JOURNAL OF INQUIRY-BASED LEARNING IN MATHEMATICS
No. **18**, (Mar. 2010)

An Introduction to Probability and Statistics

Lisa Orloff Clark

Susquehanna University

Contents

To the Student

The style of learning in this class will probably be quite a bit different from what you are used to. Most of your time in class will be spent working through the problems presented in these notes.

This is a math class; thus, you will be doing math. One thing you will not be doing is watching the instructor do math while you take notes. To ride a bike, it is best to get on the bike and attempt to ride; not watch someone else ride over and over again. In time you will see that math is not a set of rules to memorize or formulas to apply. It is not watching others do examples over and over until you can mimic what they have done. That may be what some of you expect, even enjoy, but that is not math.

To truly do math, one must think analytically. This involves considering things that are known, like definitions, and axioms, and using these known things to solve problems and construct arguments. Everything in math comes from somewhere and you too can experience the thrill of discovery!

The skills you will gain here are ones that will be useful to you for the rest of your life. This is true for math majors and non-math majors alike! Analytical thinking is something that transcends the math classroom. Statistics in particular gives you the tools you need to think rationally and make decisions.

You can do this!

Acknowledgement

Thank you to the anonymous referee for the numerous corrections and helpful suggestions.

Chapter 1

Numerical Summaries of Sample Distributions

1.1 The "Average" of a Distribution

Suppose you have a *sample distribution,* that is, a list of n numbers $\{x_1, x_2, ..., x_n\}$, listed from smallest to largest.

- A **mode** of the distribution is the number that occurs most often. The mode is not unique. A distribution can have more than one mode.
- The **median** is the "middle" of the distribution. If the number of observations n is odd then the median is the center observation in the ordered list. If the number of observations n is even then the median is the arithmetic mean (see below) of the two center observations.
- The **arithmetic mean** $\bar{x}$ is what we will often times refer to simply as the **mean**. To find the arithmetic mean, you add all of the observations and divide the sum by the total number of observations. That is $\bar{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$.

Example 1. *Suppose you have taken a poll that asked all of the faculty members of the math department how many times they brush their teeth a day. The responses are as follows:*

$$0, 1, 2, 2, 2, 2, 3, 7$$

- *The mode of this distribution is* 2.
- *The median of this distribution is* 2.
- *The arithmetic mean of this distribution is* $\dfrac{0+1+2+2+2+3+7}{7} =$ 2.43.

1.2 Exercises

Answer each of the questions below and **fully justify your answer** using complete sentences. If you answer “yes”, explain why. If you answer “no”, it may be appropriate to show a counterexample, that is, an example of a specific distribution where some property does not hold.

IMPORTANT: Using one example is not enough to show something is ALWAYS true; however, using one example can be enough to show something is NOT ALWAYS true.

1. Give an example of a distribution that has more than one mode.
2. What is the maximum number of modes a particular sample distribution can have?

   For numbers 3–7 below, suppose you are given a sample distribution $x_1, x_2, \dots, x_n$.
3. Will the mode *always* take a value that is equal to x_i for some i?
4. Will the median *always* take a value that is equal to x_i for some i?
5. Will the arithmetic mean *always* take a value that is equal to x_i for some i?
6. Will the median *always* take a value that is equal to the mode?
7. Will the median *always* take a value that is equal to the arithmetic mean?
8. Suppose 6 students take an exam and the mean score is 80%. Five of the students scores are: 95,78,85,56, 96. What is the sixth student’s score?
9. The number 15 is added to each of the biggest 150 numbers in a distribution of 301 numbers. (In other words, $n = 301$).
   (a) How does this addition affect the median of the distribution? [1, 4.2.7]
   (b) How does this addition affect the mode of the distribution?
   (c) How does this addition affect the arithmetic mean of the distribution?
10. Suppose there are n students in a class ($n \geq 4$). Some of the students appealed their scores on a certain test. The papers were reviewed and the scores of four students were raised from a 70 to a 90. How does this change affect the mean score of the class? [1, 4.2.8]

11. Suppose you have two 11th grade American History classes at a particular high school. One is taught by Mr. Knight and one is taught by Ms. Princess. Mr. Knight's class has 30 students in it and Ms. Princess's has 26 students. A common exam was given to both classes. Mr. Knight's class had a median score of 80 and Ms. Princess's class had a median score of 70. What can you say about the median score of the classes when combined? [1, 4.2.9]

12. Let a sequence of numbers have the following characteristic: starting from the third place, each value is the mean of the values that precede it.

    (a) Suppose that the sequence starts out $\{5,7,...\}$. Write out several more terms of this sequence. In other words, the first value is 5 and the second value is 7. What will the third value be? What about the fourth? The fifth? The sixth? Etc. [1, 1.1.2]

    (b) Now suppose that the sequence starts out $\{2,10,...\}$. Write out several more terms of this sequence.

    (c) Now suppose that the sequence starts out $\{a,b,...\}$. Write out several more terms of this sequence, simplify your answer as much as possible.

13. You decide you would like to know the mean length of the worms in your earthworm collection. You put all of the worms on a plate, measure and record the length of each, and return them to the box in which you keep them. You then compute the mean length of your collection to be 8 inches. As you are adding some nice moist compost to the box before putting it back into the garage, you notice one worm stuck to the lid. This little sweety had managed to elude being put on the plate and was not included in your measurements. You measure this worm and its length is 8 inches. How will this little worm's length affect the overall mean length of your collection?

14. Some years ago a well-known public official left California and moved to Alabama. A local California reporter revealed both his regional chauvinism and his feelings about the official when he remarked that "on this occasion he raised the mean IQ in both states".

    (a) Explain how this is statistically possible.

    (b) Considering this, if the mean IQ of the US is the average of the means of each state, it appears that by a mere reshuffling of populations between states, one could increase the mean IQ of the US! Is this so? Explain. [1, 1.1.16]

1.3 The Five-Number Summary of a Distribution

Suppose you have the distribution of a quantitative variable listed in order from smallest to largest with median *Me*.

- The **first quartile** Q_1 is the median of the observations smaller than *Me*.
- The **third quartile** Q_3 is the median of the observations larger than *Me*.
- The **five-number summary** of a distribution is a list containing: the smallest observation, Q_1, *Me*, Q_3, and the largest observation.

Example 2. *Suppose you are teaching a class of 11 third graders and have given them a quiz. Their scores out of 100 are as follows:*

$$55, 70, 70, 72, 85, 88, 90, 90, 92, 95, 98$$

Here $Me = 88, Q_1 = 70,$ *and* $Q_3 = 92.$

The five-number summary is: 55, 70, 88, 92, 98.

1.4 Exercises

1. The five-number summary of a sample distribution of exam scores (in percentages) for 100 students is: $0, 0, 92, 94, 100$. Discuss the performance of the 100 students overall. Give as much detail as you can. Be sure to include in your discussion what you still do not know about the students' scores.
2. The 100 students from the previous question retake the exact same exam and the five-number summary of the new scores is: $55, 75, 92, 96, 100$. Discuss the performance of the students overall on this exam.
3. Billy Jo is one of the 100 students who took the exams described above. On the first exam, he got a 26% and on the second he got a 56%. Discuss Billy Jo's performance on each exam based on how his score compares to the rest of the class. If you were responsible for assigning grades to the students, what grade would you give to Billy Jo on each exam respectively?
4. Write down a sample distribution with $n = 20$ so that the five-number summary of the distribution is: $2, 4, 6, 8, 10$.

1.5 Variance and Standard Deviation

A **population** is a group of individuals or subjects that we want to learn something about. When we talk about a sample distribution, we typically consider the list of n numbers $\{x_1, x_2, \ldots, x_n\}$ as representing *some of* the individuals or subjects from a population.

- The **variance** s^2 of a sample distribution is the mean of the square of the distance each observation is away from the mean $\overline{x}$.

  To compute the variance:

  1. Compute the mean $\overline{x}$ of the distribution.
  2. Compute the distance each observation is away from $\overline{x}$.
  3. Square each of the distances found in 2.
  4. Find the mean of the list found in 3.

  The formal equation for this is:

$$\begin{aligned} s^2 &= \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+(x_3-\overline{x})^2+\ldots+(x_n-\overline{x})^2}{n} \\ &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n} \end{aligned} \tag{1.1}$$

- The **standard deviation** s is the square root of the variance. In other words, $s=\sqrt{s^2}$.

Important Note: We are using the formulae above for variance and standard deviation. They are the “true” formulas. However, the terminology of *variance* and *standard deviation* are also used for what we will call **unbiased** estimator for a population’s variance and standard deviation. In this case, the denominator in the formula (1.1) above is $n-1$ rather than n. Be careful. For now, we will use the formula given above but eventually, we will switch to the unbiased version.

Beware: When using software (like Exel, Minitab, or a calculator) to compute variance and standard deviation, they will give you the unbiased version.

Example 3. *Suppose you chart the number of green M&M’s in each of 4 bags of M&M’s. Your results are as follows:*

$$2, 6, 12, 13$$

- *To compute the variance:*

  1. *The mean* $\overline{x}=8.25$.

2. *The distance each observation is away from the mean:*

$$\begin{aligned} 2 - 8.25 &= -6.25 \\ 6 - 8.25 &= -2.25 \\ 12 - 8.25 &= 3.75 \\ 13 - 8.25 &= 4.75 \end{aligned}$$

3. *Squaring each of the distances found in 2:*

$$\begin{aligned} (-6.25)^2 &= 39.0625 \\ (-2.25)^2 &= 5.0625 \\ (3.75)^2 &= 14.0625 \\ (4.75)^2 &= 22.5625 \end{aligned}$$

4. *Finding the mean of the numbers found in 3, we get*

$$s^2 = \frac{39.0625 + 5.0625 + 14.0625 + 22.5625}{4} = 20.1875.$$

- *The standard deviation* $s = \sqrt{20.1875} = 4.493$.

1.6 Notation

We have been using $\bar{x}$ and s to denote the mean and standard deviation of a sample distribution from a population. If we are talking about the mean and standard deviation of the entire population, we use μ and σ for the mean and standard deviation. (Greek letters mu and sigma.) At this point we will not worry too much about the distinction but it will be important later.

1.7 Exercises

1. Suppose you have a sample distribution 2,2, 6, 8, 10. Compute the mean, variance, and standard deviation of this distribution.

2. Students were asked to analyze a set of 50 nonnegative scores, not all of which were identical. The set included exactly three 0 (zero) scores. It also included two nonzero scores which were identical to the mean of all the scores.

   (a) One student decided not to include the three zero scores in his analysis, on the false assumption that zero is not a number. He correctly calculated the following measures, based upon his altered data set. For each measure, would his calculation increase, decrease, or not change the original measure, or is it impossible to tell? Explain all of your answers.

i. mean
ii. median
iii. mode
iv. range (largest value minus the smallest value)
v. variance

(b) Another student decided not to include the two scores that were identical to the mean, arguing that most measures are based on deviations from the mean, whereas those two scores did not deviate from the mean. How would the measures (i – v above) that this student obtained change in relation to the correct answers of the complete set of scores? Answer by 'increase', 'decrease', 'no change', or 'impossible to know'. Explain your answers. [1, 1.1.3]

3. Write two numbers with mean 8 and variance 4.

4. (Extra credit) Is it possible to write two numbers different from those found in number 3 with mean 8 and variance 4?

5. Write three numbers with mean 5 and variance $\frac{8}{3}$.

6. (Extra credit) Is it possible to find three numbers different from those found in number 5 with mean 5 and variance $\frac{8}{3}$.

7. Suppose you have two numbers with variance 9. What is the range of this set? (Note: the range is the largest value minus the smallest value.)[1, 1.1.4]

8. In an educational research project it is necessary to construct a control group that shares some features with the experimental group. There are 12 children in the experimental group, and the investigator decides that the control group should be the same size.

   The mean and the variance of the variable *x* in the experimental group are 6.0 and 14.00, respectively. The investigator wishes to construct the control group so that both groups will have the same mean and variance.

   After ten children have been selected for the control group, the mean of their *x*-values is 5.8 and the variance is 16.36. Two additional children will be selected for the control group.

   What should the *x*-values of these children be so that the experimental group and the control group have equal means and variances? [1, 1.1.11]

9. Eight people took a test in which one can score only 1, 2, or 3.
   (a) You know that exactly two people scored 1 and that the distribution is *symmetric* about the mean. What is that variance of the set of scores?
   (b) Let the variance of the set be 1. List the eight scores.
   (c) Given that the mean of the scores is 3, what is the standard deviation of the set of scores?[1, 1.1.13]

10. The treasury department is considering several schemes for revising its salary and employment policies for government workers.

    The following three schemes are suggested. Determine, in each case, how the suggested revision would affect each of the following measures:

    (a) Each employee will get a raise of $125 per month.
    (b) The salaries will be increased by 15% across the board.
    (c) The number of employees at each salary level will be decreased to 90% of their original number.
        i. The mean monthly salary in dollars.
        ii. The variance of the monthly salaries.
        iii. The standard deviation of the monthly salaries.
        iv. The median monthly salary.
        v. The modal monthly salary.[1, 1.1.15]

11. The mean salary in a certain plant was $1500, and the standard deviation was $400. A year later each employee got a $100 raise. After another year each employee's salary (including the above mentioned raise) was increased by 20%. What are the mean and standard deviation of the current salaries in dollars? [1, 4.2.17]

12. For those of you who have been ignoring the sigma notation, that is the "Σ's" above, no longer!

    (a) Suppose $x_1 = 1, x_2 = 2, x_3 = 3, ..., x_{10} = 10$. In other words, $x_i = i$ where i takes all integer values from 1 to 10.
        Compute each of the following:
        i. $\sum_{i=1}^{10} i$
        ii. $\sum_{i=1}^{5} i^2$
    (b) Suppose $x_1 = 2, x_2 = 4$ and $x_3 = 7$. Compute $\sum_{i=1}^{3}(x_i - 2)$.
    (c) Suppose $x_i = 2^i$. Compute $\sum_{i=1}^{6}(x_i)^2$.
    (d) Suppose $x_1 = 100, x_2 = 95, x_3 = 82, x_4 = 60$. Compute $\sum_{i=1}^{4}(x_i - 60)^2$.

Chapter 2

Standardization

2.1 Standardization

- A **standard distribution** is distribution with mean 0 and standard deviation 1. Notice that the variance of a standard distribution is also 1.

- Suppose you have a sample distribution $\{x_1, x_2, ..., x_n\}$.

  For each x_i, we can compute the **standard score** of x_i also called the **z-score** of x_i by computing

  $$z_i = \frac{x_i - \mu}{\sigma}$$

  This tells us how many standard deviations x_i is away from the mean. The distribution $\{z_1, z_2, ...\}$ is a standard distribution.

Example 4. *Suppose you give a test to a history class. After you have graded them, you compute the mean and standard deviation of the distribution of grades to be $\mu = 80$ and $\sigma = 12$. The standard score for a student who gets a 72 on the exam is:*

$$\frac{72-80}{12} = -.667.$$

2.2 Exercises

1. In example 3 on page 9, notice that $x_1 = 2, x_2 = 6, x_3 = 12, x_4 = 13$. Compute the z-scores of each of these values. After you have done this, compute the mean and standard deviation of the standardized distribution $\{z_1, z_2, z_3, z_4\}$.

2. (Extra Credit) Suppose you have a sample distribution $\{x_1, x_2, \ldots, x_n\}$ and you compute the z-scores of each x_i to get a new distribution $\{z_1, z_2, \ldots, z_n\}$. Verify that the mean and standard deviation of this new distribution is indeed 0 and 1 respectively.

3. The list of test results for students in a certain course was published in standard scores. Jill saw the number 1.5 next to her name. She decided to compute this number's standard score in relation to the list of numbers published. What was the result of her computation? [1, 4.3.1]

4. Suppose you have a sample distribution $\{x_1, x_2, \ldots, x_n\}$ and you compute the z-scores of each x_i to get a the standardized distribution $\{z_1, z_2, \ldots, z_n\}$. Determine whether each of the following claims about the set of standard scores is true or false. Justify each of your answers.

   (a) The sum of the positive scores is equal to the sum of the absolute values of the negative scores. (Hint: Consider the mean formula.)

   (b) The number of positive scores is equal to the number of negative scores.

   (c) The sum of the squared standard scores is n. (Hint: Consider the variance formula.)[1, 4.3.2]

5. All the math scores of 26 children in one class were converted to standard scores, $\{z_1, z_2, \ldots, z_{26}\}$. Determine whether each of the following is "impossible", "must be true", or "can be true". Justify your answers.

   (a) $\sum_{i=1}^{26} z_i^2 = 30$.

   (b) The largest standard score is 3.3.

   (c) $\sum_{i=1}^{26} z_i = 0$.

   (d) 15 of the standard scores are negative.

   (e) The median standard score is -0.3.[1, 4.3.3]

6. In a certain school it was decided to transform the scores of students in each class to standard scores. John is in a class of 36 students. He got a standard score of 2 in English. What conclusion can be drawn about this data. [1, 4.3.4]

   Note: You must be as precise you can. Is John the best in the class? Did he get the median grade? Is he in the top ten? Top three? Bottom 2? Etc.

Chapter 3

Correlation and Regression

3.1 Motivating Exercises

1. Using a tape measure, measure the height of each person in your group. Record your data in inches.
2. Measure the arm length of each person in your group, measuring from the breast bone to the tip of the middle finger. Record your data in inches.
3. Considering the measurement of each person as an ordered pair: (height, arm length), record the ordered pairs of the entire class on the board.
4. Plot the data of the class on the $x-y$ plane with height on the x-axis and arm length on the y-axis.
5. Does the data form a straight line? (All you need to say here is either: yes, no, or almost).
6. Based on this data, what can you say about the relationship between a person's height and their arm length?
7. Use the data to predict the arm length of the tallest human being ever to have lived, Robert Waldo. Waldo was 8 feet 11 inches tall.[1]

3.2 Review: Lines

One way to describe a line in the cartesian plane is as follows: a line is the set of ordered pairs, (x,y) where x and y satisfy the equation:

$$y = mx + b$$

[1] Thank you to Jeremy Alm of Illinois College for suggesting this kind of exercise.

where m represents the slope and b is the y-intercept of the line. Thus, if you know the slope of a particular line and you know b, that is, the y-value when $x = 0$, you can write down an equation for that line.

How do you find the slope of a line? Here is the key feature of lines that make the equation above work: the **slope of a line is constant**. In other words, you can find the slope by choosing *any two points* on the line and then taking the change in the y coordinate divided by the change in the x-coordinate. That is,

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

for any two points (x_1,y_1) and (x_2,y_2) on the line.

3.3 Exercises

1. (a) What is the slope of the line between the points $(0,5)$ and $(2,3)$?

   (b) Find an equation for the line given in part (a) and sketch the graph of this line.

   For 2–7 below, justify your answer using the slope formula.

2. Do the points $(0,0)$, $(-1,2)$, and $(5,-10)$ lie on one straight line?

3. Do the points $(0,0)$, $(-1,-2)$, and $(5,10)$ lie on one straight line?

4. Do the points $(0,0)$, $(-1,-2)$, $(5,10)$, and $(3,7)$ lie on one straight line?

5. Do the points $(1,0)$, $(-1,4)$, $(0,1)$ and $(5,10)$ lie on one straight line?

6. Do the points $(2,1)$, $(4,2)$, $(1,2)$, and $(-2,-4)$ lie on one straight line?

7. Do the points $(2,0)$, $(4,-2)$, $(1,1)$, and $(-2,4)$ lie on one straight line?

8. (a) For each of the preceding four exercises, sketch a graph of the points given.

   (b) How close to “linear” are each of the relationships of questions 4–7? Assess this by ranking them from “most linear” to “least linear”.

3.4 Linear Correlation

The **linear correlation coefficient** r_{xy} (also called Pearson's correlation coefficient) measures the linear relationship between two quantitative variables x and y. The value of r_{xy} is always between -1 and 1 inclusive. The closer r_{xy} is to the extreme values, the stronger the linear relationship. If $r_{xy} = 1$ or $r_{xy} = -1$ then the relationship is a **perfectly linear relationship** with either positive or negative slope respectively.

Suppose you have data on two quantitative variables x and y from n individuals:

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

The mean and standard deviation of the x values are $\overline{x}$ and s_x. The mean and standard deviation of the y values is $\overline{y}$ and s_y. The correlation coefficient for x and y is defined by the formula

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

Example 5. *You have three lab rats. Suppose you measure the length of each rat's head in inches and the weight of each rat in pounds. You get the following measurements:*

$$\{(2 \textit{ inches }, 5 \textit{ pounds }), (3 \textit{ inches }, 6 \textit{ pounds }), (1 \textit{ inch }, 4 \textit{ pounds })\}$$

You can see that $\overline{x} = 2$ *and* $\overline{y} = 5$. *Also,* $s_x = \sqrt{2/3}$ *and* $s_y = \sqrt{2/3}$.

Therefore

$$\begin{aligned} r_{xy} &= \frac{1}{3}\left(\left(\frac{2-2}{\sqrt{2/3}}\right)\left(\frac{5-5}{\sqrt{2/3}}\right) + \left(\frac{3-2}{\sqrt{2/3}}\right)\left(\frac{6-5}{\sqrt{2/3}}\right) + \left(\frac{1-2}{\sqrt{2/3}}\right)\left(\frac{4-5}{\sqrt{2/3}}\right)\right) \\ &= \frac{1}{3}\left(0 + \frac{1}{2/3} + \frac{1}{2/3}\right) \\ &= \frac{1}{3}\left(\frac{2}{2/3}\right) = 1. \end{aligned}$$

Now suppose on the same rats, you count how many whiskers each rat has on its face and get the following measurements. Note, the rats are listed in the same order as in the previous example.

$$\{3, 5, 10\}$$

Here, if you let x *be the length of the rat's head as before and* y *be the number of whiskers, you get* $r_{xy} = -.554$.

3.5 Exercises

1. Find the missing numbers (x and/or y values) which satisfy the requirements listed next to each table. No elaborate computations of correlations are needed to complete any of the solutions.[1, 1.2.1]

   (a)

x	y
1	7
2	12
6	32
10	
	67

   $r_{xy} = 1$

   (b)

x	y
1	
2	
3	
4	
5	

   $r_{xy} = -1$

   (c)

x	y

   $\overline{x} = \overline{y}$, $\sigma_x = \sigma_y \neq 0$, $r_{xy} < 0$

2. Use software to compute each r_{xy} for exercises 4–7 in section 3.2 and see whether your ranking of the strength of the linear relationships was correct.

3. The table below gives data concerning the number of registered boats in Florida and the number of recorded manatee deaths in Florida for the years 1977–1990. [3] Compute r_{xy} for this data and discuss your findings.

Year	Motorboats registered	Manatee Deaths
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47

3.6 Rank-order Correlation

The (Spearman) **Rank-order correlation coefficient** r_s provides a measure of correlation between ranks. Again, $-1 \leq r_s \leq 1$. It is important to note that the r_s tells us the strength of a *monotonic* relationship between two variables, and does not tell whether this relationship is linear.

The Spearman Rank-order correlation coefficient is defined by

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(d^2)}{n(n^2-1)}$$

where d is the difference in rank of corresponding variables.

Because it uses ranks, the Spearman rank correlation coefficient is much easier to compute than Pearson's correlation coefficient.

Example 6.

Suppose 2 art critics rank ten paintings from 1 – 10. We want to determine whether the two critics ranks are related. If the two critics rankings are exactly the same, we say there is a ***perfect positive correlation****. In this case the differences between the ranks d will all be 0 and* $r_s = 1$.

If their rankings are exact opposites then this is ***perfect negative correlation****. In this case it turns out that* $r_s = -1$.

3.7 Exercises

1. The dexterity of five people's left and right hands was measured, and the following scores were obtained. Answer the questions below without making any elaborate computations of correlation coefficients.

Subject	Left hand score	Right hand score
A	4	6
B	1	3
C	5	7
D	3	5
E	10	12

   (a) What is the linear correlation coefficient between scores of the right hand and those for the left hand?

   (b) What is the rank-order correlation coefficient between scores of the right hand and those for the left hand?

   (c) After a while, another subject, F, was tested and given the following scores: left hand 9, right hand 8. Will this additional datum affect the linear correlation coefficient or the rank-order correlation coefficient? If so, how?[1, 1.2.2]

2. Compute the Spearman rank order correlation coefficient for the boater registration/ manatee death data from the previous section and interpret your result.

   To do this, you must find the "ranks" for each year of both the number of motorboats registered and also the number of manatee deaths. Number each from 1 to 14 in order of abundance. The *d* in the formula is the difference in ranks for each year.

   For example, in 1977, the number of motorboats registered is the very smallest making that ranking 14. The number of manatee deaths is also at a minimum making that ranking 14 as well. Thus $d = 14 - 14 = 0$ for 1977. Continue this for each year.

3.8 Regression Line

- A **regression line** is a straight line that describes how the variable y changes in response to the variable x. We often use a regression line to predict the value of y for a given value of x. Sometimes, this is known as a **best-fit line**.
- One such regression line is called the **least-squares regression line**. Given data for x and y, we can compute an equation of a line in the form $\hat{y} = mx + b$ where
$$m = (r_{x,y})\frac{s_x}{s_y}$$
and
$$b = \overline{y} - m\overline{x}.$$
This particular regression line minimizes the vertical distances of the observed points from the line.
- Plugging in an observed x value gives us a predicted $\hat{y}$ value. If $r_{x,y} = 1$ then the least-squares regression line is 100% accurate in predicting the $\hat{y}$ values for given x values. In other words, $\hat{y} = y$ (the actual y value). In general, the square of the linear correlation coefficient, $r^2_{x,y}$ **tells us the percent of accuracy** of the least-squares regression line.

3.9 Exercises

1. Compute the least-squares regression line for the motorboat-manatee death data.
2. Suppose you are told that there were 821 motorboats registered in 2007. Use the line found in part a to predict how many manatee deaths occurred that year.
3. How accurate is the prediction made in part b?

Chapter 4

Probability

4.1 Set Theory

Here are some terminology and notation:

- $\emptyset$ denotes the empty set or the set that contains no elements. We also write this $\{\}$.
- $A, B,$ and C denote sets.
- Ω denotes the universe of all possible elements in consideration.
- $\overline{A}$ denotes the set consisting of elements that are in Ω and not in the set A. We call this A *complement*.
- $A \cup B$ is the set consisting of all elements in the set A combined with all the elements in set B. We call this A *union* B.
- $A \cap B$ is the set that contains only the elements that are in both A and B. We call this A *intersect* B.
- We denote $A \subseteq B$ to say that "A is a subset of B". This means that every element of A is also an element of set B.
- We write $A - B$ to mean the set containing elements that are in A and not in B. Notice that $\overline{A} = \Omega - A$.
- We say that two sets are **disjoint** if they have no elements in common. In other words, A and B are disjoint if and only if $A \cap B = \emptyset$.

4.2 Exercises

1. Consider the sets $A = \{1,2,3,4,5\}$ and $B = \{2,4,6,8,10\}$ where $\Omega = \{1,2,3,4,5,6,7,8,9,10\}$. Compute each of the following sets:

   (a) $A \cup B$

   (b) $A \cap B$

   (c) $\overline{A}$

   (d) $\overline{B}$

   (e) $B - A$

   (f) $A - B$

   (g) $\overline{A \cup B}$

   (h) $\overline{A} \cup \overline{B}$

   (i) $(B - \overline{A}) \cap \overline{(A \cap B)}$

2. Suppose $\Omega = \{\text{red, orange, yellow, green, blue, indigo, violet}\}, A = \{\text{red}\}$ and $B = \{\text{red, orange, blue}\}$. Compute a–i from question 1 for this example.

3. Write down an example of a specific Ω, A, B, and C so that $A \subseteq B$ and C is disjoint from both A and B.

4. Write down an example of a specific Ω, A, B, and C so that A, B, and C are all disjoint and $A \cup B \cup C = \Omega$.

5. (Extra Credit) Consider the sets $A = \emptyset$ and $B = \{\emptyset\}$ and $C = \{1, \emptyset\}$. Find each of the following sets, if it is possible. If it isn't, state why.

   (a) $C \cup B$

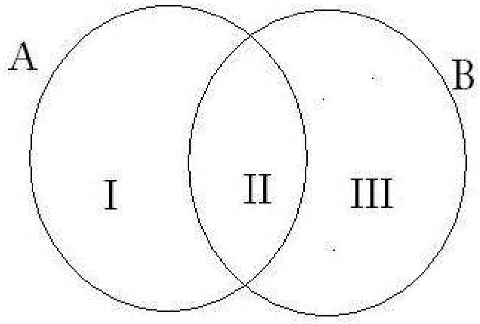
   (b) $A \cap B$

   (c) $\overline{A}$

   (d) $B - A$

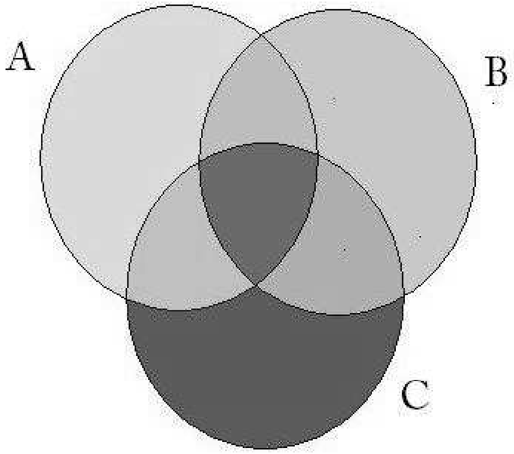
   (e) $C - B$

Venn Diagrams



Set A corresponds to regions I and II.
Set B corresponds to regions II and III.
Set $A \cap B$ corresponds to region II.
Set B-A corresponds to region III.

Set $A \cup B$ corresponds to regions I, II, and III.



Set A is yellow which includes the regions that are yellow, green, brown and orange.

The set $A \cap B \cap C$ corresponds to the brown region.

The set $A \cap B$ corresponds to the brown and orange regions.

4.3 Venn Diagrams

A **Venn Diagram** is useful in illustrating sets and their relationships to each other. At the top of the page is an example of a Venn diagram with two sets. Below that is an example of a Venn diagram with three sets.

4.4 Exercises

1. Determine whether each of the following is true or false. If you say true, show that the Venn diagram of the left-hand side is the same as the Venn diagram of the right hand side. If you say false, come up with specific sets where the equality does not hold.[1, 2.1.1]

   (a) $\overline{A \cup B} = \overline{A} \cup \overline{B}$

   (b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

   (c) $\overline{A - B} = \overline{A} \cup B$

   (d) $A - \overline{A} = \emptyset$

(e) $\overline{A \cap \overline{A}} = \Omega$

(f) $A = (A \cap B) \cup (A \cap \overline{B})$

2. Let Ω be the set of all students currently enrolled in classes at Susquehanna University. Let A be the set of all students enrolled in intro stats this term, let B be the set of all students who play a varsity sport.

   Suppose there are 2,000 total students enrolled at SU and 120 are enrolled in a section of intro stats this term and 230 play a varsity sport. Note: These numbers are not the official counts.

   Interpret each of the sets below in terms of this example and, if possible, determine how many people are in each set.

   (a) A

   (b) B

   (c) Ω

   (d) $A \cap B$

   (e) $A \cup B$

   (f) $\overline{A}$

   (g) $\overline{B}$

   (h) $\overline{A \cup B}$

   (i) $\overline{A \cap B}$

3. Repeat number 2 with the added information that there are 102 students enrolled in Introductory Statistics this term who do not play a varsity sport.

4.5 A Probability Exercise

Note to instructor: You will need to provide each student in the class with a bag of M&M's or require each of them to bring their own bag to class. You will also need to provide each student with a sandwich bag.

1. You have a bag of M&M's in front of you...if you don't, get one. Open your bag and put the M&M's into a sandwich bag. Number a piece of paper from 1–120. Draw one M&M from the sandwich bag (without peeking) and record the color of the M&M as number 1. **Put the M&M back into the bag**; shake the bag a bit; and then repeat 120 times, making sure to record your results after each draw. **DO NOT EAT** any of the M&M's until you have completed all 120 draws and (a)–(h) below.

   (a) Look at the first 10 draws. Of those drawn, what percentage was red?

   (b) Look at the first 25 draws. Of those drawn, what percentage was red?

   (c) Look at the first 50 draws. Of those drawn, what percentage was red?

   (d) Look at the first 80 draws. Of those drawn, what percentage was red?

   (e) Now consider all 120 draws. Of those drawn, what percentage was red?

   (f) Based on the previous questions, what do you estimate is the actual proportion of red M&M's in your bag of M&M's?

   (g) Repeat (a)–(f) for each of the other colors of M&M's.

   (h) Now count the M&M's in your bag, record the number of each color and see if you were right.

   (i) According to M&M's official web site (in February of 2008), the proportion of M&M's in every bag is as follows:

   - 13% of the M&M's are red.
   - 13% are brown.
   - 14% are yellow.
   - 24% are blue.
   - 20% are orange.
   - 16% are green.

   Did this hold true for your bag and the bags of your group members?

   (j) When choosing an M&M at random from your full bag of M&M's, what is the probability of drawing a red M&M?

4.6 The Vocabulary and Axioms of Probability

We will be using set theory to help us study and learn probability. Here is some of the notation and vocabulary of probability theory:

- An **experiment** is any process that allows researchers to obtain observations (outcomes).
- An **event** is a set that represents a collection of results or outcomes from an experiment. We will continue to denote these with capital letters: A, B, C.
- A **sample space** for an experiment is a set that consists of all possible outcomes. Again, this will be written Ω.
- P denotes probability.
- $P(A)$ denotes the probability of event A occurring.

Example 7. *In the exercise in section 4.5, the experiment was drawing an M&M at random from a full bag of M&M's and recording its color. One of the events we considered was the outcome of drawing a red M&M. The sample space in this case was the set of all M&M colors.*

Example 8. *Suppose that we toss a coin twice. The sample space for this experiment is $\Omega = \{HH, HT, TH, TT\}$. Let A be the event of tossing a head on the first toss. Thus, $A = \{HH, HT\}$. Suppose we let B be the event of tossing at least one head. Then $B = \{HH, HT, TH\}$ and we see that $A \subseteq B$.*

Example 9. *Suppose we select a student at random from a class of 30 adult students. Here Ω is the 30 students. Let A be the event that the student selected is a man. Let B be the event that the student selected is a woman. With the assumption that everyone in the class identifies themselves as either male or female, $A \cup B = \Omega$ and $A \cap B = \emptyset$.*

Probability Axioms

A1: $P(A) \geq 0$ for every event A.

A2: $P(\Omega) = 1$.

A3: If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$.

4.7 Exercises

1. Consider the M&M experiment described in section 4.5. Let A be the event that you draw a red M&M. Let B be the event that you draw a green M&M. Based on the data you found and the axioms of probability, what is $P(A \cup B)$ and $P(A \cap B)$?

2. Let Ω be the set of all students at a given university.

   Several subsets of Ω are represented below. Some of the sets are represented by symbols and some by the probability that a random student will belong to the set.

   Use the given symbols and probabilities to complete the missing items in the table. [1, 2.1.2]

Definition of set in words	Symbol	Probability
All social-science students	S	
All research students	R	
All social-science and/or research students		0.25
All research students who are not from social sciences		0.05
All social-science students who are not research students		0.10
All social-science research students		

3. Here is a rough description of an experiment carried out by a psychologist:

   Subjects were given the following judgment problem: “Think of a population of women with academic degrees in the social sciences. Consider a random woman from such a population. Rank order the following categories according to the probability that the woman will belong to them.”

   (a) Employed at a university.

   (b) Married and unemployed.

   (c) Owns her own business.

   (d) Unemployed.

   The subjects’ rankings, from the most probable possibility to the least probable one, matched the order in which the categories are written, that it, (a) was rated most probable and (d) least probable.

   Are these rankings compatible with the axioms of probability theory, or do they violate them in any way? Explain. (Note that you are not being asked about the state of employment of academic women. Rather, you are asked to evaluate the consistency of the judgments of the experimental subjects.)[1, 2.1.3]

4. Tom and Harriet cannot agree on whether to go to the baseball game (Tom's choice) or the movies (Harriet's choice). Flipping a coin and deciding for Harriet if 'heads' and for Tom if 'tails' appears too trivial. They discuss several chance procedures for making their decision.

   Let T be the event that 'Tom wins' and H the event that 'Harriet wins'. Find $P(T)$ and $P(H)$ for each of the suggestions described below, and determine whether each procedure is fair or biased.

   (a) Playing one round of "Rock, Paper, Scissors" to determine whose wish will be granted.

   (b) Flipping two coins. Harriet prevails if at least one outcome is heads, Tom– otherwise.

   (c) Giving each a box containing three notes numbered 1,2, and 3, and having each blindly draw one of the notes. If the sum of their draws is even– Tom wins, if it is odd – Harriet wins.

   (d) Rolling two dice and computing the absolute difference between the two numbers obtained (always a positive number). Tom wins if the outcome is either 1 or 2, Harriet–otherwise. [1, 2.1.4]

5. Suppose A and B are two *disjoint* events and that $P(A) = .22; P(B) = .33$. Calculate the following probabilities.

   (a) $P(A \cup B)$

   (b) $P(A \cap B)$

   (c) $P(\overline{A} \cup B)$

   (d) $P(\overline{A} \cap B)$

   (e) $P(\overline{A \cap B})$

   (f) $P(\overline{A} \cap \overline{B})$

   (g) $P(A - B)$. [1, 2.1.5]

6. Let A and B be events so that $A \subseteq B$. Also, let $P(A) = .30$, and $P(B) = .45$. Calculate the following probabilities.

   (a) $P(A \cup B)$

   (b) $P(\overline{A})$

   (c) $P(A \cap B)$

   (d) $P(\overline{A} \cap B)$

   (e) $P(A - B)$

   (f) $P(\overline{A \cup B})$

   (g) $P(\overline{A} - B)$

   (h) $P(\overline{A} - \overline{B})$.[1, 2.1.6]

7. Suppose A and B are events so that $P(A) = .40$ and $P(B) = .25$.
    (a) What is the very largest that $P(A \cup B)$ can be?
    (b) What is the very smallest that $P(A \cup B)$ can be?
    (c) What additional information do you need in order to compute $P(A \cup B)$ exactly?
8. Suppose that A and B are events so that $P(A) = .75$ and $P(B) = .60$. Answer questions (a) – (c) above with these probabilities.
9. Suppose that A and B are events so that $P(A) = 1$ and $P(B) = .35$. Answer questions (a) – (c) above with these probabilities.

4.8 Conditional Probability: Exercises

1. Suppose you have a class with 30 students enrolled. You know that 18 of the students are women and 12 of the students are men. In addition, 10 students wear glasses, 7 of whom are women.

   (a) How many men wear glasses?

   (b) How many women do not wear glasses?

   (c) How many students are either women or wear glasses (or both)?

   (d) Suppose you choose a student at random. What is the probability that you choose a woman?

   (e) What is the probability that the student chosen wears glasses?

   (f) What is the probability that the student chosen does not wear glasses and is not a woman?

   (g) What is the probability that the student chosen is a woman who wears glasses?

   (h) Suppose you know that you have chosen a woman. What is the probability that the chosen student wears glasses? Note: Be careful with this one. Your answer should be different than what you got in (g).

   (i) Suppose you know that the student you have chosen wears glasses. What is the probability that the chosen student also is a woman?

2. Consider the same class as in question 1. Suppose you choose a student at random from this class. Let W be the event that you choose a women, G be the event that you choose a person who wears glasses.

   (a) Use the notation of set theory and probability to describe each of the sets in (d)–(g) above. (For example, the probability that the student chosen is a man would be $P(\overline{W})$.

   (b) Use the notation of set theory and probability to write down a formula for the (h) and (i) from question 1 using the values $P(W), P(G), P(W \cap G)$ (not necessarily all three).

   Notation: $P(B|A)$ is the probability that event B will occur given that the event A has already occurred. We read "$B|A$" as "B given A."

3. Suppose that you have 10 slips of paper numbered 1-10. You put the slips of paper into a hat and draw one slip of paper at random. Let B be the event that you draw a six and let A be the event that you draw an even number. Find $P(B)$ and $P(B|A)$.

4. Find a general formula for $P(B|A)$ involving $P(A \cap B)$ and $P(A)$. Use the previous exercise to justify your answer.

4.9 Conditional Probability: Summary

As you saw in the previous exercises, the general formula for the *conditional probability* $P(B|A)$ is given by:

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

We say A and B are **independent** events if $P(B|A) = P(B)$. This means that the occurrence of A does not effect the probability of the occurrence of B. If A and B are not independent, then they are called **dependent** events.

Two useful formulas:

1. From the definition above, it is easy to see that

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

2. You should have seen in section 4.7 that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

4.10 Exercises

As always, you must FULLY justify your answer. If you are multiplying two things together, it is important to explain why. Often times, your justification will come from the formulas above.

1. You go to a casino to play a game of Texas Hold 'em. You are dealt two cards from an ordinary deck of cards.
   (a) What is the probability that the two cards are both aces?
   (b) What is the probability that the first card dealt is an ace and the second card dealt is a king of the same suit?
   (c) What is the probability that either both cards are aces or the first is an ace and the second is a king of the same suit?
   (d) What is the probability that the two cards are the same number or same face card? For example, two kings or two aces or two jacks or two 3's, etc.
   (e) What is the probability that both cards are face cards or both cards are spades?

2. Five people are standing for the first time on the edge of a cliff in Argentina "ready" to dive into the pool of water below. The five friends have been planning this trip for months but now that the big moment to jump has arrived, they are feeling pretty nervous. The cliff is *very* high. They need to come up with a method to determine who will go first. Unfortunately, standing in their bathing suits at the top of the cliff, shaking with fear, none of them have any ideas. Luckily, a nonpartisan bystander happens to be standing at the top of the cliff with them and offers a solution.

   This bystander's solution is to essentially "draw straws". The bystander is a smoker and takes out five cigarettes. Then, the bystander secretly breaks one of the cigarettes in half and then holds all five in his hand, displaying them to our fearful divers so that they are unable to tell which cigarette is the broken one.

   The five friends are then instructed to order themselves by height, shortest to tallest. Then, the shortest person is to select one of the cigarettes. If the cigarette selected is the broken one, our short friend will dive first. If not, the second shortest person draws. This process continues until one of the friends gets the broken cigarette.

   Is this procedure fair? (You need to figure out the probability that each friend will dive first.)

3. A population is distributed according to the four standard blood types as follows:

   A– 42%
   O – 33%
   B – 18%
   AB – 7%

   Assuming that people choose their mates independent of blood type, calculate the probability that a randomly sampled couple from this population will have the same blood type. [1, 2.3.7]

4. Assume that the weather can be described by one and only one of two states, *fair* or *rainy*, and there exists an unequivocal system for determining the state of the weather on any given day. Assume further that the probability that the weather is in a given state depends only on the preceding day's weather.

   For this problem, let's denote by $P(y|x)$ the probability of weather y on a given day, given that the previous day was x (both x and y assume the values *fair* or *rainy*).

The weather in a given region may described by the following conditional probabilities:
$P(rainy|fair) = 0.2$ $P(fair|fair) = 0.8$
$P(rainy|rainy) = 0.4$ $P(fair|rainy) = 0.6$
It rains Monday. A picnic is planned for Thursday. What is the probability that in three days there will be fair weather? [1, 2.3.13]

5. A doctor is called to see a sick child. The doctor knows (prior to the visit) that 90% of the children in that neighborhood are sick with the flu, denoted F, while 10% are sick with the measles, denoted M. Let us assume for simplicity's sake that M and F are complementary events.

   A well-known symptom of measles is a rash, denoted R. The probability of having a rash for a child sick with the measles is 0.95. However, occasionally children with the flu also develop a rash, with a conditional probability of 0.08.

   Upon examining the child, the doctor finds a rash. What is the probability that the child has the measles? [1, 2.4.1]

6. A man was arrested on suspicion of murder. Let us denote the event 'the man is guilty' by G. The investigating officer collected all the relevant information, added his impressions of the suspect, and arrived at the conclusion that the man's probability of guilt was 0.60.

   (a) As the investigation went on, it was learned (beyond any reasonable doubt) that the murderer's blood type was O. The relative frequency of blood type O in the population is 0.33 (that is the probability that a randomly selected person in the population has blood type O). The suspect's blood was tested and found to be O.

   Compute the 'posterior' probability of this suspect's guilt (from the officer's point of view) considering all the data.

   (b) Suppose both the murderer's and the suspect's blood types were found (with certainty) to be A. The relative frequency of blood type A in the population is 0.42. How would the posterior probability of guilt in that case compare with the same probability in part (a)? Would it be greater, smaller, or equal? Explain. [1, 2.4.2]

7. Suppose that we have two bags and each bag contains sixteen balls. Also, in each bag the sixteen balls are a combination of black balls and white balls. One bag contains three times as many white balls as black. The other bag contains three times as many black balls as white. Suppose we choose one of these bags at random.

(a) From this bag, we select five balls at random, replacing each ball after it has been selected. The result is that we find 4 white balls and one black. What is the probability that we were using the bag with mainly white balls?

(b) Now suppose instead we select five balls at random *without replacement*. The result is again that we find 4 white balls and one black. What is the probability that we were using the bag with mainly white balls?

4.11 Bayes' Rule: Summary

Bayes' Rule is basically the formula:

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

We derived this formula via the following two equations:

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

$$P(B) = P(B \cap A) + P(B \cap \overline{A}) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}).$$

4.12 Bayes' Rule: Another Exercise

1. Assume that people can be sorted unequivocally into two distinct sets according to their hair color: dark – denoted D, and blond – denoted B.

   A person's hair color is determined by two alleles of a gene, each transmitted at random by one parent. The allele for dark hair, denoted d, is dominant over that for blond hair, denoted b. Hence, of the three genotypes, dd, db, and bb, the first two would result phenotypically in D, only the third would be B. One can be certain that a blond person is homozygous (i.e., bb). However, upon observing a dark-haired person, one cannot know whether that person is genotypically homozygous (dd) or heterozygous (bd).

   Consider a couple with both mates dark haired and heterozygous for hair color. The genotypes and phenotypes of the potential offspring of the couple are given in the table below:

		Mother D	
		d	b
Father D	d	dd (D)	db (D)
	b	bd (D)	bb (B)

   The probability of such a couple giving birth to a dark-haired child is $3/4$ and to a blond child is $1/4$. The probability of a random D child of such parents being heterozygous is $2/3$.

   A couple is interested in knowing whether any of its future children could be blond. Both husband and wife are dark haired. Note that each of these prospective parents has two dark-haired parents and a blond brother.

Let H denote the genotypic event, or the hypothesis that the couple has the potential to produce a blond child. H is equal to the intersection of the events that husband and wife are heterozygotes (that is, each is a carrier of a recessive gene b). $\overline{H}$ is the event that the couple is incapable of producing a blond child (that is, at least one spouse is homozygous dd). If H is true, then the probability of that couple having a B child is $1/4$.

(a) Find the probability of H (before any children are born in that family). Denote this $P_0(H)$.

(b) A dark-haired baby is born to the couple. Denote this event D_1 (first-born child dark). Has the probability of H changed? Denote the probability of H after the birth of the first child by $P_1(H)$. What is the value of $P_1(H)$?

(c) How many dark haired children would the couple need to have in order to convince you that H is not true?

(d) After a few years the couple has another baby, which is blond. Denote this event B_2. What is the 'posterior' probability of H in light of this information? In line with the previous notations, what is $P_2(H)$? [1, 2.4.5]

Chapter 5

Probability Distributions

5.1 Definitions

- A **probability distribution** is an assignment of a probability to each *outcome* X of an experiment. We will typically use P to represent a probability distribution as this is essentially a special case of the P we've been using (which denotes the probability of events).

  In a probability distribution, each of the probabilities must be between 0 and 1 (inclusive) and the sum of all the probabilities must equal 1. We write these two conditions:

  1. $0 \leq P(X) \leq 1$ for all possible values of X,
  2. $\Sigma P(X) = 1$ where X assumes all possible values.

  Example 10. *Consider the following experiment: you flip a coin two times and record the number of heads that appeared. It should be clear that the set of all possible outcomes for this experiment is* $\{0\ heads, 1\ head, 2\ heads\}$.

  If the coin being flipped is a fair coin, then $P(2) = \frac{1}{4}$, $P(0) = \frac{1}{4}$, *and since there are two different ways of getting only one head (heads then tails and tails then heads)* , $P(1) = \frac{1}{2}$.

- The **expected value** of an experiment is denoted $E(X)$, and, in some sense, represents the outcome X we "expect" to see. We compute $E(X)$ with the formula $E(X) = \Sigma(X * P(X))$, summing over all possible values of X.

  It is important to note that $E(X)$ might not actually be equal to a possible outcome of the experiment (see the following example). However, if the experiment is performed over and over again and the outcomes X from each experiment are averaged, this value will get closer and closer to $E(X)$.

Example 11. *Suppose we give a test with four multiple choice questions. Each question has three choices and only one correct answer. We then record the number of correctly answered questions. This experiment has possible outcomes $X = 0, 1, 2, 3,$ and 4.*

Now suppose that a student takes the test and arbitrarily guesses the answer to each question. A probability distribution table for this experiment is given below with an additional column to help us compute the expected value.

X	$P(X)$	$X * P(X)$
0	$(2/3)^4 = 16/81$	0
1	$4(1/3)(2/3)^3 = 32/81$	$32/81$
2	$6(1/3)^2(2/3)^2 = 24/81$	$48/81$
3	$4(1/3)^3(2/3) = 8/81$	$24/81$
4	$(1/3)^4 = 1/81$	$4/81$

So the expected value is $E(X) = 0 + 32/81 + 48/81 + 24/81 + 4/81 = 108/81 \approx 1.33$. This means that if several people were to take this test by guessing the answers, we should expect the average number of correct answers for everyone to be about 1.33.

Example 12. *Expected values comes up a lot in gambling. Here is a simple example. Suppose you choose one card from a deck of 52 playing cards. You make a deal with a friend that if the card drawn is a king, he will pay you $50. If the card drawn is not a king, you will pay him $10. We will compute your expected loss on this bet.*

Event	X	$P(X)$	$X * P(X)$
You win	*$50*	*4/52*	*$3.85*
You lose	*-$10*	*48/52*	*-$9.23*

$E(X) = -\$5.38$. Thus, if you were to play this game over and over, in the long run, you should expect to lose $5.38 per game. Not good for you.

5.2 Exercises

1. Suppose you have a weighted coin that is twice as likely to land on heads as it is tails. You decide to flip this coin twice and record the number of times the coin lands on heads. Let X represent the possible outcomes of this experiment.

   (a) Find the probability distribution for this experiment. In other words, find $P(X)$ for all possible values of X.

   (b) Find the expected value of X, $E(X)$.

2. In a research study on animal behavior, mice are given a choice among four similar doors. One of them is the 'correct' door. If a mouse chooses the correct door, it is rewarded with food, and the experiment ends. If it chooses an incorrect door, the mouse is punished with a mild electric shock and then brought back to the starting point to choose again.

   Let X denote the number of trials in the experiment, that is the number of the trial on which the first correct choice occurs. Find the probability distribution of X in each of the two cases below. In other words, find $P(X)$ for every possible value of X. Note: You will need to use conditional probabilities.

   (a) On each trial, the mouse chooses, with equal probabilities, one of the doors that have not been chosen up to that moment. A door that has been tried is never chosen again (an intelligent mouse).

   (b) All doors are equally likely to be chosen on each trial (a dumb mouse). [1, 2.5.1]

3. An enthusiastic sports fan decides to express his support for his team by betting on it, and eventually making some profit out of that enjoyable activity.

   He knows from past statistics that the probability of his team winning in a given match is 0.75, and the probability of losing is 0.25 (let us ignore other possible outcomes).

   (a) He gets \$12 if his team wins. How much should he be willing to pay, if his team loses, in order to make and average profit of \$2 per bet?

   (b) He pays \$15 beforehand, and if his team wins he receives \$20. What is his expected gain from this bet? [1, 2.5.2]

4. An urn contains 7 balls: 3 red, 4 blue. Balls are randomly drawn from the urn, one after the other, *without* replacement.

Let X be the number of red balls drawn before drawing the first blue ball.

(a) Construct the probability-distribution table for this experiment.

(b) What is the expected value of X? [1, 2.5.4]

5. Consider an urn comprising of 3 red balls and 4 blue balls, as in the previous problem.

   Balls are randomly drawn from that urn, one after the other, *with* replacement.

   Let X be the number of red balls drawn before drawing the first blue ball.

   (a) Give the formula for the probability distribution of this experiment.

   (b) What is the expected value of X? [1, 2.5.5]

6. A medical clinic tests blood for a certain disease from which approximately one person in a hundred suffers. People come to the clinic in groups of 50. The director wonders whether he can increase the efficiency of the testing procedure by conducting pooled tests.

   Suppose that, instead of testing each individually, he would pool the 50 blood samples and test them all together. If the pooled test was negative, he could pronounce the whole group healthy. If not, he could then test each person's blood individually.

   What is the expected number of tests the director will have to perform if he pools the blood samples? [1, 2.5.9]

Chapter 6

Normal Distributions

6.1 An Exercise

Suppose that we give an exam to 200 randomly selected college students. The students' scores are rounded to the nearest 5. For example, 82 is rounded to 80, 93 is rounded to 95. The results are given in the table below.

1. The 200 students tested represent a random sample taken from the population of all college students. Now suppose we want to estimate the probability that any other randomly selected college student will receive a particular score. For example, we want to estimate the probability that a college student taking this exam will get an 80. Estimate the values in the third column of this table by using the scores of the students who have already taken the exam.

Score X	Cumulative frequency of X	$P(X)$
45	1	
50	4	
55	12	
60	28	
65	57	
70	95	
75	135	
80	167	
85	186	
90	195	
95	199	
100	200	

2. Verify that $\Sigma P(X) = 1$.

3. What is the probability that a randomly selected student will get at

least an 80 on the exam?

4. What is the probability that a randomly selected student will get a score less than 80?

5. Plot each of the points $(X, P(X))$ on a graph. Be as detailed as you can. The scales of your axes will be very different. Use one entire piece of paper for this graph.

6. Your graph has several points plotted. Now, we want to draw a curved line connecting these points so that each of the points of the curve will represent the values of $(X, P(X))$ for all possible values of X (not rounded). This curve is the **density curve** for this experiment. In fact, this density curve is of a very special type of distribution, called a **normal distribution**. (We will define these again later.)

7. Use *Minitab* to compute the mean $\overline{X}$ and the standard deviation s of the sample distribution. Plot $\overline{X}$ on the X-axis of your graph. Also plot $\overline{X} - s$ and $\overline{X} + s$.

8. Use *Minitab* to convert this distribution to a standard distribution and repeat 1–5 above with this distribution.

6.2 Density Curves and Normal Distributions

- Most of the examples we have seen have dealt with experiments where the number of outcomes is *discrete*. This means we've mostly considered experiments with a finite number of possible outcomes. We now want to shift our attention to experiments with an infinite number of outcomes, a "continuum" of possibilities.
- The graphical form of a probability distribution for a set of outcomes X to $P(X)$ is called a **density curve**. Density curves are always above (or on) the horizontal axis and have *area* 1 underneath them. (If you have had Calculus, this deals with changing a Σ in the discrete case to a $\int$ in the continuous case.) Thus, there is a correspondence between area under a density curve and probability.
- We call a probability distribution a **normal distribution** if the density curve of the distribution is *symmetric* and *bell shaped*. Normal distributions come up all the time; in fact, you have probably seen them before: good old bell curves.

  There is a more precise definition for normal distributions but it gets pretty technical... here it is just in case you are curious: The distribution of X is normal if and only if

  $$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-.5[(X-\mu)/\sigma]^2}$$

  where μ is the mean of the distributions and σ is the standard deviation.
- All normal distributions have a lot in common; however, depending on the mean and standard deviation, they can be more of less spread out. To account for this, we can *standardize* a normal distribution, which we discussed earlier in the term, to get the **standard normal distribution**. This is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. Once we do this, all standard normal distributions are the same.

6.3 More Probability

The area under the standard normal density curve to the left of an X-value a represents the probability that a randomly chosen individual will have a z-score less than (or equal to) a. You could compute this area yourself with a bit of calculus. Fortunately for you, there is a table which gives areas under the standard normal density curve to the left of all possible X-values. There are also numerous online applications that can give you these areas.

It is useful to know some of these areas by heart. In particular, the **The 68-95-99.7 Rule** says:

- 68% of all z-scores fall within 1 standard deviation of the mean.
- 95% of all z-scores fall within 2 standard deviations of the mean
- 99.7% of all z-scores fall within 3 standard deviations of the mean.

6.4 Exercises

1. Use a table or online application like:

   http://www.stat.tamu.edu/ west/applets/normaldemo.html

   to find the area under the standard normal density curve that is:

   (a) To the left of $z = 0$.
   (b) To the left of $z = 3$.
   (c) To the right of $z = -1.25$.
   (d) To the right of $z = 6.11$.
   (e) To the left of $z = -4.21$.

2. Verify the 68-95-99.7 rule using a table.

3. Find the z-score with 25% of the observations falling below it.

4. How high does a z-score need to be in order to have 99% of all the observations below it?

5. How low does a z-score need to be in order to have 99.99% of the observations above it?

6. The level of cholesterol in the blood is important because high cholesterol levels may increase the risk of heart disease. The distribution of blood cholesterol levels in a large population of people of the same age and same sex is roughly normal. For 14-year-old boys, the mean is $\mu = 170$ milligrams of cholesterol per deciliter of blood (mg/dl) and the standard deviation is $\sigma = 30$ mg/dl. Levels above 240 mg/dl may require medical attention. What percent of 14-year-old boys have more than 240 mg/dl of cholesterol?[2, 3.6]

7. An important measure of the performance of a locomotive is its "adhesion", which is the locomotive's pulling force as a multiple of its

weight. The adhesion of one 4400-horsepower diesel locomotive model varies in actual use according to a normal distribution with mean $\mu = 0.37$ and standard deviation $\sigma = 0.04$.

(a) What proportion of adhesions measured in use are higher than 0.40?

(b) What proportion of adhesions are between 0.40 and 0.50?[2, 3.11]

8. The National Collegiate Athletic Association (NCAA) requires division I athletes to score at least 820 on the combined mathematical and verbal parts of the SAT exam to compete their first college year. (Higher scores are required for students with poor high school grades.) In 2002, the scores of the 1.3 million students taking the SATs were approximately normal with mean 1020 and standard deviation 207. What percent of all students had scores less than 820? [2, 3.21]

9. The heights of women aged 20 to 29 follow an (approximately) normal distribution with mean 64 inches and standard deviation 2.7. Similarly, the heights of men aged 20 to 29 also follow an (approximately) normal distribution with mean 69.3 and standard deviation 2.8. What percent of young women are taller than the mean height of young men? [2, 3.22]

10. (Extra Credit)The heights of people of the same sex and similar ages follow a normal distribution pretty closely. Weights, on the other hand, are not normally distributed. The weights of women aged 20 to 29 have mean 141.7 pounds and median 133.2 pounds. The first and third quartiles are 118.3 and 157.3 pounds. What can you say about the shape of the weight distribution? Why? [2, 3.27]

Chapter 7

Sampling Distributions

7.1 Motivating Exercise

Note to instructor: In order to do this exercise, you must provide the class with a bag filled with a large number of small slips of paper, each slip of paper having a number between 1 and 15 on it. The entire class will share this one bag.

As part of the exercise, the students will be making two histograms out of post-it notes. To set this up for them, draw two horizontal axes across the bottom of the blackboard and evenly space the numbers from 1–15 below the line as labels for the columns of each of the histograms. Provide each group of students with two post-it notes.

This exercise should be done with the students in groups of two.

1. We would like to learn the mean of a large population. In this case, our population is the collection of numbers written on slips of paper in the bag. *You should assume there are too many slips of paper in the bag to reasonably look at all of them.*

   Randomly draw one slip of paper from the bag. Look at it then put it back. (Pass the bag to the next group.) How much information does this give you about the mean?

2. Randomly draw 10 slips of paper out of the bag, and record each number, replacing after each draw. (When you finish, pass the bag to the next group.) Compute the mean of the 10 drawn. Based on the 10 you drew, predict the mean of the entire bag.

   (a) How accurate do you think this prediction is?

   (b) What could you do to more accurately predict the mean?

3. Round the mean you computed to the nearest whole number. Using a post-it note, create a histogram on the blackboard by "plotting" your

mean one of the axes drawn on the board. If more than one person has the same mean, the post-it notes should be put on top of one another so as to make that column vertically taller.

4. Compute the standard deviation of the distribution potted in this histogram treating each mean as one data point.
5. Repeat number 2 drawing 25 slips of paper from the bag rather than 10.
6. Using a post-it note, create a new histogram on the blackboard by "plotting" your mean of the 25 slips drawn.
7. Using the histogram created by you and your classmates, predict the mean of the entire bag and discuss the accuracy of this prediction.
8. Compute the standard deviation of the distribution plotted in this new histogram, again using each mean as one data point. How does it differ from the standard deviation found in number 4? [1]

[1] Thank you to Jeremy Alm of Illinois College for suggesting this sort of exercise.

7.2 Sampling distribution

- We want to use *statistics* (which we can compute) of a random sample drawn from a population to estimate *parameters* (which we cannot compute) of the population itself.

- A **sample statistic** is a number that describes a sample. It is calculated from the observations in the sample. For example, the mean, median, and standard deviation of a sample are each sample statistics.

- A **parameter** is a number that describes a population. Often the value of a parameter is unknown because we cannot examine the entire population. The mean, median, and standard deviation of a population are all parameters.

- **The Law of Large Numbers:** Draw a sample at random from any population with finite mean μ. As the sample size increases, the mean $\bar{x}$ of the sample gets closer to the mean of the population.

- The **sampling distribution** of a sample statistic is the distribution of all values taken by the statistic in all possible samples of the same size taken from the population.

- **Theorem for the Sampling Distribution of Sample Mean**: If individual observations have a normal distribution with mean μ and standard deviation σ, then the *sampling distribution of the sample mean* for samples of size n is also normal and has mean μ and standard deviation $\sigma/\sqrt{n}$.

- If the sampling distribution of a sample statistic has a mean equal to the population parameter, then the statistic is said to be an **unbiased estimate** of the parameter. If the mean of the sampling distribution is not equal to the parameter, the statistic is said to be a **biased estimate** of the parameter.

  The mean is an unbiased estimate based on the theorem above. Our version of variance and standard deviation are actually (downward) biased estimates of the population variance and standard deviation. To estimate the population in an unbiased fashion, it turns out we just need to change the n to $n-1$ in the denominator of the variance formula. Why is this true? Good question. This is not something we can answer in this class. However, be aware, oftentimes the variance and standard deviation are *defined* with the $n-1$ in the denominator rather than a n for this very reason.

- **The Central Limit Theorem:** Draw a random sample of size n from a population with mean μ and finite standard deviation σ. When n is large, the *sampling distribution of the sample mean* is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Thus, for large enough n, the *sampling distribution of the sample mean* is **normal** even if the original distribution is not. Pretty cool.

7.3 Exercises

1. A roulette wheel has 38 slots, of which 18 are black, 18 are red, and 2 are green. When the wheel is spun, the ball is equally likely to come to rest in any of the slots. One of the simplest wagers chooses red or black. A bet of $1 on red returns $2 if the ball lands in a red slot. Otherwise, the player loses his dollar. When gamblers bet on red or black, the two green slots belong to the house.

   (a) What is the expected gain on a $1 bet on red?

   (b) One way to think of the expected value in this case is that it is the "mean loss/gain per bet". Explain what the law of large numbers tells us about what will happen if a gambler makes many bets on red. [2, 10.19]

2. A population's mean is 60 and its standard deviation is 5. In each case below, different information about the population's distribution is given. Find the probability that the result obtained will either be greater than 66 or less than 54 in each case. (Note: You might answer "not enough information" to some.)

   (a) Consider a single random individual chosen from the population. (Note that nothing is known about the population's distribution.)

   (b) The population's distribution is normal. Consider a single random individual chosen from the population.

   (c) The population's distribution is normal. Consider the mean of a random sample of 4 individuals from the population.

   (d) The population's distribution is normal. Consider the mean of a random sample of 18 individuals from the population. [1, 3.1.4]

3. Sheila's doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is a variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational diabetes if the glucose level is above 140 milligrams per deciliter one hour after a sugary drink is ingested. Sheila's measured glucose level an hour after ingesting a sugary drink varies according to the Normal distribution with $\mu = 125$mg/dl and $\sigma = 10$mg/dl. [2, 10.24]

   a.) If a single glucose measurement is made, what is the probability that Sheila is diagnosed as having gestational diabetes?

   b.) If measurements are made on 4 separate days and the mean result is compared with the criterion 140mg/dl, what is the probability that Sheila is diagnosed as having gestational diabetes?

c.) Find the level L such that there is a probability of only .05 that the mean glucose level of 4 test results falls above L for Sheila's glucose level distribution. What is the value of L?

4. What is the difference between the *Theorem for the Sampling Distribution of Sample Mean* and the *Central Limit Theorem*?

5. The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4. This distribution takes only whole-number values so is certainly not Normal.

    a.) Let $\bar{x}$ be the mean number of accidents per week during a year. What is the approximate distribution of $\bar{x}$ according to the Central Limit Theorem?

    b.) What is the approximate probability that $\bar{x}$ is less than 2? [2, 10.27]

6. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital, about 15 are born each day.

    (a) As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

    For a period of one year, each hospital records the days when more than 60% of the babies born are boys. Which hospital would you expect to record more such days?

    (b) The length of each newborn baby is measured in both hospitals.

    For a period of 1 year, each hospital records the days then the longest baby born that day exceeds 56 cm. Which hospital would you expect to record more such days? [1, 6.1.4]

7. We are interested in estimating the positive *range* of a large and finite population by the range of a random sample (of size 20) from this population. Recall that the range of a distribution is the largest value minus the smallest value. Is this estimate of the range an unbiased estimate, and *upward* biased estimate (too big), a *downward* biased estimate (too small), or is it impossible to tell? [1, 6.1.5]

8. A normally-distributed population has an unknown mean μ_x and standard deviation $\sigma_x = 20$.

    (a) What is the probability that a random observation X from this population will satisfy

$$\mu_x - 2 \leq X \leq \mu_x + 2?$$

In other words, what is the probability that X falls within the range $\mu_x \pm 2$?

(b) What is the probability that the mean of a random sample of size 144 from the population will fall within the range $\mu_x \pm 2$?

(c) What is the probability that the mean of a random sample of size 400 from the population will fall within the range $\mu_x \pm 2$?

Chapter 8

Confidence Intervals

We want to find the value of a population parameter using information gained from a sample statistic. In this section, we will focus on estimating a range in which we expect the *mean* of the population to fall using the mean of a random sample. In computing this range, we will give a level of “confidence”, which is essentially the probability that this computed range indeed captures the true mean of the population.

8.1 Exercises

1. This is basically exercise 8 from the previous section, just in reverse. Suppose that you have a normally distributed population with unknown mean μ_x and standard deviation $\sigma_x = 20$. Given that the probability that a random observation X will fall within the range $\mu_x \pm E$ is .95 or 95%, find E.

2. Let’s do the more general version. Suppose you have a normally distributed population with unknown mean μ_x and known standard deviation σ_x. Given that the probability that a random observation x will fall within the range $\mu_x \pm E$ is .95 or 95%, find E in terms of σ_x.

3. Now suppose that our population is not necessarily normally distributed with unknown mean μ_x and known standard deviation σ_x. Suppose we draw a random sample of size n, where n is pretty big. We want to use the mean of this sample $\bar{x}$ to estimate the mean of the population. Let’s consider an interval around $\bar{x}$ that is 4 standard deviations wide. That is the interval
$$\bar{x} \pm 2\frac{\sigma_x}{\sqrt{n}}.$$
What are the chances that this interval will enclose the true population mean?

4. A **confidence interval** is a range of values that is likely to contain the true value of a population parameter. The **degree of confidence**, or **confidence level** is the probability that the population parameter is contained in the confidence interval. Common confidence levels include 99%, 95%, and 90%.

   Suppose we are in the situation of the previous problem. Find a formula for the confidence interval to estimate the mean of the population for each of these confidence levels from the mean of the sample. That is, μ_x will be in the interval $\bar{x} \pm E$ with a probability of .90, find E in terms of σ_x and n. Replace .90 with .99 and .95 and repeat, finding E for each.

5. A laboratory scale is known to have a standard deviation of $\sigma = 0.001$ gram in repeated weighings. Scale readings in repeated weighings are normally distributed, with mean equal to the true weight of the specimen. three weighings of a specimen give (in grams)

   $$3.412, 3.414, 3.415$$

   Give a 95% confidence interval for the true weight of the specimen.

6. In the previous problems, we assumed that we did not know the population mean but that we did know the population standard deviation. This is not realistic. However, for large enough samples, more specifically for samples with $n \geq 30$, we can use the standard deviation of the sample in place of the standard deviation of the population. We call the standard deviation of the sample s_x rather than σ_x.[2, 13.4]

   What are the three confidence interval formulas in this situation? In other words, find E in terms of s_x and n for each level of confidence from problem (4).

   NOTE: To find confidence intervals for smaller sized samples, we need to consider t statistics instead of z statistics. We will learn about this in the next section.

8.2 Confidence Intervals and t-distributions

If we do not know mean of the population, we certainly do not know the standard deviation of the population. We must rely on the standard deviation of the sample, which we can compute. To find a confidence interval for “pretty big” sample sizes, the central limit theorem tells us that the sampling distribution of the sample mean is *normal*, even if the original distribution is not. We then find the confidence interval for the mean using the standard normal density curve to get the desired confidence interval for the mean with the formula

$$\bar{x} \pm z_* \frac{s_x}{\sqrt{n}}$$

where z_* depends on the level of confidence. For example, if we want a 95% confidence interval to estimate μ_x, $z_* = 1.96$. You should have found this in exercise (6) in the previous section.

What if our sample is not very big? If our original distribution is normal, we can still find a confidence interval with one modification. We use the **t-distribution** rather than the z-distribution. This distribution is not the standard normal distribution. It varies with the sample size n. Thus, the confidence interval will depend on both the level of confidence and the sample size. We define the **degrees of freedom= df**= n-1. The t-distribution table will tell you the area under the t-distribution density curve.

Thus, the formula for the confidence interval for the mean of a normal distribution is:

$$\bar{x} \pm t_* \frac{s_x}{\sqrt{n}} \tag{8.1}$$

where t_* depends on both the confidence level and the degrees of freedom. For sufficiently large n, the z and the t-distributions coincide.

1. Suppose you draw a random sample of size 100 (with sample mean $\bar{x}$ and standard deviation s_x) from a population with unknown mean and unknown standard deviation. Find t^* in equation (8.1) above for the confidence levels .90, .95 and, and .99.

2. Observational studies suggest that moderate consumption of red wine may reduce the risk of heart attack. One reason may be that red wine contains polyphenols, substances that do good things to cholesterol in the blood. In an experiment, healthy men were assigned at random to several groups. One group of 9 men drank a half a bottle of red wine each day for two weeks. The level of polyphenols in their blood was measured before and after the two week period. Here are the percent changes in level:

   3.5, 8.1, 7.4, 4.0, 0.7, 4.9, 8.4, 7.0, 5.5

Give a 90% t-confidence interval for the mean percent change in blood polyphenols among all healthy men if all drank this amount of red wine and discuss what your findings tell us. [2, 16.6]

3. How do confidence intervals change as n increases?

4. How do confidence intervals change as the level of confidence increases?

Chapter 9

Classical Statistics: Significance

What happens when we obtain an experimental result that our probabilistic model says happens almost never? Do we conclude that we have simply obtained a very rare result, or is it reasonable to question our model?

Let H_0 denote the **null hypotheses** and let R be an observed result of an experiment. If $P(R|H_0) \leq \alpha$, we say the result is **statistically significant** (or just **significant**) at the α-level. Worded differently, assuming that H_0 is true, if the probability of the result R occurring merely by chance is less than or equal to α, we say the result is statistically significant at the α-level.

$P(R|H_0)$ is sometimes called the ***p*-value** of the experiment. The smaller the p-value, the more evidence we have *against* the null hypotheses.

9.1 Exercises

1. An environmental group collects a liter of water from each of 45 random locations along a stream and measures the amount of dissolved oxygen in each specimen. The mean is 4.62 milligrams (mg). Is this strong evidence that the stream has a mean oxygen content of less than 5 mg per liter?

   To do this, suppose you know that the dissolved oxygen varies among locations according to a normal distribution and the standard deviation is $\sigma = 0.92$ mg.

   Let H_0 be the hypothesis that the mean of the entire stream is indeed 5 mg. We want to compute the probability that a sample will have a mean of 4.62 mg or lower. Find the p value of this experiment. Is this result significant at the $\alpha = .05$ level? Discuss your results. [2, 14.17]

2. A student group claims that first-year students must study 2.5 hours per night during the school week. A skeptic suspects that they study less

than that on average. A class survey finds that the average study time claimed by 269 students is $\bar{x} = 137$ minutes. Regard these students as a random sample of all first-year students and suppose we know that study times follow a normal distribution with standard deviation 65 minutes. Carry out a significance test of the null hypothesis

$$H_0 : \mu = 150.$$

In other words, find the p-value for this sample result. Does this result give good evidence that the mean study time of first-year students is less that 2.5 hours per week? [2, 14.26]

3. We suspect that on average students will score higher on their second attempt at the SAT mathematics exam than on their first attempt. Suppose we know that the changes in score (second try minus first try) follow a normal distribution with standard deviation $\sigma = 50$. Here are the results for 46 randomly chosen high school students:

   -30 24 47 70 -62 55 -41 -32 128 -11 -43 122 -10 56 32 -30 -28 -19 1 17 57 -14 -58 77 27 -33 51 17 -67 29 94 -11 2 12 -53 -49 49 8 -24 96 120 2 -33 -2 -39 99

   Do these data give good evidence that the mean change in the population is greater than zero? State the null hypotheses, H_0 and calculate the p-value for the results obtained. Clearly state your conclusion. [2, 14.18]

4. In a discussion of the educational level of the American workforce, someone says "The average young person can't even balance a checkbook." The National Assessment of Educational Progress (NAEP) says that a score of 275 or higher on its quantitative test reflects the skill needed to balance a checkbook. The NAEP random sample of 840 young men had a mean score of $\bar{x} = 272$, a bit below checkbook balancing level. Is this sample result good evidence that the mean for all young men is less than 275? Are your results significant at the $\alpha = .1$ level? At the $\alpha = .05$ level? [2, 14.8]

9.2 Exercises: The Classics vs. the Bayesians

1. The experiment described below is designed to clarify the difference between 'classical' inference (significance testing) and Bayesian inference.

   Suppose you confront 10 opaque urns. You know that all of them contain 7 beads, and the urns are divided into two types: There are nine urns of type A, each containing 5 white beads and 2 black beads. There is one urn of type B, containing 5 black beads and 2 white beads.

You randomly choose one of the ten urns and formulate two complementary hypotheses about it:

H_0: The urn is of type A.

H_1: The urn is of type B.

You now apply a decision procedure in line with classical statistics. You blindly draw two beads without replacement from the urn in question. If the two beads are black, H_0 is rejected and H_1 is accepted, otherwise, H_0 cannot be rejected. [1, 3.1.10]

Let R denote the event that H_0 is rejected (i.e., two black beads are drawn).

(a) Compute the level of significance of the test, that is compute the p-value $P(R|H_0)$.

(b) Compute the Bayesian posterior probability of H_0, given R, that is $P(H_0|R)$.

2. Let's revisit exercise (1) in section 4.12 where H denotes the genotypic event, or the hypothesis that the couple has the potential to produce a blond child. This time, let's approach this problem 'classically' (since many of you *did not* enjoy approaching it as a Bayesian).

   (a) Suppose the couple has two dark haired children in a row (and no blond children). What is the significance of this event? In other words, compute $P(D_2|H)$.

   (b) Suppose the couple has 8 dark haired children in a row (and no blond children). At this point, do you have enough evidence to reject the null hypothesis at the $\alpha = .01$ level?

   (c) Suppose the couple has 20 dark haired children in a row (and no blond children). At this point, do you have enough evidence to reject the null hypothesis at the $\alpha = .01$ level?

   (d) Do you prefer this method of determining whether H is true or the previous method used in section 4.12? (There is no right answer here.)

   (e) Suppose the couple has a blond baby. What is the significance of this event?

3. Two cab companies, the blues and the greens, operate in a given town: 85% of the cabs in town are blue, 15% are green.

   One night a cab was involved in a 'hit-and-run' accident. An eyewitness claimed that the cab involved was *green*. The court examined the discrimination capacity of the witness, given the illumination conditions at the time and site of the accident. They determined that the witness identifies colors correctly in 80% of the cases presented to him (and made a wrong judgment in 20% of the cases.)

(a) What is the probability that the cab involved in the accident was indeed green?

You have no information whatsoever about the driving records of the two companies, nor about the differential distributions in various parts of the town. You may therefore start by regarding the cab involved in the accident as if it were randomly picked from the totality of cabs in town.

(b) Upon further investigation, another eyewitness was found. The new witness independently supported the first witness's testimony, namely, that the car involved was a *green* cab. It turned out that this witness's capacity for color discrimination was equal to that of the first witness.

Two judges tried the case:

Judge C (the classic) decided that he would find the green cab company guilty if the probability of getting two such testimonies, given that the hit-and-run cab was blue, is significant at the .05 level.

Judge B (the Bayesian) decided to calculate the posterior probability that the hit-and-run cab was green, given the two testimonies. He decided to find the green company guilty if that probability exceeded 95%.

Carry out the calculation suggested by each judge and state the conclusion that follows from it. Which verdict would you support? Explain. [1, 3.1.11]

4. A Society for Beyond the Sensory announced the opening of a new journal, and invited investigators in the behavioral sciences to submit experimental papers in the field. The editor claimed that he would publish any paper that presented results significant at a level of $\alpha = 0.05$, indicating extra-sensory perception (ESP).

   Ten investigators responded to the challenge and, independently of each other, started experiments designed to find out whether their subjects perceive extra-sensory messages above random (guessing) level.

   Assume that none of the subjects perceive anything and that they just guess (this is the null hypothesis). What is the probability that at least one paper will be published (as a result of 10 experiments), heralding significant ESP findings? [1, 3.1.12]

5. There is wide consensus among scientists and methodologists that replications are crucial for establishing the validity of research results.

   Let us examine quantitatively some of the contingencies involving two independent replications of a given experiment.

   Two investigators conduct the same kind of experiment on two independent random samples from the same populations. Both test the same null hypothesis H_0 and, assuming H_0 is true, each of the tests is conducted at a level of significance of 0.05.

   Compute the following:

   (a) The probability that both tests will turn out statistically significant.

   (b) The probability that at least one experimenter will obtain a statistically significant result.

   (c) The probability that only one of the tests will come out statistically significant.

   (d) The probability that neither test will turn out significant.

   (e) The conditional probability that the second investigator will obtain a statistically significant result, given that the first one did so.[1, 3.1.13]

9.3 Some Culminating Review Exercises

1. Test A was given to a large class. The distribution of the scores, denoted x, is characterized by several measures. The names of these measures (and their symbols) are written in the first column of the chart below. The numerical values of these measures are written in the second column.

   The teachers' council concluded that the test had been too 'harsh' and considered two schemes for adjusting the scores of all the students.

   (a) Five points would be added to each of the student's original score.

   (b) Each score would be increased by 15% of its original value (that is, multiplied by 1.15).

   Fill the table with the measures characterizing the adjusted distribution of scores according to each of the two schemes. [1, 3.1.1]

Measure	Original scores x	scheme (a) $x+5$	scheme (b) $1.15x$
Mean $\overline{x}$	58		
Standard deviation s_x	12		
Variance s_x^2	144		
Median Me_x	59		
Lowest standard score	-2.8		
Range R_x	64		
Mode Mo_x	60		
Percentile of student i	78%		
r_{xy} with scores on test B	0.55		
r_s with scores on test C	0.61		

2. A statistician performed a squaring transformation on a set of n non-negative x values, and obtained a set of y values such that $y = x^2$.

   The following is a list of measures characterizing the distribution of the variable x. For each measure, check whether squaring the x measure gives the corresponding y measure. That is, given that m is some measure of the distribution, determine whether $m_x^2 = m_y$. [1, 3.1.2]

   (a) $\overline{x}$

   (b) s_x

   (c) Me_x

   (d) Mo_x

   (e) That largest x value

   (f) $Q_3(x) - Q_1(x)$

(g) R_x

(h) r_{xu} with some other distribution u

(i) r_s with some other distribution v

(j) the z-score of the mean

(k) the standard deviation of the standardized distribution.

3. A statistics teacher told four students their scores in a somewhat playful manner. She informed them that the distribution of the class's scores was approximately normal with mean 70 and standard deviation 9, and she gave them the following information about their own academic achievements:

   Jack: "Your score splits the class. One quarter has higher scores; three quarters have lower scores."

   Jill: "Your raw score is 88."

   John: "Your standard score is -1.4."

   Jane: "Only 10% of the class got higher scores than yours."

   Fill in the missing scores in the table below: [1, 3.1.4]

Student's name	Raw Score	Standard Score	Percentile
Jack			75
Jill	88		
John		-1.4	
Jane			90

4. Does eating more fiber reduce the blood cholesterol level of patients with diabetes? A randomized clinical trial compared normal and high-fiber diets. Here is part of the researchers' conclusion:

   The high-fiber diet reduced total cholesterol concentrations by 6.7 percent (with a p-value of .02), triglyceride concentrations by 10.2 percent (p-value .02), and very-low-density lipoprotein cholesterol concentrations by 12.5 percent (p-value 0.01).

   A doctor who knows no statistics says that a drop of 6.7% in cholesterol isn't a lot–maybe it's just an accident due to chance assignment of patients on the two diets. Explain in simple language how the p-value answers this objection. [2, 14.39]

9.4 Two Extra Credit Problems

1. It is known that Tom and Dick tell the truth only a third of the time (the probability that each is lying is $\frac{2}{3}$ every time).

   Tom makes a statement, and Dick tells us that Tom was speaking the truth.

   What is the probability that Tom was actually telling the truth? [1, 2.4.9]

   Hint: The answer is not $\frac{1}{3}$.

2. Tom, Dick, and Harry are three liars. Each of them tells the truth only a third of the time. Tom makes a statement, and Harry tells us that Dick said that Tom was speaking the truth.

   What is the probability that Tom was actually telling the truth? [1, 2.4.10]

   Hint: Derivation of the required probability is not immediate. Figure out what conditional probability you need to find then use a ‘tree’ to reach the likelihoods that should be inserted into the Bayesian calculation.

   The answer is not $\frac{1}{3}$ here either.

Bibliography

[1] R. Falk, *Understanding Probability and Statistics: A Book of Problems,* A. K. Peters, Ltd, Wellesley , MA 1993.

[2] D. S. Moore, *The Basic Practice of Statistics, third edition,* W. H. Freeman Company, New York, NY 2004.

[3] http://www.csulb.edu/ lhenriqu/AETS2005.htm