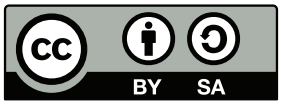
Statistics for Scientists
An inquiry based approach to learning

Ben Woodruff[1], with much work borrowed from Craig Johnson

Typeset on April 8, 2014



[1]Mathematics Faculty at Brigham Young University–Idaho, woodruffb@byui.edu

Contents

Introduction iii

I Exploring Data 1

1 Statistical Exploration 2
- 1.1 Measures of Center . . . 2
- 1.2 Measures of Spread . . . 7
  - 1.2.1 Spread in a Population . . . 10
  - 1.2.2 Spread in a Sample . . . 13
- 1.3 Measures of Position . . . 14
  - 1.3.1 z-scores . . . 14
  - 1.3.2 Percentiles . . . 16
- 1.4 Graphical Displays of Data . . . 17
  - 1.4.1 Boxplots . . . 17
  - 1.4.2 Histograms . . . 18

2 Probability 20
- 2.1 Probability Density Functions . . . 24
- 2.2 Cumulative Distribution Functions . . . 27
- 2.3 Mean, Variance, and Standard Deviation of a Discrete Random Variable . . . 30
  - 2.3.1 Discrete Random Variables . . . 30
- 2.4 Continuous Random Variables . . . 33

3 Jointly Distributed Random Variables 41
- 3.1 Conditional Probability and Bayes' Rule . . . 41
  - 3.1.1 Bayes' Rule . . . 45
- 3.2 Jointly Distributed Random Variables: Discrete Case . . . 46
  - 3.2.1 Expected Values of Functions of Jointly Distributed Discrete Random Variables . . . 47
  - 3.2.2 Conditional Probability (Revisited) . . . 52
- 3.3 Component Failure - Series Versus Parallel . . . 55

II Statistical Inference 61

4 The Central Limit Theorem 62
- 4.1 Tolerance Stack-up . . . 62
  - 4.1.1 Propagation of Error . . . 63
- 4.2 The Sample Mean . . . 68
  - 4.2.1 Bernoulli Distribution . . . 69
  - 4.2.2 Normal Distribution . . . 70

4.3 The Distribution of the Sample Mean - The Central Limit Theorem 75

5 Inference for a Single Population Mean 80
5.1 One Population Mean: σ Known 80
5.1.1 Point Estimators 80
5.1.2 Confidence Interval for One Mean, σ Known 81
5.1.3 Sample Size Calculations 84
5.2 Hypothesis Test for One Mean, σ Known 85
5.3 One Population Mean: σ Unknown 93

6 Inference for Two Population Means 99
6.1 Paired Data: Dependent Samples 99
6.1.1 Hypothesis Tests 102
6.2 Two Independent Samples 108
6.2.1 The Standard Error 108
6.2.2 Hypothesis Tests and Confidence Intervals 111

III Looking for Correlation 119

7 Analysis Of Variance (ANOVA) 120
7.1 One Way ANOVA 120

8 Linear Regression 129

IV In Progress 132
8.1 Two Way ANOVA 133
8.2 2^p Factorial Design 133

V Appendix 134

9 Supplemental Data 135
9.1 List of Data Files 135

Introduction

What is Inquiry Based Learning

This class may be unlike any math course you have taken in the past. We'll be learning through inquiry, rather than lecture. You'll have the chance to jump in and discover the big ideas. You are the artists in this course, and you'll discover what you decide to paint. My job as your teacher is to create the scaffolding that will enable you to discover centuries of learning. I will craft the problems you work on so that they start where you are at and get you to the knowledge of the old masters.

Here's what a typical day of class might look like:

- You have 8 problems to work on before coming to class. You crack 6 of them, but try on the other 2 and fail (which is OK - it will happen).
- When you come to class, during the first 10-15 minutes we'll work at the boards together in small groups to tackle some problems. These will often be related to the preparation you did for the current day, and to help prepare you for the next day's problems.
- While you're at the boards, I'll randomly select some of you to present your prep solutions to the class. Come to class with your solutions already written up (one solution per page). I'll take your solutions and make them available on the class projector.
- We'll turn the time over to you to share your work. You'll present, and then defend your work. Your peers will ask you questions. Sometimes you'll be spot on right, and sometimes you'll be wrong. As long as you can justify why you did what you did, we'll all learn and grow.
- We end class and have another set of problem to tackle for the next day. You set up a group meeting time, where you learn to work together as a team.

As you progress in this course, you'll find that you enhance your ability to (1) reason critically, (2) present and defend your results,(3) work with a team, and (4) speak the language of mathematics. The entire structure of the course is designed to build these abilities in you.

As your instructor for this courses, I'm your coach and cheerleader. My goal is to build in you the knowledge of several centuries of work. If you'll jump in and start exploring, you'll find that you can learn so much with inquiry based learning, perhaps more than you've ever learned before.

Inquiry based learning has been around for hundreds of years, and has been shown over and over again to be more effective than lecture based learning. You can read more about it on the web at

- http://www.inquirybasedlearning.org/default.asp?page=Why_Use_IBL.

Adopting inquiry based learning requires that I turn our classroom over to you, the students. I've learned that you, my students, are capable of far more than I could ever have imagined.

Deep practice

To learn through inquiry, we have to be willing to explore ideas on our own. We have to be willing to allow ourselves to fail. Sometimes we'll tackle problems where there may not be a right answer. We have to formulate ideas, try them, and learn from the results (good or bad). We have to learn from our failures.

My first exposure to inquiry based learning occurred in the fourth grade. Our teacher gave us pretend money throughout the week for good behavior and then on Friday let us decide how to spend that money on goodies at our class store. We had to use our newly acquired arithmetic skills to figure out how to get the most bang for our buck. Some weeks I would purchase things, and then see another student's purchase and realize I had bought the wrong stuff. The next week I changed my purchasing plan. This was a weekly problem that I was allows to fail at, an then try again, with no punishment at all.

In the seventh grade, we learned an algorithm for solving absolute value inequalities. For two weeks, we had to repeat the algorithm over an over again on different problems. I wanted to find a faster way to do the problems, and discovered a way that relied on distances. Every time we learned a new concept, I tried to see if I could work the new concept into this faster approach. Some concepts took a little more trial and error to master with the faster approach, but eventually they all did. The time savings was amazing. I was super happy I had discovered something. When the test came for this material, I wanted to race my teacher to see if I could finish the exam before he finished passing out the exams. I beat him.

The next day of class, Mr. Nelson said he wanted to have me share with the class how I did the problems so quickly. He gave me the chalk and I got my first chance to stand in front of people and teach. After 5 minutes (or less), Mr. Nelson thanked me and told me to take a seat. No one in the room, not even my teacher, had a clue what I was doing. I failed.

This failure changed me. It changed entirely how I studied. I started working ahead of my class, because when they hit new material, I wanted to be able to answer questions for them. Ever time I learned an idea, I asked myself how I would share it with someone else. I was never again given a chance to teach an entire class, not even for a few minutes, but I was always ready. I didn't want to fail again.

I will always be grateful for a teacher, Mr. Nelson, who trusted me enough to allow me to try, and fail, at teaching his class. Failure is a key to success.

Why is Failure So Crucial?

I'm guessing that many of you have probably had most of your math classes delivered in lecture form. Does this describe your typical math class?

> You show up to class, you take notes, you then go home and do every other odd at the end of the chapter (or something similar), and then come to class the next day to listen to a lecture again. This repeats every day, and you get used to the monotony. Your teacher comes to class and shows you the right way to do everything. You are asked to reproduce the method you were shown. If you can't do it right within a day, you get docked on your homework.

I like to refer to this as, "Monkey see, Monkey do."

Rarely did you have to struggle with the big ideas. They were handed to you already in perfect form. A few of you may have taken the time to thoroughly digest the proof in the book. More importantly, most of us don't know what doesn't work. This is the key to understanding why the current approach matters. Most valuable ideas have a plethora of important failures that helped shape the success.

These failures are perhaps the most important parts of the puzzle. Why should you do a problem the way the textbook says to? Is there a better way? If you knew the underlying issues, and tried to solve the problem yourself, you'd discover the reasons why the algorithms in your book are so awesome.

Most textbooks rarely focus on teaching you what doesn't work. They focus on success only. Perhaps our entire culture puts too much emphasis on success, and likes to forget the huge amount of knowledge that comes from our failures. This is where inquiry based learning comes in so handy. To truly master and appreciate an idea, we have to struggle with it. We have to attack the idea, struggle, fail, and then continue working. We have to learn to celebrate our failures.

Discovery takes time. It takes effort. It can at times be frustrating. But when you've struggled with an idea, failed some, and eventually crack it, there is so much joy that can come from this endeavor. This is the joy of discovery, and it fuels scientific research. My hope is that each of you will feel this joy multiple times this semester.

Our course allows you to discover. You'll be asked to try new things that you've never been shown. You'll be given a chance to try, fail, and eventually succeed. You'll come to class and present what you've done. Sometimes you'll be wrong, and we'll celebrate anyway. Your errors will have helped everyone in the class see why a certain approach does not work. Sometimes you'll be spot on right, and not even think you are before you share. The goal is to learn through inquiry, and share with each other what you've learned.

Your Peers are A Valuable Asset

I've learned over the past year that your peers might be one of your most valuable assets this semester. Every one of you comes to the course with a different background. Some of you are algebra masters. Some of you know how to do calculus really well, but don't know the right words. When you meet with peers, you'll find that you help each other over hurdles that would otherwise completely halt your progress. Your peers are perhaps your most valuable asset.

As you tackle problems this semester, you will get stuck at certain parts. I'll draw upon your knowledge from all your past classes, and almost every has holes somewhere. Some of you will want to weather this road alone. This might mean spending hours trying to relearn something you forgot. If you decide to weather the road alone, and you put in the time to master each idea that we explore, you'll have gained way more than I could ever hope if I had lectured and showed you the steps of how to do everything.

You don't have to do it all alone. If you study with peers, you can help each other over hurdles. I've seen time and time again where all you needed was a peer, not a tutor, to help you continue on. We all forget things occasionally, and our peers can be valuable assets. If you would like some suggestions about how to improve a group meeting, please see the following page online.

http://bmw.byuimath.com/dokuwiki/doku.php?id=group_study_suggestions

Part I

Exploring Data

Chapter 1

Statistical Exploration

It has never been easier for businesses to collect and manage data. Unless data can inform business decisions, their value is limited. An important part of data analysis is "data reduction". This is the process by which data are summarized in a few **statistics**,[1] which can be used to make decisions.

We will explore several summary statistics in this chapter. Many of these statistics will already be familiar to you.

DARPA (Defense Advanced Research Projects Agency) [Public domain], via Wikimedia Commons

Write your answers to the questions on loose leaf papers that can be submitted in class. If you present your solution, it will be scanned and made available to the rest of the class.

Some of the statistics in this chapter have very well-defined, commonly-accepted definitions. Other statistics have several accepted definitions. Various statistical packages may employ different definitions, and they can return slightly different results. Try not to let this bother you. The results will be close, and each is acceptable.

1.1 Measures of Center

In this first section, we'll be analyzing ways of measuring the average, or center, of a collection of data.

Definition 1.1: Sample Mean $\bar{x}$ and Population Mean μ. The mean of a sample is computed by adding up the values in the data set and dividing by the total number of observations. We often denote the **sample mean** by

$$\bar{x} = \frac{\sum x_i}{n} \tag{1.1}$$

where n is the sample size and x_i are the observed values, and $i = 1, 2, 3, \ldots, n$.

If the data we have represent a full population, we compute the **population mean** in a similar manner using

$$\mu = \frac{\sum x_i}{N} \tag{1.2}$$

where N is the population size.

[1] A **statistic** is any number that is computed based on data.

The mean is just one way of measuring the center of a collection of data. The median and mode are two other common measures of center.

Definition 1.2: Median and Mode. The median is the numerical value that separates the higher and lower halves of a collection of data. To compute the median of a sample, or population, we arrange the observations in from lowest to highest and then pick the middle one. If there are an even number of observations, then we average the two middle values (so the median of $\{1, 2, 3, 4\}$ is $(2 + 3)/2 = 2.5$).

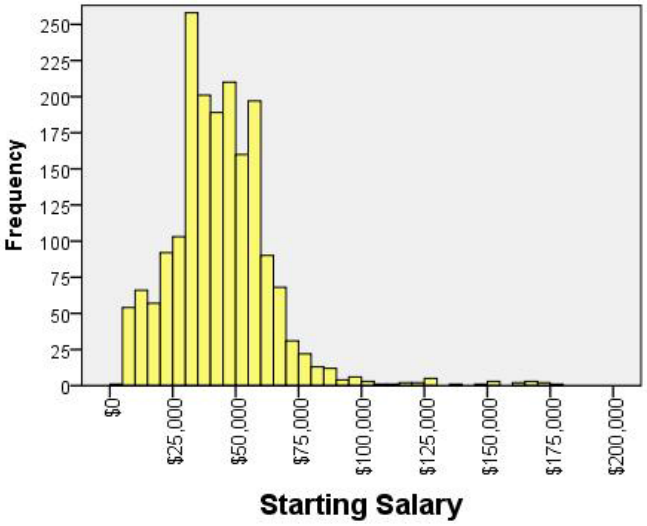
The mode is the value that occurs the most often in a collection of data.

Let's now explore the differences in these concepts as we tackle a few problems.

BYU-Idaho surveys its graduates to assess how well the University's goals are being met. Prior to a recent graduation ceremony, the population of BYU-Idaho graduates who had secured employment reported their major and their starting salary. The data are given in the file BYU-IdahoGradSalaries.xlsx.

Tip:
Data files are given as hyperlinks in this document. To download the data file, just click on the file name, which is an embedded hyperlink.

Problem 1.1 The following image is a histogram of the population of starting salaries given in the data file.



Note:
A histogram is the most popular way to visualize quantitative data. The bars of a histogram divide the x-axis into different "bins" or "classes". The height of the bars in a histogram tell the number of observations that fall in each bin.

In the histogram at left, each of the bins has a width of \$5,000. All starting salaries that are at least \$25,000 and less than \$30,000 will be in the same bin. Find this bin in the histogram at left. How many salaries are between these two values? (The answer is 103 salaries.) This is illustrated by the height of the bar in that bin. As another example, there are 258 salaries between \$30,000 and \$35,000.

For an explanation about histograms, you can watch a simple video at KhanAcademy (khanacademy.org/video/histograms) or search online for additional examples.

Use this figure to answer the following questions.

1. Describe the shape of the distribution of these salaries. Would you describe the shape as left-skewed, right-skewed, or symmetric? If you are not familar with these words, please see Wikipedia.
2. What is a "typical" starting salary for a BYU-Idaho graduate?
3. Are the salaries spread out or close together?
4. Are there any unusual starting salaries? How did you determine if a salary was "unusual"?

The university keeps track of the average starting salary of BYU-I graduates. If we know the salary of every graduate, then we have what we call the population average. If instead we only asked 20 graduates what their starting salary is, then we would have a statistic called the sample average that approximates the population average. The goal of statistics is to use samples to try to understand populations. If we know information about a sample, how can we use that information to infer things about the population.

Problem 1.2 Please download the file BYU-IdahoGradSalaries.xlsx.

1. Use Excel to find the mean starting salary for the full population of recent BYU-Idaho graduates given in the data file. You can use the command `=average()` to find the mean of an array of numbers in Excel.

2. The data file contains 1861 graduates. Let's simulate obtaining a sample of 20 students, and see how the mean of this sample of 20 students compares with the mean of the entire population. Randomly select 20 numbers between 1 and 1861. You can use excel command `=randbetween(1,1861)` to obtain a random number. If that command doesn't work, feel free to get 20 random numbers from http://www.random.org/ or use some other random number generator. Most hand held calculators have a random number generator in them.

3. Once you have 20 random numbers, make a list of the 20 salaries corresponding to each SurveyID. Then find their mean and compare it to the mean of the entire population.

Problem 1.3 Continuing the previous problem, answer the following.

Don't worry about trying to find the "right" answer. Allow yourself to explore and think critically about the questions.

1. Suppose you take many random samples (each of size $n = 60$) from the population. For each sample, imagine you compute the average of the salaries in that sample. If you were to make a histogram of the averages, how would the shape of the distribution of those averages compare to the figure in problem 1.1?

2. Suppose you took thousands of samples of size $n = 60$ from the population. Can you estimate what the average of the averages might be? How did you draw this conclusion?

3. How would the spread of the sample averages compare to the spread of the actual salaries illustrated in the histogram above? Justify your reasoning.

4. Would you be likely to observe "unusual" sample averages? Why or why not?

ComposiTite Inc., a materials company, wants to know if the yield strength of a particular material is different depending on which of two possible manufacturing processes is used (process A or process B). The company is interested in comparing and contrasting the two processes in order to select one of them as the "best" process. It is important to the company that the material has a minimum yield strength of 180 MPa.

In order to evaluate the two manufacturing process, the company sets up an experiment in which they manufacture 25 test coupons using each of the two processes. All 25 test coupons are tested to determine their yield strength. The data file YieldStrength.xlsx contains yield strength measurements for all 25 test coupons manufactured with each process (A and B). We will consider the 25 test coupons to be a sample that represents the actual behavior of each process.

A **test coupon** is an small piece of raw material to which a manufacturing process is applied. The resulting piece is tested for compliance with product specifications. This provides inexpensive and sample representative of a part of the production process.

Problem 1.4 Start by reading the two paragraphs prior to this problem. For each of the two processes, compute the mean, median, and mode for the yield strength of the process.

Measure of Center	Process A	Process B
1.		
2.		
3.		

What differences do you see among the "middle values" of these two processes?

Which measure of center is better to use? That depends entirely on what we want to show.

Problem 1.5 Two small companies are recruiting new employees. Company A has 7 employees whose annual salaries (in thousands of dollars) are

$$28, 29, 29, 30, 35, 98, 120.$$

Company B has 7 employees whose annual salaries (in thousands of dollars) are

$$26, 38, 42, 42, 44, 50, 60.$$

1. For each company, compute the mean and median salary. If you were going to work for a company, would you rather know the mean salary, or the median salary? Explain.
2. Remove the highest and lowest paid employee from each company, and then recompute the mean and median salaries from the remaining 5 employees.
3. Which measure of center is affected more by outliers? When would you want to report the mean salary? When would you want to report the median salary?

Problem: In Class Activity Find a partner, and work together to answer the following questions. Download http://tinyurl.com/JVL7PQN.

1. There are 1861 graduates represented in this population. We can use this file to take a sample of 60 randomly-selected salaries. Make sure the sample size defined in cell K2 is set to 60. The values in column H represent a particular random sample chosen from the population. (Scroll down to observe the highlighted values in column H.) The value in cell K3 is the average of the $n = 60$ randomly selected salaries. Do the following:

   Step 01: Press F9 to get a new random sample, and then write the average salary given in cell K3 in one of the green boxes. Repeat this step 45 times until you have an average written in each green box.

   Step 02: Cut the green sheet into 45 tiny boxes. Remember, each green box represent the average of a sample of size 60 from our population.

Your teacher will provide you with:

- 20 small (2" $\times$ $1\frac{1}{2}$") Post-It Notes
- Scissors
- An 11" $\times$ 17" inch poster, which you will use to build a histogram

Mac users: If F9 doesn't work, you can click on "Recalculate All" in the Formulas ribbon or turn off keyboard shortcuts.

Step 03: Glue each green box in the appropriate place on the histogram provided to you.[2]

2. What is the average of your averages? Compare this center of your distribution of the sample averages to the population average you computed in Problem 1.2. What do you notice?

3. What is the shape of the distribution of the sample averages shown in your histogram? Compare this to your answer for Problem 1.1(1)?

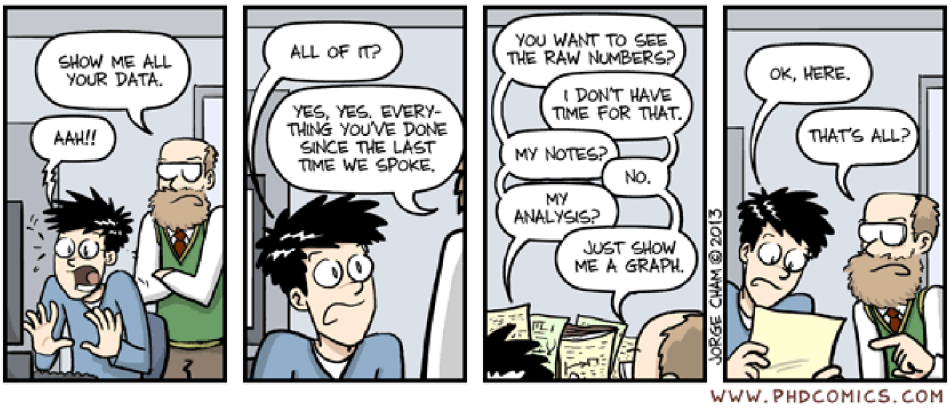
4. How does the spread of the distribution of the averages you observed compare[3] to the spread for the raw salaries you described in question 1.1(3)?

5. The author of a recent newspaper article claimed that they took a sample of $n = 60$ BYU-Idaho graduates, and the average starting salary from their sample was \$48,000. Based on your results, what is the approximate probability that a random sample of BYU-Idaho graduates would have an average starting salary that is \$48,000 or higher? Explain how you calculated your answer.

6. What does your answer to problem 5 suggest about the newspaper's result?

7. List three important "take-aways" from this activity.

"Piled Higher and Deeper" by Jorge Cham



[2]You will create a histogram. To do this, you glue the green box in the lowest unused space of the appropriate "bin" or "class". The labels on the horizontal axis give the upper and lower boundaries for each bin. If an observed value is exactly equal to a class boundary, put it in the higher bin.

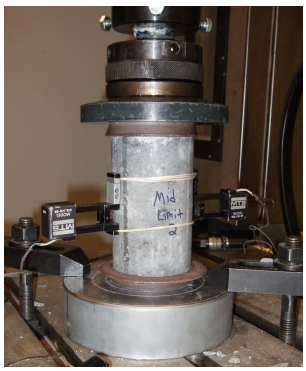
[3]Hint: Notice the range of values on the horizontal axis of your graph compared to the horizontal axis of the histogram in question 1.1.

1.2 Measures of Spread

Cylindrical Compressive Strength A team of researchers led by Dr. K. Tan measured the cylindrical compressive strength (in MPa) of $n = 11$ concrete beams [**?**]. In a separate project, a proprietary additive designed to strengthen the concrete was added to the concrete mix and $n = 10$ additional beams were poured, cured and tested. The researchers want to determine whether including the additive increases the compressive strength.

Regular ($n = 11$)	Additive ($n = 10$)
38.43	38.45
38.43	38.30
38.39	38.53
38.83	38.05
38.45	38.42
38.35	38.80
38.43	38.38
38.31	38.53
38.32	38.85
38.48	38.59
38.50	

Table 1.1: Cylindrical Compressive Strength (in MPa)



A lightweight concrete test cylinder under compression testing Source: Xb-70 at en.wikipedia [Public domain], from Wikimedia Commons

Problem 1.6 Answer the following:

1. *Visually* examine the lists of numbers in Table 1.1. Do not do any computations. Which of the two materials seems to have a higher compressive strength? How did you make this determination? Are you confident in your conclusion? Why or why not?

2. Find the mean of the cylindrical compressive strength of the observations for each of the concrete types.

3. Which of these two types of concrete has a higher compressive strength? Are you confident in your conclusion? Why or why not?

Even though there are only a few observations, it is difficult to draw a conclusion about whether this additive increases the compressive strength by scanning a list of numbers. In addition to computing the mean, we need to understand something about the spread of the data. It is helpful to create a plot.

Problem 1.7 Figure 1.1 on page 9 illustrates the cylindrical compressive strength observed with and without the additive.[4] Write a paragraph to explain your answers to the following.

1. What do you notice about the spread of the data points in the two groups?

2. What is your opinion about the effectiveness of the additive? Does it increase the strength of the concrete?

[4]The plot in this figure is called a dot plot. Dots mark the value of each observation. The marks are stacked on each other when values are repeated.

The simplest measure of the spread in a data set is the range.

Definition 1.3: Range. The range of a collection of data is defined as the maximum value minus the minimum value.

Problem 1.8 Consider again the compressive strength observed with and without the additive.

1. Find the range of the data in each of the two groups. Based on this measure, which is more spread out?
2. What is a weakness of the range? Why might it not be the best measure of the spread of a set of data?

A weakness of the median is that directly represents information from only one or two values in the data set. The mean, on the other hand, includes information about every observation in the data set.

The range only includes direct information about two observations: the maximum and the minimum. We want to find a measure of the spread in the data that includes all the observations. There are several ways to measure the spread in a data set that include all the observations. In the next few problems, you will explore a few of these.

One way to think about the spread of the data is to consider how far the data points are from the mean. Each of the arrows in Figure 1.2 on page 9 represents the distance from the mean to a particular observation in the "additive" data set.

The mean is $\bar{x} = 38.49$ MPa. We can use subtraction to determine how far each of the points is from the mean.

For example, the largest value in the "additive" data set is 38.85. If we subtract the mean from this value, we get $38.85 - 38.49 = 0.36$.
The smallest value in the data set is 38.05. Subtracting the mean from this, we get: $38.05 - 38.49 = -0.44$. We repeat this for all the observed values:

x	$x - \bar{x}$
38.45	$38.45 - 38.49 = -0.04$
38.30	$38.30 - 38.49 = -0.19$
38.53	$38.53 - 38.49 = 0.04$
38.05	$38.05 - 38.49 = -0.44$
38.42	$38.42 - 38.49 = -0.07$
38.80	$38.80 - 38.49 = 0.31$
38.38	$38.38 - 38.49 = -0.11$
38.53	$38.53 - 38.49 = 0.04$
38.85	$38.85 - 38.49 = 0.36$
38.59	$38.59 - 38.49 = 0.10$

Problem 1.9 How could you combine the numbers in this column of differences to obtain a measure of the spread? Conjecture a way to combine these values, and then compute your measure.

Problem 1.10 A diligent student suggested taking the mean of the differences from problem 1.9.

1. Compute the mean of the column of differences.

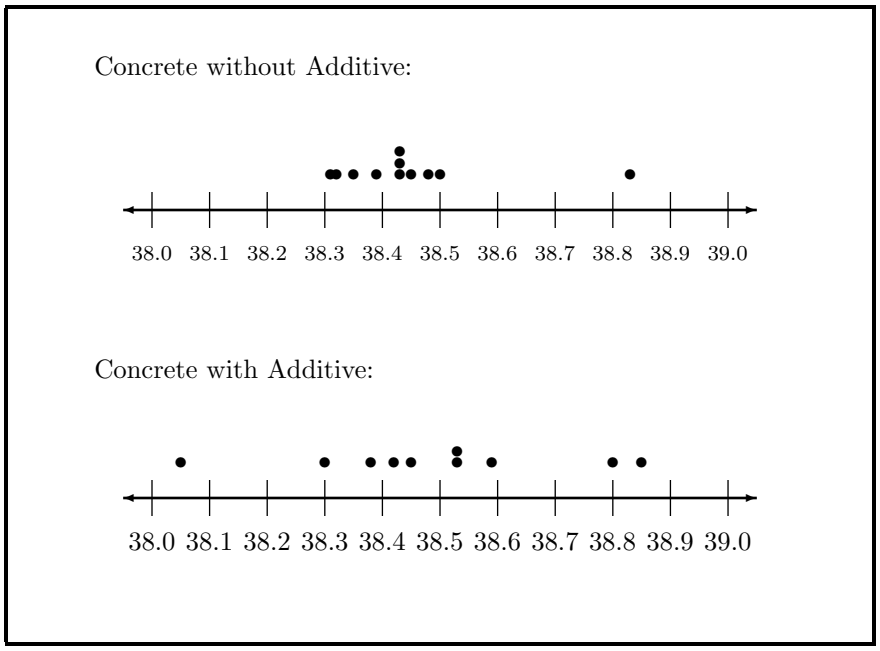


Figure 1.1: Dot plot of the cylindrical compressive strengths

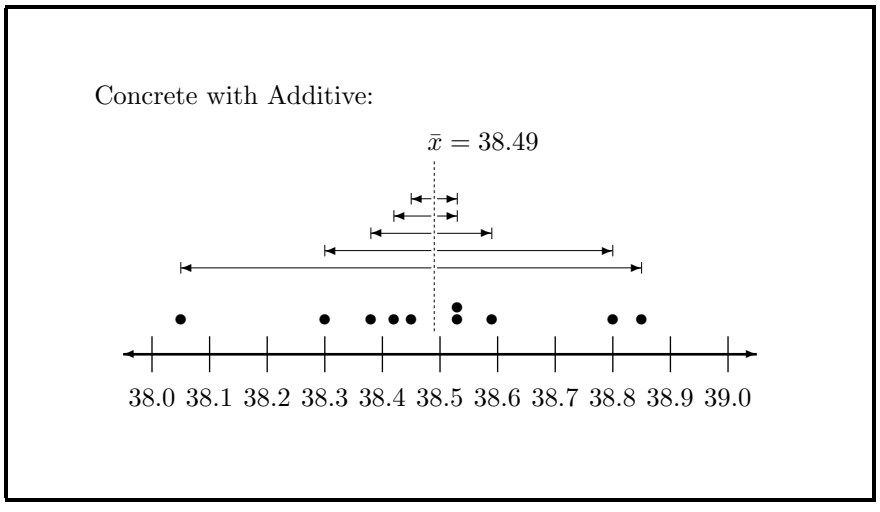


Figure 1.2: Spread of the cylindrical compressive strengths with the additive; arrows indicate the distance from the mean to each observed value

2. Is this a meaningful measure of how spread out the data are? Explain your reasoning with a sentence or two.

3. Do you think the result to the first part of this question is due to chance, or do you think you would get this result for any data set? Why?

Problem 1.11 The positive and negative differences in the previous question are canceling each other out. One way to prevent this is to make all the negative numbers positive. The absolute value function does this.

1. Compute the absolute value of each number in the column of differences.

2. We still have 10 numbers describing the spread of 10 data values. We want one number that summarizes the spread in the 10 values. Apply some operation (to the 10 numbers you computed) to get one number that summarizes the spread in the data.

Whatever statistic we choose to use to estimate the spread in a data set, we hope it will posses some optimal properties. [5] This should remind you of calculus, where we find optimum values by computing derivatives. The absolute value function is not differentiable, which means the previous method above, while it does give a way to measure the spread, is not something to which we can apply calculus. The next problem has you develop another method of computing the spread.

Problem 1.12 The function $f(x) = x^2$ is a differential function that takes any real number, either positive or negative, and return a nonnegative number. It's an alternate way to take a negative number and from it obtain a positive number.

1. Square each of the differences from problem 1.9.

2. Combine the ten values you have computed into one statistic, by computing their mean. We call this the number the variance.

3. The original data are given in units of MPa. What are the units of the statistic you just created?

4. What can you do to the variance to obtain a statistic whose units are MPa.

1.2.1 Spread in a Population

Definition 1.4: Population Variance and Standard Deviation. The **population variance**, denoted σ^2, is one measure of spread in a population. Suppose that we have a population that consists of N values, namely x_1, x_2, x_3, ..., x_N. If the population mean is denoted by μ, we can express the variance of this population as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 . \tag{1.3}$$

The population variance σ^2 is the average of the squared deviations from the mean.

The **population standard deviation**, σ, is defined as the square root of the population variance.

[5] We will not prove these optimal properties in this class. If you are interested in learning more about mathematical statistics, you should consider taking MATH 423.

Problem 1.13 Suppose the population is the mass of 300 laboratory specimens, measured in grams.

1. If we compute the variance of the population of masses, what will the units be?
2. What are the units for the standard deviation of the masses of the laboratory specimens?

Problem 1.14 W. Robert Johnston collected data on all Soviet submarine accidents that involved nuclear reactors.[**?**] His data are presented in Table 1.2 on page 11. Since all known accidents are listed here, these data represent a population. Find the population mean, variance, and standard deviation for the number of deaths in the accidents.

Problem 1.15 Continuing from the previous problem, Find the population mean, variance, and standard deviation for the number of injuries in the Soviet submarine accidents.

Date	Location	Type of Accident	Deaths	Injuries
10/13/1960	K-8 submarine, Barents Sea	reactor leak	0	3
7/4/1961	K-19 submarine, North Atlantic	reactor accident	8	31
2/12/1965	K-11 submarine, Severodvinsk, USSR	accident during refueling of naval reactor	0	7
5/24/1968	K-27 submarine, Barents Sea	naval reactor accident	9	83
1/18/1970	Sormovo, Russia, USSR	construction accident on submarine nuclear reactor	3	2
12/28/1978	K-171 submarine, Pacific Ocean	submarine reactor accident	3	0*
1979	USSR nuclear submarine, unknown location	submarine reactor accident	0	4
8/10/1985	K-431 submarine, Chazhma Bay, Vladivostok, Russia, USSR	reactor accident during refueling	10	49

* In the 1978 accident, zero injuries were reported, but there may have been some additional personnel exposed to the radiation.

Table 1.2: Soviet Nuclear Power Accidents. Use this table to answer questions 1.14 and 1.15.

Deaths	Frequency (Count)
0	3
3	2
8	1
9	1
10	1

Table 1.3: Summary of Deaths from Soviet Nuclear Power Accidents

Problem 1.16 Twenty students take a quiz. Their scores (as points) are given in the table below.

Score	Frequency (Count)
100	3
90	7
80	6
70	4

1. What is the average score? Show how you did your computations.
2. What is the variance of the scores? Do this computation by building a table of differences, squaring them, etc., rather rather than just using a formula in Excel.
3. It's possible to do this problem by creating a table with 20 rows. You can also do this problem by creating a table with only 4 rows, realizing that people with the same score will have the same difference, and hence you can just do one row and times it by the frequency of that score occurring. If you did not do the problem this way, try doing so now.

Consider the data on the number of deaths that occurred in Soviet nuclear submarine accidents. Three accidents had zero deaths. Two accidents had exactly three deaths. The remaining accidents involved 8, 9, or 10 deaths. For simplicity, we can simplify the presentation of the data by reporting the values and how many times each occurred. The data on the deaths is presented in Table 1.3 on page 11. We can use this summarized data to find the population variance of the number of deaths.

When we have a frequency table summarizing the population values, where f_i is the number of times we observed x_i, we can compute the population variance using the simplified formula

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 f_i. \tag{1.4}$$

The frequency with which a value occurs (f_i) divided by the total number of values (N) is the probability

$$p_i = \frac{f_i}{N}$$

that the value x_i will be selected at random. So if we know the probabilities p_i, then the population variance is given by the formula

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 f_i = \sum_i (x_i - \mu)^2 \frac{f_i}{N} = \sum_i (x_i - \mu)^2 p_i. \tag{1.5}$$

Many of you have seen this formula in other contexts, such as finding moments, center of mass, or computing a weighted average.

Problem 1.17 Consider the summarized data in Table 1.3 on page 11.

1. What is the probability p_i for each of the observed values of 0, 3, 8, 9, 10?
2. Give a formula for the standard deviation of a population that uses the probabilities p_i of each observed value x_i occuring (similar to equation (1.5) above), and then use your formula to compute the standard deviation of the number of deaths in the Solviet nuclear power accidents.

Figure 1.3: "Piled Higher and Deeper" by Jorge Cham

1.2.2 Spread in a Sample

Consider the data on the cylindrical compressive strength of concrete from page 7. We developed a measure for the spread of the data in Problem 1.9. We will now refine our work, if needed, to find an equation for the variance and standard deviation of a sample.

Recall that the population variance is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 .$$

How do we compute the variance when we have a sample, rather than the entire population? It seems natural to estimate σ^2 by replacing the population average μ with the sample average $\bar{x}$, and replacing the population size N with the sample size n. However, doing so produces a biased result.[6] We can get an unbiased estimator by replacing N with $n-1$, which gives the following definition.

Definition 1.5: Sample Variance and Standard Deviation. If $x_1, x_2, \ldots, x_n$ are a sample of size n from a population, then the sample variance s^2 and standard deviation s are given by the formulas

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{and} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Problem 1.18 Compute, using a table of differences (feel free to create the table in Excel), the sample variance and sample standard deviation for the cylindrical compressive strength of the concrete with the additive. Then check your answer using an Excel formula for computing standard deviation.

Problem 1.19 Write a paragraph that explains the concept of the standard deviation without using statistical jargon. Your response should make sense to someone who has never taken statistics.

[6] This issue is studied in MATH 423. If you are interested in a stronger understanding of statistical theory, consider taking this course.

Figure 1.4: "Piled Higher and Deeper" by Jorge Cham

Just as we did for populations, when we have summarized data including frequencies f_i we can compute the sample variance s^2 and sample standard deviation s with alternate formulas, similar to equations 1.4 and 1.5 on page 12.

Problem 1.20 A sample of 100 adult women was taken, and each was asked how many children she had. The results were as follows:

Taken from Navidi, *Statistics for Engineers and Scientists*, page 24.

Children	0	1	2	3	4	5
Number of Women	27	22	30	12	7	2

Please use excel to perform the computations below. Please organize your work into a table.

1. Find the sample mean number of children. (Excel can save you time.)
2. State probability p_i of selecting a woman with x_i children for each observed value x_i.
3. Give formulas for computing the sample variance and sample standard deviation that involve the probabilities p_i (similar to the formulas on page 12).
4. Compute the sample variance and standard deviation of the number of children. (Again, Excel can save you time.)
5. If a woman had 5 children, how many standard deviations above the mean is she?
6. What proportion of women were within 1 standard deviation of the mean.

1.3 Measures of Position

1.3.1 z-scores

Sometimes, we wish to compare the values of different variables directly. Sometimes this is difficult as the means and standard deviations of said variables can differ greatly.

Figure 1.5: "Piled Higher and Deeper" by Jorge Cham

Problem 1.21 The SAT is scored on a 1600 point scale (combined scores in Math and Critical Reasoning) with an average 1010[7], but the ACT is scored on a 36-point scale with a mean score of about 21 points[8]. Our goal in this problem is to determine how admissions officials can compare scores from the two tests?

Joey got a 1270 on the SAT and 29 on the ACT.

1. Without doing any computations, guess which score is better? Why?
2. How many points above the mean is the SAT score? What about the ACT score? Why does this not help you determine which score is better?
3. Suppose that you knew the standard deviation of SAT scores was 130 and the standard deviation of ACT scores is 5. How many standard deviations away from the mean is the SAT score? the ACT score?
4. Which score do you think is better? Are you more or less sure of your answer than your guess on the first part?
5. What are the units to the numbers you calculated in part 3?

Definition 1.6: z-score. A z-score for an observation is the number of standard deviations away from the mean that observation lies.

Problem 1.22: Formula for z-scores Look back at the previous problem. How did you obtain the number of standard deviations away from the mean each score was?

1. If x is an observed value, and we know the population mean μ and standard deviation σ, give a formula for the z-score for x in terms of x, μ, and σ.
2. If x is an observed value, and we know the sample mean $\bar{x}$ and standard deviation s, give a formula for the z-score for x.
3. What are the units for a z-score if the original data are presented in grams per cubic centimeter?

[7] averages obtained from collegebound.com, 2012 testing year
[8] mean obtained from www.act.org, 2012 testing year

A z-score is one of the best ways to figure out how extreme an observation may be. We'll often call an observation extreme if it lies more than 2 standard deviations away from the mean.

Problem 1.23 For the SAT, suppose $\mu = 1010$ points and $\sigma = 130$ points. For the ACT, suppose $\mu = 21$ points and $\sigma = 5$ points.

1. What scores on the SAT are considered extreme?
2. What scores on the ACT are considered extreme?
3. About 95% of people score within 2 standard deviations of the mean. What percentage of people will score above 2 standard deviations of the mean?

1.3.2 Percentiles

We've discussed measures of center and spread for a data set. We need tools to compare a data point to other values from the same distribution. For example, knowing that Earl Boykins, a player in the NBA, is only 65 inches tall isn't nearly as remarkable as knowing that he is 14 inches below the average NBA player's height in the 2007-2008 season. Knowing that your company grossed $14,000 this week is not as helpful as knowing that this is the company's best week ever.

Whenever we take a standardized test, our raw score is usually reported together with a percentile. Percentiles provide an indicator of how a student performs in comparison to their peers. The term percentile refers to $\frac{1}{100}$. The percentiles are the numbers that divide a data set into 100 equal parts.

Problem 1.24 Imagine a long road with 100 homes.

1. How many fences would be required to separate the homes? (Note: no fence is required before the first home or after the last home.)
2. Now, think of these homes as observed data values. How many percentiles are required to separate a sorted data set into 100 equal groups?
3. Is it possible to score in the 100^{th} percentile on the ACT? Justify your answer.
4. What percent of data values in a set of data have a value less than the 40^{th} percentile? What percent of data values in a set of data have a value greater than the 90^{th} percentile?

Problem 1.25 Use Excel to find the following percentiles for the BYU-IdahoGradSalaries.xlsx data.[9]

1. 5^{th} percentile
2. 25^{th} percentile
3. 50^{th} percentile

[9] You may want to search Excel's help files to find the command.

4. 75^{th} percentile
5. 90^{th} percentile
6. 98^{th} percentile

Problem 1.26 Do the following.

1. Percentiles break a data set up into 100 equally-sized collections of data. Find out what quartiles are and explain what you learned.
2. Find the quartiles for the BYU-IdahoGradSalaries.xlsx data. Compare your result to the values you found in the problem 1.25.
3. What other names have we used for the the 50^{th} percentile?

"Piled Higher and Deeper" by Jorge Cham

1.4 Graphical Displays of Data

1.4.1 Boxplots

Now that we can describe the position of data points as a percentile, let's look at one way that we can graphically display some of these values. The simplest such graph is a boxplot, where we use the min, first quartile, median, third quartile, and maximum, to create a plot.

The following **boxplot**[10] is a replica of a graph created by the quality department in a factory that fills ketchup bottles. The quality engineer was investigating the horizontal offset of the label from its desired location. The data are given in millimeters.

Problem 1.27 Use figure 1.6 to answer the following

1. In figure 1.6, there are five important numbers: -7, -3, 0, 2, and 3 mm. What do each of these values represent?[11]
2. What is the range of these data?

[10] This graph is sometimes called a box-and-whisker plot or box-and-whisker diagram.
[11] Hint: These numbers are typically called the **five-number summary**.

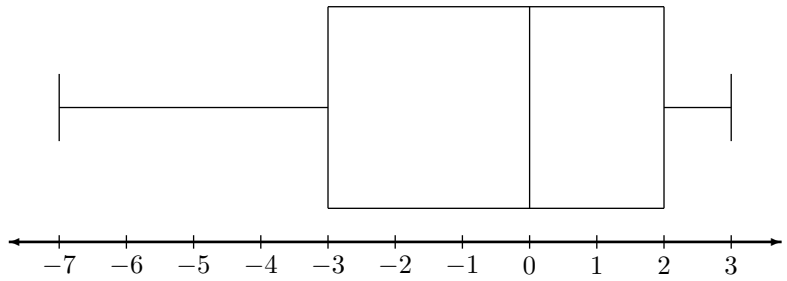


Figure 1.6: Horizontal offset of the label placement in a ketchup filling operation

3. Define the **interquartile range**, and then state the interquartile range of these data?
4. What percentage of the data lie between 2 mm and 3 mm?
5. Based on the boxplot in figure 1.6, can you state the value of the mean of these data?

One of the quick uses of a box plot is that they allow you to quickly make side by side comparisons of two groups.

Problem 1.28 Open the BYU-IdahoGradSalaries.xlsx data. Sort the data by Column B, "College." Separate the data into two groups:

Group 1: the College of Physical Sciences and Engineering,[12] and

Group 2: the remaining graduates (combined).

1. Find the five-number summary for each of the two groups.
2. Create side-by-side box plots for these two groups (by hand).[13]
3. What do you conclude about the salaries in the College of Physical Sciences and Engineering compared to the rest of the campus?

1.4.2 Histograms

We've already looked at histograms and seen how they work. Now it is your turn to create a histogram from a data set. One of the downsides of using excel is that there is no built in tool for creating histograms. I'll provide you with a histogram creator later in the semester, which is built of the ideas of the next problem.

Problem 1.29 Open the BYU-IdahoGradSalaries.xlsx data. Use that data to complete the histogram in Figure 1.7. Each bin has a width of $25,000. Please state exactly how tall each bin should be. The first bin has been done for you, and it's height is 270 students.

You will want to use the Excel command `=countif()` to find the appropriate heights of each bin.

Note: If you would like additional information on the content of this chapter, you may want to read Chapter 1 in Navidi, *Statistics for Engineers and Scientists*, third edition.

[12] This college is listed as "Physical Sci & Eng" in the data file.

[13] To create a side-by-side box plot, draw both box plots above the same number line. One box plot should be above the other.

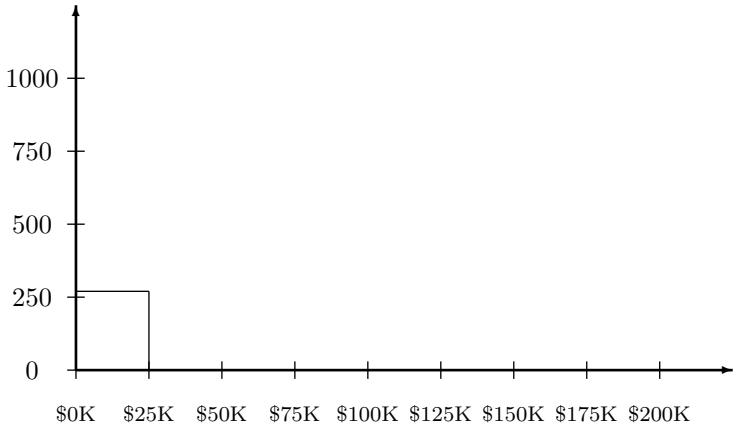


Figure 1.7: Histogram of BYU-Idaho graduate starting salaries

Chapter 2

Probability

In most applications, we need to know how a manufacturing process will behave before we begin production. For example, what will be the mean and standard deviation of a critical product characteristic? When designing a new part, we recognize that variation will occur during the manufacturing process. We can use the ideas in this chapter to better understand how that variation will affect the parts we actually produce.

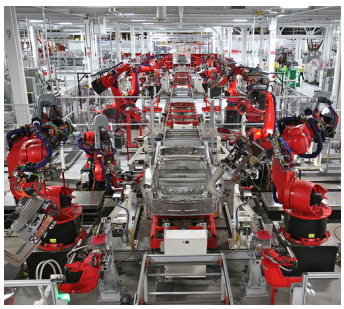


Photo by Steve Jurvetson

Source: flickr.com/photos/jurvetson/7408451314

Definition 2.1. A **random variable** is a variable that generates real numbers.

The value rolled on a die is a random variable. It is usually not possible to predict the next value generated by a random variable, but if many values of the random variable are examined, a pattern becomes apparent.

One important characteristics of a random variables depends on the type of output that we can obtain.

Definition 2.2: Discrete versus Continuous. When a random variable assumes a value from a discrete set of numbers, we say it is a **discrete** random variable. For discrete random variables, the number of possible outcomes can be counted. If a random variable takes a value from a continuous range of values, we say that it is a **continuous** random variable.

As a basic rule of thumb, we can think of continuous random variables as being measured (as opposed to counted). The next problem has you classify several random variables as either discrete or continuous.

Problem 2.1 Classify the following random variables as either discrete or continuous. Give an example of the kind of value that you could obtain from this random variable. The first has been done for you.

1. The current air temperature in Fahrenheit. [Ans: This is a continuous random variable. Possible values could be 78.4. We measure temperature, rather than count it.]
2. Your height in inches.
3. The number of students who will attend our next class meeting.

4. The total weight of all coins in your possession in grams.
5. The monetary value of the coins in your possession.
6. The sum of the digits in your best friend's cell phone number.
7. The duration of your last conversation with your best friend.
8. The number of text messages you receive in a day.
9. The number of times you smiled yesterday.
10. The total number of credits you have completed at BYU-Idaho.
11. The amount of time that has elapsed since you blinked last.

Very often, we find the need to account for processes which include some degree of uncertainty. One very obvious instance is when flipping a coin or rolling a die, but uncertainty applies to many manufacturing processes and signal processing applications. Knowing how likely your factory is to produce faulty parts is a top priority for many quality control personnel, and data sent over wireless signals is only usable if the signal encounters few or no digital errors (confusing a one for a zero or vice versa). While some of the following questions may seem a little silly, almost all of statistics is built on the foundation of probability. Knowing how to recognize and use the three basic laws of probability is somewhat akin to mastering addition and subtraction before tackling calculus: somewhat tedious, but it prevents silly errors.

Photo by Tomi Tapio
Source: flickr.com/photos/tomitapio

Before we dive into the content of probability, Let's make sure we all use the same definition of probability. The probability of an event will be understood to mean the number of ways that event can happen divided by the total number of equally likely outcomes possible. To illustrate, when rolling a fair twelve-sided die, The probability of rolling a number divisible by four is $\frac{3}{12} = 0.25$ because out of the twelve possible values one can roll, only three (4, 8, 12) are divisible by four. Note that one convenient property of this definition of probability is that if you multiply the probability by 100, the result is a percentage that behaves how we would expect. Let's examine some of the governing rules of probability.

Problem 2.2 A die was rolled 6000 times, with the the following table to summarize the results.

Value (x)	Frequency (f)
1	1036
2	968
3	961
4	1006
5	?
6	1038
Total	6000

1. What is the value of the unknown quantity? How did you find it?
2. What proportion of the rolls resulted in an even number? What proportion of the rolls yielded an odd number?

3. Theoretically, what is the probability of getting either an even number or an odd number when a die is rolled? How does this compare with your answers to the previous question?

We've previously categorized variables as either **discrete** or **continuous**. (See page 20). The previous question dealt with a discrete variable. Now, we'll look at a continuous variable.

Problem 2.3 A producer of consumer electronics claims that the cord length for a pair of headphones is between 4 and 6 feet. The manufacturer takes a random sample of 500 pairs of headphones and finds that because of the rapid feed rate of some machines in the factory, 2% of all headphone cords are actually too long, and 97% of cords have appropriate length.

1. What is the probability of a cord length being at least four feet long?
2. What is the probability of a cord length being either too short, too long, or an appropriate length? This is one of those questions that might seem a little silly.
3. Given your answers to the two previous questions, what is the probability that a randomly selected cord will have a length of less than four feet?

Problem 2.4 Suppose that we are given a random variable with n different possible values, namely $x_1, x_2, \ldots, x_n$, and each value x_i is equally likely to occur.

1. What is the probability that one of these values will occur? In other words, what is $P(x = x_i)$?
2. If we were to sum all the probabilities together, which we could write as $\sum_{i=1}^{n} P(x = x_i)$), then what value would we get. Please show how you obtained your answer.

The previous question and its accompanying answer are very important! The probability of all possible outcomes together add to one, regardless of the specific application. This is called the **Law of Total Probability**.

Theorem (Law of Total Probability)**.** *The probability of all possible outcomes together add to one, regardless of the specific application.*

Problem 2.5 In a factory that produces exercise equipment, it is known that about one-third of the products are defective.

1. What is the probability that a randomly-selected product will not be defective?
2. If the probability of getting a particular outcome (let's call it x) is p, what is the probability that x will **not** occur? Use the law of total probability to explain your answer.

   This is one of the most useful applications of the law of total probability and is called the **Complement Rule**.

Calculating probabilities is fairly simple, but if you don't keep track of your process, you could count some events twice!

Problem 2.6 Think of a 20-sided die with its faces numbered 1-20 used by players of trading card games or role-playing games.

1. What is the probability that a random roll of the die yields a number greater than 15?
2. What is the probability that a random roll of the die yields an even number?
3. What is the probability that a random roll of the die yields an even number or a number greater than 15? List out all possible rolls that give this result.[1]

Problem 2.7 In a factory producing aluminum rods, a sample of 1000 rods was taken and the length and diameter of each rod was measured. The following table summarizes critical information about the measurements:

	Diameter		
Length	Too Thin	Correct	Too Thick
Too Short	10	3	5
Correct	38	900	4
Too Long	2	25	13

Compute the following probabilities, making sure you show how you obtained your solution. Use sentences to explain what numbers you combined.

1. What is the probability that a randomly chosen rod is too thin?
2. What is the probability that a randomly chosen rod is too thick?
3. What is the probability that a randomly selected rod is either too thick or too long?
4. What is the probability that a rod chosen at random is either too thin or too short?

Problem 2.8 Using the table from the previous question, what is the probability that a randomly selected rod is out of spec (not Correct in Either Length or Diameter)? Do this in two ways.

1. Count the total number of rods that are out of spec, and then divide by the total number of rods.
2. Use the Complement Rule, as we know how many rods are in spec.

Ask me in class to show you how you can tackle the previous problems using Venn Diagrams, if you are interested.

[1] Notice that some numbers could be counted twice, if they are both even and greater than 15.

2.1 Probability Density Functions

As we've seen, calculating simple probabilities is sometimes quite easy, and once you master the rules that build a framework for these calculations, there is really quite a bit you can do with them. Many times, calculating the probability of an event is a bit more complicated. How does one calculate the probability of a work-related accident happening during any given 30-minute period of time at a factory open 24 hours a day? How can we find the probability of rolling a seven with a loaded pair of dice? What about finding the probability that a given axle will be able to withstand the force associated with a full vehicle dropping from a height of several feet? These questions and more require that we look at the functions that describe the probability of events occurring.

As is often the case, we'll look at a simple example to build one of these probability functions and try to generalize it as much as possible. The board game *The Game of Life* features a wheel (shown at the right) with areas numbered 1-10 which each player spins at the beginning of his or her turn to determine the number of spaces that the player moves forward. We'll assume that every player has an equally likely chance of spinning any number on each turn.

Problem 2.9 Start by reading the previous paragraph. Then answer the following questions.

1. What is the probability that a player will spin a four on their turn? What is the probability of spinning a seven?

2. Consider the types of function that could describe the set of events and probabilities in the previous problem. We could define such a function in a piecewise manner since we know the probability and value of each event.

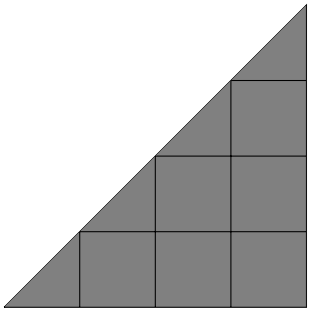
   Fill out the following table to create a function that describes the values and probabilities associated with spinning the wheel in *The Game of Life*.

Value x	*Probability* $f(x)$
1	?
2	?
3	?
4	?
5	?
6	?
7	?
8	?
9	?
10	?

3. Draw a graph of this function. This should be done by labeling the x-axis with the different values that the wheel can take and making the height of the function equal to the probability of the wheel giving that number on a random roll.

The function you developed in the previous problem is used to summarize the possible values of an event and their corresponding probability. For brevity, such functions are denoted by the letter f. We call them probability density functions. Let's create a probability density function now for a problem that involves infinitely many possible outcomes.

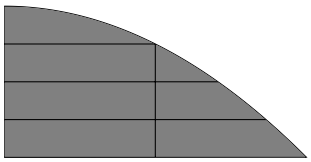
Problem 2.10 Suppose we cut out a triangular sheet of paper. The paper is 4 inches wide, 4 inches tall, and we place the paper so that it lines up nicely with the x-axis, as shown to the right.



1. If you randomly choose a point on the triangle, are you more likely to pick a point (x, y) with $x < 2$ or with $x > 2$? Explain.
2. What proportion of the triangle's area lies to the left of $x = 2$? You'll need to compute the area of the triangle that is left of $x = 2$ and compute the total area of the triangle.
3. What proportion of the trianlge's area lies between $x = 2$ and $x = 3$?
4. What is the probability of randomly picking a point (x, y) in the triangle so that $1 \leq x \leq 3$?

We defined probability already as the proportion of times an event will occur. When there are an infinitely number of possible outcomes, we can use area and integration to compute this probability.

Problem 2.11 Suppose that a manufacturer creates a metal plate shaped in the form of a half parabola. The plate occupies a region in the xy plane that lies below the curve $g(x) = 4 - x^2$, above the x-axis, for $0 \leq x \leq 2$, as shown to the right. We'll be considering the random variable X which we obtain by randomly picking a point (x, y) on the metal plate and then recording the x-value.



1. Start by computing the total area for $0 \leq x \leq 2$.
2. What proportion of the area lies to the left of $x = 1$? Use your answer to state the probability of randomly selecting a point (x, y) that lies to the left of $x = 1$ (we'd write $P(x < 1)$), and then state the probability $P(x > 1)$.
3. What is the probability of randomly selecting a point with $x = 1$ exactly, so compute $P(x = 1)$.
4. Compute $P(.5 < x < 1.5)$.
5. What probability do you obtain by computing the integral formula
$$\frac{\int_{.5}^{1.75} g(x)dx}{\int_0^2 g(x)dx}?$$
6. What formula would you use to compute the probability of selecting a point whose x value is between a and b?

In the previous two problems, we started by computing the total area, and then every probability computation afterwards required that we compute an integral, and then divide by this area. We can summarize this by noting that

$$P(a \leq x \leq b) = \frac{1}{A}\int_a^b g(x)dx.$$

If we times the function g by a scalar c, so that the total area equals 1, then we don't have to worry about dividing by the area, as then $A = 1$.

Definition 2.3: Normalizing Constant. Suppose that $g(x)$ is a nonnegative function. A normalizing constant is the constant c that we must multiply by $g(x)$ so that the area under g will equal 1. The resulting function $f(x) = cg(x)$ is called a normalized function.

Problem 2.12 Let's practice with finding normalizing constants and normalized functions.

1. In problem 2.10, a function to describe the top of the triangle is $g(x) = x$ for $0 \leq x \leq 4$. Find the normalizing constant c, i.e. find a constant c so that $\int_0^4 cxdx = 1$.
2. In problem 2.11, the function $g(x) = 4 - x^2$ for $0 \leq x \leq 2$ described the top of the metal plate. Find the normalizing constant c so that the scaled function $f(x) = cg(x)$ when integrated over the bounds $0 \leq x \leq 2$ will be 1, i.e. we want $\int_0^2 f(x)dx = 1$.
3. Find a normalizing constant for $g(x) = e^{-2x}$ for $0 \leq x \leq 10$, and the corresponding normalized function $f(x)$.

If a function $g(x)$ is already normalized, so $\int_a^b g(x)dx = 1$, then we can use the function to rapidly compute probabilities. Feel free to use Wolfram Alpha or a calculator to perform any needed integrals.

Problem 2.13 Consider the function $g(x) = e^{-.2x}$ for $0 \leq x \leq \infty$.

1. Consider the region in the xy-plane that lies below $g(x)$ and above the x-axis for $0 \leq x \leq \infty$. What is the probability of randomly selecting a point (x, y) so that $0 \leq x \leq 10$? Compute the total area under $g(x)$ for $0 \leq x \leq \infty$, and the the area under $g(x)$ for $0 \leq x \leq 10$ to obtain your answer.
2. Show that $g(x)$ is not a normalized function, and then compute the normalizing constant c and the state the normalized function $f(x)$.
3. Repeat the first part of this problem, but now use the function $f(x)$ instead to perform your computations.
4. Use the normalized function $f(x)$ to compute the probability that $5 \leq x \leq \infty$.
5. Use your answer from the previous part to state $P(0 \leq x \leq 5)$.

Beware that not every function can be normalized and then used to compute probabilities. There's an important part of the definition of a normalizing constant that many students forget to check.

Problem 2.14 Consider the function $g(x) = x^2 - 1$ for $0 \leq x \leq 3$.

1. Find a constant c so that $\int_0^3 cg(x)dx = 1$. Let $f(x) = cg(x)$.
2. Compute the integral $\int_0^1 f(x)dx$. Explain why this cannot be the probability of randomly selecting a point with $0 \leq x \leq 1$.
3. Draw a graph of $g(x)$. Explain, without doing any integrals, why you know this function cannot be used to compute probabilities by merely integrating.

Definition 2.4: Probability Density Function. A nonnegative normalized function $f(x)$ is called a probability density function (pdf). In other words, we say that $f(x)$ is a probability density function (over some region R) if

1. $f(x) \geq 0$
2. $\int_R f(x)dx = 1$, where the bounds of the integral depend on R.

Problem 2.15 Which functions below are probability density functions? If it is not a probability density function, find a normalizing constant so that $cf(x)$ is a pdf, or explain why this is not possible. Feel free to use technology to perform any integrals.

1. $$f(x) = \begin{cases} 0.2e^{-.2x}, & 0 < x < 25 \\ 0, & \text{otherwise} \end{cases}$$
2. $$f(x) = \begin{cases} .25(1 - (\frac{x}{5})^4), & 0 < x < 5 \\ 0 & \text{otherwise} \end{cases}$$
3. $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{4}}, -\infty < x < \infty$$
4. $$f(x) = \begin{cases} x/54, & 6 < x < 12 \\ 0, & \text{otherwise} \end{cases}$$

2.2 Cumulative Distribution Functions

Now that we know how probability functions work- what they look like and a few of their simple properties, lets talk about a similar function, which we'll refer to as the summative probability function, or cumulative distribution function. Recall that probability functions give the probability of a specific event (or value) occurring, the summative probability function gives the probability of the variable obtaining any value less than or equal to the specified value, and is usually denoted by the letter $F(x)$. To find such a probability, we just add up all the probabilities of all the values your variable can take on that are less than or equal to the value of interest. If the variable takes on finitely many values, then we just sum up finitely many things. If the variable takes on infinitely many values, we need integrals.

Definition 2.5: Cumulative Distribution Function CDF. Suppose that a random variable X has probabilty function $f(x)$. The summative probability function, or cumulative distribution function, is the function

$$F(x) = P(X \leq x).$$

- If $f(x)$ is a probability function for a discrete random variable, then we just sum the finitely many probabilities to obtain

$$F(x) = P(X \leq x) = \sum_{x_i < \leq x} f(x).$$

- If $f(x)$ is a probability density function defined on $a \leq x \leq b$, then we can quickly compute

$$F(x) = P(X \leq x) = \int_a^x f(x)dx.$$

Problem 2.16 Recall the spinning wheel from *The Game of Life*. Let's create a cumulative distribution function for spinning the wheel.

1. What is the probability of rolling a number less than or equal to 4 on a random spin?
2. What is the probability of rolling a number less than or equal to 6 on a random spin?
3. Complete the following table of values.

Value x	0	1	2	3	4	5	6	7	8	9	10	11
pdf $f(x) = P(X = x)$	0	1/10	1/10	1/10								
cdf $F(x) = P(X \leq x)$	0	1/10	2/10									

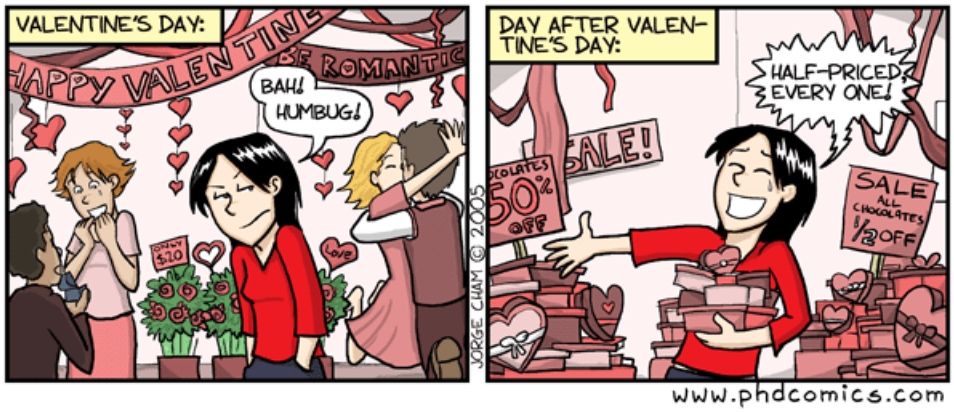
4. What is the probability of rolling a number greater than 6?

Just as we found values of the summative function for certain values on the wheel used in *The Game of Life*, we can find values similarly with functions involving a continuous variable. Remember that all we're doing is adding up all the probabilities of all values less than or equal to the value of interest. How do we usually find the sum of all the values of a function over some interval? As you may have guessed, integration is a handy tool to more readily calculate values of the summative probability function.

Problem 2.17 Consider again the parabolic metal plate from problem 2.11. The pdf for this function is $f(x) = \frac{3}{16}(4 - x^2)$ for $0 \leq x \leq 2$.

1. What's the probability of randomly choosing a point (X, Y) with $X \leq .5$?
2. What is the probability of randomly choosing a point (X, Y) with $0 \leq X \leq x$? In other words, compute the cumulative distribution function $F(x) = \int_0^x \frac{3}{16}(4 - x^2)dx$.
3. What is $F(0)$? What is $F(2)$? Construct a graph of $F(x)$.

"Piled Higher and Deeper" by Jorge Cham



Problem 2.18 A producer of oil filters employs a worker to dispose of visibly dented filters before they are painted and packaged to reduce overhead costs. The amount of time in seconds, x, that elapses between occurrences of such filters can be described with the function

$$f(x) = \begin{cases} 0.2e^{-.2x}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}.$$

1. What is the probability that a randomly selected wait time between two visibly dented filters is between 3 and 15 seconds?
2. What is the probability that a randomly selected wait time between two visibly dented filter is over 8 seconds?
3. What is the probability that a randomly selected wait time between two visibly dented filters is exactly 7 seconds?

Problem 2.19 The length of a 8-foot long piece of lumber sold at a lumber yard varies somewhat from one board to the next. The pdf of the length, X in inches, of a randomly selected board is given by the following function:

$$f(x) = \begin{cases} \dfrac{3(x-90)^3(x-100)^2}{50000}, & 90 < x < 100 \\ 0, & \text{otherwise} \end{cases}$$

Note that an 8-foot board will have a length of 96 inches.

1. What is the probability that a board chosen at random is shorter than the advertised 8 feet?
2. What is the probability that a board chosen at random is within 2 inches of the advertised length of 8 feet?

Note: If you would like additional information on the content of this chapter, you may want to read Chapter 2 in Navidi, *Statistics for Engineers and Scientists*, third edition.

2.3 Mean, Variance, and Standard Deviation of a Discrete Random Variable

2.3.1 Discrete Random Variables

To help us understand more complicated situations, we start with something familiar: rolling a die.[2] The principles that apply to dice generalize to many applications.

Problem 2.20 Imagine yourself rolling a die many, many times. After each roll, suppose you computed the mean of all the values you had observed up to that point.

1. After many, many rolls, what value will the mean approach? Make a guess before doing any computations.

2. A die was rolled 6000 times. The following table summarizes the results.

Value (x)	Frequency (Count)
1	1047
2	968
3	961
4	1006
5	980
6	1038
Total	6000

   Compute the mean, variance, and standard deviation of this data. Feel free to use Excel.

3. Let X be a random variable in which we roll a die and record the value shown. Rather than rolling the die 6000 times, we can theoretically tackle this problem by considering the following probability table.

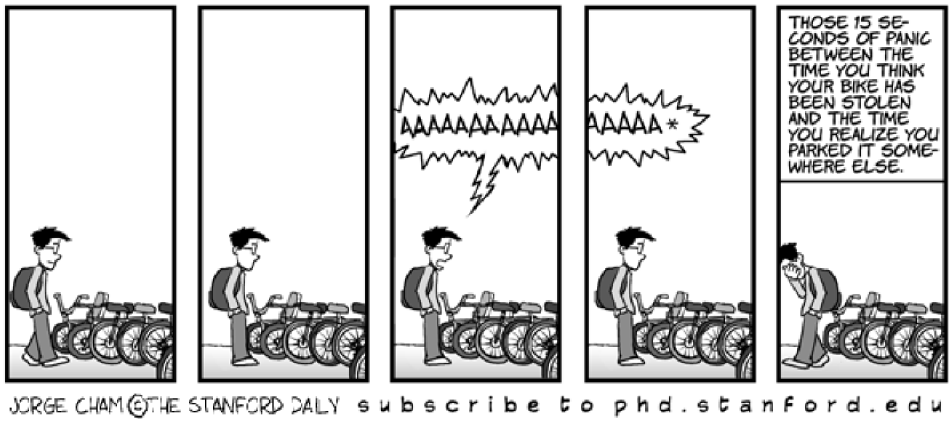
Value (x)	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

   Compute the theoretical mean, variance, and standard deviation of the random variable X.

   We often call the mean of a random variable the expected value and write $E[X]$ or μ_X.

4. A student claimed that "the mean of a random variable must be a value that the random variable can attain." Comment on the accuracy of this statement. And then make a guess for the the mean (long-run average) of a fair 10-sided die?

"Piled Higher and Deeper" by Jorge Cham



Problem 2.21 In Marilyn vos Savant's October 16, 2011, Parade Magazine column, *Ask Marilyn*, Andy G. posed the following question:

> As children, my siblings and I often settled a disagreement with a game called Evens and Odds. In this game, one side is assigned evens and the other is assigned odds. Then, on the count of three, a representative of each side reveals a number of fingers from zero to five. If the sum of the two numbers is even, the evens win; if the sum is odd, the odds win. Is this method fair? Or do the evens have an advantage?
> –Andy G., Cedar Hill, MO

We assume that each person is equally likely to show 0, 1, 2, 3, 4, or 5 fingers. The table below gives the probability for each of the 11 possible outcomes.

1. What is the probability that the sum of the fingers shown will equal 5?
2. Which is more likely, evens or odds?
3. What is the mean of this probability distribution?
4. What is the variance and standard deviation?

Total Number of Fingers x	Probability $P(X = x)$
0	1/36
1	2/36
2	3/36
3	4/36
4	5/36
5	?
6	5/36
7	4/36
8	3/36
9	2/36
10	1/36

---

[2]Whenever we simply refer to a "die", we always mean a fair, six-sided die. If the die has a different number of sides or the sides do not appear with equal probability, this will be specified.

Problem 2.22 The following question was posed in Marilyn vos Savant's December 27, 1998, Parade Magazine column, *Ask Marilyn*:

> At a monthly 'casino night,' there is a game called Chuck-a-Luck: Three dice are rolled in a wire cage. You place a bet on any number from 1 to 6. If any one of the three dice comes up with your number, you win the amount of your bet. (You also get your original stake back.) If more than one die comes up with your number, you win the amount of your bet for each match. For example, if you had a $1 bet on number 5, and each of the dice came up with 5, you would win $3. It appears that the odds of winning are 1 in 6 for each of the three dice, for a total of 3 out of 6 – or 50%. Adding the possibility of having more than one die come up with your number, the odds would seem to be in the gambler's favor. What are the odds of winning this game? I can't believe that a casino game would favor the gambler.
> –Daniel Reisman, of Niverville, NY

Unfortunately for Daniel, he figured the probabilities incorrectly. There are four possible outcomes,[3] representing the number of dice that match your number. Let X represent the profit that a gambler gets from a $1 bet when playing Chuck-A-Luck. The table below summarizes the probabilities of earning a profit of x dollars from a $1 bet.

How many dice match the chosen number	Profit x	Probability $P(X = x)$
0 dice	$−1	125/216
1 dice	$1	75/216
2 dice	$2	15/216
3 dice	$3	1/216

1. Determine the (long-run) mean of the profit for a gambler who plays Chuck-A-Luck.
2. Compute the variance and standard deviation of the profit.

Problem 2.23 A pyramid-shaped die has four sides. The numbers corresponding to each of these sides do not have to be 1, 2, 3, and 4. For one such die, the numbers on the die are: x_1, x_2, x_3, and x_4. (Notice that we are not specifying their values.) The probabilities associated with each of these values are p_1, p_2, p_3, and p_4, respectively. This information is summarized in this table on the right.

Value (x)	Probability
x_1	p_1
x_2	p_2
x_3	p_3
x_4	p_4
Total	1

1. Use this information to give equations for the mean and variance of the values that will be rolled on this die.
2. Generalize your results to give an equation for the mean and variance of the values rolled on a weighted die with m sides, where m is some positive integer. The values shown on the die can be expressed as x_1, x_2, x_3, ..., x_m. The corresponding probabilities can be represented as p_1, p_2, p_3, ..., p_m.

---

[3] The possible outcomes are: 0, 1, 2, or 3.

"Piled Higher and Deeper" by Jorge Cham

2.4 Continuous Random Variables

Now that we can compute the mean and variance for discrete random variable, where the number of possible outcomes is countable, we can rapidly extend this to continuous random variables, which can assume any real number in a range of values? If we replace the summation symbol with an integral, and the probability p_i with the pdf $f(x)$, then we obtain the following.

Definition 2.6: Mean, Variance, and Standard Deviation for Continuous Random Variables. Suppose X is a continuous random variable with probability density function $f(x)$, $a \leq x \leq b$. The mean and variance of X are

$$\mu_X = E(X) = \int_a^b x f(x) dx \qquad \text{and} \qquad \sigma_X^2 = \text{Var}(X) = \int_a^b (x - \mu_X)^2 f(x) dx.$$

The standard deviation of X is, as always, the square root of the variance. Another name for the mean μ_X is the expected value of X, or $E(X)$.

One of the simplest continuous distributions is one in which any real number, between any two, is equally likely to occur. The graph of this distribution is very similar to the graph of the distribution for the Game of Life. We call this distribution the uniform distribution. A beam of wood with a constant cross sectional area is called a uniform beam.

Problem 2.24 Imagine we have a beam of wood that is 8 feet long. We randomly pick a length between 0 and 8 ft (our random variable X is this length), and then cut the beam to this length. Suppose that every length between 0 and 8 feet is equally likely.

1. We would like to obtain a probability density function for the random variable X. Since each length is equally likely, a graph of the pdf should be a horizontal line. State such a function $f(x)$, and make sure to normalize it.

2. Compute μ_X, σ_X^2, and σ_X. Please show you integration steps.

Problem 2.25: Uniform Distribution Suppose that X is a random variable that takes on any value in the range $a \leq x \leq b$ with equally likely probability. We call X a uniform random variable.

1. We know that the pdf of X is a function of the form $f(x) = c$ for some constant c. Find the constant c, in terms of a and b, so that $f(x)$ is the pdf.

2. Show that the expected value is $\frac{a+b}{2}$, by computing the integral $\int_a^b x f(x)dx$ by hand. Show your integration steps.

3. Compute the variance of X. Again, please show your integration steps.

Definition 2.7: Uniform Distribution. A uniform random variable X is a random variable that takes on values between a and b with equal probability, which we'll write using $X \sim U(a, b)$.

Wikipedia can be an excellent place to look for information about different types of distributions. For example, the page

- http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)

gives all the details about the uniform distribution that we'll be using throughout the semester. Each time you see a new distribution, feel free to use Wikipedia to check your work.

Problem 2.26 We now know that the pdf for a uniform random variable X with $a \leq x \leq b$ is given by the function $f(x) = \frac{1}{b-a}$ provided $a \leq x \leq b$. If we pick a value outside this range, we'll just assume that $f(x) = 0$. We can combine this together using piecewise function notation to write

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}.$$

In this problem, you'll give a formula for the cdf (cumulative distribution function).

1. What is the probability of picking a number less than or equal to x where $x < a$? We'll write this as $P(X \leq a)$.

2. Compute $P(X \leq x)$ where $a \leq x \leq b$. You'll need to compute $\int_a^x f(x)dx$.

3. Compute $P(X \leq x)$ where $x > b$.

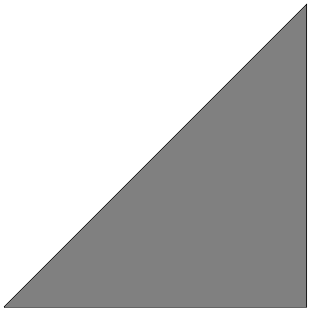
4. Combine your previous three answers, using piecewise function notation, to give a formula for the cdf of X. You can check your work with wikipedia.

The uniform distribution is perhaps the simplest distribution to work with, and rises naturally as an extension of rolling die. Let's explore another distribution, namely the triangular distribution.

Problem 2.27 Recall the triangular metal plate from problem 2.11. In this problem, we obtain the pdf $f(x) = \frac{x}{8}$ where the triangle place had corners $(0, 0)$, $(4, 4)$, and $(4, 0)$. Let's change our rectangular plate so that the width and height are now a. We randomly pick a point (x, y) from the triangle and let the random variable X be the x coordinate of that point.

1. Show that the pdf for X is $f(x) = 2x/a^2$.

2. Compute the expected value of X, i.e. the mean μ_X. Show what integral you computed, and then give its value.



3. Compute the variance of X. Show what integral you computed, and give its value. Follow this pattern always to present in class.

Problem 2.28 Continuing from the previous problem, complete the following: (Note: if you are having a hard time doing the computations because the variable a appears in the problem, then let $a = 7$ and perform the computations. Then go back through and change every 7 you see to a, and every 49 to a^2, etc.)

1. Compute $P(X \leq x)$ where $0 < x < a$. In other words, give the cdf $F(x)$ for the triangular distribution.
2. Compute $F(a/3)$ and $F(2a/3)$.
3. Use your answer from the previous part so state $P(a/3 < X \leq 2a/3)$.
4. Give a formula for $P(c < X \leq d)$ if you know $F(c)$ and $F(d)$.

We'll look at two more continuous distributions before we close this chapter, namely the normal distribution (the bell curve) and the exponential distribution.

The normal distribution is a continuous distribution defined for all real numbers x (so $-\infty < x < \infty$) whose probability density function is given by

$$f(x) = \frac{1}{b\sqrt{2\pi}} e^{-(x-a)^2/2b^2}$$

for two constants a and b. We'll quickly see that these constants are directly related to the mean and variance.

Problem 2.29 Consider the function $f(x) = \frac{1}{b\sqrt{2\pi}} e^{-(x-a)^2/2b^2}$.

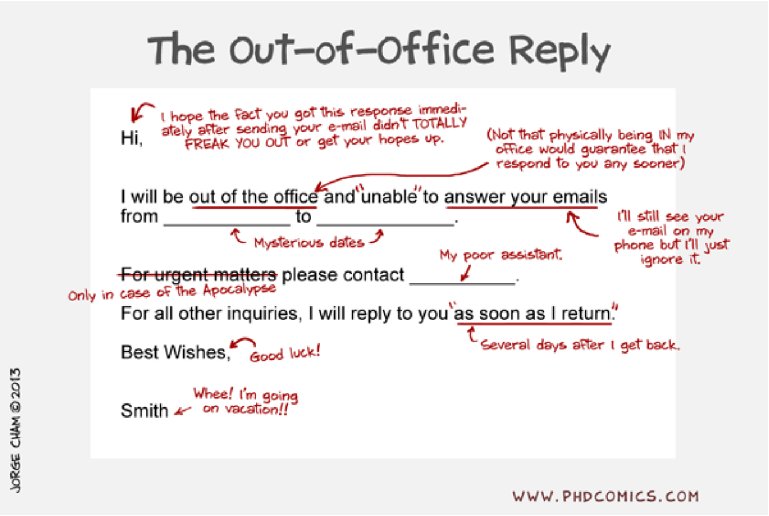
1. Let $a = 3$ and $b = 5$. Construct a plot of $f(x)$ for $-10 \leq x \leq 10$.
2. Repeat the previous part with several different values of a. Try values of a that are both positive and negative. What happens to the graph when you change a?
3. Let $a = 0$, and repeat part 1 with several values of b. Try values of b that are both positive and negative. What happens to the graph when you change b?

If you obtained a bell shaped graph in the previous problem, then you're doing it absolutely correct. If not, then double check your parenthesis and try it again.

Problem 2.30 Consider the function $f(x) = \frac{1}{b\sqrt{2\pi}} e^{-(x-a)^2/2b^2}$.

1. Compute the derivative of $f(x)$ and find the critical values (the x-values where $f'(x) = 0$). This tells you the x coordinate of the maximum of your graph.
2. Compute the second derivative of $f(x)$ and find the x-coordinates of the points of inflection (the x-values where $f''(x) = 0$).
3. Let $a = 100$ and $b = 15$. Sketch by hand, using what you now know about maximums and points of inflection, the graph of f using these values.

"Piled Higher and Deeper" by Jorge Cham



It's time to compute the expected value and variance of a normal random variable X.

Problem 2.31 Consider the function $f(x) = \dfrac{1}{b\sqrt{2\pi}} e^{-(x-a)^2/2b^2}$.

1. Use a computer algebra system (Wolfram Alpha is fine) to show that $f(x)$ for $-\infty < x < \infty$ is a pdf for a random variable X. [4]
2. Compute, by hand, the expected value of X. Do this integral by hand. See the footnotes below for a hint.[5]
3. Use a computer algebra system to compute the variance of X.

Now that we know the connection between a and b and the mean and variance, let's formally define a normal distribution.

Definition 2.8: Normal Distribution. A random variable X follows a normal distribution, and we write $X \sim N(\mu, \sigma)$, when the pdf of X is $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$. The mean and standard deviation of X are μ and σ.

The pdf for a normally distributed random variable cannot be evaluated in closed form using transcendental functions[6] that you have studied in previous math classes. We must use numerical integration to find approximate solutions to this integral. One hundred years ago, people used tables to look up values

[4] This integral cannot be done without multivariate calculus and double integrals. Feel free to ask me about it out of class.

[5] Start by doing a u-substitution, such as $u = x - a$. You should be able to split the integral up into two integrals such as

$$\int_{-\infty}^{\infty} u \frac{1}{b\sqrt{2\pi}} e^{-u^2/2b^2} + a \int_{-\infty}^{\infty} \frac{1}{b\sqrt{2\pi}} e^{-u^2/2b^2}.$$

Explain why the first integral must be zero, and the second must be 1. You'll have $0 + a \cdot 1$.

[6] Transcendental functions are the type of functions you studied in your college algebra class: powers, exponents, logarithms, and trigonometric functions.

for a normal distribution. Fortunately, computers can do the computations for us easily, and we can just use integrals as needed.

Problem 2.32 We know that ACT scores are approximately normally distributed with a mean of 21 points and standard deviation of 5 points. Let $X \sim N(21, 5)$. Remember, every time you are going to compute an integral, please show what integral you are computing, and then state its value.

1. Compute $P(-\infty \leq X \leq 21)$. What does this tell you about ACT scores?
2. Compute $P(16 \leq X \leq 26)$, the probability that a randomly chosen student will have a score within 1 standard deviation of the mean.
3. Compute the probability of randomly choosing a student whose score will be within 2 standard deviations of the mean.
4. Compute the probability of randomly choosing a student whose score will be within 3 standard deviations of the mean.

Problem 2.33 We know that SAT scores are approximately normally distributed with a mean of 1010 points and standard deviation of 130 points. Let $X \sim N(1010, 130)$. Remember, every time you are going to compute an integral, please show what integral you are computing, and then state its value.

1. Compute $P(-\infty \leq X \leq 1010)$. What does this tell you about SAT scores?
2. Compute $P(880 \leq X \leq 1140)$, the probability that a randomly chosen student will have a score within 1 standard deviation of the mean.
3. Compute the probability of randomly choosing a student whose score will be within 2 standard deviations of the mean, and then within 3 standard deviations of the mean.
4. Compare this problem with the previous. What do you notice?

Because the answers to the previous two problems are the same, we can translate questions about a normal distribution to questions about z-scores. We'll return to this more later. For now, let's examine one final distribution, namely the exponential distribution.

We've already seen the exponential distribution before in problem 2.18. The pdf for that problem was

$$f(x) = \begin{cases} 0.2e^{-.2x}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases},$$

which clearly involves an exponential function. This kind of distribution often gets used to model failure times, wait times, and times between occurrences.

Definition 2.9: Exponential Distribution. A continuous random variable X is said to be exponentially distributed with parameter λ if X takes on only nonnegative values and has probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases},$$

Problem 2.34 Suppose X is exponentially distributed with parameter λ. Show, by hand, that the $E(X) = 1/\lambda$ and $\text{Var}(X) = 1/\lambda^2$. You'll need to use integration-by-parts, perhaps several times, to compute both integrals.

Problem 2.35 Start by computing the cumulative distribution function for an exponentially distributed random variable with parameter λ.

1. Then compute the probability that X takes on a value less than or equal to its expected value $1/\lambda$, and round your answer to the nearest thousandth.
2. The median of X is the value x so that $P(X \leq x) = .5$ (so half the area under the curve is to the left, and half is to the right). Find the median of x. Your answer will be in terms of λ.

Problem 2.36 The lifespan of a school bus (in years) is modeled by the equation

$$f(x) = 0.08e^{-0.08x}$$

where x is the number of years, and $0 < x < \infty$.

1. Find the mean lifespan for a school bus.
2. Find the variance of the lifespan of a school bus.

We are often interested in the expected value of the function of a random variable. For example, a school district would want to know the expected cost to operate a school bus over its life time, in addition to the expected lifespan. To find the expected value of X, we just times each possible outcome (x years) by the probability $f(x)$ and then sum the result. If $C(x)$ represents the cost per year to operate a school bus that has been around for x years, then to find the expected value of $C(x)$ we could times each possible outcome ($C(x)$ \$/year) by the probability $f(x)$ and then sum these products.

Definition 2.10: Expected Value of a Function of A Random Variable. Suppose $f(x)$ is the pdf of the random variable X, and $g(x)$ is some function defined on the range of X. For discrete random variables, the expected value of the function $g(x)$ is given by

$$E\left[g(X)\right] = \sum_x g(x) \cdot f(x)$$

For continuous random variables, the equation is

$$E\left[g(X)\right] = \int_a^b g(x) \cdot f(x)dx$$

All we do is multiply each possible outcome $g(x)$ by the probability $f(x)$, and then sum the results.

Problem 2.37 The annual cost to operate a school bus can be modeled by a quadratic function as the bus ages. Data from one school district yields the model $C(x) = 96x^2 + 59600$ where $C(x)$ is the annual cost (in US\$ per year) to operate a bus that has been in service for x years. In question 2.36, the pdf for the lifespan of a school bus was given as $f(x) = 0.08e^{-0.08x}$.

If you would like additional information on the content of this chapter, you may want to read Chapter 2 in Navidi, *Statistics for Engineers and Scientists*, third edition.

1. Find the expected value $E[C(x)]$ of the annual cost to operate a school bus.
2. Conjecture a formula for computing the variance of the annual cost to operate a school bus, and use software to compute the integral you conjectured.

Problem: Class Activity Powerball is a multi-state lottery game. Fifty-nine white balls numbered 1–59 are placed in a drum and thoroughly mixed; five of the balls are drawn. Then, thirty-five red balls–one of which is labeled the "powerball"–are placed in another drum and thoroughly mixed, and one ball is drawn from this drum. Players win by correctly guessing some or all of the balls that will be selected in weekly drawings. The payout value of the grand prize depends on the number of tickets sold. If more than one ticket matches all the balls, the grand prize payoff is divided evenly among all the grand prize tickets.

Historical Note: The field of probability was developed to analyze games of chance. The ideas in this lesson can be used to determine the house advantage. This is the long-run average revenue the casino can expect for every dollar a gambler bets. This can be applied to Powerball to assess the mean profit from playing Powerball.

The profit is the amount a player wins minus the cost to play ($2). So, if a person wins a prize of $4, their profit is $2. If a person loses, their profit is −$2. We will assume that the grand prize is $200,000,000, although in many drawings it will be much less than this.

Table 2.1 illustrates the profit for all possible outcomes in a drawing where the grand prize is $200,000,000. The symbol "O" represents one white ball that was correctly guessed, and the symbol "P" represents that the powerball was correctly guessed.

Outcome	Payoff (Gross Winnings)	Profit x	Probability $P(x)$
OOOOOP	$200,000,000	$199,999,998	$\frac{1}{175,223,510}$
OOOOO	$1,000,000	$999,998	$\frac{34}{175,223,510}$
OOOOP	$10,000	$9,998	$\frac{270}{175,223,510}$
OOOO	$100	$98	$\frac{9,180}{175,223,510}$
OOO*P*	$100	$98	$\frac{14,310}{175,223,510}$
OOO	$7	$5	$\frac{486,540}{175,223,510}$
OO*P*	$7	$5	$\frac{248,040}{175,223,510}$
O*P*	$4	$2	$\frac{1,581,255}{175,223,510}$
P	$4	$2	$\frac{3,162,510}{175,223,510}$
Losing ticket	$0	−$2	?

Table 2.1: Powerball Probability Distrubtion (Excel Download)

Answer the following questions in a small group.

1. Based on your intuition, which of the following events is *least* likely?
   (a) You are struck by lightning twice
   (b) A plane falls on you from the sky
   (c) You win the Powerball jackpot
2. Some people say, "The probability of winning may be low, but if I don't play, that probability is zero!" How would you respond to this statement?
3. Suppose you purchase a $2 Powerball ticket. What is the probability that it is a losing ticket?
4. Use Excel to find the mean of the profit for people who play Powerball. You can use this link to download the data in an Excel file.

Problem If people played the lottery every week, and they or their friends never won anything, they might stop playing. For this reason, lottery makers want people to win something quite often, even if it's small. Let's look at this problem.

1. Let X_1 be the random variable which takes on the value 0 if you loose, and 1 if you win something. Compute both $P(X = 1)$ and $P(X = 0)$, rounding your answer to 4 decimal places.
2. Suppose you buy 1 ticket this week, and then 1 ticket the next week. Let X_2 be the random variable with returns zero if each ticket is a loosing ticket, and 1 otherwise (you won something at least once). Compute $P(X_2 = 0)$ and $P(X_2 = 1)$.
3. Now suppose you buy 1 ticket each week for 3 weeks. Let X_3 be the random variable with returns zero if each ticket is a loosing ticket, and 1 otherwise (you won something at least once). Compute $P(X_3 = 0)$ and $P(X_3 = 1)$.
4. Generalize the previous problem to 4 weeks, 5 weeks, etc. Can you come up with a formula for n weeks?
5. If you were part of a group of 20 people that played each week, and you each bought a ticket, what's the probability that at least one of you would win something?

Chapter 3

Jointly Distributed Random Variables

3.1 Conditional Probability and Bayes' Rule

Sometimes the probability of an event can be affected by other events. One simple application of this concept is the transmission of signals wirelessly from computer to computer. Imagine that computer 1 sends a long data string to computer 2; computer two sends the data on to computer 3. The probability of computer 3 receiving a data string with no errors is higher if the data string that computer 2 is sending is correct. On the other hand, if computer 2 incorrectly decodes the information from computer 1, the odds of computer 3 ending up with the right information are very low indeed. The condition of what happens between computers one and two tells us more about the probability that computer 3 will end up with the right information.

Problem 3.1 The following table summarizes data on 161 student responses to questions regarding height and favorite color.

	Height		
Color	under 68"	68"-71"	over 71"
Red	10	13	5
Blue	20	42	4
Other	29	25	13

1. What is the probability that a randomly chosen student from this group is in the height range of 68 to 71 inches?

2. What is the probability that a randomly chosen student from this group listed red as their favorite color?

3. Suppose that one of the students randomly asks you to guess his favorite color. If you can tell that he is under 68" in height, what should you guess as his favorite color? What is the probability that you will be correct?

4. Now suppose that a student wants you to guess their height and the only thing you know about the student is that they love all things Blue. Which of the three height ranges should you guess? What is the percent chance that you will be right?

We'll be using set theory notation to talk about many of the ideas that follow. As such, let's give a formal definition so we all use the same notation.

Definition 3.1: Set, Subset, Intersection, Union, Complement. A set A is a collection of objects. For A to be a set, it must be possible to determine, given an object, if the object is in the set or not.

- We say that B is a subset of A, written $B \subseteq A$, if every element in B is an element in A.
- The intersection of two sets A and B, written $A \cap B$, is the collection of objects that are in both set A and in set B.
- The union of two sets A and B, written $A \cup B$, is the collection of objects that are in set A, or in set B, or in both sets A and B. When we say that something is in A or B, we are including the possibility that it might be in either.
- Suppose that A is a subset of some larger set U. The complement of A in U, written $\overline{A}$, is the set of objects that are in U but not in A.

Let's practice using these words and symbols.

Problem 3.2 Use the chart from problem 3.1 to answer the following questions. Let R represent the set of students whose favorite color is red. Similarly define B and O. Let X, Y, and Z be the sets of students whose heights are 68", 68"-71", and over 71", respectively. So if we write $B \cap Z$, then we mean the set of students whose favorite color is blue and whose height is over 71".

1. Suppose that we randomly choose one of the 161 students. Compute the probability that the student is in $R \cap Y$, which we write as $P(R \cap Y)$. Leave your answers as a fraction (don't use decimals).
2. Now compute $P(R \cup Y)$. Show how you obtained your answer.
3. Compute $P((R \cap O) \cup X)$. Also compute $P((R \cup X) \cap (O \cup X))$. If you feel like the answer on this one is silly, then you're doing it right. However, the point here is not to obtain a number, but to show how you obtained each number. Make sure your work shows how each computation is performed.
4. Compute $P((R \cup O) \cap X)$ and $P((R \cap X) \cup (O \cap X))$. Show how each computation is performed.

Problem 3.3 Continue from the previous problem. Each of the sets R, B, O, X, Y, and Z are all subsets of the entire set of students. As such, we can talk about their complements. With each of the following, please leave your answer as a fraction. The point is to show how you obtained each answer, not just give an answer, so please make sure you work shows this.

1. Compute both $P(B)$ and $P(\overline{B})$.
2. Compute the four probabilities $P(\overline{(B \cap Z)})$, $P(\overline{(B \cup Z)})$, $P((\overline{B} \cap \overline{Z}))$, and $P((\overline{B} \cup \overline{Z}))$.

The previous problem introduced a set theory law called De Morgan's law, which states that

$$\overline{(A \cap B)} = (\overline{A}) \cup (\overline{B}) \quad \text{and} \quad \overline{(A \cup B)} = (\overline{A}) \cap (\overline{B}).$$

The complement of an intersection is the union of the complements, and the complement of a union is the intersection of the complements.

You'll see these laws show up on the FE exam (for engineers). You don't have to memorize them, rather you just need to be able to use them.

Now that we've practiced some with set theory notation, let's return to some more probability questions. We can compute the probability that an event A occurs. However, if we already know that some other event B has occurred, then how does this affect the probability of A occurring. We'll write this as $P(A \text{ given } B)$ or $P(A|B)$.

Problem 3.4 Use the chart from problem 3.1 to answer the following.

1. What is the probability of choosing someone that is both over 71" and has red as a favorite color?
2. What is the probability of someone having red as a favorite color?
3. What is the probability of someone choosing someone that picked red as a favorite color given that you know they are over 71" tall?
4. If having red as a favorite color is called set A and being over 71" tall is set B, express each of the previous three questions using set theory notation. Then express your answer to the third question purely in terms of sets. Your answer here should look like $P(A \text{ given } B) =$ some fraction with a $\cap$ or $\cup$ symbol.

In the final part of the previous questions, you found a very significant rule about conditional probability. As a reminder, we use the vertical bar symbol to denote "given that". For example, we'll write the probability of A given B as $P(A|B)$.

The next problem has you build some tables of conditional probabilities. Given some data, one of the first things we'll want to do with that data is convert it into a table of probabilities. However, when we have several variables that we are collecting then there are several ways to compute these probabilities.

Problem 3.5 Here is the table from problem 3.1.

	Height			
Color	under 68"	68"-71"	over 71"	Row Total
Red	10	13	5	28
Blue	20	42	4	?
Other	29	25	13	?
Column Total	?	80	?	161

Please use excel to complete the following.

1. One of the first things we'll often do with a table of data is add some margins to the table where we keep track of the totals in each row and in each column. Please fill in the missing boxes in the margins.

2. To obtain the probability of each event occurring, we just need to divide each entry by the grand total 161. Create such a table (with 16 entries) where you divide every number by the total. From your table, what is the probability of a student picking blue as a favorite color.

3. Sum the three probabilities in the first column, ignoring the margin. What do you get? Does this same pattern hold if you sum the three numbers in the second column?

4. What number do you get if you sum the probabilities in a given row, excluding the margin?

5. If you sum the probabilities in the margin on the right (so the row total probabilities), what do you get? Are there any other collections of numbers that you can sum to get this exact same number?

We'll often call the table you created above a joint distribution function. There are two different variables we are analyzing, so our table of values is needed to describe the probabilities of any possible collection of events. If we had three variables, we would need a three dimensional array to keep track of the data. Drawing such a table can be cumbersome, but luckily modern computers are equipped with multidimensional array capabilities that makes keeping track of large amounts of data quite simple. We'll focus on just comparing two variables, as the ideas we develop here can be used on larger problems. This next problem has you build a probability table of conditional probabilities.

Problem 3.6 We know how to compute the probability of a favorite color being red given that a student's height is between 68" and 71". Since there are 80 students in that height range, and of those 80 students only 13 have red as their favorite color, we just compute $P(\text{Red given 68"-71"}) = 13/80$. In a like manner, we can compute the conditional probabilities of picking a student with a specific favorite color given that the student's height falls in a specific category. Please do so now for every combination, completing the following table as well as the margins. I've filled in a few values for you.

P(Color given Height) **Color**	**Height** under 68"	68"-71"	over 71"	Row Total
Red		13/80		28/161
Blue				
Other	29/59			
Column Total		80/80		

Ignoring the margins, do the rows or the columns individually sum to 1? How can the margins help you remember this?

Problem 3.7 Let's swap the roles of the previous problem. Now let's create a table of conditional probabilities of picking a student with a specific height given that we know their favorite color. Please do so now, completing the following table, as well as the margins. I've filled in a few values for you.

P(Height given Color) **Color**	**Height** under 68"	68"-71"	over 71"	Row Total
Red		13/28		
Blue				
Other			13/67	
Column Total	59/161			

Ignoring the margins, do the rows or the columns individually sum to 1? How can the margins help you remember this?

3.1.1 Bayes' Rule

Now that we know the basic way that conditional probability works, let's look at one of the more convenient additional properties that conditional probability gives us. In problem 3.4, we obtained the following formula for conditional probabilities, namely that

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

provided of course that $P(B) \neq 0$. We can use this to create a multiplication rule for computing the probability of the intersection of two events.

Problem 3.8 Complete the following:

1. Write a formula for $P(B|A)$. It should look quite similar to the equation for $P(A|B)$.
2. Take the formula you just got and solve for $P(A \cap B)$. This should give you a multiplication rule for computing probabilities.
3. Now that we have an expression for $P(A \cap B)$, write an equivalent expression that doesn't use any $\cap$ or $\cup$ symbols. In other words, translate the notation into an English sentence.
4. Your answer above requires that we know $P(B|A)$ and $P(A)$. What if instead we know $P(A|B)$ and $P(B)$. How do we compute $P(A \cap B)$ using this knowledge?

The rules you develop above are called Bayes' rule. They give a way to compute the probability of A and B by multiplying together two other probabilities. We'll return to this rule later, and see its power.

Problem: Optional (Come by my office and share this one with me if you would like extra presentation points.) Use De Morgan's laws and Bayes' rule to develop a formula for $P(A \cup B)$ that depends on some of $P(\overline{A})$, $P(\overline{B})$, $P(\overline{A}|\overline{B})$, and/or $P(\overline{B}|\overline{A})$.

"Piled Higher and Deeper" by Jorge Cham

3.2 Jointly Distributed Random Variables: Discrete Case

Sometimes we sample items from a population that are linked to several random variables. For example, a quality engineer typically makes several measurements on a randomly selected product, testing length, weight, color, visible flaws, etc.

When two random variables (X and Y) are measured on each unit in a sample, the value of one random variable is typically related to another. The associated probabilities vary jointly. These are often described using a **joint probability distribution function**, or a joint pdf. If the random variables are discrete, we often summaries the joint pdf in a table.

Defects In a paint booth at an automotive manufacturing facility, two serious defects are (1) foreign object debris, FOD, in the paint and (2) runs in the paint. We'll define the following random variables:

- Let X be equal to 1 if there are runs in the paint and equal to 0 otherwise.
- Let Y be equal to 1 if FOD is present in the paint and 0 otherwise.

The joint pdf for the random variables X and Y is given in table 3.1.

$f_{X,Y}(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	0.9603	0.0297	0.99
$x=1$	0.0097	0.0003	0.01
$f_Y(y)$	0.97	0.03	

Table 3.1: Joint PDF of presence of runs (X) and FOD (Y) in paint.

The values in the center of the table give the probability of individual events. For example, the probability that a randomly selected part will have no runs and no FOD is 0.9603. This can be written using probability notation as $P(X=0, Y=0) = 0.9603$ or, equivalently as $f_{X,Y}(0,0) = 0.9603$.

The values in the right and bottom margins of the table give the **marginal probability distributions**, $f_X(x)$ and $f_Y(y)$, respectively. For example, the probability that a randomly selected part will have runs is: $P(X=1) = f_X(1) = 0.01$. The probability that there will be no FOD is $P(Y=0) = f_Y(0) = 0.97$.

Definition 3.2: Joint Probability Function Notation. Suppose X and Y are discrete random variables. We'll use the notation $f_X(x)$ to compute the

probability that $X = x$. We'll use the notation $f_Y(y)$ to compute the probability that $Y = y$. We'll use the notation $f_{X,Y}(x, y)$ to compute the probability that $X = x$ and $Y = y$. Notationally, we can summarize all this as

$$\begin{aligned} f_X(x) &= P(X = x), \\ f_Y(y) &= P(Y = y), \text{ and} \\ f_{X,Y}(x, y) &= P(X = x \text{ and } Y = y). \end{aligned}$$

Problem 3.9 Use Table 3.1 to answer these questions.

1. What is the probability that the paint in a randomly selected part will have FOD but no runs?
2. What is the probability that the paint in a randomly selected part will have no runs?
3. Find the probability that the paint on a randomly selected part will have FOD.
4. Given that the paint on a part contains FOD, what is the probability that the paint will have runs?
5. Given that the paint has runs, what is the probability that the paint contains FOD?
6. What is the product of your answer to question 3.9(2) and your answer to question 3.9(3)? Compare this result to your answer for question 3.9(1)?

Given a joint probability distribution, we can use the margins to study each variable independently. This means we can compute the mean (expected value) and variance of each random variable. The next question has you do this.

Problem 3.10 Use Table 3.1 to answer these questions.

1. Use $f_X(x)$ to find the mean of the random variable X. Interpret this result. (All you have to do is multiply each of the possible outcomes for X by its respective probability, and then sum the result. There are two possible outcomes for X, namely $x = 0$ and $x = 1$, so you should a short sum.)
2. Use $f_X(x)$ to find the variance of the random variable X. (Remember to subtract the mean from each possible X value, square the difference, times by the probability, and then sum the result.)
3. Use $f_Y(y)$ to find the mean of the random variable Y. Interpret this result.
4. Use $f_Y(y)$ to find the variance of the random variable Y.

3.2.1 Expected Values of Functions of Jointly Distributed Discrete Random Variables

If X is a random variable with probability density function $f(x)$, then we've already defined the expected value of X to be

$$E[X] = \sum xf(x) \qquad \text{for discrete random variables, and}$$

$$E[X] = \int xf(x)dx \qquad \text{for continuous random variables.}$$

If $g(X)$ is a function of the random variable X, then we've also defined the expected value of $g(X)$ to be

$$E[g(X)] = \sum g(x)f(x) \qquad \text{for discrete random variables, and}$$

$$E[g(X)] = \int g(x)f(x)dx \qquad \text{for continuous random variables.}$$

These definitions generalize to functions of jointly distributed random variables. For discrete random variables X and Y, the **expected value of a function of these random variables** is defined as

$$E\left[g(X,Y)\right] = \sum_{X,Y} g(x,y) \cdot f_{X,Y}(x,y) \tag{3.1}$$

where $g(x,y)$ is a function of X and Y, and $f_{X,Y}(x,y)$ is the joint pdf of X and Y.

Problem 3.11 Use the definition of the expected value of a function of random variables, and the information in Table 3.1 to answer these questions.

1. Let $g(X,Y) = X$. Find the expected value of $g(X,Y)$, written

   $$E[g(X,Y)] = E[X].$$

   The following table may be helpful. The sum of the values in Column 5 give the desired result.

Column 1	Column 2	Column 3	Column 4	Column 5
x	y	$f_{X,Y}(x,y)$	$g(X,Y)$	$g(X,Y) \cdot f_{X,Y}(x,y)$
0	0	0.9603	0	$0 \cdot 0.9603$
1	0	0.0097	1	$1 \cdot 0.0097$
0	1	0.0297	?	?
1	1	0.0003	?	?
$E\left[g(X,Y)\right] = \sum_{X,Y} g(x,y) \cdot f_{X,Y}(x,y) \; =$				?

   Compare your answer to the value you computed in 3.10(1).

2. Let $g(X,Y) = Y$. Find the expected value of $g(X,Y)$, so compute

   $$E[g(X,Y)] = E[Y].$$

   Compare this result to the value you obtained in 3.10(3).

3. The function $g(X,Y) = X + Y$ gives the number of different defects (of runs or FOD) in the paint. Find and interpret the expected value of $g(X,Y)$, namely compute

   $$E[g(X,Y)] = E[X+Y].$$

4. Are your three answers above connected to each other in any way? How? What do you notice?

By letting $g(X,Y) = X$, you should have seen that you just obtain the mean of X. Similarly, letting $g(X,Y) = Y$ gives us the mean of Y. We can also compute the mean number of total defects, by letting $g(X,Y) = X + Y$. Any time there is a defect, the company decides to repaint, which means an added cost to production. The next problem has you compute the expected cost associated with repainting.

Problem 3.12 Use Table 3.1 to answer these questions.

1. When a run or FOD is found in the paint, the part must be reworked. If the cost for rework of a run is \$30 and for FOD the cost is \$20, then for each piece, the cost for rework is given by the function
$$g(X, Y) = 30X + 20Y.$$
This function assumes that if a part has both FOD and a run, then the cost associated with rework is \$50. Find the expected value of $g(X, Y)$ by filling out a table similar to the one in problem 3.11. Interpret this result.
2. We have already computed $E[X]$ and $E[Y]$ in both of the previous two problems. Now compute
$$30 \cdot E[X] + 20 \cdot E[Y].$$
How does this relate to the first part of this problem.

Problem 3.13 Use Table 3.1 to answer these questions.

1. Let
$$g(X, Y) = (X - \mu_X)^2$$
Find the expected value of $g(X, Y)$. We have already given a name to this value. What's its name?

We can denote the mean of the random variable X either as $E[X]$ or μ_X. Similar notation is used for the mean of Y.

2. Let
$$g(X, Y) = (Y - \mu_Y)^2$$
Find the expected value of $g(X, Y)$. Try to do this without any additional computations, rather just look back at a previous problem (which one) and state the answer.
3. Let $g(X, Y) = (X - \mu_X)(Y - \mu_Y)$. Find the expected value of $g(X, Y)$. You'll want to create a table for this one, similar to the one in problem 3.11. The quantity
$$E[(X - \mu_X)(Y - \mu_Y)]$$
is called the **covariance** of X and Y and written $Cov(X, Y)$.
4. The **correlation coefficient** for X and Y is defined as
$$\rho = \frac{Cov(X, Y)}{\sqrt{VarX} \cdot \sqrt{VarY}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2} \cdot \sqrt{\sigma_Y^2}}.$$
Find this quantity.

Definition 3.3: Covariance and Correlation Coefficient. The **covariance** of X and Y, written $Cov(X, Y)$, is the quantity
$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$
The **correlation coefficient** for X and Y is the quantity
$$\rho = \frac{Cov(X, Y)}{\sqrt{VarX} \cdot \sqrt{VarY}} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2} \cdot \sqrt{\sigma_Y^2}}.$$

The previous problems shows us that variance is just an expected value of a function of a random variable. As we progress through the semester, we'll see that many questions about probability are all connected to the concept of finding the expected value of some function of the random variables.

Jump Drive Dimensions Table 3.2 gives the joint pdf for the length L and width W of a jump drive in millimeters (mm). The length can be either 49, 50, or 51 mm. The width can be either 17 or 18 mm.

$f_{(L,W)}(\ell, w)$	$w = 17$	$w = 18$	$f_L(\ell)$
$\ell = 49$	0.14	0.06	0.20
$\ell = 50$	0.40	0.30	0.70
$\ell = 51$	0.06	0.04	0.10
$f_W(w)$	0.60	0.40	

Table 3.2: Joint pdf of the length (L) and width (W) of a jump drive

Problem 3.14 Refer to Table 3.2 as you complete this problem.

Compute each of the following sums, and explain the significance of each result. As you expand each sum, please show the intermediate steps (which numbers are you adding), before stating the answer.

The point to this problem is to help you get more comfortable with summation notation, and notice many patterns about summing rows, columns, etc., in a joint distribution table.

1. $\displaystyle\sum_{w=17}^{18}\sum_{\ell=49}^{51} f_{(L,W)}(\ell, w)$
2. $\displaystyle\sum_{\ell=49}^{51} f_L(\ell)$
3. $\displaystyle\sum_{w=17}^{18} f_W(w)$
4. $\displaystyle\sum_{\ell=49}^{51} f_{(L,W)}(\ell, 17)$
5. $\displaystyle\sum_{\ell=17}^{18} f_{(L,W)}(w, 50)$

Problem 3.15 Refer to Table 3.2 as you complete this problem.

1. What is the probability that a randomly selected jump drive will have a length of 50 mm and a width of 18 mm? In other words, compute $P(L = 50 \text{ and } W = 18)$.
2. What is the probability that a randomly selected jump drive will have a length of 50 mm?
3. What is the probability that a randomly selected jump drive will have a length that is less than 51 mm?
4. What is the probability that a randomly selected jump drive will have a width of 18 mm?

Problem 3.16 Refer to Table 3.2 as you complete this problem.

1. Given that the jump drive has a width of 17 mm, what is the probability that the length will be 51 mm? In other words, compute $P(L = 51|W = 17)$.
2. Compute $P(W = 17|L = 51)$, i.e given that the jump drive has a length of 51 mm, compute the probability that the the width will be 17 mm.

3. Bayes' rule tells us that $P(A \cap B) = P(A|B)P(B)$. In problem 3.9 we saw that sometimes $P(A \cap B) = P(A)P(B)$. Let's see if that rule holds true for the jump drive scenario.

   Let A represent the event that the jump drive has a width of 17mm. Let B represent the event that the jump drive has a length of 51mm. Is it true that $P(A \cap B) = P(A)P(B)$?

4. Now let A represent the event that the jump drive has a width of 17mm, and let B represent the event that the jump drive has a length of 49mm. Is it true that $P(A \cap B) = P(A)P(B)$?

Problem 3.17 Continue to refer to Table 3.2 as you complete this problem.

1. Let $g(L, W) = L$. Find the expected value of $g(L, W)$, in other words compute
$$E[g(L, W)] = E[L] = \mu_L.$$
   Please use excel to simplify your computations.

2. Let $g(L, W) = W$. Find the expected value of $g(L, W)$, namely
$$E[g(L, W)] = E[W] = \mu_W$$

3. Compute the expected value of the perimeter of the jump drive. (You'll need a function for the perimeter, so $g(X, Y) =?$.)

Problem 3.18 Continue to refer to Table 3.2 as you complete this problem.

1. Compute the expected value of the surface area (so $g(L, W) = LW$) of one side of the jump drive. Make sure you show us how you computed this from a table.

2. The cost of the plastic casing for the jump drive is:

   - \$0.015 per linear millimeter of the perimeter,
     plus
   - \$0.0065 per square millimeter for *each* of the two faces (the top and bottom).

   Write a function $C(\ell, w)$ that gives the total cost of the plastic casing for the jump drive, given the dimensions.

3. Find the expected cost of the plastic casing for the jump drive.

Problem 3.19 Continue to refer to Table 3.2 as you complete this problem.

1. Let $g(X, Y) = (X - \mu_X)^2$. Find $E[g(X, Y)] = E\left[(X - \mu_X)^2\right] = \sigma_X^2$.

2. Let $g(X, Y) = (Y - \mu_Y)^2$. Find $E[g(X, Y)] = E\left[(Y - \mu_Y)^2\right] = \sigma_Y^2$.

3. Find the covariance of the random variables X and Y.

4. Compute the correlation coefficient for the random variables X and Y.

3.2.2 Conditional Probability (Revisited)

Problem 3.20 A coin will be tossed three times and the result of each toss will be recorded. If we let "H" represent *heads* and let "T" represent *tails* on each toss of the coin, then the possible outcomes on the three tosses are: HHH, HHT, HTH, THH, HTT, THT, TTH, and TTT. Each of these is equally likely to occur. We define two random variables as follows:

- Let X be the number of times *heads* occurs.
- Let Y be equal to 1 if the first toss results in *heads* and 0 if the first toss is *tails*.

We can organize this information into a joint probability distribution for the random variables X and Y:

Column 1 Column 2 Column 3 Column 4

$f_{X,Y}(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	0.125	?	0.125
$x=1$	0.250	0.125	?
$x=2$	?	?	0.375
$x=3$	0	?	?
$f_Y(y)$	0.5	?	

1. Fill in the missing values in the table above.
2. Suppose that the coin has been tossed once and it came up *tails*. What is the probability that no *heads* will occur in a total of three tosses? (In other words, given that $y=0$, what is the probability that $x=0$?)
3. Continue the calculations you did in question 3.20(2) to find the probabilities that $x=1$, $x=2$, or $x=3$, given that $y=0$. Fill your results in column 2 of the table below.[1] After filling in this column, compute the sum of these probabilities and write it at the bottom of Column 2.

Column 1 Column 2 Column 3 Column 4

$P(X \text{ given } y)$	$y=0$	$y=1$	
$x=0$		0	–
$x=1$	.5		–
$x=2$			–
$x=3$			–
f			–

4. Now, suppose the first toss had been *heads* (i.e., $y=1$). Given that this has already occurred, find the probability that $x=0$, $x=1$, $x=2$, and $x=3$. Also, compute the sum of these values. Fill this information in Column 3 above.

Problem 3.21 Continue from the previous problem.

[1] Notice that we are only working with Column 2. In this case, we know that $y=0$. Since the first toss has occurred, it is impossible for y to be equal to 1.

1. Suppose you know that in three tosses of a coin, *heads* never appeared. What is the probability that the first toss yielded *tails*?[2] What is the probability that the first toss yielded *heads*?[3]

2. Starting with the probabilities you computed in question 3.20(1), fill in the entries in the following table with the probability that Y equals the value specified in the column, given that X equals the value stated in each row. Compute the sum for each row and write it in Column 4.

	Column 1	Column 2	Column 3	Column 4
$P(Y \text{ given } x)$	$y=0$	$y=1$	
$x=0$	?	?	?
$x=1$	?	?	?
$x=2$	?	?	?
$x=3$	?	?	?
	–	–	

Let's work through one more examples with conditional probabilities, before we summarize what we have learned and draw some conclusions.

Problem 3.22 Suppose in an investigation into the quality of Girl Scout cookies, it was determined that some of the boxes had defects. Some contained broken cookies, others were under weight. Three different categories of cookies were considered: Thin mints, shortbread, and other. The following table summarizes the probability that each type of defect would occur in the three categories of cookies.

$f_{X,Y}(x,y)$		Broken $y=1$	Low weight $y=2$	No defect $y=3$	$f_X(x)$
Thin Mints	$x=1$	0.03	0.006	0.264	0.3
Shortbread	$x=2$	0.01	0.002	0.088	0.1
Other	$x=3$	0.06	0.012	0.528	0.6
	$f_Y(y)$	0.10	0.02	0.88	

Use this table to answer the following questions.

1. The probability that $X = x$ given $Y = y$ is called the **conditional probability** of X given Y. We denote this as: $P(x|y)$. This is also written as $f_{X|Y}(x|y)$. Fill in the following table with the conditional probabilities that X will equal x, given that $Y = y$.

$f_{X\|Y}(x\|y)$		Broken $y=1$	Low weight $y=2$	No defect $y=3$	
Thin Mints	$x=1$	?	?	?	–
Shortbread	$x=2$	?	?	?	–
Other	$x=3$	?	?	?	–
		?	?	?	

[2] This is the same as saying, what is the probability that $y = 0$ given that $x = 0$.
[3] This is the same as saying, what is the probability that $y = 1$ given that $x = 0$.

2. Now, fill in the following table with the values of the conditional probability of Y given X.

$f_{Y\|X}(y\|x)$		Broken $y=1$	Low weight $y=2$	No defect $y=3$	
Thin Mints	$x=1$	?	?	?	?
Shortbread	$x=2$	?	?	?	?
Other	$x=3$	?	?	?	?
		–	–	–	

3. What do you observe in the previous questions?

Problem 3.23 In problem 3.20, you completed a table of probabilities. In the right and bottom margins of this table are the **marginal probability distributions**. These are denoted $f_X(x)$ and $f_Y(y)$, respectively.

1. Choose a value of x and a value of y. Compute $f_X(x) \cdot f_Y(y)$. Is this value equal to $f_{X,Y}(x,y)$? Do you draw the same conclusion for other values of x and y?
2. In problem 3.22, a table of probabilities is given to you. Choose a value of x and a value of y from the table. Compute $f_X(x) \cdot f_Y(y)$. Is this equal to $f_{X,Y}(x,y)$? Do you draw the same conclusion for other values of x and y?
3. How do the conclusions to the first part of this problem compare? In one of these problems, the random variables are **independent** and in the other problem the random variables are **not independent**. Can you determine which is which?
4. If you were to make a definition for the term **independent**, what would it be?

Definition 3.4: Independent Random Variables. Let A and B be two events. We say that A and B are independent if the occurrence of B does not affect the probability of A, which means that $P(A|B) = P(A)$. In other words, we have

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B),$$

which shows that events are independent if and only if $P(A \cap B) = P(A)P(B)$. If two events are not independent, then we say they are not independent (we do NOT say that they are dependent).

Suppose that X and Y are discrete random variables. We say that the random variables are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. In other words, the random variable Y does not affect the probability distribution of the random variable X, so we have $f_{X|Y}(x|y) = f_X(x)$.

Problem 3.24 Suppose we know that two independent random variables X and Y have the probability density functions $f_X(x)$ and $f_Y(y)$ shown in the table below.

$f_{X,Y}(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	?	?	0.5
$x=1$	?	?	0.2
$x=2$	?	?	0.3
$f_Y(y)$	0.7	0.3	

Use this information to fill in the the joint probability distributions for $f_{X,Y}(x, y)$. Then create two more probability tables for $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. When you are done doing this, you should see how nice the tables look when variables are independent, as well as which percentages should be equal when looking at comparing conditional probabilities.

3.3 Component Failure - Series Versus Parallel

When considering the reliability of computer systems, the analysis of the lifetimes of components is crucial. Companies need to understand the expected lifetimes of critical system components. This broad area is known as survival analysis. (Think of the survival of the component.)

Problem 3.25 Two components in an electrical circuit are connected in series. The circuit will fail when either of these components fails. Figure 3.1 illustrates a part of this circuit. The lifetimes of the two components in this

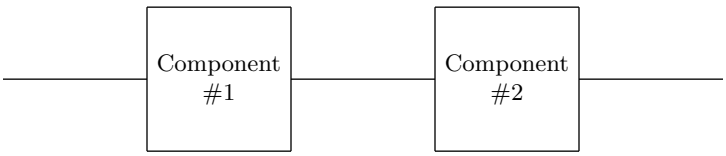


Figure 3.1: Two Series Components in an Electrical Circuit

circuit can be modeled as random variables. Let X represent the lifetime of component #1 (in years). Let Y be the lifetime of component #2 (in years). Let's assume that X and Y are independent random variables. Suppose that the corresponding probability density functions for X and Y are

$$f_X(x) = \frac{1}{3}e^{-(x/3)} \quad \text{and} \quad f_Y(y) = \frac{1}{4}e^{-(y/4)},$$

so each is exponentially distributed (hence we know $x > 0$ and $y > 0$).

1. For each component, compute the probability that that component will last at least five years. Remember, this just means you must compute the integrals

$$P(X \geq 5) = \int_5^\infty f_X(x)\,dx \quad \text{and} \quad P(Y \geq 5) = \int_5^\infty f_Y(y)\,dy.$$

2. (No Multivariable Calc Version) Compute the probability that the circuit will not fail in the first 5 years. This requires that both components make it at least 5 years.

2. (Multivariable Calc Version) Compute the probability that the circuit will not fail in the first 5 years. This requires that both components make it at least 5 years. Do so by calculating the double integral

$$\int_5^\infty \int_5^\infty f_{X,Y}(x, y)\,dx\,dy = \int_5^\infty \int_5^\infty f_X(x) f_Y(y)\,dx\,dy.$$

(Notice, you just have to times the pdf's together to get the joint pdf, because we have assume that the random variables are independent.

3. (If you have had multivariate calculus, Math 215) The probability that component #1 will fail before component #2 fails is computed by evaluating the integral [4]

$$\int_0^\infty \int_0^y f_{X,Y}(x,y)\,dx\,dy.$$

Evaluate this integral and interpret the result. Then set up an appropriate integral that would give the probability that component #2 fails before component #1, and state this probability.

Problem 3.26 Figure 3.2 is a schematic of a portion of a system where two components are in parallel. The system will function properly if either one of the two components are functional.

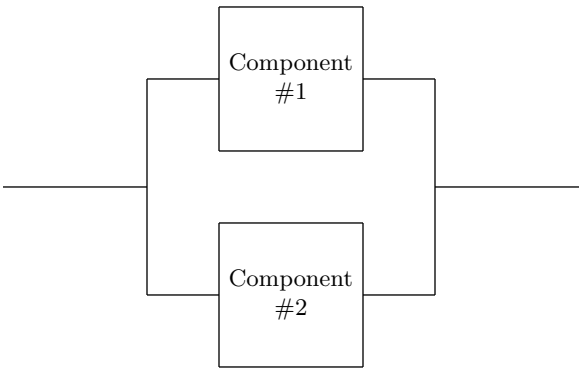


Figure 3.2: Two Parallel Components in an Electrical Circuit

The lifetimes of the two components in this circuit can be modeled as random variables. Let X represent the lifetime of component #1 (in years). Let Y be the lifetime of component #2 (in years). The pdf's for these random variables are

$$f_X(x) = \frac{1}{5}e^{-(x/5)} \quad \text{and} \quad f_Y(y) = \frac{1}{2}e^{-(y/2)},$$

where $x > 0$ and $y > 0$.

1. For each component, compute the probability that the component will last longer than 5 years.

2. Find the probability that the system will last longer than 5 years. [Hint: what's the probability that both fail before 5 years?]

Problem 3.27 Suppose that three components are connect together in some system. We'll call them components A, B, and C. Suppose that the probability that component A will last at least 5 years is $p_A = .9$, the probability that component B will last at least 5 years is $p_B = .8$, and the probability that component C will last at least 5 years is $p_C = .7$.

[4] This is the multiple integral in the X, Y plane of the function $f_{X,Y}(x,y)$ over the region in the first quadrant where $y > x$.

1. If these components are all connected in a series, so that each component must work for the system to work, then compute the probability that the entire system will work for at least 5 years.

2. If these components are all connected in a parallel manner, so that only of the components must function for the entire system to function, then compute the probability that the entire system will work for at least 5 years.

3. How would you generalize your work to each of the previous parts if there were 4 components instead of just 3?

That completes the ideas in this chapters. The next problem asks you to show an alternate formula for computing the variance, which allows you to compute the expected value of X and X^2, instead of $(X - \mu_X)^2$, and then combine these expected values to get the variance.

Problem 3.28 We've been using the formula

$$\text{Var}[X] = E[(X - \mu_x)^2] = \sum (x - \mu_X)^2 f_X(x)$$

to compute the variance of a random variable X. There is another way to perform this computation, that is sometimes much simpler. Show that we can also compute the variance using the formula

$$\text{Var}[X] = E[X^2] - (E[X])^2 = \left(\sum x^2 f_X(x)\right) - (\mu_X)^2.$$

Problem: Class Activity Suppose in a certain town that there is a probability of .005 that a person will have a specific disease. There are medical tests available to check if a person has this disease. If someone has the disease, then there is a .99 probability that the test will correctly say the person has the disease. If the person does not have disease, then there is a .02 probability that the test will incorrectly say the person has a disease (called a false positive).

- Let D be a random variable where $d = 0$ means you do not have the disease, and $d = 1$ means you do have the disease.
- Let P be a random variable where $p = 0$ means the test returns a negative result, and $p = 1$ means the test returns a positive result (meaning the test says you have the disease).

We have already been given several probabilities above, namely we have been told $P(d = 1)$, $P(p = 1|d = 1)$, and $P(p = 1|d = 0)$.

1. Use the information above to create a joint probability distribution for D and P. Make sure your work shows how you obtained the numbers in your table. You'll need to make use of Bayes' rule $P(A \cap B) = P(A|B)P(B)$.

$f_{D,P}(d,p)$	$p=0$	$p=1$	$f_D(d)$
$d=0$	?	?	?
$d=1$	?	?	?
$f_P(p)$	?	?	

2. What is the probability that you have the disease but the test tells you you do not. Also, what is the probability that you do not have the disease but the test tells you that you do have the disease? You should be able to get both of these numbers straight out of your distribution.
3. Suppose a test comes back positive. What is the probability that the person has the disease? In other words, compute $P(d = 1|p = 1)$. Hopefully this result is rather surprising, as being told you have a disease by a test, even though you don't have it, can be quite disruptive to life.
4. Compute the expected value of P. Do so by letting $G(D, P) = P$ and then computing $E[G(D, P)]$.
5. State the expected value of D.
6. Compute the variance of P. Do so by computing $E[(P - E[p])^2]$.
7. Find the variance of D.
8. Compute the covariance of D and P. Remember, this means you are computing $E[(D - \mu_D)(P - \mu_P)]$.

Exam Review

To prepare for the exam, we will spend one day of class focused on review. The goal of this class period is to have each student find/create examples to illustrate each of the big ideas from the material we have been studying. You'll see a list of the big ideas on the next page. To prepare for class, here is your assignment.

1. For each concept on the next page, find or create an example that illustrates the concept.

2. Organize your work into a "lesson plan" where you include the problem and key steps to solving that problem on your lesson plan.

3. Come to class and spend 1 hour with a partner teaching from the ideas in your lesson plan. The idea here is to let each person share an example, then swap roles. If you both feel like you need more practice in a specific area, spend your time there.

4. Report in I-Learn that you have completed your lesson plan and taught it to your peer, and upload your document to I-Learn.

If you are gone for the STEM fair, then you can still complete this assignment by finding another student from our class who is at the stem fair, and you can spend 1 hour teaching each other from your examples.

Chapter 1 - Statistical Exploration

1. What's the difference between the mean and median. Compare them.
2. Describe the shape of a distribution. When you look at a histogram of averages, what shape do you get?
3. Be able to use several ways to measure the spread of a distribution, in particular know how to find the range, interquartile range, max, min, variance, and standard deviation.
4. Compute the mean, variance, and standard deviation given raw data, or given summarized data.
5. Compute and interpret z-scores and percentiles, both of which are measures of position in a distribution.
6. Create and interpret box plots and histograms.

Chapter 2 - Probability

1. Explain what a random variable is, and the difference between continuous and discrete.
2. Use the basic laws of probability to compute various probabilities, in particular the law of total probability and the complement rule.
3. Determine if a function is a probability density function, and obtain normalizing constants when it is not.
4. Use pdfs to compute probabilities, both from a table for discrete random variables and using integrals for continuous random variables.
5. Show how to obtain a cumulative distribution function from a pdf, and show how to use the cdf to compute probabilities.
6. Find the expected value and variance of both discrete and continuous random variables.
7. Be comfortable with performing probability computations using integrals, especially as they relate to normal, exponential, uniform, and triangular distributions.

Chapter 3 - Jointly Distributed Random Variables

1. Use and interpret expressions given in set notation, as they relate to probability computations.
2. Explain how to compute conditional probabilities, and use them to explain Bayes's Rule.
3. Find the expected value, variance, and covariance of joint distributions, and functions $G(X, Y)$ of joint distributions.
4. Create probability distributions for $P(X = x, Y = y)$, for $P(X = x|Y = y)$, and for $P(Y = y|X = x)$.
5. Define what it means for random variables to be independent, and show how to recognize this in several ways.
6. Analyze complex systems that are connected in series and in parallel.

Part II

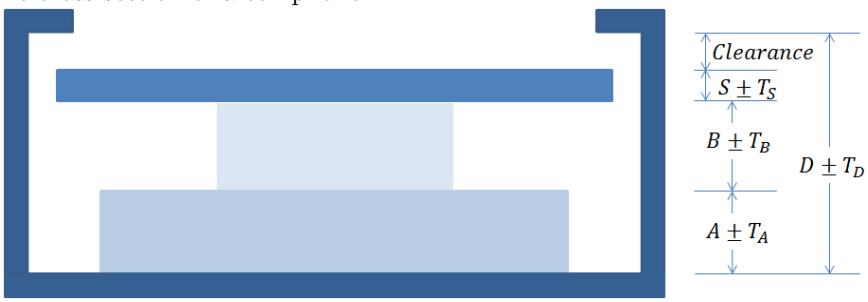
Statistical Inference

Chapter 4

The Central Limit Theorem

4.1 Tolerance Stack-up

Most products consist of several components. Each of the components has characteristics that affect the performance of the product. For example, a cell phone screen is held in place by components below the screen that press upward on the back of the screen while it is held in place by the edges of the cover. The specifications for the phone require a *negative* gap between the edge of the cover and the top of the components behind the screen. The screen is bent slightly by this pressure, which keeps it firmly in place. The following figure is a schematic of the cross section of a cell phone.



The design specifications are summarized here:

Screen ($S \pm T_S$)**:** 0.100 ± 0.005 inches

Component ($A \pm T_A$)**:** 0.200 ± 0.01 inches

Component ($B \pm T_B$)**:** 0.225 ± 0.012 inches

Inner Depth of the Case ($D \pm T_D$)**:** 0.500 inches

Clearance: Depends on the dimensions of the other components

Note that the clearance must be a negative distance to provide the necessary tension, and the components are press fit together.

When the components are manufactured, their actual dimensions are random. Some of them will be a little smaller and some a little larger. What happens to a product if all the components are unusually large or unusually small?

Have you ever had a nut and bolt that should be able to fit together that did not? This can happen if the nut's inner diameter was too small or the bolt's outer diameter was too large. A similar problem can occur if, for example, a

cell phone has a case that is smaller than normal and the internal components are larger than normal.

In this chapter, you will discover how to use the variability in the components' dimensions to find the total variability in a product's dimensions. For example, what proportion of the cell phones illustrated above will have a cross section that fits together properly? What proportion will be defective?

4.1.1 Propagation of Error

Definition 4.1. The difference between a measured value and true value is called the **error in the measured value**. We can think of the error as consisting of two parts, namely the **systematic error or bias** (we're always off by this amount) and the **random error**. We say that our measurement technique is **unbiased** if the systematic error is zero. The smaller the bias, the more **accurate** our measurement. If our measuring technique has a small random error, we say it is **precise**. The smaller the random error, the more precise our measurement technique. We'll use the standard deviation σ to help us determine precision. Scientists generally refer to σ as the **uncertainty**. We can summarize the above by writing

$$\text{measured value} \;=\; \text{true value} \;+\; \text{bias} \;+\; \text{random error}.$$

Problem 4.1 Suppose someone weighs themselves once on a bathroom scale and observed a weight of 184 pounds. After they stepped off the scale, it showed a value of 6 pounds.

1. If possible, estimate the *uncertainty* in this measurement. If it is not possible, explain why not.
2. If possible, estimate the *bias* in this measurement. If it is not possible, explain why not.
3. Suppose now that someone weighed themselves repeatedly on a bathroom scale and observed the following values (in pounds):

   $$182, 185, 182, 183.$$

   Each time the person steps off the scale, the scale still shows a value of 6 pounds.

   (a) If possible, estimate the *uncertainty* in these measurements. If it is not possible, explain why not.

   (b) If possible, estimate the *bias* in these measurements. If it is not possible, explain why not.

Problem 4.2 Researchers made five measurements of the carbon content (in ppm) of a silicon wafer whose true carbon content was known to be 1.1447 ppm [**?**]. The observed values are:

$$1.0979, 1.0870, 1.0711, 1.0825, \text{ and } 1.0730$$

1. If possible, estimate the *uncertainty* in these measurements. If it is not possible, explain why not.
2. If possible, estimate the *bias* in these measurements. If it is not possible, explain why not.

"Piled Higher and Deeper" by Jorge Cham

The next few problems will give us some needed formulas to help us combine means, variances, and standard deviations of several random variables. Let's start by looking at a constant multiple of a single random variable.

Problem 4.3 Let X be a random variable. Let a be a constant. We need to understand how multiplying a random variable by a constant affects the mean, variance, and standard deviation.

1. Show that $E[aX] = aE[X]$, assuming X is a discrete random variable.
2. Show that $\text{Var}[aX] = a^2\text{Var}[X]$, assuming X is a discrete random variable. Remember that the variance is $\text{Var}[aX] = E[(aX - E[aX])^2]$.
3. Show that $\sigma_{aX} = |a|\sigma_X$.
4. How would your computations above change if X were a continuous random variable instead of a discrete random variable.

What happens if we decide to add two random variables together?

Problem 4.4 Let X and Y be random variables. Feel free to assume that they are discrete random variables, as the computations for continuous random variables will follow immediately.

1. Show that $E[X + Y] = E[X] + E[Y]$.
2. Show that $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$.
3. If two variables are independent, then the covariance is always zero. Suppose that X and Y are independent, and then give a formula for the variance of $X + Y$, and give a formula for the standard deviation of $X + Y$.

We would really like to know what happens when we average together several random variables. For this reason, we need to be able to compute the mean, variance, and standard deviation of expressions such as $\frac{1}{3}X + \frac{1}{3}Y + \frac{1}{3}Z$. This next problem has you obtain these values for $aX + bY$, and then for $aX + bY + cZ$.

Problem 4.5 Let X, Y, and Z be independent random variables, and suppose that a, b, c are constants.

"Piled Higher and Deeper" by Jorge Cham

1. Compute $E[aX + bY + cZ]$. Give your answer in terms of μ_X, μ_y, and μ_Z.
2. Show that $\text{Var}[aX + bY + cZ] = a^2\sigma_X^2 + b^2\sigma_Y^2 + c^2\sigma_Z^2$. Then give the standard deviation of $aX + bY + cZ$.
3. Generalize your work. If we have n independent random variables, then state the expected value, variance, and standard deviation of

$$\sum_{i=1}^{n} c_i X_i = c_1X_1 + c_2X_2 + c_3X_3 + \cdots + c_nX_n.$$

Problem 4.6 During an assembly process for a part, there are three tasks that must occur in sequence. The time (in minutes) to complete each of the three tasks is modeled by the random variables T_1, T_2, and T_3, respectively. The expected values of these random variables are:

$$E[T_1] = 2.1$$
$$E[T_2] = 3.7$$
$$E[T_3] = 3.3$$

Use this information to answer these questions.

1. The costs (in US$ per minute) to perform the three tasks are $2.01, $7.11, and $0.98, respectively. What is the expected cost to assemble one part?
2. Find the expected value of the linear combination:

$$\frac{1}{3}T_1 + \frac{1}{3}T_2 + \frac{1}{3}T_3$$

   Interpret the result.

"Piled Higher and Deeper" by Jorge Cham

In the previous problems, you should have shown that if a is a constant and X is a random variable, then the uncertainty of aX is

$$\sigma_{(aX)} = |a|\sigma_X$$

If there are several measurements $(X_1, X_2, \ldots, X_n)$ that are independent and the values $c_1, c_2, \ldots c_n$ are constants, then you should have obtained the uncertainty of the linear combination $c_1X_1 + c_2X_2 + \cdots + c_nX_n$ to be

$$\sigma_{(c_1X_1+c_2X_2+\cdots+c_nX_n)} = \sqrt{\sum_{i=1}^{n} c_i^2\sigma_{X_i}^2} = \sqrt{c_1^2\sigma_{X_1}^2 + c_2^2\sigma_{X_2}^2 + \cdots + c_n^2\sigma_{X_n}^2}.$$

If the measurements $(X_1, X_2, \ldots, X_n)$ are not independent, then it is very difficult to accurately assess the uncertainty of a linear combination.[1] It is possible to compute an upper bound on the uncertainty of this linear combination, namely an upper bound is

$$\sigma_{(c_1X_1+c_2X_2+\cdots+c_nX_n)} \leq \sum_{i=1}^{n} |c_i|\,\sigma_{X_i} = |c_1|\,\sigma_{X_1} + |c_2|\,\sigma_{X_2} + \cdots + |c_n|\,\sigma_{X_n}. \quad (4.1)$$

Let's look at some applications of what we've discovered.

Problem 4.7 A surveyor needs to measure the perimeter of a rectangular property.

1. One surveyor measures the lengths of two adjacent sides of the property and doubles their sum, so $P = 2L + 2W$. If the length is measured as 102 with uncertainty 0.3 meters and the width is measured as 75 with uncertainty 0.2 meters, find the estimate of the perimeter and the uncertainty of this measurement. Assume that the measurements are independent.

2. A different surveyor measures the length of each side, and then sums the 4 results, so $P = L_1 + L_2 + W_1 + W_2$. They obtain the same measurements of 102 m for each length with uncertainty .3 m, and 75 m for each width with uncertainty .2 m. Show that the expected perimeter is exactly the same, but the variance is smaller.

---

[1] This can be done, in theory, but it requires a lot of information to which you will probably not have access in practice.

Problem 4.8 The body temperature of a healthy person is a random variable with $\mu = 36.8$ Celsius with an uncertainty of 0.1 Celsius. Recall that the formula for converting from Celsius to Fahrenheit is $F = \frac{9}{5}C + 32$. You could look at this as a random variable $Y = aX + b$.

1. If we add 32 to every number, how will the mean and standard deviation change?
2. Find the expected body temperature in degrees Fahrenheit.
3. Give the uncertainty of the measurement in degrees Fahrenheit.

Recall the cell phone design described on page **??**.

Screen ($S \pm T_S$): 0.100 ± 0.005 inches

Component ($A \pm T_A$): 0.200 ± 0.01 inches

Component ($B \pm T_B$): 0.225 ± 0.012 inches

Inner Depth of the Case ($D \pm T_D$): 0.500 inches

Clearance: Depends on the dimensions of the other components

For each pair of numbers above, consider each as $\mu \pm \sigma$.

Problem 4.9 Let C be the clearance.

1. Write an equation for C in terms of S, A, B, and D, and then find the expected clearance.
2. Suppose the measurements of S, A, B, and D are independent. Give the uncertainty of the estimate of the clearance.
3. Suppose the measurements of S, A, B, and D are *not* independent. Give an upper bound for the uncertainty of the estimate of the clearance.
4. What do you conclude about this product?

Problem 4.10 The grades in a particular Engineering course are determined as follows:

Activity	Weight
Preparation	10%
Homework	20%
Quizzes	5%
Group Project	15%
Tests	50%

The teacher estimates a student's grades will follow the pattern below:

Activity	Score $\pm$ Uncertainty
Preparation	$80\% \pm 12\%$
Homework	$75\% \pm 20\%$
Quizzes	$60\% \pm 15\%$
Group Project	$90\% \pm 10\%$
Tests	$70\% \pm 20\%$

Estimate the expected grade in the class for a particular student and give the uncertainty of this estimate.

4.2 The Sample Mean

Problem 4.11: The Mean of Several Random Variables Let $X_1, X_2, \ldots, X_n$ be independent random variables with the common mean μ and the same variance σ^2. We want to consider the random variable which is the average of these, namely

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

1. Compute the mean of the random variable $\bar{X}$. Give your answer in terms of μ.
2. Compute the variance of the random variable $\bar{X}$, and give your answer in terms of n and σ^2.
3. Show that the standard deviation of $\bar{X}$ is $\dfrac{\sigma}{\sqrt{n}}$.

Problem 4.12 The length of a uranium rod was measured 4 times. The mean of the measurements was $\bar{X} = 2.178$ meters. The standard deviation of the observed data was determined to be: $s = 0.006$ meters.

1. Which of the following is closest to the uncertainty of the measurement of $\bar{X}$? Justify your answer.
   - 0.006
   - 0.003
   - 0.0015
   - 0.001
2. The length of a different uranium rod was measured once. The measured value was 2.155. Which of the following is closest to the uncertainty of the measurement of $\bar{X}$? Justify your answer.
   - 0.006
   - 0.003
   - 0.0015
   - 0.001
3. If we measured the length of a rod 25 times, what would you use as an approximate for the uncertainty of the measurement of $\bar{X}$?

Problem 4.13 Using capture-recapture techniques, on nine sequential days, biologists estimated the number of fish in a particular pond to be: 1275, 1143, 1500, 1214, 1200, 1333, 1412, 1417, and 1159.

1. The biologists will use the mean as an estimator of the number of fish in the pond. Give the value of this estimator.
2. Find the uncertainty in the estimate of this mean.

Problem 4.14 The article "On Pulsar Distance Measurements and Their Uncertainties" gives the estimated distance to some pulsars, together with their uncertainties.[**?**]

1. The distance to the pulsar J0141+6009 was measured 12 times using a process that has an uncertainty of 0.7 *kpc*. Let $\bar{X}$ represent the mean of the measurements. Find the uncertainty in the estimate of the mean, $\bar{X}$.

2. New technology allows researchers to measure the distance to this pulsar with an uncertainty of 0.35 *kpc*. Suppose this process is repeated 6 times, and $\bar{Y}$ is the mean of the six observations. Find the uncertainty in $\bar{Y}$.

Problem 4.15 Continuing from the previous problem, astronomers hope to reduce the variability in the estimates by combining the information. Two astronomers propose the following ways to combine the data:

Scientist	Proposed Statistic
Astronomer #1:	$\frac{1}{2}\bar{X} + \frac{1}{2}\bar{Y}$
Astronomer #2:	$\frac{12}{18}\bar{X} + \frac{6}{18}\bar{Y}$

- Astronomer #1 argues that the two estimates were done separately, and they should each contribute equally to the final estimate.
- Astronomer #2 claims that since $\bar{X}$ is based on 12 observations and $\bar{Y}$ is only based on 6 observations, weights of $\frac{12}{18}$ and $\frac{6}{18}$ are appropriate.

1. Find the uncertainty in each of the proposed estimates, and determine which is smaller.

2. A third astronomer says that we should use the formula

$$k\bar{X} + (1-k)\bar{Y}$$

where k is some number between 0 and 1. Find the value of k such that the uncertainty of this estimator is minimized. What is the minimum uncertainty of this weighted average?

4.2.1 Bernoulli Distribution

The **Bernoulli distribution** was named for Jacob Bernoulli,[2] a Swiss scientist. This distribution applies where there is a discrete probability with only two outcomes. For clarity, we will call these outcomes "success" and "failure." Success indicates that a particular event has occurred. Failure indicates that it did not occur.

Success and failure are not value judgments. Success does not imply that the event is desirable, only that it occurred. For example, the event may be that a component in a computer system fails.

[2] b. 27 December 1654 – d. 16 August 1705

We define a Bernoulli random variable Y so that

$$Y = \begin{cases} 1, & \text{if the event occurred} \\ 0, & \text{if the event did not occur} \end{cases}$$

The PDF of a Bernoulli distribution is summarized in the following table, where p is a number between 0 and 1:

y	$P(Y = y)$
1	p
0	?

Problem 4.16 Do the following.

1. Fill in the missing probability in the table above.
2. Use the PDF in the table above to compute the mean and variance of a Bernoulli random variable. Your answer should be in terms of p.
3. Use your formula to compute the mean and variance of a Bernoulli random variable representing the occurrence of a defect in a computer chip, where an average of 4 chips per million are defective.

Problem 4.17 Let $X_1, X_2, \ldots, X_n$ be n independent Bernoulli random variables, each with the exact same probability of success p. Let S be the sum of the X_i's, and let $\bar{X}$ be the average of the X_i's, so we have

$$S = \sum_{i=1}^{n} X_i \quad \bar{X} = \sum_{i=1}^{n} \frac{1}{n} X_i.$$

1. Find the mean and variance of the random variables S and X. Your answers will be in terms of p, $(1-p)$, and n.
2. Suppose that $n = 100,000$ computer chips are tested. Assume the probability of observing a defective chip is 4 per million. Find the mean and variance of the number of defects observed. Also find the mean and variance of the average number of defects observed.

4.2.2 Normal Distribution

Recall we a say a random variable X is normally distributed with mean μ and standard deviation σ, and we write $N(\mu, \sigma)$ if the probability density function for X is the function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

Remember, we can compute the probability that $a \leq X \leq b$ by computing the integral

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

It can be shown that a linear combination of two (or more) normal random variables is normally distributed. Use this information to answer the following questions.

Problem 4.18 Complete each of the following:

1. If X is a normal random variable with mean $\mu = 3$ and variance $\sigma^2 = 4$, what is the distribution, mean, and variance of the random variable $W = 5X + 7$?

2. Compute the probability $P((3-2) \leq X \leq (3+2)) = P(1 \leq X \leq 5)$ using an integral, and then compute the probability $P((22-10) \leq W \leq (22+10)) = P(12 \leq W \leq 32)$ using an integral.

3. Now generalize your result. Suppose X is a normal random variable with mean μ and variance σ^2, and the values a and b are constants. What is the distribution, mean and variance of the random variable $W = aX + b$?

4. Make a guess for the probabilities $P(\mu_X - \sigma_X \leq X \leq \mu_X + \sigma_X)$ and the probability $P(\mu_W - \sigma_W \leq W \leq \mu_W + \sigma_W)$.

Problem 4.19 Let $X_1, X_2, \ldots, X_n$ be independent normal random variables with common mean μ and variance σ^2.

1. Find the distribution of the random variable:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

   In addition to the name of the distribution, be sure to state the mean and variance.

2. How does the mean of the random variable $\bar{X}$ compare to the mean of the random variable X?

3. How does the standard deviation of the random variable $\bar{X}$ compare to the standard deviation of the random variable X?

Problem 4.20 Suppose that X is a normal random variable with mean μ_X and variance σ_X^2. Assume that Y is a normal random variable with mean μ_Y and variance σ_Y^2.
If X and Y are independent, find the distribution (indicate the distribution, the mean, and the standard deviation) of the following:

1. $X + Y$

2. $X - Y$

Problem 4.21 The monthly revenue, R, for a company follows a normal distribution with mean $\mu_R = \$152\,000$ and standard deviation $\sigma_R = \$17\,500$. The monthly costs, C, is modeled as a normal random variable with mean $\mu_C = \$127\,000$ and standard deviation $\sigma_R = \$31\,000$. The monthly profit P is given by the equation

$$P = R - C$$

Use this information to answer the following questions.

1. Find the mean and standard deviation of the monthly profit, P.

2. Find the probability that the next month's profit will be negative.[3] You'll need to compute an integral for this one. Feel free to use WolframAlpha to compute the integral.

Problem 4.22 Computer code is said to be *cloned* if there are multiple segments of code that perform the same function. Clones are a big concern in programming. Cloned code adds to the complexity of a system and increases maintenance and testing costs.
The process of detecting clones follows a two-step algorithm, where the steps are performed sequentially.

> First, source code is transformed into an internal format. Second, a more or less sophisticated comparison algorithm is then performed on the internal data.[**?**]

Suppose the time required to perform the first step on a particular computer is normally distributed with a mean of 10 milliseconds (ms) and a variance of 5 ms^2. The time required to perform the second step follows a normal distribution with mean 14 ms and variance 9 ms^2.

1. Find the mean and standard deviation of the total time required to run the algorithm.

2. Find the probability that the time required to run the algorithm exceeds 30 ms. Again, you'll need an integral for this one.

For much of the remainder of the semester, we'll be needing to compute probabilities for normal distributions. The integral involved in the normal probability computations cannot be evaluated by hand, but instead can be approximated using numerical schemes. If we have access to a numerical integration tool, we can just do all our computations using that tool. Because no such tools existed 100 years ago, instead people decided to create a way to standardize the computations and compute probabilities by looking up numbers on a table. They created z-scores as a way to take any normal distribution with mean μ and standard deviation σ and return a new distribution with a simpler mean and standard deviation.

If X is a random variable, then we can consider the random variable Z given by the formula

$$Z = \frac{X - E[X]}{\sqrt{Var[X]}} = \frac{X - \mu_X}{\sigma_X}.$$

This is just a z-score, which means it measures the number of standard deviations away from the mean.

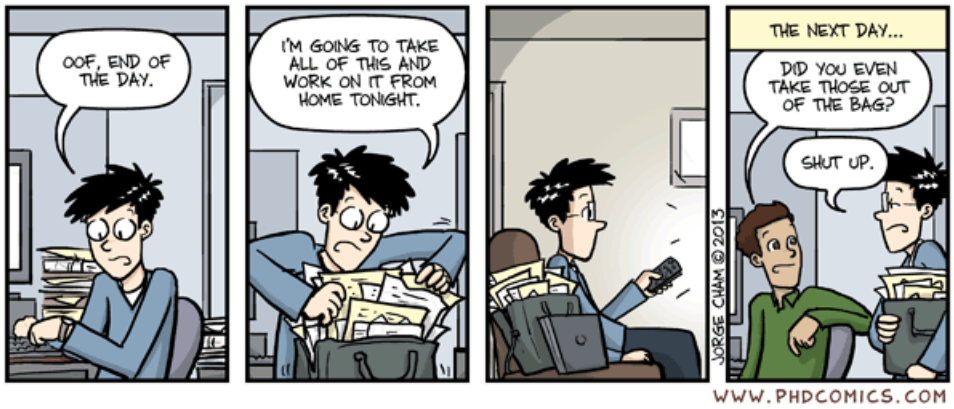
Problem 4.23 Suppose X is a random variable. Let Z be the random variable given by $Z = \dfrac{X - \mu_X}{\sigma_X}$.

1. Find the mean, variance, and standard deviation of the random variable Z.

2. If X is a normal random variable, then Z is a normal random variable. Draw a graph of the pdf for the normal random variable Z. You can check if you are correct by heading to the URL The http://byuimath.com/apps/normprob.html. The random variable Z is called the standard normal random variable.

[3] What does it mean if the profit is negative?

"Piled Higher and Deeper" by Jorge Cham



Given any random variable X, we can always standardize the random variable using z-scores. This will shift the mean to zero, and make the variance equal to 1. If we're working with a normal distribution, then instead of having to integrate $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$ to perform probability computations, we can instead integrat $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The integral still cannot be done by hand, but now we only have to focus on one function, instead of a different function for any given μ and σ. This simplification allows us to quickly make probability computations using z-scores.

We have several ways to find these probabilities. An applet developed at BYU-Idaho is an excellent tool for doing calculations with the normal probability distribution. Go to and explore this applet. Explore its capabilities before proceeding.

The normal probability applet will play a significant role in this class. You may consider bookmarking the page in your web browser.

Problem 4.24 Look at the distribution shown in the applet at http://byuimath.com/apps/normprob.html.

1. At what value is this function maximized? Identify any points of inflection.
2. What is the mean of the distribution shown in the applet? What is the standard deviation? Explain how you made this determination.
3. Suppose random variable Z follows a standard normal distribution. Use the applet to do the following:
   (a) Find $P(Z > 1)$, $P(Z < 1)$, $P(Z = 1)$, and $P(Z < -1)$.
   (b) Compute $P(Z \leq 0.4)$. What percentile corresponds to $Z = 0.4$?

Problem 4.25 Suppose random variable Z follows a standard normal distribution. Use the applet to do the following:

1. Find the value of z such that $P(Z < z) = 0.75$.
2. Find the value of z such that $P(Z < z) = 0.5$.
3. Find the value of z such that $P(Z > z) = 0.95$.
4. Find the value of z such that $P(Z > z) = 0.05$.

5. Find the value of z that corresponds to the 90^{th} percentile of the curve.

Problem 4.26 Suppose random variable Z follows a standard normal distribution.

1. Find $P(-1 < Z < 1)$. Give your answer as a decimal, accurate to four decimal places.
2. Find $P(-2 < Z < 2)$. Give your answer as a decimal, accurate to four decimal places.
3. Find $P(-3 < Z < 3)$. Give your answer as a decimal, accurate to four decimal places.
4. Express your answers above as percentages, rounded to the nearest percent (though round the third one to the nearest tenth of a percent). Then fill in the blanks below.
   - _____% of the values are within 1 standard deviation of the mean.
   - _____% of the values are within 2 standard deviations of the mean.
   - _____% of the values are within 3 standard deviations of the mean.

Let's now compute probabilities for a normal distribution that is not the standard normal distribution. The key is to just compute z-scores.

Problem 4.27 Ball bearings are manufactured in a process where the diameters are normally distributed.[4] The mean of the diameters is 2.755 cm and the standard deviation is 0.007 cm. Let X be a variable representing the diameter of a randomly selected ball bearing. Use this information to answer the following questions.

1. If a ball bearing has a diameter of 2.769 cm, how many standard deviations *above* the mean is this value? What is the z-score associated with a ball bearing that has a diameter of 2.769 cm?
2. What is the probability that a randomly selected ball bearing will have a diameter that exceeds 2.769 cm? You can use the applet to do this computation. You'll need a z-score first.
3. What is the probability that a randomly selected ball bearing will have a diameter smaller than 2.750 cm?
4. What is the value of z that corresponds to the 10^{th} percentile?
5. What is the 10^{th} percentile of the diameters of these ball bearings? In other words, if you know z from the previous part, and you know μ and σ, what is x?

We can use the standard normal distribution to work with z-scores, and then we can use the formulas $z = \dfrac{x - mu}{\sigma}$ and/or $x = \mu + z\sigma$ to help us work with x values. If we need a specific percentile, the normal probability applet will give us the corresponding z-score instantly. We then just have to convert this z-score to an x-value to interpret the results.

[4] The normal distribution occurs naturally in many processes, including in manufacturing.

Problem 4.28 Continue from the previous problem.

1. What is the 95^{th} percentile of the diameters of these ball bearings?
2. The upper specification limit[5] for the diameter of these bearings is 2.778 cm. What proportion of the ball bearings are too large?
3. The process can be re-calibrated so that the mean is set to 2.750 cm. If this is done, what proportion of the ball bearings will exceed the upper specification limit?

Problem 4.29 In a paint application process, the thickness of the paint is normally distributed with mean 120 microns and variance 21 microns. The lower specification limit is 90 microns.[6]

1. In what proportion of the products will the paint be too thin?
2. We can recalibrate the paint application process so that we can change the mean from 120 microns to 130 microns. If we do this, in what proportion of the products will the paint be too thin?
3. Find a value μ so that if we recalibrate the machine to have this value of μ, then 99.9% of the products will be above the lower specification limit.

Problem 4.30 In a bottling operation for a cola manufacturer, soda is injected into bottles that are labeled to contain 12 fluid ounces (fl. oz.) The mean amount of soda injected into the bottles is 12.050 fl. oz., and the standard deviation of the process is 0.022 oz. You can assume that the distribution of amount of soda is normal.

1. If we assume that the machine is properly calibrated so the mean is as claimed, what is the probability that a randomly selected bottle will be under-filled (i.e., the volume of soda in the bottle is less than 12 fl. oz.)?
2. Suppose we gather 100 bottles and find that 20 are under filled. One reason this could occur is that the machine is improperly calibrated. Give another reason why this could occur.

4.3 The Distribution of the Sample Mean - The Central Limit Theorem

When quality control wants to check if a machine is working properly, they will test several products. In the soda can example above, quality control might grab 100 bottles, and then measure the contents in each bottle. They'll average the results together, and then use that information to decide if the machine need to be recalibrated. Wait, this means they are looking at the distribution of averages, namely $\bar{X}$, not the distribution of the actual volume X. We've already noted that the average of a collection of independent normal distributions is

[5] This is the maximum allowable value, based on the engineering specifications for the product.

[6] This is the minimum acceptable paint thickness.

still normally distributed, and we've shown that the mean of $\bar{X}$ is the same as the mean of X. However, the standard deviation of $\bar{X}$ is different, namely

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}.$$

We can use this information to perform probability computations with $\bar{X}$ instead of with X. Remember that $\bar{X}$ is normally distributed with mean $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$. This means z-scores are now

$$Z_{\bar{X}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}}.$$

Problem 4.31 In a bottling operation for a cola manufacturer, soda is injected into bottles that are labeled to contain 12 fluid ounces (fl. oz.) The mean amount of soda injected into the bottles is 12.050 fl. oz., and the standard deviation of the process is 0.022 oz. You can assume that the distribution of amount of soda is normal.

1. What is the probability that 1 randomly selected bottle will have a volume that is under 12.04 fl. oz?
2. What is the probability that 4 randomly selected bottles will have a mean volume that is under 12.04 fl. oz?
3. What is the probability that 25 randomly selected bottles will have a mean volume that is under 12.04 fl. oz?
4. What is the probability that 100 randomly selected bottles will have a mean volume that is under 12.04 fl. oz?
5. Suppose we sample 100 bottles and find that their mean volume is 12.043. If the machine is calibrated correctly, what's the probability of observing a sample mean that is this low or lower? Do you think the machine is calibrated correctly?

The problems above are doable because we know that the distribution of the means of normally distributed random variables will still be normally distributed. What if we don't start with a random variable X that is normally distributed? What will the distribution of $\bar{X}$ look like? This question is the most important question in all of statistics. Let's repeat it.

> If X is any random variable, what do we know about the distribution of the means $\bar{X}$?

We've already shown, using our expected value computations, that $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$, regardless of the distribution of X. The surprising answer to the question above, and the entire reason statistics exists, is that the answer to the question above can be given. The normal curve is the answer.

Theorem 4.2 (The Central Limit Theorem)**.** *Let X_1, X_2, ..., X_n be independent random variables with the same distribution. Then the mean $\bar{X}$ of these random variables, under appropriate conditions, will be approximately normally distributed. The appropriate conditions are either of the following:*

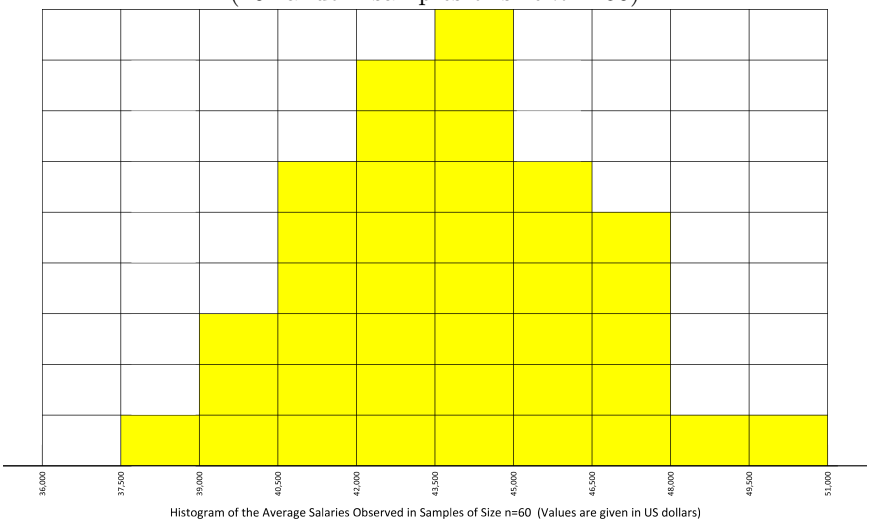
- *The original distribution of the X_i's is normal. In this case, regardless of the sample size n the distribution of $\bar{X}$ is normal.*

- *If the original distribution is large enough, then as the sample size n gets larger, the distribution gets closer and closer to normal.*

How large of a sample size n is large enough to use the central limit theorem. The real answer to that questions requires more mathematics than we will introduce in this class. However, for most applications if we require $n \geq 30$, then the conclusions of the central limit theorem will apply. This means that if our sample size is 30 or larger, then regardless of what the original distribution looks like, we can use the central limit theorem to perform probability computations.

We did this during the first week of class. Recall the histogram on page 3 that was right skewed, and clearly not normal. However, as part our in class activity we took 40 random samples of size $n = 60$ from the population of starting salaries. An example of a histogram of their averages is shown below.

Mean salaries of recent BYU-Idaho graduates
(40 random samples of size $n = 60$)



Histogram of the Average Salaries Observed in Samples of Size n=60 (Values are given in US dollars)

Notice that the histogram of the averages is bell shaped. If we took 1000 samples of size $n = 60$ and made a histogram of these, we would see an even better bell shape appear. We'll explore this in class together by looking at the applet at

- http://onlinestatbook.com/stat_sim/sampling_dist/.

Problem 4.32 In the BYU-I student salaries problem, the population mean and standard deviations are $\mu = \$43,662$ and $\sigma = \$20,558$.

1. If we use samples of size $n = 60$, then what is the shape, mean, and standard deviation of the random variable $\bar{X}$?
2. If we repeated sampled group so 60 students, about what proportion of the time would we observe a mean salary that lies above \$48,000? (You'll want to compute a z-score, and then find an area.)
3. Joe collected one sample of size 60 and obtained a mean of $\bar{x} = \$39,984$. What is the probability of obtaining a salary that is this far, or further, from the actual mean?

Problem 4.33 In the BYU-I student salaries problem, the population mean and standard deviations are $\mu = \$43,662$ and $\sigma = \$20,558$.

1. Use the normal probability applet to give a z-score so that $P(-z \leq Z < z) = 0.90$.
2. Give a dollar amount m so that 90% of all sample means from samples of size $n = 60$ will lie between $\mu - m$ and $\mu + m$.
3. Compute the lower bound $\mu - m$ and upper bound $\mu + m$ between which 90% of all average salaries, from samples of size 60, should occur.
4. What formula did you use to get from the z-score to the dollar amount m?

Problem 4.34 Repeat parts 1 though 3 of the previous problem twice, but this time use probabilities of 0.95 and 0.99 instead of 0.90. When you are done, you should have two different z-scores, as well as two different values for m (called the margin of error) and two different intervals with lower and upper bounds.

In the previous two problems, we were able to take a probability p, from it compute a z-score, and then from that obtain a value m, called the margin of error, and then from that give an interval $(\mu - m, \mu + m)$ centered about the population mean. This interval gives us a lower and upper bound for knowing what sample means we should expect.

Problem 4.35 A supplier sells synthetic fibers that have an average breaking strength of 28 lbs with a standard deviation of 8 lbs. When a large shipment of the fibers arrive at Jimmy's plant, he has his quality control randomly select 30 of the fibers for testing. They want to be 95% confident that the fibers they paid for have the stated breaking strength.

1. Find a z-score so that $P(-z \leq Z \leq z) = 0.95$, and then use this to give an interval of acceptable values between which 95% of all samples of size 30 fibers
2. His quality control team find the average breaking strength of these 30 fibers to be 26 lbs.

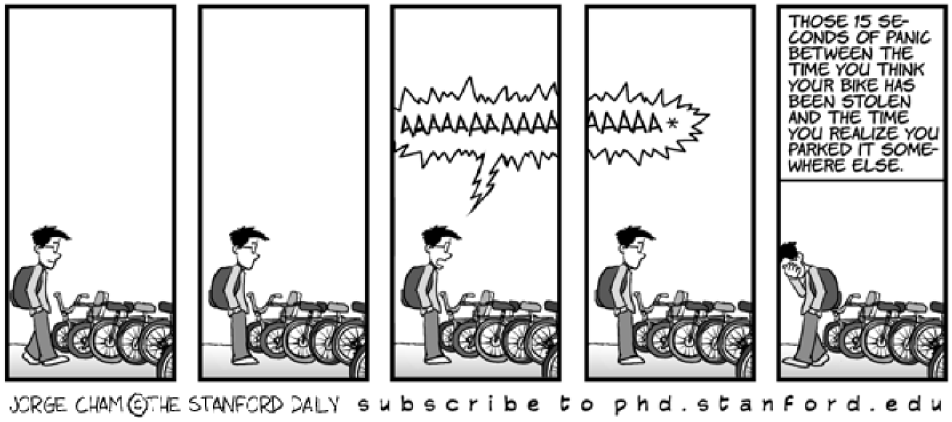
Remember, the central limit theorem applies whenever the original population is normally distributed, or the sample size is large enough. Once we have checked to make sure either of these conditions apply, we can use the normal distribution to perform probability computations. The next problem summarizes these ideas and concludes the chapter.

Problem 4.36 In each problem below, compute the requested probability, or explain why you cannot.

1. In a random sample of 53 concrete specimens, the average perocity (in percent) was found to be 20.3. The process is known to have an expected perocity of $\mu = 21.3$ with a standard deviation of 3.2. What is the probability of obtaining an expected perocity as low, or lower, than what was observed.
2. In a factory that makes batteries, history has shown that the capacities of the batteries are normally distributed with an expected capacity of 175 ampere-hours and a standard deviation of 14. A quality control team randomly samples 10 batteries and finds the average capacity to be 170. What's the probability that a quality control team would observe a sample average of 170 or below, assuming that the machine properly creating batteries.

3. A supplier sells synthetic fibers that have an average breaking strength of 28 lbs with a standard deviation of 8 lbs. When a large shipment of the fibers arrive at Jimmy's plant, he grabs 4 of the fibers and breaks them, finding they have an average breaking strength of 30 lbs. What's the probability of obtaining an average breaking strength of 30lbs or higher.

"Piled Higher and Deeper" by Jorge Cham



Chapter 5

Inference for a Single Population Mean

5.1 One Population Mean: σ Known

During an election in the United States, many polls are conducted to determine the attitudes of likely voters. Poll results are usually reported as percentages. For example, a poll might state that 49% favor the Republican candidate and 51% favor the Democratic candidate.

Polls always include a margin of error. The **margin of error** gives an estimate of the variability in the responses. A common value for the margin of error in political polls is 3%.

When we incorporate the margin of error, we estimate that the true proportion of people who favor the Republican candidate is 49% 3%, or in other words between 46% and 52%. For the Democratic candidate, we get 48% to 54%. There is a lot of overlap in these numbers. In this case, the political race is too close to know who might win.

In this chapter, we will explore the margin of error and its role in estimating a parameter.

5.1.1 Point Estimators

We have learned about several statistics. Remember, a statistic is any number computed based on data. The sample statistics we have discussed are used to estimate population parameters.

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Variance	s^2	σ^2
Standard Deviation	s	σ
$\vdots$	$\vdots$	$\vdots$

The statistics above are called **point estimators** because they are just one number (one point on a number line) that is used to estimate a parameter. Parameters are generally unknown values. Think about the mean. If μ is unknown, how do we know if $\bar{x}$ is close to it?

The short answer is that we will never know "for sure" if $\bar{x}$ is close to μ. This does not mean that we are helpless. The laws of probability and the normal distribution provide a way for us to create a range of plausible values for a parameter (e.g. μ) based on a statistic (e.g. $\bar{x}$). This is called an **interval estimator**.

5.1.2 Confidence Interval for One Mean, σ Known

A point estimator gives one specific value as an estimate of a parameter. An **interval estimator** is a range of plausible values for a parameter. We can create an interval estimate by starting with a point estimate and adding or subtracting the margin of error.

In the political poll mentioned above, the point estimate for the support of the Republican candidate is 49%. By adding and subtracting the margin of error, we get the interval estimate: 46% to 52%.

Problem 5.1 Answer the following:

1. What is the **68-95-99.7% Rule**?
2. Approximately what percentage of data from a bell-shaped distribution will lie within two standard deviations of the mean?
3. Recall that use the notation $\bar{X}$ to talk about the sample mean. The two conditions under which we know that $\bar{X}$ will be approximately normal are

   (i) If the population is approximately _____, or

   (ii) If the _____ _____ is large.
4. If the original population from which the data (X) were drawn has mean μ and standard deviation σ, what will be the mean and standard deviation of $\bar{X}$? In other words, state $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ in terms of μ and σ.

Note:
The **68-95-99.7% Rule** is also known as the **Emperical Rule**.

Problem 5.2 The sample mean, $\bar{X}$, estimates μ. We need to develop a way to estimate how close $\bar{X}$ is expected to be to μ. Let Z be a standard normal random variable.

1. Use the normal probability applet to find the value of z^* that makes the following statement true:

   $$P\left(-z^* < Z < z^*\right) = 0.95.$$

   How does this compare with the 68-95-99.7% Rule?
2. Recall that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. Rather than writing $P\left(-z^* < Z < z^*\right)$, we could write

   $$P\left(-z^* < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z^*\right).$$

   Solve the inequality above for μ so that we can instead write is as the probability $P(\ ?\ <\ \mu\ <\ ?\ ) = 0.95$.

   If you get stuck, break the expression into two inequalities and work with each separately. Isolate μ in each and then combine your results.
3. The lower and upper bounds you found in question 5.2(2) define the **95% confidence interval**.[1] Assume that $\bar{x} = 11.36$, $\sigma = 1.70$ and $n = 5$.

[1] We typically write a confidence interval as the lower and upper limits, separated by commas, enclosed in parentheses. If the lower limit was 35.97 and the upper limit was 41.26, we would give the confidence interval as: (35.97, 41.26).

Evaluate the upper and lower limits you found in the previous problem. These values give the 95% confidence interval for the population mean. Present your result in the form:

$$(lower,\ upper)$$

Problem 5.3 Do the following.

1. Use the normal probability applet to find the value of z^* that makes the following statement true:

$$P\left(-z^* < Z < z^*\right) = 0.90. \tag{5.1}$$

2. Use the mean, standard deviation, and sample size given in problem 5.2(3) to compute the 90% confidence interval for the population mean.

3. Compare this confidence interval to the one you obtained in question 5.2(3). Which interval is larger?

Problem 5.4 Do the following.

1. Use the normal probability applet to find the value of z^* that makes the following statement true:

$$P\left(-z^* < Z < z^*\right) = 0.99 \tag{5.2}$$

2. Use the mean, standard deviation, and sample size given in question 5.2(3) to compute the 99% confidence interval for the population mean.

3. Compare this confidence interval to the one you obtained in question 5.2(3). What do you conclude?

Problem 5.5 Generalize the previous three problems to write a general expression for a confidence interval for a mean μ. Write z^* instead of a number. Your answer should be of the form

$$\left(\text{Point Estimate} - \text{Margin of Error}, \text{Point Estimate} + \text{Margin of Error}\right).$$

What is the point estimate of this confidence interval. What is the margin of error.

"Piled Higher and Deeper" by Jorge Cham

Problem 5.6 In questions 5.2(1), 5.3(1), and 5.4(1), you found values for z^* that led to various confidence intervals. Summarize the values of z^* that you found by completing the following table, and then add a few more.

Confidence Level	z^*
90%	?
95%	?
99%	?
80%	?
93%	?
97%	?

Problem 5.7 For a fair, six-sided die, the standard deviation of the values rolled is

$$\sigma = \sqrt{\frac{35}{12}}$$

If a die is rolled several times, and the observed values are:

3 1 2 1 1 2 4 1 1 2 6 5 6 3 4 2 2 5 6 2 6 4 1 1 5

1. Compute a 93% confidence interval for the population mean. Interpret your result.
2. Does the 93% confidence interval you computed contain the true population mean? Explain your reasoning.

Problem 5.8 For a fair, six-sided die, the standard deviation of the values rolled is

$$\sigma = \sqrt{\frac{35}{12}}$$

If a die is rolled several times, and the observed values are:

3 1 2 1 1 2 4 1 1 2 1 5 1 3 4 2 2 5 2 2 6 4 1 1 5

1. Compute a 93% confidence interval for the population mean. Interpret your result.
2. Does the 93% confidence interval you computed contain the true population mean? Explain your reasoning.

Problem 5.9 What are the assumptions (or requirements) that must be satisfied to compute a confidence interval for the mean of a population? Are these assumptions satisfied for the confidence interval you computed in question 5.7?

5.1.3 Sample Size Calculations

It is important for health care administrators to know the mean hospital costs for patients who have coronary artery bypass graft (CABG) surgery. A large hospital is planning a study to determine their mean costs for patients who have CABG surgery.

A study will be conducted in which the charts of patients who had CABG surgery will be sampled, and their hospital costs will be recorded. For budgetary reasons, the hospital administrators do not want to collect a sample that is too large. However, if the sample size is not large enough, the confidence interval will be too wide to be useful as a planning tool.

After a discussion among the senior administration, they have determined that they want to estimate the mean hospital costs of CABG surgery within \$2000 (i.e., plus or minus \$2000.) In other words, they want the confidence interval for the true mean to have a margin of error of \$2000 dollars.

Recall the equation for the confidence interval is:

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right) \tag{5.3}$$

The part of the equation that is added to and subtracted from $\bar{x}$ is called the **margin of error**. We will denote the margin of error by the letter m.

$$m = z^* \frac{\sigma}{\sqrt{n}} \tag{5.4}$$

Notice that a confidence interval is of the form:

$$\begin{array}{ccc} \textit{Point} & \pm & \textit{Margin} \\ \textit{Estimate} & & \textit{of Error} \end{array}$$

To use this formula, the parameter σ must be given to us. It is the true standard deviation of the data you are observing. If you do not know σ, you can estimate it using the standard deviation reported in a previous study or by conducting a pilot study. A study published by another hospital [**?**] reported that the standard deviation of the costs for CABG surgery was \$28,705.

Problem 5.10 In the following questions, you will compute the margin of error, m, for a future study of the hospital costs for CABG surgery. The hospital administrators want to use a 95% level of confidence. Assume the standard deviation can be estimated to be $\sigma = \$28,705$.

1. If the hospital administrators want to be 95% confident in the results, what should the value of z^* be?
2. If the hospital collects a sample of $n = 100$ patients' costs, what would the margin of error be?
3. If the hospital collects a sample of $n = 1000$ patients' costs, what would the margin of error be?
4. What's the smallest sample size needed to obtain a margin of error of at least \$2000? [Hint: solve for n in the margin of error equation.]
5. Should you round up or round down when computing sample sizes?

Problem 5.11 The administration at BYU-Idaho is concerned about the possibility of grade inflation in their courses. Grades in a course are coded on a scale from 0 to 4, as given in the following table:

Grade Point Scale at BYU-Idaho											
A	$A-$	$B+$	B	$B-$	$C+$	C	$C-$	$D+$	D	$D-$	F
4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0	0.7	0.0

Historical records indicate that the standard deviation of the numerical value for the grades students' receive in a course is $\sigma = 0.68$.

Due to the confidential nature of grades, the administration will not release the current value of the mean grade earned on campus. It would be interesting to estimate the current mean grade point average at BYU-Idaho.

You have been asked to help study the issue of grade inflation at BYU-Idaho. A certain number of individual grades will need to be sampled to estimate the mean grade earned at BYU-Idaho.

1. What sample size would be required to estimate the true mean grade earned at BYU-Idaho with 95% confidence and a margin of error of 0.2 grade points?

2. What sample size would be required to estimate the true mean grade earned at BYU-Idaho with 98% confidence and a margin of error of 0.2 grade points?

3. What sample size would be required to estimate the true mean grade earned at BYU-Idaho with 95% confidence and a margin of error of 0.1 grade points?

4. What sample size would be required to estimate the true mean grade earned at BYU-Idaho with 98% confidence and a margin of error of 0.1 grade points?

Problem 5.12 Suppose you are going to collect data from a population in which $\sigma = 2$ units, and you plan to create a confidence interval whose margin of error is no less than .4 units.

1. If you want to be %95 confident, then what's the smallest sample size you can collect.

2. If your budget is limited and you can only collect a sample of size 35, then what confidence level will result in a margin of error of .4 units. (Hint: Start by finding the corresponding z^*, and from that compute the confidence level.)

5.2 Hypothesis Test for One Mean, σ Known

Ethan Allen

A tragic accident on Lake George in New York, USA, called into question the safety regulations for commercial tour boats. On October 5, 2005, a full boat of 47 passengers and 1 crew member began a routine one-hour tour of Lake George. As the operator initiated a turn, the tour boat *Ethan Allen* listed (tipped) enough to take water aboard. The force caused by dipping beneath the surface caused the vessel to list, shifting the passengers to one side of the boat. After this shift in the weight distribution, the boat capsized killing 20 passengers and injuring 9 others.

At the time of the accident, the stability requirements were based on the Coast Guard criteria of a mean of 140 pounds per person. So, the Ethan Allen was supposed to be able to safely transport passengers and crew with a mean weight of 140 pounds. We want to investigate if 140 pounds is a reasonable

value for the mean weight of tour boat passengers. The research question is: "Is the mean weight of tour boat passengers greater than 140 pounds?"

We can rewrite the research question in a declarative sentence to obtain a hypothesis, or a testable statement about a population.

The first hypothesis we will write is that the Coast Guard criteria is appropriate: "The mean weight of tour boat passengers is 140 pounds." We call this the null hypothesis. The **null hypothesis** is a statement of the "status quo," or the value typically considered to be appropriate. Notice that the null hypothesis is expressed with a statement involving equality (=).

H_o : $\mu = 140$ pounds

In contrast to the null hypothesis, we write the alternative hypothesis. This is typically the statement that a researcher suspects is the actual truth. In our case, we suspect that "The mean weight of tour boat passengers is greater than 140 pounds."

H_a : $\mu > 140$ pounds

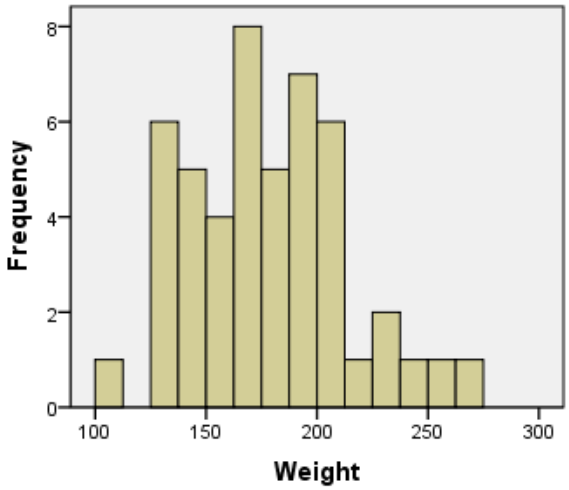
We label the null hypothesis H_o and the alternative hypothesis H_a. In every hypothesis test in this class, the null hypothesis will be a statement involving equality. The alternative hypothesis can include greater than ($>$), less than ($<$), or not equal ($\neq$).

When we test hypotheses, we assume the null hypothesis is true. Because of this requirement, whenever we need to use in a calculation, we can use the value specified in the null hypothesis. When we conduct a hypothesis test, we gather evidence against the requirement that the null hypothesis is true. If we get enough evidence against the null hypothesis, we reject it. If we do not have sufficient evidence against the null hypothesis, we do not reject it.

How do we gather evidence against a null hypothesis? We collect data.

The marine accident report gives the weight (in pounds) of each of the passengers and the crew member. These values are given in the file EthanAllen-Passengers.xlsx.[**?**] According to the CDC, the standard deviation of the weights of individuals in the United States is $\sigma = 26.7$ pounds.[**?**]

To help you visualize the the data, here is a histogram summarizing the weights of the passengers.



Considering the data as a random sample of all possible tour boat passengers, it appears that the true mean weight of tour boat passengers might be greater than 140 pounds. However, we need to check this with a formal test of our hypotheses. It is not sufficient to gain an intuitive sense for the data. We will test if there is sufficient evidence to reject the null hypothesis that the true mean weight of tour boat passengers is 140 pounds.

Problem 5.13 Answer the following questions, using the data EthanAllen-Passengers.xlsx.[?].

1. Find the mean and the sample size.
2. Find the z-score corresponding to the sample mean.
3. Assuming the null hypothesis is true, what is the probability that we would observe a sample mean as extreme or more extreme than the values we observed?[2]
4. Interpret the previous result. What do you conclude? Give your answer in the context of the original problems.

As a result of this accident, the United States government took several actions. The Coast Guard stability regulations were changed, and the assumed average weight per person was increased to 185 pounds.[3] As a result, the safety of public vessels has been improved.

Problem 5.14 An apple processing facility packages apples into 3 lb bags prior to shipping them to supermarkets. The bags have a 3 lb label on them, but clearly not every bag can weigh 3 lbs. The machine that packages the apples can be adjusted to have a mean weight $\mu = 3.1$ lbs, but over time the machine needs to be recalibrated. The standard deviation is historically known to be about $\sigma = .05$ lbs, and doesn't change even if μ is off. If the μ drops below 3.1, then the company runs the risk of shipping out too many low weight bags and getting a fine or lawsuit. If μ rises above 3.1, then the company is loosing money as they are still going to charge the same for heavier apples.

1. The company would like to run a hypothesis test to determine if the mean weight of the apples is $u = 3.1$ lbs or not. What are the null and alternative hypotheses?
2. Every day, a sample of 30 bags is pulled from the belt and weighed. One day, the average weight is computed to be 3.09 lbs. Assuming the null hypothesis is correct, what is the probability of observing a sample mean that is at least as extreme as this one. In other words, what is the probability of observing a sample mean that is either below $\bar{x} = 3.09$ or above 3.11. We call this a two-sided P-value.
3. Compute a %95 confidence interval for μ using the sample mean $\bar{x} = 3.09$. Does this confidence interval contain $\mu = 3.1$?
4. Give the smallest possible $\bar{x}$ so that a %95 confidence interval for μ would still contain $\mu = 3.1$. Any sample mean smaller than this would suggest that the machine need recalibration.

---

[2] This is the same as the probability of observing a value of $\bar{x} = 177.6$ or more pounds, given that the true mean really is $\mu = 140$ pounds.

[3] See http://www.uscg.mil/hq/cg5/cg5212/docs/secg12142010.pdf.

Body Temperatures

Have you ever wondered how it was determined that the true mean body temperature of healthy adults is 98.6°?

It is not exactly clear who first reported this value, but this temperature has been used since the 1800's.[**?**, **?**] One of the most influential researchers in this area is Carl Reinhold August Wunderlich. He reported measuring over 1,000,000 body temperatures on over 20,000 patients.[**?**, **?**] Based on his research, Wunderlich stated, "The axillary[4] temperature of 98.6° F = 37° C. . .is considered the central thermic point of health".[**?**] In other words, the mean body temperature of healthy adults is 98.6° F (or 37° C.)

A group of researchers led by Philip A. Mackowiak, MD, conducted a study to assess the true mean body temperatures of healthy adults.[**?**] They selected $n = 148$ subjects between the ages of 18 and 40 years old, representative of the general population. Each volunteer was given a physical to assure that they were not ill at the time of the data collection. Their axillary body temperature was measured and reported in a paper published in the *Journal of the American Medical Association*.[**?**] These data were extracted and are presented in the file BodyTemp.xlsx. The body temperatures in the file are given in degrees Fahrenheit.

Problem 5.15 Do the following.

1. Create a histogram illustrating the body temperatures of the individuals in the Mackowiak study.
2. Based on the mean of the observations and the histogram of the data, does it appear that the mean body temperature of healthy adults is significantly different from 98.6° F?

Problem 5.16 Continue from the previous problem.

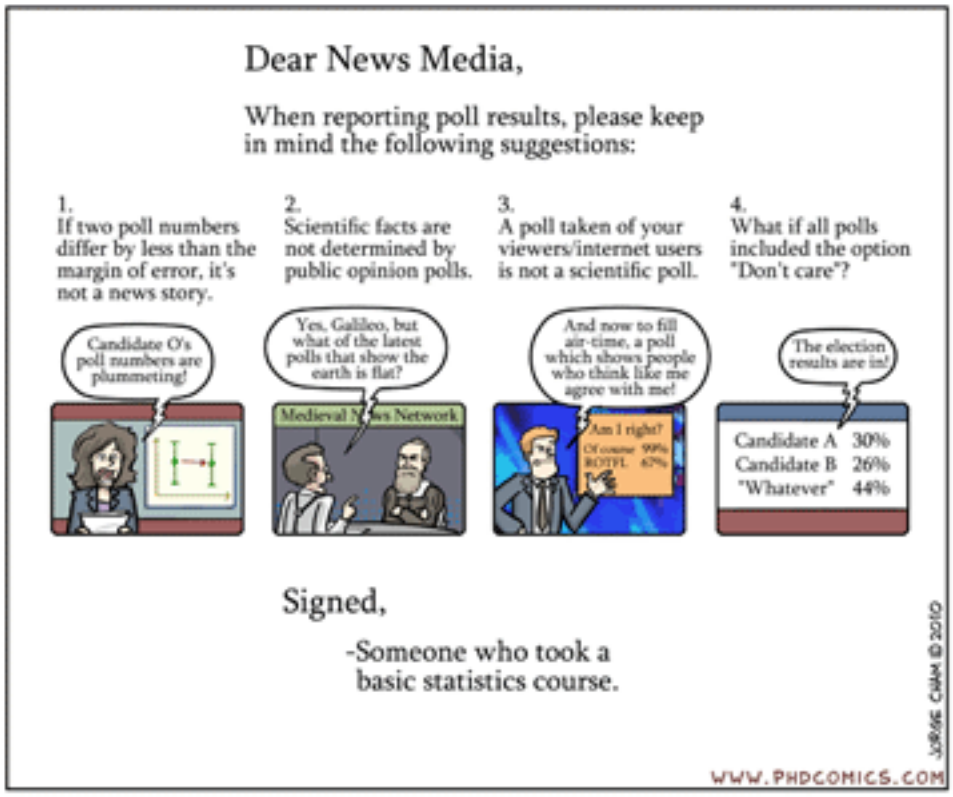
1. We'd like to know if the true mean body temperature of healthy adults is really $\mu = 98.6°$ F. Write a null hypothesis H_0 and alternative hypothesis H_a for this study.
2. Assuming that the true mean body temperature of healthy adults is $\mu = 98.6°$ F, and the population standard deviation is $\sigma = 0.675°$ F, [**?**] find the mean and standard deviation of the random variable $\bar{X}$.
3. Use the information in the previous part to find the z-score for the sample mean. Is this sample mean extreme?

Problem 5.17 Continue from the previous problem.

1. What is the probability of observing a z-score that is as extreme or more extreme (further away from 0) than the z-score you calculated in the previous question?
2. Assuming the mean body temperature really is 98.6° F, how likely would it be for a random sample of $n = 148$ people in the population to have a mean body temperature that is at least as extreme as the one observed here?

[4] When the temperature is measured in the arm pit, it is called an axillary temperature measurement.

"Piled Higher and Deeper" by Jorge Cham



3. Results as unlikely as this demand an explanation. What do you think is the reason for a z-score as extreme as this?

4. What do you conclude about the mean body temperature of healthy adults?

When we make a confidence interval and use that interval to make decision, we run a risk of making an error. For simplicity, let's assume that we are making a 95% confidence interval. In this case, our interval could be one of the rare 5% of all confidence intervals for μ that does not contain the true mean μ. We might use this interval to incorrectly decide the mean is somewhere it is not.

When we run a hypothesis test, we again run the risk of making an error. There are two types of errors that can occur in a hypothesis test.

Type I Error: We reject the null hypothesis when the null hypothesis is true.

Type II Error: We fail to reject the null hypothesis when the null hypothesis is false.

What we would like to do is reduce the probability of making the two errors above. There is a catch though, as we reduce the probability of a Type I error, we increase the probability of a Type II error, and vice versa. We can't reduce the risk of both errors simultaneously.

The power of a test is related to the probability of a type II error. If you are interest in learning more, search online for "the power of a test."

Problem 5.18 Answer the following using complete sentences.

1. We collect some data and make a 95% confidence interval for μ. What is the probability that our interval contains the true mean μ of the population.

2. We collect some data and make a 95% confidence interval for μ. What is the probability that our interval contains the sample mean $\bar{x}$ of our sample.

3. If we were to conduct many studies (thousands) and in each study we constructed a 95% confidence interval for μ, about what percent of the intervals would contain μ.

4. We wish to run a hypothesis test $H_0 : \mu = 3$ versus $H_a : \mu \neq 3$. We'll collect data and create a 95% confidence interval for μ and reject the null hypothesis if our interval does not contain 3. Under this scenario, what is the probability of making a Type I error. [Hint: Read the paragraphs before this problem if you're stuck.]

The previous problem should have shown you there is a connection between the confidence level and the probability of making a Type I error. Let's make a formal definition.

Definition 5.1: Significance Level. The significance level α of a hypothesis test is the probability of making a type I error. In other words, the significance level α is probability of rejecting the null hypothesis when the null hypothesis is actually true.

If we are using 90% confidence intervals to make decisions, then a two-side hypothesis test would have a significance level of $\alpha = 0.10$. The confidence level and significance level are complimentary for two sided tests. However, the same is not true for one-sided hypothesis tests. Let's look at an example with a two-side test, and a one-side test, before we generalize the idea.

Problem 5.19 Suppose we would like to test the hypothesis $H_0 : \mu = 5$ versus $H_a : \mu \neq 5$. We collect a sample of size $n = 100$ from a population with $\sigma = 2$ and find $\bar{x} = 4.82$.

1. Compute a 95% confidence interval for μ. Based of this interval, should we reject, or fail to reject, the null hypothesis?

2. We observed a sample mean that was 0.18 units away from the hypothesized value of μ. If we assume the null hypothesis is correct, compute the probability of observing a sample mean that is further from μ than what we observed. In other words, state the P-value corresponding to this hypothesis test.

3. We want to make decisions in this two-sided test with 95% confidence. What should we use for our significance level α?

4. How does the P-value compare to α? From just the P-value alone, should we reject, or fail to reject, the null hypothesis?

Problem 5.20 Suppose we would like to test the hypothesis $H_0 : \mu \geq 5$ versus $H_a : \mu < 5$. We collect a sample of size $n = 100$ from a population with $\sigma = 2$ and find $\bar{x} = 4.82$. Because we see that $\bar{x}$ is less than 5, we might have evidence to suggest that $\mu < 5$. In our test below, let's use a significance level of $\alpha = 0.05$.

1. We observed a sample mean that was 0.18 units less than the hypothesized value of μ. If we assume the null hypothesis is correct, compute the probability of observing a sample mean that is even less than what we observed. In other words, state the P-value corresponding to this hypothesis test.

2. Because we are using $\alpha = 0.05$, we will only reject the null hypothesis if we observed something that should happen in less than 5% of samples, assuming that the null hypothesis is true. Should we reject or fail to reject the null hypothesis in this example?

3. In a one-sided hypothesis test, we only look at the area in one tail of the normal distribution, whereas confidence intervals use both tails. If we use a significance level of $\alpha = 0.05$ to make our decisions in a one-tailed test, then what's the corresponding confidence level C we should use to construct confidence intervals.

4. Generalize the previous result. If we know this significance level α for a one-sided test, then what's the corresponding confidence level C we should use to construct confidence intervals. Your answer should be in terms of α.

Definition 5.2: P-value in a Hypothesis Test for a Mean with Sigma Known. The test-statistic in a hypothesis test for a mean with sigma known is the z-score $z = \dfrac{\bar{x} - \mu}{\sigma}$. The P-value is the probability, assuming the null hypothesis is true, of observing a sample mean as extreme or more extreme than what was observed.

We have already seen the P-value show up in several problems. Let's review how we computed the P-value.

- If it's a two-side hypothesis test (so something like $H_0 : \mu = 5$ versus $H_a : \mu \neq 5$) then we compute the test statistic z and then find the area to the left of $-|z|$ and to the right of $|z|$, so the area in both tails.
- If it's a left-tailed test (so something like $H_0 : \mu \geq 5$ versus $H_a : \mu < 5$) then we compute the test statistic z and then find the area to the left of z, so the area in the left tail.
- If it's a right-tailed test (so something like $H_0 : \mu \leq 5$ versus $H_a : \mu > 5$) then we compute the test statistic z and then find the area to the right of z, so the area in the right tail.

In all cases, the key is to look at the alternative hypothesis and then find the area in the tail(s) that would suggest evidence in favor of the alternative hypothesis. Once you have found a P value, the next step is to decide to either reject, or fail to reject, the null hypothesis.

Problem 5.21: The Decision Rule Suppose we are performing a hypothesis test and we have decided to use a level of significance of $\alpha = .10$.

1. What does it mean to say $\alpha = .10$.
2. If we compute a P value of $P = .07$, should we reject or fail to reject the null hypothesis? Explain your answer by writing down several sentences.
3. If we compute a P value of $P = .11$, should we reject or fail to reject the null hypothesis? Explain your answer by writing down several sentences.
4. State a general decision rule for deciding to reject or fail to reject the null hypothesis. Suppose that the level of significance of α and the P-value is P. This should be a simple rule that others can remember for helping them decide when to reject or fail to reject the null hypothesis.

Problem 5.22: Mean age of people who read the New York Times online

The New York Times is a large newspaper that is available both in the traditional print format and online. It is known that the mean age of readers of the print edition is 42 years. The research question is: Is the mean age of readers of the online version of the Times less than 42 years?

The population is all people who read the Times online. The observational units are the individual customers. The age of each person who was selected and agreed to participate was measured. The customers age is a quantitative random variable.

Using a simple random sample, data were collected on $n = 25$ people who read the Times online. Their ages were recorded and the mean age of the subjects was $\bar{x} = 35.32$. Assume the standard deviation of the population is $\sigma = 12$ years.

1. State an appropriate null and alternative hypothesis.
2. Compute the test statistic and P-value. If our level of significance is $\alpha = 0.05$, then what decision should we make (reject or fail to reject)?
3. What should you tell the manager of the Times about the average age of their online readers? In other words, interpret the results of your hypothesis test in a way that business leaders can make informed decisions.
4. What assumptions must we check if we want to use the results above.

Problem 5.23: Mean GPA of early risers In Doctrine and Covenants 88:124, the Lord commands, "...cease to sleep longer than is needful; retire to thy bed early, that ye may not be weary; arise early, that your bodies and your minds may be invigorated." Are there academic benefits from obeying this commandment? People who sleep late tend to be stereotyped as slackers. It is not fair to make these claims without any scientific evidence. The purpose of this study is used to address whether early risers earn better grades.

Researcher K. Clay and others presented a paper at the Associated Professional Sleep Societies meeting [4]. They suggested that the mean GPA of students who are early risers tends to be higher than average. The mean grade point average (GPA) of all students at BYU-Idaho is $\mu = 3.15$. The population standard deviation for the grades of students at BYU-Idaho is $\sigma = 0.68$.

A simple random sample of $n = 378$ early-rising BYU-Idaho students was collected. The students were asked to report their BYU-Idaho GPA. The GPA of each person who was selected and signed the informed consent statement (to authorize the use of their data) was recorded. The mean GPA of the subjects was $\bar{x} = 3.21$.

1. State an appropriate null and alternative hypothesis.
2. Compute the test statistic and P-value. If our level of significance is $\alpha = 0.05$, then what decision should we make (reject or fail to reject)?
3. Present your conclusion in an English sentence, relating the result to the context of the problem.
4. What assumptions must we check if we want to use the results above.

"Piled Higher and Deeper" by Jorge Cham

The next two problems have you build an excel sheet that will perform all the statistical computations above by just typing in the relevant summary statistics. The excel commands `=normsdist()` will help you obtain an areas from a z-scores, and the command `=normsinv()` will help you get a z^* from an area.

Problem 5.24: Automate The Computations for Confidence Intervals

Construct an excel sheet that will perform all the confidence interval computations for you by just updating the values of n, $\bar{x}$, σ, and the confidence level C.

In other words, if you know $n = 40$, $\bar{x} = 23$, $\sigma = 11$, and have a confidence level of $C = 95\%$, then set up an excel worksheet that will compute the margin of error m, the value of z^*, and the lower and upper bounds for your confidence interval for μ.

You should get (19.59, 26.41). You'll need to use the command `=normsinv()`.

You should then be able to obtain the confidence interval if $n = 50$, $\bar{x} = 57$, $\sigma = 8$, $C = 93\%$, by just changing these 4 values, and not doing any other computations.

Problem 5.25: Automate The Computations for Hypothesis Testing

Construct an excel sheet that will perform all the hypothesis testing computations in a two tailed test by just updating the values of the null hypothesis $H_0 : \mu = \mu_0$, sample size n, sample mean $\bar{x}$, population standard deviation σ, and the significance level α.

In other words, if you want to perform a hypothesis test on $H_0 : \mu = 20$ versus $H_a : \mu \neq 20$ where we know $n = 40$, $\bar{x} = 23$, $\sigma = 11$, and $\alpha = .05$, then set up an excel worksheet that will compute the z-score and P-value automatically.

You should get $P = .084549$. You'll need to use the command `=normsdist()`.

You should then be able to obtain the P-value if you change $H_0 : \mu = 60$, $n = 50$, $\bar{x} = 57$, $\sigma = 8$, $\alpha = 0.07$, by just changing these 5 values, and not doing any other computations.

5.3 One Population Mean: σ Unknown

As you may have already surmised, we almost never know σ. How can we create a confidence interval, if σ is unknown?

In practice, we almost never know the population standard deviation, σ. So, it is generally not appropriate to use the formula

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

In 1908, William Sealy Gosset published a solution to this problem [**?**]. He found a way to appropriately compute the confidence interval for the mean when σ is not known. The basic idea is to use the sample standard deviation s in the place of the true population standard deviation σ. If σ is not known, we cannot base the calculations on the standard normal distribution, and we cannot use the formula above to conduct hypothesis tests. This is because in general our sample standard deviation will not take into account extreme observations, and hence will more often be smaller than the true population standard deviation.

William Sealy Gosset (1876-1937) was a British industrial scientist and statistician best known for his discovery of the t-distribution.

In a remarkable piece of work, Gosset found the appropriate distribution to use when σ is unknown. At the time of this discovery, Gosset worked for the Guinness brewery. To avoid problems with industrial espionage, Guinness prohibited employees from publishing any research results. Knowing his work provided a significant contribution to statistics, Gosset chose to publish his results anyway. He chose the pseudonym "Student". Gosset's test statistic was denoted by the letter t, and this distribution has come to be known as **Student's t-distribution**. We write

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

The t-distribution is bell-shaped and symmetrical. The t-distribution has a mean of 0, but it has more area in the tails than the standard normal distribution, so there's a greater change of getting larger test statistics, as we divide by potentially smaller standard deviations.

The exact shape of the t-distribution depends on a parameter called the **degrees of freedom** (abbreviated df). The degrees of freedom is related to the sample size. As the sample size goes up, the degrees of freedom increase accordingly. For the procedures discussed in this chapter, the degrees of freedom equal the sample size minus one, so we have $df = n - 1$. Figure **??** shows the graph of several t distributions for $df = 1, 5, 15$ (the black curve) together with the normal distribution (the red curve). As the degrees of freedom increase, the distribution approaches the normal bell curve.

Other than a slightly different shape, and a new test statistics t, we perform computations for confidence intervals and P-values in the exact same way. Instead of computing z or z^*, we now just compute t and $t*$. We then need an appropriate tool to work between values of t and areas under a t-distribution. We'll use Excel. The commands `=normsdist()` and `=normsinv()` allow us to work between z-scores and areas. Typing `=normsdist(2)` will give us the area to the left of $z = 2$, and typing `=normsinv(.5)` will give us the value of z that has half the area to the left. We now explore the functions `=tdist()` and `=normsinv(tinv()`.

Problem 5.26 Complete the following in Excel (these commands will not work in Google docs, but they should in Open Office and several other spread sheet programs).

1. Given a value of t and the degrees of freedom df, the Excel command `=tdist(t,df,2)` returns an area under the two tailed student's t-distribution. Use excel to compute `=tdist(0,6,2)`, `=tdist(.1,6,2)`, `=tdist(1,6,2)`,

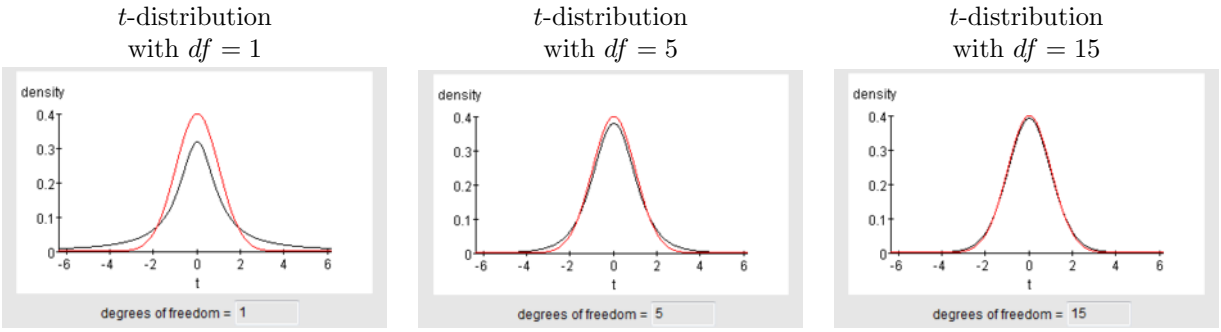


Figure 5.1: Each t-distribution is bell shaped but has more area in the tails than the normal distribution. The black curves above are t-distributions, while the red curve is the standard normal distribution. Note that as the degrees of freedom increase, the black curve approaches the red curve. (Image credit: Webster West, http://www.stat.tamu.edu/~west/applets/tdemo1.html)

and `=tdist(2,6,2)`. For each of the four computations, please draw a bell shaped curve (representing the t-distribution) and shade in the appropriate area that the Excel computation gave.

2. Given an area p and degrees of freedom df, the Excel command `=tinv(p,df)` returns a value of t. Beware, this command does not work the same way as `=normsinv(p)` which always gives a z-score with area p to the left. Play around with several values of p in `=tinv(p,6)`. How does the t relate to the area p?

3. Excel gives the result `=tinv(0.2,6)` $= 1.4398$. Draw a bell curve, shade the appropriate region with area p, and mark the value $t = 1.4398$ on your picture.

Language Translation: BLEU

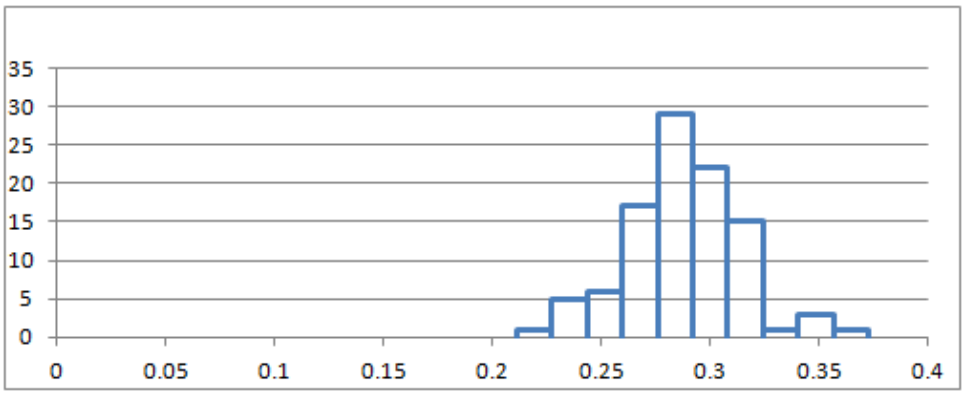
Computer software is commonly used to translate text from one language to another. As part of his Ph.D. thesis, Philipp Koehn developed a phrase-based translation program called Pharaoh. [**?**]

The quality of the translation can vary. A good translation system should match a professional human translation. [**?**] It is important to be able to quantify how good the translations produced by Pharaoh are.

The IBM T. J. Watson Research Center developed methods to measure the quality of a translation from one language to another. [**?**] One of these is the the BiLingual Evaluation Understudy (BLEU). [**?**] BLEU is a score ranging from 0 to 1 that indicates how well a computer translation matches a professional human translation of the same text. Higher scores indicate a better match. BLEU helps companies who develop translation software "to monitor the effect of daily changes to their systems in order to weed out bad ideas from good ideas." [**?**]

To test Pharaoh's ability to translate, Koehn took a random sample of 100 blocks of Spanish text, each of which contained 300 sentences, and used Pharaoh to translate each of these to English. The BLEU score was calculated for each of the 100 blocks. The data were extracted from Figure 2 in a paper Koehn published. [**?**] The 100 BLEU scores are given in BLEU-Scores.xlsx.

Koehn wants to find an estimate of the true mean BLEU score for text translated by the Pharaoh computer program. He would like to compute a confidence interval, but he does not know the true population standard deviation σ. The summary statistics are $\bar{x} = 0.2876$, $s = 0.0264$, and $n = 100$. The following histogram illustrates the data using 10 bins.



The requirements for creating a confidence interval for a mean with σ unknown are the same as the requirements for this procedure when σ is known, namely

1. A simple random sample was drawn from the population, and
2. The sample mean $\bar{x}$ is normally distributed.

It is reasonable to treat the data as representative of the population. The passages selected are a simple random sample of all possible texts.

Recall the requirement of normality is satisfied if the data are approximately normally distributed or if the sample size is large. The data are bell-shaped and fairly symmetric. So, the sample mean, $\bar{x}$, is approximately normally distributed.

The formula for the confidence interval where σ is known is

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}},\ \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right).$$

It is impossible to know the true standard deviation of the BLEU scores for a new translation program like Pharaoh. Replacing σ with s and replacing z^* with t^*, we get the confidence interval formula

$$\left(\bar{x} - t^* \frac{s}{\sqrt{n}},\ \bar{x} + t^* \frac{s}{\sqrt{n}}\right).$$

The value of t^* depends on the level of confidence and the sample size. It must be computed using Excel or looked up on a table. The other numbers ($\bar{x}$, s, and n) can all be obtained directly from your data.

Problem 5.27 Answer the following questions using the data set BLEU-Scores.xlsx, which gives the BLEU scores for $n = 100$ translations from Spanish to English by the computer program Pharaoh.

1. Use the Excel function `=TINV(p,df)` to find the t^* for a 95% confidence interval for a sample of size $n = 100$. (Remember the degrees of freedom is $df = n - 1$). You should obtain $t^* = 1.9842$.
2. The sample statistics for these data are: $\bar{x} = 0.2876$, $s = 0.0264$, $n = 100$. Give the 95% confidence interval for μ.

3. Interpret this confidence interval in an English sentence. What do these numbers mean? Be sure to relate this to the context of the problem.

Problem 5.28 Continue from the previous problem.

1. Now give a 90% confidence interval for the true mean BLEU score for translations by the Pharaoh program. Give your answer accurate to 4 decimal places and interpret this confidence interval in a complete sentence.
2. Repeat this computation to give a 99% confidence interval for the true mean BlEU score for translations by the Pharaoh program. You can make this really fast if you create an excel sheet to do your computations.
3. What do you notice about the confidence interval as the confidence level increased from 90% to 95% to 99%?

Euro Coin Weights

A group of statisticians measured the weights of 2000 Belgian one Euro coins in eight batches. Each batch contains coins that were all minted together. [**?**] You can learn more about these data at:
http://www.amstat.org/publications/jse/datasets/euroweight.txt

The coins were borrowed from a bank in Belgium, one batch at a time. The weights (in grams) of the coins are given in the file EuroWeight.xlsx.

Problem 5.29 Answer the following questions using the data in the file EuroWeight.xlsx.

1. Give the relevant summary statistics.
2. Make an appropriate graph to illustrate the data.
3. Verify the requirements (or assumptions) have been met.
4. Create a 99% confidence interval for the true mean.
5. Present your observations in an English sentence, relating the result to the context of the problem.

Hypothesis testing is pretty much the same as it was when we knew σ. The only difference now is that we compute a t-score from $\bar{x}$, the hypothesized value of μ, and s. Let's revisit the body temperature problem, but this time use a t test instead of a z test.

Problem 5.30 We want to conduct a hypothesis test to determine if the mean body temperature is different from 98.6° Fahrenheit. Previously, we assumed that we knew the value of σ. Actually, this value is not known. From a simple random sample of $n = 148$ subjects, Dr. Mackowiak and his colleagues obtained body temperatures whose summary statistics were $\bar{x} = 98.234$ and $s = 0.738$.

State an appropriate null and alternative hypothesis. Compute the test statistic t, state the degrees of freedom df, and give the corresponding p-value, and decide to reject or fail to reject the null hypothesis. Finally, interpret your results in the context of the original question.

A companies can use a t test to assess if their product contains a desired amount of some compound, to prove that their product performs better than some industry standard, and much more. Let's look at two more examples.

Problem 5.31 A manufacturing process is supposed to create a composite alloy that contains 25% aluminum by weight. Thirty objects are created and then the aluminum content is measured to have an average of 25.2% aluminum by weight, with a standard deviation of $s = 0.53\%$. Create an appropriate null and alternative hypothesis, state the test statistics, degrees of freedom, and p-value, and then give a conclusion of the test that would help the plant manager know how their manufacturing process is doing.

The next problem is silly, but you can replace widget with any product that you want to test if it is faster than some known standard (like a microprocessor, max speed in a car, better gas mileage, etc.)

Problem 5.32 You have designed a new widget to perform faster than the other widgets on the market. The fastest widget on the market currently takes 0.336 min to travel 1 mile. You grab 40 of your new widgets and have each travel 1 mile, recording the time each takes. The average time is $\bar{x} = 0.320$ min, with a standard deviation of $s = 0.068$ min. Conduct an appropriate hypothesis test to determine if the new widget performs faster than the current industry standard.

Chapter 6

Inference for Two Population Means

If you want to prove that your product performs better than another, show that two manufacturing processes are producing different products (when they should be the same), compare 4 different processing chips, etc., then we need tools that allow us anaylze more than just one mean. This chapter will give us the tools to do this.

We'll explore how to compute confidence intervals and conduct hypothesis tests when comparing several means. The concepts from the previous chapter will apply as well. In the previous chapter we always started by computing an observed difference ($\bar{x} - \mu$), divided it by an appropriate reference ($\sigma/\sqrt{(}n)$ or $s/\sqrt{(}n)$) to determine if the difference was large (obtaining the test statistic z or t), and then computed a P-value by looking for the area in a tail or both tails of the corresponding distribution. This process is exactly the same for the tests we'll examine here, except that now we will be considering more than one mean.

6.1 Paired Data: Dependent Samples

In education, it is very common for researchers to conduct studies in which they administer a pre-test, provide some instruction, and then give a post-test. The difference between the post- and pre-test scores is a measure of the student's progress. In this case, it would not make much sense to only look at the mean score on the pre-test and compare it to the mean score on the post-test.

This is called a **matched-pairs** design or we say we have **dependent samples**. Matched-pairs (or **paired-data**) designs typically involve only one population, and a pair of observations is drawn on the individuals selected for the sample. In the context of the educational study, the two observations are student's scores on (1) the pre-test and (2) the post-test. If a student is selected to participate in the pre-test (i.e., they are selected to be part of group 1), they are automatically selected to participate in the post-test (i.e., they are chosen to be in group 2 automatically.)

There is a lot of merit in subtracting the individual scores and looking at the mean *gain*. The researchers are not really interested in the students knowledge before the instruction. This is used as a baseline to measure how much was gained during the instruction. There is great value in looking at the difference. This removes the effect of the individual students' ability, and it measures their learning during the unit.

To analyze the data, the researchers first find the difference in the post- and

pre-test scores. At that point, the data have been reduced to a list of numbers (representing the increase in scores.) Now, the researchers can conduct inference on the mean of these values. In other words, they can do a hypothesis test for the mean of the difference in the post- and pre-test scores.

A hypothesis test for two means with paired data (dependent samples) is conducted in the same way as a hypothesis test for a single mean with σ unknown. The only exception is that the pairs of data must be subtracted before you start any computations. From a practical perspective, after you subtract, then you apply the one-sample procedures you learned in Chapter **??**.

The assumptions for creating a confidence interval for the paired difference of means are the same as the for the hypothesis test. We assume the following:

- A simple random sample was drawn from the population.
- The mean of the differences is normally distributed.

We will first explore an application of pre- and post-testing in forestry.

Mountain Pine Beetle Attacks

Mountain pine beetles are small insects that bore into the bark of trees. The female beetles that first infest the tree emit pheromones to attract other beetles. In response to the pheromones, many beetles bore into the tree and ultimately kill it. The insects can destroy large tree stands within one year.

Lodgepole pine (*Pinus contorta* Dougl.ex Loud.) are particularly susceptible to mountain pine beetle (*Dendroctonus ponderosae* Hopkins) outbreaks. The image above shows the destruction that can be caused by these insects. The large brown patches are pines that have been killed by the beetles.

The mountain pine beetle threatens many forests in the United States. These tiny insects are only 0.5 cm long–about the size of a grain of rice. This photo of a mountain pine beetle is magnified greatly. These little creatures can destroy a large, healthy forest. Can you give a spiritual parallel? If you would like a good read, look up the analogy President Hinckley gave about an axe head in the notch of a tree.

Ron Long, Simon Fraser University, Bugwood.org

In a study conducted in the Arapaho National Forest in Colorado, researchers from the USDA Forest Service studied the effect of pine beetle outbreaks on the average number of trees in an area. [**?**] The researchers counted the number of established trees per hectare before a pine beetle outbreak and seven years after an outbreak. (One hectare is an area of 100 meters by 100 meters.) Data representative of their observations are given in the file PineBeetle.xlsx.

Problem 6.1 Do the following

1. Find the mean and standard deviation of the number of trees per hectare "before" the pine beetle outbreak. How would you describe the density of the trees in this forest? Express this in terms that make sense to you.
2. Repeat question 1 for the number of trees per hectare *after* the outbreak.
3. Create a new column of data in the file PineBeetle.xlsx by subtracting the "before" counts from the "after" counts:

   $$Difference = After - Before$$

   For these differences, report the mean, the standard deviation, and the sample size.
4. Create a histogram of the differences in the density of the trees, and verify the requirements have been met.
5. Find the 95% confidence interval for the difference. Present your observations in an English sentence, relating the result to the context of the problem.

Sleep Inducing Drugs

In William Sealy Gosset's landmark paper on the t-distribution, he cites data on a sleep-inducing drug. In a paper published in 1905, Arthur R. Cushny and A. Roy Peebles reported the effect of Lvorotary Hyoscyamine Hydrobromate (L-Hyoscyamine) on the length of time that people sleep before waking. [**?**] The primary research question is: does L-Hyoscyamine impact the mean amount of time that people sleep before waking? We will compute a 90% confidence for the true mean difference in the times.

Eleven subjects were included in the study. At the start of the study, the researchers observed the average length of time that each of the participants slept before waking. Later, each subject was given 0.6 mg of L-Hyoscyamine and the duration of uninterrupted sleep was again measured.

The difference in the amount of time each person slept was computed by subtracting the amount of time the subjects slept when taking the drug minus the sleep duration with no drug. The data are summarized in the table below.

Subject	Control (no drug)	L-Hyoscyamine	Difference
1	0.6	1.3	0.7
2	3.0	1.4	−1.6
3	4.7	4.5	−0.2
4	5.5	4.3	−1.2
5	6.2	6.1	−0.1
6	3.2	6.6	3.4
7	2.5	6.2	3.7
8	2.8	3.6	0.8
9	1.1	1.1	0.0
10	2.9	4.9	2.0
11	–	6.3	–

Notice that the "control" data for Subject #11 is missing. It is not possible to compute a difference for this person, so their data will be omitted from our analysis. For this analysis, we will use the remaining $n = 10$ observations.

You may find it easier to copy and paste the data from the following table. The last row has been omitted.

Difference
0.7
-1.6
-0.2
-1.2
-0.1
3.4
3.7
0.8
0.0
2.0

Problem 6.2 Do the following

1. For the differences, report the mean, the standard deviation, and the sample size.
2. Create a histogram of the differences in the hours of sleep and verify the requirements have been met.
3. Find the confidence interval. Use the 90% level of confidence. Present your observations in an English sentence, relating the result to the context of the problem.

6.1.1 Hypothesis Tests

Mahon's Weight Loss Study

Annie Mahon and other researchers in Wayne Campbell's nutrition lab studied the weight loss of $n = 27$ middle aged women who consumed a prescribed low-calorie diet. [**?**] The women's weights were recorded (in kilograms) at the beginning of the study and after the nine-week diet period. The data are given in the file Mahon.xlsx. An excerpt of the data is given below.

Subject	Pre	Post
1	62.5	56.1
2	88.8	80.2
3	74.7	70.8
⋮	⋮	⋮
26	76.3	73.8
27	82.1	77.9

Notice the structure of the data. The weight of each subject was measured before the study and at the conclusion of the study. Each person provided a pre-study weight and a post-study weight. Stated differently, the pre-study weights and the post-study weights are paired. For each row of data, both of these numbers came from the same person. When we collect two observations of the same measurement on each subject, we call it **paired data**. Sometimes paired data are called **dependent samples**.

Problem 6.3 Answer the following questions.

1. The researchers measured the initial weights of the women prior to the study, even though they were not particularly interested in this value. What was the purpose of measuring the pre-study weights?

2. Annie Mahon and her research team are interested in the difference of the weights after the study compared with before:

   $$\text{Difference} = \text{Post} - \text{Pre}$$

   The researchers are not interested in the weights of the women, they are more interested in the "change" in the women's weights. This will give them a measure of the effectiveness of the low-calorie diet. Notice that in this weight loss study, the change in the weights is negative. This indicates that the final weight was lower than the initial weight.
   Compute the difference in the women's weights by subtracting the post-study weights from the pre-study weights using software. Call this new column "Difference".
   What is the mean of the values in the "Difference" column? Interpret this value.

After you have subtracted the pre-study weights from the post-study weights, you are left with a column of differences. We will denote the pre-study weights by x_1 and the post-study weights by x_2. Then, the differences can be denoted as $d = x_2 - x_1$. The difference, d, is defined as the change in the volunteer's weight during the study.

After computing the differences, we do not use the data for the individual groups at all. The researchers are not interested in the values of the women's weights at the beginning of the study or at the end of the study. They are mostly interested in the difference in the weights after the participants complete the study.

After we subtract, we can conduct a hypothesis test to determine if the mean of the differences is less than zero. We use the symbol μ_d to represent the true mean difference in the weights of the women who follow the diet prescribed in this study. The null hypotheses is that the true mean difference is zero ($\mu_d = 0$). The alternative hypothesis is that there is a decrease in the weights, in other words, that the true mean difference is less than zero ($\mu_d < 0$).

Notice that this is essentially a one-sample t-test where the data are the differences in the women's weights. We have one column of data, the differences. We are testing whether the true mean difference is less than zero. After subtracting, a test for a difference of two means with paired data is just like a test for one mean with σ unknown.

In the hypothesis test, we will refer to the variable representing the differences as d. We will use this notation throughout the hypothesis test. For example, the true population mean will be labeled μ_d and the sample mean will be labeled $\bar{d}$. The sample standard deviation of the differences is denoted s_d.

Twenty-seven women participated in a nine week weight loss study. During the study period, the participants were provided a reduced calorie diet. Their weights were recorded at the beginning of the study and nine weeks later. The difference of the weights is defined as the post-study weights minus the pre-study weights. The researchers expected that the mean difference in the weights would be negative–in other words, that the women would tend to lose weight.

The women's weights were recorded at the beginning of the study. The women were provided a reduced calorie diet for nine weeks. Then, their weights were again at the end of the study. A calibrated scale was used to provide an accurate weight.

Problem 6.4

1. State the null and alternative hypotheses and the level of significance. We will use the $\alpha = 0.05$ level of significance.
2. Give the relevant summary statistics.
   Report the number of subjects (n), the mean difference ($\bar{d}$), and the standard deviation of the differences (s_d).
3. Make an appropriate graph (histogram) to illustrate the data, and verify the requirements have been met.
4. Give the test statistic and its value. State the degrees of freedom. Also give the P-value.
5. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

Nosocomial Infections

Matched-pairs designs are not just used in pre- and post-test situations. They are often used in situations where it is not possible to randomly assign subjects to groups (for example, by a coin toss.) Nosocomial (pronounced: NO-suh-KOH-MEE-uhl) infections are infections that occur in hospitals, but are not a result of the original condition. An example of a nosocomial infection is when a heart attack patient develops a staph infection at the site of an IV injection. The infection was not caused by the heart attack, but it was acquired in the hospital. Nosocomial infections are very dangerous and may result in longer recovery times or increased death rates.

This AP photo of a chest x-ray shows pneumonia of the left lower lobe of the lung.
Pneumonia is an example of a possible nosocomial infection.

(Photo credit: Dr.Thomas Hooten, CDC)

Health care providers suspect that nosocomial infections increase the amount of time required to recover from an illness or injury. In controlled experiments, subjects (e.g., patients) are randomly assigned to treatments. However, it is not ethical to give patients a nosocomial infection in order to determine if it increases the duration of their hospital stay! At best, we can collect information on the duration of hospital stays for patients who acquire nosocomial infections and compare them to the duration of the stays for patients who do not.

There are many factors that affect the amount of time that a patient will need to stay in the hospital, including: nature of illness, types of procedures conducted, overall health, gender, age, etc. How can health care practitioners assess the effect of a nosocomial infection in the presence of so many other variables?

One way is to match a patient who develops a nosocomial infection with another one who has similar characteristics (illness, procedures, health, gender, age group, etc.) but does not develop a nosocomial infection. Now, the patients are matched into pairs with similar characteristics, where the principle difference between the members of each pair is whether or not they acquired a nosocomial infection.

By pairing the patients according to specific characteristics, the researchers can now subtract to observe a difference in their recovery times. In this way, it is possible to assess if nosocomial infections increase the mean duration of a hospital stay. Some researchers conducted such a study in which 52 pairs of patients were matched based on clinical characteristics. A patient with a nosocomial infection was matched as closely as possible to a similar case where there was no nosocomial infection. Patients who died were excluded from the study. [**?**] The lengths of the hospital stays (in days) for these patients are given in the file NosocomialInfections.xlsx.

The difference, d, is defined as the duration of the hospital stay of the individual in the pair with the nosocomial infection minus the duration of the stay for the individual who did not get a nosocomial infection:

$$Difference = Infected - NotInfected$$

After computing the differences, we do not use the data for the individual groups any more. In fact, after we subtract, the hypothesis test is conducted

(essentially) like a one-sample test for a single mean with σ unknown.

Data were collected by matching hospital records of individuals who were admitted to the hospital. Patient records were matched based on their overall health and the reason they were admitted to the hospital. In each pair, one patient developed a nosocomial infection and one did not. Since the characteristics of the patients in the first group determined which patients would be paired with them in the second group, the data represent dependent samples.

Problem 6.5 Answer the following.

1. State the null and alternative hypotheses and the level of significance.
2. Give the relevant summary statistics.
3. Make an appropriate graph to illustrate the data, and verify the requirements have been met.
4. Give the test statistic and its value, state the degrees of freedom, and give the P-value.
5. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

Effect of Stressful Classical Music on Your Metabolism

Obesity is a growing problem worldwide. Many scientists are seeking creative solutions to trim down this epidemic. Reduced energy expenditure is a potential cause of obesity.

Resting Energy Expenditure (REE) is defined as the amount of energy a person would use if resting for 24 hours. In essence, this is the amount of energy that a person's body will consume if they do not do any physical activity. REE is measured in terms of kilo-Joules per day (kJ/d).

REE accounts for approximately 70 to 80% of all energy that a person will expend in a day. [**?**] If researchers can find simple, enjoyable activities that will increase REE, it may be possible to minimize the spread of obesity around the world.

Ebba Carlsson and other researchers in Sweden investigated whether listening to stressful classical music increases a person's REE. [**?**] Each subject's REE was measuring during silence and again while listening to stressful classical music. Data representing their results are given in the file REE-ClassicalMusic.xlsx.

Notice that this is not a pre- and post-test, but it is still a test involving paired data. Two REE measurements were made for each subject: (1) in silence (REE_1) and (2) while listening to stressful classical music (REE_2).

The REE was measured by a technique called "indirect calorimetry" using a Deltatrac II Metabolic Monitor. [**?**] The REE was measured twice for each person: while the person was (1) resting in silence or (2) resting while listening to stressful classical music. These trials were conducted in random order. Some of the subjects had the "silence" treatment first, and others had the "stressful" treatment first.

Let $\alpha = 0.1$.

Problem 6.6 1. State the null and alternative hypotheses and the level of significance.

2. We will define the difference in REE by subtracting the REE in silence from the REE while listening to stressful classical music. If listening to stressful classical music actually increases the mean REE, would you expect the value of the difference to be typically positive or negative?

3. Compute the difference in REE for each person. (You should get 50 kJ/d for the first person). Then give the relevant summary statistics.

4. Make an appropriate graph to illustrate the data. Verify the requirements have been met. Give the test statistic and its value. State the degrees of freedom. Finally find the P-value and compare it to the level of significance.

5. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

In the problem above, we do not say we "accept" the null hypothesis. We do not know that listening to stressful classical music has no effect on a person's REE. Based on the data available to us, we were not able to reject the requirement that this type of music does not increase the mean REE.

Cost of Airline Tickets

Pressures of supply and demand act directly on the prices for an airline ticket. As the seats available on the plane begin to fill, airlines raise the price. If seats on a flight do not sell well, an airline may discount the tickets or even cancel the flight. Business travelers frequently demand travel booked on short notice. They must pay the current price. Typically, tourists book their flights well in advance, hoping to buy tickets before the price rises. We will consider the cost of a one-way ticket from London's Heathrow Airport to a variety of destinations in Europe.

Allie Henrich, a BYU-Idaho student, compared the lowest published ticket prices of one-way flights from Heathrow to various destinations in Europe. Using Travelocity.com, she recorded the lowest published fares for nonstop midweek flights booked either 14 days in advance or 90 days in advance. The prices (in US dollars) are given in the file DirectFlightCosts.xlsx. Notice that for some destinations, flights were not available.

The data are paired, because measuring the costs twice for each city. The 14-day ticket price is paired with the 90-day price for each city.

We will conduct a hypothesis test to determine if there is a difference in the cost of the nonstop flights when tickets are purchased 14 days in advance compared to 90 days in advance. We will use the 0.01 level of significance. The data were collected using the website Travelocity.com. The lowest advertized ticket prices were recorded for nonstop flights from Heathrow Airport. All prices were recorded in US dollars. Data are provided on the cost of a nonstop ticket purchased with 14 days notice compared to 90 days notice. ::We will compute the difference in the costs for each destination. Some destinations did not include both flight options. In this case, the difference is not computed and the data are omitted from the analysis.

Problem 6.7 Conduct an appropriate hypothesis test to address the problem above. Your work should follow the same format as all the other problems you've worked on prior to this one.

6.2 Two Independent Samples

Study Tip: Samples are dependent (or represent paired data) if knowing which subjects will be in the first group determines which will be in the second group. If knowing which subjects are in the first group gives you no information about the second group, we say the samples are independent.

In the previous section, we studied confidence intervals and hypothesis tests for the difference of two means, where the data are paired. One example of paired data is pre- and post-test scores, such as Mahon's weight loss study. [**?**] Another example is paired comparisons, like the nosocomial infection study. [**?**] How can you tell if data are paired? The key characteristic of dependent samples (or matched pairs) is that knowing which subjects will be in Group 1 determines which subjects will be in Group 2. The data for each subject in Group 1 is "paired" with the data for a corresponding subject in Group 2. In the case of the weight loss study, the same subject provided weight data for both groups: once in the pre-test (group 1) and once in the post-test (group 2).

In contrast to dependent samples, two samples are independent if knowing which subjects are in Group 1 tells you nothing about which subjects will be in Group 2. With **independent samples**, there is no pairing between the groups. Suppose you want to compare the incomes of men and women in the general population. A random sample of men would be collected, and each would be asked to report their income. Similarly, a random sample of women would be drawn, and they would also be asked to report their income. Notice that the groups are independent. Knowing the names of the men who are selected tells you nothing about which women would be selected. This is an example of independent samples.

We can compare the mean income of men to the mean income of women using the procedures of this section. We will conduct hypothesis tests and compute confidence intervals for the difference in the true population means of two groups ($\mu_1 - \mu_2$).

Some students make the association that samples are independent if they do not affect each other. This is a false notion. Instead, remember that *samples are independent if knowing who was selected for Group A tells you nothing about who will be selected for group B.*

6.2.1 The Standard Error

When we perform hypothesis tests, we compute a test statistic such as z or t by comparing an observed difference to the standard deviation in a sampling distribution. The standard error is often used as a word to describe the standard deviation of the sampling distribution. We could write our observations as follows:

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\ t &= \frac{\bar{x} - \mu}{s/\sqrt{n}} \\ \text{test statistic} &= \frac{\text{difference}}{\text{standard error}}. \end{aligned}$$

The standard error shows up in our confidence intervals as well, which gives us the following formulas:

$$\begin{aligned} \bar{x} &\pm z \cdot (\sigma/\sqrt{n}), \\ \bar{x} &\pm t \cdot (s/\sqrt{n}), \\ \text{(point estimate)} &\pm \text{(critical value)} \cdot \text{(standard error)}. \end{aligned}$$

We would now like to extend the ideas above to allow us to compare two means. As example, perhaps we would like to know if two different manufacturing processes produce similar products. We would need to measure some quantity in

each process from which we could compute a mean. If we called these quantities X and Y, then we would like to know if $\mu_X = \mu_y$. As these are population means, there isn't any way we can ever truly answer the questions, but we can use statistics to help us determine if they are equal. Our null hypothesis would be $H_0 : \mu_X = \mu_Y$ or alternately $H_0 : \mu_X - \mu_Y = 0$. This means our point estimate would be $\bar{x} - \bar{y}$. What we need now is a formula for the standard error.

Problem 6.8 In chapter 5, we showed that the variance of the random variable $X - Y$ is $\sigma_X^2 + \sigma_Y^2$, and that the variance of a mean $\bar{X}$ is $\sigma_X/\sqrt{n}$. Recall also that the formula for a sample variance is $s^2 = \sum(x - \bar{x})^2/(n-1)$.

1. Explain why the standard deviation of $\bar{X} - \bar{Y}$ is $\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$. This is the standard error of the difference of two means.
2. Explain why

$$\sum_{i=1}^{n_X}(x_i - \mu_X)^2 + \sum_{i=1}^{n_Y}(y_i - \mu_Y)^2 = (n_X - 1)s_X^2 + (n_Y - 1)s_Y^2.$$

The quantity above is called the error sum of squares, written SSE, and is just the sum of the squares of the differences between each observation and its respective mean.

You just developed the ideas behind both of the formulas use for the standard error when comparing two means. The most commonly used formula, and the one you should basically always use except in a few select situations, is the computation for the standard deviation of $\bar{X} - \bar{Y}$, in the first part of the problem above. With this formula for the standard error, we can compute test statistics and confidence intervals using the two formulas

$$z = \frac{(\bar{x} - \bar{y}) - (0)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}, \quad \text{and} \quad (\bar{x} - \bar{y}) \pm z^* \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}.$$

If we don't know the populations standard deviations (which is pretty much always), then again we replace z with t and each population variance with the sample variance to obtain the formulas

$$t = \frac{(\bar{x} - \bar{y}) - (0)}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}, \quad \text{and} \quad (\bar{x} - \bar{y}) \pm t^* \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

where the degrees of freedom for the t-distribution is now given using the Satterthwaite approximation given by

$$df = \nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

The derivation of this formula for the degrees of freedom is beyond the scope of our course.

Note that excel rounds the degrees of freedom to the nearest integer, so the computations in excel could be slightly different than computations obtained with statistical software. As long as the sample sizes are not extremely small, this should not produce a significant difference.

Problem 6.9 There is another formula for the standard error that is occasionally used when we know that the two population standard deviations are equal. Please head to the FE (Fundamentals of Engineering) exam website at

- http://ncees.org/exams/study-materials/download-fe-supplied-reference-handbook/.

Supply them your email and they will send you a passcode that allows you to download the FE exam reference book. This book is the only reference material you may take with you into the FE exam. Look on page 39 for the relevant information, and then answer the following questions.

1. If we assume the unknown population variances are equal, what is the appropriate formula for the standard error (you'll need to replace S_p with what it equals), and what are the appropriate degrees of freedom.
2. Suppose we have collected some data and have obtained $\bar{x} = 10, n_x = 30, s_x = 4$ and also $\bar{y} = 11, n_y = 40, s_y = 5$. If we assume the population variances are not equal, compute the standard error and degrees of freedom.
3. Using the same numbers as the previous part, now assume that the population variances are equal and compute the standard error and degrees of freedom.

Now that we have a formula for the standard error, we can create confidence intervals and perform hypothesis tests. As usual, the assumptions required to perform these tests are that (1) we have obtained a simple random sample from each population, and (2) that each population is normally distributed or that the sample sizes are large.

Problem 6.10 Suppose we have collected some data and have obtained $\bar{x} = 10, n_x = 30, s_x = 5$ and also $\bar{y} = 11, n_y = 40, s_y = 4$, the same numbers from the previous problem.

1. Assume that the population variances are not equal, and compute a 90% confidence interval for the difference $\mu_X - \mu_Y$. Make sure you show what you used for the point estimate, the standard error, the degrees of freedom, and the critical value t^*. Any time you create a confidence interval, you should always show this information.
2. Assume that the population variances are equal, and compute a 90% confidence interval for the difference $\mu_X - \mu_Y$. Again, make sure you show what you used for the point estimate, the standard error, the degrees of freedom, and the critical value t^*.
3. Which interval does a better job of estimating the mean?
4. Which assumption is generally better to assume, that the variances are equal or not equal?

Problem 6.11 Again suppose we have collected some data and have obtained $\bar{x} = 10, n_x = 30, s_x = 5$ and also $\bar{y} = 11, n_y = 40, s_y = 4$. We collected this data because we wanted to test the hypothesis $H_0 : \mu_X - \mu_y = 0$ versus $H_a : \mu_X - \mu_y \neq 0$.

1. Assume that the population variances are not equal. Compute the test statistic, degrees of freedom, and then give a P-value for this test as well as what decision should be made.

2. Assume that the population variances are equal. Compute the test statistic, degrees of freedom, and then give a P-value for this test as well as what decision should be made.

In this course, we do not go very deep into statistical theory. For those students who are interested, there is a lot of theory undergirding statistical practice. If the variances of the two groups are equal, then traditional statistical theory suggests that you combine or "pool" the information about the variance in the two groups. If the variances are not equal, you do not combine the information about the spread. These two techniques usually lead to slightly different values for the t-statistic, degrees of freedom, and P-value.

If the variances observed in the sample data are very different from each other, you assume unequal variances and do not pool the data. However, if the variances are very similar to each other, the results of the two procedures will be nearly identical. In this case, it does not really matter which you choose. So, if the sample variances differ significantly, we should not assume equal variances. If the variances do not differ significantly, it doesn't really matter if you assume equal variances or not, as the results will be very close.

It's time to now use our new found ideas about comparing two means to answer several questions. Let's practice.

6.2.2 Hypothesis Tests and Confidence Intervals

Reading Practices of Children with Developmental or Behavioral Problems

Is there a difference in the amount of reading done by children with problematic behavior compared to other children?

Researchers led by Arlene Butz published a study on the reading practices of children. [**?**] They wanted to know if there was a difference in the reading practices of children with developmental or behavioral problems (the DEV group or Group 1) compared to children in the general population who do not have developmental problems (the GEN group or Group 2.) One of the factors they considered was the number of nights each week that the children participated in reading in the home. Data representative of their results are given in the file ReadingPractices.xlsx.

A group of children were enrolled in the study. Children who were identified to have developmental or behavioral problems were labeled as Group 1 (the DEV group). Children who did not display developmental or behavioral problems were labeled as Group 2 (the GEN group). A survey was administered to a parent of each of the children. One of the questions on the survey asked the number of nights that either their child read or that they read to their child during the week. This data is found in the file ReadingPractices.xlsx.

The null hypothesis is that there is no difference in the mean number of nights each week in which the two groups of children participate in reading in the home. The alternative hypothesis is that there is a difference in the mean number of nights that the children in the two groups participate in reading in the home. These hypotheses are expressed mathematically as:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 \\ H_a &: \ \mu_1 \neq \mu_2 \end{aligned} \qquad \text{or} \qquad \begin{aligned} H_0 &: \ \mu_1 - \mu_2 = 0 \\ H_a &: \ \mu_1 - \mu_2 \neq 0. \end{aligned}$$

We will use the $\alpha = 0.05$ level of significance.

Problem 6.12 Answer the following questions:

1. Do the data published by Arlene Butz and her colleagues represent paired data or independent samples? How can you tell?

2. We will use $\bar{x}_1$ to denote the mean of Group 1. Similarly, we use s_1 and n_1 for the standard deviation and sample size of Group 1. For Group 2, we indicate the mean, standard deviation and sample size with the symbols: $\bar{x}_2$, s_2, and n_2, respectively. Find the mean, standard deviation and sample size for the two groups, separately. In other words, find $\bar{x}_1$, s_1, n_1, $\bar{x}_2$, s_2, and n_2.

3. Verify the requirements have been met to perform a hypothesis test, and then give the standard error, test statistic, degrees of freedom, and P-value. Even though the sample standard deviations are quite close, for your computations DO NOT assume that they are the equal. Do all your computations by assuming unequal variances.

4. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

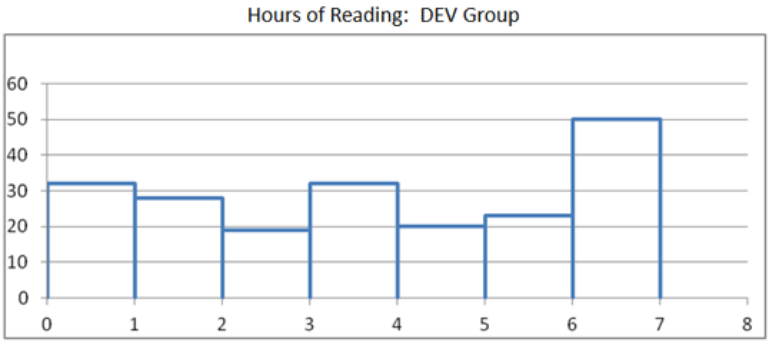
We could have addressed the hypothesis test above by instead computing a confidence interval. Let's do so now.

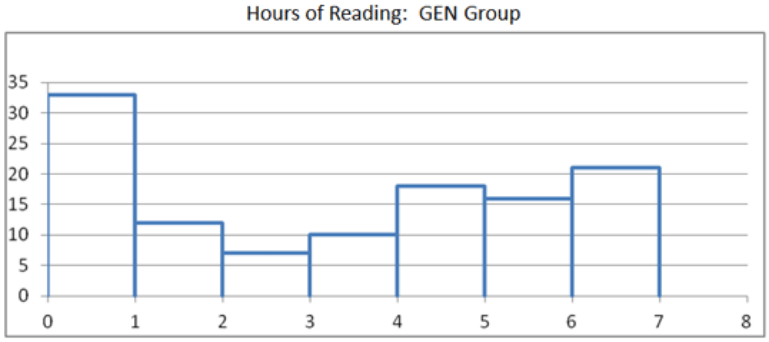
Problem 6.13 Continue from the problem above. You should have obtained the summary statistics

Summary Statistics:

	DEV Group Group 1	GEN Group Group 2
Mean:	$\bar{x}_1 = 4.1$	$\bar{x}_2 = 3.7$
Standard Deviation:	$s_1 = 2.4$	$s_2 = 2.5$
Sample Size:	$n_1 = 204$	$n_2 = 117$

An appropriate graph to illustrate the two populations is found below.





1. Find the 95% confidence interval for the difference of the means.
2. Based off your interval, what conclusions can you draw about the difference between the two groups? Present your observations in an English sentence, relating the result to the context of the problem.

World Cup Heart Attacks

Do intense sporting events increase the probability of a person having a heart attack? We will consider this question in the next example.

The FIFA Football (Soccer) World Cup is held every four years and is one of the biggest sporting events in the world. In 2006, Germany hosted the World Cup. A study was conducted by Dr. Wilbert-Lampen, et. al. to determine if the stress of viewing a soccer match would increase the risk of a heart attack or another cardiovascular event. [**?**]

The 2006 World Cup was held from June 9, 2006 to July 9, 2006. The number of patients suffering cardiovascular events (e.g. heart attacks) was obtained from medical records of patients in the Greater Munich (Germany) area during this time period. To provide a control group, counts of patients suffering cardiovascular events was recorded from May 1 to June 8 and July 10 to July 30, 2006, as well as May 1 to July 30 in 2003 and 2005. The year 2004 was avoided, due to the European Soccer Championships held in Portugal. These data were extracted from Figure 1 in the article by Wilbert-Lampen, [**?**] and are given in the file WorldCupHeartAttacks.xlsx.

We will use the data on cardiovascular problems during the World Cup to test the hypothesis that the mean number of cardiovascular events is greater during the World Cup than during the control period. Let $\alpha = 0.01$.

Problem 6.14 Perform an appropriate hypothesis test to answer the question above. The steps below serve as a guide of what you should do in any

hypothesis test. Eventually I'll stop giving these guides, and you should know what needs to be reported without a reminder.

1. State an appropriate null and alternative hypotheses and give the level of significance. Note that we want to determine if intense sporting events increase a probability.

2. Give the relevant summary statistics, and verify the requirements have been met. Then give the test statistic and its value, state the degrees of freedom, and give the P-value.

3. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

Problem 6.15 Use the Data Analysis Toolpack[1] in Excel to compute the summary statistics, the test statistic, and the P-value for this problem. In addition, obtain a confidence interval for the difference in the means. Compare these results to your answers in the previous question.

Chronic Obstructive Pulmonary Disease (COPD)

The National Heart Lung and Blood Institute gives the following explanation of COPD [**?**]:

> COPD, or chronic obstructive pulmonary (PULL-mun-ary) disease, is a progressive disease that makes it hard to breathe. "Progressive" means the disease gets worse over time.
> COPD can cause coughing that produces large amounts of mucus (a slimy substance), wheezing, shortness of breath, chest tightness, and other symptoms.
> Cigarette smoking is the leading cause of COPD. Most people who have COPD smoke or used to smoke. Long-term exposure to other lung irritants, such as air pollution, chemical fumes, or dust, also may contribute to COPD.

A study was conducted in the United Kingdom to determine if there is a difference in the effectiveness of community-based rehabilitation program compared to hospital-based rehabilitation [**?**]. Patients suffering from COPD were randomly assigned to either the community or hospital group. Twice a week for six weeks, they participated in two-hour educational and exercise sessions. Patients were also encouraged to exercise between sessions.

The effectiveness of the program was measured based on the total distance patients could walk at one time at a particular pace. This is called the endurance shuttle walking test (ESWT). This was measured at the beginning of the study and again at the end of the six week rehabilitation period. Data representing the improvement of the patients in each group is given in the file COPD-Rehab-stacked.xlsx. The data represent the increased distance (in meters) that each patient could walk. Negative values indicate that the patient was not able to walk as far at the end of the rehabilitation treatment as at the beginning.

Because hospital-based rehabilitation tends to be more expensive, the researchers wanted to assess if there is a significant difference in the patients'

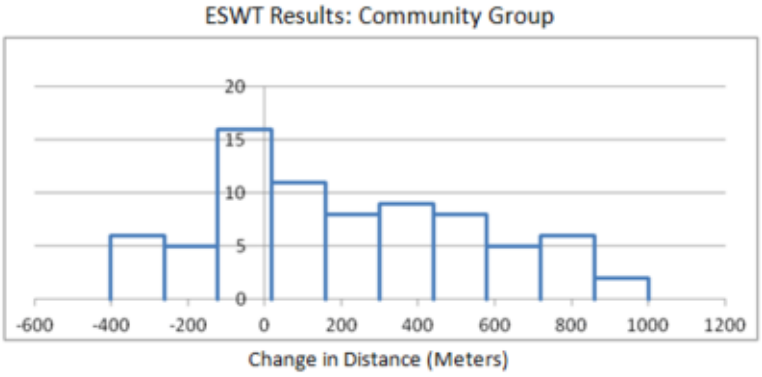
[1] The help files in Excel give step-by-step instructions for installing and using the Data Analysis Toolpack.

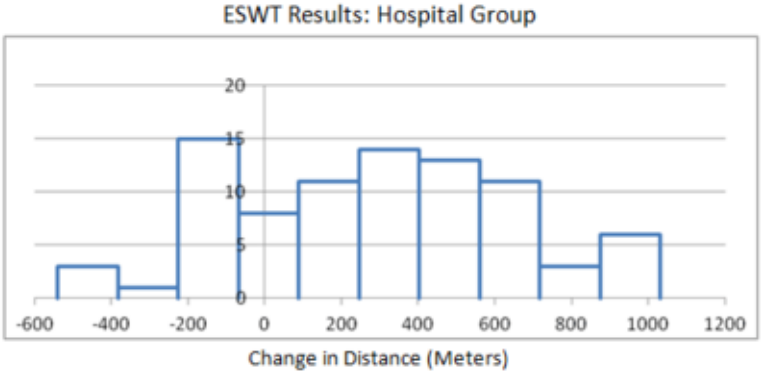
improvement under the two programs. If not, then it makes sense to refer patients to the less expensive treatment option. The purpose of this study was to determine if pulmonary rehabilitation in a community setting is as effective as rehabilitation in a hospital setting.

Problem 6.16 The relevant summary statistics are given below.

Location	
Community	Hospital
Group 1	Group 2
$\bar{x} = 216.1$	$\bar{x}_2 = 283.4$
$s = 339.9$	$s = 359.9$
$n = 76$	$n = 85$

Here is an appropriate graph to illustrate the data.





1. Based on the graphs, does there appear to be a difference in the mean responses of the two groups? Note that this is just a visual observation. We need to conduct a hypothesis test or create a confidence interval to verify any conclusion we draw from a test. It's easy to be deceived by a graph, which is why we create intervals and/or conduct hypothesis tests.

2. Verify the requirements have been met, and then find the 95% confidence interval for the difference in the mean responses of the two groups.

3. Present your observations in an English sentence, relating the result to the context of the problem. In particular, does there appear to be a difference in the mean improvement observed between the two groups?

4. Repeate the above, but now with a 90% confidence interval for the difference in the mean responses of the two groups, instead of a 95% confidence interval.

5. Why is the 95% confidence interval wider than the 90% confidence interval?

Problem 6.17 Use the Data Analysis Toolpack in Excel to compute the confidence intervals in the previous problem.

Exam Review

To prepare for the exam, we will spend one day of class focused on review. The goal of this class period is to have each student find/create examples to illustrate each of the big ideas from the material we have been studying. You'll see a list of the big ideas on the next page. To prepare for class, here is your assignment.

1. For each concept on the next page, find or create an example that illustrates the concept.

2. Organize your work into a "lesson plan" where you include the problem and key steps to solving that problem on your lesson plan.

3. Come to class and spend 1 hour with a partner teaching from the ideas in your lesson plan. The idea here is to let each person share an example, then swap roles. If you both feel like you need more practice in a specific area, spend your time there.

4. Report in I-Learn that you have completed your lesson plan and taught it to your peer, and upload your document to I-Learn.

Chapter 4 - The Central Limit Theorem

1. Compute the expected value of sums and constant multiples of random variables.
2. Explain how to compute the variance of sums and multiples of independent random variables. Use this to compute the uncertainty of problems involving sums and multiples of random variables.
3. Explain why $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}} = \sigma_X/\sqrt{n}$, and be able to apply this fact in appropriate situations.
4. Use the normal probability applet to address various types of problems related to normal distributions.
5. Explain what the central limit theorem says, and what assumptions we must make to apply the central limit theorem.

Chapter 5 - Inference For a Single Population Mean

1. Construct confidence intervals for a single population mean when σ is either known or unknown.
2. Appropriately conduct hypothesis tests for a single mean when σ is either known or unknown.
3. Compute the sample size needed to obtain a desired margin of error.
4. Set up appropriate null and alternative hypotheses. Explain when you should choose to use a one-sided versus two-sided test.
5. Define the level of significance α. Define the p-value of a hypothesis test. Explain how we use these two numbers to make decisions in a hypothesis test.
6. Explain what it means to say that we are 95% confident that the mean lies in some interval. In other words, explain what a confidence interval for μ represents.

Chapter 6 - Inference for Two Population Means

1. Compare and contrast a matched pairs design versus an independent two sample design.
2. Appropriately conduct a hypothesis test in a matched pairs design.
3. Construct and interpret confidence intervals for the change in some value in a matched pairs design.
4. Explain how to obtain the standard deviation of $\bar{X} - \bar{Y}$ when we assume X and Y are independent.
5. Appropriately conduct a hypothesis test in an independent samples design.
6. Construct and interpret confidence intervals for difference between two means in an independent samples design.

Part III

Looking for Correlation

Chapter 7

Analysis Of Variance (ANOVA)

ANalysis Of VAriance (ANOVA) is a test for equality of several means. This procedures provides a way for us to compare the means for several groups in one hypothesis test. In ANOVA, we compare the ratio of the variability between groups (from one group to another) to the variability within the groups (observations in the same group). If there are large differences in the means of the groups compared to the variability in each group, we conclude the means differ. Since we are only conducting one test, the probability of committing a Type I error is controlled at the $\alpha = 0.05$ level. We'll introduce the idea in class using the Wolfram Demonstrations applet at http://demonstrations.wolfram.com/VisualANOVA/.

7.1 One Way ANOVA

The Effects of Gratitude

Robert Emmons and Michael McCullough investigated the effects of gratitude on people's perception of life as a whole [**?**]. In a study of $n = 192$ undergraduates, the participants were randomly assigned to one of three groups.

- Group 1 (Gratitude): The participants in this group were asked to record five things each week for which they were grateful or thankful.
- Group 2 (Hassles): The volunteers in this group recorded five irritants that had occurred to them in the previous week.
- Group 3 (Events): The people in the events group recorded five things that occurred in the past week that had an impact on them.

In addition to the weekly record of the five things they recorded their level of satisfaction with life in general. (Higher values are more favorable.) Reports were collected for nine weeks, and the overall level of satisfaction with life as a whole was recorded for each individual.

The researchers wanted to determine if there was a difference in the perception of life as a whole between the subjects assigned to each of the three groups. Stated differently, they wanted to determine if expressing gratitude affects a person's view of life in general.

Here is an excerpt of data representing the results of this study:

President Gordon B. Hinckley said, "My plea is that we stop seeking out the storms and enjoy more fully the sunlight. I am suggesting that as we go through life, we 'accentuate the positive.' I am asking that we look a little deeper for the good, that we still our voices of insult and sarcasm, that we more generously compliment and endorse virtue and effort" (*Standing for Something*, 2000, p.101).

Grateful	Hassles	Events
6.1	5.9	4.9
6.2	5.4	4.2
3.7	3.7	4.1
4.4	4.4	4.8
5.7	3.8	6.7
7.5	3.4	3.6
5.4	5.3	4.8
4.3	5.4	4.7
4.8	4.2	4.4
6	5.5	5.1
5.6	6.1	5.4
4.3	4.6	4

Higher values indicate a greater level of satisfaction with life as a whole.

How might we analyze these data? One possible method would be to conduct separate t-tests for all the possible pairs of groups in the study. If we did this, we would need to conduct a separate t-test to compare groups 1 & 2, 1 & 3 and 2 & 3. If the probability of committing a Type I error is $\alpha = .05$ on each of these tests, then the probability that we would commit a Type I error on at least one of the tests is much greater than 0.05. We need a hypothesis test that we can use to compare all the groups at once. The procedure that allows us to do this is called Analysis of Variance (ANOVA).

Null and Alternative Hypotheses

ANOVA is a test for equality of several means. It allows us to compare the means for several groups–in one hypothesis test. It is based on a comparison of the spread of the data within each of the groups compared to the spread of the means of the groups.

In an ANOVA test, the null hypothesis is typically expressed in words:

$$H_o : \text{All the means are equal}$$

The alternative hypothesis is given as:

$$H_a : \text{At least one of the means is different than the others.}$$

If the means differ from each other in comparison to the variability in each group, then we conclude that the means are not all equal. If the means do not differ by much (when compared to the spread of the data in each group) then we do not reject the hypothesis that all the means are equal. We will use the level of significance, α, and the P-value just as we have in the other hypothesis tests.

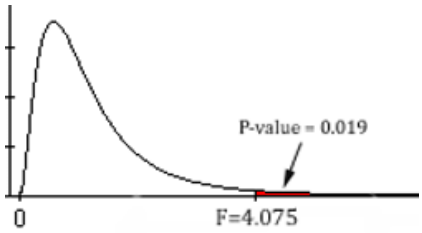
F-distribution

The sample test statistic follows an F-distribution. This is the first time we have encountered this distribution. In previous tests, we have used z and t as sample test statistics. For the ANOVA test, we always use the F-statistic.

Here is a brief summary of the characteristics of the F-distribution:

- It is right skewed.

• The values of F are never negative.
• The P-value for the ANOVA test is the area in the *right tail*. We will never divide the area in the tail.



Problem 7.1 Fill in the blanks in the following questions.

1. The shape of the F-distribution is typically illustrated as skewed to the _____.
2. The F statistic is always greater than or equal to _____.
3. The t-statistic has one number that represents the degrees of freedom. An F-statistic has _____ numbers to represent its degrees of freedom (you'll need to look this one up).
4. The P-value in an ANOVA test is always an area in the _____ (right/left) tail of the F-distribution.
5. Use Excel to compute the P-value if $F = 2.3$ when we want to compare the means of 3 groups and have collected data from 38 different subjects (spread across the three groups). You should obtain something between 0.11 and 0.12.

Assumptions of ANOVA

There are two assumptions of ANOVA that must be checked:

- The data are normally distributed in each group.
- The variances are equal for each group.

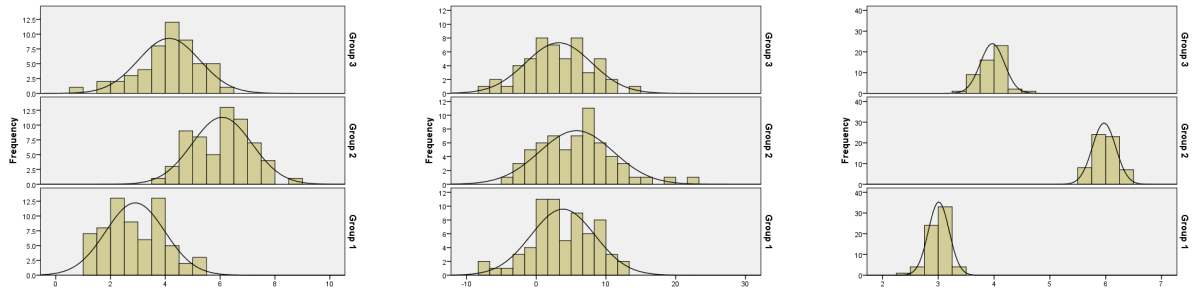
To check the first requirement, we can make a histogram of the data from each group. To check the second, we examine the sample standard deviations. If the standard deviations are close together, then the variances must be close together as well. Many people use the following rule: *if the largest standard deviation is less than or equal to two times the smallest standard deviation, then we will assume that the variances are close enough to use ANOVA.*

If done by hand, the calculations for one simple ANOVA problem can easily require an hour of hard work. Spreadsheets can rapidly decrease the time needed, and statistical software automates all the computations.

ANOVA Model

ANOVA is modeled under the assumption that the observed data values are based on two components: (*i*) the population mean group from which the data were drawn and (*ii*) a random "error" term.[1] The notation used to depict the

[1]The term "error" is antiquated and slightly misleading. It comes from early measurements of the position of stars. Early astronomers made repeated measurements on the position of stars. Recognizing the variation in the measurements was not due to star movement, but imprecision in their measurements, they called the deviation from the expected value "error". The term is still in use today. A component of a model that indicates deviations from a mean is often called the error term.



Moderate Standard Deviation Large Standard Deviation Small Standard Deviation

Figure 7.1: In the three figures above, the population means from which the three groups of data were drawn are $\mu_1 = 3$, $\mu_2 = 6$, and $\mu_3 = 4$. The only parameter that differs from one plot to another is the standard deviation. A large standard deviation makes it difficult to detect if there is a difference in the means. A small standard deviation makes it extremely simple to see that the means differ. ANOVA helps us make decisions when the standard deviation within groups is moderate, by giving a P-value that tells us the probability of observing a difference in the means at least as large as what we observed, assuming that the the means are equal.

ANOVA model can vary, but the elements are the same. For one-way ANOVA,[2] the response of the j^{th} unit[3] (subject) in the i^{th} group is denoted Y_{ij}. In an equation, this value is given as:

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{7.1}$$

where μ_i is the mean of group i, and ϵ_{ij} is the random error associated with this observation.

The error terms are assumed to follow a normal distribution with mean 0 and a common standard deviation, σ. So in ANOVA we assume the distribution of the three populations is the same, except for the mean. This is illustrated in the side-by-side histograms in Figure 7.1.

Imagine collecting data from three three groups. By creating a side-by-side histogram, you might be able to detect that the data come from populations with different means. This would depend on how large or small the population standard deviations are, in relation to how varied the means are.

Notice that it would be more difficult to detect a difference in the means if we were examining data from the three populations illustrated in the middle picture of Figure 7.1. Looking at data with a large standard deviation relative to the difference of the means, we could easily conclude that the there is no reason to believe the means differ.

Now, if the standard deviation was very small, as in the third picture in Figure 7.1, notice how easy it would be to detect a difference in the means. When the standard deviation is very small, comparing the histograms of the three groups should lead to the conclusion that the means are different.

The principal underpinning of ANOVA is a comparison of the variability between groups[4] to the variability within groups[5]. We take a ratio of these two

[2]One-way ANOVA is sometimes called single-factor ANOVA.

[3]An experimental unit, or unit for short, is the individual from which the data are gathered. Sometimes we refer to units as subjects. The datum collected from the unit is called an observation.

[4]This is estimated by finding the variability of the group means. In other words, we estimate how spread out the population means are from one another.

[5]This is estimated by computing a pooled estimate of the variance within the groups. In

estimators.

- If the ratio is large, then we conclude that the variability between groups is large compared to the variability within groups. This suggests that the means are significantly different from each other. This is illustrated in the histogram labeled "smaller standard deviation". The three distributions appear to be very distinct from each other.
- If the ratio is smaller, then the variability in the group means is small compared to the variability within each group. This is illustrated by the histogram labeled "larger standard deviation". In this case, we would not conclude that there is a difference in the means. Notice how similar the three distributions appear to be in this plot.

Formulas

Suppose we collect observations from $J = 3$ groups, where the sample size of each group is n_1, n_2, and n_3, and we let $n = n_1 + n_2 + n_3$ be the total number of observations. For each group, we compute the sample means $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$ and standard deviations s_1, s_2, s_3. We also compute the sample mean of all observations $\bar{x}$ which we could have done using the weighted average formula $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_2\bar{x}_3}{n}$. These are our summary statistics. We'll use the notation $x_{j,i}$ to represent the ith observation from the jth group. We've summarized this information in the table below. We use the abbreviation SS to stand for "Sum of Squares."

Group	Sample Size	Sample Mean	Sample Variance
1	n_1	$\bar{x}_1$	$s_1^2 = \frac{\sum_1^{n_1}(x_{1,i}-\bar{x}_1)^2}{n_1-1} = \frac{SS_{Group1}}{n_1-1}$
2	n_2	$\bar{x}_2$	$s_2^2 = \frac{\sum_1^{n_2}(x_{2,i}-\bar{x}_2)^2}{n_2-1} = \frac{SS_{Group2}}{n_2-1}$
3	n_3	$\bar{x}_3$	$s_3^2 = \frac{\sum_1^{n_3}(x_{3,i}-\bar{x}_3)^2}{n_3-1} = \frac{SS_{Group3}}{n_3-1}$
Total	$n = n_1 + n_2 + n_3$	$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_2\bar{x}_3}{n}$	$\frac{\sum_{j=1}^{3}\sum_1^{n_j}(x_{j,i}-\bar{x})^2}{n-1} = \frac{SS_{Total}}{n-1}$

Notice that every variance is computed by considering the sum of the square, and then dividing by an appropriate degrees of freedom. We measure variability by summing the squares. There are three variabilities we measure in ANOVA.

Definition 7.1. The treatment sum of squares, error sum of squares, and total sum of squares, are defined below. Note that the variable J represents the number of groups in our experiment.

1. ($\text{SS}_{\text{Treatement}}$) We want to know how far the sample means $\bar{x}_j$ are from the grand mean $\bar{x}$. If one group has more observations than another group, than this needs to be taken into account. The corresponding formula is

$$\text{SS}_{\text{Treatement}} = \text{SSTr} = \sum_{j=1}^{J} n_j(\bar{x}_j - \bar{x})^2.$$

   We call this the treatment sum of squares.

other words, we estimate how spread out the y's are within each group.

2. (SS_{Error}) We want to know how much each observation varies from its respective mean. The formula is

$$\text{SS}_{\text{Error}} = \text{SSE} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{x}_{j,i} - \bar{x}_j)^2 = \sum_{j=1}^{J}(n_j - 1)s_j^2.$$

   We call this the error sum of squares. The second formula above is easier to use if we have already compute the sample standard deviations of each group.

3. (SS_{Total}) We want to know how much each observation varies from the grand mean $\bar{x}$. The formula is

$$\text{SS}_{\text{Total}} = \text{SST} = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(\bar{x}_{j,i} - \bar{x})^2.$$

   We call this the total sum of squares.

Problem 7.2 Show that the sum of squares terms defined above satisfy the equation

$$\text{SST} = \text{SSTr} + \text{SSE}.$$

This shows that in general we only need to consider two of the terms above, as the third can quickly be obtained from the other two. [If you get stuck on this one, then move on.]

Hardiness of A Welded Joint

A common task in engineering design or scientific research is to determine if a certain factor has a significant effect on a particular outcome of interest. For example, suppose that the hardness of a welded joint is an important characteristic of a particular design. Suppose also that welds can be made with 4 different types of welding flux. A natural question would be: "Does the type of welding flux affect the hardness of the welded joint?" This type of question can be answered by running an ANOVA experiment in which welds from the different fluxes are made and hardness readings are taken and analyzed.

Hardness values of 5 samples made from each of the 4 different fluxes are shown below.

Flux	Sample Values
A	250, 264, 256, 260, 239
B	263, 254, 267, 265, 267
C	257, 279, 269, 273, 277
D	253, 258, 262, 264, 273

The question of interest is the following:

> Are the differences in sample means due to random variation in the experimental procedure only, or is there a real difference in sample means for the different fluxes?

This is precisely what statistics helps us answer.

Problem 7.3 There 4 groups in which we computed hardiness values. For each of the 4 groups, state the sample size, the sample mean, and sample standard deviation, in other words give the summary statistics. Then give the values of SSTr and SSE, and sum them to give the value of SST. Put your results in the table below.

Flux	Group #	n_j	$\bar{x}_j$	s_j
A B C D	1 2 3 4			
Total	-	20	$\bar{x} =$	

SSTr	
SSE	
SST	

We are almost done with our ANOVA test. Notice that when we compute a standard deviation, we compute a sum of squares and then divide by a degrees of freedom. This is precisely the next step. The degrees of freedom of SSTr, SSE, and SST are precisely $J-1$, $n-J$, and $n-1$. When divide a sum of square by its degrees of freedom, we call it a mean square. This gives us the treatment mean square, error mean square, and total means square using the formulas

$$\text{MSTr} = \frac{\text{SSTr}}{(J-1)}, \quad \text{MSE} = \frac{\text{SSE}}{(n-J)}, \quad \text{and} \quad \text{MST} = \frac{\text{SST}}{(n-1)}.$$

The F statistic is the ratio

$$F = \frac{\text{MSTr}}{\text{MSE}},$$

and the degrees of freedom of the F statistics are $J-1$ and $N-J$, namely the degrees of freedom of the numerator and the degrees of freedom of the denominator.

Problem 7.4 Recall that the null hypothesis in an ANOVA test is that all the population means are equal. Suppose we collected some data and actually observed that the sample means were identically the same, and the sample standard deviations were the same.

1. What does MSTr to equal if the sample means are exactly the same. In this case, what value would we obtain for F
2. Why should we expect MSE to equal the common variance σ^2, regardless of whether the means are equal or not?
3. The F statistic is the ratio MSTr/MSE. If the error between treatments (MSTr) significantly greater than the error within treatments (MSE), so we have a large F statistics, then what conclusion should we draw?

Problem 7.5 Let's continue problem 7.3. You should have obtained SSTr = 743.4 and $SSE = 1023.6$.

1. Give the degrees of freedom of each of these two quantities.
2. Compute the treatment mean square MSTr and error mean square MSE.
3. State the F-statistic and use Excel to obtain the corresponding P-value (you should get about $P = 0.029$).
4. What answer would you give to the original question, "Does the type of welding flux affect the hardness of the welded joint?"

More Gratitude Give Me

The students in the gratitude study were randomly assigned to one of the three treatments. They wrote in a weekly journal, according to their group assignment. At the end of the semester, they completed a questionnaire that asked about their attitude toward life. The responses on the survey were coded into a number, where higher numbers represent a more positive outlook. The data are given in the file Gratitude.xlsx.

We will conduct a hypothesis test to determine if the mean responses of the individuals in the three groups differ.

Problem 7.6 Do the following.

1. State the null and alternative hypotheses. We will use the $\alpha = 0.05$ level of significance.
2. Give the relevant summary statistics.
3. Verify the requirements have been met.
4. Give the test statistic and its value, as well as the degrees of freedom.
5. Find the P-value and compare it to the level of significance.
6. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem

---

If we take a closer look, we see that the *Hassles* and *Events* groups had means that were fairly close together. However, the *Grateful* group appears to have a significantly higher mean level of satisfaction than the other two groups.

"Piled Higher and Deeper" by Jorge Cham

Soccer Shoes

Nike, a company that makes sporting goods including shoes, funded a study to compare five soccer shoe designs. [**?**, **?**] The objective of the research was to determine if there is a difference in the mean accuracy soccer players achieve using different Nike shoe designs.

As part of the research, they asked trained soccer players to kick a ball at a target. The target was placed 115 cm above the ground and at a distance of 10 m from the players. Using electronic equipment, the researchers recorded the distance from the center of the target to the point where the ball hit. The objective of the research was to assess if footwear could affect the accuracy of a soccer player.

The subjects wore five different soccer shoes and for one treatment they kicked the ball in stocking feet. Due to the proprietary nature of the data, the shoes are only labeled "A," "B," "C," "D," and "E" in the article. Data representing the results of this study are given in the file SoccerShoes.xlsx. Let $\alpha = 0.10$.

Problem 7.7 Answer the following questions.

1. State the null and alternative hypotheses and the level of significance.
2. Give the relevant summary statistics.
3. Verify the requirements have been met.
4. Give the test statistic and its value.
5. State the degrees of freedom.
6. Find the P-value and compare it to the level of significance.
7. State your decision. State your decision. Present your conclusion in an English sentence, relating the result to the context of the problem.

Chapter 8

Linear Regression

Suppose we have collected several data points (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n). In class we have already defined and shown the following:

$$\begin{aligned}
S_{xx} &= \sum (x_i - \bar{x})^2 && = \sum (x_i^2) - n\bar{x}^2 && = \sum (x_i^2) - \frac{1}{n}\left(\sum x_i\right)^2 \\
S_{yy} &= \sum (y_i - \bar{y})^2 && = \sum (y_i^2) - n\bar{y}^2 && = \sum (y_i^2) - \frac{1}{n}\left(\sum y_i\right)^2 \\
S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) && = \sum (x_i y_i) - n\bar{x}\bar{y} && = \sum (x_i y_i) - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)
\end{aligned}$$

We also defined the correlation coefficient to be

$$r = \frac{1}{n-2} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

The least squares regression line is the line $\hat{y} = \hat{a} + \hat{b}x$ that minimize the sum of the squares of the residuals, which are

$$e_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i).$$

To find the coefficients $\hat{a}$ and $\hat{b}$, we compute the partial derivatives of

$$\sum e_i^2 = \sum (y_i - (\hat{a} + \hat{b}x_i))^2$$

with respect to both $\hat{a}$ and $\hat{b}$, set each equal to zero, and solve the corresponding system of equations to obtain

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \quad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

This gives us the ability to find a line that best approximates the linear trend in a sample of data. The real question in statistics lies in being able to not just find such a line, but also being able to give a confidence interval for our estimate of the coefficients. In other words, the line $\hat{y} = \hat{a} + \hat{b}x$ is a line that approximates the trend in our sample. We would like to find the line $y = \beta_0 + \beta_1 x$ that best approximates the linear trend in the entire sample. Since it's probably impossible to obtain the entire population's data, we use the statistics $\hat{a}$ and $\hat{b}$ to approximate the parameters β_0 and β_1. We need confidence intervals to tell us a plausible range of values for these unknown parameters.

Problem 8.1 Show that the sum of the residuals is zero. In other words, show that $\sum e_i = 0$. This shows that the expected residual is zero.

When we were performing ANOVA tests, one of our assumptions was that the variance for each group was equal. In regression, each different x value gives us a different group. This generally means there are an infinite number of groups. For each x value we could compute the variance of possible y values that correspond to this fixed x value. One of the assumptions needed to create confidence intervals is that this variance is the same regardless of which x value we choose. More formally, for each x value we assume that the corresponding y values are normally distributed with a mean that lines on the line (so $\mu_y = \beta_0 + \beta_1 x$) and common standard deviation σ.

To approximate this standard deviation σ, the residuals are the key. The residuals measure the distance from the sample regression line. So if we square the residuals and then average them, we obtain an approximation for the variance σ^2. However, the residuals measure the distance to the sample regression line, not the population regression line. As such, using the formula $\sigma^2 \approx \sum e_i^2/n$ leads to a biased estimate of σ, and is in general too small. The appropriate correction is to divide by $n-2$ instead of n when averaging. As a point estimate of σ^2, we'll use the formula

$$S_e^2 = \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}\sum (y_i - (\hat{a} + \hat{b}x_i))^2$$

Problem 8.2 Show that we can instead estimate the common variance using the formula

$$S_e^2 = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}(n-2)}.$$

This is the formula that shows up on the FE Exam packet. You'll see that they use MSE instead of S_e^2 in their work.

In your work, you'll need the fact that

$$\sum (y_i - (\hat{a} + \hat{b}x_i))^2 = (1 - r^2)\sum (y_i - \bar{y})^2.$$

You can assume this is true in this problem. As an optional problem, prove that this fact is true.

Now that we have an estimate of for the common variance σ, we can start using this to obtain our confidence intervals for the intercept β_0 and slope β_1. We'll start with the slope, as it's simpler.

Problem 8.3 We have already shown that $\hat{b} = \frac{S_{xy}}{S_{xx}}$. Show that we could instead write

$$\hat{b} = \sum_i \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} y_i = \sum_i \frac{(x_i - \bar{x})}{S_{xx}} y_i.$$

The problem above shows that we can think of $\hat{b}$ as a random variable that we create by multiplying each y_i by some constant and then summing the results. We've already shown that if we have a collection of random variables Y_i together with some constants c_i, then the expected value and variance of $c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n = \sum c_i Y_i$ are

$$E\left[\sum c_i Y_i\right] = \sum c_i E[Y_i]$$
$$Var\left[\sum c_i Y_i\right] = \sum c_i^2 Var[Y_i]$$

In regression, we create our regression line $\hat{y}_i = \hat{a} + \hat{b}x_i$ as an estimate for the population regression line $\mu_{\hat{y}_i} = \beta_0 + \beta_1 x_i$. For a given x_i, the expected value of any y corresponding to that particular x_i value is assumed to be on the line. This means that we have $E[y_i] = \beta_0 + \beta_1 x_i$.

Problem 8.4 In the previous problem you showed that

$$\hat{b} = \sum \frac{(x_i - \bar{x})}{\sum (x_j - \bar{x})^2} y_i = \frac{(x_1 - \bar{x})}{S_{xx}} y_1 + \frac{(x_2 - \bar{x})}{S_{xx}} y_2 + \cdots + \frac{(x_n - \bar{x})}{S_{xx}} y_n.$$

Use this to compute the expected value $E[\hat{b}]$ and show that $\mu_{\hat{b}} = \beta_1$. In other words, the expected value of our sample slope actually is the slope of the population regression line.

Problem 8.5 Recall that we have assumed that $Var[y_i] = \sigma$ regardless of which i we choose. Use this to compute the variance of $\hat{b}$, and show that

$$Var[\hat{b}] = \frac{\sigma^2}{S_{xx}}.$$

This gives us the standard error we need to compute confidence intervals. Since we don't actually know the value σ, we instead use the point estimate S_e. Our confidence interval for β_1 is simply

$$\hat{b} \pm t^* \frac{S_e}{\sqrt{S_{xx}}}.$$

where the degrees of freedom for t^* is $df = n - 2$.

Part IV

In Progress

8.1 Two Way ANOVA

8.2 2^p Factorial Design

"Piled Higher and Deeper" by Jorge Cham

Part V

Appendix

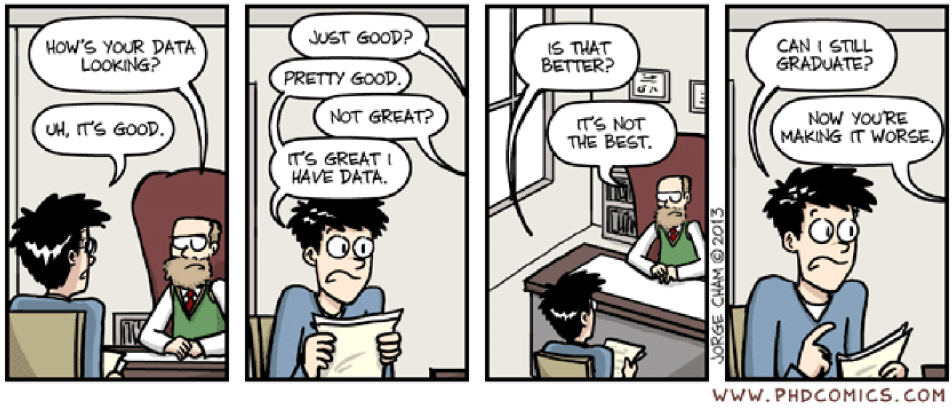
Chapter 9

Supplemental Data

9.1 List of Data Files

- http://emp.byui.edu/johnsonc/Data/M330/BLEU-Scores.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/BodyTemp.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/BYU-IdahoGradSalaries.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/BYU-IdahoGradSalaries-Sampling.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/DirectFlightCosts.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/EthanAllenPassengers.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/EuroWeight.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/Mahon.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/MysteryRVs.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/NosocomialInfections.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/PineBeetle.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/ReadingPractices.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/SoccerShoes.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/WorldCupHeartAttacks.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/REE-ClassicalMusic.xlsx
- http://emp.byui.edu/johnsonc/Data/M330/YieldStrength.xlsx

"Piled Higher and Deeper" by Jorge Cham



Bibliography

[1] K. Tan, K. Tong, and C. Tang. Direct strut-and-tie model for prestressed deep beams. *Journal of Structural Engineering*, pages 1076–1084, 2001.

[2] W. Robert Johnston. Naval reactor accidents causing radiation casualties. Website, accessed Sept 2013. `http://www.johnstonsarchive.net/nuclear/radevents/radevents3.html`.

[3] P. Pancratz. *Statistical Casde Studies for Industrial and Process Improvement*. SIAM–ASA, 1997.

[4] J. P. W. Verbiest, J. M. Weisberg, A. A. Chael, K. J. Lee, and D. R. Lorimer. On pulsar distance measurements and their uncertainties. *The Astrophysical Journal*, 755(39), 2012.

[5] Stéphane Ducasse, Matthias Rieger, and Serge Demeyer. A language independent approach for detecting duplicated code. In *Software Maintenance, 1999.(ICSM'99) Proceedings. IEEE International Conference on*, pages 109–118. IEEE, 1999.

[6] Danny Chu, Faisal G. Bakaeen, Tam K. Dao, Scott A. LeMaire, Joseph S. Coselli, and Joseph Huh. On-pump versus off-pump coronary artery bypass grafting in a cohort of 63,000 patients. *The Annals of Thoracic Surgery*, 87(6):1820–1827, 2009.

[7] National Transportation Safety Board. Capsizing of new york state-certificated vessel *ethan allen*, lake george, new york, october 2, 2005. Technical Report NTSB PB2006-916403, National Transportation Safety Board, Washington, D.C., 2006.

[8] Centers for Disease Control and Prevention. 2000 CDC growth charts for the United States: Methods and development. website, May 2002. `http://www.cdc.gov/growthcharts/2000growthchart-us.pdf`.

[9] P. A. Mackowiak, S. S. Wasserman, and M. M. Levine. A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association*, 268(12):1578–1580, September 1992.

[10] Steven M. Horvath, H. Menduke, and George Morris Piersol. Oral and rectal temperatures of man. *Journal of the Americal Medical Association*, 144(18):1562–1565, 1950.

[11] Carl Reinhold August Wunderlich. *Das Verhalten der Eigenwärme in Krankheiten*. Otto Wigand, Leipzig, Germany, 1868.

[12] Carl Reinhold August Wunderlich. *Medical Thermometry and Human Temperature*. William Wood & Co, New York, NY, USA, 1871.

[13] Märtha Sund-Levander, Christina Forsberg, and Lis Karin Wahren. Normal oral, rectal, tympanic and axillary body temperature in adult men and women: a systematic literature review. *Scandinavian Journal of Caring Sciences*, 16:122–128, 2002.

[14] W. S. Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[15] Philipp Koehn, editor. *Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models*, pages 115–124. Lecture Notes in Computer Science. Springer Berlin, Heidelberg, Germany, 2004.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, July 2002.

[17] P. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. Rossin. A statistical approach to machine translation. *Computational Linguistics*, 16:76–85, 1990.

[18] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain, July 2004.

[19] Shkedy, Ziv and Aerts, Marc and Callaert, Herman. Journal of statistics education–data archive. Website. `http://www.amstat.org/publications/jse/datasets/euroweight.txt`.

[20] Jennifer G. Klutsch, José F. Negrón, Sheryl L. Costello, Charles C. Rhoades, Daniel R. West, John Popp, and Rick Caissie. Stand characteristics and downed woody debris accumulations associated with a mountain pine beetle (Dendroctonus ponderosae Hopkins) outbreak in Colorado. *Forest Ecology and Management*, 258:641–649, 2009.

[21] Arthur R Cushny and A Roy Peebles. The action of optical isomers ii. hyoscines. *The Journal of physiology*, 32(5-6):501–510, 1905.

[22] A. K. Mahon, M. G. Flynn, H. B. Iglay, L. K. Stewart, C. A. Johnson, B. K. McFarlin, and W. W. Campbell. Measurement of body composition changes with weight loss in postmenopausal women: a comparison of methods. *J Nutr Health Aging*, 11(3):203–213, 2007.

[23] A Asensio Vegas, V Monge Jodra, and M Lizn Garca. Nosocomial infection in surgery wards: A controlled study of increased duration of hospital stays and direct cost of hospitalization. *European Journal of Epidemiology*, 9(5):504–510, September 1993.

[24] Ebba Carlsson, Hannah Helgegren, and Frode Slinde. Resting energy expenditure is not influenced by classical music. *Journal of Negative Results in BioMedicine*, 4:6, 2005.

[25] Arlene M. Butz, Michael Crocetti, Richard E. Thompson, and Paul H. Lipkin. Promoting reading in children: Do reading practices differ in children with developmental problems? *Clinical Pediatrics*, 48(3):275–283, 2009.

[26] Ute Wilbert-Lampen, David Leistner, Sonja Greven, Tilmann Pohl, Sebastian Sper, Christoph Völker, Denise Güthlin, Andrea Plasse, Andreas Knez, Helmut Küchenhoff, and Gerhard Steinbeck. Cardiovascular events during world cup soccer. *British Medical Journal*, 358(5):475–483, January 31, 2008.

[27] National Heart Lung and Blood Institute, National Institutes of Health, U.S. Department of Health & Human Services. What is chronic obstructive pulmonary disease (COPD)? Website, March 2009. `http://www.nhlbi.nih.gov/health/dci/Diseases/Copd/Copd_WhatIs.html`.

[28] J. C. Waterhouse, S. J. Walters, Y. Oluboyede, and R. A. Lawson. A randomised 2×2 trial of community versus hospital pulmonary rehabilitation for chronic obstructive pulmonary disease followed by telephone or conventional follow-up. *Health Technology Assessment*, 14(6), 2010.

[29] Robert A. Emmons and Michael E. McCullough. Counting blessings versus burdens: An experimental investigation of gratitude and subjective well–being in daily life. *Journal of Personality and Social Psychology*, 84(2):377–389, 2003.

[30] Ewald M. Hennig, Katharina Althoff, and Ann-Kathrin Hoemme. Soccer footwear and ball kicking accuracy. *Footwear Science*, 1(S1):85–87, 2010.

[31] Ewald M. Hennig and Thorsten Sterzing. The influence of soccer shoe design on playing performance: a series of biomechanical studies. *Footwear Science*, 2(1):3–11, 2010.

"Piled Higher and Deeper" by Jorge Cham