Lecture 13 (MWF) Confidence intervals for the mean and margin of error

Suhasini Subba Rao
Review of previous lecture

A student did several experiments (measuring the size of some cells), due to the variability in conditions the numbers she got varied from from experiment to experiment. Here is a summary of her data:

0.025, 0.025, 0.057, 0.064, 0.054, 0.035, 0.047, 0.059, 0.045.

(see http://www.stat.tamu.edu/~suhasini/teaching651/msa.txt).

• The raw data does not convey any meaningful information.

• One is interested in understanding some feature in the ‘population’ it came from. The mean, $\mu$, gives us some idea of its center.

• The sample mean of this data set is $\bar{X} = 0.046$. On the face it, this number appears meaningless, since a different set of experiments will
give you a different sample mean - it is not a fixed value. However, we have learnt that this number comes from a distribution which is close to normal, that has a mean $\mu$ (the mean of the population) and with spread/standard error $\frac{\sigma}{\sqrt{9}}$.

• However, the population standard deviation is unknown. But it can be estimated from the data using the formula:

$$ s = \sqrt{\frac{1}{7} ([0.025 - 0.046]^2 + \ldots + [0.045 - 0.046]^2)} = 0.15. $$

We can use this sample standard deviation to replace the unknown population standard deviation. However, we need to ‘correct’ for the fact that it is an estimate - we cover this in lecture 14.

• Using this correction we can:
– Evaluate probabilities.
– Construct confidence intervals.
– To do statistical tests (later on).
Example 1

A forester wishes to estimate the average number of trees per acre over a 2000-acre plantation. She can use this information to determine the total timber volume in the plantation. The standard deviation for the distribution of the number of trees in an acre is 12.1. A random sample of $n = 50$ 1-acre plots are selected and examined. It is found that the average number of trees per acre (based on this sample) is 27.3.

(a) Use this information to construct a 95% CI for the mean number of trees per acre.

(b) Use this information to construct a 99% CI for the mean number of trees per acre.
(c) Here we twist the story. Suppose that it is known that the mean number of trees per acre is 32. You want to know the number trees in 2000 acres. Construct an interval, centered about the mean, which will contain this number with 95%.
Solution 1

The sample average $\bar{X} = \frac{1}{50} \sum_{k=1}^{50} X_i$. The variance for one acre is known to be $12.1^2$, hence the standard error for the sample mean is $12.1/\sqrt{50}$. The sample size $n = 50$, which is large enough to assume normality.

(a) Therefore the 95% CI is:

$$[27.3 - 1.96 \frac{12.1}{\sqrt{50}}, 27.3 + 1.96 \frac{12.1}{\sqrt{50}}].$$

The length of interval is $2 \times 1.96 \frac{12.1}{\sqrt{50}}$.

(b) The 99% CI is

$$[27.3 - 2.57 \frac{12.1}{\sqrt{50}}, 27.3 + 2.56 \frac{12.1}{\sqrt{50}}].$$
The length of interval is $2 \times 2.57 \frac{12.1}{\sqrt{50}}$.

(c) The sample average over 2000 acres will be approximately normally distributed with mean 32 (this is given in the question) and standard deviation $\frac{12.1}{\sqrt{2000}} = 0.27$ (since this the standard error of the sample mean) ($\bar{X} \sim \mathcal{N}(32, 0.27)$). Therefore, by centering about 32, there is a 95% chance the sample mean will lie in the interval 

$[32 \pm 1.96 \frac{12.1}{\sqrt{2000}}] = [31.47, 32.52]$ (return to lectures 9 and 11 to see this calculation).

Since the sample mean is the number of number of trees in 2000 acres divided by 2000, the total number of trees in 2000 acres is the sample mean TIMES 2000.

$$\text{average} = \frac{\text{Total in hundred acres}}{2000}.$$
Therefore, there is a 95% chance the number of tree that are in 100 acres will be in the interval $2000 \times [31.47, 32.52] = [6400 \pm 1.96 \times 12.1 \times \sqrt{2000}] = [62940, 65040]$.

- **Advanced (ignore if it is too difficult)** Of course we do not know what the mean is. Instead we need to replace the mean $\mu = 32$ (which in reality is unknown) with the sample mean $\bar{X} = 27.3$ in this case the interval where we are 95% confident most tree counts will lie is

$$
\left[ 27.3 \times 200 \pm 1.96 \times 12.1 \times \sqrt{1 + \frac{2000}{50}} \right],
$$

notice this interval is wider than the interval given in (c) and this is too account for the uncertainty in the sample mean (which is not the true population mean). Explanation: This calculation is because $2000(\bar{Y}_{2000} + \bar{X}_{50}) \sim N(0, 12.1^2(2000^{-1} + 50^{-1}))$. 

A social worker is interested in estimating the average time outside prison a first time offender spends outside prison before they re-offend (if at all). A random sample of $n = 150$ first time offenders are considered. Based on this data it is found that the average time they spend 3.2 years away from prison. The sample standard deviation is 1.1 years. Stating all assumptions construct a 99% CI for the true average $\mu$.  

**Example 2**
Solution 2

• The sample mean is $\bar{X} = 3.2$. The sample standard deviation is 1.1.

• The sample size is large $n = 150$, hence we can assume normality of the sample mean $\bar{X}$. Moreover, since we have estimated the standard deviation $s = 1.1$ using 150 observations (relatively large sample), we can assume it is a good estimator of the true sample standard deviation $\sigma$. Hence in our calculations we will use $s = 1.1$ in place of the true standard deviation $\sigma$.

Therefore the sample variance of the sample mean $\bar{X}$ is $1.1^2/150$.

• The 99% CI is $[3.2 - 2.56 \frac{1.1}{\sqrt{150}}, 3.2 + 2.56 \frac{1.1}{\sqrt{150}}]$. The length is $2 \times 2.56 \frac{1.1}{\sqrt{150}}$. 
Example 3 (JMP) Calf weights at birth

Now we analyse the birth weights of calves using JMP (wt 0), it can be found at http://www.stat.tamu.edu/~suhasini/teaching651/cow_birth_weights.csv.

(a) Construct a 90% CI for the population mean (the average birth weight of all newborn calves).

(b) Evaluate the probability of getting a sample mean of 93.21 or over, given that the true mean is 91.
The data is summarised below:

**Summary Statistics**

- Mean: 93.215909
- Std Dev: 7.7784719
- Std Err Mean: 1.1726488
- Upper 95% Mean: 95.560781
- Lower 95% Mean: 90.851037
- N: 44
• By using the distribution options in JMP we can see that the number of observation is 44, the sample mean is 93.21 pounds (remember this is the average of the data we have been given), the standard deviation is 7.44 and the standard error (the amount of variability in the sample mean) is $7.44/\sqrt{44} = 1.07$ (see how you can get the standard deviation from the standard error - it is simply $s.d. = 1.07 \times \sqrt{44}$).
Solution 3(a)

• To construct the 90% CI confidence interval for the mean, we have to see whether it is sensible to assume that the sample mean is close to a normal distribution. There are two reasons this assumption appears to be plausible:
  
  – The sample size is large, \( n = 44 \), thus by the central limit theorem regardless of the original distribution the sample mean for such a large sample size is close to normal.
  
  – Looking at the histogram of the original data, it appears to be symmetric without a large skew. This tells us that an average (sample mean) based on a random sample taken from this distribution will easily converge to a normal even for relatively small sample sizes.

• Based on the above, a 90% CI is for the population mean is \([93.21 - 1.64 \times 1.07, 93.21 + 1.64 \times 1.07] = [91.45, 94.96]\).
Solution 3(b)

• To evaluate the probability of getting a sample mean of 93.21 or over, given that the true mean is 91 we use normality of the sample mean.

• Since the sample mean is close to normal, with standard error of 1.07, then under the assumption it has a population mean of 91 we can say $\bar{X} \sim N(91, 1.07^2)$. Thus probability $P(\bar{X} > 93.21) = P(Z > \frac{93.21 - 91}{1.07}) = P(Z > 2.06) = 1 - 0.98 = 0.02$ (using tables).

• How to interpret this result. It basically tells is that we have a 2 percent chance of observing a sample mean of 93.21 or less, given that the true population mean weight is 91 pounds. Since we have a sample mean of 93.21, this means that the true mean weight being 91 pounds seems only slightly plausible, it is more likely the true mean is larger. Notice that 90% CI for the mean does not contain 91.
Margin of Error

• According to a survey, Americans walk on average 40 miles a week with a margin of error of 2.5 miles.

• What does this mean in terms of confidence intervals?

• 40 miles corresponds to the average number of miles walked in the sample, the margin of error is the plus and minus in the confidence interval.

In other words for a 95% confidence interval

\[
\text{Margin of Error} = 1.96 \times \frac{\sigma}{\sqrt{n}}.
\]

• The smaller the margin of error the more precisely we can pin point the population mean. Of course it is worth bearing in mind that we can
never be sure that our confidence interval contains the mean, which is why we prescribe a level (such as a 95%) to the interval. In other words, we can never be sure that the population mean is within the prescribed margin of error of the sample mean.
Sample size and margin of error

• We compare the 95% confidence intervals for $n = 9$ and $n = 25$. We see

\[
\begin{align*}
n=9 & \quad [\bar{X} - 1.96 \frac{\sigma}{\sqrt{9}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{9}}] \\
n=25 & \quad [\bar{X} - 1.96 \frac{\sigma}{\sqrt{25}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{25}}]
\end{align*}
\]

What are the lengths of the above intervals?

- For $n=9$ it is $2 \times 1.96 \frac{\sigma}{\sqrt{9}}$ the margin of error is $1.96 \times \frac{\sigma}{\sqrt{9}}$.
- For $n=25$ it is $2 \times 1.96 \frac{\sigma}{\sqrt{25}}$ the margin of error is $1.96 \times \frac{\sigma}{\sqrt{25}}$.

Observe that the length and margin of error does not depend on $\bar{X}$. 
Example: \( \bar{X} = 10.38, \sigma^2 = 33. \)

\[
\begin{align*}
\text{n=9} & \quad [10.38 - 1.96\frac{\sqrt{33}}{\sqrt{9}}, 10.38 + 1.96\frac{\sqrt{33}}{\sqrt{9}}] = [6.63, 14.13] \\
\text{n=25} & \quad [10.38 - 1.96\frac{\sqrt{33}}{\sqrt{25}}, 10.38 + 1.96\frac{\sqrt{33}}{\sqrt{25}}] = [8.12, 12.63]
\end{align*}
\]

- We see that the second interval is smaller than the first interval. The second interval has a smaller margin of error.

- When the sample size is large our estimator is more reliable. We ‘see’ this in the confidence interval. The confidence interval will be narrower so we can locate the mean more precisely and the margin of error is smaller.
The standard deviation and the margin of error

• We see that the variability in the sample measured by the standard deviation $\sigma$ will have an impact of the reliability of an estimator and it’s margin of error.

• Example: Suppose $\bar{X} = 10.38$, $n = 9$, but the variability in the populations are different:

\[
\sigma^2 = 33 \quad [10.38 - 1.96\frac{\sqrt{33}}{\sqrt{9}}, 10.38 + 1.96\frac{\sqrt{33}}{\sqrt{9}}] = [6.63, 14.13]
\]

\[
\sigma^2 = 100 \quad [10.38 - 1.96\frac{\sqrt{100}}{\sqrt{9}}, 10.38 + 1.96\frac{\sqrt{100}}{\sqrt{9}}] = [3.38, 16.91]
\]

• The more variable in the population the more variability there is in the
sample mean. The only way we can compensate for this variability is to use a larger sample size (recall that the standard error is $\sigma/\sqrt{n}$).
**How large an interval to use?**

- You read in a newspaper that the proportion of the public that supports gay marriage is now 55% ± 15%.

- This means a survey was done, the proportion in the survey who supported gay marriage was 55% and that confidence interval for the population proportion is $[55 - 15, 55 + 15]\% = [40, 70]\%$.

- This is an extremely large interval, it is so wide, that it is really not that informative about the opinion of the public.

- The reason it is too wide is that the sample size is too small.

- This experiment was not designed well.
• Typically, before data is calculated, we need to decide how large a sample to collect.

• This is usually done by deciding how much ‘above and below the estimator seems reasonable. For example, an interval of the type \([55-3, 55+3]\% = [52, 58]\%) is more information.

• The 3% is is the margin of error. Given a margin of error we can then determine the sample size to collect.
Choosing the sample size for estimating $\mu$

- In an ideal world we would have a very large sample size. A large sample size gives a small standard error, which in turn gives a narrow confidence interval and a smaller margin of error.

- However, obtaining very large samples can be impossible for many different reasons:
  - To have a very large sample size would be nice, but often it can be too costly or infeasible.
  - A sample size which is too small is not informative.

- How can one determine the number of observations to be included in a sample? How to choose the sample size $n$?

Answer: Usually we have a margin of error in mind. We can accept the...
reliability of an estimator up to a certain margin or error. Once we know what margin of error is acceptable we can then choose the sample size.
Formula for choosing the sample size

• To choose the sample size according to the margin of error, we need to know (or guess apriori) the standard deviation $\sigma$ (if we don’t know what it is, then we err on the cautious side and use a value that seems reasonable but large).

• We recall that in the confidence interval:

$$[\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}]$$

the margin of error is the $E = 1.96 \frac{\sigma}{\sqrt{n}}$. In other words, the length of the confidence interval is $2 \times$ Margin of Error $= 2 \times 1.96 \frac{\sigma}{\sqrt{n}}$. Therefore, if we want to choose the sample size such that the margin of error is $E$ we need to solve for $E = 1.96 \frac{\sigma}{\sqrt{n}}$. Ie. $n = \left(\frac{1.96\sigma}{E}\right)^2$. 
• Example: Suppose we guess that \( \sigma^2 = 3 \) and we want the margin of error \( E = 0.25 \). The confidence interval is

\[
[\bar{X} - 1.96 \frac{\sqrt{3}}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sqrt{3}}{\sqrt{n}}]
\]

and we solve \( 1.96 \frac{\sqrt{3}}{\sqrt{n}} = 0.25 \). This gives

\[
n = \left( \frac{2 \times 1.96 \sqrt{3}}{0.5} \right)^2 = 184.4
\]

• Of course, a larger value of \( n \) will give a smaller margin of error, so we round up and use \( n = 185 \). In other words, for this experiment we need to choose a sample size of at least 185 to get a margin of error of 0.25.
General CIs and tolerable error

The above was for 95% CI.

- How should we choose $n$ for the 99% CI, in this case $\alpha = 1\%$?

- If we want to use a 99% CI, we first look up 0.5% in the z-tables, $z_{0.5\%} = 2.57$. Then we need to solve $E = 2.57\frac{\sigma}{\sqrt{n}}$

$$n = \left(\frac{2.57\sigma}{E}\right)^2.$$

- In general, for the $(100 - \alpha)\%$ CI use the formula $E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and solve it to give $n = \left(z_{\alpha/2} \frac{\sigma}{E}\right)^2$. 

Example 4: Heights

• Researchers want to estimate the mean height of students at a university (in feet) within a margin of error of 0.04 (using a 95% CI level). They know that the standard deviation of a randomly sampled student is 0.113. How many students must they sample to achieve their specifications?

• Solution 4:
  – Use the formula.
  – $E = 0.04$.
  – Using the formula we have
    $$n = \frac{(1.96)^2(0.113)^2}{(0.04)^2} = 30.65.$$  
  – Hence they must sample 31 people such that a 95% confidence interval has length $2 \times 0.04 = 0.08$. 
Example 5: Caffeine content

- The caffeine content is coffee is being analysed and it is known that standard deviation of a randomly selected coffee is 7.1mg.

Suppose 100 cups of coffee are analysed, and the total weight of caffeine in all the cups is \( \sum_{i=1}^{100} X_i = 11000 \text{mg} \), construct a 95% CI for the mean caffeine content.

Construct an 80% CI for the mean caffeine.

Find the minimum number of coffees which must be analysed for the 95% CI to have length 1mg (in other words the margin of error is 0.5mg)?

- Solution 5:

- The total weight of caffeine for the 100 cups is 11000 mg. Therefore the sample average of caffeine per cup is \( \bar{x} = 11000/100 = 110 \).
• Calculating the 95% CI (use the formula or calculate yourself): \( z_{0.05/2} = z_{0.025} = 1.96 \), \( n = 100 \), \( \sigma^2 = 7.1^2 \) and \( \bar{X} = 110 \).

• The CI is:

\[
\left[ 110 - 1.96 \times \frac{7.1}{\sqrt{100}}, 110 + 1.96 \times \frac{7.1}{\sqrt{100}} \right] = [108.6, 111.4].
\]

• To construct an 80% CI only one thing has to change, that is we only have to replace the 1.96 above with another number. To find this value go to the normal table and look inside it for 0.1, you should see –1.28. Replace 1.96 with 1.28 to give

\[
\left[ 110 - 1.28 \times \frac{7.1}{\sqrt{100}}, 110 + 1.28 \times \frac{7.1}{\sqrt{100}} \right] = [109.1, 110.9].
\]
• If we want the length to be 1mg, then the interval

\[ 110 - 1.96 \times \frac{7.1}{\sqrt{n}}, \quad 110 + 1.96 \times \frac{7.1}{\sqrt{n}} \]

must have length 1. This means that

\[ 1 = 2 \times 1.96 \times \frac{\sigma}{\sqrt{n}} \]

(if you want to write it in terms of \( E \), then \( E = 1/2 \) ) Solve this (or use the formula) to give

\[ n = \frac{1.96^2 \times 7.1^2}{0.5^2} = 775. \]
• Hence we need to sample at least 775 cups to get such a small interval with 95% CI. This makes sense - the more in our sample the more accurate the estimator.
Example 6

How large a sample size do we require such that the margin of error for a 95% confidence interval for the mean of human heights is maximum 0.25 inch. The standard deviation is unknown, but it is believed that $\sigma$ lies somewhere between 2-5 inches.

• Why this question matters: In general the standard deviation will be unknown. But we can ‘guess’ limits on how large or small it is based on own expertise.
Solution 6

The more variable the data the larger the confidence interval. Therefore, if we given a range of standard deviations and our aim is that the margin of error should be no larger than 0.25 (ie. 0.25 or less), then we need to use the largest standard deviation in the margin of error calculation.

- MoE = $1.94 \times \frac{5}{\sqrt{n}}$. Solving for this leads to $n = 1537$.

- For any other $\sigma < 5$, using $n = 1537$ will lead to a Margin of Error which is smaller than 0.25.

- Note that if we use $\sigma = 2$, and solve for the equation will lead to $n = 246$. However, for any $\sigma > 2$ using $n = 246$ will lead to a Margin of Error which is larger than 0.25.
Example 7

A confidence interval for the length of parrots is [4, 10] inches. It is based on a sample size $n$. By what factor should the sample size increase such that the margin of error reduces to 1?
Solution 7

• The original margin of error is 3. Thus $1.96 \times \frac{\sigma}{\sqrt{n}} = 3$.

• We want to increase the sample size such that it decreases to 1.

\[
1.96 \times \frac{\sigma}{\sqrt{\text{Factor} \times n}} = 1
\]

\[
\Rightarrow 1.96 \times \frac{\sigma}{\sqrt{\text{Factor} \times \sqrt{n}}} = \frac{3}{\sqrt{\text{Factor}}} = 1.
\]

Solving for this we see that we need to increase the sample size by factor 9 in order to decrease the margin of error by a factor 3.

• An extremely large increase in sample size has to be made for a moderate reduction in margin of error.