Análisis de regresión

12 de agosto
(semana 1)

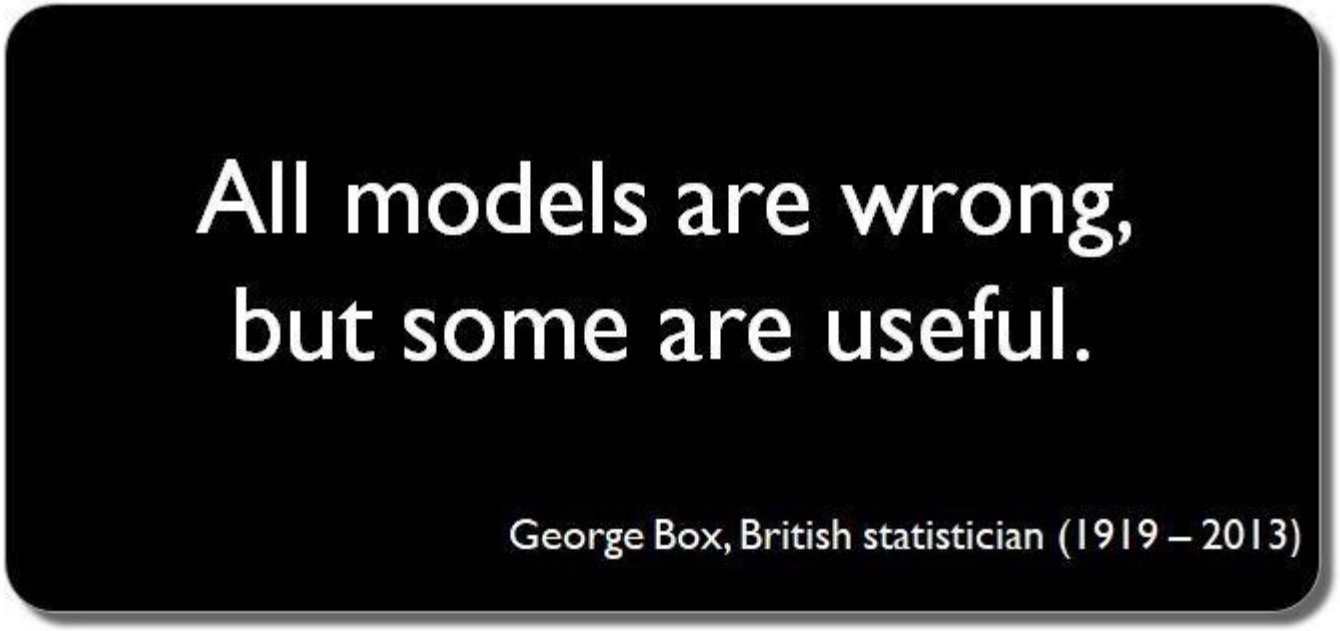
Plan de trabajo

1. Relaciones entre variables
2. Estudio de pares de variables cualitativas
3. Estudio de pares de variables cualitativa-cuantitativa
4. Estudio de pares de variables cuantitativas

Nota:

- Ya está disponible el taller 1.
- Mismo monitor: Jesús David Castro, jecastroa@unal.edu.co. En correos poner asunto: "DUDA REGRESION" al inicio del asunto.
- **No olviden mi horario de atención**: miérc./viernes de 11 a 12:30pm en mi oficina (325-404). Martes 10-12 virtual (con cita previa).
- La primera entrega del trabajo final será para el domingo 4 de septiembre.
- En la semana del 29 de agosto al 2 de septiembre no habrá clase. Pendientes de una actividad a desarrollar.

Motivación



Fuente
https://twitter.com/ithinkwellhugh/status/1283227327628038146

Estudio de las relaciones entre variables

- Seguimos en el contexto de la inferencia frecuentista basada en el modelo.

- Hasta este momento (salvo en muestras pareadas) hemos asumido que las variables de interés para todas las unidades estadísticas forman una muestra aleatoria (**independientes e idénticamente distribuidas**).

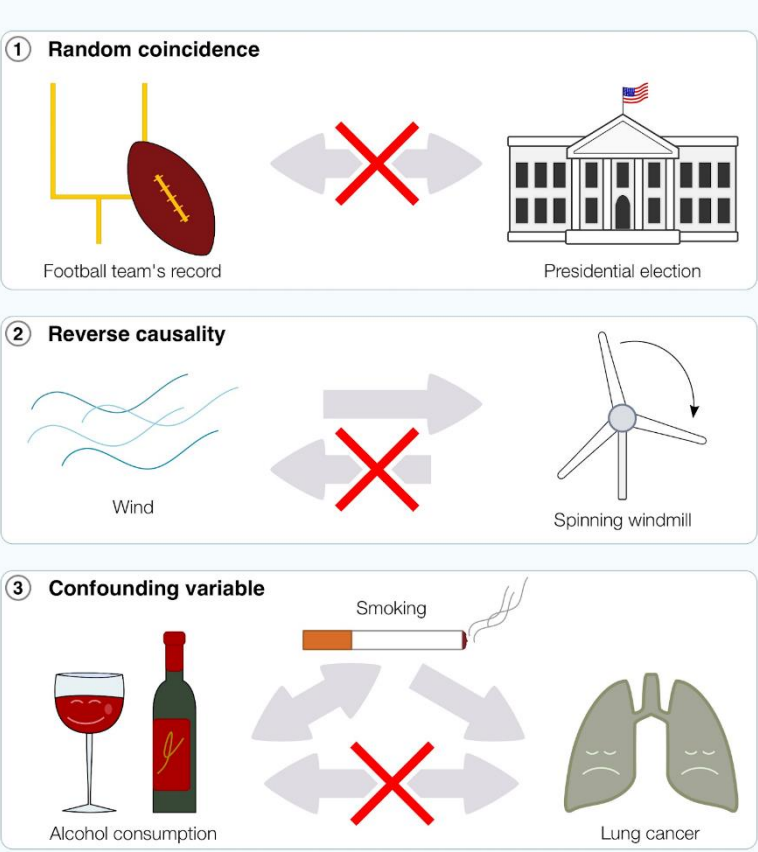
- Sin embargo, cuando se dispone de otras variables, es posible que el segundo supuesto pueda ser relajado y que se pueda construir un modelo que aproveche dicha relación para capturar mejor la dinámica de las variables.

  - Impacto del nivel educativo en el salario.
  - Impacto de la edad en la respuesta a una vacuna.

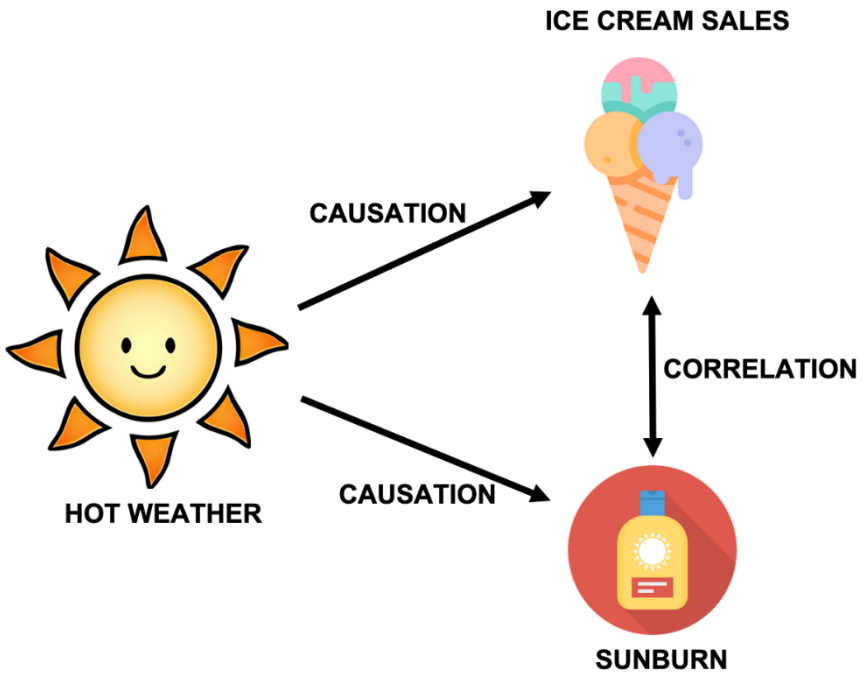
¿Qué tipos de relaciones se pueden evidenciar en los datos?

- **Relación causal:** Cuando se logra determinar que los cambios en una variable inducen cambios en la otra variable siguiendo un modelo causa-efecto.
  - Se requiere de un diseño experimental para determinar relaciones causales (control factores externos).
- **Asociación (correlación o covarianza)*:** Se observa en los datos que, a medida que una de las variables varía, la otra suele hacerlo de una manera "predecible". Este comportamiento no es indicio de causalidad.
  - Al igual que la anterior, debe estar apoyada y soportada por el conocimiento de los expertos.
- **Relación espuria:** Cuando los datos del estudio exhiben relaciones que en la realidad no tienen sentido.

Correlación no implica causalidad

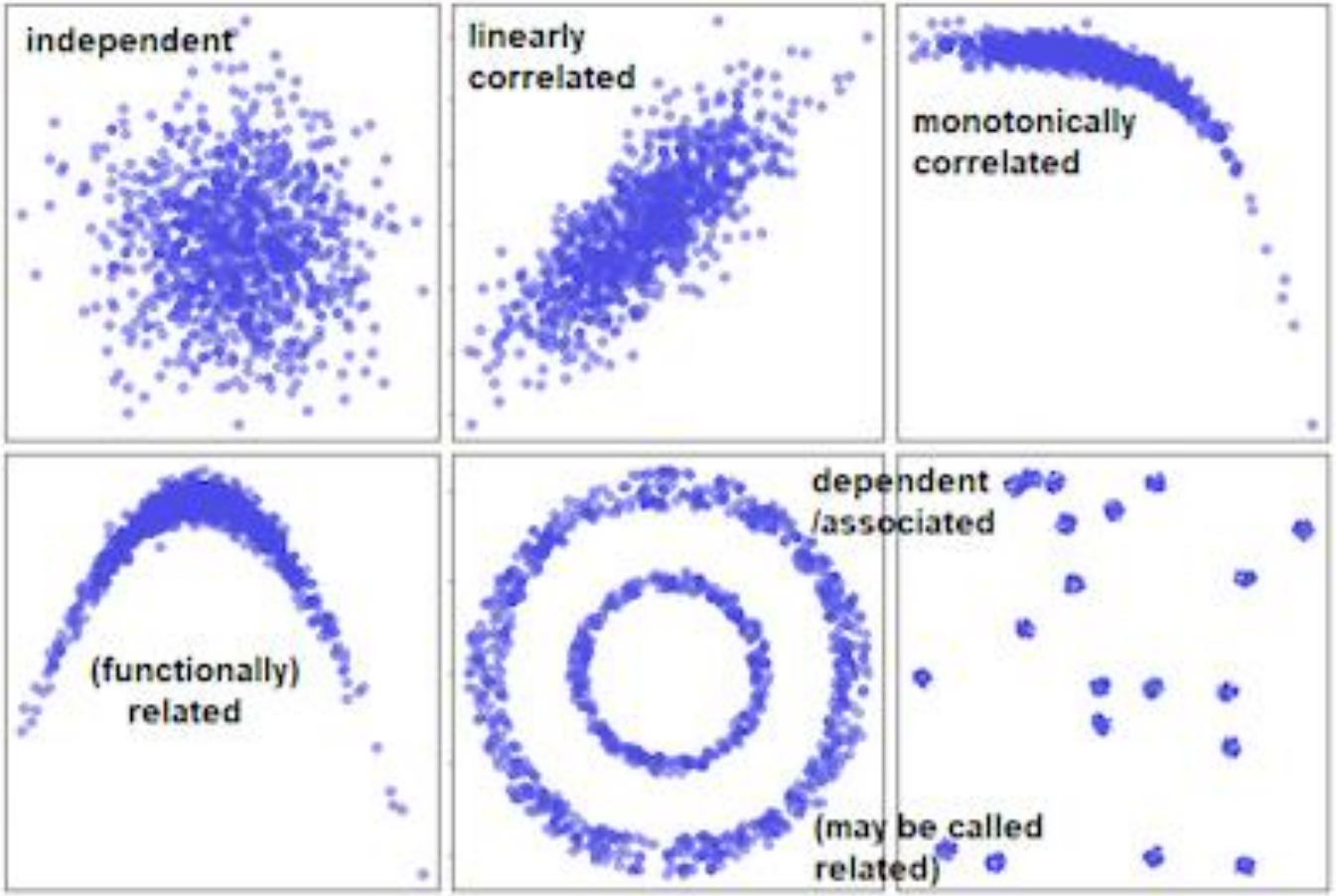


Fuente:
https://sitn.hms.harvard.edu/flash/2021/when-correlation-does-not-imply-causation-why-your-gut-microbes-may-not-yet-be-a-silver-bullet-to-all-your-problems/



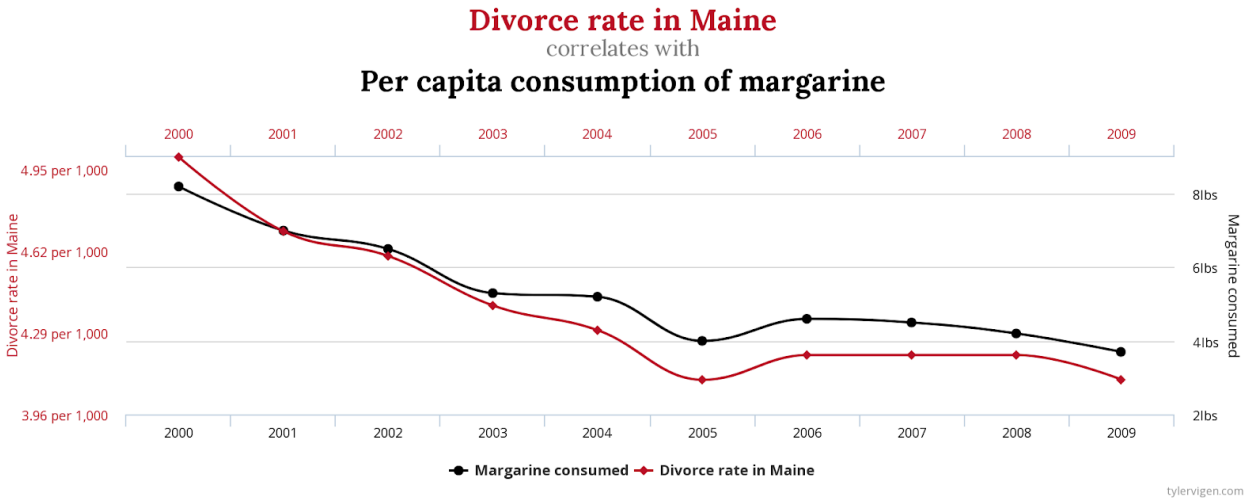
Fuente:
https://www.royriachi.com/2019/02/correlational-and-causal-relationships.html

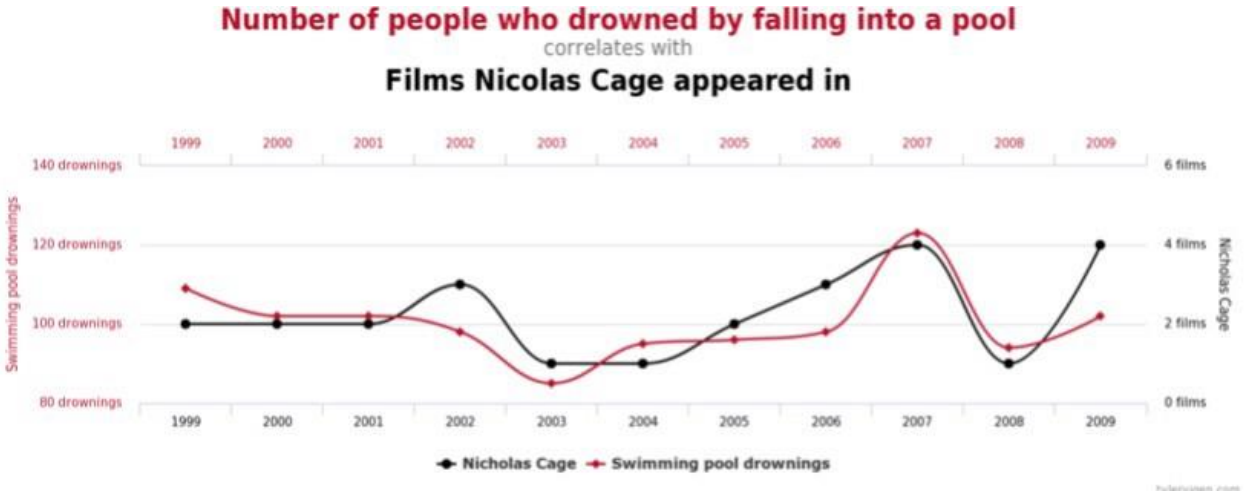
Correlación es simplemente asociación de tipo lineal



Fuente: http://statcalculators.com/the-difference-between-association-and-correlation/

Relaciones espurias





Fuente: https://sitn.hms.harvard.edu/flash/2021/when-correlation-does-not-imply-causation-why-your-gut-microbes-may-not-yet-be-a-silver-bullet-to-all-your-problems/

Preguntas

- *¿Cómo es posible determinar si dos variables exhiben una relación entre ellas?*

		Variable(s) explicativa(s) ***(x)***		
		Cualit. nominal	*Cualit. Ordinal*	*Cuantit.*
Variable de interés ***(y)***	*Cualit. nominal*	-Prueba chi c. de independencia -Coeficiente V de Crámer*	-Prueba chi c. de independencia -Coeficiente V de Crámer*	-ANOVA a una vía (o similares) - Porcentaje de varianza explicado*
	Cualit. Ordinal	-Prueba chi c. de independencia -Coeficiente V de Crámer*	-Prueba chi c. de independencia -Tau de Kendall* -Rho de Spearman*	-ANOVA a una vía (o similares) -Tau de Kendall* -Rho de Spearman*
	Cuantit.	-ANOVA a una vía (o similares) - Porcentaje de varianza explicado*	-ANOVA a una vía (o similares) -Tau de Kendall* -Rho de Spearman*	-Inferencia sobre regresión/corr./asoc. -Correlac. de Pearson* -Tau de Kendall* -Rho de Spearman* -Coef. Xi*

*: Mide qué tan fuerte es la relación.

Preguntas

- *¿Cómo es posible determinar si una variable dependiente (**y**) exhibe relación con dos o más variables independientes?*

- *Si existe una relación, ¿cómo se puede modelar?*
  - *Con fines explicativos del fenómeno*
  - *Con fines predictivos (así no haya relación causal)*

		Variable(s) explicativa(s) ***(x)***		
		Cualit.	*Cualit.+Cuant.*	*Cuantit.*
Variable de interés ***(y)***	*Cualit.*	(1) Modelos lineales generalizados (2) Modelos de aprendizaje de máquina		
	Cuantit.	(1) Modelos lineales (regresión paramétrica) (2) Modelos semiparamétricos (3) Modelos lineales generalizados (4) Modelos de aprendizaje de máquina		

Distribuciones muestrales

Distribución Normal o Gaussiana

Descripción: La variable X puede tomar cualquier número real. Es la distribución más comúnmente usada (y abusada).

Posibles valores (soporte): (-∞,∞)

Notación: $X \sim N\left(\mu, \sigma^2\right)$

Parámetros:

$\mu \in \mathbb{R}$: Valor esperado

$\sigma^2 \in \mathbb{R}^+$: Varianza

Función de densidad:

$$f_X\left(x\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(x-\mu\right)^2}{2\sigma^2}}$$

Función de distribución: No tiene expresión analítica

Valor esperado: $E\left[X\right] = \mu$

Varianza: $Var\left[X\right] = \sigma^2$

Función g. mom. y caract.:

$$m_X\left(t\right) = \exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right)$$

$$\phi_X\left(t\right) = \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right)$$

Código R. Para la f. d para *d*: **dnorm(x=*d*, mean= μ, sd= σ)**

Para la f. d. p. para *d*: **pnorm(q=*d*, mean= μ, sd= σ)**

Para el percentil *alpha*: **qnorm(p=*alpha*, mean= μ, sd= σ)**

Para generar *m* números aleatorios: **rnorm(n=*m*, mean= μ, sd= σ)**

Distribución normal multivariada

Descripción: La definición general es que cualquier combinación lineal tiene distribución normal (así sea degenerada). Si la matriz de varianzas y covarianzas es invertible, se tienen estas propiedades.

Notación: $\mathbf{X} = \left(X_1, X_2, \ldots, X_n \right)^T \sim MVN\left(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}} \right)$

Posibles valores (soporte): $\mathbb{R}^n$

Función g. mom. y caract.:

$$m_{\mathbf{X}}\left(\mathbf{t}\right) = \exp\left(\mathbf{t}^T \boldsymbol{\mu}_{\mathbf{X}} + \tfrac{1}{2} \mathbf{t}^T \Sigma_{\mathbf{X}} \mathbf{t} \right)$$

$$\phi_{\mathbf{X}}\left(\mathbf{t}\right) = \exp\left(i\mathbf{t}^T \boldsymbol{\mu}_{\mathbf{X}} - \tfrac{1}{2} \mathbf{t}^T \Sigma_{\mathbf{X}} \mathbf{t} \right)$$

Parámetros:

$\boldsymbol{\mu}_{\mathbf{X}} \in \mathbb{R}^n$: Vector de valores esperados.

$0 \prec \Sigma_{\mathbf{X}} \in \mathbb{R}^{n \times n}$: Matriz de varianzas y covarianzas

Función de densidad de probabilidad:

$$f_{\mathbf{X}}\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{n/2}} \left|\Sigma_{\mathbf{X}}\right|^{-1/2} \exp\left[\frac{-1}{2} \left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} \right)^T \Sigma_{\mathbf{X}}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} \right) \right]$$

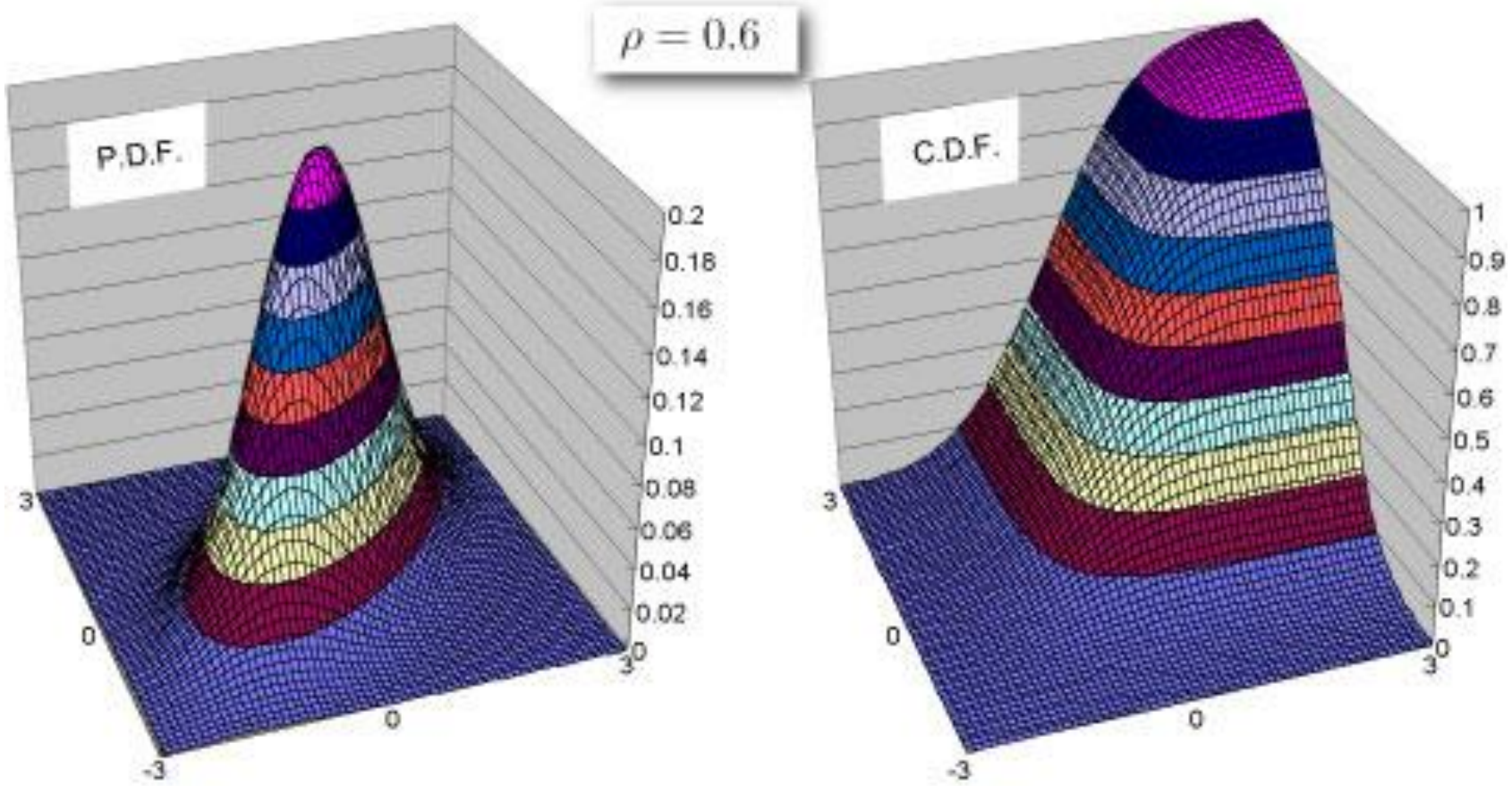
Función de distribución: (No fórmula general)

Valor esperado: $E\left[\mathbf{X}\right] = \boldsymbol{\mu}_{\mathbf{X}} = \left(\mu_1, \mu_2, \ldots, \mu_n \right)^T$

Varianza: $Var\left[X_i \right] = \left(\Sigma_{\mathbf{X}} \right)_{ii}$, $\operatorname{cov}\left(X_i, X_j \right) = \left(\Sigma_{\mathbf{X}} \right)_{ij} \quad \left(i \neq j \right)$

***Todas las distr. marginales y condicionales son normales o normales multivariadas.**

Distribución normal multivariada



Fuente: http://www.ntrand.com/images/functions/plot/binormdist.jpg

Características de la distribución normal

Teo.: propiedades útiles de la distribución normal

(1) Si $X \sim N\left(\mu, \sigma^2\right)$, $Y = aX + b \sim N\left(a\mu + b, a^2\sigma^2\right)$, $a \neq 0$.

(1.1) En particular, $Z = \left(\dfrac{X - \mu}{\sigma}\right) \sim N(0,1)$.

(2) Si $X_1, X_2, \ldots, X_n$ son v.a. independientes con $X_i \sim N\left(\mu_i, \sigma_i^2\right)$,

$\mathbf{X} = \left(X_1, X_2, \ldots, X_n\right)^T \sim MVN\left(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}}\right)$ donde

$\boldsymbol{\mu}_{\mathbf{X}} = \left(\mu_1, \mu_2, \ldots, \mu_n\right)^T$ y $\Sigma_{\mathbf{X}} = \text{diag}\left\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right\}$.

(3) Si $\mathbf{X} \sim MVN\left(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}}\right)$, $\mathbf{a}^T\mathbf{X} + b \sim N\left(\mathbf{a}^T\boldsymbol{\mu}_{\mathbf{X}} + b, \mathbf{a}^T\Sigma_{\mathbf{X}}\mathbf{a}\right)$.

$(X_1, X_2, \ldots, X_n) \sim MVN \Rightarrow X_i \sim N \quad \forall i$

$\not\Leftarrow$

Distribución Chi cuadrado

Descripción: La variable X puede tomar valores positivos. Muy útil en inferencia estadística

Posibles valores: $[0,\infty)$

Notación: $X \sim \chi^2(n)$

Parámetros: $n \in \mathbb{N}^+$: Grados de libertad

Función de densidad:

$$f_X(x) = \begin{cases} \dfrac{\left(1/2\right)^{n/2}}{\Gamma\left(n/2\right)} x^{n/2-1} e^{-x/2}, & x \geq 0 \\ 0, & \text{en otro caso} \end{cases}$$

Función de distribución: No tiene expresión analítica

Valor esperado: $E[X] = n$

Varianza: $Var[X] = 2n$

Función g. mom. y caract.:

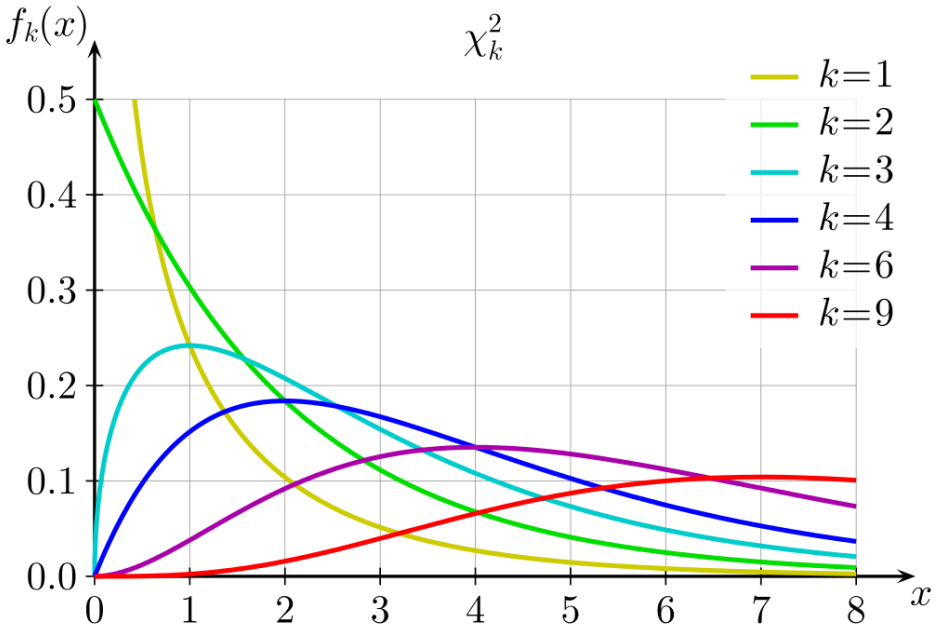
$$m_X(t) = (1-2t)^{-n/2}, \; t < 1/2$$

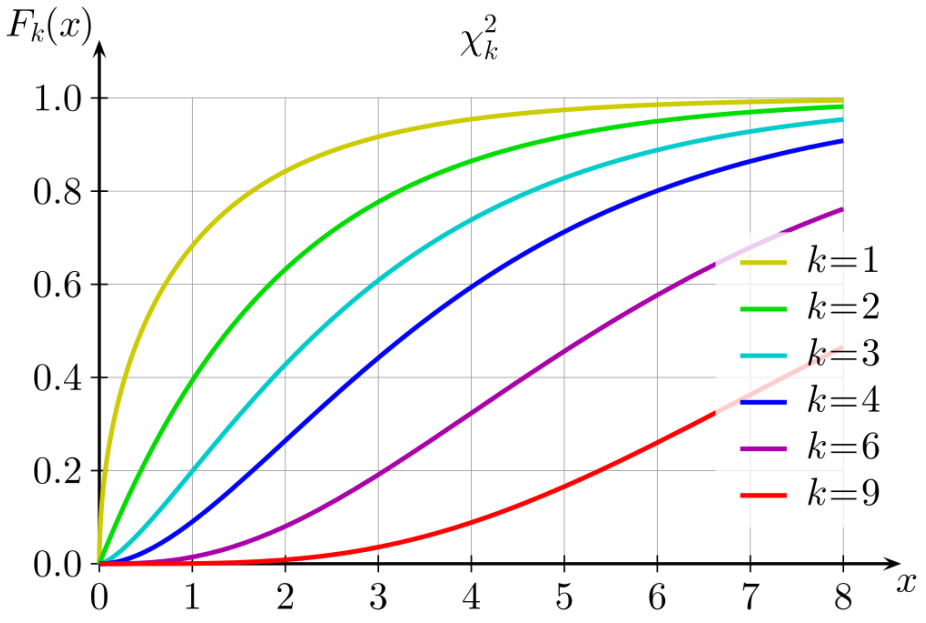
$$\phi_X(t) = (1-2it)^{-n/2}$$

Código R. Para la f. d para d: **dchisq(x=d, df=n)**
Para la f. d. p. para d: **pchisq(q=d , df=n)**
Para el percentil $alpha$: **qchisq(p=$alpha$, df=n)**
Para generar m números aleatorios: **rchisq(n=m , df=n)**

Características de la distribución chi cuadrado

- Algunos posibles resultados de la f. d., dependiendo del valor de los parámetros





Fuente: Wikipedia

Características de la distribución Chi cuadrado

Teo.: propiedades útiles de la distribución Chi cuadrado

(1) Si $Z \sim N(0,1)$, $Y = Z^2 \sim \chi^2(1)$.

(2) Si $Z_1, Z_2, \ldots, Z_n$ son una m. a. $N(0,1)$, $\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$.

(3) En general, la suma de Chi-cuadrados independientes
tiene distribución Chi-cuadrado con la suma de los grados de libertad.

(4) $\chi^2(n) \overset{d}{=} \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.

Distribución t de Student

Descripción: La variable X puede tomar cualquier número real. Muy útil en inferencia estadística

Posibles valores: (-∞, ∞)

Notación: $X \sim t(n)$

Parámetros: $n \in \mathbb{N}^+$: Grados de libertad

Función de densidad: $$f_X(x) = \frac{1}{\sqrt{n\pi}}\Gamma\left(\frac{n+1}{2}\right)\left[\Gamma\left(\frac{n}{2}\right)\right]^{-1}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

Función de distribución: No tiene expresión analítica

Valor esperado: $E[X] = 0$

Varianza: $Var[X] = \frac{n}{n-2}$, si $n > 2$

Función g. mom. y caract.:

$m_X(t)$ no existe

$\phi_X(t)$ no es sencilla

Código R. Para la f. d para *d*: **dt(x=*d*, df=*n*)**

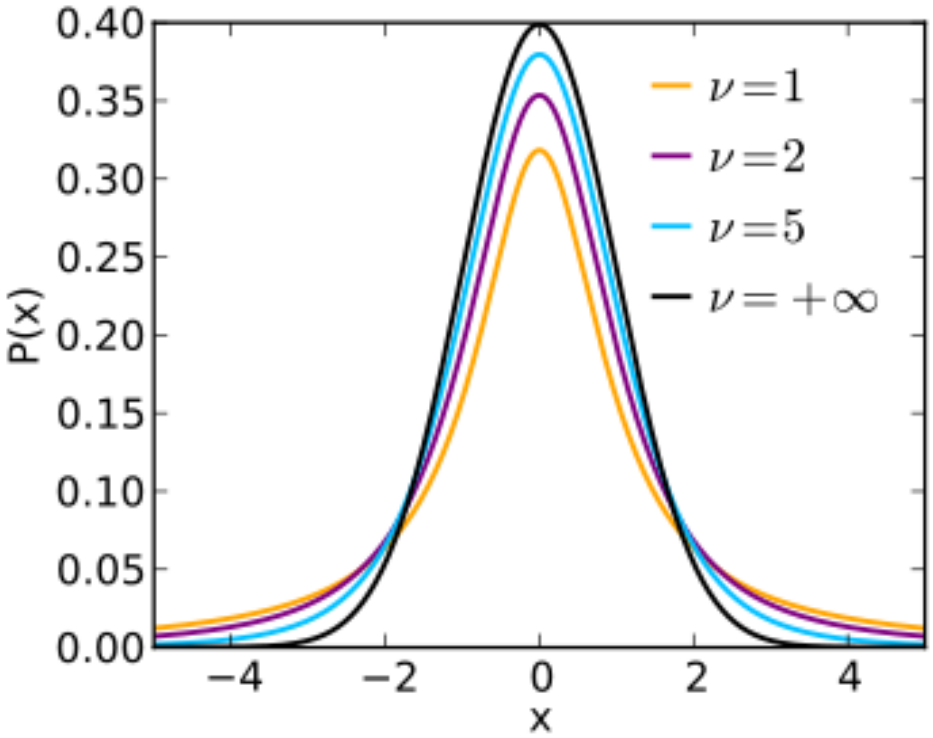
Para la f. d. p. para *d*: **pt(q=*d* , df=*n*)**

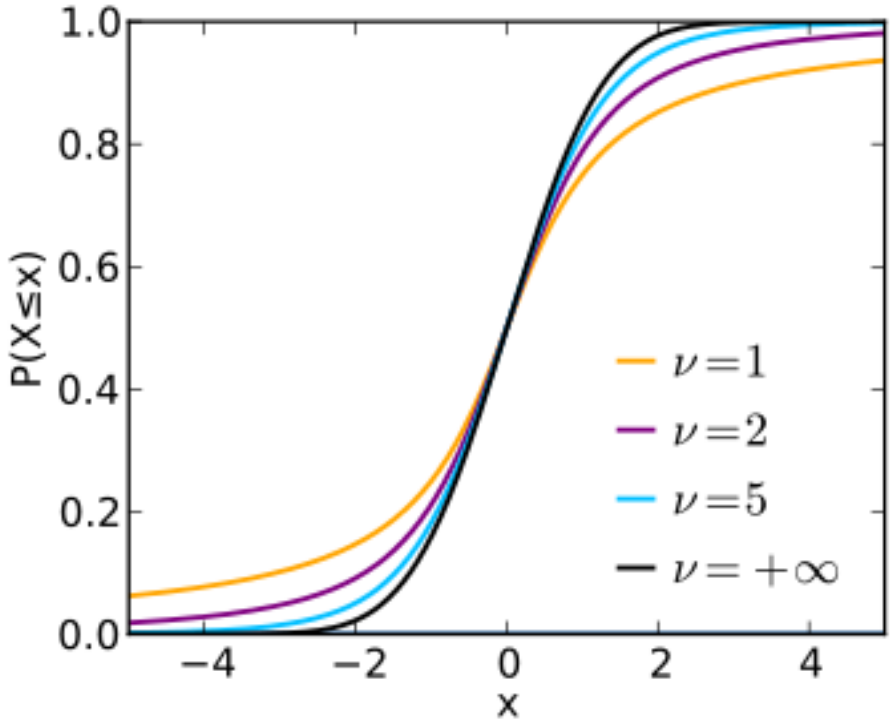
Para el percentil *alpha*: **qt(p=*alpha* , df=*n*)**

Para generar *m* números aleatorios: **rt(n=*m* , df=*n*)**

Características de la distribución t de Student

- Algunos posibles resultados de la f. d., dependiendo del valor de los parámetros





Fuente: Wikipedia

Características de la distribución t de Student

Teo.: propiedades útiles de la distribución t de Student

(1) Si $Z \sim N(0,1)$, $Y \sim \chi^2(n)$ independientes, entonces

$$\frac{Z}{\sqrt{Y/n}} \sim t(n).$$

(2) Sea $\{T_n\}$ tal que $T_k \sim t(k)$, entonces $T_n \xrightarrow{d} Z \sim N(0,1)$

Distribución F de Fisher

Descripción: La variable X puede tomar cualquier número real positivo. Muy útil en inferencia estadística

Posibles valores: $(0,\infty)$

Notación: $X \sim F\left(m,n\right)$

Parámetros: $m,n \in \mathbb{N}^{+}$: Grados de libertad del numerador/del denominador

Función de densidad: $$f_X\left(x\right)=\Gamma\left(\frac{m+n}{2}\right)\left[\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1}\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{x^{(m-2)/2}}{\left[1+\left(m/n\right)x\right]^{(m+n)/2}}I_{(0,\infty)}\left(x\right)$$

Función de distribución: No tiene expresión analítica

Valor esperado: $E\left[X\right]=\frac{n}{n-2}$, si $n>2$

Varianza: $Var\left[X\right]=\frac{2n^2\left(m+n-2\right)}{m\left(n-2\right)^2\left(n-4\right)}$, si $n>4$

Función g. mom. y caract.:

$m_X\left(t\right)$ no existe

$\phi_X\left(t\right)$ no es sencilla

Código R. Para la f. d para *d*: **df(x=*d*, df1=*m* , df2=*n*)**

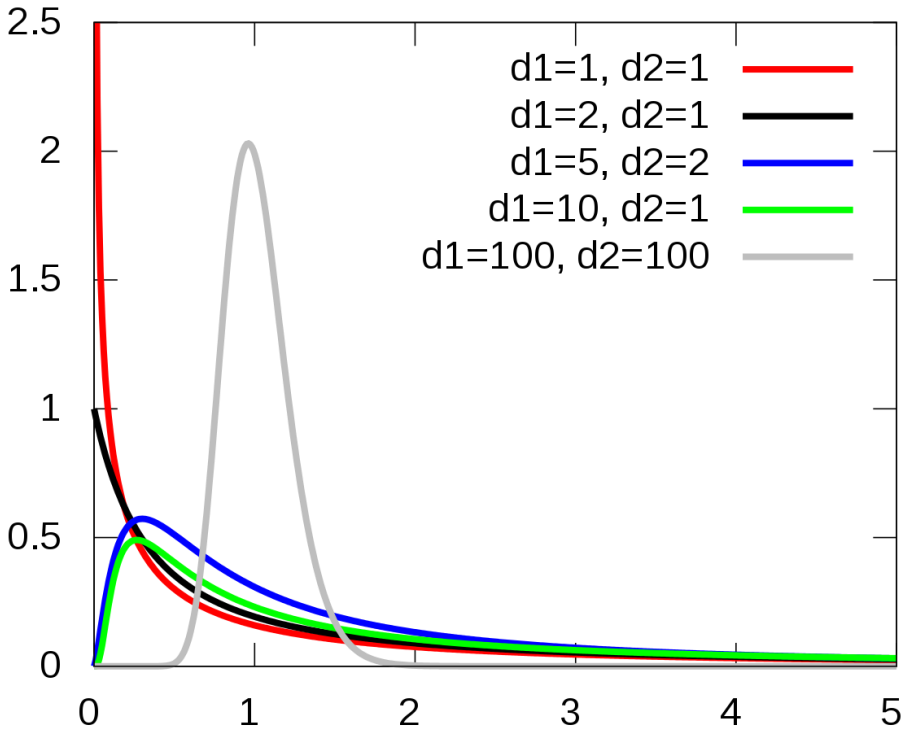
Para la f. d. p. para *d*: **pf(q=*d* , df1=*m* , df2=*n*)**

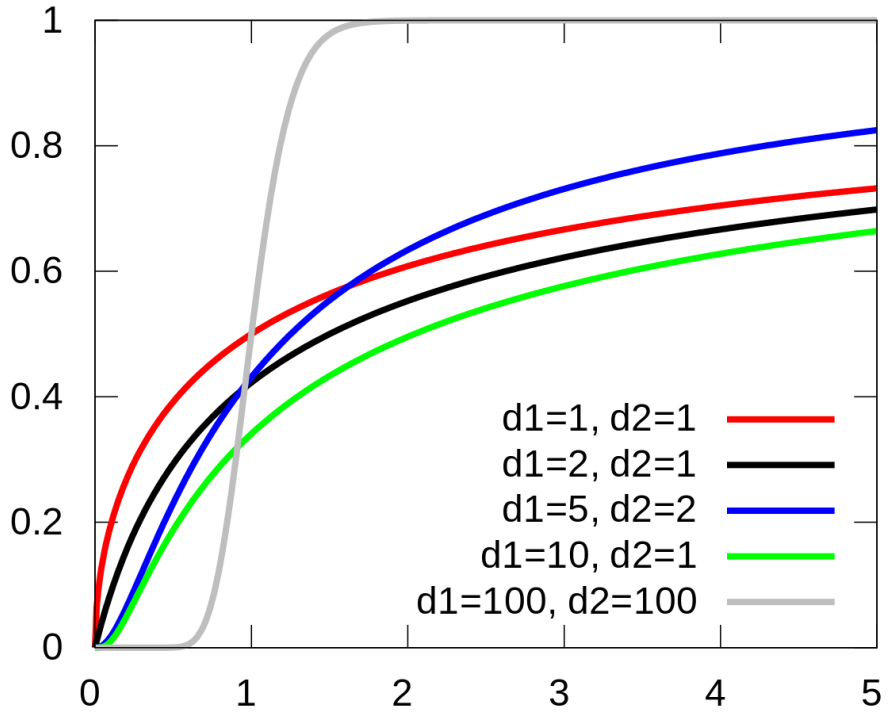
Para el percentil *alpha*: **qf(p=*alpha* , df1=*m* , df2=*n*)**

Para generar *r* números aleatorios: **rf(n=*r* , df1=*m* , df2=*n*)**

Características de la distribución F de Fisher

- Algunos posibles resultados de la f. d., dependiendo del valor de los parámetros





Fuente: Wikipedia

Características de la distribución F de Fisher

Teo.: propiedades útiles de la distribución F de Fisher

(1) Si $W \sim \chi^2(m)$, $Y \sim \chi^2(n)$ independientes, entonces

$$\frac{W/m}{Y/n} \sim F(m,n).$$

(2) Si $F_1 \sim F(m,n)$, entonces $1/F_1 \sim F(n,m)$.

(3) Si $T \sim t(n)$, entonces $T^2 \sim F(1,n)$.

(4) Sea $\{F_n\}$ tal que $F_k \sim F(m,k)$, entonces $mF_n \xrightarrow[n\to\infty]{d} \chi^2(m)$

Relación entre 2 variables cualitativas

Exploración

- **Tablas de contingencia:** Muestran las frecuencias conjuntas de ocurrencia de cada una de las posibles combinaciones de categorías entre las dos variables.

Género	Estado civil
H	Soltero
M	Casado
H	Casado
H	Casado
M	Soltero
H	Casado
H	Casado
M	Viudo
M	Soltero
M	Soltero

Frecuencias absolutas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	1	4	0	5
M	3	1	1	5
Total col.	4	5	1	10

Frecuencias relativas

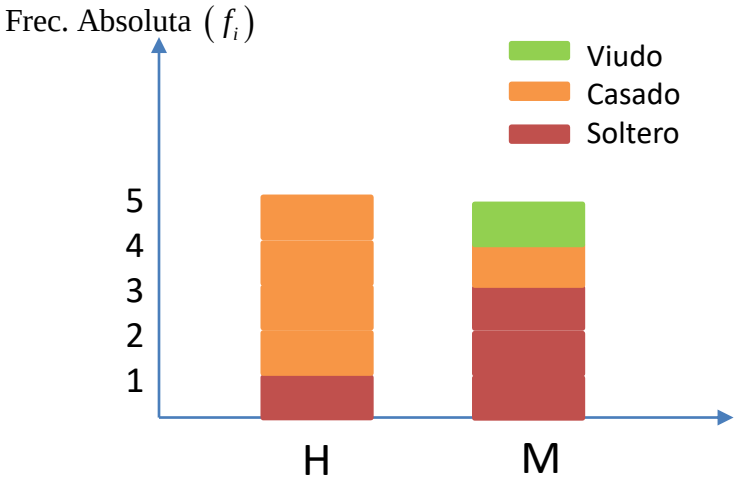
Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	10%	40%	0%	50%
M	30%	10%	10%	50%
Total col.	40%	50%	10%	100%

Exploración (II)

- **Perfiles fila:** Dividen a cada celda por el total de la fila.
- **Perfiles columna:** Dividen a cada celda por el total de la columna.

Frecuencias absolutas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	1	4	0	5
M	3	1	1	5
Total col.	4	5	1	10



Perfiles fila

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	20%	80%	0%	100%
M	60%	20%	20%	100%

Perfiles columna

Género/ Est. civil	Soltero	Casado	Viudo
H	25%	80%	0%
M	75%	20%	100%
Total col.	100%	100%	100%

Pruebas de independencia para variables cualitativas

- Los tamaños de muestra anteriores eran simplemente ilustrativos porque son muy bajos para hacer una prueba de hipótesis. Considere esta situación:

Género/ Est. civil	Soltero	Casado	Viudo	**Total fila**
H	32	20	2	54
M	18	19	9	46
Total col.	50	39	11	100

- ¿Cómo sabemos si la distribución en estados civiles es diferente para los géneros en la población de donde se tomaron los datos? Para ello, se puede realizar la siguiente prueba de hipótesis:

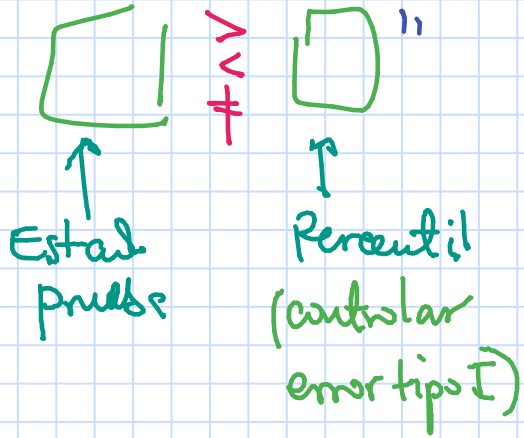
$$\begin{cases} H_0: \text{Las variables género y estado civil son independientes} \\ \textit{versus} \\ H_1: \text{Las variables género y estado civil } \mathbf{no} \text{ son independientes} \end{cases}$$

¿Cómo hacer una prueba de hipótesis?

① Sistema de hipótesis $\begin{cases} H_0: \\ H_1: \end{cases}$

② Criterio o test:

Σ: "Rechazar H_0 si



Pruebas de independencia para variables cualitativas

- Si fuesen independientes y se seleccionase un individuo de la población al azar:

$$p\left(\text{soltero y hombre}\right) = p\left(\text{soltero}\right) \cdot p\left(\text{hombre}\right)$$

- Y así sucesivamente con las demás combinaciones de variables. Entonces, bajo la hipótesis nula, es de esperarse que:

$$\hat{p}\left(\text{soltero y hombre}\right) \approx \hat{p}\left(\text{soltero}\right) \cdot \hat{p}\left(\text{hombre}\right)$$

- Y se puede hacer un cálculo de las frecuencias esperadas a partir de este supuesto.

Pruebas de independencia para variables cualitativas

Frecuencias observadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	0.32	0.20	0.02	0.54
M	0.18	0.19	0.09	0.46
Total col.	0.5	0.39	0.11	1

Frecuencias esperadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**0.27**	**0.21**	**0.06**	0.54
M	**0.23**	**0.18**	**0.05**	0.46
Total col.	0.5	0.39	0.11	1

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	32	20	2	54
M	18	19	9	46
Total col.	50	39	11	100

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**27**	**21**	**6**	54
M	**23**	**18**	**5**	46
Total col.	50	39	11	100

Prueba de independencia

Test Chi-cuadrado de independencia

Considere el sistema: $\begin{cases} H_0: \text{ Las variables son independientes} \\ \textit{versus} \\ H_1: \text{ Las variables } \textbf{no} \text{ son independientes} \end{cases}$.

Si se consideran $c \geq 2$ categorías para las columnas $\{a_1, a_2, \ldots, a_c\}$ y $r \geq 2$ categorías para las filas $\{b_1, b_2, \ldots, b_r\}$,

$$J_n := \sum_{i=1}^{c} \sum_{l=1}^{r} \frac{\left(O_{a_i b_l} - E_{a_i b_l}\right)^2}{E_{a_i b_l}} \xrightarrow[n \to \infty]{d\ (H_0 \text{ cierta})} J \sim \chi^2\left((r-1)(c-1)\right).$$

El test τ:"Rechazar H_0 si $j_n > \chi^2_{1-\alpha}\left((r-1)(c-1)\right)$" es un test aprox. del $100 \cdot \alpha\%$ de significancia.

El p-valor de este test es $p - value = p\left[J > j_n\right]$.

Pruebas de independencia para variables cualitativas

Frecuencias observadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	0.32	0.20	0.02	0.54
M	0.18	0.19	0.09	0.46
Total col.	0.5	0.39	0.11	1

Frecuencias esperadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**0.27**	**0.21**	**0.06**	0.54
M	**0.23**	**0.18**	**0.05**	0.46
Total col.	0.5	0.39	0.11	1

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	32	20	2	54
M	18	19	9	46
Total col.	50	39	11	100

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**27**	**21**	**6**	54
M	**23**	**18**	**5**	46
Total col.	50	39	11	100

Observados: 12

- $j_{100} \approx 7.81$, y $\chi^2_{0.95}(2 \cdot 1) = 5.99$; luego, con un nivel de significancia del 5%, se rechaza la hipótesis nula.

$j_{100}: \frac{(32-27)^2}{27} + \frac{(20-21)^2}{21} + \ldots$

Algunas observaciones

- La función `chisq.test` de R calcula el test de independencia. Sin embargo, hay que pasarle los vectores de frecuencias observadas como una matriz.

```
> cont<-matrix(c(32,20,2,18,19,9),ncol=3,nrow=2,byrow=T)
> chisq.test(cont)

        Pearson's Chi-squared test

data:  cont
X-squared = 7.8102, df = 2, p-value = 0.02014
```

- Este test también requiere frecuencias esperadas mayores a 5. El software generará un error o una advertencia si no es así y se deberá reagrupar.

- Ver (p.132, Sáez-Castillo.)

Relación entre variables cualitativas nominales

- Una vez que se rechaza la hipótesis nula, ¿qué tan fuerte se puede decir que la relación es?

Coeficiente V de Crámer

$$V := \sqrt{\frac{J_n / n}{\min\{r-1, c-1\}}}.$$

V toma valores entre 0 y 1, donde 0 representa ausencia de relación y 1 representa perfecta relación.

- En R, se encuentra la función `cramerV()` del paquete `rcompanion` o del paquete `lsr`. Este último también ofrece una versión de corrección de sesgo.

Relación entre variables cualitativas nominales

Frecuencias observadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	0.32	0.20	0.02	0.54
M	0.18	0.19	0.09	0.46
Total col.	0.5	0.39	0.11	1

Frecuencias esperadas

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**0.27**	**0.21**	**0.06**	0.54
M	**0.23**	**0.18**	**0.05**	0.46
Total col.	0.5	0.39	0.11	1

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	32	20	2	54
M	18	19	9	46
Total col.	50	39	11	100

Género/ Est. civil	Soltero	Casado	Viudo	Total fila
H	**27**	**21**	**6**	54
M	**23**	**18**	**5**	46
Total col.	50	39	11	100

- $$v = \sqrt{\frac{j_{100}/100}{\min\{2-1, 3-1\}}} \approx \sqrt{\frac{7.81/100}{1}} \approx 0.279.$$

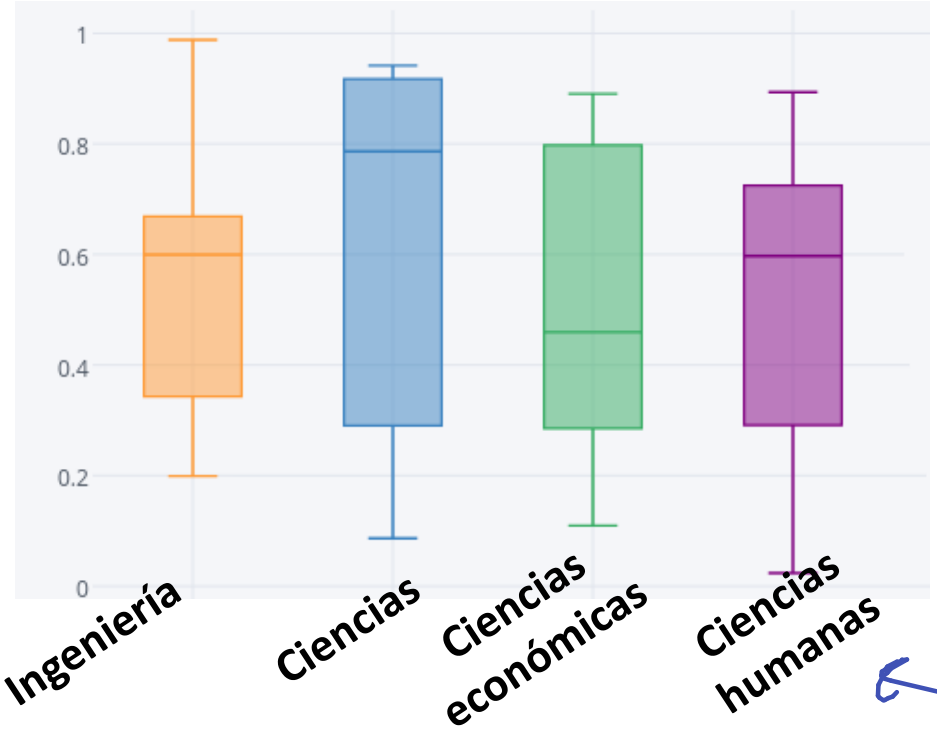
La relación es débil-moderada.

Relación entre una variable cualitativa y una cuantitativa

Exploración (1 v. cuantitativa y 1 v. cualitativa)

- Las categorías de la variable cualitativa dividen al conjunto de análisis en **clases o subgrupos**.
- Se hace un diagrama (histograma o boxplot) para los valores de la variable cuantitativa para cada clase o subgrupo y se comparan entre sí (medidas de tendencia central, dispersión, simetría, curtosis).

Porcentaje de tiempo del día que el estudiante está estudiando



Fuente: https://stackoverflow.com/questions/53767621/boxplot-with-pandas-in-python