STATISTICS 151 SECTION 2 FINAL EXAM
DECEMBER 14 2006

Your name (print!):

Identification number:

Honor pledge: On my honor, I have neither given nor received unauthorized aid in this exam.

Sign here:

This is an open book exam. Course text, personal notes and calculator are permitted. You have 3 hours to complete the exam. Personal computers are not allowed. If you have any queries about the meaning of the questions, ask the instructor for assistance.

All answers are to be written on this sheet and you are to hand the sheet in on completion. SHOW ALL WORKING — even correct answers will not get full credit if it's not clear how they were obtained. Incorrect answers will gain substantial credit if the method of working is substantially correct. If the space provided for working is insufficient, you may also use the back of the sheet.

Answer SIX of the eight questions. Full credit may be gained by answering six questions, and it is not recommended that you attempt more than six unless you are sure you have time to do so. However, all questions you do attempt will be graded, and the best six scores will be used to compute your final total.

The total score available on each question is 20. Points for each part are given in parentheses.

In cases where part of a question depends on the answer to an earlier part of the same question, an incorrect answer to the first part will not prevent you gaining full credit for the second part, if the method used for the second part is correct and complete.

1. The year is 2011 and for the fifth year in a row, UNC football is ranked in the top 25. A sports journalist reckons that the probability that UNC will qualify for a bowl game of some description is 0.8, and the probability that UNC will qualify for a BCS bowl game is 0.2. In a BCS bowl game, the probability that they win is 0.3; in a non-BCS bowl game, the probability that they will win is 0.7. If they don't qualify for a bowl game at all, then by default the probability that they win is 0.

   (a) Represent all this information by means of a tree diagram, that represents (i) whether UNC get to a BCS, non-BCS or no bowl game, (ii) if they get to a bowl game, whether they win. **(7)**

(b) What is the probability that UNC win a bowl game at the end of the season? **(5)**

(c) Given that they win, what is the probability that it is a BCS game? **(8)**

2. Ozone is a chemical pollutant which is regulated by the Environmental Protection Agency. Table 1 (*back of exam — feel free to tear off and discard at end of exam*) represents the mean July ozone level, in parts per billion (ppb) for the years 1987–2000 in 95 U.S. cities.

   (a) Draw a stem and leaf plot of the data, choosing suitable units for the stem and the leaves. (Use the back of the page if you prefer to take more space.) **(4)**

   (b) Find the five-number summary and the inter-quartile range. **(4)**

   (c) Draw a histogram, making your own choice of suitable intervals. **(4)**

(d) Are there any outliers by the 1.5IQR rule? **(3)**

(e) You are given that $\bar{x} = 51.8$, $s = 11.5$. Are there any outliers by the alternative rule that uses $\bar{x}$ and s? **(2)**

(f) Summarize the overall shape of the distribution, including such issues as whether it is skewed or whether you believe there are outliers. **(3)**

3. An airline estimates that 80% of the passengers who reserve tickets actually show up for their flights. Based on this information, it has to decide how many tickets it will sell for each flight, which is typically more than the number of seats actually available. To answer the following questions, make sure you identify which probability distribution you are using and explain your justification.

   (a) In the first-class section of a particular aircraft, 12 seats are available. The airline sells 14 seats. What is the probability that more passengers will show up than there are seats for? **(7)**

   (b) In the economy section of a particular aircraft, 200 seats are available. The airline sells 225 seats. What is the probability that more passengers will show up than there are seats for? **(7)**

(c) Continuing part (b), the airline would like to sell 250 seats. To what number must it increase the number of seats on the aircraft, so that the probability of passengers having to be turned away is less that 0.05? **(6)**

4. A regular *New York Times* poll includes the question, "Do you approve or disapprove of the way George W. Bush is handling foreign policy?" In a recent poll (September 15 2006), there were 1,131 respondents, of whom 37% answered "approve", and 56% answered "disapprove" (the other 7% either didn't answer or said they had no opinion).

   (a) Construct a 90% confidence interval of the true proportion of the population who approve of Mr. Bush's foreign policy. **(8)**

   (b) Does this poll contain convincing evidence that more people disapprove than approve? Explain. **(5)**

   (c) In future, the New York Times decides its survey must have a margin or error no greater than 1.2%, with confidence level 95%. What sample size do they need to achieve this? **(7)**

5. A recent survey in *Consumer Reports* (December 2006) listed the prices of 11 models of bicycles, as follows:

Model	Price ($)
Klein Reve v	1800
Specialized Roubaix	1300
Trek Pilot 2.1	1320
Canondale Synapse 4	1050
Raleigh Cadent 1.0	650
Schwinn Super Sport GS	700
Schwinn Sierra GS	340
Mongoose Switchback SX	280
Brompton C-type C3E	675
Bike Friday Stock Pocket Tourist	830
Dahon Speed 7	380

($\bar{x} = 848;\ s = 476$)

(a) Compute a 99% confidence interval for the average selling price of all bicycles. **(14)**

(b) What are the assumptions behind the procedure in (a)? Highlight *two* features of the problem which might suggest the assumptions are inappropriate in this case. **(6)**

6. Many people believe that when the defendant in a jury trial is black, the prosecuting attorneys use peremptory challenges systematically to exclude blacks from the jury. To study this, a defendants' rights group followed the composition of 20 juries (240 jurors) in a particular county. Over the period of the study, 18% of the jury pool were black. However, of the 240 jurors actually chosen to serve on a jury (excluding alternates), only 14 were black.

(a) Formulate this as a hypothesis testing problem. Say what is the parameter of interest, and define an H_0 and H_a that would correspond to saying that the jury selection process either is, or is not, biased against blacks being chosen. **(5)**

(b) Compute the standard error, the test statistic, and the P-value. **(12)**

(c) Summarize your conclusions in everyday language. **(3)**

7. A chemical plant is discharging a certain toxic chemical into a river. Officials from the EPA determine that the maximum safe discharge of this chemical is 1.5 ounces per week. In an attempt to monitor the plant's compliance, they measure the actual amount of discharge each week. In five weeks, the actual discharges are 2.8, 1.3, 4.1, 1.8 and 2.5 ounces.

   (a) Compute the mean and the standard deviation of these discharges. What is the standard error of the mean? **(5)**

   (b) What would be an appropriate hypothesis test to determine whether the plant is in compliance with the standard? State H_0 and H_a. **(3)**

   (c) Compute the relevant test statistic and determine the P-value (as accurately as you are able to determine what it is). Would you accept or reject H_0? **(9)**

(d) Explain in layman's terms what this shows. **(3)**

8. The following table (taken from the General Social Survey) represents the answers by a randomly selected group of respondents to the question "How willing would you be to accept cuts in your standard of living in order to protect the environment?" The same question was asked in 1994 and again in 2000, with the following results:

Response	Very willing	Fairly willing	Neither willing nor unwilling	Not very willing	Not at all willing
1994	57	363	306	390	214
2000	64	280	312	240	274

(a) Complete the table by computing the row and column totals. What was the total number of respondents? **(3)**

(b) Compute the χ^2 statistic, the number of degrees of freedom, and the P-value. Would you accept or reject the hypothesis of independence? **(13)**

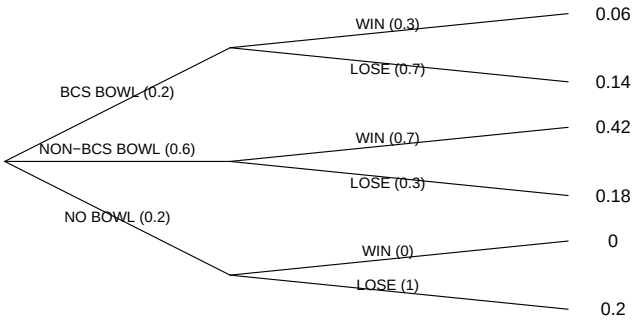
(c) Explain in layman's language what this means in terms of people's attitude to the environment. Is there evidence that this changed between 1994 and 2000? In what respect has the change been greatest? **(4)**

City	Ozone	City	Ozone	City	Ozone
Akron	63	Grand Rapids	54	New York	55
Albuquerque	54	Greensboro	61	Oakland	31
Arlington	67	Honolulu	15	Oklahoma City	53
Atlanta	63	Houston	41	Omaha	40
Austin	41	Huntsville	58	Orlando	42
Bakersfield	79	Indianapolis	59	Philadelphia	59
Baltimore	62	Jackson	45	Phoenix	55
Baton Rouge	45	Jacksonville	45	Pittsburgh	59
Biddeford	49	Jersey City	64	Portland	39
Birmingham	49	Johnstown	58	Providence	50
Boston	46	Kansas City, MO	53	Raleigh	62
Buffalo	54	Kansas City, KS	47	Riverside	79
Cedar Rapids	40	Kingston	52	Rochester	52
Charlotte	67	Knoxville	63	Sacramento	62
Chicago	47	Los Angeles	64	Salt Lake City	61
Cincinnati	57	Lafayette	41	San Antonio	38
Cleveland	54	Las Vegas	56	San Bernardino	88
Columbus, GA	49	Lexington	57	San Diego	48
Columbus, OH	58	Lincoln	39	San Jose	38
Colorado Springs	49	Lake Charles	40	Seattle	38
Corpus Christi	25	Louisville	57	Shreveport	51
Coventry	57	Little Rock	51	Spokane	45
Dayton	59	Madison	48	Santa Ana/Anaheim	49
Washington	61	Memphis	59	St. Louis	49
Denver	55	Miami	27	Stockton	56
Des Moines	31	Milwaukee	51	St. Petersburg	38
Detroit	51	Mobile	48	Syracuse	51
Dallas/Fort Worth	51	Modesto	63	Tacoma	38
El Paso	50	Muskegon	63	Tampa	40
Evansville	62	Nashville	49	Toledo	55
Fresno	76	New Orleans	38	Tucson	50
Fort Wayne	58	Newark	52		

Table 1: July average ozone levels in 95 U.S. cities

SOLUTIONS AND COMMENTS

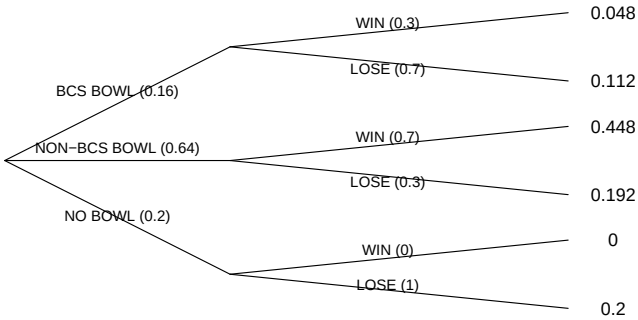
1. (a) See the following:



(b) .06+.42=.48

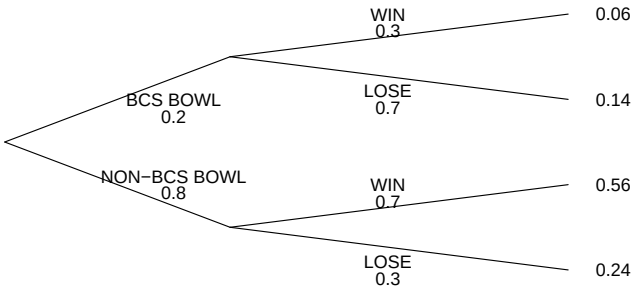
(c) $\frac{.06}{.48} = \frac{1}{8} = .125$

Remarks. Although I maintain that the above was the only correct answer to the question, there were two other solutions that proved popular. One was based on the following tree diagram (or an equivalent representation):



This would have been the correct tree diagram if part of the question had read, "Given that UNC qualify for a bowl game, the conditional probability that it is a BCS bowl game is 0.2." However, the question as worded did not specify that the 0.2 for a BCS bowl game was conditional on UNC being in a bowl game. With this interpretation of the question, the answers to (b) and (c) were .496 and .097 (or $\frac{3}{31}$). I gave 17/20 for this solution, so long as the rest of the question was correct.

The other solution that some people gave was



under which scenario the possibility "no bowl game at all" has probability 0. This seems less plausible to me, but I gave 16/20 if you did the rest of the question correctly including .62 as the answer to (b) and again .097 or $\frac{3}{31}$ for (c).

2. (a) Either of the following:

```
1 | 5
2 |
2 | 57
3 | 11
3 | 88888899
4 | 00001112
4 | 55556778889999999
5 | 000111111222334444
5 | 5555667777888899999
6 | 11122223333344
6 | 77
7 |
7 | 699
8 |
8 | 8
```

or

```
1 | 5
2 | 57
3 | 1188888899
4 | 000011125555678889999999
5 | 000111111222334444555566777888899999
6 | 1112222333334477
7 | 699
8 | 8
```

I actually think the first of these plots is more informative about the distribution, but the second one obeys all the "rules" for stem and leaf plots that we discussed in class, so I gave it equal credit.

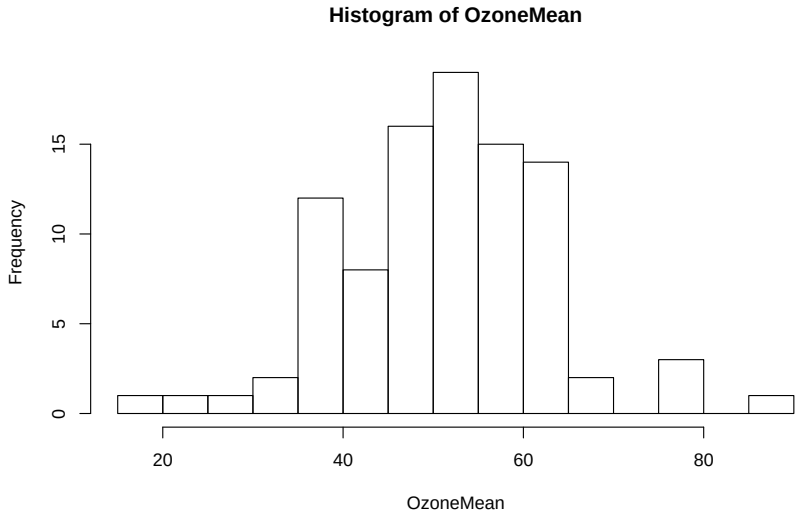
(b) (15,45,52,59,88). The IQR is 59–45=14.

(c) See diagram at top of next page (obviously there are many different choices for the interval widths and endpoints here but I was looking for something that was not *too* scraggly drawn and gave a reasonable representation of the shape of the distribution).

(d) 1.5IQR is 21 so the boundaries for outliers are 45–21=24 and 59+21=80. The smallest and largest values (15 in Honolulu, 88 in San Bernardino) are both outliers.

(e) This rule would declare something an outlier if it was more than 3 standard deviations from the mean. Thus the boundaries are $51.8-3\times11.5 = 17.3$ and $51.8+3\times11.5 = 86.3$. The same two cities are outliers.

(f) The distribution is reasonably symmetric (possibly slightly left-skewed, but not strongly so). However the presence of outliers at both ends shows it's not a perfect normal curve.



3. (a) Let X be the number of passengers who book tickets who actually show up for the flight. X has a binomial distribution with $n = 14$ and $p = 0.8$. The aircraft is overbooked if $X = 13$ or 14. We calculate

$$\begin{aligned} \Pr\{X = 13\} &= 14 \times (0.8)^{13} \times 0.2 = .1539, \\ \Pr\{X = 14\} &= (0.8)^{14} = .0440, \end{aligned}$$

resulting in a probability .1539+.0440=.1979.

(b) In this case $n = 225$, $p = 0.8$. The normal approximation is justified because $np = 180$, $n(1-p) = 45$, both of which are greater than 15. We have $\mu = np = 180$, $\sigma = \sqrt{np(1-p)} = \sqrt{36} = 6$ so the z score associated with $x = 200$ is $z = \frac{200-180}{6} = 3.33$. The left-tail probability associated with that 0.9996, so the probability that the number of passengers who show up is greater than 200 is 1–0.9996=0.0004.

(c) $z = 1.96$ would correspond to a left-tail probability of 0.95 or a right-tail probability of 0.05. For $n = 250$ we have $\mu = 250 \times 0.8 = 200$ and $\sigma = \sqrt{250 \times 0.8 \times 0.2} = 6.325$ so the number of seats that must be available is $200 + 1.96 \times 6.325 = 212.4$, or 213 to the next-highest whole number.

4. (a) The standard error is $\sqrt{\frac{.37 \times .63}{1131}} = .0144$. For 90% confidence, the margin of error is $1.645 \times 0.0144 = .0236$. The confidence interval is $.37 \pm .0236 = (.3464, .3936)$.

(b) Both the upper and lower bound of the confidence interval are well below 50%, implying that we have very high confidence that $p < 0.50$ (where p is the proportion of the population who approve Bush's policy). This comparison is slightly complicated by the 7% don't knows, but even if the don't knows were included in with those who approve, it's still clear that the proportion is well below those who disapprove. (Another way to do the calculation would be to calculate a confidence interval for the proportion who disapprove, which leads to the interval (.536,.584), strong evidence that the proportion of the whole population who disapprove is greater than 0.5.)

Remark added after grading. A lot of students equated the phrase "more people disapprove than approve" to $p < 0.5$. But this may not be true! For example, if the poll had

shown 40% approve, 40% disapprove, 20% don't know, there would be no basis at all to say "more people disapprove than approve". So you must say *something* about how the don't knows affect your answer, but I was satisfied if you simply said you didn't believe the 7% don't knows would change the result.

(c) Since we're talking about all future polls, it would not be reasonable to assume that the proportion (0.37) remains the same, so we assume it is $p = 0.5$ according to the argument that this gives a conservative choice of sample size. The margin of error calculation then leads to $m = 1.96\sqrt{\frac{p(1-p)}{n}}$ or $n = p(1-p)\left(\frac{1.96}{m}\right)^2$. With $p = 0.5,\ m = 0.012$ we have $n = 0.25\left(\frac{1.96}{0.012}\right)^2 = 6669.4...$, or a sample size of 6670. (*Remark:* In spite of the above disclaimer, some students did assume p would remain at .37, in which case the correct answer is 6219.)

5. (a) The standard error is $\frac{476}{\sqrt{11}} = 143.5$. For $df = 11 - 1 = 10$, the critical value for a 99% confidence interval is 3.169 (Table B). Therefore, the margin of error is $3.169 \times 143.5 = 455$. The confidence interval is $848 \pm 455 = (393, 1303)$.

(b) The assumptions include (i) the data are a random sample from the population of all bicycles, and (ii) the underlying distribution is normal. Both assumptions seem dubious here, (i) because there is no evidence that any kind of random selection is involved (more likely, Consumer Reports just chose bikes in a number of different categories to give what they thought was a representative sample, but we have seen that human judgment of what is random is often highly suspect); (ii) because the distribution seems right-skewed (e.g. the most expensive bike is exactly 2 standard deviations above the mean, but the least expensive is only 1.19 standard deviations below).

6. (a) Let p be the proportion of blacks (among those already in the jury pool) who are selected for an actual jury. The null hypothesis is $H_0:\ p = p_0 = 0.18$ against the alternative $H_a:\ p < p_0$. A one-sided alternative is appropriate since the stated objective is to look for bias against blacks.

(b) Since this is a hypothesis testing problem rather than a confidence interval calculation, the standard error is $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.18\times0.82}{240}} = .0248$. The test statistic is $z = \frac{\hat{p}-p_0}{S.E.} = \frac{0.0583-0.18}{.0248} = -4.91$. The P-value is the left-tail probability associated with $z = -4.91$. This is (with certainty) less than the P-value associated with $z = -4.5$, which is .000003. Therefore, the actual P-value is smaller than .000003 which implies very strong evidence *against* the null hypothesis.

(c) The study provides very strong evidence that the jury selection process is biased against blacks.

7. (a) The table is

Observation	x	$x - \bar{x}$	$(x - \bar{x})^2$
1	2.8	0.3	0.09
2	1.3	–1.2	1.44
3	4.1	1.6	2.56
4	1.8	–0.7	0.49
5	2.5	0	0
Total	12.5	0	4.58

The mean is $\bar{x} = \frac{12.5}{5} = 2.5$ and the standard deviation is $\sqrt{\frac{4.58}{4}} = 1.07$. The standard error of $\bar{x}$ is $\frac{1.07}{\sqrt{5}} = 0.478$.

(b) If μ is the mean discharge per week, then a one-sided test would test $H_0 : \mu = 1.5$ against $H_a : \mu > 1.5$. One-sided is appropriate, because the EPA would not be worried if the discharge was less than 1.5.

(c) $t = \frac{\bar{x}-1.5}{S.E.} = \frac{2.5-1.5}{0.478} = 2.09$. According to Table B with $df = 4$, the right-tail probability is between 0.10 and 0.05 (but closer to 0.05). Since it is a one-sided test, the right-tail probability is also the P-value. Since the P-value is greater than 0.05, we cannot reject H_0.

(d) The data don't prove that the company is breaking the law. However, this could simply mean there are not enough data, and the EPA might want to continue the monitoring process for a few more weeks.

8. (a) The complete table with row and column totals is

Response	Very willing	Fairly willing	Neither willing nor unwilling	Not very willing	Not at all willing	Total
1994	57	363	306	390	214	1330
2000	64	280	312	240	274	1170
Total	121	643	618	630	488	2500

(b) The table of expected values is

Response	Very willing	Fairly willing	Neither willing nor unwilling	Not very willing	Not at all willing	Total
1994	64.372	342.076	328.776	335.16	259.616	1330
2000	56.628	300.924	289.224	294.84	228.384	1170
Total	121	643	618	630	488	2500

The contributions to the χ^2 statistic are

0.844	1.280	1.578	8.973	8.015
0.960	1.455	1.794	10.200	9.111

The total is 44.209. The degrees of freedom are $(5-1) \times (2-1) = 4$ and the most extreme value for the χ^2 statistic in the table with $df = 4$ is 18.47, corresponding to a 0.001 tail probability. Since $44.209 > 18.47$, it follows that the P-value is less than 0.001.

(c) The result of the chi-squared test is highly significant, in other words, the null hypothesis of independence is rejected. Looking at the table of contributions to the χ^2 statistic, it is seen that the highest contributions are from the last two columns, in particular, a substantial shift of people from the "not very willing" category into "not at all willing". Politically, this probably means that people on the right of the political spectrum, who opposed Clinton's/Gore's environmentalist views, hardened their opinions and became more opposed while Clinton was president, while people on the left of the political spectrum basically agreed with what Clinton was doing and did not change their views.

Remark. Part (c) was the worst answered part here! A lot of students did the chi-squared test correctly without looking at the table to see where the independence assumption was breaking down. But there's no point knowing how to do a chi-squared test (or any kind of statistical test) if you don't know how to interpret it in the light of the actual data!