

1. En este ejercicio estudiaremos la distribución del promedio de variables independientes e idénticamente distribuidas y a través de los histogramas correspondientes analizaremos el comportamiento de estas distribuciones a medida que promediamos un número creciente de variables aleatorias. Es decir, trataremos de ver qué quieren decir empíricamente los resultados de la Ley de los Grandes Números y el Teorema Central del Límite.

Para ello generaremos una muestra de variables aleatorias con una distribución dada y luego calcularemos el promedio de cada muestra. Replicaremos esto mil veces, es decir, generaremos una muestra aleatoria de la variable  $\bar{X}_n$  de tamaño 1000. Observe que, en principio, desconocemos la distribución de  $\bar{X}_n$ . A partir de todas las repeticiones realizaremos un histograma para los promedios generados para obtener una aproximación de la densidad o la función de probabilidad de  $\bar{X}_n$ .

- (a) Comencemos por tomar un primer conjunto de datos de variables aleatorias  $X_1, \dots, X_{1000}$  independientes con distribución  $U(0, 1)$ . Le pedimos al R que nos genere una muestra de ellas y luego hacemos un histograma. O sea, estamos haciendo un promedio de una sola observación y repitiéndolo 1000 veces. ¿A qué densidad se parece el histograma obtenido?

- (b) Considerar dos variables aleatorias  $X_1$  y  $X_2$  independientes con distribución  $U(0, 1)$  y el promedio de ambas, es decir,

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

Generando una muestra de dos variables aleatorias con distribución  $U(0, 1)$  computar la variable promedio. Replicar 1000 veces y a partir de los valores replicados realizar un histograma. ¿Qué características tiene este histograma?

- (c) Aumentemos a cinco las variables promediadas. Considerar ahora 5 variables aleatorias uniformes independientes, es decir  $X_1, X_2, \dots, X_5$  i.i.d. con  $X_i \sim U(0, 1)$  y definir

$$\bar{X}_n = \frac{1}{5} \sum_{i=1}^5 X_i.$$

Generando muestras de cinco variables aleatorias con distribución  $U(0, 1)$  computar la variable promedio. Repetir 1000 veces y realizar un histograma para los valores obtenidos. Comparar con el histograma anterior. ¿Qué se observa?

- (d) Aumentemos aún más la cantidad de variables promediadas. Generando muestras de 30 variables aleatorias con distribución  $U(0, 1)$  repetir el ítem anterior. ¿Qué se observa?

- (e) Ídem anterior generando muestras de 500 variables aleatorias. ¿Qué pasa si se aumenta el tamaño de la muestra? Observar que para poder comparar los histogramas de los distintos conjuntos de datos será necesario tenerlos dibujados en la misma escala tanto para el eje horizontal como para el vertical. Por eso, en general son más cómodos otros gráficos para comparar distintos conjuntos de datos.

- (f) Finalmente hacerlo también para 1200, y hacer un gráfico de las estimaciones de la densidad (una versión continua del histograma) de los 6 conjuntos de datos en el mismo gráfico. En este gráfico se verá que a medida que aumenta el  $n$  los valores de los promedios tienden a concentrarse, ¿alrededor de qué valor? Calcule media y varianza muestral para cada conjunto de datos. ¿Puede dar los valores teóricos a los que deberían parecerse?

- (g) El teorema central del límite nos dice que cuando hacemos la siguiente transformación con los promedios,  $\frac{\bar{X}_n - E(X_1)}{\sqrt{\frac{Var(X_1)}{n}}}$ , la distribución de estas variables aleatorias se aproxima a la de la normal estándar, cuando  $n$  es suficientemente grande. Para comprobarlo empíricamente, hagamos esta transformación en los 6 conjuntos de datos (es razonable hacerlo para valores de  $n$  suficientemente grandes, lo realizaremos en todos los casos para comparar) y luego comparemos los datos transformados mediante histogramas.

- (h) Repetir los ítems anteriores generando ahora variables con distribución  $\mathcal{C}(0, 1)$ . Comparar los resultados obtenidos. Recordar que la densidad de una Cauchy es

$$f_X(x) = \frac{1}{\pi(1+x^2)},$$

que es una densidad simétrica alrededor del cero, con colas que acumulan más probabilidad que la normal estándar, y que no tiene esperanza ni varianza finitas.

2. Expliquemos ahora el Tablero de Galton (Galton's board, o bean machine o quincunx).

<http://www.jcu.edu/math/iseq/quincunx/quincunx.html>  
<http://www.math.uah.edu/stat/applets/GaltonBoardExperiment.html>

Georgii, Hans-Otto. *Stochastics. Introduction to Probability and Statistics*. Walter de Gruyter, First edition, 2008, pág. 33. (es el texto y la imagen que figuran a continuación).

En R: <http://vis.supstat.com/2013/04/bean-machine/>

```
library(animation)
ani.options(interval = 0.03, nmax = 213)
quincunx()
```

The significance of the binomial distribution as the distribution of the number of successes in a Bernoulli experiment can be illustrated physically, in the case  $p = 1/2$ , by the *Galton board*, see Figure 2.2. A ball is dropped at the top of the board and passes  $n$  rows of pins, at which it is deflected either to the right or to the left, each with probability  $1/2$ . The probability that it takes the right option  $k$  times is given by  $\mathcal{B}_{n,1/2}(\{k\})$ ,  $0 \leq k \leq n$ . In this case it ends up in compartment  $k$ . If the compartments are big enough to accommodate a large number of balls, the relative frequency of the balls in compartment  $k$  is close to  $\mathcal{B}_{n,1/2}(\{k\})$ , by the law of large numbers (Theorem (5.6)).

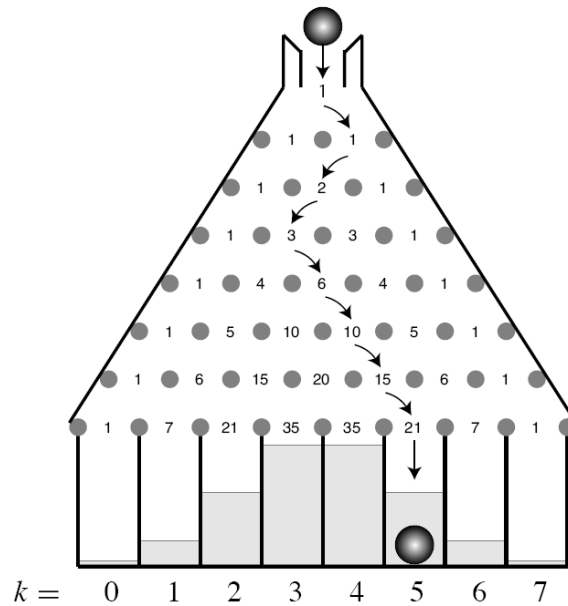


Figure 2.2. The Galton board for  $n = 7$ . The arrows mark a possible path for the ball. The numbers indicate the total number of paths leading to the respective position. Note that they form Pascal's triangle for the binomial coefficients. In the compartments at the bottom, the bar chart for  $\mathcal{B}_{7,1/2}$  is indicated in grey.