POLI210: Political Science Research Methods

Lecture 9.2: Means and distributions

Olivier Bergeron-Boutin

October 28th, 2021

Boring admin stuff

- First quiz: how did it go? (POLLING)
  - I will release the grades shortly
- Grading in pset1: some accommodations
- I will attend labs tomorrow and announce in-person OHs

Plan for this lecture

1. Descriptive statistics

   - Measures of central tendency: mean, median, mode
   - Measures of dispersion: standard deviation, variance

2. Confidence intervals
3. How should I describe distributions?

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

- μ is simply a letter that represents the mean

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

- μ is simply a letter that represents the mean
- This is a summation operator

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

- μ is simply a letter that represents the mean
- This is a summation operator
- This is what the summation operator does

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

- μ is simply a letter that represents the mean
- This is a summation operator
- This is what the summation operator does

$$\sum_{i=1}^{5} a_i = 1 + 2 + 3 + 4 + 5$$

Measures of central tendency: the mean

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Wait, what?

- μ is simply a letter that represents the mean
- This is a summation operator
- This is what the summation operator does

$$\sum_{i=1}^{5} a_i = 1 + 2 + 3 + 4 + 5$$

$$\sum_{i=10}^{12} a_i = 10 + 11 + 12$$

Measures of central tendency: the mean

Take the following vector: $(7, 1, 10) = (a_1, a_2, a_3)$

Measures of central tendency: the mean

Take the following vector: $(7, 1, 10) = (a_1, a_2, a_3)$

- The mean is:

$$\mu = \frac{1}{5} \sum_{i=1}^{3} a_i$$

Measures of central tendency: the mean

Take the following vector: $(7, 1, 10) = (a_1, a_2, a_3)$

- The mean is:

$$\mu = \frac{1}{5} \sum_{i=1}^{3} a_i$$

$$= \frac{7 + 1 + 10}{3} = 6$$

Measures of central tendency: the mean

Take the following vector: $(7, 1, 10) = (a_1, a_2, a_3)$

- The mean is:

$$\mu = \frac{1}{5} \sum_{i=1}^{3} a_i$$

$$= \frac{7 + 1 + 10}{3} = 6$$

Measures of central tendency: the mean

Why do we like the mean? It's often a good one-number summary of the data

- But not always: the mean is sensitive to outliers
- What's the mean here? $(23, 28, 97)$
- $\mu = \frac{23 + 28 + 96}{3} = 49$

But there's no one even close to 49!

- That's because of the "outlying" value of 97
- Outlier: a value that is far from rest of distribution
- Example:
  - The mean income is in this zoom meeting? Probably around 5,000$
  - What if Elon Musk walks in? (net worth: 255 billion)
  - $\mu = \frac{255 \text{ billion+ whatever we make}}{100 + 1} = 2.5 \text{ billion}$

Measures of central tendency: the median

Observation $\frac{n+1}{2}$ (once ordered)

Consider this vector of values:

```
set.seed(123)
incomes_5 <- sample(1:100, size = 5, replace = TRUE)
incomes_5
```

```
## [1] 31 79 51 14 67
```

```
# Let's order them:
incomes_5[order(incomes_5)]
```

```
## [1] 14 31 51 67 79
```

The median is the: $\frac{5+1}{2} = 3^{\text{rd}}$ value $= 42$

Measures of central tendency: the median

```
# Imagine that there are 9 students and me; our incomes:
incomes <- sample(1000:15000, size = 10, replace = TRUE)

# Let's order them:
incomes[order(incomes)]
```

```
##  [1]  2841  3985  4370  5760  7745 10333 12637 14325 14540 14555
```

```
median(incomes)
```

```
## [1] 9039
```

```
# Elon Musk walks in...
incomes <- c(incomes, 255000000000)
# Let's order them:
incomes[order(incomes)]
```

```
##  [1]         2841         3985         4370         5760         7745
##  [6]        10333        12637        14325        14540        14555
## [11] 255000000000
```

```
median(incomes)
```

```
## [1] 10333
```

Measures of central tendency: the mode

The mode is pretty simple: the value that appears most often

- Not so useful for continuous variables, e.g. income
- Pretty useful for nominal variables
    - Nominal variables: values cannot be ordered

Measures of dispersion: the standard deviation

If you picked a random value from a distribution, how far away from the mean would you expect it to be?

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Procedure:

- For each value, you compute its distance from the mean
- You square that distance (e.g. $4^2 = 16$)
- You sum up all of those squared distances
- You divide by $n - 1$
- You take the square root

Measures of dispersion: the standard deviation

```
incomes_5
```

```
## [1] 31 79 51 14 67
```

```
mean(incomes_5)
```

```
## [1] 48.4
```

```
incomes_5 - mean(incomes_5)
```

```
## [1] -17.4  30.6   2.6 -34.4  18.6
```

```
distance_from_mean_sq <- (incomes_5 - mean(incomes_5))^2
distance_from_mean_sq
```

```
## [1]  302.76  936.36    6.76 1183.36  345.96
```

```
sum_distance <- sum(distance_from_mean_sq)
sum_distance
```

```
## [1] 2775.2
```

```
sqrt(sum_distance/(5-1))
```

```
## [1] 26.34008
```

Confidence intervals

Uncertainty due to sampling

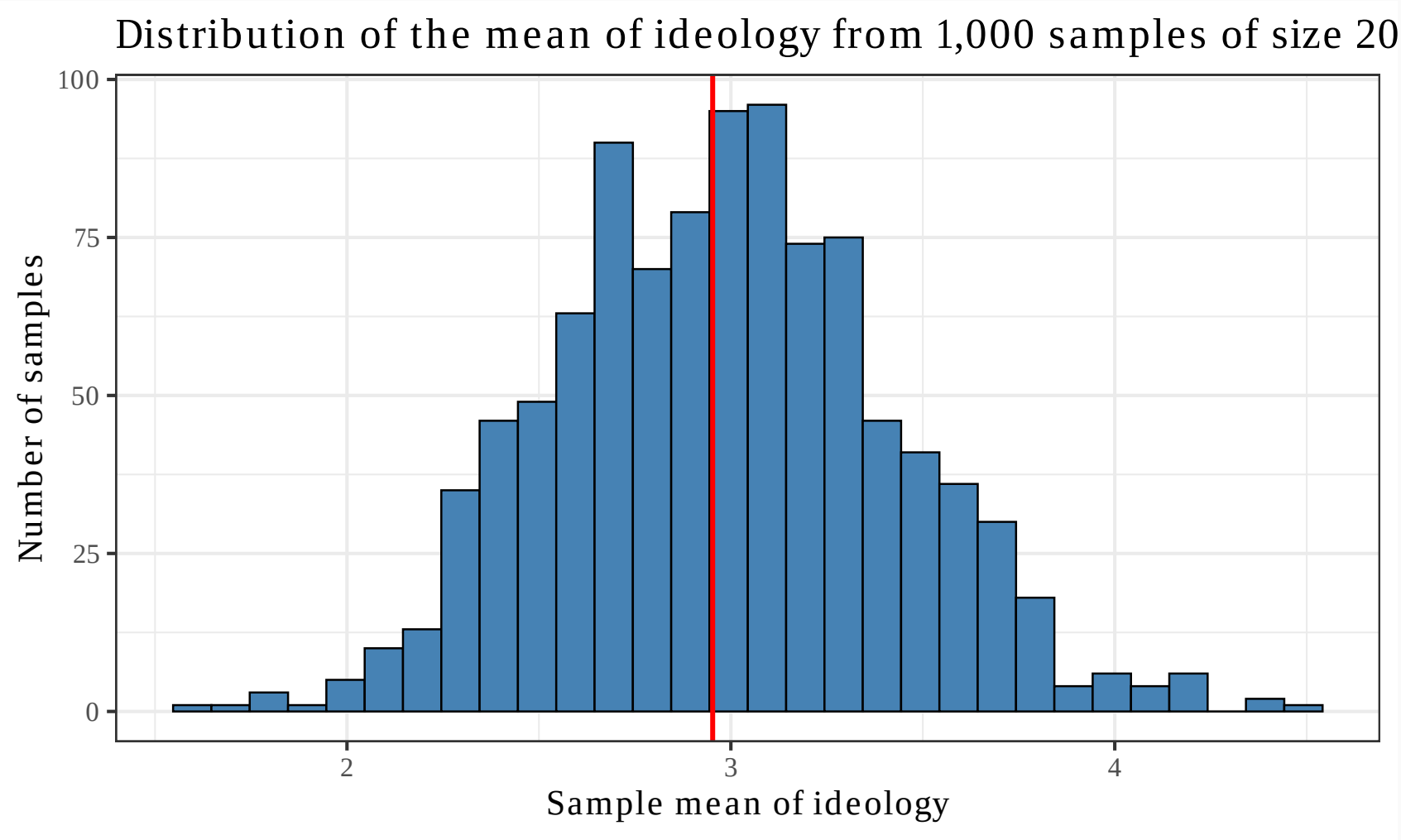
We draw a sample

- Ideally by randomly drawing from the population
- We compute some sample statistic of interest, e.g. the mean height
- We want to **infer** the population parameter
- But we know that there is **sampling variation**
    - Even if we draw a random sample, we may be more or less far from the true population parameter

Remember the **central limit theorem**

- Under repeated (random) sampling, the distribution of sample means (the sampling distribution) will approximate a normal distribution
- No matter the underlying shape of the population distribution!

CLT, again


Distribution of the mean of ideology from 1,000 samples of size 20
Number of samples
100
75
50
25
0
2
3
4
Sample mean of ideology

Confidence intervals

We don't know if we drew a "good sample"

- A good sample: sample mean is close to true population parameter
- Given CLT, we are more likely to be close than to be far
- But there is a possibility that we're way off!

This is where confidence intervals come in

```
# true parameter that sampling is trying to infer
mean(survey$ideology, na.rm = T)
```

```
## [1] 2.952555
```

```
# Making one sample of 30 students
survey_30 <- survey[sample(1:nrow(survey), 30),]
mean(survey_30$ideology, na.rm = T)
```

```
## [1] 2.9
```

```
t.test(survey_30$ideology)$"conf.int"
```

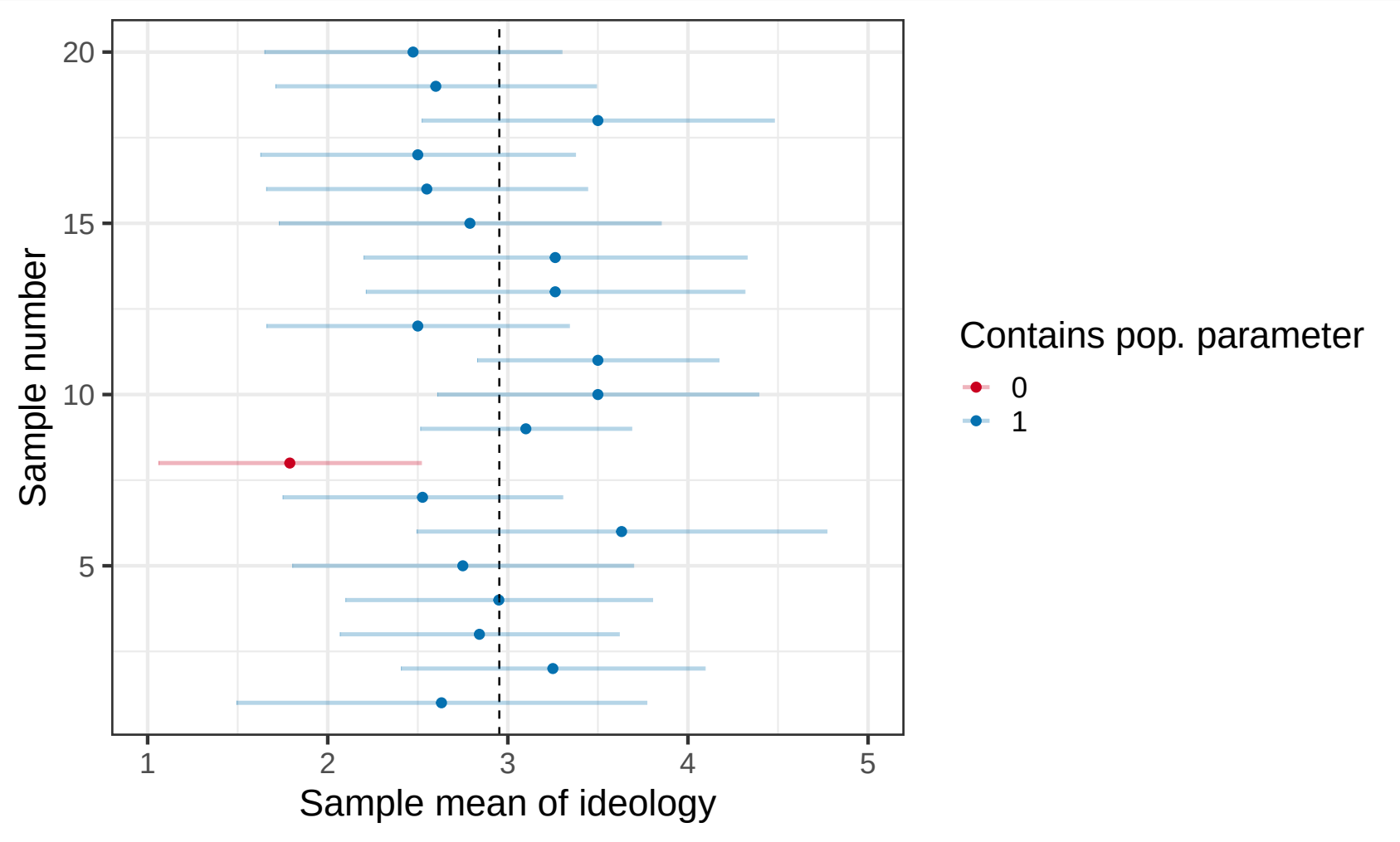
```
## [1] 2.177068 3.622932
```

What your confidence interval says and does not say

Confidence intervals have a confidence "level"

- Generally 95%, but sometimes 90% or 99%
- If we were to repeatedly sample random units and computed the mean and associated confidence interval for each sample, I would expect 95% of confidence intervals to include the true population parameter
- Does **this specific CI in this specific sample** contain the true population parameter?
    - **We don't know!** It's more likely that it does than it doesn't! But it might not!
- Do confidence intervals replace the need for appropriate sampling strategies?
    - **NO!!!** Confidence intervals are only valid under random sampling from the population
    - If there is sampling bias, the confidence interval is NOT VALID

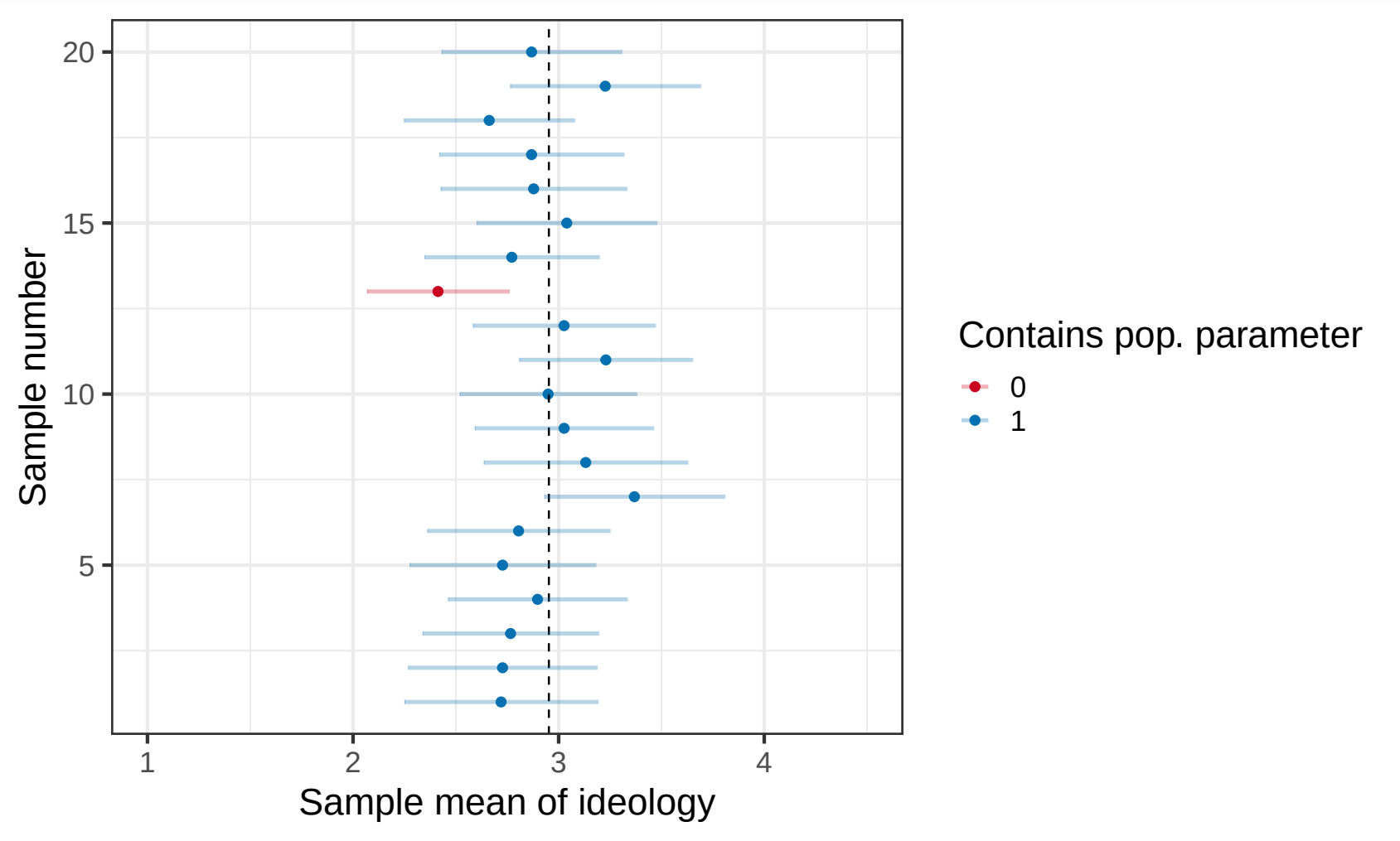
Confidence intervals from the class survey



Each sample is size 20

- Note the quite large CIs! (but coverage looks fine!)

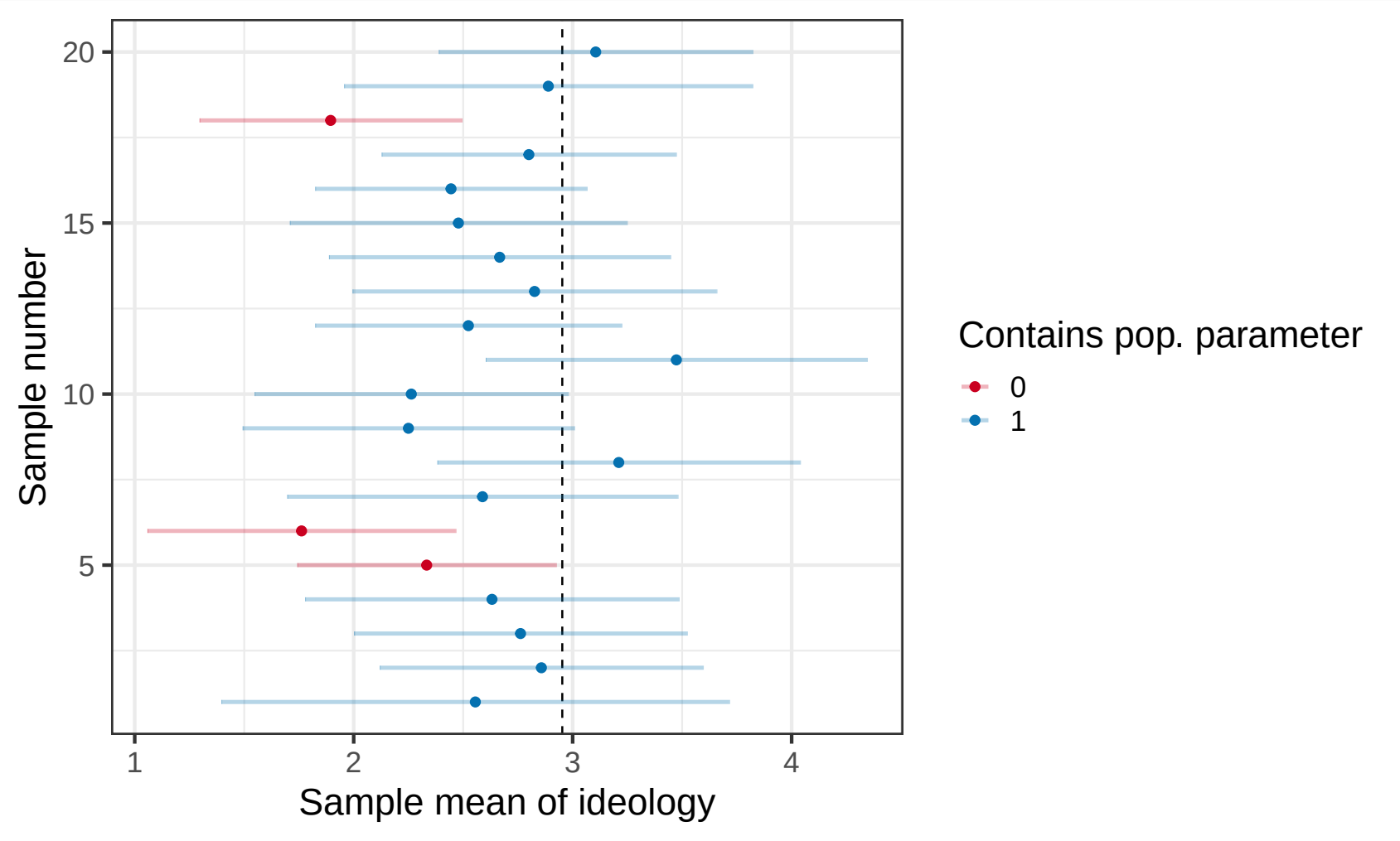
Confidence intervals with larger samples



Each sample is size 80

- The CIs are much narrower (same x scale as before)

Confidence intervals with sampling bias



Non-response bias: men have 0.2 probability of answering; women 0.9

- 3 of my confidence intervals (15%) do not include the true parameter

How is the confidence interval computed?

For the 95% confidence interval, a helpful rule of thumb is:

$$\text{CI}_{\text{lower}} = \hat{\mu} - 2 * SE$$

- What's the hat on top of mu? Simply means it's an estimate from our sample
- What's SE?
  - The **standard error** of the mean
  - It's our estimate of the **standard deviation of the sampling distribution**
  - If I could draw many samples, compute the mean each time, and plot the distribution of sample means...what would be its standard deviation?
- The idea: the larger the SE, the more dispersed the sampling distribution
  - The more likely I draw a sample that is far from true parameter
  - And thus the less confidence we have in our sample estimate

How is the confidence interval computed?

$$SE = \frac{\hat{\sigma}}{\sqrt{n}}$$

$\hat{\sigma}$: our estimate of the population standard deviation

```
sample <- survey[sample(1:nrow(survey), 30),]
mean(sample$ideology, na.rm = T)
```

```
## [1] 2.689655
```

```
sd(sample$ideology, na.rm = T)
```

```
## [1] 1.64975
```

```
se <- sd(sample$ideology, na.rm = T) / sqrt(30)
se
```

```
## [1] 0.3012018
```

```
mean(sample$ideology,na.rm = T) + 2*se
```

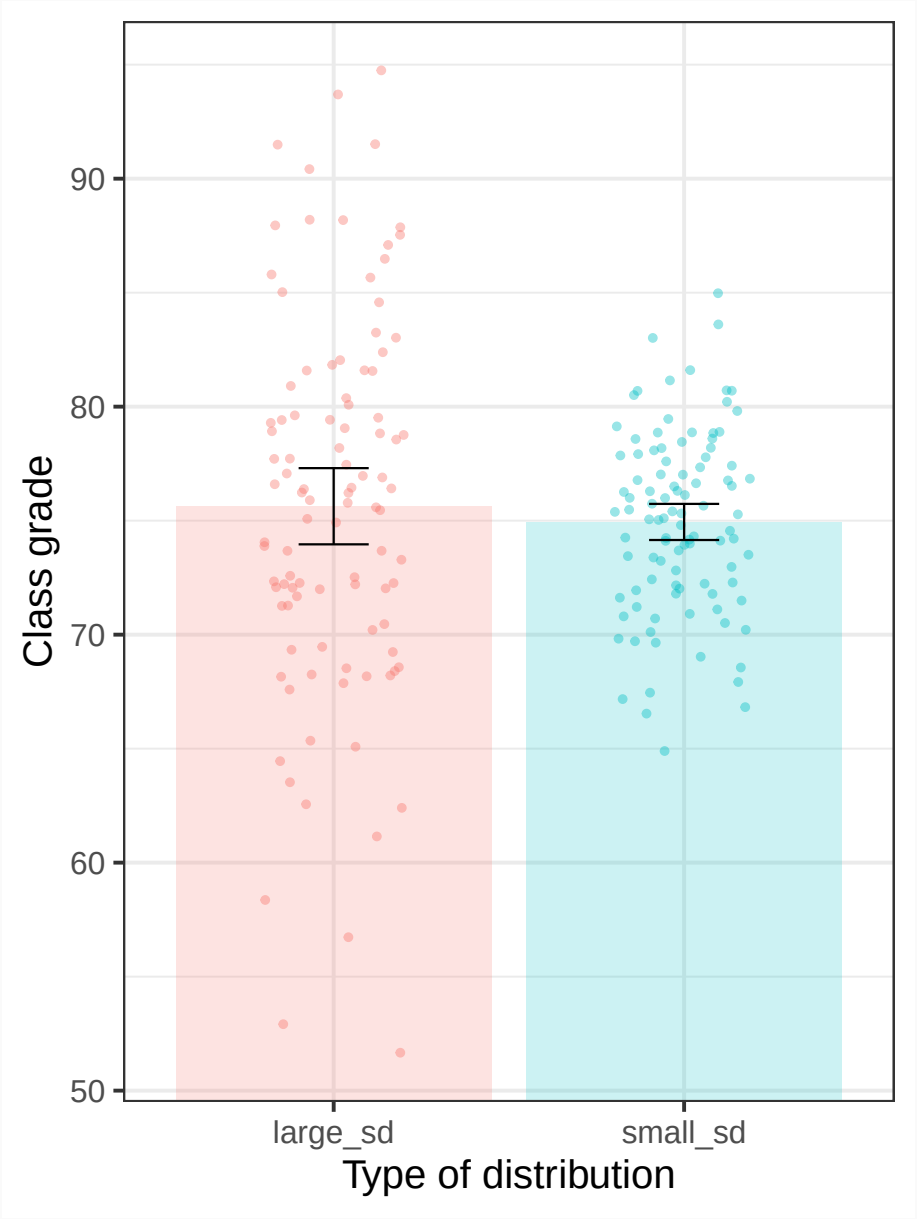
```
## [1] 3.292059
```

What affects the confidence intervals?

Two parameters influence the width of the confidence interval:

- The standard deviation of the data in the sample
  - More spread out means less certainty
  - Think of opposite situation: I draw values of 2,3,2,3,3,3...
  - My mean is still pretty close to 2.95
  - But the sample's SD is very small; the CI will be narrower
- The sample size
  - More people in the sample $\rightsquigarrow$ more precise estimates
  - BUT: notice the square root?
  - There are diminishing returns to sample size
  - Going from 100 to 1,000: great!
  - Going from 1,000 to 10,000: not a great resource expenditure!

SD of the data and CIs



NBA scoring

```
zion_harden <- read.csv("lectures/lecture_9.2/nba_data.csv")

# A tidyverse function to sample from dataframe
sample_n(zion_harden, 6)
```

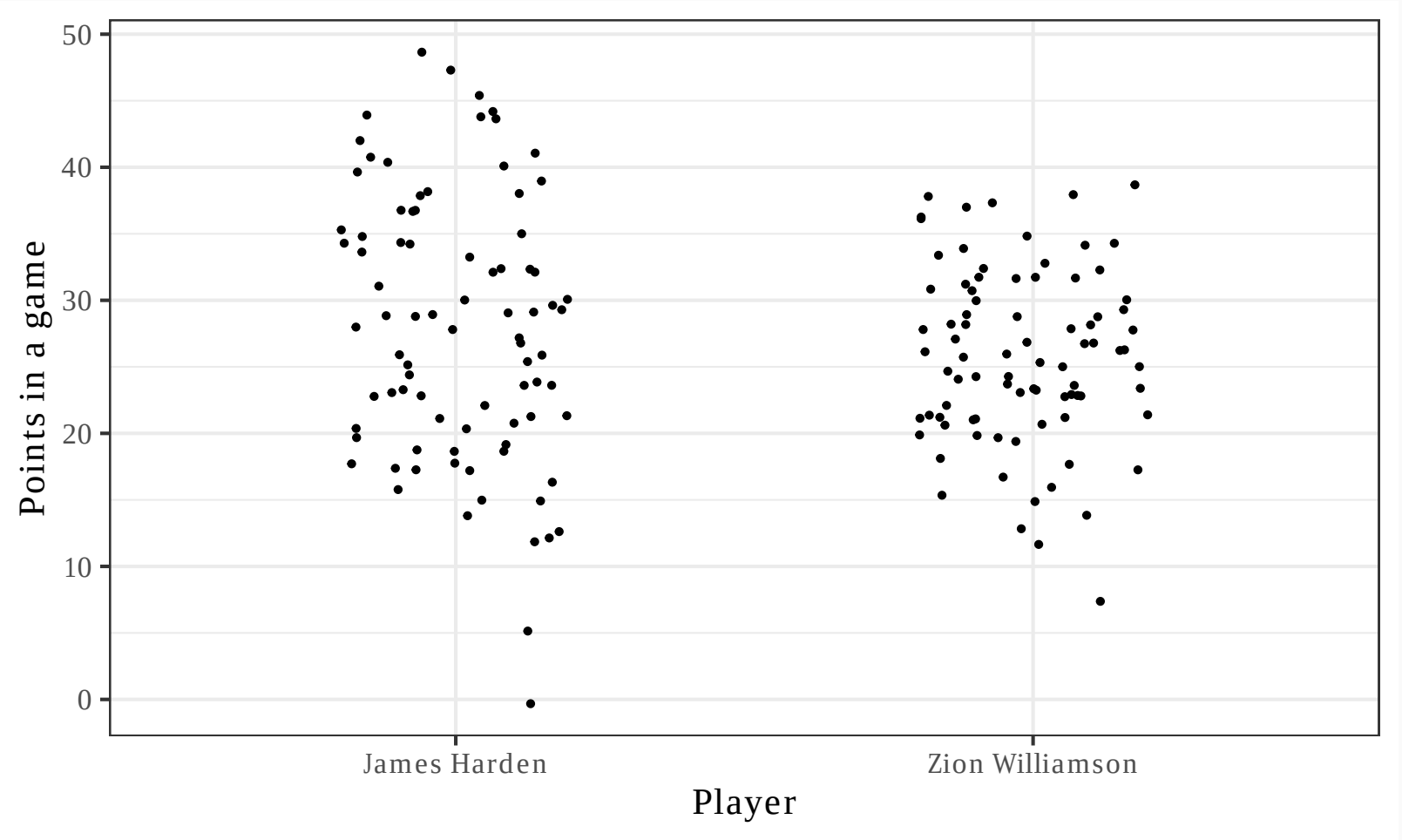
```
##     X yearSeason   dateGame      namePlayer pts
## 1 159       2021 2021-04-14 Zion Williamson  25
## 2  48       2020 2020-08-06    James Harden  39
## 3   4       2021 2021-03-31    James Harden  17
## 4 102       2020 2020-03-04 Zion Williamson  21
## 5  76       2020 2020-01-11    James Harden  32
## 6  46       2020 2020-08-12    James Harden  45
```

Figure 1: Zion Williamson

Figure 2: James Harden

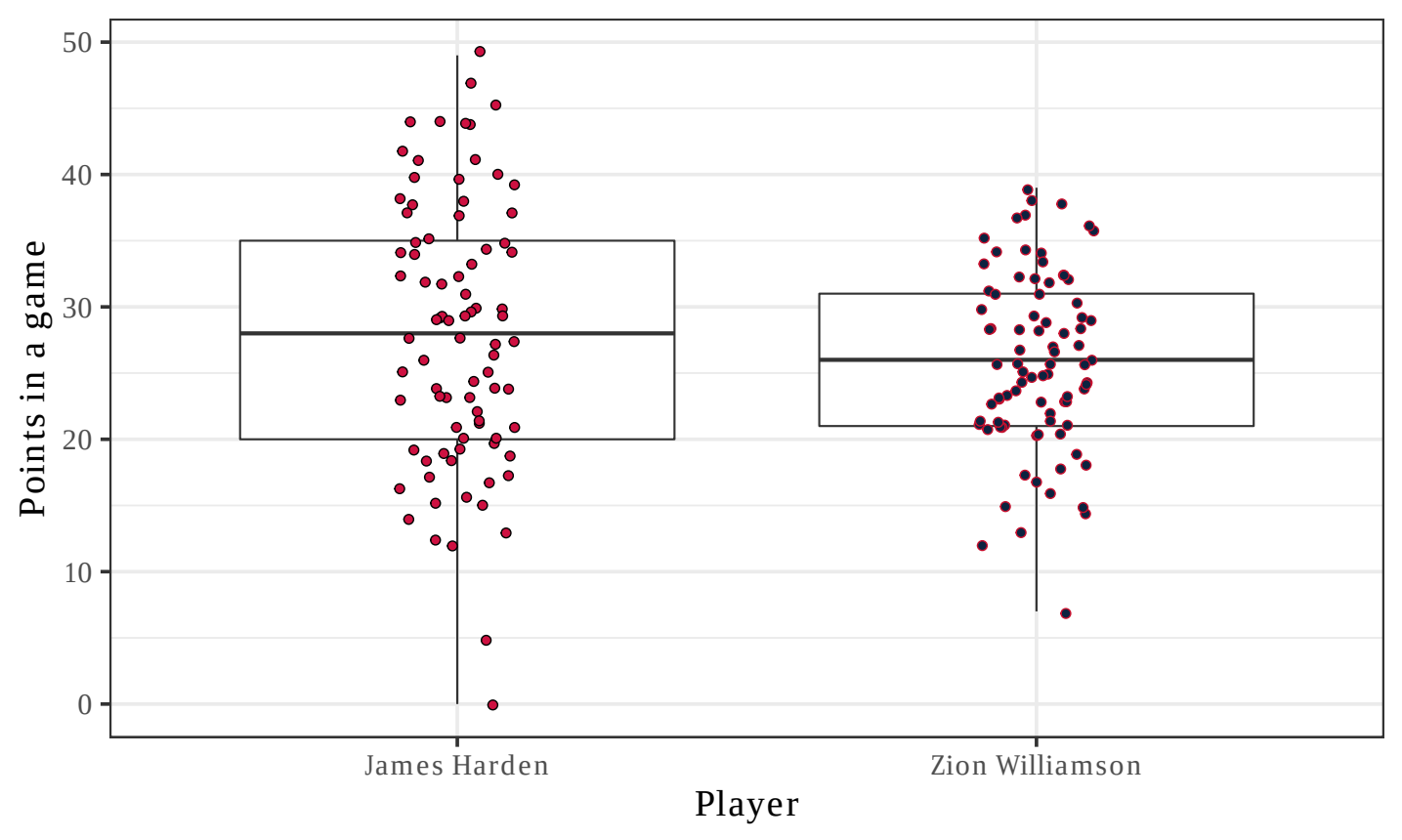
Zion vs Harden

```
ggplot(zion_harden, aes(x = namePlayer, y = pts)) +
  geom_jitter(width = 0.2) +
  labs(x = "Player", y = "Points in a game") +
  theme_bw(base_size = 19, base_family = "Fira Sans")
```



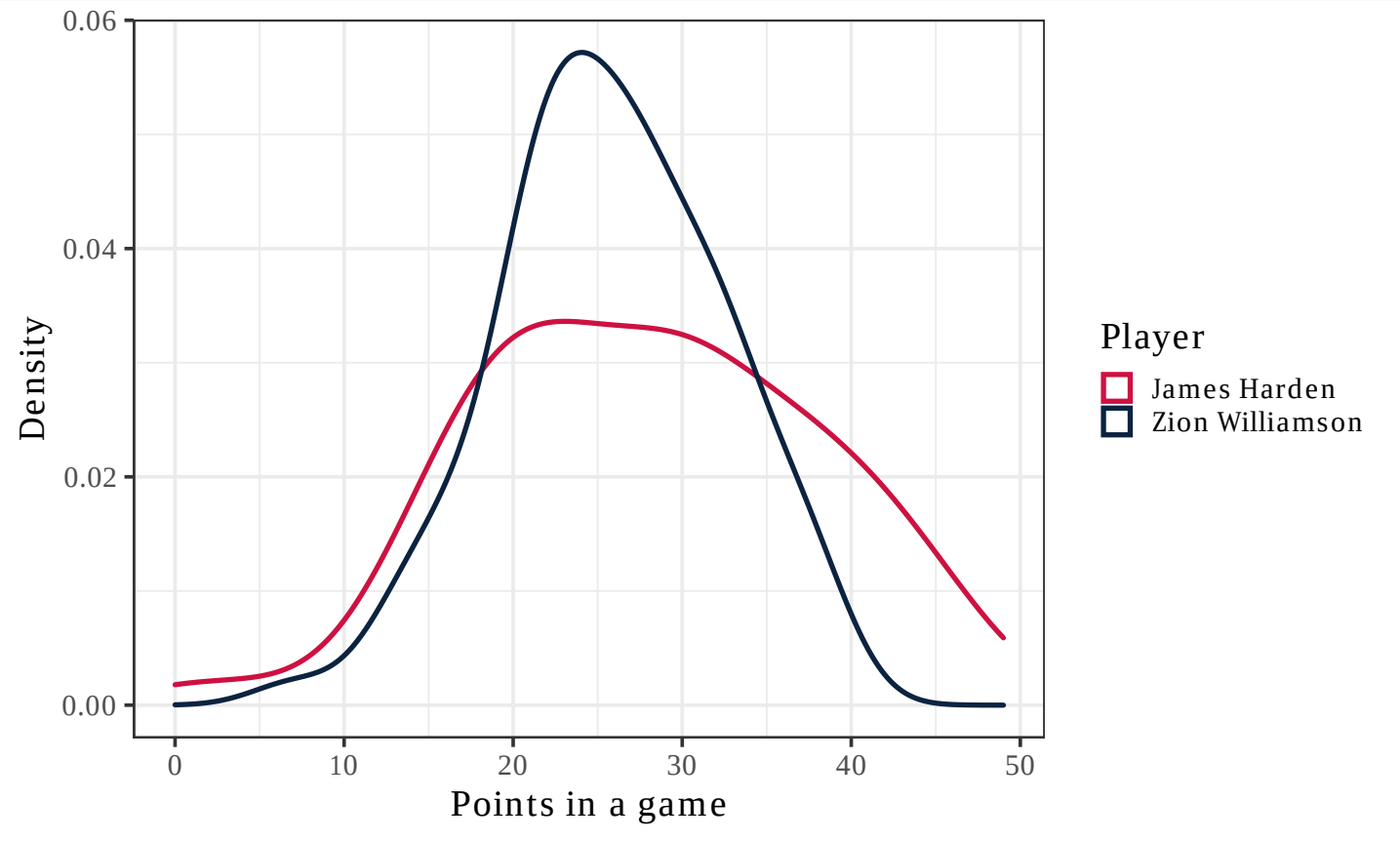
Visualizing with a boxplot

```
ggplot(zion_harden, aes(x = namePlayer, y = pts)) +
  geom_boxplot() +
  geom_jitter(width = 0.1, aes(col = namePlayer, fill = namePlayer), shape = 21, size = 2) +
  labs(x = "Player", y = "Points in a game") +
  theme_bw(base_size = 19, base_family = "Fira Sans") +
  scale_fill_manual(values =  c("#CE1141", "#0C2340")) +
  scale_color_manual(values = c("#000000", "#C8102E")) +
  guides(fill = FALSE, col = FALSE)
```



Visualizing with a density plot

```
ggplot(zion_harden, aes(x = pts, col = namePlayer)) +
  geom_density(size = 1.25) +
  labs(x = "Points in a game", y = "Density") +
  theme_bw(base_size = 19, base_family = "Fira Sans") +
  scale_color_manual(values =  c("#CE1141", "#0C2340"), name = "Player")
```



Confirming our intuition

```
zion_harden %>%
  group_by(namePlayer) %>%
  summarise(mean_pts = mean(pts),
            sd_pts = sd(pts))
```

```
## # A tibble: 2 x 3
##   namePlayer      mean_pts sd_pts
##   <chr>              <dbl>  <dbl>
## 1 James Harden        27.9  10.1
## 2 Zion Williamson     25.7   6.59
```

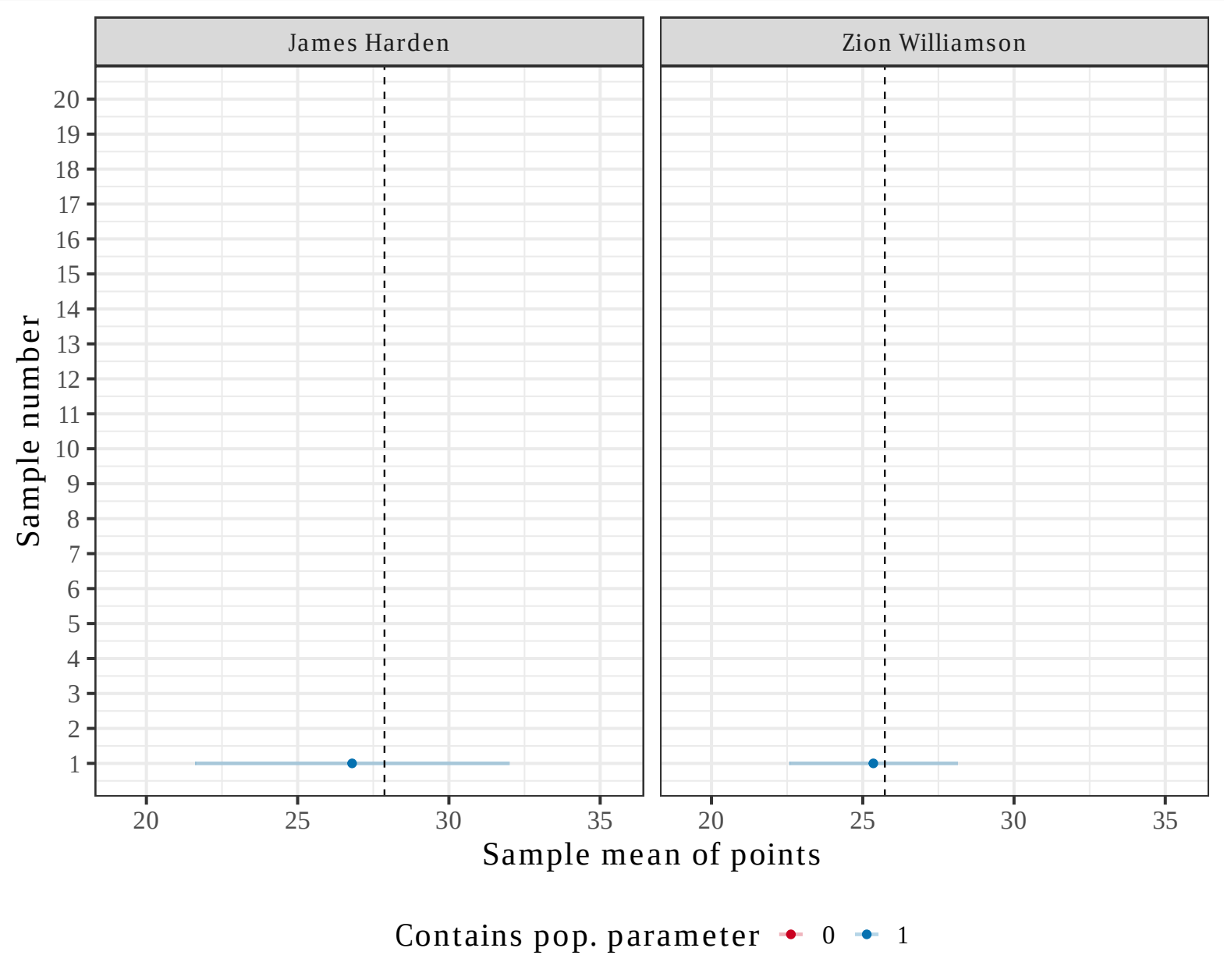
Zion scores 25.7 points per game, on average

- On any given night, he's likely to be pretty close to that

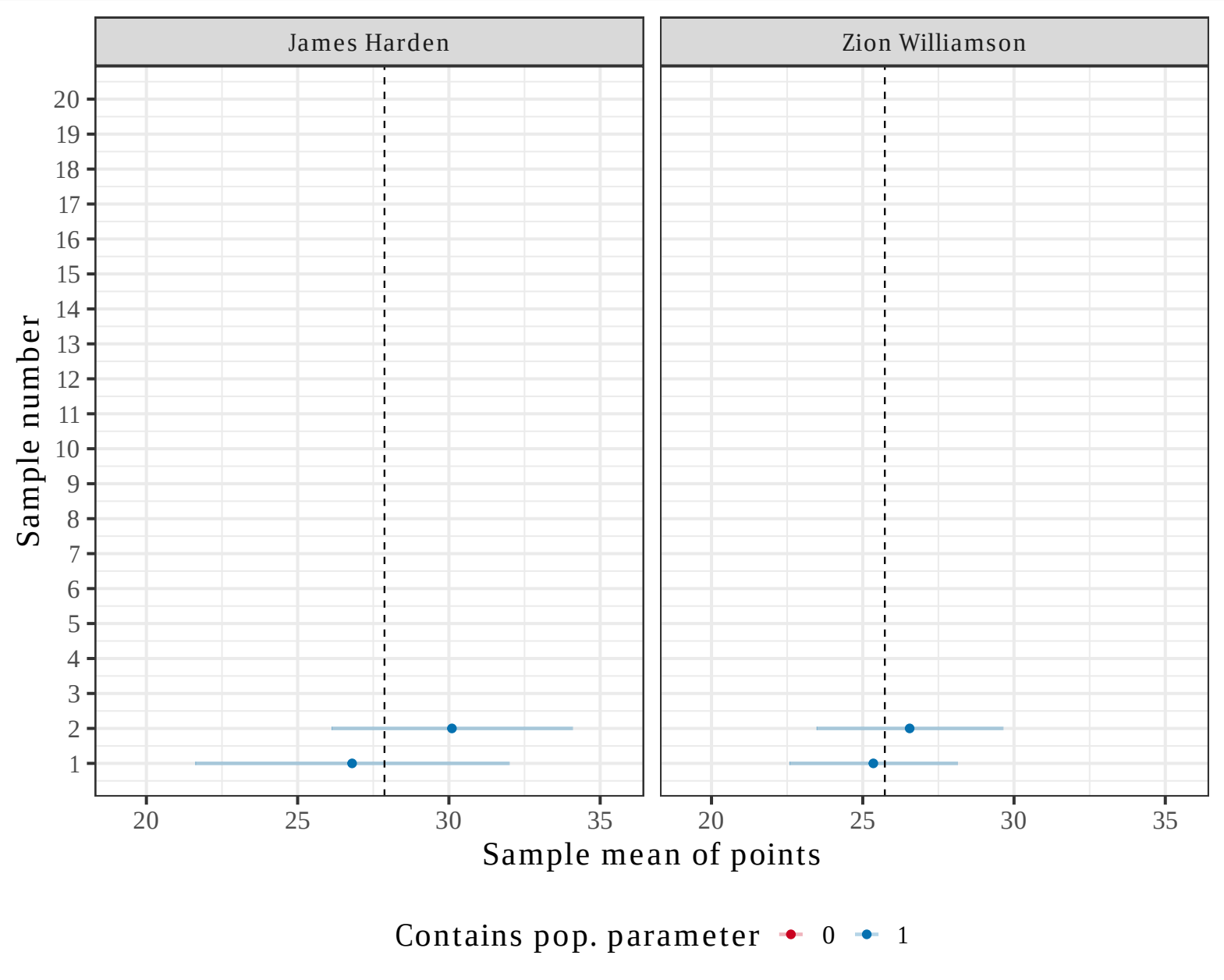
Harden scores 27.9 points per game, on average

- On any given night, he may disappear or blow up and score 50

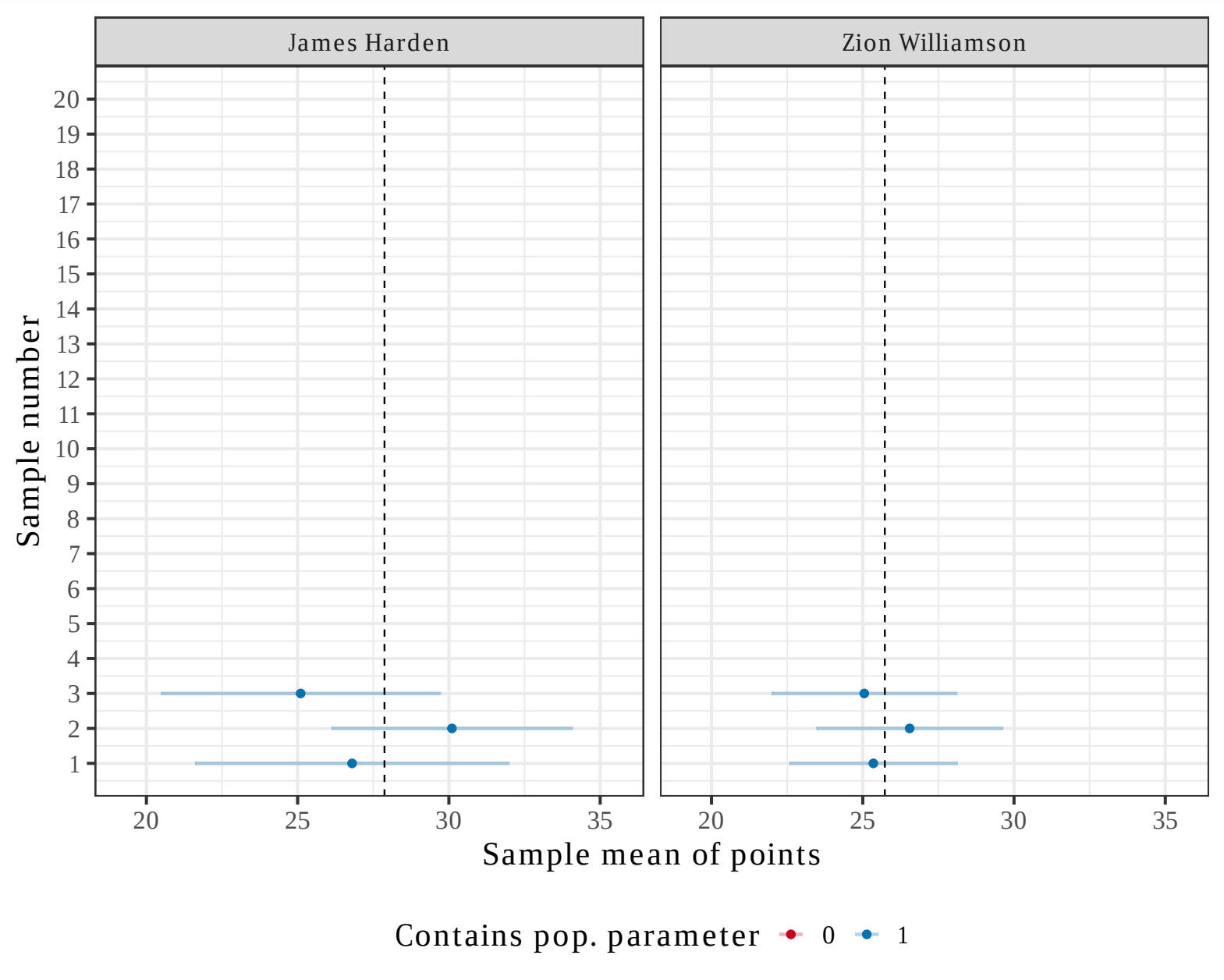
Drawing a first sample of 20 games



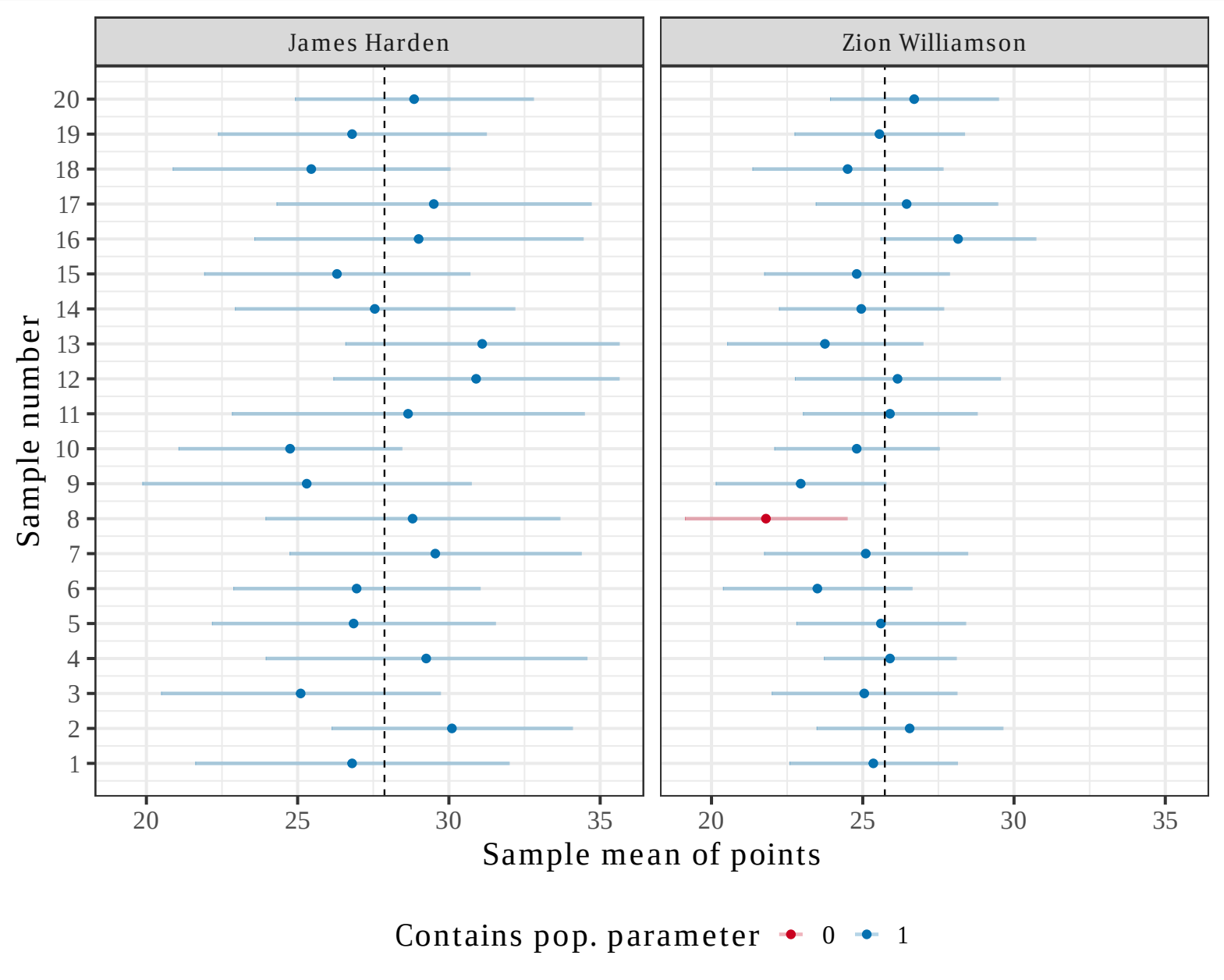
Adding a second sample of 20 games



And a third...


James Harden
Zion Williamson
Sample number
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
20
25
30
35
Sample mean of points
Contains pop. parameter
0
1

All 20 samples

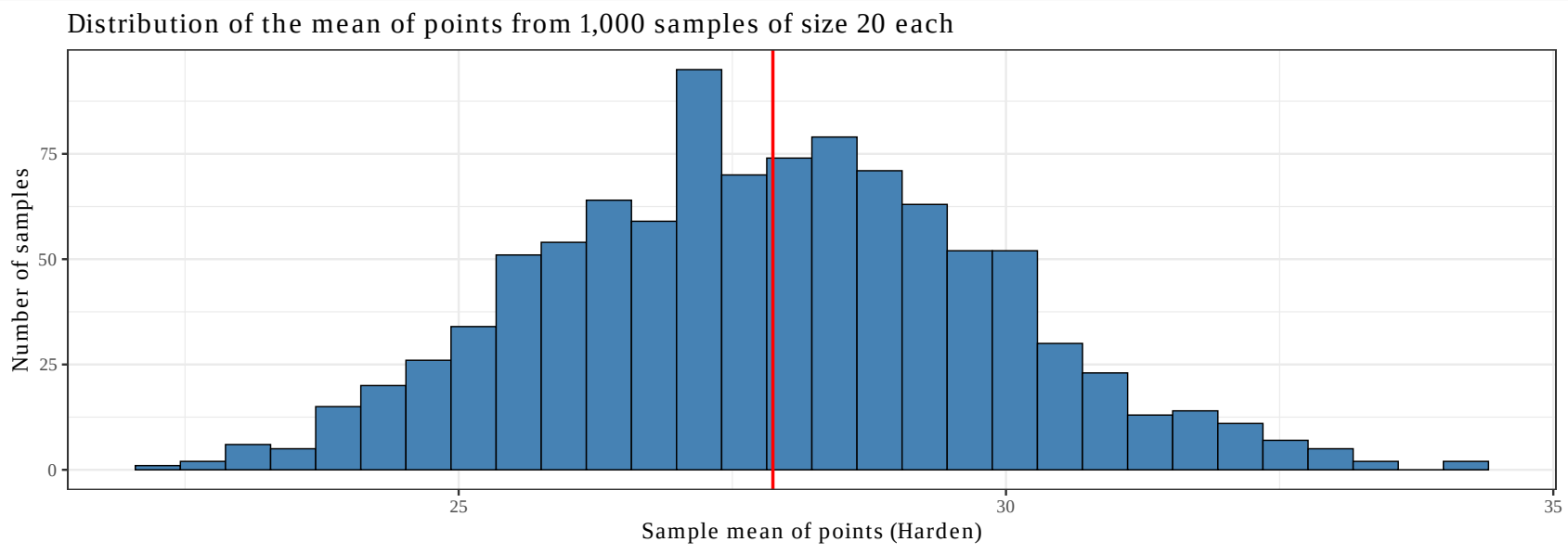


Harden vs Williamson

Harden samples have wider CIs; why?

- The standard deviation is higher for Harden
- Thus, the standard error of the mean is higher
  - $SE = \frac{\hat{\sigma}}{\sqrt{n}}$
- If I draw many samples...
  - The sampling distribution has a larger standard deviation
    - i.e. $\uparrow$ standard error of the mean

Sampling distribution for Harden




```
sd(harden_mean_20$sample20)
```

```
## [1] 2.013424
```

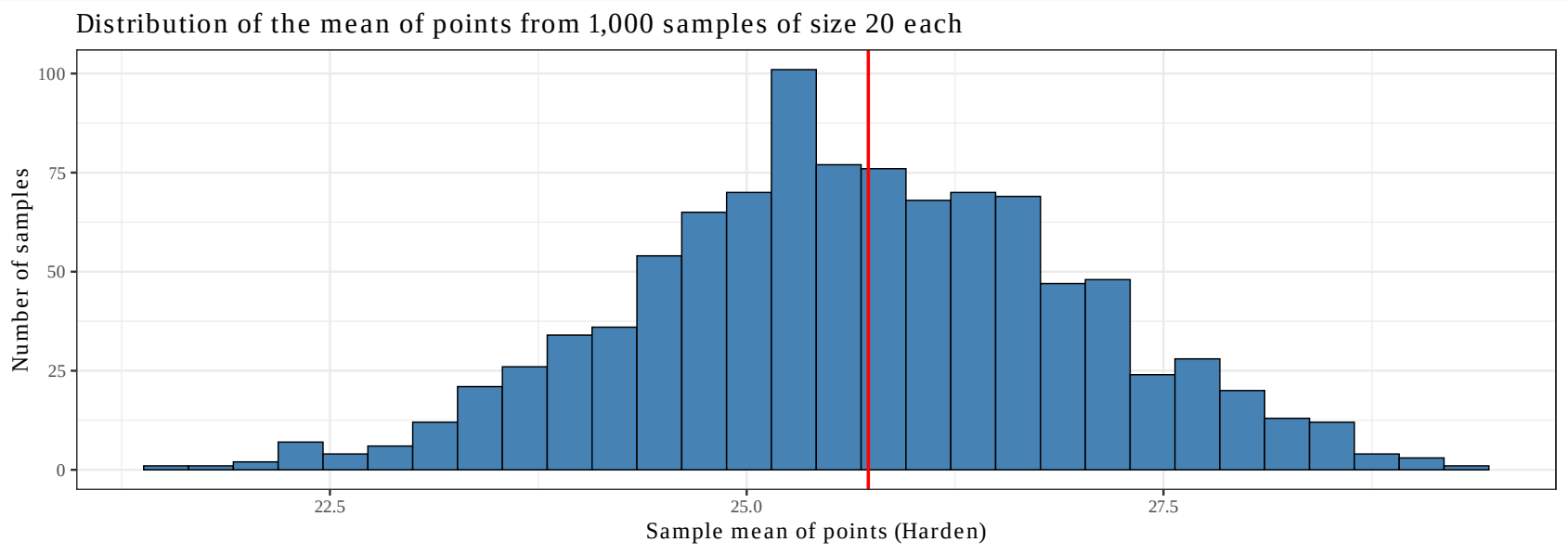
```
harden_sample <- sample_n(harden_last85, size = 20)$pts
harden_sample
```

```
##  [1] 19 34 40 29 33 25 13 44 27 26 41 32 17 31 16 47 17 32 35 39
```

```
sd(harden_sample)/sqrt(20)
```

```
## [1] 2.200807
```

Sampling distribution for Zion




```
sd(zion_mean_20$sample20)
```

```
## [1] 1.308247
```

```
zion_sample <- sample_n(zion_first85, size = 20)$pts
zion_sample
```

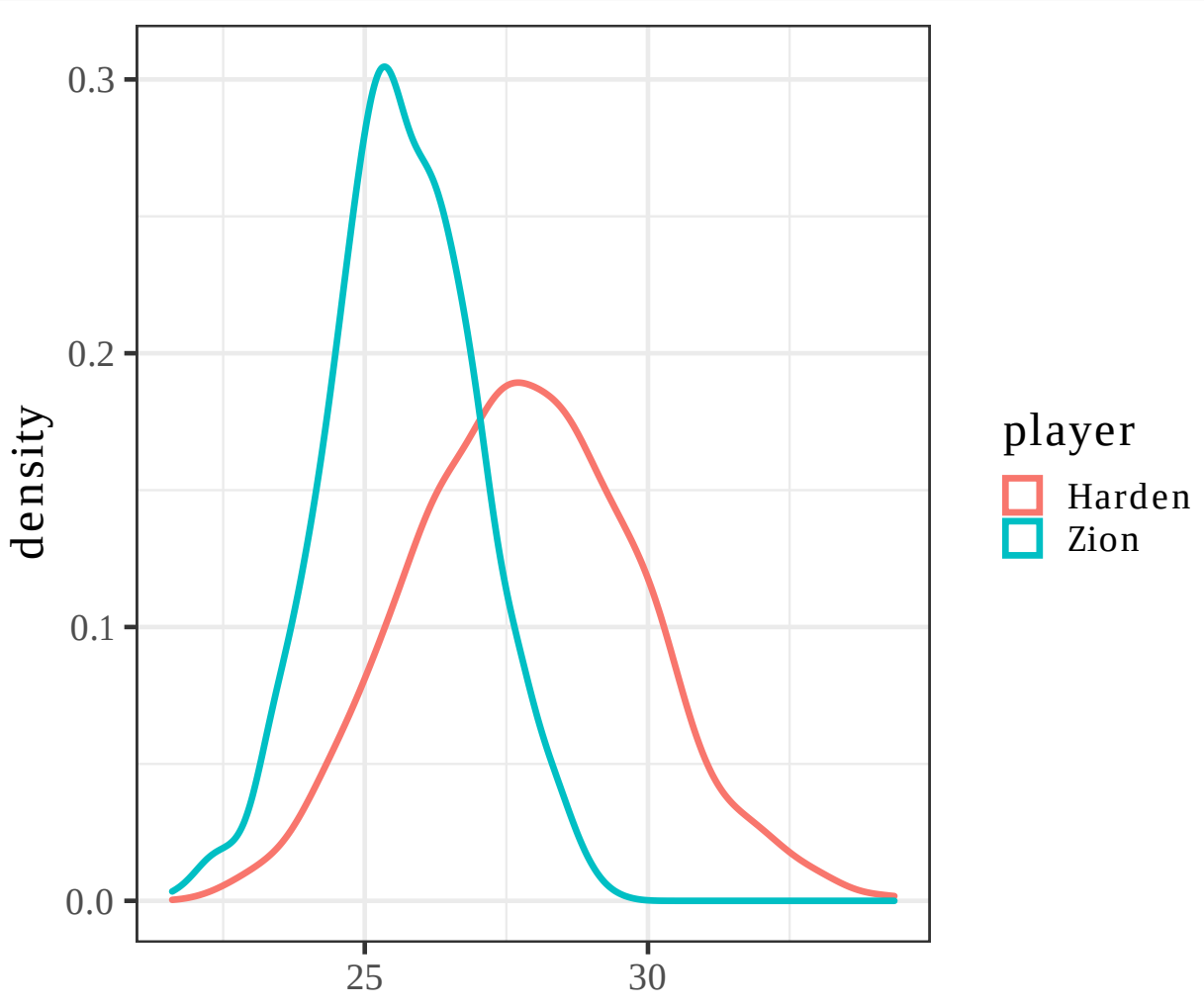
```
##  [1] 31 37 38 24 17 20 25 33 29 32 23 30 14 21 27 32 33 29 28 36
```

```
sd(zion_sample)/sqrt(20)
```

```
## [1] 1.471617
```

Two sampling distributions together

```
rbind(zion_mean_20, harden_mean_20) %>%
  ggplot(aes(x = sample20, col = player)) +
  geom_density(size = 1.25) +
  theme_bw(base_size = 19, base_family = "Fira Sans")
```

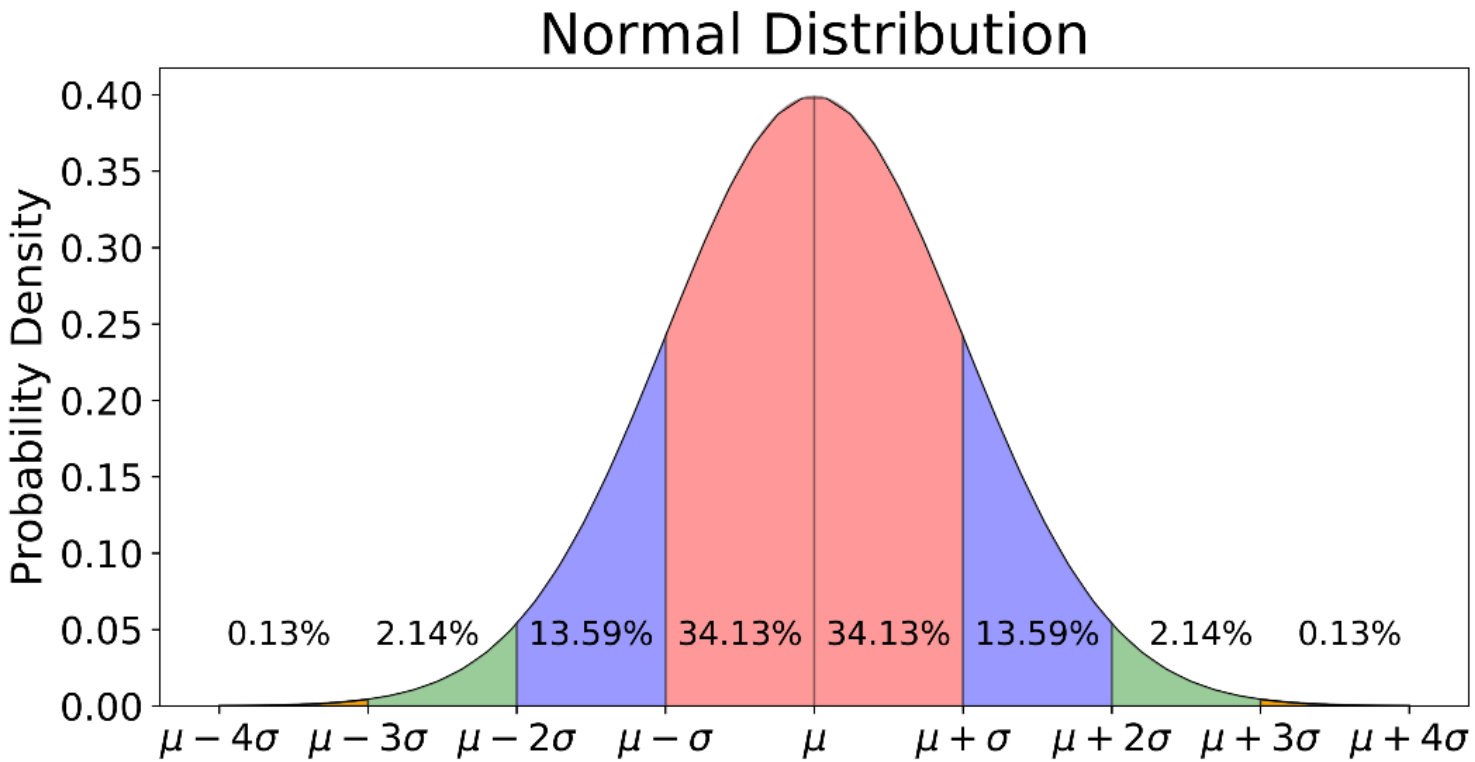


Describing distributions

The normal distribution

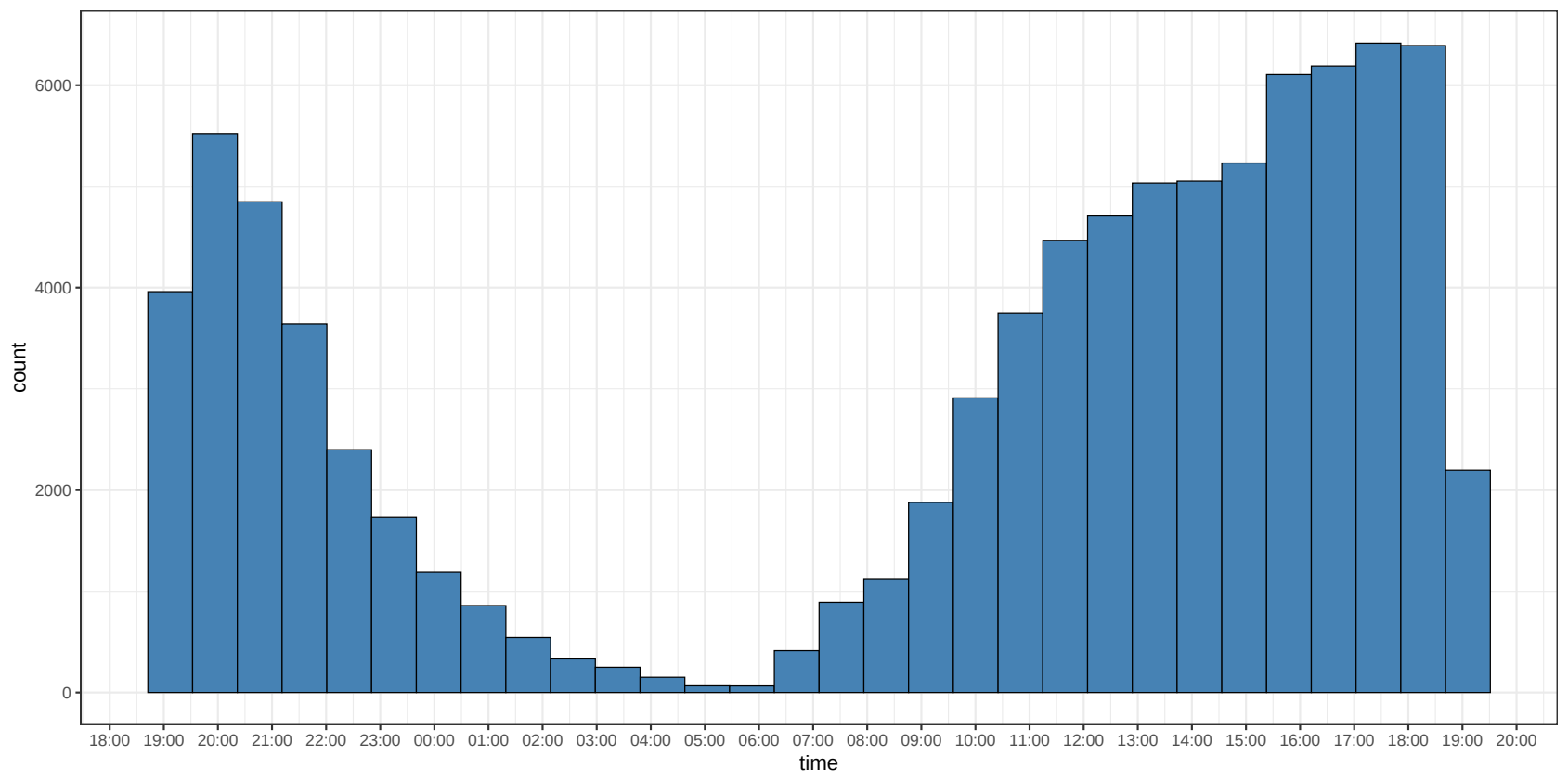
We'll start with something we've seen before: the normal distribution

- The shape is commonly known: the "bell curve"
- It is perfectly symmetrical: mean = mode = median
- Turns out, a lot of things naturally follow the normal curve!



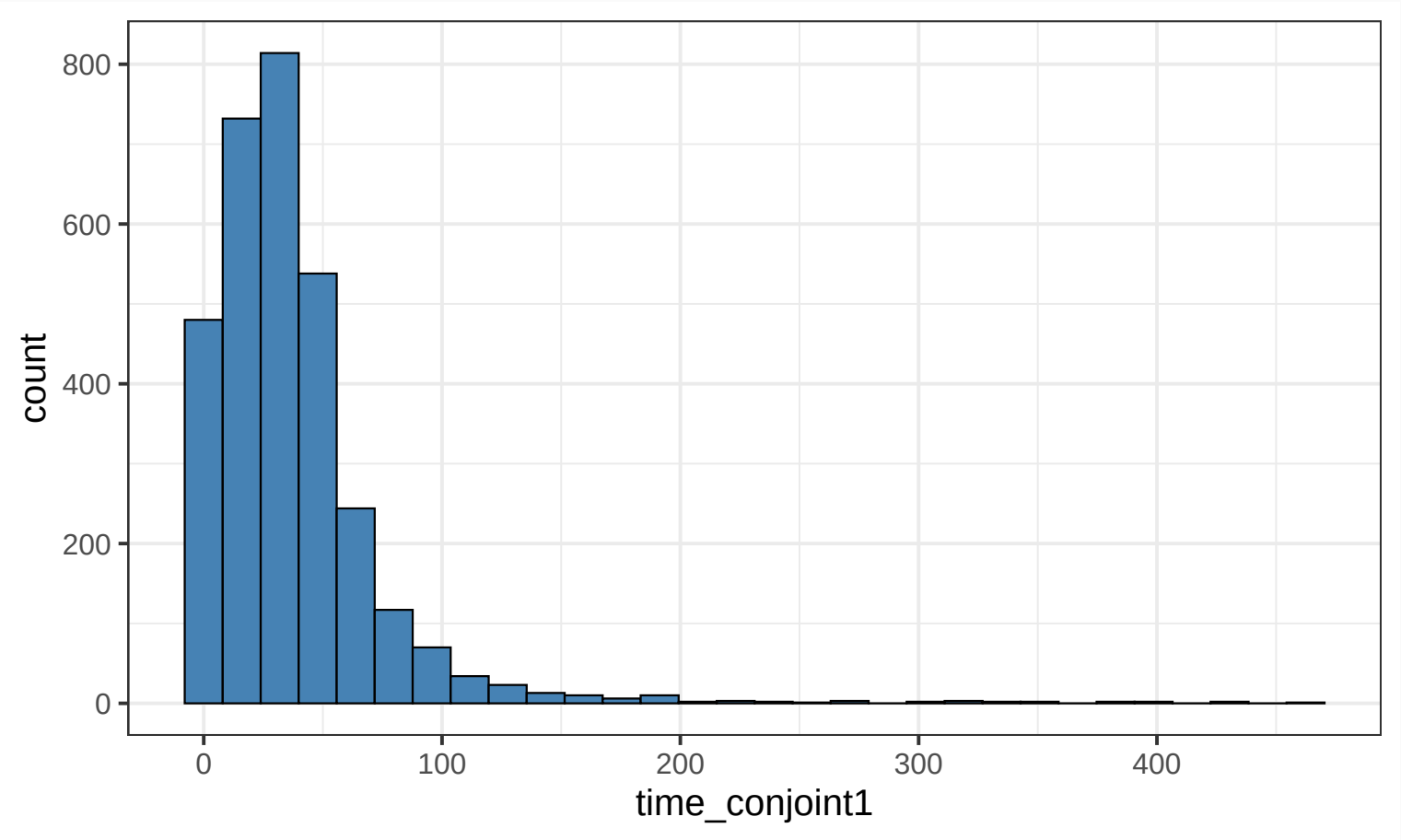
Normality in my listening habits

```
ggplot(music, aes(x = time)) +
  geom_histogram(fill = "steel blue", col = "black") +
  scale_x_datetime(breaks = scales::breaks_width("60 min"),
                   date_labels = "%H:%S") +
  theme_bw(base_size = 19)
```



Skewed distributions: right-skew

```
load("lectures/lecture_9.1/survey.RData")
survey_500s <- subset(survey_full, time_conjoint1 < 500)
ggplot(data = survey_500s, aes(x = time_conjoint1)) +
  geom_histogram(fill = "steel blue", col = "black") +
  theme_bw(base_size = 19)
```



Skewed distributions: right-skew

Skewed distributions are asymmetric

- In the example above, the right-tail is much longer
  - It's a **right-skewed** distribution
  - Also called a positively-skewed distribution
- Mean $\neq$ Median

Skewed distributions: right-skew

Skewed distributions are asymmetric

- In the example above, the right-tail is much longer
  - It's a **right-skewed** distribution
  - Also called a positively-skewed distribution
- Mean $\neq$ Median

```
mean(survey_500s$time_conjoint1)
```

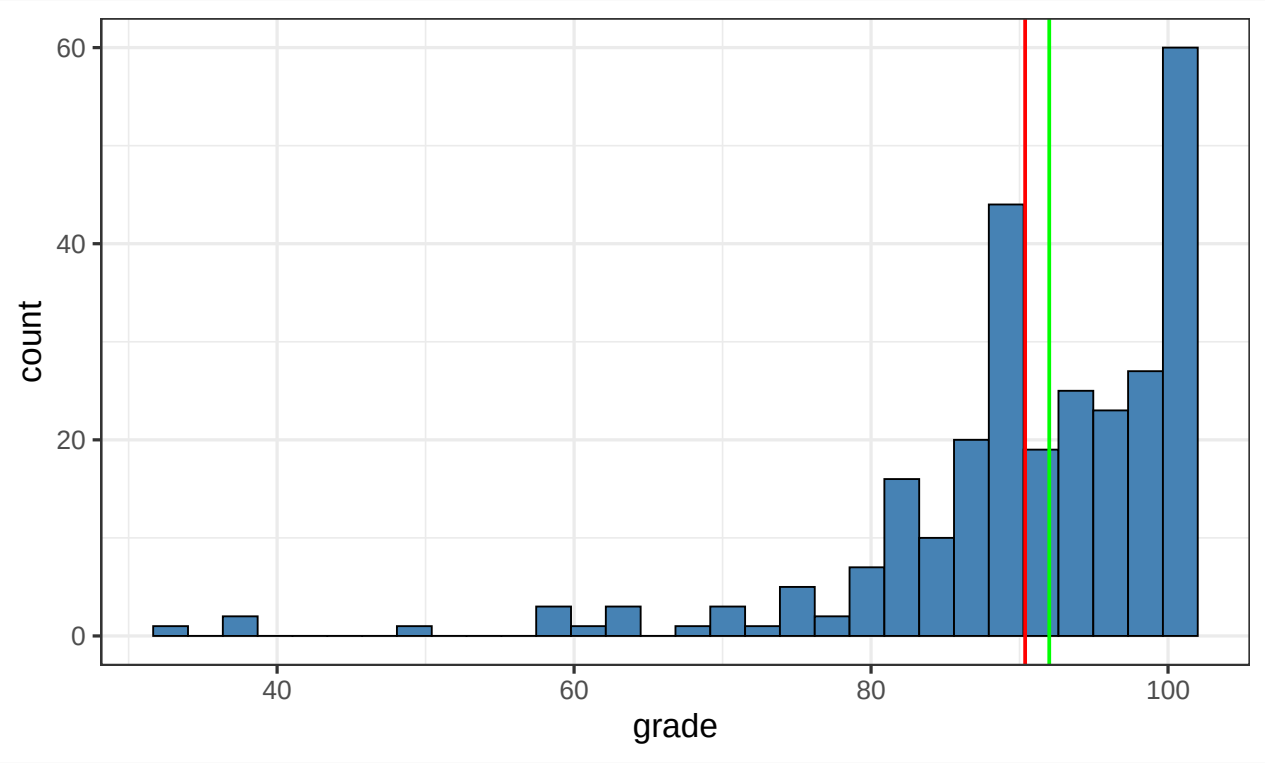
```
## [1] 37.72603
```

```
median(survey_500s$time_conjoint1)
```

```
## [1] 30.319
```

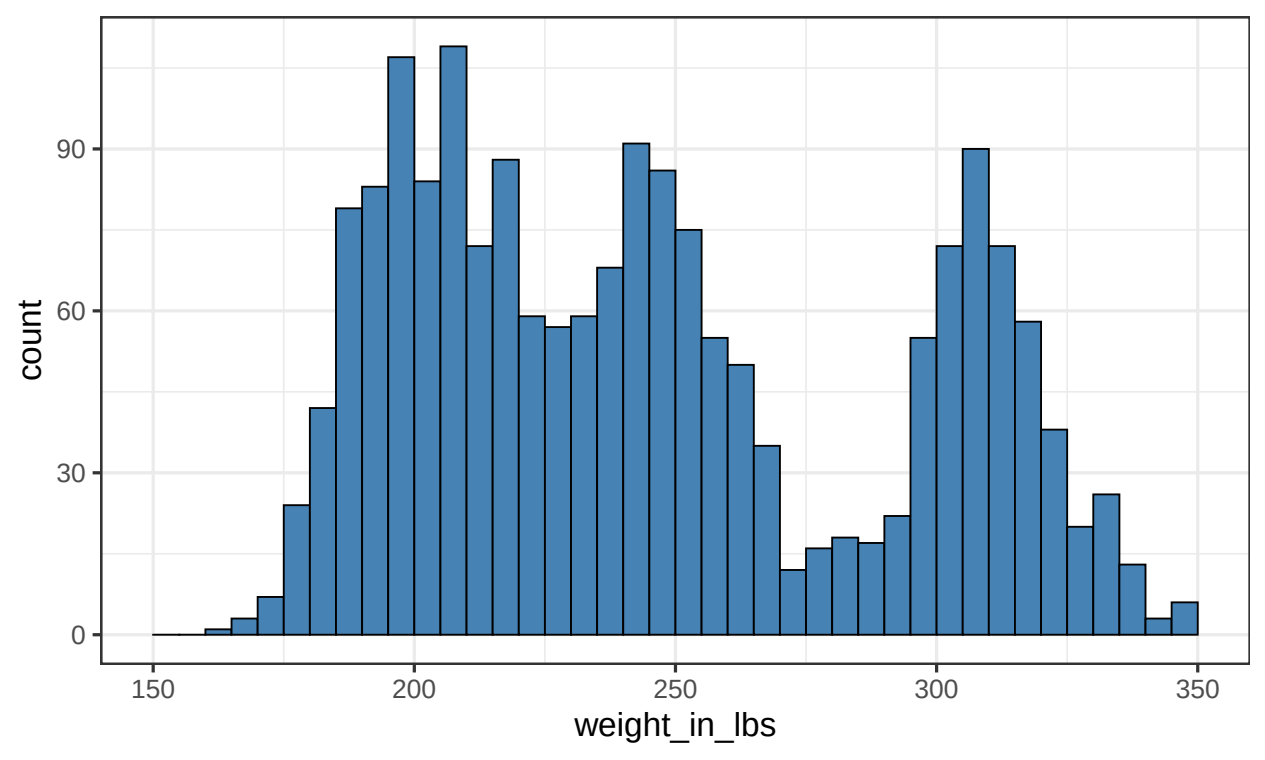
Skewed distributions: left-skew

```
grades <- read.csv("lectures/lecture_9.2/pset1_grades.csv")
ggplot(grades, aes(x = grade)) +
  geom_histogram(fill = "steel blue", col = "black") +
  theme_bw(base_size = 19) +
  geom_vline(xintercept = mean(grades$grade,na.rm = T),
             col = "red", size = 1) +
  geom_vline(xintercept = median(grades$grade,na.rm = T),
             col = "green", size = 1)
```

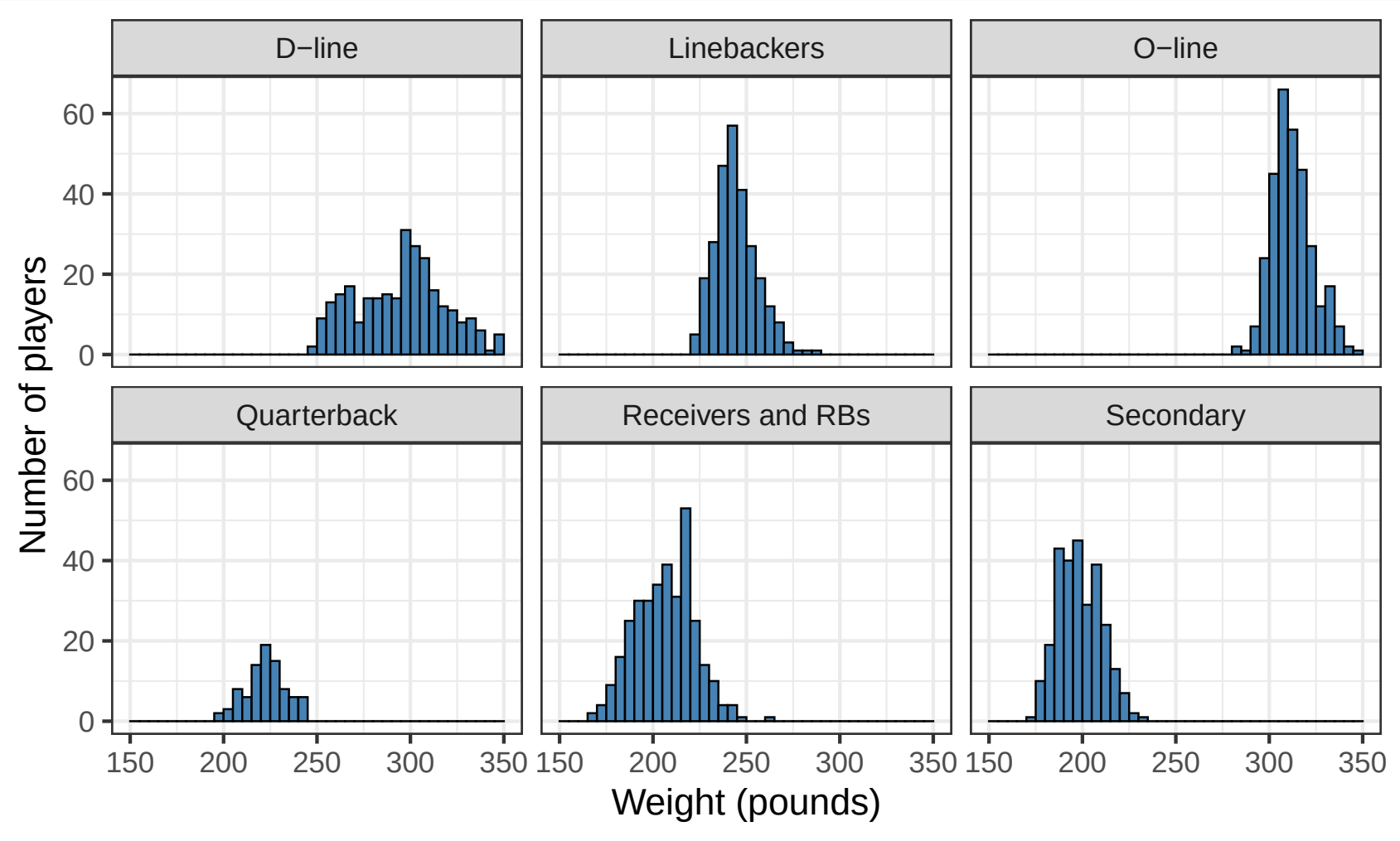


Multimodal distributions: NFL players

```
nfl <- read.csv("lectures/lecture_9.2/nfl_height_weight.csv")
ggplot(data = nfl, aes(x = weight_in_lbs)) +
  geom_histogram(breaks = seq(150, 350, 5),
                 fill = "steel blue", col = "black") +
  theme_bw(base_size = 19)
```



Multimodal distributions



Skewed distribution: what kind of skewness?

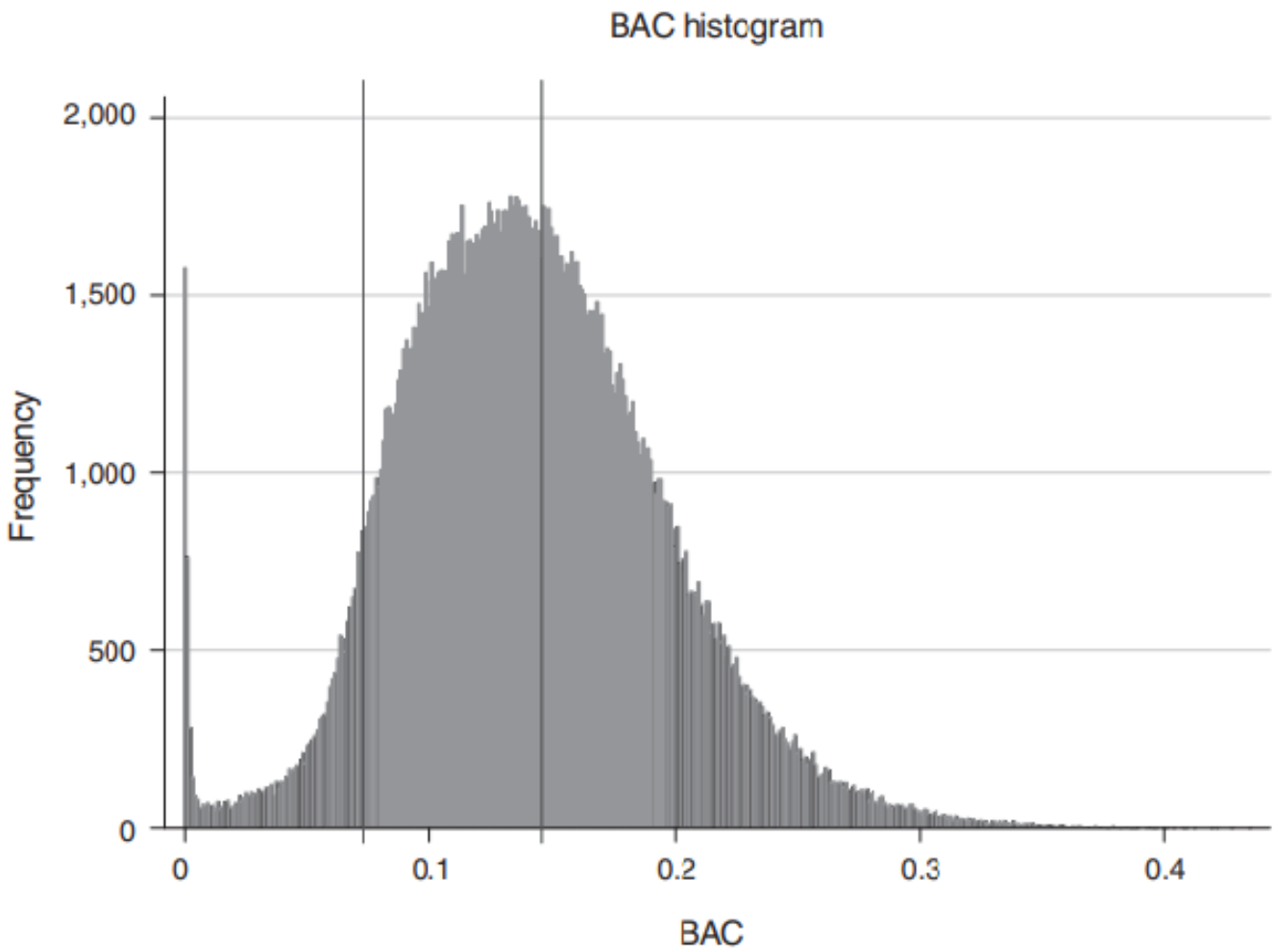


FIGURE 1. BAC DISTRIBUTION

Notes: Based on administrative records from the Washington State Impaired Driver Testing Program, 1999–2007. The histogram height on the vertical axis is based on frequency of observations, with BAC on the horizontal axis. The vertical black lines represent the two legal thresh-

References i

Hansen, Benjamin. 2015. “Punishment and Deterrence: Evidence from Drunk Driving.” American Economic Review 105 (4): 1581–1617. https://doi.org/10.1257/aer.20130189.