Uncertainty

How confident are we?

Prof. Bisbee

Vanderbilt University

Slides Updated: 2024-01-09

Agenda

1. Uncertainty
2. More NBA data
3. Bootstrap Sampling

The Missing Ingrediant

- Thus far we have:

  1. Tested whether **selective** schools have **higher SAT scores**: Yes
  2. Tested Trump's theory that **polls were biased against him**: No
  3. Tested whether RDD polls **contact more Trump supporters**: No
  4. Tested whether state polls **accurately predicted the president**: No

- We want to do more than say "Yes" or "No" when answering a Research Question or making a Prediction

- We want to express our **confidence**

What is "confidence"?

- In frequentist statistics:
    - How often your conclusion would be correct if you were able to run an "experiment" many times
    - How often your conclusion would be correct if you were able to observe the world many times
- Research Question: Are NBA players in their rookie season more prone to turnovers?
    - Theory: ??
    - Hypothesis: ??
- Analysis: compare `tov` by `isRookie`

NBA Example

```r
require(tidyverse)
nba <-
read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/nba_pl
glimpse(nba %>% select(tov,isRookie))
```

```
## Rows: 530
## Columns: 2
## $ tov      <dbl> 144, 4, 135, 14, 121, 8, 33, 6, 28, 2, 72…
## $ isRookie <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, …
```

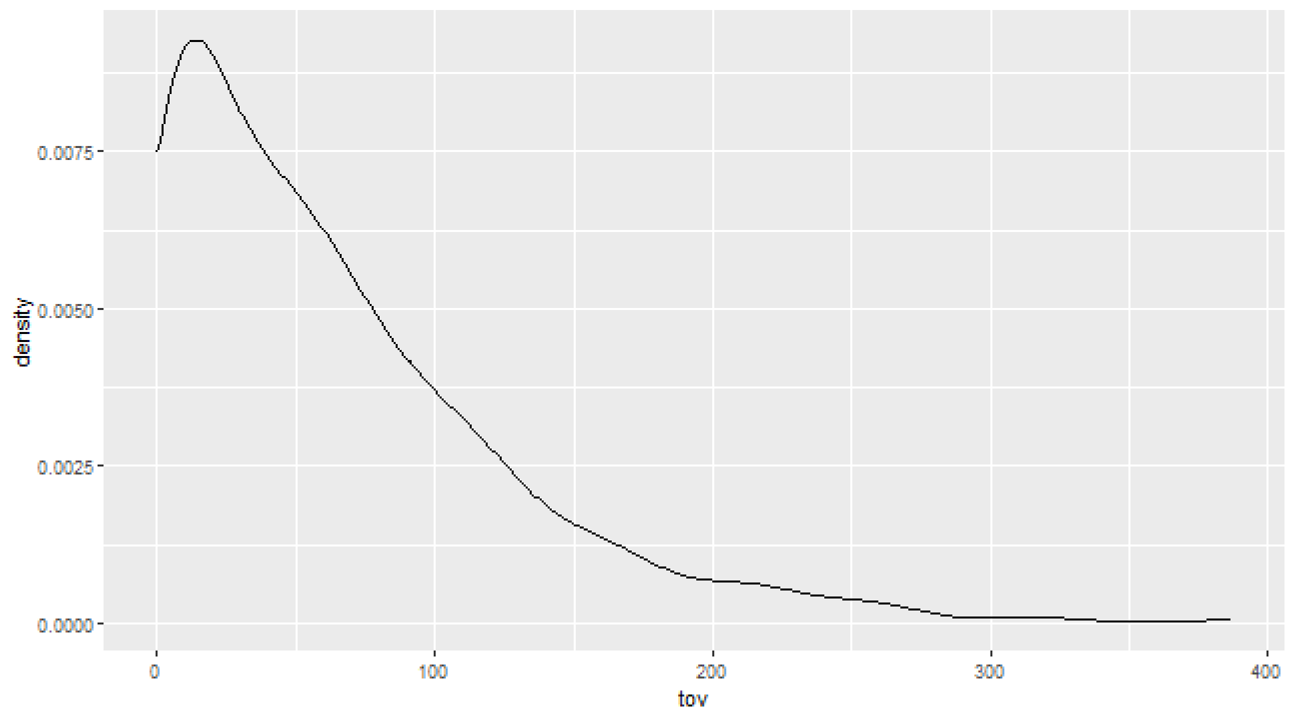
Look

```
summary(nba %>% select(tov,isRookie))
```

```
##       tov            isRookie      
##  Min.   :  0.00   Mode :logical  
##  1st Qu.: 14.25   FALSE:425      
##  Median : 47.00   TRUE :105      
##  Mean   : 62.82                  
##  3rd Qu.: 91.75                  
##  Max.   :387.00                  
```

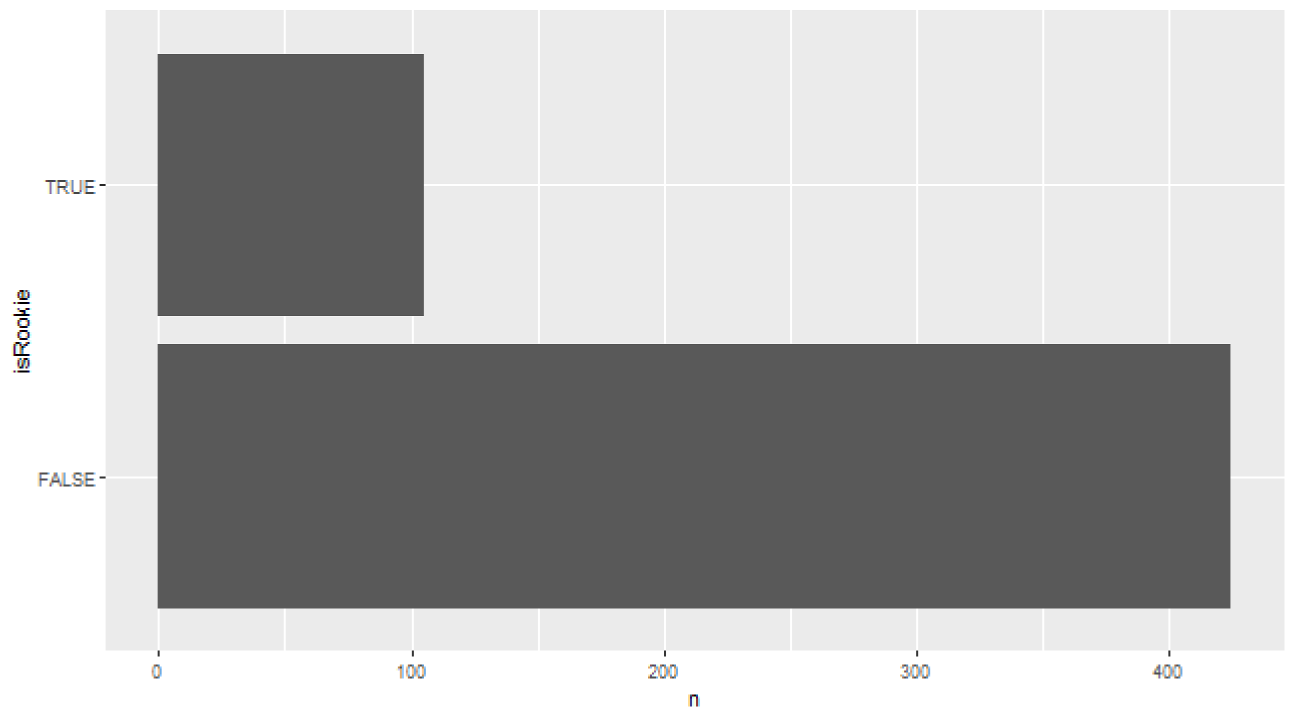
Visualize: Univariate Y

```
nba %>%
  ggplot(aes(x = tov)) +
  geom_density()
```



Visualize: Univariate X

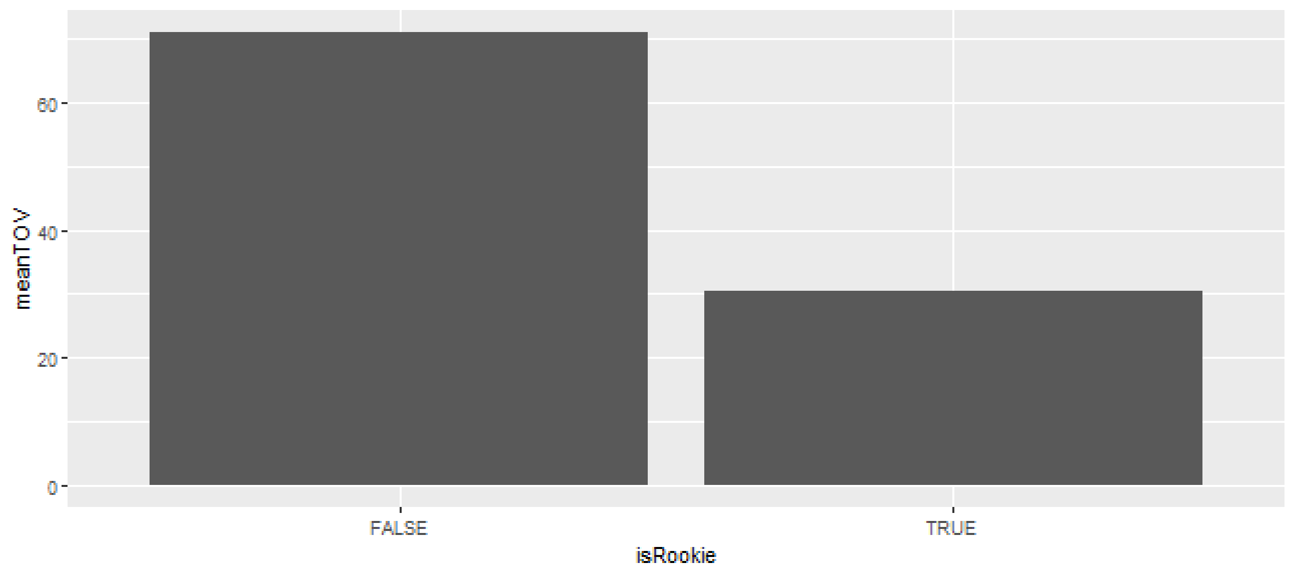
```
nba %>%
  count(isRookie) %>%
  ggplot(aes(x = n,y = isRookie)) +
  geom_bar(stat = 'identity')
```



Visualize: Multivariate

- Option #1: `summarise()` data prior to plotting

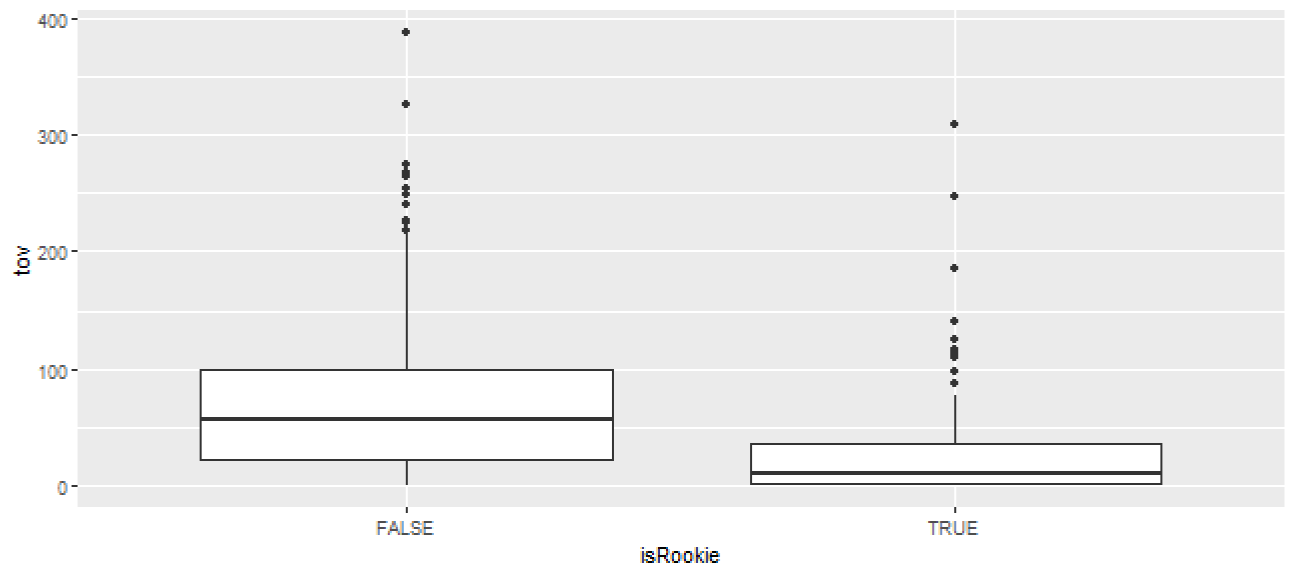
```r
nba %>%
  group_by(isRookie) %>%
  summarise(meanTOV = mean(tov,na.rm=T)) %>%
  ggplot(aes(x = isRookie,y = meanTOV)) +
  geom_bar(stat = 'identity')
```



Visualize: Multivariate

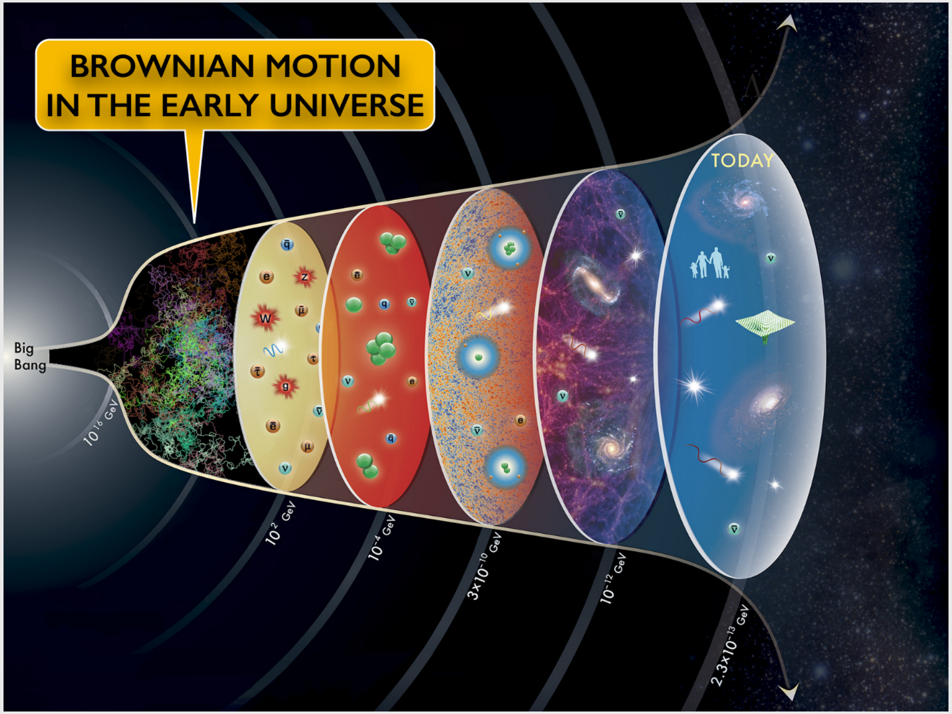
- Option #2: plot raw data

```
nba %>%
  ggplot(aes(x = isRookie,y = tov)) +
  geom_boxplot()
```



Uncertainty

- Are rookies **better** than more senior players?
- Big philosophical step back
    - We live in a stochastic universe!



Uncertainty

- Are rookies **better** than more senior players?
- Populations versus samples
    - Intro stats: uncertainty due to **sample**

Uncertainty

- Big philosophical step back
  - We live in a stochastic universe!
- What does **better** mean?
  - Theory: An innate quality in greater abundance
  - Prediction: If we had to bet on who turns over the ball less, who do we choose?
- How **confident** would we be with this bet?

Uncertainty

- If the universe is inherently stochastic, we are inherently uncertain
  - We THINK rookies are more careful passers, but not 100% certain
- How to measure this?
  - Run 100 experimental seasons
  - Record turnovers for rookies and non-rookies for each season
  - Calculate how many times rookies turned the ball over less than non-rookies
- 90 seasons out of 100 → 90% confident / certainty
- 100 seasons out of 100 → 100%?
- **FUNDAMENTAL STOCHASTIC NATURE OF REALITY (FSNoR)**

Uncertainty

- Running 100 experimental seasons is impossible

    1. We are not Adam Silver
    2. Even if we were Adam Silver, 100 seasons = a century of basketball!

Uncertainty

- Running 100 experimental seasons is impossible

  1. We are not Adam Silver
  2. Even if we were Adam Silver, 100 seasons = a century of basketball!
  3. If we were God? 100 seasons with the same players?

- *STILL wouldn't be 100% certain due to* ***FSNoR***

  - (**F**undamental **S**tochastic **N**ature **o**f **R**eality)

Uncertainty

- But we are data scientists
- Take 1 season of basketball but sample it randomly
- **Bootstrap sampling**
- Theory: By mimicking the sampling process, we can simulate a God experiment
    - (NB: this goes much deeper. Uncertainty from bootstrap combines FSNoR + sampling uncertainty.)
- Practice: `sample_n()` + `for()` loops

Bootstrap Demo Step 1

- One randomly sampled player via `sample_n(size,replace)`
  - `size`: how many samples (from 1 to all observations)
  - `replace`: whether to put the sample back (`TRUE` or `FALSE`)

```r
set.seed(123) # Ensure we can reproduce results exactly

nba %>%
  sample_n(size = 1,replace = T) %>%
  select(namePlayer,slugSeason,isRookie,tov)
```

```
## # A tibble: 1 × 4
##   namePlayer    slugSeason isRookie   tov
##   <chr>         <chr>      <lgl>    <dbl>
## 1 Moritz Wagner 2018-19    TRUE        39
```

Bootstrap Demo Step 2

- Two randomly sampled players

```
set.seed(123)
nba %>%
  sample_n(size = 1,replace = T) %>%
select(namePlayer,slugSeason,isRookie,tov)
```

```
## # A tibble: 1 × 4
##   namePlayer     slugSeason isRookie   tov
##   <chr>          <chr>      <lgl>    <dbl>
## 1 Moritz Wagner 2018-19    TRUE        39
```

```
nba %>%
  sample_n(size = 1,replace = T) %>%
select(namePlayer,slugSeason,isRookie,tov)
```

```
## # A tibble: 1 × 4
##   namePlayer slugSeason isRookie   tov
##   <chr>      <chr>      <lgl>    <dbl>
## 1 Sam Dekker 2018-19    FALSE       24
```

Bootstrap Demo Step 2

- OR two randomly sampled players

```
set.seed(123)

nba %>%
  sample_n(size = 2,replace = T) %>%
select(namePlayer,slugSeason,isRookie,tov)
```

```
## # A tibble: 2 × 4
##   namePlayer    slugSeason isRookie   tov
##   <chr>         <chr>      <lgl>    <dbl>
## 1 Moritz Wagner 2018-19    TRUE        39
## 2 Sam Dekker    2018-19    FALSE       24
```

Bootstrap Demo Step 3

- Randomly sample all players: `size = nrow(nba)` (or `nrow(.)`)

```
set.seed(123)

nba %>%
  sample_n(size = nrow(nba),replace = T) %>% # Same as nrow(.)
  select(namePlayer,slugSeason,isRookie,tov)
```

```
## # A tibble: 530 × 4
##    namePlayer        slugSeason isRookie   tov
##    <chr>             <chr>      <lgl>    <dbl>
##  1 Moritz Wagner     2018-19    TRUE        39
##  2 Sam Dekker        2018-19    FALSE       24
##  3 Joe Harris        2018-19    FALSE      121
##  4 Jonas Valanciunas 2018-19    FALSE       90
##  5 John Holland      2018-19    FALSE        0
##  6 Angel Delgado     2018-19    TRUE         0
##  7 Donovan Mitchell  2018-19    FALSE      218
##  8 Damian Jones      2018-19    FALSE       16
##  9 Luke Kornet       2018-19    FALSE       25
## 10 Justin Anderson   2018-19    FALSE       23
## # ℹ 520 more rows
```

Bootstrap Demo Step 4

- Linking to **confidence**: Do we draw the same conclusion twice?

```
set.seed(123)

# Bootstrapped Season #1
bsSeason1 <- nba %>%
  sample_n(size = nrow(.),replace = T) %>%
  select(isRookie,tov) %>%
  mutate(bsSeason = 1)

# Bootstrapped Season #2
bsSeason2 <- nba %>%
  sample_n(size = nrow(.),replace = T) %>%
  select(isRookie,tov) %>%
  mutate(bsSeason = 2)
```

Bootstrap Demo Step 4

- Linking to **confidence**: Do we draw the same conclusion twice?

```
bsSeason1 %>%
  group_by(isRookie) %>%
  summarise(mean_tov = mean(tov))
```

```
## # A tibble: 2 × 2
##   isRookie mean_tov
##   <lgl>       <dbl>
## 1 FALSE        68.6
## 2 TRUE         36.9
```

```
bsSeason2 %>%
  group_by(isRookie) %>%
  summarise(mean_tov = mean(tov))
```

```
## # A tibble: 2 × 2
##   isRookie mean_tov
##   <lgl>       <dbl>
## 1 FALSE        65.6
## 2 TRUE         28.5
```

Bootstrap Demo Step 5

- Want to do this 100 times!
- Use a `for()` loop to make it cleaner
- A `for()` loop repeats the same code multiple times
    - Benefit: don't need to copy and paste a chunk of code 100 times
    - Just put a chunk of code in a loop that repeats 100 times!

```r
set.seed(123) # Ensure you'll get the same results each time
bsSeasons <- NULL # Instantiate empty object
for(bsSeason in 1:100) { # Repeat 100 times
  tmpSeason <- nba %>%
    sample_n(size = nrow(.),replace = T) %>% # Sample the data
    select(isRookie,tov) %>% # Select variables of interest
    mutate(bsSeasonNumber = bsSeason) # Save the simulation ID
  bsSeasons <- bind_rows(bsSeasons,tmpSeason) # Append to the empty
object!
}
```

Bootstrap to measure Confidence

- Compare rookie versus non-rookie turnovers each season

```
bsSeasons %>%
  group_by(bsSeasonNumber,isRookie) %>%
  summarise(mean_tov = mean(tov),.groups = 'drop')
```

```
## # A tibble: 200 × 3
##    bsSeasonNumber isRookie mean_tov
##             <int> <lgl>       <dbl>
##  1              1 FALSE        68.6
##  2              1 TRUE         36.9
##  3              2 FALSE        65.6
##  4              2 TRUE         28.5
##  5              3 FALSE        62.5
##  6              3 TRUE         26.5
##  7              4 FALSE        67.5
##  8              4 TRUE         29.9
##  9              5 FALSE        74.8
## 10              5 TRUE         31.3
## # ℹ 190 more rows
```

Bootstrap to measure Confidence

- Compare rookie versus non-rookie turnovers each season

```
bsSeasons %>%
  group_by(bsSeasonNumber,isRookie) %>%
  summarise(mean_tov = mean(tov),.groups = 'drop') %>%
  spread(isRookie,mean_tov)
```

```
## # A tibble: 100 × 3
##    bsSeasonNumber `FALSE` `TRUE`
##             <int>   <dbl>  <dbl>
##  1              1    68.6   36.9
##  2              2    65.6   28.5
##  3              3    62.5   26.5
##  4              4    67.5   29.9
##  5              5    74.8   31.3
##  6              6    70.7   31.6
##  7              7    73.7   19.8
##  8              8    73.7   33
##  9              9    65.0   24.3
## 10             10    72.2   28.0
## # ℹ 90 more rows
```

Bootstrap to measure Confidence

- Compare rookie versus non-rookie turnovers each season

```
bsSeasons %>%
  group_by(bsSeasonNumber,isRookie) %>%
  summarise(mean_tov = mean(tov),.groups = 'drop') %>%
  spread(isRookie,mean_tov) %>%
  filter(complete.cases(.)) %>%
  mutate(rookieBetter = ifelse(`FALSE` > `TRUE`,1,0))
```

```
## # A tibble: 100 × 4
##    bsSeasonNumber `FALSE` `TRUE` rookieBetter
##             <int>   <dbl>  <dbl>        <dbl>
##  1              1    68.6   36.9            1
##  2              2    65.6   28.5            1
##  3              3    62.5   26.5            1
##  4              4    67.5   29.9            1
##  5              5    74.8   31.3            1
##  6              6    70.7   31.6            1
##  7              7    73.7   19.8            1
##  8              8    73.7   33              1
##  9              9    65.0   24.3            1
## 10             10    72.2   28.0            1
```

Bootstrap to measure Confidence

- Compare UVA and UT's FT percentages in each season

```
(conf <- bsSeasons %>%
  group_by(bsSeasonNumber,isRookie) %>%
  summarise(mean_tov = mean(tov),.groups = 'drop') %>%
  spread(isRookie,mean_tov) %>%
  filter(complete.cases(.)) %>%
  mutate(rookieBetter = ifelse(`FALSE` > `TRUE`,1,0)) %>%
  summarise(rookieBetter = mean(rookieBetter)))
```

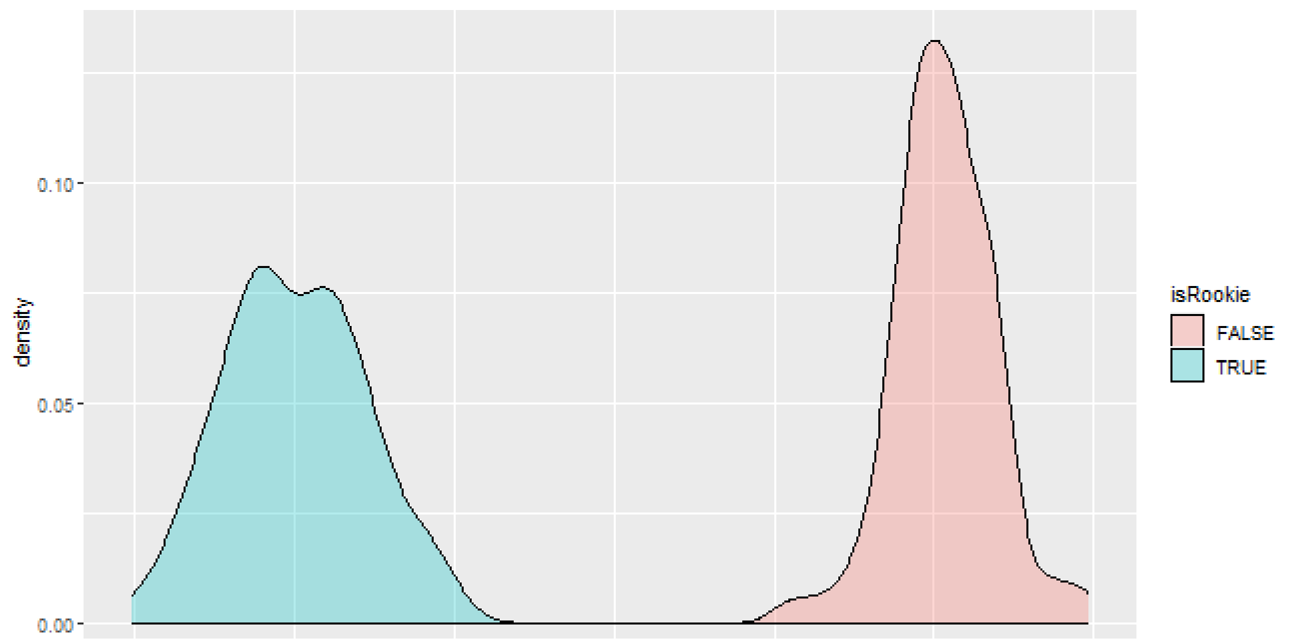
```
## # A tibble: 1 × 1
##   rookieBetter
##          <dbl>
## 1            1
```

- Rookies have fewer turnovers 100% of the time! (How much do you bet on next season?)

Other ways to use bootstraps

- Could plot the **distributions** for each school

```
bsSeasons %>%
 group_by(bsSeasonNumber,isRookie) %>%
 summarise(mean_tov = mean(tov),.groups = 'drop') %>%
 ggplot(aes(x = mean_tov,fill = isRookie)) +
 geom_density(alpha = .3)
```



Other ways to use bootstraps

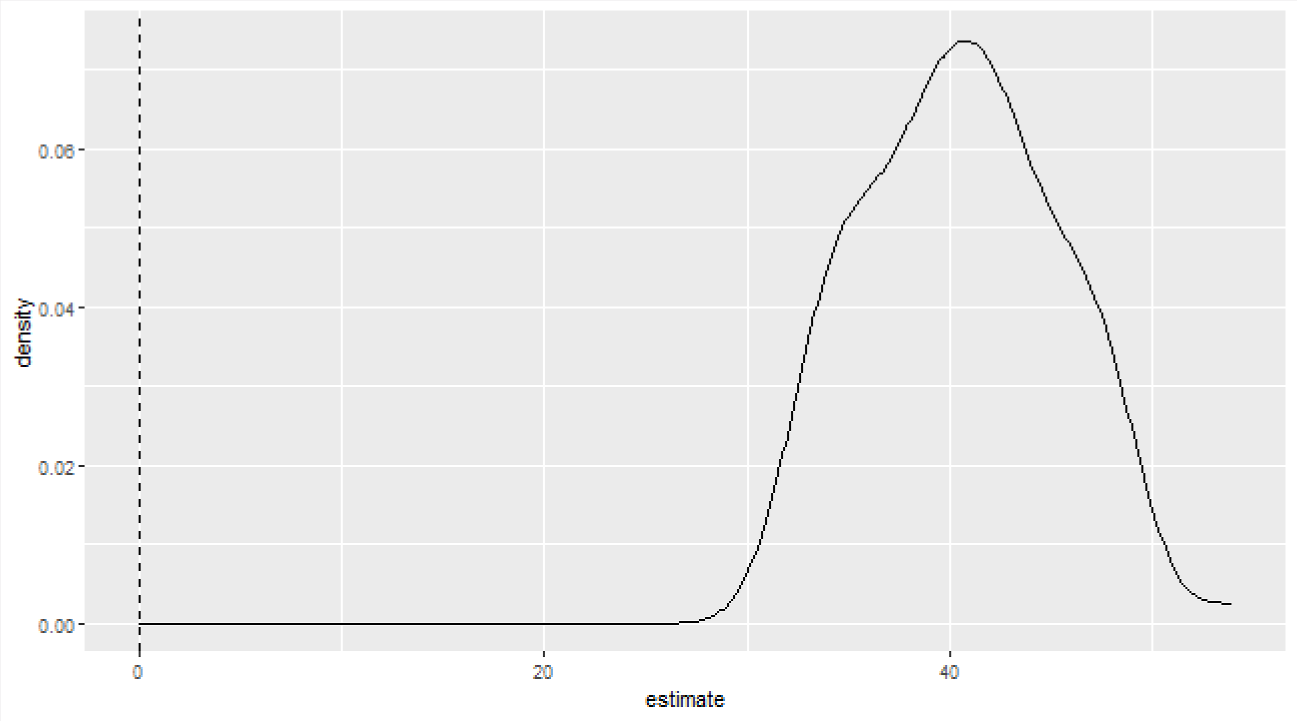
- Could plot the **distributions** of the "estimate"

```r
p <- bsSeasons %>%
  group_by(bsSeasonNumber,isRookie) %>%
  summarise(mean_tov = mean(tov),.groups = 'drop') %>%
  spread(isRookie,mean_tov) %>%
  mutate(estimate = `FALSE` - `TRUE`) %>%
  ggplot(aes(x = estimate)) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0,linetype = 'dashed')
```

Other ways to use bootstraps

- Could plot the **distributions** of the "estimate"

```
p
```



Where to calculate the "estimate"

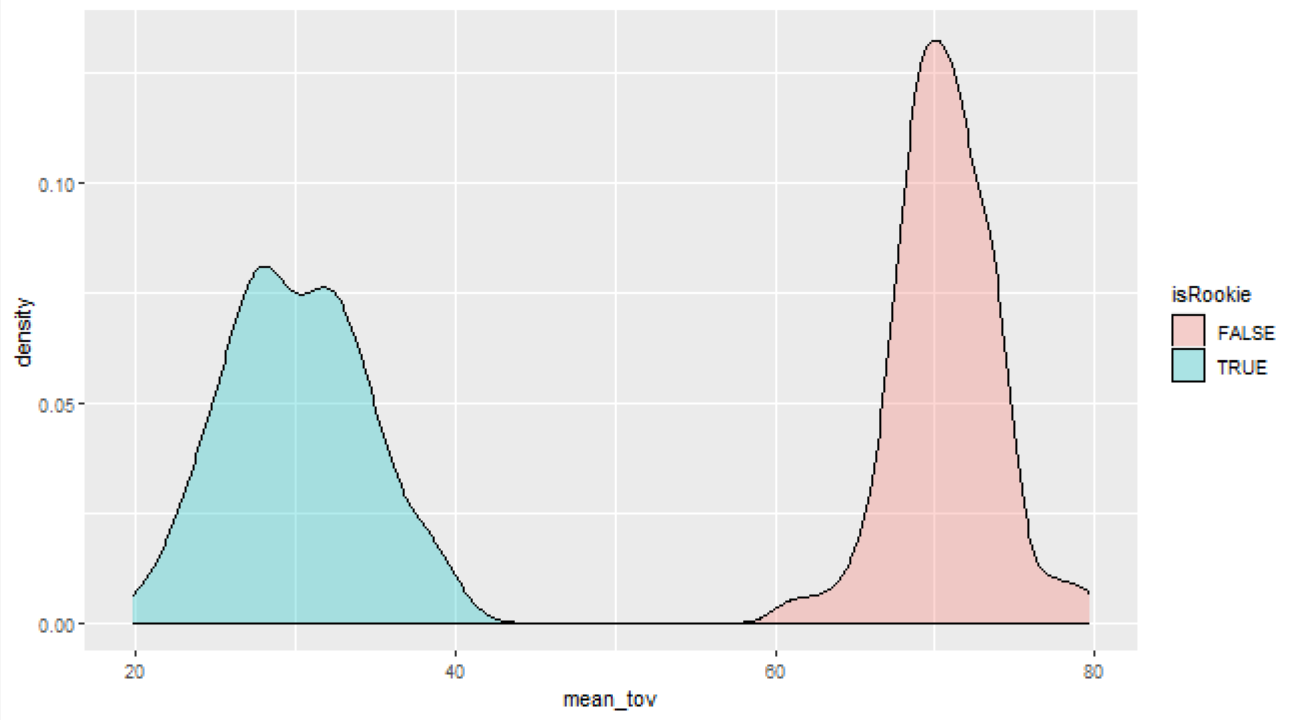
- **First** we created a new dataset of 100 simulated seasons
- **Then** we calculate average FT % for TN and UVA for each simulation
- **Finally** we calculate proportion of times average is higher for TN
- **BUT!** It is equally valid to calculate the "estimate" *within* the `for()` loop

```r
set.seed(123)
bsRes <- NULL
for(counter in 1:100) {
  tmpEst <- nba %>%
    sample_n(size = nrow(.),replace = T) %>%
    group_by(isRookie) %>%
    summarise(mean_tov = mean(tov,na.rm=T)) %>%
    mutate(bsSeason = counter)

  bsRes <- bind_rows(bsRes,tmpEst)
}
```

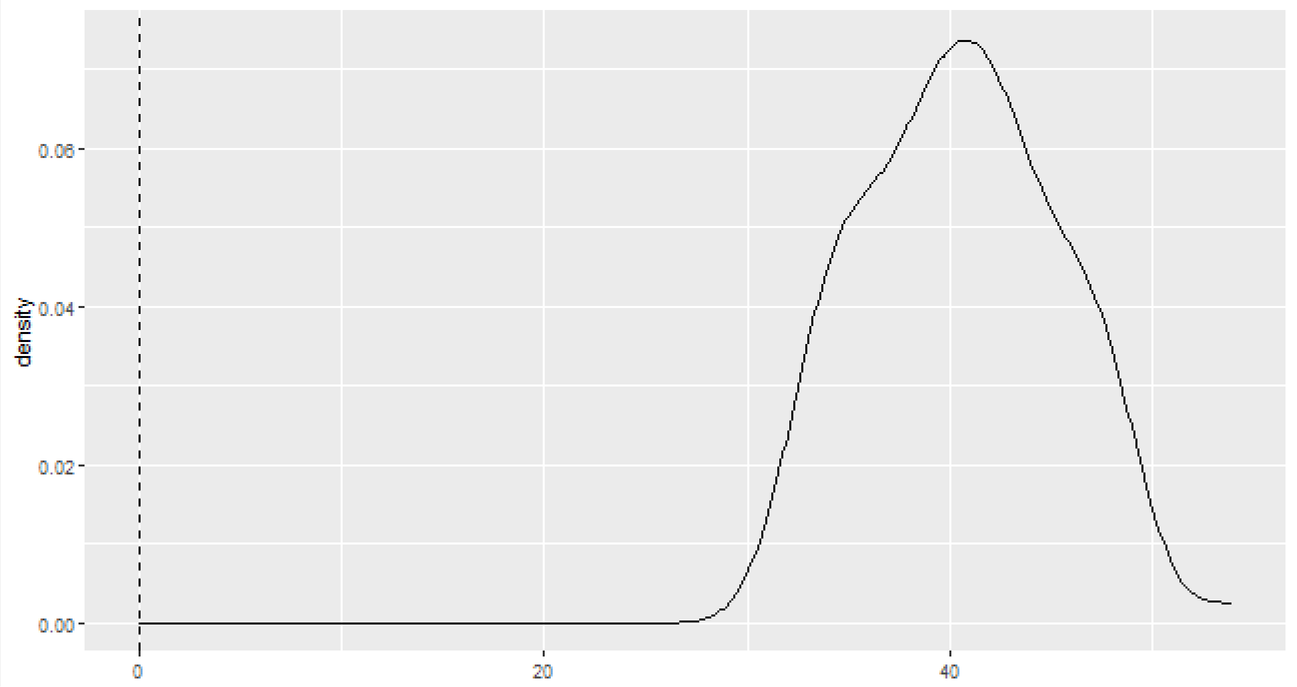
Where to calculate the "estimate"

```
bsRes %>%
  ggplot(aes(x = mean_tov,fill = isRookie)) +
  geom_density(alpha = .3)
```



Where to calculate the "estimate"

```
bsRes %>%
  spread(isRookie,mean_tov) %>%
  mutate(rookieBetter = `FALSE` - `TRUE`) %>%
  ggplot(aes(x = rookieBetter)) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0,linetype = 'dashed')
```



Where to calculate the "estimate"

- Same confidence measure

```
bsRes %>%
  spread(key = isRookie,value = mean_tov) %>%
  mutate(rookieBetter = ifelse(`FALSE` > `TRUE`,1,0)) %>%
  summarise(confidence = mean(rookieBetter,na.rm=T))
```

```
## # A tibble: 1 × 1
##   confidence
##        <dbl>
## 1          1
```

Interpreting Confidence

- **Is this high?**
    - What value reflects the minimum confidence?
    - A coin flip → 50%
- What does a confidence level of 0.1 (or 10%) mean?
    - We are 100% confident?

Do we believe this?

- Why might this conclusion be **spurious**?
- Rookies get less playing time
- Therefore fewer opportunities to turn the ball over
- Solution? Turnovers per minute (or hour)

Re-evaluating

```
nba <- nba %>%
  mutate(tov_hr = tov*60 / minutes)

nba %>%
  group_by(isRookie) %>%
  summarise(tov_hr = mean(tov_hr))
```

```
## # A tibble: 2 × 2
##   isRookie tov_hr
##   <lgl>     <dbl>
## 1 FALSE      3.24
## 2 TRUE       2.78
```

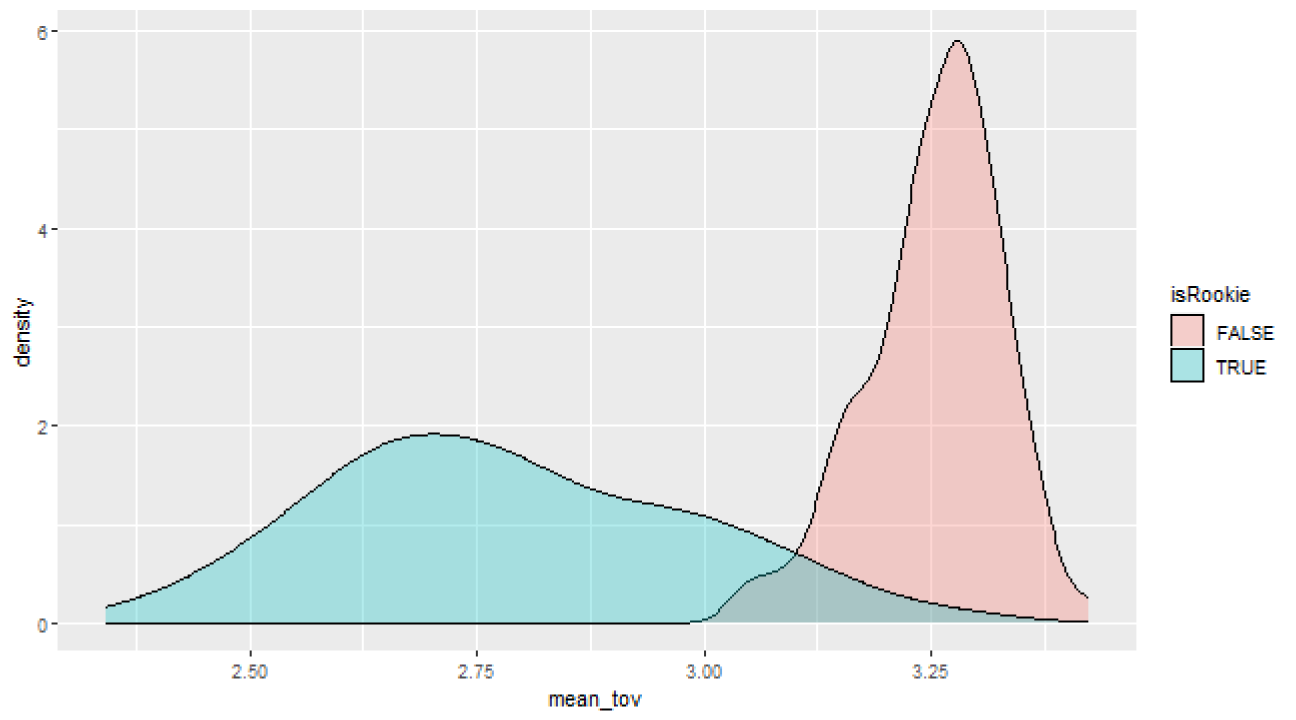
Re-evaluating

```r
set.seed(123)
bsRes <- NULL
for(counter in 1:100) {
  tmpEst <- nba %>%
    sample_n(size = nrow(.),replace = T) %>%
    group_by(isRookie) %>%
    summarise(mean_tov = mean(tov_hr,na.rm=T)) %>%
    mutate(bsSeason = counter)

  bsRes <- bind_rows(bsRes,tmpEst)
}
```

Re-evaluating

```
bsRes %>%
  ggplot(aes(x = mean_tov,fill = isRookie)) +
  geom_density(alpha = .3)
```



Re-Evaluating

```r
bsRes %>%
  mutate(isRookie = ifelse(isRookie == TRUE,'Rookie','Not Rookie'))
%>%
  spread(isRookie,mean_tov) %>%
  summarise(conf = mean(`Not Rookie` > Rookie))
```

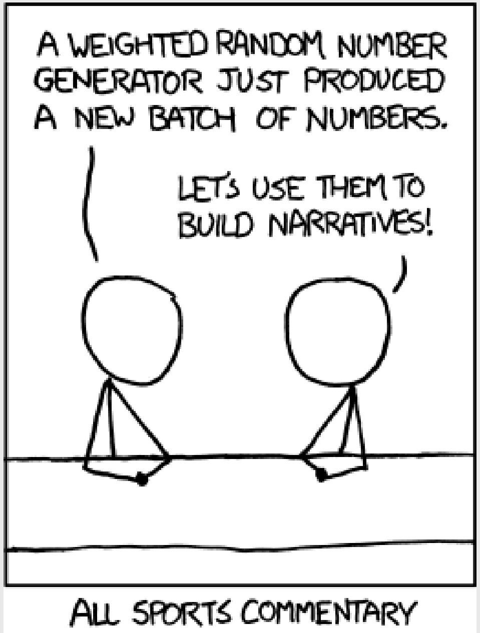
```
## # A tibble: 1 × 1
##    conf
##   <dbl>
## 1  0.99
```

Other Applications

- Could do the same to express **confidence** in conclusions about:
  - The relationship between SAT scores and selective admissions
  - The relationship between MSM polls and anti-Trump bias
  - Whether state polls are good at predicting the 2020 president

Conclusion

- Anyone can spit stats


A WEIGHTED RANDOM NUMBER GENERATOR JUST PRODUCED A NEW BATCH OF NUMBERS.
LET'S USE THEM TO BUILD NARRATIVES!
ALL SPORTS COMMENTARY

- Data scientists are comfortable with **uncertainty**

Quiz & Homework

- Go to Brightspace and take the **9th** quiz
    - The password to take the quiz is ####
- **Homework:**
    1. Work through ds1000_hw_10.Rmd (regression!)
    2. Problem Set 6 (on Brightspace)