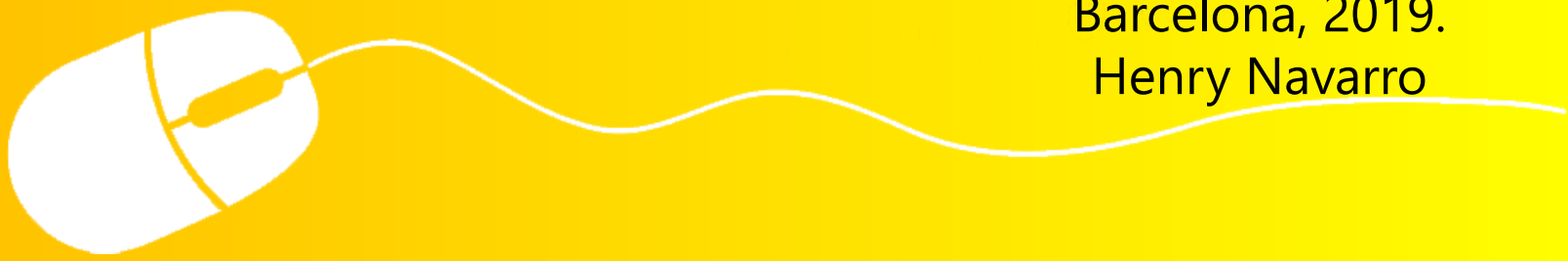
KSCHOOL

Máster Data Science

Barcelona, 2019.

Henry Navarro



KSCHOOL

About me

- Ahora: Lead Data Scientist – Altran Innovation.
- Antes: Data Scientist – Equifax, Solutio, Enefgy.
- Profesor Universidad Carlos III de Madrid – Grupo ML4DS & GISC. (https://goo.gl/9eeHAz)
- Profesor Escuela de Organización Industrial.
- Data Analyst – Ministerio de Justicia (Venezuela).
- Máster en Ingeniería Matemática – UC3M.
- Lienciado en Matemática – UCV.

KSCHOOL

Intro Estadística en R

- R es considerado el mejor lenguaje para análisis estadísticos.
- R fue creado y orientado a las estadísticas, esto hace que proporcione un amplio abanico de herramientas.
- Entre otras características de R, podemos nombrar su capacidad gráfica, que permite generar gráficos con alta calidad, con sólo utilizar las funciones de graficación.
- R también puede usarse como herramienta de cálculo numérico y a la vez ser útil para la minería de datos.

Motivación Estadística en R

- **Estadística = King** : es un lenguaje desarrollado por dos estadísticos que se propusieron crear un sistema para la computación estadística y la visualización de datos.
- **Paquetes:** al ser código abierto, tiene muchísimos colaboradores y muy respaldado. Los paquetes en R para estadísticas y ciencia de datos van desde muy específicos hasta generalmente versátiles. Ejemplo: tidyverse.
- **Visualización = Reina:** R tiene uno de los conjuntos más completos de herramientas de visualización de datos que puede haber. Ejemplo: https://www.r-graph-gallery.com/.
- **Puerta de entrada a data science:** la evolución de los datos y el interés por el aprovechamiento de éstos, ha hecho que por supuesto R se adapte a la ciencia de datos que conduce obviamente a lo que enseñamos en este master

“R se está quedando atrás”, “R va a morir dentro de poco”, ...

Recientemente, se ha hablado de que R "no es suficiente", en términos de la velocidad de procesamiento, memoria o cómo maneja sus datos y la vectorización. También ha habido sugerencias para eliminar R en su totalidad y desarrollar un nuevo lenguaje mejor desde cero.

Nada más alejado de la realidad, hay dos puntos importantes a considerar:

- R se está adaptando (a través de más vectorización, uso de multinúcleos, etc.).
- R es todavía el líder en ser “the data language".

Intro Probabilidad y Estadística

El Cálculo de Probabilidades nos permite calcular el grado de fiabilidad o error de las conclusiones obtenidas mediante inferencia estadística.

La probabilidad mide o cuantifica la incertidumbre que tenemos sobre el resultado de un experimento aleatorio.

Un experimento es determinista cuando existe un conjunto de circunstancias que, antes de su ejecución, determinan completamente su resultado.

Un experimento es aleatorio si no podemos predecir su resultado de antemano:

- Se conocen previamente y con exactitud los posibles resultados del experimento.
- Es imposible saber su resultado antes de su realización.
- Se puede repetir indefinidamente, en las mismas condiciones iniciales, obteniendo resultados distintos.

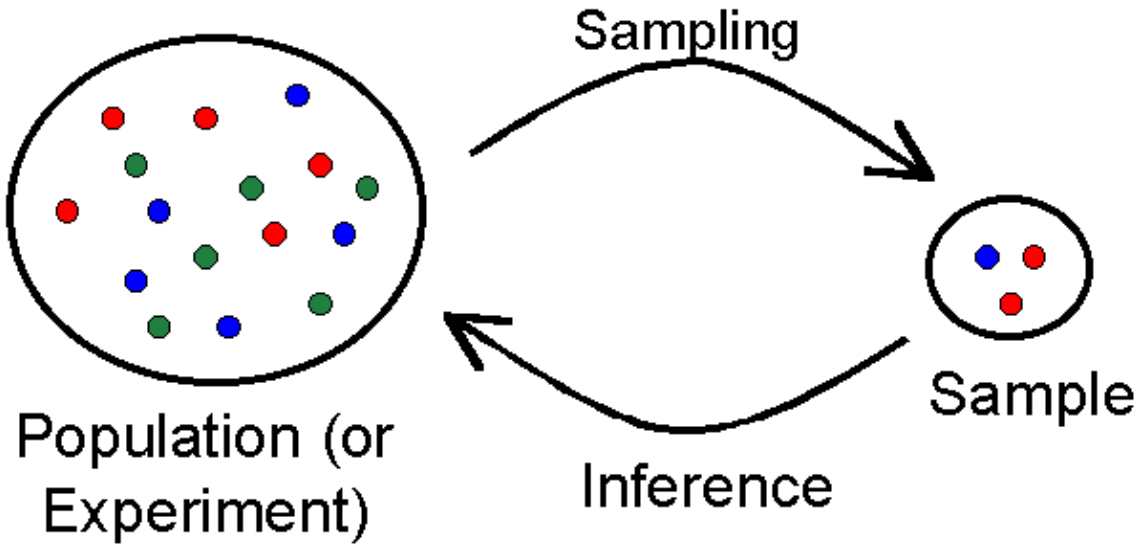
KSCHOOL

Intro Probabilidad y Estadística

La Estadística es la ciencia que estudia cómo debe emplearse la información para luego dar respuesta a situaciones prácticas que entrañan incertidumbre.

La estadística se ocupa de obtener conclusiones de investigaciones mediante el uso de modelos matemáticos, proporcionando una metodología a través de la cual se pueda evaluar y juzgar las discrepancias entre lo observado y lo previsto por el modelo (inferencia), todo esto basado en la información contenida en un conjunto de datos.

Inferencia estadística vs estadística descriptiva



Inferencia estadística vs estadística descriptiva

La **Estadística descriptiva** aporta las técnicas para resumir y presentar la información extraída de una muestra. Sin embargo, rara vez nos interesa la muestra como tal, sino que interesa por su capacidad para aportar información acerca de otros sujetos o otras situaciones.

La **Estadística Inferencial** aporta las técnicas para extraer conclusiones a partir de una muestra. En la inferencia distinguimos:

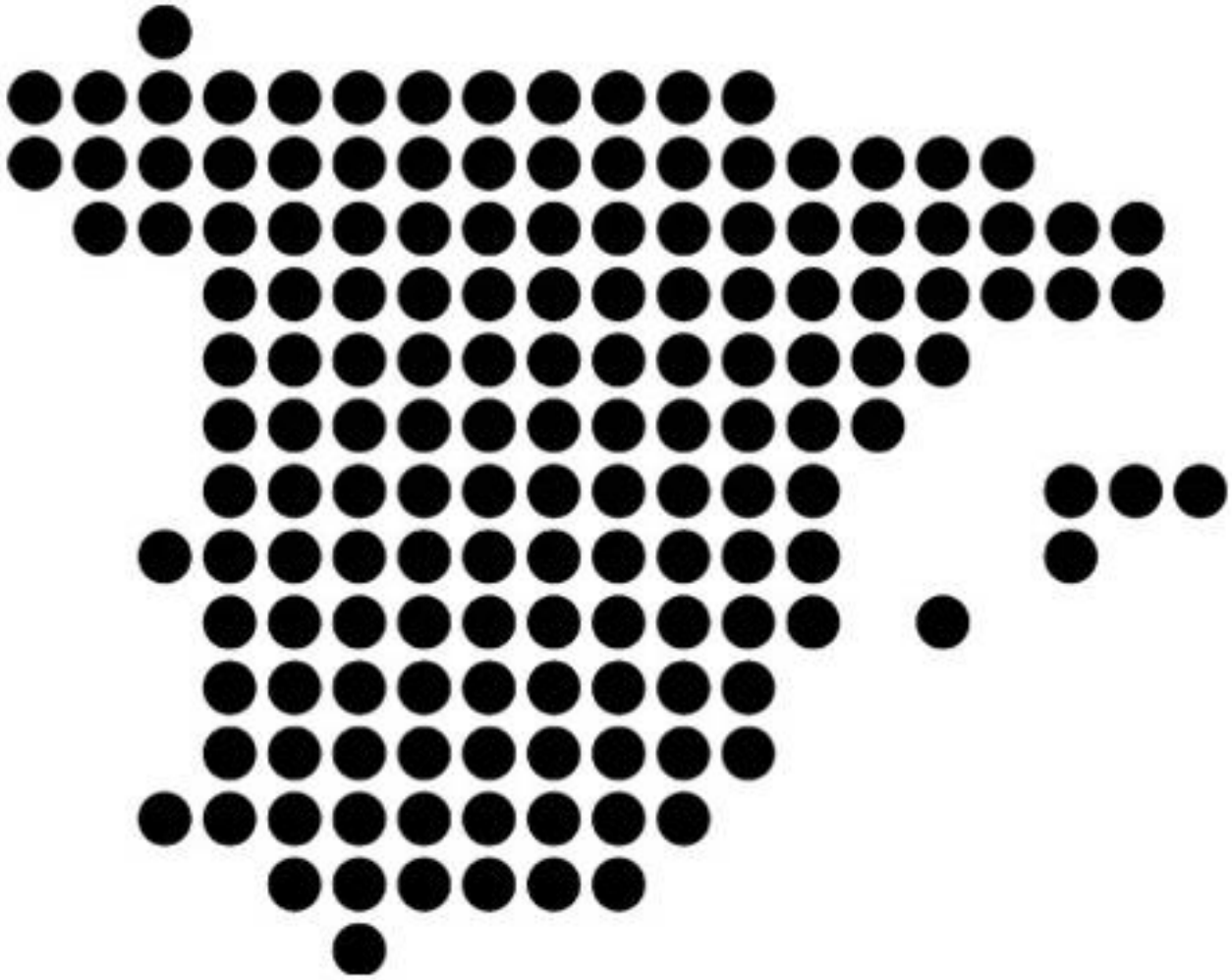
- La Estimación: permite estimar los parámetros de la población a partir de la muestra (Ej.: ¿Qué volumen de envases recicla un hogar español?).
- El Contraste de hipótesis: permite tomar una decisión sobre los parámetros de la población (Ej.: ¿Se recicla mejor que hace 10 años?)

Inferencia estadística vs estadística descriptiva

- ¿Cuántos envases (en Kg) recicla un hogar al año?
- Este volumen de envases es una variable (X) porque varía de un hogar al otro.
- Es aleatoria porque no controlamos (del todo) sus variaciones; no podemos prever sus valores de manera determinista.
- ¿Cómo describir una variable aleatoria?

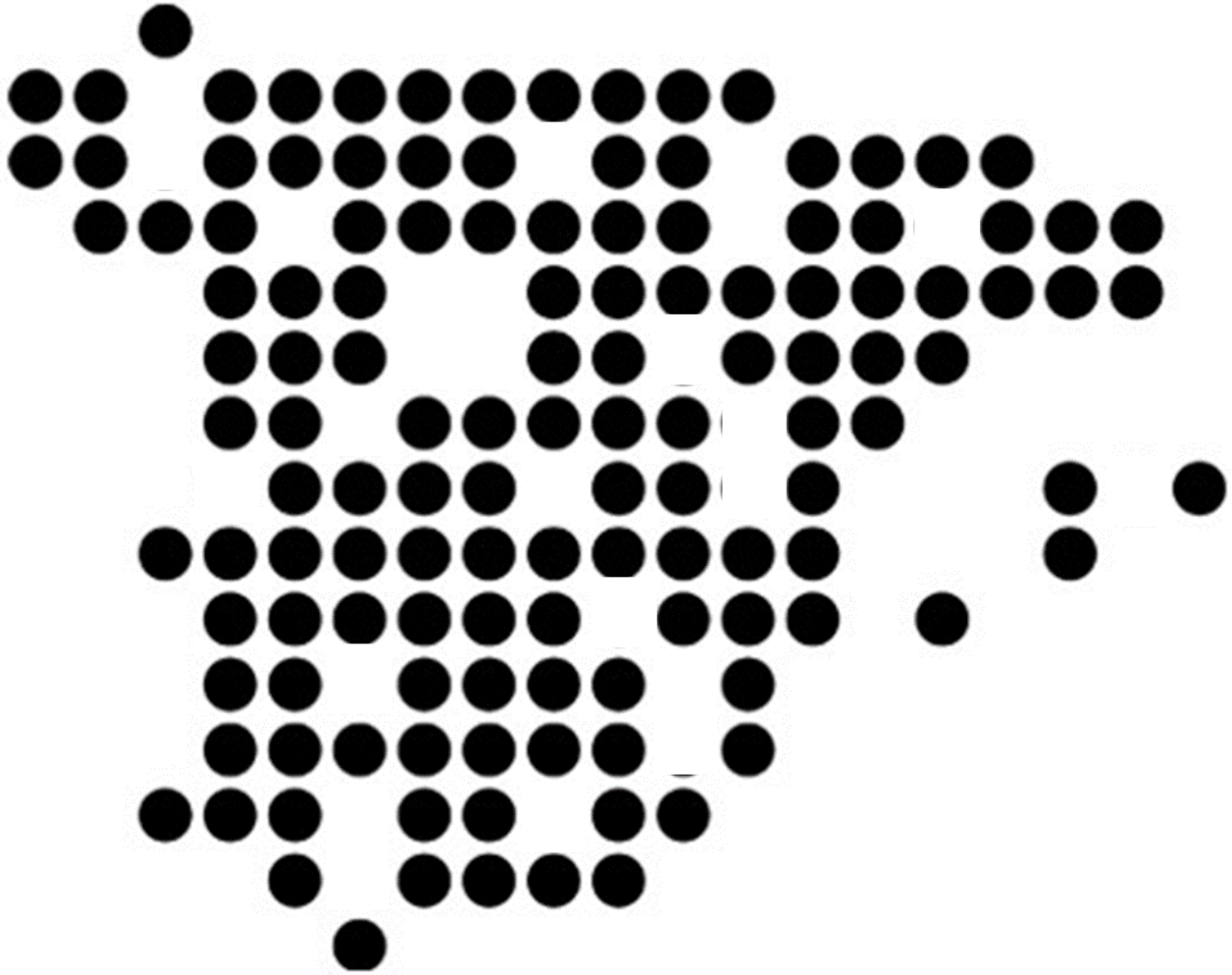
KSCHOOL

Muestra vs. Población



KSCHOOL

Muestra vs. Población



Variable aleatoria.

Variable aleatoria.

Una variable aleatoria es una función que asocia un valor numérico a cada posible resultado de un evento aleatorio.

Ejemplo.

Lanzar un dado una vez. Sea la v.a. X = resultado de la tirada. ¿cuántos sucesos elementales hay? ¿qué valores puede tomar X? Se suelete denotar las v.a. con letras mayúsculas, y sus posibles valores con letras minúsculas.

Variable aleatoria.

Variable aleatoria discreta.

Una variable aleatoria es discreta si toma un número finito o numerable de valores.

Variable aleatoria continua.

Una variable aleatoria es continua si toma un número infinito no numerable de valores (por ejemplo, en un intervalo de $\mathbb{R}$).

Ejemplos

- X =“resultado al tirar un dado” es una variable discreta.
- Y = “altura de un alumno elegido al azar” es una variable continua.

Función de probabilidad

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\}$. Se llama función de probabilidad o función de masa, al conjunto de probabilidades con las que X toma cada uno de sus valores, es decir, $p_i = P[X = x_i]$, para cada $i = 1, 2, \dots$

Ejemplo $X =$ resultado de lanzar un dado. La función de probabilidad es

x	1	2	3	4	5	6
$P[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Función de probabilidad

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots\}$. Se llama función de probabilidad o función de masa, al conjunto de probabilidades con las que X toma cada uno de sus valores, es decir, $p_i = P[X = x_i]$, para cada $i = 1, 2, \dots$

Ejemplo $X =$ resultado de lanzar un dado. La función de probabilidad es

x	1	2	3	4	5	6
$P[X = x]$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

KSCHOOL

Función de distribución

$$F(x) = \begin{cases} \sum_{x_i \leq x} f(x_i), & si\ es\ una\ v.a.\ discreta \\ \int_{-\infty}^{x} f(t)dt, & si\ es\ una\ v.a.\ continua \end{cases}$$

KSCHOOL

Ejemplos v.a. discretas

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Discrete uniform	$p(x) = \frac{1}{n}$ $x = 1, 2, \ldots, n$		$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Hyper-geometric	$\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$ $\text{Max}[0, n-(N-N_1)]$ $\leq x \leq \text{Min}(n, N_1)$		$\mu = n\theta$ $\theta = \frac{N_1}{N}$	$\sigma^2 = \frac{N-n}{N-1} n\theta(1-\theta)$ $\theta = \frac{N_1}{N}$
Bernoulli	$\theta^x (1-\theta)^{1-x}$ $x = 0, 1 \quad 0 \leq \theta \leq 1$	$\theta e^t + (1-\theta)$	θ	$\theta(1-\theta)$

KSCHOOL

Ejemplos v.a. continuas

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Binomial	$\binom{n}{x}\theta^x 1-\theta^{n-x}$ $x=0,1,\ldots,n;\ 0\le\theta\le 1$	$(\theta e^t+(1-\theta))^n$	$n\theta$	$n\theta(1-\theta)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$ $x=0,1,\ldots;\ \lambda>0$	$e^{\lambda(e^t-1)}$	λ	λ
Uniform	$f(x)=\frac{1}{b-a}$ $a\le x\le b$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$ $-\infty<x<\infty,-\infty<\mu<\infty,$ $\sigma>0$	$e^{\mu t+(\sigma^2t^2/2)}$	μ	σ^2
Chi-square	$\frac{1}{2^{n/2}\Gamma(n/2)}w^{n/2-1}e^{-w/2}$ $w\ge 0, n>0$	$(1-2t)^{-n/2}$	n	$2n$
Student-t	$f(t)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}$ $\left(1+\frac{t^2}{n}\right)^{(n+1)/2}$ $-\infty<t<\infty$		0	$\frac{n}{n-2}$

Estadísticos básicos

Media:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Varianza:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{X})^2$$

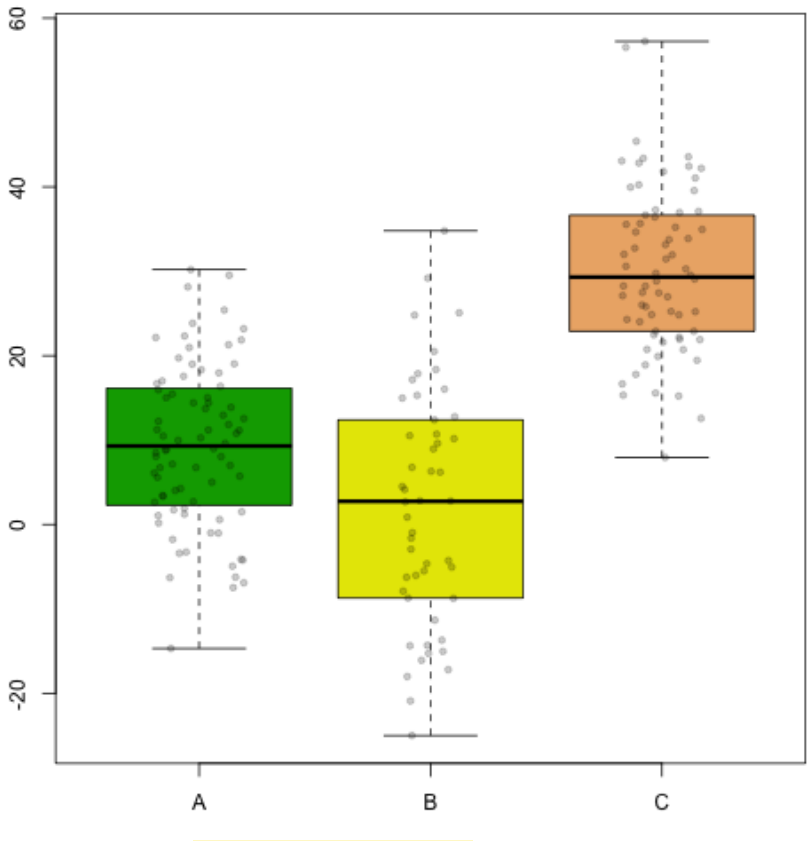
Desviación típica/estándar:

$$sd(X) = \sqrt{Var(X)}$$

Moda: el valor que más se repite en una muestra, puede estimarse en datos agrupados.

Boxplots

No es más que la representación de los principales estadísticos de una variable.

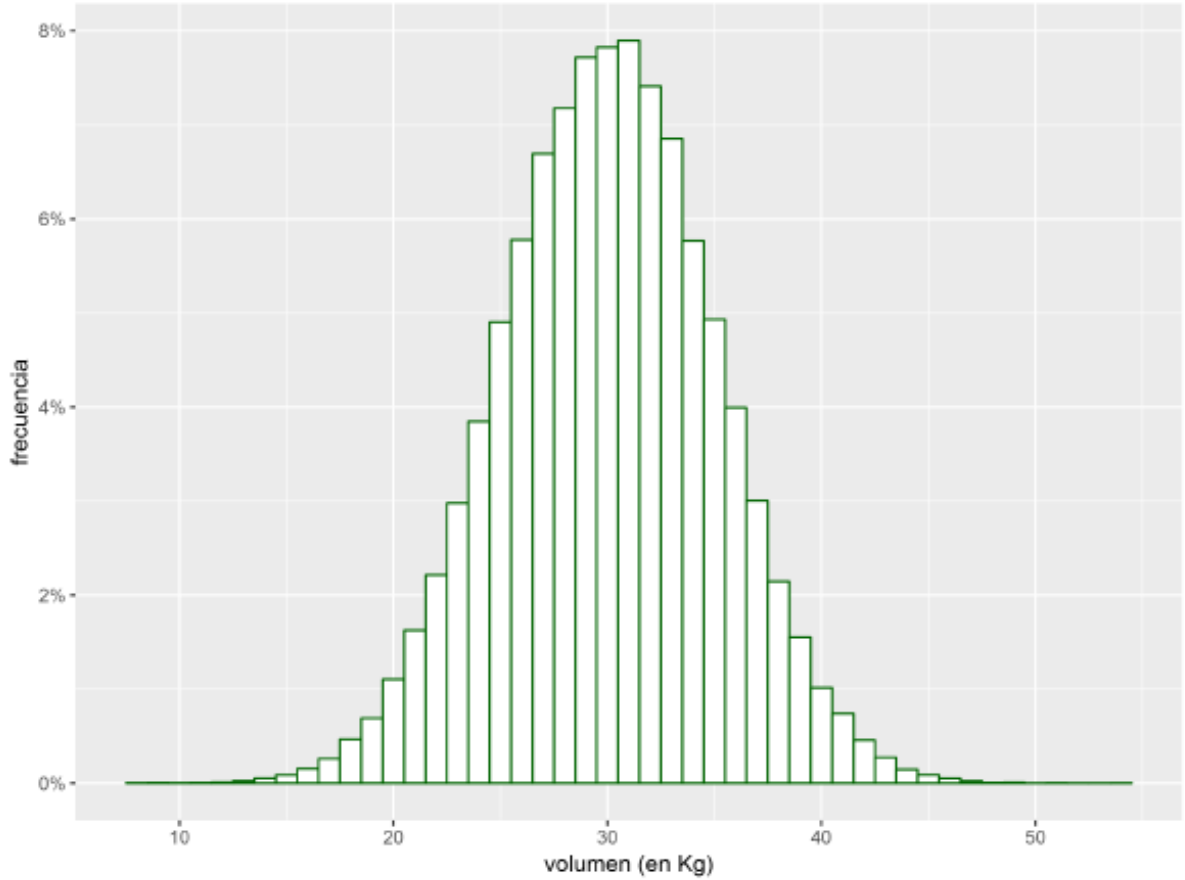


Histograma

- Es una gráfica de la distribución de un conjunto de datos. Es un tipo especial de gráfica de barras, en la cual una barra va pegada a la otra, es decir no hay espacio entre las barras. Cada barra representa un subconjunto de los datos.
- Un histograma muestra la acumulación o tendencia, la variabilidad o dispersión y la forma de la distribución.
- Un histograma es una gráfica adecuada para representar variables continuas, aunque también se puede usar para variables discretas. Es decir, mediante un histograma se puede mostrar gráficamente la distribución de una variable cuantitativa o numérica. Los datos se deben agrupar en intervalos de igual tamaño, llamados clases.

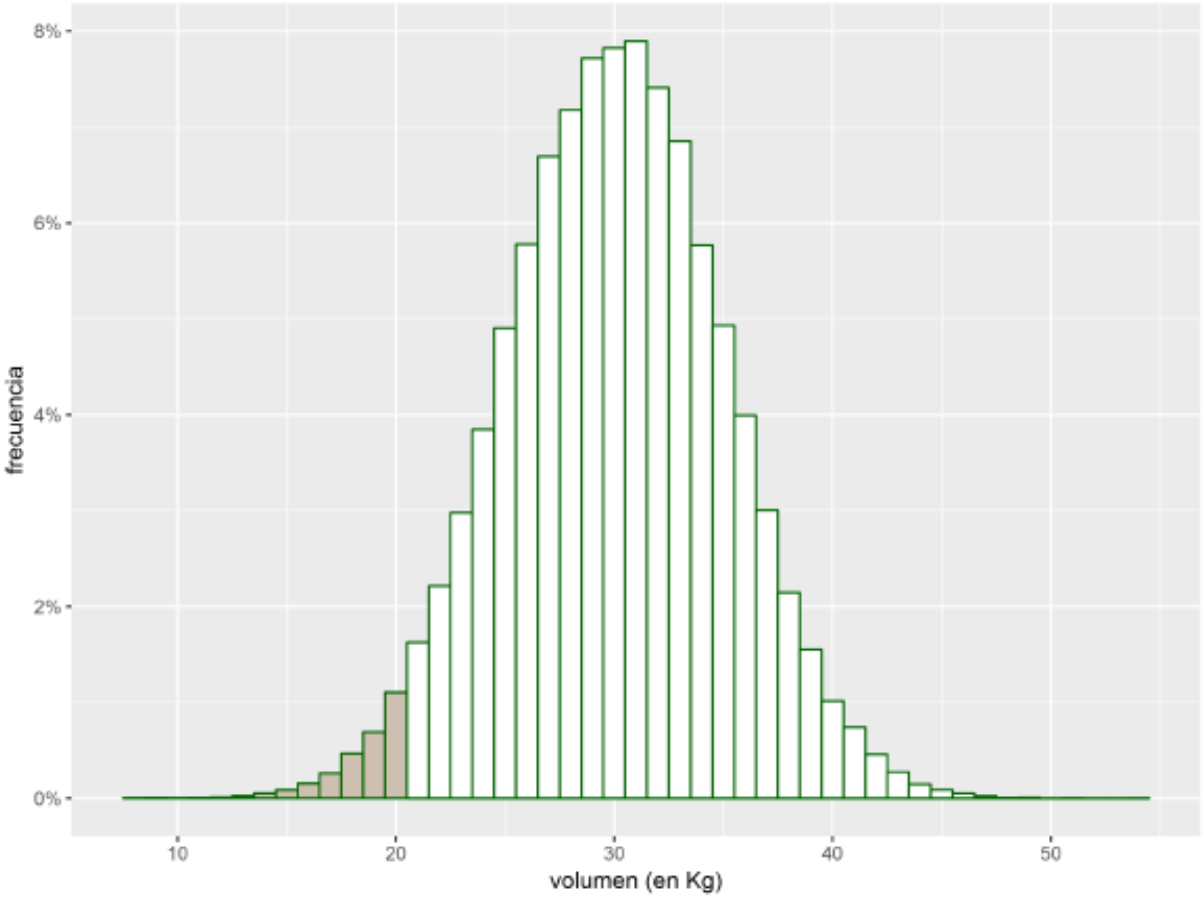
KSCHOOL

Histograma



KSCHOOL

Quantiles



KSCHOOL

Let's do it...

KSCHOOL

Covarianza y correlación

Covarianza poblacional:

$$\sigma(x, y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

Covarianza muestral:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})$$

Correlación entre dos variables:

$$corr(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

KSCHOOL

Let's do it...

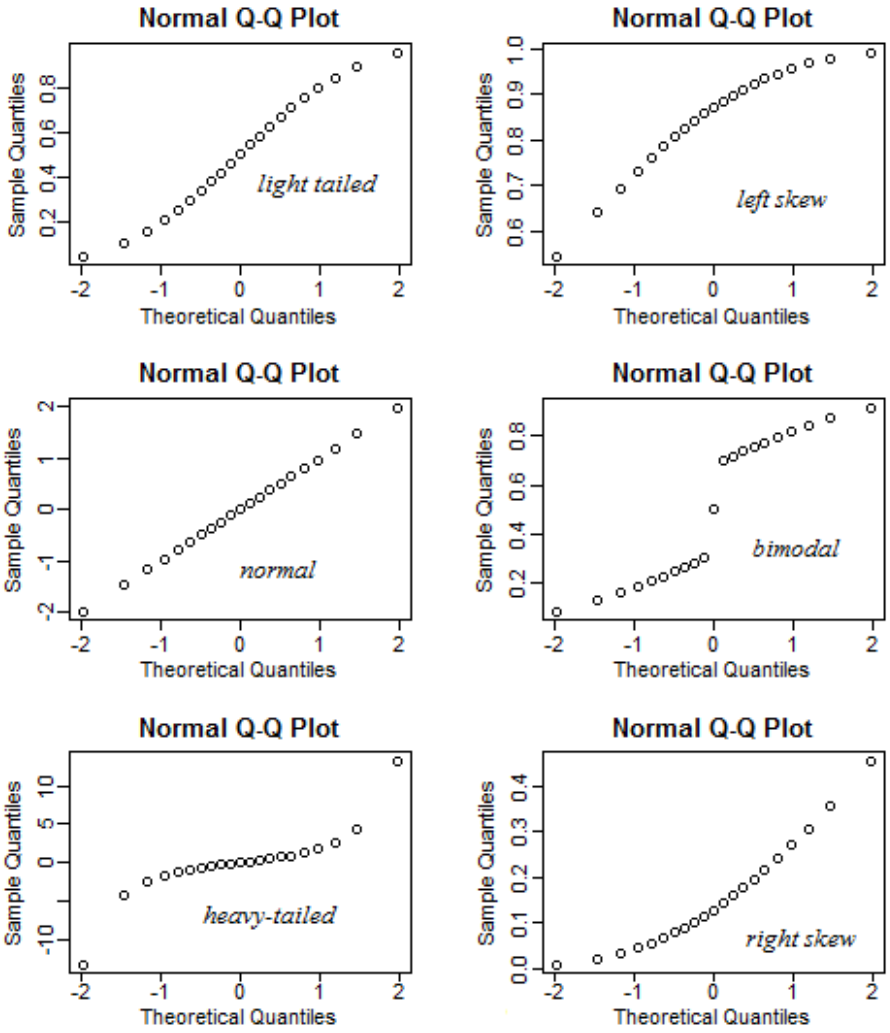
Quantile Quantile Plot

- El quantile quantile plot es una técnica gráfica para determinar si dos conjuntos de datos provienen de poblaciones con una distribución común.
- No es más que una gráfica de los cuantiles del primer conjunto de datos contra los cuantiles del segundo conjunto de datos.
- Recordemos que por un cuantil, nos referimos a la fracción (o porcentaje) de puntos por debajo del valor dado. Es decir, el cuantil de 0.3 (o 30%) es el punto en el que el 30% de los datos cae por debajo y el 70% cae por encima de ese valor.
- Particularmente, utilizaremos el Q-Q plot para determinar si unos datos tienen distribución normal.

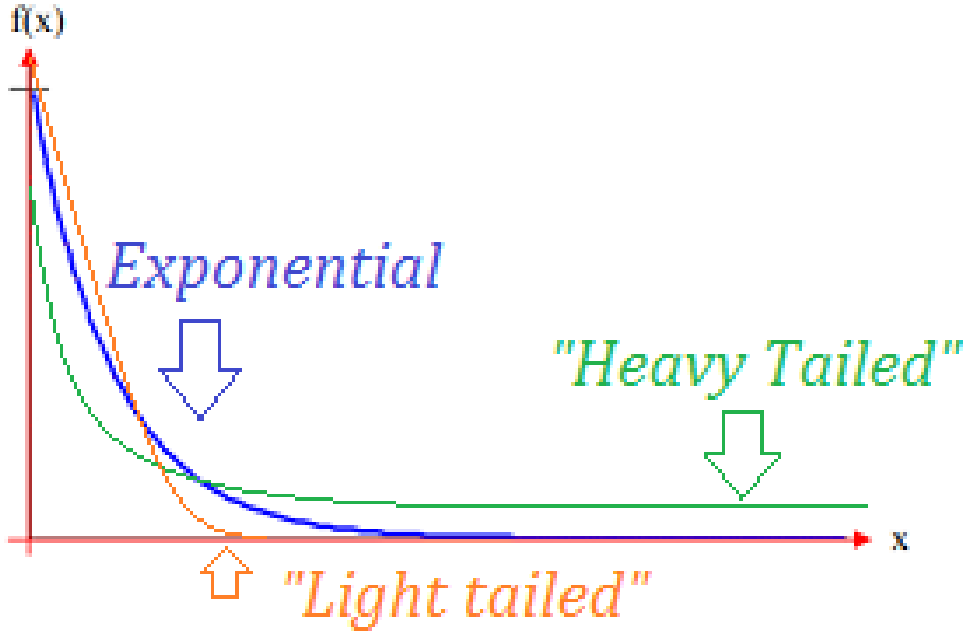
Quantile Quantile Plot

- También se suele traza una línea de referencia de 45 grados. Si los dos conjuntos provienen de una población con la misma distribución, los puntos deberían caer aproximadamente a lo largo de esta línea de referencia.
- Cuanto mayor sea la desviación de esta línea de referencia, mayor será la evidencia de la conclusión de que los dos conjuntos de datos provienen de poblaciones con diferentes distribuciones.
- Los tamaños de muestra no necesitan ser iguales.
- Muchos aspectos distributivos pueden ser probados simultáneamente. Por ejemplo, los diferencias en la "ubicación", los diferencias de escala, diferencias en la simetría y la presencia de valores atípicos se pueden detectar desde esta gráfica.

Quantile Quantile Plot



Quantile Quantile Plot



KSCHOOL

Let's do it...

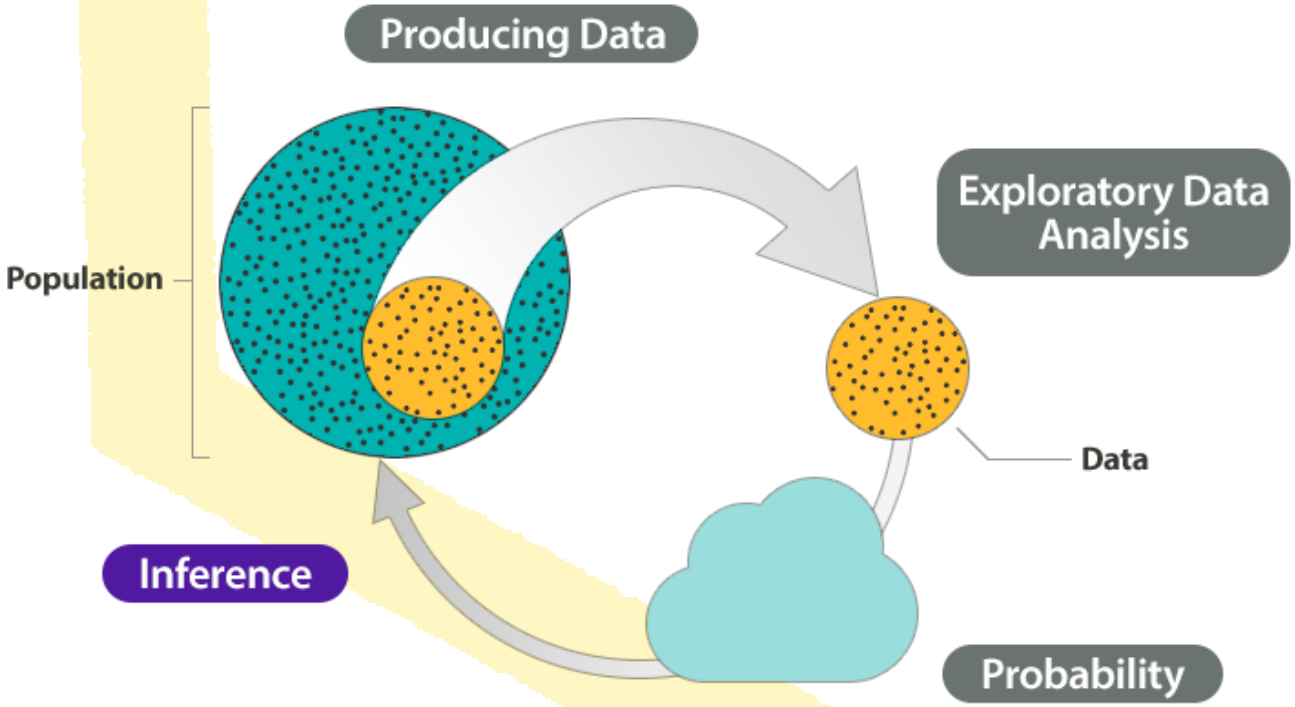
Estimadores

- Si X tiene muchos valores posibles, describir su distribución mediante un histograma puede ser laborioso.
- La mayoría de los modelos de distribución se caracterizan por el conocimiento de su media y su varianza (Ej.: la distribución Normal).
- La media y la desviación típica son respectivamente medidas de localización y de dispersión de la variable.

Estimadores

- Una forma natural de estimar un parámetro poblacional consiste en utilizar su equivalente muestral.
- Con un muestreo equiprobabilistico, $\bar{X}$ es un estimador insesgado.
- Curiosamente, $Var(X)$ no es un estimador insesgado de la varianza σ^2
- En algunos casos, conviene que el estimador sea máximo verosímil.

Inferencia Estadística, o como estimar sobre una población a partir de una muestra...



Intervalos de confianza

- Un intervalo de confianza es un intervalo de números que contiene los valores más plausibles para nuestro parámetro de población.
- Proporcionan el valor de un estadístico mediante un intervalo, bajo una confianza α (nivel de confianza).
- Ecuaciones que permiten mediante un estadístico realizar una estimación para toda la población.

KSCHOOL

Ejemplo

Se seleccionó una muestra aleatoria de 225 cursos de estadísticas del primer año y se registró el número de estudiantes ausentes de cada uno. Los resultados de ausencias fueron $\mu = 11.6$ y $s = 4.1$.

Calcule el número medio de ausencias por curso con un 95% de confianza.

Un intervalo de confianza para μ viene dado por:

$$\bar{x} \pm 1.645\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$11.6 \pm 1.645\left(\frac{4.1}{\sqrt{225}}\right) = 1.6 \pm 0.45 = (11.15, 12.05)$$

Interpretación

Es incorrecto decir que existe una probabilidad de 0.95 de que μ esté entre 11.15 y 12.05. De hecho, esta probabilidad es 1 o 0 (μ está o no está en el intervalo).

El 95% se refiere al porcentaje de todos los intervalos posibles que contienen μ, es decir, al proceso de estimación en lugar de un intervalo en particular.

También es incorrecto decir que el 95% de todos los cursos tenían entre 11.15 y 12.05 alumnos ausentes.

KSCHOOL

Let's do it...

Prueba de hipótesis

Otra manera de hacer inferencia es haciendo una afirmación acerca del valor que el parámetro de la población bajo estudio puede tomar.

Esta afirmación puede estar basada en alguna creencia o experiencia pasada que será contrastada con la evidencia que nosotros obtengamos a través de la información contenida en la muestra. Esto es a lo que llamamos **Prueba de Hipótesis**

KSCHOOL

Prueba de hipótesis

Elementos De Una Prueba de Hipótesis

- a) Hipótesis Nula, H0.
- b) Hipótesis Alternativa, Ha.
- c) Estadístico de Prueba.
- d) Región de Rechazo, RR.

Ejemplo

Elementos De Una Prueba de Hipótesis

- a) Hipótesis Nula, H0.
- b) Hipótesis Alternativa, Ha.
- c) Estadístico de Prueba.
- d) Región de Rechazo, RR.

KSCHOOL

Tipos de errores

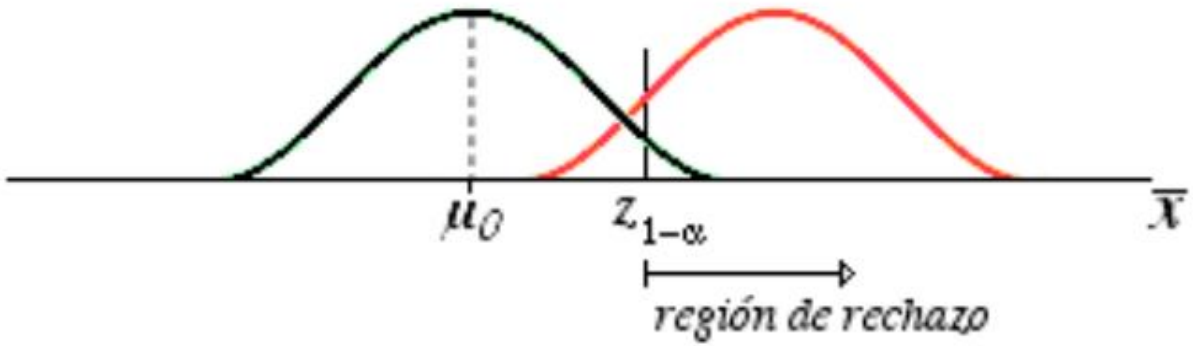
	H_0 Verdadera	H_0 Falsa
Rechazamos H_0	Error Tipo I P(error Tipo I) = α	Decisión Correcta
No Rechazamos H_0	Decisión Correcta	Error Tipo II P(error Tipo II) = β

Conclusiones en prueba de hipótesis

- Si rechazamos la Hipótesis Nula, concluimos que "hay suficiente evidencia estadística para inferir que la hipótesis nula es falsa".
- Si no rechazamos la Hipótesis Nula, concluimos que "no hay suficiente evidencia estadística para inferir que la hipótesis nula es falsa"

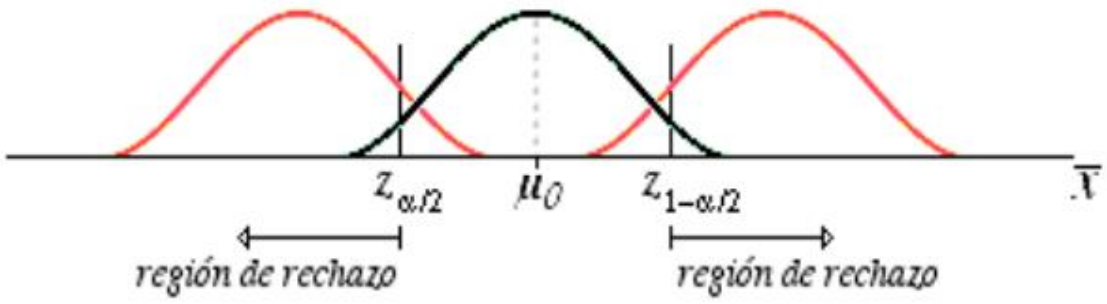
Regiones de rechazo

$$H_1 : \mu > \mu_0$$



Regiones de rechazo

$$H_1 : \mu \neq \mu_0$$



Ejemplos “a mano”

- http://www.estadistica.net/Algoritmos2/guia-pvalor.pdf

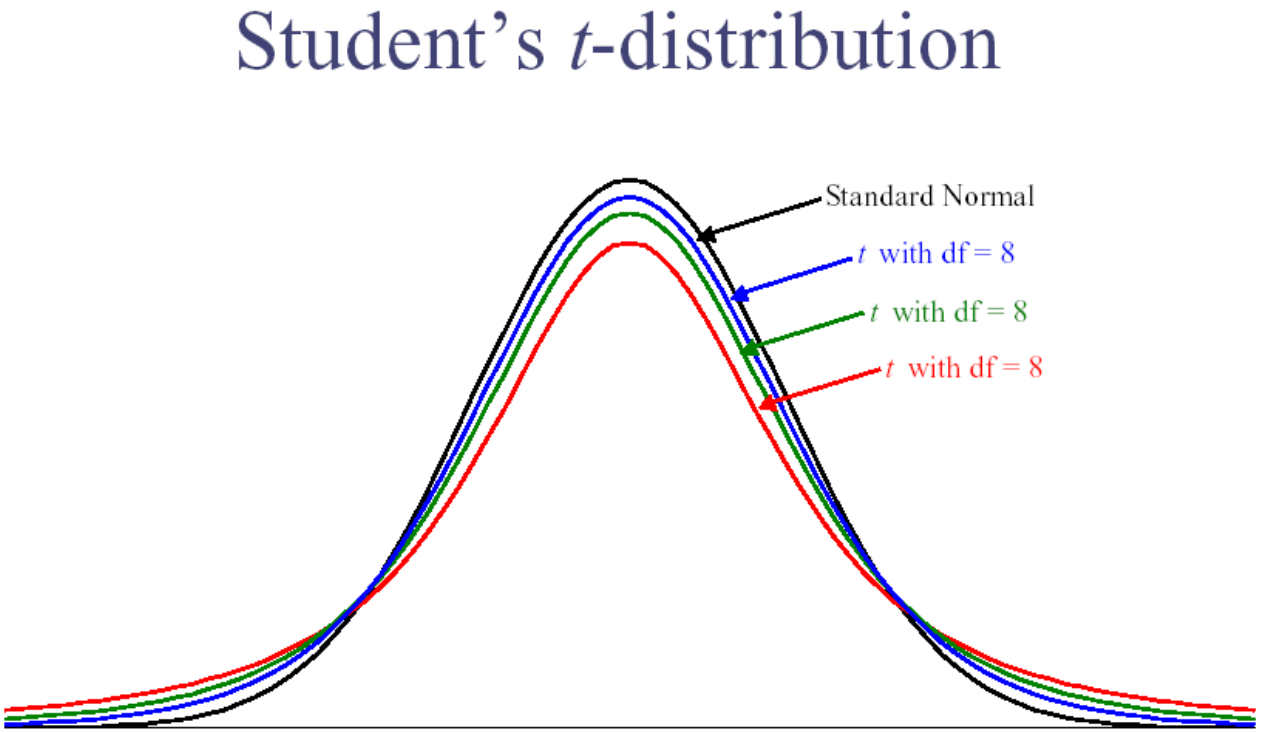
P-valor

- El p-valor es un concepto a menudo mal entendido. A pesar de todas las interpretaciones erróneas, un p-valor es la probabilidad, si la hipótesis nula fuera correcta, de obtener un resultado tan extremo o más extremo.
- Es una medida de cuán extrema es la estadística, en este caso, la media estimada. Si el estadístico es demasiado extremo, concluimos que la hipótesis nula debe ser rechazada.
- Ronald A. Fisher, el padre de las estadísticas modernas, decidió que deberíamos considerar que un p-valor más pequeño que 0.10, 0.05 o 0.01 es demasiado extremo.
- Si bien esos valores de p han sido el estándar durante décadas, fueron elegidos arbitrariamente, lo que llevó a algunos científicos de datos modernos a cuestionar su utilidad.

P-valor

- El valor p es el tamaño más pequeño a para el que se rechaza H0.
- El valor p expresa evidencia contra H0: cuanto más pequeño es el valor p, más fuerte es la evidencia contra H0.
- Generalmente, el valor de p se considera pequeño cuando p <0.01 y grande cuando p> 0.1.
- El valor p no es la probabilidad de que el H0 sea verdadero.

T-Tests: una sola distribución para múltiples pruebas...



One-Sample t-test

Esta prueba esencialmente calcula la media de los datos y crea un intervalo de confianza. Si el valor que estamos probando cae dentro de ese intervalo de confianza, podemos concluir que es el verdadero valor de la media de los datos; de lo contrario, llegamos a la conclusión de que no es el verdadero medio.

Two-Sample t-test

La mayoría de las veces, la prueba t se usa para comparar dos muestras.

Sin embargo, antes de ejecutar la prueba t, primero debemos verificar la varianza de cada muestra.

Una prueba t tradicional requiere que ambos grupos tengan la misma varianza, mientras que la prueba t de dos muestras de Welch puede manejar grupos con variaciones diferentes.

Analysis **O**f **VA**riance

Después de comparar dos grupos, el siguiente paso natural es comparar varios grupos. Cada año, incalculables estudiantes en clases introductorias de estadísticas se ven obligados a aprender el test de ANOVA (análisis de varianza) y memorizar su fórmula, que es

$$F = \frac{\frac{\sum_i n_i (\overline{Y}_i - \overline{Y})^2}{K-1}}{\frac{\sum_{ij} \left(\overline{Y_{ij}} - \overline{Y}_i\right)^2}{N-K}}$$

donde n_i es el número de observaciones en el grupo i, $\overline{Y}_i$ es la media del grupo i, $\overline{Y}$ es la media general, Y_{ij} es la observación j en el grupo i, N es el número total de observaciones y K es el número de grupos.

KSCHOOL

Let's do it...

Bibliografía.

- Ejericios a mano de contraste de hipótesis: http://www.estadistica.net/Algoritmos2/guia-pvalor.pdf
- Q-Q Plot: https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot
- Teoría contraste de hipótesis: http://halweb.uc3m.es/esp/Personal/personas/aarribas/esp/docs/estI_grado/estIG_tema8.pdf
- Libro de estadística digerido: https://www.academia.edu/36701678/Practical_Statistics_for_Data_Scientists
- Guía de probabilidades:
- http://www.matematica.ciens.ucv.ve/pregrado/Probabilidades/Curso%20de%20Probalidad%20por%20M.%20Arriojas.pdf

KSCHOOL

Fin...