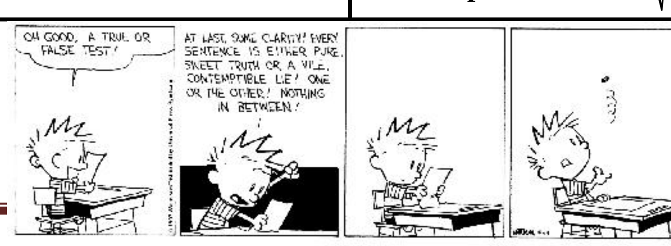
AP Statistics

Chapter 17 examines two probability models based on a random experiment called 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.

Flipping a coin is a Bernoulli trial. Rolling a die and noting whether or not it came up as a six is a Bernoulli trial. Asking a person if they like to use Internet Explorer is probably a Bernoulli trial (you might have to make some assumptions). The words "success" and "failure" are just labels—they don't necessarily mean that something good (or something bad) has happened.

Binomial Setting	**Geometric Setting**
❑ A Binomial model tells us the probability for a random variable that counts the number of successes in a **fixed number** of Bernoulli trials. Suppose the random variable ***X*** = the number of successes in ***n*** observations. Then *X* is a binomial random variable if: 1) There are only two outcomes: success or failure. 2) The probability of success, *p*, is the same for each observation. 3) The ***n*** observations are independent. 4) There is a **fixed number** of ***n*** observations. ❑ Binomial models require two parameters, ***n***, the number of trials; and ***p***, the probability of success. We denote this by **binom (n, p).** ❑ In ***n*** trials, there are $_nC_k = \frac{n!}{k!(n-k)!}$ ways to have ***k*** successes (i.e., total number of ways to arrange ***k*** out of ***n*** objects) • Read $_nC_k$ as "***n*** choose ***k***." ❑ ***Note***: ***n*! = *n* × (*n*-1) × ... × 2 × 1**, and ***n*!** is read as "***n* factorial.**"	❑ A Geometric probability model tells us the probability for a random variable that counts the number of Bernoulli trials **until the first success**. Suppose the random variable ***X*** = the number of trials required to obtain the first success. Then *X* is a geometric random variable if: 1) There are only two outcomes: success or failure. 2) The probability of success, *p*, is the same for each observation. 3) The ***n*** observations are independent. 4) The variable of interest is the number of trials required **until the first success**. Because ***n*** is not fixed, there could be an infinite number of ***X*** values. However, the probability that ***X*** is a very large number is more and more unlikely. Therefore the probability histogram for a geometric distribution is always **right skewed**. ❑ Geometric models are completely specified by one parameter, ***p***, the probability of success, and are denoted by **Geom(*p*).**

Binomial probability model for Bernoulli trials: Binom(n,*p*)	**Geometric probability model for Bernoulli trials: Geom(p)**
n = number of trials ***p*** = probability of success ***q*** = 1 – ***p*** = probability of failure ***X*** = # of successes in ***n*** trials $P(X = x) = {}_nC_x \cdot p^x \cdot q^{n-x}$ « $= np$ ± $= \sqrt{npq}$	***p*** = probability of success ***q*** = 1 – ***p*** = probability of failure ***X*** = # of trials **until the first success** occurs $P(X = x) = q^{x-1} \cdot p$ « $= \frac{1}{p}$ ± $= \sqrt{\frac{q}{p^2}} = \frac{\sqrt{q}}{p}$



AP Statistics

Probability Model for Bernoulli (Ber-noo-lee) trials:

1) There are only two possible outcomes (called success or failure) on each trial.
2) The probability of success, denoted p, is the same on every trial.
3) The trials are independent.

   NOTE: The independence assumption is violated whenever the population is finite and we sample without replacement. When we don't have an infinite population, the trials are not independent. If that assumption is violated, it is still okay to proceed as long as **the sample is smaller than 10% of the population.**

Calculator Skills:

The **probability density function** (or **pdf**) assigns a probability to **<u>each value</u>** of ***X*.**

The **cumulative density function** (or **cdf**) calculates the **<u>sum of the probabilities</u>** up to ***X***.

geometpdf(found under 2nd Dist, used to find the probability that it takes a certain number of trials to get a success. When using, you put in **geometpdf(probability, # of trials)**

geometcdf(cumulative density function, **finds the sum** of the probabilities of several possible outcomes. It calculates the probability of finding the first success *on or before* the *x*th trial....

binompdf(located under 2nd Dist and allows us to find the probability of an individual outcome. Known as **binompdf(n, p, k),** where ***n*** stands for number of trials, ***p*** stands for probability of success, and ***k*** is the desired number of successes.

binomcdf(located under 2nd Dist and allows us to find the total probability of getting ***x*** or fewer successes among ***n*** trails. Known as **binomcdf(n, p, X),** ***n*** stands for number of trials, ***p*** stands for probability, and ***X*** stands for number of successes from ***x*** number to 0.

Example 1) 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.

a) Is this situation a Bernoulli trial? Explain.

b) What's the probability that the first speckled candy is the fourth one we draw from the bag?

$$\boldsymbol{P}\left(\mathbf{1}\boldsymbol{st}\ \textbf{speckled is }\mathbf{4}\boldsymbol{th}\right) =$$
$$=$$

c) What's the probability that the first speckled candy is the tenth one?

$$\boldsymbol{P}\left(\mathbf{1}\boldsymbol{st}\ \textbf{speckled is }\mathbf{10}\boldsymbol{th}\right) =$$
$$=$$

d) Write a general formula.

$P(X=x)=q^{x-1}\bullet p$ **where:** p = probability of success

q = probability of failure

x = # of trials until 1st success

e) What's the probability we find the first speckled one among the first three we look at?

=

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

$E(X)=$

Example 2) Suppose each child born to Jay and Kay has probability 0.25 of having blood type O. If Jay and Kay have 5 children, what is the probability that exactly 2 of them have type O blood?

Let X = number of children with type O blood in the 5 children.

1) There are only two outcomes: success (has type O) or failure (not type O).
2) The probability of success, type of blood, p, is 0.25 for each of the 5 observations.
3) Each of the 5 observations is independent, since one child's blood type will not influence the next child's blood type.
4) There is a fixed number of observations, n = 5.

So X is a **binomial random variable**.

The following table shows the probability distribution function, pdf, for the binomial random variable, X.

x	0	1	2	3	4	5
$P(X=x)$						

$P(X=0)=P(\text{none have blood type O})=(0.75)^5=binompdf\,(5,0.25,0)=0.2373$

$P(X=1)=P(\text{exactly 1 has blood type O})=binompdf$

$P(X=2)=P(\text{exactly 2 has blood type O})=binompdf$

$P(X=3)=P(\text{exactly 3 has blood type O})=binompdf$

$P(X=4)=P(\text{exactly 4 has blood type O})=binompdf$

$P(X=5)=P(\text{exactly 5 has blood type O})=binompdf$

The following table shows the cumulative distribution function, cdf, for the binomial random variable, X.

x	0	1	2	3	4	5
$P(X=x)$						
$F(X)$	$P(X\le 0)$	$P(X\le 1)$	$P(X\le 2)$	$P(X\le 3)$	$P(X\le 4)$	$P(X\le 5)$

Binomcdf (5, 0.25, 0) = Binomcdf (5, 0.25, 1) = Binomcdf (5, 0.25, 2) =
Binomcdf (5, 0.25, 3) = Binomcdf (5, 0.25, 4) = Binomcdf (5, 0.25, 5) =

Example 3) Suppose that 20% of the light bulbs in a strand of Christmas tree lights are defective.

a) What is the probability of getting the first defective light bulb on the third try?

$\boldsymbol{P}(\boldsymbol{X}=\mathbf{3})=$

b) What's the probability that it will take at least 5 trials to get the first defective bulb?

$\boldsymbol{P}(\boldsymbol{X}\geq\mathbf{5})=$

c) What's the probability that it will take at most 3 tries to get the first defective light bulb?

$\boldsymbol{P}(\boldsymbol{X}\leq\mathbf{3})=$

d) What's the probability that there will be no defective light bulbs among the first five light bulbs selected?

$\boldsymbol{P}(\boldsymbol{X}\geq\mathbf{6})=$

Example (p. 390): Spam and the Geometric Model

Postini is a global company specializing in communication security. The company monitors over 1 billion Internet messages per day and recently reported that 91% of e-mails are spam!

Let's assume that your e-mail is typical – 91% spam. We'll also assume you aren't using a spam filter, so every message gets dumped in your inbox. Since spam comes in many different sources, we'll consider tour messages to be independent.

Overnight your inbox collects e-mail. When you first check your e-mail in the morning, about how many spam e-mails should you expect to have to wade through and discard before you find a real message? What's the probability that the 4^{th} message in your inbox is the first one that isn't spam?

There are two outcomes: a real message (success) and spam (failure)
The probability of success, *p* = 1 – 0.91 = 0.09
Assume that the messages arrive independently and in a random order, so I can use the model Geom(0.09).

Let *X* = number of e-mails

$\boldsymbol{E}(\boldsymbol{X})=$

$\boldsymbol{P}(\boldsymbol{X}=\mathbf{4})=\boldsymbol{P}(\textbf{spam}\cap\textbf{spam}\cap\textbf{spam}\cap\textbf{real})$

$=$

On average, I expect to check just over e-mails before I find a real message. There's slightly less that a % chance that my first real message will be the 4^{th} one I check.

Example (p. 395): Spam and the Binomial Model

Using the information from the previous example, suppose your inbox contains 25 messages. What are the mean and standard deviation of the number of real messages you should expect to find in your inbox? What is the probability that you will find only 1 or 2 real messages?

Assume that messages arrive independently and at random, with the probability of success (real message) p = 1 – 0.91 = 0.09. Let X = number of real messages among 25, so I can use the model Binom(25, 0.09).

$E(X) = n$ $SD(X) =$

$P(X = 1 \text{ or } 2) =$

Among 25 e-mail messages, I expect to find an average of that aren't spam, with a standard deviation of messages. There's just % chance that 1 or 2 e-mails will be real messages.

The Normal Model

- When dealing with a large number of trials in a Binomial situation, making direct calculations of the probabilities becomes tedious (or outright impossible).
- Fortunately, the Normal model comes to the rescue…
- As long as the Success/Failure Condition holds, we can use the Normal model to approximate Binomial probabilities.
- **Success/failure condition:** A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures: $np \geq 10$ and $nq \geq 10$

Continuous Random Variables

- When we use the Normal model to approximate the Binomial model, we are using a continuous random variable to approximate a discrete random variable.
- So, when we use the Normal model, we no longer calculate the probability that the random variable equals a *particular* value, but only that it lies *between* two values.

Example (p. 398): Spam and the Normal approximation to the Binomial
From the *Postini* example, we know that 91% of e-mail messages are spam. Recently you installed a spam filter. You observe that over the past week it okayed only 151 of 1422 e-mails you received, classifying the rest as junk. Should you worry that the filtering is too aggressive? What is the probability that no more than 151 of 1422 e-mails is a real message?

Assume that messages arrive randomly and independently with the probability of success (real message) p = 1 – 0.91 = 0.09. The model Binom(25, 0.09) applies but will be hard to work with. Checking conditions for the Normal approximation:

- **These messages represent less than 10% of all e-mail traffic.**
- **np = (1422)(0.09) = 127.98 real messages and nq = (1422)(0.91) = 1294.08 spam messages, both far greater than 10.**
- **It is okay to approximate this binomial probability by using a Normal model.**

$u =$ $\pm =$

$P(X \leq 151) =$

Among my 1,422 e-mails, there's % chance that no more than 151 of them were real messages, so the filter may be working properly.

JUST CHECKING (p. 399)

From a previous chapter, we noted that the Pew Research Center reported that they are able to contact only 76% of the randomly selected households drawn for a telephone survey.

1) Explain why these phone calls can be considered Bernoulli trials.

2) Which of the models in this chapter (Geometric, Binomial, Normal) would you use to model the number of successful contacts from a list of 1000 sampled households? Explain.

   u = **and** ± =

3) Pew further reports that even after they have contacted a household, only 38% agree to be interviewed, so the probability of getting a completed interview for a randomly selected household is only 0.29. Which of the models of this chapter would you use to model the number of households Pew has to call before they get the first completed interview?