The building blocks of statistics

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

January 15, 2025

Recap

This week + next: Conceptual foundations of statistical analysis of causality

Last time we discussed:

1. Causal statements as counterfactual statements
2. Ingredients of a causal analysis
   - Unit of analysis and population of interest
   - Outcome variable to explain
   - Treatment variable we suspect affects outcome (versus some comparison group)

Today’s agenda

Correlation does not imply causation

… but all our statistical evidence about causation is built on correlations

So we need to understand the basics of statistical correlations

1. Population mean (expectation) and sample mean
2. Variance and standard deviation
3. Covariance and correlation
4. Conditional means and differences in means

⚠ **This is gonna require some math**

So be sure to ask questions if you’re confused about anything! I’m here to help!

Motivating example

Mostly we'll be talking about abstract stats today

Question to keep in mind if you get lost in the abstraction

What's the relationship between watching Tucker Carlson's show and opinions of Donald Trump among registered voters in 2020?

⚠ **This is not a causal question**

We're not (yet) asking what is the effect of watching Tucker on one's opinion toward Trump. We'll get there soon!

One-variable statistics

Population versus sample

We are interested in a **population**[1] of units, $i = 1, \ldots, \mathcal{N}$ ("fancy N")

- → e.g., all U.S. registered voters in 2020

We only have a **sample** from the population, $i = 1, \ldots, N$

- → e.g., the 7845 voters interviewed for the 2020 ANES

For now, no assumptions on how the sample was drawn

Random sampling is important for **inference** — drawing conclusions about the population from a sample

Expected value

A **variable** Y_i is a numerical attribute of population unit i

- → e.g., registered voter i's opinion of Trump on a 0–100 scale

The **expected value** of a variable, $\mathbb{E}[Y_i]$, is its average in the population

- → e.g., average opinion of Trump among all registered voters in 2020

Expected value for binary variables

We often code "yes"/"no" variables like "Whether the respondent watches Tucker Carlson" using 1 = yes, 0 = no.

For these **binary variables**, the expected value is the **proportion** (between 0 and 1) of the population with a "yes".

e.g., if $\mathbb{E}[Y_i] = 0.043$, that means we have "yes" for 4.3% of the population and "no" for the remaining 95.7%.

The mathematics of expected value

The mathematical formula:

$$\mathbb{E}[Y_i] = \frac{1}{\mathcal{N}}[Y_1 + Y_2 + \cdots + Y_{\mathcal{N}}] = \frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}} Y_i$$

Linearity of expectation

Important property: If X_i and Y_i are variables, and a and b are constants,

$$\mathbb{E}[aX_i + bY_i] = a\mathbb{E}[X_i] + b\mathbb{E}[Y_i].$$

Sample mean

Without data on the full population, we don’t know the expected value

But we can always calculate the **sample mean** of the data we have

We’ll use $\textbf{avg}[Y_i]$, aka $\bar{Y}$, to denote the sample mean of Y_i

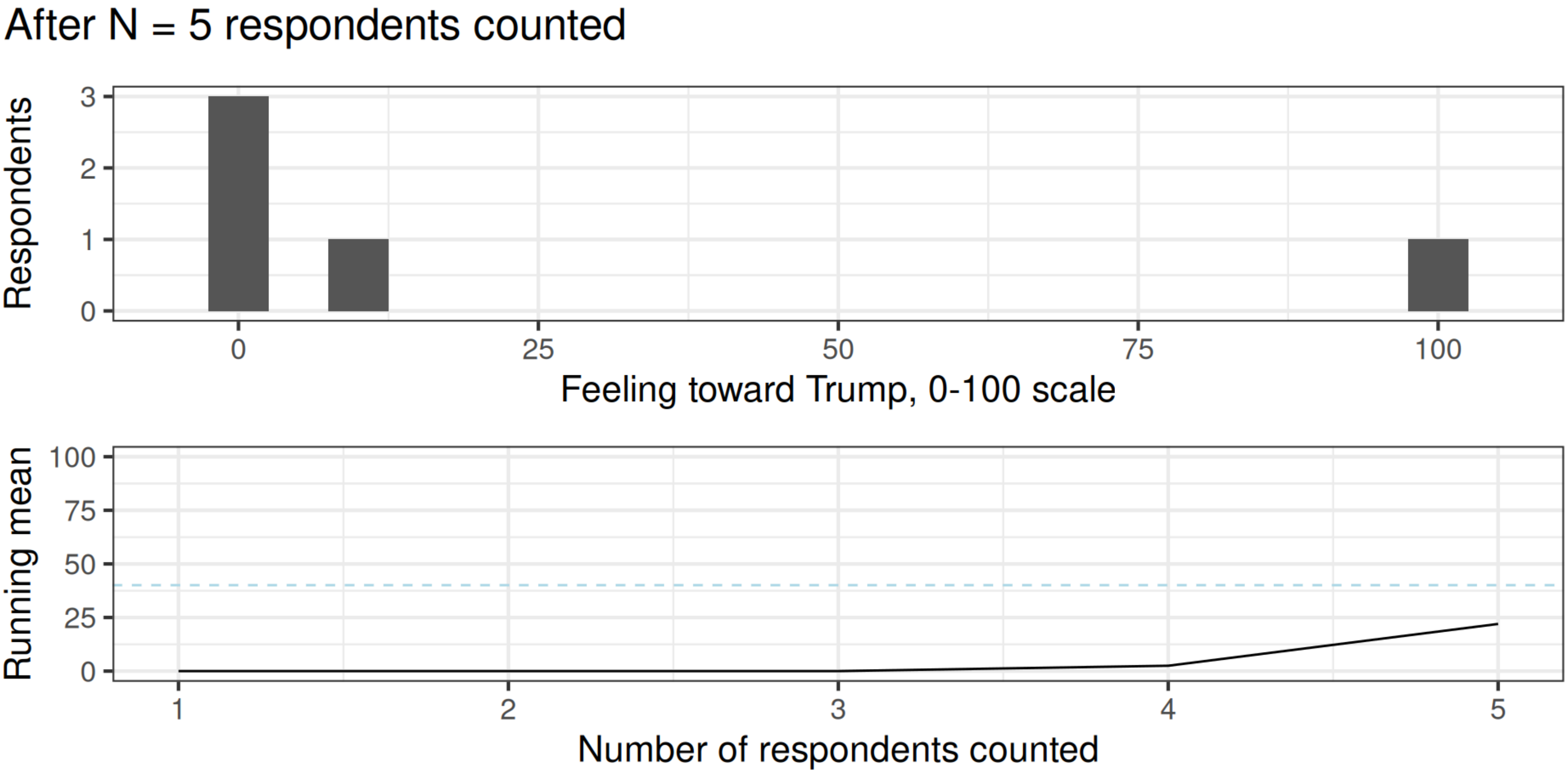
$$\textbf{avg}[Y_i] = \bar{Y} = \frac{1}{N}[Y_1 + Y_2 + \cdots + Y_n] = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Same as $\mathbb{E}[Y_i]$, but summing over observed data instead of full population

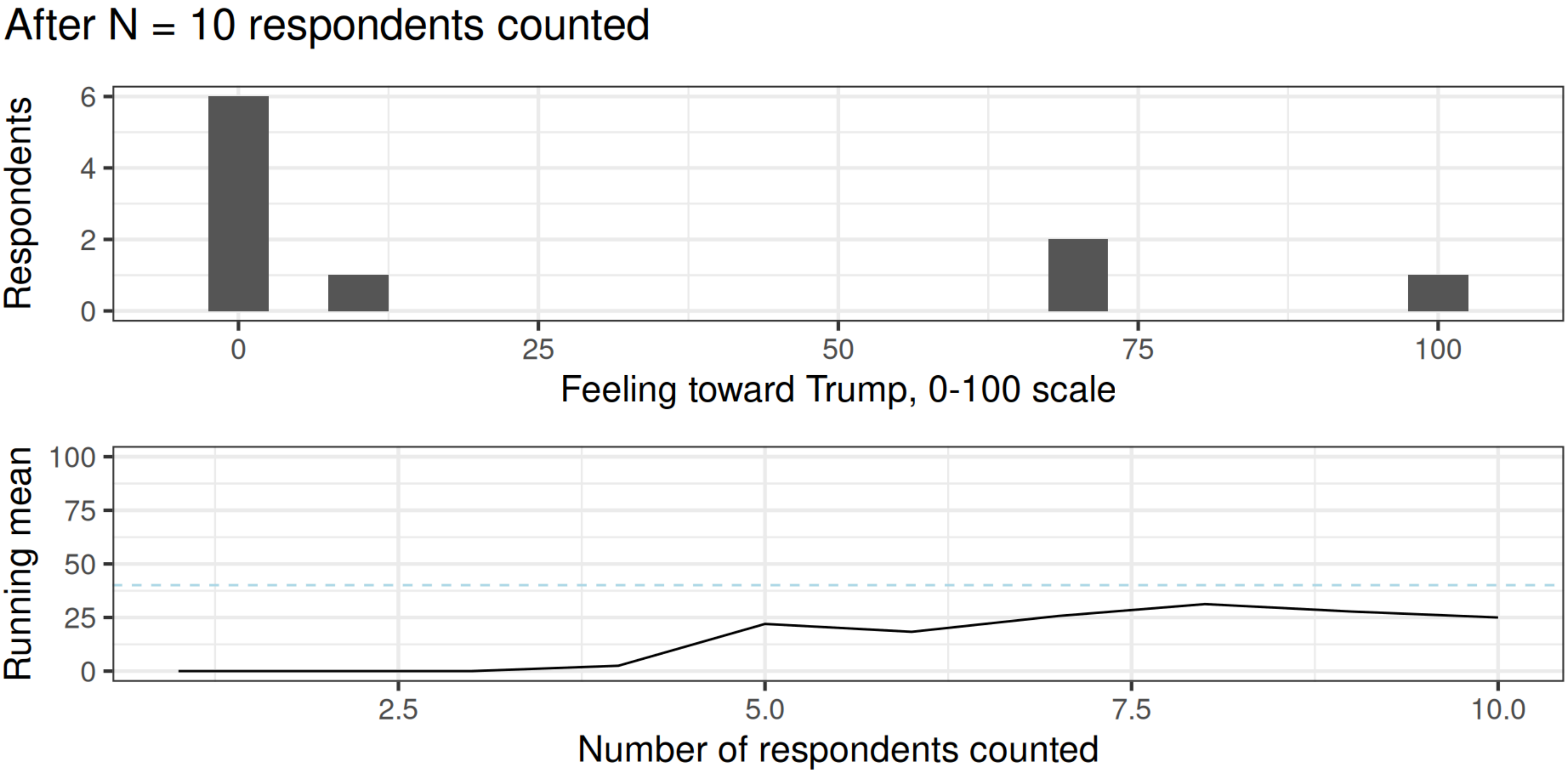
Linearity of sample mean

Just like with the expected value, $\textbf{avg}[aX_i + bY_i] = a\,\textbf{avg}[X_i] + b\,\textbf{avg}[Y_i]$.

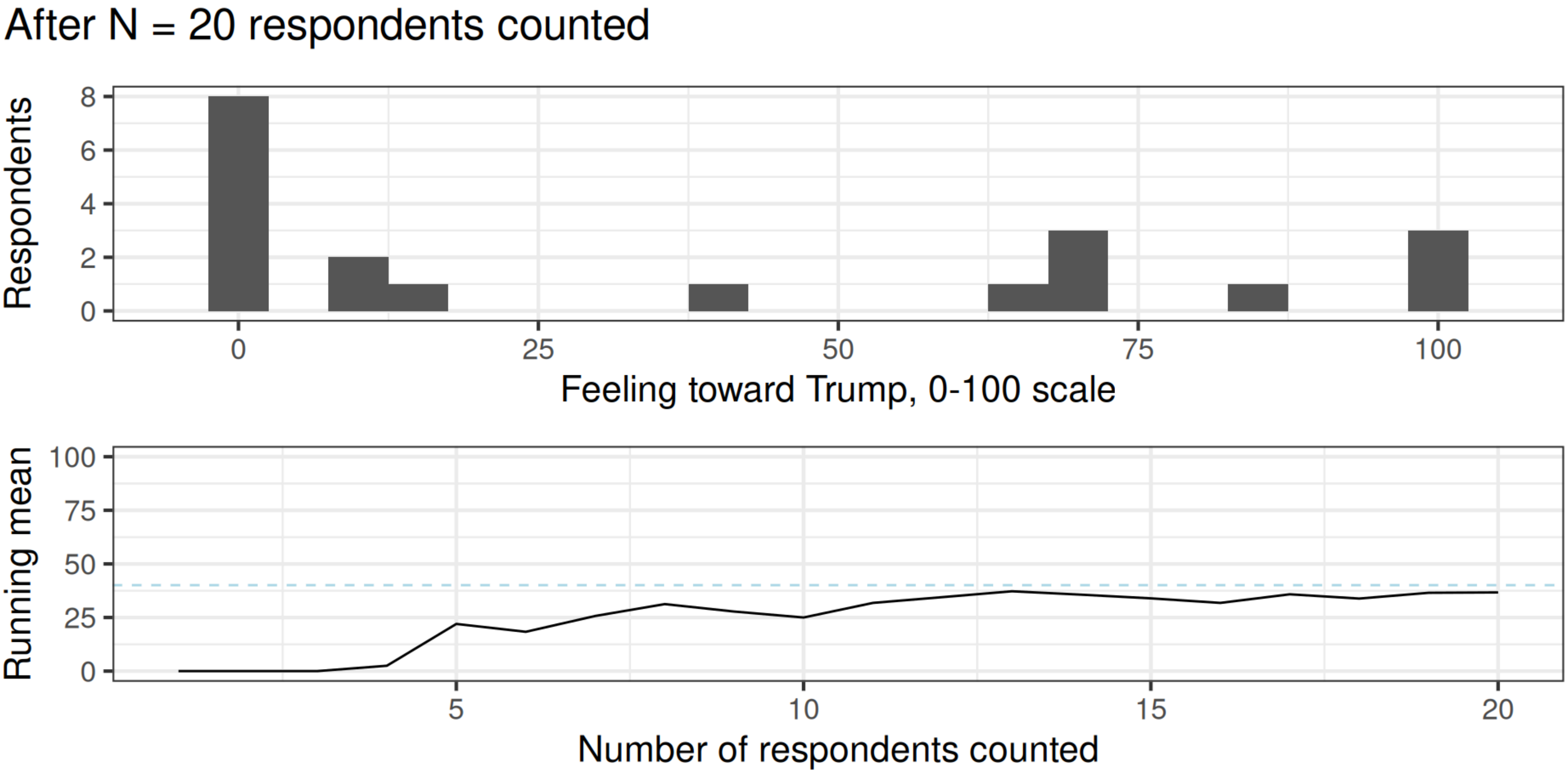
Stability of the sample mean



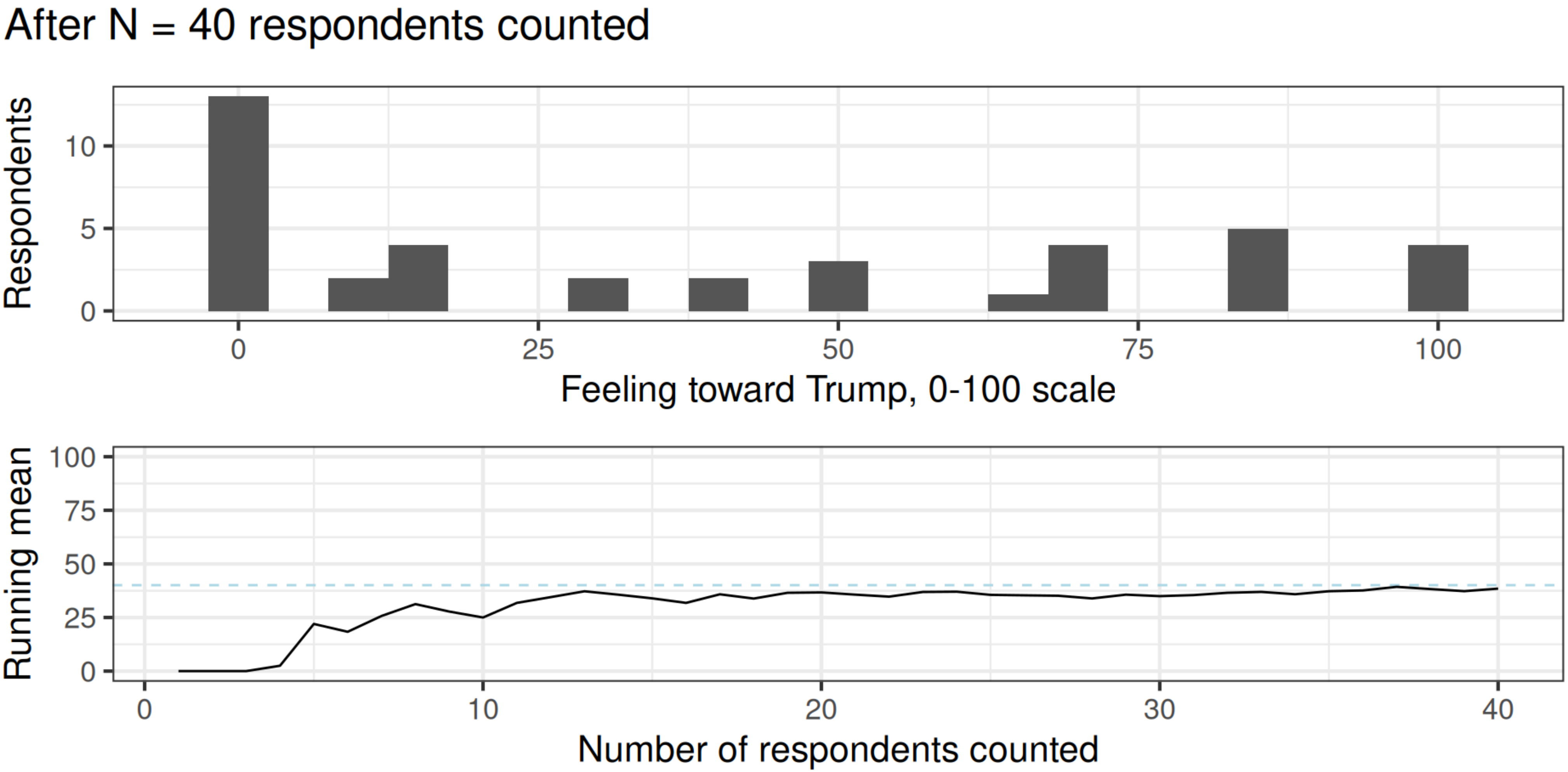
Stability of the sample mean



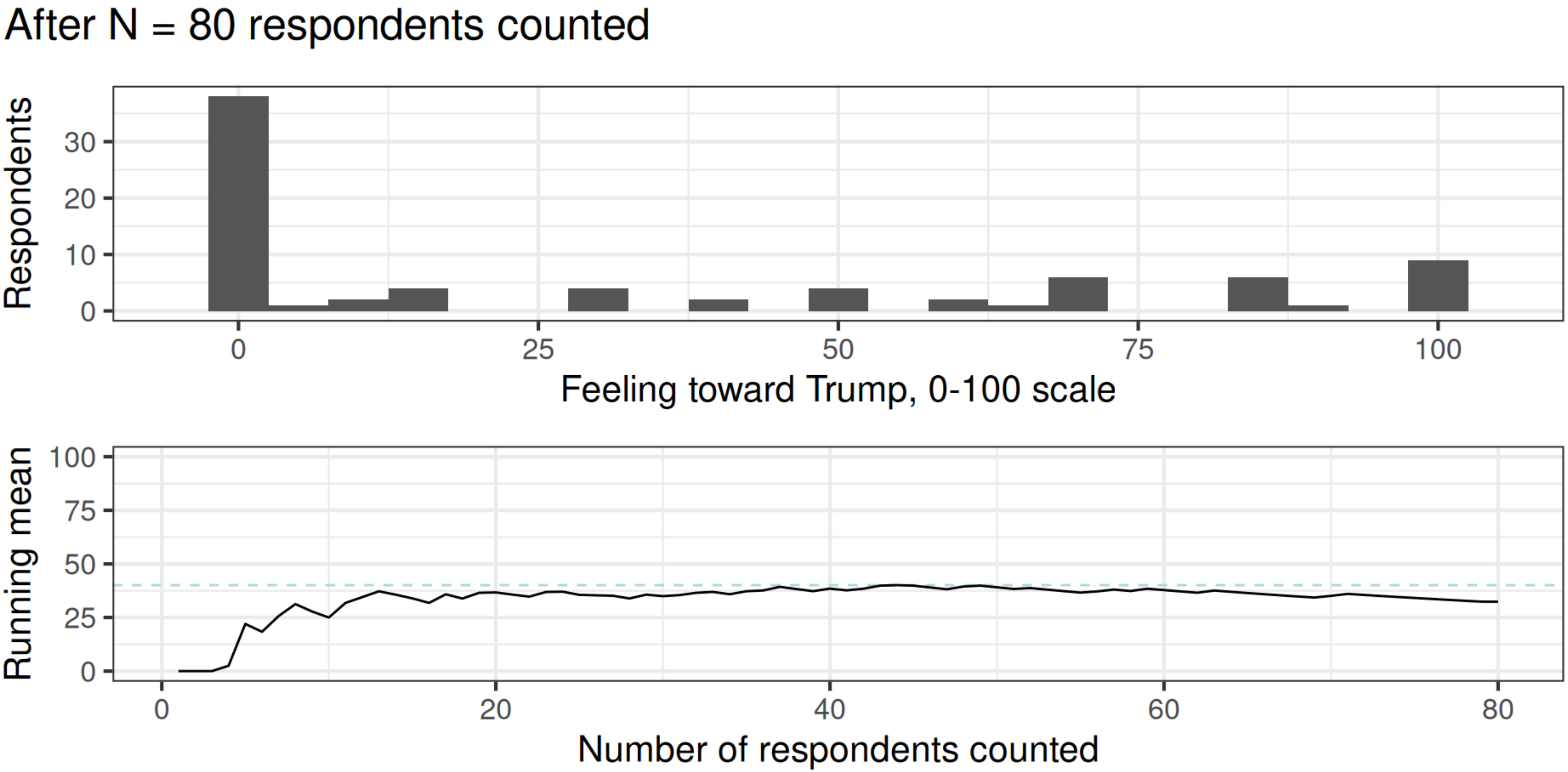
Stability of the sample mean



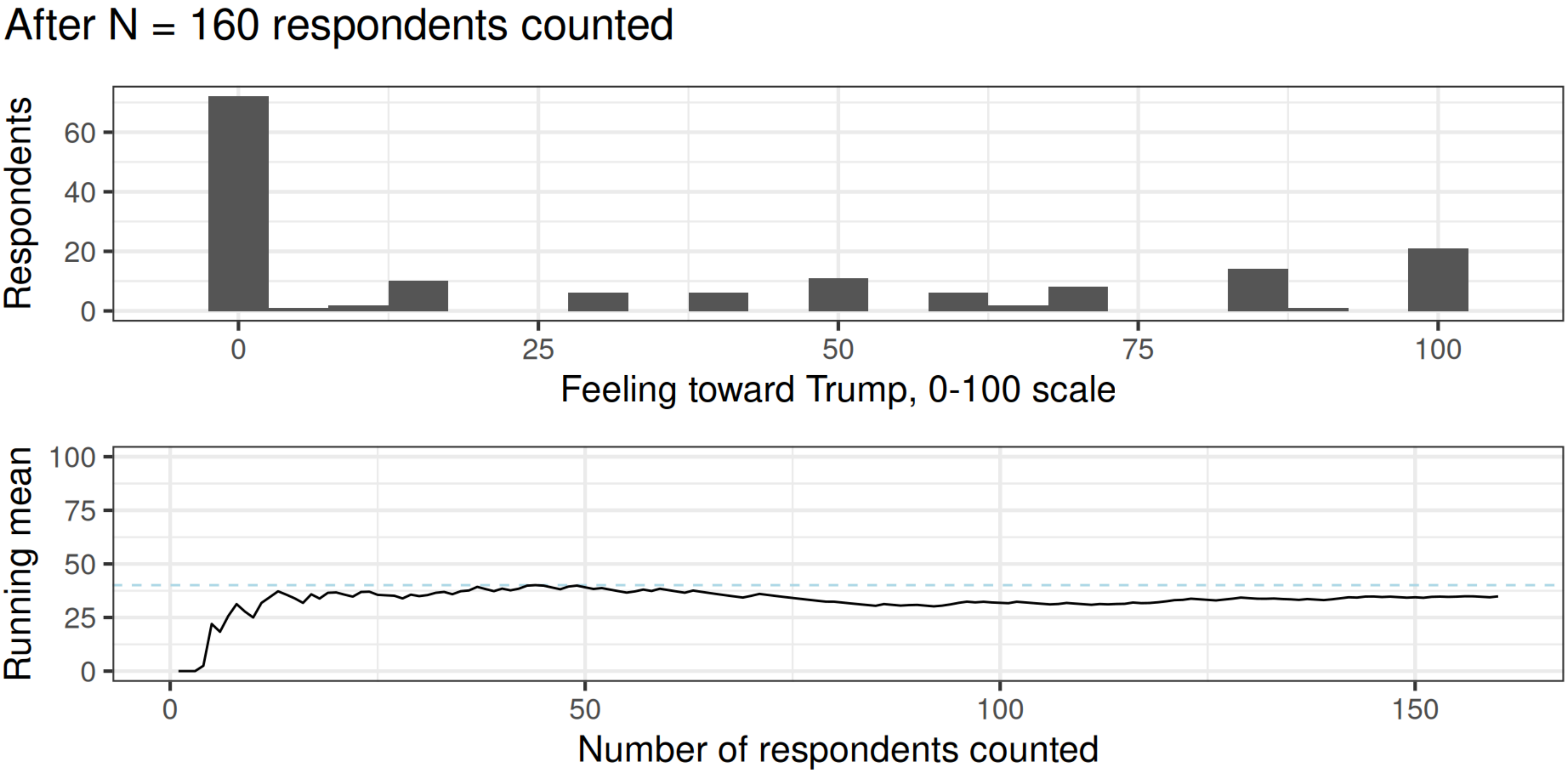
Stability of the sample mean



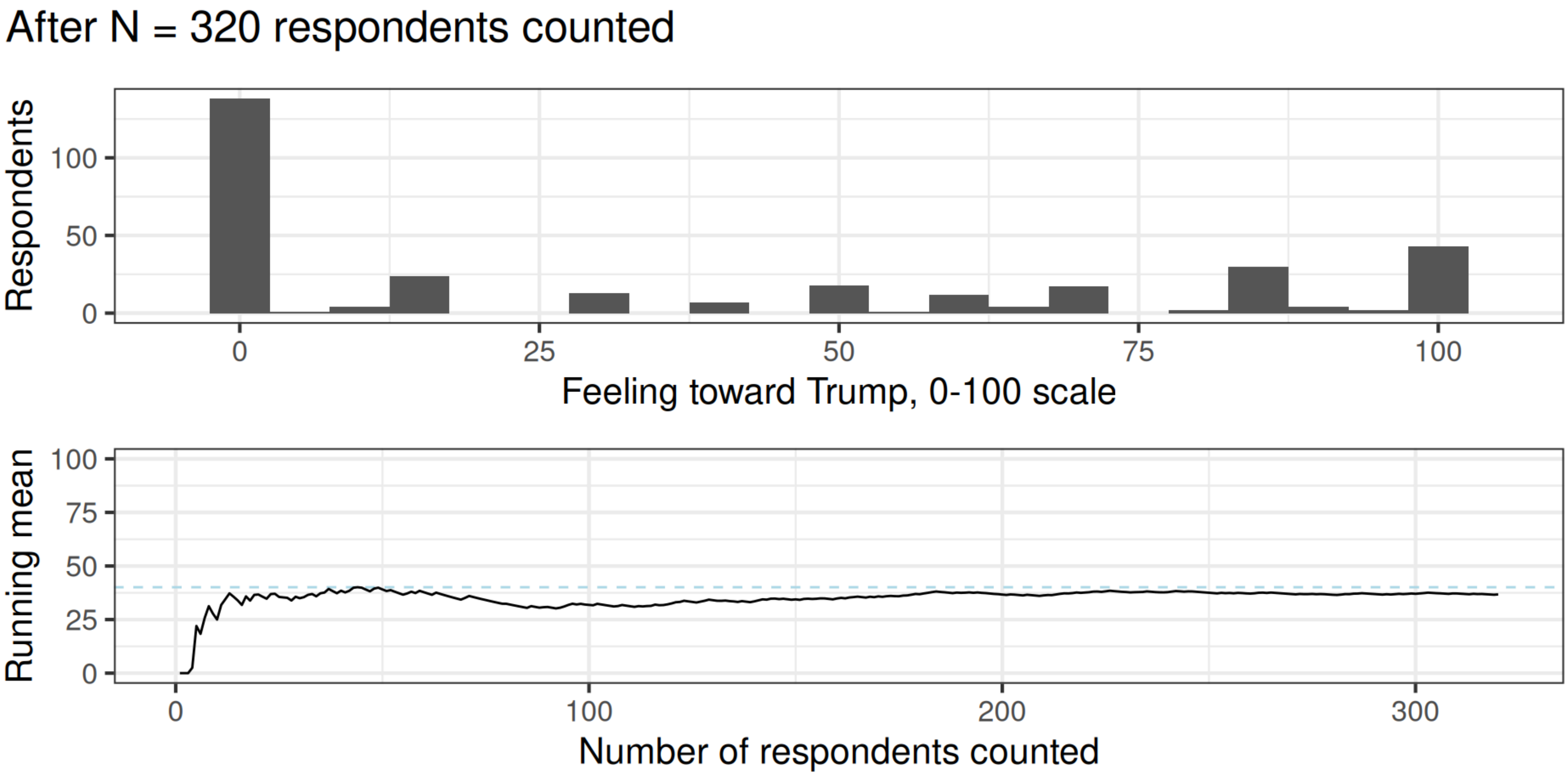
Stability of the sample mean



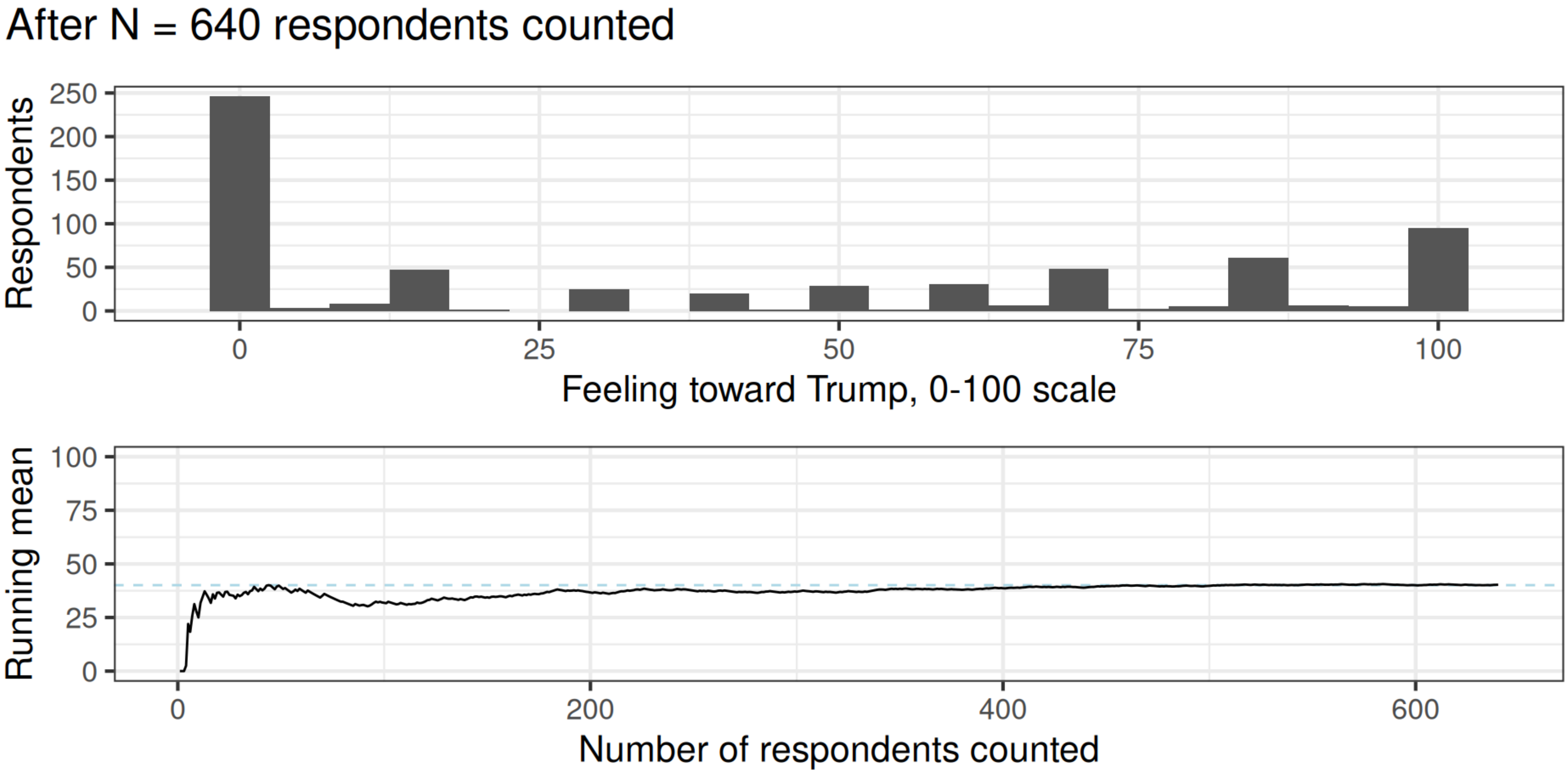
Stability of the sample mean



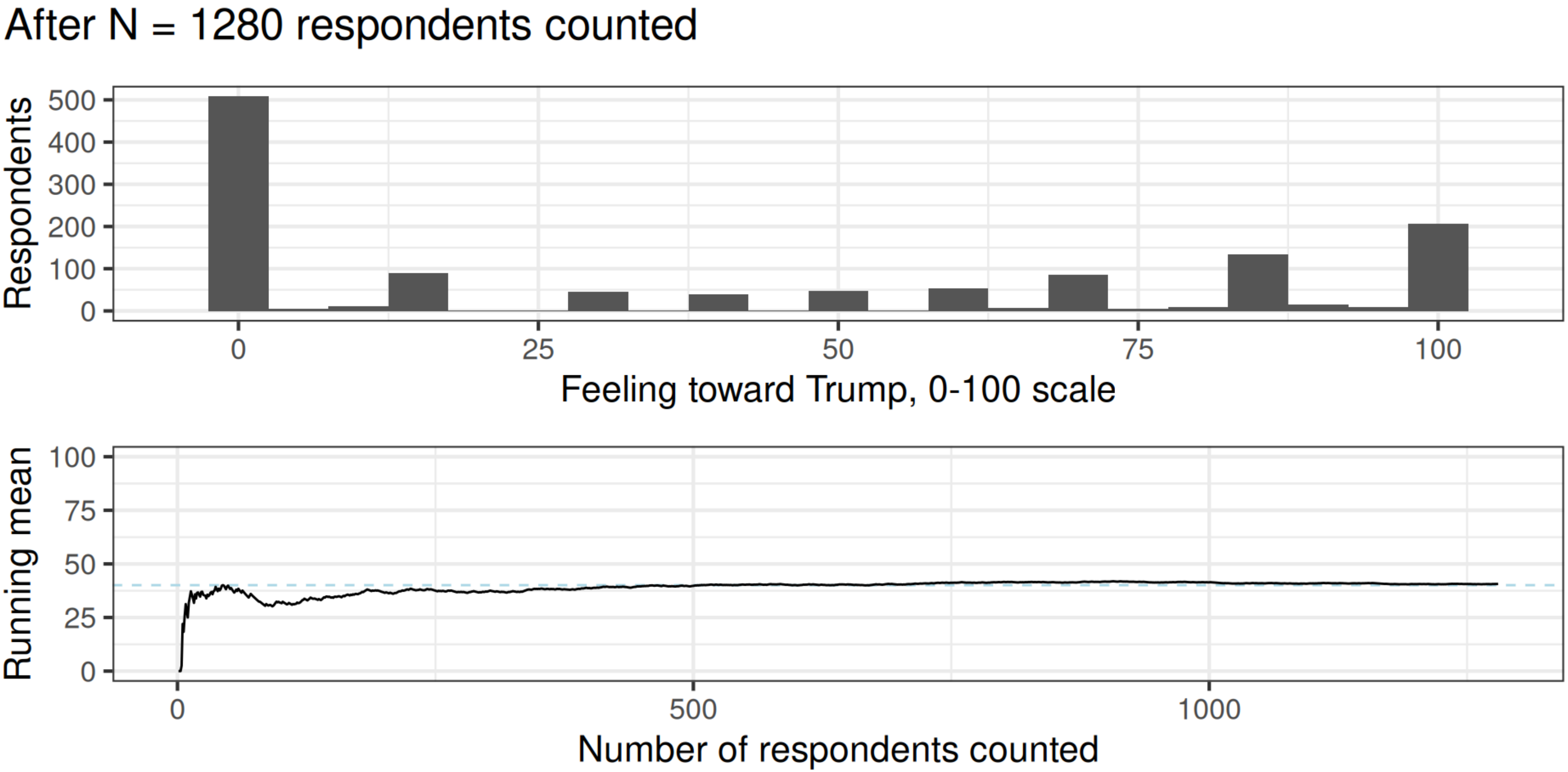
Stability of the sample mean



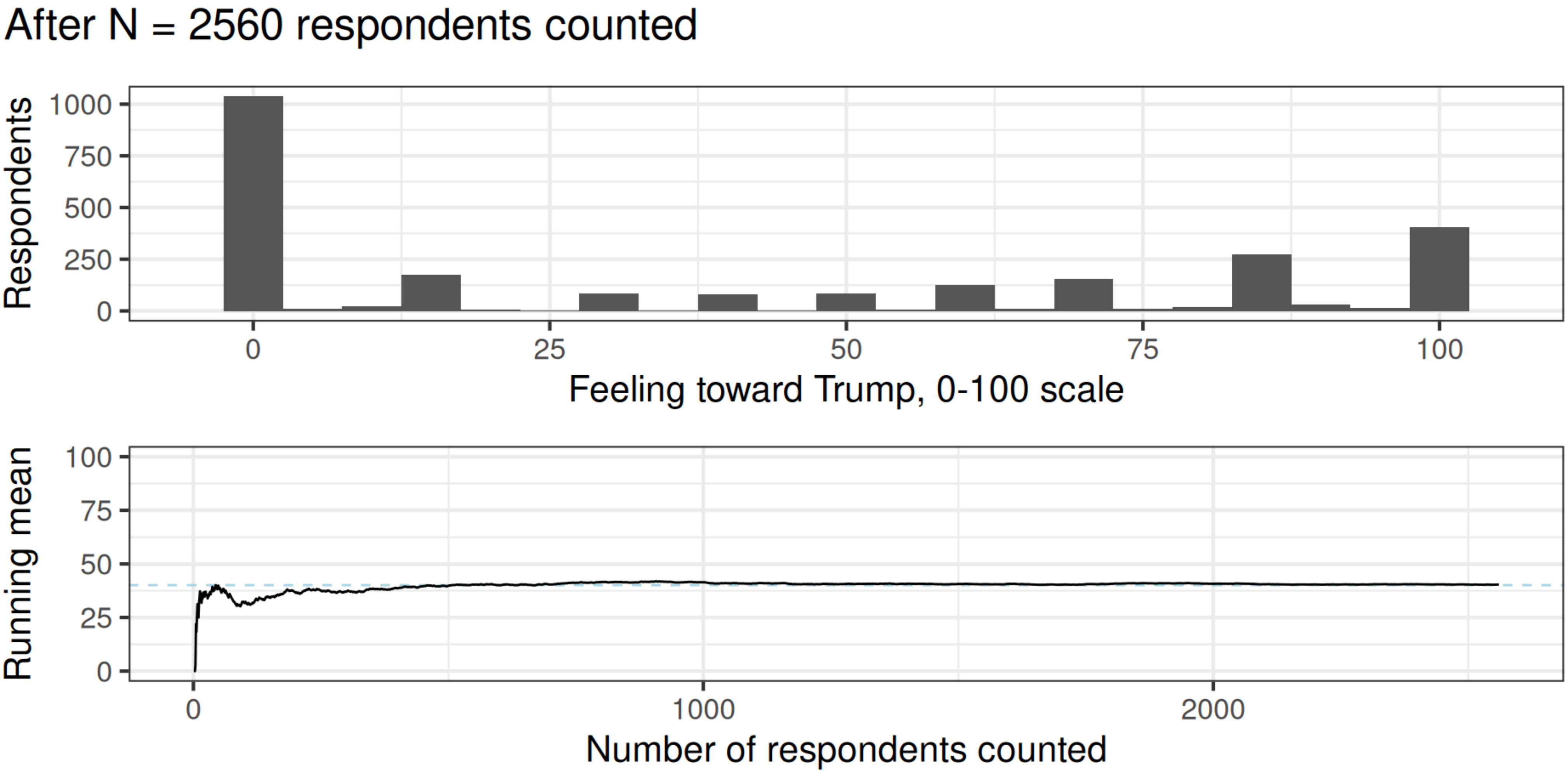
Stability of the sample mean



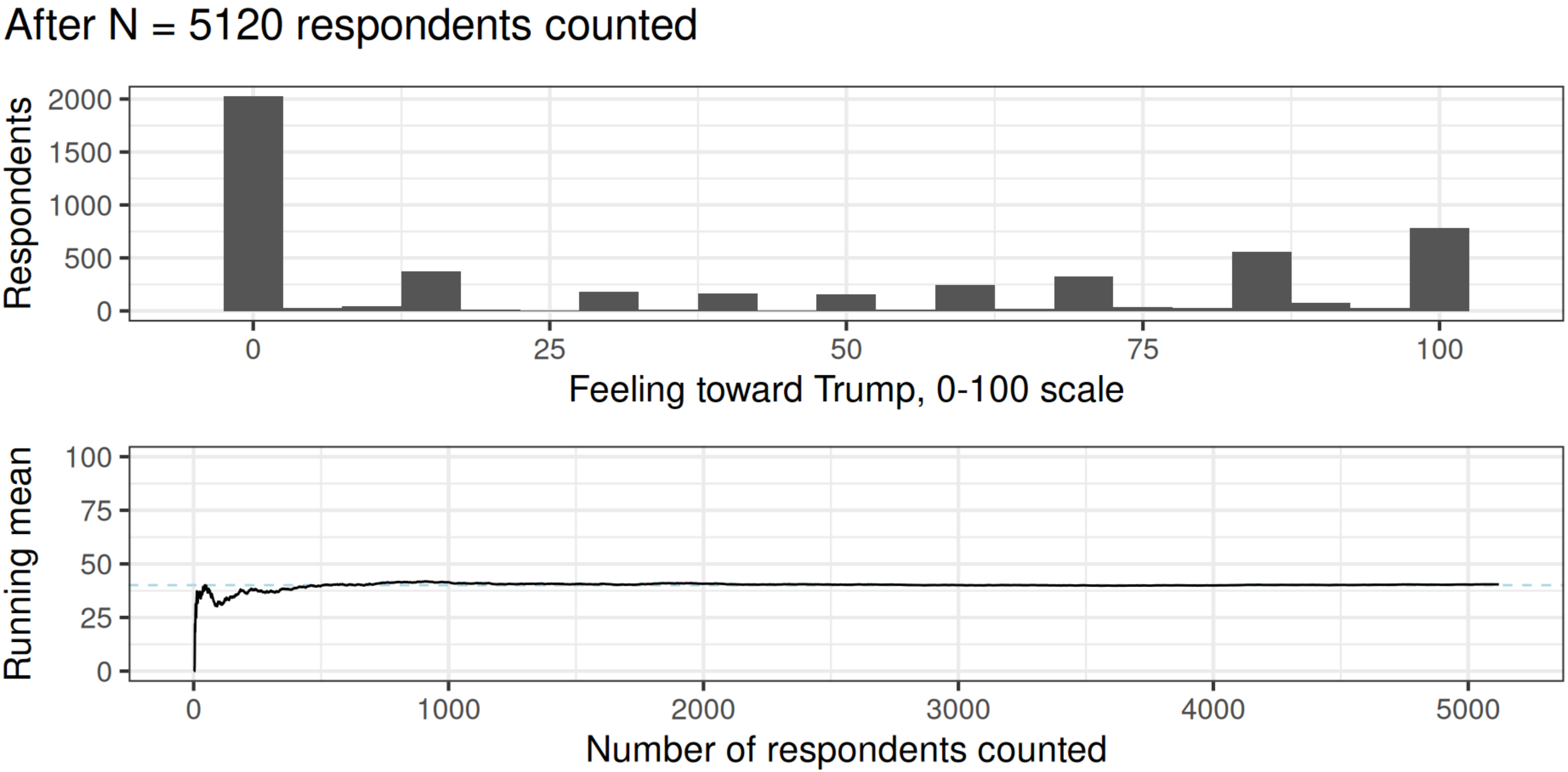
Stability of the sample mean



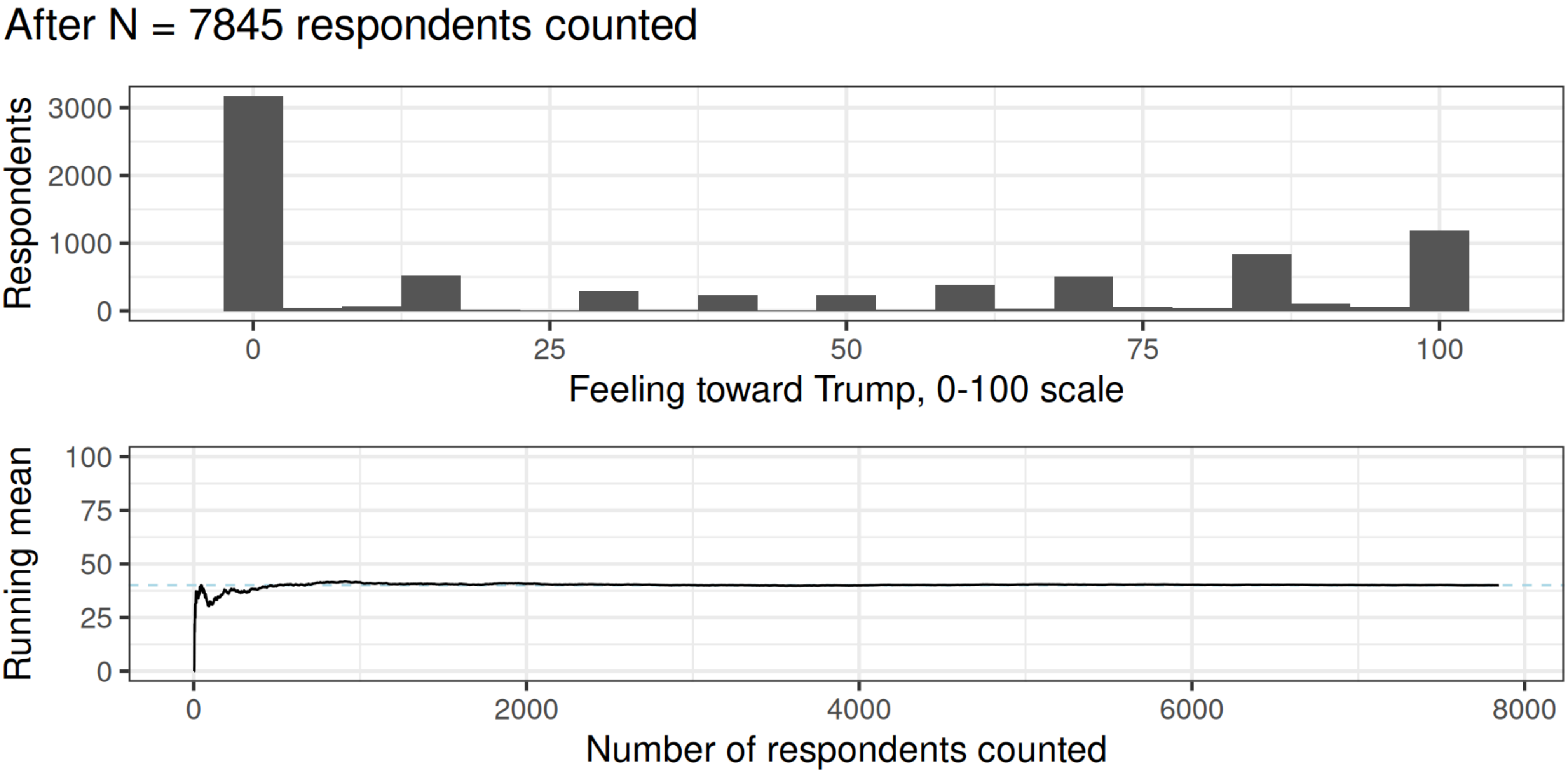
Stability of the sample mean



Stability of the sample mean



Stability of the sample mean



After N = 7845 respondents counted

Population variance

Variance is the typical (squared) distance b/w an observation and the mean

Mathematical formula:

$$\mathbb{V}[\boldsymbol{Y_i}] = \mathbb{E}\left[(\boldsymbol{Y_i} - \mathbb{E}[\boldsymbol{Y_i}])^2\right]$$

How to think about this:

1. Calculate each observation's distance from mean: $\boldsymbol{Y_i} - \mathbb{E}[\boldsymbol{Y_i}]$
2. Square the distances: $(\boldsymbol{Y_i} - \mathbb{E}[\boldsymbol{Y_i}])^2$
    - Now they're all positive
    - Bigger differences matter more-er
3. Take the average of the squared distances: $\mathbb{E}[(\boldsymbol{Y_i} - \mathbb{E}[\boldsymbol{Y_i}])^2]$

Sample variance

We go from population variance to **sample variance** in much the same way we go from expected value to sample mean:

$$\mathrm{var}[Y_i] = \mathrm{avg}\left[\left(Y_i - \bar{Y}\right)^2\right]$$

ⓘ **The average in the sample variance**

Most of the time, including by default in R, sample variance is calculated with a slight modification to the standard averaging formula:

$$\mathrm{var}[Y_i] = \frac{N}{N-1}\,\mathrm{avg}\left[\left(Y_i - \bar{Y}\right)^2\right] = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2.$$

This is for a *bias correction*, a technical issue we won't get into. You should usually work with large enough samples that it doesn't matter if you divide by N or by $N - 1$.

Standard deviation

Variance can be hard to interpret because it’s in *squared* units

We often work with the **standard deviation**, the square root of the variance:

$$\text{sd}[Y_i] = \sqrt{\text{var}[Y_i]} = \sqrt{\text{avg}[(Y_i - \bar{Y})^2]}.$$

```r
var(df_anes$therm_trump)
```

```
[1] 1624.415
```

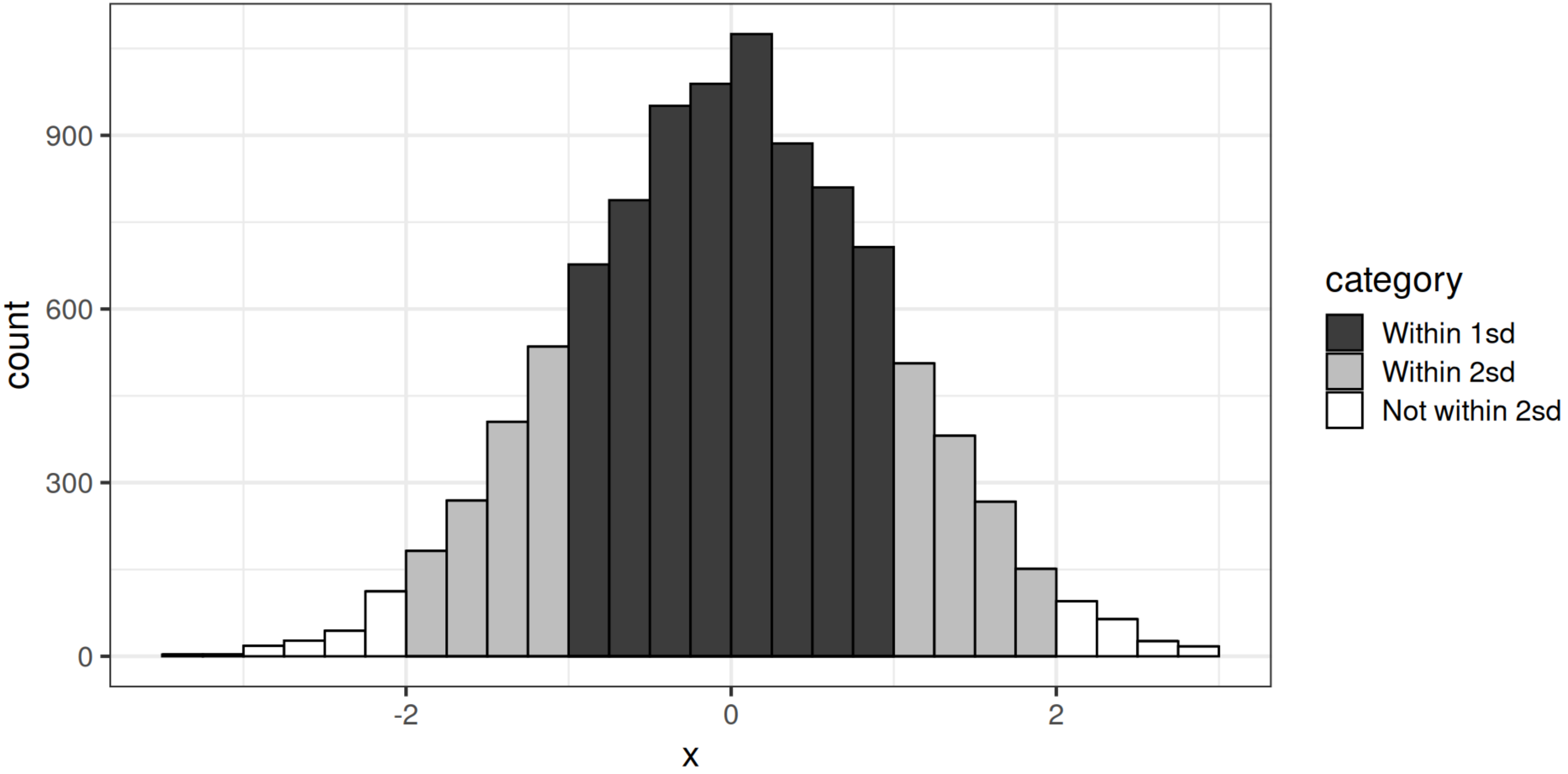
```r
sd(df_anes$therm_trump)
```

```
[1] 40.30403
```

Interpreting standard deviation

Standard deviation is a “typical” distance from the mean

Normally distributed data ⇝ ~68% within 1sd, ~95% within 2sd



Variance and standard deviation of a binary variable

Think again about a binary variable Y_i

→ e.g., $Y_i = 1$ for those who watch Tucker Carlson, $Y_i = 0$ for those who don't

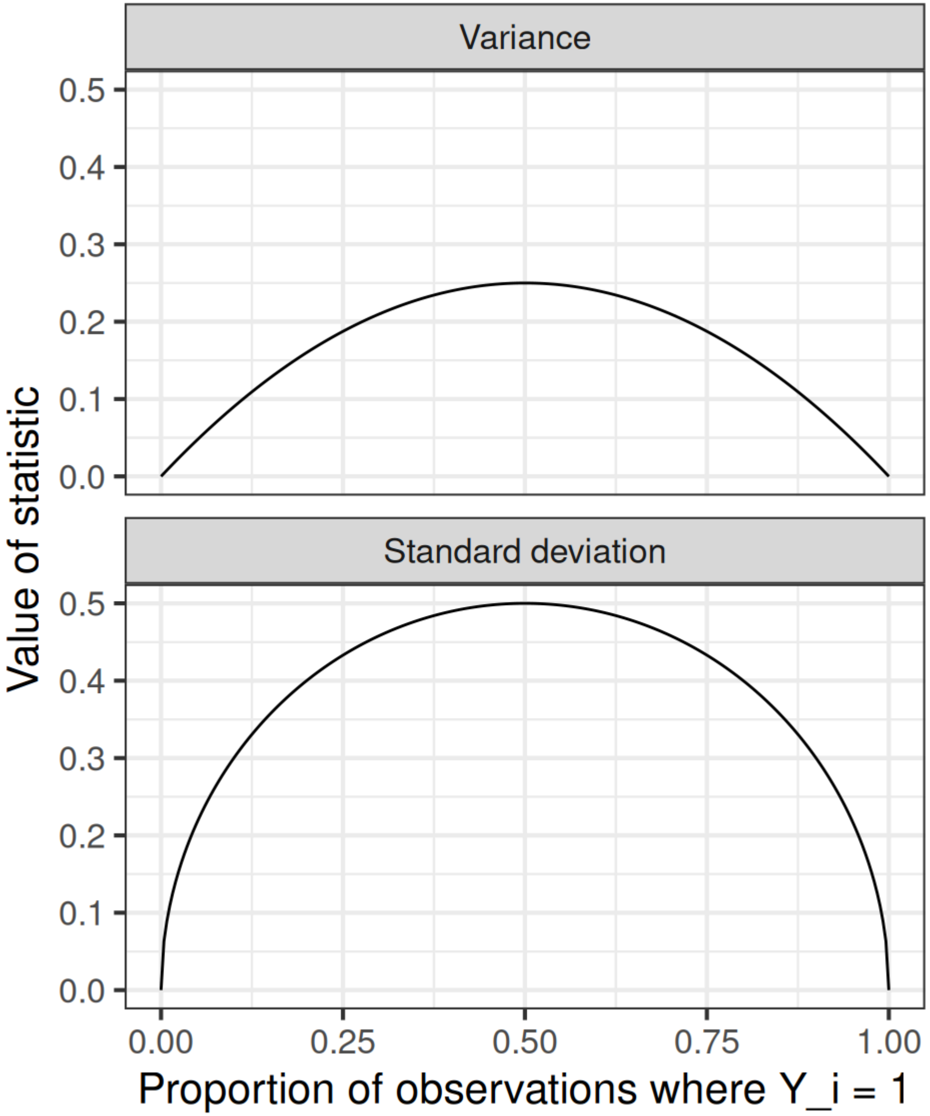
Sample mean $\bar{Y}$ is proportion of obs where $Y_i = 1$

Simple formulas for the sample variance and standard deviation in this case:

$$\begin{aligned} \text{var}[Y_i] &= \bar{Y}(1 - \bar{Y}) \\ \text{sd}[Y_i] &= \sqrt{\bar{Y}(1 - \bar{Y})} \end{aligned}$$

Variance and standard deviation of a binary variable

- $\bar{Y} = 0$
  - → every $Y_i = 0$
  - → no variance
- $\bar{Y} = 1$
  - → every $Y_i = 1$
  - → no variance
- $\bar{Y} = 0.5$
  - → max uncertainty about each individual observation
  - → highest possible variance for binary variable



Covariance and correlation

Covariance

Now suppose we have two variables, X_i and Y_i

- → e.g., X_i = does this person watch Tucker?, Y_i = opinion of Trump, 0–100

Covariance: how much above-average X_i predicts above-average Y_i

$$\mathbb{C}[X_i, Y_i] = \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])\right]$$

Loose guide to interpreting covariance:

Covariance sign	Interpretation
$\mathbb{C}[X_i, Y_i] > 0$	above-average X_i predicts above-average Y_i
$\mathbb{C}[X_i, Y_i] = 0$	no average relationship b/w X_i and Y_i
$\mathbb{C}[X_i, Y_i] < 0$	above-average X_i predicts below-average Y_i

The trouble with covariance

Remember with the variance $\mathbb{V}[Y_i]$:

- Hard to interpret directly because in squared units
- We took the square root to ease interpretation

Interpreting covariance on its own is even harder

Measured in (units of X_i) $\times$ (units of Y_i)

```r
cov(df_anes$watch_tucker, df_anes$therm_trump, use = "complete")
```

```
[1] 2.344265
```

What does this number mean, besides being positive? I've taught PhD-level stats for a decade and can't honestly say

From covariance to correlation

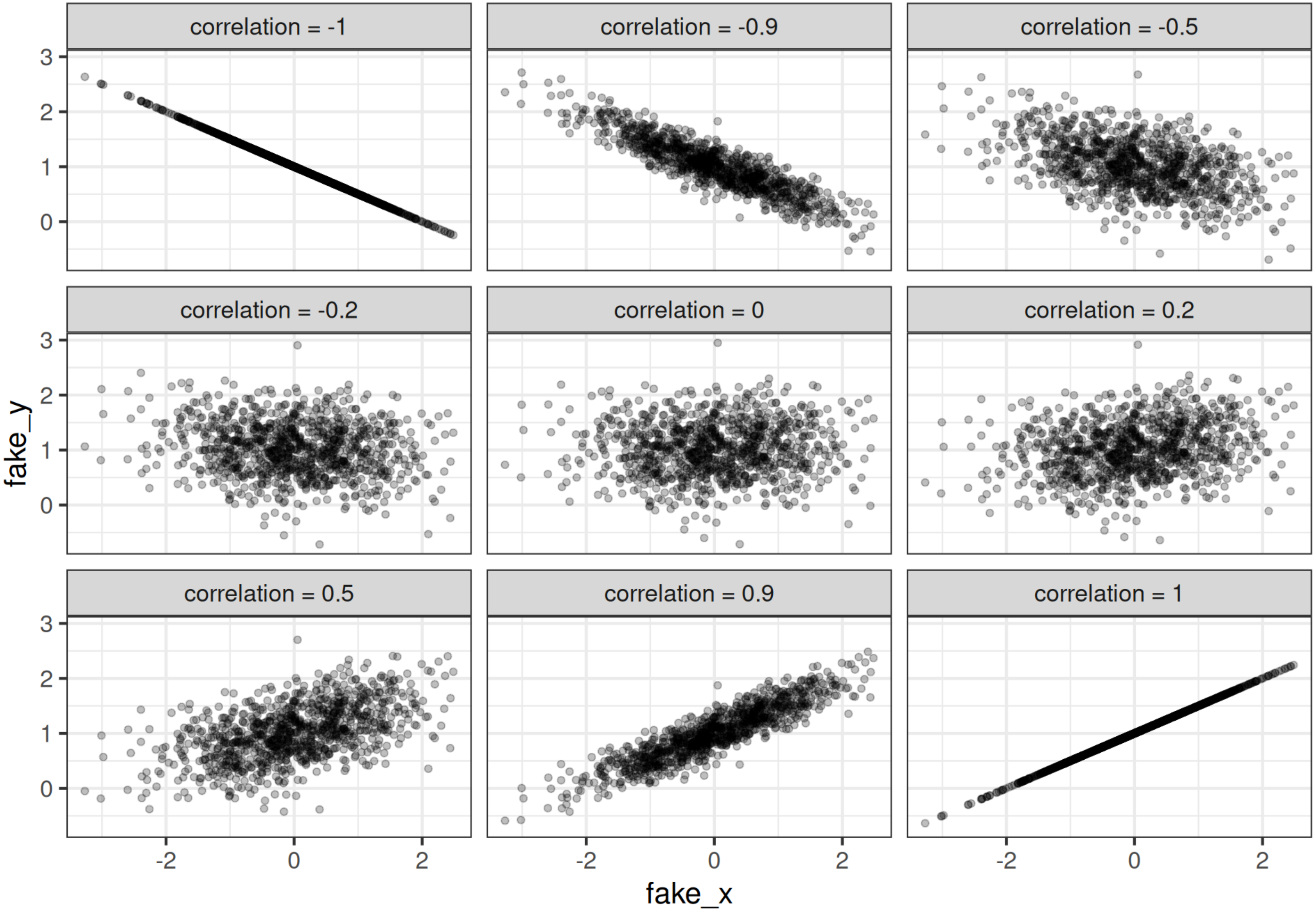
More common measure of relationship strength is the **correlation coefficient**

$$\frac{\mathbb{C}[X_i, Y_i]}{\sqrt{\mathbb{V}[X_i] \times \mathbb{V}[Y_i]}}$$

Only takes values from -1 to 1

- Sign — direction of relationship
  - → Positive: Above-average X_i predicts above-average Y_i
  - → Negative: Above-average X_i predicts below-average Y_i
- Magnitude — strength of relationship
  - → Close to 0: X_i not very predictive of Y_i (and vice versa)
  - → Close to -1 or 1: X_i highly predictive of Y_i (and vice versa)

The correlation coefficient



Sample covariance and correlation

Shouldn’t surprise you at this point, but just for completeness…

Sample covariance:

$$\mathrm{cov}[X_i, Y_i] = \mathrm{avg}[(X_i - \bar{X})(Y_i - \bar{Y})]$$

Sample correlation coefficient:

$$\mathrm{cor}[X_i, Y_i] = \mathrm{avg}\left[\frac{X_i - \bar{X}}{\mathrm{sd}[X_i]} \times \frac{Y_i - \bar{Y}}{\mathrm{sd}[Y_i]}\right] = \frac{\mathrm{cov}[X_i, Y_i]}{\mathrm{sd}[X_i]\,\mathrm{sd}[Y_i]}$$

Difference of means and regression

The difference of means

How much more or less do Tucker watchers like Trump, compared to non-Tucker-watchers?

⚠ **Still not causal!**

We are not (yet) asking if watching Tucker causes changes in opinion toward Trump.

This is a question about **conditional means**

- $\mathbb{E}[Y_i \mid X_i = 1]$: average Trump opinion among watchers
- $\mathbb{E}[Y_i \mid X_i = 0]$: average Trump opinion among non-watchers
- $\mathbb{E}[Y_i \mid X_i = 1] - \mathbb{E}[Y_i \mid X_i = 0]$: difference of means

Sample difference of means

Average opinion among watchers in our sample: $\textbf{avg}[Y_i \mid X_i = 1]$

```r
mean(df_anes$therm_trump[df_anes$watch_tucker == 1])
```

```
[1] 82.64583
```

Average opinion among non-watchers in our sample: $\textbf{avg}[Y_i \mid X_i = 0]$

```r
mean(df_anes$therm_trump[df_anes$watch_tucker == 0])
```

```
[1] 37.59949
```

Sample difference of means: $\textbf{avg}[Y_i \mid X_i = 1] - \textbf{avg}[Y_i \mid X_i = 0]$

```r
mean(df_anes$therm_trump[df_anes$watch_tucker == 1]) -
  mean(df_anes$therm_trump[df_anes$watch_tucker == 0])
```

```
[1] 45.04635
```

An amazing fact about the difference of means

aka: Why I taught you about covariance even though it's hard to interpret

If X_i is a binary (0/1) variable:

$$\text{avg}[Y_i \mid X_i = 1] - \text{avg}[Y_i \mid X_i = 0] = \frac{\text{cov}[X_i, Y_i]}{\text{var}[X_i]}$$

▶ Mathematical proof

Example with our ANES data:

```
trump_tucker_cov <- cov(df_anes$therm_trump, df_anes$watch_tucker)
tucker_var <- var(df_anes$watch_tucker)
trump_tucker_cov / tucker_var
```

```
[1] 45.04635
```

A related amazing fact about bivariate regression

You hopefully remember the bivariate regression formula

$$\mathbb{E}[Y_i \mid X_i = x] = \alpha + \beta x$$

where α is the **intercept** and β is the **slope**

Using ordinary least squares, our estimated slope is

$$\hat{\beta} = \frac{\textbf{cov}[X_i, Y_i]}{\textbf{var}[X_i]}$$

Same as difference of means formula for the special case when X_i is binary

A related amazing fact about bivariate regression

```r
lm(therm_trump ~ therm_biden, data = df_anes)
```

```

Call:
lm(formula = therm_trump ~ therm_biden, data = df_anes)

Coefficients:
(Intercept)  therm_biden  
    85.8744      -0.9315  
```

```r
trump_biden_cov <- cov(df_anes$therm_trump, df_anes$therm_biden)
biden_var <- var(df_anes$therm_biden)
trump_biden_cov / biden_var
```

```
[1] -0.9315374
```

Wrapping up

What we did today

Our most important statistical calculations are built on a few key statistics:

- The **mean**, $\bar{Y} = \text{avg}[Y_i] = \frac{1}{N}\sum_{i=1}^{N} Y_i$
- The **variance**, $\text{var}[Y_i] = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- The **covariance**, $\text{cov}[X_i, Y_i] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$

The correlation, difference of means, and bivariate regression slope can all be computed from these underlying statistics

To do for next time

We'll cover the potential outcomes framework, creating a statistical model of cause-effect relationships

1. Read pages 1–11 of *Mastering 'Metrics*.
2. Read "Statistics and Causal Inference". Most importantly:
    - Model for associational inference and "Rubin's model" (§1–3)
    - The fundamental problem of causal inference (§3)
    - What can be a cause? (§7)
3. Problem Set 1 to be posted by today, due next Fri 1/24

Appendix

Proving the difference of means formula (1/2)

First we'll prove a useful property of covariance, namely that $\mathbf{cov}[X_i, Y_i] = \mathbf{avg}[X_i Y_i] - \bar{X}\bar{Y}$. (This is true even when X_i is not binary.)

$$
\begin{aligned}
\text{cov}[X_i, Y_i] &= \text{avg}[(X_i - \bar{X})(Y_i - \bar{Y})] \\
&= \frac{1}{N} \sum_{i=1}^{N} [(X_i - \bar{X})(Y_i - \bar{Y})] \\
&= \frac{1}{N} \left[\sum_{i=1}^{N} X_i Y_i - \bar{X} \sum_{i=1}^{N} Y_i - \bar{Y} \sum_{i=1}^{N} X_i + N\bar{X}\bar{Y} \right] \\
&= \frac{1}{N} \left[\sum_{i=1}^{N} X_i Y_i - N\bar{X} \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} Y_i \right)}_{=\bar{Y}} - N\bar{Y} \underbrace{\left(\frac{1}{N} \sum_{i=1}^{N} X_i \right)}_{=\bar{X}} + N\bar{X}\bar{Y} \right] \\
&= \frac{1}{N} \left[\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y} \right] \\
&= \text{avg}[X_i Y_i - \bar{X}\bar{Y}].
\end{aligned}
$$

▶ Back to main slides

Proving the difference of means formula (2/2)

Now consider the case where X_i is binary. Assume the observations are ordered so that $X_i = 1$ for $i = 1, \ldots, N_1$ and that $X_i = 0$ for $i = N_1 + 1, \ldots, N$. Define N_0 as the number of observations for which $X_i = 0$, i.e., $N_0 = N - N_1$.

By definition, the sample mean of X_i is $\bar{X} = \frac{N_1}{N}$. Additionally, we have $\text{avg}[Y_i \mid X_i = 1] = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i$ and $\text{avg}[Y_i \mid X_i = 0] = \frac{1}{N_0}\sum_{i=N_1+1}^{N} Y_i$.

$$\begin{aligned}
\frac{\text{cov}[X_i, Y_i]}{\text{var}[X_i]} = \frac{\text{avg}[X_iY_i - \bar{X}\bar{Y}]}{\bar{X}(1-\bar{X})} &= \frac{\frac{1}{N}\sum_{i=1}^{N} X_iY_i - \left(\frac{N_1}{N}\right)\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)}{\frac{N_1}{N} \times \frac{N_0}{N}} \\
&= \frac{N[\sum_{i=1}^{N_1}(1)Y_i + \sum_{i=N_1+1}^{N}(0)Y_i] - N_1\sum_{i=1}^{N} Y_i}{N_1 \times N_0} \\
&= \frac{\cancel{(N - N_1)}}{N_1 \times \cancel{N_0}}\sum_{i=1}^{N_1} Y_i - \frac{\cancel{N_1}}{\cancel{N_1} \times N_0}\sum_{i=N_1+1}^{N} Y_i \\
&= \text{avg}[Y_i \mid X_i = 1] - \text{avg}[Y_i \mid X_i = 0].
\end{aligned}$$

▶ Back to main slides