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

AST0212 – 2016-1

Introducción al análisis de datos

Instituto de Astrofísica

Facultad de Física

Pontificia Universidad Católica de Chile

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 ¿Entrega Tarea 1?		22 **C6 – SM1**	23
24 Semana 8	25 **TL7**	26 **TM7**	27		29 **C7 – SM2**	30
1 May Semana 9	2 **TL8**	3 **TM8**	4	5	6 **C8 – SM3**	7
8 Semana 10	9 **TL9**	10 **TM9**	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

← ¿Control 2?

Tutorías día lunes
Módulo 4:
Nicolás Castro

Tutorías día martes
Módulo 6:
Francisco Aros

El control

AST 0212 – Introducción al análisis de datos – 2016-1

Control 1 – 18/M

Valor medio, dispe

Pregunta: Medidas repetidas y ci

Usted recibió un set de 9 medicione
de un péndulo.

1. Calcule el valor medio y la disp
significativas que tengan sentido.
la tabla provista para que los c
hubiera errores de arrastre). Por s

Set de datos número: 24	
i	
1	
2	
3	
4	
5	
6	
7	
8	
9	
Σ	

5/6

AST 0212 – Introducción al análisis de datos – 2016-1

Control 1 – 18/Marzo/2016 – 15 minutos

Valor medio, dispersión, cifras significativas

Pregunta: Medidas repetidas y cifras significativas

Usted recibió un set de 9 mediciones del intervalo de tiempo, en segundos, que dura la oscilación de un péndulo.

1. Calcule el valor medio y la dispersión del set de datos y repórtelo con el número de cifras significativas que tengan sentido. Justifique el número de cifras significativas usado. Utilice la tabla provista para que los correctores puedan chequear sus cálculos intermedios (por si hubiera errores de arrastre). Por seguridad, en las columnas 3 y 4 anote cinco dígitos decimales.

5/6

EL NUMERO DE CIFRAS SIGNIFICATIVAS QUE USO ES 2 DADO QUE EN ESTE USO LOS CEROS A LA DERECHA DE LOS DATOS SERIAN MUCHOS Y NO SERIAN DE MAYOR RELEVANCIA

Set de datos número: 9			
i	x_i	$(x_i - \bar{x}_{est})$	$(x_i - \bar{x}_{est})^2$
1	19.74	-0.29777	0.08866
2	19.89	-0.14777	0.02183
3	20.00	-0.03777	0.00142
4	20.28	0.24223	0.05867
5	20.18	0.14223	0.02022
6	20.09	0.05223	0.00272
7	20.03	-0.00777	0.00006
8	20.17	0.13223	0.01748
9	19.96	-0.07777	0.00604
Σ	$20.03 = \bar{x}_{est}$	N/A	$0.16 = \sigma_{est}$

Clase previa (Clase 2):

1. Temas pendientes de la Clase 1
    1. Datos para Tarea 1
        1. ¿Status de toma de datos?
        2. “Fake” data y ejemplo de uso de herramientas Linux

¿Datos listos?
¿Tablas listas?

Sistemas Linux: Seguir practicando. ¡Es el futuro!

2. Vueltas de tuerca sobre la Tarea 1
    1. ¿Cómo visualizar fácilmente cientos de datos?
    2. ¿Cuál es la mejor balanza?

Temas del día: Histogramas, errores aleatorios y sistemáticos.

Esta clase (Clase 3):

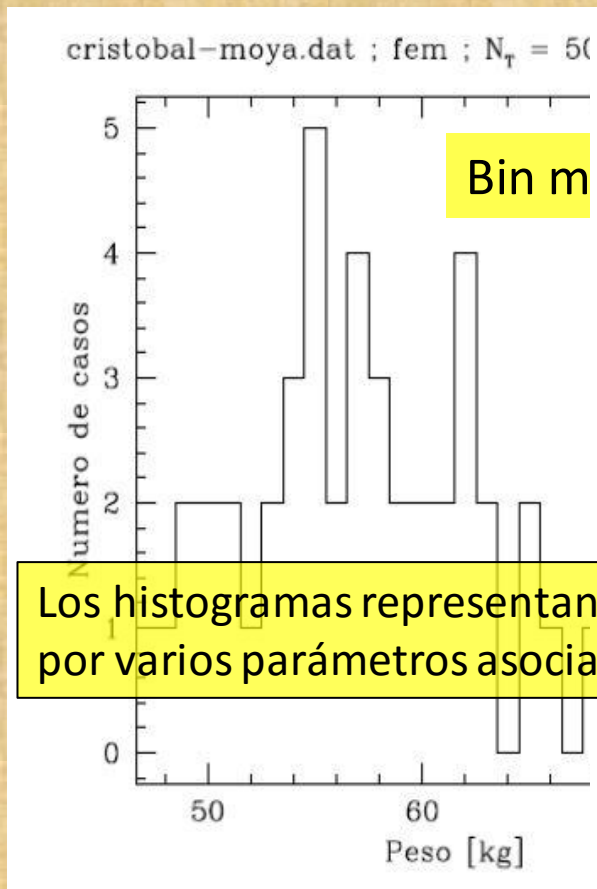
1. Temas pendientes
    1. Datos para Tarea 1 ¡Grupos 1, 6 y 8 no enviaron sus datos!
    2. Una vuelta de análisis sobre el Control 1
2. Vueltas de tuerca sobre la Tarea 1
    1. Herramientas Linux de selección de datos en archivos de texto simple organizados en columnas: *awk*

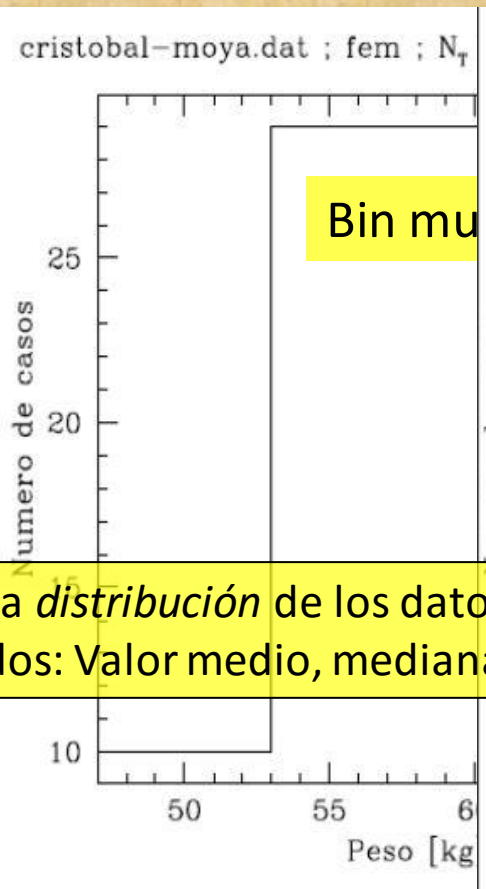
Temas del día: Visualización cualitativa de histogramas. Histogramas y funciones de distribución de probabilidad. Uso de la FDP para calcular parámetros de la distribución.

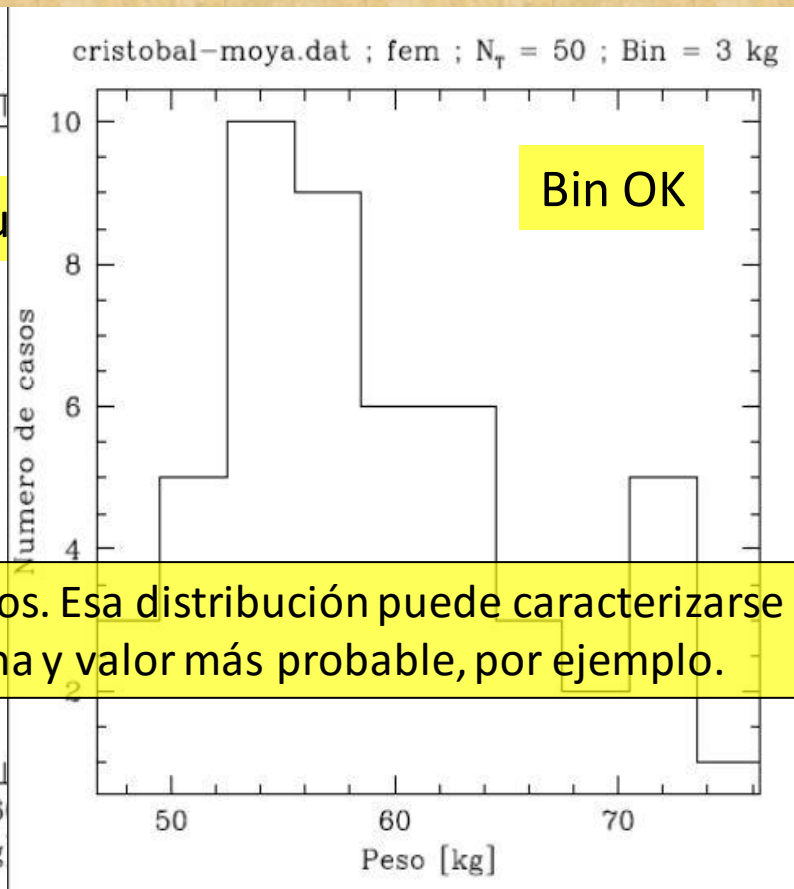
Histogramas

REPASO

El objetivo de los histogramas es proporcionar una visión rápida y compacta de una gran cantidad de datos *directamente comparables* y ver como se organizan de acuerdo a su valor. Hay algo de arte en esto de construir un histograma:



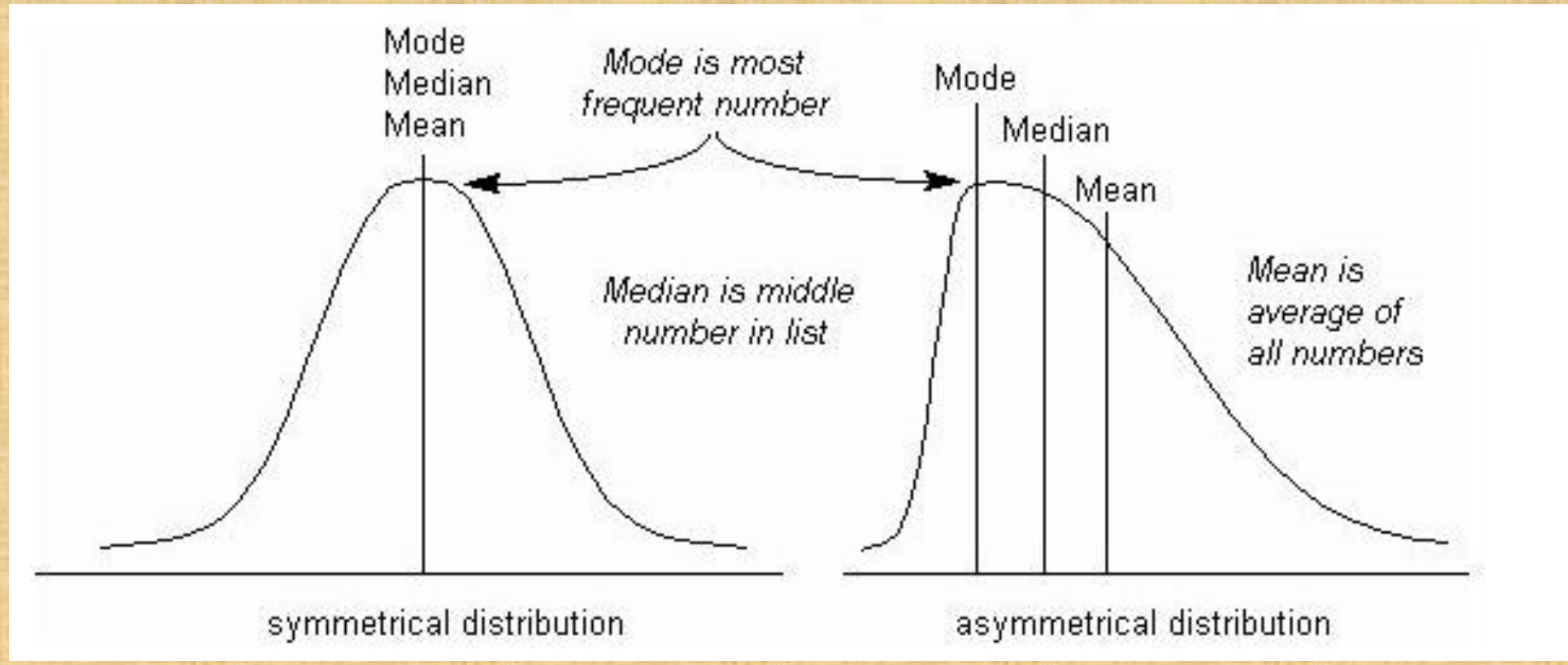




Los histogramas representan la *distribución* de los datos. Esa distribución puede caracterizarse por varios parámetros asociados: Valor medio, mediana y valor más probable, por ejemplo.

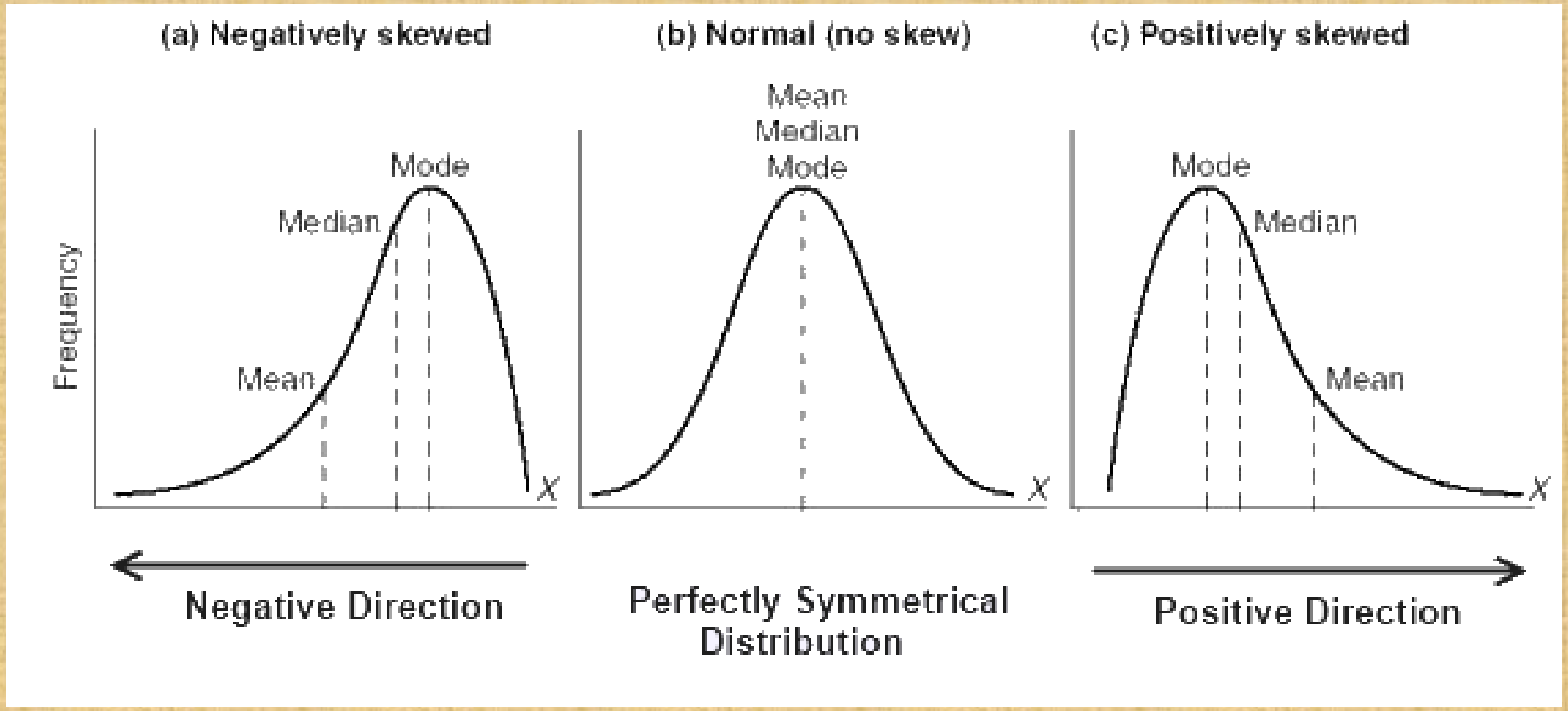
Histogramas: Media, mediana y moda

REPASO



Histogramas: Media, mediana y moda

REPASO



Las distribuciones, y sus histogramas, pueden ser simétricas o estar sesgados hacia un lado u otro. Ésto se define arbitrariamente como sesgo positivo (a la derecha) o negativo (a la izquierda).

REPASO

Histogramas

Calcular la mediana y la moda requiere contar valores.

Calcular la media requiere aplicar fórmulas:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

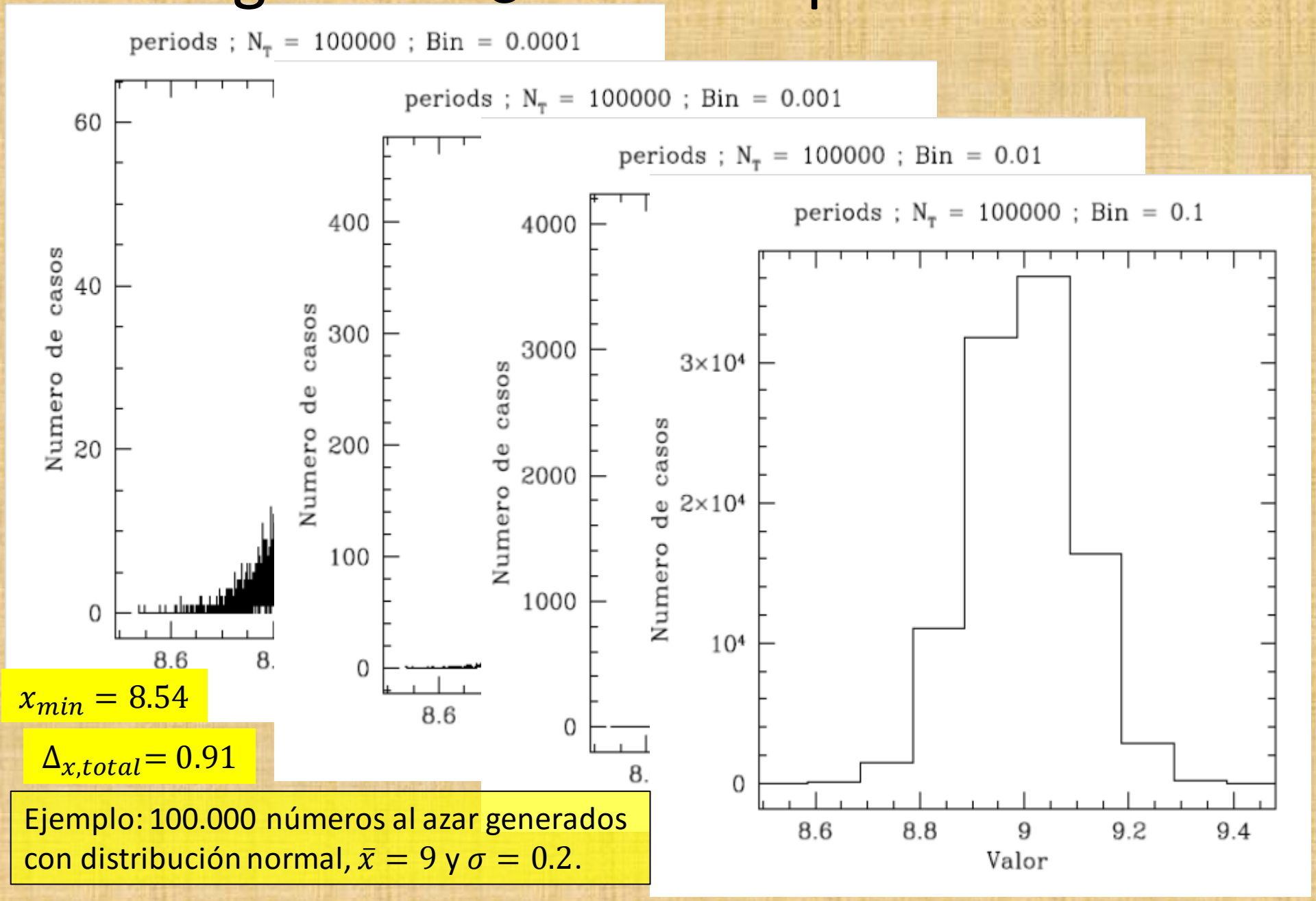
$$\overline{x_g} = \frac{1}{N}\sum_{j=1}^{M} n_j \overline{x_j}$$

"*i*" recorre todos los datos, del primero al último (1 a *N*).

"*j*" recorre el número de bins, del primero al último (1 a *M*). $\overline{x_j}$ es el valor medio del j-ésimo bin.

Pregunta para pensar en casa: ¿Son consistentes estas definiciones? ¿Dan el mismo valor medio?

Histogramas: ¿Tamaño óptimo del bin?



$x_{min} = 8.54$

$\Delta_{x,total} = 0.91$

Ejemplo: 100.000 números al azar generados con distribución normal, $\bar{x} = 9$ y $\sigma = 0.2$.

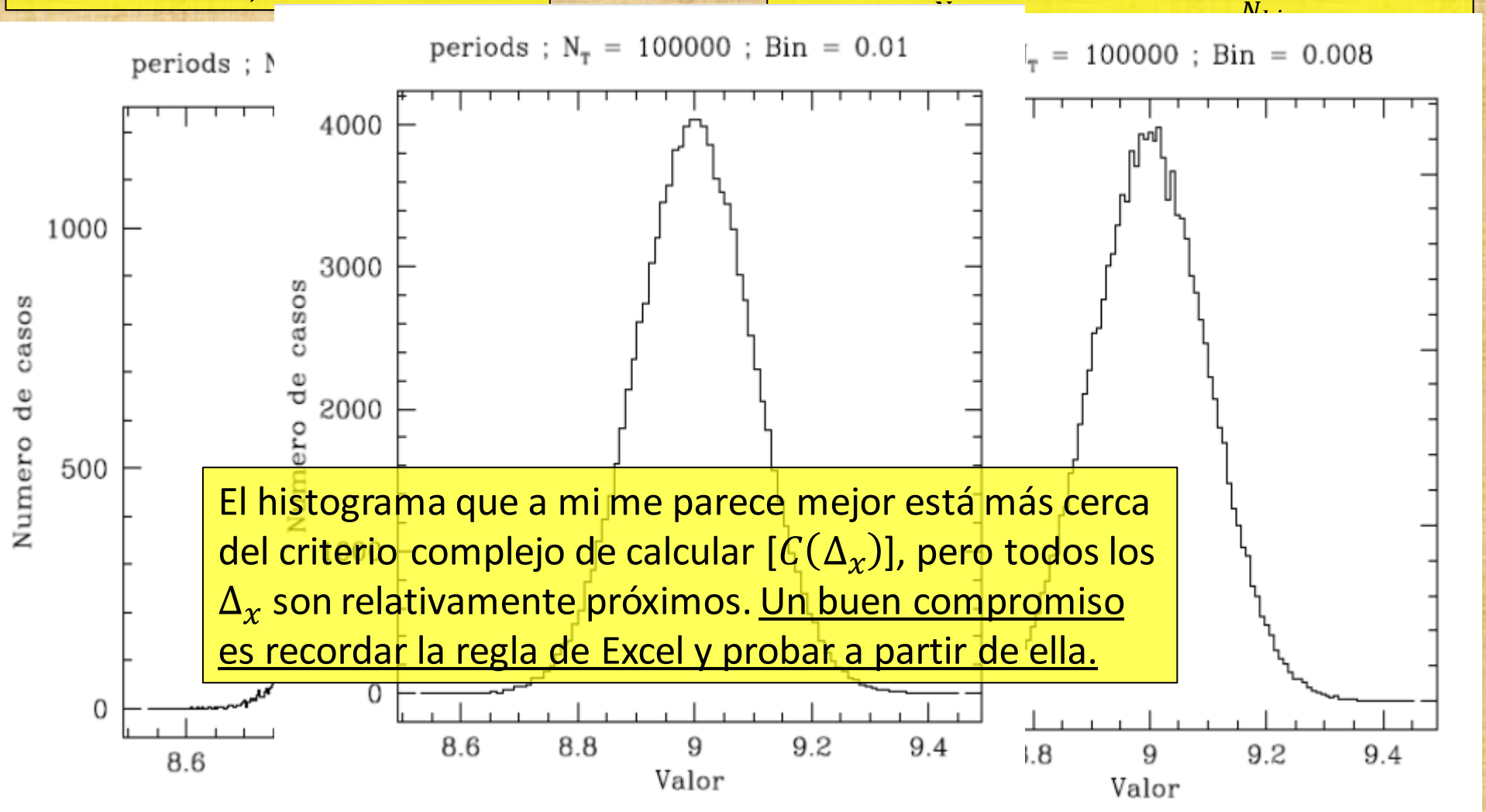
Histogramas: ¿Tamaño óptimo del bin?

Regla de "Excel": $N_{bin} \cong \sqrt{N_{total}}$

$$\Longrightarrow \Delta_x = \frac{0.91}{316{,}228} = 0{,}00288$$

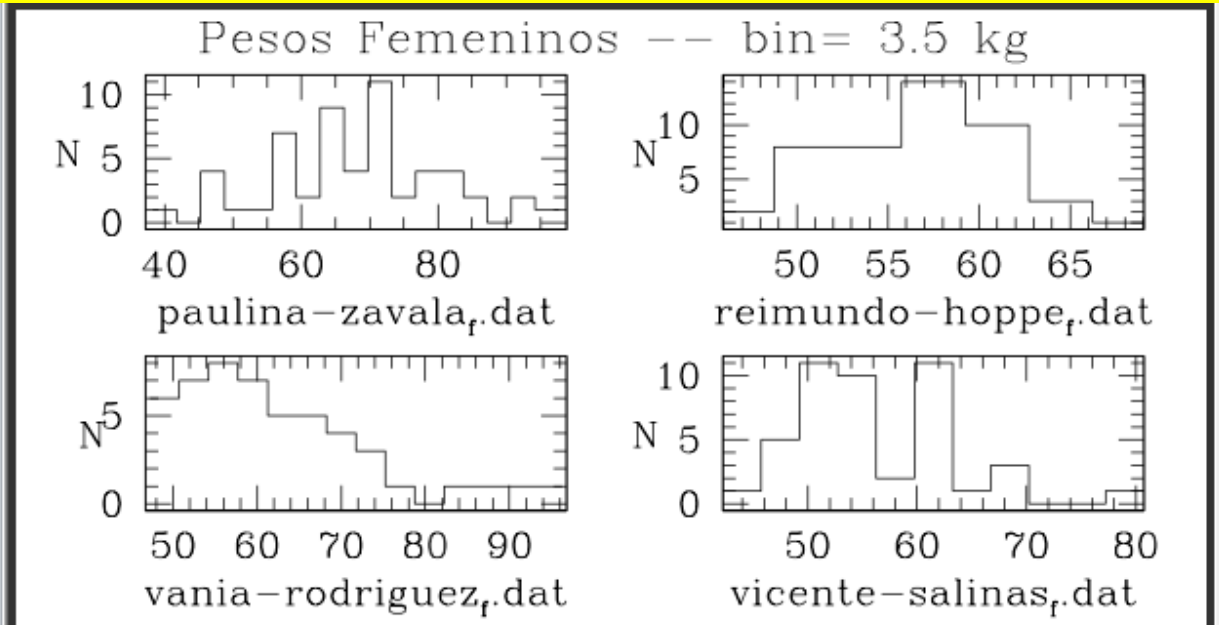
Regla de Shimazaki & Shinomoto (2007): El Δ_x que minimiza $C(\Delta_x)$

$$C(\Delta_x) = \frac{\left(2\bar{h} - v_h\right)}{{\Delta_x}^2}$$

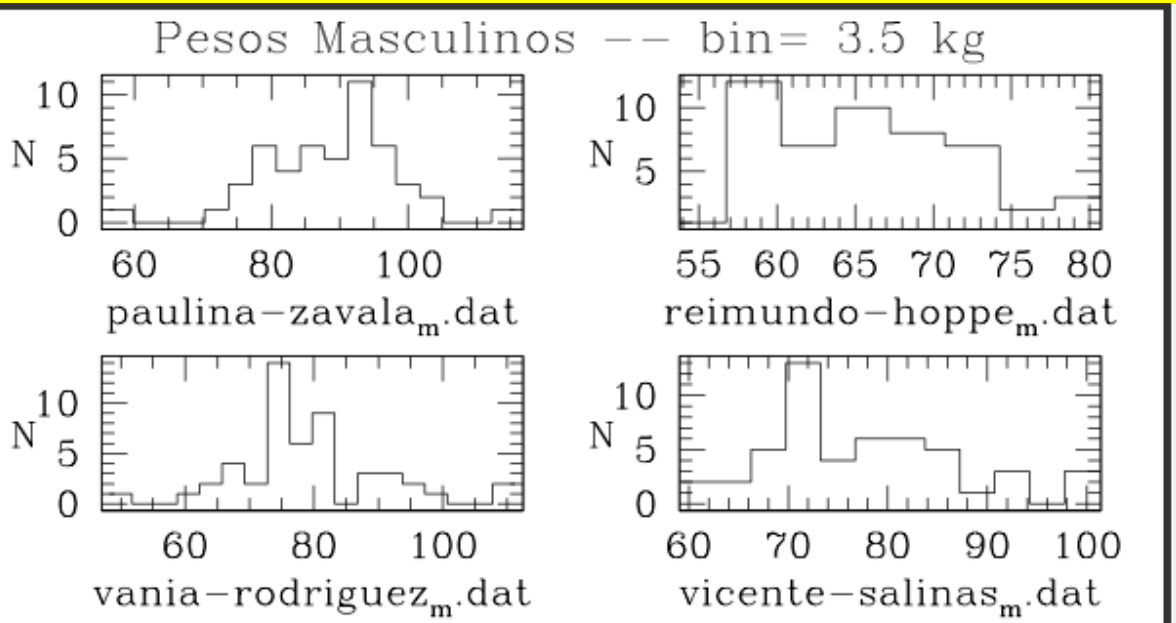


El histograma que a mi me parece mejor está más cerca del criterio complejo de calcular $[C(\Delta_x)]$, pero todos los Δ_x son relativamente próximos. Un buen compromiso es recordar la regla de Excel y probar a partir de ella.

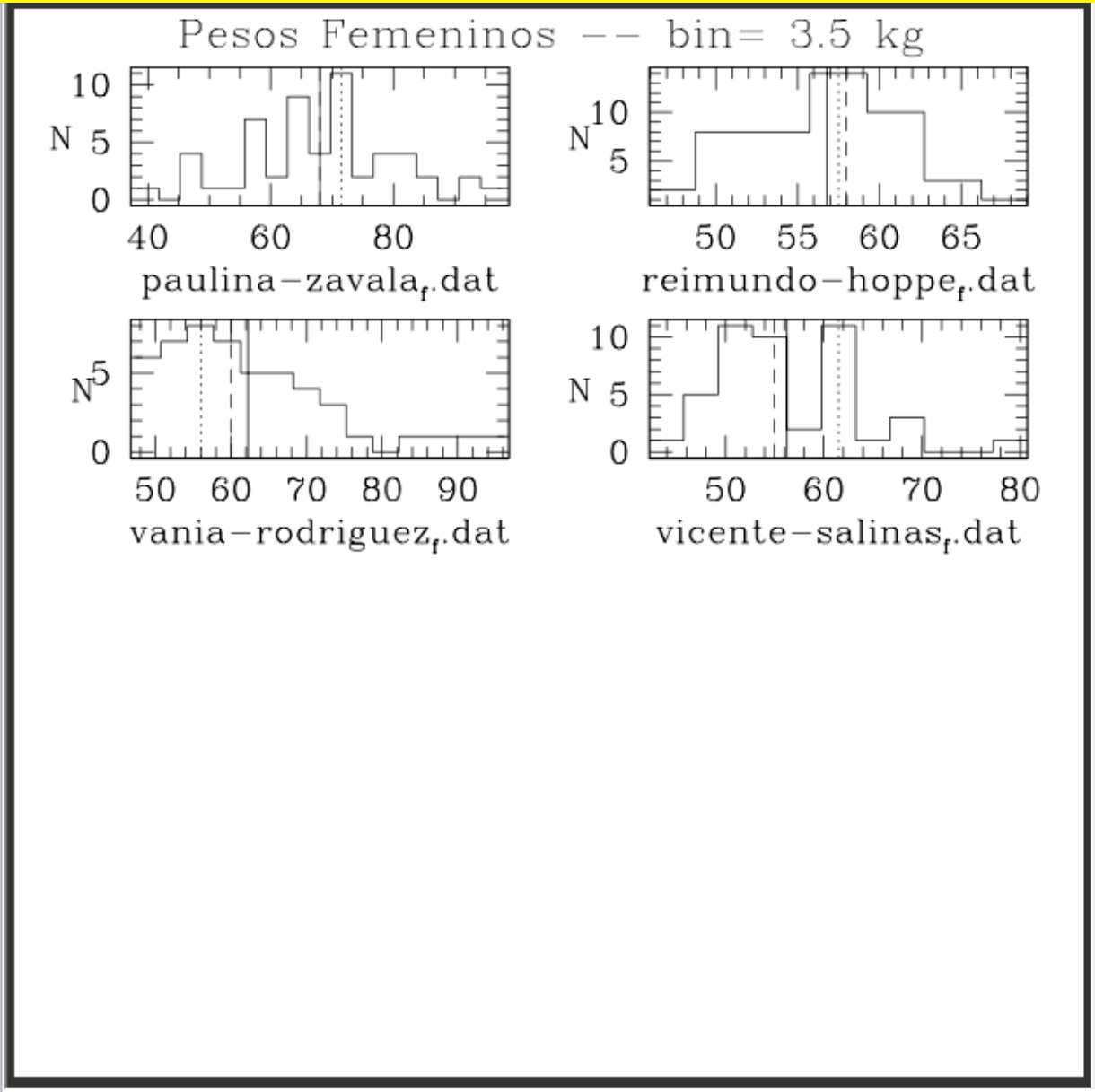
Histogramas de datos inventados



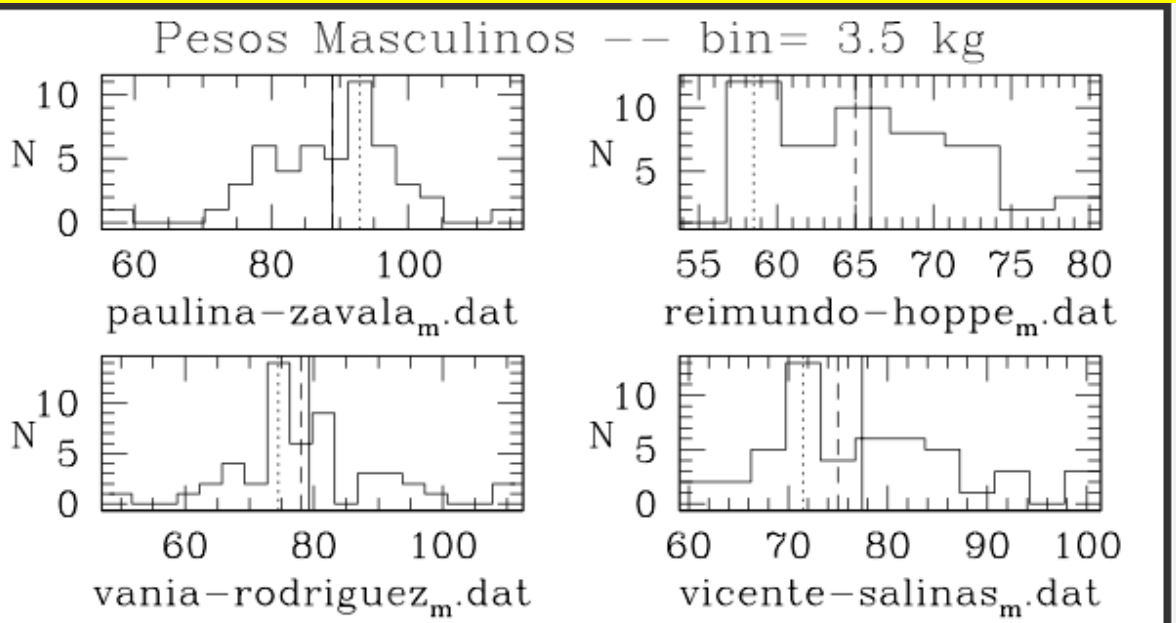
Histogramas de datos inventados



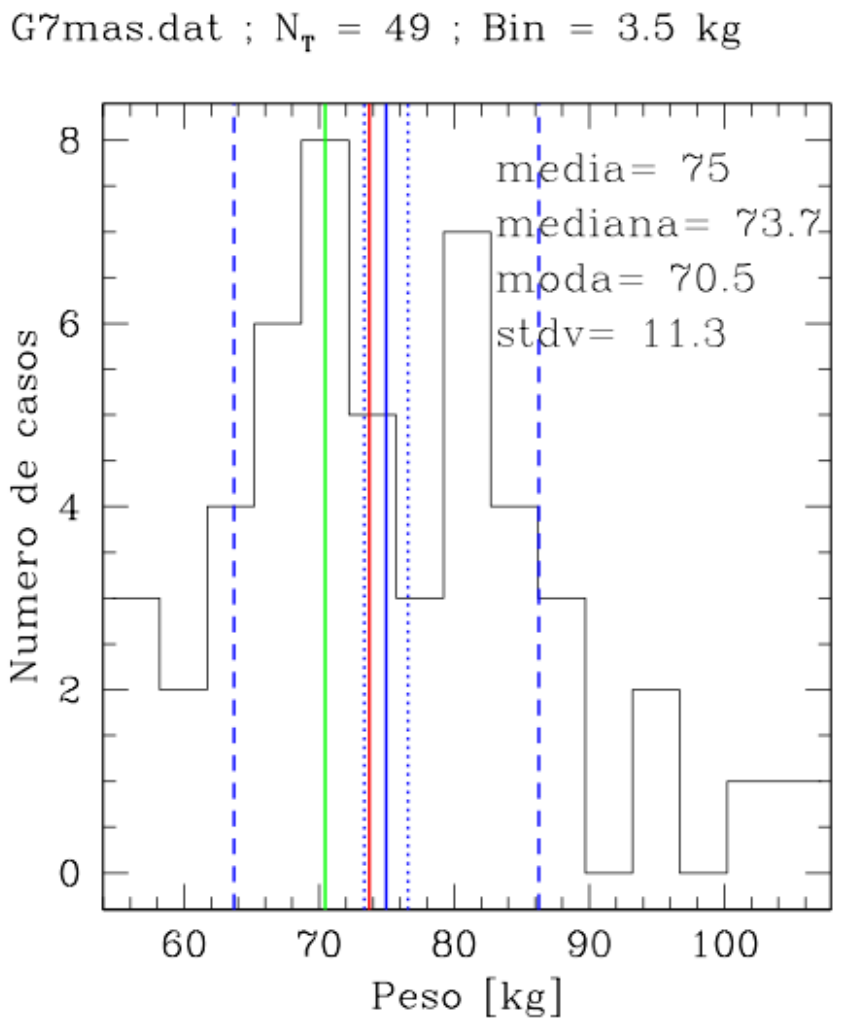
Histogramas de datos inventados



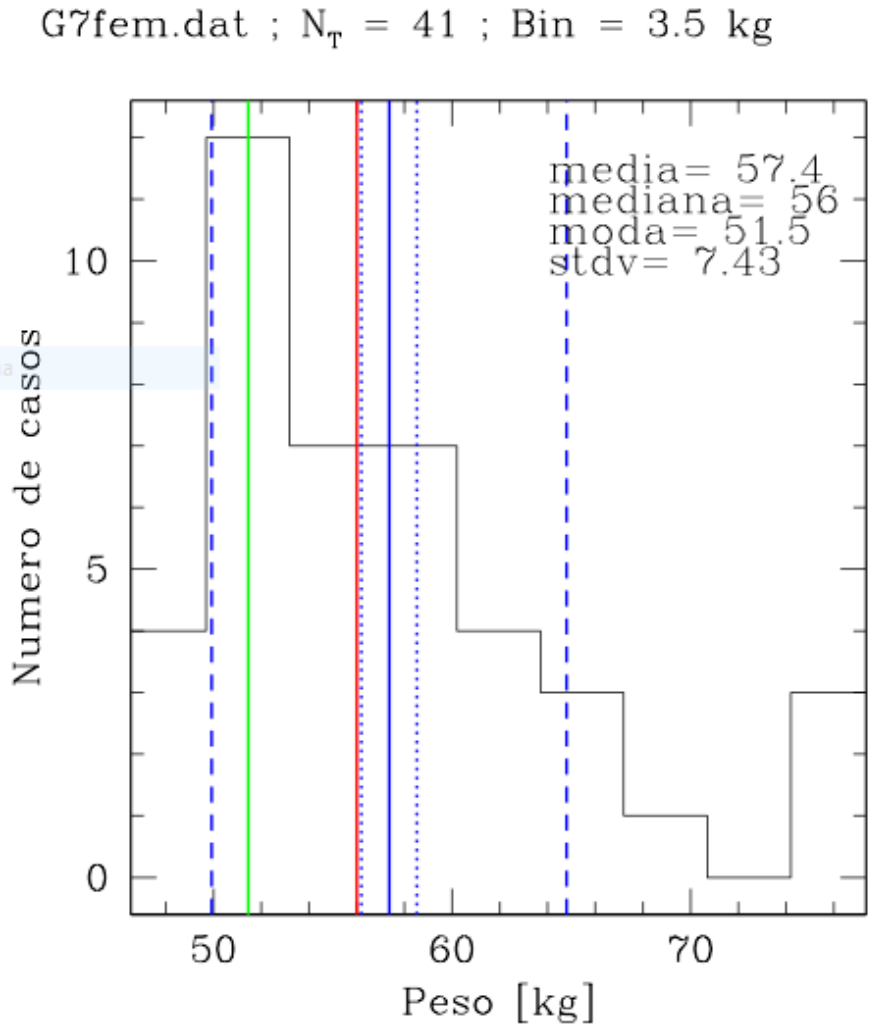
Histogramas de datos inventados



Histogramas de datos reales

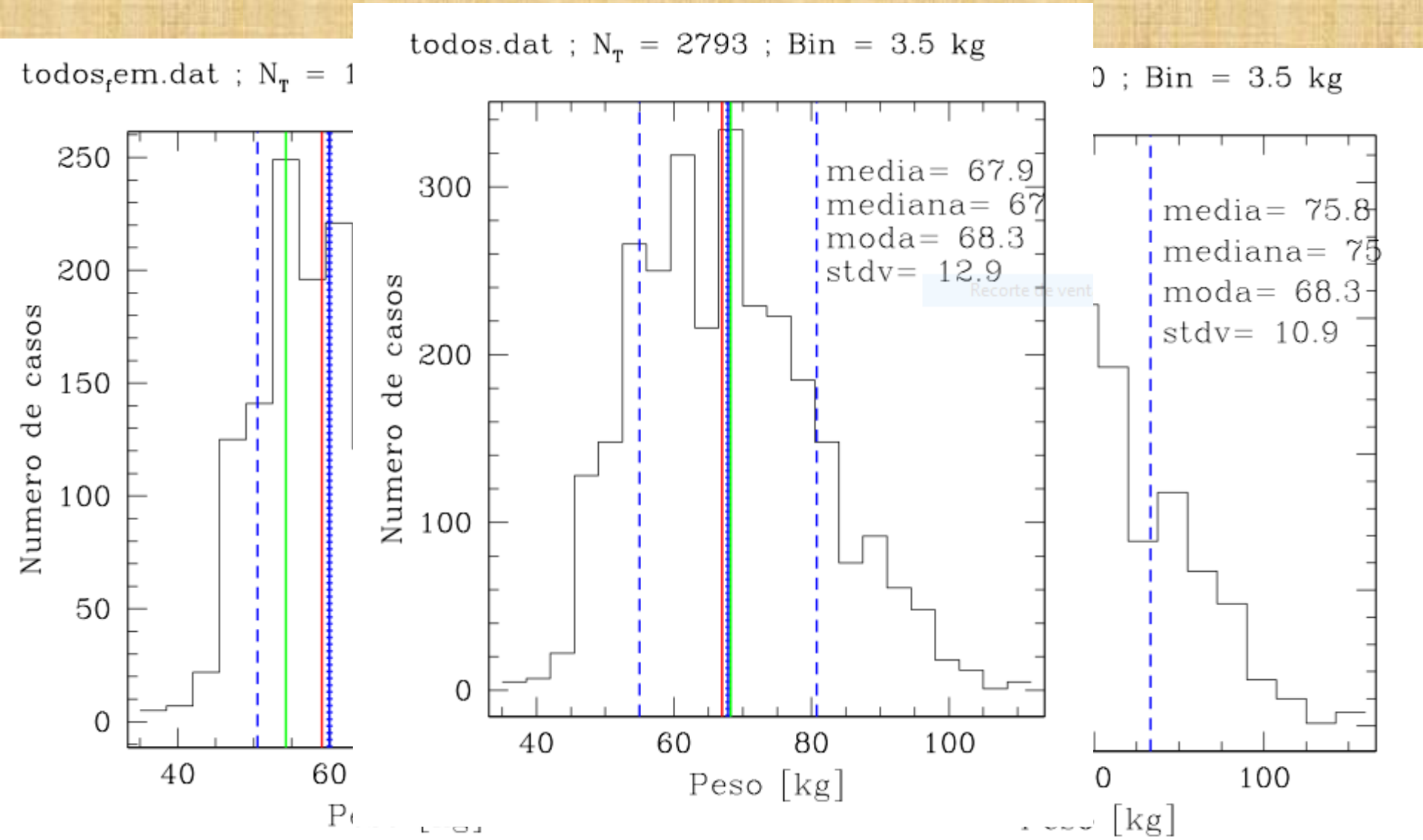


Histogramas de datos reales



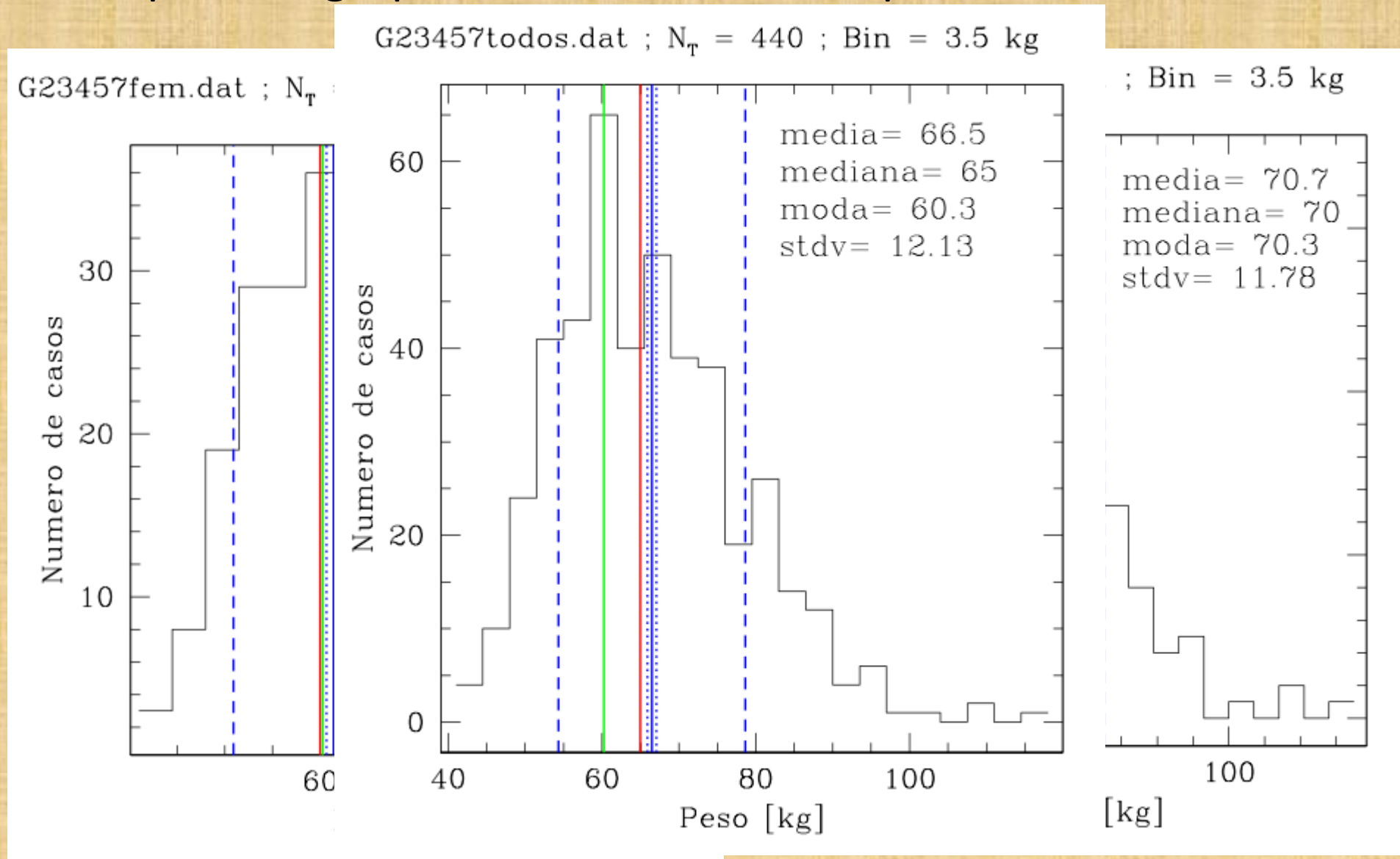
Histogramas de datos inventados

¿Qué pasa si agrupamos todos los datos que *imaginaron* ustedes?



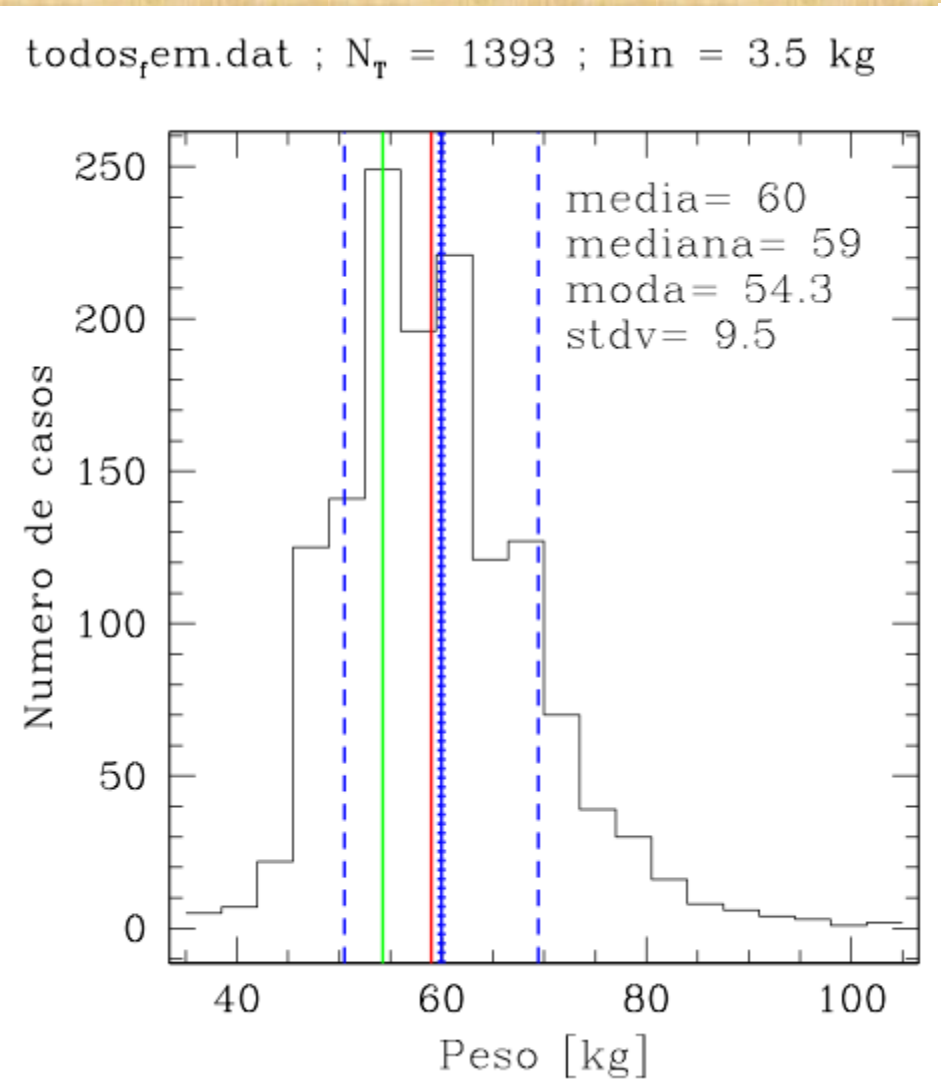
Histogramas de datos observados

¿Qué pasa si agrupamos todos los datos que *tomaron* ustedes?

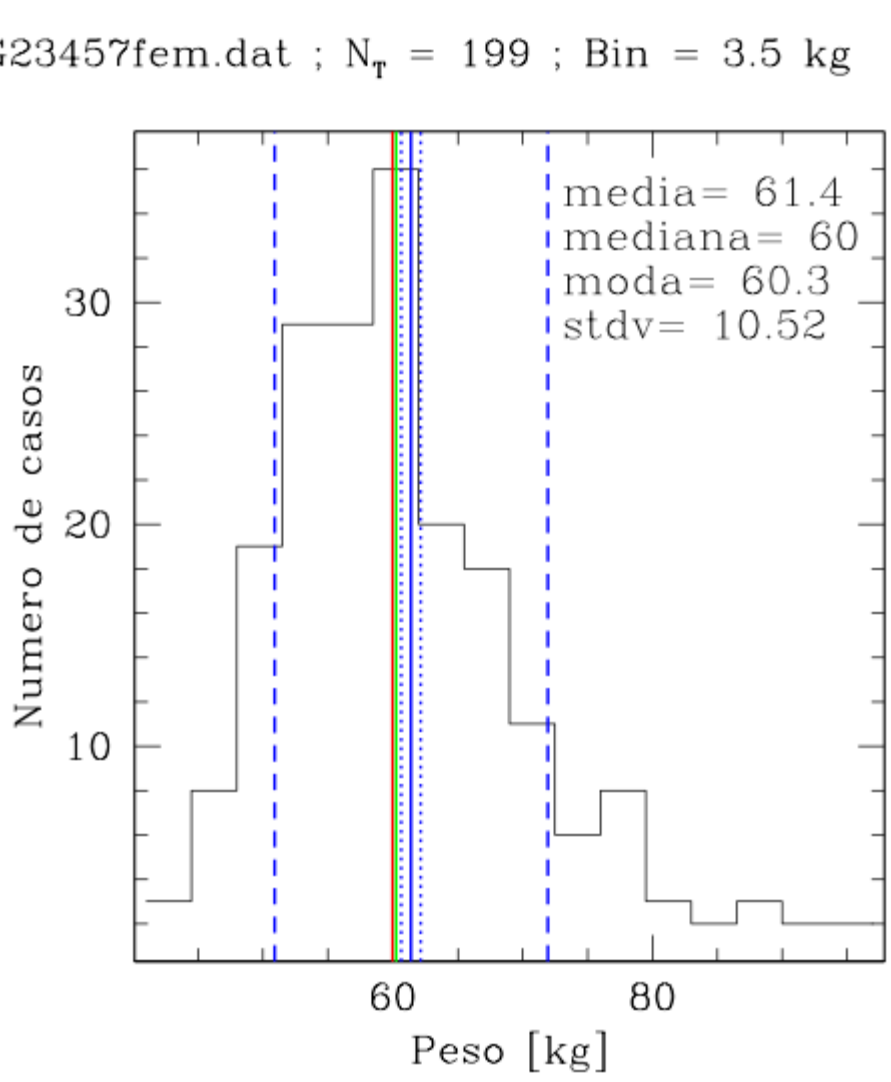


Histogramas imaginados vs. observados

Imaginado

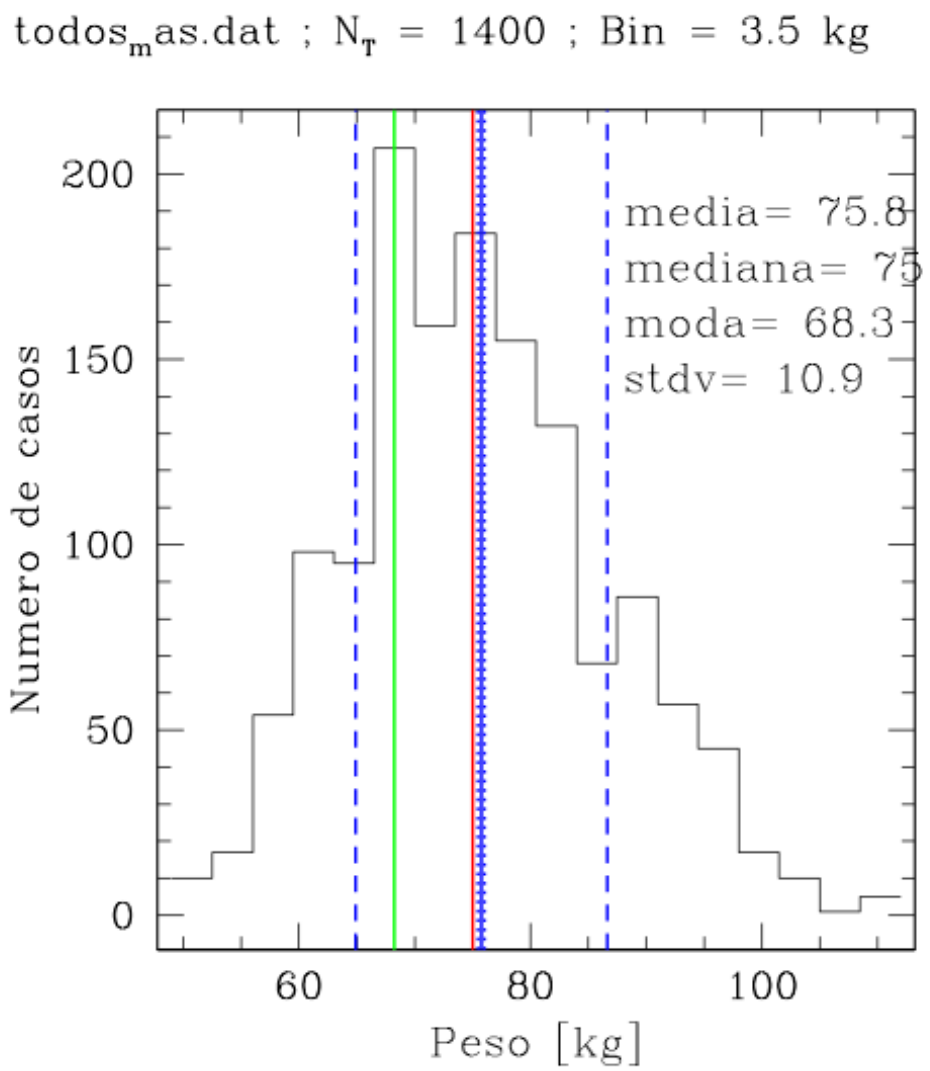


Observado

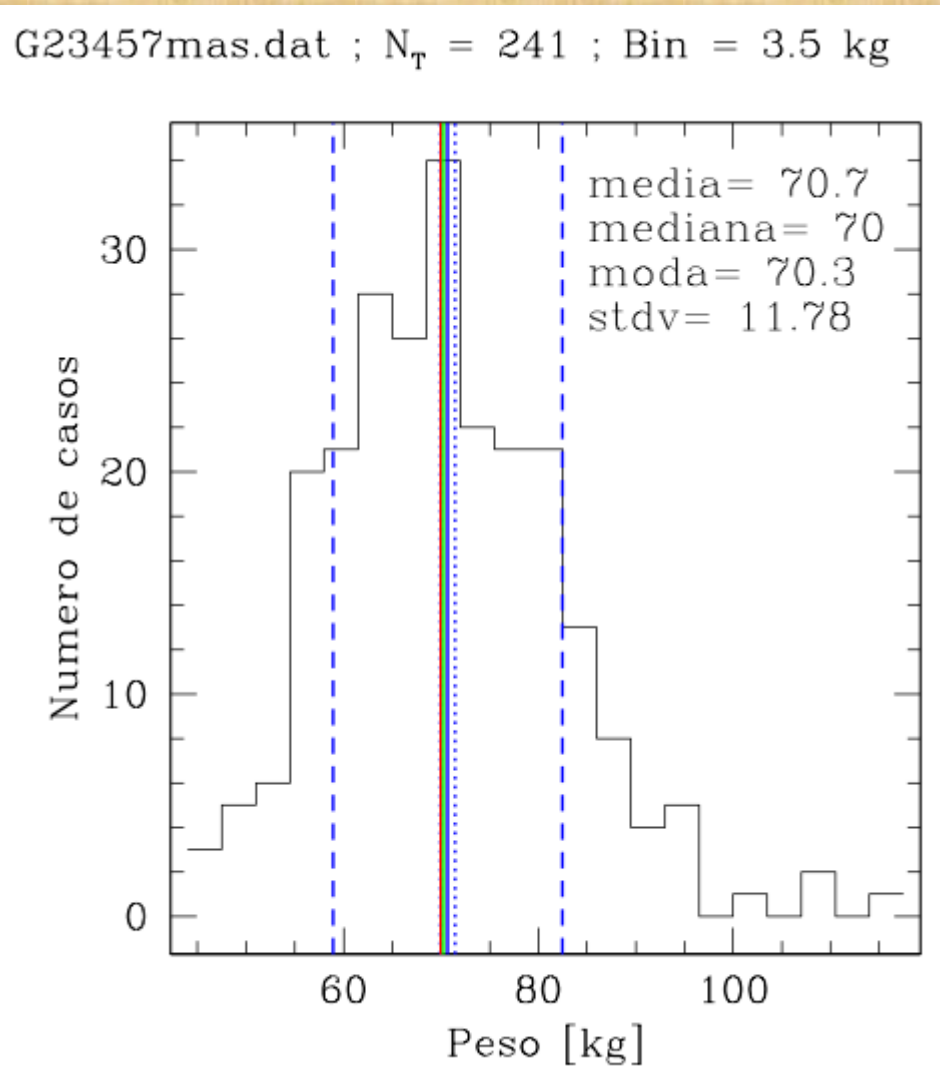


Histogramas imaginados vs. observados

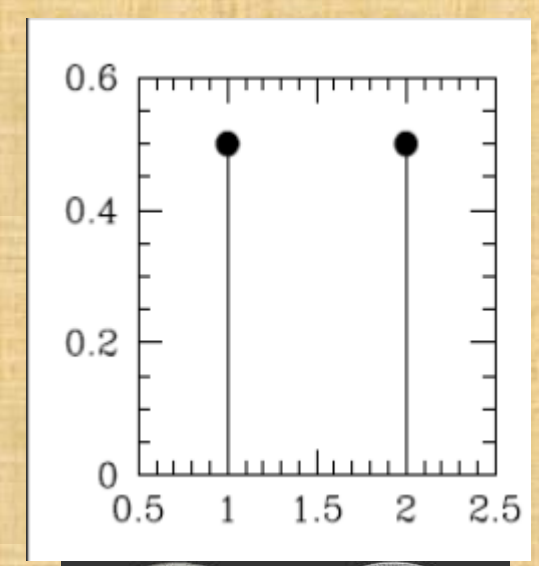
Imaginado

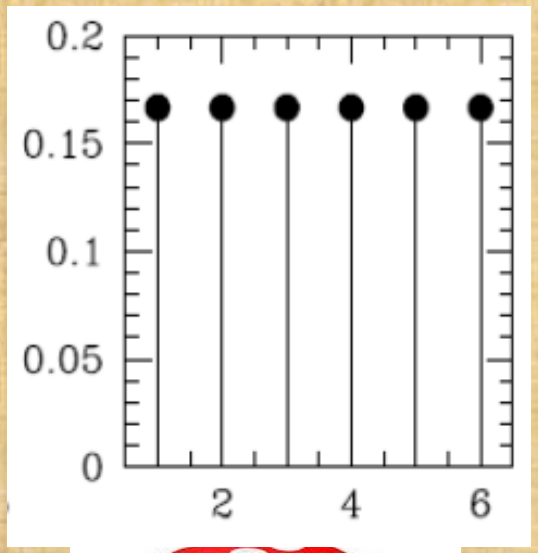


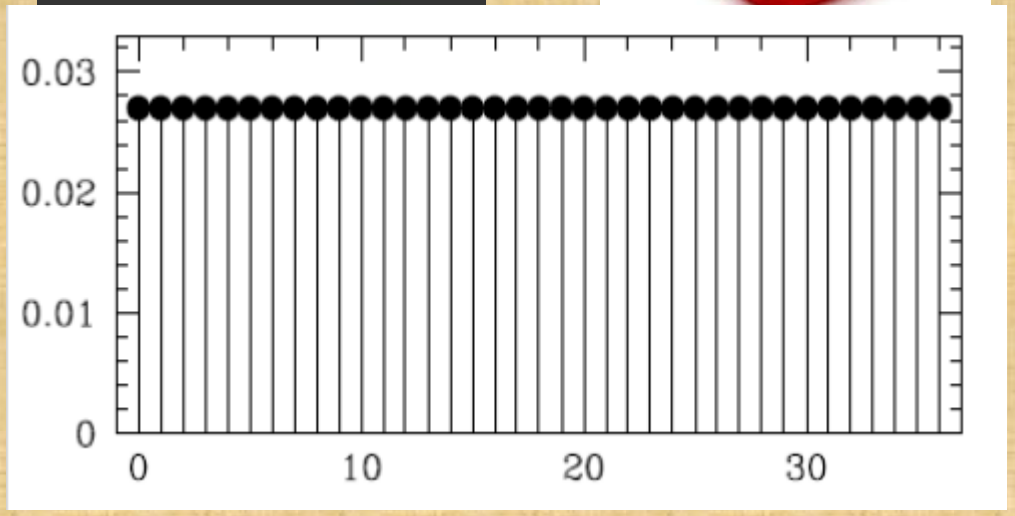
Observado



Funciones de distribución de probabilidad

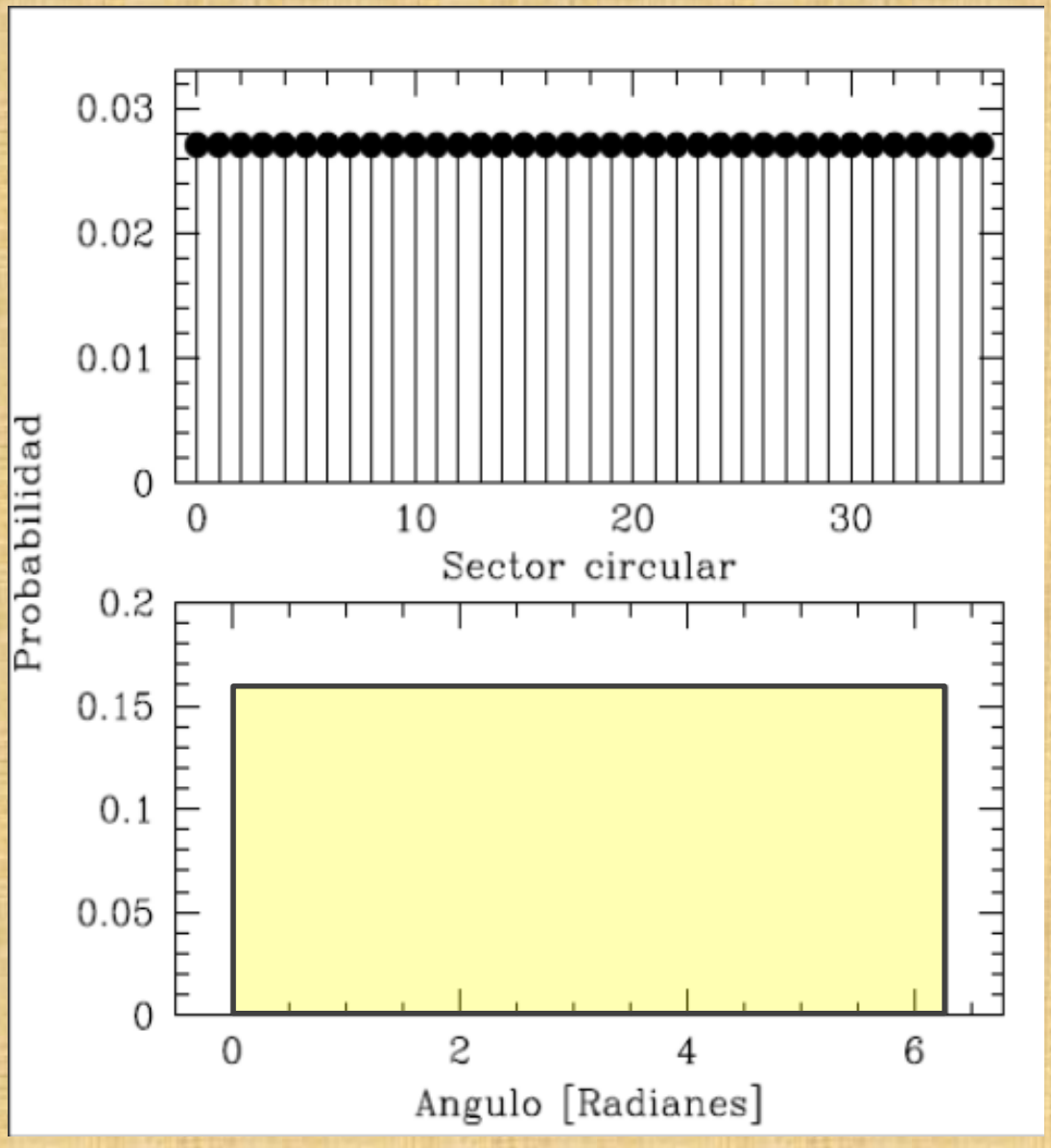






¿?

Funciones de distribución de probabilidad

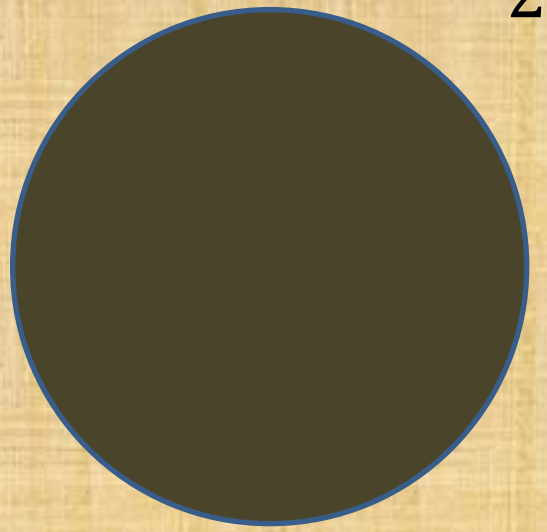


$$P_i = \frac{1}{37} \qquad \sum_{1}^{37} P_i = 1$$

$$dP_\theta = Cd\theta$$

$$P_{\theta_1<\theta<\theta_2} = \int_{\theta_1}^{\theta_2} Cd\theta$$

$$\Rightarrow C = \frac{1}{2\pi}$$



Funciones de distribución de probabilidad

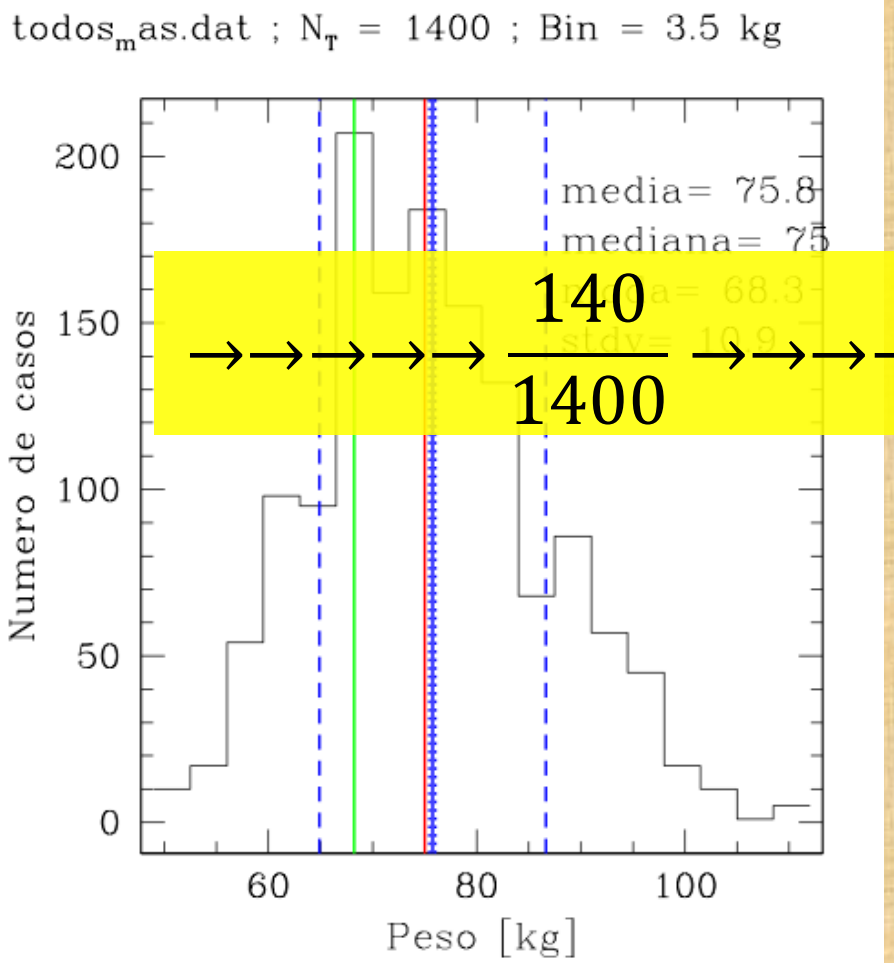
La FDP para el resultado del experimento de rotar un disco y tomar nota del ángulo en el que se detiene es la forma más simple de una FDP continua.

Puedo usar la forma de la FDP para calcular los valores que tienen los parámetros teóricos de la distribución, por ejemplo valor medio y varianza.

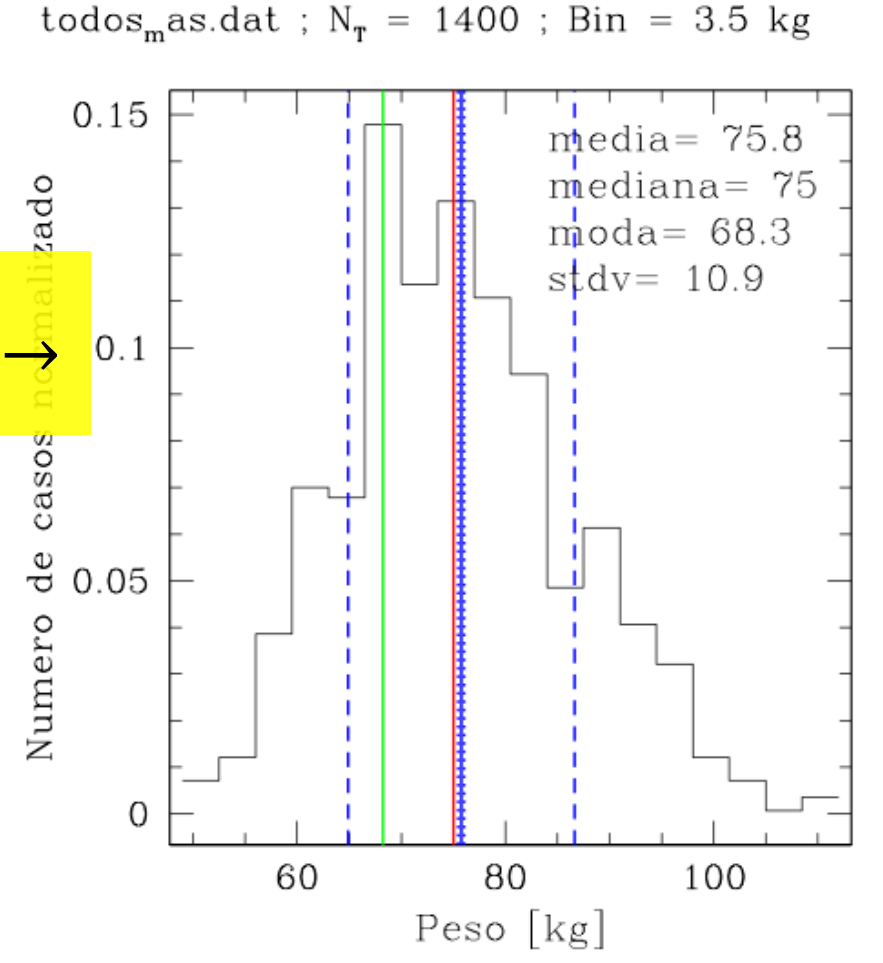
Para entender esto un poco mejor, miremos de nuevo un histograma y tratemos de verlo como una aproximación a una FDP.

Histogramas como FDP discretas

Un histograma puede ser entendido como una FDP unidimensional discreta que asigna una cierta probabilidad a que el valor de la variable (x) en consideración esté comprendido en el intervalo Δx en torno al centro del j-ésimo *bin*. Para ilustrar esto sólo tenemos que dividir el histograma completo por el número total de casos:



$$\rightarrow\rightarrow\rightarrow\rightarrow\rightarrow \frac{140}{1400} \rightarrow\rightarrow\rightarrow\rightarrow\rightarrow$$



Histogramas como FDP discretas

Para calcular el valor medio de la variable *x* cuando la teníamos clasificada dentro de los intervalos de un histograma (lo llamamos antes "caso de datos agrupados"), teníamos:

$$\overline{x_g} = \frac{1}{N}\sum_{j=1}^{M} n_j \overline{x_j}$$

De ésta →

$$\overline{x_g} = \sum_{j=1}^{M} \frac{n_j}{N}\overline{x_j} = \sum_{j=1}^{M} P_j \overline{x_j}$$

donde $P_j = \frac{n_j}{N}$ es la probabilidad de que la variable x esté en el *bin j.*

Para el caso de una variable continua, la sumatoria tiende a una integral, exactamente igual que para la definición de integral como límite de una sumatoria (notar que $M \to \infty, \Delta x = (x_j - x_{j-1}) \to 0$):

$$\overline{x_g} = \sum_{j=1}^{M} P_j \overline{x_j} \to \int_{x_1}^{x_N} x P_x dx$$

x_1 y x_N son los límites entre los cuales la variable x está definida.

Fin de ppt de Clase 3