Comparing two populations
These notes relate with the chapter 9 of your text book.

AUTHOR: ATUL ROY

We are going to look at comparison of two population means. We shall do very basic level calculations, to understand the idea because in practice mostly such work is done by using computational packages.

Example 1:

A big apartment rental company with many buildings uses a battery of type "A" for their fire alarm. Another brand "B" claims that their batteries will last longer than "A." To test this claim at 1% level of significance, 40 batteries of brand "A" and 42 batteries with brand "B" are tested in similar environment and the results are (time in hours)

	mean	standard deviation	sample size
A	1015	89	40
B	1032	91	42

If μ_A is the overall mean for the life of the brand "A"
and
If μ_B is the overall mean for the life of the brand "B"
H_o: $\mu_A = \mu_B$ $\quad\quad$ $\mu_A - \mu_B = 0$
H_A: $\mu_A < \mu_B$ $\quad\quad$ $\mu_A - \mu_B < 0$

Test statistic:

$$z = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

sample sizes are larger than 30
We may use the sample standard deviation for the population standard deviation
z $= \frac{(1015-1032)-(0)}{\sqrt{\frac{89^2}{40}+\frac{91^2}{42}}} = -0.855\,155\,380\,2$
P_value is the area under the z-curve to the left of $z = -0.86$

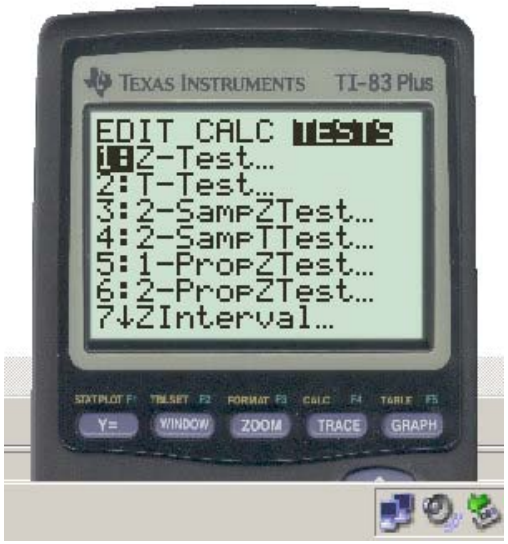
FRom the z-table, this area is
0.1949

P_value is not less than 0.01
Do not reject the null
The results are not significant at 1% level

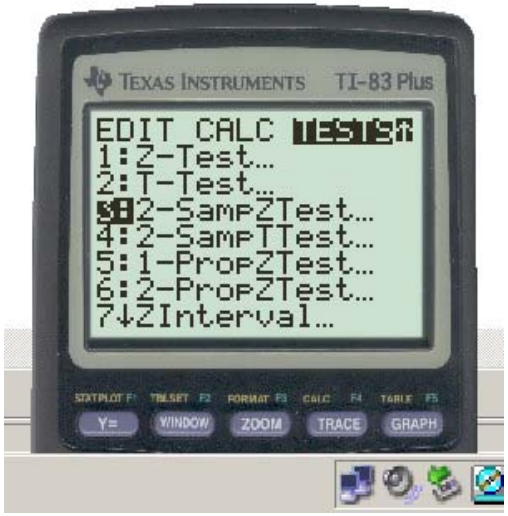
The above is an Example of a Two sample test for two population means,
when the samples are independent.

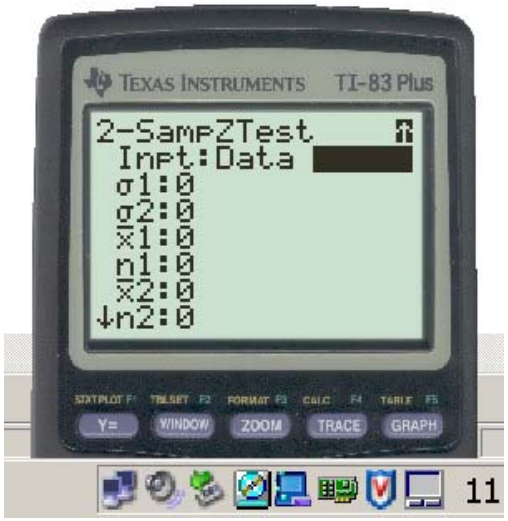
Using a TI-83plus

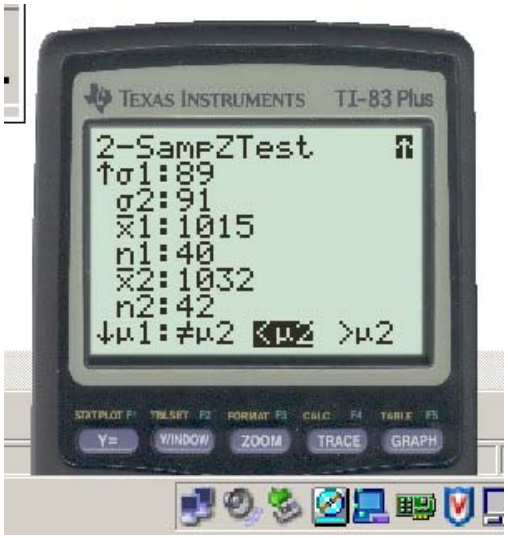
Use STAT and select TESTS


TEXAS INSTRUMENTS TI-83 Plus
EDIT CALC TESTS
1:Z-Test…
2:T-Test…
3:2-SampZTest…
4:2-SampTTest…
5:1-PropZTest…
6:2-PropZTest…
7↓ZInterval…
Y=
WINDOW
ZOOM
TRACE
GRAPH

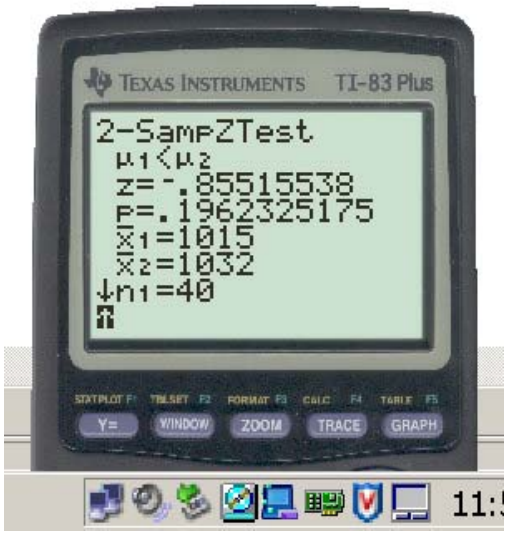
Select


TEXAS INSTRUMENTS TI-83 Plus
EDIT CALC TESTS
1:Z-Test…
2:T-Test…
3:2-SampZTest…
4:2-SampTTest…
5:1-PropZTest…
6:2-PropZTest…
7↓ZInterval…
Y=
WINDOW
ZOOM
TRACE
GRAPH





Ask it to calculate



.....

What if one (or both) of the sample sizes is smaller than 30

May use t-test
under the assumption that both the populations are normal
and that the samples are independent.

Example 2: (This is a corrected version)

A fruit juice bar uses mangoes of brand ”A” to make its Mango Juices.
Another brand ”B” claims that their mangoes will give more pulp.
To test this at 5% level of significance, the pulps from 16 randomly selected magoes of the brand ”A” are weighed and also the pulps from 18 randomly selected mangoes of brand ”B” are weighed.

Here are the results (weights in grams)

	mean	standard deviation	sample size
A	**89.7**	**7.6**	**16**
B	**97.8**	**6.9**	**18**

Assuming a normal distribution for the weights of the mangoes and independent selection of samples

H_o: $\mu_A = \mu_B$ $\quad\quad$ $\mu_A - \mu_B = 0$

H_A: $\mu_A < \mu_B$ $\quad\quad$ $\mu_A - \mu_B < 0$

$t = \dfrac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right)}}$ **the degrees of freedom here is** $n_A + n_B - 2$

(using the methods on the page 377 in the text)

t $= \dfrac{(89.7 - 97.8) - (0)}{\sqrt{\dfrac{(16-1)7.6^2 + (18-1)6.9^2}{16+18-2}\left(\dfrac{1}{16} + \dfrac{1}{18}\right)}} = -3.257\,688\,747\,495\,732\,303$

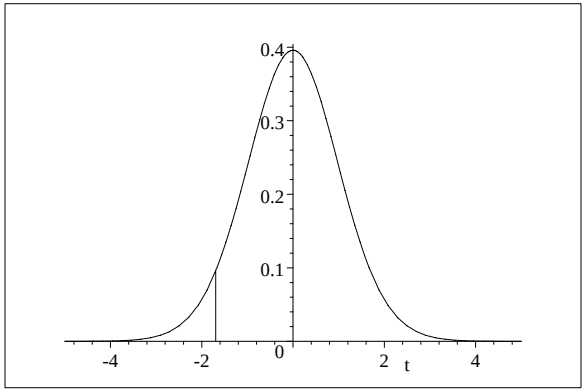
df is $16 + 18 - 2 = 32$

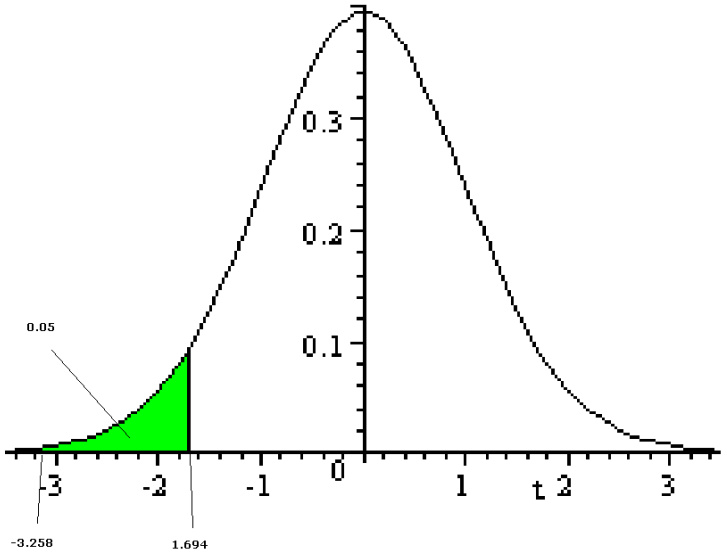
Go to the t-table and note that

$$\begin{array}{ccc} & & .05 \\ & & \downarrow \\ 32 & \rightarrow & 1.694 \end{array}$$

This means that the area under the t-curve to the left of -1.694
at 32 degrees of freedom is 0.05

$s(t)$





The P_value is smaller than the green area which is 0.05, therefore reject the null hypothesis.

Example 3:

Picture: by Atul Roy

Ron runs a landscaping business.

Ron would like to determine if the flowers treated by a fertilizer (that he can he can obtain at a good price) would be taller than the flowers that are left untreated.

Ron divdes 80 plants of such flowers randomly into two groups of 40 each and plants them on identical but separate lots. At the end of the season, Ron measures the heights of the plants in both the lots and finds that

	Mean	Standard Deviation
Treated by the fertilizer	14.5”	1.96”
Untreated by the fertilizer	8.7”	2.12”

To determine if the above sample shows a good evidence that the flowers in the fertilizer group will be taller.

Note that the above two may be treated as independent samples from the populations that grew in the treated and untreated lots.

Population #1 : The fetrilizer or the treatment group

Population #2: The control group

μ_1**: The mean for the population #1**

μ_2**: The mean for the population #2**

$H_o : \mu_1 - \mu_2 = 0$ **or** $(\mu_1 - \mu_2 \leq 0)$

H$_A$: $\mu_1 - \mu_2 > 0$

The test statistic:

Large samples

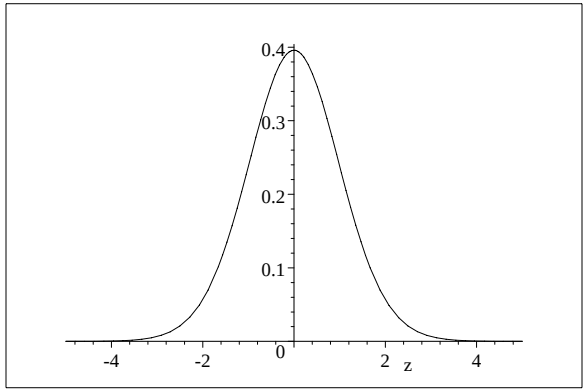
$$\mathbf{z} = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

or

$$\mathbf{z} = \frac{(14.5 - 8.7) - (0)}{\sqrt{\frac{1.96^2}{40} + \frac{2.12^2}{40}}} = \mathbf{12.705}$$

The P_value is the area to the right of z=12.705 under the z-curve

$s(z)$



The P_value is approximately 0.

or

$$\int_{12.705}^{\infty} s(z)dz = 2.7735 \times 10^{-37}$$

The P_value is very small, reject the null hypothesis.

See a good evidence that the plants using the treatment are taller.

Example 4:

To estimate difference between the reading comprehension levels of the 8th graders from two large counties a standardized exam is administered

to two randomly selected group of students from these two counties (1 and 2.) Here are the results

	mean	standard deviation	sample size
County 1	858	49	21
County 2	949	51	25

To compute a 95% confidence interval for the difference between the mean scores for such an assessment test

for these two counties.
We pretending as if all the students are taking this test.

Estimate: $\overline{x_2} - \overline{x_1} = 949 - 858 = 91$

For the margin of error

m=$t_{.025}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$ **,** $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
degrees of freedom is $n_1 + n_2 - 2$
21 $+$ **25** $-$ **2** $=$ **44**

$\mathbf{s}_p^2 = \frac{(21-1)49^2+(25-1)51^2}{21+25-2} =$ **2510. 090 909**

.025

↓

44 → **2. 015**

m=$2.015\sqrt{\frac{2510.090\,909}{21} + \frac{2510.090\,909}{25}} = 29.882\,645\,31$

a 95% confidence interval is

$(91 - 29.89, 91 + 29.89) = (61.11, 120.89)$

Remember that we assumed normal populations here

Example 5:

Example of a Matched Pair Design for comparing Two Population Means

This procedure is also called testing with Related Samples

A faculty training program claims that it will improve the level of student perception of faculty. A big private university that employs 100s of adjunct faculty, randomly selects 31 faculty members and send them to this training program. The student ratings of these faculty members is taken before and after they participate in such a training. We would like to test whether the mean rating after the program is higher than the mean rating before the program.

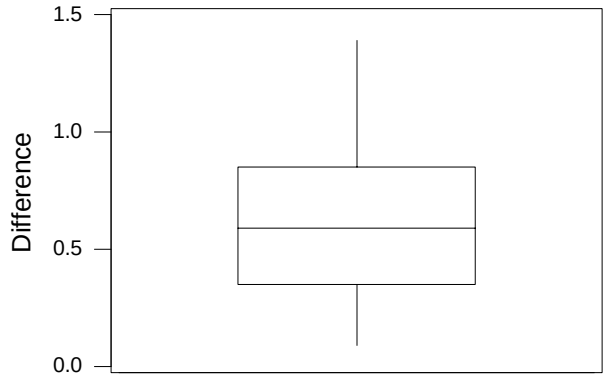
Here are the data

Before	After	Difference
7.04	7.89	0.85
5.76	6.51	0.75
7.75	8.18	0.43
9.1	9.49	0.39
9	9.41	0.41
9.7	9.79	0.09
4.57	5.6	1.03
7.12	8.1	0.98
8.86	9.12	0.26
6.01	6.75	0.74
6.53	7.15	0.62
5.22	6.21	0.99
5.01	5.5	0.49
6.15	7.2	1.05
6.42	7	0.58
4.66	5.3	0.64
6.69	7.3	0.61
6.9	7.1	0.2
7.61	8.2	0.59
6.95	7.15	0.2
7	7.2	0.2
6.94	7.31	0.37
9.19	9.56	0.37
7.32	8.2	0.88
7.19	8.12	0.93
6.76	8.15	1.39
9.3	9.6	0.3

The differences that are obtained in the above manner are called ”paired differences.” and the summary statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Before	31	7.234	7.000	7.260	1.452	0.261
After	31	7.826	7.890	7.867	1.302	0.234
Difference	31	0.5926	0.5900	0.5793	0.3186	0.0572

Now that we have a sample that contains 31 differences and that a boxplot of the sample is



which does not show any strong skewness.

Here we are testing (the subscript d stands for difference)

H_o: $\mu_d = 0$
H_A: $\mu_d > 0$

The test statistic is

$$t = \frac{\overline{d} - \mu_d}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

here

$$\mathbf{t} = \frac{.5926 - 0}{\left(\frac{.3186}{\sqrt{31}}\right)} \cong \mathbf{10.36}$$

Such a high value of t means that the P_value is almost 0, therefore the sample shows a good evidence that the mean is higher after the training.

We could also compute a confidence interval for the mean of the differences.

For example, if we would like a 95% confidence interval for the mean of the differences, first we look for the critical value $t_{.025}$ at 30 ($31 - 1 = 30$**) for a 95% confidence interval**

		$.025$
		$\downarrow$
30	$\rightarrow$	**2.0423**

margin of error given by the above sample is

$$\mathbf{2.0423} \times \frac{.3186}{\sqrt{31}} \cong \mathbf{0.1169}$$

an estimate $\overline{d} = .5926$

therefore a 95% confidence interval for the mean of the paired differences is

$$(.5926 - .1169, .5926 + .1169) = (0.4757, 0.7095)$$

.....

Testing for two population proportions:

Independent Samples

Example 5:

There are two neighboring states ”O” and ”K” sharing the same freeway.
”O” has no tolerance for exceeding the speed limit (of course they can not fine each vehicle exceeding the speed limit but they would ticket any body who shows up doing that on the radar.) ”K” is more liberal in this regard.
To test (1% level) if the proportion of drivers exceeding the speed limit is higher in the state ”K” as compared to the state ”O”, a study finds the following data
regarding their observations of randomly spotted vehicles.

	Exceeding	Total
O	213	1002
K	654	1143

$\widehat{p_O} = \frac{213}{1002}$ **Sample Proportion for ”O”**
$\widehat{p_K} = \frac{654}{1143}$ **Sample Proportion for ”K”**

H_o**:** $p_O = p_K$
$H_A : p_O < p_K$

Under the assumption that there is no difference between the two proportions, we can pool the two estimates
$p = \frac{213+654}{1002+1143} = \frac{867}{2145}$ **pooled proportion**

z $= \frac{p_O - p_K}{\sqrt{p(1-p)\left(\frac{1}{n_O}+\frac{1}{n_K}\right)}}$

z $= \frac{\frac{213}{1002} - \frac{654}{1143}}{\sqrt{\frac{867}{2145}\left(1-\frac{867}{2145}\right)\left(\frac{1}{1002}+\frac{1}{1143}\right)}} =$ **−16. 932 446 52**

The P_value is the area under the z-curve to the left of $-16.\,932\,446\,52$

The P_value is less than 0.01
reject the null
We have significant evidence that lower proportion of the drivers in ”O” are exceeding the speed limit as compared to ”K”

Visiting the website

http://www.pfizer.com/download/uspi_zyrtec.pdf

Please read the prescribing information

For children between 6-11 years old

	% feeling Somnolence	Total number in the sample
Zyrtec 10 mg (group1)	**4.2% (the actual number** $.042 \times 215 = 9.03$	**215**
Placebo (group2)	**1.3% (**$.013 \times 309 = 4.017$**)**	**309**

Does this sample show a good evidence at 5% level of significance that the proportion of chidren experiencing somnolence is higher in the medication group?

p_1**: the proportion of chidren experiencing somnolence in the zyrtec 10 mg group**
p_2**: the proportion of chidren experiencing somnolence in the placebo group**

H$_o$: $p_1 - p_2 = 0$
H$_A$: $p_1 - p_2 > 0$
The sample information: $\overline{p_1} = .042$ $\overline{p_2} = .013$

The test statistic:

$$z = \frac{\left(\overline{p_1} - \overline{p_2}\right) - (p_1 - p_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Pooled estimate $\overline{p} = \frac{\overline{p_1} n_1 + \overline{p_2} n_2}{n_1 + n_2}$

(conditions: $\overline{p_1} n_1 \geq 5$**,** $\overline{p_2} n_2 \geq 5$**,** $(1-\overline{p_1}) n_1 \geq 5$**,** $(1-\overline{p_2}) n_2 \geq 5$ **)**

(caution: one of these numbers is only 4, still let us see)

$$\overline{p} = \frac{.042 \times 215 + .013 \times 309}{215 + 309} = 2.489\,885\,496\,183\,206\,106\,87 \times 10^{-2} \cong .025$$

(the pooled estimate under the assumption that there is no difference between the two groups, this gives the

empirical probability that a child in the entire group of 215+309=524 will get somnolence.)

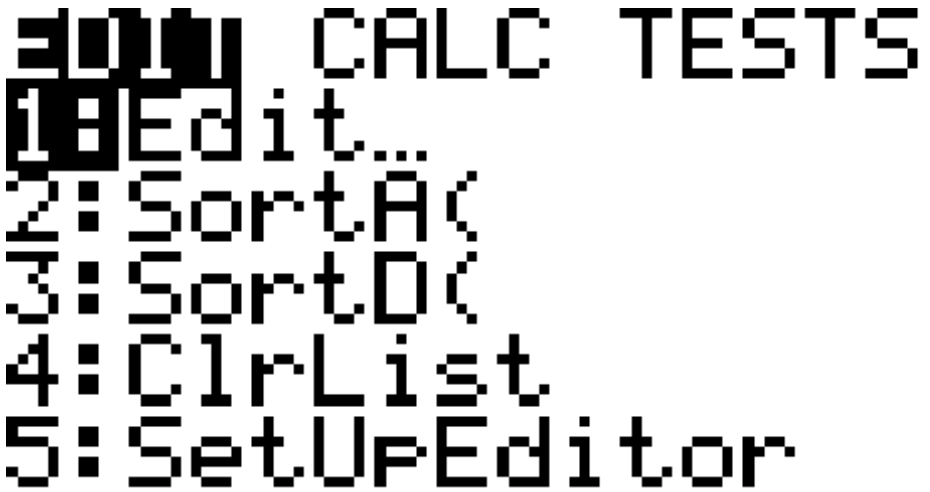
z $= \dfrac{(.042 - .013) - (0)}{\sqrt{.025(1-.025)\left(\frac{1}{215} + \frac{1}{309}\right)}} =$ **2. 091 503 096 479 581 085 42**

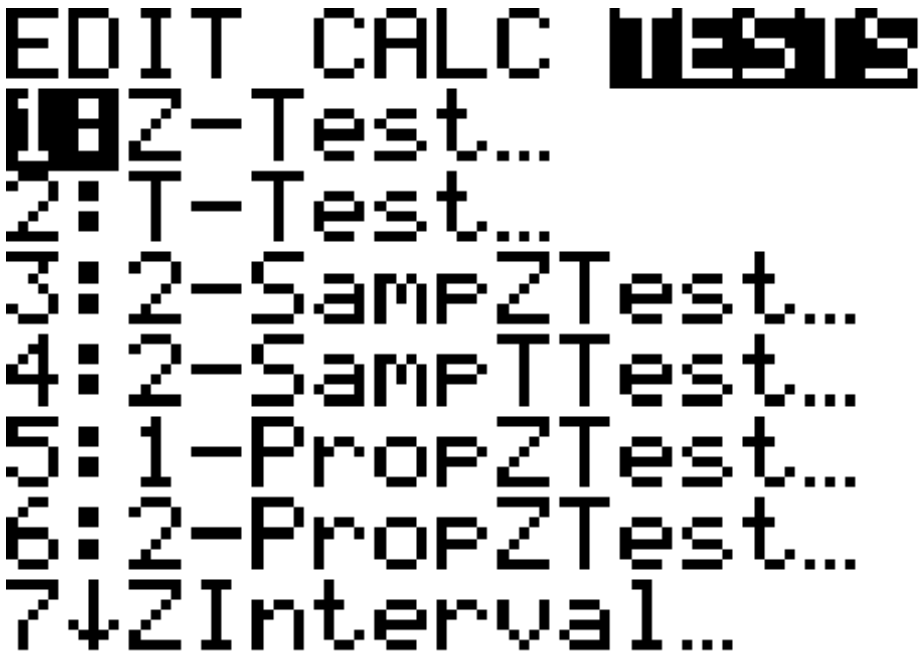
The P_value is the area under the z-curve to the right of z=2.09, which is $.5 - .4817 = 0.0183 < .05$

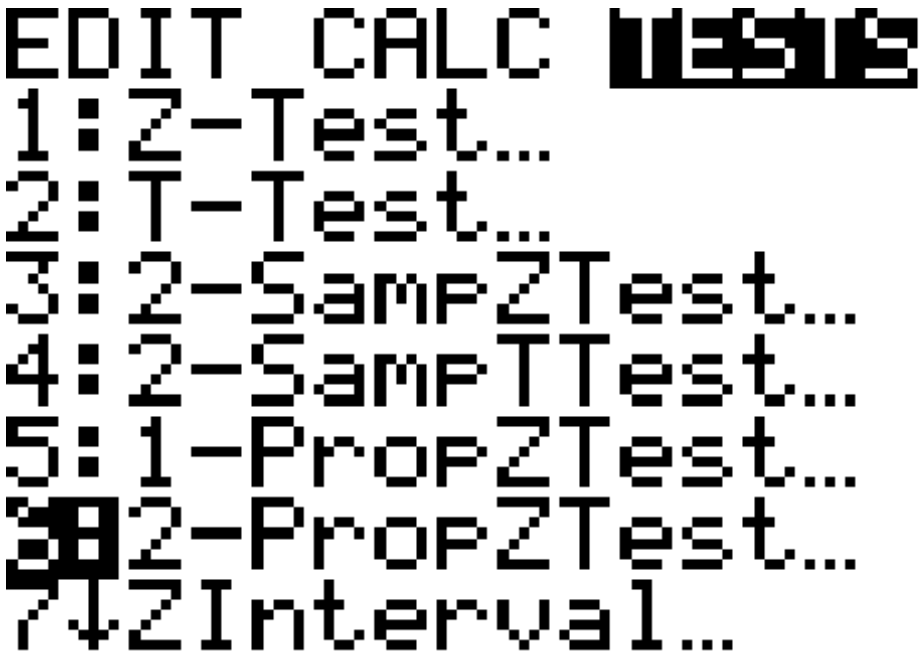
may reject the null but note that one of requirements is not met.

TI83plus

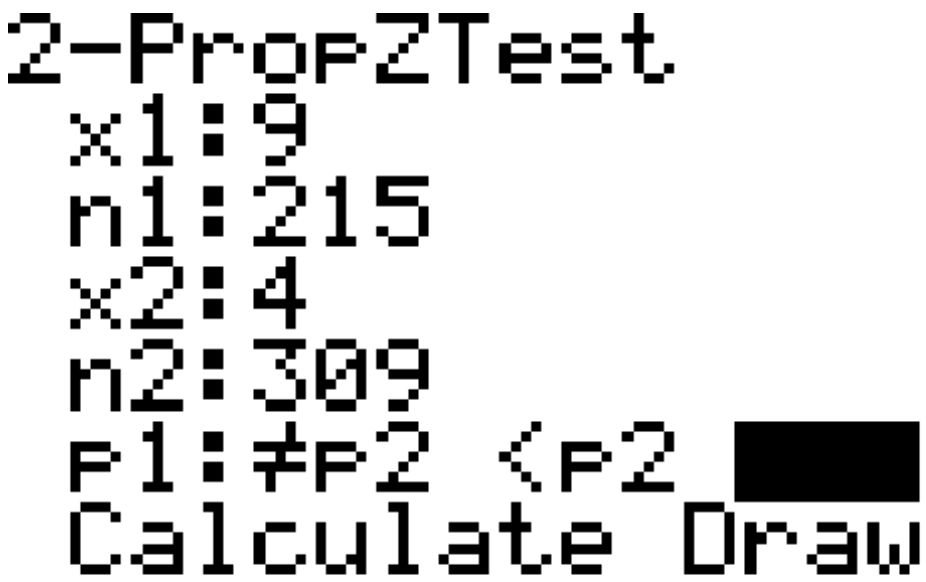
STAT

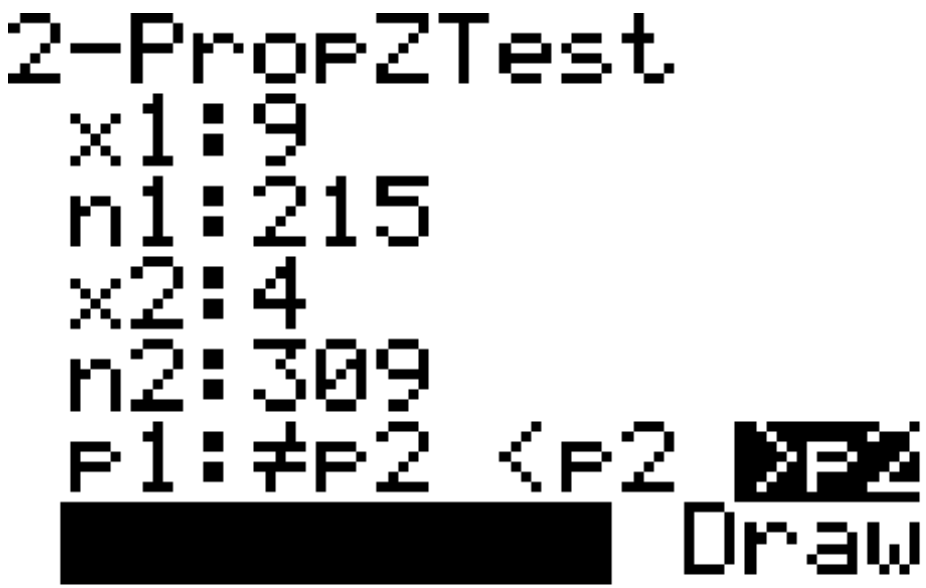

EDIT CALC TESTS
1:Edit…
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor

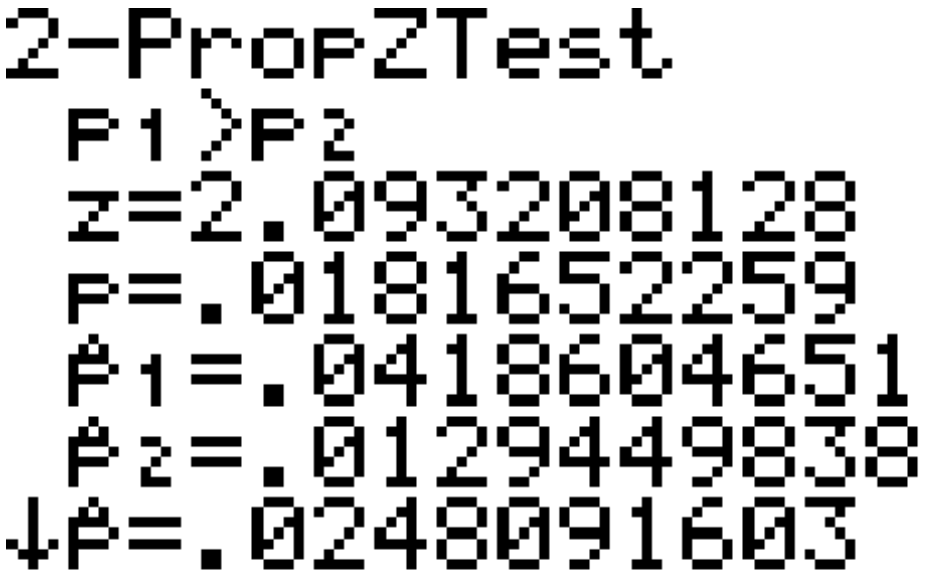

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7↓ZInterval...


EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7↓ZInterval...

ENTER


2-PropZTest
x1:9
n1:215
x2:4
n2:309
p1:≠p2 <p2
Calculate Draw


2-PropZTest
x1:9
n1:215
x2:4
n2:309
p1:≠p2 <p2 >p2
Draw



Again from the same prescribing information:
For people 12 years or older

	% feeling Somnolence	Total number in the sample
Zyrtec 10 mg (group1)	**13.7% (the actual number** $.137 \times 2034 = 278.658$	**2034**
Placebo (group2)	**6.3 (the actual number** $.063 \times 1612) = 101.556$	**1612**

Test if the sample shows a good evidence that the proprtion feeling somnolence is higher in the zyrtec group.

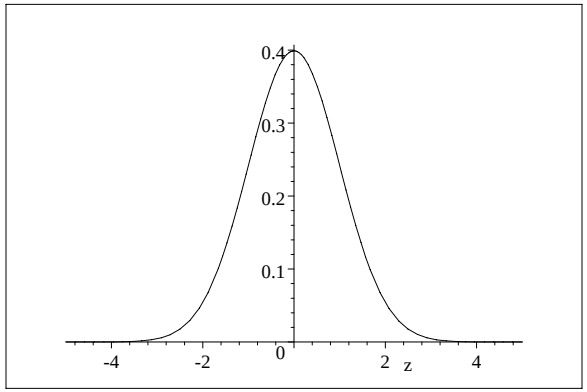
$H_o : p_1 - p_2 = 0$ $\quad p_1 = p_2$

H_A: $p_1 - p_2 > 0$ $\quad$ **$p_1 > p_2$**

$\overline{p} = \frac{278.658 + 101.556}{2034 + 1612} =$ **0.104 282 501 371 365 880 417**

$$z = \frac{.137 - .063}{\sqrt{.1043(1 - .1043)\left(\frac{1}{2034} + \frac{1}{1612}\right)}} = 7.260356386478242793$$

$$\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$



The P_value is the area under the above curve to the right of z=7.26 approximately 0,

$$\int_{7.26}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1.935451618009 6 \times 10^{-13}$$

Reject the null hypothesis

.....

To use this data (older than 12 years) to compute a 95% confidence interval for the difference in the proportions for somnolence between the two groups.

Estimate: $\overline{p_1} - \overline{p_2} = .137 - .063 = 0.074$

Margin of error is

$$\mathbf{z}_{.025}\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1}+\frac{\overline{p_2}(1-\overline{p_2})}{n_2}}$$

$$\mathbf{1.96}\sqrt{\frac{.137(1-.137)}{2034}+\frac{.063(1-.063)}{1612}}=\mathbf{1.907826812665184972}\,\mathbf{19}\times\mathbf{10}^{-2}\cong\mathbf{0.019}$$

a 95% confidence interval is

$$(.074-.019,.074+.019)=(0.055,0.093)$$

.....

Please ignore the following

$$\mathbf{u}(\mathbf{t},\mathbf{n})=\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\mathbf{s}(\mathbf{t})=\mathbf{u}(\mathbf{t},\mathbf{32})$$

...

...

$$\mathbf{u}(\mathbf{t},\mathbf{n})=\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{\pi n}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$$\mathbf{m}(\mathbf{t})=\mathbf{u}(\mathbf{t},\mathbf{24})$$

m(**t**)