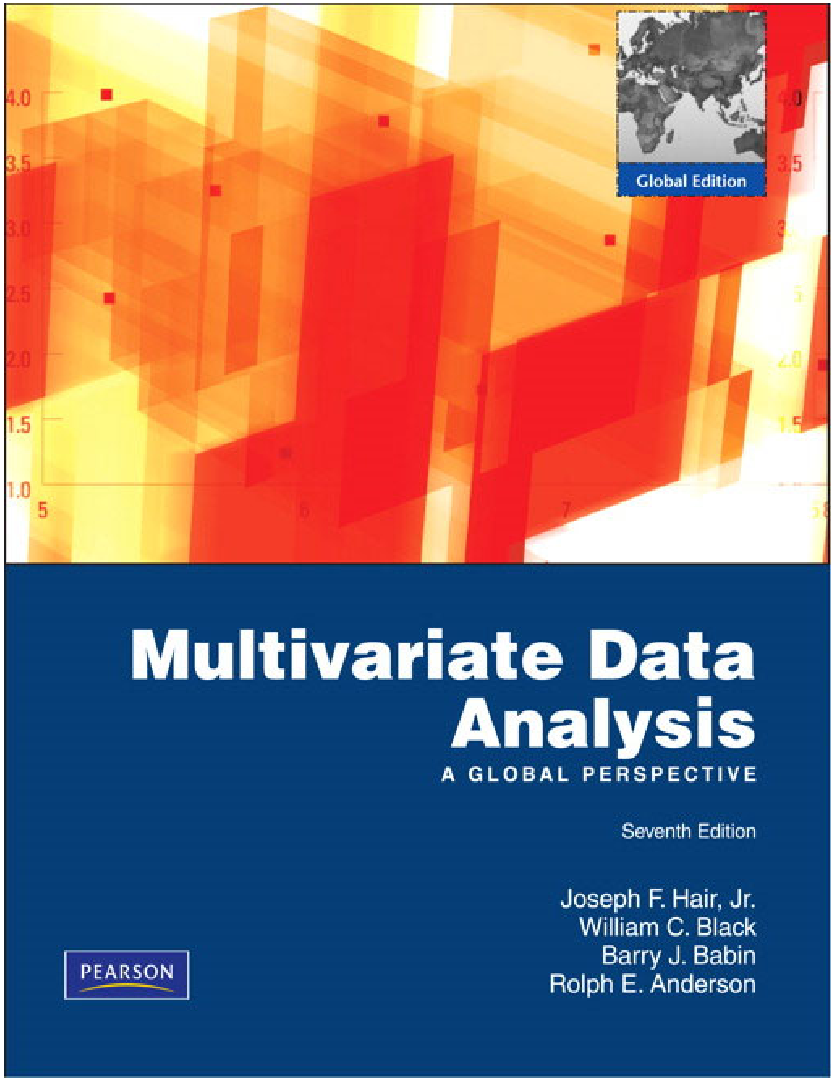
Chapter 2 Cleaning and Transforming Data



By Yen-I Chiang (江彥逸),

Department of Information Management

Chapter 2 Cleaning & Transforming Data

LEARNING OBJECTIVES

Upon completing this chapter, you should be able to do the following:

- Select the appropriate graphical method to examine the characteristics of the data or relationships of interest.
- Assess the type and potential impact of missing data.
- Understand the different types of missing data processes.
- Explain the advantages and disadvantages of the approaches available for dealing with missing data.

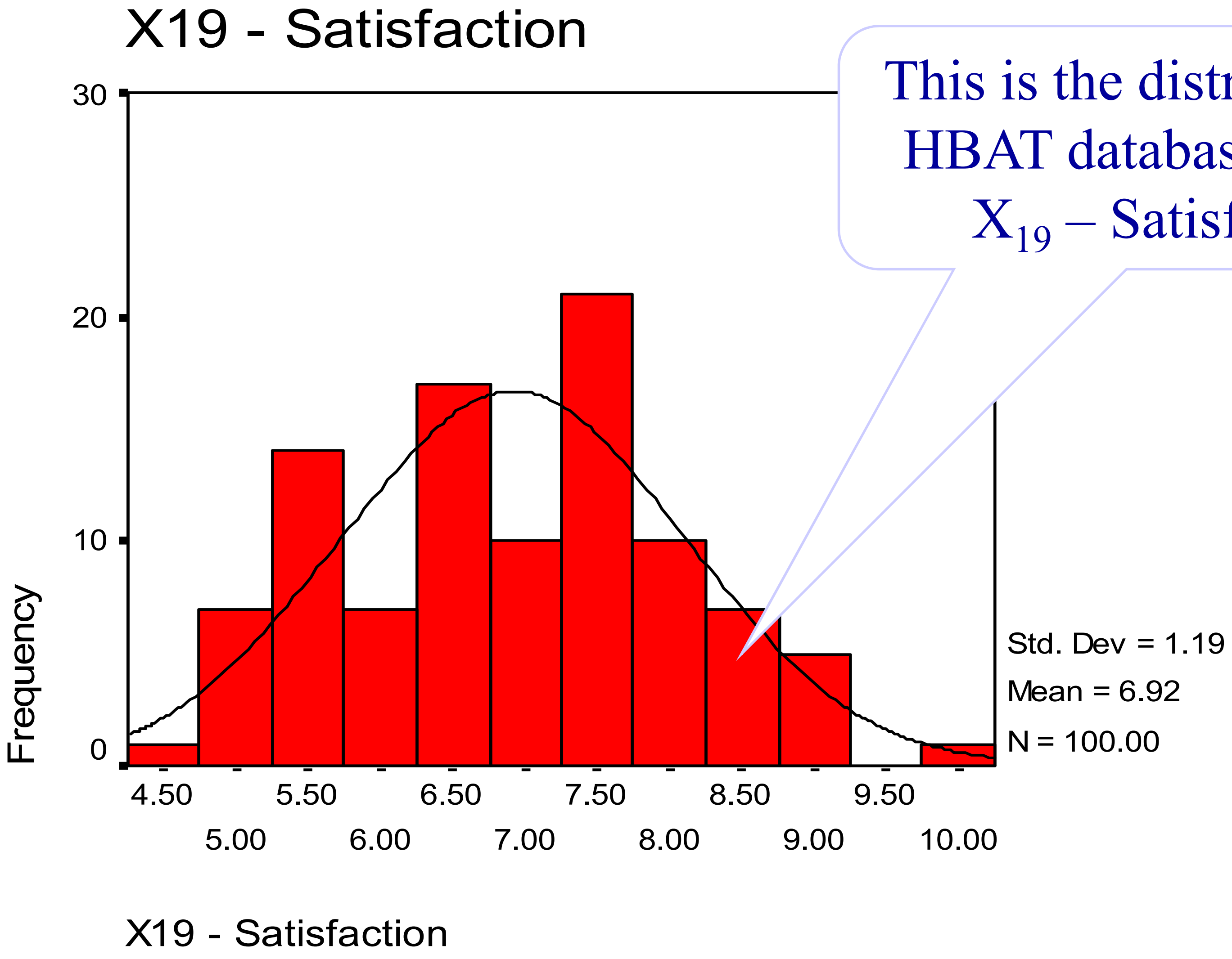
Cleaning Your Data

- Graphical examination.
- Identify and evaluate missing values.
- Identify and deal with outliers.
- Check whether statistical assumptions are met.
- Develop a preliminary understanding of your data.

Graphical Examination

- Shape:
  - ✓ Histogram
  - ✓ Bar Chart
  - ✓ Box & Whisker plot
  - ✓ Stem and Leaf plot
- Relationships:
  - ✓ Scatterplot
  - ✓ Outliers

Histograms and The Normal Curve

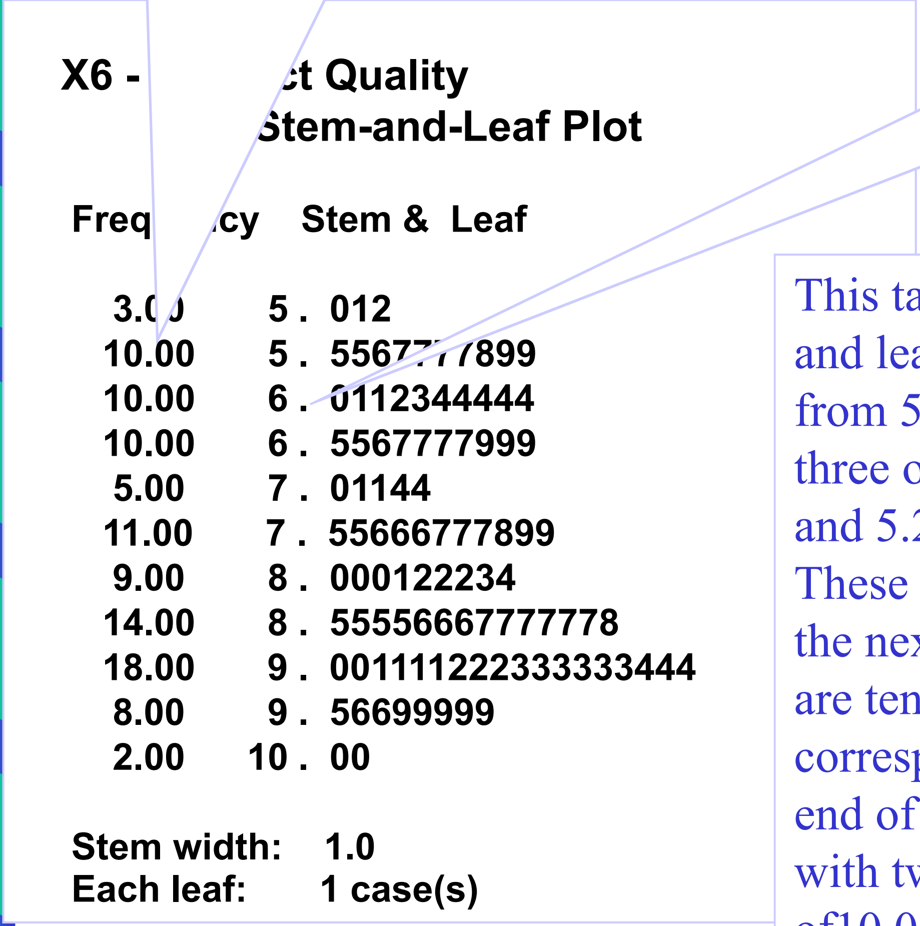


This is the distribution for HBAT database variable X_{19} – Satisfaction.

Stem & Leaf Diagram – HBAT Variable X_6

Each stem is shown by the numbers, and each number is a leaf. This stem has 10 leaves.

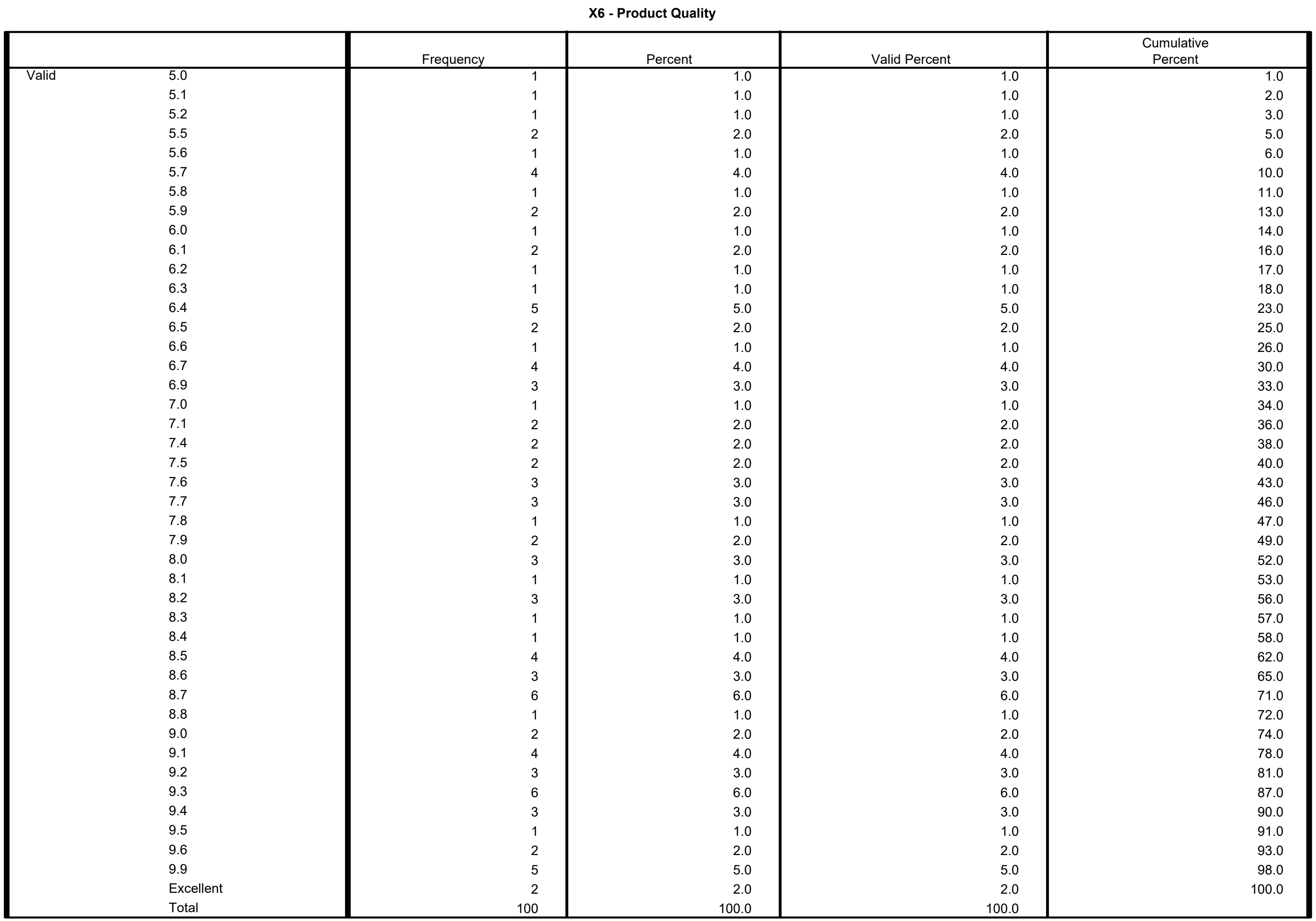
The length of the stem, indicated by the number of leaves, shows the frequency distribution. For this stem, the frequency is 14.



```
X6 -          ct Quality
              Stem-and-Leaf Plot

 Freq     ıcy    Stem &  Leaf

     3.00         5 .  012
    10.00         5 .  5567777899
    10.00         6 .  0112344444
    10.00         6 .  5567777999
     5.00         7 .  01144
    11.00         7 .  55666777899
     9.00         8 .  000122234
    14.00         8 .  55556667777778
    18.00         9 .  001111222333333444
     8.00         9 .  56699999
     2.00        10 .  00

 Stem width:     1.0
 Each leaf:      1 case(s)
```

This table shows the distribution of X6 with a stem and leaf diagram (Figure 2.2). The first category is from 5.0 to 5.5, thus the stem is 5.0. There are three observations with values in this range (5.0, 5.1 and 5.2). This is shown as three leaves of 0, 1 and 2. These are also the three lowest values for X6. In the next stem, the stem value is again 5.0 and there are ten observations, ranging from 5.5 to 5.9. These correspond to the leaves of 5.5 to 5. 9. At the other end of the figure, the stem is 10.0. It is associated with two leaves (0 and 0), representing two values of10.0, the two highest values for X6.

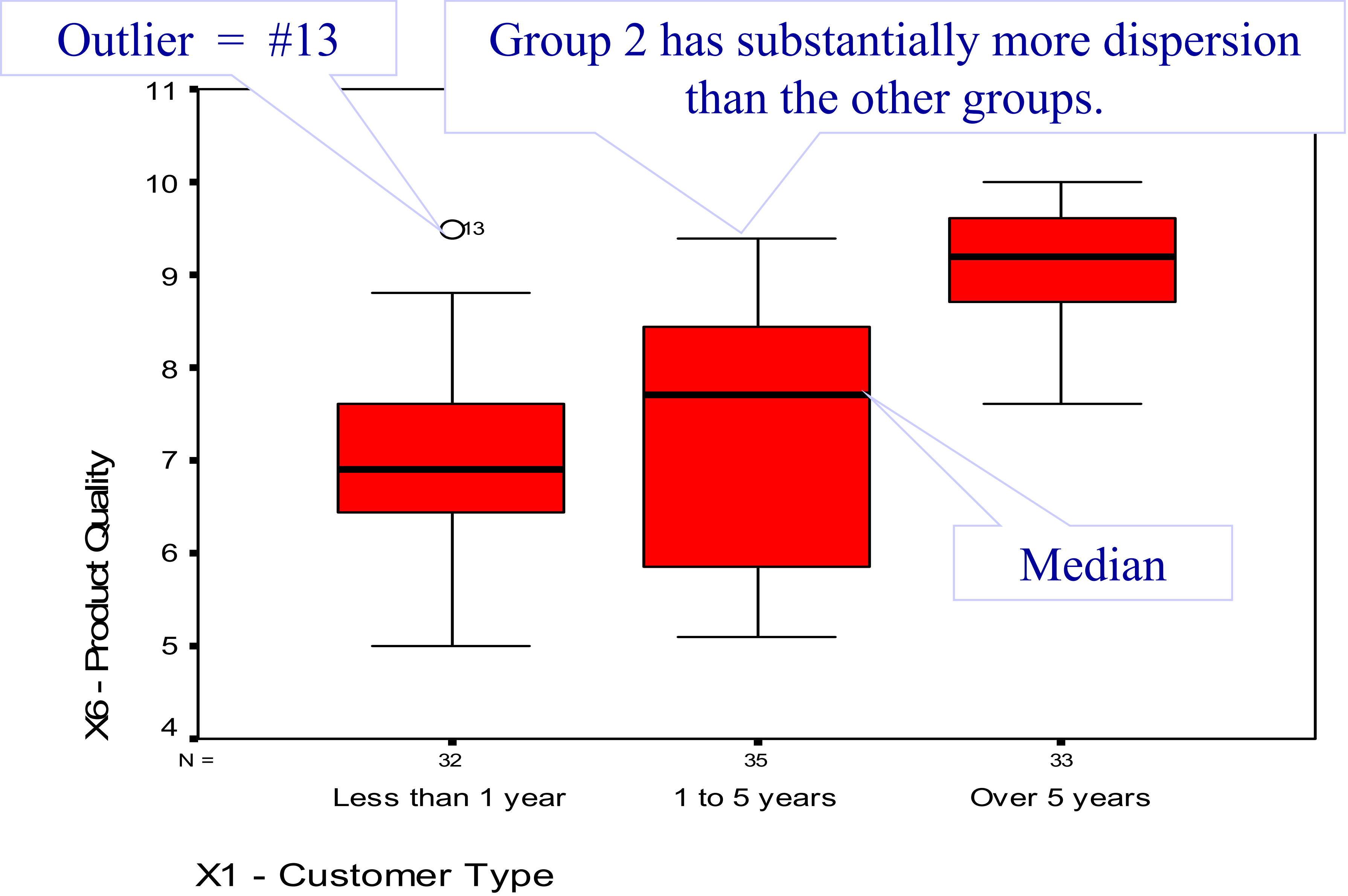
Frequency Distribution: Variable X_6 – Product Quality



X6 - Product Quality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5.0	1	1.0	1.0	1.0
	5.1	1	1.0	1.0	2.0
	5.2	1	1.0	1.0	3.0
	5.5	2	2.0	2.0	5.0
	5.6	1	1.0	1.0	6.0
	5.7	4	4.0	4.0	10.0
	5.8	1	1.0	1.0	11.0
	5.9	2	2.0	2.0	13.0
	6.0	1	1.0	1.0	14.0
	6.1	2	2.0	2.0	16.0
	6.2	1	1.0	1.0	17.0
	6.3	1	1.0	1.0	18.0
	6.4	5	5.0	5.0	23.0
	6.5	2	2.0	2.0	25.0
	6.6	1	1.0	1.0	26.0
	6.7	4	4.0	4.0	30.0
	6.9	3	3.0	3.0	33.0
	7.0	1	1.0	1.0	34.0
	7.1	2	2.0	2.0	36.0
	7.4	2	2.0	2.0	38.0
	7.5	2	2.0	2.0	40.0
	7.6	3	3.0	3.0	43.0
	7.7	3	3.0	3.0	46.0
	7.8	1	1.0	1.0	47.0
	7.9	2	2.0	2.0	49.0
	8.0	3	3.0	3.0	52.0
	8.1	1	1.0	1.0	53.0
	8.2	3	3.0	3.0	56.0
	8.3	1	1.0	1.0	57.0
	8.4	1	1.0	1.0	58.0
	8.5	4	4.0	4.0	62.0
	8.6	3	3.0	3.0	65.0
	8.7	6	6.0	6.0	71.0
	8.8	1	1.0	1.0	72.0
	9.0	2	2.0	2.0	74.0
	9.1	4	4.0	4.0	78.0
	9.2	3	3.0	3.0	81.0
	9.3	6	6.0	6.0	87.0
	9.4	3	3.0	3.0	90.0
	9.5	1	1.0	1.0	91.0
	9.6	2	2.0	2.0	93.0
	9.9	5	5.0	5.0	98.0
	Excellent	2	2.0	2.0	100.0
	Total	100	100.0	100.0	

HBAT Diagnostics: Box & Whiskers Plots

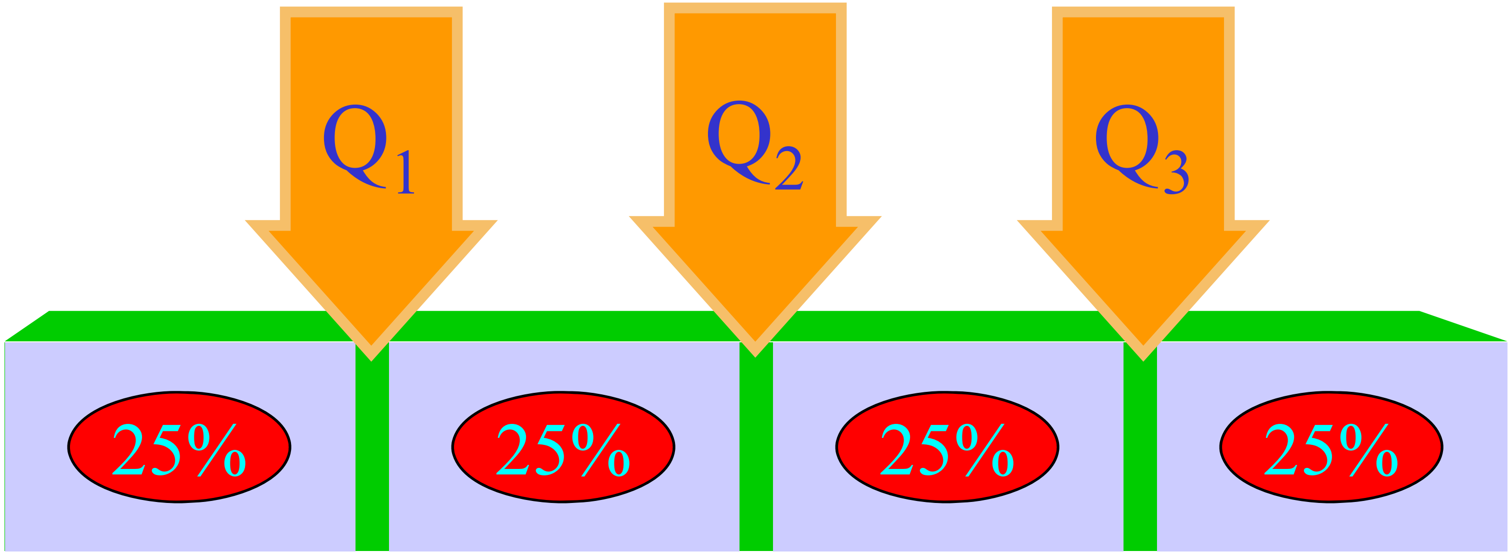


Outlier = #13

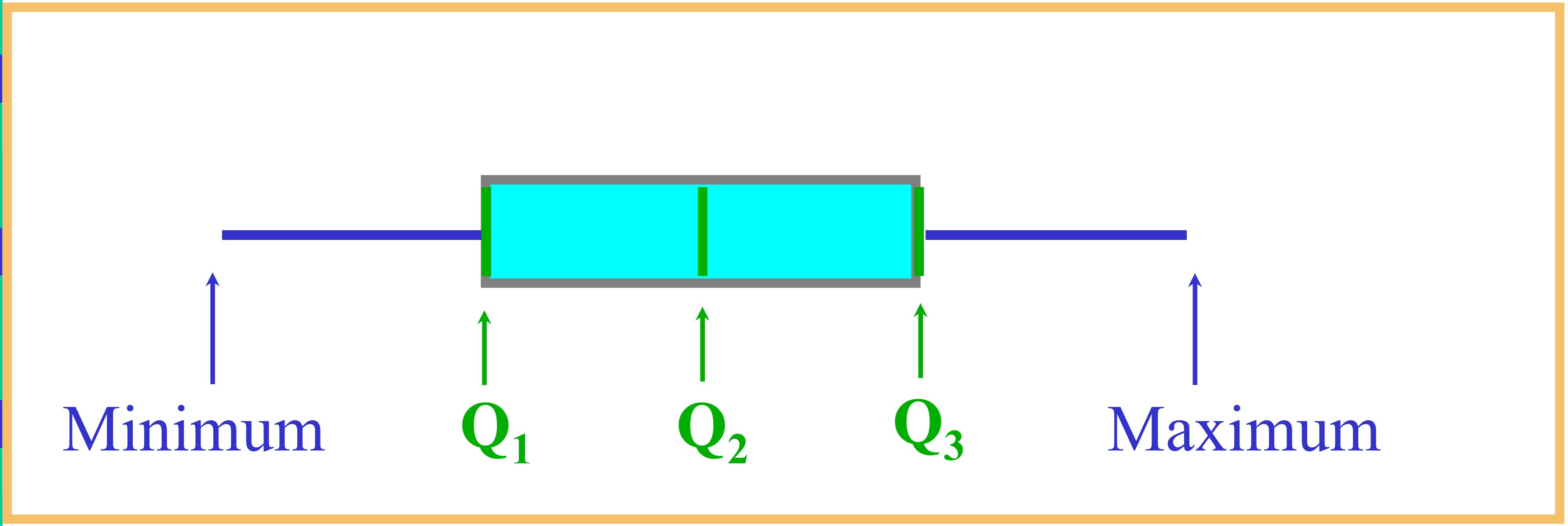
Group 2 has substantially more dispersion than the other groups.

Median

Quartiles



Box and Whisker Plot (Simple version)

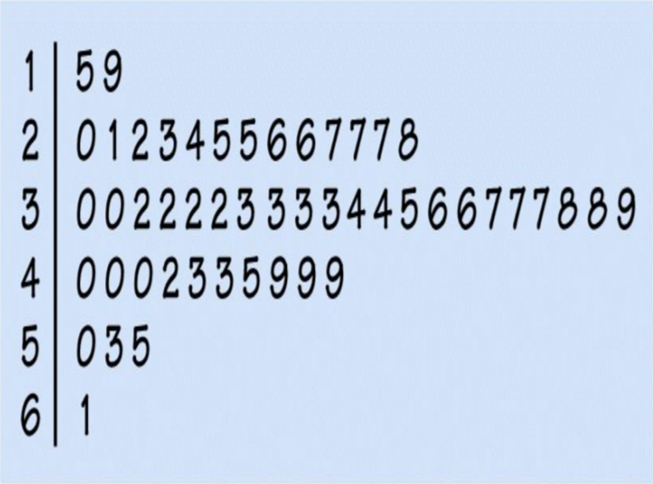


Box and Whisker Plot

- Five specific values are used:
  - Median, Q_2
  - First quartile, Q_1
  - Third quartile, Q_3
- Inner Fences (use dots)
  - $IQR = Q_3 - Q_1$
  - Lower inner fence = $Q_1 - 1.5$ IQR
  - Upper inner fence = $Q_3 + 1.5$ IQR
- Outer Fences (use stars)
  - Lower outer fence = $Q_1 - 3.0$ IQR
  - Upper outer fence = $Q_3 + 3.0$ IQR

An Example for Box and Whisker Plot

A different calculator setting gives the box-and-whisker plot with the outliers specially marked (in this case, with a simulation of an open dot), and the whiskers going only as far as the highest and lowest values that aren't outliers:



```
1 | 59
2 | 0123455667778
3 | 002222333344566777889
4 | 0002335999
5 | 035
6 | 1
```

(a)

Median, $Q_2 = 33.5$ (P $_{25}$)

First quartile, $Q_1 = 27$ (P $_{13}$)

Third quartile, $Q_3 = 40$ (P $_{38}$)

Minimum value in the data set= 15

Maximum value in the data set= 61

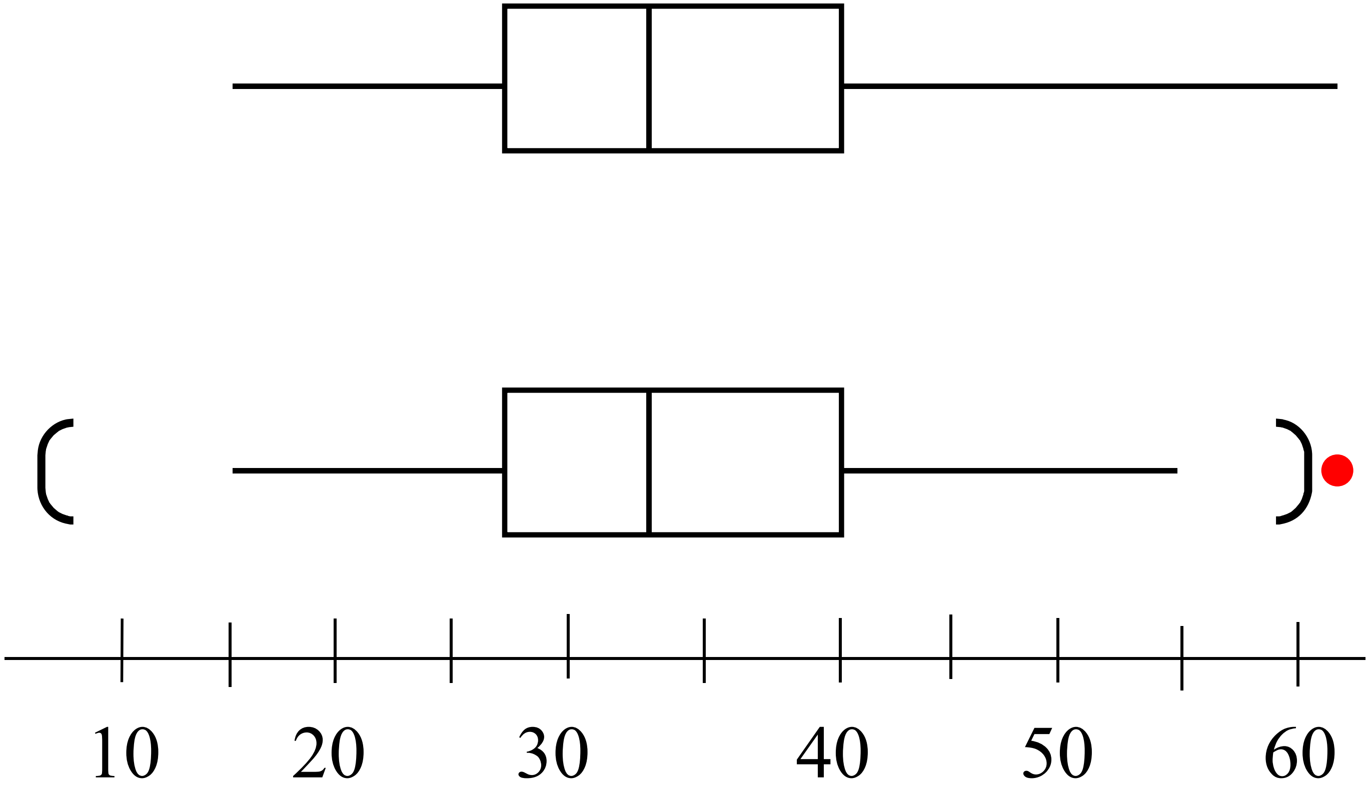
Inner Fences

IQR = Q_3 - Q_1 (= 13)

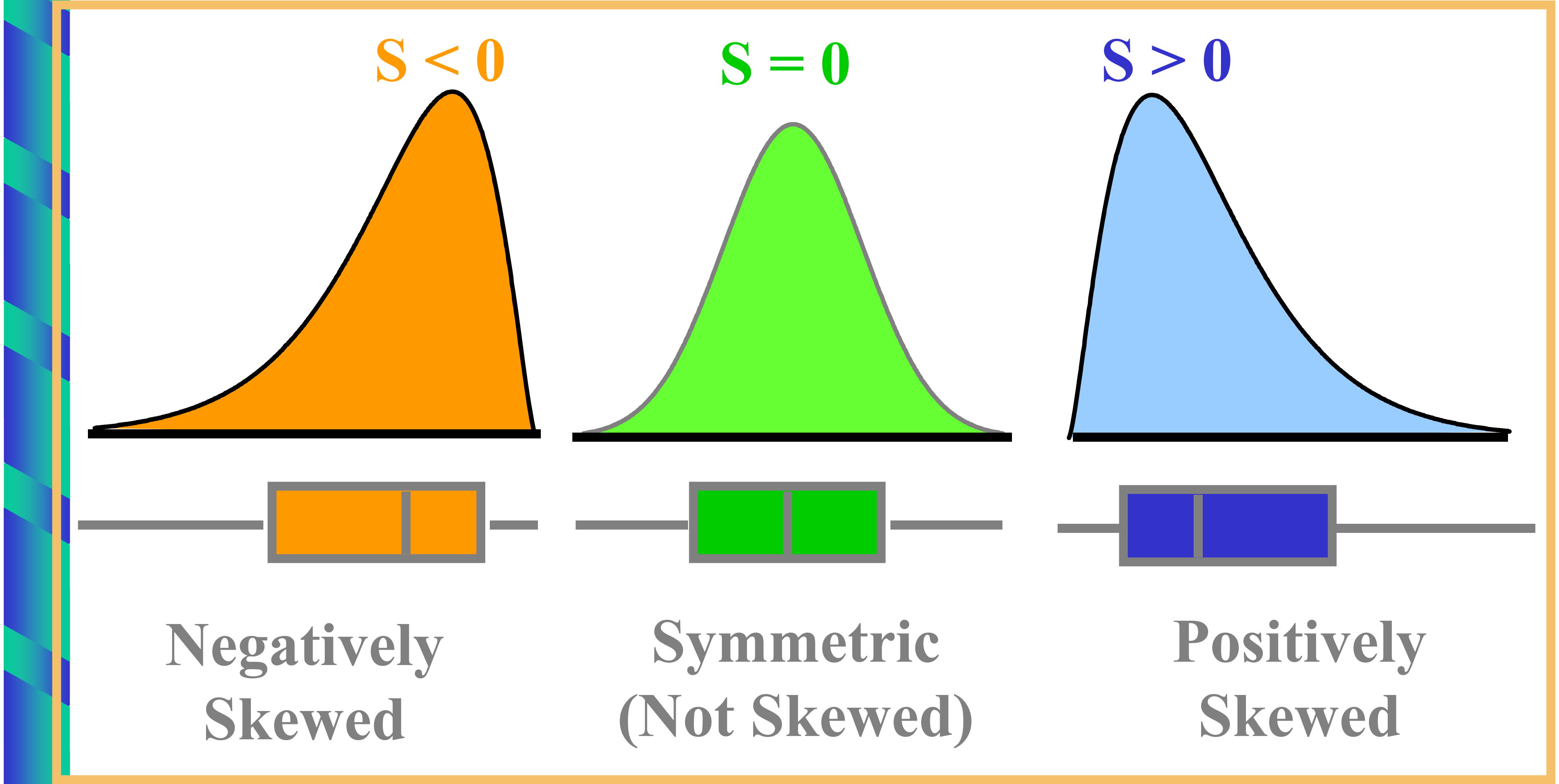
Lower inner fence = Q_1 - 1.5 IQR
= 7.5

Upper inner fence = Q_3 + 1.5 IQR
= 59.5

An Example for Box and Whisker Plot



Skewness: Box and Whisker Plots, and Coefficient of Skewness



$S < 0$

$S = 0$

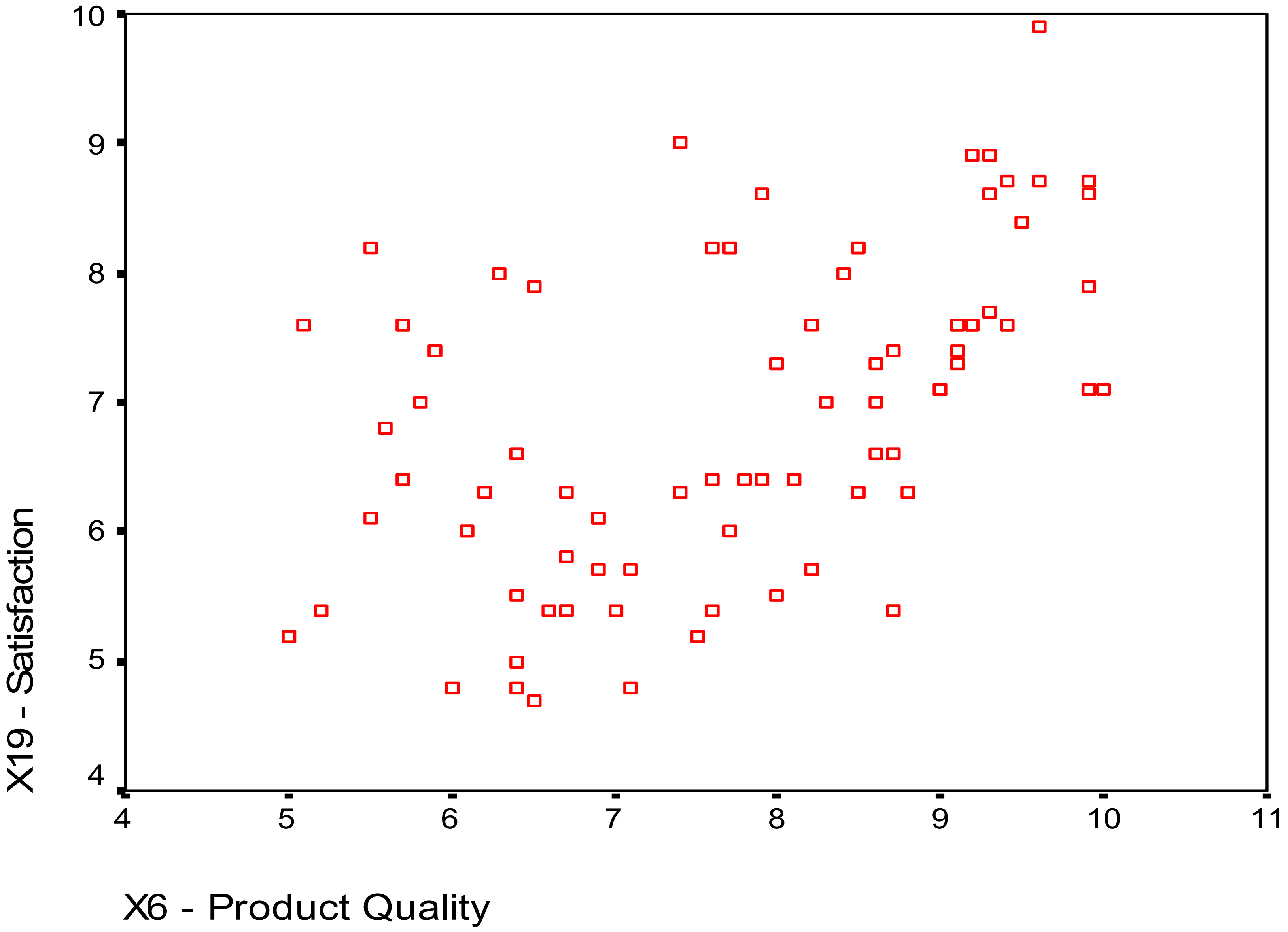
$S > 0$

Negatively Skewed

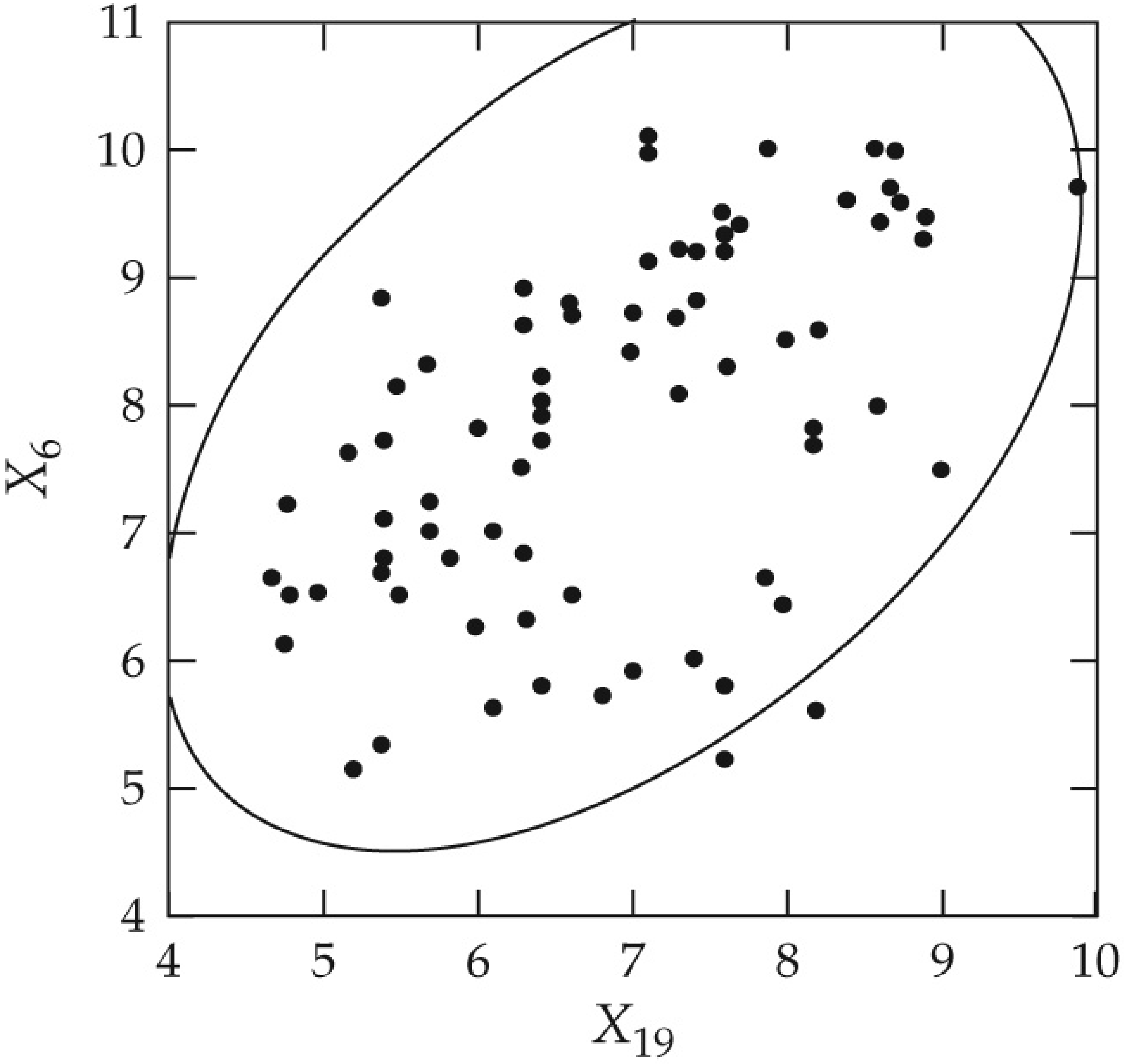
Symmetric (Not Skewed)

Positively Skewed

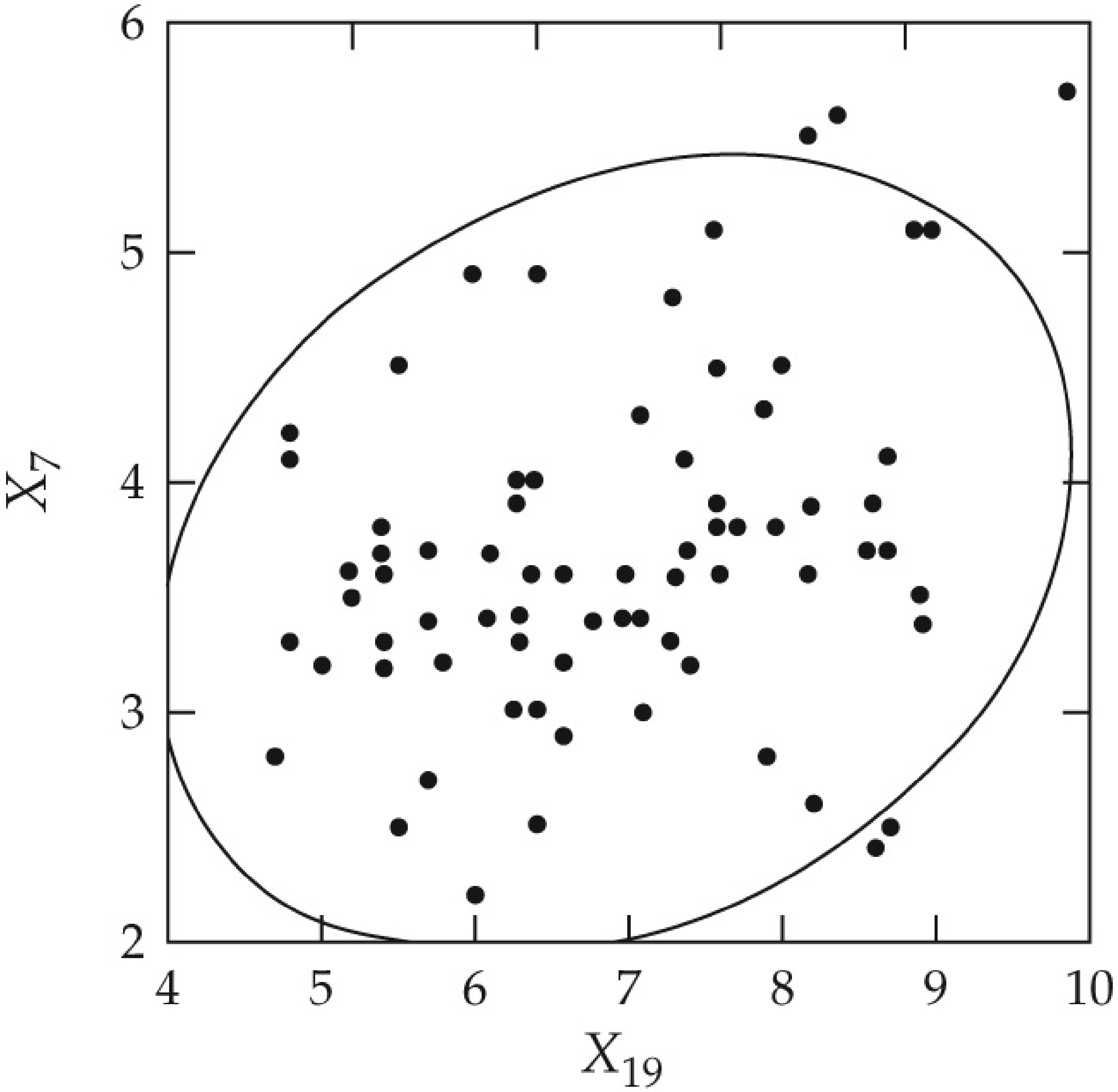
HBAT Scatterplot: Variables X_{19} and X_6



HBAT Scatterplots:



HBAT Scatterplots:



Missing Data

- Missing Data = information not available for a subject (or case) about whom other information is available. Typically occurs when respondent fails to answer one or more questions in a survey.
  - ✓ Systematic?
  - ✓ Random?
- Researcher's Concern = to identify the patterns and relationships underlying the missing data in order to maintain as close as possible to the original distribution of values when any remedy is applied.

Effects of Ignoring Missing Data

- Impact . . .
  - Reduced sample size (i.e. Reduces sample size available for analysis) - loss of statistical power
  - Data may no longer be representative (distort results) - introduces bias
  - Difficult to identify effects
- **Reasons for Missing Data**
  - ✓Refusals (question sensitivity)
  - ✓Don’t know responses (cognitive problems, memory problems)
  - ✓Not applicable
  - ✓Data processing errors
  - ✓Questionnaire programming errors
  - ✓Design factors
  - ✓Attrition in panel studies

Four-Step Process for Identifying Missing Data

Step 1: Determine the Type of Missing Data

Step 2: Determine the Extent of Missing Data

Step 3: Diagnose the Randomness of the Missing Data Processes

Step 4: Select the Imputation Method

Missing Data

Strategies for handling missing data . . .

- use observations with complete data only;
- delete case(s) and/or variable(s);
- estimate missing values.

Methods of Handling Missing Data

- **Listwise (casewise) deletion:** uses only complete cases
- **Pairwise deletion:** uses all available cases
- **Dummy variable adjustment:** missing value indicator method (constant value or a dummy variable)
- **Mean substitution:** substitute mean value computed from available cases (cf. unconditional or conditional)
- **Regression methods:** predict value based on regression equation with other variables as predictors
- **Hot deck:** identify the most similar case to the case with a missing and impute the value
- **Maximum likelihood methods:** use all available data to generate maximum likelihood-based statistics.
- **Multiple imputation:** combines the methods of ML to produce multiple data sets with imputed values for missing cases

Rules of Thumb 2–1

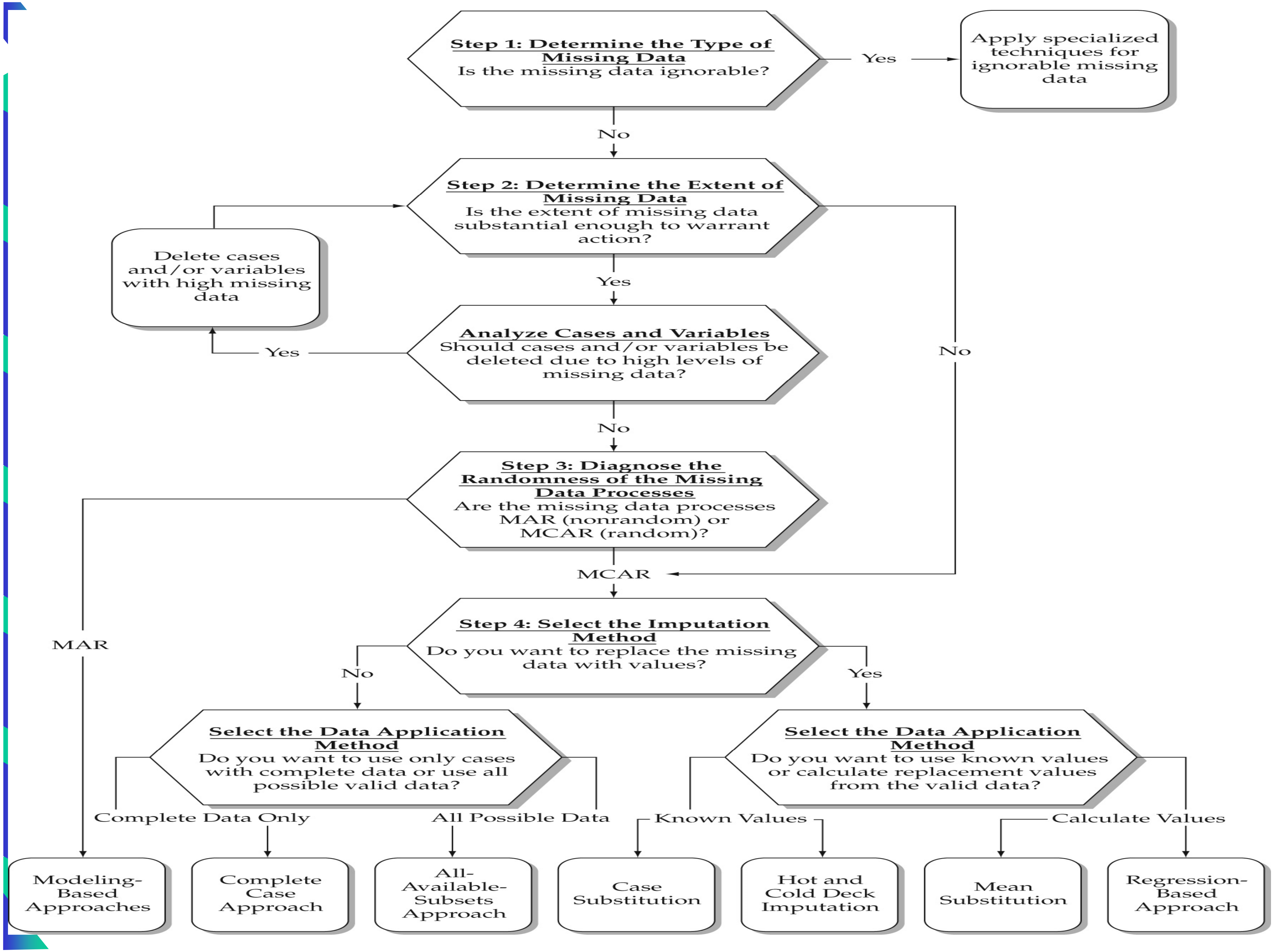
How Much Missing Data Is Too Much?

- Missing data under 10% for an individual case or observation can generally be ignored, except when the missing data occurs in a specific nonrandom fashion (e.g., concentration in a specific set of questions, attrition at the end of the questionnaire, etc.).
- The number of cases with no missing data must be sufficient for the selected analysis technique if replacement values will not be substituted (imputed) for the missing data.

Rules of Thumb 2–3

Imputation of Missing Data

- Under 10% – Any of the imputation methods can be applied when missing data is this low, although the complete case method has been shown to be the least preferred.
- 10 to 20% – The increased presence of missing data makes the all available, hot deck case substitution and regression methods most preferred for MCAR (Missing Completely At Random) data, while model-based methods are necessary with MAR (Missing At Random) missing data processes
- Over 20% – If it is necessary to impute missing data when the level is over 20%, the preferred methods are:
  - the regression method for MCAR situations, and
  - model-based methods when MAR missing data occurs.



Step 1: Determine the Type of Missing Data
Is the missing data ignorable?

Yes

Apply specialized techniques for ignorable missing data

No

Step 2: Determine the Extent of Missing Data
Is the extent of missing data substantial enough to warrant action?

Yes

Analyze Cases and Variables
Should cases and/or variables be deleted due to high levels of missing data?

Yes

Delete cases and/or variables with high missing data

No

Step 3: Diagnose the Randomness of the Missing Data Processes
Are the missing data processes MAR (nonrandom) or MCAR (random)?

MAR

MCAR

No

Step 4: Select the Imputation Method
Do you want to replace the missing data with values?

No

Yes

Select the Data Application Method
Do you want to use only cases with complete data or use all possible valid data?

Complete Data Only

All Possible Data

Select the Data Application Method
Do you want to use known values or calculate replacement values from the valid data?

Known Values

Calculate Values

Modeling-Based Approaches

Complete Case Approach

All-Available-Subsets Approach

Case Substitution

Hot and Cold Deck Imputation

Mean Substitution

Regression-Based Approach

Outlier

Outlier = an observation/response with a unique combination of characteristics identifiable as distinctly different from the other observations/responses.

Issue: "Is the observation/response representative of the population?"

Why Do Outliers Occur?

- Procedural Error.
- Extraordinary Event.
- Extraordinary Observations.
- Observations unique in their combination of values.

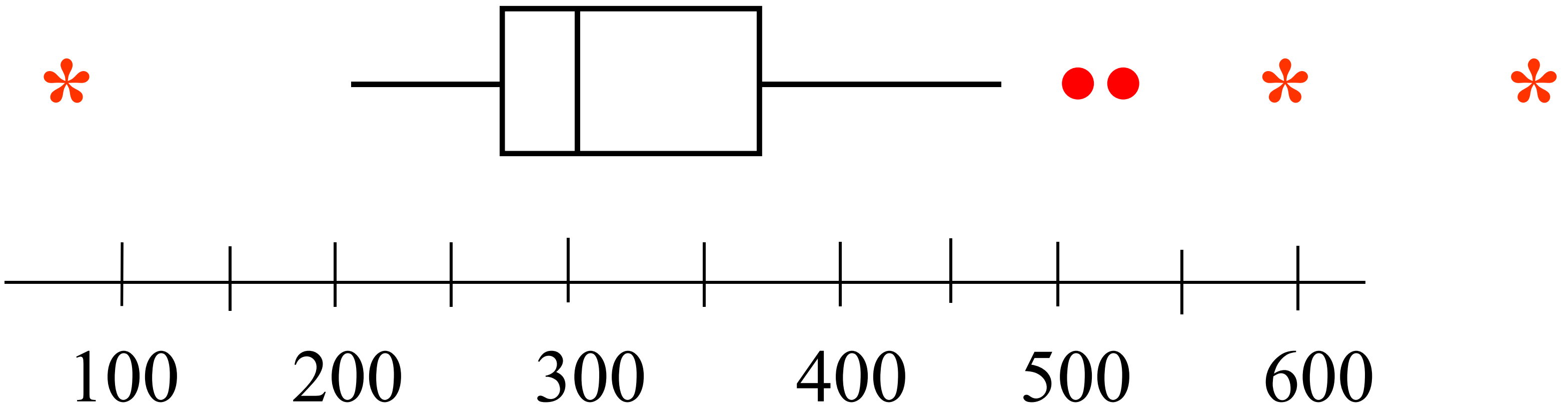
Dealing with Outliers

- Identify outliers.
- Describe outliers.
- Delete or Retain?

Identifying Outliers

- Standardize data and then identify outliers in terms of number of standard deviations.
- Examine data using Box Plots, Stem & Leaf, and Scatterplots.
- Multivariate detection (D^2).

An Example of Outliers



Rules of Thumb 2–4

Outlier Detection

- Univariate methods – examine all metric variables to identify unique or extreme observations.
- For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.
- For larger sample sizes, increase the threshold value of standard scores up to 4.
- If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.

Rules of Thumb 2–4

Outlier Detection

- Bivariate methods – focus their use on specific variable relationships, such as the independent versus dependent variables:
  - use scatterplots with confidence intervals at a specified Alpha level.
- Multivariate methods – best suited for examining a complete variate, such as the independent variables in regression or the variables in factor analysis:
  - threshold levels for the D2/df measure should be very conservative (.005 or .001), resulting in values of 2.5 (small samples) versus 3 or 4 in larger samples.

Multivariate Assumptions

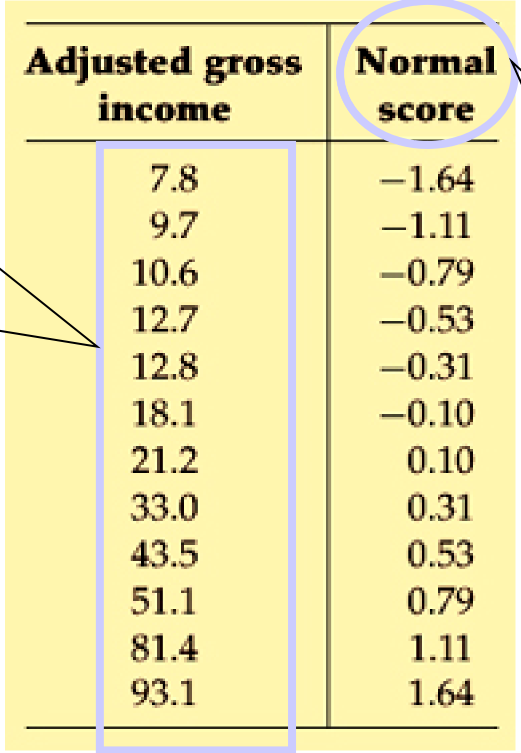
- Normality
- Linearity
- Homoscedasticity
- Non-correlated Errors
  - ✓ Data Transformations?

Testing Assumptions

- Normality assumptions
  - Visual check of histogram.
  - Kurtosis.
  - Normal probability plot.
- Homoscedasticity
  - Equal variances across independent variables.
  - Levene test (univariate).
  - Box’s M (multivariate).

Normal probability plot

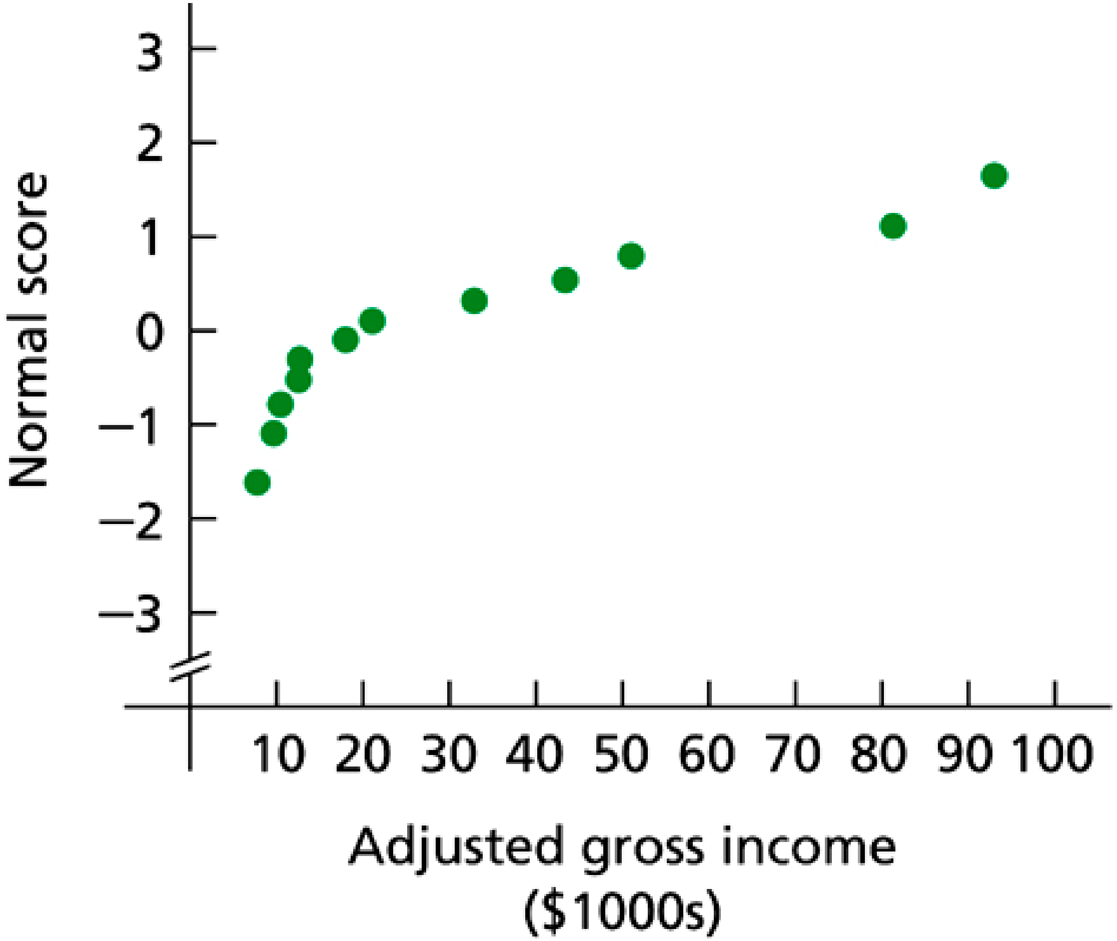
First arrange the data in increasing order



Adjusted gross income	Normal score
7.8	−1.64
9.7	−1.11
10.6	−0.79
12.7	−0.53
12.8	−0.31
18.1	−0.10
21.2	0.10
33.0	0.31
43.5	0.53
51.1	0.79
81.4	1.11
93.1	1.64

We then obtain the normal scores from tables or derived from calculation

Normal probability plot



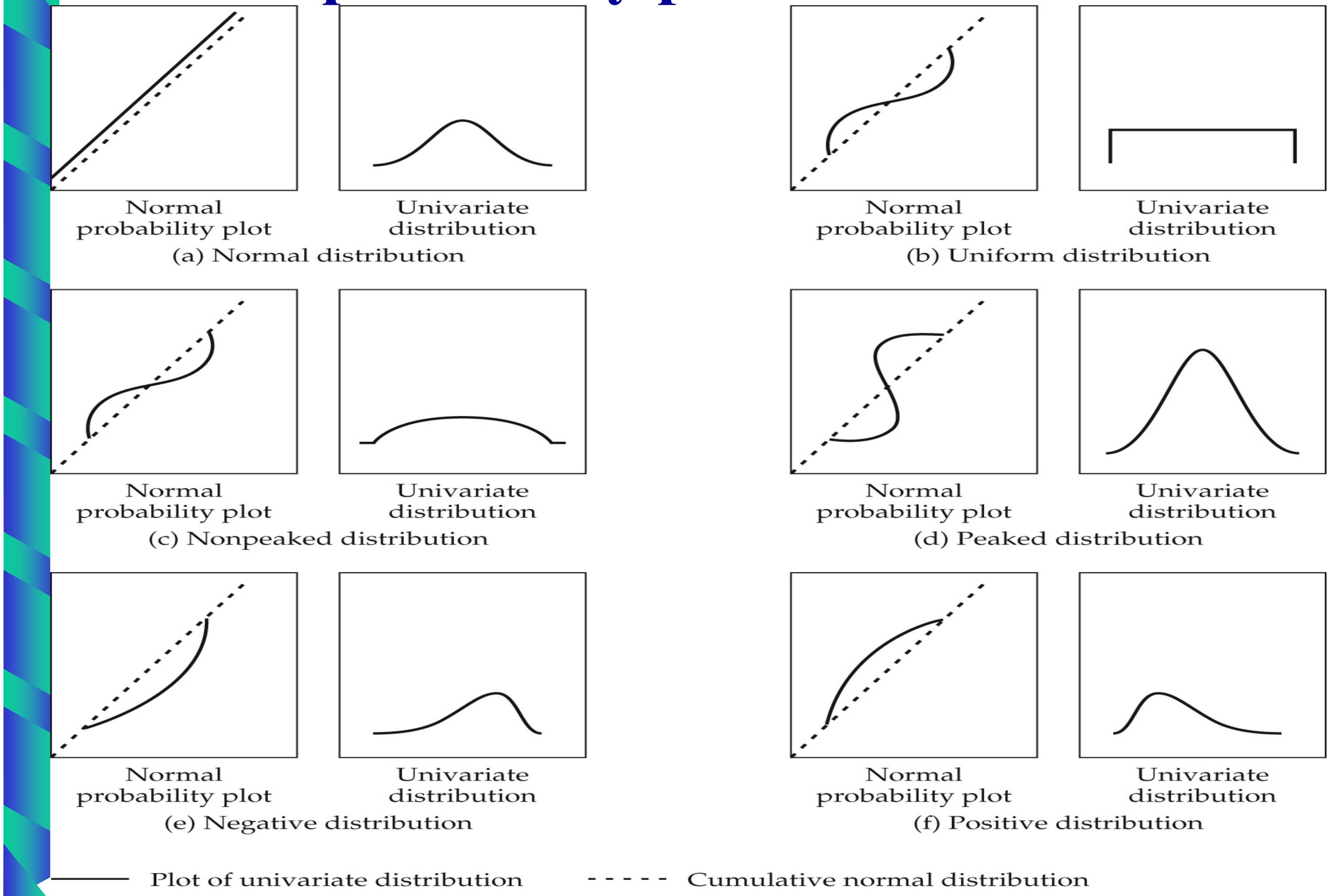
Guidelines for assessing Normality Using a Normal Probability Plot

To assess the normality of a variable using sample data, construct a normal probability plot.

- If the plot is roughly linear, you can assume that the variable is approximately normally distributed.
- If the plot is not roughly linear, you can assume that the variable is not approximately normally distributed.

There guidelines should be interpreted loosely for small samples, but usually strictly for large samples.

Normal probability plot



(a) Normal distribution

(b) Uniform distribution

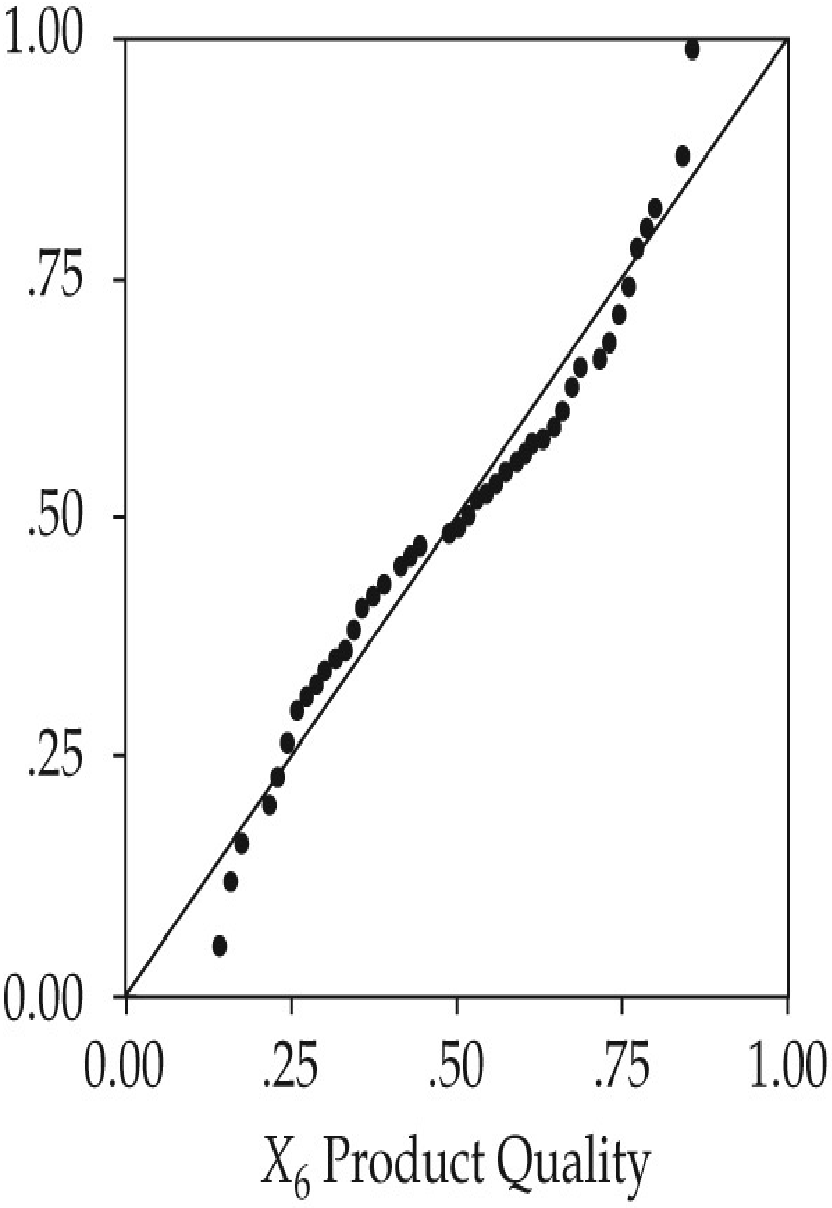
(c) Nonpeaked distribution

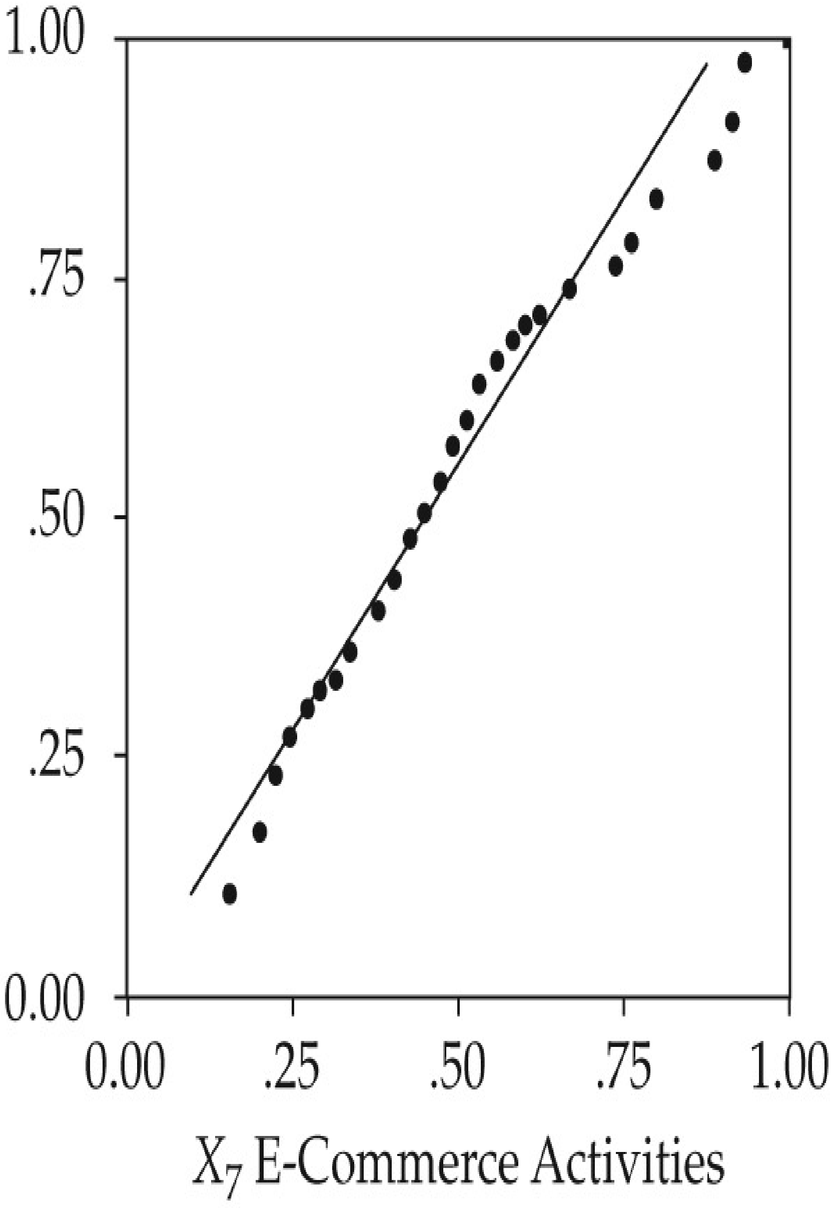
(d) Peaked distribution

(e) Negative distribution

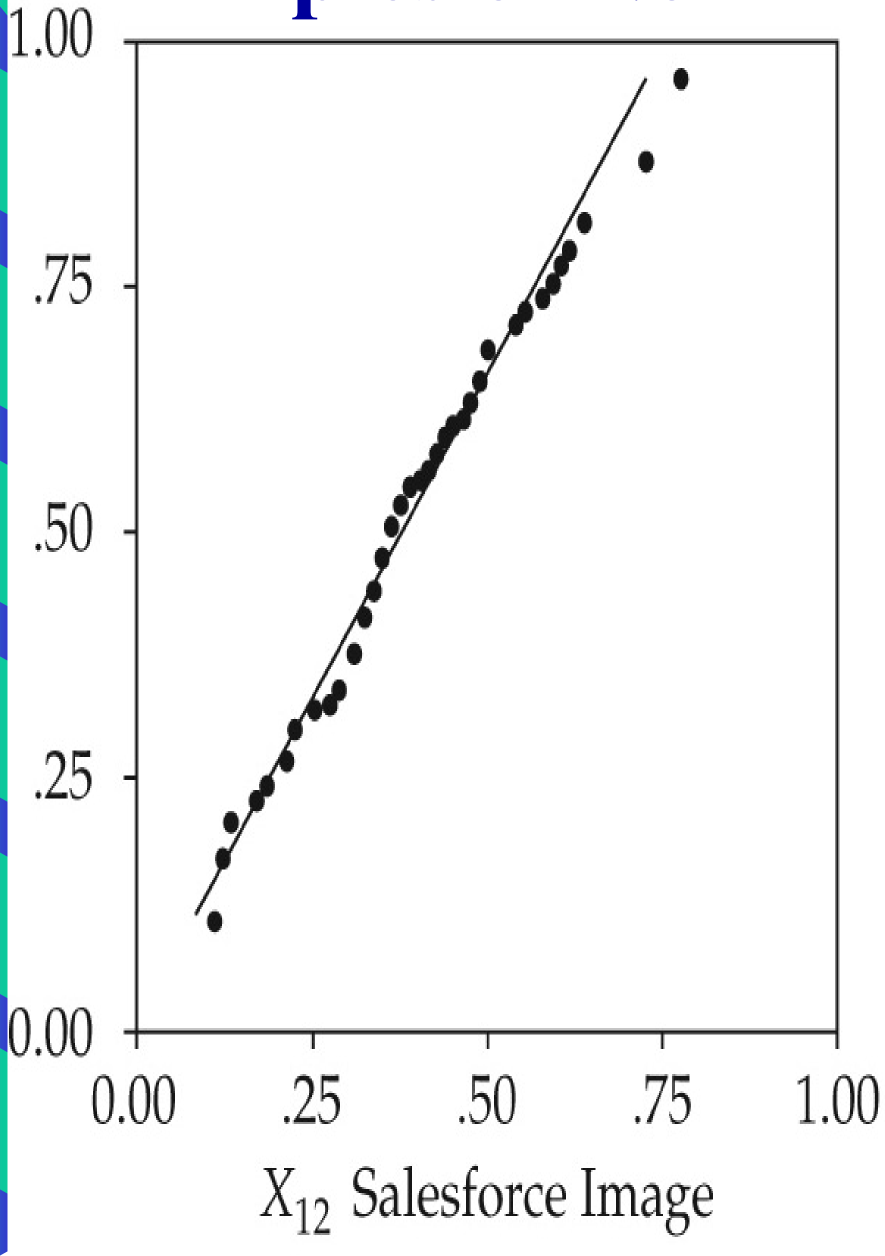
(f) Positive distribution

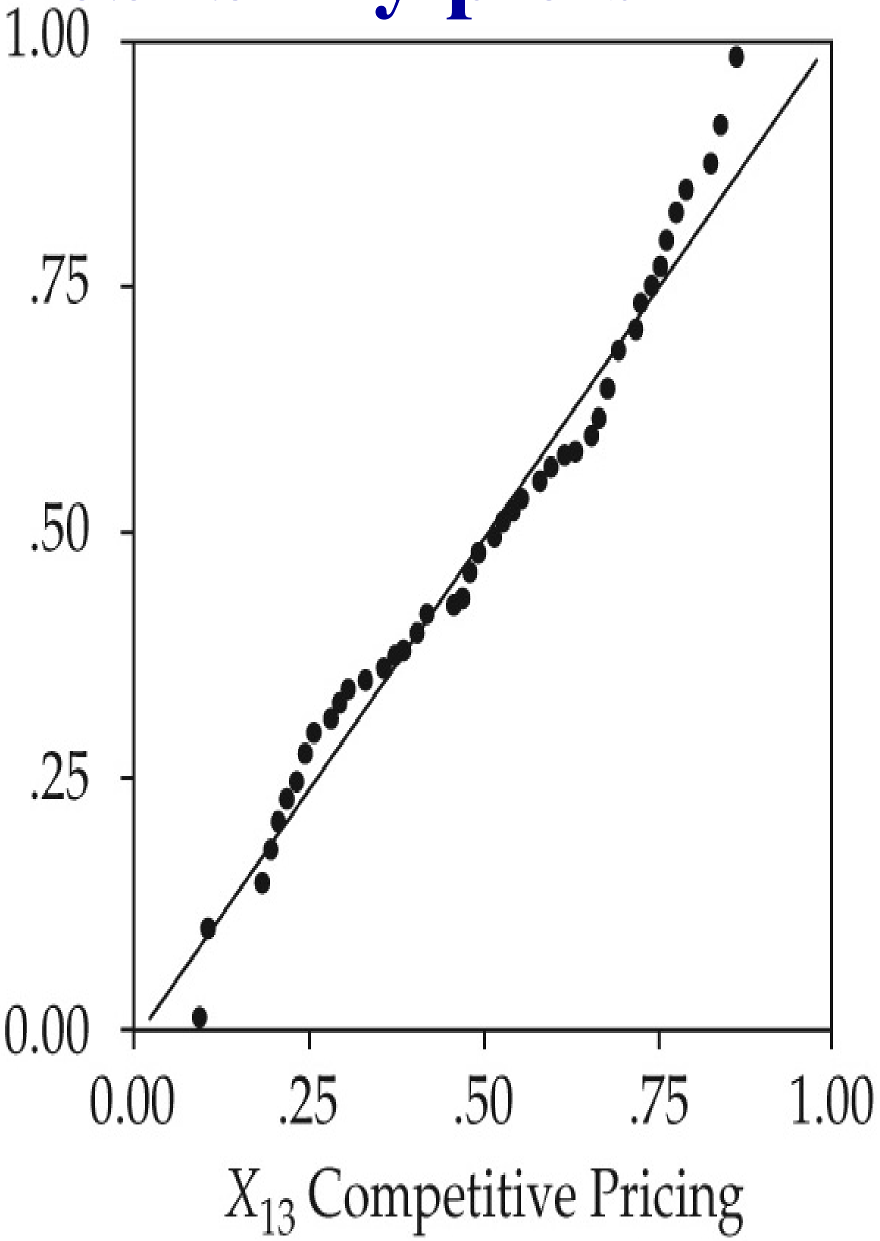
Examples of Normal probability plots



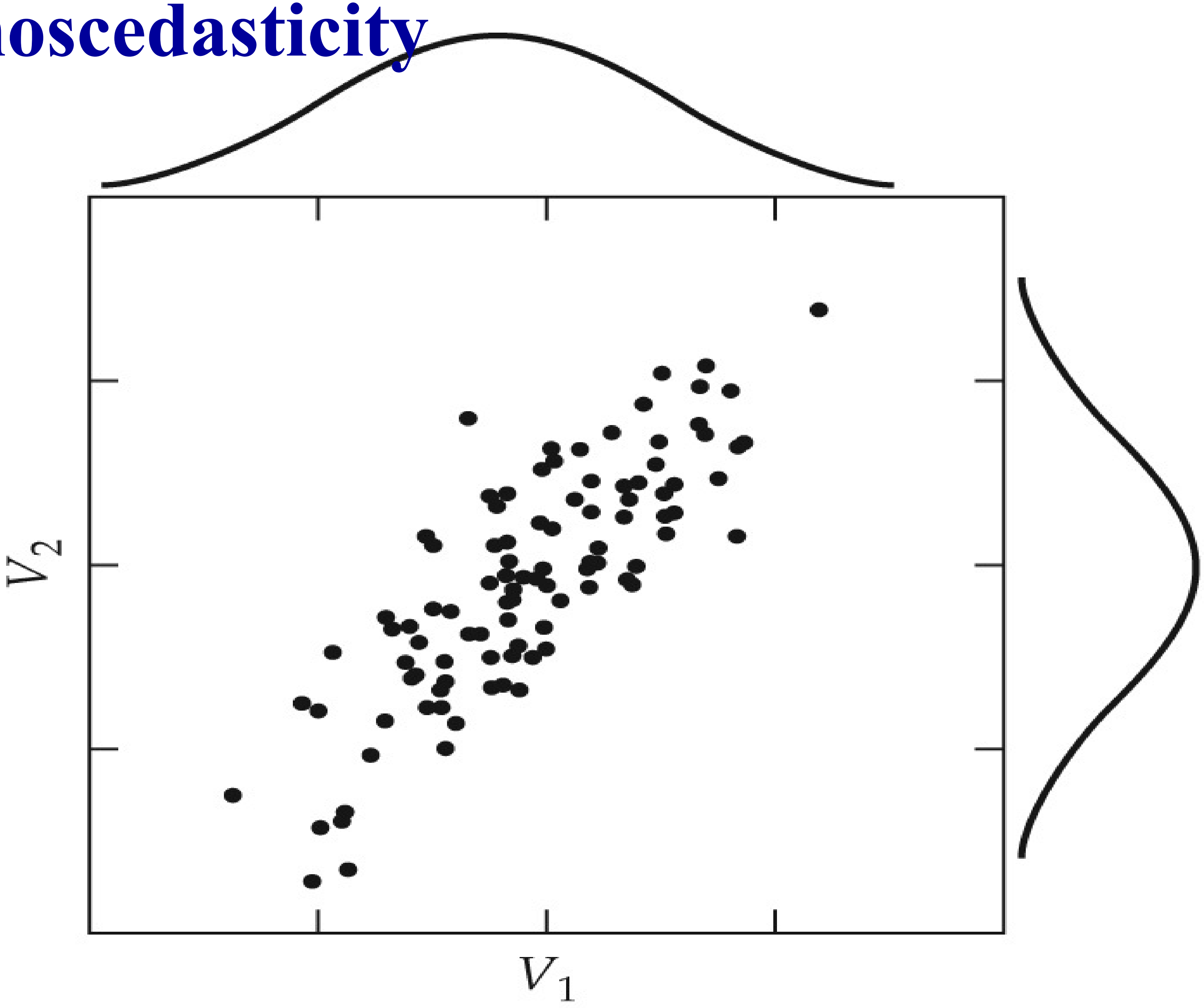


Examples of Normal probability plots



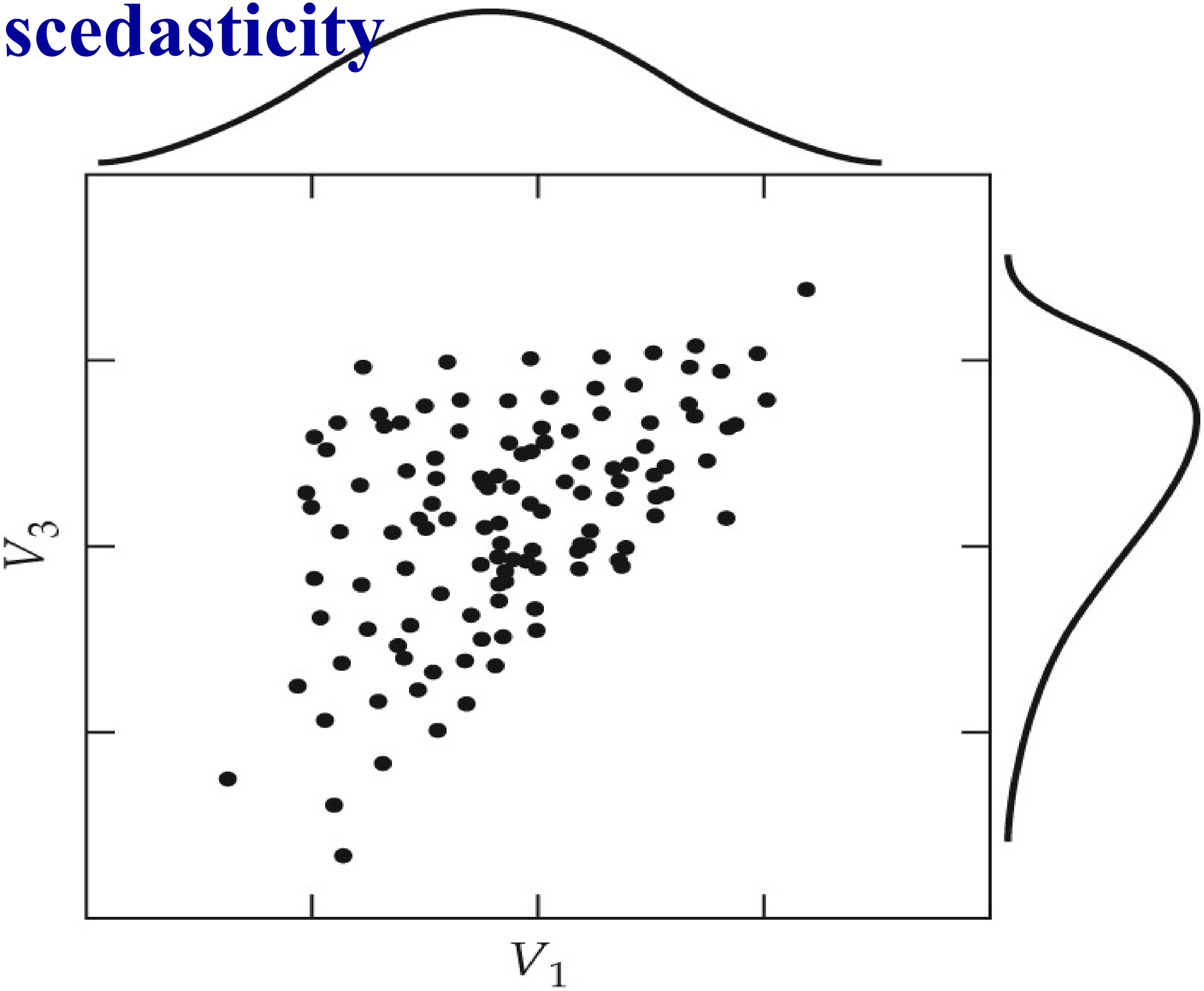


Homoscedasticity



(a) Homoscedasticity

Heteroscedasticity

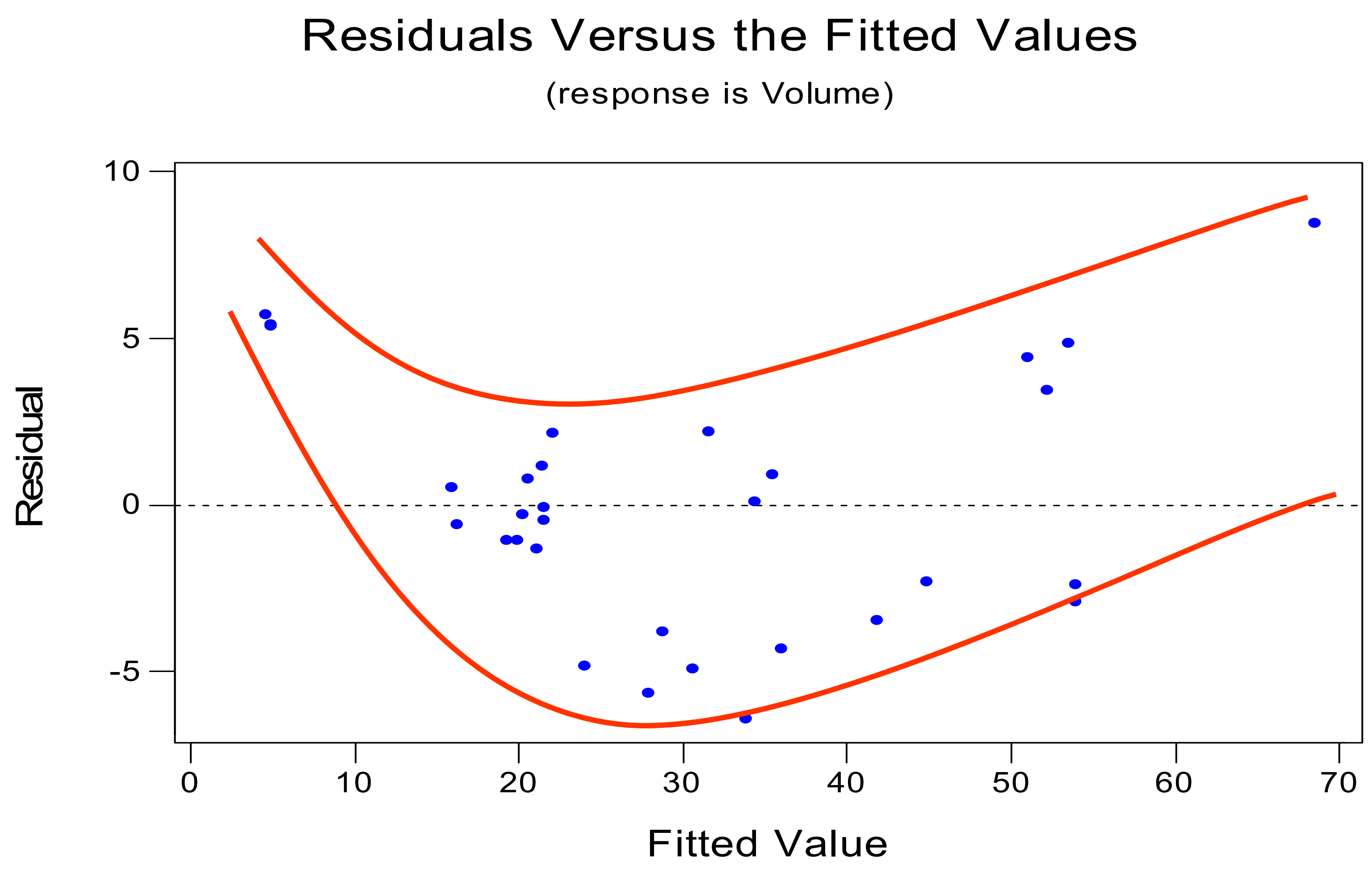


(b) Heteroscedasticity

Rules of Thumb 2–5

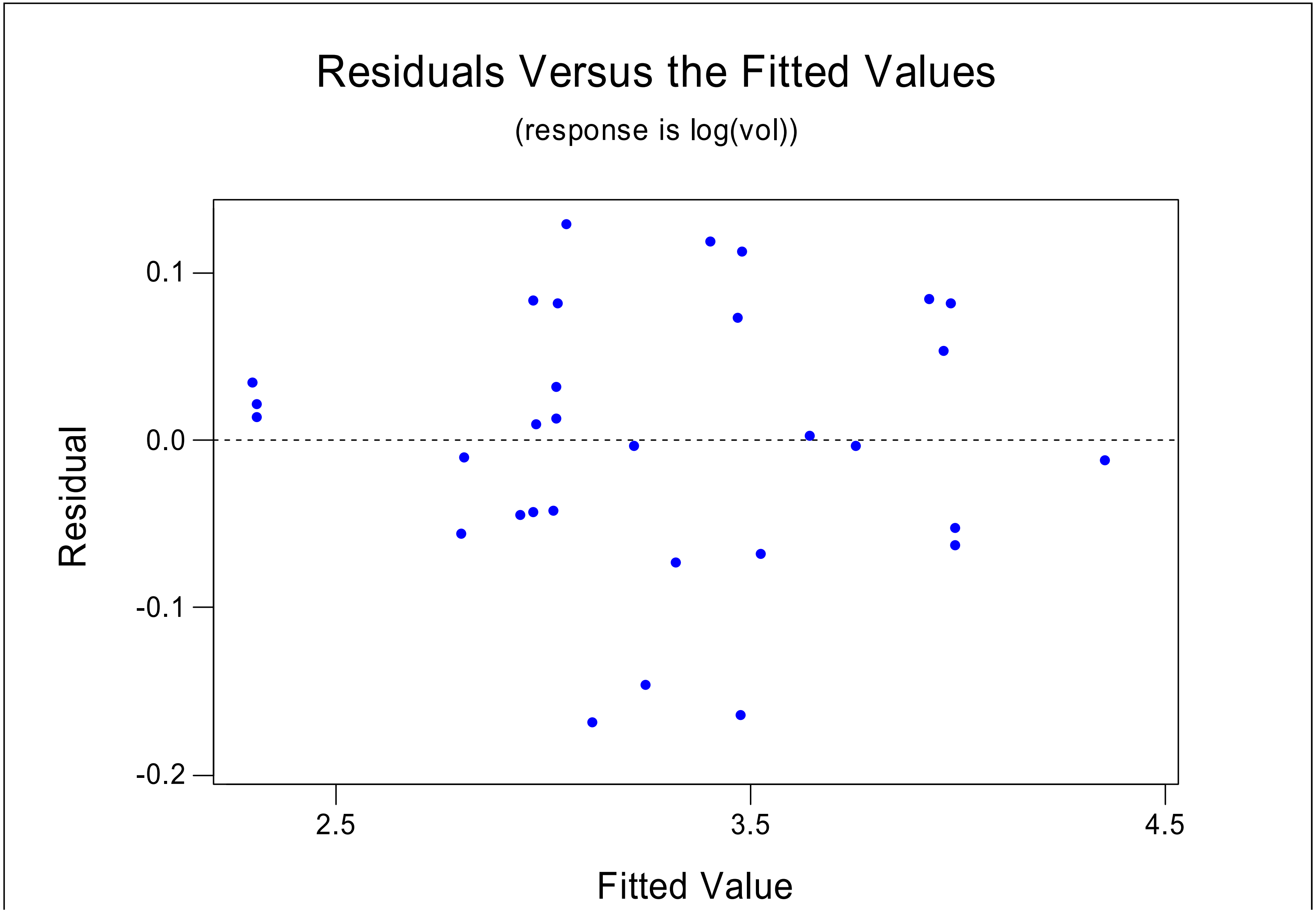
Testing Statistical Assumptions

- Normality can have serious effects in small samples (less than 50cases), but the impact effectively diminishes when sample sizes reach 200 cases or more.
- Most cases of heteroscedasticity are a result of non-normality in one or more variables. Thus, remedying normality may not be needed due to sample size, but may be needed to equalize the variance.



The pred-res plot tells us something is definitely out of whack; the "Nike Swoosh" shape tells us

1. that our variance assumption does not hold
2. There is some non-linear trend left in the residuals



The predicted vs. residual plot shows none of the banding we saw before – everything looks okay

Rules of Thumb 2–5

Testing Statistical Assumptions

- Nonlinear relationships can be very well defined, but seriously understated unless the data is transformed to a linear pattern or explicit model components are used to represent the nonlinear portion of the relationship.

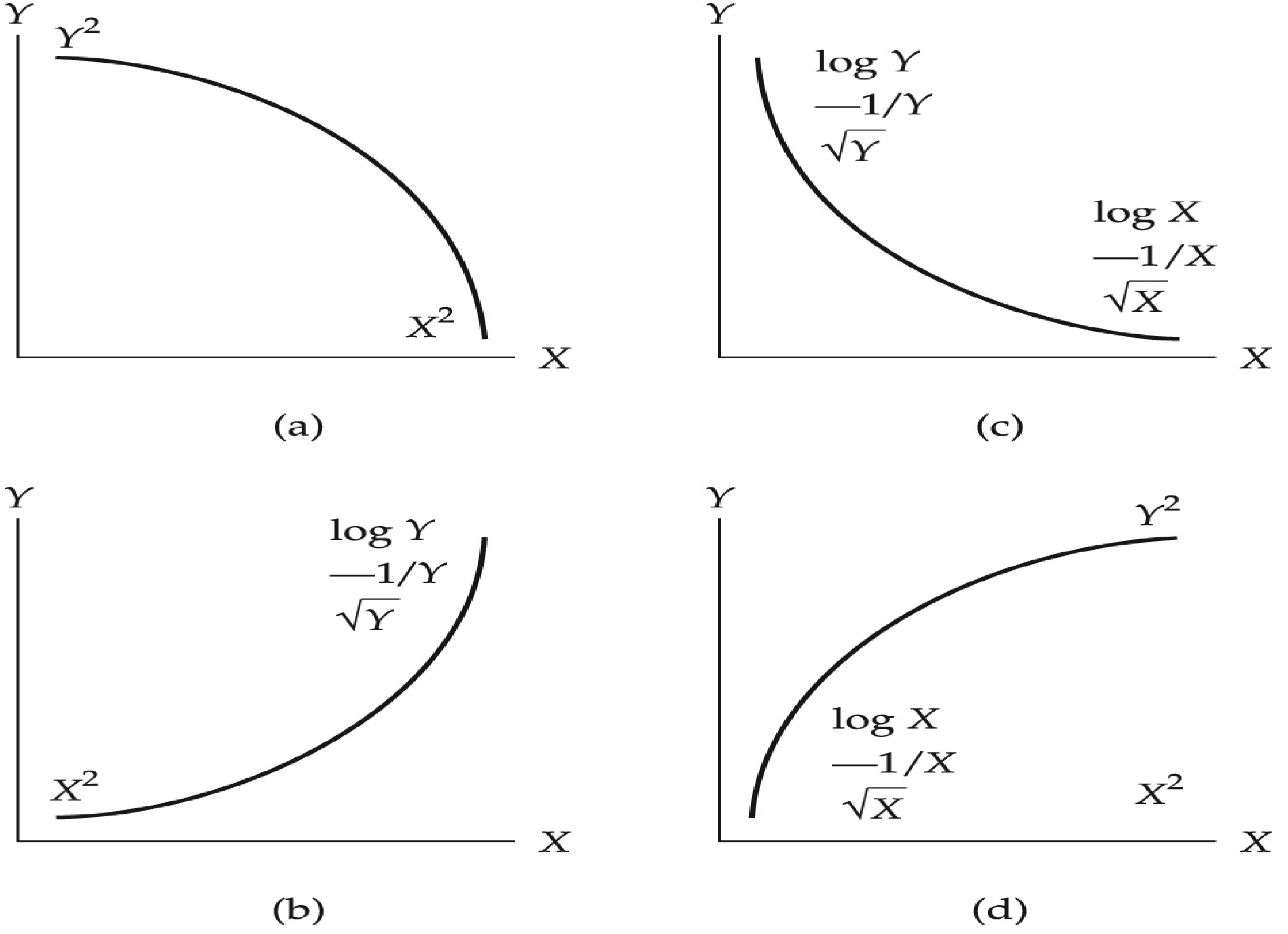
- Correlated errors arise from a process that must be treated much like missing data. That is, the researcher must first define the "causes" among variables either internal or external to the dataset. If they are not found and remedied, serious biases can occur in the results, many times unknown to the researcher.

Data Transformations ?

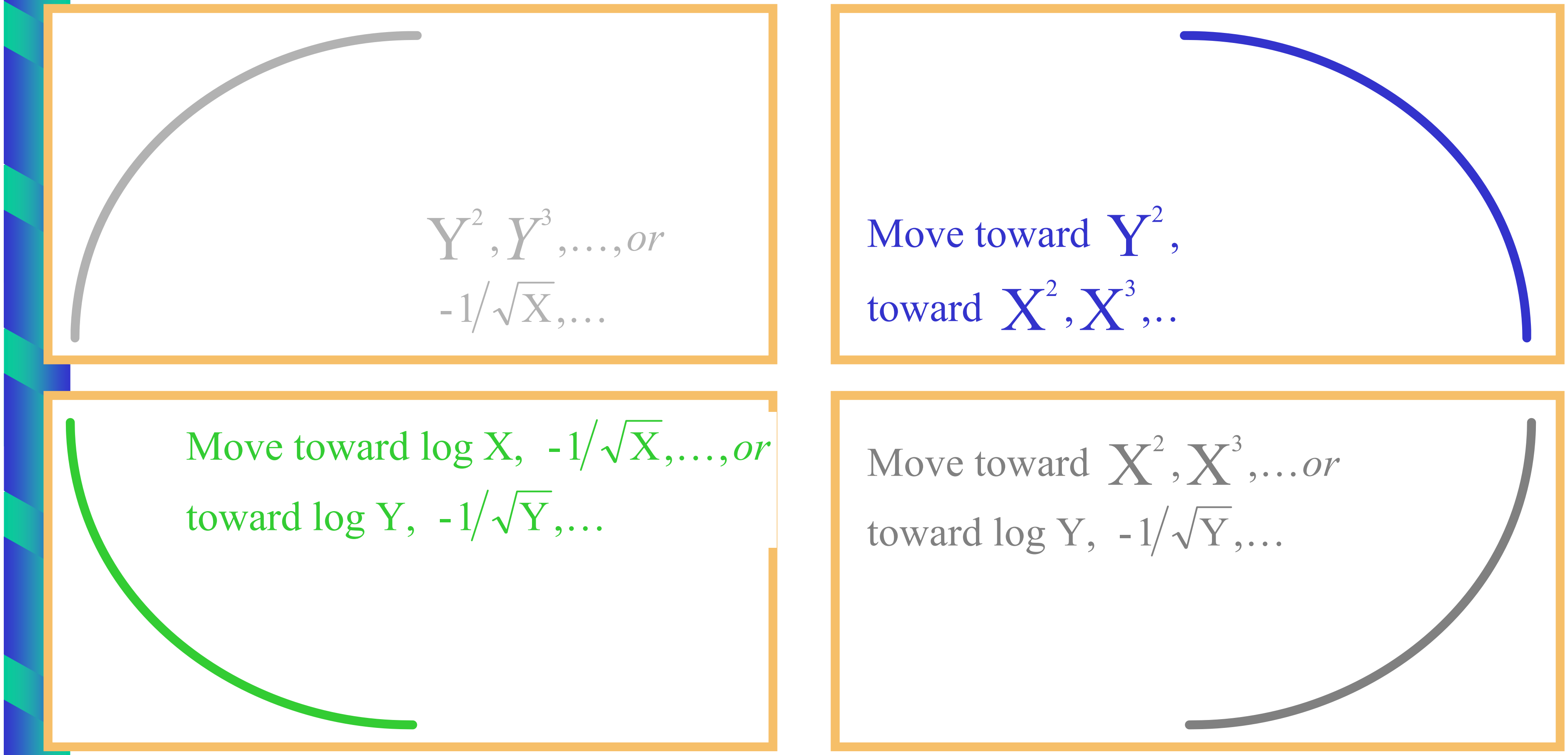
Data transformations . . . provide a means of modifying variables for one of two reasons:

1. To correct violations of the statistical assumptions underlying the multivariate techniques, or
2. To improve the relationship (correlation) between the variables.

How to Do Data Transformations ?



Tukey’s Four Quadrant Approach



$Y^2, Y^3, \ldots, or$

$-1/\sqrt{X}, \ldots$

Move toward Y^2,

toward $X^2, X^3, \ldots$

Move toward log X, $-1/\sqrt{X}, \ldots, or$

toward log Y, $-1/\sqrt{Y}, \ldots$

Move toward $X^2, X^3, \ldots or$

toward log Y, $-1/\sqrt{Y}, \ldots$

Models or Prediction Equations

- Some examples of various possible relationships:

Linear: $\hat{y} = b_0 + b_1 x$

Quadratic: $\hat{y} = (a + bx)^2$

Exponential: $\hat{y} = a(b^x)$

Logarithmic: $\hat{y} = a \log_b x$

Reciprocal: $\hat{y} = \dfrac{1}{a + bx}$

Note: What would a scatter diagram look like to suggest each relationship?

Nonlinear Regression Models: Model Transformation

$$\hat{y} = ab^{x}$$

$$\Rightarrow \log(\hat{y}) = \log(a) + x\log(b)$$

$$\Rightarrow \hat{y}' = a' + b'x$$

where:

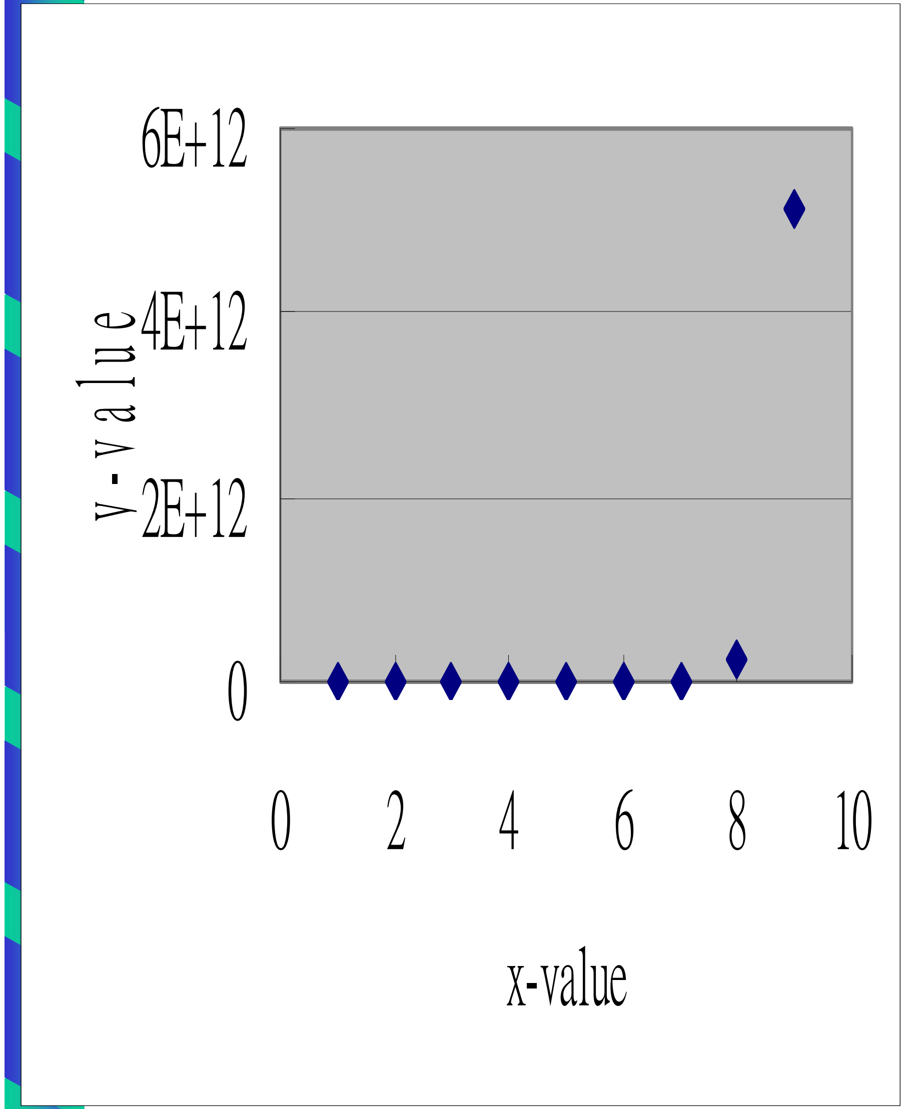
$$\hat{y}' = \log(\hat{y})$$

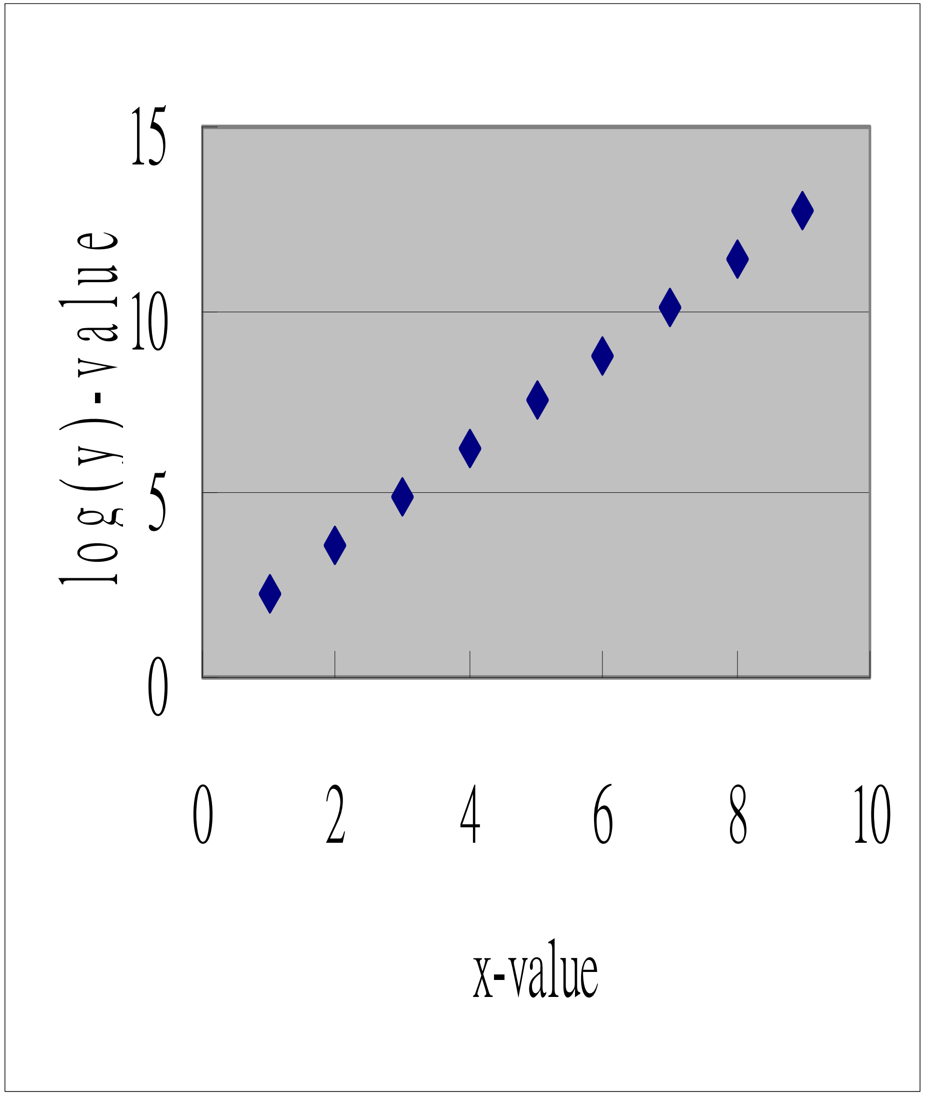
$$a' = \log(a)$$

$$b' = \log(b)$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = (a + bx)^2$$

$$\Rightarrow \sqrt{\hat{y}} = a + bx$$

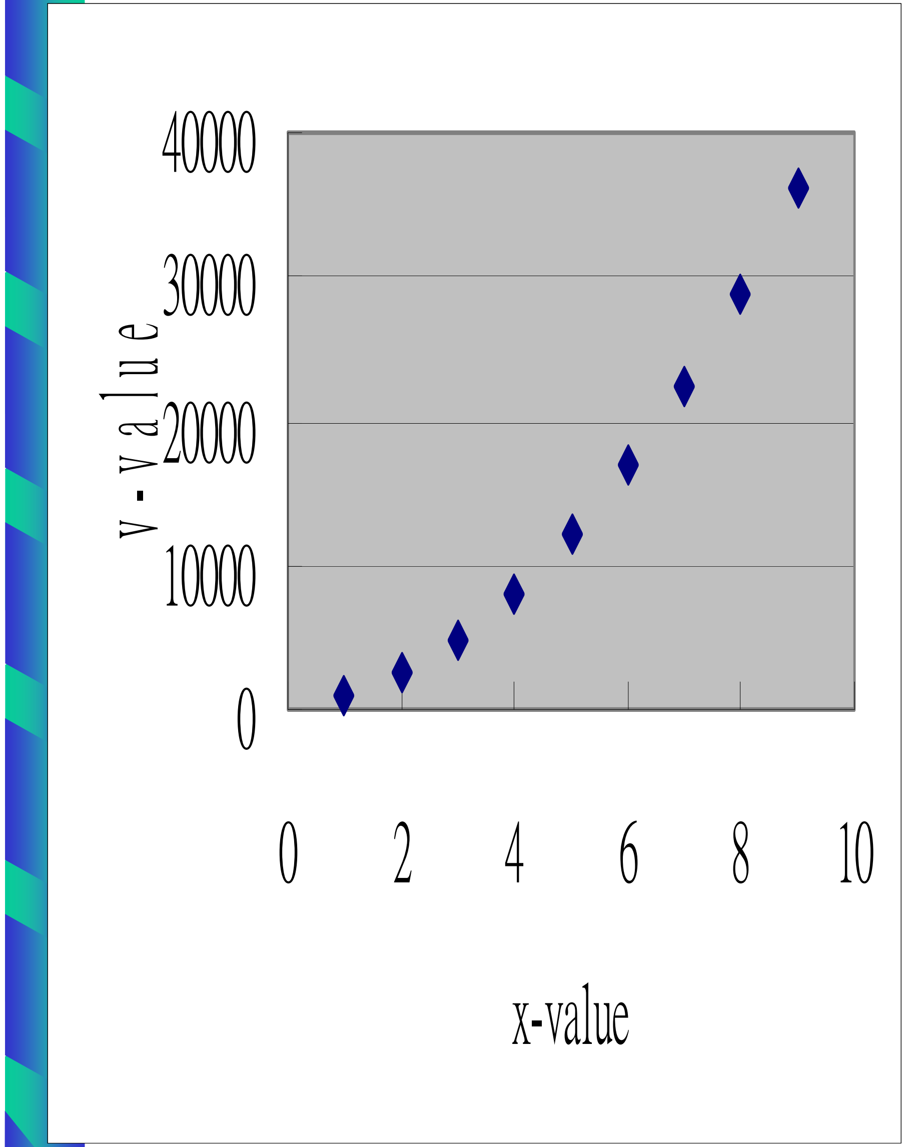
$$\Rightarrow \hat{y}' = a + bx$$

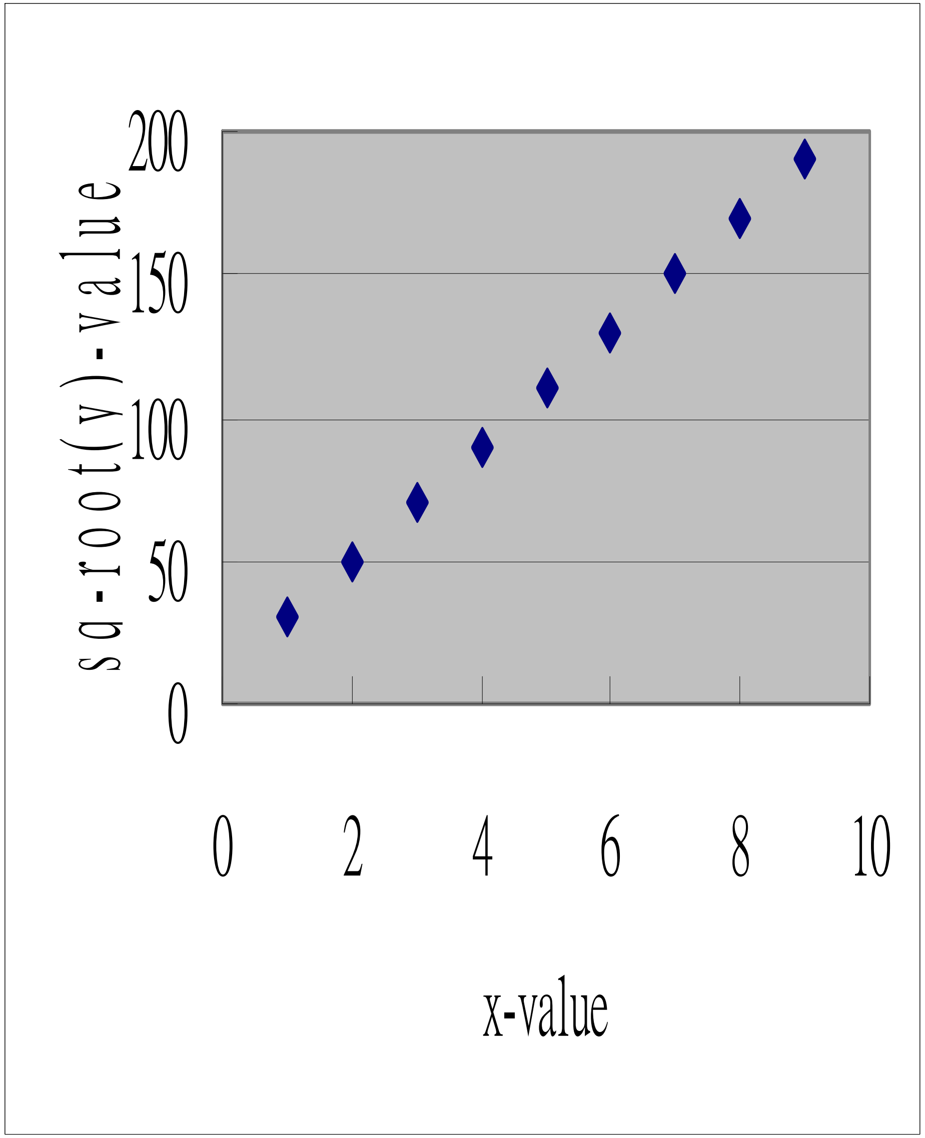
where:

$$\hat{y}' = \sqrt{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

Corresponding Scatter Plot





Nonlinear Regression Models: Model Transformation

$$\hat{y} = \frac{1}{a + bx}$$

$$\Rightarrow \frac{1}{\hat{y}} = a + bx$$

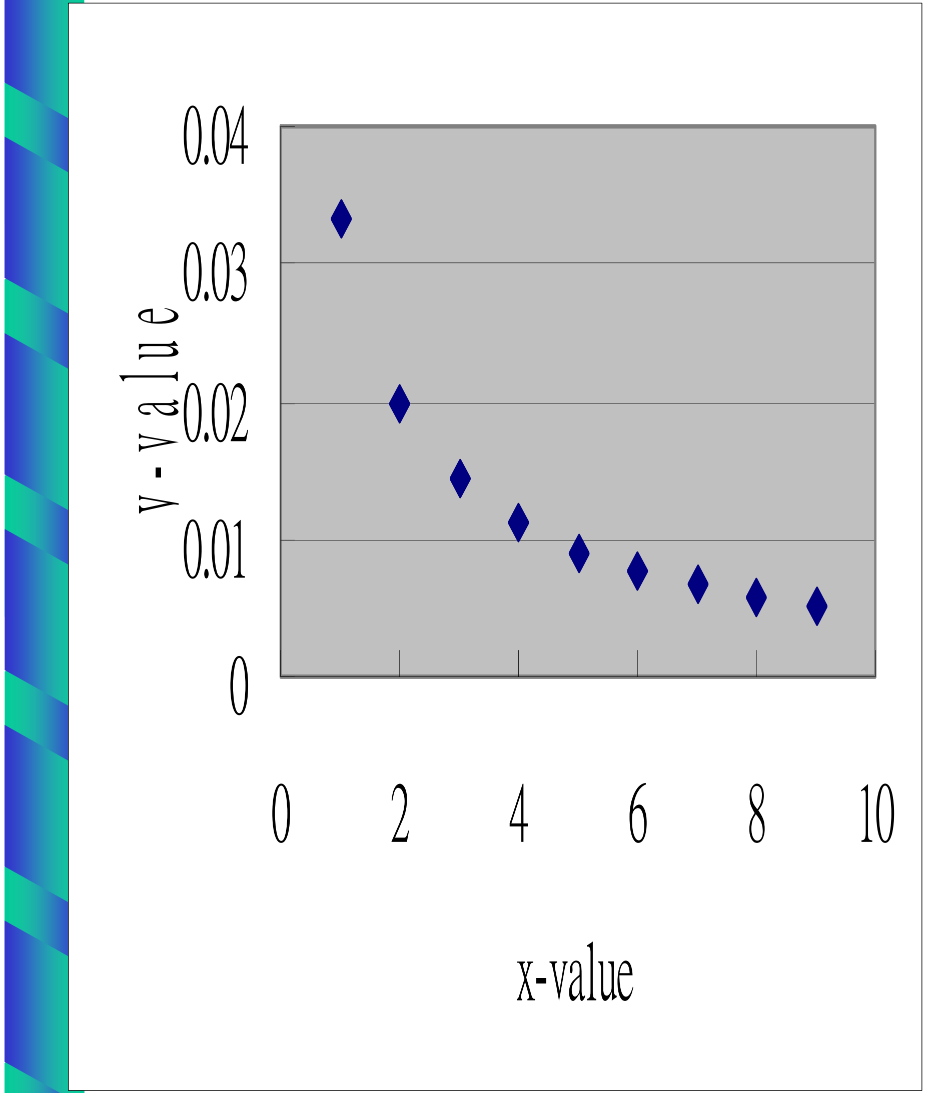
$$\Rightarrow \hat{y}' = a + bx$$

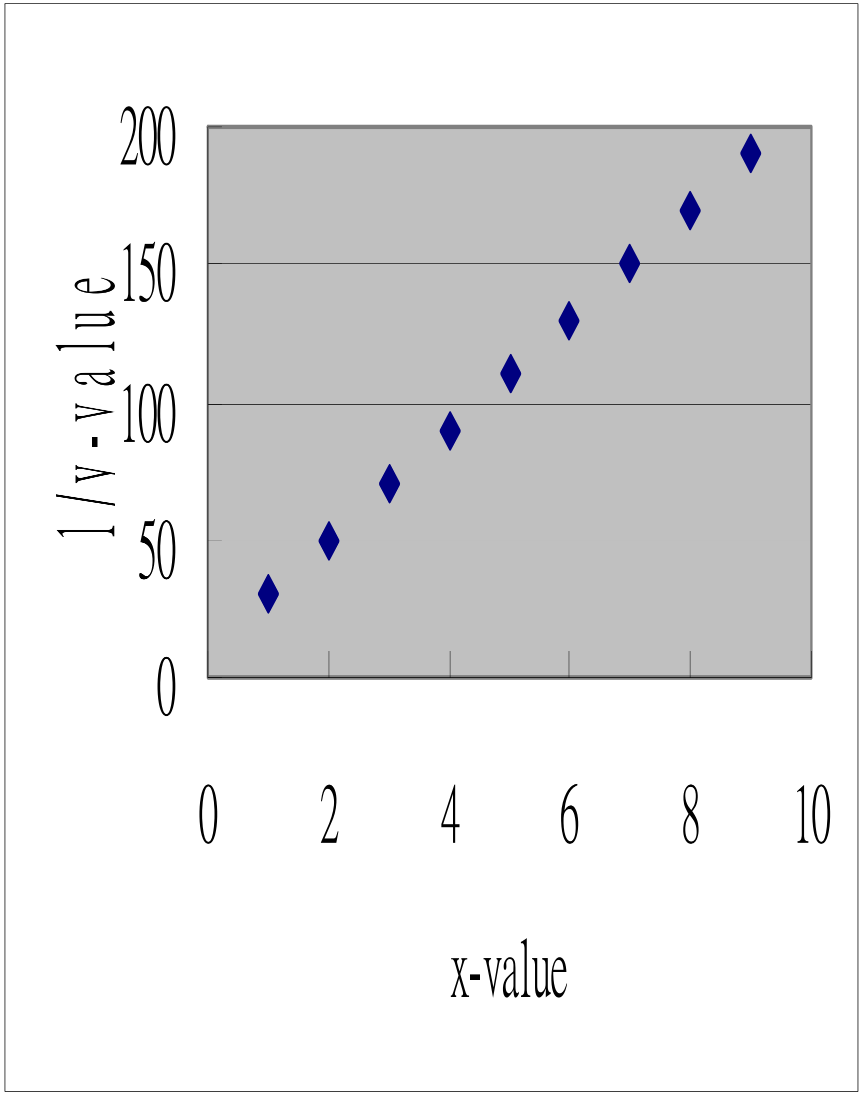
where :

$$\hat{y}' = \frac{1}{\hat{y}}$$

Hence we map $\hat{y}'$ vs. x

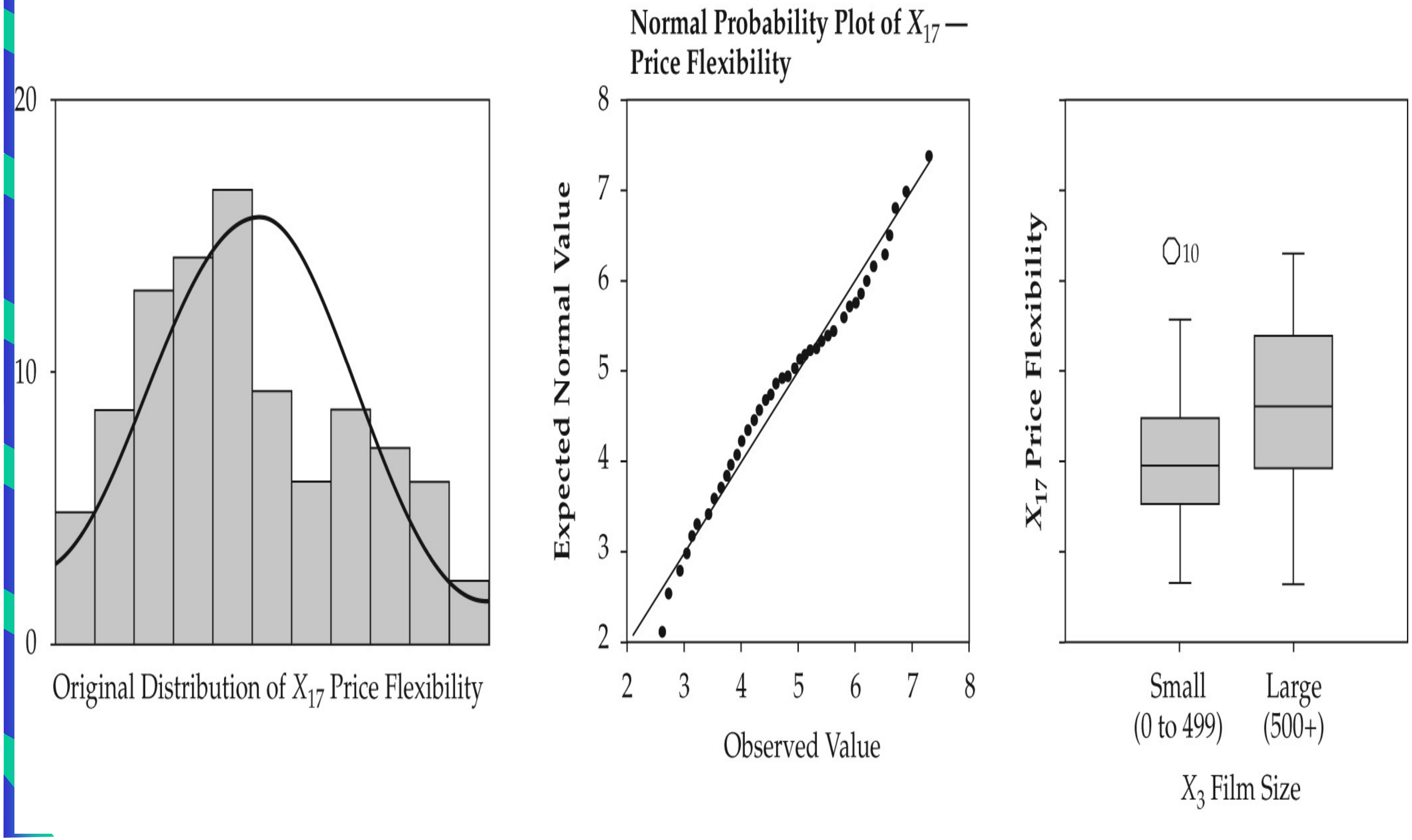
Corresponding Scatter Plot





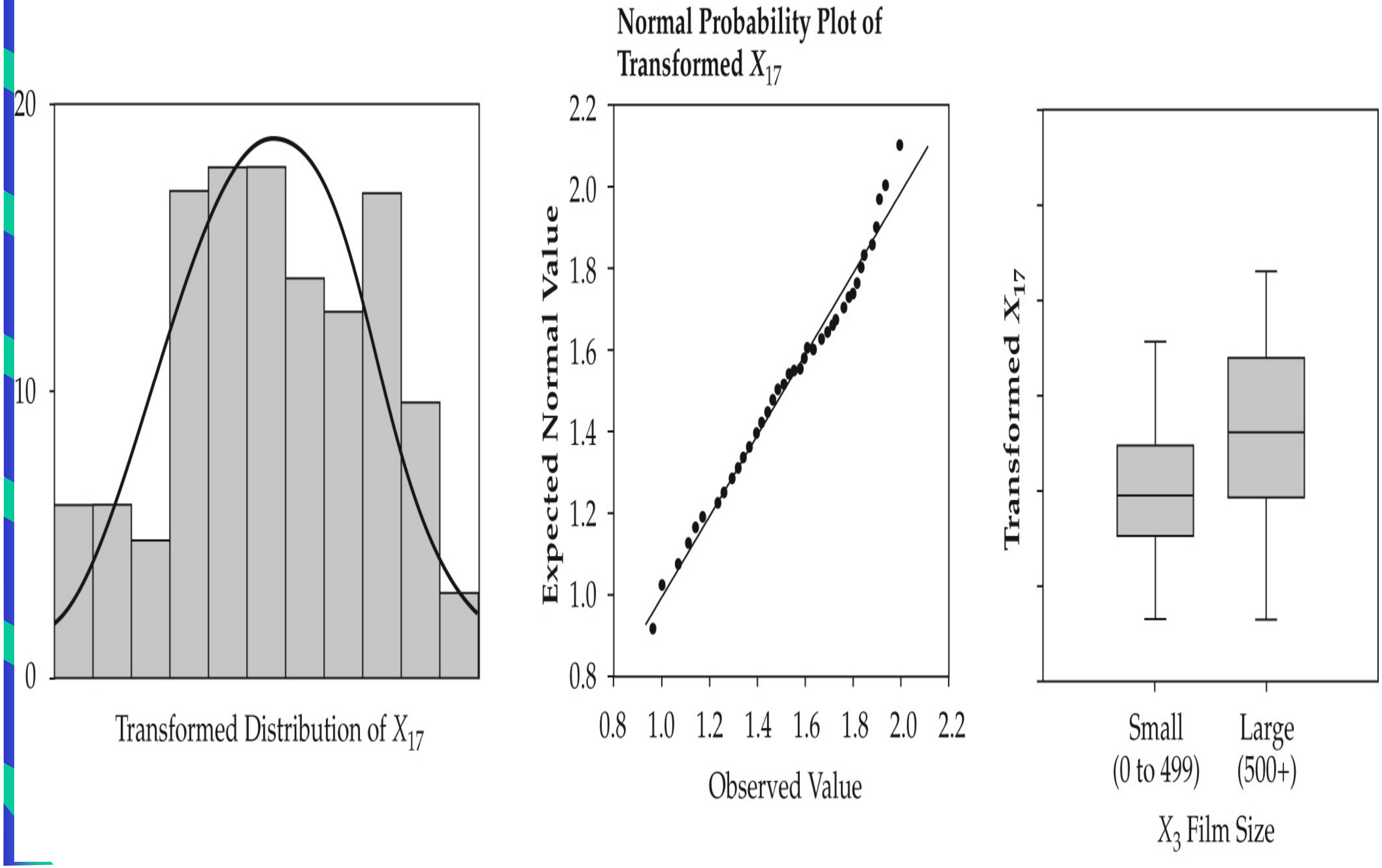
An example, for variable X_{17}

ORIGINAL VARIABLE



An example, for variable X_{17}

TRANSFORMED VARIABLE



Rules of Thumb 2–6

Transforming Data

- To judge the potential impact of a transformation, calculate the ratio of the variable's mean to its standard deviation:
  - Noticeable effects should occur when the ratio is less than 4.
  - When the transformation can be performed on either of two variables, select the variable with the smallest ratio .
- Transformations should be applied to the independent variables except in the case of heteroscedasticity.

Rules of Thumb 2–6

- Heteroscedasticity can be remedied only by the transformation of the dependent variable in a dependence relationship. If a heteroscedastic relationship is also nonlinear, the dependent variable, and perhaps the independent variables, must be transformed.
- Transformations may change the interpretation of the variables. For example, transforming variables by taking their logarithm translates the relationship into a measure of proportional change (elasticity). Always be sure to explore thoroughly the possible interpretations of the transformed variables.
- Use variables in their original (untransformed) format when profiling or interpreting results.

Dummy Variable

Dummy variable . . . a non-metric independent variable that has two (or more) distinct levels that are coded 0 and 1. These variables act as replacement variables to enable non-metric variables to be used as metric variables.

Dummy Variable Coding

Category	X_1	X_2
Physician	1	0
Attorney	0	1
Professor	0	0

Simple Approaches to Understanding Data

- Tabulation = a listing of how respondents answered all possible answers to each question. This typically is shown in a frequency table.

- Cross Tabulation = a listing of how respondents answered two or more questions. This typically is shown in a two-way frequency table to enable comparisons between groups.

Simple Approaches to Understanding Data

- Chi-Square = a statistic that tests for significant differences between the frequency distributions for two (or more) categorical variables (non-metric) in a cross-tabulation table. Note: Chi square results will be distorted if more than 20 percent of the cells have an expected count of less than 5, or if any cell has an expected count of less than 1.

- ANOVA = a statistic that tests for significant differences between two means.

Cleaning and Transforming Data

Learning Checkpoint

1. Why examine and clean your data?
2. What are the principal aspects of data that need to be examined?
3. What approaches would you use?
4. Why would you transform data?

The end, thank you.