# L15: Binomial distributions

# Today's objectives

- Introduce the binomial distribution
    - What kinds of outcomes follow a binomial
    - Understanding the probability space for a binomial
    - What is the theoretical distribution for binomials

# Types of outcomes

The first part of probability we talked about events with binary our categorical possibilities.

We've now seen how we can use a normal probability distribution to help us evaluate continuous variables.

What about expected values(probabilities) for sets of outcomes that have only 2 possibilities?

# Example: Dice, one roll vs 10 rolls

Roll 1 die one time, what is the set of outcomes?

Roll 1 die one time, what is the probability that you roll a 6?

Roll 20 die, what is the set of possible outcomes?

What if we are interested in the probability that we roll 20 dice and we get 5 rolls where the die is 6?

What other kinds of questions follow this setup?

# The binomial setting and binomial distributions

- An elementary school administers eye exams to 800 students. How many students have perfect vision?
- A new treatment for pancreatic cancer is tried on 250 patients. How many survive for five years?
- You plant 10 dogwood trees. How many live through the winter?

# What are the common threads to each of these questions?

- Something is done *n* number of times.
- The outcome of interest for each question is categorical (binary - two levels)

# Binomial Distributions

# Binomial Distributions

# Binomial Probability Distributions and notation

- Bernoulli Random Variable: The variable must assume one of two possible mutually exclusive outcomes
- Each trial of the BRV results in either a success or failure of the event happening
- Derived from the experiment: counting the number of occurrences of an event in n independent trials
- Random Variable: X = number of times the event happens in the fixed number of trials (n)
- Parameters
  - n = number of trials
  - p = probability of success (event happening)

# Bernoulli Process

- The *n* observations are independent. Knowing the result of one observation does not change the probabilities assigned to other observations
- Each observations is either a “success” or a “failure” (usually noted with 0 or 1). These terms are used for convenience.
- The probability of success, call it $p$ is the same for each observation.

$$P(0 \cup 1) = P(0) + P(1) = 1$$

# Example 1

A researcher has access to 40 men and 40 women and selects 10 of them at random to participate in an experiment. The number of women selected can be represented by X. Is X binomially distributed?

- Read the question carefully. What is the probability of selecting a woman when there are 40 individuals. If a woman is chosen, what is the probability of selecting a women the second time?

# Example 2

A pharmaceutical company inspects a simple random sample of 10 empty plastic containers from a shipments of 10,000. They are examined for traces of benzene. Suppose that 10% of the containers in the shipment contain benzene. Let $X$ represent the number of containers contaminated with benzene. Is $X$ binomially distributed?

- Issue: Each time you sample one bottle, it affects the change that the next bottle will be contaminated. However given that the population is size 10,000 and the sample size is 20, the effect of one sample's success status on the next bottle's success status is negligible.
- Here the distribution of $X$ is *approximately* Binomial:

$$X \dot{\sim} Binom(10000, 0.10)$$

where $\dot{\sim}$ is read as "approximately distributed as".

# Binomial probability

If $X$ has the binomial distribution with $n$ observations and probability $p$ of success on each observation, the possible values of $X$ are 0, 1, 2, . . . , n. If $k$ is any one of these values,

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

- Read $\binom{n}{k}$ as “n choose k”. It counts the number of ways in which $k$ successes can be arranged among $n$ observations.
    - The binomial probability is this count multiplied by the probability of any one specific arrangement of the $k$ successes.

# Sampling distribution of binomial

# Definition: sampling distribution

A sampling distribution is shown as the distribution (with a histogram) of a sample statistic after taking many samples.

The distribution of the number of successes across many samples is called the sampling distribution for X. with mean # successes denoted by $\bar{x}$

The distribution of the proportion of successes across many samples is called the sampling distribution for $p$ with mean proportion successes denoted by $\hat{p}$

# Binomial approximation when $N$ is much larger than $n$

Choose a simple random sample of size $n$ from a population with proportion $p$ of successes. When the population size ($N$) is much larger than the sample, the count $X$ of successes in the sample has approximately the binomial distribution with parameters $n$ and $p$.

# Example Trial of 2

# Smoking status

In 1987, 29% of the adults in the United States smoked cigarettes, cigars, or pipes.

Let Y be a random variable that represents smoking status.

- $Y = 1$, an adult is currently a smoker
- $Y = 0$, an adult is not a current smoker

The two values of Smoking status are mutually exclusive and exhaustive.

What is the probability a randomly selected person is a smoker? $P(Y=1)$

What is the probability a randomly selected person is a non-smoker? $P(Y=0)$

# Smoking status

Suppose that we randomly select two individuals from the population of adults in the United States.

The random variable X represents the number of persons in the pair who are current smokers.

| First Person($Y_1$) | Second Person ($Y_2$) | Probability | Number of Smokers (X) |
|---|---|---|---|
| 0 | 0 | | 0 |
| 1 | 0 | | 1 |
| 0 | 1 | | 1 |
| 1 | 1 | | 2 |

# Smoking status

Suppose that we randomly select two individuals from the population of adults in the United States.

The random variable X represents the number of persons in the pair who are current smokers.

| First Person($Y_1$) | Second Person ($Y_2$) | Probability | Number of Smokers (X) |
|---|---|---|---|
| 0 | 0 | $(1-p)\times(1-p)$ | 0 |
| 1 | 0 | $p\times(1-p)$ | 1 |
| 0 | 1 | $(1-p)\times p$ | 1 |
| 1 | 1 | $p\times p)$ | 2 |

Remember that we can multiply to get the probabilities here because the events are independent

# Smoking status

Recall that for one trial, P(event) + P(no event) = 1

So P(smoker) + P(not smoker) =1

We know P(smoker) = 29% so :

1 - 0.29 = 0.71 (probability of non smoker)

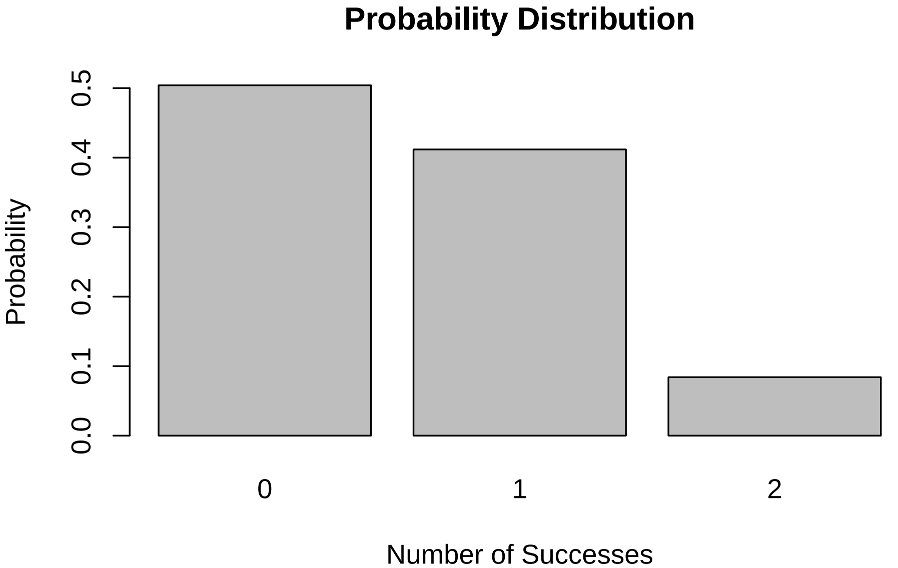
# Calculate by hand using a table

If 29% of US adults smoked, p=0.29, what are the values of these probabilities?

The random variable X represents the number of persons in the pair who are current smokers.

| First Person($Y_1$) | Second Person ($Y_2$) | Probability | Number of Smokers (X) |
|---|---|---|---|
| 0 | 0 | $(.71) \times (.71) = .5041$ | 0 |
| 1 | 0 | $.29 \times (.71) = .2059$ | 1 |
| 0 | 1 | $(.71) \times .29 = .2059$ | 1 |
| 1 | 1 | $.29 \times .29) = .0841$ | 2 |

# Probability distribution for 2 selected individuals



# Binomial Probability Distributions

In the binomial distribution the sum of all probabilities of potential outcomes equals 100% (1.0)

If you have a certain number of events with probability of success (p),

What is probability that X is occurs at least once P(X>=1)?

$$P\ (X >= 1) = 1 - P\ (X = 0)$$

If you have a certain number of events with probability of success (p),

what is probability that X occurs fewer than twice P(X<2)?

$$P(X < 2) = P\ (X = 1) + P\ (X = 0)$$

# Binomial Probability Distributions: for n trials

What if we are interested in the expected outcomes if select a larger group of individuals? It starts to get cumbersome to write out that table by hand. The general expression of the probability distribution of a binomial random variable X where x is the number of successes in a sample of size n.

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

where n = 1,2,3,. . . and x = 0,1,. . . n.

# Binomial Combinations: How many combinations of n people give x successes?

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

$$\binom{2}{0} = \frac{2!}{0!(2-0)!} = 1$$

$$\binom{2}{1} = \frac{2!}{1!(2-1)!} = 2$$

**remember that 0! $= 1$

# Binomial Probability Distributions

So for 1 success in 2 individuals (n=2, x=1) where 29% are smokers (p=0.29)

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

$$P(X = 1) = \binom{2}{1} 0.29^1 (1-0.29)^{2-1} = 0.4118$$

# Binomail Probability Distributions

We can use the formulas as shown to calculate the probability of a given number of successes from a binomial by hand

Or we could use R with ‘dbinom(#successes,size,probability of success)’

This function calculates the probability of observing `x` successes when $X \sim Binom(n, p)$

```
dbinom(1,size=2,prob=0.29)
```

```
## [1] 0.4118
```

let’s look further at sampling distributions using our container example. . .

# Example trial of size 10

# Sampling distribution of a count in R

First, set up a large population of size 10,000 where 10% of the containers are contaminated by benzene. We call `benzene` a “success” since it is coded as 1. We can see that 10% of the containers are contaminated and 1000 bottles are “successes”

We simulate these data:

```r
container.id <- 1:10000
benzene <- c(rep(0, 9000), rep(1, 1000))
pop_data <- data.frame(container.id, benzene)
```

# Sampling distribution of a count in R

```
# Calculate the population number of bottles contaminated by benzene and the
# population mean proportion
pop_stats <- pop_data %>% summarize(pop_num_successes = sum(benzene),
                                    pop_mean = mean(benzene))

pop_stats
```

```
##   pop_num_successes pop_mean
## 1              1000      0.1
```

# Sampling distribution of a count in R

Take a sample of size 10 from the population. Note that 10 is much smaller than 10,000.

- How many contaminated bottles are we expecting in the sample?
- Given that we sample 10, what is the full range of possible values we could see for X, the number of successes and $p$ the proportion of successes?
- Which values are most likely?

```
# first sample
set.seed(1)
sample_data <- pop_data %>% sample_n(10)
sample_data %>% summarize(sample_num_successes = sum(benzene),
                          sample_mean = mean(benzene))
```

```
##   sample_num_successes sample_mean
## 1                    1         0.1
```

# Sampling distribution of a count in R

We only took one sample, and got 2 successes for a sample mean of 20%. Is that usual or unusual?

To see what is most likely, we need to imagine repeatedly taking samples of size 10 from the population and calculating the sample number of successes and proportion of successes for each sample.

For the next few slides, we focus on the sampling distribution for X.

# Sampling distribution of a count in R

The embedded code takes 1000 samples each of size 10.

It then calculates the mean sample proportion and number of successes for each sample and stores all the results in a data frame.

You don't need to know how the code works.

# Sampling distribution of a count in R

Here are the first rows of the data frame we made on the previous slide. Each row represents an independent sample from the population.

```
head(many.sample.stats)
```

```
##   sample_proportion sample_num_successes sample.id
## 1               0.1                    1         1
## 2               0.1                    1         2
## 3               0.1                    1         3
## 4               0.3                    3         4
## 5               0.2                    2         5
## 6               0.1                    1         6
```

# Sampling distribution of a count in R

We want to know: Of the 1000 samples, what percent observed 0 contaminated bottles? What percent observed 1 contaminated bottle? And so on. We can used dplyr functions to calculate this and plot the results in a histogram.

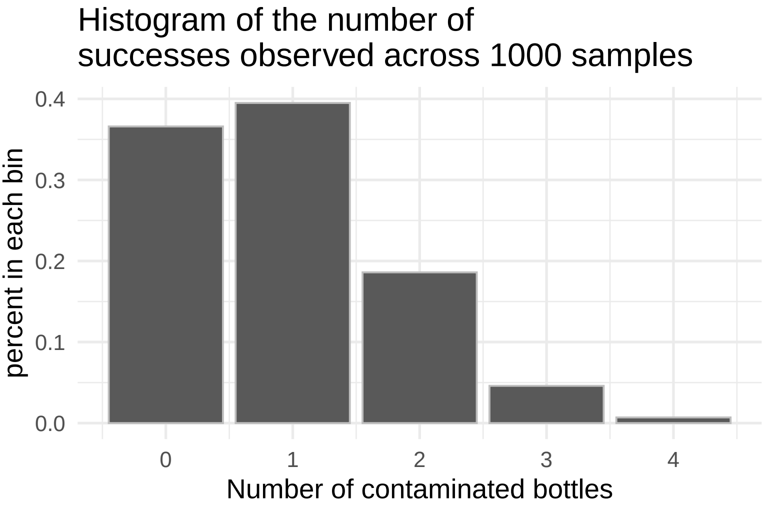
```
aggregated.stats <- many.sample.stats %>%
  group_by(sample_num_successes) %>%
  summarize(percent = n()/1000)
```

# Sampling distribution of a count in R

```
aggregated.stats
```

```
## # A tibble: 5 x 2
##   sample_num_successes percent
##                  <dbl>   <dbl>
## 1                    0   0.366
## 2                    1   0.395
## 3                    2   0.186
## 4                    3   0.046
## 5                    4   0.007
```

# Sampling distribution of a count in R



# Sampling distribution of a count in R

As we will see in a moment, this histogram *approximates* the shape of the binomial distribution with $n = 10$ and $p = 0.1$. Observing one success is the most likely outcome. Why is that?

# Worked probabilities, $x = 0$

We sampled n=10 bottles where the probability of success on any one pick is 10%.

- What is the chance of observing zero contaminated bottles?
- This means the first bottle is not contaminated and the second bottle is not contaminated, and . . . and the tenth bottle is not contaminated

$P(X_1 = 0 \text{ and } X2 = 0 \text{ and...and } X_{10} = 0)$

$= P(X_1 = 0) \times P(X_2 = 0) \times ... \times P(X_{10} = 0)$ , using the multiplication rule for independent events

$= (0.90)^{10}$

$= 0.3486784 = 34.9\%$

# Worked probabilities, x = 1

- What is the chance of observing exactly one contaminated bottle?
- Suppose that the first bottle was contaminated, then all the rest had to be not contaminated. What is the probability of observing this specific sequence of events?

$P(X_1 = 1 \text{ and } X2 = 0 \text{ and } X3 = 0 \text{ and...and } X_{10} = 0)$

$= P(X_1 = 1) \times P(X_2 = 0) \times P(X_3 = 0) ... \times P(X_{10} = 0)$

$= (0.1)^1 (0.90)^9$

$= 0.03874205 = 3.87\%$

But we're not done. This is only one specific way of observing exactly one contaminated bottle. What is another way? How many ways are there to observed exactly one contaminated bottle when there are ten bottles?

# Worked probabilities, $x = 1$

There are ten ways to observe exactly one contaminated bottle:

- 1, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 1, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 0, 1, 0, 0, 0, 0, 0, 0, 0
- 0, 0, 0, 1, 0, 0, 0, 0, 0, 0
- 0, 0, 0, 0, 1, 0, 0, 0, 0, 0
- . . .
- 0, 0, 0, 0, 0, 0, 0, 0, 0, 1

# Worked probabilities, x = 1

Remember

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

$$\binom{10}{1} = \frac{10!}{1!(10-1)!} = 10$$

# Worked probabilities, x = 1

Each of these ten ways has the same probability of occurring.

$P$(observed exactly 1 contaminated bottle) $=$

$P$(1st bottle is contaminated, and rest are not OR 2nd bottle is contaminated, and rest are not

$= (0.1)^1(0.9)^9 + (0.1)^1(0.9)^9 + ... + (0.1)^1(0.9)^9$, using the addition rule for disjoint events

$= 10 \times (0.1)^1(0.9)^9$

$= 0.3874205 = 38.7\%$

# Worked probabilities, $x = 1$

We can check our calculations using the `dbinom()` function in R.

```r
dbinom(x = 1, size = 10, prob = 0.1)
```

```
## [1] 0.3874205
```

This is exactly the answer we obtained.

# Worked probabilities, $x = 2$

What is chance of observing exactly two contaminated bottles?

Following the same line of thinking, suppose that the first two bottles were contaminated. The chance of this happening is:

$(0.1)^2(0.9)^8 = 0.004303672$

But how many ways are there to observe exactly two contaminated bottles?

# Worked probabilities, x = 2

Remember

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

$$\binom{10}{2} = \frac{10!}{2!(10-2)!} = 45$$

You could write out all the possibilities like last time, but there are a lot more!

# Worked probabilities, $x = 2$

Note: we can get our calculators or R to perform this calculation for us. On our calculator, we need the button $\binom{n}{k}$, pronounced "n choose k", and asks how many ways are there to have $k$ successes when there are $n$ individuals? In R we need the function `choose(n, k)`

```r
choose(10, 2)
```

```
## [1] 45
```

There are 45 ways to observe exactly two contaminated bottles when you have ten bottles observed.

Make sure you can also perform this calculation on your calculator!

# Worked probabilities, x = 2

To get the probability of observing exactly 2 contaminated bottles, because all of the possible combinations are equally possible, we can multiply 45 by the probability of observing the first two bottles as being contaminated:

$45 \times (0.1)^2(0.9)^8 = 0.1937102 = 19.4\%$

Check using R:

```
#fill in during class
```

# All of the combinations with 10 bottles

Each of these is written as $\binom{10}{k}$, where k is 0, 1, 2, . . . , 10. This is known as the binomial coefficient.

Let's compute `choose(n, k)`, for n=10, and k=0, 1, 2, . . . , 10:

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
## [1,]    1   10   45  120  210  252  210  120   45    10     1
```

Notice the symmetric structure of `choose(n, k)`. Why is it symmetric?

# R code recap

We have seen two important new functions in R today:

dbinom() to calculate the discreet probability of an outcome in a given binomial distribution

choose() to calculate the number of ways we can have k successes from n trials

# Comic Relief