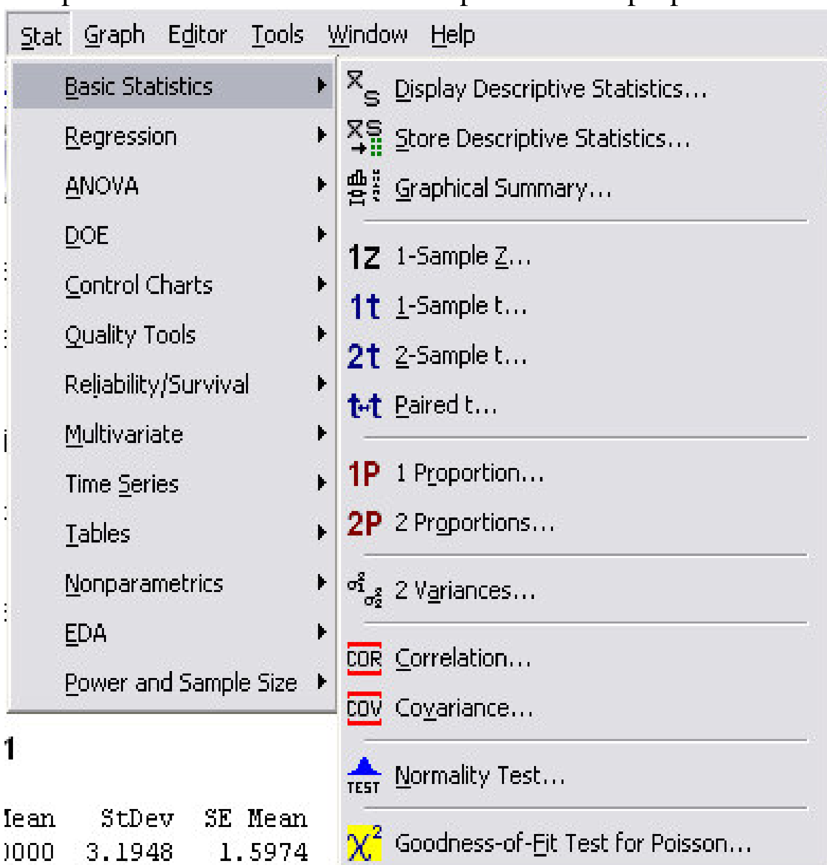
Stat 538 - Biostatistics I - Fall 2005

Lab 5

Confidence intervals of a population mean and population proportion

Minitab can calculate the Confidence intervals by choosing the Stat/Basic Statistics/1t 1-Sample t for the mean and 1P 1 Proportion for a proportion.



But before we let Minitab do them for us, we'll make sure we can do them by hand (mostly).

1. Confidence Intervals for μ

For continuous data, we can easily obtain confidence intervals for the population mean, μ. The formula is summarized on page 214. I repeat it here:

$$\bar{y} \pm t_{.025,df} \frac{s}{\sqrt{n}}, \text{ where}$$

$\bar{y}$ is the sample mean

$t_{.025,df}$ is the critical value of t from the t-distribution with right-tail probability .025 with df=n-1 degrees-of-freedom.

s is the sample standard deviation

$\sqrt{n}$ is the square-root of the sample size n.

Example 1

From these four values, calculate the sample statistics required for the confidence interval, then construct a 95% confidence interval.

Below are the number of pages for the last 3 labs (note that lab 1 had two parts).

3 5 4 8

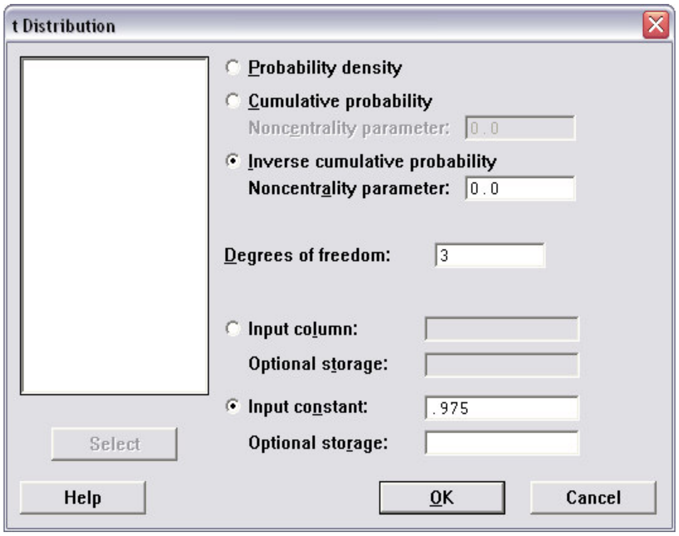
The summary statistics are given below from Minitab.

Notice that SE Mean = $s/\sqrt{n}$, that is, 1.08 = 2.16/√4 = 2.16/2.

Descriptive Statistics: C1

```
Variable  N  Mean  SE Mean  StDev
C1        4  5.00     1.08   2.16
```

We now have $\bar{y}$ = 5, s=2.16, n=4. Lastly, we need the value of t, which we can get from the Calc/Probability Distributions/t, choose "**Inverse cumulative probability**", input the degrees of freedom as n-1, in this case 4-1=**3**. The only part you need to think about is for the input constant. If we have a 95% interval, then 5% is in the tails, 2.5% on the left and 2.5% on the right. Thus, we want a cumulative probability of .95+.025 on the left for a 95% interval, so input **.975**. Below our value of t(.975,df=3) = 3.18245.



Inverse Cumulative Distribution Function

```
Student's t distribution with 3 DF
P( X <= x )        x
      0.975  3.18245
```

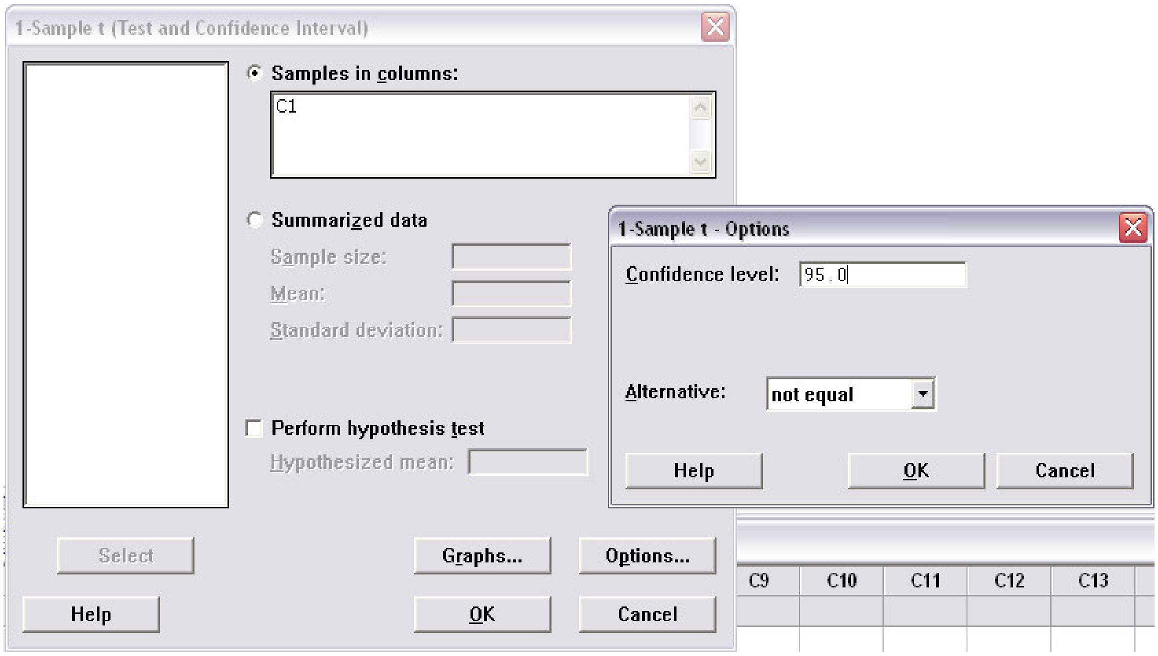
At last we're ready to construct our interval.

$$\bar{y} \pm t_{.025,df} \frac{s}{\sqrt{n}}$$, where

5.00 ± 3.18245*1.08 gives an interval **(1.56257, 8.43743)**.

Minitab can calculate this interval for you.
Stat/Basic Statistics/1t 1-Sample t for the mean

There are two ways. Either input the data in a column and specify the column name with the data (C1 below), or input the Summarized data (sample size *n*, mean, and standard deviation). Under the options button, input the confidence level you want (in percentage, eg. 95.0 for a 95% CI) with Alternative not equal. Below is the result, note that it is the same as above.



One-Sample T: C1

```
Variable  N     Mean    StDev  SE Mean         95% CI
C1        4  5.00000  2.16025  1.08012  (1.56257, 8.43743)
```

In practice, we will test our assumptions given on page 214. That is, the data is a random sample from a large population, the observations are independent, and the population is normal. We generally assess whether the population is normal by looking at a histogram of the sample and seeing whether it appears normal. Normal probability plots help us here as well, by checking if we see a straight line between the data and normal scores.

Interpreting a Confidence Interval

Before the data is collected, the interval is unknown and is viewed as random since it will depend on the actual sample selected. The population parameter (the mean) is unknown and fixed. The confidence interval is determined once the data is collected and the confidence level (or coefficient, such as 95%) is specified. Hence, different samples give different intervals.

The interpretation of a 95% confidence interval is as follows. If intervals are constructed accordingly, over and over again, 95% of them will contain the true value for the parameter of interest. Note that the interval constructed will either cover the parameter or will not. That is, for a 95% confidence interval, 5% of the intervals will be wrong (not covering the parameter).

For the interval we just constructed, provided our assumptions are met, here is the interpretation. 95% of intervals constructed in this way will contain the true population mean number of pages for labs for Stat 538.

Test your understanding

Now, repeat this exercise with these data. In 5 corner-to-corner shots that I missed while practicing pool, with the object ball in the center of the table, these are the number of centimeters from the target pocket that I missed. For example, I missed the first shot 5 cm to the left, and the second shot 8 cm to the right.

-5 8 3 -1 1

Construct a 90% confidence interval for the mean distance that I miss the pocket by when I miss. Use Minitab to calculate the sample statistics, the *t* critical value, then calculate the interval by hand (you can use the calculator in windows from Start/Run… then type "calc", or from the accessories menu). At last, compare the answer you got by hand with the value from Minitab by using the 1 sample t test. Interpret your interval.

Work it out, then check your answer below.

Inverse Cumulative Distribution Function

```
Student's t distribution with 4 DF
P( X <= x )        x
       0.95  2.13185                  (this is the t value)
```

One-Sample T: C1

```
Variable  N     Mean    StDev  SE Mean          90% CI
C1        5  1.20000  4.81664  2.15407  (-3.39214, 5.79214)
```

Interpretation: 95% of intervals constructed in this way will contain the true population mean of distance from the pocket in cm that I miss pool shots by from across the table.

2. Confidence Intervals for Population Proportion *p*

The text provides three confidence intervals for a population proportion, Minitab provides an exact interval also:

1. The first is the one based on $\hat{p}$ in the footnote on page 208. This is the one most often used when the sample size is large enough.
2. The second one is based on $\tilde{p}$ on page 208 with its standard error on page 207. It is the same as the phat one, except that you add 2 to the number of successes and add 4 to the sample size.
3. The third one in the text with the *Z* value we will not use.
4. Minitab provides an exact interval based on the binomial distribution which will always be correct. All of the others are based on an approximation using the normal distribution.

When we do confidence intervals by hand, we use the normal approximation. Because the first one based on $\hat{p}$ is what you will see most often in journals, that is the one we will concentrate.

To make confidence intervals for population proportions, the sample size needs to be large enough. A simple rule of thumb is to check that $np > 5$ and $n(1 - p) > 5$ for the method to be suitable, meaning that you need at least 5 successes and at least 5 failures in your sample. If the population proportion *p* is unknown, you should use the sample estimate in these formulas to check the suitability of the CI. Additionally, the data should be a random sample from a large population where the observations are independent (see page 214 in text).

By hand

We will use interval 1 in the list above detailed in the footnote on page 208. A 95% confidence interval for the population proportion *p* is

$$\hat{p} \pm z_{.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\text{, where}$$

$\hat{p}$ is the sample proportion, the number of successes divided by the sample size.
$z_{.025} = 1.96$ is the critical value of z from the normal distribution with right-tail probability .025 giving a 95% interval. In general, we can use any significance level.
n is the sample size.

Imagine that I ask you all for your eye color, and that 14 of the 21 of you have brown eyes. If we accept that you are a random sample from the population of medical and public health students, then this confidence interval constructed from the sample data will estimate the true population proportion of brown eyes in the population. Construct a 95% confidence interval. Below are the necessary values we need in the above equation.

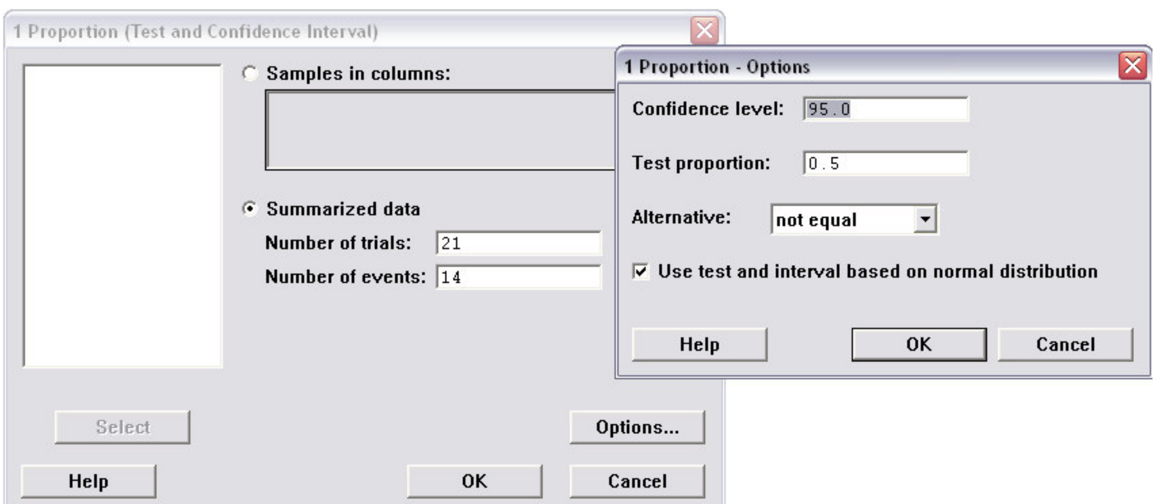
$\hat{p}$ = 14/21 = 2/3 = 0.667
n=21
z=1.96

the 95% CI is $0.667 \pm 1.96\sqrt{\frac{0.667(1-0.667)}{21}}$ giving the interval (0.464, 0.868).

95% of intervals constructed in this way (from samples of this size using the normal approximation) will contain the true value of the population proportion of medical and public health students having brown eyes.

Minitab can calculate this interval for you.
From the menu, select Stat/Basic Statistics/1P 1 Proportion for a proportion.

There are two ways. Either input the data in a column of 0s and 1s (failures and successes) and specify the column name with the data, or input the Summarized data (number of trails n and the number of events (successes)) as I did below. Under the options button, input the confidence level you want (in percentage, eg. 95.0 for a 95% CI) with Alternative not equal. For the normal approximation select "Use test and interval based on normal distribution". Otherwise, it will give the exact interval. Below is the result, note that it is the same as above (except for rounding error).



Test and CI for One Proportion

```
Test of p = 0.5 vs p not = 0.5
Sample   X   N  Sample p        95% CI            Z-Value  P-Value
1       14  21  0.666667  (0.465047, 0.868286)     1.53    0.127
```

Compare the exact interval below (I just unchecked the normal distribution checkbox). As the sample size increases, these two intervals become more and more the same. These are pretty close.

Sample	X	N	Sample p	95% CI	Exact P-Value
1	14	21	0.666667	**(0.430325, 0.854123)**	0.189

To get the $\tilde{p}$ interval in the book, simply add 4 to the sample size and 2 to the number of successes. Therefore, we would make the number of successes 14+2=16 with a number of trials 21+4=25. The resulting interval based on the normal distribution is not much different from the two above.

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	16	25	0.640000	**(0.451843, 0.828157)**	1.40	0.162

Test your understanding

Now, repeat this exercise with these data. In a series of corner-to-corner shots, I potted 34 shots (successes) and missed 16 (failures). Create a 97% confidence interval for the proportion of shots I sink. Use the first type of interval based on $\hat{p}$, and based on the normal distribution.

First calculate the interval by hand, then use Minitab to confirm your answer. Interpret your interval.

Work it out, then check your answer below.

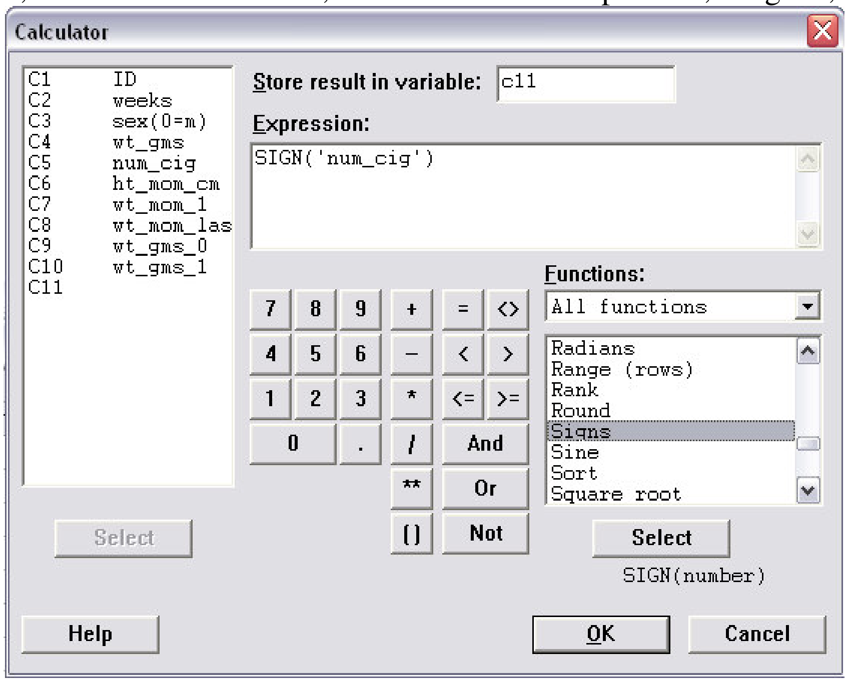
Test and CI for One Proportion

Sample	X	N	Sample p	97% CI	Z-Value	P-Value
1	34	50	0.680000	**(0.536840, 0.823160)**	2.55	0.011

Interpretation: 97% of intervals constructed in this way will contain the true population proportion of balls I pot from across the table.

A hint for problem 8 in homework 4. You are asked to create a variable that indicates whether a mother smoked during pregnancy. We have the number of times she smoked, we need to create a 1 for those who smoked, and a 0 for those who did not. Such a variable is called an *indicator variable* since it simply indicates a presence/absence of a certain condition.

The best way to do this is to use the calculator and the SIGN function. For data that are 0, the new column will 0, but when the data is positive, we get 1, and negative we get -1.



With the new column, calculate from the Stat/Basic Statistics/1 proportion menu a 95% confidence interval from the data in the new column (which I have labeled 'smoke_indicator'). The resulting confidence interval is given below. The interpretation is that 95% of intervals constructed in this way will contain the true population proportion of mothers who smoke during their first pregancy.