STAT 305 – Handout 9
Binomial Distributions

Next we will examine several special cases of discrete random variables that arise often enough to be widely studied and to have their own names. We start with the most important and widely used discrete probability distribution.

Example 9-1: Quiz Guessing

Suppose that you take a multiple choice quiz for which you guess randomly among the options on every question (well, not really you, of course, I mean some clueless student instead). Suppose there are 5 questions, with 3 options to choose (guess) from on each question.

a) Here are 5 questions with 3 options each. Circle your answers. (Be careful: #3 is not as straight-forward as it looks!)
   1. A   B   C
   2. A   B   C
   3. A   B   C
   4. A   B   C
   5. A   B   C

I’ll reveal the correct answers and ask you to grade your own quiz on the honor system. Then we’ll have a show of hands for the number of questions answered correctly.

Let the random variable X be the number of questions that you answer correctly.

b) What are the possible values of X?

c) Determine the (exact, theoretical) probability that you answer all 5 questions correctly. Also indicate what rule you use to do this calculation.

d) Determine the probability that you answer 0 questions correctly.

e) Determine the probability that you answer exactly 2 questions correctly. [Hint: First find the probability of one particular sequence with exactly 2 correct. Then multiply by the number of such sequences.]
f) Determine a general expression for the probability that you answer $x$ questions correctly by random guessing on this 5-question, 3-option exam.

g) Now generalize this situation to an exam with $n$ questions and probability $p$ of answering any one question correctly. Determine a general expression for the probability of answering $x$ questions correctly in this case.

- A random variable $X$ is said to have a **binomial** distribution with parameters $n$ and $p$ if:
  - $X$ can be thought of as the number of successes in $n$ trials
  - Each trial has two possible outcomes (S, F)
  - The trials are independent
  - The probability of success is the same ($p$) on every trial
- Notation: $X \sim \text{Bin}(n, p)$
- The pmf for a binomial random variable $X$ is: $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \ldots, n$.
- It can be shown that $E(X) = np$ and $\text{Var}(X) = np(1-p)$.
- Calculations in R:
  - pmf: `dbinom(x, n, p)`; cdf: `pbinom(x, n, p)`, inverse cdf: `qbinom(prob, n, p)`
  - simulation: `rbinom(numreps, n, p)`

h) Use R to graph the pmf of the binomial distribution above and then the cdf at these 6 points.

i) Suppose that you must answer at least half of the questions correctly in order to pass the 5-question quiz with 3 options per question. Determine the probability of passing, assuming that you guess randomly on each question. First calculate this by hand from the pmf, and then use both the pmf and cdf in R.
j) Would you expect this probability of passing (getting at least half correct) to be larger, the same, or smaller, if the quiz consists of 15 questions rather than 5? Explain your reasoning.

k) Determine the probability of passing the quiz if it consists of 15 questions rather than 5. Use good notation to express this probability, and then use R to calculate it. Is your intuition from the previous question supported?

l) How would you expect the probability of passing the 15-question quiz to change if each question has 5 options rather than 3? Explain your reasoning.

m) Make this change (from 3 options per question to 5), and recalculate the probability of passing the 15-question quiz. How does this compare to the previous one?

Example 9-2: Naughty or Nice?
We all recognize the difference between naughty and nice, right? What about children less than a year old—do they recognize the difference and show a preference for nice over naughty? In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual’s actions towards others in evaluating that individual as appealing or aversive, perhaps laying for the foundation for social interaction (Hamlin, Wynn, and Bloom, 2007). In one component of the study, 10-month-old infants were shown a “climber” character (a piece of wood with “google” eyes glued onto it) that could not make it up a hill in two tries. Then they were alternately shown two scenarios for the climber’s next try, one where the climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with. The researchers found that the 14 of the 16 infants chose the helper over the hinderer. (You can see the videos at: http://www.yale.edu/infantlab/socialevaluation/Helper-Hinderer.html.)
a) If the infants really had no preference between the toys and so were just choosing a toy at random, what is the probability that 14 or more of the 16 infants would have chosen the nice toy? Explain how your answer is based on the binomial distribution. Perform the calculation both by hand and with R.

b) Is this probability small enough to cast strong doubt on the “no preference” assumption, which would suggest that infants really do have a preference for the nice toy? Explain.

- This probability is a binomial p-value.
  - Small p-values provide strong evidence against the “no preference” (null) hypothesis.
    - P-values less than .05 are generally considered to provide strong evidence against the null hypothesis.
    - P-values less than .01 are generally considered to provide very strong evidence against the null hypothesis.

c) How many of the 16 infants would have to choose the nice toy, in order for the probability to be no more than .05 of obtaining that many or more choosing the nice toy under the “no preference” assumption? Show how to use R to determine these values, first by trial-and-error and then with the qbinom command.

- These value are called the rejection region of the test.
  - The value .05 is called the significance level of the test.

d) Suppose that the infants actually have a .75 probability of choosing the nice toy. Determine the probability that the result for 16 infants would fall in the rejection region with a significance level of .05. (Indicate how to calculate this probability by hand, but do the calculation with R.) Also interpret the probability. Finally, show how to use R to do this all in one command that combines both pbinom and qbinom.
- The **power** of a test is the probability that the test would successfully detect an effect when an effect is actually present.

e) Suppose that the study had involved twice as many infants (32), with the same proportion choosing the nice toy. How would you expect the p-value to change, if at all? Explain your reasoning.

f) Calculate the p-value for the scenario in e). Indicate how to do this with the binomial distribution, and use R to perform the calculation. Is your intuition in e) confirmed or refuted?

g) Continue to assume that the study had involved twice as many infants (32). Determine the rejection region with a significance level of .05. Also determine the power of the test when the actual probability of choosing the nice toy is .75. Describe how these have changed and why that makes intuitive sense.

**Example 9-3: Flat Tire?**
A legendary story on college campuses concerns two students who miss a chemistry exam because of excessive partying but blame their absence on a flat tire. The professor allowed them to take a make-up exam, and he sent them to separate rooms to take it. The first question, worth five points, was quite easy. The second question, worth ninety-five points, asked: Which tire was it? I will ask each of you to indicate which tire you would pick. Do not confer with anyone else before answering.

a) Which tire would you pick?

b) Name the tire that I predict to be the most popular choice.
c) Record the counts for the class data below.

<table>
<thead>
<tr>
<th>Left front</th>
<th>Left rear</th>
<th>Right front</th>
<th>Right rear</th>
</tr>
</thead>
</table>


d) The null model asserts that my conjecture is wrong and there is nothing special about the right front tire, so it is equally likely to be picked as any other tire. Let the random variable \( X \) be the number of students in our class who would choose the right front tire. Under the null model of “nothing special” about this tire, describe the probability distribution of \( X \).

e) Indicate how to calculate the p-value, and then use R to calculate it.

f) Is this p-value small enough that the class data provide fairly strong or very strong evidence against the null model? Explain the reasoning process behind your answer.

Example 9-4: Random Babies (cont.)
Consider yet again the “random babies” process with 4 mothers/babies. Suppose that you simulate this process 10 times. Let the random variable \( X \) be the number of times that you obtain 0 matches.

a) What probability distribution does \( X \) have? (Specify its parameter values as well as its name.)

b) Determine and interpret the expected value of \( X \).

c) Determine the variance and standard deviation of \( X \).

d) Determine the most likely value of \( X \) and its probability. [Hint: Use R to look at the entire pmf.]

e) Determine the probability that you obtain 0 matches more than half the time in these 10 repetitions.