Statistics 101, Section 002: Practice Exam C for Final Exam 2004

Instructions: Write your answers on the exam in the spaces after the questions. For maximum credit, show all work.

You are permitted to use two sheets of notes, front and back. Any other form of aid is not permitted. If you need clarification on any part of the exam, contact Prof. Jang.

Provide the information requested below in the adjacent empty spaces.

NAME (print): ___________________________ LAB TIME: ___________________________

Honor Pledge: ``I have not given or received assistance on this exam while taking the exam.''

SIGNATURE:

<table>
<thead>
<tr>
<th>Page</th>
<th>Points Possible</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>160</td>
<td></td>
</tr>
</tbody>
</table>

QUESTIONS 1 – 7 REFER TO THE DATA SET DESCRIBED BELOW
Psychologists are interested in measuring perception of risk, since it is an important component in any decision-making process. Carlstrom et al. (2000) asked 611 participants to provide a numerical value of risk for several activities using a scale from 0-100 (0 being no risk, and 100 being high risk). The participants were also asked questions to identify their world view. Participants are classified either as hierarchicalists (“Everyone has his/her place in society, and societal status is hierarchical”), individualists (“I control my environment and destiny”), or egalitarians (“I have little respect for any decisions not made by the group”). For this study, participants who were not classified in one of these groups were called “unclassifiable.”

DESCRIPTION OF THE DATA
========================

The 611 participants in this study were recruited between 1997 and 1998 from five sources: UCLA psychology undergraduate classes, campus and community organizations, community and college newspaper advertisements, a paid consultant, and posted flyers. The following are five activities each of the participants were asked to rate on risk value:

- **DOC**: Work as a family physician in rural area
- **SWAT**: Work as a member of a SWAT police team
- **POOL**: Swim in indoor public pool each weekend
- **NUC**: Live near nuclear power station
- **PLANE**: Fly on commercial airplanes every month

Other variables in the data set include:

<table>
<thead>
<tr>
<th>Race</th>
<th>1 = Caucasian, 2 = African-American, 3 = Mexican-American, 4 = Taiwanese-American.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0 = Female, 1 = Male.</td>
</tr>
<tr>
<td>Age</td>
<td>Age of participant.</td>
</tr>
<tr>
<td>Worldview</td>
<td>0 = Unclassifiable, 1 = Individualist, 2 = Hierarchicalist, 3 = Egalitarian</td>
</tr>
</tbody>
</table>

There are no problems on this page. The next two pages display output from exploratory data analyses that you should use to answer exam questions. The questions begin on page 5.
Women: 385
Men: 226

Caucasian: 158
African-American: 147
Mexican-American: 140
Taiwanese-American: 166

Unclassifiable: 384

Individualist: 51
Hierarchicalist: 98
Egalitarian: 78

---

Distribution of NUC
Mean = 77.86,
Median = 90.00,
SD = 26.30

Distribution of SWAT
Mean = 72.25
Median = 80.00
SD = 22.60

Distribution of NUC minus SWAT
Mean = 5.61
Median = ?
SD = 29.60

Correlations among selected variables

<table>
<thead>
<tr>
<th></th>
<th>DOC</th>
<th>NUC</th>
<th>PLANE</th>
<th>POOL</th>
<th>SWAT</th>
<th>AGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOC</td>
<td>1.0000</td>
<td>0.0372</td>
<td>0.1474</td>
<td>0.2844</td>
<td>0.0535</td>
<td>-0.0792</td>
</tr>
<tr>
<td>NUC</td>
<td>1.0000</td>
<td>0.2540</td>
<td>0.2040</td>
<td>0.2682</td>
<td>-0.1390</td>
<td></td>
</tr>
<tr>
<td>PLANE</td>
<td>1.0000</td>
<td>0.4121</td>
<td>0.1707</td>
<td>-0.0781</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POOL</td>
<td>1.0000</td>
<td>0.0878</td>
<td>-0.0332</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWAT</td>
<td>1.0000</td>
<td>0.0141</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

42 rows not used due to missing values.
Data on world views

246 women are unclassifiable, 37 women are individualists, 64 women are hierarchicalists, and 38 women are egalitarian.

Contingency table of world view by race

<table>
<thead>
<tr>
<th></th>
<th>Unclassifiable</th>
<th>Individualist</th>
<th>Hierarchicalist</th>
<th>Egalitarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasians</td>
<td>79</td>
<td>10</td>
<td>24</td>
<td>45</td>
</tr>
<tr>
<td>African Americans</td>
<td>111</td>
<td>10</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Mexican Americans</td>
<td>91</td>
<td>9</td>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>Taiwanese-Americans</td>
<td>103</td>
<td>22</td>
<td>23</td>
<td>18</td>
</tr>
</tbody>
</table>

There are no exam problems on this page. Exam problems begin on the next page.
EXAM PROBLEMS BEGIN HERE

1. (2 points per part) For problems 1a and 1b, write numbers for each answer. Ranges (e.g., “between 64.3 and 68.5”) will receive no credit. For parts 1c-1e, circle the correct answer.

a) Estimate the average age of the study participants. ____

b) Estimate the percentage of people in the study under age 25. ____

c) Circle the number that is closest to the SD of age: 1, 5, 15, 25, 35, 45.

d) The study has no one under age 17. True False Cannot tell without more information

e) About 68% of the participants’ ages are within 1 SD of the average age.
   True False Cannot tell without more information

2. (2 points per part). For 2a – 2d, circle the appropriate answer. For 2e, write a number for your answer. Ranges will receive no credit.

a) Which one of the following three scatter plots portrays the relationship between NUC and SWAT most accurately. Circle the letter of the correct plot.

The scatter plots were drawn by hand and so are hard to put up. But, one had a very strong correlation much bigger than 0.25. Another had absolutely zero correlation. The third had the right amount of correlation.

b) Which variable has the strongest linear association with AGE in these data:
   DOC, NUC, PLANE, POOL, SWAT, Cannot tell without more information

c) When used as the predictor in a simple regression, which variable explains the most variation in PLANE scores in these data:
   DOC, NUC, POOL, SWAT, AGE, Cannot tell without more information

d) When using a simple regression with NUC as the outcome and AGE as the predictor, the typical deviation around the regression line is at least 26.3.
   True, False, Cannot tell without more information.

e) Predict the SWAT score for a person who rates NUC as a 90. Show your work for full credit.
3. (3 points) Like many psychology studies, this is not a random sample from a well-defined target population. List one way that the researchers’ methods of selecting study participants could result in untrustworthy conclusions about the relationships between risk perception and the other variables in the study (e.g., with age, gender, or world view). Assume the wording of questions is not a problem, i.e. focus on the data collection, not the question wording.

For the remaining problems on this exam concerned with this data set, consider the study participants a random sample from the target population of all people affiliated with the UCLA psychology department.

4. **Comparisons of the perceived risks of living near a nuclear power plant (NUC) to the those of being on a police SWAT team (SWAT).**

   a) (2 points) Estimate the median for the difference variable, “NUC minus SWAT.”

   b) (3 points) Give an interval for the average ranking of SWAT in the target population of all people affiliated with the UCLA psychology department. Use a 95% confidence level.

   c) (14 points) Test whether people in this target population on average rate SWAT and NUC differently. Write the null and alternative hypotheses, the value of the test-statistic, the p-value, and the conclusions. Write conclusions in at most two sentences using language that someone who doesn't know statistics would understand. Consider p-values less than 0.10 as small.

   d) (5 points) In addition to random sampling, what condition must hold for the test in part c to be valid? Do you think this condition holds in these data? Make sure to address both questions in your answer.
5. Comparisons of risk perceptions across gender for DOC, POOL, and PLANE.

a) (3 points) Give an interval for the difference in the population average of DOC for women minus the population average of DOC for men. Use a 99% confidence level.

b) (3 points) Is there enough evidence to say with 99% confidence that, in this target population, on average women consider being a doctor in a rural area to be a riskier activity than men consider it to be? Justify your answer based on part a.

c) (2 points) The summary statistics for POOL by gender and for PLANE by gender are not reported on page 4. They are reported here in two tables. Write the variable name (POOL or PLANE) that corresponds to each set of statistics to the right of each table.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>38.46</td>
<td>26.15</td>
</tr>
<tr>
<td>Men</td>
<td>31.45</td>
<td>27.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>25.73</td>
<td>22.60</td>
</tr>
<tr>
<td>Men</td>
<td>21.91</td>
<td>22.60</td>
</tr>
</tbody>
</table>

d) (7 points) Which null hypothesis has stronger evidence against it in these data: (i) the population average of PLANE for women is equal to the population average of PLANE for men; or, (ii) the population average of POOL for women is equal to the population average of POOL for men? Defend your answer with statistical arguments.
6. World views of different groups

a) (3 points) Give an interval for the percentage of women in this target population who are egalitarian. Use a 95% confidence level.

b) (15 points) Test whether the population percentage of men who are egalitarians differs from the population percentage of women who are egalitarians. Show your null and alternative hypotheses, the value of the test statistic, the p-value, and conclusions. Write conclusions in at most two sentences using language that someone who doesn’t know statistics would understand. Consider p-values less than 0.10 as small.

IF YOU CANNOT DETERMINE THE SAMPLE PERCENTAGES OF EGALITARIANS, use 36% for women and 25% for men. These are made-up (not correct) percentages, and you should use the correct ones for full credit. Using the made-up percentages can earn a max of 10 points.

c) (5 points) A chi-squared test is performed to test whether or not race and world view are independent. The chi-squared test statistic equals 69.0, and the p-value is less than 0.0001. Explain (i) how to interpret the p-value, and (ii) your conclusion about the relationship between race and world view. Make sure to explain both (i) and (ii) in your answer.

d) (5 points) How much does the entry for Mexican-American hierarchicalists contribute to the value of the chi-squared test statistic? That is, calculate the piece of the chi-squared test statistic derived from this cell of the relevant contingency table.
7. Some conceptual questions on the results.

a) (4 points) For each of the following data points, what happens to the correlation between NUC and SWAT when it alone is added to the data? Circle the appropriate answer for each point.

<table>
<thead>
<tr>
<th>SWAT</th>
<th>NUC</th>
<th>Effect on correlation when data point is added (circle one answer for each)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>decreases slightly, stays exactly the same, increases slightly</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>decreases slightly, stays exactly the same, increases slightly</td>
</tr>
<tr>
<td>0</td>
<td>100</td>
<td>decreases slightly, stays exactly the same, increases slightly</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>decreases slightly, stays exactly the same, increases slightly</td>
</tr>
</tbody>
</table>

b) (6 points) Suppose you sample thirty more men and thirty more women. Amazingly, all thirty men rank DOC as a 21, and all thirty women rank DOC as a 23. Using all 611+60 = 671 people, you obtain correctly the p-value for a test for the difference in average DOC score between men and women. Here’s the question: would this p-value be larger or smaller than the p-value for a similar hypothesis test based on only the original 611 people? Justify your answer, using numerical arguments in your defense.

d) (5 points) Suppose you make a contingency table of just Caucasians and Taiwanese Americans who are unclassifiable or individualists. The table is thus:

<table>
<thead>
<tr>
<th></th>
<th>Unclassifiable</th>
<th>Individualist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasians</td>
<td>79</td>
<td>10</td>
</tr>
<tr>
<td>Taiwanese Americans</td>
<td>103</td>
<td>22</td>
</tr>
</tbody>
</table>

Add 10 people however you want so that there is clearly no association between race and worldview. For your answer, fill in the contingency table below, showing your new counts.
8. If you used these dice in Vegas...well, let's just say I wouldn't recommend it.

"Ace-six flats" are a type of crooked dice where the cube is shortened in the one-six direction, the effect being that the 1s and the 6s are more likely than 2s, 3s, 4s, and 5s. Suppose that

\[ \Pr(\text{roll a 1}) = \Pr(\text{roll a 6}) = 1/4, \text{ and } \Pr(\text{roll a 2}) = \Pr(\text{roll a 3}) = \Pr(\text{roll a 4}) = \Pr(\text{roll a 5}) = 1/8. \]

For the ace-six flats dice described, the chance that the sum of two dice is 7 equals 0.1875. For regular, fair six-sided dice, the chance that the sum of two dice is 7 equals 0.1667.

a) (5 points) You can choose to roll two ace-six flats dice 1000 times, or to roll two regular dice 100 times. If you roll more than 20% sevens, you win one million dollars. Which choice gives you the better chance of winning the million dollars? Justify your answer.

b) (4 points) In the casino game craps, you roll two dice. You win if the sum of the two dice is a seven or an eleven. You roll a pair of dice one time. Calculate the chances that you win with (i) the ace-six flats dice, and (ii) fair dice. Show the chances and your work for both types of dice.

c) (5 points) Pretend that you are the owner of the casino. You see a gambler who you suspect is using ace-six flats dice rather than regular ones. She has played 100 times and obtained 30 wins by throwing a seven or eleven on the first roll of the dice. For the ace-six-flats dice, calculate the chance she would get at least 30 wins. Show work.

d) (1 point) Do you think the person in part c is using the ace-six flats dice or the fair dice? Very briefly say why.
9. Come on… be a Bayesian. Everyone is doing it.

A Stat 101 savvy student seeks to learn about the percentage of current Duke undergraduate students who have jobs this summer. He surveys a random sample of 50 Duke undergraduate students. Of the 50 students, 35 say that they have a summer job, and 15 say they do not. Because the sample size is reasonably large, the student uses a normal curve to approximate the likelihood function, with mean and SE based on those from the data.

Based on information from the Duke career counselors, the student has a prior belief that the percentage of Duke students with summer jobs will be around 50%, give or take 10%. To represent his prior beliefs, he uses a normal curve with a mean of 0.50 and an SD of 0.10. He then proceeds to use Bayesian statistics to estimate the percentage of current Duke students who have summer jobs.

a) (8 points) The graph below shows four curves: the likelihood function, the prior distribution, the posterior distribution, and a completely unrelated curve. Label the likelihood function (the curve for the data) with the phrase “Like”, the prior with the phrase “Prior”, the posterior with the phrase “Post”, and the unrelated curve with the phrase “Unrel.” Write labels at the top of the curve.

The plots were drawn by hand. There was a curve centered at 50% (the prior), a curve centered at 70% (the likelihood), a curve around 65% with about the right SD (the posterior), and a curve centered around 65% with a huge SD (the unrelated curve).

b) (3 points) Can you determine the maximum likelihood estimate of the population percentage of Duke students who have a summer job? For your answer, either write your estimate, or say you need more information.

c) (4 points) Suppose the student wants to reduce the influence of his prior beliefs on the estimates. Which of the following steps would accomplish this? Circle ALL the steps that would reduce the influence of his prior beliefs.

Increase the sample size.
Decrease the sample size.
Increase the SD of the normal curve for the prior beliefs.
Decrease the SD of the normal curve for the prior beliefs.

10. True or False (6 points per part).

For each statement, if you think the statement is always true, just say it is true. If you think the statement is always false or sometimes false, say it is false and explain why or when it is false in two or less sentences.

a) When the p-value is 0.015, you should reject the null hypothesis because there is only a 1.5% chance that the null hypothesis is true.

b) The senior survey at Duke is sent to all 1500 seniors, who are asked to respond to various questions about their Duke experience. Out of the 500 seniors who return it, 400 cite parking as a "serious problem that negatively affected my Duke experience." True or false: A 95% confidence interval for the percentage of Duke seniors rating parking as a serious problem is: 0.80 ± 0.035.

c) In the study on risk perception presented on this exam, the correlation between DOC and RACE is a valid summary of the association between DOC and RACE. (The correlation is not shown, but you still can answer the question as true or false.)

d) In the U.S., a 95% confidence interval for the difference in population average years of education for men and population average years of education for women stretches from 0.25 to 0.29. True or false: The p-value for a hypothesis test of whether there is any difference in the population averages for men and women is smaller than 0.05.