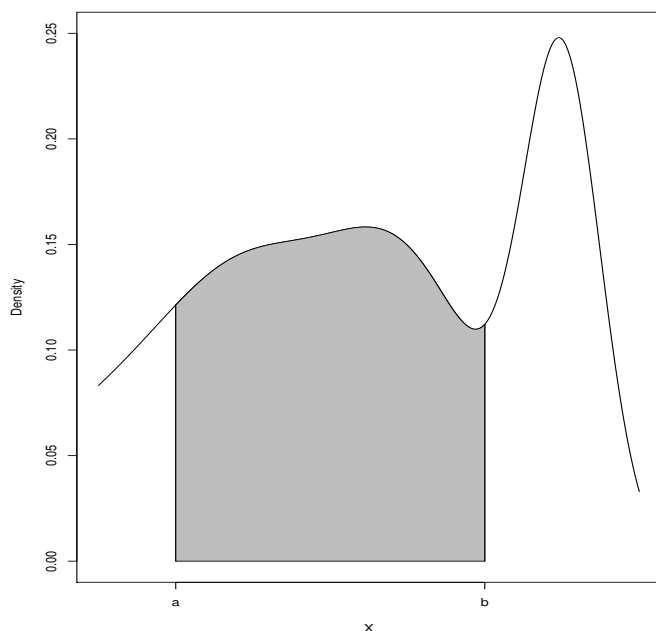


## Continuous distributions

In contrast to discrete random variables, like the Binomial distribution, in many situations the possible values of a random variable cannot be counted. For example, the measurement of temperature of a room is only limited by the accuracy of a thermometer, and the measurement of a person's height is only limited by the accuracy of a ruler. For random variables of this type, the probability of a particular number is 0:

$$P(\text{Temperature} = 70.00000 \dots) = 0.$$

To handle this situation, we define the *density function*, where probabilities are defined as areas under a curve:



## The Normal (Gaussian) distribution

The most important continuous random variable (indeed, the most important random variable of any type) is the **Gaussian** random variable. Carl Friedrich Gauss first described most of its properties in the early 1800's. Sir Francis Galton was the first person to call it the **normal** distribution in 1889 (this was a bit unfortunate, since the name carries the implication that nonnormal distributions are somehow abnormal).

The Gaussian random variable is so important for two reasons:

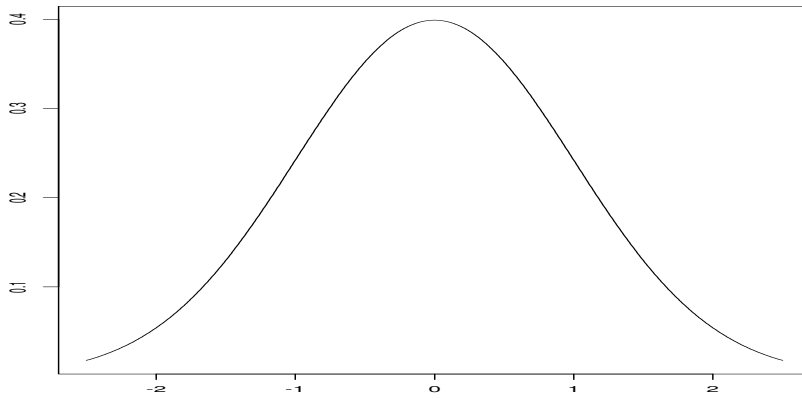
- (1) It is often a good approximation for random processes. That is, in many situations the distribution of random variables is roughly Gaussian.
- (2) For large samples, many statistics are roughly Gaussian (this is the statement of the Central Limit Theorem). This is important because it means that we will be able to make inferences about unknown population parameters using sample statistics without making many strong assumptions.

In fact, the second reason above is related to the first. Many statistics take the form of sums of things (think of  $\bar{X}$  or  $s^2$ , for example), and the Central Limit Theorem says that almost all sums of things become normally distributed as the number of items being summed gets larger. Many other random variables are also sums of things, which is why they appear normally distributed. The annual rainfall in New York City is the sum of hourly rainfalls for the entire year, so it's not surprising that it looks Gaussian; a person's height at age 40 is the sum of all of the increments in height they've had during their lives, so it too looks Gaussian.

The density function for a Gaussian random variable has three important properties:

- (1) It is symmetric about its mean (that is, the density function is a mirror image of itself around its mean).
- (2) It is completely determined by its mean  $\mu$  and its variance  $\sigma^2$ .
- (3) It extends to  $\pm\infty$ , never touching zero (although it gets vanishingly close to zero).

Here is a picture of the Gaussian density function:



This density function equals

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

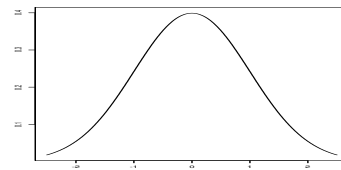
Calculating probabilities would require integrating this function. Fortunately, the normal  $(\mu, \sigma^2)$  distribution can be *standardized* to the **standard normal** ( $\mu = 0, \sigma^2 = 1$ ) by

$$Z = \frac{X - \mu}{\sigma}.$$

Consider the following example. SAT verbal scores are normally distributed with mean  $\mu = 500, \sigma = 100$ .

- (a) What is the probability of scoring less than 600?

$$X_1 = 600: Z_1 = \frac{600-500}{100} = 1$$



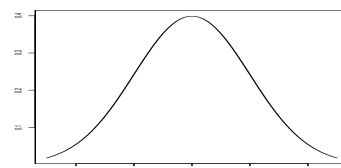
So it equals the probability of a standard normal being less than 1

$$\Rightarrow \text{prob} = .8413$$

- (b) What is the probability of scoring between 400 and 700?

$$X_1 = 400: Z_1 = \frac{400-500}{100} = -1$$

$$X_2 = 700: Z_2 = \frac{700-500}{100} = 2$$

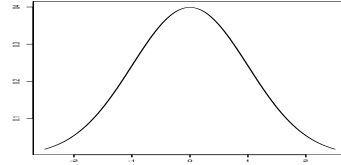


So it equals the probability of a standard normal being between  $-1$  and  $2$

$\Rightarrow$  prob =

(c) What is the probability of scoring greater than 700?

$$X_1 = 700: Z_1 = \frac{700-500}{100} = 2$$



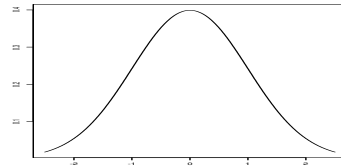
So it equals the probability of a standard normal being greater than 2

$\Rightarrow$  prob =

(d) What SAT score corresponds to the 75<sup>th</sup> percentile of scores (third quartile)?

$$z_{.25} = .675: .675 = \frac{X-500}{100}$$

$$\Rightarrow X = 500 + 67.5 = 567.5$$



The normal distribution provides a simple connection between probability and the standard deviation  $\sigma$ . Note that

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= P\left(\frac{\mu - \sigma - \mu}{\sigma} < Z < \frac{\mu + \sigma - \mu}{\sigma}\right) \\ &= P(-1 < Z < 1) \\ &= .6826 \\ &\approx \frac{2}{3} \end{aligned}$$

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &= P\left(\frac{\mu - 2\sigma - \mu}{\sigma} < Z < \frac{\mu + 2\sigma - \mu}{\sigma}\right) \\ &= P(-2 < Z < 2) \\ &= .9554 \\ &\approx .95 \end{aligned}$$

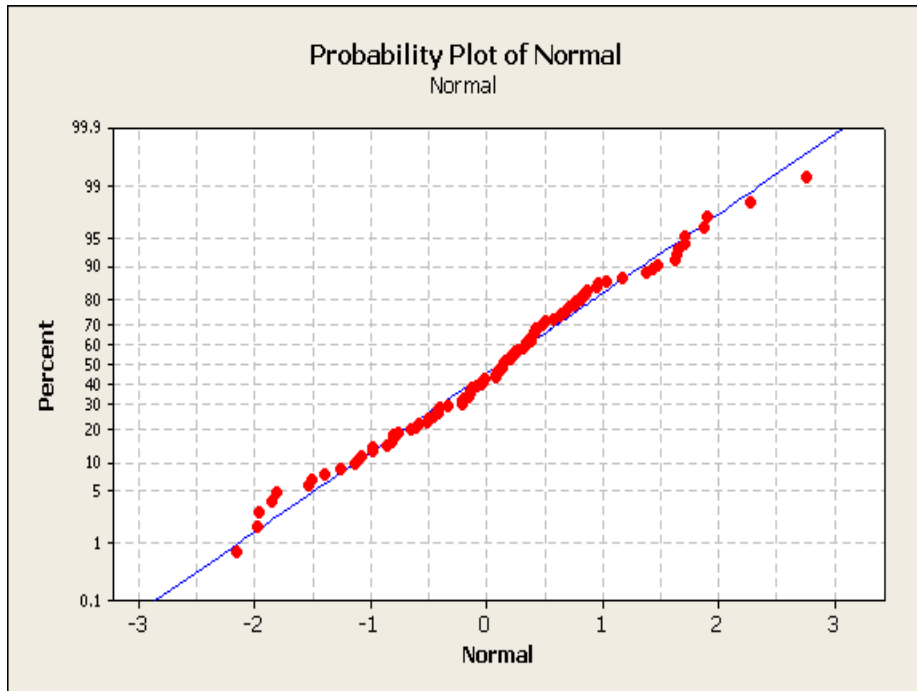
That is, for normally distributed data, about  $\frac{2}{3}$  of the time the value falls within 1 standard deviation of the mean; about 95% of the time the value falls within two standard deviations of the mean.

*Example.* Historical data suggest that the daily change in the Dow Jones Industrial Average is normally distributed, with standard deviation about 1.5% of the DJIA value. In a “flat” market, the mean daily change is zero.

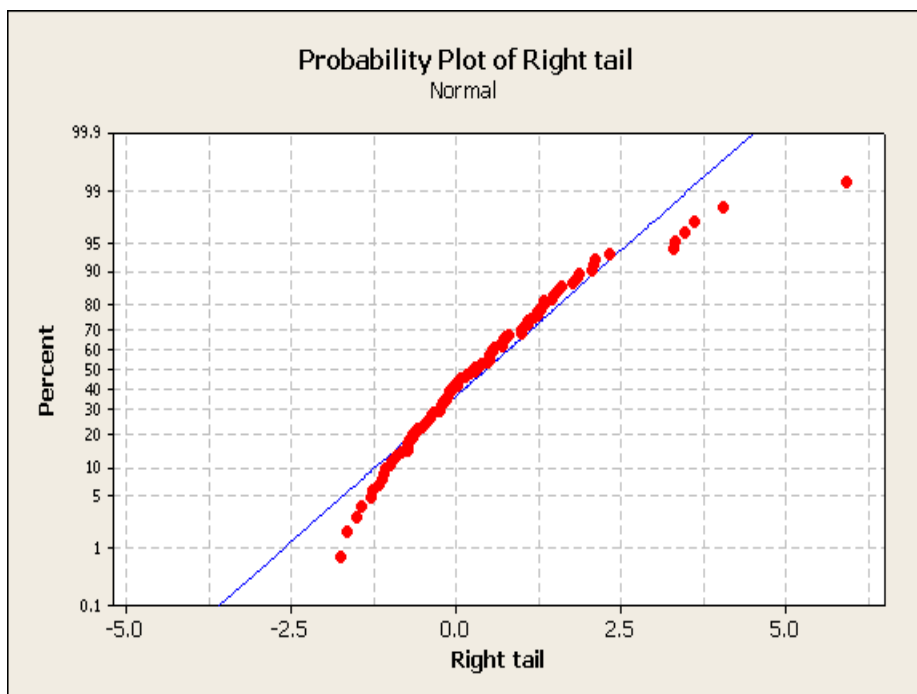
- (a) If the current DJIA is 11000, and the market is currently flat, what is the probability that tomorrow’s closing DJIA will be over 11100?
- (b) On January 1, 1995, the DJIA was 3834. What would the probability have been of at least a 100 point gain then assuming a flat market?

A key point here, of course, is that these probability calculations are only appropriate if the underlying random variable really is normally distributed. Is there any way to check that based on a sample from that random variable? A histogram gives some guidance, but isn’t really the answer, since it is very dependent on the number and position of the bin edges. There are better estimates of density functions (see my book *Smoothing Methods in Statistics*), but generally speaking, people are not good at assessing the shapes of curves. They are, on the other hand, very good at assessing the straightness of a line, and that is the idea behind *probability plots*. A probability plot looks like a scatter plot of two variables, but it is actually a plot of only one variable. The ordered values (termed the *order statistics* in the variable (from smallest to largest) are plotted along the horizontal axis, and the values that would have been expected for that order statistic if the data came from the specified distribution are plotted along the vertical axis. If the data do come from that distribution, the plotted points should look like a straight line. Of course, the most important probability plot is the normal plot, which is used to assess normality of a variable.

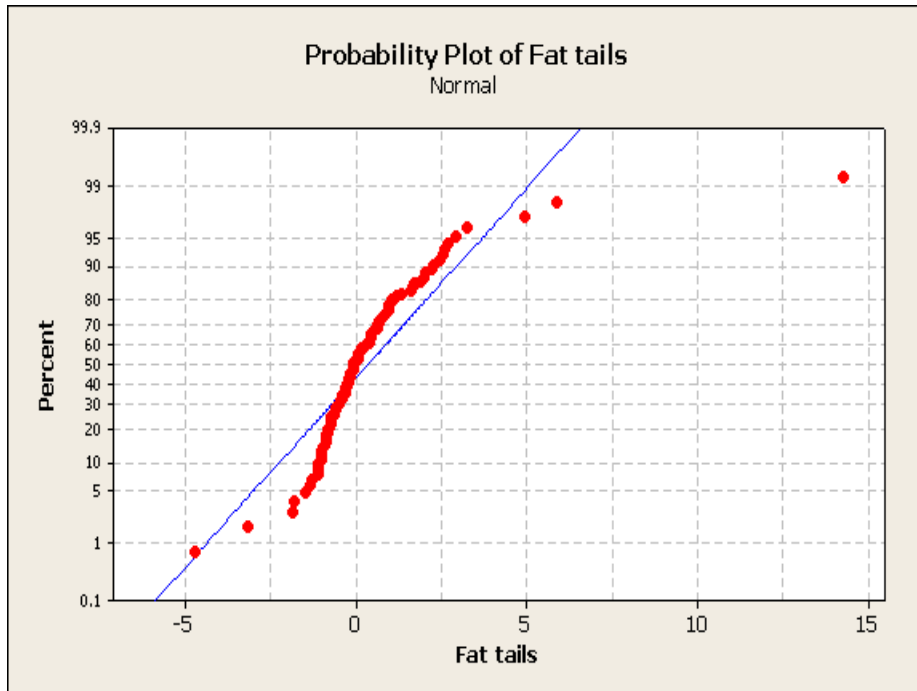
Here is a normal plot for a sample drawn from a normal distribution. Note that while the points don’t follow the expected (blue) line exactly (as they shouldn’t, because of random fluctuation), the line is reasonably straight, indicating no problem with a normality assumption.



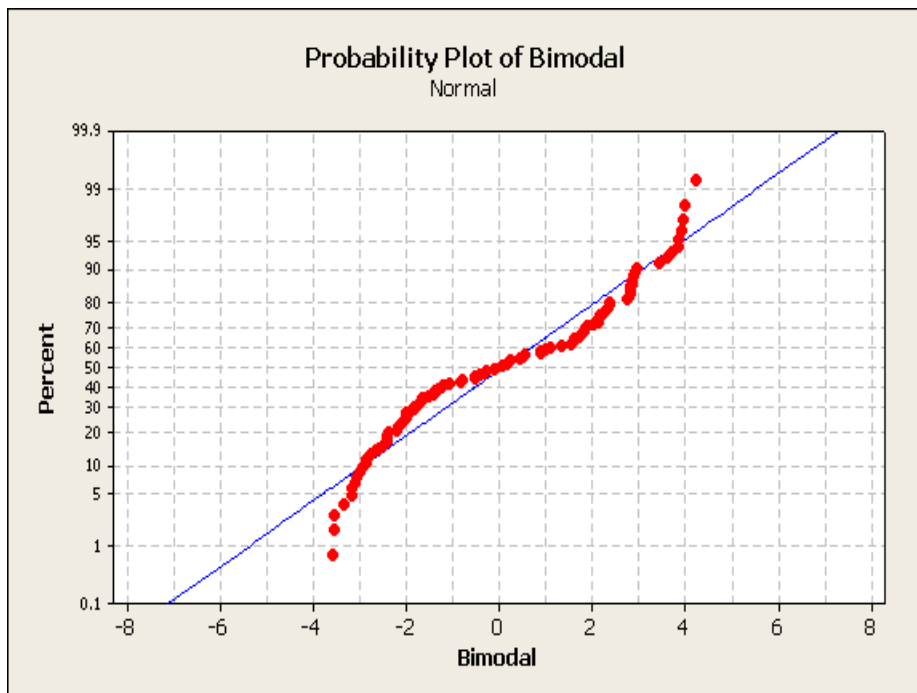
Now consider the following plot:



It is clear that the points don't follow a straight line here. The observed points have a tail that is shorter than the normal on the left, and longer than the normal on the right; that is, this is right-tailed data. Now consider the following plot:



For these data observations are both “too negative” and “too positive,” implying fat tails in both directions. Finally, consider the following plot:



This plot doesn't exhibit long tails in either direction, but rather shows two parallel lines, suggesting two separate normals; that is, a bimodal distribution.

### **Minitab commands**

To construct a normal plot, click on **Graph** → **Probability Plot**. Enter the variable you are plotting under **Graph variables:** and click **OK**.

The  
normal  
law of error  
stands out in the  
experience of mankind  
as one of the broadest  
generalizations of natural  
philosophy \* It serves as the  
guiding instrument in researches  
in the physical and social sciences and  
in medicine agricultural and engineering \*  
It is an indispensable tool for the analysis and the  
interpretation of basic data obtained by observation and experiment

— W.J. Youden