Lecture 7

Last time

The connection between randomness and probability.

Each time we roll a die, the outcome is random (we have no idea and no control over which number, 1-6, will land face up). In the long-term, if we flip the die forever we can see the probability of each outcome (each number has a 1/6 chance of occurring).

Does one roll of a 6 mean that the chance of getting a 1-5 increases on my next roll?

Differences in sampling distribution and population distributions.

> We use statistics like $\bar{X}$ in order to infer something about the parameter μ. We call $\bar{X}$ a **random variable** since it changes with each new random sample chosen.

> Things to notice about how $\bar{X}$ relates to :
> $\bar{X}$ is the statistic that estimates μ
> μ, in practice, is fixed but rarely known
> If a different SRS is taken, we'd have a slightly different $\bar{X}$. $\bar{X}$varies from sample to sample
> The larger the sample size, the closer $\bar{X}$ is to μ

Also, we discussed looking at a distribution as a *probability* distribution.

X value	**Probability**
x_1	p_1
x_2	p_2
x_3	p_3
.	.
.	.
.	.
x_k	p_k

Note that these probabilities must satisfy two requirements:
1. all probabilities, p_i, must be between 0 and 1
2. all the probabilities add up to one: $p_1 + p_2 + p_3 + ... + p_k = 1$

Binomial Distribution for count variables

In Lecture 6 we talked about finding the probability of getting 23 6s when we rolled a die 100 times. We simulated the results by doing this action 109 times and seeing how often we got 23 6s. We were simulating a **binomial distribution** where the number of trials was 100 (rolls of the dice) and the probability (of getting a 6) was 1/6. Anytime we have independent trials, such as rolling a die, we can use the binomial distribution to quickly find the probability of a certain number of occurrences.

In many applications, an observed outcome is **binary**: it has one of two possible outcomes. For instance, a person may:
- accept, or decline, an offer from a bank for a credit card
- have, or not have, health insurance
- vote yes or no in a referendum, such as whether to recall a governor from office

With a sample, we summaries such variables by counting the number and proportion of cases with an outcome of interest.

Example:
Say we have a sample of size n=5. Let the random variable X (it is random because it changes with each sample) denote the number of people who vote "yes" on some issue in a referendum. The possible values of X are 0, 1, 2, 3, 4 and 5. If certain conditions are met, this random variable X that counts the number of observations of a particular type has a binomial probability distribution.

Conditions for a binomial distribution

- Each of the n trials has two possible outcomes. The outcome of interest is called a "success" and the other outcome is called a "failure"

- Each trial has the same probability of a success. This probability is denoted by p.

- The probability of a failure is denoted by $1 - p$.

- The n trials are independent. That is, the result for one trial does not depend on the results of other trials.

If these conditions are met, then X has a binomial distribution with parameters n and p, abbreviated $B(n, p)$.

NOTE: The sampling distribution of a count variable is only well-described by the binomial distribution is cases where the population size is significantly larger than the sample size. As a general rule, the binomial distribution should not be applied to observations from a simple random sample (SRS) unless the population size is at least 20 times larger than the sample size.

Example:

Suppose individuals with a certain gene have a 0.70 probability of eventually contracting a certain disease. If 100 individuals with the gene participate in a lifetime study, then the distribution of the random variable describing the number of individuals who will contract the disease is distributed $B(100, 0.7)$.

Example:

Daisy the psychic claims to possess extrasensory perception (ESP). An experiment is conducted in which a person in one room picks one of the integers 1, 2, 3, 4, 5 at random and concentrates on it for one minute. In another room Daisy identifies the number he believes was picked. The experiment is done with three trials. After the third trial, the random numbers are compared with Daisy's predictions. Daisy got the correct result twice.

If Daisy does not actually have ESP and is merely guessing the number, what is the probability that he's make a correct guess on two of the three trials?

Example:

Classify each of the following as binomial or not binomial:

1. Roll a die 4 times and count the sum of the "up" faces.

2. Roll a die 3 times and count the number of times the "up" face is two or less.

3. Draw two cards from a standard deck of playing cards (one at a time) and count the number of jacks.

4. Flip a fair quarter twice and count the number of heads.

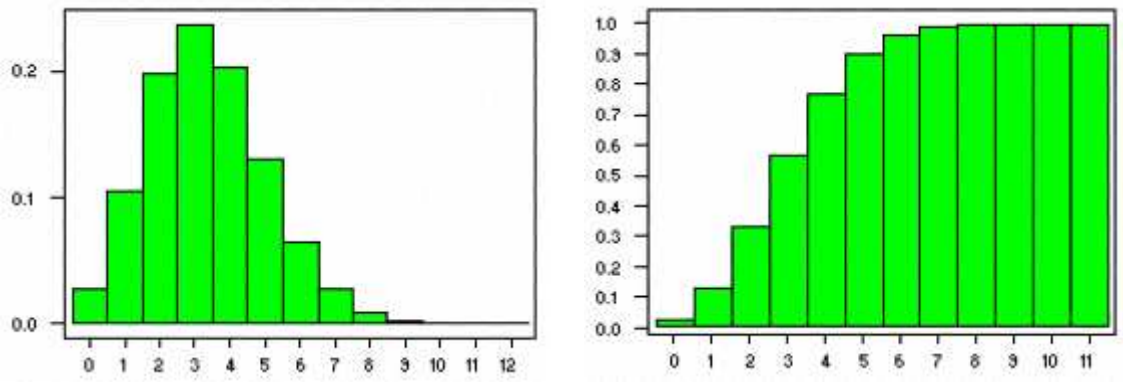
To find probabilities from a binomial distribution, one may either calculate them directly, use a binomial table(Table C on pages T-6 to T-10), or use a computer.

Example:
The number of 6s rolled by a single die in 20 rolls has a $B(20, 1/6)$ distribution. The probability of rolling more than 2 6s in 20 rolls, $P(X > 2)$, is equal to $1 - P(X < 2) = 1 - (P(X = 0) + P(X = 1) + P(X = 2))$.

```
Binomial with n = 20 and p = 0.166667

        x       P( X = x)        P( X <= x)
        0         0.0261             0.0261
        1         0.1043             0.1304
        2         0.1983             0.3287
        3         0.2378             0.5665
        4         0.2022             0.7687
        5         0.1295             0.8982
        6         0.0644             0.9629
        7         0.0258             0.9887
        8         0.0085             0.9972
        9         0.0088             0.9994
```

The corresponding graphs for the probability density function and cumulative distribution function for the $B(20, 1/6)$ distribution are shown below:



Since the probability of 2 or fewer sixes is equal to 0.3287, the probability of rolling more than 2 sixes = 1 - 0.3287 = 0.6713

Denote the probability of success on a trial by p. For n independent trials, the probability of k successes equals:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, for\ k = 0, 1, 2, ...n \tag{1}$$

Example:

Go back to the last example of counting the number of 6s in 20 rolls of a single die. Here X has $B(20, 1/6)$ distribution. Let's find the probability of rolling 2 6s in 20 rolls using the formula:

First, there are $\binom{20}{2} = \frac{20!}{2!(20-2)!} = 190$ ways to get 2 6s in 20 rolls

Second, the probability of getting 2 6s on any 2 rolls you make is $(\frac{1}{6})^2 = .02778$

Third, the probability of getting the other 18 rolls to be anything but a 6 is $(\frac{5}{6})^{18} = .03756$

Last, multiply all these together. Is this the same probability as in the table from the previous example?

Example:

Suppose that is known that the probability of an A&M student being Republican is 70% and I select a random sample of 12 students

1. Find the probability that there are exactly 8 Republicans.
2. Find the probability that there are at least 10 Republicans.

3. Find the probability that there are more than 9 Republicans.

4. Find the probability that there are no more than (at most) 7 Republicans.

5. Calculate $P(R < 4)$.

6. Calculate $P(8 \leq R \leq 10)$.

Mean and variance of a binomial random variable

The binomial probability distribution for n trials with a probability p of success on each trial has the following mean μ and standard deviation σ:

$$\mu_X = np \tag{2}$$

$$\sigma_X = \sqrt{np(1-p)} \tag{3}$$

The formula for the mean makes sense. If the probability of success is p for a given trial, then we expect about a proportion p of the n trials to be successes, or about np total. If we sample n=12 students from the entire A&M population in which 70% are Republican, then we expect that about np=12(.70)=8.4 (so about 8 or 9) in the sample to be Republican.

****The binomial distribution can be approximated by a normal distribution N(np, $np(1-p)$) if n is large enough than the expected number of successes, np, and the expected number of failures, $n(1-p)$, are both at least 15.****

Example:

In the 1990s, the US Justice Department and other groups studied possible abuse by Philadelphia police officers in their treatment of minorities. One study, conducted by the ACLU, analyzed whether African-American drivers were more likely than others in the population to be targeted by police for traffic stops. They studied the results of 262 police car stops during one week in 1997. Of those 207 of the drivers were African American, or 79% of the total. At that time, Philadelphia's population was 42.2% African American. Does the number of African Americans stopped suggest possible bias, being higher than we would expect if we take into account ordinary random variation?

Treat the 262 police car stop as trials so n=262.

We will suppose that the percentage of drivers in Philadelphia during that week who were African American was the same as their representation in the population (42.2%). Then, with no bias, the chance for any given police car stop where the driver is African American is p=.422 (other things being equal, such as the rate of violating traffic laws, ...).

We must also suppose that successive police car stops are independent. (they would not be, for example, if once a car was stopped, the police followed that car and stopped it repeatedly).

Under these assumptions, for the 262 police car stops, the number of African Americans stopped has a binomial distribution with n=262 and p=.422.

$\mu_X = np = 262(.422) = 110.564 \approx 111$
$\sigma_X = \sqrt{np(1-p)} = \sqrt{262(.422)(.578)} = \sqrt{63.906} \approx \sqrt{64} = 8$

We know that the binomial distribution can be approximated by the normal distribution under certain assumptions ($np > 15$ *and* $np(1-p) > 15$). So, because the normal distribution is a symmetric distribution, the probability within 3 standard deviations from the mean is about 99.7%. This interval is between:

$\mu_X \pm 3\sigma_X = 111 \pm 3(8) = (87, 135)$

If no racial profiling is happening, we would not be surprised if between 87 and 135 of the 262 people stopped were African Americans. However, the actual number stopped (207) is well above these values. This suggests that the number of African Americans stopped by be too high, even taking into account random variation.

Proportion as a statistic

There are many statistics that we can compute from the data. We can also look at count data, X, as the sample proportion, $\hat{p} = \frac{X}{n}$ that represent the proportion of successes that occured out of n trials.

Example: Suppose we wish to find out the true proportion of voting Americans who will support one of two political candidates. We will select a sample of 974 voting Americans and find that 499 support candidate A. In this case the statistic, , is the count X=499 of voters who support candidate A divided by the total number of individuals in the group n=974. This provides an estimate of the parameter p, the proportion of individuals who support the candidate in the entire population. With repeated sampling we can also predict the number (or count = X) of people who will vote for Candidate A.

Mean and variance of sample proportions

If we know that the count X of "successes" in a group of n observations with success probability p has a binomial distribution with mean np and variance $np(1-p)$, then we are able to get information about the distribution of the sample proportion, the count of successes X divided by the number of observations n.

$$\mu_p = p \tag{4}$$

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} \tag{5}$$

NOTE:This formula indicates that as the size of the sample increases, the variance decreases.

Example: You are rolling a 6-sided die 20 times. We know that the probability p of rolling a 6 on any roll is 1/6. (Note: this probability is the average probability we would get in an infinite number of rolls). What proportion of 6s should we expect to get after 20 rolls?