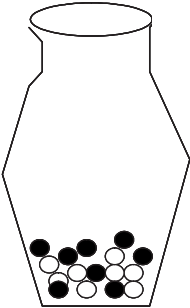
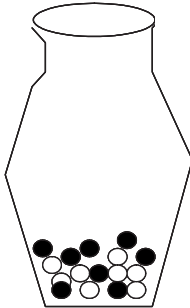
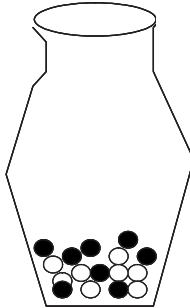
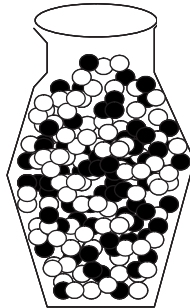


Midterm SI 544 fall 2006 solution

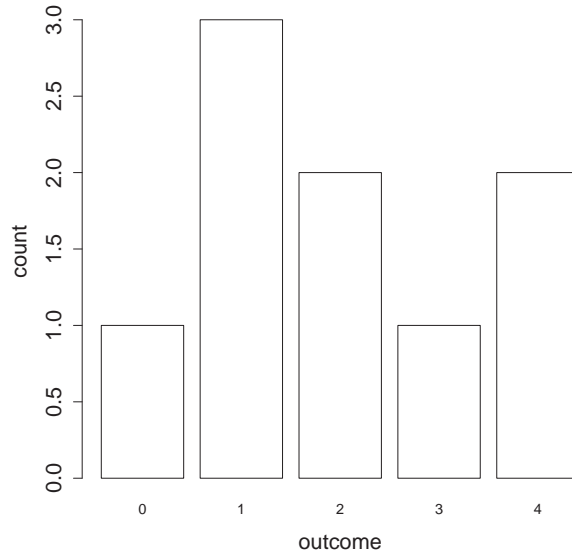
Name: _____

- Imagine you have an urn which contains some black and some white balls. Suppose you will put your hand in the urn without looking at the balls and draw, one by one, n balls from the urn and note their color. If you make a draw without replacement, you will not place the ball back in the urn, but draw the next ball from those that remain in the urn. If you draw with replacement, each time you will put the ball back in the urn, before drawing the next one. For each of the 4 scenarios below, list from these three distributions (hypergeometric, binomial, normal), which, if any, exactly describe your experiment and which, if any, are approximations. For example, if the hypergeometric is the exact solution, and the normal is a good approximation, I would write “hypergeometric (exact)” on the first line, and “normal (approx.)” on the second line below the appropriate urn.

draw $n=10$ balls <i>without</i> replacement	draw $n=10$ balls <i>with</i> replacement	draw $n=100$ balls <i>with</i> replacement	draw $n=10$ balls <i>without</i> replacement
			
N=16	N=16	N=16	N=200
=====	=====	=====	=====
=====	=====	=====	=====
=====	=====	=====	=====

hypergeometric (exact)	binomial (exact)	binomial (exact)	hypergeometric (exact)
		normal (approx.)	binomial (approx.)
(5pts)	(5pts)	(5pts)	(5pts)

2. You surveyed several of your classmates about how many books they have on loan from the library. You binned and plotted your results, obtaining the following histogram:



Please answer the following, writing down, where appropriate, the equation you used to get your answer. Your final answer need not be a single number (if you don't have R or a calculator handy to calculate it), but you should substitute numerical values into the equation. If you can't compute one value, but need it for another, e.g. using the mean when calculating the variance, just write the word 'mean' in the appropriate place in the equation.

- (3 pts) What is the number of observations made? 9
- (3 pts) What is the mode? 1
- (4 pts) What is the median? 2
- (5 pts) What is the mean?

$$m = \frac{\sum x * f(x)}{n} = (0 * 1 + 1 * 3 + 2 * 2 + 3 * 1 + 4 * 2) / 9 = 18 / 9 = 2$$

- (5 pts) What is the variance?

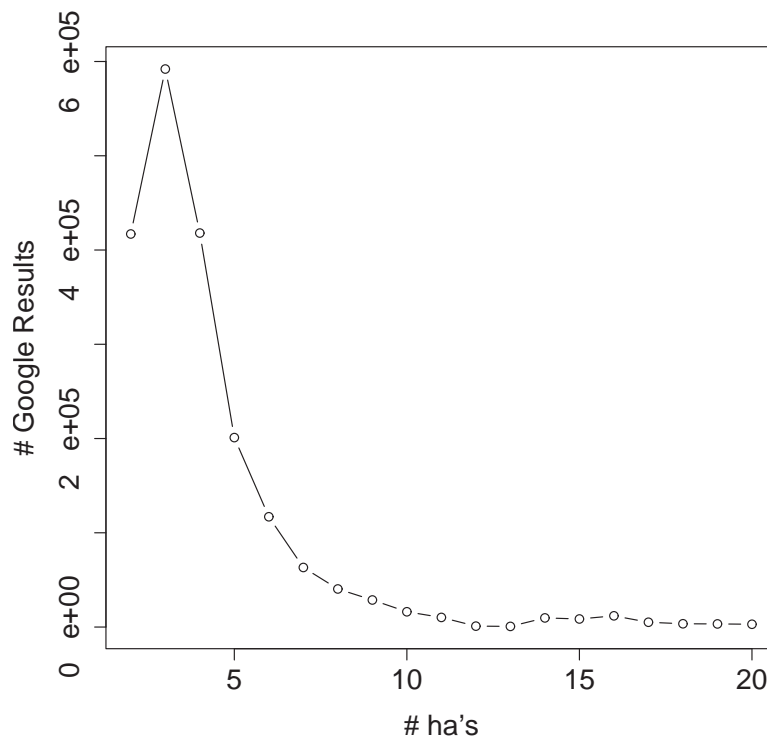
$$s^2 = \frac{\sum (x-m)^2 * f(x)}{n-1} = (2^2 * 1 + 1 * 3 + 0 * 2 + 3 * 1 + 2 * 4) / 8 = 16 / 8 = 2$$

- (5 pts) What is the sample estimate of the standard error of the mean?
 $SEM = s / \sqrt{9} = \sqrt{2} / 3$

- (5 pts) If you were to survey 4 times as many people, the standard error of the mean would
 (a) decrease by 50%

3. You are researching the occurrence of ‘evil laughs’ (with or without stroking a white cat) online, and so have issued the query ‘bwahaha...’ with varying numbers of ha’s to Google. Your results are in the table below:

query	# of ha’s	# Google results
bwahaha	2	417000
bwahahaha	3	592000
bwahahahaha	4	418000
bwahahahahaha	5	201000
bwahahahahahaha	6	117000
bwahahahahahahaha	7	63200
bwahahahahahahahaha	8	40400
bwahahahahahahahahaha	9	28700
bwahahahahahahahahahaha	10	16200
bwahahahahahahahahahaha	11	10200
bwahahahahahahahahahaha	12	967
bwahahahahahahahahahaha	13	690
bwahahahahahahahahahaha	14	9690
bwahahahahahahahahahaha	15	8560
bwahahahahahahahahahaha	16	11900
bwahahahahahahahahahaha	17	5090
bwahahahahahahahahahaha	18	3480
bwahahahahahahahahahaha	19	3310
bwahahahahahahahahahaha	20	2970



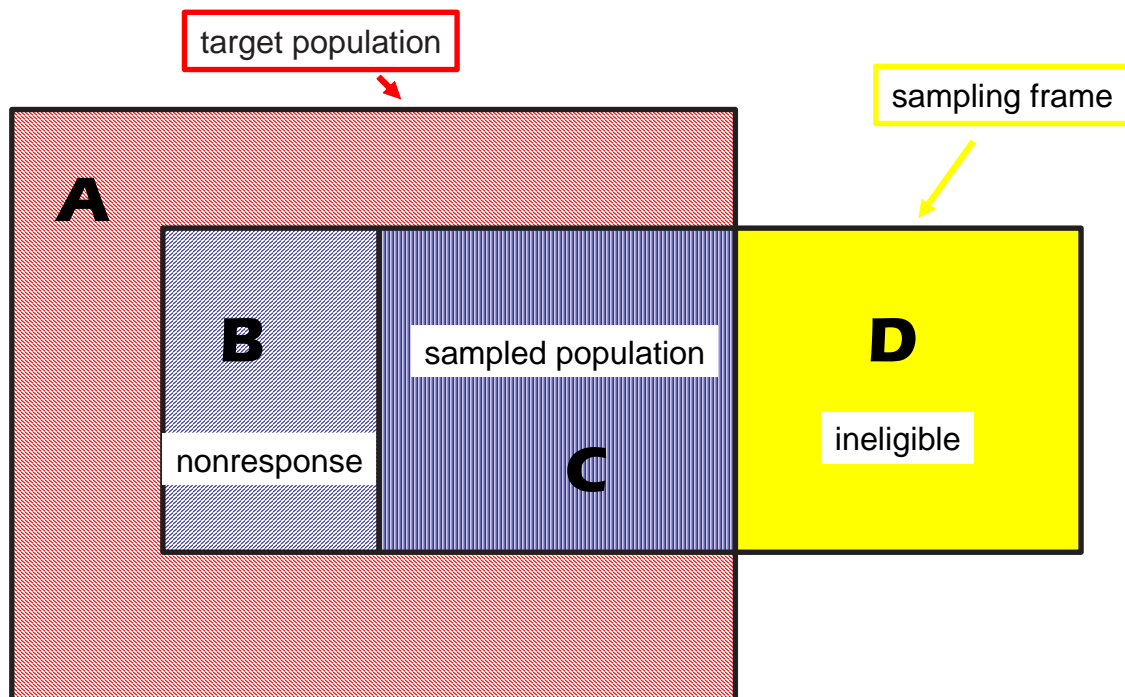
You've plotted them (the figure is on the previous page), and you are getting a sense of the

- mode 3 (fill in) (3pts)
- skew: (b) right-skewed (5pts)

But you're not really sure what is going on for higher numbers of ha's (10-20), because the curve appears nearly flat. How would you modify the plot in order to visualize the trend for higher numbers of ha's? (there are several OK choices, one will suffice) (7pts).

One option is to plot all the data on a log-log scale. Another is to plot just the tail using `clim()`. Yet another is to use `lowess` to draw a trend curve (probably still while limiting the y-axis using `clim()`).

4. (5 pts) You change your decision rule to reject the null hypothesis when $p < 0.01$ instead of when $p < 0.05$. By making this change, you are increasing your chances of committing what type of error?
(b) type II
5. You want to conduct a survey of SI students' study habits. You decide to go to the UGLi (undergrad library at the University of Michigan) and conduct surveys via face-to-face interviews with any students you find there.



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

Describe which students correspond to which area of the figure. I've filled in two of them.

Area in Figure	students
target population (includes A,B&C)	<u>--- all SI students ---</u>
(4 pts) sampling frame (includes B,C&D)	students in the library
(4 pts) area A	SI students who were not in the library
(4 pts) area B	SI students who were in the library but did not take the survey
(4 pts) area C	SI students who were in the library and took the survey
area D	<u>-- students who were in the library but are not SI students --</u>

(5 pts) Give one reason why your sample might be biased.

Students who study in the library may not have the same study habits of students who don't.

(4 pts) If the probability that any particular SI student is in the library at the time of your survey is 0.05, and the probability that a student will take a survey when asked is 0.8, what is the probability that an SI student will take your survey?

0.05*0.8

(b) 0.04