Urban Design and Planning
URBDP 520 Quantitative Methods in Urban Design and Planning

Lecture Notes 6: Interval Estimation

Introduction

In the last chapter we talked about the probability that a sample mean would fall into some range. This chapter turns the question around and asks, if we know a sample mean and standard deviation, how big a range about that do we have to have to be fairly certain that the population mean lies within that range?

For example, you examine 100 hogs and find the mean and the standard deviation for the weight of the sample. If the sample mean is 500 pounds and the sample standard deviation is 50, within what range would the actual population mean lie, say, 95% of the time?

As another example, imagine that you select a random sample of 200 people who voted in the last election and asked them if they plan on voting for or against Proposition 666, which would change state government in a subtle but important way. The proportion of people who support the proposition is 49%. Into what range does the actual population (of people who voted in the last election) proportion fall with 90% certainty? With 95% certainty? With 99% certainty?

Say you've taken a sample and found a sample mean and sample standard deviation. Into what range will the actual population mean fall X% of the time?

The range is given a confidence interval. An X% confidence interval is based on a sample and will contain the actual population value with a probability of X/100.

The formula for a confidence interval is given by one of four formulas. They're scattered throughout the book, but I'm giving them to you here, all in one place, so you can see that they're all really the same and not at all difficult or intimidating.

	Large Sample[1]	**Small Sample**[2]
Population Mean	$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$
Population Proportion	$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	$\bar{p} \pm t_{\alpha/2,n-1} \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

where

$\bar{x}$ is the sample mean

$\bar{p}$ is the sample proportion

s is the sample standard deviation

n is the sample size

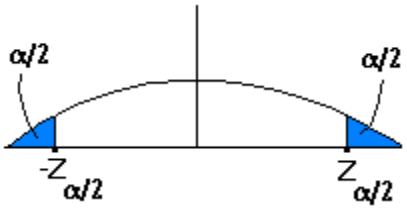
$z_{\alpha/2}$ is the value z such that the area to the right of z under the standard normal distribution is $\alpha/2$

$t_{\alpha/2,n-1}$ is the value t such that the area to the right of t under the t distribution with n-1 degrees of freedom is $\alpha/2$

[1] A large sample, for our purposes, will be any sample where $n > 29$. In reality, there is no perfect break point.

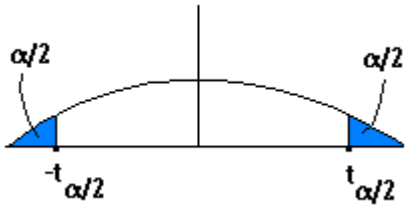
[2] The small sample formulas assume that the sample is drawn from a distribution which is approximately normal. I think we have called this "mound shaped" in the past.

Here are tables for $z_{\alpha/2}$ and $t_{\alpha/2,n-1}$



Confidence Level	**α**	**α/2**	**$Z_{\alpha/2}$**
90%	0.10	0.050	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.575

The idea is that the area under the curve between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$ is $1-\alpha$ and the probability that the population mean is between these two values is $1-\alpha$. That is why it is called a $(1-\alpha)$x100% confidence interval.



	80%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
	α=0.2	α=0.1	α=0.05	α=0.025	α=0.01	α=0.005	α=0.001	α=0.0005
	α/2=0.1	α/2=0.05	α/2=0.025	α/2=0.0125	α/2=0.005	α/2=0.0025	α/2=0.0005	α/2=0.00025
v	$t_{0.1,v}$	$t_{0.05,v}$	$t_{0.025,v}$	$t_{0.0125,v}$	$t_{0.005,v}$	$t_{0.0025,v}$	$t_{0.0005,v}$	$t_{0.00025,v}$
1	3.078	6.314	12.706	25.452	63.656	127.321	636.578	1273.155
2	1.886	2.920	4.303	6.205	9.925	14.089	31.600	44.703
3	1.638	2.353	3.182	4.177	5.841	7.453	12.924	16.326
4	1.533	2.132	2.776	3.495	4.604	5.598	8.610	10.305
5	1.476	2.015	2.571	3.163	4.032	4.773	6.869	7.976
6	1.440	1.943	2.447	2.969	3.707	4.317	5.959	6.788
7	1.415	1.895	2.365	2.841	3.499	4.029	5.408	6.082
8	1.397	1.860	2.306	2.752	3.355	3.833	5.041	5.617
9	1.383	1.833	2.262	2.685	3.250	3.690	4.781	5.291
10	1.372	1.812	2.228	2.634	3.169	3.581	4.587	5.049
11	1.363	1.796	2.201	2.593	3.106	3.497	4.437	4.863
12	1.356	1.782	2.179	2.560	3.055	3.428	4.318	4.717
13	1.350	1.771	2.160	2.533	3.012	3.372	4.221	4.597
14	1.345	1.761	2.145	2.510	2.977	3.326	4.140	4.499
15	1.341	1.753	2.131	2.490	2.947	3.286	4.073	4.417
16	1.337	1.746	2.120	2.473	2.921	3.252	4.015	4.346
17	1.333	1.740	2.110	2.458	2.898	3.222	3.965	4.286
18	1.330	1.734	2.101	2.445	2.878	3.197	3.922	4.233
19	1.328	1.729	2.093	2.433	2.861	3.174	3.883	4.187
20	1.325	1.725	2.086	2.423	2.845	3.153	3.850	4.146
21	1.323	1.721	2.080	2.414	2.831	3.135	3.819	4.109
22	1.321	1.717	2.074	2.405	2.819	3.119	3.792	4.077
23	1.319	1.714	2.069	2.398	2.807	3.104	3.768	4.047
24	1.318	1.711	2.064	2.391	2.797	3.091	3.745	4.021
25	1.316	1.708	2.060	2.385	2.787	3.078	3.725	3.997
26	1.315	1.706	2.056	2.379	2.779	3.067	3.707	3.974
27	1.314	1.703	2.052	2.373	2.771	3.057	3.689	3.954
28	1.313	1.701	2.048	2.368	2.763	3.047	3.674	3.935
29	1.311	1.699	2.045	2.364	2.756	3.038	3.660	3.918
30	1.310	1.697	2.042	2.360	2.750	3.030	3.646	3.902
40	1.303	1.684	2.021	2.329	2.704	2.971	3.551	3.788
50	1.299	1.676	2.009	2.311	2.678	2.937	3.496	3.723
60	1.296	1.671	2.000	2.299	2.660	2.915	3.460	3.681
70	1.294	1.667	1.994	2.291	2.648	2.899	3.435	3.651
80	1.292	1.664	1.990	2.284	2.639	2.887	3.416	3.629
90	1.291	1.662	1.987	2.280	2.632	2.878	3.402	3.612
100	1.290	1.660	1.984	2.276	2.626	2.871	3.390	3.598
110	1.289	1.659	1.982	2.272	2.621	2.865	3.381	3.587
120	1.289	1.658	1.980	2.270	2.617	2.860	3.373	3.578

EX: You take a sample of size 81 from a population. For the sample, you find a sample mean of 50 and a sample standard deviation of 27. Determine a 90% and a 95% confidence interval for the population mean.

90%: The appropriate value from the standard normal table is 1.645, so the range will be given by

$$\bar{x} \pm z_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 1.645\left(\frac{s}{\sqrt{n}}\right)$$

$$50 \pm 1.645 * \left(\frac{27}{\sqrt{81}}\right)$$

$$50 \pm 1.645 * \left(\frac{27}{9}\right)$$

$$50 \pm 1.645 * 3$$

$$50 \pm 4.935 \Rightarrow (45.065, 54.935)$$

So, 90% of the time, the population mean will fall into this range.

95%: The appropriate value from the standard normal table is 1.96, so the range will be given by

$$\bar{x} \pm z_{0.025}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right)$$

$$50 \pm 1.96 * \left(\frac{27}{\sqrt{81}}\right)$$

$$50 \pm 1.96 * \left(\frac{27}{9}\right)$$

$$50 \pm 1.96 * 3$$

$$50 \pm 5.88 \Rightarrow (45.12, 55.88)$$

So, 95% of the time, the population mean will fall into this range.

EX: A simple random sample of 50 items resulted in a sample mean of 32 and a sample standard deviation of 6.

A. Provide a 90% confidence interval for the mean.

A 90% confidence interval for the mean is given by

$$\bar{x} \pm z_{0.05}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 1.645\left(\frac{s}{\sqrt{n}}\right)$$

$$32 \pm 1.645 * \left(\frac{6}{\sqrt{50}}\right)$$

$$32 \pm 1.645 * \left(\frac{6}{7.07}\right)$$

$$32 \pm 1.396 \Rightarrow (30.604, 33.396)$$

A 95% confidence interval for the mean is given by

$$\bar{x} \pm z_{0.025}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right)$$

$$32 \pm 1.96 * \left(\frac{6}{\sqrt{50}}\right)$$

$$32 \pm 1.96 * \left(\frac{6}{7.07}\right)$$

$$32 \pm 1.663 \Rightarrow (30.337, 33.663)$$

A 99% confidence interval for the mean is given by

$$\bar{x} \pm z_{0.005}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 2.576\left(\frac{s}{\sqrt{n}}\right)$$

$$32 \pm 2.576 * \left(\frac{6}{\sqrt{50}}\right)$$

$$32 \pm 2.576 * \left(\frac{6}{7.07}\right)$$

$$32 \pm 2.186 \Rightarrow (29.814, 34.186)$$

Notice that as the confidence level increases, the width of the interval increases as well. This is because the interval needs to be wider in order for you to be more certain that the population mean will fall within it.

A confidence interval may also be expressed as a function of the standard error of the sample mean, which is approximately equal to $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$. This is just another way of saying the same thing.

The margin of error, or the difference between the sample mean and the population mean, can be expressed as a function of the standard error of the sample mean and of the level of confidence that the interval will contain the population mean.

EX: s=5.5, n=36
A. Standard Error of the Sample Mean is

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{5.5}{\sqrt{36}} = \frac{5.5}{6} = 0.917$$

B. Margin of error is $z_{\alpha/2}\sigma_{\bar{x}} = z_{\alpha/2}\frac{s}{\sqrt{n}} = z_{\alpha/2} \cdot 0.917$

With probability 0.75, the margin of error will be
$z_{0.25/2} \cdot 0.917 = 1.15 \cdot 0.917 = 1.05455$ or less.

With probability 0.90, the margin of error will be
$z_{0.10/2} \cdot 0.917 = 1.645 \cdot 0.917 = 1.508465$ or less

With probability 0.99, the margin of error will be
$z_{0.01/2} \cdot 0.917 = 2.575 \cdot 0.917 = 2.361275$ or less

C. The 99% confidence interval for the population mean if the sample mean is 48.6...

$$\bar{x} \pm z_{0.005}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm 2.575\left(\frac{s}{\sqrt{n}}\right)$$

$$48.6 \pm 2.575 * \left(\frac{5.5}{\sqrt{36}}\right)$$

$$48.6 \pm 2.575 * \left(\frac{5.5}{6}\right)$$

$$48.6 \pm 2.36 \Rightarrow (46.24, 50.96)$$

Small Samples: The t-distribution, a.k.a. Student's t-distribution
When n is less than 30, the sample mean won't be normally distributed, so we must use another distribution to determine the confidence interval for the population mean from a sample mean. This distribution is the t-distribution, a table for which is given on page 290 and 811 of the text. Note that the t-table is set up differently than the standard normal table.

The numbers across the top of the table are the area to the right of number of the table.
The numbers along the left hand side of the table are "degrees of freedom."

Operationally, the number of degrees of freedom you will want to look at will be equal to your sample size minus 1, n-1.

The area you want to choose to be to the right of the number in the table is $\alpha/2$.

For example, if you had a sample of size 20 and wanted a 95% confidence interval for the mean, you would choose the entry in the t-distribution table with 19=20-1 degrees of freedom and 0.025=[(1-0.95)/2] as the upper tail area, to give a value of 2.093.

Incidentally, there is also the assumption that the small sample is drawn from a population which is approximately normally distributed.

Basically, the small sample stuff is just like the large sample except that you need to use the t-distribution table with the appropriate number (n-1) of degrees of freedom.

You should notice that the t-distribution with infinite degrees of freedom is identical to the standard normal distribution, and it's pretty close with 60 degrees of freedom.

EX:
A simple random sample of 20 items from a normal population resulted in a sample mean of 17.25 and a sample standard deviation of 3.3.
A. Develop a 90% confidence interval for the population mean.

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{0.05,19}\left(\frac{s}{\sqrt{n}}\right)$$

$$17.25 \pm 1.729 * \left(\frac{3.3}{\sqrt{20}}\right)$$

$$17.25 \pm 1.729 * \left(\frac{3.3}{4.472}\right)$$

$$17.25 \pm 1.276 \Rightarrow (15.974, 18.526)$$

B. Develop a 95% confidence interval for the population mean

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{0.025,19}\left(\frac{s}{\sqrt{n}}\right)$$

$$17.25 \pm 2.093 * \left(\frac{3.3}{\sqrt{20}}\right)$$

$$17.25 \pm 2.093 * \left(\frac{3.3}{4.472}\right)$$

$$17.25 \pm 1.545$$

C. Develop a 99% confidence interval for the population mean

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{x} \pm t_{0.005,19}\left(\frac{s}{\sqrt{n}}\right)$$

$$17.25 \pm 2.861 * \left(\frac{3.3}{\sqrt{20}}\right)$$

$$17.25 \pm 2.861 * \left(\frac{3.3}{4.472}\right)$$

$$17.25 \pm 2.111$$

Determining the Necessary Sample Size

It is possible to determine how many observations are necessary to assure that the population mean will be within some given distance of the sample mean some percentage of the time.

That is, how many items do you have to sample to get as close to the real value as the boss wants without wasting a lot of money and effort by sampling unnecessary observations?

To do this, consider the previous equations for confidence intervals...

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

The confidence intervals have boundaries that are

$$z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

from the sample mean. That is, you can be (1-α)% certain that the population mean will lie within this distance of the sample mean. Call this distance E.

$$E = z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$E = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Now, if you need an E of a certain size, you can determine what n is necessary to achieve that. That is, if you need to be (1-α)% certain that you will be within E of the population mean, you can know what sample size you need to have.

$$E = z_{\alpha/2}\frac{s}{\sqrt{n}} \Rightarrow n = \left(z_{\alpha/2}\frac{s}{E}\right)^2$$

$$E = t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \Rightarrow n = \left(t_{\alpha/2,n-1}\frac{s}{E}\right)^2$$

Part of the problem is that you don't really know s until you take the sample.

EX:
Determine the sample size necessary so that there will be 95% probability, for a sample with s=7.2, that the margin of error will be E = 2, 1.5, 1

E=2: $n = \left(z_{\alpha/2}\frac{s}{E}\right)^2 = \left(z_{0.025}\frac{s}{E}\right)^2 = \left(1.96\frac{7.2}{2}\right)^2 = 49.787$ or n=50

E=1.5: $n = \left(z_{\alpha/2}\frac{s}{E}\right)^2 = \left(z_{0.025}\frac{s}{E}\right)^2 = \left(1.96\frac{7.2}{1.5}\right)^2 = 88.510$ or n=89

E=1: $n = \left(z_{\alpha/2}\frac{s}{E}\right)^2 = \left(z_{0.025}\frac{s}{E}\right)^2 = \left(1.96\frac{7.2}{1}\right)^2 = 199.158$ or n=200

So, to assure E is no greater than 2, 1.5 and 1, (with 95% probability) they would have to have samples of at least 50, 89 and 200.

There's a bit of a trick here in that you don't know what n will wind up being, so you don't really know what number of degrees of freedom to use. I like to start with the standard normal (z) distribution and, if the answer is sufficiently large, I just stick with that.

Confidence Intervals for Population Proportions
In talking about sampling distributions, we mentioned the proportion of a sample which had some characteristic, $\bar{p}$. From a sample proportion, we can construct a confidence interval for the population proportion.

The place you most often see this stuff is with public opinion polls where the people discussing the margin of error for the poll. What they don't tell you is the confidence level associated with that margin of error.

	Large Sample	**Small Sample**
Population Proportion	$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	$\bar{p} \pm t_{\alpha/2,n-1}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

EX: A few years back, a poll was taken of 400 Seattle-ites who would vote in the upcoming election. Of the sample, 52% said they would vote for Charlie Chong. Calculate a 95%
confidence interval for the portion of the population that would vote for Chong.

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$
$$0.52 \pm 1.96\sqrt{\frac{0.52 \cdot 0.48}{400}}$$
$$0.52 \pm 1.96 \cdot 0.025$$
$$0.52 \pm 0.049$$

EX: In a poll of 400 likely voters, 248 said they'd like to see more mayonnaise at the polls. Construct a 90% confidence interval for the proportion of all likely voters who would like to see more mayonnaise at the polls.

$n = 400, \bar{p} = \frac{248}{400} = 0.62$

The 90% confidence interval for p is

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$
$$0.62 \pm 1.645\sqrt{\frac{0.62 \cdot 0.38}{400}}$$
$$0.62 \pm 1.645 \cdot 0.02427$$
$$0.62 \pm 0.0399$$
$$(0.580, 0.660)$$

When p is small

Don't ask me to explain this, but according to the book, when the probability of a member of the sample having a particular characteristic is small or large (close to zero or close to one, something like 0.01 or 0.99 or so) the performance of the confidence interval is much better if you use the following:

$$p^* \pm z_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n+4}}$$

where

$$p^* = \frac{x+2}{n+4}$$

x is the number in the sample with the characteristic
n is the sample size

EX: Let's see how this adjustment affects these results. Imagine that you take a sample of size 80 and find that, of the people in the sample, 1 has a birthmark suggesting some type of fruit. Calculate a 95% confidence interval for the proportion of the population which has a fruit-shaped birthmark.

$$\bar{p} = \frac{1}{80} = 0.0125 \qquad \bar{p}^* = \frac{1+2}{80+4} = \frac{3}{84} = 0.0357$$

Using the original technique gives:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$0.0125 \pm 1.96\sqrt{\frac{0.0125 \cdot 0.9875}{80}}$$

$$0.0125 \pm 0.0243$$

using the modified technique gives

$$p^* \pm z_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n+4}}$$

$$0.0357 \pm 1.96\sqrt{\frac{0.0357 \cdot 0.9643}{84}}$$

$$0.0357 \pm 0.0397$$

Necessary Sample Size

As with determining the necessary sample size to get a confidence interval of a particular width for a population mean, we can also calculate necessary sample sizes for a population proportion confidence interval of a certain width.

The trick is that the answer depends on the sample proportion $\bar{p}$.

The confidence interval says that you're (1-α)*100% certain that the population proportion lies within

$$z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

If you want to select a sample size n so as to be (1-α)*100% certain that the population proportion lies within E of the sample proportion, the solution for n is

$$E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$n = z_{\alpha/2}^2 \frac{\bar{p}(1-\bar{p})}{E^2}$$

The trick here is that you don't actually know the sample proportion $\bar{p}$ until after you've taken the poll.

If you want to be safe, you can work with a *planning value* (a made up value of p used in choosing a sample size) of p=0.50. This will give you the largest value of n possible.

EX: How large a sample should we take to be 95% certain that the sample percentage is within +/-2.5% of the actual population percentage? Use p=0.25 and p=0.50 as planning values.

p=0.25

$$n = z_{\alpha/2}^2 \frac{\bar{p}(1-\bar{p})}{E^2}$$

$$n = 1.96^2 \frac{0.25 \cdot 0.75}{0.025^2}$$

$$n = 1152.48$$

You would need 1153 people.

p=0.50

$$n = z_{\alpha/2}^2 \frac{\bar{p}(1-\bar{p})}{E^2}$$

$$n = 1.96^2 \frac{0.50 \cdot 0.50}{0.025^2}$$

$$n = 2919.6$$

You would need 2920 people.

EX: You work for a political polling firm. Your clients want poll results correct to within some percentage 95% or the time. Assuming that (for safety) you assume p=0.50, how many people do you need to poll to get to within...

A. +/- 3% with 95% confidence?

$$n = 1.96^2 \frac{0.50 \cdot 0.50}{0.03^2} = 1067.1$$ or 1068.

B. +/- 2% with 95% confidence?

$$n = 1.96^2 \frac{0.50 \cdot 0.50}{0.02^2} = 2401$$

C. +/- 1% with 95% confidence?

$$n = 1.96^2 \frac{0.50 \cdot 0.50}{0.01^2} = 9604$$

Note that using p=0.50 gives the maximum value necessary as p(1-p) is maximized at p=0.50.