Analyzing Measurement Data

Chapter 6

Probability Distribution

- A probability distribution for an experiment is the assignment of probability values to each of the possible outcomes.
- A **probability density function** (PDF) is a mathematical function that describes a continuous probability distribution. It provides the **probability density** of each value of a variable, which can be greater than one.

Classical Data Analysis Techniques

- After collecting relevant data we must analyze it appropriately. We need a proper understanding of the following notions:
  - Measure of central tendency
  - Measure of dispersion
  - Distribution of data
  - Student’s *t*-test
  - *F*-statistic
  - Level of significance
  - Confidence limits…. and more

Nature of Data

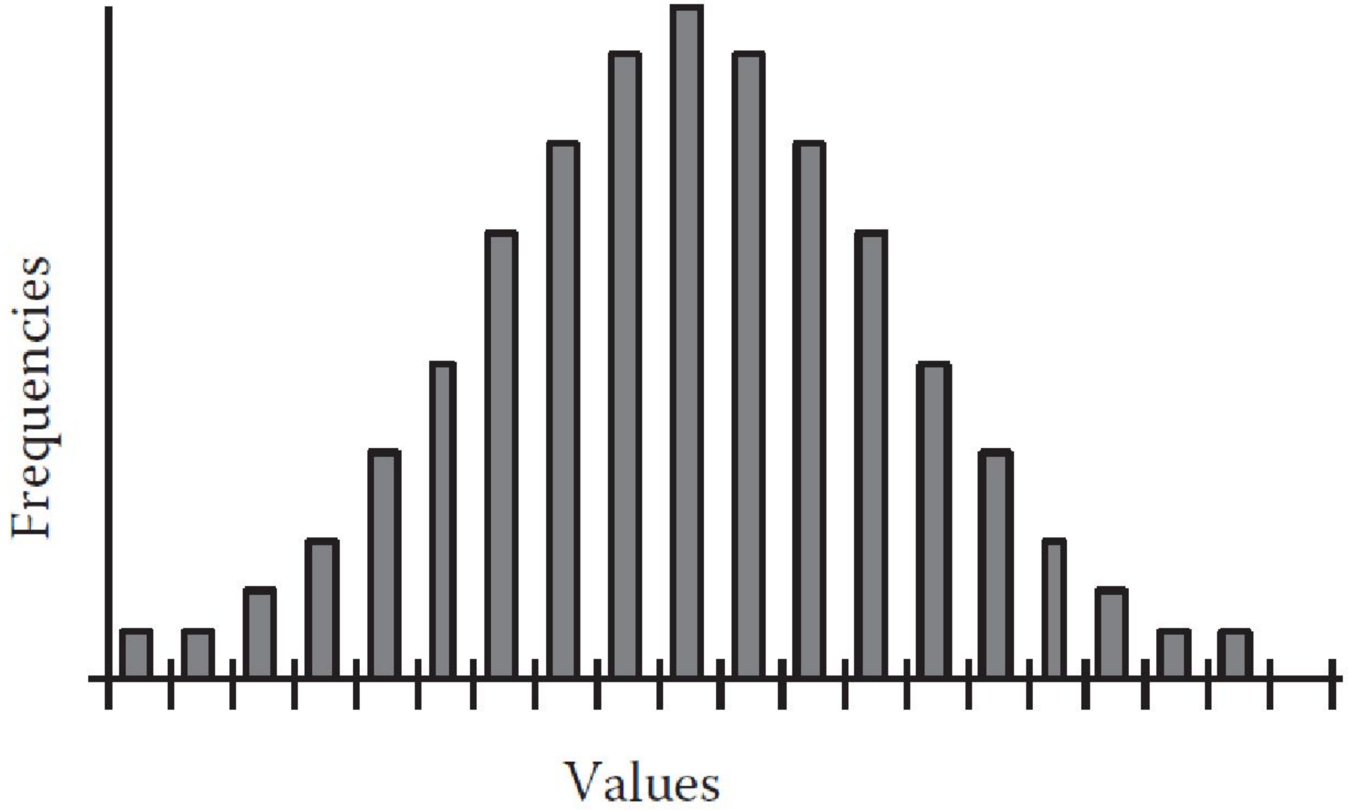


FIGURE 6.4 Data resembling a normal distribution.

Nature of Data

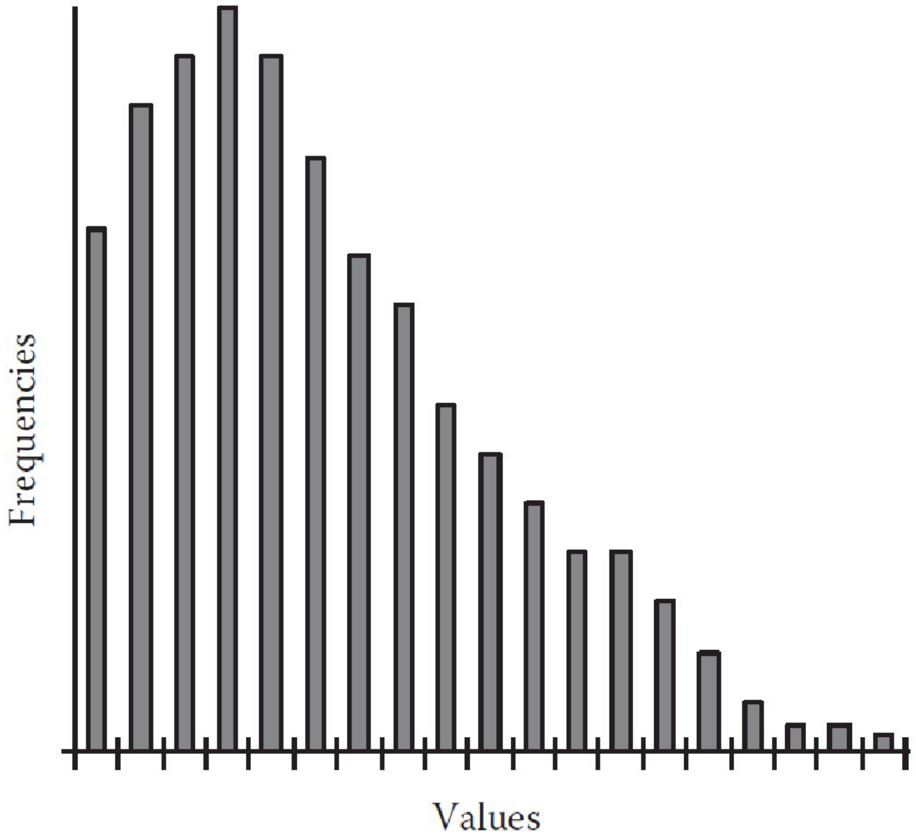


FIGURE 6.5 Distribution where data are skewed to the left.

Nature of Data

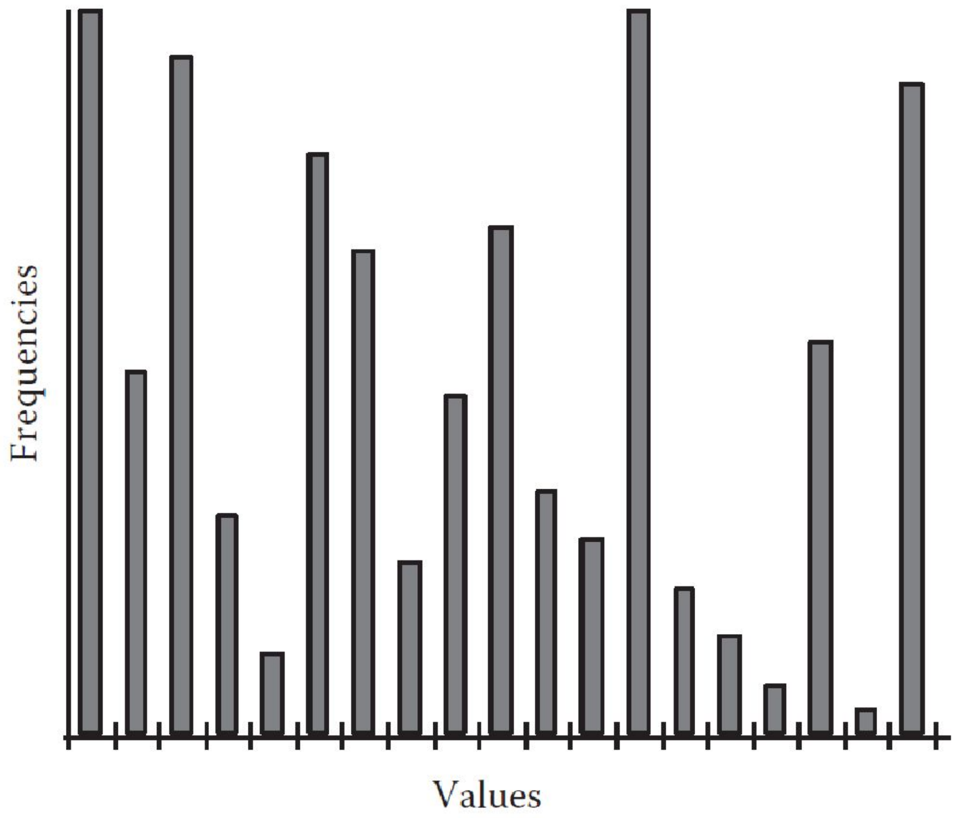


FIGURE 6.6 Nonnormal distribution.

Distribution of software measurements

- Data used for most of the software measurements may follow the different types of distribution. The normal distribution is common.
- When we do not know anything about the distribution, there are a number of approaches to dealing with our lack of knowledge:
  - *Robust statistics and nonparametric methods*. Regardless of whether the data are normally distributed or not, *robust methods* yield meaningful results. *Nonparametric statistical techniques* allow us to test various hypotheses about the dataset without relying on the properties of a normal distribution.
  - In particular *nonparametric techniques* often use properties of the *ranking* of the data.
  - Attempt to transform basic measurements into a scale in which the measurements conform more closely to the normal distribution

Hypothesis Testing Approaches

- A key criterion for testing a hypothesis is a "test of significance" which evaluates the probability that a relationship was due to chance.
- The classical approaches examine whether or not the null hypothesis can be refuted with some predetermined confidence level, often 0.05 (5%).
- Using the 0.05 confidence level, we can refute the null hypothesis only if our evidence is so strong that there is only a probability of 5% that, in spite of an apparent relationship, the null hypothesis is really true.

Type-I and Type-II errors

- Rejecting the null hypothesis when it is true is a *Type-I error*
- Accepting the null hypothesis when it is actually false is called a *Type-II error*

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

Example Datasets

TABLE 6.1A Dataset 1

Project Effort (Months)	Project Duration (Months)	Product Size (Lines of Code)
16.7	23.0	6050
22.6	15.5	8363
32.2	14.0	13,334
3.9	9.2	5942
17.3	13.5	3315
67.7	24.5	38,988
10.1	15.2	38,614
19.3	14.7	12,762
10.6	7.7	13,510
59.5	15.0	26,500

Example Datasets

TABLE 6.1B Dataset 2

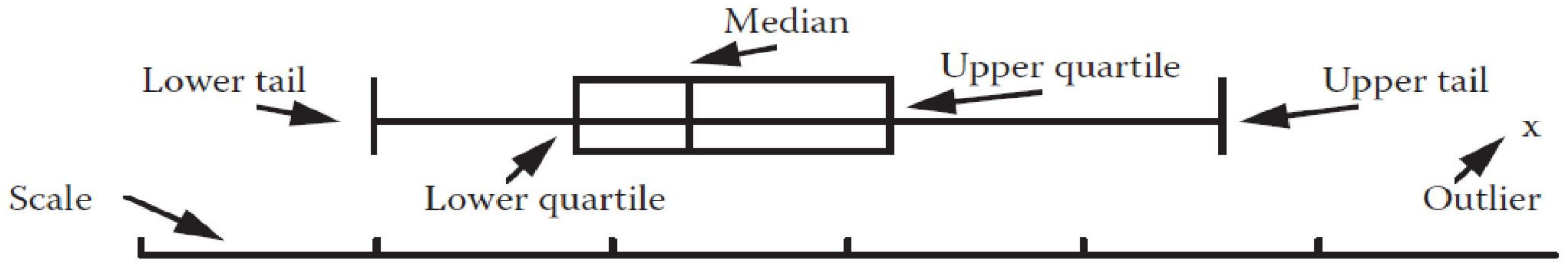
Module Size	Module Fan-Out	Module Fan-In	Module Control Flow Paths	Module Faults
29	4	1	4	0
29	4	1	4	2
32	2	2	2	1
33	3	27	4	1
37	7	18	16	1
41	7	1	14	4
55	1	1	12	2
64	6	1	14	0
69	3	1	8	1
101	4	4	12	5
120	3	10	22	6
164	14	10	221	11
205	5	1	59	11
232	4	17	46	11
236	9	1	38	12
270	9	1	80	17
549	11	2	124	16

Analysis Techniques

- Many robust techniques are useful with software measurement data, *regardless of the distribution* (i.e., normal, non-normal)
- You can implement them using simple spreadsheets or statistical packages.

Example of Simple Analysis Techniques

- Box plots
  - Software measurement datasets are often not normally distributed, and the measurements may not be on a ratio scale. Box plots use to represent non-normal data.
  - Box plots are constructed from three summary statistics: the median (*m*), the upper quartile (*u*), and the lower quartile (*l*). Where (*u) is defined as* the median of the values more than *m. (l) defined as the median of the values less than m*
  - The box length (*d*) is the distance from the upper quartile (*u*) to the lower (*l*).
  - The tails are the theoretical bounds between which we are likely to find all the data points if the distribution is normal



Box plots

Upper tail value = *u+1.5d*
Lower tail value = *l*-1.5*d*

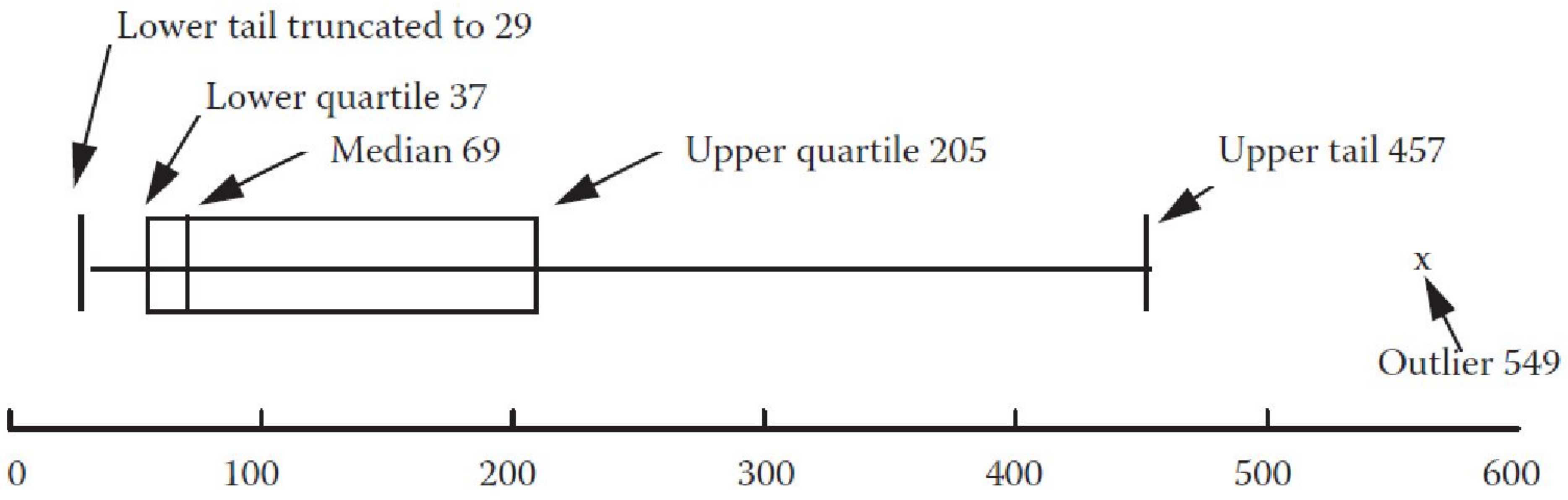
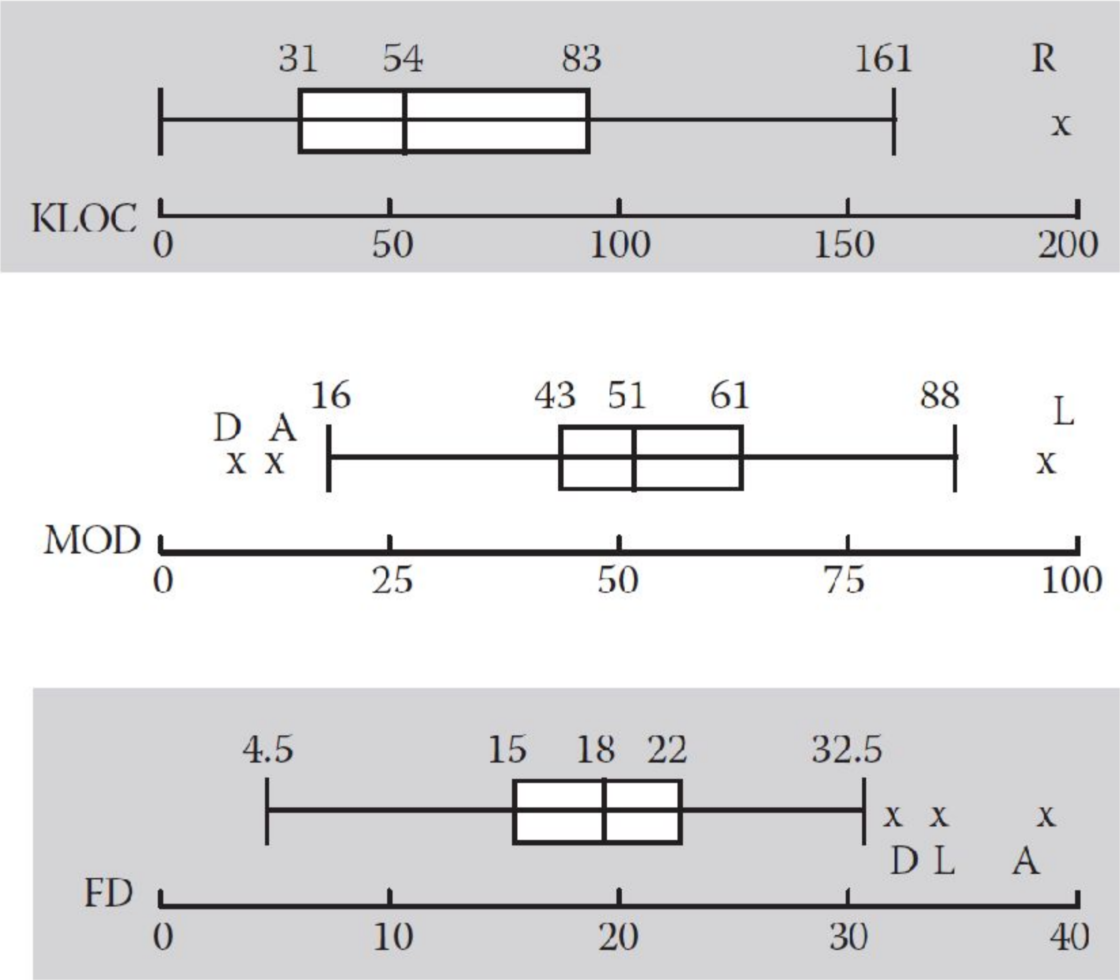


FIGURE 6.9 Box plot of lines of code (17 procedures) for dataset 2 of Table 6.1b.

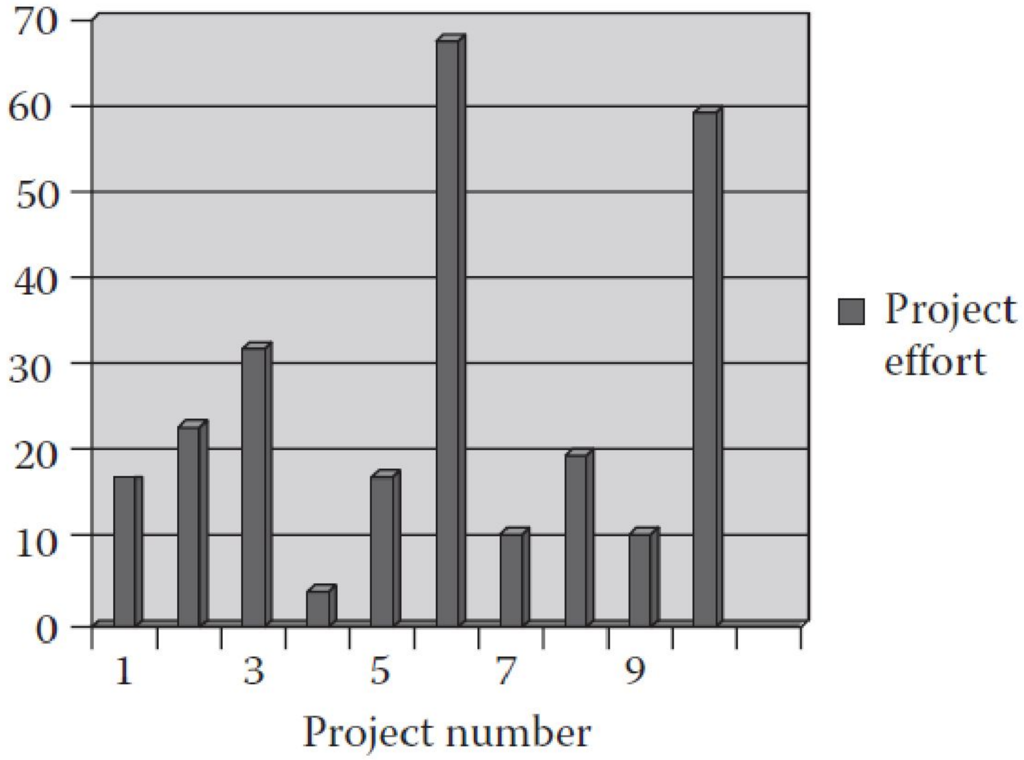
Box plots

System	KLOC	MOD	FD
A	10	15	36
B	23	43	22
C	26	61	15
D	31	10	33
E	31	43	15
F	40	57	13
G	47	58	22
H	52	65	16
I	54	50	15
J	67	60	18
K	70	50	10
L	75	96	34
M	83	51	16
N	83	61	18
P	100	32	12
Q	110	78	20
R	200	48	21



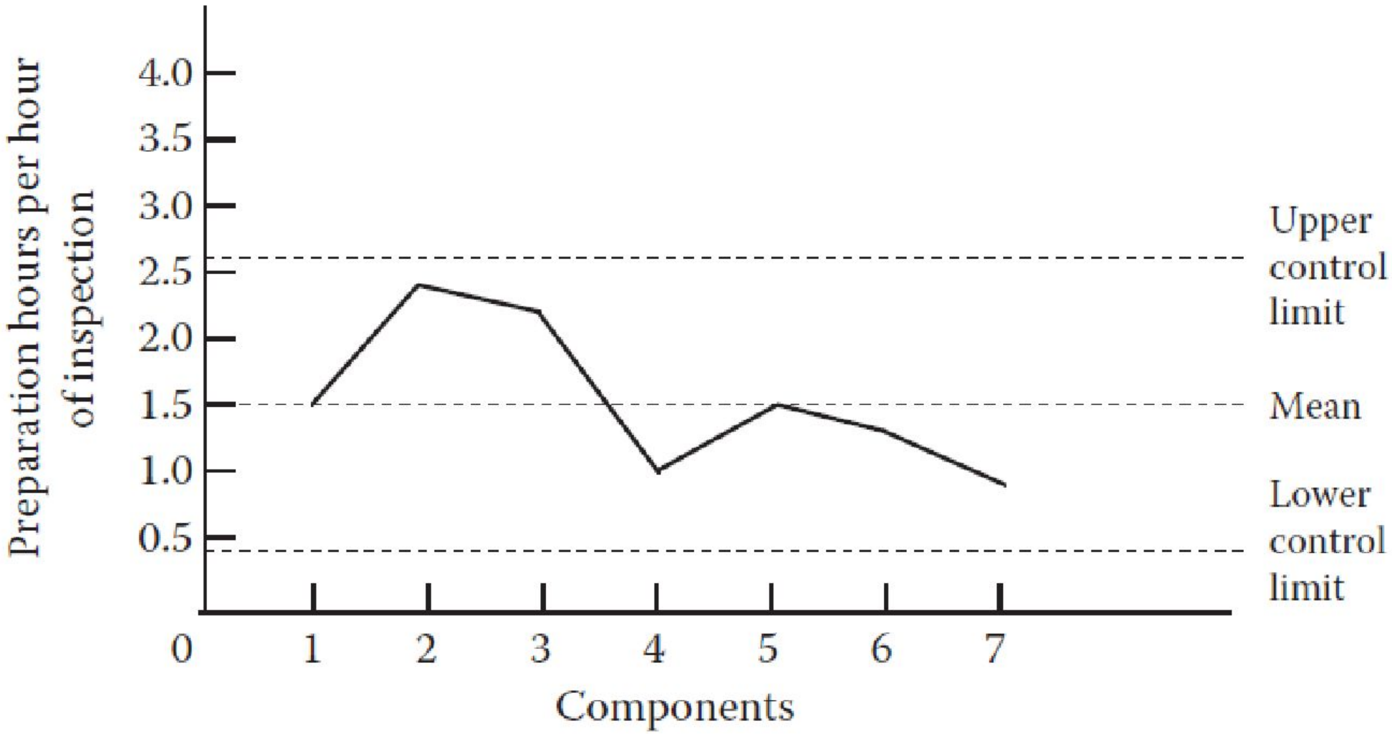
Bar Charts

- Unlike box plots, bar charts allow us to readily identify the entity associated with each measured value.



Control Charts

- Helps to see when your data are within acceptable bounds.



Scatter plots

- Useful to represent relationship between two attributes (pair)
- Not helpful for more than two variables

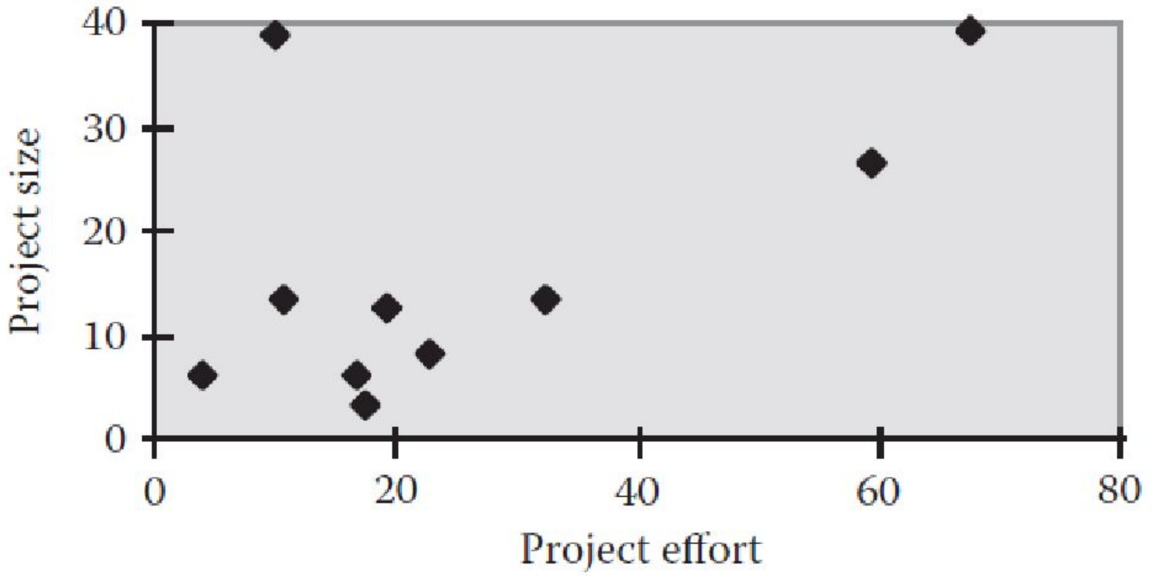


FIGURE 6.13 Scatter plot of project effort against project size for dataset 1.

Measures of Association

- Scatter plots depict the behavior of two attributes, and sometimes we can determine that the two attributes are related.
- The appearance of a relationship is not enough evidence to draw conclusions.
- Statistical techniques that can help us evaluate the likelihood that the relationship seen in the past will be seen again in the future.
- We call these techniques ***measures of association***, and the measures are supported by statistical tests that check whether the association is significant.

Pearson correlation coefficient

- For **normally distributed attribute** values, the *Pearson correlation coefficient* is a valuable measure of association
- Let's consider two attributes, say *x (size of a module)* and *y (number of faults in the module)*.
- If the datasets of *x* and *y* values are normally distributed (or nearly), then we can form pairs (x_i, y_i), where there are *i* software items and we want to measure the association between *x* and *y*.
- The total number of pairs is *n*, and for each attribute, we calculate the *mean* and *variance*. We represent the *mean* of the *x* values by m_x, and the *mean* of the *y* values by m_y. Likewise, var(*x*) is the variance of the set of *x* values, and var(*y*) is the variance of the *y* values. Finally, we calculate

$$r = \sum_{i=1}^{n} \frac{(x_i - m_x)(y_i - m_y)}{\sqrt{n \operatorname{var}(x) \operatorname{var}(y)}}$$

Pearson correlation coefficient

- The value of r, called the correlation coefficient, varies from −1 to 1.
- When r is 1, then x and y have a perfect positive linear relationship; that is, when x increases, then so do y in equal linear steps.
- Similarly, −1 indicates a perfect negative linear relationship (i.e., when x increases, y decreases linearly), and 0 indicates no relationship between x and y
- Statistical tests can be used to check whether a calculated value of r is significantly different from zero at a specified level of significance; in other words, the computation of r must be accompanied by a test to indicate how much confidence we should have in the association (strength of association).
- **Chi-square test?**

Linear Regression

-
- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- Linear regression expresses an association (between dependent and independent variables) as a linear formula
- The simple linear model is expressed using the following equation:

$$y = a + bx + r$$

$$b = \frac{\sum(x_i - m_x)(y_i - m_y)}{\sum(x_i - m_x)^2}$$

$$a = m_y - bm_x$$

Where:

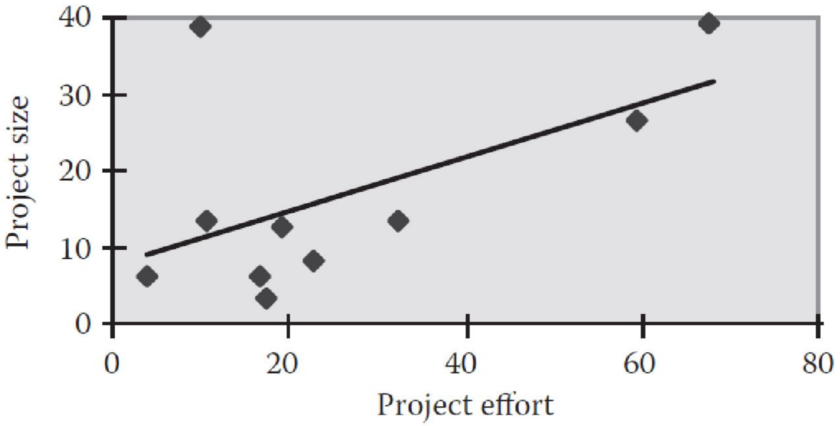
y – Dependent variable

x – Independent (explanatory) variable

a – Intercept (constant)

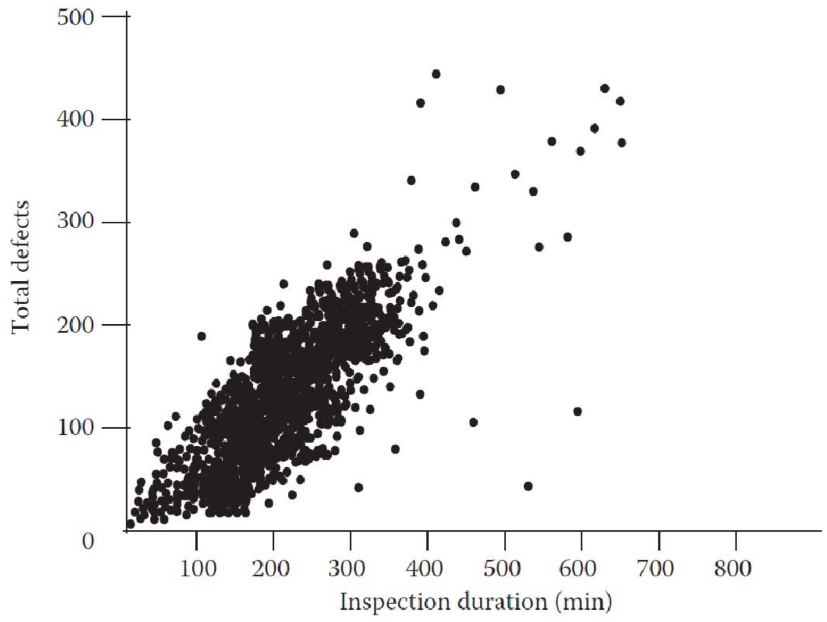
b – Slope

r – Residual (error)



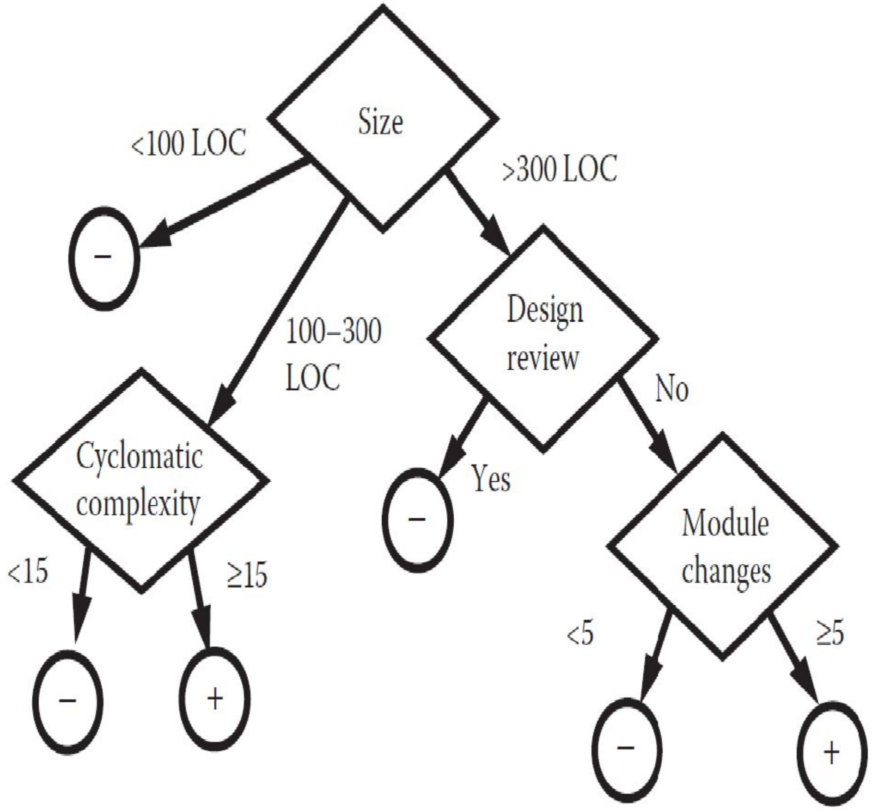
Linear Regression

- The vertical distance from each point to the trend line represents the discrepancy between the data and the line. We need to keep the distance as small as possible.
- The discrepancy for each point is called the residual, and the formula for generating the linear regression line minimizes the sum of the squares of residuals
- The residuals values can be plotted (scatter plot) in a graph and can demonstrate if any values are unusually large.

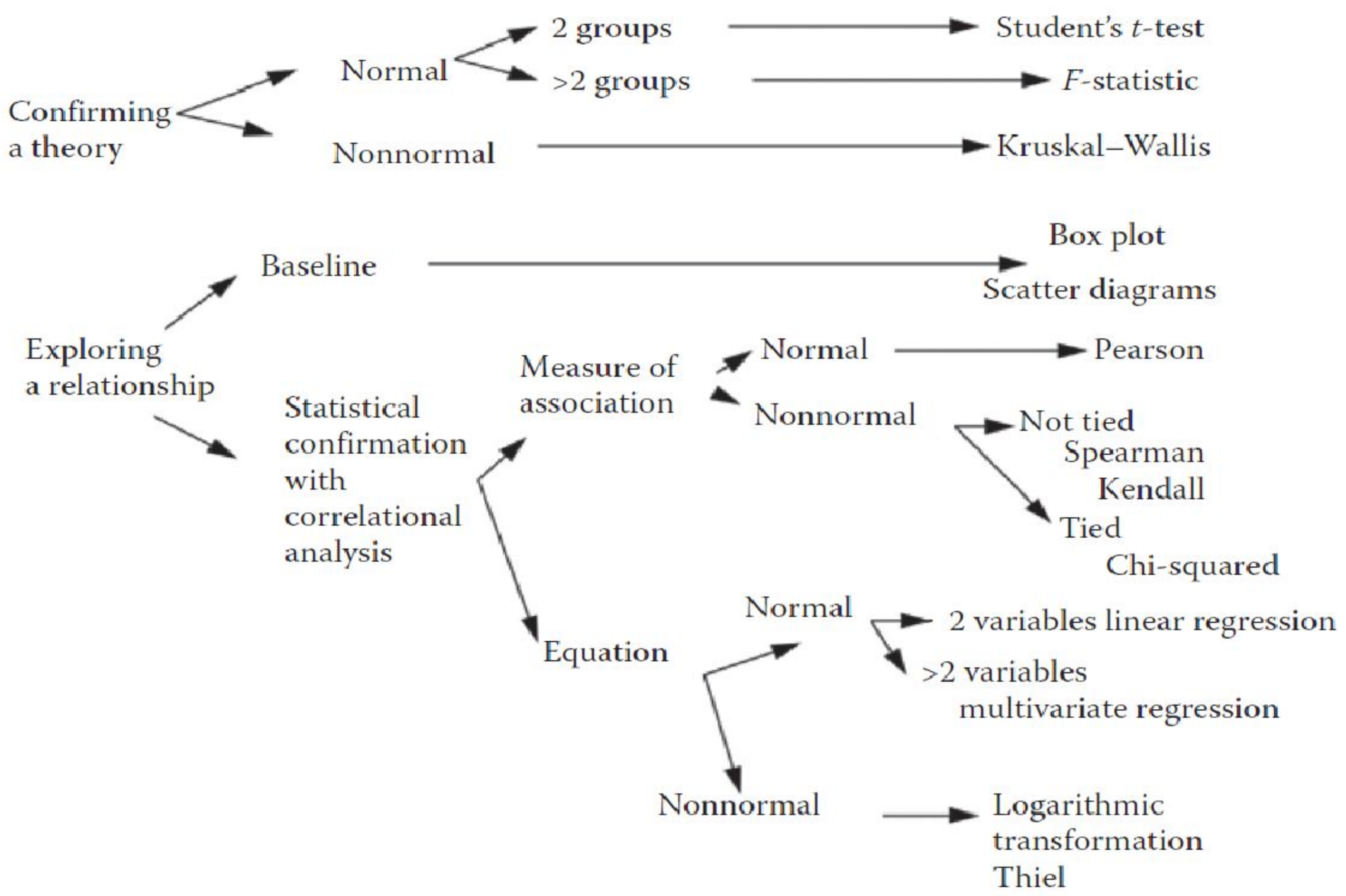


Classification Tree Analysis

- Often we want to know which measures provide the best information about a particular goal or behavior. That is which measures are best predictors of the behavior in a given attribute.
- Classification tree analysis can be used to address this problem.
- Suppose you want to determine which of the metrics are the best predictors of poor quality code modules among a large number of collected code module data.



+ poor quality;
- good quality


Confirming a theory
Normal
2 groups
Student's t-test
>2 groups
F-statistic
Nonnormal
Kruskal–Wallis
Exploring a relationship
Baseline
Box plot
Scatter diagrams
Statistical confirmation with correlational analysis
Measure of association
Normal
Pearson
Nonnormal
Not tied
Spearman
Kendall
Tied
Chi-squared
Equation
Normal
2 variables linear regression
>2 variables
multivariate regression
Nonnormal
Logarithmic transformation
Thiel

Principal Component Analysis (PCA)

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster.
- So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.