EXPLORATIONS

Chapter 9

Sampling Distributions

Topic 19 covers the distribution of sample proportions with a simulation. For a small sample size, a skewed distribution results because the population proportion used is not 0.5. As the sample size increases, the distribution becomes more normal. (This will be very useful when you calculate confidence intervals and test hypotheses about a population proportion from a sample in Topics 22 and 26.) Topic 20 covers the distribution of a sample mean taken from a uniform distribution.

Topic 19—Sampling Distribution of a Sample Proportion (Simulation) and the Normal Distribution as an Approximation to the Binomial

Example: The following exercise simulates a sampling from a very large population with 33% Caucasians and 67% other races. Topic 16 shows that using a die or randBin works very well for this simulation. (See Topic 16, screens 20 to 24 for sample size 5.)

Create a folder named **RACE** and change to that folder.

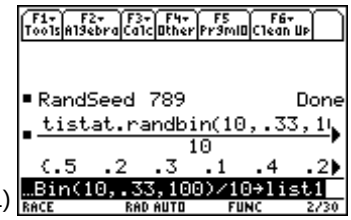
(See Topic 1, *Creating a New Folder* section, and *Changing Folders While in the Stats/List Editor* section.)

Sample Size n = 10, p = 0.33

Extend Topic 16, screen 22, to a sample size of 10 and change to proportions of Caucasians instead of numbers of Caucasians in each sample.

1. From the Home screen, set **RandSeed 789** if you want to repeat these results.
2. Calculate **tistat.randbin(10,.33,100)/10→list1** for a result of **{.5, .2, .3, .1, . . . }** (screen 1).
3. From the Stats/List Editor, set up and define **Plot 1** as Plot Type: **Histogram**, x: **list1**, Hist. Bucket Width: **0.1**, and Use Freq and Categories?: **NO**.

(1)

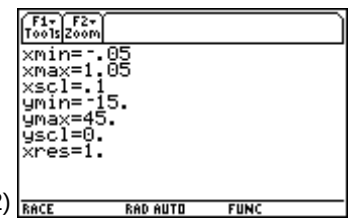


Note: **tistat.randbin** *from* **randBin** *under* CATALOG, F3 **Flash Apps**. *The first random sample of size 10 had five Caucasians, or 50% Caucasian; the second sample 20% Caucasian, etc.*

4. Set up the window using ♦ [WINDOW] with the following entries:
   - **xmin = -.05**
   - **xmax = 1.05**
   - **xscl = .1**
   - **ymin = -15**
   - **ymax = 45**
   - **yscl = 0**
   - **xres = 1**

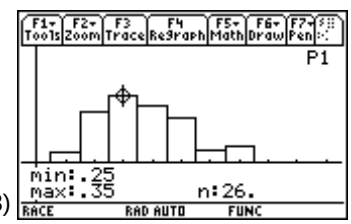
   (See screen 2.)

(2)



5. Press ♦ [GRAPH], F3 **Trace**, and ▶ a few times for the cell that contains **p = .33**, which has 26 of 100 sample proportions (screen 3).

(3)

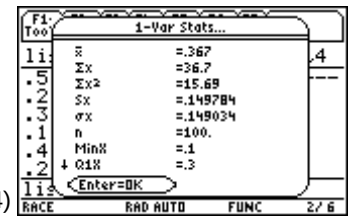


Note: *The distribution is skewed to the right, with no samples with zero Caucasians or with eight, nine, or ten Caucasians. There are primarily samples with two to five Caucasians.*

6. From the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1**, and Freq: **1**, for the results $\bar{x}$ = **0.367 ≈ p = .33**, and s_x = **0.149784 ≈ 0.1498 =** σ_p

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.33*.67}{10}} = 0.1487 \text{ (screen 4).}$$

(4)

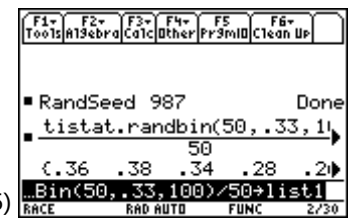


Sample Size n = 50, p = .33

Repeat steps 1 through 5 of the last example, but with:

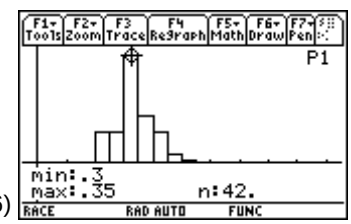
1. From the Home screen, set **RandSeed 987** if you want to repeat these results.
2. Calculate **tistat.randbin(50,.33,100)/50→list1**, change Hist.Bucket Width: **0.05**, with no change in the window (screen 5).

(5)



Results are displayed in screen 6 with 42 of the 100 sample proportions in the cell of half the width as before, but containing **p = .33**. The shape is more symmetric and normal. All but four values are between **0.20** and **0.45**, with **p = 0.33** in the middle of these values.

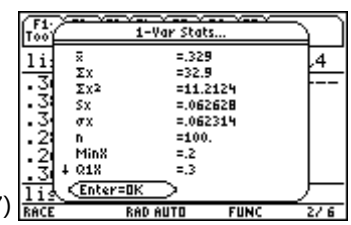
(6)



3. From the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1** and Freq: **1** for the results $\bar{x}$ = **.329 ≈ p = .33**, and s_x = **.062628 ≈ 0.063 =** σ_p

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.33*.67}{50}} = 0.063 \text{ (screen 7).}$$

(7)



***Note**: Even though screen 7 uses σ_x to represent the standard deviation of the proportions, σ_p would be a more meaningful symbol.*

Comparing the Previous Examples

You probably have noticed that as the sample size increases, the distribution of sample proportions appears less spread out and it has more observations close to the population value of **p = 0.33**. σ_p decreases from 0.149 to 0.066. Because the mean of the sample proportions is the same as the population proportion, the sample proportion is said to be an *unbiased estimator* of the population proportion.

The property of getting a more normal distribution with a smaller standard deviation as sample size increases can be explained by the *Central Limit Theorem*. The *Central Limit Theorem* is usually discussed in terms of sample means (see Topic 20). If you consider a success "1" and a failure "0", then the proportion is indeed a mean. (See the last paragraph of Topic 15.)

Checking for Normality with a Normal Probability Plot

Using **list1** of screen 5 on the previous page, check the previous distribution for normality.

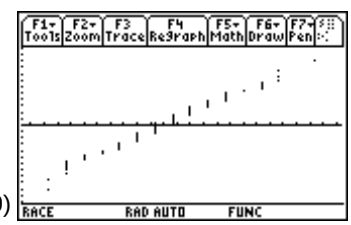
1. From the Stats/List Editor, turn off all functions and plots with F2 **Plots**, **4:FnOff** and F2 **Plots**, **3:PlotsOff**.
2. Press F2 **Plots**, **1:Plot Setup**, highlight **Plot 1**, press F3 CLEAR, and then press ENTER.
3. Press F2 **Plots**, **2:Norm Prob Plot**, with Plot Number: **Plot1**, List: **list1**, Data Axis: **X**, Mark: **Dot**, Store Zscores to: **statvars\zscores** (screen 8).
4. Press ENTER to return to the Stats/List Editor.
5. Press F2 **Plots**, **1:Plot Setup** for the Plot Setup screen.

(8)

6. Press F5 **ZoomData** (screen 9).

The normal probability plot is fairly straight, indicating a normal distribution, but there are groups of points. There are two 0.20's (or 10/50) at the lower left of the plot. (Press F3 **Trace**, and use ▷ to see the coordinates.) There are six 0.22's (or 11/50), sixteen 0.32's (or 16/50), and up to one .50 (or 25/50).

(9)



Note: *There are no .48's.*

Having stacks of points is called *granularity*. In this case, the granularity occurs because between 10/50 and 25/50 there are only 16 possibilities, so with 100 simulations, you are forced to have multiple values. The distribution becomes more normal as both ***n * p*** and ***n * (1 - p)*** become larger, assuring more possible values.

Normal Approximation to the Binomial

The sample size is considered large enough to use a normal distribution to approximate a binomial distribution if ***n * p*** and ***n * (1 - p)*** are both greater than 10 (some texts say 5). From the example above, $n * p = 50 * 0.33 = 16.5 > 10$ and $n * (1 - p) = 50 * 0.67 = 33.5 > 10$.

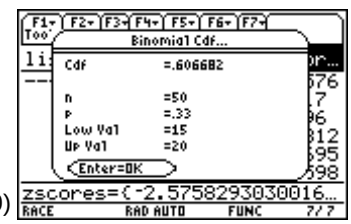
Example: What is the probability that a random sample of size 50 from a population with p = .33 = 33% Caucasians will have from 30 to 40% Caucasians?

30% of 50 = 15 > 10, 40% of 50 = 20 > 10:

1. From the Stats/List Editor, press [F5] **Distr**, **C:BinomialCdf** for inputs: n: **50**, p: **.33**, Low Val: **15**, and Up Val: **20**.

2. Press [ENTER] to display 0.606682 ≈ 61% chance of getting from 30 to 40% Caucasians, which is the true binomial probability (screen 10).

(10)



The area under a normal continuous curve to contain the discrete outcome of 15 must go from 14.5 to 15.5, therefore the approximation area must extend from 14.5 to 20.5. Changing scales by dividing by 50, you need the area from

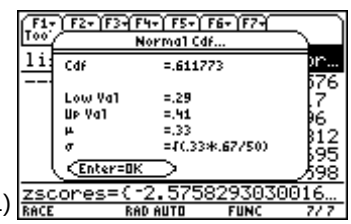
$$\frac{14.5}{50} = .29 \text{ to } \frac{20.5}{50} = .41$$

(this is called the *continuity correction*).

3. Press [F5] **Distr**, **4:NormalCdf**, with Low Val: **.29**, Up Val: **.41**, μ = **.33**, and $\sigma = \sqrt{\frac{.33*.67}{50}}$.

4. Press [ENTER] to display .611773 ≈ 61% (screen 11). This is very good agreement with the binomial result (screen 10).

(11)



If the **NormalCdf** used Low Val: **.30** and Up Val: **.40**, the answer would have been 53%, which is not very accurate. As the sample size increases to n = **500**, p = **.33**, Low Val: **500 * .30 = 150**, and Up Val: **500 * .40 = 200**, the binomial probability = 0.930261 = 93%. The **Normal Cdf** from .30 to .40 with p = **.33**, $\sigma = \sqrt{\frac{.33*.67}{500}}$ gives a probability = 0.922721 ≈ 92%, which is a very good approximation, even without the continuity correction. The larger the sample, the better the approximation.

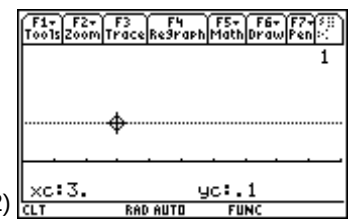
With the continuity correction, the Low Val: **.299** (or 149.5/500) and the Up Val: **.401** (or 200.5/500), with p = **.33**, $\sigma = \sqrt{\frac{.33*.67}{500}}$ gives 0.9294 ≈ 93%.

Topic 20—Sampling Distribution of a Sample Mean and Simulations of the Central Limit Theorem

Example: The distribution of sample means will be simulated from a continuous uniform distribution with all possible values between 0 and 10. See screen 12 with height equal to 0.1 and base 10 for the area under the distribution equal to 1. The mean of this distribution is $\mu = \frac{10+0}{2} = 5$ and the standard deviation is $\sigma = \frac{10-0}{\sqrt{12}} = 2.89$.

Note: *For those who prefer something more hands on, you will also simulate throwing dice and show that the distribution of the means of dots behaves in the same manner.*

(12)



Central Limit Theorem

You will take different sample sizes from the population and show that as the sample size increases, the distribution of the means of the samples becomes more normally distributed with mean = $\mu = 5$ and the standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{2.89}{\sqrt{n}}$.

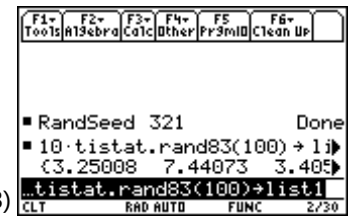
Note: *The sample mean is an unbiased estimator of the population mean.*

Sample Size n = 1

Create a folder named **CLT** and change to that folder. From the Home screen:

1. Set **RandSeed 321**.
2. Type **10*****, then **tistat.rand83(100)→list1**, with **tistat.rand83** from [CATALOG].
3. From the Stats/List Editor, set up and define **Plot 1** as Plot Type: **Histogram**, x: **list1**, and Hist. Bucket Width: **1**.

(13)

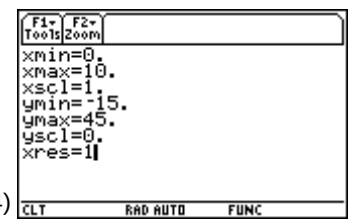


Note: **rand83(100)** *gives 100 random values between 0 and 1. Multiplying by 10 transforms these to 100 values between 0 and 10, starting with 3.25008, 7.44073, 3.40556, 6.22548 (screen 13). (Highlight the output and use ▶ to check the third and fourth values.)*

4. Set up the window using ◆ [WINDOW] with the following entries:
   - **xmin = 0**
   - **xmax = 10**
   - **xscl = 1**
   - **ymin = -15**
   - **ymax = 45**
   - **yscl = 0**
   - **xres = 1**

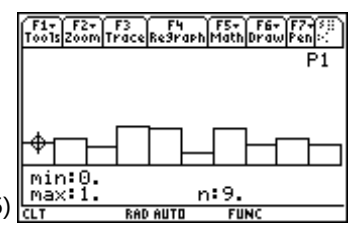
   (See screen 14.)

(14)



5. Press ◆ [GRAPH], and then [F3] **Trace** (screen 15). The first class has nine values.

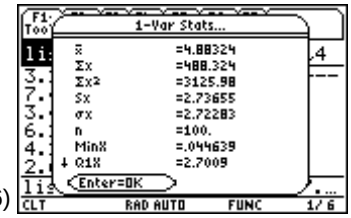
(15)



Note: *The 10 classes have the following frequencies: 9, 10, 7, 15, 14, 5, 14, 8, 10, and 8.*

6. From the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1** and Freq: **1** for the results $\bar{x}$ = **4.88324** ≈ **5** = μ, and s_x = **2.73655** ≈ $\sigma_{\bar{x}} = \frac{2.89}{\sqrt{1}} = 2.89$ (screen 16).

(16)

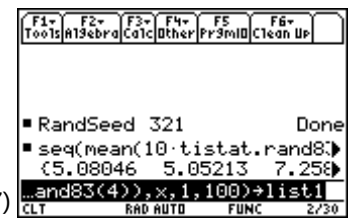


Sample Size n = 4

From the Home screen:

1. Set **RandSeed 321** as explained in Topic 14.
2. To generate 100 sample means of size 4 from your population and store the results in **list1**, enter **seq(mean(10*tistat.rand83(4)),x,1,100)→list1**, with the first value of **5.08046** (screen 17).

(17)

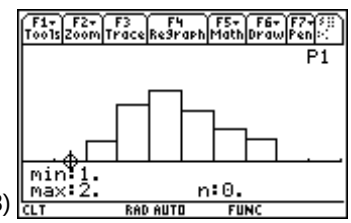


Note: **5.08046 = (3.25008 + 7.44073 + 3.40556 + 6.22548)/4**, *the mean of the first four values generated in step 1 above with sample size* n = **1**.

With the Plot and the Window set up as in steps 4 and 5 on the previous page:

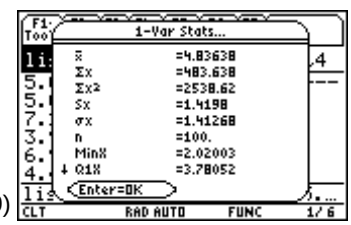
3. Press ◆ [GRAPH], F3 **Trace**, and ▶ (screen 18).
   Previously (screen 15) there were nine values in the first class and 10 in the second, but now there are none. It is very unlikely that all four values in a sample would be so small that the mean is less than two. Samples are more likely to be like the one in step 2. For every small value like **3.25008**, there is probably a larger value like **7.44073** that brings the mean closer to μ = 5.00.

(18)



4. From the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1** and Freq: **1** for results of $\bar{x}$ = **4.83638** ≈ **5** = μ, and s_x = **1.4198** ≈ $\frac{2.89}{\sqrt{4}} = 1.45 = \frac{\sigma}{\sqrt{4}} = \sigma_{\bar{x}}$ (screen 19).

(19)

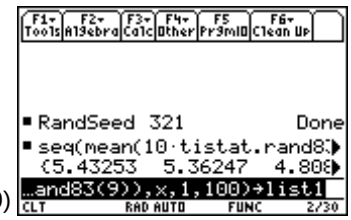


Sample Size n = 9

From the Home screen:

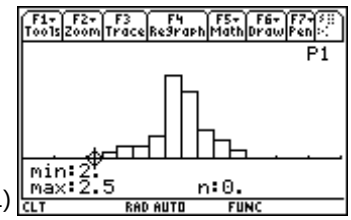
1. Set **RandSeed 321** as explained in Topic 14.

2. To generate 100 sample means of size 9 from your population and store the results in **list1**, enter **seq(mean(10*tistat.rand83(9)),x,1,100)→list1**, with the first value of **5.43253** (screen 20).

(20)


3. From the Stats/List Editor, press F2 **Plots**, **1:Plot Setup** and define **Plot 1** as Plot Type: **Histogram**, x: **list1**, and Hist. Bucket Width: **0.5**.

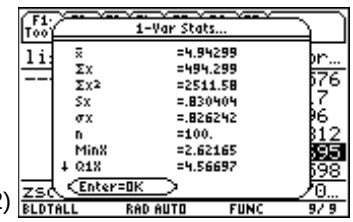
4. With the window set up as above, press ♦ [GRAPH], F3 **Trace**, and ▶ (screen 21).

(21)


5. From the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1** and Freq: **1** for the results $\bar{x}$ = **4.94299** ≈ **5** = μ, and

s_x = **.830404** ≈ $\frac{2.89}{\sqrt{9}} = 0.96 = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{x}}$ (screen 22).

The sample means are squeezed even closer to the population mean 5.

(22)


Check on Normality

n = **9**

1. From the Stats/List Editor, turn off all functions with F2 **Plots**, **4:FnOff**.

2. Press F2 **Plots**, **1:Plot Setup**, highlight **Plot 1**, press F3 [CLEAR], and then press [ENTER].

3. Press F2 **Plots**, **2:Norm Prob Plot**, with Plot Number: **Plot1**, List: **list1**, Data Axis: **X**, Mark: **Dot**, Store Zscores to: **statvars\zscores** (screen 23).

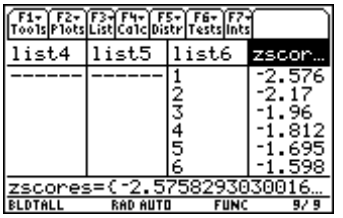
(23)


Note: *If* **Plot 1** *was not cleared in step 1, it could not be used here.*

4. Press ENTER to return to the Stats/List Editor that now has **List zscores** pasted to the end of the list (screen 24).

5. Press F2 **Plots**, **1:Plot Setup** for the Plot Setup screen (not shown) and observe that **Plot 1** has been automatically set up with Plot Type: **Scatter**, Mark: **Dot**, X List: **npplist**, and Y List: **zscores**.

(24)

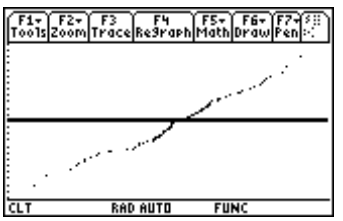


Note: *This is a scatterplot with list* **npplist** *(***list1***) sorted in ascending order. List* **zscores** *is also a list, in order from low to high. If you wish to make a second normal probability plot but need to save the above results, you must store lists* **npplist** *and* **zscores** *to other list names.*

6. Press F5 **ZoomData** to display screen 25.

   The data are close to lying on a straight line, which is easier to eyeball than normality (as described in Topic 18, screen 17). Linearity in a normal probability plot is an indication that the data come from a normal distribution.

(25)

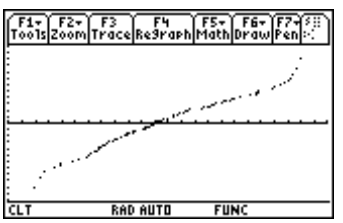


The normal probability plot (Topic 18, screens 18, 19, and 20) for screen 21 data is given in screen 25 using dots. The plot is fairly straight and denser in the middle, as would be expected for data from a normal distribution.

n = **4**

Screen 18 data has a normal probability plot shown in screen 26, which is somewhat normal in the center, but the tails are not long enough.

(26)

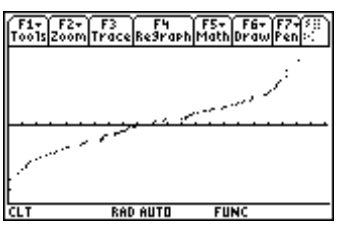


Theory also says that if these samples were from a population that was normally distributed, then even the distribution of means of sample size 4 could be normally distributed. You might want to investigate this.

n = **1** (uniform distribution)

The data in screen 15 has a normal probability plot as shown in screen 27, which is *not* denser in the middle and the tails are not long enough. You should not expect to see normality here because the points are uniformly distributed.

(27)



Tossing Dice

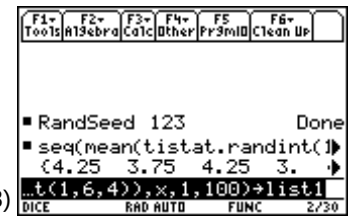
A die has values of 1, 2, 3, 4, 5, or 6, each with the probability of 1/6 occurring. Use **1:1-Var Stats** to find μ = **3.5** and σ = **1.70783**.

Note: *This could be done as a group activity.*

Create a folder named **DICE** and change to that folder. From the Home screen:

1. Set **RandSeed 123**.
2. Simulate tossing four dice or one die four times, calculate the mean, and repeat for 100 experiments with **seq(mean(tistat.randint(1,6,4)),x,1,100)→list1** (screen 28).

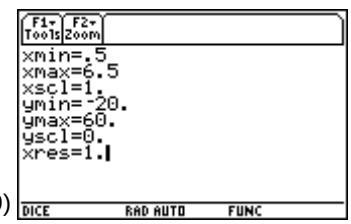
(28)



3. Set up the histogram. In the Stats/List Editor, press F2 **Plots**, **1:Plot Setup**. Set up and define **Plot 1** as Plot Type: **Histogram**, x: **list1**, and Hist. Bucket Width: **1**.
4. Set up the window using ◆ [WINDOW] with the following entries:
   - **xmin = .5**
   - **xmax = 6.5**
   - **xscl = 1**
   - **ymin = -20**
   - **ymax = 60**
   - **yscl = 0**
   - **xres = 1**

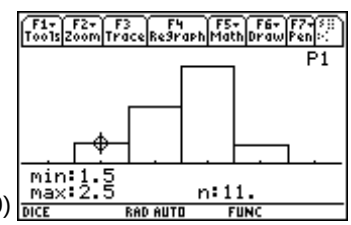
   (See screen 29.)

(29)



5. Press ◆ [GRAPH], and then F3 **Trace** (screen 30).

(30)

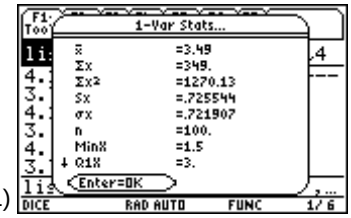


6. In the Stats/List Editor, press F4 **Calc**, **1:1-Var Stats**, with List: **list1** and Freq: **1** for the result of $\bar{x} = \mathbf{3.49} \approx \mu_x = \mu = 3.50$ and

   $s_x = \mathbf{0.725544} \approx \dfrac{1.70783}{\sqrt{4}} = 0.85 = \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{4}}$ (screen 31).

7. Repeat steps 1 through 6 from Topic 19, *Checking for Normality with a Normal Probability Plot* section.

(31)



The normal probability plot in screen 32 looks fairly straight, but with the granularity associated with limited possibilities, as seen in Topic 19, screen 9.

(32)