**Introduccion Estadística con R**

La **estadística** estudia muestras o poblaciones enteras e intenta explicar la correlación y causalidad entre hechos. Importante: deben ser ambas porque puede tener correlación pero no causalidad y eso no nos valdría por ejemplo en predicciones.

**Estadística puede** **ser :**

1. **Descriptiva:** Explica los datos
2. **Inferencial:** Son los modelos para predicciones.

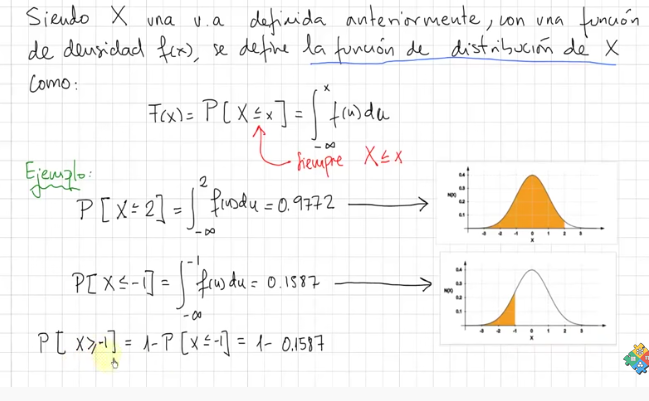
a-La Estimación: permite estimar los parámetros de la población a partir de la muestra (Ej.: ¿Qué volumen de envases recicla un hogar español?).

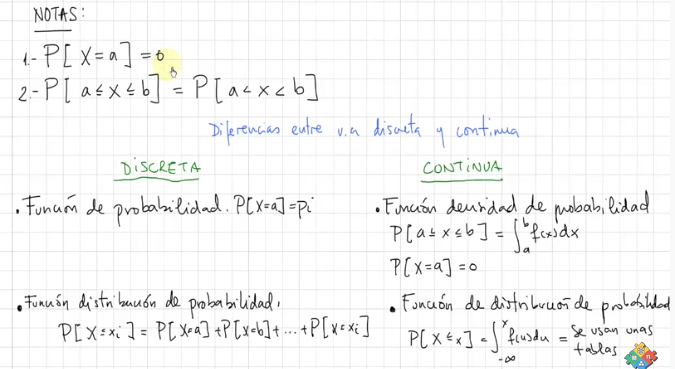
b- El Contraste de hipótesis: permite tomar una decisión sobre los parámetros de la población (Ej.: ¿Se recicla mejor que hace 10 años?)

Un experimento puede ser determinista si existe una causa que determina un resultado o aleatorio que con las mismas causas siempre tendremos resultados distintos (tirar un dado)

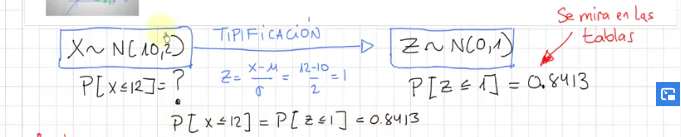
La función de **probabilidad aleatoria** asigna a cada valor una probabilidad. Las **variables aleatorias** pueden ser:

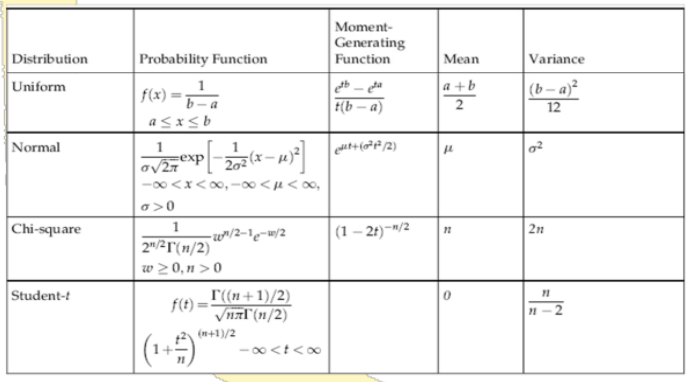
* **Continuas:** Cuando dentro del rango min-max(tiempo de espera en una cola) puede tomar cualquier valor. Se representa como una función integral





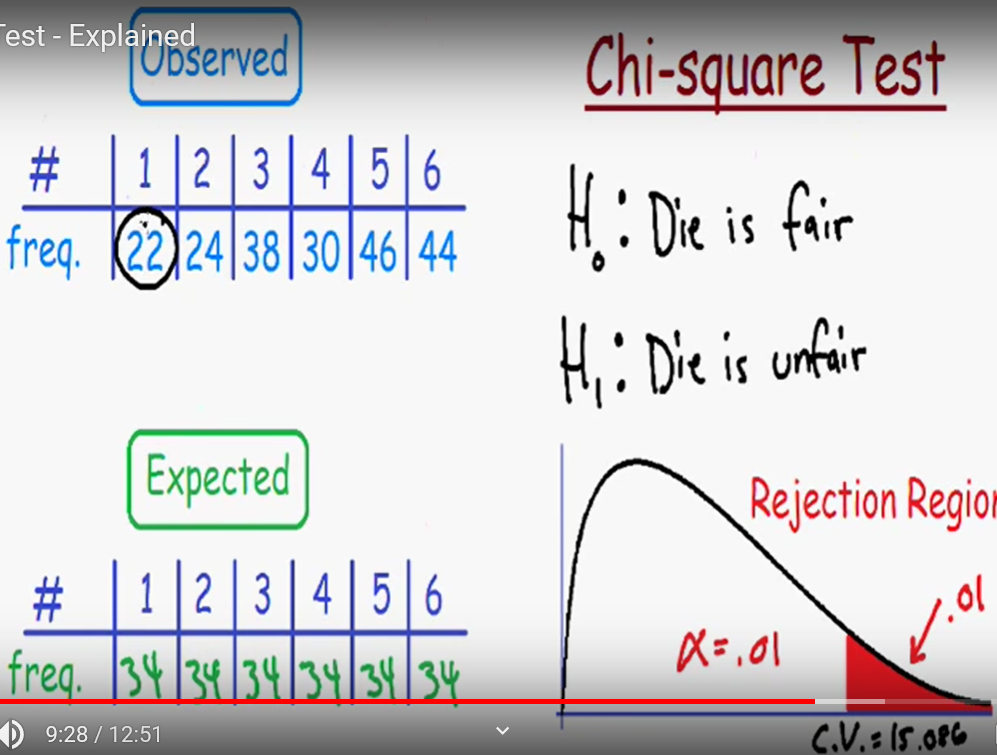
* + Distribución normal: Muchas veces hay que tipificar para convertir las variables en una normal (0,1)

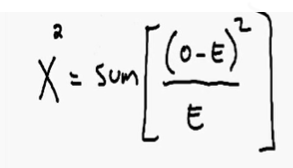


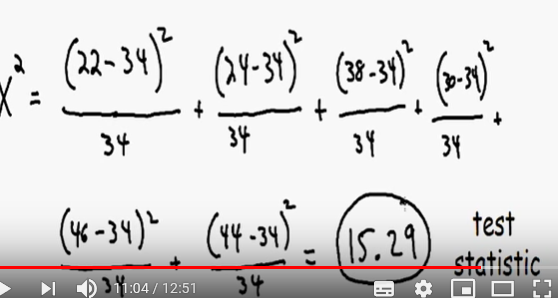


**Chi-square**: <https://www.youtube.com/watch?v=1Ldl5Zfcm1Y>

(ejemplo si un dado esta trucado o no) Sirve para comprobar si lo que se observa es igual a lo esperado. Como en el ejemplo si tiro un dado varias veces para ver si tirándolo un numero determinado de veces la t es mayor o menor al valor critico(cv,se mira en tablas) para el nivel de significación que había elegido(0,01 por ejemplo).





 rechazamos la hipótesis nula. El dado está trucado.

**t-student**: to compare two population(ei two fields of grain): <https://www.youtube.com/watch?v=pTmLQvMM-1M>

Deben tener una distribución normal, similar varianza,similar numero de observaciones, si es más de 20/30 observaciones usamos el z test.

* **Discretas:** Solo pueden tomar unos valores concretos. Una tirada de dados puede tener como resultado 1 a 6 pero nunca 1,5.Se representa como una función sumatorio
  + **Binomial**: La variable aleatoria binomial, expresa el número de éxitos obtenidos en cada prueba del experimento. Sólo puede tomar los valores suponiendo que se han realizado pruebas.

*1 En cada prueba del experimento sólo son posibles dos resultados: el suceso {A}(el cual llamamos éxito) y su contrario {\overline{A}}(el cual llamamos fracaso).*

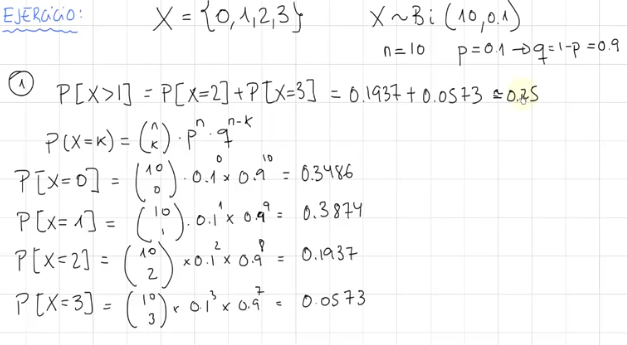
*2 La probabilidad de que ocurra el suceso {A}es constante, es decir, que no varía de una prueba a otra; esta probabilidad se representa por {p}.*

*3 El resultado obtenido en cada prueba es independiente de los resultados obtenidos anteriormente.*

De una urna, la variable \*Sacar una bola roja\* sería un proceso de Bernouilli. Y la respuesta a la pregunta: \*Si saco 2 bolas al azar, cual es la probabilidad de que las 2 sean rojas?\* Seria una Binomial.

*La distribución binomial se suele representar por {B(n, p)}, donde {n} es el número de pruebas de que consta el experimento, {p} es la probabilidad de éxito. La probabilidad de fracaso es {1-p}y la representamos por {q}.*





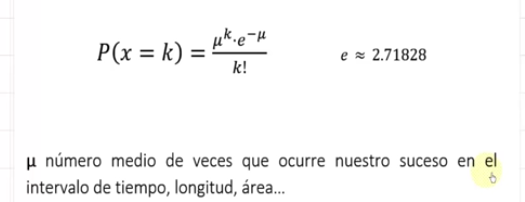
Ejemplo: La variable aleatoria \*Me tocará la loteria?\* tiene dos posibles respuestas: Si (1) o No (0)

Está variable aleatoria se llama Bernouilli y es una de las distribuciones que siguen más sucesos de la vida. La probabilidad de acierto en una Bernouilli es un número \*p\* entre 0 y 1 y la probabilidad del contrario (no sacar una bola roja) es \*1-p\*. \*Esa es una de las características de una función de probabilidad: la suma de las probabilidades de los posibles outcomes de la variable es 1\*. Así podemos modelar la probabilidad de que alguien devuelva un préstamo a un banco o no, la probabilidad de que alguien tenga una enfermedad o no o, como decíamos al principio, la probabilidad de que al lanzar una moneda al aire obtenga una cara o una cruz.

Si en realidad lo que queremos es saber cuál es la probabilidad de que me toque la loteria si tengo 10 décimos y van a sacar 1000 bolas, tenemos una distribución \*binomial\*.

dbinom(10,1000,10^-5) donde p=1/100.000, x=10, n=1000

* + **Poisson:** Nos proporciona la probabilidad de que ocurra un determinado suceso un numero de veces k en un intervalo de tiempo, longitud, área, etc.

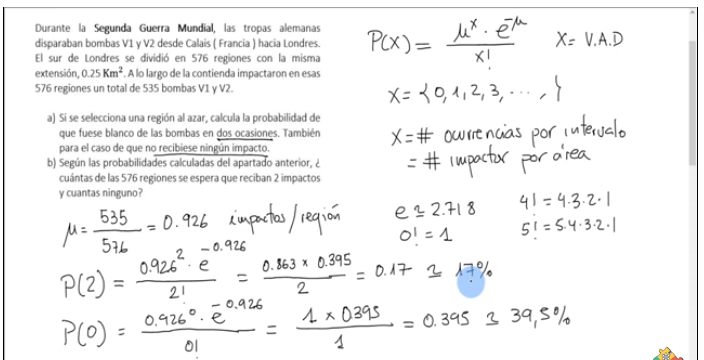


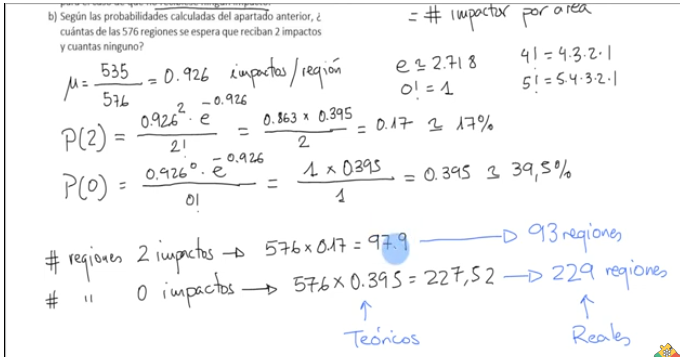
X es el suceso

K es el número de veces que se repite nuestro suceso. K! k factorial

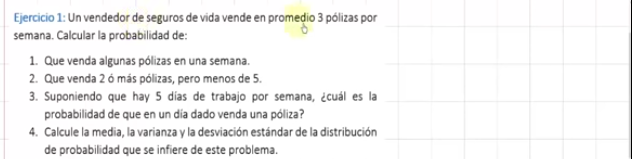
Los sucesos tienen que ser aleatorios. Los sucesos deben ser independientes unos de otros. (no podría estudiar la probabilidad de tener gripe en un colegio porque no son independientes al contagiarse unos a otros).

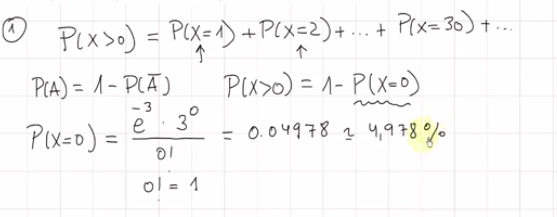
Caso real

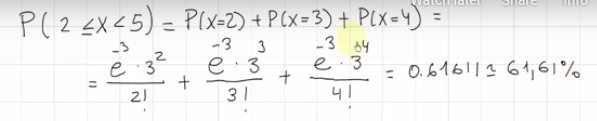


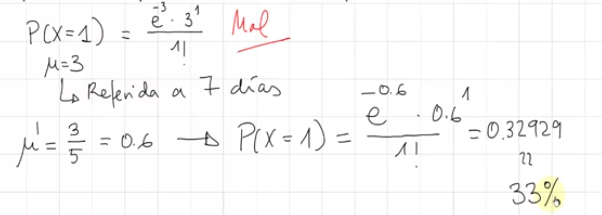


Ejercicio

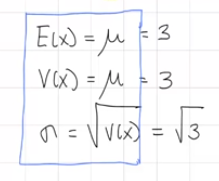


2

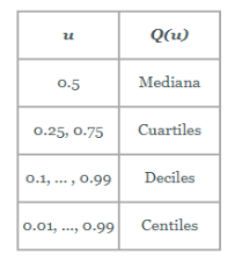
3



4



**Medidas de posición**



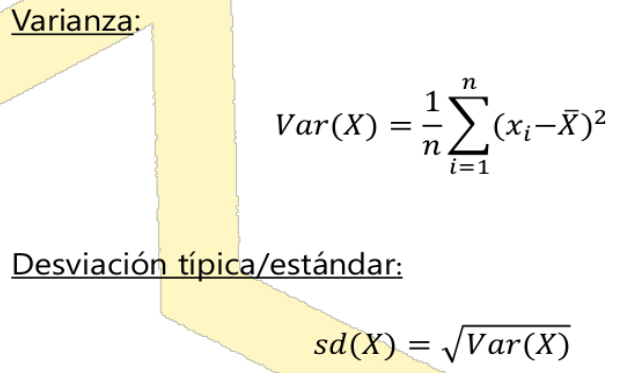
Media

Mediana

Cuartiles

Moda

**Medidas de dispersión**



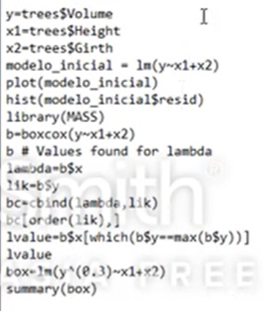
En muchas ocasiones hay que realizar **transformación de variables aleatorias :**

a-Tipificación(canónica): Es la operación de realizar una transformación lineal en la variable para que la (nueva) variable transformada, tenga por media, cero, y por desviación típica, uno. Consiste en restar a la variable la media y dividirla por la desviación típica.

*La tipificación permite la comparación de valores puntuales (puntuaciones) de dos distribuciones distintas de distintas medias y varianzas.*

*Es extraordinariamente útil para construir indicadores de comparación universal (univariantes: coeficiente de asimetría y coeficiente de curtosis y multivariantes : coeficientes de correlación).*

**b- y la transformación de Box-Cox:** para corregir sesgos en la distribución de errores, para corregir varianzas desiguales (para diferentes valores de la variable predictora) y principalmente para corregir la no linealidad en la relación (mejorar correlación entre las variables)

Una vez realizado el modelo se revosan los residuos y si no siguen la normal se realizan las transformaciones. Después se pasa el modelo de nuevo.

**YOU TUBE BUENO PARA EJEMPLOS**



**Tipos de muestreo:**

**1-Probabilistico:**

El muestreo probabilístico representa un tipo de [muestra estadística](https://enciclopediaeconomica.com/muestra-estadistica/), en este caso el método **se concentra en analizar y estudiar grupos específicos de la** [**población estadística**](https://enciclopediaeconomica.com/poblacion-estadistica/)**, pero utilizando la selección aleatoria.**

Además, **el muestreo probabilístico tiene como requisito principal que la población estudiada cuente con igualdad en las oportunidades para ser seleccionados.**

## Tipos de muestreo probabilístico

A la hora de clasificar los **tipos de muestreo probabilístico** nos encontramos con los siguientes posibles: **Muestreo aleatorio simple, muestreo sistemático, muestreo estratificado y muestreo por conglomerados.**



### **a-Muestreo aleatorio simple**

* Artículo principal: [Muestreo aleatorio simple](https://enciclopediaeconomica.com/muestreo-aleatorio-simple/).

El muestreo aleatorio simple quizás sea el método de selección más sencillo para una muestra. **Se basa en elegir los individuos de la muestra a investigar de forma aleatoria**. Es decir, se le otorga un número a cada sujeto de la población y posteriormente, a modo de sorteo, se seleccionan números de forma aleatoria, conformando de esta manera la muestra.

**Por ejemplo:** Se debe tomar una muestra de 10 escuelas sobre una población de 200 escuelas que hay en el país. Por lo tanto, cada escuela obtiene un número al azar. A modo de sorteo se seleccionan 10 para conformar dicha muestra.

### **b-Muestreo sistemático**

* Artículo principal: [Muestreo sistemático](https://enciclopediaeconomica.com/muestreo-sistematico/).

En el muestreo sistemático s**e lleva a cabo dividiendo la población en sub-grupos en forma de lista, finalmente se selecciona aleatoriamente un número**. El número seleccionado aleatoriamente será la posición del individuo de cada grupo que conforme la muestra.

**Por ejemplo:** Tenemos una población de 10.000 personas y debemos conformar una muestra de 100 individuos. Por lo tanto, se dividen la población en 200 grupos de 50. Luego se selecciona un número al azar, por ejemplo, el número 20. Por lo tanto, la muestra de 100 individuos ser irá conformando por el individuo número 20 de cada grupo.

### **c-Muestreo estratificado**

* Artículo principal: [Muestreo estratificado](https://enciclopediaeconomica.com/muestreo-estratificado/).

En el muestreo estratificado, **el investigador divide la población en estratos o sub-grupos que, generalmente, comparten las mismas características**. Luego, de forma al azar y utilizando el método del muestreo aleatorio simple, se elige a un individuo de cada estrato para conformar la muestra.

**Por ejemplo:** Se realiza una investigación en donde la población va a estar conformada por personas que trabajan en una empresa. Por lo tanto, el investigador divide dicha población en los siguientes estratos:

* Hombre con media jornada
* Hombre con jornada completa
* Mujer con media jornada
* Mujer con jornada completa

De esta manera el investigador divide la población en 4 diferentes estratos en donde las personas que componen cada estrato comparten características.

### **d-Muestreo por conglomerados**

Este método es utilizado específicamente cuando no es práctico o es imposible elaborar una lista exacta de los elementos y detalles con los que cuenta una población. **Para llevarlo a cabo, los detalles de dicha población ya deben estar agrupados con sus poblaciones y las listas ya deben existir o se pueden crear**.

**Por ejemplo:** Un investigador decide realizar una investigación sobre México, por lo que dividir a toda la población del país en grupos sería difícil e impráctico. Por lo tanto, aprovecha y utiliza la separación natural por ciudades. A partir de cada ciudad, elige los individuos de forma aleatoria y conforma así la muestra.

## Ventajas y desventajas del muestreo probabilístico

A la hora de emplear el muestreo probabilístico nos encontramos con ciertas **ventajas y desventajas** que presenta este tipo de muestra.

### Ventajas

Las ventajas del muestreo probabilístico son las siguientes:

* Es la forma más sencilla de armar una muestra.
* La **representatividad aumenta** y para ponerlo en práctica resulta más fácil, además no se requiere marco de muestreo.
* Se toman en cuenta las subpoblaciones y tiene precisión.
* Este método **resulta económico** y finalmente ponerlo en práctica no es un problema.

### Desventajas

Las desventajas del muestreo probabilístico son las siguientes:

* A veces resulta **poco preciso** y calcularlo o interpretar los resultados es difícil.
* Resulta complicado elegir las variables de estratificación y es **costoso**.
* En algunos casos puede **disminuir el aspecto de la representatividad**.
* Muchas veces es complicado especificar el marco de muestreo, pues hay menor precisión y en muchos casos **la** **representatividad no es**

**2-NO Probabilistico**

El muestreo no probabilístico es **aquella técnica utilizada en la muestra estadística que, al contrario de la** [**muestra probabilística**](https://enciclopediaeconomica.com/muestreo-probabilistico/)**, se basa en un proceso que no les permite a todos los individuos de una** [**población**](https://enciclopediaeconomica.com/poblacion-estadistica/) **investigada tener la misma oportunidades de ser seleccionados**.

Es decir, este tipo de muestreo predomina a aquellos individuos que al cumplir con cierta cualidad o característica benefician a la investigación, entonces pueden ser parte de la muestra.

A la hora de **clasificar al muestreo no probabilístico**, podemos encontrar los siguientes tipos: **Muestreo por cuotas, conveniencia, bola de nieve o discrecional**.



Tipos de muestreo no probabilístico

### Muestreo por conveniencia

* Artículo principal: [Muestreo por conveniencia](https://enciclopediaeconomica.com/muestreo-por-conveniencia/).

El muestreo no probabilístico por conveniencia es aquel que, tal como lo indica su nombre, **el investigador realiza la muestra en base a las selección de individuos que considera accesibles, fácil y de rápida investigación**. Esto, generalmente, lo hace por proximidad al investigador.

**Por ejemplo:** El investigador decide realizar un estudio sobre la opinión de un profesor en un aula determinada. Al utilizar el muestreo por conveniencia, conforma su muestra para encuestar en base a los primeros 5 alumnos de la lista del aula.

### Muestreo por cuotas

* Artículo principal: [Muestreo por cuotas](https://enciclopediaeconomica.com/muestreo-por-cuotas/).

En este, **el investigador se asegura de formar la muestra de forma que ésta sea equitativa y proporcional de acuerdo a las características, cualidades o rasgos de la población a estudiar**.

**Por ejemplo**: El investigador debe realizar una muestra sobre empleados de una empresa en donde 60% son mujeres y 40% son hombres. Por lo tanto, el investigador debe realizar la muestra seleccionando individuos para que dicha muestra sea proporcional a la población. Dicha selección la hace a través de un muestreo por conveniencia o a elección del investigador.

### Muestreo de bola de nieve

También conocido como muestreo en cadena, este método de muestra no probabilística es a**quella en donde el encargado de la investigación le exige al primer sujeto identificar o señalar a otra persona que tenga potencial y que cumpla con los requisitos de dicha investigación**.

**Por ejemplo:** Un investigador decide realizar una investigación en donde la muestra va a estar conformada por individuos que posean una rara enfermedad. Por lo tanto, al encontrar un individuo que posea dicha enfermedad, el investigador le pide ayuda a este individuo para encontrar otros con sus mismas características, conformándose de esta manera la muestra.

### Muestreo discrecional

También conocida como **muestreo por juicio o intencional**. En este caso **se consiguen las muestras y los sujetos se eligen para llegar a ser un grupo en donde la muestra busca un objetivo específico**. En este particular, **el encargado de la investigación sabe que algunas personas resultan más adecuadas para el análisis que otras**.

**Por ejemplo:** Se desea realizar una investigación sobre el comportamiento de los padres con sus hijos. Por lo tanto, el investigador seleccionara los individuos para su muestra en base a personas que tengan hijos, ya que lo considera aptos de conocimiento para formar parte de la muestra deseada.

## Ventajas y desventajas del muestreo no probabilístico

### Ventajas

Las ventajas que suele presentar el muestreo no probabilístico son:

* Menor costo para realizar la investigación.
* Se pueden controlar las características de la muestra.
* Conlleva menor tiempo, ya que se conoce el individuo que va a formar la muestra.
* Se pueden conocer las características poco comunes.

### Desventajas

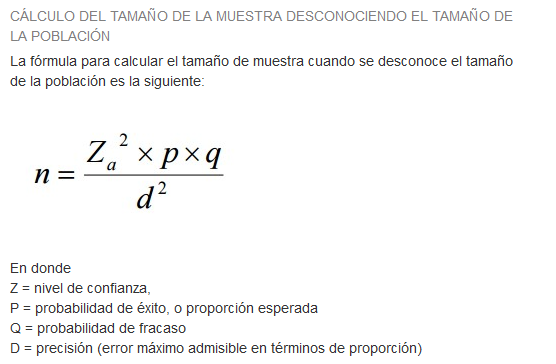
Las desventajas de utilizar el muestreo no probabilístico son:

* No es una muestra que asegure la representación total de la población.
* No generaliza y es subjetivo.
* No es recomendable en el caso de que la investigación sea causal o descriptiva.

**Como determinar el tamaño de la muestra:**

1. **Si no conozco la población total**

**Para expresarlo en proporciones. n amas de casa a encuestar para comrpobar si compran un producto x**

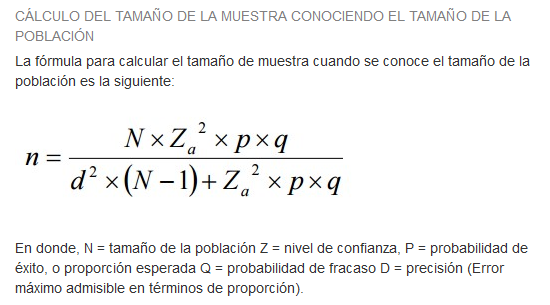


**Ejemplo: para un nivel de confianza del 95% y un error del 5%(0,05) cuántas amas de casa debería encuestar?**

El nivel de confianza para una confianza del 95 en tablas es 1,96. Además si no me dicen nada o no hay encuesta previa p=q= 0,5.

Por tanto quedaría 1.96^2 x 0,5^2 / 0,05^2 = 384,16 que seria mi tamaño de muestra necesario

1. **Si conozco el tamaño población**



**Ejemplo: para un nivel de confianza del 95% y un error del 5%(0,05) para una poblacion de 2500 familias. Cuántas amas de casa debería encuestar?**

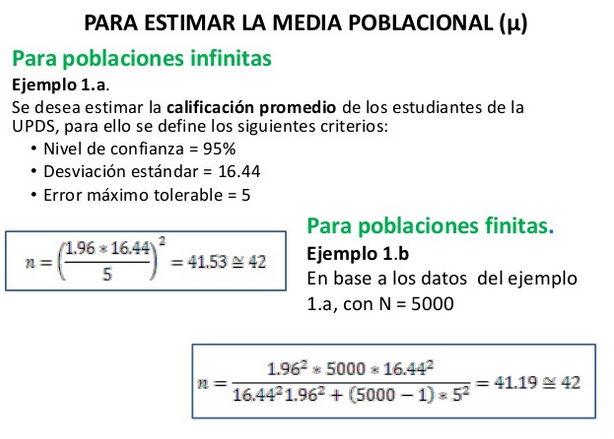
**En una encuesta anterior p=12% por tanto q=0,88**

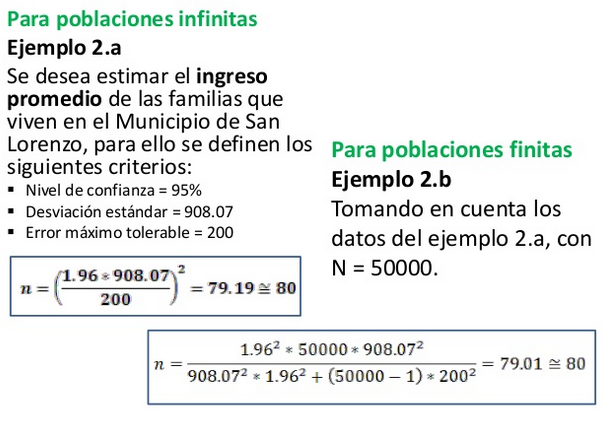
El nivel de confianza para una confianza del 95 en tablas es 1,96. Además si no me dicen nada o no hay encuesta previa p=q= 0,5.

Por tanto quedaría 2500x1.96^2 x 0,12x0,88 / 2499x0,05^2+1.96^2x0,12x0.88 = 153 que seria mi tamaño de muestra necesario

**Ejemplos para expresarlo de forma cuantitativa y no en porcentaje. Promedio de gasto de estudiantes en una poblacion**

**Ejemplos**





**Esperanza matematica**

También llamado valor esperado. Es lo que nos permitira saber si jugar o no a algo por su probabilidad de ganar o perder. Si es cero el juego es equitativo para todos.

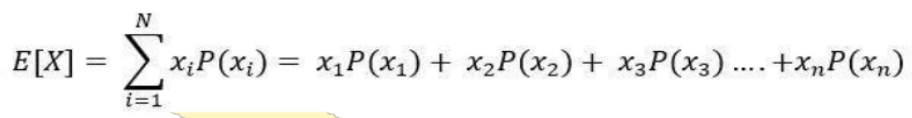
**Ejemplo:** Si tengo una bolsa con 8 bolas. 6 rojas y 2 negras. Si sale roja pierdo 50€ y si sale negra gano 200€

La probabilidad de perder sería 6/8 y la de ganar 2/8

Xi sería -50 o 200 dependiendo si sale roja o negra.

La esperanza matematica sería 6/8\*(-50) + 2/8\*(200)= 12,5€ en media ganaría

En distribuciones discretas con la misma probabilidad en cada suceso la media aritmética es igual que la esperanza matemática



Dónde x es el valor del suceso, P la probabilidad de que ocurra, i el periodo en el que se da dicho suceso y N el número total de periodos u observaciones.

**Teoremas de Bayes**

Dados dos sucesos A y B con P(B)>0, definimos la probabilidad de A condicionada a B como la probabilidad de que ocurra A sabiendo que ha ocurrido B.

P(A|B) Probabilidad que se de A dado que se ha dado B

P(A|B)= P(A)\*P(B|A) / P(B) Orden facil. Si primero aparece A, en la formula primero pondremos A y despues seguimos el orden ABAB.

La probabilidad tiene que ser distinto a cero.

**Ejemplo:** en un consultorio el 40% son consultas de cliente que fingen estar enfermos. El 10% son hombres. La probabilidad de que un paciente finja estar enfermo dado que es hombre es del 50%

P(F) = 0,4

P(H)=0,1

P(F|H)=0,5

P(H|F)=? Ojo no es lo mismo que P(F|H)

P(H|F)= 0,1\*0,5/0,4 = 0,125 por tanto 12,5% de probabilidad de sea hombre si alguien ha fingido una enfermedad.

**Repetición clase con Fatima**

La distribución empírica es la distribuciónde los datos de la muestra.

Los procesos estadisticos independientes (Sacar un bola de una urna, ponerla de vuelta y sacar otra) serían simulaciones de Montecarlo.

Siempre que nos enfrentemos a un situación donde tenemos que medir el numero de éxitos o fracasos de una muestras estamos trabajando con una binomial.

Cada vez que saco una bola de una urna es una bernoulli y si quiero saber cuantas rojas y cuantas azules me han salido de la urna entonces es uan binomial.

Probabilidad me toque la loteria bernoulli y si llevo varios numeros la probabilidad que me toque alguno de ellos sería binomial.

Si tienes varias posibilidades sería multinomial.(Ejemplo a quien votas en las elecciones). O con el ejemplo anterior que en lugar de una bola azul y un roja tuviera en la urna bolas de mas colores.

**Función de densidad** se utiliza para varibles categoricas. En las funciones continuas la probabilidad de que ocurra un evento concreto sería cero.

P(X=K)

dbim()

**Función de distribución.**

(Px<=k)

pbinom()

pnorm()

para los quantiles.

qbinom()

**La random** rbinom()

**Inferencia estadísticas**

Siempre tendremos una hipótesis que queremos descartar o no según los datos.

Cualquier suceso cuyo resultado no pueda ser predicho con total seguridad es un proceso estocástico.

El resultado cuantitativo de un proceso estocástico es una \*variable aleatoria\*. El hecho de que el resultado de un proceso tenga una componente estocástica, de incertidumbre no significa que sea un proceso totalmente azaroso: se puede modelar la incertidumbre y controlarla. \*\*Este es uno de los principales trabajos de un data scientist: cuantificar el azar, la estocasicidad. La inferencia estadística propone un marco teórico para hacerlo\*\*

Un proceso aleatorio puede ser de dos tipos dependiendo de lo que está intentando medir: una variable \*\*continua\*\* o una \*\*categórica/discreta\*\*. La medida del segmento es una variable aleatoria continua y tirar la moneda al aire es una variable aleatoria categórica.

**## Errores sistemáticos y errores aleatorios en los datos**

\* Errores sistemáticos son aquellos que se pueden estimar y posteriormente transformar los datos para eliminarlos.Por ejemplo, si la regla con la que medimos nuestro segmento está mal calibrada podemos calibrarla y corregir nuestras medidas con estos valores

\* Errores aleatorios son aquellos introducidos de manera aleatoria en nuestro datos. No podemos estimarlos con total certeza pero puede que sepamos como modelarlos. Por ejemplo, los errores de medicion de los que hablábamos antes siguen una distribución normal, que en ocasiones se denomina "ruido blanco". Otra fuente importante de error aleatorio que podemos modelar y tener en cuenta en nuestros datos es el error de muestreo.

**## Población y muestras aleatorias: simulaciones de Monte Carlo**

En general, con la inferencia estadística intentamos llegar a conclusiones generales a partir de una muestra de datos limitada. Esto introduce una incertidumbre en nuestras conclusiones (qué pasaría si en lugar de haber llamado a estas 1000 casas para preguntar su intención de voto hubiera llamado a otras 1000?). Pero esta incertidumbre se puede modelar y tener en cuenta en nuestra toma de decisiones.

**## Probabilidad discreta: Procesos estocásticos categóricos**

**### Distribución Binomial**