

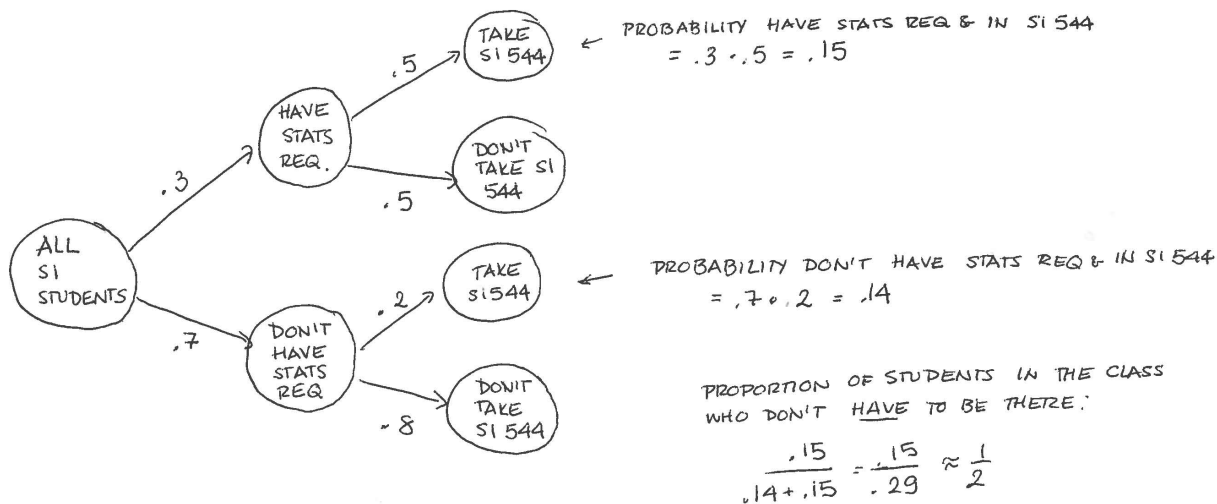
SI 544 Midterm Winter 2008

NAME: _____

Feb. 18, 2008

1 Probability (10 pts)

The probability that an SI masters student chooses a specialization with a statistics requirement is 0.3. The probability that a student who needs to fulfill the stats requirement takes SI 544 is 0.5 (the other 50% have either taken statistics before or elect to take stats in another department). The probability that a student who does not have to fulfill a stats requirement takes SI 544 is 0.2. What is the probability that a student sitting in SI 544 is there because they have to fulfill the stats requirement? Draw a probability tree and show your calculation. You can leave your answer as a fraction.



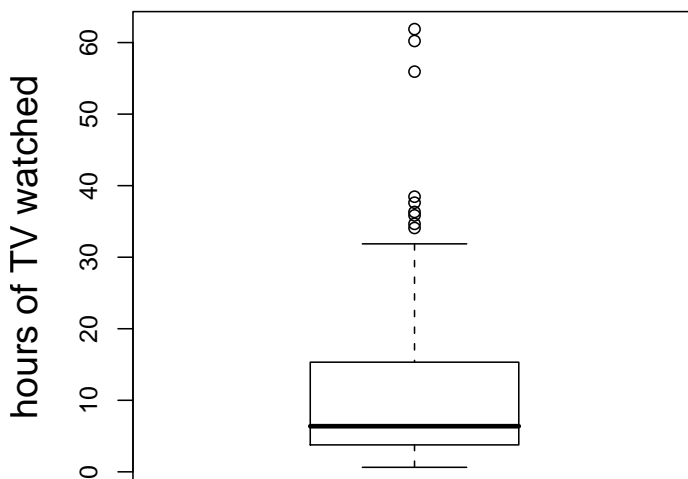
2 Probability distributions

(16pts) For each of the following, place a single letter (B, H, P or N), corresponding to the most suitable distribution to use in each case: **(B) binomial**, **(H) hypergeometric**, **(P) poisson**, **(N) normal**.

- **_P_** (2pts) The average number of goals scored in a World Cup soccer game is 2.51. Assuming that the probability of a goal occurring in any given time interval is the same across games, what distribution would you use to figure out the probability of attending a game where 6 or more goals are scored?
- **_B_** (2pts) You flip an unbiased coin 10 times and wonder what the probability is that 8 of the tosses turn up the same (either 8 heads or 8 tails).
- **_H_** (2pts) You have shown up for a game of ultimate frisbee. There are 13 other players besides you, 7 players on each team. There are 10 players you'd like to be on your team, and the other 3 you'd rather not have on your team. If players are allocated randomly between the two teams, what is the likelihood that all 3 of the ones you don't like end up on your team?
- **_B_ (P is also a good approximation)** (2pts) Each user visiting your website has a 1% chance of leaving a comment. What is the probability that out of 100 users, 5 or more leave a comment?
- **_H_** (2pts) You purchase 4 rose bushes from nursery A and 6 rose bushes from nursery B. All 10 rose bushes are of the same hybrid tea rose variety. You planted the rose bushes along your front white picket fence (in random order), and treated them all with equal care. At the end of the year, only 5 of the rose bushes survived, all of them from nursery B. Should you be suspicious about the quality of rose bushes from nursery A?
- **_P_** (2pts) Your gardening book says that an average of 5 Japanese beetle grubs per square foot of lawn is OK. You find you have 20 grubs in the square foot of lawn you sampled, should you be worried?
- **_B_** (2pts) You claim that you have a psychic power that allows you to guess the winning contestant on Jeopardy before the game starts. There are 3 contestants in each game. Out of 10 games, you picked the winner correctly 5 times. Dumb luck or not?
- **_N_** (2pts) For all of the above, this distribution would describe the mean of 30 or more samples (e.g. the average number of goals scored in a sample of 50 soccer games, the average number of heads in 50 experiments having 10 coin tosses each, etc.)

3 Box plots and t-tests

The UofM provost has funded your study to figure out the TV watching habits of undergraduates. In exchange for \$20, you get 100 undergraduates to log their total hours of TV watching during 1 week (or rather, they note the start and stop times of all their TV watching sessions, and you end up having to add them up). At the start of the week you have the participants give you their best estimate of how many hours of TV they watch per week. You also collect their GPAs and majors. The following is a boxplot of the number of hours of TV watched.



(a) (3 pts) Is the distribution left or right skewed?

Right skewed.

(b) (3 pts) Give the approximate (within 2 hours) median number of hours of TV watched in your sample.

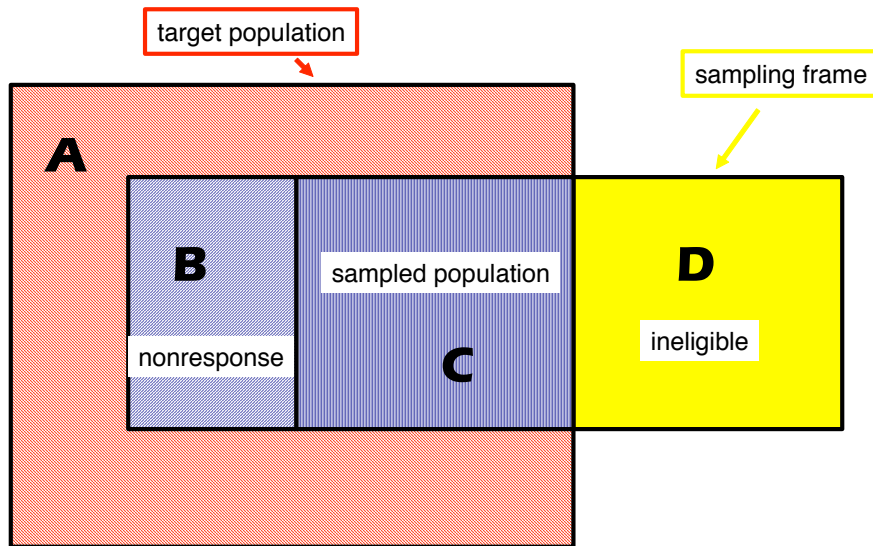
The median looks like it might be 7.

(c) (8pts) Check those hypotheses for which a t-test would be appropriate

- The mean number of hours watched by the undergraduate student population is 7.
- Chemistry majors watch more TV on average than history majors do.
- Students tend to underreport on average the amount of TV watching they do. (I had actually intended to ask whether students *underestimate* the amount of TV they watch).
- Each additional hour of TV watched per week corresponds to an average decrease of 0.05 in GPA.

4 Sampling (35 pts)

You are interested in figuring out the average amount of money donated by just the students in the penny war. You camp out on the West Hall stairwell one day and ask any live being going by to tell you how much they donated to the campaign.



Note that areas A,B,C and D are mutually exclusive.

(a) Describe which students correspond to which area of the figure. I've filled in the first one.

Area in Figure	students
target population (includes A,B&C)	--- all SI students ---
(3 pts) sampling frame (includes B,C&D)	-- people using the WH stairwell --
(3 pts) area A	--- SI students who did not use the stairwell ---
(3 pts) area B	--- SI students who used the stairwell but declined to answer the question ---
(3 pts) area C	--- SI students who used the stairwell and answered your question ----
(3 pts) area D	--- non SI students who were using the stairwell ----

In the first 1/2 hour, you have answers from 9 SI students. From those 9 students you have a mean of \$8.45 donated with a standard deviation of \$6.

(b) (5pts) Construct an approximate 95% confidence interval for the mean amount donated by a student during the penny war (your confidence interval should be numeric, but you can leave fractions as fractions)

$$SEM = s/\sqrt{n} = 6/\sqrt{9} = 6/3 = 2$$

I am constructing an interval that is $[m - 2*SEM, m + 2*SEM] = [8.45 - 4, 8.45 + 4] = [4.45, 12.45]$

(since the sample is small, an even more accurate solution would use the 95% from the t-statistic).

(c) (5pts) You decide to keep surveying students and to stop once you have (roughly) 80 students in your sample. At this point you expect the standard deviation in the amount donated by an individual student in your sample to be (relative to the 9-student sample):

(i) approximately unchanged - I am asking for the standard deviation of an *individual observation*, and that won't change much, because for the variance, you have n terms $(X_i - \bar{X})$, and you have $(n - 1)$ in the denominator. Or differently said, your sample standard deviation is approximately the standard deviation of the underlying population, no matter the sample size.

(d) (5pts) With a sample of 80 students, you expect the width of your 95% confidence interval for the population mean to be (relative to the confidence interval of the 9-student sample):

(v) one third as wide - the SEM has \sqrt{n} in the denominator, and I've gone from 9 to a sample size that is roughly 9 times as large. So my SEM drops to a third and therefore my confidence interval does too.

(e) (5pts) Think back to your sampling technique. Should you trust your estimate? Explain briefly.

My sample is probably biased. First of all, students who go up the west hall stairwell often are both more likely to have contributed to the penny war and to be there when I'm conducting a survey.

Second, students who donated money are more likely to answer my survey.

5 Inference errors (5pts)

The pairwise t-test inflates the p-values in order to lessen what kind of error?

(i) type I - with higher p-values I am less likely to reject the null when I really should have kept it (especially since I am making multiple comparisons, making spurious significant results more likely to occur at least once). A type I error is when I reject the null, but I should have kept it.

6 Experiments and observations (15pts)

You are kindergarden teacher.

You would like to figure out if children who take a 1 hour nap do better in spelling tests. For your experiment, specify the following:

(a) H_0 (3pts):

The mean spelling test score of kids who take naps is no different from the mean score of kids who don't.

(b) independent variable (2pts):

Whether or not the kid takes a nap.

(c) dependent variable (2pts):

A kid's score on a spelling exam.

(d) specify how you would set up your experiment to get a statistically significant result (if there is one to be had) and, if a significant result occurs, to demonstrate causality between napping and spelling performance (5pts) :

There are two choices, one is to randomly assign some kids to take a nap, and others to stay awake. The other choice is to have the same kids nap or not nap, but to randomly assign some kids to nap on the first day, and the others to nap on the second day. This way I'd control for kids getting better at spelling from day to day, while being able to do a paired t-test on the matched samples for the same kids under two conditions (nap and no nap).

(e) Given your setup in (d), which statistical test would you use(3pts) (circle one)

:

- (i) one sample t-test (ii) ordinary two-sample t-test
(iii) paired two-sample t-test (iv) pairwise multiple sample t-test

If I don't have the same kid try both napping and not napping, then I would use the ordinary t-test (ii).

If I do have the same kid napping and not napping, then I would use a paired t-test (iii).