SIMULATIONS FOR UNDERSTANDING INFERENCE FOR REGRESSION

Written by: Julie Hicks
Woodbridge High School
Woodbridge, VA
jjhicks@home.com

Basketball Simulation Inspired by:
Floyd Bullard
NCSSM
Durham, NC
bullard@ncssm.edu

SIMULATIONS FOR UNDERSTANDING INFERENCE FOR REGRESSION

Inference for regression is typically the last topic covered in AP Statistics classes. Students study linear regression earlier in the year and fit "the" regression line to a set of data. When studying inference for regression, they must view the slope as a random variable with a sampling distribution rather than a static value. These activities are simulations designed to aid students in understanding the sampling distribution of the slope of the regression line, as well as what happens when some of the conditions for inference for regression are violated.

Simulations may be implemented with manipulatives, random digit tables, graphing calculators, and computer programs. It is assumed that students have performed simulations in different contexts earlier in the course and are familiar with the basic concept of simulation. The simulations for inference for regression require some use of technology. While general instructions are given to teachers, some of the student activity sheets are based on the TI-83 calculator. Print-outs of some of the programs are included so teachers can modify them for other calculators or for computer simulation.

The following is a brief description of each activity. More detailed teacher instructions and student activity sheets are available for each simulation.

1. Simulation of the Sampling Distribution of the Slope of the Regression Line - Sampling from a Finite Population

   Students will sample form a sample of heights and weights from a given population of professional basketball players in the NBA. The distribution of the slope of the regression lines of weight on height will be examined. TI-83 commands and programs are provided.

2. Simulation of the Sampling Distribution of the Slope of the Regression Line - Sampling from an Infinite Population

   Students will sample from a population generated from a known regression line to compare sample slopes to the known true slope. TI-83 commands and program are provided.

3. Examination of the Conditions for Inference for Regression

   Students will repeat the above simulation, but change the error structure to illustrate violation of the conditions for inference for regression. TI-83 commands and programs are provided.

SIMULATION OF THE SAMPLING DISTRIBUTION OF THE SLOPE OF THE REGRESSION LINE
Teacher Notes

Sampling from a Finite Population:
In these activities, students will investigate the behavior of the slope of the regression line calculated from a sample of points from a given population. After their earlier experiences fitting regression lines to data, students view the slope as a static number. Later, in the context of inference, students need to understand that the slope is a random variable with a sampling distribution. The purpose of this simulation is twofold. The first is to demonstrate that the slope of the regression line varies depending upon the particular sample chosen. Secondly, we want to verify the credibility of the assertion that if the response variable is normally distributed for each value of the explanatory variable, the distribution of slopes from different random samples of the same size from the same population is normal with mean equal to the slope of regression line for the whole population. The sampling distribution of the slope is the basis for the t-test for the slope.

In this simulation class generated data will be used to create a distribution of the slope of regression lines based on samples of size 5 from a given population. The population is a set of 278 professional basketball players in the NBA and the variables will be their heights and weights (data from sportingnews.com). One run will consist of choosing a random sample of 5 basketball players, without replacement, and recording the slope of the regression line of weight on height. Each student will complete 6 runs and class generated values will be used to build a graphical display of the distribution. (One way to do this is to put each value for the sample slope, b, on "sticky notes" and build a histogram on a piece of poster board. Another way would be to enter all values for b in one calculator to display on the overhead.) The mean and standard deviation of this distribution may also be calculated. Subsequently use all the data to find the least squares regression line for the population (the "true" regression line) and compare its slope, β, to the center of values shown on the histogram.
.

Extensions:
This activity may be implemented with a computer or a programmable graphing calculator with statistical capabilities. It is important, however, that students understand the simulation and perform several runs by hand before allowing the technology to generate many runs.

The accompanying programs are for the TI-83 graphing calculator. The data must be entered in two lists, áBBHT (height) and áBBWT (weight) prior to running the programs. (Program BBALL will do this for you.) Program SIMSHELL is an overall program that will repeat a run a given number of times. It calls program SIMONCE which performs one run with a given sample size, chosen without replacement.

Similar programs can be written for other calculators. When using technology, the simulation process is faster. Also, students can change both the sample size and the number of runs to see how the sampling distribution of the slope is affected. (Note: When using many runs or large sample sizes, the TI-83 will take several minutes. This might be a good activity for homework or when you have something else to do while the calculators are running.)

TI-83 programs for the basketball simulation: The data must be entered in two lists, áBBHT (height) and áBBWT (weight) prior to running the programs. The player numbers do not have to be entered.

```
prgmSIMSHELL                  This program will perform several runs of the simulation.
ClrHome
ClrList áSIMS
Disp "NUMBER OF"
Input "NO. OF RUNS?",D        Set the number of runs
Input "SAMPLE SIZE?",N        Set the sample size
ClrHome
For(I,1,D)
Output(1,1,"   ")
Output(1,1,D+1-I)             Count down to keep track of runs remaining
prgmSIMONCE                   Execute program SIMONCE (see below).
SüáSIMS(I)                    Enter the values for the slope in list SIMS.
End
If D=1                        If doing 1 run, plots a scatterplot with sample
Then                            regression line
Plot2(Scatter,áSAMPH,áSAMPW,Ð)
ZoomStat
Else
Plot3(Histogram,áSIMS)        If doing more than one run, plots a histogram of the
ZoomStat                        distribution of slopes.
Pause
1-Var Stats áSIMS             These five lines calculate and display the mean and
Output(1,1,"Ë =")               standard deviation of the sampling distribution of
Output(1,6,Ë)                   the slopes.
Output(3,1,"Sx =")
Output(3,7,Sx)
END
```

```
prgmSIMONCE                   This program performs one run of the simulation.
dim(áBBHT)üP                  Store dimension of data list to a variable
ClrList áIDS,áSAMPH,áSAMPW
Lbl A
randInt(1,P,N)üáIDS           These three lines allow you to sample without
SortA(áIDS)                     replacement.
If prod(¾List(áIDS))=0
Goto A
For(B,1,N)                    These three lines put your sample values into new lists.
áBBHT(áIDS(B))üáSAMPH(B)
áBBWT(áIDS(B))üáSAMPW(B)
End
LinReg(a+bx) áSAMPH,áSAMPW,Y,      Calculate the least squares regression line for the sample
büS                                Store the slope to the variable S.
```

Note: When doing only one run, the program will create a scatterplot of the sample using the largest mark and display the regression line. An interesting activity would be to graph the population (in Plot 1) and population regression line in (Y_1) using the small marks. When doing more than one run, the program will skip the scatterplot and plot a histogram of the sample slopes.

Sampling from an Infinite Population:
An alternative to sampling from a finite population would be to sample points from a hypothetical population with a given "true" regression line. The advantage to this is that you know the "answer" and can study the behavior of the estimated slope when all conditions are met or when one or more are violated.

Start by choosing an equation for the "true" regression line, for example, $y = -2 + 5x$. The data in this population will be randomly scattered about the line with a given standard deviation, σ. This may be simulated by adding an error term to the regression equation and calculating the coordinates for several ordered pairs. For example, the integers from 1 to 12 could be the explanatory variables and $y = -2 + 5x + \text{error}$ would be the equation to generate the responses. Since one of the necessary conditions for inference for regression is that for a fixed value of the explanatory variable, the response values are normally distributed, the error should be generated from a normal distribution. Calculate the least squares regression line for these data and record the slope. Repeat several times and construct a graphical display from the class generated values of the slopes. Calculate the mean and standard deviation for the class data.

This simulation requires technology and can be implemented on either a calculator or a computer. It is important that students understand the simulation, therefore they should work through a few runs before allowing the calculator or computer to perform many runs at once.

The following general instructions will assume a sample size of 12 and a population whose "true" regression equation is $y = -2 + 5x$. The data points will be distributed about the line with error distribution N(0, 3).

1. Enter integers from 1 to 12 into a column. (These are the explanatory variables.)
2. Generate 12 random numbers from the N(0, 3) distribution. (These are the errors.)
3. Calculate the response variables using the equation $y = -2 + 5x + \text{error}$.
4. Calculate the least squares regression line for your sample and record the slope.
5. Repeat steps 2-4 many times and plot a histogram of the distribution of the slope. Calculate the mean and standard deviation of the sample slopes.

For further investigations, students could change the sample size, number of runs, or the standard deviation of the errors.

Instructions for Implementing Sampling from an Infinite Population on the TI-83

The following describes how to implement this activity on the TI-83 using only commands from the home screen. It is important for the students to understand what is happening in each step before they are given a program to accomplish the simulation.

1. Initialize a counter: `1→C`.
2. Enter the numbers from 1-12 in L_1. `seq(X,X,1,12,1)→`L_1
3. Generate the errors and store in L_2: `randNorm(0,3,12)→`L_2
4. Calculate the response variables and store in L_3: `-2+5*`L_1`+`L_2`→`L_3
5. Calculate the regression line for your sample: `STAT-CALC-8:LinReg` L_1, L_3
6. Store the slope in L_4: `b→`L_4`(C)` (Find b under the VARS-Statistics-EQ menu.)
7. Update the counter: `C+1→C`
8. Repeat steps 3-7 many times. This may be done by using 2nd ENTRY or by entering lines 3-7 all on one line, separated by colons. Every time ENTER is selected the process is repeated. The numbers generated on the home screen are the last value for the counter (one larger than the number of repetitions you have done).
9. Create a histogram of the slopes and calculate the mean and standard deviation.

For those with programming experience, the same commands could also be entered into a program, using a FOR loop instead of the counter. The following is a sample TI-83 program that will randomly generate a given number of points and perform a given number of runs of the simulation.

```
prgmSLOPDIST
Float
ClrHome
ClrList L•,L‚,Lƒ,L„,L…,L†
Disp "NUMBER OF"
Input "REPETIONS?",D        Set the number of runs
Input "SAMPLE SIZE?",N      Set the sample size
ClrHome
1üC                         Initialize the counter
seq(X,X,1,N,1)üL•           Enter the explanatory variables into List 1
For(I,1,D)                  Start of loop that performs the runs
randNorm(0,3,N)üL‚             Enter the errors into List 2
ú2+5*L•+L‚üLƒ                  Store the response variables into List 3
LinReg(a+bx) L•,Lƒ             Calculate the least squares regression line
büL„(C)                        Store the slope to List 4
Output(1,1,"  "
Output(1,1,D+1-C)              Count down to keep track of runs remaining
C+1üC                          Update the counter
End                         End of loop
Plot3(Histogram,L„)         Plot a histogram of the slopes
ZoomStat
Pause
1-Var Stats L„              Calculate mean and standard deviation of the slopes
Fix 3                       Set decimal places to 3
Output(1,1,"Ë =")           These four lines display the mean and standard
Output(1,6,Ë)                 deviation of the simulated sampling distribution
Output(3,1,"Sx =")            of the slopes
Output(3,7,Sx)
```

EXAMINATION OF CONDITIONS FOR INFERENCE FOR REGRESSION
Teacher Notes

One of the conditions for inference for regression is that at any particular x-value the errors are normally distributed with a constant standard deviation. This activity examines what happens to the distribution of the slopes when this condition is violated. When sampling from an infinite population, students generated data with an error structure that was normal and had a constant standard deviation throughout. Now they will generate data that violate the conditions of constant standard deviation and of normality.

Steps for sampling from an infinite population: (repeated from previous activity)

1. Enter integers from 1 to 12 into a column. (These are the explanatory variables.)
2. Generate 12 random numbers from the N(0, 3) distribution. (These are the errors.)
3. Calculate the response variables using the equation $y = -2 + 5x + \text{error}$.
4. Calculate the least squares regression line for your sample and record the slope.
5. Repeat steps 2-4 many times and plot a histogram of the distribution of the slope. Calculate the mean and standard deviation of the sample slopes.

What happens to the distribution of slopes when the condition of constant standard deviation is violated?

Have students repeat the activity involving sampling from an infinite population, but change the error structure to **error = explanatory variable * a number generated from N(0, 3)**. The syntax for this on a TI-83 calculator is L_1`* randNorm(0,3,12)`$\rightarrow L_2$. This will generate data that "fan out" as x increases. In this case the errors are still normally distributed at each value of the explanatory variable, but the standard deviation increases as the explanatory variable increases. What happens to the distribution of the slopes?

What happens to the distribution of slopes when the condition of normality of responses for a given explanatory variable is violated?

Another condition for inference for regression is that the errors are normally distributed. What happens when the errors have a constant standard deviation, but are not normally distributed? Repeat the activity involving sampling from an infinite population, but change the error structure in to a highly skewed distribution, **error = (c^{2} with df = 1) – 1**. The syntax for this on a TI-83 is a little different: `(randNorm(0,1,12))`^{2}`-1`$\rightarrow L_2$. (The square of the standard normal distribution is the χ^2 distribution with one degree of freedom. Since this distribution has a mean of 1, we must subtract 1 to make the expected value of the errors equal to zero.) What happens to the distribution of the slopes?

For further explorations, see the document "Violating Conditions in Significance Testing". Through simulation on the TI-83 calculator students will examine how the probability of a Type I error compares to the stated significance level α when conditions for the *t*-test are violated.

SIMULATION OF THE SAMPLING DISTRIBUTION OF THE SLOPE OF THE REGRESSION LINE

Sampling from a Finite Population

What is the relationship between the height and weight of professional basketball players in the NBA? We would suspect that there may be a linear association and that the slope would be positive. However, if we just sample a few players, the slope of the regression line of weight on height will vary depending on the players chosen. In this activity you will use class generated data to examine the sampling distribution of the slope of the least squares regression line.

The back of this sheet contains data on the heights and weights of 278 NBA basketball players (data from sportingnews.com). One run will consist of choosing a random sample of 5 players, calculating the least squares regression line of weight on height, and recording the slope, y-intercept, and standard deviation about the line. Perform a total of 6 runs.

Sample	Slope	y-intercept	Standard Deviation
1			
2			
3			
4			
5			
6			

Make a graphical display of the slopes generated by the class, as directed by your teacher. Describe the distribution of the slopes and record the mean and standard deviation.

mean = _______________ standard deviation = _______________

What is your guess for the slope of the regression line for the entire population?

Link and get the programs BBALL, SIMSHELL, and SIMONCE if you have not already done so. Run program BBALL to enter the data for the entire population into lists BBHT and BBWT. Make a scatterplot (using the smallest mark) of the entire population and graph the population regression line. How close was your guess for the slope to the population slope?

Run program SIMSHELL and enter one run. The program will perform one run of the simulation you just completed, store the regression equation in Y_2, and graph the points in Plot 2. Run the program several times and observe how the regression line changes from sample to sample. Did you get any very unusual lines?

HEIGHTS AND WEIGHTS OF NBA PLAYERS

Player	Ht	Wt	Player	Ht	Wt	Player	Ht	Wt	Player	Ht	Wt	Player	Ht	Wt	Player	Ht	Wt
1	70	173	51	76	200	101	78	220	151	80	220	201	83	230	251	84	230
2	70	171	52	76	190	102	78	205	152	80	235	202	83	225	252	84	240
3	71	182	53	76	190	103	78	200	153	80	224	203	83	280	253	84	250
4	71	180	54	76	190	104	78	226	154	80	220	204	83	220	254	84	240
5	71	163	55	76	205	105	78	225	155	80	220	205	83	230	255	84	255
6	71	170	56	76	208	106	78	190	156	81	235	206	83	230	256	84	265
7	72	198	57	76	212	107	79	245	157	81	245	207	83	264	257	84	245
8	72	170	58	76	194	108	79	255	158	81	230	208	83	265	258	84	290
9	72	160	59	76	194	109	79	215	159	81	220	209	83	242	259	84	245
10	72	165	60	76	190	110	79	193	160	81	230	210	83	261	260	84	245
11	72	180	61	76	203	111	79	250	161	81	240	211	83	220	261	84	277
12	72	195	62	76	190	112	79	260	162	81	220	212	83	245	262	84	248
13	73	200	63	76	195	113	79	215	163	81	225	213	83	260	263	84	249
14	73	189	64	76	205	114	79	210	164	81	254	214	83	240	264	84	235
15	73	175	65	76	190	115	79	215	165	81	240	215	83	245	265	84	240
16	73	180	66	77	209	116	79	249	166	81	230	216	83	246	266	85	300
17	73	185	67	77	200	117	79	215	167	81	225	217	83	235	267	85	250
18	73	195	68	77	194	118	79	236	168	81	225	218	83	270	268	85	260
19	73	168	69	77	210	119	79	235	169	81	255	219	83	230	269	85	244
20	73	183	70	77	210	120	79	225	170	81	245	220	83	234	270	85	260
21	73	180	71	77	185	121	79	218	171	81	248	221	83	235	271	86	292
22	73	200	72	77	216	122	79	220	172	81	245	222	83	235	272	86	260
23	73	185	73	77	215	123	79	263	173	81	255	223	83	225	273	86	280
24	73	185	74	77	210	124	79	255	174	81	240	224	83	256	274	86	285
25	73	180	75	77	205	125	79	228	175	81	256	225	83	260	275	87	238
26	73	175	76	77	179	126	79	240	176	81	260	226	83	235	276	87	292
27	73	185	77	77	205	127	79	215	177	81	225	227	83	246	277	88	265
28	73	195	78	77	220	128	79	250	178	81	255	228	84	240	278	91	303
29	73	190	79	77	185	129	79	240	179	81	245	229	84	250			
30	74	177	80	77	209	130	79	185	180	81	250	230	84	260			
31	74	190	81	77	208	131	79	215	181	81	225	231	84	253			
32	74	175	82	77	190	132	79	205	182	81	258	232	84	250			
33	74	191	83	77	205	133	79	215	183	81	245	233	84	245			
34	74	185	84	77	183	134	79	215	184	81	230	234	84	245			
35	74	205	85	77	215	135	79	228	185	81	240	235	84	240			
36	74	182	86	77	220	136	80	215	186	81	220	236	84	260			
37	74	195	87	78	218	137	80	250	187	81	260	237	84	260			
38	74	180	88	78	200	138	80	210	188	81	235	238	84	245			
39	75	190	89	78	220	139	80	228	189	81	235	239	84	245			
40	75	185	90	78	195	140	80	215	190	81	230	240	84	232			
41	75	195	91	78	208	141	80	230	191	81	255	241	84	240			
42	75	195	92	78	230	142	80	225	192	81	220	242	84	230			
43	75	210	93	78	195	143	80	205	193	81	220	243	84	265			
44	75	210	94	78	228	144	80	256	194	81	235	244	84	220			
45	75	180	95	78	228	145	80	220	195	81	230	245	84	220			
46	75	190	96	78	216	146	80	210	196	81	206	246	84	250			
47	75	200	97	78	252	147	80	210	197	83	250	247	84	255			
48	75	202	98	78	225	148	80	250	198	83	225	248	84	255			
49	75	185	99	78	216	149	80	230	199	83	230	249	84	240			
50	76	200	100	78	210	150	80	250	200	83	260	250	84	265			

SIMULATION OF THE SAMPLING DISTRIBUTION OF THE SLOPE OF THE REGRESSION LINE

Sampling from an Infinite Population

The regression line calculated for a given situation will vary depending on the particular observations made. In this activity you will examine the sampling distribution of the slope of the least squares regression line. We will assume the true population regression line is y = –2 + 5x and draw random samples from this population. Remember the data in this population do not all fall on the given line, rather for each value of the explanatory variable the responses vary normally with mean 0 and standard deviation 3.

#1 – 7: For each calculator command, state what is being done:

1. `1→C` ____

2. `seq(X,X,1,12,1)→L1` ________________________________

3. `randNorm(0,3,12)→L2` _______________________________

4. `-2+5*L1+L2→L3` ____________________________________

5. `STAT-CALC-8:LinReg L1,L3` ___________________________

6. `b→L4(C)` (Find b under the VARS-Statistics-EQ menu.) ______________

7. `C+1→C` ____

8. Steps 3-7 constitute one run of the simulation. Repeat steps 3-7 many times. This may be done by using 2nd ENTRY or by entering lines 3-7 all on one line, separated by colons. Every time ENTER is selected the process is repeated. The numbers generated on the home screen are the last value for the counter (one larger than the number of repetitions you have done).

9. Create a histogram of the slopes and calculate the mean and standard deviation.

   Histogram:

   $\overline{x}$ =

   s_x =

10. Based on the histogram, what guess would you make for the slope of the population regression line? How does your guess compare to the "true" slope?

EXAMINATION OF CONDITIONS FOR INFERENCE FOR REGRESSION

What happens to the distribution of slopes when conditions are violated?
Previously you examined the distribution of the slope of the regression line from samples from an **infinite** population by using the program SLOPDIST or the following steps on your calculator.

1. Enter integers from 1 to 12 into a column. (These are the explanatory variables.)
2. Generate 12 random numbers from the N(0, 3) distribution. (These are the errors.)
3. Calculate the response variables using the equation y = –2 + 5x + error.
4. Calculate the least squares regression line for your sample and record the slope.
5. Repeat steps 2-4 many times and plot a histogram of the distribution of the slope. Calculate the mean and standard deviation of the sample slopes.

One of the conditions for inference for regression is that at any particular x-value the errors are normally distributed with a constant standard deviation. This activity examines what happens to the distribution of the slopes when this condition is violated.

What happens to the distribution of slopes when the errors for each value of the explanatory variable are normally distributed, but condition of constant standard deviation is violated?
Repeat the activity involving sampling from an infinite population, but change the error structure to **error = explanatory variable * a number generated from N(0, 3)**. The syntax for this on a TI-83 calculator is `L1* randNorm(0,3,12)→L2`. This will generate data that "fan out" as x increases. In this case the errors are still normally distributed at each value of the explanatory variable, but the standard deviation increases as the explanatory variable increases. Describe the distribution of the slopes when the condition of constant standard deviation is violated.

What happens to the distribution of slopes when the errors for each value of the explanatory variable have a constant standard deviation, but the condition of normality is violated?
Repeat the activity involving sampling from an infinite population, but change the error structure in to a highly skewed distribution, **error = (c^{2} with df = 1) – 1**. The syntax for this on a TI-83 is a little different: `(randNorm(0,1,12))`^{2}`-1→L2`. (The square of the standard normal distribution is the χ^2 distribution with one degree of freedom. Since this distribution has a mean of 1, we must subtract 1 to make the expected value of the errors equal to zero.) What happens to the distribution of the slopes?