5. INTRODUCTION TO THE NORMAL DISTRIBUTION

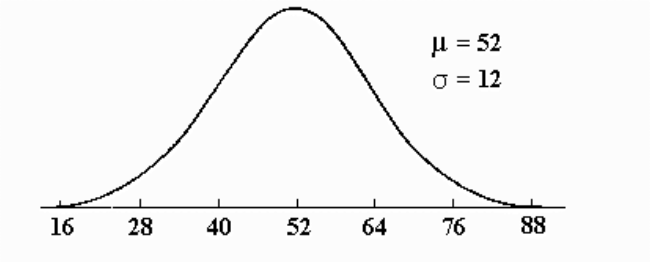
The Normal Curve

The normal distribution, the theoretical "bell-shaped curve", is one of the most important continuous distributions in statistics. This is because many types of data, especially biological data, have a distribution that is approximately normal in the population. Even when a variable is not normally-distributed, the distribution of sample means is approximately normal if sample size is sufficiently large (Central Limit Theorem). These facts have been used to great advantage in the development of methods for analysing biological data.

This week we will use JMP IN to take random samples from normal and non-normal distributions, calculate area (probability) under the normal curve, and test the goodness of fit between real data and the theoretical normal curve. The equation for the normal curve is (don't memorize this, please):

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

X is a continuous variable with mean μ and standard deviation σ. $P(X)$ is "probability density" rather than frequency, and probability is given by area under the curve (rather than height of the curve). Here is an example of the normal distribution:



The normal distribution is symmetric, centred over the mean, with tails that extend to positive and negative infinity.

The existence of an infinite number of normal distributions, each one with a different standard deviation and mean, would make it difficult to calculate areas under the normal curve. We solve this problem by converting X for any normal distribution to the **standard normal distribution** Z having a mean of 0 and standard deviation of 1. The formula for converting is:

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

Areas under the standard normal curve are provided in statistical tables in the back of your textbook and are those given by JMP IN.

Using the Standard Normal Random Number Generator

To sample from the standard normal distribution you will again need to use JMP IN's calculator. Open a new data table and then use **Add Rows** to add rows to the first data column. Select **Column Info** and then **New Property -> Formula** (a shortcut is to select **Formula** directly instead of **Column Info**). In the calculator window that pops up, Select **Random -> Random Normal** in the **Functions** box. This will generate samples from a standard normal distribution (having a mean of 0 and a standard deviation of 1). To generate random samples from a normal distribution with mean μ=10 and standard deviation σ=5, apply the above conversion formula in reverse (i.e., $X = Z\sigma + \mu$), by typing **Random Normal() × 5 + 10** into the formula window (take care to put the 5 and 10 in the correct position, otherwise you might end up with the standard normal ×15).

Obtaining Probabilities Under the Normal Curve

JMP IN can calculate exact probabilities under the normal curve. Open a new data set and create a single row. Then choose **Formula** from the columns menu to open the calculator window. Under **Functions** choose **Probability -> Normal Distribution**. Enter a number between the brackets in the formula (e.g., 1.96) and click "Apply". The row you created in the data table will contain Prob($Z \leq 1.96$).

Evaluating Fit to the Normal Distribution

Because the normal distribution is an assumption of so many methods for analysing data, a way to evaluate the assumption is needed. JMP IN has two visual tools for assessing the goodness of fit between the data and a normal distribution.

The first tool is the simplest, and involves comparing the histogram of the data with the normal curve having the same mean and variance as the data. Click the "▼" symbol next to the variable name above the histogram and choose **Fit Distribution -> Normal**. This will result in a normal curve superimposed on the histogram, to allow a visual comparison of shape. To supplement the visual impression, get JMP IN to calculate skewness and kurtosis of the data by clicking the "▼" symbol again and selecting **Display Options -> More Moments**. The normal curve has zero skew and kurtosis, and departures from zero in the data inform us about departures from normality.

The second tool is the normal quantile plot. This plot is easier to explain if you have chosen the horizontal layout (Click the "▼" symbol next to the variable name above the histogram and choose **Display Options -> Horizontal Layout**). Then click the "▼" again and choose **Normal Quantile Plot**. This will generate a new plot next to the histogram that compares quantiles of the data on the *X*-axis with *Z*-values corresponding to each quantile of the normal distribution on the *Y*-axis. The plot is basically a cumulative relative frequency distribution, with which you are already familiar, but here cumulative relative frequency (given on the *Y* axis; see the numbers ranging from 0.01 to 0.99 along the inside edge of this axis) is plotted on a normal probability scale. The numbers on the outside edge of the *Y*-axis are *Z*-values corresponding to successive values for cumulative relative frequency. If the data are normally distributed, then the points in the figure will lie on a straight line. Departures from a straight line indicate departures from normality.

JMP IN will also carry out a goodness of fit test of normality. To carry out this test you will first need to **Fit Distribution -> Normal** as explained above. Then go to the results and click the "▼" symbol next to the **Fitted Normal** heading and select **Goodness of Fit**. For very large samples (>2000), JMP IN carries out the Kolmogorov-Smirnov-Lilliefors test, which is a modification for the normal distribution of the familiar KS test (recall this test compares the goodness of fit between observed and expected cumulative relative frequency distributions). For smaller samples JMP IN provides the Shapiro-Wilks test, which tests the adequacy of a linear fit in the normal quantile plot.

Problems

1. Generate a random sample of 1000 individuals from a normal distribution with mean 12 and standard deviation 4 (call it "Norm"). Begin with 1000 rows.

   a) Plot a histogram and boxplot of the data and describe the distribution. Is the sample distribution symmetric? Does it have any outliers?

   b) Note the mean and standard deviation of the distribution. Do they match the values you supplied when generating the random numbers? Also obtain values for skewness and kurtosis. Are the last two values close to those expected for a normal distribution?

   c) Generate a normal quantile plot for the variable Norm. Is the fit linear?

   d) Carry out the Shapiro-Wilks test of normality. State clearly the null and alternate hypotheses for this test. What is the probability of rejecting Ho?

2. Determine the following probabilities under the normal curve.

   a) What is the probability of obtaining a *Z* value less than or equal to -1.00?

   b) What is the probability of obtaining a *Z* value less than or equal to -1.96?

   c) What is the probability of obtaining a *Z* value greater than or equal to 2.50? What is the probability of obtaining a Z value greater than 2.50?

   d) What is the probability of obtaining a *Z* value greater than -0.65?

   e) What is the probability of obtaining a *Z* value between -2.3 and 0.7?

   f) What is the probability of obtaining a *Z* value less than –1.2 **or** greater than 0.2?

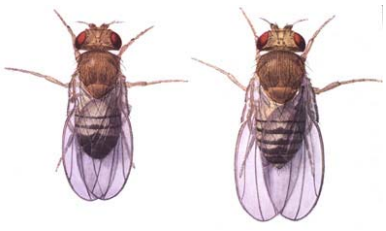
   g) What is the probability of obtaining a *Z* value less than –1.2 **and** greater than 0.2?

   h) Using **Probability->Normal Quantile**, the normal quantile function, what value of *Z* corresponds to an area of 0.05 on the left tail of the standard normal distribution?

   i) What value of *Z* corresponds to an area 0.01 in the upper (right) tail of the standard normal distribution?

   j) What values of *Z* correspond to a total area of 0.25 spread evenly between both tails?

3. Open the **fruitflies.jmp** file in the shared directory. These data were collected by Partridge (1981, Nature 294: 580-581) to test whether male flies suffered a survival cost from mating. The life spans of individual males supplied with 1 or 8 receptive virgin females per day were compared with life spans of three types of control males. The first two types of control males were supplied with either 1 or 8 newly inseminated females (newly inseminated females will not re-mate for at least two days, but they control for other effects females have on males (e.g., competition for food). The third type of control males were kept alone (i.e., 0 females were added). The four variables are:



Number of female partners supplied daily to males (0,1 or 8)
Treatment (0, 1 or 8 virgin females, 1 or 8 newly inseminated females)
Male lifespan, in days
Male thorax length, in mm

a) For now, ignore the fact that males are in different treatment groups and consider all of them together. Visually compare the histogram of *male lifespan* with a normal distribution. Is the fit reasonably close? Add the normal quantile plot and reassess the fit of the data to a normal distribution. Is the fit reasonably good?

b) Test whether *male lifespan* fits a normal distribution. Show all steps. If the null hypothesis is not rejected, does this mean the data are from a population having a normal distribution?

c) Repeat steps (3a) and (3b) for the variable *male thorax*. How do the results compare with those for *male lifespan*?

4. In fact the males in the data set **fruitflies.jmp** come from different treatment groups, and treating them as though they constitute a single sample is not valid. We will deal with analysis of multiple treatment groups later in the course. Here we briefly explore the distributions of multiple groups. Use **Distribution** to plot a histogram of lifespan separately for each treatment group (this can be done all at once by selecting the variable *Treatment* in the **By** box of the **Distribution** popup window). Click the "▼" symbol next to the label **Distributions** at the top of the results window and select **Stack** for easier comparison of histograms.

a) Produce a normal quantile plot for lifespan separately for each treatment group. Does each sample conform reasonably well with a normal distribution? Test each fit using the Shapiro-Wilks test. Do any of the groups depart significantly from the normal distribution?

b) Return to the fly data table. This time, use **Fit Y by X** to plot male lifespan (*Y*) against treatment (*X*). This will produce a plot in which the lifespans of males are plotted separately for each treatment group. Click the "▼" symbol at the top of the result window and select **Display Options -> Mean Error Bars**. This illustrates the mean ± 1 standard error for each treatment group. Which treatment group appears to have the shortest mean lifespan (for now, refrain from testing differences)?

c) Click the "▼" symbol again at the top of the result window and select **Display Options -> Box Plots**. This generates a quantile box plot for each group (these are simpler than the outlier box plots you are already familiar with, indicating only the median, quartiles, and range). How informative are the box plots regarding the fit of lifespans to the normal distribution within each treatment group?

d) Click the "▼" symbol again at the top of the result window and select **Normal Quantile Plot -> Plot Actual by Quantile**. This produces a normal quantile plot for each treatment group separately.

5. Open the file **cntrlmt.jmp**. This file has a set of 5 columns, each of which uses random numbers between 0 and 1 that have been raised to the fourth power. This distribution is highly skewed, with most observations lying near 0. The first column of the data table is simply a random sample from this distribution (each value is a random number between 0 to 1 that has been raised to the fourth power). Each row of the second column is a mean of a random sample of n=5 observations from this distribution. The third column reports means of random samples of n=10 observations. The fourth and fifth columns are means of samples of n=50 and n=100 observations. In reality, you will rarely have the opportunity to take multiple samples from a population to examine the distribution of sample means; the idea that you *would* get a different value for the sample mean each time represents a "thought experiment". Here, the computer is used to illustrate the outcome of such a "thought experiment".

a) Add 500 rows to each of the 5 columns. Display histograms for each of the 5 columns and compare their general shapes. Which ones appear to have the best fit to a normal distribution?

b) Compare skew and kurtosis of each of the 5 columns. What happens to these values as sample size increases?

c) Carry out a goodness of fit test to the normal distribution on each of the columns. Which ones are significantly non-normal according to the test?

d) What principle is illustrated by the fact that successive columns provide increasingly better fits to the normal distribution even though the underlying distribution is not normal?