

Estadística



Estadística

- La **estadística** es una ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.
- **Características**
 - Hace uso de la matemática
 - Trabaja sobre datos adquiridos
 - Busca explicar o interpretar situaciones donde hay incertidumbre y variación.

Estadística

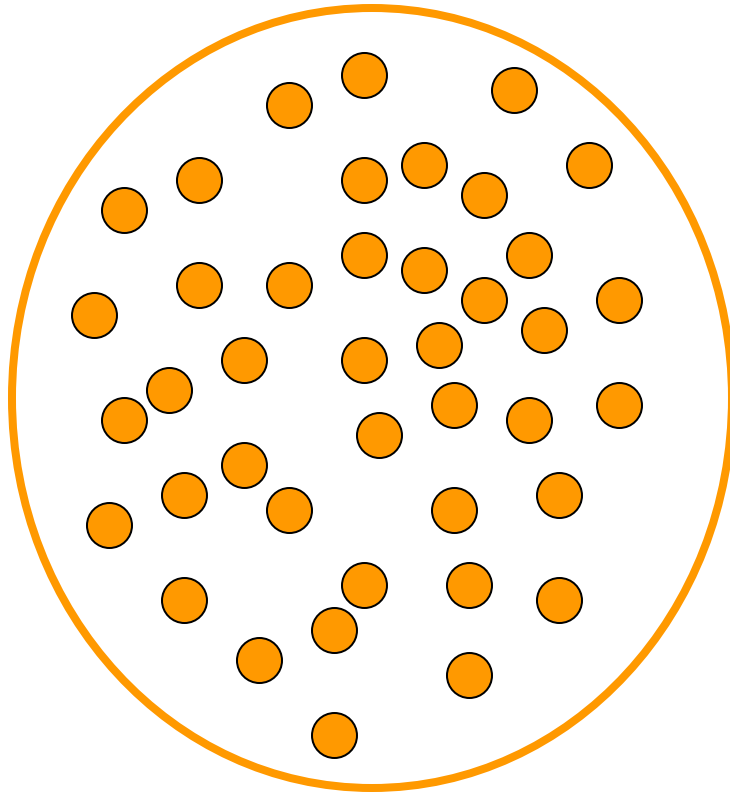
- La estadística proporciona métodos para **organizar y resumir datos**.
 - Ej: utilizando medidas generales como el valor medio, la mediana y la desviación estándar.
- También para **sacar conclusiones** a partir de la información que contienen.
 - Ej: A partir de las pruebas realizadas en pacientes a los que se les aplicó cierta droga puede estimarse si hubo mejora en su salud o no.

Población

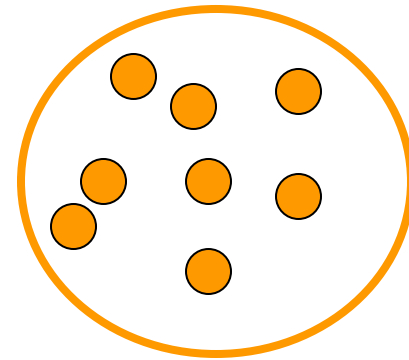
- Los datos utilizados se refieren a la **población** de interés.
- Ejemplos de población
 - Todos los egresados de la Facultad de Informática durante los últimos 5 años.
 - Todos los autos fabricados por Chevrolet Argentina durante 2007 y 2008.
- Si se dispone de la misma información para todos los objetos de la población, lo que se tiene es un **censo**.

Población y muestra

POBLACION



MUESTRA



Por cuestiones prácticas se trabaja con una **muestra** (un subconjunto de la población)

Ramas de la Estadística

□ **Estadística Descriptiva**

- Se dedica a los métodos de recolección, descripción, visualización y resumen de los datos.

□ **Inferencia Estadística**

- Se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones.

Estadística Descriptiva

- Medidas de Resumen Numéricas
 - Media y Mediana
 - Cuartiles
 - Medidas de dispersión

- Representaciones gráficas
 - Diagramas de caja.
 - Histograma.

Media Muestral

- La media muestral \bar{x} de un conjunto de observaciones x_1, x_2, \dots, x_n está dada por

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

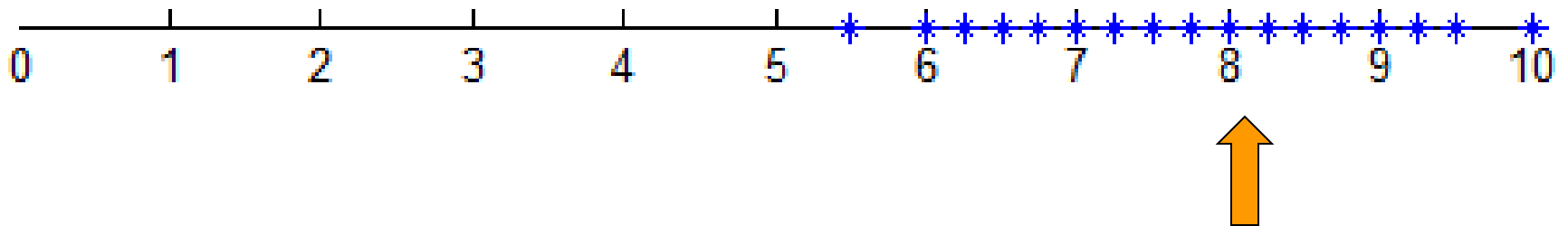
Media Muestral

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

Alumno	Nota
Angioni Formia, Bruno Gabriel (bgangioni)	8.75
Aparicio, Ivan Agustin (in5f480)	9.00
Apezteguia, Matias (matiasap)	8.25
Archuby, Federico (archu)	9.00
...	...
Trujillo, Leticia Vanesa (truleto5p)	9.25
Vallejos, Fernando (yabran)	9.00

Media Muestral

$$\overline{X} = \frac{\sum_{i=1}^n x_i}{n} = 8.1571$$



8.1571

(medida de resumen numérica)

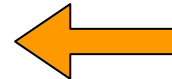
Mediana

- Ordenar las muestras de menor a mayor y tomar como valor para la mediana \tilde{x}
 - El valor del medio de la lista si la cantidad de elementos es impar.
 - El promedio de los valores centrales si la cantidad de elementos de la lista es par.

Mediana

$$\tilde{x} = 8,25$$

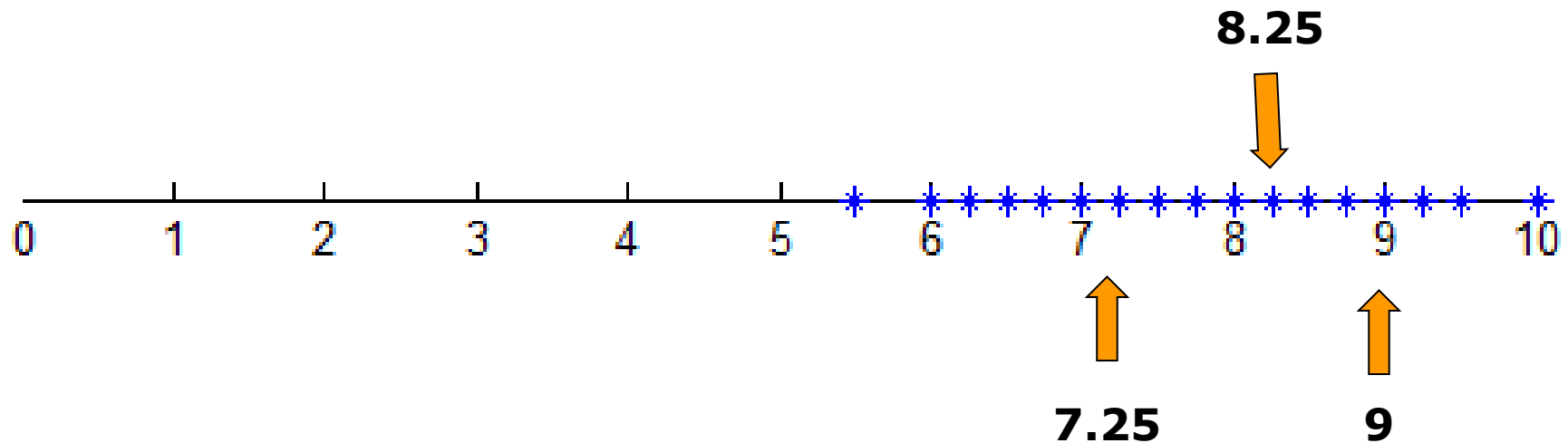
Nota	# Orden
5,5	1
...	...
8,25	52
8,25	53
8,25	54
...	...
10	105



Ordenar las
notas de menor
a mayor y tomar
la nota del que
está al medio

Cuartiles

- La mediana divide al conjunto de datos en dos partes iguales.
- Los **cuartiles** dividen los datos en 4 partes con la misma cantidad de valores

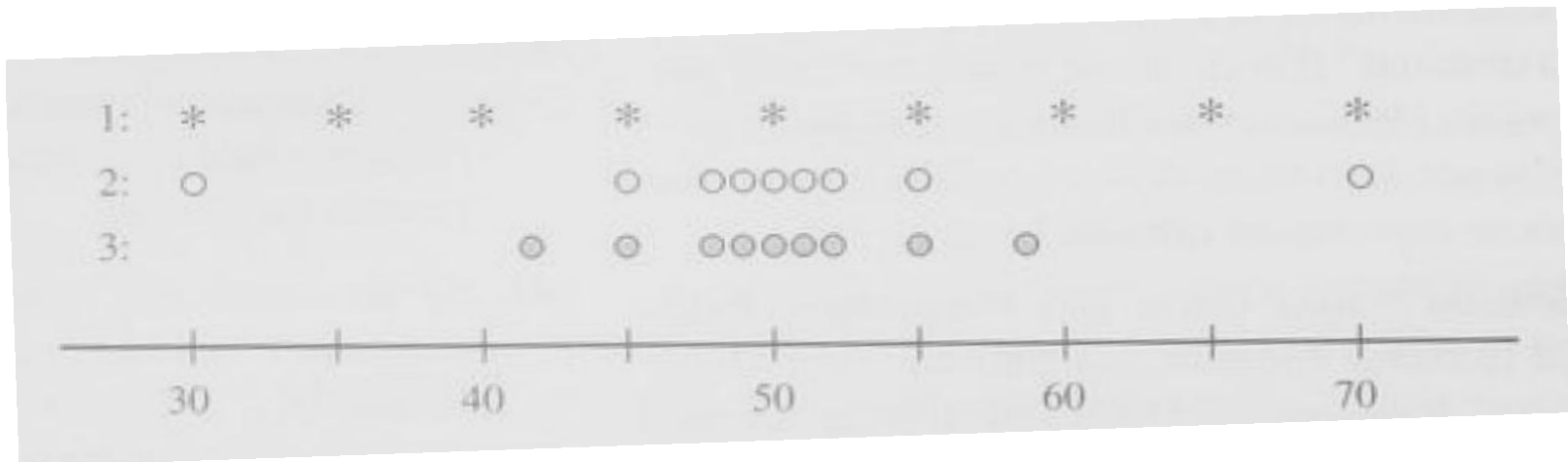


Medidas en Excel

Medidas numéricas	Valor	Función Excel
Media	8,1571	=AVERAGE(B2:B106)
Mediana	8,25	=MEDIAN(B2:B106)
Primer Cuartil	7,25	=QUARTILE(B2:B106; 1)
Segundo Cuartil	8,25	=QUARTILE(B2:B106; 2)
Tercer Cuartil	9	=QUARTILE(B2:B106; 3)

Medidas de Dispersión

- Estas tres muestras tienen la misma media pero distinta dispersión

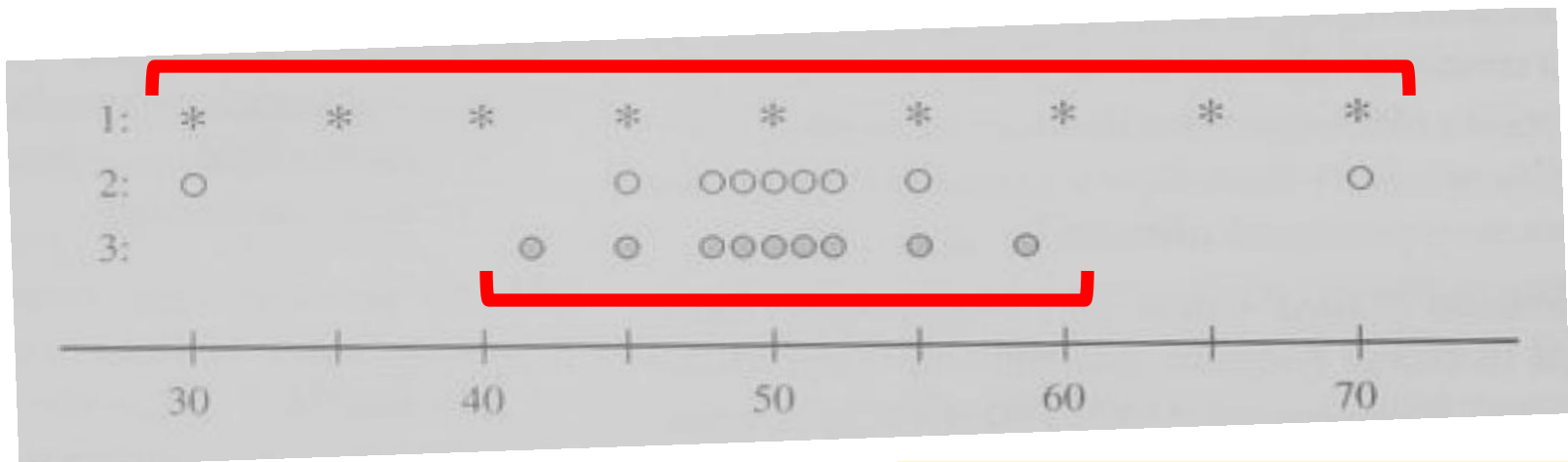


- La muestra 1 es la que tiene mayor variación y la muestra 3 es la más compacta.

Medidas de Dispersión para datos muestrales

- La más simple es el **rango** o **recorrido** que es la diferencia entre los valores extremos

Ej: La muestra 1 tiene rango $70-30=40$ mientras que la muestra 3 tiene un recorrido menor



¿ Desventajas ?

Medidas de Dispersión

- Las principales medidas utilizan las desviaciones a partir de la media.
- Es decir que se consideran las diferencias de cada muestra con la media

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$$

- Una opción natural parece ser la suma


$$\sum_{i=1}^n (x_i - \bar{x})$$

¿ PROBLEMAS ?

Medidas de Dispersión


□ Suma de las desviaciones

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x}$$


 $n \cdot \bar{x}$

Medidas de Dispersión

□ Suma de las desviaciones

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n \cdot \bar{x} = 0$$


¿Cómo cambiar las desviaciones a cantidades no negativas?

$\frac{1}{n} \sum_{i=1}^n x_i$

Medidas de Dispersión

□ ¿Cómo cambiar las desviaciones a cantidades no negativas?

□ **Opción 1**

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

□ **Opción 2**

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

La función valor absoluto tiene algunas dificultades teóricas entonces usamos esta

Medidas de Dispersión

- ¿Cómo cambiar las desviaciones a cantidades no negativas?

- **Opción 1**

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- **Opción 2**

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

En realidad, se divide por ***n-1***

Varianza Muestral

- La **varianza muestral** se denota por S^2 y se define como

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- La **desviación muestral** se denota por **S** y es la raíz cuadrada positiva de S^2

Varianza Muestral

- La varianza muestral también puede expresarse como

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

donde

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

Varianza Muestral

- Volviendo al ejemplo de las calificaciones de la primera autoevaluación

Medida numéricas	Valor	Función Excel
Varianza	1,2094	=VAR(B2:B106)
Desviación	1,0997	=STDEV(B2:B106)

Estadística Descriptiva

□ **Medidas de Resumen Numéricas**

- Media y Mediana
- Cuartiles
- Medidas de dispersión

□ **Representaciones gráficas**

- Diagramas de caja.
- Histograma.

Veamos
estas

Diagramas de Caja

- Características del conj.de datos
 - Centro
 - Dispersión
 - Desviación respecto a la simetría
 - Identificación de valores atípicos (alejadas del grueso de las observaciones).
- Utiliza medidas "*resistentes*" a los datos atípicos: la mediana y la cuarta dispersión

Diagramas de Caja

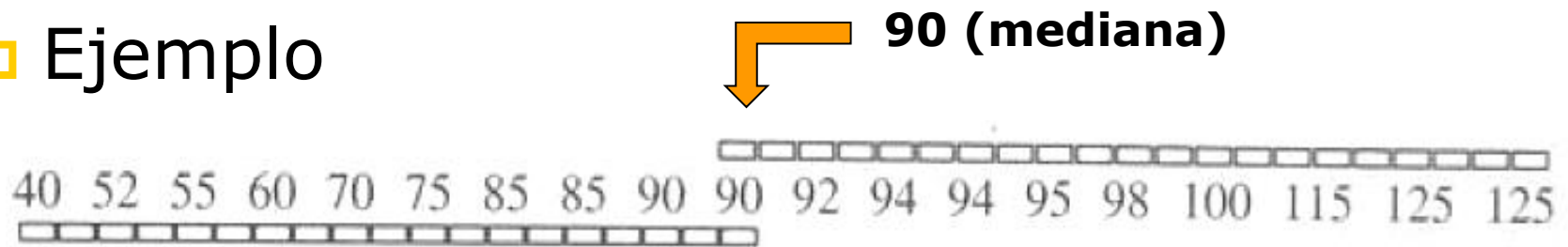
- Primero se ordenan las muestras de menor a mayor.
- Luego, la **cuarta dispersión** f_s está dada por

$$f_s = \text{cuarto superior} - \text{cuarto inferior}$$

donde el cuarto inferior es la mediana de la primera mitad y el cuarto superior la mediana de la segunda mitad.

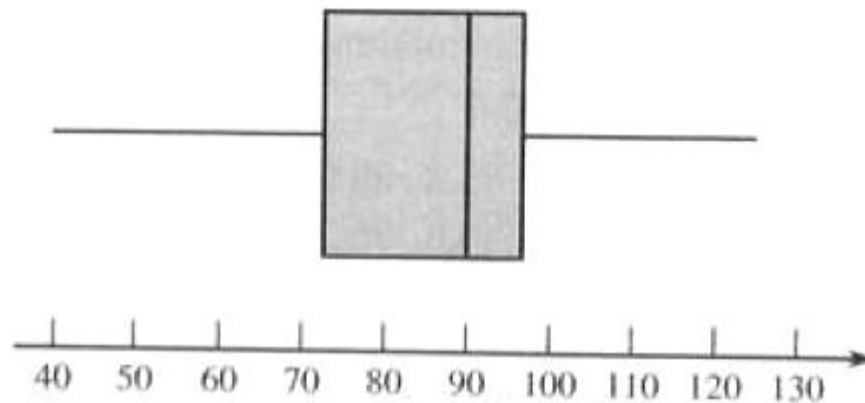
Diagramas de Caja

□ Ejemplo



72.5
(cuarto inferior)

96.5
(cuarto superior)



El ancho de
la caja es f_s

Valores atípicos

- Cualquier muestra más allá de $1.5 f_s$ desde el cuarto más cercano es un **valor atípico**.
- Un valor atípico es **extremo** si está a más de $3 f_s$ del cuarto más cercano.
- Entre $1.5 f_s$ y $3 f_s$ es se considera un valor atípico **moderado**.

Diagrama de Caja

- Estos son los valores correspondientes a las calificaciones de la primera autoevaluación

Medida	Notas
Cuarto inferior	7,25
Valor mínimo	5,5
Mediana	8,25
Valor máximo	10
Cuarto Superior	9

Diagrama de Caja con Excel

- Seleccionar la tabla anterior e insertar un gráfico de línea con marcadores
>Insert > Line Chart > Line with Markers

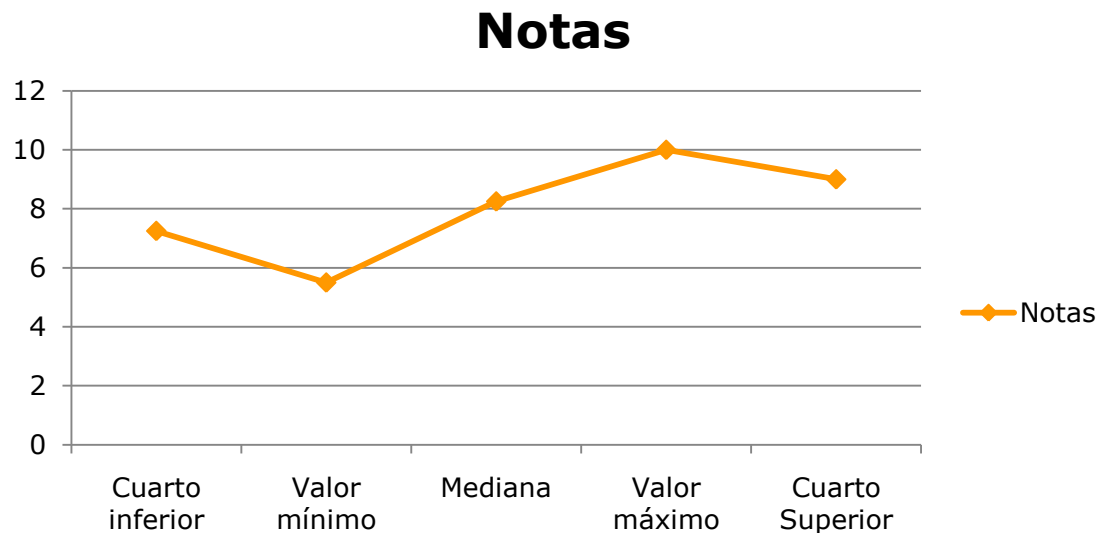


Diagrama de Caja con Excel

- Clickear con el botón derecho del mouse sobre cualquier de los marcadores y seleccionar

>Format Data Serie > line Color > No Line

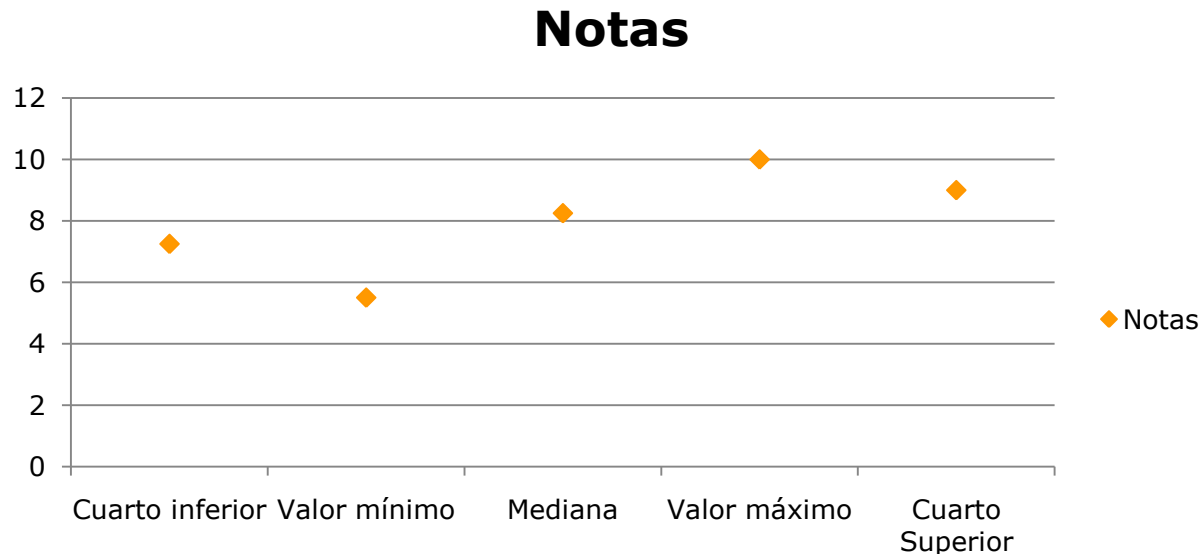


Diagrama de Caja con Excel

- Invertir filas por columnas: Ir a la solapa **Design** y elegir **Switch Row/Column**

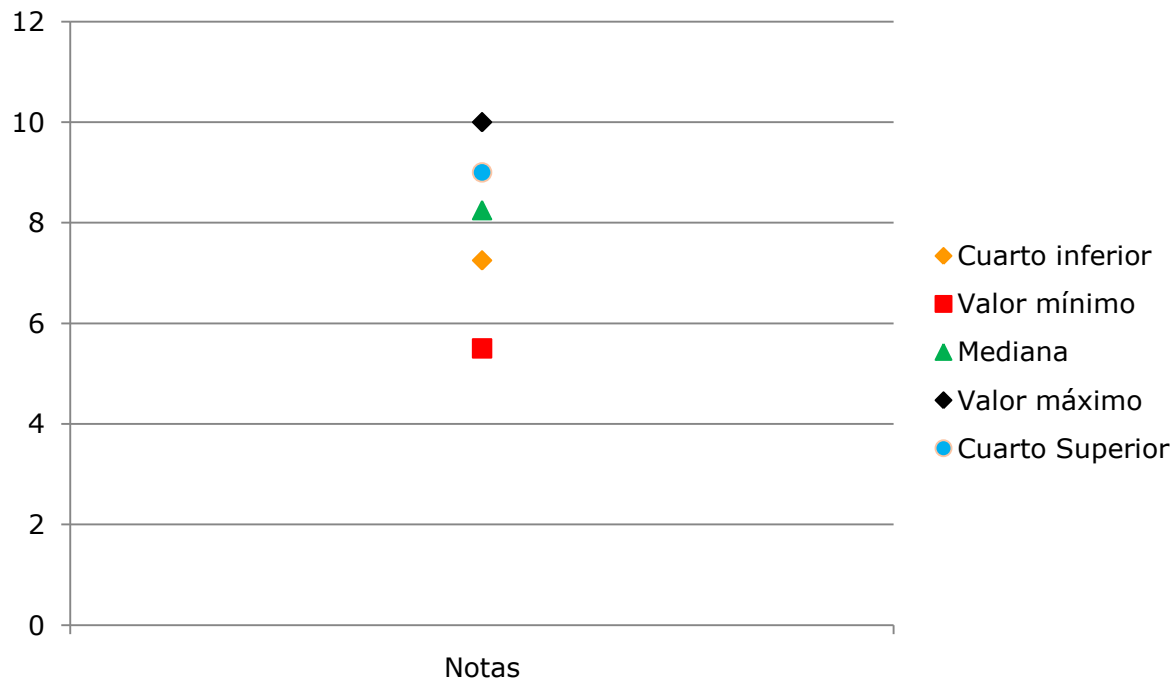


Diagrama de Caja con Excel

- En la solapa **Layout** elegir
> **Lines** > **High-Low Lines**

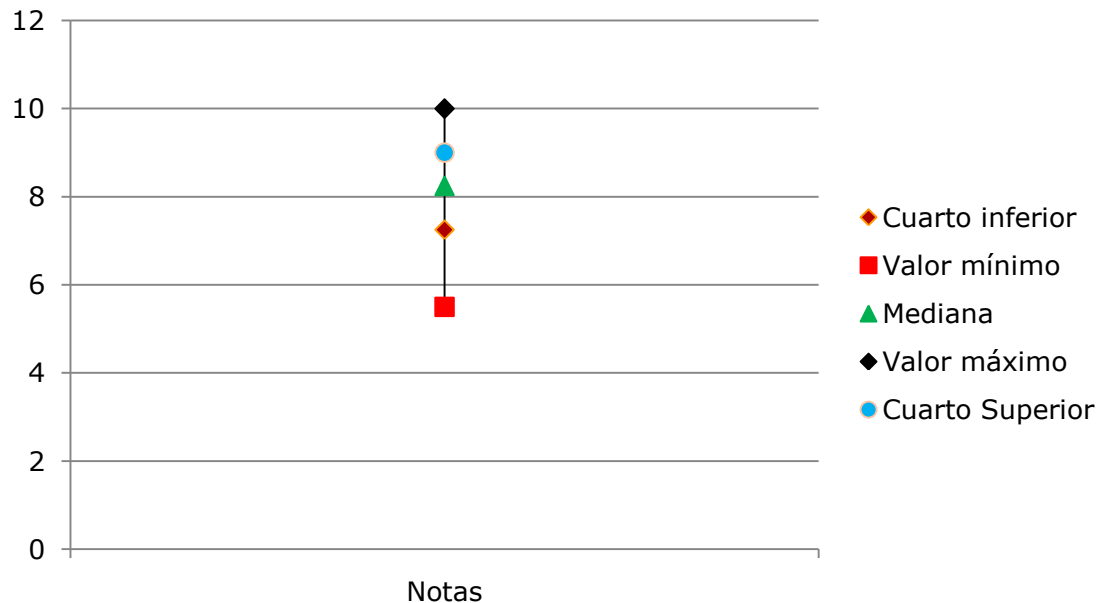
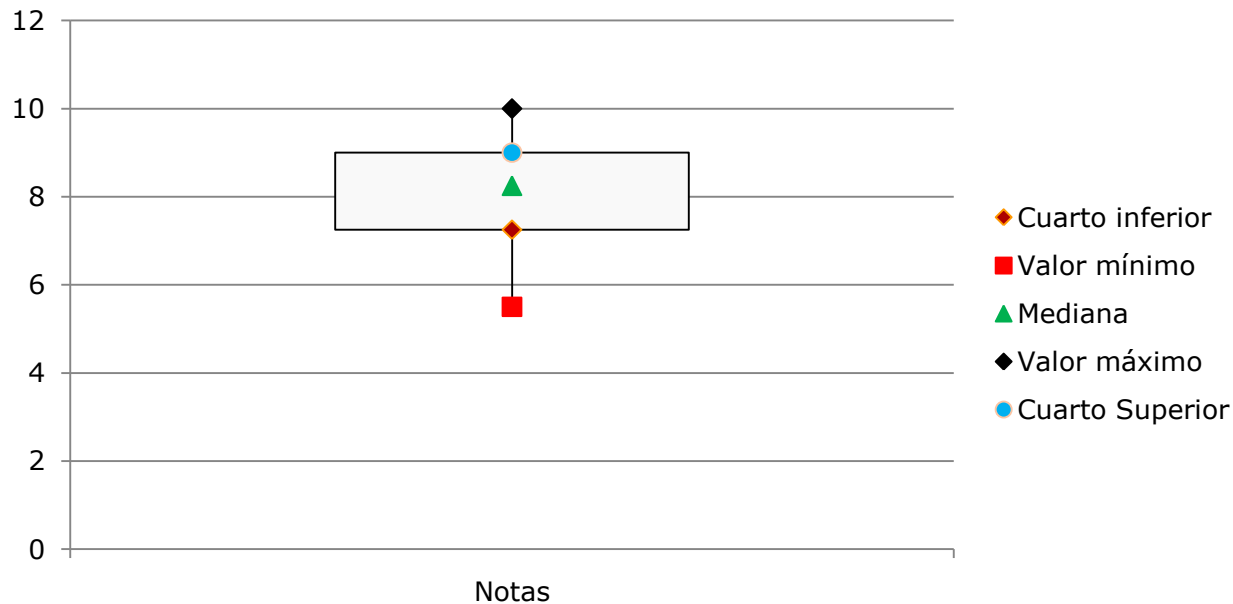


Diagrama de Caja con Excel

- En la solapa **Layout** elegir
- **>Up/Down Bars > Up/Down Bars**



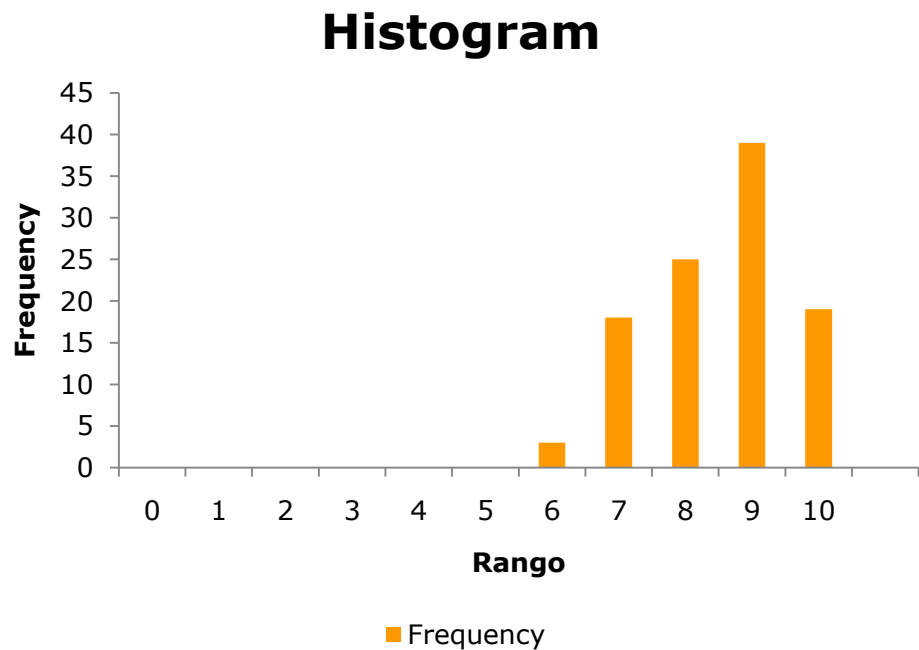
Histograma

- Permite apreciar la frecuencia con que aparecen los distintos valores de una v.a.
- Representación
 - Sobre el eje X se indican los valores de la variable.
 - Sobre el eje Y se representa la frecuencia relativa o la frecuencia con la que cada valor aparece.
- Si la variable es continua es preciso discretizarla.

Histograma

- El histograma de las calificaciones de la primera autoevaluación es

<i>Rango</i>	<i>Frequency</i>
0	0
1	0
2	0
3	0
4	0
5	0
6	3
7	18
8	25
9	39
10	19



Muestras Aleatorias



Muestra

- Considere elegir dos muestras distintas de tamaño n de la misma distribución poblacional
- **Ejemplo:** Consumo de combustible de 3 autos

	Muestra 1	Muestra 2
x_1	30.7	28.8
x_2	29.4	30.0
x_3	31.1	31.1

Antes de obtener los datos hay incertidumbre acerca del valor de cada x_i , por lo tanto cada observación se ve como una v.a.

Cada muestra se representa mediante X_1, X_2, \dots, X_n (en este ejemplo $n=3$)

Muestra

- Considere elegir dos muestras distintas de tamaño n de la misma distribución poblacional
- **Ejemplo:** Consumo de combustible de 3 autos

	Muestra 1	Muestra 2
x_1	30.7	28.8
x_2	29.4	30.0
x_3	31.1	31.1



Casi siempre los valores de la 2da.muestra serán un poco distintos a los de la 1ra.

Muestra

	Muestra 1	Muestra 2
x_1	30.7	28.8
x_2	29.4	30.0
x_3	31.1	31.1
$\bar{\mathbf{X}}$	30.4	29.97
\mathbf{S}	0.89	1.15

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}}$$

- Las variaciones entre muestras hacen que cualquier función de las observaciones muestrales (ej: media muestral $\bar{\mathbf{X}}$, desviación estándar muestral \mathbf{S} , etc) cambie de una muestra a otra.

Ejemplo

- El tiempo que tarda un conductor en reaccionar a las luces de freno de un vehículo en desaceleración tiene una distribución normal con valor medio 1.25 segundos y desviación estándar 0.46 segundos.
- Analizar 6 muestras formadas por el tiempo de respuesta de 10 conductores cada una.

Ejemplo

$$\mu = 1.25 \text{ seg.}$$

$$\sigma = 0.46 \text{ seg.}$$

Nro.	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6
1	2,1230	1,4750	1,3066	2,2239	1,4109	1,7338
2	0,9071	1,0563	1,5586	1,3302	0,8332	1,5900
3	1,2096	1,3621	1,5233	1,5126	0,7559	1,3493
4	0,9780	2,0770	1,0361	0,7811	1,3896	0,7226
5	1,0719	0,9561	1,6450	0,5599	1,5712	1,0043
6	1,8029	1,2427	1,7326	1,7384	1,2130	1,6203
7	1,2374	1,1200	1,1444	1,0040	0,0208	1,1875
8	0,6790	1,6861	1,5334	1,3784	1,6667	1,0838
9	1,5620	2,2011	1,4218	0,9477	1,1868	1,3452
10	1,4220	2,1642	1,8251	2,0382	1,5130	1,0518
\bar{X}	1.2993	1.5341	1.4727	1.3515	1.1561	1.2689
S	0.4374	0.4728	0.2502	0.5422	0.4986	0.3183

Note que la media muestral y la desviación estándar muestral difieren de una muestra a otra

Estadístico

- Un **estadístico** es cualquier cantidad cuyo valor se calcula a partir de los datos de la muestra (ej: media muestral \bar{X})
- Antes de obtener los datos hay incertidumbre con respecto al valor que se obtendrá para un estadístico en particular. Por lo tanto, un **estadístico es una v.a.**
- Cualquier estadístico, que es una v.a., tiene una distribución de probabilidad también llamada **distribución de muestreo.**

Muestra aleatoria

- Se dice que las v.a. X_1, X_2, \dots, X_n forman una ***muestra aleatoria (simple)*** de tamaño n si
 - Las X_i son v.a. independientes.
 - Todas las X_i tienen la misma distribución de probabilidad.

Distrib.de muestreo de un estadístico

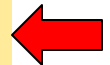
- Hay dos métodos generales para obtener la distribución de muestreo de un estadístico
 - Haciendo cálculos a partir de las reglas de probabilidad.
 - Puede usarse si se trata de una función muy simple de las X_i y hay pocos valores distintos de X en la población.
 - Realizando un experimento de simulación.

Ejemplo

- Un taller cobra 40, 45 y 50 u\$s por una afinación de autos de 4, 6 y 8 cilindros, respectivamente. Si 20% de las afinaciones se hacen en autos de 4 cilindros, 30% en autos de 6 cilindros y 50% en los de 8, entonces la distribución de probabilidad del ingreso de una sola afinación elegida al azar está dada por

x	40	45	50
p(x)	0.2	0.3	0.5

$$\mu = 46.5$$



$$\sigma^2 = 15.25$$

$$\mu = E(X) = \sum_{x \in R_x} x \cdot p(x) = 40 \cdot 0.2 + 45 \cdot 0.3 + 50 \cdot 0.5 = 46.5$$

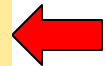
Ejemplo

- Un taller cobra 40, 45 y 50 u\$s por una afinación de autos de 4, 6 y 8 cilindros, respectivamente. Si 20% de las afinaciones se hacen en autos de 4 cilindros, 30% en autos de 6 cilindros y 50% en los de 8, entonces la distribución de probabilidad del ingreso de una sola afinación elegida al azar está dada por

x	40	45	50
p(x)	0.2	0.3	0.5

$$\mu = 46.5$$

$$\sigma^2 = 15.25$$



$$\sigma^2 = V(X) = \sum_{x \in R_x} (x - \mu)^2 \cdot p(x) =$$

$$(40 - 46.5)^2 * 0.2 + (45 - 46.5)^2 * 0.3 + (50 - 46.5)^2 * 0.5 = 15.25$$

Ejemplo

- Suponga que en un determinado día sólo dos servicios requieren afinación. Sea X_1 el ingreso obtenido de la 1ra. afinación y X_2 el de la 2da.
- Suponga que X_1 y X_2 son independientes, cada una con la distribución de probabilidad anterior.
- Es decir que X_1 y X_2 *constituyen una muestra aleatoria* de la distribución.

Ejemplo

x	40	45	50
p(x)	0.2	0.3	0.5

x1	x2	p(x1,x2)	\bar{x}	s ²
40	40	0.04	40	0
40	45	0.06	42.5	12.50
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.50
50	40	0.10	45	50
50	45	0.15	47.5	12.50
50	50	0.25	50	0



$$\bar{X} = \frac{X_1 + X_2}{2}$$

Ejemplo

x	40	45	50
p(x)	0.2	0.3	0.5

x1	x2	p(x1,x2)	\bar{x}	s ²
40	40	0.04	40	0
40	45	0.06	42.5	12.50
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.50
50	40	0.10	45	50
50	45	0.15	47.5	12.50
50	50	0.25	50	0



$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^2 (x_i - \bar{X})^2}{2-1} = \sum_{i=1}^2 (x_i - \bar{X})^2$$

Ejemplo

x	40	45	50
p(x)	0.2	0.3	0.5

x1	x2	p(x1,x2)	\bar{x}	s ²
40	40	0.04	40	0
40	45	0.06	42.5	12.50
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.50
50	40	0.10	45	50
50	45	0.15	47.5	12.50
50	50	0.25	50	0

Ej: $p_{\bar{X}}(45) = P(\bar{X} = 45) = ?$

Ejemplo

Para obtener la distribución de probabilidad de la media muestral hay que calcular la probabilidad de cada valor



x1	x2	p(x1,x2)	\bar{x}	s ²
40	40	0.04	40	0
40	45	0.06	42.5	12.50
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.50
50	40	0.10	45	50
50	45	0.15	47.5	12.50
50	50	0.25	50	0

Ej: $p_{\bar{X}}(45) = P(\bar{X} = 45) = 0.10 + 0.09 + 0.10 = 0.29$

Ejemplo

x1	x2	p(x1,x2)	\bar{x}	s²
40	40	0.04	40	0
40	45	0.06	42.5	12.50
40	50	0.10	45	50
45	40	0.06	42.5	12.50
45	45	0.09	45	0
45	50	0.15	47.5	12.50
50	40	0.10	45	50
50	45	0.15	47.5	12.50
50	50	0.25	50	0

\bar{x}	40	42.5	45	47.5	50
$p_{\bar{x}}(\bar{x})$	0.04	0.12	0.29	0.30	0.25

Ejemplo

x	40	45	50
p(x)	0.2	0.3	0.5

$$\mu = 46.5$$

$$\sigma^2 = 15.25$$

Dos afinaciones se realizan el día seleccionado

\bar{x}	40	42.5	45	47.5	50
$p_{\bar{x}}(\bar{x})$	0.04	0.12	0.29	0.30	0.25

$$\mu_{\bar{x}} = 46.5 = \mu$$

$$\sigma_{\bar{x}}^2 = 7.625 = \frac{\sigma^2}{2}$$

Cuatro afinaciones se realizan el día seleccionado

\bar{x}	40	41.25	42.5	43.75	45	46.25	47.5	48.75	50
$p_{\bar{x}}(\bar{x})$	0.0016	0.0096	0.0376	0.0936	0.1761	0.2340	0.2350	0.1500	0.0625

$$\mu_{\bar{x}} = 46.5 = \mu$$

;

$$\sigma_{\bar{x}}^2 = 3.8125 = \frac{\sigma^2}{4}$$

La media se mantiene pero la varianza se reduce

Ejercicio

- Una marca de harina se vende al por mayor en bolsas de tres tamaños: 25, 40 y 65 kilos. 20% de los compradores elige la de 25 kg, 50% la de 40 kg y el 30% la de 65 kg. Sean X_1 y X_2 los pesos de las bolsas que eligen dos compradores seleccionados de manera independiente.
 - Determinar la distribución de muestreo de \bar{X} , calcular $E(\bar{X})$ y comparar con μ .
 - Determinar la distribución de muestreo de la varianza S^2 , calcular $E(S^2)$ y comparar con σ^2 .

Ejercicio

x	25	40	65
p(x)	0.2	0.5	0.3

$$\mu = 44.5$$

$$\sigma^2 = 212.25$$

x1	x2	p(x1,x2)	\bar{x}	s ²
25	25	0.04	25	0
25	40	0.10	32.5	56.25
25	65	0.06	45	400
40	25	0.10	32.5	56.25
40	40	0.25	40	0
40	65	0.15	52.5	156.25
65	25	0.06	45	400
65	40	0.15	52.5	156.25
65	65	0.09	65	0

$$\bar{x} = \frac{x_1 + x_2}{2}$$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2}{2}$$

Ejercicio

x	25	40	65
p(x)	0.2	0.5	0.3

$$\mu = 45.5$$

$$\sigma^2 = 212.25$$

x1	x2	p(x1,x2)	\bar{x}	s ²
25	25	0.04	25	0
25	40	0.10	32.5	56.25
25	65	0.06	45	400
40	25	0.10	32.5	56.25
40	40	0.25	40	0
40	65	0.15	52.5	156.25
65	25	0.06	45	400
65	40	0.15	52.5	156.25
65	65	0.09	65	0

\bar{x}	25	32.5	40	45	52.5	65
$p_{\bar{x}}(\bar{x})$	0.04	0.2	0.25	0.12	0.30	0.09

Ejercicio

x	25	40	65
p(x)	0.2	0.5	0.3

$$\mu = 45.5$$

$$\sigma^2 = 212.25$$

x1	x2	p(x1,x2)	\bar{x}	s ²
25	25	0.04	25	0
25	40	0.10	32.5	56.25
25	65	0.06	45	400
40	25	0.10	32.5	56.25
40	40	0.25	40	0
40	65	0.15	52.5	156.25
65	25	0.06	45	400
65	40	0.15	52.5	156.25
65	65	0.09	65	0

S ²	0	56.25	156.25	400
p _{S²} (S ²)	0.38	0.20	0.30	0.12

Ejemplo

□ Distribución original

x	25	40	65
p(x)	0.2	0.5	0.3

$$\mu = 44.5$$
$$\sigma^2 = 212.25$$

□ Muestras de tamaño 2

\bar{x}	25	32.5	40	45	52.5	65
$p_{\bar{x}}(\bar{x})$	0.04	0.2	0.25	0.12	0.30	0.09

S^2	0	56.25	156.25	400
$p_{S^2}(S^2)$	0.38	0.20	0.30	0.12

$$\mu_{\bar{x}} = 44.5 = \mu \quad ; \quad \sigma_{\bar{x}}^2 = 106.1250 = \frac{\sigma^2}{2}$$

Experimento de simulación

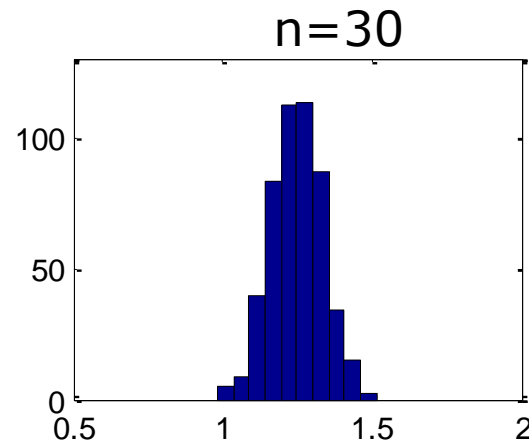
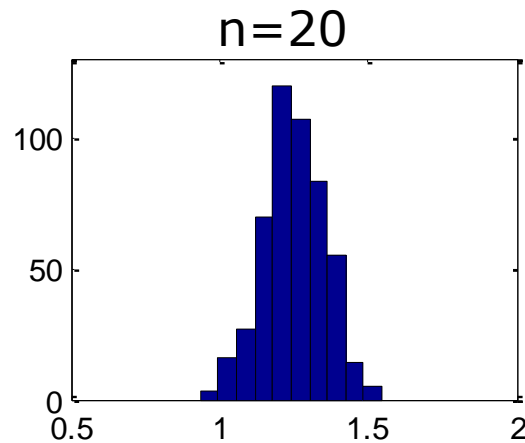
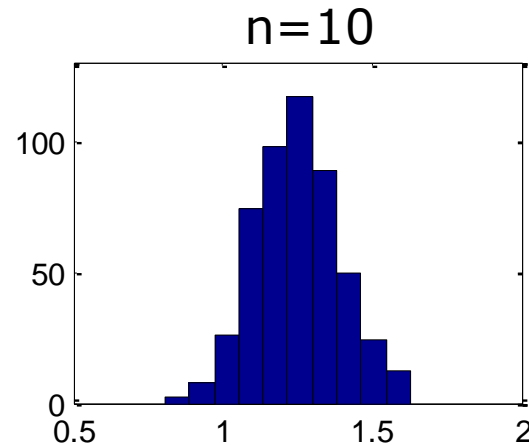
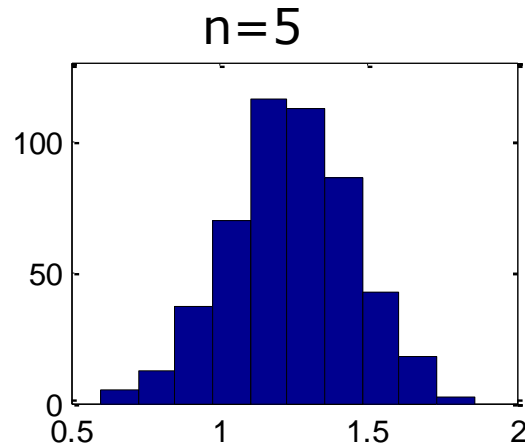
- El 2do.método para obtener información sobre la distribución de un estadístico es realizar un experimento de simulación. Debe indicarse
 - El estadístico de interés (ej: \bar{X})
 - La distribución poblacional (ej: normal con $\mu=100$ y $\sigma=15$).
 - El tamaño de la muestra n (ej: $n=10$)
 - El número de réplicas k ; es decir la cantidad de muestras a considerar (ej: $k=500$)

Experimento de Simulación

Estadístico : \bar{X}

Distribución : $N(1.25, 0.46^2)$

k = 500 (nro. de muestras)



Distribución de la media muestral

- Sea X_1, X_2, \dots, X_n una muestra aleatoria de una distribución con media μ y desviación estándar σ , entonces

$$1. E(\bar{X}) = \mu_{\bar{X}} = \mu$$

$$2. V(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \quad y \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Además, $T_o = X_1 + X_2 + \dots + X_n$

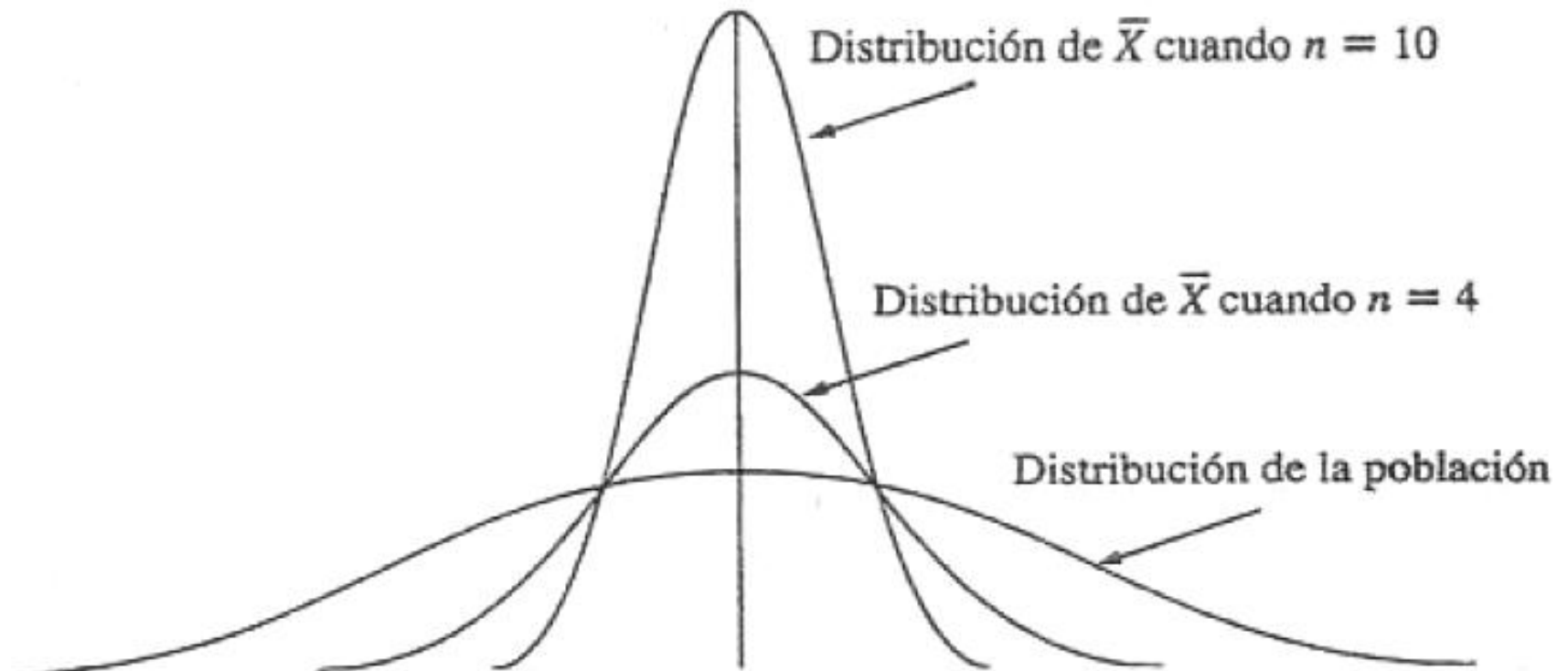
$$E(T_o) = n\mu$$

$$V(T_o) = n\sigma^2 \quad y \quad \sigma_{T_o} = \sqrt{n}\sigma$$

Caso de una distribución normal

- Sea X_1, X_2, \dots, X_n una v.a. de una distribución normal con media μ y desviación estándar σ , entonces para **cualquier n** ,
 - \bar{X} tiene una distribución normal con media μ y desviación estándar σ/\sqrt{n}
 - T_o también tiene una distribución normal pero con media $n\mu$ y desviación estándar $\sqrt{n}\sigma$

Caso de una distribución normal



Ejemplo

- El tiempo que tarda una rata de cierta especie en encontrar su camino por un laberinto es una v.a. con distrib.normal con $\mu=1.5$ min y $\sigma=0.35$ min.
- Se eligen 5 ratas. Sean X_1, X_2, \dots, X_5 sus tiempos en el laberinto.
- ¿Cuál es la probabilidad de que el tiempo total $T_0 = X_1 + X_2 + \dots + X_5$ para las 5 ratas esté entre 6 y 8 minutos?

Ejemplo

- Si X_1, X_2, \dots, X_5 tienen distribución normal entonces T_o también. Sus parámetros son

$$\mu_{T_o} = n\mu = 5(1.5) = 7.5$$

$$\sigma_{T_o}^2 = n\sigma^2 = 5(0.1225) = 0.6125 \quad \therefore \quad \sigma_{T_o} = 0.783$$

- Luego

$$\begin{aligned} P(6 \leq T_o \leq 8) &= P\left(\frac{6-7.5}{0.783} \leq Z \leq \frac{8-7.5}{0.783}\right) \\ &= P(-1.92 \leq Z \leq 0.64) = \phi(0.64) - \phi(-1.92) = 0.7115 \end{aligned}$$

Teorema Central del Límite

- Sea X_1, X_2, \dots, X_n una muestra aleatoria de una distribución con media μ y varianza σ^2 , entonces ***si n es suficientemente grande***
 - \bar{X} tiene una distribución normal con media μ y varianza σ^2/n
 - T_o también tiene una distribución normal pero con media $n\mu$ y varianza $n\sigma^2$

Regla empírica

Si **$n > 30$** se puede usar el Teorema Central del Límite

Resumen

\bar{X}	$\frac{1}{n} \sum_{i=1}^n X_i$	$E(\bar{X}) = \mu$ $V(\bar{X}) = \frac{\sigma^2}{n}$
T_o	$\sum_{i=1}^n X_i$	$E(T_o) = n\mu$ $V(T_o) = n\sigma^2$

Ejemplo

- Cuando se prepara un lote de cierto producto, la cantidad de determinada impureza en el lote es una v.a. con valor medio 4 g y una desviación estándar de 1.5 g.
- Si 50 lotes se preparan de forma independiente ¿cuál es la probabilidad (aproximada) de que la cantidad promedio muestral de impureza \bar{X} esté entre 3.5 y 3.8 g?
- Según el TCL, la distribución de \bar{X} se aproxima a una normal con
$$\mu_{\bar{X}} = 4 ; \sigma_{\bar{X}} = 1.5 / \sqrt{50} = 0.2121$$

Ejemplo

- Si $\bar{X} \approx N(4, (0.2121)^2)$ la probabilidad (aproximada) de que la cantidad promedio muestral de impureza esté entre 3.5 y 3.8 g es

$$\begin{aligned} P(3.5 \leq \bar{X} \leq 3.8) &\approx P\left(\frac{3.5-4}{0.2121} \leq Z \leq \frac{3.5-4}{0.2121}\right) \\ &= \phi(-0.94) - \phi(-2.36) = 0.1645 \end{aligned}$$

Ejercicio

- La densidad del sedimento de cierto líquido (g/cm^3) tiene una distribución normal con media 2.65 y desviación estándar 0.85.
- Si se dispone de una muestra aleatoria formada por 6 observaciones de dicho líquido
 - ¿Cuál es la probabilidad de que la densidad promedio muestral sea a lo sumo 3?
 - y de que esté entre 2.65 y 3.00?

Ejercicio

$$X \approx N(\mu, \sigma^2)$$

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8
x_1	3,3793	3,2365	1,6279	3,8529	2,5997	2,1029	2,3798	3,1412
x_2	3,7159	3,3433	2,6332	1,9657	1,7910	2,9733	3,5808	2,6843
x_3	1,2953	3,2551	2,5168	3,0994	3,1723	1,7923	1,0571	3,2255
x_4	1,4252	3,7467	1,2865	2,8364	3,0816	2,6334	3,0140	3,1336
x_5	3,1355	3,2183	2,8687	1,8664	4,0886	2,6090	3,4113	2,4327
x_6	2,3101	3,6622	1,7520	0,8049	3,1526	2,6500	3,2713	2,3292
\bar{x}	2,5435	3,4104	2,1142	2,4043	2,9809	2,4602	2,7857	2,8244

$$P(\bar{X} < 3) = ?$$

$$\bar{X} \approx N(\mu, \sigma^2/n)$$

Ejercicio

$$X \approx N(\mu, \sigma^2) \quad \bar{X} \approx N(\mu, \sigma^2/n)$$

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

$$\begin{aligned} P(\bar{X} < 3) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{3 - 2.65}{0.85/\sqrt{6}}\right) \\ &= P(Z < 1.0086) = 0.8435 \end{aligned}$$

El 84,35% de las muestras tendrán un valor de densidad del sedimento promedio inferior a 3.

Verifique este porcentaje con las muestras de la transparencia anterior

$$X \approx N(\mu, \sigma^2) \quad \bar{X} \approx N(\mu, \sigma^2/n)$$

Ejercicio

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

- Qué tan grande debería ser el tamaño de la muestra para asegurar que la probabilidad de que la densidad promedio muestral sea a lo sumo 3 (la que calculamos antes) sea por lo menos 0.99?
- Es decir, cuánto debe valer ***n*** para que

$$P(\bar{X} < 3) \geq 0.99$$

Ejercicio

$$X \approx N(\mu, \sigma^2) \quad \bar{X} \approx N(\mu, \sigma^2/n)$$

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

$$P(\bar{X} < 3) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{3 - 2.65}{0.85/\sqrt{n}}\right) = 0.99$$

$$= P\left(Z < \frac{3 - 2.65}{0.85/\sqrt{n}}\right) = 0.99$$



2.33

Ejercicio

$$X \approx N(\mu, \sigma^2) \quad \bar{X} \approx N(\mu, \sigma^2/n)$$

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

$$P(\bar{X} < 3) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{3 - 2.65}{0.85/\sqrt{n}}\right) = 0.99$$

$$= P\left(Z < \frac{3 - 2.65}{0.85/\sqrt{n}}\right) = 0.99$$

$$\frac{(3 - 2.65)}{0.85/\sqrt{n}} = 2.33$$

Falta despejar
n

Ejercicio

$$X \approx N(\mu, \sigma^2)$$

$$\bar{X} \approx N(\mu, \sigma^2/n)$$

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

$$\frac{(3 - 2.65)}{0.85/\sqrt{n}} = \frac{\sqrt{n} * (3 - 2.65)}{0.85} = 2.33$$

$$\sqrt{n} = \frac{2.33 * 0.85}{(3 - 2.65)} \quad \Rightarrow \quad \sqrt{n} = 5.6586$$

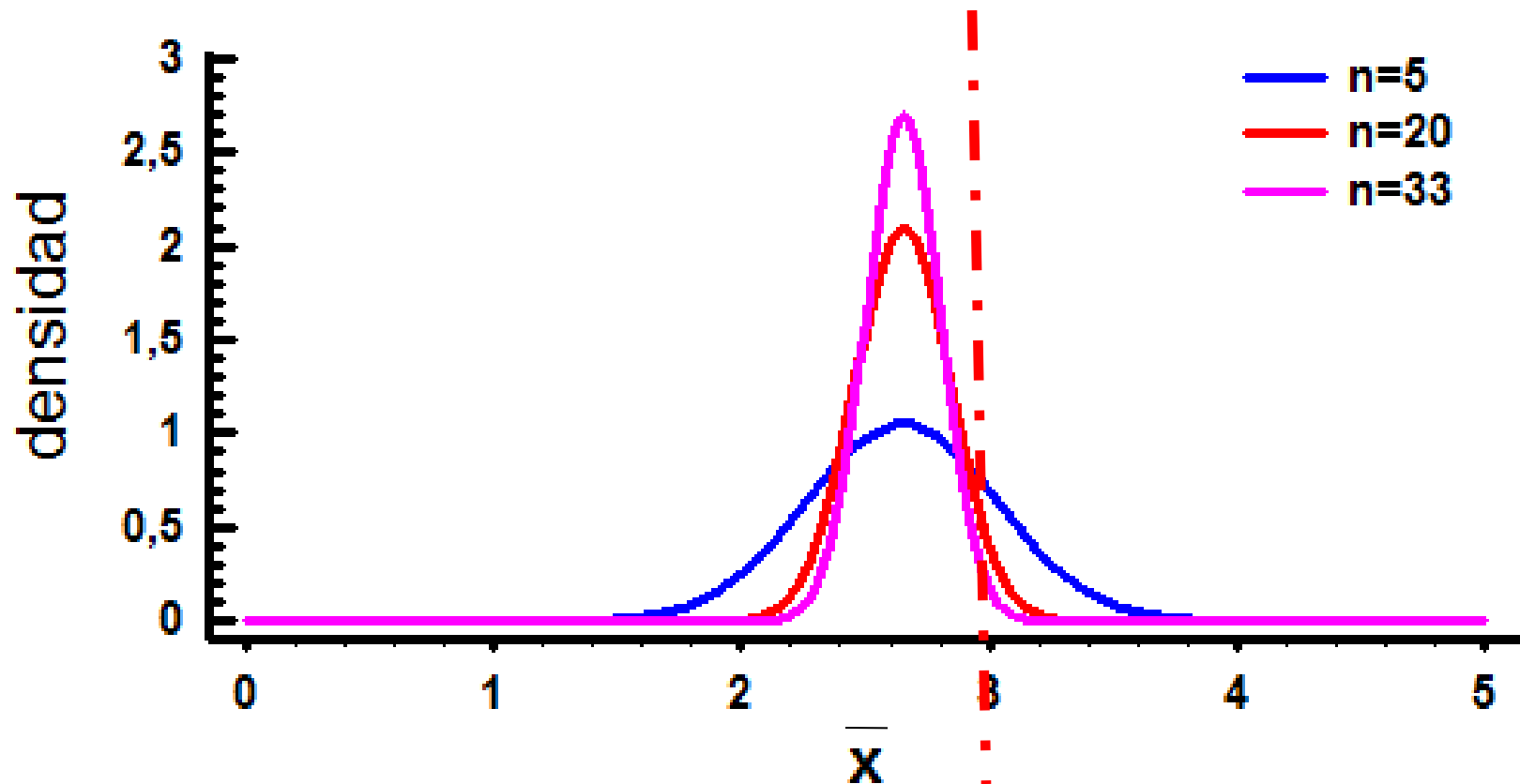
$$n = 5.6586^2 \quad \Rightarrow \quad n = 32.0198$$

El tamaño mínimo de la muestra debería ser **33**

$$\bar{X} \approx N(\mu, \sigma^2/n)$$

Ejercicio

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$



Ejercicio 2

- La densidad del sedimento de cierto líquido (g/cm^3) tiene una distribución normal con media 2.65 y desviación estándar 0.85.
- Se realizan 6 observaciones de dicho líquido. Sean X_1, X_2, \dots, X_6 sus densidades de sedimento.
- ¿Cuál es la probabilidad de que la densidad total $T_o = X_1 + X_2 + \dots + X_6$ para las 6 observaciones esté entre 14 y 16 g/cm^3 ?

$$X \approx N(\mu, \sigma^2) \quad T_0 \approx N(n\mu, n\sigma^2)$$

Ejercicio 2

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$

- Si X_1, X_2, \dots, X_6 tienen distribución normal entonces T_0 también. Sus parámetros son

$$\mu_{T_0} = n\mu = 6(2.65) = 15.90$$

$$\sigma_{T_0}^2 = n\sigma^2 = 6(0.85^2) = 4.3350 \quad \therefore \quad \sigma_{T_0} = 2.0821$$

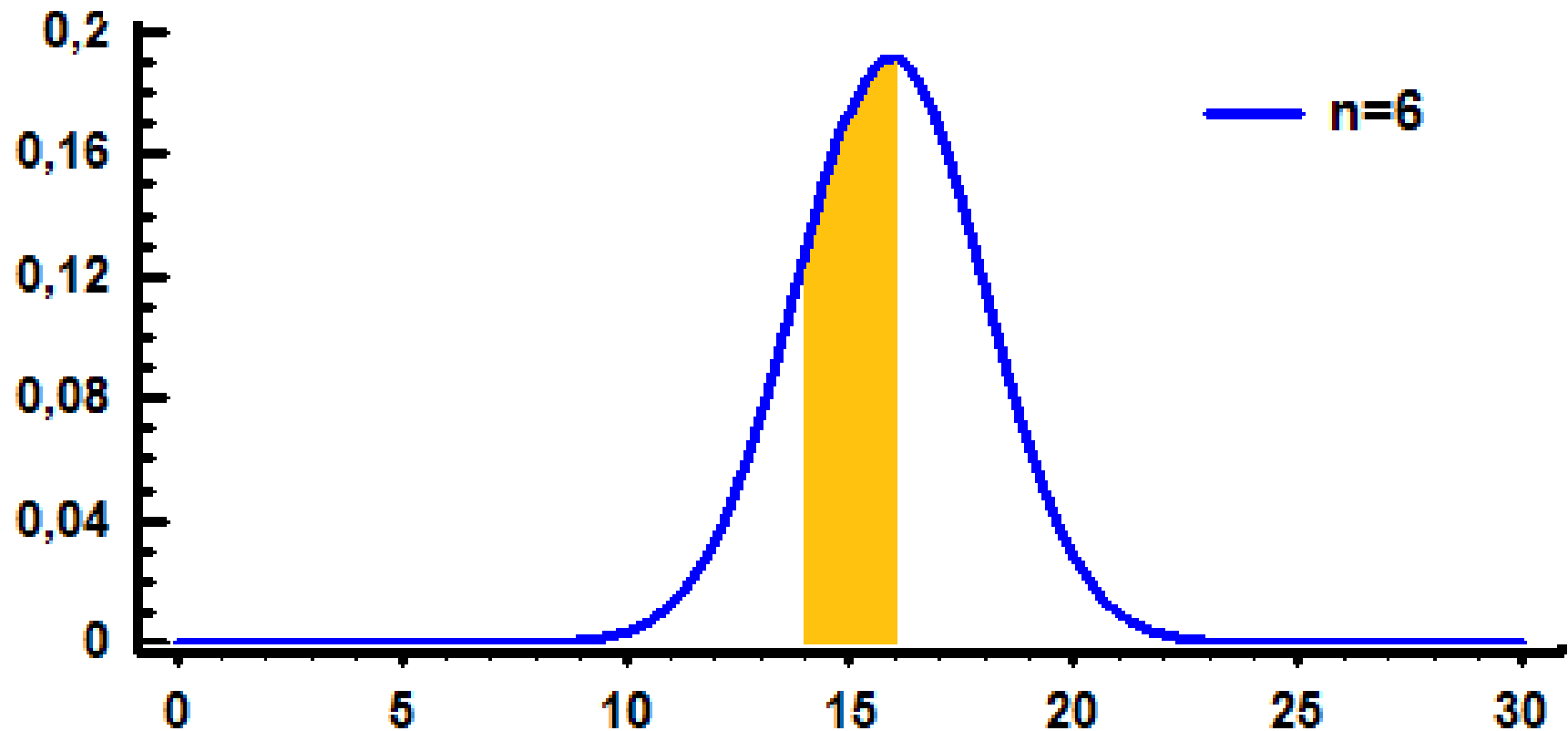
- Luego

$$\begin{aligned} P(14 \leq T_0 \leq 16) &= P\left(\frac{14 - 15.9}{2.0821} \leq Z \leq \frac{16 - 15.9}{2.0821}\right) \\ &= P(-0.9125 \leq Z \leq 0.048) = \phi(0.048) - \phi(-0.9125) = 0.3384 \end{aligned}$$

$$T_0 \approx N(n\mu, n\sigma^2)$$

Ejercicio 2

$$\mu = 2.65 \quad ; \quad \sigma = 0.85$$



Aprox.Normal a la Distrib.Binomial

- El TCL se puede utilizar para aproximar las probabilidades de algunas v.a. discretas cuando es difícil calcularlas exactamente para valores grandes de los parámetros.
- Si $X \sim B(n,p)$ hay dos formas de calcular $P(X \leq k)$

- $$P(X \leq k) = \sum_{i=0}^k P(X = i)$$

- Usando las tablas de *fda*; pero no existen para valores grandes de n lo que nos obliga a hacer la suma anterior.

Aprox.Normal a la Distrib.Binomial

- Como una opción podemos considerar a X como suma de v.a. más simples, específicamente, si definimos

$$X_i = \begin{cases} 1 & \text{si en la } i\text{-ésima repetición de } \varepsilon \text{ ocurre éxito} \\ 0 & \text{en caso contrario} \end{cases} \quad i=1,2,\dots, n$$

entonces cada $X_i \sim B(1,p)$ y además X_1, X_2, \dots, X_n son independientes

Aprox.Normal a la Distrib.Binomial

- Si X_1, X_2, \dots, X_n tienen distribución $B(1,p)$, por el TCL, T_o tiene distribución normal con media **np** y varianza **$np(1-p)$** .
- El tamaño de la muestra necesario para que la aproximación funcione depende de p .
- Note que la distribución de cada X_i es simétrica cuando p es cercana a 0.5 y sesgada cuando está cerca de 0 o 1.
- ***Se recomienda usar la aproximación cuando $np \geq 10$ y $n(1-p) \geq 10$***

Corrección por continuidad

- Según el TCL, si $X \sim B(n, p)$ para n suficientemente grande puede usarse

$$X \sim N(np, np(1-p))$$

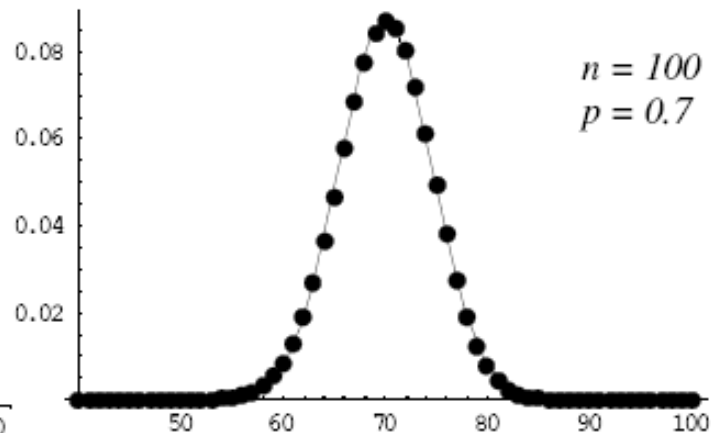
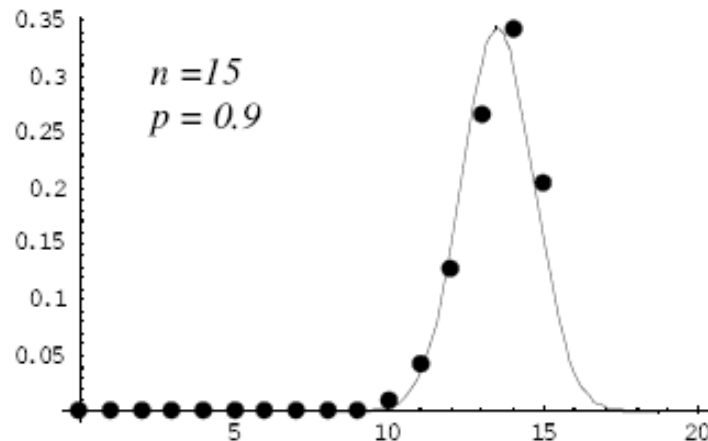
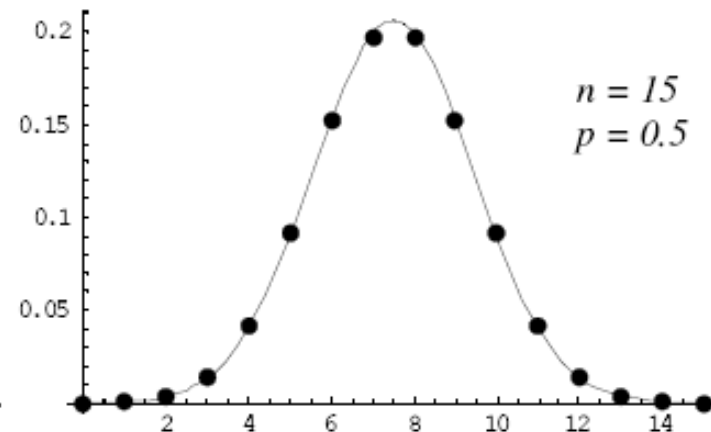
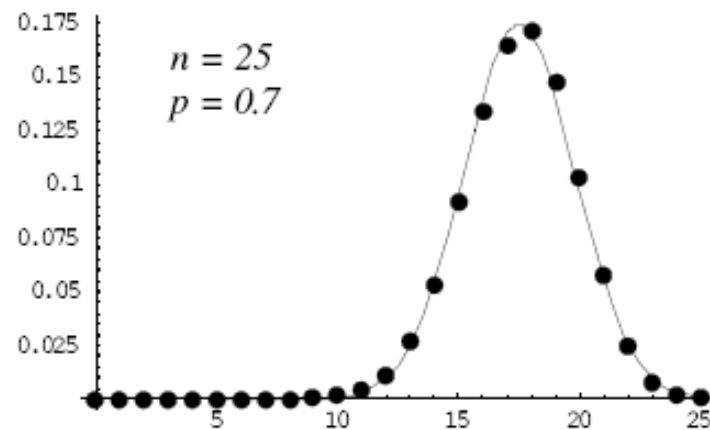
- Dado que la binomial es discreta y la normal continua, deben hacerse algunas correcciones

$$P(X = k) \cong P\left(k - \frac{1}{2} \leq X \leq k + \frac{1}{2}\right)$$

$$P(X \leq k) \cong P\left(X \leq k + \frac{1}{2}\right)$$

$$P(X \geq k) \cong P\left(X \geq k - \frac{1}{2}\right)$$

$B(n,p)$ aprox. por $N(np, np(1-p))$



Combinación lineal de v.a.

- Dadas n v.a. X_1, X_2, \dots, X_n y n constantes numéricas a_1, a_2, \dots, a_n la v.a.

$$Y = a_1X_1 + \dots + a_nX_n = \sum_{i=1}^n a_iX_i$$

se llama **combinación lineal** de las X_i

- Si $a_1 = \dots = a_n = 1$, $Y = T_o$ y si $a_1 = \dots = a_n = 1/n$, $Y = \bar{X}$.
- Note que las X_i podrían tener distribuciones distintas y por lo tanto, medias y varianzas distintas. Tampoco tienen que ser independientes.

Distrib.de una combinación lineal

- Si X_1, X_2, \dots, X_n tienen valores medios $\mu_1, \mu_2, \dots, \mu_n$ respectivamente y varianzas $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectivamente :

1.- Si las X_i son independientes o no

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) \\ &= a_1\mu_1 + \dots + a_n\mu_n \end{aligned}$$

Distrib.de una combinación lineal

- Si X_1, X_2, \dots, X_n tienen valores medios $\mu_1, \mu_2, \dots, \mu_n$ respectivamente y varianzas $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ respectivamente :

2.- Si las X_i son independientes

$$\begin{aligned} V(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n) \\ &= a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2 \end{aligned}$$

3.- Para cualquier X_1, X_2, \dots, X_n tienen

$$V(a_1X_1 + a_2X_2 + \dots + a_nX_n) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j)$$

Ejemplo

- Una estación de servicio vende tres tipos de nafta: común, super y super premium. Estas se venden a 2.4\$, 3.1\$, 3.5\$ por litro.
- Sean X_1 , X_2 y X_3 las cantidades de estos tipos de naftas vendidas en un día en particular.
- Suponga que las X_i son independientes con $\mu_1=1000$, $\mu_2=500$, $\mu_3=300$, $\sigma_1=100$, $\sigma_2=80$, $\sigma_3=50$.
- El ingreso obtenido por estas ventas es
$$Y=2.4X_1+3.1X_2+3.5X_3$$

Ejemplo

□ Si $Y = 2.4X_1 + 3.1X_2 + 3.5X_3$ entonces

$$E(Y) = 2.4\mu_1 + 3.1\mu_2 + 3.5\mu_3 = 5000\$$$

$$V(Y) = (2.4)^2 \sigma_1^2 + (3.1)^2 \sigma_2^2 + (3.5)^2 \sigma_3^2 = 31689$$

$$\sigma_Y = \sqrt{149729} = 386.95\$$$

Diferencia entre dos v.a.

- Un caso especial de la combinación lineal resulta de tomar $n=2$, $a_1=1$ y $a_2=-1$

$$Y = a_1 X_1 + a_2 X_2 = X_1 - X_2$$

- Aplicando lo anterior se obtiene

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$

y si son independientes

$$V(X_1 - X_2) = V(X_1) + V(X_2)$$

Ejemplo

- Sean X_1 y X_2 los rendimientos de combustibles para autos de 6 y 4 cilindros, respectivamente, seleccionados de manera independiente y aleatoria; con $\mu_1=22$, $\sigma_1=1.2$, $\mu_2=26$ y $\sigma_2=1.5$

$$E(X_1 - X_2) = \mu_1 - \mu_2 = 22 - 26 = -4$$

$$V(X_1 - X_2) = \sigma_1^2 + \sigma_2^2 = (1.2)^2 + (1.5)^2 = 3.69$$

$$\sigma_{X_1 - X_2} = \sqrt{3.69} = 1.92$$

Note que si hubieramos utilizado X_1 para referirnos a los autos de 4 cilindros, $E(X_1 - X_2) = 4$ pero la varianza hubiera sido la misma.

Resumen

- **Muestra aleatoria**

- **Estadístico**

 - Distribución

 - Cálculos

 - Experimento de simulación

- **Teorema Central del Límite**

- **Aprox.Normal**

 - Binomial

- **Combinación lineal de v.a.**

 - Distribución

 - Diferencia