AP Stats Chapter 10: Estimating with Confidence

How long can you expect a AA battery to last? What proportion of college undergraduates have engaged in binge drinking? Is caffeine dependence real? Rather than collect every battery or ask all college undergraduates, we use a sample to answer our questions. We will then use our sample data to make some conclusions about the population.

Statistical Inference

Statistical inference provides methods for drawing conclusions about a population from sample data.

Confidence Intervals

Example: The admissions director at a university wants to market his school using the IQ score of current students. He chooses a SRS of 50 students from the school's 5000 freshmen and administers an IQ test. The mean of the sample is 112. What can the director say about the mean score of the population of all 5000 freshmen?

Is the mean IQ of all freshmen 112? Somewhere close to 112? How do we know?

How close to 112 is μ likely to be? How would the sample mean \overline{x} vary (suppose standard deviation is 15 for the population) if we took many samples of 50 freshmen from this same population?

Putting these facts together gives us the reasoning of statistical estimation in a nutshell.

1. To estimate the unknown population mean μ , use the mean \overline{x} of our random sample.

2. Although \overline{x} is an unbiased estimate of μ , it will rarely be exactly equal to μ , so our estimate has some error.

3. In repeated samples, the values of \overline{x} follow approximately a normal distribution with mean μ and standard deviation 2.1.

4. The 95 part of the 68-95-99.7 rule for Normal distributions says that in about 95% of all samples, the mean IQ score for the sample will be within 4.2 of the population mean.

5. When ever \overline{x} is within 4.2 points of μ , μ is within 4.2 points of \overline{x} . This happens in about 95% of all possible samples. So the unknown μ lies between \overline{x} - 4.2 and \overline{x} + 4.2 in about 95% of all samples.

6. If we estimate that μ lies somewhere in the interval from

We would be calculating this interval using a method that "captures" the true mean in about 95% of all possible samples.

Sampling Distribution for the mean IQ score of an SRS of size 50.



In 95% of all samples, the mean lies within 4.2 of the unknown population mean.



To say that \overline{x} +/- 4.2 is a 95% confidence interval for the population mean μ is to say that, in repeated samples, 95% of these intervals capture μ .



Conclusion: Our sample of 50 freshmen gave $\overline{x} = 112$. The resulting interval is 112 ± 4.2 . Which can be written as (107.8, 116.2). We say that we are 95% confident that the unknown mean IQ score for all university freshmen is between 107.8 and 116.2.

The interval we give is called the **confidence interval** and the 95% is the **confidence level.** The plus or minus 4.2 is the **margin of error**.



Twenty five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that contains the population mean μ .



Assignment: p. 624-626 10.1 to 10.5

Confidence Interval for a Population Mean (When σ is known)

The calculation of our interval is based on three assumptions:



1. SRS, 2. Normality, 3. Independence Check these three before you construct a confidence interval

Example to find z*:

To construct an 80% confidence interval, we must catch the central 80% of the Normal sampling distribution of \overline{x} . In catching the central 80%, we leave out 20% or 10% in each tail. So z^* is the point with 0.1 area to the right and 0.9 area to the left under the standard Normal curve. Looking at Table A the closest entry with 0.9 to the left is $z^{*}=1.28$. There is 0.8 area under the normal curve between -1.28 and 1.28.

z	.07	.08	.09
1.1	.8790	.8810	.8830
1.2	.8980	.8997	.9015
1.3	.9147	.9162	.9177







The critical values are generally $-z^*$ and $+z^*$.

Confidence Interval for a Population Mean (sigma known)

Choose an SRS of size n from a population having unknown mean and known standard deviation. A level C confidence interval for μ is

Here z^* is the value that determines the area C between $-z^*$ and z^* under the standard Normal curve. This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

Example: Constructing a confidence interval

A manufacturer of high resolution video terminals must control the tension on the mesh of fine wires that lies behind the surface of the viewing screen. Some variation is inherent in the production process. The standard deviation of the tension readings is $\sigma = 43mV$. Here are the readings from an SRS of 20:

233.5 327.4 304.8 269.5 297.0 269.6 283.3 280.4 257.4 317.5 374.6 343.3 264.7 307.7 310.0 328.1 342.6 338.8 340.1 336.1

Construct and interpret a 90% confidence interval for the mean tension μ of all the screens produced on this day.

Step 1: Parameter-Identify the population of interest and the parameter you want to draw conclusions about.

Step 2: Conditions-Choose the appropriate inference procedure. Verify conditions for using it. SRS:

Normality:

Independence:

Step 3: Calculations-If the conditions are met, carry out the inference procedure. What is the sample mean?

What is the critical value z* for a 90% confidence level?

Step 4: Interpretation-Interpret your results in the context of the problem. Remember: Conclusion, connection and context.

Just a reminder...Larger samples give smaller intervals because the standard deviation is being divided by a larger value.



 1.
 ______ gets smaller

 2.
 ______ gets smaller

3. _____ gets larger

Suppose the manufacturer of the screens in yesterday's example wants 99% confidence rather than 90% confidence. The

critical value for 99% confidence is $z^* =$ _____. The 99% confidence interval for μ based on an SRS of 20 with mean = 306.3 is



Demanding 99% confidence instead of 90% confidence has increased the margin of error from ______ to _____.

Sample Size for a Desired Margin of Error

To determine the sample size n that will yield a confidence interval for a population mean with a specified margin of error m, set the expression for the margin of error to be less than or equal to m and solve for n

Researchers would like to estimate the mean cholesterol level of μ of a particular variety of monkey that is often used in laboratory experiments. They would like their estimate to be within 1 milligram per deciliter of blood of the true value of

 μ at a 95% confidence level. A previous study involving this variety of monkey suggests that the standard deviation of cholesterol level is about 5 mg/dl. Obtaining monkeys is time-consuming and expensive, so the researchers want to know the minimum number of monkeys they will need to generate a satisfactory estimate.

For 95% confidence, the table gives $z^* = 1.96$. We know that sigma = 5. Set the margin of error to be at most 1:

Always round up to the next whole number when finding n.

Notice that the size of the sample determines the margin of error. The size of the population does not influence the sample size. (this is true as long as the population is much larger than the sample)

Assignment: p. 637-639 10.13 to 10.18 Section 10.1 Exercises p. 640 10.19-10.20, 10.23, 10.26

Estimating a Population Mean

How can we construct a confidence interval for an unknown population mean μ when we don't know the population standard deviation σ ? We have to estimate σ from the data even though we want to know the mean.

Conditions for Inference about a Population Mean: 1.

2.

3.

To estimate the population standard deviation we will use the sample standard deviation s. This is called the Standard Error:

t Distribution

When we use the standard error rather than the standard deviation, we do not have a normal distribution, we have a t distribution. Unlike the standard Normal distribution, there is a different t distribution for each sample size n. We specify a particular t distribution by giving **degrees of freedom (df)**. When we perform an inference about μ using a t distribution, the degrees of freedom are found by n - 1.

Comparing z and t distributions:

1. Define Y1 = normalpdf(X) and change the graph style to thick. Define Y2 = tpdf(X, 2).



2. Change your window to look like the following. Then graph.

WINDOW Xmin=-3 Xmax=3 Xscl=1 Ymin=-.1 Ymax=.4 Yscl=.1 Xres=1 3. Sketch what you see on your screen.

	•
<u> </u>	<u> </u>
	L

4. Change Y1 to a dotted line and change Y2 to tpdf(X, 9), graph and sketch what you see.



5. Change Y2 to tpdf(X, 30), graph and sketch what you see.

·
l

6. What appears to be happening to the shape of the t distribution curve as the number of degrees of freedom increases?

Table C in the back of your book gives critical values t* for the t distributions.

Suppose you want to construct a 95% confidence interval for the mean μ of a population based on an SRS of size n = 12. What critical value t* should you use?

df = _____

One-Sample t Confidence Intervals

To construct a confidence interval for μ based on a sample from a Normal population with unknown σ , replace the standard deviation σ / \sqrt{n} of \overline{x} by its standard error s / \sqrt{n} in the formula $\overline{x} \pm z * \sigma / \sqrt{n}$. Use critical values from the t-distribution with n - 1 degrees of freedom in place of the z critical values.

The final form of the t distribution interval is:

Example: Constructing a one sample t interval

Environmentalists, government officials, and vehicle manufacturers are all interested in studying the auto exhaust emissions produced by motor vehicles. The major pollutants in auto exhaust from gasoline engines are hydrocarbons, monoxide, and nitrogen oxides (NOX). The following table gives the NOX levels (in grams per mile) for a random sample of light duty engines of the same type.

1.28	1.17	1.16	1.08	0.60	1.32	1.24	0.71	0.49	1.38	1.20	0.78
0.95	2.20	1.78	1.83	1.26	1.73	1.31	1.80	1.15	0.97	1.12	0.72
1.31	1.45	1.22	1.32	1.47	1.44	0.51	1.49	1.33	0.86	0.57	1.79
2.27	1.87	2.94	1.16	1.45	1.51	1.47	1.06	2.01	1.39		

Construct and interpret a 95% confidence interval for the mean amount of NOX emitted by light duty engines of this type.

Step 1: Parameter-

Step 2: Conditions— SRS:

Normality:

Independence:

Step 3: Calculations—

Step 4: Interpretation—

When the actual df does not appear in your table, use the greatest df available that is less than your desired df.

Paired t Procedures

To compare the responses to the two treatments in a matched pairs design or before and after measurements on the same subjects, apply one sample t procedures to the observed differences.

The parameter μ in a paired t procedure is...

--

--

--

Example: Is caffeine dependence real?

Our subjects are 11 people diagnosed as being dependent on caffeine. Each subject has been barred from caffeine. Instead they took capsules containing their normal caffeine intake. During a different time period, they took placebo capsules. The order in which the subject took placebo and caffeine was randomized. The table contains information on two of several tests performed on the subjects. We are interested in whether caffeine affects these outcomes. Construct and interpret a 90% confidence interval for the mean change in depression score.

Subject	Depression	Depression	Number of Beats	Number of Beats		
	(Caffeine)	(Placebo)	(Caffeine)	(Placebo)		
1	5	16	281	201		
2	5	23	284	262		
3	4	5	300	283		
4	3	7	421	290		
5	8	14	240	259		
6	5	24	294	291		
7	0	6	377	354		
8	0	3	345	346		
9	2	15	303	283		
10	11	12	340	391		
11	1	0	408	411		

Step 1: Parameter-

Step 2: Conditions— SRS:

Normality:

Independence:

Step 4: Interpretation—

Random Selection:

Random Assignment:

Robust Procedures:

Example: t procedures are not robust against outliers. If we remove the outlier from the NOX example, we get the following results.

Using the t Procedures:

- -----
- ---
- ---

Assignment: p. 657-658 10.33 to 10.10.36 Section 10.2 Review p. 659-661 10.37, 38, 39, 40, 42, 43, 44

Using the calculator for t intervals:

1. Enter the following numbers into L1.

		C											
-1	10	3	-3	-31	4	-12	-3	-7	-10	-22	-4	-1	-3
2. Go t EDI 2: T 3: 2 4: 2 5: 1 6: 2 7↓2 3. Cho Ca	io STAT T Cf T Cf Tes Sar Sar Pro Pro Inte Inte iter Star Sea Iter Star Star Sar Inte Inte Inte Star Sar Sar Sar Sar Sar Sar Sar Sar Sar S	ALC ALC St St APZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE OPZTE	est est est est st st the Tim Stat	e 8:TInte	shown								
4. Sele TIN (- x= Sx n=	et Calcu 11.8 -5.7 =10. 14 sults tell	late and p /a1 31, . 3 71428 . 5642 us that th	8536 5714 9816	>	interval	for the tru	le mean p	oopulation	n differer	nces in he	ealing rate	e is betwe	een -
11.81 a	and 0.38	5 microm	eters per l	our.			-						

Estimating a Population Proportion

Thus far we have been concerned with using inference to find out about population means. But we often want to answer questions about the proportion of some outcome in the population. For instance,

- What proportion of US adults are unemployed right now?
- What proportion of teenagers have a computer with internet access in their bedroom?
- What proportion of college students pray on a daily basis?
- What proportion of preteens have a cell phone?
- What proportion of Californians approve of President Bush's handling of the situation in Iraq?

Example: Estimating a population proportion

Alcohol abuse has been described by college presidents as the number one problem on campus, and it is an important cause of death in young adults. How common is it? A 2001 survey of 10,904 US college students collected information on drinking behavior and alcohol related problems. The researcher defined "frequent binge drinking" as having five or more drinks in a row three or more times in the past two weeks. According to this definition, 2486 students were classified as frequent binge drinkers. That's 22.8% of the sample. Based on these data, what can we say about the proportion of all college students who have engaged in frequent binge drinking?

Conditions for Inference about a Proportion:

1.	2.	3.

Are the conditions met in the binge drinking example? SRS:

Normality:

Independence:

A Confidence interval for a Population Proportion

Draw an SRS of size n from a large population with unknown proportion p of successes. An approximate level C confidence interval for p is

Where z^* is the upper (1-C)/2 critical value of the standard normal distribution.

Example: 2486 out of 10,904 college students said they had engaged in frequent binge drinking. P hat = 0.228. We will act as if the sample is an SRS.

A 99% confidence interval for the proportion p of all college students who admit to frequent binge drinking uses the critical value $z^* = 2.576$, this value can be found in the bottom of Table C. The confidence interval is then...

We are 99% confident that the proportion of college students who engaged in frequent binge drinking lies between ______ and _____.

Assignment: p. 669 10.45 to 10.50

Choosing the Sample Size

In planning a study we may want to choose a sample size that will allow us to estimate the parameter within a given margin of error.

The margin of error in the approximate confidence interval for p is:

Here z* is the critical value for the level of confidence we want. Because the margin of error involves the sample proportion of successes p hat, we need to guess this value when choosing n. Call our guess p*. Here are two ways to get p*:

1.

2.

Sample Size for Desired Margin of Error:

To determine the sample size n that will yield a level C confidence interval for a population proportion p with a specified margin of error *m*, set the following expression for the margin of error to be less than or equal to *m*, and solve for *n*:

Where p^* is a guessed value for the sample proportion. The margin of error will be less than or equal to *m* if you take the guess p^* to be 0.5.

Example: Determining sample size

A company has received complaints about its customer service. They intend to hire a consultant to carry out a survey of customers. Before contacting the consultant, the company president wants tome idea of the sample size that she will be required to pay for. One critical question is the degree of satisfaction with the company's customer service, measured on a five point scale. The president wants to estimate the proportion p of customers who are satisfied (that is, who choose either "satisfied" or "very satisfied," the two highest levels on the scale). She decides that she wants the estimate to be within 3% at a 95% confidence level.

Since we have no idea of the true proportion p of satisfied customers, we decide to use $p^* = 0.5$. The sample size required is given by