Lecture 12 (MWF) Distribution of the sample mean

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

- The principle behind a confidence interval is that usually an observation will be within a “few” standard deviations of the population mean \( \mu \).

- This is the same as saying that “usually” the mean will be within a “few” standard deviations of the observations.

- We have used the word *usually* and *few* very loosely. But they can be made precise if we know the distribution from which the observations was taken from.

- If the observations is normal, then with a very large probability (99.6%) the observation will be within 3 standard deviations of the mean.

- Equivalently the mean will be within 3 standard deviations of the
observations. I.e. with probability 99.6% the mean $\mu$ will lie in the interval $[X - 3\sigma, X + 3\sigma]$. $X$ and $\sigma$ are given numbers.

- 3 standard deviation makes the interval very wide. So often we compromise on the level to make the standard deviation narrower.

**Example Cls for the mean and different levels of confidence**

Suppose we observe $X$ which is known to follow a normal distribution $N(\mu, \sigma)$, where the mean $\mu$ is unknown but the standard deviation $\sigma$ 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.
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

(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.

- Keeping $\sigma$ fixed, you will observe that the smaller the confidence level (95% is smaller than 99%), the smaller the interval. Conversely the larger the level of confidence, the longer the confidence interval.

- 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
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

$X$ (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!

- **Confidence interval for different $\sigma$**

  (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 $\mu$. 


There are two problems with we encountered in the discussion at the end of lecture 11 and the above.

(a) We cannot change the standard deviation of the population (this is not possible, the variation is what it is). But we would like the interval to be narrow, so that it is more informative.

We will show below that this can be done by using the sample mean rather than just one observations.

(b) We are assuming the the data is normal, which in reality will only be true for certain populations. Without normality, normality the stated confidence level does not match the true level of confidence. It is analogous to one just making-up the level of confidence!
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.

- **Main objective of this lecture**: To understand why, under certain conditions

\[
[ \bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}]
\]

is a 95% confidence interval for the mean \( \mu \) (regardless of whether the observations are normal or not). And in general

\[
[ \bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}]
\]

is a \((1 - \alpha) \times 100\%\) confidence interval for the mean.
The Height data

- Recall the height data from lecture 4:

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>66.4</td>
<td>63.8</td>
<td>65.8</td>
<td>68</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68</td>
<td>74</td>
<td>68</td>
<td>61</td>
</tr>
<tr>
<td>Sample 2</td>
<td>65</td>
<td>62</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>Sample 3</td>
<td>67</td>
<td>64</td>
<td>68</td>
<td>66</td>
</tr>
<tr>
<td>Sample 4</td>
<td>69</td>
<td>62</td>
<td>71</td>
<td>66</td>
</tr>
</tbody>
</table>

- Here four samples each of size 5 were drawn.
- Our objective is to understand the behaviour of the averages. So far 31 samples have been collected. The task is cumbersome. Therefore I make a computer do it.
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

One height

This is what the distribution of heights looks like.

- The software below will draw 5 heights from this distribution and calculate the average. It will do this 1000s of times and plot the histogram of all the averages.
- It always draws the sample with *replacement*. This means the population can be finite (in this example only 80). If the sampling is done without replacement the population size must be very large with respect to the sample size. This again is due to the independence assumption we make about the sample (see Example 4, in Lecture 7).
One sample mean based on 5 heights

Here the computer is drawing 5 heights from the distribution. This sample is given in the middle plot. It is 60, 66, 68, 69 and 69 and the sample mean is 66.4.

The average of this sample is the green box on the lower plot.
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

Histogram of sample mean based on 5 heights

Rather than the 31 samples I managed to draw in class, here is 30,000 averages. Each based on a sample size of 5. The histogram is the bottom plot. The standard error ($=\text{standard deviation of sample mean}$) is $1.8 = 4.3/\sqrt{5}$. 

\begin{tabular}{|l|}
\hline
Population & \\
\hline
Mean & 66.7125 \\
Median & 66 \\
Unadjusted Std. dev. & 4.1958 \\
\hline
\end{tabular}

\begin{tabular}{|l|}
\hline
Samples & \\
\hline
Sample size & 5 \\
Mean & 67.4 \\
Median & 66 \\
Std. dev. & 3.0725 \\
\hline
\end{tabular}

\begin{tabular}{|l|}
\hline
Sample means & \\
\hline
# of Samples & 30000 \\
Mean & 66.7197 \\
Median & 66.6 \\
Std. dev. & 1.8828 \\
\hline
\end{tabular}
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

The QQplot of the averages based on $n = 5$

This is a QQplot of all those averages (not the original heights). This is equivalent to the QQplot of the average of heights given in Lecture 10. But the difference there is that I only could make it using 31 samples. If you compare the plot in Lecture 10 with this plot, they will look quite similar.
The Central Limit Theorem and confidence intervals where \( \sigma \) is known

Histogram of sample mean based on 25 heights

Each based on a sample size of 10. The computer took many samples, the histogram is the bottom plot. Observe how “normal” it looks.

The standard error (\( = \) standard deviation of sample mean) is 0.84 = 4.3/\( \sqrt{25} \). This corresponds to the distribution being far narrower than when \( n = 5 \).
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where \( \sigma \) is known

**The QQplot of the averages based on \( n = 25 \)**

This is a QQplot of all those averages (not the original heights). Observe how normal the averages are.
The average number in five bags of M&Ms

Another example, we sample five bags of M&Ms and take the average. Observe that the distribution of the number of M&Ms in a bag is far from normal.

The standard error (standard deviation) in this case is $2.07 = 4.6 / \sqrt{5}$. 
The average number in ten bags of M&Ms

In this example, we sample ten bags of M&Ms and take the average. Observe that the distribution of the number of M&Ms in a bag is far from normal.

The standard error (standard deviation) in this case is $1.46 = \frac{4.6}{\sqrt{10}}$. 

<table>
<thead>
<tr>
<th>Population</th>
<th>Mean</th>
<th>13.5412</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>Unadjusted Std. dev.</td>
<td>4.6452</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Std. dev.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample means</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Samples</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Std. dev.</td>
</tr>
</tbody>
</table>
**Mean and standard error for M&M example**

For M&M example, the variability (measured by standard deviation) for one bag of M&Ms was 4.65. However, we saw that the variability decreased when we considered the sample mean (average). Indeed as we increased the sample size the variability decreases in a predictable fashion:

<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 the sample the smaller the standard error of the sample mean

- To recollect suppose $X_i$ is a random variable with mean $\mu$ and standard deviation $\sigma$. 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 normally distributed observations

- If $X_1, X_2, \ldots X_n$ are independent normally distributed random variables with mean $\mu$ and standard deviation $\sigma$. Then the sample mean $\bar{X}$, will be normally distributed 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 ($X \sim N(64.5, 2.5)$).

  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 $2.5/\sqrt{4} = 1.25$, i.e. $\bar{X} \sim N(64.5, 1.25)$.

- Of course, in reality the population mean is UNKNOWN, and we want to locate it based on a sample. So let us suppose that women's heights are normally distributed with unknown mean $\mu$ and standard deviation 2.5 $N(\mu, 2.5)$. I collect a random sample of 4 women, their heights are 63, 64, 66.2, 68.3, the sample mean is $\bar{X} = 65.375$ ($\bar{X}_4 \sim N(\mu, 1.25)$).
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where \( \sigma \) is known

- 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. Below we construct 100 CIs for the mean. Using the sample mean drawn from this distribution. We see that we really do have 95\% confidence in the interval.
The central limit theorem

The central limit theorem:

- Suppose $X_1, \ldots, X_n$ is an independent sample from a population with mean $\mu$ and standard deviation $\sigma$.
- 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

$$\bar{X} \sim N \left( \mu, \frac{\sigma}{\sqrt{n}} \right).$$

- We already “showed” this to be true using the height and M&M example on the previous page.
- The independent assumption is important (this is why we assume the sampling is with replication, if not, the population size must be far larger than the sample size, such that it is almost independent).
If there is dependence in the observations, then the standard error will not be $\sigma / \sqrt{n}$. If there is extreme dependence, the normal distribution will not hold.

- **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.
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

**Example**

- Consider the heights of students from the 301 class. The standard deviation for this data set is $\sigma = 4.2$, but as we saw in the previous slides, the distribution of heights is not normal.

- A sample of size 5 is drawn 60, 66, 68, 69 and 69 and the sample mean is 66.4. The 95% CI is

$$
\left[ 66.4 - 1.96 \times \frac{4.2}{\sqrt{5}}, \, 66.4 + 1.96 \times \frac{4.2}{\sqrt{5}} \right] = [62.7, 70.1].
$$
To see if we really do have 95% confidence in this interval, we run a simulation (where we draw from the distribution of heights). Observe that about 95% of the intervals cross the vertical black line, which means we have 95% confidence in the intervals.
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 hold. However, like all “rule of thumbs” it is not a very reliable method.
  - If the data is close to normal - then for a far smaller sample size the sample mean is close to normal. The implication of this is that reliable confidence intervals can be constructed with the stated level of confidence.
  - On the other hand if the data is highly non-normal (you can check this by making a QQplot of the data), 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 sample mean is normal.
  If the sample size is quite small and the distribution of the data does not appear to
be normal this can severely compromise the reliability of the confidence interval. The implication is that 95% of the time, the mean will not lie in the interval.
Lecture 12 (MWF) The Central Limit Theorem and confidence intervals where $\sigma$ is known

**Skewed distributions: Baseball salaries**

- The top left plot is the distribution of baseball salaries.
- 30 baseball salaries are drawn from this distribution. The green plot is a histogram of sample means.
- The original data is highly RIGHT skewed (top plot). The distribution of the sample mean, using 30 observations, is less skewed. But it is still skewed - the QQplot of the right is of the sample mean (the right skew is still apparent).
- This means in the tails of the distribution of the sample mean it will differ from normality. We use the tails of the normal distribution to calculate the confidence interval.
Impact of non-normality of sample mean on CI

- Here we construct over 100,000 99% CIs for the mean using observations drawn from baseball salaries with sample size 30.
- We saw on the previous page that the distribution of the sample mean deviated slightly from normality in the tails.
- You see the very small effect here. Where rather than having 99% confidence in the interval we have 98.94% confidence.
- This difference is so small, it hardly matters. This shows that the CLT works even for when the original data is highly non-normal.

However, if the original data has an even more pronounced skew. The difference between the stated level of confidence and the true level of confidence will be large for even moderate sample sizes.
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. Construct a 95% confidence interval for the mean price of one apartment rental. The JMP output is given below.
Solution 1: Checking for normality

Before constructing a 95% confidence interval for mean we check to see if the sample mean is close to normal. The plot below is sampling from the observed data, so the mean given in the plots below are not the true population mean. However, the distribution (the green histogram), does give us an idea of the actual distribution of the sample mean (this is called bootstrap).

We see that that the green plot is sufficient close to normal that the we can be sure that we really do have 95% confidence in the interval.
Solution 1: The CI

- Since the sample mean is estimating the mean it will be centered about the unknown mean price $\mu$ with standard error $60/\sqrt{10} = 19$. The 95% confidence interval for the mean is

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

- 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).

- Important The above interval DOES NOT tell us that 95% of the rental 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 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 assume that the sample mean based on 10 is sufficiently close to being normally distributed. The distribution of the sample mean based on 10 is

\[
\bar{X}_{10} \sim N \left(3.58, \frac{0.4}{\sqrt{10}} = 0.126 \right).
\]

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.
(b) The sample size is 49. The sample mean is more likely to be 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 distribution of the sample mean based on 49 is

$$\bar{X}_{10} \sim N(3.58, \frac{0.4}{\sqrt{49}} = 0.04).$$

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. As the sample size grows, the chance of misdiagnoses based on the sample mean decreases.
Example 3

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 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 taken and the sample mean evaluated. Calculate and interpret 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
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.
Example 4

A social worker is interested in estimating the average time outside prison a first time offender spends 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 $\mu$. 
Solution 4

- 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 $\sigma$.

The standard error of the sample mean $\bar{X}$ is $1.1/\sqrt{150}$.

- The 99% CI is
  \[ [3.2 - 2.57 \frac{1.1}{\sqrt{150}}, 3.2 + 2.57 \frac{1.1}{\sqrt{150}}] \]

  The length of the CI $2 \times 2.57 \frac{1.1}{\sqrt{150}}$.  

Example 5: Calf weights at birth

Now we analyse the birth weights of calves using JMP, 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 for birth weight of newborn calves.
(b) Evaluate the probability of getting a sample mean of 93.21 or over, given that the true mean is 91.
Solution 5

- In JMP Analyze > Distribution > Highlight variable of interest (same as before). A window should pop-up with histogram and summary statistics.
- From 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 $7.44/\sqrt{44} = 1.07$ (you can get the standard deviation from the standard error; $s.d. = 1.07 \times \sqrt{44}$).
- JMP also gives the 95% CI for the mean [90.8, 95.6].
To construct the 90% CI confidence interval for the mean, we have to check 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 \([93.21 - 1.64 \times 1.07, 93.21 + 1.64 \times 1.07] = [91.45, 94.96]\).
Changing the confidence level in JMP

The default confidence level in JMP is 95%. But this can be changed. Over the red arrow in Summary Statistics, right click on Customize Summary Statistics. The following window will pop-up. Change the level at the bottom of the page. Observe there is a slight difference in our answer and the answer derived by JMP. This is because JMP is using the t-distribution to calculate the CI. The t-distribution is used because the standard deviation used in the construction is not the population standard deviation but the sample standard deviation.
• To evaluate the probability of getting a sample mean of 93.21 or over, given that the true mean is $\mu = 91$ 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 $\mu = 91$ we can say $\bar{X} \sim \mathcal{N}(91, 1.07)$. This gives

$$P(\bar{X} > 93.21) = P \left( Z > \frac{93.21 - 91}{1.07} \right) = P(Z > 2.06) = 1 - 0.98 = 0.02$$

• How to interpret this result. We are calculating the plausibility of the mean $\mu \leq 91$.

• The result says, there is 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.

• What we did was test $H_0 : \mu \leq 91$ vs $H_A : \mu > 91$. 2% corresponds to the p-value; the plausibility of the data when the $\mu \leq 91$. As 2% is relatively small, a more likely explanation is that $\mu > 91$. 
