CHAPTER 10

Binomial Distributions, Geometric Distributions, and Sampling Distributions

IN THIS CHAPTER

Summary: In this chapter we finish laying the mathematical (probability) basis for inference by considering the binomial and geometric situations that occur often enough to warrant our study. In the last part of this chapter, we begin our study of inference by introducing the idea of a sampling distribution, one of the most important concepts in statistics. Once we've mastered this material, we will be ready to plunge into a study of formal inference (Chapters 11–14).

KEY IDEA

Key Ideas

- Binomial Distributions
- Normal Approximation to the Binomial
- Geometric Distributions
- Sampling Distributions
- Central Limit Theorem

Binomial Distributions

A binomial experiment has the following properties:

- The experiment consists of a fixed number, n, of identical trials.
- There are only two possible outcomes (that's the "bi" in "binomial"): success (S) or failure (F).

- The probability of success, p, is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **binomial random variable *X*,** which is the count of successes in n trials. The probability distribution of X is the **binomial distribution**.

(Taken together, the second, third, and fourth bullets above are called *Bernoulli trials*. One way to think of a binomial setting is as a fixed number n of Bernoulli trials in which our random variable of interest is the count of successes X in the n trials. You do not need to know the term Bernoulli trials for the AP exam.)

The short version of this is to say that a *binomial experiment* consists of n independent trials of an experiment that has two possible outcomes (success or failure), each trial having the same probability of success (p). The *binomial random variable* X is the count of successes.

In practice, we may consider a situation to be binomial when, in fact, the independence condition is not quite satisfied. This occurs when the probability of occurrence of a given trial is affected only slightly by prior trials. For example, suppose that the probability of a defect in a manufacturing process is 0.0005. That is, there is, on average, only 1 defect in 2000 items. Suppose we check a sample of 10,000 items for defects. When we check the first item, the proportion of defects remaining changes slightly for the remaining 9,999 items in the sample. We would expect 5 out of 10,000 (0.0005) to be defective. But if the first one we look at is *not* defective, the probability of the next one being defective has changed to 5/9999 or 0.0005005. It's a small change but it means that the trials are not, strictly speaking, independent. A common rule of thumb is that we will consider a situation to be binomial if the population size is at least 10 times the sample size.

Symbolically, for the *binomial random variable* X, we say X has $B(n, p)$.

example: Suppose Dolores is a 65% free throw shooter. If we assume that that repeated shots are independent, we could ask, "What is the probability that Dolores makes exactly 7 of her next 10 free throws?" If X is the binomial random variable that gives us the count of successes for this experiment, then we say that X has $B(10,0.65)$. Our question is then: $P(X = 7) = ?$.

We can think of $B(n,p,x)$ as a particular binomial probability. In this example, then, $B(10,0.65,7)$ is the probability that there are exactly 7 successes in 10 repetitions of a binomial experiment where $p = 0.65$. This is handy because it is the same syntax used by the TI-83/84 calculator `(binompdf(n,p,x))` when doing binomial problems.

If X has $B(n, p)$, then X can take on the values $0,1,2, \ldots, n$. Then,

$$B(n,p,x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

gives the *binomial probability* of exactly x successes for a binomial random variable X that has $B(n, p)$.

Now,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

On the TI-83/84,

$$\binom{n}{x} = {}_{\mathrm{n}}\mathrm{C}_{\mathrm{r}},$$

and this is found in the `MATH PRB` menu. $n!$ ("n factorial") means $n(n-1)(n-2)\ldots(2)(1,)$, and the factorial symbol can be found in the `MATH PRB` menu.

example: Find $B(15,.3,5)$. That is, find $P(X=5)$ for a 15 trials of a binomial random variable X that succeeds with probability 3.

solution:

$$P(X=5)=\binom{15}{5}(0.3)^5(1-0.3)^{15-5}$$
$$=\frac{15!}{5!10!}(0.3)^5(0.7)^{10}=.206.$$

(On the TI-83/84, $\binom{n}{r}$ = $_n C_r$`r` can be found

in the `MATH PRB` menu. To get $\binom{15}{5}$, enter `15`$_n$C$_r$`5.`)

Calculator Tip: On the TI-83/84, the solution to the previous example is given by `binompdf(15,0.3,5)`. The `binompdf` function is found in the `DISTR` menu of the calculator. The syntax for this function is `binompdf(n, p, x)`. The function `binomcdf(n, p, x)` = $P(X=0)+P(X=1)+\ldots P(X=x)$. That is, it adds up the binomial probabilities from $n=0$ through $n=x$. You must remember the "npx" order—it's not optional. Try a mnemonic like "never play xylophone."

example: Consider once again our free-throw shooter (Dolores) from an earlier example. Dolores is a 65% free-throw shooter and each shot is independent. If X is the count of free throws made by Dolores, then X has $B(10, 0.65)$ if she shoots 10 free throws. What is $P(X=7)$?

solution:

$$P(X=7)=\binom{10}{7}(0.65)^7(0.35)^3=\frac{10!}{7!3!}(0.65)^7(0.65)^3$$
$$=\texttt{binompdf(10,0.65,7)}=0.252.$$

example: What is the probability that Dolores makes *no more than* 5 free throws? That is, what is $P(X\leq 5)$?

solution:

$$P(X\leq 5)=P(X=0)+P(X=1)+P(X=2)+P(X=3)$$
$$+P(X=4)+P(X=5)=\binom{10}{0}(0.65)^0(0.35)^{10}+\binom{10}{1}(0.65)^1(0.35)^9$$
$$+\cdots+\binom{10}{5}(0.65)^5(0.35)^5=0.249.$$

There is about a 25% chance that she will make 5 or fewer free throws. The solution to this problem using the calculator is given by `binomcdf(10,0.65,5)`.

example: What is the probability that Dolores makes at least 6 free throws?

solution: $P(X \geq 6) = P(X = 6) + P(X = 7) + \ldots + P(X = 10)$
$= 1-$`binomcdf(10,0.65,5)`$=0.751$.

(Note that $P(X > 6) = 1 -$ `binomcdf(10,0.65,6)`).

The **mean and standard deviation of a binomial random variable** X are given by $\mu_X = np$; $\sigma_X = \sqrt{np(1-p)}$. A binomial distribution for a given n and p (meaning you have all possible values of x along with their corresponding probabilities) is an example of a *probability distribution* as defined in Chapter 7. The mean and standard deviation of a binomial random variable X could be found by using the formulas from Chapter 7,

$$\left(\mu_x = \sum_{i=1}^{n} x_i p_i \text{ and } \sigma_x = \sqrt{\sum_{i=1}^{n} (x-\mu_x)^2 p_i}\right),$$

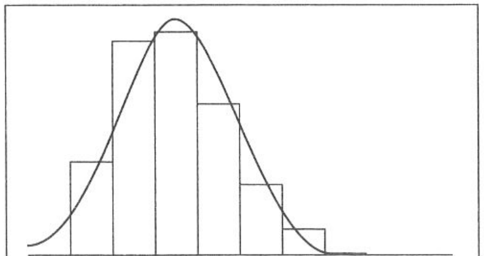
but clearly the formulas for the binomial are easier to use. Be careful that you don't try to use the formulas for the mean and standard deviation of a binomial random variable for a discrete random variable that is *not* binomial.

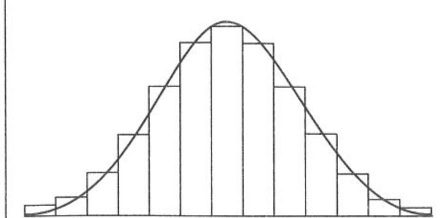
example: Find the mean and standard deviation of a binomial random variable X that has $B(85, 0.6)$.

solution: $\mu_X = (85)(0.6) = 51; \sigma_X = \sqrt{85(0.6)(0.4)} = 4.52$.

Normal Approximation to the Binomial

Under the proper conditions, the shape of a binomial distribution is approximately normal, and binomial probabilities can be estimated using normal probabilities. Generally, this is true when $np \geq 10$ and $n(1-p) \geq 10$ (some books use $np \geq 5$ and $n(1-p) \geq 5$; that's OK). These conditions are not satisfied in Graph A (X has $B(20, 0.1)$) below, but they are satisfied in Graph B (X has $B(20, 0.5)$)


Graph A: B (20, 0.1)


Graph B: B (20, 0.5)

Another way to say this is: If X has $B(n, p)$, then X has approximately $N(np, \sqrt{np(1-p)})$, provided that $np \geq 10$ and $n(1 - p) \geq 10$ (or $np \geq 5$ and $n(1 - p) \geq 5$).

example: Nationally, 15% of community college students live more than 6 miles from campus. Data from a simple random sample of 400 students at one community college are analyzed.

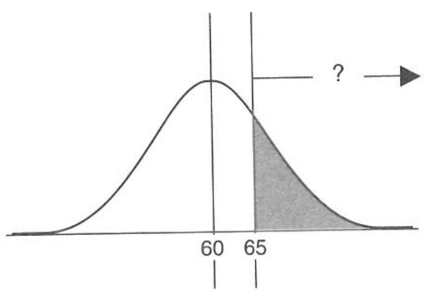
(a) What are the mean and standard deviation for the number of students in the sample who live more than 6 miles from campus?

(b) Use a normal approximation to calculate the probability that at least 65 of the students in the sample live more than 6 miles from campus.

solution: If X is the number of students who live more than 6 miles from campus, then X has $B(400, 0.15)$.

(a) $\mu = 400(0.15) = 60$; $\sigma = \sqrt{400(0.15)(0.85)} = 7.14$.

(b) Because $400(0.15) = 60$ and $400(0.85) = 340$, we can use the normal approximation to the binomial with mean 60 and standard deviation 7.14. The situation is pictured below:



Using Table A, we have $P(X \geq 65) = P\left(z \geq \frac{65-60}{7.14} = 0.70\right) = 1 - 0.7580 = 0.242$.

By calculator, this can be found as `normalcdf(65,1000,60,7.14) = 0.242`.

The exact binomial solution to this problem is given by

`1-binomcdf(400,0.15,64) = 0.261` (you use $x = 64$ since $P(X \geq 65) = 1 - P(X \leq 64)$).

In reality, you will need to use a normal approximation to the binomial only in limited circumstances. In the example above, the answer can be arrived at quite easily using the exact binomial capabilities of your calculator. The only time you might want to use a normal approximation is if the size of the binomial exceeds the capacity of your calculator (for example, enter `binomcdf(50000000,0.7,3250000)`. You'll most likely see `ERR:DOMAIN`, which means you have exceeded the capacity of your calculator, and you didn't have access to a computer. The real concept you need to understand the normal approximation to a binomial is that another way of looking at binomial data is in terms of the *proportion* of successes rather than the count of successes. We *will* approximate a distribution of sample proportions with a normal distribution and the concepts and conditions for it are the same.

Geometric Distributions

In Section 8.1, we defined a binomial setting as a experiment in which the following conditions are present:

- The experiment consists of a fixed number, n, of identical trials.
- There are only two possible outcomes: success (S) or failure (F).
- The probability of success, p, is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **binomial random variable** X, which is the count of successes in n trials. The probability distribution of X is the **binomial distribution**.

There are times we are interested not in the count of successes out of n fixed trials, but in the probability that the first success occurs on a given trial, or in the average number of trials until the first success. A **geometric setting** is defined as follows.

- There are only two possible outcomes: success (S) or failure (F).
- The probability of success, p, is the same for each trial.
- The trials are independent (that is, knowledge of the outcomes of earlier trials does not affect the probability of success of the next trial).
- Our interest is in a **geometric random variable** X, which is the number of trials necessary to obtain the first success.

Note that if X is a *binomial*, then X can take on the values 0, 1, 2, ..., n. If X is *geometric*, then it takes on the values 1, 2, 3,.... There can be zero successes in a binomial, but the earliest a first success can come in a geometric setting is on the first trial.

If X is geometric, the probability that the first success occurs on the *nth trial* is given by $P(X = n) = p(1 - p)^{n-1}$. The value of $P(X = n)$ in a geometric setting can be found on the TI-83/84 calculator, in the `DISTR` menu, as `geometpdf(p, n)` (note that the order of p and n are, for reasons known only to the good folks at TI, reversed from the binomial). Given the relative simplicity of the formula for $P(X = n)$ for a geometric setting, it's probably just as easy to calculate the expression directly. There is also a `geometcdf` function that behaves analogously to the `binomcdf` function, but is not much needed in this course.

example: Remember Dolores, the basketball player whose free-throw shooting percentage was 0.65? What is the probability that the first free throw she manages to hit is on her fourth attempt?

solution: $P(X = 4) = (0.65)\ (1-0.65)^{4-1} = (0.65)\ (0.35)^3 = 0.028$. This can be done on the TI-83/84 as follows: `geometpdf(p,n)` = `geometpdf(0.65,4)` = 0.028.

example: In a standard deck of 52 cards, there are 12 face cards. So the probability of drawing a face card from a full deck is $12/52 = 0.231$.

(a) If you draw cards with replacement (that is, you replace the card in the deck before drawing the next card), what is the probability that the first face card you draw is the 10th card?

(b) If you draw cards without replacement, what is the probability that the first face card you draw is the 10th card?

solution:

(a) $P(X = 10) = (0.231)\ (1 - 0.231)^9 = 0.022$. On the TI-83/84: `geometpdf(0.231,10)=0.0217)`.

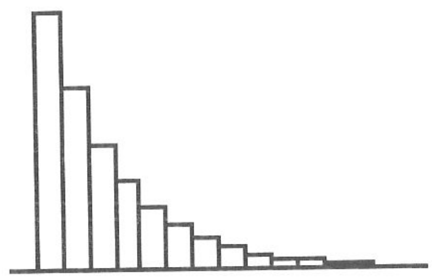
(b) If you don't replace the card each time, the probability of drawing a face card on each trial is different because the proportion of face cards in the deck changes each time a card is removed. Hence, this is not a geometric setting and cannot be answered the by techniques of this section. It can be answered, but not easily by the techniques of the previous chapter.

Rather than the probability that the first success occurs on a specified trial, we may be interested in the average wait until the first success. The average wait until the first success of a geometric random variable is $1/p$. (This can be derived by summing $(1)\cdot P(X = 1) + (2)\cdot P(X = 2) + (3)\cdot P(X = 3) + \ldots = 1p + 2p(1 - p) + 3p(1 - p)^2 + \ldots$, which can be done using algebraic techniques for summing an infinite series with a common ratio less than 1.)

example: On average, how many free throws will Dolores have to take before she makes one (remember, $p = 0.65$)?

solution: $1/0.65 = 1.54$.

Since, in a geometric distribution, $P(X = n) = p(1-p)^{n-1}$ the probabilities become less likely as n increases since we are multiplying by $1 - p$, a number less than one. The geometric distribution has a step-ladder approach that looks like this:



Sampling Distributions

Suppose we drew a sample of size 10 from a normal population with unknown mean and standard deviation and got $\bar{x} = 18.87$. Two questions arise: (1) what does this sample tell us about the population from which the sample was drawn, and (2) what would happen if we drew more samples?

Suppose we drew 5 more samples of size 10 from this population and got $\bar{x} = 20.35$, $\bar{x} = 20.04$, $\bar{x} = 19.20$, $\bar{x} = 19.02$, and $\bar{x} = 20.35$. In answer to question (1), we might believe that the population from which these samples was drawn had a mean around 20 because these averages tend to group there (in fact, the six samples were drawn from a normal population whose mean is 20 and whose standard deviation is 4). The mean of the 6 samples is 19.64, which supports our feeling that the mean of the original population might have been 20.

The standard deviation of the 6 samples is 0.68 and you might not have any intuitive sense about how that relates to the population standard deviation, although you might suspect that the standard deviation of the samples should be less than the standard deviation of the population because the chance of an extreme value for an average should be less than that for an individual term (it just doesn't seem very likely that we would draw a *lot* of extreme values in a single sample).