Simulaciones - Distribución uniforme

Iniciamos hoy una serie, que nos tomaremos con calma, sobre simulaciones elementales de variables aleatorias. Nos basaremos en las prácticas de nuestro curso de Estadística, (http://www.hojamat.es/estadistica/iniestad.htm) adaptándolas al formato de un blog. Usaremos nuestro Simulador implementado para hojas de cálculo, el cual puede sufrir cambios a lo largo de la serie, por lo que se aconseja su recarga en caso de duda.

Comenzaremos con la distribución uniforme. Si no tienes claro el concepto puedes acudir a la Teoría correspondiente (http://www.hojamat.es/estadistica/tema6/teoria/teoria6.pdf). Por ahora te basta con la idea de que representa experimentos aleatorios en los que todos los elementos presentan una misma probabilidad de ocurrir, como tiradas de dados o las loterías. Se suele distinguir entre distribución uniforme discreta, cuando sólo existe un número finito de posibilidades (dados o monedas), o continua, cuando pueden aparecer infinitos sucesos, o al menos, tantos, que sea preferible tratarlos como infinitos.

Distribución uniforme discreta

En ella se trabaja sobre un conjunto finito de **n** elementos con la hipótesis de que todos ellos poseen la misma probabilidad de aparecer, que será, por tanto, **1/n**. Un ejemplo es el de una tirada de dados. La distribución uniforme más útil es aquella en la que el conjunto está formado por los números comprendidos entre **a** y **b** ambos inclusive. En los dados a=1 y b=6.

Puedes consultar en cualquier manual los valores de los principales estadísticos de esta distribución.

Media:

$$m = \frac{a+b}{2}$$

Varianza:

$$var = \frac{(b-a+1)^2-1}{12}$$

En el caso de las tiradas de dados m=3,5 y var=35/12=2,92 y su desviación típica 1,7321.

Podemos comprobar estos valores mediante nuestro simulador, alojado en estas direcciones:

http://www.hojamat.es/estadistica/tema1/open/simulador.ods (versión LibreOffice Calc)

http://www.hojamat.es/estadistica/tema1/open/simulador.xlsm (versión Excel)

Contiene dos hojas, la de *criterios y simulación* y la de los *estadísticos*. La mejor forma de aprender su funcionamiento es proceder a la primera simulación.

Deseamos saber si con 1000 tiradas de dados su media y varianza se acercan suficientemente a la teoría. Para ello, en la primera página del simulador concretamos lo siguiente:

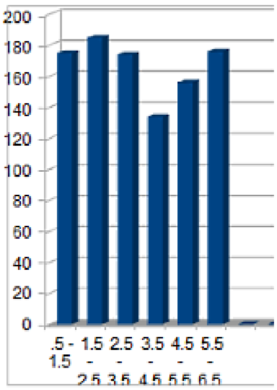
Simulador	A. Roldán - Versión 2.2 - Año 2016		
	Repeticiones de la simulación		5
	Número de filas		200
	(de 1 a 1000)		
	Número de columnas (si procede)		1
	(de 1 a 12)		
Tipo de simulación		Extremos (si procede)	
Uniforme			
		Mínimo	1
		Máximo	6
		Parámetros	
Decimal – Entero			
Entero		Media	4
		Sigma	2
Criterios			
Máximo-Mínimo		Otros parámetros	
		A	32,000
		B	0,5
		C	1,2
Número intervalos		D	
	6		

Cinco repeticiones de 200 filas y una columna, para que se acumulen 1000 tiradas, mínimo=1 y máximo=6. Como criterios, "Uniforme", "Entero", para que la distribución sea discreta, y "Máximo-Mínimo". También es conveniente fijar, unas celdas más abajo, el número de intervalos en 6. El resto de parámetros se puede ignorar. Con ello, al dar al botón **Simulador** obtendrás los resultados en la siguiente hoja de Estadísticos. En nuestro caso serían:

Media	3,4390
Desviación típica	1,7419
Asimetría	0,0886
Curtosis	-1,3161
Valores teóricos	
Media	3,5000
Desviación típica	1,7078
Asimetría	0,0000
Kurtosis	-1,2686

Se ha obtenido una aproximación apreciable. La herramienta también nos proporciona la asimetría y la kurtosis, pero prescindimos de ellas en esta simulación. A la derecha puedes observar la tabla de frecuencias y el diagrama de barras, que presenta una uniformidad de alturas bastante aceptable para el número de simulaciones que hemos fijado:

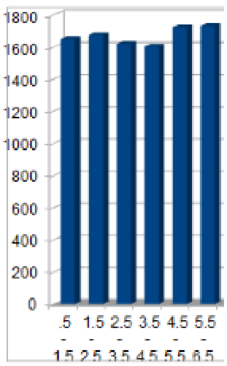
.5 - 1.5	175
1.5 - 2.5	185
2.5 - 3.5	174
3.5 - 4.5	134
4.5 - 5.5	156
5.5 - 6.5	176



Como ya sabrás, es probable que, si aumentamos el número de repeticiones, el ajuste mejore, pero no lo des por seguro, que sólo existe una probabilidad. Si aumentamos a 50 repeticiones obtenemos:

Media	3,5270
Desviación típica	1,7196
Asimetría	-0,0189
Curtosis	-1,2910
Valores teóricos	
Media	3,5000
Desviación típica	1,7078
Asimetría	0,0000
Kurtosis	-1,2686

.5 - 1.5	1650
1.5 - 2.5	1673
2.5 - 3.5	1619
3.5 - 4.5	1604
4.5 - 5.5	1723
5.5 - 6.5	1731

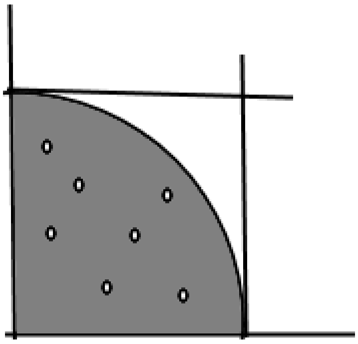


Ha mejorado bastante el ajuste. En general, las simulaciones comienzan a ser útiles si las repites miles de veces. Con unas pocas no son útiles.

Distribución uniforme continua

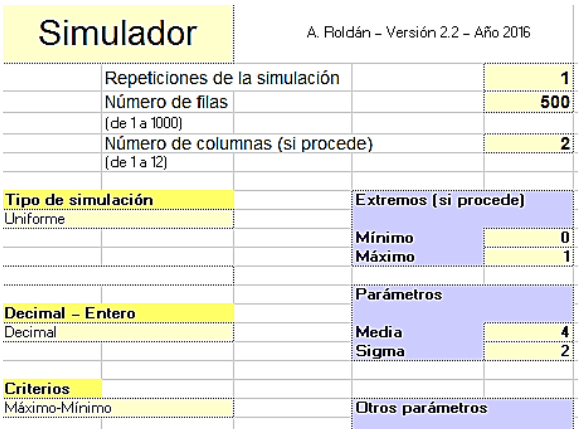
En esta modalidad los datos se distribuyen de forma continua (en la práctica, con todos los decimales que deseemos) entre dos extremos **a** y **b**. Prácticamente no hay ejemplos en la vida diaria de distribuciones uniformes, ya que son más frecuentes otras, como la normal. Suelen aparecer en experimentos diseñados o en instrumentos creados por nosotros, como puede ser el movimiento de las manecillas de un reloj, que recorre de manera uniforme toda la circunferencia.

Un ejemplo de distribución uniforme continua es el experimento de calcular π mediante simulación (método de Montecarlo). Consiste en simular dos coordenadas X e Y de manera uniforme entre 0 y 1 y contar aquellos pares en los que X^2+Y^2<1. De esa forma, su frecuencia relativa deberá ser π/4=0,7854 aproximadamente. En la imagen sería como contar todos los puntos que caen dentro de la zona sombreada.



Lo organizaremos así:

Planteamos los criterios contenidos en la imagen:



Simulador		A. Roldán – Versión 2.2 – Año 2016		
	Repeticiones de la simulación			1
	Número de filas			500
	(de 1 a 1000)			
	Número de columnas (si procede)			2
	(de 1 a 12)			
Tipo de simulación			**Extremos (si procede)**	
Uniforme				
			Mínimo	0
			Máximo	1
			Parámetros	
Decimal – Entero				
Decimal			**Media**	4
			Sigma	2
Criterios				
Máximo-Mínimo			**Otros parámetros**	

Tomamos 500 filas y dos columnas (que representarán X e Y). No planteamos repeticiones porque añadiremos una columna nueva a la simulación. Concretamos un mínimo de 0 y un máximo de 1. En los restantes criterios elegimos “Uniforme”, “Decimal” (por ser continua) y “Máximo-Mínimo”.

Con ello obtenemos dos columnas de 500 valores de X e Y. Ahora, en una columna paralela, por ejemplo comenzando en J5, escribimos =SI(G5^2+H5^2<1;1;0). Esto significa que obtendremos un 1 si el punto (X,Y) pertenece al sector circular sombreado, y 0 si está fuera. Extendemos esa fórmula hacia abajo hasta abarcar los 500 valores:

0,135988891	0,129352689	1
0,883682072	0,886273146	0
0,409658372	0,573923469	1
0,815548956	0,356251717	1
0,734143436	0,334889054	1
0,798231423	0,142625093	1
0,375023305	0,146670222	1
0,097308695	0,979814053	1
0,332297981	0,274846196	1
0,712780774	0,398977041	1
0,605967462	0,719416976	1
0,644647658	0,868955612	0
0,196031749	0,214803338	1
0,627330124	0,593012094	1
0,774594724	0,229846835	1
0,864090502	0,430201054	1
0,7318694	0,358022809	1
0,479562581	0,849364042	1
0,005538881	0,802593589	1
0,411429465	0,319342613	1

Ahora basta sumar esa columna nueva y deberemos obtener un valor próximo a 500π/4≈393. Esto no se suele obtener en una simulación con tan pocos datos. En nuestro caso se ha obtenido 389. Podemos repetir el trabajo (de forma manual) varias veces y encontrar la media de resultados. Aquí tienes un ejemplo, con 7 repeticiones o 3500 casos:

389, 408, 383, 389, 392, 386, 384

Sumo y obtengo 2731, divido entre 3500 (para encontrar la frecuencia relativa) y multiplico por 4 para aproximar a π: 2731/3500*4=3,1211. No es una extraordinaria aproximación a π, pero resulta aceptable si tenemos en cuenta las herramientas utilizadas.

Análisis de intervalos

En una de las actualizaciones del Simulador hemos añadido una simulación entre intervalos. En el caso de la distribución uniforme nos servirá para analizar la igualdad aproximada de las frecuencias en intervalos de igual longitud.

En la segunda hoja *Estadísticos* figura una tabla y un botón para obtener frecuencias **f** entre dos extremos **a** y **b**. Estos extremos se consideran alcanzables, por lo que en las distribuciones discretas estarán incluidos. En la última fila se calcula la frecuencia relativa **h**. Sólo se puede usar para tres columnas o menos. Basta fijar los extremos en cada columna, pulsar el botón *Intervalos* y comparar resultados.

La imagen corresponde a una simulación uniforme continua entre 10 y 20, con tres columnas. En cada una de ellas hemos fijado extremos con una diferencia de 4, con lo que las tres frecuencias relativas se acercan a 0,4.

Intervalos

a	10	11	12
b	14	15	16
f	407	417	432
h	0,407	0,417	0,432

Otros ejemplos de distribución uniforme

El seis triple

En muchos juegos de mesa, sacar un 6 tres veces seguidas es un acontecimiento importante. Según la teoría, la probabilidad de que esto ocurra es de 1/(6*6*6)=1/216. Podemos esperar que de cada 216 tiradas triples que efectuemos, una de ellas presente 6-6-6. Podíamos simularlo para ver qué ocurre.

Para ello concretamos el Simulador:

- Distribución **uniforme entera** (es decir, discreta).
- Mínimo 1 y máximo 6
- 200 filas y 3 columnas (las filas son opcionales. Podrían ser 216 o el doble)
- Criterio **Máximo-mínimo**

Simulador		A. Roldán – Versión 2.2 – Año 2016	
	Repeticiones de la simulación		1
	Número de filas (de 1 a 1000)		200
	Número de columnas (si procede) (de 1 a 12)		3
Tipo de simulación		**Extremos (si procede)**	
Uniforme		**Mínimo**	**1**
		Máximo	**6**
		Parámetros	
Decimal – Entero			
Entero		**Media**	**200**
		Sigma	**8**
Criterios			
Máximo-Mínimo		**Otros parámetros**	

Así, a simple vista percibiremos bien si resulta un 6 triple.

Nosotros hemos efectuado dos simulaciones sin obtener el seis triple, que sólo ha aparecido en el tercer intento:

5	1	6		
4	3	4		
6	6	6		1
2	5	3		
3	3	5		
5	5	1		

¿Por qué, entonces, creemos que es más fácil conseguirlo? Todo el que ha jugado recuerda haber enlazado tres 6 seguidos. Ocurre que recordamos mejor las jugadas favorables, y no las desfavorables, como cuando en el juego de la Oca nos toca la cárcel o la muerte, o las veces que necesitamos un cinco para salir en el parchís y no nos sale en muchas jugadas.

La lotería primitiva

En esta lotería española pueden salir los números del 1 al 49, y los apostadores eligen entre varios tipos de apuestas. Se puede tratar como distribución uniforme porque todos los resultados presentan la misma probabilidad (1/13983816), aunque tenemos que acudir a un muestro sin reemplazamiento, que refleja mejor el proceso. En el sorteo aparecerán seis números y se gana más o menos dinero según las coincidencias entre nuestra apuesta y el resultado del sorteo. Suponemos la apuesta más sencilla, compuesta de seis números.

< **Sorteo 100, 15/12/2016**

Combinación ganadora

16 17 21 26 31 38

Nuestro objetivo es el de comprobar la dificultad de acertar varios resultados, que suelen tenerse en cuenta entre tres o seis aciertos, cambiando la cuantía del premio según el número de aciertos.

Abrimos el ***Simulador.***

Le concretamos los siguientes datos:

Simulador		A. Roldán – Versión 2.2 – Año 2016	
	Repeticiones de la simulación		1
	Número de filas (de 1 a 1000)		6
	Número de columnas (si procede) (de 1 a 12)		1
Tipo de simulación		**Extremos (si procede)**	
Muestreo sin		**Mínimo**	1
		Máximo	49
		Parámetros	
Decimal – Entero		**Media**	10
Entero		**Sigma**	2
Criterios			
Máximo-Mínimo		**Otros parámetros**	

- Máximo y mínimo, de 1 a 49, que son los números del sorteo
- Distribución **Muestreo sin** y con **enteros**
- Seis filas y una columna, para simular bien el sorteo
- Criterios **Máximo-Mínimo**

Es importante que actives exactamente lo que se ha sugerido. Al pulsar el botón **Simulación** aparecerán **seis números distintos** del 1 al 49. Si no lo logras, repasa bien todos los criterios necesarios.

Para hacer ver la dificultad de acertar, hemos adjuntado nuestra apuesta a la izquierda de la simulación:

Apuesta	Resultado
3	10
9	18
12	21
22	6
29	7
40	23

Ahora se trata de usar reiteradamente el botón de simulación y comprobar cuántos números hemos acertado. Nosotros hemos estado un buen rato jugando a simular y siempre hemos obtenido menos de tres aciertos. Valga esta experiencia para enseñarnos a jugar con prudencia.

Uso de la distribución uniforme para probabilidades dadas por una tabla

Con el Simulador y el uso de la distribución uniforme continua (con decimales) se pueden simular probabilidades dadas por una tabla. Por ejemplo, supongamos que disponemos de tres bolas rojas, dos verdes y dos amarillas para efectuar experimentos aleatorios. Lo normal sería introducirlas en una bolsa e ir sacando una a una con reposición. Esto, en las aulas, podría ser divertido y algo caótico. Una alternativa es simularlo con ordenador.

Partimos de la tabla de probabilidades

Roja	3/7
Verde	2/7
Amarilla	2/7

En primer lugar, la cambiamos a probabilidades acumuladas, es decir, que cada una sea la actual más todas las anteriores. Así:

Roja	3/7
Verde	5/7
Amarilla	7/7

Esto es para organizar unas desigualdades más adelante.

El proceso es como sigue:

En el simulador organizamos una distribución uniforme **con decimales**, **mínimo 0 y máximo 1** (los 7/7). Como criterio le damos que use el **Máximo y mínimo**, y fijamos, por ejemplo, 100 filas y una columna. En el parámetro A se ha escrito 3/7 y en el B 5/7 (esto no es obligatorio).

Copiamos la imagen de la hoja:

Simulador	A. Roldán – Versión 2.2 – Año 2016		
	Repeticiones de la simulación		1
	Número de filas		100
	(de 1 a 1000)		
	Número de columnas (si procede)		1
	(de 1 a 12)		
Tipo de simulación		**Extremos (si procede)**	
Uniforme			
		Mínimo	**0**
		Máximo	**1**
		Parámetros	
Decimal – Entero			
Decimal		**Media**	**10**
		Sigma	**2**
Criterios			
Máximo-Mínimo		**Otros parámetros**	
		A	**0,429**
		B	**0,71428571**

Si iniciamos la simulación obtendremos una columna con 100 números entre 0 y 1 expresados con decimales:

0,01
0,18
0,34
0,64
0,9
0,83
0,35
0,52
0,95
0,04
0,74
0,68
0,7
0,49
0,65
0,14

Estos números (y esto no lo consigue la herramienta) debemos convertirlos en **Rojo, Verde y Amarillo**. Para eso añadimos una fórmula a la derecha de cada número, de forma que si este no llega a 3/7 (que hemos escrito en la celda E21) , es que se trata de la Roja, y en caso contrario, si tampoco llega a 5/7 (probabilidad acumulada escrita en E22), se trata de la Verde, y si no, es Amarilla. La primera vez que se organiza esto resulta complicado, pero hay que insistir y buscar otros ejemplos. Esta fórmula, en nuestro caso, es:

=SI(G5<E$21;"Roja";SI(G5<E$22;"Verde";"Amarilla"))

(hemos copiado la de G5)

En ella se ve claramente el mecanismo: Si el número contenido en G5 es menor que el parámetro A (3/7), se trata de Roja. Si no es Verde o Amarilla según quede menor o mayor que el otro parámetro. El resultado lo tienes en la imagen:

0,01		Roja				
0,18		Roja		38	0,38	0,43
0,34		Roja		33	0,33	0,29
0,64		Verde		29	0,29	0,29
0,9		Amarilla				
0,83		Amarilla		100		
0,35		Roja				
0,52		Verde				
0,95		Amarilla				
0,04		Roja				
0,74		Amarilla				
0,68		Verde				

Los números que no llegan a 3/7 producen una roja, los siguientes hasta 5/7 la verde y el resto amarilla. Para ver si la simulación es razonable hemos contado los colores, resultando 38 rojas, 23 verdes y 29 amarillas. A la derecha ves la comparación con la teoría: las amarillas han resultado muy aproximadas y las otras con errores de centésimas, luego aceptamos el procedimiento.

Media y desviación típica

En teoría (no en una simulación concreta), si la distribución uniforme es continua entre los extremos a y b, su media es (a+b)/2 y su varianza (b-a)^2/12. Con el simulador podríamos aproximarnos a esos valores. Construimos un ejemplo:

Estimar la media y desviación típica de una distribución uniforme generada entre los extremos 10 y 20

Concretaremos distribución uniforme con decimales, **extremos 10 y 20**, y para obtener una buena aproximación, usaremos 100 repeticiones del experimento, con una sola columna y 20 filas (todo esto ha sido opcional. Puedes cambiarlo)

Simulador		A. Roldán – Versión 2.2 – Año 2016		
	Repeticiones de la simulación			100
	Número de filas			20
	(de 1 a 1000)			
	Número de columnas (si procede)			1
	(de 1 a 12)			
Tipo de simulación			Extremos (si procede)	
Uniforme				
			Mínimo	10
			Máximo	20
			Parámetros	
Decimal – Entero				
Decimal			Media	4
			Sigma	2
Criterios				
Máximo-Mínimo			Otros parámetros	
			A	32,000
			B	0,5
			C	1,2
Número intervalos			D	
	10			

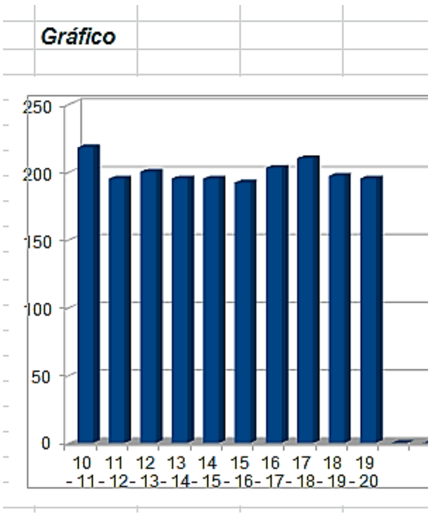
En este caso también concretaremos la opción de usar **máximo y mínimo** y fijaremos los intervalos en **10**:

De esta forma generaremos 2000 datos uniformes entre 10 y 20. Su media debería ser (10+20)/2=15 y su varianza (20-10)^2/12=8,33. La desviación típica esperada será la raíz cuadrada de esta cantidad, 2,88.

Iniciamos la simulación y obtenemos:

Estadísticos de momentos	
Media	14,9677
Desviación típica	2,9008
Asimetría	-0,0086
Curtosis	-1,2125
Valores teóricos	
Media	15,0000
Desviación típica	2,8868
Asimetría	0,0000
Kurtosis	-1,2000

Ha resultado bastante aproximado: media 14,96 y desviación típica 2,9008. En el gráfico se percibe bastante bien la uniformidad:



Obtención de muestras

Otro de los objetivos de una simulación es el de obtener muestras para después usarlas en otras cuestiones. Por ejemplo, se pueden usar para comprobar algunos teoremas de Inferencia Estadística, al relacionar los estadísticos de una muestra con los parámetros de la población. También pueden ser la base de otras cuestiones elementales. La ventaja de usar el Simulador es que podemos dar forma rectangular a la muestra y analizarla por filas y columnas. En este documento nos pueden ser útiles también para repasar el funcionamiento del Simulador, al intentar reproducir las muestras sugeridas.

Vemos algún ejemplo:

Tiradas sucesivas de cinco dados

Muchas familias juegan al póker de dados. Una idea bastante buena de lo que se espera en cada tirada es construir una muestra de cinco columnas y las filas que deseemos:

1	1	5	2	1
6	1	2	6	6
5	3	6	3	4
6	4	6	4	5
2	6	6	5	2
1	4	6	3	4
3	5	5	1	3
4	5	6	5	3
4	1	3	5	4
3	1	2	2	2
2	1	6	1	4
6	6	1	2	2
1	4	6	5	3
1	1	4	1	1
3	4	6	2	6
5	2	5	4	2
6	2	6	6	3
6	2	1	3	3
5	2	5	4	3
5	1	2	5	3
5	1	1	4	4
2	4	5	6	5
2	3	3	3	4
1	1	2	6	1
1	3	3	4	3
6	1	6	5	3

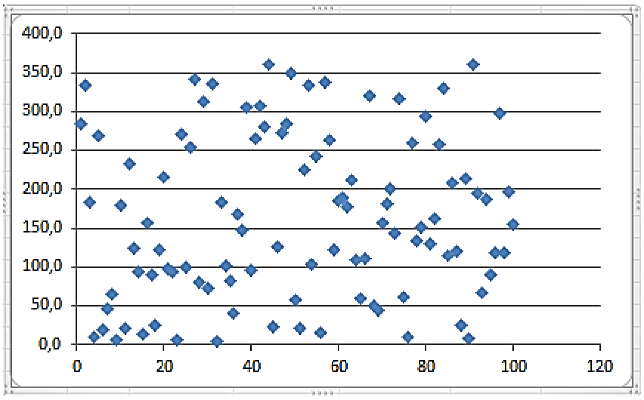
Detectamos en ellos bastantes tríos y dobles parejas, junto con parejas aisladas, pero no parece que haya ningún "full" ni póker. Estas muestras aclaran la dificultad de estas jugadas.

Ángulos en la circunferencia

281,9	331,9	182,0	7,5	266,6	18,1	43,5	62,8	5,1	178,1	19,0	232,0
	121,9	92,1	12,8	155,0	88,2	24,3	120,4	214,5	95,3	92,9	4,4
	269,8	98,2	252,9	339,9	79,0	310,6	71,3	333,7	2,1	181,3	99,1
	81,4	39,1	165,9	145,3	302,9	94,3	264,5	305,3	278,4	359,1	21,3
	123,6	272,2	282,2	347,6	55,9	19,8	223,6	332,2	102,0	241,4	14,5
	335,1	262,1	121,3	183,7	187,6	176,0	210,6	106,8	58,3	108,3	318,3

Para construir esta muestra en modo texto se ha realizado una simulación de una sola columna, copiando después mediante transposición a una sola fila, y desde ella se ha copiado en el documento en modo texto. Resultan unas tabulaciones que se pueden corregir posteriormente. Se ha reducido el número de decimales a uno.

La inserción de un gráfico de dispersión destaca la relativa uniformidad de la muestra, al producir una nube de datos sin tendencia definida.



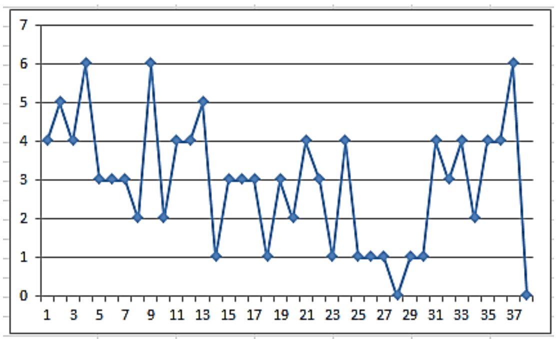
Tiradas en una ruleta

Existe la creencia popular de que, si ha salido un número en la ruleta, tardará en salir de nuevo, por lo que conviene memorizar resultados. También se cree que unos números salen más que otros. Probamos fortuna.

Hemos creado una muestra de tiradas en la ruleta francesa (números del 0 al 36) una sola vez, para ver qué puede ocurrir en la realidad. Confeccionaremos una muestra de 111=37*3 tiradas, para descubrir repeticiones.

1	36	15	30	11	5	2	11	23	33	12	3
	10	26	23	7	8	31	8	12	18	8	31
	0	1	36	3	20	24	29	14	23	23	34
	0	2	33	14	25	19	8	32	34	11	32
	35	8	5	30	3	21	21	3	9	18	1
	15	0	18	30	12	30	1	34	10	35	14
	3	10	15	12	20	32	6	11	12	19	36
	32	16	17	4	8	22	28	20	6	2	3
	1	5	31	36	4	16	35	34	9	2	4
	35	16	36	21	20	7	36	13	6	0	10

Vemos que la banca sólo gana cuatro veces, que son los ceros que han salido. Los números más repetidos han sido el 23 y el 30, con seis repeticiones cada uno (parecen muchas) y no ha salido el 20 en 111 tiradas. Mucha irregularidad, como se ve en el gráfico:



Las repeticiones más frecuentes son las de tres o cuatro veces, por lo que el hecho de repetir tirada o no resulta poco esclarecedor a la hora de volver a apostar.