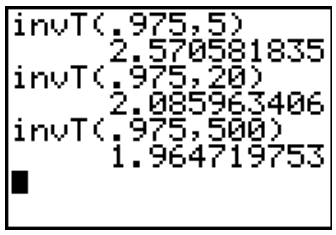
Chapter 11 – Inference for Means

Inference for means is a little different than that for proportions. Most introductory statistics texts base this on standard normal models, which is truly appropriate only if the population standard deviation, σ, is known. In most cases this is not true; the only time one might really believe σ is known is in the case of quality control sampling where a production line has been tracked for a long time. If σ is not known confidence intervals and hypothesis tests should be based on *t* distributions. T distributions have larger critical values (multipliers in confidence intervals) than the standard normal curve to allow for the additional uncertainty in having estimated two parameters of the population – the mean and standard deviation, instead of just one (as with proportions). These distributions become the standard normal distribution when the sample size is very large (infinite).

On the TI-84 or 89, we can see the impact of sample size on these critical values. From the DISTR menu (2nd VARS, or F5 on a TI-89), choose 4:invT(. As with InvNorm which is used to find percentiles of the normal distribution, the parameters for this command are the area to the left of the desired point and the degrees of freedom ($n - 1$ for a single sample). For a 95% confidence interval, we saw before that z^* is 1.96. The *t* critical values at right are also for a 95% interval (if there is 95% in the middle of the curve, there is 97.5% to the left of the high end of the region) and represent samples of size 6, 21, and 501. Notice that as degrees of freedom or sample size get larger, these numbers get closer to 1.96.



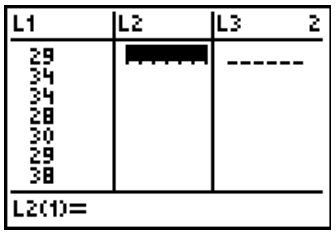
Small sample sizes give rise to their own problems. If the sample size is less than about 30, the Central Limit Theorem does not apply, and one cannot merely assume the sample mean has a normal distribution. In the case of small samples, you must check that the data come from a (at least approximately) normal population, usually by normal probability plots or boxplots since histograms are not useful with small samples.

CONFIDENCE INTERVALS FOR A MEAN

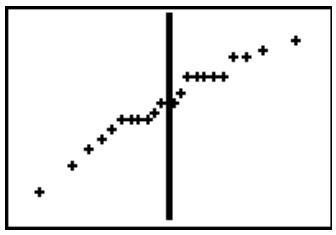
Residents of Triphammer Road are concerned over vehicles speeding through their area. The posted speed limit is 30 miles per hour. A concerned citizen spends 15 minutes recording the speeds registered by a radar speed detector that was installed by the police. He obtained the following data:

29	34	34	28	30	29	38	31	29	34	32	31
27	37	29	26	24	34	36	31	34	36	21	

We want to estimate the average speed for all cars in this area, based on the sample. Enter the data in a list. Here, I have entered them into list L1. This is a small sample – there are only 23 observations, so we should check to see if the data looks approximately normal.

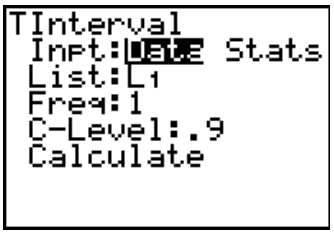


A normal plot of the data looks relatively straight, with no outliers, so it's reasonable to continue. This plot shows some granularity (repeated measurements of the same value) but no overt skewness or outliers. If you've forgotten how to create normal probability plots, return to Chapter 4 of this manual.



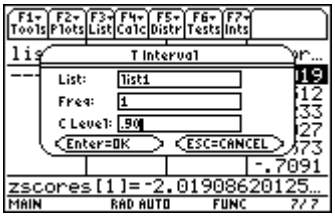
Copyright 2010 Pearson Education, Inc.

TI-83/84 Procedure

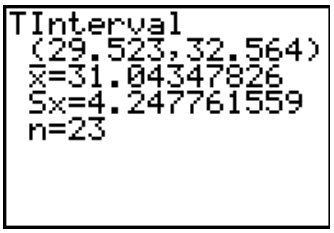


Press [STAT], arrow to TESTS then select choice 8:Tinterval. You have two choices for data input: using data in a list such as we have or inputting summary statistics from the sample. Move the cursor to DATA and press [ENTER] to move the highlight. Enter the name of the list with the data ([2nd][1] for L1). Each observation occurred once, so leave Freq as 1. If there were a separate list of frequencies for each data value, that would be entered here. Enter the desired amount of confidence (here, 90%, but in decimal form) and finally press [ENTER] to perform the calculation.

TI-89 Procedure



From the Statistics/List editor, press [2nd][F2] =[F7] (Ints). Select choice 2:Tinterval. You have two choices for data input: using data in a list such as we have or inputting summary statistics from the sample. Pressing the right arrow allows you to make the selection. Press [ENTER] to get the next input screen. Enter the name of the list with the data ([2nd][−] takes you to the [VAR-LINK] screen). Each observation occurred once, so Freq should be 1. If there were a separate list of frequencies for each data value, that would be entered here. Enter the desired amount of confidence (here, 90%, but in decimal form) and finally press [ENTER] to perform the calculation.

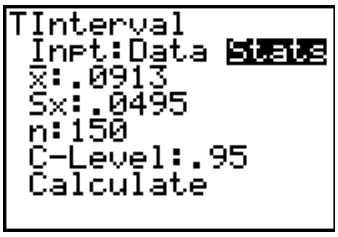


Here are the results. Based on this sample, we are 90% confident the average speed for all cars on this road is between 29.5 and 32.6 miles per hour. There are two caveats here: the first is that this was not a truly random sample but a convenience one (only one 15 minute period was sampled). Also, the presence of the radar speed detector may have influenced the drivers at that time. Drivers may be driving over the posted 30 miles per hour limit, but since 30 is included in the interval, we have not shown conclusively that drivers are speeding, on average.

What about the extra decimal places? The general rule here, as with reporting means and standard deviations in general, is to report one more decimal place than in the original data. Our data were in integer miles per hour, so report one decimal place. Your instructor may have a different rule for this, so please listen to him or her.

What if we don't have the data? In the case of a small sample size, one must assume the data comes from an approximately normal population. If the sample is "large," the Central Limit Theorem will apply and $\bar{x}$ will be approximately normal.

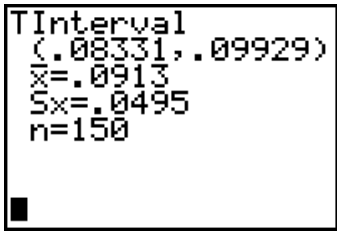
Another Example



In 2004 a team of researchers published a study of contaminants in farmed salmon. Fish from many sources were analyzed for several contaminants, one of which was the insecticide mirex. After outliers from one particular farm were removed, the remaining 150 fish averaged 0.0913 ppm with stancard deviation s = 0.0495 ppm. What is a 95% confidence interval for the mean mirex contamination? Here we have moved the highlight from Data to Stats. When this is done, the input screen changes to ask for the sample mean, standard deviation, sample size and confidence level.

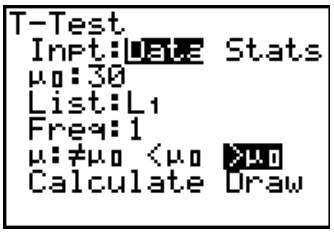
Copyright 2010 Pearson Education, Inc.

Pressing ENTER to calculate the interval tells us we are 95% confident, based on this sample the mean mirex contamination in farm-raised salmon is between 0.083 and 0.099 ppm.

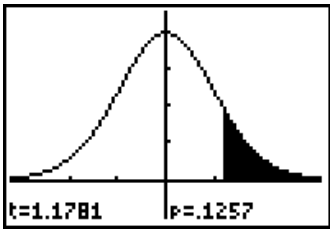


A ONE SAMPLE TEST FOR A MEAN

We can also do a hypothesis test to decide whether the mean speed is more than 30mph. From the STAT TESTS menu, select choice 2:Ttest. We are still using data in list L1. μ_0 is set to 30 since that's the speed limit we're comparing against. The alternate has been selected as $> \mu_0$ since we want to know if people are going too fast, on average. Notice we have the options of Calculate and Draw here, just as we did on tests of proportions.



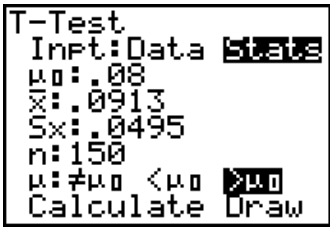
Selecting Draw yields the screen at right. We can clearly see the shaded portion of the curve which corresponds to the p-value for the test of 0.1257. The calculated test statistic is $t = 1.178$. The p-value indicates we'll expect to see a sample mean of 31.04 (the mean from our sample) or higher by chance about 12.5% of the time by randomness when the mean really is 30. That's not very rare. We fail to reject the null and conclude these data do not show motorists on the street are speeding on average.



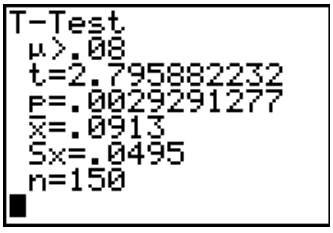
Another Example

Researchers tested 150 farm-raised salmon for organic contaminants. They found the mean concentration of the carcinogenic insecticide mirex to be 0.0913 parts per million (ppm), with standard deviation 0.0495 ppm. The Environmental Protection Agency's recommended "screening value" for mirex is 0.08 ppm. Do farm-raised salmon appear to be contaminated beyond the level permitted by the EPA?

The salmon were randomly selected, and raised and purchased in many places, so they should be independent of each other.. Further, these represent a really small fraction of the potential salmon available for sale. With a sample size of 150, the actual shape of the distribution is of small concern (it's actually somewhat right skewed, with no outliers.) Since all the conditions are met, we may proceed to the test. We want to know if these salmon appear to have contaminant levels that exceed the EPA permitted, so the form of the alternate hypothesis is "$>\mu_0$."



We see on the results screen that if the level were indeed .08 (or less) that the observed mean of 0.0913 is 2.80 standard deviations above that level. The probability of our observed sample mean or something higher is 0.0029. This is an extremely small p-value, so we have very strong evidence that these fish do indeed exceed the EPA screening value. One might want to think twice about eating farm-raised salmon.



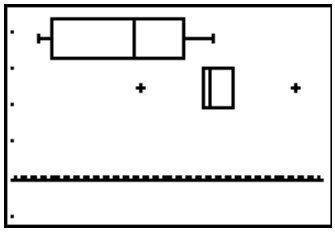
Copyright 2010 Pearson Education, Inc.

COMPARING TWO MEANS – CONFIDENCE INTERVALS

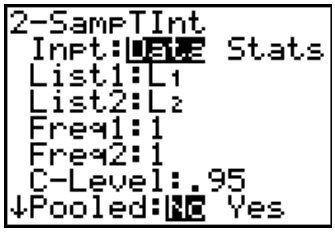
Should you buy name brand or generic batteries? Generics cost less, but if they do not last as long on average as the name brand, spending the extra money for the name brand may be worthwhile. Data were collected for six sets of each type of battery, which were used continuously in a CD player until no more music was heard through the headphones. The lifetimes (in minutes) for the six sets were:

Brand Name:	194.0	205.5	199.2	172.4	184.0	169.5
Generic:	190.7	203.5	203.5	206.5	222.5	209.4

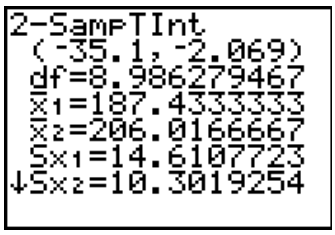
The first step in performing a comparison such as this one (or any!) should always be to plot the data. Here, a side-by-side boxplot is natural.



We have entered the data into list L1 for the Brand name batteries, and list L2 for the generics. We defined two boxplots to identify outliers on the STAT PLOT menu (2nd Y=). For more on these plots, see Chapter 3 of this manual. From the plot, it certainly appears the generics last longer than the name brand batteries; they also seem more consistent (they have a smaller spread). There are two outliers for the generic batteries, but with a sample size this small the outlier criteria are not very reliable. Neither of the extreme values are unreasonable, so it's safe to continue.



From the STAT TESTS menu select choice 0:2-SampTInt (option 4 on the F7 Ints menu on a TI-89). Our data are already entered, so move the highlight (if necessary) to Data and press ENTER. The data were in lists L1 and L2, and each value in the lists occurred once. The confidence level has been set to 95% (entered as always in decimal form). The next option is new. Pooled: refers to whether the two groups are believed to have the same standard deviation. Visually, this is not true for our two battery samples. In general, unless there is some reason to believe the groups have the same spread, it's safest to answer this question with No. Reasoning behind this question has to do with computing a "pooled standard deviation" (or not) and the number of degrees of freedom for the test. Before the advent of computers (and statistical calculators) there were many recipes for handling this question, since the calculation of degrees of freedom in the unpooled case is complex. Luckily, we just let the calculator do the work.



Pressing ENTER to calculate the interval gives the screen at right. We see we are 95% confident the average life of the name brand batteries is between 35.1 and 2.1 minutes *less* than the average life of the generic batteries. (Remember, it's always *group*1 – *group*2 in the interval, just as we found with proportions). The next line gives the degrees of freedom for the interval – notice they're not even integer-valued. We also see the two sample means and standard deviations. The ↓ at the bottom left indicates more output can be obtained (the sample sizes). Assuming generic batteries are cheaper than name brand ones, it certainly would make sense to buy them.

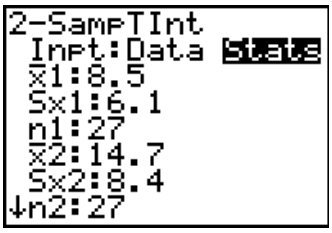
The Subtly Refilling Soup Bowl

Do people take visual or internal cues when they eat? Researchers wanted to examine this question. Twenty-seven people were each assigned randomly to eat soup – one group from regular bowls, and the other to eat soup from bowls that were secretly refilled. Which group ate more? How much more? The results of the experiment are summarized in the table below.

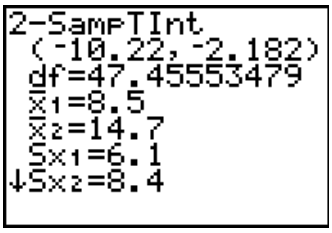
Copyright 2010 Pearson Education, Inc.

	Ordinary bowl	Refilling Bowl
n	27	27
$\bar{y}$	8.5 oz.	14.7 oz
s	6.1 oz.	8.4 oz.

It appears that the people with the refilling bowl are more, but is the difference statistically significant? What does a 95% confidence interval say about the difference? In the screen at right, I have entered the summary statistics given above.



Group 1 was the group with the ordinary bowl. We notice that both ends of the confidence interval are negative. This means that we are 95% confident based on this experiment that people with ordinary bowls will eat between 2.18 and 10.22 oz. *less* than people with the refilling bowl. It would seem that fullness of the bowl (rather than the stomach) is the more important cue.

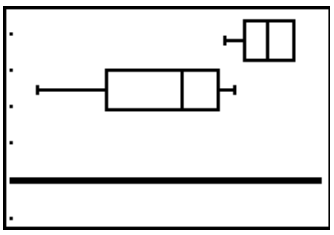


TESTING THE DIFFERENCE BETWEEN TWO MEANS

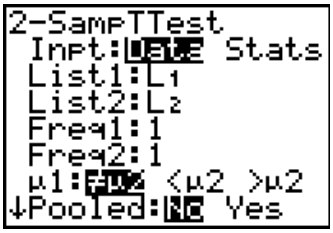
If you bought a used camera in good condition, would you pay the same amount to a friend as to a stranger? A Cornell University researcher wanted to know how friendship affects simple sales such as this.[1] One group of subjects was asked to imagine buying from a friend whom they expected to see again. Another group was asked to imagine buying from a stranger. Here are the prices offered.

Friend	$275	300	260	300	255	275	290	300
Stranger	260	250	175	130	200	225	240	

Here are side-by-side boxplots of the data. There certainly looks to be a difference. Prices to buy from strangers seem lower and much more variable than the prices for buying from a friend. As with the battery example, looking at skewness or outliers for these small samples is difficult, but the plots look reasonable.



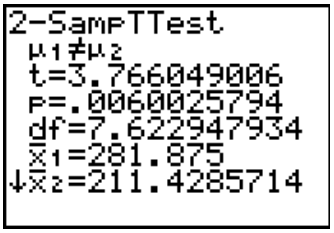
From the STAT TESTS menu, select choice 4:2-SampTTest. Again, we have the data in two lists, so Data is highlighted as the input mechanism, we have indicated the data are in lists L1 and L2, and each data value has a frequency of 1. The alternate hypothesis is $\mu_1 \neq \mu_2$ since our original question was "would you pay the same amount." Again, we have indicated No in regards to pooling the standard deviations (the spreads of the distributions do not look equal and there is no reason to believe they should be the same).



[1] Halpern, J.J. (1997). The transaction index: A method for standardizing comparisons of transaction characteristics across different contexts, *Group Decision and Negotiation*, 6(6), 557-572.

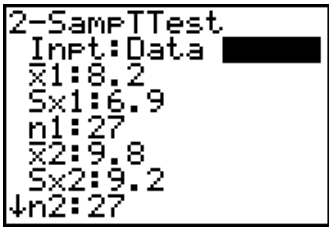
Copyright 2010 Pearson Education, Inc.

Pressing ENTER to calculate the test gives the screen at right. The computed test statistic is t = 3.766, and the p-value is 0.006. From these data we conclude that not only are people not going to pay the same amount to a friend as to a stranger, they're willing to pay more. We might even go so far as to warn people not to pay *too much* to friends.

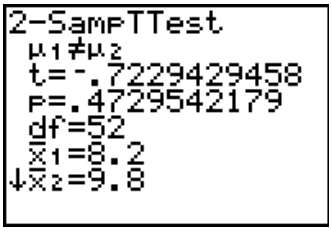


Back to the Soup

This manual (and the authors of your text) have said that unless there is some reason to believe the spreads of the two samples should be the same, it's best to use the non-pooled test. There are occasions where answering "Yes" to the pooled question makes sense. The individuals in the soup experiment not only had the actual amount eaten measured, but they were asked how much they thought they had eaten. If the two groups really were equivalent before the soup experiment, they should have similar standard deviations, and in fact, the standard deviation for the ordinary bowl was 6.9 oz and the standard deviation for the refilling bowl was 9.2 oz. We want to examine the question of whether or not there is a significant difference in the amount of soup the people thought they'd eaten. Since the two groups should have similar spreads (and the statistics are similar), this is a good argument to use a pooled standard deviation with these data.



The results indicate no significant difference in the amount the subjects thought they'd eaten. The p-value for the test is 47.3%. Notice the degrees of freedom here are different from those used in the confidence interval.



PAIRED DATA

The two sample problems considered above used two *independent* samples. Many times data which might seem to be for two samples are naturally paired (say, examining the ages of married couples – each couple is a natural pair) or are even two observations on the same individuals. In such cases one works with the *differences* in each pair, and not the two sets of observations. The reason for this is to eliminate variability among the pairs and focus on the difference within the pairs.

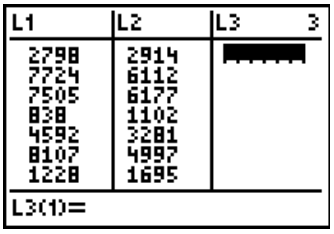
Do flexible schedules reduce the demand for resources? The Lake County (IL) Health Department experimented with a flexible four-day week. They recorded mileage driven by 11 field workers for a year on an ordinary five-day week, then they recorded the mileage for a year on the four-day week.[2] The data are below. The first important fact to realize is that we have data on the same individuals under the two different schedules. These are *not* independent samples, but rather *paired data*.

Name	5 day mileage	4 day mileage
Jeff	2798	2914
Betty	7724	6112
Roger	7505	6177
Tom	838	1102
Aimee	4592	3281
Greg	8107	4997
Larry G	1228	1695
Tad	8718	6606
Larry M	1097	1063
Leslie	8089	6392
Lee	3807	3362

[2] Catlin, Charles S. Four-day Work Week Improves Environment*, Journal of Environmental Health*, Denver, March 1997 **59**:7.

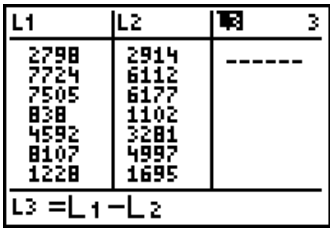
Copyright 2010 Pearson Education, Inc.

Cursory examination reveals that after the change, some drove more, and some less. It is also easy to see there are large differences in the miles driven by the different workers. It is this variation between individuals that paired tests seek to eliminate.



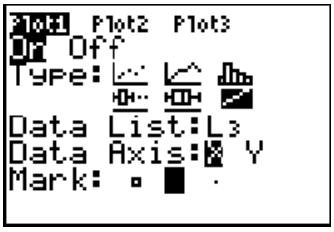
We have entered the data into the calculator; the five-day week mileages are in list L1, and the four-day mileages are in list L2.

We need to find the differences. On the home screen, one could press 2nd 1 − 2nd 2 STO▸ 2nd 3 which results in the command L1 – L2 →L3. However, since we are in the editor, an easier way is to move the cursor to highlight the name of an empty list and enter the command.

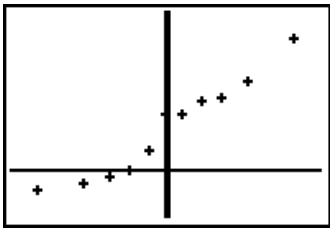


The command will look as at right. On a TI-89, the command is the same, with the exception of using [VAR-LINK] to access list names. Pressing ENTER to complete the calculation will display the first few values. If you want to see the entire list, scroll through it using the down and up arrows.

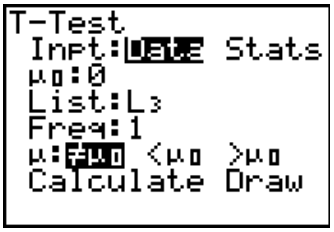
We need to check if the *differences* are approximately normal (or certainly at least have no strong skewness or outliers). For this type of test, we are using the differences as the data so the Nearly Normal condition applies to them and not the original data. We define the normal plot as at right to use the differences which were just created. Press ZOOM 9 to display the plot.



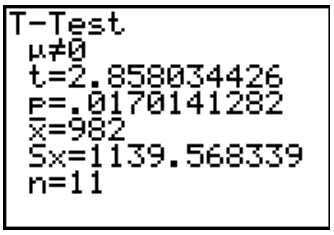
The plot at right is not perfectly straight. However, there are no large gaps, so no extreme outliers.



We now proceed to the test. We will perform a one-sample test using the differences as the data. From the STAT TESTS menu, select 2:T-Test. If the change in work week made no difference, the average value of the computed differences should be 0, so this is the value for μ_0. We are using the data from list L3 as the input, and have selected the alternative hypothesis as $\mu \neq \mu_0$.

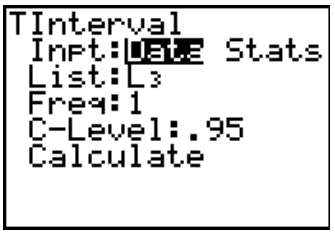


Pressing ENTER when the cursor is over Calculate displays the results. The computed test statistic is $t = 2.85$ and the p-value is 0.017. We conclude that these data do indicate a difference in driving patterns between a five-day work week and a 4-day work week. Further since the average difference is positive (982 miles) it seems that employees drove less on the four-day week (the subtraction was five-day – four-day mileages). It's hard to say if the difference is meaningful to the department (remember, statistical significance is not necessarily practical significance). If so, they may want to consider changing all employees to four-day weeks.

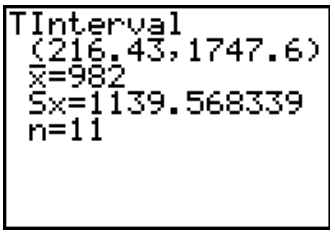


Copyright 2010 Pearson Education, Inc.

We can go further and compute a confidence interval for the average difference. Select `8:TInterval` from the `STAT TESTS` menu, and define the interval as at right.



Pressing ENTER to calculate the interval, we find we are 95% confident the five-day work week will average between 216.4 and 1747.6 more yearly miles than a 4-day work week.

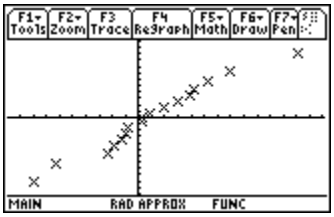


Speed Skaters

In the 2006 Olympics, there were allegations that the outer lane in the speed skating competition was "faster" than the inner lane. The racers are randomly assigned to race in pairs throughout the day of competition, and conditions may vary as the day goes on. Even though the skaters switch lanes halfway through the race, it was believed that those who started in the outer land had an advantage. Is this so? The table below gives the times (and differences) for each pair of skaters in the women's 1500m race, according to which lane was the first.

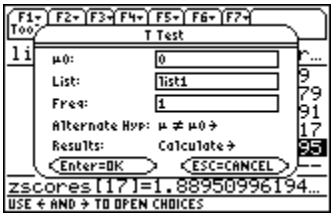
Skating Pair	Inner Time	Outer Time	Difference
1	125.75	122.34	3.41
2	121.63	122.12	-0.49
3	122.24	123.35	-1.11
4	120.85	120.45	0.40
5	122.19	123.07	-0.88
6	122.15	122.75	-0.60
7	122.16	121.22	0.94
8	121.85	119.96	1.89
9	121.17	121.03	0.14
10	124.77	118.87	5.90
11	118.76	121.85	-3.09
12	119.74	120.13	-0.39
13	121.60	120.15	1.45
14	119.33	116.74	2.59
15	119.30	119.15	0.15
16	117.31	115.27	2.04
17	116.90	120.77	-3.87

The normal plot at right of the differences shows no overt skew or outliers. A histogram of the differences also appears relatively symmetric.

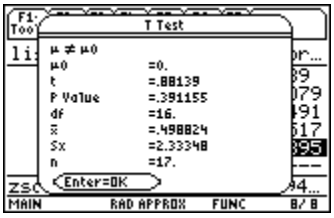


Copyright 2010 Pearson Education, Inc.

Here is my input screen for the test using the TI-89. The differences are in `list1`, and our question of interest is whether (or not) the two lanes are different, so the alternate is "not equal." Again, if the two lanes are equivalent (fair), the average difference in the pairs times should be 0.



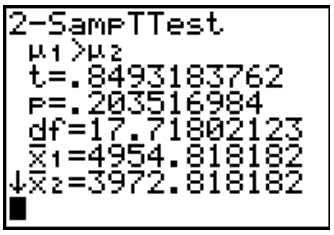
The results indicate there was no advantage from either lane. With a *t* statistic of 0.88 and a p-value of 0.3912, we fail to reject the null hypothesis of no difference. Our data do not indicate any unfairness due to lane assignment.



WHAT CAN GO WRONG?

Not Pairing Paired Data

This is a critical mistake. One needs to think carefully if there is some natural pairing of data that might (possibly) come from independent samples. Clearly, if the samples sizes are not the same, the data cannot have been paired. If one fails to pair data that should be paired, wrong conclusions will usually be made, due to overwhelming variability between the subjects. Here, we have the output if we had (wrongly) not used the paired test on the mileage data. Notice we would have made the opposite conclusion – the large p-value of more than 20% would indicate no difference in mileage due to shortening the work week.



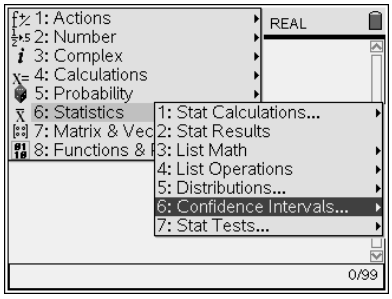
Bad Conclusions

The biggest thing to guard against is bad conclusions. Think about the data and what they show. Do not let conclusions contradict a decision to reject (this means we believe the alternate is true) or not reject (this means we have failed to show the null is wrong) a null hypothesis.

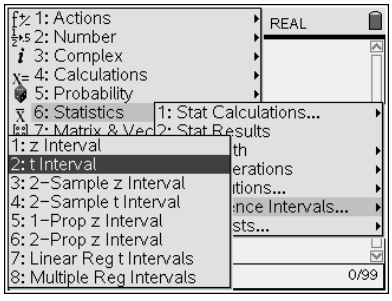
Other than that, there is not much that hasn't already been discussed – trying to subtract lists of differing length will give a dimension mismatch error. Having more plots "turned on" than are needed can also cause errors.

Commands for the TI-Nspire™ Handheld Calculator

For statistical inference, start on a calculator page. Press (menu) , then select Statistics and Confidence Intervals.

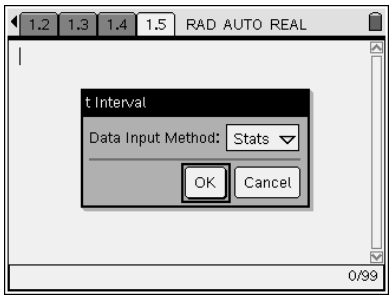


For a one sample mean problem, select t Interval.

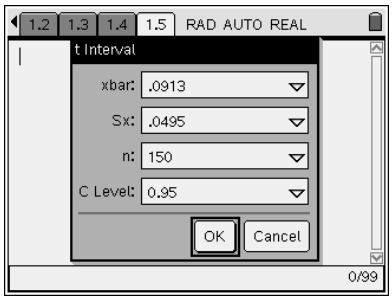


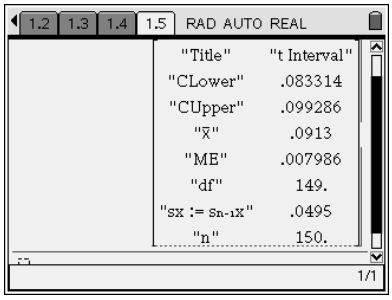
Copyright 2010 Pearson Education, Inc.

For the salmon example, select Stats.

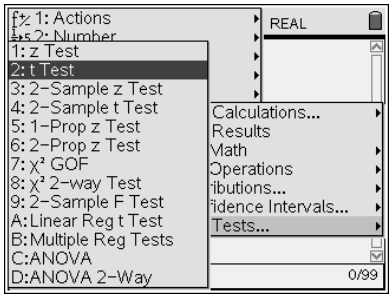


In the input box, type the sample mean, the sample standard deviation the sample size, and the confidence level.

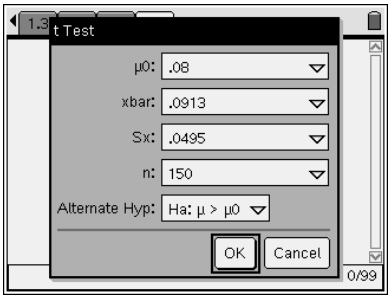




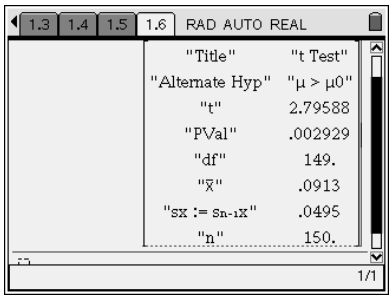
For a one sample mean hypothesis test, press (menu) , select Statistics, Stat Tests, and t Test.



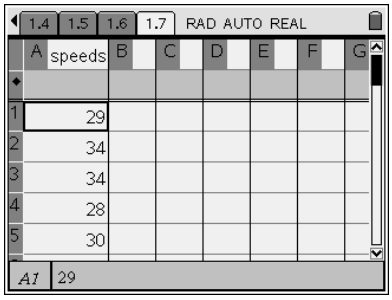
To conduct the test on the salmon data, select Stats as before, and complete the input box.



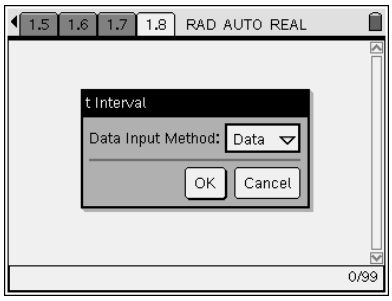
Copyright 2010 Pearson Education, Inc.



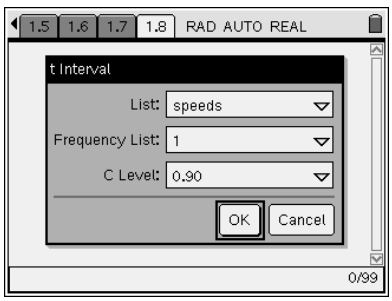
For some inference problems, you will enter all of the sample data. For the auto speeds, start with a list of the data.

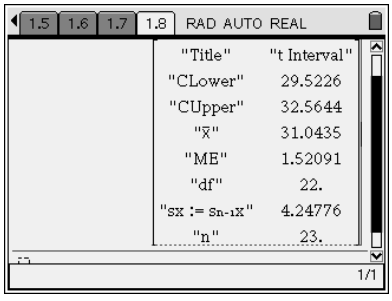


Select the t interval as above, but select Data rather than Stats at this screen.



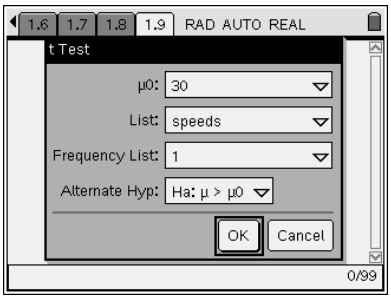
Select the list name and the confidence level.

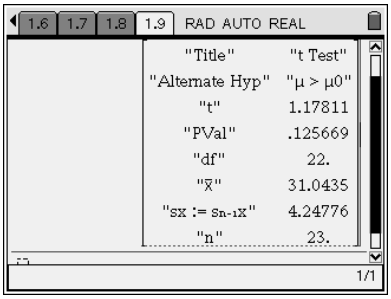




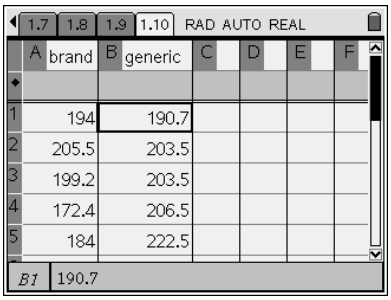
Copyright 2010 Pearson Education, Inc.

You can conduct the hypothesis test in a similar manner.

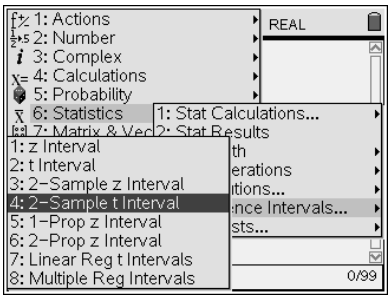




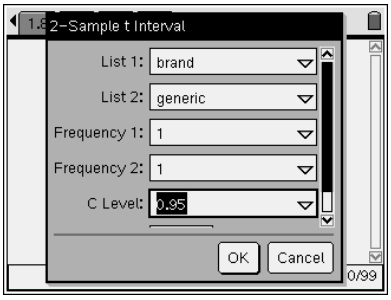
For the two sample battery problem, place the data in two lists.



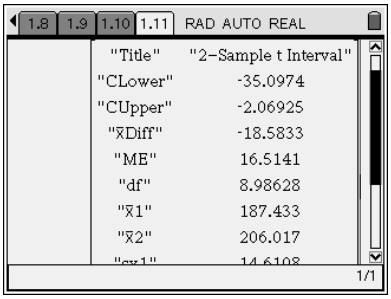
From the Confidence Interval menu, select 2-sample t Interval.

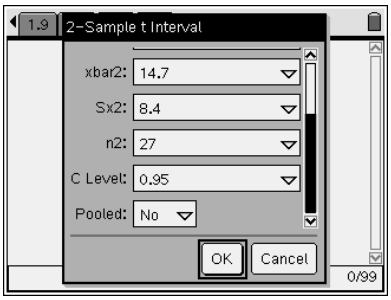


Select Data for the first input box, and then the list names and confidence level in the next input box.

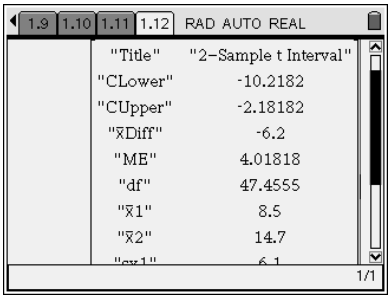


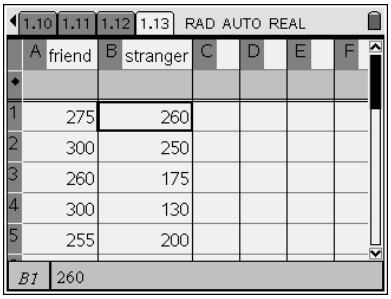
Copyright 2010 Pearson Education, Inc.



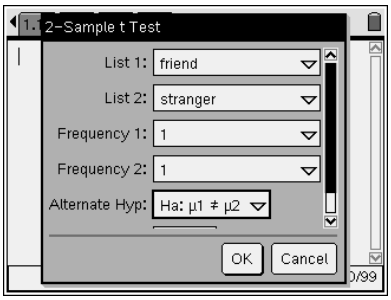


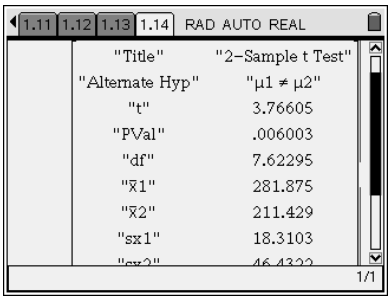
For the soup example, select Stats in the first input box, and complete the second input box.





A two sample t test is conducted in a similar manner. The selling the camera is shown.





Copyright 2010 Pearson Education, Inc.