MTH 243 Supplement to OpenStax College
Introductory Statistics

July 12, 2016

Download *Introductory Statistics* for free at http://cnx.org/contents/30189442-6998-4686-ac05-ed152b91b9de@18.10.

Preface

This MTH 243 Supplement was produced with grant support from Open Oregon for PCC students using *Introductory Statistics* by Barbara Illowsky and Carol Dean from https://openstax.org/.

For complete PCC Course Content Outcome Guide, see https://www.pcc.edu/ccog/default.cfm?fa=ccog&subject=MTH&course=243.

Contents

Chapter 1

Sampling & Data

1.1 Identify the differences between experiments and observational studies. Identify the elements of experiments and observational studies including: factors.

Observational studies and experiments are important surveying methods in collecting data with clear differences that should be examined.

In an observational study, researchers do not have control over what occurs, they merely identify variables of interest and collect data. Data may already exist, or it may be gathered as time passes. A study using existing data is called a *retrospective study*; a study collecting data as time passes is called a *prospective study*.

Example 1. *A community college investigated the highest math class taken by any student finishing with a one or two year certificate within the last ten years. To do this, they examined all records and recorded the information. Examining existing records makes this a retrospective study.*

Example 2. *A new class is created at a high school and faculty want to collect information about the class, such as grades of students. Since data will be collected as time passes, this is a prospective study.*

In an experiment, researchers do have control over what occurs.

Example 3. *Consider the simplified description of an experiment*

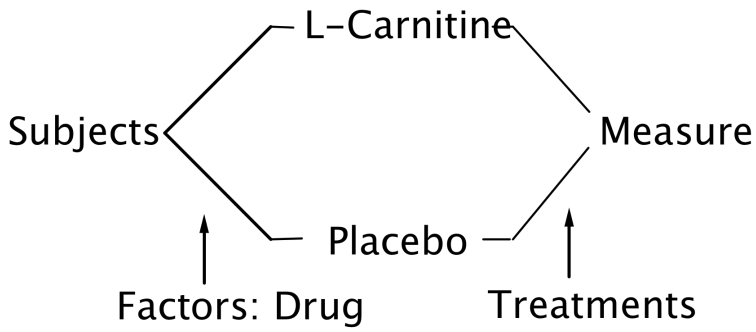

FIGURE 1.1: Diagram of L-Carnitine experiment.

http://www.ncbi.nlm.nih.gov/pubmed/18065594.

Sixty-six centenarians (people aged 100 or more) were split into two groups by researchers. Once group received a daily dose of L-Carnitine, the other received a placebo. After the study was complete, the research team used the Mini-Mental State Examination (MMSE) to measure each centenarian's mental fatigue and cognitive function. Results showed a statistically significant increase in capacity for cognitive activity for those centenarians given L-Carnitine compared to those give the placebo.

An experimental diagram of this situation can be seen in Figure 1.1

In experiments, researchers look for a relationship between at least two variables. Assuming we only have two variables, we begin by identifying *factors* that we want to manipulate (the dosage received) and *response variables* to measure (mental fatigue or cognitive function). When factors are manipulated, their specific values are known as *levels* (if they receive a dose of L-Carnitine or the placebo). We then find *subjects* to participate in the experiment (people aged 100 or more). Those subjects are then randomly assigned to a combination of levels from all factors, known as a *treatment* (they receive a dose that is a placebo.)

1.2 Identify sample designs including: multistage sample.

Multistage sampling combines more than one sampling method.

Example 4. *During election cycles, polling agencies will do a multistage samples. As a first stage, they may do an SRS of a large number of election districts, perhaps 1000. As a second stage, they would sample 10 household within each of those districts.*

Example 5. *The Portland City of Portland wishes to investigate neighborhood street conditions and will use a multistage sample. As a first stage, they do a cluster sample of Portland neighborhoods. As a second stage, instead of sampling (or checking) every street in the clusters, they do a random sample of streets and investigate those.*

1.3 Identify and describe terminology: Biased vs. Unbiased

In one dictionary, bias is defined as "prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair." Wikipedia defines bias to be "an inclination towards something, or a predisposition, partiality, prejudice, preference, or predilection." These definitions match many people's general ideas of the meaning of bias. This kind of bias can be conscious or unconscious, and it may affect people's lives and livelihoods.

In statistics, *bias* is defined as "the systematic distortion of a statistic" (Wikipedia). Systematic distortion of a statistic often arises from the sampling process. A statistic from a voluntary response sample, for example, nearly always has bias toward extreme views.

A biased sample is not representative of the population from which the sample was drawn. That means the statistics of the sample do not reflect the parameters of the population. Good sampling methods avoid bias. An unbiased sample represents its population well in the sense that each sample statistic is a good estimate of the corresponding population parameter. Unfortunately, sources of bias in a sample are not always clear to researchers, but ethical researchers do their best to use a representative sample, minimizing the effect of bias.

1.4 Homework

1. Decide if each of the following situations is a prospective or retrospective study.

   (a) The City of Portland wants to find out how many one bedroom, two bedroom and three bedroom units there are in the city.

   (b) The DMV has participants sleep for 2, 4 or 6 hours and then has then drive a simulator to measure reaction time to dangerous situations.

   (c) A large college wants to know the distribution of race / ethnicity of students applying to the college in 2017, 2018 and 2019.

   (d) A local farm tries various planting and irrigation methods to see if it helps with crop growth.

2. Astronauts are randomly assigned a set of pills to see if it influences incidence of kidney stones in outer space. They do not know if they are given pills containing potassium citrate or placebos. They then collect their urine sample while in outer space.

   (a) Create a diagram that illustrates this experiment.

   (b) Identify the subjects of the experiment.

   (c) Identify the factors, levels, and treatments in the experiment.

3. Decide if the following sampling methods are biased or not. If they are biased, explain why.

   (a) A news network uses a Twitter survey to report opinions on a current topic.

   (b) The survey agency Gallup does a national survey by calling randomly dialed mobile phones.

   (c) Portland Tri-Met randomly surveys Red Line trains to collect people's opinions of train services.

   (d) PCC does student online evaluations before grades are released at the end of every term to have students give opinions on their classes.

Chapter 3

Probability

3.1 Bayes Theorem

The tree diagram in Figure 3.1 shows a situation where an individual leaves their car at a mechanic's shop. There are three mechanics: Miranda, Jose, and Tanya. Miranda is the most experienced mechanic, while Tanya is the least experienced. Miranda repairs 55% of the cars; Jose repairs 30% of the cars; and Tanya repairs the remaining 15% of the cars. Miranda's repairs fail for 5% of cars on which she works; Jose's fail for 10%; and Tanya's repairs fail for 20%.

This presents an interesting application of conditional probability. Using the tree diagram above we can compute some probabilities.

Example 6. *Selecting one car at random that has been repaired at this shop, use the tree diagram to compute the probability that*

1. *Miranda repaired the car, $P(M)$,*

2. *the repair failed if Tanya worked on the car, $P(F|T)$,*

3. *Jose repaired the car and the repair was successful, $P(J \text{ AND } S)$.*

Notice that we can work top to bottom through the "branches" of the tree diagram to consider different probability questions. The top branch contains the "simple" probabilities of Miranda, Jose, and Tanya repairing a randomly selected car, like $P(M)$ above. Working down, the second branch contains conditional probabilities of failure or success, depending on which mechanic

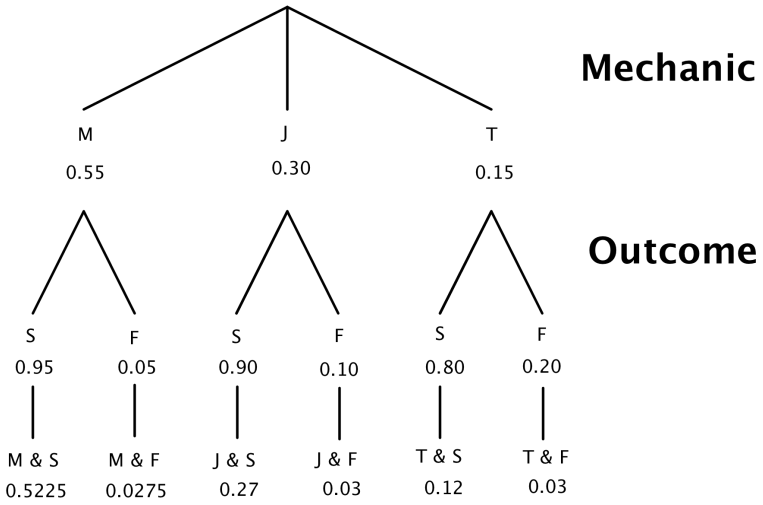

FIGURE 3.1: Tree diagram for repairs at the mechanic's shop.

performed the repair like $P(F|T)$ above. We multiply simple and conditional probabilities along the tree diagram's branches to calculate joint probabilities of various outcomes, like $P(J\ AND\ S)$ above. With a complete tree diagram, we can ask and answer questions like those above quickly.

But what about probability questions that the diagram does not answer directly?

Example 7. *Selecting one car at random, what is the probability that Tanya repaired the car, if that repair failed, $P(T|F)$? This question "reverses the condition" from conditional probability problem above. Though $P(T|F)$ is not in the diagram, we can use the conditional probability formula as follows:*

$$P(T|F) = \frac{P(T\ AND\ F)}{P(F)}$$

Now from the tree diagram, we see

$$P(T\ AND\ F) = P(T) \cdot P(F|T) = (0.15)(0.20) = 0.03.$$

Since there are three mechanics, we can calculate $P(F)$:

$$\begin{aligned} P(F) &= P(M\ AND\ F) + P(J\ AND\ F) + P(T\ AND\ F) \\ &= (0.55)(0.05) + (.30)(0.10) + (0.15)(0.20) \\ &= 0.0875 \end{aligned}$$

Then we have $P(T|F) = \frac{P(T \text{ AND } F)}{P(F)} = \frac{0.03}{0.0875} \approx 0.343$*, so there is a 34.3% chance that if a random repair failed, Tanya did that repair.*

Bayes Theorem is a mathematical formula that encodes this process. In this case, *Bayes Theorem* combines our work above into a formula like so:

$$P(T|F) = \frac{P(T) \cdot P(F|T)}{P(M \text{ AND } F) + P(J \text{ AND } F) + P(T \text{ AND } F)}$$

Computation using this formula is as complicated as the computations we did above using the tree diagram to guide our work.

At the introductory level, it is wise to focus on using the tree diagram and the conditional probability formula to "reverse the condition," rather than to worry about memorizing and using the Bayes Theorem formula.

3.2 Calculate and interpret marginal distribution

Consider the contingency table shown below for the number of males and females in a class using an OER (Open Educational Resource) or a textbook.

	OER	Textbook
Male	10	25
Female	20	5

Since this table shows counts, we would also refer to it as a frequency table. Since the totals for each sex and for each class type are also of interest, we can include a row and a column for the respective totals, as shown below.

	OER	Textbook	Total
Male	10	25	35
Female	20	5	25
Total	30	30	60

The totals end up in the margins of the contingency table. The Total column shows the *marginal distribution* of sex, and the Total row shows the *marginal distribution* of class type.

The contingency table can also be represented using proportions as shown below.

	OER	Textbook	Total
Male	0.17	0.42	0.58
Female	0.33	0.08	0.42
Total	0.50	0.50	1

Here we can see the relative marginal distributions of sex and of class type in the Total column and row, respectively.

3.3 Homework

1. In some states, police used to set up DUI checkpoints where they would stop drivers and try to determine if each driver had been drinking. Assume on a given night 15% of people are drinking. If you have been drinking, police will stop you and give you a breath check 85% of the time. If you have not been drinking, police will stop you and give you a breath check 10% of the time.

   (a) Create a tree diagram for this scenario. Use B = Breath check and D = Drinking.

   (b) What is the probability that you are drinking, given that you did have a break check, that is, find $P(D|B^c)$.

2. Confidence is of the utmost importance for soccer players. A good player will score on 80% of their penalty kicks. If they score on their first penalty kick, there is a 90% chance they will score on a second penalty kick. If they miss their first penalty kick, there is a 70% chance they will not score on a second penalty kick.

   (a) Create a tree diagram for this scenario. Use S = scores and S^c = Does not score.

   (b) Find $P(S)$

   (c) Find $P(S \text{ 2nd}|S^c \text{ 1st})$

   (d) Find $P(S \text{ 1st}|S^c \text{ 2nd})$

3. A food packaging factory has both machines and workers on assembly line. First, machines do part of the packaging and then the worker

finishes it by hand. There is a 1% chance a machine makes a defect in the package. If the machine does make a defect, there is a 70% chance the worker also makes a defect. If the machine does not make a defect, there is a 20% chance the worker makes a defect.

(a) Create a tree diagram for this scenario. Use M = Machine Defect and W = Worker Defect.

(b) Find $P(M \text{ AND } W)$

(c) Find $P(W|M)$

(d) What is the probability the machine made a defect given that a worker made a defect?

4. A computer company is investigating its sales. Customers buy laptops or desktops, which are either new or used. The company knows about 10% of people buy new laptops, about 60% of people buy laptops and about 30% of people buy new computers.

(a) Complete the two way table below.

	New	Used	Total
Laptop	0.10		0.60
Desktop			
Total	0.30		1

(b) Give the marginal distribution of computer style.

(c) Find P(Laptop AND Used).

(d) Find P(New OR Desktop).

(e) Find P(Laptop|New).

5. Consider the table below for products at a grocery store.

	Sale item	Non-sale item	Total
Organic			
Non-organic			
Total			

(a) Assume 40% of products are organic, 20% of products are non-organic and on sale, and 90% of products are neither on sale nor organic. Complete the two way table.

(b) Find P(Non-organic).

(c) Find P(Non-organic AND Non-sale).

(d) Find P(Organic|Sale).

Chapter 4

Discrete Random Variables

4.1 Variance

We can use expected value, μ, and standard deviation, σ, to summarize a random variable X.

Recall that

$$\mu = \sum_{x \in X} xP(x) \qquad \text{and} \qquad \sigma = \sqrt{\sum_{x \in X} (x - \mu)^2 P(x)}$$

The square of the standard deviation, σ^2, is called the *variance* of the random variable. The variance is the sum of the products of the squared differences from the mean and their corresponding probabilities:

$$\sigma^2 = \sum_{x \in X} (x - \mu)^2 P(x)$$

Statisticians use the variance σ^2 when considering mathematical transformations of all the outcomes of a random variable, or when combining more than one random variable.

4.2 Transformations of Random Variables

Refer to the "Try It" Exercise in Section 4.1 of *Introductory Statistics*. The table gives the number of times a post-op patient rings the nurse during a 12-hour shift. The table for the situation, denoted as the random variable X, is shown below.

x	$P(x)$
0	$\frac{4}{50}$
1	$\frac{8}{50}$
2	$\frac{16}{50}$
3	$\frac{14}{50}$
4	$\frac{6}{50}$
5	$\frac{2}{50}$

Using this table, it was found that the expected value is $\mu = 2.32$, the variance is $\sigma^2 = 1.4976$ and the standard deviation is $\sigma = 1.2238$.

Example 8. *Assume there is a malfunction with the button that rings the nurse. This malfunction causes the nurse to be rung one more time than each patient intends, so if a patient rings the nurse zero times, the malfunction causes one ring. With this malfunction, we can denote the situation as the random variable $X + 1$. The table for the situation is shown below.*

$x+1$	$P(x+1)$
1	$\frac{4}{50}$
2	$\frac{8}{50}$
3	$\frac{16}{50}$
4	$\frac{14}{50}$
5	$\frac{6}{50}$
6	$\frac{2}{50}$

We find the expected value of the random variable $X + 1$ to be

$$\mu = 1(\tfrac{4}{50}) + 2(\tfrac{8}{50}) + 3(\tfrac{16}{50}) + 4(\tfrac{14}{50}) + 5(\tfrac{6}{50}) + 6(\tfrac{2}{50}) = 3.32.$$

Note that all the outcomes of the random variable X were increased by one and this made the expected value increase by one, as we see in the random variable $X + 1$.

We calculate the variance σ^2 of the random variable $X+1$ like so:

$$\begin{aligned}\sigma^2 &= (1-2.32)^2(\tfrac{4}{50}) \\ &\quad + (2-2.32)^2(\tfrac{8}{50}) \\ &\quad + (3-2.32)^2(\tfrac{16}{50}) \\ &\quad + (4-2.32)^2(\tfrac{14}{50}) \\ &\quad + (5-2.32)^2(\tfrac{6}{50}) \\ &\quad + (6-2.32)^2(\tfrac{2}{50}) \\ &= 1.4976.\end{aligned}$$

Then we calculate the standard deviation σ of the random variable $X+1$ by finding the square root of the variance σ^2:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.4976} = 1.2238.$$

Note this is the same variance and standard deviation as the original random variable X. We have not changed the distance of any outcome from the expected value, therefore we have not changed the spread.

To summarize, for any random variable X, if we add or subtract the same value to every outcome (without affecting their likelihoods) we denote that transformation by $X \pm c$. Our work above shows that $E(X \pm c) = E(X)$.

Example 9. *Imagine a different ringer malfunction: Every time the patient presses the button to ring the nurse, it rings twice! This means the people who press the button once actually ring the nurse twice! With this malfunction, we can denote the situation as the random variable $2X$. The table for the situation:*

$2x$	$P(2x)$
0	$\frac{4}{50}$
2	$\frac{8}{50}$
4	$\frac{16}{50}$
6	$\frac{14}{50}$
8	$\frac{6}{50}$
10	$\frac{2}{50}$

We calculate the expected value μ to be

$$\mu = 0(\tfrac{4}{50}) + 2(\tfrac{8}{50}) + 4(\tfrac{16}{50}) + 6(\tfrac{14}{50}) + 8(\tfrac{6}{50}) + 10(\tfrac{2}{50}) = 4.64$$

Note this is twice the expected value of the original random variable X, since all the distances from the original expected value have been doubled.

We calculate the variance σ^2 of the random variable $2X$ to be

$$\begin{aligned}\sigma^2 &= (0-4.64)^2(\tfrac{4}{50}) \\ &\quad + (2-4.64)^2(\tfrac{8}{50}) \\ &\quad + (4-4.64)^2(\tfrac{16}{50}) \\ &\quad + (6-4.64)^2(\tfrac{14}{50}) \\ &\quad + (8-4.64)^2(\tfrac{6}{50}) \\ &\quad + (10-4.64)^2(\tfrac{2}{50}) \\ &= 5.9904\end{aligned}$$

We calculate the standard deviation σ of the random variable $2X$ to be

$$\sigma = \sqrt{\sigma^2} = \sqrt{5.9904} = 2.4475.$$

Example 10. *The malfunction has been fixed! A new nurse begins their shift and has two patients, X_1 and X_2, in post op. The nurse is curious about the average number of times the patients will ring in total, that is any patient X_1 and X_2.*

We denote the number of times both patients ring as the random variable $X_1 + X_2$. The table for this combined variable would have outcomes ranging from 0 to 10, since both patients could ring up to 5times, but determining the probability of each outcome would be quite a challenge. (That'd be 11 probability problems!) Luckily, statisticians have methods to compute expected value and standard deviation of $X_1 + X_2$ without reference to a new table.

For expected value, the nurse expects patient X_1 on average to ring them $\mu_{X_1} = 2.32$ times and patient X_2 to ring them $\mu_{X_2} = 2.32$ times, since both random variables X_1 and X_2 have the same expected value, μ. Thus the expected value for $X_1 + X_2$ is $\mu_{X_1} + \mu_{X_2} = 2.32 + 2.32 = 4.64$

The variance and standard deviation calculations are more complicated. One would hope we could simply add the standard deviations for the random variables X_1 and X_2 together. The bad news is that will not work. The good news is that statisticians have learned that as long as two random variables are independent—the outcome of one variable does not affect the outcome of the other—we can add their variances. Here, since one patient's needs do not affect the other patient's needs, X_1 and X_2 are independent.

Then to calculate the standard deviation of the sum, $\sigma_{X_1+X_2}$, we first calculate the variance using the sum of the variances of X_1 and X_2, which is, $\sigma^2_{X_1+X_2} = \sigma^2_{X_1} + \sigma^2_{X_2} = 1.2238^2 + 1.2238^2 = 2.9954$. Then take the square root for the standard deviation of the sum: $\sigma_{X_1+X_2} = \sqrt{2.9954} = 1.7307$

To summarize, for a random variable X_1 and X_2, if we add two random variables X_1 and X_2 together, we denote that transformation by $X_1 + X_2$. For the expected value of $X_1 + X_2$, we can add the expected values of X_1 and X_2. For the standard deviation of $X_1 + X_2$, we must first find the variance. As long as X_1 and X_2 are independent, we can add their variances.

4.3 Homework

1. Refer to Example 4.1 on page 229. The table gives the number of times a post-op patient rings the nurse during a 12-hour shift. The table for the situation, denoted as the random variable X, is shown below.

   Using this table, it was found that the expected value is $\mu = 2.32$, the variance is $\sigma^2 = 1.4976$ and the standard deviation is $\sigma = 1.2238$. Find the expected value and standard deviation for the following transformations on the random variable X

   (a) $X + 7$

   (b) $3X$

   (c) $X - 2$

2. Your instructor has offered extra credit! They take 10 fair dice and put them into a hat. There are 6 white dice, 2 green dice, 2 red dice and 1 pink die. The instructor then has everyone draw one die. If you pick a white die, you receive one free point on the next test, if you pick a green die, you receive two free points, five free points for the red die, but if you pick the pink die, you lose 15 points.

   (a) Let X be the random variable describing the number of free points won or lost as described above. Construct the probability model for X.

   (b) Find the expected value, variance and standard deviation for the random variable X.

(c) Should the students accept their instructor?s offer? Why or why not?

(d) In this situation, what is the meaning of $X + 5$? $2X$? $X + Y$?

3. Las Vegas hotels clean rooms quickly once people have left. The Belagio hotel created the table below for the number of minutes takes to clean a normal room N and a suite room S.

	N	S
μ	50	75
σ	20	25

Find the expected value and standard deviation (and state the meaning in context) for the following transformations to the given random variables

(a) $N - 3$

(b) $2S$

(c) $S + N$

(d) $N + 2S$

Chapter 6

The Normal Distribution

6.1 Approximate a binomial probability using a normal distribution.

There is an interesting relationship between discrete and continuous probability models. As the number of trials increases, each discrete model looks more and more continuous.

Example 11. *Flipping a fair coin a set number of times can be modeled using* $B(n, 0.50)$*. Observe how the binomial distribution changes as we increase* n *from smaller to larger values. In Figures 6.1 through 6.6, we see the distributions for* $n = 1$*,* $n = 5$*,* $n = 20$*,* $n = 50$*,* $n = 100$ *and* $n = 500$*, respectively.*

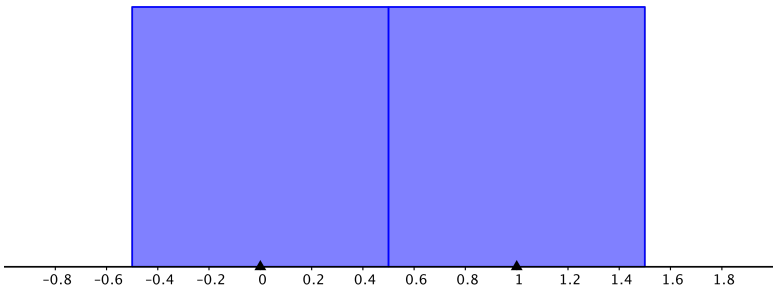

FIGURE 6.1: $B(1, 0.50)$

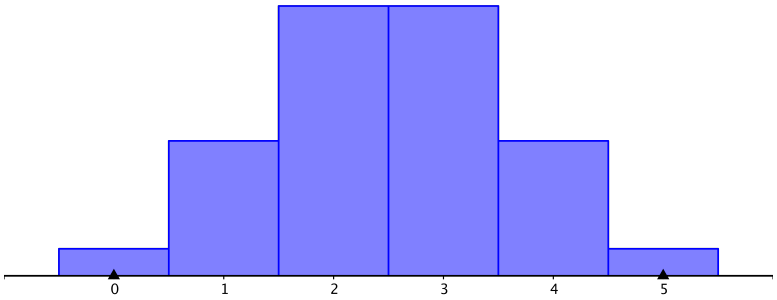

FIGURE 6.2: $B(5, 0.50)$

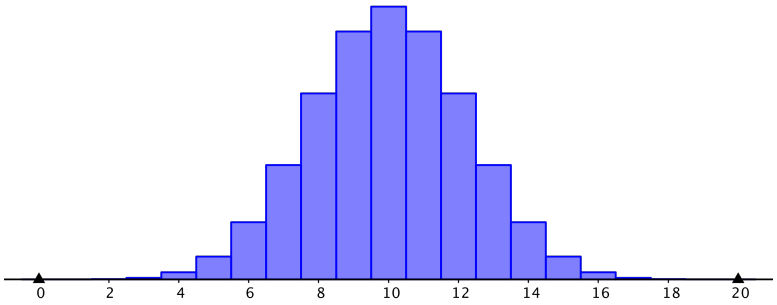

FIGURE 6.3: $B(20, 0.50)$

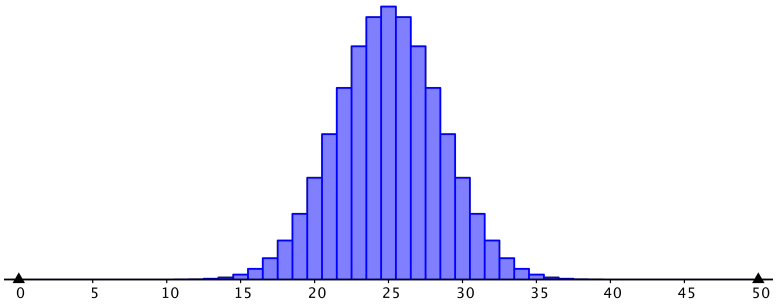

FIGURE 6.4: $B(50, 0.50)$

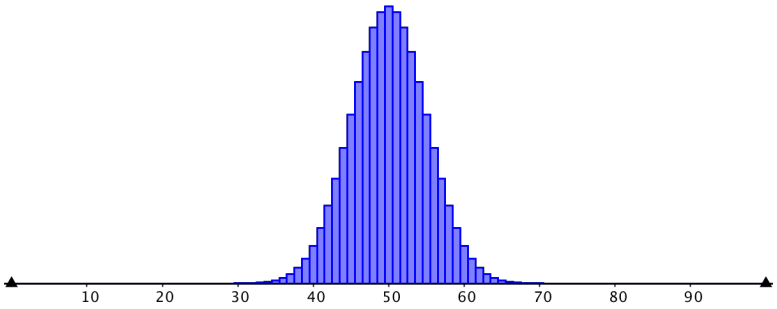

FIGURE 6.5: $B(100, 0.50)$

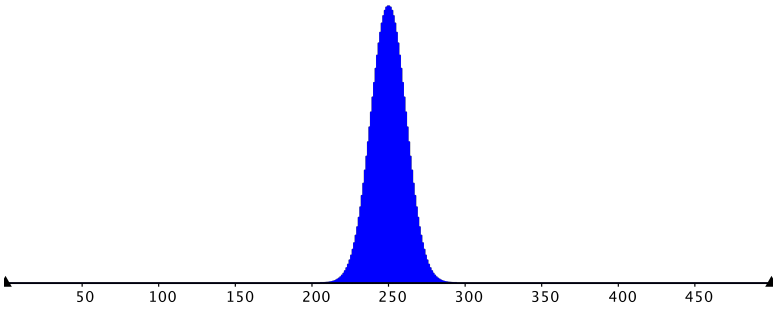

FIGURE 6.6: $B(500, 0.50)$

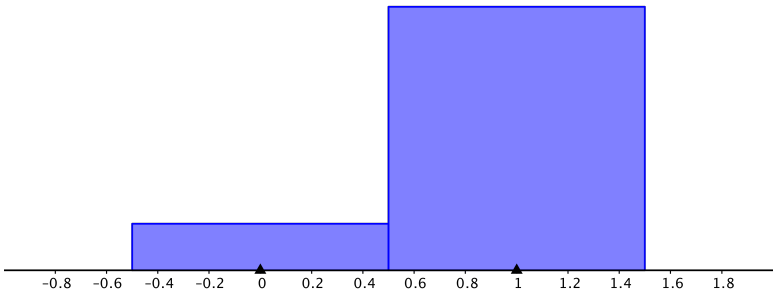

FIGURE 6.7: $B(1, 0.85)$

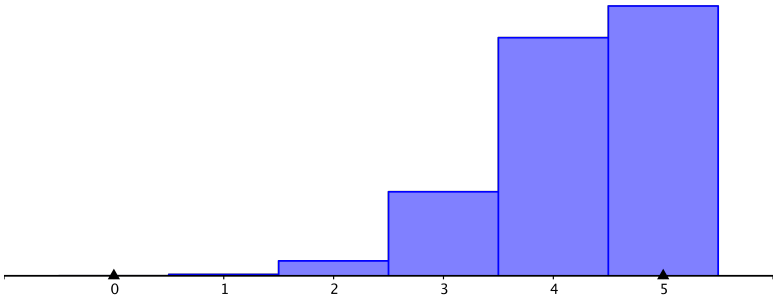

FIGURE 6.8: $B(5, 0.85)$

We see that as n increases, the discrete binomial distribution looks more and more like the continuous normal model. This means that if we conduct enough trials, that is if n is large, a binomial distribution can be approximated using a normal model.

Example 12. *In basketball, an excellent player can make* 85% *of their free throws. We can view that player's free throw attempts as Bernoulli trials with success probability* $p = 0.85$. *Let's consider* $B(n, 0.85)$ *and observe how the binomial distribution changes as we increase* n *from smaller to larger values. In Figures 6.7 through 6.12, we see the distributions for* $n = 1$, $n = 5$, $n = 20$, $n = 50$, $n = 100$ *and* $n = 500$, *respectively.*

For $n = 1$, $n = 5$, and $n = 20$, the binomial distributions are skewed left. As $n = 1$, $n = 5$, and $n = 20$, the binomial distributions are skewed left. As n increases, the distributions look increasingly unimodal and symmetric. If

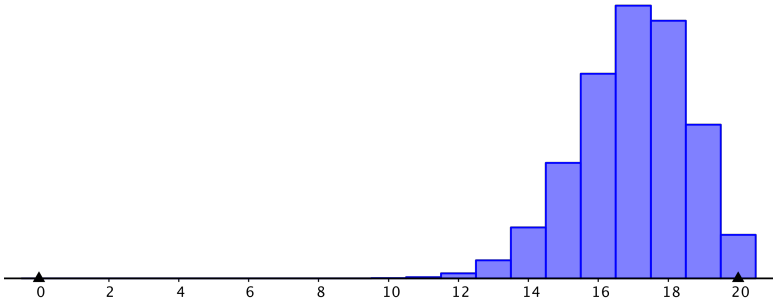

FIGURE 6.9: $B(20, 0.85)$

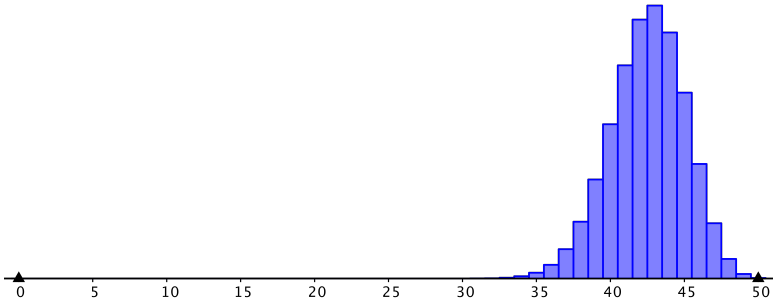

FIGURE 6.10: $B(50, 0.85)$

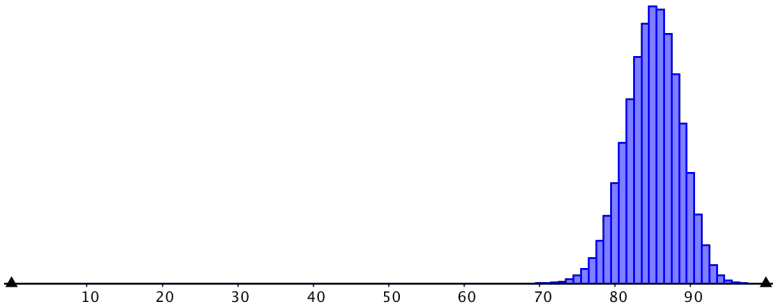

FIGURE 6.11: $B(100, 0.85)$

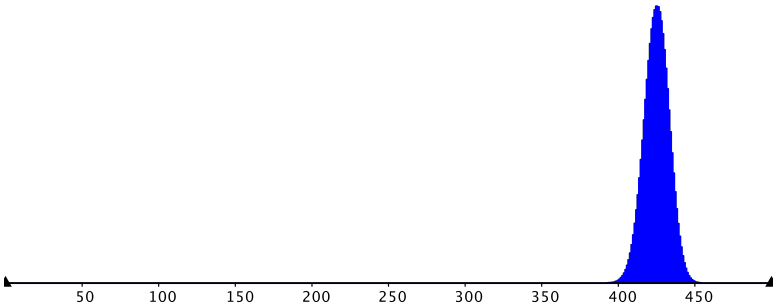

FIGURE 6.12: $B(500, 0.85)$

n is large enough, we can use a normal model to approximate the binomial distribution. (This was a useful approximation before the advent of powerful computing.) If n is large enough, we can use a normal model to approximate the binomial distribution, even when p is near 0 or 1.

How large must n be for the distribution to become unimodal and symmetric? Statisticians have agreed that when both the expected number of "successes" and the expected number of "failures" are both 10 or greater, that the normal approximation makes sense.

We can calculate the expected number of successes by multiplying the likelihood of success p by the number of trials n, so we check if $np \geq 10$. For failures, we check if $nq \geq 10$. Table 6.1 shows these calculations for Example 6.2.

Notice that when the $np \geq 10$ and $nq \geq 10$ thresholds are met between $n = 50$, when $np = 42.5$ and $nq = 7.5$, and $n = 100$, when $np = 85$ and $nq = 15$ the distributions look both unimodal and symmetric. Before

n	np	nq
1	$(1)(0.85) = 0.85$	$(1)(0.15) = 0.15$
5	$(5)(0.85) = 4.25$	$(5)(0.15) = 0.75$
20	$(20)(0.85) = 17$	$(20)(0.15) = 3$
50	$(50)(0.85) = 42.5$	$(50)(0.15) = 7.5$
100	$(100)(0.85) = 85$	$(100)(0.15) = 15$
500	$(500)(0.85) = 425$	$(500)(0.15) = 75$

TABLE 6.1

that point, say for $n = 20$, the distribution is visibly skewed, so a normal approximation would not make sense.

6.2 Homework

1. Consider a "rare event", such as $p = 0.15$.
   (a) Create histograms of the binomial distributions for $B(n, 0.15)$ using $n = 1$, $n = 5$, $n = 20$, $n = 50$, $n = 100$ and $n = 500$, respectively.
   (b) Describe the shape of each distribution.
   (c) As n increases, how the the shape change?
   (d) To use a normal approximation, what number of trials, n, is reasonable?
2. PCC's spam filter is expected to allow only 3 out of 100 spam email messages to make it to your inbox.
   (a) Use the binomial distribution to calculate the probability of having more than 6 spam emails out of 100 spam emails in your inbox.
   (b) What are the mean μ and standard deviation σ of the binomial distribution?
   (c) Use the normal distribution with μ and σ from (b) to calculate the probability from (a).
   (d) Does the normal distribution approximate the binomial distribution well here? Why or why not?
3. PCC's spam filter is expected to allow only 3% of spam email messages to make it to your inbox. What is the probability of having more than 35 spam emails if you were to receive 1000. Would you recommend to someone to use the binomial model or the normal approximation of the binomial? Provide a reason.
4. PCC's spam filter is expected to allow only 3% of spam email messages to make it to your inbox. Assume you can use the normal approximation of the binomial. What is the probability of having more than 35 spam emails if you were to receive 1000?

Chapter 9

Hypothesis Testing

9.1 Comparing information a confidence interval provides versus a significance test

Introductory Statistics Example 9.17

Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, she uses a 1% level of significance.

A hypothesis test is designed to test the likelihood of an assumption (the null hypothesis H_o in light of collected sample data. In Example 9.17, the test returned p-value 0.5485. Here, the p-value means that if the proportion of first-time brides in the U.S. who are younger than their partners is 50%, then the likelihood of observing the sample proportion, 53%or greater, (or of observing 47% or lower) would be 54.85%. See Figure (picture of p-value below). That's not unlikely, so these data do not throw doubt on the assumption that half of first-time brides are younger than their partners.

Another way to consider the likelihood of the assumption, H_o, is to construct a confidence interval. Using the same data, we can construct a 95% confidence interval to estimate the proportion of first-time brides who are younger than their partners. With $p' = 0.53$and $n = 100$, we have, $0.53 \pm 1.95\sqrt{\frac{(0.53)(0.47)}{100}}$ which yields the interval $(0.48, 0.58)$.

This says that, based on sample data, we estimate with confidence level 95% that the proportion of all first-time brides who are younger than their

partners is between 48% and 58%.

Like the hypothesis test, this confidence interval does not throw any doubt on the assumption, since the assumed value, $p_o = 0.50$, from the hypothesis test falls within the confidence interval.

Why have two methods here?

The subtle difference between the hypothesis test and the confidence interval is the standard deviation for the sampling model. For a hypothesis test, we calculate standard deviation using the assumed proportion, p_o and the formula $\sqrt{\frac{(p_0)(q_o)}{n}}$. But for a confidence interval, we calculate standard deviation using the sample proportion, p' and the formula $\sqrt{\frac{(p')(q')}{n}}$. When the assumption and the statistic do not differ much, these models are very similar.

9.2 Statistical Significance vs. Practical Value

Once a significance level, α, has been chosen, a hypothesis test's results may or may not have statistical significance. Sometimes, evidence is statistically significant, but that evidence is of little practical value.

Example 13. *The National Institutes of Health reports that approximately* 6%*of people suffer from Insomnia. A researcher tests a new surgery procedure with the hypothesis* H_o *: No change in Insomnia / amount of sleep and* H_a *: positive change in Insomnia / increased amount of sleep. They set a significance level of* 0.5% *and after their research concludes they obtain a p-value of* 0.1%*. They claim their evidence has statistical significance. The surgery is then brought into the private health marketplace at a cost of* \$50000*.*

While the research claims to have evidence of statistical significance, that evidence is of little practical value to most people since the cost for the surgery is \$50000*.*

Example 14. *A consumer organization is testing what type of fuel you should use for your car, unleaded or premium. They test the hypothesis* H_o*: no change in MPG and* H_a*: an increase in MPG. They set a significance level of* 1% *and after their testing concludes they obtain a p-value of* 0.5%*. They claim their evidence has statistical significance. They print their information in their next magazine as a reason to buy premium fuel.*

The organization claims to have evidence of statistical significance, but the average price for a gallon of unleaded fuel is approximately \$2.05 *and the average price for a gallon of premium fuel is approximately* \$2.50 *(as of March 2016, Source: AAA). This means that while they claim to have evidence of statistical significance that average mileage increases using more expensive fuel, we also know that using more expensive fuel makes your increased miles per gallon more expensive. This makes our statistical significance a little less practical.*

9.3 Homework

1. Introductory Statistics Example 9.18

   Suppose a consumer group suspects that the proportion of households that have three cell phones is 30%. A cell phone company has reason to believe that the proportion is not 30%. Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three cell phones. Using the data in this question, the hypothesis test return a p-value 0.7216. That?s not unlikely, so these data do not throw doubt on the assumption that half of first-time brides are younger than their partners.

   Construct a confidence interval to consider the likelihood of the assumption, H_o.

2. Introductory Statistics 9.5 Homework 100

   Recall that question 100 from Homework section 9.5, "According to an article in Bloomberg Businessweek, New York City's most recent adult smoking rate is 14%. Suppose that a survey is conducted to determine this year?s rate. Nine out of 70 randomly chosen N.Y. City residents reply that they smoke. Conduct a hypothesis test to determine if the rate is still 14% or if it has decreased."

   After conducting a hypothesis test, construct a confidence interval to consider the likelihood of the assumption, H_o.

3. You take a 9:00am bus to go to your 10:00am class and are worried about arriving on time or if you need to take an earlier bus. You decide

to test the hypothesis $H_o : \mu = 50$ minutes and $H_a : \mu > 50$ minutes, in regards to how long the bus ride will take. You collect data by riding the bus 13 times, shown below in Figure 1 and preliminary statistics, shown in Figure 2. You then do a hypothesis test and find a p-value of 0.0016, shown in figure 3, thus you declare statistical significance and reject the null hypothesis that the average bus ride time is 50 minutes. You then create a confidence interval based on your data, shown in figure 4, which shows a plausible range for the true average time on the bus. This range is from 51.17 minutes to 54.69 minutes. While the data has statistical significance, what is the practical meaning?