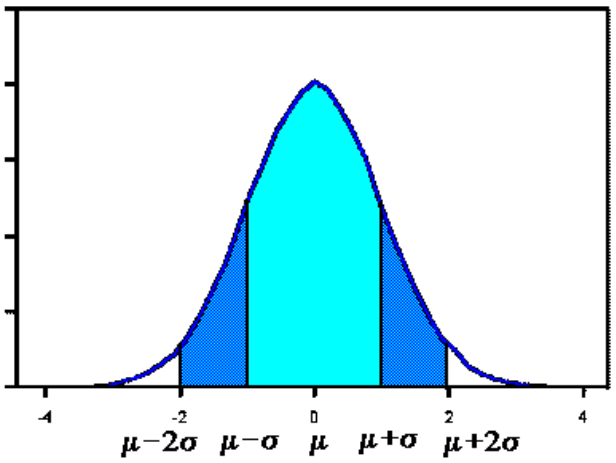
The Normal Distribution

A normal distribution has a bell-shaped density curve described by its mean μ and standard deviation σ. The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation. The height of a normal density curve at a given point x is given by

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



The **Standard Normal** curve, shown here, has mean 0 and standard deviation 1. If a dataset follows a normal distribution, then about 68% of the observations will fall within σ of the mean μ, which in this case is with the interval (-1,1). About 95% of the observations will fall within 2 standard deviations of the mean, which is the interval (-2,2) for the standard normal, and about 99.7% of the observations will fall within 3 standard deviations of the mean, which corresponds to the interval (-3,3) in this case. Although it may appear as if a normal distribution does not include any values beyond a certain interval, the density is actually positive for all values, $(-\infty, \infty)$. Data from any normal distribution may be transformed into data following the standard normal distribution by subtracting the mean μ and dividing by the standard deviation σ.

Example

The dataset used in this example includes 130 observations of body temperature. The MINITAB "DESCRIBE" command produced the following numerical summary of the data:

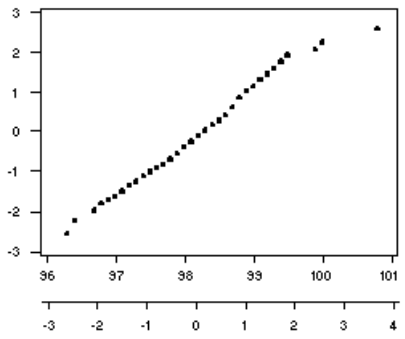
```
Variable         N     Mean   Median  Tr Mean    StDev  SE Mean
BODY TEMP      130   98.249   98.300   98.253    0.733    0.064

Variable       Min      Max       Q1       Q3
BODY TEMP   96.300  100.800   97.800   98.700
```

The spread of the data is very small, as might be expected.

The normality of the data may be evaluated by using the MINITAB "NSCORES" command to calculate the normal scores for the data, then plotting the observed data against the normal quantile values. For the first 10 sorted observations, the table below displays the original temperature values in the first column, standardized values in the second column (calculated by subtracting the mean 98.249 and dividing by the standard deviation 0.733), and corresponding normal scores in the third column.

```
96.3    -2.65894       -2.58163
96.4    -2.52251       -2.24352
96.7    -2.11323       -1.98066
96.7    -2.11323       -1.98066
96.8    -1.97681       -1.80820
96.9    -1.84038       -1.71725
97.0    -1.70396       -1.63847
97.1    -1.56753       -1.50561
```

```
97.1     -1.56753        -1.50561
97.1     -1.56753        -1.50561
```



The standardized values in the second column and the corresponding normal quantile scores are very similar, indicating that the temperature data seem to fit a normal distribution. The plot of these columns, with the temperature values on the horizontal axis and the normal quantile scores on the vertical axis, is shown to the right (the two scales in the horizontal axis provide original and standardized values). This plot indicates that the data appear to follow a normal distribution, with only the three largest values deviating from a straight diagonal line.

Data source: Derived from Mackowiak, P.A., Wasserman, S.S., and Levine, M.M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick," Journal of the American Medical Association, 268, 1578-1580. Dataset available through the JSE Dataset Archive.

Like any continuous density curve, the probabilities of observing values within any interval on the normal density are given by the area of the curve above that interval. For example, the probability of observing a value less than or equal to zero on the standard normal density curve is 0.5, since exactly half of the area of the density curve lies to the left of zero. There is no explicit formula for that area (so calculus is not of much help here). Instead, the probabilities for the standard normal distribution are given by tabulated values (found in Table A in Moore and McCabe or in any statistical software).

To compute the probability of observing values within an interval, one must subtract the cumulative probability for the smaller value from the cumulative probability for the larger value. Suppose, for example, we are interested in the probability of observing values within the standard normal interval (0,0.5). The probability of observing a value less than or equal to 0.5 (from Table A) is equal to 0.6915, and the probability of observing a value less than or equal to 0 is 0.5. The probability of the normal interval (0, 0.5) is equal to 0.6915 - 0.5 = 0.1915.

Example

Assuming that the temperature data are normally distributed, converting the data into standard normal, or "Z," values allows for the calculation of cumulative probabilities for the temperatures (the probability that a value less than or equal to the given value will be observed). These data are standardized by first subtracting the mean, 98.249, and then dividing by the standard deviation, 0.733. In MINITAB, the "CDF" command calculates the cumulative probabilities for standard normal data, or the probability that a value less than or equal to a given value will be observed. Here are some of the body temperature observations, their normalized values, and their relative frequencies:

```
VALUE          Z-VALUE        CDF
96.7           -2.11302       0.017299
98.0           -0.33993       0.366955
98.3             0.06924       0.527603
98.5             0.34203       0.633835
98.8             0.75120       0.773735
99.9             2.25151       0.987823
```

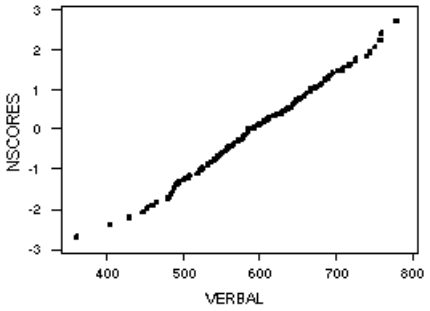
The values below the observed mean, 98.249, have negative standardized values and relative frequencies less than 0.5, while values above the mean have positive standardized values and relative frequencies greater than 0.5. Notice that the probability of observing a value smaller than 96.7 is very small, as is the probability of observing a value greater than 99.9 (this probability is 1- (the probability of observing a value less than 99.9) = 1-0.9878 = 0.0122). Both of these values lie outside of the (-2,2) interval, which includes 95% of the data in a standard normal distribution.

RETURN TO MAIN PAGE.

The chi-square goodness of fit test may also be applied to continuous distributions. In this case, the observed data are grouped into discrete bins so that the chi-square statistic may be calculated. The expected values under the assumed distribution are the probabilities associated with each bin multiplied by the number of observations. In the following example, the chi-square test is used to determine whether or not a normal distribution provides a good fit to observed data.

Example

The MINITAB data file "GRADES.MTW" contains data on verbal and mathematical SAT scores and grade point average for 200 college students. Suppose we wish to determine whether the verbal SAT scores follow a normal distribution. One method is to evaluate the normal probability plot for the data, shown below:



The plot indicates that the assumption of normality is not unreasonable for the verbal scores data.

To compute a chi-square test statistic, I first standardized the verbal scores data by subtracting the sample mean and dividing by the sample standard deviation. Since these are estimated parameters, my value for *d* in the test statistic will be equal to two. The 200 standardized observations are the following:

```
  [1] -2.11801 -2.69073  0.76066  1.04702  0.91138 -0.09842  0.23316  1.04702  0.65516  0.
 [11] -0.53549 -1.39457 -0.58071  0.77573 -0.58071  0.47430  0.05230 -2.25365 -0.21899 -0.
 [21] -0.30942 -0.38478  0.23316  1.12238  1.45396  0.05230 -0.67114 -1.25893  1.12238  0.
 [31]  1.19774 -0.58071  0.50445  1.92118 -0.67114  0.05230  0.36880 -0.23406 -0.73142  0.
 [41] -1.54529  1.55946  0.03723  0.21809  0.21809 -0.71635 -1.39457 -1.81658  0.98674 -0.
 [51]  0.17287  0.64009  0.33866 -3.14288  1.19774  0.47430  1.92118 -0.17378  0.77573  0.
 [61]  0.64009  0.91138  1.33338 -0.17378  0.33866 -0.67114 -0.53549 -0.29435 -0.95750  0.
 [71]  0.47430 -0.03813 -0.53549  0.29344  0.36880  0.21809  0.12766  1.31831  2.26782  0.
 [81] -1.24386  1.83075  1.04702  1.58960  0.03723  0.33866 -0.30942 -0.58071 -0.71635 -0.
 [91] -0.03813  0.83602  0.27837  0.77573 -0.03813 -1.00271 -0.85200 -0.73142 -0.29435  0.
[101] -0.09842 -0.71635  0.23316 -1.15343  1.04702 -0.71635 -2.02758  1.27310  0.05230  0.
[111]  1.72524 -0.67114 -0.71635 -0.71635  0.68531  1.86089  0.91138 -1.40965  0.09751 -0.
[121]  0.64009 -0.06827 -0.53549  0.36880 -2.40437  1.99653 -1.12329  0.41402  1.12238 -0.
[131] -0.61085  0.91138 -0.38478 -1.33429 -0.47521  0.91138  0.76066 -0.09842 -0.44506 -1.
[141] -0.35463 -0.44506 -0.42999  0.23316 -0.21899  0.91138 -0.23406  0.09751  0.50445 -0.
[151] -0.98764 -1.12329  0.23316 -0.95750  1.48410 -0.17378 -1.39457 -0.85200 -0.58071  1.
[161] -0.42999  1.19774  0.54966 -1.12329  1.45396 -0.30942  0.18794 -0.86707 -0.38478 -1.
[171] -0.09842 -1.42472  1.31831 -0.71635 -1.83165  2.26782 -0.00799 -1.12329 -0.42999  1.
[181] -1.86179 -1.10821  0.41402  1.31831  0.64009  1.12238  0.48937 -0.00799 -0.30942 -0.
[191]  1.72524 -1.10821 -0.38478  0.41402 -0.03813 -1.68093 -1.86179  0.33866  2.20754  0.
```

I chose to divide the observations into 10 bins, as follows:

```
Bin             Observed Counts
(< -2.0)                  6
(-2.0, -1.5)              6
(-1.5, -1.0)             18
(-1.0, -0.5)             33
```

(-0.5, 0.0)	38
(0.0, 0.5)	38
(0.5, 1.0)	28
(1.0, 1.5)	21
(1.5, 2.0)	9
(> 2.0)	3

The corresponding standard normal probabilities and the expected number of observations (with n=200) are the following:

Bin	Normal Prob.	Expected Counts	Observed - Expected	Chi-Value
(< -2.0)	0.023	4.6	1.4	0.65
(-2.0, -1.5)	0.044	8.8	-2.8	-0.94
(-1.5, -1.0)	0.092	18.4	-0.4	-0.09
(-1.0, -0.5)	0.150	30.0	3.0	0.55
(-0.5, 0.0)	0.191	38.2	-0.2	-0.03
(0.0, 0.5)	0.191	38.2	-0.2	-0.03
(0.5, 1.0)	0.150	30.0	-2.0	-0.36
(1.0, 1.5)	0.092	18.4	2.6	0.61
(1.5, 2.0)	0.044	8.8	0.2	0.07
(> 2.0)	0.023	4.6	-1.6	-0.75

The chi-square statistic is the sum of the squares of the values in the last column, and is equal to 2.69.

Since the data are divided into 10 bins and we have estimated two parameters, the calculated value may be tested against the chi-square distribution with 10 -1 -2 = 7 degrees of freedom. For this distribution, the critical value for the 0.05 significance level is 14.07. Since 2.69 < 14.07, we do not reject the null hypothesis that the data are normally distributed.