A Review of Statistical Methods for Medical and Allied Health Professionals

A Review of Statistical Methods for Medical and Allied Health Professionals

By

Antoine Al-Achi

Cambridge
Scholars
Publishing

A Review of Statistical Methods for Medical and Allied Health Professionals

By Antoine Al-Achi

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-8858-0
ISBN (13): 978-1-5275-8858-5

To my wife Pamela and my children Elias Gabriel, Anthony William, and John Peter

To Bailey and Charley

To my beloved parents, Elias Antoun Al-Achi and René Nassif Tadros Al-Achi

To my siblings and their families, Peter, Claudette, and Kamil

To my students, past, present, and future

TABLE OF CONTENTS

Preface and Acknowledgments x

Lecture One 1
Descriptive Statistics
The basics 1
Grouped frequency table 2
Measurement scales 3
Measures for location 3
Measures for dispersion 5
Measures for location and dispersion from a grouped frequency table 7
The box plot 9
Skewness and kurtosis 12
Handling outlying values 13
Precision, accuracy, and bias 14
Useful rules 15
Practice examples 15
Practice problems 23
A quick review 37

Lecture Two 42
Inferential Statistics
The normal distribution 42
The central limit theorem 43
Hypothesis testing 44
Confidence intervals 49
The binomial distribution 51
Probability rules 59
Bayes' rule 61
T-scores 66
Practice examples 66
Practice problems 74
A quick review 88

Lecture Three 94
Hypothesis Testing: Parametric Tests Z-Test and t-Test
Introduction 94
Z-test for two samples 94
The t-test 95
Two-sample t-test 98
Are the variances homogenous? 103
The paired t-test 108
Concluding remarks 112
Practice examples 112
Practice problems 120
A quick review 126

Lecture Four 130
Hypothesis Testing: Parametric Tests Analysis of Variance (ANOVA)
Introduction 130
Are the variances homogenous? 139
Repeated measures ANOVA 140
Repeated measures cross-over design 140
Repeated measures over time design 142
Practice examples 151
Practice problems 156
A quick review 158

Lecture Five 160
Hypothesis Testing: Nonparametric Tests
Non-parametric tests 160
The Chi-square test 160
Single-sample chi-square: An extended example 162
Two or more samples chi-square: An extended example 163
Mann-Whitney U test 166
Kruskal-Wallis H test 167
Wilcoxon Matched-Pairs (Signed Rank Test) 168
Friedman's Matched Samples 169
McNemar's test 171
Practice examples 172
Practice problems 185
A quick review 187

Lecture Six 189
Clinical Parameters
Introduction 189
Clinical parameters 190
A quick review 194

Lecture Seven 196
Linear Correlation and Regression
Introduction 196
Methodology 196
Practice examples 201
A quick review 207

Lecture Eight 209
Experimental Design
Introduction 209
Determining of the sample size 209
Defining the outcome(s) of the experiment 210
Identifying input variables 212
Full factorial design 212
DoE general statistical concepts 228
Fractional factorial designs 233
Blocking, anti-drift, and randomization designs 237
Other designs 250
Concluding remarks 250
Practice examples 250
Practice problems 259
A quick review 266

Lecture Nine 273
Advanced Topics
Multivariate analysis 273
Principal components analysis 282
Discriminant analysis 298
Cluster analysis 326
Survival analysis 333
Multiple Linear Regression 336
Meta-analysis 345
Heterogeneity 348
Reliability 349
Practice problems 358
Practice exercises 364
A quick review 370

Epilogue 374

General Exercises by Selected Topics 376
Descriptive statistics 376
Inferential statistics 387
Hypothesis testing (Z- and t- tests) 398
Analysis of variance test 404
Non-parametric tests/clinical parameters 405
Linear correlation and regression 412
Multivariate analysis, cluster analysis, discriminant analysis, reliability, multiple linear regression 415

General Exercises 430

Answers to General Exercises 501

Cumulative Quick Review Concepts 562

Answers to Cumulative Quick Review Concepts 573

Equations 579

Bibliography/References 588

Index 590

PREFACE

Statistics is a subject that benefits many other disciplines in its applications. Medical science statistical applications have tremendously contributed to the advancement of medicine, particularly during the 20th Century and beyond. In recognition of the central role of statistics in health fields, certification agencies have incorporated this science into their members' requirement of knowledge acquisition. This centrality of statistical knowledge is reflected in the board exams, particularly those taken for clinical board specialty certification tests. This text's primary objective is to reinforce statistical principles for those who have taken a statistics course during their years of education. It is structured so that health students, including medicine, can read the chapters, solve the problem sets, and answer the theoretical questions related to health science's statistical methods. The text is not meant to be exhaustive in its content. Rather it reintroduces the reader to concepts that they may find helpful in their preparation for tests or reading a scientific paper with statistical applications included in it. I envision that this text will benefit students in healthcare sciences, healthcare professionals, graduate students conducting research, and people who desire to learn more on this subject.

The discussion herein serves as a quick and ready guide to statistical tests. The scientific methods are the foundation on which experimental research is conducted. Having observed a phenomenon that needed an explanation, a hypothesis is generated to test it via experimentation. Statistics is a science that includes planning, gathering, organizing, and analyzing observations. During the planning phase of the research, it is important to decide the type of data to be collected and the statistical tests to be used for the analysis. Once the data is collected, it must then be organized in various forms such as summary tables, graphs, and numerical designations. Tables are often used to group the raw data systematically. This data organization helps find and detect trends in the collected data. Graphs offer a visual representation of the information that also shows trends in the data and the presence of gaps and outlying values. The numerical summaries are perhaps the most important because they allow groups to be compared concerning these summaries.

Researchers may compute various measures to describe the data. These are measures for central tendency, measures for variability, kurtosis measures, and measures for skewness. While kurtosis and skewness represent the deviation of a bell-shaped curve from symmetry, measures of central tendency and variations are used to provide markers or points on the scale of the variable being studied and the dispersion of the observation about a central value, respectively. The arithmetic mean, the median, and the mode are collectively known as averages. Besides, the quantiles (quartiles and percentiles) are also used to measure a central tendency. Measures for variability describe the data's dispersion, the most important of which is the standard deviation (or the variance, the square of the standard deviation). Other variability measures include the range (the most straightforward), interpercentile ranges, and the coefficient of variability percent (%CV). The latter is often used in the United States Pharmacopeia and is the relative standard deviation percent (%RSD) (i.e., %CV = %RSD). The coefficient of variation percent is simply the standard deviation ratio to the arithmetic mean multiplied by 100. It compares data concerning their variability about the mean of two or more variables. For example, to answer the following question, "which of the two factors is more variable concerning their means, the volume of distribution or the average steady-state concentration? Since these two variables have different units, L/kg for the volume of distribution and µg/mL for steady-state concentration, they cannot be compared directly using the standard deviation value. Thus, %CV is used to standardize the mean value variability and express it as a percent. Once both values are expressed as a percent, the investigator can compare them. Although the arithmetic mean and the standard deviation are sensitive to extreme values, they are considered the most reliable measures for comparing groups.

The science of statistics allows the researcher to reach a conclusion on the tested hypothesis. This conclusion is always made with a given confidence level. Commonly, the 95% or the 99% confidence levels are chosen to report a study's findings. The uncertainty in writing the conclusion stems from the fact that experimentations are conducted using samples rather than populations. A sample is considered to be a small part of a population. A random sample is obtained by a random selection from a population. Random samples are considered to be representative of their populations. For example, people vary in their rate of metabolizing drugs. If a pharmacokinetics study intends to test drug metabolism, then a representative sample must reflect the actual variability in drug metabolism existing in the population. A non-randomly selected sample might bias the results and yield conclusions that either over-estimate or under-estimate the rate of metabolizing drugs. Similar examples apply to drug absorption, distribution, and elimination. Thus, it is of the highest importance that scientifically conducted studies employ randomly obtained samples. Random sampling allows for fair comparison among groups. Incidentally, the best mean value to represent a "process rate" is the geometric mean which is the nth root of the product of all observations in the sample.

Moreover, random samples must be large enough to detect the hypothesized differences among groups. In planning the experiment, one of the first tasks for the researcher is to select a sample size. This size selection of a random sample requires knowledge of specific statistical components. The first is the uncertainty by which the conclusions are going to be reported. As stated earlier, the degree of uncertainty is often chosen to be either at the 5% or 1% level (or at the 95% or the 99% confidence level). This uncertainty is called the level of significance, also known as alpha. The level of significance is a probability term. We examine our hypothesis against the null hypothesis (H_0) in statistical testing. The statement of H_0 is a declaration of no difference. Therefore, the conclusion of any study is about H_0. Once the statistical analysis is performed, the decisions drawn from the analysis will focus on either rejecting or not rejecting the null hypothesis. Although the null hypothesis' true nature cannot be reached with 100% certainty, one can declare the study's conclusion with a given uncertainty, as stated above. The probability that the null hypothesis is rejected, whereas it is, in reality, true, is known as alpha (α), or the level of significance, also known as the probability of Type I Error. This probability is kept very low, generally 5% or below. Another term related to the null hypothesis is beta (β), or the probability of Type II Error. This term is defined as the probability of not rejecting a false null hypothesis. In science, beta is kept at 20% or less. The complement of β is the power of the test. It is defined as the probability of rejecting the null hypothesis, given the null hypothesis is false. The power of the test is commonly set at 80% or higher. Whereas β and power complement each other, the complement of α is known as the confidence coefficient or the confidence level.

In general, statistical tests are broadly classified into two main categories, parametric and nonparametric. The term "parametric" implies that these tests depend on evaluating parameters obtained from a distribution. Nonparametric tests do not include parameters in comparing groups; instead, they rely on comparing groups for counts, proportions, or ranks. Data, in general, can be classified in one of the following four scales: nominal, ordinal, interval, or ratio. For instance, most clinical variables are either interval or ratio in nature. For example, tissue concentration of digoxin in the heart muscle is a ratio scale variable. The temperature of the patient in degrees Celsius is an interval-type variable. The main difference between these two scales is that ratio scale variables have an actual zero point, whereas, in interval variable scales, the zero point is determined arbitrarily. Collectively, interval and ratio scale data are continuous. Ordinal scale data have "order" or "rank" to them, similar to those of continuous variables. However, the distance between the categories in an ordinal scale variable cannot be assumed to be equal.

When the data consists of only named categories, the data is classified as a nominal scale. For example, the concentration of the drug in biological tissues is a nominal scale variable. In this case, the drug concentration may be in plasma, urine, liver, brain, or other tissues. These "levels" are named categories, and each concentration belongs to one and only one of them. For example, the drug concentration in brain tissues is always found under that category and cannot be listed under the other categories. Thus, these categories are mutually exclusive. It should be emphasized here that the concentration value is ratio-type data. However, "the category" as a variable is nominal. Another example of nominal scale data is the drug type. For example, in a pharmacokinetics study, researchers may be interested in testing the difference in the oral absorption of a generic versus a branded form of a drug. In this case, "generic or branded" is a nominal scale type variable, whereas expressing the "rate of absorption" numerically is a ratio type scale.

Parameters are characteristics that belong to a population, whereas statistics are characteristics obtained from samples. For example, the mean volume of distribution of a drug is 0.48 L/kg with a standard deviation of 0.05 L/kg. If these two values represent the population, they are parameters; otherwise, they are called statistics if obtained from samples. To differentiate between parameters and statistics, Greek letters are used for parameters, and Latin letters are used for statistics. For example, the population mean is (μ = 0.48 L/kg). The symbol for the standard deviation (a measure of the variability) in the population is sigma (σ). In testing hypotheses using parametric tests, the mean and the standard deviation are used for comparing groups. For example, the t½ of phenobarbital in the population is 5 days with a standard deviation of 0.5 days.

The simplest form of analysis of variance test (ANOVA) is a one-way ANOVA applicable when the data has one dependent variable and one independent variable. A two-way ANOVA has two independent variables and a single dependent variable. Other types of ANOVA are also used, such as multifactorial ANOVA (multiple dependent variables and multiple independent variables) and repeated measures ANOVA applied to cross-over design experiments. The repeated measures ANOVA sometimes is known as the "within subjects design."

Nonparametric tests are mainly used when the outcome is either a nominal or an ordinal scale variable. These tests do not mandate that the outcome's distribution be normally distributed (they are often labeled as "distribution-free" tests). In studies that result in a severe skewness in the dependent variable distribution and where the number of observations is relatively small, nonparametric tests are preferable over the parametric tests in analyzing the data. It should be noted that as a group, nonparametric tests are less powerful and less flexible than parametric tests. One of the most applicable nonparametric tests is the Chi-square test (χ^2) which is appropriate when the outcome is nominal and the groups being compared are independent. The chi-square test outcome is based on count data converted to a proportion of the sample that meets a specific attribute. The Chi-square distribution has positive values only (0 to $+\infty$) with a shape resembling a positively skewed bell-shaped curve. If the groups are dependent on a nominal outcome, then

McNemar's test is applicable. McNemar's test uses the Chi-square distribution. Suppose the output is ordinal and the groups are independent. Two possible nonparametric tests can be used, the Mann-Whitney U test (for comparing two groups) or the Kruskal-Wallis H test (for two or more groups' comparison). The Mann-Whitney U test uses the normal distribution as its statistics are z values, while the Kruskal-Wallis H test depends on the Chi-square distribution. Friedman's test may be used when the outcome is ordinal, and the groups are dependent. Friedman's test applies when more than two groups are evaluated, and two independent variables are present simultaneously. In essence, Friedman's test is a nonparametric two-way ANOVA. Its statistics follow the Chi-square distribution. Wilcoxon's Sign Rank test results in a calculated z statistics value compared to the standard normal distribution. The procedure for all of these tests is very similar:

1) State the null and alternate hypotheses at a given alpha value

2) Calculate test statistics (a computed value for Chi-square or z, depending on the test being used)

3) Compare the calculated value with a tabulated value

4) Build a confidence interval on the true proportion that is expected in the population

5) Decide whether or not to reject the null hypothesis

Many statistical software programs perform the above or similar tests found in the literature. Computer programs calculate a *p-value* for the test to determine whether or not the results are significant. This decision, of course, is accomplished by comparing the computed *p-value* with a predetermined α value. JMP® Pro Statistical Discovery Software, <Versions 14.0 and 15.0>, SAS Institute Inc., Cary, North Carolina is the program used in this text. It is a "user-friendly" approach to data analysis for beginners and advanced users alike.

This text is divided into nine lectures. The content of these lectures is herein summarized:

Lecture One
This lecture covers essential information on how raw data collected from experiments can be organized into tables, graphs, and numerical measures. The latter includes measures for central tendency, variations, skewness, and kurtosis. Analysis of outlying values is also presented. Numerical and theoretical examples are introduced throughout the text.

Lecture Two
The notion of statistical application is further explored with emphasis on hypothesis testing. The reader will review concepts related to Type I and Type II errors and their related concepts of confidence level and test power. The Gaussian and binomial distributions are covered in some detail. Readers will find the concepts related to confidence intervals helpful in relating to the testing hypothesis. Examples of how to approach hypothesis testing are also included throughout the lecture.

Lecture Three
In Lecture Three, hypothesis testing's theoretical basis is further examined and expanded using the Z- and t-test platforms. Also, the homogeneity of variances test is covered in this lecture. Finally, numerous examples of applications of the two tests are introduced.

Lecture Four
The various forms of the Analysis of Variance test (ANOVA) are introduced herein. These include one-way, multifactorial, repeated measure ANOVA, and the Analysis of Covariance (ANCOVA) method. Examples, both numeric and theoretical, are being used to emphasize the study concepts.

Lecture Five
Nonparametric tests are covered herein as opposed to parametric tests covered in the earlier lectures. These tests include Chi-square, Mann-Whitney, Kruskal-Wallis, Friedman's, etc. This lecture emphasizes and discusses when to use these tests versus the parametric test. Numerical and theoretical concept examples are included as well.

Lecture Six
The clinical parameters such as the odds ratio, relative risk, absolute (or attributable) risk reduction, relative risk reduction, and number needed to treat are discussed in this lecture.

Lecture Seven
The lecture's subject is the development of a linear function that describes a single input variable on its effect on a single dependent variable. Statistical analysis of the line's slope, the y-intercept, and the correlation coefficient are discussed, along with numerical and theoretical examples.

Lecture Eight
The discussed subjects include general concepts in designing an experiment, full factorial designs, fractional factorial and screening designs, blocking or trial number methods, mixture designs, and optimal designs.

Lecture Nine
Various multivariate analysis concepts, including principal components analysis, discriminant analysis, cluster analysis, and reliability assessment, are covered in this lecture. *Meta-analysis* covers developing confidence interval estimates from reported data in the literature, building a Forrest and a Funnel plot, and calculating a *p-value* for the overall combined literature data. *Survival analysis* pertaining to mortality and morbidity data is introduced. Finally, the hazard ratio and Kaplan-Meier method will be discussed. Multiple linear regression is also included to expand on the concepts discussed in Lecture Seven.

This text serves as a review tool for those who plan to take a board exam in their medical specialty that requires statistics knowledge. Libraries throughout the U.S.A. and in the western world may find this book an excellent addition to their selection. It includes numerous practice questions, both theoretical and computational. Graduate students can also use it in science and related fields to improve their ability to handle empirically collected data. Undergraduate and graduate students taking a college course in statistics will find this text to be most valuable for its content of practice exercises and theoretical review. There has been a plethora of scientific and clinical investigations during the latter part of the 20th Century up until now that provided reliable scientific and clinical applications and discoveries using statistical methods. To that end, medical and allied health programs are increasingly interested in teaching their students more information about these statistical methodologies. The concise presentation and the repetition of ideas throughout the text help solidify learning and retention of knowledge of the various topics presented. In this way, students in health care fields, graduate students, and practitioners can use this book as a tool to reinforce their understanding of data handling and analysis.

General disclaimer: The information provided in this text is intended for informational use only and is not intended to provide any medical or health recommendation or advice. No information in this text is meant to diagnose, mitigate, treat, prevent, cure, alter, or manage a disease state or any health matter. The information presented herein cannot be used as a substitute for a licensed physician's or another healthcare professional's recommendations.

Acknowledgments: The author wishes to thank Cambridge Scholars Publishing (CSP) and its employees for making this text a reality. Thanks to Ms. Rebecca Gladders, Senior Commissioning Editor, for working with me from the beginning on this project. I am also thankful to Ms. Sophie Edminson (Designer) and Ms. Amanda Millar (Typesetting Manager) from CSP for their help during the production phase of the text. Finally, I am also appreciative of the support from my employer, Campbell University, for their commitment to scholarships. Thank you.

LECTURE ONE

DESCRIPTIVE STATISTICS

The basics

Statistics as science has two major branches, descriptive statistics and inferential statistics. The latter will be covered in Lecture Two. The subject of this Lecture is descriptive statistics, the first component of statistics. The purpose is to describe raw data collected from experimentation by organizing the data in tables, graphs, and numerically. By doing so, much of the details in the raw data are lost; however, it is replaced by the revelation of trends. The numerical descriptive statistics obtained from a sample drawn from a population are known collectively as *statistics*. These statistics are representative of their counterpart in the population, the *parameters*. By convention, statistics are symbolized by Latin letters, whereas parameters have Greek symbols. For example, the arithmetic mean as a statistics has the symbol of $\bar{X}$, and μ when describing the mean of the population. In this Lecture, descriptive statistics for samples is discussed.

Consider the data collected from a clinic which is specialized in weight loss. For the last week, 40 new patients attended the clinic, and their initial weight was recorded (Table 1-1).

Table 1-1. The initial weight of 40 new patients who attended a weight loss clinic is shown in pounds.

213	316	198	286	196	205	307	274	197	243
309	298	195	265	200	210	219	310	200	209
211	240	219	286	273	243	249	288	281	201
215	307	222	237	293	333	276	290	293	278

The data in Table 1-1 are known as raw data. They were obtained by measuring the patient's weight using a scale. It is important to note that all the patients must have been weighed using the same scale type. Using different scales does not necessarily guarantee that the collected weights represent a *homogenous* set of values. Six months later, the same patients returned to the clinic having followed a calorie restriction diet. Their weights were measured again using the same scale used initially, and the data is recorded in Table 1-2.

Table 1-2. The patient's weight following a diet for six months recorded in pounds.

200	315	197	280	190	200	305	270	190	243
290	276	190	260	186	200	199	270	178	200
210	210	200	280	265	238	220	280	280	189
200	297	190	211	293	289	264	280	290	243

It appears that most of the patients have lost a significant number of pounds. Plotting the data in Tables 1-1 and 1-2 on a rectangular graph paper resulted in Figure 1-1.

Figure 1-1 shows the initial and six-month weights for every patient. Graphs, in general, are good visual representations of the data. Graphically, this data set does not lead to any conclusion other than that the diet was not effective. It appears from the graph that there was no trend in reducing the weight. A better way to represent the data is to plot the weights side by side using a bar graph (Figure 1-2). Note that a histogram is a bar graph for continuous variables (interval and ratio scale variables; see below). A continuous variable can have any value on a scale (limited to the instrument's sensitivity), whereas a discrete variable can only have whole numbers. Variables can also be qualitative or quantitative. The latter can take on numerical values (e.g., serum cholesterol level), whereas the former can only have categorical values (e.g., a person's nationality).

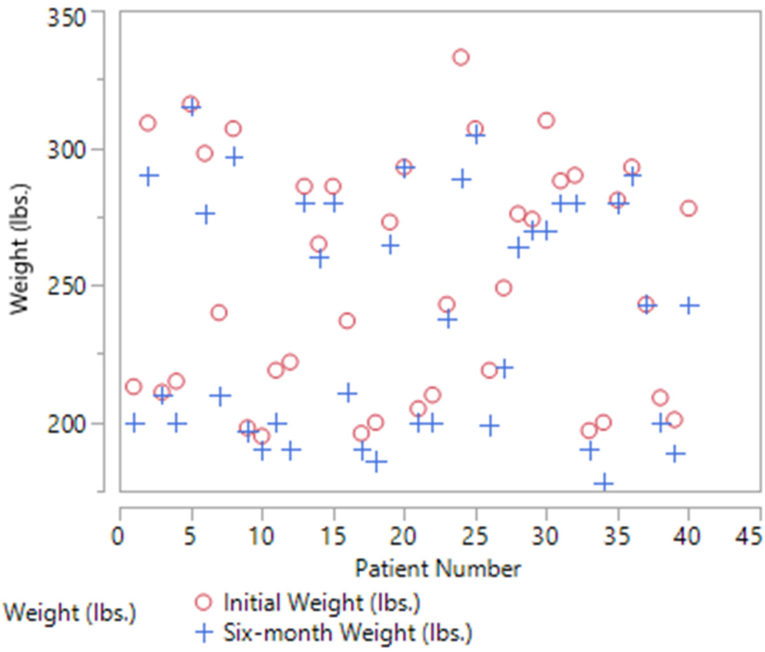


Figure 1-1. Patient's weight: (o) initial weight; (+) six-month weight.

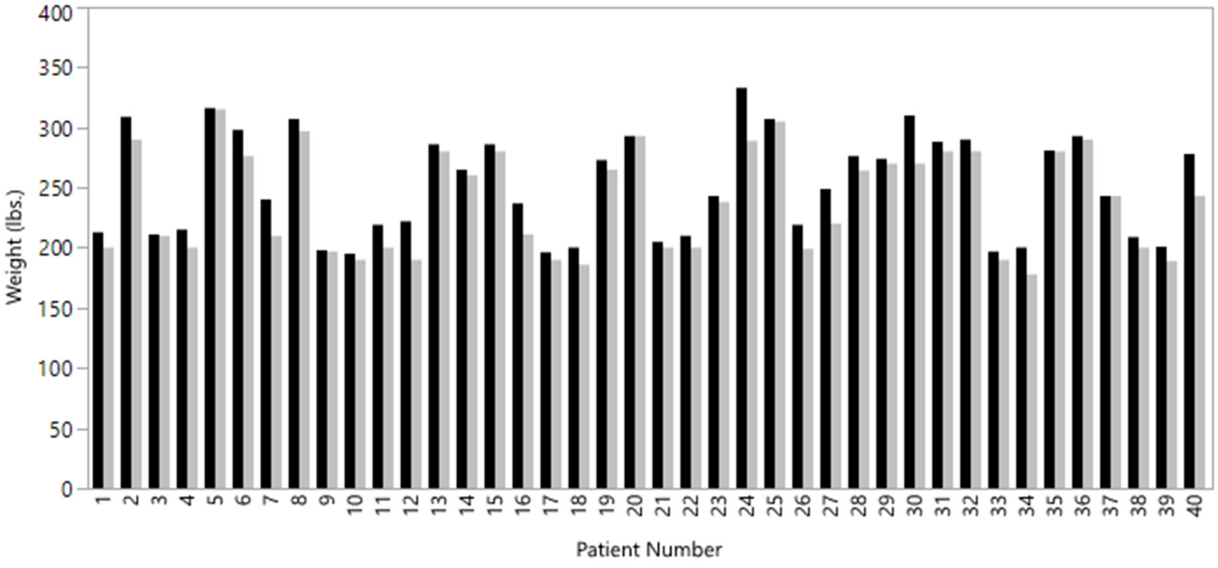


Figure 1-2. A bar graph for the initial and six-month weight is shown for all the patients in the sample. Black bars are for initial weight, and gray bars are for six-month weight.

Grouped frequency table

Raw data, such as the ones presented in Tables 1-1 and 1-2, can be summarized in a "grouped frequency table." To build this table, arrange the data in ascending order, calculate the range value, and choose the number of intervals between 5 and 20. For example, for the six-month weights data, Table 1-3 shows the grouped frequency table for the data in Table 1-2.

1. Raw data

200	315	197	280	190	200	305	270	190	243
290	276	190	260	186	200	199	270	178	200
210	210	200	280	265	238	220	280	280	189
200	297	190	211	293	289	264	280	290	243

2. Ranked data

178	186	189	190	190	190	190	197	199	200
200	200	200	200	200	210	210	211	220	238
243	243	260	264	265	270	270	276	280	280
280	280	280	289	290	290	293	297	305	315

3. The range is (315 – 178 = 137 lbs.)

4. The range of 137 is divisible by 6 (the number of intervals chosen) to give an interval width of 23 units.

Table 1-3. A grouped frequency table for the data in Table 1-2 is presented.

Six-month Weights (lbs.) (from, but less than)	**Frequency**	**Relative Frequency = Frequency/Total**	**Relative Frequency (%) = Relative Frequency x 100**	**Cumulative Frequency (%)**
178 - 201	15	0.375	37.5	37.5
201 - 224	4	0.1	10	47.5
224 - 247	3	0.075	7.5	55.0
247 - 270	3	0.075	7.5	62.5
270 - 293	11	0.275	27.5	90.0
293 - 316	4	0.1	10	100.0
Total	**40**			

If the sample is chosen by a random method, it would be assumed to represent its population. In this case, the relative frequency (%) can be read as the "*probability*" of occurrence. Thus, it can be assumed that after six months of diet, the probability that a patient who is chosen by random to have a body weight between 270 lbs. but less than 293 lbs. is 27.5%. And the likelihood that the weight of the patient is less than 293 lbs. would be 90%.

Measurement scales

Variables can have numeric or character values. The patient's name, gender, and marital status are examples of character type variables. The patient's weight and systolic blood pressure are numeric variables. Also, variables can be one of the following scales: nominal, ordinal, interval, or ratio. The latter two scales are numeric, with the interval scale having an arbitrary zero point while the ratio scale has a true zero point. These two scales are often referred to as "continuous." The ordinal scale can contain character or numeric values, and nominal variables can have only character values.

Measurement Scale	Examples
Nominal	Yes/No; Type of Cancer; Organ Systems; Drug Type
Ordinal	Likert's Scale; Satisfied/Unsatisfied;
Interval	Temperatures in degrees Celsius or Fahrenheit; Psychological Scales
Ratio	Weight; Height; Length; Concentration; Temperature in Kelvin

Investigators may choose to write ratio or interval scales as nominal or ordinal type. For instance, patient's weight in kilograms (a ratio type variable) can be written as categories (A: 150-200 kg; B: 201-250 kg) (a nominal type) instead, by asking the patients to specifically list their exact weights.

Measures for location

Even though Figure 1-2 is clearer in its representation of the data than Figure 1-1, the graph is "too busy" to draw a firm conclusion. A better way to summarize the data in Tables 1-1 and 1-2 is to use a bar graph representing the *arithmetic mean* of the weight at both time points, initial and six-month. Figure 1-3 illustrates the data using the average values.

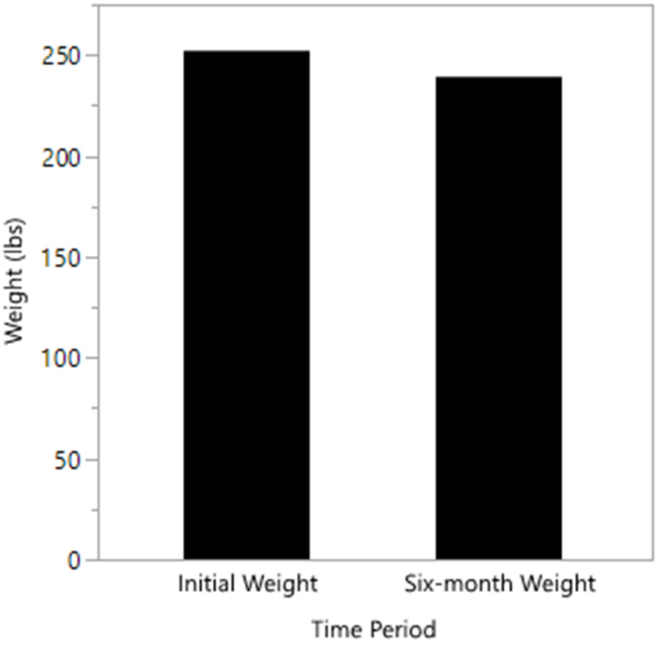


Figure 1-3. A slight decrease in the average weight was observed in 40 patients on a six-month diet.

Figure 1-3 is much more illustrative than Figures 1-1 and 1-2. The diet produced a slight decrease in weight on average following a six-month regimen. Numerically, the averages were estimated by using Equation 1-1.

$$Average = \bar{X} \text{ or } \bar{Y} = \left(\sum X_i\right)/n \quad (1\text{-}1)$$

Where X_i represents the data values and n is the total number of data points. The symbol 'Σ' stands for the "sum of." Thus, to calculate the arithmetic mean for the weights in Tables 1-1 or 1-2, the 40 weights are summed up, and then their sum is divided by their number 40. Table 1-3 shows the calculation steps.

Table 1-4. Calculation steps for the arithmetic mean.

		Results	
Step	Initial Weight (lbs.)	Six-month Weight (lbs.)	Overall/Combined for the Two Sets
$\sum X_i$	10085	9568	10085 + 9568 = 19653
N	40	40	80
Arithmetic Mean (lbs.)	10085/40 = 252.125	9568/40 = 239.2	19653/80 = 245.6625

Based on the calculations in Table 1-4, there was a (252.125 – 239.2) 12.925 lbs. weight reduction, on average, following the diet over six months. If the diet does not affect weight, the two sets can be combined. The overall arithmetic average for the two sets is 245.6625 lbs. (Table 1-4). Note that the overall average can only be calculated if the two data sets are considered homogenous (i.e., the diet did not affect weight). The "average" weight can also be expressed as a "*median"* or a "*mode*." The median value is centered in the middle of the distribution, with 50% of the values below the median value. To obtain the median for the initial weights, the values are ranked from the smallest to highest, and then the median is determined from the most centered value. The median value for the initial and six-month weights is 246 lbs. and 240.5 lbs., respectively. The mode is the most frequent value in a set of data. The mode values are given in Table 1-5. For the initial weight, multiple values occurred twice. The value of 200 lbs. was the most frequent for the six-month weight, followed by 280 and 190. The values 210, 243, 270, and 290 appeared twice each. Therefore, there were multiple mode values for the initial weight. There was a primary mode (200 lbs.), a secondary mode (280 lbs.), and a tertiary mode (190 lbs.) for the six-month weight. Note that if all the values occur with equal frequency, there will be no value for the mode.

Table 1-5. The mode value for the data in Tables 1-1 and 1-2 is found by identifying the highest frequency.

Initial Weight (lbs.)		Six-month Weight (lbs.)	
Level	Frequency	Level	Frequency
195	1	178	1
196	1	186	1
197	1	189	1
198	1	190	4
200	2	197	1
201	1	199	1
205	1	200	6
209	1	210	2
210	1	211	1
211	1	220	1
213	1	238	1
215	1	243	2
219	2	260	1
222	1	264	1
237	1	265	1
240	1	270	2
243	2	276	1
249	1	280	5
265	1	289	1
273	1	290	2
274	1	293	1
276	1	297	1
278	1	305	1
281	1	315	1
286	2	Total	40
288	1		
290	1		
293	2		
298	1		
307	2		
309	1		
310	1		
316	1		
333	1		
Total	40		

The arithmetic mean, median, and mode collectively are known as "averages." While the arithmetic mean is sensitive to extreme values, the median and mode are not. This is because every value in the sample contributes to calculating the mean. Therefore, the median is usually used to describe the average value in severely skewed distributions. Also, these three measures represent a center point in a data set. Thus, they are termed measures for central tendency or location.

Measures for dispersion

Several measures can be used to describe the variability or dispersion in a data set. Whereas measures for location are points on a scale, those for dispersion are distances on the scale. The simplest of all is known as the *range* (W). It is the algebraic difference between the largest and smallest values in the data. For the data in Table 1-1, the range is calculated from the maximum value of 333 lbs. and the minimum value of 195 lbs. (333 – 195 = 138 lbs.). The larger the value for W, the greater is the variability. The W for the six-month data is 137 lbs. (315 – 178), which is the same magnitude as that obtained from the initial weights. The range is also sensitive to extreme values existing in the data set. The sample range is considered unreliable in estimating the population range because it is highly unlikely that the population's minimum and maximum values will exist together in the same sample obtained from the population. Note that a sample is collected from the population, usually by a random method, to avoid bias. A random sample eliminates bias and allows a fair assessment of the outcome. In addition, random samples are representative of their populations which is an important feature to have in any investigation. Unless the research is epidemiological, which involves populations, academic and clinical studies rely on obtaining samples from their populations. What is learned from a sample is then applied to the entire population. Thus, a representative sample is of utmost importance to ensure applicability to the population from what is learned from the sample.

Another measure for variability is the *interquartile range* (IQR). This type of statistics belongs to a class of measures known as the *interpercentile ranges* (IPRs). A percentile is a value below which a population percentage is found. For example, 50% of the distribution lies below the median. Thus, the median is known as the 50th percentile (P_{50}). The median is also known as the second quartile, with the first quartile being P_{25} and the third quartile being P_{75}. The IQR defines the middle 50% of the distribution. It is the distance that spans from P_{25} to P_{75}. The interquartile range (P_{75} – P_{25}) for the initial weights and six-month weights is 79.25 lbs. (289.5 – 210.25) and 80 lbs. (280 – 200), respectively. To calculate other interpercentile ranges, subtract the percent from 100 and divide the result by 2. The corresponding values that delineate the distribution's upper and lower tails are the percentiles to be used in the calculation. For example, the following steps are taken to calculate the 95th interpercentile range: 100 – 95 = 5 and 5/2 = 2.5. Thus, the boundary of the upper tail of the distribution (2.5%) is $P_{97.5}$ and that for the lower tail is $P_{2.5}$. Therefore, the 95th interpercentile range is $P_{97.5}$ – $P_{2.5}$. To compute the 95% IPR for the data in Tables 1-1 and 1-2 yields 137.55 lbs. (332.575 – 195.025) and 136.55 lbs. (314.75 – 178.2), respectively. The percentiles and IPRs are insensitive to extreme values. Higher values for IPRs signify more variation in the data and less precision. Percentiles are found for an ordered sample (ascending) from Equations 1-2 and 1-3.

$$Rank\ Percentile = \left(\frac{Percentile}{100}\right)(n+1) \qquad (1\text{-}2)$$

$$Percentile = [(Y_{i+1} - Y_i)\ x\ fraction] + Y_i \qquad (1\text{-}3)$$

Equation 1-2 calculates the rank for the percentile. If the rank is a whole number, then the percentile is the corresponding value to the rank of the percentile value. When Equation 1-2 produced a rank with an integer and a fraction, Equation 1-3 is used to find the percentile (i = integer). If the rank number is greater than the highest rank or smaller than 1, the percentile is the maximum or the minimum value, respectively. For example, the following sample was taken from a capsules batch showing their weight (mg). The investigator was interested in calculating P_{75} and P_{90}.

124	120	122	120	123	125	120

The data is first arranged in ascending order:

Data	120	120	120	122	123	124	125
Rank #	1	2	3	4	5	6	7

Applying Equation 1-2,
Rank P_{75} = (75/100) (7 + 1) = 6
Thus, the value 124 mg is the third quartile.
Rank P_{90} = (90/100) (7 + 1) = 7.2
Since the rank for P_{90} is greater than the highest rank (7), then P_{90} = 125 mg.

Perhaps the most used measure for variability is the *standard deviation* (SD). The symbol of the standard deviation of a population is σ. The SD refers to the "*average distance*" that each value in the distribution has from the mean. Similar to the range, the SD is sensitive to extreme values. The equation to calculate the sample's standard deviation is:

$$[\sum(X_i - \bar{X})^2/(n-1)]^{0.5} \qquad (1\text{-}4)$$

The units for the SD are those of the data points. To calculate σ, Equation 1-4 is used; however, instead of using (n – 1) as a denominator, the population size N is used. The SD for the initial weight and six-month weight is 42.78 lbs. and 42.83 lbs., respectively. More variation in the distribution is reflected by a higher value for the SD. A closely related measure of variability to SD is the *variance* (SD^2). The units for the variance are those of the data points squared. The variance values for the data in Table 1-1 and 1-2 are 1830.32 $lbs.^2$ and 1834.22 $lbs.^2$, respectively. Both the SD and the variance are sensitive to extreme values. For the clinical variables encountered in medicine, the clinically acceptable normal values for the variable are found within the middle 95% of the distribution. Values outside this range are clinically labeled "abnormal." The middle 95% of the distribution is found in the range (μ ± 1.96 σ). For example, the normal serum sodium concentration is in the range of 136-145 mEq/L. This represents the middle 95% of the serum sodium level in normal adults. If the mean serum sodium concentration is 140 mEq/L, then,

$$\mu + 1.96\ \sigma = 145$$

$$140 + 1.96\ \sigma = 145 \text{ or } \sigma = 2.55 \text{ mEq/L}$$

The *coefficient of variation* (%CV) (also known as relative standard deviation; %RSD) is another measure of dispersion. It is used to standardize the variability to be compared among different groups or within the same group, as when comparing two variables for their dispersion profile. The equation for calculating (%CV) is:

$$(SD/Mean)*100 \tag{1-5}$$

For example, the 40 patients who attended the weight loss clinic had their blood pressure measured during their initial visit to the clinic. Table 1-6 shows the systolic blood pressure readings (mmHg) of the 40 patients.

Table 1-6. Systolic blood pressure readings of 40 patients from the weight loss clinic are recorded in mmHg.

145	165	140	160	150	158	160	162	142	125
148	153	167	180	128	144	175	180	140	165
154	130	160	166	178	120	127	135	170	130
165	175	145	175	150	160	145	167	122	120

The mean and the SD for systolic blood pressure readings are 152.025 mmHg and 17.88 mmHg, respectively. The %CV for systolic blood pressure was 11.76%. For the same patients, the %CV for their weight was 16.97%. Therefore, it can be stated that the patients were more variable with respect to their weight than with respect to their systolic blood pressure.

The systolic blood pressure of the 40 patients from the weight loss clinic was compared to that of 12 patients attending a weight loss program at a neighboring community hospital (Table 1-7). The calculated %CV for the 12 hospital patients was 16.03%. Therefore, it can be concluded that patients attending the weight loss clinic were less variable with respect to systolic blood pressure than those at the hospital.

Table 1-7. Systolic blood pressure readings of 12 patients from a community hospital are recorded in mmHg.

120	125	143	120
176	128	122	180
155	120	150	165

The *median absolute difference (or deviation)* (MAD) is another measure of variability. An example for calculating MAD follows:

Data points: 3, 12, 5, 9, 7 seconds

1) Arrange the data in ascending order and find the median value:

	3	5	7	9	12
Rank	1	2	3	4	5

Median = 7 seconds

2) Calculate the absolute differences from the median;
$|3 - 7| = 4$
$|5 - 7| = 2$
$|7 - 7| = 0$
$|9 - 7| = 2$
$|12 - 7| = 5$

3) Arrange the absolute differences in ascending order:

	0	2	2	4	5
Rank	1	2	3	4	5

MAD = 2 seconds

Measures for location and dispersion from a grouped frequency table

When grouped frequency table is formed, specific information pertaining to the raw data set is lost and replaced by gaining information on trends in the data. Calculating measures for location and dispersion from a grouped frequency table is possible; however, the statistics calculated are only approximate.

To find the grouped mean, Equation 1-6 is used.

$$Grouped\ Mean = [(m_i x\ f_i)]/n \qquad (1\text{-}6)$$

Where m_i and f_i are the midpoint and frequency for the ith interval, respectively, the midpoint for any interval is calculated by taking the arithmetic mean of the interval's two limits. Once the first interval's midpoint is calculated, the remaining midpoints can be calculated by adding the interval width to each value to get the subsequent value.

The midpoints for the data in Table 1-3 are presented below.

Six-month Weights (lbs.) (from, but less than)	**Frequency**	**Midpoints (m_i)**
178 - 201	15	(178+201)/2 = 189.5
201 - 224	4	189.5 + 23 = 212.5
224 - 247	3	235.5
247 - 270	3	258.5
270 - 293	11	281.5
293 - 316	4	304.5
Total	**40**	

Grouped Mean = [(189.5 x 15) + (212.5 x 4) + … + (304.5 x 4)]/40 = 9489/40 = 237.225 lbs.

The grouped modal interval corresponds to the highest frequency. The data in Table 1-3 has a modal interval of (178-201) lbs.

The grouped median is found by locating the most centered data point. This is accomplished by:

1. Calculate the location of the median: [(n/2) + 0.5] = [(40/2) + 0.5] = 20.5th
2. Determine the interval where the median is located: 20.5 – (15 + 4) = 1.5th; the median is located in the third interval, and it is observation 1.5th in that interval.
3. Estimate the median value: The third interval has a width of 23 units and a frequency of 3. Thus, 3 observations occupy a width of 23 units. By proportion, observation 1.5th is located (1.5 x 23/3) 11.5 units away from the beginning of the third interval. Therefore, the median value is (224 + 11.5) 235.5 lbs.

The grouped standard deviation is calculated from Equation 1-7.

$$Grouped\ SD = \{[(m_i - \ grouped\ mean)^2 x\ f_i]/(n-1)\}^{0.5} \qquad (1\text{-}7)$$

$Grouped\ SD = \{[(189.5 - \ 237.225)^2\ x\ 15 + (212.5 - \ 237.225)^2\ x\ 4 + \cdots + \ (304.5 - \ 237.225)^2\ x\ 4]/(40 - 1)\}^{0.5}$

Grouped SD = $(77643.975/39)^{0.5}$ = 44.62 lbs.

Grouped variance = $(\text{Grouped SD})^2 = (44.62)^2$ = 1990.87 lbs.^{2}

Grouped range ≈ 4 x (Grouped SD) = 4 x 44.62 = 178.48 lbs.

Contrasting the actual statistics with the grouped statistics, the values obtained from the grouped frequency table are within an acceptable range (4-6%).

Statistics	Actual	Grouped	Difference (%)
Arithmetic mean	252.125	237.225	-5.90
Median	246	235.5	-4.27
SD	42.78	44.62	4.30

The box plot

This graph is also known as the *box and whiskers* plot. A box plot is used to identify various statistics on the graph. These are the IQR, $P_{25,}$ P_{50}, P_{75}, the range, and the presence of outliers. There are two types of outlying values; those that exist beyond [(3 x IQR) + P_{75}] and [P_{25} – (3 x IQR)] are known as extreme outlying values. There are those which exist above [(1.5 x IQR) + P_{75})] to and including the value equals [(3 x IQR) + P_{75}] or below [P_{25} – (1.5 x IQR)] to and including [P_{25} – (3 x IQR)] are known as minor outliers. The "whiskers" on both sides of the box extend to the largest and smallest values that are not outlying values. Box plots are useful for comparing groups with their dispersion when placed side by side on the same graph (Figure 1-4).

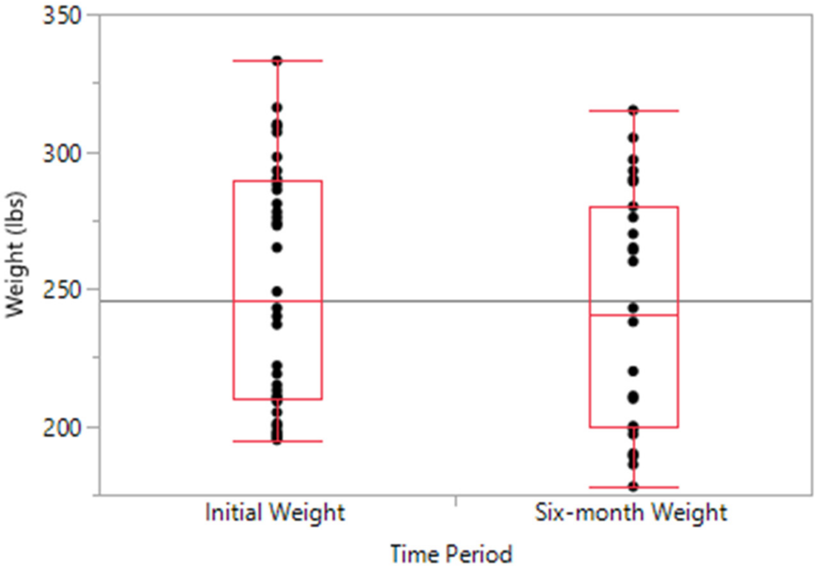


Figure 1-4. A box plot for patients' weight (lbs.) is shown.

Based on the properties of the normal distribution (Lecture Two), a relationship exists between the range (W) and the standard deviation. *Approximately*, the SD equals one-fourth of the range value.

$$SD \approx \left(\frac{1}{4}\right) W \qquad (1\text{-}8)$$

For example, the approximate SD for the initial weights (lbs.) is [(1/4) x (333 – 195)] = 35 lbs. (the actual SD value for the initial weights (lbs.) is 42.78 lbs.).

Incidentally, JMP® Pro Statistical Discovery Software generates box plot along with other descriptive statistics. Take for instance the following data for two groups:

Group	Observations
1	21
1	120
1	220
1	320
1	420
1	470
1	480
1	490
1	500
1	505
1	510
1	1000
2	9
2	45
2	440
2	670
2	870
2	880

2	890
2	900
2	902
2	905
2	910
2	2000

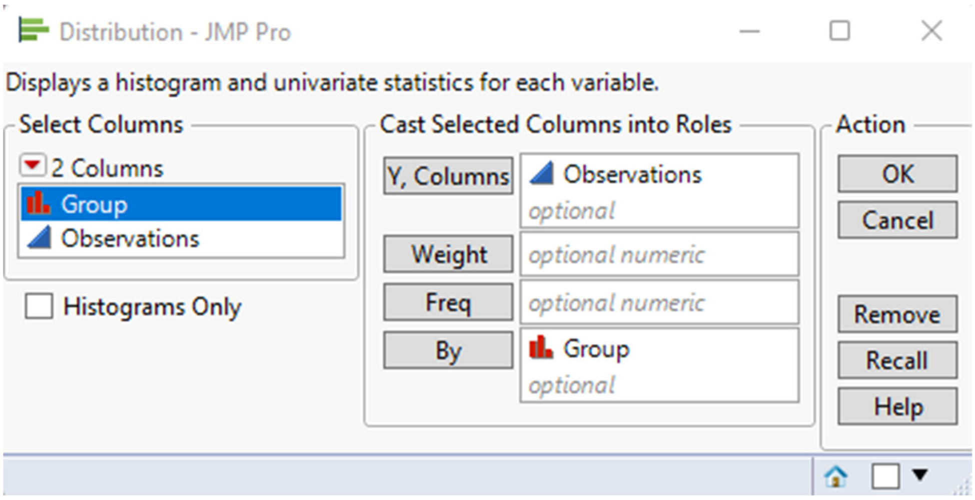
Using JMP® Pro Statistical Discovery Software for the analysis,

Analyze → Distribution

Select Observations

By → Group

OK



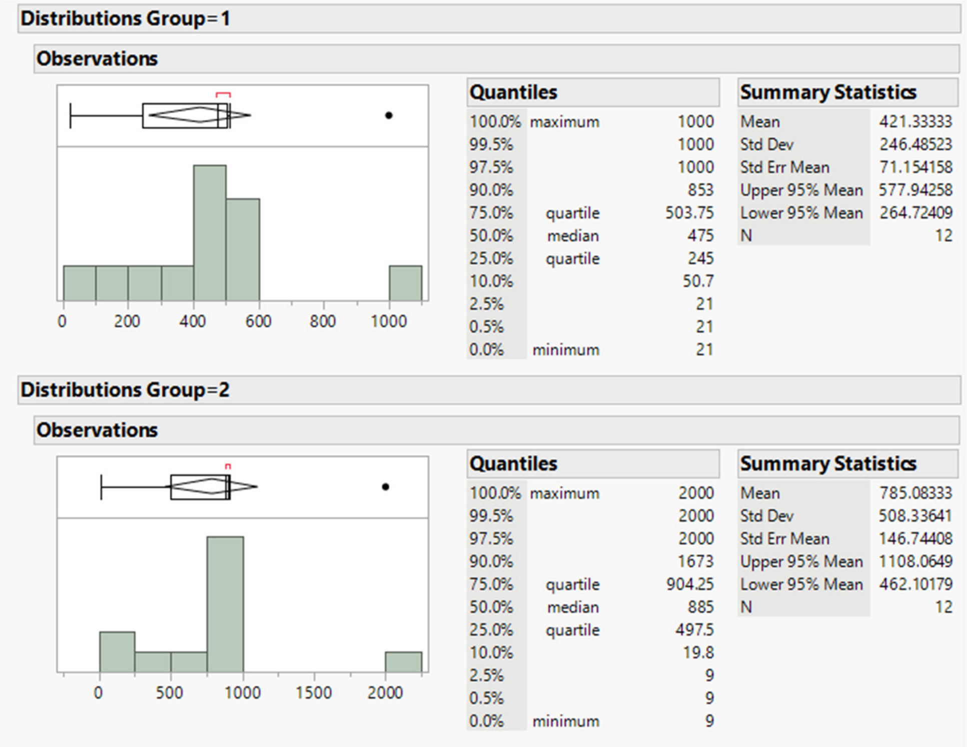
Distribution - JMP Pro

Displays a histogram and univariate statistics for each variable.

Y, Columns: Observations

By: Group

Two box plots (along with histograms and descriptive statistics) are generated.



Distributions Group=1

Observations

Quantiles

100.0%	maximum	1000
99.5%		1000
97.5%		1000
90.0%		853
75.0%	quartile	503.75
50.0%	median	475
25.0%	quartile	245
10.0%		50.7
2.5%		21
0.5%		21
0.0%	minimum	21

Summary Statistics

Mean	421.33333
Std Dev	246.48523
Std Err Mean	71.154158
Upper 95% Mean	577.94258
Lower 95% Mean	264.72409
N	12

Distributions Group=2

Observations

Quantiles

100.0%	maximum	2000
99.5%		2000
97.5%		2000
90.0%		1673
75.0%	quartile	904.25
50.0%	median	885
25.0%	quartile	497.5
10.0%		19.8
2.5%		9
0.5%		9
0.0%	minimum	9

Summary Statistics

Mean	785.08333
Std Dev	508.33641
Std Err Mean	146.74408
Upper 95% Mean	1108.0649
Lower 95% Mean	462.10179
N	12

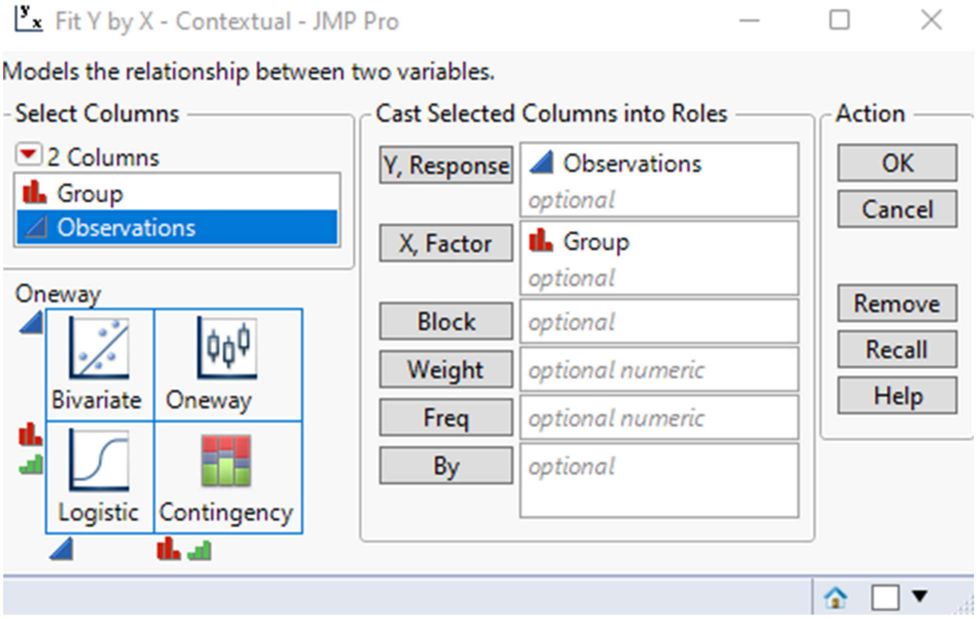
Another way to do this box plot formation is to plot the groups side by side on the same graph.

Analyze → Fit Y by X

Group → X, Factor

Observations → Y, Response

OK



Fit Y by X - Contextual - JMP Pro

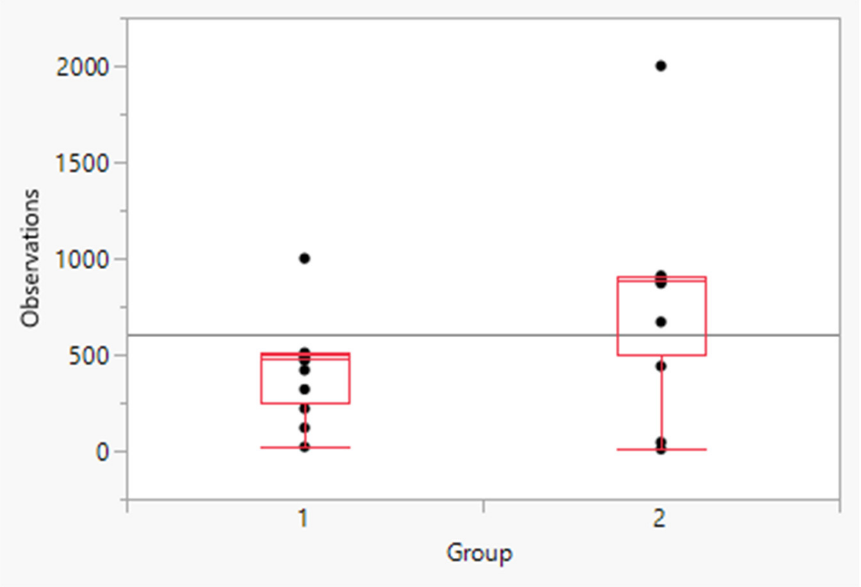
Models the relationship between two variables.

Select Columns: 2 Columns — Group, Observations

Cast Selected Columns into Roles:

- Y, Response: Observations
- X, Factor: Group
- Block: optional
- Weight: optional numeric
- Freq: optional numeric
- By: optional

Click on the red triangle adjacent to “Oneway Analysis of Observations By Group,” and select Means/Anova/Pooled t-test. Select from the drop-down menu (red triangle) and hover with the computer mouse over Display Options, from the side menu choose Box Plots (and deselect Mean Diamonds).



From the box plot above, both groups have outlying values located beyond the whiskers. Note the location of the three quartiles by reading their values from the y-axis. It is apparent that Group 2 has a higher value for the median and a greater variability.

Skewness and kurtosis

A distribution can be symmetrical or skewed. The skewness can be positive or negative. Distributions with arithmetic mean equals the median value (or the mode) are symmetrical (skewness = 0). Those distributions with their arithmetic mean greater than the median are positively skewed; if the median is greater than the arithmetic mean, then the distribution is negatively skewed. Skewness can also be severe. This approach for evaluating skewness is credited to Karl Pearson.

$$Pearson's\ Skewness\ Coefficient = (Mean - Mode)/SD \qquad (1\text{-}9)$$
$$Pearson's\ Skewness\ Coefficient = [3\ x\ (Mean - Median)]/SD \qquad (1\text{-}10)$$
$$Pearson's\ Skewness\ Coefficient = [(Mean - Median)]/SD \qquad (1\text{-}11)$$

Equation 1-11 is commonly used for its simplicity. If the value obtained from Equation 1-11 exceeds ±0.2, the distribution is considered to have a severe skewness.

Kurtosis is a measure of "peakedness" or how heavy the distribution's tails are. A flat curve has a negative kurtosis, while a peaked or pointed curve has a positive kurtosis. If the value for kurtosis is zero (0), then the distribution does not exhibit kurtosis. The greater the deviation of kurtosis or skewness away from the value zero, the greater the degree of kurtosis or skewness.

The averages are ranked in ascending order for negatively skewed distributions as mean < median < mode. When the distribution is positively skewed, the reverse is true (mode < median < mean) (Figure 1-5).

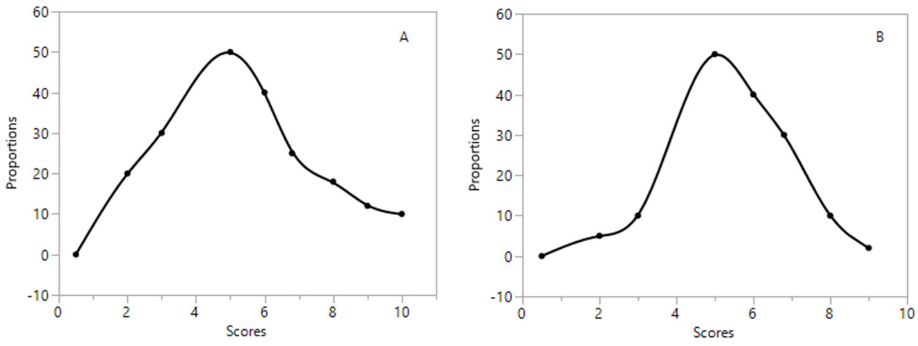
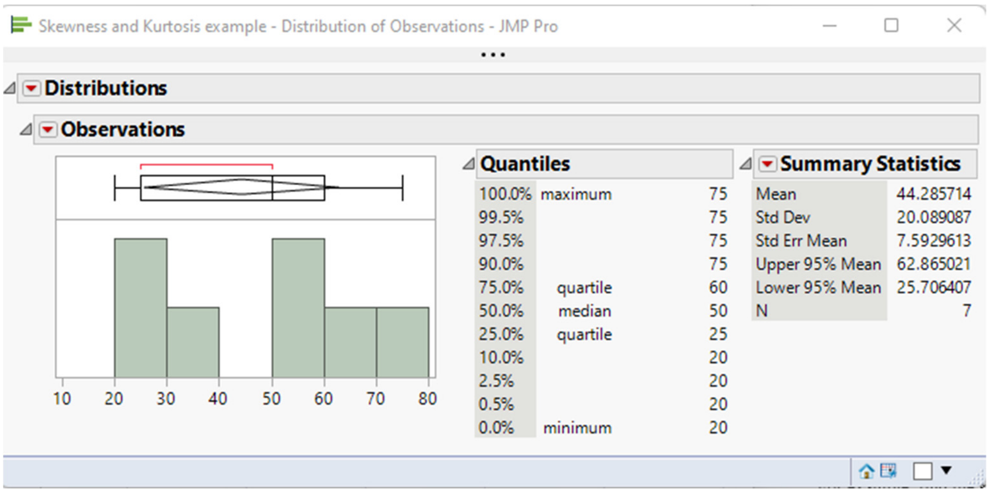


Figure 1-5. The graphs show positive skewness (A) and negative skewness (B).

Both skewness and kurtosis can be requested from JMP® Pro Statistical Discovery Software from the descriptive statistics utility.

For example, find the skewness and kurtosis for the following data:

50 30 25 50 20 60 75



Skewness and Kurtosis example - Distribution of Observations - JMP Pro

Distributions

Observations

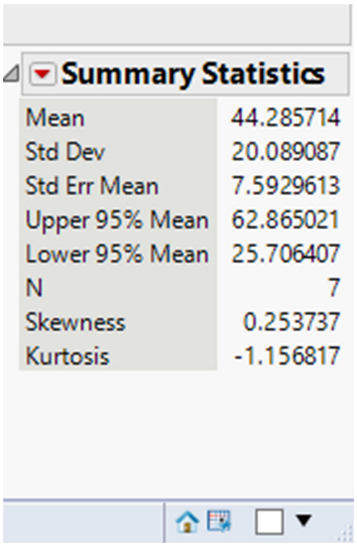
Quantiles

100.0%	maximum	75
99.5%		75
97.5%		75
90.0%		75
75.0%	quartile	60
50.0%	median	50
25.0%	quartile	25
10.0%		20
2.5%		20
0.5%		20
0.0%	minimum	20

Summary Statistics

Mean	44.285714
Std Dev	20.089087
Std Err Mean	7.5929613
Upper 95% Mean	62.865021
Lower 95% Mean	25.706407
N	7

Click on the red triangle next to Summary Statistics. Select Customize Summary Statistics. Check the boxes next to Skewness and Kurtosis. These are added to the report above.



Summary Statistics	
Mean	44.285714
Std Dev	20.089087
Std Err Mean	7.5929613
Upper 95% Mean	62.865021
Lower 95% Mean	25.706407
N	7
Skewness	0.253737
Kurtosis	-1.156817

This data set has a positive skewness (the tail of the distribution is pointing toward the right) and a negative kurtosis (a flat distribution).

Handling outlying values

Outlying values are often encountered in real data that are collected empirically. It behooves a researcher to ask, "Why did they occur?" before trying to discard them. The answer to that question might lead to discoveries. However, if the decision has been reached to discard them, their removal from the data must follow a systematic way that ensures consistency. To that end, it is possible in large enough sets of data that investigators discard the observations occurring at both extremes of the distribution (the top and bottom 5%). For a small number of observations (between 3 to 10 observations), the Q test is applicable if, and only if, one datum is an outlying value in a data set, *and* it is unique. To run the Q test, the data must be arranged and ranked, the range value is determined, and the distance between the outlying value and that next to it in magnitude is established (d). For example, the following data points are collected for aspirin concentration in a bottle of tablets (labeled as 325 mg Aspirin/tablet).

1. Raw data

324	320	325	340	324	325	322

2. Ranked data

320	322	324	324	325	325	340

3. The value of 340 mg appears to be an outlier. The distance between 340 mg and 325 mg is d = 15 mg.
4. The range = W = 340 – 320 = 20 mg
5. Q-calculated = d/W = 15/20 = 0.75
6. Q-table = 0.568 (for n = 7)
7. Since the value calculated exceeded the tabulated value, the datum 340 mg can be discarded from the data set (with 95% confidence; at α = 0.05) (Lecture Two).

The presence of outlying values may be detected by graphing a box plot and then observing the outlying values located beyond the whiskers. Another way is to run an outlier's analysis on the data using JMP® Pro Statistical Discovery Software.

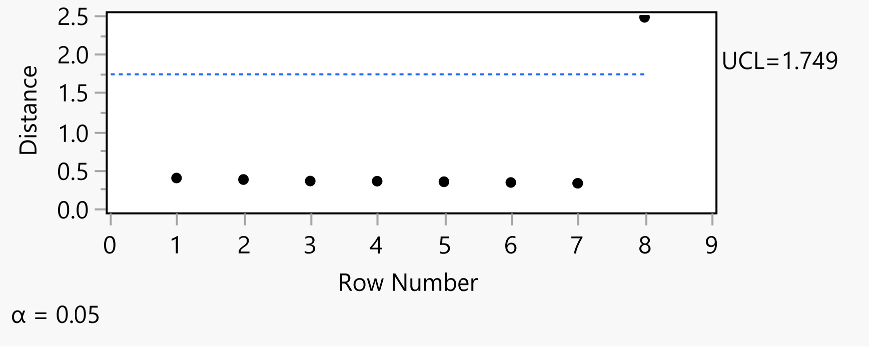
For example, consider the following set of observations:

180 190 200 201 205 210 215 1670

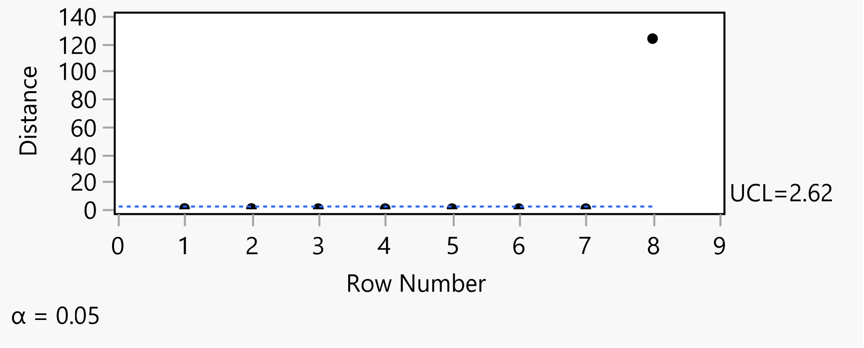
Analyze → Multivariate Methods → Multivariate → Red Triangle (pull-down menu) → Outlier Analysis

Outlier Analysis

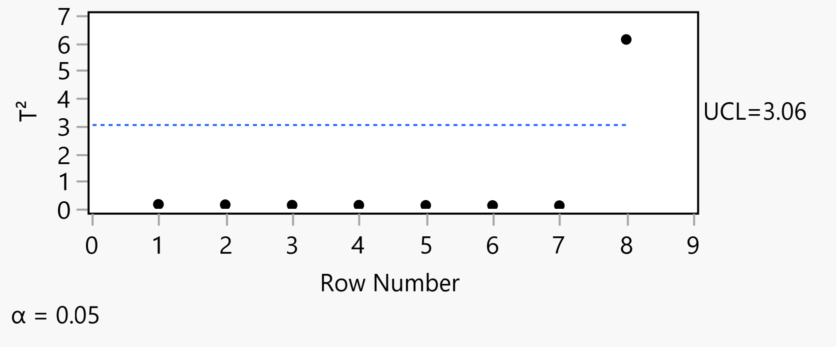
Mahalanobis Distances



Jackknife Distances



T^2



Values that are found beyond the UCL (upper control limit) are considered outliers (see Lecture Nine).

Precision, accuracy, and bias

A precise measurement has low variability associated with it. Thus, precision is evaluated by a measure for dispersion. The customarily measure that is used for precision is %CV. On the other hand, accuracy is assessed by how far a measurement is from a "true" value. A biased set of measurements is when *all* its values are located below or above a true value.

Consider the following data sets for analyzing vitamin C packets, each labeled to contain 500 mg of vitamin C (the true value). The amount of vitamin C per packet was determined using two analytical methods.

	Vitamin C (mg)

Packet #	Analytical Method 1	Analytical Method 2
1	501	499
2	498	496
3	504	498
4	495	490
5	508	499
%CV	1.01%	0.76%
Arithmetic Mean	501.2	496.4

Analytical method 2 is more precise (based on a lower value for %CV). However, analytical method 1 is more accurate because the average value for vitamin C is closer to the true value of 500 mg. In addition, analytical method 2 is biased because each value in the sample was below the true value.

Useful rules

There are a couple of rules to consider when modifying the values in a data set by a constant.

Rule one: Adding or subtracting values to or from a constant.

Modifying values by adding or subtracting a constant from every value in the data set will affect the mean value, but the standard deviation value remains unchanged. For example, consider the following data set: 30, 60, 21, 74, and 80 mg. The arithmetic mean of the set is 53 mg with a standard deviation value of 26.32 mg. If each value increases by 2 mg, the mean becomes 55 mg, but the SD value remains the same. Note that while the range value does not change the value of % CV changes accordingly. The %CV for the original and modified set is 49.66% and 47.85%, respectively.

Rule two: Multiplying or dividing by a constant.

If the values of a data set are modified so that each is multiplied or divided by a constant, both the arithmetic mean, and the standard deviation will change accordingly. For example, the 30, 60, 21, 74, and 80 mg data set was modified by dividing each original value by 5. The resulting data set was 6, 12, 4.2, 14.8, and 16 mg. As a result, the new data set will have a mean of 10.6 mg and a standard deviation of 5.264 mg. The value of the range also changes accordingly, but not that of %CV, which remains the same.

Practice examples

1. The %CV for the set [90, 87, 65, 44, 33] is ___.

   A. 23.11%
   B. 31.08%
   C. 49.80%
   D. 39.71%

   Answer: The arithmetic mean = 63.8; SD = 25.33; %CV = (25.33/63.8) x 100 = 39.71%.

2. Find the variance of the following data set:

   78, 92, 45, 81, 55, 105

   A. 505.6
   B. 4.7
   C. 130.2
   D. 436.4

   Answer: SD = 22.49; SD^2 = Variance = 505.6

3. Consider the following data set:

3, 5, 7, 9, 12, 55

Calculate the arithmetic mean value.

A. 11.73
B. 12.82
C. 15.17
D. 29.95

Answer: Using Equation 1-1,

$Average = \bar{X} = (\sum X_i)/n$

$\bar{X} = 15.17$

4. If the mean is 200 units and the median is 210 units, it is most likely that the distribution is positively skewed.

A. True
B. False

Answer: Since the median is larger than the mean, this distribution is negatively skewed.

5. If no minor outlying values are found in a sample, the sample could not contain extreme outlying values.

A. True
B. False

Answer: Extreme outlying values can be present regardless of whether or not minor outliers exist.

6. Consider the following data for the weight of tablets:

264.8, 244.1, 279.4, 233.6, 247.9, 219.9, 249.1, 225.6, 227.4 mg

The first quartile = ___.

A. 226.5 mg
B. 230.5 mg
C. 244.1 mg
D. 256.95 mg

Answer: The first quartile is P_{25}. The data is first rank ordered in an ascending fashion.

Data	219.9	225.6	227.4	233.6	244.1	247.9	264.8	279.4
Rank #	1	2	3	4	5	6	7	8

Applying Equation 1-2: Rank P_{25} = (25/100) (8 + 1) = 2.25

Applying Equation 1-3: P_{25} = [(227.4 – 225.6) x 0.25] + 225.6 = 226.05 mg

7. Consider the following data for the weight of tablets:

264.8, 244.1, 279.4, 233.6, 247.9, 219.9, 249.1, 225.6, 227.4 mg

The data belong to the ____ scale.

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Answer: The ratio scale.