

A photograph of two lion cubs in a savanna setting. One cub is on the left, facing right, and the other is on the right, facing left. They are both on their hind legs, reaching towards each other with their front paws. The background is a blurred green field with some taller grass. The text 'AST0212 – 2016-1' is overlaid in the center in a large, bold, yellow font.

AST0212 – 2016-1

Introducción al análisis de datos

Instituto de Astrofísica

Facultad de Física

Pontificia Universidad Católica de Chile

Nuestro Semestre 2016-1

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
6 Mar 2016 Semana 1					C1 ✓	
13 Semana 2	TL1	TM1			C2 ✓	← Control 1 Reparto Tarea 1
20 Semana 3	TL2	TM2			Feriado	
27 Semana 4	TL3	TM3			C3 ✓	
3 Semana 5	TL4	TM4			C4 ✓	
10 Semana 6	TL5	TM5			C5 ✓	← Control 2
17 Semana 7	TL6	TM6			C6 ✓ – SM1	← Reparto T2
24 Semana 8	TL7	← Entrega Tarea 1			C7 – SM2	
1 May Semana 9	TL8	TM8			C8 – SM3	
8 Semana 10	TL9	← Entrega Tarea 2			C9 – SM4	
15 Semana 11	TL10	TM10			C10	
22 Semana 12	TL11	TM11			C11	
29 Semana 13	TL12	TM12	1 Jun		Feriado	
5 Semana 14	TL13	TM13			C12	
12 Semana 15	TL14	TM14			C13	
19 Tutorías día lunes Módulo 4: Jorge Castro	↑	↑	Tutorías día martes Módulo 6: Francisco Aros		Notas	

Clase previa (Clase 6):

REPASO

1. Herramienta Linux de selección de datos en archivos organizados en columnas: *awk* ✗
2. Repaso de temas críticos de la clase previa ✓
 1. FDPs que hay que conocer: Constante, Poisson y Gauss.
 2. Modelos de la realidad, distribución subyacente.
 3. Test modelo vs. realidad: χ^2 explicado. ✓
 4. FDP de χ^2 . ✓
3. Viaje sin escalas a la propagación de errores. ✓
4. Correlación. ✓

Esta clase (Clase 7):

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.

¿Qué es χ^2 ?

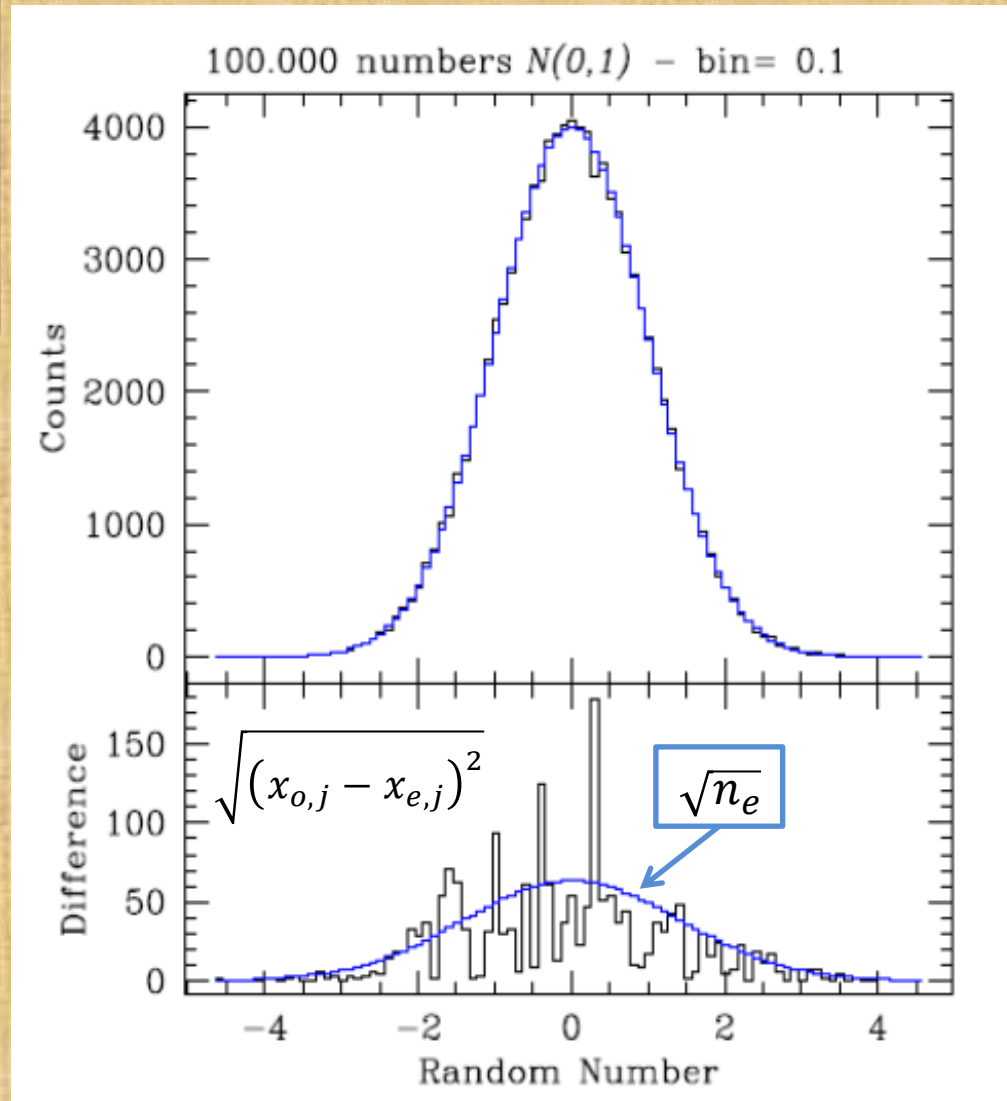
REPASO

En buena medida, esto es lo que ya veíamos en la figura anterior. Las diferencias $|n_o - n_e|$ siguen, aproximadamente la forma $\sqrt{n_e}$. Tenemos entonces:

$$\chi^2 = \sum_{j=1}^M \left(\frac{n_{o,j} - n_{e,j}}{\sigma_{e,j}} \right)^2$$

$$\chi^2 = \sum_{j=1}^M \left(\frac{n_{o,j} - n_{e,j}}{\sqrt{n_{e,j}}} \right)^2$$

$$\chi^2 = \sum_{j=1}^M \frac{(n_{o,j} - n_{e,j})^2}{n_{e,j}}$$

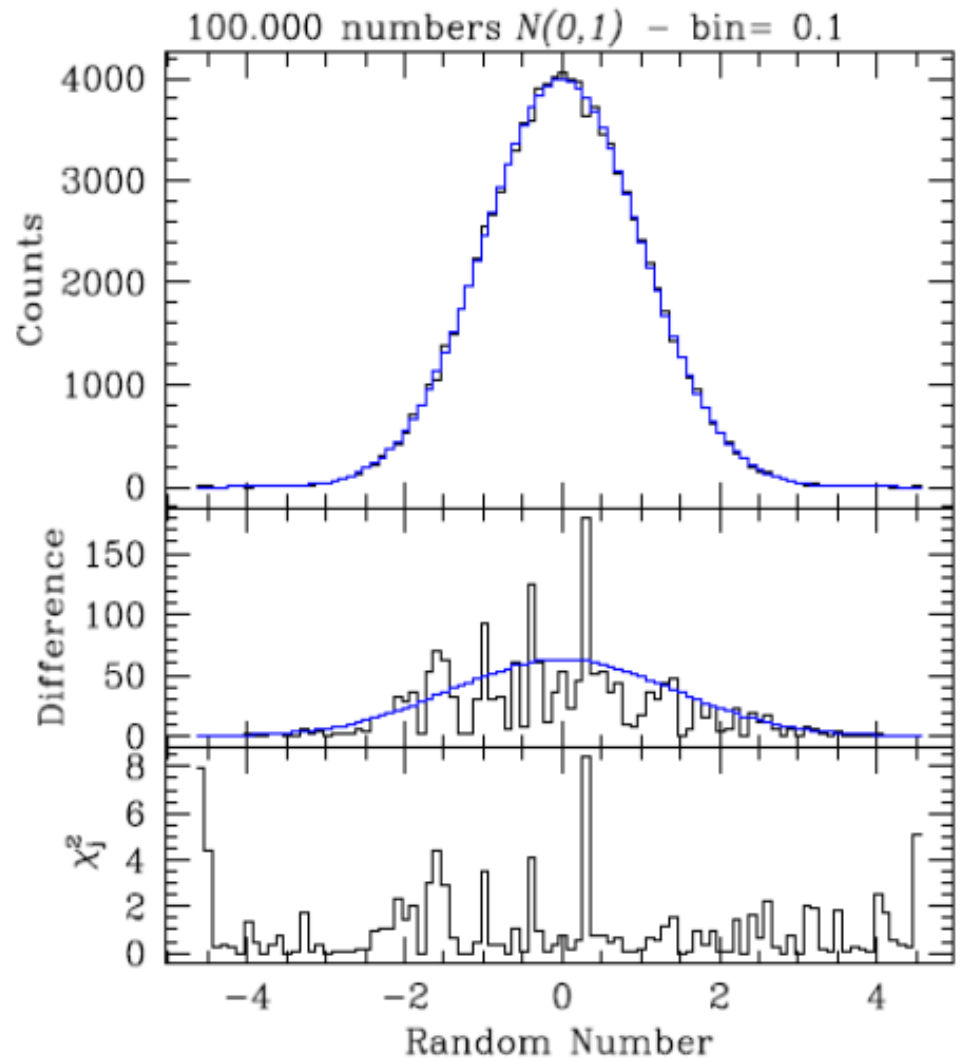


Esta última es la expresión que había puesto como “receta” al final de la Clase 4.

¿Qué podemos esperar de χ^2 ? REPASO

$$\chi^2 = \sum_{j=1}^M \left(\frac{n_{o,j} - n_{e,j}}{\sqrt{n_{e,j}}} \right)^2 = \sum_{j=1}^M \chi_j^2$$

Esperamos, en principio, que el valor de promedio de $(n_{o,j} - n_{e,j})$ esté bien representado por $\sigma_{e,j}$. Por lo tanto el valor esperado de cada uno de los sumandos del χ^2 es $(n_{o,j} - n_{e,j}) \approx 1$. Si $\chi^2 \gg M$ deberíamos concluir que el modelo de la realidad que codificamos dentro de los $n_{e,j}$ no es una buena representación de los datos. Por otro lado, si tuviéramos $\chi^2 \ll M$ deberíamos concluir que estamos ajustando la realidad por encima de la expectativa estadística (el típico caso de algo “demasiado bueno para ser cierto”).



$$\chi_o^2 = 92.86 ; M = 92$$

Variable poissoniana
observada.

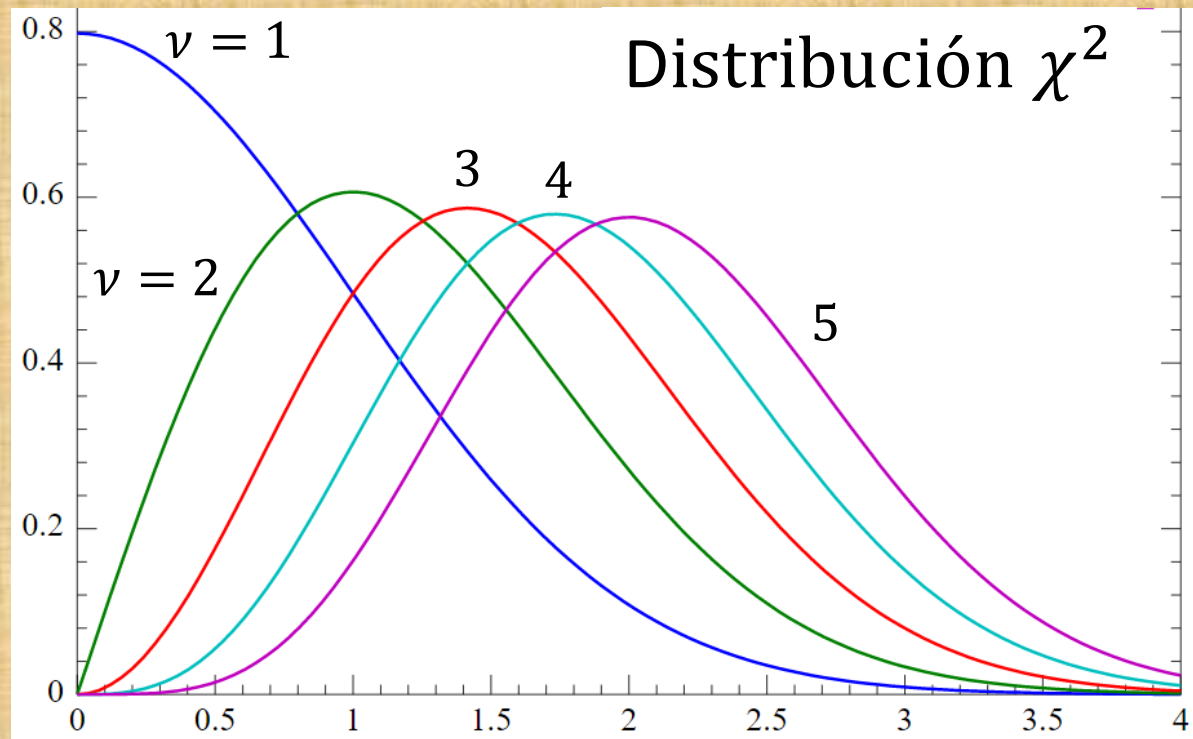
Variable poissoniana
teórica.

¿Qué podemos esperar de χ^2 ? REPASO

$$\chi^2 = \sum_{j=1}^M \left(\frac{n_{o,j} - n_{e,j}}{\sqrt{n_{e,j}}} \right)^2$$

$$P_{\chi}(\chi^2, \nu) = \frac{(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\chi^2/2}}{2^{\nu/2} \Gamma(\nu/2)}$$

Si repitiéramos muchas veces el experimento de generar números aleatorios y comparar los histogramas observados y teóricos, tendríamos una distribución de valores de χ^2 , debido al factor azar. La combinación de poissonianas le da a χ^2 su propia FDP, que puede calcularse, y nos permite saber cual es la probabilidad de que un cierto valor de χ^2 haya sido obtenido por azar.



A medida que ν aumenta, la probabilidad mayor se corre a valores de χ^2 más grandes. Pero ¿Qué es ν ?

¿Qué es ν ?

REPASO

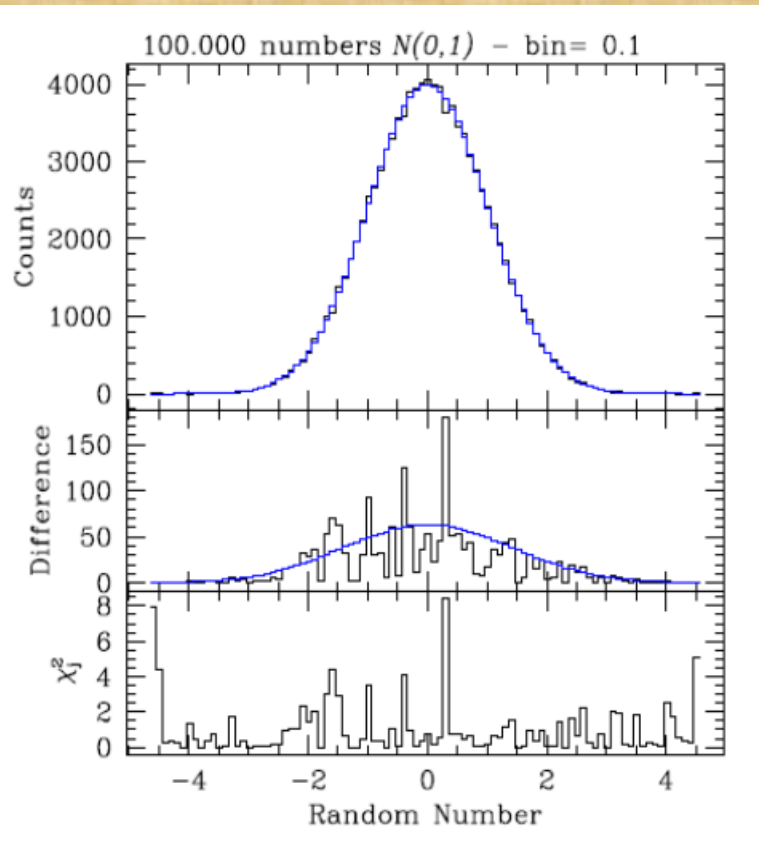
En este contexto ν , los “grados de libertad”, es el número de bins independientes que puedo llenar con la distribución de probabilidad modelo.

Si hay N bins y no se han ajustado parámetros de la distribución con los mismos datos, el número de grados de libertad es $\nu = N - 1$ (¿Por qué?). En el caso del experimento gaussiano anterior, hay 92 bins, pero se han ajustado dos parámetros de la distribución (μ y σ). En este caso $\nu = (N - 1 - 2) = (N - 3) = (92 - 3) = 89$

Con todo esto, estamos en condiciones de cuantificar estadísticamente las diferencias entre la distribución observada y el modelo de la realidad subyacente de la cual se han sacado los datos. Dado un par de valores (χ^2, ν) : ¿Cuál es la probabilidad de obtener ese χ^2 , o un valor mayor, por azar? Si la probabilidad es muy baja, el experimento nos permite rechazar la hipótesis sobre la realidad. Si la probabilidad es alta, podemos decir que el modelo es consistente con las observaciones, aunque nunca podremos *probar* que la realidad es como el modelo.

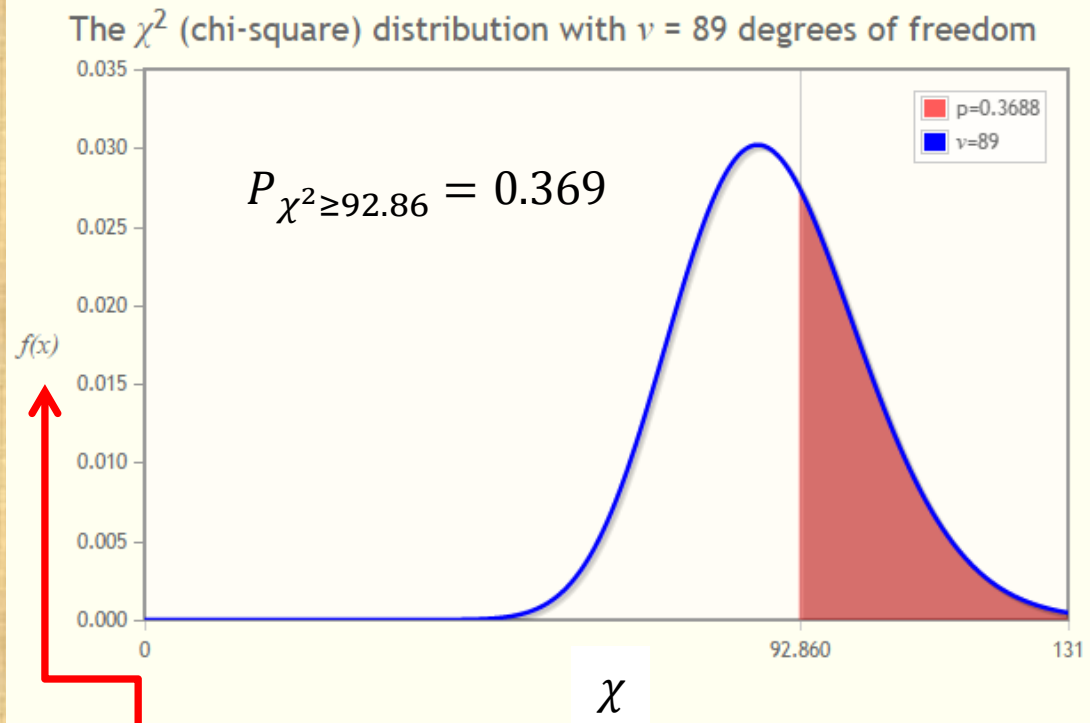
$$P_{\chi}(\chi^2, \nu) = \frac{(\chi^2)^{\frac{1}{2}(\nu-2)} e^{-\chi^2/2}}{2^{\nu/2} \Gamma(\nu/2)} \quad P_{\chi^2 \geq \chi_0^2} = \int_{\chi_0^2}^{\infty} \frac{(x^2)^{\frac{1}{2}(\nu-2)} e^{-x^2/2}}{2^{\nu/2} \Gamma(\nu/2)} dx$$

$$P_{\chi^2 \geq \chi_o^2}?$$



$$\chi^2 = 92.86 ; M = 92 ; \nu = 89$$

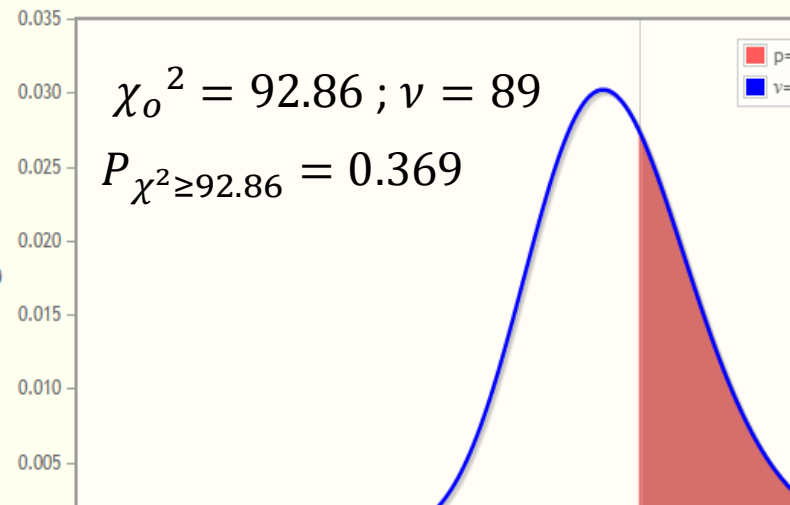
$$P_{\chi^2 \geq \chi_o^2} = \int_{\chi_o^2}^{\infty} \frac{(x^2)^{\frac{1}{2}(\nu-2)} e^{-x^2/2}}{2^{\nu/2} \Gamma(\nu/2)} dx$$



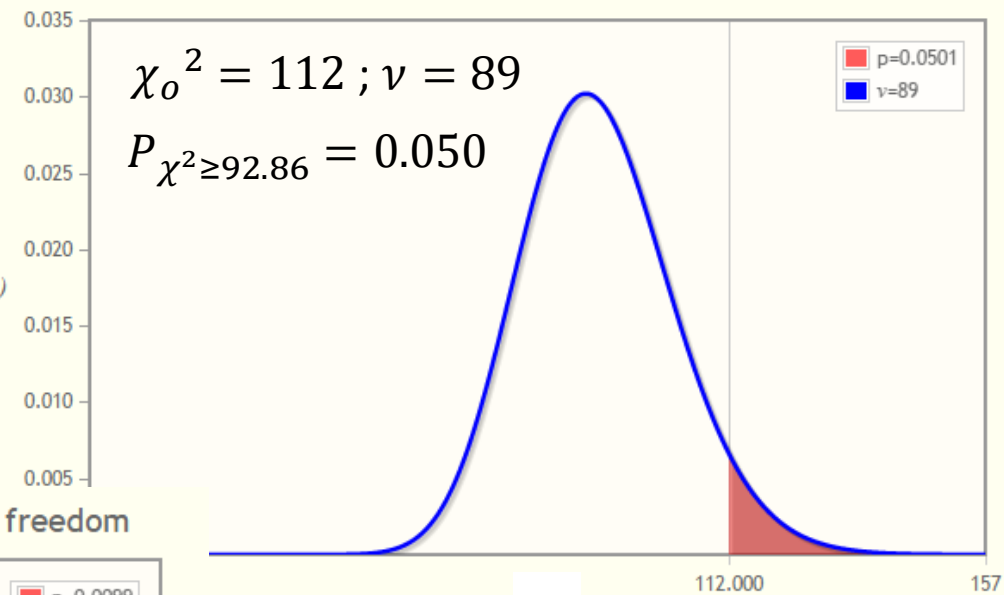
Esto cierra el círculo de interpretación: El χ^2 obtenido tiene $P_{\chi^2 \geq \chi_o^2} \approx 0.37$. Este valor es lo suficientemente grande como para ser *consistente* con la hipótesis de que la distribución subyacente era $N(0,1)$.

¿Cuándo es $P_{\chi^2 \geq \chi_o^2}$ suficientemente baja?

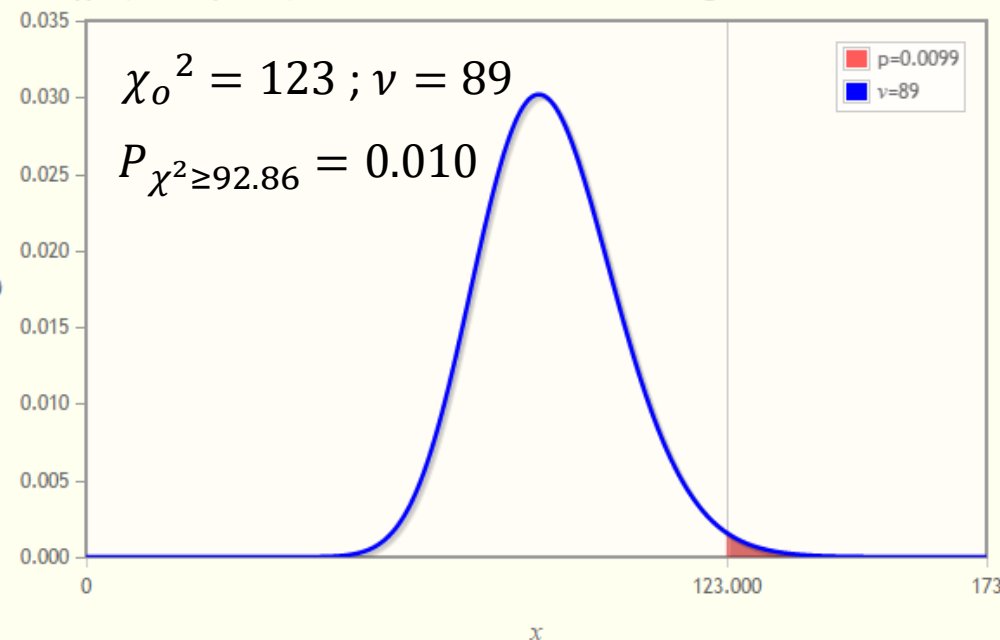
The χ^2 (chi-square) distribution with $\nu = 89$ degrees of freedom



The χ^2 (chi-square) distribution with $\nu = 89$ degrees of freedom



The χ^2 (chi-square) distribution with $\nu = 89$ degrees of freedom



$P_{\chi^2 \geq \chi_o^2} \approx 0.05$, el χ^2 tiene una probabilidad lo suficientemente baja como para rechazar la hipótesis $N(0,1)$ con el 5% de nivel de certeza (caso de arriba). En el caso de la izquierda, la hipótesis $N(0,1)$ se rechaza con el 1% de nivel de certeza.

Resumen:

1. Hacer observaciones (o simularlas).
2. Construir histogramas con los datos.
3. Hacer un modelo de la distribución subyacente.
4. Calcular el histograma esperado.
5. Calcular el χ_o^2
6. Calcular $P_{\chi^2 \geq \chi_o^2}$
7. Decidir si modelo y realidad son consistentes, o puede rechazarse el modelo.

Tarea 1

Tarea 2

Paso de la Tarea 1 a Tarea 2:

1. Estimar los errores, o incertezas, de las medidas
2. Traducir todas las medidas a la misma escala
3. Asegurar que la escala sea absoluta

Requiere:

1. Propagación de errores
2. Correlación

Propagación de incertezas

REPASO

El problema puede plantearse de muchas formas y los libros de estadística más básica suelen poner mucha energía en mostrar, demostrar o justificar la forma de calcular las incertezas de cantidades inciertas cuando son combinadas (incerteza de una suma, un producto, un cociente, etc.). Pero, para personas que ya tienen conocimiento de cálculo, el planteo puede ser conceptualmente trivial: Usar primero las herramientas de cálculo diferencial y aproximar luego el resultado con “diferencias finitas”. Si tenemos una cierta función $z = f(x, y)$ podremos escribir el diferencial de z como:

$$dz = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy$$

Podemos preguntarnos cuál será el “largo” de este dz , pensándolo en un sentido geométrico (imaginando a dx y dy orientados por sus respectivos vectores unitarios \hat{i} y \hat{j} , por ejemplo). Tendremos:

$$|dz| = \sqrt{\left(\frac{\partial f}{\partial x} dx\right)^2 + \left(\frac{\partial f}{\partial y} dy\right)^2}$$

$$\Delta z = \sqrt{\left(\frac{\partial f}{\partial x} \Delta x\right)^2 + \left(\frac{\partial f}{\partial y} \Delta y\right)^2}$$

Propagación de incertezas

REPASO

Si hubiera más variables, tendríamos más términos bajo la raíz en el dz .

$$\Delta z = \sqrt{\left(\frac{\partial f}{\partial x} \Delta x\right)^2 + \left(\frac{\partial f}{\partial y} \Delta y\right)^2 + \left(\frac{\partial f}{\partial t} \Delta t\right)^2 + \dots, \text{etc.}}$$

← Esta fórmula es válida en general en tanto $\Delta x, \Delta y, \Delta t$, etc., sean pequeños e independientes.

Ejemplo, incerteza de una suma o diferencia : $z = x \pm y$

$$\frac{\partial f}{\partial x} = 1 ; \frac{\partial f}{\partial y} = \pm 1 ; \Delta z = \sqrt{\Delta x^2 + \Delta y^2}$$

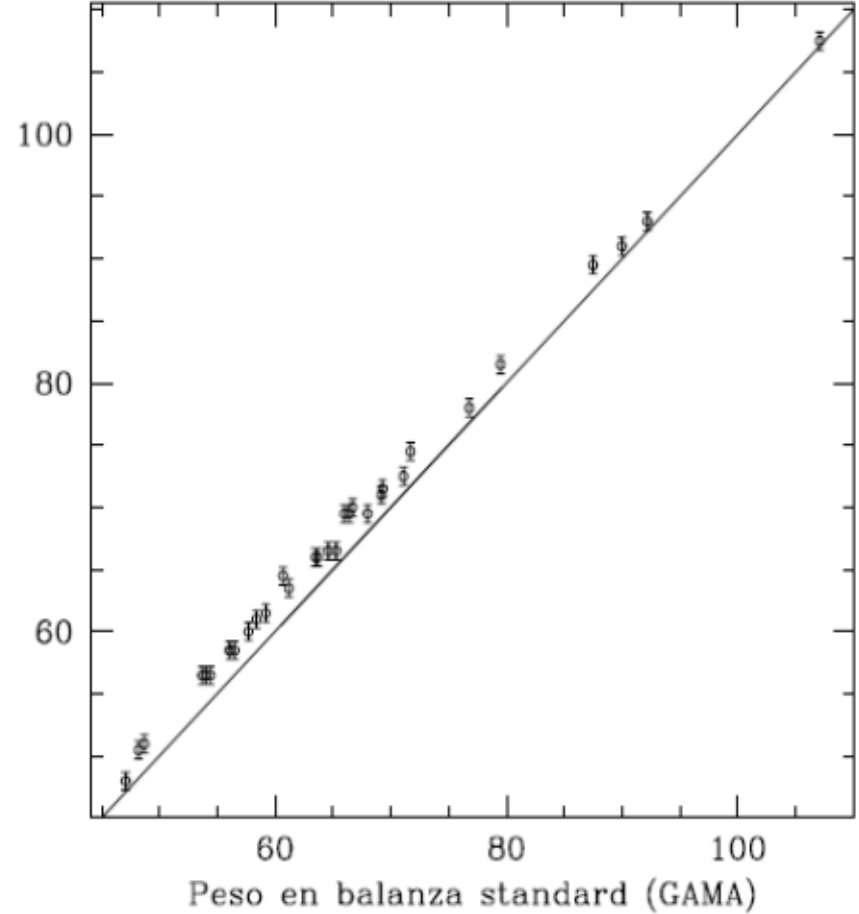
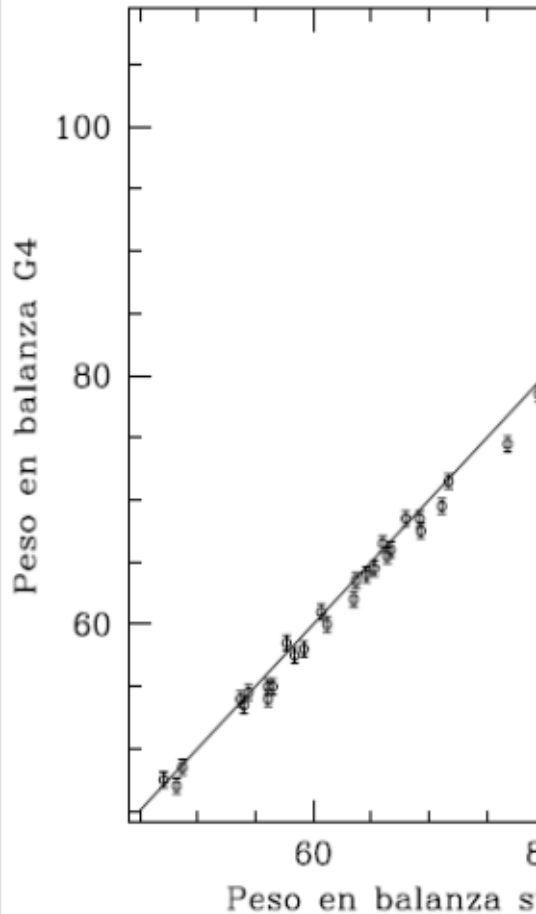
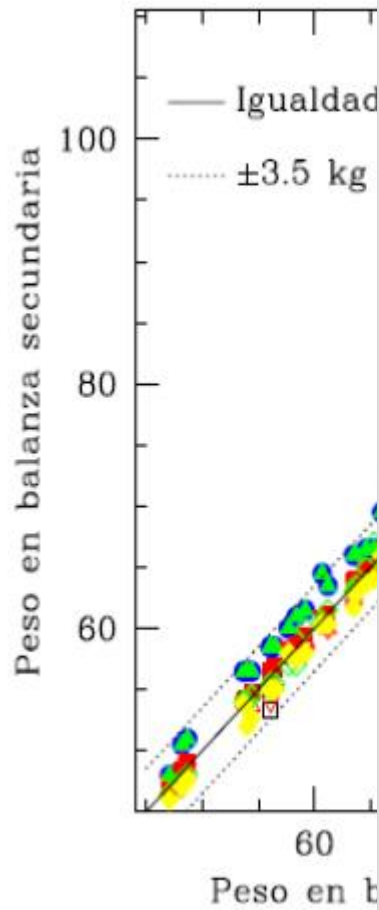
Ejemplo, incerteza de un cociente: $z = x/y$

$$\frac{\partial f}{\partial x} = \frac{1}{y} ; \frac{\partial f}{\partial y} = -\frac{x}{y^2} ; \Delta z = \sqrt{\frac{\Delta x^2}{y^2} + \frac{(x\Delta y)^2}{y^4}} \rightarrow \frac{\Delta z}{z} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta y}{y}\right)^2}$$

Ejemplo, incerteza de un producto: $z = xy$

$$\frac{\partial f}{\partial x} = y ; \frac{\partial f}{\partial y} = x ; \Delta z = \sqrt{(y\Delta x)^2 + (x\Delta y)^2} \rightarrow \frac{\Delta z}{z} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta y}{y}\right)^2}$$

Correlación: Peso medido con \neq balanzas



Desarrollo en pizarra:
Método de “cuadrados mínimos”

Método de “cuadrados mínimos”

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

$$y_i = ax_i + b$$

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

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

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

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

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

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

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

Método de “cuadrados mínimos”

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

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

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

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

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

Incerteza en los parámetros calculados

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

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

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

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

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

Incerteza en los parámetros calculados

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

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

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

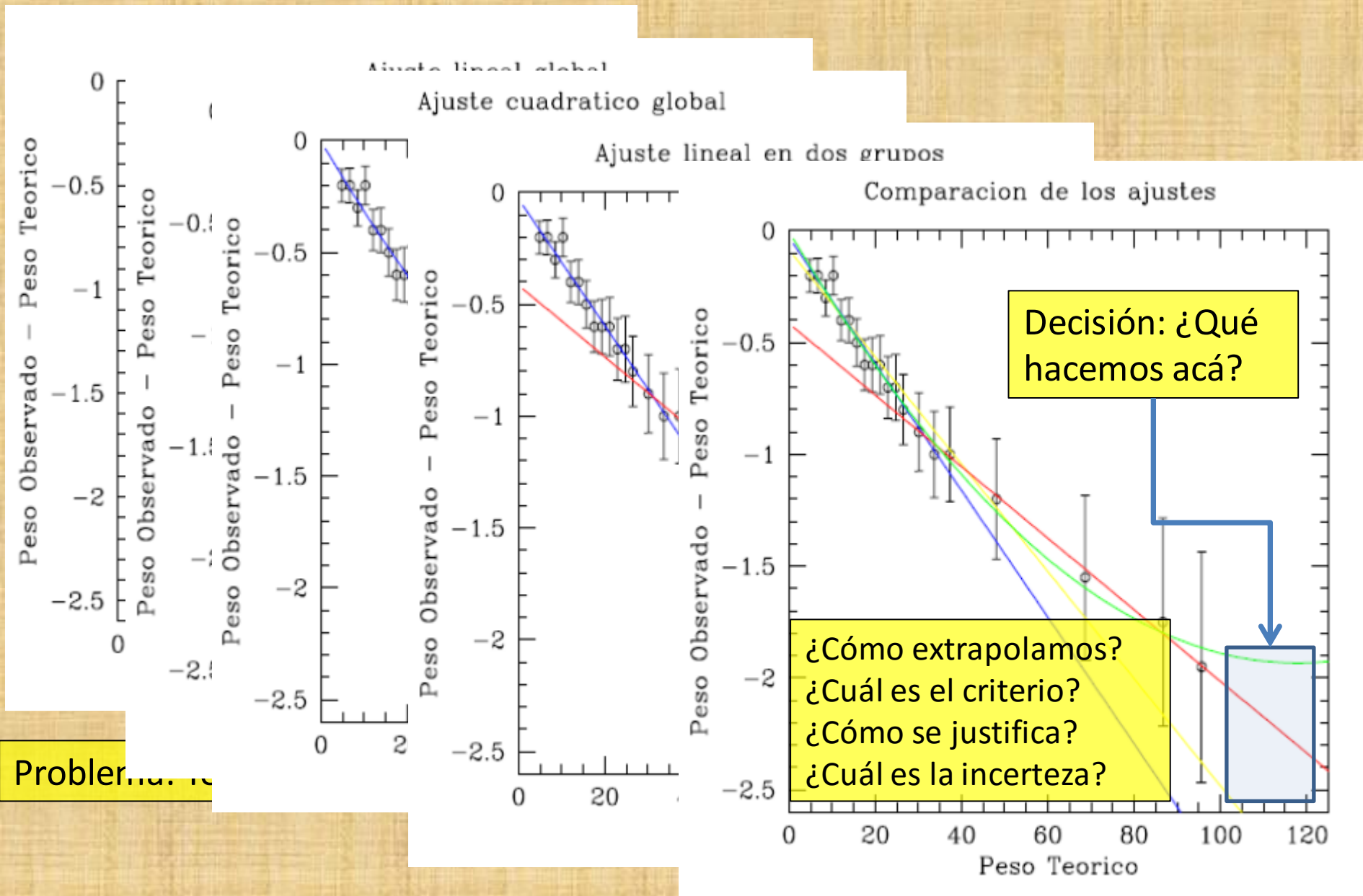
$$\sigma_a^2 = \frac{1}{\Delta} \sum_{i=1}^N \frac{1}{\sigma_i^2}$$

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

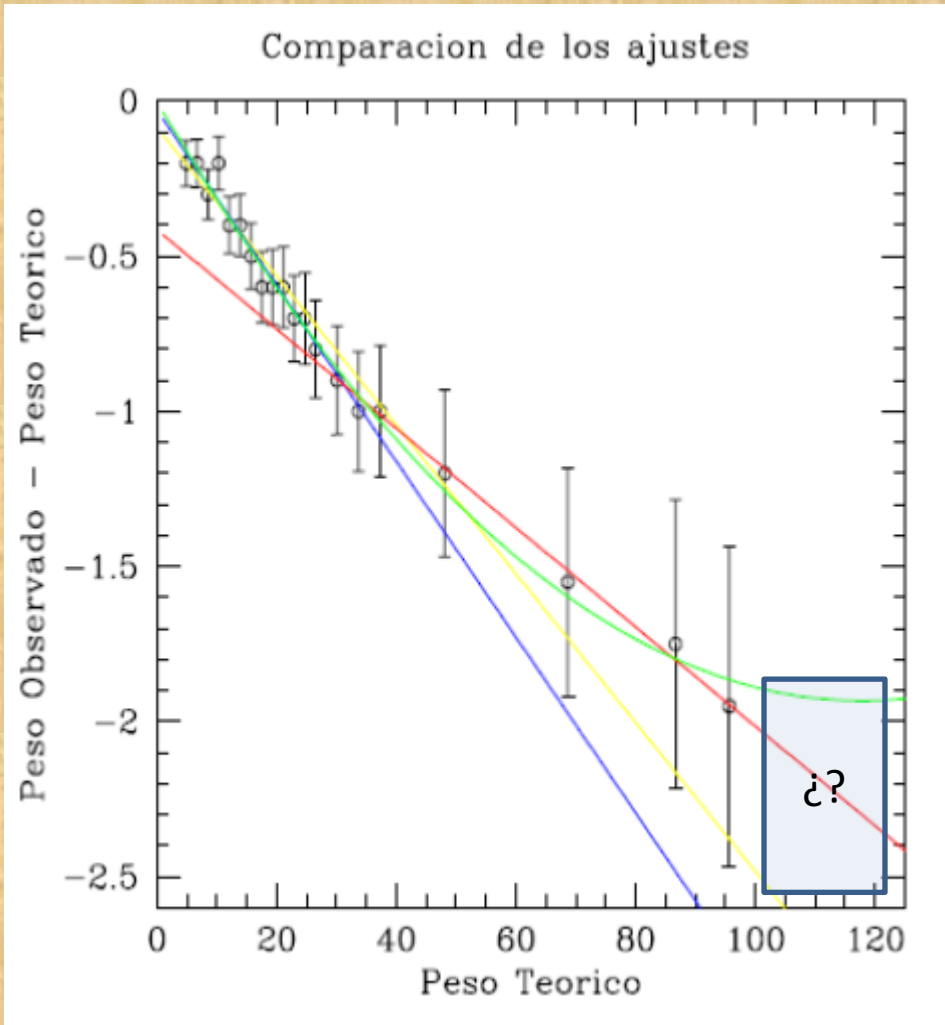
Correlación: Error sistemático de GAMA



Correlación: Error sistemático de GAMA



Correlación: Error sistemático de GAMA



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?).

Comparación de dos distribuciones observadas

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

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.*

Teatro municipal de Estocolmo, Diciembre del 2011



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

ADAM G. RIESS,¹ AL
 PETER M. GARNAVI
 B. LEIBUNDGUT,

1022

RIESS ET AL.

Vol. 116

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

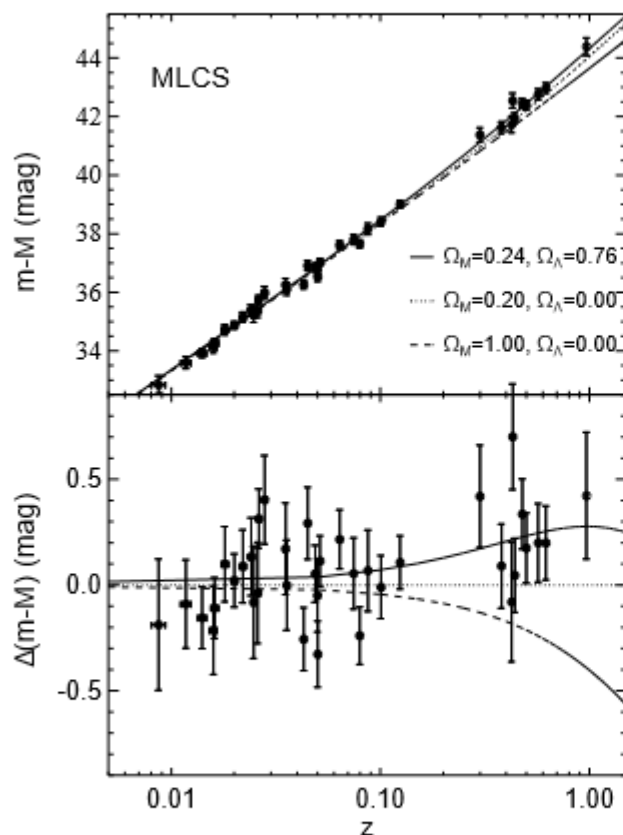


FIG. 4.—MLCS SNe Ia Hubble diagram. The upper panel shows the Hubble diagram for the low-redshift and high-redshift SNe Ia samples with distances measured from the MLCS method (Riess et al. 1995, 1996a; Appendix of this paper). Overplotted are three cosmologies: “low” and “high” Ω_M with $\Omega_\Lambda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

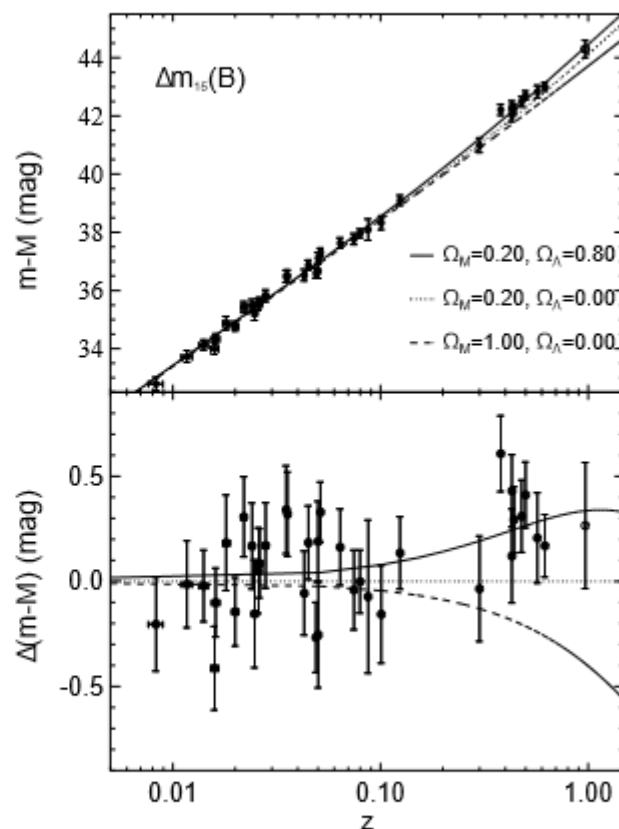


FIG. 5.— $\Delta m_{15}(B)$ SN Ia Hubble diagram. The upper panel shows the Hubble diagram for the low-redshift and high-redshift SNe Ia samples with distances measured from the template-fitting method parameterized by $\Delta m_{15}(B)$ (Hamuy et al. 1995, 1996d). Overplotted are three cosmologies: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

Fin de ppt de Clase 7