Homework #5: Linear Relationships between Variables (Correlation)

****Due on day of Exam 1. If you have questions, please see either me or your TA for SPSS questions.**

Confidence Intervals: Questions 1-2

Correlation: Questions 3-5

SPSS: Bivariate Correlation

After completing the problems, do this SPSS practice problem by entering the data from Problem 4 (Psychotherapist Empathy and Patient Satisfaction) into SPSS.

A. Use either

Graphs → Legacy Dialogs → Simple Scatter

OR

Graphs → Chart Builder → Simple Scatter

to create a scatterplot of the relationship between the predictor and criterion variables. Add the regression line to the graph in order to familiarize yourself with the concept of linear models.

a. Indicate whether computing a correlation coefficient it would be appropriate.
b. Explain the appropriateness of using a linear model to further examine with the Pearson correlation procedure.

B. Use **Analyze → Correlate → Bivariate** to run a bivariate correlation in SPSS. Examiner the **Pearson Correlation Coefficient** in the output.

a. On your correlation output, identify and label the correlation coefficient (*r*) as well as the significance level (*p*-value).
b. Explain what the numbers actually mean in terms of the variables of the study. In other words, state you interpretation that is specific to these data.

Confidence Intervals

1. Sports scientists sometimes talk of a "red zone", which is a period during which players in a team are more likely to pick up injuries because they are fatigued. When a player hits a red zone, resting for a game or two is a good idea to prevent injury. A sports team measured how many consecutive games the players could play before hitting the red zone. The number of games played before hitting the red zone for 11 players were:

6, 17, 7, 3, 8, 9, 4, 13, 11, 14, 7

Using these data:

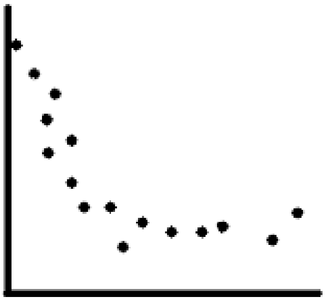
a. Calculate the *sample mean*.
b. Calculate the *standard deviation for the sample*.
c. Calculate the *95% confidence interval* for the mean. Assume $\sigma_{\bar{X}} = 1.31$

2. You and a friend are discussing a research project. She tells you that a 95% CI means there is a 95% chance the mean of the population is within the parameters of your calculated CI. Is she correct?

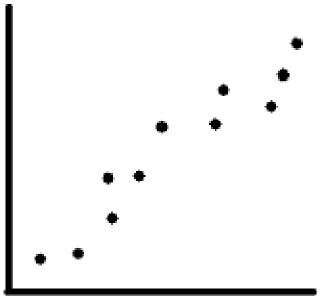
Correlation

3. For each of the following scatter diagrams, indicate whether the pattern likely indicates a linear relationship, curvilinear relationship, or no relationship; if the relationship is linear, indicate whether the relationship is either positive or negative and then specify the approximate strength (weak, moderate, strong) of the correlation.

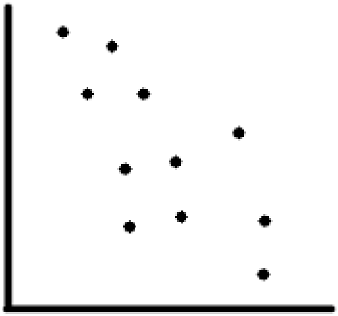
a)



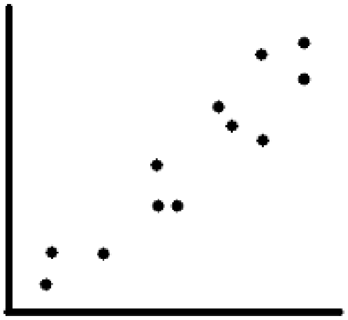
b)



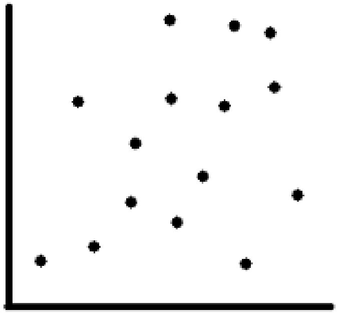
c)



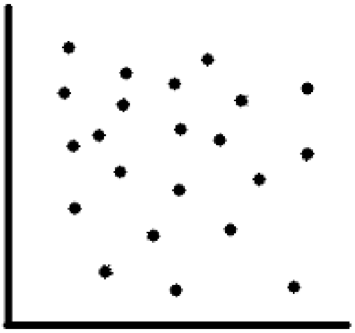
d)



e)



f)



4. A researcher studied the relationship between psychotherapists' degree of empathy and their patients' satisfaction with therapy. Four patient-therapist pairs were studied. Here are the data from:

Pair Number	**Therapist Empathy**	**Patient Satisfaction**
1	70	4
2	94	5
3	36	2
4	48	1

   a. Make a *scatterplot* of the scores by hand. This is based on only 4 scores, so it's easy.
   b. Based on your scatterplot, describe in words the general pattern of relationship (*kind* and *magnitude*) or correlation, if any.
   c. Calculate the Pearson *r correlation coefficient* using the formulae.
   d. Note the *magnitude* of the correlation coefficient and the sample size. Make some guess about it being statistically significant (or not) for a two-tailed hypothesis test with $\alpha = 0.05$.
   e. Provide a measure of *effect size* for these data and identify another name for it.
   f. Provide three logically possible directions of causality, saying for each whether it is a reasonable direction in light of the variables involved (and why).

5. For the following situation, indicate why the correlation coefficient might be a distorted estimate of the true correlation. Then provide the name that describes this problem. What effect might this problem have on a correlation?
   a. *Comfort of living situation* and *happiness* are correlated among a group of millionaires.

Homework #5: Linear Relationships between Variables (Correlation)
Answers

Confidence Intervals

1. **Red Zone**
   a. $\overline{X} = 9.00$
   b. $SD = \sqrt{\dfrac{SS}{N}} = \sqrt{\dfrac{SS}{11}} = \sqrt{\dfrac{188}{11}} = \sqrt{17.09091} = 4.134115$
   c. 95% CI

   Remember, we need to use the *standard error of the mean* for our CI.

   Standard error $\sigma_{\overline{X}} = 1.31$

   $95\%\ \text{CI}: \overline{X} \pm (1.96)(\sigma_{\overline{X}})$

   Upper limit = $\overline{X} + (1.96)(\sigma_{\overline{X}}) = 9.00 + (1.96)(1.31) = 11.5676$ or 11.57

   Lower limit = $\overline{X} - (1.96)(\sigma_{\overline{X}}) = 9.00 - (1.96)(1.31) = 6.4324$ or 6.43

2. **Define a CI**

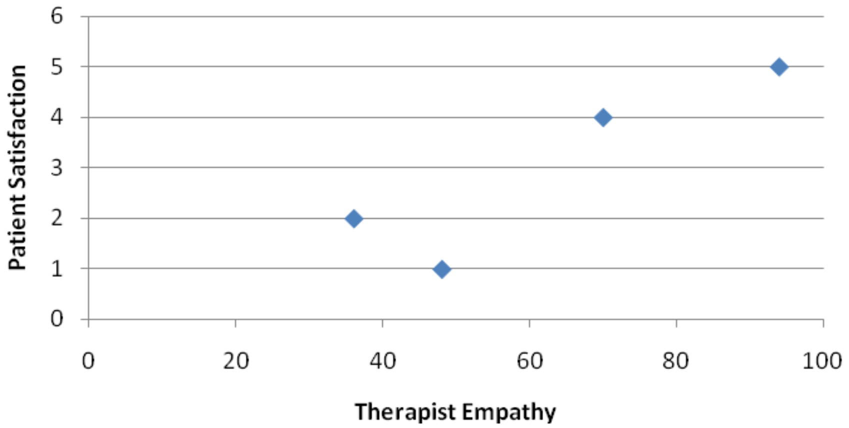
   A good response would be something like: "The correct way to understand the level of confidence, usually 95%, is in relation to indefinitely many replications of an experiment, all identical except that a new random sample is taken each time. If the 95% CI is calculated for each [replication of the] experiment, in the long run 95% of these intervals will include the population mean **μ**, or other parameter being estimated. For our sample, or any particular sample, the interval either does or does not include μ…" (Cumming & Fidler, 2005, p. 88).

Correlation

3. **Scatter Diagrams**
   a. Curvilinear
   b. Linear, positive, fairly strong
   c. Linear, negative, fairly strong
   d. Linear, positive, fairly strong
   e. If anything, linear, positive, weak to moderate
   f. No apparent linear correlation

4. Psychotherapists and Patient Satisfaction

a. Create a scatter diagram.



b. Describe the correlation.

The general pattern is *linear* and indicates a *positive correlation between therapist empathy and patient satisfaction*. This example represents a unique case for which correlating variables makes sense because the data are associated; in this case there is a "pair" of data. Even though the people measured here are different individuals, the "XY data pair" provides measure for both X (therapist empathy) and Y (patient satisfaction) variables, which are linked. If we had measured husbands and wives on their happiness, we could derive a correlation for them as well.

c. Calculate the correlation coefficient.

Therapist Empathy (X)	$\left(X-\bar{X}\right)$	$\left(X-\bar{X}\right)^2$	Patient Satisfaction (Y)	$\left(Y-\bar{Y}\right)$	$\left(Y-\bar{Y}\right)^2$	$\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$
70	8	64	4	1	1	8
94	32	1024	5	2	4	64
36	-26	676	2	-1	1	26
48	-14	196	1	-2	4	28
$\bar{X}=62$		$SS_X=1960$	$\bar{Y}=3$		$SS_Y=10$	$SP=126$

$$r=\frac{\Sigma\left[\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)\right]}{\sqrt{SS_X SS_Y}}=\frac{126}{\sqrt{(1960)(10)}}=\frac{126}{140}=.90$$

d. Determine if the correlation is significant.

Compare the *p* value provided in the SPSS table to your chosen *alpha level*.

e. Effect size for bivariate correlation is simply r^2, or the *coefficient of determination*. This is the complement of the $1-r^2$, or the *coefficient of non-determination/alienation*.

f. Provide three plausible directions of causality.

i. If a therapist has more empathy, this causes the patient to feel more satisfied (greater empathy might cause satisfaction to increase).

ii. If a patient feels more satisfied, this causes the therapist to feel more empathetic toward the patient (greater satisfaction might cause empathy to increase). **Correlation coefficients do not provide information about the direction of causality.**

iii. Some third factor, such as a good match of the patient's problem with the therapist's ability, causes both patients to be more satisfied and therapists to be more empathic toward their patients (some third factor causes both empathy and satisfaction). **The "third-variable problem".**

If your answers were not identical with those provided above, that is fine. The point of the problem is to get you to think critically about the relationship between variables in terms of causality and directionality. When using correlation rather than conducting an experiment, we cannot easily determine what variables are the causes and which are the effects.

5. Distorted Correlations

a. When calculating the correlation coefficient for *comfort* and *happiness* for millionaires, you have the problem of a **restriction in range (aka truncated range**). Among millionaires, there may not be a lot of variability in comfort of living situation (they probably all have quite comfortable situations). The correlation between *comfort of living* with any variable (including *happiness*) can be limited as a function of a restriction in the range of values for the millionaire population.

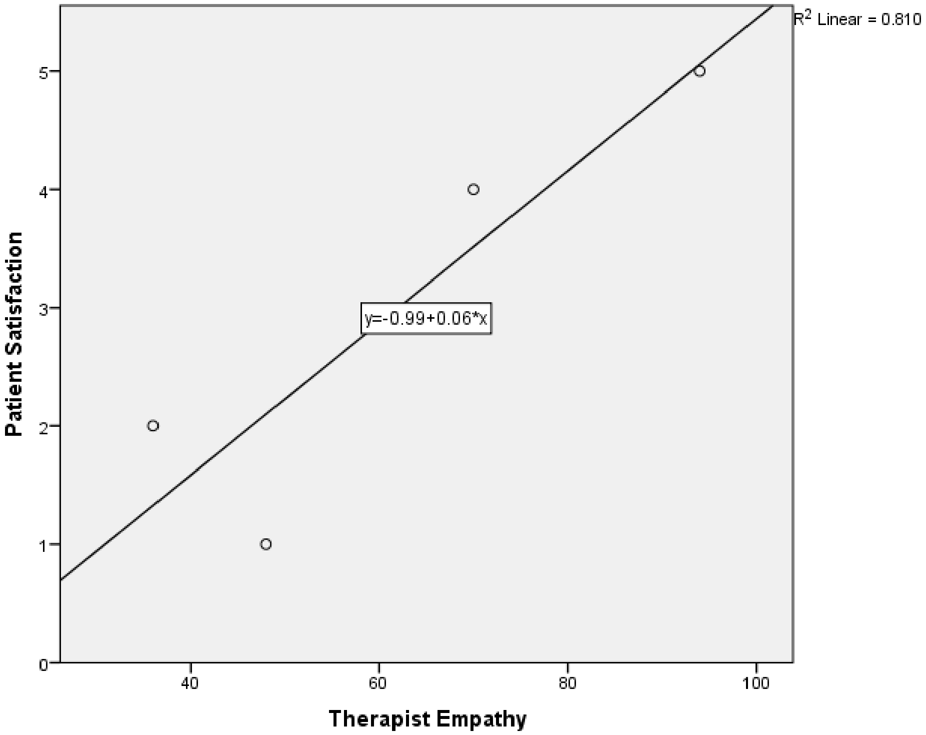
SPSS Bivariate Correlation

Syntax:

```
GRAPH
 /SCATTERPLOT(BIVAR)=TherapistEmpathy WITH PatientSatisfaction
 /MISSING=LISTWISE.

CORRELATIONS
 /VARIABLES=TherapistEmpathy PatientSatisfaction
 /PRINT=TWOTAIL NOSIG
 /MISSING=PAIRWISE.
```

A.



The above scatterplot indicates that a *linear* relationship might exist between the two variables. More specifically, higher values of *patient satisfaction* correspond with higher values of *therapist empathy*; lower values of *therapist empathy* correspond with lower values of *patient satisfaction*. The data points cluster around the regression line, indicating the linearity and strength of the relationship. Based on the scatterplot, we conclude that computing a correlation coefficient (measuring a linear relationship) is appropriate in order to determine whether a statistical linear correlation exists between therapist empathy and patient satisfaction.

B.

Correlations

		Therapist Empathy	Patient Satisfaction
Therapist Empathy	Pearson Correlation	1	.900
	Sig. (2-tailed)		.100
	N	4	4
Patient Satisfaction	Pearson Correlation	.900	1
	Sig. (2-tailed)	.100	
	N	4	4

Correlation Coefficient
r = 0.90

Statistical Significance level
p = 0.10

The *correlation* between therapist empathy and patient satisfaction is .90, which indicates a strong positive relationship. The *p-value* is .10. If we chose an alpha level of either .05 or .01, we would retain the null hypotheses because $p > \alpha$. Note also that the test of statistical significance is 2-tailed, which is recommended. We would then claim that the size of the correlation, given the sample size, likely comes from a population with a mean *r*=0, which we would expect under the null hypothesis.

This example also illustrates the fact that rejecting the null hypothesis is difficult when you have small sample sizes. In other words, because small sample sizes are often biased measures, we need a lot of evidence (e.g., very, very big correlations) to reject the null hypothesis.