Lecture_2_13_2023_notes

Prof. Bisbee, Vanderbilt University

2023-02-13

```
require(tidyverse)
```

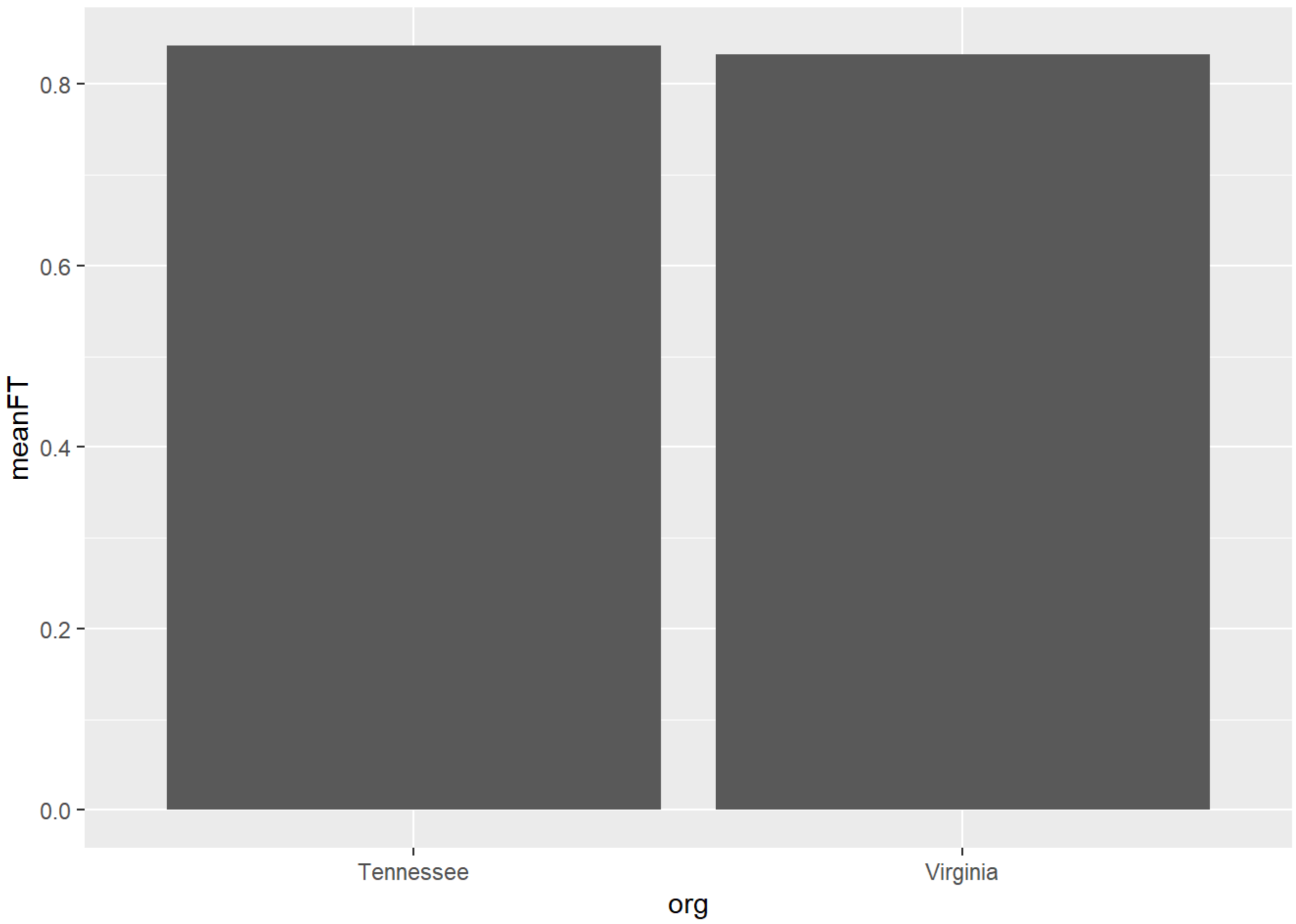
```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

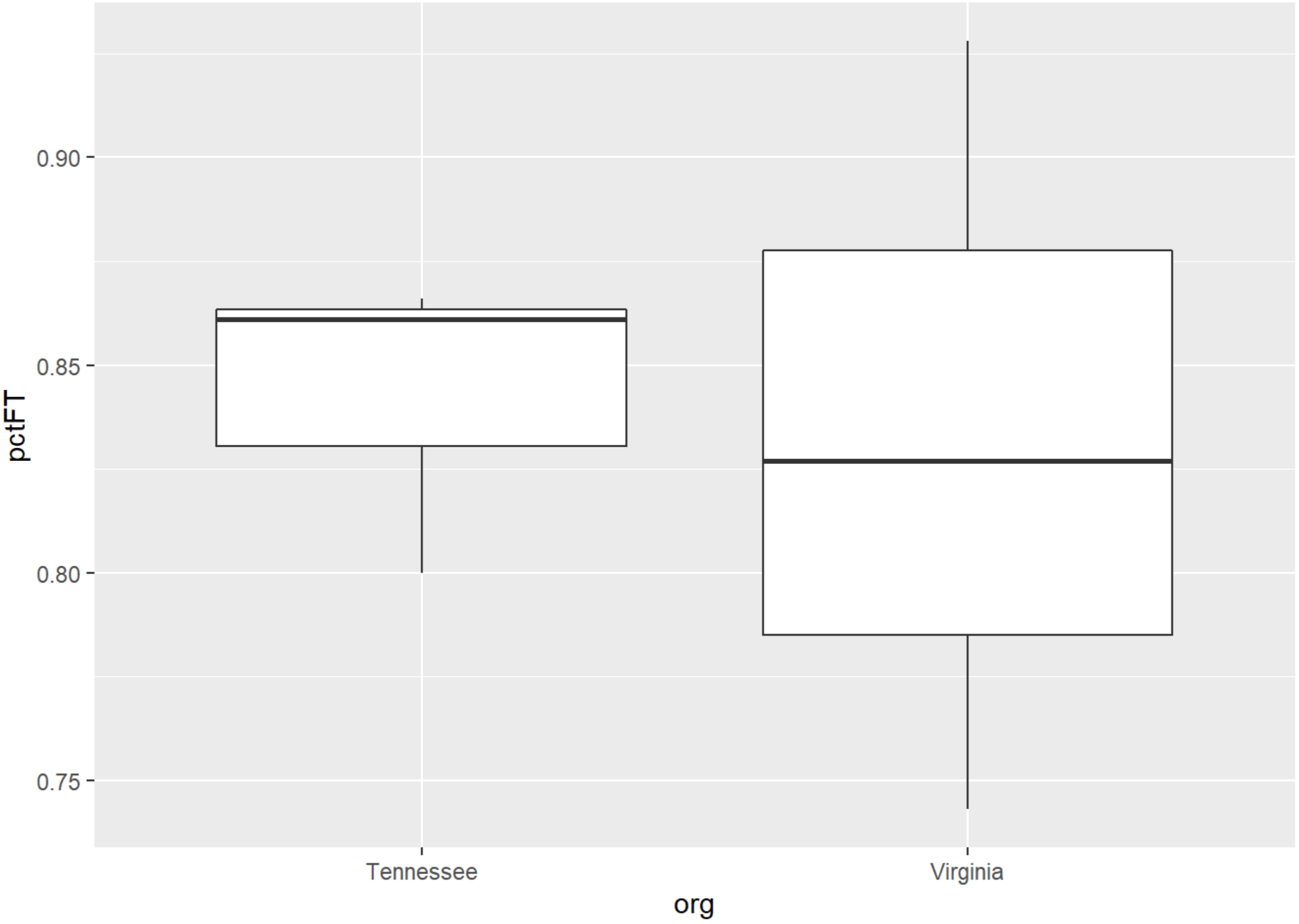
```
nba <- read_rds('https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Multivariate/data/nba_players_2018.Rds?raw=true')

gms <- read_rds('https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Multivariate/data/game_summary.Rds?raw=true')
```

Multivariate Analysis

```
nba %>%
  filter(org %in% c('Tennessee','Virginia')) %>%
  group_by(org) %>%
  summarise(meanFT = mean(pctFT)) %>%
  ggplot(aes(x = org,y = meanFT)) +
  geom_bar(stat = 'identity')
```



```
# Method 2 visualization
nba %>%
  filter(org %in% c('Tennessee','Virginia')) %>%
  ggplot(aes(x = org,y = pctFT)) +
  geom_boxplot()
```



Bootstrap sample example

```
set.seed(123)
nba %>%
  select(org,pctFT) %>%
  sample_n(size = 1,replace = T)
```

```
## # A tibble: 1 × 2
##   org      pctFT
##   <fct>    <dbl>
## 1 Michigan 0.811
```

```
# Get a simulated season
simSeason1 <- nba %>%
  select(org,pctFT) %>%
  sample_n(size = nrow(nba),replace = T)

simSeason1 %>%
  filter(org %in% c('Tennessee','Virginia')) %>%
  group_by(org) %>%
  summarise(meanFT = mean(pctFT))
```

```
## # A tibble: 2 × 2
##   org       meanFT
##   <fct>      <dbl>
## 1 Tennessee  0.866
## 2 Virginia   0.785
```

Repeating a chunk with `for()` loop

```
bootstrap_result <- NULL
for(indexNumber in 1:1000) {
  # stop()
  simSeason <- nba %>%
  select(org,pctFT) %>%
  sample_n(size = nrow(nba),replace = T) %>%
    mutate(bootstrap_number = indexNumber)

  bootstrap_result <- bootstrap_result %>%
    bind_rows(simSeason)
}

prepared_bootstrap <- bootstrap_result %>%
  filter(org %in% c('Tennessee','Virginia')) %>%
  group_by(bootstrap_number,org) %>%
  summarise(meanFT = mean(pctFT),.groups = 'drop') %>%
  spread(key = org,value = meanFT) %>%
  mutate(diff = Tennessee - Virginia)


# Calculate confidence
prepared_bootstrap %>%
  summarise(confidence = mean(diff > 0,na.rm=T))
```

```
## # A tibble: 1 × 1
##   confidence
##        <dbl>
## 1      0.583
```

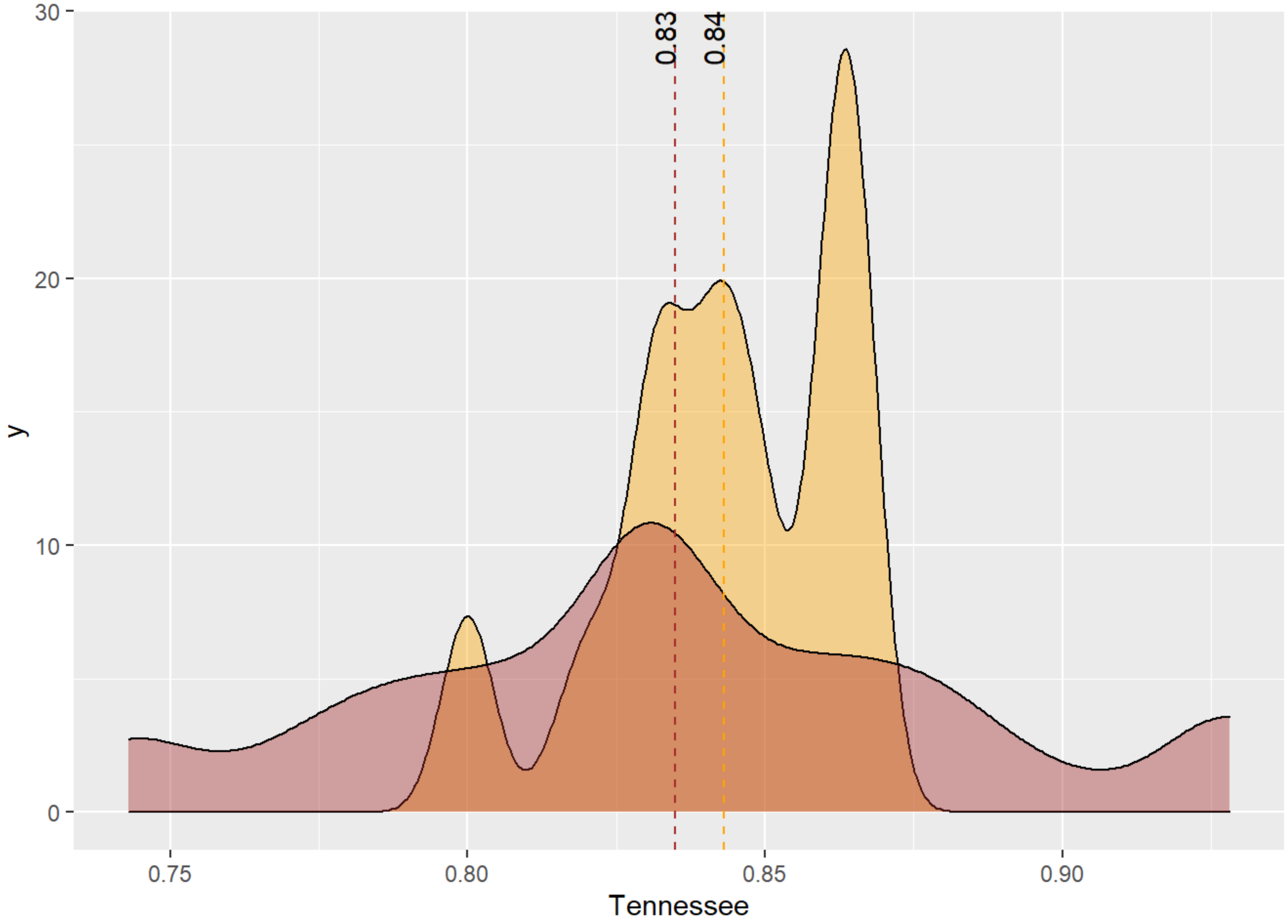
Visualizing Uncertainty

- Method 1: plot the **outcomes**

```
prepared_bootstrap %>%
  ggplot() +
  geom_density(aes(x = Tennessee),fill = 'orange',alpha = .4) +
  geom_density(aes(x = Virginia),fill = 'brown',alpha = .4) +
  geom_vline(xintercept = mean(prepared_bootstrap$Tennessee,na.rm=T),
             linetype = 'dashed',color = 'orange') +
  geom_vline(xintercept = mean(prepared_bootstrap$Virginia,na.rm=T),
             linetype = 'dashed',color = 'brown') +
  annotate(geom = 'text',x = mean(prepared_bootstrap$Tennessee,na.rm=T),y = Inf,label = round(mean(prepared
_bootstrap$Tennessee,na.rm=T),2),angle = 90,hjust = 1,vjust = 0) +
  annotate(geom = 'text',x = mean(prepared_bootstrap$Virginia,na.rm=T),y = Inf,label = round(mean(prepared_
bootstrap$Virginia,na.rm=T),2),angle = 90,hjust = 1,vjust = 0)
```

```
## Warning: Removed 58 rows containing non-finite values (stat_density).
```

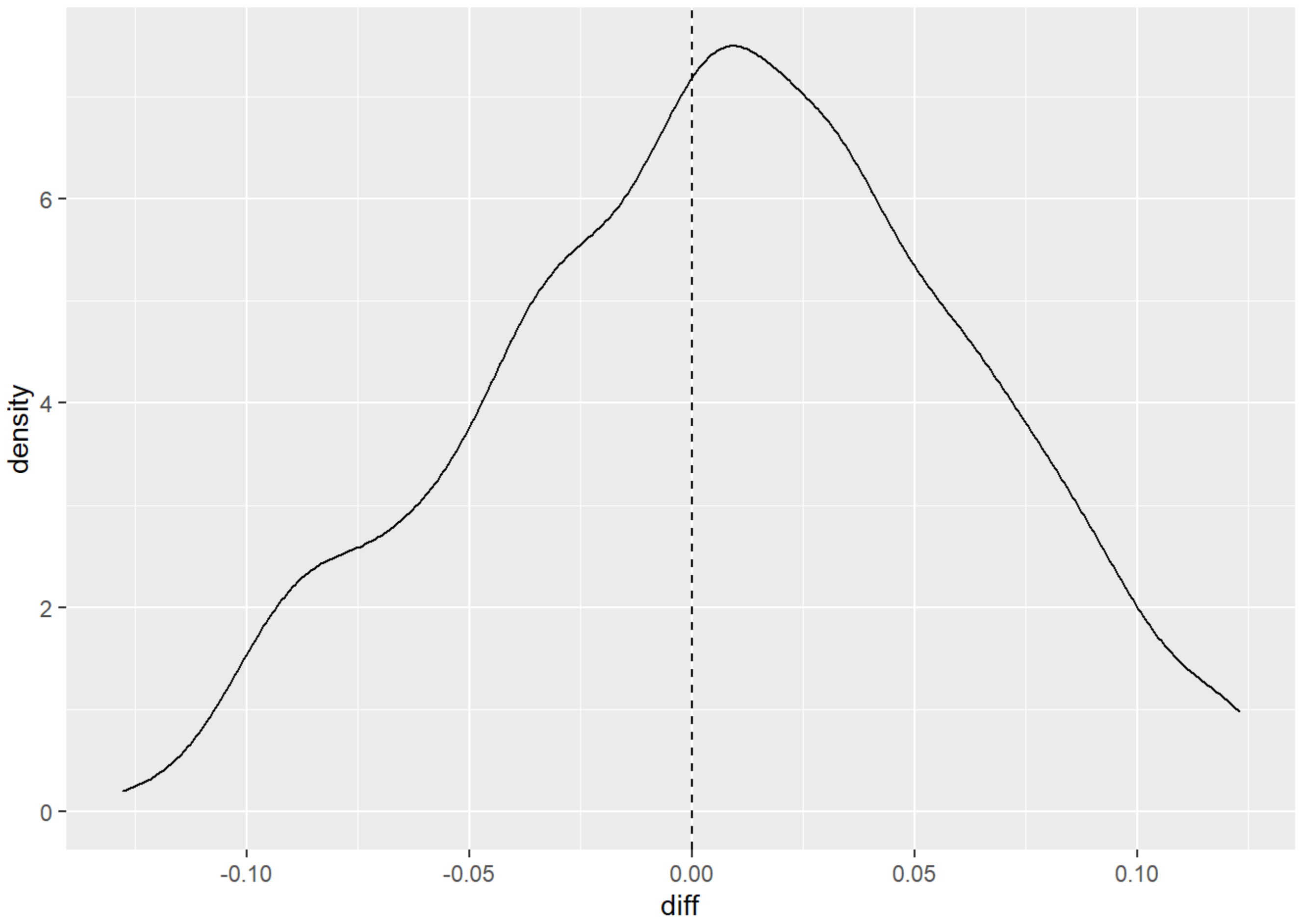
```
## Warning: Removed 49 rows containing non-finite values (stat_density).
```



- Method 2: plotting the **estimates**

```
prepared_bootstrap %>%
  ggplot(aes(x = diff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed')
```

```
## Warning: Removed 107 rows containing non-finite values (stat_density).
```


density
0
2
4
6
-0.10
-0.05
0.00
0.05
0.10
diff