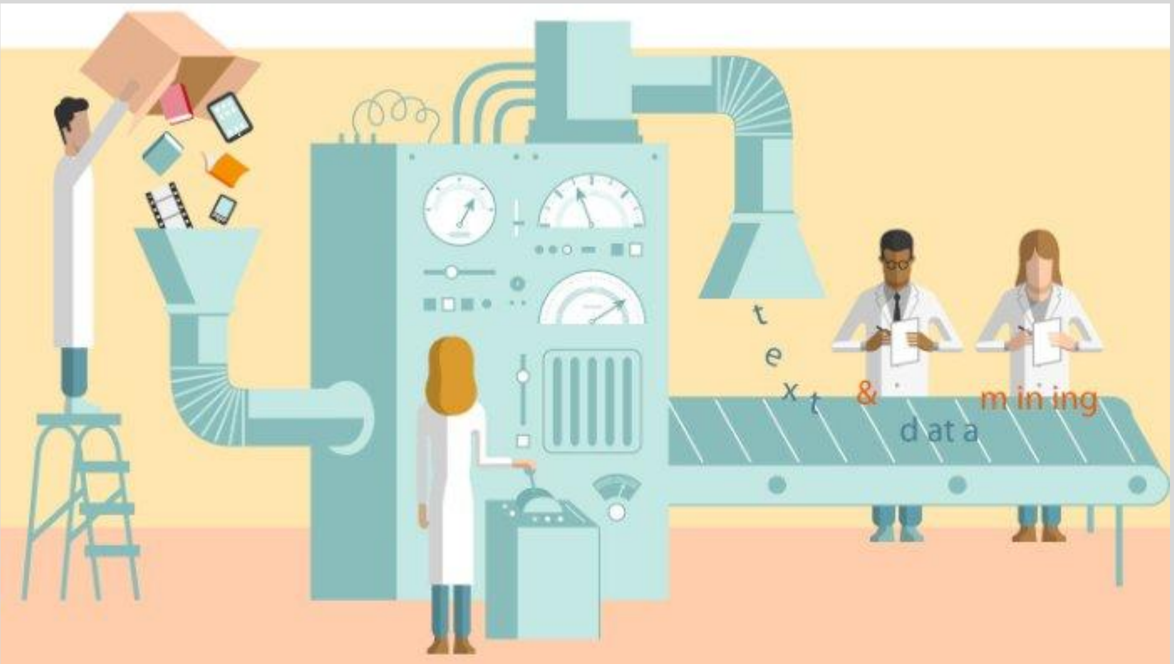


DigitalHouse >
Coding School

DATA SCIENCE

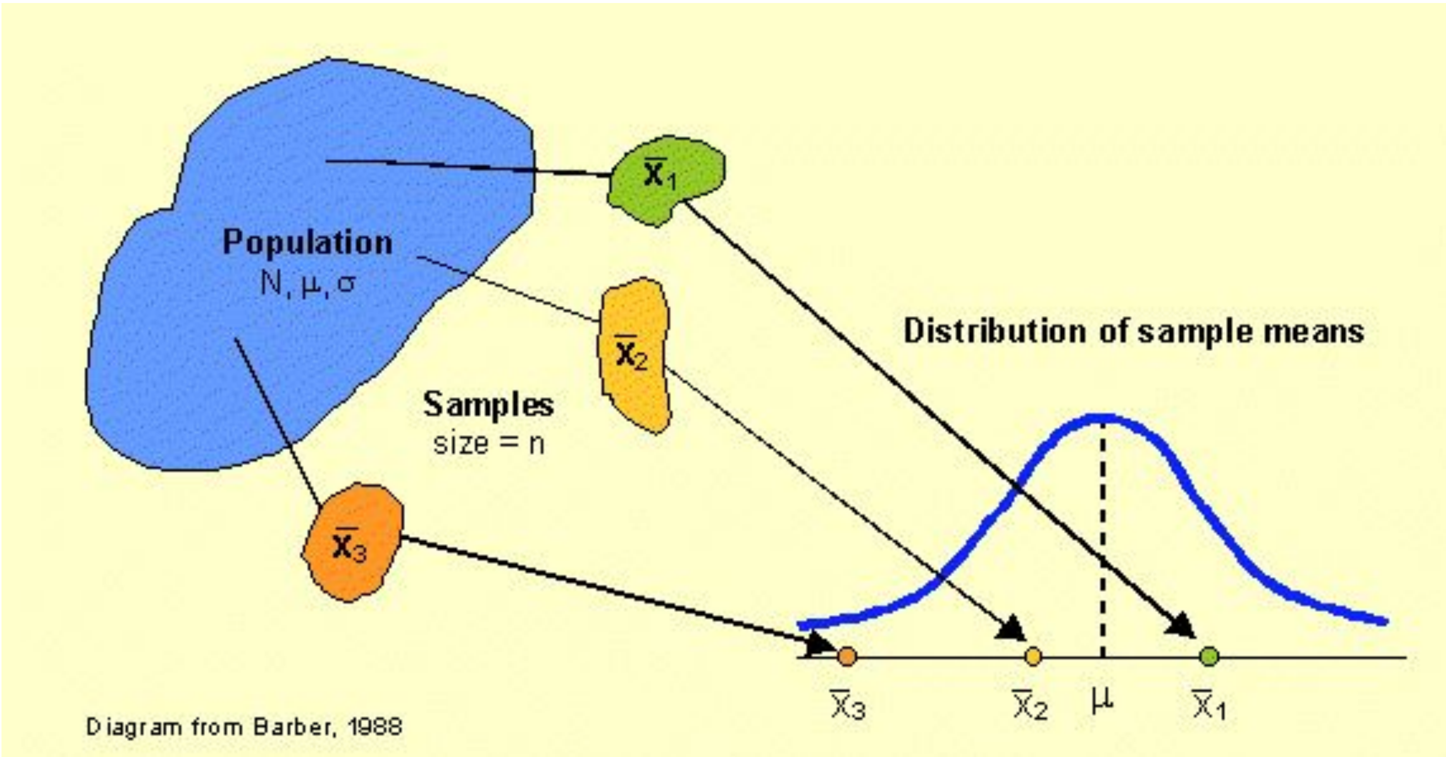
MÓDULO 2

Inferencia Estadística II

1. Pruebas de hipótesis para una población.
2. Prueba de hipótesis para dos poblaciones. Caso de grupo de tratamiento y de control.
3. Aplicación A/B Testing.

PRUEBA DE HIPÓTESIS

Distribución de las medias muestrales:
Simulación:
https://gallery.shinyapps.io/CLT_mean/



Teorema (del límite central): Sea $X_1, X_2, ..., X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas de una distribución con media **μ** y varianza **σ²≠0**. Entonces, si **n** es suficientemente grande, la variable aleatoria

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Condiciones del TCL

- **Independencia**: las observaciones tienen que provenir de un **muestreo aleatorio** y deben ser **independientes**. Si el muestreo es sin reemplazo, entonces n < 10% de la población.
- Si la distribución de una población es **muy asimétrica**, **n deberá ser muy grande** (a veces se menciona un n=30 como aproximación, pero en realidad depende de cuán asimétrica es la población). Si la población tiene distribución normal, no hay condiciones sobre n.

Un investigador de mercados y hábitos de comportamiento afirma que el tiempo que los niños de tres a cinco años dedican a ver la televisión cada semana es en promedio **22 horas.** Frente a este estudio, una empresa de investigación de mercados cree que la media es mayor y para probar su hipótesis toma una **muestra de 64 observaciones** procedentes de la misma población, obteniendo como resultado **una media de 25 y desvío estándar de 6 horas**. Verifique si la afirmación del investigador es realmente cierta utilizando un nivel de significación del 5%.

¿Qué sabemos?

- Tenemos una hipótesis sobre la media poblacional: $\mu = 22$
- Tenemos una muestra: $n = 64, \bar{X} = 25\ hs, S = 6\ hs$

Queremos saber si la muestra nos puede proveer **evidencia para rechazar nuestra hipótesis** sobre la media poblacional.

¿Qué sabemos?

- Tenemos una hipótesis sobre la media poblacional: $\mu = 22$
- Tenemos una muestra: $n = 64, \bar{X} = 25\ hs, S = 6\ hs$

Queremos saber si la muestra nos puede proveer **evidencia para rechazar nuestra hipótesis** sobre la media poblacional.

Planteamos nuestra hipótesis:

$$H_o: \boldsymbol{\mu} = 22$$
$$H_a: \boldsymbol{\mu} > 22$$

¿Qué sabemos?

- Tenemos una hipótesis sobre la media poblacional: $\mu = 22$
- Tenemos una muestra: $n = 64, \bar{X} = 25\ hs, S = 6\ hs$

Proponemos un estadístico de prueba:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

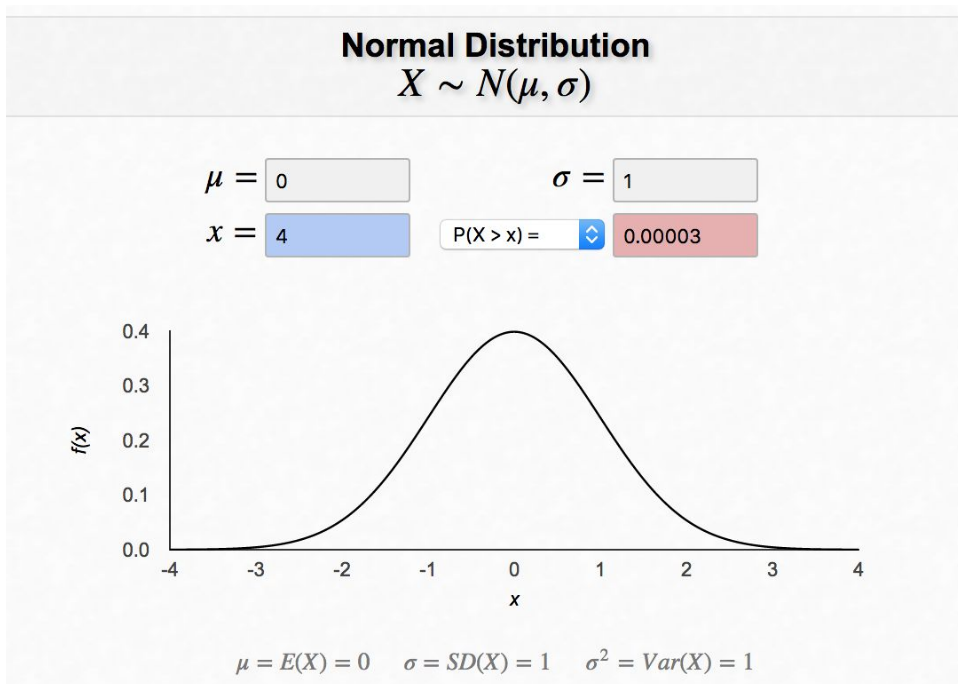
¿Qué sabemos?

- Tenemos una hipótesis sobre la media poblacional: $\mu = 22$
- Tenemos una muestra: $n = 64, \bar{X} = 25\ hs, S = 6\ hs$

Proponemos un estadístico de prueba:

$$Z_{prueba} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{25 - 22}{6/\sqrt{64}} = 4$$

Vemos que la **probabilidad** de haber obtenido una muestra aleatoria de tamaño 64 y media muestral 25 hs, asumiendo que la media poblacional sean 22 hs, es **casi nula**. **Tenemos argumentos para rechazar la hipótesis nula** y afirmar que muy probablemente la media poblacional sea mayor a 22 hs semanales.



1. Las hipótesis: siempre sobre un **parámetro poblacional.**

2. Hipótesis **nula** vs. hipótesis **alternativa.**

3. Nivel de significación (**alpha**).

- Planteo más general. Pasos en un test de hipótesis:

1. Formular las hipótesis nula y alternativa.
2. Identificar el estadístico de prueba apropiado y su distribución bajo la hipótesis nula (asumiendo que ésta es cierta).
3. Fijar el nivel de significación (la probabilidad de equivocarme al rechazar H_0 que estoy dispuesto a tolerar).
4. Establecer la regla de decisión (dado el nivel de significación, tendré valores críticos que separan la región de no rechazo de la de rechazo).
5. Recolectar datos y calcular el valor muestral del estadístico de prueba.
6. Tomar la decisión estadística.

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Value of the population parameter under } H_0}{\text{Standard error of the sample statistic}}$$

Definición de **p-value**:

- “La probabilidad de obtener el valor observado o más extremos del estadístico de prueba si la hipótesis nula fuese cierta”

Si p-value < nivel de significación, entonces rechazo H_0. Forma alternativa de fijar la regla de decisión. No necesito buscar valores críticos en una tabla.

- Supongamos que, dado un conjunto de datos, se computa el estadístico de prueba y el p-value resulta de 0.001. Partamos de que H0 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, sólo 1 de cada 1000 investigadores puede obtener un valor del estadístico tan extremo como el obtenido.

- 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.

TESTS PARA UNA POBLACIÓN

- Especificaciones de las hipótesis nula y alternativa:
  - Hipótesis unilaterales

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$$
$$H_a: \boldsymbol{\mu} > \boldsymbol{\mu}_0$$

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$$
$$H_a: \boldsymbol{\mu} < \boldsymbol{\mu}_0$$

  - Hipótesis bilaterales

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$$
$$H_a: \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

- Estadístico de prueba:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Bajo H_o, manteniendo que H_o es verdadera, este estadístico tiene distribución normal estándar (con media 0 y desvío 1).

Nota: si la varianza no es conocida pero podemos asumir la normalidad de la distribución de la media muestral, podemos utilizar al desvío estándar de la muestra como estimador de sigma.

Ejemplo:

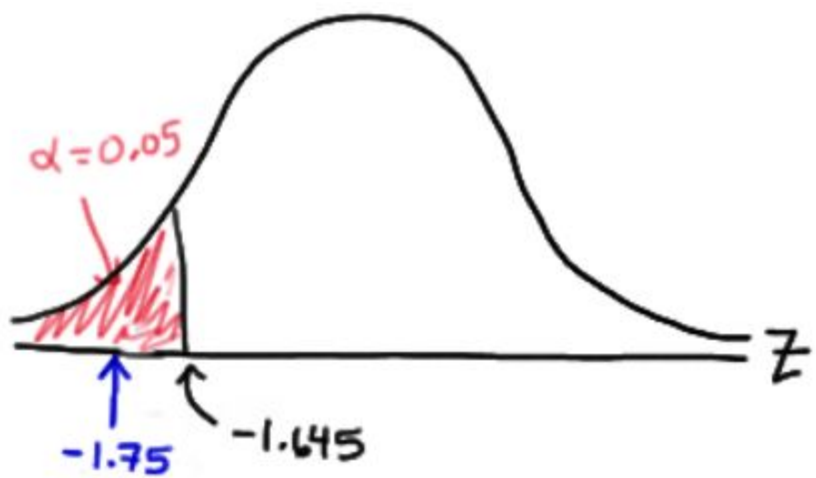
- Fijamos nivel de significación (**alpha**) en 5%.
- Se obtiene una muestra de **n=25** donde la **media muestral** es **80.94.** Por otra parte, se sabe que el desvío poblacional es 11.6.

  H_o: $\boldsymbol{\mu} = 85$
  H_a: $\boldsymbol{\mu} < 85$

- El valor del estadístico resulta:

$$Z = \frac{80.94 - 85}{11.6/\sqrt{25}} = -1.75$$

- El enfoque de la región crítica nos dice que rechacemos la hipótesis nula en el nivel α = 0.05 si Z < -1.645. Por lo tanto, rechazamos la hipótesis nula porque Z = -1.75 < -1.645 y, por lo tanto, cae en la región de rechazo:



- Otra forma de tomar la decisión es mediante el cálculo del p-value:

- En este enfoque rechazamos H_0 si el p-value es menor que el nivel α = 0.05. **En este caso, el p-value es Probabilidad(Z <- 1.75) = 0.0401 < 0.05**

Si no conocemos la varianza poblacional y no podemos asegurar las condiciones de normalidad de la distribución de la media muestral, entonces utilizamos **el estadístico de prueba T-Student con n - 1 grados de libertad**:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Ejemplo:

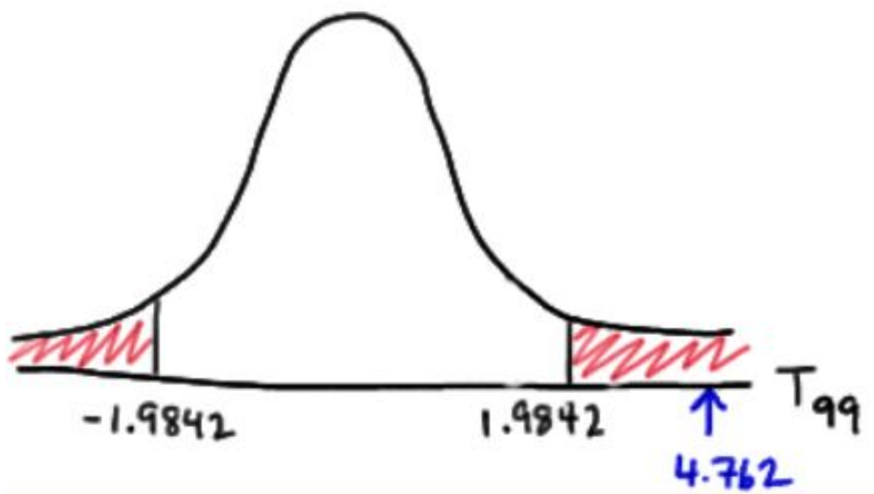
- Muestra de **n = 100**
- Media muestral de 130.1 con una desviación estándar de 21.21.

Ho: $\boldsymbol{\mu} = 120$
Ha: $\boldsymbol{\mu} \neq 120$

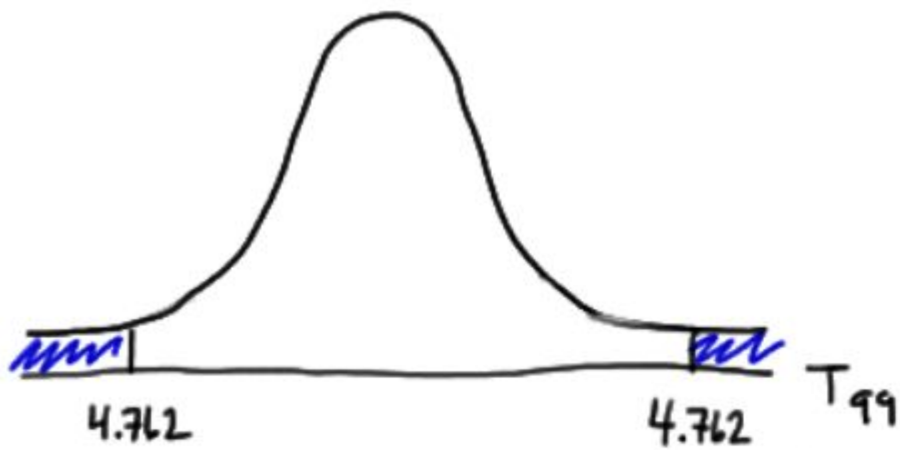
- El valor del estadístico resulta:

$$T = \frac{130.1 - 120}{21.21/\sqrt{100}} = 4.762$$

- El enfoque de la región crítica nos dice que rechacemos la hipótesis nula al nivel $\alpha = 0.05$ si $T \geq T_{0.025,\,99} = 1.9842$ o si $T \leq T_{0.025,\,99} = -1.9842$. Por lo tanto, rechazamos la hipótesis nula porque $T = 4.762 > 1.9842$, y por lo tanto cae en la región de rechazo:



- Sacamos la misma conclusión al usar el enfoque del p-value. El enfoque del p-value nos dice que rechacemos la hipótesis nula al nivel $\alpha = 0.05$ si el valor p-value $< \alpha = 0.05$.
- En este caso, el p-value es $2 \times P(T_{99} > 4.762) < 2 \times P(T_{99} > 1.9842) = 2(0.025) = 0.05$. Rechazamos H0 porque p-value < 0.05.



- Se tira un dado con 4 caras (tetraedro) 1000 veces y se registra que salen 290 veces el número 4.
- ¿Hay evidencia para concluir que el dado está sesgado, es decir, que se observan más veces el número 4 que lo esperado?

- Si el dado está equilibrado, cada cara tiene probabilidad 0.25 de aparecer. Supongamos inicialmente que este es el caso:

$$H_0: p = 0.25$$

- En base a la muestra la proporción observada resulta:

$$\hat{p} = \frac{y}{n} = \frac{290}{1000} = 0.29$$

- Según el Teorema Central del Límite, la proporción muestral

$$\hat{p} = \frac{y}{n}$$

tiene distribución asintóticamente normal con media (bajo H_0)

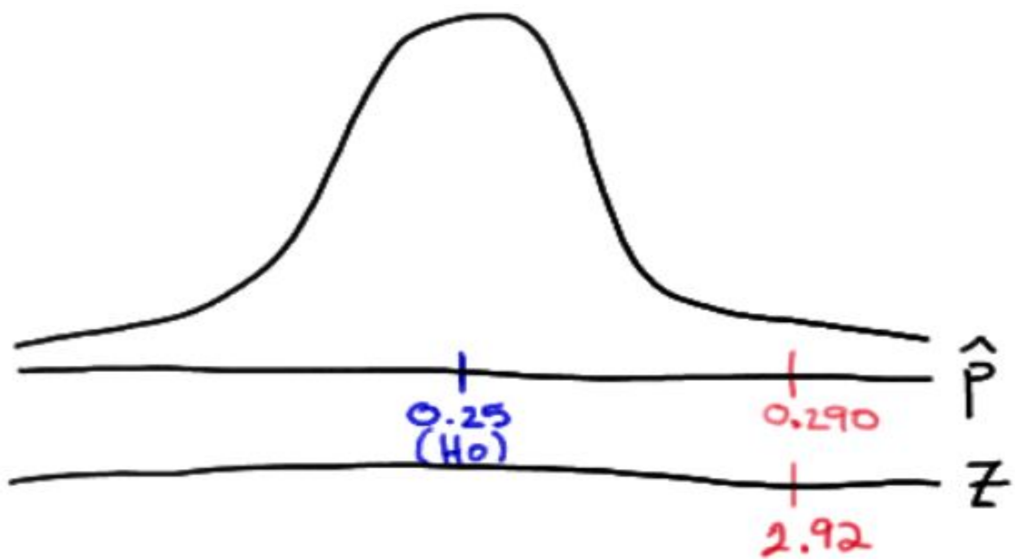
$$p_0 = 0.25$$

y desvío estándar (bajo H_0)

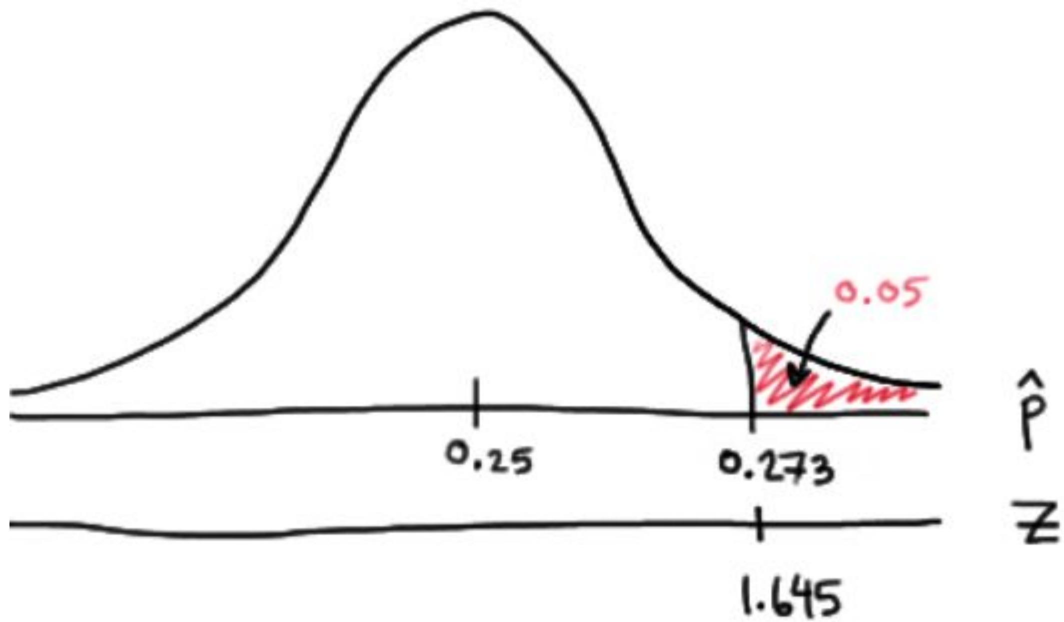
$$\sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.25(0.75)}{1000}} = 0.01369$$

- Por lo que, bajo H_0, manteniendo que H_0 es verdadera, este estadístico tiene distribución normal estándar (con media 0 y desvío 1):

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1 - p_0)}{n}}}$$



- Buscamos testear H_0 que la probabilidad, en la población, de obtener un 4 es igual igual a 0.25 versus H_1, que esta probabilidad es mayor que 0.25.
- Si fijamos el nivel de significación en 0.05, entonces rechazamos si Z > 1.645.



¿Qué pasa si no podemos aplicar el TCL para realizar un test de hipótesis para una proporción?

- Imaginemos que sospechamos que una moneda está cargada. Queremos testearlo. Para eso, decidimos hacer 10 tiradas.
- Una estrategia es contrastar hipótesis.
- Puedo construir una **distribución muestral** de cada una de las tiradas de la moneda suponiendo que la moneda está equilibrada (esa sería la H_o).
- De hecho, podemos hacerlo utilizando la distribución binomial...

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

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

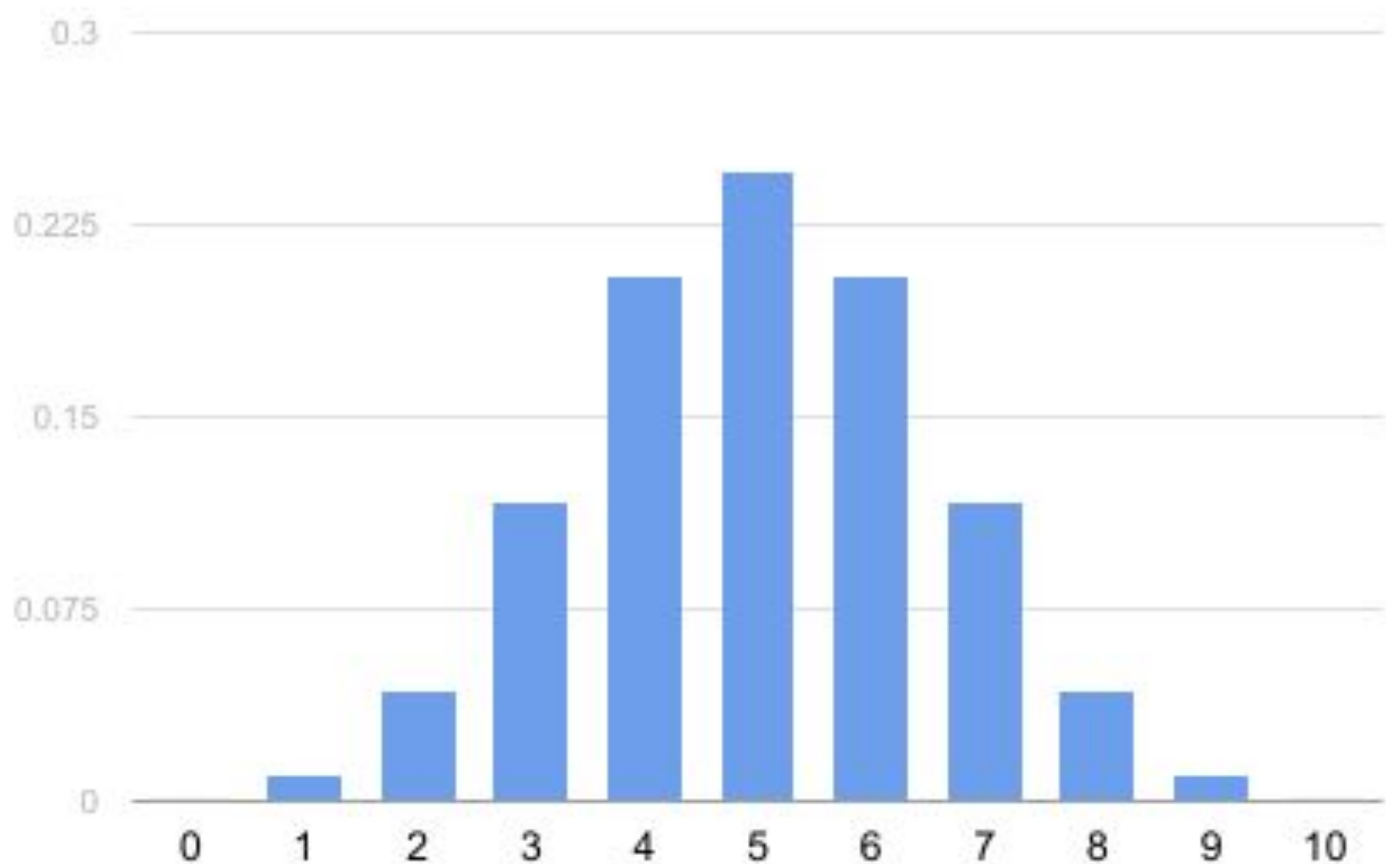
Para 0 caras = 1 x 0.5 **0 x 0.5**10
Para 1 cara = 10!/1! x 0.5 **1 x 0.5**9
etc..

Nota: 0.5 ** 10 = 1/1024

- 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

EJEMPLO: TIRADA DE MONEDAS



- 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 alpha de 0.1
  - Tiramos la moneda 10 veces y obtenemos 7 caras…
  - ¿Qué dirían sobre la H_o?
  - ¿Y si hubiéramos obtenido 1 cara en 10 tiradas?

- H_o: $\boldsymbol{p} = 0.5$
  H_a: $\boldsymbol{p} \neq 0.5$
  $\boldsymbol{\alpha} = 0.1$
- Como n es chico, en vez de trabajar con la estimación de *p*, trabajamos con X.
- Rechazamos H_o si $X < X_{C1}$ o $X > X_{C2}$, donde X_{C1} es el mayor entero que acumula una probabilidad menor o igual a $\boldsymbol{\alpha/2}$, y X_{C2} es el menor entero que acumula una probabilidad mayor o igual a $\boldsymbol{1 - \alpha/2}$.
- En este caso, 7 no es menor a 1 ni mayor que 8, por lo que no podemos rechazar H_o.

#	Prob.	Prob. acum.
0	0,001	0,001
1	0,010	0,011
2	0,044	0,055
3	0,117	0,172
4	0,205	0,377
5	0,246	0,623
6	0,205	0,828
7	0,117	0,945
8	0,044	0,989
9	0,010	0,999
10	0,001	1,000

TIPOS DE ERRORES EN UN TEST DE HIPÓTESIS

Al realizar un **test de hipótesis**, podemos incurrir en dos tipos de errores:

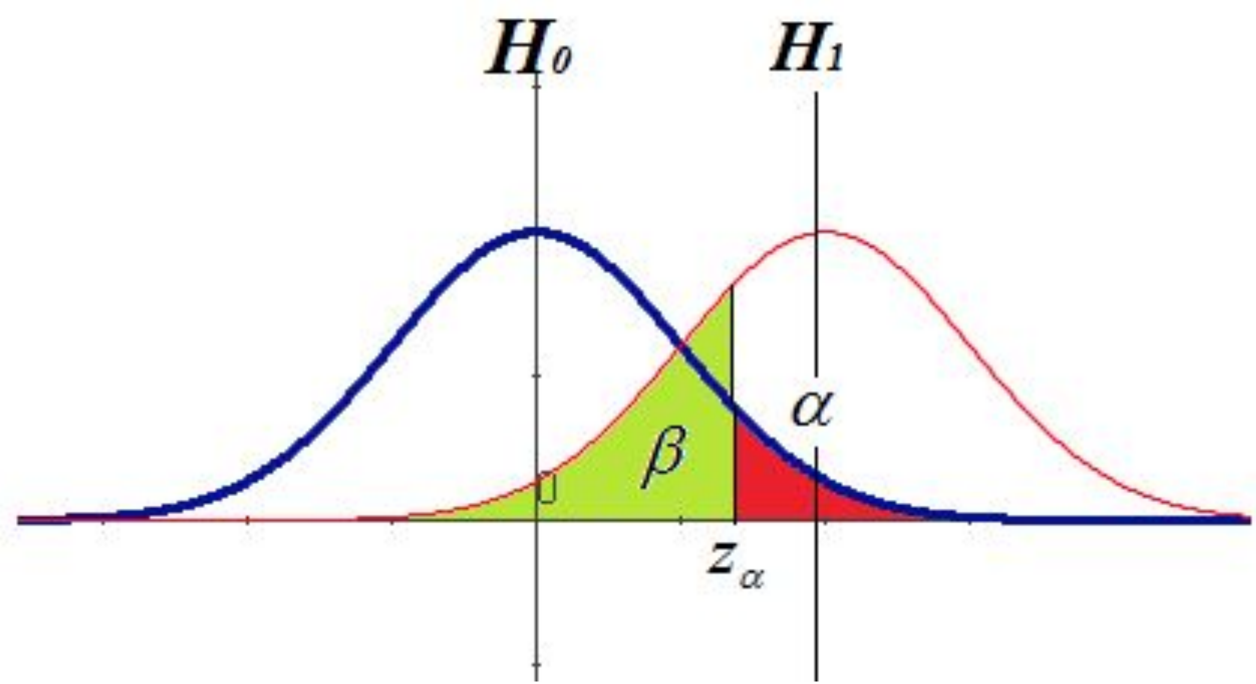
1. **Error de Tipo I**
2. **Error de Tipo II**

Decisión Adoptada	Hipótesis H_0	
	Cierta	Falsa
No rechazar H_0	***Decisión Acertada*** **$(1 - \alpha)$**	***Error de Tipo II*** **(β)**
Rechazar H_0	***Error de Tipo I*** **(α)**	***Decisión Acertada*** **$(1 - \beta)$**

Decisión Adoptada	Hipótesis H_0	
	Cierta	Falsa
No rechazar H_0	***Decisión Acertada*** **$(1 - \alpha)$**	***Error de Tipo II*** **(β)**
Rechazar H_0	***Error de Tipo I*** **(α)**	***Decisión Acertada*** **$(1 - \beta)$**

- 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?

- Alpha y Beta varían en forma inversa… ¿por qué?



TESTS PARA DOS POBLACIONES

Ejemplo: estudio sobre el efecto de jugar videojuegos durante el almuerzo y su efecto sobre el consumo de snacks durante la tarde.

- Muestra: 44 pacientes, 22 mujeres y 22 hombres.
- Se estratifica por género y se los divide aleatoriamente en dos grupos de 22 personas.
- Al grupo 1 se le pide que juegue a un videojuego mientras almuerzan, intentando ganar la mayor cantidad de puntos posibles.
- Al grupo de control se le pide que almuerce sin distracciones.
- A los dos grupos se les provee el mismo almuerzo y los mismos snacks durante la tarde.
- El objetivo es determinar si el grupo que jugó a los videojuegos consume significativamente más snacks que el grupo de control.

Ejemplo: estudio sobre el efecto de jugar videojuegos durante el almuerzo y su efecto sobre el consumo de snacks durante la tarde.

Datos:

Consumo de snack	Media Muetral	Desv. Est	n
Videojuegos	52.1 gr	45.1 gr	22
No distracciones	27.1 gr	26.4 gr	22

Ejemplo: estudio sobre el efecto de jugar videojuegos durante el almuerzo y su efecto sobre el consumo de snacks durante la tarde.

Intervalo de confianza para la diferencia de medias muestrales:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} SE_{(\bar{x}_1 - \bar{x}_2)}$$

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$df = \min(n_1 - 1, n_2 - 1)$$

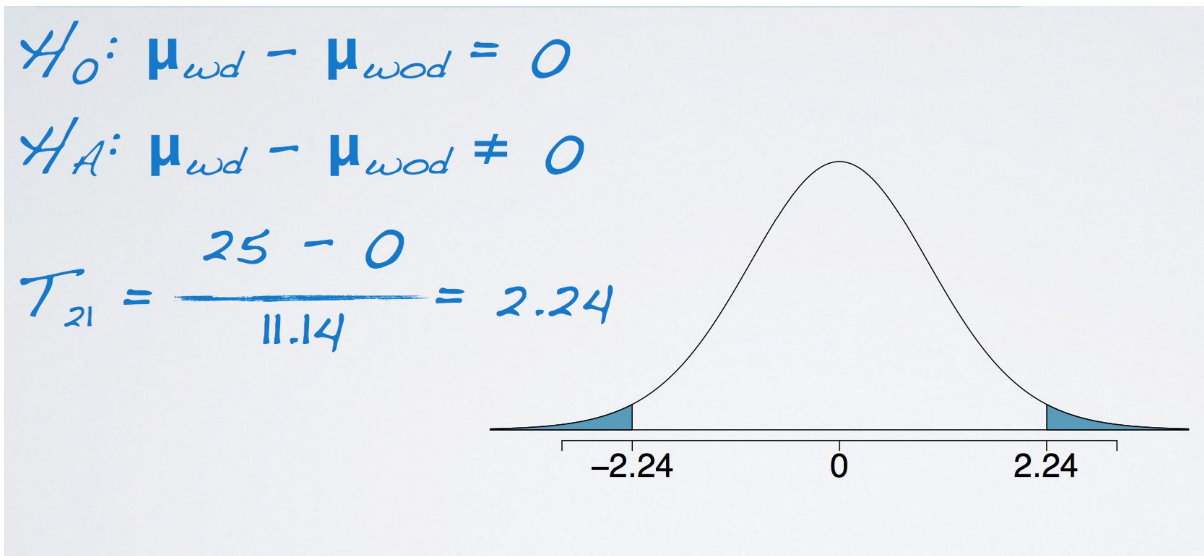
Ejemplo: estudio sobre el efecto de jugar videojuegos durante el almuerzo y su efecto sobre el consumo de snacks durante la tarde.

Intervalo de confianza con nivel de confianza de 95% para la diferencia de medias:

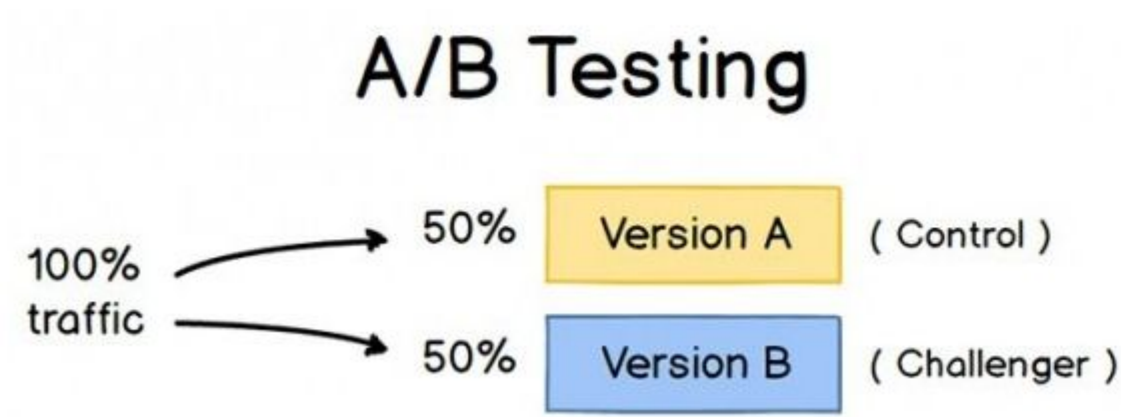
$$
\begin{aligned}
(\bar{X}_{wd} - \bar{X}_{wod}) \pm t^{*}_{df}\, SE &= (52.1 - 27.1) \pm 2.08 \times \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}} \\
&= 25 \pm 2.08 \times 11.14 \\
&= 25 \pm 23.17 \\
&= (1.83,\ 48.17)
\end{aligned}
$$

Ejemplo: estudio sobre el efecto de jugar videojuegos durante el almuerzo y su efecto sobre el consumo de snacks durante la tarde.

Test de Hipótesis:



p-value = 3.6% → Rechazamos la hipótesis nula


A/B Testing
100%
traffic
50%
Version A
(Control)
50%
Version B
(Challenger)


Visitors
20%
Original Page
20%
20%
20%
20%
Variation A
Variation B
Variation C
Variation D
ThankYou Page

Apliquemos un ejemplo...

- Una de tus principales responsabilidades es la optimización de las conversiones.
- Uno de tus anunciantes ha desarrollado un nuevo producto, y el VP de Advertisements quiere tu ayuda eligiendo entre el anuncio A y el anuncio B.
- Decidís aplicar A/B testing para definir cuál es el mejor anuncio...

- Digamos que N_A personas ven el anuncio A y que de éstas, n_A hacen click en él.
- Podemos pensar en cada vista de un anuncio como una prueba de Bernoulli donde p_A es la probabilidad que alguien haga click en el anuncio A.
- Entonces, si N_A es grande, que es el caso aquí, sabemos que n_A/N_A es aproximadamente una variable aleatoria normal con media p_A y desvío estándar $\sigma_A = \sqrt{p_A(1-p_A)/N_A}$

- Entonces, si N_B es grande, que es el caso aquí, sabemos que n_B/N_B es aproximadamente una variable aleatoria normal con media p_B y desvío estándar $\sigma_B = \sqrt{p_B(1-p_B)/N_B}$

- Si asumimos que estas dos normales son independientes (lo que parece razonable porque las pruebas Bernoulli individuales deben serlo), entonces su diferencia debe ser también normal con media $p_B - p_A$ y desvío estándar $\sqrt{\sigma_A^2 + \sigma_B^2}$

- Esto significa que podemos testear la hipótesis nula de que $p_B - p_A$ es cero (p_B y p_A son iguales) usando el estadístico:

```
def estimated_parameters(N, n):
    p = n / N
    sigma = math.sqrt(p * (1 - p) / N)
    return p, sigma
```

```
def a_b_test_statistic(N_A, n_A, N_B, n_B):
    p_A, sigma_A = estimated_parameters(N_A, n_A)
    p_B, sigma_B = estimated_parameters(N_B, n_B)
    return (p_B - p_A) / math.sqrt(sigma_A ** 2 + sigma_B ** 2)
```

que debería ser aproximadamente una normal estándar.

- Por ejemplo si A obtiene 200 clicks de 1000 vistas y B obtiene 180 clicks de 1000 vistas, el estadístico es igual a

```
z = a_b_test_statistic(1000, 200, 1000, 180)    # -1.14
```

La probabilidad de ver una diferencia tan grande si las medias fueran realmente iguales sería

```
two_sided_p_value(z)                            # 0.254
```

que es lo suficientemente grande para que no podamos concluir que existe mucha diferencia.

- Por otro lado, si B sólo obtuviera 150 clicks tendríamos

```
z = a_b_test_statistic(1000, 200, 1000, 150)    # -2.94
two_sided_p_value(z)                            # 0.003
```

lo que significa que sólo existe una probabilidad de 0.003 de ver una diferencia tan grande si los anuncios fueran igualmente efectivos.

- ¿Quién hace esto?

DigitalHouse> Coding School

- Análisis

Combinations (24) | **Page Sections (2)**

Download: XML CSV TSV | Print

Relevance Rating	Variation	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Button 5 / 5	Original	**7.51%** ± 0.2%	—	—	5851 / 77858
	Learn More	**8.91%** ± 0.2%	100%	18.6%	6927 / 77729
	Join Us Now	**7.62%** ± 0.2%	73.5%	1.37%	5915 / 77644
	Sign Up Now	**7.34%** ± 0.2%	13.7%	-2.38%	5660 / 77151
Media 5 / 5	Original	**8.54%** ± 0.2%	—	—	4425 / 51794
	Family Image	**9.66%** ± 0.2%	100%	13.1%	4996 / 51696
	Change Image	**8.87%** ± 0.2%	92.2%	3.85%	4595 / 51790
	Barack's Video	**7.76%** ± 0.2%	0.04%	-9.14%	3992 / 51427
	Sam's Video	**6.29%** ± 0.2%	0.00%	-26.4%	3261 / 51864
	Springfield Video	**5.95%** ± 0.2%	0.00%	-30.3%	3084 / 51811

- And the winner is...

https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/