Módulo 2: Métodos Estadísticos para IA

Test de hipótesis

Ramiro M. Irastorza, Marcelo Cappelletti

Diplomatura Superior en Inteligencia Artificial

UNAJ

Objetivos de esta presentación

- Introducir conceptos de inferencia estadística.
- Describir aspectos básicos de test de hipótesis.

Resumen

Introducción

Ejemplo: máquina de chicles

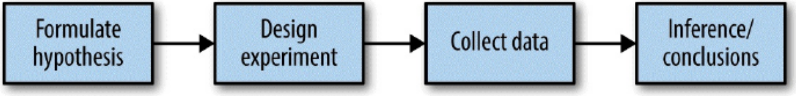
Formalizar con un modelo

Otro ejemplo: test sobre media con varianza conocida

Comentarios y definiciones

Introducción: inferencia estadística[1]

- Algo vimos en la clase pasada



[1] P. Bruce y A. Bruce, Practical Statistics for Data Scientists, 2017

Importancia del Diseño de Experimentos

El **diseño de experimentos** es un elemento clave en la práctica de la estadística, con aplicaciones en prácticamente todos los ámbitos.

El objetivo principal es diseñar un experimento para **confirmar o rechazar una hipótesis**.

En ciencia de datos existe la necesidad de realizar experimentos de forma continua, especialmente en aspectos como:

- Interfaz de usuario
- Estrategias de marketing de productos

El Proceso Clásico de Inferencia Estadística

Al hablar de **significancia estadística**, *t-tests* o *p-valores*, se hace referencia al flujo clásico de la inferencia estadística:

1. Se formula una **hipótesis**
   - "El fármaco A es mejor que el fármaco estándar"
   - "El precio A es más rentable que el precio B"
2. Se diseña un **experimento** (ej.: prueba A/B) para ponerla a prueba
3. Se **recogen y analizan** los datos
4. Se extraen **conclusiones** aplicables a una población más amplia

El término **inferencia** refleja la intención de extender los resultados del experimento a un conjunto mayor de procesos o individuos.

Pruebas A/B

Una prueba A/B se construye típicamente a partir de una **hipótesis**.

Ejemplo de hipótesis:

- “El precio B produce una mayor ganancia que el precio A”.

Preguntas clave:

- ¿Por qué necesitamos una hipótesis?
- ¿Por qué no simplemente observar el resultado y elegir la opción más favorable?

¿Por qué necesitamos una hipótesis?

La respuesta está en la **tendencia humana a subestimar el alcance del azar**:

- Dificultad para anticipar **eventos extremos**.
- Tendencia a interpretar **patrones ilusorios** en sucesos aleatorios.

La prueba de hipótesis estadística se inventó para proteger al analista de ser engañado por el azar y tomar conclusiones incorrectas.

Resumen

Introducción

Ejemplo: máquina de chicles

Formalizar con un modelo

Otro ejemplo: test sobre media con varianza conocida

Comentarios y definiciones

Introducción: El Misterio de la Máquina de Chicles [1][2]

El detective estadístico

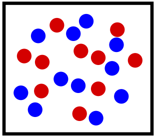
- Un detective formula primero una hipótesis.
- Busca pruebas que respalden o contradigan su teoría.
- Basándose en la evidencia, emite un veredicto.

La prueba de hipótesis en estadística funciona igual: una lógica para **decidir** *bajo incertidumbre.*

[2] R. Park, The American Statistician (2018)

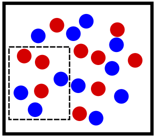
Presentación del caso

- Máquina de chicles con rojos y azules.
- Pregunta central:
  ¿Cuál es la proporción real de chicles rojos?
- La máquina completa = **población (N)**.
- No podemos contar todo $\Rightarrow$ usamos muestras.



Presentación del caso

- Máquina de chicles con rojos y azules.
- Pregunta central:
  ¿Cuál es la proporción real de chicles rojos?
- La máquina completa = población (N).
- No podemos contar todo $\Rightarrow$ usamos **muestras (n)**.



El plan de investigación

1. Formular una sospecha (Hipótesis).
2. Recoger y analizar pistas (Muestreo y eventos raros).
3. Emitir un veredicto (Decisión estadística).

Paso 1: formular una sospecha

¿Qué es una Hipótesis?

- Es una suposición inicial.
- Afirmación sobre la población.
- Punto de partida de la investigación.

Nuestra primera sospecha: Hipótesis Nula

- Supongamos: proporción 50:50 entre rojos y azules.
- En estadística: **hipótesis nula**.
- Es el *status quo*: mantenido salvo evidencia en contra.

Paso 2: Recoger y analizar pistas

Muestreo

- No contamos todos los chicles: tomamos muestras.
- Ejemplo: 10 chicles $\Rightarrow$ pequeña representación.
- Una muestra puede variar mucho.

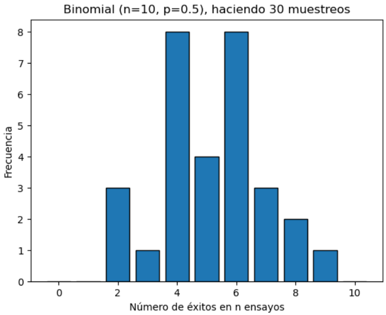
En este ejemplo simulamos la extracción de 10 chicles, varias veces (1000) y anotamos cuántos rojos sacamos en cada una.

```
n = 10        # número de ensayos
p = 0.5       # probabilidad de éxito en cada ensayo
muestras = 30  # cantidad de muestras
datos = np.random.binomial(n, p, muestras)# Generar las muestras
```

codigo/ejemplo9.py

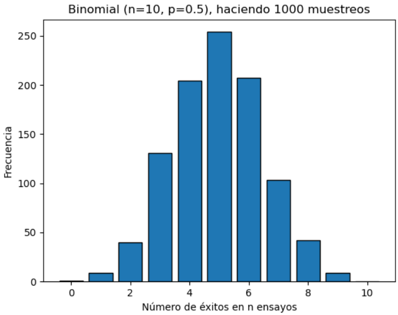
Distribución de Muestreo

- Repetimos muestreos muchas veces.
- Aprendemos dos cosas:
  1. Los resultados varían (**variabilidad natural**).
  2. Los resultados se agrupan alrededor del valor esperado.
- Patrón general: **distribución de muestreo**.



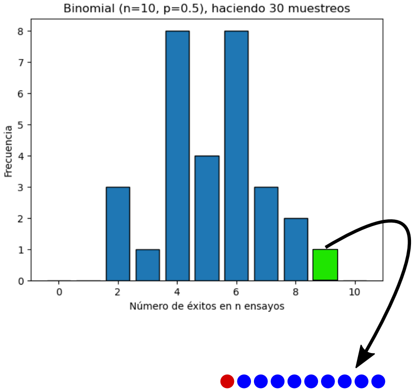
Distribución de Muestreo

- Repetimos muestreos muchas veces.
- Aprendemos dos cosas:
  1. Los resultados varían (variabilidad natural).
  2. **Los resultados se agrupan alrededor del valor esperado**.
- Patrón general: **distribución de muestreo**.



Eventos raros y zona de sospecha

- Eventos muy poco probables si la hipótesis es cierta.
- Ejemplo: menos de 5 % de probabilidad.
- Zona de sospecha = región de resultados inusuales.
- Si caemos allí, dudamos de la hipótesis inicial.



Paso 3: Emitir un veredicto

1. Reafirma la hipótesis (50:50).
2. Define la zona de sospecha.
3. Obtené la muestra real.
4. Compará con la zona.
5. Decidí: aceptar o rechazar hipótesis.

Ejemplos de veredicto

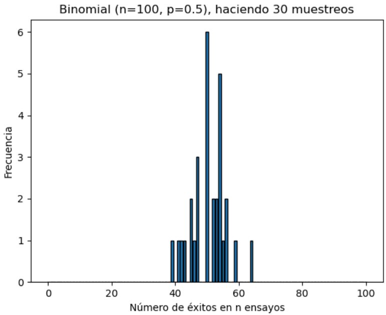
- Resultado: 9 rojos de 10.
- Evento raro bajo 50:50.
- Conclusión: **rechazar hipótesis nula**.

  Nuestro resultado (9 rojos) cae directamente dentro de nuestra "zona de sospecha". Es tan improbable que obtengamos 9 rojos por pura casualidad si nuestra hipótesis de 50:50 fuera cierta, que tenemos una fuerte evidencia en su contra. Concluimos que la proporción real de chicles rojos no es 50:50.

- *En cambio, con 6 rojos: no hay evidencia suficiente para rechazar.*

Tamaño de la muestra

- Más grande = más precisa la estimación.
- Distribución de muestreo más estrecha.
- Mayor poder para detectar hipótesis falsas.



Conclusión: con una distribución más precisa, es mucho más fácil detectar si un resultado es verdaderamente raro". Una pista más grande nos da más poder para identificar correctamente una hipótesis falsa y llegar a un veredicto más seguro.

Conclusión: del misterio al mundo real

- Significancia estadística = evento raro para la hipótesis.
- Prueba de hipótesis = herramienta de razonamiento crítico.

Resumen

Introducción

Ejemplo: máquina de chicles

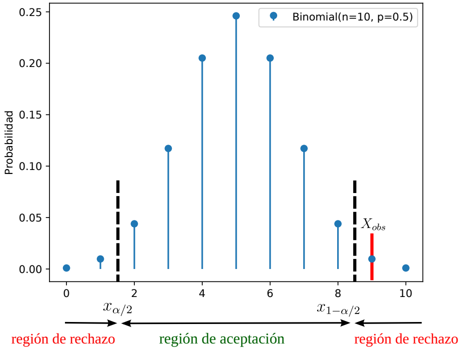
Formalizar con un modelo

Otro ejemplo: test sobre media con varianza conocida

Comentarios y definiciones

Zonas de aceptación y rechazo

La decisión se basa en la "ubicación" de la observación de la muestra (X_{obs}). En el ejemplo se la figura se **rechaza la hipótesis nula** (salió 9 veces un chicle rojo).



Las preguntas ahora son: ¿Cómo definimos la región de aceptación (es 6 adecuado)? ¿y la de rechazo es 9 demasiado?

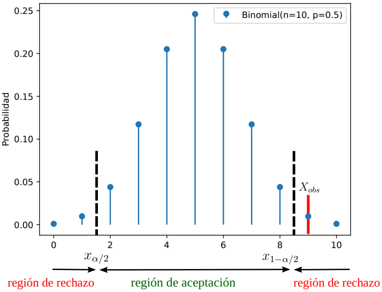
Muestra aleatoria y Estadístico: modelo matemático

- Si quisiera cuantificar probabilísticamente lo graficado anteriormente ¿cómo lo haría?
- Si consideramos la v.a. definida como la **cantidad de éxitos** X (ahora con mayúscula porque es una v.a.) cuando extraemos una muestra de tamaño n sobre una población y suponiendo que nuestro éxito tiene una proporción poblacional p_0 entonces X tiene distribución Binomial con parámetros n y p_0. Lo denotamos así $X \sim \mathbf{B}\,(n, p_0)$.
- Podemos pensar a X como una suma de n variables aleatorias Y_i (debemos asegurarnos de que sean estadísticamente iid) donde cada $Y_i \sim \mathbf{B}\,(1, p_0)$ (experimento de Bernoulli). Y_i cuenta "1" para un éxito con probabilidad p_0 y "0" para fracaso con probabilidad $1 - p_0$. Entonces:

$$X = \sum_{i=1}^{n} Y_i \tag{1}$$

Error de tipo I y significancia

- Ahora podemos definir la región de rechazo con alguna **significancia** (α) con la Ec. (2).
- Se define el error **error de tipo I** como: la probabilidad de equivocarme al rechazar la hipótesis nula cuando ésta es verdadera.

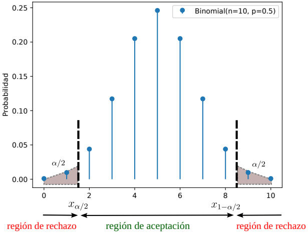


$$P\{X_{obs} < x_{\alpha/2} \text{ o } X_{obs} > x_{1-\alpha/2} |\ H_0 \text{ verdadera}\} = \alpha \qquad (2)$$

¿Qué significa H_0 verdadera? (condición: que H_0 sea verdadera).

Respuesta: que $p = 0{,}5$.

Error de tipo I y significancia



Si elegimos una significancia de $\alpha =$0,05, la decisión fuerte será:

$$\boxed{\text{Rechazar } H_0 \text{ si } x_{obs} > 8 \text{ o si } x_{obs} < 2}$$

```python
from scipy.stats import binom
n = 10        # número de ensayos
p = 0.5       # probabilidad de éxito
alfa = 0.05     # cuantil deseado
x_alfa_2 = binom.ppf(alfa/2, n, p)
x_1_alfa_2 = binom.ppf(1-alfa/2, n, p)
```

codigo/ejemplo9.py

La salida es:

$>$ Valor crítico menor[3] ($x_{\alpha/2}$): 2

$>$ Valor crítico mayor ($x_{1-\alpha/2}$): 8

[3]Nota: ¡Observar que es menor estricto! (y mayor estricto).

Tomamos la decisión: Ejemplo 1

Veamos cuál es el valor de la observación de nuestra muestra:

- Ejemplo 1: Si tomamos una muestra y sacamos 9 chicles rojos entonces...

Se dice que es una **decisión fuerte** porque se le da muchas chances de salir a la observación en la región de aceptación ($1-\alpha$, es decir, 95 % de probabilidad) y el que diseña el test maneja esa probabilidad. Dicho de otro modo,

- El analista selecciona la significancia del test α de tal forma que sea poco probable equivocarse cuando H_0 es verdadera.
- Pero... ¿Por qué puedo cometer un error?

Tomamos la decisión: Ejemplo 2

Veamos cuál es el valor de la observación de nuestra muestra:

- Ejemplo 2: Si tomamos una muestra y sacamos 6 chicles rojos entonces...

Se dice que es una **decisión débil**, porque no hay evidencia para rechazar.

Es crucial entender lo que esto significa. No estamos ”probando”que la hipótesis de 50:50 sea cierta. Simplemente estamos diciendo que la evidencia que hemos recogido no es lo suficientemente fuerte como para descartarla. El caso sigue abierto, por así decirlo.

Error de tipo II

- Se puede cometer otro error y calcular su probabilidad: equivocarme al aceptar la hipótesis nula cuando ésta **NO** es verdadera.
- La probabilidad del denominado **error de tipo II** (β), queda determinada por:

$$\beta = P\{\underbrace{x_{\alpha/2} \leq X_{obs} \leq x_{1-\alpha/2}}_{\text{región de aceptación}} \mid H_0 \text{ falsa}\} \tag{3}$$

Que es lo mismo que:

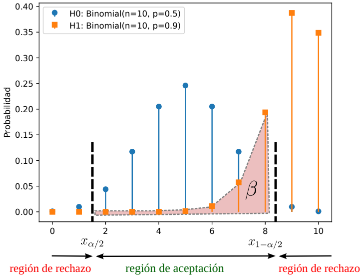
$$\beta = P\{x_{\alpha/2} \leq X_{obs} \leq x_{1-\alpha/2} \mid H_1 \text{ verdadera}\}. \tag{4}$$

- ¿Pero que p usamos? Respuesta: ¡debemos proponerla!
- Importante: Notar que ahora los símbolos son "$\leq$".

Cálculo de β

Entonces podemos ahora calcular el error de tipo II:

- Proponemos $H_1 : p = 0{,}9$
- La zona coloreada es la probabilidad β del error de tipo II.

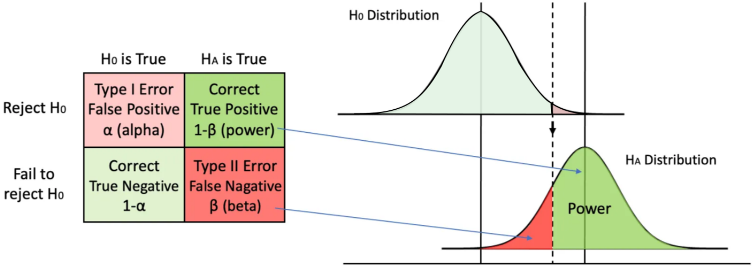



```python
p1 = 0.9
beta = binom.cdf(8, n, p1)-binom.cdf(1, n, p1)# P(2<= x <= 8)
```

codigo/ejemplo9.py

El resultado es $\beta = 0{,}2639$.

Potencia del test (π)



La **potencia** de un test es la probabilidad de que el test rechace correctamente la hipótesis nula.

Definición del p-valor

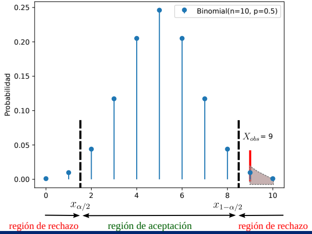
Veamos la definición a partir de un ejemplo:

- Ejemplo 1: tomamos una muestra y obtenemos 9 chicles rojos y uno azul.
- Supongamos que H_0 es verdadera, calculemos la probabilidad dibujada:

```
p = 0.5
p_valor_2 = 1-binom.cdf(8, n, p)#P(X>=9)
```

`codigo/ejemplo9.py`

- El resultado de esa probrabilidad es: 0,0107.
- Notar que si lo comparamos con $\alpha/2$ es un valor inferior.
- Notar que también que cae en zona de rechazo.



Definición del p-valor

- La idea anterior se puede generalizar: podemos decidir con el p-valor, no será necesario calcular los valores críticos:
  Si p-valor $< \alpha$ entonces rechazaremos H_0.
- Cuantifica la cercanía del valor observado a la región de aceptación.
- Mientras más pequeño sea más lejos de la región de aceptación estará.
- Se define entonces **p-valor** como el nivel de significancia más pequeño que conduce al rechazo de la hipótesis nula H_0.

Definición del p-valor

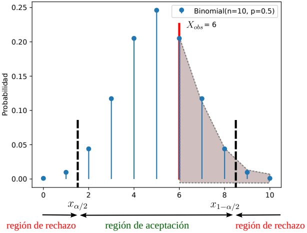
Tomemos otro ejemplo:

- Ejemplo 2: tomamos una muestra y obtenemos 6 chicles rojos y 4 azules.

```
p_valor_2 = 1-binom.cdf(5, n, p)#P(X>=6)
```

codigo/ejemplo9.py

- El resultado de esa probrabilidad es: 0.3769.
- Notar que si lo comparamos con $\alpha/2$ es un valor superior.
- Notar que también que cae en zona de aceptación.
- Si p-valor $> \alpha$ entonces no hay evidencia para rechazar H_0.



Resumen

Introducción

Ejemplo: máquina de chicles

Formalizar con un modelo

Otro ejemplo: test sobre media con varianza conocida

Comentarios y definiciones

Otro ejemplo: Ahora con una v.a. continua

Supongamos que estamos en casa y deseamos poder decidir con algún criterio que la tensión (voltaje) promedio que recibimos de la empresa que suministra electricidad es o no 220V. Disponemos de un voltímetro, por lo que podemos hacer $n = 25$ mediciones en el transcurso del día y el valor medio ($\bar{x}$) nos arroja 230,3 V[4].
El interés será entonces contrastar:

$$\begin{aligned} H_0 &: \mu = \mu_0 = 220V \text{ Hipótesis nula} \\ H_1 &: \mu \neq \mu_0 = 220V \text{ Hipótesis alternativa} \end{aligned} \tag{5}$$

Como se ve de la ecuación anterior, se estableció lo que se denomina **hipótesis nula** y la **alternativa**.

[4] ¡Notar que ahora la variable aleatorio X es continua!

Otro ejemplo: Ahora con una v.a. continua

- ¿Qué cambió respecto del ejemplo de la máquina de chicles anterior?
- ¿Tengo un modelo matemático que represente esto? (el de la máquina de chicles era un proceso Binomial).

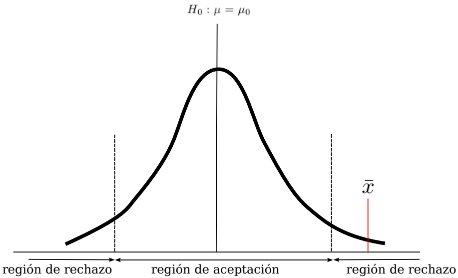
Definición de las hipótesis

- La **hipótesis nula** puede ser resultado de la experiencia en el proceso, de alguna teoría o provenir de consideraciones externas (tales como especificaciones de diseño, ingeniería u obligaciones contractuales).
- Una obligación contractual sería el caso de nuestro ejemplo: la empresa nos debería ofrecer 220V.

- El propósito del test de hipótesis es establecer si la evidencia experimental soporta el rechazo de la hipótesis nula.

Zonas de aceptación y rechazo

La decisión se basa en la “ubicación” de la observación de la muestra ($\bar{x}$). Por ejemplo: en la figura se muestra una observación que cae fuera de la región de aceptación. Por lo tanto se **rechaza la hipótesis nula**.



Las preguntas ahora son: ¿Cómo definimos la región de aceptación? ¿y la de rechazo? ¿Está 230,3V lo suficientemente lejos de 220V como para rechazar esa hipótesis?

Muestra aleatoria y Estadístico

Para poder continuar, debemos suponer algunas cosas:

- la muestra que tomamos es en realidad una **muestra aleatoria** $X_1, X_2, \ldots, X_n$. Si tomamos $\bar{X}$, supondremos que tiene una **distribución aproximadamente Normal**[5].
- **Conocemos la varianza** de la variable aleatoria (σ^2) o si n es grande la podríamos reemplazar por la observación muestral (s^2).

Estas suposiciones nos permiten establecer un **estadístico** de prueba de la forma:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{6}$$

La elección de la Ec.(6) se realizó porque si la **hipótesis nula es verdadera**, este estadístico tiene distribución Normal estandar $N(0, 1)$.

[5]Se cumple el **Teorema Central del Límite** (tomar $n >$20 como suficiente).

Error de tipo I y significancia

Por consiguiente ahora podemos definir la región de rechazo con alguna **significancia** (α):

$$P\{|Z_0| > z_{\alpha/2} \mid H_0 \text{ verdadera}\} = \alpha \tag{7}$$

esto es por definición el **error de tipo I**[6]. Cuyo significado es:

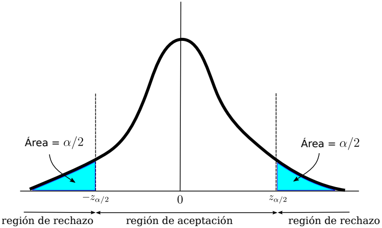
- la probabilidad de equivocarme al rechazar la hipótesis nula cuando ésta es verdadera. Esto quiere decir que si el valor observado de mi estadístico de prueba (z_0) cae fuera del intervalo $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ tendré evidencia como para rechazar la hipótesis nula.

[6]La Ec. (7) es una probabilidad condicionada a que la hipótesis nula sea verdadera.

Error de tipo I y significancia: uso de scipy.stats

Si elegimos una significancia de $\alpha =$0,05, la decisión fuerte será:

$$\boxed{\text{Rechazar } H_0 \text{ si } z_0 > 1,96 \text{ o si } z_0 < -1,96}$$



Para calcular esos valores podemos utilizar:

```
alfa = 0.05
zalfa_2_neg = stats.norm.ppf(alfa / 2)  #cuantil x_{0,025}
zalfa_2_pos = stats.norm.ppf(1 - alfa / 2)  #cuantil x_{0,975}
```

codigo/ejemplo10.py

La salida es:

$>$ Valor crítico negativo: -1.959964

$>$ Valor crítico positivo: 1.959964[7]

[7] Dan el mismo módulo porque la *fdp* es simétrica.

Tomamos la decisión

Veamos cuál es el valor de la observación de nuestra muestra. Para esto debemos conocer σ (raíz cuadrada de la **varianza**), supongamos que nos dieron ese dato y vale 20 V. Entonces:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{230,3 - 220}{20/\sqrt{25}} \approx 2,57$$

como $z_0 > 1,96$ entonces tenemos evidencia suficiente como para **rechazar** H_0 con una significancia de 0,05.

Se dice que es una **decisión fuerte** porque se le da muchas chances de salir a la observación en la región de aceptación ($1 - \alpha$, es decir, 95 % de probabilidad) y el que diseña el test maneja esa probabilidad. Dicho de otro modo,

- el analista selecciona la significancia del test α de tal forma que sea poco probable equivocarse cuando H_0 es verdadera.

Error de tipo II

Se puede considerar otro error. La probabilidad del denominado **error de tipo II** (β), queda determinada por la probabilidad de equivocarme al aceptar la hipótesis nula cuando ésta **NO** es verdadera.

Considerando el desarrollo anterior supongamos que la hipótesis nula es falsa y que el verdadero valor de la media es $\mu = \mu_0 + \delta$ con $\delta > 0$. Veamos qué distribución tiene el estadístico de prueba:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma} - \frac{\delta\sqrt{n}}{\sigma} = \frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma} \quad (8)$$

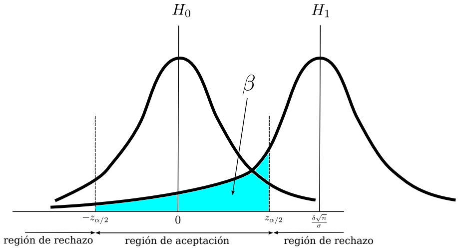
En la Ec. (8) sumamos y restamos $\frac{\delta\sqrt{n}}{\sigma}$ con lo cual no se altera el estadístico de prueba. Aunque ahora la distribución del mismo será una Normal con otra media: $Z_0 \sim N\left(\frac{\delta\sqrt{n}}{\sigma}, 1\right)$.

Cálculo de β

Entonces podemos ahora calcular el error de tipo II:

$$\beta = P\{|Z_0| \leq z_{\alpha/2} \mid H_0 \text{ falsa}\} = P\{|Z_0| \leq z_{\alpha/2} \mid H_1 \text{ verdadera}\}. \quad (9)$$

La zona coloreada es la probabilidad β del error de tipo II.



Esta probabilidad será entonces:

$$\beta = P\left(Z \leq z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - P\left(Z \leq -z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \qquad (10)$$

aquí estamos integrando la Normal estandarizada.

Elección de tamaño de muestra

Veamos qué se puede hacer con la Ec. (10). Si $\delta > 0$ entonces si $-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} < -3$ (ver [8]), podemos decir que su probabilidad tenderá a cero. La Ec. (10) se reducirá a:

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \tag{11}$$

lo que nos permitirá hacer un cálculo del tamaño de la muestra (n) para un valor de β particular (si δ y α son dato). Si definimos $-z_\beta = z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}$, entonces el resultado será:

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2} \tag{12}$$

[8] En la Normal standard es casi nula la probabilidad por encima de 3 (o $<$ -3).

Tamaño de muestra y potencia del test (π)

Si la media verdadera es 232,0V (la tomamos como la Hipótesis alternativa H_1), entonces $\delta = 12$ V. Si además deseamos $\beta = 0{,}1$ entonces:

$$n \approx \frac{(1,96 + 1,28)^2 20^2}{12^2} \approx 30$$

Hemos calculado $-z_\beta$ utilizando:

```
beta = 0.1
zbeta_neg = stats.norm.ppf(beta)
```

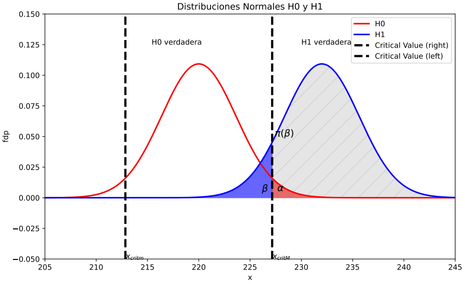
codigo/ejemplo10.py

También se puede definir la **potencia del test**: $\pi(\beta) = 1 - \beta$, que en este caso será 0,9.

Ejemplo de código

El ejemplo 10 muestra todo lo calculado hasta ahora. Si lo ejecutan obtendrán la siguiente figura:

- Debe comentarse que en la figura se muestran todas las variables sin normalizar, por ejemplo (utilizando $n = 30$):
  $x_{critM} = \mu_0 + z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}} = 227,15$ V.
- De esta forma se puede concluir directamente con la variable observada, es decir, si se midió nuevamente (ahora $n = 30$) y el promedio es $\overline{x} =$ 229,6 V entonces cae fuera de la región de aceptación por lo tanto H_0 es rechazada.



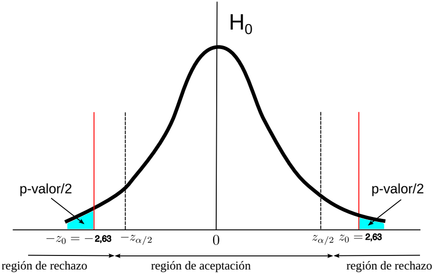
Definición del p-valor

Para la definición del **p-valor** volvamos al ejemplo. Supongamos que H_0 es verdadera, calculemos la probabilidad dibujada. El valor calculado es 0,0086 (ver ejemplo10.py).

```
x_media_observada = (229.6 - mu0) / sigma   # Estandarizo
p_valor = 2 * (1 - stats.norm.cdf(x_media_observada, 0, 1))
```

codigo/ejemplo10.py

- Cuantifica la cercanía del valor observado a la región de aceptación.
- Mientras más pequeño sea más lejos de la región de aceptación estará.



Definición del p-valor

- Finalmente, como el p-valor = $0,0086$ entonces se puede decidir (si $\alpha = 0,05$) que se rechaza la Hipótesis nula. ¡Note que también se rechaza si $\alpha = 0,01$!
- Se rechazará H_0 siempre que se cumpla que p-valor $< \alpha$.
- Por esta razón, en general, es el único valor que se informa en los estudios. Por ejemplo, la referencia de Stone y otros [2]:

RESULTS

Of the 614 patients who were enrolled in the trial, 302 were assigned to the device group and 312 to the control group. The annualized rate of all hospitalizations for heart failure within 24 months was 35.8% per patient-year in the device group as compared with 67.9% per patient-year in the control group (hazard ratio, 0.53; 95% confidence interval [CI], 0.40 to 0.70; P<0.001). The rate of freedom from device-related complications at 12 months was 96.6% (lower 95% confidence limit, 94.8%; P<0.001 for comparison with the performance goal). Death from any cause within 24 months occurred in 29.1% of the patients in the device group as compared with 46.1% in the control group (hazard ratio, 0.62; 95% CI, 0.46 to 0.82; P<0.001).

Resumen

Introducción

Ejemplo: máquina de chicles

Formalizar con un modelo

Otro ejemplo: test sobre media con varianza conocida

Comentarios y definiciones

Procedimiento general

Una vez que ya se diseñó el experimento y se conoce la hipótesis nula se puede establecer un procedimiento para el diseño de un test de hipótesis:

1. Especificar el estadístico de prueba (en el ejemplo Ec. (6))
2. Definir si es un test unilateral (una cola) o bilateral (dos colas, este es el caso del ejemplo)
3. Definir cuál será la significancia

1-Estadístico de prueba

- Un Estadístico es una **función de variables aleatorias**. Cuando se busca un Estadístico para test de hipótesis, necesitamos conocer su función de distribución de probabilidades, dado que debemos computar α, β, p-valor, etc.
- En el ejemplo supusimos que la muestra aleatoria proviene de una distribución Normal. Sabemos que la suma de Normales es Normal (propiedad reproductiva). Si σ es conocido entonces el estadístico de prueba (Ec. 6) tiene distribución conocida.
- Aunque no provenga de una distribución Normal, si n es grande, se podría decir que es aproximadamente Normal por el Teorema Central del Límite.
- Pero, si n es pequeño y la m.a. es Normal entonces se puede estimar σ por s, aunque tendremos un estadístico de prueba que no tiene distribución Normal sino *t-student* con $n-1$ grados de libertad.

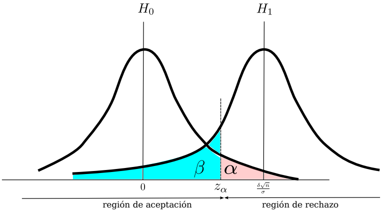
2-Test unilateral o bilaterales

En el test que vimos se contrastó la hipótesis nula con una hipótesis alternativa de "dos colas", es decir, la media verdadera puede ser mayor o menor que μ_0. Los mismos razonamientos pueden ser aplicados a tests unilaterales. Retomando el ejemplo, podemos plantear:

$$H_0 : \mu = \mu_0 = 220\ V \text{ Hipótesis nula}$$

$$H_1 : \mu > \mu_0 = 220\ V \text{ Hipótesis alternativa}$$

Las probabilidades α y β deben ser recalculadas[9].



[9]Similar para el caso en que H_1 sea $\mu < \mu_0$.

3-Seleccionar la significancia práctica y p-valor

- Con la significancia lo que buscamos es ser conservativos respecto de la Hipótesis nula. Esto no siempre es así, en nuestro ejemplo, si el valor observado fuera: $\overline{x} =$ 229,6 V, podemos pensar que no es prácticamente un valor lejano de H_0, sin embargo la hemos rechazado porque cae fuera de la región de aceptación.
- Imaginemos que la H_1 es $\mu = 229,6$ V, que el valor observado es $\overline{x} =$ 229,6 V (fijo) y, además que podemos variar n. Calculamos p-valor y potencia del test ($\pi = 1 - \beta$):

$$n = 10 \longrightarrow \text{p-valor} = 0,13 \longrightarrow \text{potencia} = 0,33$$

$$n = 20 \longrightarrow \text{p-valor} = 0,032 \longrightarrow \text{potencia} = 0,57$$

$$n = 30 \longrightarrow \text{p-valor} = 0,009 \longrightarrow \text{potencia} = 0,75$$

$$n = 100 \longrightarrow \text{p-valor} = 1,6 \times 10^{-6} \longrightarrow \text{potencia} = 0,998$$

3-Seleccionar la significancia práctica y p-valor

- A medida que aumentamos el número de muestras el p-valor disminuye y, para $n > 30$ rechazamos casi siempre H_0. Ver ejemplo 10.
- Conclusión: para n grande el test es muy sensible y detectará la diferencia con H_0, aunque la separación con μ_0 sea muy pequeña.
- Note que esto se evidencia con la potencia, a medida que crece n también lo hace la potencia.

$$n = 10 \longrightarrow \text{p-valor} = 0,13 \longrightarrow \text{potencia} = 0,33$$

$$n = 20 \longrightarrow \text{p-valor} = 0,032 \longrightarrow \text{potencia} = 0,57$$

$$n = 30 \longrightarrow \text{p-valor} = 0,009 \longrightarrow \text{potencia} = 0,75$$

$$n = 100 \longrightarrow \text{p-valor} = 1,6 \times 10^{-6} \longrightarrow \text{potencia} = 0,998$$

Bibliografía

1. D. C. Montgomery and G. C. Runger, “Applied Statistics and Probability for Engineers Third Edition”, John Wiley & Sons, Inc.
2. A. Agresti and M. Kateri, “Foundations of statistics for data scientists: with R and Python”, Chapman and Hall/CRC (2021).
3. P. Bruce and A. Bruce, “Practical Statistics for Data Scientists”, O’Reilly Media (2017).

Referencias

Ryoungsun Park.
Practical teaching strategies for hypothesis testing.
The American Statistician, 2019.

Gregg W Stone, JoAnn Lindenfeld, William T Abraham, Saibal Kar, D Scott Lim, Jacob M Mishell, Brian Whisenant, Paul A Grayburn, Michael Rinaldi, Samir R Kapadia, et al.
Transcatheter mitral-valve repair in patients with heart failure.
New England Journal of Medicine, 379(24):2307–2318, 2018.