Análisis cuantitativo avanzado

Clase 6: Tests de hipótesis

Lic. Lucio José Pantazis

Motivación

Disclaimer

Los datos con los que trabajaremos en esta presentación son **simulados**.

En general, es preferible ver las herramientas estadísticas desde los datos de un problema real, pero no es fácil encontrar datos reales que permitan retratar con claridad algunos conceptos.

Por lo tanto, vale aclarar que estos datos no tienen por qué representar la realidad de forma fidedigna.

Un cuento chino

Para introducir esta temática, vamos a ver el siguiente extracto de la película “Un cuento chino”, del año 2005.

Un cuento chino

Imaginemos cómo puede haberse dado la previa a este intercambio.

- Hay una **hipótesis previa**: si bien la cantidad de tornillos por caja es a veces es mayor o menor, su **media** es de 350 tornillos. Esta hipótesis es sostenida por el proveedor.
- Como cliente, Ricardo tiene una sospecha que esta hipótesis es falsa. Por lo tanto, plantea una **hipótesis alternativa**, contrapuesta a la hipótesis planteada por el proveedor, que establece que la **media** de los tornillos por caja es **menor** que 350 tornillos.
- Para poner a prueba estas hipótesis enfrentadas, Ricardo decide contar la cantidad de tornillos en las cajas. Es decir, recolecta **evidencia** que pueda respaldar alguna de las hipótesis.
- Luego de recolectar evidencia, debe **decidir** por una de las hipótesis para saber cómo actuar. En esta decisión puede cometer **errores**:
  - Puede decidir que el proveedor se equivoca, cuando éste tiene razón. Esto lo llevaría a **cambiar** de proveedor de forma innecesaria. Por lo tanto, ante la duda, estará más reacio a cometer este error.
  - Puede decidir que el proveedor tiene razón, cuando no la tiene. Esto lo llevaría a **mantener** a un proveedor que no cumple con su palabra.
- Para decidir por cuál de las dos hipótesis se inclina, debe tener en cuenta que la cantidad de tornillos es **variable**. Por lo tanto, no cualquier valor menor que 350 es necesariamente indicio de que la hipótesis del proveedor es falsa. Por lo tanto, debe elegir con cuidado el **límite**.
- Una vez recolectada la evidencia, mientras más alejada esté de 350 la cantidad de tornillos, más sustento tiene la decisión de rechazar la hipótesis del proveedor, ya que la evidencia es **fuerte**.

Analogía judicial

Podemos hacer una analogía con un proceso judicial.

- "Toda persona es inocente hasta que se demuestra lo contrario". Por lo tanto, la hipótesis previa (se suele llamar **hipótesis nula**) es la presunción de inocencia.
- La **hipótesis alternativa** es la hipótesis de culpabilidad, enfrentada a la hipótesis nula.
- Se presentan todas las **evidencias** correspondientes al caso.
- En base a las evidencias, se toma una **decisión**, que puede acarrear 2 tipos de errores:
  - Se puede declarar culpable a alguien que es inocente. Esta decisión **cambia** el estado del acusado en el caso de que sea encarcelado. En casos dudosos, se trata de evitar cometer este error.
  - Se puede declarar inocente a alguien culpable. Con esta decisión, si antes estaba libre, **mantiene** esa condición.
- La decisión deberá contemplar que no cualquier evidencia incriminatoria pueda ser comprobación de culpabilidad. Por lo tanto, la decisión está sujeta a **dudas razonables**.
- En caso de ser declarado culpable, se puede establecer además un **grado de responsabilidad**. Mientras más fuerte sea la evidencia, mayor será pena impuesta.

Hipótesis

Volviendo al cuento chino, vamos a repensar el problema de Ricardo. Consideremos μ a la cantidad **media** de tornillos por caja.

Este parámetro μ es **poblacional**, es decir, hace referencia a lo que pasa con **todas** las cajas de tornillos. Como la discusión es sobre esta cantidad media, las hipótesis que se van a testear, involucran a μ.

Entonces, tenemos las siguientes hipótesis enfrentadas:

- **Hipótesis nula**: $H_0 : \mu = 350$ (Hipótesis del proveedor, la asumida hasta ahora)
- **Hipótesis alternativa**: $H_1 : \mu < 350$ (Hipótesis de Ricardo, que se encuentra a prueba)

Estadístico de prueba

Para juntar evidencia, Ricardo decide tomar una **muestra** de $n = 100$ cajas y tomar el **promedio muestral** $\overline{x}$ de esas cajas.

A partir de la **media muestral** $\overline{x}$, intentará sacar conclusiones sobre la **media poblacional** μ.

Regla de decisión

Luego de tomar la muestra, Ricardo toma una decisión: Diremos que "Rechaza H_0" si las evidencias apoyan su sospecha y "Acepta H_0" si las evidencias no la apoyan lo suficiente.

Por lo tanto, como nunca tendremos certeza absoluta si H_0 es cierta o falsa, se pueden dar 4 escenarios:

	Rechaza H_0	**Acepta H_0**
H_0 **Verdadera**	Error tipo I	Decisión correcta
H_0 **Falsa**	Decisión correcta	Error tipo II

Como dijimos, está reacio a cambiar de proveedor de forma innecesaria porque le involucra nuevos esfuerzos y quemar un puente que luego potencialmente necesitará cruzar.

Por lo tanto, querrá además que la probabilidad de cometer el error tipo I (se denomina **nivel de significación**, notada α), sea pequeña. Se suele tomar por default $\alpha = 0.05$.

Comentario: Decimos que se "Acepta H_0" porque no significa que sea cierta. Puede que no hayamos juntado las evidencias suficientes para rechazarla.

Del mismo modo que si no se junta demasiada evidencia para probar la culpabilidad de una persona, no significa que la persona sea inocente.

Valores críticos

Como hemos dicho, para decidirse por alguna de las hipótesis, **no** debe tomar **cualquier** valor menor a μ_0=350 (valor de referencia) como prueba de que la hipótesis del proveedor es falsa.

Rechazará la hipótesis si el valor obtenido es *mucho menor* a 350 (contemplando la variabilidad del proceso). Dicho de otro modo, se rechaza la hipótesis nula si el valor observado de $\overline{x}$ es **significativamente** menor que 350.

La pregunta es: ¿A partir de qué valor consideramos que el promedio es significativamente menor que 350? Es decir, ¿cómo calculamos un **valor crítico** (llamémoslo x_c) de forma que se rechace H_0 para todos los promedios **menores** que x_c.

Además, para no verse influido por el resultado final, el valor crítico debería establecerse **previo a tomar la muestra**, ya que en base a este límite, se toma la decisión.

Es decir, **previo a tomar la muestra** se determina que:

$$\begin{array}{lcl} \text{Si } \overline{x} < x_c & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si } \overline{x} \geq x_c & \Rightarrow & \text{Se acepta } H_0 \end{array}$$

Distribución

Hay herramientas teóricas que permiten avalar que, bajo ciertas hipótesis, $\overline{x}$ tiene distribución aproximadamente normal.

Más aún, **asumiendo la hipótesis nula**

$$Z = \frac{\text{promedio} - \text{valor de referencia}}{\text{desvío}/\sqrt{\text{tamaño muestral}}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{\overline{x} - 350}{s/\sqrt{100}}$$

tiene distribución normal **estándard**.

Comentarios:

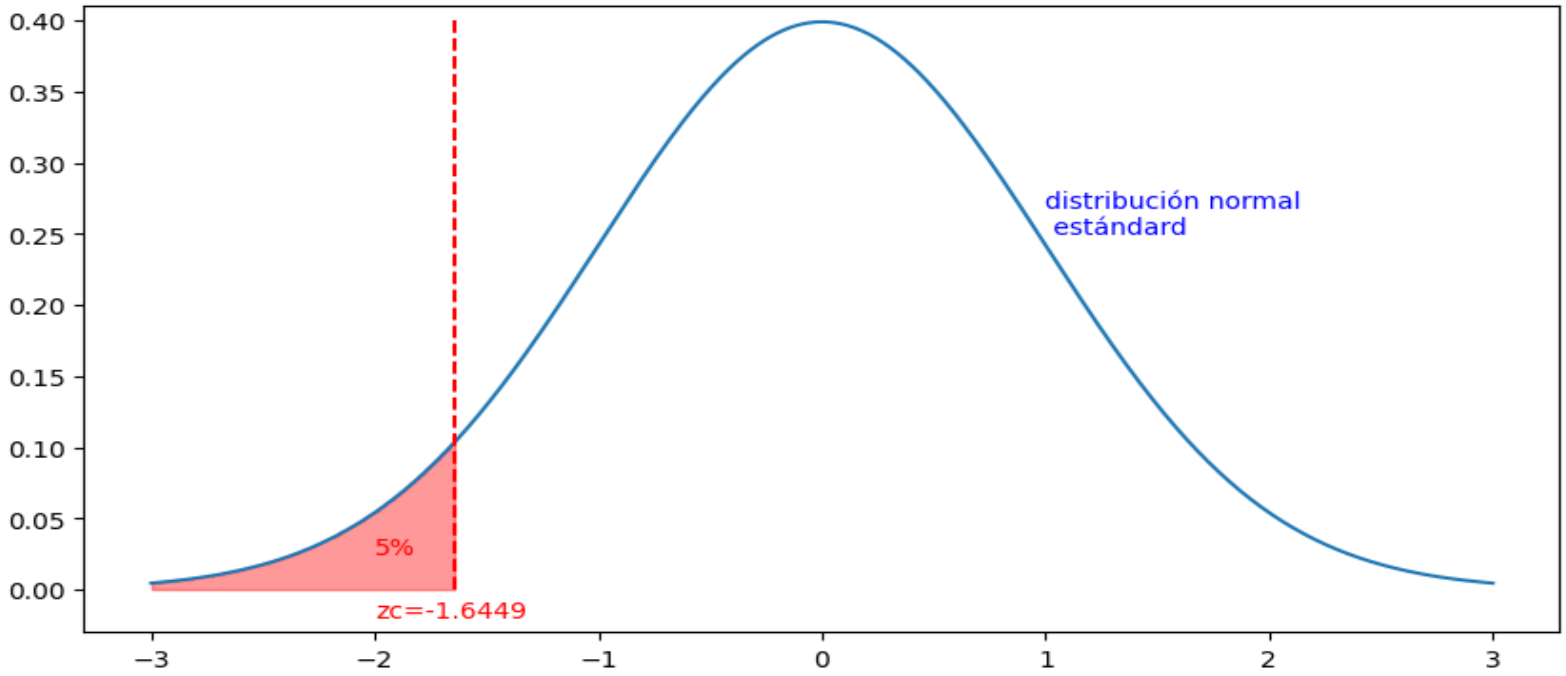
- Considerar que la *distribución* hace referencia a qué valores pueden tomar esas combinaciones de *promedios* y *desvíos*, y con qué **frecuencia** lo hace. Es decir, permite diferenciar los valores **probables** de los **improbables**.
- Incluir el desvío en el cálculo es importante ya que también se considera la **variabilidad** del proceso para identificar valores improbables.
- Notar que el promedio y el desvío no fueron reemplazados porque, al no querer que la decisión se vea influida por los resultados, esta distribución se plantea **previo** a tomar la muestra.

Región de rechazo

Por lo tanto, ya que esta distribución nos permite identificar lo improbable, podemos utilizarla para determinar el valor crítico de la decisión x_c. Es decir, se rechaza la hipótesis nula si el promedio observado es **improbablemente menor** que el valor de referencia.

Por otro lado, como usamos la distribución normal estándard, podemos cambiar el valor crítico x_c por un valor improbable para la normal estándard z_c y rechazar la hipótesis nula si $Z < z_c$

Utilizando el nivel de significación $\alpha = 5\%$, y asumiendo la hipótesis nula, se puede poner un límite que permita identificar el 5% menos probable:



Región de rechazo

Es decir, **previo a tomar la muestra** se determina que se rechaza la hipótesis nula si $Z < z_c = -1.6449$, en caso contrario, se acepta H_0.

Recordando las nociones probabilísticas, esto significa que de **todos** los casos en los que el proveedor tiene razón (H_0 cierta), en el 5% de ellas tomaremos una decisión equivocada.

Valor observado y decisión

Hasta ahora, todos los cálculos son **previos a tomar la muestra**. Estos cálculos nos permiten identificar valores extremos para lo que se pueda llegar a observar al tomar la muestra, y en base a ellos, tomar una decisión por alguna de las hipótesis enfrentadas.

Después de tomar la muestra, se observa un valor del promedio $\overline{x}_{obs}$ y otro para el desvío s_{obs}. Con estos datos, podemos construir el valor de z_{obs} que permita comparar con el valor crítico z_c:

$$z_{obs} = \frac{\text{promedio} - \text{valor de referencia}}{\text{desvío}/\sqrt{\text{tamaño muestral}}} = \frac{\overline{x}_{obs} - \mu_0}{s_{obs}/\sqrt{n}}$$

La regla para decidir es la siguiente:

- si $z_{obs} < z_c$, se rechaza H_0
- si $z_{obs} \geq z_c$, se acepta H_0.

Supongamos que después de tomar la muestra, se observa:

- un promedio **muestral** $\overline{x}_{obs} = 346.4$
- un desvío **muestral** $s_{obs} = 10$
- Con estos valores, el estadístico de prueba toma el valor $z_{obs} = \frac{346.4 - 350}{10/\sqrt{100}} = -3.6$

Como este valor es menor que $z_c = -1.6499$, se rechaza H_0.

Por lo tanto, se puede decir que Ricardo tiene **suficiente evidencia** para rechazar la hipótesis del proveedor. Más aún, tiene evidencia para **cambiar** de proveedor.

p-Valor

Con los datos observados, como $z_{obs} < z_c$, se rechaza la hipótesis nula. Es decir, una decisión **binaria**. Pero este binarismo omite que el valor observado z_{obs} es **mucho menor** que el valor crítico z_c.

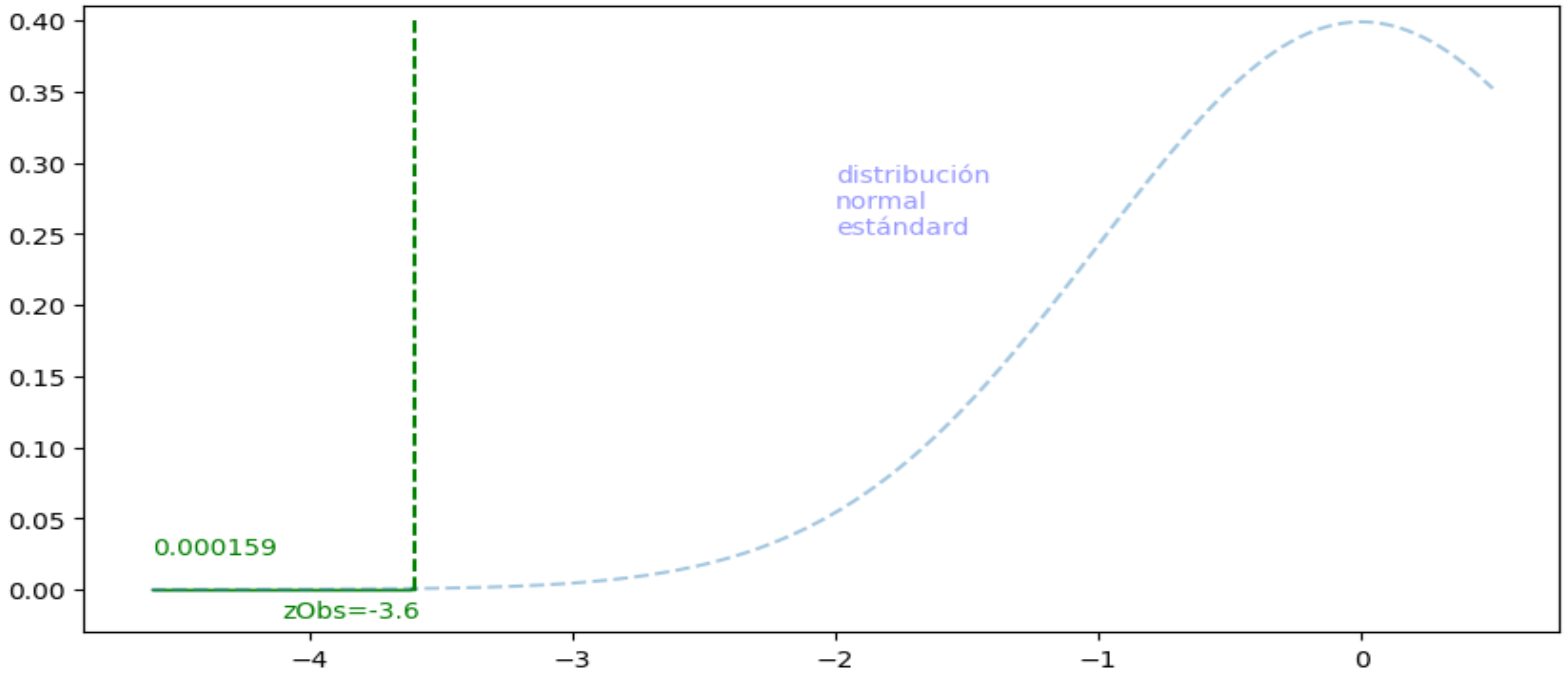
Por lo tanto, una vez que se toma la muestra, se puede **cuantificar** la distancia entre el valor observado y el valor crítico. Mientras más alejado esté z_{obs} de z_c (siempre y cuando sea **menor**), más evidencia hay **en contra** de la hipótesis nula.

Así como el valor crítico fue elegido utilizando la probabilidad, en este caso también. Definimos el p-Valor como la probabilidad de observar algo tan extremo que lo observado, **asumiendo la hipótesis nula verdadera**.

Es decir, aún dando el beneficio de la duda, mientras más improbable sea lo observado, más evidencia hay en contra de la hipótesis nula. Por lo tanto, mientras más **chico** sea el p-Valor, más evidencia hay **en contra** de la hipótesis nula.

Representación gráfica

Gráficamente, al ser una probabilidad que viene de una distribución normal, el p-Valor se representa con un área bajo la curva bajo una normal estándard:

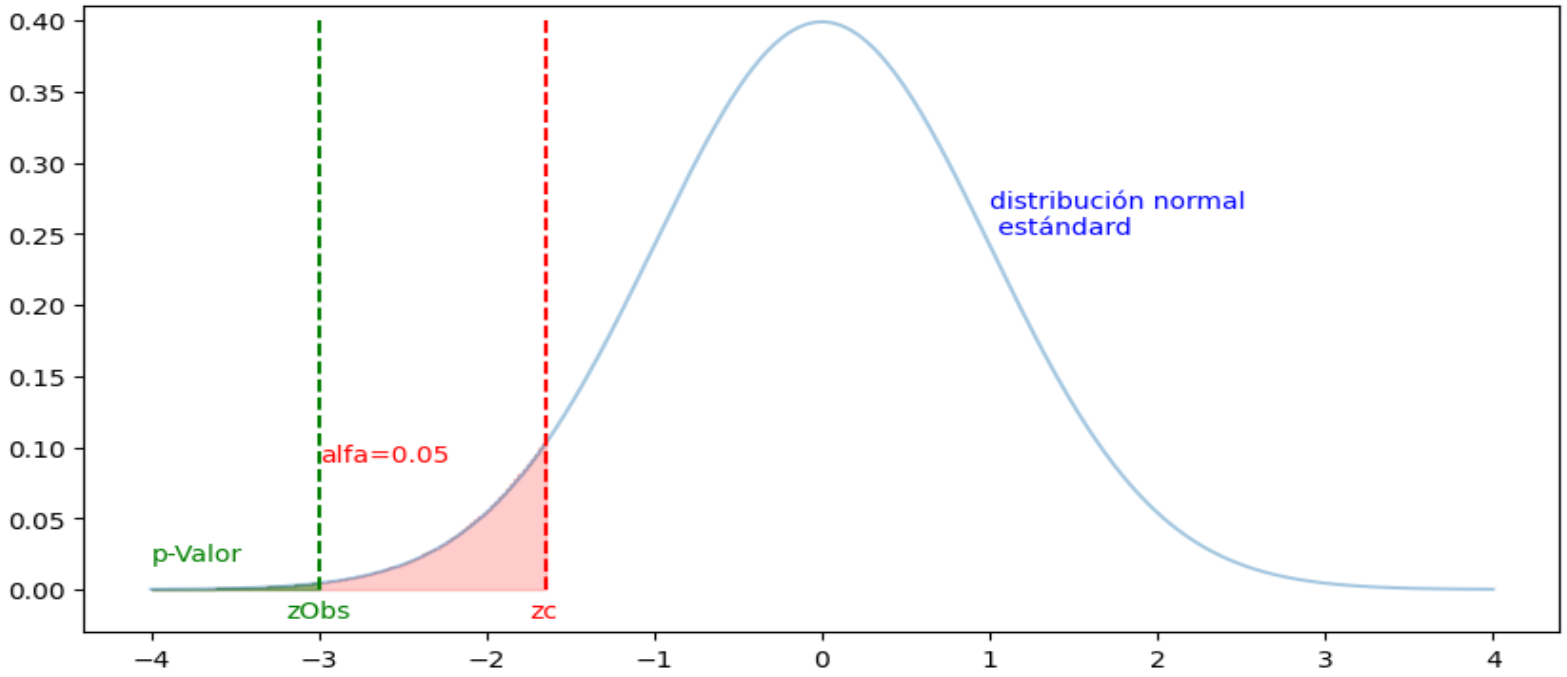


Es decir, asumiendo la hipótesis del proveedor, hay un 0.0159% de probabilidades de observar algo similar a lo observado. Esto es **mucho menor** que el valor establecido anteriormente. Por lo tanto, hay **más evidencia** para rechazar la hipótesis del proveedor.

Otra perspectiva

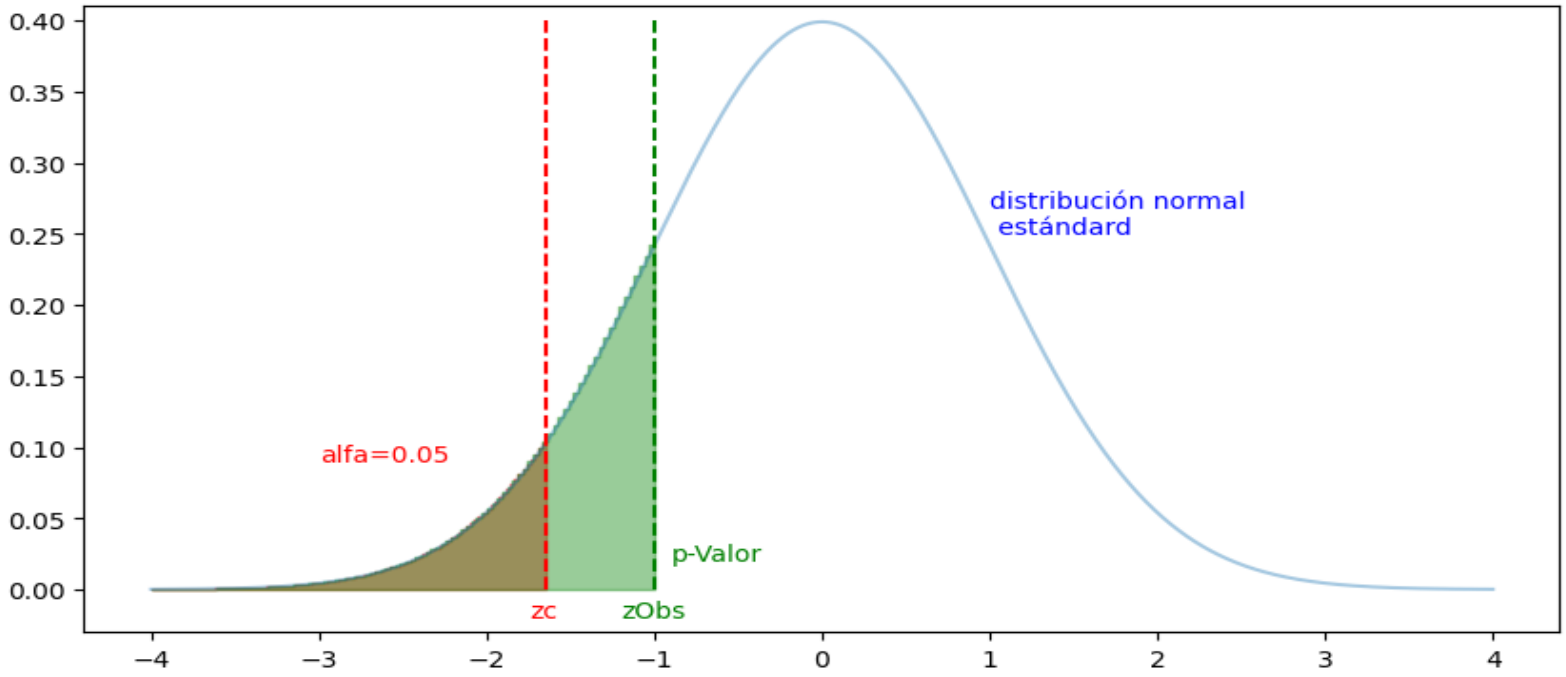
Se había establecido que si $z_{obs} < z_c$, se rechaza H_0.

Gráficamente, eso significa que el área bajo la curva hasta el valor de z_{obs} es **menor** que el nivel de significación α. Es decir, que el p-Valor es **menor** que $\alpha = 0.05$.



Otra perspectiva

Del mismo modo, si $z_{obs} \geq z_c$, entonces, el p-Valor será mayor o igual que α:



Por lo tanto, también podemos tomar una decisión respecto de H_0 comparando el p-Valor con el nivel de significación α.

Resumen test de hipótesis

Es decir, en este test de hipótesis sobre la media tenemos las siguientes características:

- **Previo a tomar la muestra:**
  - **Valor de referencia:** $\mu_0 = 350$
  - **Tamaño muestral:** $n = 100$
  - **Nivel de significación:** $\alpha = 0.05$
  - **Hipótesis nula:** $H_0 : \mu = \mu_0 = 350$
  - **Hipótesis alternativa:** $H_1 : \mu < \mu_0 = 350$
  - **Estadístico de Prueba:** $Z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$
  - **Valor crítico:** $z_c = -1.6499$
  - **Regla de decisión:** $\left\{ \begin{array}{lcl} \text{Si } Z < z_c & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si } Z \geq z_c & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$ ó
    $\left\{ \begin{array}{lcl} \text{Si p-Valor} < \alpha & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si p-Valor} \geq \alpha & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$
- **Posterior a tomar la muestra:**
  - **Valor muestral observado:** $z_{obs} = -3.6$
  - **Decisión:** Como $z_{obs} = -3.6 < -1.6499 = z_c$ y $p - Valor = 0.000159 < 0.05 = \alpha$, se rechaza H_0.

Perspectiva del proveedor

Previo a esta discusión acalorada, el proveedor hizo un análisis de costos y concluyó que si la máquina expendedora ofrecía una cantidad media de tornillos (μ) **mayor** que el estipulado, dejaba de ser rentable su negocio.

Por lo tanto, planteó otro test de hipótesis, en el que deseaba **cambiar** la máquina en caso de que ofreciera tornillos de más.

Es decir, asumió las siguientes características para su test, con un tamaño muestral menor:

- **Previo a tomar la muestra:**
  - **Valor de referencia:** $\mu_0 = 350$
  - **Tamaño muestral:** $n = 25$
  - **Nivel de significación:** $\alpha = 0.05$
  - **Hipótesis nula:** $H_0 : \mu = \mu_0 = 350$
  - **Hipótesis alternativa:** $H_1 : \mu > \mu_0 = 350$
  - **Estadístico de Prueba:** $Z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$
  - **Valor crítico:** $z_c = 1.6499$ (se da el opuesto al anterior por la simetría de la normal)
  - **Regla de decisión:** $\left\{ \begin{array}{lcl} \text{Si } Z > z_c & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si } Z \leq z_c & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$ ó

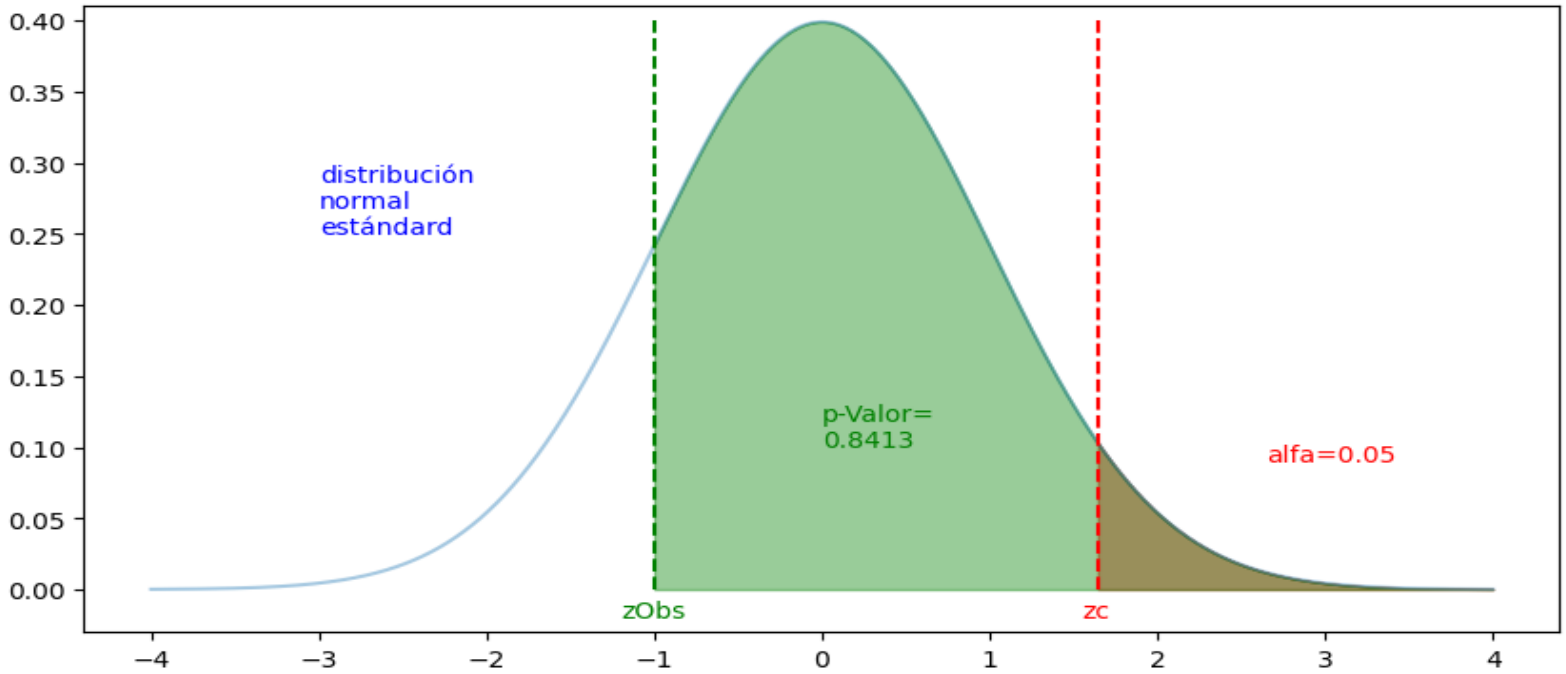
    $\left\{ \begin{array}{lcl} \text{Si p-Valor} < \alpha & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si p-Valor} \geq \alpha & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$

Después de la muestra

Supongamos que observa un valor de promedio muestral $\overline{x}_{obs} = 348$ y desvío muestral $s_{obs} = 10$. Entonces, $z_{obs} = \dfrac{348 - 350}{10/\sqrt{25}} = -1$.

Es decir, como $z_{obs} = -1 \leq 1.6499 = z_c$, se acepta H_0.

Siguiendo la hipótesis alternativa, ahora el p-Valor se calcula con el área hacia la derecha.



Es decir, también, como el p-Valor $=0.8413 \geq 0.05 = \alpha$, se **acepta** H_0. Por lo tanto, el proveedor decidió **mantener** la máquina con dicho funcionamiento.

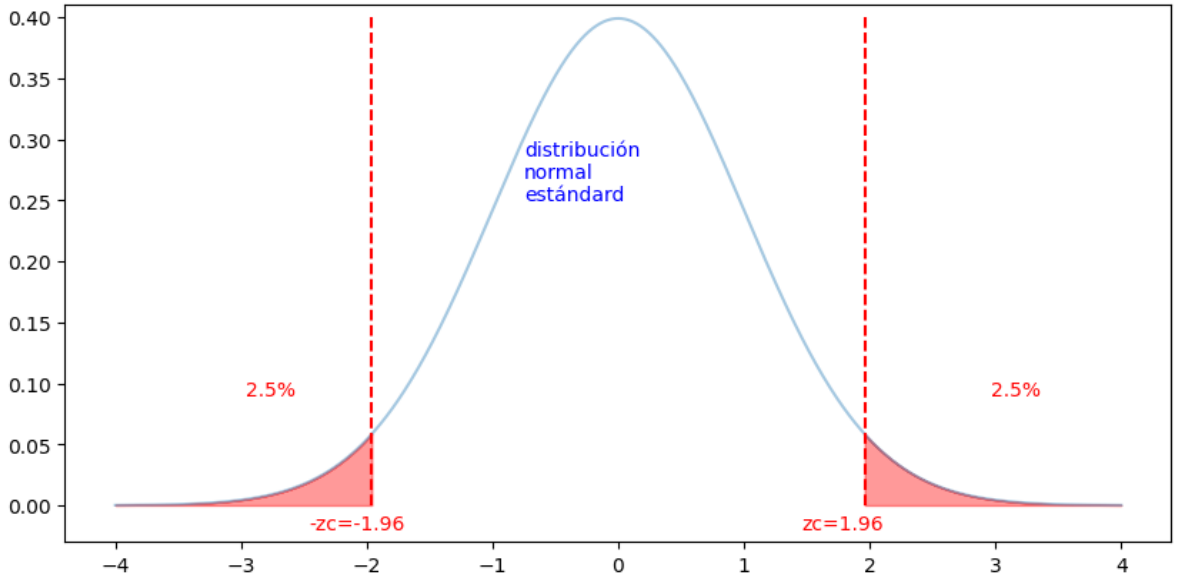
Mediación

Luego de la disputa entre Ricardo y su proveedor, se llama a una mediación para que ninguna de las dos partes resulte perjudicada, donde el proveedor se compromete a mejorar el funcionamiento de la máquina.

Por lo tanto, el organismo mediador analiza los datos, y decide someter a revisión la máquina expendedora si la media de los tornillos μ es **distinta** al valor de referencia $\mu_0 = 350$.

La mayor diferencia en este caso es que se rechaza la hipótesis nula cuando el valor observado es **significativamente mayor** o **significativamente menor** al valor de referencia. Es decir, asumiendo un nivel de significación $\alpha = 5\%$, ahora se debe buscar "repartir" esa probabilidad en 2.5% a cada lado del valor de referencia.

Gráficamente:



Antes de la muestra

Por lo tanto, asumió las siguientes características para su test:

- **Previo a tomar la muestra:**
  - **Valor de referencia:** $\mu_0 = 350$
  - **Tamaño muestral:** $n = 25$
  - **Nivel de significación:** $\alpha = 0.05$
  - **Hipótesis nula:** $H_0 : \mu = \mu_0 = 350$
  - **Hipótesis alternativa:** $H_1 : \mu \neq \mu_0 = 350$
  - **Estadístico de Prueba:** $Z = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}}$
  - **Valor crítico:** $z_c = 1.96$ (el valor utilizado en intervalos de confianza).
  - **Regla de decisión:** $\left\{ \begin{array}{lcl} \text{Si } Z > z_c \text{ o } Z < -z_c & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si } -z_c \leq Z \leq z_c & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$ ó

    $\left\{ \begin{array}{lcl} \text{Si p-Valor} < \alpha & \Rightarrow & \text{Se rechaza } H_0 \\ \text{Si p-Valor} \geq \alpha & \Rightarrow & \text{Se acepta } H_0 \end{array} \right\}$

Después de la muestra

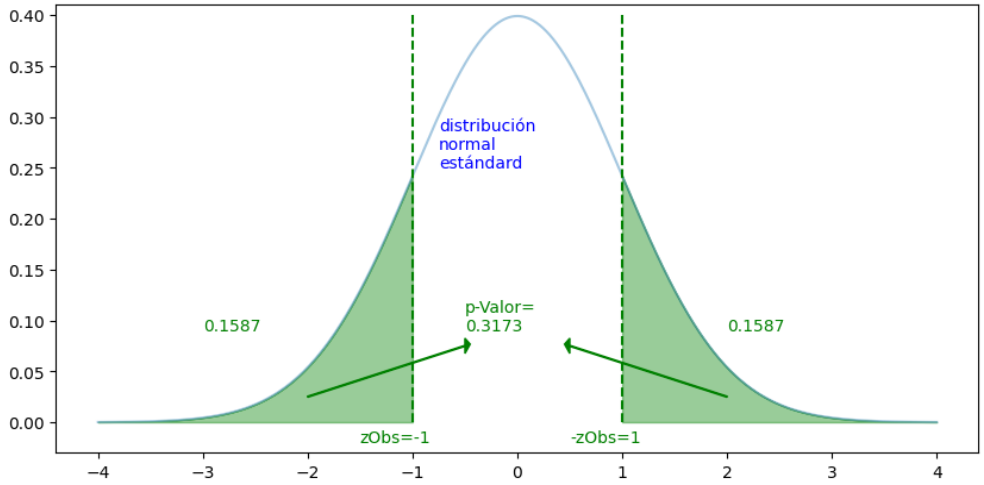
Supongamos que después de la muestra, se observa un promedio de $\overline{x}_{obs} = 348.5$ y un desvío $s_{obs} = 15$. Como

$$z_{obs} = \frac{\overline{x}_{obs} - \mu_0}{s_{obs}/\sqrt{n}} = \frac{348.5 - 350}{15/\sqrt{100}} = -1$$

está entre -1.96 y 1.96, se **acepta** H_0.

Respecto del p-Valor, ahora deben calcularse ambas "colas" de la distribución normal, ya que el concepto de "más extremo" utilizado en el p-Valor se refiere a la hipótesis alternativa.

Por lo tanto, dado este valor observado $z_{obs} = -1$, el p-valor se calcula del siguiente modo:



Notar que el p-Valor es mayor que $\alpha = 0.05$ y por lo tanto, es consistente con que se **acepta** H_0.

Tests de hipótesis para medias

Tests de hipótesis para una media

Hemos visto tres tipos de test de hipótesis para una media **poblacional** μ con **valor de referencia** μ_0:

- **Test unilateral a izquierda:** $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$
- **Test unilateral a derecha:** $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$
- **Test bilateral:** $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

Si bien todos los tests planteaban distintas regiones de rechazo y valores críticos, todos se ven unificados en lo siguiente: Se rechaza la hipótesis nula si el p-Valor es **menor** que el nivel de significación α.

Además, mientras más chico sea este p-Valor, más evidencia hay para rechazar la hipótesis nula.

Cálculos en Python

Para calcular estos tres test de hipótesis, generaremos una simulación de tamaño 100 centrada en μ_0=350:

```
from scipy.stats import norm
import numpy as np
normal = norm()
np.random.seed(0)
n=100
mu0=350
s=10
x1=mu0+s*normal.rvs(n)
zObs1=(np.mean(x1)-mu0)*np.sqrt(n)/np.std(x1)
print("zObs="+str(zObs1))
```

```
## zObs=0.5934028091872832
```

Cálculos en Python

Se puede utilizar el mismo comando ztest en python, cambiando el argumento alternative:

```
import statsmodels.stats.weightstats as smsw
# Test unilateral a izquierda
testUI=smsw.ztest(x1,value=mu0,alternative="smaller",ddof=0)
print("pValor="+str(testUI[1]))
```

```
## pValor=0.7235441953355044
```

```
# Test unilateral a derecha
testUD=smsw.ztest(x1,value=mu0,alternative="larger",ddof=0)
print("pValor="+str(testUD[1]))
```

```
## pValor=0.2764558046644956
```

```
# Test bilateral
testBi=smsw.ztest(x1,value=mu0,alternative="two-sided",ddof=0)
print("pValor="+str(testBi[1]))
```

```
## pValor=0.5529116093289912
```

Notar que en todos los casos, el p-Valor no es suficentemente chico. Eso se debe a que la muestra fue generada con una media de $\mu_0 = 350$, es decir, el valor de referencia utilizado.

Por lo tanto, tiene sentido que no se detecten diferencias en ninguno de los tests.

Cálculos en Python

Para notar diferencias, veamos qué sucede con los p-Valores cuando las muestras se centran en un valor sensiblemente menor. Por ejemplo, tomando como centro $\mu_2 = 340$.

```
np.random.seed(0)
n2=81
mu2=340
s=10
x2=mu2+s*normal.rvs(n2)
zObs2=(np.mean(x2)-mu0)*np.sqrt(n2)/np.std(x2)
print("zObs="+str(zObs2))
```

```
## zObs=-9.32900347941551
```

Ya podemos observar que el valor de z_{obs} es más lejano al cero que en el caso anterior, por lo que es más probable que sea un valor extremo.

Cálculos en Python

```
# Test unilateral a izquierda
testUI=smsw.ztest(x2,value=mu0,alternative="smaller",ddof=0)
print("pValor="+str(testUI[1]))
```

```
## pValor=5.34354749329709e-21
```

```
# Test unilateral a derecha
testUD=smsw.ztest(x2,value=mu0,alternative="larger",ddof=0)
print("pValor="+str(testUD[1]))
```

```
## pValor=1.0
```

```
# Test bilateral
testBi=smsw.ztest(x2,value=mu0,alternative="two-sided",ddof=0)
print("pValor="+str(testBi[1]))
```

```
## pValor=1.068709498659418e-20
```

Al compararlo con $\mu_0 = 350$, los tests que consideran el valor menor como alternativo (unilateral a izquierda y el bilateral) rechazan la hipótesis nula porque el p-Valor es muy pequeño. Esto se debe a que los datos se tomaron con una media **menor** al valor de referencia.

Por este mismo motivo también se observa un valor alto del p-Valor en el test unilateral a derecha, ya que no hay ninguna evidencia de que la media pueda ser **mayor** que 350.

Comentarios:

Notar que en todos los casos se obtiene el mismo valor de z_{obs}, pero cambian los p-Valores. Esto se debe a que el área acumulada para el p-Valor varía según la hipótesis alternativa:

- Notar que el p-Valor del test unilateral a izquierda se obtiene restándole a 1 el p-Valor del test unilateral a derecha.
- Notar que el p-Valor del test bilateral es el doble del p-Valor del test unilateral a derecha.

Comparación de 2 medias

Supongamos que la muestra simulada `x1` proviene de su anterior proveedor y `x2` proviene de un proveedor nuevo, con medias μ_1 y μ_2 respectivamente.

Ricardo quiere testear cuál de los proveedores ofrece mayor cantidad media de tornillos. Por lo tanto, puede testear las siguientes hipótesis:

- **Test unilateral a izquierda:** $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$. Aquí se busca detectar si el primer proveedor ofrece menos que el nuevo proveedor.
- **Test unilateral a derecha:** $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$. Aquí se busca detectar si el primer proveedor ofrece más que el nuevo proveedor.
- **Test bilateral:** $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. Aquí se busca detectar si hay diferencias entre ambos proveedores.

Comentarios:

- Todos estos tests se pueden replantear considerando que la diferencia de ambas medias tiene como referencia el cero. Es decir, por ejemplo, $H_0 : \mu_1 - \mu_2 = 0$ y cambiando todas las hipótesis alternativas de forma correspondiente.
- Al igual que los intervalos de confianza de comparación, estos tests incluyen una fórmula para el desvío que incluye la variabilidad y el tamaño muestral de cada muestra.

Comparación de 2 medias (Python)

Para realizar estos cálculos en Python, se puede usar el mismo comando pasando dos conjuntos de datos y cambiando el valor de referencia por 0:

```
# Test unilateral a izquierda
testUIDif=smsw.ztest(x1,x2,value=0,alternative="smaller",ddof=0)
print("pValor="+str(testUIDif[1]))
```

```
## pValor=0.9999999999998854
```

```
# Test unilateral a derecha
testUDDif=smsw.ztest(x1,x2,value=0,alternative="larger",ddof=0)
print("pValor="+str(testUDDif[1]))
```

```
## pValor=1.1461207764421363e-13
```

```
# Test bilateral
testBiDif=smsw.ztest(x1,x2,value=0,alternative="two-sided",ddof=0)
print("pValor="+str(testBiDif[1]))
```

```
## pValor=2.2922415528842726e-13
```

Vemos aquí que el único test que no rechaza es el unilateral a izquierda, ya que el p-Valor es grande. Esto se debe a que la primer muestra fue generada con $\mu_1 = 350$ y la segunda con $\mu_2 = 340$. Por lo tanto, no los datos no sostienen que $\mu_1 < \mu_2$.

Comparación de dos medias apareadas

Generemos una nueva muestra que suponemos que es del **mismo** proveedor, pero con arreglos en la **misma** máquina expendedora. Podríamos querer ver si los cambios en la máquina afectaron el valor medio de tornillos.

Es decir, considerando μ_A la media antes de los arreglos y μ_D la media después de los arreglos, se consideraría:

$$H_0 : \mu_A = \mu_D \text{ vs. } H_1 : \mu_A \neq \mu_D$$

Matemáticamente, esto sería restar las cantidades de los tornillos en cada caja y hacer un test de una única variable $d = x_D - x_A$ para ver si su media es cero. En Python:

```
np.random.seed(1)
n=100
mu3=352
s3=15
x3=mu3+s3*normal.rvs(n)
d=x3-x1
testBiPair=smsw.ztest(d,value=0,alternative="two-sided",ddof=0)
print("pValor="+str(testBiPair[1]))
```

```
## pValor=0.1340974961206925
```

Es decir, estos arreglos no cambiaron significativamente la media de tornillos dispensados por la máquina.

Comparación de más de dos medias

Supongamos el proveedor recibe presentaciones de tres fabricantes de máquinas expendedoras, cada una con un período de prueba gratuito. Para elegir entre las tres opciones, decide testear si hay diferencia entre las medias de las máquinas. Es decir, si μ_1, μ_2 y μ_3 son las medias de cada máquina, se testea como hipótesis nula:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

Para la hipótesis alternativa, hay que considerar que la negación de "todas las medias son iguales" es "hay algún par de medias distintas". Es decir,

$$H_1 : \mu_1 \neq \mu_2 \text{ o } \mu_1 \neq \mu_3 \text{ o } \mu_2 \neq \mu_3$$

Bajo ciertas condiciones, para testear esta igualdad se puede utilizar un procedimiento llamado ANOVA (**AN**alyisis **O**f **VA**riance). El procedimiento se puede extender a muchos más grupos que tres, pero también complejiza los análisis posteriores.

ANOVA en Python

El procedimiento ANOVA tiene variantes, pero la estándard es la denominada "one-way ANOVA", y en python se utiliza con el comando `anova_oneway`:

```
import statsmodels.stats.oneway as smsow
smsow.anova_oneway([x1,x2,x3])
```

```
<class 'statsmodels.stats.base.HolderTuple'>

statistic = 37.61193521075858

pvalue = 2.1461445941328226e-14

df = (2.0, 181.70011652455145)

df_num = 2.0

df_denom = 181.70011652455145

nobs_t = 281.0

n_groups = 3

means = array([350.59808016, 339.55450215, 352.90874278])

nobs = array([100.,  81., 100.])

vars_ = array([102.60874942, 102.81768673, 178.06852793])

use_var = 'unequal'

welch_correction = True

tuple = (37.61193521075858, 2.1461445941328226e-14)
```

Notemos aquí la importancia del p-Valor, porque los resultados son muy complejos y hay muchas cosas para analizar. Sin embargo, con ver el pequeñísimo tamaño del p-Valor, hay evidencia que sostiene que por lo menos algún par de las 3 medias son diferentes.

Comparaciones múltiples

El hecho de rechazar la hipótesis nula da cuenta de que las medias no son iguales, pero no especifica cuáles de ellas son diferentes entre sí.

Por eso, una vez que se rechazó H_0, se puede visualizar cuáles de ellas son diferentes considerando un test de comparaciones múltiples de Tukey:

```
import statsmodels.sandbox.stats.multicomp as smsmc
lisDat=[i for l in [x1,x2,x3] for i in l]
lisG=[i for l in [["1"]*len(x1),["2"]*len(x2),["3"]*len(x3)] for i in l]
dat=np.array(lisDat)
datG=np.array(lisG)
MC=smsmc.MultiComparison(dat,datG)
print(MC.tukeyhsd())
```

```
##  Multiple Comparison of Means - Tukey HSD, FWER=0.05
## ====================================================
## group1 group2 meandiff p-adj   lower    upper  reject
## ----------------------------------------------------
##      1      2 -11.0436    0.0 -15.0526 -7.0346   True
##      1      3   2.3107 0.3241  -1.4821  6.1034  False
##      2      3  13.3542    0.0   9.3452 17.3633   True
## ----------------------------------------------------
```

Esto especifica que la diferencia se observa en la media 2 (recordemos que se utilizó como base una media de 340), mientras que las diferencias entre μ_1 y μ_3 no fueron significativas (recordemos que se utilizaron como base medias de 350 y 352, respectivamente).

Tests para proporciones

Tests para una proporción o porcentaje

El proveedor quiere realizar otro análisis. Considera que si la cantidad de tornillos en una caja es menor que 340, la diferencia en el peso hace que se sospeche de su contenido y que se revise la cantidad, poniendo su trabajo en riesgo.

Por lo tanto, si el porcentaje de cajas con cantidad reducida está por debajo del 5%, considera que es un riesgo aceptable. En caso contrario, deberá hacer cambios a la máquina.

Es decir, dado π el porcentaje **poblacional** de cajas que tienen menos de 340 tornillos, se pueden plantear las siguientes hipótesis:

$$H_0 : \pi \leq 5\% \text{ o } H_1 : \pi > 5\%$$

Tests para una proporción en Python

Para hacer este cálculo en Python, podemos apelar al comando `proportions_ztest`:

```
import statsmodels.stats.proportion as smsp
fObs1=np.sum(x1<340)
n1=len(x1)
print("pObs="+str(fObs1/n1))
TestPropUD=smsp.proportions_ztest(fObs1,n1,value=0.05,alternative= "larger")
print("zObs="+str(TestPropUD[0]))
print("pValor="+str(TestPropUD[1])).
```

```
pObs=0.15

zObs=2.800560168056019

pValor=0.0025506998230024 96
```

Es decir, en este caso, el riesgo que corre el proveedor es mayor que lo aceptable y debe cambiar la máquina.

Comparación de porcentajes

Comparemos el porcentaje de cajas con menos de 340 tornillos en la primer máquina (media=350 tornillos) con la tercera (media=352 tornillos).

El porcentaje de cajas con menos de 340 tornillos debería bajar, pero no sabemos si ese porcentaje es **significativamente** menor.

Entonces, se puede plantear el siguiente test:

$$H_0 : \pi_1 = \pi_3 \text{ vs. } H_1 : \pi_1 > \pi_3$$

- π_1 es el porcentaje de cajas de cantidad reducida provenientes de la primer máquina.
- π_3 es el porcentaje de cajas de cantidad reducida provenientes de la tercer máquina.

Comparación de porcentajes en Python

Para realizar el cálculo en Python, se utiliza el mismo comando pero agregando una lista de valores

```
fObs3=np.sum(x3<340)
n3=len(x3)
print("pObs3="+str(fObs3/n3))
TestPropDif=smsp.proportions_ztest([fObs1,fObs3],[n1,n3],value=0,alternative
print("zObs="+str(TestPropDif[0]))
print("pValor="+str(TestPropDif[1]))
```

```
pObs3=0.14
```

```
zObs=0.20082507770959226
```

```
pValor=0.4204176766060942
```

Por lo tanto, si bien bajó el porcentaje de 15% a 14%, esa diferencia no resulta significativa.

Otros tests

Test de normalidad

Hay tests para considerar si una distribución efectivamente viene de una variable normal. Es decir:

$$H_0 : \text{la variable es normal vs. } H_1 : \text{ la variable NO es normal}$$

Este test en realidad permite descartar normalidad, ya que si se rechaza H_0, hay evidencia para decir que la variable **no** proviene de una distribución normal.

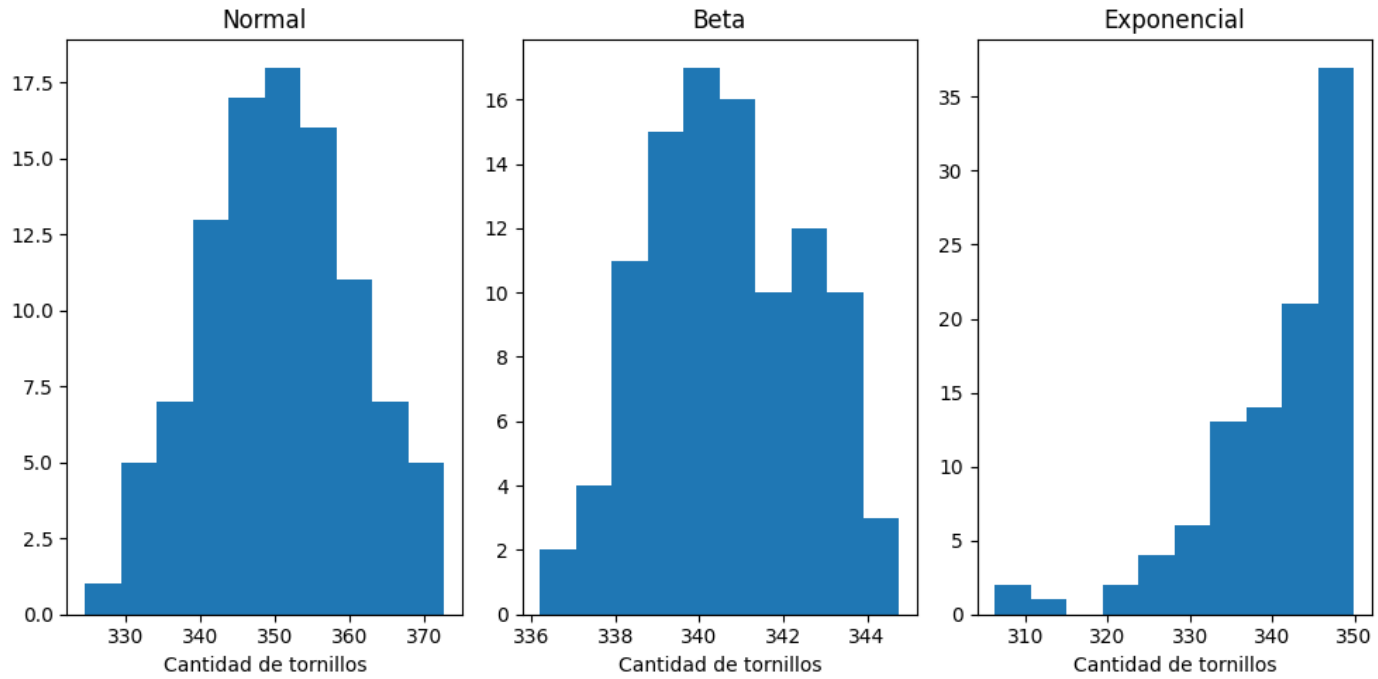
Sin embargo, si no se rechaza la hipótesis nula, **no confirma** la normalidad de la variable.

Diferentes distribuciones

Vamos a ver cómo funciona este test con una de las variables que ya generamos y sabemos que es normal.

Para generar variables que **no** sean normales, tomaremos una variable con una asimetría moderada (se denomina distribución beta), y una de fuerte asimetría (distribución exponencial) para ver en qué casos el test termina descartando la normalidad.

```
from scipy.stats import beta
bet=beta(3,4);y1=345-10*bet.rvs(n)
from scipy.stats import expon
expo=expon(scale=10);y2=350-expo.rvs(n)
```



Test de normalidad en Python

En Python este test (denominado Shapiro-Wilk) se utiliza con el comando `shapiro`:

```
import scipy.stats as stats
print("pValor="+str(stats.shapiro(x1)[1]))
# Caso normal
```

```
## pValor=0.868945837020874
```

```
print("pValor="+str(stats.shapiro(y1)[1]))
# Caso beta
```

```
## pValor=0.336654305458068 85
```

```
print("pValor="+str(stats.shapiro(y2)[1]))
# Caso exponencial
```

```
## pValor=6.053584744591944e-09
```

Notemos que en el caso normal, el p-valor da alto. Además, en el caso exponencial se rechaza la hipótesis de normalidad, lo cual es consistente.

Sin embargo, la variable beta no es normal y el test no logra rechazarlo. Aquí se demuestra que el test no sirve para **confirmar** normalidad, sirve para descartarla en algunos casos concretos.

Test de correlación

También hay tests para ver si dos variables están correlacionadas. Esto es importante porque muchos procedimientos estadísticos que requieren que las variables no tengan influencia entre sí.

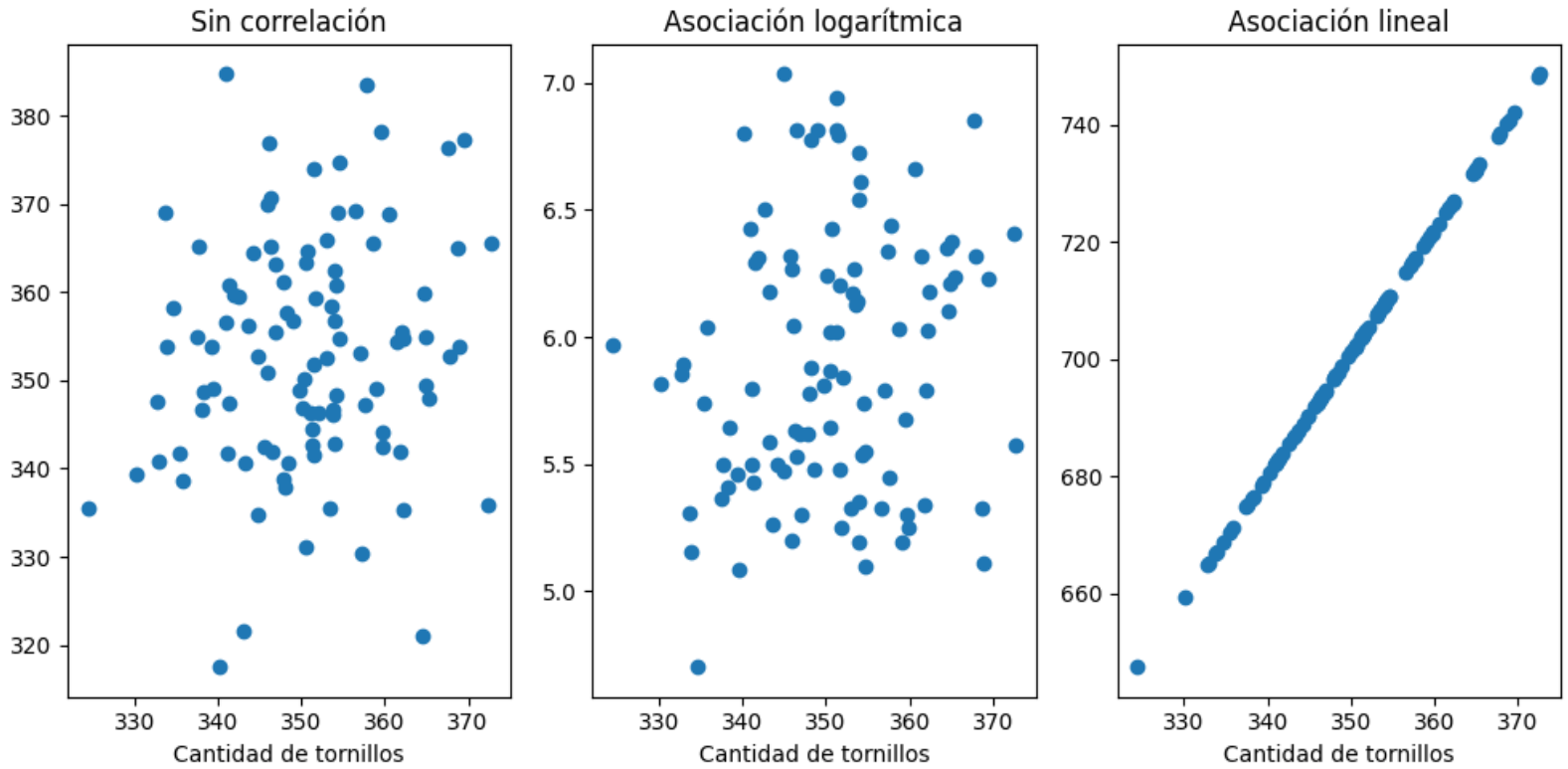
Es decir, considerando dos variables X e Y, y su correlación ρ, se plantean las siguientes hipótesis:

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

Diferentes asociaciones

Podemos visualizar tres asociaciones posibles, por ejemplo, comparando ambas distribuciones normales ya generadas `x1` y `x3` (que no se generaron mediante ningún vínculo), `x1` y su logaritmo (deberían tener una asociación lineal más débil) y `x1` y una función lineal aplicada a ella (debería dar correlación cercana a 1):

```
np.random.seed(0)
z2=2*x1+1+normal.rvs(n)
z1=np.log(x1)+0.5*normal.rvs(n)
```



Test de correlación en Python

En Python se utiliza el comando `pearsonr`

```
print("pValor="+str(stats.pearsonr(x1,x3)[1]))
# Sin correlación
```

```
pValor=0.13793984548034918
```

```
print("pValor="+str(stats.pearsonr(x1,z1)[1]))
# Asociación logarítmica
```

```
pValor=0.09874405876890886
```

```
print("pValor="+str(stats.pearsonr(x1,z2)[1]))
# Asociación lineal
```

```
pValor=0.0
```

Vemos nuevamente lo mismo, el test ayuda a identificar casos de fuerte asociación lineal, pero cuando la asociación no tiene esa estructura deja de detectar el vínculo.

Por lo tanto, recordar que aceptar la hipótesis nula **no confirma** que no hay asociación.

Resumen

- Los tests de hipótesis no son tan informativos como los intervalos de confianza. Por ejemplo,
  - si hacemos un test para ver si hay diferencia estadística entre dos poblaciones, podremos decidir si **hay o no hay** diferencia.
  - si hacemos un intervalo para ver si hay diferencia estadística entre dos poblaciones, podremos saber **cuanta** diferencia hay.
- Esto nos podría llevar a pensar que los intervalos de confianza son más útiles que los tests de hipótesis. Sin embargo, los tests de hipótesis tienen un valor de referencia que simplifica algunos cálculos. Esto permite contestar preguntas que los intervalos de confianza no pueden responder. Como por ejemplo, si un conjunto de datos proviene de una distribución normal o no.
- Además de esta versatilidad, provee una forma de elegir que se aplica a todo el mundo de los tests. Si el p-Valor es pequeño, podemos decir que hay evidencia estadística suficiente para rechazar la hipótesis nula.
- De todas formas, vale aclarar que no cualquier test es válido en cualquier circunstancia. Cada test trae aparejado una serie de hipótesis que de no cumplirse, las conclusiones no serán válidas.