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

Equipo docente:

Profesor: Alejandro Clocchiatti

Ayudantes:

Francisco Aros (TM6)

Nicolás Castro (TL4)

TM6: Tutoría del martes en módulo 6

TL4: Tutoría del lunes en módulo 4

Nuestro Semestre 2016-1

AST0212

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 Semana 11	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 C1×3	18
19	20	21	22	23	24	25
26	27	28	29	30	1 Jul	2
3	4	5	6	7	8 Notas	9

PDF Calendar by www.pdfcalendar.com

C0 ✓

← Control 1
Reparto Tarea 1

← Control 2
← Reparto T2

← Entrega Tarea 1

← Entrega Tarea 2

Tutorías día lunes
Módulo 4:
Nicolás Castro

Tutorías día martes
Módulo 6:
Francisco Aros

Clase previa (Clase 5):

REPASO

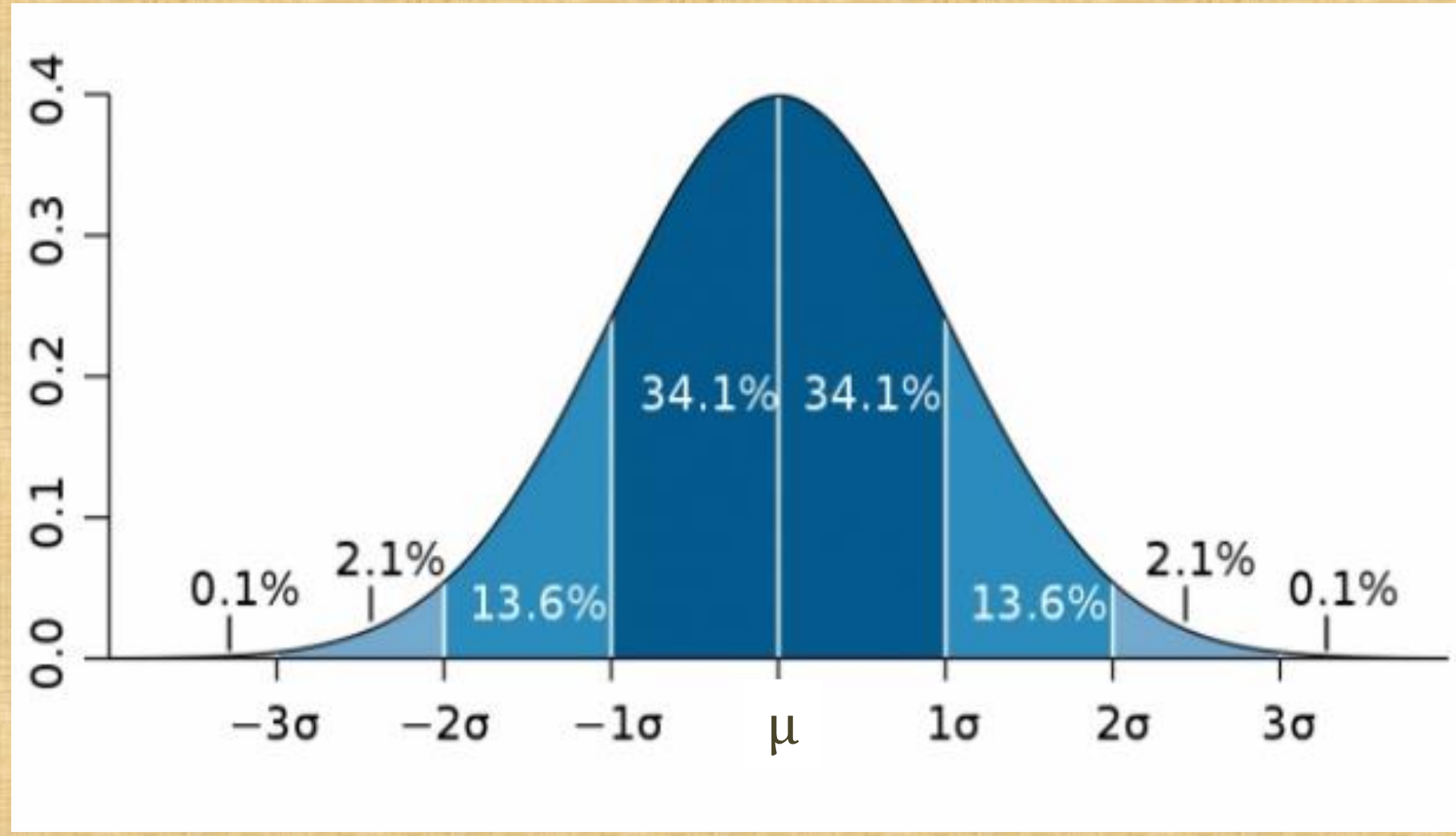
1. Herramienta Linux de selección de datos en archivos organizados en columnas: *awk* ✗
2. Repaso de temas críticos de la clase previa ✓
   1. FDPs que hay que conocer: Constante, Poisson y Gauss.
   2. Modelos de la realidad, distribución subyacente.
3. Test modelo vs. realidad: χ^2 explicado. ✓
4.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

Esta clase (Clase 6):

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.

FDP de Gauss

REPASO



$$N(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

FDP de Gauss

REPASO

La forma usual de llevar a cabo ese cálculo es convirtiendo la $N(\mu,\sigma)$ en $N(0,1)$ con el cambio de variables $x' = (x-\mu)/\sigma$

$$P_{(x_1<x<x_2)} = \frac{1}{\sqrt{2\pi}}\int_{x'_1}^{x'_2} e^{-\frac{x^2}{2}}dx = Erf(x'_2) - Erf(x'_1)$$

donde *Erf(x)* es

$$Erf(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{x^2}{2}}dx$$

La *erf(x)* nos permite calcular un primer "modelo de la realidad" para comparar con nuestras observaciones o con nuestra imaginación.
Si el modelo de la realidad correcto es una distribución $N(\mu,\sigma)$, entonces el valor esperado de casos que caerán en el intervalo de ancho Δx centrado en x_j será:

$$n_{e,j} = N_T\{Erf(x_j + \Delta x/2) - Erf(x_j - \Delta x/2)\}$$

donde N_T es el número total de casos (es decir, la suma total del histograma).

FDP de Gauss como modelo de la realidad

REPASO

Habiendo hecho esto puedo desarrollar un test cuantitativo para medir cuán diferente es el histograma observado del que predice el modelo. El test que vamos a usar es el llamado χ^2, que se construye sumando los cuadrados de las diferencias entre las cuentas real, observadas, y las predichas por el modelo (o esperadas), bin a bin, normalizadas por las cuentas esperadas. Hay razones fundadas para usar este estimador, pero por hoy (clase 4) sólo tomaremos nota de la receta:

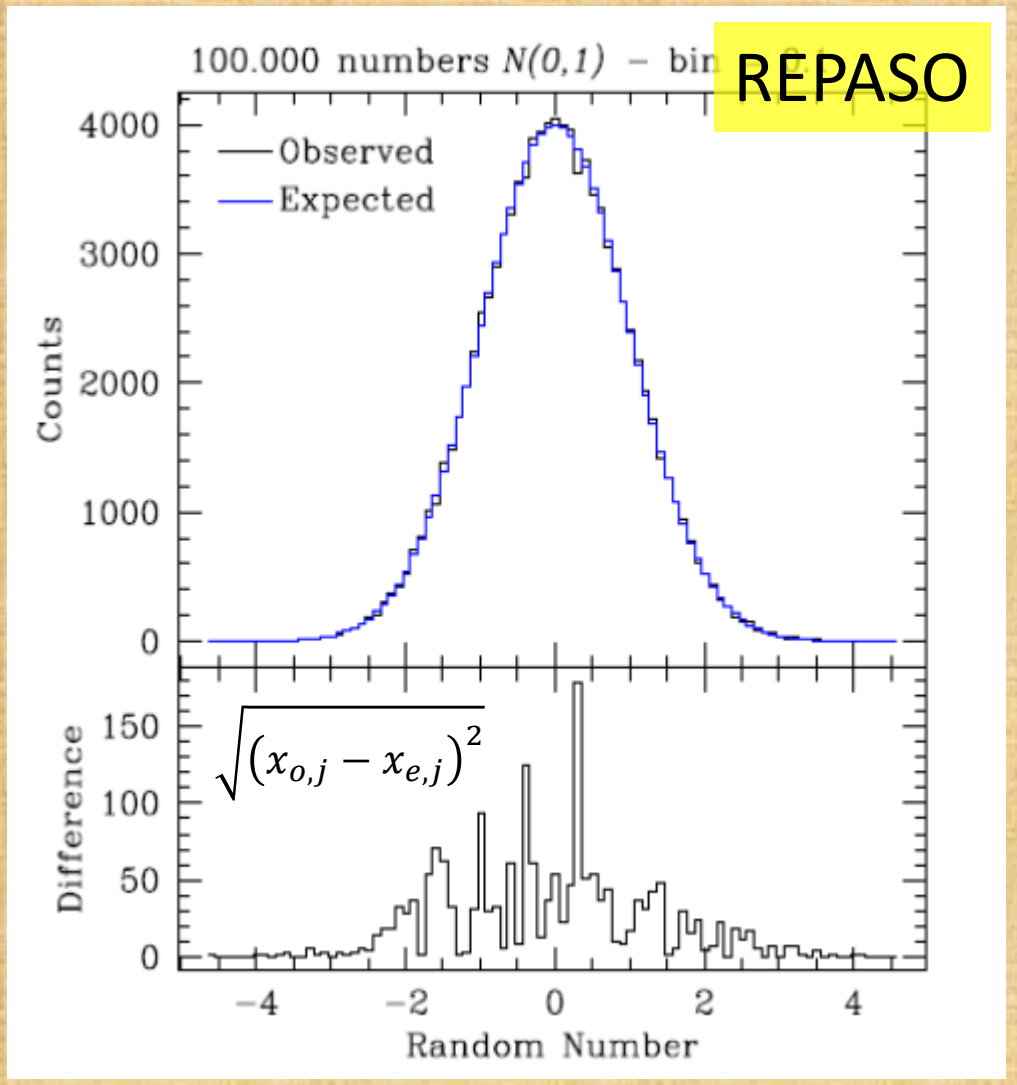
$$\chi^2 = \sum_{j=1}^{M} \frac{\left(n_{o,j} - n_{e,j}\right)^2}{n_{e,j}}$$

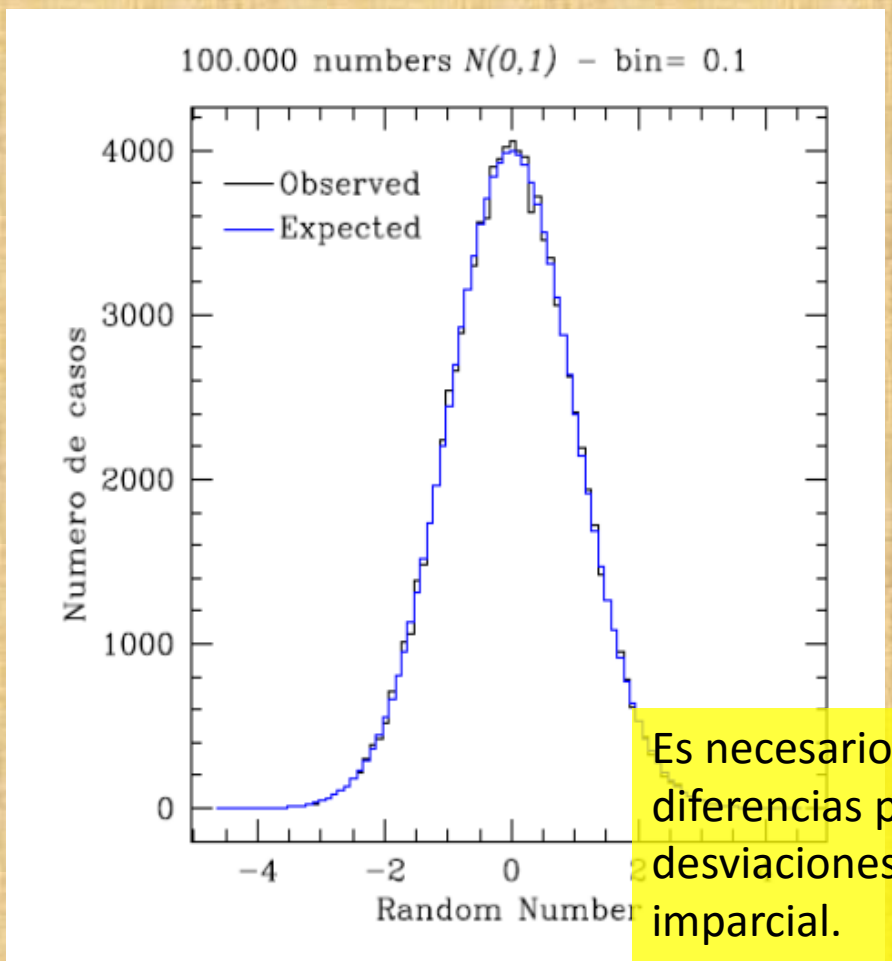
$n_{o,j}$ es el número de casos observados en el bin *j*, $n_{e,j}$ es el número de casos esperados en el bin *j*, calculados como se indica en la imagen previa, de acuerdo a la distribución $N(\mu, \sigma)$, donde μ y σ son el valor medio y la dispersión (o desviación estándar), medidas para el histograma que estamos tratando de representar, y la suma se extiende a los *M* bins que tenga el histograma.

¿ Qué es χ^2?

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

REPASO

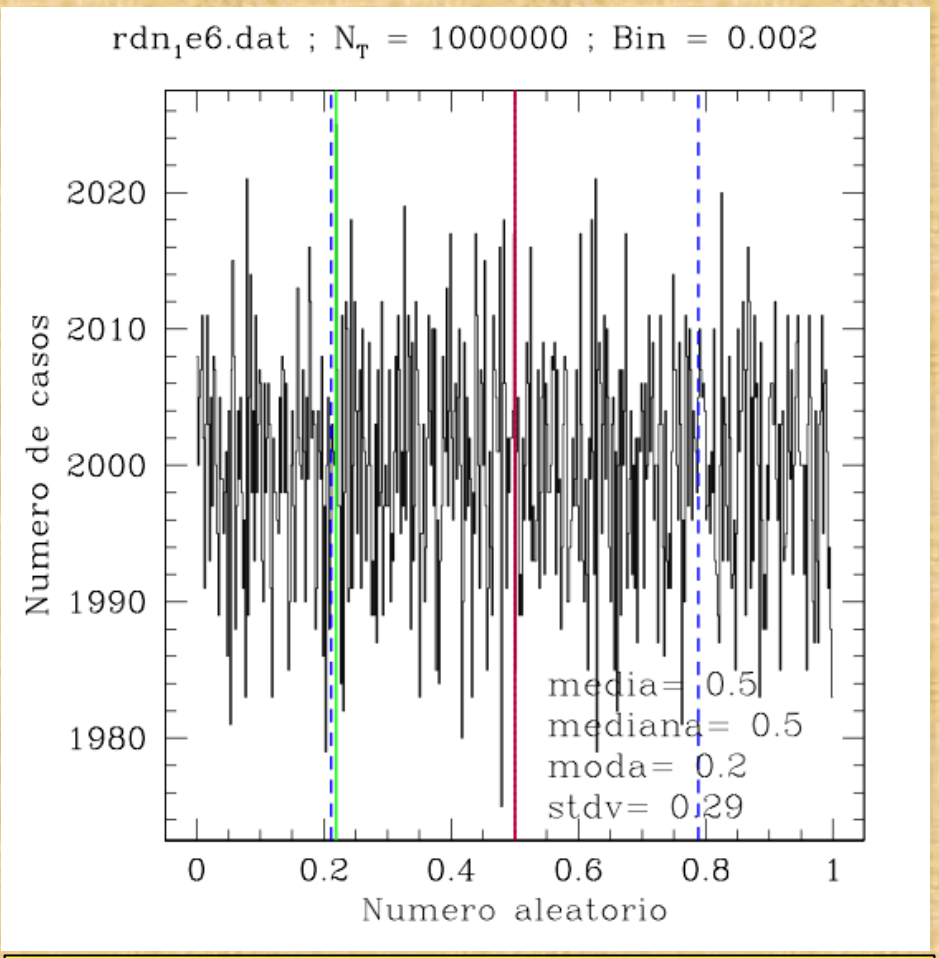




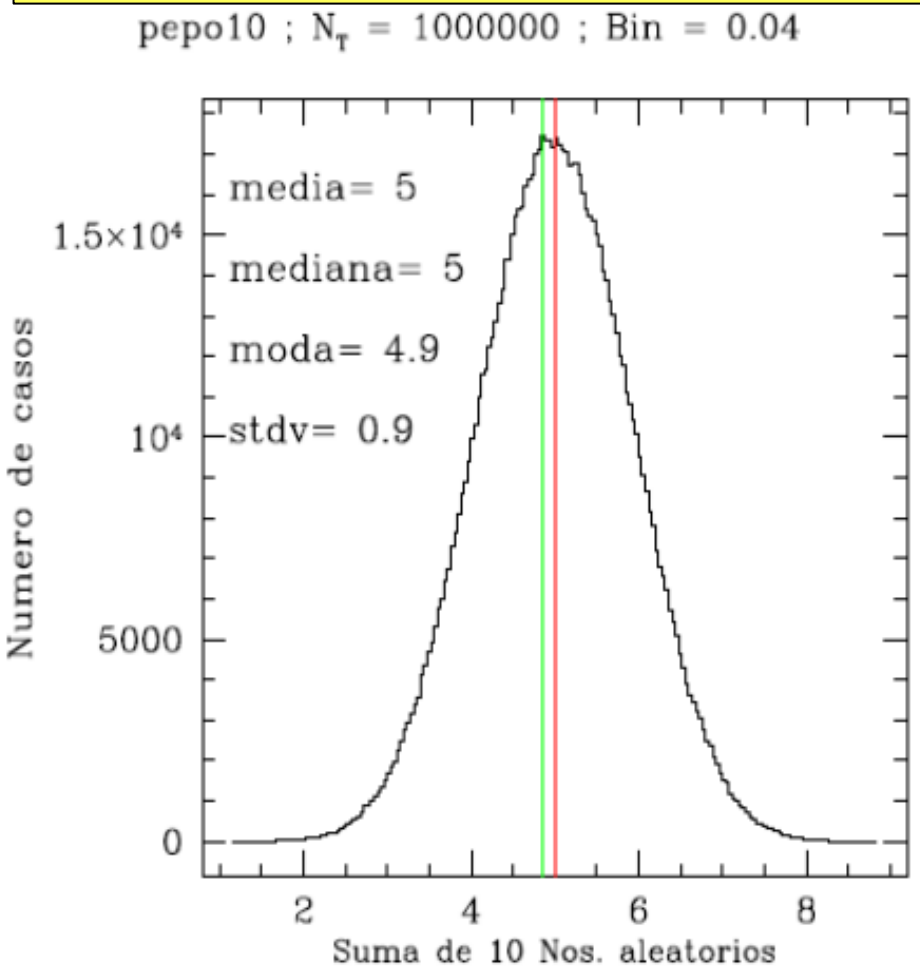
Es necesario normalizar las diferencias para medir las desviaciones de manera imparcial.

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

¿Cuál es un buen modelo para σ_e ? REPASO

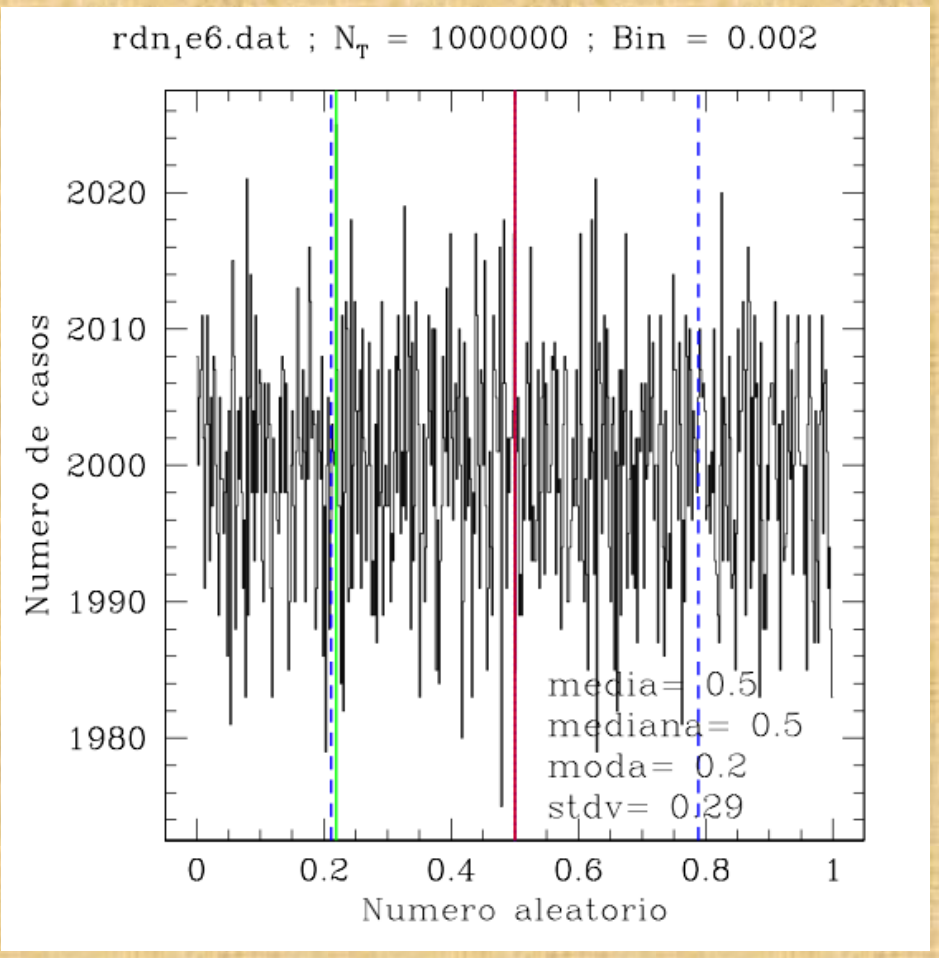


El proceso de llenar bins con números sacados al azar una distribución subyacente sigue <u>siempre</u> una FDP de Poisson. μ (y σ) dependerán en general del bin.

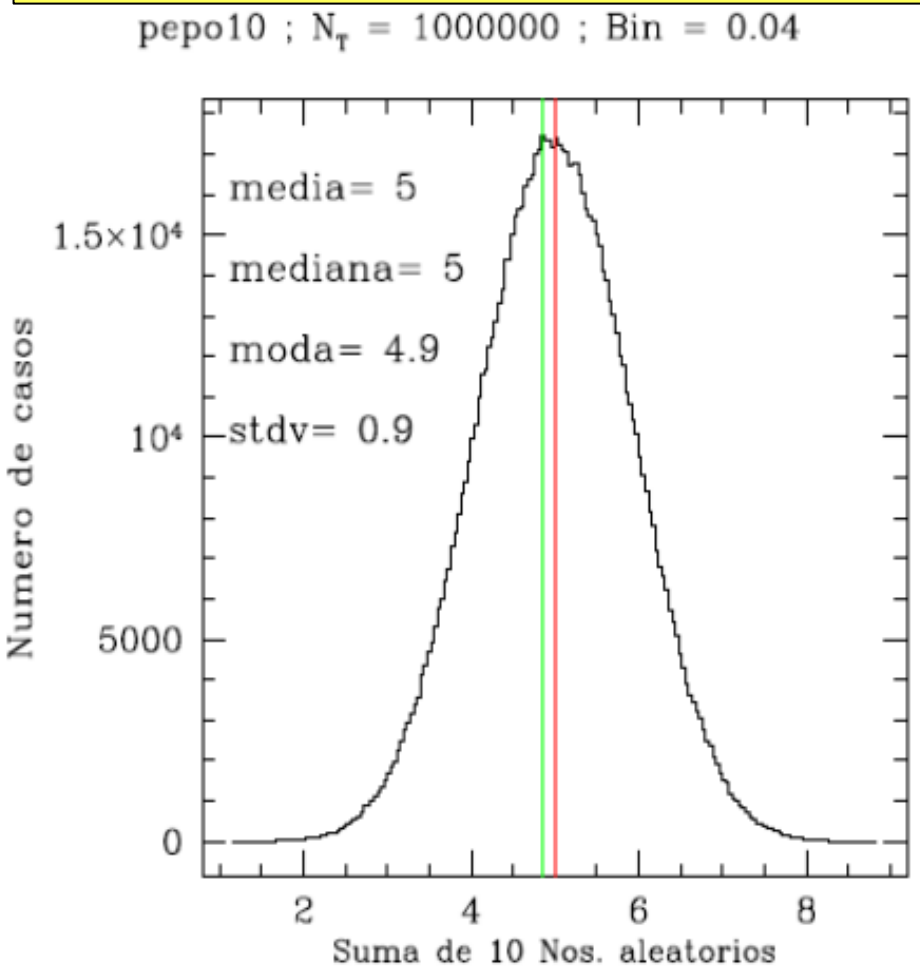


El número de cuentas en cada bin sigue una distribución de Poisson, con valor medio μ y dispersión $\sigma = \sqrt{\mu}$. $n_{e,j} = \mu$, $\forall$ *j, pero esto es propiedad de la PDF, <u>no del proceso que llena los bins</u>.*

¿Cuál es un buen modelo para σ_e?



El proceso de llenar bins con números sacados al azar una distribución subyacente sigue <u>siempre</u> una FDP de Poisson. μ (y σ) dependerán en general del bin.



El número de cuentas en cada bin sigue una distribución de Poisson, con valor medio μ y $\sigma = \sqrt{\mu}$. En este caso $n_{e,j} = \mu$ (cte.), $\forall$ *j, pero esto es propiedad de la PDF, <u>no del proceso que llena los bins</u>.*

¿ Qué es χ^2?

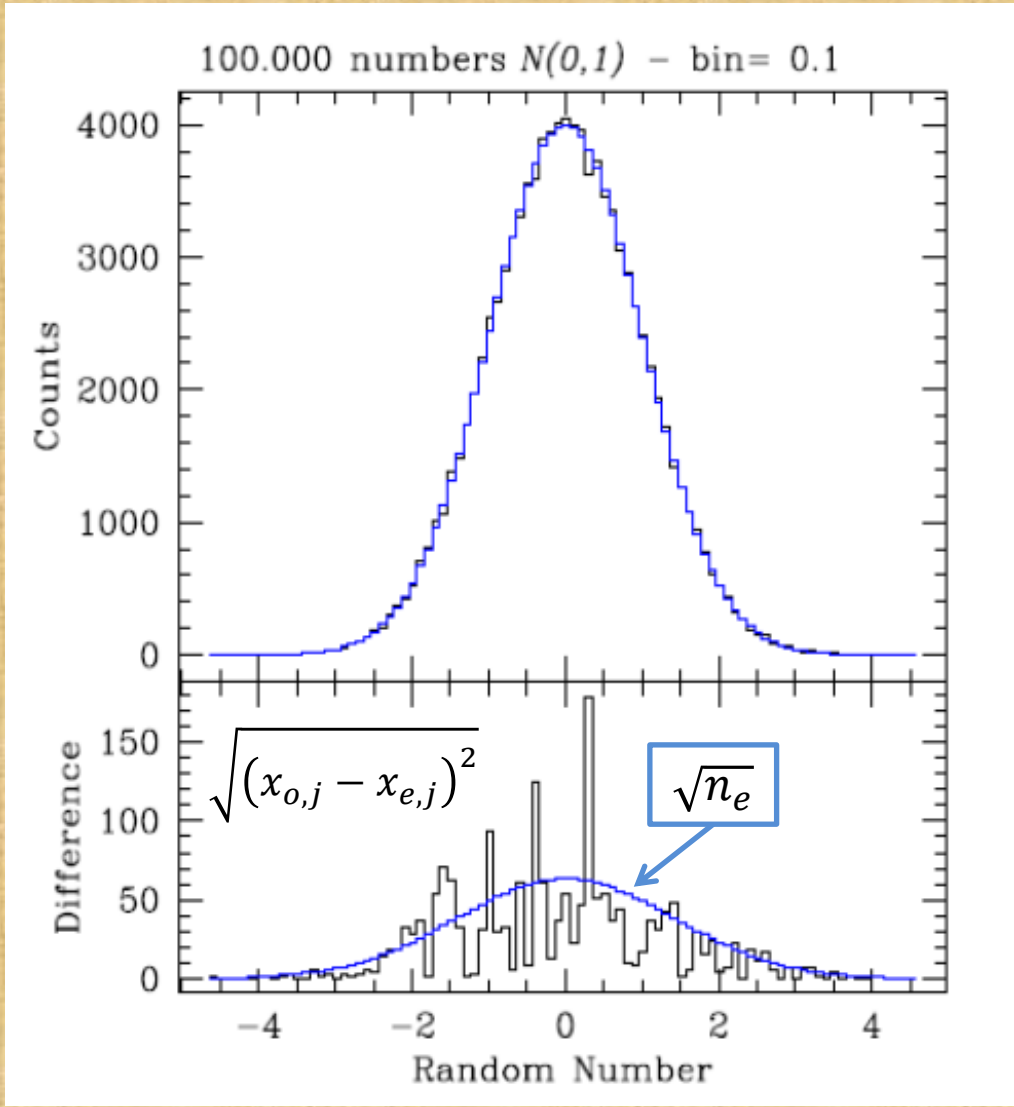
REPASO

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:

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

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

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



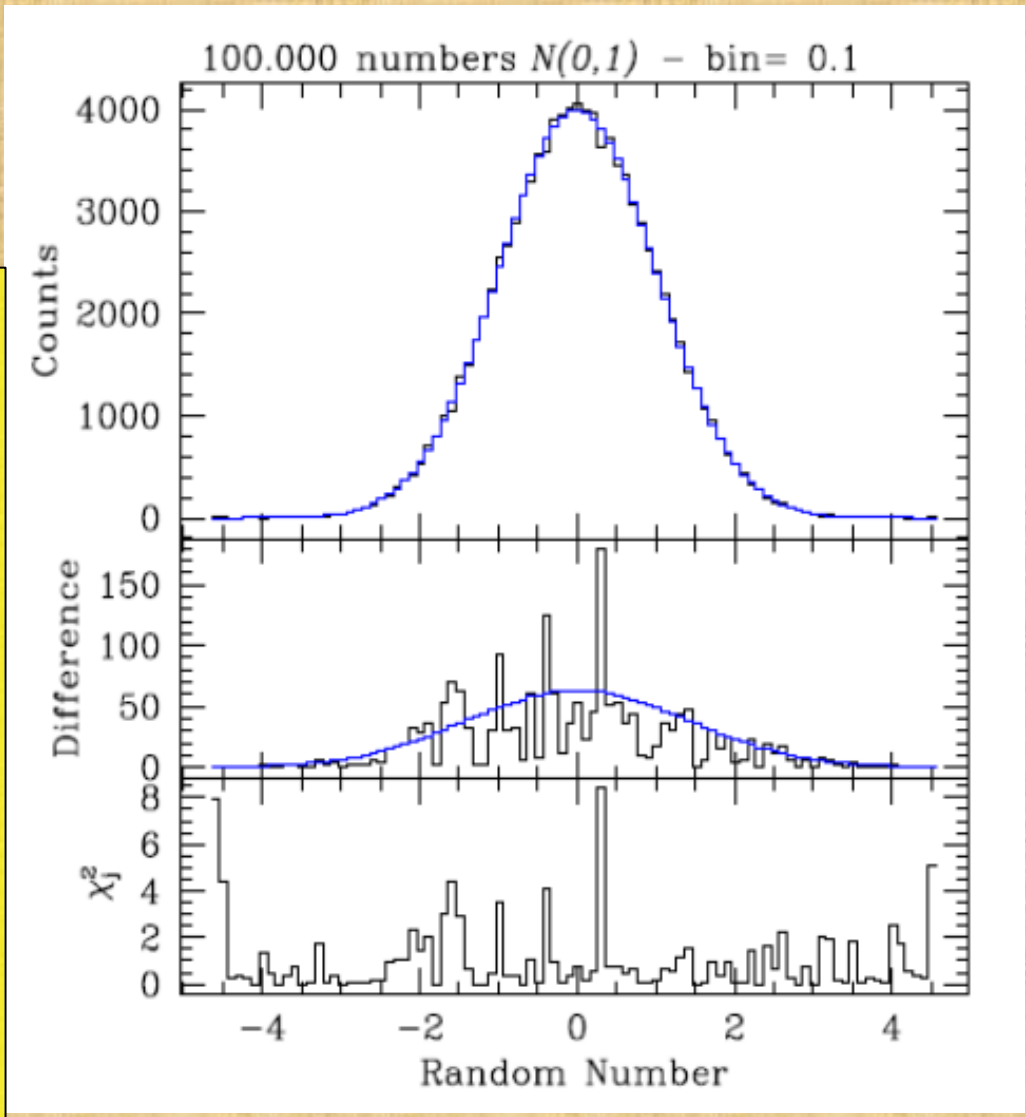
Esta última es la expresión que había puesto como “receta” al final de la Clase 4.

¿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 expectación estadística (el típico caso de algo "demasiado bueno para ser cierto").



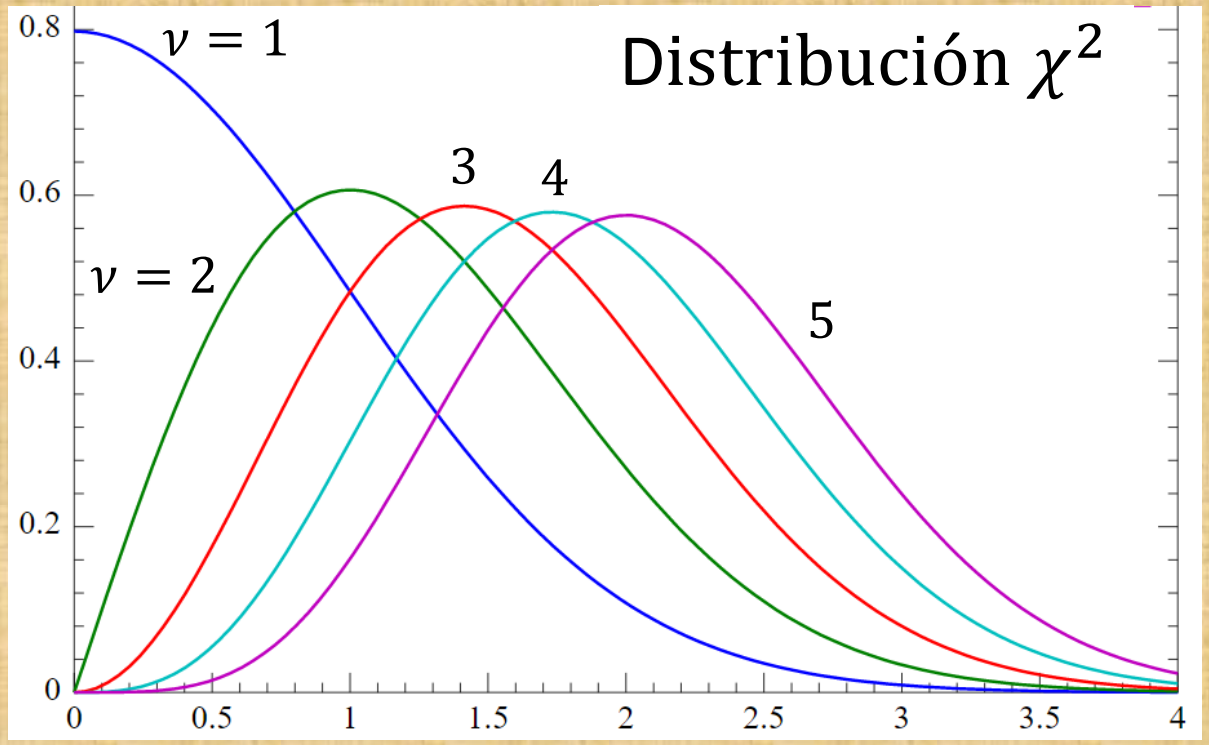
$${\chi_o}^2 = 92.86 \; ; M = 92$$

¿Qué podemos esperar de χ^2?

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

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

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?

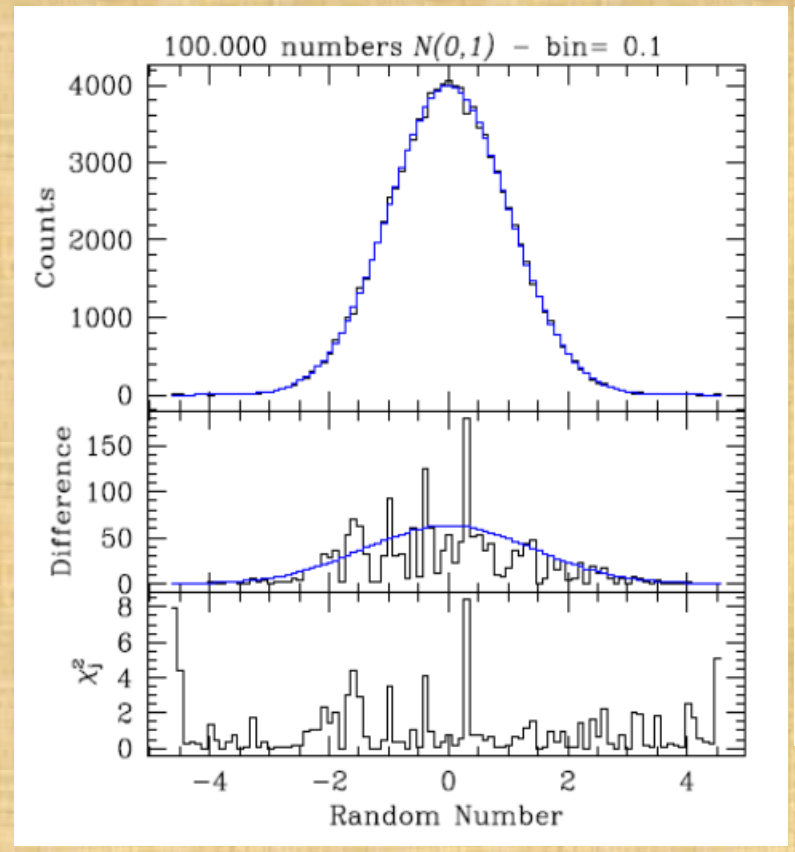
En este contexto ν, los "grados de libertad", es el número de bins independientes que puedo llenar con la distribución de probabilidad modelo.

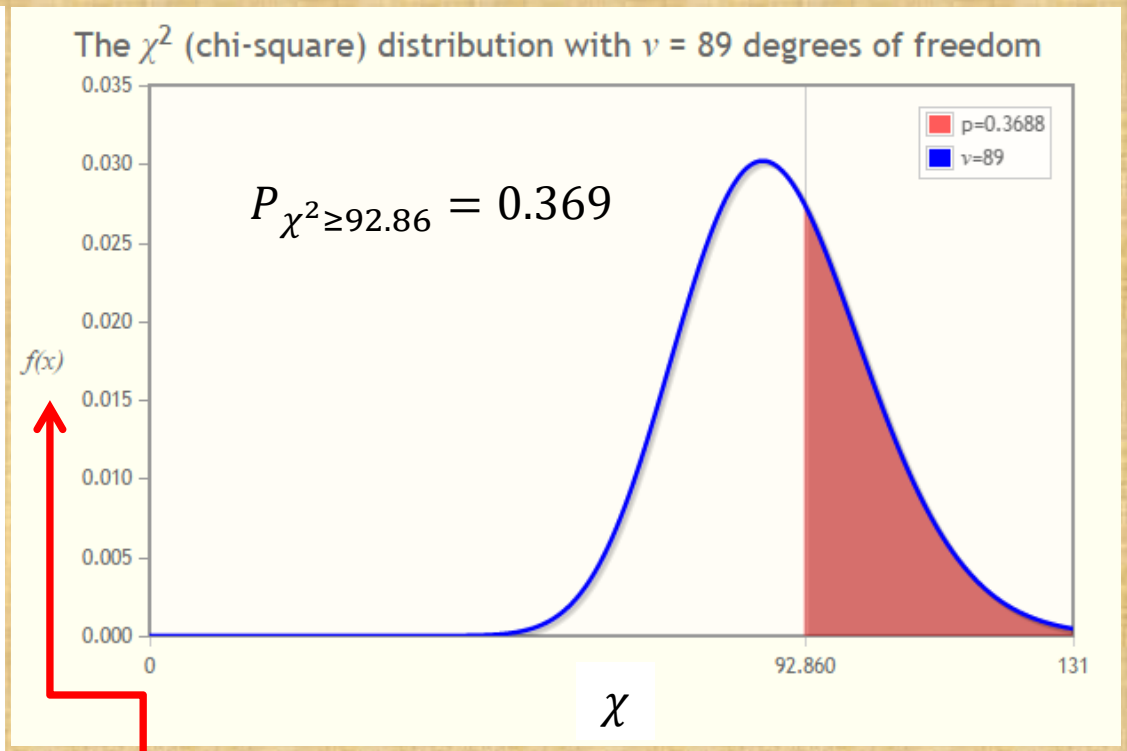
Si hay N bins y no se han ajustado parámetros de la distribución con los mismos datos, el número de grados de libertad es $\nu = N - 1$ (¿Por qué?). En el caso del experimento gaussiano anterior, hay 92 bins, pero se han ajustado dos parámetros de la distribución (μ y σ). En este caso $\nu = (N - 1 -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

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.

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

$¿\, P_{\chi^2 \geq \chi_o{}^2}?$





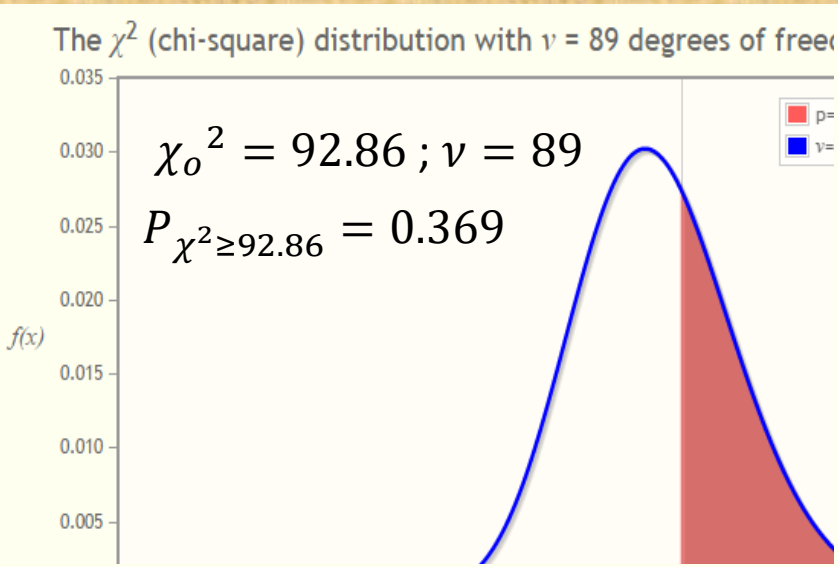
$\chi^2 = 92.86\,;\, M = 92\,;\, \nu = 89$

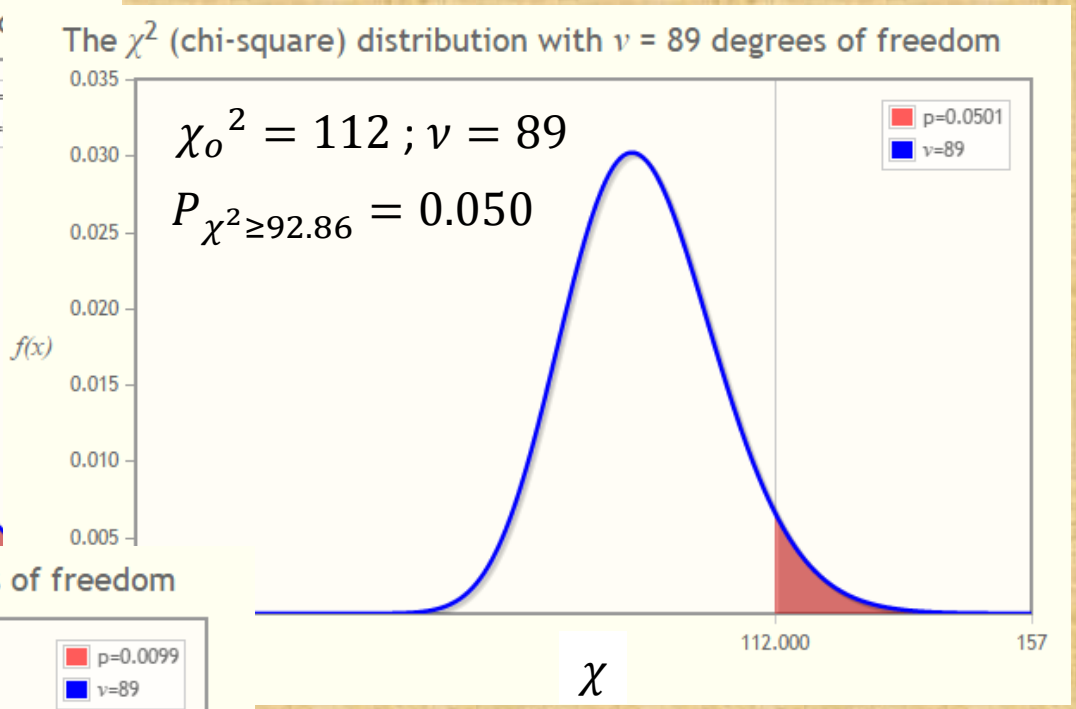
$$P_{\chi^2 \geq \chi_o{}^2} = \int_{\chi_o{}^2}^{\infty} \frac{(x^2)^{\frac{1}{2}(\nu-2)} e^{-x^2/2}}{2^{\nu/2}\Gamma(\nu/2)}\, dx$$

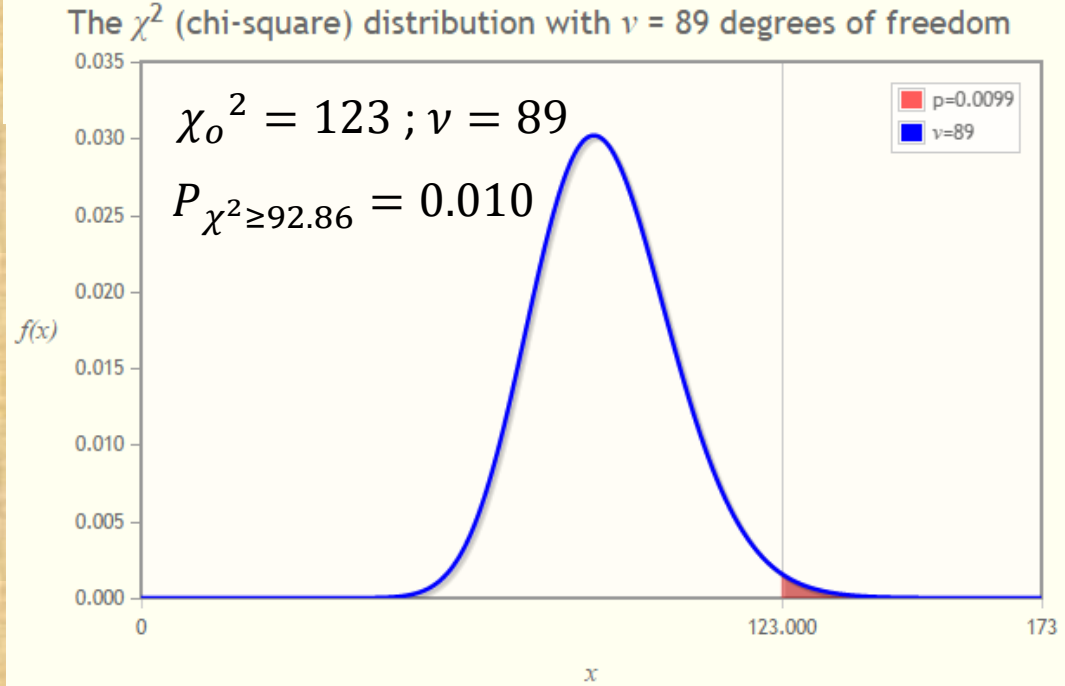
Esto cierra el círculo de interpretación: El χ^2 obtenido tiene $P_{\chi^2 \geq \chi_o{}^2} \approx 0.37$. Este valor es lo suficientemente grande como para ser *consistente* con la hipótesis de que la distribución subyacente era $N(0,1)$.

http://www.di-mgt.com.au/chisquare-calculator.html

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







$P_{\chi^2 \geq \chi_o{}^2} \approx 0.05$, 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 $\chi_o{}^2$

Tarea 1

6. Calcular $P_{\chi^2 \geq \chi_o{}^2}$
7. Decidir si modelo y realidad son consistentes, o puede rechazarse el modelo.

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

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 $\check{\imath}$ y $\check{\jmath}$, por ejemplo). Tendremos:

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

Propagación de incertezas

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

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

← 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 : $z = x \pm y$

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

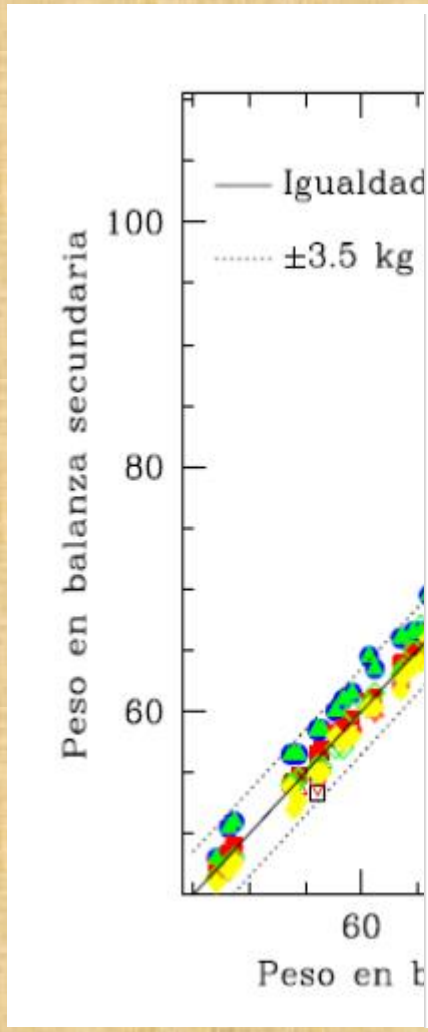
Ejemplo, incerteza de un cociente: $z = x/y$

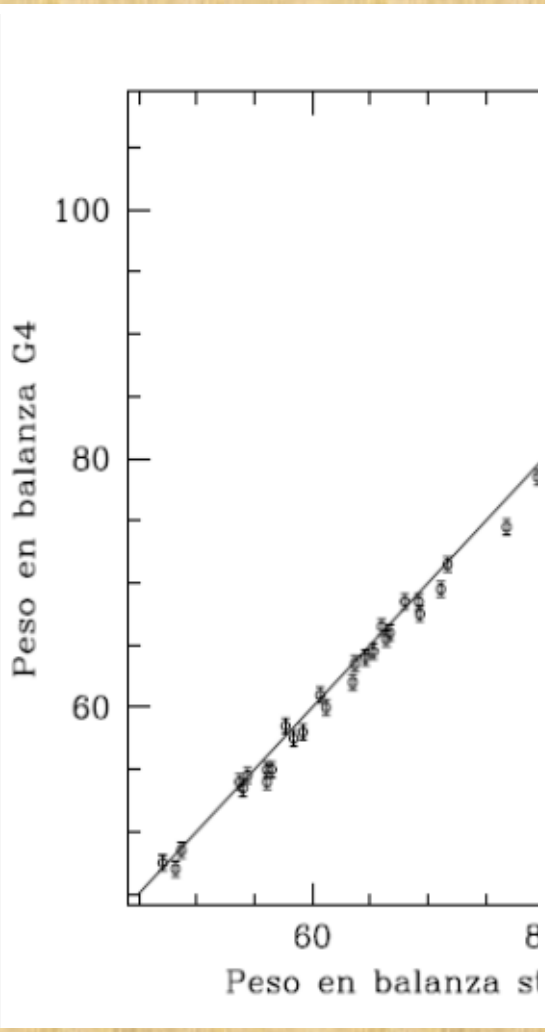
$$\frac{\partial f}{\partial x} = \frac{1}{y}; \frac{\partial f}{\partial y} = -\frac{x}{y^2}; \Delta z = \sqrt{\frac{\Delta x^2}{y^2} + \frac{(x\Delta y)^2}{y^4}} \rightarrow \frac{\Delta z}{z} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta y}{y}\right)^2}$$

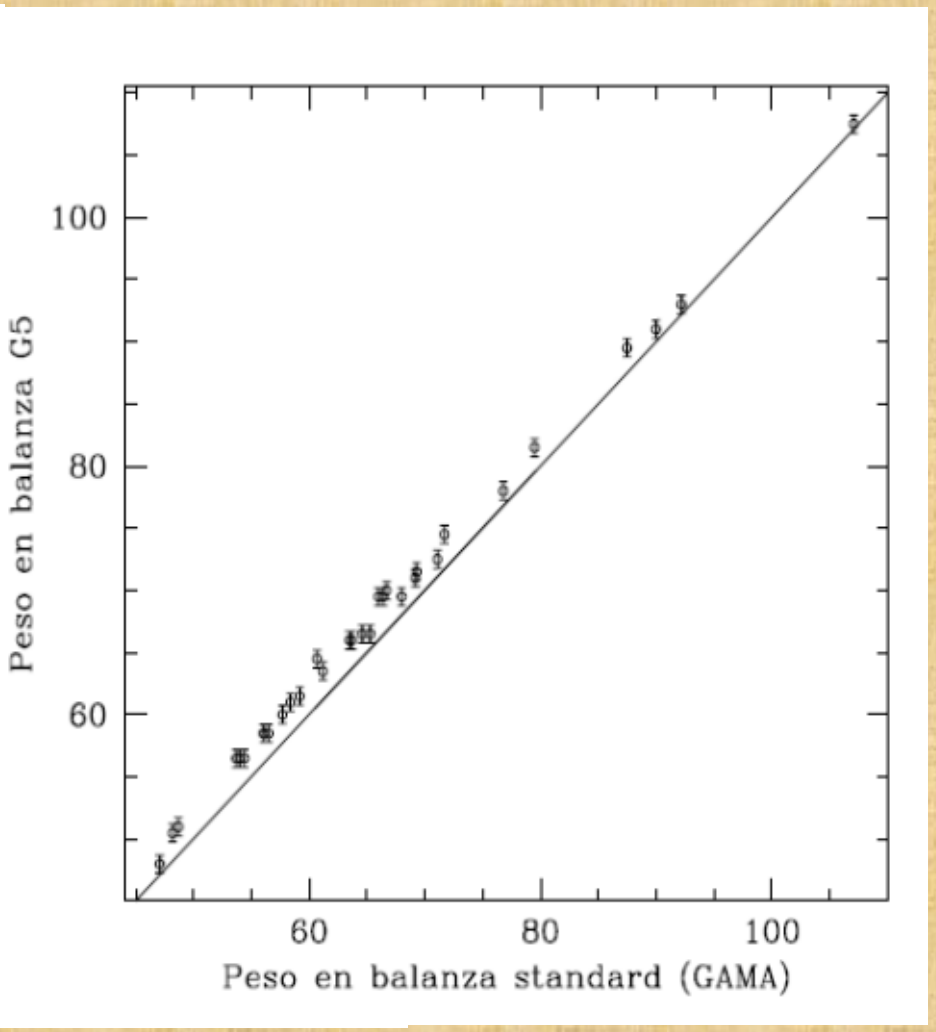
Ejemplo, incerteza de un producto: $z = xy$

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

Correlación: Peso medido con ≠ balanzas







Desarrollo en pizarra:
Método de “cuadrados mínimos”