

**DigitalHouse** >  
Coding School

# DATA SCIENCE

UNIDAD 1  
MÓDULO 2

Nociones de Estadística  
Inferencial

Agosto 2017

1

**Definir los términos población-muestra;  
parámetro-estimador**

2

**Comprender el concepto de muestreo y de  
distribución muestral**

3

**Calcular estimaciones puntuales y por  
Intervalos de Confianza bajo Muestreo  
Aleatorio Simple**

4

**Presentar algunas nociones generales acerca  
de las pruebas de hipótesis**

# POBLACIÓN-MUESTRA



- En los problemas de diferentes disciplinas se estudia el comportamiento de varias variables definidas sobre un conjunto de objetos. El conjunto de objetos será denominado población.
- Sobre esos elementos se observan variables, indicadas  $X_1, X_2, \dots, X_k$ , que son características que cambian de individuo a individuo.
- Una forma de simbolizar las “N” unidades de la población es:
  - $\{U_i \text{ donde } i = 1 \dots N\} = (U_1, U_2, \dots, U_i, \dots, U_N)$
- La población se define en relación al problema de investigación a abordar.
  - Consideremos el conjunto P de votantes en una determinada elección donde se presentan 3 candidatos, que denominamos 1, 2 y 3. Podemos notar  $X(a)$  como el número del candidato votado por a.
  - Por ejemplo: asalariados de la CABA, explotaciones de la región pampeana, empresas textiles de la provincia de Chaco

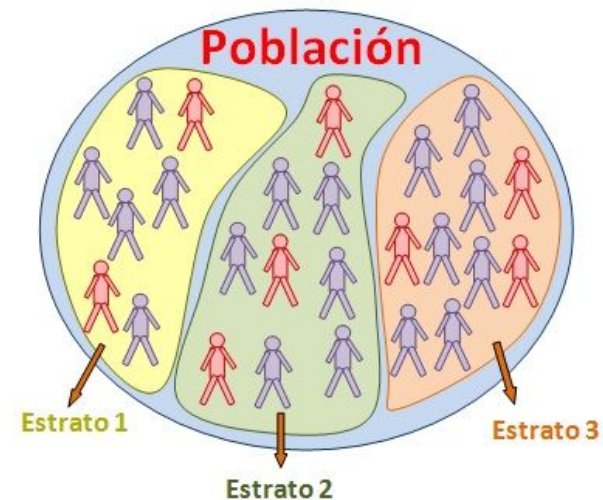
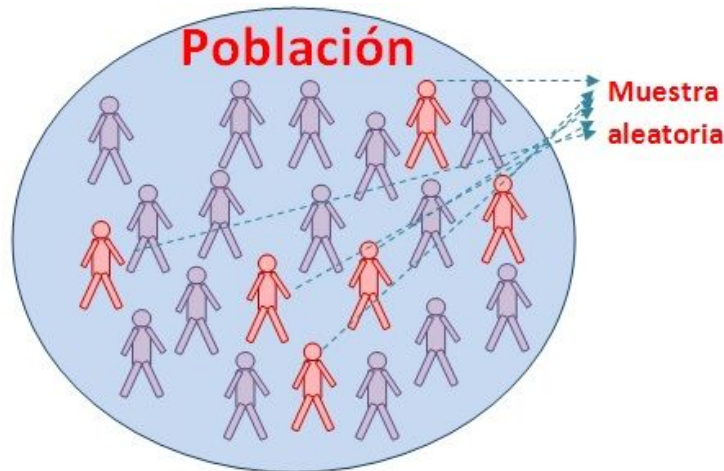
- En muchos problemas interesa la distribución de una variable aleatoria  $X$  que se observan cada vez que se repite un mismo experimento
- En estos casos, cada elemento a estudiar corresponde al resultado de un experimento y al menos teóricamente se puede repetir el experimento tantas veces como se quiera.
- Se puede pensar entonces en una población infinita compuesta por los infinitos posibles experimentos que se pueden realizar, aunque tal población no tiene existencia real.
  - Por ejemplo: El experimento consiste en tirar una moneda y  $X$  vale 0 ó 1 según caiga ceca o cara.

- **Muestra:** es seleccionada de la población (definida en relación al problema de investigación). Es el subconjunto de unidades seleccionadas de la población definida.
- En ésta recae la realización de las observaciones, mediciones, etc. Las “n” unidades o Muestra seleccionada de una Población de “N” se simbolizan:
  - $\{ u_i \text{ donde } i = 1...n \} = (u_1, u_2, \dots, u_i, \dots u_n)$

- **Probabilísticas:** puedo calcular la probabilidad de selección de cada una de las unidades de la muestra. => Puedo calcular una medida del error

### Algunos tipos:

1. [Muestreo aleatorio simple](#)
2. [Muestreo aleatorio estratificado](#)



- **No probabilísticas:** la muestra no probabilística no es un producto de un proceso de selección aleatoria. Los sujetos en una muestra no probabilística generalmente son seleccionados en función de su accesibilidad o a criterio personal e intencional del investigador.
  - **Muestreo por conveniencia**
  - **Muestreo discrecional**

Ejemplo Un hospital desea hacer un estudio para testar la eficacia de su nueva vacuna contra la gripe que acaba de patentar un laboratorio farmacéutico. **Realizan el estudio sobre sus pacientes porque así al hospital le supone menos costes económicos.**





# PARÁMETROS, ESTIMACIONES Y ESTIMADORES



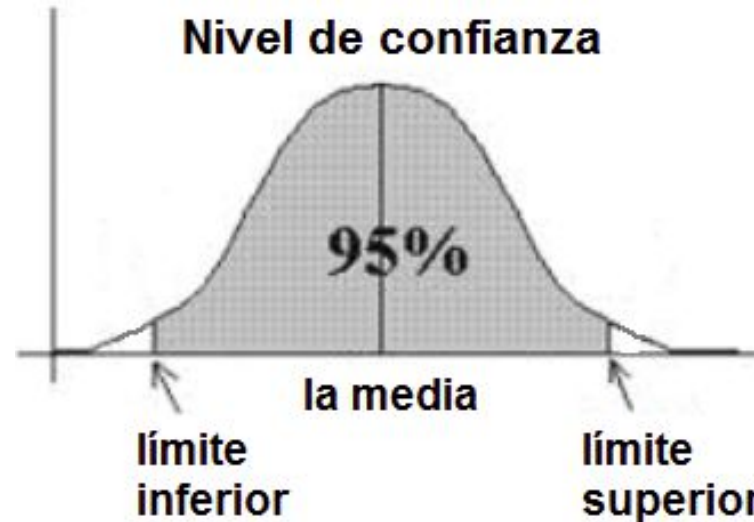
- El objetivo siempre es “estimar” alguna característica de la población. que se supone fija (no aleatoria).
- Pueden ser características simples como:
  - una media, una proporción, una varianza
- O medidas más complejas, como por ejemplo:
  - los coeficientes de una regresión o la asociación entre variables
- A esta característica se la llama **parámetro**.
- En general, podemos considerar a los parámetros como relativamente “constantes” (en tiempo y espacio).
  - Esto los diferencia de los **estimadores** que veremos a continuación.

- Un estimador es un estadístico (esto es, una función de la muestra) usado para estimar un parámetro desconocido de la población.
- Por ejemplo, la **media muestral** es un estimador de la **media poblacional** que se calcula tomando el promedio de los datos.
- Qué estimador usar depende (entre otras cosas) de dos factores:
  - Del parámetro a estimar
  - Del diseño muestral (muestreo aleatorio simple, muestreo estratificado, etc.)

- Los estimadores son **variables aleatorias**. No conozco cuánto va a valer la media muestral antes de tomar la muestra.
- Puedo conocer qué valores puede tomar e incluso con qué probabilidad (según el diseño muestral del experimento, el muestrista “elige” las probabilidades de observar cada muestra).

- **Estimación puntual:** La estimación puntual consiste en utilizar el valor de un estadístico (alguna función de los datos) que denominaremos estimador para calcular el valor de un parámetro desconocido de una población. En la estadística clásica esos parámetros se consideran fijos (no aleatorios).
  - Por ejemplo, cuando usamos la media muestral para estimar la media de una población, o la proporción de una muestra para estimar el parámetro de una distribución binomial.
- Una estimación puntual de algún parámetro de una población es un solo valor obtenido a partir de un estadístico.

- En general, hay :
  - **Estimación por intervalos de confianza:** se da una “franja” de valores posibles. Generalmente, se da un límite inferior (Li) y otro superior (Ls) tal que la probabilidad de que el parámetro se encuentre entre Li y Ls es conocida.



- Dos grandes tipos de errores afectan a una muestra:
  - **Errores debidos al muestreo (o muestrales):** consecuencia de observar parcialmente a la población.
    - Si  $n \rightarrow N$  los errores tienden a disminuir.
    - Imposibles de anular al menos que  $n=N$ .
  - **Errores NO debidos al muestreo:** todos aquellos que no provienen del hecho de estar trabajando con una muestra. En general, afectan tanto a muestras como a censos. Ejemplos:
    - Información, suministrada u obtenida, errónea.
    - Falta de respuesta, en general se presenta por: negativa, respondente ausente o desconocer la respuesta, etc.
    - Información mal cargada en la base de datos.
    - Errores de procesamiento de la información de la base.
    - Errores de diseño de la muestra, por ejemplo marcos de selección incompletos que hacen surgir errores de cobertura.

# DISTRIBUCIONES MUESTRALES





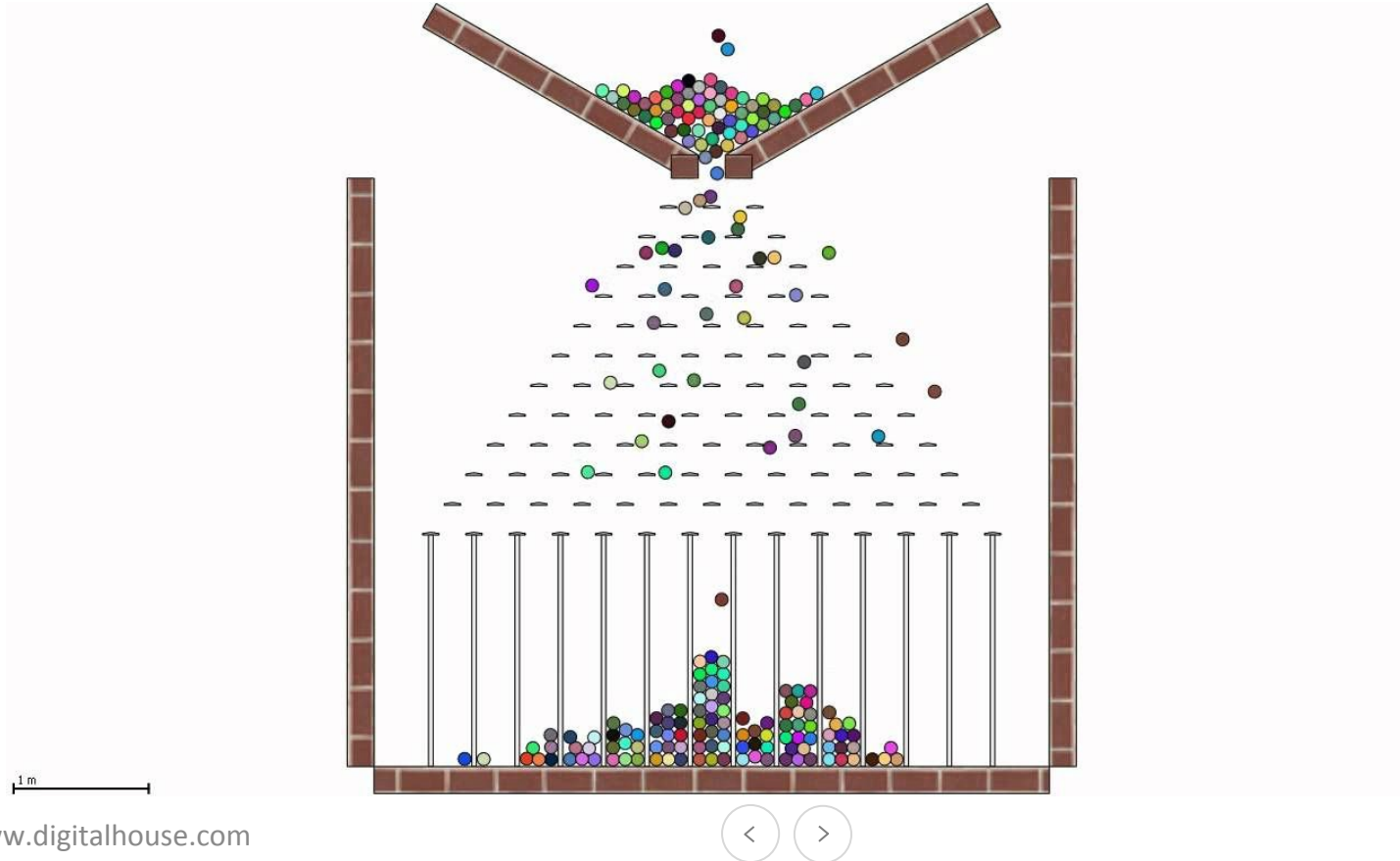
- **Concepto clave: distribuciones muestrales** de... estimadores (medias, proporciones, coeficientes de regresión, de correlación... etc.).
- Distingamos tres tipos de distribuciones:
  - Distribución de la variable (X) en la población. En muestreo se consideran no aleatorias a las características (X) de las unidades.
  - Distribución de la variable (x) en la muestra
  - Distribución muestral -de un estimador- de todas las muestras posibles de tamaño n fijo de una población. Es inducido por el esquema de muestreo.
- Es un **concepto TEÓRICO**: Para construir, por ejemplo, la distribución muestral de medias necesitaríamos todos los individuos de la población, con todos los valores de la variable y poder extraer **TODAS** las muestras posibles. Esto es imposible. Para una población de  $N=10.000$ , hay aproximadamente  $6,5208E+241$  muestras posibles de tamaño 100 (sin reposición)

# Práctica Guiada 1: extrayendo muchas muestras y “aproximando” una distribución muestral de medias

# TEOREMA CENTRAL DEL LÍMITE



- <http://vis.supstat.com/2013/04/bean-machine/>
- <https://www.mathsisfun.com/data/quincunx.html>



- Si se obtienen sucesivas muestras de tamaño  $n$  (fijo), de una población cuya variable en estudio ( $X$ ) se distribuye de forma normal, la **distribución muestral de medias** de esa distribución será normal con media igual a la media de la población y con desvío estándar igual  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ 
  - Este resultado no depende del TCL y es exacto (no asintótico)

- **Ley de los Grandes Números:** ¿Por qué tiene sentido estimar la media poblacional (valor esperado) con la media muestral?
- Supongamos que las observaciones son extracciones de la misma distribución (en particular con igual valor esperado) e independientes entre si
- La LGN nos garantiza que a medida que el tamaño muestral aumenta la media muestral se acerca a la media poblacional.

- Teorema Central del Límite. Conexión con muestreo.
  - El teorema central del límite nos da una distribución para nuestro estimador de la media poblacional dado por la media muestral.
  - Si las observaciones venían de una distribución normal entonces la distribución será exactamente normal con *media*  $\mu$  y *varianza*  $\sigma^2 / n$  sin importar si el  $n$  es suficientemente grande o no.
  - Recuerde que si estimamos la media muestral y el desvío estándar muestral entonces el error estándar de la media muestral es desvío estándar muestral dividido la raíz del tamaño muestral.
  - El error estándar nos permite medir dispersión respecto a la media muestral pero ajustando por el tamaño muestral (podemos comparar dispersión para estimadores de la media que usan diferentes tamaños muestrales).

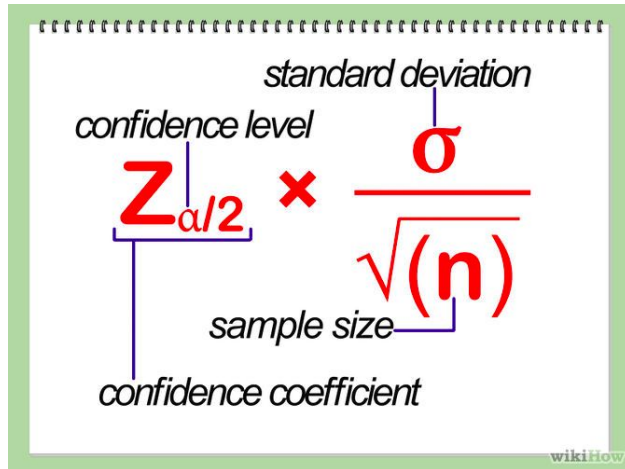
# INTERVALOS DE CONFIANZA



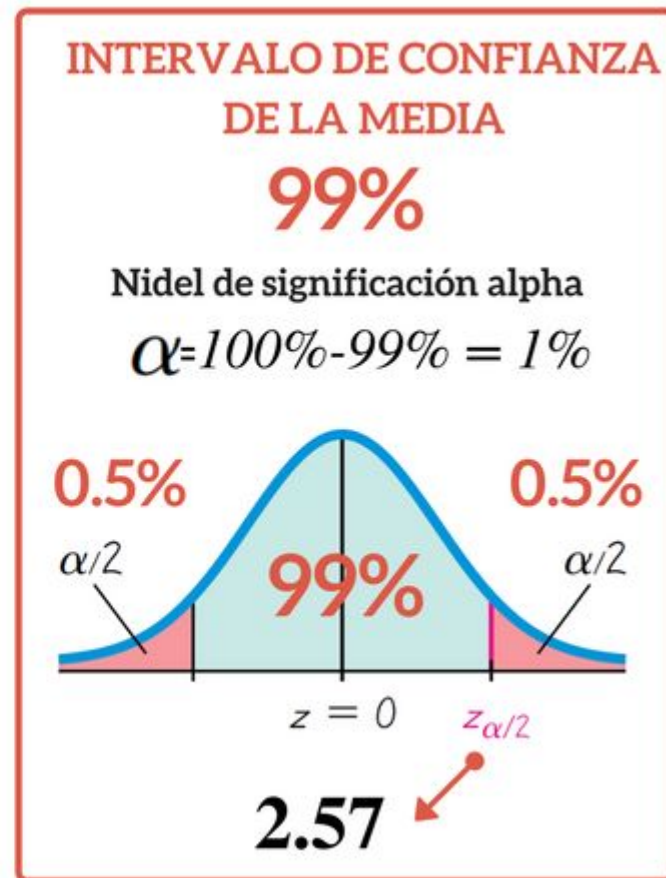
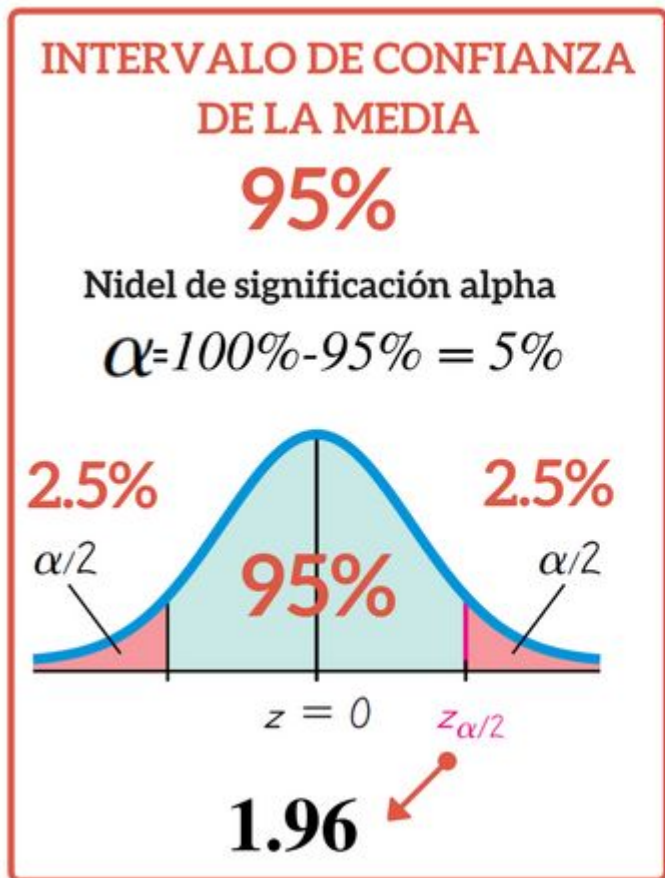


- **Media de la muestra:** estimación puntual => es el valor del estadístico en esa muestra
  - No sirve de mucho. ¿Por qué?
- **Estimación por intervalos:** la idea es dar un rango de valores posibles para el parámetro con un valor de probabilidad asociado.
  - «El parámetro de la población está entre 4,5 y 8,2 con una probabilidad de 95%»
- **¿Cómo se logra?** A partir de poder conocer (o estimar) la forma de la distribución muestral del estimador.

$$\bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{(n)}}$$



A diagram illustrating the components of the confidence interval formula. The formula is shown as  $Z_{\alpha/2} \times \frac{\sigma}{\sqrt{(n)}}$ . Labels with lines pointing to the corresponding parts are: 'confidence level' pointing to  $Z_{\alpha/2}$ , 'confidence coefficient' pointing to  $Z_{\alpha/2}$ , 'standard deviation' pointing to  $\sigma$ , and 'sample size' pointing to  $(n)$  under the square root. The diagram is enclosed in a green border with a spiral top edge, and a 'wikiHow' watermark is visible at the bottom right.



- Se desea analizar el peso medio al nacer de los terneros Shorton. A partir de una muestra al azar de 100 terneros cuyo promedio es de 30 Kg. y desvío estándar de aproximadamente 7 Kg. , se desea construir un intervalo de confianza del 90% para el peso promedio poblacional.
- Dado que n puede ser considerado

“grande” =>  $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$   $IC(\mu; k = 0.90) = 30kg \pm 1.645 * \frac{7kg}{\sqrt{100}}$

$$IC(\mu; k = 0.90) = 30kg \pm 115kg$$

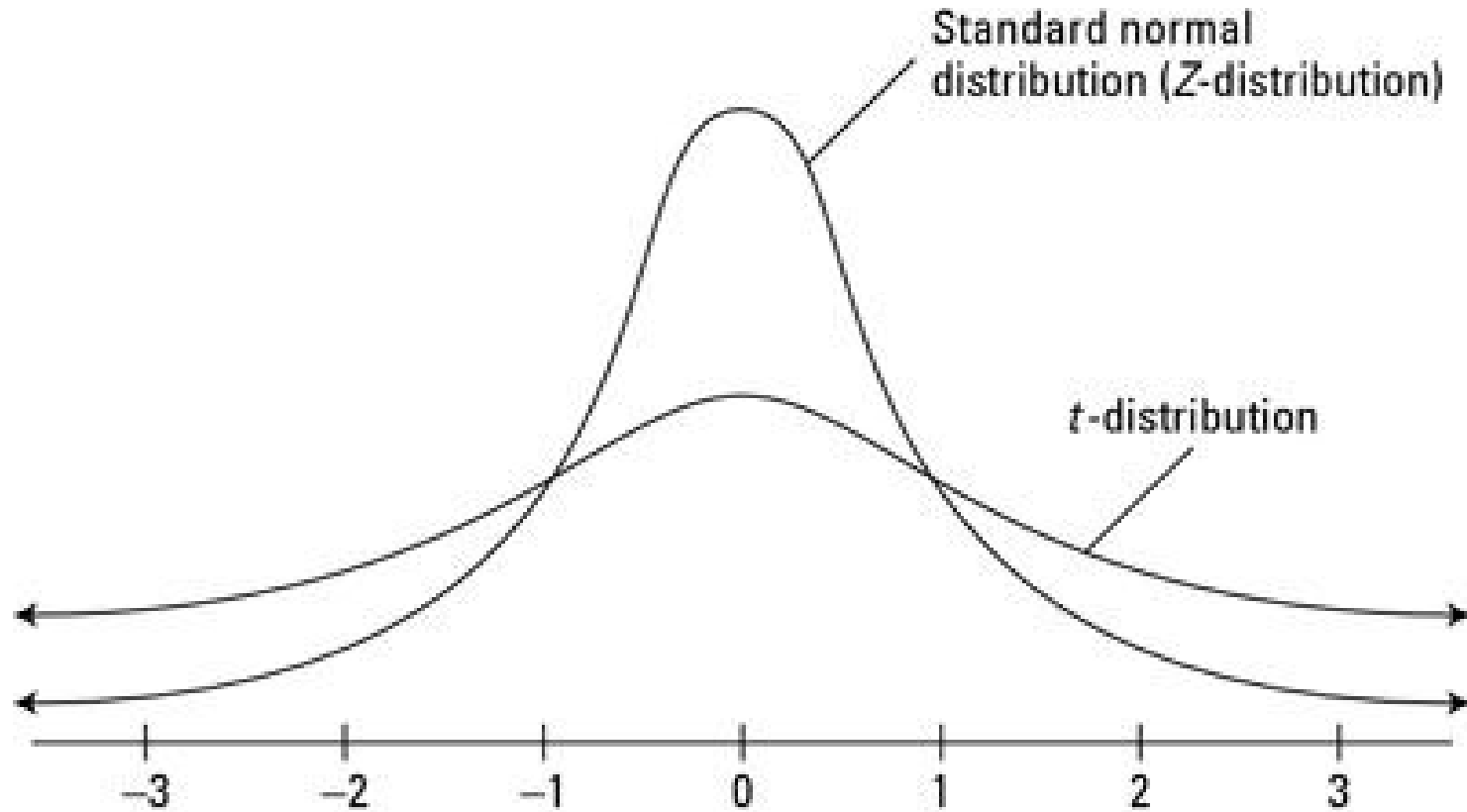
$$L_i = 28.85kg$$

$$L_s = 31.15kg$$

$$P(28.85kg \leq X \leq 31.15kg) = 0.90$$

- Observaciones.

- ¿Y si no conocía la varianza poblacional? Es una fuente más de incertidumbre que debo incorporar si en la derivación anterior cambiamos el desvío estándar poblacional por el muestral la distribución de la variable normalizada pasa a ser una T-Student.
- La distribución T-Student es simétrica pero dado un valor  $\alpha/2$  que se busca acumular en cada cola el valor de la distribución T-Student que acumula esa probabilidad es mayor en valor absoluto que en la normal estándar. ¿Intuición? Esto genera intervalos más anchos para incorporar el hecho de que el desvío estándar muestral es un estimador del poblacional.



- Observaciones.

- El intervalo de confianza es siempre función de la muestra que es aleatoria. Entonces los bordes del intervalo también son variables aleatorias! En este caso el borde del intervalo depende de la media muestral (que cambia con cada muestra).
- El intervalo de confianza no puede depender de parámetros que sean desconocidos. En nuestro ejemplo usamos un estimador del desvío poblacional o asumimos conocido el desvío poblacional.

# PRUEBA DE HIPÓTESIS





- ¿Sabemos que las estimaciones que hacemos con base a una muestra pueden ser considerados como variables aleatorias (¿por qué?).
- Hipótesis sobre la media de los Klout Scores
- Estimar la intención de voto (proporción de personas que votarían a un candidato) en dos momentos del tiempo (mediante dos muestras diferentes)
  - ¿Cuál es la probabilidad de que la diferencias observadas entre esos valores provengan de poblaciones con parámetros diferentes?
  - ¿Las diferencias observadas son producto del azar o las poblaciones son diferentes?

- Siempre se realizan sobre algún parámetro de la población.
  - “el otorgamiento de becas de estudio ha logrado disminuir el porcentaje de estudiantes que abandonan la escuela secundaria”
    - ¿Cuál es el parámetro?
    - ¿Cuál la variable?
- Para testear (de forma estadística) una hipótesis se desdobra en dos:
  - Hipótesis nula ( $H_0$ ): basada en los conocimientos previos de la población que se desean comprobar.
  - Hipótesis alternativa ( $H_a$ ): es la que se tomará como (probablemente) cierta en caso de que a partir de los datos de la muestra se derive en el rechazo de la  $H_0$

## — Hipótesis unilaterales

### ○ Izquierda:

■  $H_0: \mu = \mu_0$

■  $H_a: \mu = \mu_1 < \mu_0$

### ○ Derecha:

■  $H_0: \mu = \mu_0$

■  $H_a: \mu = \mu_1 > \mu_0$

## — Hipótesis bilaterales

■  $H_0: \mu = \mu_0$

■  $H_a: \mu = \mu_1 \neq \mu_0$

1. Formulación de  $H_0$  y  $H_a$
2. Selección de una prueba/test apropiado
3. Determinación de un nivel de error/significancia (probabilidad de rechazar  $H_0$  o  $\alpha$ )
4. Suponer que  $H_0$  es cierta y asumir una determinada distribución muestral “centrada” en la hipótesis nula
5. Establecer la regla de decisión (dado el nivel de significatividad tendré valores críticos que separan la región de no rechazo de la de rechazo).
6. Calcular el estadístico de prueba bajo la  $H_0$  y bajo los valores observados
7. Contratastar ambos
8. “Decidir”

Decisión Adoptada	Hipótesis $H_0$	
	Cierta	Falsa
No rechazar $H_0$	<i>Decisión Acertada</i> $(1 - \alpha)$	<i>Error de Tipo II</i> $(\beta)$
Rechazar $H_0$	<i>Error de Tipo I</i> $(\alpha)$	<i>Decisión Acertada</i> $(1 - \beta)$

- Nivel máximo de error que estamos dispuesto a tolerar.
- Dado que nos encontramos trabajando con muestras nunca podemos estar 100% seguros de si hemos tomado la decisión correcta al rechazar o no una  $H_0$ .  
¿Por qué?

Decisión Adoptada	Hipótesis $H_0$	
	Cierta	Falsa
No rechazar $H_0$	<i>Decisión Acertada</i> $(1 - \alpha)$	<i>Error de Tipo II</i> $(\beta)$
Rechazar $H_0$	<i>Error de Tipo I</i> $(\alpha)$	<i>Decisión Acertada</i> $(1 - \beta)$

- alfa y beta varían en forma inversa... ¿por qué?

- Imaginemos que sospechamos que una moneda está cargada (y un amigo nos está trampeando). Queremos testarlo... para eso, decidimos hacer 10 tiradas.
- Una estrategia es contrastar hipótesis
- Puedo construir una **distribución muestral** de cada una de las tiradas del dado suponiendo que la moneda está equilibrada (esa sería la  $H_0$ ).
- De hecho, podemos hacerlo utilizando la distribución binomial...

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq p \leq 1$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

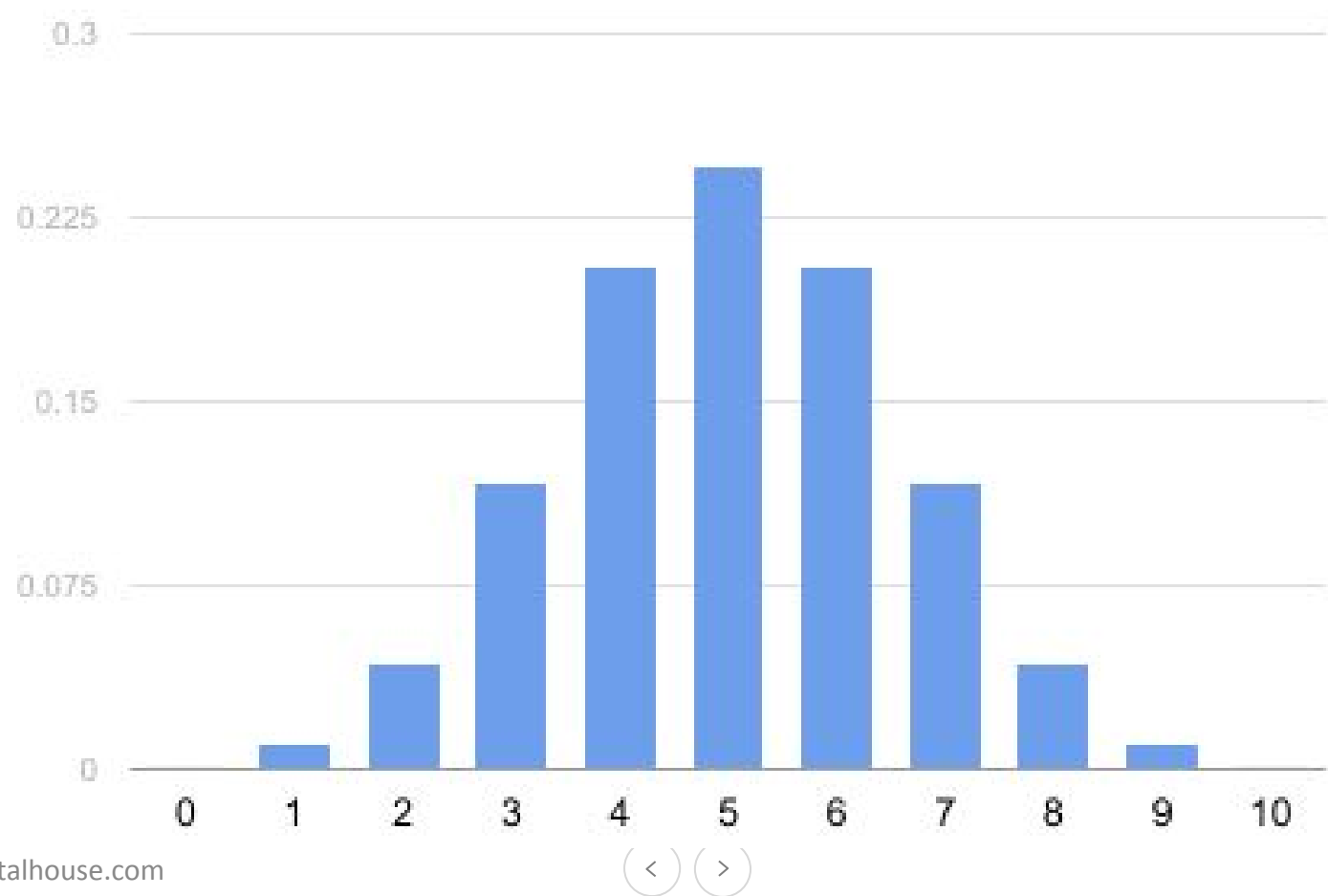
Para 0 caras = 1 sola posibilidad

Para 1 cara =  $10!/1! \cdot 0.5^{**1} * 0.5^{**9}$



- Puedo calcular las probabilidades de obtener 0, 1, 2,..., 10 caras en 10 tiradas de monedas.
- La forma es a través de una distribución binomial que permite calcular las probabilidades de obtener x éxitos en n repeticiones de un experimento -¿cuál sería el experimento, cuál el éxito y cuántas repeticiones habría en este caso?
- ¿Qué suceso es más probable (**si la moneda NO estuviese cargada**)?
- ¿Cuándo rechazaríamos que la moneda **NO está cargada**?

#	Fav / Pos	Prob
0	1 / 1024	0,001
1	10 / 1024	0,010
2	45 / 1024	0,044
3	120 / 1024	0,117
4	210 / 1024	0,205
5	252 / 1024	0,246
6	210 / 1024	0,205
7	120 / 1024	0,117
8	45 / 1024	0,044
9	10 / 1024	0,010
10	1 / 1024	0,001



- La anterior es la distribución de probabilidad de obtener 0 a 10 caras en 10 tiradas de monedas si la moneda no estuviera cargada... es decir...

### **SI LA HIPÓTESIS NULA FUERA VERDADERA**

- Ahora, nosotros realizamos el experimento (o extraemos la muestra).
  - Definimos un alfa de 0.1
  - Tiramos la moneda 10 veces y obtenemos 7 caras...
  - ¿Qué dirían sobre la  $H_0$ ?
  - ¿Y si hubiéramos obtenido 2 caras en 10 tiradas?
- ¿Qué pasaría si la  $H_0$  NO fuese verdadera?
- Las pruebas de hipótesis se basan en el supuesto de que  $H_0$  es verdadera...

- También conocido como “el concepto más citado, menos entendido y peor explicado de la estadística”.
- Tests de hipótesis: argumento por contradicción
  - mostrar que mantener  $H_0$  (que se asume siempre como cierta) lleva a conclusiones absurdas y  $\Rightarrow$  debe ser rechazada.
- Problema: determinar cuánto de la variación entre datos observados y  $H_0$  puede atribuirse a variación aleatoria (porque trabajo con una muestra).

- Muy relacionado a la distribución de probabilidades (muestral) de un estadístico bajo el supuesto de que  $H_0$  es verdadera.
  - “La probabilidad de obtener el estimador observado o más extremos si la hipótesis nula fuese cierta”
- Medida de qué tan probable serían los datos observados si la hipótesis nula fuera cierta.
- Supongamos que dado un conjunto de datos se computa el estadístico de prueba y el p-value resulta de 0.001. Partamos de que  $H_0$  es cierta e imaginemos a otros investigadores repitiendo el experimento en idénticas condiciones.
- Ese valor del p-value dice que si  $H_0$  es cierta solo 1 de cada 1000 investigadores puede obtener un valor del estadístico tan extremo como el obtenido.

- El p-valor es el menor valor de significación para el que rechazamos  $H_0$  para una muestra dada (que genera cierto valor observado del estadístico de prueba)
- El cálculo del p-valor para una muestra depende tanto de la distribución del estadístico bajo  $H_0$  como de la regla de decisión elegida.
- Si  $p\text{-value} < \text{nivel de significación}$  entonces rechazo  $H_0$ . Forma alternativa de fijar la regla de decisión. No necesito buscar valores críticos en una tabla.
- Notemos que el p-value no es la probabilidad de que  $H_0$  sea cierta. Sin importar la cantidad de repeticiones del experimento  $H_0$  es siempre cierta o falsa.
- El p-value da la probabilidad de obtener evidencia en contra de  $H_0$  (aleatoriamente) asumiendo que  $H_0$  es cierta. Cuanto menor sea el p-value más evidencia en contra de  $H_0$  tenemos siempre asumiendo que  $H_0$  es cierta.

- En un pueblo de la provincia de Catamarca, hasta hace 10 años la población joven había ido paulatinamente emigrando, con lo cual el total había ido disminuyendo y envejeciendo. La promoción industrial parece haber modificado esa situación, pero no se tienen datos fehacientes sobre el tema. Previo a la realización de un estudio demográfico que lo incluye, se plantea el supuesto de que en dicho pueblo el promedio de personas por vivienda es de 2,5, mientras que en la Intendencia Municipal consideran que esa cifra es muy baja y que en realidad el valor medio por vivienda es superior. Con el fin de verificar la  $H_0$  se realiza una muestra de 256 viviendas con la que se obtienen los siguientes resultados:
  - Media=2,68 personas
  - Desv media=0,8 personas

$$H_o : \mu = \mu_o = 2,5$$

$$H_a : \mu = \mu_1 > 2,5$$

- Se desea realizar la prueba con un nivel de significación del 10% ( $\alpha=0,10$ )
- n puede ser considerado grande ( $n=256$ ) =>
- Si  $H_o$  es cierta =>  $\bar{x} \sim N(\mu = 2,5; \frac{0,8}{\sqrt{256}})$
- Se desconoce la dispersión de la población y se utiliza la de la muestra como estimación
- $P(x > x_c) = 0,10$ ;
- Entonces, calculamos los puntajes z necesarios.
  - $Z_c = 1,28$
  - $X_c = 2,5 + 1,28 * 2,5 = 2,568$  personas =>  $X_c < X_o$  => rechazo  $H_o$ .



- ¿Qué significado tiene el  $\alpha = 0,10$  ?
- Si el valor de  $\mu$  es 2,5, y si se tomasen muchas muestras de tamaño  $n = 256$  de la población, debe esperarse que en un 10 % de las veces ( $100 \cdot \alpha$ ), se encuentre un valor del estadístico de prueba y en los casos en que esto se da debe rechazarse  $H_0$ . La probabilidad  $\alpha$  también se conoce como el nivel de significación. Esto implica que la evidencia muestral es tal que garantiza el rechazo de  $H_0$  a un nivel dado de  $\alpha = 0,10$ .
- Qué se hubiese hecho si la muestra observada hubiese dado un promedio de personas por familia menor a 2,5 (por ejemplo 2,38) Dada la  $H_1$   $\mu = \mu_1 > 2,5$  ; si la muestra hubiese dado menor no sería necesario realizar la comparación entre  $x$  y  $c_x$  , ya que lógicamente el valor muestral “cae” en la zona de No Rechazo.
- Si el nivel de riesgo  $\alpha$  hubiese sido menor (por ejemplo del 5 %), ¿hubiese podido cambiar la conclusión del problema ?

- Para definir sobre la prueba solicitada se compara:  $x = 2,68 > x_c = 2,568$  ; es decir que la media muestral supera el Valor Crítico, y que se está dando un resultado poco probable si fuese cierta  $H_0$ , por lo que se concluye que NO ES CIERTA Y QUE DEBE SER RECHAZADA.
- Esto quiere decir que el promedio de personas por vivienda en el pueblo es superior a las 2,5 personas propuestas como  $H_0$ , y que en realidad la población del pueblo no parece haber disminuido tanto como se suponía (si bien esto no fue verificado, puede suponérselo como resultante).

# Práctica Guiada 3: Una prueba de hipótesis de medias