Problems: BIVARIATE REGRESSION ANALYSIS

1. Your hypothesis is that when parents regularly use physical punishment to discipline their children, that these children will develop low self-esteem (i.e., a belief about oneself as being of little value). You obtain data on the self-esteem of 50 adolescents. The parents of half of these adolescents used physical punishment to discipline their children (including the adolescents in this study); the parents of the other half did not. You ask each adolescent to write a one-page essay entitled “Who am I?” For each essay you then count the number of times that the adolescent refers to him/herself as being “of value.” (Each of these counts represents the self-esteem score for one of the adolescents, where a large count indicates high self-esteem and a low count indicates low self-esteem.) The following are means and variances on the self-esteem measure for each group:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parents used physical punishment</td>
<td>8.3</td>
<td>16.5</td>
</tr>
<tr>
<td>Parents did NOT use physical punishment</td>
<td>12.3</td>
<td>9.8</td>
</tr>
</tbody>
</table>

a. Give the appropriate null and alternative hypotheses.

b. Assume for the moment that the means in your data are exactly equal to the population means to which they correspond. Given this assumption, what would be the power of the hypothesis test set up in part a? (Use the .05 significance level throughout.)

c. Of course, the above means and variances are only estimates. Use them to perform the hypothesis test set up in part a. What do you conclude? (Your conclusion must indicate more than “the theory is [not] supported” or “the hypothesis is [not] rejected.” And again, use the .05 $\alpha$-level.)

d. Reanalyze the above data by fitting a regression model with the self-esteem measure as the dependent variable and with a dummy variable as the independent variable. If the dummy variable were coded with a value of 1 for adolescents with parents who disciplined them using physical punishment and with a value of 0 for other adolescents, what would be the constant and the slope in the resulting bivariate regression equation?

e. What proportion of the variance in the self-esteem measure is explained by the dummy variable in the regression model described in part d?
2. In a study of child abuse in 10 countries, you collect the following data:

<table>
<thead>
<tr>
<th>Country:</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population density (X):</td>
<td>7.7</td>
<td>9.4</td>
<td>12.0</td>
<td>7.3</td>
<td>9.1</td>
<td>11.1</td>
<td>8.0</td>
<td>8.2</td>
<td>9.6</td>
<td>18.6</td>
</tr>
<tr>
<td>(in 10,000 people per square mile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child abuse (Y):</td>
<td>0.5</td>
<td>1.9</td>
<td>4.2</td>
<td>1.3</td>
<td>2.4</td>
<td>3.4</td>
<td>1.3</td>
<td>1.0</td>
<td>2.7</td>
<td>0.0</td>
</tr>
<tr>
<td>(instances per 10,000 people)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Plot the data and fit the model \( \hat{Y} = \hat{a} + \hat{b}X \).

b. Plot your regression line and state in words what your regression model suggests is the relation between population density and instances of child abuse in the ten countries. (Be sure to indicate both the direction of the relation and the units of all measures.)

c. Why might you believe that this model does not accurately reflect the true relation between the two variables?

d. Upon reexamining your data, you learn that country J reports no instances of child abuse because child abuse is legal there. Fit a new regression model in which you ignore the data from country J.

e. Plot this new regression line and state what this new regression model suggests is the relation between population density and instances of child abuse in the remaining nine countries.

f. What would you predict would be the amount of child abuse in country J, if they were to report actual child abuse rates?

g. Express \( \hat{a} \) and \( \hat{b} \) (from the regression model fit in part d) in words.

Some help:

data list records=1/popden 1-4 abuse 6-8.
begin data.
  7.7 0.5
  9.4 1.9
  12.0 4.2
  7.3 1.3
  9.1 2.4
  11.1 3.4
  8.0 1.3
  8.2 1.0
  9.6 2.7
  18.6 0.0
end data.
plot plot=abuse with popden.
regression variables=popden,abuse/dependent=abuse/enter.
3. Using the 1984 NORC data, find the correlation between “rincome” and “agewed”.

   a. Use algebra to convert this correlation coefficient to the regression coefficient from the regression of “rincome” on “agewed” (i.e., \( \hat{b} \) from \( \hat{Y} = \hat{a} + \hat{b}X \)).

   b. Standardize “rincome” and “agewed” and rerun the regression of “rincome” on “agewed” only using standardized variables this time. (You can get means and standard deviations using the “frequencies” command.) Verify that the unstandardized regression coefficient from this regression equals the correlation coefficient. Why does the constant in the standardized regression equal zero?

   c. Give a 95% confidence interval for the correlation.

Some help: When you standardize a variable, you subtract its mean and divide by its standard deviation. Although the following statement will NOT work for this problem, you will want to use a statement like it:

\[
\text{compute newvar} = (\text{oldvar} - 123) / 456.
\]

To get a correlation coefficient, you will need the command:

\[
\text{pearson corr agewed with rincome}.
\]

4. You are doing a content analysis of Hitler’s speeches during the last nine months of the Third Reich. You have counted the number of occurrences of the words “Jew” or “Jewish” in the speeches and wish to discover what relation these occurrences have to the number of deaths of Jews in German-run concentration camps.

\[
\begin{array}{cccccccc}
	ext{Occurrences (X):} & 74 & 95 & 121 & 74 & 92 & 110 & 83 & 83 & 98 \\
\text{Jews killed (Y):} & 3 & 22 & 40 & 9 & 22 & 30 & 31 & 32 & 42 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
\end{array}
\]

   a. Fit the model \( \hat{Y} = \hat{a} + \hat{b}X \) and test the null hypothesis that \( b = 0 \) both at the .05 and at the .01 levels.

   b. State in words what your regression model suggests is the relation between Hitler’s mentions of Jews and the numbers of Jews killed. (How many Jews were killed every time Hitler mentioned “Jew” or “Jewish” in his speeches?)
c. Historians working at the German National Archives in Koblenz have discovered a draft of Hitler’s “final” speech (never given) in which Jews were mentioned 180 times. What would you estimate would have been the number of Jews killed at the time of the speech? Calculate a 95% prediction interval for this estimate.

d. A second, revised version of the speech was found a few years later. In it Jews were mentioned only 90 times. Again, estimate the number of Jews that would have been killed at the time of the speech and calculate a 95% prediction interval for the estimate. Which of the two intervals (from ‘c’ or ‘d’) is larger? Why?

e. After doing all this analysis, a colleague points out that your data violate one of the assumptions of regression analysis. What assumption is this? Explain how your data violate the assumption.

5. “Polarization Theory” argues that mass media influence the public by raising issues that different segments of society feel strongly about. When members of society are (through the media) made aware of issues that they feel strongly about, they will attend more to the media. (This results as people prepare themselves to defend their feelings in conversations with people whose feelings differ from their own.)

You have data on how much money each of 49 subjects donated to agencies that take an active stand on the Nicaragua issue. Subjects were instructed only to indicate money that was donated during the year prior to the beginning of the Iran-Contra hearings. Your thinking is that the more people donated money to such agencies, the stronger were their feelings about whether the United States should provide military support to the rebel forces in Nicaragua. (That is, you believe “money donated” is a good measure of “strength of feelings.”) Second, you have data on the amount of time each subject has watched the Iran-Contra hearings on television. This is your measure of subjects’ attention to the media. Your data on these variables are as follows:

Table 1: Cell Frequencies on “Amount of Money Donated toward Agencies that Take an Active Stand on the Nicaragua Issue” and “Amount of Time Watching the Iran-Contra Hearings.”*

<table>
<thead>
<tr>
<th>Hours per week watching Hearings</th>
<th>Money Donated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$100-199</td>
</tr>
<tr>
<td>5-9</td>
<td>3</td>
</tr>
<tr>
<td>10-14</td>
<td>4</td>
</tr>
<tr>
<td>15-19</td>
<td>9</td>
</tr>
</tbody>
</table>

* Hypothetical data.
a. How much of the variance in hours of watching the Iran-Contra hearings is explained by money donated to agencies that take an active stand on the Nicaragua issue?

b. Is the proportion of explained variance large enough to be statistically significant at the .10 level of significance? SHOW YOUR WORK!

c. According to polarization theory, which of the two variables in Table 1 is independent? Which variable is dependent? (In each case explain how you can tell.)

d. Do the data support polarization theory? (Explain your answer.)

e. Find an equation that depicts the number of hours subjects would have watched on the basis of money donated to agencies that take an active stand on the Nicaragua issue. Express the constant and slope of this equation in words.

f. How many hours of the Iran-Contra hearings would you predict would have been watched by someone who donated $350 to agencies that take an active stand on the Nicaragua issue?

6. You are in charge of marketing research for Pepsi-Cola (a soft drink company with $120 billion dollars in assets). The company president has asked you to make a recommendation either for or against advertising Pepsi-Cola during the upcoming Olympic games in Korea. To help you in making your decision, the president tells you that advertising should only be recommended if the company’s annual profit (currently $1,500 for every advertising dollar spent in Korea) has increased from the current amount by at least 60% in the third year following the Korean Olympics. Based upon data from the Moscow Olympics (which took place in Russia), you have found that among companies that advertised at the games, large companies profit more per advertising dollar than smaller companies.

You have data on one hundred companies. Your SPSS data set includes two variables, namely “size” (size of company, measured in millions of dollars in assets) and “profit” (dollars profit earned in Russia during the third year after the Moscow Olympic games for each advertising dollar spent in Russia during the year of the Moscow Olympic games). You run the following SPSS commands:

```
compute q = size * profit.
frequencies general = size,profit,q / statistics = mean,variance.
```

Parts of your output look as follows:
### Statistics

<table>
<thead>
<tr>
<th></th>
<th>SIZE</th>
<th>PROFIT</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>53.84</td>
<td>12.97</td>
<td>897.45</td>
</tr>
<tr>
<td>Variance</td>
<td>451.20</td>
<td>1098.39</td>
<td>321119.31</td>
</tr>
</tbody>
</table>

a. How much of the variance in profit is explained by company size?

b. Is this a statistically significant amount of variance at the .05 level?

c. Give a 95% confidence interval for the correlation between profit and size.

d. Give the unstandardized regression equation appropriate to your research question and say in words what each regression coefficient means.

e. What profit would you project for Pepsi-Cola in three years if the company advertises in the Korean Olympic games?

f. Would you recommend to the president that Pepsi-Cola advertise in the Korean Olympics? (Use the .05 level of significance and be sure to take into account the president’s criterion for recommending advertisement.)

7. Rap music is believed to “corrupt the minds of youth.” This is because it contains so many curse words. As one community leader argues, “Once kids start to use that filthy language, their corrupted language then develops into corrupt behavior.” You decide to investigate whether or not young people are more likely to speak curse words if such words are used in their music.

You randomly sample 23 high school freshmen from a local school and obtain parents’ permissions (1) to obtain a listing of the titles of all records that the students own and (2) to tape record one hour of the students’ unsupervised conversations with friends in the student’s home. You derive two measures from these data:

**RAPCURSE** = The average number of curse words per hour in the lyrics of the student’s records.

**KIDCURSE** = The number of curse words mentioned by the student during the hour of unsupervised conversations.

You also create a new variable, **RAPKID**, by multiplying each student’s **RAPCURSE** score by his/her **KIDCURSE** score. Means and standard deviations on these three variables are as follows:
a. How much of the variance in KIDCURSE is explained by RAPCURSE?

b. Is this amount of variance statistically significant at the .05 level?

c. Give a 95% confidence interval for the correlation between RAPCURSE and KIDCURSE.

d. Give the unstandardized regression equation appropriate to your research question and say in words what each regression coefficient means.

e. Are the data consistent with the contention that curse words in lyrics may lead to youths’ cursing? Justify your answer in the light of your findings. (Hint: What two pieces of information are needed to justify your answer?)

8. You wish to know whether greater use of knee braces by football players has an effect upon the number of serious football-related injuries suffered by members of college football teams. Because football teams usually have policies concerning knee braces, you find that use of knee braces is NOT a random variable among members of the same football team. Since use of knee braces is (less arguably) random among football teams, you use “the football team” as your unit of analysis.

You have data on thirty football teams. Your SPSS data set includes two variables, namely “pbrace” (proportion of the football team comprised of players that regularly use a knee brace) and “injuries” (number of serious football-related injuries per 100 players suffered by the football team last year). You run the following SPSS instructions:

```plaintext
compute j = pbrace * injuries.
frequencies general = pbrace,injuries,j / statistics = mean,
variance.
```

Parts of your output look as follows:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>PBRACE</th>
<th>INJURIES</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.682</td>
<td>2.97</td>
<td>2.23</td>
</tr>
<tr>
<td>Variance</td>
<td>0.013</td>
<td>28.17</td>
<td>18.84</td>
</tr>
</tbody>
</table>

a. How much of the variance in injuries is explained by use of knee braces?
b. Is this (i.e., the amount found in part a) a statistically significant amount of variance at the .05 level?

c. Give the unstandardized regression equation appropriate to your research question and say in words what each regression coefficient means.

d. Would you recommend use of a knee brace to prevent serious football-related injuries? (Justify you answer by referring to your data.)

e. Test the null hypothesis that there is a slope of zero between the proportion of football teams comprised of players that regularly use a knee brace and the number of serious football-related injuries suffered by the teams. (Use the .05 level of significance.)

9. You supervise 20 social workers in a national welfare agency. During the past few years you have gained the impression that when your workers closely monitor child-abuse-families (i.e., families within which child abuse is suspected to be taking place), more child abuse may be taking place than when workers monitor these families less closely. (That is, monitoring and abuse appear to be positively associated.) Each of your workers keeps a log of when they visit each of the 75 child-abuse-families under your jurisdiction. Your workers also keep track of admissions of children from these families into local hospitals and health clinics for injuries that may be due to abuse. You have information on 23 hospital/clinic admissions of this kind. From these data you construct the following two measures:

CHANGE—the increase in monitoring from two weeks prior to one week prior to a child’s admission into a hospital or clinic (in number of uninvited visits more during the first week prior than during the second week prior to the admission) I.e., a score of 5 means that the admitted child’s family had five more uninvited visits by a social worker(s) during the week just prior to admission than it had during the week prior to this.

SEVERITY—the number of independently-inflicted injuries to the admitted child, according to the doctor who treated the child in the hospital/clinic.

These two variables have a correlation of $r = .4$ and their means and variances are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHANGE</td>
<td>3.5</td>
<td>5.0</td>
</tr>
<tr>
<td>SEVERITY</td>
<td>2.1</td>
<td>3.8</td>
</tr>
</tbody>
</table>

a. What is the unit of analysis in this analysis?
b. How much of the variance in injury severity is linearly explained by an increase in monitoring prior to hospital/clinic admittance?

c. Is the amount of variance found in part b statistically significant at the .05 level of significance?

d. Give the unstandardized regression equation appropriate to your research problem and state in words what each regression coefficient means.

e. How many independently-inflicted injuries would you estimate to be found on a child admitted to a hospital/clinic, whose child-abuse-family was visited without invitation by a social worker once during the second week prior to the admission and three times during the week prior to the admission?

f. Do your data provide evidence at the .05 significance level that there is an increase in the severity of child abuse within child-abuse-families when social workers’ monitoring of these families increases? (Hints: Although this has many similarities to the question asked in part c, it is not identical to it. Also, be sure and give a statistical justification for your answer.)

10. “Project Home” was organized to investigate the influence that family interactions have on the children of poor Iowa families. The theoretical underpinnings of the project lie in modeling theory and the idea of “supportive interaction.” A person is said to interact supportively with another if the person conveys the other’s importance to this other. For example, parents exhibit supportive interaction if the father communicates the mother’s importance to her and the mother communicates the father’s importance to him. According to modeling theory children learn how to interact with people by modeling (i.e., by imitating) the interactions of their parents. If parents frequently exhibit supportive interaction, their children will come to interact supportively with others. By modeling its parents’ supportive interaction with others, a child will stimulate others to model (i.e., interact supportively with) it. For example, the child’s teachers will be stimulated to model the child’s supportive interaction back to it. One form this can take is in the teachers’ conveying the child’s importance to it by assigning it high grades in school. Your theoretical argument is thus that it is through a two-step process of “children modeling parents and teachers modeling children” that parents’ supportive interaction leads to their children’s stellar performance in school.

You draw a simple random sample of 150 families from the Iowa State Labor Bureau’s list of all Iowa families on welfare. After initial telephone contacts with each of these families you eliminate 47 of the families from your sample, because they are single-parent families or families in which the oldest child was not of school age. Because some parents were either unwilling or unable to cooperate
with you in your research, you lose an additional 22 families from the sample. Your final sample size is thus of the remaining 81 (150 - 47 - 22) three-member family subgroups consisting of mother, father, and oldest child. From videotapes of parents’ interactions you obtain data on parents’ supportive interaction; from school records you obtain data on their oldest child’s school performance. Using an ordinary least squares regression program to regress school performance on supportive interactions you find an unstandardized slope of 0.5 between these two variables. Means and variances on the variables are as follows:

<table>
<thead>
<tr>
<th>Number of supportive interactions per hour</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oldest child’s school performance (on a 100 point scale)</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>

a. According to your theoretical argument which is the independent and which is the dependent variable in your analysis?

b. Find the constant in the unstandardized regression equation, and describe in words the meanings of the slope (i.e., 0.5 given above) and this constant.

c. What proportion of the variance in supportive interaction is explained by school performance? What proportion of the variance in school performance is explained by supportive interaction? (Hint: Be sure to answer both questions.)

d. Give null and alternative hypotheses for testing your theoretical argument.

e. Using the .05 significance level, test the hypotheses given in part d.

f. Using the regression equation described in part b (and assuming the child is an oldest child of school age) what would be the school performance of a child of Iowa welfare parents who exhibit 45 supportive interactions per hour?

g. Is the school performance found in part f an estimate or a prediction? Explain your answer.

h. Find the 95% confidence interval for the value of school performance calculated in part f.

11. One occupational hazard of being an exotic dancer is that members of your audience will often attempt to touch (or grope at) you while you are dancing. Such groping behavior is likely to occur during late evening hours when audience members’ drunkenness undermines the propriety of their behaviors. It is only as a last resort that dancers will use physical force (i.e., that they will strike an audience member) as a way to stop these attempts at groping. Instead, dancers
usually begin by trying to persuade the audience member not to touch the dancer.

Exotic dancers generally use one of two strategies of persuasion. On the one hand, the perception strategy is an attempt to alter how the audience member perceives the dancer. For example, the audience member might be persuaded that the dancer is like a relative whom one would never contemplate gropering. ("You wouldn’t speak to your sister in that way now, would you?") On the other hand, the threat strategy is an attempt to persuade the audience member that gropering will have dire consequences. ("See that big guy in the corner? He’s my boy friend . . .")

Your data are from field notes that you took while observing interactions between exotic dancers and audience members in two strip bars located in a small Midwestern city. You have information on the following two variables for each of 56 interactions:

**GROPES**  the number times of that an audience member attempts to grope the dancer

You have the following data on this variable (i.e., on GROPES):

<table>
<thead>
<tr>
<th>number of cases(n)</th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strip bar A:</td>
<td>31</td>
<td>2.80</td>
</tr>
<tr>
<td>Strip bar B:</td>
<td>25</td>
<td>4.06</td>
</tr>
<tr>
<td>Strip bars A and B combined:</td>
<td>56</td>
<td>3.50</td>
</tr>
</tbody>
</table>

**PERCEIVE/THREAT**  the number of perception strategies attempted by the dancer minus the number of threat strategies attempted by the dancer  (Note: This is a measure of how much more the dancer uses perception strategies than threat strategies to persuade an audience member not to grope.)

You have the following data on this variable (i.e., on PERCEIVE/THREAT):

<table>
<thead>
<tr>
<th>number of cases(n)</th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strip bar A:</td>
<td>31</td>
<td>-1.48</td>
</tr>
<tr>
<td>Strip bar B:</td>
<td>25</td>
<td>-1.52</td>
</tr>
<tr>
<td>Strip bars A and B combined:</td>
<td>56</td>
<td>-1.50</td>
</tr>
</tbody>
</table>

a. Do you have statistically significant evidence at the .05 significance level that audience members in one of the strip bars was more likely to attempt to grope dancers than audience members in the other strip bar?  (Hint: Note that
this question does not request any information about the PERCEIVE/THREAT variable.)

b. What is the probability that a Type II error was made in the conclusion drawn in part a regarding the frequencies at which groping is attempted in the two bars? (Hint: It is not necessary to know the true mean of the sampling distribution to answer this question.)

c. The correlation between GROPES and PERCEIVE/THREAT equals -.40. Given this finding, what proportion of the variance in audience members’ groping behavior is explained by dancers’ perception strategies versus threat strategies toward them? Is this proportion statistically significant at the .05 significance level?

d. Find the unstandardized regression equation for the regression of GROPES on PERCEIVE/THREAT, and express the meaning of each coefficient in words.

e. Imagine that you are giving advice to someone who is just beginning a career as an exotic dancer. You explain to her that your research suggests that perception strategies should be used and threat strategies avoided by exotic dancers who wish not to be groped. In response she asks you, “How many more perception strategies than threat strategies should I use to be sure that audience members will not try to grope me more than once?” You respond that you cannot answer her question with a number that will work 100% of the time, but that you are happy to give her one that should work 95% of the time. Note that in her question the dancer gives you the number, 1, when she uses the word, “once.” Also note that she asks you for a number, namely “how many more perception strategies than threat strategies” she should use. Indicate which of the following 8 statistical expressions corresponds to (1) the number she gives you and (2) the number she asks for. (Hint: You are being asked to circle two of the below expressions, and to label one as “the number given” and to label the other as “the number asked for.”)

\[ X_e \] \[ \hat{Y}_e \] \[ \hat{Y}_e + \Delta \] \[ \hat{Y}_e - \Delta \]
[ \[ X_p \] \[ \hat{Y}_p \] \[ \hat{Y}_p + \Delta \] \[ \hat{Y}_p - \Delta \]

12. In 1990 Iowa Beef Packers Inc. (IBP) opened a beef processing plant in Lexington, Nebraska (pop. 10,075). The city of Lexington had suffered economically during the agricultural crisis of the mid-1980s, leaving its city leadership easily convinced by IBP’s claims that residents would benefit from better-paying jobs and that local industries would benefit from increased demand for their products. Today the plant currently employs 2,400 and has an annual payroll of $58 million.
You are interested in exploring the economic impact that IBP’s plant has had on Lexington’s economy. Although the plant’s $58 million payroll seems large, when spread among 2,400 employees—and taking managers’ high salaries and workers’ overtime into account—one finds that the vast majority of the plant’s employees earn little more than minimum wage. Lexington residents complain that (in contrast to IBP’s promises prior to 1990) the plant has not provided better-paying jobs. Instead, it created thousands of minimum-wage jobs that have been filled by migrant workers, not by the long-term residents of Lexington.

Your first research question addresses the residents’ claim that there has been an increase in the number of migrant workers who work in Lexington. Data in the following table are from two random samples of Nebraska residents collected by the Gallup Organization’s Lincoln, Nebraska, office:

Table 2. Long-term and migrant residents of Lexington, Neb., in 1990 and 2000.

<table>
<thead>
<tr>
<th>Resident type</th>
<th>1990</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>long-term</td>
<td>120</td>
<td>54</td>
</tr>
<tr>
<td>migrant</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Be sure that you read this table correctly: For example, the upper left-hand cell in the table indicates that there were 120 respondents to Gallup’s 1990 survey who were long-term residents of Lexington, Nebraska.

a. State the null and alternative hypotheses for testing whether there has been an increase from 1990 to 2000 in the proportion of Lexington residents who are migrants.

b. If the percent of migrants among Lexington’s residents was 10% more in 2000 than in 1990, what is the probability that a Type II error might be committed in testing the hypotheses stated in part a? (Hint: Use the .05 significance level, plus use the data in Table 2 to estimate the variance of your sampling distribution.)

c. Given the hypotheses stated in part a, find the P-value for your findings as reported in Table 2. Do you have statistically significant evidence at the .05 level that there has been an increase in the proportion of migrant residents in Lexington? Justify your answer by using the P-value that you have just found. (Hint: In this part and in part b, please assume that the central limit theorem holds regarding the normal shape of your sampling distribution.)
Your second research question addresses IBP’s claim that building the beef processing plant would increase Lexington’s long-term residents’ incomes. You find that many residents’ incomes did increase in the early 1990s when their occupational status changed from unemployed (no income) to employed (some income from jobs at the IBP plant). However, you are interested in evaluating whether (after adjusting their incomes for inflation) workers’ incomes have increased during the years after they were first hired.

You randomly sample 84 long-term Lexington residents who were hired at least one year ago, and who have worked at the IBP plant continuously from the day they were hired up to the present. For each resident, you obtain data on the following two variables:

YEARSEMPLOYED  The number of years that the resident worked at the IBP plant in Lexington (Values on this variable range from 1 to 10 years.)

AVERAGEINCREASE  The resident’s average annual salary increase (in inflation-adjusted dollars) while working at the plant (Values on this variable range from -$5,000 to $5,000. Note that a negative average salary “increase” is possible when one’s salary increases have not “kept up with” inflation.)

A quick look at the data reveals a correlation of .23 between YEARSEMPLOYED and AVERAGEINCREASE. Moreover, YEARSEMPLOYED has a mean of 5 and a variance of 4, and AVERAGEINCREASE has a mean of 1,000 and a variance of 2,500.

d. Find the 95% confidence interval for the correlation between YEARSEMPLOYED and AVERAGEINCREASE.

e. Find the unstandardized regression equation for the regression of AVERAGEINCREASE on YEARSEMPLOYED. State the meaning of the unstandardized slope in words.

13. According to DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition), children with Attention Deficit Disorder (ADD) are recognized according to such symptoms as inattention to details, careless mistakes, not following directions, not finishing schoolwork or chores, being disorganized, and loosing things frequently. When complicated by hyperactivity, Attention Deficit Hyperactivity Disorder (ADHD) children are also likely to interrupt or intrude on others, talk excessively, and frequently fidget. In brief, ADD and ADHD children have difficulty controlling their behaviors within socially acceptable bounds. As
most research bears out, the cure of these children’s disorders lies in training
them to control themselves.

Given this research, one would expect that parents and teachers would work
together to cure ADD and ADHD children of their disorders. Instead, the
disorders have provided parents and teachers with a way to “explain” the
misbehaviors. (“The disorder, not my bad parenting/teaching, makes the child
misbehave.”) In other words, whereas some parents and teachers have viewed
these disorders as temporary conditions that can be cured, many parents and
teachers view them as permanent conditions to which they must adjust. Your
theory is that this latter view results as parents and teachers each absolve
themselves of blame for ADD/ADHD children’s misbehaviors by holding the other
responsible for it (i.e., parents blame teachers and teachers blame parents,
leaving both blamers free of responsibility for curing ADD/ADHD). Of course,
parents’ and teachers’ views of ADD and ADHD are not exclusively divided
between fleetingly temporary and resolutely permanent. To capture this range of
views, you ask a sample of 56 parents of ADD or ADHD children and a sample of
24 teachers (both from your school district) to answer the following questions:

“Children having an attention deficit disorder (i.e., ADD or ADHD) lack self-
discipline. On a scale from 0 to 9, where 0=totally ineffective and 9=totally
effective, please indicate how effective you believe that parents and teachers are
in instilling self-discipline into the ADD or ADHD children under their care?”

<table>
<thead>
<tr>
<th>Totally Ineffective</th>
<th>Totally effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parents</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>teachers</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

The three variables in your analysis are thus PAEFFECT and TEEFFECT
(respectively, parent’s and teacher’s effectiveness on a 10-point scale), plus WHO
(0=parent; 1=teacher). Your data on these variables are as follows:

<table>
<thead>
<tr>
<th>Correlation Coefficients</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAEFFECT</td>
<td>1.00</td>
</tr>
<tr>
<td>TEEFFECT</td>
<td>.31</td>
</tr>
<tr>
<td>WHO</td>
<td>.20</td>
</tr>
</tbody>
</table>

(Make sure you read the data in the above table correctly. For example, the zero-
order correlation between parents’ and teachers’ effectiveness is .31. The mean
and standard deviation on the parent’s effectiveness variable are 6.76 and 2.766
respectively.)

a. Find the unstandardized regression equation for the regression of parent’s
effectiveness (PAEFFECT) on teacher’s effectiveness (TEEFFECT). Express
the meanings of the slope and constant from this equation in words that a nonstatistician could understand.

b. Referring to the slope found in part a, find the $P$-value associated with the hypothesis that when teachers are viewed as more effective in instilling self-discipline into the ADD or ADHD children under their care, parents are also viewed as more effective in this way.

c. Find a 95% confidence interval for the parent’s effectiveness score being given by parents and teachers from your school district who report a teacher’s effectiveness score of 6.

d. Calculate two means using the data on correlations, means, and standard deviations: First, obtain the mean parent’s effectiveness score among the 56 parents in the sample. Second, obtain the mean parent’s effectiveness score among the 24 teachers in the sample. (Hints: You may wish to find the unstandardized regression equation for the regression of parent’s effectiveness on WHO [i.e., the regression of PAEFFECT on the dummy variable indicating whether it was a parent or a teacher who judged parent’s effectiveness]. Also please note that data on the TEEFFECT variable is not used here nor in either of the remaining parts of this problem.)

e. As given above, the standard deviation for the variable, PAEFFECT, equals 2.766. However, when you calculate the standard deviation for PAEFFECT just among the 56 parents, it equals 3.2. When you calculate the standard deviation for PAEFFECT just among the 24 teachers, it equals 1.9. Given these two within-group standard deviations and the means obtained in part d, test whether there is significant evidence that teachers view parents as more effective than parents view themselves. (Hints: Use the .01 [NOT .05] significance level throughout. And, please remember that this question asks nothing about teacher’s effectiveness as measured by TEEFFECT.)

f. State (in notation, please) the null and alternative hypotheses tested in part e. If in fact the average parent’s effectiveness score among teachers in your school district is 2 points greater than the average parent’s effectiveness among parents in the district, what is the power of this hypothesis test?

14. Status Quo Theory states that society provides an opportunity structure within which members of society achieve statuses commensurate with their abilities (i.e., incapable people end up in low-status jobs, whereas capable people end up in high-status jobs). Yet occupations held by women and members of ethnic minorities are typically low-status occupations, suggesting that women and minorities are less capable than nonminority men to achieve high-status jobs within a society’s opportunity structure. Although accepting the premise that societies offer their members occupational positions within an opportunity
structure, feminists (among others) refuse to accept the argument that ability determines these members’ ultimate positions within this structure.

Catastrophe Theory provides an alternative theoretical mechanism to account for how occupational positioning occurs: Life can be viewed as a series of catastrophes—events that interrupt goal-oriented activity. The more often behavior toward a particular goal is interrupted, the more implausible attainment of this goal will appear. In turn, this appearance of implausibility undermines people’s self-confidence, and (regardless of their innate abilities) undermines their motivation to achieve. Thus if people with identical abilities invested the same amount of time in achieving the same goal, those who were most interrupted in this achievement-activity would be less likely than the others to achieve the goal by virtue of their lower self-confidence (i.e., due to their having convinced themselves of their incapacity to achieve the goal). Feminists argue that this is why executives have secretaries (and husbands “have” wives). Women protect men from interruptions that could undermine their self-confidence (but unwittingly do so at a cost to their own self-confidence). Society-wide differences in achievement are self-perpetuated as boys are left “uninterruptedly” to be boys, but girls are continually interrupted in order that they might “responsibly” subsume their goals to (i.e., allow their goals to be interrupted by) those of others.

You research investigates Catastrophe Theory based on data from two samples of female alumni, each of whom gave birth to a child while attending graduate school at ISU during the 1990s. Your thinking is that these infant children are likely to have interrupted the women’s goal-attainment (namely, their achieving a graduate degree). One sample is of 31 alumni who completed their graduate degrees; the other sample is of 31 alumni who did not complete their graduate degrees. The two samples are “matched” so that neither group differs on the type of degree (MA, PhD, etc.) sought, the area of study (psychology, physics, etc.), number of children, etc. During face-to-face interviews, you ask each of the 62 women the following questions:

DEMAND: “How would you characterize your child’s demands on you when s/he was an infant? On a scale from 0 to 9, where 0=no demands (e.g., all care was provided by another parent or relative) and 9=constant demands, please indicate how demanding your infant was.”

HOURS: “Not including any time that you may have been interrupted for parenting or other household responsibilities, how many hours did you spend studying on a typical weekday when you were a graduate student?”

You also construct a third variable, DEGREE, for which 1=completed graduate degree and 0=did not complete graduate degree. (Note that 31 of your subjects
Alumni who completed degree (i.e., DEGREE=1) | Mean | Standard Deviation | Sample Size
--- | --- | --- | ---
DEMAND | 3.5 | 2.4 | 31
HOURS | 10.6 | 5.4 | 31
Alumni who did not complete degree (i.e., DEGREE=0) | Mean | Standard Deviation | Sample Size
--- | --- | --- | ---
DEMAND | 6.3 | 2.6 | 31
HOURS | 7.0 | 2.8 | 31
All alumni (i.e., DEGREE=1 or DEGREE=0) | Mean | Standard Deviation | Sample Size
--- | --- | --- | ---
DEMAND | 4.9 | 2.5 | 62
HOURS | 8.8 | 4.6 | 62

(Make sure you read the data in the above table correctly. For example, whereas the 31 alumni with a graduate degree spent an average of 10.6 hours studying on a typical weekday [with a standard deviation of 5.4 hours], all 62 alumni spent an average of 8.8 hours studying on a typical weekday [with an overall standard deviation of 4.6 hours].)

a. Do your data provide statistically significant evidence (at the .05 significance level) that alumni who completed their degrees had less demanding infants than alumni who did not complete their degrees?

b. A colleague points out to you that the statistical test performed in part a is inappropriate for evaluating Catastrophe Theory. Why would it be wrong to use the test in this way? What “more appropriate” test might your colleague have in mind?

c. Combining your data on all 62 alumni, you find that DEMAND and HOURS have a correlation of -.094 with each other. Using this correlation coefficient and data provided above, find the unstandardized slope from the regression of HOURS on DEMAND and express the meaning of this slope in words that a nonstatistician could understand.

d. Of course, alumni with demanding infants had less time to study. If you were to test this statement using the slope obtained in part c, what would be your null and alternative hypotheses?

e. State the rejection rule appropriate for testing the hypotheses that you gave in part d at the .05 significance level. Apply the rejection rule, and state your conclusion.

f. What would be the power of the hypothesis test performed in part e, if among all “female ISU alumni who gave birth to babies while in graduate school” (i.e., if among the population of all alumni from which your sample of
62 alumni was randomly drawn) the hours of these alumni’s study time actually decreased one hour for each additional point they reported on the DEMAND scale?

15. Couples argue. They sometimes fight. On occasion these fights escalate to the point that one member of the couple (usually a male) physically injures the other member (usually a female). When the injured (or about-to-be-injured) member contacts the police for protection, police officers refer to this as a potential case of “domestic violence.” The use of the word, “potential,” is central here, because upon arriving at the scene police officers must decide if they are faced with a “real” case of domestic violence or not.

Prior research has established that police officers’ identification with one or the other members of a couple is central in their decision that they are faced with a “real” case of domestic violence. For example, in cases in which a female alleges that a male threatens her with violence, female police officers (who presumably identify more with the female than the male member of the couple) are more likely than male police officers to decide that the case is (really) one of domestic violence. Some theorists have followed up on this line of thinking by noting that most police officers are themselves members of couples. Accordingly, if threats of violence are commonplace between them and the other member of their couple, they would be more likely than other police officers to identify with couples who threaten each other with violence.

There are two theories regarding how such “identification with the couple” might influence police decisions: Normative Theory suggests that such identification blinds police officers from seeing violence threats as nonnormative (e.g., as cases of domestic violence), because—generalizing from their own couple-experiences—such threats appear typical between members of couples. On the other hand, Sensitivity Theory suggests that such identification makes police officers more able to recognize the likelihood that threats will escalate into violent behavior. Thus if the former theory is true, identification will yield decisions that cases are NOT ones of domestic violence. If the latter theory is true, identification will yield decisions that they ARE ones of domestic violence.

The Los Angeles Police Department has recently initiated a program to reduce the influence of its police officers’ domestic problems on their job performance. As part of this initiative, you were hired to provide compulsory marital counseling to all of the department’s married male police officers and their wives. You collect your data from 94 officer-wife couples during interviews performed during this counseling, and from other police records. During the interviews you obtain data on the following variable:

THREATEN “All couples occasionally get mad at each other. Although these occasions may be unpleasant to recall, please write on this sheet of
paper the ten worst things you can recall having said to each other during the past year.” (Responses range from 0=“none of the ‘worst things’ comprise threats of violence” to 10=“all ten of the ‘worst things’ comprise threats of violence”.)

From police records, you obtain data on another variable:

DECIDE the proportion of “cases of potential domestic violence that a police officer decided on during the past year” that he decided to be real cases of domestic violence (For example, .5 means that half of the time the police officer decided that the case was a real case of domestic violence.)

There is a correlation of $r_{TD} = -.15$ between THREATEN and DECIDE. The variables’ means and standard deviations are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>THREATEN</td>
<td>3.12</td>
<td>1.214</td>
</tr>
<tr>
<td>DECIDE</td>
<td>.35</td>
<td>.377</td>
</tr>
</tbody>
</table>

a. Find the unstandardized OLS (Ordinary Least Squares) regression equation for the regression of DECIDE on THREATEN, and express the meanings of the slope and constant in words that a lay person could understand.

b. Assume that your objective is to evaluate BOTH Normative Theory and Sensitivity Theory based on the slope estimated in part a. Find the $P$-value associated with this slope.

c. Does the slope between THREATEN and DECIDE provide evidence in support of Normative Theory? Does it provide evidence in support of Sensitivity Theory? Is this evidence statistically significant at the .10 level of significance? Give arguments in support of your answers. (Hint: Keep in mind that theoretical support does not require statistical significance, and that statistical significance does not necessarily entail theoretical support.)

d. What would you estimate to be “the proportion of an officer’s cases that he decided were real cases of domestic violence” among officers who themselves or their wives mentioned 4 threats of violence among the ten worst things they could recall having said to each other during the past year?

e. Find a 95% confidence interval for the estimate found in part d.

16. As in high schools throughout the United States, Hispanic students in Iowa high schools tend to have lower grades than non-Hispanic students. Cultural
Impairment Theory (CIT) argues that this is because Hispanic students have a hard time understanding both their English-speaking teachers and their textbooks written in English. That is, their limited English skills impair their performance in high school. In contrast, Culture of Poverty Theory (CPT) argues that Hispanic students’ life circumstances lead them to expect that they, like their parents, will remain poor laborers in agriculture-related industries. As a result, they conclude that Hispanic high school students perform poorly because their aspirations are low (i.e., they do not work hard at their studies because they do not believe that these studies are relevant to the skills required by such laborers).

Your research objective is to evaluate the relative merits of these two theories. You obtain a random sample of 40 Hispanic Iowa high school students. In face-to-face interviews you obtain a response from each student to the following question:

ASPIRE “What I’m learning in high school will help me have a better life than the life my parents have now.” (Possible responses to this variable are 0=“strongly disagree”, 1=“disagree”, 2=“not sure”, 3=“agree”, and 4=“strongly agree”.)

During the interview you also obtain data on each student’s English proficiency. This variable, named PROFICIENT, has scores ranging from (1=no proficiency to 10=complete fluency). You also obtain each student’s grade point average from administrators at each school. These GPAs have the usual scores ranging from 0.0 to 4.0 (i.e., F to A). Below is a table of correlations, means, and standard deviations among these three variables:

<table>
<thead>
<tr>
<th>---Correlation Coefficients---</th>
<th>ASPIRE</th>
<th>PROFICIENT</th>
<th>GPA</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASPIRE</td>
<td>1.00</td>
<td>.10</td>
<td>.20</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>PROFICIENT</td>
<td>.10</td>
<td>1.00</td>
<td>.30</td>
<td>5.0</td>
<td>2.0</td>
</tr>
<tr>
<td>GPA</td>
<td>.20</td>
<td>.30</td>
<td>1.00</td>
<td>2.5</td>
<td>.8</td>
</tr>
</tbody>
</table>

(Notes: Please be sure that you understand these data. For example, the correlation between ASPIRE and PROFICIENT is .1. The mean value of ASPIRE is 2 agreement points; the standard deviation of ASPIRE is 1 agreement point. Moreover, you may assume that ASPIRE, PROFICIENT, and GPA are all interval-level scales with agreement points, proficiency points, and grade points as their respective units.)

a. What proportion of the variance in ASPIRE is explained by PROFICIENT?

b. Find the 95% confidence interval for the correlation between PROFICIENT and GPA. (Show your work!)
c. Assuming PROFICIENT to be a measure of students’ “English skills” and GPA to measure their “learning abilities,” do your data provide statistically significant evidence (at the .05 significance level) in support of Cultural Impairment Theory?

d. Find the unstandardized regression equation for the regression of GPA on ASPIRE.

e. Express the slope and constant found in part d in words that a nonstatistician could understand.

f. If you were to evaluate Culture of Poverty Theory based on the regression slope obtained in part e, what would be your null and alternative hypotheses?

g. Using the .05 significance level, give the rejection rule for testing the hypotheses given in part f.

h. If the slope between GPA and ASPIRE were actually b=.25, what would be the power of testing the hypotheses given in part f at the .05 significance level?

17. Getting hired involves making a good impression during a job interview. Wear jeans, arrive late, lack interest, ... and you’ll not be hired. In the eyes of the person conducting the interview, you will lack the critical quality of “professionalism.” And no matter how stellar one’s résumé, no company wants to hire someone who is unprofessional.

A haircut is also part of one’s professional self-image. In the business world, a man’s professional haircut is relatively uniform: clean, combed, and short (but not shaved). To appear professional, a woman’s hair should also be clean and combed. However, there is considerable variety in professional women’s hair styles and lengths. And some hair styles are closely linked to a woman’s race. For example, without investing the time and expense needed to straighten her hair, an African American woman with long hair will almost inevitably have “Afro” as her hair style.

You obtain access to the personnel records at a large financial corporation in Chicago. In each employee’s file is a photograph of the employee at the time she or he was hired. You randomly sample files on 41 African American women and 41 Caucasian women from among the company’s records. Based on the photographs in these files you estimate each woman’s hair length when she was hired. The average hair length among the African American women is 5 inches (with a standard deviation of 2 inches), and the average hair length among the Caucasian women is 6.5 inches (with a standard deviation of 3 inches).
Your premise is that only women with “professional-looking hair” were hired (i.e.,
the hair in the fotos of all 82 women in your sample are professional-looking).
Moreover, you hypothesize that “Afro” haircuts tend not to be professional-
looking, so the African American women who were hired will tend to have had
shorter (and thus non-Afro) hair than did the Caucasian women who were hired.

a. What would be the null and alternative hypotheses used in testing whether
or not at the time when each woman was hired, African American women in
the company had shorter hair than did the company’s Caucasian women-
employees.

b. Test the hypotheses given in part a, using the .05 significance level
throughout.

c. What is the $P$-value associated with your finding in part b that the African
American women’s average hair length was $1\frac{1}{2}$ inches shorter than the
Caucasian women’s average hair length?

d. The $P$-value found in part c is a probability. (More specifically, it is the
probability that something happens under specific conditions.) State the
meaning of this $P$-value in one or more sentences that contain the number that
you calculated in part c.

CONTEXTUAL INFORMATION FOR PARTS e THROUGH g:
After examining the personnel folders more closely, you discover that the 41
Caucasian women in your sample were hired during the last 25 years, whereas
nearly all of the African American women were hired fairly recently. You decide
to examine the effects of “year hired” on “hair length,” and continue your
analysis with only the Caucasian women. (That is, from this point forward your
sample size is n=41, and your data are only on female Caucasian employees.)
The correlation between “year hired” and “hair length” is .10 and the year hired
variable has a mean of 1991 and a standard deviation of 4 years.

e. Based on your data, what was the hair length of Caucasian women who
were hired into the company in 1989?

f. Find a 95% confidence interval for the hair length found in part e.

g. After examining a plot of your data, a colleague notes that, following the
fashions of the times, the Caucasian women’s hair lengths tended to be
shorter in the early 1990s and longer in both the early 1980s and early 2000s.
Given this, she claims that one of the assumptions of regression analysis is
violated. Do you agree? If so, which assumption is violated? What aspects of
the plot suggest that this is (or is not) the case?
18. When one thinks of the Ku Klux Klan or of other white supremacist organizations, one tends to think of poor, uneducated, Southern, fundamentalist Caucasians who find scapegoats (Jews, African Americans, Washington liberals, etc.) for their problems. Yet if one looks beyond this stereotype, one finds that many white supremacists are actually quite wealthy. Moreover, recent studies have shown the wealthiest members of supremacist organizations to be the source of funds for the organizations’ recruitment of new members.

Of course, recruiting people into an organization is not as simple as “paying them to sign up.” Recruitment typically begins as a member befriends a young person (the member’s “target”) who is disenchanted with her or his life situation. Only after having first built a strong friendship with the target, does the member (or “recruiter”) gradually start implicating scapegoats. Thus friendship-building is the first step toward recruitment into a white supremacist organization. And money is an important means with which members build friendships with their targets.

Your research is based on face-to-face interviews with a random sample of 62 current members of the KKK (Ku Klux Klan) who recruited fellow high school students into the KKK when they were teenagers. (You obtained these members’ names and addresses from archives at the KKK National Headquarters in Harrison, Arkansas.) Your analysis is of these members’ responses to the following three interview questions:

**WAIT (W)** “After deciding to recruit a fellow high school student to the KKK, how many weeks did you typically wait before starting to talk with this student about how Jews, Blacks, and liberals are the source of their problems?” (Responses were coded in units of “weeks,” such that 0=“within one week,” 10=“during the 10th week,” and 52=“about one year.”)

\[
\bar{W} = 30 \\
\sigma_W^2 = 400
\]

**SUCCESS (S)** “Of the students that you tried to recruit to the KKK during high school, what percentage of these students actually joined the KKK as a result of your recruitment efforts?” (Responses ranged from 0=“none” to 50=“half” to 100=“all.”)

\[
\bar{S} = 65 \\
\sigma_S^2 = 200
\]

**MONEY (M)** “Did either of your parents give you money for the purpose of convincing your friends to join the KKK?” (Responses were either 1=“yes” or 0=“no.”)

\[
\bar{M} = 0.500 \\
\sigma_M^2 = 0.254
\]

These variables’ means and variances are provided under their variable names.
a. A colleague hypothesizes that parents only provide their children with money for recruitment to the KKK after these children have shown a propensity to successfully recruit fellow high school students to the KKK. You point out that the literature suggests that parents’ money is the cause not the effect of their children’s successful recruitment efforts. Nonetheless, you give your colleague access to your data so that she can test her own hypothesis. What statistical technique would you recommend that she use?

b. Among the 31 KKK members whose parents gave them money (i.e., for whom M=1), the mean and variance estimates of SUCCESS are $\bar{S}_1 = 70$ and $\hat{\sigma}^2_{S_1} = 125$ respectively. Among the other 31 KKK members (i.e., those for whom M=0), the mean and variance estimates of SUCCESS are respectively $\bar{S}_0 = 60$ and $\hat{\sigma}^2_{S_0} = 225$. Using the .05 significance level throughout, test whether or not high school recruiters for the KKK were more successful in their recruitment efforts if their parents gave them money than if their parents did not give them money for this recruitment.

c. You find that there is a correlation of .2 between WAIT and SUCCESS (i.e., $r_{WS} = .20$). Given this, obtain the unstandardized regression equation appropriate for testing whether or not the longer that high school KKK recruiters waited before implicating scapegoats (e.g., Jews, Blacks, and liberals), the more successful they were in their recruitment efforts.

d. State in words the meaning of the slope estimate obtained in part c.

e. Working with the slope estimate obtained in part c, test the hypothesis that the longer that high school KKK recruiters waited before implicating scapegoats (e.g., Jews, Blacks, and liberals), the more successful they were in their recruitment efforts. (Use the .05 significance level.)

f. If the true value of the slope estimated in part c were actually $b = .2$, what would be the power of the hypothesis test performed in part e?

19. People generally presume continuity. For example, if I notice a dog sitting in my front yard at various times during an afternoon, I might conclude, “That dog has been sitting in my yard all afternoon,” even though I may have only seen it sitting there 3 or 4 times within a 6 hour period. The principle here is that people tend to presume “no change” occurs between similar experiences. Psychologists call this tendency, repetition blindness (RB)—a phenomenon they investigate by presenting subjects with two consecutive occurrences of the same item within a rapid serial visual presentation (RSVP). Evidence of RB consists in subjects’ failure to detect (i.e., their blindness to) the second of these occurrences. Similarly, semantic repetition blindness (SRB) is a failure to detect two conceptually related items (e.g., a picture of a tree and the word, “tree”).
You are conducting a SRB study in which the items are words, and the type of conceptual relation among these items is the words’ language (namely, Spanish versus English). For example, the word, “libro” has the same meaning in Spanish as “book” does in English. An instance of SRB occurs when a subject fails to detect the second in such a pair of words during RSVP. Your hypothesis is that SRB will be more prevalent among people who are bilingual in Spanish and English, than among people who are monolingual in either Spanish or English. Your reasoning is that since monolingual subjects are incapable of conceptually relating such pairs of words, they will be less likely to experience semantic blindness than your bilingual subjects.

In a preliminary analysis of 31 bilingual (in Spanish and English) and 41 monolingual (in English only) subjects, you obtain the following means and variances for these subjects’ SRB instances during your study’s RSVP:

Table 1. Means and variances in numbers of “SRB instances during RSVP” among 31 bilingual and 41 monolingual subjects

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Variance</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual (B)</td>
<td>8</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>Monolingual (M)</td>
<td>6</td>
<td>2</td>
<td>41</td>
</tr>
</tbody>
</table>

a. Give the null and alternative hypothesis appropriate for testing your hypothesis (from this problem’s second paragraph) using the data in Table 1.

b. Test your hypothesis (from the second paragraph) in accordance with the null and alternative hypotheses that you gave in part a. (Hint: Use $\alpha = .05$ throughout.)

IMPORTANT: The rest of this problem only deals with your 31 bilingual subjects. Please make no more use of (i.e., ignore, forget, stop using, etc.) the data on your monolingual subjects.

c. Each of the 31 bilingual subjects in your study is a student who studies Spanish at a local community college. Consequently, you believe that there is likely to be a lot of variation in how “bilingual” these students actually are. (This variation may be why the variance in SRB instances is larger among your bilingual than your monolingual subjects.) To measure these 31 students’ bilingual ability, you obtain data on their grade point averages (GPAs) from the community college. (Whereas a student with a GPA of 4.0 in Spanish may be strongly bilingual and thus susceptible to SRB, a student with a zero GPA in Spanish may be nearly monolingual and thus not susceptible to
SRB.) After obtaining GPAs on all 31 of your bilingual subjects, you find that the correlation between their GPAs and their SRB instances is .15 (i.e., $r_{GS} = .15$). Find a 95% confidence interval for this correlation.

d. What proportion of the variance in the bilingual subjects’ SRB instances is explained by their GPAs? Is this proportion significantly large at the .05 significance level? (Explain your answers!)

e. Given that the mean GPA score is 2.5 and the variance is 0.7 among the 31 bilingual subjects, calculate the unstandardized regression equation (i.e., constant and slope) for the regression of their SRB instances on their GPAs.

f. State in plain English (i.e., as explained in class) what the constant and slope from part e mean.

20. In 1995 the Iowa State Legislature established the Iowa Sex Offender’s Registry to allow background checks on people, who apply for jobs at schools and other organizations that work with children. After a prospective employer receives a job application, the application is forwarded to the Iowa Division of Criminal Investigation (IDCI) for processing. Workers at the IDCI’s Identification Section look for matches between data on the application and data in their records. Of course, it is not uncommon for a convicted sex offender to use an alias rather than his or her own name. As a consequence, there may be many things that match between application and records, although the names on each may not match. In such cases employers are asked to request other surnames from the applicant. The hitch is that background checks cost $43.25 per surname! So if the employer (presumably to contain costs) refuses to inquire for additional surnames or if the applicant (possibly to prevent detection) provides no additional ones, a “near match” may be ignored and a sex offender may gain employment that provides access to children.

In an attempt to encourage employers to provide additional surnames when such “near match cases” arise, the IDCI changed its fee system three years ago so that $43.25 is charged per job application. This way many surnames might be provided for the same application, and the cost of checking a single job applicant’s criminal history never increases beyond $43.25.

You are an employee at the IDCI’s Identification Section. After discovering that you have been studying statistics, your boss asks you to investigate whether or not the recent change in the IDCI fee system has increased the proportion of employers who provide additional surnames when requested. During the year
before the fee system change you randomly sample 196 job applications for which the IDCI requested additional surnames; during the year after the change you sample 144 such applications. You search your records for data on these 340 applications and record whether or not in each case the employer provided additional surnames for the applicant.

a. Give the null and alternative hypothesis appropriate for testing whether or not the recent change in the IDCI fee system has increased the proportion of employers who provide additional surnames when requested.

b. You find that during the year before the fee system change there were 49 applications for which employers provided additional surnames and 147 applications for which they did not. During the year after the change you find 45 applications for which employers provided additional surnames and 99 for which they did not. Using the .05 significance level, state the rejection rule for testing the hypotheses stated in part a.

c. Apply the rejection rule obtained in part b. Do you have statistically significant evidence that the recent change in the IDCI fee system has increased the proportion of employers who provide additional surnames when requested? Explain your answer.

d. Imagine that the true proportions of employers who provide additional surnames are actually .23 before the recent change in the IDCI fee system and .33 after the change. What is the power of the hypothesis test performed in parts a to c? Show your work.

IMPORTANT: The rest of this problem deals with new data. Please make no more use of (i.e., ignore, forget, stop using, etc.) the data on employers’ providing additional surnames.

e. The IDCI is concerned that online companies are often used to perform background checks on prospective employees. For this reason the IDCI has developed a series of educational programs to increase attendees’ awareness of the advantages of IDCI background checks over those available online. Its most recent program was developed for Iowa elementary school principals. Last month the program was piloted (i.e., tried out) on a group of 28 attendees at a meeting for elementary school principals. Before the program began you asked each of these principals, “How many times during the past year was an online background check performed on a prospective employee at your elementary school?” During follow-up phone calls you later asked each of the
principals, “Of the background checks performed last year on prospective employees at your elementary school, what percent of these checks do you believe should have been performed by the IDCI?” You find a correlation of .40 (i.e., $r_{TP}=.40$) between “times online background check was performed” (T) and “percent of background checks that IDCI should have performed” (P). Find a 95% confidence interval for this correlation.

f. Find the standardized regression equation for the regression of P on T (i.e., for the regression of the standardized dependent variable, $Z_P$, on the standardized independent variable, $Z_T$).

g. State in plain English (i.e., as explained in class) what the slope from part f means.

21. There are approximately 165 organic farmers in Iowa who provide production animals (pigs, chickens, cows, sheep, goats, …) for the food industry. Like all farmers with livestock, these farmers often require veterinary services. Yet there are at least two reasons why they hesitate to request such services. First, few veterinarians are trained in organic methods to care for production animals. For example, organic farmers’ veterinarians must be familiar with alternatives to antibiotics when treating bacterial infections. Second, farmers’ fear of losing their organic certification makes them hesitant to report any problems on their farms.

As a researcher affiliated with ISU Extension, you have been asked to investigate factors associated with Iowa farmers’ use of veterinary services. Your study uses data from a survey mailed to random samples of 25 Iowa organic pig farmers and 31 Iowa nonorganic pig farmers. Your initial analysis makes use of the following two questions from the survey:

REQUEST (R) = During the past year, how many times did you request veterinary services for one or more of your pigs?

NUMSOLD (N) = How many pigs did you sell last year?

Means and standard deviations on Iowa organic vs. nonorganic farmers’ responses to these two questions are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Means</th>
<th>Standard Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Organic</td>
<td>Non-organic</td>
</tr>
<tr>
<td>REQUEST (R)</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>NUMSOLD (N)</td>
<td>252</td>
<td>963</td>
</tr>
</tbody>
</table>
a. What null and alternative hypotheses would you use to test whether or not among Iowan pig farmers, organic farmers requested veterinary services less frequently than nonorganic farmers?

b. Using the .05 significance level throughout, test the hypotheses stated in part a. Be sure to state your conclusion in clear English!

c. Assuming that last year among Iowan pig farmers, nonorganic farmers requested veterinary services an average of once per month more than organic farmers, what is the power of the test evaluated in part b?

d. Of course, it is likely that the more pigs farmers have, the more veterinary services they will use. And so after restricting your analysis to the organic-farmers-subsample, you find the correlation between REQUEST and NUMSOLD to be .34 (i.e., $r_{RN} = .34$) among these 25 farmers. Find the unstandardized equation for the regression of REQUEST on NUMSOLD among Iowan organic farmers only.

e. Referring to the table provided at the beginning of this problem, an average of 963 pigs were sold by the 31 nonorganic farmers in your other sample. Based on the equation obtained in part d, how many requests for veterinary services would an Iowan organic farmer who sold 963 pigs have made last year?

f. Obtain a 95% confidence interval for the number of requests calculated in part e. Show your work!

g. Applying the confidence interval obtained in part f, would 36 requests for veterinary services (i.e., would the average number of requests from the nonorganic farmers) be more than one would expect to result from sampling error among random samples of 25 Iowan organic pig farmers? Explain your answer!