AP Statistics
Unit 05 – Distributions
Day 01 Notes – Binomial & Geometric Random Variables

Name Key
Period

EXAMPLE: Refresher on Discrete Random Variables

A large auto dealership keeps track of sales made during each hour of the day. Let X = the number of cars sold during the first hour of business on a randomly selected Friday. Based on previous records, the probability distribution is as follows:

Cars Sold	0	1	2	3
Probability	0.3	0.4	0.2	0.1

a. Compute and interpret the mean of X.

$E(X) = 0\cdot 0.3 + 1\cdot 0.4 + 2\cdot 0.2 + 3\cdot 0.1 = \boxed{1.1 \text{ cars sold on average}}$

b. Compute and interpret the standard deviation of X.

$\sigma^2 = (0-1.1)^2(0.3) + (1-1.1)^2(0.4) + (2-1.1)^2\cdot 0.2 + (3-1.1)^2\cdot 0.1 = 0.89$

$\sqrt{\sigma^2} = \boxed{\sigma = 0.9434 \text{ cars sold}}$

Binomial Settings & Binomial Random Variables

What do the following scenarios have in common?

- Toss a coin 5 times. Count the number of heads.
- Spin a roulette wheel 8 times. Record how many times the ball lands in a red slot.
- Take a random sample of 100 babies born in US hospitals today. Count the number of females.

In each case, we're performing repeated *trials* of the same chance process. The number of trials is fixed in advance. Knowing the outcome of one trial tells us nothing about the outcome of any other trial. Our chances of "success" (getting a specific result or event) are the same on each trial. This is called a **binomial setting**.

Binomial Setting: A binomial setting arises when we perform several independent trials of the same chance process and record the number of times that a particular outcome occurs.

Four conditions:

B Binary: the possible outcomes of each trial can be classified as "success" or "failure."

I Independent: trials must be independent. Knowing the result of one trial must not tell us anything about the result of any other trial.

N Number: The number of trials n must be fixed in advance. (of the chance process

S Success: There is the same probability p of success on each trial.

*Note: There can be multiple outcomes (not just two) but we have to define a success and failure.

Binomial Random Variable: The count X of successes in a binomial setting.

Binomial Distribution: The probability distribution X with parameters n and p, where n is the # of trials of the chance process and p is the probability of success in any one trial. The possible values of X are whole numbers from 0 to n.

EXAMPLE: From Blood Types of Aces

Here are three scenarios involving chance behavior. In each case, determine whether the given random variable has a binomial distribution. Justify your answer.

1. Genetics says that children receive genes from each of their parents independently. Each child of a particular set of parents has probability 0.25 of having type O blood. Suppose these parents have 5 children. Let X = the number of children with type O blood.

B: success = has type O blood failure = does not have type O blood

I: knowing one child's blood type tells you nothing about another's because children inherit blood type genes independently from parents.

N: there are n = 5 trials

S: probability of success p = 0.25 per trial. (on each trial)

This is a binomial setting.

2. Shuffle a deck of cards. Turn over the first 10 cards, one at a time. Let Y = the number of aces you observe.

B success = get an ace. failure = don't get an ace.

I not independent. If the first card is an ace, the second card is less likely to be an ace (3/51 instead of 4/52) + vice versa.

N

S

∴ not a binomial setting.

3. Shuffle a deck of cards. Turn over the top card. Put the card back in the deck, and shuffle again. Repeat this process until you get an ace. Let W = the number of cards required.

B

I

N the # of trials is not set in advance ∴ this is not a binomial setting.

S

EXAMPLE: Inheriting Blood Type

Each child of a particular set of parents has probability 0.25 of having type O blood Genetics says that children receive genes from each of their parents independently. If these parents have 5 children, the count X of children with type O blood is a binomial random variable with n = 5 trials and probability p = 0.25 of success on each trial. In this setting, a child with type O blood is a "success" (S) and a child with another type of blood is a "failure" (F).

What is P(X = 0)? That is, what is the probability that *none* of the 5 children has type O blood? It's the chance that all 5 children *don't* have type O blood. The probability that any one of this couple's children doesn't have type O blood is 1 – 0.25 = 0.75 (complement rule). By the multiplication rule for independent events,

$$P(X = 0) = P(FFFFF0 = (0.75)(0.75)(0.75)(0.75)(0.75) = (0.75)^5 = 0.2373$$

How about P(X = 1)? There are several ways in which exactly 1 of the 5 children could have type O blood. For instance, the first child born might have type O blood, while the remaining 4 children don't have type O blood. The probability that this happens is

$$P(SFFFF) = (0.25)(0.75)(0.75)(0.75)(0.75) = (0.25)(0.75)^4$$

Alternatively, Child 2 could be the one that has type O blood. The corresponding probability is

$$P(SFFFF) = (0.75)(0.25)(0.75)(0.75)(0.75) = (0.25)(0.75)^4$$

There are three more possibilities to consider:

$$P(SFFFF) = (0.75)(0.75)(0.25)(0.75)(0.75) = (0.25)(0.75)^4$$
$$P(SFFFF) = (0.75)(0.75)(0.75)(0.25)(0.75) = (0.25)(0.75)^4$$
$$P(SFFFF) = (0.75)(0.75)(0.75)(0.75)(0.25) = (0.25)(0.75)^4$$

In all, there are five different ways in which exactly one child would have type O blood, each with the same probability of occurring. As a result,

$$\begin{aligned} P(X = 1) &= P(\text{exactly 1 child with type O blood}) \\ &= P(SFFFF) + P(FSFFF) + P(FFSFF) + P(FFFSF) + P(FFFFS) \\ &= 5(0.25)(0.75)^2 = 0.39551 \end{aligned}$$

There's about a 40% chance that exactly 1 of the couple's 5 children will have type O blood.

What if we wanted to find P(X = 2)? The general rule is that we find the number of arrangements for a particular number of successes, then multiple by the probability of that many successes.

$$P(X = k) = P(\text{exactly k successes in n trials}) = \text{number of arrangements} \times p^k(1 - p)^{n-k}$$

Counting all of the arrangements every time would be super tedious and frustrating to calculate, so here's a nifty formula that does it for you!

Binomial Coefficient:

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

For k = 0, 1, 2, …, n where

$$n! = n(n-1)(n-2)\ldots(3)(2)(1)$$

and 0! = 1.

On the calculator…

1. Type in the number of trials you will do
2. Press MATH → PRB
3. Choose nCr
4. Type in the number of successes you want and press ENTER

If you want to calculate the total probability of k successes among n observations, you would use the **Binomial Probability Formula**.

Binomial Probability Formula:

If X has the binomial distribution with n trials and probability p of success on each trial, the possible values of X are 0, 1, 2, …, n. If k is one of these values,

$$P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$$

EXAMPLE: Inheriting Blood Type

Each child of a particular set of parents has probability 0.25 of having type O blood. Suppose the parents have 5 children.

1. Find the probability that exactly 3 of the children have type O blood.

let X = the # of children with type O blood. X has a binom. dist. with n = 5 and p = 0.25

$P(X=3) = \binom{5}{3}(0.25)^3(0.75)^2 = 0.08789$ There is about a 9% chance that exactly 3 of 5 children have type O blood.

2. Should the parents be surprised if more than 3 of their children have type O blood? Justify your answer.

$P(X>3) = P(X=4) + P(X=5) = \binom{5}{4}(0.25)^4(0.75)^1 + \binom{5}{5}(0.25)^5(0.75)^0$

$= 0.01563$

Because there's only about a 1.6% chance of having more than 3 children with type O blood, the parents should definitely be surprised if this happens.

On the calculator...

There are two commands that we use to calculate binomial probabilities on the calculator:

binompdf(n,p,k) calculates $P(X = k)$ binomcdf(n,p,k) calculates $P(X \leq k)$

1. Press 2ND, VARS and choose binompdf or binomcdf
2. Enter your values in the order (n, p, k)
3. Press ENTER

Sometimes, we're looking for $P(X \geq k)$, which means we need to do a little maneuvering of our equation.

$$P(X \geq k) = 1 - P(X < k) = 1 - P(X \leq k - 1)$$

Or

$$P(X > k) = 1 - P(X \leq k)$$

Rules for finding the binomial probability:

Step 1: State the distribution and the values of interest. Specify a binomial distribution with the number of trials n, success probability p, and the values of the variable clearly identified. State/Plan

Step 2: Perform calculations – show your work! Do one of the following:

i) Use the binomial probability formula to find the desired probability
ii) Use the binompdf or binomcdf command and label each of the inputs. DO

Step 3: Answer the question. conclude

EXAMPLE: Free Lunch?

A local fast-food restaurant is running a "Draw a three, get it free" lunch promotion. After each customer orders, a touch-screen display show the message "Press here to win a free lunch." A computer program then simulates one card being drawn from a standard deck. If the chosen card is a 3, the customer's order is free. Otherwise, the customer must pay the bill.

a. All 12 players on a school's basketball team place individual orders at the restaurant. What is the probability that exactly 2 of them win free lunch?

Step 1: let X = # of players who win a free lunch. there are 12 independent trials each with a success probability of 4/52, so X has a binomial distribution with $n = 12$ and $p = 4/52$. we want to find $P(X=2)$.

Step 2: $P(X=2) = \binom{12}{2}\left(\frac{4}{52}\right)^2\left(\frac{48}{52}\right)^{10} = 0.1754$ or binompdf(trials = 12, p = 4/52, x-value = 2) = 0.1754

Step 3: There is about a 17.5% chance that exactly 2 players will win a free lunch.

b. If 250 customers place lunch orders on the first day of the promotion, what's the probability that fewer than 10 win a free lunch?

Step 1: Let Y = # of customers who win a free lunch. ∃ 250 independent trials each with a success probability of 4/52, so Y has a binomial distribution with n = 250 and p = 4/52. We want to find P(Y < 10).

Step 2: $P(Y<10) = P(Y \le 9) = \text{binomcdf(trials}=250, p=4/52, \text{xvalue}=9)$

$= 0.00613$

Step 3: There is less than a 1% chance that fewer than 10 customers will win a free lunch. If this actually happened, the customers should be suspicious about the restaurant's claim.

Mean and standard deviation of a binomial random variable:

If a count X of successes has the binomial distribution with number of trials n and probability of success p, the mean and standard deviation of X are

Mean: $\mu = np$
Standard Deviation: $\sigma = \sqrt{npq}$

Where q = 1 – p

* *These formulas only work for binomial distributions and cannot be used for other distributions.*

EXAMPLE: Tastes as good as the real thing?

The makers of a diet cola claim that its taste is indistinguishable from the taste of the full-calorie version of the same cola. To investigate, an AP Statistics student named Emily prepared small samples of each type of soda in identical cups. Then she had volunteers taste each cola in random order and try to identify which was the diet cola and which was the regular cola. Overall, 23 or the 30 subjects made the correct identification.

If we assume that the volunteers really couldn't tell the difference, then each one was guessing with a ½ chance of being correct. Let X = the number of volunteers who correctly identify the colas.

a. Explain why X is a binomial random variable.

B guesses are either correct or incorrect
I the results of one volunteer's guess tells us nothing about the results of another volunteer's guess.
N n = 30 trials
S p = 50% for every trial.

b. Find the mean and the standard deviation of X. Interpret each value in context.

$\mu_X = np = 30 \cdot 0.5 = 15$

$\sigma = \sqrt{npq} = \sqrt{30 \cdot 0.5 \cdot 0.5} = \sqrt{7.5} = 2.74$

If this experiment were repeated many times and the volunteers were randomly guessing, the average # of correct guesses would be about 15 and would typically vary from the mean by about 2.74 correct guesses.

c. Of the 30 volunteers, 23 made correct identifications. Does this give convincing evidence that the volunteers can taste the difference between the diet and regular colas?

$P(X \geq 23) = 1 - P(X \leq 22) = 1 - \text{binomcdf}(\text{trials} = 30, p = 0.5, \text{xvalue} = 22)$

$= 0.0026$

There is a very small chance that there would be 23 or more correct guesses if the volunteers couldn't tell the difference in the colas. Therefore, we have convincing evidence that the volunteers can taste the difference.

EXAMPLE: Bad Flash Drives

A supplier inspects an SRS of 10 flash drives from a shipment of 10000 flash drives. Suppose that (unknown to the supplier) 2% of the flash drives in the shipment are defective. Count the number X of bad flash drives in the sample.

This is not quite a binomial setting. Removing 1 flash drive changes the proportion of bad flash drives remaining in the shipment. The conditional probability that the second flash drive chosen is bad changes when we know whether the first is good or bad. But removing 1 flash drive from a shipment of 10000 changes the makeup of the remaning 9999 flash drives very little. The distribution of X is very close ot eh binomial distribution with n = 10 and p = 0.02. To illustrative this, let's calculate the probability that none of the 10 flash drives is defective:

$$P(X = 0) = 0.8171$$

The actual probability of getting no defective flash drives is:

$$P(\text{no defectives}) = 0.8170$$

The two probabilities are quite close!

In practice, the binomial distribution gives a good approximation as long as we don't sample more than 10% of a population. We refer to this as the 10% Condition.

10% Condition: when taking an SRS of size n from a population of size N, we can use a binomial distribution to model the count of successes in the sample as long as $n < \frac{1}{10}N$.

↑ because w/o replacement (usually) violates independence.

Normal Model Distributions

When dealing with large number of trials, making direct calculations of probabilities becomes tedious and therefore we use the "normal curve" to simplify the process.

Binomial Probability Model for Bernoulli Trials   Binom (n,p)
n= number of trials
p = probability of success (and q=1-p = probability of failures)
X= number of success in n trials

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, where \binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

Mean: $\mu = np$
Standard Deviation: $\sigma = \sqrt{npq}$

To determine if you should use the normal model distribution you need to check the Success/Failure Condition, also known as the Large Counts Condition.

Large Counts Condition:

A binomial model is approximately Normal if we expect at least 10 successes and 10 failures:

$np \geq 10 \ and \ nq \geq 10$

EXAMPLE:

The communications monitoring company Postini has reported 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? In order to answer this question, you need to find what the probability that no more than 151 of 1422 e-mails is a real message.

x

1. STATE: what do you know?
   $n = 1422$
   $p = 0.09$ OK emails
   $q = 0.91$ spam emails
2. PLAN: Check conditions
   Large Counts: $np > 10$? $nq > 10$?
   $1422 \cdot 0.09 = 127.98 > 10$ ✓ $1422 \cdot 0.91 = 1294.02 > 10$ ✓
3. DO: $\mu = np = 1422 \cdot 0.09 = 127.98$ OK emails
   $\sigma = \sqrt{npq} = \sqrt{1422 \cdot 0.09 \cdot 0.91} = \sqrt{116.46} = 10.79$ OK emails
   $P(X \leq 151)$ $z = \frac{x-\mu}{\sigma} = \frac{151 - 127.98}{10.79} = 2.13$ $P(z \leq 2.13) = 0.9834$
4. CONCLUDE:
   Among the 1422 emails, there's about a 98.34% chance that no more of them were real messages, so the filter may be working properly.

EXAMPLE:

It is generally believed that nearsightedness affects about 12% of all children. A school district tests the vision of 169 incoming kindergarten children. How many would you expect to be nearsighted? With what standard deviation?

State: $n = 169$
$p = 12\%$
$q = 88\%$

DO: $\mu = np = 169 \cdot 0.12 = 20$
$\sigma = \sqrt{npq} = 4.22$

Conclude: we expect about 20 kids to be nearsighted with a standard deviation of 4.22 kids.

EXAMPLE:

A newly hired telemarketer is told he will probably make a sale on about 12% of his phone calls. The first week he called 200 people, but only made 10 sales. Should he suspect he was misled about the true success rate? Explain.

State: $n = 200$
$p = 0.12$
$q = 0.88$

Plan: Large counts:
$np > 10$
$200 \cdot 0.12 = 24 > 10$ ✓
$nq > 10$
$200 \cdot 0.88 = 176 > 10$ ✓

DO: $\mu = np = 200(0.12) = 24$ sales
$\sigma = \sqrt{npq} = \sqrt{200 \cdot 0.12 \cdot 0.88} = 4.6$ sales
$P(X \leq 10)$ $z = \frac{x - \mu}{\sigma} = \frac{10 - 24}{4.6} = -3.046$
$P(z \leq -3.046) = 0.116\%$

Conclude: yes, because making 10 sales puts him in the 0.1%ile for sales, which is extremely unlikely.

EXAMPLE:

Police estimate that 80% of drivers now wear their seatbelts. They set up a safety roadblock, stopping cars to check for seatbelt use.

a. What's the probability that the first unbelted driver is in the 6th car stopped?

$p = 0.8$ seatbelt
$q = 0.2$ no seatbelt

$(0.8)^5 (0.2)^1 = 0.065536$

There is a 6.55% chance that the 6th car is the first unbelted driver.

b. What's the probability that the first 10 drivers are all wearing their seatbelts?

$p = 0.8$ seatbelt
$q = 0.2$ no seatbelt

$(0.8)^{10} = 0.107$

There is a 10.7% chance that the first ten drivers are all wearing their seat belts.

c. If they stop 30 cars during the first hour, find the mean and standard deviation of the number of drivers expected to be wearing seatbelts?

$p = 0.8$ seatbelt
$q = 0.2$ no seatbelt
$n = 30$

$\mu = np = 0.8 \cdot 30 = 24$
$\sigma = \sqrt{npq} = \sqrt{0.8 \cdot 0.2 \cdot 30} = 2.19$

We expect 24 of 30 drivers to be wearing seatbelts, with a standard deviation of 2.19 drivers.

d. If they stop 120 cars during the safety check, what's the probability that they find at least 20 drivers NOT wearing their seatbelts?

State: $p = 0.2$ no seatbelt
$q = 0.8$ seatbelt
$n = 120$

Plan: Large counts: $np > 10$: $0.2 \cdot 120 = 24 > 10$ ✓ $nq > 10$: $120 \cdot 0.8 = 96 > 10$ ✓

DO: $\mu = np = 0.2 \cdot 120 = 24$
$\sigma = \sqrt{npq} = \sqrt{.2 \cdot 0.8 \cdot 120} = 2.19$
$P(X \geq 20)$
$z = \frac{x - \mu}{\sigma} = \frac{20 - 24}{2.19} = -0.913$
$P(z \geq -0.913) = 81.93\%$

Conclude: there is an 81.93% chance that they'll find at least 20 drivers NOT wearing seatbelts.

Geometric Random Variables

Geometric Setting:

- Repeat a chance process ***until a success occurs***
- Trials are independent
- No set number of trials
- The probability of success remains the same for all trials

The number of trials Y that it takes to get a success in a geometric setting is a **geometric random variable.** The probability distribution of Y is a **geometric distribution** with parameter *p*, the probability of a success on any trial. The possible values of Y are 1, 2, 3, …

Geometric Probability Formula

If Y has the geometric distribution with probability *p* of success on each trial, the possible values of Y are 1, 2, 3… If *k* is any one of these values,

$P(Y = k) = (1 - p)^{k-1}p$

On the calculator:

geometpdf(p,k) computers $P(Y = k)$

* geometcdf(p,k) computes $P(Y \le k)$

EXAMPLE:

Mrs. De Marre decides she wants a pink gumball from the gumball machine at her local grocery store. The probability of getting a pink gumball is 1/5 per quarter put in the machine. She keeps putting quarters in the machine until she gets a pink gumball.

a. Find the probability that she spends exactly 3 quarters.

$p = 0.2$ (pink)
$q = 0.8$ (notpink)
$K = 3$

$P(Y=3) = (1-0.2)^{3-1} \cdot 0.2 = (0.8)^2 \cdot 0.2$

geometpdf$(p=0.2, k=3) = \Rightarrow$ $= 12.8\%$

There is a 12.8% chance that she spends exactly 3 quarter.

b. Find the probability that she spends exactly \$4.25.

17 quarters

$p = 0.2$ pink
$q = 0.8$ not pink
$K = 17$

$P(Y=17) = (0.8)^{16}(0.2) = 0.563\%$

geometpdf$(0.2, 17) = 0.563\%$

There is a 0.563% chance that she spends exactly \$4.25.

c. Find the probability that she spends no more than 7 quarters.

$p = 0.2$ pink
$q = 0.8$ not pink
$K = 7$

$P(Y \le 7) =$ geometcdf$(0.2, 7) = P(Y=1) + P(Y=2) + \dots + P(Y=7)$

$= 0.7903$

There is a 79.03% chance that she will spend no more than 7 quarters to get a pink gumball.