Presentación TP2

Base de Datos
DC - FCEN - UBA

2do Cuat 2014

Conceptos útiles para el tp

1. Media, Desvio estándar
2. Histograma, Error Bar
3. Distribuciones Uniforme y Normal
4. Test de hipótesis

Comparen las dos poblaciones/series de valores

a=[27, 92, 41, 11, 84, 94, 56, 4, 74, 4, 92, 11, 99, 30, 66, 87, 50, 58, 70, 5, 88, 41, 35, 97, 48, 91, 74, 62, 11, 31, 75, 1, 17, 91, 48, 74, 93, 18, 24, 26, 32, 29, 8, 60, 22, 18, 40, 63, 70, 72, 66, 53, 50, 31, 29, 39, 5, 99, 99, 91, 18, 74, 5, 100, 5, 96, 33, 22, 91, 28, 35, 2, 97, 63, 13]

b=[24, 21, 32, 56, 51, 2, 10, 82, 38, 52, 84, 15, 18, 60, 16, 59, 85, 94, 66, 7, 44, 45, 92, 31, 43, 50, 80, 48, 95, 7, 98, 62, 78, 3, 77, 74, 46, 47, 35, 17, 16, 81, 99, 99, 31, 63, 33, 48, 31, 62, 2, 91, 84, 4, 70, 73, 42, 57, 67, 39, 14, 100, 28, 58, 40, 92, 8, 80, 94, 22, 97, 24, 87, 29, 58]

Puedo concluir algo?

Comparen las dos poblaciones/series de valores (cont.)

Es poco probable, necesitamos cuantificadores que nos permitan **comparar**, **entender**, etc..

Comparen las dos poblaciones/series de valores (cont.)

Es poco probable, necesitamos cuantificadores que nos permitan **comparar**, **entender**, etc..

(O tener algun super poder, ej:)

Media (aka promedio)

Es una forma de caracterizar una serie de valores

Dado $s = a_1, a_2, ..., a_n$ la media de esta serie de valores esta definida como:

$$media(s) = \frac{1}{n} * \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + ... + a_n}{n}$$

Media (aka promedio)

Es una forma de caracterizar una serie de valores

Dado $s = a_1, a_2, ..., a_n$ la media de esta serie de valores esta definida como:

$$media(s) = \frac{1}{n} * \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + ... + a_n}{n}$$

La media caracteriza bien una serie de valores?

Media (aka promedio)

Es una forma de caracterizar una serie de valores

Dado $s = a_1, a_2, ..., a_n$ la media de esta serie de valores esta definida como:

$$media(s) = \frac{1}{n} * \sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + ... + a_n}{n}$$

La media caracteriza bien una serie de valores?

Es muy sensible a los valores extremos. Otra posibilidad: Mediana

Desvio Estándar (σ, $\sqrt{varianza}$, std)

Desvio Estándar (σ, $\sqrt{varianza}$, std)

Es una forma de caracterizar la disperción que tiene una serie, tratando de medir cuanto se alejan los valores su media.

Dado $s = a_1, a_2, ..., a_n$ el desvio estándar esta definida como:

$$std(s) = \frac{1}{n} * \sum_{i=1}^{n} (a_i - media(s))^2$$

Desvio Estándar (σ, $\sqrt{varianza}$, std)

Es una forma de caracterizar la disperción que tiene una serie, tratando de medir cuanto se alejan los valores su media.

Dado $s = a_1, a_2, ..., a_n$ el desvio estándar esta definida como:

$$std(s) = \frac{1}{n} * \sum_{i=1}^{n} (a_i - media(s))^2$$

Ejemplo, sea:
$a = [0, 0, 0, 0, 0, 0, 0]$ y $b = [30, 20, 10, 5, 0, -5, -10, -20, -30]$

Desvio Estándar (σ, $\sqrt{varianza}$, std)

Es una forma de caracterizar la disperción que tiene una serie, tratando de medir cuanto se alejan los valores su media.

Dado $s = a_1, a_2, ..., a_n$ el desvio estándar esta definida como:

$$std(s) = \frac{1}{n} * \sum_{i=1}^{n} (a_i - media(s))^2$$

Ejemplo, sea:
$a = [0, 0, 0, 0, 0, 0, 0]$ y $b = [30, 20, 10, 5, 0, -5, -10, -20, -30]$

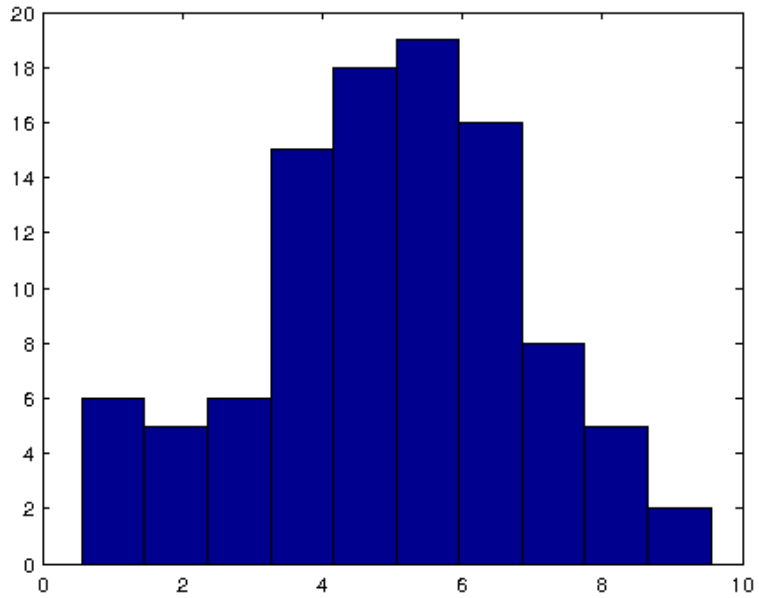
Las medias de a y de b son iguales (0), sin embargo:
$std(a) = 0$ y $std(b) = 17{,}795$

Histograma

Histograma

Representación gráfica de una variable en forma de barras, donde **la altura de cada barra es proporcional a la frecuencia de los valores** representados.

Ej: Serie de 100 valores, bins = [[0,1),[1,2),...[9,10)]

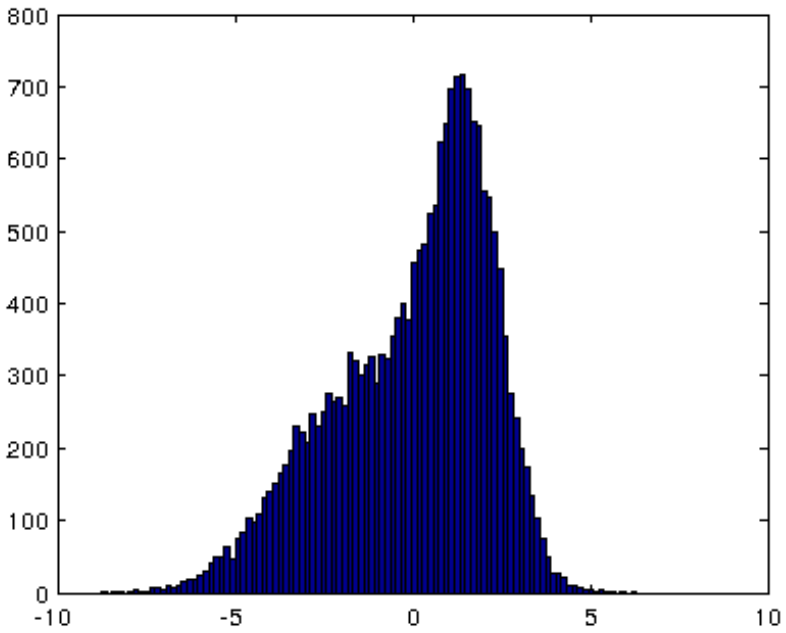


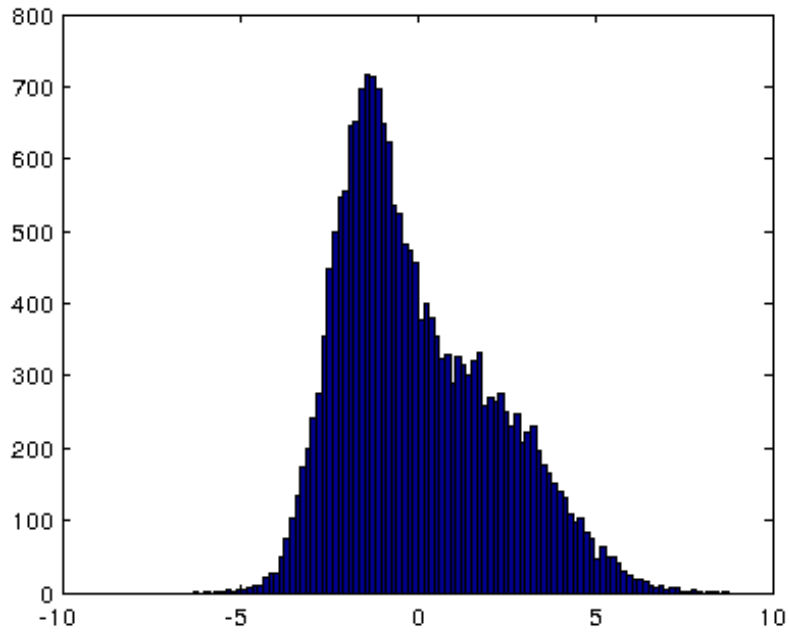
Histograma

Dos series, misma media, mismo desvio estándar, pueden tener diferente histograma?

Histograma

Dos series, misma media, mismo desvio estándar, pueden tener diferente histograma?

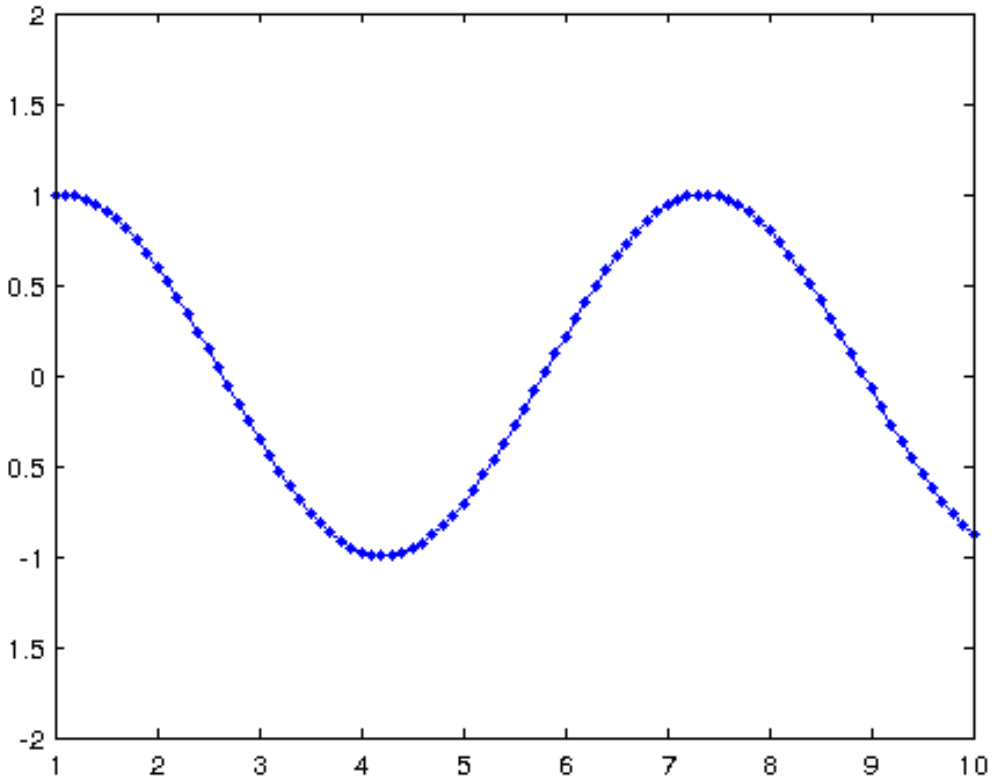




Mirar el histograma de datos reales nos ayuda a entender cómo son los datos

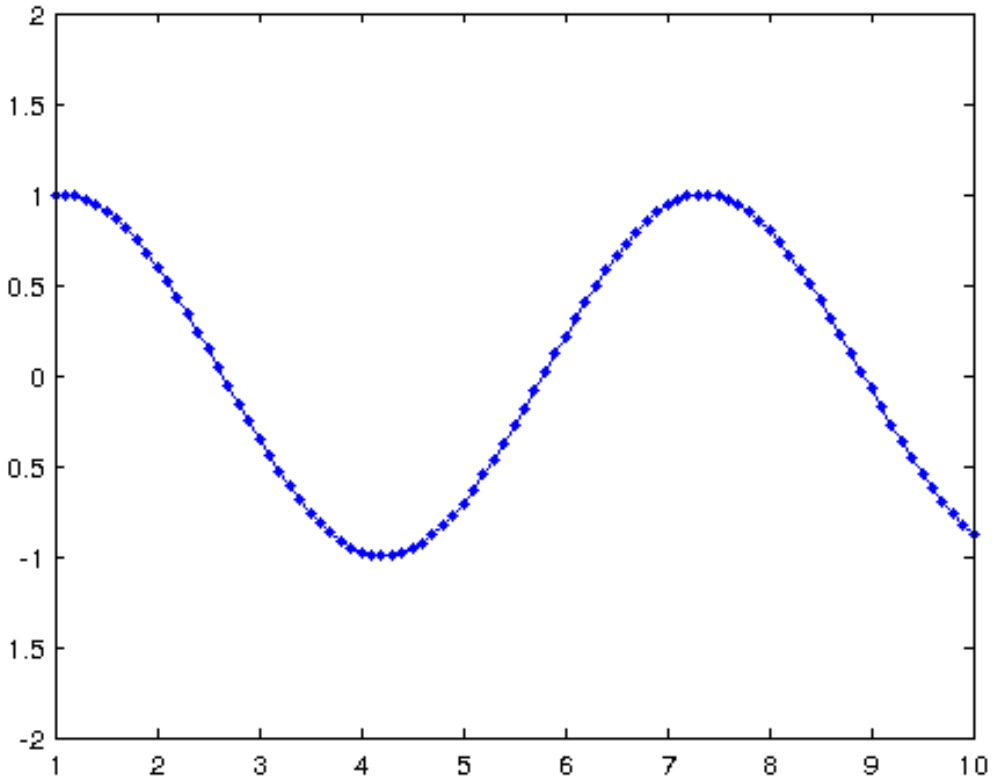
Error bar

Cada punto representa la media de los valores de alguna población para su valor en el eje x. Es **informativo** este grafico?



Error bar

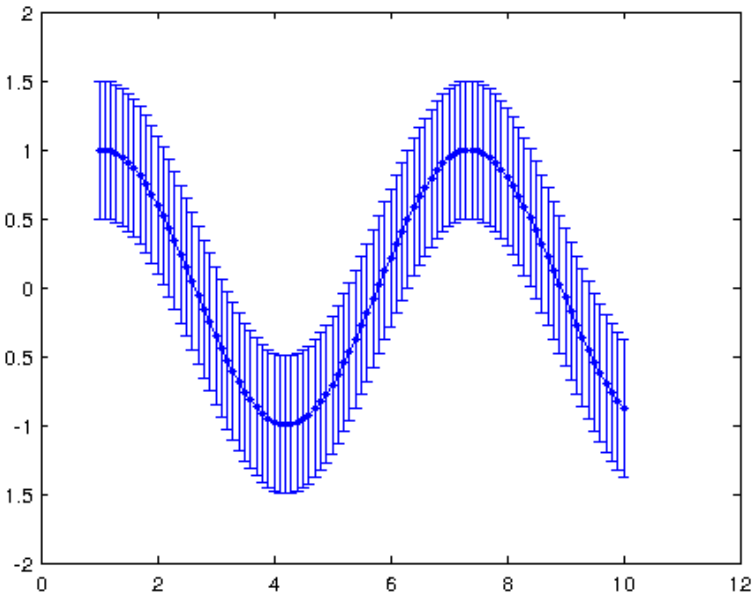
Cada punto representa la media de los valores de alguna población para su valor en el eje x. Es **informativo** este grafico?

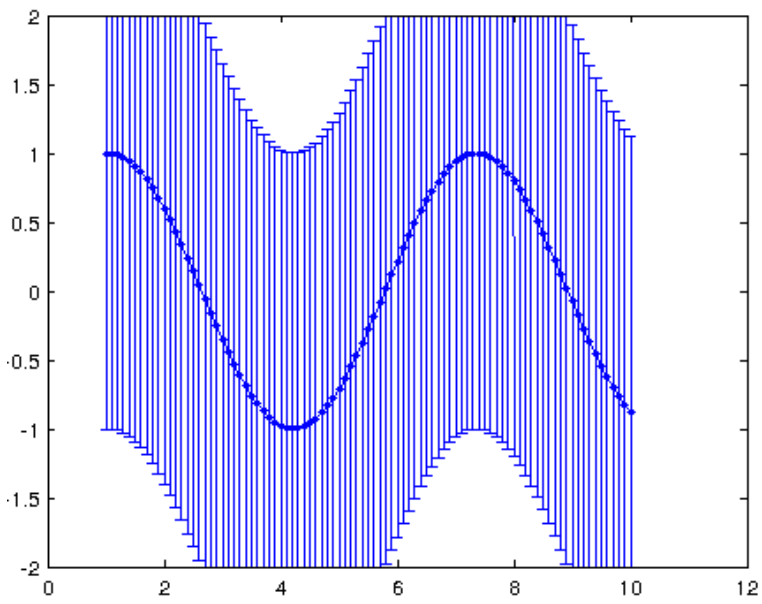


Qué pasa con el desvio? Con la cantidad de muestras que representa cada punto?

Error bar

El Error Bar agrega al gráfico anterior unas barres de *errores* o *incertidumbre*, que aporta más intuición a lo que estamos viendo.

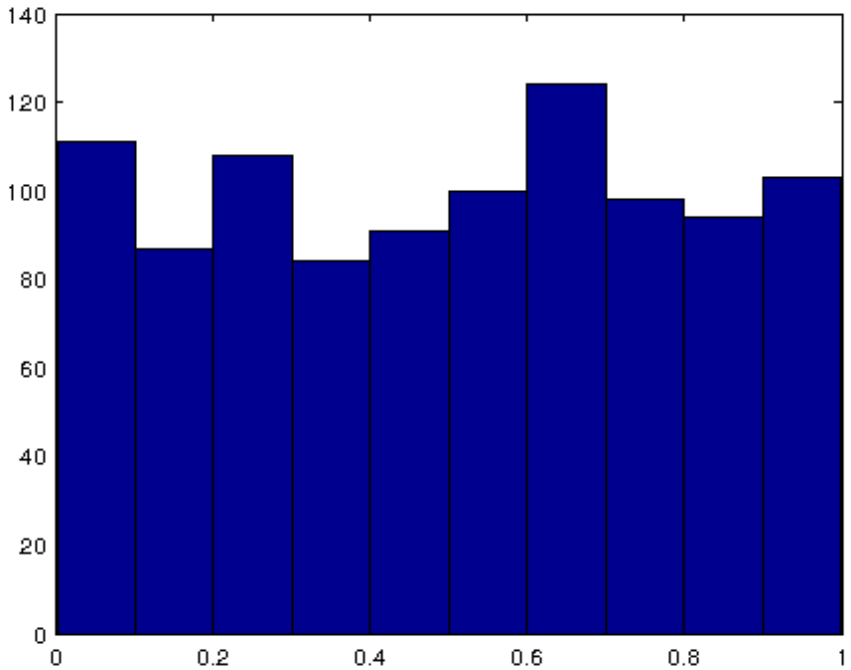




Distribución Uniforme

Distribución Uniforme

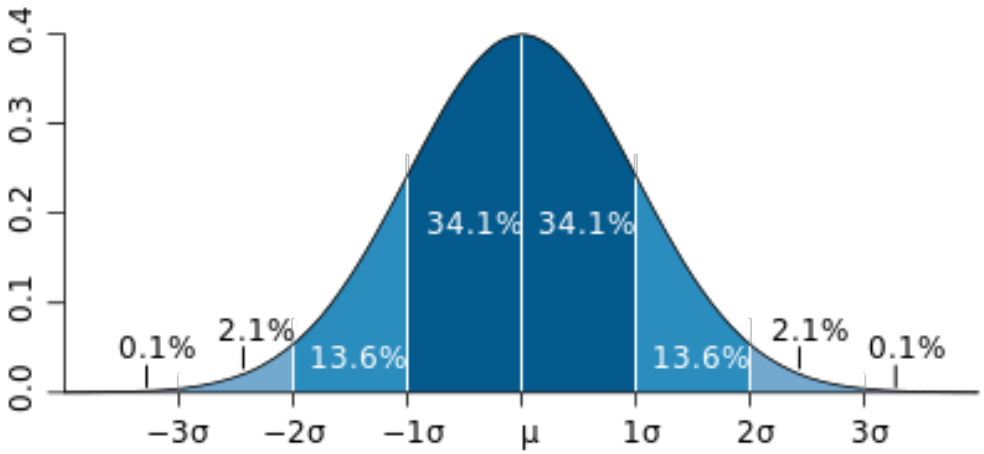
Representa un tipo de distribución donde para cualquier subintervalo del mismo tamaño la frecuencia es la misma.



Ej: El último dígito del numero de la puerta de su casa

Distribución Normal (gaussiana)

Representa un tipo de distribución que tiene una típica forma acampanada. Se define por su media μ y su varianza σ^2.



Ej: El tiempo en que tardan habitualmente para llegar a la facultad

Fuente imagen: `http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram_micro.svg`

Test de hipótesis

Test de hipótesis

¿Cómo cuantificamos la significancia de un fenómeno?

Test de hipótesis

¿Cómo cuantificamos la significancia de un fenómeno?

Test de hipótesis

- Hipótesis Nula (H_0): queremos rechazarla
- Hipótesis Alternativa (H_1):

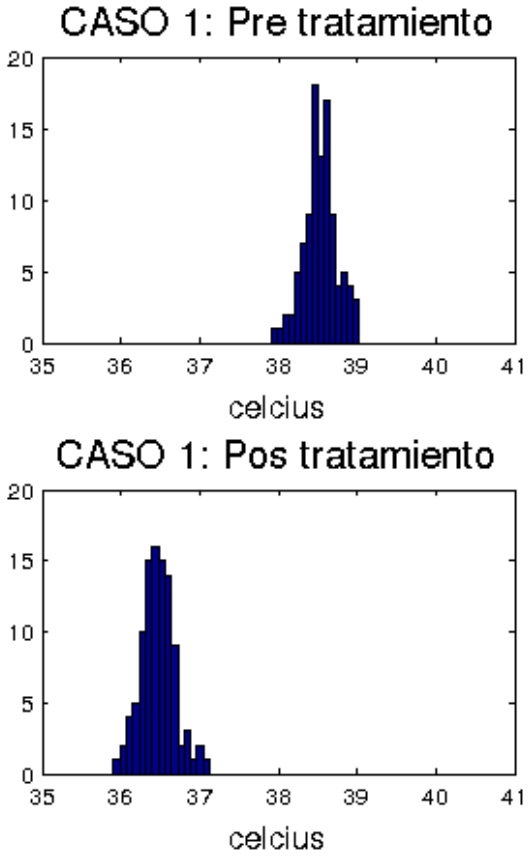
Vayamos a un ejemplo y como se usa

Test de hipótesis

Ejemplo: Un laboratorio esta probando un fármaco para bajar la temperatura corporal de los pacientes. Para esto, toma la temperatura a 100 pacientes antes de del tratamiento y luego.

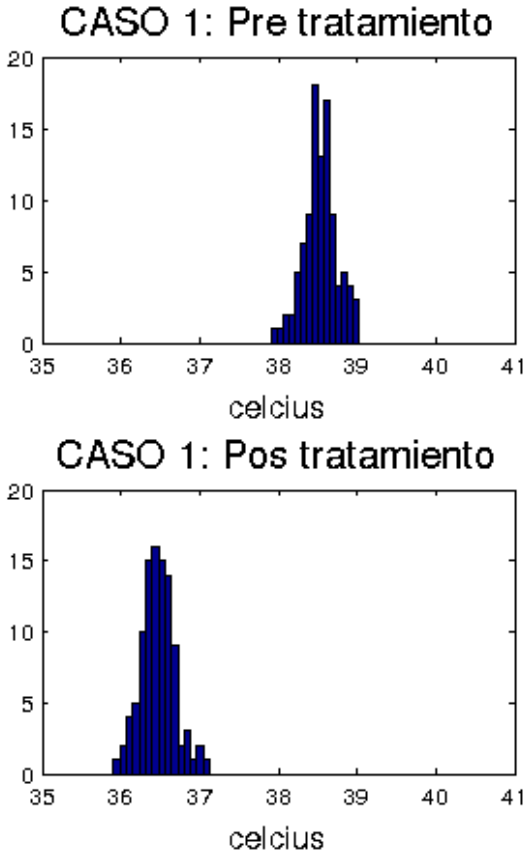
- H_0: El fármaco es inocuo para bajar la fiebre
- H_1: El fármaco baja la fiebre

Test de hipótesis



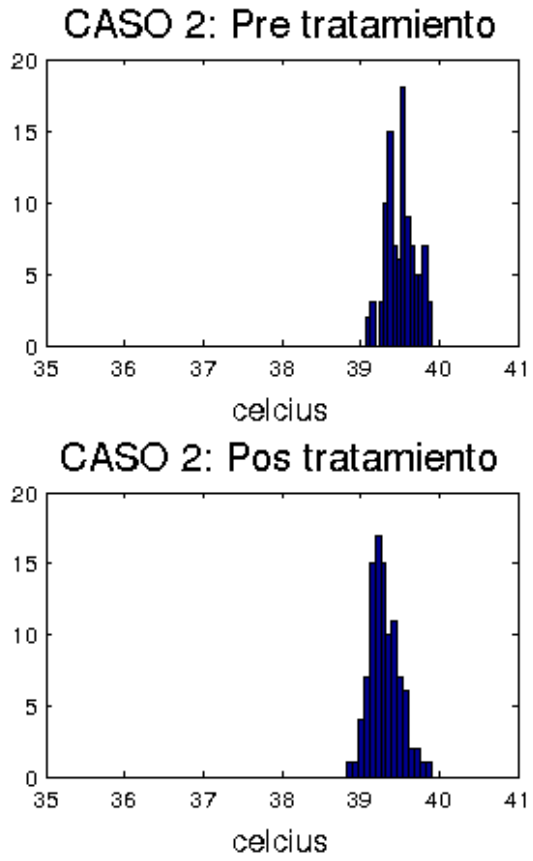
Funciono el tratamiento?

Test de hipótesis



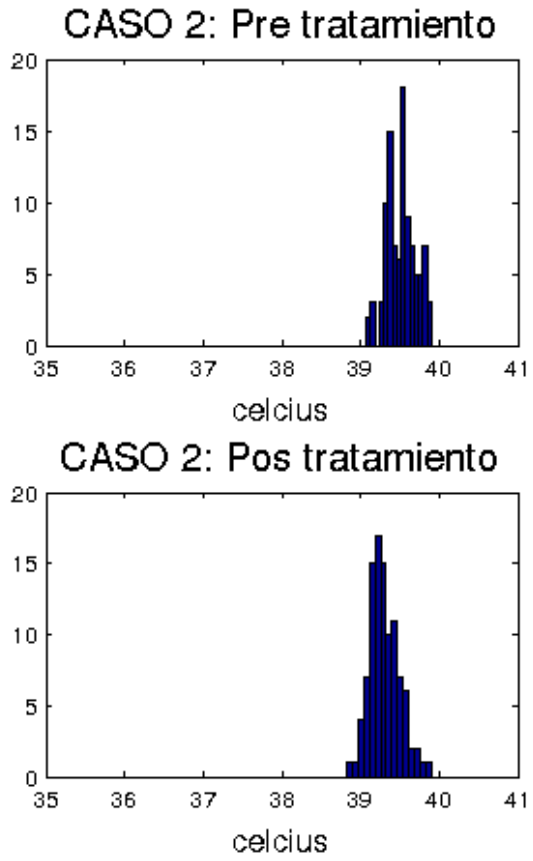
Funciono el tratamiento? $pvalue = 1{,}1848 \times 10^{-85}$

Test de hipótesis



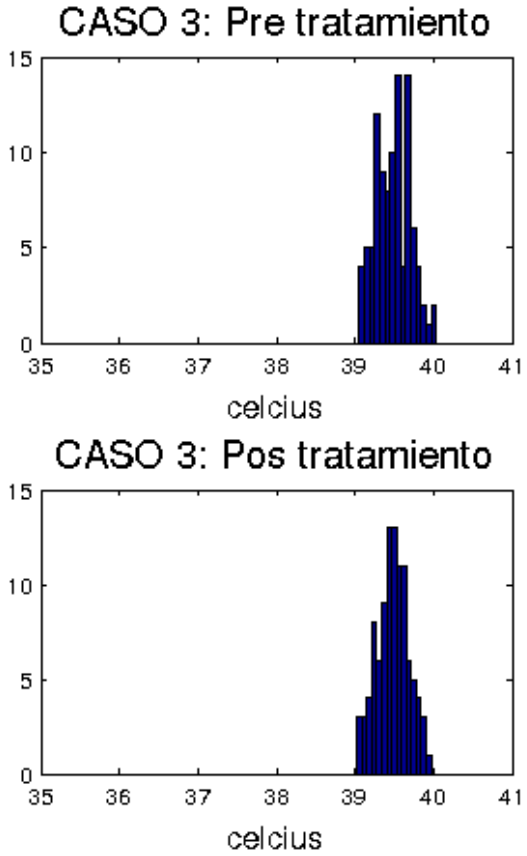
Funciono el tratamiento?

Test de hipótesis



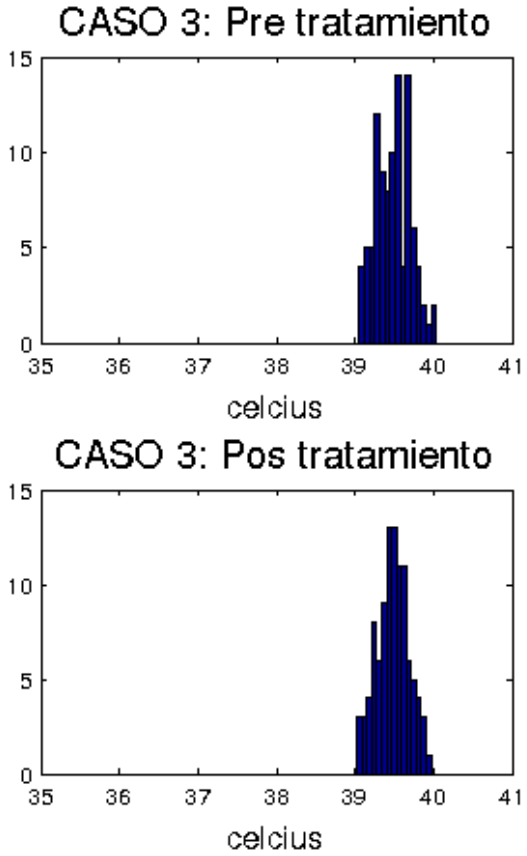
Funciono el tratamiento? $pvalue = 0{,}00033022$

Test de hipótesis



Funciono el tratamiento?

Test de hipótesis



Funciono el tratamiento? $pvalue = 0{,}4$

Test de hipótesis

Cómo calcularlo?

- En Scipy: scipy.stats.ttest_rel
- En Matlab: ttest

Test de hipótesis

Cómo calcularlo?

- En Scipy: scipy.stats.ttest_rel
- En Matlab: ttest

Si $p < 0{,}01$ o $p < 0{,}05$ podemos rechazar H_0 y aceptar H_1

En resumen...

Todo esto fue un repaso de herramientas útiles para resolver el TP2, este consiste en:

- Leer 1 Paper
- Implementar métodos
- Hacer análisis de los métodos
- Reportar resultados en un completo informe

Entrega: Viernes **14 de Noviembre**
Recuperatorio: Viernes **12 de Diciembre**

En un rato en la web de materia

¿Preguntas?

¿Preguntas?

(si... ya vamos a entregar los parciales)