PROBABILITY PLOTS

Many tests or other procedures in statistics assume a certain (e.g., normal) distribution. Some procedures are *robust* (i.e., still work pretty well) to some departures from assumptions, but often not to dramatic ones.

This raises the question: How to judge whether data come from a given distribution?

Histograms don't serve this purpose well -- e.g., bin sizes, samples sizes, and their interaction cause problems.

Probability plots (also known as *Q-Q plots* or *quantile plots*) are not perfect, but πsomewhat better. The idea:

- Order the data: $y_1 \le y_2 \le \ldots \le y_n$.
- Compare them with $q_1 \le q_2 \le \ldots \le q_n$, where

  q_k = the expected value (as approximated by computer) of the kth smallest member of a simple random sample of size n from the distribution of interest.

If the data come from this distribution, we expect $y_k \approx q_k$, so the graph will lie approximately along the line y = x.

Variation often used to test for normality:

Take the q_k's from the *standard normal* distribution. So if the y_k's are sampled from an $N(\mu,\sigma)$ distribution, then the transformed data $\frac{y_k - \mu}{\sigma}$ comes from a standard normal distribution, so we expect

$$\frac{y_k - \mu}{\sigma} \approx q_k$$

In other words, if the y_k's are sampled from an $N(\mu,\sigma)$ distribution, then

$$y_k \approx \sigma q_k + \mu,$$

so the graph should lie approximately on a straight line with slope and intercept σ and μ, respectively.