Appendix U

Chapter 2. Frequencies

The Frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables. The Frequencies procedure is a good place to start looking at your data.

For a frequency report and bar chart, you can arrange the distinct values in ascending or descending order, or you can order the categories by their frequencies. The frequencies report can be suppressed when a variable has many distinct values. You can label charts with frequencies (the default) or percentages.

Example. What is the distribution of a company's customers by industry type? From the output, you might learn that 37.5% of your customers are in government agencies, 24.9% are in corporations, 28.1% are in academic institutions, and 9.4% are in the healthcare industry. For continuous, quantitative data, such as sales revenue, you might learn that the average product sale is $3,576, with a standard deviation of $1,078.

Statistics and plots. Frequency counts, percentages, cumulative percentages, mean, median, mode, sum, standard deviation, variance, range, minimum and maximum values, standard error of the mean, skewness and kurtosis (both with standard errors), quartiles, user-specified percentiles, bar charts, pie charts, and histograms.

Frequencies Data Considerations

Data. Use numeric codes or strings to code categorical variables (nominal or ordinal level measurements).

Assumptions. The tabulations and percentages provide a useful description for data from any distribution, especially for variables with ordered or unordered categories. Most of the optional summary statistics, such as the mean and standard deviation, are based on normal theory and are appropriate for quantitative variables with symmetric distributions. Robust statistics, such as the median, quartiles, and percentiles, are appropriate for quantitative variables that may or may not meet the assumption of normality.

To Obtain Frequency Tables

1. From the menus choose:
   Analyze > Descriptive Statistics > Frequencies...
2. Select one or more categorical or quantitative variables.

Optionally, you can:

- Click **Statistics** for descriptive statistics for quantitative variables.
- Click **Charts** for bar charts, pie charts, and histograms.
- Click **Format** for the order in which results are displayed.

Frequencies Statistics

Percentile Values. Values of a quantitative variable that divide the ordered data into groups so that a certain percentage is above and another percentage is below. Quartiles (the 25th, 50th, and 75th percentiles) divide the observations into four groups of equal size. If you want an equal number of groups other than four, select **Cut points for n equal groups.** You can also specify individual percentiles (for example, the 95th percentile, the value below which 95% of the observations fall).

Chapter 4. Explore

The Explore procedure produces summary statistics and graphical displays, either for all of your cases or separately for groups of cases. There are many reasons for using the Explore procedure--data screening, outlier identification, description, assumption checking, and characterizing differences among subpopulations (groups of cases). Data screening may show that you have unusual values, extreme values, gaps in the data, or other peculiarities. Exploring the data can help to determine whether the statistical techniques that you are considering for data analysis are appropriate. The exploration may indicate that you need to transform the data if the technique requires a normal distribution. Or you may decide that you need nonparametric tests.

Example. Look at the distribution of maze-learning times for rats under four different reinforcement schedules. For each of the four groups, you can see if the distribution of times is approximately normal and whether the four variances are equal. You can also identify the cases with the five largest and five smallest times. The boxplots and stem-and-leaf plots graphically summarize the distribution of learning times for each of the groups.

Statistics and plots. Mean, median, 5% trimmed mean, standard error, variance, standard deviation, minimum, maximum, range, interquartile range, skewness and kurtosis and their standard errors, confidence interval for the mean (and specified confidence level), percentiles, Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, Tukey's biweight estimator, the five largest and five smallest values, the Kolmogorov-Smirnov statistic with a Lilliefors significance level for testing normality, and the Shapiro-Wilk statistic. Boxplots, stem-and-leaf plots, histograms, normality plots, and spread-versus-level plots with Levene tests and transformations.

Explore Data Considerations

Data. The Explore procedure can be used for quantitative variables (interval- or ratio-level measurements). A factor variable (used to break the data into groups of cases) should have a reasonable number of distinct values (categories). These values may be short string or numeric. The case label variable, used to label outliers in boxplots, can be short string, long string (first 15 bytes), or numeric.

Assumptions. The distribution of your data does not have to be symmetric or normal.

To Explore Your Data

1. From the menus choose:
   Analyze > Descriptive Statistics > Explore...
2. Select one or more dependent variables.

Optionally, you can:

- Select one or more factor variables, whose values will define groups of cases.
- Select an identification variable to label cases.
- Click **Statistics** for robust estimators, outliers, percentiles, and frequency tables.
- Click **Plots** for histograms, normal probability plots and tests, and spread-versus-level plots with Levene's statistics.
- Click **Options** for the treatment of missing values.

Explore Statistics

Descriptives. These measures of central tendency and dispersion are displayed by default. Measures of central tendency indicate the location of the distribution; they include the mean, median, and 5% trimmed mean. Measures of dispersion show the dissimilarity of the values; these include standard error, variance, standard deviation, minimum, maximum, range, and interquartile range. The descriptive statistics also include measures of the shape of the distribution; skewness and kurtosis are displayed with their standard errors. The 95% level confidence interval for the mean is also displayed; you can specify a different confidence level.

M-estimators. Robust alternatives to the sample mean and median for estimating the location. The estimators calculated differ in the weights they apply to cases. Huber's M-estimator, Andrews' wave estimator, Hampel's redescending M-estimator, and Tukey's biweight estimator are displayed.

Outliers. Displays the five largest and five smallest values with case labels.

Percentiles. Displays the values for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles.

Explore Plots

Boxplots. These alternatives control the display of boxplots when you have more than one dependent variable. **Factor levels together** generates a separate display for each dependent variable. Within a display, boxplots are shown for each of the groups defined by a factor variable. **Dependents together** generates a separate display for each group defined by a factor variable. Within a display, boxplots are shown side by side for each dependent variable. This display is particularly useful when the different variables represent a single characteristic measured at different times.

Descriptive. The Descriptive group allows you to choose stem-and-leaf plots and histograms.

Normality plots with tests. Displays normal probability and detrended normal probability plots. The Kolmogorov-Smirnov statistic, with a Lilliefors significance level for testing normality, is displayed. If non-integer weights are specified, the Shapiro-Wilk statistic is calculated when the weighted sample size lies between 3 and 50. For no weights or integer weights, the statistic is calculated when the weighted sample size lies between 3 and 5,000.

Spread vs. Level with Levene Test. Controls data transformation for spread-versus-level plots. For all spread-versus-level plots, the slope of the regression line and Levene's robust tests for homogeneity of variance are displayed. If you select a transformation, Levene's tests are based on the transformed data. If no factor variable is selected, spread-versus-level plots are not produced. **Power estimation** produces a plot of the natural logs of the interquartile ranges against the natural logs of the medians for all cells, as well as an estimate of the power transformation for achieving equal variances in the cells. A spread-versus-level plot helps to determine the power for a transformation to stabilize (make more equal) variances across groups. **Transformed** allows you to select one of the power alternatives, perhaps following the recommendation from power estimation, and produces plots of transformed data. The interquartile range and median of the transformed data are plotted. **Untransformed** produces plots of the raw data. This is equivalent to a transformation with a power of 1.

Explore Power Transformations

These are the power transformations for spread-versus-level plots. To transform data, you must select a power for the transformation. You can choose one of the following alternatives:

- **Natural log.** Natural log transformation. This is the default.
- **1/square root.** For each data value, the reciprocal of the square root is calculated.
- **Reciprocal.** The reciprocal of each data value is calculated.
- **Square root.** The square root of each data value is calculated.

Explore

A1 - 37 Sales: Greater than 10000sqft and less than 65000sqft

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Total Net Area	37	100.0%	0	0.0%	37	100.0%

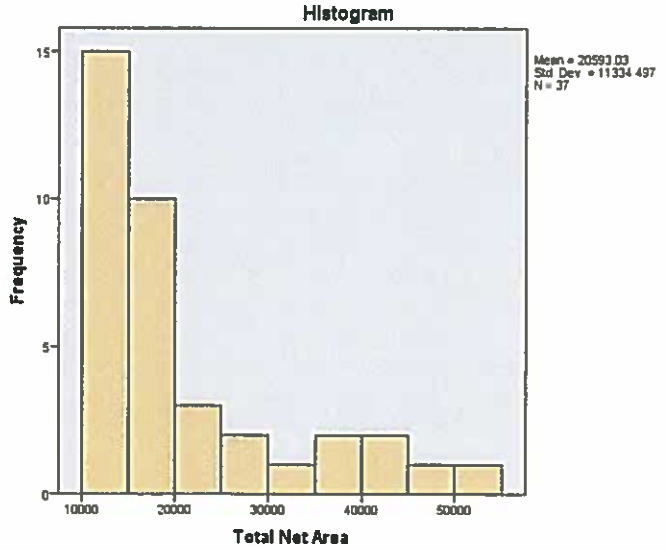
Descriptives

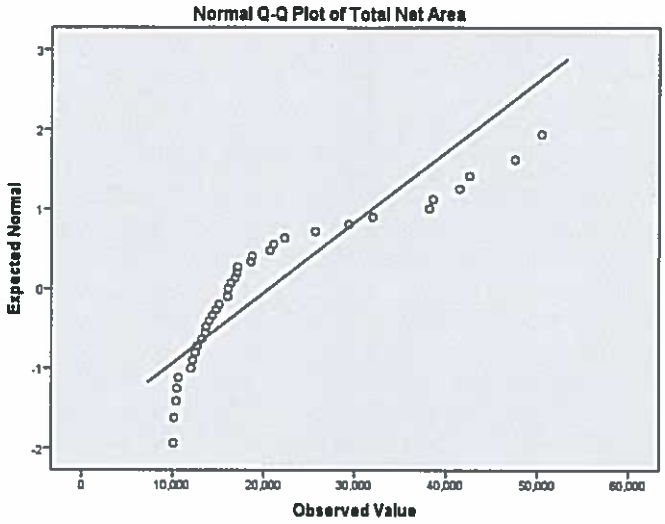
			Statistic	Std. Error
Total Net Area	Mean		20593.03	1863.380
	95% Confidence Interval for Mean	Lower Bound	16813.92	
		Upper Bound	24372.14	
	5% Trimmed Mean		19591.19	
	Median		16050.00	
	Variance		128470820.900	
	Std. Deviation		11334.497	
	Minimum		10057	
	Maximum		50462	
	Range		40405	
	Interquartile Range		11033	
	Skewness		1.394	.388
	Kurtosis		.818	.759

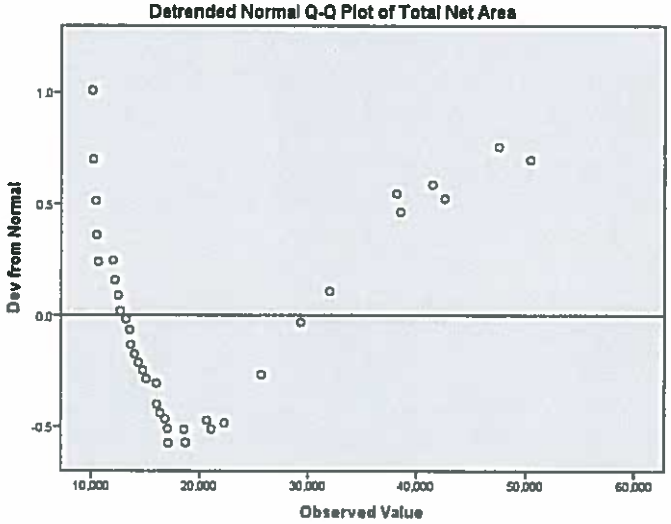
Tests of Normality

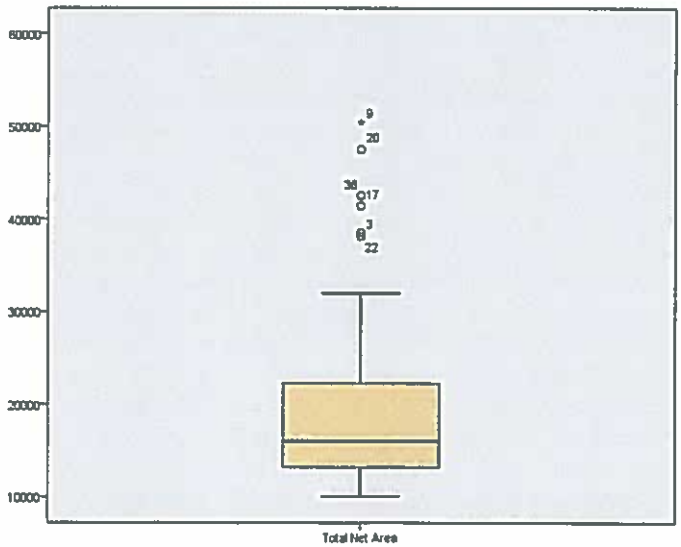
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Total Net Area	.243	37	.000	.795	37	.000

a. Lilliefors Significance Correction




Normal Q-Q Plot of Total Net Area
Expected Normal
Observed Value
3
2
1
0
-1
-2
0
10,000
20,000
30,000
40,000
50,000
60,000


Detrended Normal Q-Q Plot of Total Net Area
Dev from Normal
Observed Value
1.0
0.5
0.0
-0.5
10,000
20,000
30,000
40,000
50,000
60,000


60000
50000
40000
30000
20000
10000
9
20
38
17
3
22
Total Net Area