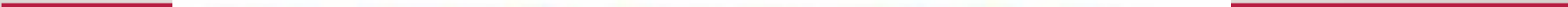




# MULTIVARIATE ANALYSIS



**TESTING STATISTICAL  
ASSUMPTIONS**

## 2 STATISTICAL ASSUMPTIONS

---

All statistical procedures have underlying assumptions, some more stringent than others. In some cases, violation of these assumptions will not change substantive research conclusions. In other cases, violation of assumptions will undermine meaningful research. Establishing that one's data meet the assumptions of the procedure one is using, is an expected component of all quantitatively-based journal articles, theses, and dissertations.

### 3 STATISTICAL ASSUMPTIONS

---

- Data Related Assumptions
- Model Related Assumptions

## 4 DATA RELATED STATISTICAL ASSUMPTIONS

---

- SOUND MEASUREMENT
- All forms of statistical analysis assume sound measurement, relatively free of coding errors. It is good practice to run descriptive statistics on one's data so that one is confident that data are generally as expected in terms of means and standard deviations, and there are no out-of-bounds entries beyond the expected range.
- Avoiding Attenuation
- When the range of the data is reduced artificially, as by classifying or dichotomizing a continuous variable, correlation is attenuated, often leading to underestimation of the effect size of that variables
- Sample Size Adequacy
- Randomness (RUN TEST) (SPSS)
- Descriptive Statistics: Central Tendency, Dispersion, Skewness and Kurtosis (SPSS)

## 5 MODEL RELATED STATISTICAL ASSUMPTIONS FOR MULTIVARIATE ANALYSIS

---

- Normality of Variables
- Zero Expected Residuals
- Normality of Residuals
- No Multicollinearity
- Homoscedasticity
- No Autocorrelation
- Linearity
- Model Specification
- Model Identification

# TESTS OF NORMALITY

---

- There are several methods of assessing whether the data are normally distributed or not. They fall under two categories:
- GRAPHICAL METHODS
  - Histogram (SPSS)
  - BOX PLOT (SPSS)
  - Q-Q Probability Plots (SPSS)
  - P-P (Cumulative Probability (or Frequency) Plots (SPSS)
- STATISTICAL TESTS
  - W/S Test (SPSS)
  - Jarque Bera Test (EViews)
  - Shapiro-Wilk's Test (SPSS)
  - Kolmogorov-Smirnov Test (SPSS)
  - D' Agostino's Test (NCSS)
  - Chi-Square Test (SPSS)
  - Anderson Darling Test (MINITAB) (EXCEL AD Calculator)

# 7 TESTING FOR NORMALITY

## Graphical methods of assessing normality

A *histogram* of a variable shows rough normality, and a histogram of residuals, if normally distributed, is often taken as evidence of normality of all the variables.

A *graph of empirical by theoretical cumulative distribution functions (cdf's)* simply shows the empirical distribution as, say, a dotted line, and the hypothetical distribution, say the normal curve, as a solid line.

A *P-P plot* is found in SPSS under Graphs, P-P plots. One may test if the distribution of a given variable is normal (or beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, Pareto, Student's t, Weibull, or uniform). The P-P plot plots a variable's cumulative *proportions* against the cumulative proportions of the test distribution. The straighter the line formed by the P-P plot, the more the variable's distribution conforms to the selected test distribution (ex., normal). Options within this SPSS procedure allow data transforms first (natural log, standardization of values, difference, and seasonally difference).

A *quantile-by-quantile or Q-Q plot* forms a 45-degree line when the observed values are in conformity with the hypothetical distribution. Q-Q plots plot the *quantiles* of a variable's distribution against the quantiles of the test distribution. From the SPSS menu, select Graphs, Q-Q. The SPSS dialog box supports testing the following distributions: beta, chi-square, exponential, gamma, half-normal, Laplace, Logistic, Lognormal, normal, pareto, Student's t, Weibull, and uniform. Q-Q plots are also produced in SPSS under Analyze, Descriptive Statistics, Explore, Plots when one checks "Normality plots with tests."

## 8 SPSS COMMANDS FOR GRAPHS

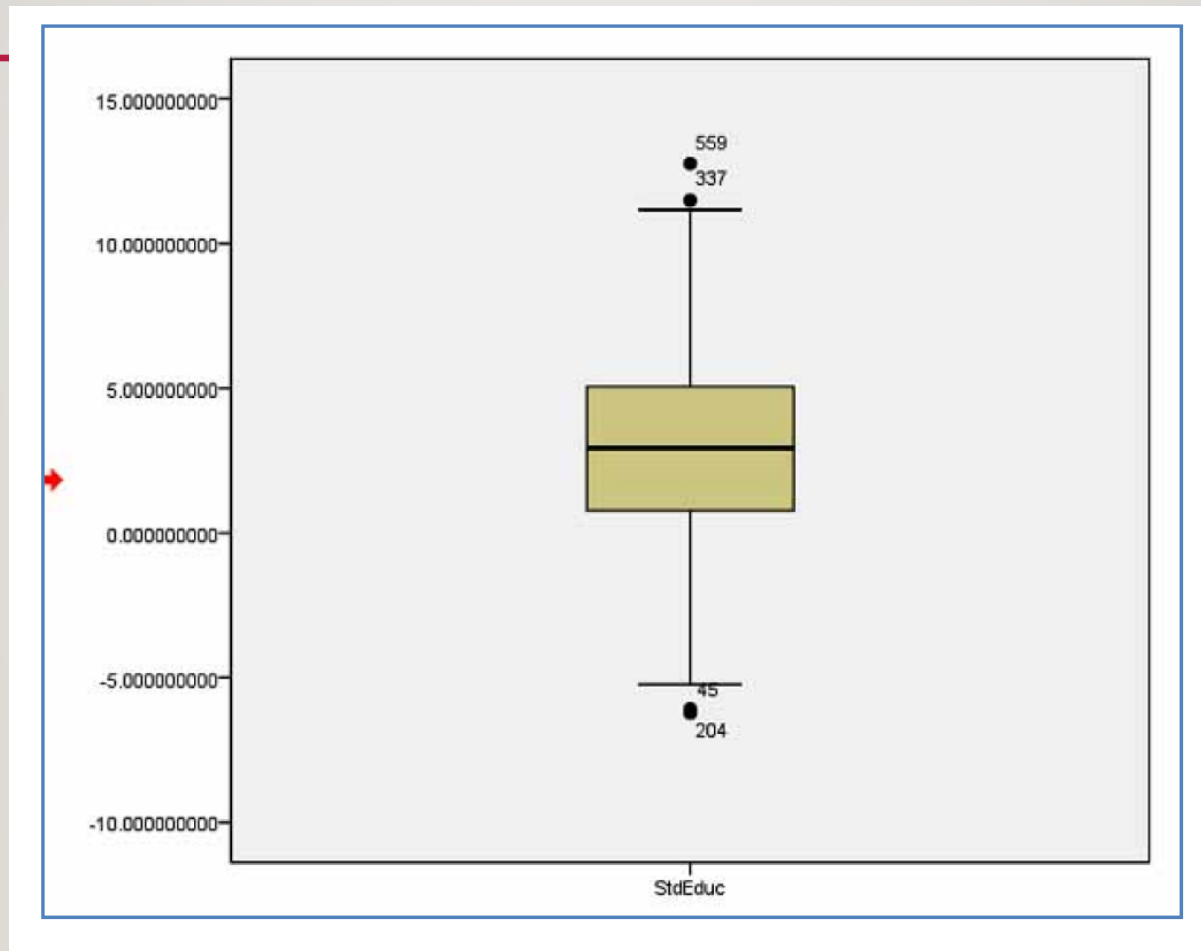
---

- Histogram: GRAPH>LEGACY DIALOGUE>HISTOGRAM>Display Normal Curve
- Box Plot, Q-Q Plot and Normality Tests (K-S test and Shapiro-Wilks Test):
- ANALYZE>DESCRIPTIVE STATISTICS>EXPLORE>PLOTS>HISTOGRAM and NORMALITY PLOTS WITH TESTS
- For P-P Plots: ANALYZE>DESCRIPTIVE STATISTICS>P-P Plots or Q-Q Plots

## 9 TESTING FOR NORMALITY

**Boxplot tests of the normality assumption.** Outliers and skewness indicate non-normality, and both can be checked with boxplots. The SPSS boxplot output option (also under Analyze > Descriptive Statistics > Explore, "Plots" button, check "Normality plots with tests"). For a single variable being tested for normality, a box plot is a chart with that variable on the X axis and with the Y axis representing its spread of values (in the illustration below, the values of standardized education, StdEduc). Inside the graph, for the given variable, the height of the rectangle indicates the spread of the values for the variable. The horizontal dark line within the rectangle indicates the mean. In the illustration it is 0 since zero is the mean for standardized variables, which StdEduc is. If most of the rectangle is on one side or the other of the mean line, this indicates the dependent is skewed (not normal). Further out than the rectangle are the "whiskers," which mark the smallest and largest observations which are not outliers (defined as observations greater than 1.5 inter-quartile ranges [IQR's = boxlengths] from the 1st and 3rd quartiles). Outliers are shown as numbered cases beyond the whiskers. ( Note you can display boxplots for two factors (two independents) together by selecting Clustered Boxplots from the Boxplot item on the SPSS Graphs menu.)

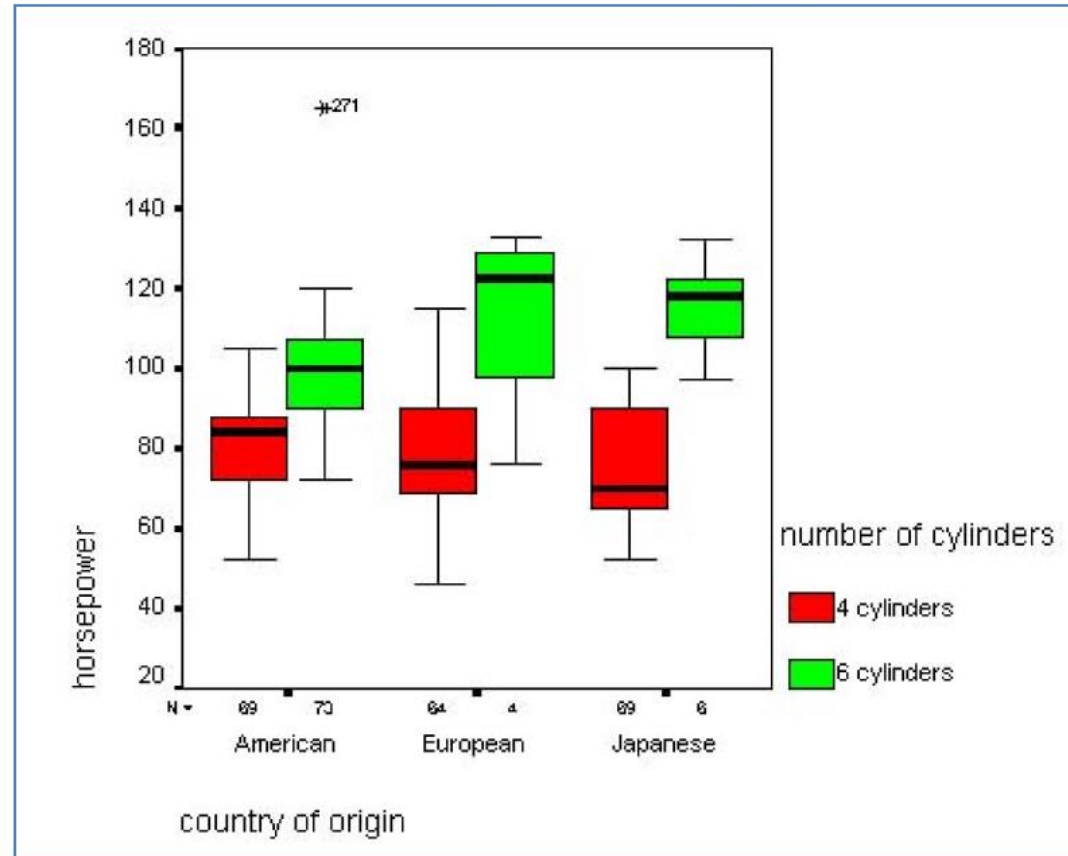
## 10 TESTING FOR NORMALITY



## II TESTING FOR NORMALITY

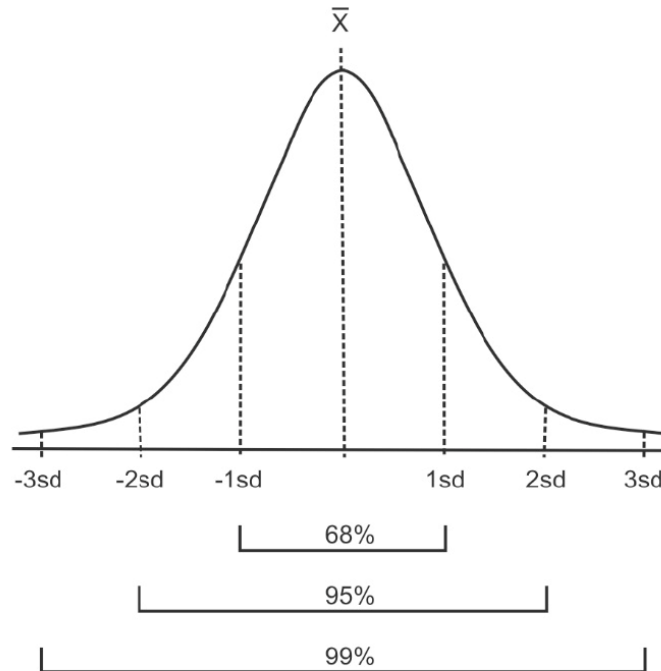
- To take a more complex example, a box plot can also be a chart in which categories of a categorical independent or of multiple independents are arrayed on the X axis and values of an interval dependent are arrayed on the Y axis. In the example below there are two categorical independents (country of car manufacture, number of cylinders) predicting the continuous dependent variable horsepower.
- Inside the graph, for each X category, will be a rectangle indicating the spread of the dependent's values for that category. If these rectangles are roughly at the same Y elevation for all categories, this indicates little difference among groups. Within each rectangle is a horizontal dark line, indicating the mean. If most of the rectangle is on one side or the other of the mean line, this indicates the dependent is skewed (not normal) for that group (category). Whiskers and outliers are as described above for the one-variable case.

## 12 TESTING FOR NORMALITY



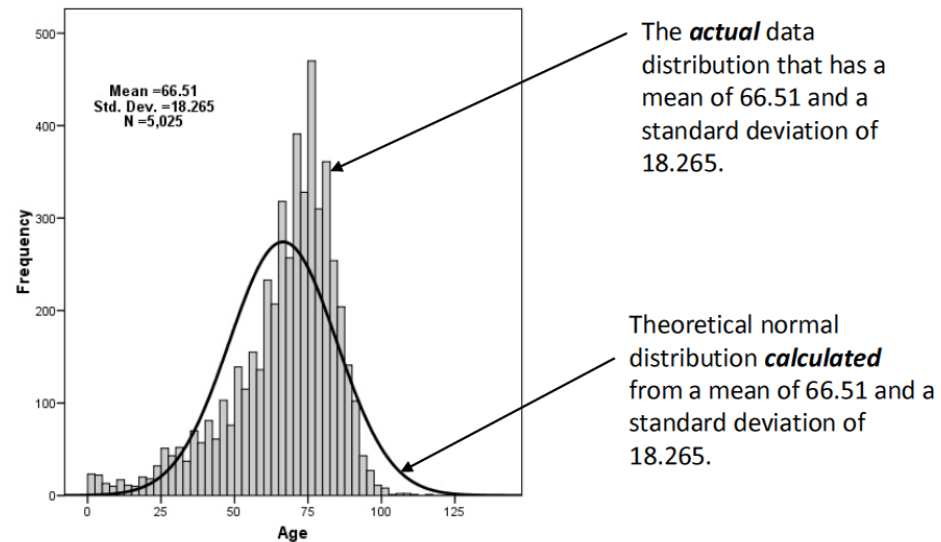
# 13 TESTING FOR NORMALITY

For each mean and standard deviation combination a theoretical normal distribution can be determined. This distribution is based on the proportions shown below.



## 14 TESTING FOR NORMALITY

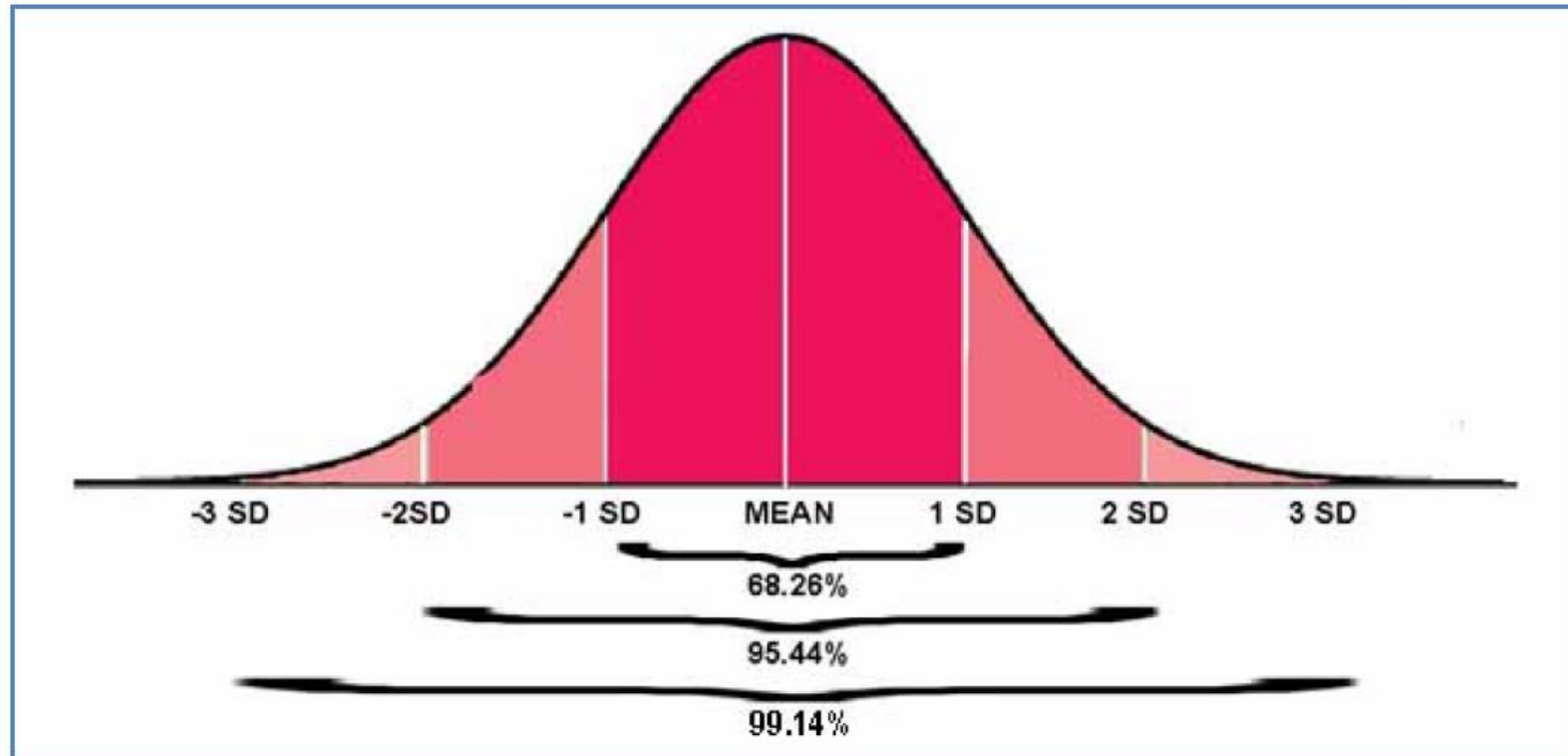
This theoretical normal distribution can then be compared to the actual distribution of the data.



Are the actual data statistically different than the computed normal curve?

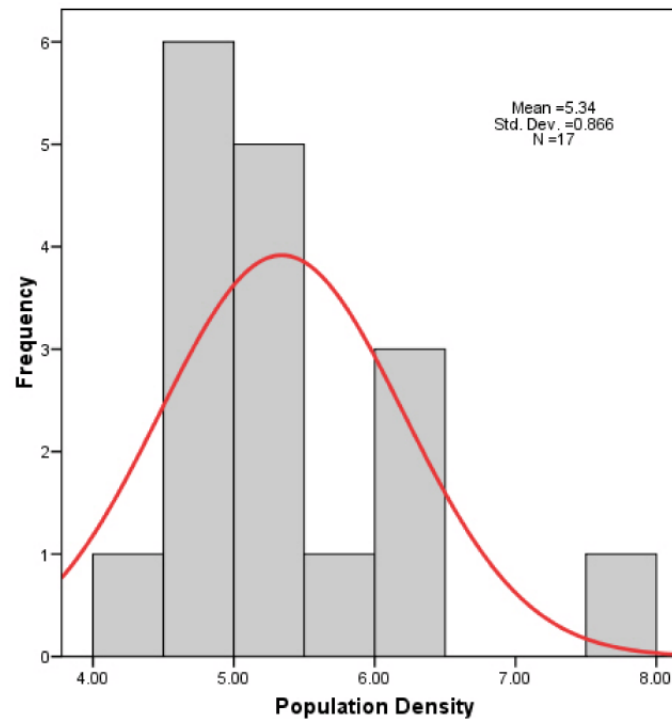
## 15 DATA RELATED STATISTICAL ASSUMPTIONS

that it will lie outside 3 standard deviations. This statement is analogous to statements pertaining to significance levels of .05 and .01, for two-tailed tests. .



## 16 TESTING FOR NORMALITY

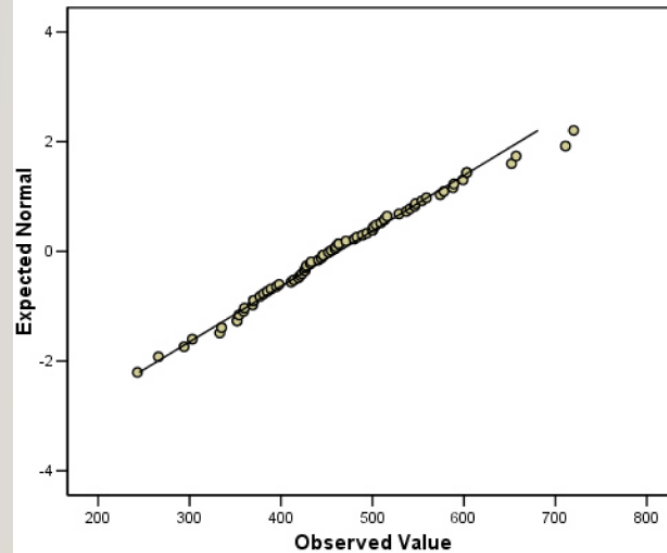
Graphical methods are typically not very useful when the sample size is small. This is a histogram of the last example. These data do not 'look' normal, but they are not statistically different than normal.



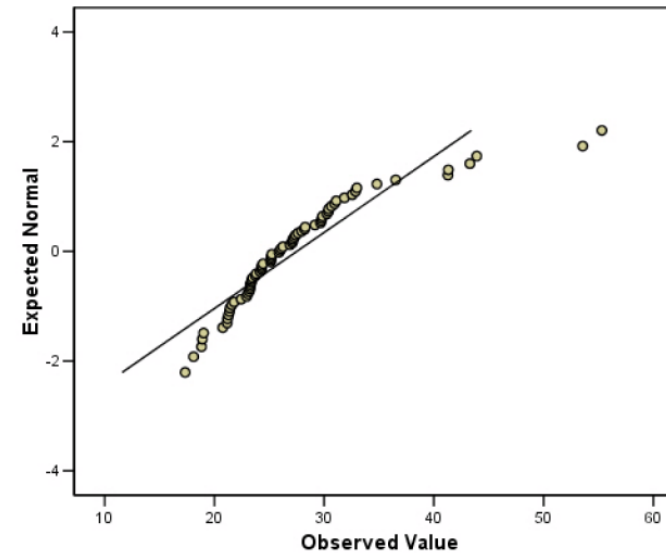
# 17 TESTING FOR NORMALITY

Q-Q plots display the observed values against normally distributed data (represented by the line).

Q-Q Plot: Normally Distributed Data

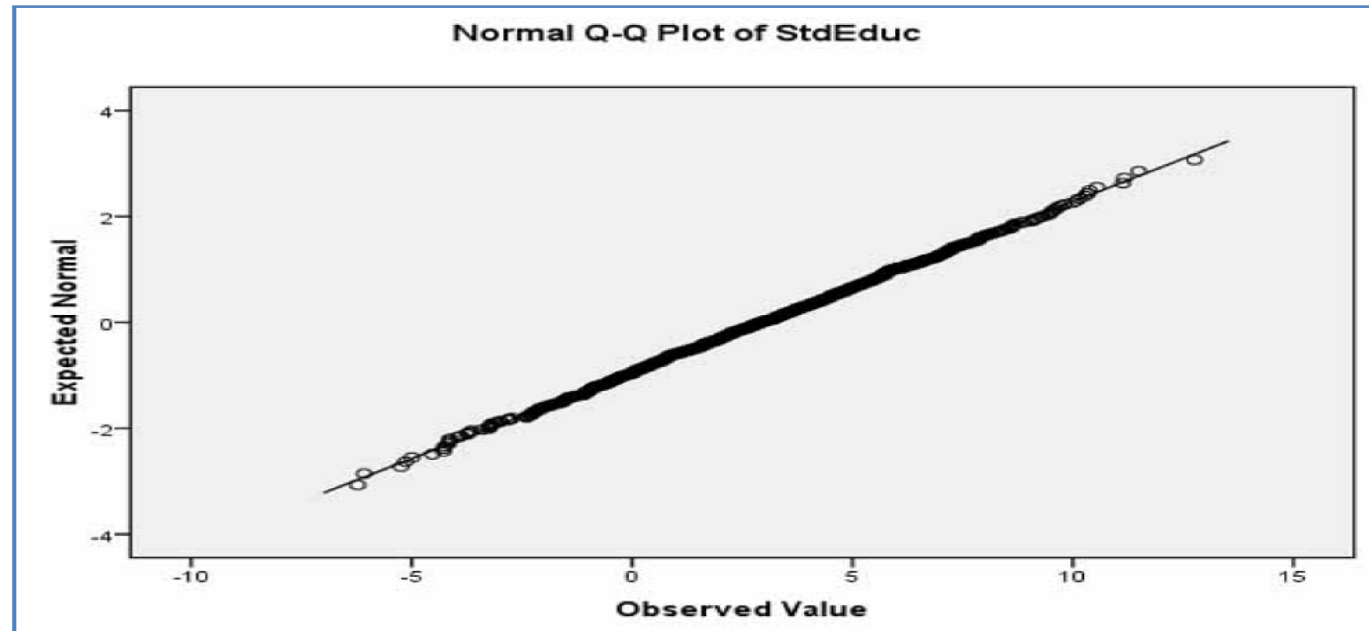


Q-Q Plot: Non-normally Distributed Data



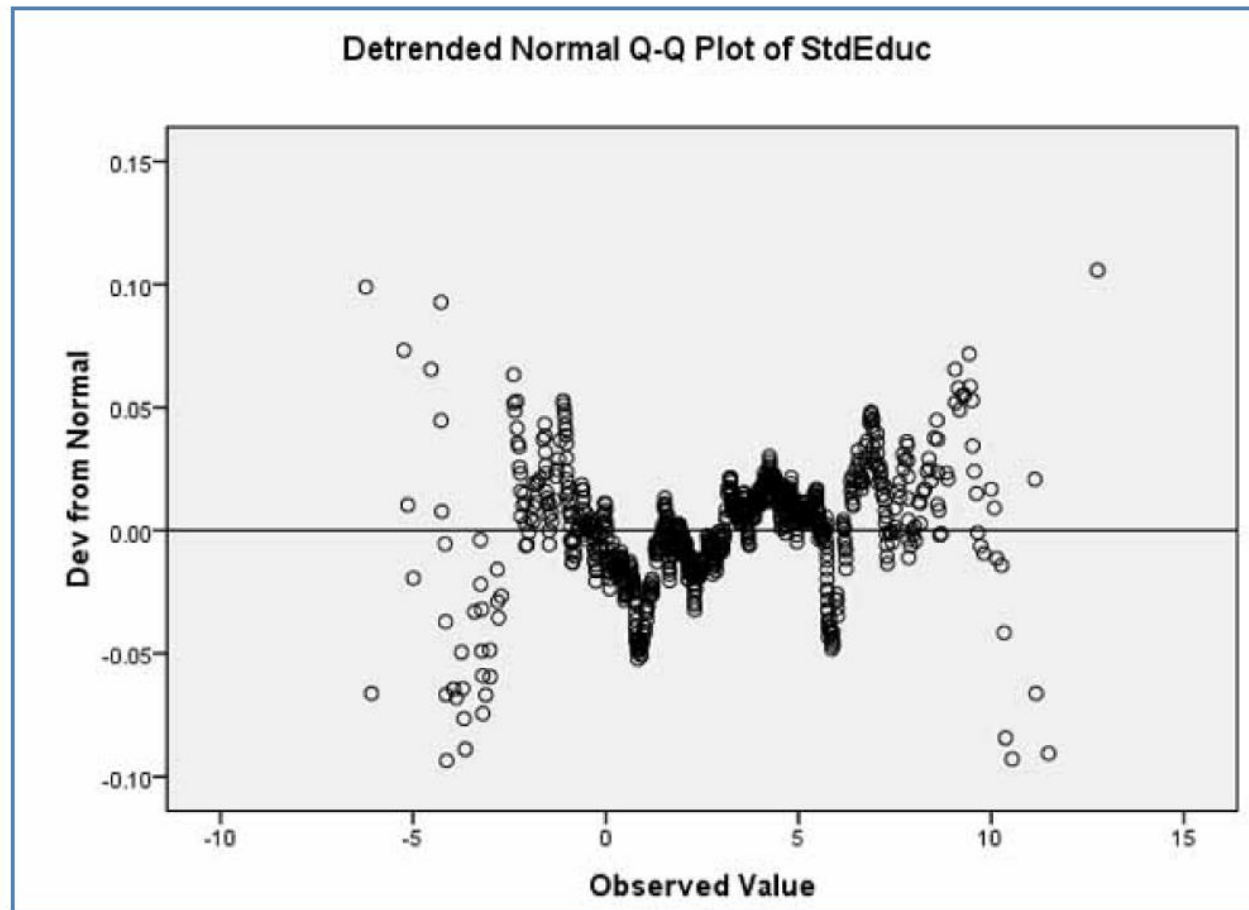
Normally distributed data fall along the line.

## 18 TESTING FOR NORMALITY



A *detrended Q-Q plot*, obtained in the same way in SPSS, provides similar information. If a variable is normally distributed, cases in the detrended Q-Q plot should cluster around the horizontal 0 line representing 0 standard deviations from the 45-degree line seen in the non-detrended Q-Q plot above. The detrended Q-Q plot is useful for spotting outliers. For the illustration below, however, there are no outliers in the sense that there are no cases more than  $\pm .12$  standard deviations away. Cases more than  $\pm 1.96$  standard deviations away are outliers at the .95 confidence level.

## 19 TESTING FOR NORMALITY



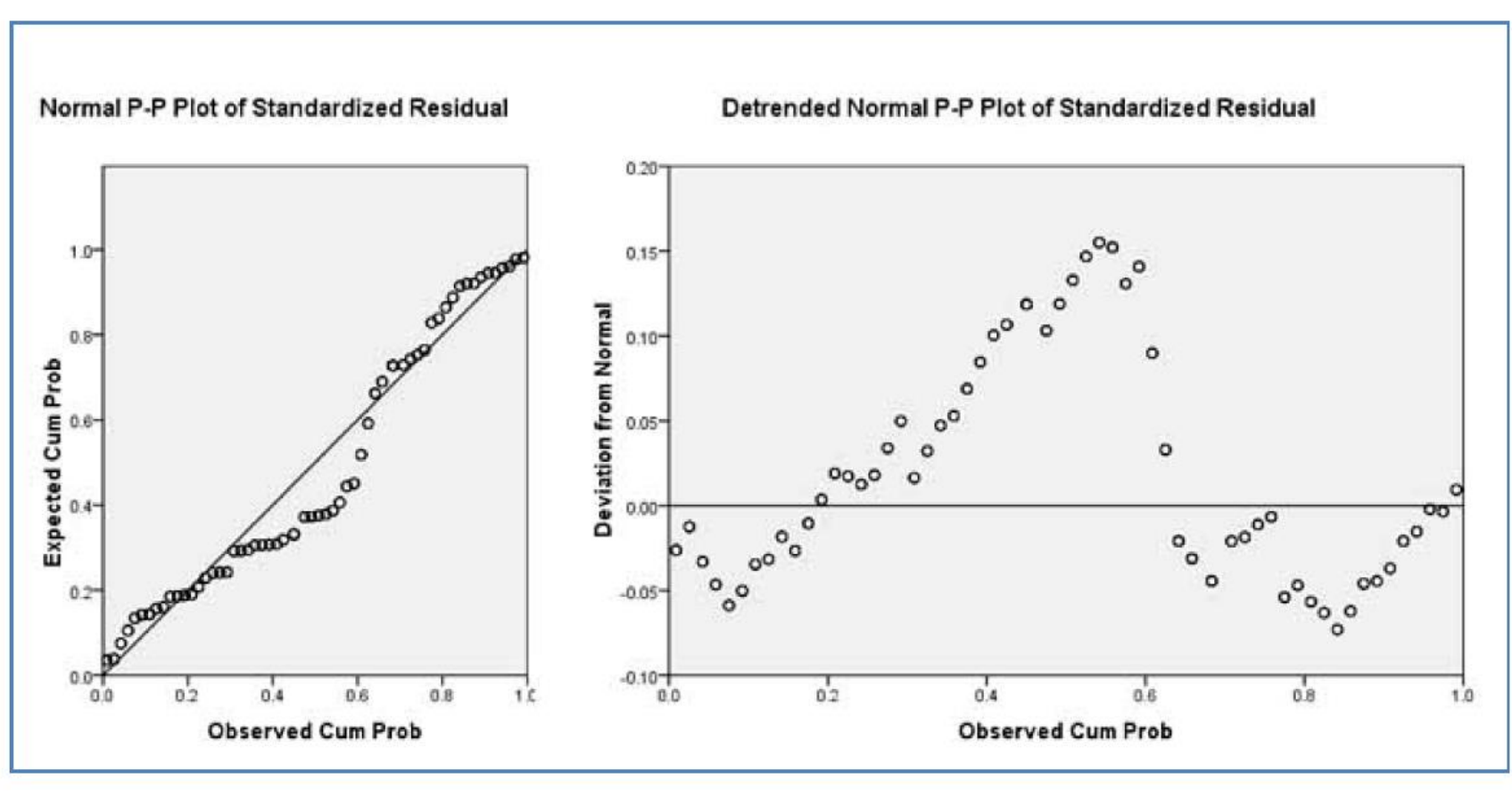
## 20 NORMAL PROBABILITY PLOT (P-P PLOT)

---

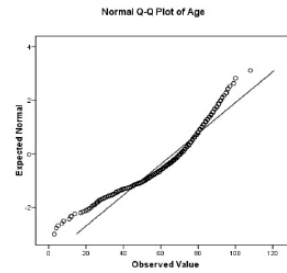
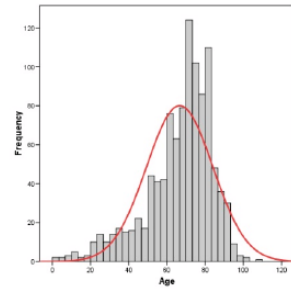
### **A normal probability plot**

The normal probability plot, also called a P-P Plot, is an alternative method, plotting observed cumulative probabilities of occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree line will appear when the observed conforms to the normally expected and the assumption of normally distributed error is met. In the detrended version, the pattern of points above and below the horizontal 0 line should be random (unlike the illustration below). *P-P plots* are found in SPSS under Graphs, P-P plots. Or select Analyze, Descriptive Statistics, P-P plots. Or select Analyze, Regression, Linear; click Plots; check Normal probability plot.

## 21 NORMAL PROBABILITY PLOT (P-P PLOT)



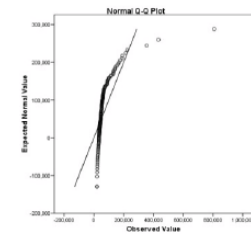
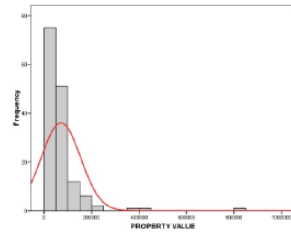
## 22 TESTING FOR NORMALITY



Tests of Normality

|     | Kolmogorov-Smirnov <sup>a</sup> |      |      | Shapiro-Wilk |      |      |
|-----|---------------------------------|------|------|--------------|------|------|
|     | Statistic                       | df   | Sig. | Statistic    | df   | Sig. |
| Age | .110                            | 1048 | .000 | .931         | 1048 | .000 |

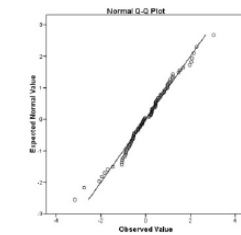
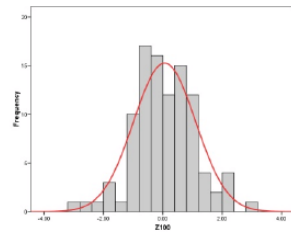
a. Lilliefors Significance Correction



Tests of Normality

|            | Kolmogorov-Smirnov <sup>a</sup> |     |      | Shapiro-Wilk |     |      |
|------------|---------------------------------|-----|------|--------------|-----|------|
|            | Statistic                       | df  | Sig. | Statistic    | df  | Sig. |
| TOTAL_VALU | .283                            | 149 | .000 | .463         | 149 | .000 |

a. Lilliefors Significance Correction



Tests of Normality

|      | Kolmogorov-Smirnov <sup>a</sup> |     |       | Shapiro-Wilk |     |      |
|------|---------------------------------|-----|-------|--------------|-----|------|
|      | Statistic                       | df  | Sig.  | Statistic    | df  | Sig. |
| Z100 | .071                            | 100 | .200* | .985         | 100 | .333 |

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

## 23 TESTING FOR NORMALITY

Statistical tests for normality are more precise since actual probabilities are calculated.

Tests for normality calculate the probability that the sample was drawn from a normal population.

The hypotheses used are:

$H_0$ : The sample data are not significantly different than a normal population.

$H_a$ : The sample data are significantly different than a normal population.

## 24 TESTING FOR NORMALITY

Typically, we are interested in finding a difference between groups. When we are, we 'look' for small probabilities.

- If the probability of finding an event is rare (less than 5%) and we actually find it, that is of interest.

When testing normality, we are not 'looking' for a difference.

- In effect, we want our data set to be NO DIFFERENT than normal. We want to accept the null hypothesis.

So when testing for normality:

- Probabilities  $> 0.05$  mean the data are normal.
- Probabilities  $< 0.05$  mean the data are NOT normal.

## 25 TESTING FOR NORMALITY

---

### **Shapiro-Wilk's W test**

This test is a formal test of normality offered in the SPSS EXAMINE module or the SAS UNIVARIATE procedure. This is the standard test for normality. In SPSS, select Analyze, Descriptive statistics, Explore. Note the Explore menu choice pastes the EXAMINE code. Click the Plots button and check "Normality plots with tests." Output like that below is generated:

## 26 TESTING FOR NORMALITY

Tests of Normality

|         | Kolmogorov-Smirnov <sup>a</sup> |     |       | Shapiro-Wilk |     |      |
|---------|---------------------------------|-----|-------|--------------|-----|------|
|         | Statistic                       | df  | Sig.  | Statistic    | df  | Sig. |
| StdEduc | .017                            | 932 | .200* | .999         | 932 | .960 |

a. Lilliefors Significance Correction

\*. This is a lower bound of the true significance.

For a given variable, W should not be significant if the variable's distribution is not significantly different from normal, as is the case for StdEduc in the illustration above. W may be thought of as the correlation between given data and their corresponding normal scores, with  $W = 1$  when the given data are perfectly normal in distribution. When W is significantly smaller than 1, the assumption of normality is not met. Shapiro-Wilk's W is recommended for small and medium samples up to  $n = 2000$ . For larger samples, the Kolmogorov-Smirnov test is recommended by SAS and others.

## 27 TESTING FOR NORMALITY

Remember that LARGE probabilities denote normally distributed data. Below are examples taken from SPSS.

Normally Distributed Data

|              | Kolmogorov-Smirnov <sup>a</sup> |    |       | Shapiro-Wilk |    |      |
|--------------|---------------------------------|----|-------|--------------|----|------|
|              | Statistic                       | df | Sig.  | Statistic    | df | Sig. |
| Asthma Cases | .069                            | 72 | .200* | .988         | 72 | .721 |

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Non-Normally Distributed Data

|              | Kolmogorov-Smirnov <sup>a</sup> |    |      | Shapiro-Wilk |    |      |
|--------------|---------------------------------|----|------|--------------|----|------|
|              | Statistic                       | df | Sig. | Statistic    | df | Sig. |
| Average PM10 | .142                            | 72 | .001 | .841         | 72 | .000 |

a. Lilliefors Significance Correction

## 28 TESTING FOR NORMALITY

Normally Distributed Data

|              | Kolmogorov-Smirnov <sup>a</sup> |    |       | Shapiro-Wilk |    |      |
|--------------|---------------------------------|----|-------|--------------|----|------|
|              | Statistic                       | df | Sig.  | Statistic    | df | Sig. |
| Asthma Cases | .069                            | 72 | .200* | .988         | 72 | .721 |

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

In SPSS output above the probabilities are greater than 0.05 (the typical alpha level), so we accept  $H_0$ ... these data are not different from normal.

## 29 TESTING FOR NORMALITY

| Non-Normally Distributed Data |                                 |    |      |              |    |      |
|-------------------------------|---------------------------------|----|------|--------------|----|------|
|                               | Kolmogorov-Smirnov <sup>a</sup> |    |      | Shapiro-Wilk |    |      |
|                               | Statistic                       | df | Sig. | Statistic    | df | Sig. |
| Average PM10                  | .142                            | 72 | .001 | .841         | 72 | .000 |

a. Lilliefors Significance Correction

In the SPSS output above the probabilities are less than 0.05 (the typical alpha level), so we reject  $H_0$ ... these data are significantly different from normal.

**Important:** As the sample size *increases*, normality parameters becomes **MORE** restrictive and it becomes harder to declare that the data are normally distributed. So for very large data sets, normality testing becomes less important.

# Three Simple Tests for Normality

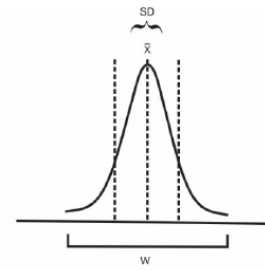
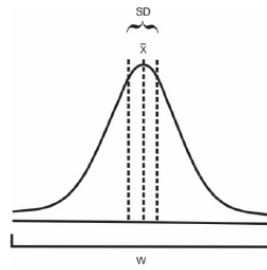
## W/S Test for Normality

- A fairly simple test that requires only the sample standard deviation and the data range.
- Based on the  $q$  statistic, which is the 'studentized' (meaning  $t$  distribution) range, or the range expressed in standard deviation units. Tests kurtosis.
- Should not be confused with the Shapiro-Wilks test.

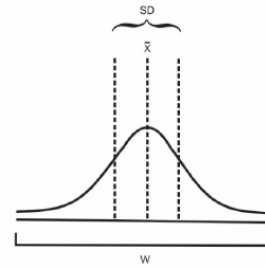
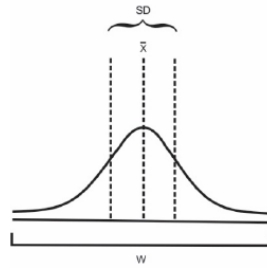
$$q = \frac{w}{s}$$

where  $q$  is the test statistic,  $w$  is the range of the data and  $s$  is the standard deviation.

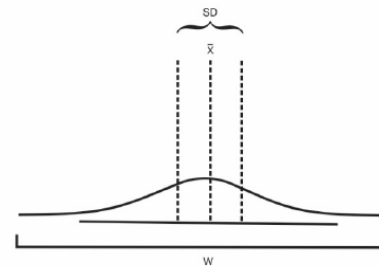
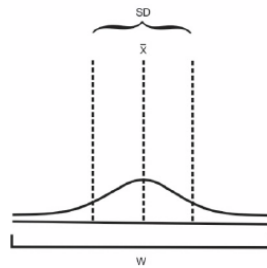
# 3 | TESTING FOR NORMALITY



Range constant,  
SD changes



Range changes,  
SD constant



## 32 TESTING FOR NORMALITY

| Village         | Population Density |
|-----------------|--------------------|
| Aranza          | 4.13               |
| Corupo          | 4.53               |
| San Lorenzo     | 4.69               |
| Cheranatzicurin | 4.76               |
| Nahuatzen       | 4.77               |
| Pomacuaran      | 4.96               |
| Sevina          | 4.97               |
| Arantepacua     | 5.00               |
| Cocucho         | 5.04               |
| Charapan        | 5.10               |
| Comachuen       | 5.25               |
| Pichataro       | 5.36               |
| Quinceo         | 5.94               |
| Nurio           | 6.06               |
| Turicuaro       | 6.19               |
| Urapicho        | 6.30               |
| Capacuaro       | 7.73               |

Standard deviation (s) = 0.866

Range (w) = 3.6

n = 17

$$q = \frac{w}{s}$$

$$q = \frac{3.6}{0.866} = 4.16$$

$$q_{\text{Critical Range}} = 3.06 \text{ to } 4.31$$

The W/S test uses a critical range. IF the calculated value falls WITHIN the range, then accept  $H_0$ . IF the calculated value falls outside the range then reject  $H_0$ .

Since  $3.06 < q=4.16 < 4.31$ , then we accept  $H_0$ .

# 33 TESTING FOR NORMALITY

Critical Values of  $q$  for the W/S Normality Test

Taken from Kanji, 1994 Table 14

Columns  $a$  denote the lower boundaries or the left-sided critical values.  
Columns  $b$  denote the upper boundaries or the right-sided critical values.

| $n$  | Level of significance $\alpha$ |       |       |       |       |       |       |       |       |       |       |       |
|------|--------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|      | 0.000                          |       | 0.005 |       | 0.01  |       | 0.025 |       | 0.05  |       | 0.10  |       |
|      | $a$                            | $b$   | $a$   | $b$   | $a$   | $b$   | $a$   | $b$   | $a$   | $b$   | $a$   | $b$   |
| 3    | 1.732                          | 2.000 | 1.735 | 2.000 | 1.737 | 2.000 | 1.745 | 2.000 | 1.758 | 1.999 | 1.782 | 1.997 |
| 4    | 1.732                          | 2.449 | 1.82  | 2.447 | 1.87  | 2.445 | 1.93  | 2.439 | 1.98  | 2.429 | 2.04  | 2.409 |
| 5    | 1.826                          | 2.828 | 1.98  | 2.813 | 2.02  | 2.803 | 2.09  | 2.782 | 2.15  | 2.753 | 2.22  | 2.712 |
| 6    | 1.826                          | 3.162 | 2.11  | 3.115 | 2.15  | 3.095 | 2.22  | 3.056 | 2.28  | 3.012 | 2.37  | 2.949 |
| 7    | 1.871                          | 3.464 | 2.22  | 3.369 | 2.26  | 3.338 | 2.33  | 3.282 | 2.40  | 3.222 | 2.49  | 3.143 |
| 8    | 1.871                          | 3.742 | 2.31  | 3.585 | 2.35  | 3.543 | 2.43  | 3.471 | 2.50  | 3.399 | 2.59  | 3.308 |
| 9    | 1.897                          | 4.000 | 2.39  | 3.772 | 2.44  | 3.720 | 2.51  | 3.634 | 2.59  | 3.552 | 2.68  | 3.449 |
| 10   | 1.897                          | 4.243 | 2.46  | 3.935 | 2.51  | 3.875 | 2.59  | 3.777 | 2.67  | 3.685 | 2.76  | 3.57  |
| 11   | 1.915                          | 4.472 | 2.53  | 4.079 | 2.58  | 4.012 | 2.66  | 3.903 | 2.74  | 3.80  | 2.84  | 3.68  |
| 12   | 1.915                          | 4.690 | 2.59  | 4.208 | 2.64  | 4.134 | 2.72  | 4.02  | 2.80  | 3.91  | 2.90  | 3.78  |
| 13   | 1.927                          | 4.899 | 2.64  | 4.325 | 2.70  | 4.244 | 2.78  | 4.12  | 2.86  | 4.00  | 2.96  | 3.87  |
| 14   | 1.927                          | 5.099 | 2.70  | 4.431 | 2.75  | 4.34  | 2.83  | 4.21  | 2.92  | 4.09  | 3.02  | 3.95  |
| 15   | 1.936                          | 5.292 | 2.74  | 4.53  | 2.80  | 4.44  | 2.88  | 4.29  | 2.97  | 4.17  | 3.07  | 4.02  |
| 16   | 1.936                          | 5.477 | 2.79  | 4.62  | 2.84  | 4.52  | 2.93  | 4.37  | 3.01  | 4.24  | 3.12  | 4.09  |
| 17   | 1.944                          | 5.657 | 2.83  | 4.70  | 2.88  | 4.60  | 2.97  | 4.44  | 3.06  | 4.31  | 3.17  | 4.15  |
| 18   | 1.944                          | 5.831 | 2.87  | 4.78  | 2.92  | 4.67  | 3.01  | 4.51  | 3.10  | 4.37  | 3.21  | 4.21  |
| 19   | 1.949                          | 6.000 | 2.90  | 4.85  | 2.96  | 4.74  | 3.05  | 4.56  | 3.14  | 4.43  | 3.25  | 4.27  |
| 20   | 1.949                          | 6.164 | 2.94  | 4.91  | 2.99  | 4.80  | 3.09  | 4.63  | 3.18  | 4.49  | 3.29  | 4.32  |
| 25   | 1.961                          | 6.93  | 3.09  | 5.19  | 3.15  | 5.06  | 3.24  | 4.87  | 3.34  | 4.71  | 3.45  | 4.53  |
| 30   | 1.966                          | 7.62  | 3.21  | 5.40  | 3.27  | 5.26  | 3.37  | 5.06  | 3.47  | 4.89  | 3.59  | 4.70  |
| 35   | 1.972                          | 8.25  | 3.32  | 5.57  | 3.38  | 5.42  | 3.48  | 5.21  | 3.58  | 5.04  | 3.70  | 4.84  |
| 40   | 1.975                          | 8.83  | 3.41  | 5.71  | 3.47  | 5.56  | 3.57  | 5.34  | 3.67  | 5.16  | 3.79  | 4.96  |
| 45   | 1.978                          | 9.38  | 3.49  | 5.83  | 3.55  | 5.67  | 3.66  | 5.45  | 3.75  | 5.26  | 3.88  | 5.06  |
| 50   | 1.980                          | 9.90  | 3.56  | 5.93  | 3.62  | 5.77  | 3.73  | 5.54  | 3.83  | 5.35  | 3.95  | 5.14  |
| 55   | 1.982                          | 10.39 | 3.62  | 6.02  | 3.69  | 5.86  | 3.80  | 5.63  | 3.90  | 5.43  | 4.02  | 5.22  |
| 60   | 1.983                          | 10.86 | 3.68  | 6.10  | 3.75  | 5.94  | 3.86  | 5.70  | 3.96  | 5.51  | 4.08  | 5.29  |
| 65   | 1.985                          | 11.31 | 3.74  | 6.17  | 3.80  | 6.01  | 3.91  | 5.77  | 4.01  | 5.57  | 4.14  | 5.35  |
| 70   | 1.986                          | 11.75 | 3.79  | 6.24  | 3.85  | 6.07  | 3.96  | 5.83  | 4.06  | 5.63  | 4.19  | 5.41  |
| 75   | 1.987                          | 12.17 | 3.83  | 6.30  | 3.90  | 6.13  | 4.01  | 5.88  | 4.11  | 5.68  | 4.24  | 5.46  |
| 80   | 1.987                          | 12.57 | 3.88  | 6.35  | 3.94  | 6.18  | 4.05  | 5.93  | 4.16  | 5.73  | 4.28  | 5.51  |
| 85   | 1.988                          | 12.96 | 3.92  | 6.40  | 3.99  | 6.23  | 4.09  | 5.98  | 4.20  | 5.78  | 4.33  | 5.56  |
| 90   | 1.989                          | 13.34 | 3.96  | 6.45  | 4.02  | 6.27  | 4.13  | 6.03  | 4.24  | 5.82  | 4.36  | 5.60  |
| 95   | 1.990                          | 13.71 | 3.99  | 6.49  | 4.06  | 6.32  | 4.17  | 6.07  | 4.27  | 5.86  | 4.40  | 5.64  |
| 100  | 1.990                          | 14.07 | 4.03  | 6.53  | 4.10  | 6.36  | 4.21  | 6.11  | 4.31  | 5.90  | 4.44  | 5.68  |
| 150  | 1.993                          | 17.26 | 4.32  | 6.82  | 4.38  | 6.64  | 4.48  | 6.39  | 4.59  | 6.18  | 4.72  | 5.96  |
| 200  | 1.995                          | 19.95 | 4.53  | 7.01  | 4.59  | 6.84  | 4.68  | 6.60  | 4.78  | 6.39  | 4.90  | 6.15  |
| 500  | 1.998                          | 31.59 | 5.06  | 7.60  | 5.13  | 7.42  | 5.25  | 7.15  | 5.47  | 6.94  | 5.49  | 6.72  |
| 1000 | 1.999                          | 44.70 | 5.50  | 7.99  | 5.57  | 7.80  | 5.68  | 7.54  | 5.79  | 7.33  | 5.92  | 7.11  |

Source: Sachs, 1972

## 34 TESTING FOR NORMALITY

---

Since we have a critical range, it is difficult to determine a probability range for our results. Therefore we simply state our alpha level.

The sample data set is not significantly different than normal ( $W/S_{4.16}$ ,  $p > 0.05$ ).

## 35 TESTING FOR NORMALITY

### Jarque–Bera Test

A goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

$$k_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3} \quad k_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{ns^4} - 3$$

$$JB = n \left( \frac{(k_3)^2}{6} + \frac{(k_4)^2}{24} \right)$$

Where  $\mathbf{x}$  is each observation,  $\mathbf{n}$  is the sample size,  $\mathbf{s}$  is the standard deviation,  $\mathbf{k}_3$  is skewness, and  $\mathbf{k}_4$  is kurtosis.

## 36 TESTING FOR NORMALITY

$$\bar{x} = 5.34$$

$$s = 0.87$$

| Village         | Population<br>Density | Mean<br>Deviates | Mean<br>Deviates <sup>3</sup> | Mean<br>Deviates <sup>4</sup> |
|-----------------|-----------------------|------------------|-------------------------------|-------------------------------|
| Aranza          | 4.13                  | -1.21            | -1.771561                     | 2.14358881                    |
| Corupo          | 4.53                  | -0.81            | -0.531441                     | 0.43046721                    |
| San Lorenzo     | 4.69                  | -0.65            | -0.274625                     | 0.17850625                    |
| Cheranatzicurin | 4.76                  | -0.58            | -0.195112                     | 0.11316496                    |
| Nahuatzen       | 4.77                  | -0.57            | -0.185193                     | 0.10556001                    |
| Pomacuaran      | 4.96                  | -0.38            | -0.054872                     | 0.02085136                    |
| Sevina          | 4.97                  | -0.37            | -0.050653                     | 0.01874161                    |
| Arantepacua     | 5.00                  | -0.34            | -0.039304                     | 0.01336336                    |
| Cocucho         | 5.04                  | -0.30            | -0.027000                     | 0.00810000                    |
| Charapan        | 5.10                  | -0.24            | -0.013824                     | 0.00331776                    |
| Comachuen       | 5.25                  | -0.09            | -0.000729                     | 0.00006561                    |
| Pichataro       | 5.36                  | 0.02             | 0.000008                      | 0.00000016                    |
| Quinceo         | 5.94                  | 0.60             | 0.216000                      | 0.12960000                    |
| Nurio           | 6.06                  | 0.72             | 0.373248                      | 0.26873856                    |
| Turicuaro       | 6.19                  | 0.85             | 0.614125                      | 0.52200625                    |
| Urapicho        | 6.30                  | 0.96             | 0.884736                      | 0.84934656                    |
| Capacuaro       | 7.73                  | 2.39             | 13.651919                     | 32.62808641                   |
|                 |                       |                  | 12.595722                     | 37.433505                     |

## 37 TESTING FOR NORMALITY

$$k_3 = \frac{12.6}{(17)(0.87^3)} = 1.13$$

$$k_4 = \frac{37.43}{(17)(0.87^4)} - 3 = 0.843$$

$$JB = 17 \left( \frac{(1.13)^2}{6} + \frac{(0.843)^2}{24} \right) = 17 \left( \frac{1.2769}{6} + \frac{0.711}{24} \right)$$

$$JB = 17(0.2128 + 0.0296)$$

$$JB = 4.12$$

## 38 TESTING FOR NORMALITY

The Jarque-Bera statistic can be compared to the  $\chi^2$  distribution (table) with 2 degrees of freedom (df or  $\nu$ ) to determine the critical value at an alpha level of 0.05.

Critical Values of the Chi-Squared ( $\chi^2$ ) Distribution  
*Taken from Rholf and Sokal Table 14*

| $\nu \backslash \alpha$ | .995  | .975  | .9    | .5    | .1     | .05    | .025   | .01    | .005   | .001   | $\alpha / \nu$ |
|-------------------------|-------|-------|-------|-------|--------|--------|--------|--------|--------|--------|----------------|
| 1                       | 0.000 | 0.000 | 0.016 | 0.455 | 2.706  | 3.841  | 5.024  | 6.635  | 7.879  | 10.828 | 1              |
| 2                       | 0.010 | 0.051 | 0.211 | 1.386 | 4.605  | 5.991  | 7.378  | 9.210  | 10.597 | 13.816 | 2              |
| 3                       | 0.072 | 0.216 | 0.584 | 2.366 | 6.251  | 7.815  | 9.348  | 11.345 | 12.838 | 16.266 | 3              |
| 4                       | 0.207 | 0.484 | 1.064 | 3.357 | 7.779  | 9.488  | 11.143 | 13.277 | 14.860 | 18.467 | 4              |
| 5                       | 0.412 | 0.831 | 1.610 | 4.351 | 9.236  | 11.070 | 12.832 | 15.086 | 16.750 | 20.515 | 5              |
| 6                       | 0.676 | 1.237 | 2.204 | 5.348 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 | 22.458 | 6              |
| 7                       | 0.989 | 1.690 | 2.833 | 6.346 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 | 24.322 | 7              |
| 8                       | 1.344 | 2.180 | 3.490 | 7.344 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 | 26.124 | 8              |
| 9                       | 1.735 | 2.700 | 4.168 | 8.343 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 | 27.877 | 9              |
| 10                      | 2.156 | 3.247 | 4.865 | 9.342 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 | 29.588 | 10             |

The critical  $\chi^2$  value is 5.991. Our calculated Jarque-Bera statistic is 4.12 which falls between 0.5 and 0.1, which is greater than the critical value.

Therefore we accept  $H_0$  that there is no difference between our distribution and a normal distribution (Jarque-Bera  $\chi^2_{4.15}$ ,  $0.5 > p > 0.1$ ).

## 39 TESTING FOR NORMALITY

### D'Agostino Test

- A very powerful test for departures from normality.
- Based on the D statistic, which gives an upper and lower critical value.

$$D = \frac{T}{\sqrt{n^3 SS}} \quad \text{where} \quad T = \sum \left( i - \frac{n+1}{2} \right) X_i$$

where  $D$  is the test statistic,  $SS$  is the sum of squares of the data and  $n$  is the sample size, and  $i$  is the order or rank of observation  $X$ . The df for this test is  $n$  (sample size).

- First the data are ordered from smallest to largest or largest to smallest.

# 40 TESTING FOR NORMALITY

| Village                  | Population Density | i  | Mean Deviates <sup>2</sup> |
|--------------------------|--------------------|----|----------------------------|
| Aranza                   | 4.13               | 1  | 1.46410                    |
| Corupo                   | 4.53               | 2  | 0.65610                    |
| San Lorenzo              | 4.69               | 3  | 0.42250                    |
| Cheranatzicurin          | 4.76               | 4  | 0.33640                    |
| Nahuatzen                | 4.77               | 5  | 0.32490                    |
| Pomacuaran               | 4.96               | 6  | 0.14440                    |
| Sevina                   | 4.97               | 7  | 0.13690                    |
| Arantepacua              | 5.00               | 8  | 0.11560                    |
| Cocucho                  | 5.04               | 9  | 0.09000                    |
| Charapan                 | 5.10               | 10 | 0.05760                    |
| Comachuen                | 5.25               | 11 | 0.00810                    |
| Pichataro                | 5.36               | 12 | 0.00040                    |
| Quinceo                  | 5.94               | 13 | 0.36000                    |
| Nurio                    | 6.06               | 14 | 0.51840                    |
| Turicuaro                | 6.19               | 15 | 0.72250                    |
| Urapicho                 | 6.30               | 16 | 0.92160                    |
| Capacuaro                | 7.73               | 17 | 5.71210                    |
| Mean = 5.34 SS = 11.9916 |                    |    |                            |

$(4.13 - 5.34)^2 = -1.21^2 = 1.46410$   
 $\frac{n+1}{2} = \frac{17+1}{2} = 9$   
 $T = \sum (i-9)X_i$   
 $T = (1-9)4.13 + (2-9)4.53 + (3-9)4.69 + \dots + (17-9)7.73$   
 $T = 63.23$   
 $D = \frac{63.23}{\sqrt{(17^3)(11.9916)}} = 0.26050$   
 $D_{Critical} = 0.2587, 0.2860$

Df = n = 17

If the calculated value falls within the critical range, accept  $H_0$ .

Since  $0.2587 < D = 0.26050 < 0.2860$  accept  $H_0$ .

The sample data set is not significantly different than normal ( $D_{0.26050}$ ,  $p > 0.05$ ).

# 41 TESTING FOR NORMALITY

Use the next lower  $n$  on the table if your sample size is NOT listed.

Critical Values for the D'Agostino D Normality Test  
Taken from Zar, 1981 Table B.22

| $n$  | $\alpha = 0.20$ | 0.10           | 0.05           | 0.02           | 0.01           |
|------|-----------------|----------------|----------------|----------------|----------------|
| 10   | 0.2632, 0.2835  | 0.2573, 0.2843 | 0.2513, 0.2849 | 0.2436, 0.2855 | 0.2379, 0.2857 |
| 12   | 0.2653, 0.2841  | 0.2598, 0.2849 | 0.2544, 0.2854 | 0.2473, 0.2859 | 0.2420, 0.2862 |
| 14   | 0.2669, 0.2846  | 0.2618, 0.2853 | 0.2568, 0.2858 | 0.2503, 0.2862 | 0.2455, 0.2865 |
| 16   | 0.2681, 0.2848  | 0.2634, 0.2855 | 0.2587, 0.2860 | 0.2527, 0.2865 | 0.2482, 0.2867 |
| 18   | 0.2690, 0.2850  | 0.2646, 0.2855 | 0.2603, 0.2862 | 0.2547, 0.2866 | 0.2505, 0.2868 |
| 20   | 0.2699, 0.2852  | 0.2657, 0.2857 | 0.2617, 0.2863 | 0.2564, 0.2867 | 0.2525, 0.2869 |
| 22   | 0.2705, 0.2853  | 0.2670, 0.2859 | 0.2629, 0.2864 | 0.2579, 0.2869 | 0.2542, 0.2870 |
| 24   | 0.2711, 0.2853  | 0.2675, 0.2860 | 0.2638, 0.2865 | 0.2591, 0.2870 | 0.2557, 0.2871 |
| 26   | 0.2717, 0.2854  | 0.2682, 0.2861 | 0.2647, 0.2866 | 0.2603, 0.2870 | 0.2570, 0.2872 |
| 28   | 0.2721, 0.2854  | 0.2688, 0.2861 | 0.2655, 0.2866 | 0.2612, 0.2870 | 0.2581, 0.2873 |
| 30   | 0.2725, 0.2854  | 0.2693, 0.2861 | 0.2662, 0.2866 | 0.2622, 0.2871 | 0.2592, 0.2872 |
| 32   | 0.2729, 0.2854  | 0.2698, 0.2862 | 0.2668, 0.2867 | 0.2630, 0.2871 | 0.2600, 0.2873 |
| 34   | 0.2732, 0.2854  | 0.2703, 0.2862 | 0.2674, 0.2867 | 0.2636, 0.2871 | 0.2609, 0.2873 |
| 36   | 0.2735, 0.2854  | 0.2707, 0.2862 | 0.2679, 0.2867 | 0.2643, 0.2871 | 0.2617, 0.2873 |
| 38   | 0.2738, 0.2854  | 0.2710, 0.2862 | 0.2683, 0.2867 | 0.2649, 0.2871 | 0.2623, 0.2873 |
| 40   | 0.2740, 0.2854  | 0.2714, 0.2862 | 0.2688, 0.2867 | 0.2655, 0.2871 | 0.2630, 0.2874 |
| 42   | 0.2743, 0.2854  | 0.2717, 0.2861 | 0.2691, 0.2867 | 0.2659, 0.2871 | 0.2636, 0.2874 |
| 44   | 0.2745, 0.2854  | 0.2720, 0.2861 | 0.2695, 0.2867 | 0.2664, 0.2871 | 0.2641, 0.2874 |
| 46   | 0.2747, 0.2854  | 0.2722, 0.2861 | 0.2698, 0.2866 | 0.2668, 0.2871 | 0.2646, 0.2874 |
| 48   | 0.2749, 0.2854  | 0.2725, 0.2861 | 0.2702, 0.2866 | 0.2672, 0.2871 | 0.2651, 0.2874 |
| 50   | 0.2751, 0.2853  | 0.2727, 0.2861 | 0.2705, 0.2866 | 0.2676, 0.2871 | 0.2655, 0.2874 |
| 60   | 0.2757, 0.2852  | 0.2737, 0.2860 | 0.2717, 0.2865 | 0.2692, 0.2870 | 0.2673, 0.2873 |
| 70   | 0.2763, 0.2851  | 0.2744, 0.2859 | 0.2726, 0.2864 | 0.2708, 0.2869 | 0.2687, 0.2872 |
| 80   | 0.2768, 0.2850  | 0.2750, 0.2857 | 0.2734, 0.2863 | 0.2713, 0.2868 | 0.2698, 0.2871 |
| 90   | 0.2771, 0.2849  | 0.2755, 0.2856 | 0.2740, 0.2862 | 0.2721, 0.2866 | 0.2707, 0.2870 |
| 100  | 0.2774, 0.2849  | 0.2759, 0.2855 | 0.2745, 0.2860 | 0.2727, 0.2865 | 0.2714, 0.2869 |
| 120  | 0.2779, 0.2847  | 0.2765, 0.2853 | 0.2752, 0.2858 | 0.2737, 0.2863 | 0.2725, 0.2866 |
| 140  | 0.2782, 0.2846  | 0.2770, 0.2852 | 0.2758, 0.2856 | 0.2744, 0.2862 | 0.2734, 0.2865 |
| 160  | 0.2785, 0.2845  | 0.2774, 0.2851 | 0.2763, 0.2855 | 0.2750, 0.2860 | 0.2741, 0.2863 |
| 180  | 0.2787, 0.2844  | 0.2777, 0.2850 | 0.2767, 0.2854 | 0.2755, 0.2859 | 0.2746, 0.2862 |
| 200  | 0.2789, 0.2843  | 0.2779, 0.2848 | 0.2770, 0.2853 | 0.2759, 0.2857 | 0.2751, 0.2860 |
| 250  | 0.2793, 0.2841  | 0.2784, 0.2846 | 0.2776, 0.2850 | 0.2767, 0.2855 | 0.2760, 0.2858 |
| 300  | 0.2796, 0.2840  | 0.2788, 0.2844 | 0.2781, 0.2848 | 0.2772, 0.2853 | 0.2766, 0.2855 |
| 350  | 0.2798, 0.2839  | 0.2791, 0.2843 | 0.2784, 0.2847 | 0.2776, 0.2851 | 0.2771, 0.2853 |
| 400  | 0.2799, 0.2838  | 0.2793, 0.2842 | 0.2787, 0.2845 | 0.2780, 0.2849 | 0.2775, 0.2852 |
| 450  | 0.2801, 0.2837  | 0.2795, 0.2841 | 0.2789, 0.2844 | 0.2782, 0.2848 | 0.2778, 0.2851 |
| 500  | 0.2802, 0.2836  | 0.2796, 0.2840 | 0.2791, 0.2843 | 0.2785, 0.2847 | 0.2780, 0.2849 |
| 600  | 0.2804, 0.2835  | 0.2799, 0.2839 | 0.2794, 0.2842 | 0.2788, 0.2845 | 0.2784, 0.2847 |
| 700  | 0.2805, 0.2834  | 0.2800, 0.2838 | 0.2796, 0.2840 | 0.2791, 0.2844 | 0.2787, 0.2846 |
| 800  | 0.2806, 0.2833  | 0.2802, 0.2837 | 0.2798, 0.2839 | 0.2793, 0.2842 | 0.2790, 0.2844 |
| 900  | 0.2807, 0.2833  | 0.2803, 0.2836 | 0.2799, 0.2838 | 0.2795, 0.2841 | 0.2792, 0.2843 |
| 1000 | 0.2808, 0.2832  | 0.2804, 0.2835 | 0.2800, 0.2838 | 0.2796, 0.2840 | 0.2793, 0.2842 |
| 1250 | 0.2809, 0.2831  | 0.2806, 0.2834 | 0.2803, 0.2836 | 0.2799, 0.2839 | 0.2797, 0.2840 |
| 1500 | 0.2810, 0.2830  | 0.2807, 0.2833 | 0.2805, 0.2835 | 0.2801, 0.2837 | 0.2799, 0.2839 |
| 1750 | 0.2811, 0.2830  | 0.2808, 0.2832 | 0.2806, 0.2834 | 0.2803, 0.2836 | 0.2801, 0.2838 |
| 2000 | 0.2812, 0.2829  | 0.2809, 0.2831 | 0.2807, 0.2833 | 0.2804, 0.2835 | 0.2802, 0.2837 |

For each significance level,  $\alpha$  is given a pair of critical values. If the calculated  $D$  is  $\leq$  the first member of the pair, or  $\geq$  the second, then, the null hypothesis of population normality is rejected.

## 42 TESTING FOR NORMALITY

Different normality tests produce vastly different probabilities. This is due to where in the distribution (central, tails) or what moment (skewness, kurtosis) they are examining.

| Normality Test     | Statistic | Calculated Value | Probability     | Results    |
|--------------------|-----------|------------------|-----------------|------------|
| W/S                | q         | 4.16             | > 0.05          | Normal     |
| Jarque-Bera        | $\chi^2$  | 4.15             | $0.5 > p > 0.1$ | Normal     |
| D'Agostino         | D         | 0.2605           | > 0.05          | Normal     |
| Shapiro-Wilk       | W         | 0.8827           | 0.035           | Non-normal |
| Kolmogorov-Smirnov | D         | 0.2007           | 0.067           | Normal     |

## 43 TESTING FOR NORMALITY

Normality tests using various random normal sample sizes:

| <u>Sample<br/>Size</u> | <u>K-S<br/>Prob</u> |
|------------------------|---------------------|
| 20                     | 1.000               |
| 50                     | 0.893               |
| 100                    | 0.871               |
| 200                    | 0.611               |
| 500                    | 0.969               |
| 1000                   | 0.904               |
| 2000                   | 0.510               |
| 5000                   | 0.106               |
| 10000                  | 0.007               |
| 15000                  | 0.005               |
| 20000                  | 0.007               |
| 50000                  | 0.002               |

Notice that as the sample size increases, the probabilities decrease. In other words, ***it gets harder*** to meet the normality assumption as the sample size increases since even small differences are detected.

## 44 TESTING FOR NORMALITY

Which normality test should I use?

**Kolmogorov-Smirnov:**

- Not sensitive to problems in the tails.
- For data sets  $> 50$ .

**Shapiro-Wilks:**

- Doesn't work well if several values in the data set are the same.
- Works best for data sets with  $< 50$ , but can be used with larger data sets.

**W/S:**

- Simple, but effective.
- Not available in SPSS.

**Jarque-Bera:**

- Tests for skewness and kurtosis, very effective.
- Not available in SPSS.

**D'Agostino:**

- Powerful omnibus (skewness, kurtosis, centrality) test.
- Not available in SPSS.

## 45 TESTING FOR NORMALITY

---

Final Words Concerning Normality Testing:

1. Since it IS a test, state a null and alternate hypothesis.
2. If you perform a normality test, do not ignore the results.
3. If the data are not normal, use non-parametric tests.
4. If the data are normal, use parametric tests.

AND MOST IMPORTANTLY:

- 5. If you have groups of data, you MUST test each group for normality.**

## 46 BOOTSTRAPPING

### Resampling

Resampling, which is a form of bootstrapping, is a way of doing significance testing while avoiding parametric assumptions like multivariate normality. The assumption of multivariate normality is violated when dichotomous, dummy, and other discrete variables are used. In such situations, where significance testing is appropriate, researchers may use a resampling method.

### Normalizing Transformations

Various transformations are used to correct skew:

- Square roots, logarithmic, and inverse ( $1/x$ ) transforms "pull in" outliers and normalize right (positive) skew. Inverse (reciprocal) transforms are stronger than logarithmic, which are stronger than roots.
- To correct left (negative) skew, first subtract all values from the highest value plus 1, then apply square root, inverse, or logarithmic transforms.

# ANDERSON-DARLING TEST

---

- Anderson-Darling Test will determine if a data set comes from a specified distribution, in our case, the normal distribution. The test makes use of the cumulative distribution function. The Anderson-Darling statistic is given by the following formula:

$$AD = -n - \frac{1}{n} \sum_{i=1}^n (2i - 1) [\ln F(X_i) + \ln(1 - F(X_{n-i+1}))]$$

- where  $n$  = sample size,  $F(X)$  = cumulative distribution function for the specified distribution and  $i$  = the  $i^{\text{th}}$  sample when the data is sorted in ascending order. You will often see this statistic called  $A^2$ .

# ANDERSON-DARLING TEST

---

- The calculation of the p value is not straightforward. The reference most people use is R.B. D'Augostino and M.A. Stephens, Eds., 1986, Goodness-of-Fit Techniques, Marcel Dekker. There are different equations depending on the value of  $AD^*$ . These are given by:
- If  $AD^* \geq 0.6$ , then  $p = \exp(1.2937 - 5.709(AD^*) + 0.0186(AD^*)^2)$
- If  $0.34 < AD^* < .6$ , then  $p = \exp(0.9177 - 4.279(AD^*) - 1.38(AD^*)^2)$
- If  $0.2 < AD^* < 0.34$ , then  $p = 1 - \exp(-8.318 + 42.796(AD^*) - 59.938(AD^*)^2)$
- If  $AD^* \leq 0.2$ , then  $p = 1 - \exp(-13.436 + 101.14(AD^*) - 223.73(AD^*)^2)$

# ANDERSON-DARLING TEST IN MINITAB

---

To display a legend showing the Anderson-Darling test statistic and p-value each time you create a normal probability plot of the residuals:

- Choose **Tools > Options > Individual Graphs > Residual Plots for Time Series**
- and **Tools > Options > Linear Models > Residual Plots**
- Check **Include Anderson-Darling test with normal plot.**
- Click **OK.**
- Minitab does not display the test when there are less than 3 degrees of freedom for error.

## 50 NORMALIZING TRANSFORMATIONS

---

- For power and root transforms, finer adjustments can be obtained by adding a constant,  $C$ , where  $C$  is some small positive value such as .5, in the transform of  $X$ :  $X' = (X + C)^P$ . When this researcher's data contain zero values, the transform using  $C$  is strongly recommended over straight transforms (ex.,  $\text{SQRT}(X+.5)$ , not  $\text{SQRT}(X)$  ), but the use of  $C$  is standard practice in any event. Values of  $P$  less than one (roots) correct right skew, which is the common situation (using a power of  $2/3$  is common when attempting to normalize). Values of  $P$  greater than 1 (powers) correct left skew. For right skew, decreasing  $P$  decreases right skew. Too great reduction of  $P$  will overcorrect and cause left skew. When the best  $P$  is found, further refinements can be made by adjusting  $C$ . For right skew, for instance, subtracting  $C$  will decrease skew.

## 51 NORMALIZING TRANSFORMATIONS

- Logs vs. roots: logarithmic transformations are appropriate to achieve symmetry in the central distribution when symmetry of the tails is not important; square root transformations are used when symmetry in the tails is important; when both are important, a fourth root transform may work (fourth roots are used to correct extreme skew).
- Logit and probit transforms. Schumacker & Lomax (2004: 33) recommend probit transforms as a means of dealing with skewness. See Lipsey & Wilson (2001: 56) for discussion of logit and probit transforms as a means of transforming dichotomous data as part of estimating effect sizes.
- Percentages may be normalized by an arcsine transformation, which is recommended when percentages are outside the range 30% - 70%. The more observations outside this range or the closer to 0% and/or 100%, the more normality is violated and the stronger the recommendation to use arcsine transformation. However, arcsine transformation is not effective when a substantial number of observations are 0% or 100%, or when sample size is small. The usual arcsine transformation is  $p' = \arcsin(\text{SQRT}(p))$ , where  $p$  is the percentage or proportion.
- Poisson distributions may be normalized by a square root transformation.

## 52 NORMALIZING TRANSFORMATIONS

---

- Other strategies to correct for skew include collapsing categories and dropping outliers.

*Warnings.* Transformations should make theoretical sense. Often, normalizing a dichotomy such as gender will not make theoretical sense. Also note that as the log of zero is undefined and leads to error messages, researchers often add some arbitrary small value such as .001 to all values in order to remove zeros from the dataset. However, the choice of the constant can affect the significance levels of the computed coefficients for the logged variables. If this strategy is pursued, the researcher should employ sensitivity analysis with different constants to note effects which might change conclusions.

*Transforms in SPSS:* Select Transform - Compute - Target Variable (input a new variable name) - Numeric Expression (input transform formula)

## 53 NORMALIZING TRANSFORMATIONS

---

### **Box-Cox Transformations of Dependent Variables**

Box & Cox proposed a maximum likelihood method in 1964 for determining the optimal power transform for purposes of normalization of data. Power transformations of dependent variables were advanced to remedy model lack of normal distribution, lack of homogeneity of variances, and lack of additivity. In a regression context, Box-Cox transformation addresses the problem of non-normality, indicated by skewed residuals (the transformation pulls in the skew) and/or by lack of homoscedasticity of points about the regression line. In an ANOVA context, the Box-Cox transformation addresses the problem of lack of homogeneity of variances associated with the correlation of variances with means in the groups formed by the independent factors and indicated by skewed distributions within the groups (the transformation reduces the correlation).

## 54 NORMALIZING TRANSFORMATIONS

---

*Procedure.* In general, the Box-Cox procedure is to (1) Divide the independent variable into 10 or so regions; (2). Calculate the mean and s.d. for each region; (3). Plot  $\log(\text{s.d.})$  vs.  $\log(\text{mean})$  for the set of regions; (4). If the plot is a straight line, note its slope,  $b$ , then transform the variable by raising the dependent variable to the power  $(1 - b)$ , and if  $b = 1$ , then take the log of the dependent variable; and (5) if there are multiple independents, repeat steps 1 - 4 for each independent variable and pick a  $b$  which is the range of  $b$ 's.

## 55 NORMALIZING TRANSFORMATIONS

*Lambda.* In practice, computer packages apply an iterative maximum-likelihood algorithm to compute lambda, a Box-Cox parameter used to determine the exact power transformation which will best de-correlate the variances and means of the groups formed by the independent variables. As a rule of thumb, if lambda is 1.0, no transformation is needed. A lambda of +.5 corresponds to a square root transform of the dependent variable; lambda of 0 corresponds to a natural log transform; -.5 corresponds to a reciprocal square root transform; and a lambda of -1.0 corresponds to a reciprocal transform. The Box-Cox transformation is not yet supported in SPSS.

*Computer packages.* However, applying the Box-Cox transformation in SPSS and SAS is discussed in detail in Jason W. Osborne's article, "Improving your data transformations: Applying the Box-Cox transformation", *Practical Assessment, Research & Evaluation* 15(12), October, 2010.

## 56 TESTING FOR NORMALITY

### MULTIVARIATE NORMALITY

#### **Multivariate normality**

Multivariate normality is when each variable under consideration is normally distributed with respect to each other variable. Multiple analysis of variance (MANOVA), for instance, requires multivariate normality in the response (dependent) variables. Structural equation modeling and certain other procedures also assume multivariate normality.

#### **Mardia's statistic**

Mardia's statistic is a test for multivariate normality. Based on functions of skewness and kurtosis, Mardia's PK should be less than 3 to assume the assumption of multivariate normality is met. PRELIS (companion software for LISREL) outputs PK. SPSS does not yet support Mardia's PK.

## 57 TESTING FOR NORMALITY

### **Univariate screening for multivariate normality**

As a "quick and dirty" method, some researchers test for normality of each variable in the model, then assume if all are normally distributed, multivariate normality exists. This approach does not assure correct conclusions.

### **Bivariate screening for multivariate normality**

A bivariate scatterplot for any pair of variables in the model should yield an oval-shaped array of points if both variables are linearly related and normally distributed. While a step up, this is considered an exploratory approach.

### **Residuals test.**

One approach is to regress each variable in the model on all other variables in the model, then save the residuals. If all the residual variables are normally distributed (ex., by Q-Q plots, acceptable skew and kurtosis, etc.), then it is assumed that the data are multivariate normal.

## 58 NORMALLY DISTRIBUTED ERRORS

---

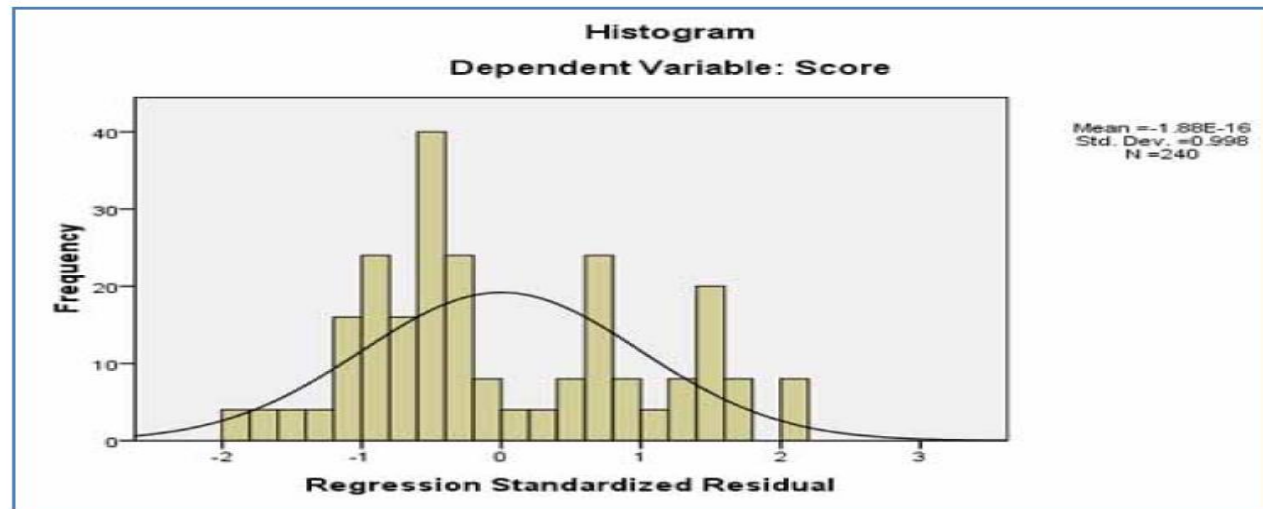
### NORMALLY DISTRIBUTED ERROR

#### Histogram of standardized residuals

The histogram of standardized residuals should show a roughly normal curve when the assumption of regression and most other techniques is met that error terms are normally distributed. In any predictive technique, the expectation is normal distribution of error, with the largest number of predictions being at or near zero and then trailing off into "high prediction" and "low prediction" tails. SPSS output shows the normal curve superimposed on such histograms. In SPSS, select Graphs, Histogram. Or select Analyze, Regression, Linear; click Plots; check Histogram.

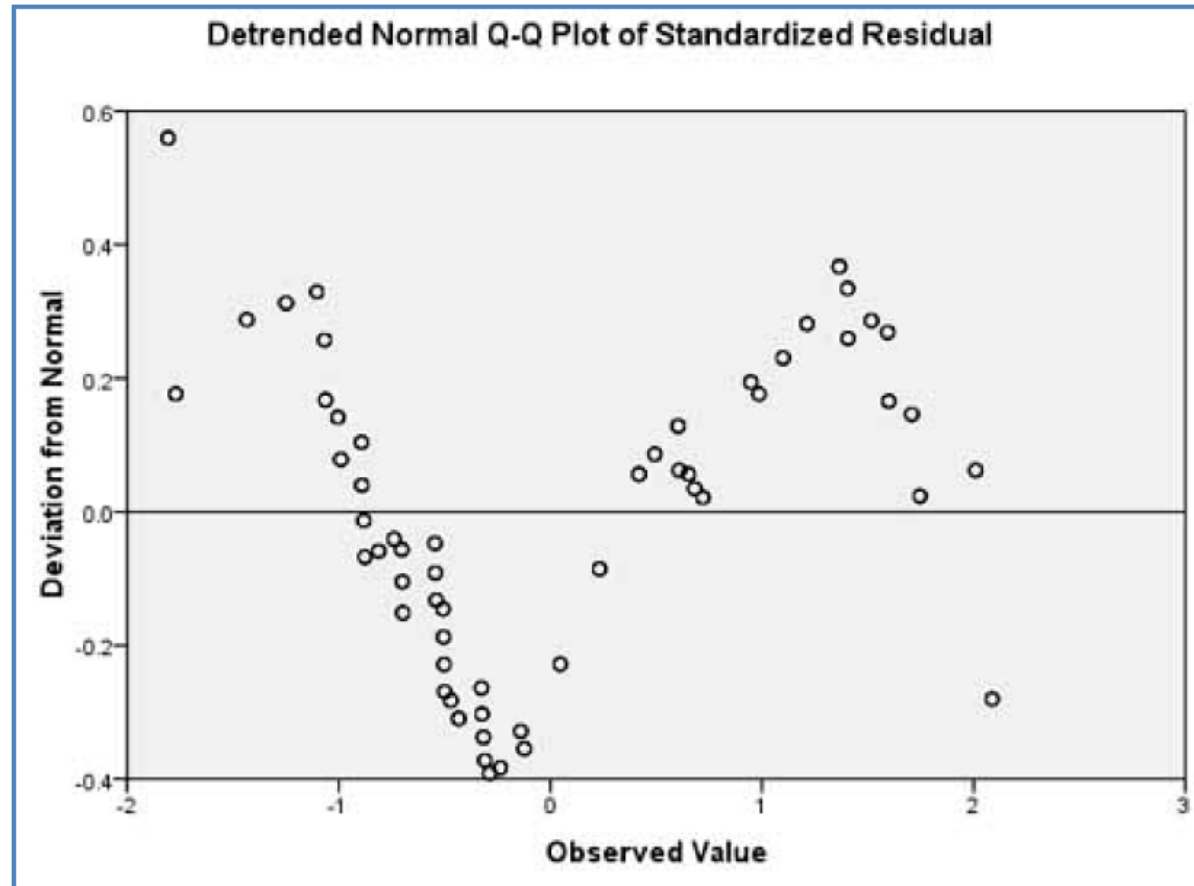
## 59 NORMALLY DISTRIBUTED ERRORS

In the example below, Score is predicted from Seniority. In this example, graphical inspection leaves doubt whether error is normally distributed.



When the residuals for the example above are saved, the SPSS Analyze, Descriptive Statistics, Explore menu choices will generate, among other statistics, the skew and kurtosis, which are .436 and -.840 respectively for these data - within normal bounds. However, as the detrended Q-Q plot for these residuals shows (below), residuals are below normal expectations for middle ranges and above for high and low ranges. This is indicative of a bimodal rather than normal distribution of error. It is also an example of where the skew and kurtosis "rules of thumb" give misleading average values.

## 60 NORMALLY DISTRIBUTED ERRORS



## 6 | TESTING FOR NORMALITY

When is non-normality a problem?

- Normality can be a problem when the sample size is small ( $< 50$ ).
- Highly skewed data create problems.
- Highly leptokurtic data are problematic, but not as much as skewed data.
- Normality becomes a serious concern when there is “activity” in the tails of the data set.
  - Outliers are a problem.
  - “Clumps” of data in the tails are worse.

# TESTING FOR HOMOGENEITY OF VARIANCES/HOMOSCEDASTICITY

---

- Brown & Forsythe's Test (SPSS)
- Welch Test (SPSS, SAS, MINITAB)
- Bartlett's Test of Sphericity (for Multivariate Normality) (SPSS)
- F-max test
- Box M test (SPSS)
- ANCOVA & MANCOVA (SPSS)
- Parallelism Tests (Test of Proportional Odds) (Running Ordinal Regression in SPSS)
- Leven's Test (SPSS)

# TESTING FOR HOMOSCEDASTICITY

---

- Graphical Method (SPSS)
- Goldfeld-Quandt Test
- Glejser Test
- Park Test
- Breusch Pagan Godfrey Test
- White Test

# HOMOGENEITY OF VARIANCES

## Overview

### Brown & Forsythe's test of homogeneity of variances

This test is based on criticisms of the Levene test. It tests for equality of group means. The Brown-Forsythe test is more robust than the Levine test when groups are unequal in size and the absolute deviation scores (deviations from the group means) are highly skewed, causing a violation of the normality assumption and the assumption of equal variances. In SPSS, select Analyze, Compare Means, One-Way ANOVA; click Options; select Brown-Forsythe.

#### Example

In the example below, Zodiac (Zodiac sign) is used to predict Polviews (liberal or conservative). As expected, the ANOVA is non-significant, indicating the Zodiac does not predict Polviews. Because the Levene statistic is not significant, the researcher fails to reject the null hypothesis that the groups have equal variances. Frequencies for Zodiac, not shown here, show group sizes not be be markedly different, so the results of the Levene test are accepted. However, were the group sizes markedly different, the Brown & Forsyth test would be used. For these data,

# HOMOGENEITY OF VARIANCES

## Overview

the Brown & Forsyth test is also non-significant and thus not different in inference from Levene's test.

### ANOVA

Think of Self as Liberal or Conservative

|                | Sum of Squares | df   | Mean Square | F     | Sig. |
|----------------|----------------|------|-------------|-------|------|
| Between Groups | 22.591         | 11   | 2.054       | 1.120 | .341 |
| Within Groups  | 2601.119       | 1419 | 1.833       |       |      |
| Total          | 2623.711       | 1430 |             |       |      |

### Test of Homogeneity of Variances

Think of Self as Liberal or Conservative

| Levene Statistic | df1 | df2  | Sig. |
|------------------|-----|------|------|
| .579             | 11  | 1419 | .847 |

### Robust Tests of Equality of Means

Think of Self as Liberal or Conservative

|                | Statistic <sup>a</sup> | df1 | df2      | Sig. |
|----------------|------------------------|-----|----------|------|
| Welch          | 1.127                  | 11  | 547.334  | .337 |
| Brown-Forsythe | 1.112                  | 11  | 1358.122 | .348 |

a. Asymptotically F distributed.

# HOMOGENEITY OF VARIANCES

## Overview

### Welch test

The Welch test is an alternative to the ANOVA F test and is used when equality of group means cannot be assumed, as it cannot in this example. However, for this example, the Welch test also finds Zodiac not related to Polviews at a significant level.

### Bartlett's test of homogeneity of variance

This is an older, alternative test. Bartlett's test is a chi-square statistic with  $(k-1)$  degrees of freedom, where  $k$  is the number of categories in the independent variable. The Bartlett's test is dependent on meeting the assumption of normality and therefore Levene's test has now largely replaced it.

# HOMOGENEITY OF VARIANCES

## Overview

---

### F-max test

This test arises in the context of ANOVA. If the ratio of the largest to smallest size group in ANOVA is not very unequal (ex., is 4:1 or less), and if F-max (the ratio of the variance in the largest group to the variance in the smallest group) is 10:1 or less, then a rule of thumb is that homogeneity of variances is assumed not to be a problem. If groups are more unequal than 4:1, then the ratio of variance may need to be as low as 3:1.

# HOMOGENEITY OF VARIANCES

## Overview

---

### SPHERICITY

**Purpose.** In a repeated measures design, the univariate ANOVA tables will not be interpreted properly unless the variance/covariance matrix of the dependent variables is circular in form. Sphericity is an assumption of repeated measures MANOVA, for instance. To conclude that sphericity is not violated, *Bartlett's test of sphericity* should not be significant. In SPSS, choose Analyze, General Linear Model, Multivariate; click the Options button; check Residual SSCP matrix.

# HOMOGENEITY OF VARIANCE-COVARIANCE MATRICES

---

## Box's M test

Box's M tests the multivariate homogeneity of variances and covariances, as required by MANOVA and some other procedures. When M is not significant, the researcher accepts the null hypothesis that groups do not differ. It has been shown to be a conservative test, failing to reject the null hypothesis too often. It also is highly sensitive to violations of multivariate normality. In SPSS, choose Analyze, General Linear Model, Multivariate; click the Options button; check Homogeneity tests.

# HOMOGENEITY OF REGRESSIONS / TEST OF PARALLELISM

---

## **Analysis of variance**

In Ancova and Mancova, the slopes of the regression lines should be the same for each group formed by the categorical variables and measured on the dependents. The more this assumption is violated, the more conservative ANCOVA and

# HOMOGENEITY OF VARIANCES

## Overview

Mancova become (the more likely to make Type II errors - failing to reject null hypotheses, resulting in false negatives).

Violation of the homogeneity of regressions assumption indicates an interaction effect between the covariate(s) and the factor(s). When running an ANCOVA or MANCOVA model in SPSS, include in the model options the interactions between the covariate(s) and each independent factor -- any significant interaction effects indicate that the assumption of homogeneity of regression coefficients has been violated.

### Homogeneity of regression

Homogeneity of regression in SPSS can be tested under the Model button of Analyze, General Linear Model, Univariate; select Custom under the Model button; enter a model with all main effects of the factors and covariates and the interaction of the covariate(s) with the factor(s). These interaction effects should be non-significant if the homogeneity of regressions assumption is met.

### Parallelism tests

Tests of parallelism are similar in certain procedures, assuring that the slope of regression lines is the same for each level of a categorical grouping variable. Ordinal regression and the probit response model, for instance, come with a test of parallelism in SPSS and other packages.

# HOMOGENEITY OF VARIANCES

## Overview

### HOMOSCEDASTICITY

#### Graphical method

Homoscedasticity means the relationship under investigation is the same for the entire range of the dependent variable. Lack of homoscedasticity is shown by higher errors (residuals) for some portions of the range compared to others. When the homoscedasticity assumption is met, residuals will form a patternless cloud of dots. Lack of homoscedasticity is most easily seen in a standardized scatterplot. This scatterplot of the standardized predicted dependent variable (ZPR\_1 in SPSS) against the standardized residuals (ZRE\_1 in SPSS, or one may use studentized deleted residuals, SRE\_1; or one might use the observed dependent vs. ZRE\_1 or SRE\_1) should show a random pattern across the entire range of ZPR\_1 when, in regression, error is homoscedastic -- that is, when the regression

## HOMOGENEITY OF VARIANCES

### Overview

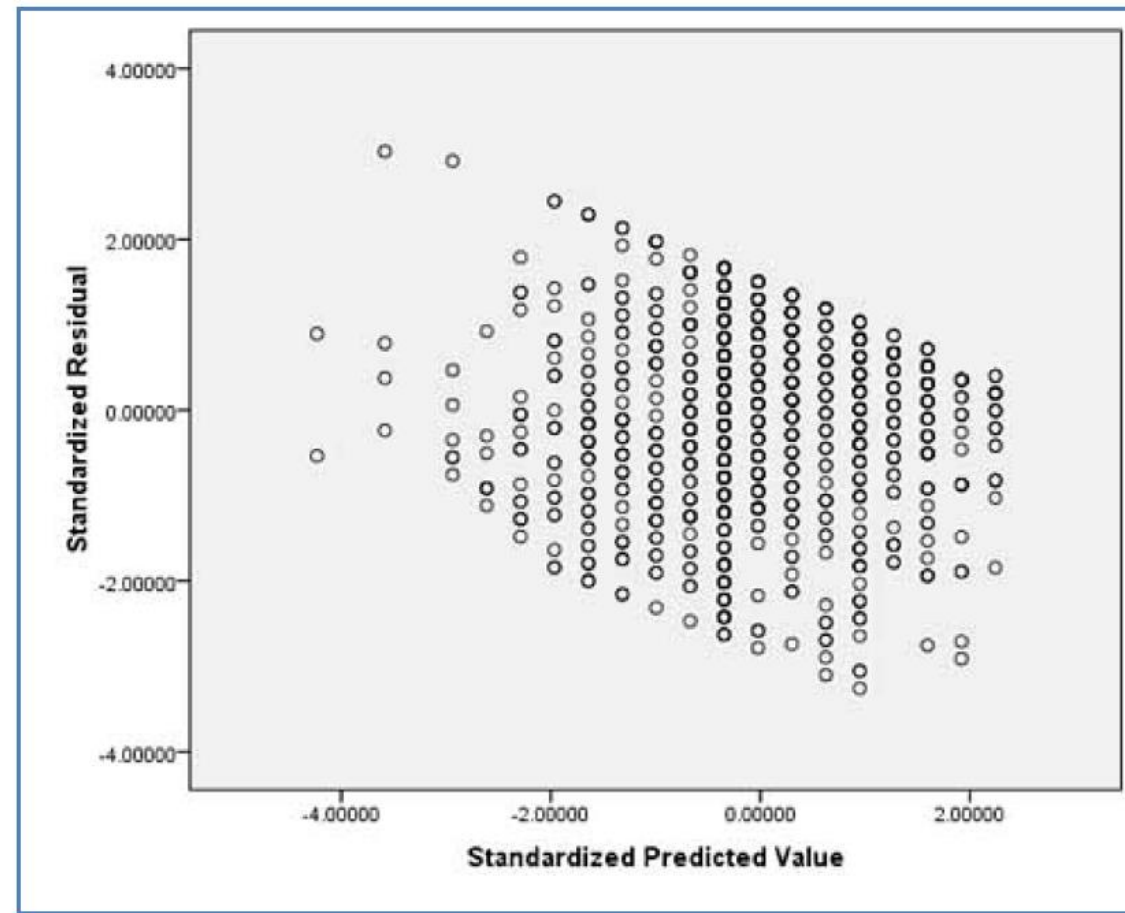
---

model is equally accurate across the range of the dependent. Sometimes a plot of residuals by the actual dependent values is used to test homoscedasticity also. In SPSS, select Analyze, Regression, Linear; click Plots.

If ZPR\_1, ZRE\_1, or other needed variables have been saved, you can also use Graphs, Legacy Dialogs, Scatter/Dot. In the output below, for instance, education was used to predict income and the standardized predicted and residual values were saved. The plot is largely a cloud (indicating homoscedasticity) but there is some pattern showing that higher predicted values have lower residuals (lack of homoscedasticity).

# HOMOGENEITY OF VARIANCES

## Overview



# HOMOGENEITY OF VARIANCES

## Overview

---

### **Weighted least squares regression**

WLS is a commonly used strategy for dealing with heteroscedasticity in a linear regression context. See the separate "blue book" volume on WLS.

### **Goldfeld-Quandt test**

This test is a formal test of homoscedasticity used when errors take a funnel (fan) shape. Though not directly supported by SPSS, this test involves running separate

# HOMOGENEITY OF VARIANCES

## Overview

---

regressions on the upper and lower observations in terms of values on the dependent, then conducting an F-test on the ratio of the error sum of squares (ESS) for the two regressions. To the extent that the ratio of the two ESS coefficients approaches 1.0, the homoscedasticity assumption is met.

### Glejser test

This test is a formal test of homoscedasticity used when errors take a bow-tie (bimodal) shape and when sample size is not small. The residuals are regressed on the independent variable (or some function of it) and if the independent has a significant b coefficient, then the researcher concludes that homoscedasticity cannot be assumed.

# HOMOGENEITY OF VARIANCES

## Overview

### Park test

The Park test is a related test of homoscedasticity. The squared residuals (or their log) is regressed on the independent variable (or its log), and if the independent has a significant  $b$  coefficient, then the researcher concludes that homoscedasticity cannot be assumed.

### Breusch-Pagan-Godfrey test

This test is another large-sample test of homoscedasticity. The squared residuals are standardized by dividing by the mean squared residual (regression sum of squares (RSS) divided by  $N$ ), giving the generalized residuals. The generalized residuals are then regressed on all independent variables ( $m$  variables) suspected of causing heteroscedasticity. The error sum of squares (ESS) divided by 2, for this regression, follows a chi-square distribution with  $(m - 1)$  degrees of freedom. A finding of significance means the null hypothesis is rejected and homoscedasticity cannot be assumed. The B-P-G test is more powerful than Goldfeld-Quandt or White's tests.

### White's test

White's test involves regressing the squared residuals on all independent variables plus all squares of independent variables plus all crossproducts of independent variables (for a total of  $t$  predictor variables in all). The resulting  $R^2$  is multiplied by  $N$ , giving a statistic with a chi-square distribution with  $t$  degrees of freedom. If the null hypothesis is rejected, homoscedasticity cannot be assumed.

## HOMOGENEITY OF VARIANCES

### Overview

---

White's test does not require prior knowledge of the form of heteroscedasticity. It is less powerful than other tests such as Goldfeld-Quandt.

# TESTS OF LINEARITY

---

- Graphical Method (SPSS)
- Curve Fitting with R-Squared Difference Test (SPSS)
- ANOVA (SPSS)
- Eta the Correlation Ratio (SPSS)
- Adding Non-Linear Terms to the Model
- Ramsey's Reset Test (STATA)

## 80 LINEARITY

### LINEARITY

Testing for nonlinearity is necessary because correlation, regression, and other members of the general linear model (GLM) assume linearity. A variety of methods are available.

#### **Graphical methods**

Simple inspection of scatterplots is a common if non-statistical method of determining if nonlinearity exists in a relationship. Even better, a plot of standardized residuals against standardized estimates (fitted values) of the dependent variable should show a random pattern when nonlinearity is absent. In regression, as a rule of thumb, an indicator of possible nonlinearity is when the standard deviation of the residuals exceeds the standard deviation of the dependent. Adding to a model nonlinear terms such as squares or cubes of an independent and then seeing if  $R^2$  in regression or fit indexes in structural equation modeling improve significantly is another way of testing for nonlinearity.

# WELCH'S TEST

---

- **Welch's ANOVA** compares two means to see if they are equal. It is an alternative to the Classic ANOVA and can be used even if your data violates the assumption of homogeneity of variances.
- **You should run Welch's test in all cases where you have normally distributed data that violates the assumption of homogeneity of variance.** ANOVA (and the non-parametric alternative Kruskal-Wallis) are very unstable for these situations, producing Type I error rates that are:
  - **Conservative** for large sample sizes and
  - **Inflated** for small sample size.
- Welch's ANOVA is fast becoming the go-to method out of the three. For normal, different-variance, and balanced data (i.e. same-size samples), Welch's has the most power and the lowest type I error rate. However, classic ANOVA still performs the best when data is normal, equal-variance, and is either balanced or unbalanced. In other words, **run Welch's if your data has unequal variances, but run a classic ANOVA if it's just an unequal sample size issue.**

# WELCH'S TEST

---

- The assumptions are pretty much the same for Welch's ANOVA as for the classic ANOVA. For example, the assumption of normality still holds. However, you *should* run Welch's when you violate the assumption of equal variances. You *can* run it with unequal sample sizes.
- In **Minitab**: the Assistant automatically runs Welch's when you choose an ANOVA test. You can also run it by clicking "ANOVA > One-Way", then clicking "Options" and unchecking "Assume equal variances."
- In **SAS**: Add a MEANS statement, the nominal variable name, and then type a slash followed by "WELCH". For example:  
PROC anova;  
CLASS A B;  
MODEL Y=A B;  
MEANS AB /WELCH;  
RUN;

# WELCH'S TEST

- In **SPSS**, click “Analyze > Compare Means > One-Way ANOVA”. Then click “Options” and check both the “Homogeneity of variance” test and the “Welch” box. This will test for homogeneity of variance and then — if the assumption is violated — you can use the Welch statistic (otherwise you can choose to use Sig. in the regular ANOVA output instead of the Sig. reported in the Welch area).
- *Welch's  $p$ -value in SPSS can be used to replace the regular ANOVA  $p$ -value.*

| ANOVA                              |                |     |             |       |      |
|------------------------------------|----------------|-----|-------------|-------|------|
| Factory Model Design – Total Score |                |     |             |       |      |
|                                    | Sum of Squares | df  | Mean Square | F     | Sig. |
| Between Groups                     | 1702.277       | 7   | 243.182     | 3.774 | .002 |
| Within Groups                      | 26473.223      | 159 | 166.435     |       |      |
| Total                              | 28175.500      | 166 |             |       |      |

| Robust Tests of Equality of Means  |                        |     |         |      |
|------------------------------------|------------------------|-----|---------|------|
| Factory Model Design – Total Score |                        |     |         |      |
|                                    | Statistic <sup>a</sup> | df1 | df2     | Sig. |
| Welch                              | 2.855                  | 7   | 159.465 | .003 |

a. Asymptotically F distributed.

## 84 LINEARITY

### Curve fitting with R-squared difference tests

The SPSS curve fitting module (Analyze, Regression, Curve Fit) will calculate R-squared for a linear model compared to a variety of nonlinear models. One can then employ an F test of R-squared difference between models to see if nonlinear model has a significantly higher R-squared than a linear one. Note that some nonlinear models cannot be fitted if there are missing values, so interpolating using the Missing Values module may be a necessary prerequisite if the researcher wishes to compare these models. Warning: this method involves multiple comparisons which make the nominal significance level more lenient than it appears, so a Bonferroni or other adjustment may be necessary depending on the number of comparisons made.

### ANOVA test of linearity

One can compute an ANOVA table for the linear and nonlinear components of any pair of variables. If the F significance value for the nonlinear component is below the critical value (ex.,  $< .05$ ), then there is significant nonlinearity. In SPSS, select

## 85 LINEARITY

Analyze, Compare Means, Means; click Options; check Linearity test. Likewise, the Contrasts option in ANOVA can be used as a test for the existence of linear, quadratic, and other polynomial relationships.

### **Eta, the correlation ratio**

Eta is a coefficient of nonlinear association. For linear relationships, eta equals the correlation coefficient (Pearson's  $r$ ). For nonlinear relationships it is greater -- hence the difference between eta and  $r$  is a measure of the extent of nonlinearity of relationship. Eta is discussed in the separate "blue book" volume on measures of association. In SPSS, select Analyze, Compare Means, Means; click Options; check Anova table and eta.

### **Adding nonlinear terms to a model**

A square term corresponds to a nonlinear pattern with a single curve, upward or downward. A cubic term corresponds to a nonlinear up-then-down or down-then-up pattern. In general, the number of curves is one less than the order of the polynomial term -- for instance, the 4th power will produce three curves. A more shotgun approach is to add all possible power terms to the model, then drop those which are non-significant.

## 86 LINEARITY

### **Ramsey's RESET test (regression specification error test)**

Ramsey's general test of specification error of functional form is an F test of differences of  $R^2$  under linear versus nonlinear assumptions. It is commonly used in time series analysis to test whether power transforms need to be added to the model. For a linear model which is properly specified in functional form, nonlinear transforms of the fitted values should not be useful in predicting the dependent variable. While STATA and some packages label the RESET test as a test to see if there are "no omitted variables," it is a linearity test, not a general specification test. It tests if any nonlinear transforms of the specified independent variables have been omitted. It does not test whether other relevant linear or nonlinear variables have been omitted.

1. Run the regression to obtain  $R_o^2$ , the original multiple correlation.
2. Save the predicted values (Y's).
3. Re-run the regression using power functions of the predicted values (ex., their squares and cubes) as additional independents for the Ramsey RESET

## 87 LINEARITY

---

test of functional form where testing that none of the independents is nonlinearly related to the dependent. Alternatively, re-run the regression using power functions of the independent variables to test them individually.

4. Obtain  $R_n^2$ , the new multiple correlation.
5. Apply the F test, where  $F = (R_n^2 - R_o^2)/[(1 - R_n^2)/(n-p)]$ , where n is sample size and p is the number of parameters in the new model.
6. Interpret F: For an adequately specified model, F should be non-significant.

# TESTING FOR MULTICOLLINEARITY

---

- Tolerance (SPSS)
- Variance Inflation Factor (SPSS)
- Condition Indices (SPSS)
- Multicollinearity in Structural Equations Modelling (SEM)
- Intra Class Correlation

## 89 MULTICOLLINEARITY

### MULTICOLLINEARITY

Multicollinearity is an unacceptably high level of intercorrelation among the independents, such that the effects of the independents cannot be separated. Under multicollinearity, estimates are unbiased but assessments of the relative strength of the explanatory variables and their joint effect are unreliable. (That is, beta weights and R-squares cannot be interpreted reliably even though predicted values are still the best estimate using the given independents). As a rule of thumb, intercorrelation among the independents above .80 signals a possible problem. Likewise, high multicollinearity is signalled when high R-squared and significant F tests of the model occur in combination with non-significant t-tests of coefficients.

#### Example

An example of multicollinearity occurred in a Bureau of Labor Statistics study of the price of camcorders. The initial model included the dummy variables Sony and 8mm, both of which corresponded to high price. However, since Sony was the only manufacturer of 8mm camcorders at the time, the Sony and 8 mm dummy variables were multicollinear. A similar multicollinearity occurred in a BLA study of washing machines, where it was found that "capacity" and "number of cycles" were multicollinear. In each study, one of the collinear variables had to be dropped from the model.

Whereas perfect multicollinearity leads to infinite standard errors and indeterminate coefficients, the more common situation of high multicollinearity leads to large standard errors, large confidence intervals, and diminished power (the chance of Type II errors is high - thinking you do not have a relationship when in fact one exists - failure to reject the null hypothesis that the coefficients are not

## 90 MULTICOLLINEARITY

different from zero). R-square is high. The coefficients and their standard errors will be sensitive to changes in just a few observations. Methods of handling high multicollinearity are discussed in the "blue book" volume on multiple regression.

### **Tolerance**

Tolerance is defined as  $1 - R\text{-squared}$ , where R-squared is the multiple R of a given independent regressed on all other independent variables. If the tolerance value is less than some cutoff value, usually .20, the independent should be dropped from the analysis due to multicollinearity. This is better than just using simple  $r > .80$  since tolerance looks at the independent variable in relation to all other independents and thus takes interaction effects into account as well as simple correlations. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

# 91 MULTICOLLINEARITY

## **Variance inflation factor, VIF**

Note, the variance-inflation factor, VIF, may be used in lieu of tolerance as VIF is simply the reciprocal of tolerance. The rule of thumb is that  $VIF > 4.0$  when multicollinearity is a problem. Some authors use the more lenient cut-off of  $VIF \geq 5$  when multicollinearity is a problem. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

## **Condition indices.**

Discussed more extensively in the "blue book" volume on multiple regression, condition indices over 15 indicate possible multicollinearity problems and over 30 indicate serious multicollinearity problems. In SPSS, select Analyze, Regression, linear; click Statistics; check Collinearity diagnostics.

## **Multicollinearity in Structural Equation Models (SEM)**

*Standardized regression weights:* Since all the latent variables in a SEM model have been assigned a metric of 1, all the standardized regression weights should be within the range of plus or minus 1. When there is a multicollinearity problem, a weight close to 1 indicates the two variables are close to being identical. When these two nearly identical latent variables are then used as causes of a third latent variable, the SEM method will have difficulty computing separate regression weights for the two paths from the nearly-equal variables and the third

## 92 MULTICOLLINEARITY

variable. As a result it may well come up with one standardized regression weight greater than +1 and one weight less than -1 for these two paths.

*Standard errors of the unstandardized regression weights:* Likewise, when there are two nearly identical latent variables, and these two are used as causes of a third latent variable, the difficulty in computing separate regression weights may well be reflected in much larger standard errors for these paths than for other paths in the model, reflecting high multicollinearity of the two nearly identical variables.

*Covariances of the parameter estimates:* Likewise, the same difficulty in computing separate regression weights may well be reflected in high covariances of the parameter estimates for these paths - estimates much higher than the covariances of parameter estimates for other paths in the model.

*Variance estimates:* Another effect of the same multicollinearity syndrome may be negative variance estimates. In the example above of two nearly-identical latent variables causing a third latent variable, the variance estimate of this third variable may be negative.

## 93 DATA INDEPENDENCE

### DATA INDEPENDENCE

#### Lack of independence

Independent observations are assumed by most statistical procedures, including multiple regression, logistic regression, and members of the general linear model (GLM) family. Lack of independence occurs in three broad classes of research situations.

- *Repeated measures data.* Before-after studies, panel studies, and paired comparison data measure the same subject at multiple times, where the subject's response at time  $t$  is correlated with the same subject's response at time  $t+1$ . Repeated measures modifications for GLM and some other statistical procedures are available.
- *Time series data.* Time series, such as budget data over time, is non-independent because of autocorrelation: data at time  $t$  is correlated with time  $t+1$  (or other lags). Time series analysis adjusts for autocorrelation.

## 94 DATA INDEPENDENCE

- *Hierarchical and grouped data.* In general, whenever data are grouped in some way there is the possibility they are non-independent because knowing the group for a case helps predict its value. Because data are apt to cluster by group, predictions will err when not taking grouping into account, leading to correlated error. Grouping may be by hierarchical layer (ex., census tract and city in multilevel sampling of individuals within tracts within cities), by observation time (ex., individual test scores grouped by year administered), and by any grouping factor (ex., religion). While non-independence may be addressed through repeated measures adaptations of conventional statistical procedures or by entering the grouping factor as a set of dummy variables, linear mixed modeling (a.k.a., hierarchical linear modeling or multilevel modeling) is the most rigorous approach to non-independence.

### **Intra-class correlation (ICC)**

A now-conventional test for non-independence is to construct a null linear mixed model with the dependent as level 1 and the grouping factor as level 2, without other predictors at either level. If the ICC is significant, there is significant non-independence in the data. Computing ICC is discussed in the separate "blue book" volume on linear mixed modeling.

# TESTS FOR AUTOCORRELATION

---

- Graphical Method (SPSS)
- Runs Test (SPSS)
- Durbin Watson d-test (SPSS)
- Breusch- Godfrey Test (Lagrange's Multiplier Test) (EXCEL)
- Akaike-Schwartz Information Criteria (EIEWS)
- Autocorrelation Function (ACF) & Partial Autocorrelation Function (PACF) (EIEWS)

## 96 AUTOCORRELATION

### **Durbin-Watson coefficient**

Independence may also be tested by the Durbin-Watson coefficient, which uses studentized residuals. The Durbin-Watson statistic should be between 1.5 and 2.5 for independent observations. In SPSS regression output, it is found in the "Model Summary" table. In SPSS, select Analyze, Regression, Linear; click Statistics; check Durbin-Watson.

### **Graphical method**

Residuals may be plotted against the case identification number, when cases are ordered by any factor potentially causing non-independence. This factor may be time, a grouping factor, or simply interview/data collection order. There should be no pattern to this plot if observations are independent. In SPSS, select Graphs, Scatterplot.

# MODEL SPECIFICATION TESTS

---

- Best Parsimonious Model:
- Searching for of best feasible model out of possible Best models for the given Data Set
- A parsimonious model is a model that accomplishes a desired level of explanation or prediction with as few predictor variables as possible.

# 98 MODEL SPECIFICATION

## PROPER MODEL SPECIFICATION

### Specification of a model

Specification refers to not omitting significant causal variables or including correlated but causally extraneous ones, and also to correctly indicating the direction of arrows connecting the variables in the model. When a misspecification error is corrected by changing the model, all parameter estimates in the model are subject to change, not only in magnitude, but sometimes even in direction. There is no statistical test for misspecification. A good literature review is important in identifying variables which need to be specified.

As a rule of thumb, the lower the overall effect (ex.,  $R^2$  in multiple regression, goodness of fit in logistic regression), the more likely it is that important variables have been omitted from the model and that existing interpretations of the model will change when the model is correctly specified. The specification problem is lessened when the research task is simply to compare models to see which has a better fit to the data, as opposed to the purpose being to justify one model and then assess the relative importance of the independent variables.

# THANK YOU

---

Have a Good Day