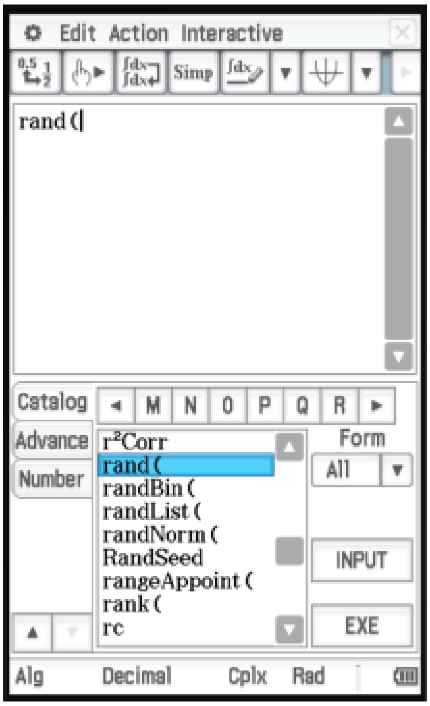
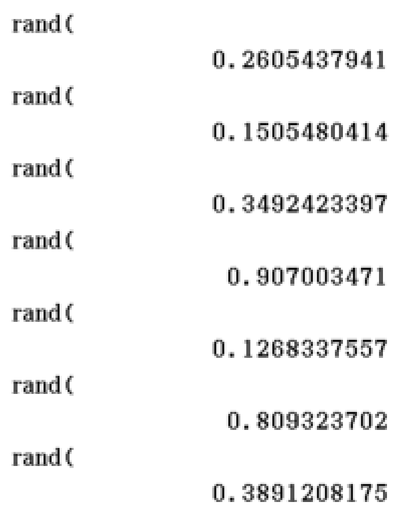
Simulation and Sampling Distributions

Consider the **rand** function on your Classpad, which generates 10-digit random decimal values in the range $0 \leq \mathbf{rand}(\,) < 1$.

This can be input from the **Catalog** of available functions (or typed with the keyboard). Should you wish to access this from the catalog, turn on your keyboard then access the second set of keyboard options using the down arrow key.





Above left is an example of the **Catalog** use to input the **rand()** command, and above right are a series of examples of the command being used.

Sketch a histogram on the following axes illustrating what you think the distribution of values would look like were this command to be used 100 times:

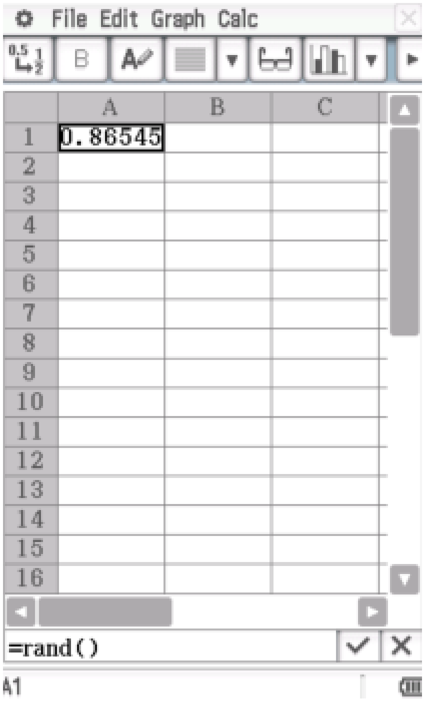
Compare your histogram with those of others around you.

A related command, **randlist(n)**, can be used to generate a list of **n** such random decimal values, a screenshot of which is shown below to generate a list of 100 values:

```
randList(100)
{0.2675262966, 0.21653754▶
```

If we would like to generate histograms of our results, perhaps the easiest and most convenient method is to do so in Spreadsheet mode.

Commands from Main can be input in spreadsheet cells, and indeed it can be easiest sometimes to copy a command from main and then paste it into a spreadsheet cell (the keyboard shortcut [SHIFT] = can be used to copy, and shortcut [SHIFT] y can be used to paste).



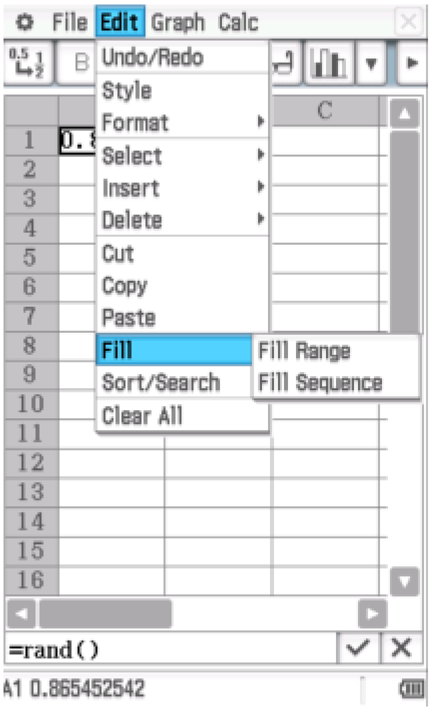
For those familiar with spreadsheets in general, the Classpad spreadsheet operation will feel quite comfortable for you – cell function commands start with the customary = sign, and have similar format including ranges of values.

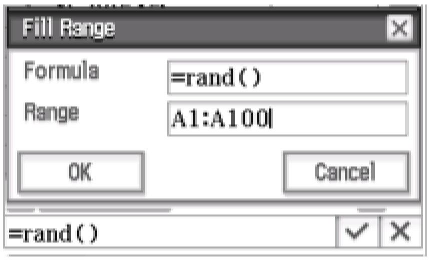
The advantage with the calculator implementation is that Classpad functionality is available to you in cell commands, so a mixture of functions can be used.

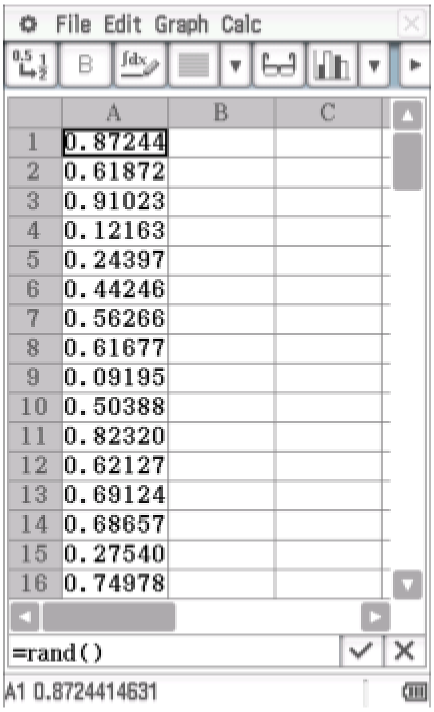
The function bar shows at the bottom, and beneath that a name box.

A command in one cell that is required to be replicated in adjacent cells can be copied via the **Fill** option residing under the **Edit** menu tab.

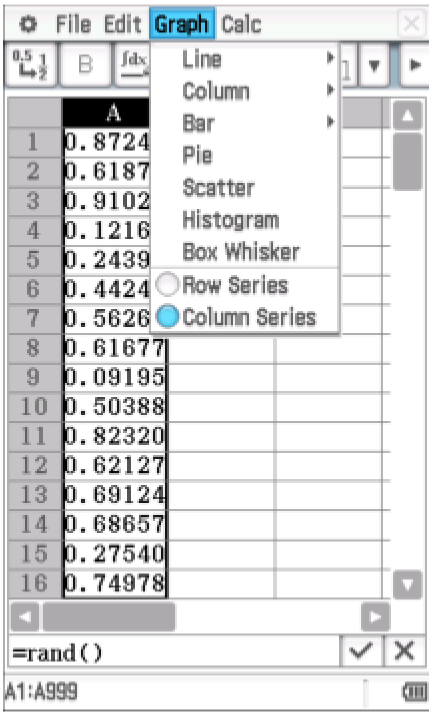
The **Fill Range** option will take a copy of the currently selected cell, and fill it into the designated range of cells. In this case we want to fill 100 cells in the first column with random decimals.







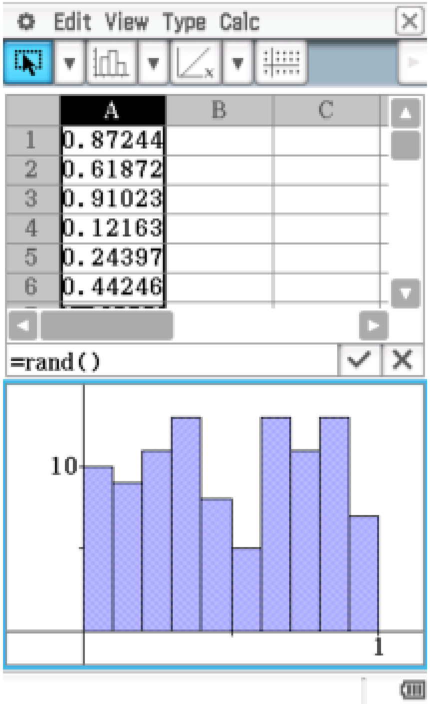
The result of our Fill directive is shown to the left.

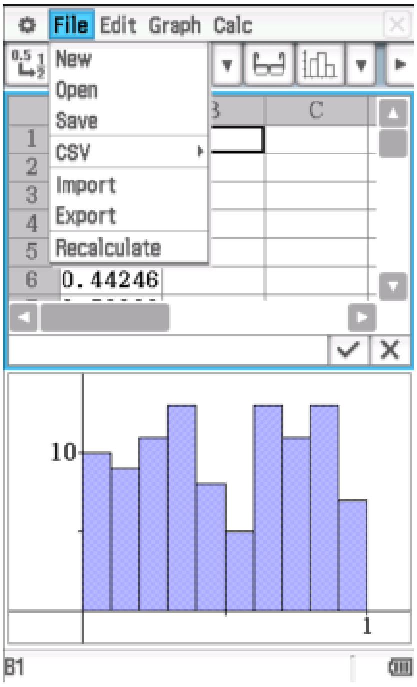


Select the column of values by clicking on the column name tab above it, just as you would ordinarily in a spreadsheet.

Under the **Graph** menu tab, select **Histogram** from the list of options.

The Classpad will now split the screen, and display a histogram of the 100 values.

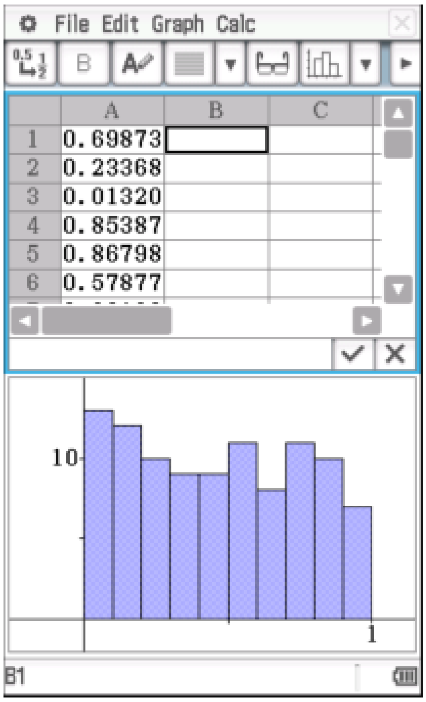


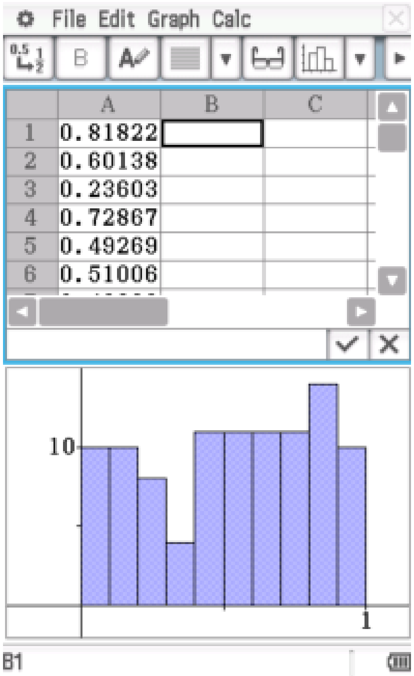


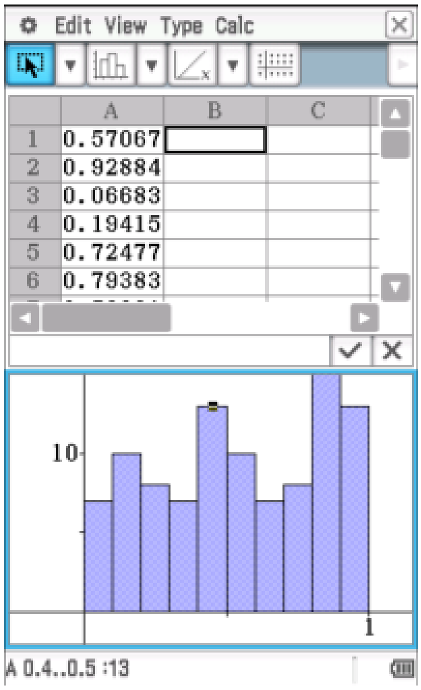
At any time the spreadsheet can be **Recalculate**d using the appropriate option under the File tab as illustrated to the left.

Note that the histogram will automatically adjust.

Three differing versions are illustrated below, illustrating he randomness of the decimal numbers generated.





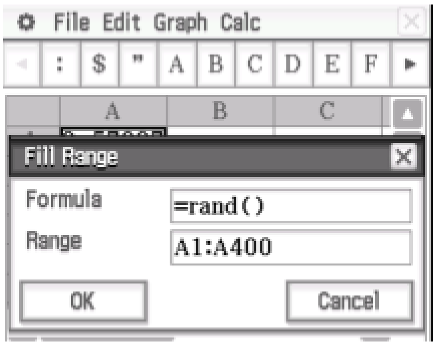


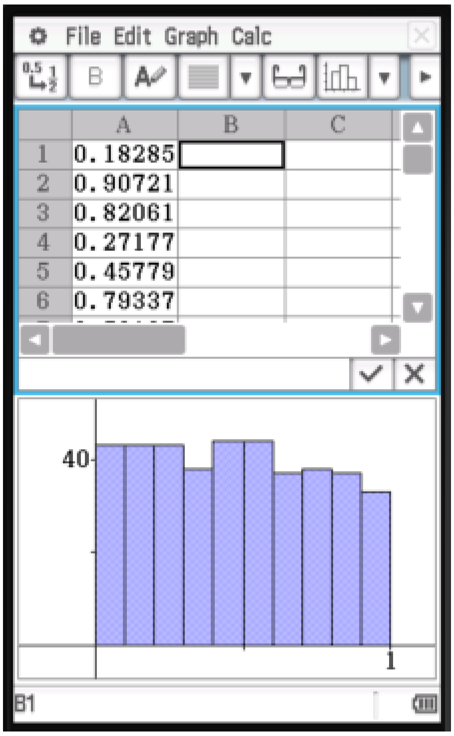
In the final example at right, two important facets are shown:

Firstly, the frequency of any interval (the height of the column) can be given by clicking on the column, after which the Name bar strip below will inform you of the column being graphed, the class interval, and the frequency – in the case to the right, there were 13 values in the range $0.4 \leq x \leq 0.5$

Also note the possibility that a column can extend beyond the top of the window.

Try it now with 400 rolls – compare your histogram with that below.

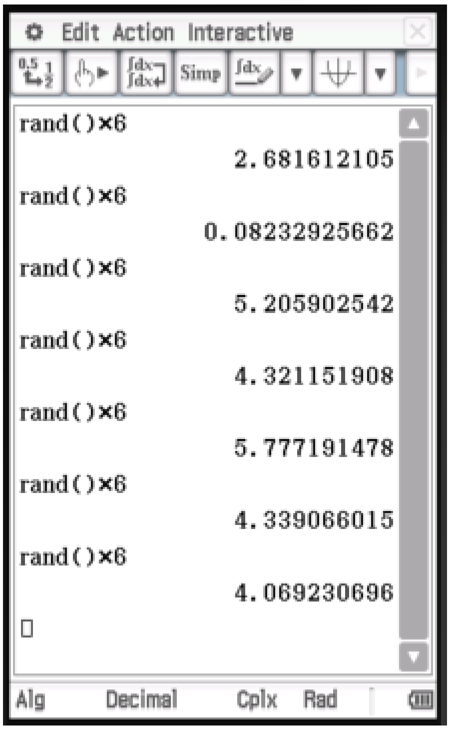




We are witnessing samples of the given size from a uniform distribution

i.e. $\mathbf{rand}(\ ) \sim U(0,1)$

Now consider the command $\mathbf{rand}(\,) \times \mathbf{6}$



What would a histogram of the command $\mathbf{rand}(\,) \times \mathbf{6}$ look like?

Experiment to confirm your suspicions, using the spreadsheet option.

Your results should confirm that $\mathbf{rand}(\ \ ) \times \mathbf{6} \sim U(0,6)$

Now consider the command $\mathbf{int}(\mathbf{rand}(\,) \times \mathbf{6})$

(where the **int** function truncates its input value to provide just the integer portion)

What would you predict its distribution to look like?

Again, experiment to confirm your suspicions (or, perhaps, to correct them!)

Your results should confirm that $\mathbf{int}(\mathbf{rand}(\,) \times \mathbf{6}) \sim U(0,5),\ \ X \in \mathbb{Z}$

Finally consider the command $\mathbf{int}(\mathbf{rand}(\,) \times \mathbf{6}) + \mathbf{1}$

Experiment to confirm your suspicions (or, perhaps, to correct them once more!)

Your results should confirm that $\mathbf{int}(\mathbf{rand}(\,) \times \mathbf{6}) + \mathbf{1} \sim U(1,6),\ \ X \in \mathbb{Z}$

Simulation and Sampling Distributions

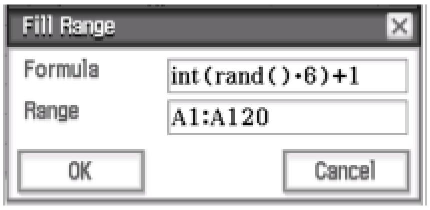
The Fair Die Simulation

We now have the means to simulate rolls of a fair die at our disposal.

There is a command for generating just random integers, **rand(1,6)**, but this is limiting in its applicability to what will follow so we wish to continue to use the manually created version.

Let's use our die simulation command from before in a spreadsheet.

Create 120 rolls of a fair die, and create yourself a histogram of the results.



Let us focus on how many 6's were rolled – 'Sixes are it!'.

Get each person to record for all to see how many 6's were rolled out of the total of 120 rolls – ensure each person gets a chance to write their value on the board: i.e. 19, 17, 24, …….

Let X = 'The number of 6's rolled in 120 rolls of a fair die"

Our aim is to complete the table below:

X	0	…	…	20	…	…	120
$P(X = x)$	≈ 0	…	…	?	…	…	≈ 0

An alternate way to view this, rather than 'How many 6's out of 120 rolls' is to consider the proportion of the total number of rolls, so that ultimately our analysis can be independent of the number of times the trial is repeated:

Let $\hat{P} = \text{the proportion of } 6'\text{s rolled in } 120 \text{ rolls}$

(note that we call this 'P Hat' when reading it aloud)

So $\hat{P} = \frac{X}{120}$ in this case.

We add a further row to our table:

X	0	…	…	20	…	…	120
$\hat{P}$	0	…	…	$1/_6$	…	…	1
$P(\hat{P} = \hat{p})$	≈ 0	…	…	?	…	…	≈ 0

Simulation and Sampling Distributions

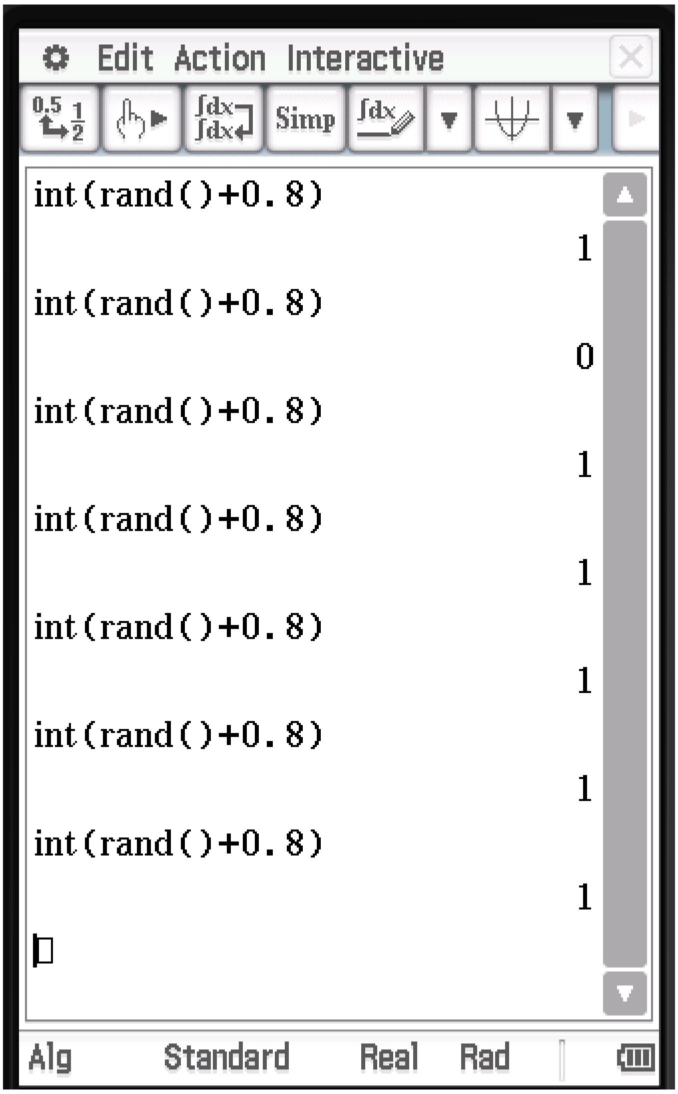
The Bernoulli Simulation

Now consider the result when the command $\mathbf{int(rand()+0.8)}$ is entered.

What will happen?

Write down what you *think* will happen – without fear of being wrong!

Now try it a few times with your Classpad, as displayed in the following screenshot:



Logically, if $\mathbf{rand()} \sim U(0, 1)$

Then $\mathbf{rand()+0.8} \sim U(0.8, 1.8)$

And $\mathbf{int(rand()+0.8)} \sim$

X	0	1
$P(x = X)$	0.2	0.8

And $\mathbf{int}(\mathbf{rand}(\,) + \mathbf{k}) \sim$

X	0	1
$P(x = X)$	1-k	k

for $0 \leq \mathbf{k} \leq 1$

Can you imagine a real context in which this type of command might be used to simulate (or model) the behaviour of the system? Write down an example which occurs to you, and share it with those around you.

Clearly examples are prolific – think of examples which are unusual probabilities – such as the probability of a condom failing being 12% (!). That could be simulated with a $\mathbf{k}$ value of 0.88, so the command returns:

0	(Broken condom)	12% of time on average
1	(Non-broken condom)	88% of time on average

We do tend to think of 1 being success in such probability trials.

Armed with these commands, and an ability to use a spreadsheet to repeat the command many times, the ability to simulate probability scenarios on your Classpad becomes quite powerful.

The Blue Smartie Activity - Proportions

Consider a container in which 10 smarties have been placed, such as in the picture at right, where 6 of the smarties are blue.

Pick one smartie out of the container at random, record its colour, and then return it.

Repeat this 10 times – so that we can investigate the number of times a blue smartie was drawn out of 10 draws.

Let $X = \text{Number of blue in 10 draws}$

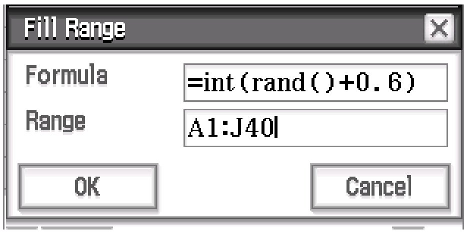
and, let $\hat{P} = \text{the proportion of the 10 draws which were blue}$

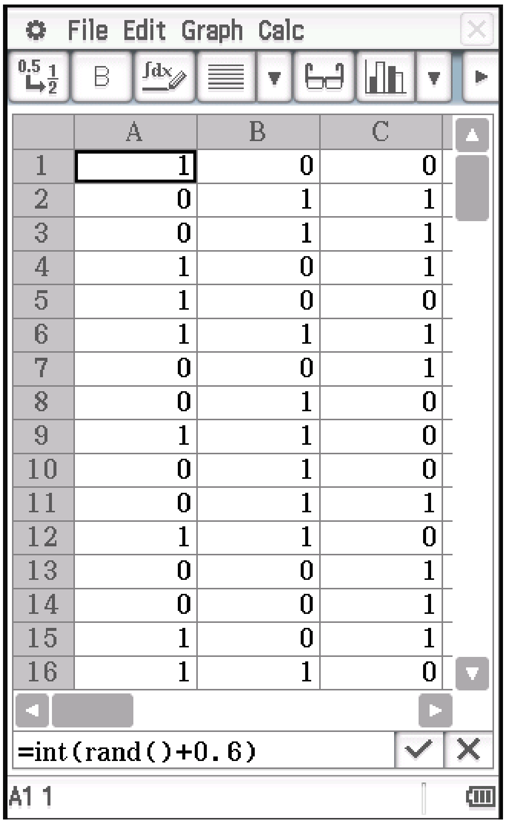
So $\hat{P} = \frac{X}{10}$

X	0	1	2	3	4	5	6	7	8	9	10
$\hat{P}$	0	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	1
$P(\hat{P} = \hat{p})$											

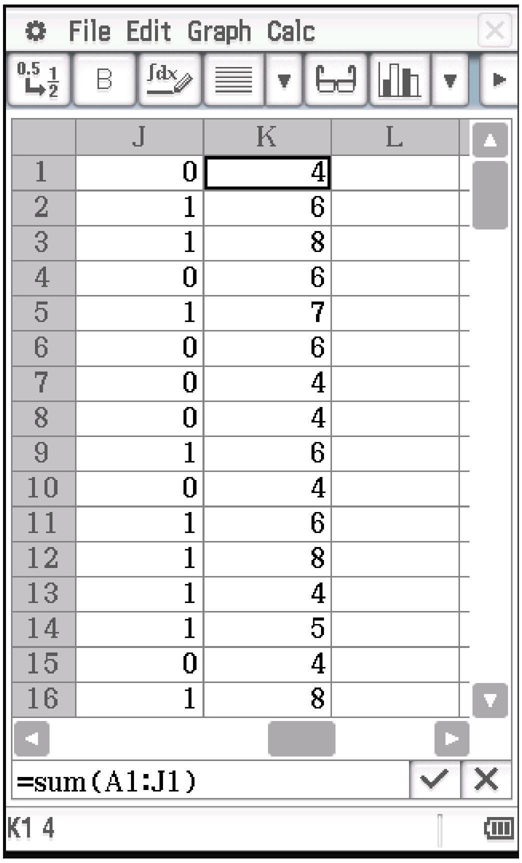
We will use the Classpad to simulate this scenario, taking samples of 10 draws and recording them to help us to better understand the probabilities for the lower part of the table. To do so we will need 10 columns of values. Because we are considering a 1 to be the random selection of a blue smartie, the sum of these 10 vales will give us the number of blue smarties out of the 10 random draws (with replacement).

From this simulation we can obtain experimental probabilities for each outcome (or Event) in our sampling distribution. Later in the session this row will be completed using theoretical probability.

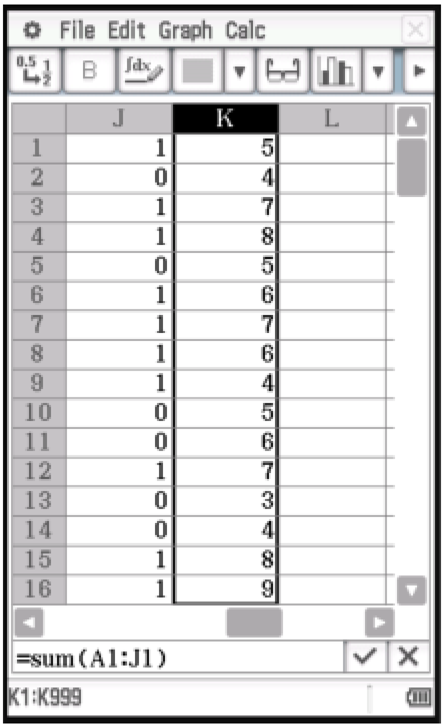


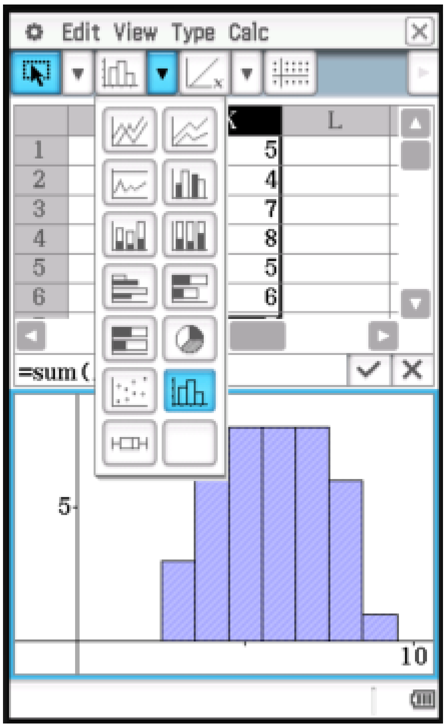


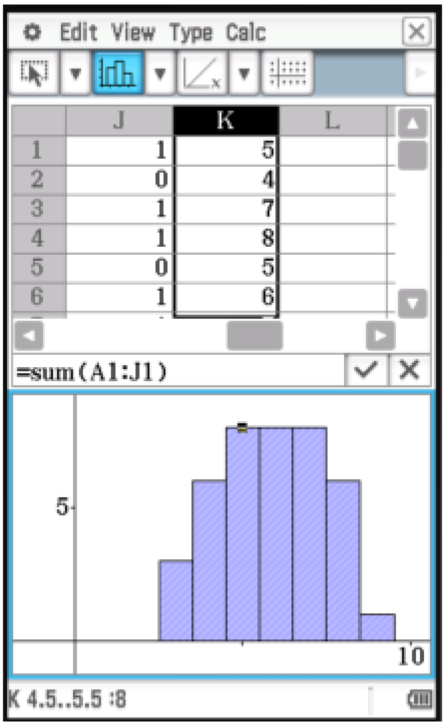
The **Fill** command above creates an array of size 40×10 alongside which a new column, seen below, has been added (and filled down) to sum the 10 values in the row to the left.



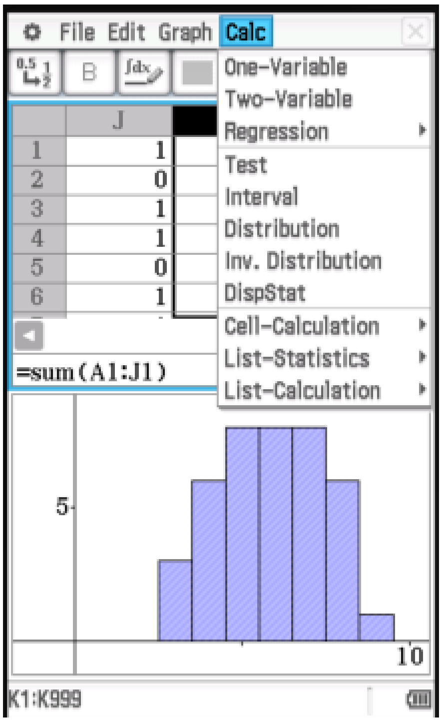
Select the extra column by clicking the heading on top and draw yourself a histogram.

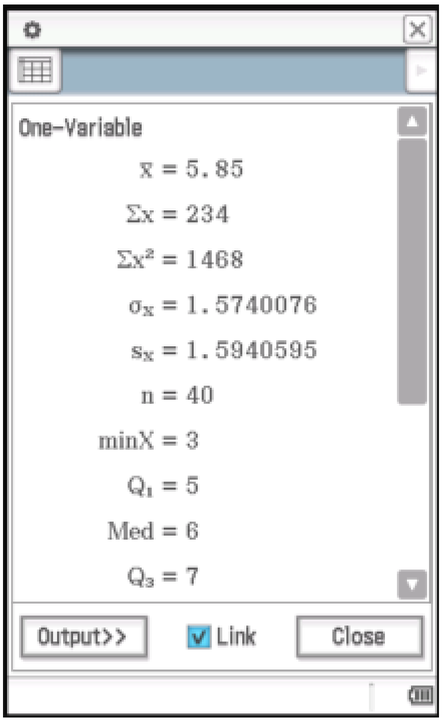






We can optionally calculate the One-Variable statistics for our samples.





If we conduct one trial (i.e. select one smartie from the container and check whether it is Blue or not) then it would be termed a Bernoulli Trial ($n = 1$, where n is the number of trials conducted).

If $n > 1$ then we are conducting a Binomial Trial. In our example above, $n = 10$ as we are repeating the trial 10 times (groups of 10 draws with replacement).

Using Theoretical probability, the probability distribution of X would be a Binomial Distribution, written as

$$X \sim B(10, \frac{6}{10})$$

The Binomial Theorem states that the probability of r successes out of a total of n binomial trials, where the fixed probability of success in each trial is p is given by

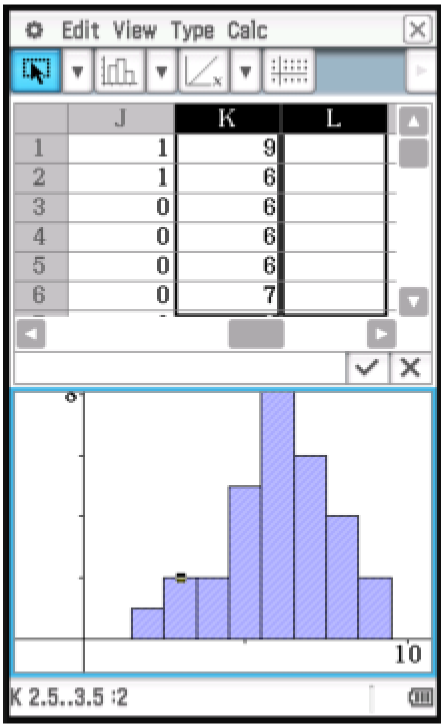
$$P(X = r) = \binom{n}{r} \times p^r \times (1-p)^{n-r}$$

So in our case the theoretical probability of getting 3 blues out of a total of 10 draws, where the probability of a blue in any draw is $\frac{6}{10}$ would be

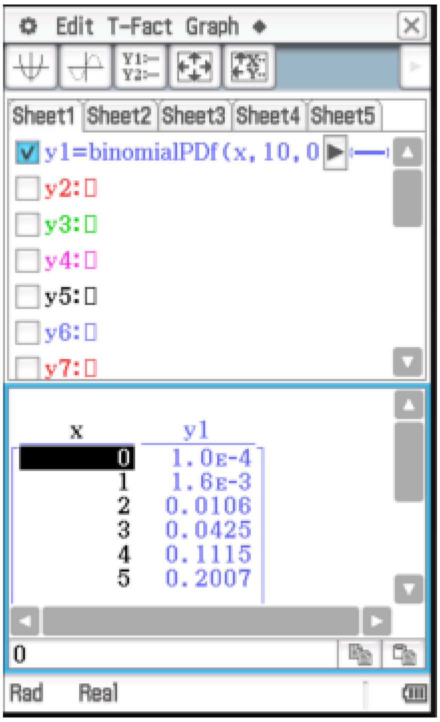
$$P(X = 3) = \binom{10}{3} \times \left(\frac{6}{10}\right)^3 \times \left(\frac{4}{10}\right)^7$$

Which, whilst possible to calculate manually, your Classpad can calculate directly using the command:

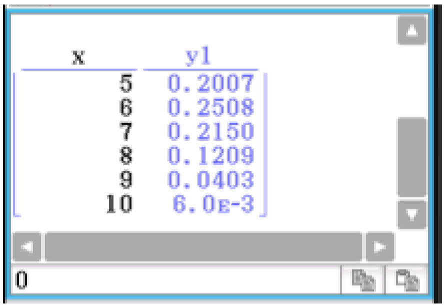
```
binomialPDf(3,10,6/10)
                    0.042467328
```



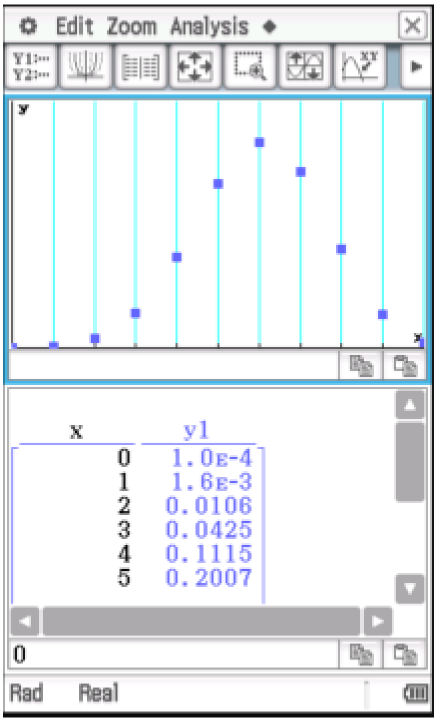
Note that in the spreadsheet example shown at left, there were 2 of the 40 instances where 3 blues were drawn, equating to an experimental probability of 0.05 (not far from the theoretical value of 0.0425 calculated above.



We can create a table of values of all theoretical probability values corresponding to each of the possible values for X, using table mode with integer values of x from 0 to 10:



And then plot these values into a scatterplot, which is termed a probability spike graph.



Note the similarity of the curve produced to a Normal Distribution. In fact, we can approximate our binomial probability distribution curve by a normal distribution curve, such that

$$X \sim B(n,p) \approx N(\,\mu(X) = n \times p\,, \sigma(X) = \sqrt{n \times p \times (1-p)}\,)$$

So that the mean of the sampling distribution will be

$$\mu(X) = n \times p$$

And the standard deviation of the sampling distribution will be

$$\sigma(X) = \sqrt{n \times p \times (1-p)}$$

For our example, with $n = 10$ and $p = \frac{6}{10}$ we have

$$\mu(X) = n \times p$$

$$= 10 \times \frac{6}{10}$$

$$= 6$$

And

$$\sigma(X) = \sqrt{n \times p \times (1-p)}$$

$$= \sqrt{10 \times \frac{6}{10} \times \frac{4}{10}}$$

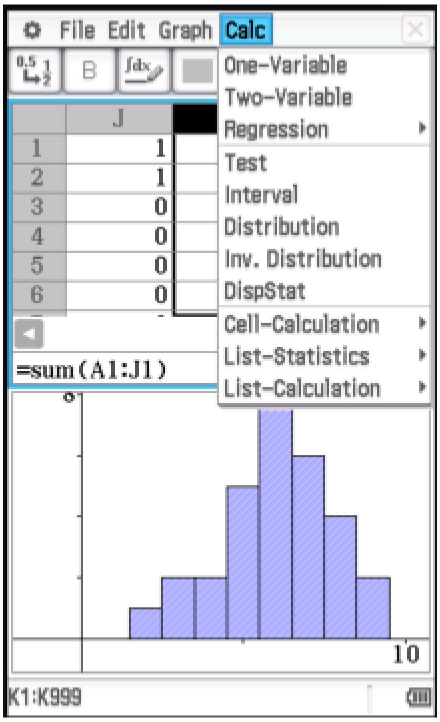
$$= \sqrt{\frac{12}{5}}$$

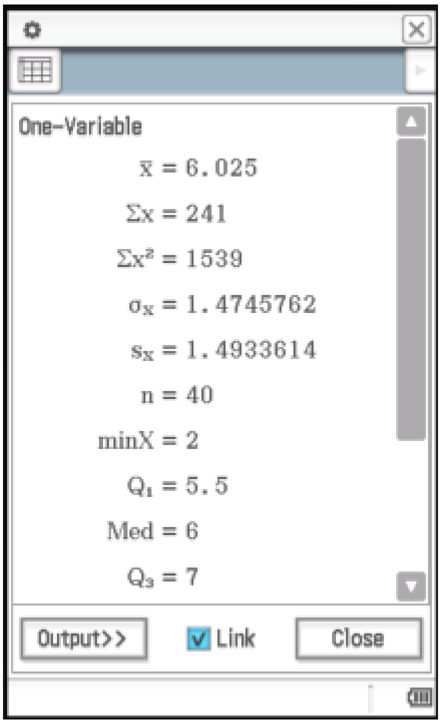
$$\approx 1.549$$

Find the mean and standard deviation of your own sample population, and compare with these values.

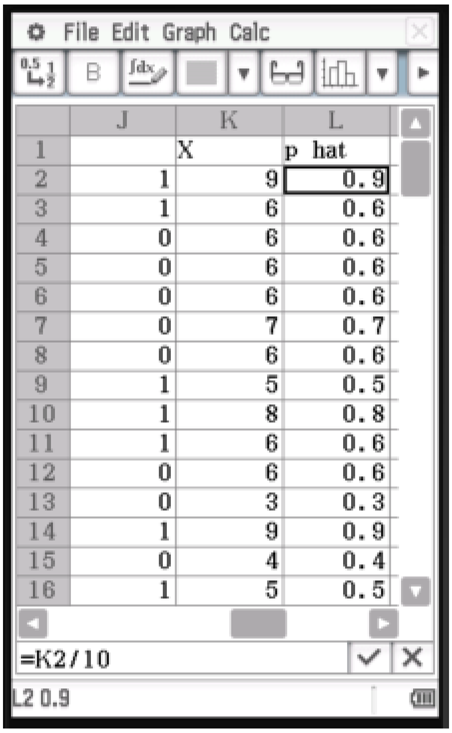
Simulation and Sampling Distributions

Here are the sample statistics for our current sample of X values, with a mean quite near to 6, as we would expect, and a shape reminiscent of a Normal Distribution (although clearly not exact):





We now repeat our earlier technique of the creation of associated $\hat{P}$ values, being $\hat{P} = \frac{X}{10}$ in this case.



These $\hat{P}$ values will also be approximately Normally Distributed, with

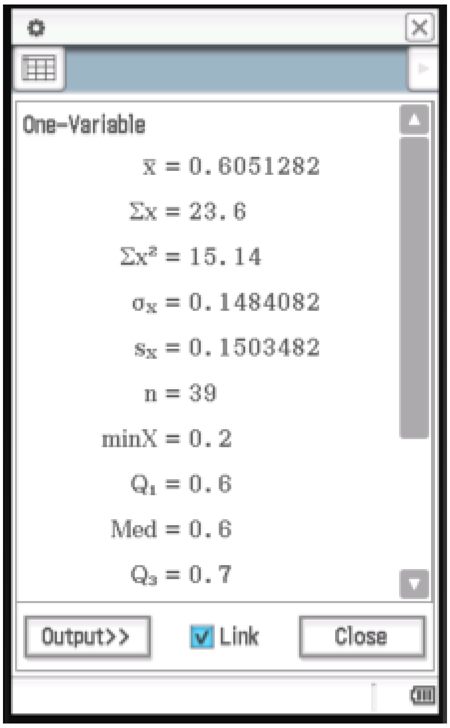
$$\mu(\hat{P}) = p$$

(i.e. the mean of our $\hat{P}$ distribution will be the original theoretical probability p, given that $\frac{6}{10}$ of the smarties in the container were blue).

And

$$\sigma(\hat{P}) = \sqrt{\frac{p \times (1 - p)}{n}}$$

The statistics for our $\hat{P}$ values from our spreadsheet are illustrated alongside, to confirm our findings above.



Simulation and Sampling Distributions

Sampling for Unknown Population Proportions

A more realistic scenario is that the population proportion is unknown, as might be the case were we to do pondering the proportion of blue Smarties in the population of all Smarties produced.

The 250g bag shown alongside contained a total of 232 smarties, of which 28 were blue.

For this sample,

$$\hat{p} = \frac{\#\ blue}{\#\ smarties} = \frac{28}{232} = 0.12$$

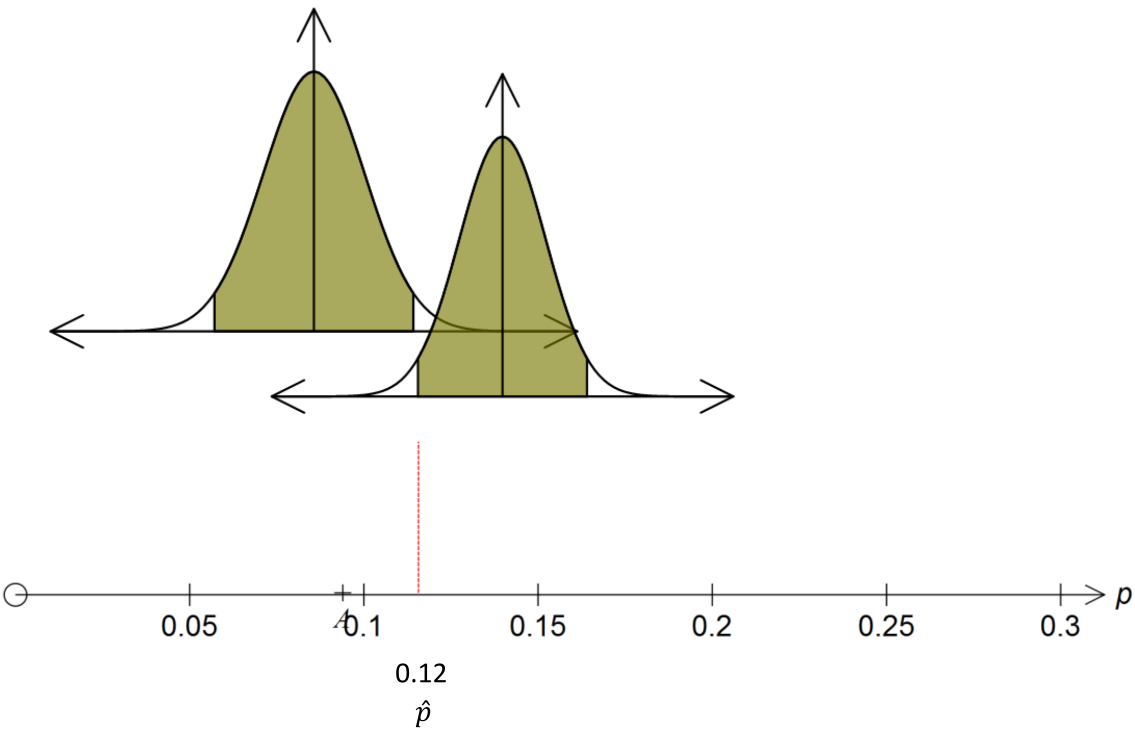
This $\hat{p}$ is one value out of a population of all possible values in the $\hat{P}$ population – the proportions of blue smarties in every 250g bag produced.

So what do we think the actual proportion of blue smarties (p) is?

If we recall and consider that the population of $\hat{P}$ values is approximately normally distributed, then clearly what we have here is one value at random drawn from that population.

Imagine that we have a normal distribution curve sitting above a number line upon which our $\hat{p}$ values are plotted.

Then what we must consider is that this $\hat{p}$ could be anywhere in the normal distribution of values, but if we consider the 'worst case scenarios' of it being in each of the tails (left and right) then we are presented with the following mental picture:



The shading within each distribution curve illustrated is the central 95% of values underneath the curve of the $\hat{P}$ population, which we know to be approximately Normally Distributed. The central 95% will be contained within a band of values stretching out from the mean by a multiple of 1.96 times of the standard deviation (i.e. $(\hat{P}) \pm 1.96 \times \sigma(\hat{P})$).

Using this as a means to understanding where the actual p lies, we come to see that if our $\hat{p}$ is within the central 95% of values under the actual curve centred around p, then we can be 95% confident that the range of values within which is lies is the width of each tail either side of our $\hat{p}$ value.

This allows us to construct from our $\hat{p}$ value an interval within which we are 95% confident that the value of p lies:

$$\hat{p} - 1.96 \times \ \sigma(\hat{P}) \leq \mu(\hat{P}) \leq \hat{p} + 1.96 \times \sigma(\hat{P})$$

We can now substitute our other known values, being that

$$\mu(\hat{P}) = p$$

And

$$\sigma(\hat{P}) = \sqrt{\frac{p \times (1-p)}{n}}$$

Since that actual p value is unknown we use our $\hat{p}$ value in its place.

$$\hat{p} - 1.96 \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} \leq \ p \ \leq \hat{p} + 1.96 \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$$

We are now in a position to construct for ourselves a 95% Confidence Interval for the proportion of blue smarties in the population of all smarties produced, based upon our sample of size 232 which it seems reasonable to assume was randomly chosen.

$$\frac{28}{232} - 1.96 \times \sqrt{\frac{\frac{28}{232} \times \left(1 - \frac{28}{232}\right)}{232}} \leq \ p \ \leq \frac{28}{232} + 1.96 \times \sqrt{\frac{\frac{28}{232} \times \left(1 - \frac{28}{232}\right)}{232}}$$

This yields us an interval within which we are 95% confident that the population proportion of blue smarties lies:

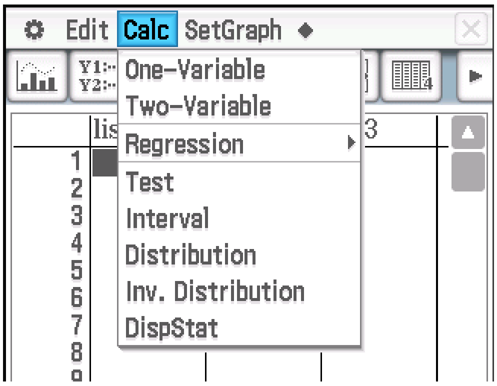
$$0.079 \leq p \leq 0.163$$

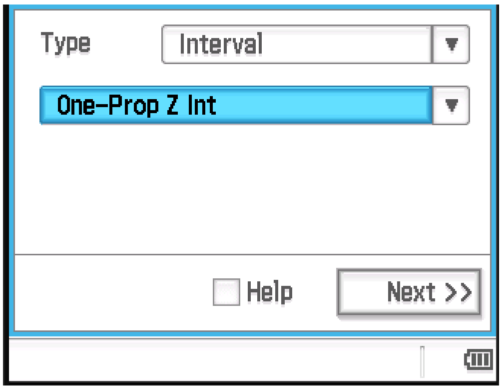
Correct to 3 decimal places.

Confidence Interval Calculations on the Classpad.

Such Confidence Interval calculations are not recommended to be manually entered into Main on a Classpad, although for completeness they have been done below.

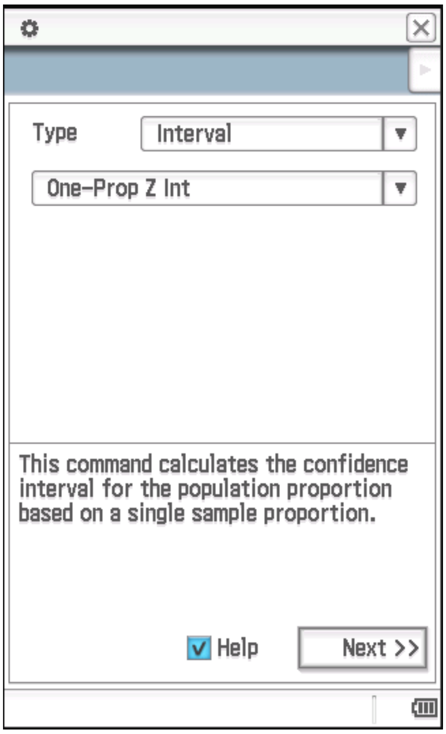
The recommended method is to use the **Interval** wizard provided within the Statistics application:

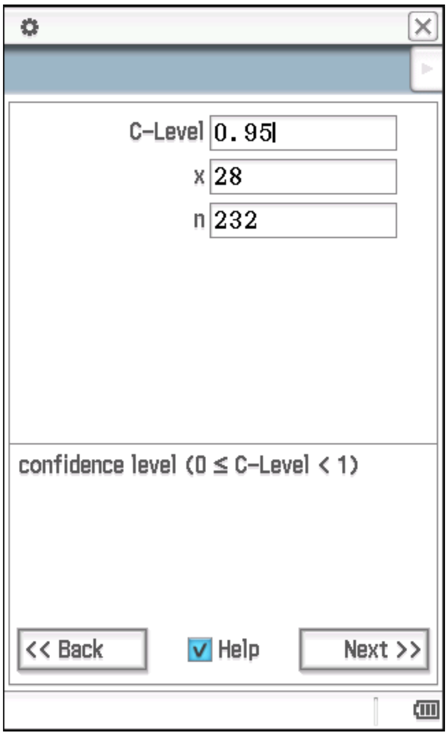




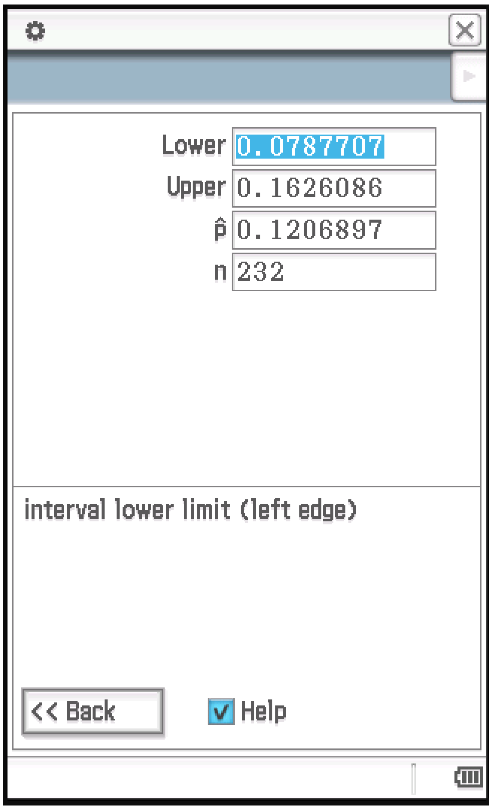
Select Interval, and from the next set of options choose the ‘One Proportion Z-Interval’ from the drop-down list.

You can optionally turn on help within the wizard, which will provide a short description of the contents of each field. For example, the CI percentage must be input as a decimal between 0 and 1.





And we are presented with the boundary values for our confidence interval:



For completeness, the manual calculation is shown also: