Chapter 17
Probability Models

What's It About?

In this chapter we examine two probability models based on Bernoulli trials: the Geometric and the Binomial. We ask students to check assumptions and conditions before using these models, and we also see what conditions allow the use of a Normal model to estimate a Binomial probability. TI Tips teach students to find Geometric and Binomial probabilities with the calculator's distribution functions.

Comments

We encountered our first special probability model, the Normal, way back in Chapter 6. In the last chapter we constructed probability models for many discrete random variables, and saw how to find means and standard deviations. Now we see two more special probability models. Students can find it difficult to distinguish between situations that are Geometric and those that are Binomial. Some of the confusion stems from the fact that all involve Bernoulli trials, so going through the checklist of those criteria fails to make the distinction. The difference lies in the definition of the random variable – are we counting the number of trials before the first success or the number of successes in a fixed number of trials, or are we working with a small probability of success and a large number of trials?

Looking Ahead

The big issue here is the Binomial model. It leads us directly to the sampling model for sample proportions in Chapter 18, the basis for inference about proportions. We want students to be able to recognize circumstances that call for a Binomial model and to be able to find the appropriate probabilities, but even more important are the assumptions and conditions that permit use of the model and the Normal approximation. They will show up repeatedly throughout the rest of the course.

Class Do's

Talk about Bernoulli trials first. Get students to recognize what characterizes Bernoulli trials: two outcomes, fixed probability of success, independence from trial to trial. They should see that such situations are common. Coins, dice, foul shots in basketball, guessing at true-false or multiple choice questions, and even sampling voter opinions seem to have the right characteristics – or are at least close enough to make the models useful.

Once students know what Bernoulli trials are, then we can define a random variable. It is the definition of that variable that determines whether a Geometric or Binomial model will prove useful. The Geometric model applies to waiting time situations—that is, counting the number of trials to achieve our first success. If instead we want to count the number of successes we achieve in a given number of trials, then we want a Binomial model. Students will make that distinction more readily when they see that it depends on what we choose to count as we conduct our Bernoulli trials.

Your decision on which (if any) of the Math Box derivations you choose to show students will depend on the interest and background of your class.

- The expected value of the Geometric model requires that students understand infinite series. Instead you can simply make the plausible argument that when rolling a die we'd expect to see

our first 4 after an average of 6 rolls, since the probability is 1/6. (The formula for the standard deviation of a Geometric model is rarely used.)

- It's similarly obvious that we'd expect to get about 100 4's in 600 rolls of a die, so the mean of the Binomial model is *np*.
- The standard deviation of the Binomial model is not obvious. The derivation is convincing, not very difficult, and uses the last chapter's revelation that we add the variances of independent random variables. This is a great opportunity to apply that important concept.

Insist that students check assumptions and conditions before using probability models. Continuing to emphasize that habit here will make the inference battles go much more smoothly. Note again the distinction we draw between assumptions and conditions. Assumptions are theoretical mathematical requirements (independence, large sample, etc). Conditions are practical guidelines that confirm (or sometimes override) assumptions.

- Sometimes there is no condition to test. For example, the question of whether the trials are independent when we toss a coin repeatedly is left to us to think about. We can't think of any reason why not, so we simply assume that they are.
- We can confirm the Large Sample Assumption with the Success/Failure Condition (*np* and *nq* at least 10).
- The Independence Assumption is violated whenever we sample without replacement, but is overridden by the 10% Condition (as long as we don't drain off more than 10% of the population, the probabilities don't change enough to matter).

For a more detailed discussion, read the article "Is That An Assumption Or A Condition" at AP Central's Statistics page.

Note that on the AP Exam students are required to <u>check</u>, not just state the conditions. In other words, it is not sufficient to simply write down $np \geq 10$ and put a little check mark there. Students must show that they checked this condition using the values given in the question, for example $np = 220(0.75) = 165 \geq 10$.

As we noted at the beginning, this probability unit is difficult for students. At the end of this chapter, they may need more work to pull it all together for a unit test. They need to develop the ability to read a question and determine which of the many approaches they have learned should be applied. The Quick Review and the Review Exercises provide lots of help. Since you won't have time to assign all of the exercises, they also provide plenty of extra examples for in-class practice. As always, having students work in groups discussing their approaches, solutions, and interpretations can be very helpful for those who are still trying to get a handle on everything.

The Importance of What You Don't Say

Don't get lost in combinatorics. There's no need to delve deeply into factorials, permutations, and combinations. If students understand the issue, that they need to know how many different ways they can get 5 successes in 7 trials, know how to write out the formula, and how to evaluate it on the calculator, that's enough. This is not a formal probability course. What's important is recognizing when and how to apply the Binomial model.

Don't worry about the continuity correction. While it is technically true that some errors in the Normal approximation to the Binomial model come from the fact that the Normal is continuous and the Binomial discrete, these are generally negligible. Nothing we do later in this course will require a level of accuracy that makes the use of the continuity correction worth the time and effort needed to teach it. Avoid this source of confusion.

Class Examples

1. Define Bernoulli trials and look at many examples and counterexamples.

2. A new sales gimmick has 30% of the M&M's covered with speckles. These "groovy" candies are mixed randomly with the normal candies as they are put into the bags for distribution and sale. You buy a bag and remove candies one at a time looking for the speckles.

- Note that this situation involves Bernoulli trials. There are two outcomes: success = speckles, failure = ordinary. The probability of success, based on the information from the candy company, is 30%. Trials can be assumed independent – there is no reason to believe that finding a speckled candy reveals anything about whether the next one out of the bag will have speckles.

- What's the probability that the first speckled one we see is the fourth candy we get? Note that the skills to answer this question come from the very first day of the probability unit. The only new issue here is formally naming the model that describes waiting time probabilities: Geometric. Remind students that they already know one useful probability model – the Normal. Now they have a second. The Normal model had two parameters: mean and standard deviation. The Geometric model has but one: probability of success.

Answer: $P(\text{First speckled is fourth candy}) = (0.7)^3(0.3) = 0.1029$

- Find other Geometric probabilities. What's the probability that the first speckled one is the tenth one? Write a general formula. What's the probability the first speckled one is one of the first three we look at? How many do we expect to have to check, on average, to find a speckled one?

Answers:

What's the probability that the first speckled one we see is the tenth candy we get?
$P(\text{First speckled is tenth candy}) = (0.7)^{10}(0.3) \approx 0.0085$

General Formula: $P(X = x) = q^{x-1}p$

What's the probability that the first speckled candy is one of the first three we look at?
$P(\text{First speckled is among first three}) = (0.3) + (0.7)(0.3) + (0.7)^2(0.3) \approx 0.657$

How many do we expect to check, on average to find a speckled one?

$\mu = \dfrac{1}{p} = \dfrac{1}{0.3} = 3\frac{1}{3}$ candies.

3. Return again to the speckled M&M's, but change the question.

- What's the probability that we'll find two speckled ones in a handful of five candies? Students will quickly come up with $(0.3)^2(0.7)^3$. You can use a tree diagram (you won't have to draw the whole thing!) to convince them that that's the correct probability for every branch that ends up with 2 of 5 speckled M&M's. They need to count the number of orders in which it can happen. Let them make a list to see that there are 10. Explain that mathematicians call such counts "combinations" and show the notation. You can also show them the $_nC_r$ operation on the TI's MATH PRB menu.

Answer: $P(\text{Two speckled candies in first five}) = \binom{5}{2}(0.3)^2(0.7)^3 = 0.3087$

- Calculate a variety of other probabilities for practice in setting them up. Students need to show the set-up on the AP Exam.

- Discuss this new model, the Binomial; it has two parameters, the number of trials and the probability of success. Students need to see that the underlying situation is the same – Bernoulli trials – whether the question asks about waiting time (Geometric) or number of successes (Binomial).

- Note that we are actually drawing M&M's without replacement from a finite population, so the Independence Condition is violated; the probability that the next one is speckled depends on how many speckled ones we have removed from the pool so far. Obviously our sample (the small bag we bought) is such a tiny subset of the vast number of M&Ms in the world that it doesn't really matter much, and that's the essence of the 10% Condition.

4. After you develop the formulas for the mean and standard deviation of a Binomial model, refer back to our earlier question (Chapter 11) about how to decide if a coin is biased. We were going to test it by flipping it 100 times. Students initially thought that they'd need to see 75–80 heads or more to suspect bias, but simulation revealed that seeing even 60–65 was pretty unusual. Now we can see that the mean is 50 heads, with standard deviation 5, so 60-65 represents 2–3 standard deviations above the mean. We're pretty sure such results are unusual IF we can apply a Normal model. This provides an opportunity to examine how Binomial models become approximately Normal for large enough *n*'s. *ActivStats* has a simulation tool that demonstrates this. There are also some decent Internet applets if you want another visual demonstration. And such discussions about whether results seem unusual foreshadow the key idea of statistical significance, coming soon.

Resources

ActivStats

- Chapter 17—"Probability Models."

Web Links

- A search for "binomial applet" will offer you several applets to look at Binomial models and the Normal approximation.

- "Is That An Assumption Or A Condition" is posted on the College Board's AP Statistics page at http://apcentral.collegeboard.com.

Assignments

You'll want to allow about four days for this chapter and another two days to review the whole probability unit before the test. Split the reading up over the first three days, then have students read the Part IV Quick Review. You can assign four problems a night from Chapter 17 along with a couple from Chapter 16 during the first two days. The next three assignments still include four problems from Chapter 17 and add a couple from the Review Exercises. On the last night assign a collection of about six varied Review Exercises. One important feature of these Review Exercises is that they are arranged in random order; students really need to *Think* about how to approach each without any hint they might have derived from where the exercise was located.

The American Red Cross says that about 11% of the U.S. population has Type B blood. A blood drive is being held at your school.

1. How many blood donors should the American Red Cross expect to collect from until it gets a donor with Type B blood?

2. What is the probability that the tenth blood donor is the first donor with Type B blood?

3. What is the probability that exactly 2 of the first 20 blood donors have Type B blood?

4. What is the probability that at least 2 of the first 10 blood donors has Type B blood?

5. The blood drive has a total of 150 donors. Assuming this is a typical number of donors for a school blood drive, what would be the mean and standard deviation of the number of donors who have Type B blood?

6. Surprised by the low number of Type B blood donors at the blood drive, the American Red Cross wonders if the 11% estimate was too high for your area. How many Type B blood donors would it take to convince <u>you</u> that this estimate might be too high? Justify your answer.

The American Red Cross says that about 11% of the U.S. population has Type B blood. A blood drive is being held at your school.

1. How many blood donors should the American Red Cross expect to collect from until it gets a donor with Type B blood?

    This is a Geometric model with $p = 0.11$.

    Expected value: $\mu = \dfrac{1}{p} = \dfrac{1}{0.11} = 9.1$ donors

2. What is the probability that the tenth blood donor is the first donor with Type B blood?

    $P\left(9 \text{ not Type B, Type B on } 10^{th}\right) = \left(0.89\right)^{9}\left(0.11\right) = 0.0385$

3. What is the probability that exactly 2 of the first 20 blood donors have Type B blood?

    $P\left(\text{exactly 2 out of 20}\right) = P\left(X = 2\right) = \dbinom{20}{2}\left(0.11\right)^{2}\left(0.89\right)^{18} = 0.2822$

4. What is the probability that at least 2 of the first 10 blood donors has Type B blood?

    $P\left(X \geq 2\right) = 1 - P\left(X \leq 1\right) = 1 - \left[\dbinom{10}{0}\left(0.11\right)^{0}\left(0.89\right)^{10} + \dbinom{10}{1}\left(0.11\right)^{1}\left(0.89\right)^{9}\right] = 1 - 0.6972 = 0.3028$

5. The blood drive has a total of 150 donors. Assuming this is a typical number of donors for a school blood drive, what would be the mean and standard deviation of the number of donors who have Type B blood?

    Using the Binomial model,

    Mean: $\mu = np = \left(150\right)\left(0.11\right) = 16.5$

    Standard deviation: $\sigma = \sqrt{npq} = \sqrt{\left(150\right)\left(0.11\right)\left(0.89\right)} = 3.83$

6. Surprised by the low number of Type B blood donors at the blood drive, the American Red Cross wonders if the 11% estimate was too high for your area. How many Type B blood donors would it take to convince <u>you</u> that this estimate might be too high? Justify your answer.

    Since $np = 16.5$ and $nq = 133.5$, we expect at least 10 successes and 10 failures. The sample size is large enough to apply a Normal model. It would be unusual to see the number of Type B donors more than 2 standard deviations below the mean. Since the standard deviation is 3.83, it would be unusual to see fewer than 9 Type B blood donors since 16.5 – 2(3.83) = 8.84. So, I would conclude that the 11% estimate is too high for my area if there were fewer than 9 Type B blood donors. NOTE: Other students might apply different criteria.

The owner of a pet store is trying to decide whether to discontinue selling specialty clothes for pets. She suspects that only 4% of the customers buy specialty clothes for their pets and thinks that she might be able to replace the clothes with more interesting and profitable items on the shelves. Before making a final decision she decides to keep track of the total number of customers for a day, and whether they purchase specialty clothes for their pet.

1. Assuming the pet store owner is correct in thinking that only 4% of her customers purchase specialty clothes for their pets, how many customers should she expect before someone buys a garment for their pet?

2. What is the probability that she does not sell a garment until the 7th customer? Show work.

3. What is the probability that exactly 3 of the first 10 customers buy specialty clothes for their pet? Show work.

4. What is the probability that at least 3 of the first 40 customers buy specialty clothes for their pet? Show work.

5. The owner had 275 customers that day. Assuming this was a typical day for his store, what would be the mean and standard deviation of the number of customers who buy specialty clothes for their pet each day?

6. Surprised by a high number of customers who purchased specialty pet clothing that day, the owner decided that her 4% estimate must have been too low. How many clothing sales would it have taken to convince you? Justify your answer.

AP Statistics Quiz B – Chapter 17 – Key

The owner of a pet store is trying to decide whether to discontinue selling specialty clothes for pets. She suspects that only 4% of the customers buy specialty clothes for their pets and thinks that she might be able to replace the clothes with more interesting and profitable items on the shelves. Before making a final decision she decides to keep track of the total number of customers for a day, and whether they purchase specialty clothes for their pet.

1. Assuming the pet store owner is correct in thinking that only 4% of her customers purchase specialty clothes for their pets, how many customers should she expect before someone buys a garment for their pet?

    Expected value of Geometric model: $\mu = \dfrac{1}{p} = \dfrac{1}{0.04} = 25$

2. What is the probability that she does not sell a garment until the 7th customer? Show work.

    $P(\text{6 no sales, sale on 7th}) = (0.96)^6 (0.04) = 0.0313$

3. What is the probability that exactly 3 of the first 10 customers buy specialty clothes for their pet? Show work.

    $P(\text{exactly 3 out of 10}) = P(X = 3) = \binom{10}{3}(0.04)^3 (0.96)^7 = 0.00577$

4. What is the probability that at least 4 of the first 40 customers buy specialty clothes for their pet? Show work.

    $P(X \geq 3) = 1 - P(X \leq 2)$

    $= 1 - \left[\binom{10}{0}(0.04)^0 (0.96)^{10} + \binom{10}{1}(0.04)^1 (0.96)^9 + \binom{10}{2}(0.04)^2 (0.96)^8 \right] = 1 - 0.9937 = 0.063$

5. The owner had 275 customers that day. Assuming this was a typical day for his store, what would be the mean and standard deviation of the number of customers who buy specialty clothes for their pet each day?

    Using the Binomial model, mean: $\mu = np = (275)(0.04) = 11$

    Standard deviation: $\sigma = \sqrt{npq} = \sqrt{(275)(0.04)(0.96)} = 3.25$

6. Surprised by a high number of customers who purchased specialty pet clothing that day, the owner decided that her 4% estimate must have been too low. How many clothing sales would it have taken to convince <u>you</u>? Justify your answer.

    Since $np = 11$ and $nq = 264$, we expect at least 10 successes and at least 10 failures. The sample size is large enough to apply a Normal model. It would be unusual to see the number of customers who purchased specialty pet clothing more than 2 or 3 standard deviations above the mean. Since the standard deviation = 3.25, it would be unusual to see more than 18 (11 + 2(3.25) = 17.5) customers who purchased specialty clothing. So, I would conclude that her 4% estimate must have been too low if more than 18 customers purchased specialty clothing for their pet.

The owner of a small convenience store is trying to decide whether to discontinue selling magazines. He suspects that only 5% of the customers buy a magazine and thinks that he might be able to use the display space to sell something more profitable. Before making a final decision he decides that for one day he'll keep track of the number of customers and whether or not they buy a magazine.

1. Assuming the owner is correct in thinking that 5% of the customers purchase magazines, how many customers should he expect before someone buys a magazine?

    _____

2. What is the probability that he does not sell a magazine until the 8th customer? Show work.

    _____

3. What is the probability that exactly 2 of the first 10 customers buy magazines? Show work.

    _____

4. What is the probability that at least 5 of his first 50 customers buy magazines?

    _____

5. He had 280 customers that day. Assuming this day was typical for his store, what would be the mean and standard deviation of the number of customers who buy magazines each day?

    _____
    _____

6. Surprised by a high number of customers who purchased magazines that day, the owner decided that his 5% estimate must have been too low. How many magazine sales would it have taken to convince <u>you</u>? Justify your answer.

The owner of a small convenience store is trying to decide whether to discontinue selling magazines. He suspects that only 5% of the customers buy a magazine and thinks that he might be able to use the display space to sell something more profitable. Before making a final decision he decides that for one day he'll keep track of the number of customers and whether or not they buy a magazine.

1. Assuming the owner is correct in thinking that 5% of the customers purchase magazines, how many customers should he expect before someone buys a magazine?

   Expected value of Geometric model: $\mu = \dfrac{1}{p} = \dfrac{1}{0.05} = 20$

2. What is the probability that he does not sell a magazine until the 8th customer? Show work.

   Using the Geometric model, Geom(0.05):

   $P(\text{number of customers } = 8) = (0.95)^7 (0.05) = 0.035$

3. What is the probability that exactly 2 of the first 10 customers buy magazines? Show work.

   Using the Binomial model, *Binom*(10,0.05):

   $P(2 \text{ of } 10 \text{ customers buy magazines}) = \dbinom{10}{2}(0.05)^2 (0.95)^8 = 0.075$

4. What is the probability that at least 5 of his first 50 customers buy magazines?

   $P(X > 5) = 1 - P(X \leq 4) = 1 - \left[\dbinom{50}{0}(0.05)^0(0.95)^{50} + \ldots + \dbinom{50}{4}(0.05)^4(0.95)^{46}\right] = 1 - 0.896 = 0.104$

5. He had 280 customers that day. Assuming this day was typical for his store, what would be the mean and standard deviation of the number of customers who buy magazines each day?

   $\mu = 280(0.05) = 14$

   $\sigma = \sqrt{280(0.05)(0.95)} = 3.65$

6. Surprised by a high number of customers who purchased magazines that day, the owner decided that his 5% estimate must have been too low. How many magazine sales would it have taken to convince <u>you</u>? Justify your answer.

   Since *np* = 14 and *nq* = 266, we expect at least 10 successes and at least 10 failures. The sample size is large enough to apply a Normal model. It would be unusual to see sales more than 2 (or 3) standard deviations above the anticipated mean. Since 14 + 2(3.65) = 21.3 (or 14 + 3(3.65) = 24.95), I would conclude the 5% estimate was probably too low if 22 (25) customers or more bought magazines.