Lecture 4

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/09/07

Slides Updated: 2023-09-06

Agenda

1. Recap of Lecture 3
2. Random variables
3. Probability distributions
4. Expectations
5. Applied example in `R`

Recap of Lecture 3

- We now have an intellectual foundation for probability
  - Probabilistic events in the context of an **experiment** and **simple events**
  - All possible events define the **sample space**
  - 3 axioms for how to assign probabilities to events (lecture 2)
  - 4 tools to decompose and compose events of interest (lecture 3)

One additional tool

- Assigning probabilities in a sample space consisting of **equiprobable events**
- S consists of n **equiprobable** events E such that $P(E_1) = P(E_2) = \cdots = P(E_n)$
  - Recall we define A as a subset of S: some subset of sample points that make up S
  - Then $P(A) = \frac{|A|}{n}$ where $|A|$ indicates the number of elements in A

Random Variables

- Experiment where events of interest are **numerical**
    - Identified in a meaningful way by numbers
    - I.e., number of seats held by Republican Party in the House after a midterm election
    - We assign a *real number* to each point in the sample space
    - Call this number the variable Y
- What is a variable?
    - A logical grouping of attributes
    - Take on values that are **exhaustive** and **mutually exclusive**
- Thus each sample point can only take on one value of Y, but the same values of Y may be assigned to multiple sample points

Functions

- We **map** numeric values using a **function**
- Thus the numeric random variable Y is a **function** of the sample points in S
- A function is a mathematical relation assigning each element of one set (the source) to one and only one element of another set (the target)
  - The function's **source** is S and its **target** is Y
  - $f : S \to Y$
  - This function (and by extension, Y) is a **random variable**
- Whenever we talk about a random variable, *we are really talking about a function* that maps each simple event in a sample space S to a meaningful number

Notation

- Random variables expressed with capital letters: i.e., Y
- Interested in the probability a random variable takes on some value
  - Probability that $Y = 0$ written as $P(Y = 0)$
- Denote observed or hypothetical values of Y with lowercase letters
  - $P(Y = y)$
- Still fundamentally interested in **events of interest** A, but denote with numbers a
  - $A \equiv \{\text{all sample points such that } Y = a\}$

Quick Detour: Random Samples

- Our experiment is the drawing of a **sample** from a population
  - **Sample**: the units selected for analysis
  - **Population**: the group of units about which we want to make inferences
- The **design** of our experiment is the method of sampling
  - Do we sample *with replacement*? Units are put back into the population after being sampled, and we might re-sample them again

Quick Detour: Random Samples

- Most common design is **random sampling**
  - Let N be the number of elements in the population and n be the number of elements in our sample
  - How many different samples without replacement can we draw?
  - $\left(\frac{N}{n}\right) = \frac{N!}{n!(N-n)!}$
  - If we draw these n elements with equal probability, this is a **random sample**

Back to RVs: Probability Distributions

- Start with **discrete** random variables
  - Y is discrete if it can only take on finite or countably infinite number of distinct values
  - "Countably infinite": a one-to-one correspondence with the integers
- To make inferences about the **population** based on a **sample**:
  - Need to know the probability of observing a particular event
  - Events are numerical events corresponding to values y of discrete random variables Y
  - $P(Y = y)$ for all the values Y can take on
  - The collection of these probabilities is a **probability distribution**

Example: dice

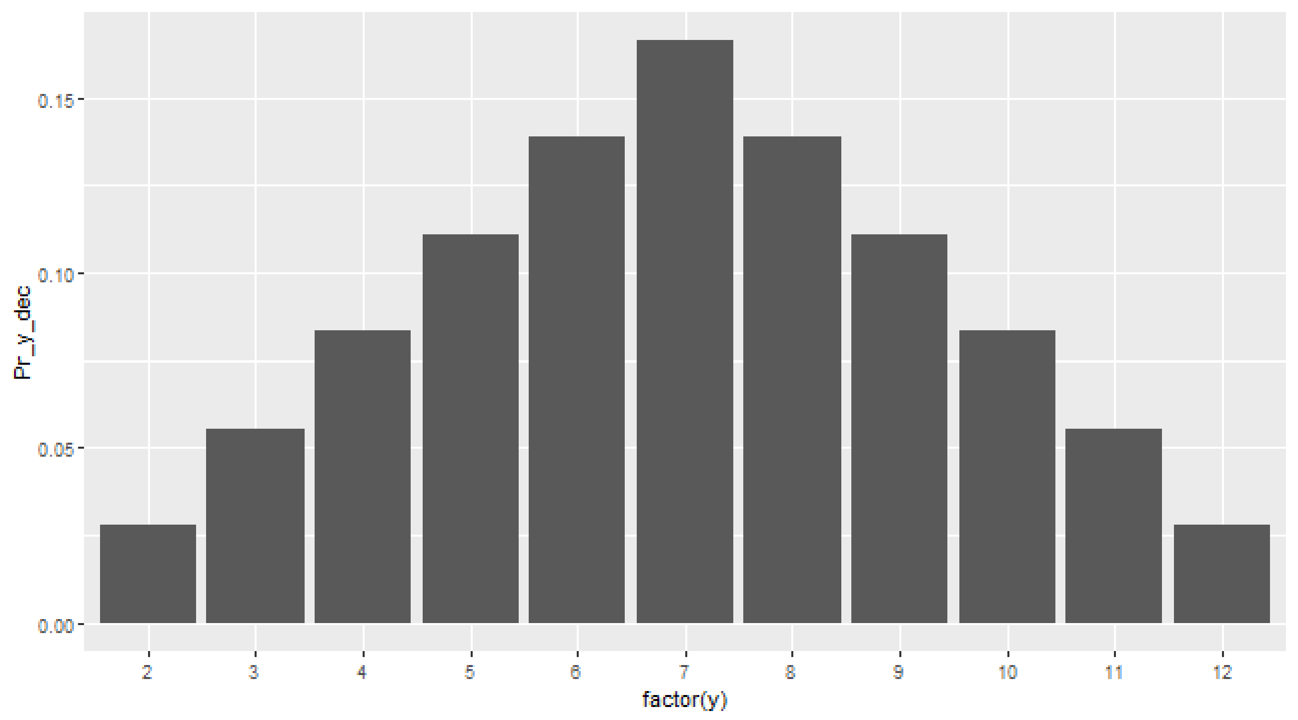
- Experiment: roll a pair of six-sided dice and record the sum of their faces
  - Sample space consists of 36 simple events
  - Random variable Y is the sum of the faces
  - $P(Y = y) = \sum_{E_i : Y(E_i) = y} P(E_i)$
  - Sometimes written as $p(y)$
- Can express Y's probability distribution as a **table**, a **graph**, or a **function**

Probability Distribution: Table

```
##     y samples Pr_y Pr_y_dec
## 1   2       1 1/36   0.0278
## 2   3       2 2/36   0.0556
## 3   4       3 3/36   0.0833
## 4   5       4 4/36   0.1111
## 5   6       5 5/36   0.1389
## 6   7       6 6/36   0.1667
## 7   8       5 5/36   0.1389
## 8   9       4 4/36   0.1111
## 9  10       3 3/36   0.0833
## 10 11       2 2/36   0.0556
## 11 12       1 1/36   0.0278
```

Probability Distribution: Graph

```r
p %>%
  ggplot(aes(x = factor(y),y = Pr_y_dec)) +
  geom_bar(stat = 'identity')
```



Probability Distribution: Function

$$P(Y = y) = p(y) = \frac{6 - |7 - y|}{36}, \; y = \{1, 2, 3, \ldots, 12\}$$

```
pdf_dice <- function(y) {
  (6 - abs(7 - y)) / 36
}
pdf_dice(y = 2:12)
```

```
##  [1] 0.02777778 0.05555556 0.08333333 0.11111111 0.13888889
##  [6] 0.16666667 0.13888889 0.11111111 0.08333333 0.05555556
## [11] 0.02777778
```

- Also called a **probability mass function** or **PMF**
- PDF is a **theoretical model** for the empirical distribution of data associated with a real population
  - If we re-roll a pair of dice multiple times, empirical distribution would look *like* the theoretical probability distribution

Expectations

- We can summarize a random variable with its central tendency and dispersion
- We can specify and manipulate formulas describing random variables using the **expectations operator**
  - **Expected value** of Y is $E(Y) \equiv \sum_y yp(y)$
  - Each possible value of Y multiplied by the probability of it appearing, summed up over all y
  - Apply this to the dice example!
- The **expected value** is how we talk about the central tendency of a random variable with a theoretical probability distribution
  - Equivalent to the concept of the *mean of an empirical frequency distribution*

Expectations

- Recall that the probability distribution of a random variable is a *theoretical model* for the empirical distribution of data **associated with a real population**
    - If the theoretical model is **accurate**, then $E(Y) = \mu$
- μ is the **population mean** which is a "parameter"
    - **Parameter**: characteristic of the distribution Y in the population that we never actually observe

Expectations

- The expected value concept can be applied to any **function of a random variable**
    - Consider any real-valued function of Y, denoted $g(Y)$
    - $E[g(Y)] = \sum_y g(y)p(y)$
- Instead of summing over the discrete values of y multiplied by their probability $p(y)$, we are summing over the discrete values of y that are transformed with the function $g(y)$
- NB: $E[g(Y)] = \sum_y g(y)p(y)$ is not a definition. We have to **prove** it.

A proof

- Denote a random variable Y taking on n values $y_1, y_2, \ldots, y_n$
- Denote a function $g(y)$ that takes on m different values $g_1, g_2, \ldots, g_m, \ m \leq n$
- Note that $g(Y)$ is itself a random variable
    - This means we can denote a new probability function p^* that describes the probability that g takes on a value g_i
    - $p^*(g_i) = P[g(Y) = g_i]$
    - $p^*(g_i) = \sum_{y_j : g(y_j) = g_i} p(y_j)$
    - Definition: $y_j : g(y_j) = g_i$ means "all y_j such that $g = g_i$ when evaluated at y_j"

Proof contd

- Definition of expected value: $E[g(Y)] = \sum_{i=1}^{m} g_i p^*(g_i)$
- Substitute: $E[g(Y)] = \sum_{i=1}^{m} g_i \left(\sum_{y_j : g(y_j) = g_i} p(y_j) \right)$
- Rearrange: $E[g(Y)] = \sum_{i=1}^{m} \left(\sum_{y_j : g(y_j) = g_i} g_i p(y_j) \right)$
- Substitute: $E[g(Y)] = \sum_{j=1}^{n} g(y_j) p(y_j)$
- Simplify: $E[g(Y)] = \sum_{y} g(y) p(y) \blacksquare$

Variance

- Using these tools, we can also define the variance of Y
- Remember that the variance of an empirical variable is $s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$
- Same idea for a random variable!
- $VAR(Y) \equiv E[(Y - E(Y))^2]$
- If Y accurately describes the population distribution, then $VAR(Y) = E[(Y - \mu)^2]$
- Denote $VAR(Y) = \sigma^2$ and the standard deviation of Y is $\sqrt{\sigma^2} = \sigma$

Example

Table 3.3 Probability distribution for Y

y	$p(y)$
0	1/8
1	1/4
2	3/8
3	1/4

- What is the mean, variance, and standard deviation of Y?
- Mean: $E(Y) = \sum_{y=0}^{3} yp(y) = (0)(1/8) + (1)(1/4) + (2)(3/8) + (3)(1/4) = 1.75$
- Variance: $\sigma^2 = E[(Y-\mu)^2] = \sum_{y=0}^{3}(y-\mu)^2 p(y)$
  - $(0-1.75)^2(1/8) + (1-1.75)^2(1/8) + (2-1.75)^2(1/8) + (3-1.75)^2(1/8) = 0.9375$
- Standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{0.9375} = 0.97$

Helpful results

- $E(c) = c$
  - Let $g(Y) \equiv c$
  - $E(c) = \sum_y cp(y)$
  - $E(c) = c \sum_y p(y)$
  - Axiom 2: $\sum_y p(y) = 1$
  - Thus $E(c) = c$ $\blacksquare$

Helpful results

- $E[cg(Y)] = cE[g(Y)]$
  - $E[cg(Y)] = \sum_y cg(y)p(y)$
  - $E[cg(Y)] = c\sum_y g(y)p(y)$
  - $E[cg(Y)] = cE[g(Y)]$ ■

Helpful results

- We can **distribute expectations**: consider $k = 2$
  - $g_1(Y) + g_2(Y)$ is a function of Y: $E[g_1(Y) + g_2(Y)] = \sum_y [g_1(y) + g_2(y)]p(y)$
  - $E[g_1(Y) + g_2(Y)] = \sum_y [g_1(y)p(y)] + \sum_y [g_2(y)p(y)]$
  - $E[g_1(Y) + g_2(Y)] = E[g_1(Y)] + E[g_2(Y)]$ ■

Demonstration in R

- Let's take a detour from this abstract work!

- Create a new `RMarkdown` file, require `tidyverse`, and load the data

```
require(tidyverse)

df <- read_rds('https://github.com/jbisbee1/PSCI_8356/raw/main/Lectures/Data/sc_debt.Rds')
```

Looking at the data

- Always *always* **always** look at your data!

```
df
```

```
## # A tibble: 2,546 × 16
##    unitid instnm        stabbr grad_…¹ control region preddeg
##     <int> <chr>         <chr>    <int> <chr>   <chr>  <chr>
##  1 100654 Alabama A &… AL       33375 Public  South… Bachel…
##  2 100663 University … AL       22500 Public  South… Bachel…
##  3 100690 Amridge Uni… AL       27334 Private South… Associ…
##  4 100706 University … AL       21607 Public  South… Bachel…
##  5 100724 Alabama Sta… AL       32000 Public  South… Bachel…
##  6 100751 The Univers… AL       23250 Public  South… Bachel…
##  7 100760 Central Ala… AL       12500 Public  South… Associ…
##  8 100812 Athens Stat… AL       19500 Public  South… Bachel…
##  9 100830 Auburn Univ… AL       24826 Public  South… Bachel…
## 10 100858 Auburn Univ… AL       21281 Public  South… Bachel…
## # … with 2,536 more rows, 9 more variables: openadmp <int>,
## #   adm_rate <dbl>, ccbasic <int>, sat_avg <int>,
## #   md_earn_wne_p6 <int>, ugds <int>, costt4_a <int>,
## #   selective <dbl>, research_u <dbl>, and abbreviated
## #   variable name ¹grad_debt_mdn
```

Looking at the data

- What are the **units of observation**?
- What are the **variables**?
    - What is the definition of a variable?

Looking at the data

- Can you find an example of a **nominal** variable? What about an **ordinal**, **interval**, and **ratio**?

```
# Some nominal variables (what is the definition?)
df %>%
  select(instnm,stabbr,control,region)
```

```
## # A tibble: 2,546 × 4
##    instnm                              stabbr control region
##    <chr>                               <chr>  <chr>   <chr>
##  1 Alabama A & M University            AL     Public  South…
##  2 University of Alabama at Birmingham AL     Public  South…
##  3 Amridge University                  AL     Private South…
##  4 University of Alabama in Huntsville AL     Public  South…
##  5 Alabama State University            AL     Public  South…
##  6 The University of Alabama           AL     Public  South…
##  7 Central Alabama Community College   AL     Public  South…
##  8 Athens State University             AL     Public  South…
##  9 Auburn University at Montgomery     AL     Public  South…
## 10 Auburn University                   AL     Public  South…
## # … with 2,536 more rows
```

Looking at the data

- Can you find an example of a **nominal** variable? What about an **ordinal**, **interval**, and **ratio**?

```r
# Is this an ordinal variable?
df %>%
  select(preddeg)
```

```
## # A tibble: 2,546 × 1
##    preddeg
##    <chr>
##  1 Bachelor's
##  2 Bachelor's
##  3 Associate
##  4 Bachelor's
##  5 Bachelor's
##  6 Bachelor's
##  7 Associate
##  8 Bachelor's
##  9 Bachelor's
## 10 Bachelor's
## # … with 2,536 more rows
```

Looking at the data

- Can you find an example of a **nominal** variable? What about an **ordinal**, **interval**, and **ratio**?

```r
# Is this an interval or a ratio variable?
df %>%
  select(sat_avg)
```

```
## # A tibble: 2,546 × 1
##    sat_avg
##      <int>
##  1     939
##  2    1234
##  3      NA
##  4    1319
##  5     946
##  6    1261
##  7      NA
##  8      NA
##  9    1082
## 10    1300
## # … with 2,536 more rows
```

Summarizing data

- Recall the different approaches to **summarizing data**

Summarizing data: Frequency tables

- Recall the different approaches to **summarizing data**

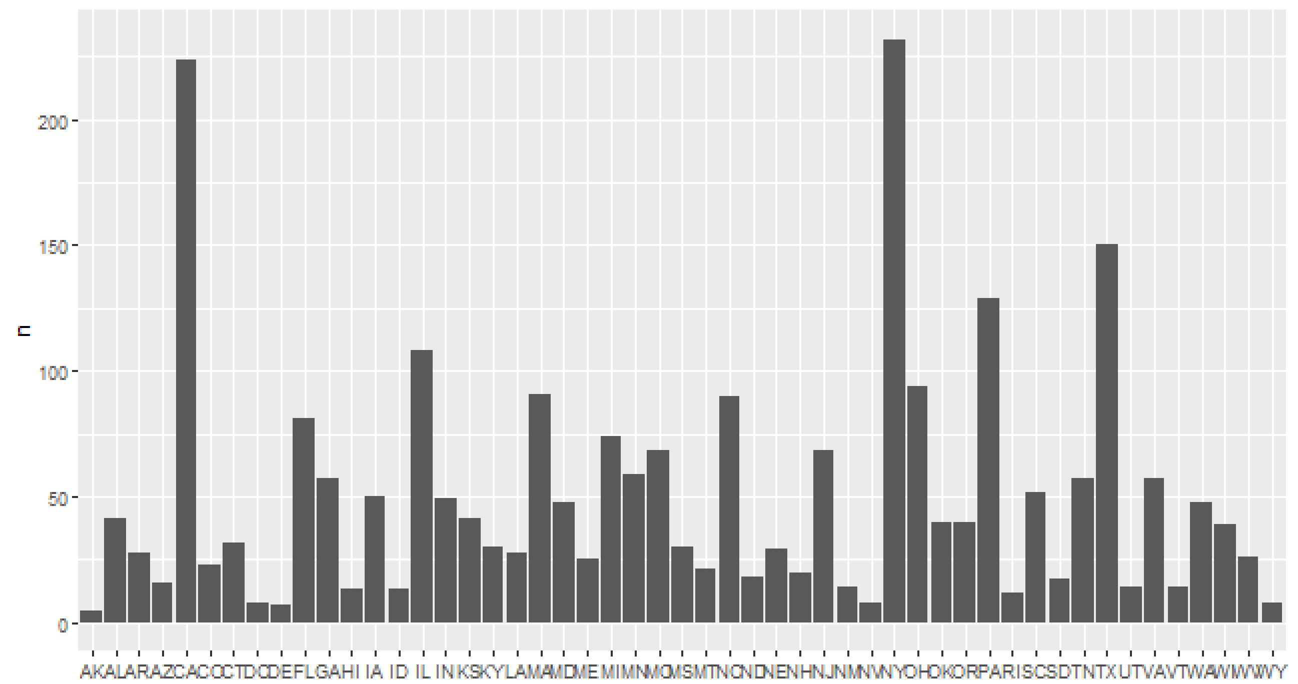
```
df %>%
  count(stabbr) %>%
  arrange(desc(n))
```

```
## # A tibble: 51 × 2
##    stabbr     n
##    <chr>  <int>
##  1 NY       232
##  2 CA       224
##  3 TX       150
##  4 PA       129
##  5 IL       108
##  6 OH        94
##  7 MA        91
##  8 NC        90
##  9 FL        81
## 10 MI        74
## # … with 41 more rows
```

Summarizing data: Plots

- Recall the different approaches to **summarizing data**

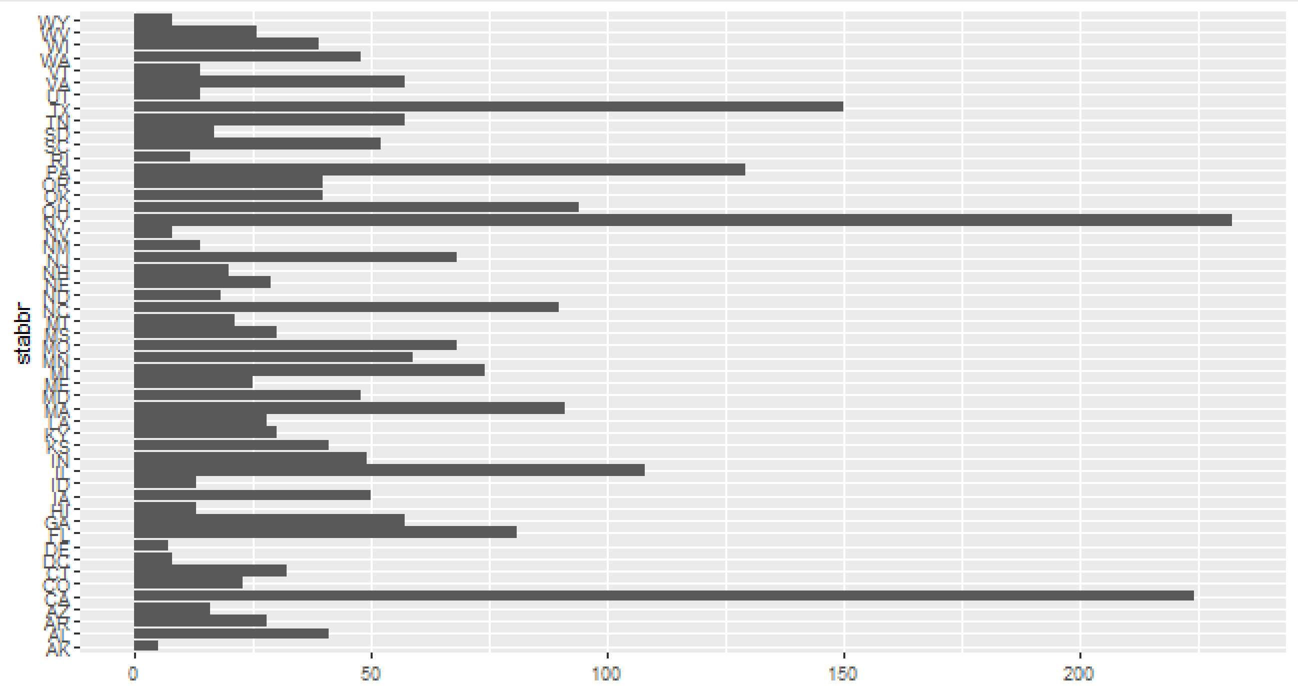
```
df %>%
  count(stabbr) %>%
  ggplot(aes(x = stabbr,y = n)) +
  geom_bar(stat = 'identity')
```



Summarizing data: Plots

- Recall the different approaches to **summarizing data**

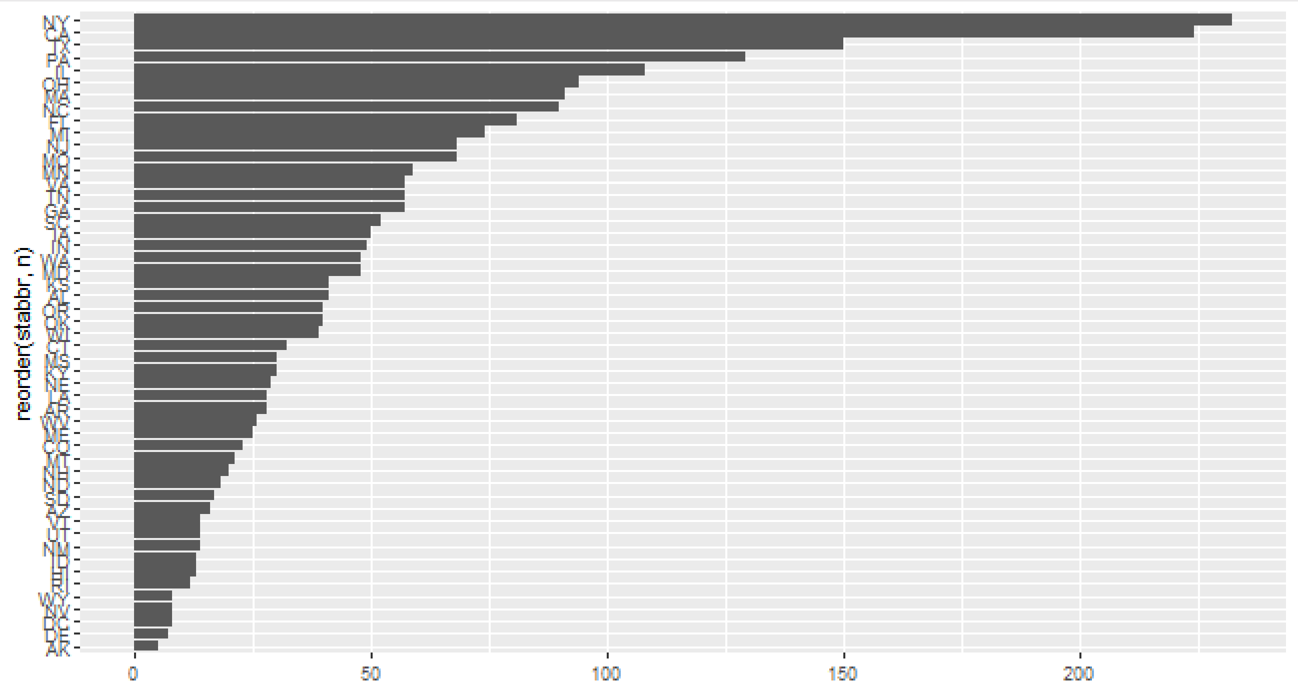
```
df %>%
  count(stabbr) %>%
  ggplot(aes(y = stabbr,x = n)) +
  geom_bar(stat = 'identity')
```



Summarizing data: Plots

- Recall the different approaches to **summarizing data**

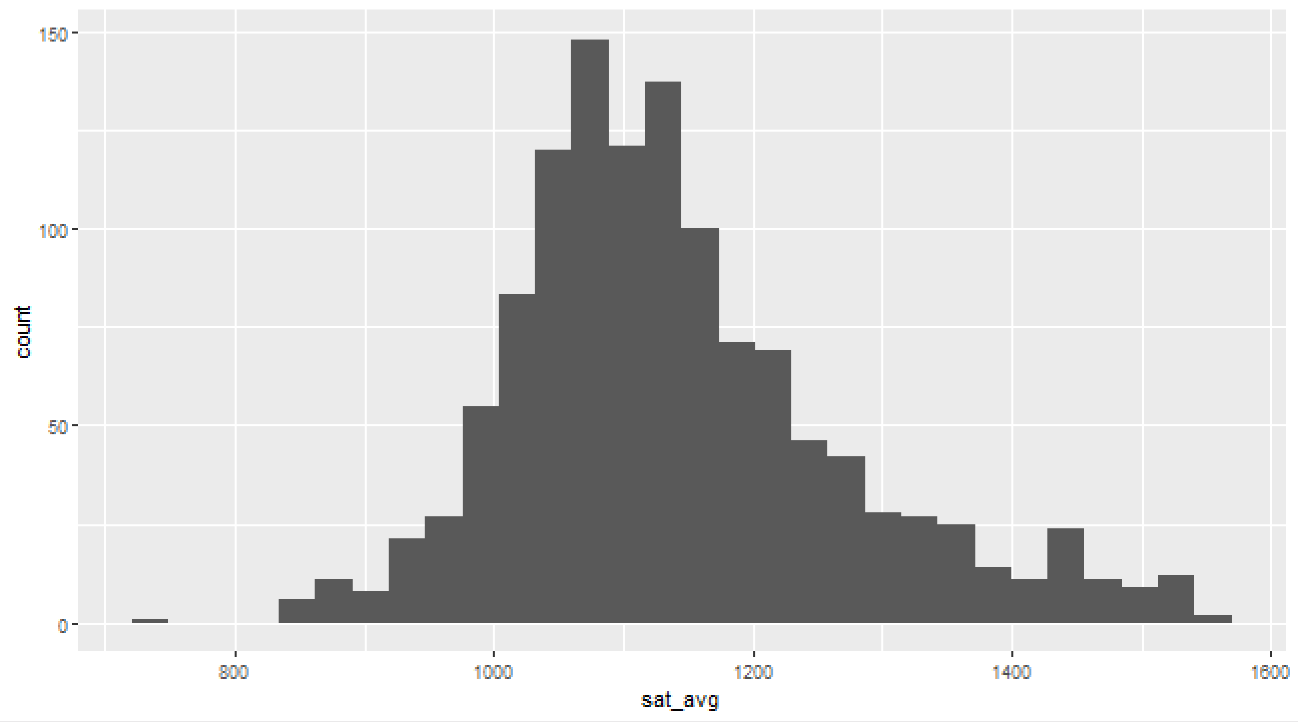
```
df %>%
  count(stabbr) %>%
  ggplot(aes(y = reorder(stabbr,n),x = n)) +
  geom_bar(stat = 'identity')
```



Summarizing Data: Plots

- What about for an interval variable?

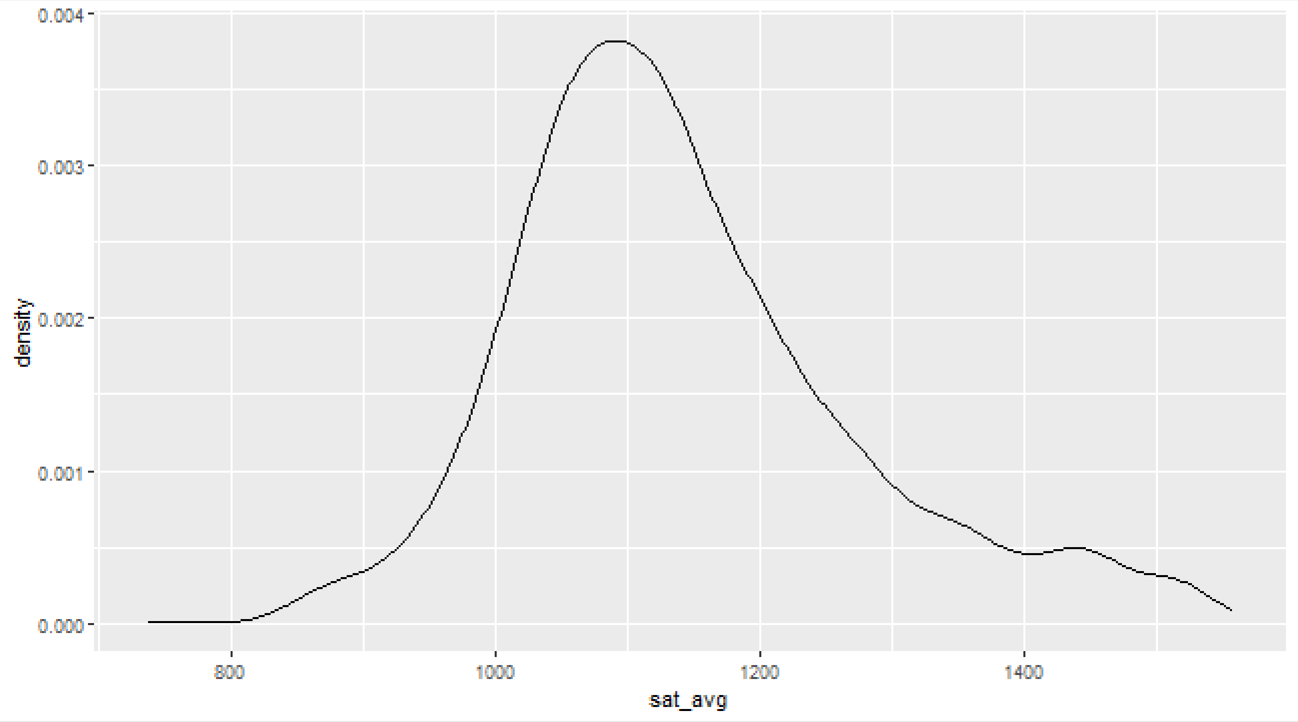
```
df %>%
  ggplot(aes(x = sat_avg)) +
  geom_histogram()
```



Summarizing Data: Plots

- What about for an interval variable?

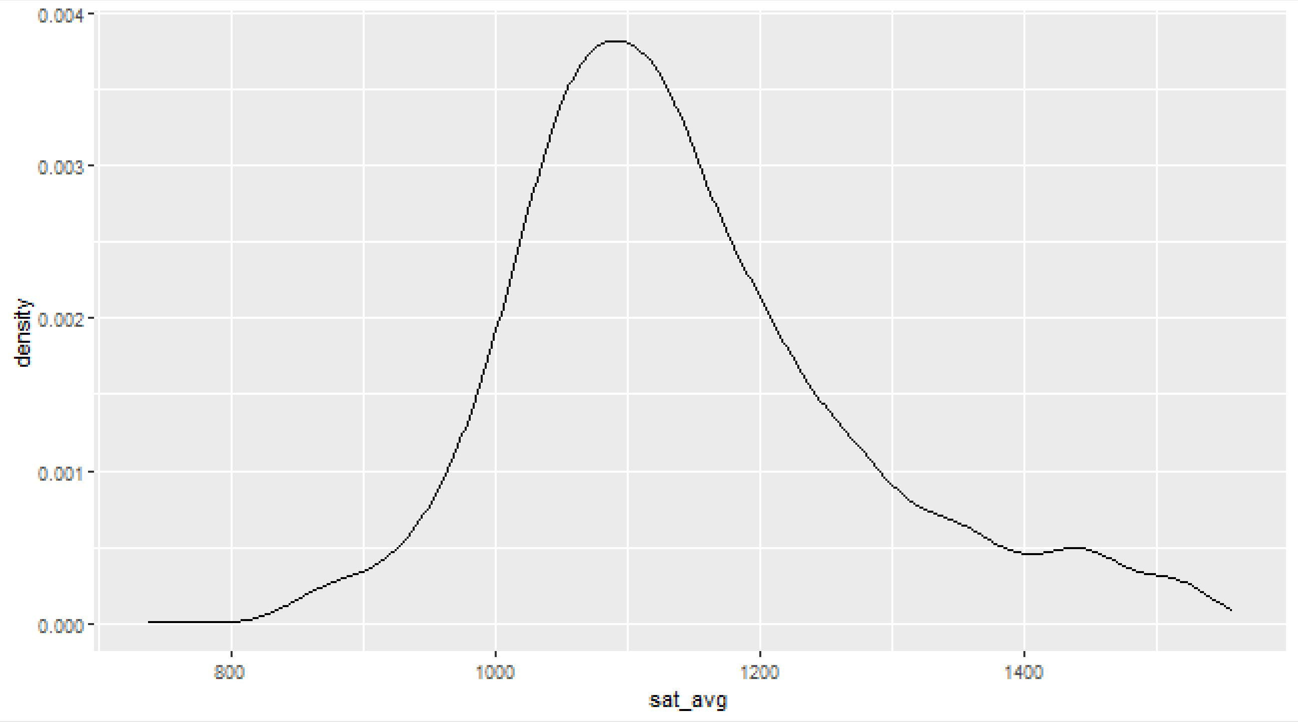
```
df %>%
  ggplot(aes(x = sat_avg)) +
  geom_density()
```



Summarizing Data: Plots

- Describe this qualitatively! Is this skewed? Bimodal?

```
df %>%
  ggplot(aes(x = sat_avg)) +
  geom_density()
```



Summarizing Data

- Recall our two summary statistics of interest
  - **Central Tendency**: The *typical value*
  - **Dispersion**: The *spread*
- What are the various measures for each?

Summarizing Data: Central Tendency

```r
# Mode
df %>%
  count(region) %>%
  filter(n == max(n))
```

```
## # A tibble: 1 × 2
##   region        n
##   <chr>     <int>
## 1 Southeast   577
```

Summarizing Data: Central Tendency

```r
df %>%
  summarise(avg_earnings = mean(md_earn_wne_p6,na.rm=T)) # Mean
```

```
## # A tibble: 1 × 1
##   avg_earnings
##          <dbl>
## 1       33028.
```

```r
df %>%
  summarise(median_sat = median(sat_avg,na.rm=T)) # Median
```

```
## # A tibble: 1 × 1
##   median_sat
##        <int>
## 1       1119
```

- Is this weird to take the median of the average SAT scores??
  - Recall what the **units** are in the data!

Summarizing Data: Dispersion

```r
df %>%
  summarise(range_sat = range(sat_avg,na.rm=T)) # Range
```

```
## # A tibble: 2 × 1
##   range_sat
##       <int>
## 1       737
## 2      1557
```

```r
df %>%
  summarise(iqr_sat = quantile(sat_avg,p = c(.25,.75),na.rm=T)) # IQR
```

```
## # A tibble: 2 × 1
##   iqr_sat
##     <dbl>
## 1    1053
## 2    1205
```

Summarizing Data: Dispersion

```r
df %>%
  summarise(var_sat = var(sat_avg,na.rm=T)) # Variance
```

```
## # A tibble: 1 × 1
##   var_sat
##     <dbl>
## 1  17052.
```

Summarizing Data: Manually

```
df %>%
  select(sat_avg) %>%
  mutate(ybar = mean(sat_avg,na.rm=T)) %>% # Calculate Y bar
  mutate(yi_ybar = sat_avg - ybar) %>% # Calculate diffs
  mutate(yi_ybar2 = yi_ybar^2) %>% # Square diffs
  mutate(yi_yibar2_sum = sum(yi_ybar2,na.rm=T)) %>% # Sum squared diffs
  mutate(N = sum(!is.na(sat_avg))) %>% # Calculate N
  mutate(var_sat = yi_yibar2_sum / (N)) # Calculate variance
```

```
## # A tibble: 2,546 × 7
##    sat_avg  ybar yi_ybar yi_ybar2 yi_yibar2_…¹     N var_sat
##      <int> <dbl>   <dbl>    <dbl>        <dbl> <int>   <dbl>
##  1     939 1141.  -202.    40667.    20939803.  1229  17038.
##  2    1234 1141.    93.3    8712.    20939803.  1229  17038.
##  3      NA 1141.    NA        NA     20939803.  1229  17038.
##  4    1319 1141.   178.    31805.    20939803.  1229  17038.
##  5     946 1141.  -195.    37893.    20939803.  1229  17038.
##  6    1261 1141.   120.    14481.    20939803.  1229  17038.
##  7      NA 1141.    NA        NA     20939803.  1229  17038.
##  8      NA 1141.    NA        NA     20939803.  1229  17038.
##  9    1082 1141.   -58.7    3441.    20939803.  1229  17038.
## 10    1300 1141.   159.    25389.    20939803.  1229  17038.
```

Quiz

- Why is the result produced by the `R` function `var()` difference from my manual attempt?
- Look up the difference between a theoretical measure and an empirical one!