Probability distributions and inference

Characteristics of a binomial distribution

Jenine Harris
Brown School

Defining and using the probability distributions to infer from a sample

- In addition to frequency distributions, information can be represented in a **probability distribution**.
- A probability distribution is the set of probabilities that each possible value (or range of values) of a variable occurs.
- An example:
  - In 2017, 26 of the 51 US states had prescription drug monitoring programs
  - If you were to put the names of US states on slips of paper and select one without looking, the probability that the state selected would have a monitoring program in 2017 would be 26 out of 51 (50 states plus Washington DC), which is 51%.
  - The probability of selected a state with no monitoring program would be 25/51 or 49%. The set of these probabilities together is the probability distribution for the PDMP policy.
- A probability distribution is the numeric or visual representation of the *set of probabilities that each value or range of values of a variable occurs*.

Characteristics of probability distributions

- Probability distributions have two important characteristics:
  - The probability of each real value of some variable is non-negative; it is either 0 or positive
  - The sum of the probabilities of all possible values of a variable is 1, representing 100%
- Each probability distribution falls into one of two categories: **discrete** or **continuous**.
  - Discrete probability distributions show probabilities for variables that can only have certain values, which includes categorical variables and variables that must be measured in whole numbers like *number of people texting during class*.
  - Continuous probability distributions show probabilities for values, or ranges of values, of a continuous variable that can take any value in some range.

Types of probability distributions

- These two categories of probability distributions are the foundation for most statistical tests in social science.
- The two probability distributions in particular that are commonly used and good examples for learning how a probability distribution works are: the *binomial distribution* and the *normal distribution*.
    - The binomial distribution is a *discrete probability distribution* and applies to probability for binary categorical variables with specific characteristics.
    - The *normal distribution* is a *continuous probability distribution* and applies to probability for continuous variables.

The characteristics and uses of a binomial distribution of a binary variable

- The binomial distribution is a *discrete probability distribution* used to understand what may happen when a variable has two possible outcomes, such as *eats brownies* and *does not eat brownies*. w
- The most common example is flipping a coin, but there are many variables that have two outcomes: alive or dead, smoker or non-smoker, voted or did not vote, depressed or not depressed.
- Whether a state adopted, or did not adopt, an opioid policy is another example of a variable with two possible outcomes: policy, no policy.

The binomial distribution and binomial random variables

- More specifically, the *binomial distribution* is used to represent the distribution of a *binomial random variable*, which has the following properties:
  - A variable is measured in the same way n times
  - There are only two possible values of the variable, often called "success" and "failure"
  - Each observed value is independent of the others
  - The probability of "success" is the same for each observation
  - The random variable is the number of successes in n measurements

Defining the binomial distribution

The binomial distribution is defined by two things:

- n - the number of observations (e.g., coin flips, people surveyed, states selected)
- p - the probability of *success* (e.g., 50% chance of heads for a coin flip, 51% chance of a state having a prescription drug monitoring program)

Using distributions to learn about populations from samples

- Researchers often work with **samples** instead of **populations**.
- In the case of the state data on opioid policies, all states are included, so this is the entire **population** of states.
- When it is feasible to measure the entire population, reporting summary statistics like percentages of the population is usually sufficient.
- However, populations are often expensive and logistically difficult or impossible to measure, so a subset of a population is often measured instead.
- This subset is a **sample**.
- For example, someone studying how the number of opioid deaths in a state relates to one or two characteristics of states might choose a sample of 20 or 30 states to examine.

Statistical properties of a binomial random variable

- If it wasn't clear which 51% of states had PDMP and a study team wanted to include exactly 10 states with PDMP, they could use the binomial distribution to decide how many states to include in a sample in order to have a good chance of picking 10 with PDMP.
- Before using the binomial distribution, ensure that the PDMP variable has the properties of a **binomial random variable**:
    - The existence of a monitoring program would be for measured each state
    - The only two possible values are "success" for having a PDMP and "failure" for not having a PDMP
    - Each state is independent of the other states
    - The probability of having a program is the same for each state
- There could be some concern about the third and fourth assumptions of states being independent of other states and having the same probability of having a program.
- State lawmakers are definitely independent of each other, but often neighboring states will often be more similar to each other than they are to states in other parts of the country.
- The influence of geography is complex and should definitely be seriously considered before publishing any research with these data, but these data are ok to explore the statistical concepts.
- Consider the states and counties in the data as independent of each other.

Expected values

- The **expected value** of a binomial random variable is np, where n is sample size and p is the probability of a *success*.
- In this example, if the sample size is 20 and the probability of success (having a PDMP) is 51%, which is formatted as a proportion rather than a percentage for the purposes of this calculation, the expected value of the binomial random variable after taking a sample of 20 states would be 20$\cdot$.51 or 10.2.
    - This means that a sample of 20 states is likely to have 10.2 states with PDMP.
- The expected value is useful, but since the value of p is a probability (not a certainty), the expected value will not occur every time a sample is taken.
- The **probability mass function** for the binomial distribution can be used to compute the probability that any given sample of 20 states would have *exactly* 10 states with PDMP.
    - A probability mass function computes the probability that an *exact* number of successes happens for a discrete random variable, given n and p.
    - In this case, the probability mass function will give them the probability of getting 10 states with a PDMP if they selected 20 states at random from the population of states where the likelihood of any one state having a prescription drug monitoring program was 51%.

Probability Mass Function (PMF) equation

- In the probability mass function equation:
  - x represents the specific number of successes of interest
  - n is the sample size
  - p is the probability of success

$$f(x, n, p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

Try substituting values into the PMF equation

Substitute the values of x, n, and p from the scenario into the PMF equation to find the probability of choosing exactly 10 states with prescription drug monitoring in a sample of 20 states from a population where 51% of states have drug monitoring:

$$f(10, 20, .51) = \binom{20}{10} \cdot .51^{10} \cdot (1 - .51)^{(20-10)} = .175$$

- There is a 17.5% probability of choosing *exactly* 10 states with drug monitoring programs when choosing 20 states at random from the population of states where 51% of states have a drug monitoring program.

Using the PMF equation in R

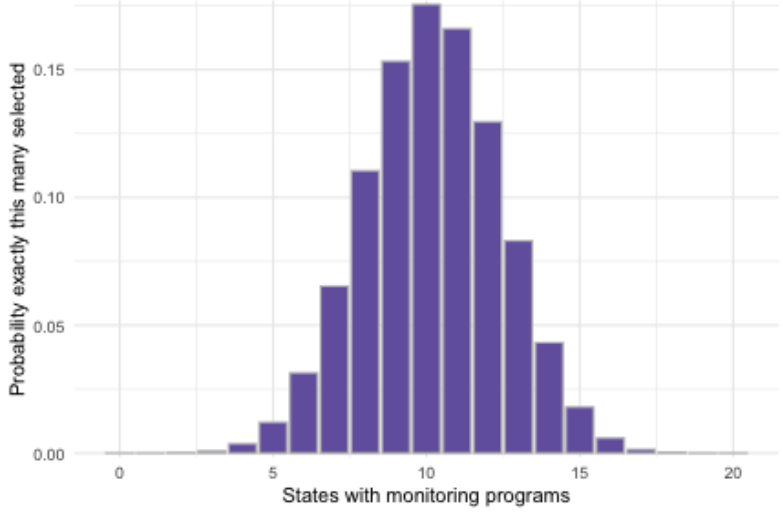
- R has a function built in for using the binomial distribution.
- The `dbinom()` function in base R uses the PMF equation to compute the probability given the number of successes (`x`), the sample size (`size =`), and the probability of successes (`prob =`):

```
# where successes = 10, n = 20, p = .51
dbinom(x = 10, size = 20, prob = .51)
```

```
## [1] 0.1754935
```

Visualizing the binomial distribution

- The binomial distribution can be displayed graphically and used to understand the probability of getting a specific number of successes or a range of successes (e.g., *10 or more successes*).
- A plot of the *probability mass function* shows the distribution of probabilities of different numbers of successes.
- For example, this graph shows the probability of getting a certain number of states with monitoring programs in a sample when 20 are selected from a population where 51% have monitoring programs.



Interpreting the graph & R code for finding probability

- Along the x-axis in the plot are the number of states selected that *have a PDMP program* (in a sample where n = 20 and p = .51).
- Along the y-axis is the probability of selecting *exactly* that many states.
- For example, it looks like there is about a one percent chance of exactly five successes.
- So, in choosing 20 states at random from all the states, there appears to be approximately a 1% chance that *exactly* five of them will have a PDMP.
- Use `dbinom()` to check this...

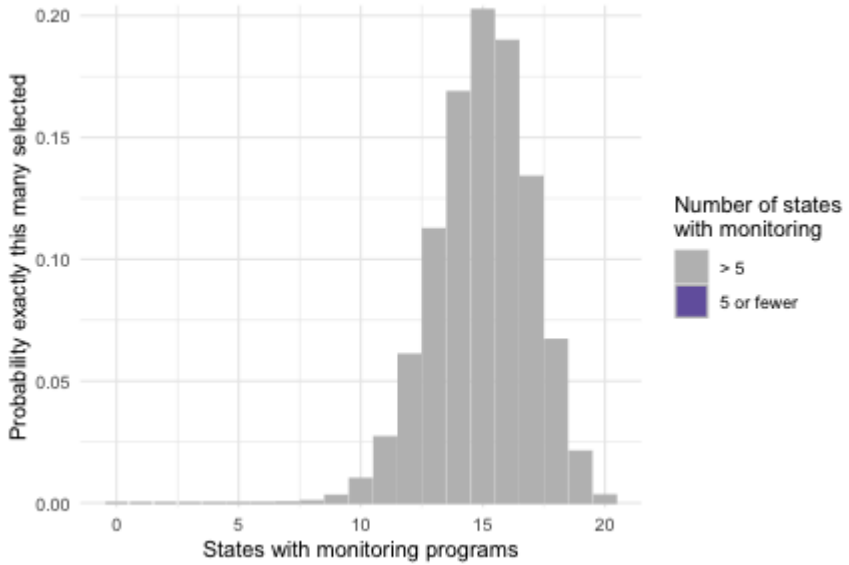
```
# 5 successes from a sample of 20 with 51% probability of success
dbinom(x = 5, size = 20, prob = .51)
```

```
## [1] 0.01205691
```

- By multiplying the result by 100, it appears there is a 1.21% chance of choosing 20 states at random from a population where 51% have PDMP and the sample has exactly five with a monitoring plan.

Probabilities in different settings

- What would happen if the percentage of states with programs increased? Or if 75% of states had programs, what would the probability be that exactly five out of 20 selected at random would have a PDMP?
- A binomial distribution for a sample size of 20 with a 75% probability of success would look like:



Interpreting the graph

- The distribution has shifted to the right on the x-axis and getting exactly five states with a monitoring program is even less probable than before when the probability of PDMP per state was 51%.
- This makes sense; it would be wierd to get so few states with a monitoring program in a sample if 75% of all states had this program.
- Use `dbinom()` to check how low the probability of five successes (states with monitoring programs) actually is:

```
# 5 successes from 20 selections with 75% probability of success
dbinom(x = 5, size=20, prob=.75)
```

```
## [1] 3.426496e-06
```

- The -06 in the output is **scientific notation**.
- When 75% of states have drug monitoring program, there is a 0.00034% chance of choosing exactly five states with PDMP out of 20 selected.

Probabilities for a range of values

- So far the probabilities were very small for scenarios of getting exactly five or exactly 10 states with monitoring programs in a sample.
- The **cumulative distribution function** for the binomial distribution can determine the probability of getting some *range of values*, which is often more useful than finding the probability of one specific number of successes.
- For example, what is the probability selecting 20 states and getting five *or fewer* states with monitoring programs?
- Likewise, what is the probability of getting 10 *or more* states with monitoring programs in a sample of 20?
- The equation for the cumulative distribution function can be used to find out:

$$f(x, n, p) = \sum_{x=0}^{x_{floor}} \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

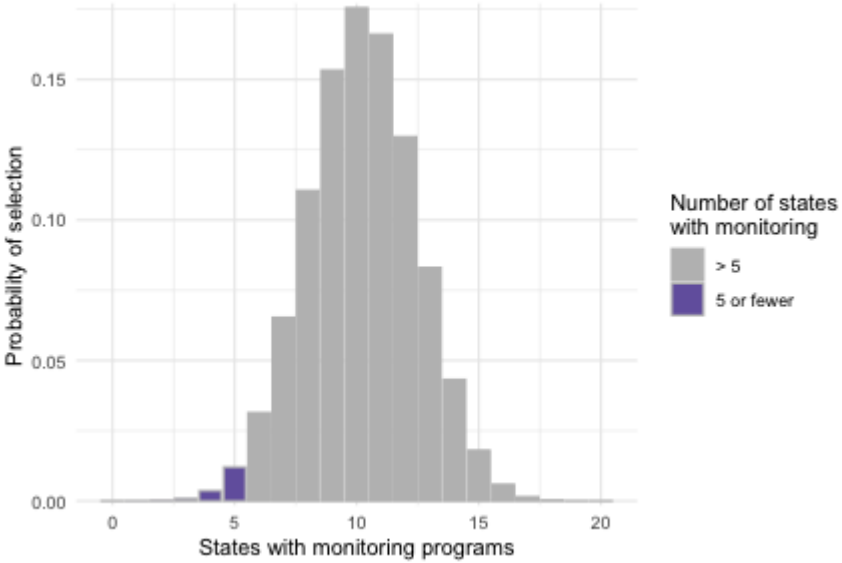
Interpreting the cumulative distribution function

$$f(x, n, p) = \sum_{x=0}^{x_{floor}} \binom{n}{x} \cdot p^x \cdot (1-p)^{(n-x)}$$

- x is number of successes
- x_{floor} is the largest integer less than or equal to x
- n is the sample size
- p is the probability of success.
- The cumulative distribution function computes the probability of getting x *or fewer* successes.

Visualizing the cumulative probability

- The graph to shows the probability of five or fewer states with PDMP in a sample of 20.



Interpreting the graph and using pbinom()

- The plot shows less than a 2% chance of five or fewer given the size of the purple section of Figure \@ref(fig:c4main10).
- Use `pbinom()` to compute the probability of five or fewer successes (`q = 5`) for a sample of 20 (`size = 20`) from a population with a 51% probability of success (`prob = .51`):

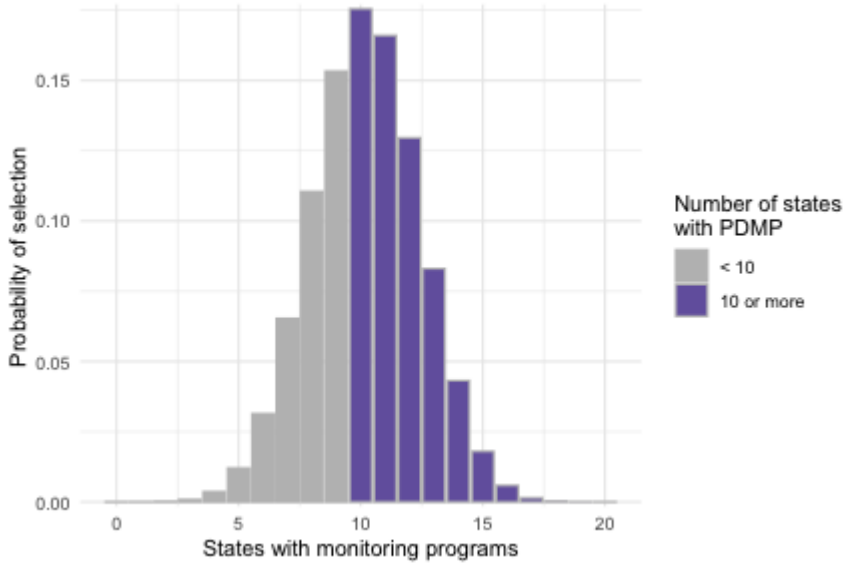
```
# 5 or fewer successes from 20 selections
# with 51% probability of success
pbinom(q = 5, size = 20, prob = .51)
```

```
## [1] 0.01664024
```

- The probability of 0.02 or 1.66% makes sense.
- The purple section of the graph appears to be quite small.

Another example

- Try computing the probability for 10 successes *or more*.



Computing the probability for 10 or more successes

- Use `pbinom()` to compute the value.
- The range of values is on the right side of the graph instead of the left side, so she looks at the help documentation and finds that there is a setting for `pbinom()` to get the right side of the graph but specifying `lower.tail = FALSE`.
- The `lower.tail =` option has the default value of true, so it did not need to be included when estimating the lower tail, but would need to be added when it is the upper tail on the right to be estimated.

```
# 10 or more successes from 20 selections
# with 51% probability of success
# get right side of graph
pbinom(q = 10, size = 20, prob = .51, lower.tail = FALSE)
```

```
## [1] 0.4474504
```

Interpreting and adjusting the calculation

- This did not seem right.
- It does not match the graph which clearly shows more than half of the histogram is shaded purple.
- The `pbinom()` calculation with the default of `lower.tail = TRUE` is for 10 or fewer successes and is computing *higher than 10* rather than *10 or higher* so it is missing some of the graph in the middle.
- Change the arguments for `pbinom()` to get the probability of *10 or more* successes:

```
# 10 or more successes from 20 selections
# with 51% probability of success
# pbinom computes left tail, so subtract from 1 for right tail percent
pbinom(q = 9, size=20, prob=.51, lower.tail = FALSE)
```

```
## [1] 0.6229439
```

- This seems like a more accurate value for the purple part of the histogram.
- The results show a 62.29% probability of selecting 10 or more states with PDMP in a sample of 20 from a population of states where 51% have PDMP.

Applying these concepts to real data

- Import data from the Kaiser Family Foundation showing the actual states and their opioid policy adoption as of 2017.

```
# bring in the opioid policy 2018 data and check it out
opioid.policy.kff <- read.csv(file = "/Users/harrisj/Box/teaching/Teachi

# check the data frame
summary(object = opioid.policy.kff)
```

```
##        X                Opioid.Quantity.Limits Clinical.Edits.in.Claim.System
##  Length:51          Length:51                 Length:51
##  Class :character   Class :character          Class :character
##  Mode  :character   Mode  :character          Mode  :character
##  Opioid.Step.Therapy.Requirements
##  Length:51
##  Class :character
##  Mode  :character
##  Other.Prior.Authorization.Requirements.for.Opioids
##  Length:51
##  Class :character
##  Mode  :character
##  Required.Use.of.Prescription.Drug.Monitoring.Programs
##  Length:51
```

Test a scenario

- Pretend to only have a budget to collect data from a sample of 25 states and would like at least 15 of these to have monitoring programs.
- Use the binomial distribution with that sample of 25 states and see if the sample has 15 or more states with monitoring programs.
- First, check what the binomial distribution suggests the probability is of getting a sample with 15 or more states with monitoring programs out of 25 states.
- For this they need the percentage of states that currently have monitoring programs (the *success* rate).
- An examination of the KFF data shows 32/51, or 63% of states had PDMP as of 2017.
- Use the `pbinom()` code and revise it for the new scenario:

```
# 15 or more successes from 20 selections
# with 63% probability of success
# pbinom computes left tail, so use lower.tail = FALSE
pbinom(q = 14, size=25, prob=.63, lower.tail = FALSE)
```

```
## [1] 0.7018992
```

Interpreting the probability and taking a random sample to compare

- The probability of selecting a sample of 25 states where 15 or more states have PDMP is 70.19%.
- Use`set.seed()` when conducting random sampling since it will result in the same sample to be taken each time the code is run, which makes sampling more reproducible.
- `sample_n()` can be used to take a sample.
    - The arguments for `sample_n()` are `size =` which is where to put the size of the sample.

```
# set a starting value for sampling
set.seed(seed = 3)
# sample 25 states and check file
opioid.policy.kff %>%
  select(Required.Use.of.Prescription.Drug.Monitoring.Programs) %>%
  sample_n(size = 25) %>%
  table()
```

```
## .
##  No Yes
```

Interpreting the output

- The output shows `No 8` and `Yes 17`, so the sample has 17 states with monitoring programs, which is more than 15.
- The binomial distribution indicated there was a 70% chance they would see 15 or more states with PDMP in a sample this big.
- Try a few more samples with different seeds just to see what happens:

```
# set a starting value for sampling
set.seed(seed = 10)

# sample 25 states and check file
opioid.policy.kff %>%
  select(Required.Use.of.Prescription.Drug.Monitoring.Programs) %>%
  sample_n(size = 25) %>%
  table()
```

```
## .
##  No Yes
##  10  15
```

Try another sample

This sample has 15 states with PDMP. Nancy thought they should try at least one more.

```
# set a starting value for sampling
set.seed(seed = 999)

# sample 20 states and check file
opioid.policy.kff %>%
  select(Required.Use.of.Prescription.Drug.Monitoring.Programs) %>%
  sample_n(size = 25) %>%
  table()
```

```
## .
##  No Yes
##  11  14
```

- This one has 14 states with PDMP.
- Out of three samples of 25 states, two samples had 15 or more states with PDMP and one sample had fewer than 15 states with PDMP.
- This is consistent with the binomial distribution prediction that 70% of the time a sample of size 25 will have at least 15 states with PDMP.