Data Analysis and Statistical Methods
Statistics 651

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

Lecture 14 (MWF) The t-distribution

Suhasini Subba Rao

Terminology: Standard deviations and errors

- The standard deviation is a measure of variation/spread of a variable (in the population). This is typically denoted as σ. See Lecture 4.

- The standard error is a measure of variation/spread of the sample mean. The standard error of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is

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

  See Lecture 12.

- Usually, σ is unknown. To get some idea of the spread, we estimate it

from the sample $\{X_i\}_{i=1}^n$ using the formula

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

We call s the sample standard deviation. It is an estimator of the standard deviation σ. Usually $s \neq \sigma$. Often $s < \sigma$ (especially, when the sample size is not large).

- Since σ is usually unknown, the standard error $\sigma/\sqrt{n}$ is usually unknown. Instead we estimate it using the sample standard error is

$$\frac{s}{\sqrt{n}}.$$

How estimating the standard deviation effects our results

- So far we have assumed that the standard deviation σ is known.
- This is sometimes a plausible assumption. There are situations when one may know the standard deviation but not the population mean.
- However in general we will not know σ. σ is unknown and has to be estimated from the data.

Motivation

- We take a SRS of 5 students and record their heights **61, 63, 65, 66, 72**. The sample mean/average is 65.4.

- Our objective is to construct a 95% confidence interval for the population mean of students. Putting numbers into the formula gives

$$\left[65.4 - 1.96 \times \frac{\sigma}{\sqrt{5}}, 65.4 + 1.96 \times \frac{\sigma}{\sqrt{5}}\right]$$

- But the population standard deviation, σ, is unknown.

- We can estimate it from the data **61, 63, 65, 66, 72**, using the sample standard deviation which is

$$s = \sqrt{\frac{1}{4-1}(61-65.4)^2 + (63-65.4)^2 + (65-65.4)^2 + (66-65.4)^2 + (72-65.4)^2} = 4.16$$

and put 4.16 into the above confidence interval.

- What we want to know is whether it changes anything. In fact it turns out that the population standard deviation is $\sigma = 4.3$...what does this tells us about the interval?

- **Constructing confidence intervals** It seems reasonable to replace σ with s when evaluations a z-transform or a 95% CI:

$$\text{z-transform} \qquad \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \to \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \quad \text{t-transform}$$

$$95\%\ \text{CI} \qquad \left[\bar{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}\right] \to \left[\bar{X} \pm 1.96 \times \frac{s}{\sqrt{n}}\right] \text{?? CI}$$

- But have we lost anything in replacing σ with s?

The effect of estimating the standard deviation

- In the discussion below we are assuming that the observations $\{X_i\}$ are independent random variables from a normal distribution with mean μ and standard deviation σ.

  What we discuss below has nothing to do with correcting for normality of the observations. It is about estimation of the population standard deviation σ.

- The sample standard deviation s is random it varies from sample to sample.

- If the sample size is relatively small it can often underestimate the true standard deviation. This can cause substantial problems.

- The z-transform is the number of standard errors that can "fit between" the mean and the sample mean. If the standard deviation has been underestimated, then the z-transform will be larger than what it is suppose to be

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \to \left.\begin{array}{l} \dfrac{\bar{X} - \mu}{s} \\ \underbrace{\dfrac{}{\sqrt{n}}}_{\text{smaller}} \end{array}\right\} \text{larger.}$$

- There is a change in terminology (when we replace the population standard error with the sample standard error) we call it

$$\text{t-transform } = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

- Equivalently, if we use the estimated standard deviation to construct the confidence interval, an underestimated standard deviation will result in a confidence interval that is too narrow. Consider the 95% confidence interval

$$\left[\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right] \rightarrow \left[\bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{s}{\sqrt{n}}\right].$$

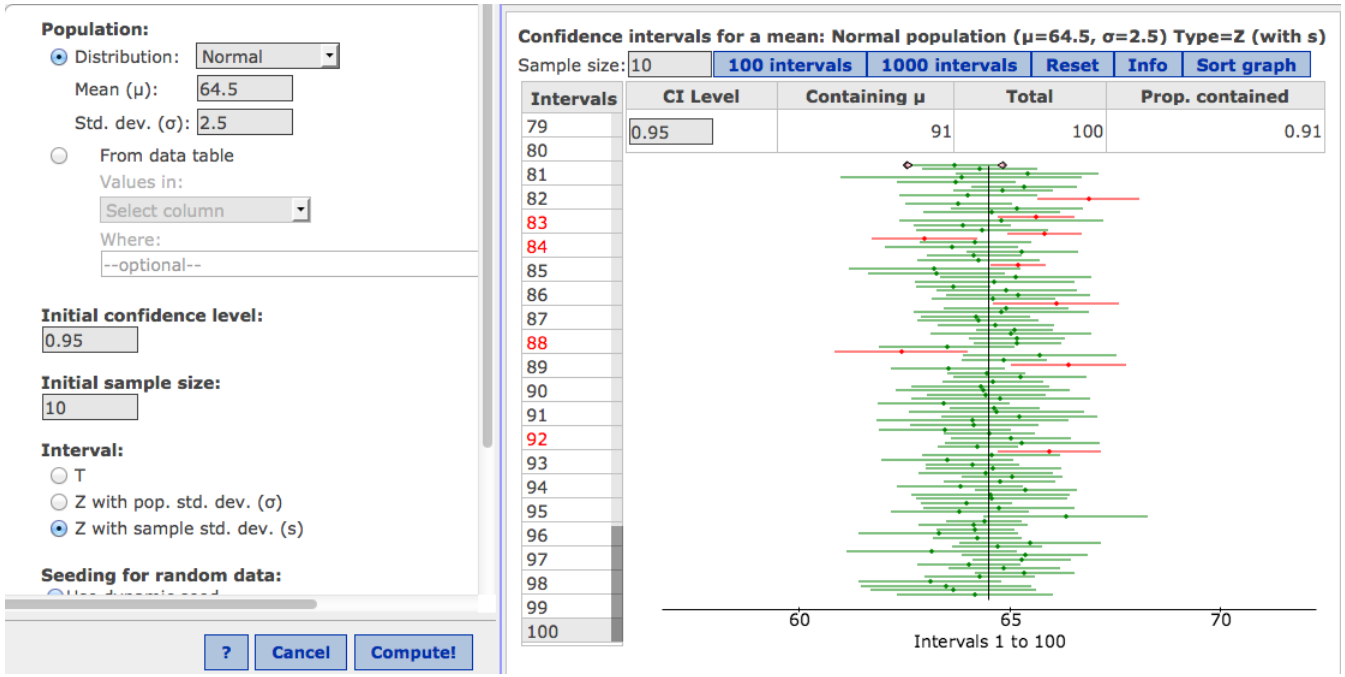
  If s is smaller than σ, then the interval will be too narrow for it to be a 95% confidence interval.

- We need to correct for the fact that s tends to underestimate the population standard deviation σ.

- Indeed, it is very simple to make the correction. All we need to do is change the distribution from a normal distribution to a t-distribution.

An illustration: Confidence intervals

We draw a sample of size 10, from a normal distribution, and estimate **both** the sample mean and standard deviation and construct a 95% CI using $z = 1.96$. Observe only 91 of the 100 confidence intervals contain the mean. We have less confidence in this interval than the stated 95% level!



The t-distribution

- The transform (which we formally called the z-transform)

$$t = \frac{\bar{X} - \mu}{\mathbf{s}/\sqrt{n}} \sim t(n-1),$$

  has a t-distribution with $(n-1)$-degrees of freedom where n is the number of observations used to estimate σ and μ.

- The term degrees of freedom is a word commonly used in statistics. It refers to the "effective" sample used to estimate the population standard deviation. The $(n-1)$- comes into play because once the sample mean is estimated the effective sample size is $(n-1)$ and not n.

- The distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ depends on the sample size.

- We call $t(n-1)$ the Student t-distribution with $(n-1)$-degrees of freedom. We use the name Student, in honor of William Gosset (he wrote all his papers under the pseudonym Student).

How does this change things?

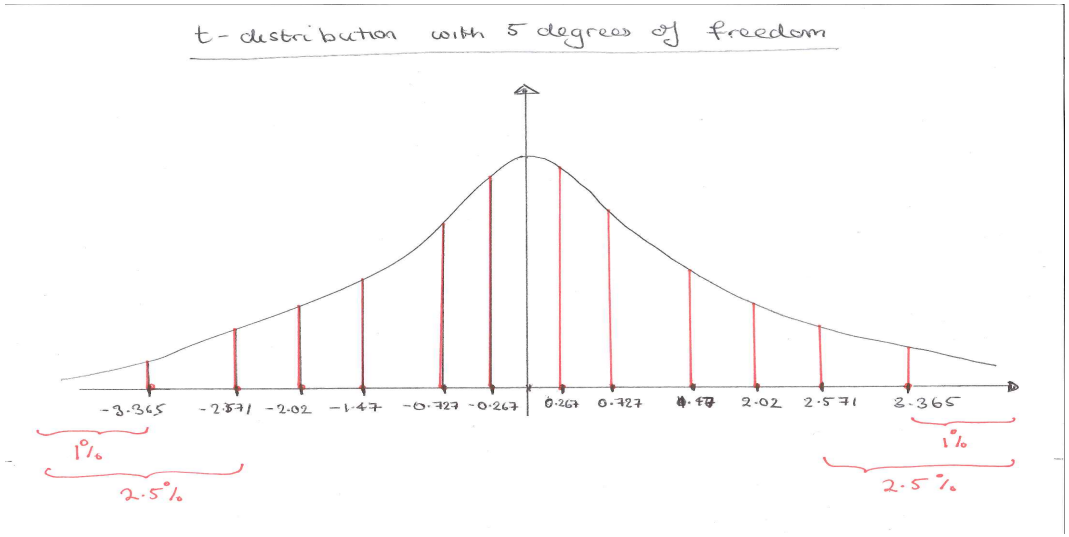
- We do almost everything as we did before, but when we estimate the standard deviation we use the t-distribution instead of the standard normal.

- The t-values are larger than the z-values to compensate for the underestimation of standard deviation.

- Rather than use the normal tables we use the t-tables which are very easy to use and can be found on my website.

- Most statistical software (such as JMP) does this automatically.

Reading t-tables (Table 2)



Percentage points of Student's t distribution

$t_{\alpha,\nu}$

df/α =	.40	.25	.10	.05	.025	.01	.005	.001	.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869

Confidence intervals using the t-distribution

- When the standard deviation σ is known. The $(1-\alpha)100\%$ CI is

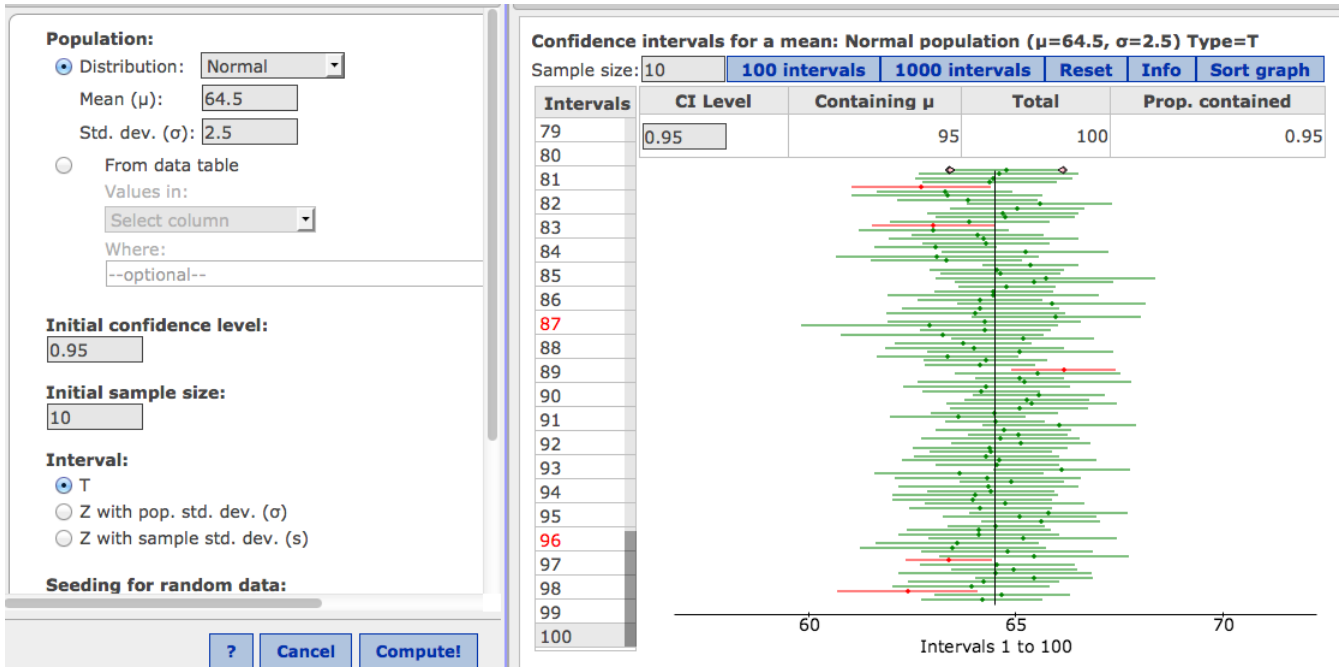
$$\left[\bar{X} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}, \bar{X} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}\right].$$

- When the standard deviation σ is unknown, we estimate it from the data $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and use the CI

$$\left[\bar{X} - |t_{\alpha/2}(n-1)|\frac{s}{\sqrt{n}}, \bar{X} + |t_{\alpha/2}(n-1)|\frac{s}{\sqrt{n}}\right].$$

An illustration: Confidence intervals

We draw a sample of size 10, from a normal distribution, and estimate **both** the sample mean and standard deviation and construct a 95% CI using $t_{0.025}(9) = 2.262$ (compare with $z = 1.96$). By using the t-distribution we have 95% confidence the interval contains the mean.



Example 1: Red Wine and polyphenols

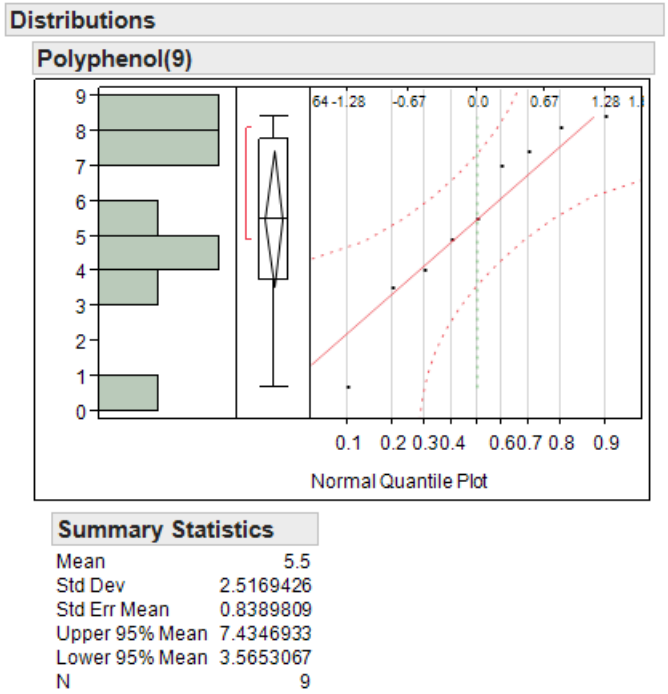
It has been suggested that drinking red wine in moderation may protect against heart attacks. This is because red wine contains polyphenols which act on blood cholesterol.

To see if moderate wine consumption does increase polyphenols, a group of nine random selected males were assigned to drink half a bottle of red wine daily for two weeks. The percentage change in their blood levels are

$$0.7, 3.5, 4, 4.9, 5.5, 7, 7.4, 8.1, 8.4$$

Here's the data: `http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt`. The sample mean is $\bar{x} = 5.5$ and sample standard deviation is 2.517. Construct a 95% confidence interval and discuss what your results possibly imply.

Solution 1: in JMP



The 95% confidence interval constructed by default in JMP is $[3.56, 7.43]$. We discuss what this means below.

Solution 1: Red Wine

- The sample size is small, therefore to construct a reliable confidence interval we need that the distribution of the blood samples does not deviate much from a normally distribution.

  Discussion of the polyphenol data set When the sample size is so small it is hard to tell from the 9 points on the QQplot whether the data has come from a normal distribution. However, the these points do not deviate too much from the line for us to believe it is skewed. Furthermore, a blood samples tend to come from a biological experiment. Based on these two facts, it seems plausible that the data does not come from a distribution severe skew or heavy tails. If this is the case, the distribution of the data is unlikely to deviate hugely from normality. Thus, the sample mean based on 9 is likely to be close to normal.

- We do not know the standard deviation and JMP estimates it from the

data. Therefore the 95% confidence interval constructed in JMP uses the t-distribution and not the normal distribution.

The exact calculation:

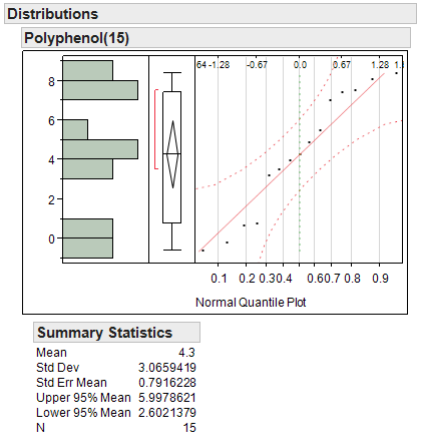
- Use the t-tables with 8df (sample size, 9, minus one) and 2.5%. This gives the critical value 2.306. Based on this the 95% CI for the mean is

$$\left[5.5 \pm 2.306 \times \frac{2.517}{\sqrt{9}}\right] = [3.57, 7.43]\,,$$

which are exactly the numbers given in the JMP output.

Example 2: Red Wine II

- We return to the same question but in order to get a smaller margin of error we include 6 extra males in our study. `http://www.stat.tamu.edu/~suhasini/teaching651/red_wine_polyphenol.txt`.



- Notice some of the new guys actually had a drop in their polyphenol levels!

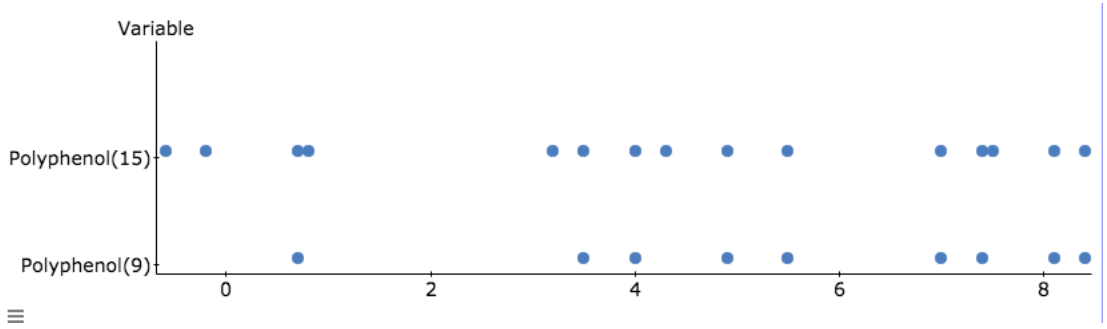
- The sample mean is 4.3 and the sample standard deviation is 3.06.

- Solution We now use a t-distribution with 14 degrees of freedom and the 95% CI for the mean level after drinking wine (for two weeks) is

$$\left[4.3 \pm 2.145 \times \frac{3.06}{\sqrt{15}}\right] = [2.1, 6]\,.$$

- The factor **2.145** has decreased from the **2.306** given in the previous example. This is because, the sample standard deviation based on $n = 15$ tends to be closer to the population standard deviation.

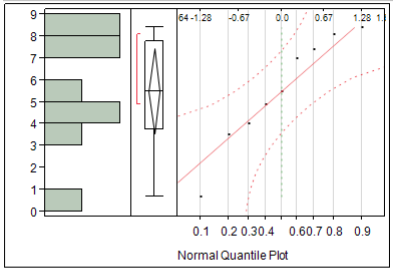
Comparing Example 1 and 2



- We see that the smaller sample size contains less 'extreme' values. As we mentioned earlier, less spread (in the smaller sample size) means that the corresponding estimated standard deviation will be less than the second sample (look at the output below and compare for $n = 9$, $s = 2.5$ whereas for $n = 15$, $s = 3.1$). We see that for smaller sample sizes the estimated standard deviation tends to underestimate the true population standard deviation.

Distributions

Polyphenol(9)

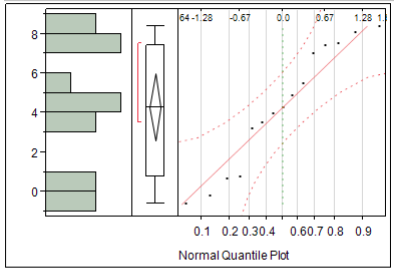


Summary Statistics

Mean	5.5
Std Dev	2.5169426
Std Err Mean	0.8389809
Upper 95% Mean	7.4346933
Lower 95% Mean	3.5653067
N	9

Distributions

Polyphenol(15)

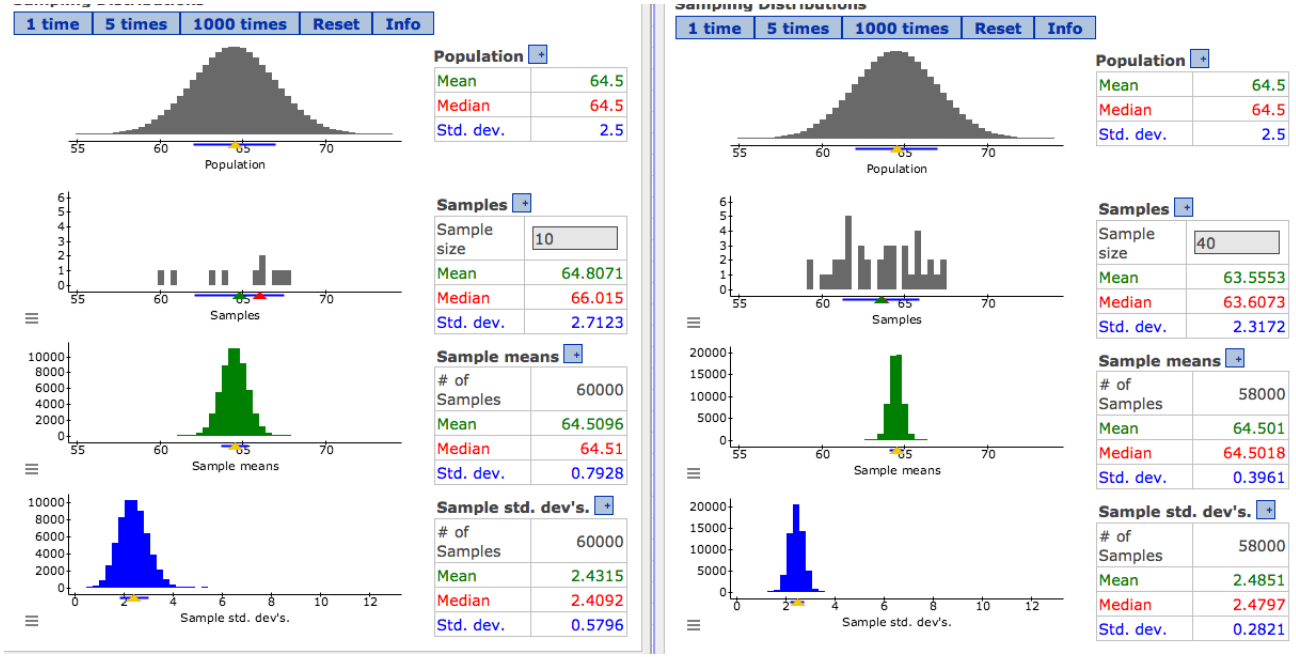


Summary Statistics

Mean	4.3
Std Dev	3.0659419
Std Err Mean	0.7916228
Upper 95% Mean	5.9978621
Lower 95% Mean	2.6021379
N	15

Sample size and the sample standard deviation

As the sample size grows, the *standard error* of the sample mean gets smaller (see green plot) and the sample standard deviation concentrates about the population mean (see blue plot). Below are plots of the distribution of sample means and standard deviation when $n = 10$ and $n = 40$, see the spread reduces as n gets larger.



Critical values and the sample size

As the sample size grows the critical values corresponding to the t-distribution get closer to the critical values of the normal distribution (in this case 1.96).

Percentage points of Student's t distribution

df/α =	.40	.25	.10	.05	.025	3
1	0.325	1.000	3.078	6.314	12.706	
2	0.289	0.816	1.886	2.920	4.303	
3	0.277	0.765	1.638	2.353	3.182	
4	0.271	0.741	1.533	2.132	2.776	
5	0.267	0.727	1.476	2.015	2.571	
6	0.265	0.718	1.440	1.943	2.447	
7	0.263	0.711	1.415	1.895	2.365	
8	0.262	0.706	1.397	1.860	2.306	
9	0.261	0.703	1.383	1.833	2.262	
10	0.260	0.700	1.372	1.812	2.228	
11	0.260	0.697	1.363	1.796	2.201	
12	0.259	0.695	1.356	1.782	2.179	
13	0.259	0.694	1.350	1.771	2.160	
14	0.258	0.692	1.345	1.761	2.145	
15	0.258	0.691	1.341	1.753	2.131	
16	0.258	0.690	1.337	1.746	2.120	
17	0.257	0.689	1.333	1.740	2.110	
18	0.257	0.688	1.330	1.734	2.101	
19	0.257	0.688	1.328	1.729	2.093	
20	0.257	0.687	1.325	1.725	2.086	
21	0.257	0.686	1.323	1.721	2.080	
22	0.256	0.686	1.321	1.717	2.074	
23	0.256	0.685	1.319	1.714	2.069	
24	0.256	0.685	1.318	1.711	2.064	
25	0.256	0.684	1.316	1.708	2.060	
26	0.256	0.684	1.315	1.706	2.056	
27	0.256	0.684	1.314	1.703	2.052	
28	0.256	0.683	1.313	1.701	2.048	
29	0.256	0.683	1.311	1.699	2.045	
30	0.256	0.683	1.310	1.697	2.042	
35	0.255	0.682	1.306	1.690	2.030	
40	0.255	0.681	1.303	1.684	2.021	
50	0.255	0.679	1.299	1.676	2.009	
60	0.254	0.679	1.296	1.671	2.000	
120	0.254	0.677	1.289	1.658	1.980	
inf.	0.253	0.674	1.282	1.645	1.960	

Common misunderstandings

As the sample size grows, two unrelated phenomena occur:

- The distribution of the sample mean gets close to the normal distribution (lecture 11 and 12). This is called the central limit theorem.

- The *sample standard deviation* tends to get closer to the population standard deviation. Consequently, the critical values of the t-distribution converge to the critical values of a normal distribution.

  The fact that the critical values of a t-distribution get closer to those of a normal distribution has **nothing** to do with the central limit theorem.

Conditions for using a t-distribution

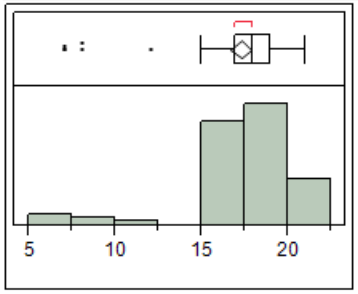
- Observations are from a Simple Random Sample.
- The sample mean is close to normally distributed.

Example: Comparing the mean number of M&Ms in a bag

We now analyse the M&M data to see whether the mean number of M&Ms in a bag vary according to the type of M&M. The data can be found here:
`http://www.stat.tamu.edu/~suhasini/teaching651/MandMs_2013.csv`
There is a proper formal method called ANOVA, which we cover in lecture 24, where we can check to see whether all three have the same mean or not. However, a crude method is to simply check their confidence intervals.

Distributions Type=M

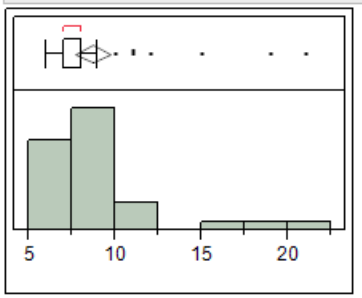
Total



Summary Statistics	
Mean	17.297619
Std Dev	2.8739787
Std Err Mean	0.3135768
Upper 95% Mean	17.921311
Lower 95% Mean	16.673927
N	84

Distributions Type=P

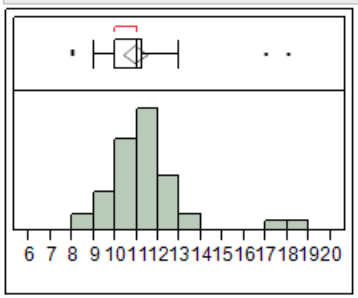
Total



Summary Statistics	
Mean	8.675
Std Dev	3.1328491
Std Err Mean	0.4953469
Upper 95% Mean	9.6769338
Lower 95% Mean	7.6730662
N	40

Distributions Type=PB

Total



Summary Statistics	
Mean	10.913043
Std Dev	1.8236238
Std Err Mean	0.2688787
Upper 95% Mean	11.454593
Lower 95% Mean	10.371494
N	46

Solution: Analysis and interpretation

- As the sample sizes used to construct each confidence interval are large (over 30 in each case), even though the distribution of M&Ms is not normal (they are integer valued!), it is safe to assume that the average is close to normal, therefore these 95% confidence intervals are reliably 95%.

- A summary of the output is given below:
  - Plain: sample mean = 17.2, standard error = 0.31, CI = [16.67,17.92].
  - Peanut: sample mean = 8.6, standard error = 0.49, CI = [7.67,9.76].
  - Peanut butter: sample mean = 10.9, standard error = 0.26, CI = [10.37,11.45].

- As none of the confidence intervals intersect our crude analysis suggests that the means are all different.

- In lecture 19 we will make the above precise (by constructing a confidence interval for the differences in the means).

Statistics in articles

This is a snap shot from the article on the influence of CO2 on diet by Eweis at. al. (2017). Below are the glucose and cholestrol levels in rats after drinking only regular water, a sugar soda, diet soda and decarbonated sugar soda (for 6 months).

Table 1 Blood glucose and cholesterol levels in rats consuming different drinks.

Group	Glucose (mg/dl)	Cholesterol (mg/dl)
Water	157 ± 22	127 ± 3
RCB	187 ± 0.4	135 ± 3
DCB	172 ± 21	135 ± 5
DgCB	192 ± 5	138 ± 3

- The table gives the [sample mean$\pm$ sample standard deviation] for each group. In each group there are 4 rats. From these numbers, we can calculate the 95% confidence intervals for the population mean under each treatment.

- When reading an article it is important to check if the $\pm$ is the margin of error (in which case the authors have given the confidence interval) of the sample standard deviation (in which case you need to construct the CI).

- In the article above the 95% confidence intervals for the mean level of water and RCB (regular soda) are

$$\begin{aligned}\left[157 \pm 3.18 \times \frac{22}{\sqrt{4}}\right] &= [121, 192] \\ \left[187 \pm 3.18 \times \frac{0.4}{\sqrt{4}}\right] &= [186.3, 187.6]\,.\end{aligned}$$

- The intervals intersect, which means we have to be cautious about saying that the different treatment groups have different means.

- However, the variation between the two data sets is very different (22 vs 0.4), which suggests that there are differences in the populations. But we need to keep in mind that these are estimated using very small sample sizes.

- Warning: Comparing the confidence intervals of several treatment groups can lead to "false positives". This is one reason we do ANOVA, which is a method for collectively comparing the means across groups. We cover this later on in the course.

IMPORTANT!!!

- A common mistake that students make is that the t-distribution is used to correct for the non-normality of sample mean (for example when the sample size is not large enough).

- NOOOOOOOOOOOOOOOOOOOOOOO

- In order to use the t-distribution we require that the sample mean is close to normal.

- THE ONLY REASON WE USE THE T-DISTRIBUTION is because the true population standard deviation is unknown and us estimated from the data. The t-distribution is used to correct for the error in the estimated standard deviation.

The t-distribution cannot correct for non-normality of the data

Here we draw a sample of size 10 from a right-skewed distribution and use the t-distribution to construct a confidence interval for the mean. We see that only 87% of the confidence intervals contain the mean. Using the