1. (1 mark) Briefly describe a p-value.

Solution: They need to mention the null, and the fact that the p-value is not the probability of the observed statistic, but of the probability of seeing something as or more extreme than the statistic. Example answers that would get full marks:

• The chance of seeing a value of a statistic that is as extreme or more extreme than what we observed, under the null hypothesis.
• The probability under the null hypothesis of data that is as or more extreme than what we observed.

2. (1 mark) What does a 95% confidence interval cover 95% of the time?

Solution: The true value of the parameter.

3. (2 marks) Briefly describe the difference between a statistic and a population parameter.

Solution: A statistic is calculated from data (1 mark; alternatives include: is a feature of a sample) A population parameter is an unknown population-level quantity (1 mark; alternatives include: is a feature of the population).

4. (6 marks) North Carolina State University looked at the factors that affect the success of students in a required chemical engineering course. Students must get a C or better in the course to continue as chemical engineering majors. There were 65 students from urban or suburban backgrounds and 52 of these students succeeded. Another 55 students were from rural or small-town backgrounds; 30 of these students succeeded in the course.

(a) (4 marks) Is there good evidence that the proportion of students who succeed is different for urban/suburban versus rural/small-town backgrounds? State hypotheses, give the p-value of a test and state your conclusion.

Solution: $H_0 : p_1 - p_2 = 0$ or $p_1 = p_2$ (either one OK) $H_a : p_1 - p_2 \neq 0$ or $p_1 \neq p_2$ (either one OK) (1 mark for both hypotheses) where $p_1$ is the proportion of urban/suburban students who succeed and $p_2$ is the proportion of rural/small-town students who succeed (or vice versa). Defining the null and alternative in words is
OK too. No mark if they use notation for population proportions without definition.

The pooled estimate of the success probability is \( \hat{p} = (52 + 30)/(65 + 55) = 0.683 \). The test statistic is

\[
\frac{52/65 - 30/55}{\sqrt{\hat{p}(1 - \hat{p})(1/65 + 1/55)}} = 2.99
\]

(1 mark) giving a p-value of \( 2 \times 0.0014 = 0.0028 \) (1 mark). We conclude that there is good evidence that the proportion of students who succeed is different for urban/suburban versus rural/small-town backgrounds (1 mark).

(b) (2 marks) Give a 95\% confidence interval for the size of the difference (use the “plus-4” method).

\[
\tilde{p}_1 = 53/67 \quad \text{and} \quad \tilde{p}_2 = 31/57,
\]

giving an estimate of \( 53/67 - 31/57 = 0.2471851 \), or about 0.247 (1 mark; 0.250 OK too)

The standard error I get is 0.0825789 and the critical value is 1.96, giving a margin of error of 0.1618546, or about 0.162. The CI is then \( 0.247 \pm 0.162 \) or \( (0.085,0.409) \) (1 mark; either the \( \pm \) or interval form of the CI is OK).

5. (5 marks) A study is being conducted to compare the yield of a genetically modified variety of corn to the yield of a standard variety. The experimenters will use a field that consists of 200 “plots”, each of which will contain one of the 200 plants (100 of each variety) available.

(a) (1 mark) Why should the experimenters use randomization to assign plants to plots?

\textit{Solution:} Any answer that shows they know randomization is one of the key principles of experimental design. E.g. Randomization tends to even out the effect of lurking variables.

(b) (2 marks) Briefly describe a strategy to randomly assign plants to the plots.

\textit{Solution:} Any reasonable strategy will do. Something like: Make a list of the plots (1 mark) and randomly choose 100 to receive the GM variety (1 mark).

(c) (2 marks) It is later determined that one half of the field has better soil than the other half. Using terminology from class, how would you describe
this “soil condition” variable? If the experiment was repeated, what should the experimenters consider doing with this variable?

*Solution:* Soil condition is a lurking (or confounding) variable (1 mark). The experimenters should consider blocking on this variable (1 mark).

6. (12 marks) The following table shows admission status by gender for students who applied to Graduate School at the University of California at Berkeley for the six largest departments in 1973.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admitted</td>
<td>1198</td>
<td>557</td>
<td></td>
</tr>
<tr>
<td>Rejected</td>
<td>1493</td>
<td>1278</td>
<td></td>
</tr>
</tbody>
</table>

(a) (2 marks) What are the row and column variables?

*Solution:* The row variable (defines the rows of the table) is Admit (1 mark) and the column variable is Gender (1 mark). This really should have been worth one mark.

(b) (2 marks) What proportion of female applicants were rejected? What proportion of male applicants were rejected?

*Solution:* Females: $1278/(557 + 1278) = 0.6964578$ or about 0.696. Males: $1493/(1198 + 1493) = 0.5548123$ or about 0.555.

(c) (1 mark) What is the estimated relative risk of rejection for females relative to males?

*Solution:* $0.696/0.555 = 1.255$ (or whatever they get if they use more decimals).

(d) (3 marks) Give an estimate and 95% confidence interval for the odds-ratio.

*Solution:* The estimated OR is

$$ \hat{OR} = \frac{1198 \times 1278}{1493 \times 557} = 1.841008$$

or about 1.841 (1 mark). The confidence interval can be calculated in one of two ways: (i) calculate the 95% CI for the log-OR and exponentiate the ends, or (ii) use the simplified formula I gave them. The 95% CI for the log-OR is

$$0.610 \pm 0.125 \quad \text{or (0.485, 0.736)}$$
and the 95% CI for the OR is about (1.62, 2.09). Two marks for the right answer. Use discretion in giving part marks. For example, a 95% CI for the log-OR would get 1 of 2 marks.

(e) (4 marks) Perform a chi-square test for association between gender and admission status at the 5% level. What do you conclude about gender equity in admissions at UC Berkeley in 1973? For your convenience, the data are reproduced below.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Admit</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admitted</td>
<td>1198</td>
<td>557</td>
</tr>
<tr>
<td></td>
<td>Rejected</td>
<td>1493</td>
<td>1278</td>
</tr>
</tbody>
</table>

**Solution:** One mark for correct expected values under the null:

\[
\begin{align*}
1043.461 & \quad 711.5389 \\
1647.539 & \quad 1123.4611 \\
\end{align*}
\]

One mark for a correct statistic value of about 92.2. One mark for correctly stating that we reject the null hypothesis at the 5% level (the p-value is less than 0.0005, the smallest value available in Table E). One mark for concluding that there was a gender bias in admissions at UC Berkeley in 1973.

7. (15 marks) Switzerland, in 1888, was entering a period known as the “demographic transition”; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries. The data collected are for 47 seven French-speaking “provinces” at about 1888. The following scatterplot examines the relationship between education (% education beyond primary school) and fertility (the Ig score, a common standardized fertility measure):
(a) (1 mark) On the plot, which variable is being considered as the response and which is the explanatory variable?

_Solution:_ Fertility is the response, education is the explanatory variable.

(b) (2 marks) The province that includes Geneva had the highest education. Circle the data point corresponding to this province on the plot and give an approximate value for its Ig score (to the nearest multiple of 5).

_Solution:_ They should circle the bottom-right-most point (1 mark). From the scatterplot, the Ig score looks to be about 35 (1 mark; it is in fact exactly 35).

(c) (1 mark) How would you describe the relationship between Education and Fertility?

_Solution:_ There is a negative relationship.

(d) (4 marks) Some summary statistics of the data are:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fertility</td>
<td>70.14255</td>
<td>12.49170</td>
</tr>
<tr>
<td>Education</td>
<td>10.97872</td>
<td>9.615407</td>
</tr>
</tbody>
</table>

_correlation between Fertility and Education = -0.6637889_
Based on these summaries, what is the slope and intercept of the least squares regression line?

_Solution:_ Slope is \(-0.6637889(12.49170/9.615407) = -0.8623506\) or about \(-0.862\) (2 marks; I don’t know why I allocated two marks). The intercept is \(70.14255 + 0.8623506 \times 10.97872 = 79.61006\) or about 79.6 (2 marks; again, I don’t know why).

(e) (2 marks) The province of Lausanne had Education of 28% and Fertility of 55.7. What is the predicted value and residual for this observation?

_Solution:_ Predicted value is 55.46424, or about 55.46 (1 mark). The residual is \(55.7 - 55.46424 = 0.23576\) or about 0.24.

(f) (1 mark) Today over 95% complete primary school. A friend who has not taken Stat 101 asks you to use your regression equation to predict fertility in Switzerland using 95% as an estimate of the education level. Using terminology from class, what would this be an example of?

_Solution:_ Extrapolation.

(g) (2 marks) Write a short sentence, using non-technical language, to explain why such a prediction would not be meaningful.

_Solution:_ They should mention two sources of extrapolation: Here are the ones I can think of:

- The relationship between education and fertility from 1888 will not necessarily hold today.
- The value of 95% is outside the range of observations in the data.

(h) (2 marks) Below is a residual diagnostic plot for the fitted model:
What assumption is being checked by this plot? Does the assumption appear reasonable?

Solution: This is checking normal errors (also accept normality of the residuals or just normality; 1 mark). The assumption appears reasonable (1 mark).