Sample Activities for Two Quantitative Variables

Beth Chance (bchance@calpoly.edu)

Activity #1: Predicting Heights

Step 1. Ask a research question. Are there competitive advantages in sports?

Based on trends observed in other sports, researchers were interested in exploring how heights of winning tennis players might be changing over time given that height may be an advantage to winning at tennis.

Step 2. Design a study and collect data. Researchers gathered data on heights of the winning tennis players at The Championships, Wimbledon, a prestigious tennis tournament held each year since the 1880s through 2019 (https://www.topendsports.com/sport/tennis/anthropometry-wimbledon.htm).

(a) Identify an observational unit in this study.

(b) What is the explanatory variable? Is it categorical or quantitative?

(c) What is the response variable? Is it categorical or quantitative?

(d) In your own words, explain what it would mean for there to be an *association* between the explanatory and response variables in this study.

(e) Is this study an experiment or an observational study? Why?

(f) What other variables might help predict someone's height?

(g) Summarize your thoughts in a "Sources of Variation diagram.

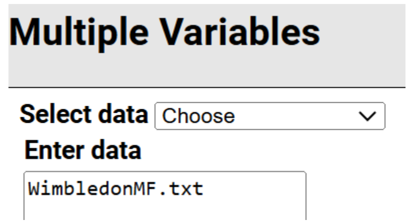
Observed Variation in:	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria *Design*		

(h) How might you represent this information in a *model*?

(i) Estimate the height of a Wimbledon champion. Provide a measure of accuracy of your estimate.

These materials were developed for the EAPOST Scholar Program and supported by the National Science Foundation under Grant NSF DUE2235355.

STEP 3: Explore the data.

Open the Multiple Variables applet. Press **Clear** to empty the data window. Type: WimbledonMF.txt Press **Use Data** to preview the data Press **Use Data** again to load in the data.	

(j) Drag the Height(cm) variable into the Response box. Summarize the behavior of the distribution. Be sure to start with "The distribution of heights of Wimbledon champs is …"

(k) What would you report as a typical height? How accurate was your prediction in (i)?

(l) The mean is 178.15 cm. Suppose I predict the 2025 winner will be 178.15 cm tall. How large of a "potential prediction error" would you estimate?

(m) To see whether heights have tended to change over time, drag the year variable to the Explanatory box. Summarize what you learn from this scatterplot. Would you say there is evidence that champions are getting taller?

When there is a lot of "noise" in the data, it can be difficult to see a pattern (the "signal"). To help focus on the overall trend, check the Show equation box to add the least squares regression line.

(n) What does this line tell you?

(o) Based on the line, what would you predict for the height of the 2025 champion? How large of a "potential prediction error" would you estimate?

Discussion: The standard deviation of a distribution can be interpreted as a typical prediction error. We start by looking at all the prediction errors $(y_i - \bar{y})$ and then squaring these values (so positive values and negative values don't cancel each other out), and then averaging them and taking the square root to return to the original units. When you fit a regression line, you can see how much smaller the average squared residual $(y_i - \hat{y})^2$ is.

(p) Check the Regression SE box. Is this value smaller than the original standard deviation? By a lot? What is the percentage reduction in the squared values?

Discussion: R^2 can be interpreted as the "percentage reduction in unexplained variation in the response variable" due to the explanatory variable.

(q) Suggest another variable that you think would explain more variation in heights than year did. Test your conjecture with the applet by removing Year from the Explanatory box and moving Sex into the Subset By box. What do you learn?

(r) Is there a *tendency* for the Men's champion to be taller than the Ladies' champion? What does that mean?

(s) If I ask you to predict the height of the Ladies champion in 2025, how accurate should this prediction be?

Now drag Year back into the Explanatory variable box.
(t) How do the *s* and R^2 values change? Is this what you would have predicted?

(u) Check the Show equation box. From the graph, does the time trend appear to differ between the Mens and Ladies champions? Explain how you are deciding.

(v) Explain why the following two models differ. What does the coefficient of male tell you in each case?

Statistical model: ☑

Term	Coeff	SE
Intercept	172.70	0.63
Sex		
M	9.71	0.85
F	(ref)	

Indicator coding

Term	Coeff	SE
Intercept	53.97	20.13
Year	0.0603	0.0102
Sex		
M	10.4898	0.7950
F	(ref)	

(w) How could we decide whether the time trend is *statistically significant*?

Possible Extensions:

- Student collected data: Height and footlength, handspan, vertical leap, haircut price, sleep, caffeine consumption
- Roller coaster height and maximum speed
- Height and finger length ("Giving the Finger to Dating Services" *Chance* magazine, 2008)

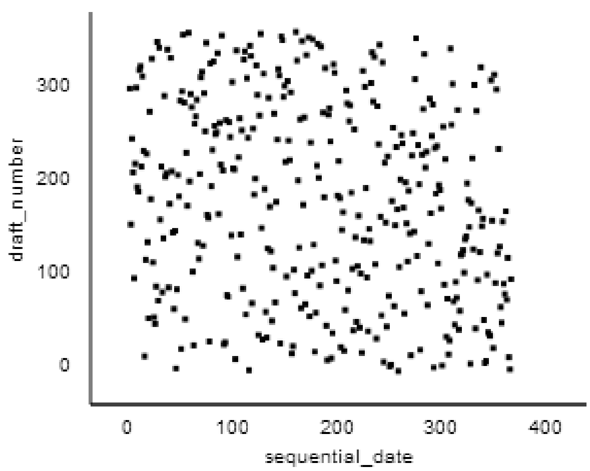
- Child height and parent height
- Height and how far can walk in a straight line while blindfolded (Axtell, 2004)

- Critique the graphs in the original article

Activity #2: Draft Lottery (randomization test for correlation coefficient)
In 1970, the United States Selective Service conducted a lottery to decide which young men would be drafted into the armed forces (Fienberg, 1971). Each of the 366 birthdays in a year (including February 29) was assigned a draft number. Young men born on days assigned low draft numbers were drafted.
CBS News Report: https://www.youtube.com/watch?v=-p5X1FjyD_g
We will regard the 366 dates of the year as observational units. We will consider two variables recorded on each date: draft number assigned to the date, and sequential date in the year (so January 31 is sequential date 31, February 1 is sequential date 32, and so on).

(a) In a perfectly fair, random lottery, what should be the value of the correlation coefficient between draft number and sequential date of birthday?

The graph below displays a scatterplot of the assigned draft numbers and the sequential dates. There are 366 dots, one for each day of the (leap) year.



(b) Does the scatterplot reveal much of an association between draft number and sequential date?

(c) Based on the scatterplot, guess the value of the correlation coefficient.

(d) Does it appear that this was a fair, random lottery?

It's difficult to see much of a pattern or association in the scatterplot, so it seems reasonable to conclude that this was a fair, random lottery with a correlation coefficient near zero.

But let's dig a little deeper …

The 1970 Draft Lottery data sheet (at the end of the exploration) shows the draft number assigned to each of the 366 birthdays.

(e) Find and report the draft number assigned to your birthday.

The second table at the end of the activity has ordered the draft numbers within each month.

(f) Use this table to determine the median draft number for your birth month.

(g) Collaborate with your classmates to determine and report the median draft number for all twelve months.

Month	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec

Median draft number												

(h) Do you notice any pattern or trend in the median draft numbers over the course of the year?

(i) The correlation coefficient for these data is r = -0.226. What does this number reveal? Is it consistent with the scatterplot?

(j) Suggest two possible explanations (hypotheses) which could have generated the value of the observed correlation coefficient.

(k) In your own words, how could we go about determining whether random chance is a plausible explanation for the observed correlation value between sequential date and draft number? Explain how the 3S strategy could be applied here, in particular identifying a simulation strategy you could conduct "by hand." Note: You do not need to actually carry out a simulation analysis.

i. What is the statistic?

ii. How would you simulate?

iii. How would you evaluate the strength of evidence?

The null hypothesis to be tested is that the lottery was conducted with a fair, random process. The null hypothesis would therefore mean that there is no association between sequential date and draft number for this process. The alternative hypothesis is that this lottery was not conducted with a fair, random process, so there is an association between sequential date and draft number.

How can we assess whether the observed correlation coefficient of r = -0.226 is far enough from zero to provide convincing evidence that the lottery process was not random? Like always, we ask how unlikely it would be for a fair, random lottery to produce a correlation value as far from zero as -0.226. Also like always, we answer that question by simulating a large number of fair random lotteries, calculating the correlation coefficient for each one, and seeing how often we obtain a correlation coefficient as or more extreme (as far from zero) as -0.226.

(l) Open the **Two Quantitative Variables** applet. Copy and paste the data from **DraftLottery.txt** (http://www.isi-stats.com/isi/data/chap10/DraftLottery.txt) into the applet (remember to include the column titles). (Or use the pull-down menu.)

1. Check the **Correlation Coefficient** box and confirm that the correlation coefficient is -0.226.

2. Check the **Show Shuffle Options** box and select the **Correlation** from the pull-down menu to keep track of that statistic. Then press **Shuffle Y-values** to simulate one fair, random lottery. (Select the **Plot** radio button to see the corresponding scatterplot.) Record the value of the correlation coefficient between the shuffled draft numbers and sequential date (in blue).

3. Press **Shuffle Y-values** four more times to generate results of four more fair, random lotteries. Record the values of the correlation coefficients in the table below.

Repetition	**1**	**2**	**3**	**4**	**5**

Correlation coefficient					

(m) Change the **Number of Shuffles** from 1 to 995 and press **Shuffle Y-values** to simulate 995 more fair, random lotteries. Look at the null distribution of these 1000 correlation coefficients. Where is this distribution centered? Why does this make sense?

(n) Use **Count Samples** pull-down menu to select **Beyond**. Specify the observed correlation coefficient (-0.226) and press **Count**. What proportion of the 1000 simulated random lotteries produced a correlation coefficient at least as extreme (as far from zero in either direction) as -0.226? Report the approximate p-value.

(o) Interpret this p-value: This is the probability of what, assuming what?

(p) What conclusion would you draw from this p-value? Do you have strong evidence that the 1970 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion.

The mixing process was changed for the 1971 draft lottery, for which the correlation coefficient turned out to be $r = 0.014$.

(q) Use your simulation results to approximate the p-value for the 1971 draft lottery. Is there any reason to suspect that this 1971 draft lottery was not conducted with a fair, random process? Explain the reasoning behind your conclusion. Also explain why you don't really need to paste in the data from the 1971 lottery first.

These materials were developed for the EAPOST Scholar Program and supported by the National Science Foundation under Grant NSF DUE2235355.

1970 Draft Lottery Data

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

The following table arranges the draft numbers in order for each month:

rank	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	17	4	29	2	31	20	13	11	1	5	9	3
2	52	25	30	14	35	22	15	21	6	7	19	10
3	58	57	33	32	37	28	23	36	8	24	34	12
4	59	68	108	62	40	60	27	44	18	38	46	16
5	77	86	122	74	55	64	42	45	49	72	47	26
6	92	89	136	81	65	69	50	48	63	79	51	39
7	101	91	139	83	75	73	67	54	71	87	66	41
8	118	144	166	90	103	85	88	61	82	94	76	43
9	121	150	169	124	112	104	93	102	113	117	80	53
10	140	152	170	147	130	109	98	106	119	125	97	56
11	159	179	200	148	133	110	115	111	149	138	99	70
12	164	181	213	191	155	134	120	114	151	171	107	78
13	186	189	217	208	178	137	153	116	158	176	126	84
14	194	205	223	218	183	180	172	141	160	192	127	95
15	199	210	239	219	197	206	187	142	161	196	131	96
16	211	212	256	231	226	209	188	145	175	201	132	100
17	215	214	258	252	250	222	190	154	177	202	143	105
18	221	216	259	253	276	228	193	167	184	220	146	123
19	224	236	265	260	278	247	227	168	195	229	156	128
20	235	285	267	262	295	249	248	198	204	234	174	129
21	238	290	268	269	296	272	270	245	207	237	182	135
22	251	292	275	271	298	274	277	261	225	241	185	157
23	280	297	293	273	308	301	279	286	232	243	203	162
24	305	299	300	312	313	335	284	291	233	244	230	163
25	306	302	317	316	319	341	287	307	242	254	266	165
26	318	338	323	336	321	353	289	311	246	264	281	173
27	325	347	332	340	326	356	303	324	255	283	282	240
28	329	363	334	345	330	358	322	333	257	288	309	304
29	337	365	343	346	357	360	327	339	263	294	310	314
30	349		354	351	361	366	331	344	315	342	348	320
31	355		362		364		350	352		359		328

Activity #3: Below is a modification of a lab assignment from 2011. There is also a newer version with a different context here. In particular see "page 4."

Lab #9: Studying and GPA
Due: Thursday, March 10 (beginning of class)

No late work will be accepted!

Directions: Working in pairs, work through Lab #9 (2011 instructions). Type the answers to each question in this document. *Please do **NOT** put your answer in bold*. When you are finished, print out and turn in a hard copy of your report.

Goals: In this lab you will explore:

- Effective ways of displaying and describing the association between two quantitative variables
- Numerical measure of the linear relationship between two quantitative variables
- A mathematical model (regression equation) for predicting GPA from study hours
- The sampling variability in regression equations from repeated sampling from a hypothesized population
- A hypothesis test for deciding when an observed relationship is statistically significant to help decide whether we have convincing evidence that students who study more hours tend to have higher GPAs than students who study fewer hours.

Name:
Section number:

Data Collection:

(a) Write a paragraph summarizing the key characteristics of this study. In particular, identify:

- **Observational units and sample size**
- **Population of interest**
- **Response variable and whether it is categorical or quantitative**
- **Explanatory variable and whether it is categorical or quantitative**
- **Type of study (observational or experimental) and why**
- **The types of randomness used in the study (random assignment, random sampling, both, neither)**

>>

Descriptive Statistics:

Open the Corr/Regression applet. Load the UOPGPA.txt data (can enter the name in the data window box). Use the pull-down menus to specify GPA on the vertical axis and StudyHoursperWeek on the horizontal axis. (Note: "snakecase") You might also want to ask students to check the Show descriptive statistics box and note the values.

(b) In describing scatterplots, our goal is to describe the association between the two variables. To do so, we focus on three characteristics: direction, strength and form.

- **Is the *direction* of the association as expected by the student researchers? State specifically what the student researchers believed the relationship would be.**
- **Would you consider the association *strong*? (In other words, does knowing a person's study hours help you to accurately predict their GPA?) Explain.**
- **Does the *pattern of the association* seem reasonably well modeled by a *line*? (Or, is there curvature or some other prominent pattern in the plot?)**

Paste copy of scatterplot here.

Numerical Summary:

See page 4 of this lab. Check the **Correlation coefficient** box.

(c) Report the value of *r*. Does the correlation coefficient indicate a positive or negative linear association here? Does it seem strong or weak? Explain.

Paste a copy of your output here.

Interpreting the Least Squares Regression Equation:

See p. 5-6 of online lab. Check the **Show Regression Line** box.

(d) Include a copy of your scatterplot with the regression line superimposed over the data. Provide a detailed (statistical) interpretation of both the slope and intercept coefficients in the context of this study.

Include a copy of your scatterplot here.

Statistical Inference - Simulation:

Simulation I:

To perform the simulation, we will (as always) assume the null hypothesis to be true, that there is no genuine association between these two variables. In other words, the pairing that exists in the data is arbitrary and we could just as easily mix up the GPA values that have been "assigned" to each Study Hours variable. To model this, we will "shuffle" the 80 GPA values and place them at random with the *x* values.

- In the applet, uncheck Show Squared Residuals if that is still checked
- Check the box next to **Show Shuffle Options**.
- Use the **Statistic** pull-down menu to select the Slope.
- Select the **Data** radio button.
- Press the **Shuffle Y-values** button.

Notice that the Study Hours values haven't changed but the GPA values have been randomly mixed up and reassigned to those Study Hours values.

- Switch from the Data radio button to the **Plot** radio button.

This scatterplot shows these shuffled values and the new regression line for this rerandomized sample (in blue). (The original regression line is in red. If not, check the Show Regression Line box again.)

(f) Did you obtain the same simulated regression line as the researchers (in red)? Is the association still positive in the shuffled scatterplot? Is the relationship stronger or weaker than the one observed by these researchers (how are you deciding)?

(g) Press **Shuffle Y-values** again. Did you obtain the same (blue) regression line?

(h) Notice that the two sample slope values you generated have been added to the dotplot on the right (the most recent in blue). Suppose you repeat this shuffling 1000 times, where do you think this distribution will be centered? Why?

(i) Press the **Shuffle Y-values** button 8 more times (for a total of 10 shuffles).

- Describe the pattern you are seeing in how the regression **lines** on the shuffled scatterplot vary from shuffle to shuffle (what pattern is there to the grey lines on the scatterplot) as if to someone who couldn't see the same picture.
- Does the observed regression line for this study (in red) appear to be extreme (compare to the simulated lines)?

To better understand the behavior of this null distribution of sample slopes, we of course need to take a lot more samples.

- Set the **Number of Shuffles** to 990 (for a total of 1000).
- Press **Shuffle Y-values**.

(j) Describe the shape, center (mean), and spread (standard deviation) of the histogram for the resulting null distribution of sample slopes. Also, make sure you describe what the standard deviation is measuring (*Hint:* What is the "variable" in this histogram?).

Now use this distribution to approximate the p-value for our research question.

- For the **Count Samples** box, use the pull-down menu to indicate whether you want a one-sided (Greater Than or Less Than) or a two-sided (Beyond) p-value, and then specify the value of the observed sample slope for this study in the box, and press **Count**.

(k) Include a screen capture of your output (null distribution and p-value).

Simulation II

The previous simulation models "random assignment," but this study (kinda) used random sampling. So let's try another simulation approach.

- In the applet, uncheck the Show Shuffle Options box.
- In the lower left, check the box next to **Create population**.
  Use the Create population pull-down menu to select **Bivariate**.
- The "Population Inputs" values have been set to match the sample (e.g., average study hours = 1.84), except the population slope has been set to zero.
- Press the **Create Population** button.

Confirm that this represents no association between GPA and study hours in a population of 20,000 individuals.

- Check the **Show Sampling Options** box.
- Set the Sample Size to 80 to match the original study.
- Press the **Draw Samples** button.

You should see one sample (in blue) and 1000 different regression lines from 1000 different random samples from the population. Check the **Show Original Regression Line** box.

(l) Describe the pattern you are seeing in how the regression **lines** on the shuffled scatterplot vary from shuffle to shuffle (what pattern is there to the grey lines on the scatterplot) as if to someone who couldn't see the same picture. Does the observed regression line for this study (in red) appear to be extreme (compare to the simulated lines)? Write the "alt text" for this image.

(m) Report the mean and standard deviation of the null distribution. How do they compare to the earlier simulation?

(n) Include a screen capture of the p-value for this null distribution. How does it compare to the earlier simulation?

(o) Using the standard deviation you found in (m), how many standard deviations is the observed sample slope (0.0894) from the hypothesized slope (0.0)?

(p) Also use the standard deviation you found in (m) to approximate a 95% confidence interval for the population slope. (Using the 2SD short-cut. Show your work.)

Extra credit: *Without hitting reset, change the sample size from 80 to 40 and generate 1,000 more random samples. Describe the changes in the sampling distribution of the slope and the new p-value. Do they change as expected? Explain including what you expected and why.*

Paste a copy of both sampling distributions here.

Statistical Inference – Central Limit Theorem:

Check the **Regression table** box

Include a copy of your output here.

(k) How does the SE Coef value corresponding to the study hours row compare to the standard deviation of the slopes found in (g)? How does this t-value compare to what you found in (i)? How does the (two-sided) p-value compare to the one-sided p-value you estimated in (h)?

(l) Interpret and evaluate the p-value you found. Should you reject or fail to reject the null hypothesis? What does this decision imply about the research question?

Conclusions:

(m) Write a few sentences summarizing your analysis, in particular, for this research study:

- **What did you learn about the relationship between study hours and GPA for this sample from the scatterplot and the correlation coefficient about the sample data?**
- **What did you learn about the null hypothesis from the test statistic and p-value?**
- **What are your final conclusions about the relationship between study hours and GPA for this population and how do they follow from these analyses?**
- **What population are you willing to generalize these results to? Are you able to draw a cause-and-effect conclusion between studying and GPA? Explain.**
- **What concerns do you have about how these data were collected and how that might impact the conclusions you can draw form this study? (Your answer should discuss at least one potential non-sampling error.)**

Keep in Mind: There is a difference between the strength of a relationship and the strength of the evidence against the null hypothesis………

Before Turning in Your Lab Report:

- Proofread!!!!!!!!!
- Staple your pages together. Reports that are not stapled will not be graded!
- Fold the report in half (lengthwise). Put both team members' names (first and last names) on the outside of the report with your section number.

Possible Extensions:

- When simulating sampling from a population, have students change each of the input values, one at a time, and ask them to predict and test how that changes the population and the sampling distribution of the slopes.