AST0212 – 2016-1

Introducción al análisis de datos

Instituto de Astrofísica

Facultad de Física

Pontificia Universidad Católica de Chile

Nuestro Semestre 2016-1

AST0212

C0 ✓

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
6 Mar 2016 Semana 1	7	8	9	10	11 C1 ✓	12
13 Semana 2	14 TL1	15 TM1	16	17	18 C2 ✓	19
20 Semana 3	21 TL2	22 TM2	23	24	25 Feriado	26
27 Semana 4	28 TL3	29 TM3	30	31	1 Apr C3 ✓	2
3 Semana 5	4 TL4	5 TM4	6	7	8 C4 ✓	9
10 Semana 6	11 TL5	12 TM5	13	14	15 C5 ✓	16
17 Semana 7	18 TL6	19 TM6	20	21	22 C6✓ – ~~SM1~~	23
24 Semana 8	25 TL7	26	27	28	29 C7✓ – ~~SM2~~	30
1 May Semana 9	2 TL8	3 TM8	4	5	6 C8✓ – ~~SM3~~	7
8 Semana 10	9 TL9	10	11	12	13 C9✓ – ~~SM4~~	14
15	16 TL10	17 TM10	18	19	20 C10	21
22 Semana 12	23 TL11	24 TM11	25	26	27 C11	28
29 Semana 13	30 TL12	31 TM12	1 Jun	2	3 Feriado	4
5 Semana 14	6 TL13	7 TM13	8	9	10 C12	11
12 Semana 15	13 TL14	14 TM14	15	16	17 C13	18
19	20	21	22	23	24	25
26	27	28	29	30	1 Jul	2
3	4	5	6	7	8 Notas	9

← Control 1
Reparto Tarea 1

← Control 2

← Reparto T2

← Entrega T1

Entrega T2 →

Reparto T3 →

Entrega T3 →

← Control 3

Tutorías día lunes
Módulo 4:
Nicolás Castro

Tutorías día martes
Módulo 6:
Francisco Aros

← Examen

PDF Calendar by www.pdfcalendar.com

Clase previa (Clase 9):

REPASO

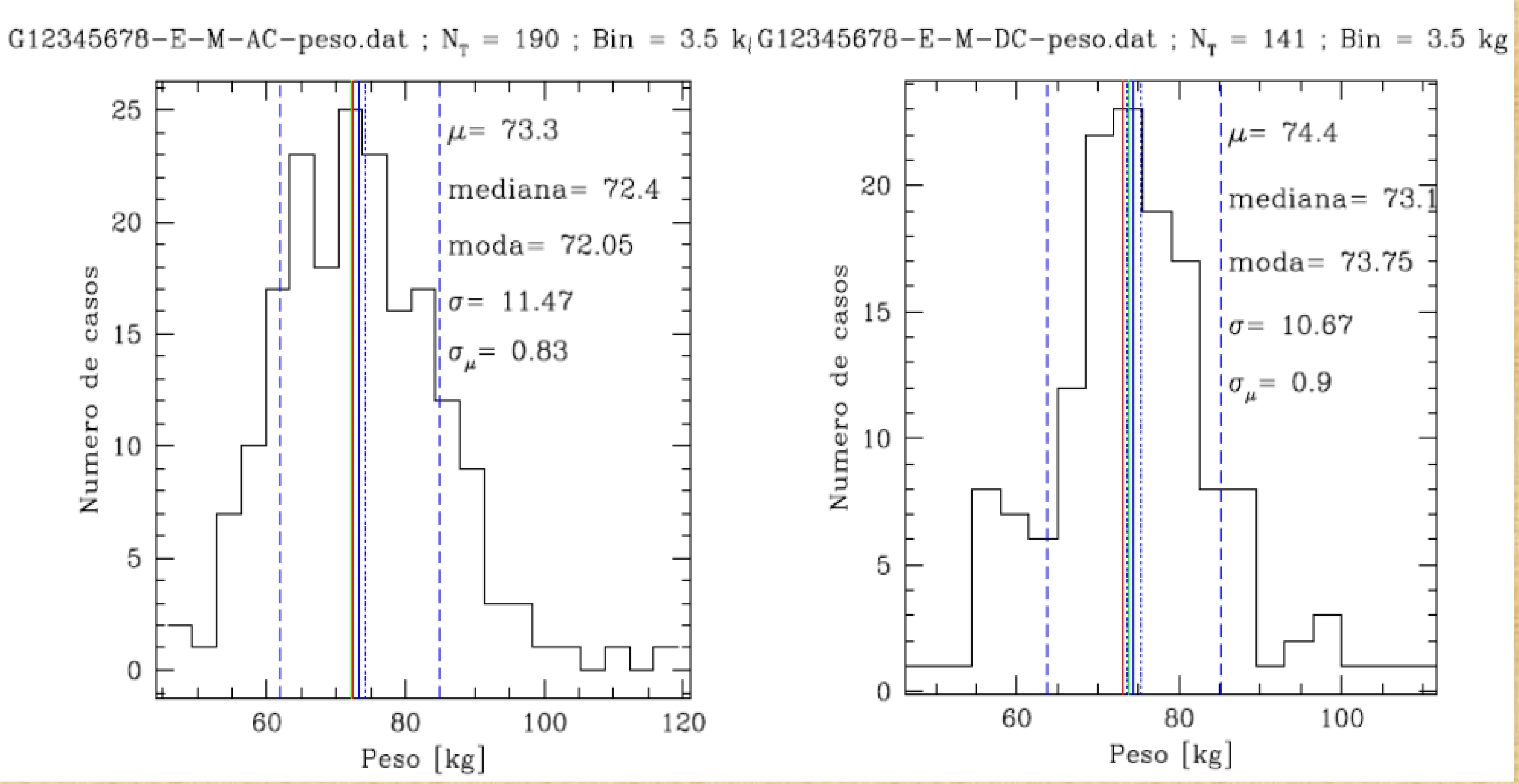
1. Repaso de temas críticos de la clase previa ✓
   1. Correlación.
   2. Incerteza de parámetros en la correlación lineal. ✓
   3. Corrección de error sistemático. Extrapolación.
2. Coeficiente de correlación. ✓
3. Significación de diferencia en media ✓

Esta clase (Clase 10):

1. Repaso de temas críticos de la clase previa
   1. Significación de diferencia en media
   2. Correlación, coeficiente de correlación
2. Significación de un coeficiente de correlación.

Comparación de dos distribuciones observadas: Significación de la diferencia de promedios

REPASO



Comparación de dos distribuciones observadas: Significación de la diferencia de promedios

REPASO

Dados:

$$\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_{1,i}$$

$$\sigma_1 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} x_{1,i}$$

$$\bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_{2,i}$$

$$\sigma_2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} x_{2,i}$$

Tendremos el error del promedio:

¡Permite la estrategia de aumentar la muestra!

$$\sigma_{\bar{x}_1} = \frac{\sigma_1}{\sqrt{N_1}}$$

(Se obtienen de aplicar propagación de errores a las definiciones de $\bar{x}_1$ y $\bar{x}_2$.)

$$\sigma_{\bar{x}_2} = \frac{\sigma_2}{\sqrt{N_2}}$$

Con estos elementos podemos construir el estimador t, con $\nu = N_1 + N_2 - 2$ grados de libertad:

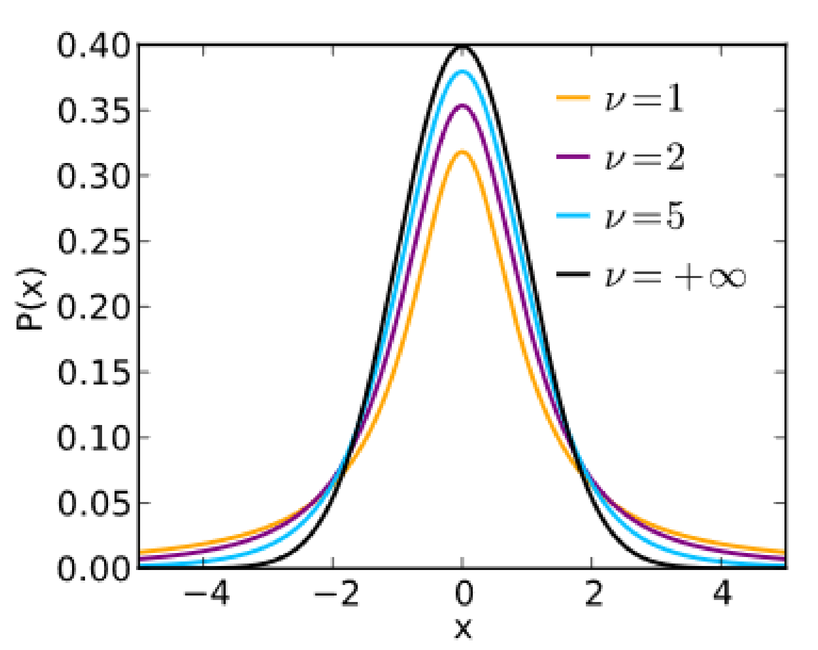
$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_D}$$

donde

$$S_D = \sqrt{\frac{\sum_{N_1}\left(x_{1,i} - \bar{x}_1\right)^2 + \sum_{N_2}\left(x_{2,i} - \bar{x}_2\right)^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

S_D es el error estándar de la diferencia de promedios. t tiene FDP tipo $t - Student$.

Distribución *t de Student*



La FDP de t, $A(t|\nu)$, denota la probabilidad de que t sea, por azar, menor que el valor medido si los promedios $\bar{x}_1$ y $\bar{x}_2$ son realmente iguales. Un valor grande (por ejemplo 0.99) indica una *alta chance de medir un valor menor que el observado si* $\bar{x}_1 = \bar{x}_2$. Esto es una indicación de que los promedios muy probablemente no sean los mismos. El valor complementario $1 - A(t|\nu)$ es la probabilidad de medir un valor tan grande como t si $\bar{x}_1 = \bar{x}_2$ (0.01 en el caso previo).

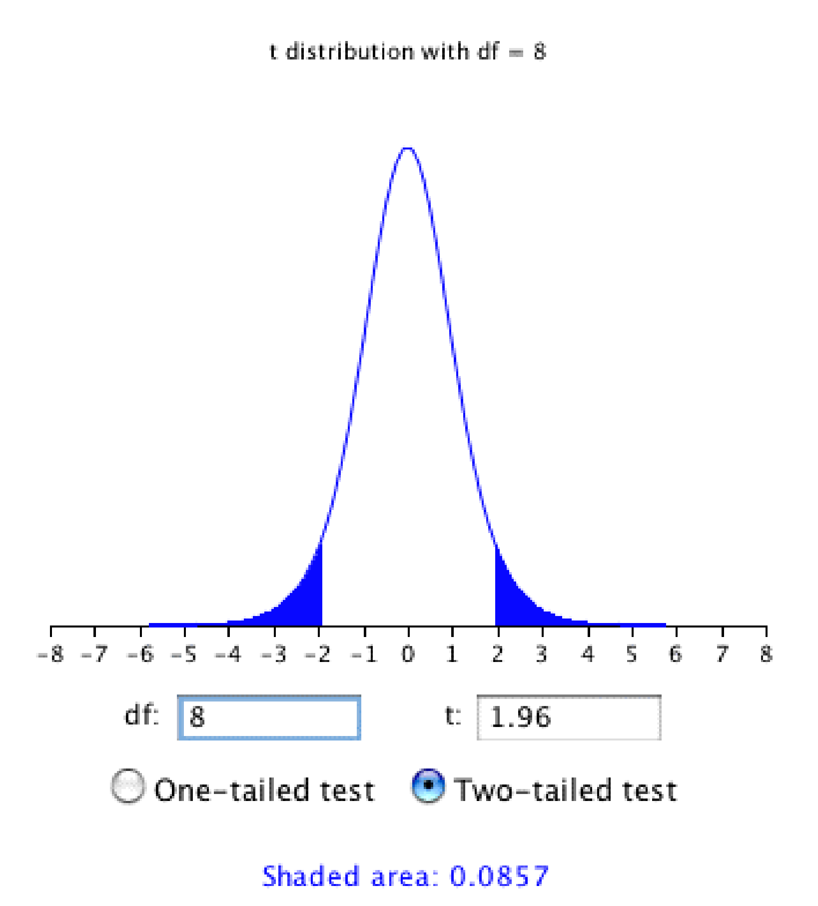
$$A(t|\nu) = \int_{-t}^{t} \frac{1}{\nu^{\frac{1}{2}} B(\frac{1}{2},\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = \int_{-t}^{t} P(x) dx = 1 - I_{\frac{\nu}{\nu+t^2}}\left(\frac{\nu}{2},\frac{1}{2}\right)$$

Donde $B(a,b)$ es la función Beta, e $I_x(a,b)$ es la función Beta incompleta (en este caso para para $x = \frac{\nu}{\nu+t^2}$, $a = \nu/2$ y $b = 1/2$).

Hay calculadores on-line para estas funciones, por ejemplo para $1 - A(t|\nu)$:
http://onlinestatbook.com/2/calculators/t_dist.html

Distribución *t de Student*

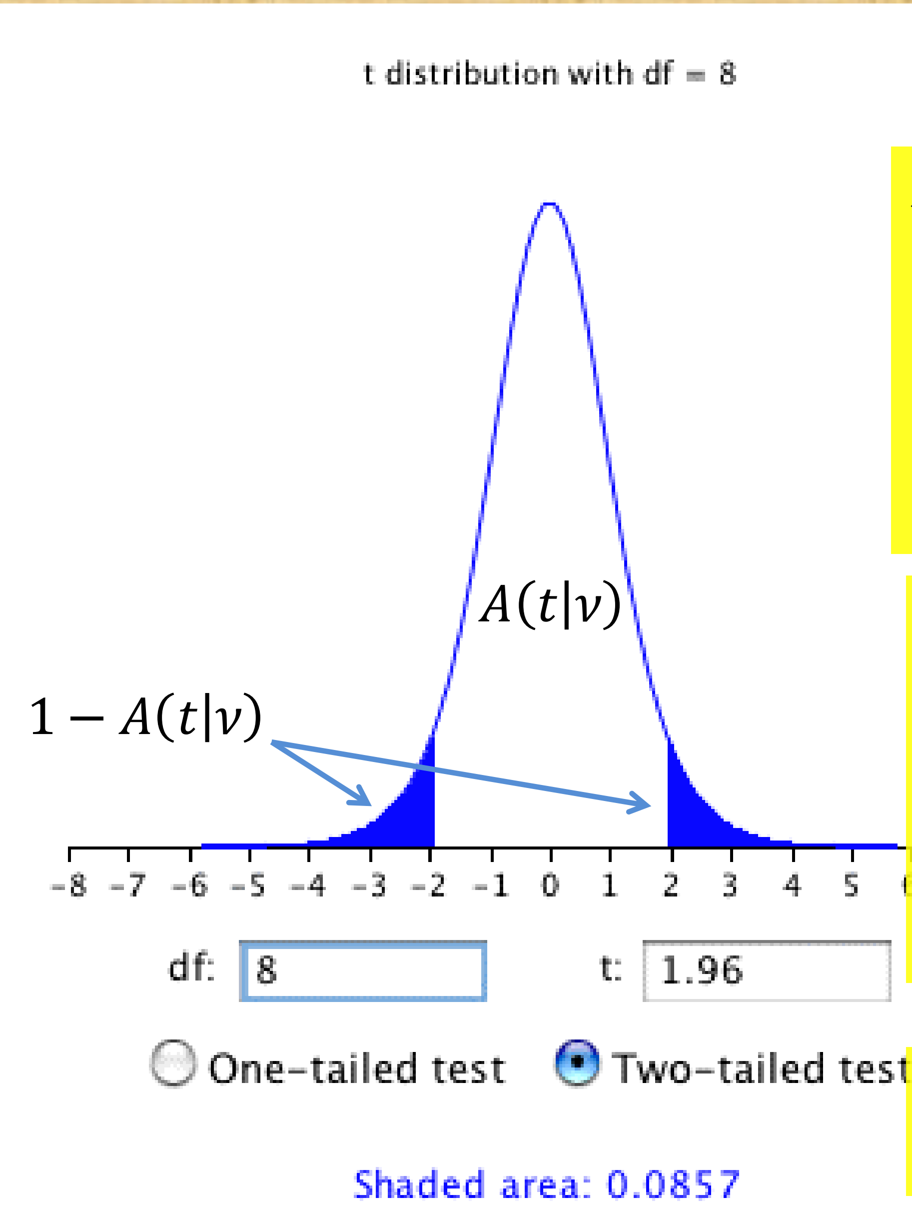
$$A(t|\nu) = \frac{1}{\nu^{\frac{1}{2}} B(\frac{1}{2},\frac{\nu}{2})} \int_{-t}^{t} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx$$



La FDP *t* de *Student* es en realidad una FDP cumulativa (la integral entre –*t* y *t*). Lo que estaba graficado en la lámina previa era el integrando.

Este gráfico muestra $A(t|\nu)$ como área blanca bajo la línea azul, y $1 - A(t|\nu)$ como área azul en los extremos derecho e izquierdo de la distribución. Por simetría, debemos considerar ambas colas (ya que el orden en que hacemos la resta en la definición de t es arbitrario. La función $1 - A(t|\nu)$ incorpora las dos colas naturalmente.

Interpretación del *t de Student*



$A(t|\nu)$: Probabilidad de que t sea así de pequeño por azar, cuando los promedios de las dos distribuciones comparadas son realmente iguales.

En este ejemplo: $A(1.96|8) = 0.9143$

$1 - A(t|\nu)$: Probabilidad de que t sea así de grande por azar, cuando los promedios de las dos distribuciones comparadas son realmente iguales.

En el ejemplo: $1 - A(1.96|8) = 0.0857$

"Pequeño" o "grande" se miden en este contexto comparando con 1 ¿Se entiende?

Coeficiente de correlación

REPASO

Partamos con un recordatorio de las ecuaciones de ajuste lineal de cuadrados mínimos:

$$y_i = ax_i + b$$

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - (ax_i + b)}{\sigma_i} \right)^2$$

$$\sum_{i=1}^{N} \frac{y_i}{\sigma_i{}^2} = b \sum_{i=1}^{N} \frac{1}{\sigma_i{}^2} + a \sum_{i=1}^{N} \frac{x_i}{\sigma_i{}^2}$$

$$\Delta = \sum_{i=1}^{N} \frac{1}{\sigma_i{}^2} \sum_{i=1}^{N} \frac{x_i{}^2}{\sigma_i{}^2} - \left(\sum_{i=1}^{N} \frac{x_i}{\sigma_i{}^2} \right)^2$$

$$\sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i{}^2} = b \sum_{i=1}^{N} \frac{x_i}{\sigma_i{}^2} + a \sum_{i=1}^{N} \frac{x_i{}^2}{\sigma_i{}^2}$$

$$a = \frac{1}{\Delta} \left(\sum_{i=1}^{N} \frac{1}{\sigma_i{}^2} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i{}^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i{}^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i{}^2} \right)$$

$$b = \frac{1}{\Delta} \left(\sum_{i=1}^{N} \frac{x_i{}^2}{\sigma_i{}^2} \sum_{i=1}^{N} \frac{y_i}{\sigma_i{}^2} - \sum_{i=1}^{N} \frac{x_i}{\sigma_i{}^2} \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i{}^2} \right)$$

Simplifiquemos para un caso sin σ (es idéntico a imaginar $\sigma = 1$)

Coeficiente de correlación (caso sin σ) REPASO

¿Tiene sentido la correlación $y_i = ax_i + b$? Prestemos atención a la pendiente.

$$a = \frac{1}{\Delta}\left(N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right) = \frac{\Delta_s}{\Delta}$$

$$\Delta = N\sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2$$

$$x_i = a'y_i + b' \qquad a' = \frac{\Delta_s}{\Delta'}$$

$$\Delta' = N\sum_{i=1}^{N} {y_i}^2 - \left(\sum_{i=1}^{N} y_i\right)^2$$

Si hay una correlación real entre x e y deberá existir una relación entre a, a', b y b'.

$$x_i = \frac{1}{a} y_i - \frac{b}{a} \Rightarrow a' = \frac{1}{a}; b' = -\frac{b}{a} \Rightarrow aa' = 1$$

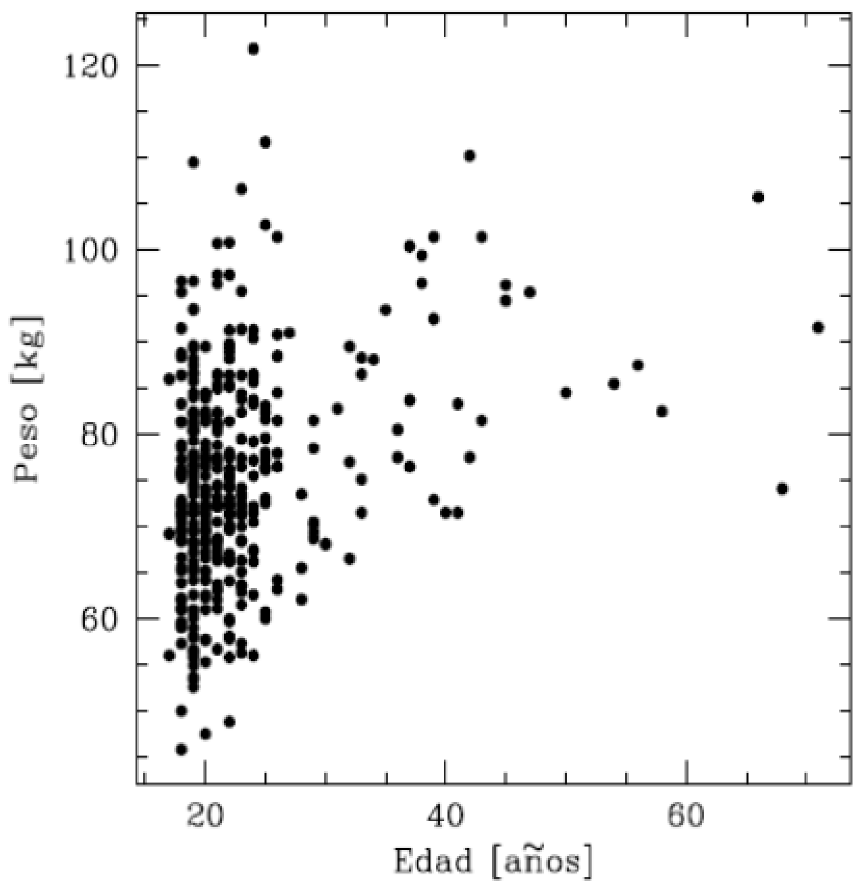
$$aa' = 1$$

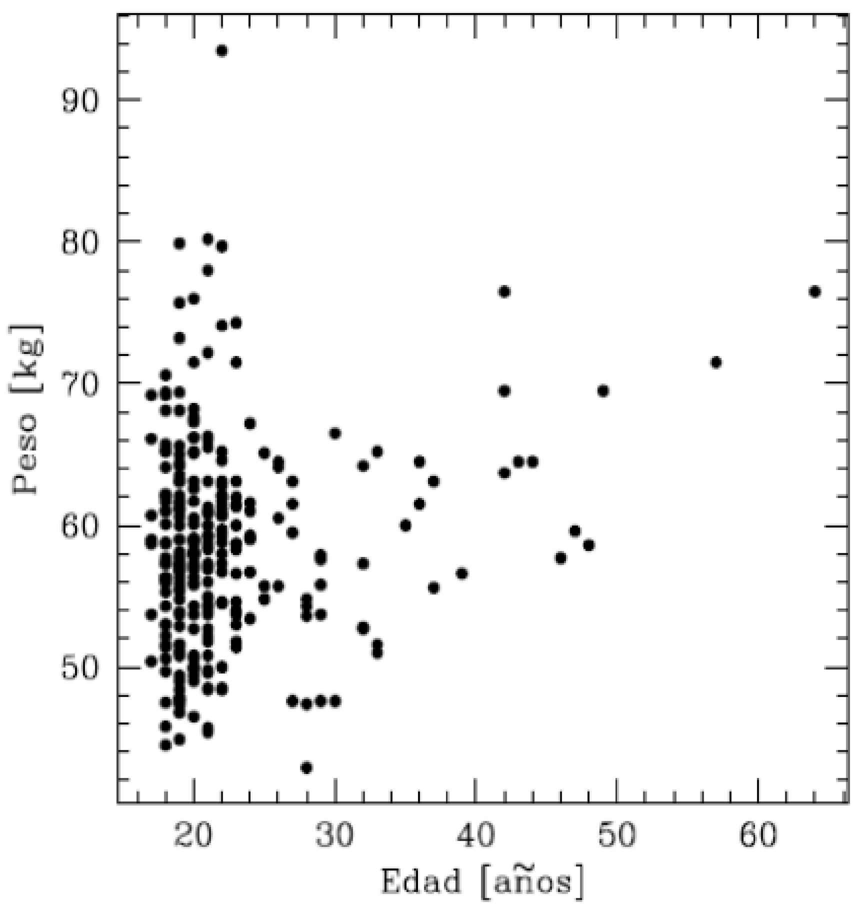
Definimos $r = \sqrt{aa'}$ cantidad llamada "coeficiente de correlación lineal", que nos da una medida experimental del grado de correlación lineal, con valor entre 0 y ± 1.

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2}\sqrt{N\sum_{i=1}^{N} {y_i}^2 - \left(\sum_{i=1}^{N} y_i\right)^2}}$$

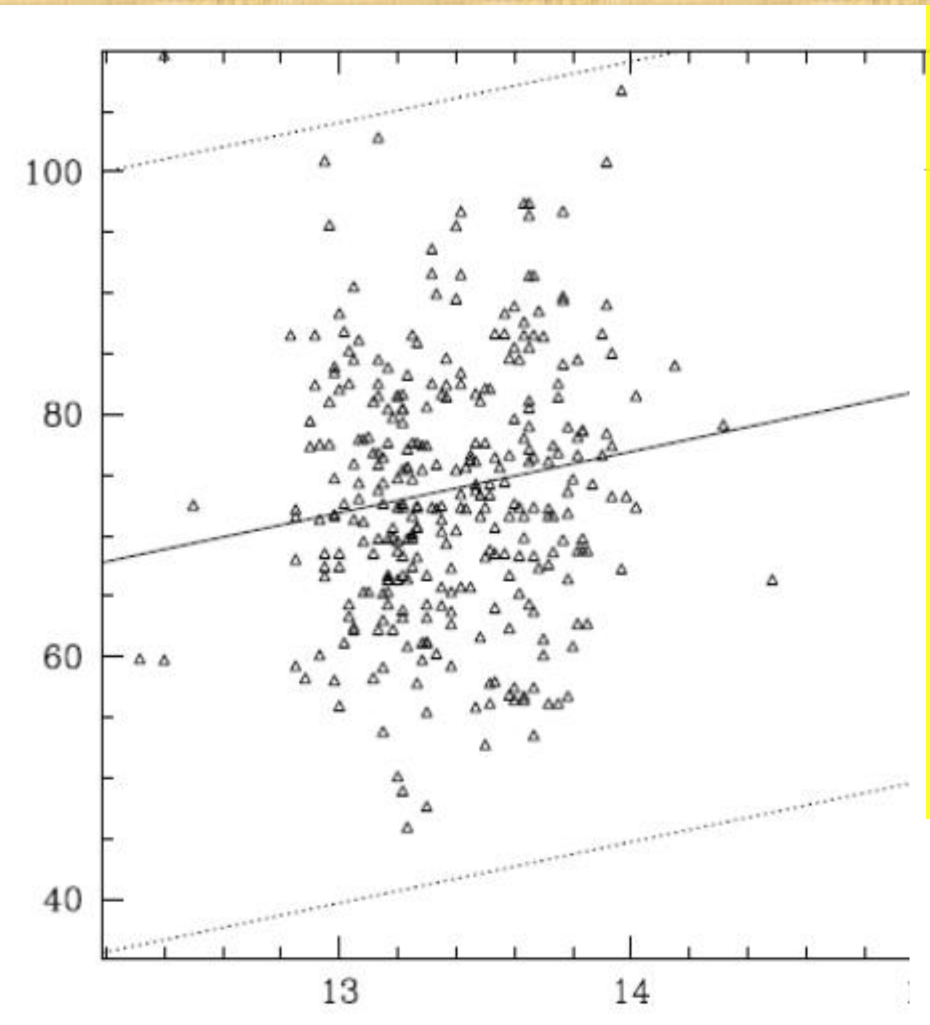
(Raimundo, antes de que tomáramos la raiz cuadrada, el numerador era un cuadrado)

¿Correlaciones casuales?





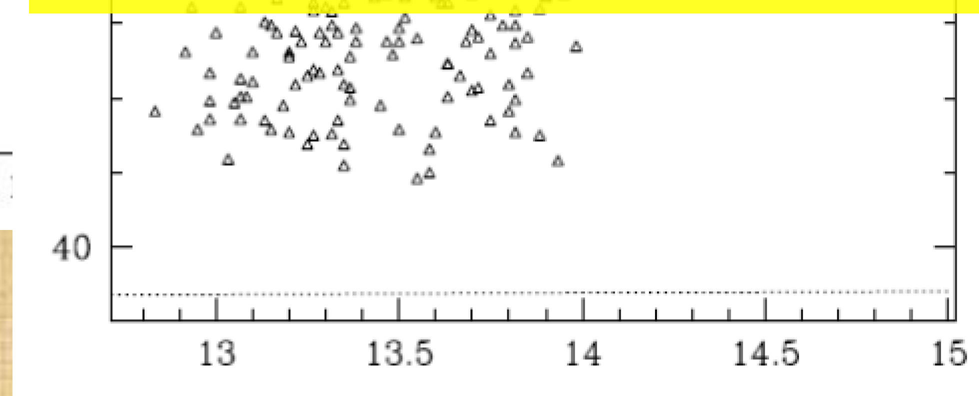
¿Correlaciones casuales?



En la hipótesis inicial (nula) de que *x* e *y* NO están correlacionados (y si los datos estás suficientemente agrupados, y si hay más que $n \approx 20$ puntos) entonces *r* tiene distribución $N(0, \sqrt{n})$. En ese caso, la probabilidad de que r sea "así de grande" por azar (es decir, con *x* e *y* no correlacionadas). Está dada por:

$$1 - erf\left(\frac{|r|\sqrt{n}}{\sqrt{2}}\right)$$

(esta $erf(x)$ es la real, no la del profe)



Condoro del profe (2)

```
[aclocchi@localhost ~/source]$ more myFx.f
      double precision FUNCTION FF(X)
c
c adapted from erfcc in NR
c
c This one assumes that x is N(0,1) and returns F(x), the integral of a
c normalized gaussian distribution between -infty and x (i.e.  c gives
c funcion P(x), eq. 26.2.2 in Abramowitz & Stegun). Note that
c
c          2.0d0*ff(x)-1.0
c
c gives the probability that the results fall between -x and x, as usual
c
c x = 1 =>  0.68268946714998124
c x = 2 =>  0.95449973904422536
c x = 3 =>  0.99730020379497075
c x = 4 =>  0.99993665751002569
c x = 5 =>  0.99999942669685327
c
c etc., etc., etc.
c
      implicit double precision (a-h,o-z)
c
      Z=DABS(X)/dsqrt(2.0d0)
      T=1./(1.+0.5*Z)
      ERFCC=T*DEXP(-Z*Z-1.26551223+T*(1.00002368+T*(.37409196+
     *    T*(.09678418+T*(-.18628806+T*(.27886807+T*(-1.13520398+
     *    T*(1.48851587+T*(-.82215223+T*.17087277)))))))))
      IF (X.LT.0.) ERFCC=2.-ERFCC
      FF=(2.0d0-ERFCC)/2.0d0
      RETURN
      END
```

Fin de ppt de Clase 10

Preguntas:

En general lo están haciendo bastante mal con la tarea:

1. ¿Qué hace falta para que esto funcione mejor?

    a) ¿Algo que pueda hacer desde la cátedra?
    b) ¿Algo que podamos hacer desde las ayudantías?
    c) ¿Algo que pueda hacer la universidad?