MAT 240 – Week 3b

Section(s) covered: 6.4, 5.3
Subject: Small sample analysis using the Student t

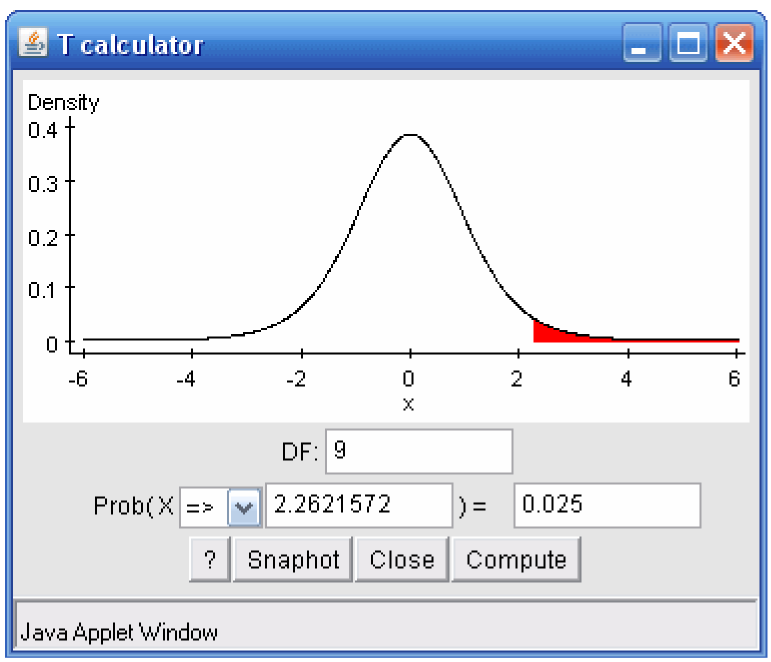
Fundamental ideas:

- We kind of skipped this as being pretty theoretical, but we never really justified why we are allowed to use the normal distribution.
    - We have a dataset, but that data probably does not have a normal distribution
    - We use the normal distribution for hypothesis tests and confidence intervals (assuming that the sample mean has a normal distribution)
    - So we kind of skipped a step as to "why can we use the normal?"
- The theoretical justification (just recognize it, we won't use it) … the important theorem here is the "Central Limit Theorem". This theorem says that
    - If the data is not too "weird"
    - If we have a lot of data points
    - Then the sample mean has approximately a normal distribution
- The question for Sections 6.4 and 5.3 … what if we don't have a lot of data points? Then we can't use the Central Limit Theorem to justify using the normal distribution … so what do we use? We use the <u>Student's t-distribution</u> instead.
- Some history – statisticians knew years ago that you can't use the normal distribution when you have too few points. Then William Gosset figured it out:
    - He figured out another distribution – the "t-distribution" – to use instead of the normal
    - He wanted to publish this in a journal to tell other statisticians, but his company (Guinness, the brewery) considered this a trade secret and would not let any of their scientists publish
    - So he published under a pen name – "Student" – as he knew the editor of one of the big statistics journals
    - Now it's called the "Student-t distribution" because of that … it really should be called the Gossett distribution after him, but …
- Characteristics of the Student t-distribution:
    - There is a different one for each number of points (i.e. a Student t for 2 data points, one for 3 data points, one for 4 data points, etc.)
    - The shapes of all those Student t's look kind of normal, but not exactly
    - For mathematical reasons, the Student t to use for 2 data points is called the one with "1 degree of freedom", for 3 data points is the one with "2 degrees of freedom", …, for 10 data points is the one with "9 degrees of freedom" …
    - You can use Excel, StatCrunch, SPSS, tables, etc. all the usual suspects to do the Student t calculations
    - As you get more and more points (more and more degrees of freedom), the Student t looks more and more like the normal

How to reach a conclusion for a hypothesis test:

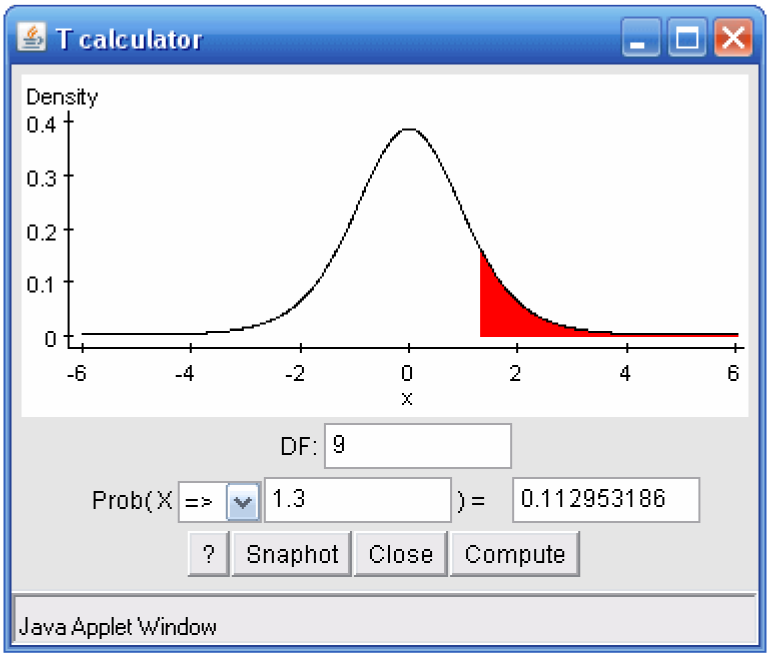
Example 1: Almost the same as Example 1 in Week2b and Week 3a … a hypothesis test. We're measuring weights of elephants. We guess that the average weight of a certain type of elephant is 15,000 pounds.

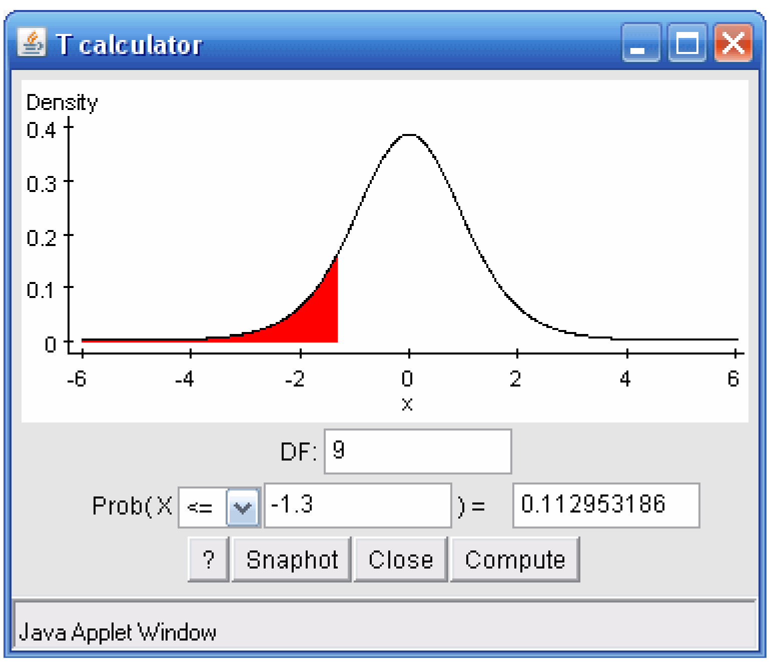
- We go look for them but we only find 10 of them. Again, their average weight is 15,500 pounds and the standard deviation of their weights is 1200 pounds.
- We're going to use the s / √n formula … big surprise … the standard deviation of the sample mean of 15,500 is 1200 / √10 = 379, 15,500 is 500 away … that's 500/379 = 1.3 steps or standard errors away. This is the "t–score" formula … (15,500 – 15,000) / 379 = 1.3. The calculation is exactly the same as the z-score.
- But the roadblock now – we can't use the normal table. The normal table doesn't apply when there are too few points. We must use a Student t table – one with 9 degrees of freedom (remember – one less than the number of points)
- The critical value method. How far are we allowed to be?
  - Using Excel – instead of NORMSINV (inverse of the normal), we use TINV (inverse of the T … makes sense?). So instead of NORMSINV(.025) = –1.96 (for the 2.5% point of the normal, we use TINV(.025,9) = 2.685. That means that we are allowed to be 2.685 steps away … more than the 1.96 that the normal allows.
  - Using StatCrunch – instead of Stat – Calculators – Normal, we use Stat – Calculators – T (T instead of Normal, makes sense?). StatCrunch gives 2.26 as the answer … here's the picture:



  - WEIRD WEIRD WEIRD! Excel and StatCrunch give different answers.
  - In both cases, though, 1.3 is less than the value (2.685 for Excel, 2.262 for StatCrunch).
- The p-value method

- Excel – use the formula TDIST(1.3,9,1) or TDIST(1.3,9,2). Here's the problem … should we use 1 or 2? 1 means to calculate only on one side … 2 means to calculate on both sides (so the 2 number is twice the 1 number … remember that we double the probability). Here, we need to use the 2 number … that's either p = 0.226 (from the TDIST(1.3,9,2) or p = 2*0.113 = 0.226 (from twice the TDIST(1.3,9,1) number)
- StatCrunch – put 1.3 into the "x" field, hit calculate, and get 0.113 in the "probability" field. The p-value for being 1.3 or more away is twice this, or 0.226. It's the 0.113 shown in red plus the one on the other side.





- Conclusion: We're not too wrong. The p value is more than 0.05, so we do not reject the null hypothesis.

Summary – it's just the same as last week … but instead of the normal distribution, we use the Student t distribution. Exactly the same but with a different lookup. We do this when we don't have too many data values to use.

A BIG warning about using Excel:

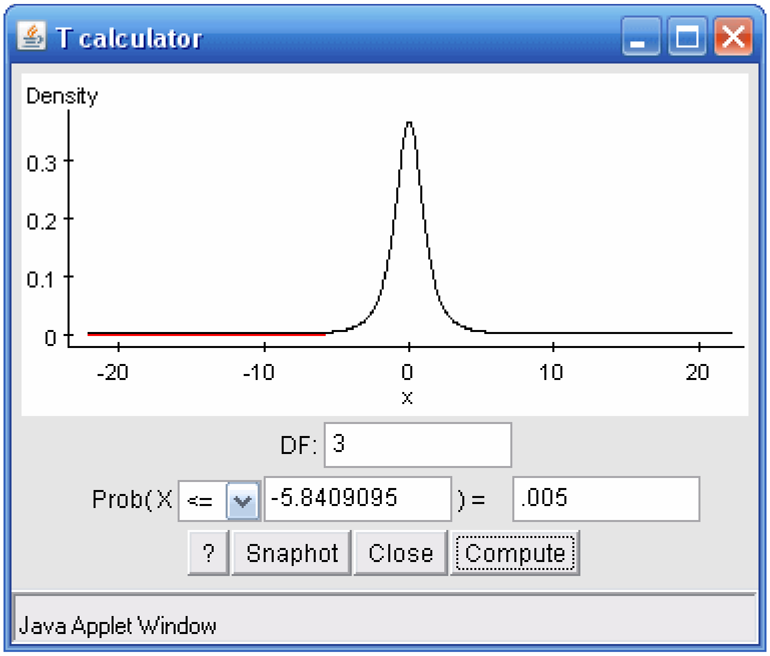
Why does Excel and StatCrunch give us different answers? For the calculation of the critical value, Excel gives =TINV(.025,9) = 2.685 and StatCrunch gives 2.262. Which one is right?

- They are both right, but Excel pulled a sneaky one on us. If you don't watch carefully, Excel will give you the wrong answer!
- Here's the problem with Excel:
    - NORMSINV(.025) gives the P[Normal < something], a .025 left-tailed calculation
    - Most of the other INV functions would also give a .025 left-tailed number
    - However, CHIINV(.025) gives the P[Chi > something], a .025 right-tailed calculation
    - AND FURTHER … that's not how TINV works! It's different from all of the other Excel functions. TINV(9,.025) = P[T < –something OR T > something], a .025 two-tailed calculation
- So TINV works differently in Excel than most of the other EXCEL stats functions. Confusing! A real good opportunity to mess up.
- What makes me more frustrated is that Excel doesn't give any diagram or anything … it just dumps a number (StatCrunch gives you that nice picture). So you don't really know that you did this calculation in a different way.

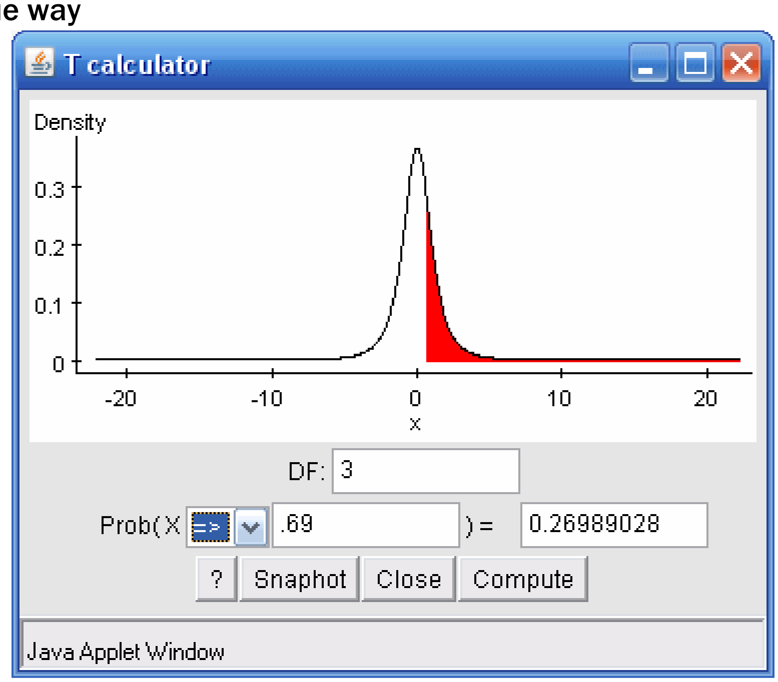
Another Example

Example 2: Almost the same as Example 2 from Week 2b and 3a … a confidence interval. We're guessing that the average amount of soda in a 2.0 liter bottle is 2.0 liters. Let's say that we want to be really sure … 99% sure instead of just 95%. But instead of 1.96, or 2.576 what do we use?

- We have a sample size of 4 bottles, a sample mean of 1.989 liters and a sample standard deviation of .032 liters.
- What's the next step? Of course it's .032 / √4 = .016, the standard error.
- The t-score is (2 – 1.989) / .016 = 0.69
- The critical value way
    - The critical value (careful using Excel!) for 3 degrees of freedom is 5.84
    - 0.69 is less than 5.84… so we do not reject



- The p-value way



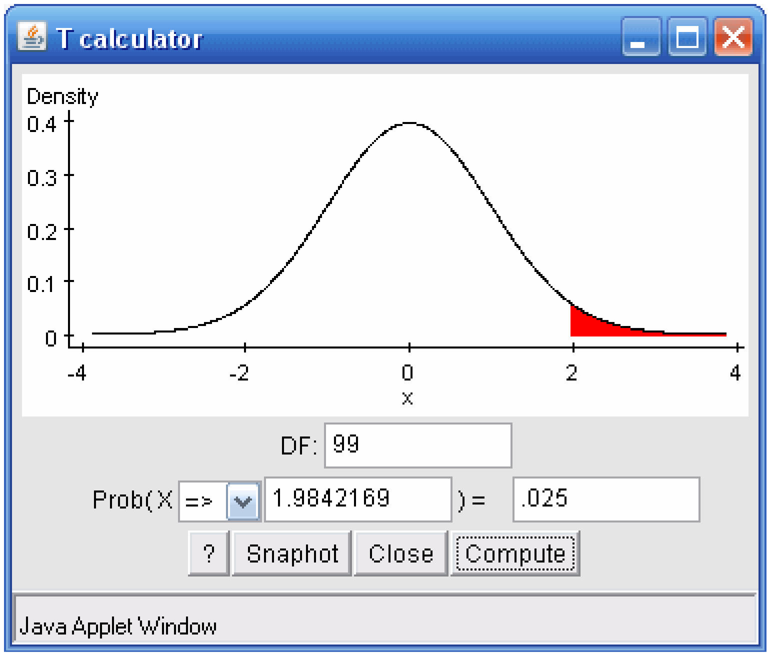
  - P[T > 0.69] = .27 … so P[T < −0.69] = .27 also, and the probability of being 0.69 or more away (either greater 0.69 or less than −0.69) is .54.
  - .54 is more than .01 … it's OK likely … so we do not reject

Comparing the z and the Student t methods

Some comparisons between the two:

- In theory, the Student t is more correct – always. That's why all statistical software give the Student t statistic, the Student t probability, etc. Since it's more theoretically correct, that's what they do.

- If you have more than 30 points or so, the calculations are very similar. For example, for the 0.05 critical value. We informally used 2.0 … 2 "steps" away. The theoretical normal calculation is 1.96, and I consider 2.0 close enough unless you were really an*l about it. The Student t for 100 data points (degrees of freedom = 99) gives a value of 1.98. So, 1.96, 1.98, 2.0 … all basically the same. So we use the normal distribution for large numbers of points because the calculation is just easier.



- For critical values, the Student t ALWAYS gives higher ones than the normal. So it makes it more strict to reject the null hypothesis using the Student t.
- For p-values, the Student t ALWAYS gives lower ones than the normal. So it makes it more strict to reject the null hypothesis using the Student t.
- A rule of thumb: If you have 30 points or more, then the normal distribution is close enough. If you have less than 30 points, particularly a very few points (like 4 or 6) and a very stringent criteria (like 0.01 chance), then the calculations are different. Remember the 0.005 calculation for a T with 3 degrees of freedom above? The critical value was 5.84. The same calculation for a normal is 2.58. 5.84 is very different from 2.58.

I didn't work any example from the text this time, I figured that you are getting the hang of it. Just use the T instead of the normal here.