Statistics 101 Summer I 2011
Midterm Examination

Adam Kapelner, Instructor

June 13, 2011, 8:30-10:30AM

First Name ______________________ Last Name __________________________________

University of Pennsylvania's Code of Academic Integrity

Since the University is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the University community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a students performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using a cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold the University of Pennsylvania's Code of Academic Integrity.

____ ____________________
signature date

Instructions

This exam is two hours and closed-book. Cheat sheets are not allowed. You may use a graphing calculator of your choice. Please read the questions carefully. I advise you to skip difficult problems until you have finished the exam, then loop back and plug in all the holes. I also advise you to use pencil.

The exam is 100 points total. Partial credit will be granted for incomplete answers on most of the questions. Good luck!

Problem 1 This short-answer section will ask basic questions about probability.

(a) [3 pt] What is the definition of probability? (you may use any of those that we discussed in class, but please explain to the best of your ability)

(b) [2 pt] Below is Boole's inequality (AKA Bonferroni's inequality):

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(A_i\right)$$

Express this inequality for $n = 2$ *i.e.* for the events A_1 and A_2.

(c) [2 pt] Under what condition(s) is Boole's inequality an equality (*i.e.* $=$ and not $<$)?

(d) [2 pt] Under what condition(s) is Boole's inequality a strict inequality (*i.e.* $<$ and not $=$)?

(e) [2 pt] Draw the sample space for a coin flip and spin of a spinner with sections for "A", "B", and "C".

Problem 2 You want to survey the residents of Philadelphia in order to find out about their dieting habits. You get a list of all apartment buildings nearby and randomly pick 40 buildings. You then make a list of every apartment in each of the 40 buildings and you randomly pick 250 from this list using a sorted random number column in JMP. You then survey those willing to talk with you (a maximum of one per apartment).

(a) [2 pt] What is the maximum size of your sample?

(b) [5 pt] If your intention is to make inference about the diets of all residents of Philadelphia, is this a properly designed survey? If not, explain the problem(s) with the survey.

(c) [2 pt] If they exist, can the problem(s) be fixed in order to make non-biased inference about dieting habits using this data?

(d) [3 pt] If instead you sampled 10 apartments from each type of building (luxury, storefront, federal housing, etc), what kind of sampling would this be called?

Problem 3 Adam has 10 shirts: 2 are red, 3 are blue, 2 are black, and 3 are white.

(a) [3 pt] When Adam dresses for a lecture, he picks a random shirt out of his closet. After the lecture, he hangs it up back in his closet. Every morning he forgets which shirt he wore the previous day. Consider each shirt unique even though they may be of the same color. If there are 23 lectures, how many different ways can he wear shirts for the entire summer session?

(b) [3 pt] Consider the unlikely situation that after a lecture, he throws the shirt in a hamper and does not wear it the next day. How many different day-shirt combinations are there in one week of lectures (*i.e.* four lectures)?

(c) [4 pt] What is the probability he wears two red shirts in a week of lectures?

(d) [4 pt] What is the probability he wears one red shirt in a week of lectures?

(e) [3 pt] Red shirts are special because they have to be dry-cleaned in a special way. Other shirts can be washed. Consider the cost of washing negligible since he has to do a wash anyway. Each red shirt worn the week prior will become dirty and costs \$3 to be dry-cleaned.

Create a r.v. X that models the cost of Adam's dry-cleaning bill. Use the $\sim$ and brace notation used in class and indicate units. If you do not know how to do this, make up a r.v. X that is *reasonable* for this model so you can continue with the rest of the problem after.

(f) [2 pt] Draw the PMF for X. Indicate axes and units clearly.

(g) [2 pt] Adam teaches for five weeks. Each week Adam drops off his dry cleaning on Thursday and picks it up on Sunday. Denote the first week's bill by X_1, the second week's bill by X_2, ..., and the fifth week's bill by X_5. Are $X_1, X_2, ..., X_5 \overset{iid}{\sim}$ with the PMF you just drew? Justify each assumption or explain why it does not hold.

(h) [5 pt] Regardless of what you wrote in the previous part, assume the weekly dry cleaning bills are $\overset{iid}{\sim}$ r.v.'s with the PMF from part (f) from now on. Find the expected total dry cleaning expenses and the standard deviation of total dry cleaning expenses.

(i) [3 pt] For those five weeks also find the expected average of his weekly dry cleaning bill and the standard deviation of the average weekly dry cleaning bill.

(j) [3 pt] Adam manages to get a 10% discount and also tips $8 at the end of the summer. Find the expected total bill and the standard deviation of the total bill.

Problem 4 Dell, Inc. was founded by Michael Dell in 1984 while he was a student at UT. By 1988, it was a publically traded company. In this question, we will be investigating a model for technical support by phone.

(a) [2 pt] In 1984, Dell got his first three customers — three unrelated business managers in the Austin, TX metro area. However, his computers weren't polished products and had many defects. During any business day, his three customers could call him personally asking for help. There was a 10% chance each would call in. How would you model these three customers' possible calls? Use the notation X_1, X_2, X_3 for the three customers and use the r.v.'s we learned about in class.

(b) [3 pt] What is the probability that two customers would call in in a single day?

(c) [4 pt] By the beginning of 1985, he had 100 customers and his computer-assembly process was greatly improved. Now there was only a 3% chance each would call in. What is the expected number of calls per day and the standard deviation of the number of calls?

(d) [4 pt] By the middle of 1985, he had 577 customers and improved his computers to the point where only his customers called in daily with 1.5% chance. What is the probability he got more than two calls per day?

(e) [2 pt] Business kept growing and growing for Dell and his ability to make better computers also improved. By 1988, Dell had 400,000 customers and the probability of each of them calling in was 0.7%. Model a r.v. N for the daily number of calls in 1988 using an approximation we learned about in class. Indicate the parameter(s) clearly.

(f) [4 pt] Using the approximation from part (e), find the probability exactly 100 people call in during one day in 1988.

Problem 5 A drivers education website requires students to read an essay on drunk driving as part of its curriculum. Below is an excerpt:

> As more alcohol is consumed the risk of getting into a vehicular accident if the person drives grows. For example, a man that weighs about 160 pounds would have a BAC of 0.04 an hour after drinking two beers. Its still way below the limit of driving under the influence but the likelihood of getting into an accident is 1.4 times more probable than [the national average]. Add two more beers then the probability goes up tenfold. Make it a six pack with two more beers, the drinker reaches the limit of 0.10 BAC and the risk is now 48 times more than [the national average]. Add two more for the road and you reach 0.15 BAC well above the legal limit and the risk is now 380 times than the [the national average]. Drunk driving is never an option...

During another part of the curriculum, they read excerpts of the National Safety Council's (NSC) report on traffic fatalities countrywide:

> The motor-vehicle death rate per 100,000,000 vehicle-miles was 1.54 in 2005...

Consider the NSC's statement to be accurate as an average across all accidents, including those under-the-influence of drugs and alcohol.

Another part of the curriculum is learning about the prevalence of drunk driving:

> WASHINGTON (AP) An estimated 17 million people have driven while drunk at least once on U.S. streets and highways in the course of a year, according to a government study released Wednesday...

We estimate that 1% of the people that "have driven drunk" drive with 0.15 BAC or over and there are 300 million people in United States.

Consider the drunk driving risk literature as well as the newswire to be ballpark-accurate for the year 2005. Consider the distance between your friend's apartment and your apartment to be 10 miles. Denote the event of getting into an accident during these 10 miles as "A". Denote the event of driving while completely drunk (*i.e.* 0.15 BAC) as "D". Please use this notation.

(a) [3 pt] Use the NSC report excerpt to find the probability of getting into a fatal accident while driving from your friend's apartment to your apartment.

(b) [3 pt] What is the probability of getting into a fatal accident while driving home from your friend's party on Saturday night after you drank heavily?

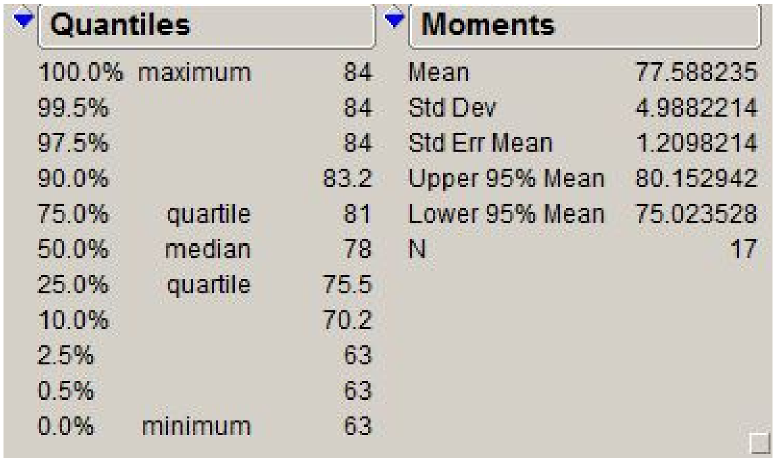
(c) [4 pt] If you got into an accident, what was the probability you were drinking heavily?

(d) [4 pt] What is the probabilty of *not* getting into a fatal accident while driving from your friend's apartment to your apartment after you did the Stat 101 homework (*i.e.* *no* drinking)? Hint: you may want to leave the values in the calculator and keep values to many significant digits.

Problem 6 Consider a the following data of heights in inches of an NBA team's roster. The data is already sorted from smallest to largest:

$$63, 72, 74, 75, 76, 76, 77, 78, 78, 79, 80, 80, 80, 82, 82, 83, 84$$

Here is an excerpt of `JMP` output for the above data.



```
Quantiles                          Moments
100.0%  maximum      84            Mean              77.588235
99.5%                84            Std Dev           4.9882214
97.5%                84            Std Err Mean      1.2098214
90.0%              83.2            Upper 95% Mean    80.152942
75.0%    quartile    81            Lower 95% Mean    75.023528
50.0%    median      78            N                        17
25.0%    quartile  75.5
10.0%              70.2
2.5%                 63
0.5%                 63
0.0%     minimum     63
```

(a) [1 pt] What is the modal value?

(b) [1 pt] What type of data is our univariate dataset?

(c) [1 pt] Begin to calculate the sample standard deviation, s. Write out the first few terms but do not compute.

(d) [1 pt] Is there a skew? If so, what type of skew?

(e) [8 pt] Draw a box and whisker plot for this dataset to scale. Use 2.4 as the half-width of the diamond. Denote the outlier(s) and the salient points of the box (do not denote the values at the ends of the whiskers). Use the JMP output; do not compute the percentiles by hand.