Advanced Statistics for Social Sciences 1

Statistical Thinking and Data Analysis with R

Instructor: Rachit Dubey

Email: rdubey@ucla.edu

Office Hours: [Thursdays 4-4:30 pm]

Summary of last week

Statistics helps us:

- **Describe** complex phenomena simply
- **Decide** under uncertainty
- **Predict** future outcomes

Descriptive statistics:

- **Mean:** Center of gravity, uses all data, sensitive to outliers
- **Median:** Middle value, robust to outliers
- **Mode:** Most frequent, useful for categorical data

Lecture 2: Understanding Variability

Measuring Spread in Data

- Why variability matters (modeling perspective)
- Measures of spread: range, IQR, variance, SD
- Degrees of freedom
- Z-scores
- What makes a good model?

Data Visualization

- Data Visualization in R
- Principles of good data visualization

Fitting models to data

Recap: What is statistical thinking?

Statistical thinking is a way of understanding a complex world by describing it in **relatively simple terms** that nonetheless capture essential aspects of its structure.

Key Characteristics:

- Simplifies complexity while preserving essential information
- Quantifies uncertainty in our knowledge

Statistical models

- “Models” are generally simplifications of things in the real world that nonetheless convey the essence of the thing being modeled.

- E.g., a model of a building conveys the structure of the building while being small & light enough to pick up with one’s hands.

- In statistics, a model is meant to provide a similarly condensed description, but for data (rather than for a physical structure).

- Like physical models, a statistical model is generally much simpler than the data being described; it is meant to capture the structure of the data as simply as possible.

Statistical models

```
data = model + error
```

Data can be expressed by:

1. A statistical model, which expresses the values we expect the data to take given our knowledge
2. The *error,* which reflects the difference between the model’s predictions and the observed data

Statistical models

```
data = model + error
```

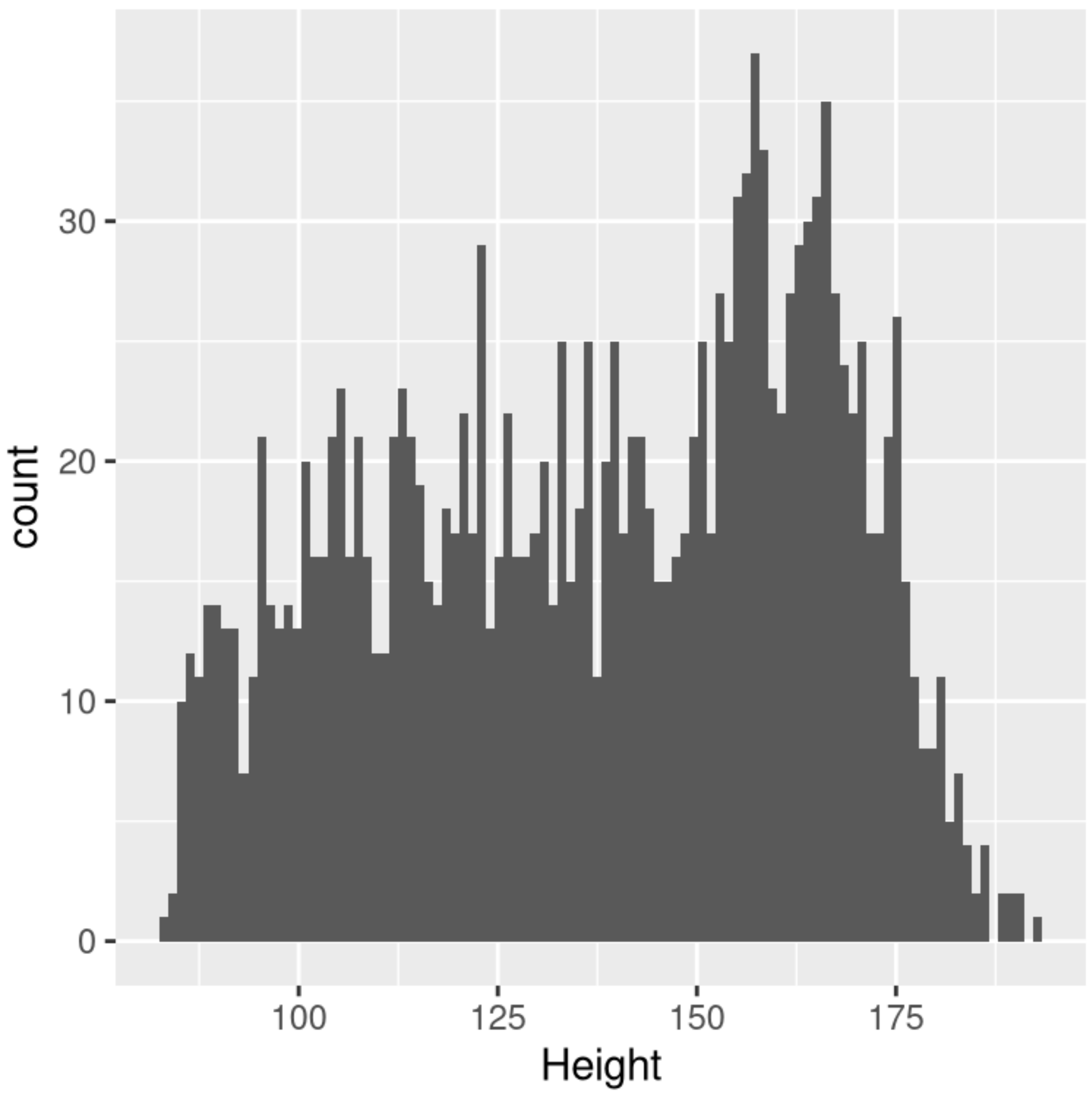
- **Last week:** We talked about central tendency (mean, median, mode).
- In essence, **these are models**!
- They aimed to provide a “condensed description” for the data we have in hand.
- The error term helps us evaluate “good” these models are

Example: Statistical Modeling

Data: Heights of 1,691 children from NHANES

Goal: Describe the data as simply as possible

Histogram of height of children in NHANES



The simplest model: predict **the same value** for every child

$y_i = \beta + \varepsilon$

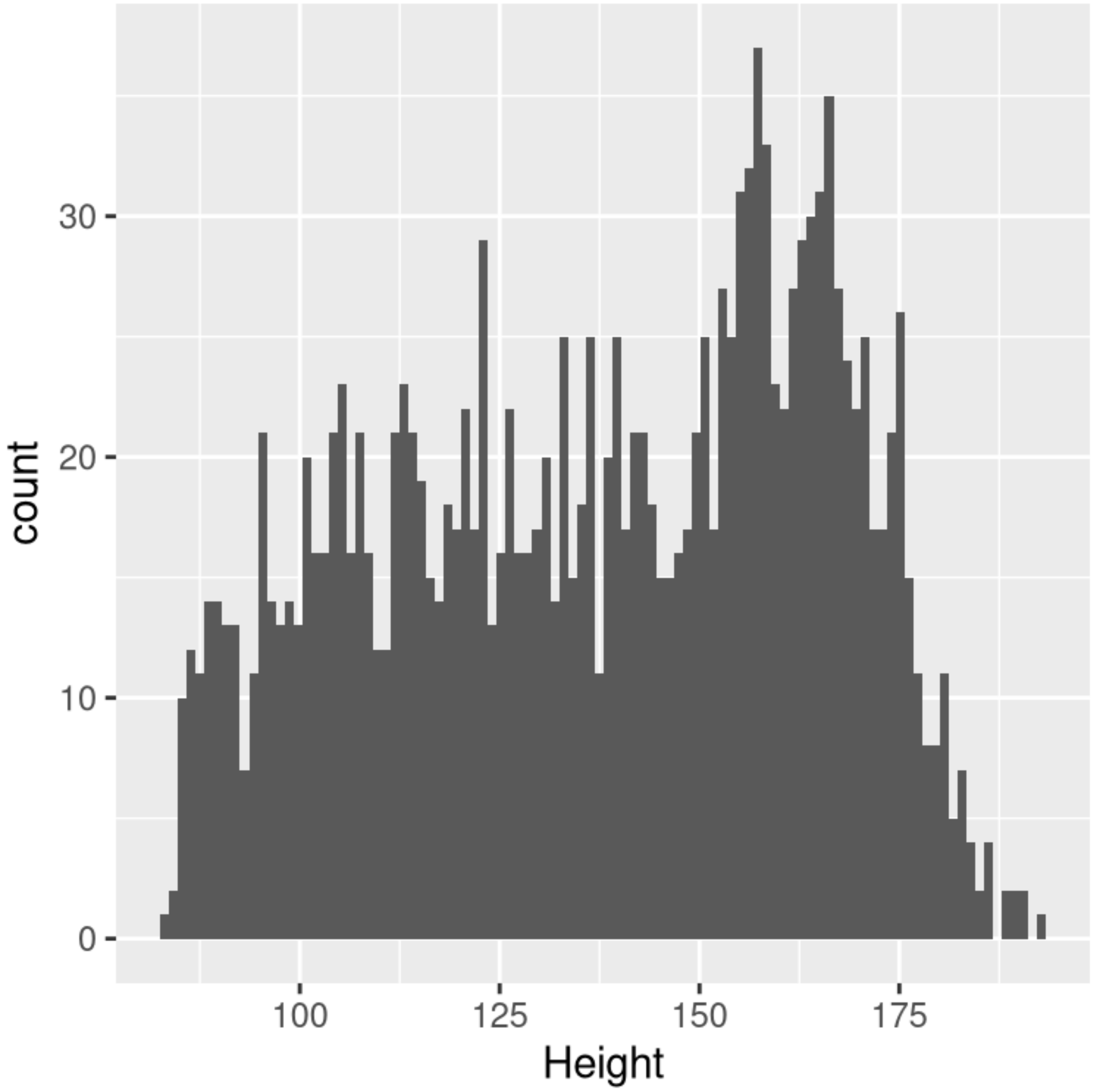
Where:

- y_i = height of child i
- β = our parameter (the single number we're choosing)
- ε = error (how far off we are for each child)

Option 1: Using the Mode

Mode: The most common value in the dataset

Histogram of height of children in NHANES



For NHANES children: Mode = 166.5 cm

Our prediction: **$\hat{y}_i$ = 166.5 cm for everyone**

The error for each child:

$\text{error}_i = y_i - \hat{y}_i$

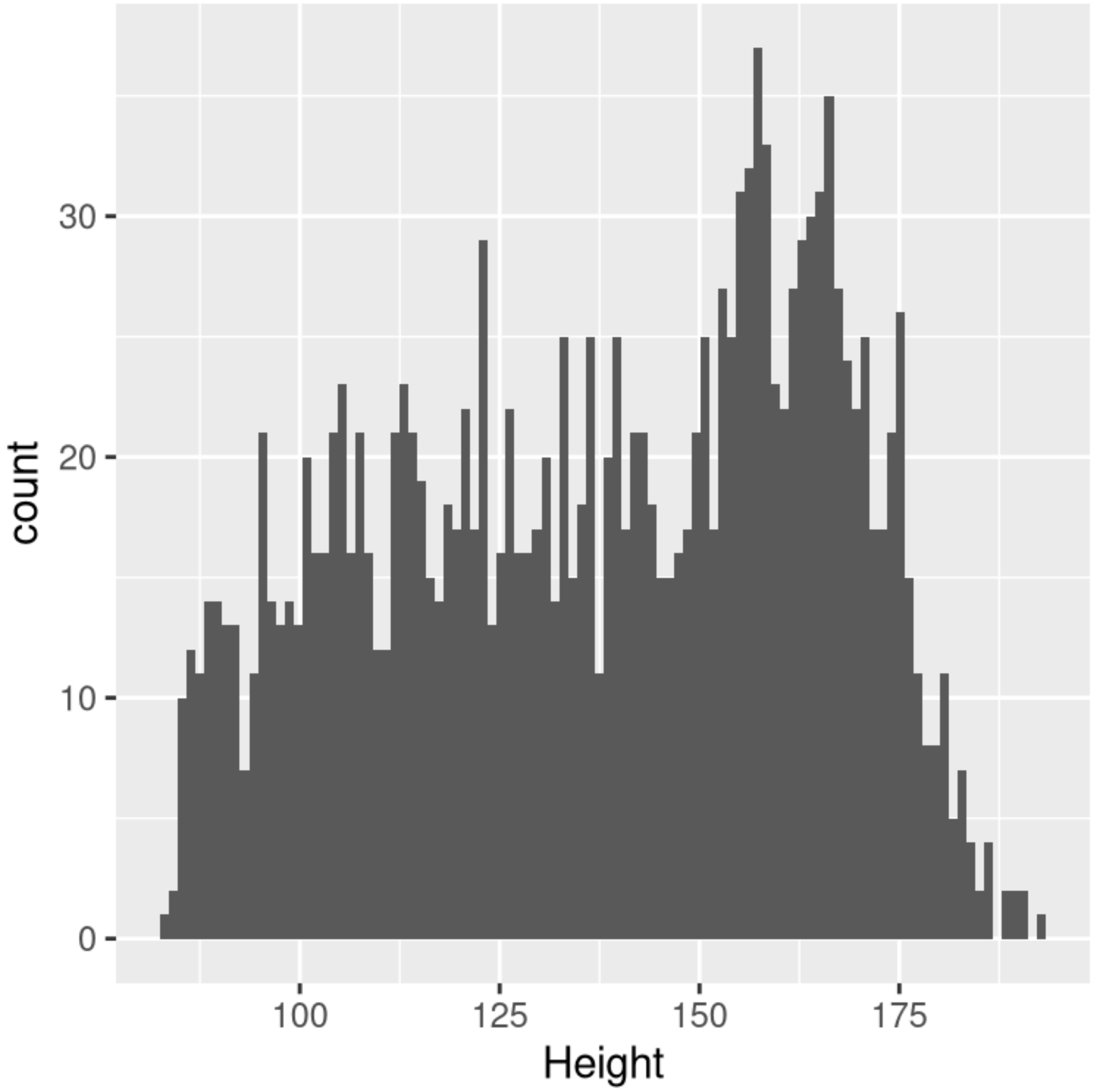
How good is this model?

- We define goodness in terms of magnitude of error
- One example is using the Average of the errors
- Average error using the mode = **-28.8 cm**
- Clearly mode isn't a great model…

Option 2: Using the Mean

Mean: Sum of all values divided by number of values, $\frac{\sum_{i=1}^{N} x_i}{N}$

Histogram of height of children in NHANES



For NHANES children: Mean = 141 cm

Our prediction: **$\hat{y}_i$ = 141 cm for everyone**

The error for each child:

$\text{error}_i = y_i - \hat{y}_i$

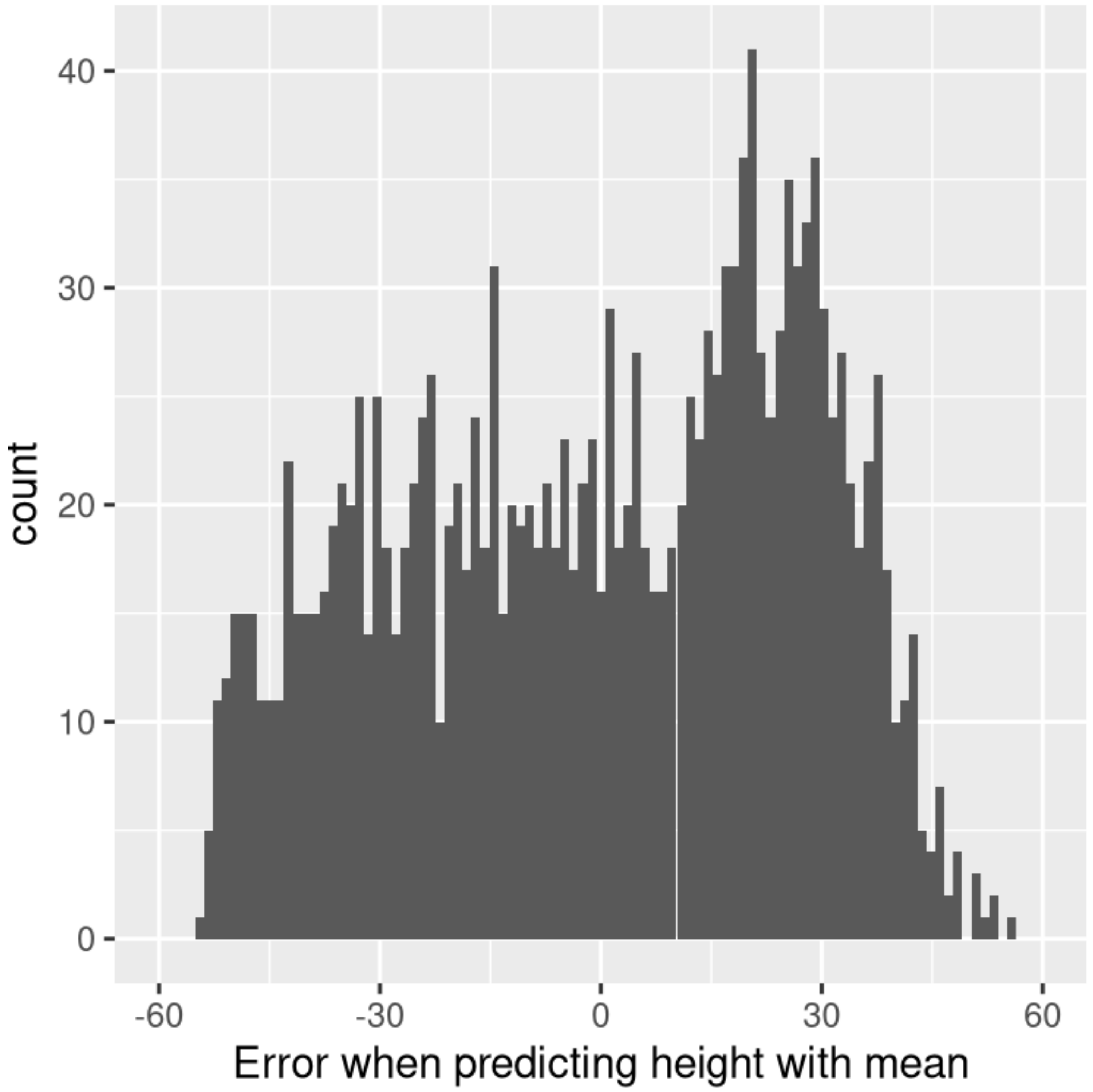
How good is this model?

- *It turns out that the average error of the mean is always 0!*

Option 2: Using the Mean

Mean: Sum of all values divided by number of values, $\frac{\sum_{i=1}^{N} x_i}{N}$

Distribution of errors from the mean



The negative and positive errors end up cancelling each other out!

Important: Individual data points still have error; positive and negative errors cancel out to give average error of zero

Why Average Error isn't enough

Average Error $= \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right) \quad = \frac{1}{N}\Big(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N}\mu\Big)$

$$= \frac{1}{N}\Big(\sum_{i=1}^{N} x_i - N\mu\Big)$$

$$= \frac{1}{N}\Big(\sum_{i=1}^{N} x_i - N\frac{\sum_{i=1}^{N} x_i}{N}\Big)$$

$$= \frac{1}{N}\Big(\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i\Big)$$

$$= 0$$

Why Average Error isn’t enough

The fact that positive and negative errors cancel each other out means:

Two models with very different error magnitudes could both have **average error = 0**

Why Average Error isn’t enough

The fact that positive and negative errors cancel each other out means:

Two models with very different error magnitudes could both have **average error = 0**

Example: Two classes with **mean = 75**

Class A: 73, 74, 75, 76, 77
Errors: -2, -1, 0, +1, +2 (average error = 0)

Class B: 45, 60, 75, 90, 105
Errors: -30, -15, 0, +15, +30 (average error = 0)

In both, `data = model + error`, implying that `data = 75` is a good statistical model for both!

Why Average Error isn’t enough

The fact that positive and negative errors cancel each other out means:

Two models with very different error magnitudes could both have **average error = 0**

We need error measures that capture the **magnitude** of errors, regardless of direction!

Class B: 45, 60, 75, 90, 105
Errors: -30, -15, 0, +15, +30 (average error = 0)

In both, `data = model + error`, implying that `data = 75` is a good statistical model for both!

Variability

```
data = model + error
```

How large are the errors? Aka how spread out is our data?

Range

Definition: Maximum Value - Minimum Value

Example, NBA Finals winning margins: 0 to 116 points

Range = 116-0 = **116 points**

Pro: Simple to calculate and easy to understand

Cons:

- Uses only two data points, ignores everything else
- *Extremely sensitive* to outliers

Example: -100, 2, 3, 4, 5, 6, 7, 8, 9, 10

Range = 110 (dominated by one outlier)

Interquartile Range (IQR)

Percentile: The value below which a certain percentage of data falls

50th percentile: 50% of the data is below this value (which is actually the median!)

10th percentile: 10% of the data is below this value

75th percentile: 75% of the data is below this value

Quartiles divide the data into equal parts:

- Q1 = 25th percentile (first quartile)
- Q2 = 50th percentile (second quartile)
- Q3 = 75th percentile (third quartile)

Interquartile Range (IQR)

Definition: Range of the “middle 50%” of data

```
IQR = Q3 - Q1 = 75th percentile - 25th percentile
```

Winning margins:

- Q1 (25th percentile) = 12.75
- Q3 (75th percentile) = 50.50
- IQR = 50.50 - 12.75 = **37.75 points**

Interpretation:

- The middle half of games had winning margins between 12.75 and 50.50 points
- **Robust** to outliers (ignores extreme 25% on each end)
- <u>**Complements the median nicely**</u>

Variance

Recall that we want to measure "typical" deviation of data from our model (where our model was the mean)

Problem: If we average the errors, they sum to zero!

Average Error $= \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \mu\right) = 0$

Solution: Square the deviations first!

Sum of Squared Errors (SSE) $= \sum_{i=1}^{N} \left(x_i - \mu\right)^2$

Variance $= \frac{\textbf{SSE}}{N} = \frac{1}{N} \sum_{i=1}^{N} \left(x_i - \mu\right)^2$

Sidenote: Biased and unbiased estimators

Population vs. sample

Population

- **All** possible observations we care about
- Parameters: μ (true mean), σ^2 (true variance)
- Usually unknown - we can't measure everyone!

Sample

- The data we actually collected
- Statistics: $\bar{x}$ (sample mean), $\hat{\sigma}^2$ (sample variance)
- Used to **estimate** population parameters

Example: NHANES children

Population: All children in the US (millions)

Sample: 1,691 children we measured

The idea is that $\bar{x}$ from our sample is a proxy estimate for the true mean, μ for the population

Population vs. sample

Variance: Two Formulas

Population variance (if our data contains all population):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

Sample variance (estimating from sample):

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{\mathbf{x}}\right)^2$$

Notice the difference: N vs (N-1). Why?

Wait… why (N-1)?

Question: Why not just divide by N, the actual number of observations?

This relates to the concept of **degrees of freedom**

Degrees of Freedom

Example: Dataset with 5 values: 3, 5, 7, 9, 11

Mean = 7

Now hide one value: **?, 5, 7, 9, 11**

Can you figure out the missing value?

Since mean = 7, the sum must be 7 × 5 = 35

We know: 5 + 7 + 9 + 11 = 32

Therefore, missing value = 35 - 32 = 3

Key insight: Once we know the mean, only **(n-1) values are "free to vary".**

The last value is determined by the constraint that the mean must equal 7

Why it matters: Bias

If we divide by N: We get a **biased** estimate

- Systematically underestimates the true population variance
- Makes data looks less variable than it really is

If we divide by (N-1): We get an **unbiased** estimate

- Correctly estimates population variance on average
- Accounts for the information “used up” by estimating the mean

General principle: df = N - (number of parameters estimated)

For variance: we estimated 1 parameter (the mean), so we “lost” a degree of freedom with df = N - 1

Okay.. let’s return back to variability

Variance

$$\textbf{Variance} = \frac{\textbf{SSE}}{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{\mathbf{x}}\right)^2$$

Example: Winning margins: 56, 31, 56, 8, 32 (mean = 36.6)

Value	Deviation	Squared Deviation
56	19.4	376.36
31	-5.6	31.36
56	19.4	376.36
8	-28.6	817.96
32	-4.6	21.16
Sum:		**1623.2**

Variance = 1623.2 / (5-1) = 1623.2 / 4 = **405.8**

Problem: Units are points^2, which kind of doesn't make much sense!

Standard Deviation

Solution: Take the square root to get back to original units

```
SD = √Variance
```

```
SD = √405 = 20.1 points
```

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{\mathbf{x}}\right)^2}$$

Interpretation: On average, winning margins deviate from the mean by about 20 points

Why SD is popular

- Same units as original data (interpretable!)
- Uses all data points
- Mathematically elegant

Summary: statistical modeling

```
data = model + error
```

- Let’s rethink of mean (and other measures) as a statistical model of the data
- Instead of saying “the data differs from the mean”, this perspective says: “*The mean differs from the data*”
- The different variability terms, are a way to capture error i.e., how much this model deviates from the data

Summary: Variability Measures

Measure	Formula	When to Use
Range	max - min	Quick sense of spread; not robust
IQR	Q3 - Q1	Robust; pairs well with median
Variance	$\Sigma(x-\bar{x})^2/(n-1)$	How well mean captures data; squared units
SD	$\sqrt{\text{Variance}}$	Most common; pairs with mean

Rule of thumb: Report SD with mean, IQR with median

Z-scores

Z-scores

Until now.. We characterized data in terms of its central tendency and variability

What if we want to see how individual datapoint sit with respect to the overall distribution?

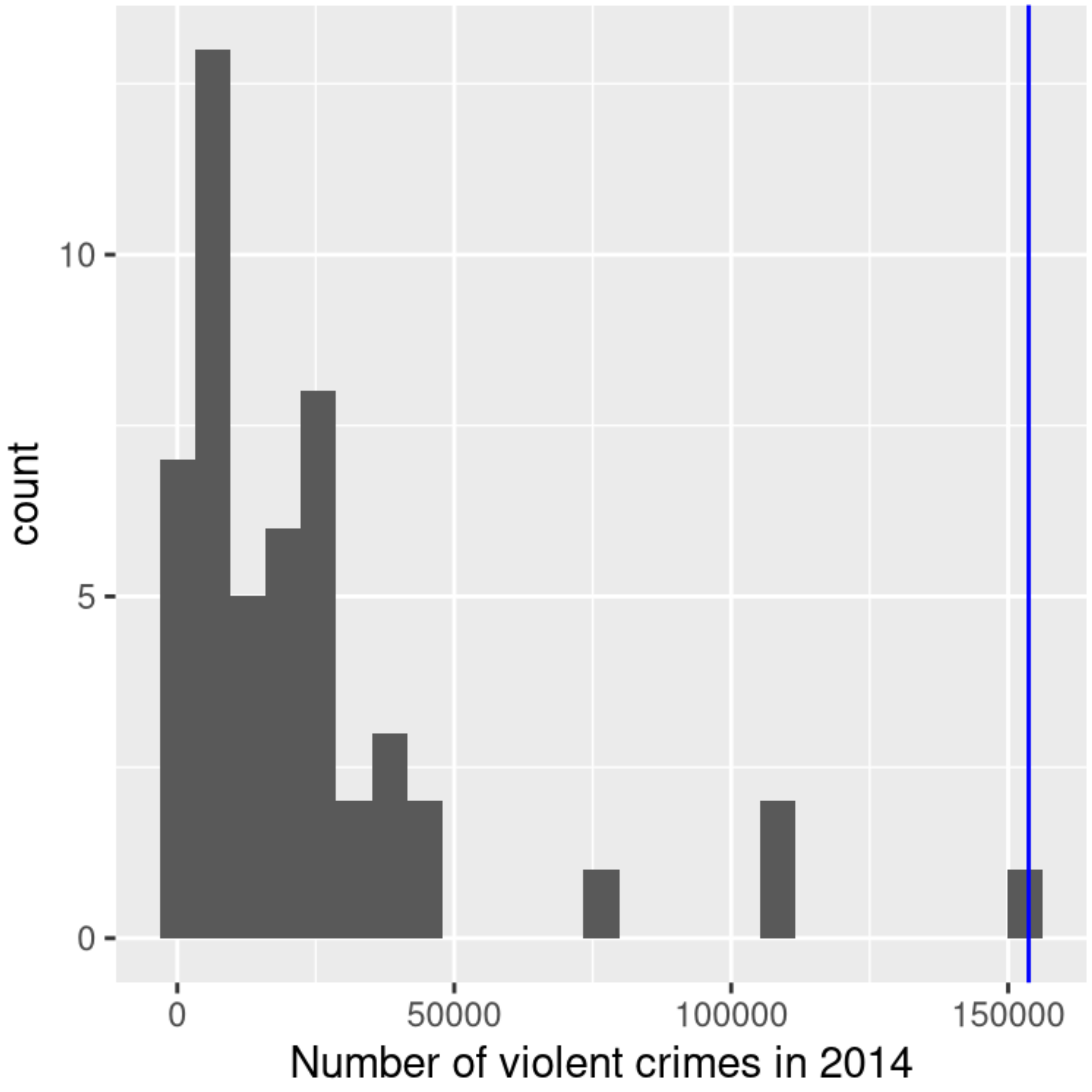
Z-score: Express each value in terms of standard deviations from mean

```
Z = (x - μ) / σ
```

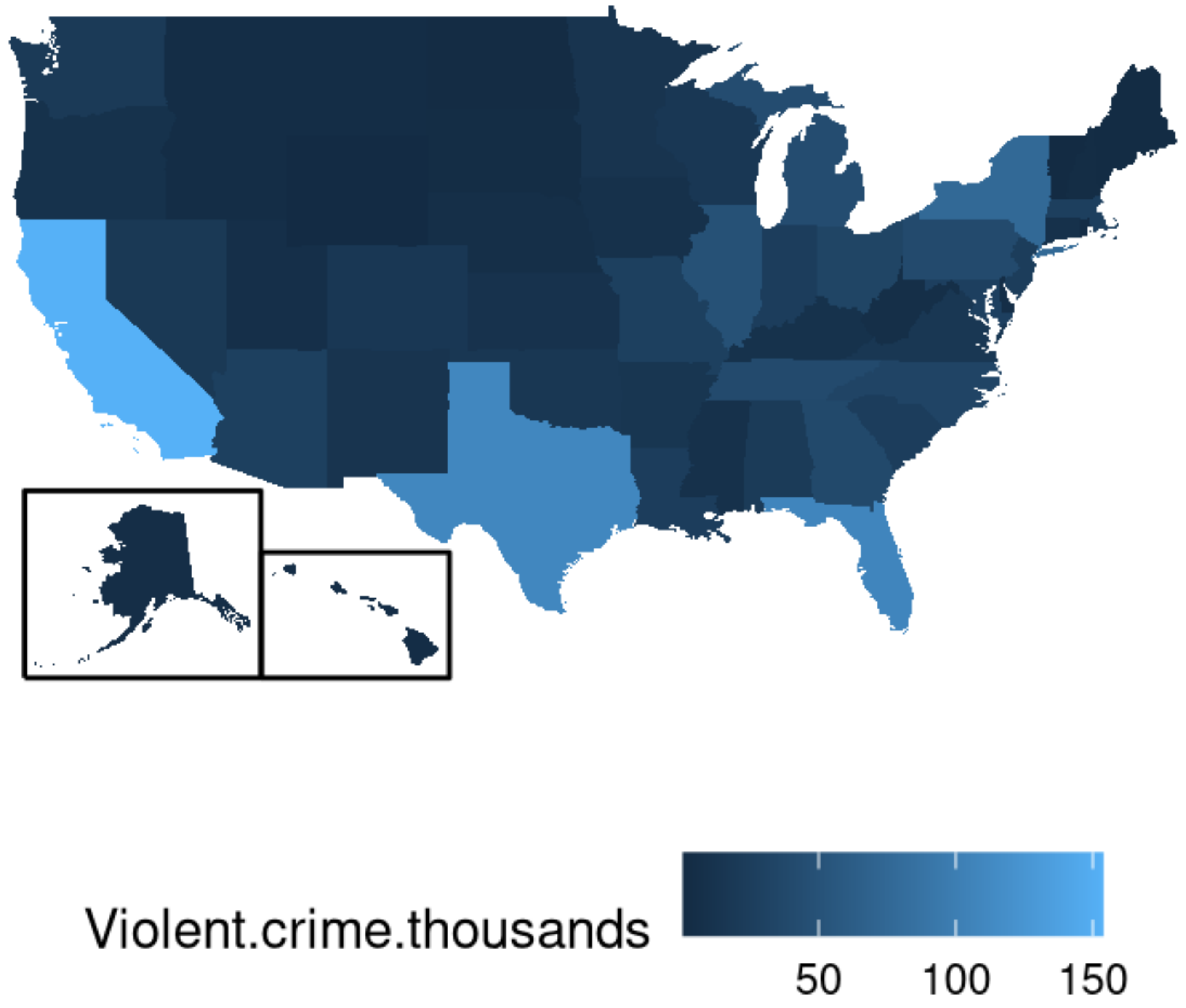
Interpretation: How far away any data point is from the mean, in units of standard deviation?

Z-scores example

Histogram of the number of violent crimes per state (California highlighted in blue)



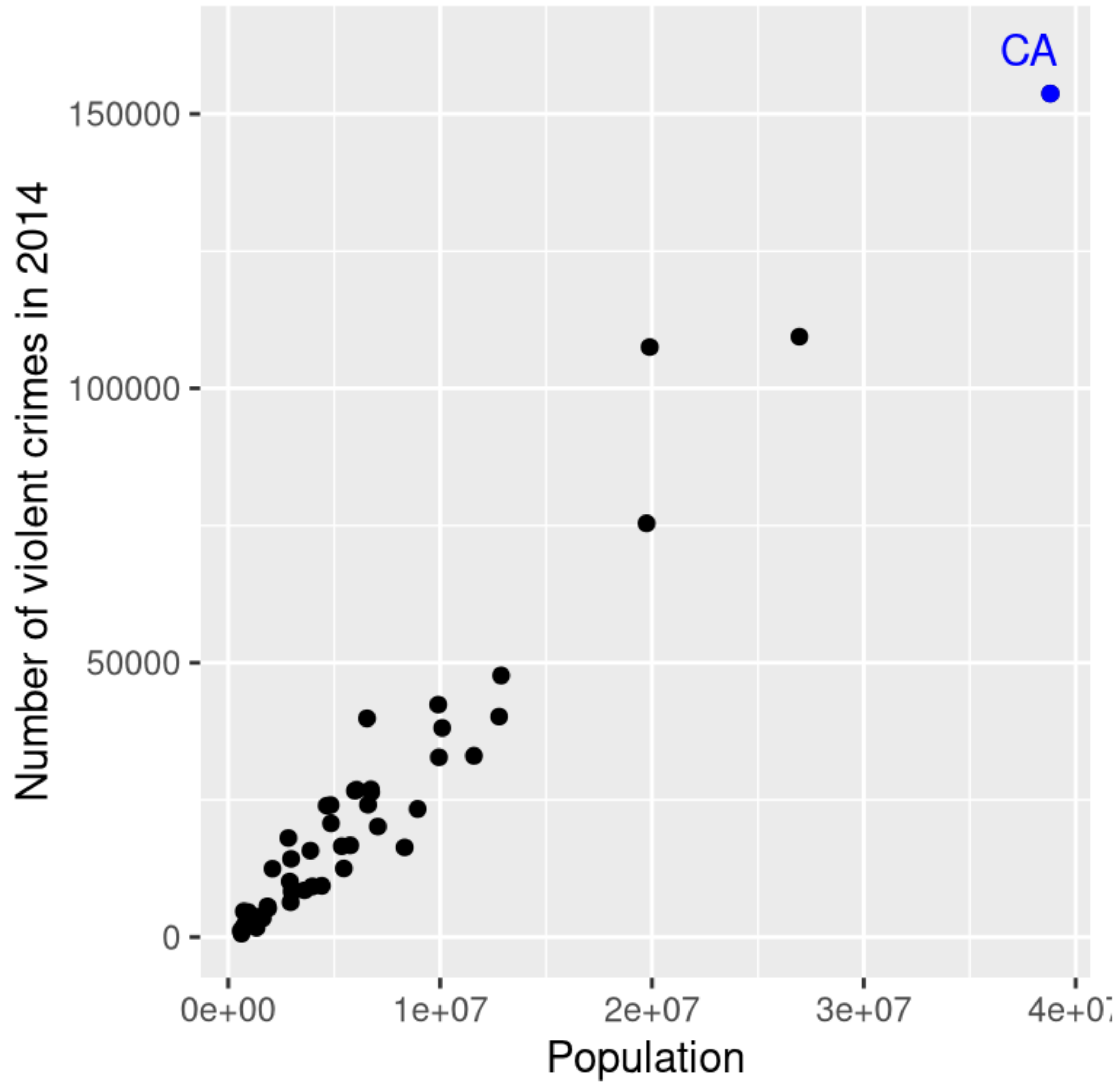
Map of the same data, with number of crimes (in thousands) plotted for each state in color.



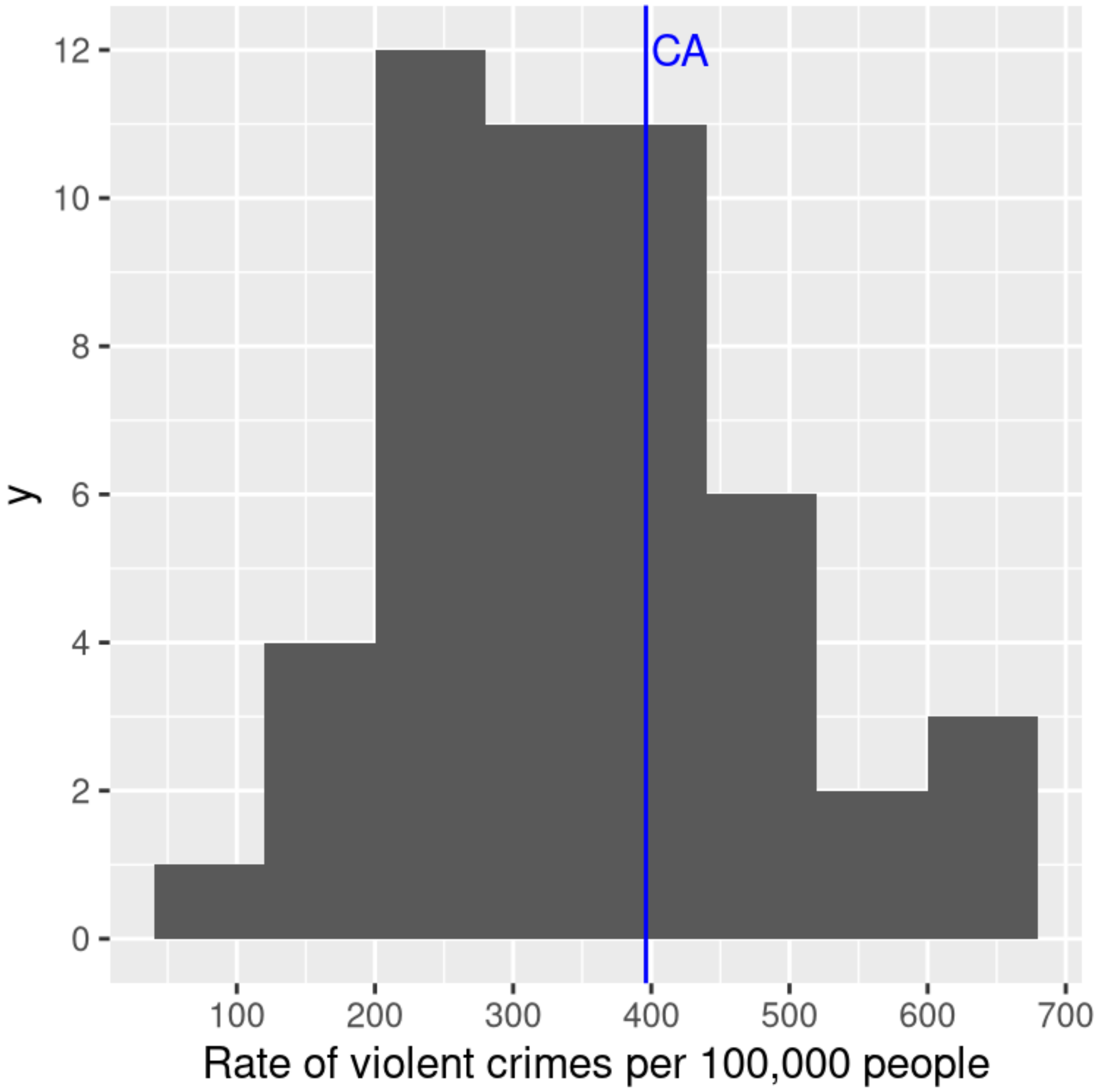
😱 **California is terribly dangerous, with 153,709 crimes in 2014!**

Z-scores example

Plot of number of violent crimes vs. population per state
(California highlighted in blue)



Histogram of per capita violent crime rates,
expressed as crimes per 100,000 people



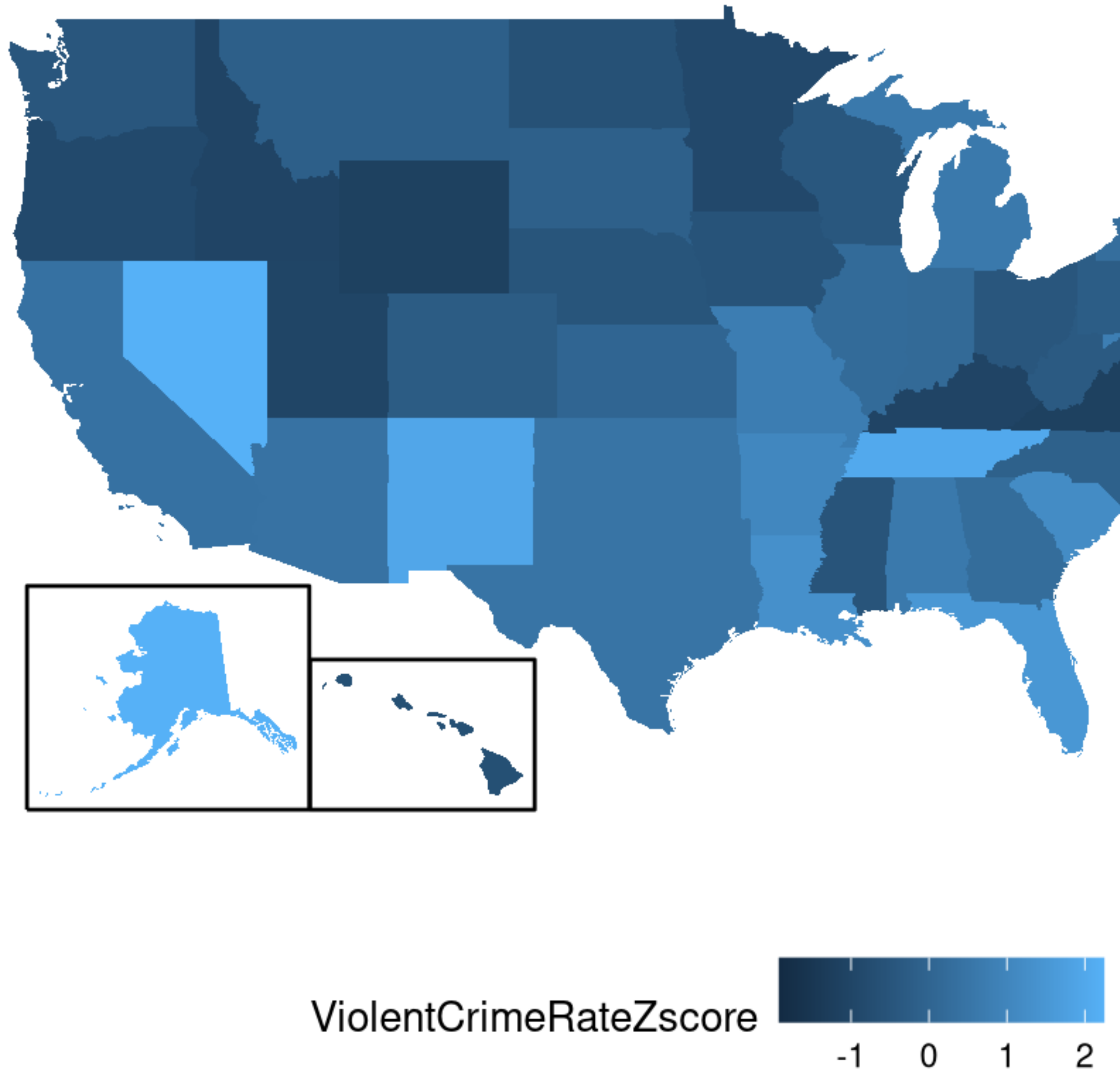
Perhaps California is not so dangerous—crime rate of 396.10 per 100,000 people is bit above the mean (346.81), but it doesn't look that bad..

Z-scores example

The *Z-score* allows us to express data in a way that provides more insight into each data point's relationship to the overall distribution

```
Z = (x - μ) / σ
```

Crime data rendered onto a US map, presented as Z-scores



California is within ~1 SD from the mean

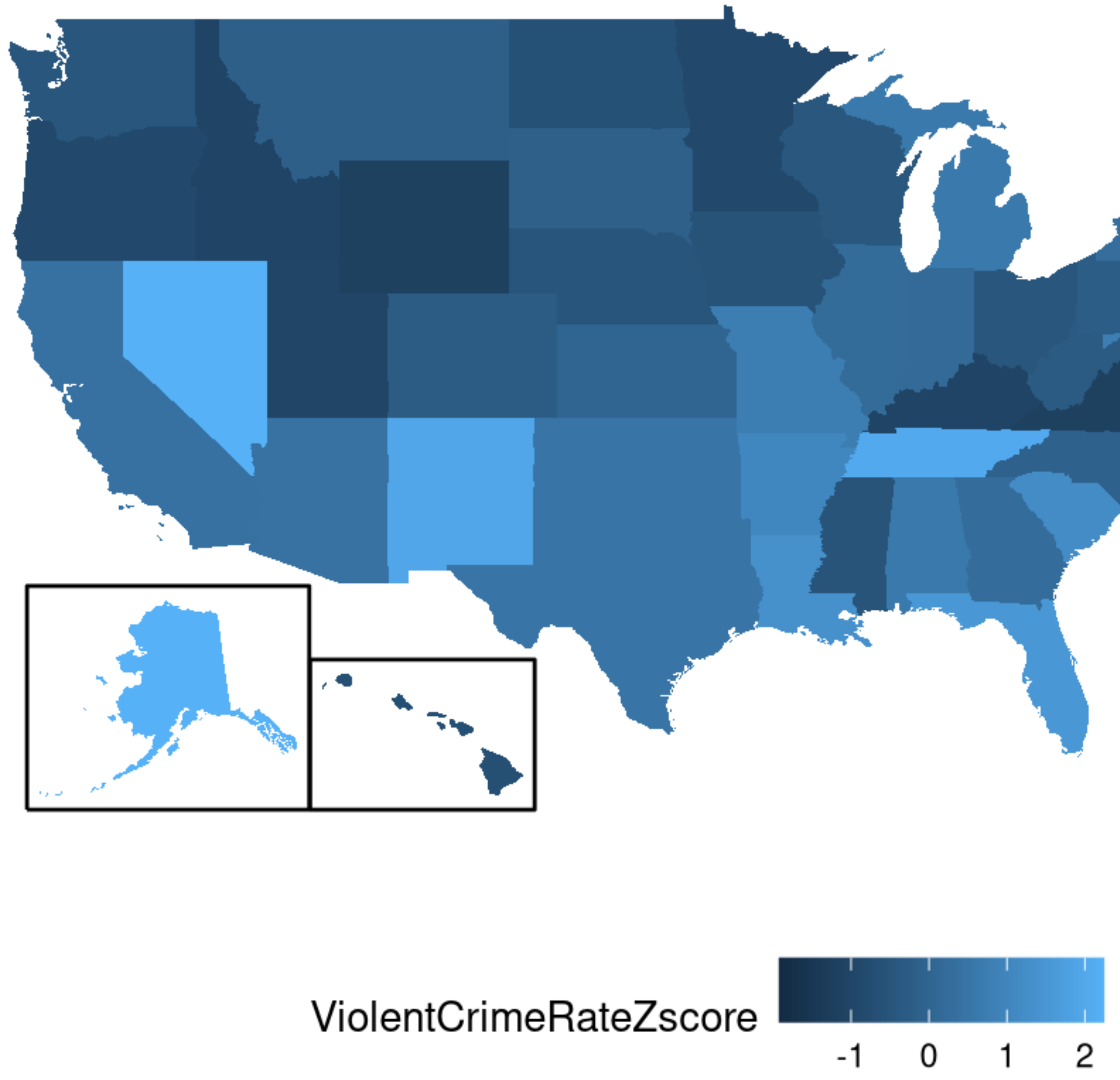
Nevada, Tennessee, and New Mexico all have crime rates that are roughly two standard deviations above the mean!

Z-scores example

The *Z-score* allows us to express data in a way that provides more insight into each data point's relationship to the overall distribution

```
Z = (x - μ) / σ
```

Crime data rendered onto a US map, presented as Z-scores



Z-score Guidelines

- $|Z| < 1$: Common
- $|Z| = 2$: Unusual (~2% of data)
- $|Z| > 3$: Very rare (potential outlier

Why Z-scores are useful

1. Comparing across different scales

Can compare different measures e.g., height (cm) to weight (kg) after standardizing

Person A:

- Height = 180 cm (mean = 170, SD = 10)
- Weight = 75 kg (mean = 70, SD = 12)

Are they more unusual in height or weight?

- Height Z: (180-170)/10 = +1.0
- Weight Z: (75-70)/12 = +0.42

More unusual in height than weight

2. Identifying outliers

Values with $|Z| > 3$ are unusual

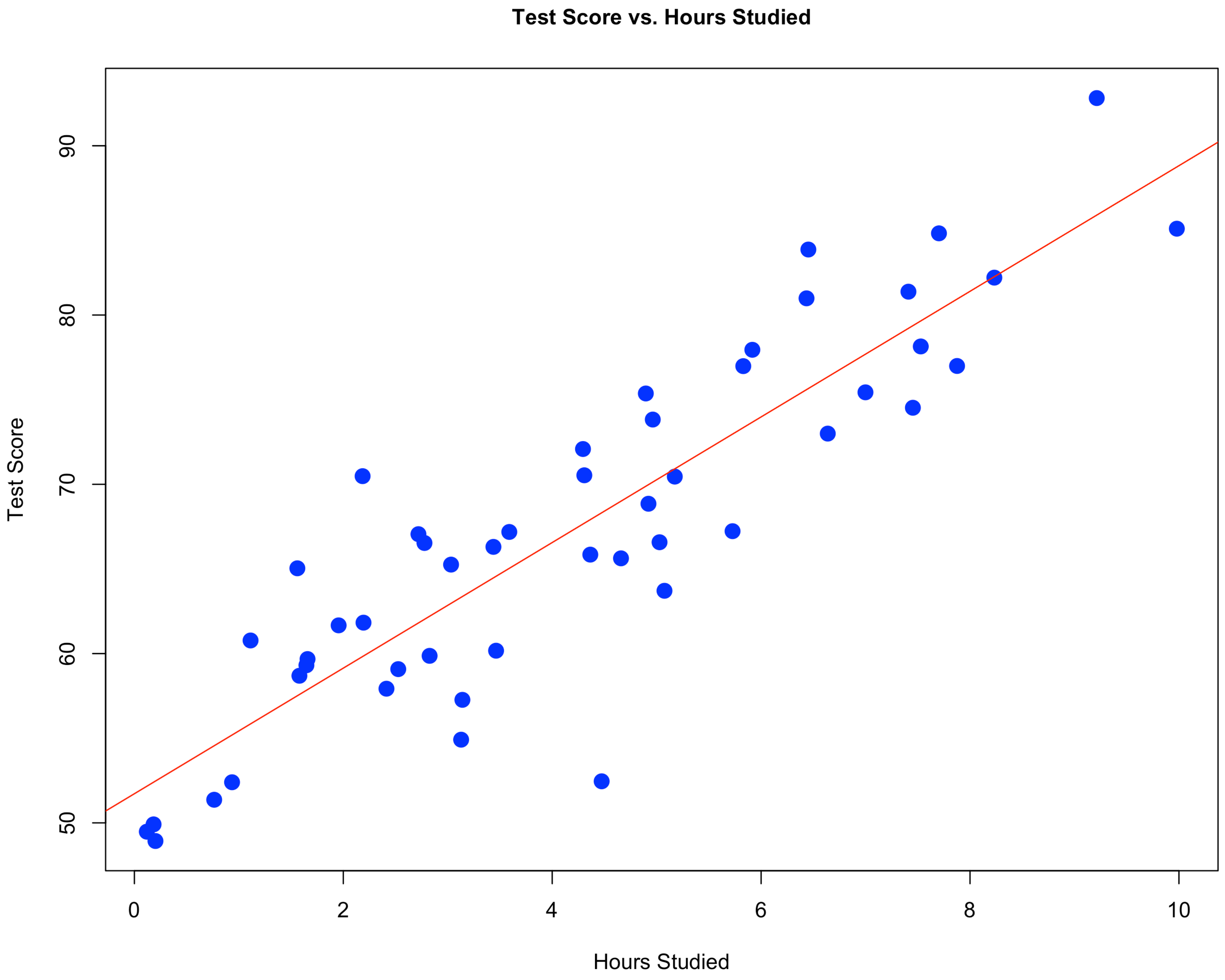
Summary

- **Variability matters:** Central tendency alone doesn't tell the full story
- **Multiple measures:** Range, IQR, variance, SD - each with strengths/weaknesses
- **Degrees of freedom:** Use (n-1) for unbiased variance estimate
- **Z-scores:** Standardize to compare across different scales

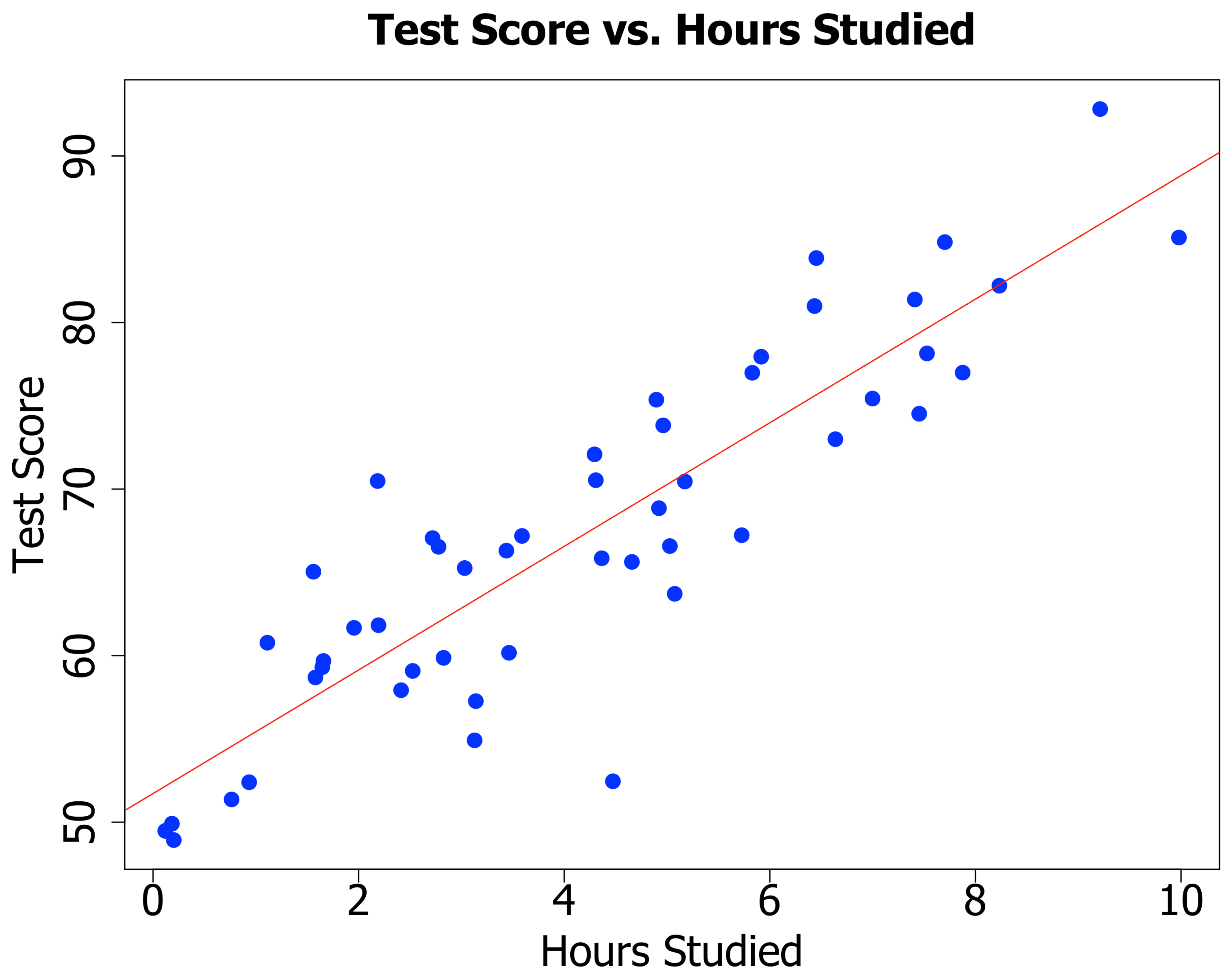
Up Next: Data visualization - make your papers standout!

Some guidelines on making pretty publication-ready data visuals

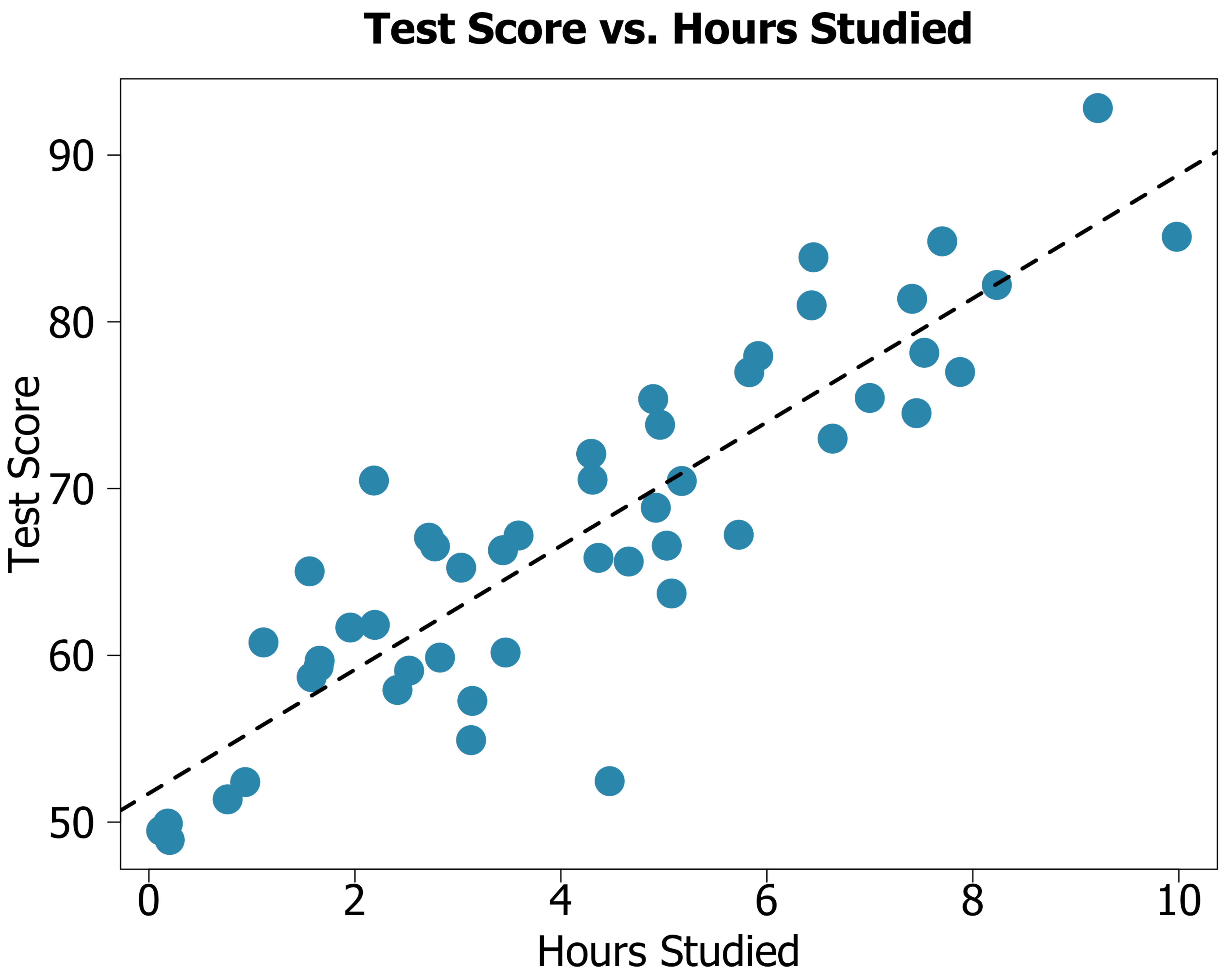
A simple scatterplot



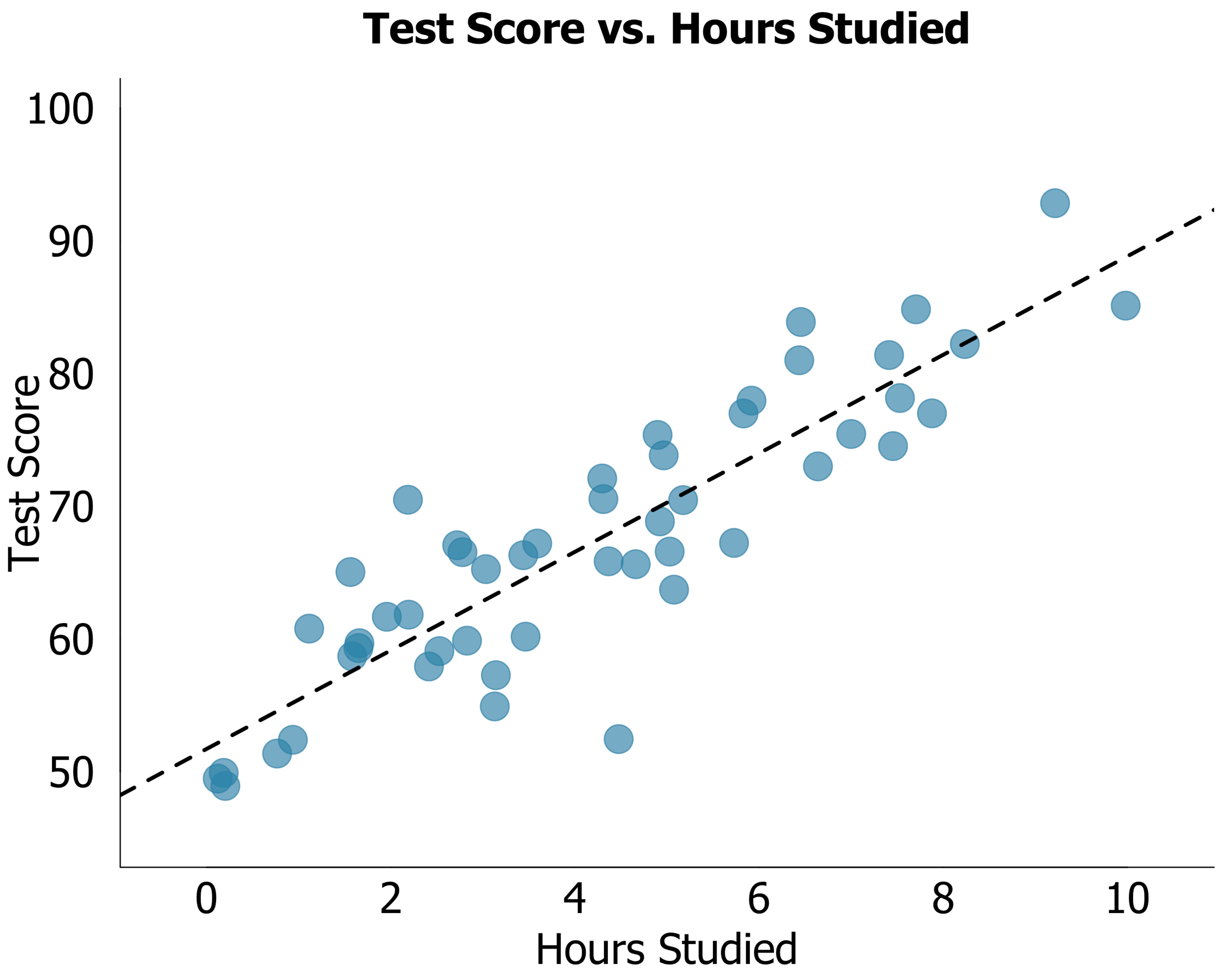
Step 1: Choose a nice font and make labels bigger



Step 2: Fix color scheme and make points bigger



Step 3: Make plot more minimal

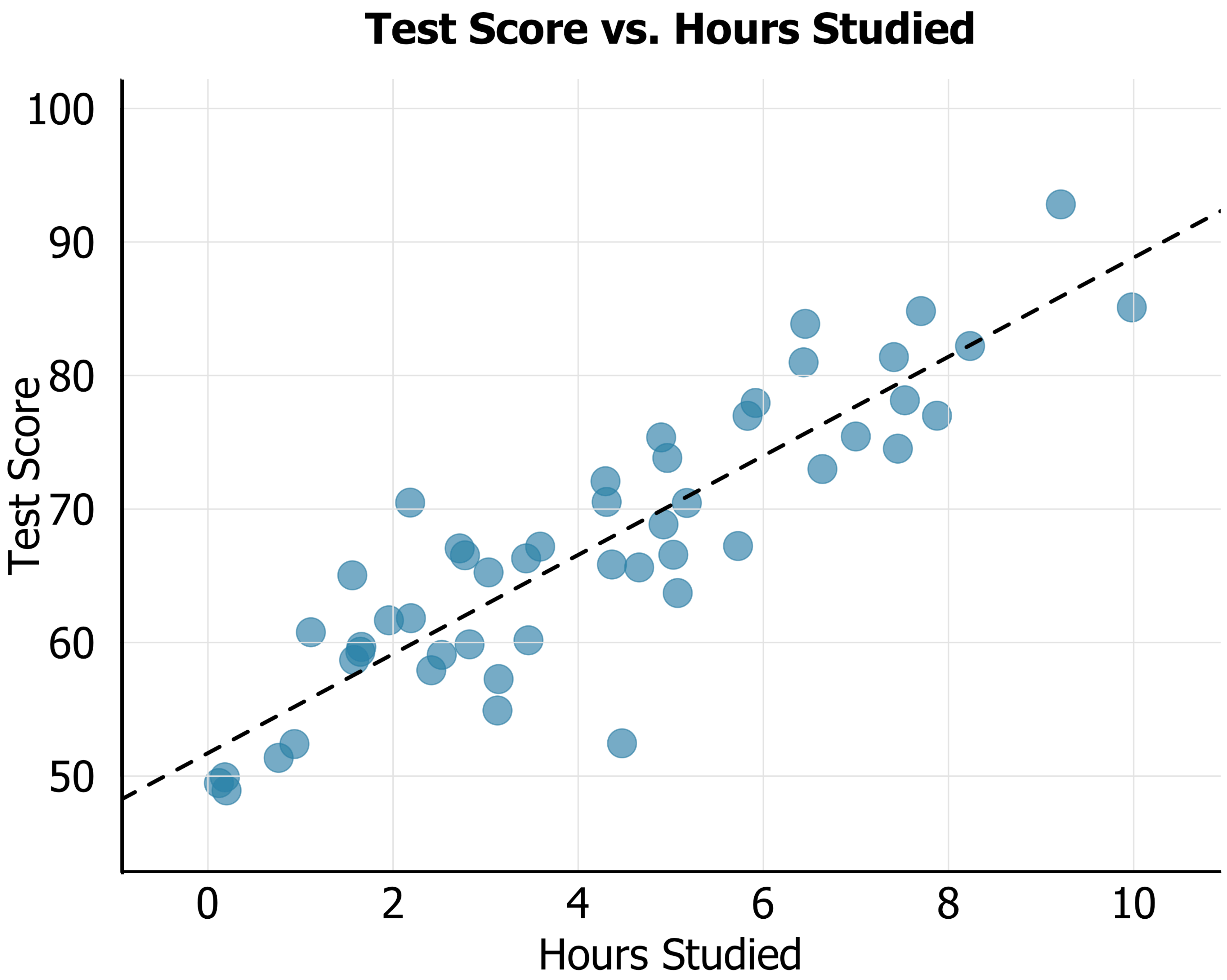


Use alpha values (when plot is very dense)

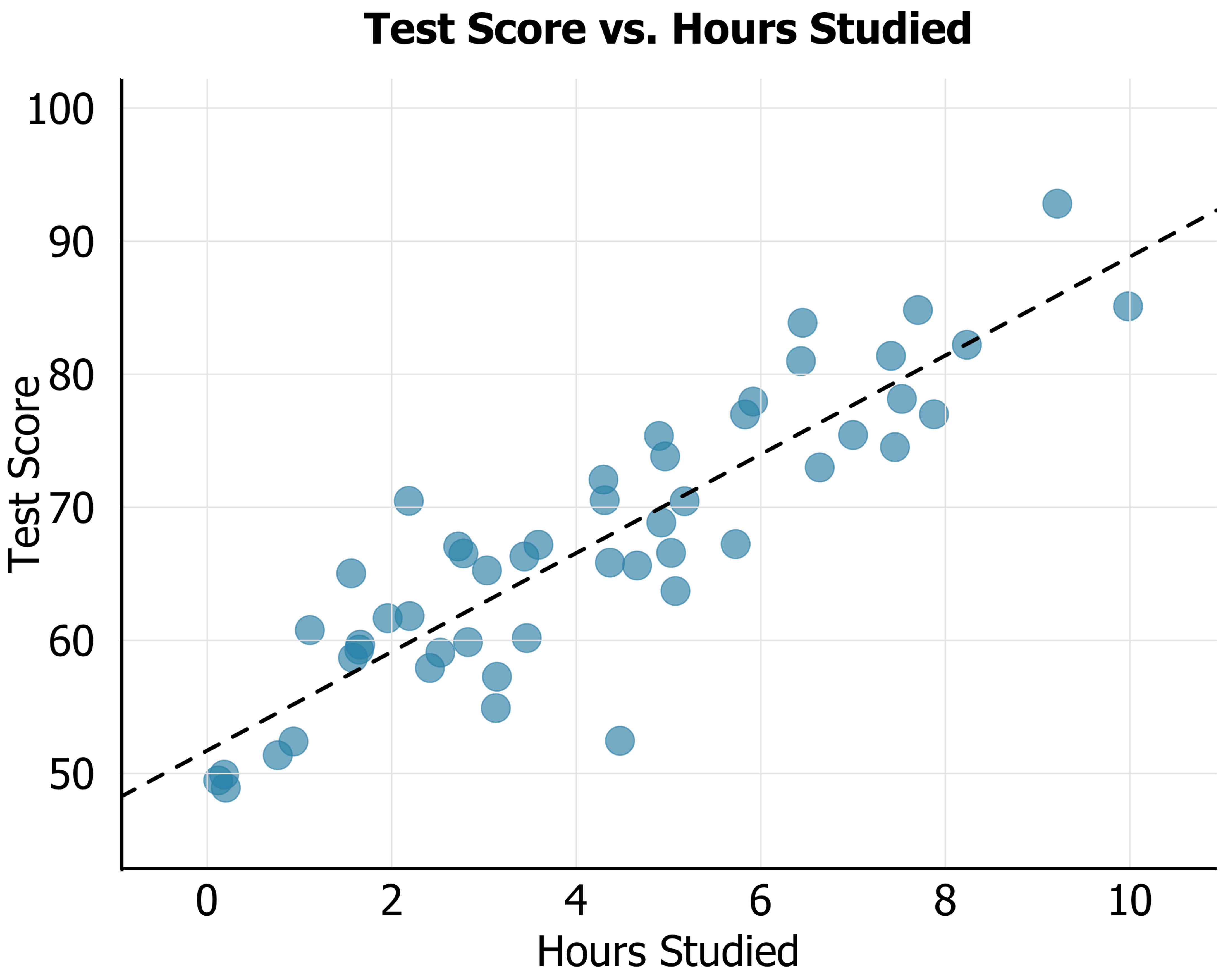
Reduce clutter (e.g., tick marks) or borders

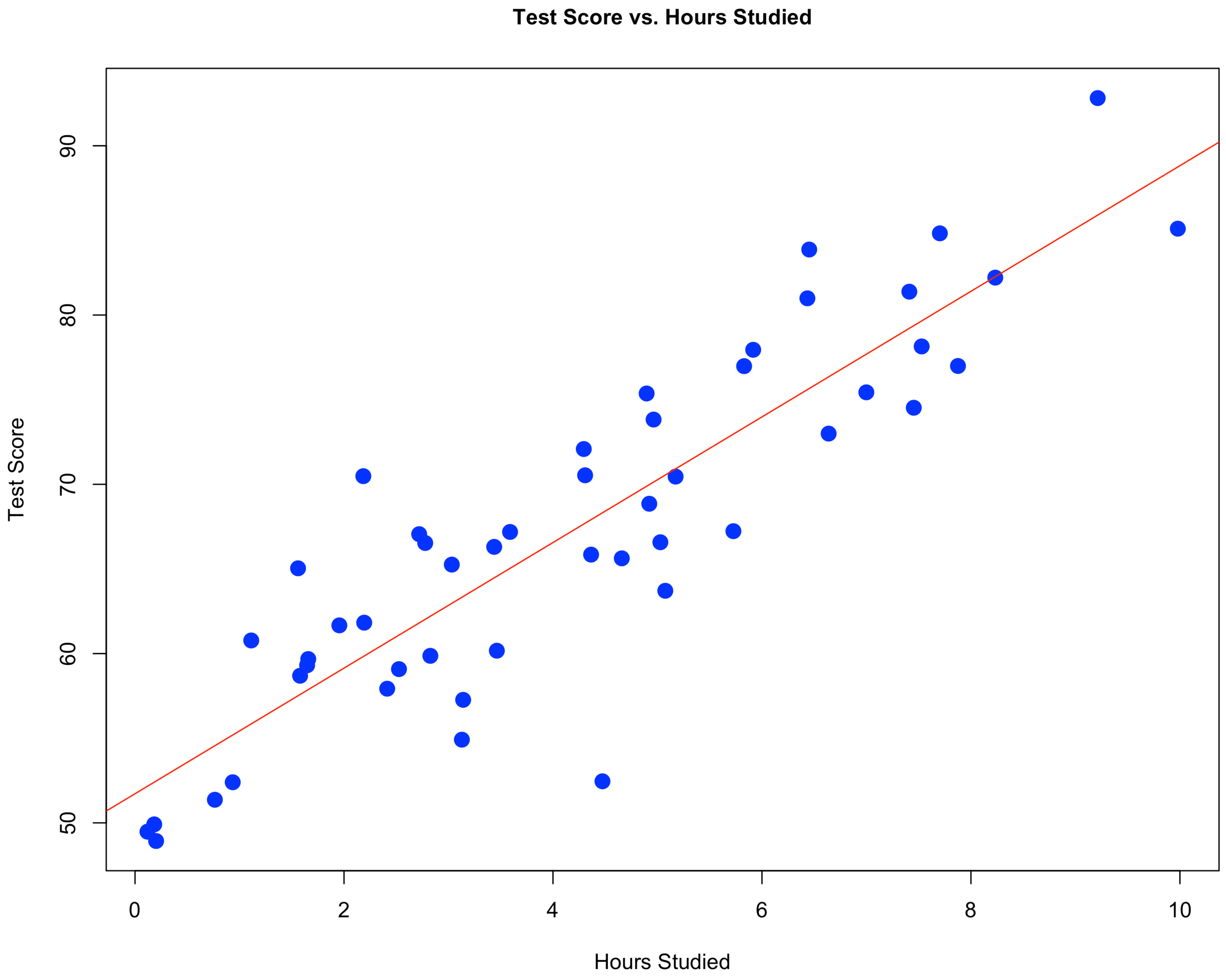
Add padding (when necessary)

Step 4: Finishing touches

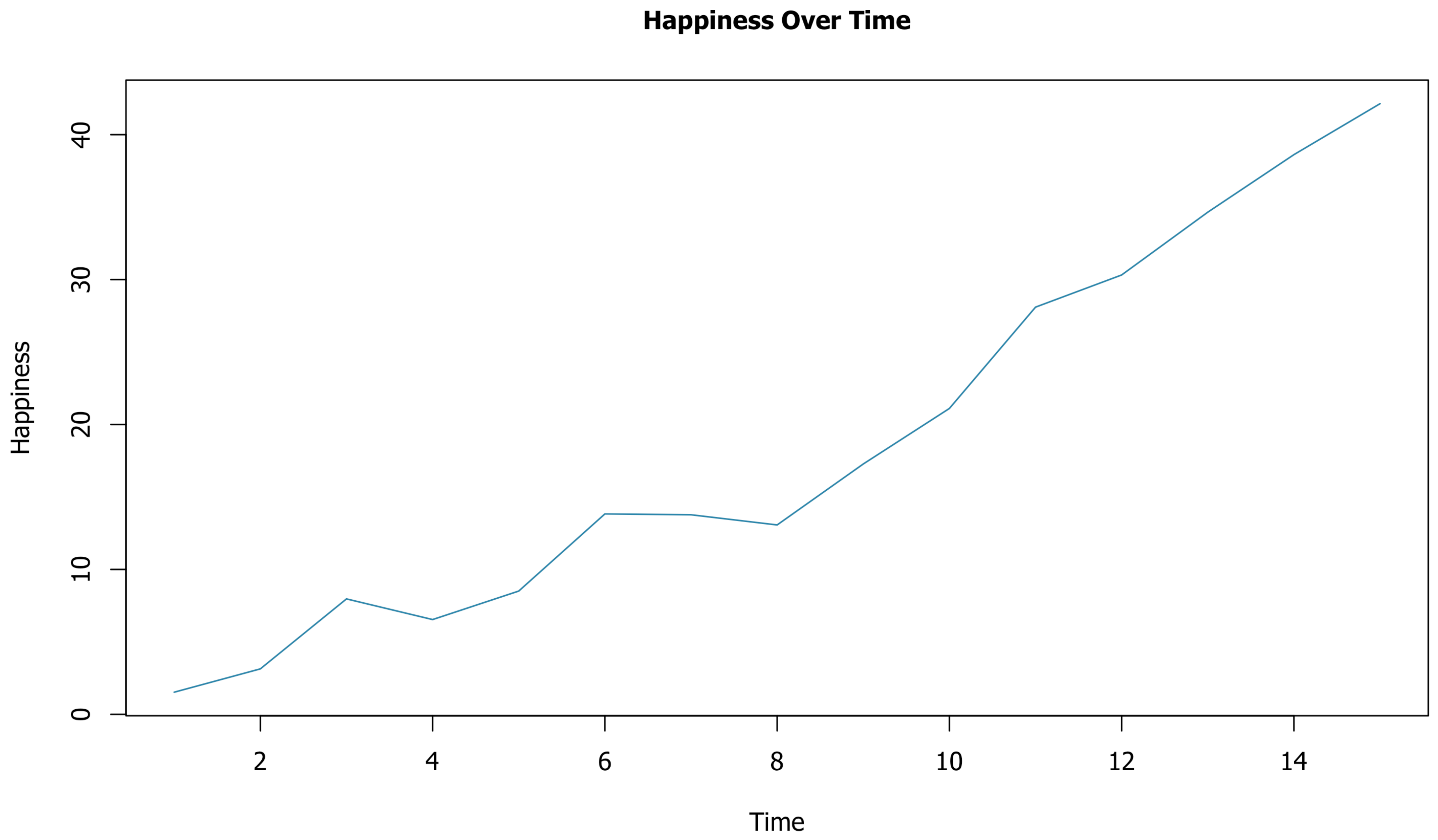


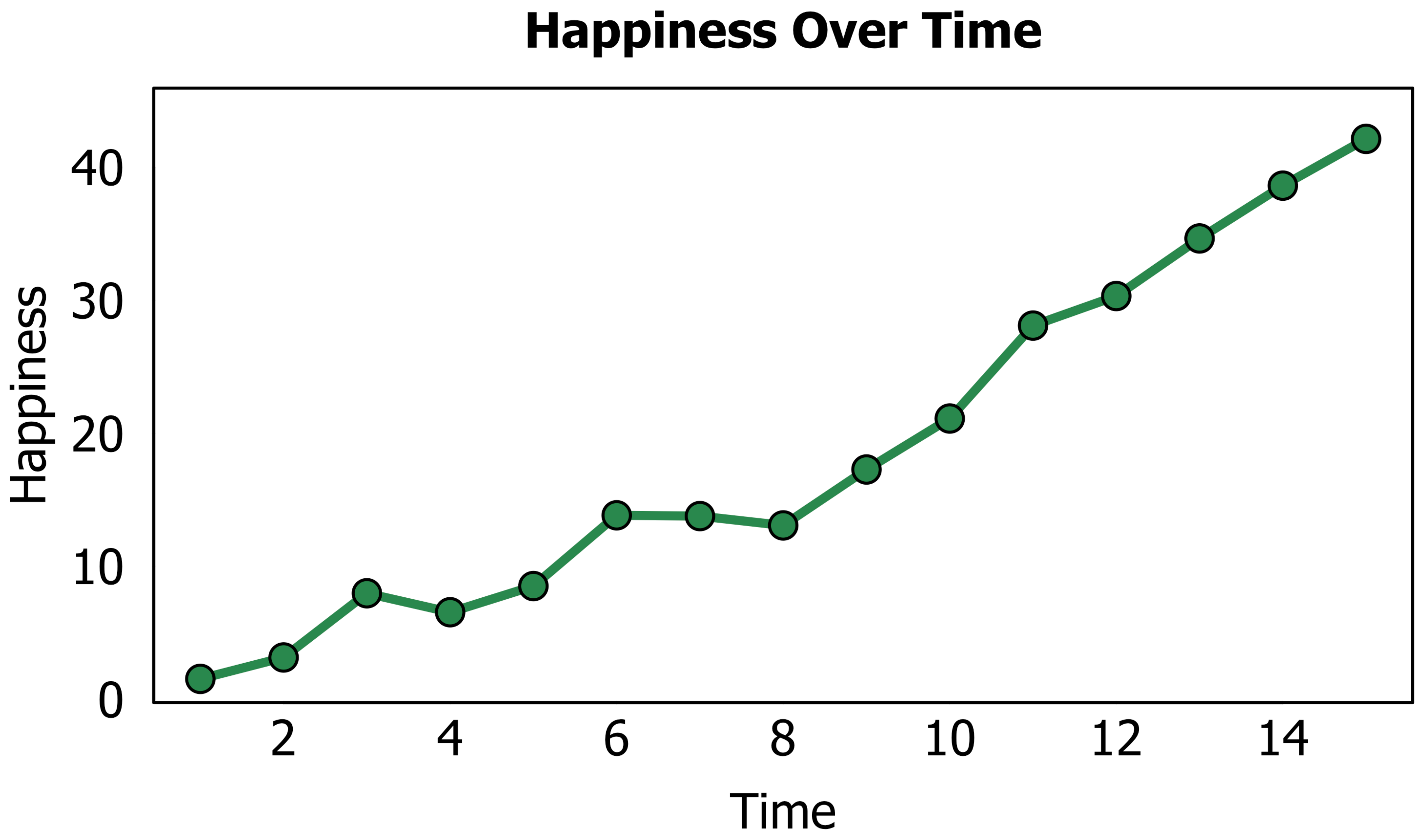
Simple things make a big difference!



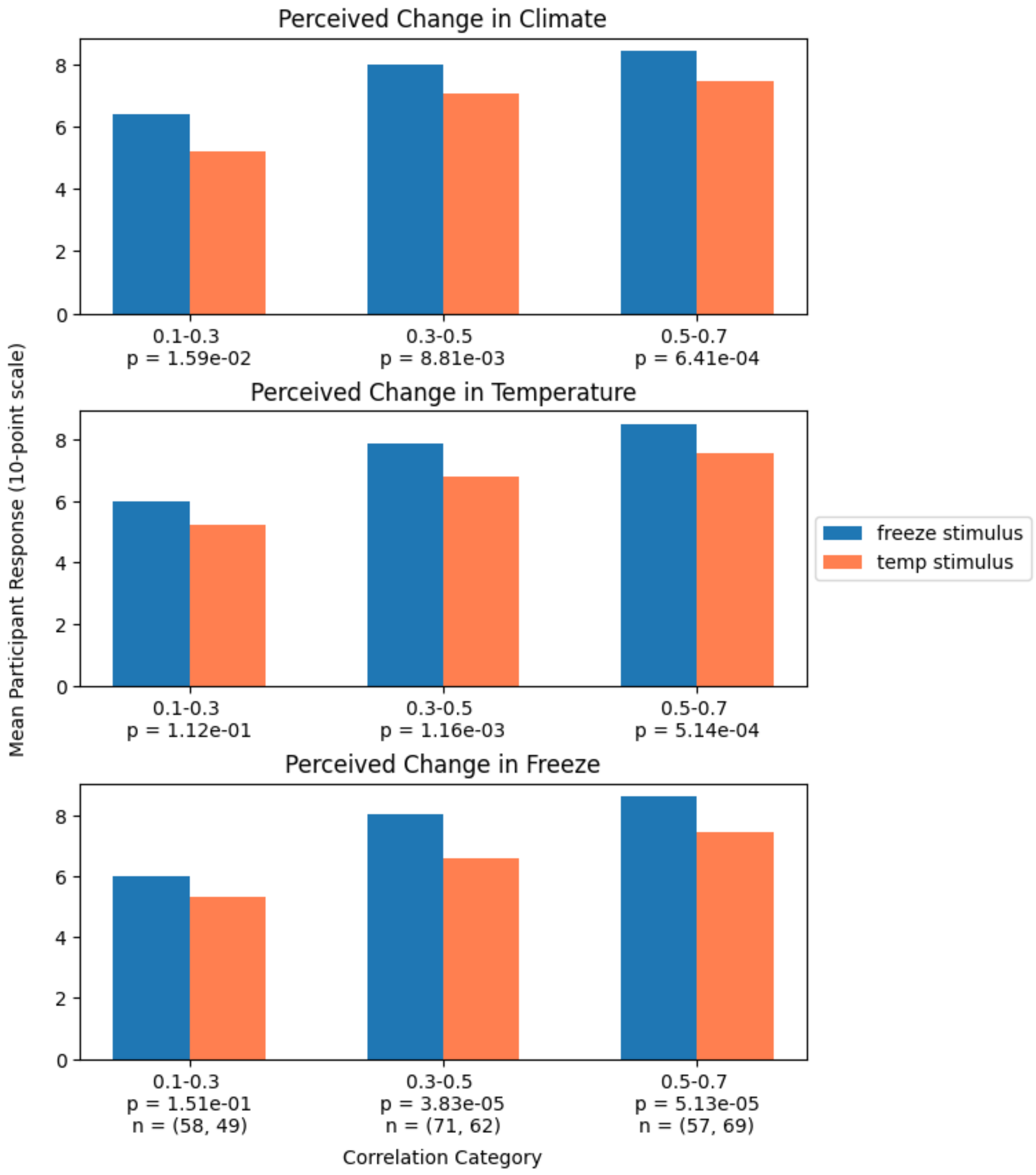


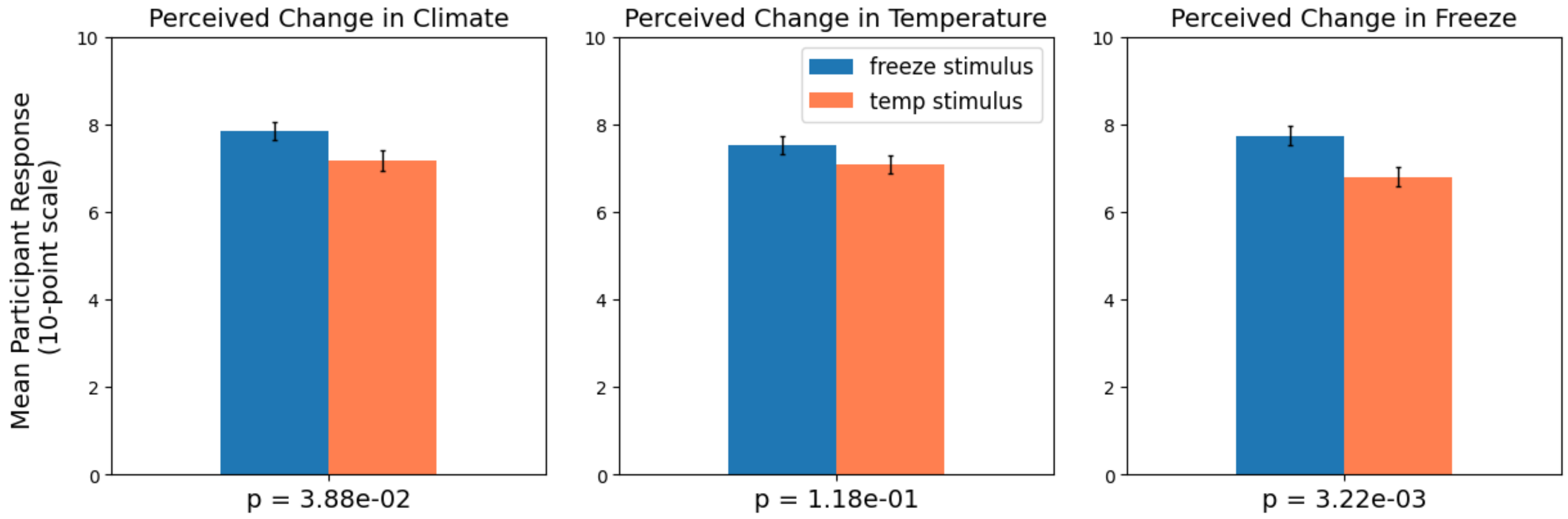
Another simple example

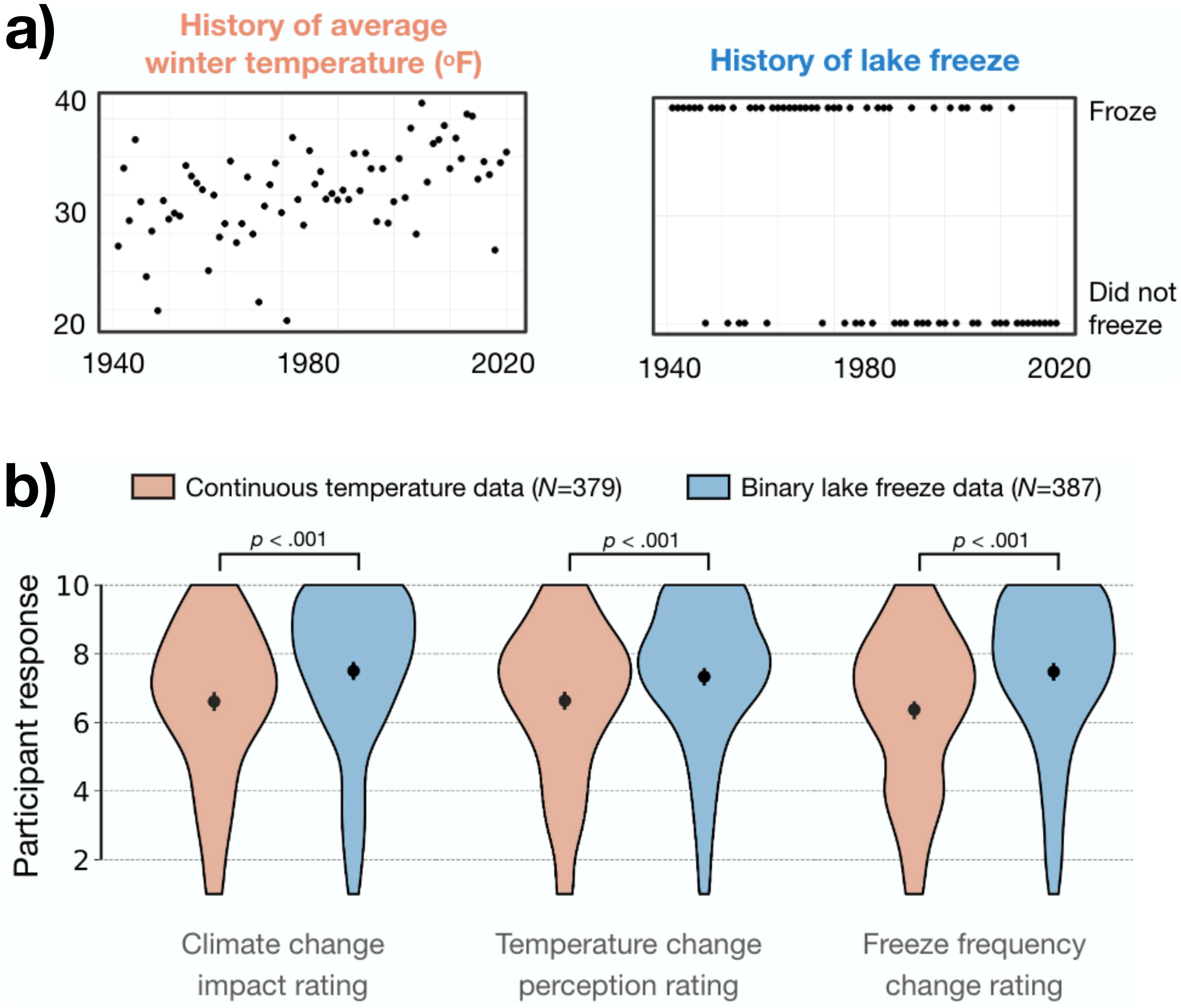




Examples from my paper

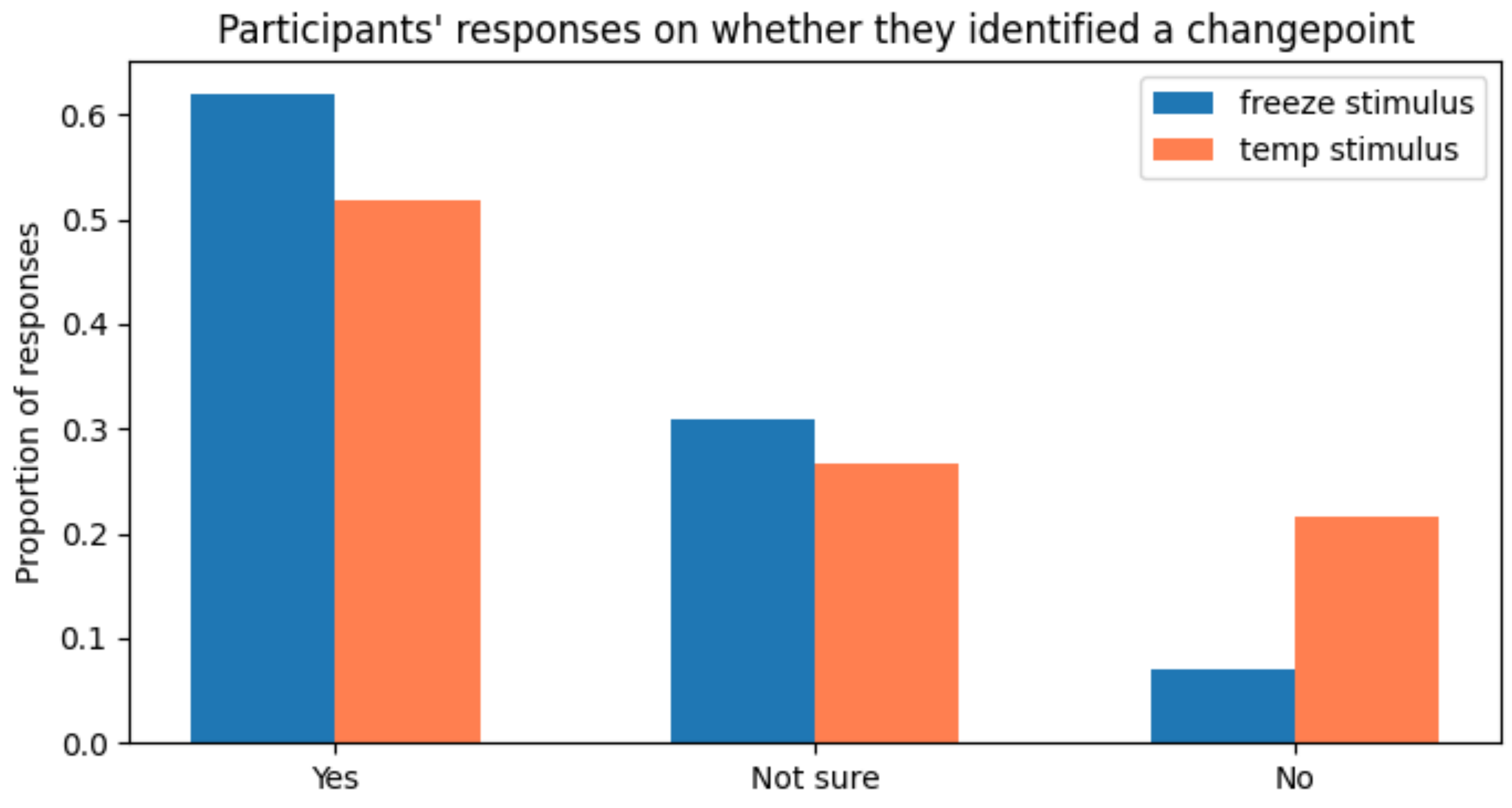



Perceived Change in Climate
Perceived Change in Temperature
Perceived Change in Freeze
Mean Participant Response
(10-point scale)
0
2
4
6
8
10
freeze stimulus
temp stimulus
p = 3.88e-02
p = 1.18e-01
p = 3.22e-03

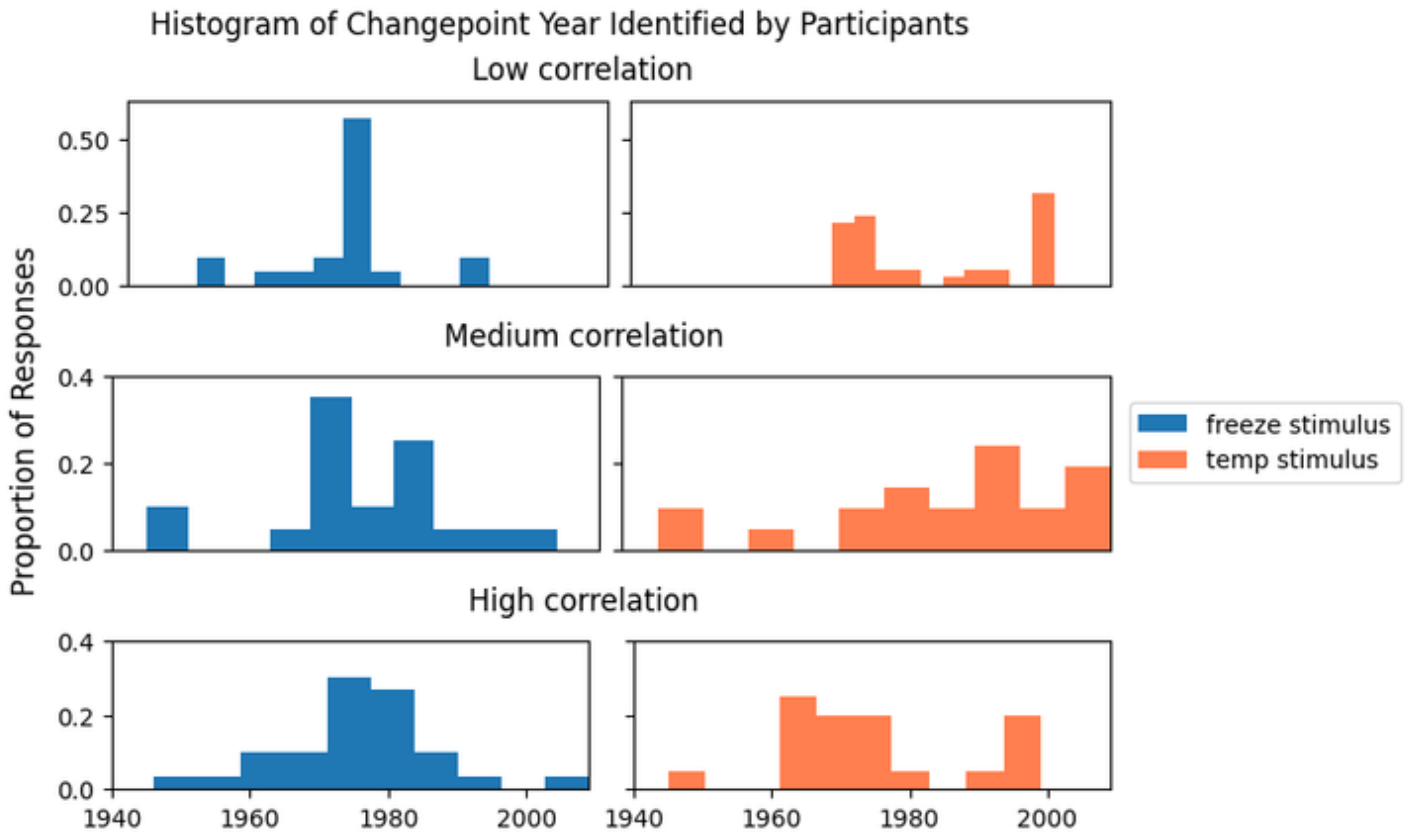

a)
History of average winter temperature (°F)
40
30
20
1940
1980
2020
History of lake freeze
Froze
Did not freeze
1940
1980
2020
b)
Continuous temperature data (N=379)
Binary lake freeze data (N=387)
p < .001
p < .001
p < .001
Participant response
10
8
6
4
2
Climate change impact rating
Temperature change perception rating
Freeze frequency change rating

Another example..

a)

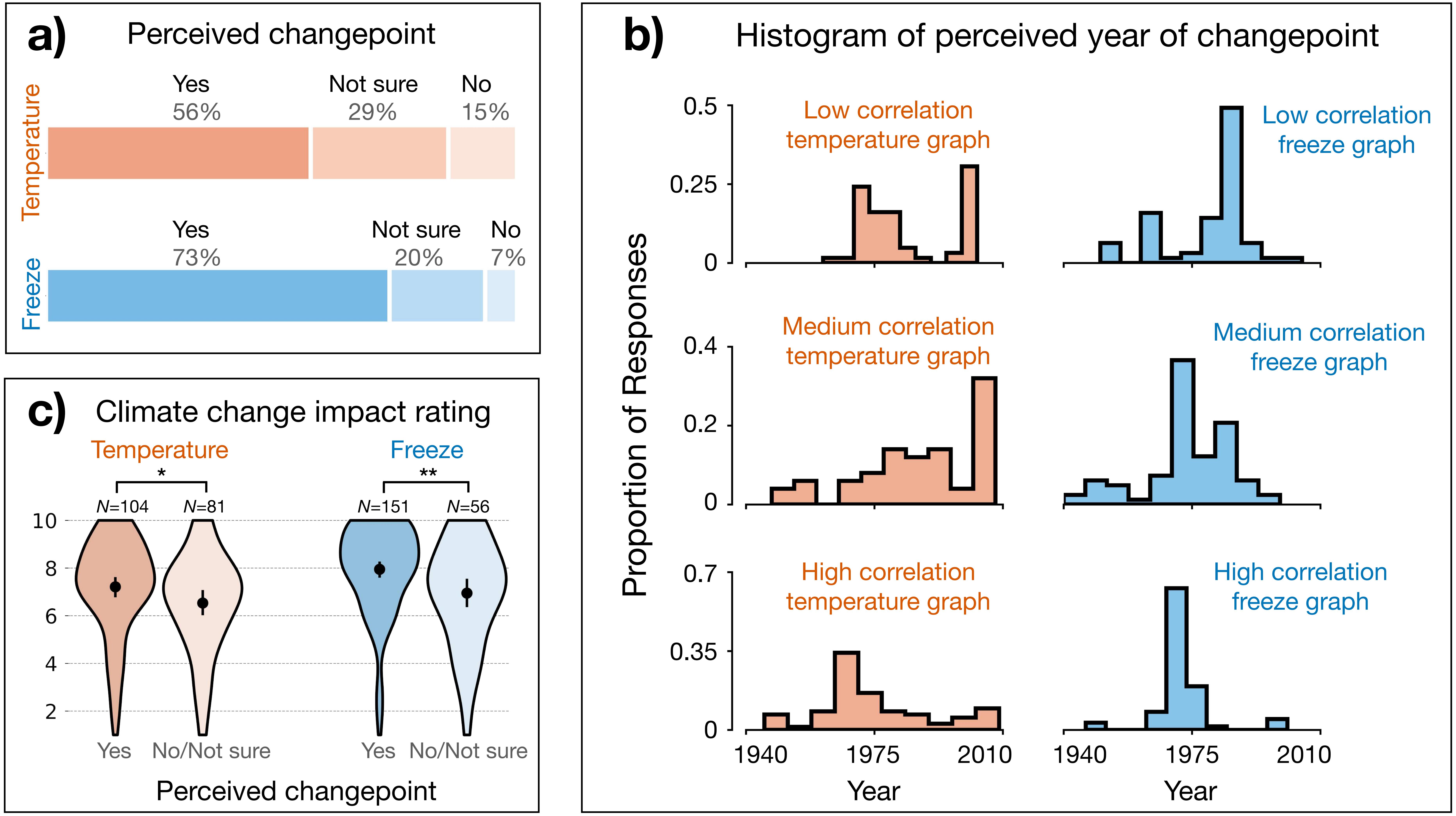


b)



c)

Restructuring graphs to fit nicely in a paper



Summary

- Use BIG and CLEAN fonts
- Make axis labels clear and easy to read
- Increase line width and marker sizes
- Avoid bar graphs (when possible)
- Group together figures nicely
- Figure panels should tell a story
- Minimize clutter, remove tick marks and borders when not needed
- Use clear headings..
- People should be able to understand your results without reading the captions!