1 Experiments and Observational Studies

A *response variable* measures an outcome of a study.

An *explanatory variable* explains or influences changes in a response variable.

Treatment group, control group, randomly assign, double-blinded, single-blinded, placebo, placebo effect.

We need a control group to cancel out the effect of the placebo effect.

In an experiment, the researcher chooses which group a subject will be in, the treatment group or control group.

In an observational study, the subjects choose, or are naturally in, either the treatment group or control group.

While an observational study can show an association, like smoking is associated with lung cancer, the study cannot show causation. To show causation, you need an experiment.

You can't always do an experiment.

- It might be unethical to assign groups. Does marijuana cause birth defects? It's not ethical to tell a pregnant women to use marijuana.
- It might not be possible to assign groups if it is genetic. Are bald men more likely to have heart problem? (Yes.) You can't assign a man to be bald; it's a genetic trait.

Confounding (lurking) variable: an inherent variable that affects the responses being studied. Oftentimes hidden.

Ex: Testing a drug on children. What is the problem with choosing 7 year olds for the treatment groups and 8 year olds for the control group when testing a new drug? What if a person's sex determines what group he or she will be in?

Historical Control: A control group that is chosen from a group of patients who were observed at some time in the past or for whom data are available through records. Historical controls are used for comparison with subjects being treated or assessed concurrently.

Definition. Bias: systematic introduction of error in a measurement. May be intentional or not.

Randomization reduces the effect of bias and also reduces the effect from confounding variables.

- Selection bias: The selection procedure excludes one or more type of people.

  Ex: Taking a door-to-door survey in the middle of the day will exclude people who leave home to work durning the day.

  Ex: Using a telephone book to randomly call people excludes people who have unlisted numbers.

- Nonresponse bias: Nonresponse bias can occur when many people who are selected for the sample either do not respond at all or do not respond to some of the key survey questions.

  Ex: Self-selected sample or a volunteer sample. To prepare for her book *Women and Love*, Shere Hite sent questionnaires to 100,000 women asking about love, sex, and relationships. 4.5% responded and Hite used those responses to write her book.

  Respondents "were fed up with men and eager to fight them"

  "The anger became the theme of the book"

  "But angry women are more likely" to respond

- Response Bias: Respondents may give socially acceptable answers (maybe not the truth!)

  Ex: A door-to-door survey asking if the person uses marijuana. Would a person answer differently if a police office were conducting the survey?

  Ex: People are asked a question about their thoughts on something they don't know about, and sometimes they pretend that they have an opinion on it. "How do you feel about the Bulls losing last night?" No such game happened last night. A man might say he felt sad about the Bulls losing and that the second half was way more exciting than the first half.

- Convenience or haphazard sample: Choosing the most convenient group available or to decide haphazardly on the spot whom to sample.

  Ex: For a final project, a Bob asks college students who were relaxing in a university quad (wide expanse of lawn) in the middle of spring day about study habits. Bob's final report states that students on campus don't have very many worries and their classes are not hard.

  Ex: Jill is at the mall and needs to ask people questions for a survey. She is a bit timid and only asks people who are well dressed and friendly looking. She does not approach mean looking people. This might affect her results.

Simpsons Paradox. The paradox is that the relationship appears to be in a different direction when the confounding variable is not considered than when the data are separated into the categories of the confounding variable.

Ex: Kidney Stone Treatment. This is a real-life example from a medical study comparing the success rates of two treatments for kidney stones. The table shows the success rates and numbers of treatments for treatments involving both small and large kidney stones, where Treatment A includes all open procedures and Treatment B is percutaneous nephrolithotomy.

	Treatment A	**Treatment B**
Small Stones	Group 1 **93% (81/87)**	Group 2 87% (234/270)
Large Stones	Group 3 **73% (192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Treatment A is more effective when used on small stones, and also when used on large stones.

Yet treatment B is more effective when considering both sizes at the same time.

In this example the confounding variable of the stone size was not previously known to be important until its effects were included.

Ex: Read about sex bias in graduate admissions on pages 17 to 20. Here, University of California, Berkeley was sued for bias against women who had applied for admission to graduate schools there.

	Applicants	**Admitted**
Men	8442	**44%**
Women	4321	35%

When broken into the different departments:

Department	**Men**		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

Conclusion: Women tended to apply to competitive departments with low rates of admission even among qualified applicants, whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants.

2 Histogram and Centers

Definition: A *variable* is a characteristic that changes from subject to subject.

Two types:

- *Quantitative*, which involve numbers. You can do arithmetic with these variables. Ex: height, weight, fastest running speed, number of steps to your front door, ...
- Qualitative or *categorical*, which places a subject into one of several groups or categories. Ex: birth month, favorite color, year in college, ...

Sometimes a variable can be both quantitative and qualitative depending on context. Ex: Shoe size or Age (in whole years).

Quantitative come in two flavors:

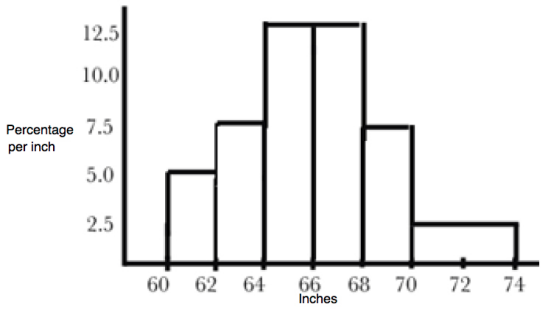
- Discrete: the variables can only differ by fixed amounts. Ex: Number of people in a family, number of steps to your front door. These differ by 0, 1, 2, ...
- Continuous: These variables can take on a range of values. Ex: Height, weight, temperature.

We will concentrate on quantitative variables.

Definition: A *histogram* is a pictorial representation of data used to summarize the data.

There are two types of histograms, one shows frequency and the other shows percentage. The book talks about percentage histograms.

Ex: Consider heights of a group of women (in inches) (Rounding errors):



The intervals (60-62, 62-64, 64-66, 66-68, 68-70, 70-74) are called *class intervals*. Note that they need not be the same width.

Area of rectangle over a class interval represent % of data in that class interval. Hence, total area of a histogram must be 100%.

Ex: What percentage of women have an height less than 64 inches? we calculate $2 \cdot 5.0 + 2 \cdot 7.5 = 25$. So, about 25%.

Ex:

192	110	195	180	170	215
152	120	170	130	130	125
135	185	120	155	101	194
110	165	185	220	180	
128	212	175	140	187	
180	119	203	157	148	
260	165	185	150	106	
170	210	123	172	180	
165	186	139	175	127	
150	100	106	133	124	

We have 53 numbers that range from 100 to 260. We make the intervals (with rounding errors)

Class Intervals	Frequency	Percentage
$100 \leq x < 120$	7	13%
$120 \leq x < 140$	12	23%
$140 \leq x < 160$	7	13%
$160 \leq x < 180$	8	15%
$180 \leq x < 200$	12	23%
$200 \leq x < 265$	6	11%

Notice, that when choosing our intervals, we include the left endpoint and not the right endpoint. You could have chosen to include the right endpoint and not the left endpoint.

We have to adjust the heights of the blocks in our histogram since our class intervals are not of width 1. For each of the first 5 class intervals, we divide the percentage by 20. For example, the first interval will have a height of $13/20 = 0.65$. The last block will have a height of $11/80 = .17$. So, our heights will be

$$0.65 \qquad 1.15 \qquad 0.65 \qquad 0.75 \qquad 1.15 \qquad 0.17$$

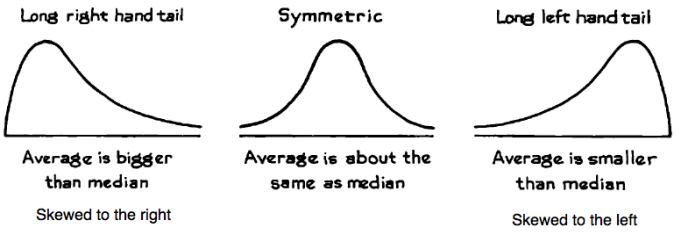
With these heights, you will get the correct percentage when you find the areas of the blocks.

(Picture of Histogram.)

Define mean and median. Examples. Do example of median with odd and even numbers.

Pictures of what median and mean represent on a histogram. The median divides the area of the histogram in half. The mean is the "balancing point" of the histogram. For a symmetric histogram, the mean and medium will be the same.

Figure 7. The tails of a histogram.



If the mean is greater than the median, then the histogram stretches out to the right.

If the mean is less than the median, then the histogram stretches out to the left.

Ex: In some city, for single-family homes, we have

$$\text{median} = \$136,000 \qquad \text{and} \qquad \text{mean} = \$149,000$$

Which do you think is more useful to someone considering the purchase of a home, the median or the mean?

$$\text{average of } -10, -5, 0, 5, 10 \text{ is } 0$$
$$\text{average of } -100, -50, 0, 50, 100 \text{ is } 0$$

The second set of numbers are more spread out that the first set, even though the mean of both sets is 0.

In most statistical studies, the objective is to use a small group of units to make an inference (or answer questions) about a larger group. The larger group of units about which inferences are to be made is called the *population*. The smaller group of units actually measured is called the *sample*.

Ex: All the students at Colgate University is the population, and the 120 students that take a short survey is the sample.

Ex: The people in Maine is the population, and the 100,000 people who are are called for a survey are the sample.

Is there a measurement for the spread of data from the mean? This measurement of spread is called the *standard deviation*.

There are two calculations for the standard deviation. Which calculation you use depends on whether your data values represent the entire population or whether your data values are from a sample of a larger population.

Below, let a denote the average of all your data values $x_1, x_2, \ldots, x_n$, and you have n data values.

- Definition. The *population standard deviation* is given by

$$SD = \sqrt{\frac{(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2}{n}}$$

You use this when you have a data value for each person in your population. Or, use this when you have a sample of a larger population, but you are only interested in this sample and do not wish to generalize your findings to the population. (Some books use σ, the greek letter sigma, in place of SD.)

An alternative form the book mentions is

$$SD = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - a^2}$$

- Definition. The *sample standard deviation* is given by

$$SD^+ = \sqrt{\frac{(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2}{n - 1}}$$

If all you have is a sample and you wish to make a statement about the population standard deviation from which the sample is drawn, then you need to use the sample standard deviation. (Some books use s in place of SD^+.)

The only small difference between these two formulas is the denominator; one formula has n while the other has $n - 1$. For right now, we will concentrate on using the formula for SD since this is what the book uses.

There is a conversion between the two:

$$SD^+ = \left(\sqrt{\frac{n}{n-1}}\right) SD \qquad \text{and} \qquad SD = \left(\sqrt{\frac{n-1}{n}}\right) SD^+$$

Ex: Find SD for the numbers 8, 11, 14, 15, 16. The average of these numbers is $a = 12.8$. The standard deviation is $SD = \sqrt{8.56} \approx 2.92$.

Rule of Thumb (The Empirical Rule): True in general, but not always.

- About 68% of data values will fall within 1 SD of the mean.
- About 95% of data values will fall within 2 SD of the mean.
- About 99.7% of data values will fall within 3 SD of the mean.

3 Measurement Error and Scaling

Read Chapter 5 for full details.

Weighting a 10 gram weight (of something called NB 10) 100 times, each trial yields a measurement of the form 9.999???. Below is a table from the book on page 99 showing the different ??? that showed up.

No.	*Result*	*No.*	*Result*	*No.*	*Result*	*No.*	*Result*
1	409	26	397	51	404	76	404
2	400	27	407	52	406	77	401
3	406	28	401	53	407	78	404
4	399	29	399	54	405	79	408
5	402	30	401	55	411	80	406
6	406	31	403	56	410	81	408
7	401	32	400	57	410	82	406
8	403	33	410	58	410	83	401
9	401	34	401	59	401	84	412
10	403	35	407	60	402	85	393
11	398	36	423	61	404	86	437
12	403	37	406	62	405	87	418
13	407	38	406	63	392	88	415
14	402	39	402	64	407	89	404
15	401	40	405	65	406	90	401
16	399	41	405	66	404	91	401
17	400	42	409	67	403	92	407
18	401	43	399	68	408	93	412
19	405	44	402	69	404	94	375
20	402	45	407	70	407	95	409
21	408	46	406	71	412	96	406
22	399	47	413	72	406	97	398
23	399	48	409	73	409	98	406
24	402	49	404	74	400	99	403
25	399	50	402	75	408	100	404

Even though care was taken each time, a slightly different measurement was recorded. There is a chance error with each measurement. The average of these numbers is 405 and the SD is 6. The SD gives us an estimate of how much chance measurement we might encounter.

Although we don't know the exact value of the weight, the average is probably a good estimate of the exact value. Each time we take a measurement, we have

$$\text{individual measurement} = \text{exact value} + \text{chance value} \approx AV + k \cdot SD$$

Scaling and its effects on the mean and SD:

Adding c to all data values:

$$\text{new avarage} = \text{old avarage} + c \qquad \text{new } SD = \text{old } SD$$

Subtracting c to all data values:

$$\text{new avarage} = \text{old avarage} - c \qquad \text{new } SD = \text{old } SD$$

Multiplying all data values by d:

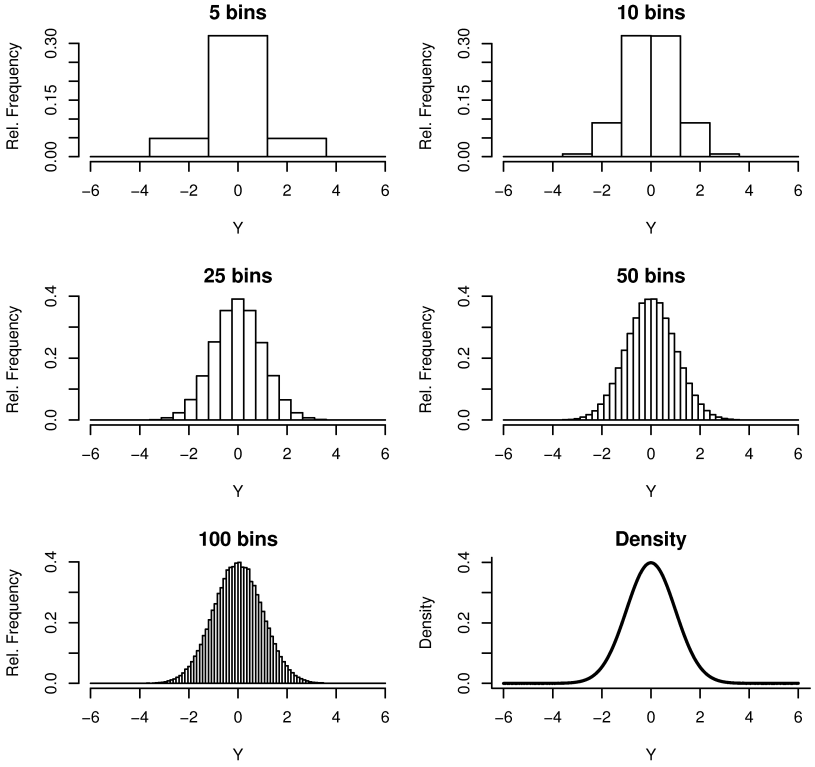
$$\text{new avarage} = d \cdot (\text{old avarage}) \qquad \text{new } SD = |d| \cdot (\text{old } SD)$$

Divide all data values by d ($d \neq 0$):

$$\text{new avarage} = \frac{1}{d} \cdot (\text{old avarage}) \qquad \text{new } SD = \left|\frac{1}{d}\right| \cdot (\text{old } SD)$$

4 The Normal Approximation for Data

Sometimes the overall pattern of a large number of observations is so regular, that we can describe it by a smooth curve, called a *density curve*.



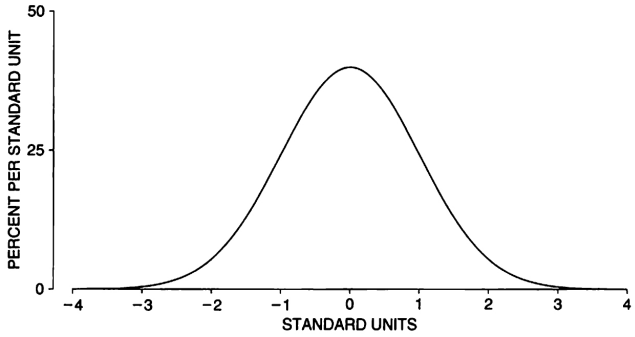
A *density curve* is a curve that

- is always on or above the horizontal axis
- has area equal to exactly 1 underneath it

We can draw a density curve over a histogram to make a density curve. In this case, the density curve becomes a mathematical model for the histogram.
(Draw picture...)

One of the most useful density curves we will use is the *(standard) normal curve*. In the figure below, the tail ends keep getting closer to the axis, but never touch it. Also, the graph is symmetric about 0. Associated with this curve is a mean of 0 and a SD or 1.

The normal curve.



The Empirical Rule applies for this curve:

- About 68% of data values will between -1 and 1.
- About 95% of data values will between -2 and 2.
- About 99.7% of data values will between -3 and 3.

Many histogram have a shape that is similar to normal. In this case, we say that the data is *normally distributive*. Making the horizontal scales match up involves *standard units*.

A value is converted to standard units by seeing how many SDs it is above or below the average. Values above the average are given a plus sign; values below the average are given a minus sign.

If your data has an average of AV and a standard deviation of SD, the conversion formula to convert your data score into the standard units is given by

$$z = \frac{\text{score} - AV}{SD}.$$

Ex: Suppose that we know that adult men have an average height of 70 inches and a SD of 2.8 inches.

1. What is Jack's height of 67.2 inches in standard units?

   We see that $67.2 = 70 - 2.8$, which says that 67.2 is 1 SD below the average, written -1 SD.

2. What is John's height of 75.6 inches in standard units?

   We see that
   $$z = \frac{\text{score} - AV}{SD} = \frac{75.6 - 70}{2.8} = \frac{5.6}{2.8} = 2,$$
   so that 75.6 is 2 SD above the average.

3. Suppose Bob has a height of 68 inches. What is his height in standard units?

We see that

$$z = \frac{\text{score} - AV}{SD} = \frac{68 - 70}{2.8} = \frac{-2}{2.8} \approx -0.714,$$

so that 68 is 0.714 SD below the average.

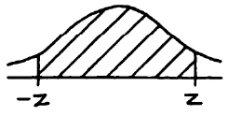
4. What height is 1.5 SD above the average?

   We solve $1.5 = \dfrac{\text{score} - 70}{2.8}$ for "score."

   $$1.5 = \frac{\text{score} - 70}{2.8} \quad \Rightarrow \quad 4.2 = \text{score} - 70 \quad \Rightarrow \quad 74.2 = \text{score}$$

   So, 74.2 inches is our the score that is 1.5 SD above the average.

Now for finding areas under the standard normal curve, with mean=0 and $SD = 1$. The table in the back of the book finds the area below when you know the value of z.



Read pages 82 through 84.

Ex: Find area under curve between $z = -1.45$ and $z = 1.45$. (Between a negative and positive version of the same number.)

Ex: Find area under curve above than $z = 0.25$. (Greater than a number.)

Ex: Find area under curve less than $z = 1.45$. (Less than a number.)

Ex: Find area under curve between $z = -1$ and $z = 2.30$. (Between a negative number and a positive number.)

Ex: Find area under curve between $z = 0.35$ and $z = 2.30$. (Between two positive numbers.)

We can find areas for arbitrary bell curves, we just have to convert to standard units.

Ex: Suppose the lengths of a certain type of adult lizard species have an average of 6.4 inches and a SD of .3 inches. Use the normal curve to estimate the percentage the lengths shorter than 5.5 inches.

With $z = \frac{\text{score}-AV}{SD} = \frac{5.5-6.4}{0.3} = -3$, we find the area under the curve less than $z = -0.30$.

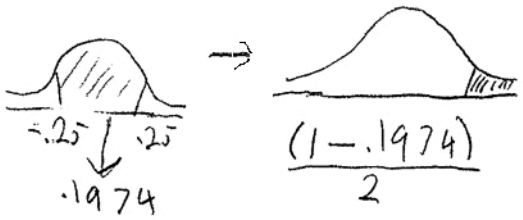
Note that the table in the back of the book is not a complete table. There are tables that have more entries and are more complete. If you find that the standard unit z you calculate is not in the table, you can estimate what you think it should be. Some graphing calculators can also find the areas for you.

Forgive the bad quality of the next page. I did it in pencil before I scanned it.

$z = -1.45$ to $z = 1.45$ is 85.29%

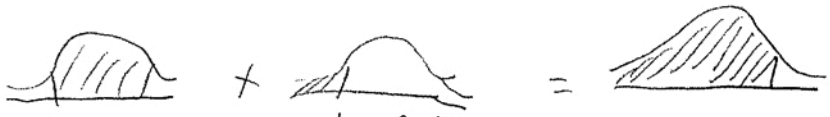
.8529

$z = .25$



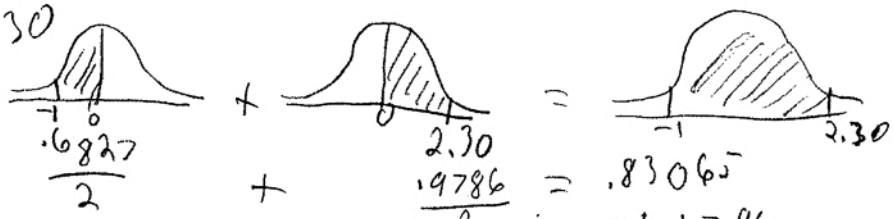
0.4013
40.13%

$z = 1.45$



$.8529 + \frac{1-.8529}{2} = .92645 \approx 92.65\%$

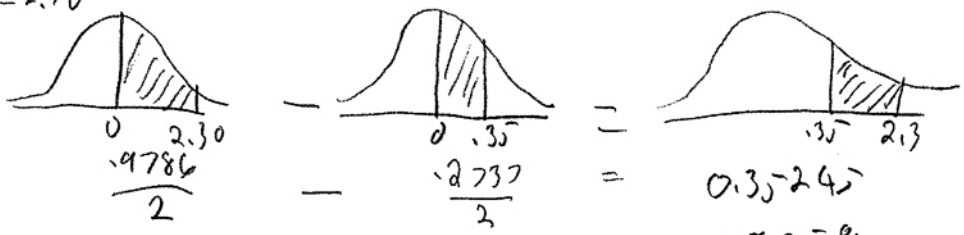
$z = -1 + 0.7 = 2.30$



$+ \frac{.9786}{2} = .8306$

83.17%

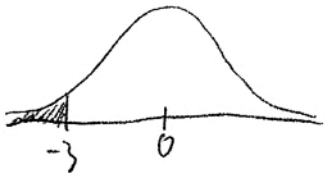
$z = .35$ and $z = 2.30$



$0.35 - 24$

35.25%

$z = -3$



$\frac{100\% - 99.73\%}{2} = 0.135\%$

5 Percentiles

The kth *percentile* is a number that has $k\%$ of the data values at or below it and $(100-k)\%$ of the data values at or above it.

The 25th percentile is called the *lower quartile*.

The 25th percentile is called the *median*.

The 75th percentile is called the *upper quartile*.

Ex: If you are told that you scored at the 90th percentile on a standardized test (such as the SAT), it indicates that 90% of the scores were at or below your score, while 10% were at or above your score.

Quartiles are useful for data that is skewed: data with long tails. For example, one might use quartiles when discussing incomes since the income scale is skewed to the right. If you only used AV and SD, then the normal approximation using the standard deviation might suggest negative incomes.

6 Lines

Read Chapter 7. Period.

y-intercept form.

$$y = mx + b, \qquad m = \text{slope} \qquad b = y\text{-intercept}$$

Point-slope form. Given two points (x_1, y_1) and (x_2, y_2), the point-slope form of the line is

$$y - y_1 = m\,(x - x_1)\,, \qquad \text{where} \qquad m = \frac{y_2 - y_1}{x_2 - x_1}$$

(Leave x and y as the variables x and y; don't plug numbers into them.

Ex: Find the y-intercept form of the line through the points (1,5) and $(-1, -1)$.

Here, $m = \frac{5-(-1)}{1-(-1)} = 3$ and $y - 5 = 3(x - 1)$ simplifies to $y = 3x + 2$.

7 Correlation and Regression

Sometimes, data comes in pairs (x, y), where (usually) x is the explanatory variable and y is the response variable. With paired data like this, we can plot

the data on the xy-plane, called a *scatterplot*. For example,

Height (in.)	Span (cm)
71	23.5
69	22.0
66	18.5
64	20.5
71	21.0
72	24.0
67	19.5
65	20.5
76	24.5
67	20.0
70	23.0
62	17.0

and so on, for $n = 167$ observations.

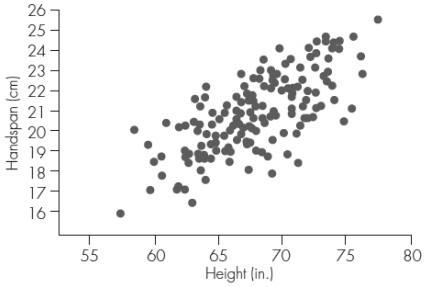
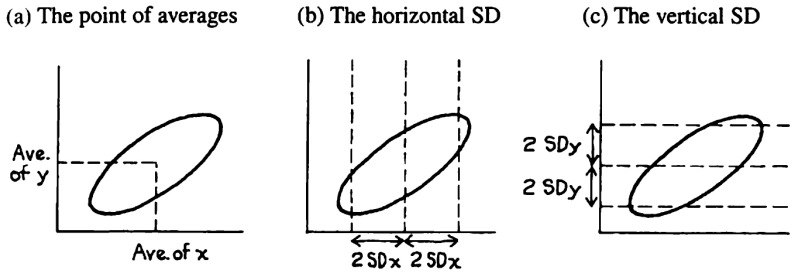


Figure 5.1 ▮ Height and handspan

It would be nice if knowing the value of one variable, x, could help us predict the value of the other variable, y.

We can use the Ave of x, Ave of y, SD of x, and SD of y to explain the area where the points are.

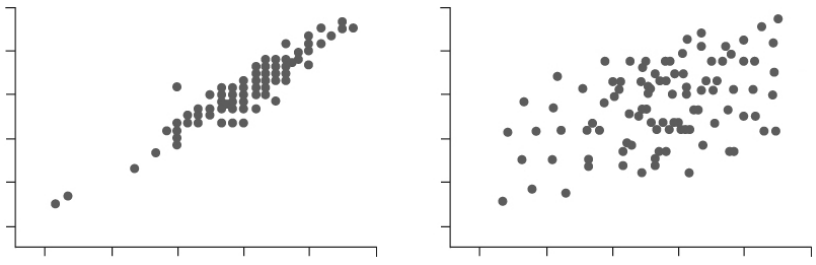
Figure 4. Summarizing a scatter diagram.



This football shape gives us an idea of where the points are. The problem is that we need another variable to explain the spread of the points. We would like some way to measure a (linear) relationship between the two variables.

Definition: *Correlation*, a statistic that measures the strength and direction of a linear relationship between two quantitative variables.

Consider the scatterplots



Both scatterplots have the same spread in both the x and y directions. The scatterplot on the left in the above picture is close to a linear relationship, and thus it has a stronger correlation. The scatterplot on the right is not close to a linear relationship, and thus it does not have a strong correlation.

A well know saying is *correlation is not causation*. This means that you might have a strong correlation (the points are close to a line), but that does *not* imply that a change in the x variable is *causing* a change in the y variable.

We want a numerical value that tells us how close to a line our points lie.

Definition: Correlation coefficient r: Two formulas

$$r = \text{average of } (x \text{ in standard units}) \times (y \text{ in standard units})$$

$$= \frac{(\text{average of } x \cdot y) - (\text{average of } x) \cdot (\text{average of } y)}{(SD \text{ of } x) \cdot (SD \text{ of } y)}$$

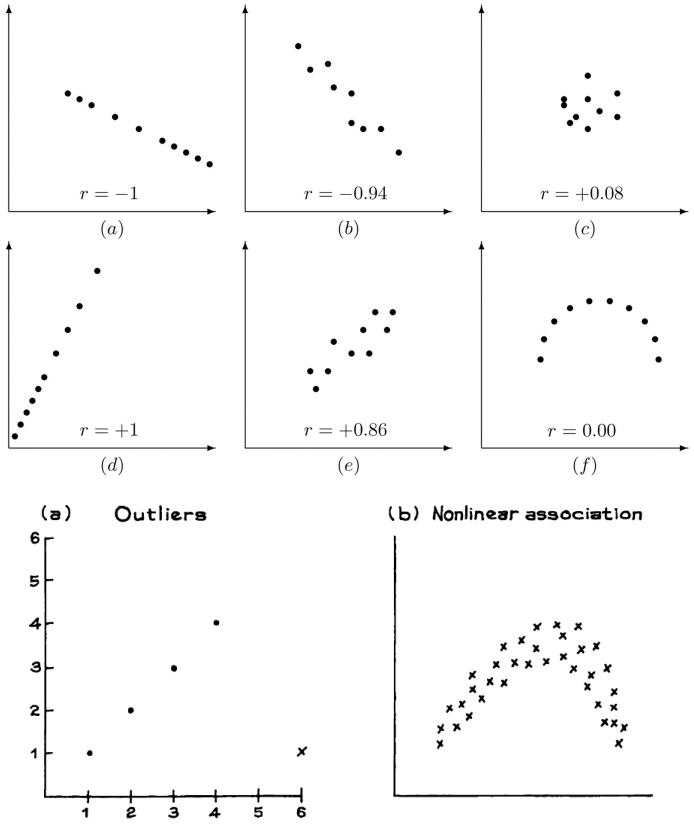
First, the properties of r:

- $-1 \leq r \leq 1$.
- If r is close to -1, then the points are close to a line that has a negative slope.
- If r is close to 1, then the points are close to a line that has a positive slope.
- If r is not close to 1 or -1, then the points are not close to a line. The points could be a cloud of points, or the points could follow another patten like a parabola.
- The correlation coefficient r does not have any units (i.e. like inches or centimeters); it is just a number. The value of r is not affected by
  - interchanging the two variables.
  - adding the same numbers to all the values of one variable.
  - multiplying all the values of one variable by the same positive number.

Ex:

	x	y	$x \cdot y$
	1	6	6
	3	7	21
	4	10	40
	5	11	55
	8	15	120
Average:	4.2	9.8	48.4
SD :	2.3152	3.1875	n/a

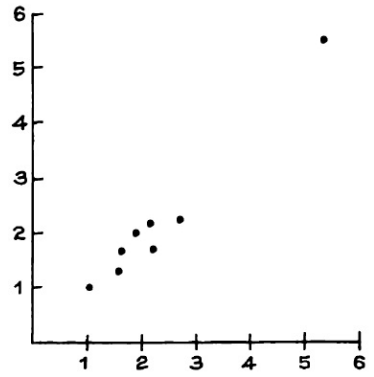
$$
\begin{aligned}
r &= \frac{(\text{average of } x \cdot y) - (\text{average of } x) \cdot (\text{average of } y)}{(SD \text{ of } x) \cdot (SD \text{ of } y)} \\
&= \frac{48.4 - 4.2 \cdot 9.8}{2.3152 \cdot 3.1875} = 0.9811
\end{aligned}
$$



The two scatterplots above have r close to zero. Figure (a) is because of the outlier. If you find good reason to remove outliers, do so. Figure (b) – and Figure (f) – is not linear-shaped data. Even though there is a trend with the data, it is not a linear trend; r measures linear association.

On the other hand, the scatter plot below has r close to 1 since the outlier is in line with the rest of the points but still a distance away. In this case, you

want to check if the outlier is a still a viable data value.



It is good practice to look at the scatterplot before crunching the numbers to find r or any of the lines we will discuss below.

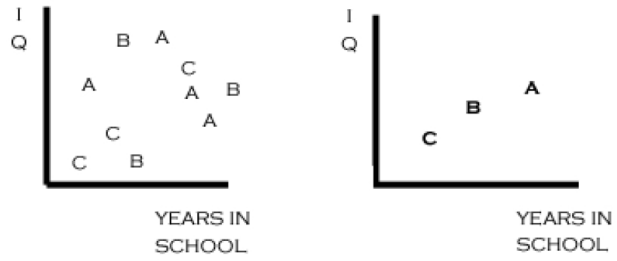
Ecological Correlations Are based on rates (e.g., averages). Used often in Poly Sci and Sociology.

They tend to overstate the strength of an association.

Reason: When we average we get rid of variations.

Ex: Consider IQ versus years in school, separated by age group:

A = 30 year olds; B = 40 year olds; C = 50 year olds.



Individuals After Averaging

Def: The scatterplot generally clusters around the *SD line*, but does not, in general, give the line of best correlation.

It goes through the ends of the football. It goes through the points (a_x, a_y) with slope $m = \pm\frac{SD_y}{SD_x}$, where the $\pm$ depends on $r > 0$ or $r < 0$. (Here, a_x and a_y are the average of the x's and y's, resp.)

The equation of the SD line is given by

$$y = \pm\frac{SD_y}{SD_x}(x - a_x) + a_y.$$

This is not the "best fit line," only a general trend line. *It is not used for predictions*, but only for location purposes on a scatterplot.

What is the best fit line? The *regression line* (or *least squares line*) is given by

$$y = r\frac{SD_y}{SD_x}(x - a_x) + a_y.$$

Note that this line passes through the point (a_x, a_y). For each increase (resp. decrease) of SD_x in x, then is an $r \cdot SD_y$ increase (resp. decrease) in y.

Ex: If x represents height in inches and y represents weight in pounds, suppose

$$\begin{array}{lll} a_x = 70 & SD_x = 3 & \\ a_y = 162 & SD_y = 30 & r = .47 \end{array}$$

If a man is chosen at random, what would you guess for his weight? Choose the average of 162, since you have no information about his height.

If you know his height is 73 inches, then what would you guess for his weight? The average weight of all 73 inch tall people in the study. We use the regression line to find this answer.

$$\text{answer} = y = .47 \cdot \frac{30}{3}((73) - 70) + 162 = 176.1$$

Since 73 is 1 $SD_x = 3$ above $a_x = 70$, we see that 176.1 is $r \cdot SD_y = .47 \cdot 30 = 14.1$ above $a_y = 162$.

Ex: We study IQ versus Math SAT score. Our group has an average IQ of 100 with a SD of 15, which obtaining an average SAT of 550 with SD of 80. We calculated the correlation coefficient to be r = .6 and found that the scatterplot was football-shaped.

If a student scores a 150 on the IQ test, what do you estimate for their SAT score?

We want to predict SAT, so we let y =SAT. Hence, the regression line is:

$$\text{answer} = y = .6 \cdot \frac{80}{15}((150) - 100) + 550 = 710$$

Plugging in $x = 150$ gives us $y = 710$, our prediction.

The regression effect: In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test – and the top group will on average fall back.

Thinking that the regression effect must be due to something important, not just the spread around the line, is the *regression fallacy.*

Ex: Preschool pre-test and post-test.

There are two regression lines:

If predicting y using x (regression line for y on x):

$$y = r\frac{SD_y}{SD_x}(x - a_x) + a_y$$

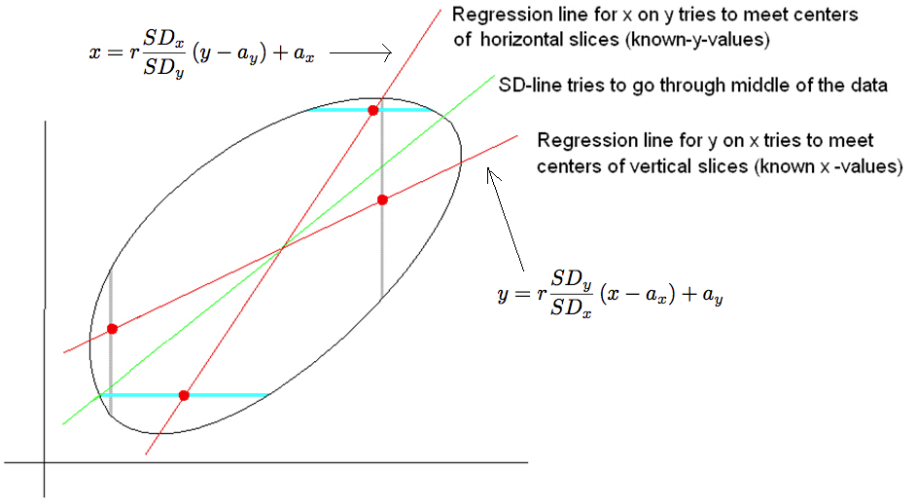
Here, x is you explanatory variable and y is you response variable.

If predicting x using y (regression line for x on y):

$$x = r\frac{SD_x}{SD_y}(y - a_y) + a_x$$

Here, y is you explanatory variable and x is you response variable.

Sometime the two variables can take on both rolls. For example, height vs. weight. Either can play the roll of the explanatory or response variable.



8 Probability

Definition: *Chance*: The percentage of time that something is expected to occur when the process is repeated many times.

If something cannot occur, its associated chance is 0%

If something always occurs, its associated chance is 100%.

Hence, chances are always between 0% and 100%.

Probability is chance as a decimal, so probabilities are between 0 and 1.

Ex: Watching male birth rates over a period of a year at a hospital.

Weeks of Watching	Total Births	Total Boys	Proportion of Boys
1	30	19	$19/30 \approx .633$
4	116	68	$68/116 \approx .586$
13	317	172	$172/317 \approx .543$
26	623	383	$383/623 \approx .615$
39	919	483	$483/919 \approx .526$
52	1237	639	$639/1237 \approx .517$

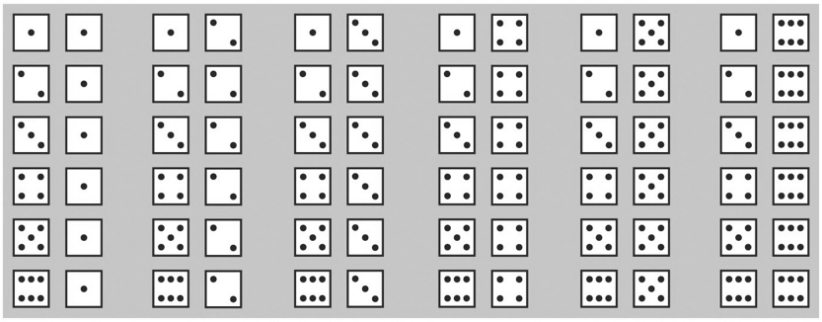
After watching this hospital for 52 weeks, the probability of having a boy is .517.

Ex: How do they get the probability for rain for an up coming day? They look at all the days in the past with recoded data that have similar conditions and look at the percentage of those days that had rain.

The *sample space* S of a random phenomenon is the set of all possible outcomes.

An *event* is an outcome or a set of outcomes (subset of the sample space).

Ex: Rolling two dice: list the possible combinations of dots on the top of the dice.



Let E be an event, and let $P(E)$ denote the probability of E. To calculate a probability E,

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in } S}$$

So, each outcome in the sample space above has a probability of $\frac{1}{36}$.

Ex: What is the probability of the event E = the sum of the dots is 5?

There are 4 outcomes where the dots add up to 5. So, the probability is $P(E) = \frac{4}{36}$.

Let E and F be two events. There are three rule we will use.

- The NOT Rule: $P(E) = 1 - P(\text{NOT } E)$
- The Multiplication Rule: $P(E \text{ and } F) = P(E) \times P(F|E)$
- The Addition Rule: $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$

The NOT Rule: No matter what the event E is, "Either event E happens or event E does not happen." Ergo,

$$P(E) + P(\text{NOT } E) = 1$$

Solving for $P(E)$ yields

$$P(E) = 1 - P(\text{NOT } E).$$

This can make the calculation for probabilities easier.

Ex: What is the probability that the sum of dots on the two dice is ≥ 4?

Let E be the event where the sum of the dots is ≥ 4. Instead of adding up all the outcomes where the sum of the dots is ≥ 4, we can use the NOT Rule with the probability for the sum of the dots < 4. There are only three outcomes where the sum of the dots are < 4. Thus,

$$P(E) = 1 - P(\text{NOT } E) = 1 - \frac{3}{36} = 0.917$$

Conditional Probability: Probabilities can change if we know additional information. We want to find the probability of an event E happing knowing that the event F happened. In math, these look like: $P(E|F)$ and we say "the probability of E given F."

Let's see if we can reason out a formula for this. Since we know that F has occurred, the possible outcomes are the number of outcomes in F. And we want: of those outcomes in F, how many of them are in E. In other words, how many are in F and E. Thus,

$$P(E|F) = \frac{\text{number of outcomes in } E \text{ AND } F}{\text{numbers of outcomes in } F}$$

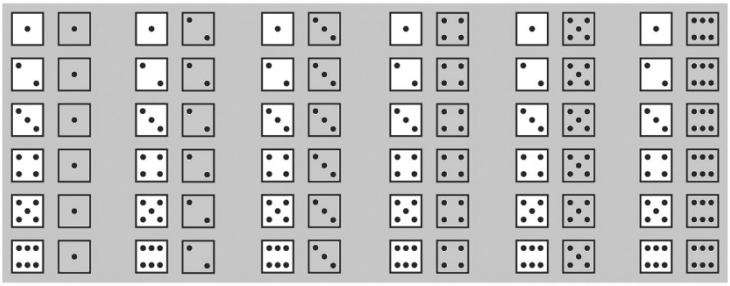
Dividing both the numerator and denominator by the total number of outcomes in the S, the sample space, gives

$$P(E|F) = \frac{(\text{number of outcomes in } E \text{ AND } F) \,/\, (\text{total number of outcomes in } S)}{(\text{numbers of outcomes in } F) \,/\, (\text{total number of outcomes in } S)}$$

which is simply written

$$P(E|F) = \frac{P\,(E \text{ and } F)}{P(F)}$$

Ex: While rolling two dice, what is the probability that the second die rolls a 2 given that the first die is a 4?



Let $E =$ "second die is a 2" and $F =$ "the first die is a 4." Using the sample space, we see that

$$\begin{aligned} P(F) &= P\,(\text{first die is 4}) = \frac{6}{36} \\ P\,(E \text{ and } F) &= P\,(\text{second die is 2 and first die is 4}) = \frac{1}{36} \end{aligned}$$

Thus, we have

$$P(E|F) = \frac{P\,(E \text{ and } F)}{P(F)} = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{6}$$

Multiplication Rule: We use this rule when both E *AND* F occur. Using the previous equation, we have *the multiplication rule*

$$P\,(E \text{ and } F) = P(E) \times P(F|E)$$

We have a special case of the multiplication rule when we consider the following definition.

Definition:

- Two events are *independent* of each other if knowing that one will occur (or has occurred) does not change the probability that the other occurs. In math notation,

$$P(F|E) = P(F) \qquad (\text{and hence, } P(E|F) = P(E))$$

- Two events are *dependent* if knowing that one will occur (or has occurred) changes the probability that the other occurs. In this case,

$$P(F|E) \neq P(F)$$

Ex: (Independence) Roll a fair die and flip a fair coin. What is the probability that your roll one dot and flip a head?

Let $E =$ "roll one dot" and $F =$ "flip a head." Note here that the outcome of either of these events does not affect the other. So, these events are independent of each other. Then the probability is

$$P(E \text{ and } F) = P(E) \times P(F|E) = P(E) \times P(F) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

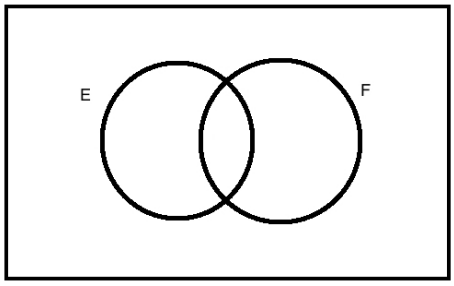
Ex: (Dependence) Shuffle a deck of 52 cards. What is the probability that the first two cards are aces?

Let $E =$ "the first card is an ace" and $F =$ "the second card is an ace." We are not replacing the first card back in the deck before we draw the second card; this makes the events dependent. The probability is

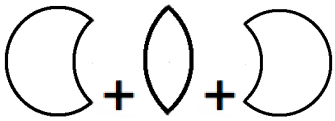
$$P(E \text{ and } F) = P(E) \times P(F|E) = \frac{4}{52} \times \frac{3}{51} = ??$$

(Note: Since E and F are dependent, $P(F|E) = \frac{3}{51} \neq P(F) = \frac{4}{52}$.)

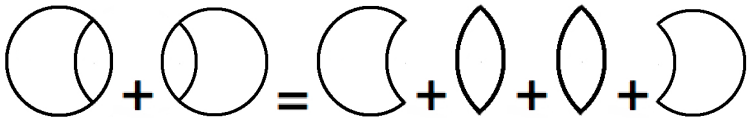
The Addition Rule: Suppose we want to find the probability of E or F (possible both). We denote this probability $P(E \text{ or } F)$. Consider the picture below.



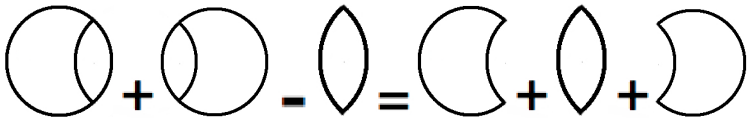
The rectangle represents all possible outcomes. The left circle represents event E, and the right circle represents event F. The circles together represent E or F. The sliver in the middle represents E and F. These can be broken down into the parts



Notice that



Subtracting the  from both sides, we get



This is written as

$$P\,(E \text{ or } F) = P(E) + P(F) - P\,(E \text{ and } F)$$

Definition: We say that two events E and F are *mutually exclusive* if the occurrence of one prevents the other from happening. In other words, E and F cannot both occur at the same time. In math notation, this means $P\,(E \text{ and } F) = 0$.

In this case, the addition rule reduces to

$$P\,(E \text{ or } F) = P(E) + P(F)$$

Note: Don't make the mistake of using this rule when events are not mutually exclusive. Your resulting answer will be wrong.

Ex: What is the probability that, when you roll two dice, both dice roll an even number or the sum of the dots on the dice is odd?

Let event E = "both dice roll even numbers" and event F = "the sum of the dots on the dice is odd." We need to find $P\,(E \text{ or } F)$.

We notice that E and F are mutually exclusive since even numbers cannot add to make an odd number. We have

$$P(E) = \frac{9}{36} \qquad \text{and} \qquad P(F) = \frac{18}{36}$$

so that

$$P\,(E \text{ or } F) = P(E) + P(F) = \frac{9}{36} + \frac{18}{36} = \frac{27}{36}$$

Ex: What is the probability that, when you roll two dice, both dice both roll an even number or the sum of the dots on the dice is equal to 8?

Let event E = "both dice roll even numbers" and event F = "the sum of the dots on the dice is equal to 8." We need to find $P\,(E \text{ or } F)$.

We notice that E and F are not mutually exclusive since even numbers can add to 8. We have

$$P(E) = \frac{9}{36} \qquad \text{and} \qquad P(F) = \frac{5}{36} \qquad \text{and} \qquad P\,(E \text{ and } F) = \frac{3}{36}$$

so that

$$P\,(E \text{ or } F) = P(E) + P(F) - P\,(E \text{ and } F) = \frac{9}{36} + \frac{5}{36} - \frac{3}{36} = \frac{11}{36}$$

Ex (One Additional Example): Pick three cards, without replacement, from a standard deck. What is the probability of getting at least one ace?

Let $E =$ "drawing at least one ace."

Solution 1 (hard): There are three ways to get at least one ace: 1 ace OR 2 aces OR 3 aces.

For one ace, we can get the ace on the first draw OR the second draw OR the third draw (and non-aces on the other two draws) with probability

$$\begin{array}{ccccccc} \text{first draw} & \text{or} & \text{second draw} & \text{or} & \text{third draw} & & \\ \frac{4}{52}\cdot\frac{48}{51}\cdot\frac{47}{50} & + & \frac{48}{52}\cdot\frac{4}{51}\cdot\frac{47}{50} & + & \frac{48}{52}\cdot\frac{47}{51}\cdot\frac{4}{50} & = & 3\cdot\frac{4}{52}\cdot\frac{48}{51}\cdot\frac{47}{50} \end{array}$$

For two aces, we can get the non-ace on the first draw OR the second draw OR the third draw (and aces on the other two draws) with probability

$$\begin{array}{ccccccc} \text{first draw} & \text{or} & \text{second draw} & \text{or} & \text{third draw} & & \\ \frac{48}{52}\cdot\frac{4}{51}\cdot\frac{3}{50} & + & \frac{4}{52}\cdot\frac{48}{51}\cdot\frac{3}{50} & + & \frac{4}{52}\cdot\frac{3}{51}\cdot\frac{48}{50} & = & 3\cdot\frac{48}{52}\cdot\frac{4}{51}\cdot\frac{3}{50} \end{array}$$

For three aces, the probability is just

$$\frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}$$

So, the probability to get at least one ace is the sum of the probabilities of getting one ace, two aces, and three aces:

$$P(E) = \underbrace{\left(3\cdot\frac{4}{52}\cdot\frac{48}{51}\cdot\frac{47}{50}\right)}_{\text{1 ace}} + \underbrace{\left(3\cdot\frac{48}{52}\cdot\frac{4}{51}\cdot\frac{3}{50}\right)}_{\text{2 aces}} + \underbrace{\left(\frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}\right)}_{\text{3 aces}} = \frac{1201}{5525}$$

Solution 2 (easy): Note that NOT $E =$ "not drawing any aces." The probability of this is simply

$$P\left(\text{NOT } E\right) = \frac{48}{52}\cdot\frac{47}{51}\cdot\frac{46}{50}$$

Using the NOT rule, we have

$$P(E) = 1 - P\left(\text{NOT } E\right) = 1 - \frac{48}{52}\cdot\frac{47}{51}\cdot\frac{46}{50} = \frac{1201}{5525}$$

9 Binomial Formula

Definition: A Binomial process is one in which the same experiment is performed repeated times.

We are typically interested in the number of successes in all repetitions.

For example, roll a fair die 10 times. What is the chance of getting exactly three 1's?

Definition: $n! = n \cdot (n-1) \cdot n-1 \cdot \cdots \cdot 3 \cdot 2 \cdot 1$. By definition, $0! = 1$.

For example, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$.

Definition: *Binomial coefficient:* $\binom{n}{k} = \dfrac{n!}{k!\,(n-k)!}$

Interpretation: This gives the number of ways, disregarding order, that k objects can be chosen from among n objects.

Ex:

$$\begin{aligned}\binom{10}{8} &= \frac{10!}{8!\,(10-8)!} = \frac{10!}{8! \cdot 2!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1} \\ &= \frac{10 \cdot 9 \cdot \cancel{8} \cdot \cancel{7} \cdot \cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}{\cancel{8} \cdot \cancel{7} \cdot \cancel{6} \cdot \cancel{5} \cdot \cancel{4} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1} \cdot 2 \cdot 1} \\ &= \frac{10 \cdot 9}{2} = 45\end{aligned}$$

Definition: Binomial Formula Gives the probability of k successes in n trials, where p is the probability of success on a single trial.

$$P\,(k \text{ successes}) = \binom{n}{k} p^k\,(1-p)^{n-k}$$

Ex: Find probability of rolling exactly 2 heads on 5 flips of a fair coin.

Let E be said event. The probability is

$$P(E) = \binom{5}{2} \cdot \frac{1}{2}^2 \cdot \left(1 - \frac{1}{2}\right)^{5-2} = \frac{5!}{2!3!} \cdot \frac{1}{32} = \frac{10}{32} = 0.3125.$$

Ex: Recall the "Hard" solution to $E =$ "drawing at least one ace":
So, the probability to get at least one ace is the sum of the probabilities of getting one ace, two aces, and three aces:

$$\begin{aligned}P(E) &= \underbrace{\left(\binom{3}{1} \cdot \frac{4}{52} \cdot \frac{48}{51} \cdot \frac{47}{50}\right)}_{\text{1 ace}} + \underbrace{\left(\binom{3}{2} \cdot \frac{48}{52} \cdot \frac{4}{51} \cdot \frac{3}{50}\right)}_{\text{2 aces}} + \underbrace{\left(\binom{3}{3} \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50}\right)}_{\text{3 aces}} \\ P(E) &= \left(3 \cdot \frac{4}{52} \cdot \frac{48}{51} \cdot \frac{47}{50}\right) + \left(3 \cdot \frac{48}{52} \cdot \frac{4}{51} \cdot \frac{3}{50}\right) + \left(1 \cdot \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50}\right) = \frac{1201}{5525}\end{aligned}$$

10 Central Limit Theorem for Sums of Draws

After many tosses of a fair coin, we should have 50% heads. This doesnt happen exactly since

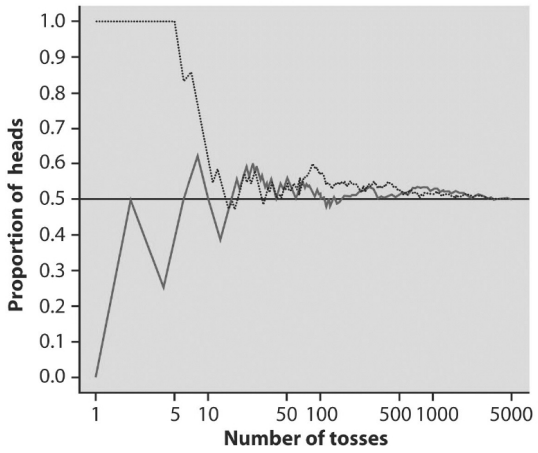
$$\text{number of heads} = \text{half the number of tosses} + \text{chance error}.$$

However, as a percentage of the number of tosses, it turns out that

$$\text{chance error as a } \% \to 0\%$$

as the number of tosses gets big.

Ex: Below is a graph of two experiments: number of tosses of a coin vs. percentage of heads.



Read pages 278 to 284.

Definition: Sum of Draws: By this it is meant that we have numbers in a box and we are picking *with replacement*. We look at the sum of the numbers picked.

Ex: Box has one 0 and one 1. Pick 100 times. Get 57 as the sum. This models flipping a fair coin 100 times and counting the number of heads (or tails).

Ex: Box has one each of the integers 1 through 6. We pick twice and find the sum (say, 8). This models rolling a pair of fair, 6-sided dice and considering the sum on the dice (we would have 8).

Making a Box Model

- Decide what numbers go in the box (sometimes called tickets)
- Decide how many of each number/ticket go in the box.
- Decide how many times are we picking from the box.

Ex: A roulette wheel has 18 red numbers, 18 black numbers, and 2 green numbers (0 and 00). If you bet \$1 on red and the ball lands on red, you get your dollar back plus another dollar. If the ball lands on black or green, you lose your dollar. The box for this is

18	+\$1	20	−\$1

The sum of draws (with replacement) from a box follows, approximately, a Normal distribution with

$$\text{mean} = EV_{\text{sum}} \qquad \text{and} \qquad \text{standard deviation} = SE_{\text{sum}}$$

where EV stands for *expected value* and SE stands for *standard error*, if the number of draws is sufficiently large. For us, sufficiently large will be at least 25.

Ex: Suppose we drew 25 tickets out of the box below with replacement. What might we expect for a sum?

0	2	3	4	6

You can expect to get each number 5 times. That gives us

$$5{\times}0 + 5{\times}2 + 5{\times}3 + 5{\times}4 + 5{\times}6 = 75$$

This is the expected value.

Of course, we will not always get that many 0's, 2's, 3's, 4's, or 6's. Which will cause our sum to be off of 75 a little bit. But by how much? This is where the standard error SE comes into play. A typical sum will be of the form

$$\text{sum} = EV \pm \text{chance error},$$

where the chance error is determined by the standard error.

Definition: The formulas are

$$\begin{aligned} EV &= (\text{number of draws}) \cdot (\text{average of box}) \\ SE &= \sqrt{(\text{number of draws})} \cdot (SD \text{ of box}) \end{aligned}$$

Ex: For the box | 0 | 2 | 3 | 4 | 6 | with 25 draws, we have

$$\begin{aligned} EV &= 25 \cdot \frac{0+2+3+4+6}{5} = 25 \cdot 3 = 75 \\ SE &= \sqrt{25} \cdot \sqrt{\frac{0^2+2^2+3^2+4^2+6^2}{5} - 3^2} = 10 \end{aligned}$$

Ex: Using these values of EV and SD, what percentage of the time should we obtain a sum between 50 and 100?

We use the standard normal curve. Convert these values to standard units:

$$\frac{50-75}{10} = -2.5 \qquad \text{and} \qquad \frac{100-75}{10} = 2.5.$$

Using the back of the book, we see that the area between -2.5 and 2.5 is 98.76%.

Ex: Consider a roulette wheel in a casino. If 10,000 people bet on red, what is the chance that the casino wins money on these 10,000 bets?

We are asking the question that the casino makes more than \$0, not the player. We need to convert this into terms of the standard normal curve using the z-score formula. Our box here is

$$\boxed{18 \;\boxed{-\$1}\;\; 20 \;\boxed{+\$1}}$$

We first find EV and SD:

$$\begin{aligned} EV &= 10,000 \cdot \frac{2}{38} \approx 526.32 \\ SE &= \sqrt{10,000} \cdot \sqrt{\frac{18 \cdot (-1)^2 + 20 \cdot 1^2}{38} - \left(\frac{2}{38}\right)^2} \approx 99.86 \end{aligned}$$

To find our chances using the standard normal curve, we need to convert to standard units. We calculate

$$\frac{0 - 526.32}{99.86} \approx -5.27$$

Checking the table, we see that 5.27 is not even on the table; the biggest number we see is 4.45. The area to the right of -4.45 is greater than 99.99955%, which implies that the area to the right of -5.27 is even greater.

SD Short-Cut: If a box has only two different numbers, a big number and a small number, then the formula for SD is much simpler:

$$SD = \left(\begin{array}{c}\text{big}\\ \text{number}\end{array} - \begin{array}{c}\text{small}\\ \text{number}\end{array} \right) \times \sqrt{\begin{array}{c}\text{fraction with}\\ \text{big number}\end{array} \times \begin{array}{c}\text{fraction with}\\ \text{small number}\end{array}}$$

Ex: For the box $\boxed{\boxed{3}\;\boxed{3}\;\boxed{3}\;\boxed{3}\;\boxed{3}\;\boxed{2}\;\boxed{2}\;\boxed{2}\;\boxed{2}}$, We have

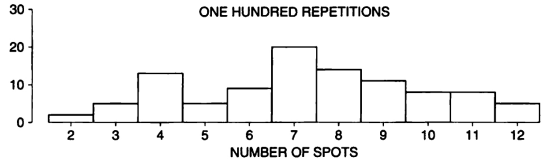
$$SD = (3-2)\sqrt{\frac{5}{9} \times \frac{4}{9}} = \frac{2}{9}\sqrt{5} \approx 0.4969$$

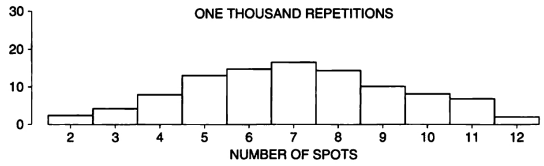
Read Pages 299-303.

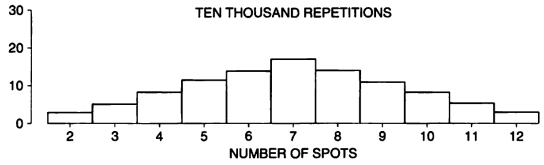
11 Probability Histogram

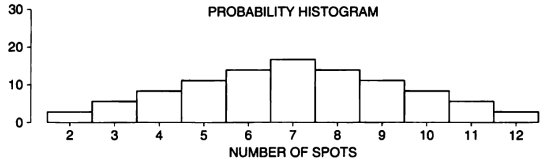
Read Chapter 18.

You roll a pair of dice 100 times, 1,000 times, or 10,000 times, and find the percentage of times each sum appears. Below are the *probability histograms*, which represents chance, not data. A probability histograms represents chance by area.







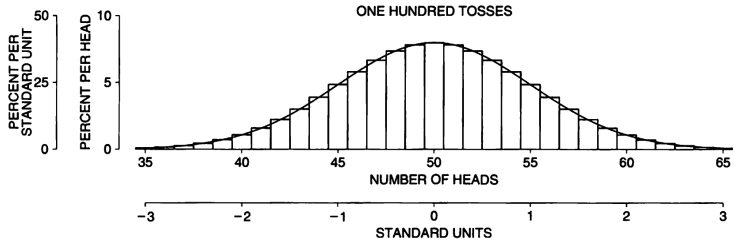


Definition: A *random variable* is a variable whose value is a numerical outcome of a random phenomenon.

An example of a random variable is the sum of a pair of dice. Another example is below: the number of heads obtained when a coin is tossed 100 times.

The probability histogram of a random variable X tells us what values X can take and the probabilities associated to those values.

What about the probability histogram for the number of heads in 100 coin tosses?



The EV is 50 and the SE is 5.

Ex: (See Example on Page 317.) A coin will be tossed 100 times. Estimate the chance of getting between 43 and 57 heads inclusive. Now exclusive.

Inclusive: The class interval over 43 goes between 42.5 and 43.5 and the class interval over 57 goes between 56.5 and 57.5. So, we are interested in finding the area from 42.5 to 57.5 that is made up by these 15 rectangles. Converting to standard units, we have

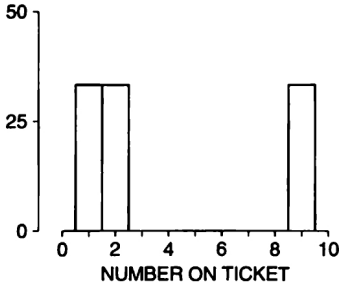
$$z = \frac{42.5 - 50}{5} = -1.5 \qquad \text{and} \qquad z = \frac{57.5 - 50}{5} = 1.5$$

Using the table, this gives us an area of 86.64%.

Exclusive: We look for heads between 43 and 57, but not counting 43 or 57. So, 44 to 56. This is the range from 43.5 to 56.5, which is -1.3 to 1.3. This gives us an area of 80.64%.

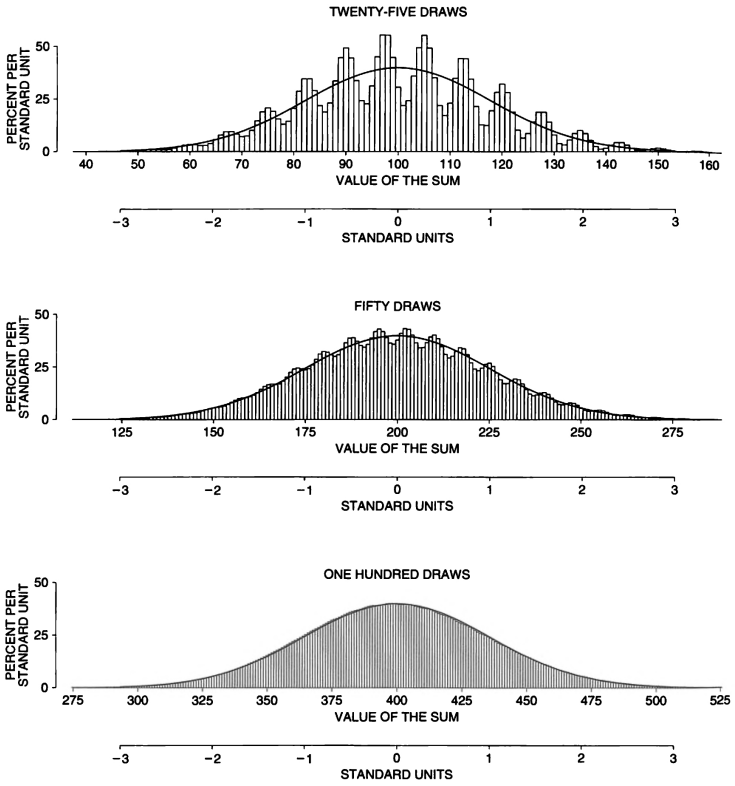
Read Pages 319 to 324.

Ex: One example from the book is a histogram that is not normal for the box $\boxed{\boxed{1}\boxed{2}\boxed{9}}$. The histogram for the box is



If we make several draws and sum up the numbers, the histogram for the sums

changes depending on the number of draws we make.



This normal curve is tied to sums. This will not work if we tried to it with products. If we did, the resulting histogram will be very right skewed.

The Central Limit Theorem: When drawing with replacement from a box, the probability histogram for the sum will follow the normal curve, even if the contents of the box do not. The histogram must be put into standard units, and the number of draws must be reasonably large.

The expected value (EV) pins the center of the probability histogram to the horizontal axis, and the standard error (SE) fixes its spread.

12 Sampling

A numerical fact about the population is called a *parameter*. The parameter is usually unknown. Investigators want to estimate this parameter using a sample; this process is called making *inferences*. Parameters are estimated by *statistics*, or numbers which can be computed from a sample.

Ex: To estimate the average height of people in Maine, we take a sample of 10,000 people. The average height of the sample is the statistic. The unknown average height of people in Maine is the parameter.

Ex: To estimate the percentage of people who watch a certain TV program, investigators take a sample of 30,000 people. The percentage of people in the sample is the statistic, and the unknown percentage of the population is the parameter.

How do we get a sample? We can use a *simple random sample (SRS)*: everybody has an equal chance of being in the sample. For example, everybody's name is written on a ticket, placed in a box, and drawn without replacement, and every ticket has an equal chance of being chosen.

But this is not always the best option. Some big polling institutions use *multistage cluster sampling*. This has several stages:

1. Break up the US into regions: Northeast, South, Midwest, and West.
2. For each region, take SRS of towns of similar size, say 50 to 250 thousand.
3. For each selected town, take SRS of wards.
4. For each selected ward, take SRS of precincts.
5. For each selected precinct, take SRS of households.

There are several different methods to select a sample. The basics are

- interviewers have no discretion at all as to whom they interview;
- there is a definite procedure for selecting the sample, and it involves the planed use of chance.

To minimize bias, an impartial and objective probability method should be used to choose the sample.

Read through Chapter 19. We have already discussed different types of bias.

13 Chance Errors in Sampling

We have previously discussed EV and SE of sums of draws. We will now denote these EV_+ and SE_+ to distinguish them from a different version of expected value and standard error we will introduce below.

You take a SRS of 100 people from a population of 6,672 people that is 46% men. What percentage of the sample is men? Will this percentage change if you take a different SRS of 100 people? You can make a histogram of the different percentages you get when you take different SRS's of size 100.

You can model taking an SRS using sum of draws. Think of a box

$$\boxed{3,091 \ \boxed{1} \ \ 3,581 \ \boxed{0}}$$

where the 1's represent men and the 0's represent women. The fractions of 1's is 0.46 and the fraction of 0's is 0.54.

With a SRS, the expected value for the sample (which we denote $EV_{\%}$) equals the population percentage. Since our box consists of just 1's and 0's, the expected value for the sample is

$$EV_{\%} = (\text{average of box}) \times 100\%.$$

Different samples will give you different sample percentages.

$$\text{sample percentage} = EV_{\%} + \text{chance error}$$

How much chance error depends on the sample size.

To simplify things, let n denote the size of your sample (also the number of draws from your box). To compute the standard error for a percentage (which we denote $SE_{\%}$), we use

$$SE_{\%} = \frac{SE_+ \text{ of box}}{n} \times 100\% = \frac{\sqrt{(n)} \cdot (SD \text{ of box})}{n} \times 100\% = \frac{SD \text{ of box}}{\sqrt{n}} \times 100\%$$

Notation: The usual notation of the population percentage is simply p. The usual notation for sample percentage is $\hat{p}$. If we let p, as a decimal, denote the percentage of 1's in the box, then the formula above becomes

$$SE_{\%} = \frac{\sqrt{p\,(1-p)}}{\sqrt{n}} \times 100\% = \sqrt{\frac{p\,(1-p)}{n}} \times 100\%,$$

which is the standard way this formula is presented. Later on when we want to use $\hat{p}$ to estimate p (which we won't know), we will most likely use the formula $SE_{\%} = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}} \times 100\%$, where we replace p with $\hat{p}$.

The number of 1's in the box follows a normal distribution with mean $EV_{\%}$ and standard deviation $SE_{\%}$. This distribution is sometimes called a *sampling distribution*.

The larger the sample size, the smaller the chance error. Consider the histograms below:

Figure 1. Histogram for the number of men in samples of size 100.

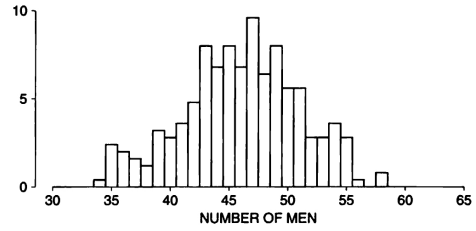
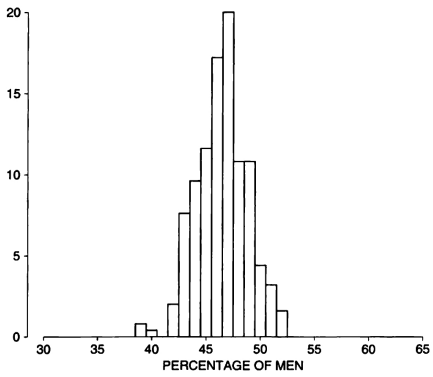


Figure 2. Histogram for the percentages of men in samples of size 400. There are 250 samples, drawn at random from the respondents to the health study.



(For us to take 250 samples from a population of size 6,672, some people will have to be in several samples.) Notice that there is less spread in the right histogram. The SD of the box is $(1-0)\times\sqrt{0.46\times 0.54} \approx 0.5$. We calculate

$$\begin{aligned} SE_{\%} \text{ for } 100 &= \frac{0.5}{\sqrt{100}}\times 100\% = 5\% \\ SE_{\%} \text{ for } 400 &= \frac{0.5}{\sqrt{400}}\times 100\% = 2.5\% \end{aligned}$$

How do SE_+ and $SE_{\%}$ behave as n grows larger? Remember that

$$SE_+ = \sqrt{n} \times (SD \text{ of box}) \qquad \text{and} \qquad SE_{\%} = \frac{SD \text{ of box}}{\sqrt{n}}\times 100\%$$

By looking at these equation, we see that

SE_+ grows as n gets larger and $SE_{\%}$ shrinks as n gets larger

Ex: A town has 3000 men and 2000 women. In a sample of 200 residents, the sample will have ____% men, give or take ____%.

Modeling the town's population, we have a box

$$\boxed{3,000 \boxed{1} \quad 2,000 \boxed{0}}$$

where the 1's represent men and the 0's represent women. Note that $\hat{p} = .6$.

We have

$$\begin{aligned} EV_+ &= (200) \times \left(\frac{3000}{5000}\right) = 120 \\ SE_+ &= \sqrt{200}\times (1-0) \times \sqrt{\frac{3000}{5000}\times\frac{2000}{5000}} = 10\sqrt{2}\frac{\sqrt{6}}{5} = 4\sqrt{3} \approx 6.9282 \end{aligned}$$

Also, we have

$$\begin{aligned} EV_{\%} &= \left(\frac{3000}{5000}\right)\times 100\% = 60\% \\ SE_{\%} &= \frac{(1-0)\times\sqrt{\frac{3000}{5000}\times\frac{2000}{5000}}}{\sqrt{200}}\times 100\% = \frac{\sqrt{.6\,(1-.6)}}{\sqrt{200}}\times 100\% = \frac{\sqrt{3}}{50} \approx 3.4641\% \end{aligned}$$

So, what is the relation between the two set of numbers? Notice that

120 out of 200 is 60%
6.9282 out of 200 is 3.4641%

Ex: Continuing on with the example, what is the chance that the sample percentage is between 54.80% and 65.20%?

Using the fact the sampling percentage follows a normal distribution, we convert these percentages to standard units:

$$z = \frac{54.80-60}{3.4641} \approx -1.5 \qquad \text{and} \qquad z = \frac{65.20-60}{3.4641} \approx 1.5$$

The table in the back of the book yields a chance of 86.64%.

Ex: A flawed example: A town has 3000 men and 2000 women. We take a sample of 100 residents to fill in the blanks: The town has ____% men, give or take ____%. Same solutions: $EV_{\%} = 60\%$ and $SE_{\%} \approx 3.4641\%$.

Some observations for the last example:

1. If we know the population makeup (i.e., whats in the box) then we do not need to do a survey. Hence, when we are doing surveys we do *not* know whats in the box. So, if we don't know the box's makeup, how do we proceed?

   We assume that our sample size is large enough to give us an accurate estimate of the population. Assuming our sample answer is correct, we then proceed to find the likely size of the chance error (how much our sample answer might be off).

2. When we survey, we do not ask the same person twice. Hence, we are picking from the box *without replacement*. Our formulas above are only for with replacement picking. How do we account for this?

   Let $SE_{w/o}$ denote the standard error without replacement and SE_{with} denote the standard error with replacement. (Note: SE can be either $SE_{\%}$ or SE_{+}, and also SE_{avg}, for average, which will come later.) The relationship is

$$SE_{w/o} = CF\times SE_{with} \qquad \text{where} \qquad CF = \sqrt{\frac{N-n}{N-1}}$$

Here, CF stands for the *Correction Factor* and

$$\begin{aligned} N &= \text{number of tickets in box} = \text{population size} \\ n &= \text{number of draws} = \text{sample size} \end{aligned}$$

The book writes this as

$$CF = \sqrt{\frac{\text{number of tickets in box} - \text{number of draws}}{\text{number of tickets in box} - \text{one}}}$$

The correction factor; the number of draws is fixed at 2,500.

Number of tickets in the box	*Correction factor (to five decimals)*
5,000	0.70718
10,000	0.86607
100,000	0.98743
500,000	0.99750
1,500,000	0.99917
15,000,000	0.99992

Notice that $CF < 1$ so that $SE_{w/o} < SE_{with}$. This makes sense. When our sample size n increases and we are picking without replacement, we are getting closer and closer to the actual answer (if $n = N$, we have polled everyone so there is 0% error; note that $CF = 0$ in this situation so $SE_{w/o}$ would be 0% if we poll everyone).

Another consequence of the correction factor is the following:

When estimating percentages, it is the absolute size of the sample which determines the accuracy, not the size relative to the population. This is true if the sample is only a small part of the population, which is usually the case.

Notice that when as N is large relative to n, the correction factor CF is nearly 1 and can be ignored. You only need to worry about the correction factor if the sample is a substantial fraction of the population.

Ex: We survey 1000 students at a university with 27,000 students to determine the percent who play a sport. Of the 1000, we find that 550 play a sport. Estimate the percent of students at the university who play a sport with an error estimate.

Using our sample, we have 55% 1's and 45% 0's. We calculate

$$\begin{aligned} EV_{\%} &= 55\% \\ SE_{\%}^{with} &= \frac{(1-0)\sqrt{.55 \cdot .45}}{\sqrt{1000}} \times 100\% = 1.57 \\ CF &= \sqrt{\frac{27,000 - 1,000}{27,000 - 1}} = 0.9813 \\ SE_{\%}^{w/o} &= CF \times SE_{\%}^{with} = 0.9813 \times 1.57 = 1.54 \end{aligned}$$

In short, if we didn't know the population size, the error is $SE_{\%}^{with} = 1.57$, which is bigger than $SE_{\%}^{w/o} = 1.54$. Hence, if we know the population size, we are able to get a better error estimate.

If we don't know the population size, and hence cannot calculate the correction factor, we assume that our population size is much larger than our sample size so that $CF \approx 1$. Alternatively, we can let $CF = 1$ and produce an error estimate that *overestimates* the error.

14 Confidence Intervals for Percentages

Definition: *Statistical inference* is the process of drawing conclusions from data obtained from a sample. These conclusions are used to gain some insight about a larger population.

Definition: A *confidence interval* is an interval of values that estimates an unknown population value, usually a percentage or average.

For right now, we will concentrate on percentages. A confidence interval for a population percentage is generically of the form

$$\text{sample } \% \pm \text{margin of error}$$

Ex: On the nightly news, you see the results of a poll that says the percentage of dog owners that own multiple dogs is

$$43\% \qquad \pm 3.5 \text{ MOE}$$

The MOE is the margin of error. You should interpret this as meaning that between 39.5% and 46.5% of dogs owners have more than one dog.

We don't know the exact population percentage; we just know there is a good chance it is between those numbers. How "good" of a chance depends on what *level of confidence*, or *confidence level*, was used to calculate the margin of error. The word *confidence* refers to the fraction or percentage of random samples for which a confidence interval procedure gives an interval that includes the unknown value of a population parameter.[1]

A more concrete formula for a confidence interval is

$$\text{sample } \% \pm z^* \times SE_{\%}$$

Note: Some books write the formula $\hat{p} \pm z^* \sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$, which gives decimals.

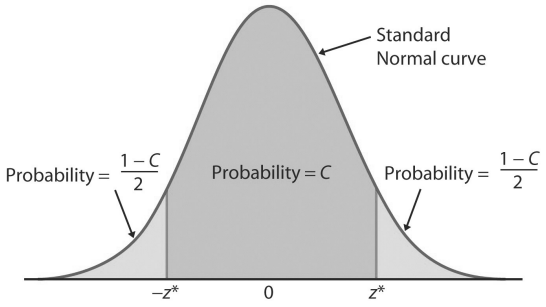
Here, z^* is called the *multiplier* and is related to the confidence level.

Ex: Here are a few examples for different values of z^* for confidence intervals (CI):

[1] Be careful when giving information about a specific confidence interval computed from an observed sample. The confidence level expresses only how often the confidence interval procedure works in the long run. It does not tell us the probability that a specific interval includes the population value.

- For a 68% CI, we have sample $\% \pm 1 \times SE_{\%}$. This means, in about 68% of random samples where we calculate this CI, the population percentage will be in this interval.
- For a 95% CI, we have sample $\% \pm 2 \times SE_{\%}$. This means, in about 95% of random samples where we calculate this CI, the population percentage will be in this interval.
- For a 99.7% CI, we have sample $\% \pm 3 \times SE_{\%}$. This means, in about 99.7% of random samples where we calculate this CI, the population percentage will be in this interval.

The confidence level and z^* is related to the normal curve as follows: For a confidence level C, we look for the area below.



The standard unit z^* is such that the area under the curve, as a percentage, between $-z^*$ and z^* is equal to C.

Ex: In a sample of 300 people from a large city, it was found that 175 were in favoring of later library hours. Find a 90% CI for the percentage of the population that are in favor of later library hours. Interpret what this interval means.

Since we have no information about the city population, we assume that $CF \approx 1$. Also, we will fill our box with our sample information: $\boxed{175\ \boxed{1}\ 125\ \boxed{0}}$. The sample percentage and $SE_{\%}$ are

$$\begin{aligned} \text{sample } \% &= \frac{175}{300} \times 100\% = 58.3\% \qquad \Rightarrow \quad \hat{p} = 0.583 \\ SE_{\%} &= \frac{SD \text{ of box}}{\sqrt{n}} \times 100\% = \frac{\sqrt{\hat{p}\,(1-\hat{p})}}{\sqrt{n}} \times 100\% \\ &= \frac{\sqrt{0.583\,(1-0.583)}}{\sqrt{300}} \times 100\% \approx 2.85\% \end{aligned}$$

We need to find z^*. Using the table and looking for the area close to 90, we see that $z^* = 1.65$. Our CI is

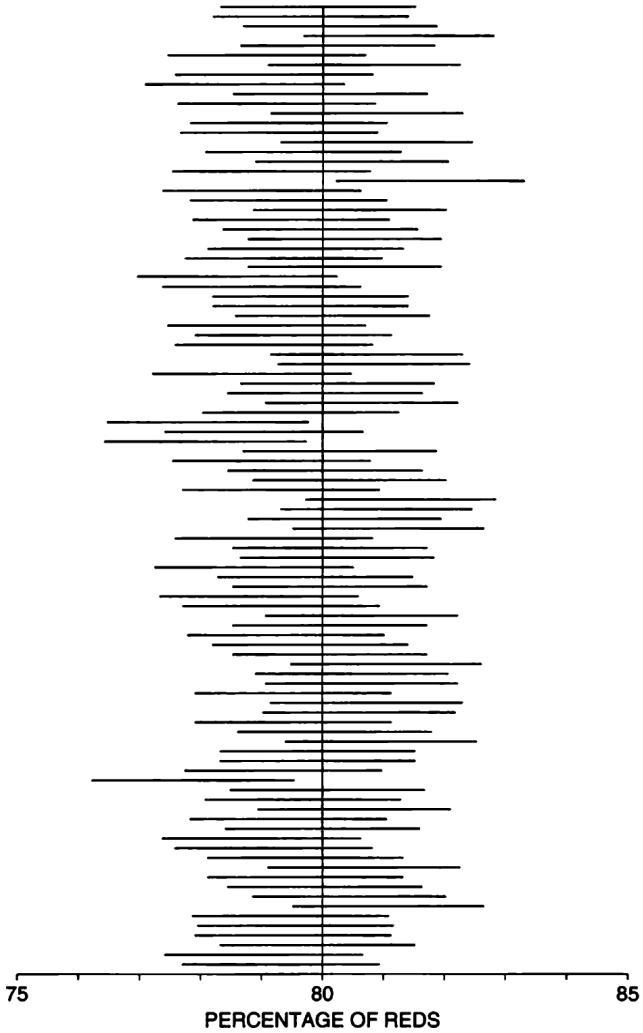
$$\text{sample } \% \pm z^* \times SE_{\%} \quad \Rightarrow \quad 58.3\% \pm 1.65 \times 2.85\% \quad \Rightarrow \quad 53.5975\% \text{ to } 63.0025\%$$

Now to interpret this interval. We are 90% confident that the population percentage is between 53.5975% and 63.0025%.

What do we mean by 90% confident? Of all samples of size 300 where we calculate the CI, the population percentage will be in the resulting interval about 90% of the time.

For a graphical interpretation of what this means, see the figure below.

Figure 1. Interpreting confidence intervals. The 95%-confidence interval is shown for 100 different samples. The interval changes from sample to sample. For about 95% of the samples, the interval covers the population percentage, marked by a vertical line.[8]



Unknown to pollsters, the true population percentage is 80% and the SD is

3. If you have 100 data values (which are sample percentages in this case) from the distribution above, then you would expect about 95 of those values to be with in $2SD$ of 80%; so, about 5 of those data values will be outside the $2SD$ range. What we see in the picture is a line drawn from each data value that emanates outward $2SD$'s in both directions; this line represents the 95% CI for that particular data value. There are 4 data values that are outside the $2SD$ range, and so the CI does not contain 80%. In about 95% of these CI's, the population percentage 80% is contained in the CI.

If your population distribution is normal, then your sample will pretty much follow a normal distribution. But if your population is skewed, you need to be careful and use a larger sample size.

Also, you need to be careful if your sample percentage is close to 0% or 100%; your sample box could be lopsided. It might not be possible to do a CI; you might not be able to trust the resulting CI. To make sure that your results are accurate, you will need a larger sample size. The larger the sample size, the better the normal approximation. Although, you might still want to consider your results if your resulting CI is still between 0% and 100%.

For example, suppose our sample % is .9% and our CI is .02% to 1.6%. Such a small sample % might be questionable, but the lower endpoint is still positive.

Things that will change the MOE $z^*SE_{\%}$:

- Having a smaller SD for the sample – and thus a smaller SD for the box – will decrease $SE_{\%}$, but we really have no control over the value of SD.
- The sample size affects the $SE_{\%}$. As n increases,

$$SE_{\%} = \frac{SD \text{ of box}}{\sqrt{n}} \times 100\%$$

  decreases, and vice versa.

- We can change the level of confidence. The more the confidence, the larger the value of z^*, and thus the margin of error will be larger. That is, the more confidence we use, the wider the CI will be.

Ex: For a sample of size 100, a 99% CI will be wider than a 95% CI.

Ex: A 95% CI for a sample of size 200 will be narrower than a 95% CI for a sample of size 100.

What sample size to use?

Suppose you want a specific value for the MOE $m\%$. Solving the formula

$$m\% = z^* \times \frac{SD \text{ of box}}{\sqrt{n}} \times 100\% \quad \Rightarrow \quad \sqrt{n} = z^* \times \frac{SD \text{ of box}}{m\%} \times 100\%$$

$$\Rightarrow \quad n = \left(z^* \times \frac{SD \text{ of box}}{m\%} \times 100\% \right)^2$$

Ex: Previous polls have shown that 40% of all voters favor Candidate X. At a 95% confidence level, find the number of people we need to sample to create a new poll for the percentage of voters who favor Candidate X with a margin of error of 3%.

Here, $z^* = 2$ so that

$$n = \left(z^* \times \frac{SD \text{ of box}}{m\%} \times 100\% \right)^2 = \left(2 \times \frac{\sqrt{.4 \times .6}}{3\%} \times 100\% \right)^2 \approx (32.66)^2 \approx 1066.67$$

Hence, we need a sample size of 1067 to give a 3% margin of error at a confidence level of 95%. In other words, with a sample size of 1067, the resulting 95% confidence interval will be approximately Sample % $\pm$3%.

Notes:

These methods (as well as the ones just below) work only for SRS. Other sampling methods require different CI methods.

When the news reports statistics with MOE, like the dog example, they typically use a 95% CI.

A CI for the difference between two population percentages.

Suppose you have two populations and you are interested in the difference between the two population percentages:

$$\text{pop 1 \%} - \text{pop 2 \%}$$

Ex: If you knew that 45% of men liked a TV program and 42% of women liked the same TV program, then the difference in the population percentages would be 45%-42%=3%.

We take separate SRS's from each population. The sample sizes do not have to be the same size. For each sample we calculate the sample % and the $SE_{\%}$.

The CI for the difference between two population percentages is

$$(\text{sample 1 \%} - \text{sample 2 \%}) \pm z^* \sqrt{(SE_{\%}1)^2 + (SE_{\%}2)^2}$$

where

$$\begin{aligned} SE_{\%}1 &= \frac{SD \text{ of box of sample 1}}{\sqrt{n_1}} \times 100\% \quad \text{and} \\ SE_{\%}2 &= \frac{SD \text{ of box of sample 2}}{\sqrt{n_2}} \times 100\% \end{aligned}$$

with n_1 and n_2 the sizes of sample 1 and sample 2, respectively.

Ex: A study was conducted to determine whether there is a relationship between snoring and risk of heart disease (Norton and Dunn, 1985). It was found that of the 1105 snorers they sampled, 86 had heart disease, while only 24 of the 1379 nonsnorers had heart disease. We will define population 1 to be

snorers and population 2 to be nonsnorers. Find a 95% CI for the difference between the percentages of the snorers and nonsnorers.

We have

$$\begin{aligned} \text{sample 1 \%} &= \frac{86}{1105}\times 100\% \approx 7.7828\% \\ \text{sample 2 \%} &= \frac{24}{1379}\times 100\% \approx 1.7404\% \end{aligned}$$

and

$$\begin{aligned} SE_{\%}1 &= \frac{\sqrt{.077828\,(1-.077828)}}{\sqrt{1105}}\times 100\% \approx 0.80592\% \\ SE_{\%}2 &= \frac{\sqrt{.017404\,(1-.017404)}}{\sqrt{1379}}\times 100\% \approx 0.35215\% \end{aligned}$$

Using $z^* = 2$, the CI is

$$\begin{aligned} & (\text{sample 1 \%} - \text{sample 2 \%}) \pm z^*\sqrt{(SE_{\%}1)^2 + (SE_{\%}2)^2} \\ \Rightarrow\quad & (7.7828\% - 1.7404\%) \pm 2\sqrt{(0.80592\%)^2 + (0.35215\%)^2} \\ \Rightarrow\quad & 6.0424\% \pm 2{\times}0.8795\% \\ \Rightarrow\quad & 4.2834\% \quad \text{to} \quad 7.8014\% \end{aligned}$$

We are 95% confident that the difference between the population is between 4.2834 and 7.8014%.

Note on notation (can ignore): Other books use the notation p to denote the population percentage and $\hat{p}$, pronounced p-hat, to denote sample percentage. With this notation, the formula for a CI estimating a population percentage p is

$$\hat{p} \pm z^*\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

and the formula for estimating the difference between two population percentages $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$$

where, $\hat{p}_1$ is the sample 1 percentage, $\hat{p}_2$ is the sample 2 percentage, n_1 is the sample 1 size, and n_2 is the sample 2 size.

15 Confidence Intervals for Averages

For this section, we assume that $n \geq 25$. We will see what to do when $n < 25$ later. (Spoiler: It involves a new type of distribution.)

When picking a large number of times (at least 25) from a box, the average of draws follows the Normal distribution with mean equal to EV_{avg} and standard deviation equal to SE_{avg}, where

$$\begin{aligned} EV_{avg} &= \frac{EV_+}{n} = \text{average of box} \\ SE_{avg} &= \frac{SE_+}{n} = \frac{SD \text{ of box}}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}} \end{aligned}$$

with n denoting the number of draws from the box. Some books use σ to denote the population standard deviation.

These formulas are for *with replacement* techniques. The formula

$$SE_{avg}^{w/o} = CF \times SE_{avg}^{with}$$

still applies in this case when you know the population size. (Remember: If you know the population size, you can use the CF to get a more accuracy.)

The general form[2] of a confidence interval for a confidence interval for a population average[3] is

$$\text{sample average} \pm z^* \times SE_{avg}$$

Ex: Find a 95% confidence interval for the average income in a town with 25,000 families. We survey 1000 families and get a total income of \$32,396,714 with a standard deviation of \$19,000.

We have

$$EV_{avg} = \frac{\text{total income}}{1000} = \frac{32,396,714}{1000} = 32,397.$$

Since we know the population size $N = 25,000$, we have

$$\begin{aligned} SE_{avg}^{w/o} &= CF \times SE_{avg}^{with} = \sqrt{\frac{N-n}{N-1}} \times \frac{SD \text{ of income}}{\sqrt{n}} \\ &= \sqrt{\frac{25,000-1000}{25,000-1}} \times \frac{19,000}{\sqrt{1000}} = 589 \end{aligned}$$

Our CI is

$$32,397 \pm 2 \times 589 \quad \Rightarrow \quad \$31219 \quad \text{to} \quad \$33575$$

We are 95% confident that the true average income for families in this town is between \$31219 and \$33575.

[2] Some books write this formula as $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$, where $\bar{x}$ denotes the sample average.

[3] This is for a sample size of $n \geq 25$.

Be careful how to interpret this. This CI does **NOT** mean that 95% of the residents of this town have incomes between \$31219 and \$33575.

Ex: We survey 400 people over the age of 25 to determine the average number of years of school attended by all people over the age of 25. We find that the 400 people had a total of 4635 years of schooling with an SD of 4.1 years. Find an 85% confidence interval for the average number of years of school attended for all people over the age of 25.

Since we don't know the population size, we assume $CF \approx 1$. We have

$$\begin{aligned} EV_{avg} &= \frac{4635}{400} = 11.59 \\ SE_{avg}^{w/o} &= CF \times SE_{avg}^{with} = 1 \times \frac{4.1}{\sqrt{400}} = .205 \end{aligned}$$

For an 85% CI, we use $z^* = 1.45$. Thus, we have

$$11.59 \pm 1.45 \times .205 \quad \Rightarrow \quad 11.29275 \text{ years} \quad \text{to} \quad 11.88725 \text{ years}$$

We are 85% confident that the true average number of years of school attended by all people over the age of 25 is between 11.29275 years and 11.88725 years.

The same methods for making the MOE smaller still work here.

If $m = z^* \times \dfrac{SD \text{ of box}}{\sqrt{n}}$ denotes the MOE, then $n = \left(z^* \times \dfrac{SD \text{ of box}}{m} \right)^2$.

Note: These methods require that we use a SRS.

A CI for the mean of differences.

You have data that comes in pairs (x_i, y_i), and you take the difference d_i of each pair:

$$\begin{aligned} x_1 - y_1 &= d_1 \\ x_2 - y_2 &= d_2 \\ &\vdots \\ x_n - y_n &= d_n \end{aligned}$$

You now have a list of the data $d_1, d_2, \ldots, d_n$, which is a list of difference. We wish to make a confidence interval for the mean of the differences. This confidence interval estimates the population of all differences; that is, if we took every pair in the population and then took the difference $x - y = d$, then the confidence interval would estimate the mean of the all these differences.

The formula of this confidence interval is the same as the formula for the confidence interval for estimating the population average:

$$\text{sample average} \pm z^* \times SE_{avg}.$$

Ex: The 25 students in a liberal arts course in statistical literacy were given a survey that included questions on how many hours per week they watched television and how many hours a week they used a computer. The responses are shown in the Table below, along with the difference d for each student. From these data, construct a 90% confidence interval for the average difference in hours spent using the computer versus watching television (Computer use − TV) for the population of students represented by this sample.

Student	Computer	TV	Difference	Student	Computer	TV	Difference
1	30	2.0	28.0	14	5	6.0	−1.0
2	20	1.5	18.5	15	8	20.0	−12.0
3	10	14.0	−4.0	16	30	20.0	10.0
4	10	2.0	8.0	17	40	35.0	5.0
5	10	6.0	4.0	18	15	15.0	0.0
6	0	20.0	−20.0	19	40	5.0	35.0
7	35	14.0	21.0	20	3	13.5	−10.5
8	20	1.0	19.0	21	21	35.0	−14.0
9	2	14.0	−12.0	22	2	1.0	1.0
10	5	10.0	−5.0	23	9	4.0	5.0
11	10	15.0	−5.0	24	14	0.0	14.0
12	4	2.0	2.0	25	21	14.0	7.0
13	50	10.0	40.0				

The average of the differences is 5.35, and the standard deviation $SD = 14.93$. We calculate $SE_{ave} = 14.93/\sqrt{25} = 2.986$. Our CI is

$$5.35 \pm 1.65 \cdot 2.986 \qquad \Rightarrow \qquad 0.4231 \text{ to } 10.2769.$$

Interpretation: We are 90% confident that the average difference between computer usage and television viewing for students represented by this sample is covered by the interval from 0.14 to 10.58 hours per week, with more hours spent on computer usage than on television viewing.

A CI for the difference between two population averages.

Suppose you have two populations and you are interested in the difference between the two population averages:

$$\text{pop 1 average} - \text{pop 2 average}$$

We take separate SRS's from each population. The sample sizes do not have to be the same size. For each sample we calculate the sample average and the SE_{avg}.

The CI for the difference between two population average is

$$(\text{sample 1 average} - \text{sample 2 average}) \pm z^* \sqrt{(SE_{avg}1)^2 + (SE_{avg}2)^2}$$

where

$$SE_{avg}1 = \frac{SD \text{ of box of sample 1}}{\sqrt{n_1}} \quad \text{and}$$
$$SE_{avg}2 = \frac{SD \text{ of box of sample 2}}{\sqrt{n_2}}$$

with n_1 and n_2 the sizes of sample 1 and sample 2, respectively.

Ex: Each of 63 students in a statistics class used their nondominant hand to print as many letters of the alphabet, in order, as they could in 15 seconds. Find a 95% CI for the difference in population means for males and females. The results are summarized below:

Gender	n	average	SD
Male	34	13.65	4.46
Female	29	12.55	4.01

Let population 1 be the male students and population 2 be the female students. We don't know the population sizes; so we assume that $CF \approx 1$. Also, we are already given the sample averages in the table. We calculate

$$SE_{avg}1 = \frac{SD \text{ of the males}}{\sqrt{n_1}} = \frac{4.46}{\sqrt{34}} \approx 0.76488$$
$$SE_{avg}2 = \frac{SD \text{ of the females}}{\sqrt{n_2}} = \frac{4.01}{\sqrt{29}} \approx 0.74464$$

This gives us

$$(13.65 - 12.55) \pm 2\sqrt{(0.76488)^2 + (0.74464)^2} \quad \Rightarrow \quad 1.1 \pm 2 \cdot 1.06749$$
$$\Rightarrow \quad -1.03498 \quad \text{to} \quad 3.23498$$

We are 95% confident that the difference between the two population averages is between -1.03498 and 3.23498.

Note here that 0 is in this interval. So it is possible that the population averages are equal. If it is the case that 0 is not in the CI, then there is less of a chance that the population averages are equal.

Note: If both of the endpoints for your CI are negative, then this just tells you that the population 2 average is probably larger than the population 1 average. An example of such a CI is

$$-5 \quad \text{to} \quad -2.$$

16 Hypothesis Testing

The *Null Hypothesis*, denoted H_0:

- The null hypothesis corresponds to the idea that an observed difference is due to chance.
- Usually the null hypothesis is a statement of "no effect" or "no difference," or it is a statement of equality.

The *Alternative Hypothesis*, denoted H_a:

- The alternative hypothesis corresponds to the idea that the observed difference is real.
- Usually the alternative hypothesis is a statement of "there is an effect" or "there is a difference," or it is a statement of inequality.

The null hypothesis and alternative hypothesis cannot both happen at the same time. Only one is true.

What do we mean by "observed difference?" It can deal with either a population average or population percentage. How do we state the null and alternative hypothesis?

Ex: Is the population average less than 20?
Null hypothesis is written: H_0: average =20.
Alternative hypothesis is written: H_a: average < 20.

Ex: Is the population percentage different from than 14%?
Null hypothesis is written: H_0: pop % =14%.
Alternative hypothesis is written: H_a: pop % $\neq$ 14%.

Note: You will always use = when you write the null hypothesis.[4]

We will deal with percentages first, and then come back to averages later.

When preforming a hypothesis test, we want to show that the alternative hypothesis is true; that is, that there is a difference. To do this, we assume that null hypothesis is true. We then find the probability that – assuming the null hypothesis is true – we got a sample percentage (or sample mean) as extreme or more extreme. If that probability if very small, the it is very unlikely that the null hypothesis is true. The alternative hypothesis is probably true.

[4]Some periodicals may instead write, for example,

$$H_0 : \text{pop average} \geq 20$$
$$H_a : \text{pop average} < 20,$$

but the equals always goes with the null hypothesis.

We need a way to assign probabilities. To do this, we first use what is called a *test statistic*:

$$z = \frac{\text{observed} - EV}{SE}$$

In our situation, the observed value is the sample percentage.

To find our probability, we look for a p-value. Assuming that the null hypothesis is true, the probability that the test statistic would take a value as extreme or more extreme than the value actually observed is called the *p-value* of the test.

That is, the p-value of a test is the chance of getting a big test statistic – assuming the null hypothesis is true. The p-value is *not* the chance of the null hypothesis being true.

How do we find this p-value depends on our alternative hypothesis.

Our null hypothesis is of the form $H_0 : \text{pop } \% = EV$. Our possible alternative hypotheses are

One-Sided

$H_a : \text{pop } \% > EV$

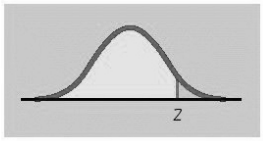
$H_a : \text{pop } \% < EV$

Two-Sided

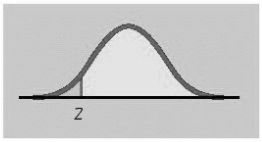
$H_a : \text{pop } \% \neq EV$

I stole the picture below from another book (or two). It shows why we label the alternative hypotheses one-sided and two-sided. The p-value is the shaded area under the bell curve. To understand the notation, $p = \text{pop } \%$ and $p_0 = EV$.

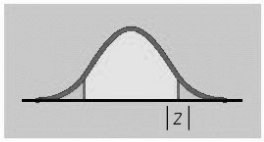
H_a: $p > p_0$ Area to the right of z (even if $z < 0$)



H_a: $p < p_0$ Area to the left of z (even if $z > 0$)



H_a: $p \neq p_0$ 2 × area to the right of $|z|$



Remember, when we are preforming the hypothesis test, we are assuming the null hypothesis is true. With this in mind, if the p-value is small enough,

then the null hypothesis is probably not true; it is evidence against the null hypothesis. Since we really want the alternative hypothesis to be true, this is good news.

But how small must a p-value be? This is the idea of *significance level*. There are two levels of significant that are usually used:

- If the p-value $< 5\%$, then the result is called *statistically significant*, or just *significant*.

  This is strong evidence against the null hypothesis.

- If the p-value $< 1\%$, then the result is called *highly significant*.

  This is very strong evidence against the null hypothesis.

If the p-value is $< 5\%$ (resp. 1%), then we say "The results are statistically (resp. highly) significant, and we can reject null hypothesis in favor of the alternative hypothesis."

Note: A note on hypothesis tests regarding one population proportion: For the best results, you want to make sure you have a sample size n large enough so that both

$$n \times EV > 10 \qquad \text{and} \qquad n \times (1 - EV) > 10.$$

Also, you need a SRS.

The Method:

1. White down the null and alternative hypothesis.
2. Calculate the test statistic.
3. Find the p-value.
4. State your conclusion in the context of the specific setting of the test.

Note: The method is for any of the hypothesis tests we encounter.

One thing to remember: When we are working with hypothesis testing for one population proportion, we use the following formula for SE when finding the test statistic:

$$SE = \frac{\sqrt{EV\,(1 - EV)}}{\sqrt{n}} \qquad \text{(use } EV \text{ as a decimal)}$$

Why do we use EV in the numerator? When running the hypothesis test, we are assuming that the null hypothesis is true; that is, that we assume that $H_0 : \text{pop } \% \ = EV$ is true. Thus, we fill our "box" with this population information and use this box when we find the SD of the box.

Ex: Question: Do you check the nutrition label when you buy food? Out of 1,003 people, 582 responded "frequently." Based on this data, is it reasonable to conclude that a majority of people frequently check nutrition food labels?

Using the Method:

1. The null and alternative hypothesis are:

$$H_0: \text{ pop } \% \ = 50\%$$
$$H_a: \text{ pop } \% \ > 50\%$$

   Here, $EV = 50\%$. We are assuming the null hypothesis is true. So we are assuming that the true population percentage is 50%.

2. The observed value is $\frac{582}{1003} \times 100\% = 58\%$, and the $SE = \frac{\sqrt{0.5(1-0.5)}}{\sqrt{1003}} \times 100\% = 1.579\%$. We see that $1003 \times .5 = 501.5 > 10$ and $1005 \times (1 - .5) = 501.5 > 10$; so we can use these methods. The test statistic is

$$z = \frac{\text{observed} - EV}{SE} = \frac{58\% - 50\%}{1.579\%} = 5$$

3. Since the alternative hypothesis is $H_a: \text{ pop } \% \ > 50\%$, the p-value is the area to the right of 5, which is virtually 0%.

4. Since $0\% < 1\%$, the results are highly significant. Thus we can reject the null hypothesis in favor of the alternative hypothesis. A majority of people frequently check nutrition food labels.

Explanation: If the null hypothesis were true, there there is basically a 0% chance that we would have gotten a sample percentage of 58%. So it is not likely that the null hypothesis is true. Therefore, the alternative hypothesis must be true.

Ex: Suppose that a pharmaceutical company wants to claim that side effects will be experienced by fewer than 20% of the patients who use a particular medication. In a clinical trial with $n = 400$ patients, they find that 68 patients experienced side effects. Perform a hypothesis test.

Using the Method:

1. The null and alternative hypothesis are:

$$H_0: \text{ pop } \% \ = 20\%$$
$$H_a: \text{ pop } \% \ < 20\%$$

   Here, $EV = 20\%$. We are assuming the null hypothesis is true. So we are assuming that the true population percentage is 20%.

2. The observed value is $\frac{68}{400}\times 100\% = 17\%$, and the $SE = \frac{\sqrt{0.20(1-0.20)}}{\sqrt{400}}\times 100\% = 2\%$. We see that $400\times .2 = 80 > 10$ and $400\times(1-.2) = 320 > 10$; so we can use these methods. The test statistic is

$$z = \frac{\text{observed} - EV}{SE} = \frac{17\% - 20\%}{2\%} = -1.5$$

3. Since the alternative hypothesis is $H_a:$ pop $\% < 20\%$, the p-value is the area to the left of -1.5, which is 6.68%.

4. Since $6.68\% > 5\%$, we conclude that there is not enough evidence to reject the null hypothesis. If this is an important issue for the company, it should consider gathering additional data.

   Note: Never say "We accept the null hypothesis." We can only say that there is not enough evidence to reject it. We need more data. Perhaps we could use a larger sample size.

Testing Hypotheses About the Difference in Two Population Proportions

The parameter of interest is pop 1 % − pop 2 %. We usually wish to test if this is different from 0.

The statistic of interest is sample 1 % − sample 2 %. We sill use this in our test.

Our null hypothesis will look like

$$H_0:\ \text{pop 1 \%} - \text{pop 2 \%} = 0, \quad \text{also written} \quad H_0:\ \text{pop 1 \%} = \text{pop 2 \%}$$

and our possible alternative hypothesis will look like

One-Sided

$$H_a:\ \text{pop 1 \%} - \text{pop 2 \%} > 0 \quad \text{also written} \quad H_a:\ \text{pop 1 \%} > \text{pop 2 \%}$$
$$H_a:\ \text{pop 1 \%} - \text{pop 2 \%} < 0 \quad \text{also written} \quad H_a:\ \text{pop 1 \%} < \text{pop 2 \%}$$

Two-Sided

$$H_a:\ \text{pop 1 \%} - \text{pop 2 \%} \neq 0 \quad \text{also written} \quad H_a:\ \text{pop 1 \%} \neq \text{pop 2 \%}$$

Here, since we are assuming the null hypothesis is true when we are running our test, $EV = 0$.

Out test statistic in this case is

$$z = \frac{\text{observed} - EV}{SE} = \frac{\text{observed}}{SE}.$$

The observed value in this case is sample 1 % − sample 2 %.

The SE in this case is a little more complicated. Define the *pooled sample proportion*

$$\hat{p} = \frac{\text{total number of successes in both samples}}{\text{total number of observation in both samples}} \quad \text{or written} \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2},$$

where X_1 denotes the number of successes in sample 1, X_2 denotes the number of successes in sample 2, n_1 is the size of sample 1, and n_2 is the size of sample 2. Note that in general, $n_1 \neq n_2$. With this notation, we have

$$SE = \left(\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \right) \times 100\%$$

Ex: Out of 156 men, 23 liked the movie. Out of 212 women, 123 liked the movie. If we let population 1 be the men and population 2 be the women, then

$$X_1 = 23, \quad n_1 = 156, \quad X_2 = 123, \quad n_2 = 212, \quad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{23+123}{156+212} = \frac{146}{368} = 0.3967$$

Side note (you can skip until the ☺): Where did we get this? For simplicity, let $\hat{p}_1$ denote sample 1 % and $\hat{p}_2$ denote sample 2 %. Instead of using $\hat{p}_1$ and $\hat{p}_2$, we use $\hat{p}$ to calculate the standard error.[5]

$$\sqrt{(SE_{\%}1)^2 + (SE_{\%}2)^2}$$

$$= \sqrt{\left(\frac{\sqrt{\hat{p}_1\,(1-\hat{p}_1)}}{\sqrt{n_1}}\right)^2 + \left(\frac{\sqrt{\hat{p}_2\,(1-\hat{p}_2)}}{\sqrt{n_2}}\right)^2} = \sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$$

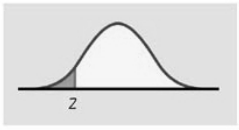
now replace $\hat{p}_1$ and $\hat{p}_2$ with $\hat{p}$

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n_1} + \frac{\hat{p}\,(1-\hat{p})}{n_2}} = \sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad ☺$$

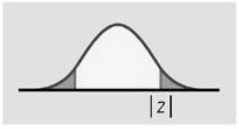
To find the p-value we look for the shaded areas shown below, where p_1 denotes pop 1 % and p_2 denotes pop 2 %. (The picture was stolen from another book.)

H_a: $p_1 > p_2$ is $P(Z \geq z)$

H_a: $p_1 < p_2$ is $P(Z \leq z)$



H_a: $p_1 \neq p_2$ is $2P(Z \geq |z|)$



[5] We have done this not because it is more convenient (it isn't – there's more calculation involved) nor because it reduces the measurement of variability (it doesn't always – often the pooled estimate is larger) but because it gives us the best estimate of the variability of the difference under our null hypothesis that the two sample proportions came from populations with the same proportion.

We use the Method that was given on Page 51 of these notes.

Note: A note on hypothesis tests regarding two population proportions: For the best results, you want to make sure the counts of successes and failures are each 5 or more in both samples. You also need both samples to be SRS's.

Ex: A university financial aid office polled a SRS of undergraduate students to study their summer employment. Not all students were employed the previous summer: 728 of the 817 men were employed and 603 of the 752 women were employed. Is there evidence that the proportion of male students employed differed from the proportion of female students employed?

1. Let population 1 be the men and population 2 be the women. Our null and alternative hypotheses are

$$H_0: \text{ pop 1 \% } - \text{ pop 2 \% } = 0$$
$$H_a: \text{ pop 1 \% } - \text{ pop 2 \% } \neq 0$$

2. We summarize the data

Sample	X	n	Sample %
1	728	817	89%
2	603	752	80%

so that $\hat{p} = \dfrac{X_1 + X_2}{n_1 + n_2} = \dfrac{728 + 603}{817 + 752} = 0.8483$

With the summarized the data, we see that there are 5 successes and failures for each sample. Also, we have

$$SE = \left(\sqrt{0.8483\,(1 - 0.8483)\left(\frac{1}{817} + \frac{1}{752}\right)} \right) \times 100\% = 1.8128\%$$

Our test statistic is

$$z = \frac{\text{observed}}{SE} = \frac{89\% - 80\%}{1.8128\%} \approx 4.96$$

3. Since the alternative hypothesis is $H_a:$ pop 1 % $-$ pop 2 % $\neq 0$, the p-value is the area to the left of -4.96 plus the area to the right of 4.96.

   The largest standard unit on our table is 4.45. The area outside of -4.45 and 4.45 is $100 - 99.9991 = 0.0009$. Since 4.96 is farther away from 0 than 4.45, the area outside of -4.96 and 4.96 is less than 0.0009. So, our p-value < 0.0009

4. Since p-value $< 0.0009 < 0.01$, the results are highly significant. Thus we can reject the null hypothesis in favor of the alternative hypothesis. The proportion of male students employed differs from the proportion of female students employed.

Testing Hypotheses about One Population Mean

We first recall the *sample standard deviation*, given by

$$SD^+ = \sqrt{\frac{(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2}{n-1}} = \left(\sqrt{\frac{n}{n-1}}\right) \times SD$$

where n is the sample size and a is the sample average.

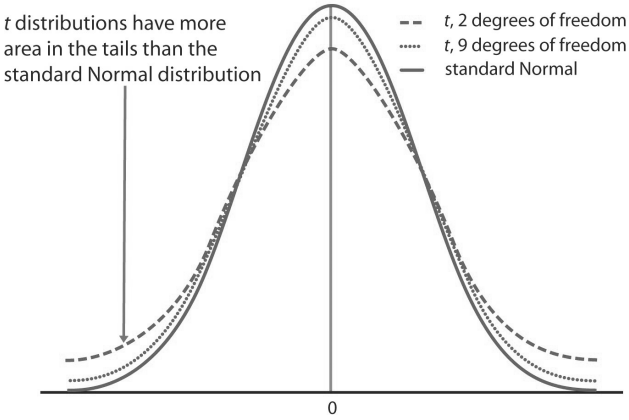
We introduce the t distribution, also called Student's curve. It is similar in shape to the standard Normal curve. They are both symmetric about 0 and bell-shaped.

The spread of the t distributions is a bit greater than that of the standard Normal curve (i.e., the t curve is slightly fatter).

There are different t curves depending on your sample size. We distinguish between these curves using *degrees of freedom* (sometimes written as *df* for shot), which is calculated

$$\text{degrees of freedom} = df = n - 1$$

How do each of these curves differ? And n increases, the t curve gets closer and closer to the standard normal curve.



How to use the table in the back of the book to fins areas under the curve.

Ex: For $df = 6$, what is the area to the right of $t = 1.94$?

Directly from the table, we see the area is 5%.

Ex: For $df = 9$, what is the area to the left of $t = -2.82$ plus the area to the right of $t = 2.82$?

Directly from the table, we see the area to the right of $t = 2.82$ is 1%. Our desired area is $2 \times 1\% = 2\%$.

Ex: For $df = 19$, find lower and upper bounds for the area to the left of $t = -1.95$.

Since the t curve is symmetric, this is tantamount finding the area to the right of $t = 1.95$, which is not in the table. But, $1.73 < 1.95 < 2.09$, so the area to the right of $t = 1.95$ is between 2.5% (which is the area to the right of $t = 2.09$) and 5% (which is the area to the right of $t = 1.73$). The bounds are 2.5% and 5%.

Our null hypothesis will look like

$$H_0 : \text{ pop average } = EV$$

and our possible alternative hypothesis will look like

One-Sided

$$H_a : \text{ pop average } > EV$$

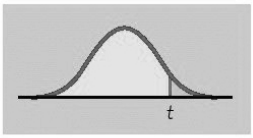
$$H_a : \text{ pop average } < EV$$

Two-Sided

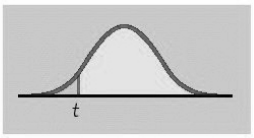
$$H_a : \text{ pop average } \neq EV$$

I stole the picture below from another book. The p-value is the shaded area under the t curve. To understand the notation, $\mu =$ pop average and $\mu_0 = EV$.

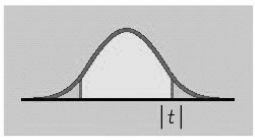
H_a: $\mu > \mu_0$ is $P(T \geq t)$



H_a: $\mu < \mu_0$ is $P(T \leq t)$



H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$



The test statistic is

$$t = \frac{\text{observed} - EV}{SE} \qquad \text{where } SE = \frac{SD^+}{\sqrt{n}}$$

We use the Method that was given on Page 51 of these notes.

Note: A note on hypothesis tests regarding one population mean: The p-values are exact if the population distribution is normal and approximately correct for large n in other cases.

If you have $n > 25$, you can use the normal approximation. If $n < 25$ and the histogram of your sample is close to normal (clustered together and not skewed), then you can use the t-distribution. If you sample data is very skewed, you cannot trust your results from the t-distribution.

Note: For simplicity in this class, we will assume that all our data is "nice enough" to use these methods, but in practice in the real world, you need to check your data to make use it is not skewed. (There are things called dotplot, boxplot, stem-and-leaf plot, and histogram you can use to check for skewness.)

Ex: The value 98.6 degrees seems to have come from determining the mean in degrees Celsius, rounding up to the nearest whole degree (37 degrees), and then converting that number to Fahrenheit using $32+(1.8)(37) = 98.6$. Rounding up may have produced a result higher than the actual average, which may therefore be lower than 98.6 degrees. With a sample of $n = 18$ body temperatures

98.2	97.8	99.0	98.6	98.2	97.8	98.4	99.7	98.2
97.4	97.6	98.4	98.0	99.2	98.6	97.1	97.2	98.5

we have a sample average of 98.217 and sample standard deviation $SD^+ = 0.684$. Is the population body temperature less than 98.6 degrees Fahrenheit?

Using the Method:

1. The null and alternative hypothesis are:

$$H_0: \text{ pop average } = 98.6$$
$$H_a: \text{ pop average } < 98.6$$

   Here, $EV = 98.6$. We are assuming the null hypothesis is true. So we are assuming that the true population average is 98.6.

2. The observed value is 98.217, and the $SE = \frac{0.684}{\sqrt{18}} = 0.161$. If you made a histogram of the 18 data points, it would mostly symmetric – only barely skewed to the right. The test statistic is

$$t = \frac{\text{observed} - EV}{SE} = \frac{98.217 - 98.6}{0.161} = -2.38$$

3. We have $df = n - 1 = 17$. Since the alternative hypothesis is H_a : pop average < 98.6, the p-value is the area to the left of $t = -2.38$, which is tantamount to the area to the right of $t = 2.38$. We have $2.57 < t < 2.90$ so that our p-value is between 0.5% and 1%.

4. Since p-value $< 1\%$, we can reject the null hypothesis in favor of the alternative hypothesis. We can conclude, on the basis of these data, that the mean temperature in the human population is actually less than 98.6 degrees.

Ex: The national mean length cell phone call is 9.2 minutes. A cell phone company believes that in a certain city, the mean length of cell phone calls is different. A SRS of 200 of current customers yields a mean and standard deviation of 8 minutes and 10 minutes, respectively. Is there evidence in the data to support the phone companys claim?

Using the Method:

1. The null and alternative hypothesis are:

$$H_0: \text{ pop average } = 9.2$$
$$H_a: \text{ pop average } \neq 9.2$$

   Here, $EV = 9.2$. We are assuming the null hypothesis is true. So we are assuming that the true population average is 9.2.

2. The observed value is 8, and the $SE = \frac{10}{\sqrt{200}} = 0.707$. The test statistic is

$$z = \frac{\text{observed} - EV}{SE} = \frac{8 - 9.2}{0.707} = -1.70$$

   We use z since our sample size is over 25.

3. Since the alternative hypothesis is $H_a:$ pop average $\neq 9.2$, the p-value is the area to the left of $z = -1.70$ plus the right of $z = 1.70$. The area is $100\% - 91.09\% = 8.91\%$.

4. Since p-value =8.91%> 5%, we cannot reject the null hypothesis. There is not sufficient evidence to conclude that the mean length of calls in this city is different from the mean for the national population.

Testing Hypotheses About the Difference in Two Population Means

The parameter of interest is pop 1 average − pop 2 average. We wish to test if this is different from 0.

The statistic of interest is sample 1 average − sample 2 average. We sill use this in our test.

Our null hypothesis will look like

$$H_0: \text{ pop 1 average } - \text{ pop 2 average } = 0,$$

also written

$$H_0: \text{ pop 1 average } = \text{ pop 2 average},$$

and our possible alternative hypothesis will look like

One-Sided

$$H_a: \text{ pop 1 average } - \text{ pop 2 average } > 0$$
$$H_a: \text{ pop 1 average } - \text{ pop 2 average } < 0$$

Two-Sided

$$H_a: \text{ pop 1 average } - \text{ pop 2 average } \neq 0,$$

which can be rewritten as

One-Sided

H_a : pop 1 average $>$ pop 2 average

H_a : pop 1 average $<$ pop 2 average

Two-Sided

H_a : pop 1 average $\neq$ pop 2 average,

respectfully. Here, since we are assuming the null hypothesis is true when we are running our test, $EV = 0$.

Out test statistic in this case is

$$t = \frac{\text{observed} - EV}{SE} = \frac{\text{observed}}{SE}.$$

The observed value in this case is sample 1 average − sample 2 average.

The SE in this case is a little more complicated. Define

$$SE^+ = \frac{SD^+}{\sqrt{n}},$$

which is the standard error based on the sample standard deviation. Thus we have

$$SE = \sqrt{(SE^+1)^2 + (SE^+2)^2}$$

where $SE^+1 = \frac{SD^+1}{\sqrt{n_1}}$ and $SE^+2 = \frac{SD^+2}{\sqrt{n_2}}$ are the sample standard errors for sample 1 and sample 2, respectfully.

To find the p-value, the area under the t-curve mimics what we do for the difference in two population percentages (see Page 54 of these notes).

Since we have two samples, what do we use for the degrees of freedom? We use the smaller of $n_1 - 1$ and $n_2 - 1$. This is a conservative approach.[6] If both of your sample sizes are bigger than 25, then (in this class) you use the standard normal curve.

We use the Method that was given on Page 51 of these notes.

Note: To be able to run this test, verify that both n's are large or that there are no extreme outliers or skewness in either sample. The samples must also be

[6] A more accurate formula is

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2},$$

where, for brevity, $s_1 = SD^+1$ and $s_2 = SD^+2$. This is called *Welch's approximation*, and we will not use this formula in this class. But it is pretty.

independent and SRS's.

Ex: Do you think that the mean time watching television is different for the population of college men than it is for the population of college women? A summary of a survey of 35 Penn State students is given below:

Gender	n	Average	SD^+
Male	19	2.37	1.87
Female	16	1.95	1.51

1. Let population 1 be the men and population 2 be the women. Our null and alternative hypotheses are

$$\begin{aligned} H_0: &\text{ pop 1 average } - \text{ pop 2 average } = 0 \\ H_a: &\text{ pop 1 average } - \text{ pop 2 average } \neq 0 \end{aligned}$$

2. Assuming that our data is not skewed, we calculate

$$\begin{aligned} SE^+1 &= \frac{SD^+1}{\sqrt{n_1}} = \frac{1.87}{\sqrt{19}} \approx 0.42901 \\ SE^+2 &= \frac{SD^+2}{\sqrt{n_2}} = \frac{1.51}{\sqrt{16}} \approx 0.37750 \end{aligned}$$

   Our standard error is

$$SE = \sqrt{(0.42901)^2 + (0.37750)^2} \approx 0.57145$$

   Our test statistic is

$$t = \frac{\text{observed}}{SE} = \frac{2.37 - 1.95}{0.57145} \approx 0.7350$$

3. Since the alternative hypothesis is $H_a:$ pop 1 average $-$ pop 2 average $\neq 0$, the p-value is the area to the left of -0.7350 plus the area to the right of 0.7350, which is tantamount to

$$2\times \text{ (the area to the right of } 0.7350).$$

   We use the smaller of $n_1 - 1$ and $n_2 - 1$ for df, which is 15. Since $t = 0.7350$ is between 0.69 and 1.34 on the row $df = 15$, our p-value is between

$$2\times 10\% \text{ and } 2\times 25\% \qquad \text{or} \qquad 20\% \text{ and } 50\%.$$

4. Since our p-value is much greater than 5%, we cannot reject the null value. On the basis of these samples, there is insufficient evidence to conclude that the mean population television viewing hours for college men and women are different.

Ex: These data were obtained in a study comparing persons with disabilities with persons without disabilities. A scale known as the Barriers to Health Promotion Activities for Disabled Persons (BHADP) Scale gave the data. We wish to know if we may conclude, at the 1% significance level, that persons with disabilities score higher than persons without disabilities.

Population	n	Average	SD^+
Disabled	132	31.83	7.93
Non-disabled	137	25.07	4.80

1. Let population 1 be disabled and population 2 be non-disabled. Our null and alternative hypotheses are

$$H_0: \text{ pop 1 average } - \text{ pop 2 average } = 0$$
$$H_a: \text{ pop 1 average } - \text{ pop 2 average } > 0$$

2. Assuming that our data is not skewed, we calculate

$$SE^+1 = \frac{SD^+1}{\sqrt{n_1}} = \frac{7.93}{\sqrt{132}} \approx 0.69021$$
$$SE^+2 = \frac{SD^+2}{\sqrt{n_2}} = \frac{4.80}{\sqrt{137}} \approx 0.41009$$

   Our standard error is

$$SE = \sqrt{(0.69021)^2 + (0.41009)^2} \approx 0.80285$$

   Since both sample sizes are greater than 25, we use the normal curve, and our test statistic is

$$z = \frac{\text{observed}}{SE} = \frac{31.83 - 25.07}{0.80285} \approx 8.42000$$

3. Since the alternative hypothesis is $H_a:$ pop 1 average $-$ pop 2 average > 0, the p-value is the area to the right of $z = 8.42$. Since $z = 8.42$ is off the chart, the area to the right is virtually 0%.

4. Since our p-value $< 1\%$, the test is highly significant. We can reject the null hypothesis in favor of the alternative hypothesis. On the basis of these data, the average persons with disabilities score higher on the BHADP test than do the non-disabled persons.

17 Confidence Intervals for Averages for Sample Smaller than 25

The formulas are the same except we use t^* instead of z^* and SE^+ instead of SE. Note that we are using the sample standard deviation SD^+ for these

computations.

Ex: Suppose that the forearm lengths (in centimeters) for a randomly selected sample of $n = 9$ men are as follows:

$$25.5,\ 24.0,\ 26.5,\ 25.5,\ 28.0,\ 27.0,\ 23.0,\ 25.0,\ 25.0$$

Find a 90% CI for the population average.

The sample average is 25.5 and $SD^+ = 1.52$. Also, $df = 9 - 1 = 8$. For a 90% CI, the area in the right tail is 5%. Looking at the t-table on the row $df = 8$ and the column 5%, we use $t^* = 1.86$. Our CI is

$$\text{sample average} \pm t^* SE^+ \quad \Rightarrow \quad 25.5 \pm 1.86 \cdot \frac{1.52}{\sqrt{9}} \quad \Rightarrow \quad 24.5576\text{cm to } 26.4424 \text{ cm}$$

We are 90% confident that the average length of a man's forearm is between 24.5576 cm and 26.4424 cm.

For a CI for the difference between two population averages when at least one of the two sample sizes is less than 25, the formula is

$$(\text{sample 1 average} - \text{sample 2 average}) \pm t^* \sqrt{(SE^+1)^2 + (SE^+2)^2}$$

where

$$\begin{aligned} SE^+1 &= \frac{SD^+ \text{ of sample 1}}{\sqrt{n_1}} \quad \text{and} \\ SE^+2 &= \frac{SD^+ \text{ of sample 2}}{\sqrt{n_2}} \end{aligned}$$

The df is the smaller of $n_1 - 1$ and $n_2 - 1$.

18 Test of Significance for Correlations

We use ρ to denote *population linear correlation*; that is, if we have a pair (x, y) for everyone in the population, what would the correlation ρ be?

Once you have calculated the sample linear correlation coefficient, r, you will want to determine whether the population linear correlation, ρ, is significant. You can do this by performing a hypothesis test. A hypothesis test for ρ can be one tailed or two tailed.

Our null hypothesis will look like

$$H_0 : \rho = 0$$

and our possible alternative hypothesis will look like

$$\begin{array}{c} \text{One-Sided} \\ H_a : \rho > 0 \\ H_a : \rho < 0 \\ \text{Two-Sided} \\ H_a : \rho \neq 0 \end{array}$$

(More often a two-sided test is used than a one-sided test.)

We will use the the t-distribution, and our test statistic is

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

where $df = n - 2$ is the number of degrees of freedom.

The appropriate areas for the p-value are the same as on page 57. And we compare our p-value with 5% and 1% for significance.

Ex: The correlation between a tree-ring index and annual precipitation is $r = 0.48$. The sample size is $n = 18$. We test

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho > 0 \end{aligned}$$

Our test statistic is

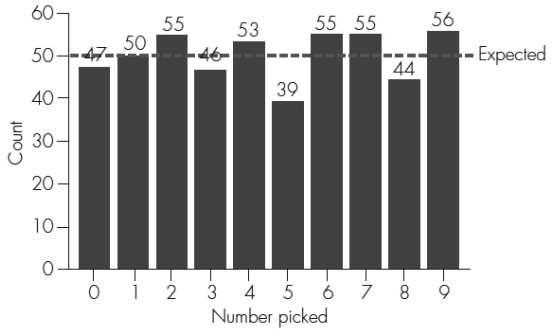
$$t_{16} = t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.48 \cdot \sqrt{18-2}}{\sqrt{1-(0.48)^2}} \approx 2.1886.$$

We have $df = n - 2 = 16$, and the $t = 2.19$ is between 2.12 and 2.58. Thus, our p-value, which is the area to the right of $t = 2.19$, is between 1% and 2.5%. With this p-value, we can reject H_0. There is a positive correlation in the population.

19 Chi-Square Test

Goodness-Of-Fit Test

Ex: The daily lottery draws three digits. For example, if 3, 6, and 3 are drawn, the number for the day is 363. We will focus on the first digit, which could be any number from 0 through 9. If these number are equally likely, the probability for each number is 10%. We keep track of the draws from the first container for the $n = 500$ days between July 19, 1999, and November 29, 2000. The expected count for each of the 10 possible outcomes is $0.10 \times 500 = 50$, and the actual observed counts were



We wish to know if these observed counts are with reason and are due to chance, or if there is something more going on here – something nefarious. Maybe, not enough 5's?

We introduce a significance test called the *chi-square goodness-of-fit test*:
We have k categories of a categorical variable.
The Mothod:

1. The hypotheses:

   H_0 : The probabilities for the k categories are given by $p_1, p_2, \ldots, p_k$.
   H_a : Not all the probabilities specified in H_0 are correct.

   (Note: $p_1 + p_2 + \cdots + p_k = 1$)

2. Calculate chi-square statistic:

$$\chi^2 = \text{sum of } \left(\frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} \right)$$

3. Assuming that the null hypothesis is true, find the p-value. Find the p-value using the chi-square table using the appropriate degrees of freedom given by

$$df = \text{ number of categories } - 1 = k - 1.$$

   The area in the tables is the area to the right of the χ^2 number.

4. Check you p-value with 5% and 1%.

Note: You need a SRS, and the sample needs to be large enough so that every expected count is at least 5.

Ex: Test the daily lottery example from above.

1. The hypotheses are

   H_0 : Probability of 10% for each of the 10 possible digits in first container.
   H_a : The probability of one digit is not 10%.

2. For the chi-square statistic, observed frequency is what we saw for each number over the 500 days, and the expected frequency is $.10 \times 500 = 50$. We calculate

$$\begin{aligned}
\chi^2 &= \text{sum of } \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}} \right) \\
&= \frac{(47-50)^2}{50} + \frac{(50-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(46-50)^2}{50} + \frac{(53-50)^2}{50} \\
&\quad + \frac{(39-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(44-50)^2}{50} + \frac{(56-50)^2}{50} \\
&= 6.04
\end{aligned}$$

3. Since there are 10 digits, there are $k = 10$ possible outcomes so that $df = 10 - 1 = 9$. Looking on the Chi-Square table on row 9, $\chi^2 = 6.04$ is between 4.17 and 6.39. Thus, the p-value, the area to the right of $\chi^2 = 6.04$, is between 70% and 90%.

4. Since, p-value $> 5\%$, the result is not statistically significant; the null hypothesis is not rejected.

Ex: As recently as 2008, 70% of users of social networking sites such as Facebook were 35 years old or younger. Now the age distribution is much more spread out. The age distribution of 975 users of social networking sites from a survey reported is June 2011 is below:

Age	18-22	23-35	36-49	50-65	65+
Frequency	156	312	253	195	59

We will test the assumption that each of the users are equally likely to be in each of the five age groups.

1. Since there are $k = 5$ groups, then there is a $\frac{1}{5} = 0.2 \Rightarrow 20\%$ to be in each group. The hypotheses are

   H_0 : Probability of 20% for each container.
   H_a : The probability of one container is not 20%.

2. For the chi-square statistic, observed frequency is what we saw for each number over the 500 days, and the expected frequency is $.20 \times 975 = 195$. We calculate

$$\begin{aligned}\chi^2 &= \text{sum of } \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right) \\ &= \frac{(156-195)^2}{195} + \frac{(312-195)^2}{195} + \frac{(253-195)^2}{195} + \frac{(195-195)^2}{195} \\ &\quad + \frac{(59-195)^2}{195} = 7.8 + 70.2 + 17.3 + 0 + 94.9 = 190.2\end{aligned}$$

3. Since there are 5 groups, there are $k = 5$ possible outcomes so that $df = 5 - 1 = 4$. Looking on the Chi-Square table on row 4, $\chi^2 = 190.2$ is to the right of 13.28. Thus, the p-value, the area to the right of $\chi^2 = 190.2$, is basically 0.

4. Since , p-value $< 1\%$, the result is highly significant; we can reject the null hypothesis in favor of the alternative hypothesis. There is strong evidence that users of social networking sites are not equally distributed among these age groups.

Ex: Is birth date related to lunar cycle? We have the data below

Category	Lunar Phase	Number of days	Number of Births
1	New moon	24	7,680
2	Waxing crescent	152	48,442
3	First quarter	24	7,579
4	Waxing gibbous	149	47,814
5	Full moon	24	7,711
6	Waning gibbous	150	47,595
7	Last quarter	24	7,733
8	Waning crescent	152	48,230
	Total	699	222,784

If there is no relationship between the number of births and the lunar cycle, then the number of births in each lunar cycle category should be proportional to the number of days included in that category. That is,

$$p_1 = \tfrac{24}{699} = 0.0343, \qquad p_2 = \tfrac{152}{699} = 0.2175,$$

$$p_3 = \tfrac{24}{699} = 0.0343, \qquad p_4 = \tfrac{149}{699} = 0.2132,$$

$$p_5 = \tfrac{24}{699} = 0.0343, \qquad p_6 = \tfrac{150}{699} = 0.2146,$$

$$p_7 = \tfrac{24}{699} = 0.0343, \qquad p_8 = \tfrac{152}{699} = 0.2175$$

Our null hypothesis H_0 is that the probabilities are as above. Our alternative hypothesis H_a is that not all the probabilities in H_0 are correct.

Using the proportions above, we can get the expected counts with $n = 222,784$. To calculate this for each category, we multiply the total number of births by the proportion for that category: np_k. For example

$$\begin{aligned} np_1 &= 222,784 \cdot 0.0343 = 7641.49 \\ np_2 &= 222,784 \cdot 0.2175 = 48455.52 \\ &\vdots \end{aligned}$$

We have

Category	Lunar Phase	Observed Number of births	Expected Number of Births
1	New moon	7,680	7,641.49
2	Waxing crescent	48,442	48,455.52
3	First quarter	7,579	7,641.49
4	Waxing gibbous	47,814	47,497.55
5	Full moon	7,711	7,641.49
6	Waning gibbous	47,595	47,809.45
7	Last quarter	7,733	7,641.49
8	Waning crescent	48,230	48,455.52

We calculate the chi-square statistic:

$$\begin{aligned} \chi^2 &= \text{sum of } \left(\frac{(\text{observed} - \text{expected})^2}{\text{expected}} \right) \\ &= \frac{(7,680 - 7,641.49)^2}{7,641.49} + \frac{(48,442 - 48,455.52)^2}{48,455.52} + \cdots + \frac{(48,230 - 48,455.52)^2}{48,455.52} \\ &= 0.194 + 0.004 + 0.511 + 2.108 + 0.632 + 0.962 + 1.096 + 1.050 \\ &= 6.557 \end{aligned}$$

With $df = 8 - 1 = 7$, we see that $\chi^2 = 6.557$ is between 6.35 and 8.38. So, our p-value is between 30% and 50%.

Since out p-value $> 5\%$, our test is not significant, and there is not sufficient evidence to conclude that birth date and lunar cycle are related.

Tests for Independence in a Two-Way Table

Chi-square test can be used to investigate association between two categorical variables in a single population. When there is an association, knowing the value of one variable provides information about the value of the other variable. Where there is no association between two categorical variables, they are said to be independent.

Ex: This is related to a paper that studied the relationship between a nurse's assessment of patients with dementia and their facial expressions associated with pain and the patient's self-reported level of pain. There were 89 patients studied.

An example of a two-way table is the 2×2 table below (row×column). There are 2 rows and 2 columns – counting the non-bold numbers.

	Self-Report		
Facial Expressions	No Pain	Pain	Row Total
No Pain	17	40	**57**
Pain	3	29	**32**
Column Total	**20**	**69**	**89**

The non-bold numbers are called *observed cell counts*. There were 89 total patients studied, and 20 were self-reported no pain. The proportion of self-reported no pain is

$$\frac{20}{89} = 0.2247.$$

If there were no difference in response for the different groups of facial expressions, you would expect about

- 22.47% of no pain facial expressions to have self-responded no pain.
- 22.47% of pain facial expressions to have self-responded no pain.

The expected count for no pain facial expressions *and* self-reported no pain cell =(0.2247)(57)=12.81

The expected count for pain facial expressions *and* self-reported no pain cell =(0.2247)(32)=7.19

In general, the *expected cell count* for each cell is given by

$$\text{expected cell count} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

We can rewrite out table with the expected cell counts written in parentheses

	Self-Report		
Facial Expressions	No Pain	Pain	Row Total
No Pain	17(12.81)	40(44.19)	**57**
Pain	3(7.19)	29(24.81)	**32**
Column Total	**20**	**69**	**89**

For an example of how we calculated some of these, the second column is

$$44.19 \approx \frac{57\times69}{89} \qquad 24.81 \approx \frac{32\times69}{89}$$

after rounding.

The Method (for chi-square test for independence):

1. Hypotheses:

   H_0 : The two variables are independent
   (or equivalently, there is no association between the two variables)
   H_a : The two variables are not independent
   (or equivalently, there is an association between the two variables)

2. The chi-square statistic in this situation is

$$\chi^2 = \text{sum of} \left(\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \right)$$

3. You look up the p-value in the chi-square table using

$$df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

   So, if you have a two-way table of size $m \times n$, you have $df = (m-1) \times (n-1)$.

4. Conclusion. Check your p-value with 5% and 1%.

Note: Need one SRS and a sample large enough so that every expected count is at least 5. (If some expected counts are less than 5, rows and columns of the table may be combined to achieve a table with satisfactory expected counts.)

Ex: For the table above, our hypotheses are

H_0 : Facial expression and self-reported pain are independent.
H_a : Facial expression and self-reported pain are not independent.

We calculate the chi-square statistic:

$$\chi^2 = \frac{(17-12-81)^2}{12.81} + \frac{(40-44.19)^2}{44.19} + \frac{(3-7.19)^2}{7.19} + \frac{(29-24.81)^2}{24.81} = 4.92$$

and we have $df = (2-1) \times (2-1) = 1$. On the 4th row of the chi-square table, $\chi^2 = 4.92$ is between 3.84 and 6.64 so that our $1\% < p\text{-value} < 5\%$. Thus, our test is significant, and we can reject our null hypothesis. There is convincing evidence of an association between a nurse's assessment of facial expressions and self-reported pain.

Ex: Is there an association between survival after a stroke and level of education? Medical records of 2,333 people are examined and shown below with expected counts in parentheses. Test at the 1% significance level.

	No Basic Education	Secondary School Graduation	Technical Training/ Apprenticed	Higher Secondary School Degree	University Graduate
Died	13(17.40)	91(77.18)	196(182.68)	33(41.91)	36(49.82)
Survived	97(92.60)	397(410.82)	959(972.32)	232(223.09)	279(265.18)

The row totals are 369 and 1964, and the first column total is 110. The first column expected counts are

$$\frac{369\times 110}{2333} = 17.4 \qquad \text{and} \qquad \frac{1964\times 110}{2333} = 92.60.$$

The hypotheses are

H_0 : Survival and level of education are independent.

H_a : Survival and level of education are not independent.

We calculate

$$\begin{aligned}\chi^2 &= \frac{(13-17.40)^2}{17.40} + \frac{(91-77.18)^2}{77.18} + \frac{(196-182.68)^2}{182.68} + \cdots + \frac{(279-265.18)^2}{265.18} \\ &= 12.219\end{aligned}$$

Also, $df = (2-1)\times(5-1) = 4$. We see that $\chi^2 = 12.219$ is between 9.49 and 13.28 so that our p-value is between 1% and 5%. Since we are testing at the 1% significance level, we cannot reject the null hypothesis. Thus, there is not sufficient evidence to conclude that an association exists between level of education and survival.

20 Some Details about Hypothesis Testing

Recall: Assuming that the null hypothesis is true, the probability that the test statistic would take a value as extreme or more extreme than the value actually observed is called the p-value of the test.

If the null hypothesis is true, there is a 5% chance of getting a difference which the test will call "statistically significant."

If the null hypothesis is true, there is a 1% chance of getting a difference which the test will call "highly significant."

So, it's possible that the null hypothesis could be true, but by pure chance, you get a sample with such extreme qualities that your test statistic yields a small enough p-value for you to reject the null hypothesis. In this case, the sample percentage (or average) is very extreme and far away from the population percentage (or average). In most samples, the sample percentage would be close to the population percentage.

Definition: A *type I error* is made when we reject the null hypothesis when in fact the null hypothesis is true.

Ex: As an example, we want to test cat weights and have

H_0 : pop average = 8lbs

H_a : pop average > 8lbs

Suppose, unbeknownst to us, that the true population average is 8lbs (so H_0 is true), but the SRS of cats we gather just happens to consists of the fattest cats in the neighborhood, which gives a sample average of 10.2lbs. This sample average gives us statistical significance. So, we reject H_0, even though H_0 is true: a type I error.

Ex: Suppose you ran 100 hypothesis tests with SRS of size n from the same population where the null hypothesis is true. With a significant level of 5%, you should expect about 5 tests to yield a p-value small enough for you to reject the null hypothesis. That's about 1 out of every 20 tests. With a significant level of 1%, you should expect 1 test to yield a p-value small enough for you to reject the null hypothesis. *CLICK HERE!*

On the other hand, It's possible that the alternative hypothesis could be true, but by pure chance, you get a sample with such extreme qualities that your test statistic yields a large enough p-value so you cannot reject the null hypothesis. In this case, the sample percentage (or average) is very extreme and close to the EV. In most samples, the sample percentage would be far away from the EV.

Definition: A *type II error* is made when we fail to reject the null hypothesis when in fact the alternative hypothesis is true.

Ex: As an example, we want to test dog weights and have

$$H_0 : \text{pop average} = 40\text{lbs}$$
$$H_a : \text{pop average} > 40\text{lbs}$$

Suppose, unbeknownst to us, that the true population average is 44.5bs (So H_a is true), but the SRS of dogs we gather just happens to consists of the smaller dogs in the neighborhood, which gives a sample average of 41.2lbs. This sample average does not give us statistical significance. So, we cannot reject H_0, even though H_a is true: a type II error.

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Cannot Reject H_0	Correct decision	Type II error

Ex: Medical Tests. Imagine that you are tested to determine whether you have a disease. The lab technician or physician who evaluates your results must

make a choice between two hypotheses:

$$H_0 : \text{You do not have the disease.}$$
$$H_a : \text{You have the disease.}$$

Unfortunately, many laboratory tests for diseases are not 100% accurate. There is a chance that the result is wrong.

- A type 1 error occurs when the diagnosis is that you have the disease when actually you do not. In other words, a type 1 error is a *false positive*.
- A type 2 error occurs when the diagnosis is that you do not have the disease when actually you do. In other words, a type 2 error is a *false negative*.

Definition: When the alternative hypothesis is true, the probability of making the correct decision is called the *power* of a test.

- *The power increases when the sample size is increased.* This makes sense because when the sample size is increased, the standard error is decreased, leading to larger values of the test statistic. Also, the sample statistic is a more accurate estimate of the population value, making it easier to detect a difference between the true population value and the null value.
- *The power increases when the difference between the true population value and the null hypothesis value increases.* This makes sense because the probability of detecting a large difference is higher than the probability of detecting a small difference. However, remember that the truth about the population is not something that the researcher can control or change.

The probability of a type I error is equal to the level of significance.
The probability of a type II error is equal to one minus the power (as a decimal) of the test.

Statistical Significance vs. Practical Significance

- When the sample size is very large, tiny deviations from the null hypothesis (with little practical consequence) will be statistically significant.
- When the sample size is very small, large deviations from the null hypothesis (of great practical importance) might go undetected (statistically insignificant).

Ex: An article describes an Austrian study of the heights of 507,125 military recruits. In an article in the journal Nature, the researchers reported their finding that men born in the spring were, on average, about 0.6 centimeter taller than men born in the fall. This is a small difference; 0.6 centimeter is only about $1/4$ of an inch. The sample size for the study is so large that even a very small

difference will earn the title statistically significant. Do you think the practical import of this difference warranted the headline?

Ex: Small samples are taken. The sample percentages show that 33% of men and 36% of women favor Candidate X, but when we run a test for a difference of population percentages, the test is not significance (because of the small sample sizes). Even though 33% and 36% is a practical significance, the test is not statistically significance.

Mail to rstephens@colgate.edu