Review: Describing Variables

The Effect Book: Chapter 3

Outline

Read: Chapter 3- Describing Variables

- Descriptions of Variables
- Types of Variables
- The Distribution
- Summarizing the Distribution
- Theoretical Distributions

This Class

- We'll be discussion how we *describe variables* and *describe relationships*
- With a bit of an R reminder
- This class is much more concerned with *How is the data utilized?*

Today

- We'll start with ways of discussing how we can *describe variables*
- And then move on to ways of discussing how we can *describe relationships*
- Secretly, pretty much all statistical analysis is just about doing one of those two things
- *Causal* analysis is *purely* about *knowing exactly which variables and relationships to describe*

Variables

- A statistical variable is a recorded observation, repeated many times
- "Number of calories I ate for breakfast this morning" is one observation
- "The number of calories I ate each breakfast in the past week" is a variable with seven observations

The Distribution of a Variable

- Variables have *distributions*
- The distribution of a variable is simply the description of *how often each value of the variable comes up*
- So for example, the statement "10% of people are left-handed" is just a partial description of the distribution of the handedness variable.
- If you observe a bunch of people and record what their dominant hand is, 10% of the time you'll write down "left-handed," 1% of the time you'll write down "ambidextrous," and 89% of the time you'll write down "right-handed." That's the full description of the distribution

Looking Straight at a Distribution

- The distribution of a variable contains *everything we know* about that variable from empirical observation
- Any description we make will be a *summary* of that distribution
- So we may as well look at it directly!

Distributions of Kinds of Variables

- There are two main kinds of variables for which the distributions look different: discrete and continuous
- Discrete variables take a finite set of values
- Examples: left-handed, right-handed, ambidextrous. Or "lives in Seattle" vs. "Doesn't" or "Number of kids"
- Continuous variables take any value (many many values)
    - Examples: `income`, `height`, `KwH of electricity used each day`
- (Sometimes, "ordinal" discrete variables with many values are treated as continuous for simplicity)

Discrete Distributions

- To fully describe the distribution of a discrete variable, just give the proportion of time it takes each value.
- Give a table with the proportions (or counts), or show a graph with the proportions

Discrete Distributions

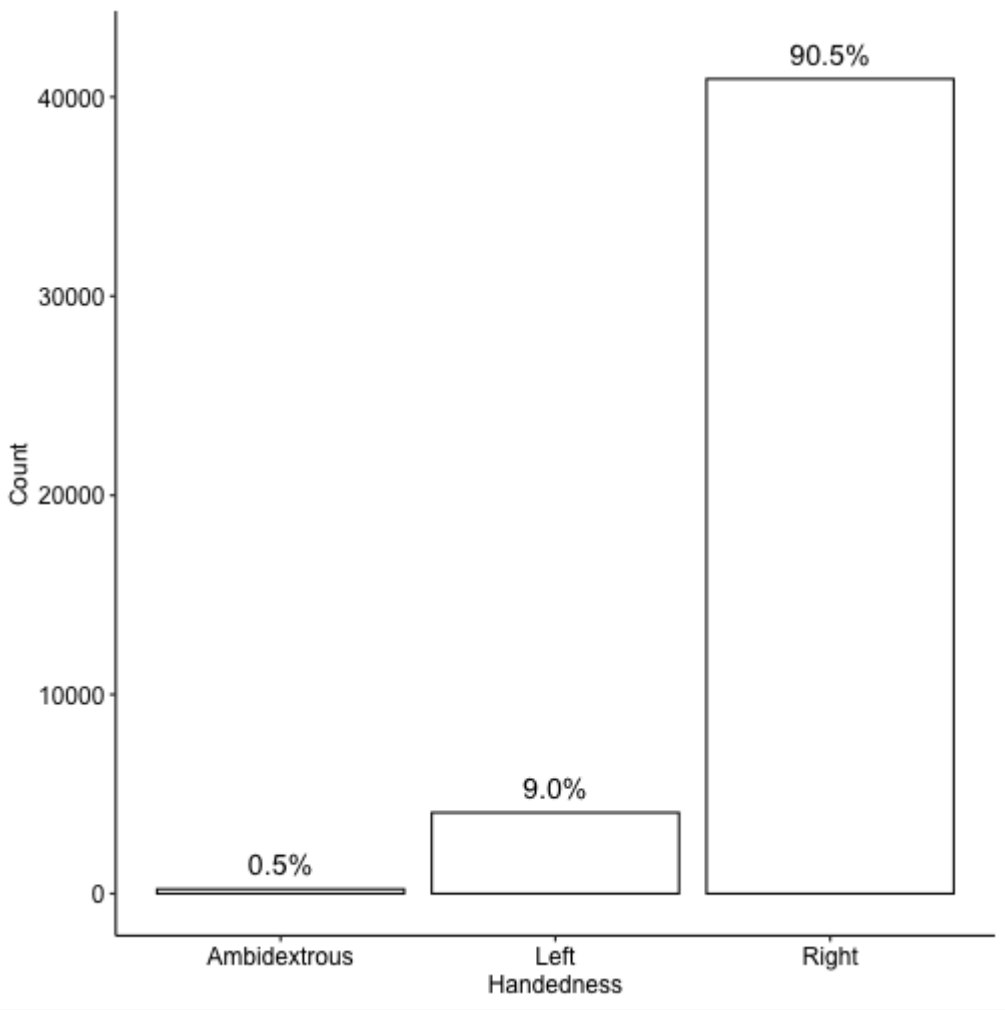
```
handedness_data ← data.frame(hand = c(rep('Left', 9*452),rep('Ambidextrous',.5*452),rep('Right',90.5*452))) #create
library(tidyverse)
handedness_data %>% # dataset
  pull(hand) %>%  # choose the variable
  table() %>%
  prop.table() # creates table with proportions
```

```
## .
## Ambidextrous         Left        Right
##        0.005        0.090        0.905
```

Discrete Distributions

```
ggplot(handedness_data, aes(x = hand)) +
  geom_bar(fill = 'white', color = 'black') + # These two lines are important
  stat_count(geom = "text", size = 5,
             aes(label = scales::percent(..count../nrow(handedness_data)),
                 y = ..count.. + 1300)) +
  ggpubr::theme_pubr() +
  labs(x = 'Handedness', y = 'Count') # The rest is just decoration
```

Discrete Distributions



Using Discrete Distributions

- What can we use a discrete distribution to say?
- X% of observations are in category A
- (X+Y)% of observations are in category (A or B)
- If it's "ordinal" (the values have an order), we can describe the median, max, min, etc.

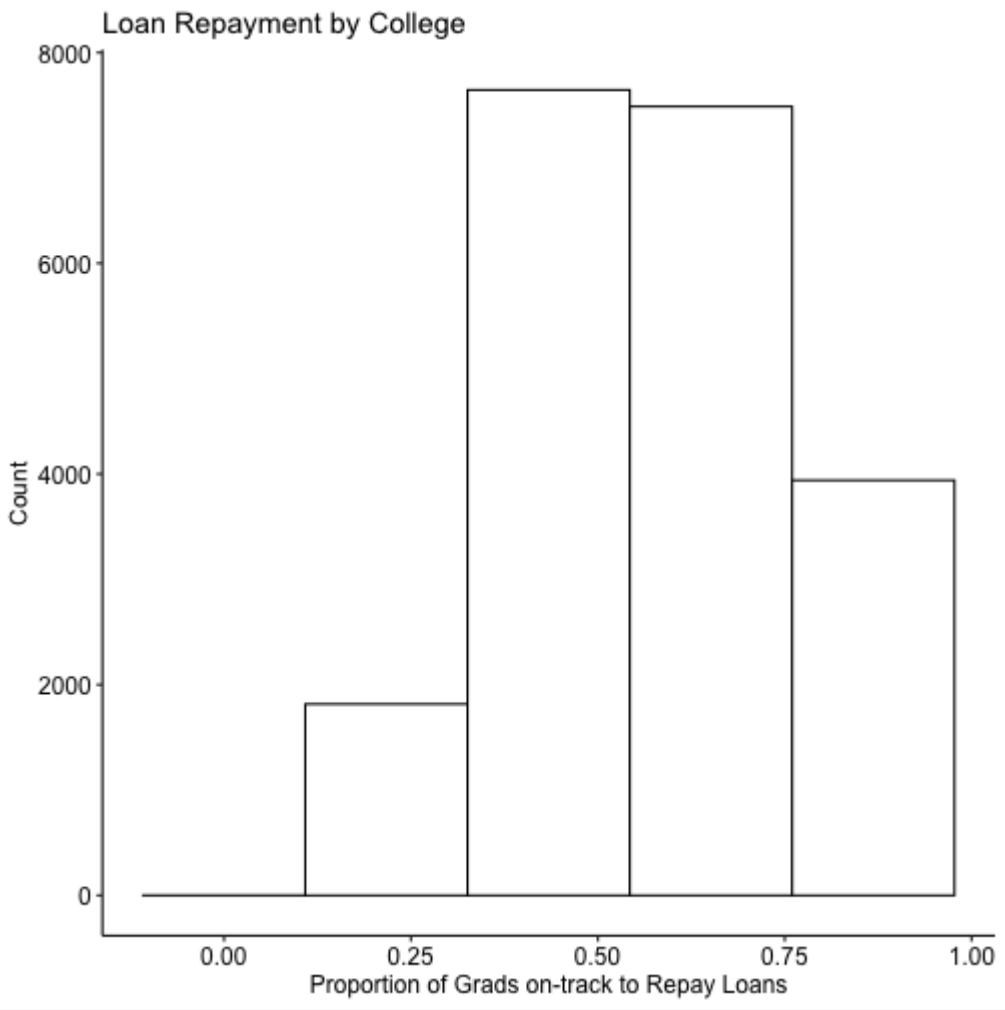
Continuous Distributions

- Variables that are numeric in nature and take *many* values have a continuous distribution
- Their distributions can be presented in one of two main ways - using a *histogram* or using a *density distribution*
- A *histogram* splits the range of the data up into bins and then just treats it like a ordinal discrete distribution
- A *density distribution* uses a rolling average of the proportion of observations within each window

Continuous Distributions

```
#devtools::install_github("NickCH-K/pmdplyr")
library('pmdplyr')
data(Scorecard, package ="pmdplyr")
ggplot(Scorecard, aes(x = repay_rate)) +
  geom_histogram(bins = 5, fill = 'white', color = 'black') +
  ggpubr::theme_pubr() +
  labs(x = 'Proportion of Grads on-track to Repay Loans', y = 'Count', title = 'Loan Repayment by College')
```

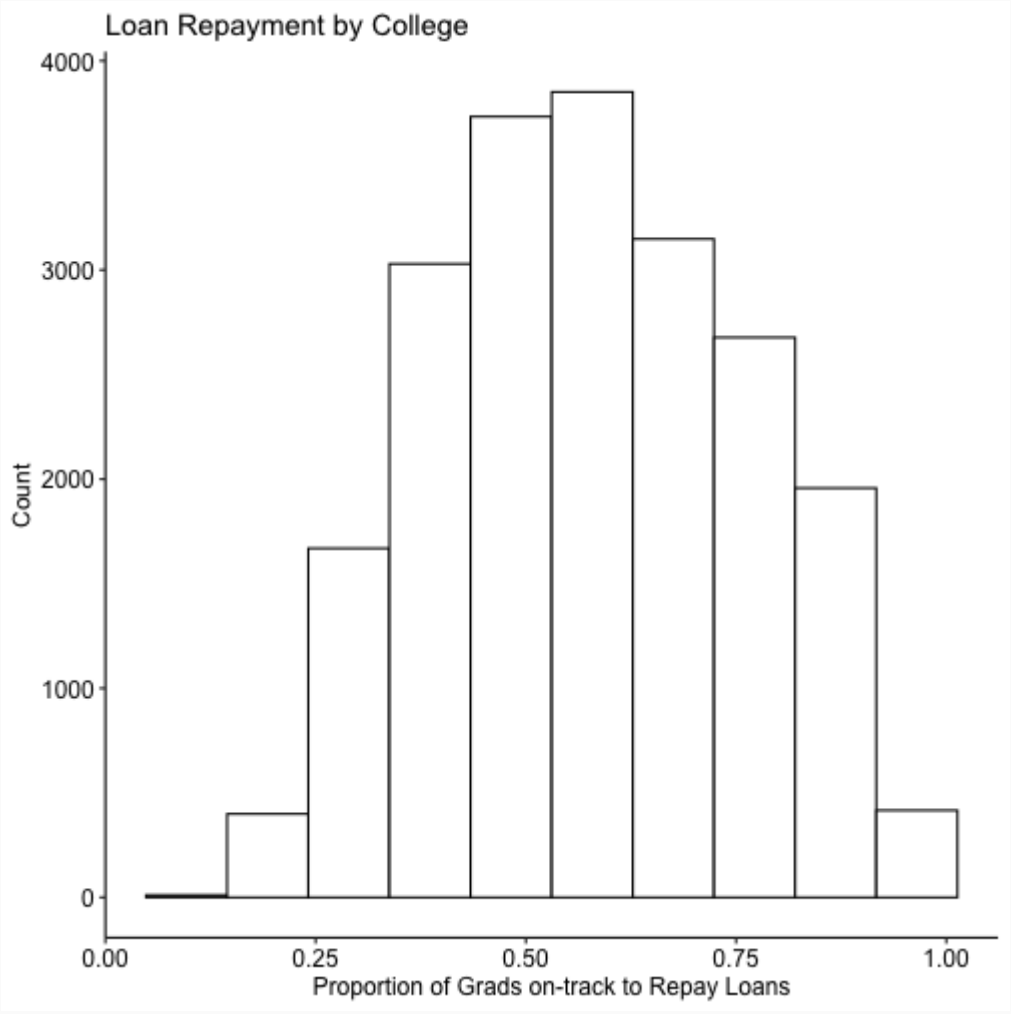
Continuous Distributions



Continuous Distributions

```
ggplot(Scorecard, aes(x = repay_rate)) +
 geom_histogram(bins = 10, fill = 'white', color = 'black') +
 ggpubr::theme_pubr() +
 labs(x = 'Proportion of Grads on-track to Repay Loans', y = 'Count', title = 'Loan Repayment by College')
```

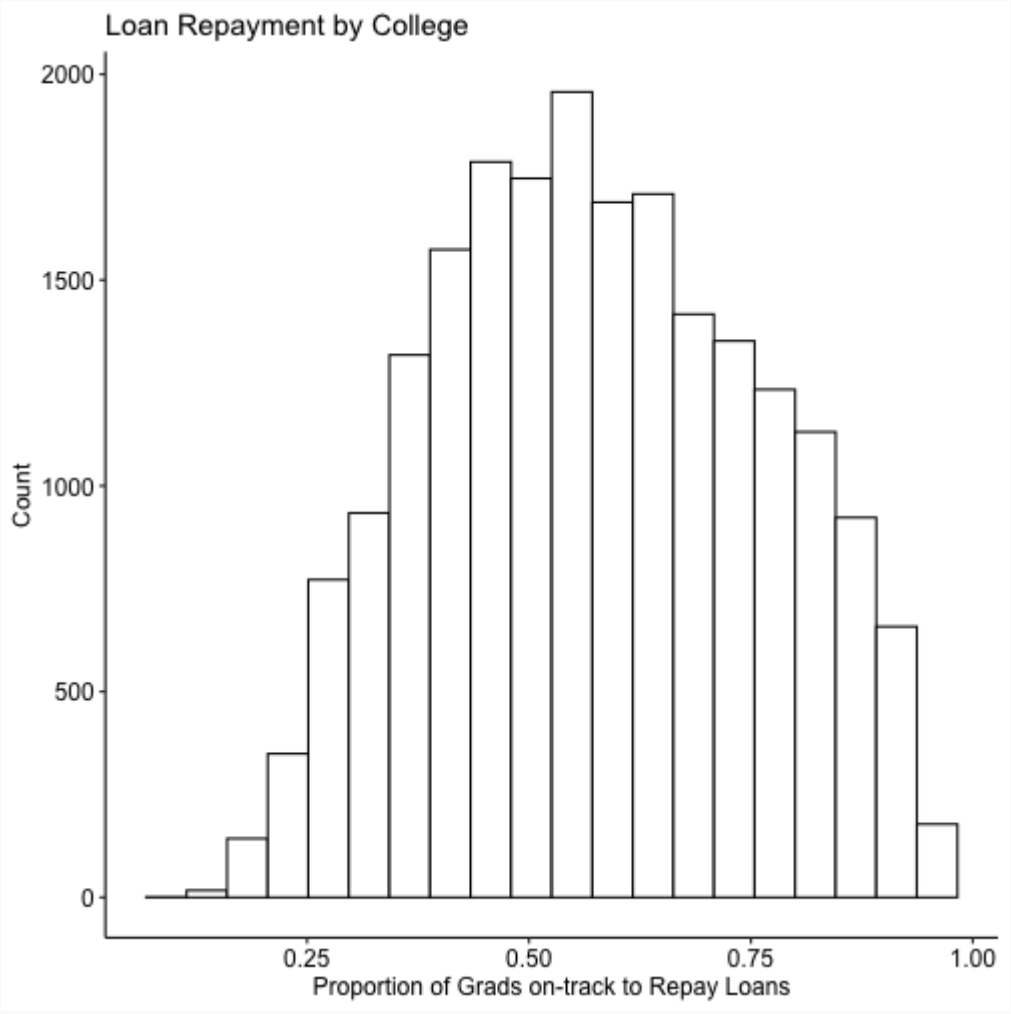
Continuous Distributions



Continuous Distributions

```
ggplot(Scorecard, aes(x = repay_rate)) +
  geom_histogram(bins = 20, fill = 'white', color = 'black') +
  ggpubr::theme_pubr() +
  labs(x = 'Proportion of Grads on-track to Repay Loans', y = 'Count', title = 'Loan Repayment by College')
```

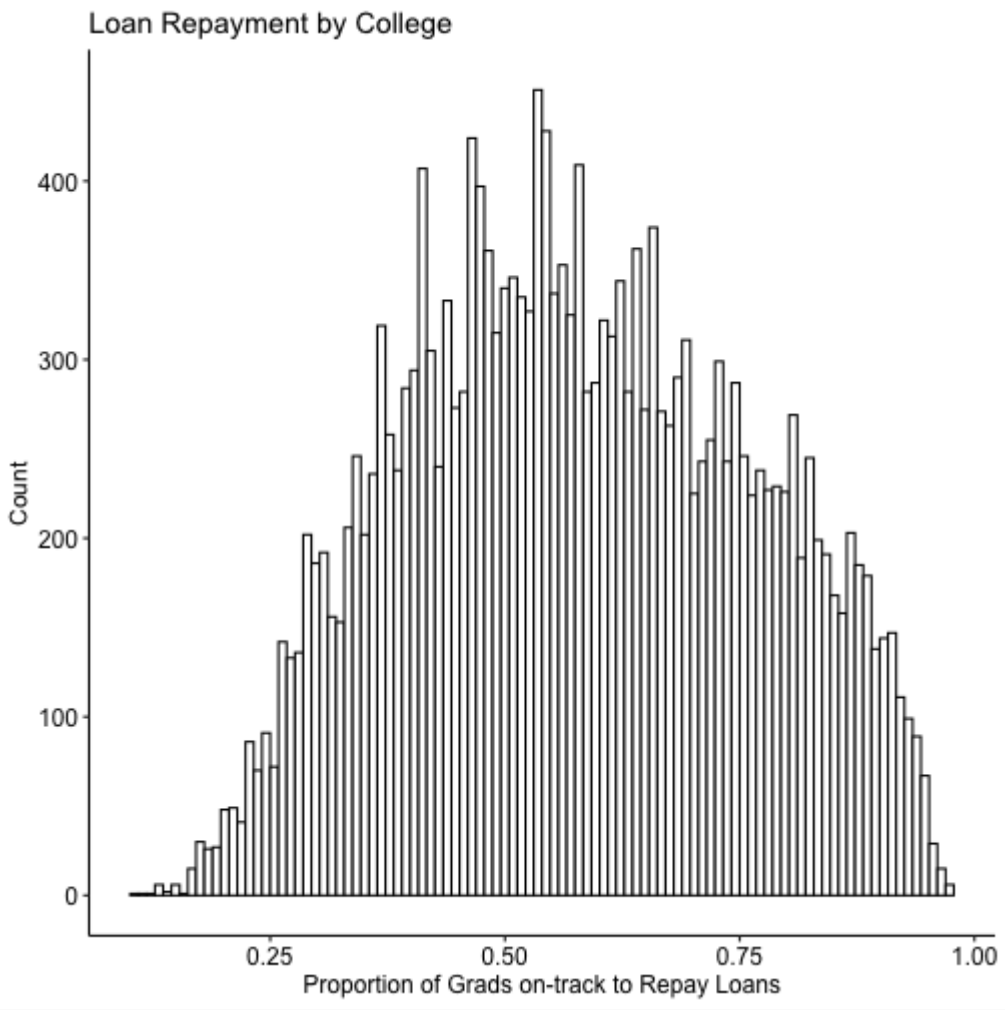
Continuous Distributions



Continuous Distributions

```
ggplot(Scorecard, aes(x = repay_rate)) +
  geom_histogram(bins = 100, fill = 'white', color = 'black') +
  ggpubr::theme_pubr() +
  labs(x = 'Proportion of Grads on-track to Repay Loans', y = 'Count', title = 'Loan Repayment by College')
```

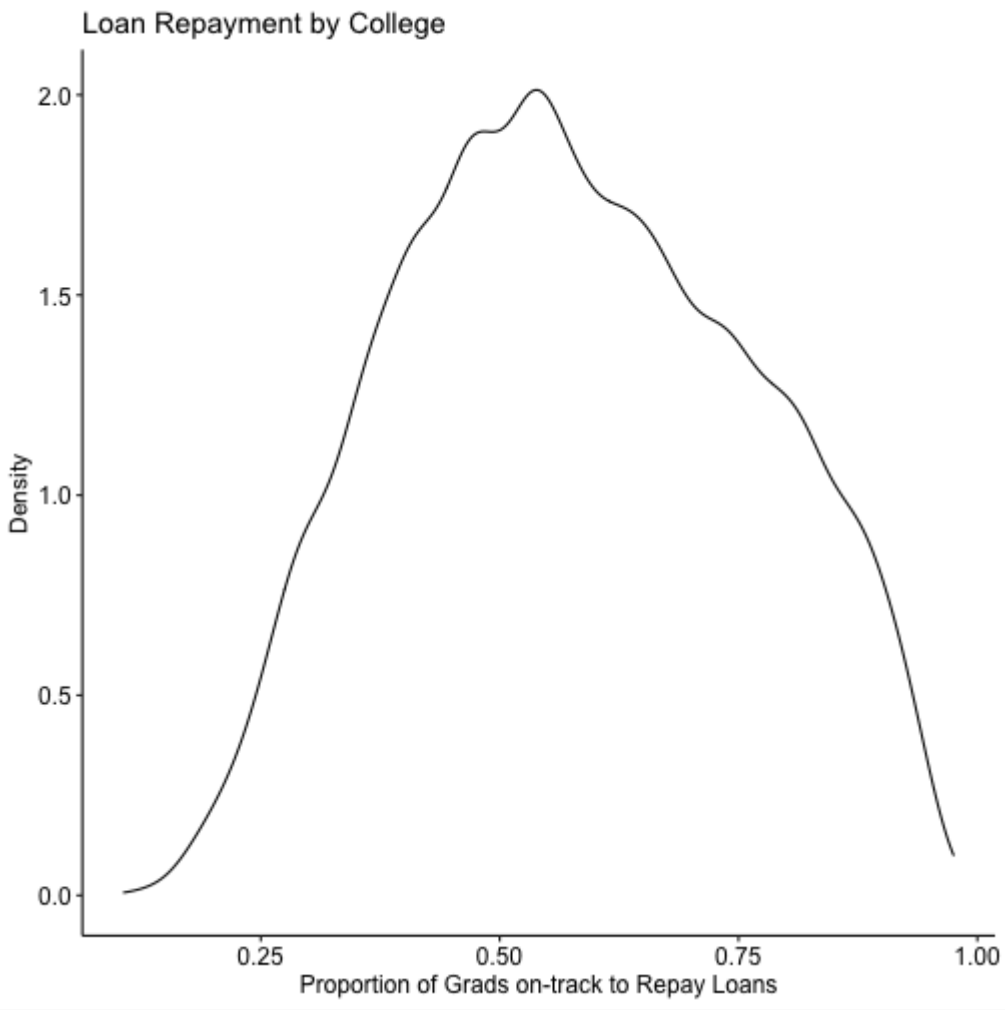
Continuous Distributions



Continuous Distributions

```
ggplot(Scorecard, aes(x = repay_rate)) +
  geom_density(color = 'black') +
  ggpubr::theme_pubr() +
  labs(x = 'Proportion of Grads on-track to Repay Loans', y = 'Density', title = 'Loan Repayment by College')
```

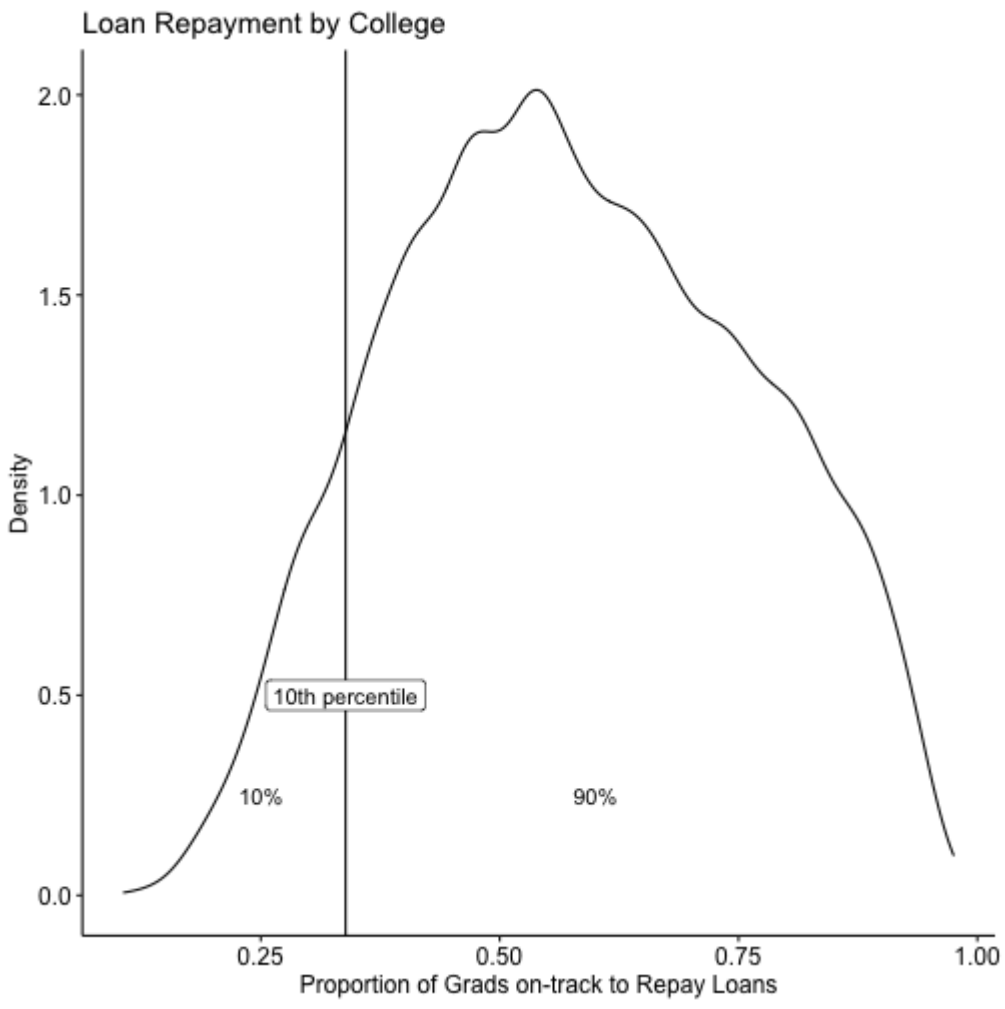
Continuous Distributions



Continuous Distributions

- We can describe these distributions fully using *percentiles*
- The X^{th} percentile is the value for which X% of the sample is less than that value
- Taken together, you can describe the entire sample by just going through percentiles

Continuous Distributions



Summarizing Continuous Data

- Commonly we want to describe these distributions much more compactly, while still telling us something about them

```
Scorecard %>%
  select(repay_rate) %>%
  sumtable()
```

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
repay_rate	20890	0.576	0.182	0.107	0.437	0.718	0.975

Summarizing Continuous Data

- Every "summary statistic" of a given variable is just a way of describing some aspect of these distributions
- Commonly we are focused on just a few important features of the distribution:
  - The central tendency
  - Dispersion

The Central Tendency

- Central tendencies are ways of picking a single number that represents the variable best
- Often: the mean
  - The median (50th percentile)
  - For categorical data, sometimes the mode

The Central Tendency

- The median is good at being representative of *a typical observation*, and is not sensitive to outliers
- The mean can be better thought of as a betting average. If you "bet the mean" and drew an infinite number of observations, you'd break even
- If Jeff Bezos walks in the room, mean income shoots through the roof (because if you're randomly drawing people in the room, sometimes you're Jeff Bezos!), but the median largely remains unchanged (because Jeff Bezos isn't anywhere near being a typical person)

The Central Tendency

- So why use the mean at all? It makes sense to think about those betting odds if you are, say, trying to predict something
- It also has a bunch of nice statistical properties
- Meaning, we *understand the mean* fairly well, and *we know how the mean changes as we go from sample to sample*
- In other words, it's handy when we're trying to learn about the *theoretical distribution* our data comes from (more on that in a bit!)

Dispersion

- Measures of dispersion tell us how *spread out* the data is
- Some of these are percentile-based measures, like the inter-quartile range (75th percentile minus 25th) or the range (Max - Min, or 100th percentile minus 0th)
- Most commonly we will use standard deviation and variance

```
library(vtable)
Scorecard %>%
  select(repay_rate) %>%
  sumtable(out = 'kable')
```

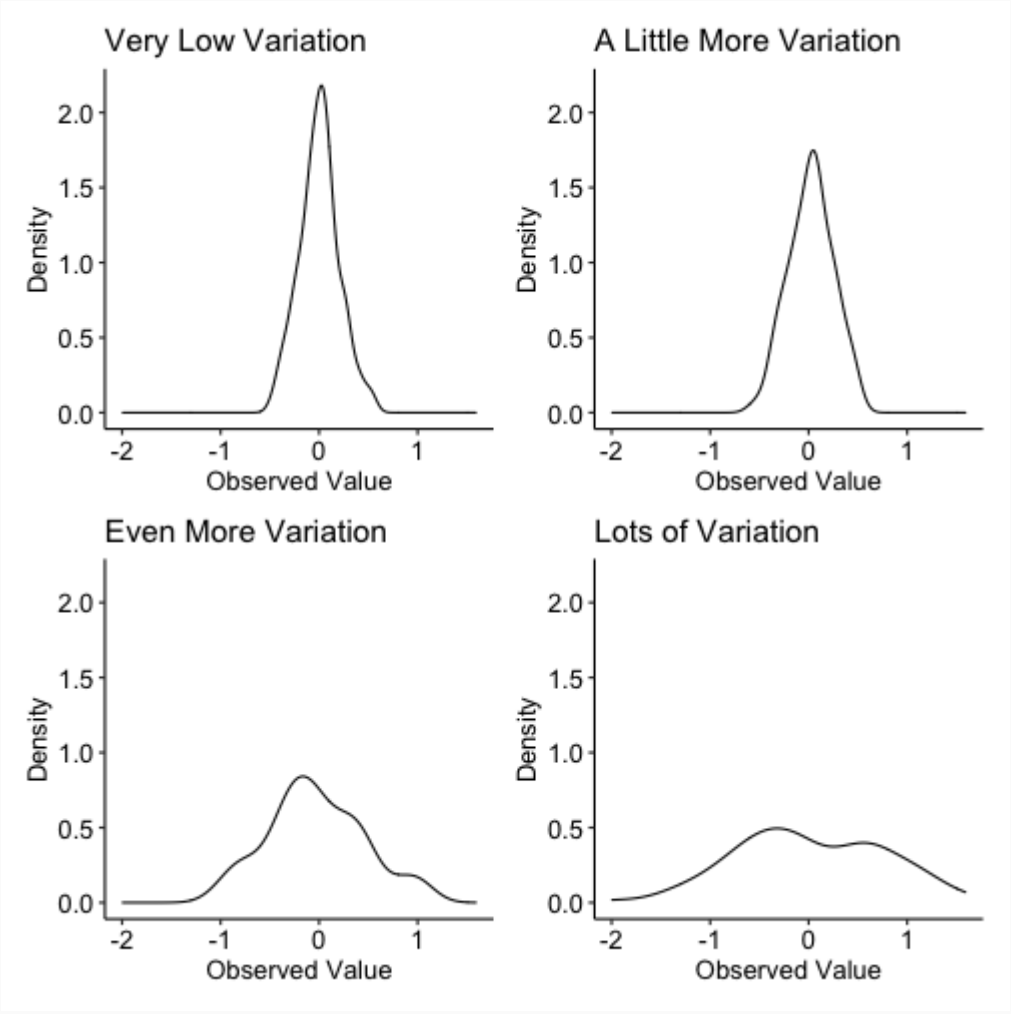
Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
repay_rate	20890	0.576	0.182	0.107	0.437	0.718	0.975

Dispersion

- Variance is *average squared deviation from the mean*.
- Take each observation, subtract the mean, square the result, and take the mean of *that* (times $n/(n-1)$)
- Standard deviation is the square root of the variance

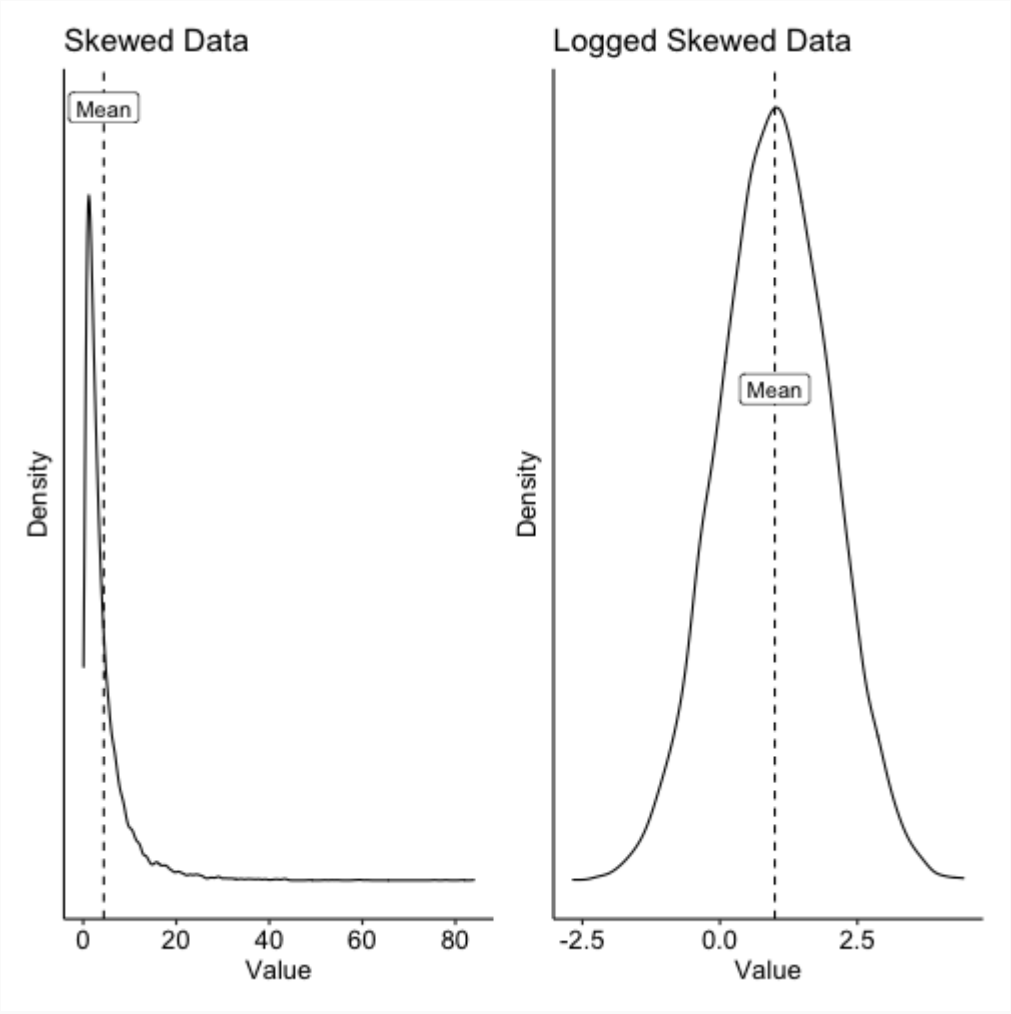
Dispersion


Very Low Variation
A Little More Variation
Even More Variation
Lots of Variation
Density
Observed Value
0.0
0.5
1.0
1.5
2.0
-2
-1
0
1

Skew

- One other aspect of a distribution we sometimes consider is *skew*
- Skew is sort of "how much the distribution *leans to one side or the other*
- This can be a problem if the skew is extreme
- Extreme right skew can make means highly unrepresentative as a few big observations pull the mean upwards
- This can sometimes be helped by taking a log of the data

Skew


Skewed Data
Mean
Density
0
20
40
60
80
Value
Logged Skewed Data
Mean
Density
-2.5
0.0
2.5
Value

Theoretical Distributions

- Now for the good stuff!
- We *rarely actually care what our data is* or the distribution of it!
- What we actually care about are *what broader inferences can we draw from the data we see!*
- The mean of your variable is just *the mean of the observations you happened to sample*
- But what can we learn about *how that variable works overall* from that?

Theoretical Distributions

- There is a "population distribution" that we cannot see - it's theoretical - we just get a sample
- If we had infinite data, we would see the theoretical distribution
- To learn about that theoretical distribution, we take what we know about *sampling variation* and use it to rule out certain theoretical distributions as unlikely

Theoretical Distributions

- For example, if we flip a coin 1000 times and get heads 999 times, can we rule out that it's a fair coin?
- (the "theoretical distribution" here is a discrete one: the coin is heads 50% of the time and tails 50%)
- We *assume that the coin is fair* (null/prior hypothesis) and see how unlikely the data is. If the coin is fair, we take what we know about sampling variation for a binary variable and calculate that 999/1000 heads has a `dbinom(999, 1000, .5)` chance of happening
- So that's probably not the real theoretical distribution!

Theoretical Distributions

Reminders:

- All we've shown is that *a particular theoretical distribution is unlikely*, not *anything else*
- We don't know what the proper theoretical distribution *is*
- We haven't shown that our result is *important*
- We have effectively calculated a `p-value` here - if it's low enough, we say "Statistically significant" but please don't get fooled into thinking that means anything other than what we have said here - a particular theoretical distribution is statistically unlikely to have generated this data

Sampling Variation

- Often when trying to generalize from a sample to a theoretical distribution we will focus on the mean
- This is because the sampling variation of the mean is very well-understood. It follows a normal distribution with a mean at the population mean, and a standard deviation of the population standard deviation, scaled by $1/\sqrt{n}$

Statistical Inference

- By using *what we know about sampling variation*, we can *make inferences* about a variable's theoretical distribution (i.e. what mean that distribution is likely to have)
- In this way we can use what we have - data - to learn about what we actually care about - population distributions
- We have to leverage what we know, and what we have, to make that theoretical inference that we really care about
- This will echo very strongly once we start talking about causality!

Next Time

- Not just single variables, but relationships!
- What are those *population relationships?*