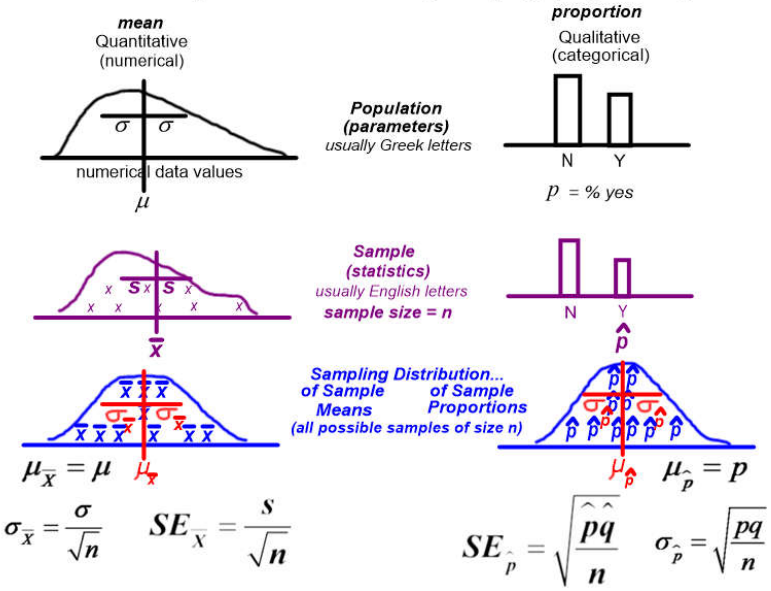
AP Statistics – Lesson Notes - Chapter 23: Inferences about Means

Groups - Work Ch23 Required Practice #1 (first page)...you may want this info...



Your answers are slightly wrong (not your fault - there is something you didn't know)

Let's not use means, let's use proportions first to analyze this situation.

Instead of considering the numerical speeds, let's just ask: what proportion (percentage) of the cars are speeding on this road (% going over 30 mph)?

The percentage speeding in this sample is: $\hat{p} = \frac{5}{10} = 0.5$

Our sample is one of many possible samples, so to say anything about the population of all cars on this road, we need to bound this with a confidence interval (say, 90% confidence level):

$$CI = \hat{p} \pm z^* \sqrt{\frac{pq}{n}}$$

$$= (0.5) \pm (1.645) \sqrt{\frac{(0.5)(1-0.5)}{10}}$$

$$= (0.5) \pm 0.26$$

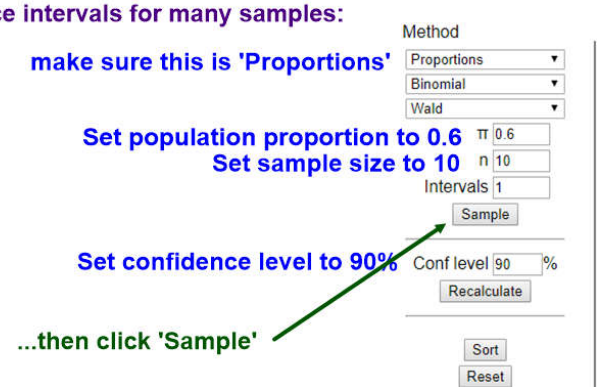
$$= (0.24,\ 0.76)$$

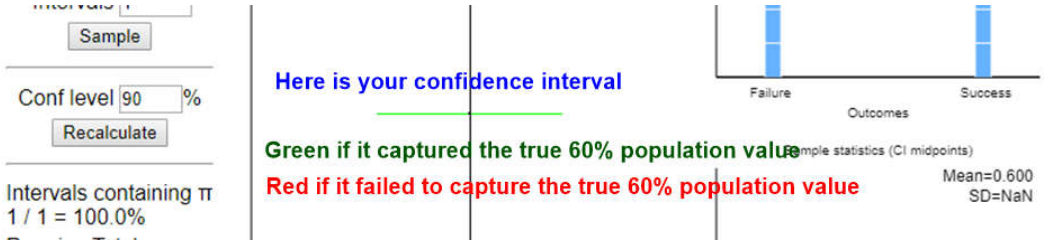
...and there is a 90% chance this interval captures the true population proportion of all cars on this road who speed.

This is just one of many possible samples. Let's see what happens if we took many samples of 10 cars...

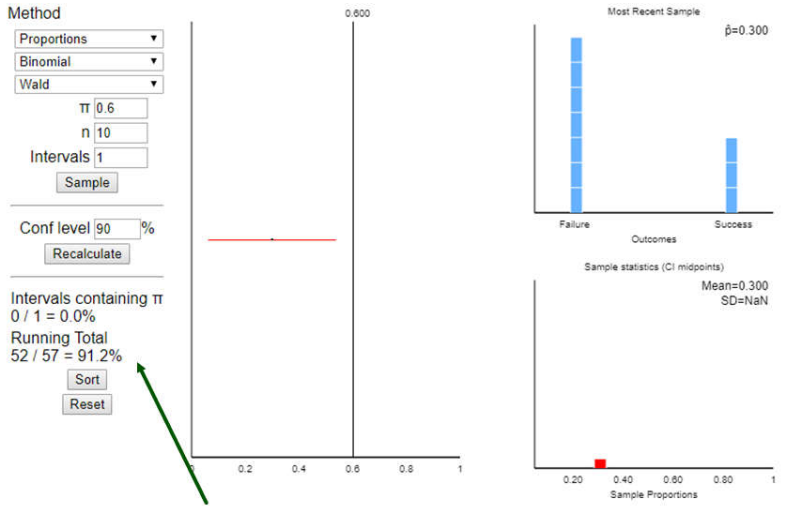
Use your phone to browse to: www.rossmanchance.com/applets/ConfSim.html

We're going to assume the true percentage of cars who speed here is 60% and look at the confidence intervals for many samples:





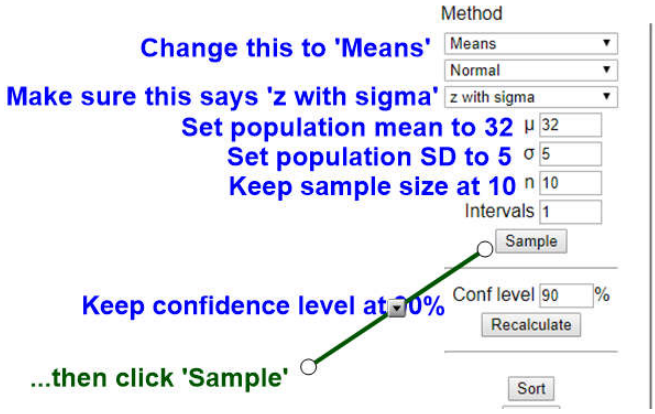
Click 'Sample' many times. Since we set confidence level to 90%, most of these confidence intervals should be green (captured the 60% population value).



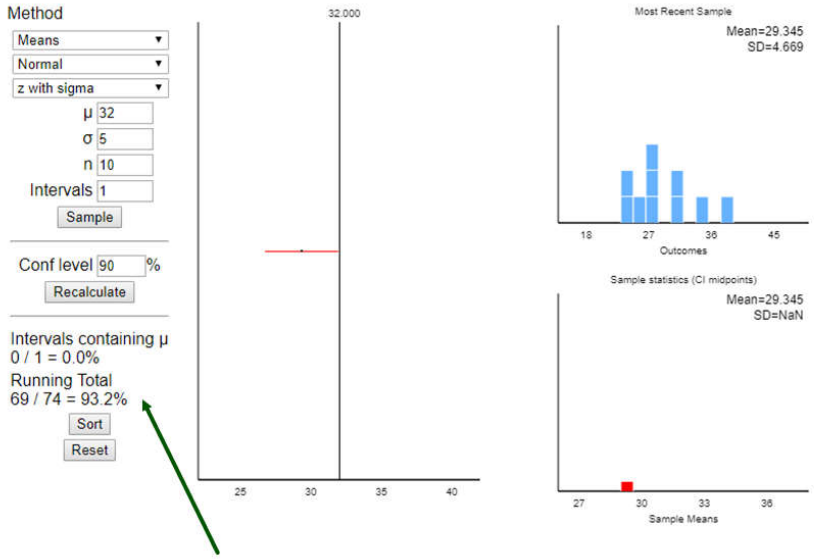
...this shows a running total percentage of all your samples whose confidence intervals captured the population value - pretty close to 90% confidence level

Now let's go back to considering the numerical speed data and the mean speed of cars on the road. Let's assume the population speed for all cars on this road is 32 mph with a standard deviation of 5 mph.

Change the applet set up as follows:

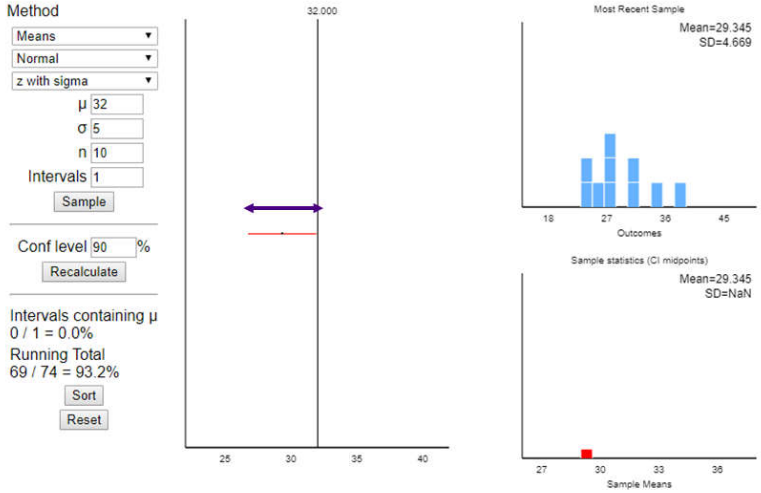


Click 'Sample' many times. Since we set confidence level to 90%, most of these confidence intervals should be green (captured the 32 mph population value).



...this shows a running total percentage of all your samples whose confidence intervals captured the population value - fairly close to 90% confidence level

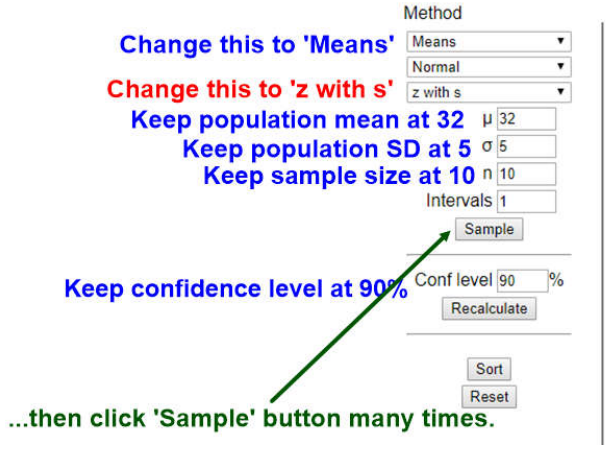
Keep clicking, and notice that the width of the confidence intervals is staying constant



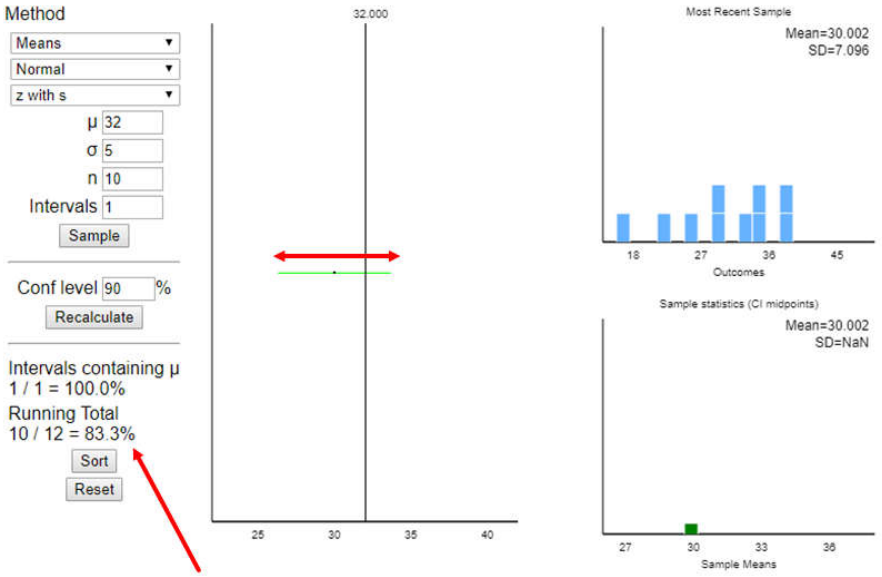
...this is because we are using 'z with sigma' which means we are using the population SD to compute the standard error for the confidence interval.

But this isn't realistic. We actually wouldn't know the true population mean or the standard deviation, so we need to use the sample's SD for the calculation.

Change the applet set up as follows:



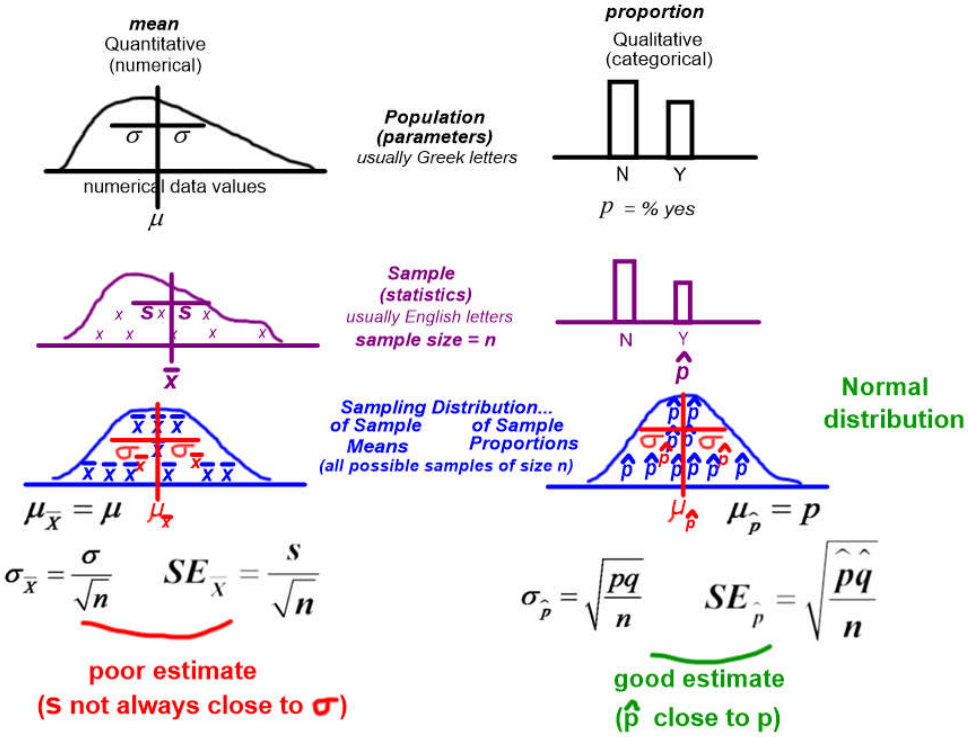
Notice that the confidence interval widths are changing now with each sample



...and the percentage of CIs which capture the true population value is not very close to the 90% confidence level (typically lower percentage).

Why does this happen?

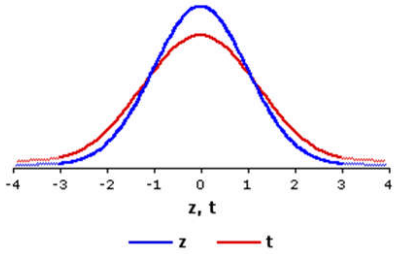
Estimated standard error for means is more variable than for proportions



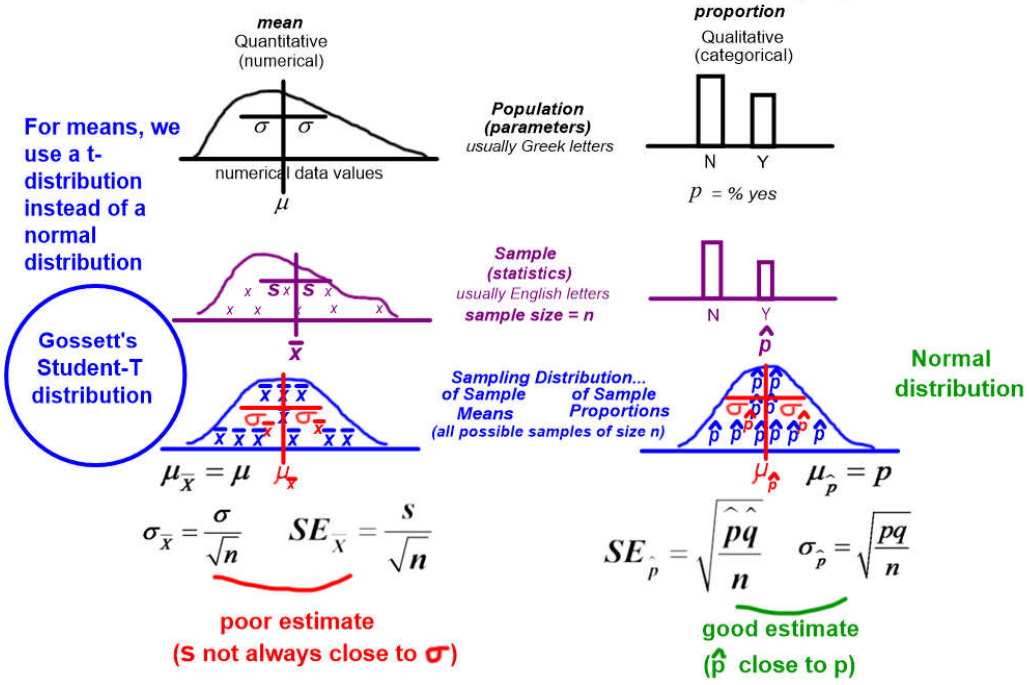
Gosset's Student-t distribution

This was first investigated by William S. Gosset (the quality control engineer at Guinness Brewery in Dublin, Ireland - an interesting story, discussed in our textbook). By using simulations (done by hand, this was before computers were available) he discovered that the sampling distributions for sample means followed a unimodal, symmetric, bell-shaped curve which is shaped similarly to Normal curves, but ***with more probability in the 'tails'***.

The distribution is referred to as Gosset's Student-t distribution (or just t-distribution).



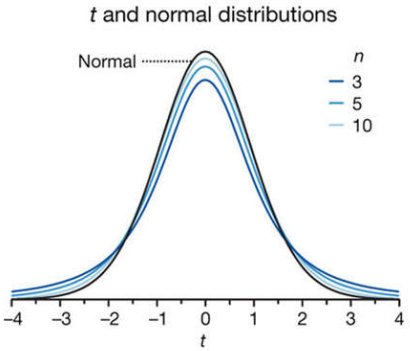
Estimated standard error for means is more variable than for proportions



t-distribution depends upon sample size - Degrees of Freedom

In fact, Gosset discovered that there is an entire family of t-distribution curves that depend on 'degrees of freedom' which is related to sample size.

The t-distribution is 'wider' than the Normal distribution because of the increased variation due to having to estimate the standard deviation of the population with the sample standard deviation. But the higher the sample size, the less likely an individual sample will cause an error in this estimate, so as n increases, the t-distribution becomes more like the Normal distribution:



When we use a t-distribution, we need to specify the appropriate curve according to the degrees-of-freedom (which depends upon the sample size, df=n-1).

Degrees of Freedom

Each data value in the sample is free to vary independently of any other data value. Recall that the definition of standard deviation is:

$$s = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n-1}}$$

Because this calculation includes the mean, if you have, for example, 10 data values, and you are using the mean calculated from these 10 values, then 9 of them can vary independently, but the 10th data value is not really free to vary (in fact, you could calculate it from the mean).

This is why we divide by n-1, not n, in the standard deviation formula for a sample, and is why the number of degrees of freedom (in statistics) is one less than the number of samples:

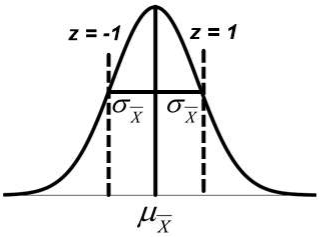
$$degrees\ of\ \ freedom:\ \ df = n-1$$

(for this particular type of data - other data may have different ways to compute *df*)

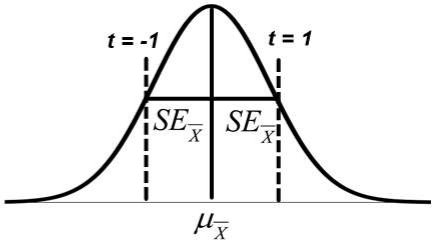
t-value and z-score

For the standardized Normal curve, the z-score gives the number of standard deviations a particular value is above or below the mean. For the t-distribution there is a corresponding **t-value** which represents the number of 'standard errors' a value is above or below the mean:

Normal distribution



t-distribution



...for proportions

$$z-score = \frac{\bar{X} - \mu_{\hat{p}}}{SE_{\hat{p}}}$$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

...for means if you know the population SD

$$z-score = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

...for means if you are using the sample SD

$$t-value = \frac{\bar{X} - \mu_{\bar{X}}}{SE_{\bar{X}}}$$

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Confidence interval for one-sample means (one sample t-interval)

A Confidence Interval for sample means is performed similarly to proportions:

1) Select a confidence level.
2) Verify that the conditions are met:

SRS	n<10%population	'Nearly Normal'
unbiased	probabilities independent	unimodal, approx. symmetrical with no outliers (histogram to check)

3) Determine $n, \bar{x}$ and estimate $SE_{\bar{X}} = \frac{s}{\sqrt{n}}$
4) Determine critical value t*

t* depends upon df=n-1 and is found from a T-table for various confidence levels or, if your calculator supports it, you can use invT(leftarea, df)

5) Calculate margin of error, $ME = t^* SE_{\bar{X}}$

and Confidence Interval, $CI = \bar{X} \pm ME$

6) State your result in the context of the problem.

Inference test for one-sample means (one-sample t-test)

An inference test for sample means is performed similarly to proportions:

1) State the hypotheses: $H_0 : \mu = \mu_0$

$H_A : (1-sided)\ \mu < \mu_0 \ or\ \mu > \mu_0 \ or\ (2-sided)\ \mu \neq \mu_0$

2) Verify that the conditions are met:

SRS	n<10%population	'Nearly Normal'
unbiased	probabilities independent	unimodal, approx. symmetrical with no outliers (histogram to check)

3) Determine $n, \overline{x}$ and estimate $SE_{\overline{x}} = \frac{s}{\sqrt{n}}$

4) Convert the problem's boundary value(s) to t-value(s): $t = \frac{\overline{X} - \mu_0}{SE_{\overline{x}}}$

5) Sketch and find $p - value = tcdf\left(lowerbound, upperbound, df\right)$

6) Decision: is p-value small enough to reject H_0 ?

7) State conclusion in the context of the problem.

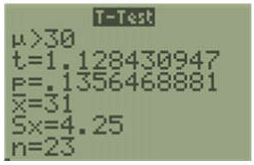
Calculator test functions

Confidence Interval

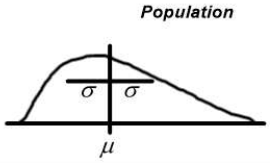
8: TInterval

Inference test

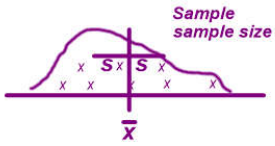
2: T-Test



Checking for 'Nearly Normal' condition - different for population, sample, model



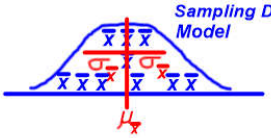
Some populations you can assume are normal (when established by natural processes, e.g. head circumference, arm length)



If sample size is large enough you can assume normal, otherwise must examine the sample. Our textbook's rules:

Must check histogram and Normal Probability plot (NPP) — **n = 15** — **Must check histogram** — **n = 40** — **Can assume nearly normal**

(Should also use a boxplot if you suspect outliers)



If n isn't large enough, you can't do inference without having a probability model (usually from a simulation)

n = 30
np, nq ≥ 10

For means: Normal by the Central Limit Theorem (CLT)

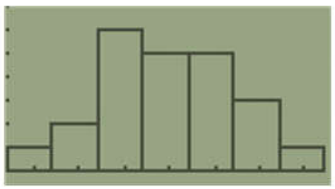
For proportions: by Normal approximation for Binomial for proportions.

Histogram and Normal Probability Plot (NPP) calculator functions

1) Graph histogram:

2nd, Y=, StatPlot1

Zoom, 9:ZoomStat



2) Graph Normal Probability Plot (NPP)

2nd, Y=, StatPlot1

Zoom, 9:ZoomStat