
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:

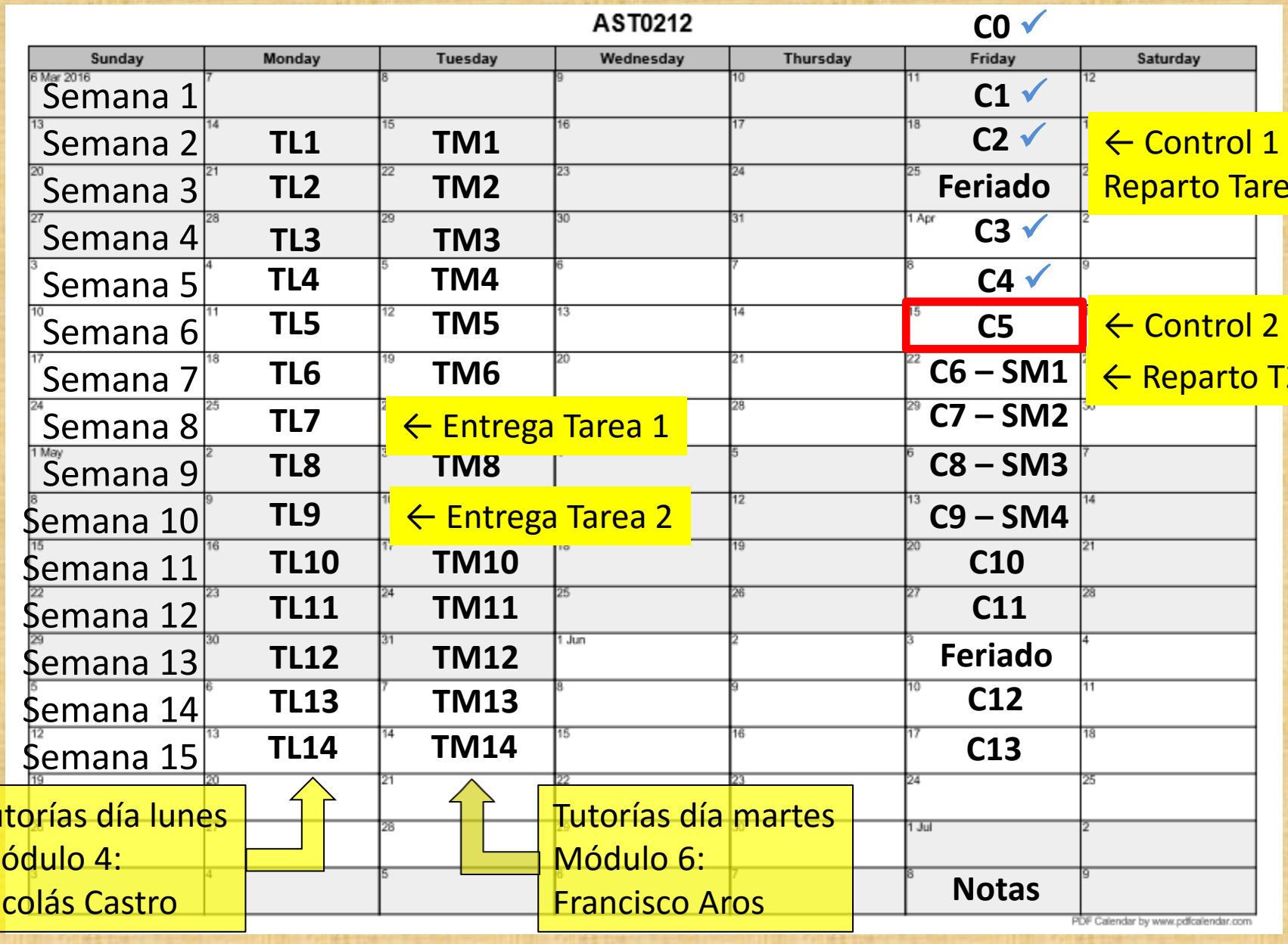
Francisco Aros (TM6)

Nicolás Castro (TL4)

TM6: Tutoría del martes en módulo 6

TL4: Tutoría del lunes en módulo 4

Nuestro Semestre 2016-1



AST0212

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 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 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

PDF Calendar by www.pdfcalendar.com

← 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 4):

REPASO

1. Temas pendientes
    1. Observaciones desde Santa Martina
1. Herramienta Linux de selección de datos en archivos organizados en columnas: *awk*
2. Breve repaso de la clase previa
    1. Visualización cualitativa de histogramas.
    2. Histogramas y funciones de distribución de probabilidad.
    3. Uso de la FDP para calcular parámetros de la distribución.

Temas del día: Segunda vuelta sobre FDP constante. Otras FDP que hay que conocer: Poisson y Gauss. Modelos de la realidad, distribución subyacente. Test modelo vs. realidad.

Esta clase (Clase 5):

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 e histogramas de histogramas

REPASO

10^6 números. FDP cte. entre 0 y 1.

$\bar{x}$ =0.5001, σ_x =0.2886, bin=0.0001

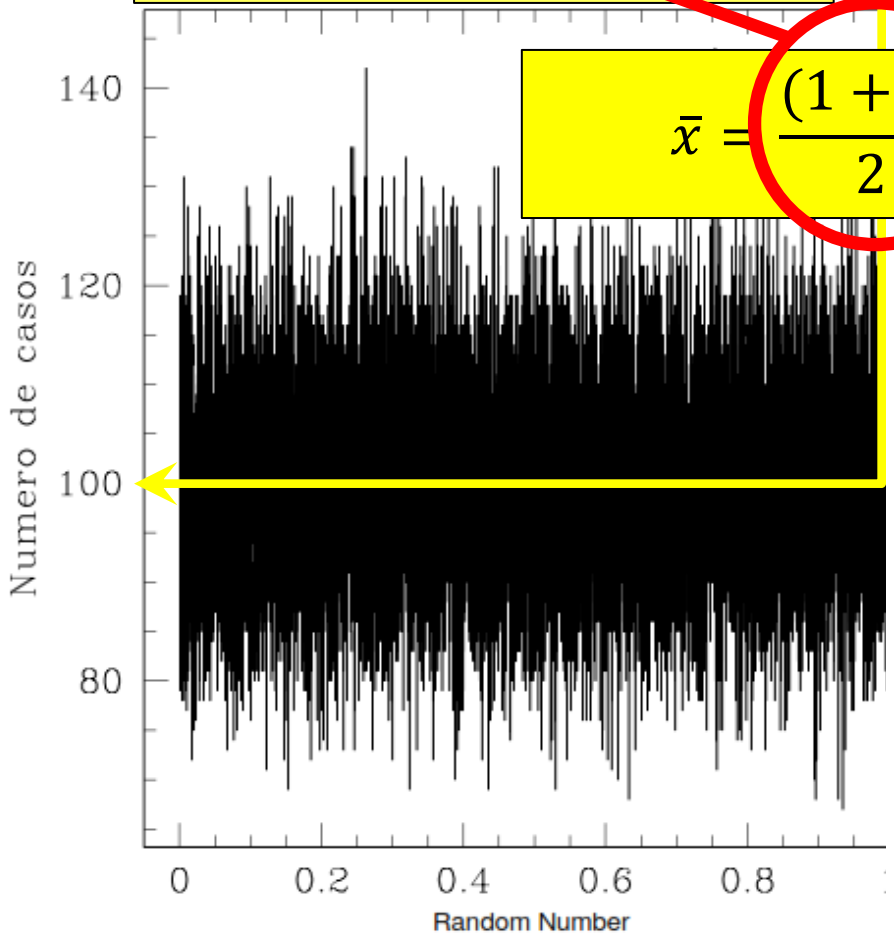
10^6 números. FDP cte. entre 0 y 1.

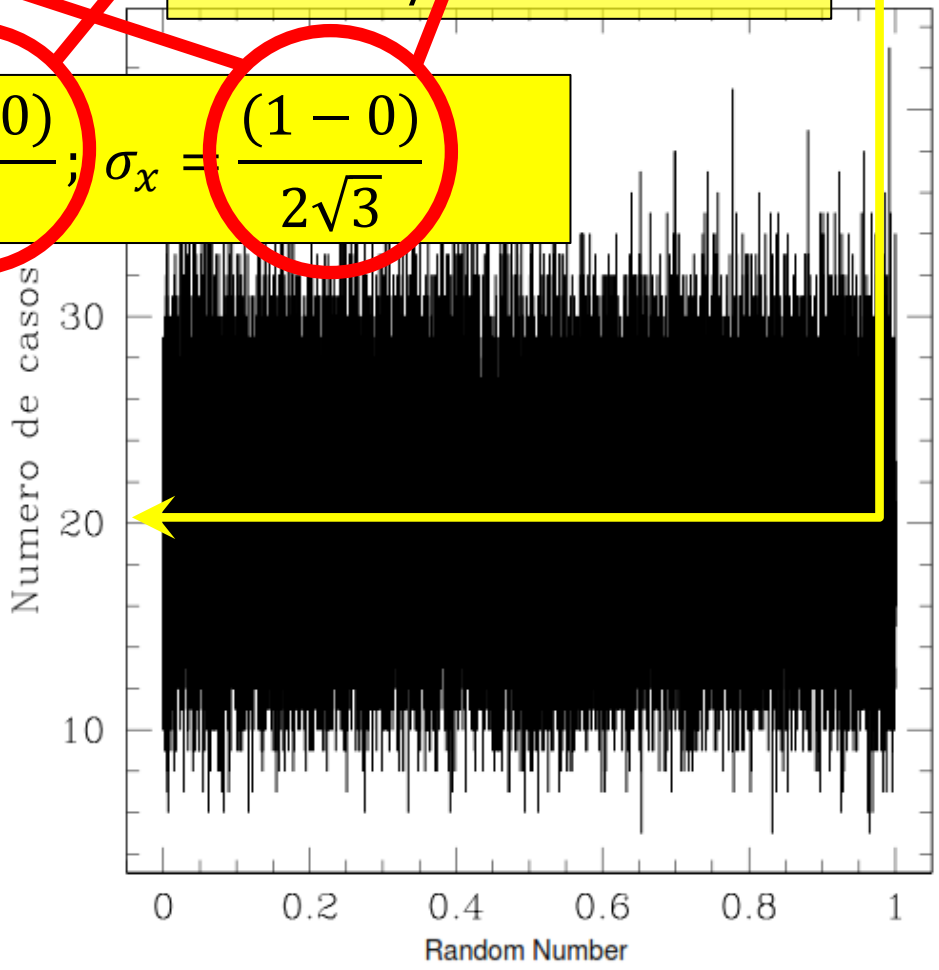
$\bar{x}$ =0.5001, σ_x =0.2886, bin=0.00002

$1 \times 10^6 / 1 \times 10^4 = 100$

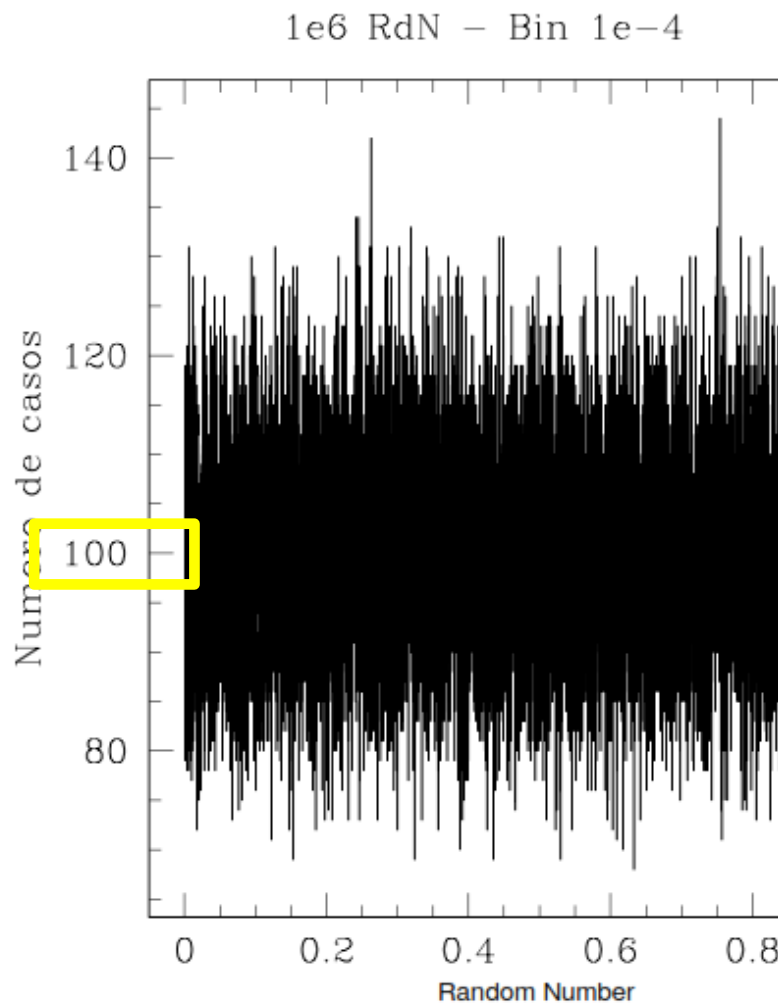
$1 \times 10^6 / 5 \times 10^4 = 20$

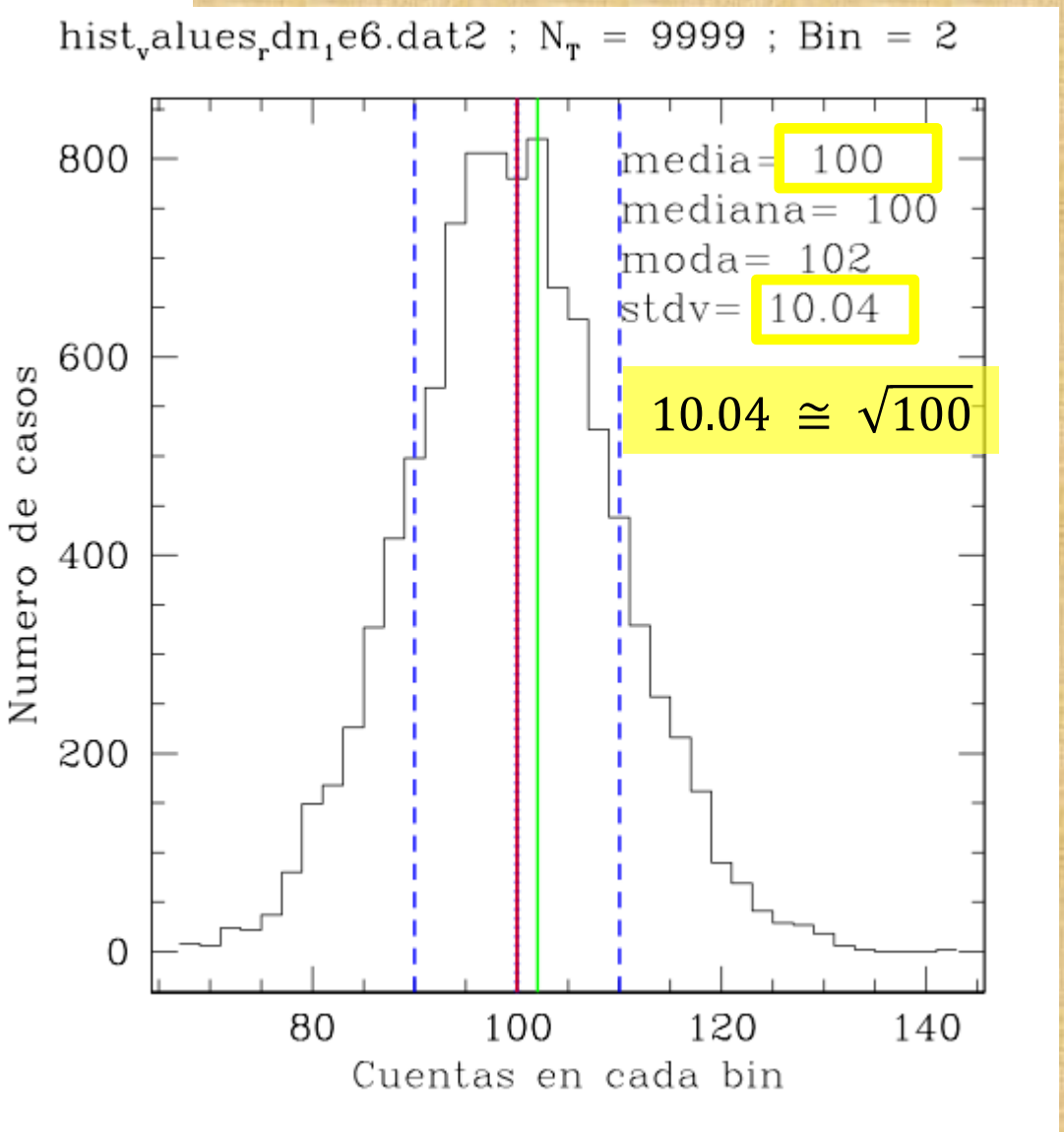
$$\bar{x} = \frac{(1+0)}{2}; \quad \sigma_x = \frac{(1-0)}{2\sqrt{3}}$$





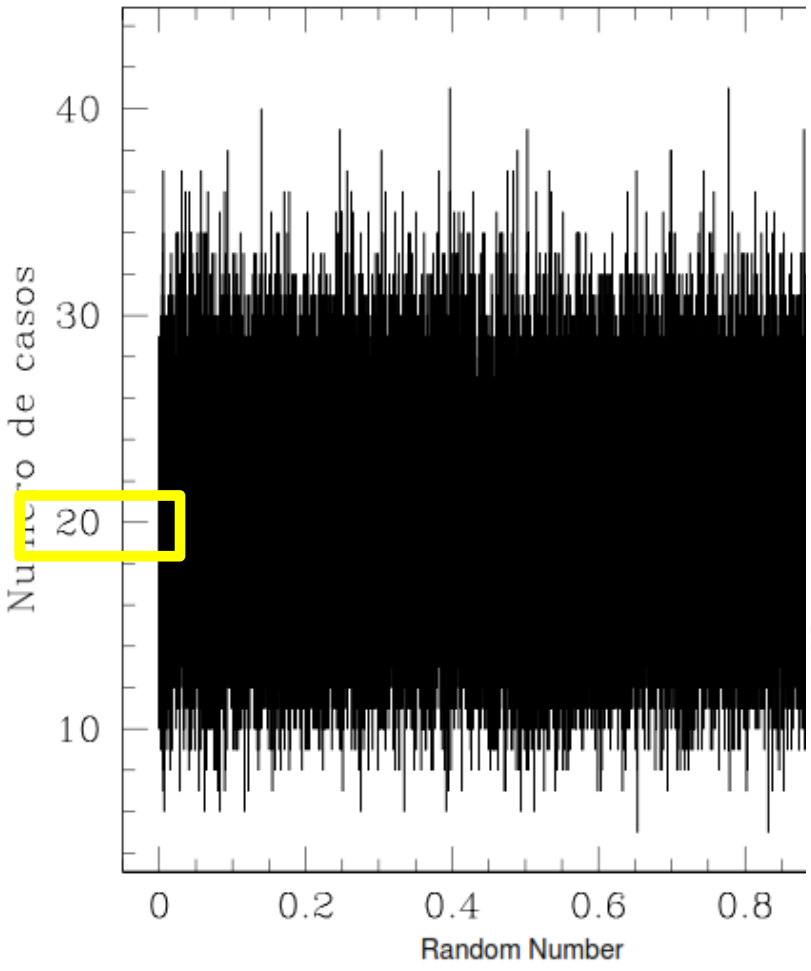
FDP e histogramas de histogramas REPASO

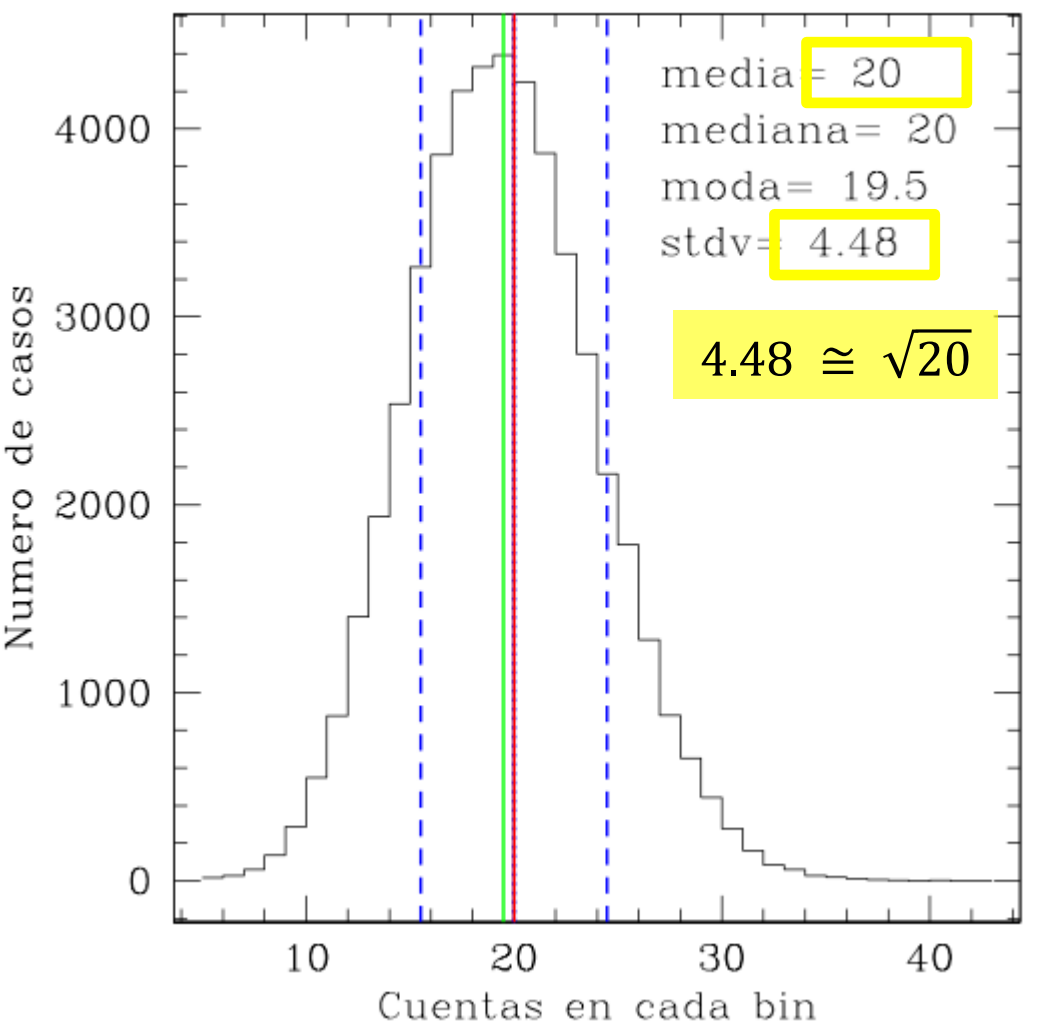




FDP e histogramas de histogramas REPASO

Si clasifico a los números aleatorios en bins más chicos, la FDP que obtengo será la misma, con parámetros diferentes:





FDP de Poisson

REPASO

La FDP que está detrás de todo esto es la llamada *Distribución de Poisson,* que resulta de contar eventos que suceden en un intervalo (de tiempo o espacio) dado, definido, cuando la probabilidad individual de cada evento es muy baja. Por ejemplo:

1. Decaimiento radioactivo de núcleos atómicos por segundo.
2. Explosiones de SN en un volumen del universo en un intervalo de tiempo.
3. Cantidad de gotas de lluvia que caen en un vaso en un intervalo de tiempo.
4. Número de fotones que llegan a un pixel de un CCD en una exposición.
5. Cantidad de números aleatorios que caen en un bin específico.

La FDP de Poisson, está dada por:

$$P_\mu(\nu) = e^{-\mu}\frac{\mu^\nu}{\nu!} \quad ;\text{con } \mu > 0$$

que es, específicamente, la probabilidad de contar ν eventos en el intervalo dado (la ecuación anterior está normalizada).

Puede mostrarse que para esta FDP $\bar{\nu} = \mu$ y ${\sigma_\nu}^2 = \mu$, o sea $\sigma_\nu = \sqrt{\mu}$.

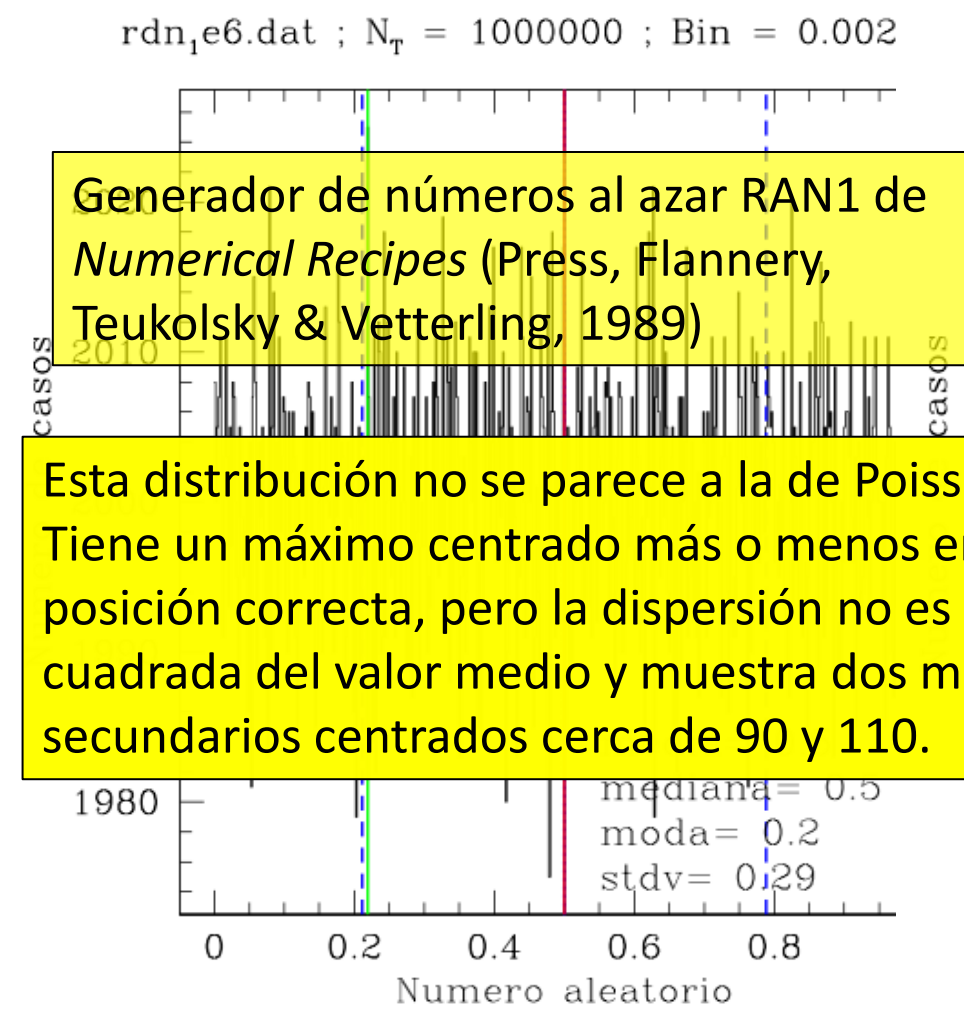
Entonces, si la tasa de ocurrencia es R (probabilidad del evento por unidad de intervalo), entonces $\mu = RT$, donde T es el largo del intervalo. Estas ecuaciones aclaran todas las coincidencias anteriores.

FDP de Poisson

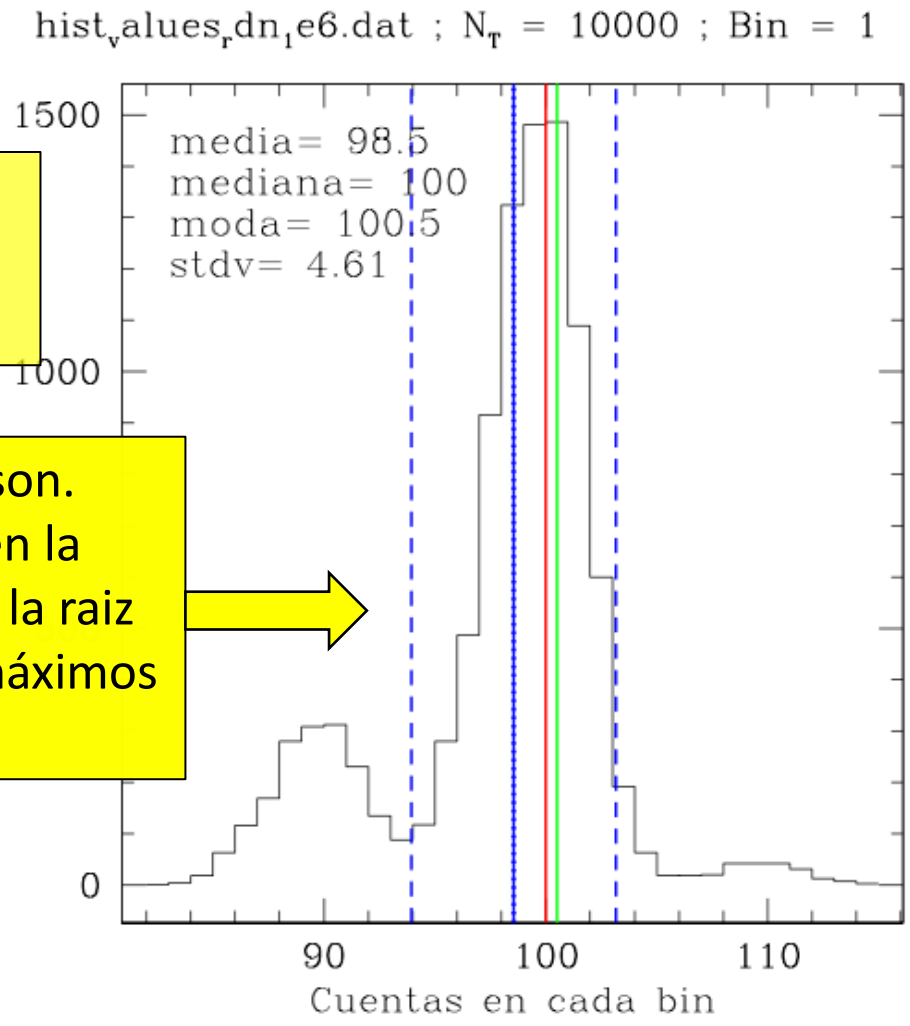
REPASO

Esta clase de análisis provee una herramienta muy buena para testear el software que estamos usando y asegurarnos que hace lo que dice que hace:

¡No todos los generadores de números al azar que andan por ahí son buenos!



Generador de números al azar RAN1 de *Numerical Recipes* (Press, Flannery, Teukolsky & Vetterling, 1989)

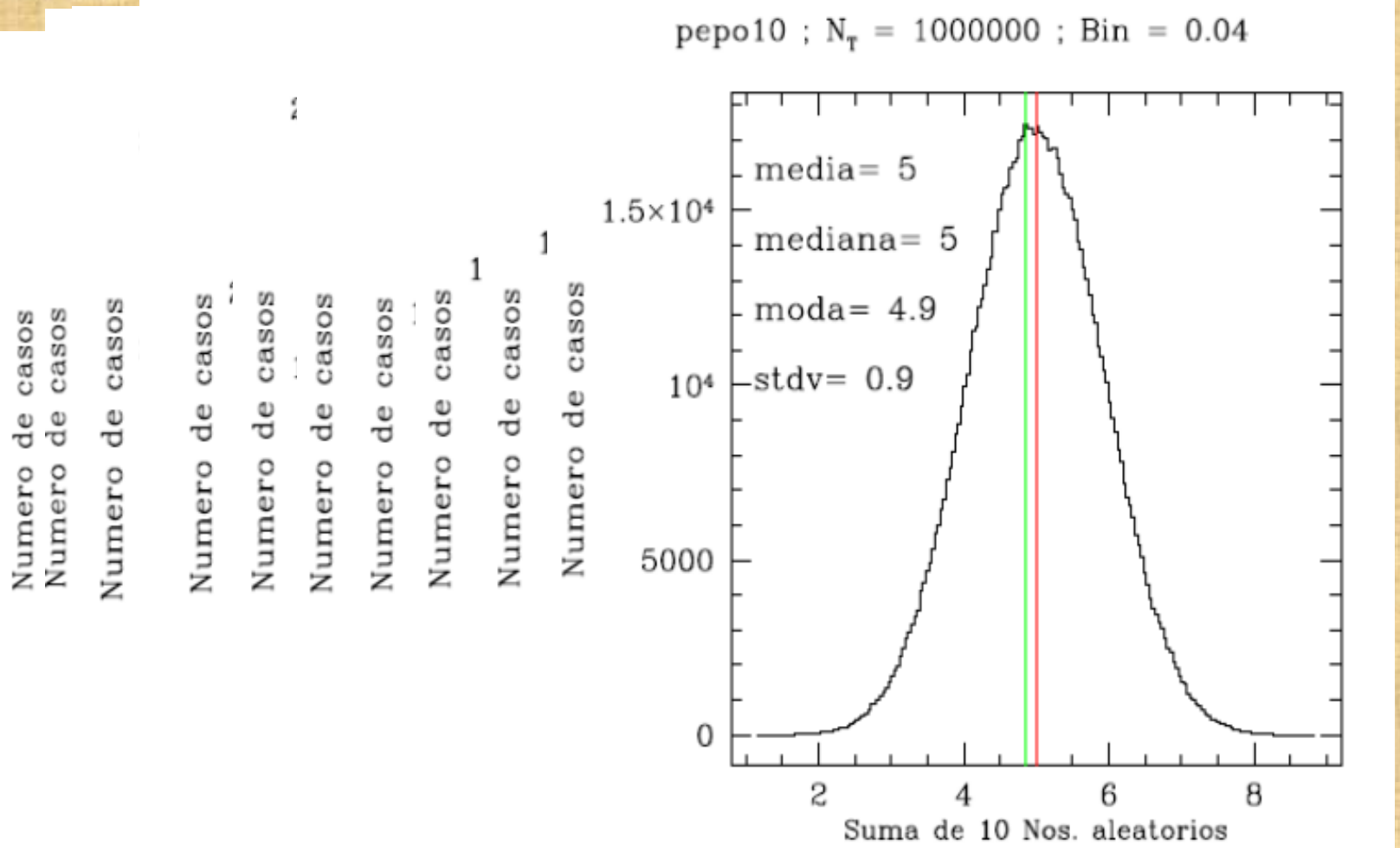


Esta distribución no se parece a la de Poisson. Tiene un máximo centrado más o menos en la posición correcta, pero la dispersión no es la raiz cuadrada del valor medio y muestra dos máximos secundarios centrados cerca de 90 y 110.

FDP de Gauss

REPASO

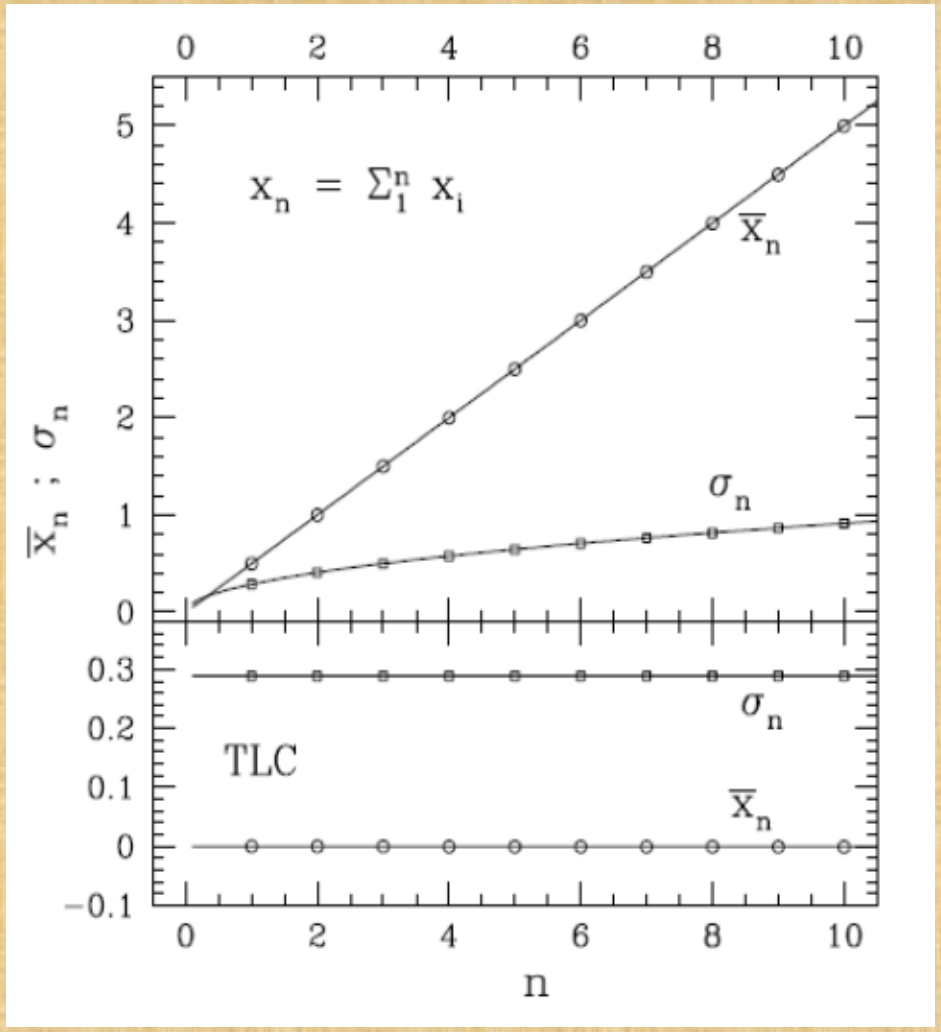
Una forma de presentar la FDP de Gauss podría ser “la distribución que se obtiene a partir de la de Poisson, en el límite $\mu \gg 1$”. Otra es jugar con las distribuciones de números con FDP constante: ¿Qué sucede si los sumamos? ¿Cómo es la FDP de $x_N = \sum_{i=1}^{N} x_i$, si cada uno de los x_i tiene FDP cte distribuida en (0,1)?



FDP de Gauss

REPASO

Parece haber algún secreto escondido ¿no?



Teorema del Límite Central

Si $x_1, x_2, \ldots, x_N$ son variables aleatorias independientes, y cada una de ellas tiene una FDP arbitraria $P_i(x_i)$, con valor medio μ_i y dispersión $\sigma^2{}_i$ entonces

$x_N = \frac{\sum_{i=1}^N x_i - \sum_{i=1}^N \mu_i}{\sqrt{\sum_{i=1}^N \sigma^2{}_i}}$, se aproxima a una distribución normal para $N \to \infty$.

$$\lim_{N\to\infty} P(x_N) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

El caso que mostré es un caso particular de esto, ya que las P_i son siempre la misma PDF, y por lo tanto $\mu_i = \mu = 0.5$ y $\sigma_i = \sigma = 1/\sqrt{12}$.

FDP de Gauss

REPASO

Teorema del Límite Central

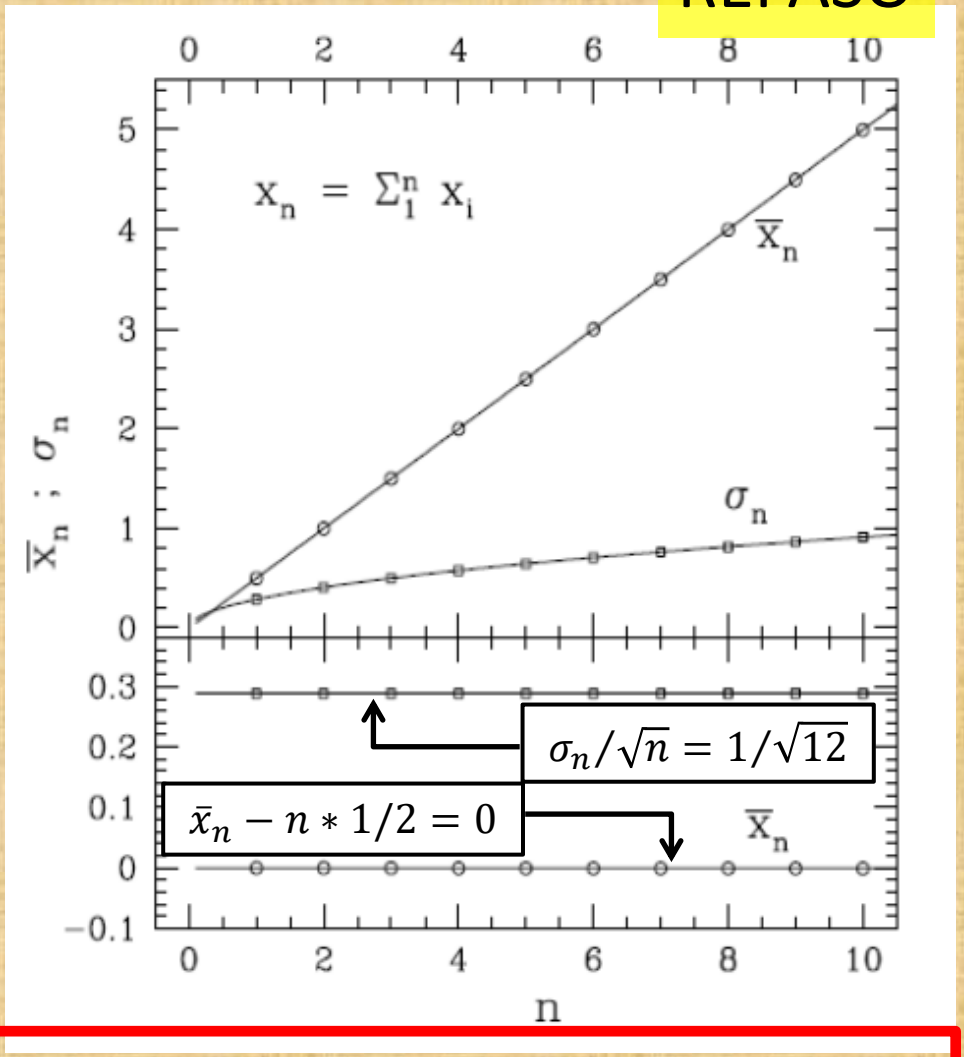
Entonces, para el caso específico de nuestra variable x_N,

$$\sum_{i=1}^{N} \mu_i = N * \mu = N * \frac{1}{2}$$

$$\sqrt{\sum_{i=1}^{N} \sigma^2{}_i} = \sigma\sqrt{N} = \sqrt{\frac{N}{12}}$$

entonces

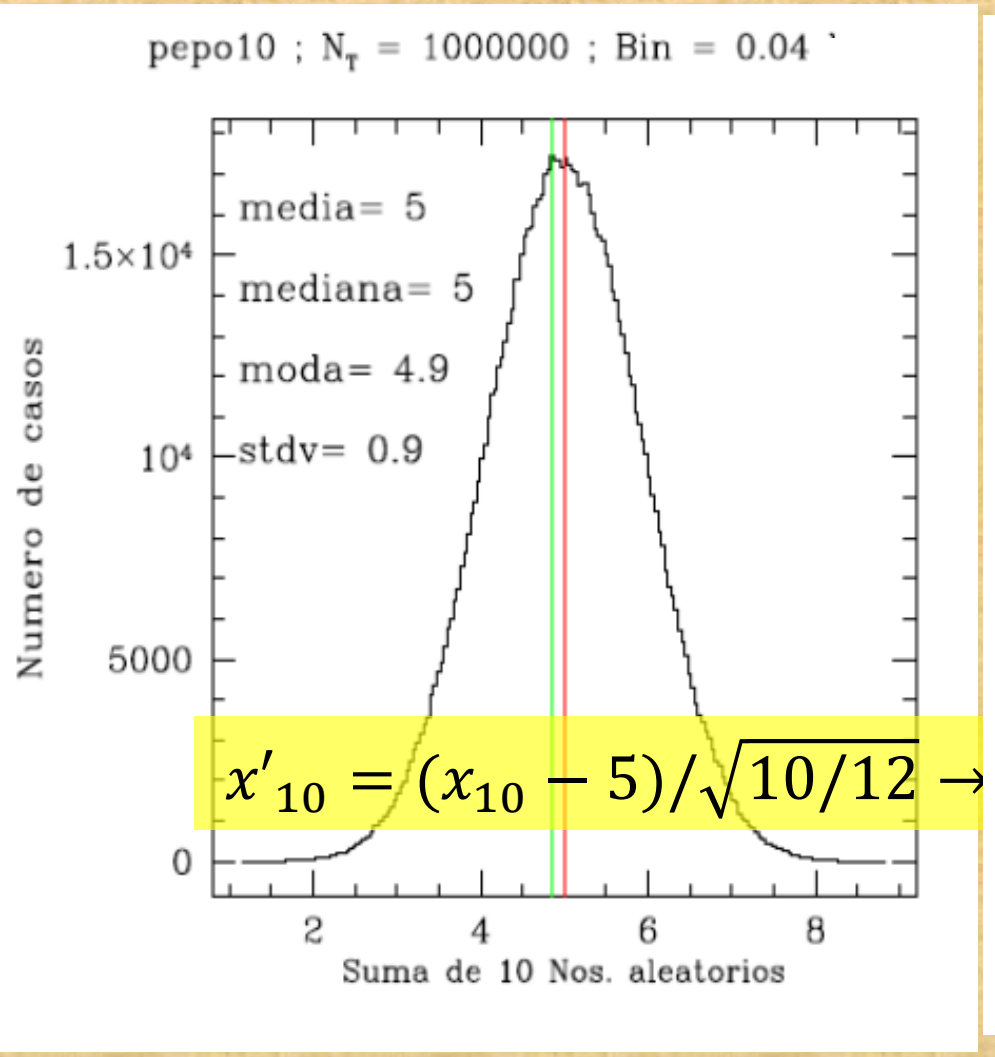
$x_N = \frac{\sum_{i=1}^{N} x_i - N/2}{\sqrt{N/12}}$, se aproxima a una distribución normal para $N \to \infty$.



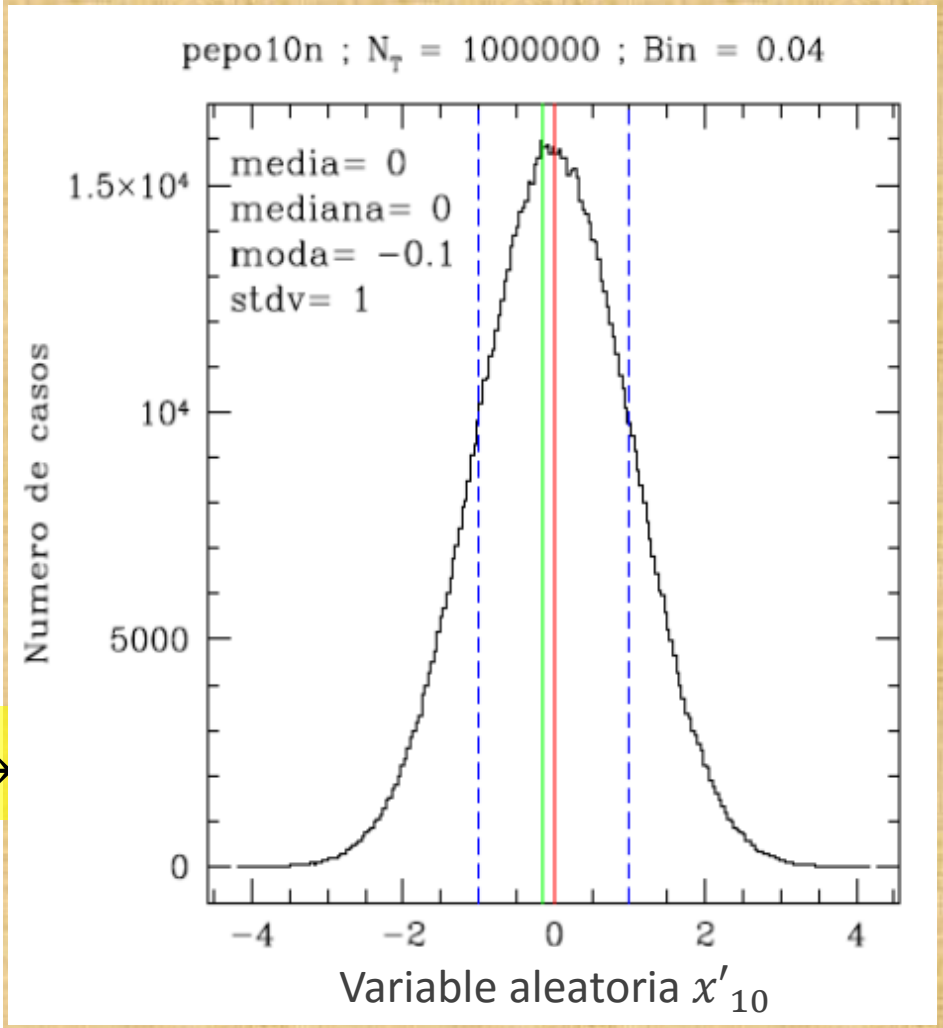
$$\lim_{N\to\infty} P(x_N) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\sum_{i=1}^{N} x_i \to \left(\frac{1}{\sqrt{N/12}\sqrt{2\pi}} e^{-\frac{(x-N/2)^2}{2}} \right)$$

FDP de Gauss



$$x'_{10} = (x_{10} - 5)/\sqrt{10/12} \rightarrow$$

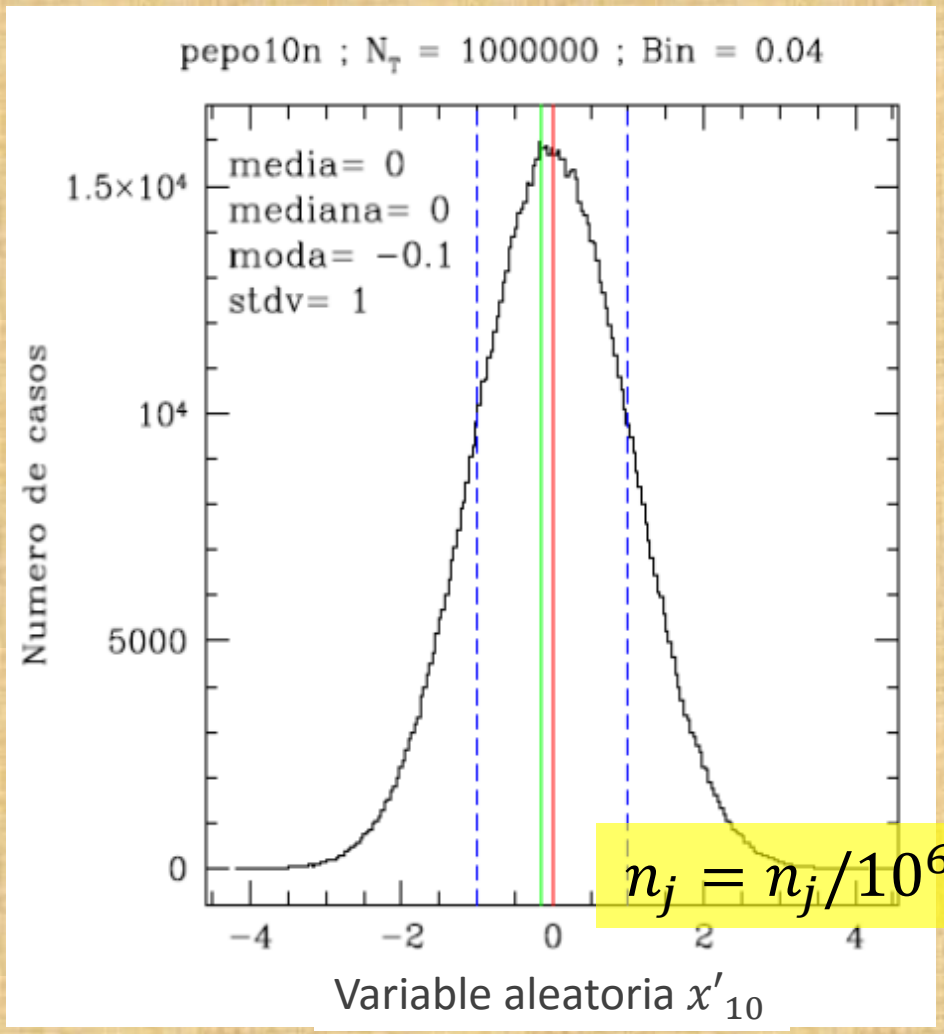


$$P(x_{10}) = N(5, \sqrt{10/12}) = \frac{1}{\sqrt{5\pi/3}} e^{-\frac{(x-5)^2}{2\sqrt{5/6}}}$$

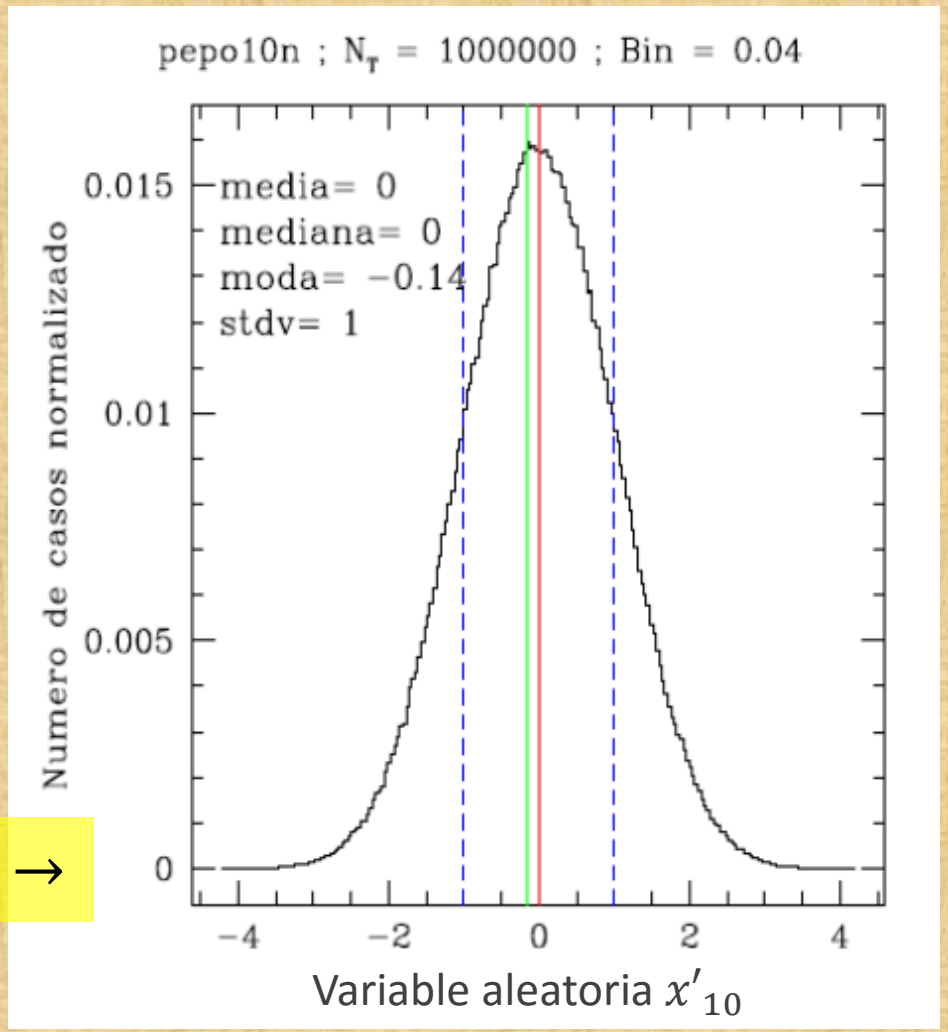
$$P(x_{10}') = 10^6 N(0,1) = \frac{10^6}{\sqrt{2\pi}} e^{-\frac{x'^2}{2}}$$

FDP de Gauss

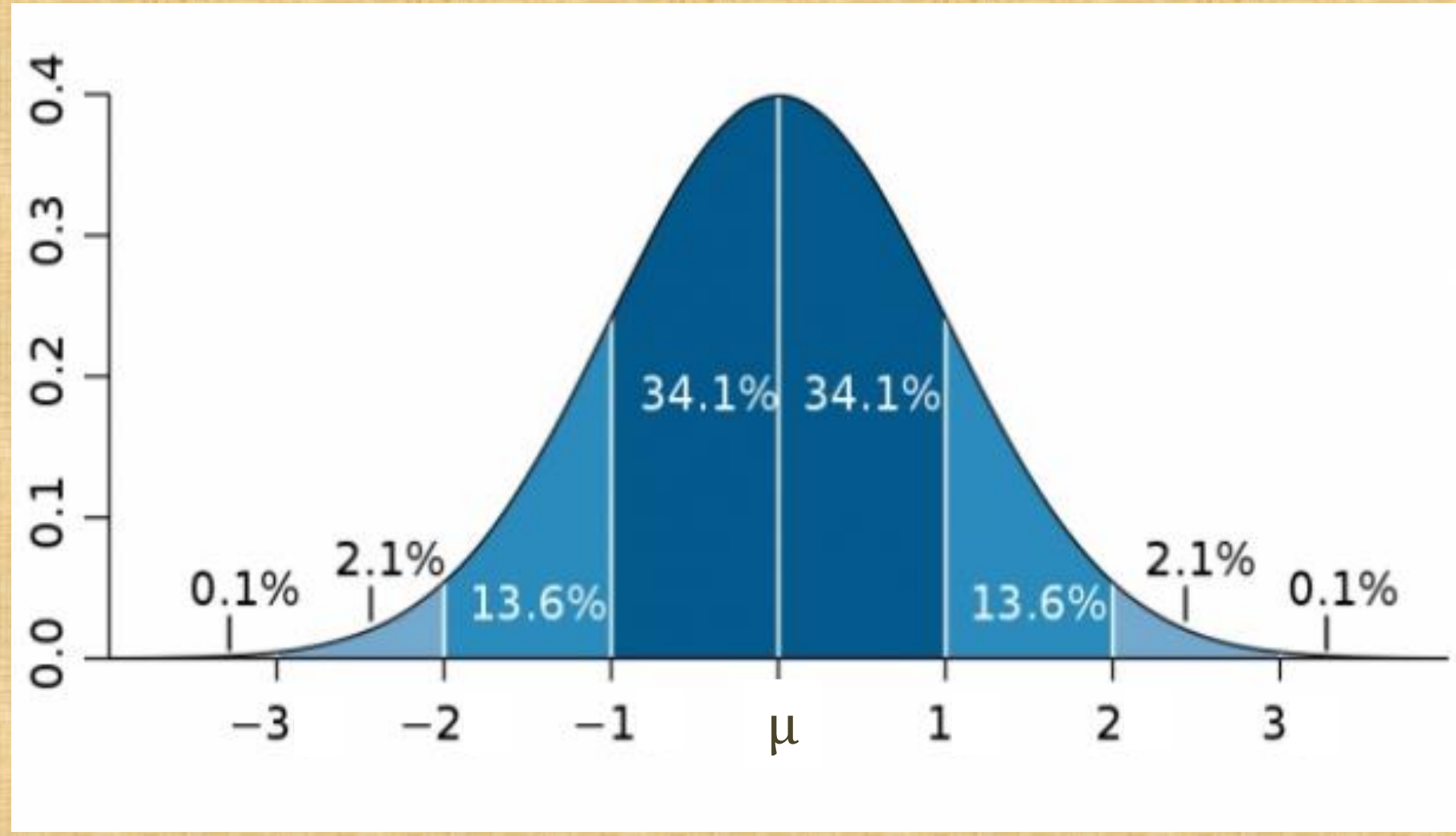
La transformación a un histograma de FDP Gaussiano de *N(0,1)* todavía no está completa porque el número de casos sigue reflejando la población de 10^6 números. Para corregir esto, dividimos las cuentas del histograma final por 10^6.



$n_j = n_j/10^6 \rightarrow$

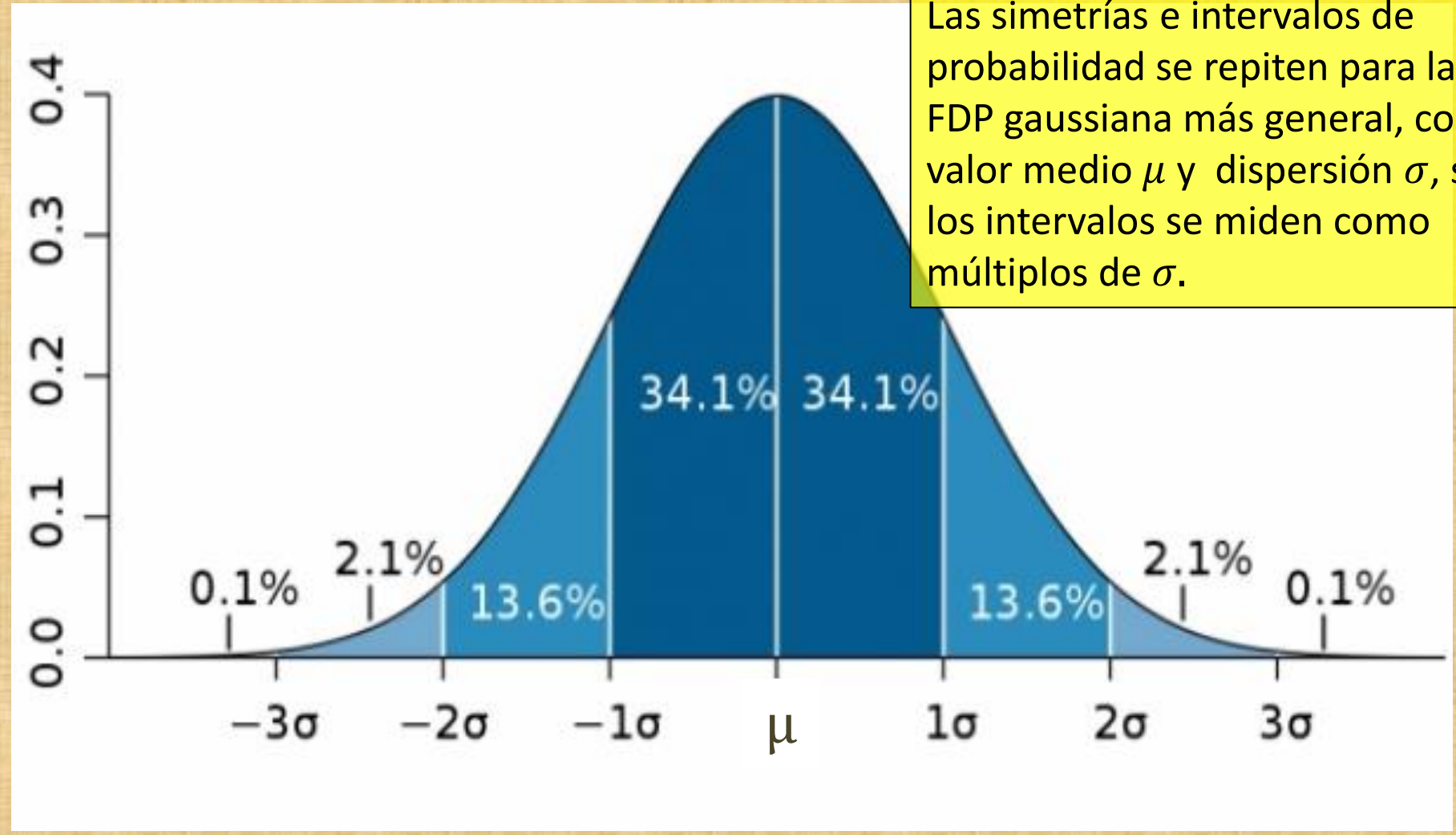


FDP de Gauss



$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

FDP de Gauss



Las simetrías e intervalos de probabilidad se repiten para la FDP gaussiana más general, con valor medio μ y dispersión σ, si los intervalos se miden como múltiplos de σ.

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

FDP de Gauss

La forma anterior es la más general de la distribución normal:

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

forma que también está normalizada de forma que su integral en el espacio completo de definición de la probabilidad, $(-\infty,\infty)$ es 1:

$$P_{(-\infty<x<\infty)} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

La FDP de Gauss puede usarse para predecir la probabilidad de que un valor de x esté en un cierto rango de la variable (x_1, x_2):

$$P_{(x_1<x<x_2)} = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{x_1}^{x_2} N(\mu,\sigma) dx$$

FDP de Gauss

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, $n_{e,j}$, 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

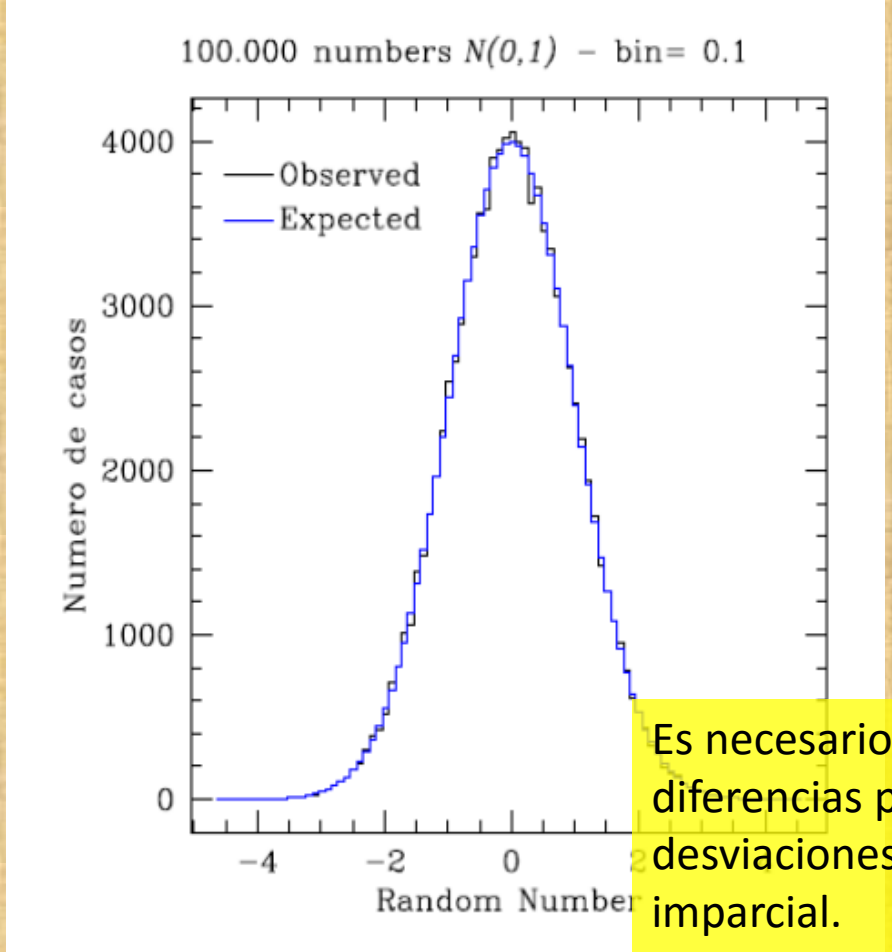
Habiendo hecho esto puedo desarrollar un test cuantitativo para medir cuán diferente es el histograma observado del que predice el modelo. El test que vamos a usar es el llamado χ^2, que se construye sumando los cuadrados de las diferencias entre las cuentas real, observadas, y las predichas por el modelo,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

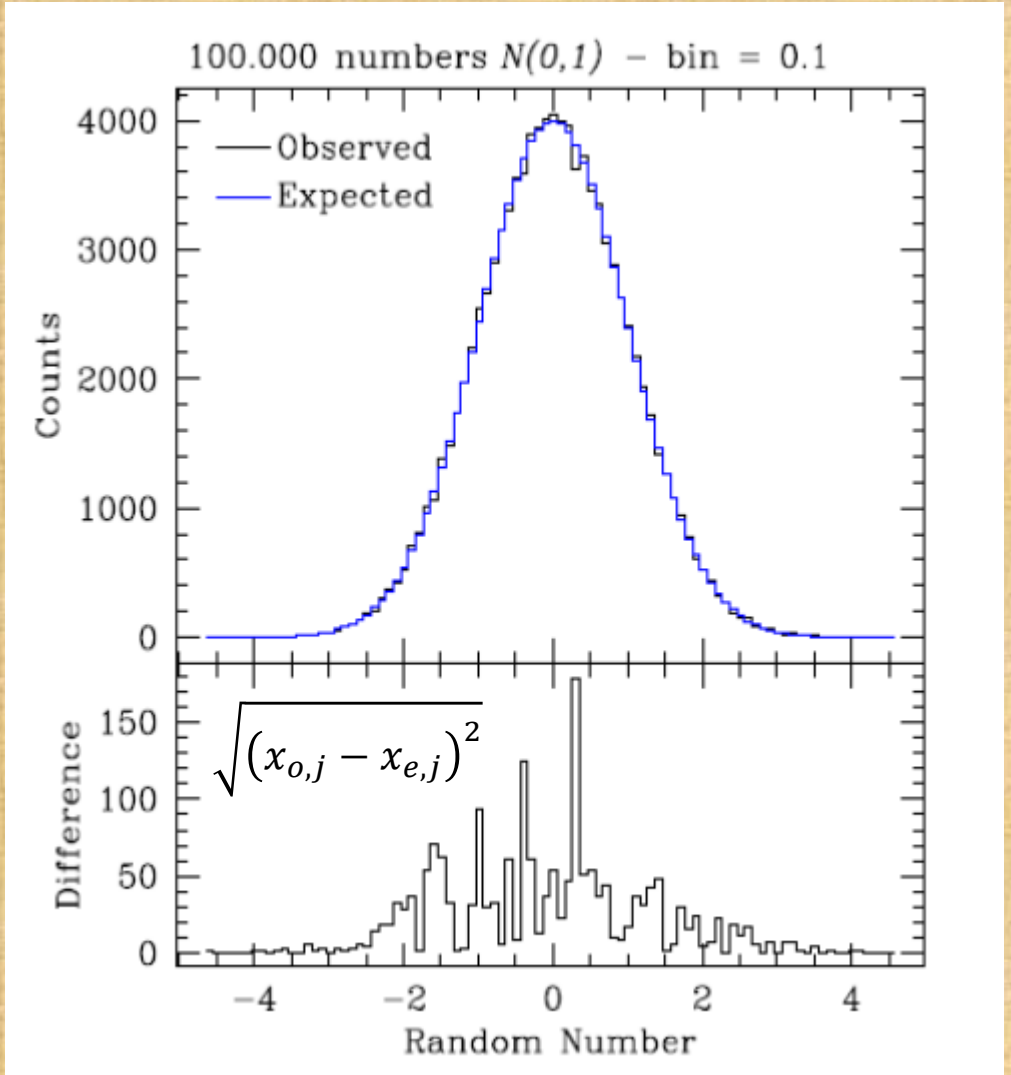
$$\chi^2 = \sum_{j=1}^{M} \frac{\left(n_{o,j} - n_{e,j}\right)^2}{n_{e,j}}$$

$n_{o,j}$ 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?

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

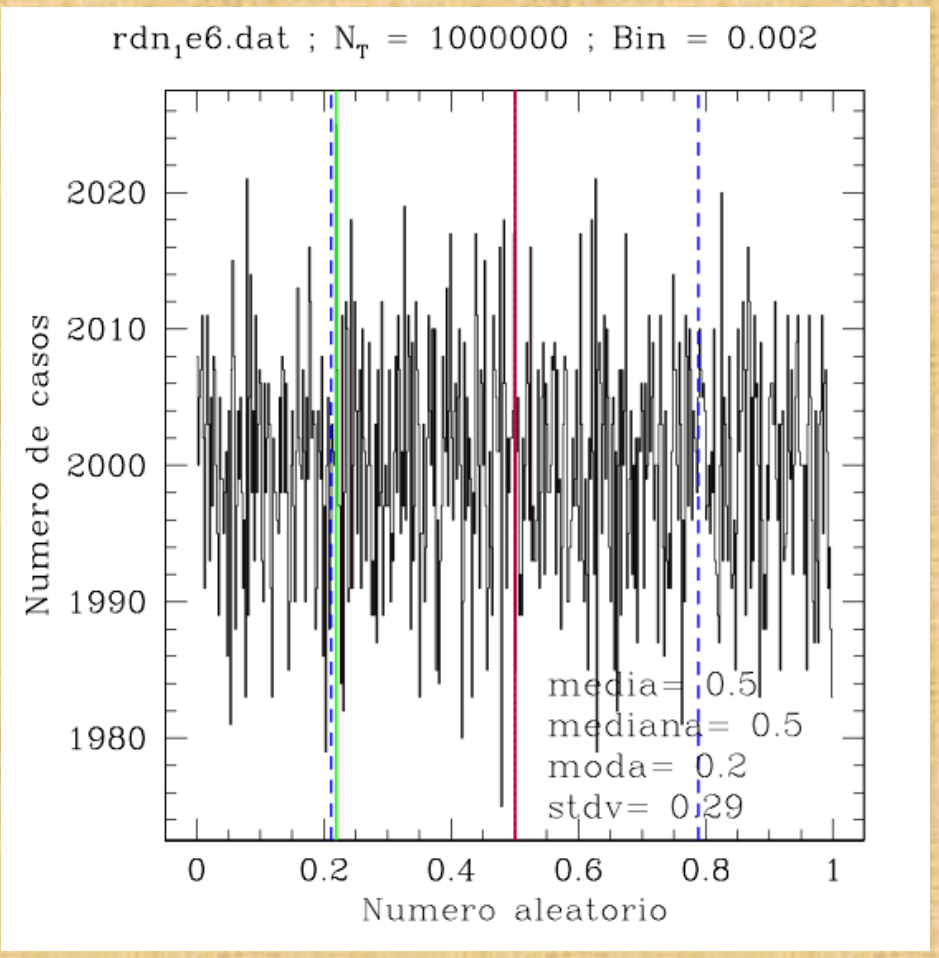




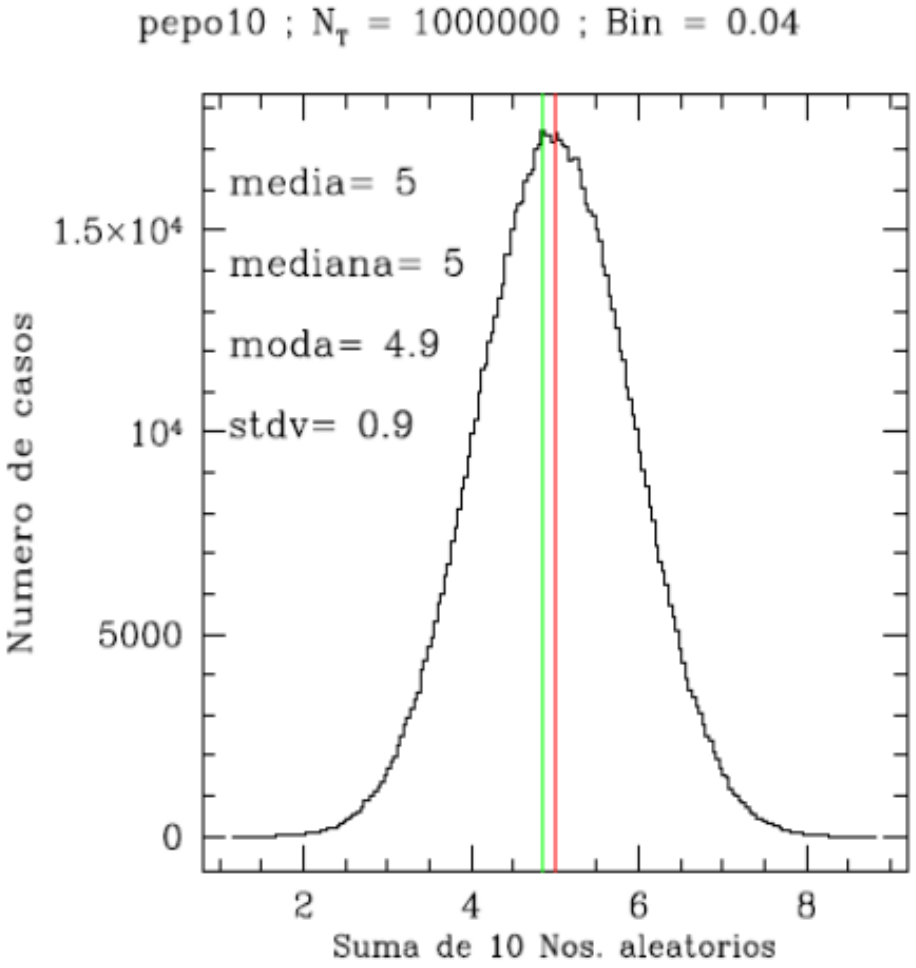
Es necesario normalizar las diferencias para medir las desviaciones de manera imparcial.

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

¿Cuál es un buen modelo para σ_e?



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>.*

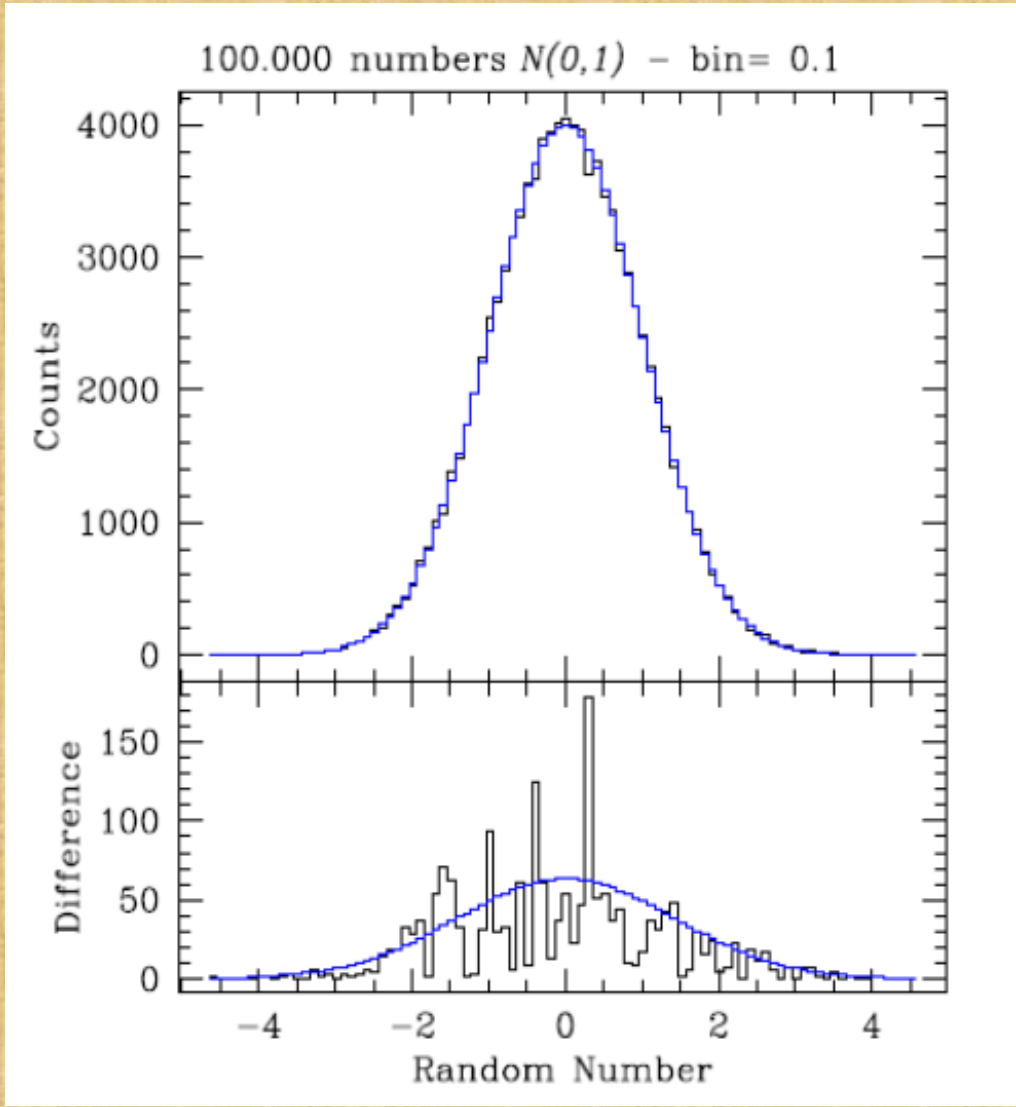
¿Qué es χ^2?

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:

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

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

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

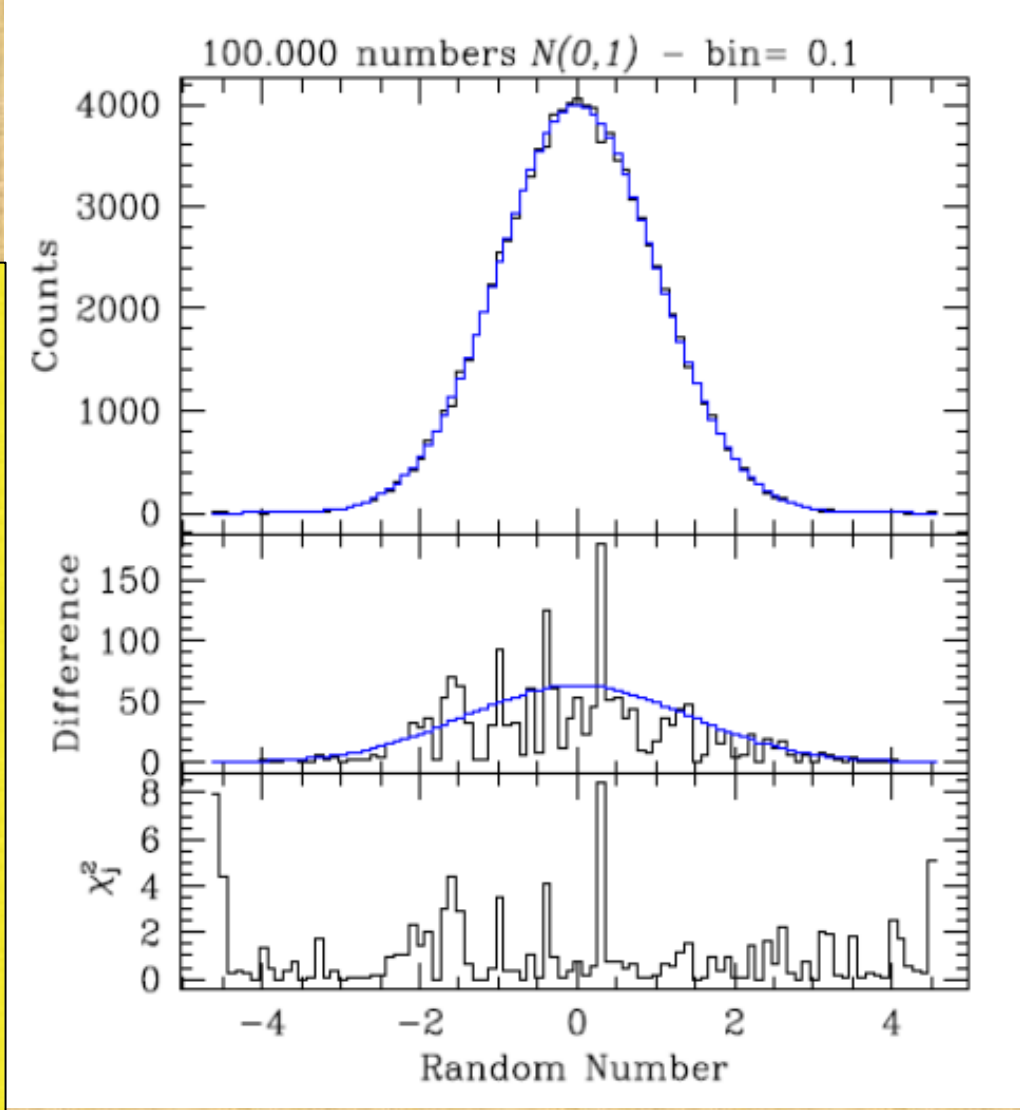


Esta última es la expresión que había puesto como “receta” al final de la clase pasada.

¿Qué podemos esperar de χ^2?

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

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").



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

Fin de ppt de Clase 5