chapter 2-3

Testing Normality

Introduction

In the previous chapters we discussed a variety of descriptive statistics which assume that the data are normally distributed. This chapter focuses upon testing if a distribution is normally distributed and then possible ways of transforming the data in order to have a distribution that better approximates the normal distribution. The two most common deviations from normality, **skewness** and **kurtosis** will be discussed here. Figure 2-3.1 shows the typical shapes of skewed distributions in comparison to the normal distribution. Positively skewed data has a long tail towards the positive or higher scores side of the distribution. This is because there are a few very high scores that are "skewing" the distribution in this direction. In Biomedical Physiology and Kinesiology it is not uncommon to find positively skewed variables. Variables such as skinfold thicknesses and weight are usually positively skewed. Even muscle girths tend to be positively skewed as a few people tend to want to go into the gym and train excessively to produce very large muscles. Negative skewness is not as common in the types of variables we might encounter in Biomedical Physiology and Kinesiology. An obvious example is the height of basketball players in the NBA. There are very few short players in the leagues. Some do exist however, and because there are only a few of them and they are extremely small in comparison to the rest of the players, they cause the distribution to be skewed towards the small side.

Figure 2-3.1: Normal, Positively Skewed and Negatively Skewed distributions

Another form of deviation from normality is Kurtosis. Figure 2-3.2 shows different kurtic distributions. The normal distribution is referred to as Mesokurtic. Rather than asymmetry as described by skewness, kurtosis is a measure of how centrally located the data are within the distribution. In a leptkurtic distribution the data is bunch more towards the centre causing the distribution to look thinner and more peaked. In the Platykurtic distribution, the shape looks

more flattened as the data are more spread out around the centre. Kurtosis is often the forgotten deviation from normality. Researchers will concern themselves with skewness before they will consider kurtosis. That said, skewness is also often overlooked. Weight and skinfold measures are usually skewed, but rarely will researchers correct the problem before applying parametric statistics. You can find thousands of papers in the scientific literature where parametric statistics have been applied to skinfold and weight data, regardless of the skewness. The good news, however, is that although skewness is a violation of the assumption of normality in these parametric tests, the significance of findings is not profoundly affected.

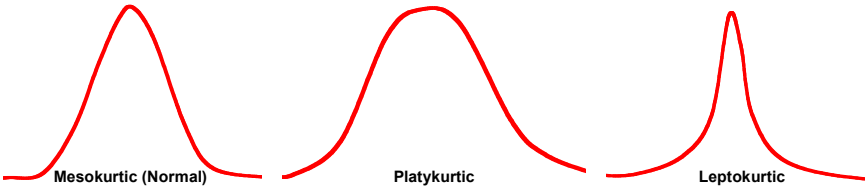


Figure 2-3.2: Mesokurtic (Normal), Platykurtic and Leptokurtic distributions

Coefficient of Skewness

A normal distribution, by definition is symmetrical; that is, the distribution looks the same either side of the centre line. Positive and negatively skewed distributions are asymmetrical. The Coefficient of Skewness quantifies this asymmetry.

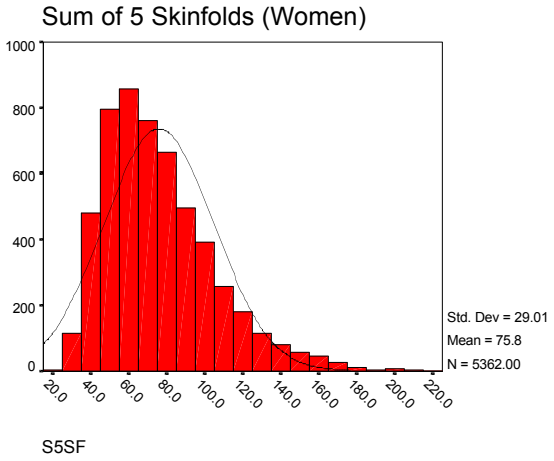
$$Coefficient\ of\ Skewness = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^3}{(N-1)s^3}$$

where $\overline{X}$ is the mean, s is the standard deviation, and N is the number of data points.

If the data are normally distributed the coefficient of skewness is zero. Infact, any symmetric data will have a coefficient of skewness near zero. The sign of the coefficient tells the type of skewness. A positive coefficient of skewness means positive skewness, and the opposite for negative skewness. A coefficient greater than 1 is regarded as significant positive skewness, whereas a coefficient less than -1 is regarded a significant negative skewness. The coefficient of skewness is an option for selection on the SPSS Descriptive statistics dialog box. Figure 2-3.3 shows the SPSS histogram of the Sum of 5 Skinfolds (S5SF) in 5,362 women from the Canada Fitness Survey (CFS) of 1981. The red bars show the distribution of the data whereas a superimposed black line shows a normal distribution with the same mean and standard deviation as the S5Sf data. This superimposed line allows you to visually appraise how deviant from the normal distribution your data are. In these data the distribution is positively skewed. A

quantification of the degree of skewness is seen in the coefficient of skewness listed in the SPSS Descriptive Statistics output for Weight (WT), Height (HT) and Sum of 5 Skinfolds (S5SF) in the same data, also shown in Figure 2-3.3. The coefficient of skewness for S5SF is 1.043 (significantly skewed). Interestingly Height (HT) is not skewed (0.09) but Weight (WT) is more skewed than S5SF with a coefficient of 1.297. The standard error of the coefficient (Std. Error in output) gives your measure of confidence in the coefficient. Coefficient of Skewness ±1.96 x Standard Error of the Coefficient gives the 95% confidence interval of the coefficient. For weight the 95% confidence interval for the coefficient of Skewness would therefore be:

1.297 ±(1.96 x 0.032) = 1.234 to 1.360



Descriptive Statistics

	N	Mean		Std.	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
WT	5704	61.9210	.1474	11.1361	1.297	.032	2.643	.065
HT	5782	161.0457	8.183E-02	6.2225	.092	.032	.090	.064
S5SF	5362	75.7820	.3961	29.0066	1.043	.033	1.299	.067
Valid N (listwise)	5347							

Figure 2-3.3: SPSS Histogram of Sum of 5 Skinfolds (S5SF) in 5362 Females from the Canada Fitness Survey (1981) and SPSS Descriptive Statistics output for Weight (WT), Height (HT) and Sum of 5 Skinfolds (S5SF) in the same data.

Coefficient of Kurtosis

As illustrated in Figure 2-3.2 a Platykurtic distribution is more flattened, while a Leptokurtic distribution is more peaked than the Mesokurtic or Normal distribution. The degree of Kurtosis is quantified by the Coefficient of Kurtosis

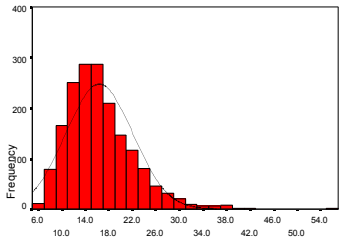
$$Coefficient\ of\ Kurtosis = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^4}{(N-1)s^4}$$

where $\overline{X}$ is the mean, s is the standard deviation, and N is the number of data points.

Normalizing Data

Many statistical tests are based on the assumption of normally distributed data. As discussed previously, many real data sets are in fact not approximately normal. However, an appropriate transformation of a data set can often yield a transformed data set that does follow approximately a normal distribution. This increases the applicability and usefulness of statistical techniques based on the normality assumption. A simple data transformation applicable to moderately positive or right skewed data is the $\log_{10}$ transformation. Figure 2-3.4 shows the frequency distribution for Triceps Skinfold (TPSF) for the CFS data set of 1,765 women aged 20-30 years. The coefficient of skewness shows significant skewness at 1.17 and the histogram illustrates this positive skewness. The lower panel of Figure 2-3.4 shows the distribution of the $\log_{10}$ transform of the data. A new variable was produced ($\log_{10}$TPSF) by calculating the $\log_{10}$ of each TPSF measure. The new distribution is more normally distributed with a coefficient of skewness of 0.02. In this case the transformation worked well; however, it is not perfect for all situations. It tends to work better in moderately rather than extremely skewed data. A better but more complex transform is the Box-Cox transform, which will be described later in this chapter.

TPSF: μ = 16.4 σ= 5.7 Skewness = 1.17



Log_{10}TPSF: μ = 1.19 σ= 0.15 Skewness = 0.02

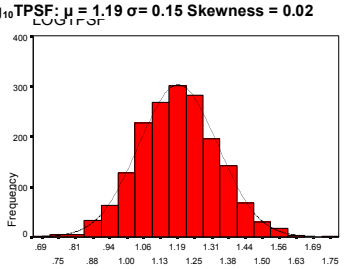


Figure 2-3.4: SPSS Histograms of Triceps Skinfold (TPSF) and $\log_{10}$TPSF in 1,765 females aged 20-30 years from the Canada Fitness Survey (1981)

Normal Probability Plots

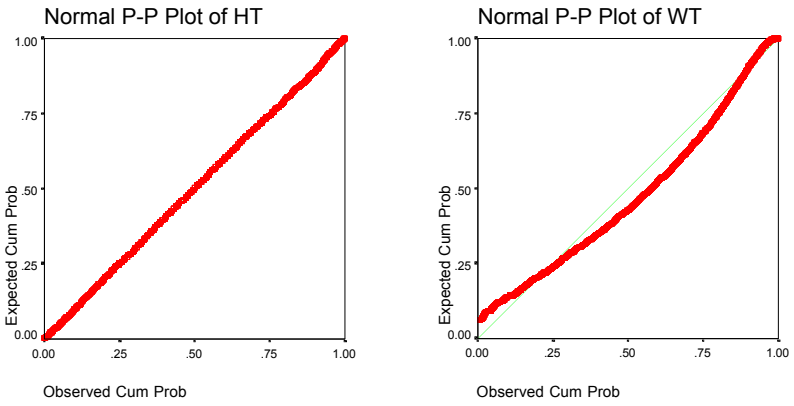


Figure 2-3.5: SPSS Expected Cumulative Probability vs Observed Cumulative Probability Plots for Height (HT) and Weight (WT) in women of data depicted in Figure 2-3.3

The normal probability plot is a useful tool in determining how normal your distribution is. In the normal probability plot, the cumulative probability for the data (observed) is plotted against the cumulative probability of the data if it were normally distributed (expected), as shown for weight (WT) and height (HT) in Figure 2-3.5. The approximately normally distributed variable, height, can be seen to have a linear relationship between observed and expected values. If the two sets of values agreed perfectly (a correlation of 1) then height would be perfectly normal. The correlation between observed and expected is therefore a measure of normality of the observed scores.

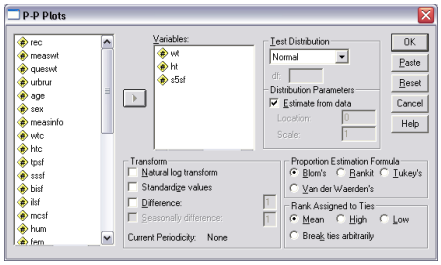


Figure 2-3.6: SPSS P-P Plots option of the GRAPH menu to produce normal probability plots

The skewed variable, weight, can be seen to have divergent observed scores of cumulative probability as shown by the bend in the normal probability plot for weight. The normal probability plots can be called up in SPSS by using the P-P option of the GRAPH menu. Figure 2-3.6 shows the dialog box for this option. The variables to be tested for normality are moved over to the Variables box. Ensure that Normal is selected in the Test Distribution box. SPSS can produce plots to test more than the normal distribution.

Box-Cox Transformation

The Box-Cox transformation is a family of transformations, being defined as:

$$T(X) = (X^{\lambda} - 1)/\lambda$$

where Y is the response variable and λ is the transformation parameter.

For λ= 0, the natural log of the data is taken instead of using the above formula.

Net Admin 12/12/11 9:48 PM
Comment: Richard, is T(X) the same as Y, the response variable?

As discussed earlier, the normal probability plot gives us an appreciation of the degree of normality of the distribution as the values of observed cumulative frequency distribution are plotted against the expected normal cumulative frequency distribution of a variable with the same mean and standard deviation. The correlation between the expected and observed values is a measure of agreement of the observed data to the normal distribution. This correlation coefficient can be used as the criterion for judgement of the value of λ that best normalizes the distribution. Figure 2-3.7 shows a typical curve of the correlation coefficients found for different values of λ. In this case -0.6 was the value of λ that gives the highest correlation (0.91) between the observed and expected values of the cumulative frequency. -0.6 would therefore be chosen as the value of λ to best transform the data to a normal distribution. Unfortunately SPSS does not carry out the Box-Cox analysis, but we can find the best value of λ using MS EXCEL, as described below.

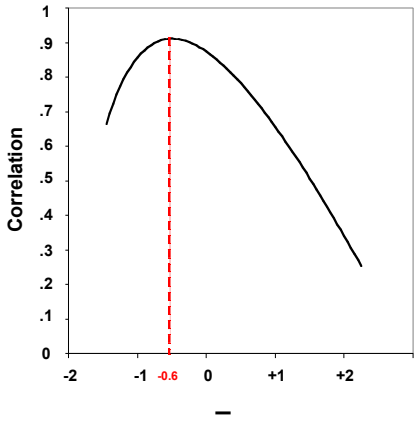


Figure 2-3.7: Plot of correlations of expected and observed values of cumulative probability curve for different values of λ. Maximum correlation found for λ=-0.6.

Calculating the Box-Cox λ using MS EXCEL

Rather than using the correlation between expected and observed cumulative frequency values as the criterion of normality, we will use the coefficient of skewness, which will approach 0 the closer the distribution is to normal. Figure 2-3.8 shows an EXCEL set up for the calculation of the best value of λ using the SOLVER function. The data being analysed are the Sum of 5 Skinfolds on 273 women, aged 18 to 19 years from the Canada Fitness Survey data set. The coefficient of skewness for this variable is 1.19; therefore, the data are significantly skewed and a Box-Cox transformation would be in order.

The first steps in calculating the best fitting value of λ are as follows:

- Calculate the column of transformed scores for Sum5SF based upon the value of λ entered in cell E1. The value in cell E1 can be any number. Choose a small number similar to the likely answer for λ. In this case 1 was used. It matters little exactly what this number is since it is only a starting point for SOLVER. The equation entered in B2 is **=((A2^E1)-1)/E1**, which is the Box-Cox transform equation shown earlier in the chapter but written in EXCEL computational form including specific cell references.

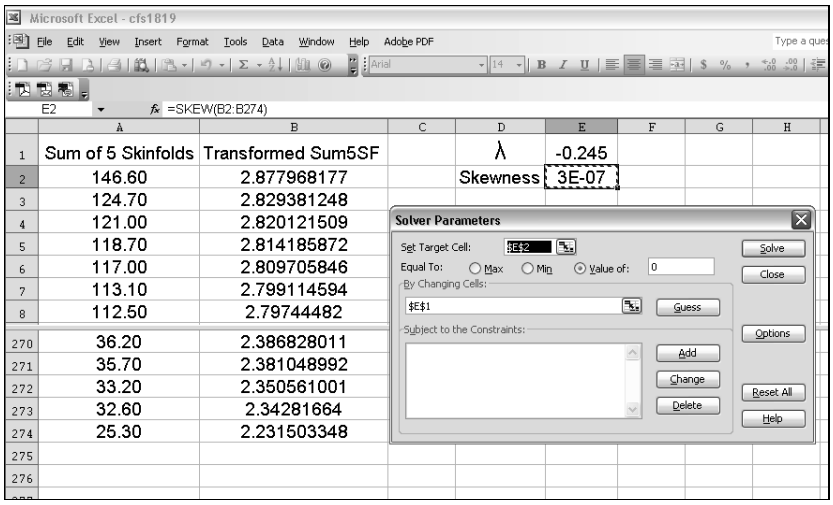


Figure 2-3.7: MS EXCEL SOLVER set up for Box-Cox transformation calculation.

- Calculate the coefficient of skewness for the transformed scores. In Figure 2-3.8 this was placed in cell E2. This is achieved using the SKEW() function of EXCEL which returns the coefficient of skewness of the data in the selected range of cells. In Figure 2-3.8 the equation typed in E2 was **=SKEW(B2:B274)**.
- Choose the SOLVER function from the TOOLS menu. Figure 2-3.8 shows the SOLVER dialog box. SOLVER requires you to give the address of the target cell. In this case we give the cell address of the coefficient of skewness E2. Now you need to check whether you want SOLVER to seek a maximum, minimum or value closest to 0. In this case we want the coefficient of skewness to get closest to 0. SOLVER needs to change one or more cells that change the target cell E2. Thus E1 (the cell containing the value of λ) is entered in the 'by changing cells' box. SOLVER is now set up. If you click solve now, SOLVER will go through a high speed process of changing the value of λ in cell E1, checking on the value of E2, changing E1 again until E2 reaches the closest possible

value to 0. In this case the value of -0.245 brought the coefficient of skewness closest to 0. Therefore $\lambda = -0.245$ would be used to transform the sum of 5 skinfold data to best approximate a normally distributed variable.