{desafío}
latam_

Hipótesis y Correlación _

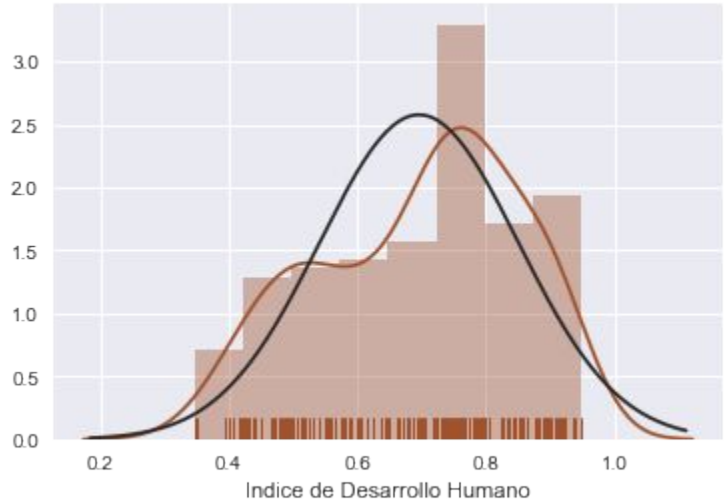
Seaborn

Objetivo de Seaborn

- Una librería orientada a sintetizar buenas prácticas respecto a la visualización de datos, considerando el marco de análisis en ciencia de datos con pandas, numpy y matplotlib.
- De esta manera nos centramos más en el análisis que en el código para realizar gráficos estándares.

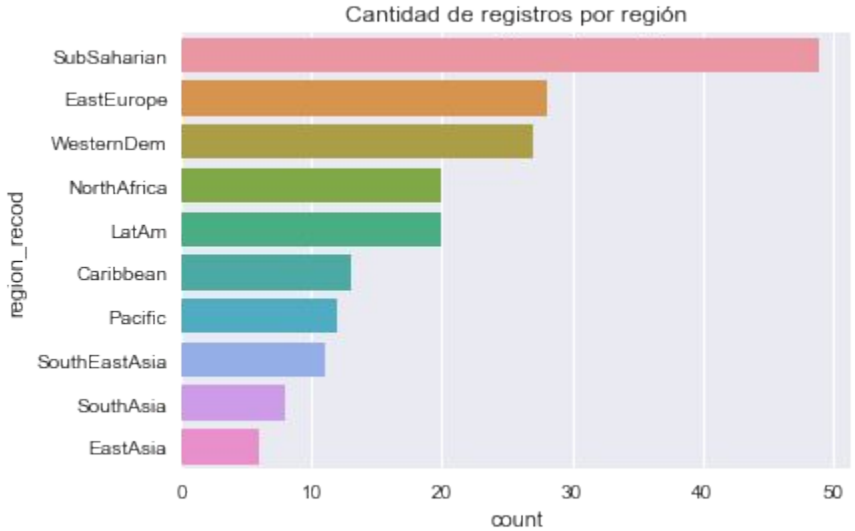
Gráfico distributivo

```
import seaborn as sns

sns.distplot(df[‘undp_hdi’],rug=True, axlabel=”Indice de desarrollo humano”,
             fit=stats.norm, color=’sienna’)
```



Conteo de frecuencias

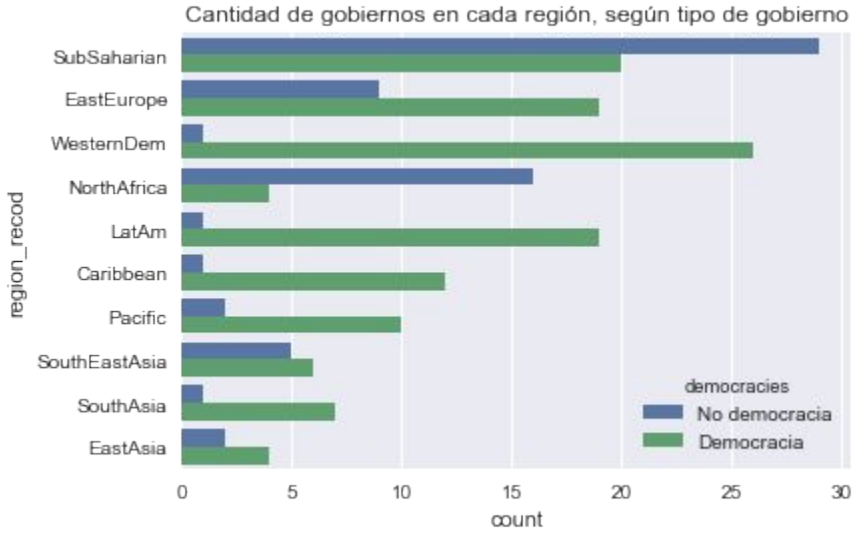
```
sns.countplot(y=df[‘region_recod’],
              order=df[‘region_recod’].value_counts().index)
```



{desafío} latam_

Conteo de frecuencias condicional a un factor

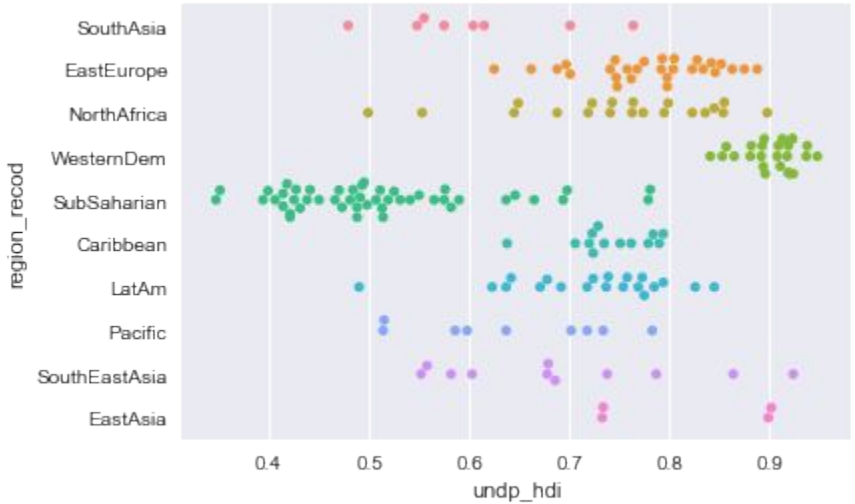
```
sns.countplot(y=df['region_recod'], hue=df['democracies'],
              order=df['region_recod'].value_counts().index)
```



{desafío} latam_

Swarmplot

```
sns.swarmplot(y=df['region_recod'], x=df['undp_hdi'])
```



FacetGrid

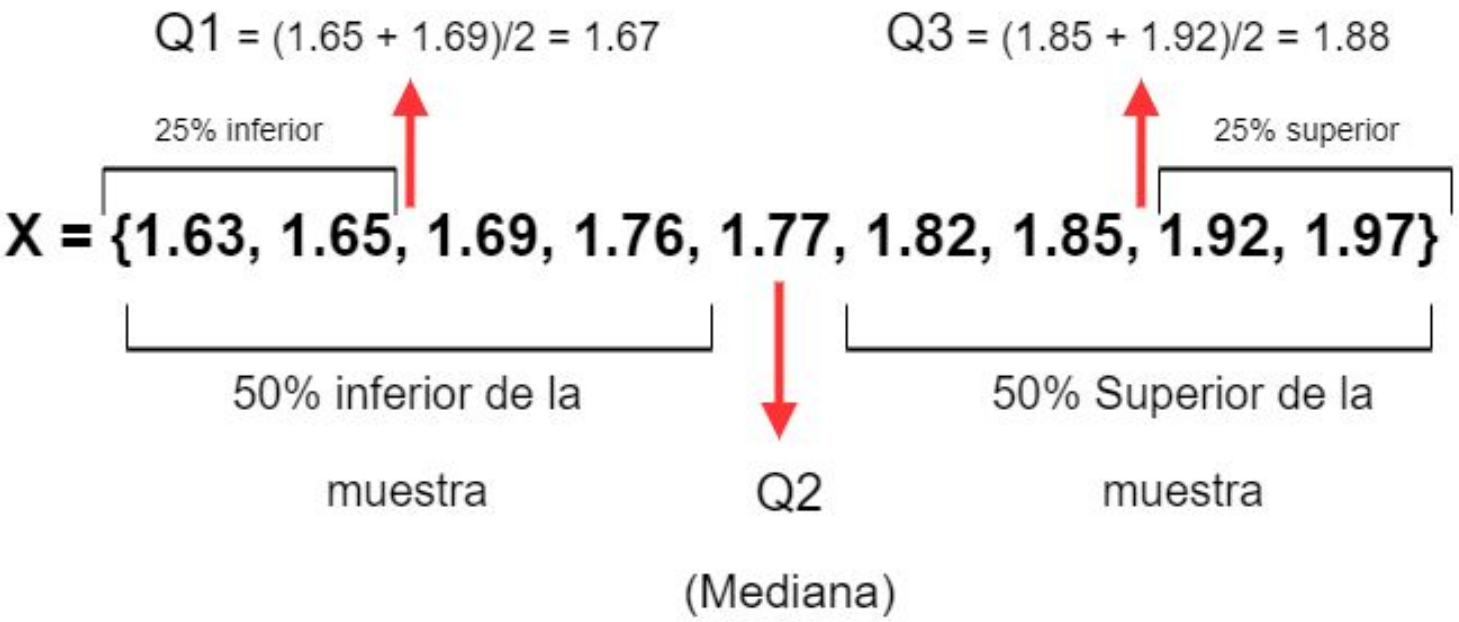
- FacetGrid nos permite graficar múltiples figuras condicional a un valor en específico.
- Flujo de trabajo con FacetGrid:
  - Iniciar un objeto con `sns.FacetGrid` declarando el DataFrame y las variables.
  - Aplicar una o más funciones mediante `map` o `map_dataframe`.

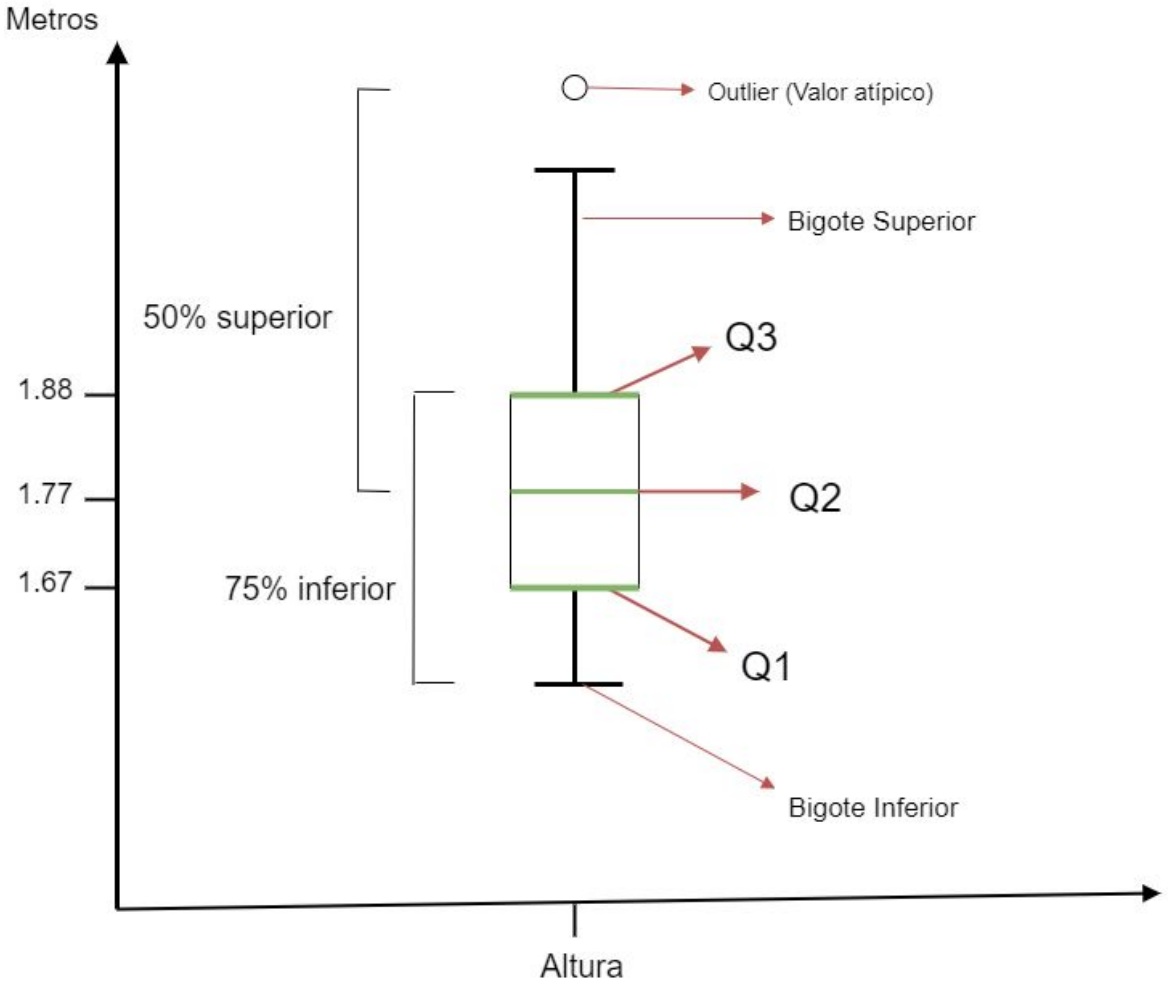
Scatterplots

- Un scatterplot o diagrama de dispersión resume el comportamiento entre dos variables.
- Para ello, asume que cada registro se conforma de coordenadas (x, y).
- Mediante los diagramas de dispersión podemos observar “qué tan juntas” viajan dos variables.

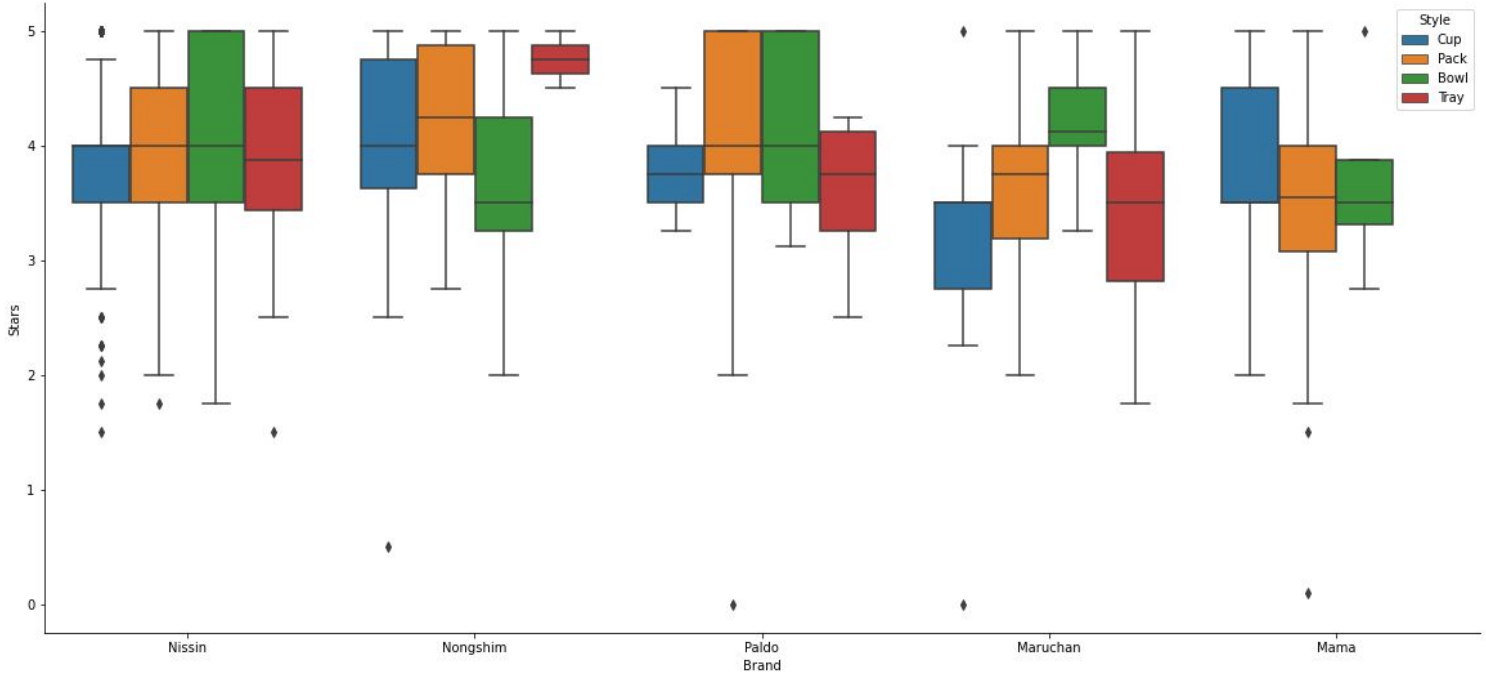
Boxplots

- Son un tipo de gráfico que permite mostrar gráficamente los cuartiles de una serie de datos.
- Permiten detectar de forma visual la presencia de valores atípicos (outliers) univariados.


Q1 = (1.65 + 1.69)/2 = 1.67
Q3 = (1.85 + 1.92)/2 = 1.88
25% inferior
25% superior
X = {1.63, 1.65, 1.69, 1.76, 1.77, 1.82, 1.85, 1.92, 1.97}
50% inferior de la muestra
Q2
(Mediana)
50% Superior de la muestra


Metros
Outlier (Valor atípico)
Bigote Superior
50% superior
Q3
1.88
1.77
Q2
1.67
75% inferior
Q1
Bigote Inferior
Altura

```
sn.boxplot(x = 'Brand', y = 'Stars', data = rated_brandStyle, hue = 'Style')
sn.despine()
```



Correlación

Preliminares

- La correlación y covarianza son piedras angulares para métodos más sofisticados.
- Covarianza:

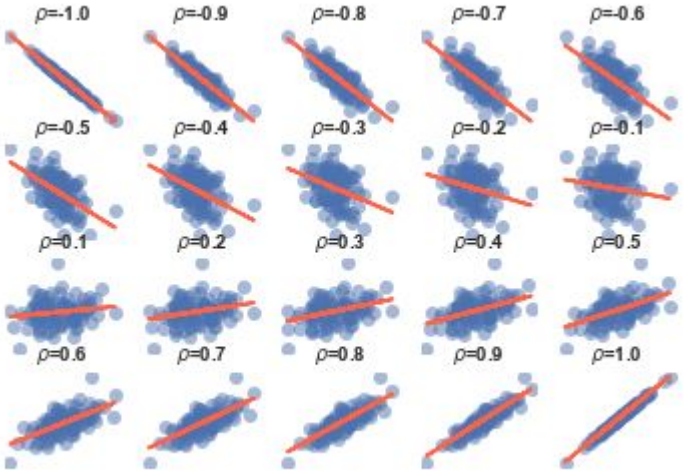
$$\text{Covarianza}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

- Correlación:

$$\text{Correlacion}(x, y) = \frac{\text{Covarianza}(x, y)}{\sqrt{\text{Varianza}(x)} \sqrt{\text{Varianza}(y))}}$$

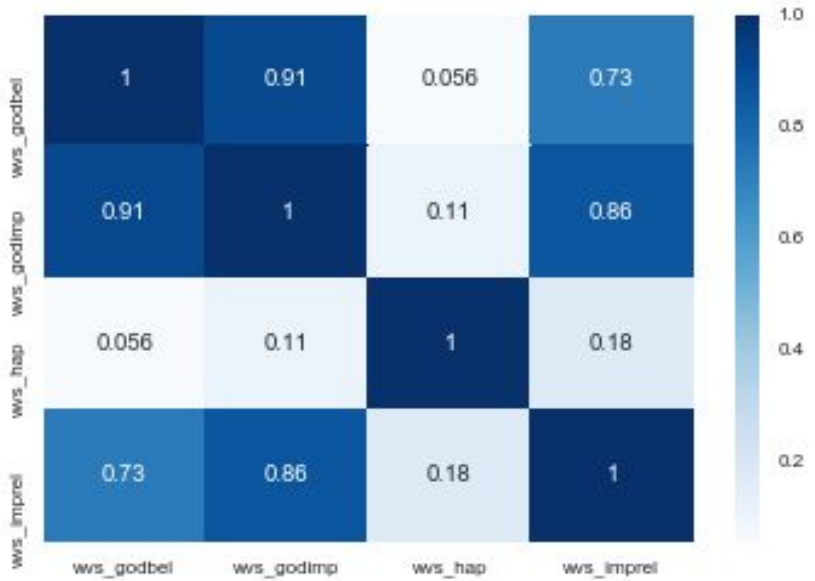
Intuición

- Dos elementos a tomar en cuenta:
  - La pendiente de la recta.
  - La dispersión de los puntos respecto a la recta.
- Valores positivos = Asociación proporcional directa entre x e y.
- Valores negativos = Asociación proporcional inversa entre x e y.



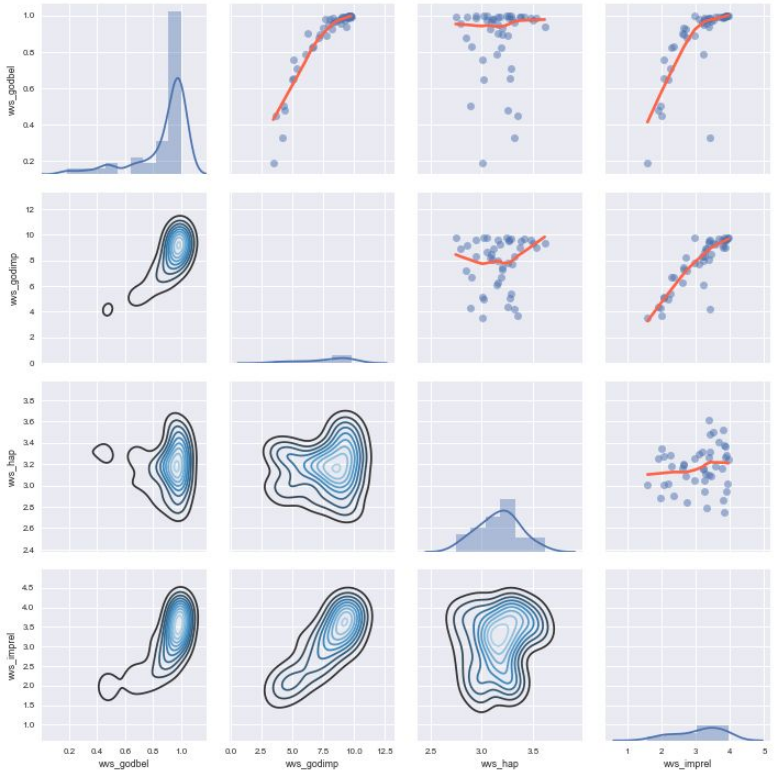
Formas de visualizar correlaciones: Heatmap

```
# Dado un subset de variables,
extraemos su matriz de correlación

corr_mat = working_subset.corr()

# implementamos el método sns.heatmap

sns.heatmap(corr_mat, cmap='Blues',
annot=True);
```



Formas de visualizar correlaciones: PairGrid

```
grid = sns.PairGrid(working_subset)
grid = grid.map_diag(sns.distplot)
grid = grid.map_lower(sns.kdeplot)
grid = grid.map_upper(sns.regplot)
```



Hipótesis

Objetivo

- Idea fundacional de la inferencia estadística: **tomar decisiones o esclarecer juicios en base a información limitada.**
- Definición de hipótesis: **juicio empíricamente comprobable sobre la relación entre dos o más.**
- En el contexto estadístico, para realizar pruebas de hipótesis debemos seguir una serie de pasos:
    - Definir enunciados de hipótesis.
    - Definir un estadístico de prueba.
    - Definir una distribución de la hipótesis nula
    - Definir un puntaje de corte.

Definición de hipótesis

- Hipótesis nula: Es la hipótesis que establece que nuestro punto estimado es nulo (o en términos generales, que no hay efecto)
- Hipótesis alternativas: Es la hipótesis que nosotros como investigadores conjeturamos

Calculando estadísticos de prueba

- Con las hipótesis declaradas, ahora podemos obtener un estadístico de prueba
- Existe una forma canónica para evaluar hipótesis:

$$Z = \frac{\hat{\theta} - \Theta}{\sigma / \sqrt{n}} \Rightarrow \frac{\underbrace{\hat{\theta}}_{\text{Estimador Muestral}} - \underbrace{\Theta}_{\text{Estimador Muestral}}}{\underbrace{\sigma / \sqrt{N}}_{\text{Error Estandar}}}$$

Evaluando nuestra hipótesis

- Hipótesis nula: Es la hipótesis que establece que nuestro punto estimado es nulo (o en términos generales, que no hay efecto)
- Hipótesis alternativas: Es la hipótesis que nosotros como investigadores conjeturamos

Prueba de Hipótesis frente a una constante

- Definimos las hipótesis

$$H_o : \bar{x}_{confianza} = .7$$
$$H_a : \bar{x}_{confianza} \neq .7$$

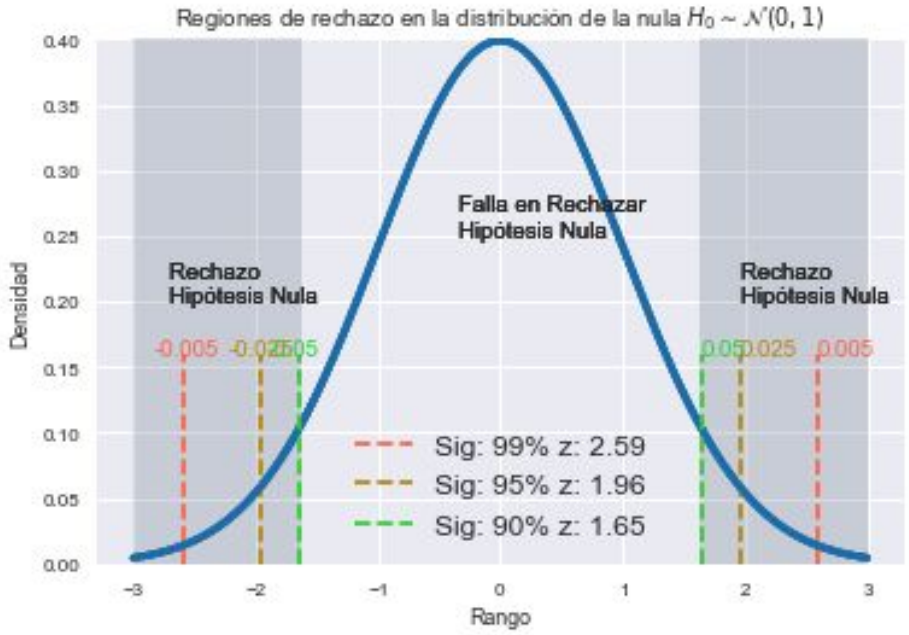
- Definimos el estimador de prueba:

$$t = \frac{\bar{x}_{confianza} - .7}{\sqrt{\frac{\sigma^2}{N}}}$$

- Posteriormente, definimos frente a qué puntaje de corte contrastamos

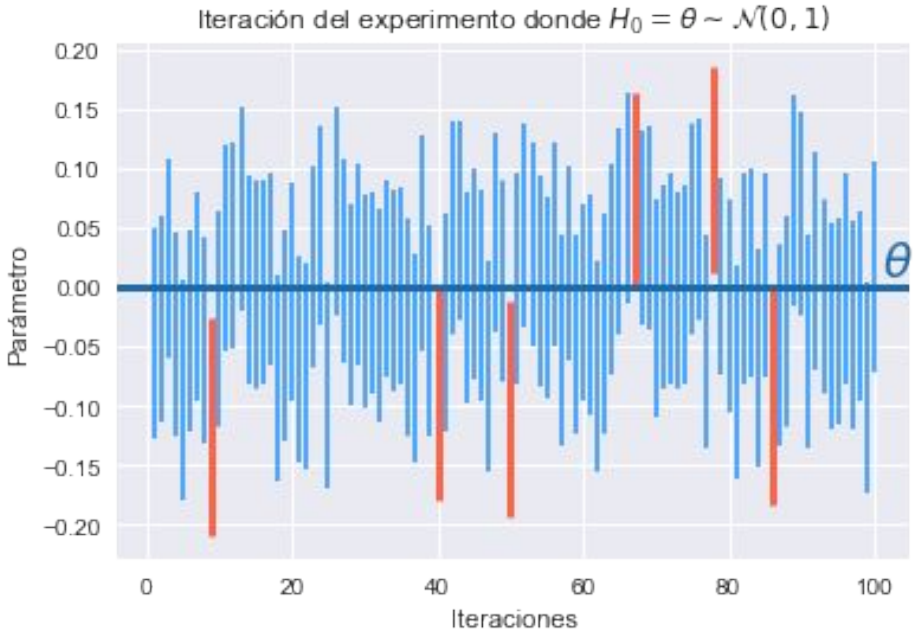
Regiones críticas

- Resulta que la distribución de la nula es la normal estandarizada.
- En esta buscamos evaluar nuestro estimador de prueba respecto a si se posiciona en regiones de rechazo o no rechazo de la nula.



Noción de Significancia Estadística

- Si tenemos evidencia para rechazar la nula, significa que en un 95% de las iteraciones de experimento, tendremos un resultado similar.



Puntajes críticos

- Si tenemos evidencia para rechazar la nula, significa que en un 95% de las iteraciones de experimento, tendremos un resultado similar.

Valor	Cobertura	Significado
2.58	99%	Si replicamos 100 un experimento bajo condiciones similares, tendremos 99 ocasiones donde el resultado será similar.
1.96	95%	Si replicamos 100 un experimento bajo condiciones similares, tendremos 95 ocasiones donde el resultado será similar.
1.68	90%	Si replicamos 100 un experimento bajo condiciones similares, tendremos 90 ocasiones donde el resultado será similar.

Prueba de hipotesis para muestras independientes

- Definimos las hipótesis:

$$H_o : \bar{x}_1 = \bar{x}_2$$
$$H_a : \bar{x}_1 \neq \bar{x}_2$$

- Definimos el estimador de prueba:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

{desafío} latam_

Academia de talentos digitales

www.desafiolatam.com