Eric Slud, Stat 430 Spring 2009

Scaled Relative Frequency Histograms

This handout is about the construction and motivation of **scaled relative frequency histograms** for purposes of comparison and graphical overlay with theoretical density functions. This is one of the main ways that histograms are used descriptively: to ask whether the shape of the distribution of observed data are ‘sufficiently normal’ or sufficiently close to the `gamma` or `student’s t` or whatever other distribution family is conjectured to provide a good fit to the data.

First we define what is meant by *histogram*, successively modeified by the adjectives ‘frequency’, ‘relative frequency’, and ’scaled relative frequency’. Imagine that a dataset $X_1, X_2, \ldots, X_n$ of observations has been drawn *iid* (indepently and with identical distribution) from a density function $f(x)$. Let $(a, b]$ be an interval on the measurement axis large enough to contain all of the points X_i, i.e. such that

$$a \;<\; \min_{1 \le i \le n} X_i \;<\; \max_{1 \le i \le n} X_i \;<\; b$$

Then divide this interval into a number m of equal length *class intervals*, each of width $w = (b-a)/m$, where the k’th such interval, J_k, is $(a+(k-1)w,\, a+kw]$. (Note that each of these intervals is defined to contain its right endpoint but not its left endpoint. Also note that by definition, the rightmost endpoint of J_m is $a + mw = b$.) All of the different kinds of histograms are pictorial representations of the frequency counts

$$n_k \;=\; \text{count in k'th interval} \;=\; \sum_{i=1}^{n} I_{[X_i - a \in J_k]} \qquad (1)$$

$$=\; \#\{i:\; 1 \le i \le n,\; (k-1)w < X_i - a \le kw\}$$

The counts $n_1, n_2, \ldots, n_m$ of numbers of observations X_i falling in the respective intervals J_k, could be arranged simply in a table,

Interval	J_1	J_2	$\cdots$	J_k	$\cdots$	J_m
Count	n_1	n_2	$\ldots$	n_k	$\ldots$	n_m

In a **frequency histogram**, a vertical bar is drawn (and sometimes filled in) over each interval J_k (with base w), with height labelled equal to the number n_k. A **relative frequency histogram** differs only with respect to the vertical units: the height of the bar over J_k is now labelled as n_k/n, which means that the total of bar-heights which was n in the frequency histogram is 1 in the relative-frequency histogram. However, since totals of bar-heights is not visually meaningful, the **scaled relative frequency histogram** redefines the vertical units in such a way that the *area* of the bar over J_k is the relative frequency n_k/n, which has the consequence that the total area within the bars of the scaled relative-frequency histogram is 1, making it just like a probability density function. Thus the *definition of the scaled relative frequency histogram* is:

$$\textbf{bar-height over} \quad J_k \quad = \frac{n_k}{w\,n} \quad \Longrightarrow \quad \text{Area} \;=\; w \cdot \frac{n_k}{wn} \;=\; \frac{n_k}{n} \tag{2}$$

and the important consequence of the definition is that the function equal to the tops of bars, with value $f_{hist}(x) \;=\; n_k/(nw)$ for $x \in J_k$ (and is 0 for $x \notin (a,b]$), is a legitimate probability density function. We argue next that with large-sample data, when the number of bars is chosen large but much smaller than the number of data points, this density is actually very close to the true but generally unknown density function of the X_i random variables.

Remark 1 *The histogram-based probability density function* $f_{hist}(x)$ *assigns area and probability* n_k/n *to the interval* $J_k = (a+(k-1)w, a+kw]$ *: a random variable* X *with density* f_{hist} *would fall in the interval* J_k *with probability* n_k/n*, and the precise value of* X *would be generated uniformly between* $a+(k-1)w$ *and* $a+kw$*.* □

Now for the argument that histograms should track densities closely. Suppose that the true density for the large number n of observed data values $X_1, \ldots, X_n$ is f, and suppose that $w = (b-a)/m$ is narrow but is fixed before we choose a large value of n. Consider the relation of the scaled histogram bar to the density over $J_k = (a+(k-1)w,\, a+kw]$. First, if n_k denotes the number of observations X_i falling within the interval J_k then the histogram bar has height $n_k/(nw)$ and area n_k/n, while the area under the true density function f over J_k is

$$\int_{a+(k-1)w}^{a+kw} f(x)\,dx \;\approx\; w \,\cdot\, f(a+(k-\frac{1}{2})w) \tag{3}$$

The approximate equality in this equation holds first of all because the interval has been assumed narrow, so that the continuous function $f(x)$ is nearly constant over the extent of J_k and the area is nearly that of the rectangle with base w and height equal to the value of f at the midpoint of the interval. A more precise argument, based on Taylor series (with mean-value-theorem form of remainder), shows that if the second derivative of f has absolute value bounded above by the constant C (everywhere on some interval containing J_k), then the difference between the right-hand and left-hand sides of (3) is $\leq \frac{1}{2} C\, w^2$. (This is a very small number if m has been chosen large, making w small.)

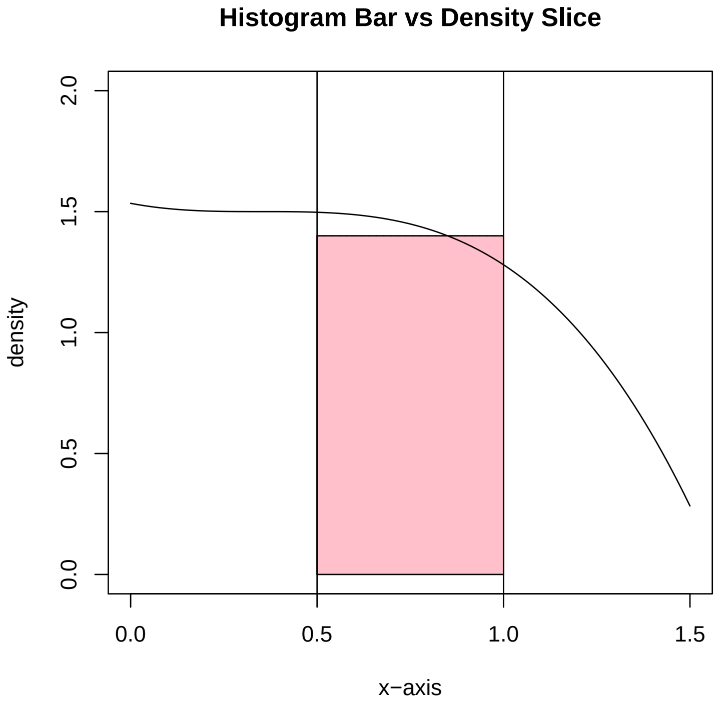

Figure 1: Area under the true density f (shown solid) versus the area of the scaled relative histogram bar (with top shown dashed) over an interval J_k.

The areas over J_k under the scaled histogram and the true density are necessarily very close with high probability (when w is small and n large). This is a consequence of the **Law of Large Numbers** applied to sequences of coin-toss outcomes. That theorem says that the relative frequency over independent identically distributed trials indexed $i = 1, 2, \ldots, n$ of the events $[X_i \in J_k]$ is close (with high probability, for all large n) to $P(X_1 \in J_k)$. By definition (1), the relative frequency is n_k/n, the area of the bar in the scaled histogram. The true probability is given by the left-hand side of (3). Again appealing to (3) (and see also Figure 1), we conclude that the density height $f(a + (k - 1/2)w)$ at the midpoint of the interval must be very close with very high probability to $n_k/(nw)$.

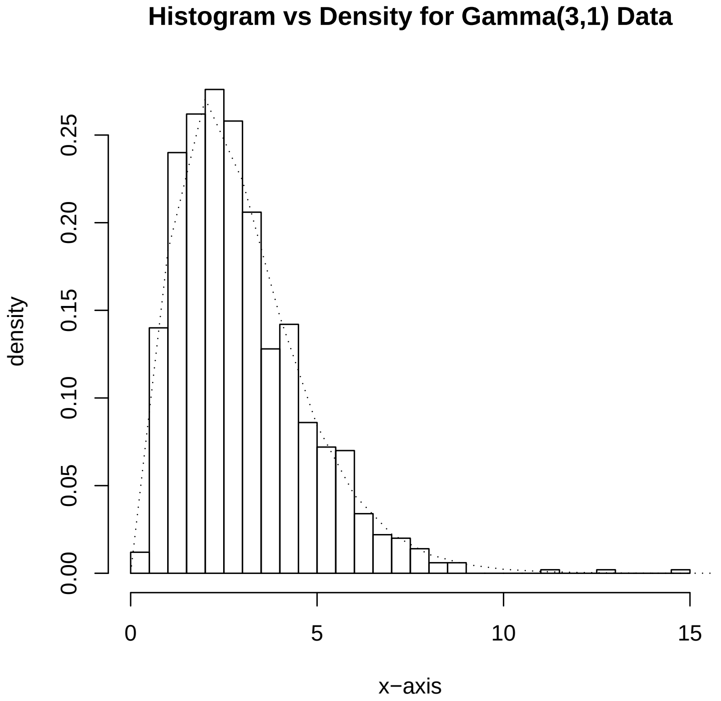

Figure 2: Scaled relative frequency histogram with m=32 class intervals for $n = 1000$ Gamma(4,1) variates generated in **R**, with true density function overlaid as dashed curve.

Figure 2 (generated in the statistical package **R**, but we will later do the same thing in `SAS`) shows the scaled relative histogram (based on $m = 32$) generated from 1000 data points generated from the $Gamma(4, 1)$ density, overlaid with the actual theoretical density plotted as a dashed curve.