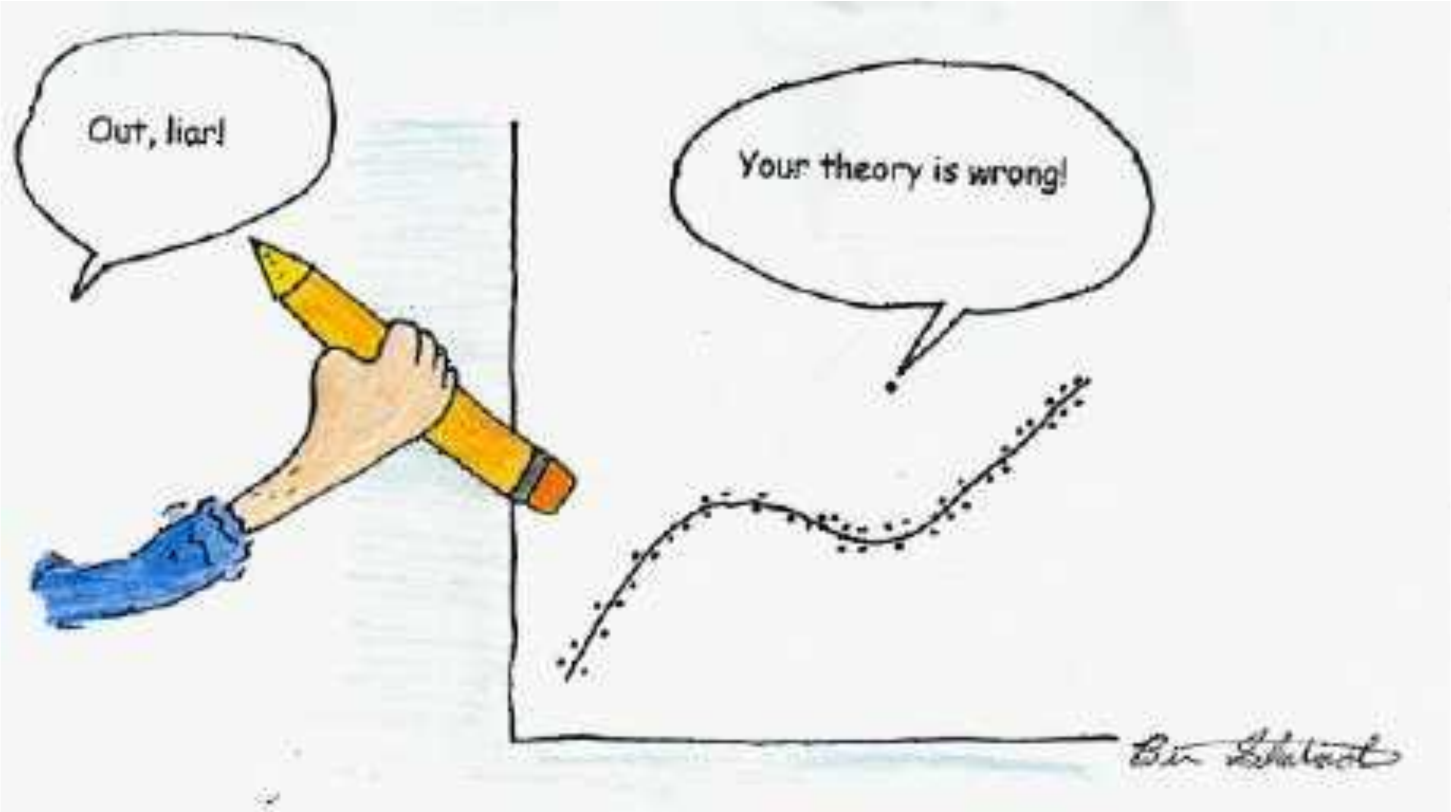
INTRODUCCIÓN A ESTADÍSTICA

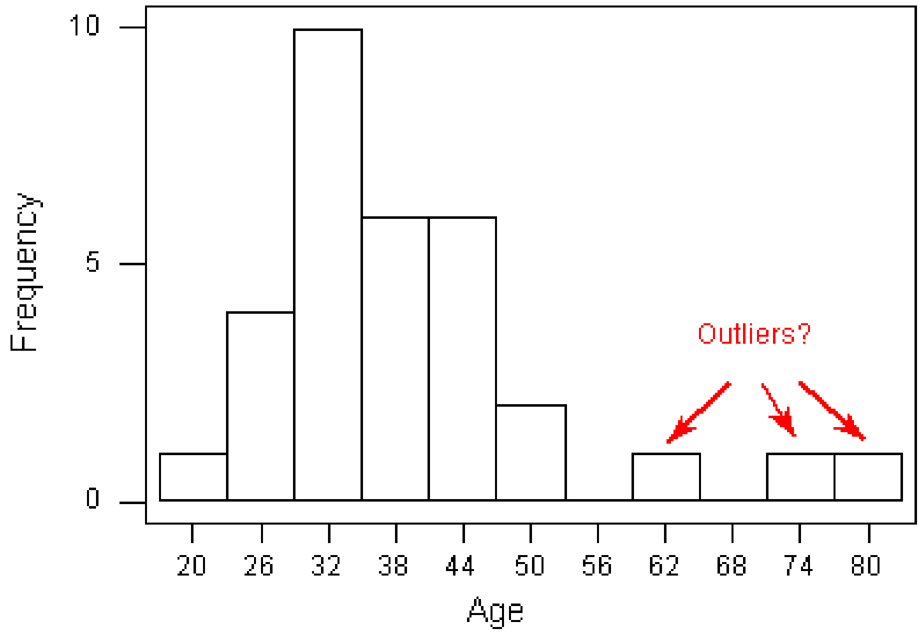
Mayo 2018

DATOS ATÍPICOS



EDADES

Nombre	Edad
José	22
María	21
Marcos	24
Susana	20
Tomás	211
Jacobo	23

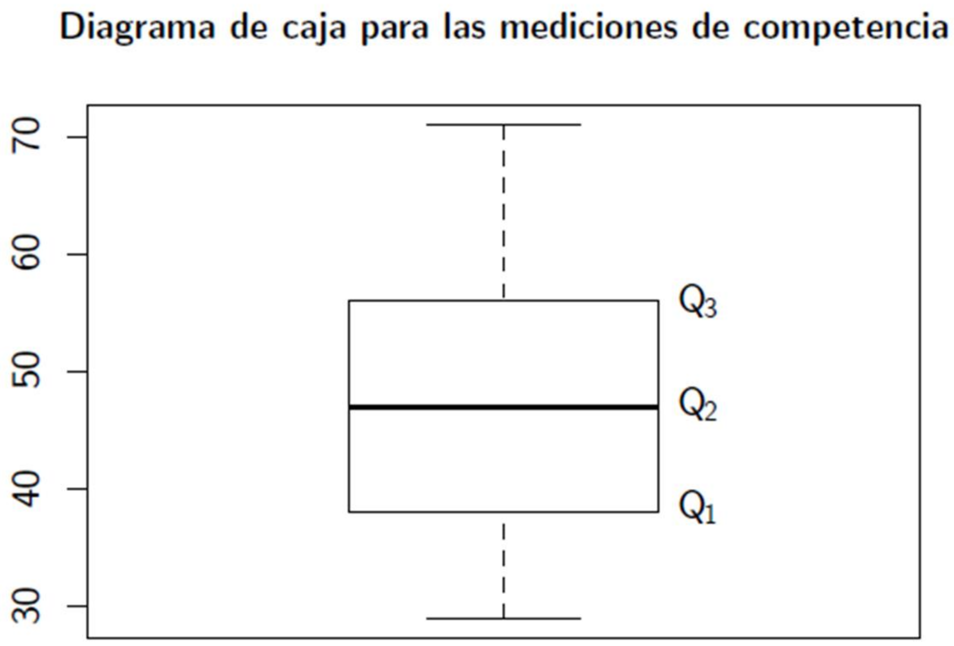


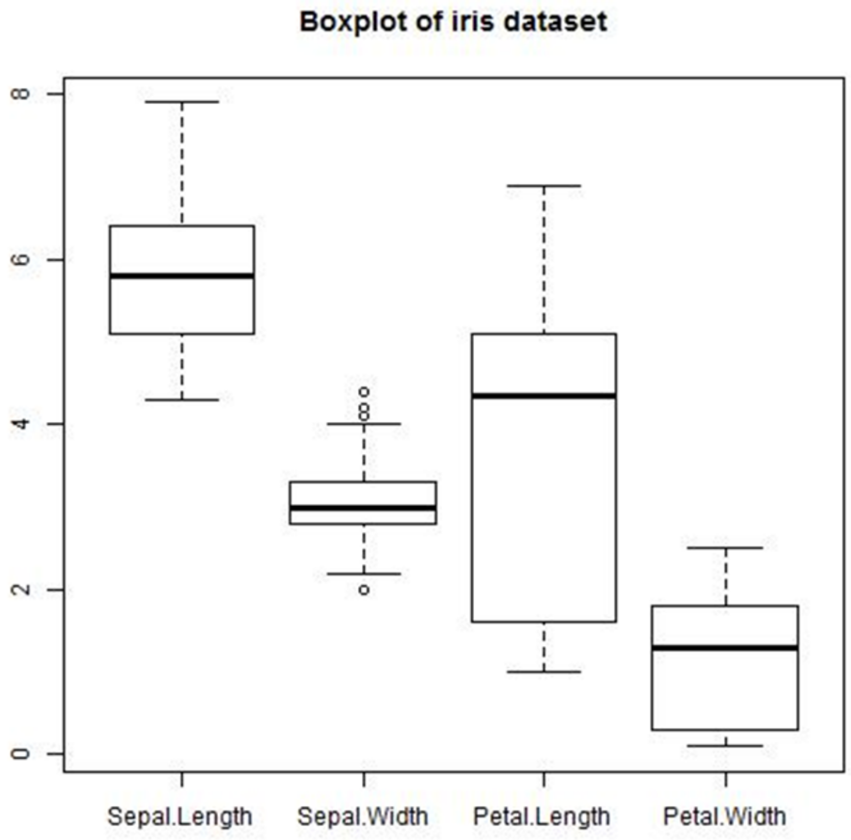
- ¿Existen datos atípicos?.

CUANTILES

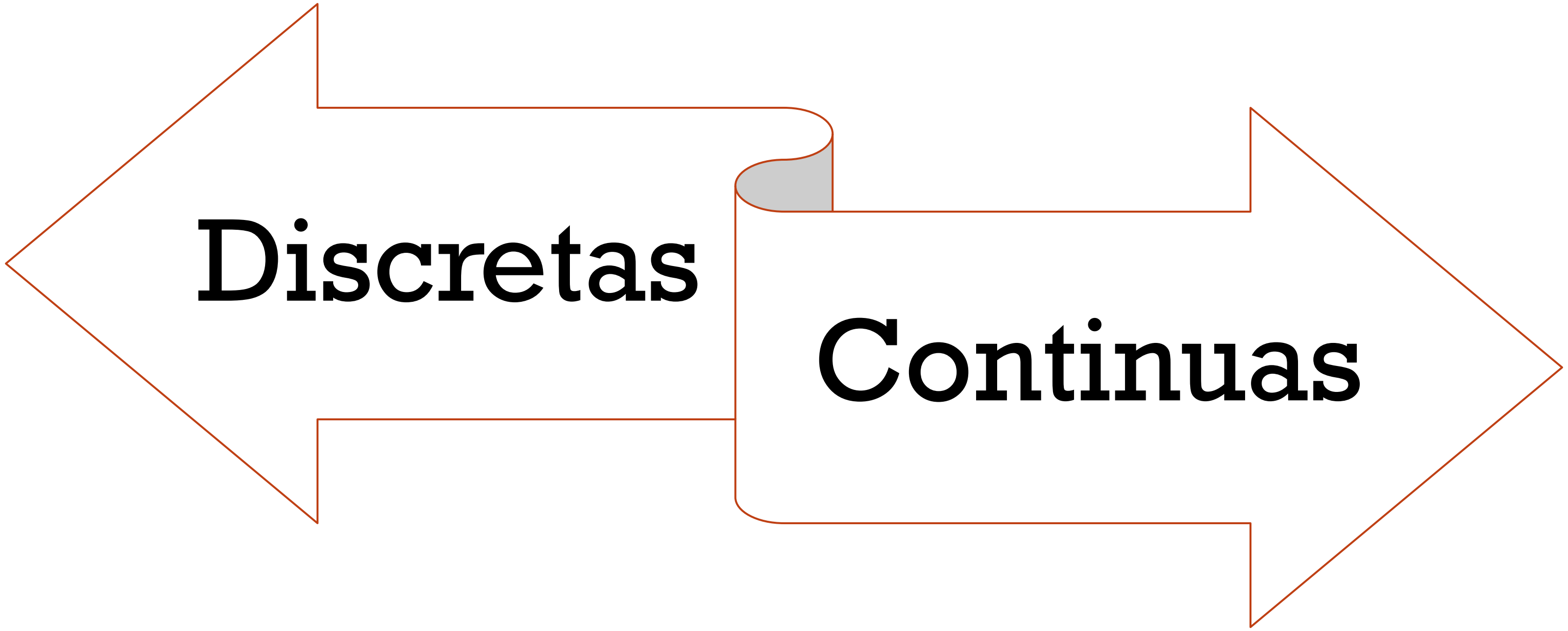
- Los cuantiles son medidas de localización que dividen un conjunto de datos ordenados en cierto número de grupos o partes que contienen la misma cantidad de datos.
- De manera más formal, sea $x_1, x_2, x_3, \dots, x_n$ un conjunto de n mediciones ordenadas en forma creciente, se define su percentil p como el valor x tal que $p\%$ de las mediciones es menor o igual a x, y el $(100-p)\%$ mayor o igual.
- Al percentil 25 también se le conoce como $primer\ curatil\ o\ cuartil\ inferior, C_i$; mientras que la mediana que es el percentil 50 corresponde al $cuartil\ medio, C_m$; y el percentil 75 es $cuartil\ superior\ o\ tercer\ cuartil, C_s$

DIAGRAMA DE CAJA





TIPOS DE VARIABLES



LANZAR UNA MONEDA

- Si lanzamos dos monedas, cuantos casos tenemos en que el # de caras sea igual al # de colas.
- Si lanzamos cuatro monedas, cuantos casos tenemos en que el # de caras sea igual al # de colas.
- Ahora bien, si lanzamos cinco monedas, cuantos casos tenemos en que obtengamos 2 caras.
- Y si lanzamos cinco monedas, cuantos casos tenemos en que obtengamos 3 caras.
- Y si lanzamos cinco monedas, cuantos casos tenemos en que obtengamos 5 caras.
- Si lanzáramos 125 monedas, ¿cuantos casos tendríamos con 3 caras?
- ¿Cuál es la probabilidad de que al lanzar cinco monedas, obtengamos una sola cara?
- ¿Cuál es la probabilidad de que al lanzar cinco monedas, obtengamos 3 caras?

DISTRIBUCIÓN BINOMIAL

- Se dice que una variable aleatoria X tiene una distribución binomial basada en n pruebas con probabilidad p de éxito si y sólo si:

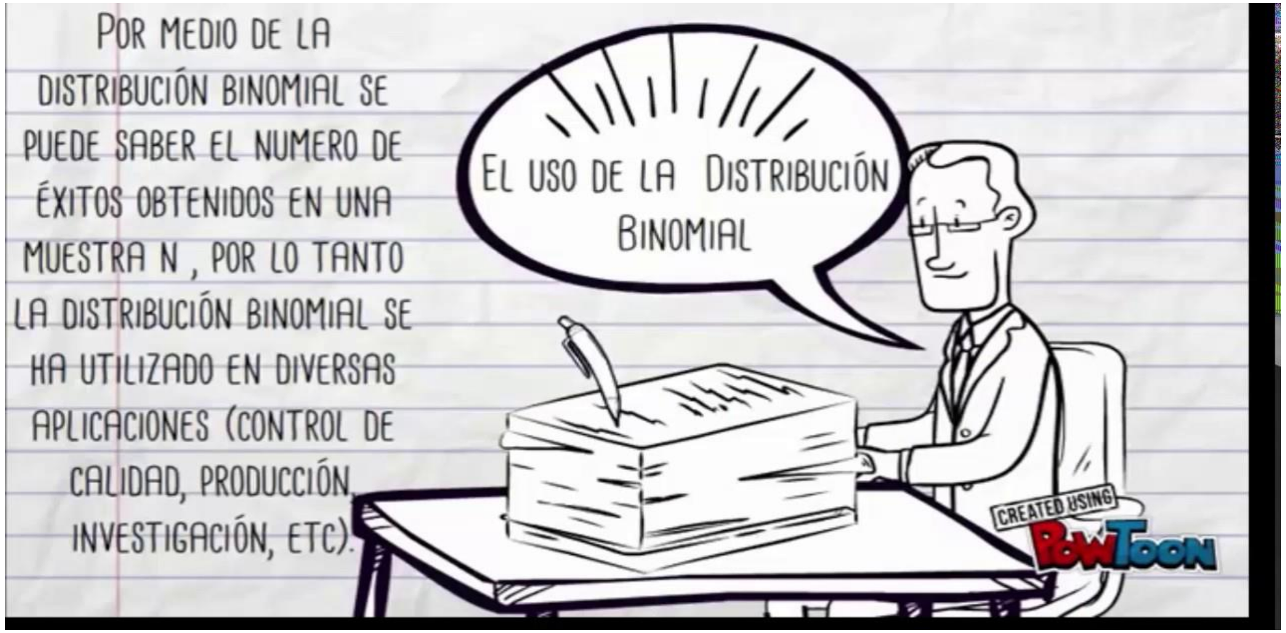
$$P(x) = \binom{n}{x} p^x q^{n-x}, x = 0,1,2,\ldots n \text{ y } 0 \leq p \leq 1.$$

- Sea X una variable aleatoria binomial basada en n pruebas y probabilidad p de éxito. Entonces:

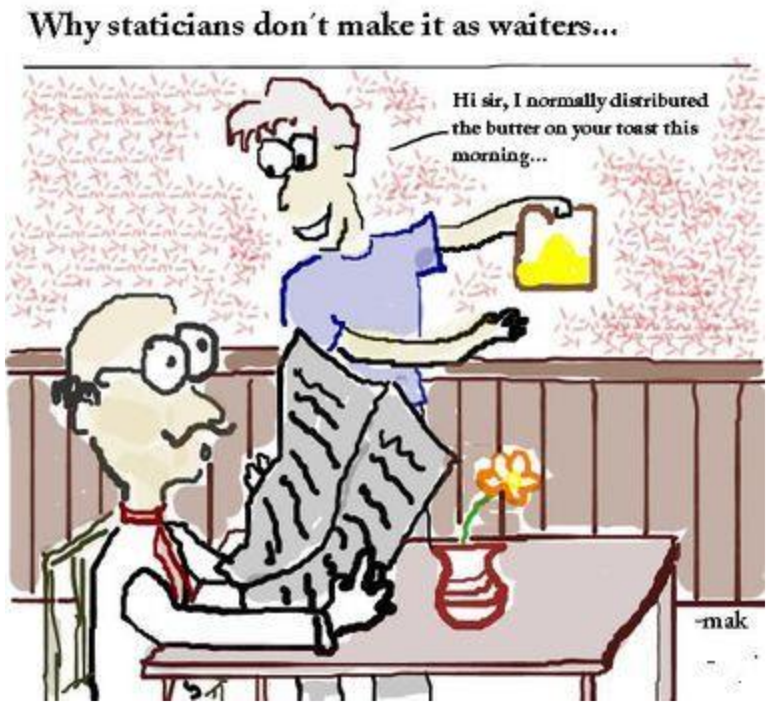
$$\mu = np \text{ y } \sigma^2 = npq$$

PIEZAS DEFECTUOSAS

- Actividad 3: Suponga que un lote de 5,000 fusibles eléctricos contiene 5% de piezas defectuosas. Si se prueba una muestra de 5 fusibles, encuentre la probabilidad de hallar al menos uno defectuoso.



TEOREMA DEL LÍMITE CENTRAL



DISTRIBUCIÓN NORMAL

- Una v.a. X continua se distribuye Normal con media (poblacional) μ y varianza (poblacional) σ^2, si su función de densidad esta dada por

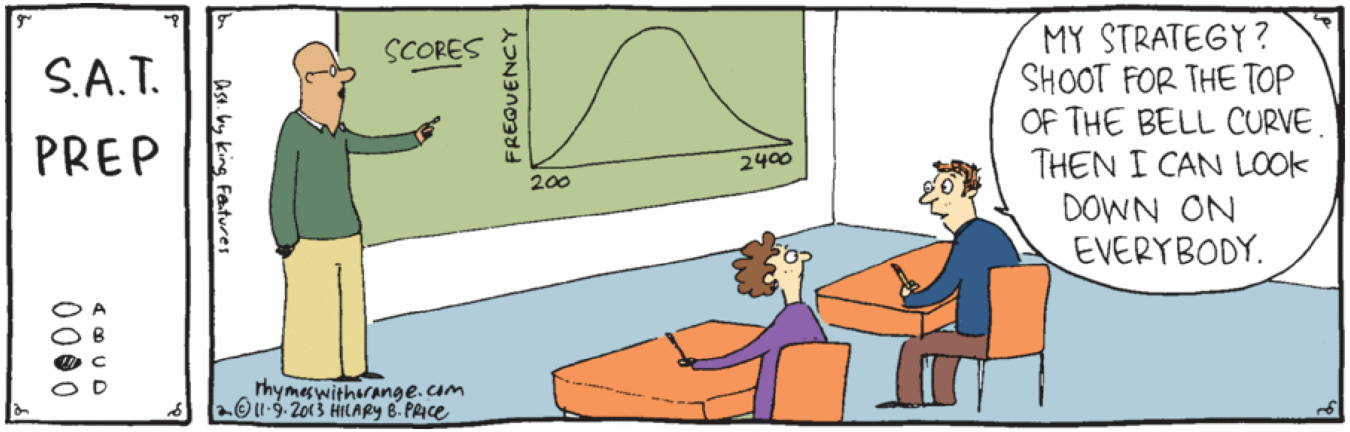
$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, x, \mu \in \mathbb{R}, \sigma^2 > 0$$

1. Muchas conjuntos de datos se pueden modelar adecuadamente con esta distribución.
2. El promedio muestra, cuando la muestra es grande, se distribuye normalmente.
3. Es la base de muchas técnicas de inferencia estadística.

- Si $X \sim N(\mu,\sigma$^2) entonces $Z=(x-\mu)/\sigma \sim N(0,1)$

CALIFICACIONES

- **Actividad 6: Las calificaciones para un examen de admisión a una universidad están normalmente distribuidas con media de 75 y desviación estándar 10. ¿Que fracción de las calificaciones se encuentra entre 80 y 90?**

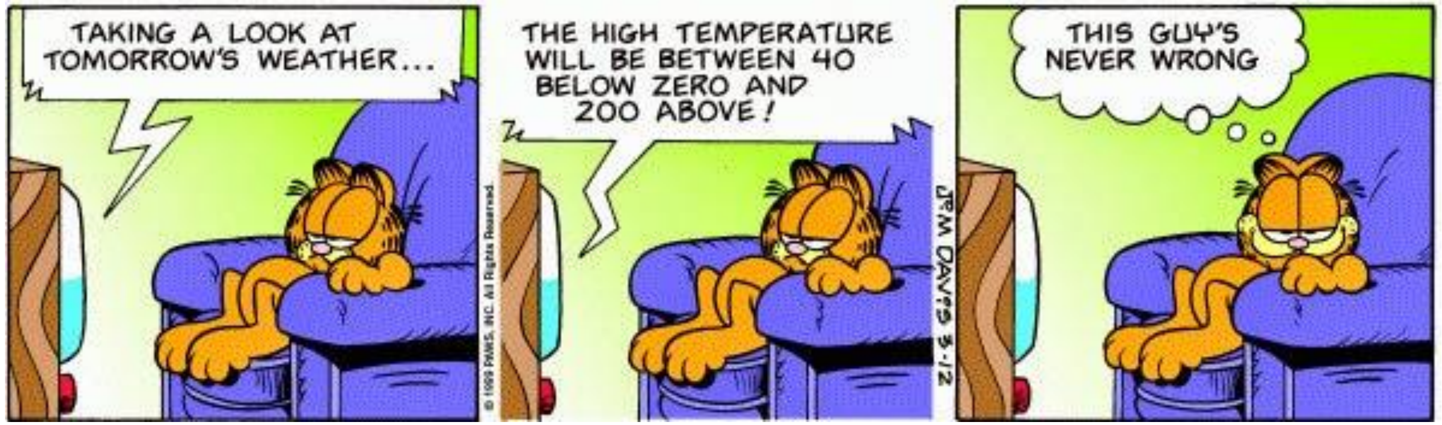


NORMAL ESTÁNDAR

- Si $X \sim N(\mu, \sigma^2)$ entonces $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$
- Es simétrica alrededor de la media y unimodal, es decir, la media, la mediana y la moda coinciden μ.
- En consecuencia:
  - $P(X \leq \mu) = P(X \geq \mu) = \frac{1}{2}$.
  - $P(Z \geq z) = P(Z \leq -z)$.

INTERVALO DE CONFIANZA

- Un intervalo de confianza del $(1-\alpha)*100$, es un conjunto de valores de la forma [a, b] el cual tiene una probabilidad asociada de $(1-\alpha)$ de contener el valor del parámetro poblacional.
- A a y b se les llama límite superior y límite inferior respectivamente



PROBABILIDAD DE INTERVALOS DE CONFIANZA

- Todos los intervalos de confianza se interpretan como un intervalo de valores los cuales tienen cierta probabilidad de contener al parámetro y no como intervalos con cierta probabilidad de que el parámetro de encuentre entre sus limites.
- Un intervalo de confianza nos brinda una idea de la dispersión (o varianza) del estimador. Cuando la varianza aumenta, el intervalo se ensancha y la estimación es menos precisa.
- Considerando un cierto nivel de confianza, la amplitud del intervalo depende de dos factores:
  - La variabilidad de las observaciones.
  - El tamaño de la muestra.

REGLA EMPÍRICA

- Muchas distribuciones de datos de la vida real se pueden aproximar por medio de una distribución normal. Tienen características definidas de variación, como se expresa en el enunciado siguiente.
- Regla empírica
- Para una distribución de mediciones que sea aproximadamente normal (forma de campana), se deduce que el intervalo con puntos extremos
  - $\mu \pm \sigma$ contiene aproximadamente 68% de las mediciones.
  - $\mu \pm 2\sigma$ contiene aproximadamente 95% de las mediciones.
  - $\mu \pm 3\sigma$ contiene casi todas las mediciones.

EXPERIMENTO DEL TÉ

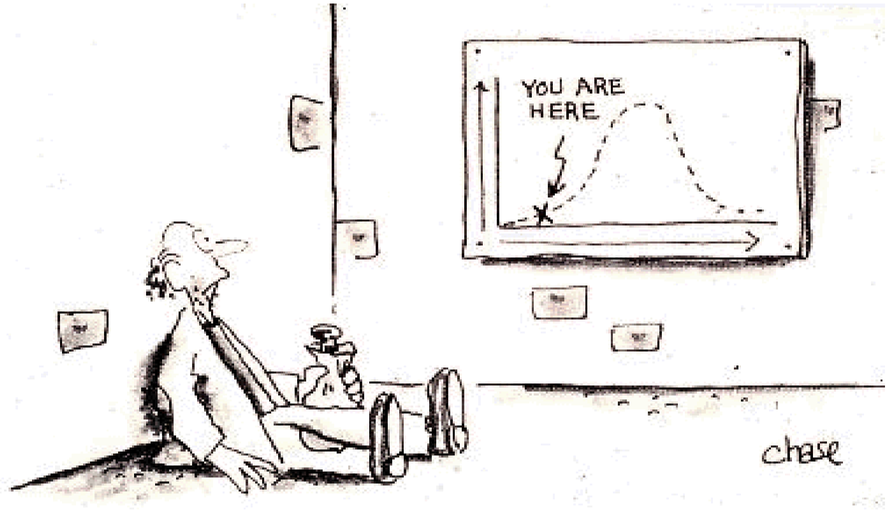
- Fisher diseño el experimento para probar si Ms. B. podía distinguir entre distintas tazas de té, unas hechas con leche primero y otras con el té primero, tomando en cuenta 3 factores:
  - Control de variabilidad aleatoria
  - ¿Cuántas tazas debe probar Ms. B.? ¿Deben ser pareadas? ¿En qué orden deben ser presentadas?
  - ¿Qué conclusión debe hacerse si no se equivoca ni una vez al identificar el orden de preparación?

HIPÓTESIS

- Hipótesis. Una hipótesis es un enunciado acerca de un parámetro poblacional. Ejemplo: H:p=0.5.
- Hipótesis nula H_0. Es una hipótesis que representa el estado de conocimientos previo a nuestra investigación.
- Hipótesis alternativa H_A. Es una hipótesis que representa la hipótesis de trabajo del investigador.
- Estadístico de prueba: Es un estimador, o alguna operación con uno o más estimadores, que usamos para determinar la plausibilidad de la hipótesis nula.

REGIÓN DE RECHAZO

- Una prueba de hipótesis rechaza la hipótesis nula si el estadístico de prueba calculado a partir de la muestra que se ha recolectado toma un valor en la región de rechazo. Si el valor del estadístico de prueba calculado está en la región de no-rechazo, la hipótesis nula se mantiene.



METODOLOGÍA

- Los pasos que seguimos anteriormente en el ejemplo, se pueden resumir en los siguientes puntos, aplicables en cualquier prueba de hipótesis:

1. Planteamiento de las hipótesis nula y alternativa.
2. Selección del estadístico de prueba.
3. Determinación del nivel de significancia y la región de rechazo.
4. Cálculo del estadístico de prueba.
5. Decisión.
6. Conclusión.

EJEMPLO: PROBLEMA

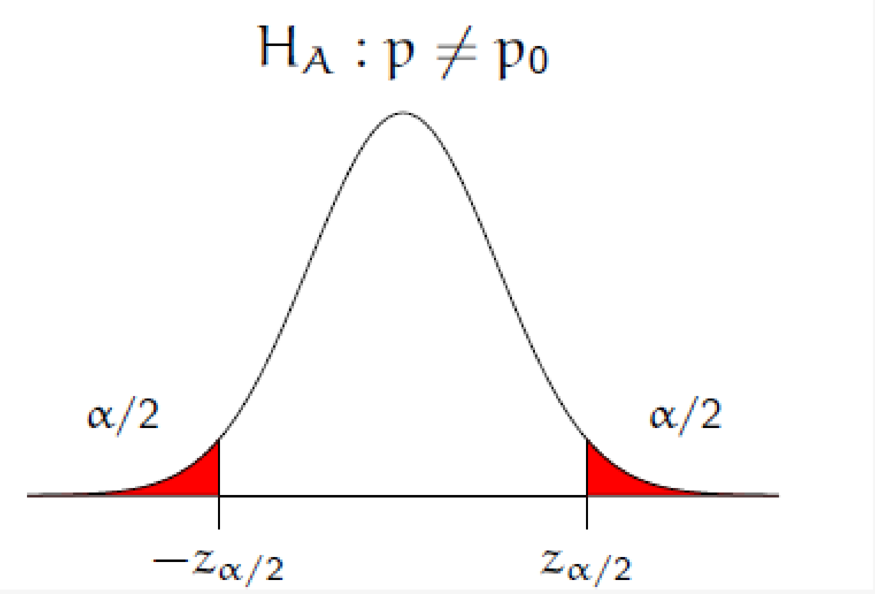
- Supongamos que deseamos verificar la afirmación de la OMS de que la proporción de individuos que han padecido algún trastorno mental en su vida es $p = 0.3$. Deseamos saber sí esta afirmación es cierta para la población de Monterrey.
- Antes de llevar a cabo cualquier recopilación de información, la situación que prevalece es que $p = 0.3$, por lo tanto $H_0: p = 0.3$.
- Dado que no tenemos más información o un objetivo más preciso, rechazaríamos la hipótesis nula usando evidencia que indique que $p > 0.3$ ó $p < 0.3$. Por lo tanto, $H_A: p \neq 0.3$.

EJEMPLO: ESTADÍSTICO

- Es mejor trabajar con $Z = \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{pq}}$, porque este estadístico
- Ahora, es necesario fijar el nivel de significancia de la prueba. Ésta es decisión del investigador y depende qué tan estricto o laxo quiera ser con su procedimiento. Un de 0.05 (5%) es lo más usado, aunque para ser más estrictos se puede usar = 0.01 (nos equivocamos menos veces si rechazamos H_0). Fijemos, el en 0.05.
- Fijemos la región de rechazo usando la distribución del estadístico de prueba, el nivel y la hipótesis alternativa.

EJEMPLO: REGIÓN DE RECHAZO

- De la grafica se pueden inferir la región de rechazo corresponde con los valores de Z que están por arriba de $z_{\alpha/2}$ y por debajo de $-z_{\alpha/2}$. (1.96 y -1.96 por simetría)



EJEMPLO: CÁLCULOS

- Supongamos que obtenemos una muestra de tamaño 100 de habitantes de Monterrey mayores a 50 años. Consideramos este tamaño de muestra grande, porque usamos un estadístico de prueba cuya distribución es $N(0,1)$ cuando n es grande.
- Calculamos el estadístico de prueba a partir de la muestra. Supongamos que en la muestra hay 18 individuos que han presentado algún trastorno mental en su vida, es decir, $\hat{p} = 0.18$. Como el valor hipotético de $p = 0.3$, el estadístico queda:

$$Z = \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{pq}} = \frac{\sqrt{100}(0.18 - 0.3)}{\sqrt{0.3 * 0.7}} = -2.619$$

EJEMPLO: CONCLUSIONES

- Comparamos el estadístico de prueba obtenido con los **puntos críticos** $z_{\alpha/2}$ y $-z_{\alpha/2}$. Como $z_{obt} = -2.619 < -1.96 = -z_{\alpha/2}$, rechazamos $H_0: p = 0.3$.
- Concluimos que existe evidencia estadística que la población de Monterrey no presenta un 30% de individuos (mayores a 50 años) que hayan padecido algún trastorno mental, sino que es menor.

1. Una prueba como ésta tiene regiones críticas o de rechazo en las dos colas, por lo que recibe el nombre de prueba de dos colas.
2. La regla de decisión queda:

$$Rechazar\ H_0\ si\ |z_{obt}| > z_{\alpha/2}.$$

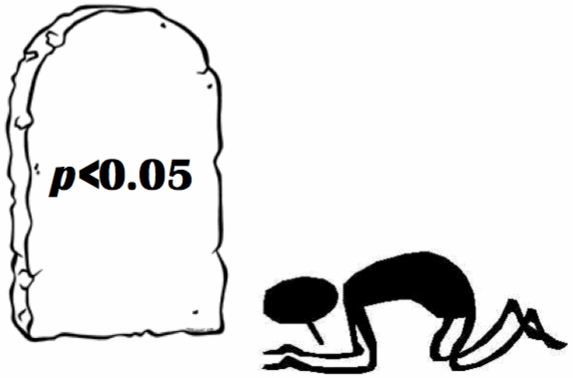
P-VALOR

- Existe una equivalencia entre la decisión de rechazar (o no) H_0 y el p-valor:

$$\text{Si } p\text{–}valor \leq \alpha \text{ entonces se rechaza } H_0$$

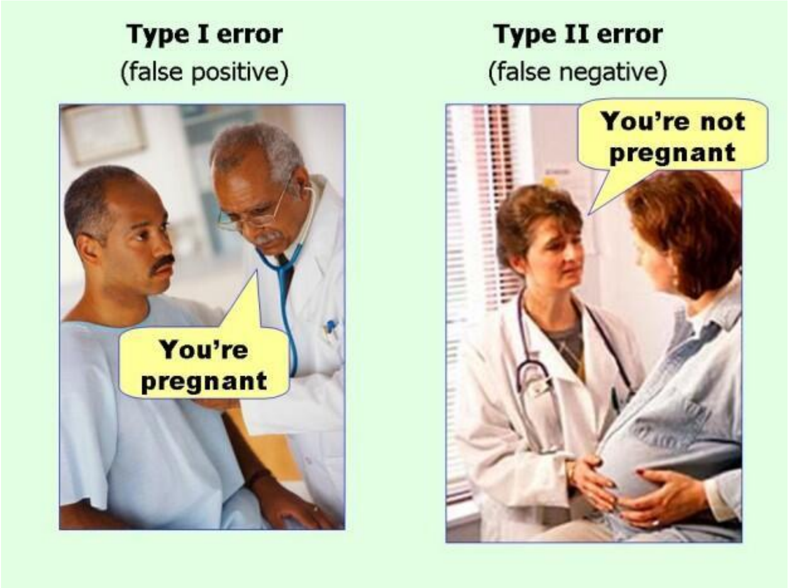
$$\text{Si } p\text{–}valor > \alpha \text{ entonces no se rechaza" } H_0$$

- El p-valor no es la probabilidad de que H_0 sea cierta.

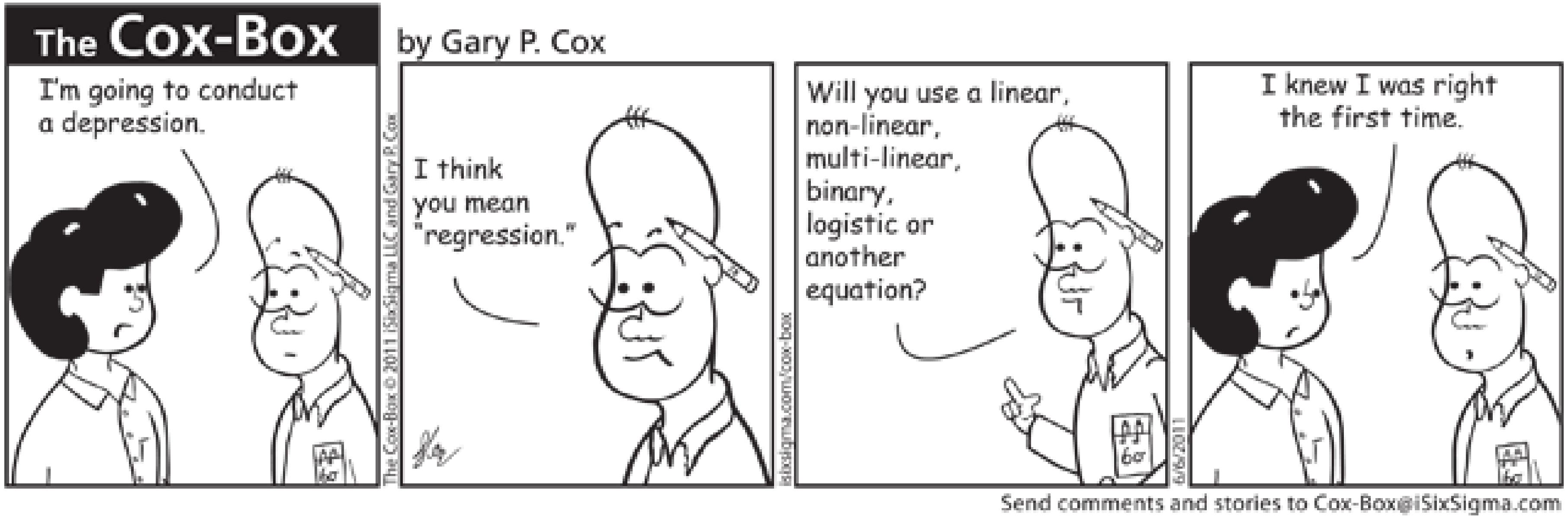


ERRORES

- La probabilidad de rechazar H_0 cuando es cierta es el α de la prueba, nivel de significancia.
- La probabilidad de no rechazar H_0 cuando es falsa es la β de la prueba.



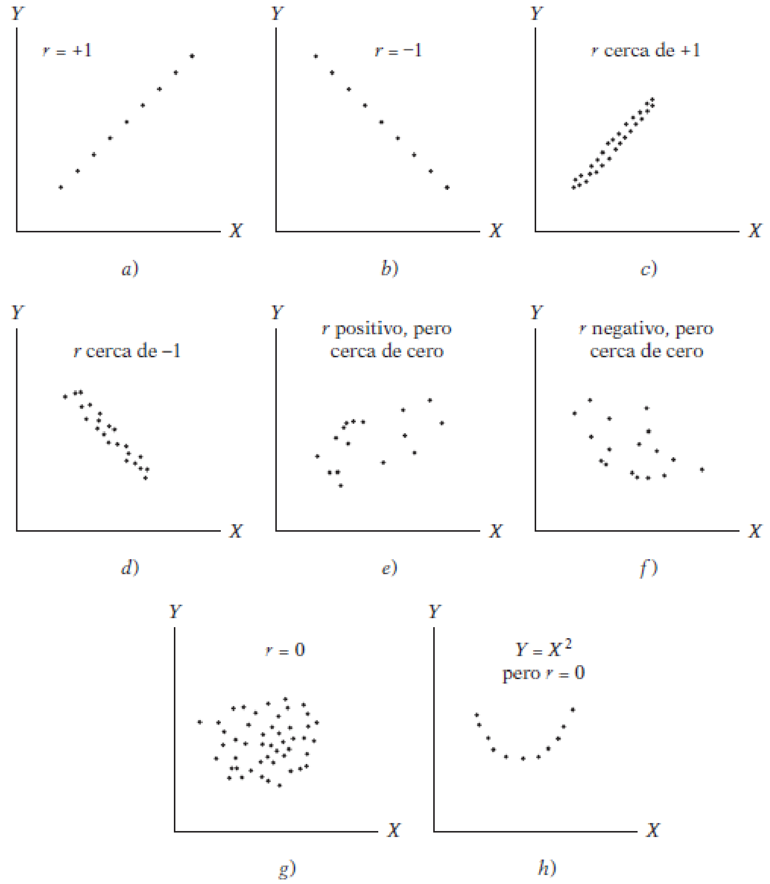
REGRESIÓN



The Cox-Box
by Gary P. Cox
I'm going to conduct a depression.
I think you mean "regression."
Will you use a linear, non-linear, multi-linear, binary, logistic or another equation?
I knew I was right the first time.
The Cox-Box © 2011 iSixSigma LLC and Gary P. Cox
isixsigma.com/cox-box
6/6/2011
Send comments and stories to Cox-Box@iSixSigma.com

REGRESIÓN Y CORRELACIÓN

PATRONES DE CORRELACIÓN



MODELO DE REGRESIÓN

- Consideremos el problema de modelar el nivel medio de una variable Y a partir de los niveles de otra variable x.
- Cuando modelamos el nivel medio de una variable atendiendo a su relación lineal con otra, consideramos una regresión lineal.

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

- donde $e_i \sim N(0, \sigma^2), i = 1, \ldots, n$ e independientes entre sí.
- La formula $\beta_0 + \beta_1 x_i$ describe una recta, la recta de regresión.
- El componente estadístico esta dado por los errores e_i's, que se distribuyen normalmente alrededor de la recta.

SUPUESTOS

- El modelo de regresión lineal plantea 7 supuestos:.

1. Supuesto 1: El modelo de regresión es **lineal en los parámetros**, aunque puede no serlo en las variables, es decir, el modelo que se muestra en la ecuación $Y_i = \beta_0 + \beta_1 X_i + e_i$, puede extenderse para incluir más variable explicativas.
2. Supuesto 2: Valores fijos de X, o valores de X independientes del término de error.
3. El valor medio de los errores es 0, $E(e_i) = 0$

SUPUESTOS

4. Supuesto 4: La varianza del error es la misma sin importar el nivel de la variable independiente, σ^2 (Homoscedasticidad), $Var(e_i) = E\left[e_i^2 \middle| X_i\right] = \sigma^2$
5. Supuesto 5: No existe auto correlación entre los errores, $Cov(e_i, e_j | X_i, X_j) = 0$ (No autocorrelación)
6. Supuesto 6: El número de observaciones n debe ser mayor al número de parámetros por estimar.
7. Supuesto 7: No todos los valores de la variable independientes deben ser iguales y no deben de existir datos atípicos.

MÍNIMOS CUADRADOS

- Recordemos que el modelo de regresión está dado por:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

- Sin embargo, la gran mayoría de las veces trabajamos con muestras, por lo tanto, el modelo de regresión muestral está dado por

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i = \hat{Y}_i + \hat{e}_i$$

$$\Rightarrow \hat{e}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- Con este objetivo nos interesa que la suma de los errores sea lo menor posible, es decir:

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)$$

- Sin embargo, aunque este criterio es intuitivamente atractivo, no es muy bueno ya que la sumatoria tenderá a 0 independientemente de peso especifico de cada $\hat{e}_i$.

MELI

- Se dice que el estimador de Mínimos Cuadrados Ordinario (MCO) es el Mejor Estimador Lineal Insesgado (MELI) si cumple con lo siguiente:

1. Es lineal.
2. Es insesgado.
3. Es eficiente (varianza mínima)

- Teorema Gauss-Markov: Dados los supuestos del modelo clásico de regresión lineal, los estimadores de mínimos cuadrados, dentro de la clase de estimadores lineales insesgados, tienen varianza mínima, es decir, son MELI.