Lecture 12 (MWF) Distribution of the sample mean

Suhasini Subba Rao
Review of previous lecture: Why confidence intervals?

- **General Example** Suppose we observe $X$ which is known to follow a normal distribution $N(\mu, \sigma^2)$, where the mean $\mu$ is unknown but $\sigma^2$ is known. We want to locate $\mu$ given the observation $X$. Suppose $X = 3$, this by itself does not really tell us much about the location of the true mean $\mu$. But

  (i) $[3 - 1.96\sigma, 3 + 1.96\sigma]$ tells us with 95% confidence the unknown mean $\mu$ lies in this interval.
  
  (ii) $[3 - 1.64\sigma, 3 + 1.64\sigma]$ tells us with 90% confidence the mean $\mu$ lies in this interval.
  
  (iii) $[3 - 2.56\sigma, 3 + 2.56\sigma]$ tells us with 99% confidence the mean $\mu$ lies in this interval.

- By placing values for $\sigma$, you will see that the smaller the confidence level (95% is smaller than 99%), the smaller the interval. Conversely the
larger the confidence we require the larger the interval needs to be. Hence, there is a trade off between pin-pointing the location of the mean and how much confidence we want in the interval. If we want to pin-point the mean, the interval should be smaller but then the confidence we have in that interval will be less. If we want more confidence that the mean lies in that interval, then the interval should be larger. But a larger interval is not very informative about the location of the mean.

An extreme example is an interval which goes over the entire range of $X$ (which if it is normal is $-\infty$ to $\infty$!). The mean is definitely inside this interval (100% confidence), but it’s not very informative about the location of the true mean!

- **Example II** Consider the example above.

(i) In the case that the standard deviation is $\sigma = 1$, the 95% CI is $[3 - 1.96, 3 + 1.96]$. 

(i) In the case that the standard deviation is $\sigma = 100$, the 95% CI is $[3 - 196, 3 + 196]$.

- The larger the variance the wider the CI. When the variance is large we need a larger interval to ensure that it includes the unknown mean $\mu$.

- **Objective of this class**: To understand why, under certain conditions, $[\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}]$ is a 95% confidence interval for the mean.
• There are two problems with what we have done so far.

(a) We cannot change the standard deviation of the population, but we would like the interval to be narrow. How can we make the interval narrower (without changing the confidence level or the standard deviation of the population - which in general is NOT possible)?

**Answer** We will show below that this can be done by estimating the mean using a larger sample size.

(b) We are assuming the the data is normal, which in reality will only be true for some populations. As we saw in the previous lecture, without normality the confidence level does not match the real level of confidence.

**Answer** We will see that the object of interest is the not the ‘parent’ population, but the population of the sample mean. The distribution
of the sample mean will look more normal as the size of the sample from which it was evaluated grows. This is known as the central limit theorem.
One bag of M&Ms

Suppose that I have only one M&M bag and can only use this as an estimator of the mean. We now the distribution for the number of M&Ms in a bag looks like:

The standard error of this estimator is (=standard deviation) is 4.65. Because there is a lot of variability in the number of M&Ms in one bag.
Five bags of M&Ms

Suppose that I sample five bags of M&Ms. I take the average of these 5 bags. The average of these 5 bags is an estimate of the mean and is random.

The standard error (=standard deviation of sample mean) is 2.09.
Ten bags of M&Ms

Suppose that I sample ten bags of M&Ms. I take the average of these 10 bags. The average of these 10 bags is an estimate of the mean and is random.

The standard error (standard deviation) in this case is 1.46.
The standard error

The examples above were based on just one sample. However, we know that one sample will be highly variable (as measured by the sample mean). Recall that the M&M example, the variability (measured by standard deviation) for one bag of M&Ms was 4.65. However, we know (in fact we saw) that the variability decreased when we considered the sample mean (average). Indeed as we increased the sample size the variability decreased.

The numbers are given in the Table:

<table>
<thead>
<tr>
<th></th>
<th>original population</th>
<th>sample mean (n= 5)</th>
<th>sample mean (n= 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>13.45</td>
<td>13.45</td>
<td>13.45</td>
</tr>
<tr>
<td>stand. dev.</td>
<td>4.65</td>
<td>2.09</td>
<td>1.46</td>
</tr>
</tbody>
</table>

We know that the variability decreases. Does it decrease in a predictable way?
• If the standard deviation, $\sigma$, in the original population is known (say $\sigma = 4.65$). Then the standard error (the standard deviation of the sample mean) follows the formula:

$$\frac{\sigma}{\sqrt{n}},$$

where $n$ is the size of the sample.
Applied to the M&M example

<table>
<thead>
<tr>
<th></th>
<th>original pop.</th>
<th>sample mean (5)</th>
<th>sample mean (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>13.45</td>
<td>13.45</td>
<td>13.45</td>
</tr>
<tr>
<td>data stand. err.</td>
<td>4.65</td>
<td>2.09</td>
<td>1.46</td>
</tr>
<tr>
<td>stand. err.</td>
<td>$4.65 = \frac{4.65}{\sqrt{1}}$</td>
<td>$2.09 = \frac{4.65}{\sqrt{5}}$</td>
<td>$1.46 = \frac{4.65}{\sqrt{10}}$</td>
</tr>
</tbody>
</table>

Thus the formula $\sigma/\sqrt{n}$ is correctly predicting the variability in the sample mean.
The larger sample, the smaller the standard error of the sample mean

- To recollect suppose $X_i$ is a random variable with mean $\mu$ and variance $\sigma^2$. We have a sample $X_1, \ldots, X_n$. The sample mean $\bar{X}$ with the following properties:
  - The mean of the sample mean $\bar{X}$ is $\mu$ (in other words the distributions of the data and the sample means share the same center).
  - If the sample size is $n$ the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$.

- What else do we notice about the relationship between standard error and sample size?

- As the sample size $n$ gets larger, $\sigma$ stays the same (you cannot change the variability of a population) BUT the standard error gets smaller.
• This is what we would expect, the larger the sample size, the more reliable the sample mean is as estimator of the population mean.

• Having figured out the mean and standard error of the sample mean (center and spread), we now have to figure out it’s distribution.
The sample mean of a population which is normal

- If the original population is normally distributed, then the sample mean (of a random sample taken from that population) will be normal with the same mean as the original data and standard error $\sigma/\sqrt{n}$.

Example

- Suppose the heights of female are normally distributed with mean 64.5 inches and standard deviation 2.5 inches ($X \sim \mathcal{N}(64.5, 2.5)$).

  I randomly sample 4 women and evaluate their sample mean. The sample mean is estimating the population mean which is 64.5. The sample mean will be normally distributed with mean 64.5 inches and standard error $2.5/\sqrt{4} = 1.25$, ie. $\bar{X}_4 \sim \mathcal{N}(64.5, 1.25)$.

- Of course, in reality the population mean is UNKNOWN, and we want to locate it based on a sample. So let us suppose that womens heights
are normally distribution with unknown mean $\mu$ and standard deviation $2.5 \ N(\mu, 2.5^2)$. I collect a random sample of 4 women, their heights are 63, 64, 66.2, 68.3, the sample mean is $\bar{X} = 65.375$ ($\bar{X}_4 \sim N(\mu, 1.25)$).

- Since the average is normal we can use the normal distribution to construct the confidence interval for the mean height. The 95% confidence interval for the mean $\mu$ is

$$\left[65.375 \pm 1.96 \times \frac{2.5}{\sqrt{4}}\right] = [62.925, 67.825]$$

Therefore with 95% confidence we believe the mean female height lies in that interval.

- **Question** What happens if original population is not normal?

- Then a remarkable thing happens...
Lecture 12 (MWF) Standard errors, the Central Limit Theorem and confidence intervals

The central limit theorem

The central limit theorem:

• Suppose \( X_1, \ldots, X_n \) is a sample from a population with mean \( \mu \) and variance \( \sigma^2 \).

• If the sample size \( n \) is large, then the sample mean

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,
\]

(approximately) has the distribution

\[
\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).
\]
This means that the random variable $X_i$ has mean $\mu$ and variance $\sigma^2$ (that is the height of a randomly chosen person has mean height $\mu$ and variance $\sigma^2$), then the average taken of a sample of $n$ individuals, let us call this $\bar{X}$, has mean $\mu$ and variance $\sigma^2/n$.

- Observe how the variance does from $\sigma^2$ (variance of individual) to $\sigma^2/n$ variance of average of $n$ people.


- **IMPORTANT** The data DOES not become more normal as you increase the sample size. I.e. As you increase the sample size the QQplot DOES NOT magically become more normal looking. The distribution of the data is fixed. If human heights are bimodal they will bimodal if the sample size is large, the distribution of M&Ms is integer valued and
multimodal regardless of the sample size. Increasing the sample size does not make it more normal.

- **So what becomes normal?** It’s the sample mean, it is the average based on the sample. The larger the sample, the more normal will the distribution of the sample mean be.
How large is large?

- How large is large is a difficult question, and varies from data to data. The ‘rule of thumb’ is that the sample size should be about $n = 30$ for the CLT of the sample mean to kick in. However, this is a rule of thumb. Below we give some details, these can be checked with the applet.

  - If the data is close to normal - then a far smaller sample size is required such that the sample mean is close to normal and you can construct reliable confidence intervals.

  - On the other hand if the data is highly non-normal (you can check this by making a QQplot), more observations are required for the sample mean to be normal. What you are looking for:

    - If the data is highly skewed then a far larger sample size is required for the CLT to kick in.
If the data takes just a few numerical discrete values, then a far larger sample size is required for the CLT to kick in.

- What does tell us about the reliability of CIs? Remember the 95% CI for the mean is constructed under the assumption the data is normal.

If the sample size is quite small and the distribution of the data does not appear to be normal (which can be checked using the QQplot), then the sample mean will not be very normal. This means that the 95% CI for the mean is not reliably a 95% CI (recall the M&M examples in the start of lecture 11). In other words the sample mean will not lie within the confidence interval 95% of time - usually it will be a lower percentage. This means the information you are trying to convey, that is there is 95% confidence in the interval is wrong.

- Go to the Statcrunch to see how distribution influences the sample size. Try sampling from these three populations using
different sample sizes: http://www.stat.tamu.edu/~suhasini/teaching651/CLT-check.csv
• The original data is highly RIGHT skewed (top plot). The distribution of the sample mean, based on 30, is less skewed. But it is still skewed - the QQplot of the right is of the sample mean (the right skew is quite apparent).

• Therefore the confidence interval based on a sample size of 30 and the
normal distribution will not be reliable.
• The original data is numerical discrete - taking only a few values (top plot). The distribution of the sample mean, based on 30 - the QQplot of the right is of the sample mean (the horizontal lines are still there but not that apparent.

• Therefore the confidence interval based on a sample size of 30 and the
normal distribution will be fairly reliable.
Example 1

You want to rent an unfurnished one bedroom place in Dallas and you would like to know the mean monthly rent. You know that the variability (standard deviation) of apartment prices in Dallas is 60 dollars ($\sigma = 60$). You take a random sample of 10 apartments and this has a sample mean of 1009.27 dollars. http://www.stat.tamu.edu/~suhasini/teaching651/apartment_dallas.dat Construct a 95% for the mean price of one apartment rental. The JMP output is given below.

![JMP output image]
Solution1: Checking for normality

Before constructing a 95% confidence interval for mean we check to see if the sample mean is close to normal. Below we see a plot of the sample mean (based on sampling from the data) and its corresponding QQplot. We see that it is sufficient close to normal that the we can be sure that we really do have 95% confidence in the interval.
Solution 1

• From previous knowledge our general view is that home prices are right skewed. So this is probably true of one bedroom prices in Dallas too (though the skew may not be too much, since we are focusing on only one bedroom prices, which cannot get too expensive and this is Dallas, where accommodation is reasonable).

• A histogram and QQplot of the data is given, there seems to be a slight right skew.

• We want to construct a 95% confidence interval for the mean based on a sample of 10. An average based on 10 will be closer to normal. This means that the 95% confidence we will construct will be something we have close to 95% confidence in.
• **Observe** The theoretical (or population) standard deviation is assumed to be 60, whereas the sample standard deviation calculated from the data is 49.92. For now we will use 60 in our calculations. However, it is worth noting that in most applications the true population standard deviation is unknown and we need to use the standard deviation calculated from the data. If we are using the estimated standard deviation to construct the CI then we need to make some adjustments in the confidence interval (we discuss this in Lecture 14).

• Since the sample mean is estimating the mean it will be centered about the unknown mean price \( \mu \) with standard error \( \frac{60}{\sqrt{10}} = 19 \). Using the 95% confidence interval for the mean is

\[
[1009 \pm 1.96 \times 19] = [972, 1046].
\]

• **Important** The above interval DOES **NOT** tell us that 95% of the rental
Lecture 12 (MWF) Standard errors, the Central Limit Theorem and confidence intervals

prices lie in this interval (a common misconception). It is simply an interval where we believe with 95% confidence the mean apartment rental price lies.
Example 2: Evaluating probabilities

A patient is classified as having low potassium if her level is below 3.5. A patient’s mean potassium level is 3.58 with standard deviation 0.4. This means she does not have low potassium, however, her true level is unknown to doctors, so it needs to be diagnosed from her blood samples. A doctor decides to take the average of her blood samples and diagnose low potassium if her sample mean level is below 3.5.

(a) Suppose that 10 blood samples are taken. Calculate the probability of her being wrongly diagnosed with low potassium.

(b) Suppose that 49 blood samples are taken, calculate the probability of her being wrongly diagnosed.

(c) What happens to the chance of wrong diagnoses when we increase the sample size.
Solution 2

(a) We do not know if the distribution of potassium in the blood samples is normally distributed or not. However, the question asks for a probability based on the sample mean calculated from 10 blood samples. Though 10 is a relatively small sample size we will hope that the average based on 10 is sufficiently close to being normally distributed.

– The sample mean based on 10 is

\[ \bar{X}_{10} \sim N(3.58, \frac{0.4^2}{10}) = N(3.58, 0.126^2) \]

– She is wrongly diagnosed if her sample mean is below 3.5. This means we need to evaluate \( P(\bar{X}_{10} < 3.5) \). We make the z-transform

\[ z = \frac{3.5 - 3.58}{0.126} = -0.63 \]

From the tables we see this corresponds to 26.3%. In other words there is a 26.3% chance of a wrong diagnoses based on 10 samples. A plot is given below:
Lecture 12 (MWF) Standard errors, the Central Limit Theorem and confidence intervals

The area to the left is the chance of the patient being misdiagnosed.
(b) The sample size is now 49, we can be more sure that the average based on 49 observations is closer to the normal distribution than in the previous sample. Therefore, the probabilities calculated for (b) will be closer to the truth than those calculated for (a).

– The sample mean based on 49 is

$$\bar{X}_{10} \sim N\left(3.58, \frac{0.4^2}{49}\right).$$

– We make the z-transform

$$z = \frac{3.5 - 3.58}{0.057} = -1.4.$$ From the tables we see this is 8%. In other words there is a 8% chance of a wrong diagnoses based on 49 samples. A plot is given below:
The area to the left is the chance of the patient being misdiagnosed.

(c) Because her true mean level of 3.58 is quite close to the 3.5 there is large chance of misdiagnoses with small sample sizes. However, as the sample size grows the chance of a misdiagnoses goes down.
Example 3: Constructing confidence intervals

Let us return to the potassium set-up above. The above method of using 3.5 as the threshold has certain disadvantages. It does not take into account variation in the sample mean. If there is a lot of variation in the sample mean, the people with low potassium may not be diagnosed (because their sample mean is above 3.5) and people with normal potassium levels may be falsely diagnosed. A more effective method is to construct a confidence interval for the mean and use this as a means of diagnoses. If the the interval is all below 3.5 it suggests the patient may have low potassium.

Suppose the standard deviation in potassium levels is known to be 0.4, 20 blood samples are take and the sample mean evaluated. Calculate and interprete the 95% confidence intervals in each of the following cases:

(i) The sample mean is 3.3.
(ii) The sample mean is 3.4.

(iii) The sample mean is 3.6.

(iv) The sample mean is 3.9
Lecture 12 (MWF) Standard errors, the Central Limit Theorem and confidence intervals

**Solution 3**

(i) The confidence interval is \( [3.3 \pm 1.96 \times 0.4/\sqrt{20}] = [3.12, 3.47] \). Since [3.12,3.47] is trying to locate the mean, and this interval contains all \( \mu < 3.5 \). This suggests the patient has low potassium.

(ii) The confidence interval is \( [3.4 \pm 1.96 \times 0.4/\sqrt{20}] = [3.22, 3.57] \). Since [3.22,3.57] is trying to locate the mean, and this interval contains both \( \geq 3.5 \) and \( < 3.5 \). A diagnoses is unclear with this sample.

(iii) The confidence interval is \( [3.6 \pm 1.96 \times 0.4/\sqrt{20}] = [3.32, 3.77] \). Since [3.32,3.77] is trying to locate the mean, and this interval contains both \( \geq 3.5 \) and \( < 3.5 \). A diagnoses is unclear with this sample.

(iv) The confidence interval is \( [3.9 \pm 1.96 \times 0.4/\sqrt{20}] = [3.62, 4.07] \). Since [3.62,3.67] is trying to locate the mean, and this interval contains only \( \geq 3.5 \). This suggests the patient has normal levels of potassium.
Class Experiment II

• You may remember that in our last class experiment we tried to construct a 95% confidence interval for the mean height. We did this based on just one sample (!!), fortunately heights are close to normal so our CI which was constructed under the assumption of normality was quite accurate (recall that 92% of the students in the class had the mean height lying in this interval).

• But the interval was quite wide \([\text{yourheight} - 1.96 \times 3.8, \text{yourheight} + 1.96 \times 3.8]\), which is an interval of length 14.9. This is because there is a lot of variability in height, which corresponds to a large standard deviation (of 3.8).

• Suppose we want to locate the mean more accurately.

• To do this, pair up with your neighbour and find the average height.
• We know from the above that the standard error of this sample mean is $3.8/\sqrt{2}$.

• Construct a confidence interval for the mean height using your average

\[
[your\ average\ height\ (you\ and\ your\ buddy) - 1.96 \times 3.8/\sqrt{2}, \ your\ average\ height - 1.96 \times 3.8/\sqrt{2}] = [average - 5.3, average + 5.3].
\]

• Check to see whether the true mean height in the population, which is 67.1 lies in this interval.

• Notice that the interval is narrower than the interval with just one person.

This is because the standard error (now we use the term standard error because this is the standard deviation of the sample mean) is smaller than the standard deviation of the population ($3.8/\sqrt{2}$ verses 3.8).
• However, the big assumption is the use of 1.96, which assumes normality of the average. Is the normality assumption okay???