Nivelatorio: Estadística para la analítica

Sesión 5: Inferir relaciones entre dos variables, a partir de la información recolectada en una muestra

Diego Antonio Bohórquez Ordóñez
dabohorquez@Icesi.edu.co

Maestría en Ciencia de Datos

Agenda

1. ¿Qué más podemos inferir de la muestra?
2. ¿Existe relación entre una variable cualitativa y una cuantitativa?
3. ¿Existe relación entre dos variables cualitativas?
4. ¿Existe relación entre dos variables cuantitativas?
5. ¿Qué limitación tienen las herramientas vistas?
6. Práctica en Python

¿Qué más podemos inferir de la muestra?

¿Qué más podemos inferir de la muestra?

- En la sesión anterior vimos cómo hacemos inferencia sobre la media o la proporción poblacional a partir de una muestra.
- Podemos usar un estimador puntual que sabemos que, aunque no será exactamente igual al valor poblacional, será cercano.
- Por el teorema del límite central sabemos que la distribución muestral de la media se aproxima a una distribución normal, donde el centro de la distribución es la media poblacional.
- Sabemos también que a medida que aumenta el tamaño de muestra, la estimación de la media será más precisa.

¿Qué más podemos inferir de la muestra?

- Sin embargo, las preguntas de investigación suelen ir más lejos que esto.
- Por ejemplo:
  - ¿Existe diferencia en el puntaje global del Saber 11 entre estudiantes de colegios públicos y privados?
  - ¿Existe diferencia en los goles anotados en un partido entre locales y visitantes?
  - ¿El nivel de ingresos laborales está relacionado con la probabilidad de pago de un crédito?
  - ¿El peso está relacionado con el riesgo de padecer diabetes?
- ¿Qué se está buscando en estas preguntas?

¿Existe relación entre una variable cualitativa y una cuantitativa?

¿Existe relación entre una variable cualitativa y una cuantitativa?

- Ejemplos de preguntas de investigación:
  - ¿Existe diferencia en el puntaje global del Saber 11 entre colegios públicos y privados?
  - ¿Existe diferencia en los goles anotados en un partido entre locales y visitantes?
  - ¿Existe diferencia entre la producción total de un cultivo entre usar o no un fertilizante?
  - ¿Existe diferencia en el peso de un recién nacido entre madres fumadores y no fumadoras?

- ¿Cómo podríamos entonces abordar estas preguntas?:
  - La variable cualitativa tiene dos categorías: **Prueba t** (o sus variantes)
  - La variable cualitativa tiene dos o más categorías: **ANOVA** (o sus variantes)

¿Existe relación entre una variable cualitativa y una cuantitativa?:

Validando los supuestos

¿Existe relación entre una variable cualitativa y una cuantitativa?

Supuestos:

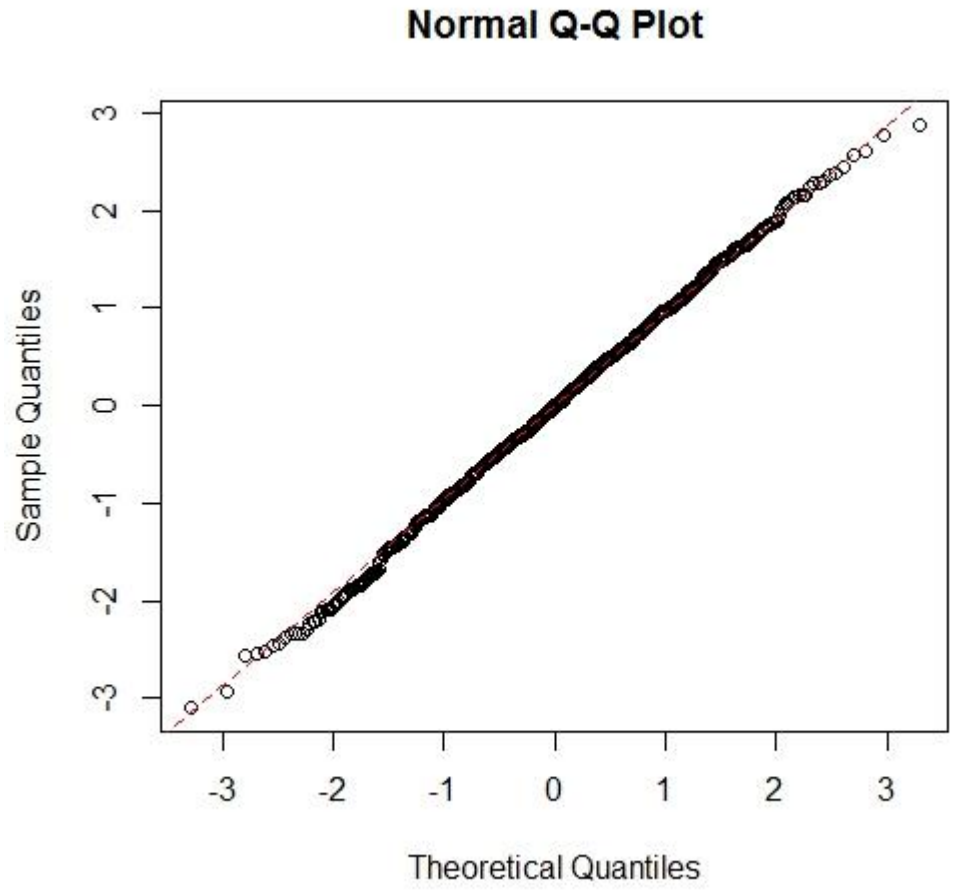
1. **Normalidad**: Las poblaciones siguen una distribución normal (también funciona si son lo suficientemente grandes para que se cumpla el TLC, suele usarse n>30).

2. **Homogeneidad de varianza**: Las varianzas de las poblaciones son iguales.

Normalidad: En la práctica existen algunas herramientas que reflejan síntomas de tener o no normalidad.

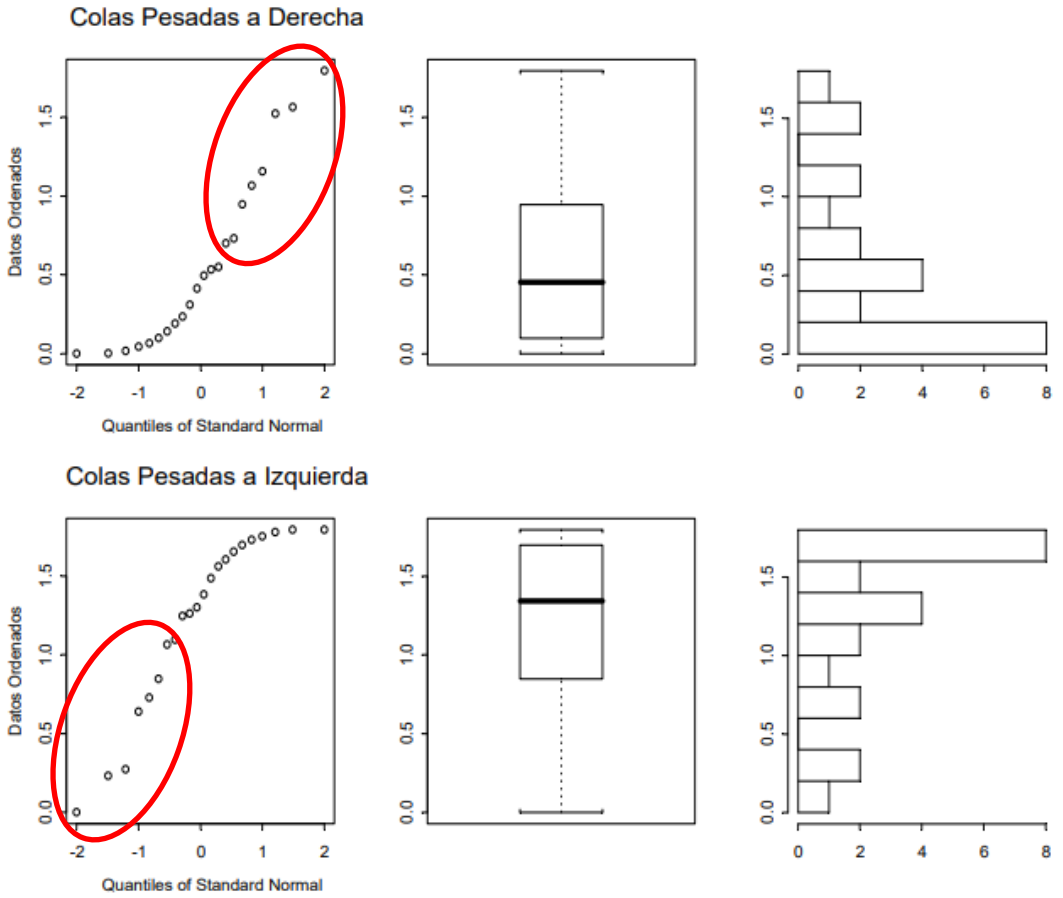
1. Qué tan cerca están la media y la mediana.
2. Coeficiente de asimetría y curtosis.
3. Graficar diagrama de cajas y el histograma
4. Graficar un Q-Q (Cuantil-cuantil)

- Antes de continuar, hablemos del gráfico Q-Q.
- Esta herramienta permite comparar los cuantiles muestrales vs los cuantiles teóricos de una distribución normal.
- Si tenemos una distribución normal, observaremos que los puntos están encima de la línea recta (cuantiles muestrales = cuantiles teóricos) y no tiene puntos dispersos en los extremos.



Pruebas de hipótesis para normalidad

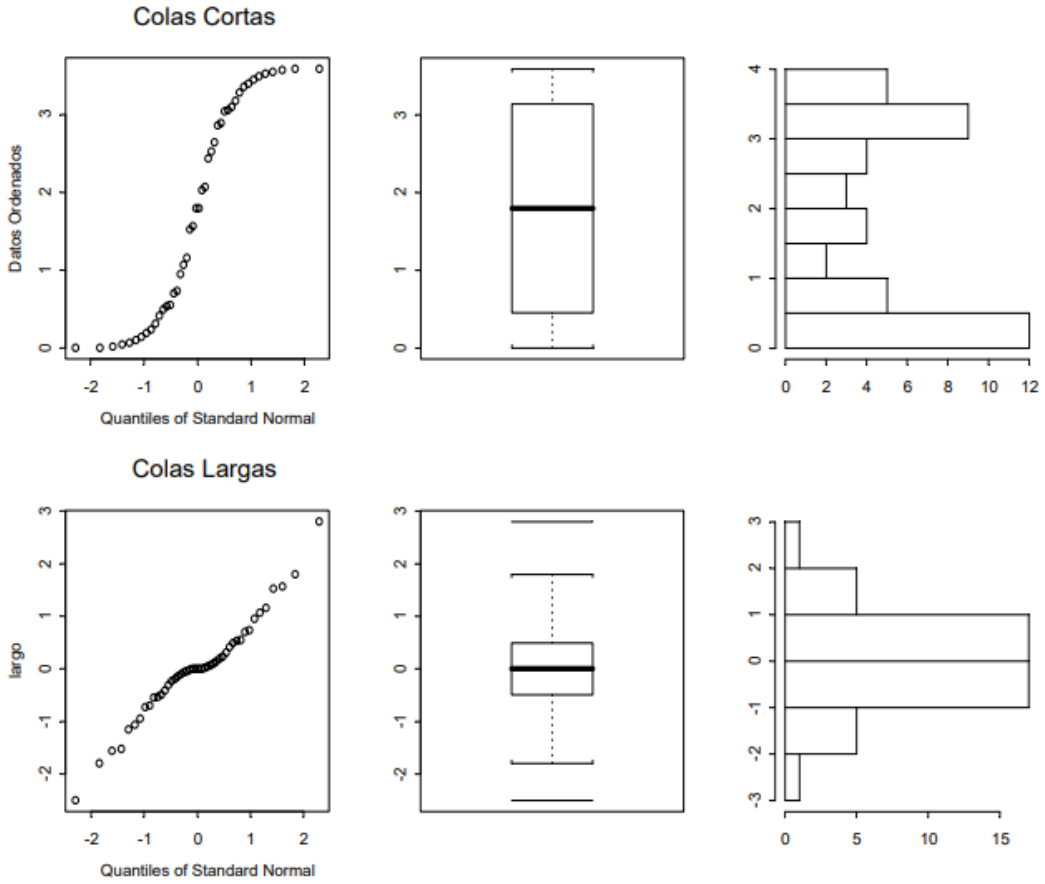
- Miremos algunos ejemplos de síntomas para creer que no se cumple la normalidad.



Tomado de:
http://www.dm.uba.ar/materias/analisis_expl_y_conf_de_datos_de_exp_de_marrays_Mae/2006/1/teoricas/Teor4.pdf

Pruebas de hipótesis para normalidad

- Miremos algunos ejemplo de síntomas para creer que no se cumple la normalidad.



Tomado de:
http://www.dm.uba.ar/materias/analisis_expl_y_conf_de_datos_de_exp_de_marrays_Mae/2006/1/teoricas/Teor4.pdf

Pruebas de hipótesis para normalidad

- Para estar más seguros, se emplean pruebas de hipótesis de normalidad, como la Shapiro-Wilks y Anderson-Darling.

- En ambas pruebas, las hipótesis son:

Ho: La muestra proviene de una población que sigue una distribución normal

Ha: La muestra NO proviene de una población que sigue una distribución normal

- Cada prueba tiene su estadístico particular, el cual se compara contra un valor crítico. También es posible tomar la decisión a partir del valor-p.

- En la práctica en Python veremos cómo emplear cada una de ellas y cómo tomar la decisión.

Comparación de varianzas

- **Homogeneidad de varianzas:** La pregunta clave es saber si la varianza de una población es igual o no a la de otra población, a través de la información de las muestras.

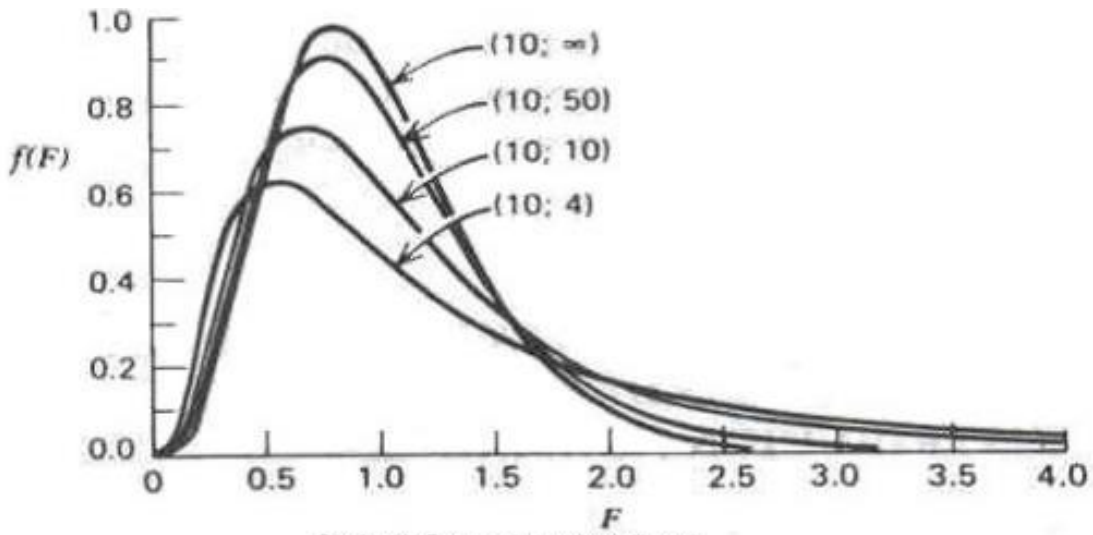
- ¿Qué pruebas puedo realizar?
  - La variable cualitativa tiene dos categorías: **Prueba F o de Levene**
  - La variable cualitativa tiene dos o más categorías: **Prueba de Levene**

- Pero antes de entrar en el detalle de la prueba, es pertinente hablar de la distribución F, ya que es un insumo importante para desarrollarla.

Comparación de varianzas

Distribución F:

- Es una distribución continua.
- No toma valores negativos.
- Tiene sesgo positivo.
- Es asintótica.
- La forma de la distribución se determina por dos parámetros: los grados de libertad en el numerador y los grados de libertad en el denominador. $F_{gln,gld}$



Distribución F para varios grados de libertad.

Comparación de varianzas

Prueba F (dos poblaciones):

- Hipótesis:

Ho: la varianza en ambas poblaciones es igual

Ha: la varianza difiere entre poblaciones

- Estadístico:

$$F = \frac{s_1^2}{s_2^2}$$

- Valor crítico:

$$F_{\alpha/2;n_1-1;n_2-1}$$

- Decisión:

Si $F = \frac{s_1^2}{s_2^2} > F_{\alpha/2;n_1-1;n_2-1}$ (o el valor_p < α), rechazo

Prueba de Levene (dos o más poblaciones):

- Hipótesis:

Ho: la varianza en todas las poblaciones es igual

Ha: la varianza de al menos una población es diferente

- Estadístico:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2}$$

- Valor crítico:

$$F_{\alpha;k-1;N-k}$$

- Decisión:

Si $W > F_{\alpha;k-1;N-k}$ (o el valor_p < α), rechazo

Supuestos de normalidad en las pruebas de comparación de varianza

- El supuesto de normalidad es muy importante en la **prueba F**. Si no se demuestra normalidad, no se puede utilizar (apelar al TLC, no es opción aquí).

- Por su parte, la **Prueba de Levene**, de la forma clásica (media como centro), requiere normalidad pero es robusto si la muestra es lo suficientemente grande para que el TLC se pueda asumir).

- La Prueba de Levene utilizando como centro la mediana, también conocida como la **prueba de Brown-Forsythe**, funciona bien en distribuciones no normales.

¿Existe relación entre una variable cualitativa y una cuantitativa?:

Comparación de medias para dos poblaciones

Comparación de varianzas:
Hay normalidad: Prueba F o de Levene
Aplica el TLC: Prueba de Levene

Comparación de medias de dos poblaciones

- Si hay normalidad (o aplica el TLC) y varianzas iguales
  - Prueba t conjunta
- Si hay normalidad (o aplica el TLC) y NO varianzas iguales
  - Prueba t con varianzas desiguales
- No hay normalidad, ni aplica el TLC
  - Prueba Mann-Whitney-Wilcoxcon

Supuestos:

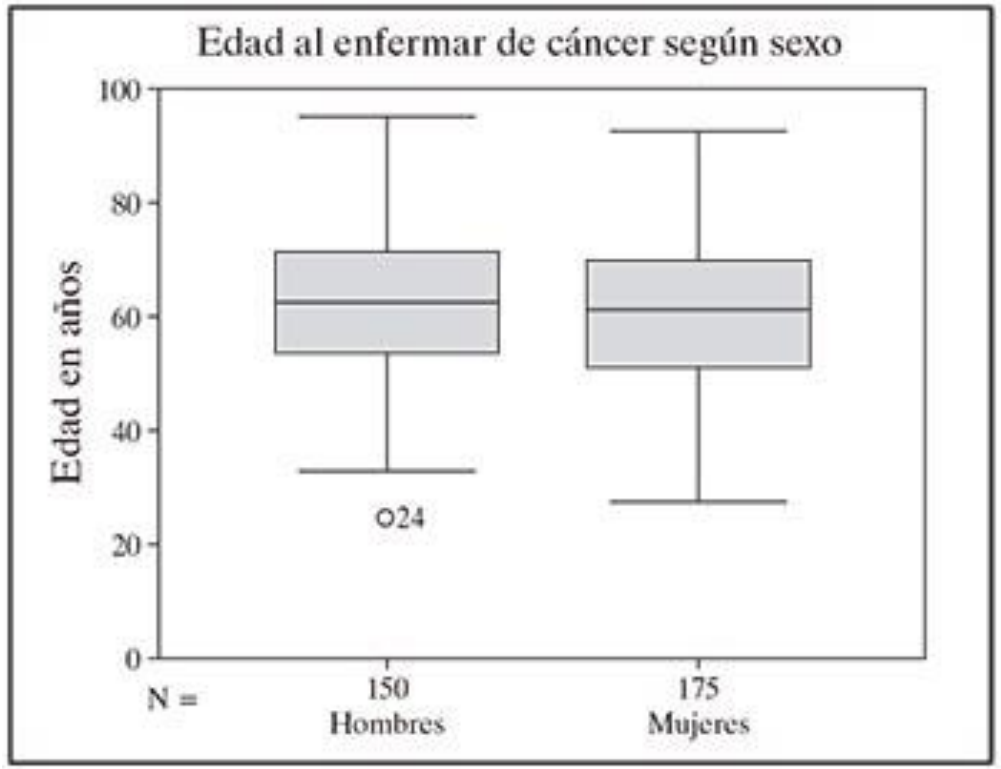
1. Las poblaciones siguen una distribución normal (también funciona si son lo suficientemente grandes para que se cumpla el TLC).

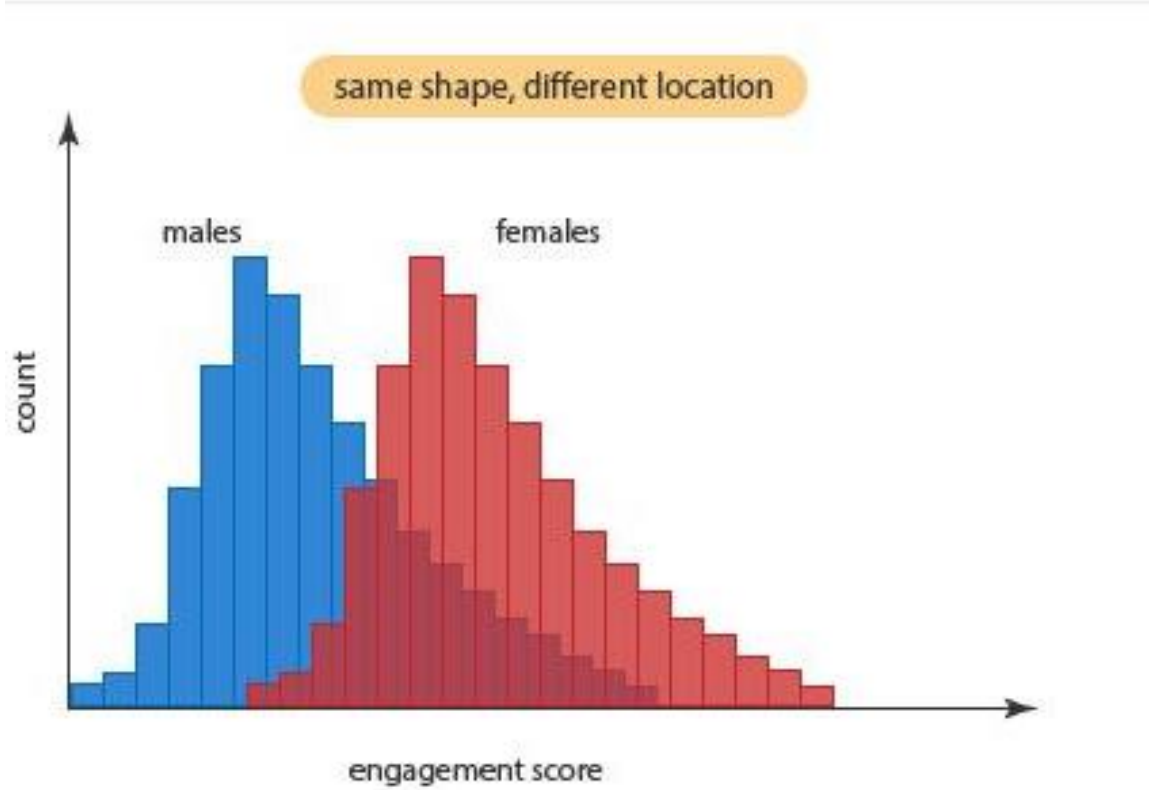
2. Las poblaciones muestreadas son independientes.

3. Las varianzas de las poblaciones son iguales (existe una variación de la prueba, cuando esto no se cumple).

La herramienta gráfica por excelencia es el diagrama de cajas comparativo. Aunque también se usan histogramas.

Comparación de medias con dos poblaciones





Comparación de medias con dos poblaciones

Prueba t conjunta (**homogeneidad de varianzas**):

- Hipótesis:

  Ho: la media en ambas poblaciones es igual
  Ha: la media difiere entre poblaciones

- Estadístico:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ donde s_p^2 es la varianza conjunta

- Valor crítico:

$$t_{\alpha/2;n_1+n_2-2}$$

- Decisión:

Si $|t| > t_{\alpha/2;n_1+n_2-2}$ (o el valor_p < α), rechazo

Comparación de medias con dos poblaciones

Prueba t (**varianzas desiguales**):

- Hipótesis:

  Ho: la media en ambas poblaciones es igual
  Ha: la media difiere entre poblaciones

- Estadístico:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Valor crítico:

  $t_{\alpha/2;gla}$ donde los grados de libertad se calculan a través de una fórmula compleja

- Decisión:

Si $|t| > t_{\alpha/2;gla}$ (o el valor_p < α), rechazo

¿Qué hacer cuando no hay normalidad y no es posible apelar al TLC?

Se utilizan pruebas no paramétricas como la de Mann-Whitney-Wilcoxon.

- Hipótesis:

  Ho: La distribución de ambas poblaciones es la misma

  Ha: La distribución de las poblaciones es diferente

- El estadístico sigue una distribución normal, por lo que se contrasta contra un valor crítico $Z_{\alpha/2}$.

Comparación de varianzas:
Hay normalidad: Prueba F o de Levene
Aplica el TLC: Prueba de Levene

Comparación de medias de dos poblaciones

- Si hay normalidad (o aplica el TLC) y varianzas iguales
  - Prueba t conjunta
- Si hay normalidad (o aplica el TLC) y NO varianzas iguales
  - Prueba t con varianzas desiguales
- No hay normalidad, ni aplica el TLC
  - Prueba Mann-Whitney-Wilcoxcon

¿Existe relación entre una variable cualitativa y una cuantitativa?:

Comparación de medias para dos o más poblaciones

Comparación de varianzas: Prueba de Levene

Comparación de medias de dos o más poblaciones

Si hay normalidad (o aplica el TLC) y varianzas iguales

Prueba Anova (si rechazo, uso la prueba de Tukey)

Si hay normalidad (o aplica el TLC) y NO varianzas iguales

Prueba Anova de Welch (si rechazo, uso la prueba de Games-Howell)

No hay normalidad, ni aplica el TLC

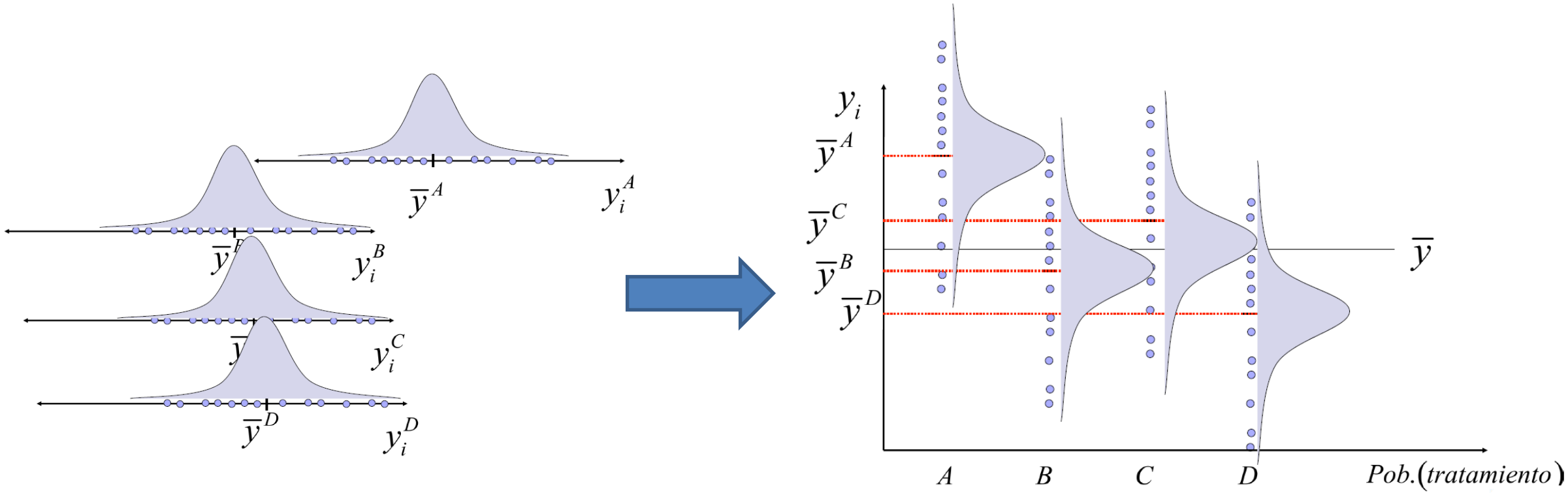
Prueba Kruskal-Wallis (si rechazo, uso la prueba de comparaciones por parejas de Wilcoxon)

Comparación de medias con más de dos poblaciones

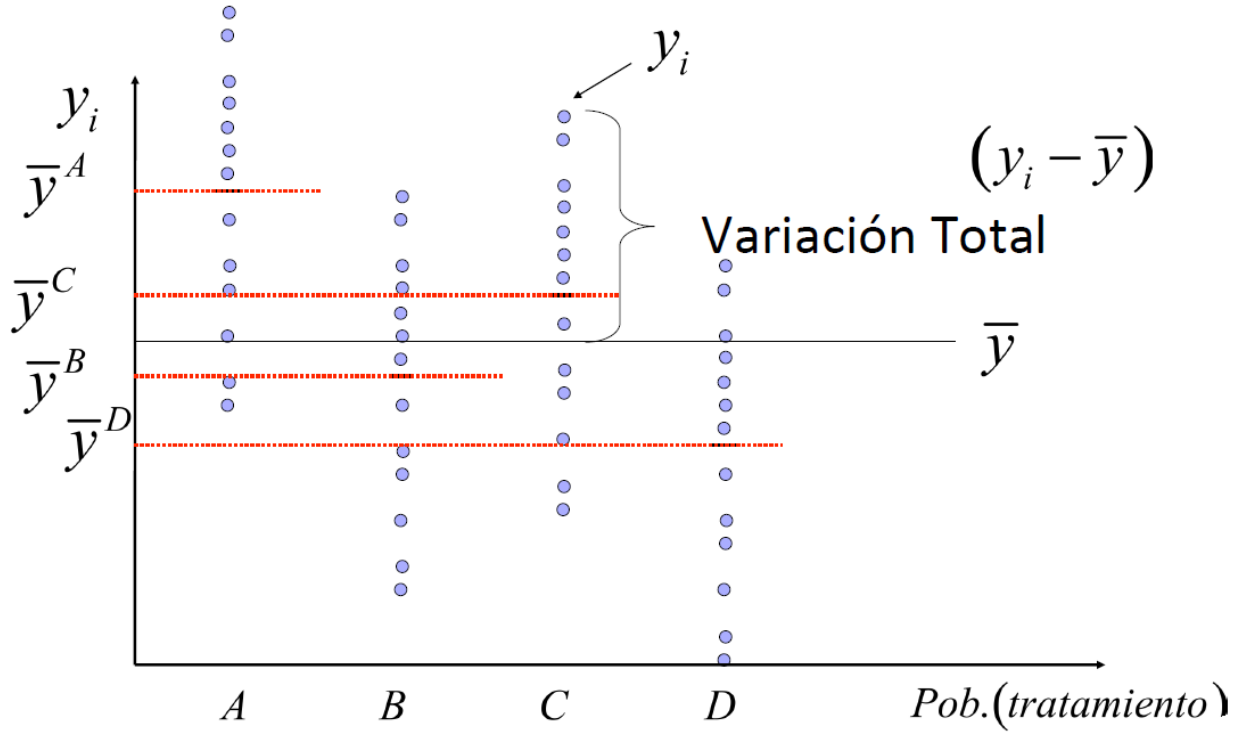
- La herramienta principal para este tipo de análisis es la ANOVA.
- Su nombre proviene de la abreviación en inglés de Analysis of Variance.
- Supuestos:

1. Las poblaciones siguen una distribución normal, aunque también funciona si son lo suficientemente grandes para que se cumpla el TLC.
2. Las poblaciones muestreadas son independientes.
3. Las varianzas de las poblaciones son iguales (existe una variación de la prueba, cuando esto no se cumple).

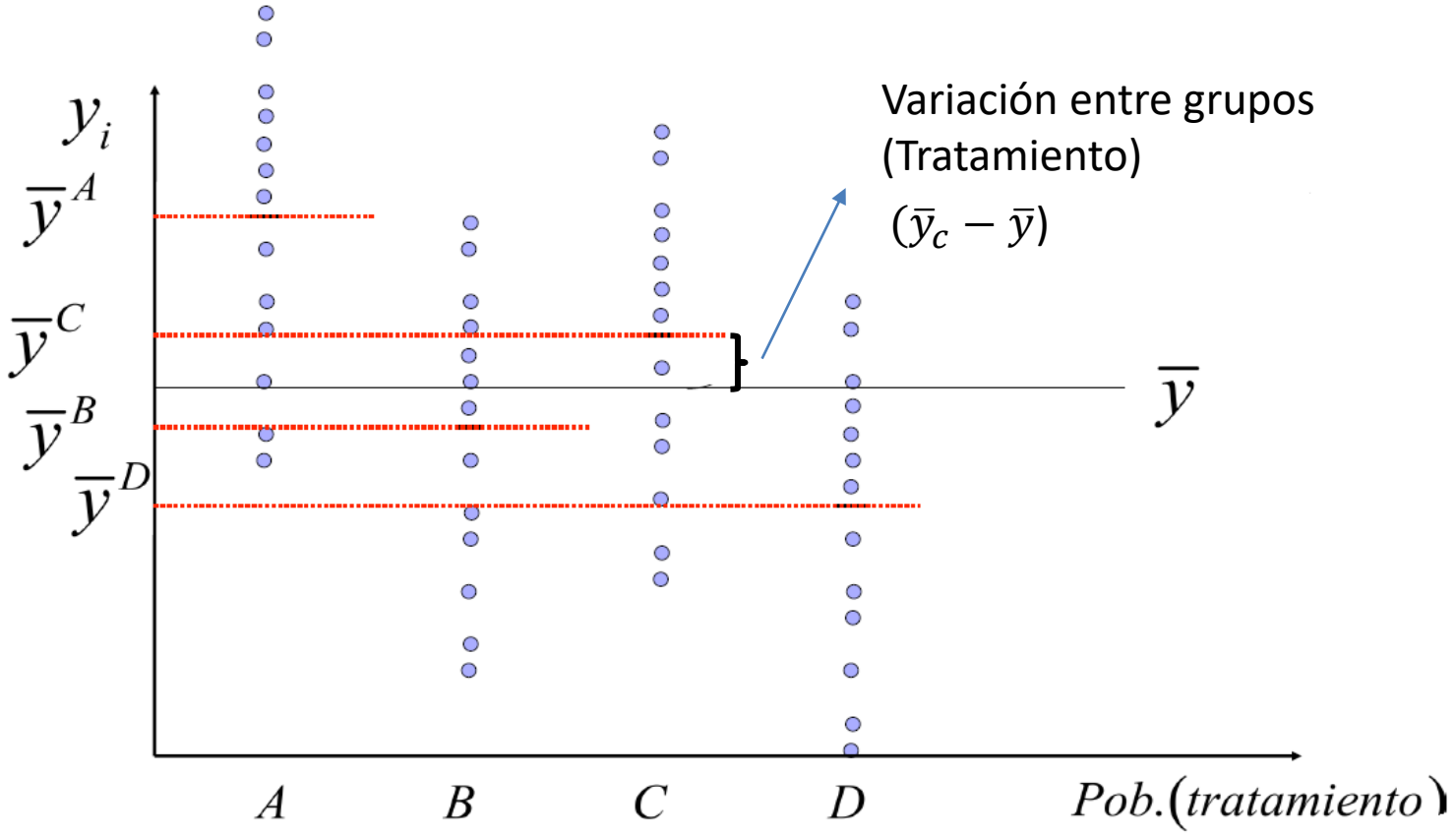
Intuitivamente, ¿qué hace la prueba?



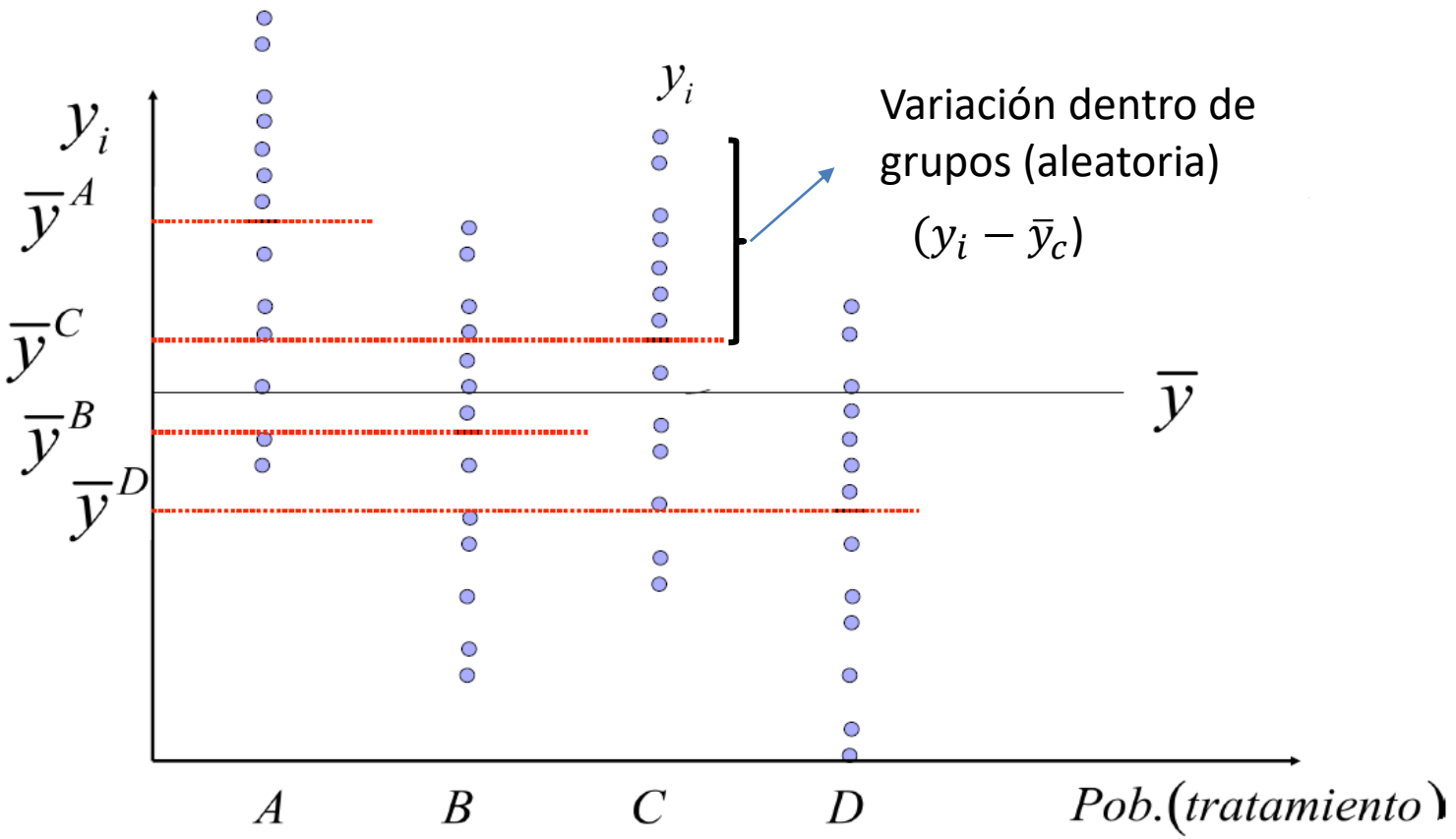
Intuitivamente, ¿qué hace la prueba?



Intuitivamente, ¿qué hace la prueba?



Intuitivamente, ¿qué hace la prueba?



En tal sentido:

$$(y_i - \bar{y}) = (y_i - \bar{y}_c) + (\bar{y}_c - \bar{y})$$

Intuitivamente, ¿qué hace la prueba?

$$SSTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SST = \sum_{j=1}^{k} (\bar{y} - \bar{y}_j)^2 \qquad k = \#\ \text{tratamientos}$$

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y}_{k \in i})^2$$

Comparación de medias con más de dos poblaciones

- El procedimiento funciona comparando la varianza entre los grupos versus la varianza dentro de los grupos, para determinar si los grupos son más distintos entre sí que dentro de sí.

- Hipótesis:

  Ho: la media en todas las poblaciones es igual

  Ha: la media de al menos una población es diferente

Tabla ANOVA

Fuente de Variación	SS	G de L	MS
Tratatamiento (Entre Grupos)	$\sum_{j=1}^{k}(\bar{y}-\bar{y}_k)^2$	$k-1$	$MST=\frac{SST}{k-1}$
Errores (Dentro de los grupos)	$\sum_{i=1}^{n}(y_i-\bar{y}_{k\in i})^2$	$n-k$	$MSE=\frac{SSE}{n-k}$
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	

Comparación de medias con dos poblaciones

- Estadístico:

$$F = \frac{MST}{MSE}$$

- Valor crítico:

$$F_{\alpha;k-1;n-k}$$

- Decisión:

Si $F > F_{\alpha;k-1;n-k}$ (o el valor_p < α), rechazo

- ¿Qué hacemos si encontramos que se rechaza la hipótesis nula?:

Prueba HSD de Tukey

¿Qué hacer cuando encontramos que no hay varianzas iguales?

Se utiliza la Prueba Anova de Welch.

- Hipótesis:

  Ho: la media en todas las poblaciones es igual

  Ha: la media de al menos una población es diferente

- ¿Qué hacemos si encontramos que se rechaza la hipótesis nula?:

  Prueba de Games-Howell

Comparación de medias con dos poblaciones

¿Qué hacer cuando encontramos que no hay normalidad?

Podemos usar pruebas no paramétricas, como la de Kruskal-Wallis.

- Hipótesis:

  Ho: La distribución de las poblaciones es la misma
  Ha: La distribución de al menos una población es diferente

- El estadístico sigue una distribución chi-cuadrado.

- ¿Qué hacemos si encontramos que se rechaza la hipótesis nula?

  Prueba de Wilcoxon para comparaciones por parejas.

Comparación de varianzas: Prueba de Levene

Comparación de medias de dos o más poblaciones

Si hay normalidad (o aplica el TLC) y varianzas iguales

Prueba Anova (si rechazo, uso la prueba de Tukey)

Si hay normalidad (o aplica el TLC) y NO varianzas iguales

Prueba Anova de Welch (si rechazo, uso la prueba de Games-Howell)

No hay normalidad, ni aplica el TLC

Prueba Kruskal-Wallis (si rechazo, uso la prueba de comparaciones por parejas de Wilcoxon)

¿Existe relación entre dos variables cualitativas?

¿Qué es una tabla de contingencia?

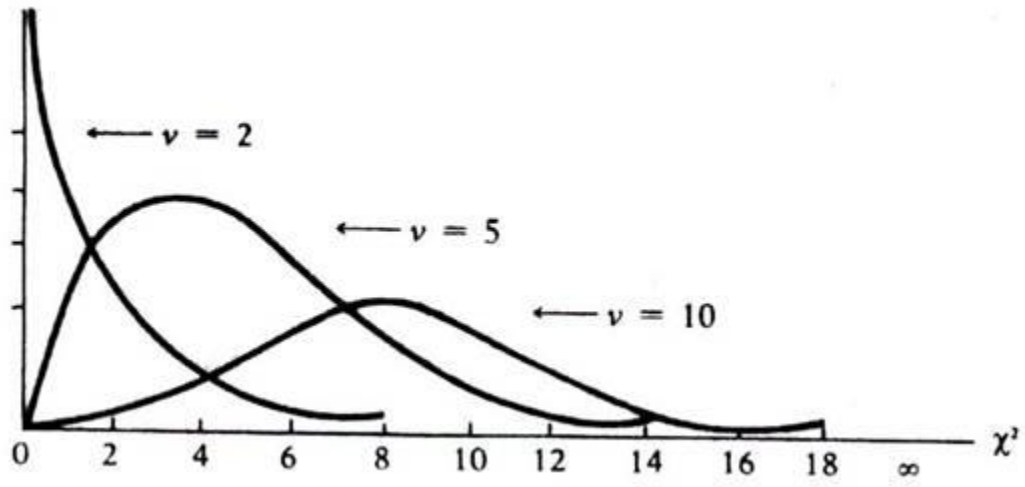
- Tabla utilizada para clasificar observaciones de acuerdo con dos características cualitativas.

Nivel	1	2	3	4	5	6	All
Aprob_NV							
No	31	47	77	33	83	53	324
Sí	110	188	335	239	312	271	1455
All	141	235	412	272	395	324	1779

- En el ejemplo anterior nos gustaría saber si existe relación entre la aprobación y el nivel de inglés.
- Es decir, en estos casos quisiéramos probar si existe relación entre dos variables cualitativas.
- Existe una herramienta muy valiosa llamada la Prueba Chi-Cuadrado de Pearson, la cual permite llevar a cabo un test de independencia.
- Pero primero, miremos la característica de la distribución Chi-Cuadrado.

Distribución Chi-Cuadrado:

- Es una distribución continua.
- No toma valores negativos.
- Tiene sesgo positivo.
- La forma de la distribución se determina por los grados de libertad.



Análisis de tablas de contingencia

Prueba Chi-Cuadrado de independencia:

- Hipótesis:

  Ho: No existe relación entre las variables
  Ha: Existe relación entre las variables

- Estadístico:

$$\chi^2 = \sum [\frac{(f_o - f_e)^2}{f_e}]$$

- Valor crítico:

$$\chi^2{}_{\alpha;(filas-1)(columnas-1)}$$

- Decisión:

Si $\chi^2 > \chi^2{}_{\alpha;(filas-1)(columnas-1)}$ (o el valor_p < α), rechazo

- La frecuencia esperada se calcula:

$$f_e = (total\ de\ columnas) * (\frac{total\ filas}{gran\ total})$$

Ejemplo:

Nivel	1	2	3	4	5	6	All
Aprob_NV							
No	31	47	77	33	83	53	324
Sí	110	188	335	239	312	271	1455
All	141	235	412	272	395	324	1779

$$141 * \left(\frac{324}{1779}\right) = 26$$

Nivel	**1**		**2**		**3**		**4**		**5**		**6**		**Total**	
Aprobación	**fo**	**fe**	**fo**	**fe**	**fo**	**fe**	**fo**	**fe**	**fo**	**fe**	**fo**	**fe**	**fo**	**fe**
No	31	26	47	43	77	75	33	50	83	72	53	59	324	324
Sí	110	115	188	192	335	337	239	222	312	323	271	265	1455	1455
Total	**141**	**141**	**235**	**235**	**412**	**412**	**272**	**272**	**395**	**395**	**324**	**324**	**1779**	**1779**

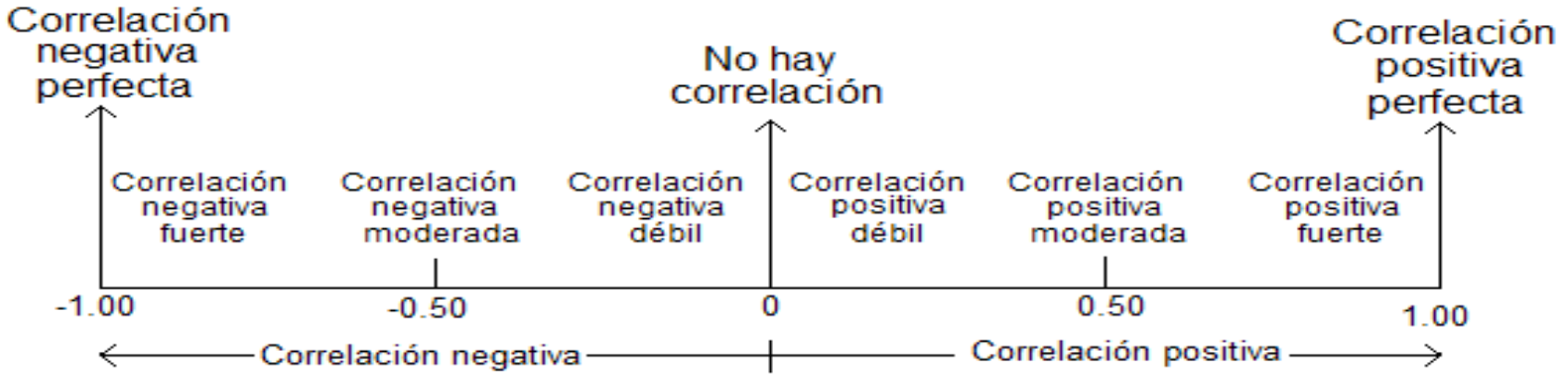
¿Existe relación entre dos variables cuantitativas?

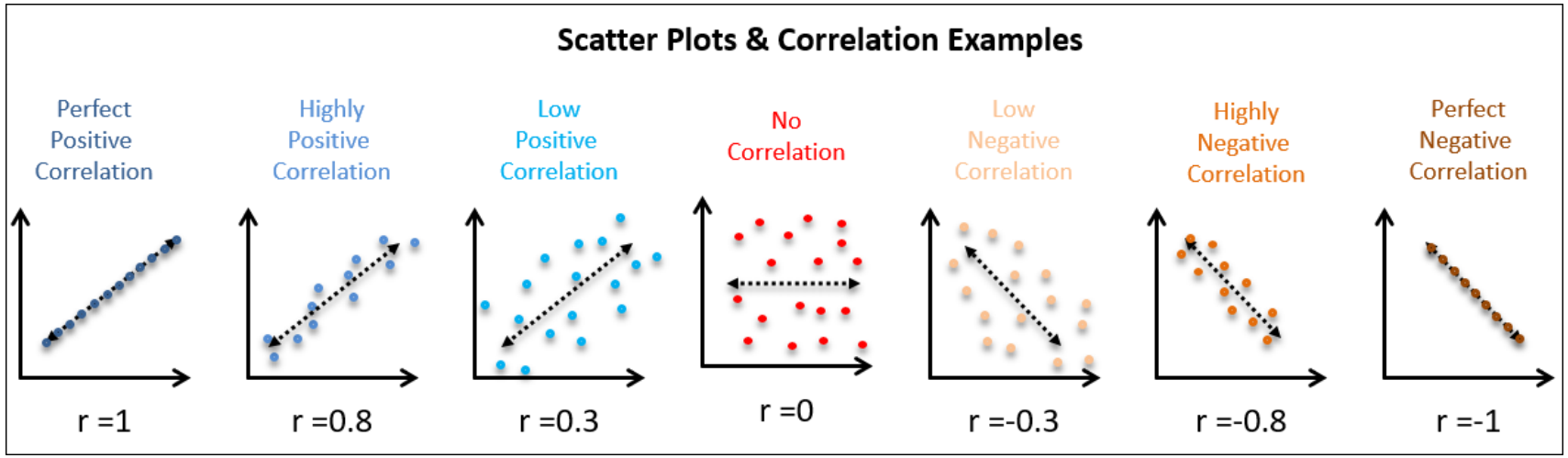
¿Existe relación entre dos variables cuantitativas?

¿Qué es?

- Es una medida que describe la fuerza de la relación o asociación lineal entre dos variables cuantitativas.
- Puede tomar cualquier valor entre -1 y 1.
- Un coeficiente igual 1, implicaría una correlación perfecta en un sentido lineal directo.
- Un coeficiente igual a -1, implicaría una correlación perfecta en un sentido lineal inverso.

- Por su parte, una medida de 0, implicaría que no existe correlación lineal entre las variables.





¿Existe relación entre dos variables cuantitativas?

- ¿Cómo se calcula?

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

- Dado que r es un estadístico, entonces conviene realizar una prueba que permita comprobar que la correlación de la población es estadísticamente diferente de cero.

  Ho: No hay correlación entre las variables ($\rho = 0$)

  Ha: Existe correlación entre las variables ($\rho \neq 0$)

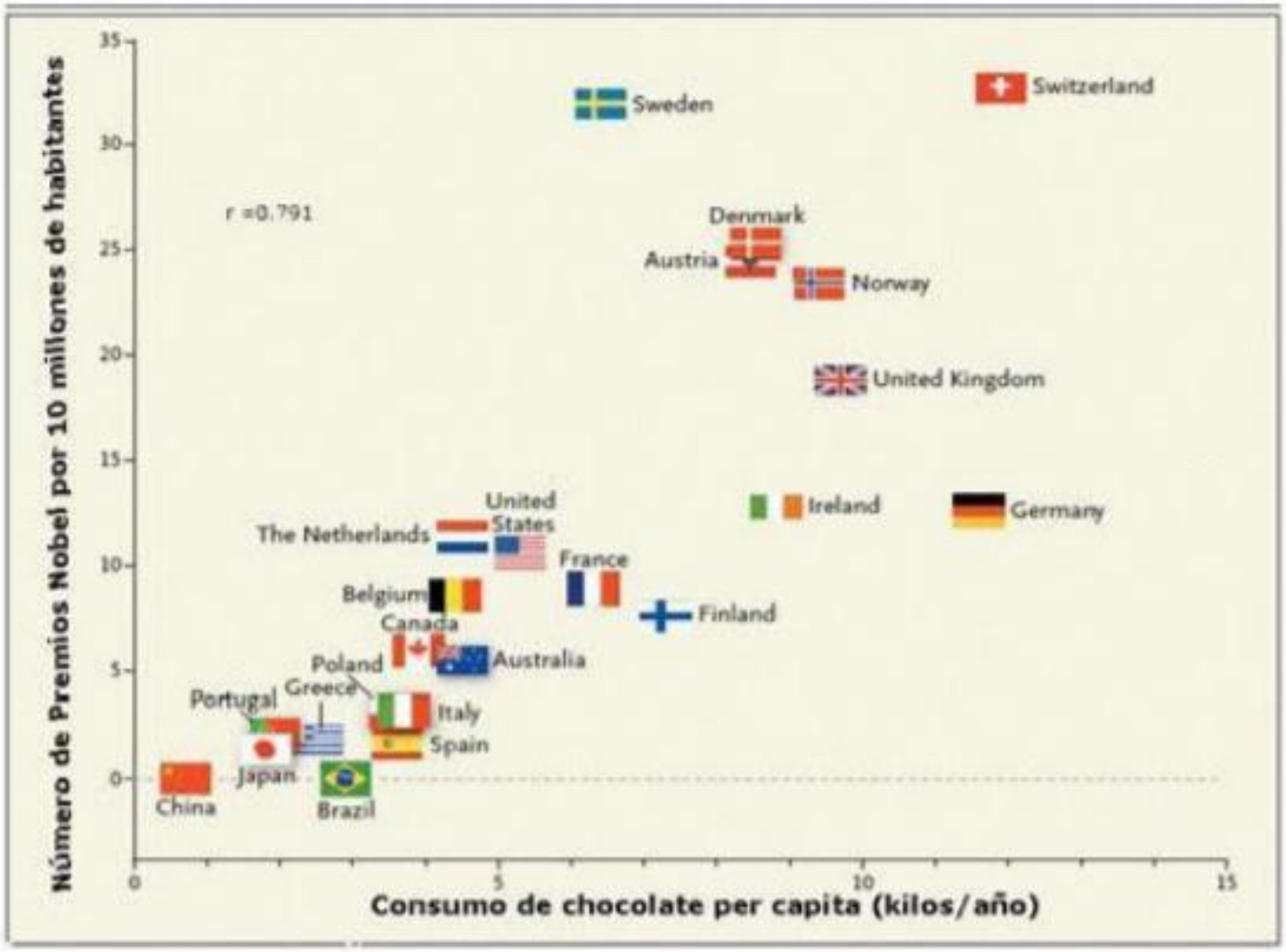
- Estadístico:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ con n-2 grados de libertad

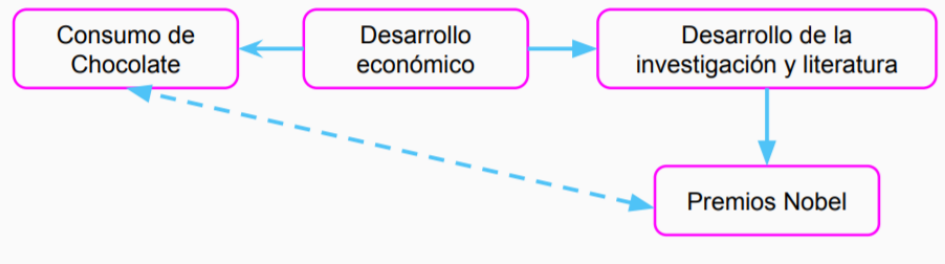
¿Qué es una correlación espuria?

- ¡Correlación no implica causalidad!
- Aunque encontremos que existe correlación fuerte entre dos variables, no podemos asegurar que un cambio en una genera un cambio en la otra.
- En muchas ocasiones se calcula la correlación entre variables que no tienen conexión lógica encontrando correlaciones fuertes. Por tanto, no debe confundirse el hecho de que haya correlación con el que haya causalidad.

¿Existe relación entre dos variables cuantitativas?



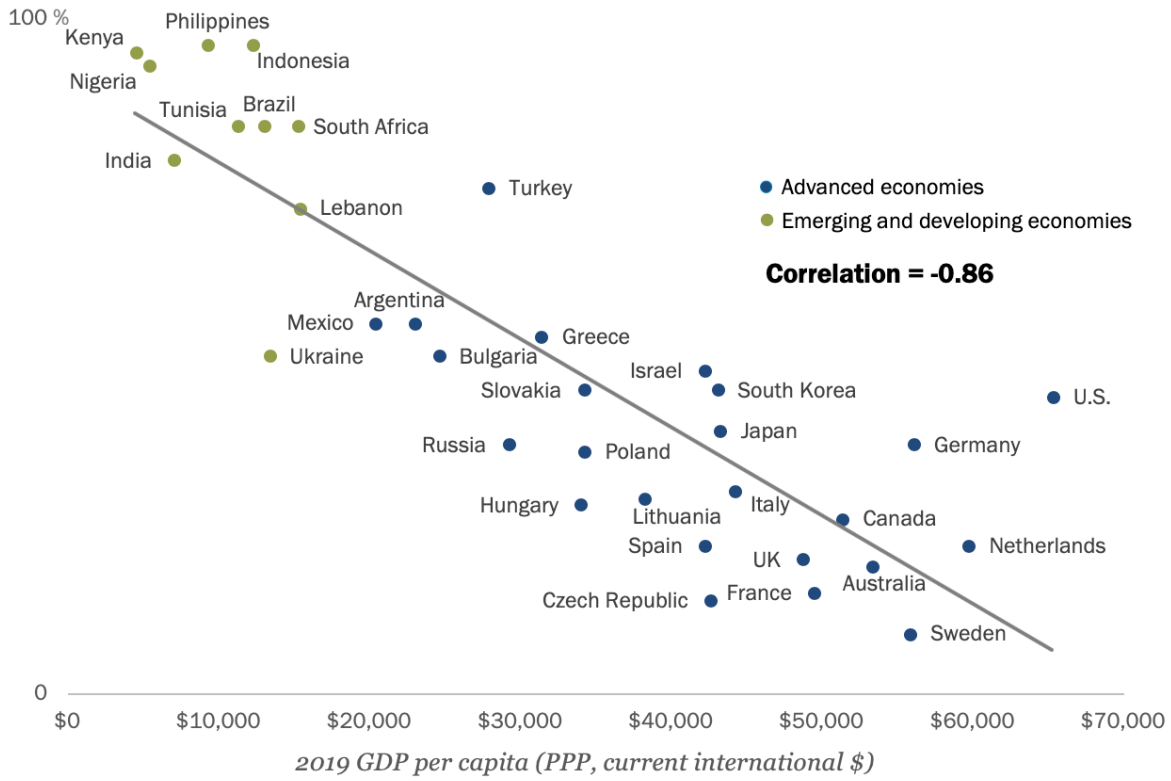
¿Nos falta comer más chocolate para tener premios Nobel?



¿Existe relación entre dos variables cuantitativas?

Countries with higher GDP per capita less likely to tie belief in God to morality

% who say it is necessary to believe in God in order to be moral and have good values

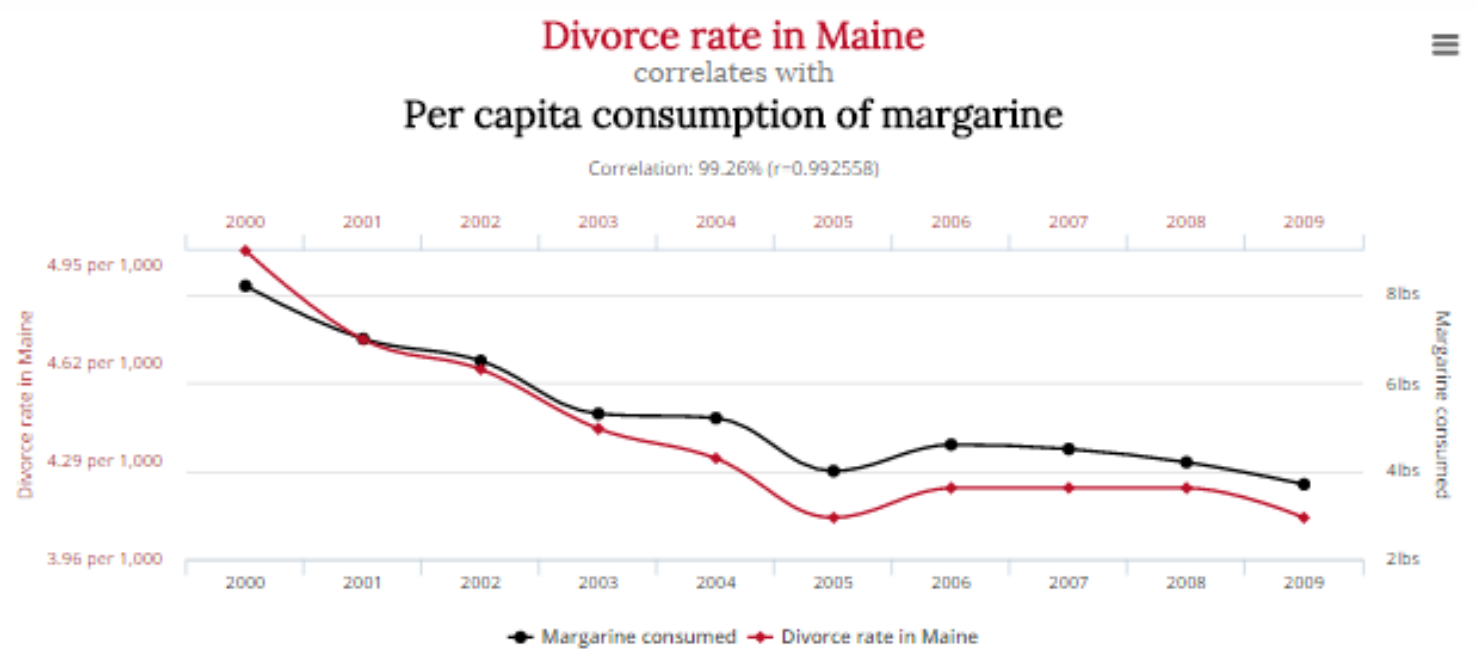


Note: Figures for gross domestic product per capita, measured by purchasing power parity (PPP) in current international $ from the World Development Indicators database, World Bank. Data accessed July 6, 2020. For more details, see Appendix B.
Source: Spring 2019 Global Attitudes Survey. Q30.

PEW RESEARCH CENTER

¿Nos falta ser menos religiosos para progresar?

¿Existe relación entre dos variables cuantitativas?



¿La clave de un matrimonio duradero es comer menos margarina?

¿Qué limitación tienen las herramientas vistas?

¿Qué limitación tienen las herramientas vistas?

- Solo están considerando la posible relación entre dos variables, ignorando todo lo demás a su alrededor.

- Podrían existir otras variables que también tienen influencia sobre la variable dependiente.

- Por tanto, se requieren metodologías que analicen la posible relación, pero controlando (teniendo en cuenta) por la presencia de otras variables.

¿Qué limitación tienen las herramientas vistas?

- Ejemplo:

¿Existe diferencia en el puntaje global del Saber 11 entre estudiantes de colegios públicos y privados?

¿Qué otras variables podrían influir en el puntaje global que obtiene un estudiante?

Género, educación de la madre, estrato, etnia, etc.

Práctica en R