Probability and Distributions, pdf^4

by **Keith G. Calkins, Ph.D.**
with assistance by
Shirleen Luttrell

2009–2010

Berrien County Math & Science Center
Andrews University
Berrien Springs, MI 49104-0140

TABLE OF CONTENTS

Note: your page numbers will vary.

Probability & Dist. Lesson 0

Preface and Stat Review

98% of all statistics are made up. Unknown

0.0 Preface and Stat Intro Review

The main sequence of math courses at the Berrien County Math & Science Center leads students presumably from an Algebra experience in eighth grade into AP Calculus AB by their senior year. Said Algebra experience varies greatly as does each student's aptitude and interests. Although a main focus of their freshman Geometry experience is to prepare them for Algebra II, we take time to develop descriptive statistics sufficient for their science project needs.

Similarly during their sophomore Algebra II experience we study probability and distributions, covering topics which lay the foundation for hypothesis testing. Although some hypothesis testing is presented, both in the freshman and sophomore coverage, it is not the major focus. That is left for a series their junior year (which has yet to be developed). The intent is to cover about half of the AP Statistics curriculum in these three units.

This **Probability and Distribution** booklet only slowly evolved. Although some work was done summer of 1998 when this textbook writing craze began in earnest, it specifically didn't really gain momentum until the next summer when it became clear that Aurora Burdick, the math teacher for Algebra II and one other subject (Geometry, then Precalculus) during 1998–2000 wanted our sophomores to have their statistics in the spring instead of the fall. We also didn't want to issue the Triola textbooks to them and that second year I taught AP Statistics to a dozen juniors. Also, at that time Statistics was a weak area for Shirleen, so her work with the project at that time served to help bring her up to speed in that area. She and Rita Sircar produced some early drafts of the web pages which I drew on whenever a particular lesson got pushed into publication.

I fell short of an initial goal of five new sections each year, but the contents and arrangement has been fairly stable since 2003—a reduced load spring 2002 was helpful in that regard. Spring 2007 I converted everything from `html` into LaTeX which allows much better content/format control and `pdf` output. Spring 2008 we further improved and enhanced the booklet, finally eliminating the non-existent lessons 4 and 5. Spring 2009 we added sections on experimental design, non-parametric statistics, some biographies, and some quotes. The needed enhancements are much fewer but it still is a work in progress.

The May 2004 Reader's Digest told the tale of a fellow who put into a cup one spoonful of many other chili's at a chili contest and ended up winning. My intent somewhat is for my lecture notes to do the same, blending the best features of some, while avoiding others. We hope you enjoy this romp through probability and distributions as much as I have enjoyed preparing the materials.

We start our series of probability and distributions lessons with a quick review of the freshmen introduction to statistics material.

0.1 Lesson 1, Data & Measurement

1. **Descriptive** Statistics characterizes or describes a data set.
2. **Inferential** Statistics tries to infer information about a population from a sample.
3. Statistics is a **collection of methods** used in planning an experiment and analyzing data. It is also the plural of statistic (see below).
4. **Population** is the complete set of data elements.
5. A **sample** is a selected portion of a population.
6. **Parameters** characterize a population, whereas a **statistic** is a sample measure.
7. **Accuracy** is a measure of rightness, whereas **precision** is a measure of exactness.
8. Statistics can be **misused** in a variety of ways to prove most anyone's point of view.
9. **Data are plural,** whereas **datum** is singular.
10. **Qualitative** data are nonnumeric: good, better, best.
11. **Quantitative** data are numeric and can be either discrete (quantized) or continuous.
12. Being able to do simple math: fractions, percentages, etc., is important.
13. There are four **levels of measurement:** nominal, ordinal, interval, ratio.
14. **Ratio** is the highest level, data are interval and has a starting point (zero, like Kelvin).
15. **Interval** data have meaningful intervals between measurements (Celsius, Fahrenheit).
16. **Ordinal** data have order but lack meaningful intervals: (strongly) agree, disagree, etc.
17. **Nominal** data have names only: brown, plaid, paisley.

0.2 Lesson 2, Sampling

1. **Sample size** is very important. We want a sample not a **census.**
2. **Measure**, don't ask.
3. **Random** errors are ok, **systematic** errors need to be accounted for, **sampling** errors should be designed out. No randomization, no generalization.
4. Sampling **medium** used (mail, phone, e-mail/web, personal interview) will affect accuracy.
5. Sample must be **representative** of the population (avoid bias).
6. Observational studies are more passive whereas experiments deliberately impose **treatments** on individuals. Can't always experiment. Experiments allow conclusions!
7. There are five primary **sampling methods**. Random = representative = proportionate.
8. In **random** sampling any population member has a equal chance of being measured.
9. In **systematic** sampling every k^{th} member of the population is sampled.
10. In **stratified** sampling the population is divided into two or more strata and each subpopulation is sampled.
11. In **cluster** sampling a population is divided into clusters and a few of these clusters are exhaustively sampled.
12. In **convenience** sampling the element can often select whether or not it is sampled.
13. Be very **wary** of convenience sampling since it is prone to bias.
14. Questions may be **open** ended (essay) or **closed** (multiple-choice, true/false).
15. Studies may be **retrospective** (looking back) or **prospective** (looking ahead).

0.3 Lesson 3, Measures of Central Tendency

1. Average is an **ambiguous** term referring often but not exclusively to the **arithmetic mean.**
2. By average we usually mean some **measure of central tendency.**
3. **Mode, median, and midrange** are additional common averages.
4. We find the arithmetic mean by summing all elements and dividing by the number of elements.
5. Although $\bar{x}$ (x-bar) is used for **sample mean,** μ (mu) is used for **population mean.**
6. Sample size is the **number of elements** and is denoted by n (lower case).
7. The **population size** is typically denoted by N (upper case).
8. Mode is the data element which occurs **most** frequently.
9. A **uniform** distribution can be said to have no mode.
10. Distributions may also be **bimodal** or multimodal.
11. The median is the **middle** element in an ordered data set.
12. When there are an even number of elements, the median is the arithmetic mean of the middle two.
13. The midrange is the arithmetic mean of the highest and lowest data elements.
14. Do not confuse midrange, a measure of central tendency, with range, a **measure of dispersion.**
15. The mean is **reliable** (uses every data element) but can be distorted by outliers.
16. While no average is the best, under certain circumstances one may be better than another.
17. We typically report the mean to one more **significant digit** than the data.
18. One should probably report the mean and standard deviation to the same precision.
19. Another common rule in science is to use three significant digits (slide rule accuracy).

0.4 Lesson 4, Various Means

1. The **arithmetic mean** is the sum of all elements divided by the number of elements.
2. The **geometric mean** is used to find average rates of growth.
3. The geometric mean is the n^{th} root of the product of the data elements.
4. n^{th} roots can be found on your calculator using fractional exponents ($\frac{1}{2}$ would be square root).
5. The **harmonic mean** is used to calculate average rates like speed. $\dfrac{n}{\sum x_i^{-1}}$
6. Harmonic mean is found by dividing n by the sum of reciprocals of the data elements.
7. **Reciprocal** means "1 over the value."
8. Speed is a **scalar,** whereas **velocity is a vector** (has both magnitude and direction).
9. The **quadratic mean** is also known as Root Mean Square (RMS). $\sqrt{\dfrac{\sum x_i^2}{n}}$
10. It is used for AC voltage and is the square root of (the sum of the squares divided by n).
11. The arithmetic mean of AC voltage is zero.
12. The 10% **trimmed mean** is the arithmetic mean without the top 10% and bottom 10%.
13. This avoids **outlier** distortion and corrects some skew.
14. A distribution is **skewed** to the right if the mean is to the right of the median.
15. **Weighted means** are most commonly encountered in GPA's where items have differing affects.
16. Sometimes none of these means suffice and some **combination** is required.
17. Be sure you can calculate means not only from a table of values, but also from a **frequency table.**

0.5 Lesson 5: Measures of Dispersion

1. **Dispersion** is how a data set is distributed.

2. Common measures of dispersion are **range, standard deviation, and variance.**

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

4. Range is easily distorted, due to its use of but two elements.

5. **Standard deviation** is by far the most important measure of dispersion.

6. Standard deviation is the **average distance** of each data element from the mean.

7. The formula for standard deviation varies depending on whether it is for a **sample or a population.**

8. Sample standard deviation is denoted by s, whereas population standard deviation is denoted by σ (sigma).

9. This use of **Roman** characters for sample and **Greek** charcters for population is standard.

10. The sample standard deviation is slightly larger because of the dependance on the sample mean.

11. **Degrees of freedom** (often $n - 1$) is an important statistic in any statistical study.

12. Standard deviation comes as the square root of the **variance.**

13. Standard deviation has the **same units** as the data so can be easier to understand.

14. In general, the range of a sample is about four times its standard deviation (**range rule of thumb**).

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

16. Variance is a **primary statistic,** standard deviation is derived, be careful with precision/accuracy.

0.6 Lesson 6: The Normal, Bell-shaped, Gaussian Distribution

1. The **Normal Distribution** has other names: **Gaussian, Bell-shaped,** and sometimes **Error.**

2. Error distributions and many other phenomena tend toward a normal distribution.

3. The normal distribution is **symmetric.**

4. A **standard normal** distribution has an area of 1, mean of 0, and standard dev. (and variance) of 1.

5. The **empirical rule** is based on the normal distribution of **68%-95%-99.7%** of a data set being within 1, 2 or 3 standard deviations of the mean.

6. IQ scores with mean of 100 and standard deviation of 15 are a common non-standard example.

7. The thin parts of a distribution are called **tails** (or sometimes wings).

8. Statistics can be interested in **one tail,** the left tail or the right tail, or both (**two tail**).

9. The Math & Science Center draws special education students from the upper tail of the IQ curve.

10. Whereas Blossomland draws special education students from the lower tail of the IQ curve.

11. In theory, the tails are of infinite extent.

12. In practice, the tails are especially difficult to measure.

13. **Chebyshev's Theorem** (C.T.) applies to any distribution.

14. C.T. **guarantees** that $1 - \frac{1}{K^2}$ of the data is within K standard deviations of the mean, for $K > 1$.

0.7 Lesson 7: Measurements of Position

1. z-scores indicate in units of standard deviation **how far** an element is from the mean.
2. Positive z-scores are **above** the mean; negative scores are **below** the mean.
3. Thus the formula is $z =$ (element - mean) / standard deviation.
4. Traditionally, z-scores are **rounded to two decimal places** and are also known as standard scores.
5. z-scores make it easier to compare scores with differing means/standard deviations.
6. An example might be test scores (70, 15), IQs (100, 15), ACT (21, 4.7), and SAT (1020, 157).
7. Data elements more (less) than 2 standard deviations from the mean are **unusual (ordinary).**
8. Data are **ranked** when arranged in [numeric] order.
9. The median divides a data set into a bottom half and a top half.
10. Similarly, the three **quartiles,** Q_1, Q_2, and Q_3 divide a data set into four quarters.
11. The left and right **hinge** correspond with Q_1 and Q_3 respectively, but definition nuances exist.
12. **Outliers** are extreme values in a data set and are often classified as mild or extreme.
13. An outlier is hard to define, but should be easy to recognize.
14. The **interquartile range** $Q_3 - Q_1$ is not sensitive to outliers.
15. The **semi-interquartile range:** $(Q_3 - Q_1)/2$ is another measure of dispersion.
16. The **midquartile** $(Q_1 + Q_3)/2$ is another measure of central tendency.
17. The 9 **deciles:** $D_1, D_2, \ldots D_9$ divide a data set into 10 parts.
18. The 99 **percentiles:** $P_1, P_2, \ldots P_{99}$ divide a data set into 100 parts.
19. Q_2, D_5, and P_{50} are synonyms for median; There is no 100^{th} percentile.
20. In the percentile **locator** formula: $L = \frac{k \cdot n}{100}$, L must be rounded UP (k is percentile).
21. The **10–90 percentile range** is another measure of dispersion: $P_{90} - P_{10}$.
22. The **5-number summary** is: minimum, Q_1, median, Q_3, maximum.

0.8 Lesson 8: Summarizing and Displaying Data

1. **Frequency tables** list data categories/classes in one column and frequencies in another.

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

3. **Class boundaries** are the numbers which separate classes—halfway between the limits.

4. **Class marks** are the midpoints of the classes.

5. Equal **class widths** are the difference between two consecutive class boundaries.

6. **Relative frequency tables** use percentages or decimal fractions instead of counts.

7. **Cumulative frequency tables** include all occurances less than the given value.

8. A **histogram or bar graph/chart** uses the vert. axis for frequency and the hor. axis for classes.

9. The skewness of a sample/population should become apparent.

10. Relative frequency histogram uses relative frequency on the vertical scale.

11. An **ogive** (*Oh Jive*) is a cumulative frequency polygon—the tops of where the bars would be are joined.

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

13. **Pie charts** are yet another way to display relative proportions of a data set.

14. **Pictographs** can be pretty but easily misleading.

15. **Stem-and-leaf diagrams** are part of **exploratory data analysis** (EDA).

16. Please **omit:** commas, not stems, horizontal lines; and put your data in order.

17. Rules for **split/combined** stems, multidigit leaves, should be reviewed.

18. Plan to have **between 5 and 20 stems** (or classes).

19. A **boxplot or box and whiskers plot** visually displays the 5-number summary.

0.9 Lesson 9: The Student t Distribution

1. Inferential statistics assumes a knowledge of descriptive statistics.
2. The test of a statistical hypothesis or hypothesis testing is fundamental to inferential statistics.
3. Conflicting (mutually exclusive) hypotheses are formed.
4. These may be simple (one value $p = \frac{1}{2}$) or composite ($p > \frac{1}{2}$).
5. One is the null hypothesis (H_0), the other the alternative (H_a) hypothesis.
6. Often one wants to reject the null and thus support the research hypothesis.
7. These hypotheses might be one-sided/-tailed/directional ($P > \frac{1}{2}$) or two-sided/-tailed/non-directional (when $p \neq \frac{1}{2}$). Establishing these is step one.
8. Type I errors, false negatives, or rejecting a true H_0 is alpha (α).
9. Type II errors, false positives, or rejecting a true H_a is beta (β).
10. Fixing a level of significance ($\alpha = 0.05$, for instance) is step two.
11. Computing the test statistic is step three.
12. Step four compares the test statistics and rejects or fails to reject the null hypothesis. Take care in wording your conclusion.
13. The Student t distribution was named for Gosset's psuedonym. $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$.
14. This Guinness Brewery chemist was disallowed publication of his small samples.
15. Degrees of freedom important (often $n - 1$). Margin of error commonly used.
16. The t test is fairly robust, but check for outliers and skewness if $n < 15$.
17. For a two sample t test very small samples can be used.

0.10 Lesson 10: Chi Square Goodness of Fit

1. χ^2 is a nonparametric test; normality and variance homogeneity is unimportant.
2. χ^2 is continuous, unimodal, always positive, the mean=ν, and variance=2ν.
3. ν here are the degrees of freedom.
4. If $\nu < 10$ it is highly skewed to the right but with $\nu > 30$ the normal can be used.
5. Generally, we form observed-expected, square it, and divide by the expected.
6. The sum is the χ^2 which can be compared with table values.
7. Identify standardized residuals: $|O - E|/\sqrt{E} > 2$ as major contributors.
8. Be sure you can check your M&M's against expected.

Probability & Dist. Lesson 1

Fundamental Definitions for Probability

> *The most important questions of life are, for the most part, really questions of probability.*[*] LaPlace

This lesson sets the stage for subsequent lessons by defining the fundamental terms used in probability, such as experiment, random experiment, event (both simple and compound), sample space, outcomes, probability, fair, impossible, certain, and random sample. We also introduce the first fundamental theorem of probability, or the law of large numbers. We start with a short biography on a man called by some the father of probability.

1.1 The Father of Probability: Girolamo Cardano

Cardano (1501–1576) is often referred to by his Latin name Cardan. He was the illegitimate son of a Milanean lawyer whom Leonardo da Vinci consulted regarding Geometry. His parents later married and lived together. Cardano learned mathematics from his father and became his father's assistant. Later he studied medicine. He was outspoken, highly critical, and hence not well liked. Cardano squandered his father's inheritance but then turned to gambling to make a living. Since he knew the odds better than his opponents, he tended to win more than he lost. However, this addiction robbed Cardano of many valuable years, money, and his reputation.

Cardano received a doctorate in medicine, and set up a practice. At first he was not given membership in the College of Physicians (due to his birth status and reputation) in Milan which limited his success. After gambling losses forced him to pawn his wife's jewelry and furniture, he obtain a lecturer position in mathematics and practiced

[*] *Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.*

medicine on the side. Some near miraculous cures and good reputation resulted in his consulting for members of the College and then his membership. Following this he gambled and played chess all day every day.

However, Cardano did extensive work, with the help of others, on solving cubic and quartic functions, which he published. This included the first work with imaginary numbers. He rose to rector of the College of Physicians with the reputation of being the greatest physician in the world. He had many offers and accepted a professorship of Medicine at Pavia.

Cardano's eldest son secretly married then poisoned a girl, after she repeatedly cuckolded him. The son was arrested and executed. Cardano himself was later imprisoned for heresy and barred from university work. In addition to his contributions to algebra, he made important contributions to hydrodynamics, mechanics, and geology. He published two encyclopædias of natural science which ranged over a broad spectrum of topics. His contributions to probability were the first to explore this topic. Supposedly, Cardano predicted his own exact date of death, but may have unduely influenced the outcome.

1.2 [Random] Experiment

An **experiment** is a method by which observations are made.

A famous example of an experiment is when Benjamin Franklin, famous American statesman and scientist, determined whether electricity is conducted. The experiment involved flying a kite in a thunder (and lightning) storm with a wire from the kite to a key in a bottle. (Don't try this at home!) (Also, questions[†] have arisen as to whether or not he actually performed this experiment. It seems others did it earlier, only his son may have been present, and his journals don't support well this event occurring.) The experimental method is now the basis of the scientific method. In statistics we often refer to a random experiment, one for which there is no way of telling beforehand what the outcome will be.

The act of rolling a fair die, flipping an honest coin, or randomly selecting a card from a deck are all considered **random experiments**.

An interesting part of mathematics is the use of common language to describe mathematical concepts. One such example is the word event. Normally, event conjures up images of special moments: the prom, banquets, fairs, weddings, births, ...

[†] http://www.sciencefriday.com/pages/2003/Jul/hour2_070403.html

In dealing with probability, event has a very precise meaning.

An **event** is the set of **outcomes** from a random experiment.
A **simple event** is an outcome which cannot be broken down.
The **sample space** is the set of all possible outcomes for a given experiment.

\	**T**	**H**
T	TT	HT
H	TH	HH

As indicated above, flipping an honest coin is a random experiment—one has no way beforehand of predicting the outcome. The sample space is a set which contains all possible outcomes. For one flip the possible outcomes are heads (H) or tails (T). For one flip the sample space contains only these two outcomes. For two flips the four possible outcomes are HH, HT, TH, or TT. Thus the sample space is {HH, HT, TH, TT}, containing four elements. Notice the difference between the events HT (heads first) and TH (tails first). The outcome of a single flip is a simple event, whereas the outcome from more than one flip is a **compound event.**

Rolling a standard six-sided (fair) die once would have a sample space with six outcomes: {1, 2, 3, 4, 5, and 6}. Rolling a pair of dice would have a sample space of six times six (6^2) or 36 possible outcomes. Let's construct below the sample space of rolling a pair of dice. In each grid location (square) we must place both the indicated outcome of the green AND the indicated outcome of the red die.

\	**1**	**2**	**3**	**4**	**5**	**6**
1						
2						
3				(4,3)		
4			(3,4)			
5						
6						

Notice that green=3 and red=4 differs from green=4 and red=3. These are like ordered pairs, with the first coordinate the horizontal component (green die) and the second coordinate the vertical (red die). [Note: this convention is in conflict with the convention of (row,column). Please be sure to generate these consistant with those already in the table.] Your homework will make further use of the outcome of the activity below by tying it in with the definition below. You will calculate various probabilities regarding the sum of pips (dots) on the two dice.

For some interactive web sites involving rolling[‡] dies,[§] flipping[¶] or spinning coins[‖]

[‡] http://www.irony.com/igroll.html

[§] http://mrsimon.tripod.com/Random.html

[¶] http://shazam.econ.ubc.ca/flip/index.html

[‖] http://www.kinkaid.org/~jburne/Statistics/Spinning/Quarters.htm check out these links. Be forewarned, however, that if cards[**] or a roulette wheel are involved your internet search is likely to lead you to gambling sites (casinos) whose legality on the web has been and is being challenged due to its addictive[††] nature and those many lives which have been ruined thereby.

1.3 Probability

Probability is denoted by P and specific events by A, B, or C. The shorthand notation used to indicate the probability that event B occurs is $P(B)$.
Empirical (Experimental) Definition of Probability: $P(A) =$ number of times A occurred divided by the times the experiment was repeated.

Classical Definition of Probability:
$P(A) =$ number of event A outcomes divided by the size of the sample space.

The probability of something occurring is related to its frequency. Specifically, when a coin is flipped twice in succession, in 1 of the 4 possible outcomes heads appeared both times. Thus the probability was $\frac{1}{4}$ or 0.25. It is important to remember that the probability of A occurring is less than or equal to one. We have tacitly assumed here that the probability of heads is equal to that of tails. Experiments have been conducted to test this. In such a case, the probability would then be an experimental rather than a theoretical result.

An event with a probability of 0 is **impossible**.
An event with a probability of 1 is **certain**.
$0 \leq P(A) \leq 1$ for any event A.

Probabilities for random events might be computed exactly. In such case we express them as fractions. Other probabilities are obtained by experiment and are thus approximations which are typically expressed to three significant digits unless there are compelling reasons for more or less precision. Probabilities are often given as percentages. In such a case, certainty corresponds with 100% and impossibility with 0%. When probabilities are expressed as percentages, **be sure to include the percent (%) symbol.**

Probability can be approximated by frequency: $P(A) =$ number of times A occurred divided by number of times experiment is repeated.

We used the term **fair** above to describe coins or dies yielding an equal likelihood for any outcome. Thus a fair coin has a 50% of turning up heads and a 50% chance of turning up tails. This is often expressed in terms of odds as 50 : 50. More on that

[**] http://www.bjmath.com
[††] http://www.gamblersanonymous.org

in Lesson 4. Each of the two outcomes is equally likely and thus had a probability of $\frac{1}{2}$. On rare occasions a coin might end up on its side, but generally we exclude such events from the set of outcomes we are considering, just as we generally consider only the genders of male and female. We would thus expect a six sided die to have a $\frac{1}{6}$ probability for any face to be on top. Again, the rare chance of balancing on an edge or corner will generally be excluded, as will be outcomes where the result cannot be determined (such as the die falling into a black hole or sewer grate).

1.4 The Law of Large Numbers

If an experiment is repeated over and over, then the empirical probability approaches the actual probability.

The above statement is often stated as a theorem known as the **Law of Large Numbers.** This is often called the first fundamental theorem of probability. Determining sample size is an exercise in optimizing tradeoffs in cost and accuracy. Large samples should be more accurate but will be more costly, whereas smaller samples cost less but provide less accuracy. Those who have not studied statistics tend to scoff at the idea that a survey of only 1000 (0.001%) people in this country of 100 million voters **can** give a good estimate of how many favor a particular candidate or position. Of course, if your sample is not random, biases will creep in, and accuracy will suffer. Later lessons will explore these concepts in greater detail.

1.5 Random Sample

In a **random sample** each element of the population has an equal chance of being chosen.

The term **random sample** is also used to denote a collection of outcomes that were selected through a representative process. Random samples and the concept of random selection is very important to inferential statistics. Impartial and unbiased sampling often requires careful and thoughtful planning. Such planning is extremely important—bad sample design has delayed many a degree completion!

Name ______________________________ Score ______

1.6 Magic Square Activity: Definitions

Directions: Match the best (numbered) definition with a corresponding (lettered) probability term. Once you have matched several, put the number in the proper space in the magic square box. If the total of the numbers are the same across, down, and both diagonals, you may have correctly matched all items! You may not use your notes/books.

Terms

____ A. Experiment

____ B. Random Experiment

____ C. Simple Event

____ D. Compound Event

____ E. Outcomes

____ F. Event

____ G. Sample Space

____ H. Impossible

____ I. Certain

Definitions

1. Scientific Results.
2. Can't foretell outcome.
3. $P(A) = 1$.
4. Repeated coin flips.
5. Set of all possible results.
6. Possible experimental results.
7. Outcome cannot be broken down.
8. Experimental results.
9. Method by which observations are made.
10. $P(A) = 0$.

A	B	C
D	E	F
G	H	I

Magic number = ____

Name ________________________ Score ______

1.7 Homework for Fundamental Def. for Prob. and Dist.

```
Using the red and green die roll outcome table
you generated in the lecture notes, calculate the
probabilities for the following compound events.
```

____ 1. The total pips on the top faces of two standard dice is 11.

____ 2. The total pips on the top faces of two standard dice is at least 11.

____ 3. The total pips on the top faces of two standard dice is less than 11.

____ 4. The total pips on the top faces of two standard dice is at most 11.

____ 5. The total pips on the top faces of two standard dice is 7.

____ 6. The total pips on the top faces of two standard dice is between 3 and 7, inclusive.

____ 7. The total pips on the top faces of two standard dice is strictly between 3 and 7.

____ 8. The total pips on the top faces of two standard dice is between 2 and 12, inclusive.

____ 9. The total pips on the top faces of two standard dice is 1.

____ 10. The pips showing on the top faces of two standard dice are 5 and 2.

____ 11. The green die has 5 and the red die has 2.

____ 12. The green die has 5 or the red die has 2.

____ 13. Either the green or the red die has a 5 or a 2.

Assume cards are drawn from a normal (no jokers) 52-card deck and **ace is low**.

____ 14. What term is used to descibe the act of randomly drawing any one card?

____ 15. What is the cardinality of (how big is) the sample space?

____ 16. How many outcomes are there where the event is "The card is less than 6"?

____ 17. Calculate P(the card is less than 6).

____ 18. Calculate P(the card is red).

____ 19. Calculate P(the card is a king).

____ 20. Calculate P(the card is between 2 and 6, inclusive).

____ 21. Calculate P(the card is the queen of spades).

____ 22. Calculate P(the card is a standard playing card).

____ 23. Calculate P(the card is a joker).

Probability & Dist. Lesson 2

Counting: Permutations and Combinations

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

In this lesson we introduce several rules used for calculating probabilities: multiplication, addition, factorial, permutations, and combinations. We explore some variations on these themes when there are indistinguishable elements, arranged in a circle, and/or order doesn't matter. Replacement is introduced.

2.1 Rigor into Calculus: Cauchy

Augustin-Louis Cauchy (1789–1857) of France contributed rigor to mathematics. His lectures and researches in analysis during the 1820's clarified the principles of calculus by developing it with limits and continuity. His theory of complex functions forms the basis of physics today. His theoretical work in optics provided a sound mathematical though physically unsatisfactory basis for the supposed pervasive ether thought to conduct light. Cauchy initiated the study of permutation groups, of which Rubik's Cube has many examples.

*Yogi Berra was well known for *non sequiturs* such as the one used above for this lesson's quote. *non sequiturs* is Latin for "it does not follow." Malapropisms (literally ill-suited) or phrases with an inappropriate word were also his forte. "It ain't over until it is over." is perhaps his most famous. Yogi was the greatest catcher in baseball, playing for the Yankees from the mid-1940's into the mid-1960's and then managed both a National League team (the Mets) and an American League team (the Yankees) into a World Series.

2.2 Fundamental Counting Rule

In the previous lesson we put two or more simple events together to create **compound events**. There are various ways of combining such events. Specifically, we might ask the number of outcomes when event A **OR** event B occurs, or we might ask the number of outcomes when event A **AND then** event B occurs. The quantity of outcomes will be used as the numerator when we calculate the probability.

Example: Assume you have 20 M&M® brand candies as follows: 5 orange, 6 yellow, 5 red, and 4 green. In one selection, how many ways can you select either 1 orange or 1 yellow M&M®? What is the corresponding probability?

Answer: Of the 20 M&M's®, 5 are orange and 6 are yellow. Hence 5+6=11 of the M&M's® are yellow or orange. The probability of selecting a yellow or orange M&M® is 11/20=0.55.

The M&M's® are either one color or another, hence getting a certain color is **mutually exclusive** of getting a different color—that is, no M&M's® are rainbow-colored, zebra-striped, or some shade such as orange-yellow or blue-green which thus might be judged different colors by different people. To clarify further the meaning of mutually exclusive, let's say that only one or another event can occur, never both at the same time.

Example: Assume you have 20 M&M's® color distributed as above. If selected **without replacement**, in how many ways can you select two red ones in two selections? What is the corresponding probability?

Answer: For the first selection, five of the 20 M&M's® are red. Since we need to get two reds in only two selections, we need only consider this successful case further, ignoring what happens if we do not get a red on this first selection. For the second selection, only four red of the 19 M&M's® remain. Hence there are $5 \cdot 4 = 20$ ways of selecting two red M&M's® in two selections. The corresponding probability would be: $\frac{5}{20} \cdot \frac{4}{19} = \frac{20}{380} = \frac{1}{19}$ or approximately 0.0526.

The first example above (OR) will be dealt with further below. We will now discuss the second example (AND then). We studied in the last lesson repeated coin flips and die rolls. The size of our sample space, that is the set of all possible outcomes, was the product of the set of possible outcomes for each event: $2 \cdot 2 = 4$ for two coin flips and $6 \cdot 6 = 36$ for rolling two dice.

This is often referred to as the **Multiplication Rule**. It can only be applied if the events are independent. For more on that subject see the next lesson.

If event A can occur in m possible ways and event B can occur in n possible ways, there are $m \cdot n$ possible ways for both events to occur.
$n(A \text{ and then } B) = n(A) \times n(B)$

This is generally expressed as event A and then event B occurring. This is an AND situation where both are performed. This calculation extends to three or more events. For example, if event C can occur in o possible ways, there are $m \cdot n \cdot o$ possible ways for these three events to turn out.

Example: How many different ways can parents have three children.

Answer: For each child we will assume there are only two possible outcomes (thus neglecting effects of extra X or Y chromosomes, or any other chromosomal/birth defects). The number of ways can be calculated: $2 \cdot 2 \cdot 2 = 8$. These can be listed: GGG, GGB, GBG, GBB, BGG, BGB, BBG, BBB where G=girl, B=boy. We could have just as well used the symbols 0 and 1: 000, 001, 010, 011, 100, 101, 110, 111. (Note that this is the same as counting in base 2, the number of Y chromosomes.) This fact can be used to more easily list outcomes or to check for missing outcomes (exactly 4 have boy first, exactly 4 have boy second, exactly 4 have boy last, *etc*). Another way to represent this information is in tree form with the branches from each node representing the possibilities for the next event. Note that this can become very large and thus listing or displaying the complete sample space is often impractical. See also Figure 2.1.

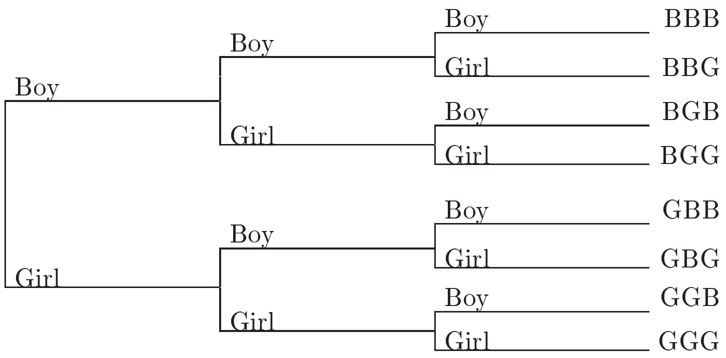


Figure 2.1: Chart showing birth order possibilities for three children.

This is often referred to as the **Addition Rule**.

If event A can occur in m possible ways and event B can occur in n possible ways, there are $m + n$ possible ways for either event A or event B to occur, but only if there are no events in common between them.
$n(A \text{ or } B) = n(A) + n(B) - n(A \cap B)$.

Because often one works with **non-overlapping** events, you will find that the last term is commonly omitted, but added later. It is better to learn the formula correctly the first time and make a special case when the intersection is indeed empty. An empty intersection might occur due to happenstance or it might occur because the events cannot occur simultaneously, *i.e.* the events are **mutually exclusive**. In the M&M® example above, the color selections were mutually exclusive.

We already saw an example of **overlapping** events when we calculated the probability of the green die having a 2 or the red die of having a 5. A careful inspecation of the diagram in the prior lesson indicates that although there are six outcomes where the green die has a 2 and six outcomes where the red die has a 5, we must be careful not to **double count** the event where both the green die has a 2 and the red die has a 5. There are thus only 11 not 12 corresponding outcomes and the probability was $11/36$ or about 0.306.

2.3 Factorial Rule

The factorial rule is used when you want to find the number of arrangements for **ALL** objects. **Example:** Suppose you have four candles you wish to arrange from left to right on your dinner table. The four candles are vanilla, mulberry, orange, and raspberry fragrances (shorthand: V, M, O, R). How many options do you have?

Solution: If you select V first then you still have three options remaining. If you then pick O, you have two candles to choose from. You can compute the number of ways to decorate your table by the factoral rule: for the first choice (event) you have 4 choices; for the second, 3; for the third, 2; and for the last, only 1. The total ways then to select the four candles are: $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$.

These types of problems occur frequently and can be summarized as follows.

Factorial Rule: For n different items, there are $n!$ **arrangements**.

Another word for arrangements is **permutations**. More commonly this word is used as in the section below when not all the objects are arranged. Please recall that the symbol ! is mathematical shorthand for factorial. $n! = n \cdot (n-1)!$ and $1! = 1$. Please also note that by definition and because it makes these types of problems easier, $0! = 1$. $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, $4! = 24$, $3! = 3 \cdot 2 \cdot 1 = 6$, and $2! = 2$.

Try solving this exercise on your own: You need to study, practice football, fix dinner, phone a friend, and go buy a notebook. How many different ways can you arrange your schedule?

2.4 Permutations

Permutation is another name for possible arrangements with **SOME** items from a given set. It is important to remember that order chosen or position arranged is taken into account. Hence permutations are similar to anagrams.[†] Given below is

[†] `http://www.mbhs.edu/~bconnell/anagrams.shtml`

the necessary equation.

$\boxed{{}_nP_r = \frac{n!}{(n-r)!}\text{, where } r \text{ is the number of items arranged from } n \text{ elements.}}$

Example: How many ways can you arrange four figurines from a set of seven? **Answer:** ${}_7P_4 = \frac{7!}{3!} = 7 \cdot 6 \cdot 5 \cdot 4 = 840$. **Alternate solution:** The figurines can be placed as follows: $\underline{\quad 7 \quad}$ $\underline{\quad 6 \quad}$ $\underline{\quad 5 \quad}$ $\underline{\quad 4 \quad}$, which is the same as the factorial notation $\frac{7!}{3!}$.

2.5 Variations on Permutations

2.5.1 Permutations with Repeated Elements

It often happens that objects which are virtually identical get arranged. Our inability to distinguish between these items reduces the number of possible permutations by the number of ways these identical items themselves can be arranged.

Example: the word MISSISSIPPI contains 11 letters of which 4 are S, 4 are I, and 2 are P. If the S's, I's, and P's are **distinguishable** there would be 11! permutations. However, the 4 S's can themselves be arranged 4! different ways as can the 4 I's. The 2 P's have 2!=2 arrangements.
Answer: Thus assuming these repeated elements are truly **indistinguishable,** the number of arrangements would be $\frac{11!}{4!4!2!} = \frac{11!}{4! \cdot 4! \cdot 2!} = 34650$.

2.5.2 Permutations on a Circle

Arrangements are also often made in a circle—we no longer have a left end and a right end. Now our first element placed merely provides a point of reference instead of having n choices. Thus with n distinguishable objects we have $(n$-1)! arrangements instead of $n!$.

Example: Consider arranging the letters ABCD. There are 4!=24 such arrangements. If considered as a circular arrangement there are but 3!=6 arrangements.

Often in circular arrangements only betweenness and not clockwise/counterclockwise is what matters. This further reduces the arrangements by a factor of 2.

2.6 Combinations

Combinations are arrangements of elements without regard to their order or position.

$\boxed{{}_nC_r = \frac{n!}{r!(n-r)!}\text{, where } r \text{ is the number of items taken from } n \text{ elements.}}$

Note that these numbers are the same as those in Pascal's Triangle, the binomial formula, and the binomial distribution. Those less than about four digits should become very familiar.

Example: You have five places left for stamps in your stamp book and you have eight stamps. How many different ways can you select five?

Answer: $\frac{8!}{5!3!} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2} = 56$.
Think of putting them in slots, the first has eight choices, the next slot has seven choices and so forth as demonstrated.

$$\underline{\quad 8 \quad}\ \underline{\quad 7 \quad}\ \underline{\quad 6 \quad}\ \underline{\quad 5 \quad}\ \underline{\quad 4 \quad}$$

Each combination of choosing 5 out of the 8 has permutations of its own. The five can be arranged in the following ways:

$$\underline{\quad 5 \quad}\ \underline{\quad 4 \quad}\ \underline{\quad 3 \quad}\ \underline{\quad 2 \quad}\ \underline{\quad 1 \quad}$$

Thus there are $\frac{8!}{3! \cdot 5!} = {}_8C_5 = 56$ ways to select five of eight, but 6720 $({}_8P_5)$ ways to arrange five of eight.

2.7 Sampling with/without Replacement

Sampling can be done with replacement or without replacement. When done with replacement, the selected object is put back before the next object is selected. When done with replacement, the events remain independent of each other, whereas if done without replacement, they become dependent. More on that in the next lesson.

Name ______________________ Score ______

2.8 Quiz over Permutations and Combinations

Closed book/notes Group. Show Work!

For problems 1-3, assume three unbiased coins are flipped and they land either face up (heads) or face down (tails) with equal probability ($\frac{1}{2}$).

____ 1. What is the sample space? (Not how big is the set, the actual set itself.)

____ 2. What is the probability of getting all heads **or** all tails?

____ 3. List all possible ways of getting exactly 1 head.

____ 4. Assume 4 standard 6-sided dies are rolled and each side has an equal probability of facing up ($\frac{1}{6}$). What is the probability of the pips totals 5? Be sure to show your work.

____ 5. Freddy Fad has 7 Abercrombie shirts and 5 Lee jeans. In how many different ways could he select a shirt-jeans combination?

For problems 6-7 assume the following information. At the MSC weekend reading race there were 10 math books and 14 science books to choose from.

____ 6. In how many different ways could a student select a science and then a math book?

____ 7. In how many different ways could a student select a math and then another (different, *i.e.* done without replacement) math book?

For problems 8-9 assume identical letters are indistinguishable.

____ 8. How many different permutations are there in the letters of: WILLIE?

____ 9. How many different **circular** permutations can be made from: CUCUMBER.

____ 10. Calculate the average growth rate for a portfolio with consecutive annual interest rates: -15%, 25%, 35%, -10%, 20%.

____ 11. Calculate by hand, showing your work, the following: ${}_9C_5$ and ${}_9P_2$.

____ 12. In a class of 28 students, 6 are left-handed, and the rest right-handed. If 9 people are selected at random from this group, what is the probability that: 2 are left-handed and 7 are right-handed? Show your set up!

Name ______________________________ Score ______

2.9 Homework for Counting: Permutations and Combinations

____ 1. Imelda Marcos has 2003 pairs of shoes. In how many different ways can she select a left shoe then a right shoe?

____ 2. Scanti Lee Clad has 15 short shorts and 11 sleeveless blouses. In how many different ways could she select a shorts-blouse combination.

____ 3. Lisa visits Pet Refuge and finds 21 dogs and 13 cats she likes. In how many ways could she select either one dog or one cat? In how many ways could she select both one dog and one cat?

____ 4. The MSC summer reading list contains 12 science books and 15 math books.

(a) In how many different ways could a student select either one science or one math book?

(b) In how many different ways could a student select one science and then one math book?

(c) In how many different ways could a student select one math and then a second math book?

____ 5. Fix Or Repair Daily manufactures light trucks with three different body styles, five different colors of paint, and six different interior colors. Compare the number of trucks necessary to exhibit an example of each with the number of possible varieties.

____ 6. Because of a two hour fog delay only 8 sophomores showed up for math class. Thus the students sit in two table groups of four students each. (Assume each chair gets numbered from 1 to 8.)

(a) In how many different ways could a student be selected to occupy chair 1?

(b) After chair 1 is occupied, how many different ways are there of seating someone in chair 2?

(c) In how many different ways could chair 1 and chair 2 be filled?

(d) If chairs 1 and 2 are occupied, how many ways could chair 3 be filled?

(e) In how many different ways could chairs 1, 2, and 3 be filled?

(f) In how many different ways could all eight chairs be filled?

____ 7. Telephone numbers in the United States and Canada have three groups of digits which meet certain requirements:

- Area Code: 3 digits, the first of which is neither 0 nor 1.
- Exchange: 3 digits, the first of which is neither 0 nor 1.
- Line Number: 4 digits, with 0000 disallowed.

(a) How many possible area codes are there?

(b) How many possible exchanges are there?

(c) How many possible line numbers are there?

(d) How many valid 10-digit phone numbers are there?

(e) What is the probability that a random 10-digit number is a valid phone number?

____ 8. What is the probability that a card drawn at random from a shuffled deck of 52 normal playing cards is a Heart or a Face card? Be sure to avoid double counting!

____ 9. In how many ways could you arrange the following?

(a) Four notebook sections from a set of six sections?

(b) Ten homeworks from a set of twelve homeworks?

(c) Five tests from a set of eight tests?

(d) All 20 questions from a set of 20 questions?

____ 10. How many arrangements can be made from the 26 letters in the English alphabet by using:

(a) 2 different letters?

(b) 3 different letters?

(c) 4 letters without replacement?

(d) 4 letters with replacement?

____ 11. Fourteen people try out for a baseball team. In how many different ways could they select:

(a) the pitcher and then the catcher?

(b) the three outfielders, after the pitcher and catcher have been selected?

(c) the four infielders (1st, 2nd, SS, and 3rd), after the other five team members have been selected?

____ 12. Teacher Thelma says "You may work these five problems in any order you choose." There are 30 students in the class. Is it possible for all 30 students to work the problems in a different order? Justify your answer. Don't just answer yes or no.

____ 13. A 6-letter permutation is selected at random from the letters UNITED. What is the probability that:

(a) The third letter is "I" and the last letter is "T"?

(b) The second letter is a vowel and the third is a consonant?

(c) The second and third letters are both vowels?

(d) The second letter is a consonant and the last letter is "E"?

(e) The second letter is a consonant and the last letter is "T"?

____ 14. Six girls start playing a volleyball game.

(a) In how many ways could the six positions be filled?

(b) In how many ways could the six positions be filled, if Nikki must be server?

(c) If the positions are selected at random, what is the probability that Nikki will be server?

(d) Express the probability in part c above as a percentage.

____ 15. Twelve sophomores line up for a fire drill.

(a) How many possible arrangements are there?

(b) How many arrangements have David and Steph next to each other?

(c) If they line up at random, what is the probability that David and Steph will be next to each other?

____ 16. How many different permutations are there in the letters of: BUBBLES?

____ 17. How many different permutations are there in the letters of: DENNIS?

____ 18. How many different circular permutations can be made from: ARITHMETIC.

____ 19. How many different ways can five boys and five girls sit alternately around a merry-go-round?

____ 20. Four boys and four girls hold hands in a circle, with boys and girls alternating. In how many different ways can they be arranged, if it doesn't matter which side someone is on?

Probability & Dist. Lesson 3

Independence, Complementary Rule, *etc.*

Life consists not in holding good cards,
but in playing those you hold well. Josh Billings

This lesson differentiates between dependent and independent events, defines complementary events, discusses the rules for complements, including the “at least one” situation, and discusses Bayesian Statistics, a type of statistics poised to fundamentally revise our approach to statistics in years to come.

3.1 Father of Inverse Probability: Thomas Bayes

Thomas Bayes (1702–1761) was a 18th century English Presbyterian minister (and statistician) who said that probabilities should be revised when we learn more about an event. Since Thomas Bayes’s father was a Nonconformist, Thomas attended the University of Edinburgh where he studied logic and theology, thus avoiding going overseas. When he studied mathematics isn’t clear, records about his early life are scarce. He served as a pastor with his father before getting his own church. Along with other publications, his essay on probability was published posthumanously. Although his conclusions were accepted by Laplace in 1781, they were challenged by Boole and have remained controversial ever since. Bayes also noted the role systematic errors could play which didn’t reduce your uncertainty the way repeated measurements reduced random measurement errors.

Bayesian probability is now a flavor or interpretation of probability based on partial information instead of complete information about a distribution. Bayes himself might not fully appreciate how his name has been applied to this. A common application now of Bayesian statistics is for classifying e-mail as spam or not spam. Filters to make such judgements have to learn or adapt their behavior based on incomplete information. An example would be the distribution of black and white balls in an urn. The probability of drawing a black ball before any are drawn would be a forward probability question. The distribution of black balls after one or more balls have been drawn is an inverse probability question.

3.2 Dependent *vs.* Independent

When working with the multiplication rule, keep in mind whether or not the events are independent. **Independent events** are those that do not affect each other. Otherwise the events are **dependent**. $P(B|A)$ represents the probability of B occurring after A has already taken place. This is known as the **conditional probability**. It is sometimes read: the probability of B, given A.

$P(A \text{ and } B) = P(A) \cdot P(B)$ if [and only if] A and B are independent.
$P(A \text{ and } B) = P(A) \cdot P(B|A)$ if A and B are dependent.

Sometimes the probability of A and B occurring ($P(A \text{ and } B)$) is given, but the question asks for the probability of B occurring after A. All that requires is solving the algebraic equation, $P(A \text{ and } B) = P(A) \cdot P(B|A)$ for $P(B|A)$, the conditional probability.

Tree diagrams are a method of double checking your work when the sample space is small.

Example: A couple plans on having 3 children. What is the probability of them having two boys and one girl?

Answer:

In Figure 2.1 there are three different ways to have two boys and one girl. Thus the probability is $3/8$ or 0.375. One can also think of the only girl being born first, second, or third. We can do it in a different way: $P(\text{GBB}) + P(\text{BGB}) + P(\text{BBG}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$. Of course, those of us who have done this awhile immediately think in terms of Pascal's Triangle and ${}_nC_r$!

Example: What is the probability of rolling a die twice and getting two sixes?

Answer: $P(6) \cdot P(6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} \approx 0.0278$.

3.3 Complementary Events

In Geometry, complementary angles summed to 90°—these angles together complete a right angle. Another widely used meaning is that **complement** is opposite, or the negation of something. In probability, the complement of event A are the outcomes which do NOT have event A occurring. The notation of the complement of A is a horizontal bar over A ($\overline{A}$). Please note that this spelling and meaning for complement is distinct from **compliment** which means a formal expression of esteem, respect, affection, or admiration.

Example: A local theater group is planning to give away a season ticket via a raffle. Eighty women dropped their ticket stubs in the bucket while only 35 men did. What is the probability of the winning ticket not going to a woman?
Solution: Thirty-five men dropped their stubs of the 115 total tickets. $P(\text{not getting a woman}) = P(\text{man}) = 35/115 = 7/23 \approx 0.304$.

3.4 At Least One

Using the complementary rule with the multiplication rule, one can find the probability of at least one event being what we want. At least one means the same as one or more. The complement of one or more is none. So the multiplication rule is used to find $P(\text{none})$ and then take the complement of it.

$$\boxed{P(\text{at least one}) = 1 - P(\text{none}).}$$

Example: A person deals you a new five card hand. What is the probability of having at least one heart?

Solution: $P(\text{at least one heart}) = 1$ - $P(\text{none}) = 1 - \frac{{}_{13}C_0 \times {}_{39}C_5}{{}_{52}C_5} = 1 - \frac{39}{52}\frac{38}{51}\frac{37}{50}\frac{36}{49}\frac{35}{48} \approx 1 - 0.222 = 0.778$. Just think how long it would have taken if instead you calculated the probabilities for getting one heart, two hearts...!

Please note, the method used above for computing **none** is very general and not well nor widely documented. I'm referring specifically to the expression: $\frac{{}_{13}C_0 \times {}_{39}C_5}{{}_{52}C_5}$. This expression is saying of the 13 hearts we choose 0, whereas of the other 39 cards we choose 5. These two items are multipied together then divided by the number of ways to choose 5 cards from 52. Thus to calculate the probability for getting one heart would be: $\frac{{}_{13}C_1 \times {}_{39}C_4}{{}_{52}C_5}$. (More on this in the section on the hypergeometric distribution.)

3.5 Rules of Complement

As we have seen before, the probability of something certain to occur (occurring 100% of the time) is one. Using the addition rule for $P(A)$ and $P(\overline{A})$, which are mutually exclusive because A and $\overline{A}$ cannot occur at the same time and knowing all that is not in A is in $\overline{A}$, we end up with $P(A) + P(\overline{A}) = 1$.

$$\boxed{P(A) + P(\overline{A}) = 1 \qquad P(\overline{A}) = 1 - P(A) \qquad P(A) = 1 - P(\overline{A})}$$

Example: A farmer expects to bring 80% of a field of wheat to market. How much of the wheat is lost by various means of destruction?

Solution: 20% is destroyed by mice, drought or other means. Remember that percentages are equivalent to probabilities: $80\% = 0.80 = P(A)$. Thus $P(\overline{A}) = 1 - 0.8 = 0.2 = 20\%$.

3.6 Bayesian Statistics and Bayes' Theorem

Bayesian statistics is very much in vogue and is considered by some a different "flavor" of statistics. Specifically, "forward probability" problems, like calculating the probability of picking green socks given the number of various colors in a sock drawer had been solved. He addressed the converse problem, given that so many socks have been drawn, what is the likely color distribution of remaining socks. His **Bayes' Theorem**, also known as **Bayes' Rule**, helped answer such a problem posed by de Moivre, with whom some speculate he studied.

Many medical tests give what are known as false positives. Bayes Theorem is commonly used in paternity suits to calculate the probability that a defendant really is the father of a child, given test results which support such a conclusion. Such tests made recent headlines in the cases of Anna Nicole Smith's daughter (Feb. 2007) and also the Fundamental Church of Jesus Christ of Latter Day Saints (April 2008).

One case of his theorem is as follows:

$$\boxed{P(A|B) = \frac{P(A)\cdot P(B|A)}{(P(A)\cdot P(B|A) + P(\overline{A})\cdot P(B|\overline{A}))}}$$

Example: Suppose two factories produce tires with 60% at factory A ($P(A) = 0.6$) and 40% at the other ($P(\overline{A}) = 0.4$). Suppose further that defect rates differ: 35% for Factory A ($P(B|A) = 0.35$), and 25% for the Factory B ($P(B|\overline{A}) = 0.25$).

Solution: The probability a defective tire came from factory A is as follows: $\frac{P(A|B)=(0.60)(0.35)}{(0.60)(0.35)+(0.40)(0.25)} \approx 0.677$.

Example: Suppose Lyme disease has a prevalence of 0.00207 in the population. Thus 0.99793 do not. 93.7% of those with Lyme disease test positive, but 6.3% give **false negatives.** 3% of those without the disease test positive (**false positives**), thus 97% of those without the disease test negative. We use Bayes' Theorem to calculate the probability that someone who actually tested positive had the disease. For more on these types of errors see Section 13.3.

actual\test	**test=yes**	**test=no**
has LD	real positive 93.7%	**false negative** 6.3%
not have LD	**false positive** 3%	real negative 97%

Solution: $\frac{(0.937)(0.00207)}{(0.937)(0.00207)+(.03)(.99793)} = 0.06085$. It should be clear that the proportion of false positives and false negatives make these test results difficult to interpret! (Amer. J. of Clinical Pathology (1993)).

Name ______________________ Score ________

3.7 Homework for Complements and Bayes

1. Showing your hand work, calculate the following arrangements (see Section 2.4):

   (a) ${}_5P_3$.

   (b) ${}_6P_4$.

   (c) ${}_9P_3$.

   (d) ${}_{10}P_4$.

   (e) ${}_9P_9$.

2. Showing your hand work, calculate the following combinations (see Section 2.6):

   (a) ${}_5C_3$.

   (b) ${}_6C_4$.

   (c) ${}_9C_3$.

   (d) ${}_{10}C_4$.

   (e) ${}_9C_9$.

3. Calculate the number of different 5-card poker hands that can be formed from a 52 card deck.

4. Calculate the number of different 13-card bridge hands that can be formed from a 52 card deck.

5. A set contains six elements. How many subsets are there with 0 elements? 1 element? 2 elements? ... 6 elements? Total subsets?

6. In a class of 24 students, 7 are left-handed, and the rest right-handed. If 8 people are selected at random from this group, what is the probability that

   (a) 3 are left-handed and 5 are right-handed?

   (b) all are right handed?

   (c) all are left-handed?

   (d) Etan & Mij, two of the left-handers, are selected?

7. A three year old tears the labels off 12 soup cans on her mother's shelf. Her mother knows there were 3 cans of tomato and 9 cans of vegetable. The mother selects 4 cans at random.

   (a) What is the probability that exactly 1 of the 4 cans is tomato?

   (b) What is the probability that none of the 4 cans are tomato?

   (c) What is the probability that at least 1 of the 4 cans is tomato?

8. The diagonals of a convex polygon are made by combining vertices two at time. However, some of the combinations are sides not diagonals. How many diagonals are there in a convex

   (a) pentagon?

   (b) heptagon?

   (c) heptadecagon?

   (d) n-gon? Simplify/generalize your answer.

9. Dee Monic is a regular customer at the Red Hot Pepper. The manager figures Dee's probability of ordering hash browns is 0.7; and eggs 0.55. What is the probability that (assume independent):

   (a) She does not order hash browns?

   (b) She does not order eggs?

   (c) She orders neither hash browns nor eggs?

   (d) She orders hash browns and eggs?

   (e) She orders either hash browns, or eggs, or both?

10. Lois Pass has the following probability of passing various courses: Chemistry, 80%; Algebra II, 75%; Computer Programming, 90%. What is the probability of (assume independent)

    (a) passing all 3?

    (b) failing all 3?

    (c) passing at least 1?

    (d) passing exactly 1?

11. Untied Harried Lines flies twin-engine airplanes on its routes. Lab tests show that any one engine has a 0.01 probability (1%) of failure during any particular flight.

    (a) If the engines operate independently, what is the probability both fail?

    (b) Records show both engines fail with probability of 0.001. What is the probability that the second engine will fail after the first has already failed?

    (c) Based on part (b) above, do the engines operate independently?

12. Three programmers are given a program which they work on independently (while occasionally playing interactive games). The probability of each programmer completing a working program by the end of the day are 1/4, 2/5, and 1/5, respectively. Find the probability that at least one working program will be available by the end of the day. (Bonus: Find the probability that only one working program will be available by the end of the day.)

Probability & Dist. Lesson 4

What are the Odds?

> *A taste for the abstract sciences . . . and . . . the mysteries of numbers is excessively rare: . . . But when a person of the sex which . . .encounter[s] infinitely more difficulties . . . to familiarize herself with these thorny researches, succeeds nevertheless in surmounting these obstacles . . . then she must have the noblest courage, quite extraordinary talents, and superior genius.*
>
> Gauss

In this lesson we explore odds, a variant method of expressing probability.

4.1 Against all Odds: Mathematician Germain

Sophie Germain (1776–1831) was born to a middle-class French merchant family in Paris, France. At age 13, the start of the French revolution, she read about Archimedes being so absorbed in his mathematics that he ignored the Roman soldier who killed him. This convinced her mathematics must be very interesting. However, her parents discouraged her by taking away all her candles and night clothes.

Germain was very interested in the teachings of Lagrange and submitted papers and assignments under the pseudonym[*] Monsieur Le Blanc. Lagrange was so impressed that he asked to meet and hence she was forced to reveal her identity. In 1804 she began corresponding with Gauss using the same pseudonym. With the Napoleonic wars she worried that Gauss might die as Archimedes did and wrote her friend the general to personally ensure his safety. The general explained to Gauss that he owed his life to Lady Germain, but he said he had never heard of her. She then wrote admitting her gender. He responded eloquently as quoted in part above.

Germain entered the French Academy of Sciences's contest several times before winning in 1816 with studies on vibrations. She was the first female, except member

[*] Actually, a stolen identity! The failing student left the new Ecole Polytechnic in Paris.

wives, to attend their sessions. Germain's major contributions were in number theory, a subject Gauss wandered away from over time. She made progress on proving Fermat's Last Theorem false for the $n = 5$ case. She also studied Sophie Germain primes, primes p where $2p + 1$ is also prime.[†]

In 1830 the University of Göttingen, at Gauss's prompting, agreed to award an honorary degree to her, but she died of breast cancer before it could be conferred.

4.2 Odds Against/Odds in Favor

Odds against is a ratio of the probability of A not occurring to the probability of A occurring.

A typical method of calculating this would be $P(\overline{A})$ to $P(A)$, where $P(\overline{A})$ is the probability of A not happening and $P(A)$ is the probability of A happening. These might possibly be expressed as a fraction, but beware that odds are not probability. Notice that if these probabilities are expressed as fractions, the denominators of both could be the same and could represent the number of outcomes in our sample space, if unreduced.

The **odds in favor/for** is the "reciprocal" of the odds against: $P(A)$ to $P(\overline{A})$.

Odds against or odds in favor [of] are sometimes left unreduced, but are typically reduced with a "denominator" of 1. They are most commonly expressed in the form $a : b$ or a to b. When one says you have a 50:50 chance of getting heads, this is a typical statement of odds, but can also be interpretted as saying you have a 50% chance of winning and a 50% chance of losing.

Example: Baseball and most such sports commonly use probabilities and not odds. Batting average, or hits per at bat, would be a typical example. Given a batting average of 0.250 and a third of those hits being for extra bases (0.083), what are the odds against getting a hit, getting an extra base hit, or the odds of a hit being for extra bases?

Solution: The odds against getting a hit would be 0.750 to 0.250 which reduces to 3 to 1. The odds of [/against] getting an extra base hit would be 0.917 to 0.083 or, using the original information or three significant figures, 11 to 1. The odds of [/against] a hit being for extra bases would be 0.167 to 0.083 or 2 to 1.

Odds are commonly used in gambling. Some common applications might be horse-racing, the lottery, or the roulette wheel. Assume a typical American-style roulette

[†] Examples: 2, 3, 5, 11, 23, 41, 53, 83, 89, ... Like twin primes, it is not known if there are an infinite number of Sophie Germain primes. Sophie Germain primes > 3 are all of the form $5 \bmod 6$, never $1 \bmod 6$. The number of Sophie Germain primes less than n is about $\frac{2C_2 \cdot n}{(\log_e n)^2}$, where $C_2 = \prod_{p \geq 3} \frac{p(p-2)}{(p-1)^2} \approx 0.660161$ is the twin prime constant.

wheel numbered 00 and 0 to 36. (European-style roulette wheels do not have the 00 and thus give better odds.) We would say the odds against selecting the correct winning slot would be 37 to 1. The odds in favor would be 1 in 37.

Although the use of odds makes it easier to deal with money exchanges resulting from gambling, odds are awkward to use in calculations. That is why they are converted into probabilities when, for example, applying the multiplication rule for combining independent events.

The odds against an event represent the ratio of net profit to the amount bet.

Example: What is the event probability and net profit for a bet which pays 50:1?

Solution: More than likely these are odds against since it was not specified. If you bet \$2, your net profit would be \$100. That is, you would collect a total of \$102. The corresponding probabilities would be 1/51.

Example: What would be the odds against rolling a total of 4 using three regular six-sided dice.

Solution: There is but one way to roll a three (all dies showing ones), but there are three ways (any die shows two, the others show one) of getting the total of four. There are 6^3 or 216 different outcomes. The odds against would be $216 - 3$ to $3 = 213$ to 3. This would be more typically expressed as 71 to 1.

To say one has **long odds** would mean it is unlikely (say, 10 to 1 or 100 to 1).

Example: A 1990's study of over 527,845 birth records noted that the adjusted odds ratio for risk of preterm birth at < 35 weeks of gestation increased as follows: white mother/black father: 1.28 [95% CI, 1.13, 1.46]; black mother/white father: 2.10 [95% CI, 1.68. 2.62]; black mother/black father: 2.28 [95% CI, 2.18, 2.39] and was even higher for extreme preterm birth (< 28 weeks of gestation) in pregnancies with a nonwhite parent.[‡]

4.3 Points Spread

Professional football now uses **points spread** as a way of assigning who is favored to win/lose any given game. These are not odds. These are not probabilities. Converting them into probabilities might be an interesting project. A typical question might indicate that the Titans are favored by 3 points to win. The Broncos are favored by -10.5 points (*i.e.* expected to lose to) over the 49'ers. The Bengals are expected to lose by 6 points. And the Chiefs are favored by 2.5 points. What is the probability of a certain combination of winning and losing occurring? This was a questions posed by a reporter about Dec. 27, 2006 and seems fairly difficult to answer. Any insights would be appreciated.

[‡] *Paternal race is a risk factor for preterm birth.* American Journal of Obstetrics and Gynecology, Vol. 197, Issue 2, Pages 152.e1–152.e7. Palomar, *et al.*

Name ______________________________ Score ________

4.4 Quiz over Probability

Open book/notes	Individual.	Show Work!

1. How large is the sample space if two 6-sided dies are rolled?

2. What term is used to describe a scientific situation where the outcome is unknown in advance?

3. What is the possible range for probabilities.

4. What does it mean for two events to be mutually exclusive?

5. What is the probability, drawing one card from a standard deck of 52 playing cards, of it being a heart or ace?

6. Calculate by hand, showing your work, the following permutation: ${}_8P_3$.

7. In a box of twelve Easter chocolates, three are bunnies. If 4 chocolates are selected at random from this box, what is the probability that two and only two are bunnies?

8. Leo Lazeee has the following (independent?) probability of completing homeworks: Computers: 90%, Algebra II: 80%; Chemistry: 70%. What is the probability of Leo completing at least 1 homework? **Bonus:** Only one?

9. What are the odds against rolling a yahtzee (all five dies the same) on one roll of five dies.

10. How many different permutations can be made from: JESSICA.

Name ______________________________ Score ______

4.5 Homework for Odds in Favor/Against

Complete the following with a short answer.

1. The __________ against Real Quiet winning are posted as 3:1.

2. That corresponds to a __________ of $\frac{1}{4}$ or 25%.

3. The odds __________ Real Quiet winning are then 1:3.

4. The use of odds makes it easier to deal with the money exchanges that result from __________.

5. For __________, the odds against an event represent the ratio of net profit to the amount bet.

6. Odds against event A occurring is the __________ $P(\overline{A})/P(A)$.

7. Odds are expressed in the form $a : b$ where a and b are integers, usually having no common __________.

8. There is no easy __________ rule for calculating odds when combining independent events.

Calculate the **odds in favor** and the **odds against** each event given. Assume regular, fair six-sided dice, an American-style roulette wheel (38 positions numbered 00 and 0 to 36), standard card deck (52 cards in 4 suites, no jokers), *etc.* Be sure to label which is which.

9. Rolling a dice total of 30 with 5 dice.

10. Rolling a dice total of 29 with 5 dice.

11. Being dealt a 4 card hand of cards, all being aces.

12. Being dealt a 5 card hand of cards, none of which are ace or face cards. (Express this first using ${}_{36}C_5$ and ${}_{52}C_5$, rewrite it using factorials, expand these factorials but cancelling common terms, then cancel common factors. *I.e.* reduce these symbolically, without your calculator, showing your work.)

13. The roulette number selected is prime (remember, one is not prime).

14. The roulette number selected is green (the 00 or the 0).

15. The roulette number selected is red (half those not 00 nor 0).

16. Suppose you go to the race track and see the following odds [against] posted. Calculate the corresponding probabilities. White (2 to 1); Black (3 to 1); Blue (5 to 1); and Green (11 to 1).

17. For the odds in the previous problem, sum the probabilities.

Probability & Dist. Lesson 5

Simulating Experiments

If you're trying to train a pilot, you can simulate almost the whole course.
You don't have to get in an airplane until late in the process. Roy Romer

Finding the correct probabilities of events may be difficult to do. At times the correct results may seem to be wrong. The use of simulation can be of great benefit by saving both time and money over the alternative of solving the problem by applying only the abstract principles of probability theory. We define a simulation as follows.

A **simulation** of an experiment is a process that behaves the same way as the experiment, so that similar results are produced.

Computer simulations are now so pervasive that we don't give them much thought, for example, weather forecasting. One of the first computer simulations was for the Manhattan Project during World War II.

5.1 The Father of the A-bomb: Robert Oppenheimer

Robert Oppenheimer (1904–1967) was an American theoretical physicist who became the scientific director of the Manhattan Project which produced the first three atomic bombs (Trinity, Little Boy, Fat Man). He was well known for reading many texts in their original language, such as Sanskrit. Right after the war Oppenheimer became the director of the Institute of Advanced Studies and eventually took Einstein's place as senior theoretical physicist.

Oppenheimer's wife had been married earlier to a communist party member. After the war Oppenheimer opposed nuclear proliferation and was asked to resign. He refused, requesting a hearing to assess his loyalty. Meanwhile, his security clearance was suspended, one day before expiring, and he was viewed by many scientists as a martyr of McCarthyism about 1953. Wernher von Braun quipped to a Congressional committee: "In England, Oppenheimer would have been knighted."

In 1963, President Kennedy awarded Oppenheimer with the Fermi award as a gesture of political rehabilitation. President Johnson presented the award just over a week after Kennedy's assassination—he still had no security clearance, however.

5.2 Let's Play Risk

Some of you may be familiar with the game of Risk, a classic board game of world dominance through aggression, now available* online. Within the game various battles are fought with the outcome determined by the roll of dice. The attacking army typically rolls three red dice to the defenders two white dice. The two highest red dice are compared with the white dice. Each high pair in turn is compared and red wins only if it is bigger. This could be analyzed either exhaustively, by trying all $6^5 = 7776$ combinations or by simulation. To do it exhaustively one might write a program to analyze each result. The results would be 2890 red wins; 2611 split; and 2275 white wins.† Expressed as probabilities: 0.372 red; 0.336 split; 0.293 white. (We classify as atypical and ignore times when the defender can only roll one die as in when he has but one army, or the attacker can roll or rolls only one or two dice.)

Alternatively, one could toss dice repeatedly and tally the results; program the TI-84+ graphing calculator to toss dice repeatedly [`int(1+6rand)`]; (or `randInt`(1,6,5) will roll 5 dice) or download a 30 day evaluation copy of MINITAB‡ and simulate it. One might be tempted to use a standard spread sheet program, but since these were not written and are not endorsed by statisticians one should be very wary of the statistical results they produce. Some have been shown to be just plain wrong, but the software giant(s?) refuse to correct these errors.

A few other common teaching packages include: Fathom§ (much like Geometry Sketchpad and also by Key Curriculum), Data Desk (packaged with ActivStat, a multimedia Statisitical education package). Some professional packages include: SAS (Statistical Analysis System), BMDP, and SPSS (Statistical Package for the Social Sciences). All these packages are produced to provide statistically valid results (unlike many spreadsheets).

A FORTRAN program to calculate the risk probabilities is given below.

```
      INTEGER TOT(0:2)/0,0,0/
      DO 500 I1=1,6
      DO 500 I2=1,6
      DO 500 I3=1,6
```

*`http://www.windowsgames.co.uk/conquest.html`

†John Burnette e-mailed the AP Statistics list server on Sunday, March 26, 2000 with the Risk probabilities and a C program which calculated them.

‡`http://www.minitab.com`

§`http://www.keypress.com/fathom`

```
        DO 500 I4=1,6
        DO 500 I5=1,6
        CALL EVAL(I1,I2,I3,I4,I5,IR)
500     TOT(IR)=TOT(IR)+1
        WRITE(*,*)'2red=',TOT(2),'  split=',TOT(1),'  2white=',TOT(0)
        STOP
        END
*
        SUBROUTINE EVAL(I1,I2,I3,I4,I5,IR)
        J1=I1;J2=I2;J3=I3;J4=I4;J5=I5;IR=0
        IF(J1.LT.J2) THEN
                JT=J1;J1=J2;J2=JT
                END IF
        IF(J1.LT.J3) THEN
                JT=J1;J1=J3;J3=JT
                END IF
        IF(J2.LT.J3) THEN
                JT=J2;J2=J3;J3=JT
                END IF
        IF(J4.LT.J5) THEN
                JT=J4;J4=J5;J5=JT
                END IF
        IF(J1.GT.J4)IR=IR+1
        IF(J2.GT.J5)IR=IR+1
        RETURN
        END
```

5.3 Let's Make Babies (not!)

Let's try another example.

Example: What is the expected average family size if a couple plans to stop having children after having one child of each gender. (No processes, such as timing, acidity, deposition depth, *etc.* are used to enhance gender selection.)

Solution: Instead of using expensive and time consuming methods such as conducting a controlled experiment or surveying a large number of families, we will toss a coin. Both sides (heads and tails) are equally likely. Heads can represent girls and tails can represent boys. For each "family," toss until you get one head and one tail: (H,H,H,T), (T,T,T,H), (H,H,H,H,H,H,T), (H,T), *etc.* After a dozen "families" or so, we will obtain a result close to 3, the theoretical results.

Historically, random number tables were commonly used as a source of random numbers. We will use here the digits of pi which are commonly available. Here we will let boys be represented by the digits 0, 1, 2, 3, and 4 and girls be represented by the digits 5, 6, 7, 8, and 9. (Since the digits of pi are uniformly but also randomly distributed, we could have just as well used even *vs.* odd or perhaps used just 1's and 0's, ignoring the rest.) Starting with 14159 26535 89793 ... we have the following families: (BBBG), (GB), (GGB), (GGGGGB).

This simple simulation will give us good results without much effort. Whenever a simulation is developed, we must be careful to ensure that the process imitates the actual process very well. One could critize this example by noting that boys (0.513) and girls (0.487) are not equally likely, nor are sibling genders necessarily independent.

5.4 Let's Make a Deal

Let's at least set up another example. An old television game show called "Let's Make a Deal," hosted by Monty Hall, generated what is known as the *Monty Hall Problem.*¶ There are three doors with a prize (red Corvette, for example) behind one. You select one door. The host opens one of the remaining doors revealing that it is empty. He then offers you the choice of keeping your door, or switching to the other unopened door. The fact that you should switch because your probability of winning then becomes $\frac{2}{3}$ is far from obvious. A possible scenerio for simulating this would be to have the digits 1, 2, and 3 represent door number 1; the digits 4, 5, and 6 represent door number 2; and the digits 7, 8, and 9 represent door number 3; the digit 0 is ignored. Before each round you would pick: 1) which door has the prize; 2) which door you pick first. This could be done by assigning doors as above and selecting digits in pairs. (We will use the digits of pi again.) Thus 14 would represent door 1 has the prize, but you picked door 2. The host would thus show door 3 as empty. The digits 15 would repeat that scenerio. The digits 92 would represent door 3 having the prize, but you picked door 1, he shows door 2. You would then track how often you win by switching and also by not switching.

One of the first applications of simulation here at Andrews University was back in the mid 1970's to analyze computer terminal usage. The results were published in the infamous 1976 self study. The director of the computing center, LeRoy Botten, had Bruce Ferris, a 14-year old high school drop out, known as The Kid or TK, do the analysis. Simulations are commonly used to forecast weather.‖ "play" war games, analyze nuclear power plants, and other applications where conducting experiments are challenging.

¶http://math.rice.edu/~ddonovan/montyurl.html

‖http://www.weather.com

5.5 Homework for Simulations

```
Note:  Be sure to make clear your procedure for obtaining your random
numbers.  Specifically, be sure to provide enough information to repeat
it EXACTLY!
```

1. Enter the expression `int(1+6rand)` on your calculator then hit `enter` 5 times (or do `randInt`(1,6,5)). Use the first three "rolls" as red dice and the last two as the white. Calculate who won based on the rules printed in the lecture. (DO NOT ADD PIPS: Compare highest red with highest white—red wins only if larger. Compare second highest red with lowest white—white wins if equal or greater.) Repeat for a total of 12 battles. Tabulate your results here. If possible, pool your results with everyone else in the class and compare with the results cited in the lecture. Please start with `0→rand` (where any number could be substituted for 0 and "→" is what the `STO▶` calculator key produces) so your results can be verified and/or repeated by someone else.

2. Using a random number table (see pi below) or random digits on your calculator, generate 15 "families," stopping once each family has both a boy and a girl. Show your work and calculate the average number of children. Be sure to specify which method you used. Please start with `0->rand` (where any number could be substituted for 0) so your results can be verified and/or repeated by someone else.

3. Either link to a *Monty Hall problem*** site and run some simulations there or conduct 25 trials as described in the lecture. The first 55 (since you are ignoring zeroes) decimal digits of pi are: 14159 26535 89793 23846 26433 83279 50288 41971 69399 37510 58209.

[**] `http://math.rice.edu/~ddonovan/montyurl.html`

4. Large values of $n!$ are occasionally needed. It may be impractical or too time consuming to calculate them by direct multiplication. A typical example might be when the Bureau of Fisheries asked Bell Labs for help finding the shortest route for getting samples from 300 locations in the Gulf of Mexico. There are 300! different possible routes. This is also known as the Travelling Salesman Problem.[††] Calculuate values for 50!, 100!, 200!, and 300! as described below and **compare** with offical results. (Give percentage error: (Observed-Expected)/Expected× 100%.
   $n! = 10^K$ where $K = (n + 0.5)\log_{10} n - 0.43429448n + 0.39908993$.
   Note that this is Stirling's Approximation[‡‡] converted to $\log_{10}$. Be sure to use ALL the significant digits since three or four will be lost in the exponent.

$$\begin{aligned} 50! &= 3.0414 \times 10^{64} \\ 100! &= 9.33262 \times 10^{157} \\ 200! &= 7.88658 \times 10^{374} \\ 300! &= 3.06058 \times 10^{614} \end{aligned}$$

5. A student guesses answers to each of the 5 true/false questions on a quiz. Use the decimal expansion of pi and even/odd (don't care which is correct) to estimate the mean number of correct responses for 10 such students.

6. Use the decimal expansion of pi (or a random number table/generator—specify which so your work can be reproduced) to estimate the average number of rolls of a single die necessary to get a 6. (*Hint:* Skip any outcomes that are not 1, 2, 3, 4, 5, or 6.)

7. A ________________________ of an experiment is a process that behaves the same way as the experiment, so that similar results are produced.

[††] `http://www.math.princeton.edu/tsp/`

[‡‡] `http://mathworld.wolfram.com/StirlingsApproximation.html`, $\log_e n! \approx (n + \frac{1}{2})\log_e n - n + \frac{1}{2}\log_e(2\pi)$.

Probability & Dist. Lesson 6

Distributions in General and Expected Value

> *Capitalism fosters a skewed distribution of wealth, whereas socialism fosters the uniform distribution of poverty.* Calkins[*]

In our review of Descriptive Statistics we noted various measures of how a data set was distributed. Special emphasis was given to measures of central tendancy (averages and/or means) and measures of dispersion or how a data set is spread. Assumptions we can make about these important measures are useful in determining how correct the inferences we might try to make about a population are. A study of the common distributions is then in order and will occur over the next several lessons. First we will turn our attention to distributions in general, the "gold standard" for distributions, the normal distribution, and expected values.

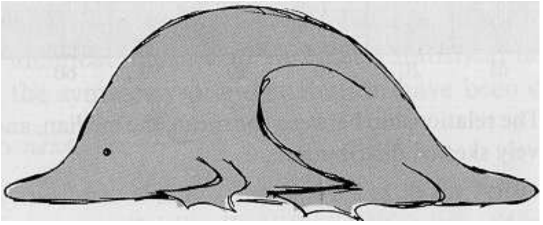
6.1 Generalizer of Integration: Lebesgue

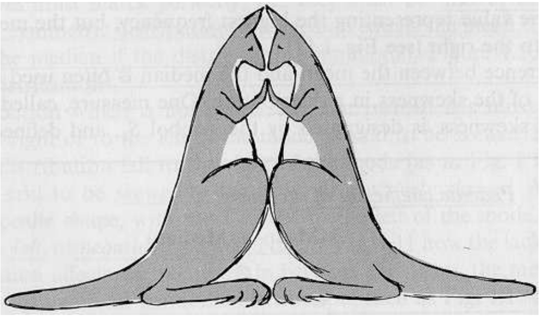
We choose Henri Lebesgue (1875–1941) for this lesson's biograpy because he developed a new theory of integration in his dissertation which can be used to develop expected or expectation value. Specifically, Lebesgue developed the theory of measure and coupled it with general issues related to finding the area under a curve. In general, the Lebesgue integral is employed to find the expected value, although as we present it here, a summation works for discrete distributions.

6.2 Distributions in General

In general, distributions often have an overall shape, center, and spread. There may be outliers, or not. Tails (wings) may be thick or thin. The distribution may be

[*] Modified from Winston Churchill and perhaps others.





skewed to the right or left. The purpose of descriptive statistics and exploratory data analysis was to quantify and/or get a feel for these distribution shapes. The normal distribution is often the "gold standard" to which data sets are compared.

Specifically, whether or not an observation is an **outliers** is, to some extent, a matter of judgment. An outlier is an individual observation that deviates from or falls outside the overall pattern. Outliers, like the old supreme court definition of pornography[†] ("You know it when you see it.") can be hard to define.

The shape of a distribution has very important consequences. Historically, grades were often given based on a curve, specifically the normal curve whereby there were mostly C's, several D's and B's, and few F's and A's. Harvard I think it was, in recent years, has limited the number of A-type grades given. In 2003–04 I think it was 50% and in 2004–05 it was 35%. They are thus forcing a change in the shape of the distribution of grades.

Distributions are commonly **symmetric**. That is, the right and left sides are approximately mirror images of each other. Even uniform or multimodal distributions can be symmetric. If they are not symmetric they are typically **heap shaped** or **mound-shaped**. We term a distribution **skewed to the right** if the right side extends much further out than the left side (usually the mean would then be to the right of the median) or **skewed to the left** if the left side extends much further out than the right side (usually the mean would then be to the left of the median). Skewness can be defined in technical terms with the third moment and exceptions to the mean/median heuristic can both be seen here.[‡]

If the distribution or mound spreads out a lot (like water!) the degree of peakedness or **kurtosis** is low and the distribution is said to be **platykurtic.** A uniform distribution is platykurtic. If the mound piles up like a stalagmite, the distribution is said to be **leptokurtic.** The reference standard is the normal distribution, which is **mesokurtic,** and how peaked a distribution is about the mean in comparison. The images above are from Gosset via Harnett (1975). The shape of a distribution is extremely important.

[†] Actually, obscenity. `http://censorware.net/essays/obscene_jt.html`

[‡] `http://www.amstat.org/publications/jse/v13n2/vonhippel.html`

We usual discuss **Probability Distribution Functions** or `pdfs`.[§] On occasion one works with **Cumulative Distribution Functions** of `cdfs`. These are especially useful for questions which ask for the probability of something being more than some value as opposed to being inside a given range. Both are commonly available, for example, on the TI-84+ graphing calculators. `Cdfs` start out at zero and end up at 1 or 100%. A uniform distribution (defined below) would do so in equal steps. The normal distribution does so by increasing quickly in the middle and slowly for the tails. Like an ogive, a `cdf` is always increasing from left to right.

6.3 Discrete *vs.* Continuous

In Statistics Lesson 1 we also noted that data can be discrete or continuous. Again, it can be hard to differentiate between the two due to quantum mechanics and uncertainties about measurement accuracy. Hence discrete distributions are commonly encountered and continuous distributions are at least possible mathematically. The normal distribution is the most important continuous distributions and whenever n is sufficiently large (generally over 30), we often make assumptions about a discrete distribution derived from the normal distribution but shown to be accurate enough.

First, all probabilities are between 0 and 1 ($0 \le P(x) \le 1$).
Second, all probabilities in a distribution sum to 1 ($\sum P(x) = 1$).
(*i.e.* it is certain your outcome is in the sample space).

We note above two fundamental rules regarding distributions.

Example: Test the following function to determine whether or not it is a probability distribution. $P(x) = \frac{5-x}{10}$ when $x \in \{1, 2, 3, 4\}$.

Solution:

x	$P(x)$
1	$\frac{2}{5} = 0.400$
2	$\frac{3}{10} = 0.300$
3	$\frac{1}{5} = 0.200$
4	$\frac{1}{10} = 0.100$

It works! All probabilities are between zero and one and summing the last column gives $10/10 = 1.000$

6.4 PDFs of a Discrete Random Variable

Consider the probability distribution of two tossed (fair) coins (where heads is x). Note the mound shape.

[§] Not to be confused with Adobe's Portable Document Format.

x	$P(x)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

Consider further the pips displayed on a (fair) die:

x	$P(x)$
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

This is a **constant** function or **uniform** probability distribution.

6.5 Normal Distributions

We will continue by assuming the student remembers several fact about the normal distribution which were reviewed before this series of lessons. Specifically, the normal distribution is also known as the error, bell-shaped, or Gaussian distribution. It is symmetric. If the mean is 0 and the standard deviation is 1, we have a **standard normal** distribution. If the mean is not 0 or the standard deviation is not 1, we have a **non-standard normal** distribution. IQ values with a mean of 100 and standard deviation of 15 is a typical example of a non-standard, approximately normal distribution which we often treat as if it were normally distributed. We use z-scores to convert non-standard normal distributions to the standard normal distribution. The empirical rule states that 68% of normally distributed data falls within 1 standard deviation of the mean, 95% falls within 2 standard deviations of the mean, and 99.7% falls within 3 standard deviations of the mean. In fact, your TI-83+ graphing calculator has the "error function" (`erf`) programmed in under `DISTR (2nd VARS)` and `normalcdf`(lower,upper), where lower and upper are the limits of the region of interest. Tables of values are also commonly available and the ability to read and interpret them is important as well.

The table below gives values for the area between z=0 and z=?, where the final z is initially read down, then the value at the top of the column is added. Alternately, the value at the top of the column can be viewed as the second decimal digit. Such tables may clarify why z-scores are so typically reported to two decimal places! Warning: Although every effort has been made to verify these numbers (on a TI-83 graphing calculator), errors may still be present.

Example: Find the probability for a data value to fall between the mean ($z = 0.00$) and one standard deviation ($z = 1.00$) above the mean, assuming the population

z	**x.x0**	**x.x1**	**x.x2**	**x.x3**	**x.x4**	**x.x5**	**x.x6**	**x.x7**	**x.x8**	**x.x9**
0.0x	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1x	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2x	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3x	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4x	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5x	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6x	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7x	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8x	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9x	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0x	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1x	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2x	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3x	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4x	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5x	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6x	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7x	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8x	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9x	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0x	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1x	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2x	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3x	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4x	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5x	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6x	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7x	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8x	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9x	.4981	.4982	.4983	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0x	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Figure 6.1: Standard Normal Distribution Table.

is normally distributed.

Solution: The table above gives the value 0.3413 or 34.13%. This is the same as what the empirical rule gives $(68 \div 2)$.

Example: Find the probability for IQ values between 75 and 130, assuming a normal distribution, mean = 100 and std = 15.

Solution: An IQ of 75 corresponds with a z-score of -1.67 and an IQ of 130 corresponds with a z-score of 2.00. We can read the value for -1.67 by remembering that the normal distribution is symmetric and then reading the value of 0.4525 off the table. For 2.00 we find 0.4772. The probability of an IQ between 75 and 130 is the same as the probability of an IQ between 75 and 100 plus the probability of an IQ between 100 and 130 or between 100 and 125 (75) plus the probability of an IQ between 100 and 130 or $0.4525 + 0.4772 = 0.9297$. Including a sketch like in Statistics Lesson 6 would be appropriate.

6.6 Expected Value

Let's look at a specific distribution so we can introduce the topic of **expected value**.

Consider the discrete random variable of the sum of pips on two rolled dies. If a random sample is taken, as our sample becomes larger, it becomes clear that the random variable, x takes on any integer value between two and twelve, inclusive. The distribution is likely to become mound-shaped, especially if n is larger than, say, 100. In theory, we would expect our distribution to approach the distribution of $\frac{1}{36}$ for two, $\frac{2}{36}$ for three, up to $\frac{6}{36}$ for seven, $\frac{5}{36}$ for eight, and down to $\frac{1}{36}$ for twelve. See Lesson 1 for the complete sample space. In practice, however, the dies could be weighted on one side, out of square, or even slightly rounded to skew the results. It would be easier, of course, to roll each die separately and verify that its distribution is uniform, with each value occurring with a probability close to $\frac{1}{6}$—see example above. However, such separation of variables is not always possible in the real world. This is often referred to as **lurking** or **confounding** variables. However, the question might arise, as to how big a sample must be taken before we can be "sure" something is amiss. Of course, random variables being what they are, in theory one could roll a million sixes in a row. But also in theory, the probability of this occurring is rather microscopically vanishing. Hence we typically set a threshhold as to how often we would like to be right. 95% is a typical threshhold for non-life threatening situations, whereas 99% or higher is a typical threshhold if more confidence is needed. We refer to these as a 95% confidence level or a 99% confidence level. These correspond with $\alpha = 0.05$ and $\alpha = 0.01$. More on this topic later. As important as these concepts are, we have wandered away from our goal.

We can calculate the **expected value** for total pips by summing the product of the value with the frequency. Thus $2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + \cdots + 12 \cdot \frac{1}{36} = \frac{252}{36} = 7.00$. The value we obtain is the expected value. In this case, it is also the mode.

Example: Find the expected value given the two coin distribution discussed above.

Solution: x takes on the values 0, 1, or 2 with frequency $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. $E = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 0 + \frac{1}{2} + \frac{1}{2} = 1.00$. Thus we expect one head when throwing two coins.

Example: Find the expected value given the one die distribution discussed above.

Solution: x takes on the values one through six with equal probability of $\frac{1}{6}$. $(1 + 2 + 3 + 4 + 5 + 6) \cdot \frac{1}{6} = \frac{21}{6} = 3.5$. Thus we expect 3.5 pips when throwing a fair, six-sided die. Obviously, since pips are discrete, we can't expect 3.5 pips on any one roll!

6.7 Activity for Odds: Word Sort

Teacher: slice the following phases apart.
Students: Sort the following phases into **MEANINGFUL** catagories.

- almost certain
- equal chance for either
- almost impossible
- ≈ 0.0
- ≈ 0.5
- ≈ 1.0
- $\approx 0\%$
- $\approx 50\%$
- $\approx 100\%$
- odds are 999:1 against
- 50:50
- the odds are even
- odds are 999:1 in favor
- winning $1,000,000 in the lottery
- winning $1.00 in the lottery
- winning $0.00 in the lottery
- Getting 10 heads in 10 flips of a fair coin
- Getting “about” 4 heads in 8 flips of a fair coin
- Getting either heads or tails when flipping a fair coin

Name ______________________________ Score ________

6.8 Homework for General Distributions

1. Consider again the random experiment consisting of rolling two dice, a red and a green, as in lesson 1. However, this time, instead of adding the pips, subtract the smaller from the larger, forming the unsigned difference. Perhaps it would be instructive to start by filling in the table below. Now generate a table tallying for each outcome (0 for doubles up to 5 for a six and a one) how many of the 36 outcomes result in that particular value. (One might note the triangular numbers involved, since a square is the sum of two consecutive triangular numbers.) From this please **form** the probability for each value, then create a **bar graph** for the distribution. Finally, note the **mode** or most probable outcome and calculate the **expected value** by summing the product of the value with its probability.

\	**1**	**2**	**3**	**4**	**5**	**6**
1						
2						
3						
4						
5						
6						

2. Statistics show that about 8% of all human males exhibit some form of color-blindness.[¶] Suppose 20 males are selected at random. Let x be the number who are color-blind, and let $P(x)$ be the probability that x of them are color-blind. Using the binomial distribution, TI-83+ program `binomial`, or TI-83+ function `binompdf`(20,.08), calculate $P(0)$, $P(1)$, $P(2)$, $P(3)$, and $P(4)$. Plot a graph of this probability distribution. What is the probability that 5 or more males in a group of 20 are color-blind. How might color-blindness affect the military?

3. Tom Bone plays a musical solo. He is quite good and figures his probability of playing any one note right is 99%. The solo has 50 notes. What is his probability of:

   (a) Getting every note right?

   (b) Making exactly one mistake?

[¶] `http://webexhibits.org/causesofcolor/2.html`

(c) Making exactly two mistakes?

(d) Making at least two mistakes?

(e) What probability per note is necessary for Tom to have a 95% probability of getting all 50 notes correct? (Logs or roots can be used to reduce an exponent.)

4. Suppose a table group decides to do a little penny gambling. Cards will be drawn at random from a standard 52-card deck. For each card drawn you pay a quarter ($0.25). The following payoffs apply: Ace (get $1.75), Face card (get $0.50), Any other card (get nothing). Calculate the expected value for this game. Does this mean you would expect to gain or lose money, in the long run?

5. Dudley Do Wright will get paid by his grandfather for good grades in his college prep classes. His grandfather, however, requires a contract, in advance. Grandpa offers to pay $100 if Dudley makes all A's, or $10 per A. Dudley estimates his probabilities of making A's as follows: Algebra II: 0.75, Chemistry: 0.65, Computers: 0.86, English: 0.96.

   (a) Calculate Dudley's expected value if he chooses $10 per A.

   (b) Calculate Dudley's probability of making all A's.

   (c) Calculate Dudley's expected value if he chooses $100 for all A's.

   (d) Which offer should Dudley choose and why?

6. Many tests have guessing penalties to correct for random guessing. The Advanced Placement (AP) exams, for which our seniors prepare specifically for the AP Calculus AB is a fine example. Other tests, such as SAT and the GRE may be similar.[‖]

   (a) Each question is multiple choice with 5 choices. If you guess randomly, what is the probability of getting a correct answer?

   (b) Suppose you are awarded 1 point for each correct answer but lose $\frac{1}{4}$ point for each wrong answer. What is your **expected value** for any question you randomly guess on?

   (c) Suppose you can eliminate one of the choices as clearly wrong. Now what is your expected value if you randomly guess between the remaining four?

   (d) Repeat the prior subquestion, eliminating two of the five choices.

   (e) Repeat the prior subquestion, eliminating three of the five choices.

   (f) Is it really worth while guessing?

[‖] On the ACT you are scored based on the number right so always answer every question!

Probability & Dist. Lesson 7

The Binomial Distribution and Experiments

To be or not to be, that is the question; Hamlet, Act III, Scene 1

Probability distributions may be either discrete or continuous. The normal (Gaussian) and Lorentzian distributions are good examples of continuous distributions—the random variable can take on any value. Examples of discrete distributions include the Binomial, the Hypergeometric, and the Poisson. We will concentrate on the Binomial and its cousin the Hypergeometric today, and defer discussion on its distant relative, the Poisson, until later.

7.1 Law of Large Numbers: Jacob Bernoulli

The Bernoulli family had many prominent mathematicians but we will concentrate here on Jacob, also known as James or Jacques. Jacob (1654–1705) followed his father's wishes and studied for the ministry. However, he also studied mathematics and astronomy, learning about the work of Boyle, Hooke, and many others during six years of travel throughout Europe. He corresponded with Leibniz and thus became familiar with calculus. In 1682 he returned to Switzerland and founded a school for Math and Science at Basel. Five years later he became a professor of mathematics at the university there, where he remained the rest of his life.

Bernoulli described basic probability theory in *The Art of Conjecture* published eight years after his death. He applied probability to games of chance and introduced the *law of large numbers*. The term Bernoulli trial, or a random experiment with two possible outomes, is named after him. His gravestone was supposed to have a logarithmic spiral on it but instead contains an Archimedian spiral.

7.2 What Makes a Binomial Experiment?

The requirements to be a binomial experiments are as follows:

1. All outcomes of trials must be in one of **two categories**.
2. Trials must be **independent**. One trial's outcome cannot affect the probabilities of other trials.
3. Probabilities must remain **constant** for each trial.
4. There must be a **fixed number of trials**.

The prefix **bi-** has the usual meaning of two in this context, just like bicycle, bifocal, and bigamist. This distribution is related to what happens when you study the expansion of the binomial $(1 + x)^n$. Here it means there are two and only two distinct categories. For instance, students either pass or they fail a test. In dining out at fast food restaurants, people either have or haven't eaten at McDonald's. Requirement 2 specifically implies **with replacement** if we are selecting something, unless the change from not replacing it is slight.

Some notation has become very standard when working with the binomial distributions. S (success) and F (failure) denote possible categories for all outcomes; whereas, p and $q = 1 - p$ denote the probabilities $P(S)$ and $P(F)$, respectively. The term success may not necessarily be what you would call a desirable result. For example, you may want to find the probability of finding a defective chip, given the probability 0.2 that a chip is defective. Here the term success might actually represent the process of selecting a defective chip. The important thing here is to correlate $P(S)$ with p. Some authors avoid q, but the formulae seem clearer using it rather than the awkward expression $1 - p$. The variable x is commonly used for the number of successes.

- $P(S) = p$.
- $P(F) = q = 1 - p$.
- n indicates the fixed number of trials.
- x indicates the number of successes (any whole number $[0, n]$).
- p indicates the probability of success for any one trial.
- q indicates the probability of failure (not success) for any one trial.
- $P(x)$ indicate the probability of getting exactly x successes in n trials.

7.3 Formulas for The Binomial

The formula for calculating $P(x)$ is as follows:

$$\boxed{P(x) = {}_nC_x \cdot p^x \cdot q^{n-x} \text{ where } x \in \{0, 1, 2, ..., n\}.}$$

Here ${}_nC_x$ has the usual definition of entries from Pascal's Triangle and was defined as $\frac{n!}{x! \cdot (n-x)!}$ in Section 2.6. The symbol !, the factorial symbol, has already been introduced in Section 2.3 as shorthand for the product of all the natural numbers up to that number. Thus, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. By definition and convention, $0! = 1$. Note that if $p = q = \frac{1}{2}$, the distribution will be symmetric due to the symmetry in Pascal's Triangle.

Example: Find the probability of having five left-handed students in a class of twenty-five, given $p = 0.1$ ($n = 25$, $x = 5$, $q = 0.9$).

Solution: $P(5) = \frac{25!}{20! \cdot 5!} \cdot (0.1)^5 \cdot (0.9)^{20} = 0.064593$.

Thus, the probability that 5 of the 25 students will be left-handed is about 6%. You should all already have the program `BINOMIAL` on your calculator and be able to recognize and calculate such probabilities. As usual, it is important to set up your solution logically. Carefully identify the important values (n, x, p, *etc.*) before cranking out the numbers and presenting your answer. The TI-83/84 series calculators also have `bionompdf` which, if given the two arguments of n and p, in that order, will output a list of $n + 1$ probabilities for each value of x, with the first one being for $x = 0$. `BINOMCDF` is similar but gives cumulative frequency. Both are under the `2nd` `VARS` or `DISTR` button (were entries 0 and A, so you may need to scroll up/down,; newer calculators have `invt` about 4 and everything else moves down).

It can be shown that the mean, variance, and standard deviation of a binomial distribution can be expressed in simple formulae as follows:

- mean: $\mu = n \cdot p$
- variance: $\sigma^2 = n \cdot p \cdot q$.
- std. dev.: $\sigma = \sqrt{n \cdot p \cdot q}$.

Example: 20 coins are flipped and each coin has a probability of 50% of coming up heads. Find the mean and standard deviation for this binomial experiment.

Solution: $n = 20$, $p = \frac{1}{2}$, so $q = \frac{1}{2}$. $\mu = n \cdot p = 20 \cdot \frac{1}{2} = 10$. This is as expected, we expect heads to come up about half the time. $\sigma = \sqrt{npq} = \sqrt{20 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{5} \approx 2.236$.

Example: Again assume 20 coins are flipped and each coin has a probability of 50% of coming up heads. This time calculate the probability of getting exactly 10 heads.

Solution: $n = 20$, $p = \frac{1}{2}$, so $q = \frac{1}{2}$, and $x = 10$. $P(x) = {}_{20}C_{10}/2^{20} =$

$184756/1048576 \approx 0.1762$.

7.4 The Hypergeometric Distribution

Often sampling will be done *without replacement* from a small finite population. A classic example might be a lottery where 6 different numbers from 54 are selected. Because of the lack of replacement we no longer have independence, thus our probabilities are not constant for each trial. However, the other conditions of the binomial are met. This is a classic application of the **hypergeometric distribution.**

If a population has A objects of one type and B objects of the other type, and if n objects are sampled without replacement, then the probability of getting x objects of type A and $n - x$ objects of type B is:

$$\boxed{P(x) = \frac{{}_AC_x \cdot {}_BC_{n-x}}{{}_{A+B}C_n}.}$$

We already encountered this formula when we found the probability for left-handers and soup cans in Homework for Lesson 3!

Example: A typical state lottery[*] allows a person to select 6 different numbers from 1 to 54 inclusive. Later, a 6-number combination is selected as winning. Various similar results are also awarded prizes. To get the probability of matching all 6 winning numbers, set $A = 6$; $B = 48$; $n = 6$; and $x = 6$. To find the probability of matching exactly 5 winning numbers, leave A and B unchanged and set $x = 5$. The probability of not matching any numbers would be similar with $x = 0$.

Solution: is left as homework questions.

If the population is large compared to the sample size (maybe more than 10 times, that is to say, the sample is less than 10% of the population), the hypergeometric is usually approximated by the binomial and approximated well.

7.5 Normal Approximation for the Binomial Distribution

In 1733 Abraham de Moivre's book *The Doctrine of Chances* introduced a huge time-saver for appoximating binomial distributions for large n. For $n > 69$, one quickly finds 70! exceeds 10^{99} or the limit on many calculators. Historically, $n > 57$ exceeds 16^{63} or the limit on most mainframe computers during the 1960's and 1970's. Thus alternatives were often used when calculating probabilities for such large values of n. In the precomputer age, large tables were constructed to look up probabilities.

It is instructive to examine the binomial distribution for large n and note how it compares with the normal distribution, especially when $p = q = \frac{1}{2}$. As n increases,

[*] `http://www.michigan.gov/lottery`

the probability distribution for values of p and q even further away from $\frac{1}{2}$ looks approximately normal. The common rule is that you can approximate the binomial with the normal when np and nq both exceed some magic number. That magic number is variously stated as 5, 10 or 15, depending on the conservative nature of the statistician, the higher the magic number, the more conservative the statistician. For these notes we will adopt the value 10.

Approximate a binomial distribution by the normal when both $np > 10$ and $nq > 10$.

Since the normal distribution is continuous and the binomial distribution is discrete, we often must apply a **continuity correction**. That is to say, x is no longer represented by a single value, but takes on a range of values from $x - \frac{1}{2}$ to $x + \frac{1}{2}$.

Example: This first example is on the edge of our magic number. Calculate the probability of getting 10 heads when 20 fair coins are flipped, but using the normal approximation to the binomial.

Solution: $n = 20$, $x = 10$, $p = q = \frac{1}{2}$ and as noted above, $np = nq = 20 \cdot \frac{1}{2} = 10$. Using the continuity correction: $x - 0.5$ to $x + 0.5$ and values for the mean (10) and standard deviation (2.236) calculated above, we find z-scores of -0.2236 and $+0.2236$. Using either `normalcdf(-0.2236,0.2236)` under `DISTR` on your TI-84+ graphing calculators, or a table of values, we obtain an answer of 0.1769 or 0.1742, which compare well with the 0.1762 obtained before. (`normalcdf` gives the cumulative area under the normal distribution function between the two z-values given.

Example: Based on U.S. Census data, 12% of U.S. men have earned bachelor's degrees. If 150 U.S. men are randomly selected, find the probability that at least 25 of them have a bachelor's degree.

Solution: $n = 150$; $p = 0.12$; $x > 24.5$. Thus the mean is $np = 150 \cdot 0.12 = 18$; and the standard deviation is $\sqrt{150 \cdot 0.12 \cdot 0.88} = 3.98$. We quickly note that nq is bigger than np since q is bigger than p and note that both are larger than 10. We can thus calculate a z-score of: $(24.5 - 18) \div 3.98 = 1.63$. It is because of the continuity correction that 24.5 is used. We can thus calculate the area under the normal curve by **normalcdf(1.63,9E99)** as 0.05. It can be accessed via `DISTR` (`2nd VARS`) on the TI-84+ calculator.

A JAVA applet to run further examples (and read someone else's notes) can be found here.[†]

[†] `http://www.ruf.rice.edu/~lane/stat_sim/normal_approx/index.html`

Name ______________________________ Score ________

7.6 Magic Square Activity: Binomial

Directions: Match the best (numbered) definition with a corresponding (lettered) probability term. Once you have matched several, put the number in the proper space in the magic square box. If the total of the numbers are the same across, down, and both diagonals, you may have correctly matched all items! You may not use your notes/books.

	Terms	Definitions
____	A. Normal Distribution	1. Has success/failure, fixed, independent trials.
____	B. Binomial Experiment	2. approximating a discrete x with continuous $x \pm \frac{1}{2}$
____	C. Binomial Distribution	3. ${}_nC_x$
____	D. Pascal's Triangle	4. $P(x) = \frac{{}_AC_x \cdot {}_BC_{n-x}}{{}_{A+B}C_n}$
____	E. Binomial mean	5. $n \cdot p$
____	F. Binomial variance	6. $P(x) = {}_nC_x \cdot p^x \cdot q^{n-x}$
____	G. Hypergeometric Dist.	7. $n \cdot p \cdot q$
____	H. Magic number	8. A continuous distribution.
____	I. Continuity Correction	9. both $np > 10$ and $nq > 10$.

A	B	C
D	E	F
G	H	I

Magic number = ____

Name ______________________________ Score ________

7.7 Homework for Binomial/Hypergeometric

1. Separately calculate using the binomial formula the probabilities of getting 0, 1, 2, 3, or 4 left-handed students in a class of 25, given a probability of 0.1. Compare your results with those obtained by doing binompdf(25,.1) (`DISTR 0`) or running `BINOMIAL` on your TI-84+ graphing calculator.

2. Using only the data from the problem above, and the data from the example in the lecture, find the probability of getting more than four left-handed students in a class of 25. Compare your results with those obtained by doing `1-binomcdf(25,.1)` (`DISTR A`) on your TI-84+ graphing calculator.

3. Check the assumptions carefully and see if we are justified in using the binomial (and not the hypergeometric) distribution for the problems above.

4. Calculate the probability described in the text for winning the lottery by matching all 6 of 54 numbers.

5. Calculate the probability described in the text for winning the lottery by matching 5 of the 6 selected numbers from 54.

6. Calculate the probability described in the text for losing the lottery by not matching any of the 6 selected numbers from 54.

7. Use the normal approximation for the binomial to calculate the probability of getting 11 heads in 20 attempts from a fair coin (ignore the magic number test). Be sure to use the continuity correction and calculate the area under the probability density curve from 10.5 to 11.5. Compare this carefully with the results from the binomial formula.

8. Use the normal approximation for the binomial to calculate the probability of getting 12 heads in 20 attempts from a fair coin (ignore the magic number test). Compare this carefully with the results from the binomial formula. Is this the same as the probability of getting 8 heads?

9. Use the normal approximation for the binomial to calculate the probability of getting 13 heads in 20 attempts (ignore the magic number test). Compare this carefully with the results from the binomial formula. Is this the same as the probability of getting 7 heads?

10. How likely is it to get 15 or more heads in 20 attempts, if the coin is fair?

11. A common rule is that you can approximate the binomial with the normal when both ________ and ________ exceed the magic number of ____.

Name ______________________ Score ________

7.8 Penny Activity for Binomial Distribution

There are various ways to flip coins. Some have been shown to be fairer than others. The common alternatives are an **up-in-the-air** flip, **spinning** on the table top, and the **on-edge** procedure used here. These each have their own unique problems. Another procedure, using the pseudo-random number generator on the TI-83, is also subject to any bias the pseudo-random number generator might have. (`int(2*rand)` or `RandInt(0,1,5)`) (and it does have a bias!).

If the coins are equally heavy on both sides, and the rim isn't beveled, it would be reasonable to expect the long-term proportion of heads from this on-edge procedure to be about 0.5. We can state this as:

Our null hypothesis: H_0: $p = 0.5$ The alternative hypothesis: H_a: $p \neq \frac{1}{2}$.

7.8.1 Procedure

It is imperative that these direction be followed carefully. Results not conforming to these specifications will be discarded after an appropriate scientific inquiry. The judgment of the instructor shall be final.

1. Each student (or group of students) places 20 pennies on edge on a flat surface (table).

2. Bang the table just hard enough to cause all the pennies to fall down.

3. Count how many pennies fall **heads up** and record **only** this value on your recording sheet.

4. Accumulate your results with the other students in the class by handing in the group's sheet.

Note: it might be easier for a pair of students to repeat this experiment twice times without regard to which student actually set the pennies upright each time. This also places smaller demands on the total number of pennies required. Also, you may need to skip coins which were mismanufactured or abused.

Note also: Please do not use any foreign (including Canadian) or pre-1959 (Indian head, wheat-back) pennies. It might be instructive to also compare pre-1982 and post-1982 pennies. In 1982 most of the copper was replaced by zinc, resulting in lighter (and cheaper) pennies. Since pennies dated 1982 were of both types (primarily copper and zinc) AND were produced both at Philadephia (plain) and Denver (D) AND both large and small date varieties exist, up to 8 total varieties of pennies may be extant for this date! Some types may be exceedingly rare. Once sufficient post-2008 style pennies become available, it will be interesting to test those separately as well!

Are these results about what you expected, or are you surprised by the results? Do you think it is likely, by chance alone, to obtain results like these results you observed? See typical **penny** results below.

April 13, 2000				April 19, 2001				4/17/2002		4/15/2003		May 5, 2004	
Calkins		Burdick		Calkins		Luttrell		Calkins		Calkins		Calkins	
0		0		0	77	0		0		0		0	
0	9	0	9	0	9	0		0		0		0	9
1	1	1		1	1	1	0	1	1	1		1	011
1	22	1	2222333	1	2233	1	223	1	23	1	23	1	2233333
1	445	1	4	1	5	1	455	1	5	1	4555	1	5555
1	66667777	1		1	666666	1	7	1	66777	1	6666667	1	66
1	889	1		1	8999	1	8	1	88	1	899	1	899
2		2		2		2	0	2		2		2	

May ?, 2005?		May ?, 2006		May 4, 2007		May 5, 2009		5/6–12/2010	
Calkins		Calkins		Calkins		Calkins		Calkins	
0		0	5	0		0		0	
0		0		0	6	0		0	
0		0	9	0		0	9	0	
1		1	001	1	0	1	1	1	001
1	3	1	2223	1	223333	1	2233	1	23
1	44	1	455	1	4445	1	445555	1	4555
1	777	1		1	666	1	66667777	1	66777
1	889	1	9	1		1	889	1	8
2		2		2		2	2		

April 13, 2000: 15.2 ± 2.71 ($n = 18$) and 12.2 ± 1.39 ($n = 9$) combined: 14.2 ± 2.74
April 19, 2001: 14.2 ± 3.78 ($n = 19$) and 14.6 ± 3.06 ($n = 10$) combined: 14.3 ± 3.50
April 17, 2002: 15.5 ± 2.42 ($n = 11$). April 15, 2003: 15.8 ± 1.91 ($n = 16$).
May 5, 2004: 13.9 ± 2.81 ($n = 20$). May 17, 2005?: 16.3 ± 2.12 ($n = 9$).
May ?, 2006: 12.1 ± 3.40 ($n = 13$). May 4, 2007: 13.1 ± 2.59 ($n = 15$).
May 5, 2009: 15.0 ± 2.47 ($n = 23$). May 6&12, 2010: 14.4 ± 2.55 ($n = 15$).

Note: this data is presented in stem-and-leaf format, but with **split stems**. Thus 10 and 11 are together, 12 and 13, 14 and 15, *etc.* Since there might be differences in the sampling techniques, we presented the data and summaries separately before combining them. Note also, some data could not be used because the instructions were not followed. (banging the table too hard making it a coin flip, jiggling the table side to side, introducing a bias, *etc.*)

Probability & Dist. Lesson 8

Queuing Theory, the Poisson and Geometric Distributions

An Englishman, even if he is alone, forms an orderly queue of one.
George Mikes

In this lesson we explore two discrete distributions, the Poisson and the Geometric. The Poisson distribution is closely related to queuing theory which is also discussed briefly. The Geometric distribution is closely related to the binomial, the major difference being the number of trials being of interest, rather than fixed in advance.

8.1 The Father of Queuing Theory: Simeon Poisson

The French mathematician Siméon Denis Poisson[*] (1781–1840) was 8 when the French Revolution started and his military father became president of a district about 50 miles south of Paris. His family apprenticed Siméon to a surgeon uncle, but Siméon lacked the coordination and career interest and soon returned home. Poisson achieved academic success very quickly after he starting studying mathematics in 1798 at the École Polytechnique. His teachers Laplace and Lagrange early recognized his mathematical talents and his paper written about this time attracted the attention of Legendre. Poisson could not draw so avoided descriptive geometry then in vogue. Poisson made important contributions to planetary motion, electrostatics, and heat. He also published important work on definite intregals and made several advances with Fourier series. His Poisson bracket notation for differential equations lives on in quantum mechanics. Although he published over 300 mathematical works, he only worked on one topic at a time, but kept future topics listed in his wallet.

In 1837 he described in a paper the discrete distribution which bears his name. The original applications tended to be rather morbid, the probability of deaths in the

[*] `http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Poisson.html`

Prussian army from the kick of a horse or the number of suicides among women and children. As discussed below, more recent applications have been arrivals at service facilities or the rates at which these services are provided.

8.2 Queuing Theory

Although the term queue isn't in as common usage on the west side of the Atlantic, nonetheless it still has an impact on our daily lives. In technical usage there is a difference between a queue and a line which is often studied in computer science. How this technical difference affects customer satisfaction is the major focus of this lesson.

In many retail stores/banks, management has tried to reduce the frustration of customers by somehow increasing the speed of the checkout/cashier lines. Although most grocery stores seem to have retained the multiple line/multiple checkout system, many banks, credit unions, and fast food providers have gone in recent years to a queuing system where customers wait for the next available cashier. The frustrations of "getting in a slow line" are removed because that one slow transaction does not affect the throughput of the remaining customers.

Walmart,[†] Lowe's,[‡] and McDonald's[§] are other examples of companies which open up additional lines when there are more than about three people in line. In fact, Walmart and Sam's Club have roaming clerks now who can total up your purchases and leave you with a number which the cashier enters to complete the financial aspect of your sale. Disney[¶] is another company where they face thousands of people a day. One method of ameliorating the problem has been to use queuing theory. It has been proved that throughput improves and customer satisfaction increases when queues are used instead of separate lines. Queues are also used extensively in computing—web servers and print servers are now common. Banks of 1-800 service phone numbers (and telemarketers) are a final example I will cite.

Queuing theory leads one directly to the Poisson Distribution[‖] discussed here. As hinted above, operations research has applied it to model random arrival times.

8.3 Poisson Distribution

The Poisson distribution is the continuous limit of the discrete binomial distribution. It depends on the following four assumptions:

[†] `http://www.walmart.com`

[‡] `http://www.lowes.com`

[§] `http://www.mcdonalds.com`

[¶] `http://www.disneyworld.com`

[‖] `http://mathworld.wolfram.com/PoissonDistribution.html`

1. It is possible to divide the time interval of interest into many small subintervals (like an hour into seconds).

2. The probability of an occurrence remains constant thoughout the large time interval (random).

3. The probability of two or more occurrences in a subinterval is small enough to be ignored.

4. Occurrences are independent.

Clearly, bank arrivals might have problems with assumption number four where payday, lunch hour, and car pooling may affect independence. However, the Poisson Distribution finds applicability in a surprisingly large variety of situations.

The equation for the Poisson Distribution is:

$$\boxed{P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}}$$

The number e in the equation above is the base of the natural logarithms or approximately 2.71828182845904523... **The variance for the Poisson Distribution is equal to the mean.** In fact, this can be a quick check to see if this distribution can be applied. Traditionally the Greek letter lambda (λ) is often used for the mean instead of μ. The differences between the Poisson distribution and the binomial distribution are:

1. The binomial distribution is affected by the sample size and the probability while the Poisson distribution is ONLY affected by the mean.

2. The binomial distribution has values from $x = 0$ to n but the Poisson distribution has values from $x = 0$ to infinity.

Example: On average there are three babies born a day at Hospital A with hairy backs. **A.** Find the probability that in one day two babies are born hairy. **B.** Find the probability that in one day no babies are born hairy.

Solution: A. $P(2) = \frac{3^2}{2} \cdot e^{-3} = .224$ **B.** $P(0) = 3^0 \cdot e^{-3} = .0498$.

Example: Suppose a bank knows that on average 60 customers arrive between 10 A.M. and 11 A.M. daily. Thus on average one customer arrives per minute. Find the probability that exactly two customers arrive in a given one-minute time interval between 10 and 11 A.M.

Solution: Let $\mu = 1$ and $x = 2$. $P(2) = \frac{1}{2}e^{-1} = \frac{1}{2}0.3679 = 0.1839$.

Example: Other examples include, the number of typographical errors on a page, the number of white blood cells in a blood suspension, or the number of imperfections in a surface of wood or metal. I assume one could apply it to finding four-leaf clovers, but a corresponding class activity has not yet been developed.

For various Java applets, include one for the Poisson Distribution, visit this location.[**]

8.4 Geometric Distribution

The requirements to be a geometric experiments are as follows:

1. All outcomes of trials must be in one of **two categories**.
2. Trials must be **independent**. One trial's outcome cannot affect the probabilities of other trials.
3. Probabilities must remain **constant** for each trial.
4. The variable of interest is the number of trials required to obtain the first success.

Note that it is the last item above the number of trials, which differentiates this from a binomial experiment. Like the binomial, however, the geometric distribution is discrete.

Example: Suppose we are to roll a die until 6 pips are up. We will call this event a success, any other number a failure. The random variable x will be the number of rolls it takes for this to occur. We roll again with failure and quit with success.

Counterexample: Suppose we are to draw a card without replacement from a standard 52-card deck until we draw an ace. Although we have two categories, the trials are independent, and the variable of interest is the number of trials, the probability changes after each trial. See the section on the hypergeometric for how to deal with this situation.

If x has a geometric distribution with a probability p of success and $q = 1 - p$ of failure on each observation, then $x \in \{1, 2, 3, \ldots\}$ and $P(x = n) = q^{n-1}p$.

If x has a geometric distribution with a probability p of success, then $\mu = \frac{1}{p}$, and $\sigma^2 = \frac{q}{p^2}$.

Note: the mean is also the expected value.

If x has a geometric distribution with a probability p of success, then $P(x > n) = q^n$.

Example: Again, we will roll a die until a 6 appears, counting the number of rolls required. Find the mean, standard deviation, and probability it will take more than 6 rolls.

Solution: $p = \frac{1}{6}$ so $q = \frac{5}{6}$. $\mu = 6$. $\sigma^2 = \frac{5/6}{1/36} = 30$ so $\sigma = 5.477$. $P(x > 6) = q^6 = 0.3349$.

[**] `http://www.stat.vt.edu/~sundar/java/applets/Distributions.html#POISSON`

Name ______________________ Score ________

8.5 Homework for the Poisson Distribution

1. Suppose a bank knows that on average 60 customers arrive in a certain service hour. Using a time interval of 1 minute, calculate the probability of exactly **one** customer arriving in a given one minute interval within that hour. Use the example in the lecture for exactly two as a pattern.

2. Suppose a bank knows that on average 60 customers arrive in a certain service hour. Using a time interval of 1 minute, calculate the probability of **no** customers arriving in a given one minute interval within that hour.

3. Suppose a bank knows that on average 60 customers arrive in a certain service hour. Using a time interval of 1 minute, calculate the probability of exactly **three** customers arriving in a given one minute interval within that hour.

4. Suppose a bank knows that on average 60 customers arrive in a certain service hour. Using a time interval of 1 minute, calculate the probability of **more than three** customers arriving in a given one minute interval within that hour.

5. **Graph** the probability distribution determined in the problems above (and the lecture example).

6. Assume a finite population as follows: $\{1, 2, 3, 4, 5, 6\}$. Note how there are $N = 6$ possible samples of size $n = 1$ and that each element is its own sample mean. The mean of these sample means is obviously the population mean.

   (a) Although our samples are too small ($n = 1$) to form a sample standard deviation, we can calculate the population standard deviation.

   (b) Now **calculate** the 15 sample means, without replacement, for all ${}_6C_2 = 15$ samples of size $n = 2$: $\big\{$\{1,2\}, \{1,3\}, \{1,4\}, ...\{5,6\}$\big\}$.

   (c) Once you have the 15 sample means, **calculate** their mean and standard deviation.

   (d) **Compare** this with the mean and standard deviation of the original population ($N = 6$).

   (e) What is the approximate relationship between the two standard deviations (or variances)? Should we be using sample or population standard deviation?

7. Suppose we are to draw a card with replacement from a well-shuffled, standard, 52-card deck. Find the mean, standard deviation, probability that more five draws will be required for the following conditions.

   (a) Drawing an ace ($p = \frac{1}{13}$).

   (b) Drawing a face card ($p = \frac{3}{13}$).

   (c) Drawing a heart ($p = \frac{1}{4}$).

   (d) Drawing an even card ($p = \frac{5}{13}$).

Probability & Dist. Lesson 9

The Lorentzian Distribution and Voigt Profiles

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

Another commonly encountered distribution is the **Lorentzian Distribution,**[*] also known as the **Cauchy Distribution**, and apparently discovered by both men. It is continuous.

9.1 A Bright Idea: Lorentz

Hendrik Lorentz (1853–1928) of the Netherlands (jointly with Zeeman) won the 1902 Nobel Prize for Physics for his theory of electromagnetic radiation. His doctoral thesis of 1875 refined Maxwell's theory (1865) so as to better explain the reflection and refraction of light. Visible light is a narrow part of the broad electromagnetic spectrum which extends from long wavelength radio waves to short wavelength X-rays, and beyond—both ways. Electromagnetic radiation (a photon) is a precise oscillation of an electric and a magnetic field. Applied/external electric and magnetic fields have an effect on this oscillation and hence change the corresponding wavelength (λ). Wavelength and frequency are inversely related with the speed of light (c) as the proportionality constant ($\lambda = c/f$). The **Lorentzian Transformation,** with it's time dilation and length contraction superceded the law of gravity of Galileo/Newton and forms the basis of Einstein's 1905 work on Special Relativity.

[*] `http://cyberbuzz.gatech.edu/technique/issues/spring2002/2002-03-29/18.html`

9.2 The Lorentzian Distribution

The Lorentzian Distribution is often[†] used to describe **resonance** behavior, things like swings (pendula) swinging, a bow on a violin string, a thin goblet shattering when the fat lady sings, that dreaded microphone feedback, or the rhythmic wind gusts which destroyed the Takoma Narrows[‡] Bridge.[§] Soldiers learn early to break stride when crossing a bridge. A radio or TV receiver is tuned to resonate in response to a specific frequency, typically by changing the capacitance or inductance. Under resonance, energy flows rhythmically between the capacitor's electric field and the inductance's magnetic field, just like the interchange of potential (energy of position) and kinetic (energy of motion) energy in a swing. Like the Gaussian, the Lorentzian is symmetric, unimodal, and continuous. Under very general assumptions the following formula can be derived:

$$\boxed{\frac{A\Gamma/(2\pi)}{(\omega-\omega')^2+(\Gamma/2)^2}.}$$

By inspecting this equation closely we can see it is symmetric and has a maximum when $\omega - \omega'$. ω (omega) is called the **driving frequency** whereas ω' is the **resonance frequency.** A is the area under the curve and Γ (upper case Greek letter gamma) is called the **full width at half maximum** or **FWHM**, the parameter which characterizes the spread of the distribution. FWHM is also commonly called the **linewidth** or **halfwidth**.

The Lorentzian distribution tends to be lower[¶] with fatter tails (often called wings) than a Gaussian distribution with equal FWHM. In fact, the wings are so extended that the standard deviation is (and higher moments are) not defined (the integrals are unbounded or there is no average distance from the mean)!

9.3 Resonance and The Second

Resonance phenomena have become important in keeping time. In 1967 a resonance experiment using an atomic beam of cesium became the definition of the second. The transition between the ground $F = 3$ and $F = 4$ state was observed to within one part in 10^{11} (10 parts per trillion or 10 ppt), one of the most accurately measured frequencies known at the time. It is so precise, 9 192 631 770 hertz (cycles

[†] http://physicslabs.phys.cwru.edu/MECH/Manual/Appendix_VI_distributions.pdf

[‡] http://www.ketchum.org/bridgecollapse.html

[§] http://www.wsdot.wa.gov/TNBhistory/

[¶] http://www.phys.unsw.edu.au/~mgb/pg_mod3_lec5/node16.html per second), that the decision was made to make it the definition of the second.

The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

Leap seconds[*] are now occasionally used to adjust between time as kept by the atomic cesium fountain clock and time as determined by the earth's motion on its axis and around the sun. In 1997 the International Committee on Weights and Measures confirmed that the definition of the second was at 0 K, in other words, the coldest possible temperature. A common fallacy you will find in many popular writeups is that "all motion ceases" at 0 K. This is quantum mechanically absurd. We expect the definition of the second to change in the near future from the 9 GHz range to an atomic transition frequency in the 100's of Thz range.

Lorentzians are also applied to: DNA microarray data[†] used to measure gene expression; edges[‡] in MR brain images; and muon[§] spin relaxation theory.

9.4 Voigt Profiles

Often several parameters influence a line profile. Although the natural linewidth may be Lorentzian, doppler broadening, caused by the random thermal motion, will be Gaussian. Small magnetic fields may cause multiple Lorentzians to overlap. The combination[¶] of these effects results in a **convolution**[‖] of Lorentzian and Gaussian distributions and is known as a **Voigt profile**. Among other things, Voigt profiles allow us to measure the temperature and pressure of the emitting or absorbing layers in stellar atmospheres.

9.5 Calkins Research

These Voigt profiles had a strong influence on my Ph.D. research. Cesium is an alkali metal meaning there is one electron outside a closed shell (column I, the leftmost column of the periodic table). Cesium is the heaviest naturally occurring alkali metal and as noted above is used in the definition of the second. Only one isotope occurs naturally (^{133}Cs). Cesium is liquid near room temperature, one can melt it in their hand (through a glass vial, please, remember the alkali metals are highly reactive). It

[*] http://tycho.usno.navy.mil/leapsec.html

[†] http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=130571

[‡] http://dutnsic.tn.tudelft.nl:8080/main/node19.html

[§] http://musr.physics.ubc.ca/theses/Kojima/node24.html

[¶] http://musr.org/intro/ppt/GardenExport/node7.html

[‖] Formally, the convolution of functions f and g, written $f * g$, is integral of the product of the two functions after one is reversed and shifted: $(f * g)(t) = \int_a^b f(\tau)g(t - \tau)\, d\tau$.

is pale gold in color (not silvery white as many reputable sources claim). It is the most electropositive element, forms the strongest base known, and has a high vapor presure. It is a well studied system. In fact, this system, through parity nonconservation, allows atomic physicists to garner information about the weak force. Also, the nucleus has a large spin ($I = \frac{7}{2}$) so that the nuclear octopole moment was recently measured.[*]

There is a large separation between the fine and hyperfine states. Whereas the D_1 ($^3S_{1/2}$ to $^3P_{1/2}$) and D_2 ($^3S_{1/2}$ to $^3P_{3/2}$) transitions are only 6 nm apart in sodium (witness the yellow of a high pressure sodium lamp), the D_1 ($^6S_{1/2}$ to $^6P_{1/2}$) and D_2 ($^6S_{1/2}$ to $^6P_{3/2}$) transitions are 42 nm apart in cesium. However, they are in the infrared (894 nm and 852 nm). The natural lifetime of these cesium states is about 30 ns hence the natural linewidth is about 5 MHz. Although one would expect doppler broadening on the order of 400 MHz for cesium, by forming a highly collimated thermal beam we are able to reduce that to about 5 MHz. 894 nm corresponds to a frequency of about 335 THz. My research involved precisely finding the peak of this Voigt profile to less than 3 kHz. We achieved an accuracy of 7 parts per trillion (0.7 parts in 10^{11}), or an improvement of an order of magnitude over Hänsch's 1999 result. Another way to say this is that our results were precise to **twelve** significant digits! We found the D_1 centroid to be 335 116 048 748.1(2.4) kHz. This is clearly one reason for my insisting on at least 3, but no more than 5 significant digits, unless clearly indicated. Note this last notation is worth further discussion. The (2.4) given after the value is the one standard deviation error bar. We are thus 95% confident that the true value is within ±4.7 kHz of the value given, where $1.96 \cdot 2.4 = 4.7$.

The new precision on the D_1 frequency in cesium allowed an improved measurement of alpha, the fine structure constant or electromagnetic coupling constant which is a fundamental constant of nature.

Shown below are typical scans of the F4 to F3 transistion taken on April 8, 2004. A 1 Gauss magnetic field has been applied in the z (vertical) direction to the bottom scan. Below that are fits of the 1 Gauss scan using a Gaussian (left) and Lorentzian (right). A better fit is found using a combination (Voigt profile). Due to good symmetry, however, any of these approaches finds the center to within about the same 1 kHz.

This research took place in the very room at NIST in Boulder CO where the Krypton-86 wavelength was precisely measured. The Krypton-86 wavelength was used to define the meter from 1960 until 1983 when that research lead to a redefinition of the meter in terms of the speed of light. We used the femtosecond laser frequency comb which is also being used to calibrate the Mercury and Calcium frequencies which might well become the new THz time standards in the near future. My dissertation can be located here.[†]

[*] `http://www.nd.edu/vgergino/cv/thesis_Final.pdf`

[†] `http://etd.nd.edu/ETD-db/theses/available/etd-04112005-175414/`

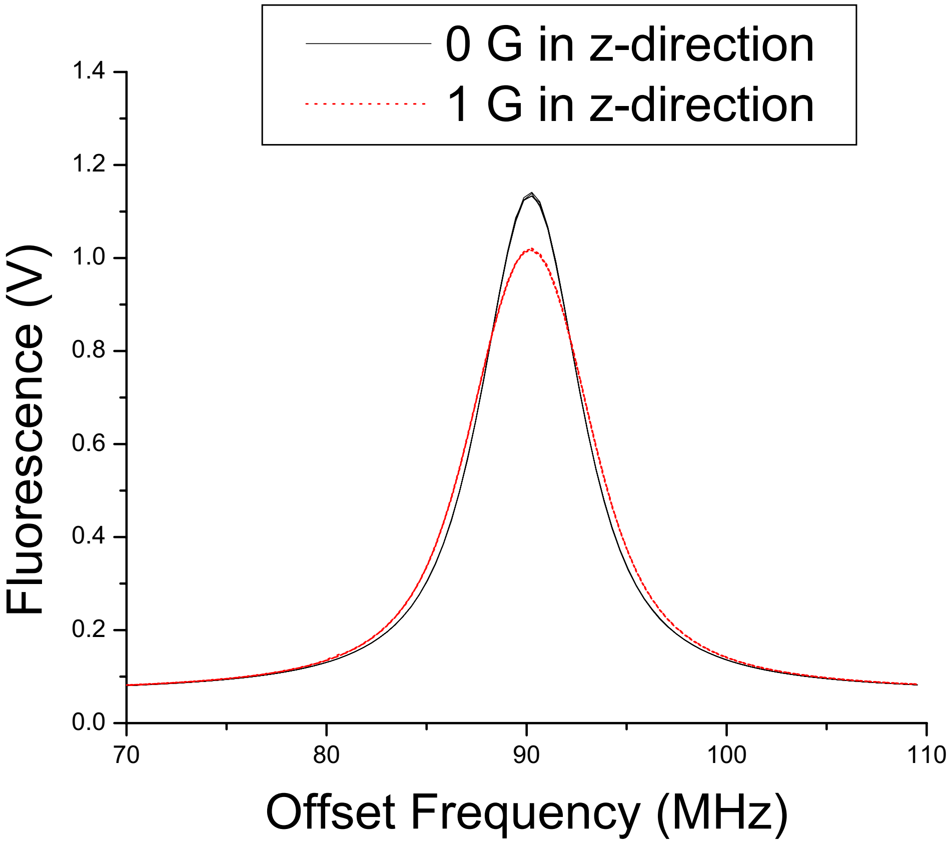


Figure 9.1: Calkins Research Data, Florensences *vs.* Offset Frequency, Comparing Runs With and Without a magnetic field.

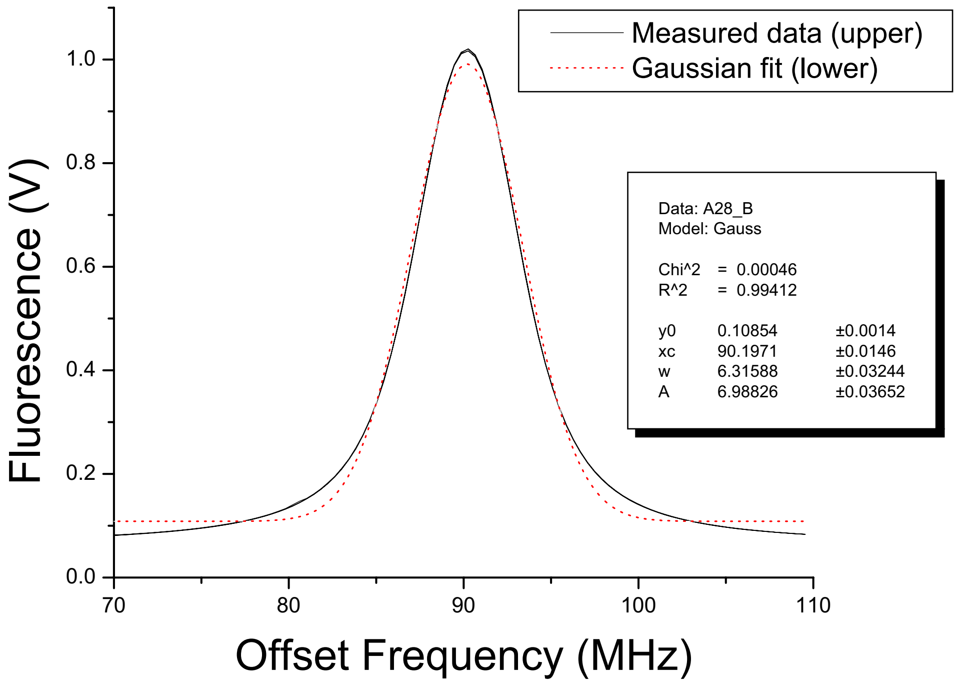


Figure 9.2: Calkins Research Data, Florensences *vs.* Offset Frequency, Fitted With Only a Guassian.

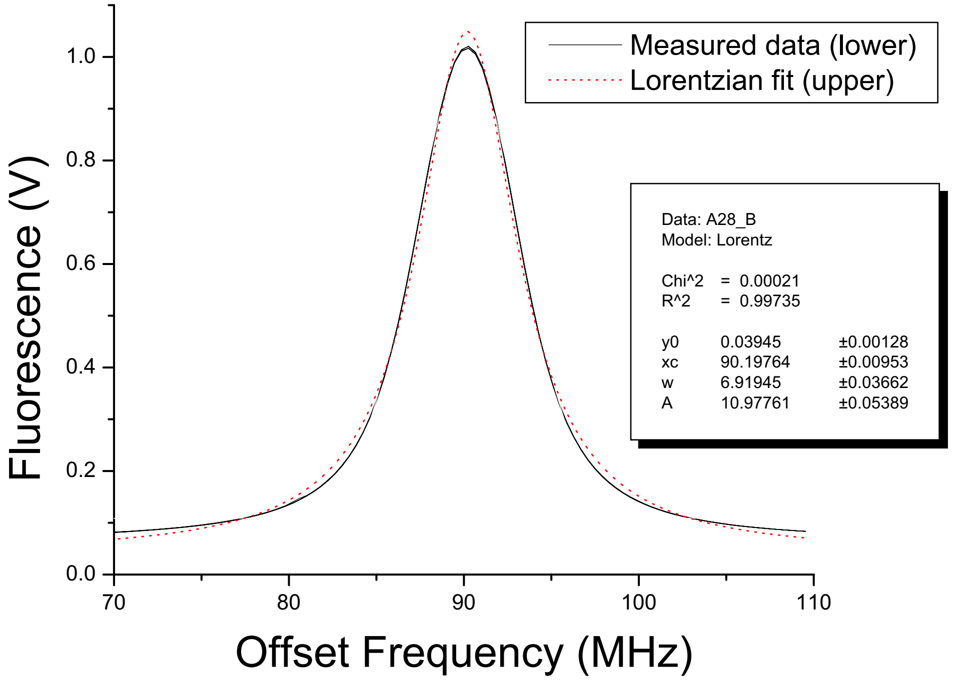


Figure 9.3: Calkins Research Data, Florensences *vs.* Offset Frequency, Fitted With a Only a Lorentzian.

9.6 Review Activity: Graphic Organizer

Directions: Each small group writes each word and its definition on an index card after discussing it. They then arrange the cards into a graphic organizer. Each student needs a list and each group a set of 24 index cards.

- Experiment
- Events
- Simple events
- Compound events
- Independent events
- Dependent events
- Mutually exclusive
- Complementary events
- Odds
- Odds for
- Odds against
- Probability
- Conditional probability
- Impossible
- Certain
- Sample space
- Outcome
- Random selection
- Combinations
- Permutations
- Chosen with replacement
- Chosen without replacement
- Simulations
- Random numbers
- Bayes' Theorem

Name ________________________________ Score _______

9.7 Homework for Lorentzians

1. Lorentz developed a ___________ used in ___________ Relativity which superceded Newton's Laws of Motion giving us time ___________ and length ___________.

2. Lorentz's theory of ___________ ________ better explained the ___________ and ___________ of light.

3. The Lorentzian Distribution is often used to describe ___________ behavior, like the interchange of ___________ and __________ energy. It is ___________, unimodal, and ___________.

4. The standard deviation of a Lorentz Distribution is ___________.

5. A Voigt Profile is a ___________ of Lorentzian and Gaussian Distributions.

6. Cesium is an _________ metal, _________ near room temperature, and used in the definition of the ________. Only ______ isotope(s) occur(s) naturally.

7. Calkins got ________ significant digits or ________ parts per trillion in his research.

8. High pressure sodium lamps give a __________ light near 590 nm. The corresponding transitions in Cesium occurs in the ___________ near 852 and 894 nm.

9. ___________ and ___________ frequencies might well become the new THz ___________ ___________.

10. Measure the FWHM for both Voigt Profiles in the first figure in the lesson, using the ends of the plot as base. Try for two significant digits.

Probability & Dist. Lesson 10

The Student t-Distribution

Fisher would have discovered it anyway. William Gosset

In this chapter we review the Student t-distribution, discuss degrees of freedom, confidence intervals, margin of error, the two sample t test, and matched pair test. There is substantial overlap between this lesson and the Introduction to Statistics, Lesson 9.

10.1 The Father of the t-Distribution: Wm. Gosset

William Gosset (1876–1937) was a Guinness Brewery employee who needed a distribution that could be used with **small samples.** Since the Irish brewery did not allow publication of research results, he published under the pseudonym of Student. You should know and we will show in the next lesson that the distribution of sample means for large samples approaches a normal distribution. 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.

10.2 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

*Such a predicament is often referred to as a Catch-22, from Heller's 1961 novel set in World War II.

made, but confidence intervals for the population mean are to be constructing using an unknown σ, then the distribution known as the **Student t-distribution** can be used.

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. 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 (unimodal and basically symmetric)

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

$$\boxed{t = \frac{\bar{x} - \mu}{s/\sqrt{n}}}$$

10.3 Vocabulary: DF, α, CI, Margin of Error

10.3.1 Degrees of Freedom

Actually, since the population mean is likely also unknown, often the t-score will be looked up based on the level of confidence desired and the **degrees of freedom** and the population mean estimated. Degrees of freedom is a fairly technical term which permeates all of inferential statistics. In many cases it is $n - 1$.

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.

10.3.2 Alpha, Type I Error

Formally, α, or the probability in hypothesis testing of rejecting the null hypothesis when it is in fact true, is known as a type I error. It is also called a false negative. For example, a woman takes a pregnancy test and it gives a negative result, but she is in fact pregnant. Alpha is also called our **level of significance.** Historically, fixed values, such as $\alpha = 0.05$ or $\alpha = 0.01$ were used.

10.3.3 Confidence Intervals

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-83+ graphing calculator. `DISTR 5` gives `tcdf`. tcdf expects three arguments, lower t value, upper t value, and degrees of freedom. Since no inverse t function is 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.

10.3.4 Margin of Error

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:

$$\boxed{\text{ME}=t_{\alpha/2}\frac{s}{\sqrt{n}}}$$

Your confidence interval should look like: $\bar{x}-\text{ME} < \mu < \bar{x}+\text{ME}$.

10.4 Table of t Values

The headings in the table below, such as .005/.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.

Deg. of Free.: 1/2 tails	**.005/.01**	**.01/.02**	**.025/.05**	**.05/.10**	**.10/.20**
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	2.132	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
$z\ (\lim_{n\to\infty})$	2.576	2.326	1.960	1.645	1.282

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 interpretted 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 outliners or strong skewness, do not use the t. Since the assumption that the samples are random is more important that the normality of the underlying population distribution, the t statistic can be safely used even when the sample indicates the population is clearly skewed, if $n > 40$.

10.5 Two sample t Tests

Often one wants to compare two treatments or populations and determine if there is a difference. This can be done either with or without matching. We will discuss first the unmatched situation. Two assumptions are used: two independent simple random samples from two distinct populations (matching would negate independence); and both populations are normally distributed with unknown means and standard deviations. Our null hypothesis would look like H_0: $\mu_1 = \mu_2$ or we might want to give a confidence interval for the difference $\mu_1 - \mu_2$. We use the sample means and standard deviations to estimate the unknown parameters. Although the statistic $\bar{x}_1 - \bar{x}_2$ has a normal distribution in terms of the combined population variance, when we use the combined sample variance, we do not obtain a t distribution. Nonetheless, we do use the t distribution for hypothesis testing in this case. The two-sample t statistic is as follows:

$$\boxed{t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}}$$

The expression in the denominator reflects the way **variances** sum (standard deviations do not sum).

There are two options for obtaining a value for the degrees of freedom. Calculate a fractional degrees of freedom as given below, or use the smaller of $n_1 - 1$ or $n_2 - 1$. This latter value always results in conservative results. As sample size increases, this latter procedure also becomes more accurate. The two-sample t procedures are more robust than the one-sample methods, especially when the distributions are not symmetric. If the sizes of the two samples are equal and the two distributions have similar shapes, it can be accurate down to sample sizes as small as $n_1 = n_2 = 5$. The two-sample t procedure is most robust against nonnormality when the two samples are of equal size. Thus when planning such a study, you should make them equal.

The fractional degrees of freedom formula is as follows:

$$\boxed{\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2 \div (n_1 - 1) + (s_2^2/n_2)^2 \div (n_2 - 1)}}$$

10.6 Matched Pair Test

Comparative studies are more convincing than single sample investigations. Thus one sample inference testing is less common. A common design compares two treatments, either before and after, or randomly picking one of each pair for treatment. In a such a matched pair design, we apply the one sample t procedures to the observed differences. Our null hypothesis would be that these differences are zero and our alternative hypothesis would be that they are not (two-tailed) or positive/negative (one-tailed).

An example might be before and after SAT scores after a high-priced course of study. Or your typical freshman practice EXPO project where peas, corn, or other seeds are grown with and without (control) a treatment. Some Biology instructors and EXPO judges have expected our freshmen to perform this calculation!

Name ______________________________ Score ________

10.7 Sampling Box Activity—Individual

You will be collecting data soon (today?) using **sampling boxes**. Each box contains many (about 400) small beads. These beads are either clear or colored. The colored beads are present in a proportion which you are attempting to determine. These boxes are not **industrial strength**, so please treat them with care. The non-clear beads are colored as follows: **green**, **red**, **blue**, **lavender**, and **orange**, in approximate increasing order of proportion. The top of the box has 20 holes used to take your sample. Notice that 20 is less than 10% of our population, so we will treat this binomially, even though we are actually sampling without replacement. Each student should plan to collect five values for each box. This value is the number of colored beads in the 20 holes. If all holes are not filled on your first attempt, try again (as in don't use that event as data). Be sure to completely empty the holes between samples. You should be able to collect this data fairly quickly and pass the box on to the next user. Record your data in the table below and calculate your $n = 5$ arithmetic **mean** and sample **standard deviation** for your results. Compare your results with your group. Record your results on the summary sheet.

Sample/color:	green	red	blue	lavender	orange
1					
2					
3					
4					
5					
mean:					
st.d.:					

10.8 Homework for Student t-Distribution

1. Find the value of t from the table which has a probability of 0.05 to the right of t when $n = 6$.

2. Use the table of t values to find a t value with probability of 0.99 to the right of t when $n = 21$.

3. What value(s) of t would you use to find a 95% confidence interval for the mean of a population if $n = 16$?

4. Use `tcdf` on your calculator to find a t value for $n = 8$ and a one-tailed $\alpha = 0.005$. You might start by comparing the results of `tcdf(4.032,9E99,5)` and `tcdf(3.169,9E99,10)` on your calculator with the corresponding entries in the table in the lesson.†

5. Suppose you have a one-sample t statistic from a sample of $n = 6$. Suppose further that you calculated a t value of $t = 2.80$ for your hypothesized population mean (H_0: $\mu = 64$ and H_a: $\mu \neq 64$). Give the two-tailed probabilities which bracket this value. Calculate the **P-value** (twice the area to the right of this t value). Should you reject or fail to reject the null hypothesis?

A university researcher placed 12 randomly selected radon detectors in a chamber that exposed them to 105 picocuries per liter of radon. The detector readings were as follows: 91.9, 97.8, 111.4, 122.3, 105.4, 95.0, 103.8, 99.6, 96.6, 119.3, 104.8, and 101.7.

†Some later TI-84 calculators have an inverse T function.

6. Construct a stem-and-leaf diagram of the above data using stems split two ways (i.e. 90–94, 95–99, ...). (Hint: it might be easier to round to integer first.)

7. Check whether the sample size and skewness allow use of a t test.

8. Calculate a t-value for the sample mean *versus* the population mean (105).

9. Calculate the areas under the curve further away from the mean for this value of t (two-tailed). Is there convincing evidence that the mean reading of all detectors of this type differ from the true value?

10. Calculate a two-sample t statistic for the data obtained from the 2000 penny experiment ($\bar{x} = 15.2$, $s = 2.71$, $n = 18$ for Calkins and $\bar{x} = 12.2$, $s = 1.39$, $n = 9$ for Burdick).

11. Calculate the fractional degrees of freedom for the above penny experiment using the formula given at the end of the lecture on two-sample t tests. ($n_1 = 18$ and $n_2 = 9$). Compare this number with that obtained from the TI-84+ calculator `STAT TESTS 4: 2-SampTTest ... not equal, not pooled, calculate.`

Probability & Dist. Lesson 11

The Central Limit Theorem

> *The weight of evidence for an extraordinary claim must be proportioned to its strangeness.*
> Principle of Laplace

In this Lesson we explore the Central Limit Theorem and its consequences. We apply its application to confidence intervals and associated margins of error. We end with some words about the Finite Population Correction Factor.

11.1 A Father of the Central Limit Theorem: Laplace

Pierre-Simeon Laplace (1749–1827) was a French mathematician and astronomer. His five volume work *Méchanique Céleste* published during the last third of his life established mathematical astronomy and shifted the classical mechanics developed by Newton from a geometric to a calculus basis. He is remembered for his development of mathematical physics as one of the greatest scientists of all time, the *Newton of France.*

Laplace was not the originator of the central limit theorem, but was rather at the right time and right place during its development. Much earlier work by De Moivre and contemporary work by Cauchy, Bessel, and Poisson are also important. Its proof is relatively simple as far as mathematical theorems go, but won't be dealt with here.

Laplace's probabilistic approach was important to solving such questions as the stability of the solar system, something Newton's cumbersome approach had to leave to occasional "divine intervention." Laplace said "I had no need for that hypothesis" when asked by Napoleon why he hadn't mentioned God in his books.

11.2 Randomization for Generalization

Remember, sampling is an important tool for determining the characteristics of a population. We usually don't know the population's parameters (mean, standard deviation, *etc.*), but often want reliable estimates of them. Ensuring random (representative) sampling free of bias and sampling errors is important. Some sources of error can be accounted for in the experimental design (blind, double blind, Latin square, *etc.*

An important rule to remember is:

No randomization, no generalization.

What this means is, your results can not be generalized if proper randomization techniques did not occur in your sampling. Many masters degree students have visited their statistician AFTER collecting their data and discovered many months or years were wasted due to poor experimental design.

11.3 The Central Limit Theorem

The **Central Limit Theorem** is often called the second fundamental theorem of probabilty—-the Law of Large Numbers is the first. It is thus a very important and useful concept in statistics. There are essentially three things we want to learn about any distribution: 1) The location of its center; 2) its width, 3) and how it is distributed. The central limit theorem helps us approximate all three.

Central Limit Theorem: As sample size (n) increases, the sampling distribution of sample means approaches that of a normal distribution with a mean ($\mu_{\bar{x}}$) the same as that of the population (μ) and a standard deviation ($\sigma_{\bar{x}}$) equal to the standard deviation of the population (σ) divided by $\sqrt{n}$ (the square root of the sample size) or $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Stated another way, if you draw **simple random samples** (SRS) of size n from any population whatsoever with mean μ and finite standard deviation σ, when n is large, the sampling distribution of the sample means $\bar{x}$ is close to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. This normal distribution is often denoted by: $N(\mu, \sigma/\sqrt{n})$.

The converse of the Central Limit Theorem, that as sample sizes decrease the distribution is more likely to become abnormal, is also true.

11.4 Confidence Intervals/Margin of Error

The value $\sigma_{\bar{x}}$ or $\sigma/\sqrt{n}$ is often termed the **standard error of the mean.** It is used extensively to calculate the margin of error which in turn is used to calculate confidence intervals.

Remember, if we sample enough times, we will obtain a very reasonable estimate of both the population mean and population standard deviation. This is true whether or not the population is normally distributed. However, normally distributed populations are very common. Populations which are not normal are often "heap-shaped" or "mound-shaped." Some skewness might be involved (mean left or right of median due to a "tail") or those dreaded outliers may be present. It is good practice to check these concerns before trying to infer anything about your population from your sample.

Since 95.0% of a normally distributed population is within 1.96 (95% is within about 2) standard deviations of the mean, we can often calculate an interval around the statistic of interest which 95% of the time would contain the population parameter of interest. We will assume for sake of discussion that this parameter is the mean.

The **margin of error** is the standard error of the mean, $(\sigma/\sqrt{n})$, multiplied by the appropriate z-score (1.96 for 95%).

A **95% confidence interval** is formed as: **estimate ± margin of error.**

We can say we are 95% confident that the unknown population parameter lies within our given range. This is to say, the method we use will generate an interval containing the parameter of interest 95% of the time. For life-and-death situations, 99% or higher confidence intervals may quite appropriately be chosen.

Example: Assume the population is the U.S. population with a mean IQ of 100 and standard deviation of 15. Assume further that we draw a sample of $n = 5$ with the following values: 100, 100, 100, 100, 150. The sample mean is then 110 and the sample standard deviation is easily calculated as $\sqrt{(10^2 + 10^2 + 10^2 + 10^2 + 40^2)/(5-1)} = \sqrt{500}$ or approximately 22.4. The standard error of the mean is $\sqrt{500}/\sqrt{5} = \sqrt{100} = 10$. Our 95% confidence interval is then formed with $z = \pm 1.96$. Thus based on this sample we can be 95% confident that the population mean lies between $110 - 19.6$ and $110 + 19.6$ or in $(90.4, 129.6)$. Suppose, however, that you did not know the population standard deviation. Then since this is also a small sample you would use the t-statistics. The t-value of 2.776 for 4 degrees of freedom and a 95% (two-tailed) confidence interval would give you a margin of error of 27.8 and a corresponding confidence interval of $(82.2, 137.8)$.

11.5 Finite Population Correction Factor

The finite population correction factor is: $\sqrt{\dfrac{N-n}{N-1}}$.

If you are sampling without replacement and your sample size (n) is more than, say, 5% of the finite population (N), you need to adjust (reduce) the standard error by multiplying it by the finite population correction factor as specified above. This reduced standard error ultimately increases your margin of error since it is used in the denominator. If we can assume that the population is infinite or that our sample size does not exceed 5% of the population size (or we are sampling with replacement), then there is no need to apply this correction factor.

Example: Suppose 40 men will be getting on a ferry and an old National Health Survey indicates that the population has a mean weight of 173 pounds with a standard deviation of 30 pounds. What is the probability that a sample of four of these men will average over 180 pounds?

Solution: $N = 40$ so our population is not infinite. Although this distribution is probably fairly normal, our sample size is not larger than 30 ($n = 4 < 30$), so a normal approximation would be inappropriate. $n/N = 4/40 = 10\%$ so we should use the finite population correction factor: $\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{40-4}{40-1}} = \sqrt{36/39} = .961$. We expect $\mu_{\bar{x}} = 173$ and $\sigma_{\bar{x}} = 30/2 = 15$. Applying the correction factor this becomes 14.4. $P(x > 179.5) = P(z > \frac{179.5-173}{14.4}) = P(z > 0.45) = 0.326$.

11.6 3-Level Study Guide for Central Limit Theorem

Directions: Check the statements which you believe say what the author says. Sometimes, the exact words are used; at other times, other words may be used.

1. Ensuring random sampling, free of sampling errors, is important.

2. The Central Limit Theorem helps us approximate the center location and width of any distribution.

3. The empirical rule states that about 95% of a normally distributed population is within 2 standard deviations of the mean.

4. The margin of error is the product of the standard error of the mean and a carefully chosen z-score.

5. Although a 95% confidence level is commonplace, life or death situations may require a higher confidence level.

6. If your sample size is more than 5% of the population you should adjust the standard error upward by a certain factor.

Directions: Check the statements which you feel represent the author's intended meaning.

1. Time spent designing an experiment is time well spent.

2. With large sample sizes, the mean of your sample means is less likely to be close to the true (population) mean.

3. For samples with sizes of more than 30, the distribution of sample means can be well approximated by a normal distribution.

4. If the empirical rule says 99.7% of a population is within three standard deviations of the mean, then a z-score of about 3 would produce a margin of error of about 99.7%.

5. Both socialogists and medical researchers use a 95% confidence interval.

6. If we replace the object before resampling, we can assume our population to be infinite.

Directions: Check the statements you agree with, and be ready to support your choices with ideas from the text as well as your own knowledge and beliefs.

1. You should have your experimental design checked by a statistician before collecting data.

2. It is important for researchers to be able to generalize their results.

3. Selecting a sample size is not important to the experimental design process.

4. Speaking in terms of a margin of error is just another way of saying "We don't know for sure."

5. A 95% confidence interval is quite acceptable for something as benign as the outcome of the presidential election in Ohio in Nov. 2004.

6. The Hypergeometric distribution and the Finite Population Correction Factor are related at a deep mathematical level.

11.7 Homework for Central Limit Theorem

1. Given a 2003 penny data sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate the margin of error (assume a 95% confidence interval will be generated).

2. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate a 95% confidence interval.

3. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate the margin of error (assume a 99% confidence interval will be generated).

4. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate a 99% confidence interval.

5. A **P-value** is a way to express the confidence of our results. For a one-tailed test, it is the area under the curve to the right (or left) of our observed mean. Calculate a t-score using our observed mean (15.8), expected mean (10.0), and standard error ($1.91/\sqrt{16}$) and sketch this region on a normal curve.

6. Calculate this area by doing a `tcdf(`t`,9E99,15)`, where t is the value calculated above, and there are 15 degrees of freedom.

7. **Alpha** (α) is the term used to express the level of significance we will accept. For 95% confidence, $\alpha = 0.05$. If our P-value is less than alpha, we can reject our null hypothesis (H_0: $\mu = 10$). Should we reject our null?

8. Try to identify sources of error or bias which might account for these (highly significant) results.

9. Do you think other coins might display similar characteristics? How many times would you have to test it to reach a significant conclusion.

10. Do you think spinning coins (especially some of the new and different state quarters) might display similar characteristics? We may hand out a data gathering sheet with very specific collection instructions.

11. How willing are you to bet money using this method of “flipping” a coin (assuming you have no scruples against such an activity)?

Probability & Dist. Lesson 12

Correlations and Regressions

> *Smoking kills. If you're killed, you've lost a very important part of your life.*
> Brooke Shields

No where has the maxim "correlation does not imply causation" been more hyped than by the tobacco industry. In this lesson we give an overview of mathematical correlation and regression. Although we concentrate on linear regression and the associated process of least squares fit, the homework explores higher order regressions on the TI-8x's series calculators. First a biography of the man who quantified how good a regression fits.

12.1 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 the man. 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-square. 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 it name given so completely. Pearson also worked on classifying distributions. Pearson was offered but refused a knighthood, among other honors.

12.2 Correlation

The common usage of the word **correlation** refers to a relationship between two or more objects (ideas, variables...). In statistics, the word correlation refers to the relationship between two variables.

Examples: one variable might be the number of hunters in a region and the other variable could be the deer population. Perhaps as the number of hunters increases, the deer population decreases. This is an example of a **negative correlation**: as one variable increases, the other decreases. A **positive correlation** is where the two variables react in the same way, increasing or decreasing together. Temperature in Celsius and Fahrenheit have a positive correlation.

How can you tell if there is a correlation? By observing the graphs, a person can tell if there is a correlation by how closely the data resemble a line. If the points are scattered about then there is may be **no correlation**. If the points would closely fit a quadratic or exponential equation, *etc.* then they have a **nonlinear correlation**. In this lesson we will restrict ourselves to linear correlation.

How can you tell by inspection the type of correlation? If the graph of the variables represent a line with positive slope, then there is a positive correlation (x increases as y increases). If the slope of the line is negative, then there is a negative correlation (as x increases y decreases).

An important aspects of correlation is how *strong* it is. The strength of a correlation is measured by the **correlation coefficient** r. Another name for r is the **Pearson product moment correlation coefficient** in honor of Karl Pearson who developed it about 1900.

$$\boxed{r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}}$$

For samples, the correlation coefficient is represented by r while the correlation coefficient for populations is denoted by the Greek letter rho (ρ which looks almost like a p).

The closer r is to $+1$, the stronger the positive correlation is. The closer r is to -1, the stronger the negative correlation is. If $|r| = 1$ exactly, the two variables are **perfectly correlated**! Temperature in Celsius and Fahrenheit are perfectly correlated.

Formal hypothesis testing can be applied to r to determine how significant a result is. The Student t distribution with $n-2$ degrees of freedom, $t = (r-0)/s_r$ (where 0 represents the expected correlation or rho), and $s_r^2 = (1-r^2)/(n-2)$. For $n = 8$, $\alpha = 0.05$ and a two-tailed test, critical values of ± 0.707 are obtained.

Remember, correlation does not imply causation.

A value of zero for r does not mean that there is no correlation, there could be a nonlinear correlation. **Confounding variables** might also be involved. Suppose you discover that miners have a higher than average rate of lung cancer. You might be tempted to immediate conclude that their occupation is the cause, whereas perhaps the region has an abundance of radioactive radon gas leaking from the subterranian regions and all people in that area are affected. Or, perhaps, they are heavy smokers....

r^2 is frequently used and is called the **coefficient of determination**. It is the fraction of the variation in the values of y that is explained by least-squares regression of y on x. This will be discussed further below after least squares is introduced.

Correlation coefficients whose magnitude are between 0.9 and 1.0 indicate variables which can be considered **very highly correlated.** Correlation coefficients whose magnitude are between 0.7 and 0.9 indicate variables which can be considered **highly correlated.** Correlation coefficients whose magnitude are between 0.5 and 0.7 indicate variables which can be considered **moderately correlated.** Correlation coefficients whose magnitude are between 0.3 and 0.5 indicate variables which have a **low correlation.** Correlation coefficients whose magnitude are less than 0.3 have little if any (linear) correlation. We can readily see that $0.9 < |r| < 1.0$ corresponds with $0.81 < r^2 < 1.00$; $0.7 < |r| < 0.9$ corresponds with $0.49 < r^2 < 0.81$; $0.5 < |r| < 0.7$ corresponds with $0.25 < r^2 < 0.49$; $0.3 < |r| < 0.5$ corresponds with $0.09 < r^2 < 0.25$; and $0.0 < |r| < 0.3$ corresponds with $0.0 < r^2 < 0.09$.

12.3 Regression

Regression goes one step beyond correlation in identifying the relationship between two variables. It creates an equation so that values can be predicted within the range framed by the data. Since the discussion is on linear correlations and the predicted values need to be as close as possible to the data, the equation is called the **best-fitting line** or **regression line**. The regression line was named after the work Galton[*] did in gene characteristics that reverted (regressed) back to a mean value.

If you go outside the original domain (x values) you are **extrapolating**, otherwise you are **interpolating**.

An equation of a line is expressed as $y = mx + b$ or $y = ax + b$ or even $y = a + bx$. As we see, the regression line has a similar equation.

$y = \beta_0 + \beta_1 x$ where y, β_0, and β_1 represents population statistics. But if a cap appears above the variable, then they represent sample statistics. Remember x is our independent variable for both the line and the data.

The y-intercept of the regression line is β_0 and the slope is β_1. The following

[*] `http://www.robertsfox.com/regression_to_mean.htm`

formulas give the y-intercept and the slope of the equation.

$$\boxed{\beta_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n \sum x^2 - (\sum x)^2} \qquad \beta_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}}$$

Notice that the denominators are the same, so that saves calculations. Also, the calculator will have values for certain portions. Also note the important difference between the sum of the squares of the x's $(\sum x_i^2)$ and the square of the sum of the x's $((\sum x_i)^2)$. Another way to write the equation is in point-slope form where the **centroid** is the point that is always on the line. The centroid is the ordered pair: (mean of x, mean of y) or $(\bar{x}, \bar{y})$.

To keep the y-intercept and slope accurate, all intermediate steps should be kept to twice as many significant digits (six to ten?) as you want in your final answer (three to five?)!

There are certain guidelines for regression lines:

1. Use regression lines when there is a significant correlation to predict values.

2. Do not use if there is not a significant correlation.

3. Stay within the range of the data. Do not extrapolate!! For example, if the data is from 10 to 60, do not predict a value for 400.

4. Do not make predictions for a population based on another population's regression line.

Example: Write the regression line for the following points:

x	y
1	4
3	2
4	1
5	0
8	0

Solution 1: $\sum x = 21$; $\sum y = 7$; $\sum x^2 = 115$; $\sum y^2 = 21$; $\sum xy = 14$. Thus $\beta_0 = [7 \cdot 115 - 21 \cdot 14] \div [5 \cdot 115 - 21^2] = 511 \div 134 = 3.81$ and $\beta_1 = [5 \cdot 14 - 21 \cdot 7] \div [5 \cdot 115 - 21^2] = -77 \div 134 = -0.575$. Thus the regression line for this example is $y = -0.575x + 3.81$.

Solution 2: On your TI-84+ graphing calculator, enter the data into L_1 and L_2 and do a `LinReg(ax+b)` L_1, L_2 (`STAT`, `CALC`, 4) or `LinReg(a+bx)` L_1, L_2 (`STAT`, `CALC`, 8). You should get a screen with
$y = ax + b$

$a = -.5746...$
$b = 3.8134...$
$r^2 = .790...$
$r = .88888...$

If the r^2 and r information is absent, do `CATALOG` (`2nd 0`) `DiagnosticOn`. `ENTER` will bring the command back to the home screen where another `ENTER` will execute it. We thus see that about 79% of the variation in y is explained by least-squares regression of y on x.

There is no mathematical difference between the two linear regression forms `LinReg(ax+b)` and `LinReg(a+bx)`, only different professional groups prefer different notations.

Note the presence on your TI-84+ graphing calculator of several other regression functions as well. Specifically, quadratic ($y = ax^2 + bx + c$), cubic ($y = ax^3 + bx^2 + cx + d$), quartic ($y = ax^4 + bx^3 + cx^2 + dx + e$), exponential ($y = ab^x$), and power or variation ($y = ax^b$). Thus an easy way to find a quadratic through three points would be to enter the data in a pair of lists then do a quadratic regression on the lists. See the homework for a specific example.

12.4 Least Squares Procedure

The method of least squares was first published in 1806 by Legendre. However, Gauss "communicated the whole matter to Olbers in 1802."

What is the Least Squares Property?

Form the distance $y - y'$ between each data point (x, y) and a potential regression line $y' = mx + b$. Each of these differences is known as a **residual**. Square these residuals and sum them. The resulting sum is called the **residual sum of squares** or SS_{res}. The line that best fits the data has the least possible value of SS_{res}.

This link[†] has a nice colorful example of these residuals, residual squares, and residual sum of squares.

Example: Find the Linear Regression line through $(3, 1), (5, 6), (7, 8)$ by brute force.

Solution:

x	y	y'	$y - y'$
3	1	$3m + b$	$1 - 3m - b$
5	6	$5m + b$	$6 - 5m - b$
7	8	$7m + b$	$8 - 7m - b$

Using the fact that $(A + B + C)^2 = A^2 + B^2 + C^2 + 2AB + 2AC + 2BC$, we can quickly find $SS_{\text{res}} = 101 + 83m^2 + 3b^2 - 178m - 30b + 30mb$. This expression is

[†] `http://www.keypress.com/sketchpad/java_gsp/squares.html`

quadratic in both m and b. We can rewrite it both ways and then find the vertex for each (which is the minimum since we are summing squares). Remember the vertex of $y = ax^2 + bx + c$ is $\frac{-b}{2a}$. $SS_{\text{res}} = 3b^2 + (30m - 30)b + (101 + 83m^2 - 178m)$. $SS_{\text{res}} = 83m^2 + (30b - 178)m + (101 + 3b^2 - 30b)$. From the first expression we find $b = (-30m+30)/6$. From the second expression we find $m = (-30b+178)/166$. These expressions give us two equations in two unknowns: $5m + b = 5$ and $83m + 15b = 89$. These can be solved to obtain $m = 7/4 = 1.75$ and $b = -15/4 = -3.75$. This is how the equations above for β_0 and β_1 were derived, from the general solution to two general equations for SS_{res}.

This link[‡] brings up a Java applet which allows you to add a point to a graph and see what influence it has on a regression line.

This link[§] brings up a Java applet which encourages you to guess the regression line and correlation coefficient for a data set.

[‡] http://www.stat.sc.edu/~west/javahtml/Regression.html

[§] http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/

12.5 Homework for Correlation and Regression

1. Suppose you remember the triangular numbers, but can't remember their formula. Enter the following values into L_1: 1, 2, 3 and L_2: 1, 3, 6. Now do a `QuadReg` L_1, L_2 using your TI-84+ calculator (`STAT`, `CALC`, `5`) and interpret the results (rewrite the formula in its usual form).

2. Suppose further you really couldn't remember if the relationship was quadratic. Try a `CubicReg` L_1, L_2 and interpret the results.

3. Add an additional point onto the end of L_1: 4 and L_2: 10. Repreform the `CubicReg` L_1, L_2 and interpret the results.

4. Try the `CubicReg` L_1, L_2 with L_1: 1, 2, 3, 4 and L_2: 1, 5, 14, 30 (sums of squares) and interpret the results. Try to express your answer in an aesthetically pleasing form (*i.e.* fractions not decimal fractions).

5. Question one can be done by solving three equations in three unknowns. Specifically, let $ax^2 + bx + c = y$. Then substitute each value of x and equate it to the corresponding y value. Solve these three equations manually by elimination (due to the regular spacing, first c, then b eliminate easily).

6. Solve the above equations using your calculator, either using augmented or inverse matrices. Record the pertinent matrices and keystrokes here.

Probability & Dist. Lesson 13

Hypothesis Testing and χ^2

> *It is easy to ... throw out an interesting baby with the nonsignificant bath water. Lack of statistical significance at a conventional level does not mean that no real effect is present; it means only that no real effect is clearly seen from the data. That is why it is of the highest importance to look at power and to compute confidence intervals.* William Kruskal

In this lesson we first review hypothesis testing by giving the four steps presented in the Introduction to Statistics, Lesson 9. We then define Power. We continue with examples using the Chi-Square distribution doing various tests for goodness of fit.

13.1 Father of Hypothesis Testing: Jerzy Neyman

Jerzy Neyman (1894–1981) was a Polish-American mathematician and statistician. His family background was Polish nobles and military heroes, but he was born in Russia. He went to Poland in 1921 and earned his PhD in 1924 under Sierpiński, among others. He studied via fellowships with Karl Pearson and other notables in London and Paris. His 1934 paper given at the Royal Statisitical Society was the event leading to modern scientific sampling. Another paper in 1937 introduced the confidence interval. In 1938 he moved to Berkeley where he advised 39 PhD students. The Neyman-Pearson Lemma is named after him and Karl Pearson's only son Egon who was born the year after Neyman but died the year before.

13.2 Hypothesis Testing

Once descriptive statistics,[*] combinatorics, and distributions are well understood, we can move on to the vast area of **inferential statistics.** The basic concept is one called **hypothesis testing** or sometimes the test of a statistical hypothesis. Here we

[*] `http://www.andrews.edu/~calkins/math/webtexts/statall.pdf`

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. However, these days it is common practice not to associate any special meaning to which hypothesis is which. (Having said that, however, I must quickly note that some researchers strongly adhere to this tradition. Check early with your research partners, just in case.)

Although simple hypotheses would be easiest to test, it is much more common to have one of each type or 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, directional), whereas if the H_a values lie on both sides of H_0, then we have a **two-sided test** (two tailed, nondirectional).

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.

13.3 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 a table will make it clearer.

Reject/Truth	H_0 **True**	H_a **True**
Reject H_a	no error	False positive, Type II, $\beta = P(\text{Reject } H_a \mid H_a \text{ true})$
Reject H_0	False negative, Type I, $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true})$	no error

The Greek letters α (alpha) and β (beta) should already be familiar. 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 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, Lyme disease, or other medical conditions, both types of errors can be a very serious matter. Formally, $\alpha = P(\text{Accept} H_a | H_0$ true), meaning the probability that we "accepted" H_a when in fact H_0 was true. This meaning for alpha is very similar to that encountered earlier and is often called the **level of significance**. Alpha and beta usually cannot both be minimized—there is a trade-off between the two. Historically, a **fixed level** of significance was selected ($\alpha = 0.05$ for the social sciences and $\alpha = 0.01$ or $\alpha = 0.001$ for medicine or 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.**

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

Now both are usually considered together when determining an adequately sized sample. Instead of testing against a fixed level of alpha, now a P-value is often reported.

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.

Obviously, the smaller the P-value, the stronger the evidence (higher significance, smaller alpha) provided by the data is against H_0.

13.4 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 in Lesson 10. 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.

13.5 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.

13.6 Power of a Test

The **power** of a test against the associated correct value is $1 - \beta$. It is the probability that a Type II error is not committed. **There is a different value of beta for each possible correct value of the population parameter.** It also depends on sample size (n), thus increasing the sample size increases the power. Power is thus important in planning and interpretting tests of significance.

It is easy to misspeak power $(1 - \beta)$ and P-value (α).

13.7 Chi Square Distributions and Tests

The χ^2 (chi-square) 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). The χ^2 distribution is important in several contexts, most commonly involving variance. Please note that χ is pronounces like a hard k sound like the Scottish Loch (lake) as in Loch Ness Monster. (I may have to disown you as a student if I hear anything which sounds soft and cuddly like a chia pet.) The χ^2 distribution is important in several contexts, most commonly involving variance.

The χ^2 distribution 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.

1. The χ^2 distribution is continuous.

2. The χ^2 distribution is unimodal.

3. The χ^2 distribution is always positive $\chi^2 > 0$.

4. The χ^2 distribution mean $= \nu$.

5. The χ^2 distribution variance $= 2\nu$.

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.

8. We can thus approximate χ^2_ν when $\nu > 30$ with the normal (see table below).

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

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

A common application of the chi-square statistics is in a **test for goodness of fit** as described in the homework. Here we compare expected with observed frequencies typically for one nominal variable. 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, sometimes a low χ^2 value is significant. An example for both is included below.

The χ^2 is also use for **tests of indepedence.** Chi-square **contingency tables** are often formed and a **contingency coefficient** may also be used, especially when working with nonparametric measurements.

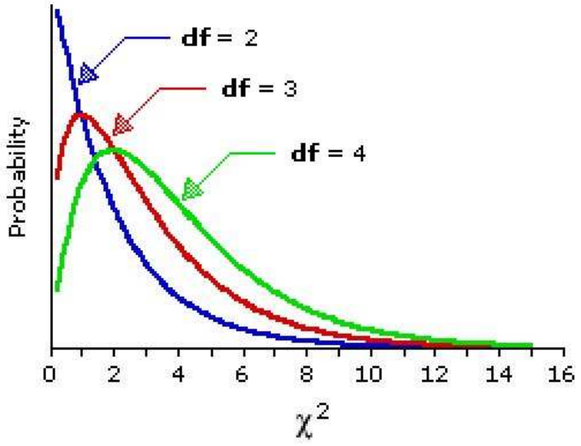
Example: On July 14, 2005 we collected 10 trials of 20 pennies each where these 20 pennies were set on edge and the table banged. We observed 145 heads. We can compare the observed with expected frequencies and test for goodness of fit as shown in the table below. There is but one degree of freedom since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

Solution: We form the χ^2 statistic by summing the $\frac{(O-E)^2}{E}$ and get $2045/100+2045/100 = 40.9$. 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$= 1.6 \times 10^{-10}$). A table of critical χ^2 values for select values is given below.

Side:	**Head**	**Tail**
Observed	145	55
Expected	100	100
(Obs-Exp)	45	-45
$(O-E)^2$	2045	2045
$(O-E)^2/E$	20.45	20.45

13.7.1 A Chi Square Distribution Table

df/upper tail area	**0.99**	**0.95**	**0.90**	**0.10**	**0.05**	**0.01**
1	0.00016	0.0039	0.016	2.706	3.841	6.635
2	0.020	0.103	0.211	4.605	5.991	9.210
3	0.115	0.352	0.584	6.251	7.815	11.34
4	0.297	0.711	1.064	7.779	9.488	13.28
5	0.554	1.145	1.610	9.236	11.07	15.09
10	2.558	3.940	4.865	15.99	18.31	23.21
15	5.229	7.261	8.547	22.31	25.00	30.58
20	8.260	10.85	12.44	28.41	31.41	37.57
25	11.52	14.61	16.47	34.38	37.65	44.31
> 30	use $z = \sqrt{2\chi^2} - \sqrt{2\text{df} - 1}$					



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

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

Solution: We form the χ^2 statistic by summing the $\frac{(O-E)^2}{E}$ and get 208/32=6.5. We can then compare this χ^2 with a critical χ^2. Only if it is 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. Such data are often said to be "cooked" as in cooked up from scratch.

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 signficant. 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.)

Name ________________________________ Score ________

13.8 Homework for Hypothesis Testing

1. Set Y_1 on your TI-84+ graphing calculator equal to χ^2`pdf(X,5)` (`2nd DISTR 6`).[†] Then adjust your viewing window to $0 < x < 10$ with `Xscl=1` and $0 < y < 0.3$ with `Yscl=0.02` and sketch the results.

2. Repeat problem 1 with Y_1 on your TI-84+ graphing calculator set to χ^2`pdf(X,2)` (`2nd DISTR 6`).[‡]

3. Repeat problem 1 with Y_1 on your TI-84+ graphing calculator set to χ^2`pdf(X,20)` (`2nd DISTR 6`),[§] but extend the domain (x) to go up to 40.

4. Since 1995, blue M&M® candies replaced tan with 30% brown, 20% yellow, 20% red, 10% orange, 10% green, and 10% 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 need to purchase one 1.69 oz bag of plain M&M®'s.** It is best if the bags are purchased at different stores and not obtained from only a few sources of supply. Complete the table below use $n = 10$ different, randomly selected M&M's for each person.

[†] The **6** may be a **7** if your calculator has `InvT` as **4**.

[‡] The **6** may be a **7** if your calculator has `InvT` as **4**.

[§] The **6** may be a **7** if your calculator has `InvT` as **4**.

Color:	brown	yellow	red	orange	green	blue
Observed						
Expected	$\frac{3n}{10}$ = ___	$\frac{2n}{10}$ = ___	$\frac{2n}{10}$ = ___	$\frac{1n}{10}$ = ___	$\frac{1n}{10}$ = ___	$\frac{1n}{10}$ = ___
$\frac{(O-E)^2}{E}$						

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: collect the information from everyone at your table and use this larger sample to complete the table below.

Color:	brown	yellow	red	orange	green	blue
Observed						
Expected	$\frac{3n}{10}$ = ___	$\frac{2n}{10}$ = ___	$\frac{2n}{10}$ = ___	$\frac{1n}{10}$ = ___	$\frac{1n}{10}$ = ___	$\frac{1n}{10}$ = ___
$\frac{(O-E)^2}{E}$						

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

6. **Bonus:** Repeat the process above but using all freshmen and sophomore data.

7. **Bonus:** After completing the count feel free to dispose of the M&M® by any method you deem appropriate.

8. Set Y_1 on your TI-84+ graphing calculator to `Fpdf(X,5,5)` (`2nd DISTR 8`).[**] Then adjust your viewing window to $0 < x < 5$ with `Xscl=1` and $0 < y < 0.7$ with `Yscl=0.1` and sketch the results.

[¶] By putting the O and E values in lists, manipulating the lists, and storing the results in another list, you can also get the calculator to sum these values easily.

[‖] *The Practice of Statistics*, 3rd edition, by Yates, Moore, Starnes, 2006, page 834 gives the following percentages: brown 13%, yellow 14%, red 13%, orange 20%, green 16%, and blue 24%. It is not known when this changed, but when all my data over the last several years were checked, these percentages yielded much better χ^2 results.

[**] The **8** may be a **9** if your calculator has `InvT` as **4**.

Probability & Dist. Lesson 14

Design of Exper., Non-parametrics

All models are wrong, some models are useful. George Box

14.1 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.

14.2 Experimental Design

Experimental design or design of expermients is a discipline important to all natural and social sciences whereby an experimenter controls how variation of a treatment (process or intervention) affects the information gathered about that treatment. The formal mathematical methodology was developed by Fisher in his 1935 book *The Design of Experiments*. A classic example, perhaps frivolous, yet nonetheless instructive was wehter a certain lady could tell by flavor alone whether tea or milk were first placed in the cup. Design of experiment was develped with analysis of variance or ANOVA which we also discuss below.

One of the first well documented experiments was done in 1747 by ship's surgeon James Lind in his quest to develop a cure for scurvy. He selected 12 sufferers, divided them into six pairs, and gave each of the six pairs a different diet variation for two weeks. All six treatments were commonly proposed remedies. The cider, sulfuric acid, seawater, garlic/mustard/horseradish, and vinigar groups had minimal improvement, whereas the group receiving two oranges and one lemon every day recovered quickly and either returned to duty or nursed the rest. Although he used replication, he did not use a control, nor randomized allocation of subjects to treatments, instead ensuring the cases *"were as similar as I could have them."* We will list and discuss here several important aspects of exerimental design.

14.2.1 Comparison

Comparison between treatments are usually more reproducible and usually preferable. This is because it is hard to exactly reproduce measured results. A standard or traditional treatment is often compared with as a baseline.

14.2.2 Randomization

Random here does not mean haphazard—some randomizing mechanism, such as random numbers, is being employed to allocate units to treatments. There are extensive mathematical theories used to calculate and manage the risks associated with getting a serious imbalance between groups. These risks depend on sample size which then must be adequate.

14.2.3 Replication

Having more than one test subject in each test unit gives us some information regarding variation. Measurements usually are subject to variation. This variation can be both between repeated measurements and between the replicated items.

14.2.4 Blocking

How the experimental units are arranged into groups is known as blocking. Some trees in an orchard are going to be on the west, others on the east, perhaps some by a road and others by a woods. Some of these variations may be irrelevant but others relavant. Blocking helps reduce these variations between units and gives a greater precision to our estimation of the source of variation.

14.2.5 Orthogonality

Orthogonal often means perpendicular but here means uncorrelated. In linear algebra, orthogonal vectors are linearly independent. Orthogonal treatments are thus independent and provides different information. Contrast is another word often used to describe these forms of comparison.

14.2.6 Factorial Experiment

When an experiment has two or more factors, it is termed a factorial experiment. These factors often have discrete levels associated with them. Often a full factorial design is infeasible so many (more than half?) of the possible combinations are omitted.

14.3 The F-Distribution and ANOVA

In prior sections we considered tests of inference about the means of various distributions. One can use the t procedure for inferences about the population means for normal populations and often for non-normal populations as well. Similarly, proportions can easily be tested. One might then be tempted to consider tests of inferences about the standard deviation of a population, but the expert advice is: **don't do it without expert advice!** The F Statistic is not robust against non-normality. Also, the F distribution and ANOVA are historically not tested on the AP Statistics Exam.

When comparing standard deviations the test is called **analysis of variance** or more commonly by its acronym **ANOVA.** The ANOVA F allows us to compare sevaral means, not just two as was done earlier with the t statistic.

Since we have use the term F several times it now behooves us to look at the underlying F distribution. The F distribution is named in honor of R. A. **Fisher** who first studied it in 1924. (As you can see by this date and Gauss's work, Statistics really only recently developed.) Specifically, the F distribution compares the variance of two normal populations. If $\sigma_1^2 = \sigma_2^2$, then we expect $s_1^2 - s_2^2$ to be distributed about zero or equivalently the ratio s_1^2/s_2^2 to be close to 1.0. However, this will depend on both sample sizes, or more precisely, on the degrees of freedom.

The ratio of the variances of two independent random samples taken from normal parent populations with equal variances has an F-distribution characterized by the degrees of freedom: $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$

1. The F distribution is always positive or zero and positively skewed (right).
2. The F distribution is characterized by two parameters, the degrees of freedom of the two samples.
3. The F distribution is the ratio of two χ^2 variables.
4. The mean and variance for the F distribution depends on the two degrees of freedom.
5. Extensive tables exist, but only for $F > 1.0$, so use the larger variance as numerator.

6. The t, χ^2, and F are all related to the **gamma distribution.**

14.4 non-Parametric Statistics

We will introduce briefly non-paremetric tests of significance for ordinal dependent variables. We already introduced the χ^2 statistic which is a non-parametric test. Specifically, the dependent variable tends to be at the nominal level and our parametric assumptions of normality and homogeneity of variance cannot be met.

14.4.1 One sample tests for Ordinal Data

Three common **one sample tests** of signficance are: 1) the one-sample sign test; 2) the Mann-Kendall test for trends; and 3) the Kolmogorov-Smirnov one-sample test. The interested reader can search online or textbooks for more information.

14.4.2 Two sample tests for Ordinal Data

We explore here two common tests of significance for the **two sample** case when the dependent variable is measured at the ordinal level. First, the **median test** tests the hypothesis that two samples have been selected from populations with equal medians. One starts by finding this **common median** by ordering the scores from both groups together and then determining the median. Then a 2 × 2 contingency table is formed with the frequency counts of how many from each group are above or below the common median. A χ^2 statistic is generated using the simplified formula:

$$\chi^2 = \frac{n(AD - BC)^2}{(A+B)(C+D)(A+C)(B+D)}$$

and compared with a critical value in the normal way.

Second, the statistically more powerful **Mann-Whitney U test,** tests not only the median, but also the total distribution (central tendancy and distribution). The null hypothesis specifies that there is no difference in the scores of the two populations sampled. Two U values are calculated and the smaller one is selected for checking in a table of critical values. The null hypothesis is reject if the computed U **is less than** the table value. The U values take into account the number of data elements in each sample (n_1 and n_2) and the sum of the ranks in each group (R_1 and R_2). Calculate $U_i = n_1 n_2 + \frac{n_i(n_i+1)}{2} - R_i$ for both $i = 1$ and $i = 2$.

When both groups are larger than 20, the sampling distribution approaches the normal distribution and a z-score can be computed from the sampling distribution mean $\mu_U = \frac{n_1 n_2}{2}$, and standard deviation $\sigma_U = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$. The z score is calculated in the usual way with $z = \frac{U - \mu_U}{\sigma_U}$.

A table of Mann-Whitney critcal U values is given below for $\alpha = 0.05$ (1-tailed=directional H_a) and selected sample sizes only. More extensive tables are readily available.[*]

n_i/n_j	2	3	4	5	6	7	8	9	10	15	20
2	0	0	0	1	1	1	2	2	2	4	5
3	0	1	1	2	3	3	4	5	5	8	12
4	0	1	2	3	4	5	6	7	8	13	19
5	1	2	3	5	6	7	9	10	12	19	26
6	1	3	4	6	8	9	11	13	15	24	33
7	1	3	5	7	9	12	14	16	18	29	40
8	2	4	6	9	11	14	16	19	21	34	48
9	2	5	7	10	13	16	19	22	25	40	55
10	2	5	8	12	15	18	21	25	28	45	63
15	4	8	13	19	24	29	34	40	45	73	101
20	5	12	19	26	33	40	48	55	63	101	139

14.4.3 Kendall's tau and Spearman's rho

Two of my favorite non-parametric statistics are Kendall's tau and Spearman's rho. Primarily because I had to write computer programs my sophomore year in college to calculate them. Kendall's τ measures the degree of correspondence between two rankings and assesses the significance of this correspondance. That is to say it measures the strength of association of the cross tabulations. Spearman's ρ is much like the Pearson product moment correlation coefficient, except the numbers are converted to ranks before it is computed.

14.4.4 K-Sample Tests for Ordinal Data

The test statistics H for the **Kruskal-Wallis one-way analysis of variance** is calculated in a similar manner to the Mann-Whitney U. The null hypothesis is that there is no difference in the distribution of data in the K populations. The alternate hypothesis would be that at least two of the K populations or a combination of populations differ. Although we won't give the formula here, the sampling distribution is the χ^2 with $K - 1$ degrees of freedom.

Tied ranks generally have minimal effect on both the Mann-Whitney U and Kruskal-Wallis H and a correction factor can be applied. However, results from either test might be questionable when there are an excessive number of tied ranks.

[*] http://www.cquest.utoronto.ca/geog/ggr270y/tables/Mann-Whitney_U_Table.htm

14.4.5 Two Sample Tests, Dependent (matched)

The **Wilcoxon matched-pairs signed-rank test** was developed for use with dependent samples and ordinal data. The null hypothesis is again stated in general terms of no difference between populations. The test statistic, termed T, is formed by ranking the pre-/post-test differences but including a sign (negative for larger post-test score). The ranks with the least frequent sign are summed and the resulting statistic compared with a table of critical values. With samples larger than 25 the sampling distribution approaches the normal distribution with $\mu_T = \frac{n(n+1)}{4}$ and standard deviation $\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$. The z score is calculated as $z = \frac{T-\mu_T}{\sigma_T}$.

14.5 Epilogue

This concludes our overview of probability and distributions. Please check your booklets for completeness and prepare them for the completeness activity (quiz) and subsequent stapling.

14.6 Conversion, Image, Copyright, Other Issues

- Generate answers for lessons 1, 2, 3, 4, 5, 6, 8, 9, 13 and homework for 14.
- Code answers for released odd/full solutions (same file).
- Lesson 6: Make skewness graphs (binomial .1, .9).
- Lesson 6: Make cdf of IQs; convert 13qz as binomial quiz.
- Lesson 8: Generate Poisson, Geometric, etc. distribution graphs.
- Lesson 13: The χ^2 image came from Wikipedia? without attribution.
- Lesson 13: Make graph of power.
- Lesson 13: Odd number of pages.
- Convert: More quizzes over Statistics?: 12qy, 34qy, 5aqz, 67qy 08qy.
- Convert: PROD03qz, PROD07ac (ran. straw samp.), PROD09ad (nickle flip).
- Resolve bio duplication of Fisher/Gosset/Pearson.
- There are htm links to numbers in 1, 5, 7, and 8

14.7 Booklet Completion Checklist

Directions: You may work together, but answer each question carefully using **your own** Probability and Distribution booklet. Take time to put the booklet in **THIS** order. Make a list by **table** of who is missing what (nonbonus) items.

1. Page *i* (front cover): Full title of booklet.
2. Page *v*: Title page title for lesson 9.
3. Page 1: quote.
4. Page 3 (0.1 Lesson 1): Singular of Data (item 9).
5. Page 5 (0.3 Lesson 3): Proper way to spell Tendancy (3 places).
6. Page 6 (0.4 Lesson 4): Another name for quadratic mean.
7. Page 8 (0.6 Lesson 6): Mean and standard deviation of IQ scores.
8. Page 10 (0.8 Lesson 8): How many stems should you have?
9. Page 11 (0.9 Lesson 9): Where did Gosset work?
10. Page 14: (1.2 Exp.) Give three common examples of random experiments.
11. Page 18: (1.6 Magic Square) Sum on magic square.
12. Page 20: (1.7 Homework) Q12. Prob. green=5 or red=2.
13. Page 23: (2.2 Counting Rule) Symbolically, what is the **addition rule**?
14. Page 28: (2.9 Homework) Q17. Different permutations in DENNIS.
15. Page 31: (3.4 At Least One) What was P(at least one heart)?
16. Page 35: (3.8 Homework) Q3. Different 5-card poker hands.
17. Page 38: (4.2 Odds) How are odds against and odds in favor related?
18. Page 40: (4.4 Quiz) Q3.
19. Page 42: (4.5 Homework) Q13. Odds against selecting a prime roulette number?
20. Page 44: (5.2 Risk) Exact probability of 3 on 2 die Risk results being split.
21. Page 48: (5.5 Homework) Q4. Sterling's approximation for 300!
22. Page 51: (6.3 D *vs.* C) Two fundamental rules regarding distributions.
23. Page 57: (6.6 Homework) Q3a. Tom Bone's P(getting every note right)?
24. Page 60: (7.2 Binomial) Requirement 4 for binomial experiments.
25. Page 64: (7.7 Magic Square) What matches I?
26. Page 65: (7.8 Homework) Q2. Probability of more than four lefties in 25 if $p = \frac{1}{10}$.

27. Page 68: (7.9 Penny) What was the 2005 average number of penny heads?
28. Page 70: (8.2 Queuing Theorey) How do the Binomial and Poisson differ?
29. Page 73: (8.5 Homework) Q3. P(3 customers in any one minute interval)?
30. Page 76: (9.2 Lorentzian) What behavior does the Lorentzian Dist. describe?
31. Page 82: (9.7 Homework): Q10. FWHM for the two Voigt Profiles.
32. Page 83: (10.1 Gosset) When did "Student" live?
33. Page 88: (10.9 Sampling Box) What was your Orange mean?
34. Page 89: (10.10 Homework) Q4. What value of t for $n = 8$ gives $\alpha = 0.005$?
35. Page 93: (11.4 CI & M of E) What is a margin of error?
36. Page 95: (11.6 3-Level SG) How does replacement affect population size?
37. Page 98: (11.7 Homework) Q7. What P-value and what α were compared?
38. Page 100: (12.2 Correlation) How are Celsius and Fahrenheit correlated?
39. Page 105: (12.5 Homework): Q2. Why didn't this work?
40. Page 108: (13.3 Error Types) Two other names for Type I & Type II errors.
41. Page 115: (13.8 Homework) Q7. Prescribed M&M® disposal method.
42. Page 117: (Chapter 14) What is the George Box quote?
43. Page 126: (14.7 Booklet Checklist) Q43. What is this question number?
44. Page 126: (A.5 Quiz over SL 5) Q8.
45. Page 122: (A.1 SL 1&2 Review Quiz) Q12.
46. Page 124: (A.2 SL 3&4 Review Quiz) Q9.
47. Page 127: (A.4 SL 6&7 Review Quiz) Q4.
48. Page 129: (A.5 SL 8 Review Quiz) Q6.
49. Page 125: (A.3 Stat Released Test) What is the date of the practice test?
50. Page 141: (A.8 Prod. Released Test) What is the date of the released test?
51. Page 146: (B.2 7 Solutions) Q6. What is the probability of losing the lottery?
52. Page 150: (B.3 10 Solutions) Q8. What is the t-value.
53. Page 152: (B.4 11 Solutions) Q6. Probability $\mu = 10$, given the penny sample?
54. Page 153: (B.5 12 C & R) Q4. Factored formula for sum of squares.
55. Page 153: (B.6 Statistics Key) z score in Q5.
56. Page 159: (B.7 Prob. Test Key) Q4.

Appendix A

Quizzes and Tests

A.1 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 ________

A.2 Stats 1&2 Review Quiz

1. What is the difference between **statistics** and **statistic**?

2. What is the difference between **descriptive** and **inferential** statistics?

3. Would you trust a Nov. 2, 2004 midnight voluntary exit poll to be accurate if it said Kerry won over Bush with a very small margin of victory in the Ohio primary? Why or why not?

4. Complete the following comparison: **Parameter** is to ?................?, as **statistic** is to ?..............?.

5. What are the two catagories of data starting with the letter **q**?

6. If numeric data is not **discrete**, then it must be ?.................?.

7. Arrange in order from **lowest to highest** the four levels of measurement.

8. If your teacher's portfolio dropped 50% in value, what percent increase (from the resulting, new value) would be required to return it back to its original value?

   The 1st ed. of a textbook contained 700 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 ed. than the 1st ed.

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

12. ?..............?% of the 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 independently of infection.) **Bonus:** Draw a Venn Diagram with numerical results for each of the four outcomes.

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

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

15. Yesterday's records for MSC attendance show that two underclassmen were absent.

16. Volvo sold 84,000 cars in the United States in 1999.

17. A 1999 Cadillac Escalade weighs 5,600 pounds.

18. The radar clocked a Nolan Ryan fastball at 98.4 mph.

Determine which *level of measurement* **is most appropriate.**

19. Colors of Skittles® brand candies.

20. Final course grades of A, B, C, D, and F.

21. Daily high and low temperatures at the Niles airport for 1998.

22. Time (in days) for a sunspot to be visible from the earth.

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

23. George went through the telephone book and called every 89th person listed.

24. Four people divided the telephone book evenly and each randomly sampling from their portion.

25. All people with a 461 telephone exchange are called.

26. 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.

27. What four letter word is associated with convenience sampling?

28. Differentiate between prospective and retrospective study.

29. Give two other names for random sampling.

Name ________________________________ Score ________

A.3 Stats 3&4 Review Quiz

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 income of select individuals related to producing this homework assignment: $40,000, $400,000, $4,000,000, $40,000,000, and $400,000,000 (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.

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 problem 17 on this quiz.

5. Find the mean temperature if the high is 20°C and the low is 12°C.

6. Find the mean temperature if the high is 6°C and the low is -7°C.

7. Find the mean temperature if the high is -1°C and the low is -9°C.

8. Give the coordinates of the midpoint of line segment TW, where $T = (-3, -3)$ and $W = (9, 3)$.

9. Give the coordinates of the midpoint of the segment with endpoints (a, b) and the origin.

10. The digits of e have been shown to be very random. Treating each of the first fifteen decimal digits as a separate element, calculate the mean, mode, median, and midrange for this sample.

11. What would you expect each of these average values to be, if say a million or billion digits of e were used?

12. Calculate the average growth rate for a portfolio with portfolio with the following consecutive annual interest rates: 5%, 15%, -25%, -10%, 20%.

13. Four students drive from Michigan to Florida (2000 km) at 110.0 kph and return at 90.0 kph. Find the average round trip speed, using the harmonic mean.

14. For the problem just above, what is their average round trip **velocity**?

15. Tom Foolery measures the voltage in a standard outlet as -120 volts, 160 volts, 90 volts, and 30 volts at random intervals. Help him calculate the RMS voltage.

16. 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 3.667); and Health, 2 credits, C (use 2.000); Debate, 2 credits, B$-$ (use 2.667). Express your results to three decimal places.

17. Using the inauguration ages from problem 4 above, calculate the 10% trimmed mean and 20% trimmed mean.

18. 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?

Name ________________________ Score ________

A.4 Stats 6&7 Review Quiz

1. Find the mean, and standard deviation for the sample data set below.

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 A.1: Fictitious Salary Data Illustrating Use of Frequency.

2. 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.

3. What does Chebyshev's Theorem say about the number of IQs between 85 and 115?

4. The Unibomber (Theodore Kaczynski) has been often cited with an IQ of 170. Calculute how many standard deviations above the mean this corresponds to. Round your answer to two decimal places.

5. Using the mean of 54.9 and the standard deviation of 6.3, list the inauguration ages for any president beyond two standard deviations from the mean.

6. Add five years ($L_1 + 5 \rightarrow L_2$) to your presidential inauguration data and recompute the mean and standard deviation. How did they each change?

7. Increase your original presidential inauguration data by 10% ($L_1 \times 1.1 \rightarrow L_2$) and recompute the mean and standard deviation. How did they each change?

8. Add 5 years then increase your original presidential inauguration data by 10% ($(L_1 + 5) \times 1.1 \rightarrow L_2$) and recompute the mean and standard deviation. How did they each change?

9. Increase your original presidential inauguration data by 10% then add 5 years ($L_1 \times 1.1 + 5 \rightarrow L_2$) and recompute the mean and standard deviation. How did they each change?

10. 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.

11. Graduating Math and Science Center students have a mean SAT score of 1279. 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 subtests—multiplying 111 by the square root of two.)

12. 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?

13. Calculate the 5-number summary (using your TI-84+ calculators) for the data set given in problem one.

14. Calculate the z-score for the largest value in the above data set. Is it an ordinary score? Is it an outlier? Which definition works best?

15. Using the data set: $\{0, 2, 4, 5, 6, 3, 6, 1, 1, 50\}$, as given in the lesson, calculate the lower and upper hinge.

16. 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. Compare these results with those of your TI-84+ calculator.

Name ____________________ Score ________

A.5 Stats 8 Review Quiz

1. Create a pie chart for the Center student distribution data given below.

Grade	**Frequency**
9 (freshmen)	30
10 (sophomores)	29
11 (juniors)	24
12 (seniors)	25

Figure A.2: Frequency Table of Center Students by Grade Level.

2. Complete the frequency and relative frequency columns on this table for the 1999 Algebra Diagnostic Test Score data which follows: 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, 56, 55, 54, 53, 53, 50, 47, 35, 32.

Test Score	**Frequency**
20 – 39	
40 – 59	
60 – 79	
80 – 99	
100 – 119	
120 – 139	
140 – 159	

Figure A.3: Frequency Table of 1999 Algebra Diagnostic Test Scores.

3. Create an ogive for the 1999 Algebra Diagnostic Test Score data given above.

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 test score data displayed below.

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

Figure A.4: Stem and Leaf Diagram for Presidential Inaugural Data.

8. List the class boundaries for the data displayed above.

9. List the class limits for the highest class in the data displayed above.

10. Find P_{10}, P_{90} and the 10–90 percentile range for the data given in problem 2. Show all your work.

Name ______________________________ Score ________

A.6 Cumulative Quiz through Binomial

Show Work! Notes only Group. Lowest score recorded?

1. When drawing one card from a standard deck, what is the probability of getting an even (2, 4, 6, 8 or 10) or a black card? (Be sure to clearly show any adjustments which must be made and why.)

2. How many circular permutations can be made from BRITTAINA?

3. The homework for lesson 7 gave a method for approximating factorials. Apply it to find the log of 75! and then clearly indicate how to obtain from this a scientific notation approximation for 75!. (Be sure to keep enough significance in the mantissa to get four significant digits in the answer.) Compare this with the correct value $2.48091408 \times 10^{109}$.

4. Pick A. Low plays a musical solo. She is quite good and figures her probability of playing any one note right is 99.7%. The solo has 56 notes. What is her probability of:

   (a) Getting every note right?

   (b) Making exactly two mistakes?

5. For Ms. Low above, what is the probability of her making at least one mistake? What is the name of the rule used to calculate this easily?

6. Which of the following can be treated as a binomial experiment? Why or why not?

   (a) Testing a sample of 5 contact lenses (with replacement) from a population of 20 contact lenses, of which 40% are defective.

   (b) Testing a sample of 5 contact lenses (without replacement) from a population of 20 contact lenses, of which 40% are defective.

   (c) Tossing an unbiased coin 500 times.

   (d) Tossing a biased coin 500 times.

   (e) Surveying 1700 TV viewers to determine whether or not they watched the Super Bowl.

7. Find the probability of getting exactly 6 girls in 10 births. (Assume male and female births are equally likely and that the birth of any child does not affect the gender of any other child.)

8. Forty percent of adult workers have a high school diploma but did not attend college. If 20 adult workers are randomly selected, find the probability that at least 12 of them have a high school diploma but did not attend college.

9. Use the normal approximation to the binomial for the previous problem and discuss its validity.

10. A quiz consists of 10 multiple choice questions, each with 5 possible answers. For someone who makes random guesses for all the answers, find the probability of passing if the minimum passing grade is 60%.

Name ______________________________ Score ______

A.7 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.

□ 25

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. $\dfrac{\Sigma(x-\mu)^2}{n}$
_____ range rule of thumb	C. 68%–95%–99.7%
_____ standard deviation	D. $1-\frac{1}{K^2}$
_____ variance	E. $\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}}$

□
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 99th 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.

□
25

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 tendancy 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 ______________________ Score ________

A.8 Released Test: Prob. & Dist. May 20, 2004

```
No textbooks allowed, but please use your two notecards and
your graphing calculator.  Each of the 21 question numbers has
equal weight (i.e. 5 points each).  Read the questions carefully.
Hand in any used scratch paper with the test.  SHOW YOUR WORK.
```

□ 10

1,2. Form the best match among the following items.

_____experiment	A.	examples: rolling die, flipping coin, drawing card
_____random experiment	B.	more than one roll, flip, or draw
_____sample space	C.	each element has an equal chance of being chosen
_____impossible	D.	method by which observations are made.
_____certain	E.	set of all possible outcomes
_____simple event	F.	where emprical approaches actual probability
_____compound event	G.	each outcome is equally likely
_____random sample	H.	$P(A) = 0$
_____law of large numbers	I.	outcome which can't be broken down
_____fair	J.	$P(A) = 1$

□ 5

____ **3**. Find the number of **circular** permutations using the letters of the word: **POISSON**.

□ 5

____ **4**. Telephone numbers in North America have three groups of digits which must meet certain requirements. Before 1995, the three digit Numbering Plan Area (NPA) code, (commonly known as an area code) had the format NBX, where N could not be 0 or 1, B had to be 0 or 1, and X could be any digit 0 through 9. How many different NPA codes were there then?

□ 5

5. Discuss the meaning of type I and type II errors within the context of an air bag switch, whether or not it triggered, and whether or not it should have.

Test continues on the back of this page.

□ 25

□ 10

6,7. Form the best match among the following items.

_____P(At least one)	A.	event outcomes/total outcomes
_____Mutually exclusive	B.	$1-\beta$
_____Exhaustive	C.	$P(A)+P(\bar{A})=1$
_____Addition rule	D.	1 - P(none)
_____Def. of probability	E.	$\sum x \cdot P(x)$
_____Expected Value	F.	$P(A \text{ or } B) = P(A)+P(B)-P(A \text{ and } B)$
_____P Value	G.	no overlap
_____complementary rule	H.	Everything enumerated
_____Bayes Theorem	I.	An area like α.
_____Power	J.	$P(A\|B) = \frac{P(A)\cdot P(B\|A)}{P(A)\cdot P(B\|A)+P(\bar{A})\cdot P(B\|\bar{A})}$

□ 5

8. Green, Blue, Red, and Plaid are running a race. The following **odds against** are listed: Green: 1 to 1; Blue: 2 to 1; Red: 8 to 1; and Plaid: 17 to 1. Give each contestant's corresponding probability of winning. Did we account for all opponents?

□ 5

____ **9**. A couple having fun one evening decide to simulate various families of three children. Help them create eight families of three children. Initialize your TI-83+ random number generator as follows: 0 → `rand` (`rand` is `math/prob/1`), then do `int(2rand)` 24 times (or `randInt(0,1,24)`) to find out whether they got a boy (1) or a girl (0). Arrange these in order into eight "families" of three children and calculate the average number of boys in them. (NonTI-83+ users, document your procedure to ensure reproducibility.)

□ 5

____ **10**. A box contains three \$1 bills, four \$2 bills, four \$5 bills, six \$10 bills, and three \$100 bills. A person is charged \$20 to select one bill. Find the expected value.

Test continues on the next page.

□ 25

□ 5

11. Farmer Calkins planted 100 hills of corn per row and 10 rows. Each hill received three kernels. Although three varieties were used, each variety had a published germination of 90%. Using only germination rate and number or kernels per hill (ignoring climatic factors, mechanical damage, crows, cut worms, gophers, *etc.*) what is the probability for no stalks, one stalk, two stalks, or three stalks **per hill.**

□ 5

____ **12**. Farmer Calkins is watching crows landing in his newly planted corn field. He estimates about one crow arrives every minute, the probability of two arriving in a minute is small enough to be ignored, and the crows arrived independently. Find the probability that exactly two crows arriving in any given one-minute interval.

□ 5

____ **13**. Scientist Calkins knows a certain resonance is convolved with a Gaussian Distribution to produce a Voigt Profile but fits it anyway with $y_1 = \frac{1}{(100-x)^2+10^2}$. Graph this in a $50 < x < 150$ by $0 < y < 0.01$ window and find the FWHM. Show a sketch of your work. (Hint: Let $y_2 = 0.005$ and use `CALC (2nd TRACE) Intersect (5)` to find the boundary values.)

□ 5

____ **14**. Gosset discovered how to model sample distributions without knowing the population standard deviation *a priori*. How many times bigger is the area under his distribution more than 1.96 standard deviations away from the mean for a sample of size 6 than that predicted by the empirical rule. Show a sketch of the region.

□ 5

____ **15**. Farmer Calkins had some old corn seed which he figured had only an 80% germination rate. Help him calculate a χ^2 for a goodness of fit test if he took the following sample:

stalks:	0	1	2	3
observed	7	100	350	543
expected	8	96	384	512

Test concludes on back of this sheet.

□ 25

□ ___ ___ **16,17**. For the old corn seed data sample in the problem above, calculate the mean number of observed stalks per hill together with the standard deviation. Give your results to four significant digits.

10

□ ___ ___ **18**. Farmer Calkins calculates the expected mean number of stalks per hill for the old seed as $\mu = np = 3(0.8) = 2.4$ and expected standard deviation as $\sigma = \sqrt{npq} = \sqrt{3(0.8)(0.2)} \approx 0.693$. Using either these values or preferably the similar values from the previous problem, calculate the (5%) margin of error and corresponding 95% confidence interval for the true average number of stalks. Clearly indicate the standard error of the mean for his $n = 1000$.

5

□ **19,20**. E.T. is sitting by the estuary with his trinoculars trained on the bank display. He/she records the following data. Help E.T. interpret this data by doing a linear regression between the first and second and between the second and third data value from each ordered triple. You suspect the first value is time so please convert it to: minutes after the first observation. (6:00,69,20), (6:31,57,14), (6:59,49,10), (7:30,40,5), (8:00,32,0). Be sure to include and interpret r and r^2 for both regressions.

10

Part II, Bonus Question, 10%, 10 points

□ ___ **21**. The problems on the previous page (test page 3) all involve different distributions. In order, give the name of each distribution.

0

□

$25 + 10$

Appendix B

Solutions to Homeworks and Tests

B.1 Summary of Sampling Box Means

Please enter your **sample means** in the space provided.

Student name/bead color:	green	red	blue	lavender	orange
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					

Name ______________________ Score ________

B.2 Solutions HW 7: Binomial/Hypergeometric

1. Separately calculate using the binomial formula the probabilities of getting 0, 1, 2, 3, or 4 left-handed students in a class of 25, given a probability of 0.1. Compare your results with those obtained by doing binompdf(25,.1) (`DISTR 0`) or running `BINOMIAL` on your TI-84+ graphing calculator.

x	$P(x)$	${}_{25}C_x p^x q^{n-x}$
0	0.07179	$1 \cdot 0.1^0 \cdot 0.9^{25}$
1	0.19942	$25 \cdot 0.1^1 \cdot 0.9^{24}$
2	0.26589	$300 \cdot 0.1^2 \cdot 0.9^{23}$
3	0.22650	$2300 \cdot 0.1^3 \cdot 0.9^{22}$
4	0.13842	$12650 \cdot 0.1^4 \cdot 0.9^{21}$

2. Using only the data from the problem above, and the data from the example in the lecture, find the probability of getting more than four left-handed students in a class of 25. Compare your results with those obtained by doing `1-binomcdf(25,.1)` (`DISTR A`) on your TI-84+ graphing calculator.

   $1 - (.07 + .20 + .27 + .23 + .14) = 0.10$ (s/b 0.09799)

3. Check the assumptions carefully and see if we are justified in using the binomial (and not the hypergeometric) distribution for the problems above.

   Sample likely less than 10% of population.

4. Calculate the probability described in the text for winning the lottery by matching all 6 of 54 numbers.

   $P(x = 6) = [6!/(0!6!)] \cdot [48!/(48!0!)] \div [54!/(48!6!)] = 3.87 \times 10^{-8}$

5. Calculate the probability described in the text for winning the lottery by matching 5 of the 6 selected numbers from 54.

   $P(x = 5) = [6!/(1!5!)] \cdot [48!/(47!1!)] \div [54!/(48!6!)] = 1.12 \times 10^{-5}$

6. Calculate the probability described in the text for losing the lottery by not matching any of the 6 selected numbers from 54.

   $P(x = 0) = [6!/(6!0!)] \cdot [48!/(42!6!)] \div [54!/(48!6!)] = 0.475$

7. Use the normal approximation for the binomial to calculate the probability of getting 11 heads in 20 attempts from a fair coin (ignore the magic number test). Be sure to use the continuity correction and calculate the area under the probability density curve from 10.5 to 11.5. Compare this carefully with the results from the binomial formula.

   mean= $n \cdot p = 10$ s.d.= $\sqrt{n \cdot p \cdot q} \approx 2.236$

x	$x-\frac{1}{2}$	$x+\frac{1}{2}$	$z(x-\frac{1}{2})$	$z(x+\frac{1}{2})$	$P(x \pm \frac{1}{2})$	${}_{20}C_x(\frac{1}{2})^{20}$
10	9.5	10.5	−0.224	0.224	0.177	0.176
11	10.5	11.5	0.224	0.671	0.160	0.160
12	11.5	12.5	0.671	1.118	0.119	0.120
13	12.5	13.5	1.118	1.565	0.0730	0.0739
14	13.5	14.5	1.565	2.013	0.0367	0.0370
15	14.5	15.5	2.013	2.460	0.0151	0.0148
16	15.5	16.5	2.460	2.907	0.00512	0.00462
17	16.5	17.5	2.907	3.354	0.00143	0.00109
18	17.5	18.5	3.354	3.801	3.26×10^{-4}	1.81×10^{-4}
19	18.5	19.5	3.801	4.249	6.13×10^{-5}	1.91×10^{-5}
20	19.5	20.5	4.249	4.696	9.42×10^{-6}	9.54×10^{-7}

   Only a few percent error until more than 2 s.d. from mean.

8. Use the normal approximation for the binomial to calculate the probability of getting 12 heads in 20 attempts from a fair coin (ignore the magic number test). Compare this carefully with the results from the binomial formula. Is this the same as the probability of getting 8 heads?

   See above. Yes.

9. Use the normal approximation for the binomial to calculate the probability of getting 13 heads in 20 attempts (ignore the magic number test). Compare this carefully with the results from the binomial formula. Is this the same as the probability of getting 7 heads?

   See above. Yes.

10. How likely is it to get 15 or more heads in 20 attempts, if the coin is fair?

    0.02069 by binomcdf
    0.0221 by normalcdf($14.5 > x \equiv 2.0 > z$)

11. A common rule is that you can approximate the binomial with the normal when both _________ and _________ exceed the magic number of _____.

    np and nq 10 (or 5 or 15...)

Name ______________________________ Score ________

B.3 Solutions HW 10: Student t-Distribution

1. Find the value of t from the table which has a probability of 0.05 to the right of t when $n = 6$.

   Answer: 2.015 df=5 1-tailed

2. Use the table of t values to find a t value with probability of 0.99 to the right of t when $n = 21$.

   Answer: -2.528 df=20 1-tail (symmetry and 0.01.)

3. What value(s) of t would you use to find a 95% confidence interval for the mean of a population if $n = 16$?

   Answer: ± 2.132 df=15 (almost always 2-sided CI)

4. Use `tcdf` on your calculator to find a t value for $n = 8$ and a one-tailed $\alpha = 0.005$. You might start by comparing the results of `tcdf(4.032,9E99,5)` and `tcdf(3.169,9E99,10)` on your calculator with the corresponding entries in the table in the lesson.*

   Answer: df $= 7$ and $t = 3.499$ by guess and check.

5. Suppose you have a one-sample t statistic from a sample of $n = 6$. Suppose further that you calculated a t value of $t = 2.80$ for your hypothesized population mean (H_0: $\mu = 64$ and H_a: $\mu \neq 64$). Give the two-tailed probabilities which bracket this value. Calculate the **P-value** (twice the area to the right of this t value). Should you reject or fail to reject the null hypothesis?

   Answer: tcdf(2.80,9E99,5)=0.018997 P-value=0.037994
   Reject at alpha=0.05, but not at alpha=0.01.
   Thus significant at .05 but not at .01 level.

   A university researcher placed 12 randomly selected radon detectors in a chamber that exposed them to 105 picocuries per liter of radon. The detector readings were as follows: 91.9, 97.8, 111.4, 122.3, 105.4, 95.0, 103.8, 99.6, 96.6, 119.3, 104.8, and 101.7.

*Some later/updated TI-84 calculators have an inverse T function.

6. Construct a stem-and-leaf diagram of the above data using stems split two ways (i.e. 90–94, 95–99, ...). (Hint: it might be easier to round to integer first.)

9	2
9	578
10	024
10	55
11	1
11	9
12	2

9	1.9
9	5.0 6.6 7.8 9.6
10	1.7 3.8 4.8
10	5.4
11	1.4
11	9.3
12	2.3

Note: the first table was rounded as well.

7. Check whether the sample size and skewness allow use of a t test.

Answer: $n = 12 < 15$ so check for skewness and outliers. no outliers slightly skewed, but not too badly.

8. Calculate a t-value for the sample mean *versus* the population mean (105).

Answer: $\frac{104.13-105}{9.397/\sqrt{12}} = -0.32$

9. Calculate the areas under the curve further away from the mean for this value of t (two-tailed). Is there convincing evidence that the mean reading of all detectors of this type differ from the true value?

Answer: $P(t < -0.32) = 0.377$ P-value=0.754 No

10. Calculate a two-sample t statistic for the data obtained from the 2000 penny experiment ($\bar{x} = 15.2$, $s = 2.71$, $n = 18$ for Calkins and $\bar{x} = 12.2$, $s = 1.39$, $n = 9$ for Burdick).

Answer: $\frac{(15.2-12.2)-0}{\sqrt{\frac{2.71^2}{18}+\frac{1.39^2}{9}}} = 3.80$

11. Calculate the fractional degrees of freedom for the above penny experiment using the formula given at the end of the lecture on two-sample t tests. ($n_1 = 18$ and $n_2 = 9$). Compare this number with that obtained from the TI-84+ calculator `STAT TESTS 4: 2-SampTTest ... not equal, not pooled, calculate.`

Answer: $\frac{(\frac{2.71^2}{18}+\frac{1.39^2}{9})^2}{\frac{\left(\frac{2.71^2}{18}\right)^2}{17}+\frac{\left(\frac{1.39^2}{9}\right)^2}{8}} = 24.9$ By calculator: 24.9

Name ______________________________ Score ________

B.4 Solutions HW 11: Central Limit Theorem

1. Given a 2003 penny data sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate the margin of error (assume a 95% confidence interval will be generated).

   Answers: $2.132 \cdot 1.91 \div \sqrt{16} = 1.018$ small n use t

2. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate a 95% confidence interval.

   Answers: $15.8 - 1.02 = 14.8$ $15.8 + 1.02 = 16.8$ $(14.8, 16.8)$

3. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate the margin of error (assume a 99% confidence interval will be generated).

   Answers: $2.947 \cdot 1.91 \div \sqrt{16} = 1.41$

4. Given a sample mean of 15.8 and a sample standard deviation of 1.91 (with $n = 16$), calculate a 99% confidence interval.

   Answers: $15.8 - 1.41 = 14.4$ $15.8 + 1.41 = 17.2$ $(14.4, 17.2)$

5. A **P-value** is a way to express the confidence of our results. For a one-tailed test, it is the area under the curve to the right (or left) of our observed mean. Calculate a t-score using our observed mean (15.8), expected mean (10.0), and standard error ($1.91/\sqrt{16}$) and sketch this region on a normal curve.

   Answers: $t = \frac{15.8 - 10.0}{1.91/\sqrt{16}} = 12.15$

6. Calculate this area by doing a `tcdf(`t`,9E99,15)`, where t is the value calculated above, and there are 15 degrees of freedom.

   Answers: $P(t > 12.15) = 1.82 \cdot 10^{-9}$

7. **Alpha** (α) is the term used to express the level of significance we will accept. For 95% confidence, $\alpha = 0.05$. If our P-value is less than alpha, we can reject our null hypothesis (H_0: $\mu = 10$). Should we reject our null?

   Answers: YES! $0.05 >> 1.82 \cdot 10^{-9}$.

8. Try to identify sources of error or bias which might account for these (highly significant) results.

   Answers: Perhaps the tail side is heavier.
   Perhaps the rim slants slightly toward the back.

9. Do you think other coins might display similar characteristics? How many times would you have to test it to reach a significant conclusion.

   Answers: Maybe. That depends on how much the results differ from 50 : 50. Perhaps many times if bias only slight.

10. Do you think spinning coins (especially some of the new and different state quarters) might display similar characteristics? We may hand out a data gathering sheet with very specific collection instructions.

    Answers: Maybe/Maybe not.

11. How willing are you to bet money using this method of "flipping" a coin (assuming you have no scruples against such an activity)?

    Answers: $50? It does seem to be a fairly good bet!

Name ______________________ Score ______

B.5 Solutions HW 12: Correlation and Regression

1. Suppose you remember the triangular numbers, but can't remember their formula. Enter the following values into L_1: 1, 2, 3 and L_2: 1, 3, 6. Now do a `QuadReg` L_1, L_2 using your TI-84+ calculator (`STAT, CALC 5`) and interpret the results (rewrite the formula in its usual form).

   Answer: $y = ax^2 + bx + c \qquad a = .5 \qquad b = .5$
   which means: $T_n = \frac{n(n+1)}{2}$

2. Suppose further you really couldn't remember if the relationship was quadratic. Try a `CubicReg` L_1, L_2 and interpret the results.

   Answer:
   ERR:DOMAIN
   1:Quit
   2:Goto
   There are too few points for the indicated fit.

3. Add an additional point onto the end of L_1: 4 and L_2: 10. Repreform the `CubicReg` L_1, L_2 and interpret the results.

   Answer:
   $y = ax^3 + bx^2 + cx + d \qquad a = d = 0 \qquad b = .5 \qquad c = .5$
   which means: $T_n = \frac{n(n+1)}{2}$

4. Try the `CubicReg` L_1, L_2 with L_1: 1, 2, 3, 4 and L_2: 1, 5, 14, 30 (sums of squares) and interpret the results. Try to express your answer in an aesthetically pleasing form (*i.e.* fractions not decimal fractions).

   Answer:
   $y = ax^3 + bx^2 + cx + d$
   $a = .3333333333$
   $b = .5$
   $c = .1666666667$
   $d = -8.2E - 12$ (Note: this is only calculator round-off errors from zero)
   $R^2 = 1$
   $y = \frac{1}{6}(2x^3 + 3x^2 + x) = \frac{x(x+1)(2x+1)}{6}$

5. Question one can be done by solving three equations in three unknowns. Specifically, let $ax^2 + bx + c = y$. Then substitute each value of x and equate it to the corresponding y value. Solve these three equations manually by elimination (due to the regular spacing, first c, then b eliminate easily).

   Answer:

$$\begin{array}{lllll}
x \longrightarrow 1 & a+b+c=1 & & & \\
 & & 3a+b=2 & & \\
x \longrightarrow 2 & 4a+2b+c=3 & & 2a=1 & a=\frac{1}{2} \\
 & & 5a+b=3 & & \\
x \longrightarrow 3 & 9a+3b+c=6 & & \swarrow & \\
 & & b=\frac{1}{2} & & \\
 & c=0 \qquad \longleftarrow & & &
\end{array}$$

6. Solve the above equations using your calculator, either using augmented or inverse matrices. Record the pertinent keystrokes here.

   Answer: $\boxed{\text{MATRIX}}$ $\boxed{\text{EDIT}}$ $\boxed{1}$ 3×4

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 4 & 2 & 1 & 3 \\ 9 & 3 & 1 & 6 \end{bmatrix} \boxed{\text{2nd}}\boxed{\text{MODE}} \text{ (QUIT)}$$

   $\boxed{\text{MATRIX}}$ $\boxed{\text{MATH}}$ $\boxed{\text{B}}$ (rref) $\boxed{\text{MATRIX}}$ $\boxed{\text{NAMES}}$ $\boxed{1}$ ([A]) $\boxed{\text{ENTER}}$ $\boxed{\text{ENTER}}$

$$\begin{bmatrix} 1 & 0 & 0 & .5 \\ 0 & 1 & 0 & .5 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

   Let $AX = Y$ and $A = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{bmatrix}$, $X = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$, and $Y = \begin{bmatrix} 1 \\ 3 \\ 6 \end{bmatrix}$. Since $A^{-1}AX = A^{-1}Y$ and $A^{-1}A = I$, $IX = A^{-1}Y$ and $X = A^{-1}Y$. The calculator, unfortunately, uses the x^{-1} key for matrix inverse and can multiple these two matrices together to give you the correct results.

Name Key Score 100/100

B.6 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 descendi
no commas or horizontal lines
no missing number
Don't omit stem 5

$\boxed{5}$ 5 **1**. Construct a **stem-and-leaf** diagram.

$\boxed{5}$ 5 70.0 **2. Midrange**. (max+min)/2=$\frac{42+98}{2}$=70.0

$\boxed{5}$ 5 70.1 **3. Arithmetic Mean**. $\frac{83+68+66+68+98+60+42+71+75}{9} = \frac{631}{9} = 70.111\ldots$

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

$\boxed{5}$ 5 15.4 **4. Standard Deviation**. $s = 15.35777, \sigma = 14.47944$

Data set is a SAMPLE so use s.

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

$\boxed{5}$ 5 -1.82 **5**. Show how to compute the z-**score** for the smallest test score. Put your answer in the proper format.

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

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

Use 2 decimal places!

$\boxed{25}$ 25

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

$\boxed{5}$ 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.

$\boxed{5}$ 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.

$$30 \cdot 70 = 2100$$
$$2100 - 90 - 22 = 1988$$
$$1988/28 = 71.0$$

$\boxed{5}$ 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/media
quadratic mean, and maximum.

$\boxed{5}$ 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.

$\boxed{5}$ 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.

$\boxed{25}$ 25

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. $\dfrac{\Sigma(x-\mu)^2}{n}$
A	range rule of thumb	C. 68%–95%–99.7%
E	standard deviation	D. $1-\frac{1}{K^2}$
B	variance	E. $\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}}$

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.

$\boxed{15}$
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.

$\boxed{5}$
5

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

See Statistics Section 3.3.

$\boxed{5}$
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
signature

Oct. 22, 2001
date

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

$\boxed{25}$
25

B.7 Key: Released Prob&Dist test: May 20, 2004

1–2. DAEHJIBCFG

3. $\frac{(7-1)!}{2!2!} = 180.$ 7 letters, reduce by one since circular, and two letters repeated (OS), each twice.

4. $8 \times 2 \times 10 = 160$.

5. Let the test be an air bag trigger and the condition being tested be a collision. Let H_0 be the air bag triggered properly and H_a be it did not. Then a type I error or false negative would be our alpha region or there was a collision and the air bag did not trigger. A type II error or false positive would be the beta region or there was not a collision and the air bag did trigger. Both kinds of errors are problematic, but for different reasons.

6–7 DGHFAEICJB

8. $\frac{1}{2} + \frac{1}{3} + \frac{1}{9} + \frac{1}{18} = \frac{9+6+2+1}{18} = \frac{18}{18} = 1.000.$ Yes.

9. Remember, if you seed the pseudo-random number generator with zero first, you will always get these values: 1 1 0; 1 0 1; 0 0 1; 0 1 1; 0 0 0; 1 0 0; 1 1 1; 0 0 0. Thus in the 8 families there are 11 boys for an average of $11/8 = 1.375$ boys per family.

10. $\frac{3(1-20)+4(2-20)+4(5-20)+6(10-20)+3(100-20)}{20} = \frac{-57-72-60-60+240}{20} = \frac{-9}{20} = -\$0.45.$ The denominator 20 is part of each probability whereas the numerator 20 is the cost per each win.

11. Using `binompdf(3,.9)` or the binomial distribution formula/program one obtains: $(0, 0.001), (1, 0.027), (2, 0.243)$, and $(3, 0.729)$. Here $n = 3$ and $p = 0.9$.

12. Using `poissonpdf(1,2)` where the first argument is the mean and the second argument is the value whose probability you seek, one obtains 0.1839. Note that this is also the same as the bank example in the text. Let $\mu = 1$ and $x = 2$. $P(2) = \frac{e^{-1}}{2!} = \frac{1}{2}0.3679 = 0.1839.$

13. The line $x = 0.005$ intersects the Lorentzian at $x = 90.0$ and $x = 110.0$ MHz. Thus the FWHM is $110.0 - 90.0$ MHz or 20.0 MHz. (Oops—there are no units given in the problem.)

14. We compare here, in relative magnitude, the right tail areas of these two distributions. Be sure to include a sketch. `tcdf(1.960,9E99,5)`÷ `normalcdf(1.960,9E99)` yields the result 2.1459 indicating the area under the probability distribution curve is over twice as big when you have such a small sample and do not know the population standard deviation.

15. Using the formula $\frac{(O-E)^2}{E}$ and summing over all observed (O) and expected (E) we obtain: $0.125 + 0.167 + 3.014 + 1.877 = 5.179.$

16–17. Using the calculator for a frequency mean: `1-Var Stats` L_1,L_2, we put 0, 1, 2, 3 in L_1 and 7, 100, 350, and 543 in L_2. This gives a mean of 2.429 and sample standard deviation 0.6982.

18. The standard error of the mean is $0.6982/\sqrt{1000} = 0.0221$. The margin of error corresponding with an alpha of 0.05 is $1.960 \times 0.0221 = 0.0433$. The confidence interval then is 2.429 ± 0.043 or $(2.386, 2.472)$.
Note 1: the order numbers are given in "interval notation" is critical. Be sure to have the left/least first and second/greatest second. Note also that interval notation for an open interval (endpoints not included) can be ambiguous with an ordered pair.
Note 2: one can use `invNorm(.975)` (one-tail) to find the z value which corresponds to a (two-tail) 95% confidence interval.

19–20. Put 0, 31, 59, 90, 120 in L_1, 69, 57, 49, 40, 32 in L_2, and 20, 14, 10, 5 0 in L_3. Do `LinReg` L_1,L_2 and obtain $y = -0.198x + 63.65$ with $r = -0.946$ and $r^2 = 0.89$. Thus the time and temperature are negatively correlated with 89% of the variation in temperature explained by the variation in time (it cools off in the evening). Do `LinReg` L_2,L_3 and obtain $y = 0.537x - 16.7$ with $r = 0.999$ and $r^2 = 0.998$. Thus the temperatures in Fahrenheit and Celsius are very well positively correlated with 100% of the variation in Celsius explained by the variation in Fahrenheit.
Note 1: it doesn't matter which linear regression you use on the calculator.
Note 2: if r and r^2 are not displayed, then you need to enable them by doing a `Diagnostics On` from `catalog`.

21. Binomial, Poisson, Lorentzian, Student t (or just t), and Chi Square.

For the May 20, 2005 test students redid their test for additional test points. Since then, a released test has been provided.

1. No erasures.
2. Use a different color (pen *vs* pencil; black *vs* blue; *etc.*).
3. The redos were due back exam week with the following diminishing returns:
   - Monday/Tuesday (5/23 or 5/24): 1/2 points
   - Wednesday (5/25): 1/3 points
   - Thursday (5/26): 1/4 points
   - Friday (5/27): 1/5 points
   - Later: 0 points

B.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 $\sum$							
2								
2								
2								
2								
	Table 2 $\sum$							
3								
3								
3								
3								
	Table 3 $\sum$							

Table	M&M® Color: **your name**	brown	yellow	red	orange	green	blue	χ^2
4								
4								
4								
4								
	Table 4 $\sum$							
5								
5								
5								
5								
5								
	Table 5 $\sum$							
6								
6								
6								
6								
6								
	Table 6 $\sum$							
7								
7								
7								
7								
	Table 7 $\sum$							

Figure B.1: Collection Point for χ^2 M&M Data.