Eric Slud, Stat 430 February 6, 2009

Handout on Empirical Distribution Function and Descriptive Statistics

The purpose of this handout is to show you how all of the common (univariate) descriptive statistics are computed and interpreted in terms of the so-called **empirical distribution** function.

Suppose that we have a (univariate) dataset $X_1, X_2, \ldots, X_n$ consisting of observed values of random variables that are *iid* or 'independent identically distributed', i.e., are independent random variables all of which follow the same probability law summarized by the

$$\textbf{distribution function} \quad F_X(t) \;=\; F(t) \;=\; P(X_1 \le t)$$

Recall that a distribution function is a nondecreasing right-continuous function with values in the interval $[0,1]$ such that $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$. This $F(t)$ is a theoretical quantity, which we can estimate in terms of data using the

$$\textbf{empirical distribution function} \quad F_n(t) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \le t]} \tag{1}$$

where I_A is the so-called *indicator random variable* which is defined to be equal to 1 when the property A holds, and equal to 0 otherwise. Thus, while the distribution function gives as a function of t the *probability* with which each of the random variables X_i will be $\le t$, the empirical distribution function calculated from data gives the *relative frequency* with which the observed values are $\le t$.

To understand why the empirical distribution function $F_n(t)$ accurately estimates the theoretical distribution $F(t)$, we must appeal to the famous Law of Large Numbers, which we next state in two versions (the second more general that the first).

Theorem 1 (**Coin-Toss Law of Large Numbers**) *Suppose that a sequence of independent coin-tosses have values* Z_i *given as* 1 *for Heads and* 0 *for Tails, with the heads-probability* $p = P(Z_i = 1)$ *the same for all* $i \geq 1$. *Then as the total number* n *of tosses gets large, for each* $\epsilon > 0$,

$$P\Big(| \frac{1}{n} \sum_{i=1}^{n} Z_i \, - \, p | \, \geq \epsilon \Big) \, \rightarrow \, 0$$

and with probability 1 *the sequence of numbers* $\bar{Z} \, = \, n^{-1} \sum_{i=1}^{n} Z_i$ *converges to* p *as* $n \rightarrow \infty$.

Theorem 2 (**General Law of Large Numbers**) *Suppose that random variables* X_i *for* $i \geq 1$ *are independent and identically distributed with distribution function* F, *and that* $g(x)$ *is any function such that* $E|g(X_1)| \, = \, \int \, |g(x)| \, dF(x) \, < \infty$. *Then as* $n \rightarrow \infty$, *for each* $\epsilon > 0$,

$$P\Big(| \frac{1}{n} \sum_{i=1}^{n} g(X_i) \, - \, E(g(X_1))| \, \geq \epsilon \Big) \, \rightarrow \, 0$$

and with probability 1 *the sequence of numbers* $n^{-1} \sum_{i=1}^{n} g(X_i)$ *converges to* $E(g(X_1) \, = \, \int g(x) \, dF(x)$ *as* $n \rightarrow \infty$.

In particular, based on large samples of data $\{X_i\}_{i=1}^{n}$, if we fix t and define the random variable $Z_i \, = \, g(X_i) \, = \, I_{[X_i \leq t]}$, then $\bar{Z} \, = \, n^{-1} \sum_{i=1} g(X_i) \, = \, F_n(t)$, and either of the two Theorems shows that $F_n(t)$ has very high probability of being extremely close to

$$Eg(X_1) \, = \, \int_{-\infty}^{t} 1 \, dF(x) \, + \, \int_{t} 0 \, dF(x) \, = \, F(t)$$

(A slightly stronger form of the law of large numbers, called the Glivenko-Cantelli Theorem, says that under the same hypotheses $\sup_{-\infty < t < \infty} |F_n(t) - F(t)|$ converges to 0 with probability 1 as $n \rightarrow \infty$.

These results tell us that the distribution function, which is generally hypothetical and unknown, can be recovered very accurately with high probability based on a large sample of independent identically distributed observations following that distribution.

Now let us think about summary values describing the underlying distribution function F, and how these summary numbers translate into **descriptive statistics** when we estimate them from the sample of data. The mean of the probability distribution function F (which may be thought to have density $F' = f$) is

$$\mu = \mu_1 = E(X_1) = \int_{-\infty}^{\infty} t\, f(t)\, dt = \int_{-\infty}^{\infty} t\, dF(t)$$

while our usual estimate of it is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i = \int t\, dF_n(t)$$

which can be thought of as the mean value of the *discrete probability distribution* of a randomly chosen member of the list $\{X_1,\, X_2,\, \ldots,\, X_n\}$. The index i of this randomly chosen member ofthe list is equally likely (i.e.,has probability $1/n$) to be any of the values $1, 2, \ldots, n$.

We recall also the formula for expectation of a discrete random variable, or a function of such a discrete variable. If W is a discrete random variable with k possible distinct values w_j, and with probability mass function

$$p_W(w_j) = P(W = w_j) = p_j \quad \text{for} \quad j = 1, \ldots, k\ , \quad \text{with} \quad \sum_{j=1}^{k} p_j = 1$$

then for any function h, the **expectation of** $h(W)$ is expressed as a weighted average of the possible values $h(w_j)$ weighted by the probabilities with which they occur,

$$E\Big(h(W)\Big) = \sum_{j=1}^{k} h(w_j)\, p_j \tag{2}$$

Think of the random variable W as the random index from $\{1, 2, \ldots, n\}$ chosen equiprobably as the position in the data column $\{X_1, \ldots, X_n\}$ from which a random data point X_W is drawn. (Here the number of distinct possible values of W is $k = n$, and the probability masses $p_j = P(W = j) = 1/n$ for $j = 1, \ldots, n$. In the present paragraph, the values X_i are viewed as fixed entries of the data column, not as random variables. The random variable Then the distributional quantities associated with X_W are called *empirical*, a terminology which makes sense because the distribution function of X_W is (for fixed $\{X_1, \ldots, X_n\}$)

$$P(X_W \le t) = P(W \in \{i :\ 1 \le i \le n,\ \ X_i \le t\}) = \frac{1}{n} \sum_{j=1}^{n} I_{[X_j \le t]} = F_n(t)$$

is just the empirical distribution function we defined before.

But now we can go further to obtain a unified understanding of descriptive statistics: the *sample moments* and quantities derived from them are obtained by finding the moments from the empirical distribution function, and the *sample quantiles* are simply defined as the quantiles from the empirical distribution function.

For simplicity and uniformity of notation, suppose that the underlying distribution function F is differentiable, with derivative $F' = f$. Then the **moments** and **quantiles** of the theoretical distribution F are defined, for integers $r \ge 1$ and probabilities $p \in (0,1)$, by

$$\mu_r \;=\; r^{th}\, \textbf{moment} \quad = \; E(X_1^r) \;=\; \int x^r\, f(x)\, dx$$

$$x_p \;=\; p^{th}\, \textbf{quantile} \quad = \; \begin{cases} F^{-1}(p) \quad \text{if} \quad F(x) = p \quad \text{has unique sol'n } x \\ \text{otherwise, midpoint of interval of solutions} \end{cases}$$

The $r = 1$ moment is simply the *mean* or expectation. The **variance** σ^2 has a well-known simple expression in terms of the first and second moments,

$$\text{Var}(X_1) \;=\; E(X_1 - \mu_1)^2 \;=\; E(X_1^2) \;-\; 2\,E(X_1)\,\mu_1 \;+\; \mu_1^2 \;=\; \mu_2 \;-\; \mu_1^2$$

and the **standard deviation** σ is defined as the square root of the variance. The **skewness** of a random variable X_1 is a measure of the asymmetry of the random variable's distribution, defined as the expectation

$$\textbf{skewness} \;=\; E\Big((X_1 - \mu_1)/\sigma\Big)^3 \;=\; (\mu_3 \;-\; \mu_1^3 \;-\; 3\,\mu_1\sigma^2)/\sigma^3$$

Finally, the **kurtosis** of X_1 is a measure of the heaviness of the tails of the density f of X_1, defined by

$$\textbf{kurtosis of}\;\; X_1 \;= E\Big((X_1 - \mu_1)/\sigma\Big)^4 \;-\; 3$$

The skewness and kurtosis were invented to compare distributional shape with the standard-normal or 'bell curve' density. Both skewness and kurtosis are easily checked to be 0 for the normal. We consider the values for two other familiar distributions, the t_k with k degrees of freedom and the Gamma$(k,1)$ with shape parameter k. Both of these distributions are close to normally distributed when k is large. The t_k is symmetric and has skewness 0 for all values of k, while the Gamma$(k,1)$ has skewness $2/\sqrt{k}$. The kurtosis is ∞ for t_k for $k \le 4$, but for larger values of k is given as follows:

Kurtosis

k=	5	6	7	8	10	20	80
t-dist (k df)	6	3	2	1.5	1	.375	.017

The quantiles of a distribution with density which is strictly positive on an interval and 0 outside that interval are just the inverse distribution function values, $x_p = F^{-1}(p)$. The p'th quantile is also called the **100p'th percentile**, and several quantiles have special names: the 1/2 quantile is the **median**, the 1/4 quantile is the **lower quartile** (designated Q1 in SAS), and the 3/4 quantile is the **upper quartile** (designated Q3).

Finally, we briefly summarize the definitions and computing formulas for the sample moments and quantiles. The r^{th} **sample moment** is the r^{th} moment of the discrete random variable X_W (where the X_i data values are again regarded as fixed) is given by

$$\hat{\mu}_r \;=\; E(X_W^r) \;=\; \sum_{i=1}^n P(W=i)\, X_i^r \;=\; \frac{1}{n}\sum_{i=1}^n X_i^r$$

The $r = 1$ sample moment is the **sample mean**, $\hat{\mu}_1 = \bar{X}$. However, the variance of X_W differs from the usual definition of **sample variance** $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$, by the factor $n/(n-1)$, since it is easy to check that

$$\mathrm{Var}(X_W) \;=\; \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \;=\; \frac{n-1}{n}\, S^2$$

(The reason this factor is introduced is to make the estimator S^2 an unbiased estimator of σ^2, a correction that does matter in small data samples but does not really concern us here.) Finally, the *sample skewness and kurtosis* are obtained from the plug-in formulas

$$\textbf{sample skewness} \;=\; (\hat{\mu}_3 + 2\,\bar{X}^3 - 3\,\bar{X}\,\hat{\mu}_2)/(\hat{\mu}_2 - \bar{X}^2)^{3/2}$$

$$\textbf{sample kurtosis} \;=\; (\hat{\mu}_4 - 4\bar{X}\hat{\mu}_3 + 6\bar{X}^2\hat{\mu}_2 - 3\bar{X}^3)/(\hat{\mu}_2 - \bar{X}^2)^2$$

Sample quantiles $\hat{x}_p$ are obtained by the same rule used to calculate quantiles, but using the empirical distribution function inverse. The main difference is that it quite often happens in the discrete empirical distribution that there is a whole interval $[c, d)$ of values (between successive ordered values of X_i observations) on which the empirical distribution has the constant value p, and in this case the second part of our definition gives $\hat{x}_p = (c+d)/2$.