STAT 135, Concepts of Statistics

Helmut Pitters

Goodness of fit: further techniques.

Department of Statistics
University of California, Berkeley

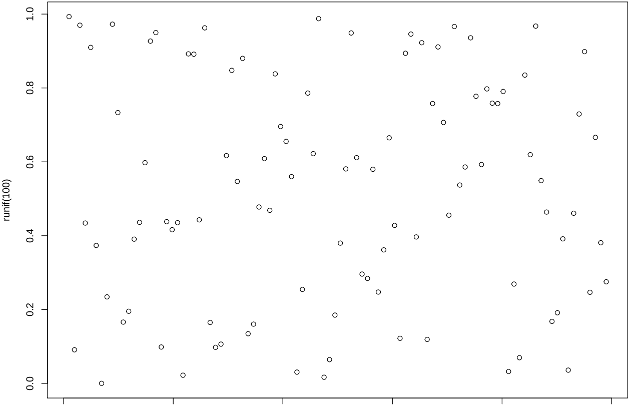
March 29, 2016

Goodness of fit: Probability plots.

Consider independent random samples $X_1, \dots, X_n$ that we conjecture to have uniform$[0,1]$ distribution.

We are interested in a graphical method that allows to check qualitatively whether our conjecture is at all reasonable.

Figure shows a plot of the pairs $(k/100, X_k)$ for $1 \le k \le 100$.



Goodness of fit: Probability plots.

Order the $X_1, \ldots, X_n$ in increasing order to obtain order statistics

$$X_{(1)} < \cdots < X_{(n)}.$$[1]

Recall from Stat 134: $X_{(k)} \sim \mathrm{beta}(k, n-k+1)$, in particular

$$\mathbb{E}X_{(k)} = \frac{k}{n+1}.$$

The points

$$(\frac{k}{n+1}, X_{(k)}) \qquad (1 \leq k \leq 100)$$

should be spread close to their averages

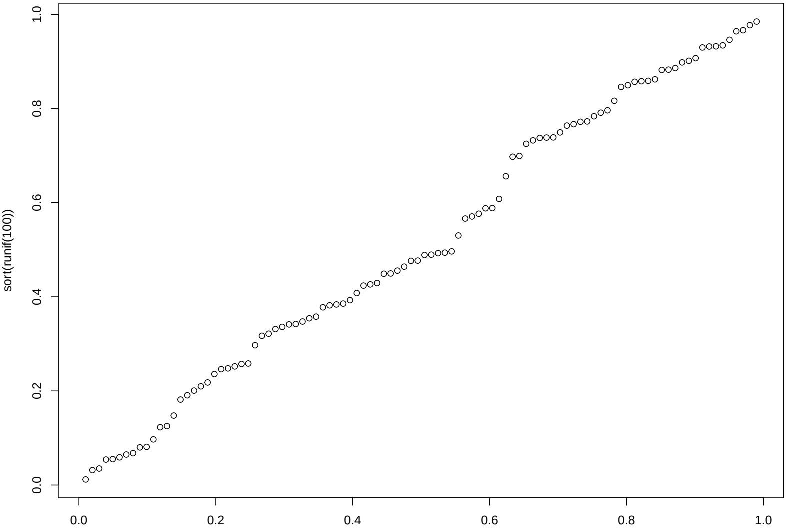
$$(\frac{k}{n+1}, \mathbb{E}X_{(k)}) = (\frac{k}{n+1}, \frac{k}{n+1})$$

which lie on the straight line through $(0,0)$ with slope 1.

[1] In particular $X_{(1)} = \min(X_1, \ldots, X_n), X_{(n)} = \max(X_1, \ldots, X_n)$.

Goodness of fit: Probability plots.

Figure shows a plot of the pairs $(k/101, X_{(k)})$ for $1 \leq k \leq 100$.



Goodness of fit: Probability plots.

What if the common distribution of $X_1, \ldots, X_n$ is not uniform$[0,1]$?
A simple observation allows us to extend this graphical method to general distributions on $\mathbb{R}$.
Let X denote a real r.v. with *cumulative distribution function (cdf)*

$$F(t) := \mathbb{P}\{X \leq t\} \qquad (t \in \mathbb{R}).$$

Goodness of fit: Probability plots.

Lemma

If X is continuous and F is strictly increasing, then $F(X)$ is a real random variable with distribution uniform$[0, 1]$.

Proof.

Notice first that $0 \leq F(t) \leq 1$ for all $t \in \mathbb{R}$. Moreover, for $0 \leq u \leq 1$

$$\mathbb{P}\{F(X) \leq u\} = \mathbb{P}\left\{X \leq F^{-1}(u)\right\} = F(F^{-1}(u)) = u,$$

where F^{-1} denotes the right inverse of F. $\square$

Goodness of fit: Probability plots.

Back to our problem: random sample $X_1, \ldots, X_n$, and we conjecture that their common cdf is F.
Provided our conjecture is true, the $F(X_1), \ldots, F(X_n)$ are i.i.d. uniform$[0,1]$. Consequently, the points

$$(\frac{k}{n+1}, F(X_{(k)})) \qquad (1 \leq k \leq n)$$

should be spread close to a linear function.
Alternatively, can plot

$$(F^{-1}(\frac{k}{n+1}), X_{(k)}) \qquad (1 \leq k \leq n)$$

where F^{-1} denotes the right inverse of F.

Goodness of fit: Probability plots.

Probability plots of theoretical distributions against each other

- uniform-uniform
- uniform-normal
- normal-normal
- normal-uniform

Probability plots of data against theoretical distributions

- percentage of manganese in iron (data: manganese.txt)
- strength of Kevlar/epoxy, a material used in space shuttle (kevlar.txt)
- Michelson's measurements of light speed (michelson.txt)
- precipitation in Illinois storms (illinois60.txt, ..., illinois64.txt)