Simulación normal

Veremos en un tema posterior en qué consiste la distribución normal, pero nada nos impide ir practicando con ella. Por ahora basta con saber que siguen esa distribución de forma aproximada muchos datos tomados de nuestra vida diaria:

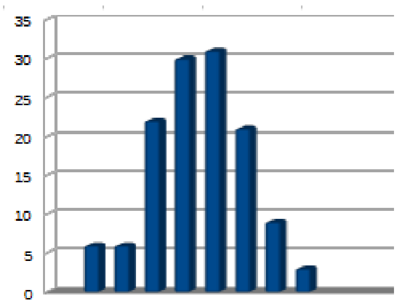
- Magnitudes que dependen de muchas causas independientes, cuyos efectos se suman y cualquiera de ellas aislada tenga efectos despreciables.

- Distribuciones de errores en las medidas.

- Medidas de tipo antropológico (estaturas, pesos, inteligencia...) y biológico (glucemia, nivel de colesterol...)

- Límite de otras distribuciones estadísticas cuando n aumenta.

Todas ellas producen gráficos con forma de campana de Gauss, más o menos aproximada.



Nuestro simulador puede producir datos aleatorios que sigan esta distribución normal.

Puedes descargarlo para Excel

http://www.hojamat.es/estadistica/tema1/open/simulador.xlsm

Y para LibreOffice Calc

http://www.hojamat.es/estadistica/tema1/open/simulador.ods

La forma más práctica de plantear una simulación de este tipo es la de dar el promedio de los datos y la desviación típica, pero también funciona conociendo el mínimo y el máximo esperados.

Lo vemos con algún ejemplo:

Los de más altura

En un Centro de Enseñanza se han tallado todos los alumnos y alumnas de un nivel, 128 en total y ha quedado como estatura mínima la de 140 cm, y como máxima, 198 cm. Si deseamos seleccionar a aquellas personas con estatura superior a 180 cm. ¿Cuántas esperaremos encontrar?

En el tema 6 aprenderemos a responder a esta pregunta mediante las propiedades de la distribución normal. Aquí lo intentaremos con el Simulador:

Simulador	A. Roldán – Versión 2.2 – Año 2016	
	Repeticiones de la simulación	1
	Número de filas (de 1 a 1000)	128
	Número de columnas (si procede) (de 1 a 12)	1
Tipo de simulación	**Extremos (si procede)**	
Normal		
	Mínimo	140
	Máximo	198
Decimal – Entero	**Parámetros**	
Decimal	**Media**	10
	Sigma	2
Criterios	**Otros parámetros**	
Máximo-Mínimo		
	A	0,429
	B	0,71429
	A	0
Número intervalos	B	
10		

Hemos concretado lo siguiente:

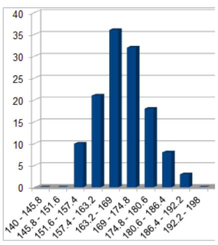
- Distribución normal con decimales (son estaturas) usando máximo y mínimo
- Mínimo 140 y máximo 198
- Una columna de 128 filas (número de alumnos y alumnas)
- Diez intervalos

Con ello la simulación se aproximará bastante a las medidas reales. Si pasas a la segunda hoja advertirás la forma típica de campana de esta distribución, y que la estatura media es aproximadamente de 169 cm., y la desviación típica cercana a 8. No podemos pretender resultados idénticos a los previstos por la teoría.

Estadísticos de momentos	
Media	169,3337
Desviación típica	8,0704
Asimetría	0,2665
Curtosis	-0,2509

Tabla de frecuencias	
140 - 145.8	0
145.8 - 151	0
151.6 - 157	10
157.4 - 163	21
163.2 - 169	36
169 - 174.8	32
174.8 - 180	18
180.6 - 186	8
186.4 - 192	3
192.2 - 198	0

Gráfico



También, de paso, hemos descubierto que esperaremos unas 11 personas con más de 180 cm. Afinamos esto más. Procede a repetir la simulación, y a la derecha del primer dato (celda G5) escribe la fórmula =SI(G5>=180;1;0), que escribe un 1 si el dato pasa de 180 y un cero si no llega a esa estatura. Rellena después esa fórmula hacia abajo hasta llegar al dato 128.

166,194	0
171,671	0
159,06	0
159,17	0
171,995	0
176,134	0
155,125	0
170,837	0
184,078	1
170,384	0
153,378	0

Después basta sumar la nueva columna y nos dará el número de datos superior a 180. En nuestro ejemplo han resultado, en varias simulaciones, 15, 11, 8, 11 y 15, por lo que juzgamos que lo más probable es que nos encontremos con unos 11, lo que nos permitirá organizar un equipo de baloncesto, si ese era el objetivo.

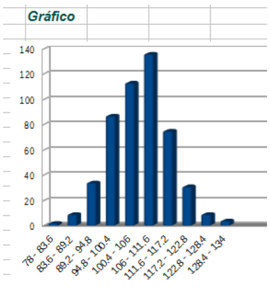
Un ejemplo con media y desviación típica

Una población de 500 personas con riesgo de diabetes en una ciudad ha presentado un promedio de 106 mg/100ml de nivel de glucosa en sangre y una desviación típica de 8 mg/100ml. Diseñar una simulación para encontrar a partir de qué nivel encontraremos las 50 personas con más riesgo.

Organizamos la simulación, pero usando ahora media y desviación típica:

Simulador	A. Roldán – Versión 2.2 – Año 2016	
	Repeticiones de la simulación	1
	Número de filas (de 1 a 1000)	500
	Número de columnas (si procede) (de 1 a 12)	1
Tipo de simulación	**Extremos (si procede)**	
Normal		
	Mínimo	**140**
	Máximo	**198**
	Parámetros	
Decimal – Entero		
Decimal	**Media**	**106**
	Sigma	**8**
Criterios		
Media-Desviación	**Otros parámetros**	

Obtendremos una columna con 500 niveles de glucosa y una distribución en forma de campana de Gauss.



En nuestra simulación se obtuvieron media y desviación bastante cercanas a las teóricas:

Media	**105,8961**
Desviación típica	**8,3871**
Asimetría	**0,0679**
Curtosis	**0,0127**

Si ahora deseamos obtener los cincuenta niveles más altos, nos bastará con ordenar la columna G de la primera hoja (de mayor a menor) y observar n qué nivel se encuentra el número 50:

116,74	48
116,71	49
116,4	50
116,31	51
116,26	

Vemos que hay que comenzar por el nivel 116,4 para así poder seleccionar los 50 posibles pacientes con más riesgo. Si repites la simulación varias veces podrás quedarte con una media más aproximada.

Medidas válidas

En una medición con mucho riesgo de errores se ha decidido rechazar aquellas medidas que se alejen de la media más de una desviación típica y media. Supongamos que en mediciones anteriores resultó una media de 65 y una desviación típica de 8. ¿Qué número aproximado de mediciones debemos efectuar para garantizarnos 200 medidas catalogadas como válidas, si la distribución en la población se puede considerar normal?

De nuevo acudimos a una simulación, ya que es prematuro acudir a la teoría. Según los datos que nos dan, las medidas válidas estarán entre 65-3*8/2 y 65+3*8/2, es decir, entre 53 y 77. Comenzamos una simulación de 250 mediciones y concretamos 14 intervalos.

Observamos, de forma aproximada, que habría que desechar unas 11 medidas inferiores y unas 13 superiores, lo que nos daría 250-11-13=226 medidas válidas.

Tabla de frecuencias	
37 - 41	0
41 - 45	1
45 - 49	3
49 - 53	7
53 - 57	20
57 - 61	40
61 - 65	50
65 - 69	54
69 - 73	44
73 - 77	18
77 - 81	9
81 - 85	2
85 - 89	2
89 - 93	0

Probamos con 230 simulaciones, pare ver si nos acercamos a 200 válidas. Después de la simulación hay que desechar 16+18=34, con lo que nos quedamos cortos, 196. Subimos y bajamos el número de simulaciones y 230 parece quedar en la media, luego es aconsejable usar muestras de 230 medidas.

Esto ha sido una especie de juego. Si acudimos a la distribución normal teórica, que estudiaremos en otro tema, descubriremos que el porcentaje esperado de medidas que se alejen más de 3/2 de desviación típica por un lado es de 0,066807201. Por los dos lados será 0,133614403, y restando de 1, el porcentaje de medidas válidas sería 0,866385597. Dividimos 200 entre ese porcentaje y obtenemos 230,844096. Nuestra simulación no estaba descaminada.

Con este experimento también hemos aprendido que los porcentajes no dependen de una media concreta sino de la medida tipificada Z, que en este caso valía Z=1,5. Por eso viene bien esta simulación para la teoría que desarrollamos en este tema.