2020

Clear-Sighted Statistics: Module 11: Confidence Intervals

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 11: Confidence Intervals

> Too many business people assign equal validity to all numbers printed on paper. They accept numbers as representing the Truth and find it difficult to work with the concept of probability. They do not see a number as a kind of shorthand for range that describes our actual knowledge of the underlying condition. -- Arthur C. Nielsen, Jr.[1]

I. Introduction

In the mid-1930s, Jerzy Neyman, one of the most important statisticians of the twentieth century, laid the foundations for confidence intervals. Confidence intervals deal with the issue of estimating unknown population parameters using sample statistics and probability. Confidence intervals are very important when we consider questions like:

- How much debt has the average college student incurred?
- How many high school and middle school students vape?
- What proportion of the American public supports stricter gun control?
- What proportion of Americans support more restrictions on a woman's right to an abortion?
- What proportion of automobile buyers intend to purchase an electric or hybrid car in the next four years?
- What proportion of people who have health insurance from their employer are satisfied with their insurance plan?
- When researchers conduct repeated experiments under identical conditions that yield varying results, how do researchers know the true parameter?[2]

We do *not* know the answer to these questions with 100 percent certainty. As the American physicist Richard Feynman wrote, "Nature permits us to calculate only probabilities. Yet science has not collapsed."[3] The answers to the questions posed above can be answered

using Neyman's innovation: ***Confidence Intervals***. Confidence intervals are based on sample statistics and probability theory.

For Neyman, the solution to not knowing the true value of parameters is to estimate these unknowns using sample statistics to construct confidence intervals, which are a range of sample values likely to contain the population parameter. When a population cannot be studied "exhaustively," when conducting a census is not a realistic option, Neyman argued, "It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population…." For Neyman, the important questions statisticians address are matters of estimation using confidence intervals.[4]

Because confidence intervals are based on sample statistics and probability, the numbers derived from them are, as A. C. Nielsen Jr. so clearly put it, "shorthand for [a] range that describes our actual knowledge of the underlying condition." Failure to understand confidence intervals can lead to foolish interpretations of the data.

In this module, we will discuss the problem of estimating unknown parameters. The two parameters we will focus on are the population mean, μ, and the population proportion, π. We will use the sample mean, $\bar{X}$, and the sample proportion, p, to estimate these unknown parameters. The resulting estimates based on confidence intervals provide the *approximate* value of the parameter. **Please note:** π is not the same π you learned about in your geometry class, 3.14159265359. Some people use "p" to symbolize the population proportion. The sample proportion is symbolized as "p" although some use "$\hat{p}$," which is called p-hat. In addition, the term 1 – p is sometimes symbolized as $\hat{q}$ (q-hat). We will follow the convention of using Greek letters to represent parameters.

Here is a typical situation: The National Center for Disease Control and Prevention and Health Promotion wants to know the extent to which teenagers and pre-teens use tobacco products. Conducting a census among this population is not an affordable option. Instead this organization conducts surveys to construct confidence intervals to estimate the unknown parameters so that it can fulfill its mission of designing, implementing, and evaluating comprehensive programs to curb the use of tobacco.[5]

After completing this module, you will understand confidence interval basics. You will be able to:

- Define confidence intervals, point estimates, levels of confidence, upper and lower confidence limits, and the margin of error (MoE).
- Construct confidence intervals for μ when σ is known using z-values.
- Construct confidence intervals for μ when σ is unknown using the student-t distribution.
- Construct confidence intervals for a proportion, π, using z-values.
- Understand the application of the finite population correction factor (FPC).

There are a number of files that accompany this module that you should download. They are:

- 11_Examples.xlsx
- 11_Exercises.xlsx
- 11_GallupMarijuana.xlsx
- Student-t_tables.pdf
- Student-t_tables.xlsx
- z-values_AreaBetweenMean&X.pdf

- z-values_AreaBetweenMean&X.xlsx

Links to these files are shown in this module.

II. Key Terms Used When Working With Confidence Intervals

With confidence intervals there is new vocabulary:

1) Confidence Interval (CI)

A confidence interval or *interval estimate* is a range of values obtained from a sample that is likely to contain the parameter we seek to estimate. Confidence intervals contain the actual parameter in the *long run a certain percentage of the time*. The confidence intervals we will construct are two-sided or two-tailed intervals that are symmetrical around the sample statistic.

Confidence intervals convey a great deal of information. They highlight values that are outside the interval, but they cannot predict values within the interval. Because of their random nature, it is unlikely that two samples from a given population will yield identical confidence intervals. But, over time a large proportion of the confidence intervals constructed from the same population will contain the parameter.

2) Point Estimate

The point estimate, or simply *estimate* or *estimator*, is the sample statistic, $\bar{X}$ or p, upon which the confidence interval is based. The confidence intervals we will focus on are split evenly around the point estimate; which is to say, half the interval will be above the point estimate and half will be below.

3) Confidence Level (Level of Confidence or CL)

Confidence intervals are probabilistic statements. The confidence level states the degree of certainty we have that our estimate that the unknown parameter is included in the

confidence interval. We use a 95 percent level of confidence most frequently, which means that we are 95 percent certain that the population parameter is contained in the confidence interval. Occasionally a 99 percent confidence interval is used, which means that we are 99 percent certain that the confidence interval contains the parameter. A 99 percent confidence interval is wider than one constructed using a 95 percent confidence level. We may even see, on relatively rare occasions, a 90 percent confidence level. Confidence intervals drawn using a 90 percent confidence level are narrower than those using a 95 or 99 percent confidence level as shown in Figure 1. Strictly speaking, the confidence level does not mean there is a 90, 95, or 99 percent probability that the population parameter is contained in the interval. The confidence level is the percentage of similarly constructed intervals that will contain the estimated parameter.

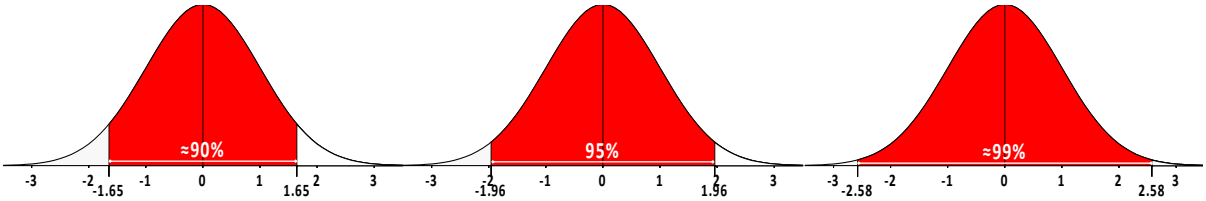

Figure 1: Confidence Interval get wider as the level of confidence increases

The confidence level sets the level of uncertainty we can tolerate. Can we ever be 100 percent certain? Yes, but the confidence interval is too wide to provide us with useful information. **Here is an example:** I am 100 percent certain I know when you are going to die! It will be before you start reading the next sentence. Or, by the time you are 969-years-old, the age Methuselah died according to the book of *Genesis in the Old Testament*.[6] You may recall that Methuselah lived longer than anyone in the *Bible*. This confidence interval, however, is absolutely useless in predicting how long you will actually live. While we can be 100 percent certain that this is the confidence interval for how long you will live, it is useless in predicting how long you will actually live. By the way, my doctor says that the

upper limit of a human life is about 120 years. So even Enoch, Methuselah's father, who lived a mere 365 years, enjoyed a longevity that was three times longer than the outer limits of human life expectancy that modern medical science has established.[7]

Closely related to the confidence level is the *significance level* or *alpha*, α, which we will discuss in detail when we get to null hypothesis significance testing. The significance level is the inverse of the confidence level. Significance levels are critically important to Null Hypothesis Testing, which is often considered the inverse of confidence intervals. You need to be familiar with significance levels because Microsoft Excel's confidence interval functions require them and not confidence levels.

Significance levels are found by this simple formula:

$$\text{Significance Level} = 1 - \text{Confidence Level}$$

Equation 1: Formula for Significance Levels

Table 1 shows the significance levels that correspond to 95 percent, 99 percent, and 90 percent levels of confidence.

Table 1: Significance Levels are the Inverse of Confidence Levels

Confidence Level (CL)	Significance Level (α)	Found By
95%	5%	1 – CL
99%	1%	1 – CL
90%	10%	1 – CL

4) Margin of Error (MoE)

The margin of error is the width of the confidence interval, that is, the distance the confidence interval extends above and below the point estimate. In opinion polls, the MoE are reported with the point estimate. *National Geographic's* "Aliens Among Us" survey conducted in May 21-29, 2012 among a random nationwide sample of 1,114 Americans,

revealed that 36 percent of Americans (point estimate) believe that UFOs exist, ± 2.9 percent (MoE).[8]

5) Confidence Limits

Confidence limits are the upper and lower limits of the confidence interval, abbreviated as UCL and LCL respectively. Table 2 shows how these confidence limits are found:

Table 2: Determining the Upper and Lower Confidence Limit

Confidence Limit	Found By
Upper Confidence Limit (UCL)	Point Estimate + Margin of Error
Lower Confidence Limit (LCL)	Point Estimate – Margin of Error

Figure 2 shows an image of a confidence interval with its point estimate, and upper and lower confidence limits.

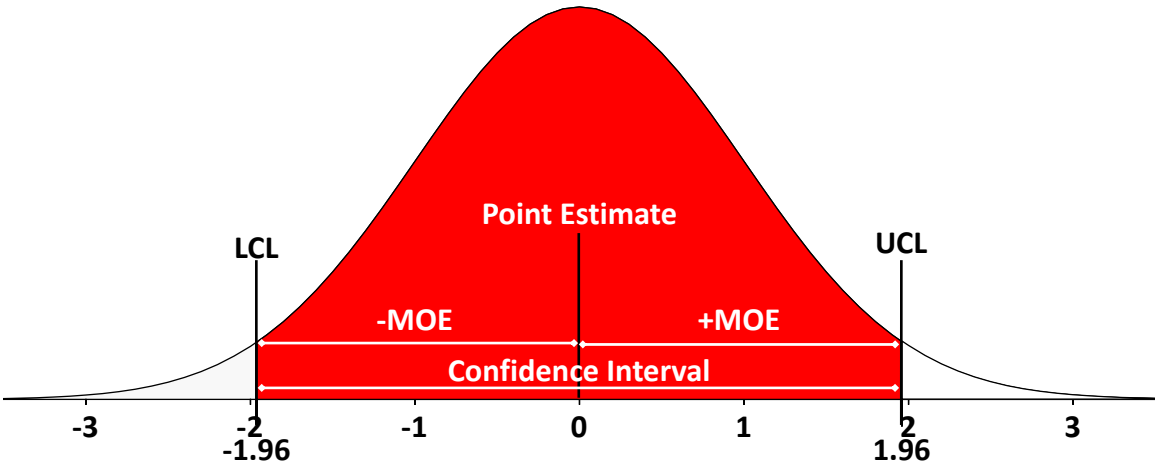

Figure 2: Image of a Confidence Interval

III. Confidence Interval Basics

Confidence intervals provide an estimate of plausible values for an unknown population parameter. Confidence intervals are often explained by stating that if we were to conduct repeated samples, with a 95 percent confidence level, 95 percent of these confidence intervals constructed should contain the population parameter. Figure 3 shows a graphic of 20 confidence intervals with one interval, 5 percent, that does not contain the parameter.

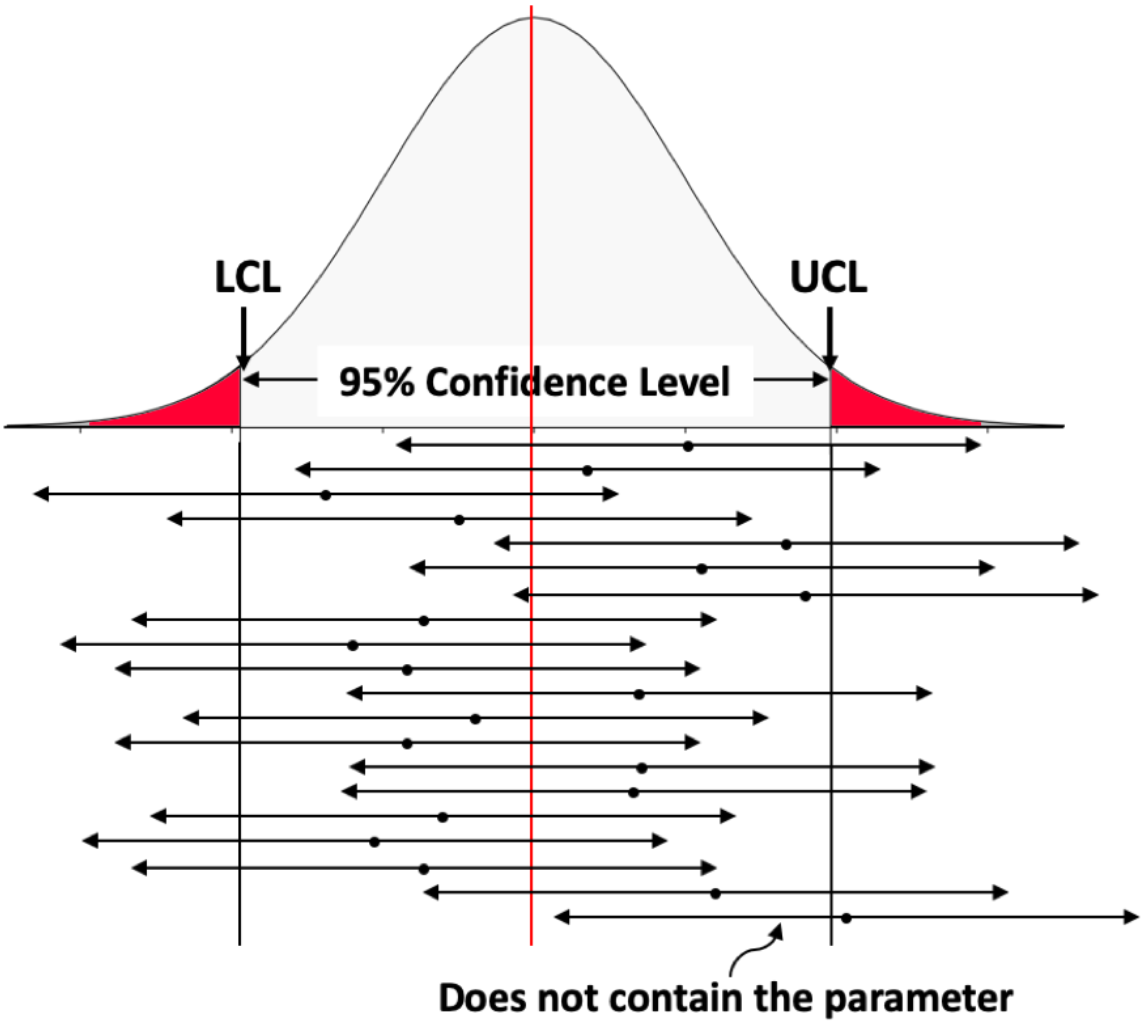

Figure 3: With a 95% Confidence Level, 5% of the Confidence Interval Do Not Contain the Parameter

This explanation of confidence intervals, however, is not quite right. It is a misconception to claim that when the first sample has a 95 percent chance of obtaining a sample statistic, that there is a 95 percent probability that future samples would obtain statistics within the first confidence interval. **This would only be true if the initial estimate is exactly equal to the parameter.** The average probability that a first 95 percent confidence interval capturing the statistic from the next sample is only around 83.4 percent.[9]

The MoE or the width of the confidence interval are affected by:

1. **The selected Confidence Level, CL:** To repeat, the higher the confidence level, the wider the confidence interval.

2. **Sample size, n:** The larger the sample the smaller the MoE and narrower the confidence interval.

3. **Variability in the data, σ:** The more variable the data the larger the MoE and wider the confidence interval. Variability of the data is measured by the standard error of the mean, or, in the case of proportions, the standard error of the proportion. These measures will be explained shortly.

We are now ready to construct three basic confidence intervals.

1. Confidence Interval for the population mean, μ, with σ known using z-values.

2. Confidence Interval for the population mean, μ, with σ unknown using t-values.

3. Confidence Interval for π using z-values.

IV. Confidence Intervals for the μ when σ is known

Imagine the following scenario: Donnie Drymph, Jr. is a young Manhattan real estate agent. He wants to specialize in renting one-bedroom apartments to young professionals. He has a big question: What is the average monthly rent for one-bedroom apartments in Manhattan? Probably no one knows with 100 percent certainty because nobody is going to invest the time and money to conduct a census.

Donnie gets a big idea. He will conduct a survey of one-bedroom rentals currently on the market in Manhattan, and based on these sample statistics, he will estimate the population mean rent with a confidence interval using a 95 percent level of confidence. He finds his old statistics textbook from college and locates the following formula for a confidence interval for the mean using z-values. Equation 2 shows this formula:

$$\bar{X} \pm z\frac{\sigma}{\sqrt{n}}$$

Point Estimate ± z-value(Standard Error of the Mean)

Equation 2: Formula for a Confidence Interval for the Mean with a Known Population Standard Deviation

Where: $\bar{X}$ = the sample mean, is the point estimate for the population mean, μ
σ = the population standard deviation
z = the z-value for the selected confidence level
n = the number of observations in the sample
$\sigma/\sqrt{n}$ is the standard error of the mean

Because he chose to construct a 95 percent confidence level, the z-value is 1.96. The confidence interval will cover the center 95 percent of the normal curve, with 2.5 percent outside the confidence interval on the left-tail and 2.5 percent on the right-tail. Figure 4 shows a normal curve drawn using a 95 percent confidence level.

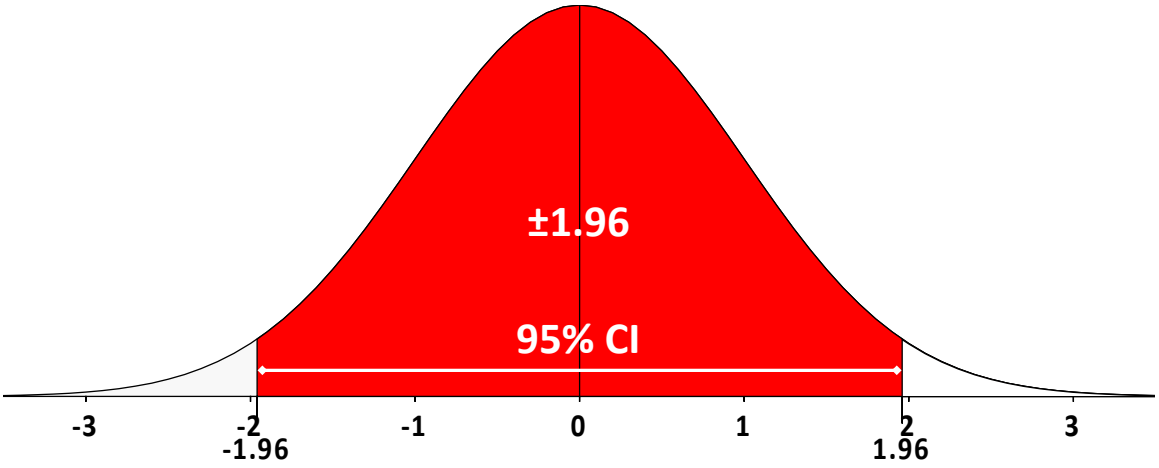

Figure 4: Normal Curve With a 95% Confidence Interval

Now Donnie needs to collect a sample of one-bedroom rental apartments in Manhattan that are currently on the market. Not having a budget, Donnie reads Craig's List for a few weeks to find the monthly rent landlords are asking for one-bedroom apartments. He collects the advertised rents for 100 apartments, n = 100. The data he collected can be found in the workbook titled "11_Examples.xlsx" under the "One-Bedroom Apts" worksheet. The average monthly rent for his sample is $2,520.22. He calculated the population standard deviation, σ. The population standard deviation for the 100 apartments is $884.92. He considers this a reasonable approximation of the real, unknown population standard deviation. He now constructs the confidence interval:

$$\$2{,}520.22 \pm 1.96 \frac{\$884.92}{\sqrt{100}}$$

$$\sigma_{\bar{x}} = \frac{\$884.92}{\sqrt{100}} = \frac{\$884.92}{10} = 88.492$$

$$\$2{,}520.22 \pm 1.96(88.92)$$

$$\$2{,}520.22 \pm \$174.28$$

Equation 3: Calculation of the Confidence Interval for One-Bedroom Apartments

The confidence interval is $2,520.22 ± $174.28. The ± $174.28 is the MoE. The lower confidence limit or LCL is $2,346.78, found by $2,520.22 - $174.28. The upper confidence limit or UCL is $2,693.66, found by $2,520.22 + $174.28. Donnie now knows the apartment he saw today that rents for $3,999 is well above the confidence interval and the apartment his younger brother Eric wants to move into with a rent of $1900, is below the confidence interval.

Microsoft Excel can calculate the confidence interval for the population mean with a known population standard deviation. See Figure 5. Here is the function:

=CONFIDENCE.NORM(alpha, standard deviation, size)

Where: alpha = significance level (1 – the confidence level)
standard deviation is the presumed population standard deviation, σ
z is the z-value for the selected confidence level
size = the sample size, n

	A	B	C	D	E
1	**One Bedroom**	**Monthly Rent**	**Confidence Level**	0.95	**Formula**
2		$2,400	**Sample Mean**	$2,520.22	=AVERAGE(B2:B101)
3		$2,699	**Pop. Std. Dev.**	$884.92	=STDEV.P(B2:B101)
4		$1,399	**n**	100	=COUNT(B2:B101)
5		$1,499	**z-value**	1.96	=ABS(NORM.S.INV(0.025))
6		$2,380	**Std. Error**	88.49	=D3/SQRT(D4)
7		$1,800	**MoE**	$173.44	=D5*D6
8		$1,550	**LCL**	$2,346.78	=D2-D7
9		$2,900	**UCL**	$2,693.66	=D2+D7
10		$2,600	**Alpha**	0.05	=1-D1
11		$3,000	**Excel**	$173.44	=CONFIDENCE.NORM(D10,D3,D4)
12		$2,600			

Figure 5: Confidence Interval Using Microsoft Excel

Donnie was feeling quite satisfied with his work. So satisfied, in fact, he decided to tell his old statistics professor of his success. She praised Donnie's solution for finding a sample that was free, but she said his solution was sophomoric. Sophomoric is a word that is an oxymoron. It means "wise fool."

Donnie asked why his solution is "foolish." She explained that you should not use z-values when you do not know the population standard deviation, σ. She then said he should have calculated the sample standard deviation, not the population standard deviation, and then calculated the confidence interval using t-values. She added that if you have to estimate the population mean, you probably do not know the population standard deviation. The more cautious solution is to use t-values and the sample standard deviation.

Let's turn to how to conduct confidence intervals for the mean using t-values.

IV. Confidence Intervals the μ when σ is unknown. Introduction to student-t Distributions

1. Student-t

Meet William S. Gosset, a chemist educated at Oxford University, who became a Brewmaster at the Guinness Brewery in Dublin, Ireland in 1900. Guinness was a very interesting place for a young chemist at the turn of the twentieth century. At that time, the brewery was the largest in the world, producing over 1.5 million barrels of its dark beer per year, which it exported around the world. Management had decided to make brewing beer scientific and hired newly graduated chemists like Gosset from the finest English universities. These young scientists turned into master brewers focusing on the raw materials for beer: Barley and hops. They asked

questions like which varieties are the best for brewing beer, how should the crops be cultivated, and how the harvested grains should be stored.[10] They ran experiments and found two factors were confounding their studies:

1. There sample sizes were tiny
2. The data were very variable, which is to say the standard deviations were large. At the time, statisticians, who worked with large samples, did not distinguish between population and sample standard deviation. As Joan Fisher Box wrote, "They always used such large samples that their estimate really did approximate the parameter value, so it did not make much difference to their results."[11]

By 1904, Gosset had become the person his colleagues consulted when they had a mathematics problem. Gosset was sent to work with Karl Pearson, who was the era's leading expert in biometrics or the measurement of the physical or behavioral characteristics of animals and plants. During the 1906-1907 academic year, Gosset worked in Pearson's laboratory at University College, London. Here he calculated the tables for a new distribution that could be used when sample sizes were small or the population standard deviation was unknown. Pearson, like most statisticians of the time, worked with very large samples, and failed to appreciate the enormous importance of Gosset's work. By 1908, Gosset was ready to publish his tables. Guinness management agreed to allow Gosset to publish as long as the brewery's name was not mentioned. Gosset chose "Student" as his pseudonym.[12] In 1912, a young Ronald A. Fisher worked on the mathematical proof for student-t distribution.[13] It is interesting that in 1919 Fisher worked as a statistician for the Rothamsted Experimental Station, one of the oldest agricultural research centers in the world.[14] Like Gosset's student-t distribution, many of Fisher's innovations arose from his agricultural studies.

2. What is the Student-t Distribution?

The student-t distribution, like the normal distribution, is continuous and symmetrical, but student-t distributions are more spread out with flatter peaks and thicker tails than normal distributions based on z-values. While normal distributions are based on two parameters: 1) the population mean, μ, and 2) population standard deviation, σ, student t is based on three dimensions: 1) sample mean, $\bar{X}$; 2) sample standard deviation, s; and 3) degrees of freedom, df, or the lowercase Greek letter nu, ν. With student-t, degrees of freedom are based on the number of observations, n, minus the number of independent samples. For confidence intervals, we have one sample, so degrees of freedom are determined by $n - 1$. The value for student t differs with the size of the sample. Student t values are always greater than z-values, but this difference diminishes as the sample size increases.

The normal distribution with its z-values is based on the Central Limit Theorem. It has certain assumptions when estimating the population mean:

1. Shape of the population is unknown, but the number of observations is 30 or more and, therefore, the sampling distribution is normal
2. The population standard deviation, σ, is known

The student-t distribution is designed to be used when these assumptions are not met. When:

1. The number of observations, n, is less than 30. Or,
2 The population standard deviation, σ, is unknown.

When using t, you must assume that the distribution is normally distributed or nearly normally distributed. If the distribution does not meet this condition, we must use a nonparametric technique, like Mann-Whitney or Kruskall-Wallis.

How do you know if the data are normally distributed? If the data have 30 or more variables, the central limit theorem suggests that sample distributions of the

population will be normally distributed. Sophisticated researchers use statistical programs like SPSS (Statistical Package for the Social Sciences) to run the Shapiro-Wilk Test for Normality. If n is > 2,000, they may run the Kolmogorv-Smirnov test. Normality can also be checked graphically using histograms or Q-Q (quantile-quantile plots), which compare two distributions by plotting their quantiles against each other. Quantiles are the same as percentiles, but are indexed by sample fractions rather than sample percentages. There is also a Chi-Square test for normality that we will cover in Module 17.

3. The Student-t Table

Figure 6 shows the student-t table. Excel and pdf versions of this table can be found in Appendix 2. Like z-tables, t-values can be positive or negative. We need not be concerned about that with confidence intervals, but it will be a concern when we turn to null hypothesis significance testing.

	Confidence Level (CL)							Confidence Level (CL)						
	80%	90%	95%	98%	99%	99.9%		80%	90%	95%	98%	99%	99.9%
	α — One-Tailed Test							α — One-Tailed Test						
	0.10	0.05	0.025	0.01	0.005	0.0005		0.10	0.05	0.025	0.01	0.005	0.0005
	α — Two-Tailed Test							α — Two-Tailed Test						
df	0.20	0.10	0.05	0.02	0.01	0.001	df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619	36	1.306	1.688	2.028	2.434	2.719	3.582
2	1.886	2.920	4.303	6.965	9.925	31.599	37	1.305	1.687	2.026	2.431	2.715	3.574
3	1.638	2.353	3.182	4.541	5.841	12.924	38	1.304	1.686	2.024	2.429	2.712	3.566
4	1.533	2.132	2.776	3.747	4.604	8.610	39	1.304	1.685	2.023	2.426	2.708	3.558
5	1.476	2.015	2.571	3.365	4.032	6.869	40	1.303	1.684	2.021	2.423	2.704	3.551
6	1.440	1.943	2.447	3.143	3.707	5.959	41	1.303	1.683	2.020	2.421	2.701	3.544
7	1.415	1.895	2.365	2.998	3.499	5.408	42	1.302	1.682	2.018	2.418	2.698	3.538
8	1.397	1.860	2.306	2.896	3.355	5.041	43	1.302	1.681	2.017	2.416	2.695	3.532
9	1.383	1.833	2.262	2.821	3.250	4.781	44	1.301	1.680	2.015	2.414	2.692	3.526
10	1.372	1.812	2.228	2.764	3.169	4.587	45	1.301	1.679	2.014	2.412	2.690	3.520
11	1.363	1.796	2.201	2.718	3.106	4.437	46	1.300	1.679	2.013	2.410	2.687	3.515
12	1.356	1.782	2.179	2.681	3.055	4.318	47	1.300	1.678	2.012	2.408	2.685	3.510
13	1.350	1.771	2.160	2.650	3.012	4.221	48	1.299	1.677	2.011	2.407	2.682	3.505
14	1.345	1.761	2.145	2.624	2.977	4.140	49	1.299	1.677	2.010	2.405	2.680	3.500
15	1.341	1.753	2.131	2.602	2.947	4.073	50	1.299	1.676	2.009	2.403	2.678	3.496
16	1.337	1.746	2.120	2.583	2.921	4.015	51	1.298	1.675	2.008	2.402	2.676	3.492
17	1.333	1.740	2.110	2.567	2.898	3.965	52	1.298	1.675	2.007	2.400	2.674	3.488
18	1.330	1.734	2.101	2.552	2.878	3.922	53	1.298	1.674	2.006	2.399	2.672	3.484
19	1.328	1.729	2.093	2.539	2.861	3.883	54	1.297	1.674	2.005	2.397	2.670	3.480
20	1.325	1.725	2.086	2.528	2.845	3.850	55	1.297	1.673	2.004	2.396	2.668	3.476
21	1.323	1.721	2.080	2.518	2.831	3.819	56	1.297	1.673	2.003	2.395	2.667	3.473
22	1.321	1.717	2.074	2.508	2.819	3.792	57	1.297	1.672	2.002	2.394	2.665	3.470
23	1.319	1.714	2.069	2.500	2.807	3.768	58	1.296	1.672	2.002	2.392	2.663	3.466
24	1.318	1.711	2.064	2.492	2.797	3.745	59	1.296	1.671	2.001	2.391	2.662	3.463
25	1.316	1.708	2.060	2.485	2.787	3.725	60	1.296	1.671	2.000	2.390	2.660	3.460
26	1.315	1.706	2.056	2.479	2.779	3.707	61	1.296	1.670	2.000	2.389	2.659	3.457
27	1.314	1.703	2.052	2.473	2.771	3.690	62	1.295	1.670	1.999	2.388	2.657	3.454
28	1.313	1.701	2.048	2.467	2.763	3.674	63	1.295	1.669	1.998	2.387	2.656	3.452
29	1.311	1.699	2.045	2.462	2.756	3.659	64	1.295	1.669	1.998	2.386	2.655	3.449
30	1.310	1.697	2.042	2.457	2.750	3.646	65	1.295	1.669	1.997	2.385	2.654	3.447
31	1.309	1.696	2.040	2.453	2.744	3.633	66	1.295	1.668	1.997	2.384	2.652	3.444
32	1.309	1.694	2.037	2.449	2.738	3.622	67	1.294	1.668	1.996	2.383	2.651	3.442
33	1.308	1.692	2.035	2.445	2.733	3.611	68	1.294	1.668	1.995	2.382	2.650	3.439
34	1.307	1.691	2.032	2.441	2.728	3.601	69	1.294	1.667	1.995	2.382	2.649	3.437
35	1.306	1.690	2.030	2.438	2.724	3.591	70	1.294	1.667	1.994	2.381	2.648	3.435

Figure 6: The Student-t Table

Notice that there are three column headers. Rows 1 and 2 are for confidence intervals at six different confidence levels. Rows 3 and 4 are for one-tailed hypothesis tests at six different

significance levels. Rows 5 and 6 are for two-tailed hypothesis tests at six different significance levels. Columns A and I, show the degrees of freedom, which is the number of observations minus the number of independent samples. We will only use t-distributions when working with one or two independent distributions. If we have three or more independent distributions, we will use F distributions, which will be introduced in Module 15.

If we were going to construct a confidence interval at a 95 percent confidence level for a sample of 26 observations, we would have 25 degrees of freedom. The value of t is not the 1.96 of a normal distribution. It is 2.060, found by looking for the intersection of the 25 degrees of freedom row and the 95% confidence level column. If the sample had 71 observations, we would have 70 degrees of freedom and the value of t would be 1.994. You can calculate the student-t values for a confidence interval using Excel. The formula is:

$$\text{student} - \text{t} = \text{TINV}(\text{significance level}, \text{degrees of freedom})$$

Equation 4: Excel's TINV Function

D) Deciding Whether to Use z-values or t-values

Figure 7 shows when to use t-values and z-values when working with means. When the population standard deviation, σ, is known and the sample size is 30 or more, we use z-values. But, when the population standard deviation, σ, is *not* known, *or* when the sample size is less than 30, we must use t-values. Whenever we use t-values, the distribution should approximate a normal distribution. In most cases the population standard deviation is not known. We will, therefore, build confidence intervals for the population mean using t-values far more often than z-values.

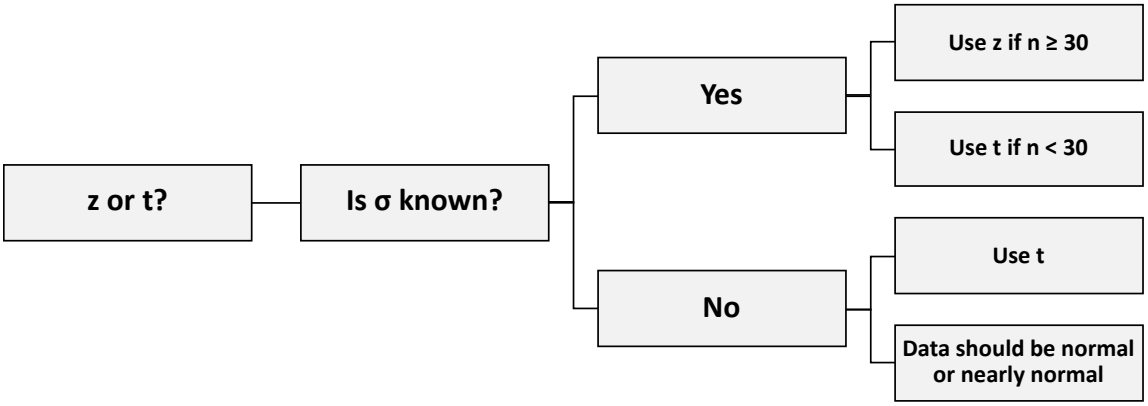

Figure 7: When to Use the Student-t Distribution

IV. Confidence Intervals for the population mean, μ, when σ is unknown

Let's return to Donnie Drymph, Jr., the young real estate executive who is trying to estimate the mean, μ, monthly rent for a one-bedroom apartment in Manhattan. The average monthly rent for his sample of 100 apartments currently on the market is $2,520.22. The sample standard deviation is $889.38 not the $884.92 he used when he calculated the population standard deviation for the 100 apartments. Using the student-t with 99 degrees of freedom, a 95 percent confidence level, the confidence interval is $2,520.22 ± $1.76.47, not the $2,520.22 ± $173.44 found using z-values. This is the proper 95 percent confidence interval for his data.

To arrive at this calculation, the formula for the confidence interval was modified for t-values:

$$\bar{X} \pm t_{df} \frac{s}{\sqrt{n}}$$

Point Estimate ± t-value(Standard Error of the Mean)

Equation 5: Confidence Interval for the Mean using student-t

Where: $\bar{X}$, the sample mean, is the point estimate for the population mean, μ
s is the sample standard deviation
t is the t-value for the selected confidence level
n is the number of observations in the sample
$s/\sqrt{n}$ is the estimate for the standard error of the mean

This first step is to find the t-value for a 95 percent confidence interval with 99 degrees of freedom, 100 – 1. We can use the critical values table for student t or we can calculate the critical value using Excel:

$$=\text{TINV}(0.05,99)$$

Equation 6: Finding the CV for t using Excel's TINV function

The critical value for t is 1.984, not 1.96. Once we have the critical value, we plug this number into the formula shown in Equation 7:

$$\overline{X} \pm t_{99} \frac{\sigma}{\sqrt{n}} = \$2{,}520.22 \pm 1.984 \frac{\$889.38}{\sqrt{100}} = \$2{,}520.22 \pm \$176.47$$

Equation 7: Confidence Interval for the Monthly Rent for 100 One-Bedroom Apartments

Figure 8 shows how to calculate the confidence interval for the population mean, μ, using student-t. The interval is slightly wider than the one we calculated using z-values. But, more importantly, it does not violate important assumptions like the one constructed using z-values.

12		$2,600	C	D	E
13		$2,400	**Using t-values**		
14		$1,750	**Sample Mean**	$2,520.22	=AVERAGE(B2:B101)
15		$1,675	**Sample Std. Dev.**	$889.38	=STDEV.P(B2:B101)
16		$1,899	**n**	100	=COUNT(B2:B101)
17		$1,595	**t-value**	1.984	=ABS(T.INV(0.025,D4-1))
18		$2,000	**Std. Error**	88.94	=D15/SQRT(D16)
19		$3,200	**MoE**	$176.47	=D17*D18
20		$4,350	**LCL**	$2,343.75	=D14-D19
21		$2,000	**UCL**	$2,696.69	=D14+D19
22		$1,699	**Alpha**	0.05	=1-D1
23		$1,675	**Excel**	$176.47	=CONFIDENCE.T(D22,D15,D16)

Figure 8: Calculating the Confidence Interval With Student-t

Remember: We construct confidence intervals to estimate unknown parameters. When we do not know the population mean, we probably do not know the population standard deviation. Calculating the confidence interval for the mean using t-values is a more cautious approach than using z-values.

V. Confidence Intervals for the population proportion, π

We also calculate confidence intervals for proportions. Proportions are represented as a fraction, decimal, ratio, or percentage of the part of a population or sample that has a certain characteristic. Proportions are binomial, which means that the outcomes fall into one of two groups. These groups can be defined as "meets the criteria" or "fails to meet the criteria." Some people call the groups "successes" and "failures." Here are some binomial questions that use proportions:

- The proportion of Americans who think President Donald J. Trump is a racist.
- The proportion of men with male pattern baldness.
- The proportion of Americans who follow the paleo diet.

We do not use student-t values to calculate confidence intervals for proportions. This is because the Central Limit Theorem often allows us to use z-values when the sample size is less than 30. We can use z-value for proportions—binomial populations—when $n\pi$ and $n(1 - \pi)$ are equal to or greater than five where "π" stands for the probability of "success" and n refers to the population size. In addition, we do not use standard deviations. So there is no issue about whether or not we know the population standard deviation.

There are three basic formulas for calculating confidence intervals for proportion:

A) Finding the sample proportion, p:

$$p = \frac{X}{n}$$

Equation 8: Formula for Proportions

Where: X = the random variable
n = the number of observations in the sample
p = the sample proportion

B) The formula for the standard error of the proportion, σ_p or SEP:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Equation 9: Formula for the Standard Error of the Proportion

C) The formula for calculating the confidence interval

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

Point Estimate ± z-value(Standard Error of the Proportion)

Equation 10: Confidence Intervals for Proportions

The Gallup poll has been tracking Americans' attitudes on marijuana for decades. A poll conducted in October 2019 showed that 66 percent of Americans favor the legalization of marijuana for recreational use compared to 53 percent of Republicans.[15]

Here is a break-out of attitudes toward the legalization of marijuana by respondents' political affiliation.

Table 3: Attitudes Toward the Legalization of Marijuana by Political Party

Party	Number Surveyed	Number Favoring Legalization	%
Republican	393	199	50.64%
Independent	655	444	67.48%
Democrat	450	352	78.22%

What is the confidence interval for Republicans favoring the legalization of marijuana using a 95 percent level of confidence? The answer is 50.64 percent ± 4.94 percent. Here are the calculations:

$$p = \frac{X}{n} = \frac{199}{393} = 0.5064 = 50.64\%$$

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

$$0.5064 \pm 1.96\sqrt{\frac{.5064(1-.5064)}{393}}$$

$$.5064 \pm 1.96\sqrt{\frac{.2500}{393}}$$

$$.5064 \pm 1.96\sqrt{0.00064}$$

$$.5064 \pm 1.96(0.0252)$$

$$.5064 \pm .0494$$

Equation 11: Confidence Interval for Republicans Favoring Legalizing Marijuana

Based on these calculations, we are 95 percent confident that between 45.69 percent and 55.58 percent of Republicans favor the legalization of marijuana. If we were to calculate a confidence interval at a 99 percent confidence level, the z-value would be 2.58, and the confidence interval would be 50.64 percent ± 6.51 percent

How do Republicans compare to Democrats and Independents?

Table 4: 95 Percent Confidence Intervals for the Legalization of Marijuana by Political Affiliation

Political Affiliation	Confidence Interval	LCL	UCL
Republican	50.64% ± 4.94%	45.69%	55.58%
Independent	67.78% ± 4.35%	63.44%	72.13%
Democrat	78.22% ± 4.31%	73.91%	82.53%

The confidence interval for Republicans does not overlap with those for Democrats and Independents. Clearly Democrats and Independents are far more likely to favor the legalization of marijuana than Republicans. The confidence intervals for Democrats and Independents also do not overlap. This means that Democrats are more likely than Independents to favor the legalization of marijuana. See Figures 9 and 10 below.

Figure 9 shows the calculations done in Excel. Excel does not have a built-in CONFIDENCE function for proportions. These calculations can be found in 11_GallupMarijuana.xlsx.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Party**	**X**	**n**	**p**	**z**	**(1 - p)**	**p(1 - p)**	**p(1 - p)/n**	**Std. Err.**	**MoE**	**LCL**	**UCL**
2	Republicans	199	393	0.5064	1.96	0.494	0.24996	0.00064	0.02522	0.0494	0.4569	0.5558
3	Independents	444	655	0.6779	1.96	0.322	0.21836	0.00049	0.02218	0.0435	0.6344	0.7213
4	Democrats	352	450	0.7822	1.96	0.218	0.17035	0.00048	0.02200	0.0431	0.7391	0.8253
5												
6	Gallup Poll											
7	23-Oct-19											
8	https://news.gallup.com/poll/267698/support-legal-marijuana-steady-past-year.aspx											

Figure 9: Confidence Interval Calculation Done in Excel

We can illustrate these confidence intervals in a chart. Figure 10 shows the three confidence intervals. This chart makes the key finding very clear. Because the confidence intervals do not overlap, Democrats are more likely to favor the legalization of marijuana that Independents who are more likely to favor the legalization of marijuana than Republicans.

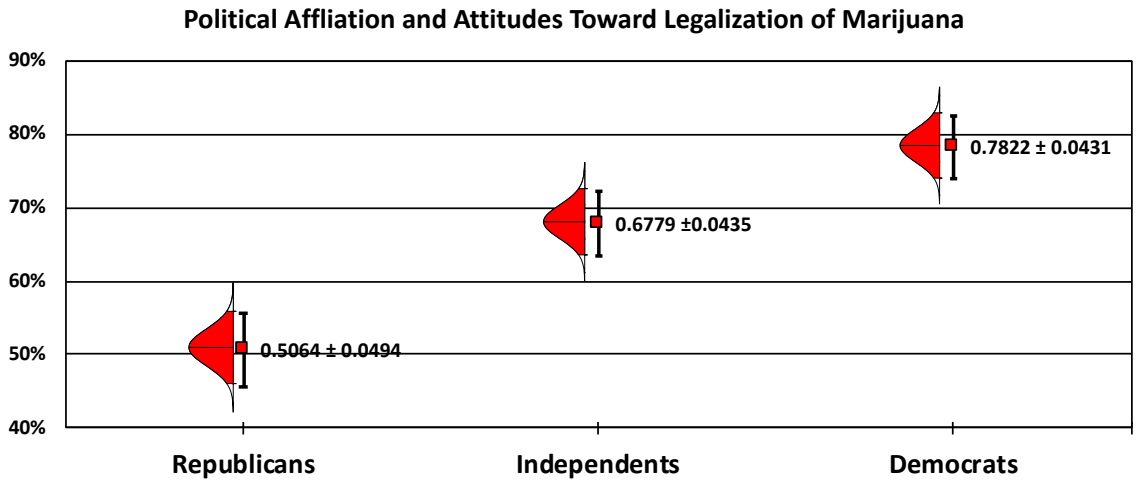

Figure 10: Political Affiliation and Attitudes Toward Legalization of Marijuana

We will look at these numbers again when we conduct NHST for proportions from two samples and a chi-square contingency table tests.

VI. Finite Population Correction Factor (FPC Factor)

The FPC factor is used when studying small (finite) populations. The FPC factor makes the confidence intervals more precise. The FPC factor is used when your sample is more than 5

percent of the finite population when the sample is performed *without replacement*. This needs clarification. When we conduct a sample, the element chosen is usually not returned to the population; which is to say, you will not select the same element twice in a single sample. This is a sample without replacement.

Equation12 show formula for the FPC factor:

$$\text{FPC} = \sqrt{\frac{\text{N} - \text{n}}{\text{N} - 1}}$$

Equation 12: Finite Population Correction Factor Formula

Here is the formula for a confidence interval for the mean using an FPC factor.

$$\overline{\text{X}} \pm \text{t}\frac{\text{s}}{\sqrt{\text{n}}}\left(\sqrt{\frac{\text{N} - \text{n}}{\text{N} - 1}}\right)$$

Equation 13: Formula for Confidence Interval for the Mean With FPC Factor

Lara Drymph is secretary of her high school graduation class. She is planning the twentieth year reunion. She wants to estimate the average number of children her classmates have. Her graduating class had 150 students, N = 150; 25 people, n, 16.67 percent of the population, responded to her survey. The sample mean is 1.60 children, and the sample standard deviation, s, is 1.20 children. What is the confidence interval using an FPC factor at a 5 percent level of confidence? The value for t with 24 degrees of freedom (25 – 1) is 2.064.

$$\overline{\text{X}} \pm \text{t}\frac{\text{s}}{\sqrt{\text{n}}}\left(\sqrt{\frac{\text{N} - \text{n}}{\text{N} - 1}}\right)$$

$$1.60 \pm 2.064\frac{1.20}{\sqrt{25}}\left(\sqrt{\frac{150 - 25}{150 - 1}}\right)$$

$$1.60 \pm 2.064\frac{1.20}{5}(0.916)$$

$$1.60 \pm 2.064(0.24)(0.916)$$

$$1.60 \pm 0.454$$

Equation 14: Confidence Interval With an FPC Factor

We now know with 95 percent certainty that members of Ms. Drumph's graduating class have on average 1.6 children ± 0.454 children with an LCL of 1.146 children and an UCL of 2.054 children.

VII. Summary

We have constructed three basic confidence intervals.

1. Confidence intervals for the population mean, μ, when σ is known
2. Confidence intervals for the population mean, μ, when σ is unknown.
3. Confidence intervals for the population proportion, π.
4. The student-t distribution was also introduced along with the Finite Population Correction Factor.

There is one more estimation problem to consider: Sample size estimation, which will be the topic for Module 12. In Module 13, we will begin our discussion of Null Hypothesis Significance Testing (NHST). In Module 19 we will discuss why many contemporary statisticians favor confidence intervals over traditional NHST.

VIII. Exercises

Answer the following questions using a handheld calculator and the tables for z-values and t-values. Check your answers using Microsoft Excel. The data for these problems are listed below and are in 11_Exercises.xlsx.

Exercise 1: Donnie Drymph's younger brother, Eric buys and sells used cars. He wants to estimate the average price—the population mean, μ, price—of a 2015 Honda Accord. He

takes a sample of 39 ads for 2015 Honda Accords, the sample mean, $\bar{X}$, is $13,466.82. The population standard deviation calculated from his sample is $2,813.82. The sample standard deviation, s, is $2,850.60.

a) What is the point estimate?
b) What is the appropriate z-value to use when constructing a 95 percent confidence interval?
c) What is the appropriate t-value to use when constructing a 95 percent confidence interval?
d) Using a 95 percent confidence level, calculate the confidence interval using z-values. Do this by hand and use Excel's CONFIDENCE.NORM function.
e) Using a 95 percent confidence level, calculate the confidence interval using student-t. Do this by hand and use Excel's CONFIDENCE.T function.
f) What does the confidence interval you constructed tell you?
g) Which standardized score is more appropriate: z-values or student-t? Explain your answer.
h) What would happen to the confidence interval if you changed the level of confidence?
i) What would happen to the confidence interval if you increase the sample size?

Exercise 2: According to a 2018 survey of student loan debt incurred to finance undergraduate education, the statistics are: $\bar{X}$ = $25,786.74, the presumed population standard deviation is $9,155.70, and n = 1,000. The t-value with 999 degrees of freedom is 1.962.

a) What is the point estimate?
b) What is the appropriate z-value to use when constructing a 95 percent confidence interval?
c) Using a 95 percent confidence level, calculate the confidence interval using z-values. Do this by hand and use Excel's CONFIDENCE.NORM function.
d) Using a 95 percent confidence level, calculate the confidence interval using student-t. Do this by hand and use Excel's CONFIDENCE.T function.
e) What does the confidence interval you constructed tell you?
f) Compare the confidence intervals created using z-values and student-t.
g) Which standardized score is more appropriate: z-values or student-t? Explain your answer.
h) What would happen to the confidence interval if you changed the level of confidence?
i) What would happen to the confidence interval if you increase the sample size?

Exercise 3: The Monmouth University Polling Institute is one of the nation's most respected pollsters. A poll conducted among 689 likely voters between August 16 and

August 20, 2019 asked, "Looking ahead to the 2020 election for president, do you think that Donald Trump should be reelected, or do you think it is time to have someone else in office? Here are the results of the survey:

Table 5: Poll On President Trump's Reelection, n = 689

Response	%
Should be reelected	39%
Someone else	57%
Don't Know/No Answer	4%

a) What is the point estimate for the position that the president "Should not be reelected; which is to say, "someone else"?
b) What is the appropriate z-value to use when constructing a 95 percent confidence interval?
c) Using a 95 percent confidence level, calculate the confidence interval using z-values for " "Someone else." Do this by hand and use Excel.
d) What do the confidence intervals you constructed tell you?
e) What would happen to the confidence interval if you changed the level of confidence?
f) What would happen to the confidence interval if you increase the sample size?

Exercise 4: Quinnipiac University Poll is a leading polling organization. In May 2019, it released a survey conducted from May 16 to May 20, 2019 of 1,078 self-identified registered voters. The poll asked a series of questions about a woman's right to have a legal abortion. Here is question 47: How about when a pregnancy was caused by rape or incest; do you think abortion should be legal in this situation or not? Here are the responses broken out by political affiliation:

Table 6: Quinnipiac Polls on the right to an abortion when rape or incest is involved.

	n = 313 Republican	n =345 Democrat	n = 323 Independent
Support	68%	92%	83%
Oppose	25%	5%	10%
Don't Know/No Answer	7%	3%	6%

Answer the following questions for the position that *supports* legal abortions in the case of rape or incest for the three political affiliations: Republican, Independent, and Democrat.

a) What are the point estimates?
b) What is the appropriate z-value to use when constructing a 95 percent confidence interval.
c) Using a 95 percent confidence level, calculate the confidence interval using z-values for the three different political affiliations? Do this by hand and use Excel.
d) What do the confidence intervals you constructed tell you?
e) What would happen to the confidence interval if you change the level of confidence?
f) What would happen to the confidence interval if you would increase the sample size?

Exercise 5: From April 4 to April 14, 2019, CBSNews conducted a nationwide poll among 1,010 adults on whether marijuana should be legal. Here are the results broken down by political affiliation. Construct three confidence intervals using a 95 percent confidence level: one for Republicans, Democrats, and Independents. Here are the results:

Table 7: CBSNews Poll on the Legalization of Marijuana

Political Affiliation	Total n	Legal %
Republican	307	56%
Democrat	321	72%
Independent	382	66%

a) What are the point estimates?
b) What is the appropriate z-value to use when constructing a confidence interval using a 95 confidence level?
c) Calculate the three confidence intervals. Do this by hand and use Excel.
d) What do the confidence intervals you constructed tell you?
e) What would happen to the confidence intervals if you changed the level of confidence?
f) What would happen to the confidence interval if you increase the sample size?

Exercise 6: Lara Drymph is secretary of her high school graduation class. She is planning the twentieth reunion dinner. In a survey she asked respondents to report how many miles

they currently live from their high school. Construct a confidence interval for the mean using t-values and the FPC Factor. Here are your data:

Table 8: FPC Factor Data

Sample Mean, $\bar{X}$	113.04
Sample Standard Deviation, s	236.51
N	150
n	25

a) What is the point estimate?
b) What is the appropriate t-value to use when constructing a confidence interval using a 95% confidence level?
c) Using a 95 percent confidence level, construct the confidence interval.
d) Comment on your findings.
e) What would happen to the confidence interval if you changed the level of confidence?
f) What would happen to the confidence interval if you increase the sample size?

Exercise 7: From July 25 to July 28, 2019, the Quinnipiac Poll asked: "Do you think that President Trump is racist or don't you think so?" Here are the "yes" responses and sample sizes for four groups. Construct a 95 percent confidence interval for the "yes" answers from each group and comment on your findings:

Group	p	n
Total	51%	1,309
Republicans	8%	392
Democrats	86%	444
Independents	56%	340

a) What are the point estimates?
b) What is the appropriate z-value to use when constructing a confidence interval using a 95 confidence level?
c) Calculate the four confidence intervals.
d) What do the confidence intervals you constructed tell you?

---

[1] Quoted by David S. Moore "Uncertainty" in Lynn Arthur Steen (Ed.). *On the Shoulders of Giants: New Approaches to Numeracy*. (Washington, DC: The National Academies Press, 1990), p. 135. Original Source, Arthur C. Nielsen, Jr. "Statistics in Marketing." In George Easton, Harry V. Roberts, and George, George C. (Eds.): *Making Statistics More Effective in Schools of Business*. Chicago, IL: University of Chicago Graduate Schools of Business, 1986.

[2] Jerzy Neyman, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society of Long. Series A. Mathematical and Physical Sciences*.

Vol. 236, No. 767. August 30, 1937, pp. 333-334. Jerzy Neyman, "On the Problem of Confidence Intervals," The Annals of Mathematical Statistics. Volume 6, No. 3. September 1935, pp. 111-116.

[3] Richard Feynman, *QED: The Stranger Nature of Light*. (Princeton, NJ: Princeton University Press, 1985), p. 19. QED is an abbreviation for the Latin phrase *quod erat demonstrandum*, which means "what was to be demonstrated."

[4] Jerzy Neyman, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society of Long. Series A. Mathematical and Physical Sciences*. Vol. 236, No. 767. August 30, 1937, p. 333.

[5] "National Youth Tobacco Survey (NYTS)," Office on Smoking and Health, National Center for Disease Control and Prevention and Health Promotion, February 28, 2018, https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm//#nyts-historical.

[6] *Genesis*, 5:27, https://www.kingjamesbibleonline.org/Genesis-Chapter-5/

[7] *Genesis*, 5:23, https://www.kingjamesbibleonline.org/Genesis-Chapter-5/

[8] Press Release from the National Geographic Channel, "Two-Thirds of Americans Think Barack Obama is Better Suited to Handle an Alien Invasion that Mitt Romney." June 29, 2012. http://spaceref.com/news/viewpr.html?pid=37603.

[9] Geoff Cumming, Jennifer Williams, and Fiona Fidler. "Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars." *Understanding Statistics*. Vol. 3, October 2004. pp. 299-311. DOI: 10.1207/s15328031us0304_5.

[10] Joan Fisher Box, "Guinness, Gosset, Fisher, and Small Samples," *Statistical Science*, Volume 2, No. 1, February, 1987, p. 47. Ms. Box is the daughter of R. A. Fisher and the widow of the statistician George E. P. Box.

[11] Joan Fisher Box, p. 45.

[12] Joan Fisher Box, p. 49.

[13] Joan Fisher Box, p. 50.

[14] Joan Fisher Box, "R. A. Fisher and the Design of Experiments, 1922-1926," *The American Statistician*, Volume 34, No. 1, February 1980, pp. 1-7.

[15] Gallup, October 2019. https://news.gallup.com/poll/267698/support-legal-marijuana-steady-past-year.aspx .