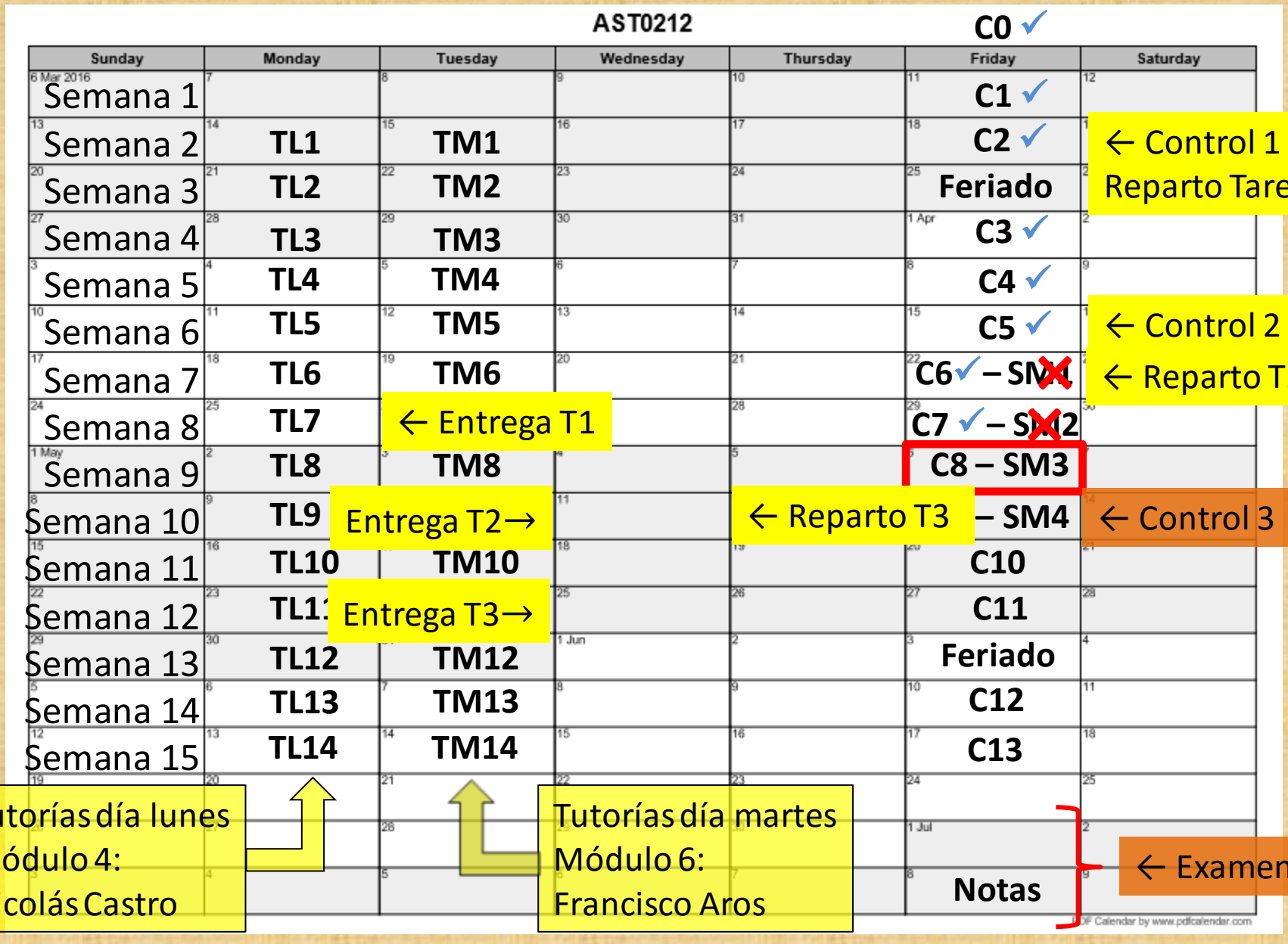

AST0212 – 2016-1
Introducción al análisis de datos
Instituto de Astrofísica
Facultad de Física
Pontificia Universidad Católica de Chile
©Albie Venter

Nuestro Semestre 2016-1



Preguntas guía para "Introducción al análisis de datos"

1) ¿Qué es un histograma? ¿Cómo se construye? ¿Cómo puede caracterizarse?
2) ¿Qué es una FDP?
3) ¿Cuál es la relación entre la FDP de una cierta variable y el histograma de valores que medimos para esta misma variable?
4) Dado un conjunto de medidas (datos) de valores directamente comparables entre sí, defina el valor medio, la mediana, la moda, y la dispersión.
5) ¿En qué clase de experimentos la dispersión de una variable observada proporciona una medida de la incerteza en la medición?
6) ¿Cuál es la diferencia entre una incerteza en la precisión y una en la exactitud?
7) Si queremos conocer una variable t, que no podemos medir, pero que se relaciona con otras variables x, y, z, que sí podemos medir directamente, por la ecuación

$$t = f(x, y, z)$$

   a. Dadas K medidas de x, N medidas de y, y M medidas de z, explique cómo haría para calcular $\bar{t}$ y σ_t.
   b. Puede imaginar una estrategia diferente para calcular $\bar{t}$ y σ_t para el caso de disponer de N medidas de x, y, z? (i.e. la misma cantidad de medidas en cada variable)

Clase previa (Clase 7):

REPASO

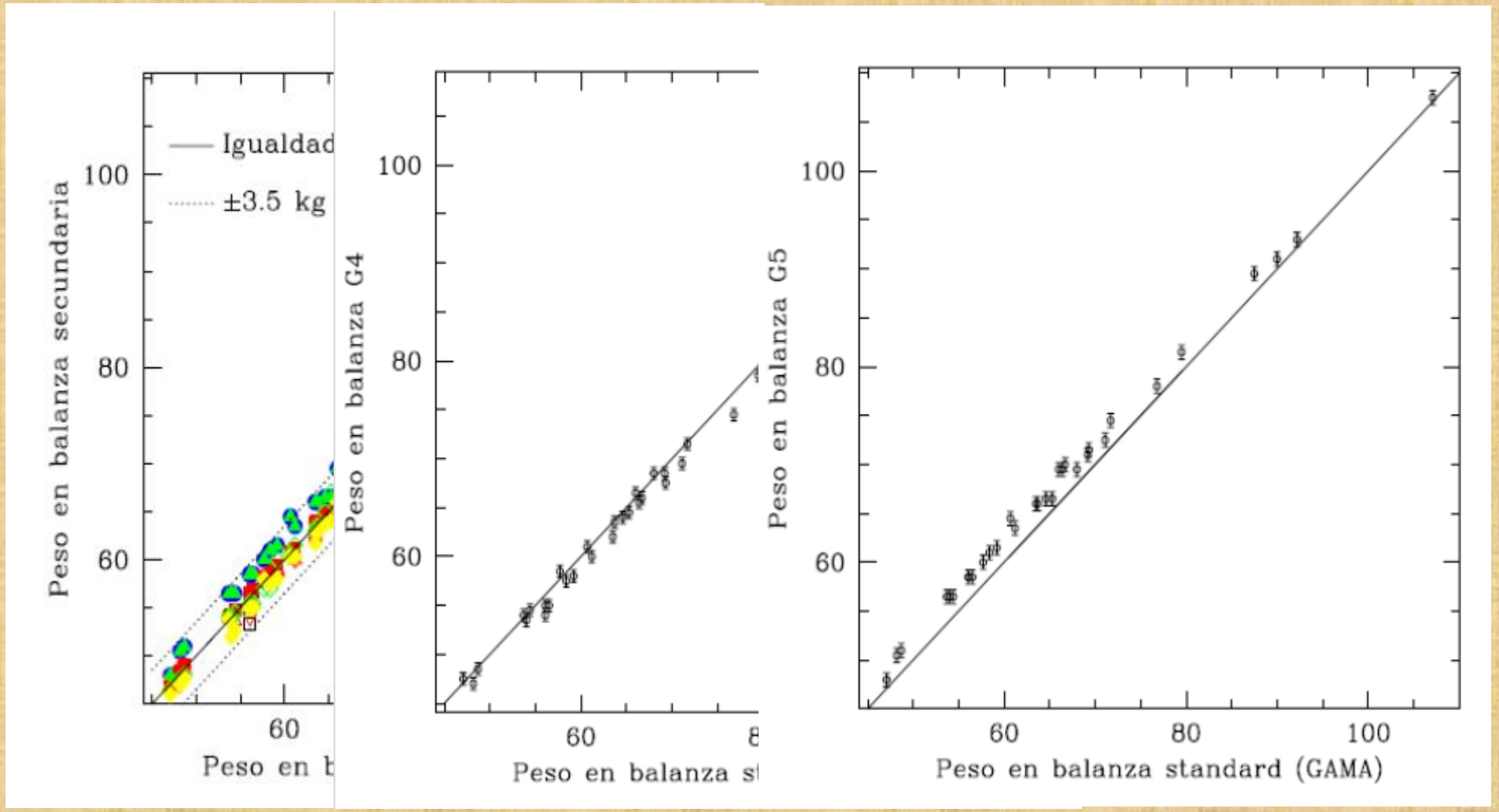
1. Herramienta Linux de selección de datos en archivos organizados en columnas: *awk* ✗
2. Repaso de temas críticos de la clase previa ✓
   1. Test modelo vs. realidad: χ^2 explicado.
   2. FDP de χ^2.
3. Viaje sin escalas a la propagación de errores. ✓
4. Correlación. ✓
5. Incerteza de parámetros en la correlación lineal ✓
6. Corrección de error sistemático. Extrapolación. ✓
7. Coeficiente de correlación. ✗

Esta clase (Clase 8):

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 y varianza

REPASO

Correlación: Peso medido con ≠ balanzas



Método de "cuadrados mínimos" REPASO

$$peso_{(balanza\ X)} = a * peso_{(balanza\ 7)} + b$$

$$y_i = ax_i + b$$

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

$$\frac{\partial \chi^2}{\partial a} = \frac{\partial}{\partial a} \left[\sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma_i} \right)^2 \right] = 0$$

$$\frac{\partial \chi^2}{\partial b} = \frac{\partial}{\partial b} \left[\sum_{i=1}^{N} \left(\frac{y_i - ax_i - b}{\sigma_i} \right)^2 \right] = 0$$

$$\sum_{i=1}^{N} \left[\frac{1}{{\sigma_i}^2} (y_i - ax_i - b) \right] = 0$$

$$\sum_{i=1}^{N} \left[\frac{x_i}{{\sigma_i}^2} (y_i - ax_i - b) \right] = 0$$

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

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

Método de “cuadrados mínimos” REPASO

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

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

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

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

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

Incerteza en los parámetros calculados

REPASO

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

$$da = \sum_{j=1}^{N} \frac{\partial a}{\partial y_j} dy_j \qquad db = \sum_{j=1}^{N} \frac{\partial b}{\partial y_j} dy_j \qquad b = \frac{1}{\Delta} \left(\sum_{i=1}^{N} \frac{{x_i}^2}{{\sigma_i}^2} \sum_{i=1}^{N} \frac{y_i}{{\sigma_i}^2} - \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \sum_{i=1}^{N} \frac{x_i y_i}{{\sigma_i}^2} \right)$$

$${\sigma_a}^2 = \sum_{j=1}^{N} \left(\frac{\partial a}{\partial y_j} \right)^2 {\sigma_j}^2 \qquad {\sigma_b}^2 = \sum_{j=1}^{N} \left(\frac{\partial b}{\partial y_j} \right)^2 {\sigma_j}^2$$

$$\frac{\partial a}{\partial y_j} = \frac{1}{\Delta} \left(\frac{x_j}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} - \frac{1}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \right) \qquad {\sigma_a}^2 = \frac{1}{\Delta^2} \sum_{j=1}^{N} {\sigma_j}^2 \left(\frac{x_j}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2} - \frac{1}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \right)^2$$

$$\frac{\partial b}{\partial y_j} = \frac{1}{\Delta} \left(\frac{1}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{{x_i}^2}{{\sigma_i}^2} - \frac{x_j}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \right) \qquad {\sigma_b}^2 = \frac{1}{\Delta^2} \sum_{j=1}^{N} {\sigma_j}^2 \left(\frac{1}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{{x_i}^2}{{\sigma_i}^2} - \frac{x_j}{{\sigma_j}^2} \sum_{i=1}^{N} \frac{x_i}{{\sigma_i}^2} \right)^2$$

Incerteza en los parámetros calculados

REPASO

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

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

$${\sigma_a}^2 = \frac{1}{\Delta} \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2}$$

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

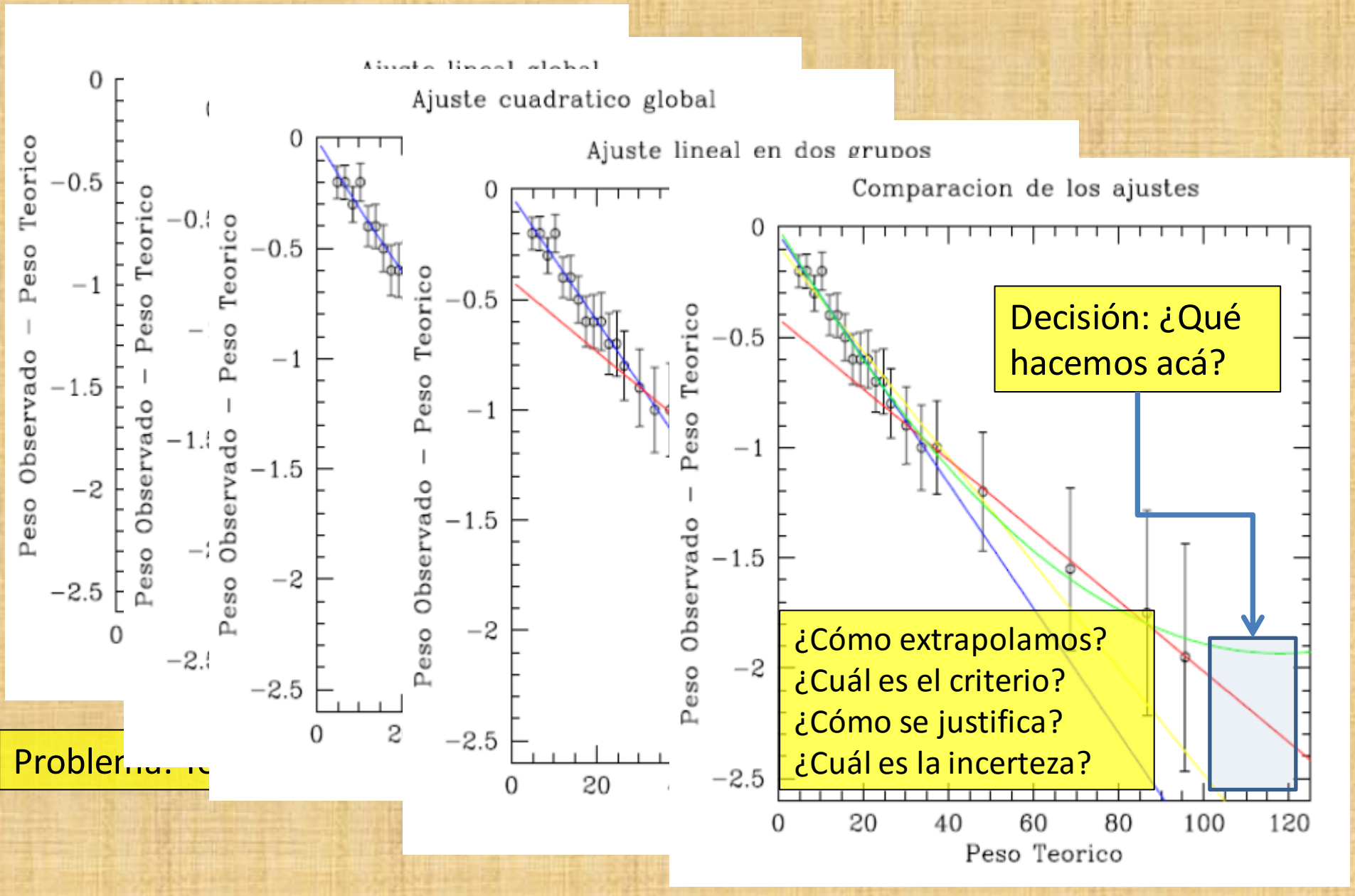
$${\sigma_b}^2 = \frac{1}{\Delta} \sum_{i=1}^{N} \frac{{x_i}^2}{{\sigma_i}^2}$$

Correlación: Error sistemático de GAMMA

REPASO

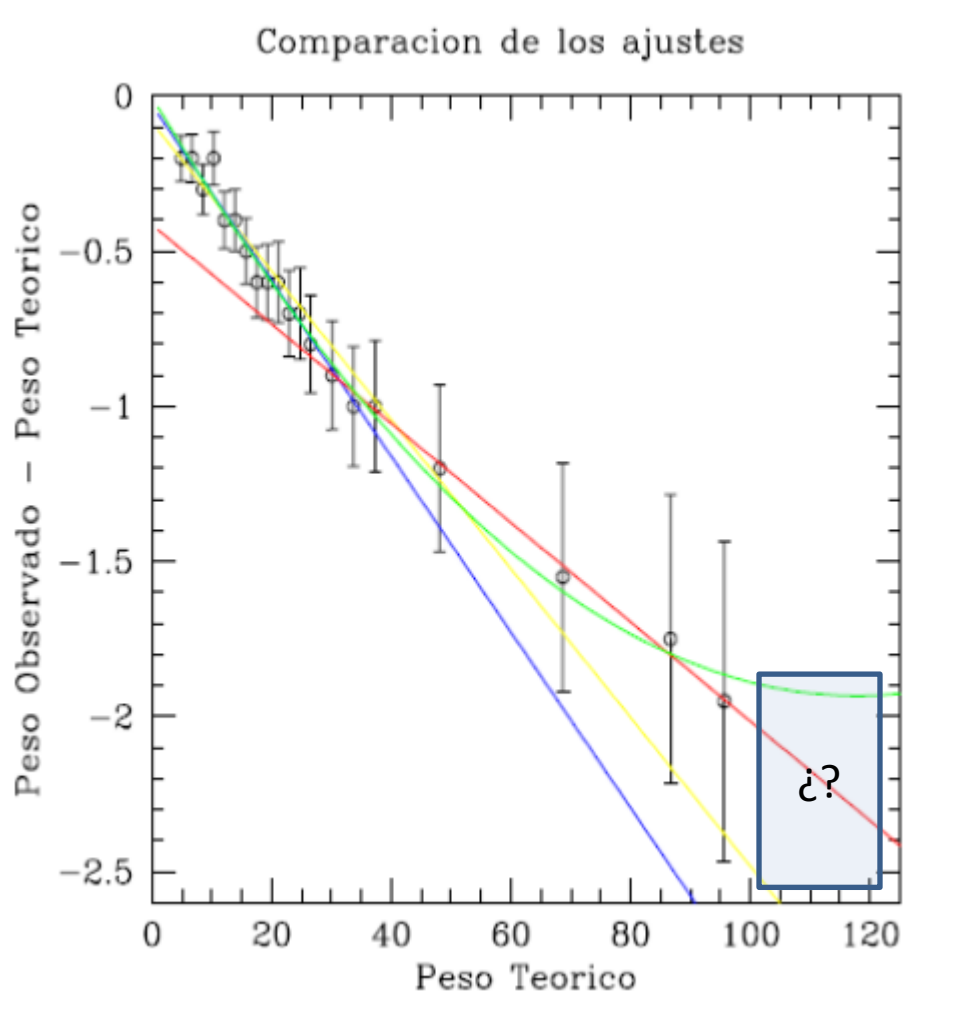
Correlación: Error sistemático de GAMMA

REPASO



Correlación: Error sistemático de GAIA

REPASO



El problema fundamental que enfrentamos acá es la ausencia de una "teoría" del instrumento. No sabemos como responde.

Un criterio de decisión posible, y usualmente aceptado en ciencias físicas es el llamado "La navaja de Occam" (Occam's razor en inglés; Lex parsimoniae en Latín). Éste es un principio general atribuido a William de Ockham (c. 1287–1347). En una formulación moderna sería: *En igualdad de condiciones, debe ser preferida la explicación que descanse en la menor cantidad de hipótesis*.

En otras palabras: **Es preferible lo simple**.

Queda a criterio de ustedes cómo se aplica ese principio en este caso particular (¿amarillo, verde o rojo?).

THE ASTRONOMICAL JOURNAL, 116:1009–1038, 1998 September

OBSERVATIONAL EVIDENCE FROM SUPERNOVAE FOR AN ACCELERATING UNIVERSE AND A COSMOLOGICAL CONSTANT

ADAM G. RIESS,[1] A
PETER M. GARNAV
B. LEIBUNDGUT,

We present spectr
range $0.16 \leq z \leq 0.6$
relations between S

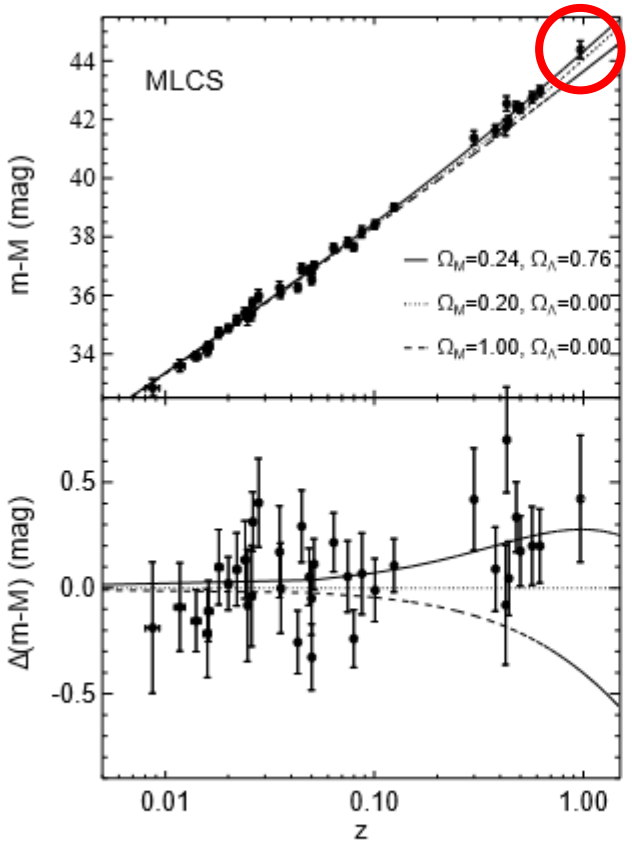


FIG. 4.—MLCS SNe Ia Hubble diagram. The upper panel shows the Hubble diagram for the low-redshift and high-redshift SNe Ia samples with distances measured from the MLCS method (Riess et al. 1995, 1996a; Appendix of this paper). Overplotted are three cosmologies: "low" and "high" Ω_M with $\Omega_\Lambda = 0$ and the best fit for a flat cosmology, $\Omega_M = 0.24$,

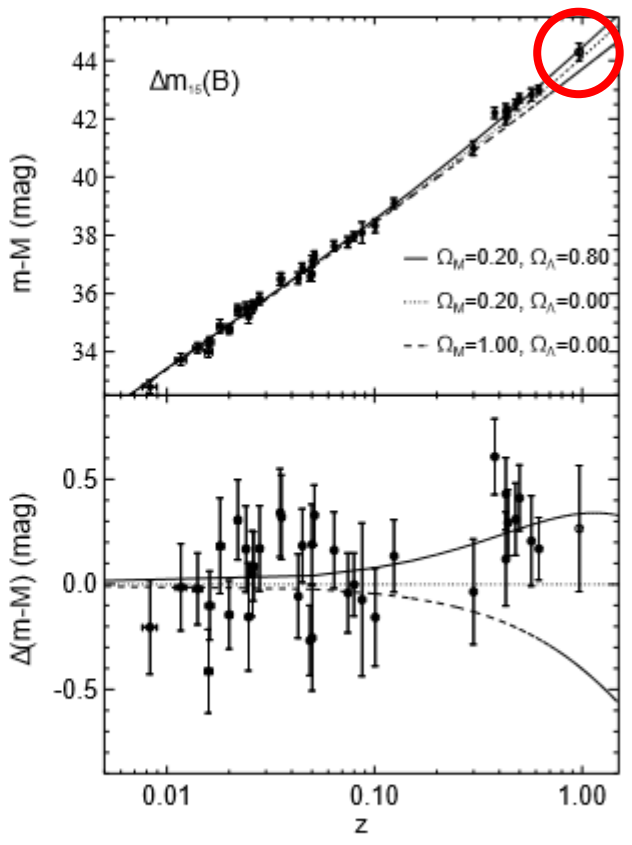


FIG. 5.—$\Delta m_{15}(B)$ SN Ia Hubble diagram. The upper panel shows the Hubble diagram for the low-redshift and high-redshift SNe Ia samples with distances measured from the template-fitting method parameterized by $\Delta m_{15}(B)$ (Hamuy et al. 1995, 1996d). Overplotted are three cosmologies: "low" and "high" Ω_M with $\Omega_\Lambda = 0$ and the best fit for a flat cosmology,

Comparación de dos distribuciones observadas

REPASO

$$\chi^2 = \sum_{j=1}^{M} \frac{\left(n_{1,j} - n_{2,j}\right)^2}{n_{1,j} + n_{2,j}}, \text{donde } M \text{ es el número de bins.}$$

La FDP de este χ^2 es la misma que mostré antes. ¿Qué es ν ahora? Si los datos son recogidos de forma tal que la suma de n_1 es necesariamente igual a la de n_2 tendremos que el número de grados de libertad es $\nu = M - 1$ (el caso usual). Si este requerimiento no existe, entonces $\nu = M$.

Ejemplo: Un observador de aves que desea comparar dos años de observaciones, tomando un bin por cada especie.

1: Base de datos es los 1000 primeros pájaros que observa cada año ($\nu = M - 1$)

2: Base de datos es todos los pájaros que vio en un número de días al azar, siendo el número de días el mismo en los dos años ($\nu = M$).

En el segundo caso puede comparar los totales. *Ese es el grado de libertad adicional*.

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

Dados:

$$\bar{x}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} x_{1,i} \qquad \sigma_1 = \frac{1}{N_1 - 1}\sum_{i=1}^{N_1} x_{1,i} \qquad \bar{x}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2} x_{2,i} \qquad \sigma_2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2} x_{2,i}$$

Tendremos el error del promedio:

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

(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:

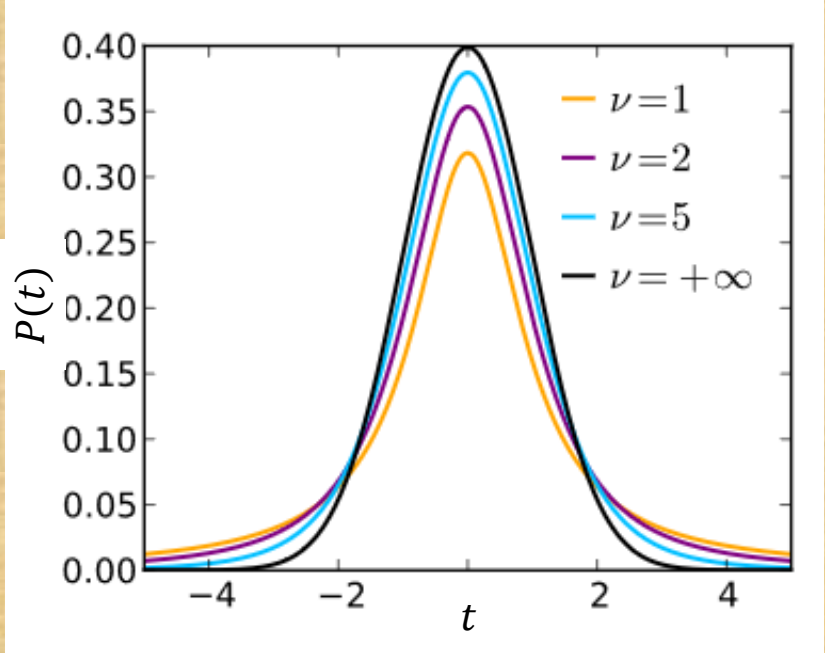
$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_D}$$

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).

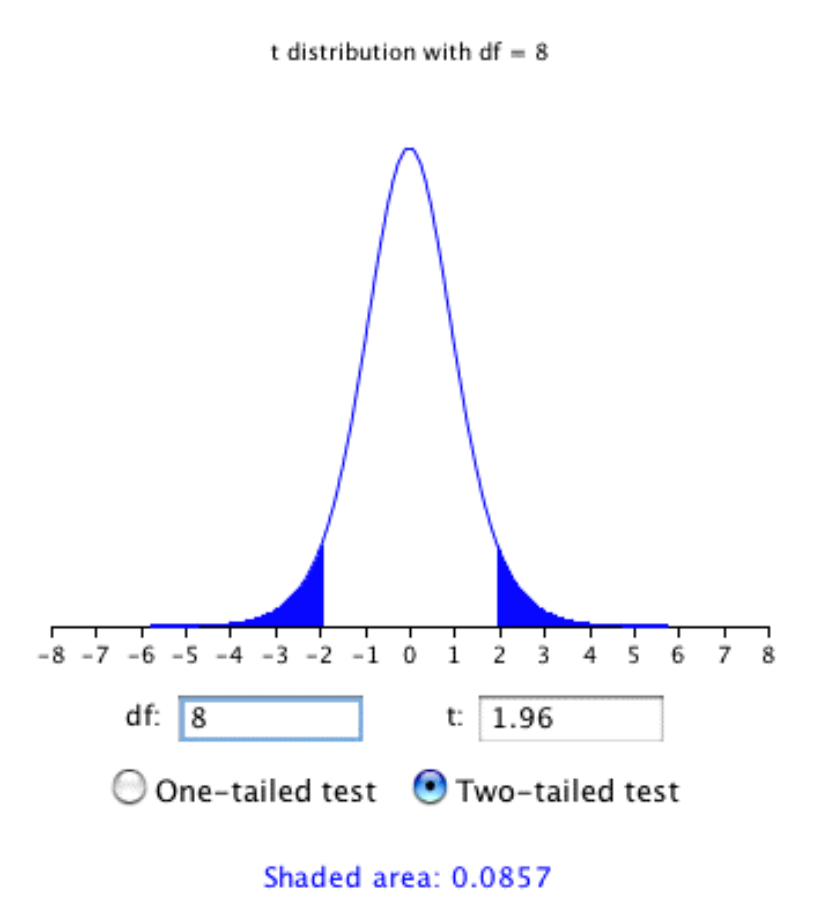
$$A(t|\nu) = \frac{1}{\nu^{\frac{1}{2}} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \int_{-t}^{t} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx = 1 - I_{\frac{\nu}{\nu+t^2}}\left(\frac{\nu}{2}, \frac{1}{2}\right)$$

Donde $I_x(a, b)$ es la función Beta incompleta 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*

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



El 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.

Coeficiente de correlación

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

$$y_i = ax_i + b$$

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

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

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

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

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

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

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

Coeficiente de correlación (caso sin σ)

¿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)

Coeficiente de correlación (caso sin σ)

Continuum Foreground Polarization and Na I Absorption in Type Ia SNe[1]

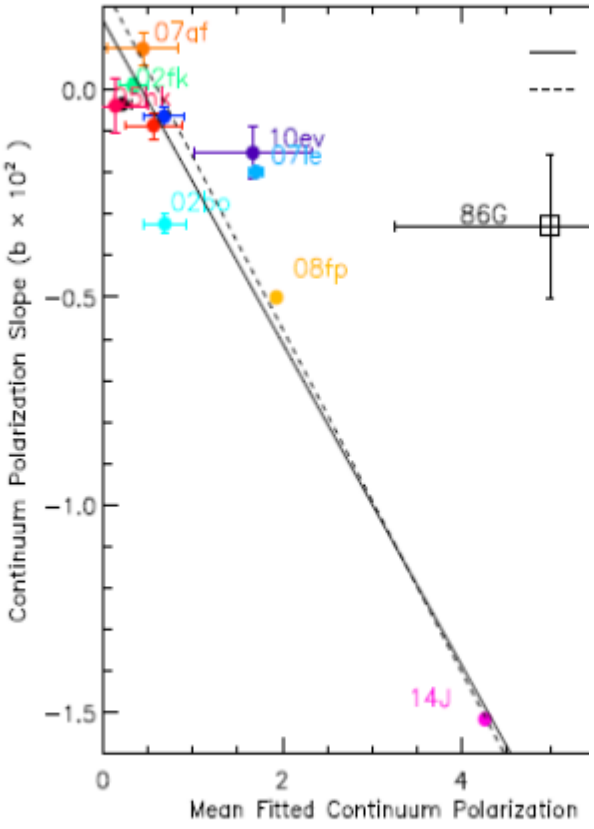


Fig. 8.— Correlation between the parameters b and $P_{\rm mean}$ th
tinuum polarization (see text and Fig. 2). The solid line sho

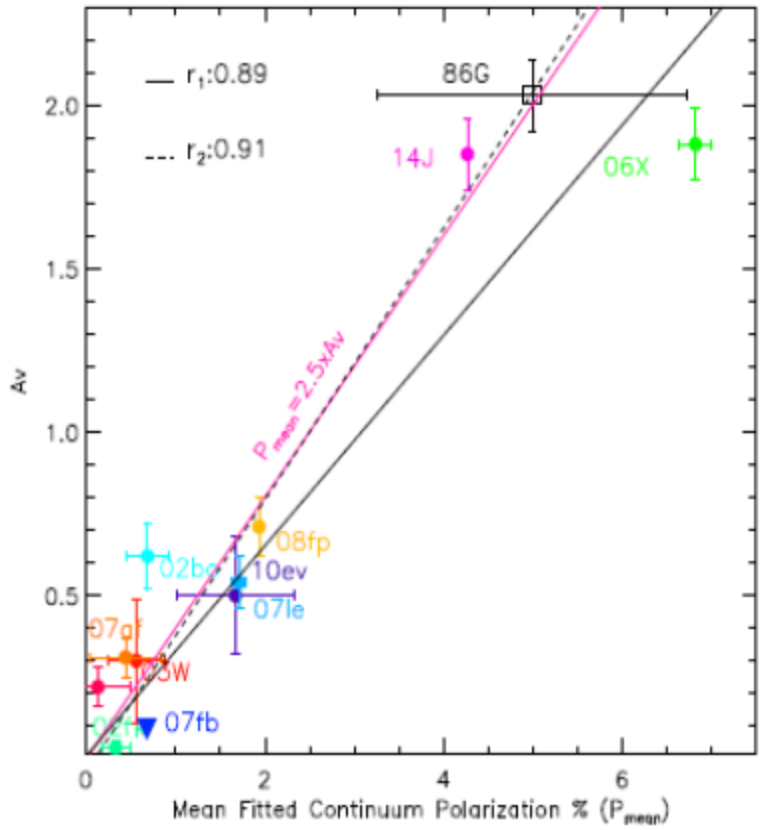
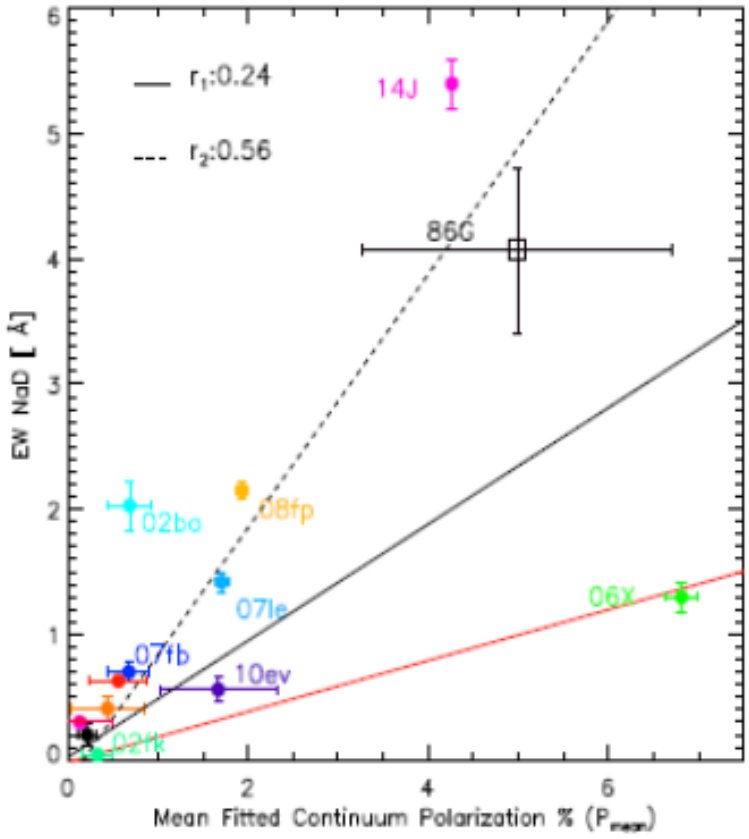


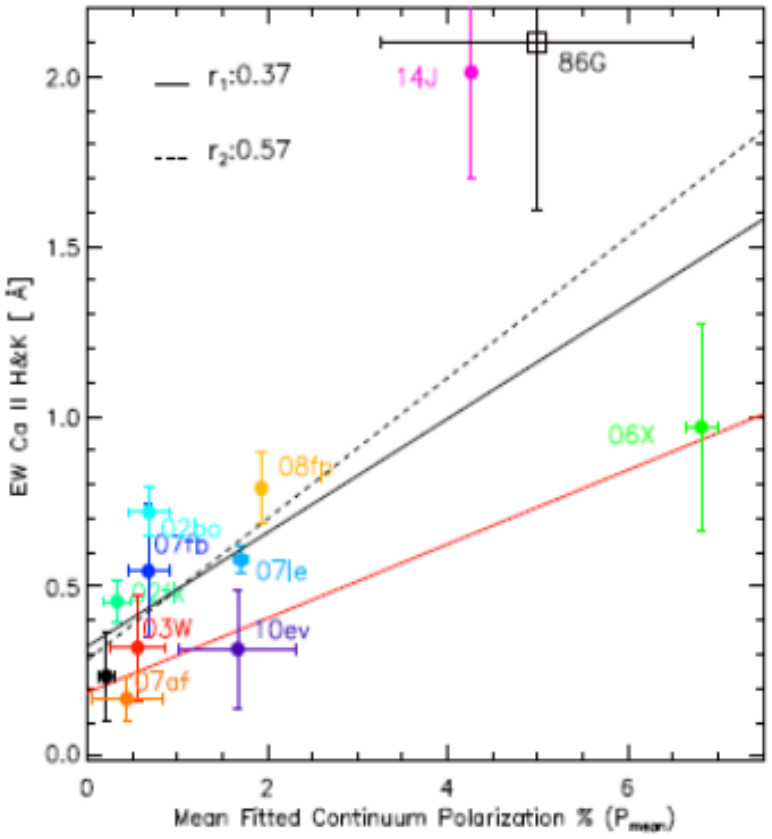
Fig. 10.— Correlation between the extinction in the visual band and $P_{\rm mean}$. The solid line shows the linear fit to all points and the dashed line the linear fit when SN 2006X is excluded (correlation coefficients r_1 and r_2 respectively). Down pointing triangles are upper limits to

Coeficiente de correlación (caso sin σ)

Continuum Foreground Polarization and Na I Absorption in Type Ia SNe[1]

P. Zelaya[2,3], A. Clocchiatti[3,2], D. Baade[4], P. Höflich[5], J. Maund[6], F. Patat[4], J.R. Quinn[3],





Fin de ppt de Clase 8