Histograms & Densities

We have seen in class various pictures of theoretical distribution functions and also some pictures of *empirical distribution functions* based on data. The definition of this concept is as follows. If $X_1, , ldots, , X_n$ represents a data sample (sequence of independent random variables which are all 'identically distributed' in the sense of having the same, possibly unknown, distribution function $F = F_X$, then the **empirical distribution function** based on the data is a distribition function which has jumps at observed values X_i of size equal to $1/n$ multiplied by the number of sample points equal to that value. In symbols,

$$\hat{F}_n(t) \; = \; \frac{1}{n} \sum_{i=1}^{n} I_{[X_i \le t]}$$

where as before, I_A is an 'indicator' equal to 1 if the property A is true, and 0 otherwise. The point of this definition is that $n\hat{F}_n(t)$ is the sum of n independent identically distributed binary ('coin-toss') random variables, all with the same probability $p = P(X_i \le t) = F(t)$ of being 1; thus

$$n \, \cdot \, \hat{F}_n(t) \; \sim \; \text{Binom}(n, \, F(t))$$

and the Law of Large Numbers (and Central Limit Theorem too) imply that $\hat{F}_n(t) \; \approx \; F(t)$ with probability close to 1, when n gets large. We have seen pictures of this in class (also on the web-page) for a few examples. Plots of empirical d.f.'s from data overlaid with theoretical d.f.'s are among the best simple ways of checking visually that the data really fit the theoretical d.f. well.

We now define a data-display concept — the **scaled relative frequency histogram** — which gives us a more refined look at the distribution of a large sample of data. Again, if we start with a sample $X_1, \, \ldots, \, X_n$, such a histogram is a picture defined through the following steps. (See also Chapter 1 of the textbook.)

Step 1. Mark off the data axis into a number L equal-length intervals which cover the whole range (from min to max) of the data points. The number L should vary with n but not be too small (5 or 6 is a kind of minimum): one theoretically based proposal is to let $L = [n^{2/5}]$ (rounded down to the nearest integer). There are now $L + 1$ points a_k and L intervals $(a_k, \, a_{k+1}]$. Let h be the width of each interval.

Step 2. Tally the numbers n_k of data points falling in the intervals,

$$n_k \;=\; \sum_{i=1}^{n} I_{[a_k < X_i \le a_{k+1}]} \quad , \qquad k = 1, \ldots, L$$

Step 3. Draw a bar-chart with a bar of constant height $n_k/(nh)$ over the interval $(a_k,\, a_{k+1}]$.

Note that the total area in the bars of such a picture is 1, and it is this scaling feature which makes them so informative in plots overlaid with theoretical density functions.

We now proceed to illustrate the usefulness of these histograms as a way of verifying that simulated data really have the distributions they are designed to or that they are supposed to according to theoretical results:

(i) in a simulation of n=1000 exponentially distributed variables (Figure 1 below);

(ii) in a simulation of n=1000 weighted sums $Z_1 + 2Z_2 - 3Z_3$ of *iid* Normal(0,1) triples (Figure 2 below); and

(iii) in a simulation of n=10000 averages of 60 Uniform[0,1] variables (Figure 3 below).

The key point about *scaled* histograms is that the area in the histogram bar over the class-interval $(a_k, a_{k+1}]$ is the base (equal to h) multiplied by the height (equal to $n_k/(nh)$), that is, the area in the bar is $h \cdot (n_k/(nh)) = n_k/n$. But when n is really large, we know from the Law of Large Numbers that the relative frequency

$$n_k/n \;=\; \frac{1}{n} \sum_{i=1}^{n} I_{[a_k < X_i \le a_{k+1}]}$$

is with high probability going to be close to the probability of any one of these indicators being 1, or $P(a_k < X_1 \le a_{k+1}) = \int_{a_k}^{a_{k+1}} f_X(x)dx$, which is equal to the area under the true density curve over the same interval. **This reasoning shows why the tops of the histogram bars must agree closely with the density function graph over each class interval: the area under both of them must be very nearly equal when n is large, according to the Law of Large Numbers.**

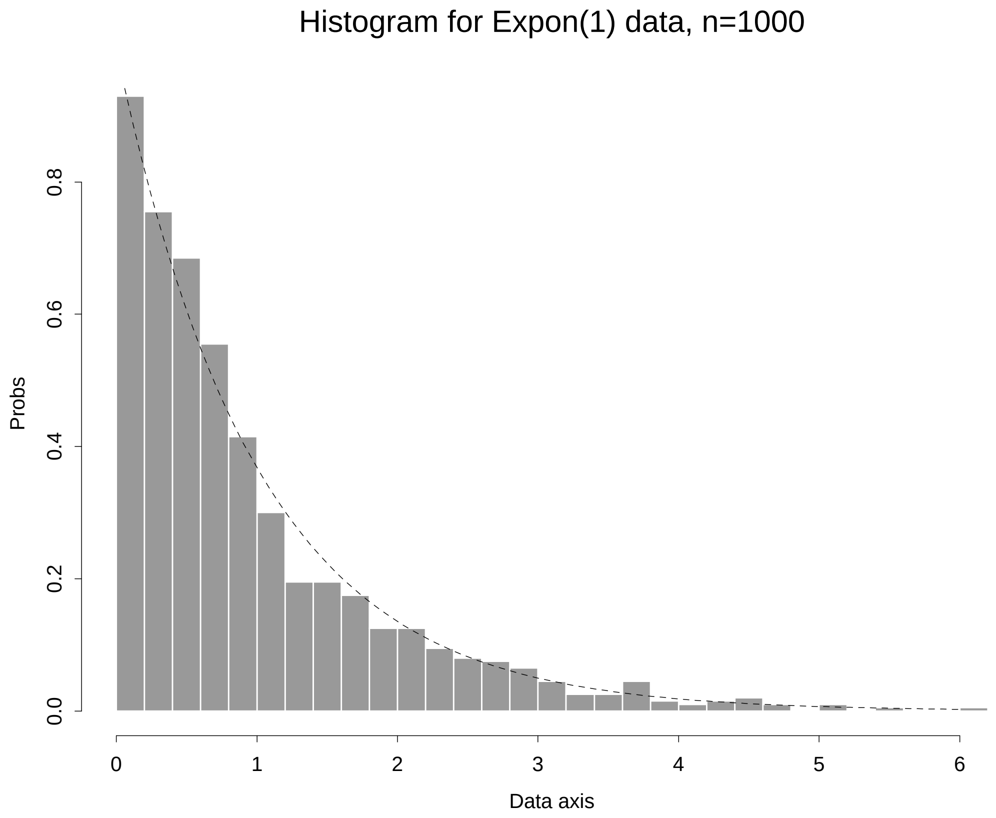

Figure 1: Scaled relative frequency histogram of 1000 Expon(1) simulated random values, broken into 36 intervals and plotted overlaid with the Expon(1) density function $f(x) = e^{-x},\ x \geq 0$.

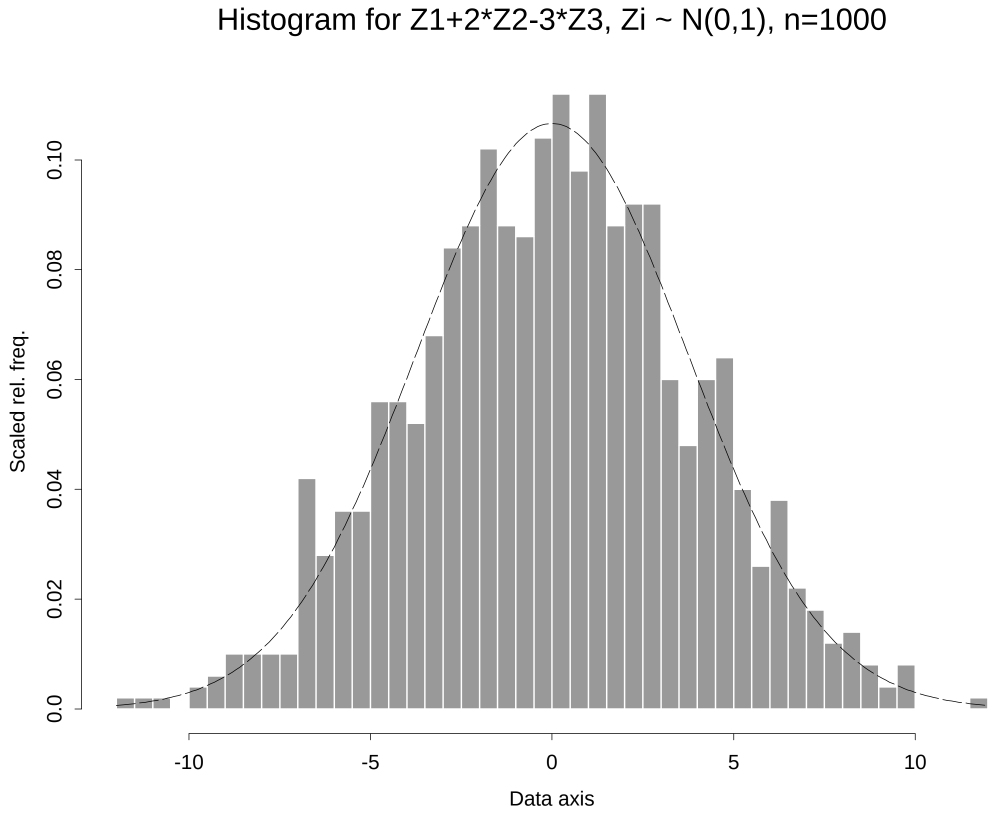

Figure 2: Scaled relative frequency histogram of 1000 simulated random values $Z_1 + 2Z_2 - 3Z_3$, $Z_i \sim \mathcal{N}(0, 1)$, broken into 36 intervals and plotted overlaid with the $\mathcal{N}(0, 14)$ density.

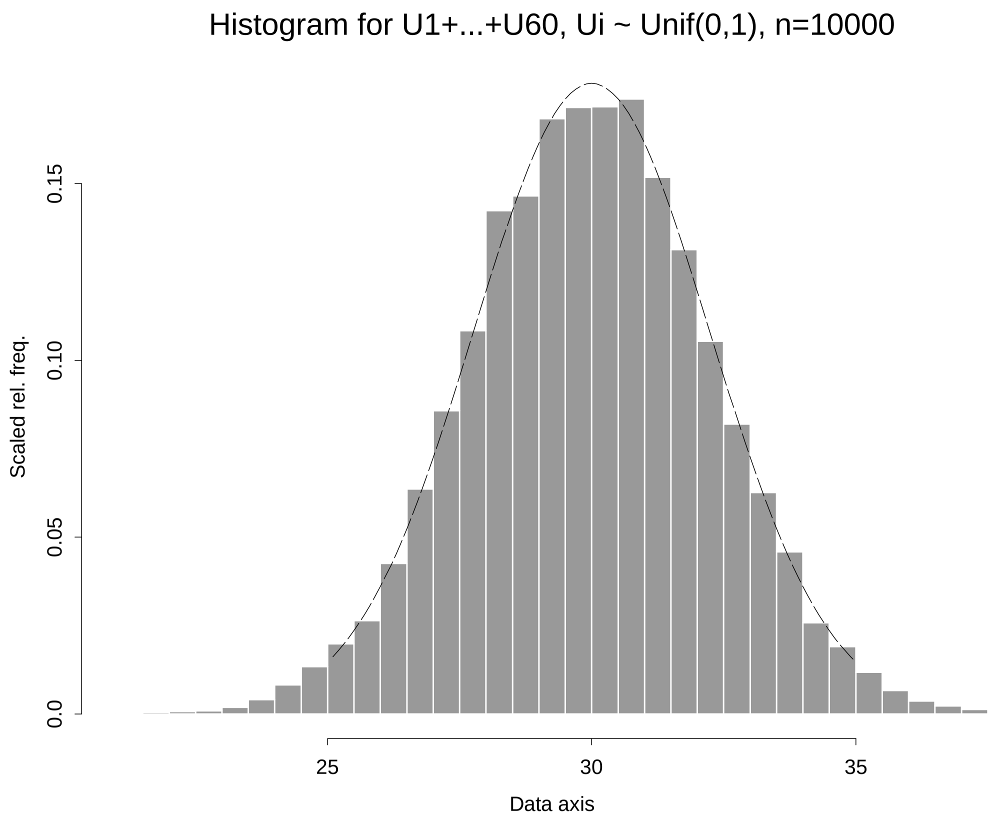

Figure 3: Scaled relative frequency histogram of 10000 simulated random values $U_1 + \cdots U_{60} \sim Unif(0,1)$, broken into 36 intervals and plotted overlaid with the $\mathcal{N}(30,5)$ density.