

An Introduction to Statistics, pdf⁵

by **Keith G. Calkins, Ph.D.**

with assistance by

Shirleen Luttrell

Fall 2010

Berrien County Math & Science Center

Andrews University

Berrien Springs, MI 49104-0140

TABLE OF CONTENTS

Preface	ix
0.1 Statistics Project	x
0.2 Statistics Project Rubric	xi
1 Definitions, Uses, Data Types/Meas.	1
1.1 Italian Father of Modern Science: Galileo Galilei	1
1.2 The Basis of Science: The Scientific Method	2
1.3 Introduction to Statistics	3
1.4 General Terms Used Throughout Statistics	4
1.5 Accuracy vs. Precision	5
1.6 Uses and Abuses of Statistics	5
1.7 Types of Data	5
1.8 Levels of Measurement	7
1.9 Homework, Stat's Introduction	9
2 How & Why of Stat. Sampling	11
2.1 Danish Father of Modern Astronomy: Tycho Brahe	11
2.2 Points to Consider	12
2.3 Methods of Sampling	13
2.4 Sampling Error	15
2.5 Question Types	15
2.6 Record Keeping	15
2.7 Experimental Design	16
2.8 Homework, Statistical Sampling	17
3 Ave.: Mean, Mode, Median, or Midrange?	19
3.1 The Father of Thermodynamics: Lord Kelvin	19
3.2 Averages	20
3.3 The Best Average	21
3.4 Round-off Rules	22
3.5 Examples	23
3.6 Activity: Calculator Averages	23

3.7	Homework, Averages	25
4	What Does He Mean?	27
4.1	English Father of Modern Science: Francis Bacon	27
4.2	Arithmetic Mean	28
4.3	Geometric Mean	28
4.4	Harmonic Mean	29
4.5	Quadratic Mean	29
4.6	Trimmed Mean	30
4.7	Weighted Mean	30
4.8	Combination Mean	31
4.9	Means from a Frequency Table	31
4.10	Activity: Exponents for Geometric Mean	31
4.11	Homework, Means	33
5	Measures of Dispersion	35
5.1	The Father of Mathematical Modelling: Siméon-Denis Poisson	35
5.2	Range	36
5.3	Standard Deviation	36
5.4	Variance	37
5.5	Range Rule of Thumb	38
5.6	More Round-off Information	38
5.7	Activity: Frequency Means/Standard Deviation	39
5.8	Homework, Dispersion	41
6	Bell-sh, Normal, Gaussian Distribution	43
6.1	The Father of Russian Mathematics: Chebyshev	43
6.2	The Bell-shaped, Normal, Gaussian Distribution	44
6.3	The Empirical Rule	45
6.4	Chebyshev's Theorem	46
6.5	Meanings of Normal	46
6.6	Other Distributions	47
6.7	Quiz over Statistics Lesson 5	48
6.8	Homework, Normal Curve	49
7	Measurements of Position	51
7.1	Father of Statistical Genetics: Sir R. A. Fisher	51
7.2	Standard or z -Scores	51
7.3	Ordinary or Unusual Scores	53
7.4	Quartiles	53
7.5	Hinges; Mild and Extreme Outliers	54
7.6	Deciles	55

7.7	Percentiles	55
7.8	5-Number Summary	56
7.9	Homework, Measures of Position	57
8	Summarizing and Presenting Data	59
8.1	The Father of Exp. Data Analysis: John Tukey	59
8.2	Frequency Tables	59
8.3	Histograms	61
8.4	Pie Charts, Pictographs, <i>etc.</i>	62
8.5	Exploratory Data Analysis	63
8.6	Stem-and-Leaf Diagrams	63
8.7	Boxplots	64
8.8	Homework, Presenting Data	65
9	The Student t Distribution	67
9.1	The Father of the t Distribution: William Gosset	67
9.2	Hypothesis Testing	68
9.3	Type I and Type II Errors	68
9.4	Computing a Test Statistic	70
9.5	Making a decision about H_0	70
9.6	The Student t Distribution	70
9.7	Degrees of Freedom, Confidence Intervals	72
9.8	Table of t Values	73
9.9	Homework, t Distribution	75
10	Chi Squared (χ^2) Goodness of Fit	77
10.1	The Father of Math. Statistics: Karl Pearson	77
10.2	Chi Squared Distributions and Tests	78
10.3	A Chi Squared Distribution Table	79
10.4	Other Applications	81
10.5	Don't Abuse Tests of Significance	82
10.6	Conclusion/Errata	82
10.7	Homework, Hypothesis Testing (χ^2)	83
10.8	Summary sheet for χ^2 Activity	85
10.9	Activity to Verify Book Before Stapling	87
A	Odd Solutions and Released Tests/Keys	89
A.1	Odd Homework Answers	90
A.2	Released Statistics Test: Oct. 19, 2001	95
A.3	Key for Released Statistics Test: Oct. 19, 2001	99

List of Figures

3.1	TI-84 Data, Statistics, 5 Number Summary.	24
4.1	Sine Curve—like US Household Current/Voltage.	30
5.1	TI-84 Graphing Calculator Display of Frequency Data.	39
5.2	TI-84 Graphing Calculator Display of 1 Variable Statistics.	40
6.1	The Standard Normal Dist. and Normally Dist. IQs	44
6.2	Data Within 1σ (left) and 2σ (right).	45
6.3	Fictitious Salary Data Illustrating Use of Frequency.	49
7.1	Fictitious Salary Data Illustrating Use of Frequency.	57
8.1	Frequency Table of Center Students by Grade Level.	60
8.2	Frequency Table of 1998 Algebra Diagnostic Test Scores.	60
8.3	TI-84 Bar Chart and Settings.	62
8.4	Stem and Leaf Diagram for Presidential Inaugural Data.	63
8.5	Stem & Leaf Diagram for Presidential Inaugural Data—Split Stems.	64
8.6	TI-84 Box-plot and Settings.	64
9.1	False Negatives/Positives and Other Names.	69
9.2	Student- t Distribution Graphs.	71
9.3	Student- t Critical Values for Various Alphas and Degrees of Freedom.	73
10.1	Chi-squared Goodness of Fit for 200 Penny Flips.	79
10.2	Table of Critical χ^2 Values for various α 's and Degrees of Freedom.	79
10.3	Graphs of the χ^2 Distribution for various Degrees of Freedom.	80
10.4	Chi ² Goodness of Fit for 192 Rolls of a Die.	80
10.5	Frequencies for Students Indicating Subject Most in Need of Change.	81
10.6	Chi-squared Goodness of Fit for 1000 Die Rolls—Real.	83
10.7	Chi-squared Goodness of Fit for 1000 Dies Rolls—Faked.	83
10.8	Chi-squared Goodness of Fit for M&M Data.	84
10.9	Collection Point for χ^2 M&M Data.	86

Preface

There are three kinds of lies: lies, damned lies, and statistics. Mark Twain*

An Introduction to Statistics, resulted in large part from the expansion of the Berrien County Math and Science Center at Andrews University from 30 students per grade level to 50 students per grade level which we endured from 1997 until 2004. The two simultaneous sections of necessity had different teachers. Fairness issues in material covered and tested and the infamous Thomas rules for “team teaching” (common gradebook, common syllabus, common tests—cowritten) exerted a strong influence.

The various editions (4th through 6th) of the college textbook by Triola which we previously used for only a few weeks in the fall had too much of a murder and mayhem slant. Being a Math and Science Center, more science and math rather than social examples seemed desirable. Teaching means and standard deviations which involve fractions and square roots seemed best preceded by at least a review of such concepts.

These statistics lessons serve as the basis for further usage by our grade nine students in their Arts and Science EXPO practice project in the fall and for the real EXPO/ISEF project in the spring. These EXPO projects have been run intermittently under the ISEF rules since 1994–95 so original research is required. This has necessitated the inclusion of two statistical tests, the Student t -test and Chi-square Goodness of Fit. The formal background needed for a theoretical understanding is deferred, however, until their sophomore year. A tenth of each semester examination for grades nine and ten has been over Descriptive Statistics. Thus, we have encouraged the students to retain the textbooks for reference all year. They can then be reluctant to give them up in the spring, or even at graduation, and thus subsequently occupying an honored place of reference in their college dorm room.

Student placement and acceleration remain concerns which in recent years has been complicated by the move of statistics from the fall into the spring for center sophomores and juniors. Starting the school year with statistics and review can bond the student with their TI-series graphing calculator (TI-84 mode for the TI-*n*spire)

*Twain notes attribution to Disraeli in his autobiography, but the concept is older. See <http://www.york.ac.uk/depts/maths/histstat/lies.htm> for a detailed history.

is the current recommendation for freshmen through juniors and the TI-89 titanium for seniors). Students joining as sophomores or juniors present extra challenges.

More college students are required to take Statistics than Calculus, yet Calculus remains the focus of our math curriculum. About half of our graduates choose majors outside the fields of mathematics, science, computers, and engineering. Students whose algebra skills are still being developed during Calculus are perhaps less motivated than optimum. AP Statistics does not fit within the Geometry, Algebra II, Precalculus prerequisites. We have been able to offer/support AP Statistics to/for some individual students ((Matt S.), Mike P., (Eric W.)) and four groups of accelerated students (12 juniors, 1999–2000; 10 sophomores, 2004–05; 17: sophomores (12), juniors (2), seniors (3), 2007–08; 15: sophomores (8), juniors (6), senior (1), 2010–11) In summary, we hope to cover about half the AP Statistics curriculum for all our students but spread over their freshmen through junior years. These 10 lessons are followed by 15 lessons their sophomore year in Probability and Distributions. The junior component on Hypothesis Testing remains ill-formed, however, at this time.

Mathematics on the web has been slow to develop. As a TeX user since 1988, I've been rather disappointed with my options. In 1995, I thought Windows 95 and the subsequent explosion of the World Wide Web would allow XML to eclipse HTML as a way to format page content. We started coding this in HTML anyway in the summer of 1998. Meanwhile, XML lagged and Adobe's PDF took a strong hold. Thus a major conversion was done during the 2006–07 school-year to convert these lessons from HTML to PDF via LaTeX, a derivative of TeX, thus retaining online access while permitting proper formatting of the material.

2007–08 saw quotes added and the biographies finished. Acetate answers were replaced with powerpoint style pdfs. In 2008–10 we concentrated on polishing things—fewer warts remain.

I offer my thanks to Sally Adkin, founder/director, and Mr. Lundgren, subsequent director of the Berrien County Math and Science Center. Innumerable internet users who came upon these lessons via a search engine have contributed in various ways. Roberto Ordóñez (1997–98), and Aurora Burdick (1998–2000) were involved in various roles. A special thanks goes to Shirleen Luttrell (1998–2007) without whom these lecture notes would never have gotten off the ground, nor flourished.

Berrien County Math and Science Center

your title

A Statistics Project
Presented in Partial Fulfillment
of the Requirements for
Integrated Geometry

by

your name

October 28, 2010

0.1 Statistics Project

Collect at least 15 newspaper/magazine clippings or articles that use descriptive statistics as a way of displaying or relating information or data (**15 points**). Include and label several which are deceptive or exaggerate discrepancies (**10 points**).

Write a few sentences describing the type of graphical representation, its meaning (**15 points**), and your opinion of its validity (**5 points**).

Use clippings/articles which are as current as practical from a variety of sources and on a variety of topics. Include enough information so that the item could be relocated (magazine name, date, page number, *etc.*) (**30 points**). The Internet may be used for SOME.

Be neat, organized, (**15 points**) and use a cover sheet (**10 points**) with the information on the back of this page centered and nicely spaced. (More than one article per sheet (or sheet per article!) tends strongly to violate neatness and organization.)

Our standard review bonus structure will apply to this assignment.

0.2 Statistics Project Rubric

Item	Score	Possible	Item	Score	Possible
Number of Articles		15	Number of Articles		15
deceptive		10	deceptive		10
description		15	description		15
validity/opinion		5	validity/opinion		5
citation		30	citation		30
neatness		15	neatness		15
cover sheet		10	cover sheet		10
bonus (extras)		± 10	bonus (extras)		± 10
subtotal		100	subtotal		100
early bonus/late demerits		$\pm 20\%$	early bonus/late demerits		$\pm 20\%$
total		$100 \pm 32?$	total		$100 \pm 32?$

Item	Score	Possible	Item	Score	Possible
Number of Articles		15	Number of Articles		15
deceptive		10	deceptive		10
description		15	description		15
validity/opinion		5	validity/opinion		5
citation		30	citation		30
neatness		15	neatness		15
cover sheet		10	cover sheet		10
bonus (extras)		± 10	bonus (extras)		± 10
subtotal		100	subtotal		100
early bonus/late demerits		$\pm 20\%$	early bonus/late demerits		$\pm 20\%$
total		$100 \pm 32?$	total		$100 \pm 32?$

Item	Score	Possible	Item	Score	Possible
Number of Articles		15	Number of Articles		15
deceptive		10	deceptive		10
description		15	description		15
validity/opinion		5	validity/opinion		5
citation		30	citation		30
neatness		15	neatness		15
cover sheet		10	cover sheet		10
bonus (extras)		± 10	bonus (extras)		± 10
subtotal		100	subtotal		100
early bonus/late demerits		$\pm 20\%$	early bonus/late demerits		$\pm 20\%$
total		$100 \pm 32?$	total		$100 \pm 32?$

Stat's Lesson 1

Definitions, Uses, Data Types, and Levels of Measurement

*Measure what is measurable and
make measurable what is not so.*

Galileo Galilei

This is the first lesson of ten in a series introducing statistics. We concentrate on descriptive statistics overall and on some basic definitions and levels of measurement in this lesson.

We will also continue to feature famous mathematicians and scientists in this series of lessons. These were picked with care. Some are important for the scientific method (Galileo, Bacon, Tycho), some for science/mathematics in general (Kelvin, Poisson, Chebyshev), and the rest due to their contribution specifically to statistics (Fisher, Tukey, Gosset, Pearson). A few scientists/politicians are noted only in passing. These include: Millikan (1), Yates (1), Fahrenheit/Rankine/Celsius (1), Gallup/Dewey/Truman (2). Except for the quote here and in lessons two and three by Tycho and Michelson/Kelvin, they are just for fun, without the burden of knowing who said it. Today's biography is on Galileo.

1.1 Italian Father of Modern Science: Galileo Galilei

Galileo Galilei (1564–1642), often referred to only as Galileo, was an Italian mathematician, astronomer, and physicist. He made several significant contributions to modern scientific thought. He is considered a founder* of the scientific or experimental method on which modern science is based. He is especially noted for being the first man to study the skies with the telescope and proving the Earth revolves around the Sun. For thousands of years many men were content to assume heavier things fell

*Francis Bacon and Tycho Brahe are also given credit for recognizing the need for using the inductive (scientific) method to discover a few powerful laws and theories about how nature works.

faster, but Galileo proved theoretically that falling bodies obey what is now known as the **law of uniformly accelerated motion**. He gathered evidence that proved the Earth revolves around the Sun and that it was not the center of the universe as was then believed. More importantly, he maintained this position despite trial in Rome and church orders to recant. He was forced to spend the last eight years of his life under house arrest. His most far-reaching achievement was the re-establishment of the scientific method against Aristotle's flawed approach.

Galileo was born at Pisa in 1564. He studied medicine at the university there starting in 1581. Supposedly it was here in the Pisa cathedral during his first year that he observed a lamp swinging and found that its period was constant, independent of the amplitude of the oscillation. Later in life he verified this observation experimentally and suggested that this principle might be used to regulate clocks. A chance Geometry lesson he overheard awakened his interest in mathematics and he began to study Mathematics and Science. In 1585, before he received a degree, he was withdrawn from the university due to lack of funds. Four years later his treatise on center of gravity earned him a post of mathematics lecturer back at Pisa. Galileo spent his childhood and the intervening years in Florence. In 1592 he was awarded the chair of mathematics at Padua where he remained for 18 years and performed the bulk of his work.

Galileo had a long-standing conflict with the Roman Catholic church regarding its teaching of a geocentric universe (Aristotle). Galileo's beliefs were supported by observations of craters on the moon, sunspots, and the heliocentric solar system (Copernicus). Galileo said "[The book of nature is written in clear mathematical form.](#)" Although often attributed to him, he may have quoted others for "[The book of scriptures was written to show us how to go to heaven, not how the heavens go.](#)"

Galileo had enough faith in the mathematical model of a moving earth to suffer condemnation by the establishment (Catholic church) until 1992! The year Galileo died (1642) was the year Newton was born.

1.2 The Basis of Science: The Scientific Method

Aristotle developed his theories of nature in the deductive style of logic. A few truths were accepted as obvious and other statements followed logically. Thus for almost 2000 years a horse had 40 teeth because Aristotle said so, not because anyone actually opened a horse's mouth and counted.[†]

The scientific method is the accurate observation of facts and the determination of order among the facts. Generally these are framed in mathematical form. This may be done by inductive reasoning, inference from a number of observed facts, or

[†]Variations on this parable are often set in ancient Greece or the year 1432, sometimes attached to a monastery, and were often attributed to Thomas Aquinas or Roger/Francis Bacon.

by deductive reasoning from a set of principles. Often, predictions are possible that are open to experimental testing.

The scientific method usually has at least five steps: (i) stating the problem; (ii) forming the hypothesis; (iii) observing the experiment (taking data); (iv) interpreting the data; and (v) drawing the conclusion by developing theory. These steps, however, often don't follow that exact order; unexpected results are often observed! These checkpoints are often used to arrange and write up an experiment.

Mathematics itself is seldom in conflict with religion. However, science, scientists, the scientific method, and the scientific theories generated often are. In general, scientific theories cannot be rigorously proved. Models are constructed which give an approximate mental picture, often giving a deeper understanding, though analogy. Well-developed models become theories and theories lead to laws about how nature works or behaves. Scientific laws generally cannot be broken, unlike political laws (the law of gravity vs. the speed limit). Laws are valid over a wide range of cases and any limitations or range of validity is clearly understood.

Scientists conduct their experiments as if the generally accepted theories/laws were true, while keeping an open mind in case new information is revealed.

Science encompasses a vast body of empirical knowledge and to try to pick and choose what to believe and what not to believe is an affront to the scientific method.

Popes Pius XII in 1951 and John Paul II in 1996 declared that Catholics may accept Evolution as more than a hypothesis and the Big Bang as a “splendid solution” without contradicting their faith. One can only hope that other religious groups will consider a similar position within a few nanohubble times.[‡] Little more can be said here while scrupulously avoiding either the fact or appearance of inadequate separation of secular and sectarian activities as required by the contract between the county and university.

1.3 Introduction to Statistics

The term **statistics** has two basic meanings. First, statistics is a subject or field of study closely related to mathematics. This two week, ten lesson unit serves as a short introduction, briefly covering the area known as descriptive statistics, and introducing two inferential statistical tests.

Descriptive statistics generally characterizes or describes a set of data elements by graphically displaying the information or describing its central tendencies and how it is distributed.

Center sophomores generally spend several weeks reviewing this material and ex-

[‡]A Hubble time is the age of the universe, with a current best estimate of 13.73(12) billion years.

tending their study of statistics with 15 lectures on probability and distributions with emphasis on the normal distribution. Those who wish to go further can study inferential statistics, and thus prepare for the AP Statistics Test.* Our intent is for juniors to have completed over half that curriculum, when this introduction and the subsequent probabilities and distribution lessons are taken into account.

Inferential statistics tries to infer information about a population by using information gathered by sampling.

Statistics: The collection of methods used in planning an experiment and analyzing data in order to draw accurate conclusions.

1.4 General Terms Used Throughout Statistics

Population: The complete set of data elements is termed the population.

The term population will vary widely with its application. Examples could be any of the following proper subsets:[†] animals; primates; human beings; *homo sapiens*; U.S. citizens; who are high school students; attending the Math & Science Center; living in Berrien County; as freshmen (class of 2013); females; home school of Niles, with one younger brother.

Sample: A sample is a portion of a population selected for further analysis.

How samples are obtained or types of sampling will be studied in the next lesson. Most any of the examples above for population could serve as a sample for the next higher level data set.

Parameter: A parameter is a characteristic of the whole population.

Statistic: A statistic is a characteristic of a sample, presumably measurable.

The plural of statistic just above is the second basic meaning of statistics.

Assume there are 30 students in a particular statistics class, with 6 going to Niles High School. Since 6 is 20% of 30, we can say 20% go to Niles. The 20% represents a *parameter* (not a *statistic*) of the class because it is based on the entire population. If we assume this class is representative of all classes, and we treat these 5 students as a sample drawn from a larger population, then the 20% becomes a statistic.

Remember: Parameter is to Population as Statistic is to Sample.

*See <http://www.collegeboard.org/ap/statistics/html/index001.html>.

[†]See Numbers Lesson 1 on sets.

1.5 Accuracy vs. Precision

The distinction between accuracy and precision, reviewed earlier in Numbers Lesson 10, is very important and the student is assumed to be familiar with it. Briefly, precision is a measure of **exactness** or repeatability. and accuracy is a measure of **rightness** or how correct the result is.

1.6 Uses and Abuses of Statistics

Most of the time, samples are used to infer something (draw conclusions) about the population. If an experiment or study was done cautiously and results were interpreted without bias, then the conclusions would be accurate. However, occasionally the conclusions are inaccurate or inaccurately portrayed for the following reasons:

- Sample is too small.
- Even a large sample may not represent the population.
- Unauthorized personnel are giving wrong information that the public will take as truth. A possibility is a company sponsoring a statistics research to prove that their company is better.
- Visual aids may be correct, but emphasize different aspects. Specific examples include graphs which don't start at zero thus exaggerating small differences and charts which misuse area to represent proportions. Often a chart will use a symbol which is both twice as long and twice as high to represent something twice as much. The area, in this case however, is **four times** as much!
- Precise statistics or parameters may incorrectly convey a sense of high accuracy.
- Misleading or unclear percentages are often used.

Statistics are often abused. Many examples could be added, (even books have been written) but it will be more instructive and fun to find them on your own.

1.7 Types of Data

A dictionary defines data as facts or figures from which conclusions may be drawn. Thus, technically, it is a collective, or plural noun. Some recent dictionaries acknowledge popular usage of the word data with a singular verb. However we intend to adhere to the traditional “English” teacher[‡] mentality in our grammar usage—sorry

[‡]My mother and step-mother were both English teachers, so clearly no offense is intended above.

if “data are” just doesn’t sound quite right! **Datum** is the singular form of the noun data. Data can be classified as either numeric or nonnumeric. Specific terms are used as follows:

1. Qualitative data are nonnumeric.

Poor, Fair, Good, Better, Best, colors (ignoring any physical causes), and types of material straw, sticks, bricks are examples of qualitative data.

Qualitative data are often termed **categorical data**. Some books use the terms **individual** and **variable** to reference the objects and characteristics described by a set of data. They also stress the importance of exact definitions of these variables, including what units they are recorded in. The reason the data were collected is also important.

2. Quantitative data are numeric.

Quantitative data are further classified as either discrete or continuous.

- Discrete data are numeric data that have a finite number of possible values.

A classic example of discrete data is a finite subset of the counting numbers, $\{1, 2, 3, 4, 5\}$ perhaps corresponding to Strongly Disagree, . . . , Strongly Agree.

Another classic is the electric charge of a single electron which was first convincingly measured in 1911 in the **Millikan Oil-drop Experiment**. **Quantum Mechanics**, the field of physics which deals with the very small, is much concerned with discrete values.

When data represent **counts**, they are discrete. An example might be how many students were absent on a given day. Counts are usually considered exact and integer. Consider, however, if three tardies make an absence, then aren’t two tardies equal to 0.67 absences?

- **Continuous** data have infinite possibilities: 1, 1.4, 1.41, 1.414, 1.4142, 1.41421, . . . , $\sqrt{2}$.

The real numbers[§] are continuous with no gaps or interruptions. Physically measurable quantities of length, volume, time, mass, *etc.* are generally considered continuous. At the physical level (microscopically), especially for mass, this may not be true, but for normal life situations is a valid assumption.

[§]See Numbers Lesson 14.

The structure and nature of data will greatly affect our choice of analysis method. By structure we are referring to the fact that, for example, the data might be pairs of measurements. Consider the legend of Galileo dropping weights from the Leaning Tower of Pisa. The times for each item would be paired with the mass (and perhaps surface area) of the item.

1.8 Levels of Measurement

The experimental (scientific) method depends on physically measuring things. The concept of measurement has been developed in conjunction with the concepts of numbers and units of measurement. Statisticians categorize measurements according to levels. Each level corresponds to how this measurement can be treated mathematically.

1. **Nominal:** Nominal data have no order and thus only gives **names** or labels to various categories. You can only count nominal data, and cannot otherwise measure it.
2. **Ordinal:** Ordinal data have **order**, but the interval between measurements is not meaningful.
3. **Interval:** Interval data have meaningful intervals between measurements, but there is no true starting point (zero).
4. **Ratio:** Ratio data have the highest level of measurement. Ratios between measurements as well as intervals are meaningful because there is a starting point (zero).

Nominal comes from the Latin root **nomen** meaning **name**. Nomenclature, nominative, and nominee are related words. **Gender** is nominal. (Gender is something you are born with,[¶] whereas **sex**^{||} is something you should get a license** for.)

Example 1: Colors

[¶]The dictionary defines gender as the fact or condition of being male or female, especially regarding how this affects or determines a person's self image, *etc.* Thus genitals may not be the only determining factor and hormonal and environmental influences may dominate.

^{||}Unfortunately, dictionaries generally note both maleness/femaleness in addition to sexual intercourse in the various definitions of sex.

**Although this statement is consistent with the federal abstinence-only program which specifies that the exclusive purpose of the education must be to "teach that a mutually faithful monogamous relationship in the context of marriage is the expected standard of human sexual activity" and that "sexual activity outside of the context of marriage is likely to have harmful psychological or physical effects," sexuality education may be more effective if contraception information is included, as the majority of parents prefer. www.ejhs.org, Volume 4, June 25, 2001.

To most people, the colors: black, brown, red, orange, yellow, green, blue, violet, gray, and white are just names of colors.

To an electronics student familiar with color-coded resistors, this data is in ascending order and thus represents at least ordinal data.

To a physicist, the colors: red, orange, yellow, green, blue, and violet correspond to specific wavelengths of light and would be an example of ratio data.

Example 2: Temperatures

What level of measurement a temperature is depends on which temperature scale is used. Specific values: $0^{\circ}\text{C} = 32^{\circ}\text{F} = 273.15\text{ K} = 491.69^{\circ}\text{R}$ $100^{\circ}\text{C} = 212^{\circ}\text{F} = 373.15\text{ K} = 671.67^{\circ}\text{R}$ $-17.8^{\circ}\text{C} = 0^{\circ}\text{F} = 255.4\text{ K} = 459.67^{\circ}\text{R}$ where C refers to Celsius (or Centigrade before 1948); F refers to Fahrenheit; K refers to Kelvin; R refers to Rankine.

Only Kelvin and Rankine have true zeroes (as starting point) and ratios can be found. Celsius and Fahrenheit are interval data; certainly order is important and intervals are meaningful. However, a 180° dashboard is not twice as hot as the 90° outside temperature (Fahrenheit assumed)! Rankine has the same size degree as Fahrenheit but is rarely used. To interconvert Fahrenheit and Celsius, see Numbers Lesson 13. (Note that since 1967, the use of the degree symbol on temperatures Kelvin is no longer proper.)

Although ordinal data should not be used for calculations, it is not uncommon to find averages formed from data collected which represented Strongly Disagree, . . . , Strongly Agree! Also, averages of nominal data (zip codes, social security numbers) is rather meaningless!

Name _____

Score _____

1.9 Homework, Stat's Introduction

Each problem is worth two points.

1. What is the difference between **statistics** and **statistic**?
2. What is the difference between **descriptive** and **inferential** statistics?
3. Would you trust a 1998 poll to be accurate when saying that Clinton needs to resign if you find out that the poll is collected by people calling in response to a newspaper article? Why or why not?
4. Complete the following comparison: **Parameter** is to ?.....?, as **statistic** is to ?.....?.
5. What are the two categories of data starting with the letter **q**?
6. If numeric data are not **discrete**, then they must be ?.....?.
7. Arrange in order from **highest to lowest** the four levels of measurement.
8. If your teacher's portfolio dropped 16% in value, show how to find what percent increase (from the resulting, new value) would be required to return it back to its original value?

The first edition of a textbook contained 600 exercises. For the revised edition, the author removed 50 of the original exercises and added 350 new exercises. Complete each of the following statements.

9. There are ?.....? exercises in the revised ed.

10. There are ?.....? more exercises in the revised edition than the 1st ed.

11. There are ?.....?% more exercises in the revised edition than the 1st ed.

12. ?.....?% of the [revised edition] exercises are new.

13. Assume 25% of the deer population is infected with TB. Suppose the total population is reduced by 10% by recurring annual methods. If the initial population was 100,000, how many infected deer are left? (Assume that the reduction methods operate independantly of infection.)

14. **In two words**, describe the difference between precision and accuracy.

15. If $0^{\circ}\text{C} = 32^{\circ}\text{F}$ and $100^{\circ}\text{C} = 212^{\circ}\text{F}$, find the temperature which is represented by the same number on both scales.

16. _____ is something you are born with, whereas _____ is something you should get a _____ for.

Stat's Lesson 2

The How and Why of Statistical Sampling

I've studied all available charts of the planets and stars and none of them match the others. There are just as many measurements and methods as there are astronomers and all of them disagree. What's needed is a long term project with the aim of mapping the heavens conducted from a single location over a period of several years. Tycho Brahe

In this lesson we will discuss various aspects of statistical sampling. We discuss four points to consider, five type of sampling, sampling errors, types of questions, and the importance of record keeping. Information on experimental design is pending.

2.1 Danish Father of Modern Astronomy: Tycho Brahe

Tycho Brahe (1546–1601) was slightly older than Galileo and made a significant contribution to our understanding of the solar system. He was a Danish nobleman famed for his accurate and comprehensive naked eye astronomical observations and alchemy (then a respected occupation). A predicted eclipse in 1560 fascinated him. Three years later at age 17 he wrote the quote given above.

Tycho, as he was commonly called in the Scandinavian tradition, was given an island estate with funding to build a research institute. His large astronomical instruments, good seeing conditions, and careful and redundant measurements were instrumental in Kepler's calculations of Mars's elliptical orbit and his laws of planetary motion which overturned the Ptolemaic (geocentric with epicycles) and established the Copernican (heliocentric) system.* However, from about 1610 when Galileo observed the phases of Venus and overturned the Ptolemaic system, until after Galileo's death when the Copernican system took over, a Tychonic system was common and

*Copernicus is known as the father of modern astronomy.

even supported by the Roman Catholic Church. In this system the planets orbited the sun, but the sun orbited the earth. In 1572 Tycho coined the term nova for a new star when he observed a supernova. This later inspired a Poe poem and is probably referenced in Shakespeare's *Hamlet* as the “[star that's westward from the pole.](#)” Kepler only had access to limited data before Tycho died and he obtained the rest only after some controversy.

Tycho's nose was cut off in a duel when he was 20. He usually wore a prosthetic nose. History records it being made of gold and silver, but in 1901 when they opened his grave they found green in his sinus area suggesting copper. Tycho at one point owned 1% of Denmark's wealth. His tame moose apparently got drunk at a party, fell down the stairs, and died. Kepler's account of Tycho's death was consistent with a bladder infection brought on by staying until the end of a banquet. He died eleven days later. Recent evidence suggests Tycho died of mercury poisoning and many doubt he would have poisoned himself.

2.2 Points to Consider

Before analyzing data statistically, it is important to consider if the data were collected appropriately. Many years of labor and even careers have been virtually wasted because of fundamental flaws in the data collection step. The statistical analysis will only likely be a minor part of the total expense of a properly conducted experiment, so time, effort, and money spent ensuring the data are collected appropriately is certainly well spent. The computer adage **Garbage In, Garbage Out** or **GIGO** is rather apropos.

Ensure that the **sample size** is large enough.

Although a large sample is no guarantee of avoiding bias, too small a sample is a recipe for disaster. How to determine minimum sample size goes beyond the scope of this introduction, but suffice it to say there are well established techniques to determine such. These techniques are based on the Central Limit Theorem and some information can be found in Probabilities and Distributions Lesson 11.

Better results are obtained by **measuring** instead of asking.

A good classroom example would be to collect people's heights. We expect such data might be randomly distributed. Asking will result in several sources of error. Perhaps the most common being exaggeration, rounding, hair style, and shoe heel variation or even complete absence of shoes. Were you instead to measure each individual, these sources of error could be reduced. You may still encounter **systematic errors**. Following are some sources of systematic error. Perhaps your measuring device is defective. Specific examples might include the common fact that rulers often don't start exactly at zero, but have a little extra margin. Maybe the measuring

tape is marked off in inches on one side and tenth's of a foot on the other and sometimes the wrong side is read. Tape measures can become kinked or even tangled (especially surveying caves). Perhaps being a Center student correlates with being shorter or taller for some unknown reason. This might only be a problem if you were to use your data to represent a larger population.

The **medium used** (mail, phone, personal interview) is important.

Surveys are a very popular method of data collection for social issues. Mail surveys tend to have a lower response rates which will distort and hence flaw a sample. Although telephone surveys may be relatively efficient and inexpensive, the more time consuming and correspondingly expensive personal interview allows more detailed and complex data to be collected. Be not called by telemarketers—the five year don't call list is expiring—you may need to reapply.

Be sure the **sample is representative** of the population.

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. An **experiment** deliberately imposes some treatment on individuals in order to observe their responses.

Observational studies are then a poor way to gauge the effect of an intervention. When our goal is to understand cause and effect, experiments are the only source of fully convincing data. However, imposing treatments may produce some ethical concerns. See more below under experimental design.

Before we move on to the next point, we should note that some studies are **retrospective**, or involve looking back at past events, whereas others are **prospective** or track groups forward in time.

2.3 Methods of Sampling

Sampling is the fundamental method of inferring information about an entire population without going to the trouble or expense of measuring every member of the population (census) . Developing the proper sampling technique can greatly affect the accuracy of your results.

Statisticians have classified sampling into five common types, as given below. A sixth type is sometimes included, the census. However, a census included every member of the population so is an improper subset, so it is technically not a sample.

Random Sampling: Members of the population are chosen in such a way that each have an equal chance to be measured.

Other names for random sampling include **representative** and **proportionate** sampling because all groups should be *proportionately represented*. Consider what might happen if a telephone directory were used as a source for randomly selecting

survey participants. Some people have no phone, others have multiple phones and corresponding listings. Still others have unlisted phone numbers. In affluent areas unlisted phone numbers were approaching half the population about the year 2000 and are certainly higher today. Now-a-days many are giving up land lines and use cell phone exclusively. Cell phone directories are controversial at best and the law disallows the use of computer dialers to access them. Pollsters commonly use computers to generate and dial phone numbers in an attempt to circumvent these problems. However, many people consider such use of the telephone as an invasion of their privacy and refusals or hang-ups may well significantly influence the outcome. Some of us have learned to recognize these computer dialers and quickly hang up. Such are the pitfalls which must be carefully considered in designing an experiment, study, or survey.

Systematic Sampling: Every k^{th} member of the population is sampled.

The historic meaning of the word **decimate**, where every 10th Roman soldier was killed, usually by his cohorts, is a gruesome example of systematic sampling.

Stratified Sampling: The population is divided into two or more strata and each subpopulation is sampled (usually randomly).

Stratum is the singular form of the word strata which means *to spread out*. One of the word's most common usage is in geology to describe the layers of sedimentary rocks which have formed during the earth's history. Gender and age groups would be commonly used strata. **Classes** is another term for strata. Each stratum must share the same characteristic. Random sampling may well be used to select a certain number of data points from each stratum. This is often the most efficient sampling method.

Cluster Sampling: A population is divided into clusters and a few of these (often randomly selected) clusters are exhaustively sampled.

Exhaustively means considering all elements. Cluster sampling is used extensively by governmental and private research organizations.

Convenience Sampling: Sampling is done as convenient, often allowing the element to choose whether or not it is sampled.

Convenience sampling is the easiest and potentially most dangerous. Often good results can be obtained, but perhaps just as often the data set may be seriously **biased**. Consider collecting GPA information from students in detention. It may be convenient, but perhaps not representative of the entire student body!

Be **wary** of convenience sampling.

2.4 Sampling Error

We have listed above several sources of **sampling error**. One of the most famous sampling errors occurred in 1948 when the Gallup poll predicted Dewey would be elected president over Truman. The day after the election, such an announcement made the front page of a major newspaper! Gallup then abandoned the quota system and instituted random sampling based on clusters of interviews nationwide. Sample subjects should be selected by the **pollster**. They should not select themselves as they do via mail or perhaps telephone surveys. The systematic errors listed above are examples of **nonsampling errors**.

Of great debate recently was what to do with the errors which arise in the decennial US Census. Considerable time was spent by all three branches of our government addressing this issue for the 2000 census.

2.5 Question Types

Some questions are classified as **open**, whereas other questions are classified as **closed**. Open questions elicit open-ended responses and thus work best in a personal interview. Multiple-choice or true/false questions are a type of closed question. Closed questions can thus more easily be coded and analyzed by a computer.

2.6 Record Keeping

In science especially, a detailed lab notebook is important for serious work. Standards will vary with the institution, level, and seriousness of the work. Some common requirements are as follows.

1. Records should be kept in a stitched notebook of quality paper.
2. Entries should be made in ink with each page numbered (and none missing).
3. Each page should be signed and dated by the principle participants.
4. A full account of each experiment should be given, including set-up, procedure, original data, analysis, and conclusions. This might include who walked into the room, when, what they were wearing, *etc.* Strong results sometimes have even stranger explanations!
5. Establish beforehand who will retain the notebook.

If surveys are used be sure to include the survey sponsor, the date the survey was conducted, the size of the sample, the nature of the population sampled, the

type of survey used, and the exact wording of the survey questions. Other important issues include: assessing the risk to those surveyed, the scientific merit of the survey, and the guarantee of the subject's consent to participate. An example of risk might be the hazard of planting ideas (rape, murder, suicide, *etc.*) in someone's head or reviving suppressed memories (abuse) while asking related questions. Nuclear poisoning/fallout from the 1950's and associated cancer deaths would be another example of risk which was recently (Sep. 2000/1998) in the news.

Lab notebooks and other sources have commonly been used to establish priority and patent claims. This may occur long after the records were made, so detail and clarity are important to remember. Patents and associated royalties for the transistor, the laser, and even computers have depended on such records, long after the fact! Lab write-ups are an essential part of any science experiment and will provide practice in this area.

The principle author has done original research in several different fields (Chemistry, Mathematics, Physics, and Computer Science) at various times in his career. Organization has been a key factor in such an achievement. Without it, the ability to move between fields would be severely hampered. Detailed records were certainly important when accusations of embezzlement arose!

Fabrication or falsification of data, although rare, is a serious breach of ethics. It can easily result in the end of a career however promising it might have been. Such record keeping can be extremely important in trying to reproduce someone else's findings. You might consider it a sort of professional diary. Classic scientific failures include evidence of a fifth force (antigravity) and cold fusion (now known as low energy nuclear reactions or LENR). Differentiating between being misled and fabrication can be very important.

2.7 Experimental Design

More information on experimental design (treatments, factors, blocking, double blind, latin square, randomized complete block, matched pairs, replication, and simulation) should be included here but isn't. Consult any good Statistics book or take the AP Statistics course for more information.

Name _____

Score _____

2.8 Homework, Statistical Sampling

Each problem is worth two points.

Identify each number as *discrete* or *continuous*.

1. Yesterday's records for MSC attendance show that two underclassmen were absent.
2. Toyota hopes to produce 100,000 Prius hybrids for the United States in 2005.
3. A 1999 Cadillac Escalade weighs 5,600 pounds.
4. The radar clocked a Justin Verlander fastball at 101.4 mph.

Determine which *level of measurement* is most appropriate.

5. Colors of SkittlesTM brand candies.
6. Final course grades of A, B, C, D, and F.
7. Daily high and low temperatures at the Niles airport for 2004.
8. Time (in days) for a sunspot to be visible from the earth.

Identify the *type* or *types of sampling* used for the following.

9. George went through the telephone book and called every 89th person listed.
10. Four people divided the telephone book evenly and each randomly sampling from their portion.
11. All people with a 461 telephone exchange are called.
12. Every 5th block of 10 students leaving the Eau Claire High School cafeteria on June 31 is exhaustively sampled about their faith in random samples.

13. What four letter words are associated with convenience sampling?

14. Differentiate between prospective and retrospective studies.

15. Give two other names for random sampling.

16. What is an open question?

17. What is the average of: 1, 1, 2, 4, 7?

18. Is this a closed question (yes or no)?

19. Give three examples where patent litigation dragged on for years and laboratory notes became very important.

20. What is the difference between an experiment and a study?

21. How important are lab notebooks? Why?

Stat's Lesson 3

Averages: Mean, Mode, Median, or Midrange?

It seems probable that most of the grand underlying principles have now been firmly established and that further advances are to be sought chiefly in the rigorous application of these principles to all the phenomena which come under our notice....future truths of physical science are to be looked for in the sixth place of decimals.

Michelson in 1894, apparently quoting Kelvin

This lesson discusses the four major types of measures of central tendency (averages), which one is best, and reviews round-off rules. It closes with examples and a calculator activity.

3.1 The Father of Thermodynamics: Lord Kelvin

Echoing the Galileo quote from Lesson 1, Kelvin said: *I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.*

William Thomson (1824–1907) was born in Ireland, worked most of his life in Scotland, but is considered a British mathematical physicist, engineer, inventor, and leader. His major works dealt with electricity and thermodynamics, including the absolute temperature scale named after him. His work in mathematical analysis did much to give physics its current form. He was also very involved as an engineer in the laying of an early transatlantic cable, first used for telegraph. This later career gave him wealth, fame, and his title of Baron, but he was commonly known as Lord Kelvin.

Three of Kelvin's early papers were written under the pseudonym P.Q.R. on the topic of heat. His analogy of heat conduction with electrostatics was ultimately described by Maxwell as one of the most valuable science-forming ideas. By age 22 he was appointed chair of natural philosophy at Glasgow where he stayed over 50 years. Kelvin was a major contributor to the development of the Second Law of Thermodynamics which states that perfect efficiency in energy exchange is impossible or **entropy** (disorder) is always increasing.

3.2 Averages

Average most often refers to the arithmetic mean, but is actually ambiguous and may be used to also refer to the mode, median, or midrange.

You should always clarify which average is being used, preferably by using a more specific term. Averages give us information about a typical element of a data set. Averages are **measures of central tendency**.

Mean most often refers to the **arithmetic mean**, but is also ambiguous. Unless specified otherwise, we will assume arithmetic mean whenever the term is used.

The **Arithmetic Mean** is obtained by summing all elements of the data set and dividing by the number of elements.

A host of other means and their method of computation will be discussed in Statistics Lesson 4. Symbolically, the arithmetic mean is expressed as $\bar{x} = \frac{\sum x_i}{n}$, where \bar{x} (pronounced "x-bar") is the arithmetic mean for a sample and Σ is the capital Greek letter sigma and indicates summation. x_i refers to each element of the data set as i ranges from 1 to n . n is the number of elements in the data set. The equation is essentially the same for finding a population mean; however, the symbol for the population mean is the small Greek letter μ (mu). As we will also see in Statistics Lesson 5, Roman letters usually represent sample statistics, whereas Greek letters usually represent population parameters.

Sample size is the number of elements in a sample.
It is referred to by the symbol n .

Be sure to use a lower case n for sample size. An upper case N refers to **Population Size**, unless being used in the context of a normally distributed population.

Mode is the data element which occurs most frequently.

A useful mnemonic is to alliterate the words mode and most. Alliterations start with the same sound like: "seven slippery slimy snakes slowly slipping southward."

Some data sets contain no repeated elements. In this case, there is **no mode** (or

the mode is the empty set). It is also possible for two or more elements to be repeated with the same frequency. In these cases, there are two or more modes and the data set is said to be **bimodal** or **multimodal**. In the rare instance of a **uniform** or nearly uniform distribution, one where each element is repeated the same or nearly the same number of times, one could term it multimodal, but some authors invoke subjectivity by specifying multimodality only when separate, distinct, and fairly high peaks (ignoring fluctuations due to randomness) occur. Thus mode can be subjective.

The **Median** is the middle element
when the data set is arranged in order of magnitude (**ranked**).

A useful mnemonic is to remember that the median is the grassy strip (in the semirural area of the midwest where I come from) that divides opposing lanes in a highway. It is in the middle.

If there are an odd number of data elements, the median is a member of the data set. If there are an even number of data elements, the median is computed as the arithmetic mean of the middle two and thus may or may not be a member.

The median has other names which will be studied in Statistics Lesson 7. The symbol \tilde{x} (pronounced “*x*-tilde”) is sometimes used for the median, but will not be used here.

The **Midrange** is the arithmetic mean of the highest and lowest data elements.

Midrange is a type of average. **Range** is a measure of dispersion and will be studied in Statistics Lesson 5. A common mistake is to confuse the two.

Symbolically, **midrange** is computed as: $\frac{x_{\max} + x_{\min}}{2}$.

3.3 The Best Average

The ambiguity of the term average can lend to deception. Statisticians may often be cast as liars as a result. Note how advertisers may distort statistics to pursue their goals.

Some basic facts regarding averages are as follows.

1. The mean, median, and midrange always exist and are unique.
2. The mode may not be unique or may not even exist.
3. The mean and median are very common and familiar.
4. The mode is used less frequently; midrange is rarely used.
5. Only the mean is “reliable” in that it utilizes every data element.

6. The midrange, and also somewhat the mean, can be distorted by extreme data elements (see Statistics Lesson 8).
7. The mode is the only appropriate average for nominal data.

3.4 Round-off Rules

The mode, if it exists, and possibly the median are elements of the data set. As such, they should be specified no more accurately than the original data set elements.

The midrange and possibly the median are the arithmetic mean of two data set elements. One additional significant digit may be necessary to accurately convey this information.

The number of significant digits for the mean should conform to one of the following rules.

1. The significant digits should be no more than the number of significant digits in the sum of the data elements. If the data have fairly consistent precision, this should be easy to determine.* Those rules were outline in Numbers Lesson 10. This is sometimes simplify as a rule of thumb[†] by stating that the mean should be given to one more decimal place than the original data. However, this assumes the data set is small ($n \ll 100$) and that the data was recorded to a consistent precision.
2. The number of significant digits should be consistent with the precision obtained for the standard deviation. This concept is expanded upon in Statistics Lesson 5 after measures of dispersion are discussed.
3. It is not uncommon in science for results to be left in and perhaps interim calculations sometimes rounded to **three significant digits**, which is about all you could get out of a slide rule. Hence, this was commonly termed **slide rule accuracy**. In pre-calculator days, this also made hand calculations easier.

The important thing to remember is not to write down ten decimal places without good reason, even though your calculator will often display such.

Presenting more than five significant digits, except on variance,
is probably a joke and points will be deducted!

Please note the quote at the beginning of the lesson regarding the state of precision in physics just before 1900. Relativity and quantum mechanics soon revo-

*Since the sample size (n) is an exact value, it has no affect on the number of significant digits obtained from the division.

[†]On a historic note, the term **rule of thumb** apparently does not come from any old English law to limit the size of stick which a husband could use to beat his wife as often stated. However, abusive relationships still remain an often hidden societal problem.

lutionalized physics and we soon were looking at details in the ninth place! The author's dissertation reported results of the cesium D_1 transition centroid frequency as: 335 116 048 748.2(2.4) kHz. The (2.4) is the one standard deviation uncertainty in this result.

3.5 Examples

Example: The homework for Statistics Lesson 2 near the end had the question: 17. What is the average of: 1, 1, 2, 4, 7?

Answer: As we have seen in this lecture, this is a rather ambiguous question and the answers 1 (mode), 2 (median), 3.0 (mean), and 4.0 (midrange) are all possible and correct!

Example: A sample of size 5 ($n = 5$) is taken of student quiz scores with the following results: 1, 7, 8, 9, 10. Find the average score.

Answer: The mean is $(1+7+8+9+10)/5 = 35/5 = 7.0$ (note one more decimal place is given). All scores occur only once, hence there is no mode. The median score is 8 (not 8.0). The midrange is $(10+1)/2 = 5.5$ (note the extra decimal place is required).

An extreme score (1) distorts the mean so perhaps the median is a better measure of central tendency. For a larger data set, this could be further defined in terms of **skewness** (median and generally mean to the left of (**negatively skewed**), right of (**positively skewed**), or same as (**zero skewness**) the mode) and **symmetry** of the data set. It is more common to be positively skewed, since exceptionally large values are easier to obtain due to lower limits. A case in point would be annual earnings. Our left **tail** is cut off by zero, whereas our right **tail** is extremely skewed by the likes of Bill Gates, Berhard Madoff, and Warren Buffett.

Example: Form the mean of the data set: 20, 1, and 1.5.

Answer: Naively, one might find $22.5/3=7.5$. However, if one were to technically follow the rules for adding significant digits, one would obtain $20/3=7$, where the sum 20 was formed using the proper rules of significant digits, and the quotient 7 similarly.

3.6 Activity: Calculator Averages

Please use your TI-84+ calculator for the following activities.

Press the **STAT** key and **ENTER** to select **EDIT**. Now enter the homework data from Statistics Lesson 2, question 17. Your screen should appear as in the left screen in Figure 3.6.

Press the **2nd MODE (QUIT)** or just go directly to **STAT**, arrow over to **CALC**, and **ENTER** to select **1-Var Stats**. Now enter **2nd 1 (L₁)**. Although this

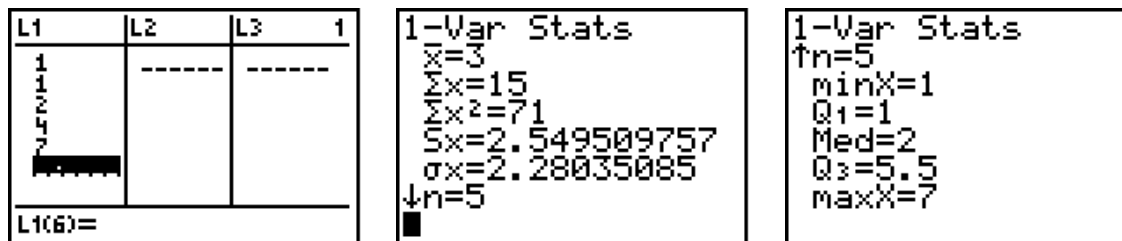


Figure 3.1: TI-84 Graphing Calculator Data in List (left); Mean and Standard Deviation from 1 Variable Statistics (middle); and Five Number Summary (right).

last aspect (L_1) is optional, it can save a lot of grief, since you never know who set the defaults to what! Your screen should now appear as in the middle screen in Figure 3.6.

\bar{x} is the arithmetic mean. Also shown are the sum of the data elements and the sum of their squares. Next is S_x . This is the **sample** standard deviation to be discussed in Statistics Lesson 5. Next is σ_x . This is the **population** standard deviation also to be discussed Statistics Lesson 5. Last is n or the sample size.

Please note the down arrow to the left of the n . If you arrow down, the **5-number summary** will be displayed in the right screen in Figure 3.6. This is the topic of Statistics Lesson 7, but note that the median (**Med**) is displayed there. For large lists, sorting it first may facilitate finding the mode (or trimming it to find trimmed means). This can be accomplished by **2nd STAT (LIST)**, then arrow over to **OPS ENTER** (to select **SortA(**, then **2nd 1 L₁**)). The closing parenthesis is optional.

The midrange can be obtained by adding the 1 and 7 and dividing by 2. If you get 4.5, you forget to use parentheses! Using **minX** and **maxX** from the 5-number summary above may be helpful.

If your lists are missing, enter **STAT**, then **5** to run the **SetUpEditor**. Normally, lists L_1 – L_6 are available. Additional lists may be named. These names are available under **2nd STAT** or **LIST**. It also might be necessary to clear a list (arrow up to the list name and enter **CLEAR**. Elements may be inserted into or deleted from lists by use of the **DEL** or **2nd DEL = INS** key. The author often rescales test scores (which were in L_5) by a command like the following: $\text{ROUND}(.9 * L_5, 0) + 11 \text{STO} \blacktriangleright L_6$.[‡] What does this do?

[‡]The store key is located just above the ON key on most TI-8x calculators. ST0 is not displayed when the key is depressed.

Name _____

Score _____

3.7 Homework, Averages

Problems 1–4 are worth four points each, each subpart of problem 9 and all other problems are worth two points each.

Find the **mean**, **mode**, **median**, and **midrange** for the following four data sets. Please use the statistics mode on your calculator only for the large data set.

1. Fabricated data based on annual earnings of select individuals related to producing this homework assignment: \$36,000, \$360,000, \$3,600,000, \$36,000,000, and \$360,000,000 (math teacher, notebook computer assembler, Netscape[®] programmer, Windows[®] programmer, Bill Gates).
2. Data set with mixed precision: 1, 1.1, 2.7, 3.14, 1.618.
3. Data set with an even number of elements: 1, 2, 3, 4, 5, 6, 7, 8.
4. Data set with lots of data (inauguration ages of U.S. presidents): 57, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 50, 48, 65, 52, 56, 46, 54, 49, 51, 47, 55, 55, 54, 42, 51, 56, 55, 51, 54, 51, 60, 62[§], 43, 55, 56, 61, 52, 69, 64, 46, 54, 47. Please use your graphing calculator and save the data for a few days.
5. Assume four students drive from Michigan to Florida (2000. km) at 100.0 kph and return at 80.0 kph. Find the **arithmetic mean** of these two speeds.
6. Use the arithmetic mean speed from the previous problem and the **total** distance travelled to obtain a false value for how long it took. (Speed is distance divided by time. Hence time is distance divided by speed.)

[§]There is some controversy regarding David Dwight Eisenhower's year of birth. The value used here is from an official web page <http://www.whitehouse.gov/history/presidents/de34.html>.

7. Velocity is displacement per unit time. Using the data from the problems just above, calculate the true value for their **time in hours** for each leg of their journey. Contrast this with the previous problem.

8. Now, divide their total distance travelled by their true total time. This is their true average speed and is the **harmonic mean**. It will be defined and derived in Statistics Lesson 4.

9. Examine lesson 11.7 in your Geometry book and do problems 1, 3a, 5, and 8.
 1. Find the mean temperature if the high was 10°F and the low was -2°F .

 - 3a. Find the second test score needed to average 90 if the first test score was 82.

 5. Find the midpoint of the segment with endpoints $(-0.09, 12)$ and $(0.3, -4)$.

 8. The center of a circle is $(4, 5)$. One diameter endpoint is $(-2, -3)$. Find the coordinates of the other endpoint.

10. Give a formula for the coordinates of the midpoint of the segment connecting (x_1, y_1, z_1) and (x_2, y_2, z_2) .

11. Give a formula for the coordinates of the midpoint of the segment connecting (x_1, y_1, z_1, ict_1) and (x_2, y_2, z_2, ict_2) .[¶]

[¶] c is the speed of light now defined as 299 792 458 m/s, t is time, $i = \sqrt{-1}$, and this is one formulation for space-time in Einstein's general relativity.

Stat's Lesson 4

What Does He Mean?

All the women are strong, all the men are good-looking, and all the children are above average.

Garrison Keillor

In this lesson we focus on the many means available, in addition to the arithmetic mean, why there are so many means, and just what we mean by the word.

4.1 English Father of Modern Science: Francis Bacon

Francis Bacon (1561–1626) developed the Baconian methodology for scientific inquiry or the scientific method. He was also an English philosopher, statesman, and is credited with inventing the English essay. Bacon was not a noted scientist but rather influenced it more by his philosophy and writings—writings he had to pay to have published. His ambition was not just to master all knowledge but to reform it, especially the process by which new knowledge was acquired and integrated.

At Trinity College/Cambridge he developed a dislike for Aristotelian philosophy still being taught which ultimately led him to conclude the methods of science of those times and corresponding results were erroneous. Discovering truth became one of his three life goals, along with service to his country and church. When he could not find work which allowed him to pursue truth, he spent two years studying law. He served in many different capacities within government but was chronically in debt.

Bacon is also often cited as a possible author of Shakespeare's works. The lack of information about Shakespeare himself, the vast vocabulary (29,000 words or 5 times what is used in the *Bible*), and breadth of topics covered point toward a committee, nobility, or well-educated man. Doubts surfaced during Bacon's and Shakespeare's lifetime and continue to this day. Many forms of statistical analysis have been used to address this question.

4.2 Arithmetic Mean

In Statistics Lesson 3, we defined arithmetic mean as the one commonly used by statisticians and as the one usually intended when we just say mean. However, there are a wide variety of other means with a variety of applications which we will review here.

4.3 Geometric Mean

The **geometric mean** is used in business to find average rates of growth. The geometric mean is the n^{th} root of the [pi] product of the data elements.

$$\text{Geometric mean} = \sqrt[n]{\prod x_i} \text{ for all } n \geq 2.$$

Example: Suppose you have an IRA (Individual Retirement Account) which earned annual interest rates of 5%, 10%, and 25%.

Solution: The proper average would be the geometric mean or $(\sqrt[3]{1.05 \cdot 1.10 \cdot 1.25})$ or about 1.13 meaning 13%.

Note that the data elements must be positive. Negative growth is represented by positive values less than 1. Thus, if one of the accounts lost 5%, the proper multiplier would be 0.95.

The geometric mean is typically first encountered in a proportion when the means are equal, as in $\frac{8}{w} = \frac{w}{4}$. Here $w^2 = 32$ and square rooting both sides gives an answer. However, in general, there may be n n^{th} geometric means. We thus cannot be sure of the sign of w above.

4.3.1 Geometric and Arithmetic Sequences

The difference between arithmetic and geometric means is similar to the difference between arithmetic and geometric sequences. In an **arithmetic sequence** (2, 4, 6, 8, ...) you **add** the same amount each time (2). In a **geometric sequence** (2, 4, 8, 16, ...) you **multiply** by the same factor each time (2). If you are given the 1st and 4th term of an arithmetic sequence (1, ?, ?, 10), you can solve for the missing terms by finding the difference of the known terms and dividing the interval by the number of gaps between missing terms: $(10 - 1)/3 = 3$ to find the **common difference** (3) and hence the full sequence: (1, 4, 7, 10). If you are given the 1st and 5th terms of a geometric sequence (2, ?, ?, ?, 32), you can solve for the missing terms by finding the ratio of the known terms and n^{th} rooting for the **common ratio**. Here again, n is determined by the number of gaps between missing terms and known terms. However, since there can be four different 4th roots, there may be up to four different sequences: (2, 4, 8, 16, 32) or (2, -4, 8, -16, 32) and, if complex numbers are allowed,

$(2, 4i, -8, -16i, 32)$ or $(2, -4i, -8, 16i, 32)$, depending on which common ratio or n^{th} root was chosen from among ± 2 and $\pm 2i$.* Hopefully, this motivates the n^{th} roots used above. Also, since fractional exponents are usually new, a section on them is included as Section 4.10 of this lesson.

4.4 Harmonic Mean

The **harmonic mean** is found by dividing the number of data elements by the sum of the reciprocals of each data element.

$$\text{Harmonic mean} = \frac{n}{\sum x_i^{-1}}$$

The harmonic mean is used to calculate average rates such as distance per time, or speed. (In physics you will learn that speed is a **scalar**, whereas velocity is a **vector**, having both magnitude and direction. Great care should be exercised to select the proper term.) Problems requiring the harmonic mean are common on contests.

Example: Suppose your grandfather walked three miles to school. Due to the terrain, for the first mile he averaged 2 mph; for the second mile 3 mph; for the final mile the average speed was 4 mph. What was the average speed for the three miles? **Solution:** The arithmetic mean of $(2+3+4)/3 = 3.0$ mph is incorrect. This would imply it took 1 hour = 60 minutes to walk to school. Breaking it down into the separate components, it takes 30 minutes (1st) + 20 minutes (2nd) + 15 minutes (3rd) to walk (each mile) or 65 minutes total. His actual speed was thus 3 miles/1.083 hour or 2.77 mph.

Another way to show our work would be:

$$\frac{3 \text{ miles}}{1/2 + 1/3 + 1/4} = \frac{3}{13/12} = \frac{36}{13} = 2.77 \text{ mph.}$$

4.5 Quadratic Mean

The **quadratic mean** is another name for Root Mean Square or RMS.

$$\text{Root Mean Square (RMS)} = \sqrt{\frac{\sum x_i^2}{n}}$$

The quadratic mean is typically used for data whose arithmetic mean is zero.

Example: Let's explore US household alternating current (AC). Alternating current typically is a sine wave like given in Figure 4.1. Note the average value of zero. However, instead of ranging from -1 to 1 , it typically ranges between ± 162 V. The

*These can be found by taking: $\sqrt[4]{32/2}$.

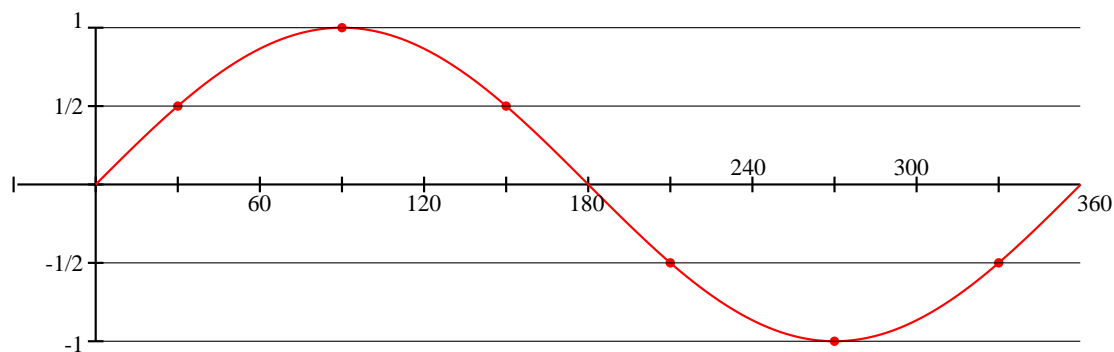


Figure 4.1: Sine Curve—like US Household Current/Voltage.

quadratic mean gives a physical measure of the average distance from zero. Suppose measurements of 120, -150 , and 75 volts were obtained.

Solution: The corresponding quadratic mean is $\sqrt{(120^2 + (-150)^2 + 75^2)/3}$ or 119 volts RMS.

4.6 Trimmed Mean

Trimmed mean usually refers to the arithmetic mean without the top 10% and bottom 10% of the ordered scores.

Technically, this is the **10% trimmed mean**. You could also find the 20% trimmed mean by only forming the mean of the middle 60% of the data. Clearly this removes extreme scores on both the high and low end of the data. Sorting the data is also clearly indicated! Section 3.6 has information about sort lists and deleting items from a list.

4.7 Weighted Mean

Weighted mean is the average of differently weighted scores. It is the sum of the weighted scores over the sum of the weights. It takes into account some measure of weight attached to different scores.

Example: Semester grades are often computed as 40% ($\frac{2}{5}$) of the 1st 9-week grade, 40% ($\frac{2}{5}$) of the 2nd 9-weeks grade, and 20% ($\frac{1}{5}$) for the semester exam. For specifics, assume Martha earned 84% for the first 9-weeks, 89% for the second 9-weeks, but only 60% on the semester exam.

Solution: In such a case, the semester grade could be computed as:

$$\frac{2 \cdot 84\% + 2 \cdot 89\% + 60\%}{5} = 81.2\%$$

Grade Point Averages is another typical example of a weighted mean, because college classes usually (and high school classes sometimes) come in a variety of credit hours. The formula for the weighted mean is given below.

$$\text{Weighted mean} = \frac{\sum w_i \cdot x_i}{\sum w_i}$$

4.8 Combination Mean

Consider if you had \$10,000 earning 6%, \$20,000 earning 12%, and \$25,000 earning 18% annual interest. Clearly some combination of weighted geometric mean would be needed to compute a proper average value![†] A similar example would involve speeds (2 mph, 3 mph, 4 mph) when applied to different distances, such as 4, 3, and 2 miles. The correct mean value would involve some weighted harmonic mean.[‡] Such problems go beyond what students are expected to master here, but may appear on contests or standardized tests.

4.9 Means from a Frequency Table

Frequency mean is the same as obtaining the arithmetic mean from a frequency table. For memory purposes, it is like the weighted mean formula.

An activity (Section 5.7) for finding the mean from a frequency table is included with Statistics Lesson 5.

4.10 Activity: Exponents for Geometric Mean

In order to do some problems in today's assignment, an expanded definition of exponents needs to be developed. Recall from Numbers Lesson 5 the definition and rules for exponentiation as follows.

$$\begin{array}{llllll} x^1 = x & x^2 = x \cdot x & x^3 = x \cdot x \cdot x & x^4 = x \cdot x \cdot x \cdot x & x^5 = x \cdot x \cdot x \cdot x \cdot x \\ \text{and } x^{-1} = 1/x & & x^{-2} = 1/x^2 & x^{-3} = 1/x^3 & x^{-4} = 1/x^4 \end{array}$$

We can extend this to define what x raised to a fractional exponent means by using the fact that when powers with common bases are multiplied, the exponents are added. Square roots were introduced in Numbers Lesson 11.

$$x^{1/2}x^{1/2} = x^{(1/2+1/2)} = x^1 = x$$

[†]It is often easiest to compute the total earnings and divide by the original principle:
 $\frac{10000 \times 0.06 + 20000 \times 0.12 + 25000 \times 0.18}{10000 + 20000 + 25000} = 0.1364$ or 13.6%.

[‡] $\frac{4+3+2}{2+1+\frac{1}{2}}$ miles/hours or $\frac{9}{3.5} = 2.57$ mph.

$$x^{1/3}x^{1/3}x^{1/3} = x^{(1/3+1/3+1/3)} = x^1 = x$$

$$x^{1/4}x^{1/4}x^{1/4}x^{1/4} = x^{(1/4+1/4+1/4+1/4)} = x^1 = x$$

In other words: $x^{1/2} = \sqrt{x}$ $x^{1/3} = \sqrt[3]{x}$ $x^{1/4} = \sqrt[4]{x}$ and
 $4^{1/2} = 2$ $8^{1/3} = 2$ $729^{1/6} = (729^{1/3})^{1/2} = 9^{1/2} = 3.$

We can define a real number x raised to rational roots such as a/b ($x^{a/b}$) to be the b^{th} root of x raised to the a^{th} power $\sqrt[b]{x^a}$. The extension of any real number to any real power goes even beyond Numbers Lesson 16 (we need the limit process from calculus).

Such roots can be calculated on the calculator three ways as follows.

- $729 \wedge 6^{-1}$ where the x^{-1} is used. This seems to be an exception to the general inability of the calculator to process exponents correctly. That can be explained by the fact that the x^{-1} is “bound rather tightly” to the 6. To be on the safe side however, parentheses should be used: $729 \wedge (6^{-1})$, as this could change!
- $729 \wedge (1/6)$
- **6 MATH 5** brings up the symbol: $\sqrt[x]{\quad}$ which you can follow with 729. This $(6 \sqrt[x]{729})$ then gives the 6th root of 729.

Name _____

Score _____

4.11 Homework, Means

Unless otherwise noted, each problem is worth two points.

1. (**Six points**) The population of 37 in the senior class of 2014, 8th grade, Algebra Diagnostic Test scores were stratified into groups of 6 taken from the data set in descending order. Using a die, one from each strata was randomly selected to obtain the following results presented in random order:

52, 127, 71, 103, 64, 87

Find the sample size, sum of the data, mean, mode, median, and midrange.

2. Discuss the sampling technique used in the problem above.
3. Compare the average results obtained in the problems above with those provided for the full population (or did you lose your summary sheet(s)?).
4. (**Four points**) The digits of e have been shown to be very random. Treating each of the first twenty **decimal** (*i.e.* omit the “2.”) digits as a separate data element, calculate the mean, mode, median, and midrange for this sample. (See Numbers Lesson 15 for e .)
5. What would you expect each of these average values to be, if say a million or billion digits of e were used?
6. Show how to calculate the average growth rate for a portfolio with the following consecutive annual interest rates: 5%, 10%, -5%, 20%, 15%.

7. Four students drive from Michigan to Florida (2000 km) at 100.0 kph and return at 80.0 kph. Show how to find the average round trip speed, using the harmonic mean.
8. For the problem just above, what is their average round trip **velocity**?
9. Tom Foolery measures the voltage in a standard outlet as 120 volts, -160 volts, 95 volts, and 10 volts at random intervals. Show how we can calculate the RMS voltage.
10. Calculate the GPA (weighted mean) for the following data: Biology, 5 credits, A $-$ (use 3.667); Chemistry, 4 credits, B $+$ (use 3.333); College Algebra, 3 credits, A (use 4.000); and Health, 2 credits, C (use 2.000); Debate, 2 credits, B (use 3.000). Show your work and express your results to three decimal places.
11. Using the inauguration ages from the previous homework, calculate the 10% trimmed mean and 20% trimmed mean.
12. A researcher finds the average teacher's salary for each state from the web. He then sums them together, divides by 50 to obtain their arithmetic mean. Why is this wrong and what should he have done?
13. Examine lesson 13.3 in your Geometry textbook and do the following problem: 13.3#1. Find the geometric mean of 2 and 50 to the nearest hundredth.
14. Examine lesson 13.3 in your Geometry textbook and do the following problem: 13.3#2. Find the geometric mean of 9 and 12 to the nearest hundredth.

Stat's Lesson 5

Measures of Dispersion

*The average [adult] human has one breast and one testicle, but those standard deviations of one effectively split the population!**

Smoking is one of the leading causes of statistics. Fletcher Knebel

Another important characteristic of a data set is how it is distributed, or how far each element is from some measure of central tendency (average). There are several ways to measure the **variability** of the data. Although the most common and most important is the standard deviation, **which provides an average distance for each element from the mean**, several others are also important, and are hence discussed here. Students should already have some familiarity with the words disperse and dispersion since they are used to describe how, for example, acorns spread out, not only under the influence of gravity, but also by squirrels, deer, *etc.*

5.1 The Father of Mathematical Modelling: Siméon-Denis Poisson

Poisson (1781–1840) was a French mathematician and physicist. He entered the École Polytechnique in Paris in 1798 and within two years had written important papers which led to his involvement with the leading mathematicians of his time. During his career he made important advances in mathematical physics in the fields of electricity, celestial mechanics, and waves. An important probability distribution is named after him. Poisson frequently said: “**Life is good for only two things, discovering mathematics and teaching mathematics.**”

The Poisson distribution applies to independent discrete events occurring in a fixed period with a known average rate. Although Poisson developed it while analyz-

*The first part of the quote should be attributed to Des McHale. The last part and word adult should be attributed to Calkins.

ing the number of Prussian soldiers who died each year from horse-kicks, it has broad applications to such situations as customer arrival times at a bank, typographical errors on a page, and web server accesses. In various limiting cases the binomial distribution can be approximated by the Poisson and the Poisson by the Gaussian. The Gaussian is introduced in the next lesson and other distributions studied sophomore year.

5.2 Range

Range is the difference between the highest and lowest data element.

Symbolically, range is computed as $x_{\max} - x_{\min}$.[†] Although this is very similar to the formula for midrange, please do not make the common mistake of reversing the two. This is not a **reliable** measure of dispersion, since it only uses two values from the data set. Thus, extreme values can distort the range to be very large while most of the elements may actually be very close together. For example, the range for the data set 1, 1, 2, 4, 7 introduced earlier would be $7 - 1 = 6$.

Recently it has come to my attention that a few books define statistical range the same as its more mathematical usage. I've seen this both in grade school and college textbooks. Thus instead of being a single number it is the interval over which the data occurs. Such books would state the range as $[x_{\min}, x_{\max}]$ or x_{\min} to x_{\max} . Thus for the example above, the range would be from 1 to 7 or $[1, 7]$. Be sure you do not say 1–7 since this could be interpreted as -6 .

5.3 Standard Deviation

The **standard deviation** is another way to calculate **dispersion**. This is the most common and useful measure because it is the **average distance** of each score from the mean. The formula for sample standard deviation is as follows.

Sample Standard Deviation:
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Notice the difference between the sample and population standard deviations. The sample standard deviation uses $n - 1$ in the denominator, hence is slightly larger than the population standard deviation which use N (which is often written as n).

Population Standard Deviation:
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

It is much easier to remember and apply these formulae, if you understand what

[†]Some books add one to this result which basically assumes the data are discrete or even integer.

all the parts are for. We have already discussed the use of Roman vs. Greek letters for sample statistics vs. population parameters. This is why **s** is used for the sample standard deviation and σ (**sigma**) is used for the population standard deviation. However, another sigma, the capital one (Σ) appears inside the formula. It serves to indicate that we are adding things up. What is added up are the **deviations** from the mean: $x_i - \bar{x}$. But the **average deviation** from the mean is actually zero—by definition of the mean! Occasionally the **mean deviation**, using average distance or using the symbols for absolute value: $|x_i - \bar{x}|$ is used. However, a better measure of variation comes from squaring each deviation, summing those squares, then taking the square root after dividing by one less than the number of data elements. If you compare this with the formula for quadratic mean (Section 4.5) you will realize we are doing the same thing, except for what we are dividing by. That $n - 1$ can be understood in terms of **degrees of freedom**—a term worth introducing here but the topic goes beyond this introduction and will be covered in Probability and Distributions Lesson 14.

Another formula for standard deviation is also commonly encountered. It is as follows.

Shortcut Formula for Standard Deviation:	$s = \sqrt{\frac{n(\sum x_i^2) - (\sum x_i)^2}{n(n-1)}}$
--	--

This formula can be algebraically derived[‡] from the former and has two primary applications. First, calculators and computer programs often employ it because less intermediate results are necessary and it can be calculated in one pass through the data set. That is, you don't have to calculate the mean first and then find the deviations. Second, it is closely related to a formula which may be used to calculate the standard deviation for a frequency table. For this course, we will rely on our graphing calculators and the appropriate activity is discussed in today's activity (Section 5.7).

5.4 Variance

Variance is closely associated with dispersion, but technically does measure dispersion. Compare the two variance formulae with their corresponding standard deviation formulae, and we see that variance is just the square of the standard deviation. Statisticians tend to consider variance a primary measure and use it extensively (ANOVA, *etc.*), whereas scientists are very happy to use standard deviation exclusively. For official information on uncertainty, please refer to the following National Institute of Standards and Technology web page. Uncertainty is another way to discuss variance and the Heisenberg's Uncertainty Principle (we cannot simultaneously

[‡]Perhaps algebra doesn't teach manipulation of summation symbols, but it can be understood at that level.

know both a particle's momentum and position more accurately than some small multiple of Planck's constant: 6.626×10^{-34} J·s) is at the very root of quantum mechanics.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} \qquad \sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

Occasionally, the abbreviations SD or Std. for standard deviation and Var for variance will be seen.

5.5 Range Rule of Thumb

It can take some time to start to understand how these measures of variation may be useful. One of the reasons we provide mean and standard deviation information regarding tests in this course is to help develop this understanding. Often your test scores will be adjusted via two different methods. Consider the following scenarios. First, if a straight five points are added to everyone's score, the mean would increase five points, say from 70.8 to 75.8 **but have no affect on the standard deviation.** It remains, say, at 10.9. Second, if each test score was multiplied by .89 and then 21 points were added, not only does this move the mean from, say, 55.4 to 70.3, but it also reduced the standard deviation from, say, 15.0 to 13.5. This can be useful if the original test scores were very variable, and could easily have resulted in more D's and F's than your efforts justified. You might consider a third common way to adjust test scores, that of dropping the possible. Technically this doesn't change either the mean or the standard deviation, but it does effectively raise everyone's percentage. This doesn't help the lower scoring students nearly as much as it helps the top students.

A commonly given **rule of thumb**[§] is that the range of a data set is approximately 4 standard deviations ($4s$). Thus the maximum data element will be about 2 standard deviations above the mean and the minimum data element about 2 standard deviations below the mean. We will explore this further in the next lesson (Statistics Lesson 6).

5.6 More Round-off Information

The standard deviation of a data set is often used in science as a measure of the precision to which a experiment has been done. It can also indicate the reproducibility of the result. **Propagation of error** will not be fully discussed here, except to note that **intermediate values in your calculations should not be rounded.** At least twice as many digits as will be used in the final answer should be retained, especially when square roots, exponentials, or logarithms are involved.

[§]See Section 3.4 for a disclaimer regarding the origins of this phrase.

It is rather meaningless to calculate the standard deviation for a data set of two elements.

Three is considered the smallest sample size where standard deviation is meaningful.

It is not uncommon for an experiment to involve millions of events and associated data. If you examine the standard deviation formula above, you will note that it depends inversely on the square root of n ($1/\sqrt{n}$). Thus with a million measurements we could expect to reduce the standard deviation of our answer by perhaps a thousand fold. It is the goal of many experiments to obtain very precise values, so great care is exercised to reduce systematic errors and also reduce the affect of random errors by increasing the repetitions.

Example: Consider a simple example of counting pennies where the outcomes 99, 100 and 101 are obtained. Find the mean and standard deviation.

Solution: We can easily calculate the mean as 100 and the standard deviation as 1.0.

Example: Consider further if this exercise were repeated 1000 times and 100 was obtained 991 times, 99 5 times and 101 4 times. Again, calculate the mean and standard deviation.

Solution: The mean is now 99.999 and the standard deviation is now 0.095. Here the additional precision is justified and the mean and standard deviation are given to the same 3 decimal place precision. It would be a mistake to report these results to only one more digit than the original data set, as in 100.0 and 0.1.

DO NOT USE a rounded s to obtain s^2 .
Variance is the primary statistic, s is a derived quantity.

Standard deviation should be reported to at least two and perhaps three significant digits. Ideally, the mean and standard deviation would have a consistent number of decimal places. They thus may have a very different number of significant digits as illustrated in the penny counting example above.

5.7 Activity: Frequency Means/Standard Deviation

L1	L2	L3	3
55	5		
60	15		
70	20		
75	25		
80	20		
90	12		
95	3		
L3(1)=			

Figure 5.1: TI-84 Graphing Calculator Display of Frequency Data.

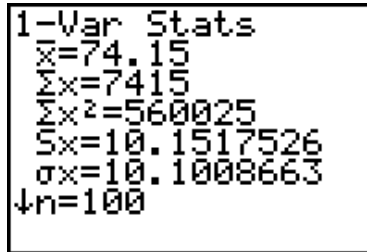
Please use your TI-84+ type calculator for the following activities.

Press the **STAT** key and **ENTER** to select **EDIT**.

In L_1 , enter in these test scores: 55, 60, 70, 75, 80, 90, and 95.

In L_2 , enter in these test frequencies: 5, 15, 20, 25, 20, 12, and 3.

These last values are how many tests there were for each of these scores.



```

1-Var Stats
x̄=74.15
Σx=7415
Σx²=560025
Sx=10.1517526
σx=10.1008663
↓n=100
  
```

Figure 5.2: TI-84 Graphing Calculator Display of 1 Variable Statistics.

Press the **2nd MODE (QUIT)** or just go directly to **STAT** arrow over to **CALC** and **ENTER** to select **1-Var Stats**. Now enter **2nd 1** (L_1), a comma (**,**), and **2nd 2** (L_2) followed by **ENTER**. Your screen should now appear as above.

When doing a frequency mean, the order of the lists is important. You need to place the score list first and then the frequency list. Thus you had **1-Var Stats** L_1, L_2 on your screen and not **1-Var Stats** L_2, L_1 . If you did it the wrong way, you can easily tell if there is an error by looking at the n value. The wrong way gave $n = 525$ instead of the correct value of $n = 100$.

Under the **1-Var Stats**, the arithmetic mean, \bar{x} is listed. **Be sure to always round this to the proper significance.** Below that is also included the sample standard deviation, denoted by a s_x . Notice that both the sample and population standard deviation (σ_x) are given. Earlier in this lesson the difference between the two was discussed. Watch out carefully for which one applies to a given data set. (Remember, standard deviation is a measure of the “average” distance each score is away from the mean.)

One last note is use of the **VARS** key followed with **5** (Statistics), to get s_x to more easily square the standard deviation to obtain the variance. This will facilitate the avoidance of rounding and increase the quality of the variance number obtained.

Name _____

Score _____

5.8 Homework, Dispersion

Problems 1–4 are worth 6 points each, the rest two points each.

Find the **range**, **sample standard deviation**, and **sample variance** for the following four data sets. Please use the statistics mode on your calculator for these four data sets.

1. Fabricated data based on annual earnings of select individuals related to producing this homework assignment: \$36,000, \$360,000, \$3,600,000, \$36,000,000, and \$360,000,000 (teacher, notebook assembler, Netscape[®] programmer, Windows[®] programmer, Bill Gates).
2. Data set with mixed precision: 1, 1.1, 2.7, 3.14, 1.618.
3. Data set with an even number of elements: 1, 2, 3, 4, 5, 6, 7, 8.
4. Data set with lots of data (inauguration ages of U.S. presidents): 57, 61, 57, 57, 58, 57, 61, 54, 68, 51, 49, 64, 50, 48, 65, 52, 56, 46, 54, 49, 51, 47, 55, 55, 54, 42, 51, 56, 55, 51, 54, 51, 60, 62, 43, 55, 56, 61, 52, 69, 64, 46, 54, 47.
5. Sort the presidential inauguration age data on your calculator. Count how many data elements are within one standard deviation of the mean (*i.e.* between[¶] $54.66 - 6.26 = 48.40$ and $54.66 + 6.26 = 60.92$). Convert this to a percentage.
6. Repeat the previous question with two standard deviations instead of one.

[¶]Since this does not indicate strictly between and these are values rounded to three decimal places, we should probably include the endpoints, although this is worthy of debate.

Consider again the **sample** data 1, 1, 2, 4, 7.

7. Explicitly calculate the standard deviation using the first formula given in the lecture.

8. Explicitly calculate the standard deviation using the shortcut formula given in the lecture.

9. One of the rounding rules listed was that if you are going to take a square root you can only retain half as many significant figures as was in the radicand. This has significance related to standard deviations. Find the square root of 5 and round the result to three significant figures. Now square this rounded root and also the rounded root ± 0.01 . Compare the three answers in terms of how a 4.5 parts per thousand (ppk) change (0.01/2.236) had an x ppk affect on the result. Find x .

10. **Bonus:** Derive the shortcut formula for the standard deviation from the first formula given.

Stat's Lesson 6

The Bell-shaped, Normal, Gaussian Distribution

The death of one man is a tragedy.

The death of millions is a statistic.

Joe Stalin

I think there is a world market for about 5 computers.

Tom Watson

This lesson introduces the gold-standard of distributions, the Gaussian Distribution. We give its other names, its defining formula, how it is normalized, non-standardized, and how much area is under various parts, *i.e.* the empirical rule. We strengthen the empirical rule with Chebyshev's Theorem. We conclude with sections on the meaning of normal and a short summary of the various types of distributions to be studied in the future.

6.1 The Father of Russian Mathematics: Chebyshev

Pafnuty Chebyshev (1821–1894) was the founding father of Russian mathematics who worked aggressively with prime numbers and other number theory. His writings covered a wide range of mathematics, including probability. We present below a probability theorem named after him which is applicable to any data set, not just normally distributed ones.

The spelling of Chebyshev's name, being transliterated from Russian, is highly variable, with Chevychov and Tschebyscheff being among the many possibilities. An unknown handicap caused Chebyshev to limp and walk with a stick. He thus played few children's games, devoting himself instead to building machines. His scientific achievements were also recognized.

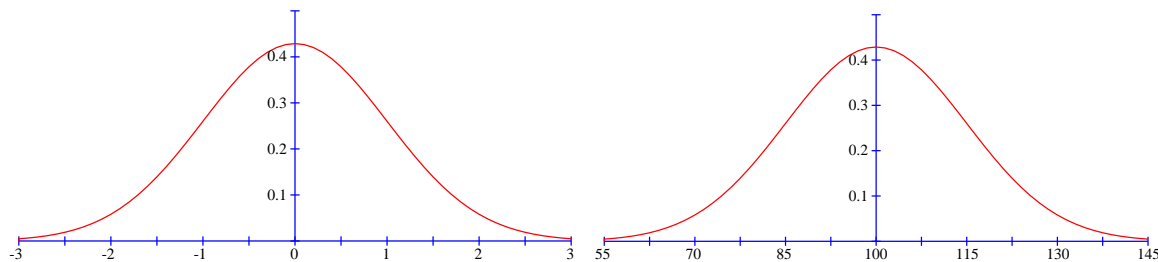


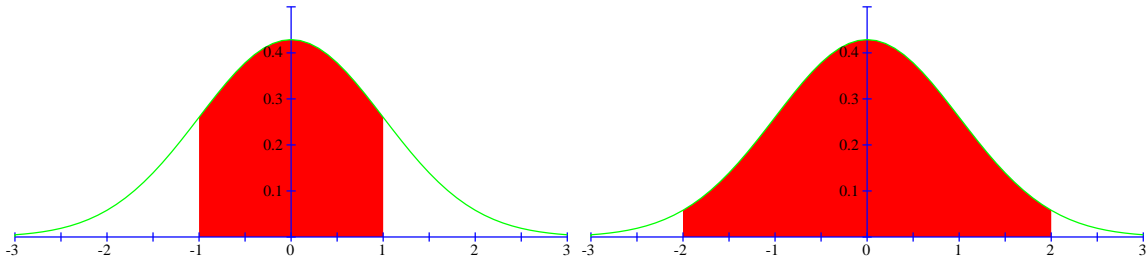
Figure 6.1: The Standard Normal Distribution (left) and Normally Distributed IQs (right).

6.2 The Bell-shaped, Normal, Gaussian Distribution

It can be shown under very general assumptions that the distribution of independent random errors of observation takes on a **normal** distribution as the number of observations becomes large. Although others were involved, Gauss was one of the first to characterize this distribution and hence it is often named after him. It is also shaped like a bell, hence yet another name. The term used in the title above is rather redundant, but serves to emphasize that the three are identical. The name **error curve**, is also possible. You can graph this curve on your calculator as seen in Figure 6.2 by entering the following function: $y = e^{-x^2/2}/\sqrt{2\pi}$ where e is the transcendental number $2.71828\cdots$ and π is the more familiar, but also transcendental number $3.14159\cdots$. The π in the formula only serves to **normalize** the total area under the curve. When we normalize something, we make it equal to some **norm** or standard, usually one (1). The word normal has several other meanings, including perpendicular and the usual/status quo which we discuss in Section 6.5.

The height of the curve represents the probability of the measurement at that given distance away from the mean. The total area under the curve being one represents the fact that we are 100% certain (probability = 1.00) the measurement is somewhere. Technically, this is the **standard normal** curve which has $\mu = 0.0$ and $\sigma = 1.0$. Other applications of the normal curve do not have this restriction. For example, intelligence has often been cast, albeit controversially, as **normally distributed** with $\mu = 100.0$ and $\sigma = 15.0$. This is represented below. Our function has been modified to $y = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$.

Other things which may take on a normal or quasi-normal distribution include body temperature, shoe sizes, diameters of trees, *etc.* It is also important to note the **symmetry** of the normal curve. Some curves may be slightly distorted or truncated beyond certain limits, but still primarily conform to a “heap” or “mound” shape (see below). This is often an important consideration when analyzing data or samples taken from some unknown population.

Figure 6.2: Data Within 1σ (left) and 2σ (right).

6.3 The Empirical Rule

For a normally distributed data set, the **empirical rule** states that 68% of the data elements are within one standard deviation of the mean, 95% are within two standard deviations, and 99.7% are within three standard deviations. Graphically, this corresponds to the area under the curve as shown in Figure 6.3 for 1 and 2 standard deviations. The empirical rule is often stated simply as **68-95-99.7**. Note how this ties in with the range rule of thumb, by stating that 95% of the data usually falls within two standard deviations of the mean.

The author usually claims an IQ of at least 145. We can see from the above information that this would put him at least three standard deviations above the population mean ($100 + 3 \times 15 = 145$). Hence, if we accept the hypothesis that IQs are normally distributed, at least 99.85% of the population would have a lower IQ and less than 0.15% a higher one. Please especially note that if 99.7% of the population is within three standard deviations of the mean, the remaining 0.3% is distributed with half beyond three standard deviations below the mean and the other half beyond three standard deviations above the mean. This is a result of the symmetry (due to the fact that x is squared, it matters not if it is positive or negative) of the curve. In practical terms, in a population of 300,000,000; 299,550,000 would have an IQ lower than 145 and 450,000 would have an IQ higher. Because of the small area of these regions, they are often referred to as **tails**. Depending on the circumstances, we may be interested in **one tail** or **two tails**.

Several societies exist which cater to individuals with high IQs. Some specific examples would be MENSA, Triple Nine, Mega, *etc.*

Another important characteristic of this distribution is that it is of **infinite extent**. In practical terms, IQs below 0 (-6.67σ) or above 210 (7.33σ) (ceiling scores such as Marilyn Vos Savant's are difficult to interpret) do not occur. A recently popularized manufacturing goal has been termed Six Sigma. Interpreted as $\pm 6\sigma$, one would think this would correspond with about 2 defects per billion, but their web site implies it is 200 per million. A typically good company operates at less than plus or minus four sigma or 99.994% perfect. This corresponds closer to 63 defects per million. If your

family has ever purchased a “lemon”* you can appreciate such striving for perfection. (Another source implies six sigma refers to $\pm 3\sigma$.)

Other similar examples would be the large increase in errors related to prescription drugs being dispensed or the case of the Florida patient who had the wrong leg amputated.

6.4 Chebyshev's Theorem

Chebyshev's theorem states that the portion of any set of data within K standard deviations of the mean is always at least $1 - \frac{1}{K^2}$, where K may be any number greater than 1.

For $K = 2$, we see that $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}$, which is 75% of the data must always be within two standard deviations of the mean.

For $K = 3$, we see that $1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9}$, which is about 89% of the data must always be within three standard deviations of the mean.

If we consider the data set 50, 50, 50, and 100, we will discover that the sample standard deviation (s) is 25, and the upper score falls exactly at $2s$ above the rest. However, since the mean is 62.5, it is well within $2s$. Added 5 more scores of 50 we find the mean is now 55.6 and the standard deviation now 16.7. We see that two standard deviations above the mean now extends to 88.9 and we have one data point outside that, but within three standard deviations.

The general concept of being able to find the mean of a data set and determine how much of it is within a certain distance (number of standard deviations) of the mean is an important one which we will continue in the next lesson.

Note: here is an example of a data set with $k = 2$ and only 75% of the data within the proscribed limits.[†] Let the discrete random variable x have probabilities $1/8, 6/8, 1/8$ at the points $x = -1, 0, 1$ respectively. $\mu = 0$ and $\sigma^2 = 1/4$. If $k = 2$, then $1/k^2 = 1/4$ and we thus attain the bound given by Chebyshev's inequality.

6.5 Meanings of Normal

The word **normal** is used extensively in math and science and generally has a very context specific meaning which differs from what is “normally” encountered. We list here several.

*A colloquialism for bad car, perhaps one built on a Monday.

[†]It comes to us from Hogg and Craig (1978, p. 70) in *Introduction to Mathematical Statistics*. (5th ed.) via the AP STAT list server on May 31, 2000.

1. Normal in mathematics and physics often means **perpendicular** to. A normal vector is perpendicular to a given line or plane. This is fact is the original meaning from the Latin word *norma* which meant carpenter's square.
2. Normal can refer to the fact that the **area** has been made **equal to one** (to normalize) so that area and probability are equivalent. This was likely the second meaning to develop for something built with a reference such as a carpenter's square.
3. Normal in statistics generally refers to the **gaussian** distribution or the "normal" way we would expect errors to be distributed. De Moivre invented it but others named it much later. Pearson popularized the name.
4. Normal in chemistry refers to the **molarity** or concentration of an acid or base. Specifically, these are related by normality equals the molarity times the number of equivalents, where equivalents refers to number of solutes. For acids like HCl and basics like NaOH the numerical values of normality and molarity are equal. However, H_2SO_4 has twice the normality for a given molarity. There are additional chemical meanings of normal related to hypothetical compounds (dehydrated acids) and unbranched hydrocarbons.
5. Of course, normal in everyday usage now tends to refer to the **ordinary** or usual, the status quo.

6.6 Other Distributions

There are many ways a data set may be distributed. The study of these ways will take up a fair section of our statistical studies next year. Of particular importance are the following: uniform distribution, binomial distribution, hypergeometric distribution, Poisson distribution, Lorentzian distribution, Student *t*-distribution, F distribution, and Chi-square distribution.

6.7 Quiz over Statistics Lesson 5

Use the sample data: 31, 32, 32, 34, 35, 43, 24, 13, 19, 23, 23, 45, 13, 13, 54, 45, 12, 75, 23, 46, 54, 87, 12, 45, 78 to answer the following questions.

1. Make a stem and leaf diagram.
2. Mean.
3. Mode.
4. Median.
5. Midrange.
6. Q1.
7. Q3.
8. Standard deviation.
9. Variance.
10. Range.

Name _____

Score _____

6.8 Homework, Normal Curve

Each problem is worth two points. (30 points possible with 3 more bonus points.)

1. Find the mean and standard deviation for the data set given in Figure 6.3.

Profession	Annual Earnings	frequency
Math Teacher	36,000	1,000,000
notebook assembler	360,000	100,000
Netscape [®] programmer	3,600,000	100
Windows [®] programmer	36,000,000	10
Bill Gates	360,000,000	1

Figure 6.3: Fictitious Salary Data Illustrating Use of Frequency.

2. Trim 10% of the data from both the top and bottom of the Figure 6.3 data set and repeat the problem above.
3. Show how to apply the symmetry of IQ distribution and the empirical rule (68–95–99.7) to find the proportion of a population with an IQ between 85 and 130.
4. What does Chebyshev's Theorem say about the number of IQs between 85 and 115?
5. The Unabomber (Theodore Kaczynski) has been often cited with an IQ of 167. Show how and calculate how many standard deviations above the mean this corresponds to. Round your answer to two decimal places.
6. Using the mean of 54.66 and the standard deviation of 6.26, list the inauguration ages for any president beyond two standard deviations from the mean.
7. What percent of the inauguration data is within two standard deviations of the mean?

8. Assume the inauguration data was presented in an earlier homework in chronological order. Which ordinal (first, second, third, *etc.*) presidents do the unusual scores above correspond with?

For one bonus point each, please supply the names of the unusual presidents in the previous problem.

9. Add five years ($L_1 + 5$ STO ► L_2) to your presidential inauguration data and recompute the mean and standard deviation.[‡] How did they each change?
10. Increase your original presidential inauguration data by 10% ($L_1 \times 1.1$ STO ► L_2) and recompute the mean and standard deviation. How did they each change?
11. Add 5 years then increase your original presidential inauguration data by 10% ($(L_1 + 5) \times 1.1$ STO ► L_2) and recompute the mean and standard deviation. How did they each change?
12. Increase your original presidential inauguration data by 10% then add 5 years ($L_1 \times 1.1 + 5$ STO ► L_2) and recompute the mean and standard deviation. How did they each change?
13. Use your TI-84+ calculator to find more precisely the percentage of data expected between -1.0000 and 1.0000 standard deviations from the mean. Use **DISTR** (2nd **VARS**) **normalcdf**($-1.0000, 1.0000$).
14. Repeat the question above for -2.0000 and 2.0000 standard deviations.
15. Repeat the question above for -3.0000 and 3.0000 standard deviations.

[‡]The STO or store key is located just above the ON key on most TI-8x calculators. STO is not displayed when the key is depressed.

Stat's Lesson 7

Measurements of Position

Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage people, you should be perfectly comfortable.

Bobby Bragan

Normal distributions are very common, but they often have different means and different standard deviations. A standard score or z -score is a way of relating scores from one normal distribution to another. Ordinary and unusual scores are discussed. Quartiles, deciles, and percentiles are introduced with an emphasis on outliers. We conclude this lesson with the 5-number summary.

7.1 Father of Statistical Genetics: Sir R. A. Fisher

Fisher (1890–1962) was an English statistician and geneticist who had a profound influence on the way the field of statistics developed, especially as it applies to biology. He is described as “a genius who almost single-handedly created the foundation for modern statistical science.” Fisher pioneered the design of experiment, analysis of variance, the technique of maximum likelihood, and began the field of non-parametric statistics. He developed ideas on sexual selection, mimicry, and the evolution of dominance. Several statistical tests and the F distribution are named after him.

7.2 Standard or z -Scores

Several times in Statistics Lesson 6, we calculated how far, in standard deviations, a data element was from the mean. This is a very widely used procedure and this measure has the name z -score. It is also termed a **standard score**. Since many data sets have a heap-shaped, if not somewhat normal distribution, it is a very helpful way to compare data elements from different populations—populations which may very well have differing means and standard deviations.

A typical example might be ACT and SAT scores. ACT scores range from 1 to 36 with a national mean of about 21.0 and standard deviation of about 4.7. (The four sections each range from 1 to 36 but are averaged.) SAT scores range from 200 to 800 (for each subtest) with a national mean of about 508 and standard deviation of about 111. Both ACTs and SATs appear to be approximately normally distributed. Math and Science Center students often take both, perhaps several times and would represent a sample. This sample would have its own mean and standard deviation, but of course, these would be statistics, not parameters. (Specifically, our Math and Science Center students average about 1050 (two section total) when they take the SAT their eighth grade year and average over 1300 (two section total) when they take it their junior year. Our average ACT score (junior) is about 29. Information about the three section (math, English, and essay) is pending. Note that historically the Math and Science Center uses SAT scores as 40% of our admission criterion. The math section now contains “higher math” (precalculus) and an alternative was investigate. However, no change was made. Also of note is the new procedure here in Michigan to replace the high school MEAP with the ACT starting with the class of 2008.)

The formulae used for z -score appear in two virtually identical forms, recognizing the fact that we may be dealing with sample statistics or population parameters. These formulae are as follows.

$$z = \frac{x - \bar{x}}{s} \qquad z\text{-score formulae} \qquad z = \frac{x - \mu}{\sigma}$$

The following important attributes should be noted about z -scores.

Negative z -scores indicate a data element's position below the mean.

Positive z -scores indicate a data element's position above the mean.

z -scores should [almost] always be rounded to two decimal places.

For the IQs of 0 and 210 referred to in Statistics Lesson 6, z -scores of -6.67 and 7.33 should be obtained respectively, based on a population mean of 100 and a standard deviation of 15.

The population does not have to be normally distributed to calculate z -scores, but that is one of its primary applications.

In summary, z -scores provide a useful measurement for comparing data elements from different [heap-shaped] data sets.

7.3 Ordinary or Unusual Scores

Now that we have defined z -score, we can define two more terms as follows.

Data elements more than 2 standard deviations away from the mean are termed **unusual**.

Data elements less than 2 standard deviations away from the mean are termed **ordinary**.

As you will recall, in a normally distributed population, 95% of the data will then be ordinary, so only 5% can be unusual. Chebyshev's Theorem guarantees at least 75% of the data to be ordinary, so no more than 25% can be unusual.

7.4 Quartiles

Yet another method of measuring how a data set is distributed is to extend the concept of median and use smaller and smaller divisions. The first division we will examine is the **quartile**.

Note first how the median divides a population into two halves: a **top half** and a **bottom half**.

The top half consists of those data elements above the median, whereas the bottom half consists of those data elements below the median. If we subdivide each of these halves yet again, we have quartered the population and each of these division points is a quartile. Although one might occasionally speak of the **bottom quartile**, **top quartile**, *etc.*, the term quartile technically refers to the three division points and not to the four divisions of the data.

Q_1 is the term used for the median of the bottom half.

Q_3 is the term used for the median of the top half.

Q_2 is another term used for the median.

The precise definition specifies that at least 25% of the data will be less than or equal to Q_1 and at least 75% of the data will be less than or equal to Q_3 . For this introduction, we will follow the conventions for calculating Q_1 and Q_3 of the TI-84+ graphing calculator, but note a similar term below under hinges.

All these measures of position assume the data is quantitative and can be put in numeric order.

Data are **ranked** when arranged in [numeric] order.

Since range is sensitive to **outliers** (defined below), sometimes the **interquartile**

range is calculated. This range is the difference between the third and first quartiles: $Q_3 - Q_1$. It is another measure of dispersion. Other common terms include: **semi-interquartile range**, $(Q_3 - Q_1)/2$, another measure of dispersion, and **midquartile** or $(Q_1 + Q_3)/2$, which is a measure of central tendency (an average).

7.5 Hinges; Mild and Extreme Outliers

Another common term is **hinge**. There is a **left hinge** and a **right hinge**. Their definition is so close to that of Q_1 and Q_3 , respectively, that for this introduction, that is what we will use. In fact, many books and software packages do not differentiate, but since some do, they are defined here. Those which do differentiate, establish the convention of replicating the median, thus including it in both halves when finding the hinge. Thus, a different value for hinge than for quartile might be found in a data set with 10 elements as illustrated in an example below.

The **upper hinge** is the median of the upper half of all scores, including the median.

The **lower hinge** is the median of the lower half of all scores, including the median.

Outliers are extreme values in a data set. Sometimes the term outlier is applied to unusual values as defined above (Triola, 5th edition). More recently, outliers are defined in terms of the hinges or quartiles. Outliers are often differentiated as **mild** or **extreme** as defined below. The interquartile range or perhaps $D = \text{upper hinge} - \text{lower hinge}$ is used. Generally, an outlier should be obvious and not borderline—right next to another element, but lying just outside some arbitrary line of demarcation.

Mild outlier are $1.5D$ to $3D$ beyond the corresponding hinge.

Extreme outlier are beyond the corresponding hinge by more than $3D$.

Example: Find any outliers in the data set: 0, 2, 4, 5, 6, 3, 6, 1, 1, 50.

Solution: Obviously, 50 is a much larger number than any of the other elements. This outlier will cause the mean and variance to be much higher. Specifically, without 50, the mean is 3.1 and standard deviation 2.3, whereas with 50, the mean is 7.8 and standard deviation 15.0. Note that the quartiles are 1 and 6, whereas the hinges are 1.5 and 5.5 for the unmodified data set. For any of these definitions, 50 is way away from the other data and is an outlier. Outliers might be legitimate data values or errors. This 50 might really have been 5.0 and was miscoded (historically, punch card input was column sensitive) or poorly recorded in a lab book, with the decimal point extremely light or missing. 50 may also represent extreme extra credit on a 5 point quiz! It is not unusual to be tempted to omit such data values. It is not considered a good practice, but if such are omitted, be sure to clearly record that fact. You will have just crossed the line between objective and subjective science. It doesn't take

many such changes to skew results to fit a preconceived notion!

In bivariant data, a data point may be termed **influential** if it has an inordinant affect on the slope or intercept of a best fit (regression) line.

7.6 Deciles

Although not nearly as common as **percentiles** which follow below, **deciles** are yet another **fractile** which serve to partition data into approximately equal parts. Hence, just as there are three quartiles which divide a population into four parts, so too are there nine deciles dividing the population into ten parts. The deciles are termed D_1 through D_9 .

D_5 is another name for the median.

7.7 Percentiles

Percentiles are also like quartiles, but divide the data set into 100 equal parts. Each group represents 1% of the data set. There are 99 percentiles termed P_1 through P_{99} .

P_{50} is yet another term for median.

Other equivalents, such as $P_{25} = Q_1$, $P_{75} = Q_3$, $P_{10} = D_1$, *etc.*, should also be obvious. Once again, the term percentile technically refers to the 99 division points, but is not uncommonly used to refer to the 100 divisions.

There is no such thing as P_{100} .

For large data sets, one can calculate the **locator** L to help find a requested percentile. It is computed as follows.

$$L = \left(\frac{k}{100}\right)n \qquad \text{Percentile Locator Formula}$$

k is the percentile being sought and n , of course, is the number of elements in our data set. Usual conventions dictate that once L is obtained, it must be checked to see if it is a whole number. If it is a whole number, the value of P_k is the mean of the L^{th} data element and the next higher data element. If it is not a whole number, L must be **rounded up** to the next larger whole number. The value of P_k is then the L^{th} data element, counting from the lowest.

There is an essential difference between rounding up and rounding off. If we round **off** π we get 3. Whereas, if we round **up** π we get 4.

One last measure of dispersion is the 10 – 90 **percentile range** which is defined to be $P_{90} - P_{10}$.

7.8 5-Number Summary

Another useful summary for a data set is known as a **5-number summary**. We have already seen the middle three members as the quartiles. The other two members, the minimum and maximum, were used earlier to calculate the range. These should be presented in ascending order. If the lower and upper hinges are defined differently from the quartiles, they should be used instead of Q_1 and Q_3 in a 5-number summary. As seen in the activity in Section 3.6, your TI-84+ graphing calculator easily provides you with a 5-number summary.

Name _____

Score _____

7.9 Homework, Measures of Position

Each problem is worth two points, unless otherwise noted. (There are 26 regular points and 3 bonus points.)

1. Graduating Math and Science Center students have a mean ACT score of 29. Calculate the z -score for their mean relative to the national mean of 21.0 and standard deviation of 4.7.
2. Graduating Math and Science Center students have a mean SAT score (math plus verbal) of 1320. Calculate the z -score for their mean relative to the national mean of 1016 and standard deviation of 157. (Note: this standard deviation was derived by quadratically combining the standard deviations of the math and verbal subtests—multiplying 111 by the square root of two.)
3. Given the fact that 50% of a normally distributed data set is within 0.675 standard deviations of the mean, estimate Q_1 , Q_3 , and the interquartile range for Center Senior ACT scores, given also a mean of 29 and standard deviation of 3.0. Would an ACT score of 36 be **unusual** for a Center student?
4. (**Five points:**) Calculate the 5-number summary (using your TI-84+ calculators) for the data set in Figure 7.1. See the activity in Section 5.7 on how to obtain this.

Profession	Annual Earnings	frequency
Math Teacher	36,000	1,000,000
notebook assembler	360,000	100,000
Netscape [®] programmer	3,600,000	100
Windows [®] programmer	36,000,000	10
Bill Gates	360,000,000	1

Figure 7.1: Fictitious Salary Data Illustrating Use of Frequency.

5. (**Four points:**) Calculate the z -score for the largest value in the Figure 7.1 data set. Is it an ordinary score? Is it an outlier? Which definition works best?

6. Using the data set: $\{0, 2, 4, 5, 6, 3, 6, 1, 1, 50\}$, as given in the lesson, calculate the left/lower and right/upper hinge.
7. Using the data set: $\{0, 2, 4, 5, 6, 3, 6, 1, 1, 50\}$, as given in the lesson, calculate its 5-number summary, using the quartiles.
8. (**Three points:**) Using the data set of the two previous problems, check if 50 is an outlier three different ways as follows.
- (a) Using the hinges and $D = \text{upper hinge} - \text{lower hinge}$.
 - (b) Using the interquartile range $= Q_3 - Q_1$ for D .
 - (c) Using the older definition of being more than 2 standard deviations from the mean.

Show all your work.

9. How low would the outlier have to be to be only 2.0 standard deviations above the mean, assuming all other numbers stayed the same?
10. Round **up** the number e to the appropriate integer.
11. (**Three bonus points:**) Using the fifty 1999 class of 2003 Algebra Diagnostic Test scores: 140, 122, 119, 99, 92, 90, 90, 88, 85, 82, 82, 81, 80, 80, 77, 74, 74, 73, 72, 71, 70, 70, 69, 69, 69, 68, 68, 68, 67, 66, 64, 64, 62, 60, 59, 59, 58, 58, 56, 56, 56, 55, 54, 53, 53, 50, 47, 35, 32, find P_{10} , P_{90} and the 10–90 percentile range. Show all your work.

Stat's Lesson 8

Summarizing and Presenting Data

*Then there is the man who drowned crossing
a stream with an average depth of six inches.*

W. I. E. Gates

There are a wide variety of ways to summarize and present data. Most of the common methods will be summarized here, along with the usual conventions and terms for each.

8.1 The Father of Exp. Data Analysis: John Tukey

Tukey (1915–2000) was an American statistician. Tukey's undergraduate and masters training was in chemistry at Brown, but he did his Ph.D. in mathematics at Princeton. He divided his time between Princeton University and AT&T Bell Labs. He was awarded the IEEE Metal of Honor “for his contribution to the spectral analysis of random processes and the fast Fourier transform (FFT) algorithm” (which is now used extensively).

Tukey's 1977 book *Exploratory Data Analysis* introduced the box plot. He also popularized stem-and-leaf diagrams. Tukey is credited with inventing the computer term bit and perhaps miscredited with inventing the term software. A statistical test, distribution, method, and lemma all bear his name.

8.2 Frequency Tables

A **frequency table** lists in one column the data categories or **classes** and in another column the corresponding frequencies.

A common way to summarize or present data is with a standard frequency table as seen in the salary data Figures 6.3 and 7.1. **Frequency** refers to the number of times each category occurs in the original data. Another example containing current Center student distribution is given in Figure 8.1.

Grade	Frequency
9 (freshmen)	30
10 (sophomores)	30
11 (juniors)	24
12 (seniors)	26

Figure 8.1: Frequency Table of Center Students by Grade Level.

Test Score	Frequency
0 – 19	2
20 – 39	11
40 – 59	9
60 – 79	11
80 – 99	8
100 – 119	7
120 – 139	2

Figure 8.2: Frequency Table of 1998 Algebra Diagnostic Test Scores.

Often, the category column will have continuous data and hence be presented via a range of values. In such a case, terms used to identify the class limits, class boundaries, class widths, and class marks must be well understood. For the following examples, use the data given in Figure 8.2.

Class limits are the largest or smallest numbers which can actually belong to each class.

For this example, the class limits are as displayed above in the left table column. For the largest class they are 120 and 139.

Each class has a **lower class limit** and an **upper class limit**.

Class boundaries are the numbers which separate classes. They are equally spaced halfway between neighboring class limits.

For this example, the boundaries would be $-0.5, 19.5, 39.5, 59.5, 79.5, 99.5, 119.5,$ and 139.5 . Note that $19.49999\dots$ is another name for and identical with $19.50000\dots$.

Class marks are the midpoints of the classes.

For this example, the class marks are $9.5, 29.5, 49.5, \dots$. It may be necessary to utilize class marks to find the mean and standard deviation, *etc.* of data summarized

in a frequency table.

Class width is the difference between two class boundaries (or corresponding class limits).

For this example, the class width is 20.0.

Following are guidelines for constructing frequency tables.

- The classes must be “mutually exclusive”—no element can belong to more than one class.
- Even if the frequency is zero, include each and every class.
- Make all classes the same width. (However, open ended classes may be inevitable.)
- Target between 5 and 20 classes, depending on the range and number of data points.
- Keep the limits as simple and as convenient as possible (multiple of width?).

If your limits are not immediately obvious based on the data, try to find an appropriate width by rounding up the range divided by the number of classes. Your lower limit should be either the lowest score, or a convenient value slightly less. Avoid irrelevant decimal places. Large data sets justify having more classes. One published guide is: number of classes = $1 + \log_2 n$. This gives you 5 classes for small data sets of 12 to 22 elements and 10 classes for larger data sets of 362 to 724 elements. The seven classes used above for 50 elements is right on target. **It is not uncommon to omit empty classes**—be alert for such guideline violations! Omitted classes do not change the class width, but can be a real source of confusion!

Relative frequency tables contain the relative frequency instead of absolute frequency.

Relative frequencies can be expressed either as percentages or their decimal fraction equivalents.

Cumulative frequency tables contain frequencies which are cumulative for subsequent classes.

In a cumulative frequency table, the words **less than** usually also appear in the left column.

8.3 Histograms

The term histogram comes from the Greek words meaning **web** and **write**. As such it is a way to untangle data. Another name for a histogram is a **bar graph** or

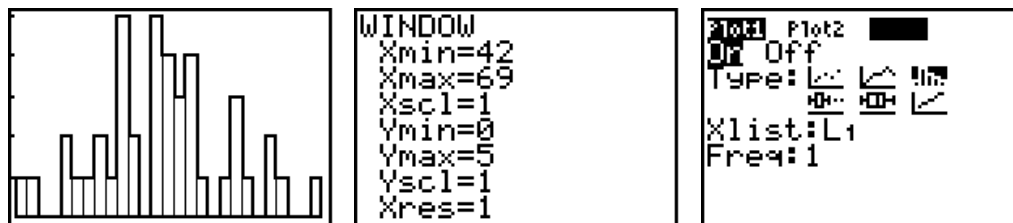


Figure 8.3: TI-84 Bar Chart and Settings.

bar chart, although some texts differentiate between the two. In a histogram the vertical axis has the frequency, while the horizontal axis has the intervals. No gaps are allowed between the bars. The distribution of the data: normal, skewed left, skewed right, should be fairly obvious from a bar graph. Histograms are quite commonly used to visually display frequency and relative frequency charts. Again, some texts indicate that a bar graph is used for categorical data and allows gaps between the bars. Illustrated in Figure 8.3 are a bar graph and the accompanying TI-84+ settings for the US presidential inauguration data.

A **relative frequency histogram** has the same shape and horizontal scale as a histogram, but the vertical scale is now the relative frequency.

A **Pareto chart** is a bar graph for qualitative data.

The bars in a pareto chart should be arranged in descending order of frequency, from left to right.

Frequency polygons are similar to histograms, but use line segments to connect the points.

When construction a frequency polygon, the class marks should be used on the horizontal scale. The graph should also be extended to the left and right so that it begins and ends with a frequency of 0.

Cumulative frequency polygons, also known as **ogives**, are also commonly encountered.

The line in an ogive (pronounced “ōh-jīve”) will always have nonnegative slope.

8.4 Pie Charts, Pictographs, etc.

Pie charts (**circle graphs**) are a common way to understandably display the relative proportions of the various data elements. This is most commonly used on unranked or qualitative data. If this is done by hand, you should use a protractor to accurately measure your angles. Remember that there are 360° in a full circle. Use proportions to convert relative frequencies (x) to angle in degrees: $x\%/100\% =$

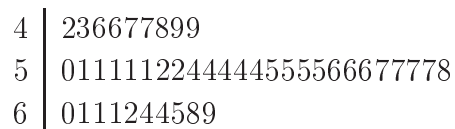


Figure 8.4: Stem and Leaf Diagram for Presidential Inaugural Data.

degrees/360°.

A **pictograph** depicts data by using pictures of an object, such as coins, money bags, airplanes, *etc.* Those which use multiple objects the same size are ok. Those which use similar objects, scaled linearly to represent data, can easily distort things. There may be many other variations, but those listed above are most common.

8.5 Exploratory Data Analysis

A recent trend in statistics has been the use of **exploratory data analysis**. It is a fundamentally different approach. Historically, statistics were used to confirm final conclusions about data. Some very important assumptions were made, calculations were complex, and graphs often unnecessary. The modern emphasis has been more on exploring data, trying to simplify the way the data are described, and gain deeper insights into its nature. Few assumptions are made, the calculations are simple, as are the graphs. The following plot types are modern in their approach.

8.6 Stem-and-Leaf Diagrams

A **stem-and-leaf diagram** has the advantage of retaining the data in its original form, but providing a visual representation. Illustrated in Figure 8.4 is the US Presidential Inauguration data. In this case, the **stem**, the tens portion of the president's age, is given on the left, and the **leaf**, the units portion of the president's age, is given on the right. Some texts advocate included a key which explains this concept. If the purpose is to present the data, that might be well and good. However, this tends to run counter to the original purpose of stem-and-leaf diagrams, to explore the data, so including a key will not be encouraged here.

The following rules should be observed when constructing stem-and-leaf diagrams.

- The leaves on the right should generally be in increasing (or decreasing) order, left to right. Some term it then an **ordered stem-and-leaf diagram**.
- No commas should appear on the right.
- If the stem/leaf break occurs at a decimal point, put the decimal point to the left with the stem.

```

4 | 23
4 | 6677899
5 | 0111112244444
5 | 555566677778
6 | 0111244
6 | 589

```

Figure 8.5: Stem & Leaf Diagram for Presidential Inaugural Data—Split Stems.

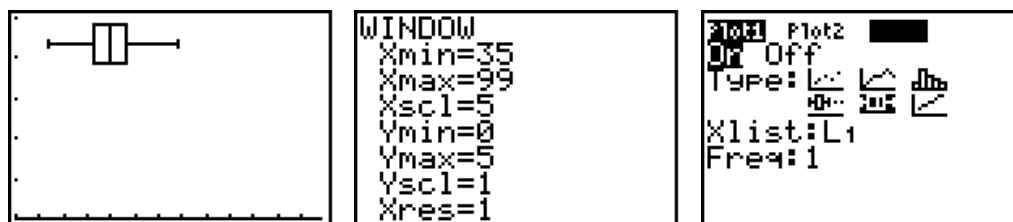


Figure 8.6: TI-84 Box-plot and Settings.

- If the leaf is double or triple digit, *etc.*, leave a [half] space between each entry.
- There should be at least five but no more than twenty rows.
- If a range is used for the stem, an asterisk (*) may be used to separate the corresponding leaves.

Reformatting the stem and leaf diagram as in Figure 8.5 with more rows (called by some books **splitting the stem**) emphasizes even more its normally distributed nature. Notice how the stem-and-leaf diagram is also somewhat like a histogram, but turned on its side. Normally, data are rounded before being put into such a diagram, but ages, for whatever reason, usually get truncated!

8.7 Boxplots

A **boxplot** or **box and whiskers plot** is a visual representation of the 5-number summary. The diagram is a quick way to spot skewed data. Illustrated in Figure 8.6 is a boxplot from the TI-84+ graphing calculator, along with the window and other settings for the US Presidential Inauguration data.

Please note that you can press **TRACE** and obtain the 5-number summary of: 42, 51, 55, 58, 69. The whiskers extend from either 1.5 inner quartile range above and below the quartiles or from the minimum to maximum values. The former is termed a **modified box plot** and will have outliers individually plotted via a symbol of your choice. They also can be traced. You may want to try the data given in Statistics Lesson 7 illustrating outliers.

Name _____

Score _____

8.8 Homework, Presenting Data

Each problem is worth 3 points.

1. Create a pie chart for the Center student distribution data given in the first table of this lesson.
2. Create a relative frequency table for the 1998 Algebra Diagnostic Test Score data given in the second table of this lesson.
3. Create an ogive for the 1998 Algebra Diagnostic Test Score data given in the second table of this lesson.
4. Create a frequency table for the first 48 decimal digits of π .
5. Create a frequency table for the first 48 decimal digits of $\frac{22}{7}$.

6. Create a stem-and-leaf diagram for the following data set:
 $\{0, 2, 4, 5, 6, 3, 6, 1, 1, 50\}$.
7. Using the class marks, find the mean and standard deviation of the 1998 Algebra Diagnostic Test Score Data contained in the second table of this lesson.
(Consider it a sample.)
8. List the class boundaries for the second stem and leaf diagram in this lesson
(presidential inauguration data with split stems).
9. List the class limits for the highest class in the second stem and leaf diagram
(Figure 8.5) in this lesson (Presidential Inauguration Data With Split Stems).

Stat's Lesson 9

The Student t Distribution

Statistics are like bikinis. What they reveal is suggestive, but what they conceal is vital.

Aaron Levenstein

Once descriptive statistics, combinatorics, and distributions are well understood, we can move on to the vast area of **inferential statistics**. Sometimes, however, such statistical tests are used without all this background and that is what this lesson is about.

9.1 The Father of the t Distribution: William Gosset

William Gosset (1876–1937) was a Guinness Brewery chemist who needed a distribution that could be used with **small samples**. Since the Irish brewery did not allow publication of research results, he published in 1908 under the pseudonym of Student. This restriction came about when some trade secrets were published in a research paper. Thus the Student t -distribution is named after Gosset, although the form used today is actually a modification due to Fisher. Gosset's application of statistics to crop development (barley) led not only to improved yields, but robust crops, crops which were hearty against a wide variety of factors. His contributions to the growing field of design of experiment are worth noting.

Gosset was the first to describe the distribution of s^2 . It is related to the χ^2 by the simple factor $(n - 1)/\sigma^2$. He wasn't able to prove mathematically how it was related to the χ^2 distribution discussed in the next lesson, but he demonstrated it by dividing a prison population of 3000 into 750 random samples of size four and used their heights.

Gosset has been described as a modest man, in contrast with both Pearson and Fisher who had massive egos. He was a friend to both, a major accomplishment since they both had a loathing for each other. Gosset once cut short an admirer by saying "**Fisher would have discovered it anyway.**" Fisher, however, considered Gosset's work a "**logical revolution.**"

9.2 Hypothesis Testing

The basic concept of inferential statistics is called **hypothesis testing** or sometimes the **test of a statistical hypothesis**. Here we have two conflicting theories about the value of a population parameter. It is very important that the hypotheses be conflicting (contradictory), if one is true, the other must be false and *vice versa*. Another way to say this is that they are **mutually exclusive** and **exhaustive**, that is, no overlap and no other values are possible. **Simple hypotheses** only test against one value of the population parameter ($p = \frac{1}{2}$, for instance), whereas **composite hypotheses** test a range of values ($p > \frac{1}{2}$).

Our two hypotheses have special names: the **null hypothesis** represented by H_0 and the **alternative hypothesis** by H_a . Historically, the null (invalid, void, amounting to nothing) hypothesis was what the researcher hoped to reject. In theory, it is now common practice not to associate any special meaning to which hypothesis is which. (In practice, this may be different, so check early with your research advisor. The **research hypothesis** becomes the alternate hypothesis and the null hypothesis or “straw man” to be knocked down is so determined.) Although simple hypotheses would be easiest to test, it is much more common to have one of each type or perhaps for both to be composite. If the values specified by H_a are all on one side of the value specified by H_0 , then we have a **one-sided test** (one-tailed), whereas if the H_a values lie on both sides of H_0 , then we have a **two-sided test** (two-tailed). A one-tailed test is sometimes called a **directional test** and a two-tailed test is sometimes called a **nondirectional test**.

The outcome of our test regarding the population parameter will be that we either **reject** the null hypothesis or **fail to reject** the null hypothesis. It is considered poor form to “accept” the null hypothesis. Not guilty (not beyond reasonable doubt) is not the same as innocent! However, when we reject the null hypothesis we have only shown that it is highly unlikely to be true—we have not proven it in the mathematical sense. The research hypothesis is **supported** by rejecting the null hypothesis. The null hypothesis locates the sampling distribution, since it is (usually) the simple hypothesis, testing against one specific value of the population parameter.

Establishing the null and alternative hypotheses is sometimes considered the **first step** in hypothesis testing.

9.3 Type I and Type II Errors

Two types of errors can occur and there are three naming schemes for them. These errors cannot both occur at once. Perhaps Figure 9.1 will make it clearer.

The term **false positive** for type II errors comes from perhaps a blood test where the test results came back positive, but it is not the case (false) that the person has

Reject ↓	Truth →	H_0 True	H_a True
		H_a False	H_0 False
Reject H_a		Not an error.	False positive, Type II, $\beta = P(\text{Reject } H_a H_a \text{ true})$
Reject H_0		False negative, Type I, $\alpha = P(\text{Reject } H_0 H_0 \text{ true})$	Not an error.

Figure 9.1: False Negatives and False Positives and Other Names.

whatever was being tested for. The term **false negative** for type I errors then would mean that the person does indeed have whatever was being tested for, but the test didn't find it. When testing for pregnancy, AIDS, or other medical conditions, both types of errors can be a very serious matter. Formally, $\alpha = P(\text{Accept } H_a | H_0 \text{ true})$, meaning the probability that we "accepted" H_a when in fact H_0 was true. **Alpha** (α) is the term used to express the **level of significance** we will accept. For 95% confidence, $\alpha = 0.05$. For 99% confidence, $\alpha = 0.01$. These two alpha values are the ones most frequently used. If our **P-value**, the high unlikeliness of H_0 being true, is less than alpha, we can reject the null hypothesis. Alpha and beta usually cannot both be minimized—there is a trade-off between the two. Ideally, of course, we would minimize both. Historically, a **fixed level** of significance was selected ($\alpha = 0.05$ for the social sciences and $\alpha = 0.01$ or $\alpha = 0.001$ for the natural sciences, for instance). This was due to the fact that the null hypothesis was considered the "current theory" and the size of **Type I errors** was much more important than that of **Type II errors**. Now both are usually considered together when determining an adequately sized sample. Instead of testing against a fixed level of alpha, now the P -value is often reported. Obviously, the smaller the P -value, the stronger the evidence (higher significance, smaller alpha) provided by the data is against H_0 .

Establishing threshold error levels is often considered **step two** in hypothesis testing.

Example: On July 14, 2005 the AU EDRM611 class took 10 samples of 20 pennies set on edge and the table banged. The resultant mean of heads was 14.5 with a standard deviation of 2.12. Since this is a small sample, and the population variance is unknown, the Student t test was selected. We calculated a t value as described below and obtained $t = 6.71 = \frac{14.5 - 10}{2.12/\sqrt{10}}$. From the Student t distribution we can find a P -value of either 8.73×10^{-5} or 4.36×10^{-5} depending on whether we do a one-tailed or two-tailed test. In either case our results are certainly **statistically significant** at the 0.0001 level.

The P -value of a test is the probability that the test statistic would take a value as extreme or more extreme than that actually observed, assuming H_0 is true.

9.4 Computing a Test Statistic

Once the hypotheses have been stated, and the criterion for rejecting the null hypothesis establish, we compute the **test statistic**. The test statistic for testing a null hypothesis regarding the population mean is a z -score, if the population variance is known (so why are we sampling?). Since this is rarely the case and samples are typically small, we often use a t -score, which is computing similarly, as shown above. When testing other sample statistics (proportion, variance, *etc.*, other test statistics will be used which have their own underlying distributions. However, the same basic procedure always applies.

Computing the test statistic is considered by some **step three** in hypothesis testing.

9.5 Making a decision about H_0

The **last step** in statistical testing is deciding whether we reject or fail to reject the null hypothesis.

Although it is common to state that we have a small chance that the observed test statistic will occur by chance if the null hypothesis is true, it is technically more correct to realize that the statement should refer to a test statistic **this extreme or more extreme** since the area under any point on the probability curve is zero. It can also be said that the difference between the observed and expected test statistic is too great to be attributed to chance sampling fluctuations. That is, 19 out of 20 times it is too great—there is that 1 in 20 chance that our random sample betrayed us (given $\alpha = 0.05$). Again, should we fail to reject the null hypothesis we have to be careful to make the correct statement, such as: the probability that a test statistic of *blah* would appear by chance, if the population parameter were *blah*, is greater than 0.05. Stated this way the level of significance used is clear and we have not committed another common error (like stating that with 95% probability, H_0 is true). It is very important for the sample to have been randomly selected, otherwise bias results make such conclusions vacuous.

9.6 The Student t Distribution

It is often the case that one wants to calculate the size of sample needed to obtain a certain level of confidence in survey results. Unfortunately, this calculation requires prior knowledge of the population standard deviation (σ). Realistically, σ , is unknown. Often a preliminary sample will be conducted so that a reasonable estimate of this critical population parameter can be made. If such a preliminary sample is not made, but confidence intervals for the population mean are to be constructing using

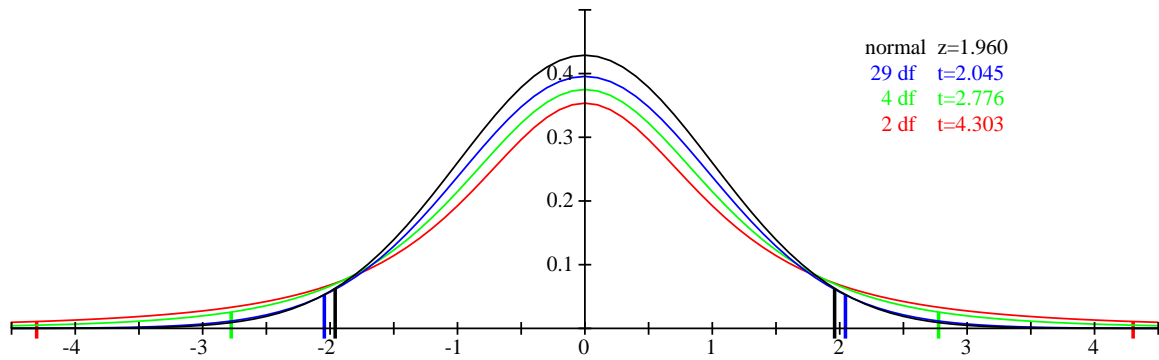


Figure 9.2: Graphs of the Student- t Distribution for various degrees of freedom (df) with the associated critical z/t values labelled for a 95% confidence interval.

an unknown σ , then the distribution known as the **Student t distribution** can be used.

Testing a hypothesis at the $\alpha = 0.05$ level or establishing a 95% confidence interval are again essentially the same thing. In both cases the critical values and the region of rejection are the same. However, we will more formally develop the confidence intervals elsewhere.

Gosset worked with small not large samples so could not use the normal distribution for his work. What Gosset showed was that small samples taken from an essentially normal population have a distribution characterized by the sample size. The population does not have to be exactly normal, only unimodal and basically symmetric. This is often characterized as heap-shaped or mound shaped.

Following are the important properties of the Student t distribution.

1. The Student t distribution is different for different sample sizes.
2. The Student t distribution is generally bell-shaped, but with smaller sample sizes shows increased variability (flatter). In other words, the distribution is less peaked than a normal distribution and with thicker tails (platykurtic). As the sample size increases, the distribution approaches a normal distribution. For $n > 30$, the differences are negligible.
3. The mean is zero (much like the standard normal distribution).
4. The distribution is symmetrical about the mean.
5. The variance is greater than one, but approaches one from above as the sample size increases ($\sigma^2 = 1$ for the standard normal distribution).
6. It takes into account the fact that the population standard deviation is unknown.

7. The population is essentially normal (at least unimodal and basically symmetric)

To use the Student t distribution, which is often referred to just as the t distribution, one calculates a t -score. This is much like finding the z -score. The formula is:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x} - \bar{x}_e}{s/\sqrt{n}}$$

Actually, since the population mean is also likely unknown, the expected sample mean must be used. The critical t -score can be looked up based on the level of confidence desired and the **degrees of freedom**.

9.7 Degrees of Freedom, Confidence Intervals

Degrees of freedom is a fairly technical term which permeates all of inferential statistics. It is usually abbreviated **df**. In this case, it is the very common value $n - 1$.

In general, the **degrees of freedom** is the number of values that can vary after certain restrictions have been imposed on all values.

Where does the term degrees of freedom come from? Suppose, for example, that you have a phone bill from Ameritech that says your household owes \$100. Your mother and father state that \$70 of it is theirs and that your younger sibling owes only \$5. How much does that leave you? Here, $n = 3$ (parents, sibling, you), but once you have the total (or mean) and two more pieces of information, the last data element is constrained. The same is true with the degrees of freedom, you can arbitrarily use any $n - 1$ data points, but the last one will be determined for a given mean. Another example is with 10 tests that averaged 55, if you assign nine people random grades, the last test score is not random, but constrained by the overall mean. Thus for 10 tests and a mean, there are nine degrees of freedom.

If the interval calls for a 90% confidence level, then $\alpha = 0.10$ and $\alpha/2 = 0.05$ (for a two-tailed test). Tables of t values typically have a column for degrees of freedom and then columns of t values corresponding with various tail areas. An abbreviated table is given below. For a complete set of values consult a larger table or your TI-84+ graphing calculator. **DISTR 5** gives **tcdf**.* **tcdf** expects three arguments: lower t value, upper t value, and degrees of freedom. Since historically no inverse t function was given on the calculator, some guessing may be involved. Note how **tcdf(9.9,9E99,2)** indicates a t value of about 9.9 for a one tailed area of 0.005 with two degrees of freedom. Please locate the corresponding value of 9.925 in the table.

*The TI-84 Silver Edition now have **invT** under **DISTR** and are not keystroke identical to the TI-83, TI-83+, and TI-84+.

α for 1 tail \rightarrow	.005	.01	.025	.05	.10
α for 2 tails \rightarrow	.01	.02	.05	.10	.20
Degrees \downarrow of Freedom					
1	63.66	31.82	12.71	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.131	1.533
5	4.032	3.365	2.571	2.015	1.476
10	3.169	2.764	2.228	1.812	1.372
15	2.947	2.602	2.132	1.753	1.341
20	2.845	2.528	2.086	1.725	1.325
25	2.787	2.485	2.060	1.708	1.316
z	2.576	2.326	1.960	1.645	1.282

Figure 9.3: Student- t Critical Values for Various Alphas and Degrees of Freedom.

As with other confidence intervals, we use the t -score to obtain the **margin of error** term which is added and subtracted from the statistic of interest (in this case, the sample mean) to obtain a confidence interval for the parameter of interest (in this case, the population mean). In this case the margin of error is defined (since you don't have population standard deviation you use the sample's) as:

$$\text{Margin of Error} = t_{\alpha/2} \cdot (s \div \sqrt{n})$$

Your confidence interval in inequality notation should look like: $\bar{x} - ME < \mu < \bar{x} + ME$, or in interval notation as: (low value, high value).

9.8 Table of t Values

The headings in Figure 9.3, such as 0.005/0.01 indicate the left/right tail area (0.005) for a one tail test or the total tail area (left+right= 0.01) for a two tailed test. In general, if an entry for the degrees of freedom you desire is not present in the table, use an entry for the next **smaller** value of the degrees of freedom. This guarantees a conservative estimate.

Although the t procedure is fairly **robust**, that is it does not change very much when the assumptions of the procedure are violated, you should always plot the data to check for skewness and outliers before using it on small samples. Here small can be interpreted as $n < 15$. If your sample is small and the data is clearly nonnormal or outliers are present, do not use the t . If your sample is not small, but $n < 40$, and there are outliers or strong skewness, do not use the t . Since the assumption that the samples are random is more important than the normality of the population

distribution, the t statistic can be safely used even when the sample indicates the population is clearly skewed, if $n > 40$.

The two sample t tests will be discussed next year.

Note: this lesson contains a heavy dose of inferential statistics. Sometimes this quantity of information is necessary for EXPO/ISEF projects. Some projects require more statistical testing than others. Testing will primarily be over the different names for types of errors, four steps of hypothesis testing, t-distribution properties, and vocabulary.

Name _____

Score _____

9.9 Homework, t Distribution

Each problem is worth three points.

1. Identify the four steps of hypothesis testing.
2. Describe Type I and Type II errors, giving alternate names as well.
3. Give alternate names for one- and two-tailed tests.
4. How are the level of significance and confidence interval related?
5. How are the critical value(s) and the region of rejection related?

6. Under what circumstances must you use the Student t distribution instead of the normal distribution?
7. Describe several characteristics of the Student t distribution.
8. What are degrees of freedom?
9. Find the 90th, 95th, and 99th percentile for the Student t distribution with 10 degrees of freedom.
10. Suppose a large college dean wishes to check for a dramatic nondirectional change in GPA in recent years. The mean for the last five years has been established as 2.95 and the mean for a random sample of 225 recent graduates is 2.85 with a standard deviation of 0.55. Test H_0 : GPA=2.95 at the $\alpha = 0.01$ level. Be sure to show your steps and state your conclusions in a professional manner. How would this change if the sample size was only 25?

Stat's Lesson 10

Chi Squared (χ^2) Goodness of Fit

*Baseball is ninety percent mental
and the other half is physical.*

Yogi Berra

*Baseball fans are junkies,
and their heroin is the statistic.*

Robert S. Weider

The test statistics used in conjunction with the normal and Student t distributions assume certain parameters about the parent populations, specifically, normality and variance homogeneity. Quite often in biological science research such restrictive assumptions cannot be made and certain **nonparametric tests** have been developed which help us analyze such data. A common distribution encountered in such nonparametric tests is the χ^2 distribution.

10.1 The Father of Math. Statistics: Karl Pearson

Karl Pearson (1857–1936) established mathematical statistics as a discipline. He started the first university statistics department in London in 1911. Although Pearson was born as Carl, this became Karl when he enrolled at a German university in 1879. He used both spellings for five years before finally adopting Karl. He eventually became universally known as KP.

Pearson worked closely with Francis Galton, a cousin to Charles Darwin. In fact, Pearson published a three volume biography on Galton. Galton worked on evolution and eugenics and upon his death funded a chair of eugenics at the University of London, which Pearson held first. Eugenics at that time was much like racism and conflicts arose between socially acceptable solutions and the scientific betterment of the race—*i.e.* Hitler's "Final Solution." Pearson's book *The Grammar of Science* affected Einstein's work.

Pearson's work in statistics was all-encompassing. We present in this lesson his Chi-squared. The **Pearson product moment correlation coefficient** is named after this Pearson because of his extensive work with correlation and regression. However, it

is unusual to find its name given so completely. Pearson also worked on classifying distributions. Pearson was offered but refused a knighthood, among other honors.

10.2 Chi Squared Distributions and Tests

The χ^2 distribution is a continuous distribution related to the normal distribution. Specifically it involves the sum of squares of normally distributed random variables. Chi is a Greek letter (χ) and is pronounced like the hard k sound in the Scottish work Loch (and **not** like those grassy chia pets). The χ^2 distribution is important in several contexts, most commonly involving variance.

The χ^2 family of distributions is characterized by one parameter called the degrees of freedom which is often denoted by ν (the Greek letter nu) and used as a subscript: χ^2_ν . The classical χ^2 distribution was developed by Fisher and Pearson.

1. The χ^2 distribution is continuous.
2. The χ^2 distribution is unimodal.
3. The χ^2 distribution is always positive (> 0).
4. The χ^2 distribution mean = ν .
5. The χ^2 distribution variance = 2ν .
6. For small ν ($\nu < 10$), the distribution is highly skewed to the right (positive).
7. As ν increases the χ^2 distribution becomes more symmetrical about ν (the mean).
8. We can thus approximate the χ^2_ν when $\nu > 30$ with the normal (see table below).

Tables of critical χ^2 values are commonly available (as below) or can be computed by a statistical package or statistical calculator.

A common application of the χ^2 distribution is in the comparison of expected with observed frequencies. When there is but one nominal variable, this is often termed **goodness of fit**. In this case we are testing whether or not the observed frequencies are within statistical fluctuations of the expected frequencies. Although one typically checks for high χ^2 values, the second example below illustrates the possible significance of a low χ^2 value.

Example: On July 14, 2005 the AU EDRM611 class collected 10 trials of 20 pennies each where these 20 pennies were set on edge and the table banged. The class observed 145 heads. We can compare the observed with expected frequencies and test for goodness of fit as shown in Figure 10.1. There is but one degree of freedom since the number of tails is dependent on the number of heads ($200 - 145 = 55$).

Side:	Head	Tail
Observed	145	55
Expected	100	100
(Obs-Exp)	45	-45
$(O - E)^2$	2025	2025
$(O - E)^2/E$	20.25	20.25

Figure 10.1: Chi-squared Goodness of Fit for 200 Penny Flips.

upper tail: degrees of f.	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	0.00016	0.001	0.0039	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21
15	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58
20	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57
25	11.52	13.12	14.61	16.47	34.38	37.65	40.75	44.31
> 30	use $z = \sqrt{2\chi^2} - \sqrt{2df - 1}$							

Figure 10.2: Table of Critical χ^2 Values for various α 's and Degrees of Freedom.

Solution: We form the χ^2 statistic by summing the $(O - E)^2/E$ and get $2025/100 + 2025/100 = 40.5$. We can then compare this χ^2 with critical χ^2 values or find an associated P -value. The critical χ^2 value for $df=1$ and one-tailed, $\alpha = 0.05$ is 3.841. Our results are far to the right of 3.841 so are VERY significant (P -value = 1.6×10^{-10}). A table of critical χ^2 values for select values is given below.

10.3 A Chi Squared Distribution Table

We also present in Figure 10.3 graphs of the χ^2 Distribution for a few Degrees of Freedom.

Example: On July 12, 2005 the AU EDRM611 class collected 192 dice rolls, each person present using a different die and each person doing 24 rolls. Were the results within the expected range?

Solution: We form the χ^2 statistic in Figure 10.4 by summing the $(O - E)^2/E$ and get $208/32 = 6.5$. We can then compare this χ^2 with a critical χ^2 . Only if it is

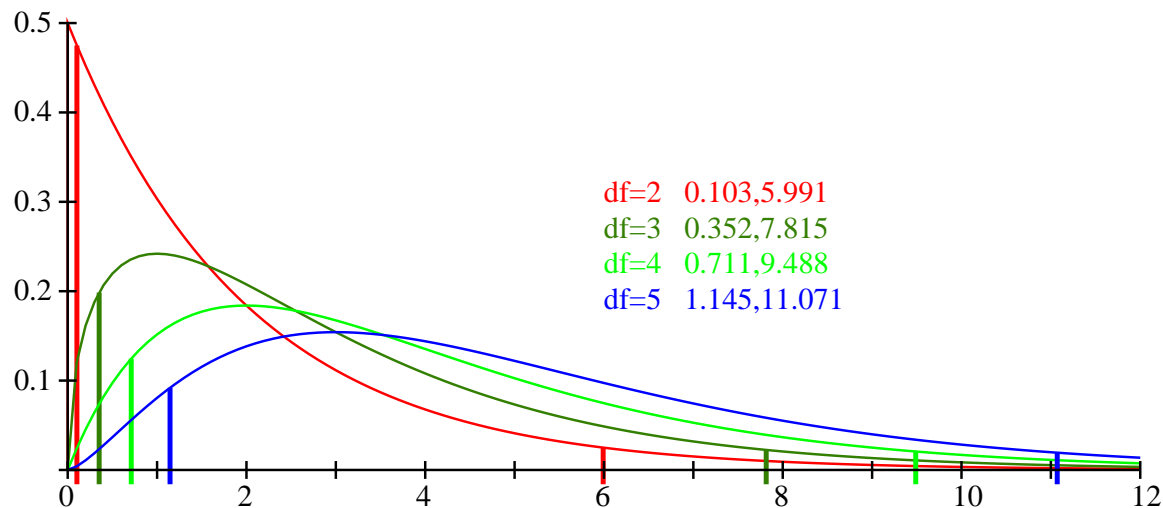


Figure 10.3: Graphs of the χ^2 Distribution for various Degrees of Freedom (df). The critical values (90% confidence interval) with 5% below or above are also indicated. The mode is $df-2$ or 0, whichever is larger. The mean is df .

Pips:	1	2	3	4	5	6
Observed	27	23	30	35	40	37
Expected	32	32	32	32	32	32
(Obs-Exp)	-5	-9	-2	3	8	5
$(O - E)^2$	25	81	4	9	64	25
$(O - E)^2/E$	0.78125	2.53125	0.125	0.28125	2.00	0.78125

Figure 10.4: Chi² Goodness of Fit for 192 Rolls of a Die.

more extreme is it worth finding a P -value. We have $6 - 1 = 5$ degrees of freedom. The critical χ^2 values for $df=5$, two-tailed, and $\alpha = 0.05$ are 1.145 and 11.07. Since our χ^2 is within this range, our results are within the range we can expect to occur by chance. Notice the lower χ^2 cut off. When people fabricate a random distribution they are likely to make it too uniform and get too small of a χ^2 which can be checked as above, but the χ^2 would likely be less than 1.145. Working backwards we see the sum of the $(O - E)^2$ would have to be less than 36 so if one were 5 or less away and the rest much closer, we might wonder.

As noted at the bottom of the table above, when the degrees of freedom are large, a z -score can be formed and compared against a standard normal distribution. Note also that the mean of any χ^2 is the degrees of freedom. This might be helpful to realize where the distribution is centered.

The χ^2 goodness of fit does not indicate what specifically is significant. To find that

out one must calculate the **standardized residuals**. The standardized residual is the signed square root of each category's contribution to the χ^2 or $R = (O - E) / \sqrt{E}$. When a standardized residual has a magnitude greater than 2.00, the corresponding category is considered a major contributor to the significance. (It might be just as easy to see which $(O - E)^2 / E$ entries are larger than 4, but standardized residuals are typically provided by software packages.)

10.4 Other Applications

The χ^2 goodness of fit test can be extended to more than one variable. It then is often termed the χ^2 **test of homogeneity**. Contingency tables are formed, expected frequencies are derived from the marginal totals, the χ^2 computed, and checked. The degrees of freedom will be $(R - 1)(C - 1)$, where R is the number of rows and C is the number of columns. The null hypothesis is that there is no statistical difference in distribution between variables one and two. Similar tests can be performed when the null hypothesis is stated somewhat differently (no relationship, form phi, test OR the proportion in state one of variable one is the same as the proportion in state two of variable one, form proportion difference, test).

Suppose we have four grade levels of students (freshman, sophomore, junior, senior) indicating which subject (English, Math, Science, Computers) is most in need of change. We collect the data in the 4×4 contingency Figure 10.5 below and have included the expected counts in parentheses based on expected frequency $= \frac{f_r \times f_c}{n}$, where f_r is the row frequency, f_c is the column frequency, and n is the total frequency (sample size).

Grade	English	Math	Science	Computers	
Freshman	35 (28.35)	8 (8.19)	10 (16.38)	10 (10.08)	63
Sophomore	30 (29.70)	8 (8.58)	20 (17.16)	8 (10.56)	66
Junior	15 (19.35)	5 (5.59)	15 (11.18)	8 (6.68)	43
Senior	10 (12.60)	5 (3.64)	7 (7.28)	6 (4.48)	28
	90	26	52	32	200

Figure 10.5: Frequencies for Students Indicating Subject Most in Need of Change.

Step 1. Our hypothesis is that there will be no difference between students in various grade levels regarding their perception of the subject most of need of change. We will use $\alpha = 0.05$. Step 2. The degrees of freedom is $(R - 1)(C - 1) = (4 - 1)(4 - 1) = 9$, giving us a critical value for the test statistic of $\chi_{cv}^2 = 16.92$. Step 3. We have calculated a $\chi^2 = 9.29$ from summing all the $\frac{(O - E)^2}{E}$. Step 4. Since 9.29 does not

exceed 16.92, our null hypothesis is not rejected and we conclude that the students are homogeneous (or rather not inhomogeneous) in their perceptions.

There are potential problems associated with small expected frequencies in contingency tables. Historically, when any cell of a 2×2 table was less than 5 a **Yates' correction of continuity** was advised. However, it has been shown that this can result in a loss of power (a tendency not to reject a false null hypothesis). Care should be exercised and advise sought. Larger contingency tables can also be problematic when more than 20% of the cells have expected frequencies less than 5 or if there are any cells with 0. One solution is to combine adjacent rows or columns, but only if it makes sense.

10.5 Don't Abuse Tests of Significance

In closing we should note the importance of focusing on a small number of well-conceived hypotheses in research rather than blindly calculating a bevy of χ^2 statistics for all variable pairs and ending up with 5% of your results being significant at the 0.05 level! You would even expect 1% of your results, due to pure random chance in your sample selection, to be significant at the 0.01 level. Since there are $n(n-1)/2$ possible pairings for n variables, one would have 4950 pairs for 100 variables of which almost 250 could look significant at the 0.05 level. Beware!

Again, this lesson was not part of my original design and contains a heavier dose of statistics than planned. However, this test has often been deemed useful for EXPO/ISEF projects and inclusion here seemed inevitable. Calculation of a Chi-squared statistic is likely on the test.

10.6 Conclusion/Errata

This concludes our introduction to statistics. We will continue next year with combinatorics, a survey of distributions, and more inferential statistics. Please collect your lectures and homework for stapling. An activity in that regard will be distributed. A list of loose ends follows.

- Formatting issues: test scoring boxes; homework 4 (yearly changes).
- References to the Numbers Lessons need to be coded with `ref` not `href` via `defs`.

Name _____

Score _____

10.7 Homework, Hypothesis Testing (χ^2)

Problems one through four are worth 5 points each. Problem 4 will be rigorously checked.

- Brian Small rolled a dice 1002 times and obtained the following results. Help him determine if the die is fair by doing a chi square goodness of fit by completing Figure 10.6. Be sure to indicate your test statistic, tails, degrees of freedom, and critical test statistic.

Pips:	1	2	3	4	5	6
Observed	181	155	141	162	153	210
Expected	167	167	167	167	167	167
(Obs-Exp)						
$(O - E)^2$						
$(O - E)^2/E$						

Figure 10.6: Chi-squared Goodness of Fit for 1000 Die Rolls—Real.

- Indicate the value of any significant standardized residuals from problem 1 above.
- Susie Agivan got tired of rolling her die and made up the data given in Figure 10.7. Help her teacher test for data fabrication by doing a chi-square goodness of fit. Be sure to indicate all four steps and values in testing this hypothesis.

Pips:	1	2	3	4	5	6
Observed	165	170	172	161	174	160
Expected	167	167	167	167	167	167
(Obs-Exp)						
$(O - E)^2$						
$(O - E)^2/E$						

Figure 10.7: Chi-squared Goodness of Fit for 1000 Dies Rolls—Faked.

4. Since 1995, blue M&M[®] candies replaced tan with* 13% brown, 14% yellow, 13% red, 20% orange, 16% green, and 24% blue candies to be expected, on average. “While we mix the colors as thoroughly as possible, the above ratios may vary somewhat, especially in the smaller bags. This is because we combine the various colors in large quantities for the last production stage (printing). The bags are then filled on high-speed packaging machines by weight, not by count.” Each student will obtain a random sample of $n = 10$ M&M's[®] from the common 14.0 oz bag.[†] Then complete the table below.

Color:	brown	yellow	red	orange	green	blue
Observed						
Expected	$\frac{13n}{100} = \underline{\quad}$	$\frac{14n}{100} = \underline{\quad}$	$\frac{13n}{100} = \underline{\quad}$	$\frac{20n}{100} = \underline{\quad}$	$\frac{16n}{100} = \underline{\quad}$	$\frac{24n}{100} = \underline{\quad}$
$(O - E)$						
$(O - E)^2$						
$(O - E)^2/E$						

Figure 10.8: Chi-squared Goodness of Fit for M&M Data.

Now add up the bottom row and call it χ^2 . Compare your value with others. Did any particular color contribute significantly to this value?

5. **Bonus:** After completing the count, feel free to dispose of the M&M's[®] by any appropriate method.

*Old values given: 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% blue.

[†]In 2009 this was reduced 10% to 12.60 oz.

10.8 Summary sheet for χ^2 Activity

Please enter your **sample data** in the space provided. Leave a blank row after each table has entered their data.

Table	M&M [®] Color: your name	brown	yellow	red	orange	green	blue	χ^2
1								
1								
1								
1								
	Table 1 Σ							
2								
2								
2								
2								
	Table 2 Σ							
3								
3								
3								
3								
	Table 3 Σ							

Table	M&M [®] Color: your name	brown	yellow	red	orange	green	blue	χ^2
4								
4								
4								
4								
	Table 4 Σ							
5								
5								
5								
5								
5								
	Table 5 Σ							
6								
6								
6								
6								
6								
	Table 6 Σ							
7								
7								
7								
7								
	Table 7 Σ							

Figure 10.9: Collection Point for χ^2 M&M Data.

10.9 Activity to Verify Book Before Stapling

Directions: You may work together, but answer each question carefully using **your own** Statistics booklet. Take time to put the booklet in **THIS** order. Make a list by **table** of who is missing what (nonbonus) items. Get booklet stapled by $Ke^{i\theta}$ on or before Oct. 22.

1. Page i (cover): Revision code/number after title.
2. Page iii (Table of Contents): Who is the Danish Father of Astronomy?
3. Page vii (List of Figures): Figure 10.1 title.
4. Page $ix-x$ (bonus): Due date (day of month) for project divided by number of clippings required (as reduced, improper fraction).
5. The project rubric listed in the table of contents as page xi will be returned with the project and is not a proper part of the statistics booklet.
6. Page 1: Lesson 1, Three founders of scientific method (bottom).
7. Page 9: Homework 1, Two data categories starting with **q** (Q5). **One bonus point** for giving synonymous names which don't start with q!
8. Page 12: Lesson 2, Second point to consider.
9. Page 17: Homework 2, How to make the answer to Q7 ratio.
10. Page 21: Lesson 3, Midrange formula.
11. Page 24: Activity 3, Keystrokes to sort a list (near middle of page).
12. Page 25: Homework 3, Official age of Eisenhower at inauguration.
13. Page 28: Lesson 4, How arithmetic and geometric sequences differ.
14. Page 33: Homework 4, Mode for first 20 **decimal** digits of e (Q4).
15. Page 36: Lesson 5, Range formula.
16. Page 41: Homework 5, Number of presidents within 2 standard deviations of mean inauguration age (Q6).
17. Page 47: Lesson 6, Second meanings for word normal.
18. Page 48: Quiz 5, Q6.

19. Page 49: Homework 6, What Chebyshev's Theorem says about IQ's between 85 and 115 (Q4).
20. Page 52: Lesson 7, Meaning and number of decimal digits in a z -score.
21. Page 58: Homework 7, Round e **up** to the appropriate integer (Q10).
22. Page 59–64: Lesson 8, Three types of graphical data representations.
23. Page 65: Homework 8, Strangeness about frequency table for $\frac{22}{7}$ (Q5).
24. Page 67: Lesson 9, Who went by the pseudonym Student?
25. Page 75: Homework 9, Give alternate names for one- and two-tailed tests (Q3).
26. Page 78: Lesson 10, The variance of the chi-square distribution.
27. Page 84: Homework 10, How many M&M's[®] in each student's sample.
28. Page 87: **Bonus:** Express $1/(\text{Section } 10.9 \text{ page number})$ **exactly** as a decimal fraction.
29. Pages 85–86, and 89 containing M&M summary and appendix header so are omitted. Strike them (pages 86 and 91) from the Table of Contents.
30. Page 92 (released test): Date on released test.
31. Page 96–97 (released test key): review before test.

In the space below draw a **BIG** smiley face. Where a nose should be put a number corresponding to how many of the above questions/activities you have answered or performed correctly

Appendix A

Odd Solutions and Released Tests/Keys

A.1 Odd Homework Answers

A.1.1 Odd Homework Answers, Stat's Introduction

1. **Statistics:** collection of methods used in planning experiments, collecting data, and analyzing it (a discipline). **Statistic:** a value, characteristic of a sample.
3. Sample: Probably Biased. Most of the callees are fed up with and do not want anything to do with Clinton.
5. Quantitative and Qualitative.
7. Ratio, Interval, Ordinal, Nomial (in that order).
9. $600 - 50 + 350 = 900$.
11. $\frac{300}{600} \times 100\% = 50\%$.
13. 90% of 25% of 100,000 is 22,500.
15. $F = \frac{9}{5}C + 32^\circ$ and $F = C$.
 $F = \frac{9}{5}F + 32^\circ$. $5F = 9F + 160^\circ$. $4F = -160^\circ$. $F = C = -40^\circ$.

A.1.2 Odd Homework Answers, Statistical Sampling

1. Discrete.
3. Continuous (except at atomic/quantum mechanical level), but probably reported fairly discretely.
5. Probably nominal (especially white, black, brown, gray, plaid, paisley, *etc.*), unless measuring rainbow color wavelengths, then ratio! There were reasons my 64 color crayon box was organized alphabetically.
7. Interval if Fahrenheit or Celsius. If converted to Kelvin or Rankine, they would be ratio!
9. Systematic.
11. Cluster.
13. Wary, bias, ("easy" isn't quite as telling).
15. Proportionate and Representative.
17. The word average is ambiguous and could refer to any of: Mode=1; Median=2; Mean=3.0; or Midrange=4.0.

19. Transistors, computers, lasers.
21. Lab notebooks can become legal documents and any information therein may help or hinder the investigation of a great breakthrough or fraud. They can be your defense, a silent witness, or your undoing.

A.1.3 Odd Homework Answers, Averages

1. Mean=\$80,000,000, if proper rules regarding significant digits are followed! Typically, students answer \$79,999,200. No mode. Median=\$3,600,000. Midrange= $\frac{\$360,000,000+\$36,000}{2} = \$180,000,000$, if proper rules regarding significant digits are followed. Typically, students incorrectly answer \$180,018,000.
3. Mean=median=midrange=4.5. No mode.
5. $\frac{100.0 \text{ kph} + 80.0 \text{ kph}}{2} = 90.00 \text{ kph}$ Watch sig. figs.!
7. To FL: $\frac{2000. \text{ km}}{100.0 \text{ kph}} = 20.00 \text{ hours}$. To MI: $\frac{2000. \text{ km}}{80.0 \text{ kph}} = 25.0 \text{ hours}$.
20.00 hours + 25.0 hours = 45.0 hours.
9. 1. $\frac{10+-2}{2} = 4^\circ \text{ F}$. 3a. $\frac{82+x}{2} = 90$. $x = 90 \cdot 2 - 82 = 98$. 5. $(\frac{-0.09+0.3}{2}, \frac{12+-4}{2}) = (0.105, 4)$.
11. $(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}, \frac{z_1+z_2}{2}, \frac{ict_1+ict_2}{2})$.

A.1.4 Odd Homework Answers, Means

1. Differs every year. 2010: $n = 6$. $\sum_{i=1}^6 x_i = 504$. $\bar{x} = 84.0$ (and not 84)
No mode. Median=79. Midrange= $\frac{52+127}{2} = 89.5$.
3. Differs every year. 2010: $\mu = 77.0$ and $\sigma = 18.1$ so $z = .39$ or the sample mean is less than half a standard deviation above the population mean.
5. Mean=Median=Midrange=4.5. No mode.
7. $\frac{2}{\frac{1}{100.0} + \frac{1}{80.0}} = 88.9 \text{ kph}$. Note: the 2 is exact.
9. $\sqrt{\frac{120^2 + (-160)^2 + 95^2 + 10^2}{4}} = \sqrt{12281.25} = 110.8 \text{ volts}$ or 111 volts, although double the usual number of significant figures out of a square root is commonly advised.
11. 54.5 for 34 ages (10% trimmed mean). 54.34 for 26 ages (20% trimmed mean).
13. $\pm\sqrt{2 \cdot 50} = \pm 10.00$. 10 not ok! Bonus for \pm .

A.1.5 Odd Homework Answers, Dispersion

1. Range=360,000,000 (not 359,964,000). $s = 160,000,000$ (not 157,249,587.4).
 $s^2 \approx 2.5 \times 10^{16}$.
3. Range=7 (possibly 7.0). $s = 2.4$. $s^2 = 6.0$.
5. 28 of 44 or 64% within [49, 60].
7. $s^2 = \frac{(1-3)^2+(1-3)^2+(2-3)^2+(4-3)^2+(7-3)^2}{5-1} = \frac{4+4+1+1+16}{4} = \frac{26}{4} = 6.5$ $s = 2.55$
9. $\sqrt{5} \approx 2.236 \approx 2.24$ $2.24^2 = 5.0176$. $2.23^2 = 4.9729$ $2.25^2 = 5.0625$.
 $\frac{.0625}{5} \approx 13 \times 10^{-3}$ or 13 ppk. $\frac{.01}{2.24} \approx 4.5 \times 10^{-3}$ or 4 ppk.

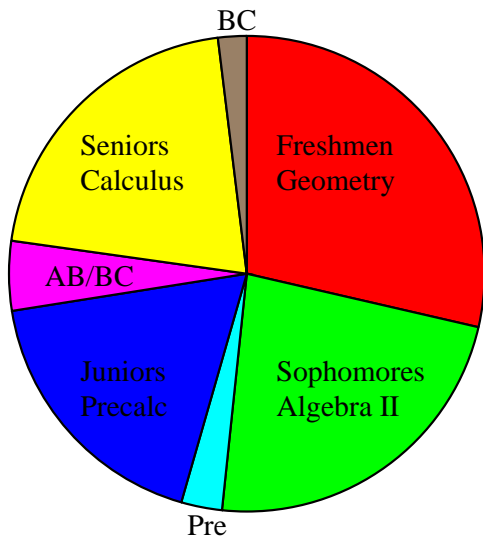
A.1.6 Odd Homework Answers, Normal Curve

1. $\bar{x} = 66400$ (not 66429.6). $s = 373,000$ (not 373243.6).
3. $\frac{68}{2} + \frac{95}{2} = 34 + 47.5 = 81.5\%$.
5. $z = \frac{167-100}{15} = 4.47$. (Remember to use 2 decimal places.)
7. 41 out of 44 or 93%.
9. $\bar{x}' = 59.8$ $s' = 6.2$ $\bar{x}' = \bar{x} + 5$ $s' = s$.
11. $\bar{x}' = 65.8$ $s = 6.8$. new $\bar{x}' = (\bar{x} + 5) \times 1.1$. $s' = s \times 1.1$.
13. $\text{normalcdf}(-1, 1) = 0.68269 = 68.269\%$.
15. $\text{normalcdf}(-3, 3) = 0.99730 = 99.730\%$.

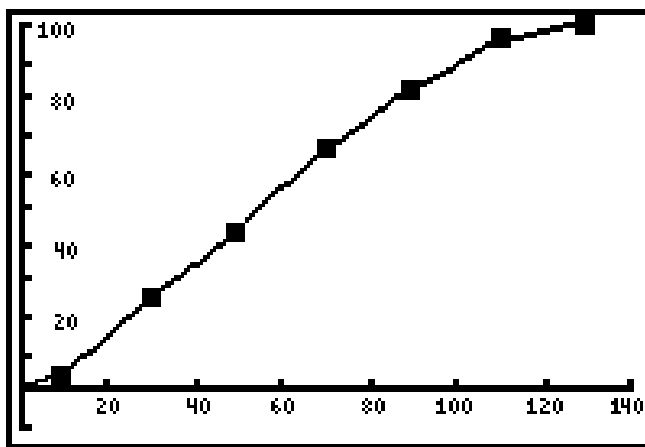
A.1.7 Odd Homework Answers, Measure of Position

1. $\frac{29-21.0}{4.7} = 1.70$.
3. $z = \pm 0.675 = \frac{Q_i - 29}{3}$. $Q_1 = 26.975 \approx 27.0$ and $Q_3 = 31.025 \approx 31.0$. $Q_3 - Q_1 = 31.0 - 27.0 = 4.0$ $29 + 2s = 29 + 6.0 = 35.0 < 36$ so yes.
5. $z = \frac{360,000,000 - 66400}{373000} = 965.00$ VERY unusual data value (outlier). The quartile and hinge definitions both fail since $Q_1 = Q_3$ and the upper and lower hinges are equal.
7. $\min X = 0$. $Q_1 = 1$. Median = 3.5. $Q_3 = 6$. $\max X = 50$.
9. Making the 50 all the way down to 9 or 10 is necessary (guess and check).
11. $L_{10} = \frac{10}{100} 50 = 5$. $\frac{53+53}{2} = 53.0 = P_{10}$.
 $L_{90} = \frac{90}{100} 50 = 45$. $\frac{90+92}{2} = 91.0 = P_{90}$. $P_{90} - P_{10} = 91.0 - 53.0 = 38.0$.

A.1.8 Odd Homework Answers, Presenting Data



1.



3.

Digits	Frequency
0	0
1	8
2	8
3	0
4	8
5	8
6	0
7	8
8	8
9	0

5.

7. The class marks are: 9.5, 29.5, 49.5, ... Enter as L_1 with the frequencies in L_2 , do 1-varstat L_1, L_2 . $\bar{x} = 65.9$. $s = 31.7$ (assuming sample).

9. 65 and 69.

A.1.9 Odd Homework Answers, t Distribution

1. A) State H_0 and H_a . B) Set α and β . C) Compute test statistics/confidence interval. D) Form conclusion (state P -value).
3. One-sided hypothesis/test is directional. A two-sided hypothesis/test is non-directional.
5. The rejection region is bounded by the critical values.
7. Varies with sample size. Generally bell-shaped but thick tails at small n . Symmetric with mean of zero. Variance > 1 , but approaches 1 as n increases.
9. `Inv-t` was added under `distr` on about the TI-84+. On earlier calculators you have to use guess and check using `tcdf(-9E99, ???)`. `invt(.90, 10) = 1.372` for $df=10$. `invt(.95, 10) = 1.812` for $df=10$. `invt(.99, 10) = 2.764` for $df=10$.

A.1.10 Odd Homework Answers, Hypothesis Testing (χ^2)

1. One-tailed. 5 degrees of freedom. $\alpha = .05$. $\chi_c^2 = 11.071$.
 $\chi^2 = 18.47$ (sum of bottom row). We can reject an H_0 that the die is fair at the $\alpha = 0.05$ level or a P -value of 0.0024.

Pips:	1	2	3	4	5	6
Observed	181	155	141	162	153	210
Expected	167	167	167	167	167	167
(Obs-Exp)	14	-12	-26	-5	-14	43
$(O - E)^2$	196	144	676	25	196	1849
$(O - E)^2/E$	1.17	0.86	4.05	0.15	1.17	11.07

3. A. H_0 : Statistically random; H_a : not statistically random.
 B. $\alpha = 0.05$ $df = 6 - 1 = 5$ 1-tailed.
 C. χ_c^2 (lower) is 1.145. Our value: $\chi^2 = 1.03$
 D. We can reject H_0 but the dean might not with $\alpha = 0.01$ and $\chi_c^2 = 0.554$.

Pips:	1	2	3	4	5	6
Observed	165	170	172	161	174	160
Expected	167	167	167	167	167	167
(Obs-Exp)	-2	3	5	-6	7	-7
$(O - E)^2$	4	9	25	36	49	49
$(O - E)^2/E$	0.024	0.054	0.150	0.216	0.293	0.293

5. Eating the M&M's is probably the most popular disposal method.

Name _____

Score _____

A.2 Released Test: Intro. to Statistics, Oct. 19, 2001

One 3"x5" notecard and your graphing calculator allowed. Place short answers on the blank provided toward the left. Leave the scoring boxes blank. **SHOW YOUR WORK.** Each of the 20 question numbers is worth 5 points. Allocate your time wisely. Read the questions carefully. Hand in all scratch paper and the cover sheet with your test.

Part I, Constructed Response, 25%, 25 points.

Given the following **sample** of test scores, perform the indicated operation or calculate the statistical quantity indicated.

$$\{83, 68, 66, 68, 98, 60, 42, 71, 75\}$$

5

1. Construct a **stem-and-leaf** diagram.

5

___ 2. **Midrange.**

5

___ 3. **Arithmetic Mean.**

5

___ 4. **Standard Deviation.**

5

___ 5. Show how to compute the **z-score** for the smallest test score. Put your answer in the proper format.**End of Part I—test continues on back side of sheet.**

Part II, Multiple Choice, 25%, 25 points.

5
___ 6. What is the mode of the data set $\{1, 1, 2, 4, 7\}$?
A. 1 B. 2 C. 2.2 D. 3.0 E. 4.0

5
___ 7. In a class of 30 students the average exam score is 70. The teacher throws out the exams with the top score (which was 90) and the bottom score (which was 22) and recomputes the average based on the remaining 28 exams. What is the new average?
A. 65.4 B. 68 C. 69 D. 71 E. Insufficient information.

5
___ 8. What is the harmonic mean of the data set $\{2, 3, 4\}$?
A. 2.77 B. 2.88 C. 3.0 D. 3.11 E. 4.0

5
___ 9. If you add 5 to each value in a data set, then the standard deviation will:
A. decrease by 5. B. stay the same C. increase by 5.
D. reduce by a factor of 2.236. E. increase by a factor of 2.236.

5
___ 10. What is the variance of the sample data set $\{1, 2, 3, 4, 5\}$?
A. 2.0 B. 2.5 C. 10 D. 15 E. 55

End of Part II—test continues on next sheet.

25

Part III, True/False, 10%, 10 points.

10

11,12. Circle **T** if the statement is true and **F** if the statement is false.

- T** **F** a. The car seat at 180°F is twice as hot as the 90°F in the shade.
- T** **F** b. A car weighing 1430 kilograms is an example of continuous data.
- T** **F** c. Three students were absent yesterday is an example of discrete data.
- T** **F** d. Colors of cars is an example of the interval level of measurement.
- T** **F** e. Ratio data have an inherent starting point.
- T** **F** f. This is an example of an open question.
- T** **F** g. Range is a measure of dispersion.
- T** **F** h. You may omit empty classes in a frequency table.
- T** **F** i. A frequency table's class width is the difference between the upper and lower class limits.
- T** **F** j. In proceeding from left to right, the graph of an ogive can follow a downward path.

Part IV, Matching, 15%, 15 points.

5

13. Form the best match among the following **dispersion terms**:

- | | |
|---------------------------|---|
| _____ Chebyshev's Theorem | A. most data is in 4 standard deviations min. to max. |
| _____ empirical rule | B. $\frac{\Sigma(x - \mu)^2}{n}$ |
| _____ range rule of thumb | C. 68%–95%–99.7% |
| _____ standard deviation | D. $1 - \frac{1}{K^2}$ |
| _____ variance | E. $\sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ |

5

14. Form the best match among the following **types of sampling**:

- | | |
|----------------------------|--|
| _____ Random sampling | A. population divided, all subpopulations sampled |
| _____ Systematic sampling | B. every k^{th} member sampled |
| _____ Stratified sampling | C. all elements have an equal chance to be measured |
| _____ Cluster sampling | D. elements might choose whether to be sampled |
| _____ Convenience sampling | E. population divided, few subpopulations exhaustively sampled |

5

15. Form the best match among the following members of a **5-number summary**:

- | | |
|---------------|---|
| _____ Minimum | A. This value is near the lower hinge. |
| _____ Q1 | B. This value is above the 99 th percentile. |
| _____ Median | C. P_{75} is another name for this value. |
| _____ Q3 | D. D_5 is another name for this value. |
| _____ Maximum | E. No score in the data set can be lower than this. |

End of Parts III and IV—test continues on back of sheet.

Part V, Short Answer/Completion, 15%, 15 points.

15

16,17,18. Complete the following sentences with one appropriate word (3 points each).

- A. **Parameter** is to population as _____ is to **sample**.
 B. _____ statistics tries to infer information about a population by sampling.
 C. Be _____ of convenience sampling.
 D. Better results are obtained by _____ instead of asking.
 E. A boxplot is also known as a box and _____ plot.

Part VI, Essay, 10%, 10 points.

5

19. Discuss which measure of central tendency is the best.

5

20. Discuss the differences in application and meaning between the empirical rule and Chebyshev's Theorem.

End of Parts V & VI.

I have been careful to not allow others to see my work and the work on this examination is completely my own. This examination is returned and associated solutions are

provided for my own personal use only. I may not share them except with concurrent classmates taking the identical course. Other uses are not condoned. I will dispose of it properly.

signature

date

End of Test.—Check your work.—Have a nice day!

25

Name Key

Score 100/100

A.3 Key for Released Statistics Test: Oct. 19, 2001

One 3"x5" notecard and your graphing calculator allowed. Place short answers on the blank provided toward the left. Leave the scoring boxes blank. **SHOW YOUR WORK.** Each of the 20 question numbers is worth 5 points. Allocate your time wisely. Read the questions carefully. Hand in all scratch paper and the cover sheet with your test.

9|8

8|3

7|51

6|8860

5|

4|2

Part I, Constructed Response, 25%, 25 points.

Given the following **sample** of test scores, perform the indicated operation or calculate the statistical quantity indicated.

{83, 68, 66, 68, 98, 60, 42, 71, 75}

in order (ascending or descending)
no commas or horizontal lines
no missing numbers
Don't omit stem 5

5

1. Construct a **stem-and-leaf** diagram.

5

70.0

2. Midrange.

$(\max + \min) / 2 = \frac{42 + 98}{2} = 70.0$

5

70.1

3. Arithmetic Mean.

$\frac{83 + 68 + 66 + 68 + 98 + 60 + 42 + 71 + 75}{9} = \frac{631}{9} = 70.111\dots$

5

15.4

4. Standard Deviation.

Round to 3 sig. fig. or 1 more than data

$s = 15.35777, \sigma = 14.47944$

5

-1.82

5. Show how to compute the **z-score** for the smallest test score. Put your answer in the proper format.

5

5

5

$z = \frac{x_i - \bar{x}}{s} = \frac{42 - 70.1}{15.4} \approx -1.82$

End of Part I—test continues on back side of sheet.

Use 2 decimal places!

25
25

Part II, Multiple Choice, 25%, 25 points.

- 5**
5
- A 6. What is the mode of the data set $\{1, 1, 2, 4, 7\}$?
A. 1 B. 2 C. 2.2 D. 3.0 E. 4.0

One occurs MOST often.
Two is the median or middle value.
2.2 is the geometric mean.
Three is the arithmetic mean.
Four is the midrange.

- 5**
5
- D 7. In a class of 30 students the average exam score is 70. The teacher throws out the exams with the top score (which was 90) and the bottom score (which was 22) and recomputes the average based on the remaining 28 exams. What is the new average?

A. 65.4 B. 68 C. 69 D. 71 E. Insufficient information.

$$30 \cdot 70 = 2100$$

$$2100 - 90 - 22 = 1988$$

$$1988/28 = 71.0$$

- 5**
5
- A 8. What is the harmonic mean of the data set $\{2, 3, 4\}$?
A. 2.77 B. 2.88 C. 3.0 D. 3.11 E. 4.0

$$\frac{3}{\frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = \frac{3}{\frac{6+4+3}{12}} = \frac{3}{\frac{13}{12}} = \frac{36}{13} = 2.77$$

Other values are: geometric mean, mean/median quadratic mean, and maximum.

- 5**
5
- B 9. If you add 5 to each value in a data set, then the standard deviation will:
A. decrease by 5. B. stay the same C. increase by 5.
D. reduce by a factor of 2.236. E. increase by a factor of 2.236.

The spread of the data doesn't change.

- 5**
5
- B 10. What is the variance of the sample data set $\{1, 2, 3, 4, 5\}$?
A. 2.0 B. 2.5 C. 10 D. 15 E. 55

$$\frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1} = \frac{4+1+0+1+4}{4} = \frac{10}{4}$$

End of Part II—test continues on next sheet.

Part III, True/False, 10%, 10 points.**10**

10

11,12. Circle **T** if the statement is true and **F** if the statement is false.

- T** **F** a. The car seat at 180°F is twice as hot as the 90°F in the shade.
T **F** b. A car weighing 1430 kilograms is an example of continuous data.
T **F** c. Three students were absent yesterday is an example of discrete data.
T **F** d. Colors of cars is an example of the interval level of measurement.
T **F** e. Ratio data have an inherent starting point.
T **F** f. This is an example of an open question.
T **F** g. Range is a measure of dispersion.
T **F** h. You may omit empty classes in a frequency table.
T **F** i. A frequency table's class width is the difference between the upper and lower class limits.
T **F** j. In proceeding from left to right, the graph of an ogive can follow a downward path.

Part IV, Matching, 15%, 15 points.**5**

5

13. Form the best match among the following **dispersion terms**:

- D Chebyshev's Theorem A. most data is in 4 standard deviations min. to max.
C empirical rule B. $\frac{\Sigma(x - \mu)^2}{n}$
A range rule of thumb C. 68%–95%–99.7%
E standard deviation D. $1 - \frac{1}{k^2}$
B variance E. $\sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$

5

5

14. Form the best match among the following **types of sampling**:

- C Random sampling A. population divided, all subpopulations sampled
B Systematic sampling B. every k^{th} member sampled
A Stratified sampling C. all elements have an equal chance to be measured
E Cluster sampling D. elements might choose whether to be sampled
D Convenience sampling E. population divided, few subpopulations exhaustively sampled

5

5

15. Form the best match among the following members of a **5-number summary**:

- E Minimum A. This value is near the lower hinge.
A Q1 B. This value is above the 99th percentile.
D Median C. P_{75} is another name for this value.
C Q3 D. D_5 is another name for this value.
B Maximum E. No score in the data set can be lower than this.

End of Parts III and IV—test continues on back of sheet.**25**

25

Part V, Short Answer/Completion, 15%, 15 points.

15

15

16,17,18. Complete the following sentences with one appropriate word (3 points each).

- A. **Parameter** is to population as statistic is to **sample**.
 B. Inferential statistics tries to infer information about a population by sampling.
 C. Be wary of convenience sampling.
 D. Better results are obtained by measuring instead of asking.
 E. A boxplot is also known as a box and whiskers plot.

Part VI, Essay, 10%, 10 points.

5

5

19. Discuss which measure of central tendency is the best.

See Statistics Section 3.3.

5

5

20. Discuss the differences in application and meaning between the empirical rule and Chebyshev's Theorem.

See Statistics Sections 6.3 and 6.4.

End of Parts V & VI.

I have been careful to not allow others to see my work and the work on this examination is completely my own. This examination is returned and associated solutions are

provided for my own personal use only. I may not share them except with concurrent classmates taking the identical course. Other uses are not condoned. I will dispose of it properly.

Keith

Oct. 22, 2001

signature

date

End of Test.—Check your work.—Have a nice day!

25

25