidad para tratar de comprenderla y representarla en un modelo con menos elementos que la misma. se usa para asignar probabilidades a los diferentes resultados posibles.

## **2.1) Modelo Binomial o de Bernoulli:**

la distribución binomial es una distribución de variable discreta que se puede aplicar a distintas situaciones de toma de decisión donde pueda verificarse las siguientes condiciones:

1. Solo son posibles dos resultados mutuamente excluyentes entre sí, éxito (p) y fracaso (q)
2. Los resultados del ensayo constituyen eventos independientes
3. La probabilidad de éxitos permanece constante de un ensayo a otro

**Formula:**

P (X, n, p) = nCX px qn-x donde nCx = n! / x! \* n!

E (X) = n \* p

V (X) = n \* p \* q

## **2.2) Distribución Poisson:**

Se utiliza para determinar la ocurrencia de un número determinado de eventos, cuando estos ocurren en un intervalo de tiempo o espacio. Se requiere un valor para determinar la probabilidad de que ocurra un número designado de eventos en un proceso de poisson, es el numero promedio a largo plazo de eventos para el tiempo o dimensión especifica de interés, denominado como lambda (λ). Condiciones para su uso

1. Son posibles dos resultados mutuamente excluyentes entre sí
2. Los resultados del ensayo constituyen eventos independientes
3. El proceso es estacionario
4. La media del proceso es siempre proporcional a la longitud del continuo de tiempo o espacio

**Fórmula:**

P (X / λ) = (λX e -λ) / X!

E (X) = λ

V (X) = λ

# **Variable Aleatoria Continua:**

Diferente a la variable aleatoria discreta, que puede asumir una cantidad finita de valores en un intervalo determinado, la variable aleatoria continua puede asumir infinitos valores dentro de un intervalo. A continuación veremos uno de los mas importantes de variable aleatoria continua.

**3.1) Distribución Normal**

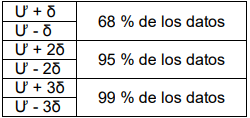
una de las más importantes en estadística por sus numerosas aplicaciones teóricas. Muchas variables aleatorias continuas tienen un comportamiento como el descripto por esta distribución.

Función de Densidad de la Normal: P(X) = (1 / σ \*√ (2∏)) \* e –((X – μ)2) /2σ

**Características:**

* Es simétrica, M (X) = Mdna (X) = Mo
* Tiene forma de campana
* Depende de dos parámetros: la media (μ) y la desviación estándar (σ)
* Es asintótica horizontalmente
* La suma del área bajo la curva es igual a 1

**Relación entre la media y la desviación estándar en la distribución normal:**



## **3.2) Distribución Normal estandarizada:**

procedimiento aplicable a cualquier variable con distribución normal que genera otra variable que vamos a llamar Z y su media es 0 y su desviación estándar es 1

Z (μ=0; σ=1) = (X - μ) / σ

# **Aproximación normal a la probabilidad binomial y poisson:**

## **4.1)** **Aproximación normal a la probabilidad binomial**

Cuando el número de casos en una variable con distribución binomial es grande se puede utilizar la distribución normal para aproximar las probabilidades binomiales. Cuando el número de casos es mayor o igual 30 y np y nq son mayores o iguales a 5, esto puede darse

* µ = np
* σ = √(npq)

**4.2) Aproximación normal a la probabilidad de poisson**

Cuando λ de una distribución poisson es grande, se puede recurrir a la distribución normal para aproximar los valores de probabilidad. Esto sucede cuando λ es mayor o igual a 10

* µ = λ
* σ = √ (λ)

# **Distribución en el Muestreo**

La estadística inferencial trata de la posibilidad de estimar valores poblacionales a partir de valores muestrales. En diversas investigaciones es habitual analizar un subconjunto de elementos y generalizar las conclusiones. Estos procedimientos nos permiten que los errores de inferir conclusiones a partir de un subconjunto de datos, puedan ser conocidos y controlados.

## **5.1) Definiciones Preliminares:**

* **Población**: nombre que le damos a un conjunto de unidades de análisis que son objeto de un estudio o una investigación particular
* **Muestra:** subconjunto de la población que comparte sus características en aspectos que son de interés para la investigación.

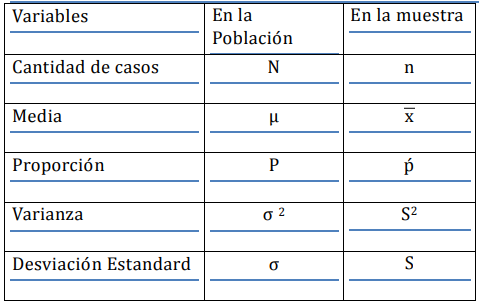
Objetivo de una investigación por muestreo: obtener información acerca de las características de la población a partir de datos de la muestra. La característica poblacional que pretende conocerse se llamarán **parámetros**. Los valores calculados de esta manera se llaman **estimadores**.

# **Estimación Puntual y por intervalos**

La importancia del carácter aleatorio en la elección de las unidades que constituirían la muestra se genera en la necesidad de generar variables aleatorias cuya probabilidad conozcamos o podamos suponer. Si podemos lograrlo conoceremos las probabilidades asociadas a resultados muestrales y podremos usarlos para inferir sobre la población.

Las medidas que calculemos a partir de datos muestrales dependeran del azar, como la media muestral, que es una variable aleatoria, su valor depende de cuales sean los casos que constituyen la muestra. Entonces nos referimos a diferentes medias cuando hablamos de media poblacional y la media muestral, siendo el mismo cálculo para las dos. La diferencia es que la media poblacional es un dato fijo, mientas que, como ya dijimos, la media muestral es una variable aleatoria que depende de que muestra hayamos tomado.

Las medidas que se refieren a la población se denominan parámetros, solo se conocen en los casos de que toda la población fue observada. Si se trabaja con variables cualitativas, es decir que se trabajará por proporciones de las categorías, se llamará proporción paramétrica o poblacional. La varianza, desviación estándar o coeficientes de correlación pueden calcularse sobre la población completa o sobre una muestra aleatoria y luego inferirlas.

Para distinguir entre las medidas de la poblacion y las de la muestra usaremos diferentes símbolos: letras latinas para identificar medidas obtenidas sobre datos muestrales y letras griegas para referirnos a los valores de la poblacion.

Al crear un grafico de frecuencias de la media muestral y analizar las probabilidades de estar cerca del verdadero valor de la media poblacional, observamos que se distribuye en forma normal, independientemente de la distribución de la variable original.

**ESTIMADORES:**

**Estimador insesgado:** cuando la esperanza matemática del estimador es igual al valor poblacional.

**Propiedades:**

**Formula de la varianza:** S2 = δ2 / n y por lo tanto S = δ / √n

**Consistencia:** Un estimador es consistente si al aumentar el tamaño muestral disminuye la dispersión resultante, es decir que el resultado de S, disminuye al aumentar n

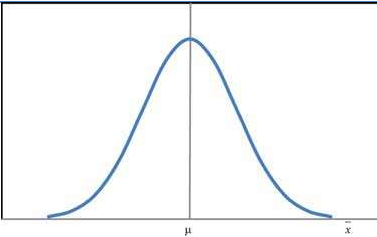
## **6.1) Tipos de Estimaciones:**

**Estimación puntual:** Se llaman así porque ofrecen un único valor como estimación del para metro de interés. Ejemplo: Una muestra de 50 medicos que egresaron en los ultimos diez años hallamos que han terminado la carrera con una nota promedio de 6,50. Si preguntamos el promedio con que terminaron la carrera todos los medicos que egresaron en los ultimos diez años, la respuesta será que “debe ser cercano a 6,50”, una respuesta imprecisa y deficiente ya que no sabemos cuan cerca puede estar la verdadera nota promedio de 6,50

**Estimación por intervalos:** Una estimación más completa de los parámetros que nos interesan. Consiste en ofrecer un intervalo acerca del cual tendremos cierta confianza que contenga al para metro. en lugar de decir que el promedio con que egresa el total de médicos de la facultad “debe ser cercano a 6,50”, construiremos un intervalo, que dirá “tenemos una certeza del 95% que el intervalo 6,10; 6,90 contiene al promedio”

**¿Cómo construyo estos intervalos para estimar la media?**

**Estimación de la media:** para construir el intervalo de confianza y para mejorar la calidad de las estimaciones puntuales uso lo conocido en distribuciones en el muestreo una muestra de 30 casos se considera suficientemente grande como para usar la distribucion normal en la estimación de la media poblacional. Si la muestra es mas pequeña no podemos usar inmediatamente la distribucion normal, sino que deberemos apelar a la distribucion t de Student.

Supongamos que el grafico de la media muestral es el siguiente:

Con esta información y recurriendo a la distribución normal estandarizada podemos estimar un intervalo de confianza donde se concentran el 95 % de los datos de esta, calculando los valores de Z, entre los cuales se encuentran el 95% de los datos en la mencionada distribución. Para ello recurrimos a la Inversa normal estandarizada y obteniendo los valores de Z.

Si queremos obtener el intervalo en la variable media muestral, recordar la definicion de la variable Z: Z = (x - µ) / δ. En este caso la variable X es M(X), por lo que nos queda Z = (X - µ) / (δ /√n), Despejando a µ que es lo que deseamos estimar, nos queda µ = X + Z \* (δ /√n) y reemplazando Z por sus respectivos valores de Z1 y Z2, así obtenemos el intervalo buscado.

# **Prueba de Hipótesis**

tiene como objetivo darnos argumentos para decidir en contextos de incertidumbre. Al realizar pruebas de hipótesis lo que haremos será evaluar cuál sería la probabilidad de hallarlo si fuera cierta la hipótesis nula. Es el caso cuando necesitamos concluir acerca de una población, a partir de información que tenemos disponible en una muestra aleatoria. El resultado de la prueba permitirá decidir si lo que se observa en la muestra es compatible con un planteo hipotético sobre la población. Podremos descartar una hipótesis por no ser compatible con lo que se observa, pero no será posible “confirmar” una hipótesis, solo podremos concluir que la evidencia no la contradice.

Una prueba de hipótesis puede compararse con un juicio: el acusado no es condenado hasta que no hay evidencia suficiente para hacerlo. En el inicio del juicio, “el acusado es inocente”, en nuestra notación llamaremos a esta condición hipótesis nula, indicada con H0. Esto indica que se trata de un estado inicial. Mientras no haya pruebas suficientes, la hipótesis nula se considerará aceptada. La decisión de condenar al acusado solo se tomará cuando haya muy poco riesgo de equivocarse. En este ejemplo, la población es el conjunto completo de información necesaria para tomar la decisión de manera certera.

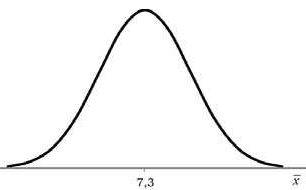
## **7.1) Prueba de Hipótesis para la media poblacional**

Situación: Para una determinada carrera universitaria, históricamente los alumnos han tardado en recibirse un promedio de 7,30 años. Históricamente para indicar que son datos acumulados en el tiempo y provienen de los registros de la facultad de años atrás. Se introdujo un cambio en el plan de estudios de la carrera y se cree que con ese cambio los alumnos tardaran un tiempo distinto en recibirse. Tenemos un promedio de la población de quienes se recibieron en las anteriores condiciones (media poblacional) y queremos inferencia sobre la media poblacional de los alumnos que cursan con el nuevo plan. Estos últimos no están todos accesibles, de esta población solo puedo conocer una muestra de los que ya han egresado y ver cuánto tiempo han tardado en recibirse.

Expresamos las hipótesis de este modo:

* H0: µ = 7.30
* H1: µ ǂ 7.30

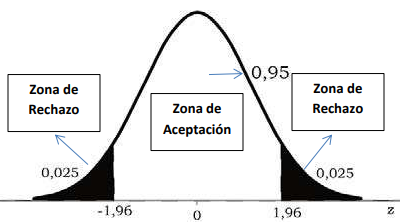
La hipótesis nula indica que la media poblacional de los alumnos que cursan con el nuevo plan es la misma que antes, pero la hipótesis alternativa afirma lo contrario. Ambas son afirmaciones sobre la población, por eso son hipótesis. Si la hipótesis nula fuera cierta, la distribución de las medias sería la siguiente:



A fin de realizar la prueba de hipótesis debemos obtener una muestra. Supongamos que seleccionamos 100 egresados y que encontramos un tiempo promedio para terminar la carrera de 7,50 años con una desviación estándar de 1,30 años. Debemos tener un criterio para decidir si este valor es compatible con la hipótesis nula o constituye evidencia para rechazarla. El criterio es el de ver cuán probable sería este valor observado si la hipótesis nula fuera cierta. Nos preguntaremos por la probabilidad que tiene la variable media muestral de asumir el valor 7,50 o uno más extremo. “Alejado” equivale a “poco probable si H0 fuera cierta”, los valores alejados se encuentran en los extremos de la distribución de la variable analizada. Para decidir si la evidencia hallada en la muestra es suficiente para rechazar la hipótesis nula, vamos a establecer un valor máximo para el a rea extrema donde consideraremos que se encuentran los valores “alejados”.

## **7.2) Criterio de toma de decisión**

H0 se rechazará si hay poca probabilidad de hallar un valor como el observado o uno más extremo que él. Ejemplo en 0,05: los valores de Z = ±1,96 delimitan un area central de 95%, es decir que dejan fuera un área de 5%. Los valores de Z superiores a 1,96 o inferiores a -1,96 tienen una probabilidad de ocurrencia de 0,05, repartida en las dos “colas” de la distribucion normal.



Los valores de la media muestral que correspondan a Z que superen a 1,96 o sean inferiores a -1,96, serán considerados como poco probables y conducirán a rechazar H0. En cambio, los valores que tengan Z entre -1,96 y 1,96 serán probables y nos llevaran a aceptar H0.

Los puntos de Z1 y Z2 se denominan valores críticos de Z y podemos indicarlo con un subíndice: Zc. En nuestro ejemplo, el valor observado es M(x) = 7,50, Lo podemos llamar M(x)aobs. El puntaje Z equivalente a ese M(x)obs se llama Z observado (Zobs) y vale:

Zobs = (M(X)obs - µ) /(S/√n) = (7,50 – 7,30) / (1,3/√100) = 1,54

Al comparar este valor con los valores extremos considerados este valor positivo es inferior a 1.96 y cae en la zona de aceptación.

El nivel de significación es la probabilidad de hallar al valor muestral en la zona de rechazo de H0, si H0 es verdadera. Se indica como α, y es elegido por el investigador.

## **7.3) Tipos de Errores en Pruebas de Hipótesis**

Siempre existe la posibilidad de tomar una decisión incorrecta. Esto sucede porque las muestras son tomadas al azar y puede que la que usamos para tomar la decisión sea una muestra extrema.

Como hemos visto, el nivel de significación mide la probabilidad de hallar un determinado resultado muestral si la H0 fuera cierta que habitualmente fijamos en 0,05 o 0,01. Si H0 es cierta y la muestra es extrema nuestra decisión será rechazar H0, esta decisión será errónea.

Hay dos tipos de errores:

Error de Tipo 1 (ETI): rechazar una H0 que es verdadera. Establecer α es afirmar que se está dispuesto a correr ese riesgo de cometer el ETI

Error de Tipo 2 (ETII): aceptemos una H0 siendo falsa. (Potencia de la Dócima)