Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 12 (MWF) Distribution of the sample mean

Suhasini Subba Rao

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 μ (regardless of whether the observations are normal or not). And in general

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

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

Return: The Height data

- Recall the height data from lecture 4:

	Sample 1	Sample 2	Sample 3	Sample 4
	68	65	67	69
	74	62	65	62
	68	60	64	71
	61	66	68	72
	61	66	65	66
Average	66.4	63.8	65.8	68

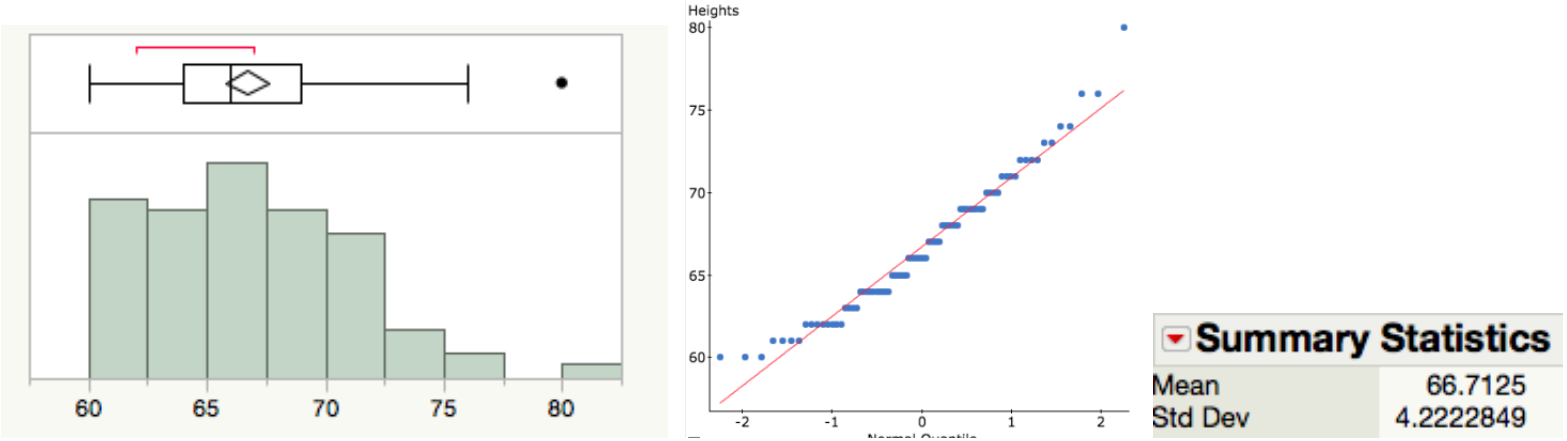
- Here four samples each of size 5 were drawn.

- Our objective is to understand the behaviour of the averages. This task

is cumbersome (and is only a thought experiment). Therefore I make a computer do it, close to an infinite number of times.

One height

This is what the distribution of heights looks like.



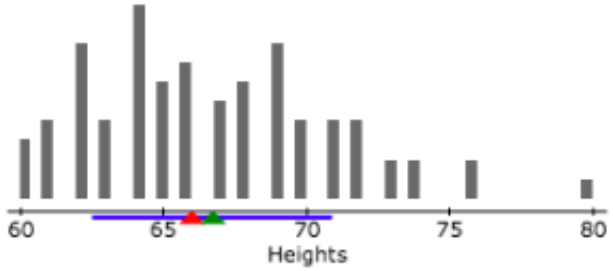
- The software below will draw 5 heights from this distribution and calculate the average. It will do this 1000s of times and plot the histogram of all the averages.

- It always draws the sample with *replacement*, and can do it many times (the number does not matter).

One sample mean based on 5 heights

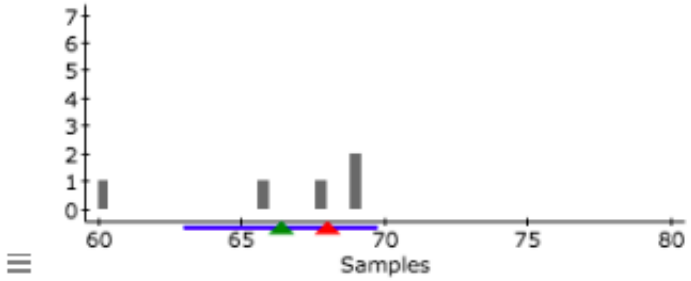
Sampling Distributions

1 time	5 times	1000 times	Reset	Analyze	Info



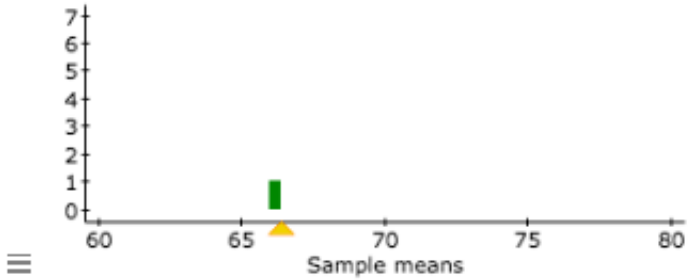
Population

Mean	66.7125
Median	66
Unadjusted Std. dev.	4.1958



Samples

Sample size	5
Mean	66.4
Median	68
Std. dev.	3.3823



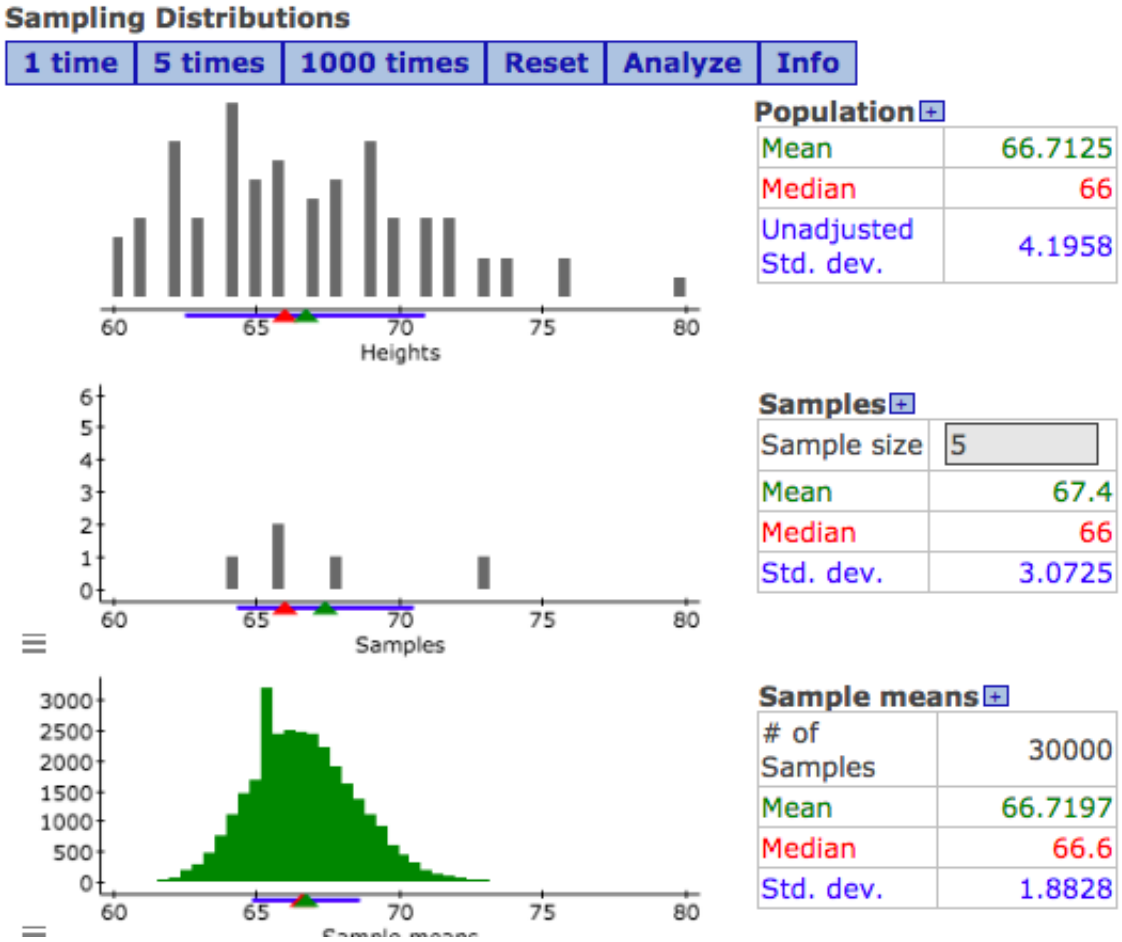
Sample means

# of Samples	1
Mean	66.4
Median	66.4
Std. dev.	0

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.

Histogram of sample mean based on 5 heights



Population

Mean	66.7125
Median	66
Unadjusted Std. dev.	4.1958

Samples

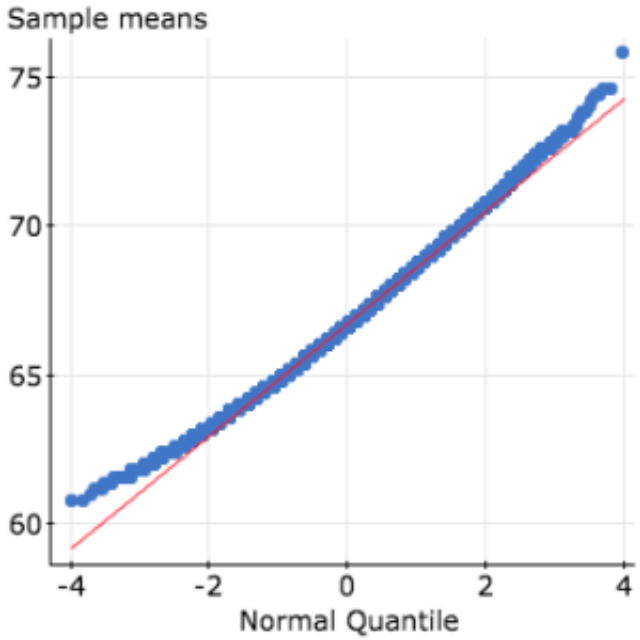
Sample size	5
Mean	67.4
Median	66
Std. dev.	3.0725

Sample means

# of Samples	30000
Mean	66.7197
Median	66.6
Std. dev.	1.8828

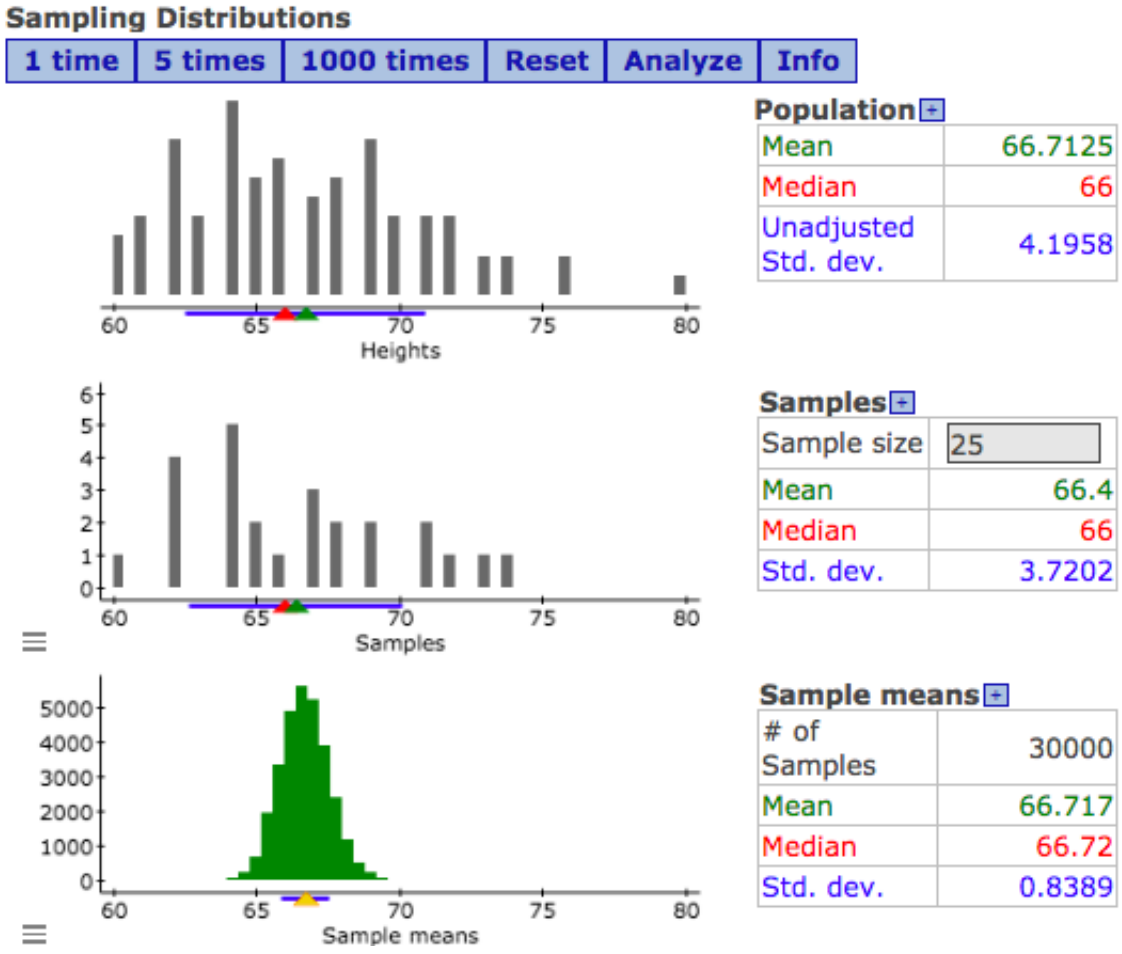
The green plot, is the histogram of all averages (ignore the 30K). Each average is an average of 5. A typical average is from take from this histogram. It is not quite normal, there is green spike. The standard error is 1.8 and sample size $n = 5$.

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



This is a QQplot of all those averages (not the original heights).

Histogram of sample mean based on 25 heights



Population

Mean	66.7125
Median	66
Unadjusted Std. dev.	4.1958

Samples

Sample size	25
Mean	66.4
Median	66
Std. dev.	3.7202

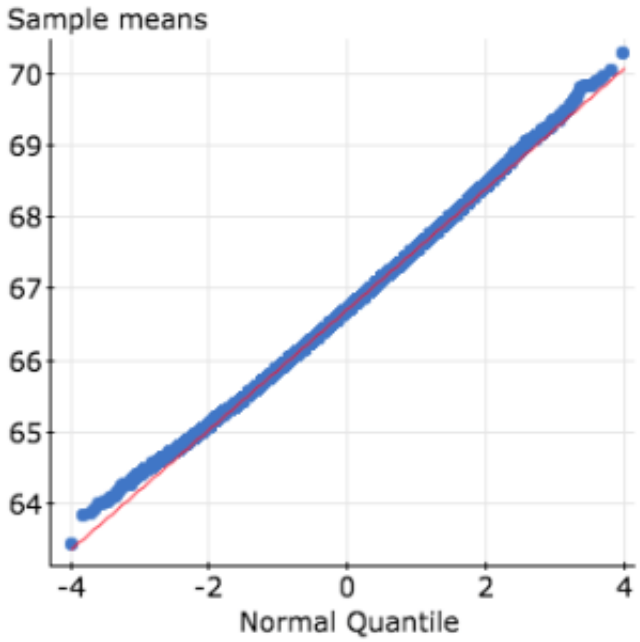
Sample means

# of Samples	30000
Mean	66.717
Median	66.72
Std. dev.	0.8389

The green plot is a histogram of the averages calculated using 25 students. Observe how "normal" it looks.

The standard error (=standard deviation of sample mean) is 0.84. The distribution is narrower than the case $n = 5$.

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.

Summary of thought experiment

- The green plots are histograms of the sample means (for example sizes 5 and 25).

- The green histograms are "centered" about the population mean, because the average is estimatng the population mean (average of averages is the population mean).

- The green histograms become more normally shaped as the sample size jumps from 5 to 25.

- The green histograms get narrower as the sample size increases from 5 to 25. This means the standard deviation/error (which is a measure of spread) has reduced.

Mean and standard error for heights

For height example, the variability (measured by standard deviation) from person to person was 4.2 inches. The variability decreased when considered the sample mean (average). As we increase the sample size the variability decreases in a predictable fashion:

	orig. population	sample mean (n= 5)	sample mean (n= 25)
mean	66.17	66.17	66.17
stand. dev.	4.2	1.88	0.83

- The average of the averages is aways the same as the population mean. Look at the green histograms, they are all centered about the mean.

- The variability decreases. We now show that it decreases in a predictable way.

- If the standard deviation, σ, in the original population is known (say $\sigma = 4.2$). 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.

- Appling formula to the height example

	orig. population	sample mean (n= 5)	sample mean (n= 25)
mean	66.17	66.17	66.17
stand. dev.	4.2	1.88	0.83
stand. err.	$4.2 = \frac{4.22}{\sqrt{1}}$	$1.88 = \frac{4.2}{\sqrt{5}}$	$0.83 = \frac{4.2}{\sqrt{25}}$

Thus the formula $\sigma/\sqrt{n}$ is correctly predicting the variability in the sample mean.

Summary: the standard error of the sample mean

- Suppose X_i is a random variable with mean μ and standard deviation σ. We observe a $X_1, \ldots, X_n$.

- Since $X_1, \ldots, X_n$ is random (changes from sample to sample) $\bar{X}$ is also random with the following properties:

- The mean of the sample mean $\bar{X}$ is μ.

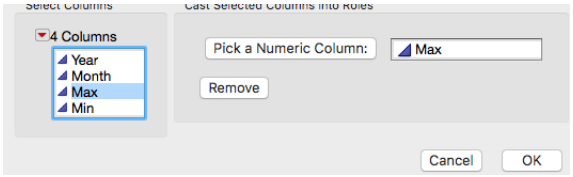
- If the sample size is n the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$.

- As the sample size n gets larger, σ stays the same; the variability of a population remains the same, BUT the standard error of the sample mean decreases. This corresponds to the green histogram getting narrower.

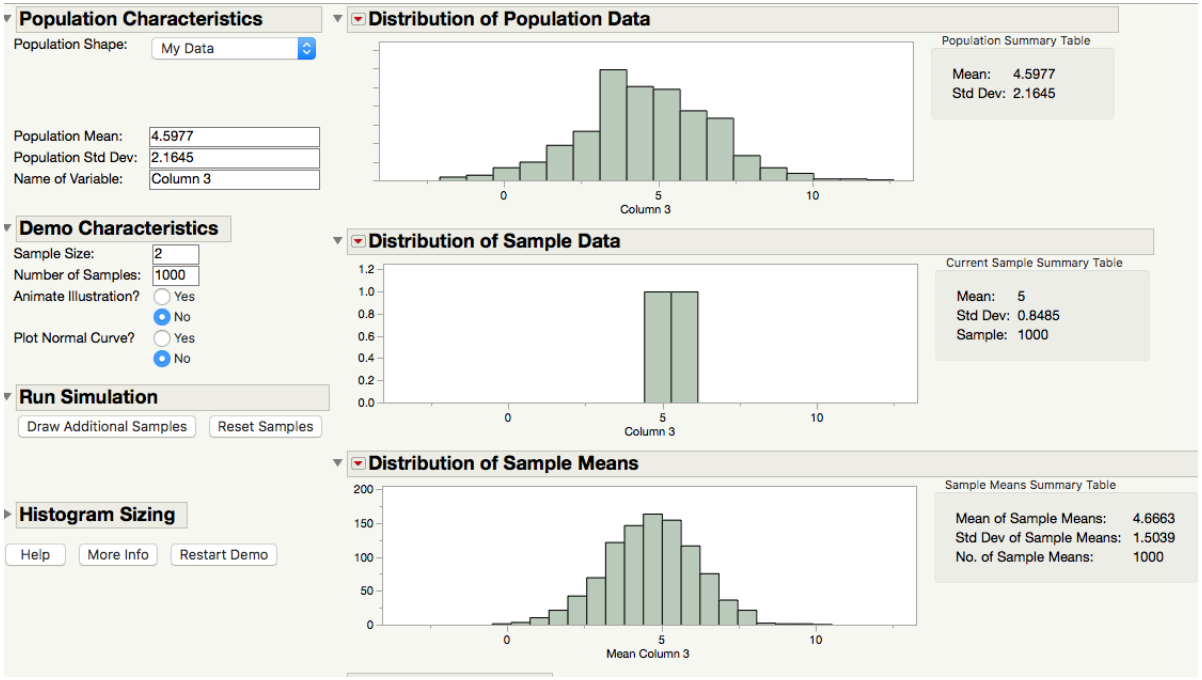
The sampling distribution of sample mean in JMP

- Load data ("population") on interest into JMP. Go to Help > Sample Data > Teaching Scripts > Interactive Modules > Sampling Distribution of Sample mean.

- For population Shape Choose My Data and "place" the data into right hand space.



- If you press the red triangle next to the Distribution of Sample means. It will make a spread sheet of the sample means (see next page right hand side).



	Sample	Sample Mean	Sample Stdv
1	1	-3.6	0.565685424949238
2	2	-16.65	14.9199530830362
3	3	-24.5	11.3137084989848
4	4	-4.4	5.51543289325507
5	5	-10.6	3.95979797464467
6	6	-14.7	8.20243866176395
7	7	-25.2	5.23259018078045
8	8	-25.1	6.08111831820431
9	9	-8.2	1.13137084989848
10	10	-13.55	10.9601551083915
11	11	-8.1	5.23259018078045
12	12	-18.3	15.6977705423414
13	13	-16.7	12.0208152801713
14	14	-18.05	15.4856385079854
15	15	-8.15	5.72756492761104
16	16	-15.35	8.41457069611992
17	17	-17.95	11.2429978208661
18	18	-21.2	11.1722871427475
19	19	-27.05	9.12167747730646

- In the above box choose the sample size (for this example I was considering the sample mean when $n = 2$ and the number of Samples (this does not matter, just choose something large)).

- Every sample of sample that is drawn from the population and average calculated is given in the spread sheet on the right.

- The top plot is the histogram of the "population". The middle plot is one sample that is drawn (in this case two numbers). The bottom plot is the histogram of the sample mean.

- Our focus is on the bottom plot.

Case 1: The distribution of sample mean of normally distributed r.v.

- If $X_1, X_2, \ldots X_n$ are independent normally distributed random variables with mean μ and standard deviation σ i.e. $X_i \sim N(\mu, \sigma)$.

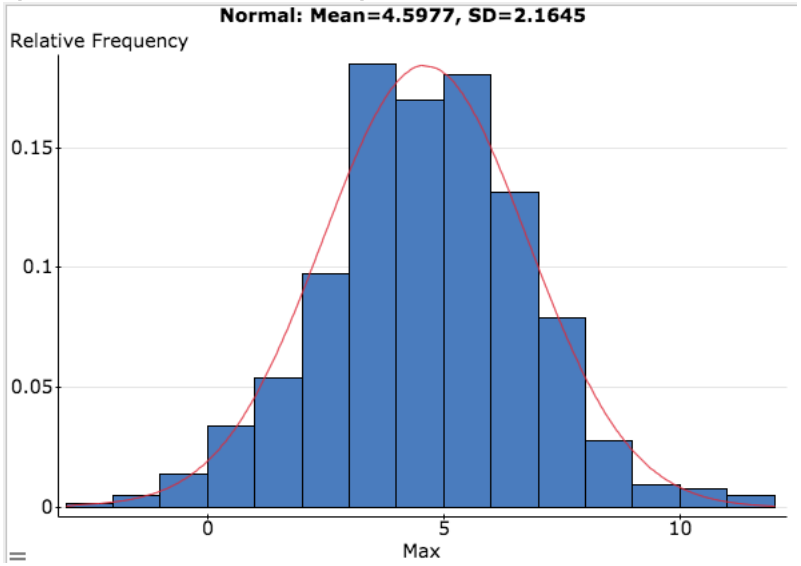
- Then the sample mean $\bar{X}$, will be normally distributed with the same mean as the orginal data and standard error $\sigma/\sqrt{n}$ i.e.

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

Since the original data is normal, then for all samples taken $1, 2, 3, \ldots$ the sample mean is normal.

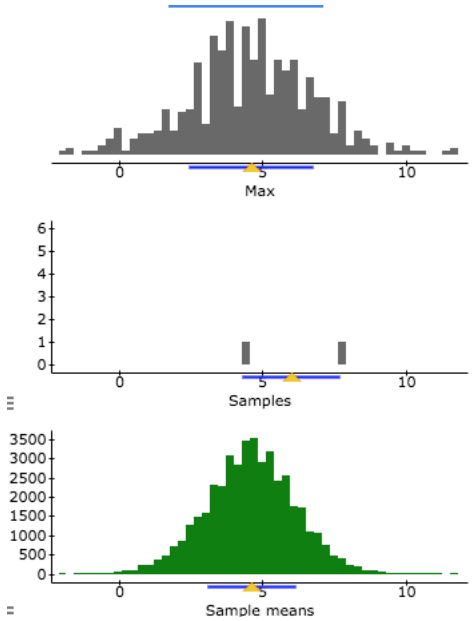
Example: Maximum temperatures in Antarctica

Recall the maximum temperatures in Antarctica is very close to normally (see Lecture 10) distributed with mean $\mu = 4.6$ and std. dev. $\sigma = 2.16$.



- Therefore any average regardless of the sample size will also be normally distributed with mean $\mu = 4.6$ and standard error $s.e. = 2.16/\sqrt{n}$.

- To see this we take samples of size 2 from the distribution and make a histogram of the average ($\bar{X} = (X_1 + X_2)/2$). It is the green plot given below. Observe the green plot looks pretty much normal, despite the sample size being only two.



Population

Mean	4.5977
Median	4.6
Unadjusted Std. dev.	2.1628

Samples

Sample size	2
Mean	6
Median	6
Std. dev.	1.7

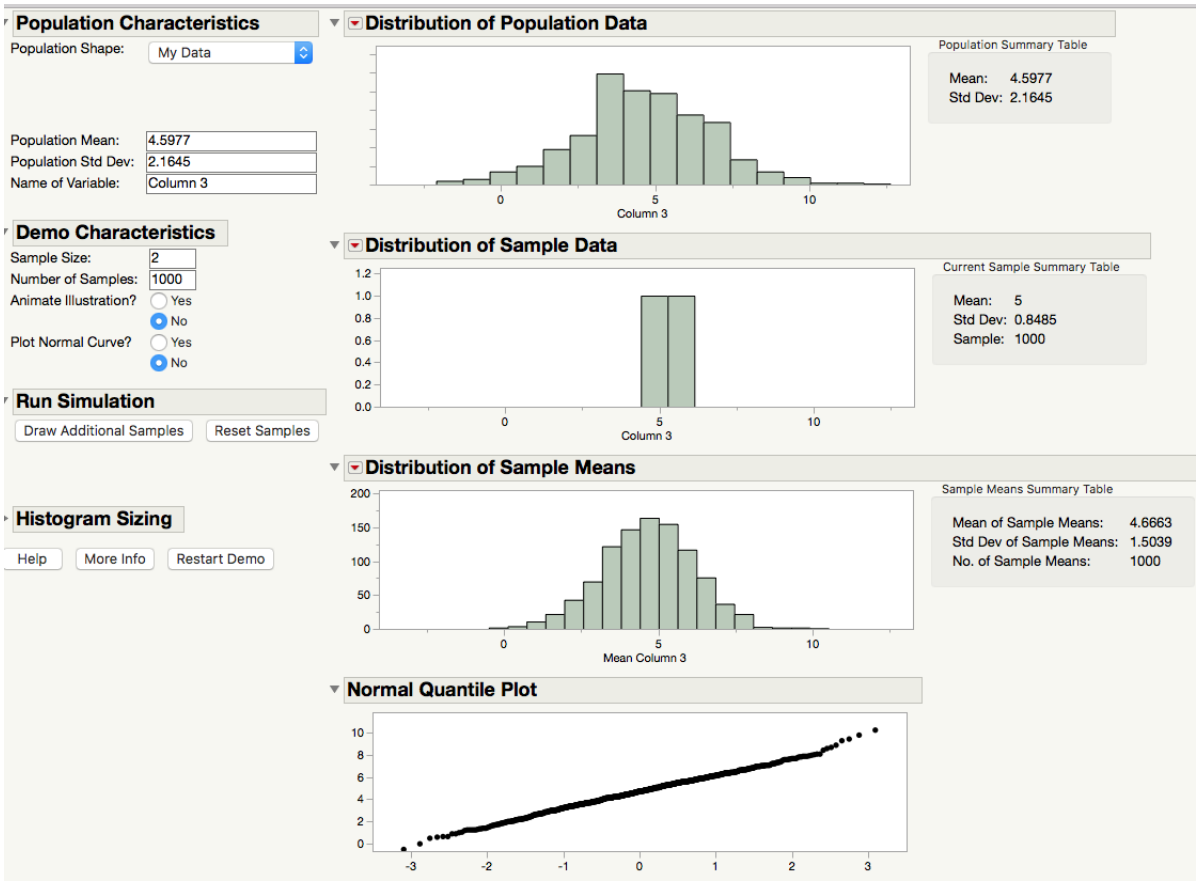
Sample means

# of Samples	44000
Mean	4.6092
Median	4.6
Std. dev.	1.5363

If a sample of size two is drawn with sample mean 6, the 95% confidence interval for the population mean is $[6 \pm 1.96 \times 1.53]$
If the green plot (distribution of sample mean) is normal, the probability of 95% assigned to this interval is correct.

The same thing in JMP

Follow the instructions a few slides back to get a similar plot:



Demonstration: Making multiple CIs in JMP

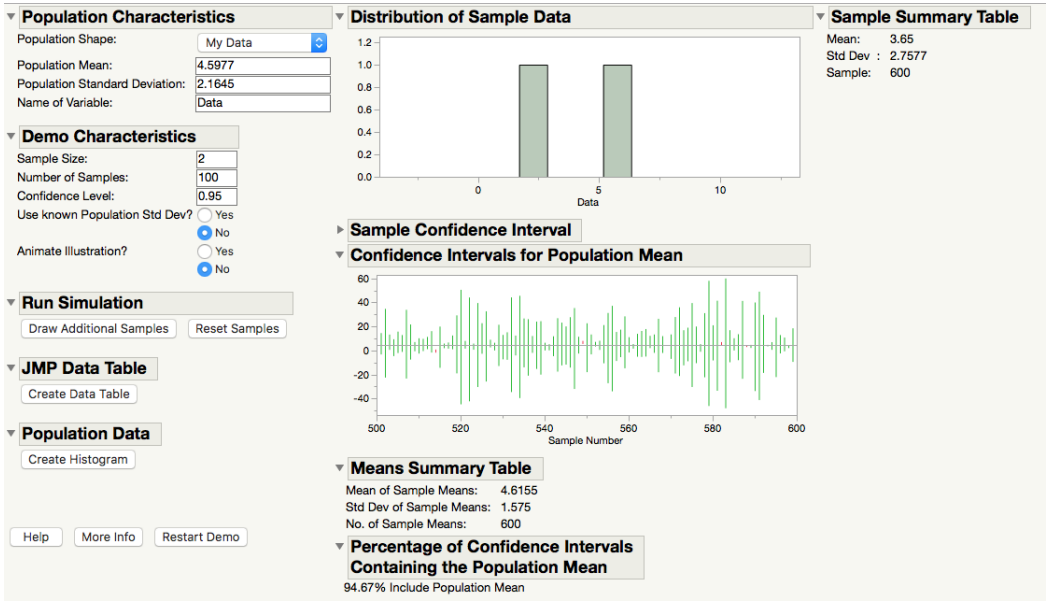
- Load data ("population") of interest into JMP. Go to Help $>$ Sample Data $>$ Teaching Scripts $>$ Interactive Modules $>$ Confidence interval for population mean. Choose My Data (like a few slides back).

- Press Draw Additional Samples. For each sample mean the computer constructs an 95% CI interval for the mean.

- Only issue: These confidence intervals are made using the t-distribution and estimated standard deviation (we return to this in Lecture 14). This is why the length of the CI varies from sample to sample.

- In practice this cannot be done, as we cannot observe the population. But the applet helps to understand, conceptionally, the information a confidence interval conveys.

- But we do observe that about 95% of the intervals contain the mean.

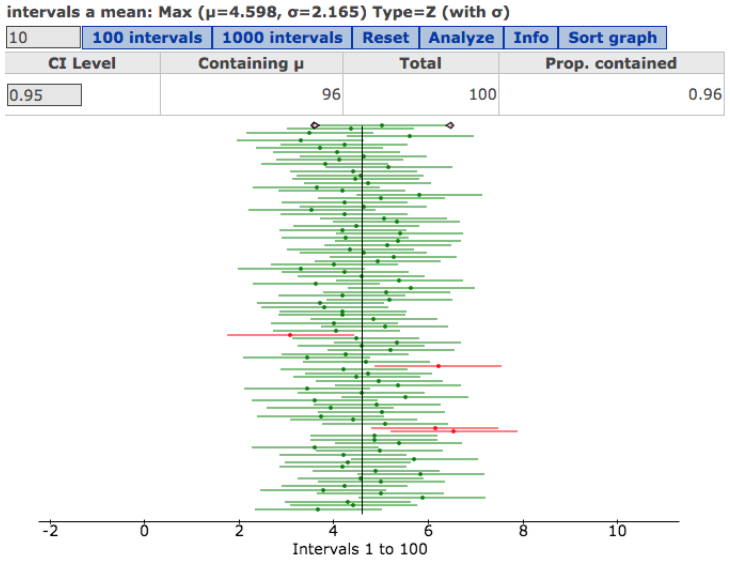


Each vertical line corresponds to one 95% CI for one sample mean. If it is green it intercepts with the population mean. If it is red it does not.

Using another software

- The CIs in this software are constructed using the normal distribution and the population standard deviation (thus for reasons we will learn in Lecture 14 each interval has the same length).

- In this demonstration, 4 out of the 100 do not contain the mean.



Conclusion: For normal data

- If the data is normally distributed, then the confidence interval for the population mean (based on the sample mean) will have the assigned level of confidence.

- Maximum temperature data This data set is normally distributed. Therefore the confidence interval for the population mean based on average of two temperatures really does have 95% confidence.

  But the interval is very wide since only two temperatures are used to construct the interval (large standard error).

Case II: The central limit theorem (nonnormal data)

The central limit theorem:

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

- 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 have already demonstrated (using the JMP and other demos) at the start of lecture 12 this result.

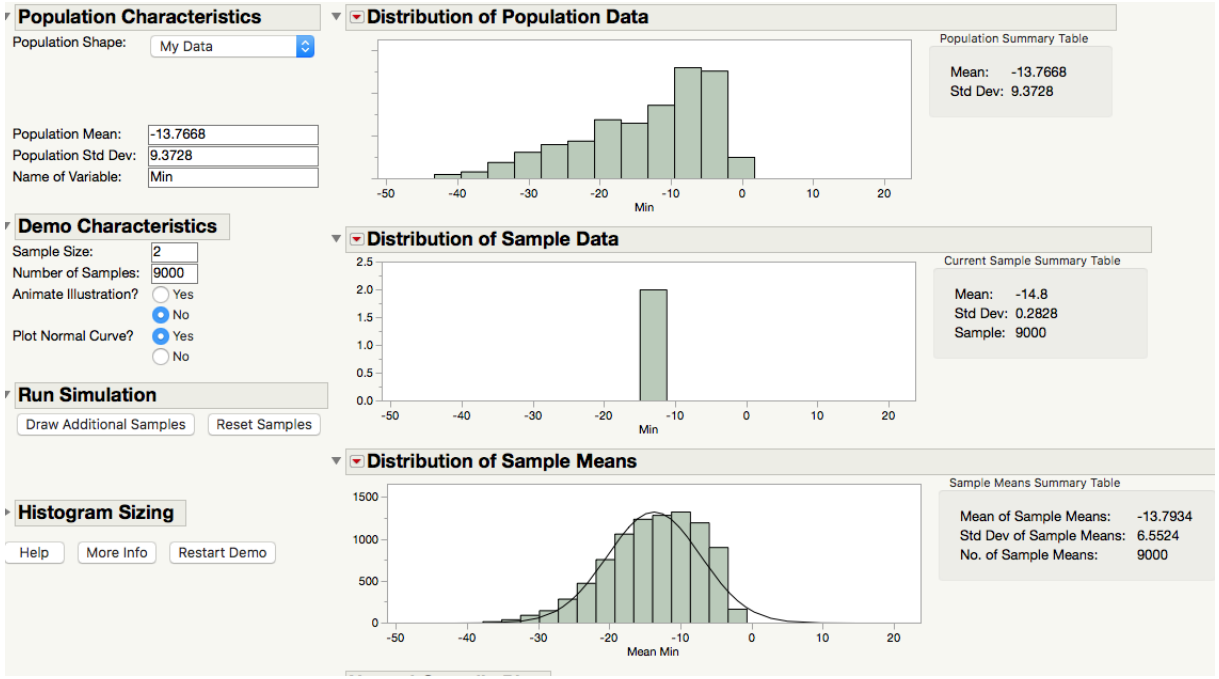
- Exceptions If there is dependence in the observations, then the standard error will *not* be $\sigma/\sqrt{n}$. If there is extreme dependence, the normal distribution will not hold.

- IMPORTANT The data DOES not become more normal as you increase the sample size. Ie. As you increase the sample size the QQplot DOES NOT magically become more normal looking. The distribution of the data is fixed. Eg. If human heights are bimodal they will bimodal even if the sample size is large. Increasing the sample size does not make the original data more normal.

- It is distribution of the sample mean that becomes normal

Example: Min temperatures in Antarctica (n=2)

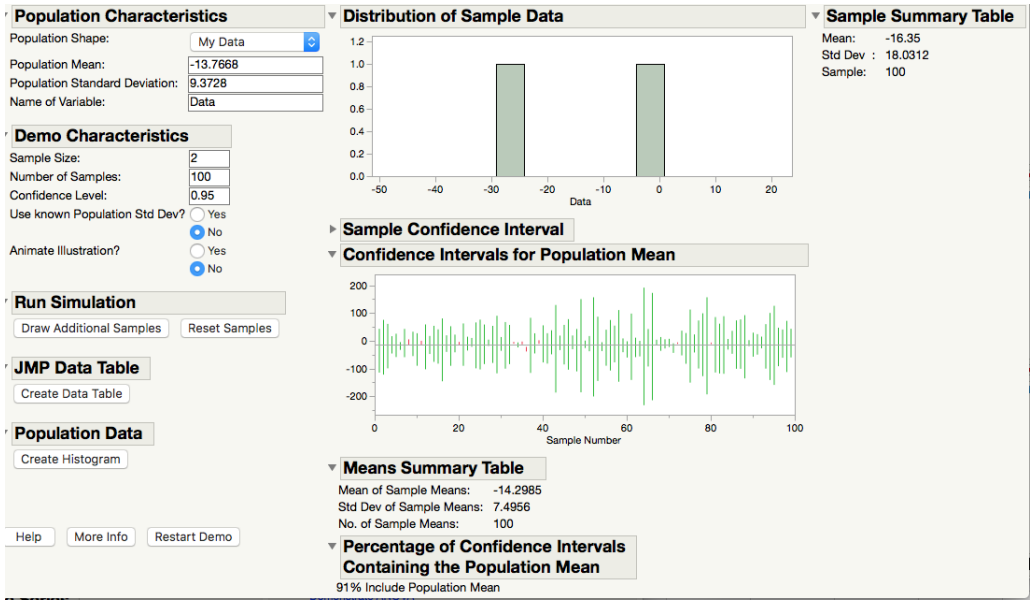
- Recall the minimum temperatures in Antarctica are left skewed.



Above, we give a plot of this data and the plot of the histogram of the sample mean based on two observations (bottom plot).

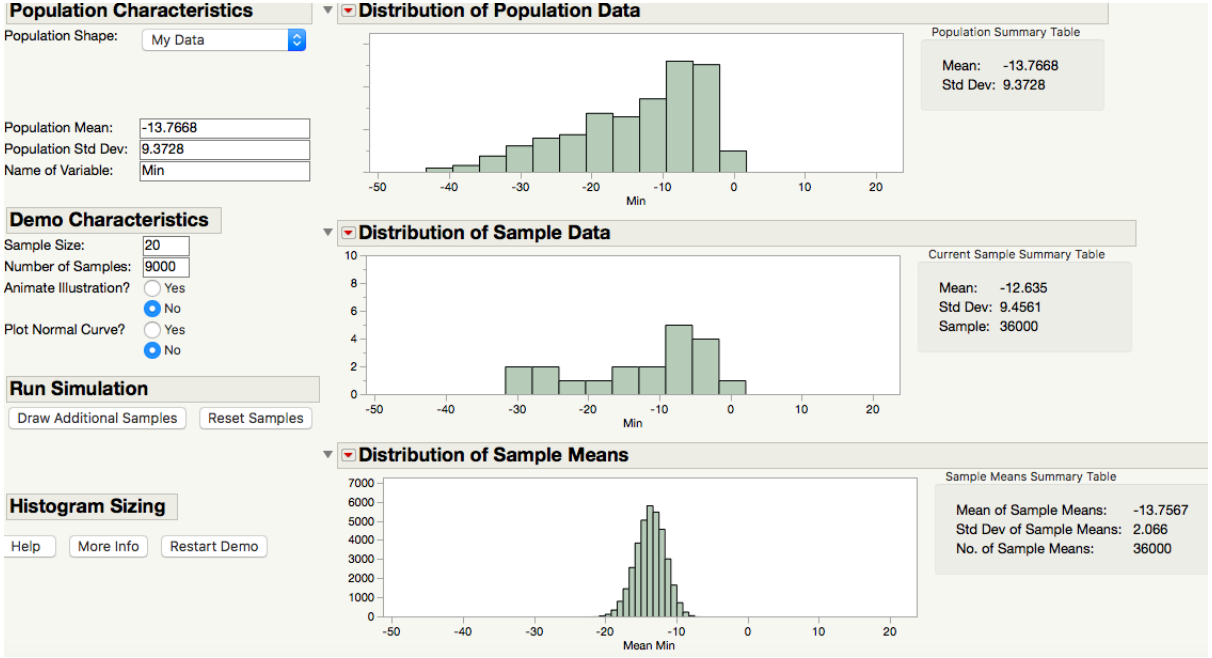
Reliability of confidence interval

- The bottom plot on the previous slide (histogram of sample mean) is still left skewed (but less so than the original population).



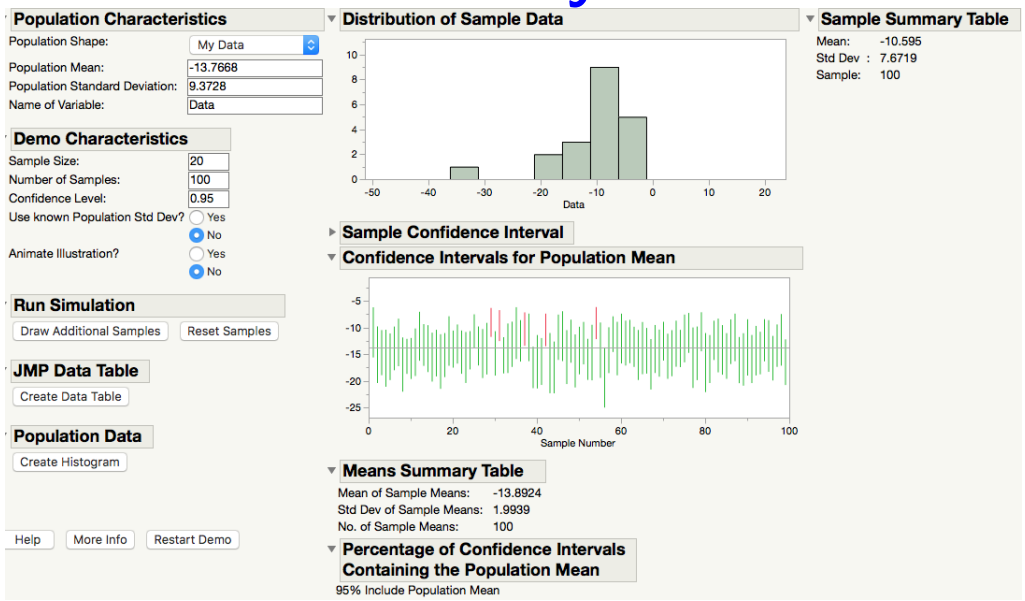
From the App, we see that a 95% confidence interval for the mean (based on the assumption of normality of the sample mean) only contains on average 91 % time the mean. The probability assigned to this interval is incorrect and unreliable.

Example: Min. temperatures in Antarctica (n=20)



The bottom plot gives the histogram of the sample mean based on 20 observations. It is closer to normal than a sample mean based on two.

Reliability of confidence interval



- From the App we see that a 95% confidence interval for the mean (based on the assumption of normality of the sample mean) contains on average 95 % time the mean. The probability assigned to a 95% confidence for this data set base based on a sample size $n = 20$ appears to be about right.

Conclusion: For non-normal data

- If the data is not normally distributed, then the confidence interval for the population mean (based on the sample mean) may not have the assigned level of confidence (when the sample size is small).

  We need to be cautious on interpreting CI for small sample sizes.

- Minimum temperature data This data set is not normally distributed. Therefore the confidence interval for the population mean based on the average of two temperatures does not have 95% confidence.

  However, if we are fortunate to have a larger sample size (say 20). The CI for the population mean based on the average of 20 obervations does have the assigned level of confidence.

Some warnings: check the assumptions

- The minimum and maximum temperature data nicely illustrate the central limit theorem and the standard errors of the sample mean.

- However, we need to be cautious.

(1) The standard error $\sigma/\sqrt{n}$ of the sample mean only holds if the observations are independent of each other (and are drawn from the same population).

(2) It only makes sense calculating the sample mean if all the observations have the same population mean.

- Application to temperature data The temperature data is recorded monthly. It is likely the temperatures over the months are **dependent**.

- Temperature data is likely to have a **trend** (the mean which changes over time).

- It is extremely important to check if the assumptions are satisfied before conducting a statistical analysis. And to understand how the results may change if an assumptions is violated.

How large is large?

- How large, is large, is a difficult question, and varies from data to data. The 'rule of thumb' given in many textbooks is that the sample size should be about $n = 30$ for the CLT of the sample mean to hold. However, this should only be take as a guideline.

  In general:

  - If the data is close to normal - then for a very small sample size, the sample mean is close to normal (look at the maximum temperature example). 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). A confidence interval for the mean based on relatively small sample sizes will not be reliable.
  - If the data is highly skewed then one needs a very large sample size for the CLT to hold true.

- If the data takes just a few numerical discrete values, then a fairly large sample size is required for the CLT to hold.

- What does this 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.

Applying the results to analysing data

- In the next few examples we apply what we have learnt to constructing confidence intervals for the mean.

- We start by checking for normality of the sample mean.

- We need to be cautious about using this method, as it is based on subsampling (which requires many hidden assumptions).

  The best advise when using this method is to be cautious about interpreting the results.

- In the "thought" experiments above, we were drawing from the "true population" to understand what the distribution of the sample mean looked like. We now attempt to "estimate" by resampling from the actual observed data and evaluating the sample mean.

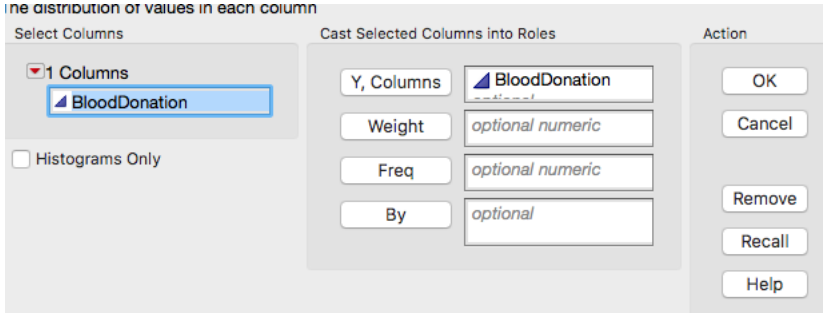
Checking normality of the sample mean using JMP

- Always make a QQplot of the data (see Lecture 10), to see how much deviates from normality.

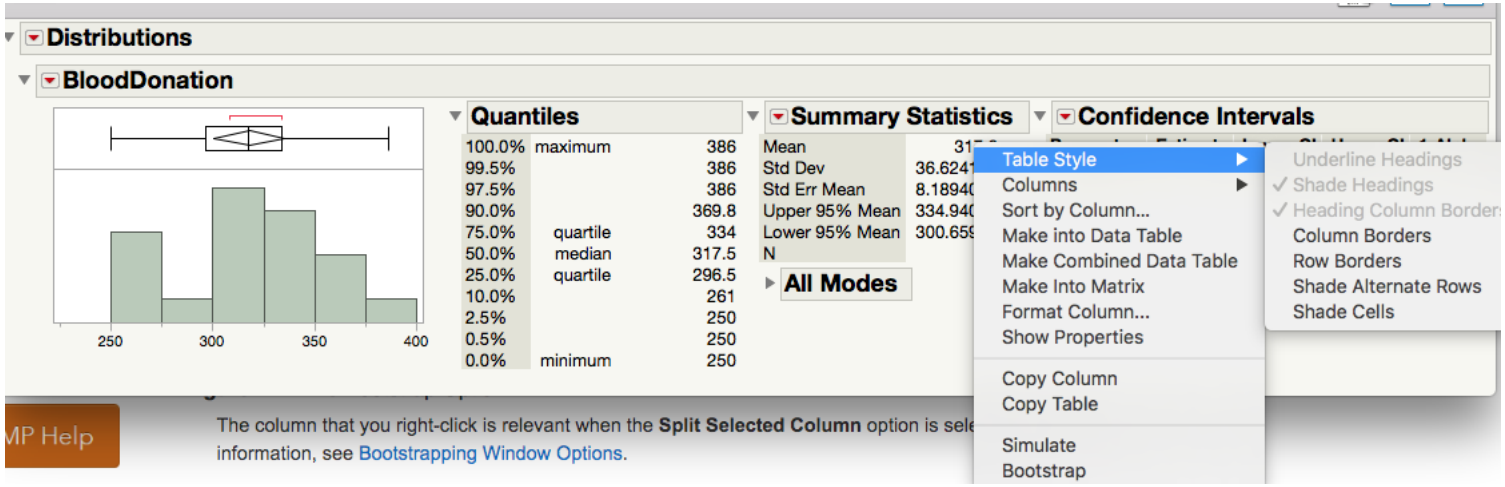
- However, we can also make a rough guess on the distribution of the sample mean by following the steps below.

- This approximation of the sample mean only makse sense if the sample size is sufficiently large, say over 20.

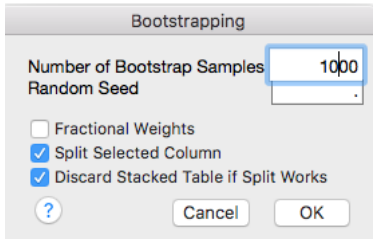
- Load data the into JMP. Go to Analyze > Distribution



- **Right click** on the number next to **Mean** and select Bootstrap.



- Choose the number of replications required. The more your choose the longer it takes (2500) is the default. I chose 1000. Press OK.



- When it is done you should get a huge table like this

	Y	BootID·	Lower 95% Mean	Mean	N	Std Dev	Std Err Mean	Upper 95% Mean
1	BloodDonation	0	300.65938326	317.8	20	36.624122048	8.1894026516	334.94061674
2	BloodDonation	1	311.50959178	327.2	20	33.525481275	7.496525511	342.89040822
3	BloodDonation	2	303.93466938	315.1	20	23.856809687	5.3345448185	326.26533062
4	BloodDonation	3	325.7707589	337.95	20	26.023218378	5.8189685286	350.1292411
5	BloodDonation	4	304.44148593	316.9	20	26.619937127	5.9523988972	329.35851407
6	BloodDonation	5	305.74651648	321.05	20	32.698744544	7.3116615579	336.35348352
7	BloodDonation	6	301.01207886	318.45	20	37.259368303	8.3314480324	335.88792114
8	BloodDonation	7	292.96407594	309.8	20	35.973089357	8.0438273163	326.63592406
9	BloodDonation	8	306.10509418	326.5	20	43.577517139	9.7442290613	346.89490582
10	BloodDonation	9	306.27201502	321.5	20	32.537427841	7.2755900466	336.72798498
11	BloodDonation	10	302.44342049	317.6	20	32.38485676	7.2414741157	332.75657951
12	BloodDonation	11	299.60029038	315.75	20	34.506864306	7.7159694278	331.89970962
13	BloodDonation	12	301.72537534	318.15	20	35.094271537	7.8473176778	334.57462466
14	BloodDonation	13	315.8507865	330.05	20	30.339265855	6.784066084	344.2492135
15	BloodDonation	14	311.46600751	322.25	20	23.042009689	5.1523500004	333.03399249
16	BloodDonation	15	299.97267161	318.1	20	38.732415365	8.6608313689	336.22732839
17	BloodDonation	16	307.00415307	323.2	20	34.605445267	7.7380128008	339.39584693
18	BloodDonation	17	288.93392791	308.15	20	41.058719184	9.1810087165	327.36607209
19	BloodDonation	18	301.79084755	315.9	20	30.146833647	6.7410369341	330.00915245
20	BloodDonation	19	311.01367226	327.7	20	35.653448936	7.9723535454	344.38632774
21	BloodDonation	20	306.02650001	319.9	20	29.643318232	6.6284474645	333.77349999
22	BloodDonation	21	297.76848926	314.95	20	36.711499704	8.2089408894	332.13151074
23	BloodDonation	22	301.74604671	322	20	43.276345794	9.6768851013	342.25395329

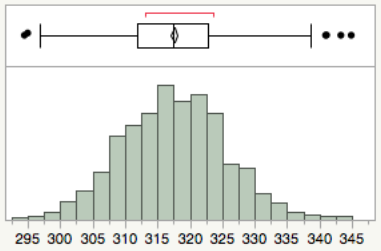
- There are 1000 rows. Each row corresponds to the summary statistics of one "bootstrap" sample (the name given for the resampling). For the blood donor example, this means 20 numbers (with replacement) are randomly drawn from the original sample (of size 20). For each "bootstrap sample" the sample mean is calculated. This is in the '**Mean** column.

- The sample means (in the mean column) are variables and behave, in some sense, like the sample means calculated using samples from the population.

- Based on the above argument the histogram of these "bootstrap" sample means should be "close" to the true histogram of the sample means.

- Make a histogram of **Mean** (Go to Analyze > Distribution)

Mean



Quantiles

100.0%	maximum	344.9
99.5%		338.546
97.5%		333.14875
90.0%		327.895
75.0%	quartile	322.85
50.0%	median	317.425
25.0%	quartile	311.9125
10.0%		307.555
2.5%		302.255
0.5%		299.20025
0.0%	minimum	294.5

Summary Statistics

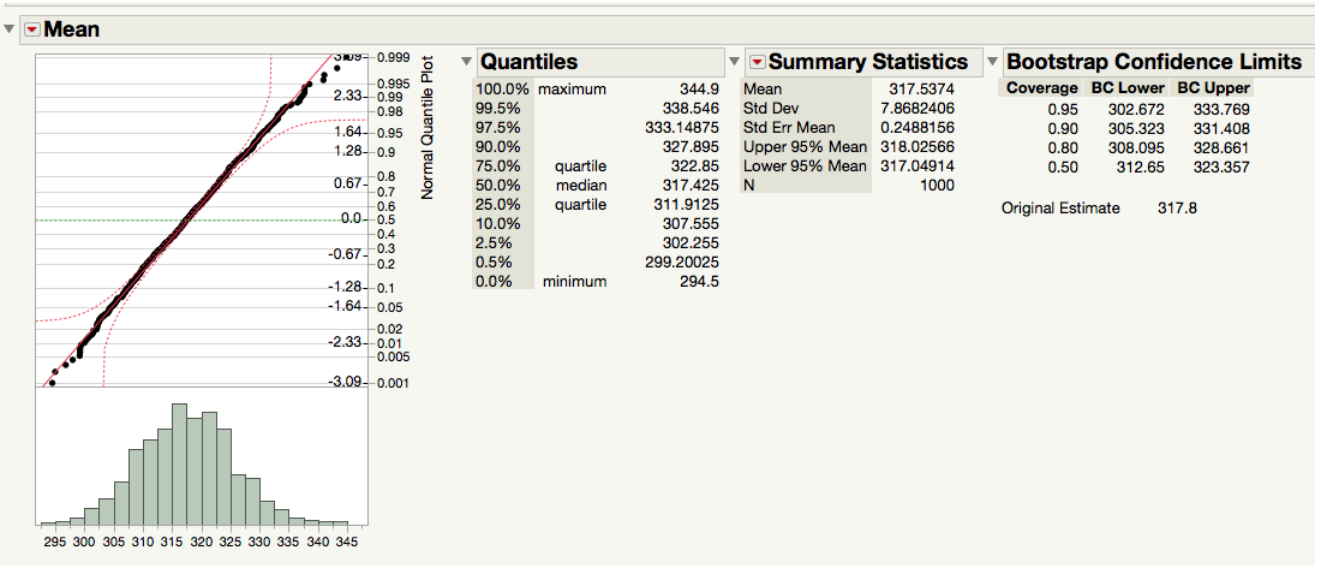
Mean	317.5374
Std Dev	7.8682406
Std Err Mean	0.2488156
Upper 95% Mean	318.02566
Lower 95% Mean	317.04914
N	1000

Bootstrap Confidence Limits

Coverage	BC Lower	BC Upper
0.95	302.672	333.769
0.90	305.323	331.408
0.80	308.095	328.661
0.50	312.65	323.357

Original Estimate 317.8

- You can also make a QQplot of the sample mean (by following the instructions in lecture 10)



- Observe from the QQplot of the bootstrap mean, that it looks close to normal. This tells us that the distribution of the sample mean should be close to normal. So using the normal distribution to construct a confidence interval will give a reliable level of confidence.

Baseball Salaries

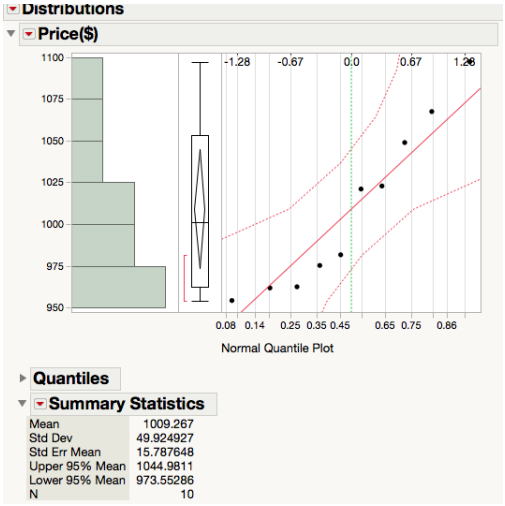
- Using the sampling method described above check if the sample mean of the Base Ball salaries are normal.

- The data can be found here:

  `http://www.stat.tamu.edu/~suhasini/teaching651/BaseBall.csv`

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 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.

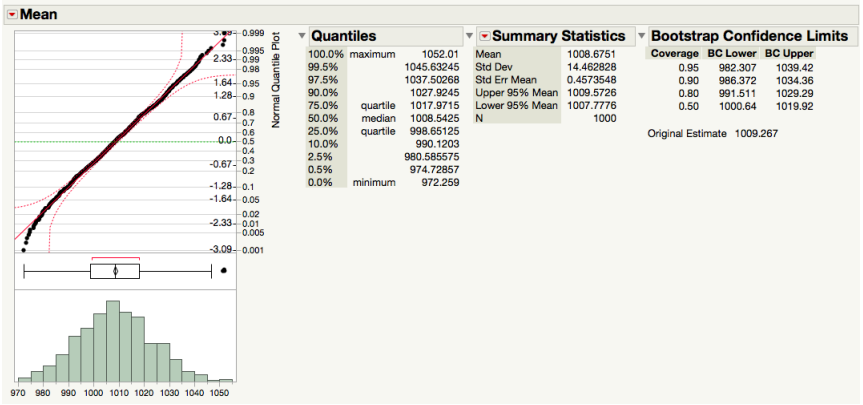


Solution1: Checking for normality

- The sample size is quite small, but the QQplot of data does not deviate hugely from normality.

- We check to see if the sample mean is close to normal using the bootstrap method described above. Beware, the sample size is only $n = 10$.

  Observe that the green plot is quite close to normal, so we can that we 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 μ 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 without thought. However, it is worth noting that 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 probabilites

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.

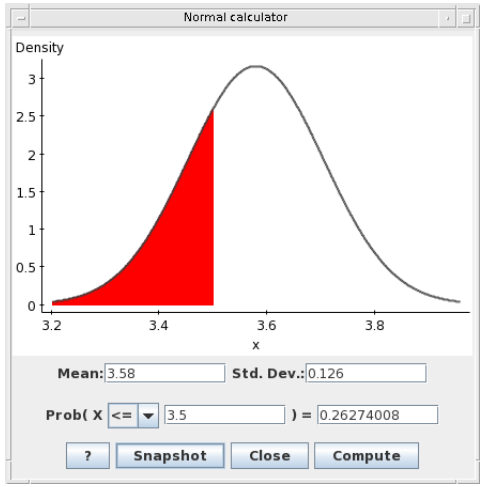
(c) What happens to the chance of wrong diagnoses when we increase the sample size?

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 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

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

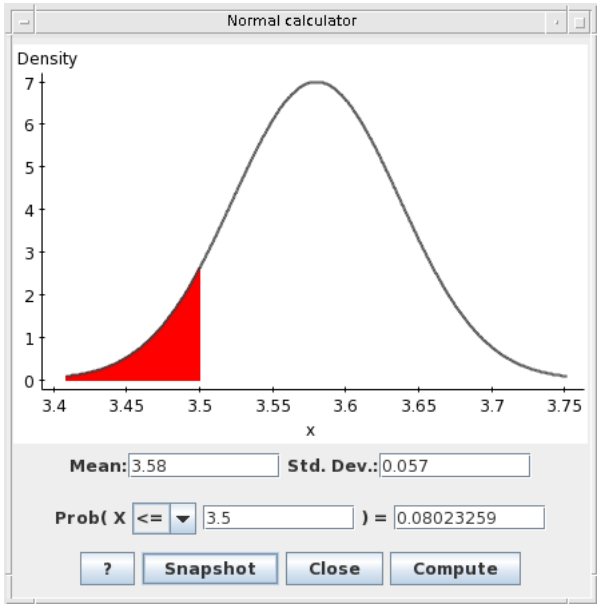
She is wrongly diagnosed if her sample mean is below 3.5. This means we need to evaluate $P(\bar{X}_{10} < 3.5)$. The z-transform $z = \frac{3.5-3.58}{0.126} = -0.63$. From the tables we see this corresponds to 26.3%. In other words there is a 26.3% chance of a wrong diagnoses based on 10 samples.



(b) The sample size is 49. The distribution of the sample mean based on 49 is

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

The z-transform is $z = \frac{3.5-3.58}{0.057} = -1.4$. 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. People with normal potassium levels (but close to 3.5) 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 complete interval lies fully 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

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 ≥ 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 ≥ 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 ≥ 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 μ.

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 σ.

  Hence in our calculations we will use $s = 1.1$ in place of the true standard deviation σ.

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

- The 99% CI is

$$\left[3.2 - 2.57\frac{1.1}{\sqrt{150}}, 3.2 + 2.57\frac{1.1}{\sqrt{150}}\right]$$

.

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

Example 5: Calf weights at birth

We analyse the birth weights of calves. It can be found at `http://www.stat.tamu.edu/~suhasini/teaching651/cow_birth_weights.csv`.

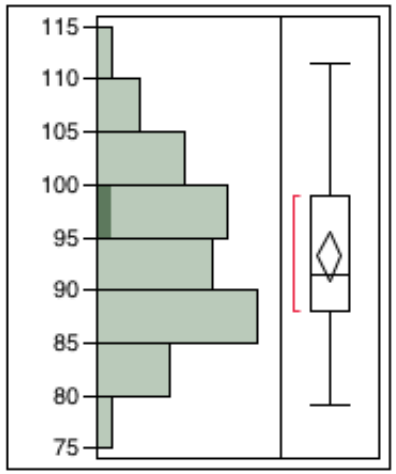
(a) Construct a 90% CI for the population mean for birth weight of newborn calves.

(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

Distributions

Wt 0



Summary Statistics

Mean	93.215909
Std Dev	7.7784719
Std Err Mean	1.1726488
Upper 95% Mean	95.580781
Lower 95% Mean	90.851037
N	44

- In JMP Analyze $>$ Distribution $>$ Highlight variable of interest (same as before). A window should pop-up with histogram and summary statistics.

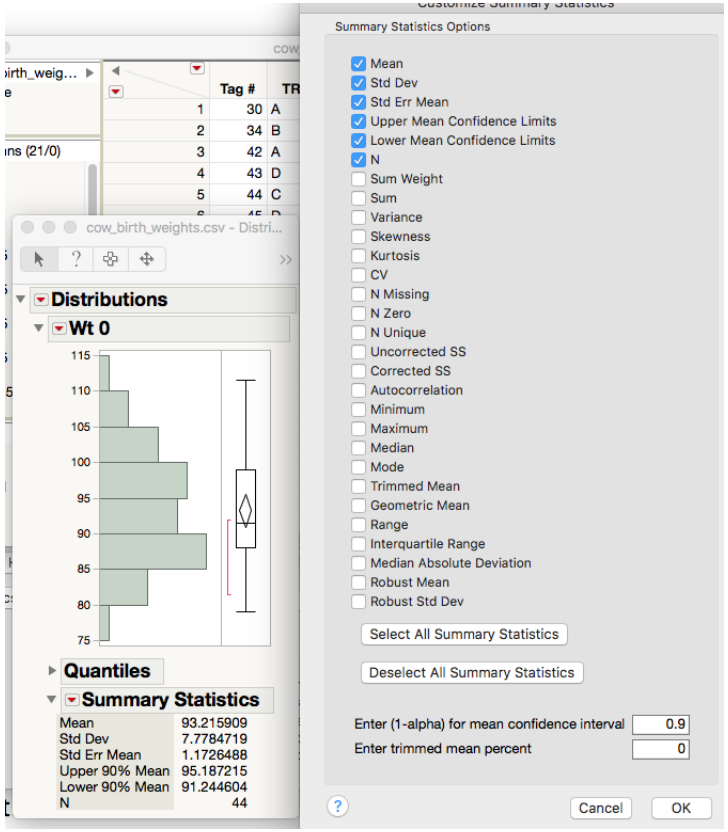
- From JMP we can see that the number of observation is 44, the sample mean is 93.21 pounds), the standard deviation is 7.44 and the standard error of sample mean is is $7.44/\sqrt{44} = 1.07$.

- 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.

Differences between JMP output and calculation

- 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.

- How the t-distribution comes into play is explained in Lecture 14.