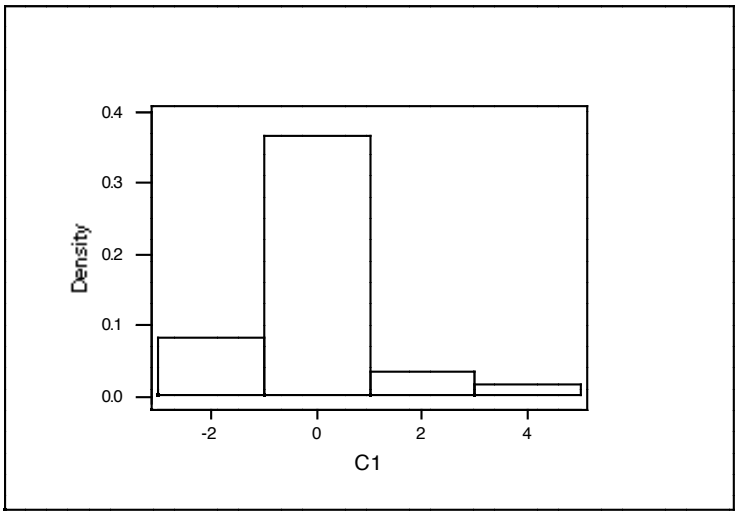
NORMAL QUANTILE PLOTS (Supplement to Section 1.3)

The problem: How do we check if it is reasonable to assume that a certain sample comes from a normal distribution? This will be important when we get to later chapters, because the theory behind many statistical procedures assumes that the data come from a normal distribution. Fortunately, many of them are *robust* (i.e., still work pretty well) if the distribution is only close to normal. But how do we tell even this?

Histograms aren't good enough. For example, here is a density histogram of a random sample of size 30 chosen from a standard normal distribution, using the default bins from Minitab. It doesn't look very normal.



Normal quantile plots aren't perfect, but work better than histograms. Here is 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 k^{th} smallest member of a simple random sample of size n from a normal distribution with the same mean and standard deviation.

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.

However, this idea doesn't work unless we know what mean and standard deviation to look for. We can get around this by using the idea of standardized normal variables:

Take the q_k's from the *standard normal* distribution. (These q_k's are what the book calls "the z-scores at these same percentiles.") So if the y_k's are sampled from a normal distribution with mean μ and standard deviation σ, then the standardized data $z_k = \dfrac{y_k - \mu}{\sigma}$

come from a standard normal distribution, so we expect (as above) $z_k \approx q_k$, or in other words, we should have

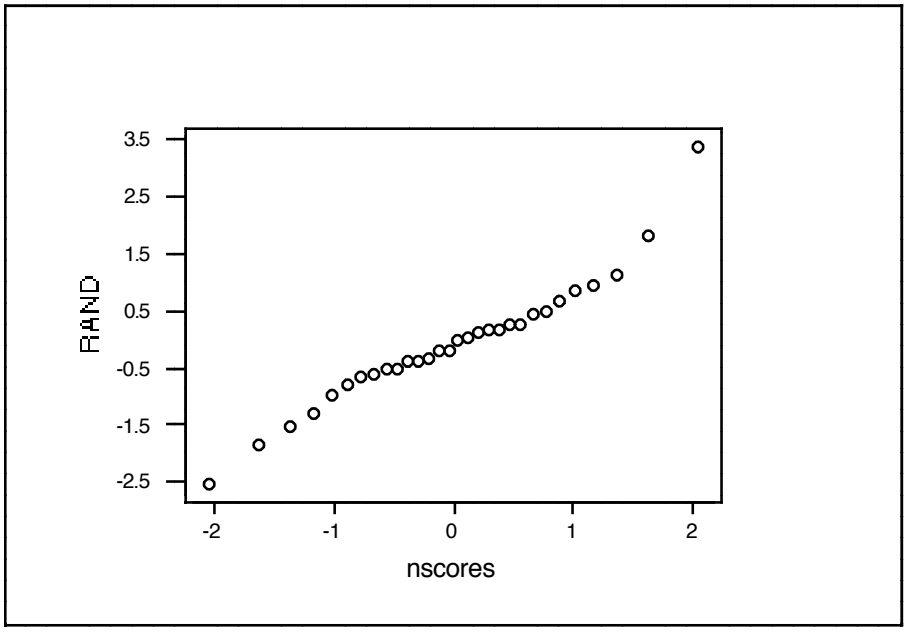
$$\frac{y_k - \mu}{\sigma} \approx q_k$$

Using a little algebra, this says that if the y_k's are sampled from a normal distribution with mean μ and standard deviation σ,, then

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

so the graph of the points (y_k, q_k) should lie approximately on a straight line with slope and intercept σ and μ, respectively.

Here is a normal quantile plot of the same sample used in the hitogram above:



Notice that most of the points lie near a straight line. The point at the right that is noticeably off the line is the point that also gave the bar farthest to the right in the density histogram. Notice that its value, 3.5, is quite unusual for a standard normal distribution.