Statistical Climatology

Empirical distribution functions

Let $X_1,\ldots,X_n$ be random variables, independent and identically distributed with continuous cumulative distribution function (cdf) $F(x)=P(X_i \le x)$, which we write $X \sim F$. For a set A define

$$1(A)=\begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Note that $E(1(A))=1\times P(A)+0\times P(A^c)=P(A)$.

The *empirical distribution function* (edf) is $F_n(x)=\frac{1}{n}\sum_{i=1}^{n}1(X_i \le x)$. It is a discrete distribution with jumps of size $1/n$ at each observed value. The random variable $\sum_{i=1}^{n}1(X_i \le x)$ counts the number of observations ≤ x, and has a binomial distribution with parameters n and $F(x)$. We can think of $F_n(x)$ as an estimate of $F(x)$ that can be used even if we do not know a functional form for $F(x)$.

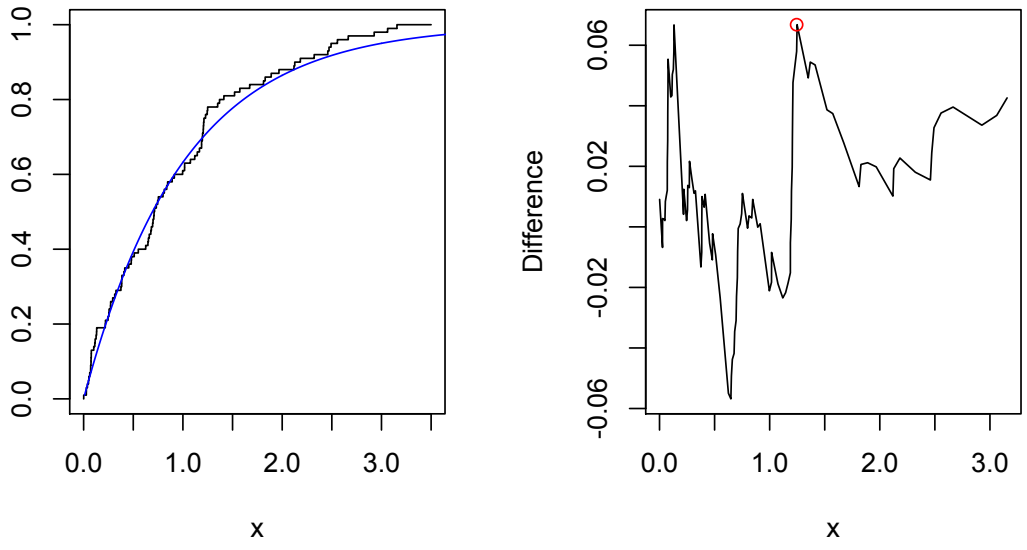

Figure 1: Left panel: Edf (black) and cdf (blue) for a sample of size 100 from an exponential distribution. Right panel: Difference between edf and cdf for the same sample. The red point indicates the largest difference, called D_n below.

Theorem 1 describes some useful properties of $F_n(x)$.

Theorem 1:

$$(a)\ E(F_n(x)) = F(x)$$
$$(b)\ Var(F_n(x)) = F(x)(1-F(x))/n$$
$$(c)\ P(|F_n(x)-F(x)| > \varepsilon) \to 0 \text{ as } n \to \infty \text{ for any } \varepsilon > 0.$$

Proof:

(a) $E(F_n(x)) = \frac{1}{n}\sum_{i=1}^{n} E(1(X_i \le x)) = \frac{n}{n}P(X_i \le x) = F(x)$

(b)

$$Var(F_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n} Var(1(X_i \le x)) = \frac{n}{n^2}\left\{E(1(X_i \le x)^2) - \left[E(1(X_i \le x))\right]^2\right\} = \frac{F(x)-F(x)^2}{n}$$

(c) By Chebyshev's inequality we have

$$P(|F_n(x)-F(x)| > \varepsilon) \le Var(F_n(x))/\varepsilon^2 = \frac{F(x)(1-F(x))}{n\varepsilon^2} \to 0.$$

The difference $|F_n(x)-F(x)|$ tells us how well our estimate describes the unknown cdf, The largest value of this difference, $D_n = \sup_x |F_n(x)-F(x)|$ is called the *Kolmogorov distance* between the two distribution functions. An interesting fact is that the distribution of D_n does not depend on F, and is therefore called *distribution free*. Note that since *F* is continuous it is strictly monotone and has a unique inverse function F^{-1} (called the *quantile function*). Now compute

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x)-F(x)| = \sup_{y=F^{-1}(x)\in(0,1)} |F_n(F^{-1}(y))-F(F^{-1}(y))| = \sup_{y\in(0,1)} |F_n(F^{-1}(y))-y|$$

since for every x there is precisely one *y* such that $F^{-1}(y) = x$. Basically what happens is that we change the x-axis but not the y-axis, so the distance between the distribution functions stays the same. Now we need a probability fact.

Theorem 2:
If $X \sim F$ then $F(X) \sim \text{Uniform}(0,1)$.

Proof: $P(F(X) \le y) = P(X \le F^{-1}(y)) = F(F^{-1}(y)) = y.$

Note that it follows that $1(X_i \le F^{-1}(y)) = 1(F(X_i) \le y)$, so that $F_n(F^{-1}(y))$ has the same distribution as the empirical distribution of a sample from the uniform distribution. Hence for all continuous cdfs *F* we can compute the distribution of D_n by assuming that the sample is from the uniform distribution. It turns out that the distribution of $\sqrt{n}D_n$ converges relatively fast to a limiting distribution, the *Kolmogorov*

distribution, that is usually employed as an approximation to the exact distribution (this, however, is in principle computable for any value of *n*).

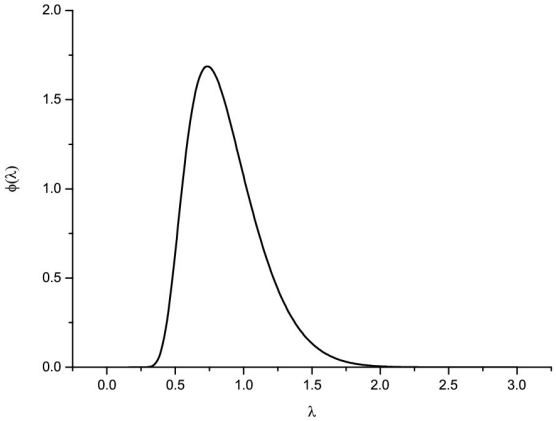

Figure 2. The asymptotic Kolmogorov density.

Let d_α be the cutoff value such that $P(Z > d_\alpha) = \alpha$ where Z follows the Kolmogorov distribution. We can use this to compute a confidence band for our unknown density *F*. We have

$$1-\alpha = P(Z \le d_\alpha) \approx P(\sqrt{n} D_n \le d_\alpha) = P\left(\sup \left| F_n(x) - F(x) \right| \le \frac{d_\alpha}{\sqrt{n}} \right)$$

$$= P\left(F_n(x) - \frac{d_\alpha}{\sqrt{n}} \le F(x) \le F_n(x) - \frac{d_\alpha}{\sqrt{n}} \text{ for all } x \right)$$

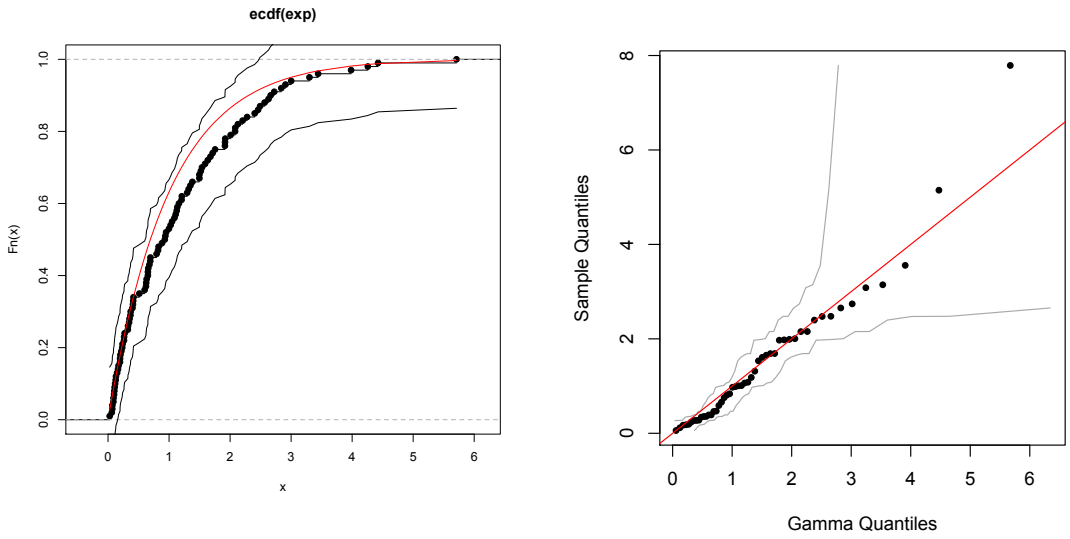

Figure 3. Simultaneous onfidence bands for cdf (left) and quantile-quantile plot (right) for another sample of size 100 from the exponential distribution. The red curve in the left panel is the true cdf that the data were generated from. The red line in the right panel is the line through the origin with slope one, corresponding to the theoretical and sample quantiles being the same.

As mentioned above, the quantile function is the inverse of the cdf. Since the edf is a step function it does not have a unique inverse. It is convenient to define the *empirical quantile function* as the right inverse of the edf, i.e. $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$. The *quantile-quantile plot* (or QQ-plot) displays the theoretical quantiles against the empirical quantiles for $p = 1/(n+1)$ ... $(n/n+1)$. If the theoretical quantiles describe the data well, a straight line with slope 1 should be close to all the points (Figure 3, right panel). To assess this better, a confidence band can be constructed much as for the cdf. The programs qqnorm and qqgamma in the R code at http://www.stat.washington.edu/peter/statclim/shiftplot.R compute these for the normal plot and the gamma distributions, respectively.

If we have two samples, one of size m from cdf F and one of size n from cdf G, with corresponding edfs F_m and G_n, we can compute the distance between the edfs as

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|$$

It turns out that if $F = G$, the distribution of $D_{m,n}$ does not depend on F, but can as in the one-sample case be computed assuming that F is uniform. Even more interesting, when m and n are large enough, $\sqrt{mn/(m+n)}D_{m,n}$ is well approximated by the same Kolmogorov distribution. Thus we compare two distributions by computing the QQ-plot of the two edfs with a simultaneous confidence band that can be computed in a similar fashion to before (and is produced by the qqplot program in the R software given above).

If we are not only interested in checking whether F and G are identical, but also allow them to be different, the *shift function* $\Delta(x)$ allows us to easily check simple relations such as shift or location-scale models. The shift function measures how much we must change F to get G, and is defined by $\Delta(x) = G^{-1}(F(x)) - x$ and can easily be estimated by plugging in the corresponding edfs F_m and G_n. Note that if $F(x-\theta) = G(x)$, i.e., we have a shift model, we can write $G(y+\theta) = F(y)$, and, by taking G^{-1} on both sides we get $y + \theta = G^{-1}(F(y))$ or $\theta = G^{-1}(F(y)) - y \equiv \Delta(y)$ so the shift function is constant and equal to the shift. Similarly, if we have a location-scale model the shift function becomes a straight line. The R software computes simultaneous confidence bands (derived from those for the QQ-plot) in the function shiftplot.

What happens if the assumptions of independence and identical distribution fails? Here is the counterpart to Theorem 1 in that case.

Theorem 3. If $X_i \sim F_i$ are not necessarily independent random variables, then

$$(a)\ E(F_n(x)) = \bar{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$$

$$(b)\ Var(F_n(x)) = \frac{1}{n^2}\sum\sum Cov\left(1(X_i \le x), 1(X_j \le x)\right)$$

$$= \frac{1}{n^2}\sum\sum\left(P(X_i \le x, X_j \le x) - F_i(x)F_j(x)\right)$$

$$(c)\ P(\left|F_n(x) - \bar{F}_n(x)\right| > \varepsilon) \to 0 \text{ as } n \to \infty \text{ for any } \varepsilon > 0 \text{ provided } Var(F_n(x)) \to 0.$$

The double sum in (b) can be worked out exactly for many models of dependence.

Some references:

Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, **63**, 421–434.

Doksum, K. A. (1974). Empirical probability plots and statistical inference for nonlinear models in the two sample case. *Ann. Statist.* **2**, 267–77.

Wilks, M. B. &. Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55**, 1–17.