Practice for Quiz #2 Story Problems

Set #1 -- Answers Follow

1. I’d like to collect data from both graduate and undergraduate students at UNL that examines the type of job they would like to have in the future (academic vs. business), their GPA, and obtain a measure of achievement motivation (scored on a scale from 1-100).
   a. Identify the variables in my study and tell the type of each (hint- there are four)
   b. **Identify the bivariate analyses** I might conduct with these variables (hint -- there are six) and tell the statistic I would use to perform the analysis. For each, state the null hypothesis.

   Based on a pilot study, I found that graduates (N=10) had a mean motivation of 85 while the undergraduates (N=10) had a mean of 80 (F(1,18) = 1.25, Mse = 100).
   c. What is my effect size for this analysis.
   d. Assuming the value in "a" is a reasonable estimate, what sample size would I need to reject H0: at p = .05 with power = .80?
   e. What is the 95% confidence interval around this mean difference? Based on this confidence interval, would you expect that the p-value for the t-test presented above would be p < .05, or  p > .05? Explain briefly.

2. Perform an outlier analysis on the following set of data. With hinges of 124.5 & 134.5.

```
109 121 123 126 126 127 129 134 135 136 152
```
   a. fourth spread
   b. lower outlier bound
   c. upper outlier bound
   d. any outliers? =

   e. Compared with the full data mean, will the screened mean be larger smaller about the same
      Explain briefly.

   f. Compared with the full data std, will the screened std be larger smaller about the same
      Explain briefly.
Set #1 Answers:

1a. GPA and motivation are quantitative; grad-undergrad and type of job are qualitative

b. gpa and motivation -- use r; H0: about linear relationship
ggrad-ugrad and jobt type -- use x²; H0: about pattern of relationship
all other pairs -- use F; H0: about mean differences on the DV

c.  \( r = .254 \)
d. using the table for \( r = .25 \), we find that for power = .8, we need S=120 or 60 in each of the 2 IV conds.
  the mean difference is 5 (85-80) so, the confidence interval = 5 +/- 9.39 (giving a range of -4.39 to 14.39)
since the CI includes 0, we would expect that p > .05, and we would retain the H0: that there is no mean
difference between the means of these populations.

2a. the fourth spread is 10
b. the upper limit is 149.5
c. the lower limit is 109.5
d. the outliers are 152 and 109
e. the mean will be about the same (slightly smaller)
f. the standard deviation is always smaller whenever an outlier(s) is removed
Set #2

1. I'd like to collect data from inpatients and outpatients at the local psychiatric hospital, including the number of previous times they have had therapy, their diagnosis (depressed or schizophrenic), and obtain a measure of their amenability to psychiatric intervention (that gives a score of 1-50).
   a. Identify the variables in my study and tell the type of each (hint- there are four)

   b. Identify one set of variables that might be examined using each of the three types of bivariate analyses (applying "conventional analyses"). Tell the H0: for each analysis.

   Pearson's Correlation ____________________ & ____________________
   H0:

   Pearson's X² ____________________ & ____________________
   H0:

   Between groups ANOVA/t-test ____________________ & ____________________
   H0:

2. The results from my pilot study comparing Schizophrenics and Depressives in terms of their performance on a social skills roll-playing task were Schizophrenics (M=41.1, S=9.59) and Depressives (M=48.48, S=7.18), with t(49) = 2.983, p = .004. I was testing my notion that depressives, because much of their cognitive capacity is taken up with ongoing rumination, will have less cognitive capacity to commit to a social interaction, and so will end up with a lower social skills score.

   a. Do these results support my notion? (Be sure to explain your answer)

   b. What is my effect size for this analysis?

   c. Planning for the next study -- assuming the values in "b" are reasonable estimates, what sample size would I need to reject H0: at p = .05 with power = .90 (be sure to tell which of the effect size estimates you are using) ?

   d. Based on the r-value computed in "b", what is the 95% confidence interval around this effect size estimate?

   e. Based on the lower-bound of the 95% CI, what sample size should we do to ensure power of .90 in our next study?

3. I've just completed the analysis of my latest data which compares the effects of practice on the performance of a series of critical thinking tasks as a function of whether or not students are given feedback following each task. As expected, there was a mean difference in the expected direction, however, the statistical result was t(27) = 2.06, p = .05. I know that finding exactly p = .05 can probably be avoided if I recalculate with more decimals, but I'd like some help… (with a brief explanation of how you arrived at each answer, please)

   a. If I reject H0: what is the probability that I'll be committing a Type II error?

   b. If I reject H0: what is the probability that I'll be committing a Type I error?

   c. If I reject H0: what is the probability that I'll be committing a Type III error?

   d. If I retain H0: what is the probability that I'll be committing a Type II error?

   e. If I retain H0: what is the probability that I'll be committing a Type I error?

   f. If I retain H0: what is the probability that I'll be committing a Type III error?

4. Perform an outlier analysis on the following set of data. Hinges = 34.5 & 43.5

   19  31  33  36  36  37  39  41  43  44  83  92

   a. fourth spread = ____________________

   b. upper outlier boundary = ____________________

   c. lower outlier boundary = ____________________

   d. any outliers? => ____________________

   e. Compared with the full data mean, will the screened mean be larger smaller about the same

   Explain briefly.

   f. Compared with the full data std, will the screened std be larger smaller about the same

   Explain briefly.
Set #2 Answers:

1. Variables & analyses
   a. variables and types -- # previous (quant), amenability (quant), patient type (qual), diagnosis (qual)
   b. each type of stat
      - don't say "in the population represented by the sample" - tell the population!!!
      - r - # previous and amenability -- There is no linear relationship between the number of previous times receiving therapy and amenability to intervention among psychiatric patients.
      - X² - diagnosis and patient type -- There is no pattern of relationship between diagnosis (schizophrenic or depressed) and patient type (in- vs. out-patient) among psychiatric patients.
      - F - any pair of 1 qual and 1 quant -- "general form": Among psychiatric patients, (one condition of a qual variable) and (other condition of the same qual variable) have the same mean (a quant variable).
      - Some folks said something that would be translated as "The mean IV is the same as the mean DV". What's a "mean IV" ??

2. Results, etc.
   a. Got an effect, but it was "backwards"
   b. r = .39
   c. Using r=.40 leads to S=58;
   d. r = .39 is .13 to .60
   e. with r = .10 (rounded down from .13) would need N = 1045

3. Errors and such
   a. 0% -- can't make a Type II error when rejecting H0:
   b. 5% -- probability of making a Type I error is the p-value used to reject the H0:
   c. Possible, but not estimable -- probability of making a Type III error when reject H0:
   d. about 50% -- need to convert t to r/d (.37/.79), look up power for the sample size of 29 (about .50), and solve for β as 1 - power
   e. 0% -- can't make a Type I error when retaining H0:
   f. 0% -- can't make a Type III error when retaining H0:

4. Outliers and results of their removal
   a. fourth spread = 98
   b. upper limit = 57
   c. lower limit = 21
   d. three outliers = 21, 83 & 92
   e. the mean will go down because removing above-the-mean outliers and only one below-the-mean outlier
   f. the standard deviation always goes down when extreme values are removed
Set #3

1. We own and operate a small company that helps students prepare for the Graduate Record Exam (GRE). We’ve been open for about 2 years, with some impressive results. Combining all the data from our first several hundred customers (most of whom were male Arts & Sciences seniors who were planning to go to graduate school) we’ve been able to boost their scores from around 520 before they take our course to about 680 after they take our course. Currently we use workbook type assignments with out customers, but we’ve been considering changing to on-line exercises. A recently published article and some of our own work suggests that we probably make the shift, but I’d like to ask some specifics…

   a. The first article I found reported better effects than we are used to seeing that were obtained by a company that used on-line exercises. Here’s the univariate stats they presented. Should I use their study to encourage my partner to consider us switching to on-line exercises? Why or why not?

   b. Do any of these variables have a skewed distribution? Explain your answer for all the points.

2. Recently we completed a pilot study that compared workbook with on-line exercises. When we compared GRE scores for those trained each way we found a mean GRE score of 627 for the workbook group (n = 12) and of 730 for the on-line group (n = 12). Statistical analysis revealed, \( F(1, 22) = 4.75, p = .040, \text{MSE} = 13467.193. \)

   a. I vaguely remember something about “statistical conclusion errors.” Could our conclusion be in error? What kind(s) of errors might we be making? How probable is the error (or errors, it there is more than one possible)?

   b. I know that pilot studies sometimes miss effects because they are too small. Should I worry about this problem for this pilot study? Why or why not?

   c. Please compute the size of this effect? Should I consider this a small, medium or large effect?

   d. It looks like we got a big mean difference. Please tell me the associated 95% confidence interval. What does my partner mean when he says that this will tell me the same thing as the p-value given above? (Now don’t go on a tirade about nhs vs. CI here – just tell how the two show the same information for this analysis!)

   e. What should I have done before running this ANOVA?

We’re planning to replicate this study but are arguing about the sample size we should use.

   e. My partner says we should just use the same sample size we used for the pilot study, because, as he said, “The sample size was big enough last time, so it’s a guarantee that it will always be big enough.” Is he right? Why or why not?

   f. What would be the power of the study if we ran it with 12 participants in each condition?

   g. Much power should we have for our replication study)? What sample size do we need for that much power?

   h. We got the following data from the on-line group.

   Perform an outlier analysis of these data (using the percentile data provided).

<table>
<thead>
<tr>
<th>610</th>
<th>620</th>
<th>680</th>
<th>700</th>
<th>700</th>
<th>720</th>
<th>740</th>
<th>740</th>
<th>750</th>
<th>750</th>
<th>880</th>
<th>890</th>
</tr>
</thead>
<tbody>
<tr>
<td>lower limit</td>
<td>upper limit</td>
<td>outliers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   i. Will the screened mean be larger or smaller than the mean computed from the full data?

   j. If we tossed the outlier, what would be the effect upon each mean? The std?

---

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>mean</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>28.9</td>
<td>3.2</td>
</tr>
<tr>
<td>Pre-training GRE score</td>
<td>664.3</td>
<td>52.8</td>
</tr>
<tr>
<td>Major</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arts &amp; Sciences</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Engineering</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>45%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR0000</td>
</tr>
<tr>
<td>N Valid</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>Missing</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>Percentile</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>690,000</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>730,000</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>750,000</td>
</tr>
</tbody>
</table>
3. There was another published study that involved a comparison of workbook and on-line training, but this time the outcome variable was whether or not the person’s score improved when they retook the GRE. The resulting data are shown below, $X^2(1) = 5.11, p = .031$.

<table>
<thead>
<tr>
<th>Workbook</th>
<th>On-line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved score</td>
<td>60</td>
</tr>
<tr>
<td>Score didn’t improve</td>
<td>30</td>
</tr>
</tbody>
</table>

a. Do these results support the notion that on-line training is superior to using workbooks? Explain your answer for all the points!

b. Do these results support the notion that more of those who do on-line assignments will improve and more of those who use the workbook will not improve? Explain your answer.

c. vaguely remember something about “statistical conclusion errors.” Could our conclusion be in error? What kind(s) of errors might we be making? How probable is the error (or errors, it there is more than one possible)?

d. I know that pilot studies sometimes miss effects because they are too small. Should I worry about this problem for this pilot study? Why or why not?

e. Please compute the size of this effect? Should I consider this a small, medium or large effect?

f. What should I have done before completing this analysis?

We're planning to replicate this study but are arguing about the sample size we should use.

g. My partner says we should just use the same sample size we used for the pilot study, because, as he said, “The sample size was big enough last time, so it’s a guarantee that it will always be big enough.” Is he right? Why or why not?

h. What would be the power of the study if we ran it with 50 participants in each condition?

i. Much power should we have for our replication study)? What sample size do we need for that much power?

4. Another question … I’ve been wondering about the number of assignments we should have our customers complete. When I looked at the relationship between the number of assignments and GRE improvement I found $r(28) = .31, p = .056$. My partner says that since I didn’t find a significant effect I’d be wasting my time to pursue this research further.

a. How would you respond to my partner’s comment?

b. My other partner points out that while this seems like a decent sized effect the .31 might be, in her words, “Overly hopeful” and suggests I compute a 99% confidence interval to “reconsider”. Please compute that for me.

c. What’s the chance that my statistical conclusion is in error? And what specific type of error(s) might this be?

d. If I do decide to pursue this research, what sample size should I use if I’m willing to take a 10% chance of making the type or error involved?

e. What should I have done before computing this correlation? Why?

5. I’m planning a study that compares the pre-training and post-training GRE scores of customers who get on-line training. I have two different research hypotheses. Tell the specific statistic to use to test each hypotheses and, also, please write the proper null hypothesis for each.

a. I think the on-line training will improve the GRE scores. Stat __________

b. I think that those who have higher GRE scores before on-line training will have the higher GRE scores after the on-line training. Stat __________
Set #3 Answers:

1a. Not a good idea, this sample is: probably older than out average customers, mostly engineering and educational majors and with initially higher GREs, and the N is only 10 !!
b. Nope: Age and pre-GRE are have +/-2std values that are possible. Skewness applies only to quantitative variables.

2a. Type I error → false alarm → p = 4.3% risk Type III → misspecification → not estimable
b. No, → you got an effect, so by definition there was “enough power”
c. r → .41, so between a medium (.30) and a large (.50) effect
d. ci95 = 1-3 +/- 90.83 or 4.17 to 201.83 CI does not include 0, so would reject null
e. Checked skewness and done outlier analyses separately for each group.
f. Not a good strategy! Sample size sufficient for “significance” isn’t always sufficient for “enough power” because the effect size estimate will vary from the population value. So need “room” to underestimate, but still be significant. In this case this works out very badly → use a power table!
g. n = 12 so N = 24 → gives us about 50% power.
h. Convention is to have .80 power (Type II error risk - .20 or 20%). With r = .40 and 80% power N should be 44, 22 in each condition.
i. Fourth spread = 750-690 = 60. lower bound = 690 – 90 = 600 upper bound = 750 + 90 = 840. 880 & 890 are outliers.
j. Because both outliers are “too large” outliers, the mean will decrease. The standard deviation will also (and always) decrease.

3a. Yes! There is a significant effect, and 50/60 passing is superior to 60/90.
b. Partial! As hypothesized, more of those who do On-Line will improve (50/60), however, contrary to the hypothesis, more of those who use the workbook did not improve (30/90)
c. Type I error → false alarm → p = 4.3% risk Type III → misspecification → not estimable
d. No, → you got an effect, so by definition there was “enough power”
e. r → .18    between small and medium
f. There is not any equivalent or outlier analysis for qualitative variables. Also, “skewness” doesn’t apply to qualitative variables.
g. Not a good strategy! Sample size sufficient for “significance” isn’t always sufficient for “enough power” because the effect size estimate will vary from the population value. So need “room” to underestimate, but still be significant. In this case, it works out pretty well, but this is not a strategy to adopt – use the power table!
h. n = 50, so N = 100, with r = .15 (rounded down from .18) power would be a little better than .30
i. Convention is to have .80 power (Type II error risk - .20 or 20%). With r = .15 (rounded down from .18) and 80% power N should be 343, 172 in each condition.

4a. This is bad advice, for several reasons… This is a medium sized effect, that was nearly significant, though it had a small sample (N=30) and poor power (<40%)
b. The CI99% is CI_{99%} = -.17 to .67
c. With power of < 40%, there’s > 60% chance of a Type II error (a “miss”)d. For an effect size of .30 and 90% power, you’ll want to use N = 109 participants.
e. Check for skewness, outliers and use a scatterplot to look for a nonlinear relationship between the variables

5. a. Improve suggests a mean difference – use WG ANOVA  
b. This language is about predictability of scores – use r
We own and operate a small company that helps students prepare for the State Bar Exam (BAR). We've been open for about 6 years, with some impressive results. Combining all the data from our first several hundred customers (with an average age of 28.4, 90% of whom were male graduates of UNL College of Law within the last year and of whom 80% planned to practice corporate law) we've been able to boost their scores from around 34 before they take our course to about 42 after they take our course (BAR has a range from 0-50). Currently we use lectures and mock tests with our customers, but we've been considering changing to a single pre-test followed by individual tutoring. A recently published article and some of our own work suggests that we should probably make the shift, but I'd like to ask some specifics…

a. The first article I found reported better effects than we are used to seeing that were obtained by a law program at a state university that used the sort of 1-on-1 tutoring with their senior students that we are considering. Here's the univariate stats they presented. Should I use their study to encourage my partner to consider us switching to 1-on-1 tutoring? Why or why not?

b. Do any of these variables have a skewed distribution? Explain your answer for all the points.

2. Recently we completed a pilot study that compared lectures with 1-on-1 tutoring. When we compared BAR scores for those trained each way we found a mean BAR score of 39.09 for the lecture group (n = 24) and of 40.67 for the 1-on-1 group (n = 24). Statistical analysis revealed, $F(1, 46) = 3.831, p = .056, \text{MSe} = 20.748$

a. I vaguely remember something about “statistical conclusion errors.” Could our conclusion be in error? What kind(s) of errors might we be making? How probable is the error (or errors, it there is more than one possible)?

b. I know that pilot studies sometimes miss effects because they are too small. Should I worry about this problem when interpreting the results of this pilot study? Why or why not?

c. I'd like to compare the results of this study to another I read. Would you please compute the size of this effect? Should I consider this a small, medium or large effect?

d. It looks like we got a big mean difference. Please tell me the associated 95% confidence interval. My partner tells me that I could use the CI instead of the F-test. How would I do that? Do the results always agree?

e. We’re planning to replicate this study but are arguing about the sample size we should use.

f. My partner says we should just use the same sample size we used for the pilot study, because, as he said, “The sample size was big enough last time, so it’s a guarantee that it will always be big enough.” Is he right? Why or why not?

g. Much power should we have (what is the “conventional” amount to shoot for)? What sample size do we need for that much power?
h. We got the following data from the 1-on-1 group. Perform an outlier analysis of these data (using the percentile data provided).

<table>
<thead>
<tr>
<th>DV</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>21.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>37.00</td>
<td>1</td>
<td>4.2</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>38.00</td>
<td>3</td>
<td>12.5</td>
<td>20.8</td>
</tr>
<tr>
<td></td>
<td>40.00</td>
<td>4</td>
<td>16.7</td>
<td>37.5</td>
</tr>
<tr>
<td></td>
<td>41.00</td>
<td>4</td>
<td>16.7</td>
<td>54.2</td>
</tr>
<tr>
<td></td>
<td>42.00</td>
<td>3</td>
<td>12.5</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>43.00</td>
<td>1</td>
<td>4.2</td>
<td>70.8</td>
</tr>
<tr>
<td></td>
<td>44.00</td>
<td>1</td>
<td>4.2</td>
<td>75.0</td>
</tr>
<tr>
<td></td>
<td>45.00</td>
<td>1</td>
<td>4.2</td>
<td>79.2</td>
</tr>
<tr>
<td></td>
<td>46.00</td>
<td>1</td>
<td>4.2</td>
<td>83.3</td>
</tr>
<tr>
<td></td>
<td>47.00</td>
<td>1</td>
<td>4.2</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>49.00</td>
<td>1</td>
<td>4.2</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>50.00</td>
<td>2</td>
<td>8.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

i. __________ lower limit __________ upper limit tell any outliers __________

j. Will the screened mean be larger or smaller than the mean computed from the full data?

Bonus: If we tossed the outlier, what would be the effect upon mean difference shown in “2” above? Why?

3. Another question … I’ve been wondering about the number of 1-on-1 sessions our customers should complete. When I looked at the relationship between the number of sessions (which ranged from 3 to 12) and BAR improvement I found \( r(29) = .32, p = .114 \). My partner says that since I found a medium sized effect I’d be wasting my time to replicate the study because we know there’s a correlation of about .3 in the population even though it wasn’t significant in this study.

a. How would you respond to my partner’s comment?

b. Please compute the 95% confidence interval for this correlation.

c. If I do decide to pursue a replication, what sample size should I use if I want to take only a 10% chance of missing an effect that is really there?

Bonus: What should I have done before computing this correlation? Why?

4. There was another published study that involved a comparison of doing 5 vs. 10 1-on-1 sessions, but this time the outcome variable was whether or not the person’s score improved when they retook the BAR. The resulting data are shown below, \( X^2(1) = 4.8, p = .032 \).

<table>
<thead>
<tr>
<th>Improved score</th>
<th>5 sessions</th>
<th>10 sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved score</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Score didn’t improve</td>
<td>20</td>
<td>4</td>
</tr>
</tbody>
</table>

b. Do these results support the notion that 1-on-1 sessions are superior to lectures? Explain your answer for all the points!

b. Does it make sense to compare these results to the study above (in 3) that showed the positive correlation between number of 1-on-1 lectures and BAR improvement? Why or why not?

c. How do you make that comparison? Are the results comparable?
Set #3 Answers:
We own and operate a small company that helps students prepare for the State Bar Exam (BAR). We’ve been open for about 6 years, with some impressive results. Combining all the data from our first several hundred customers (with an average age of 28.4, 90% of whom were male graduates of UNL College of Law within the last year and of whom 80% planned to practice corporate law) we’ve been able to boost their scores from around 34 before they take our course to about 42 after they take our course (BAR has a range from 0-50). Currently we use lectures and mock tests with our customers, but we’ve been considering changing to a single pre-test followed by individual tutoring. A recently published article and some of our own work suggests that we should probably make the shift, but I’d like to ask some specifics…

Table 1
Summary statistics for 410 customers

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>male 85% female 15%</td>
</tr>
<tr>
<td>Type of Law</td>
<td>criminal 25% corporate 75%</td>
</tr>
<tr>
<td>GPA</td>
<td>M = 3.65 S = 1.35</td>
</tr>
<tr>
<td>BAR scores (before tutoring)</td>
<td>M = 22.6 S = 7.65</td>
</tr>
<tr>
<td>Age</td>
<td>M = 28.8 S = 1.65</td>
</tr>
</tbody>
</table>

a. The first article I found reported better effects than we are used to seeing that were obtained by a law program at a state university that used the sort of 1-on-1 tutoring with their senior students that we are considering. Here’s the univariate stats they presented. Should I use their study to encourage my partner to consider us switching to 1-on-1 tutoring? Why or why not?

- Good-sized sample
- Age looks good 28.4 vs. 28.8
- Gender looks good 90% vs. 85%
- Plans looks good 80% vs. 75%
- However – take a look at the BAR pre scores – this sample has a much lower mean than our population (which might account for why they showed larger

b. Do any of these variables have a skewed distribution? Explain your answer for all the points.

This only relates to quantitative variables. If we look at GPA +/- 2 std we get 1.95 to 5.35, with a ceiling of 4.00. So the distribution of these data is probably negatively skewed.

Recently we completed a pilot study that compared lectures with 1-on-1 tutoring. When we compared BAR scores for those trained each way we found a mean BAR score of 39.09 for the lecture group (n = 24) and of 40.67 for the 1-on-1 group (n = 24). Statistical analysis revealed, F(1, 46) = 3.831, p = .056, MSE = 20.748.

i. I vaguely remember something about “statistical conclusion errors.” Could our conclusion be in error? What kind(s) of errors might we be making? How probable is the error (or errors, it there is more than one possible)?

With p = .056 we should retain H0:, and so we might be making a Type II error. The probability of this error is 1-power -- With S = df+2 = 48 and r = v(3.831/(3.831+46)) = .28, the analysis has a power of about .4 (r=.25) or about .5 (r=.30). So the Type II error estimation is between 50% & 60%, not a very convincing H0:.

Were we to reject H0: & call this a marginal effect we’d have a 5.6% chance of a Type I error and also risk a Type III.

j. I know that pilot studies sometimes miss effects because they are too small. Should I worry about this problem when interpreting the results of this pilot study? Why or why not?

Yes – this is a medium-sized effect with low power – retaining H0: is likely to be a mistake.

k. I’d like to compare the results of this study to another I read. Would you please compute the size of this effect? Should I consider this a small, medium or large effect?

This is the same r = .28 from above, which pretty close to a medium-sized effect (.30). Several said it was “small” and others “between small and medium” -- .28 is much closer to .3 than to .1 – call it “medium”.

l. It looks like we got a big mean difference. Please tell me the associated 95% confidence interval. My partner tells me that I could use the CI instead of the F-test. How would I do that? Do the results always agree?

95% CI with derror = 48, should use t-critical = 2.042 because 48 is a lot closer to 30 than to infinity

SEM D = √ [MSE error * (1/n1 + 1/n2)] = √ [20.748 * (1/24 + 1/24)] = 1.32

mean dif +/- t-critical & SEMD = 40.67 – 39.09 +/- 2.042 * 1.32 = 1.58 +/- 2.577 or -1.00 to 4.16

Since the 95% CI includes the H0: value of 0, we would retain the null and conclude there is not a significant difference between the means. The results using a 95% confidence interval and a significance test with p = .05 will always agree.
We’re planning to replicate this study but are arguing about the sample size we should use.

m. My partner says we should just use the same sample size we used for the pilot study, because, as he said, “The sample size was big enough last time, so it’s a guarantee that it will always be big enough.” Is he right? Why or why not?

This is a bad idea. The sample size used last time led to a nonsignificant results for nearlu-medium-sized effect - hard to see how that can be called “big enough last time”.

n. Much power should we have (what is the “conventional” amount to shoot for)? What sample size do we need for that much power?

A sample size of 120 would give about the “standard” of .80 (rounding down to r=.25 to be a bit safer).

o. We got the following data from the 1-on-1 group. Perform an outlier analysis of these data (using the percentile data provided).

<table>
<thead>
<tr>
<th>DV</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>37.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>8.3</td>
</tr>
<tr>
<td>38.00</td>
<td>3</td>
<td>12.5</td>
<td>12.5</td>
<td>20.8</td>
</tr>
<tr>
<td>40.00</td>
<td>4</td>
<td>16.7</td>
<td>16.7</td>
<td>37.5</td>
</tr>
<tr>
<td>41.00</td>
<td>4</td>
<td>16.7</td>
<td>16.7</td>
<td>54.2</td>
</tr>
<tr>
<td>42.00</td>
<td>3</td>
<td>12.5</td>
<td>12.5</td>
<td>66.7</td>
</tr>
<tr>
<td>43.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>70.8</td>
</tr>
<tr>
<td>44.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>75.0</td>
</tr>
<tr>
<td>45.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>79.2</td>
</tr>
<tr>
<td>46.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>83.3</td>
</tr>
<tr>
<td>47.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>87.5</td>
</tr>
<tr>
<td>49.00</td>
<td>1</td>
<td>4.2</td>
<td>4.2</td>
<td>91.7</td>
</tr>
<tr>
<td>50.00</td>
<td>2</td>
<td>8.3</td>
<td>8.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Please show your work.

4th depth = 44.75 – 40 = 4.75 -- use the percentiles you were given!!
1.5 * 4th depth = 7.125

<table>
<thead>
<tr>
<th>DV</th>
<th>Valid</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Valid</td>
<td>Missing</td>
</tr>
<tr>
<td>Percentiles</td>
<td>40.0000</td>
<td>0</td>
</tr>
<tr>
<td>50.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

i. ____________ lower limit ____________ upper limit

tell any outliers ____________

j. Will the screened mean be larger or smaller than the mean?

If we toss a “too small” outlier the screened mean will be larger than the mean from the full mean.

Bonus: If we tossed the outlier, what would be the effect upon mean difference shown in “2” above? Why?

If we toss the outlier the mean of the 1-on-1 group will increase, and the F will also increase. The standard deviation of this group will also decrease, and so the MSe will decrease and the F will increase. Given how close p is to .05, these changes might lead to a significant F for the pilot study.

Another question … I’ve been wondering about the number of 1-on-1 sessions our customers should complete. When I looked at the relationship between the number of sessions (which ranged from 3 to 12) and BAR improvement I found r(29) = .32, p = .114. My partner says that since I found a medium sized effect I’d be wasting my time to replicate the study because we know there’s a correlation of about .3 in the population even though it wasn’t significant in this study.

e. How would you respond to my partner’s comment?

While it is true that our “best guess” of the population value is .32, we also know that with this small sample size that guess could be off considerably (check the CI from the next problem…). Replication is seldom a waste, especially a replication of a lone finding that is based on a small sample size.

f. Please compute the 95% confidence interval for this correlation.

With r = .32, Zr = .332 StdZ = 1 / √(N-3) = .189 for a 95% CI we’d use 1.96 giving .332 +/- .370. This gives a Zr range from -.038 to .702, which translates to a correlation range of -.038 to .61. The large range of this CI shows just how dangerous it would be to accept the partner’s notion expressed in 3 above – we really don’t know there’s a correlation of about .3!!!
g. If I do decide to pursue a replication, what sample size should I use if want to take only a 10% chance of missing an effect that is really there?

With $r \sim .3$ and a desired power of .90, we'd want a sample size of 109.

Bonus: What should I have done before computing this correlation? Why?

Before computing the correlation it would have been a good idea to perform univariate outlier analyses for each of the variables and to compose the scatterplot, to be sure the relationship is not interestingly nonlinear. Some suggested "an a priori power analysis" but that must be done before collecting the data.

5. There was another published study that involved a comparison of doing 5 vs. 10 1-on-1 sessions, but this time the outcome variable was whether or not the person's score improved when they retook the BAR. The resulting data are shown below, $X^2(1) = 4.8$, $p = .032$.

<table>
<thead>
<tr>
<th></th>
<th>5 sessions</th>
<th>10 sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved score</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Score didn't improve</td>
<td>20</td>
<td>4</td>
</tr>
</tbody>
</table>

b. Does it make sense to compare these results to the study above (in 3) that showed the positive correlation between number of 1-on-1 lectures and BAR improvement? Why or why not?

Yes – The main question from both analyses is whether there’s a relationship between number of sessions and BAR improvement – most comparisons of this type aren’t from exact replications!

c. How do you make that comparison? Are the results comparable?

The direction/pattern of the effects of the two studies are similar – both found that more sessions was related to better improvement. We should also compute the effect size for this analysis = $\sqrt{\frac{X^2}{N}} = \sqrt{\frac{4.8}{74}} = .255$, which is not much different from the .32 we found above.

5. In another study, subjects had an average BAR1 (first time taking the BAR) score of 40 and a standard deviation of 5 and a BAR2 (second time taking the BAR) score of 48 and a standard deviation of 3. What would be the mean and standard deviation of "Y", resulting from each of the following SPSS compute statements?

<table>
<thead>
<tr>
<th>new mean</th>
<th>new standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. compute $y = (BAR1 + 5) \times 10$.</td>
<td>$(40 + 5) \times 10 = 450$</td>
</tr>
<tr>
<td>b. compute $y = (BAR1 \times 10) + 5$.</td>
<td>$(40 \times 10) + 5 = 405$</td>
</tr>
<tr>
<td>c. compute $y = (2 \times (BAR2 - BAR1)) + 10$.</td>
<td>This is a change score, change scores are non-linear transformations because they do not involve a linear transformation by a constant.</td>
</tr>
</tbody>
</table>