Section 6.1 - Estimating with Confidence

Statistics 104

Autumn 2004

Statistical Inference

Perform an experiment to find out the mean change in blood pressure when a Beta Blocker is given to the population of people with high blood pressure.

Experiment gives sample mean $\bar{x}$ (statistic)

Want to know the population mean μ (parameter)

In addition, may also be interested in variation of changes in blood pressure.

Experiment gives sample standard deviation s (statistic)

Want to know the population standard deviation σ (parameter)

$\bar{x}$ is a good guess for μ. s is a good guess for σ. (assuming a well designed experiment was performed)

Would like to be able to make stronger statements, such as

1. μ is likely between 15 and 18.

2. It is unlikely that $\sigma > 5$.

Will focus on procedures for making statement like the first one initially.

Confidence Intervals

Example: Observe 2500 women's heights

- $\bar{x} = 64$ inches
- $\sigma_x = 2.5$ inches (assume that this is known)
- $\bar{X} \sim N(\mu, 0.05)$ (approximately by the CLT)
- By the 68 - 95 - 99.7 rule, $\bar{x}$ is between $\mu - 2\sigma_{\bar{x}}$ and $\mu + 2\sigma_{\bar{x}}$ 95% of the time
- So $\bar{x}$ is within 0.1 of μ, 95% of the time.
- Can switch this to saying that μ is within 0.1 of $\bar{x}$, 95% of the time.

- So 95% of all samples will have μ in the interval

$$\bar{x} - 0.1 \text{ to } \bar{x} + 0.1$$
$$\text{e.g. } 63.9 \text{ to } 64.1$$

We say that we are **95% confident** that the mean height of women is between 63.9 and 64.1 inches.

Note that only two possibilities exist here

1. μ is in the interval calculated

   95% of all possible samples

2. The SRS taken was one where $\bar{x}$ was not within 0.1 of μ (μ not in the interval calculated).

   5% of all samples

Without outside knowledge we cannot know which of these two cases is true.

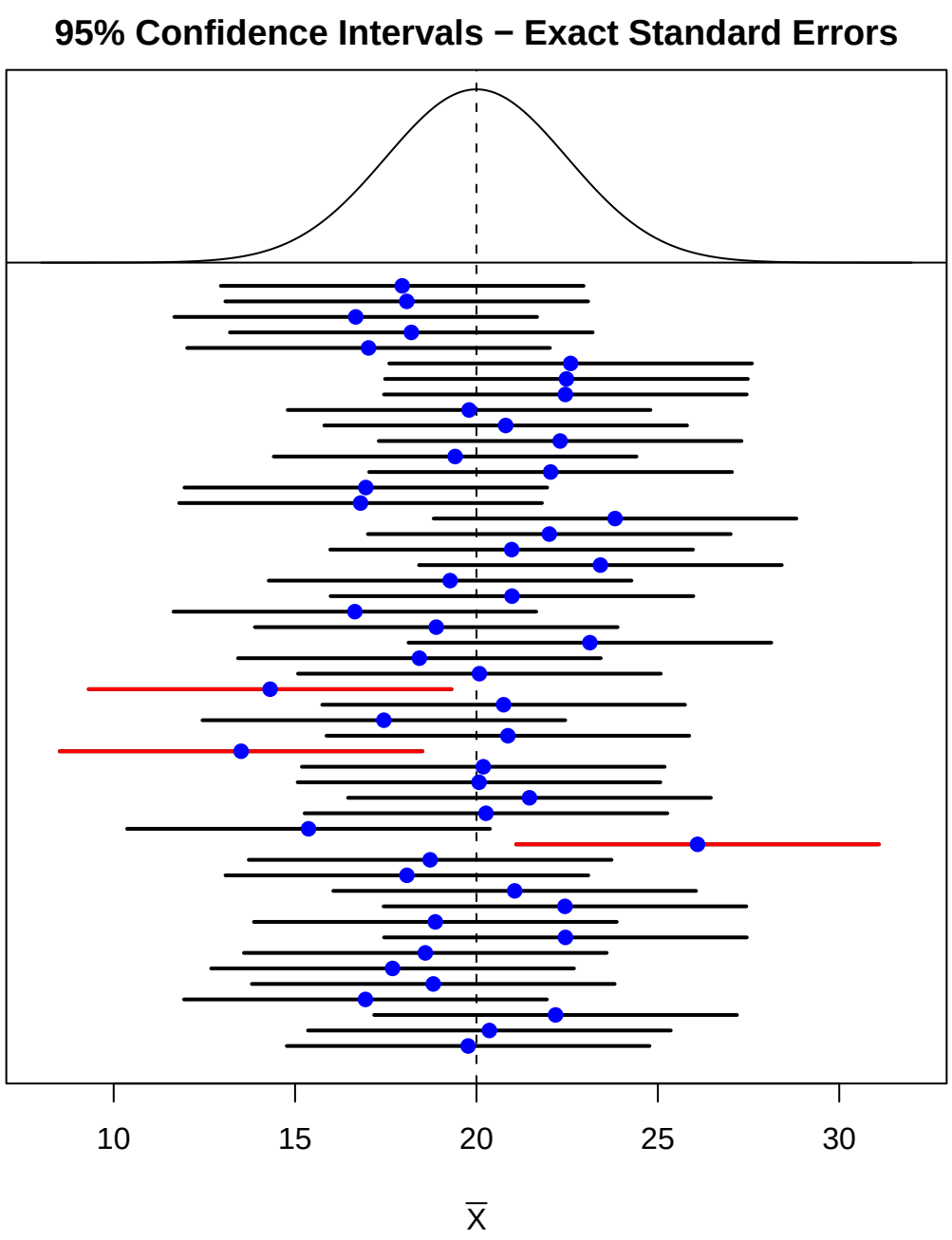
Confidence Interval

A level C confidence interval for a parameter θ is an interval computed from the data by a method that has probability C of producing an interval containing the true value of θ.

(C is known as the confidence level)

Want the truth in the interval $100C\%$ of the time, and out of the interval $100(1-C)\%$ of the time (over a large number of samples).

To display this idea, we can use simulation to examine the properties of confidence intervals. Here $\bar{X}$ is normally distributed and has a mean of 20 and a standard deviation 2.5. The 95% confidence intervals here are $\bar{x} \pm 5$



The vertical line is at the true mean

The dots are at the sample means for 50 different samples

The horizontal lines are the intervals for each of the 50 samples.

The density at the top is the normal sampling distribution of $\bar{X}$.

Confidence intervals (CI) give a set of plausible values for the unknown parameter.

Suppose that θ_1 is a value in the confidence interval for θ. This will occur if your statistic is not unlikely for the sampling distribution calculated when θ_1 is the truth.

The basic form for most confidence intervals is

$$\text{Estimate} \pm \text{Margin of Error}$$

So for the women's height example, a 95% confidence interval for the true mean women's height is

$$\bar{x} \pm 0.1$$

Constructing a confidence interval

- Choose a confidence level C - usually 90% or greater. 95% are 99% are the most popular. Most polls use 95% (19 times out of 20).

- Need a probability model for the data – e.g. a SRS of size n from $X \sim N(\mu, \sigma)$.

- From this we can get the sampling distribution of the statistic – e.g. $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

- Can use this to get the form of the interval.

When the sampling distribution for the estimator is normal, a confidence interval for μ is of the form

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

This assumes that σ, the standard deviation of population distribution of the observations, is known. If not, we need to use something slightly different.

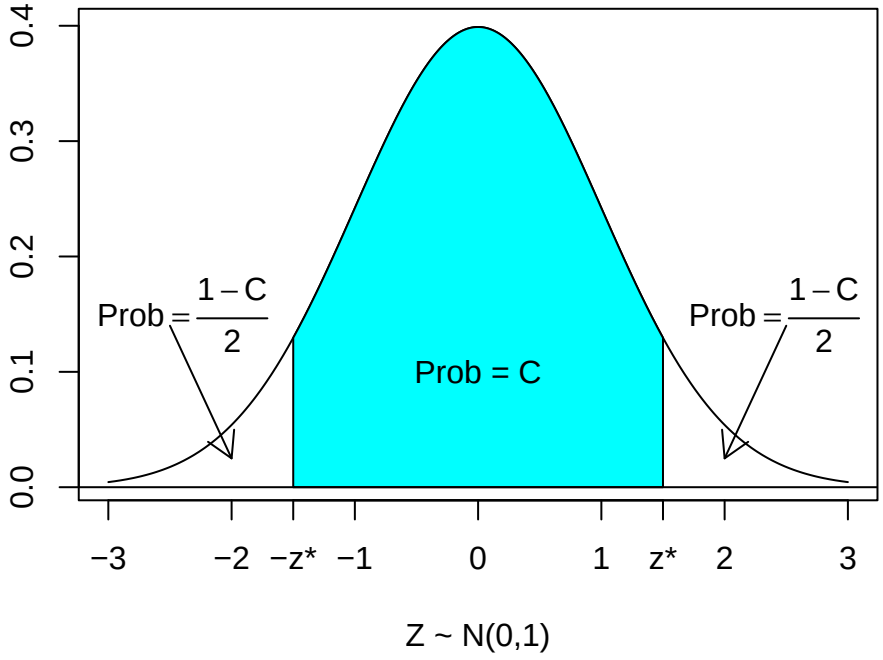
The standard deviation of the sampling distribution, $\frac{\sigma}{\sqrt{n}}$, is sometimes referred to as the standard error.

So Margin of Error = Critical Value $\times$ Standard Error

The critical value z^* comes from the standard normal distribution and depends on the choice of C. It satisfies

$$P\left[-z^* \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z^*\right] = C$$

$\bar{x}$ is within z^* standard errors of μ, $100C\%$ of the time.



These values can be determined using the normal probability table.

Find the cell in the table with the probability closest to

$$\frac{1-C}{2}$$

However its easier with Table D. Take the entry of the row denoted by z^* in the column with Confidence level C.

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

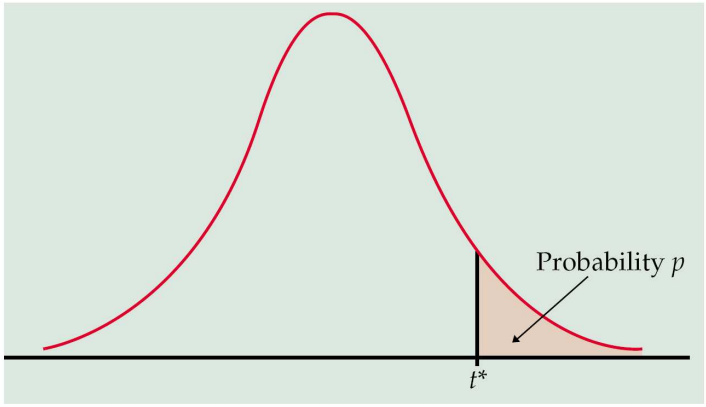


TABLE D t distribution critical values

	Upper tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

Example: VCR tape times

A company claims that their 6 hour VCR tapes on average have 360 minutes of taping time with a standard deviation of 8 minutes. A sample of 64 tapes was tested and $\bar{x} = 352$ minutes.

Does the claim appear to be valid with this data? Lets examine with a 90% confidence interval.

$$90\%\ CI \Rightarrow z^* = 1.645$$

$$\begin{aligned} CI &= 352 \pm 1.645 \times \frac{8}{\sqrt{64}} \\ &= 352 \pm 1.645 \\ &= (350.355, 353.645) \end{aligned}$$

Now lets look at a 95% CI

$$95\%\ CI \Rightarrow z^* = 1.960$$

$$\begin{aligned} CI &= 352 \pm 1.960 \times \frac{8}{\sqrt{64}} \\ &= 352 \pm 1.960 \\ &= (350.04, 353.96) \end{aligned}$$

With either interval, the manufacturer claim doesn't seem valid, based on this data.

What if someone else did a similar study, but only with 16 observations. Assume that $\bar{x} = 352$ again.

$$95\%\ CI \Rightarrow z^* = 1.960$$

$$\begin{aligned} CI &= 352 \pm 1.960 \times \frac{8}{\sqrt{16}} \\ &= 352 \pm 3.92 \\ &= (348.08, 355.92) \end{aligned}$$

90% CI, $n = 64$	95% CI, $n = 64$	95% CI, $n = 16$
352 ± 1.645	352 ± 1.960	352 ± 3.920

What affects the size of a confidence interval

1. Confidence level

   Larger $C \rightarrow$ wider interval

2. Sample size

   Larger $n \rightarrow$ narrower interval

3. σ

   Larger $\sigma \rightarrow$ wider interval

Usually a narrow interval with high confidence level is wanted.

Suppose that a further study is to be performed and that a 95% confidence interval with a margin of error of 1 is desired. How big should n be?

Desired interval of the form

$$\bar{x} \pm 1$$

Margin of error

$$z^* \frac{\sigma}{\sqrt{n}} = 1$$

So for a margin of error of 1,

$$z^* \sigma = \sqrt{n} \Rightarrow n = (z^* \sigma)^2$$
$$n = (1.96 \times 8)^2 = 245.8$$

So at least 246 tapes are needed.

In general, to get a margin of error of m, the sample size needs to be at least

$$n = \left(\frac{z^*\sigma}{m}\right)^2$$

which is gotten by solving

$$m = z^*\frac{\sigma}{\sqrt{n}}$$

So if a 95% CI with margin of error is 0.5 is desired

$$n = \left(\frac{1.96 \times 8}{0.5}\right)^2 = 4 \times (1.96 \times 8)^2 = 983.4$$

So at least 984 ($= 4 \times 246$) are needed.

To halve the width of an interval, you need 4 times as many observations.

This comes from the standard error depending on the square root of the number of observations.

What if you can't get the desired sample size?

1. Relax the desired margin of error

2. Lower the confidence level

3. Change the study design to lower the standard error. The improvement might come from a more efficient use of resources or possibly lowering σ.

   For example, it can be shown that stratified sampling with proportional allocation ($n_i = n\frac{N_i}{N}$) will always have a smaller sampling error than a SRS of the same size. Other stratified schemes can do even better than proportional allocation.

Cautions about Confidence Intervals

1. The formula given is only correct if the data is a SRS from the population (or something equivalent)

2. Other sampling schemes may have different standard errors, some lower, some higher. The formula will have to be adjusted for this.

3. Confidence intervals based on haphazardly collected data are usually meaningless. Fancy formulas cannot rescue badly produced data. (Garbage In – Garbage Out)

4. Since the interval described is based on $\bar{x}$, outliers could have a large effect. You should look for outliers before calculating the interval.

5. Non-normality. If the data is not generated from a normal distribution, $\bar{x}$ will not be exactly like a draw from a normal distribution. This implies that the actually confidence level may not be exactly C. For large samples, this should not be a big problem (due the the Central Limit Theorem), but can be for small samples. You should check for evidence of non-normality before proceeding.

6. Known σ. The interval presented assumes that we know the population standard deviation, a dubious proposition at best. However we can estimate σ with s. We can modify the procedure to use s instead of σ and calculate a valid confidence interval.

7. Bias. The confidence interval proposed can only be valid if the sampling procedure is unbiased as the margin of error only accounts for the sampling variability.