CHRISTOPHER PRENER, PH.D.
FALL, 2016

WEEK 03
LECTURE 05

QUANTITATIVE ANALYSIS

DESCRIBING DISTRIBUTIONS

AGENDA

1. Follow-up
2. Getting Organized
3. Describing Distributions
4. Visualizing Distributions
5. Anscombe's Quartet

1 FOLLOW-UP

2 GETTING ORGANIZED

KEY QUESTIONS

- How do you organize files?
- Do you keep different versions of files as your assignment or project progresses?
- If you needed your files in 5 years, could you find them?
- If you needed your files in 5 years, could you open them?
- Do you backup files ever?
- If your house was robbed or burned down, would your backup also be destroyed?

KEY QUESTIONS

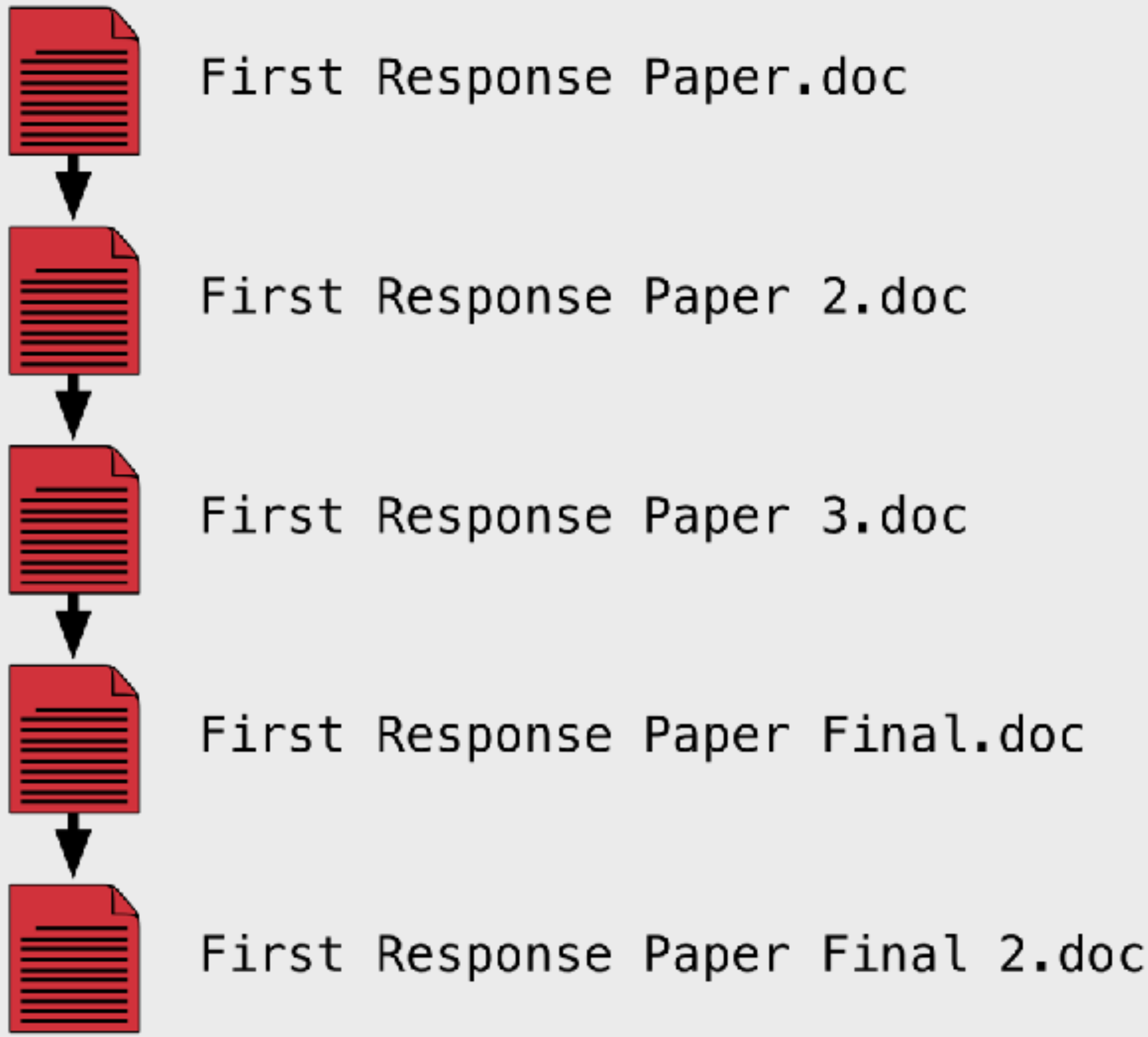
- How do you organize files?
- Do you k ... project progres
- If you ne
- If you needed your files in 5 years, could you open them?
- Do you backup files ever?
- If your house was robbed or burned down, would your backup also be destroyed?

Git & GitHub can help you address all of these key questions/issues!

GIT WORKFLOW

Top-level directories in Git are called repositories. All files placed in a "repo" are tracked unless Git is explicitly told not to track them.

TYPICAL WORKFLOW



TYPICAL WORKFLOW

GIT WORKFLOW

First Response Paper.doc

responsePaper1.md

Commits are snapshots of files that are saved at particular points in time.

TYPICAL WORKFLOW

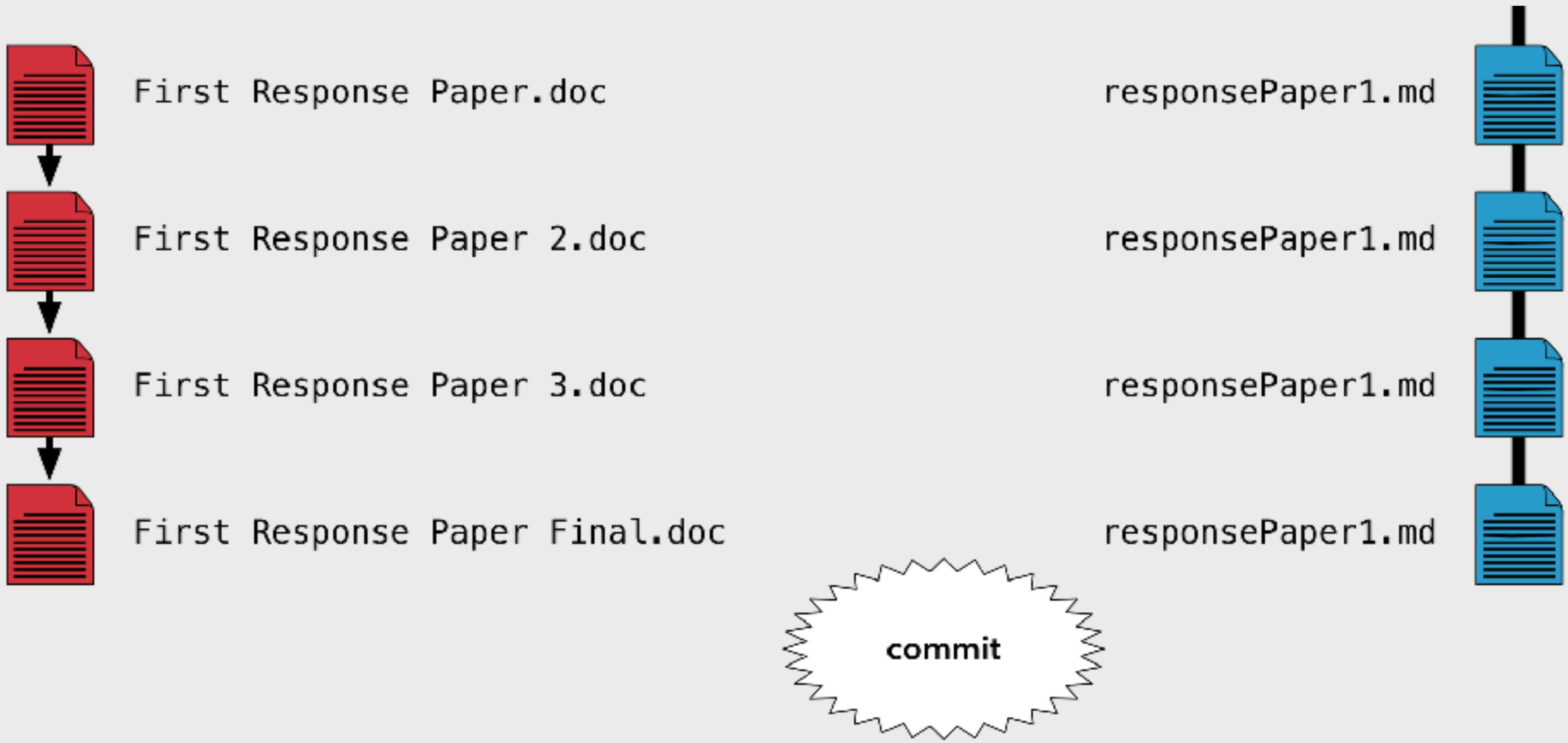
GIT WORKFLOW

TYPICAL WORKFLOW

GIT WORKFLOW

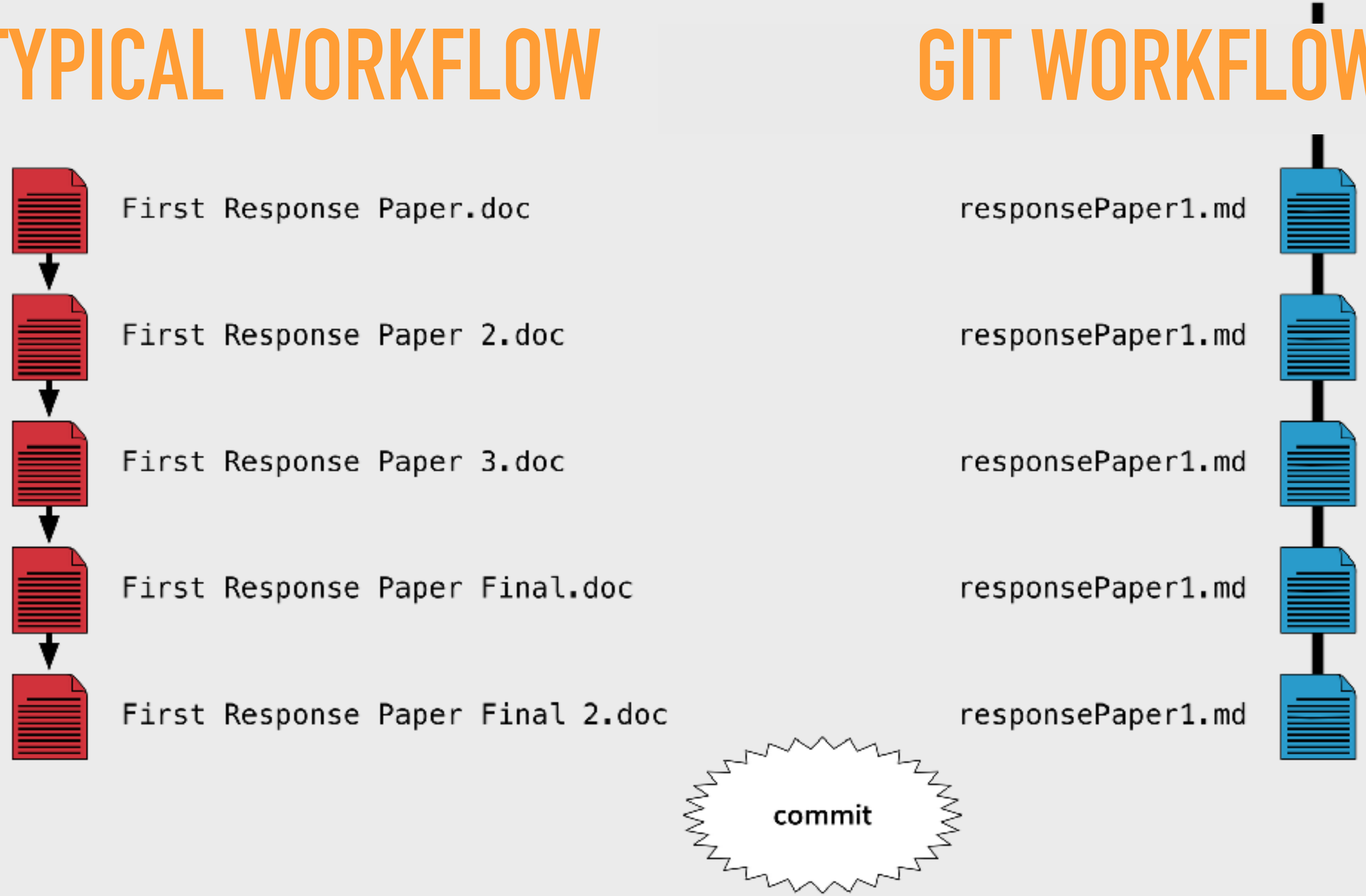
TYPICAL WORKFLOW

GIT WORKFLOW

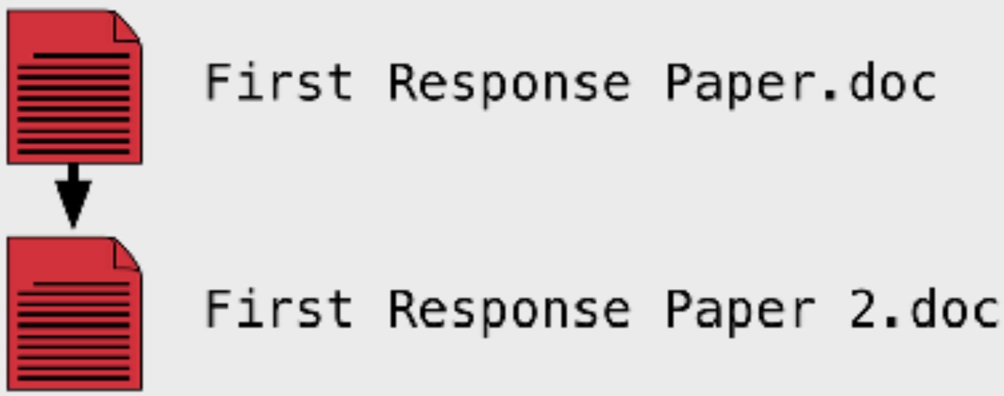


TYPICAL WORKFLOW

GIT WORKFLOW



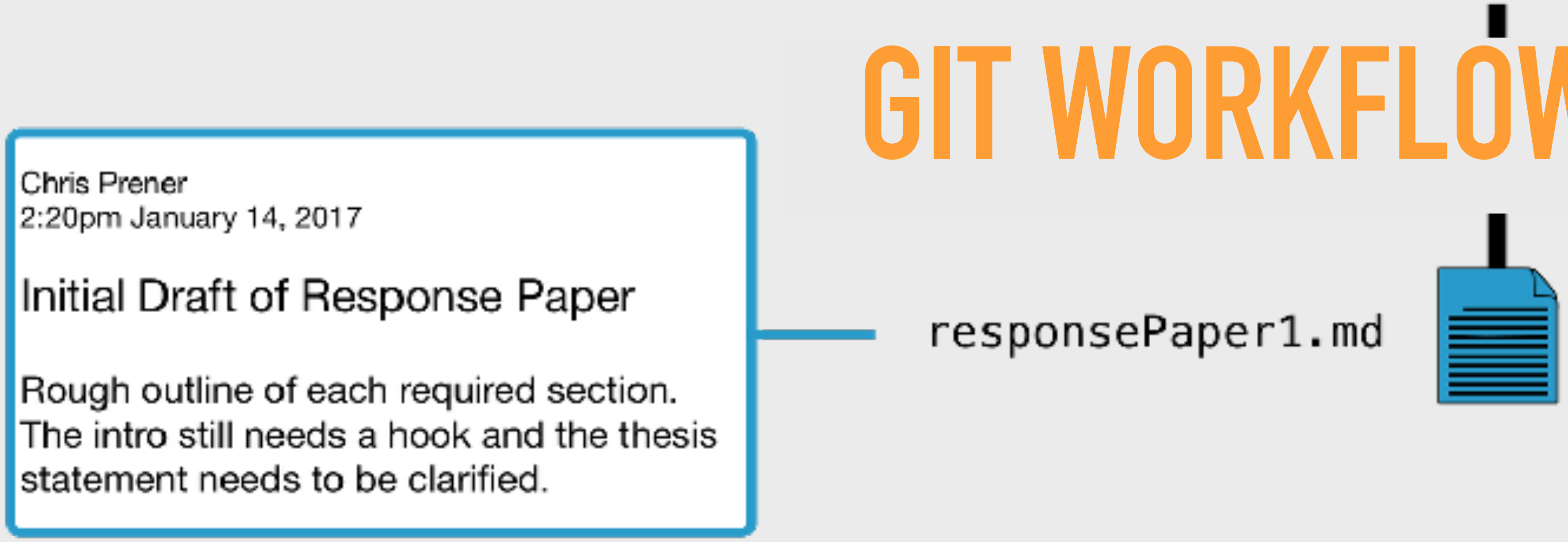
TYPICAL WORKFLOW



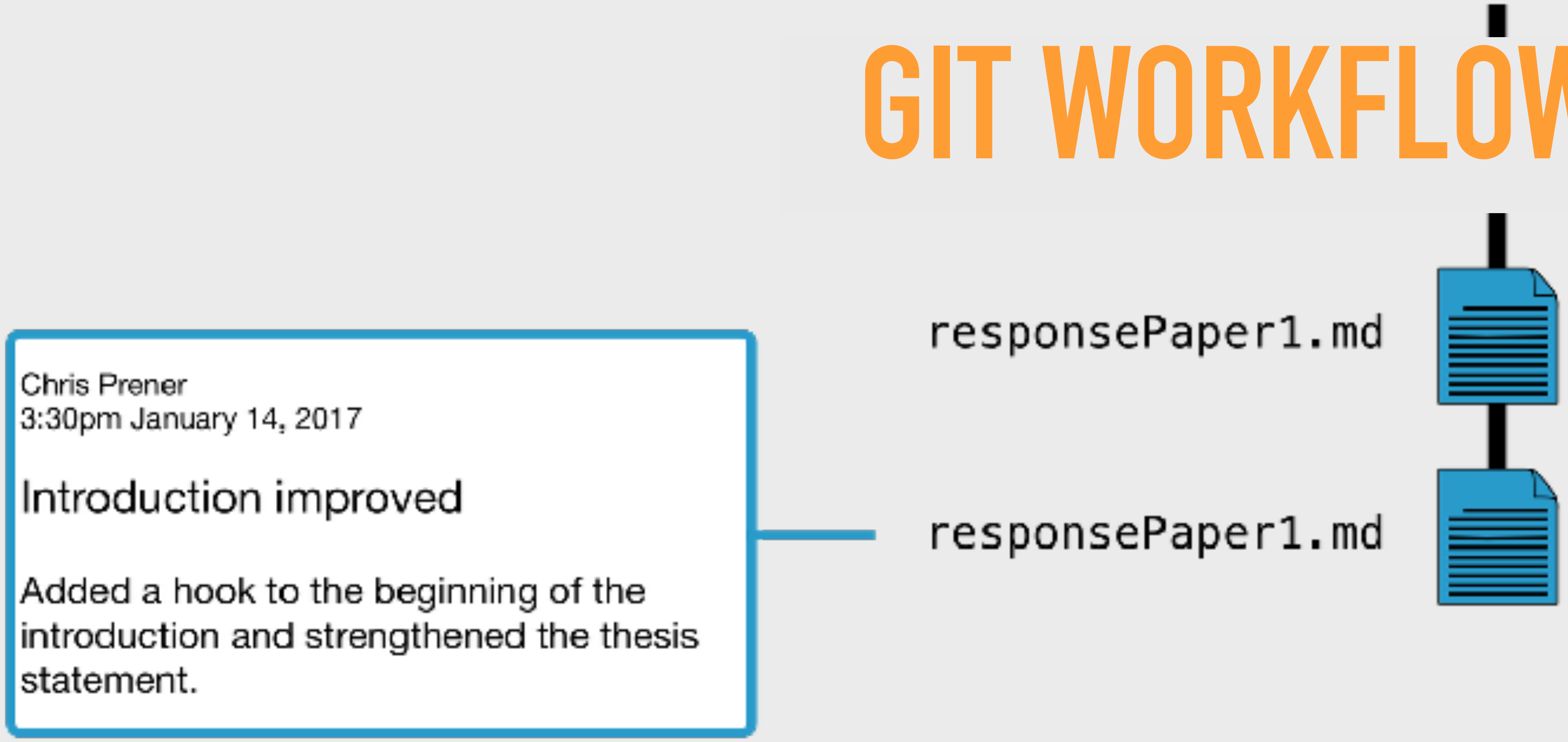
"OK, so why the $#&% did I save a second copy?!?!?!

And why the $#&% was the first copy edited *after* the second copy!?!?!?

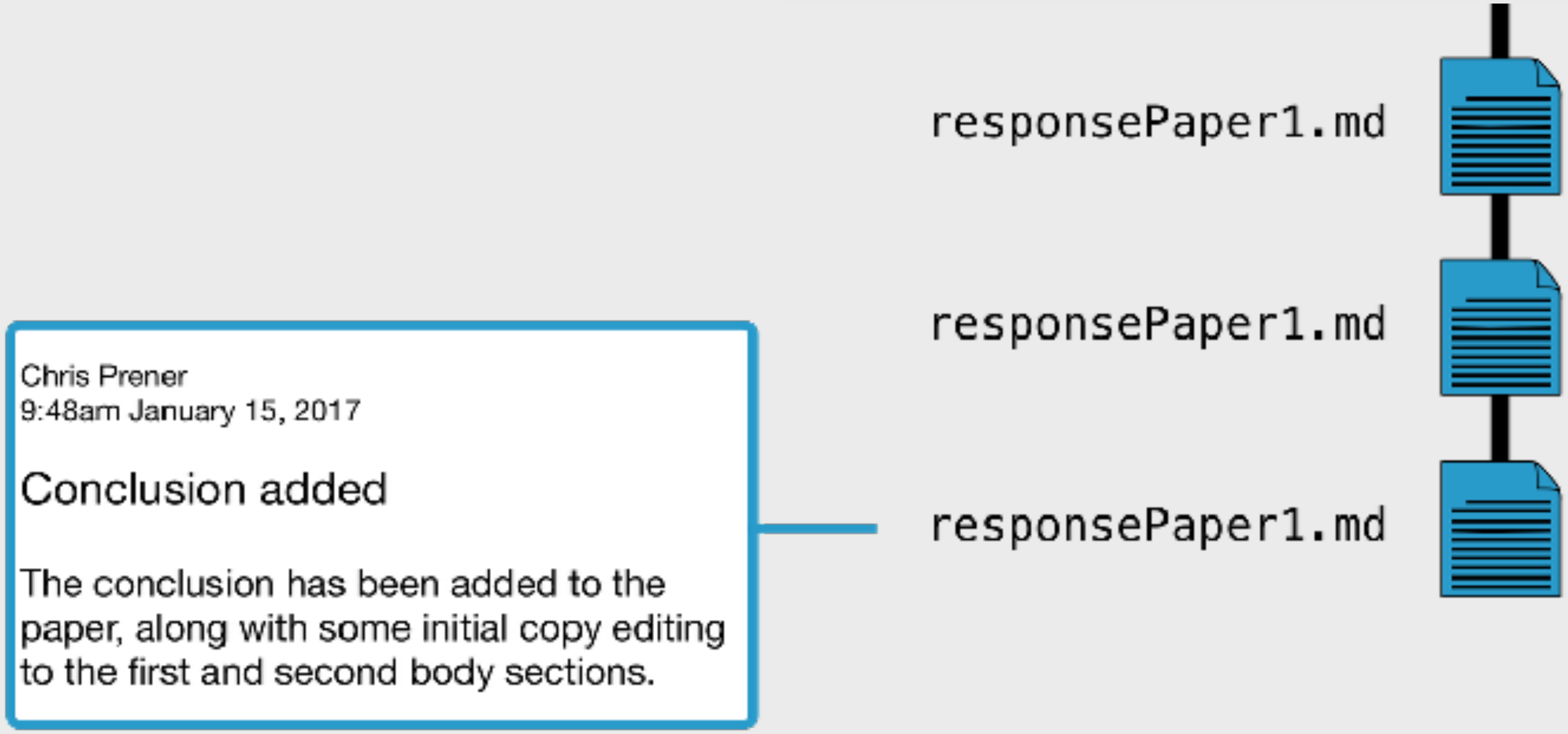
GIT WORKFLOW



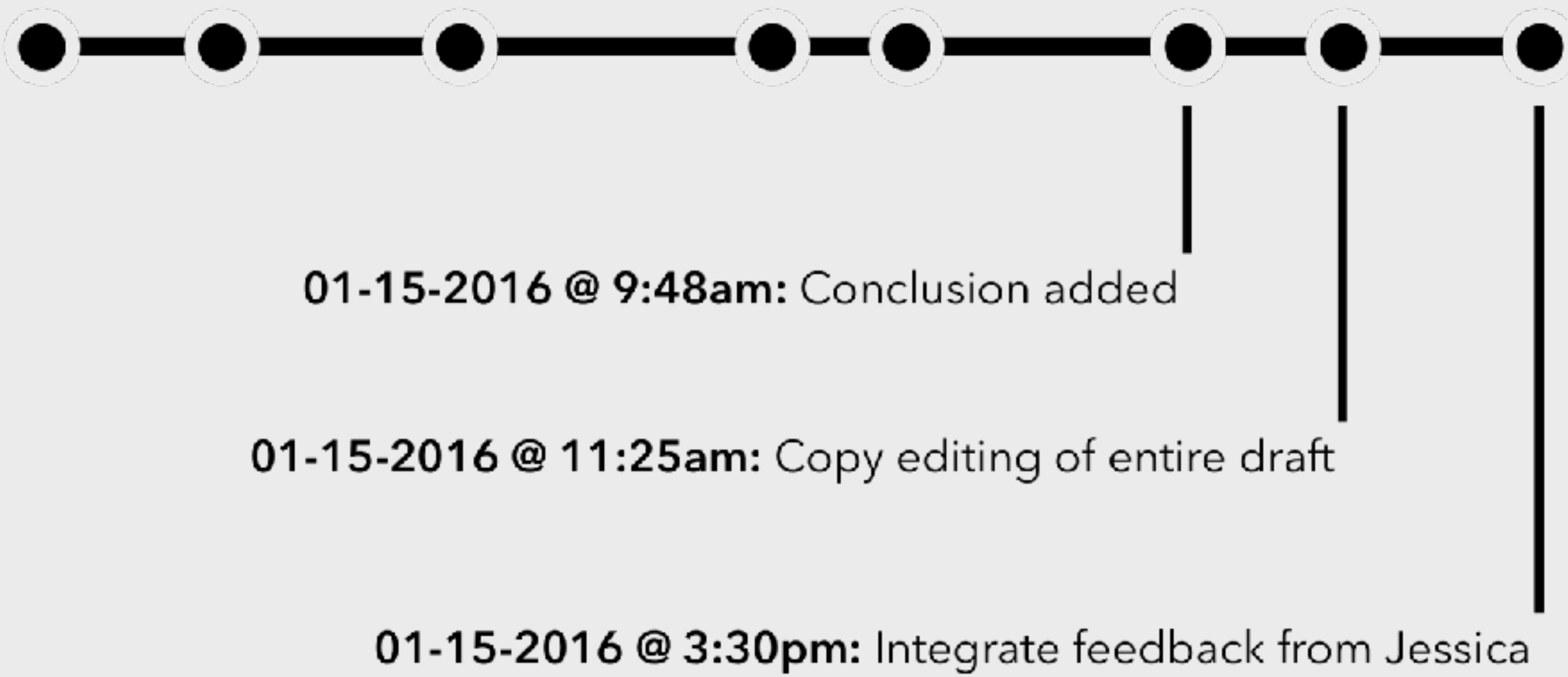
GIT WORKFLOW



GIT WORKFLOW



GIT WORKFLOW



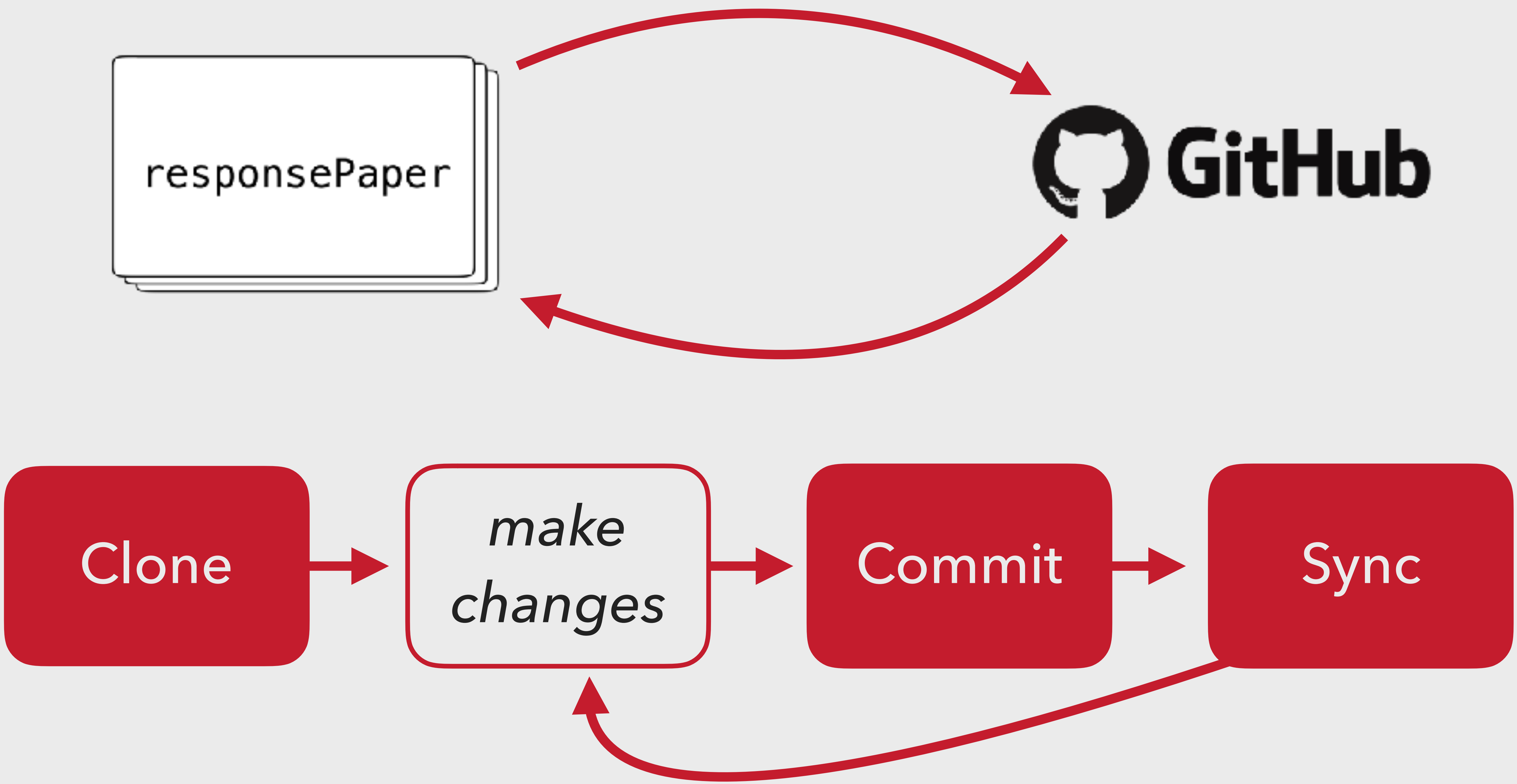
GIT WORKFLOW

Local repos can "sync" with a "remote" repo, making backup and sharing easy

GIT WORKFLOW

Copying data for the first time from GitHub is called making a "clone"

GIT WORKFLOW



USE ONE (AND ONLY ONE) COURSE DIRECTORY

```
SOC5050/
  FinalProject/
   Memo/
   DataAnalysis/
   PaperDraft/
   PaperFinal/
   PosterDraft/
   PosterFinal/
  Labs/
   Lab01/
   Lab02/
   …
   Lab16/
  Preps/
  ProblemSets/
```

USE ONE (AND ONLY ONE) COURSE DIRECTORY

```
SOC5050/
   FinalProject/
     Memo/
     DataAnalysis/
     PaperDraft/
     PaperFinal/
     PosterDraft/
     PosterFinal/
   Labs/
     Lab01/
     Lab02/
     …
     Lab16/
   Preps/
   ProblemSets/
```

Specific lab, prep, and problem set directories should have dedicated R Projects associated with them to increase organization!

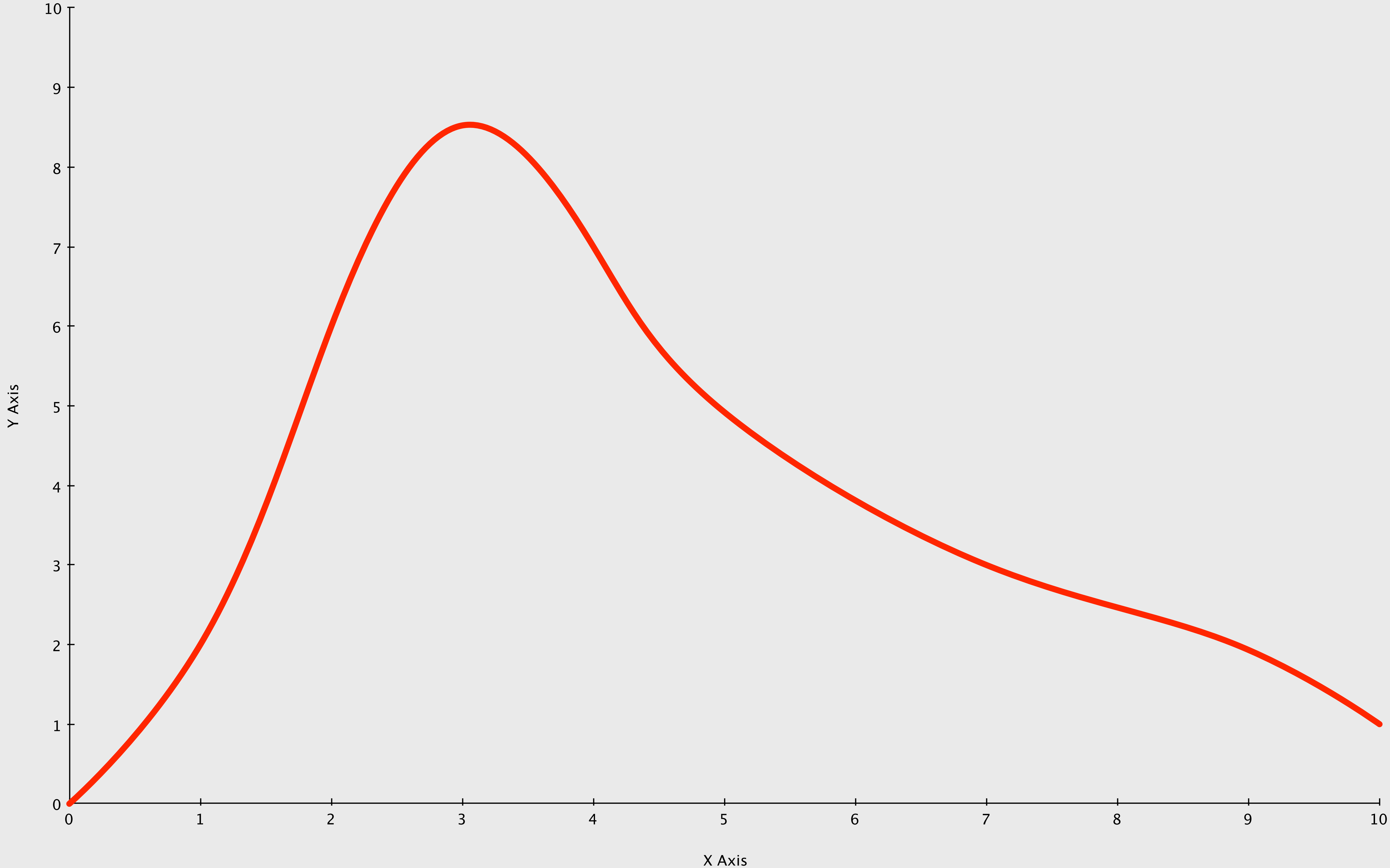
USE ONE (AND ONLY ONE) COURSE DIRECTORY

```
SOC5050/
   FinalProject/
    Memo/
    DataAnalysis/
    PaperDraft/
    PaperFinal/
    PosterDraft/
    PosterFinal/
   Labs/
    Lab01/
    Lab02/
    …
    Lab16/
   Preps/
   ProblemSets/
```

You should keep this as a GitHub repository or store it some other way (Dropbox). You can get private GitHub repos for free!

3 DESCRIBING DISTRIBUTIONS

WHAT IS A DISTRIBUTION?

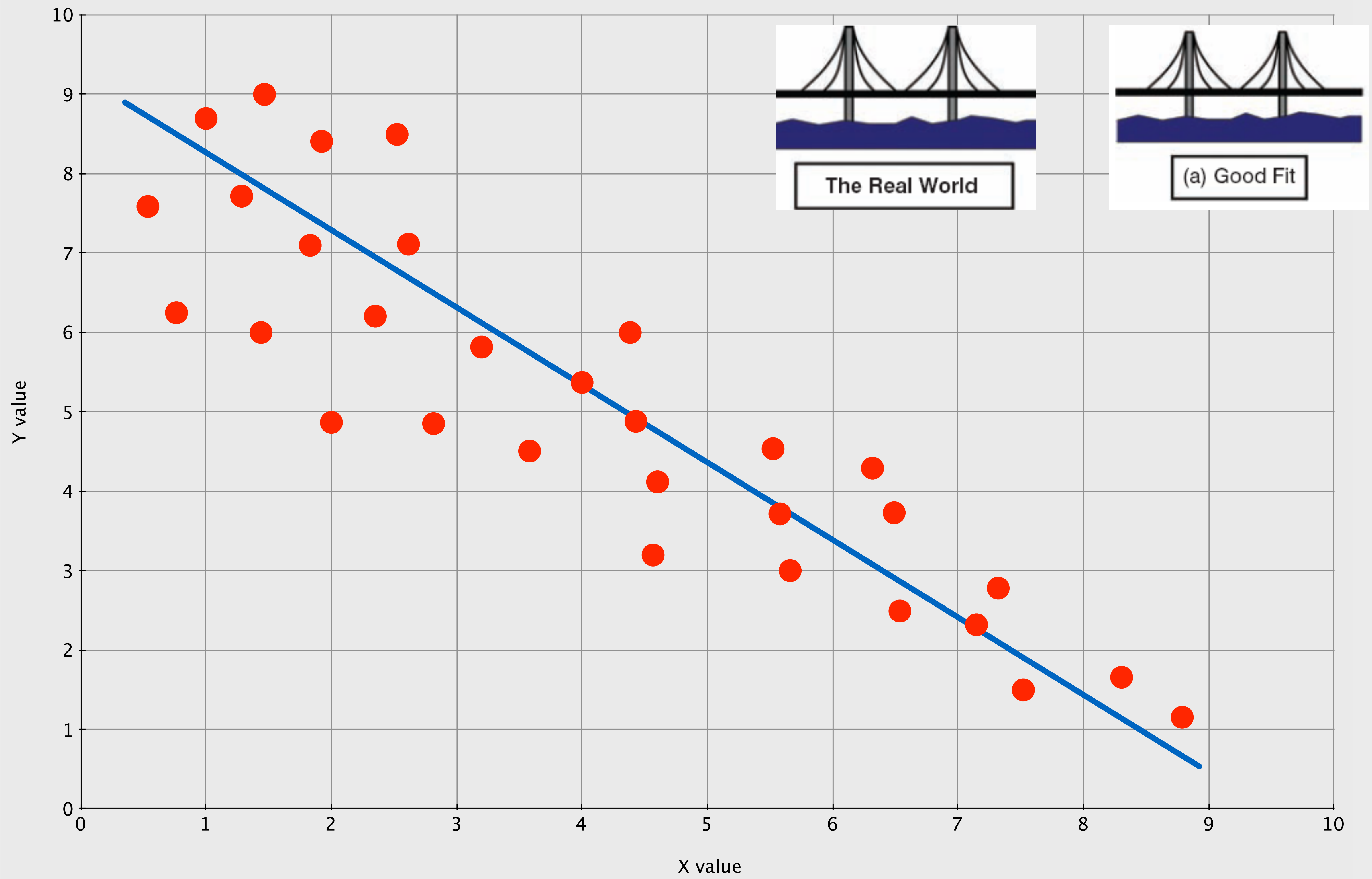


DESCRIPTIVE STATISTICS

CENTRAL TENDENCY
HOW VALUES CONGREGATE AROUND THE CENTER OF THE DISTRIBUTION

DISPERSION
HOW VALUES ARE "SPREAD"

DESCRIPTIVE STATISTICS



DESCRIPTIVE STATISTICS

CENTRAL TENDENCY
HOW VALUES CONGREGATE
AROUND THE CENTER
OF THE DISTRIBUTION

MODE

MEDIAN

MEAN

MODE

```
> library(tidyverse)

> autoData <- mpg

> table(mpg$cyl)

 4  5  6  8
81  4 79 70
```

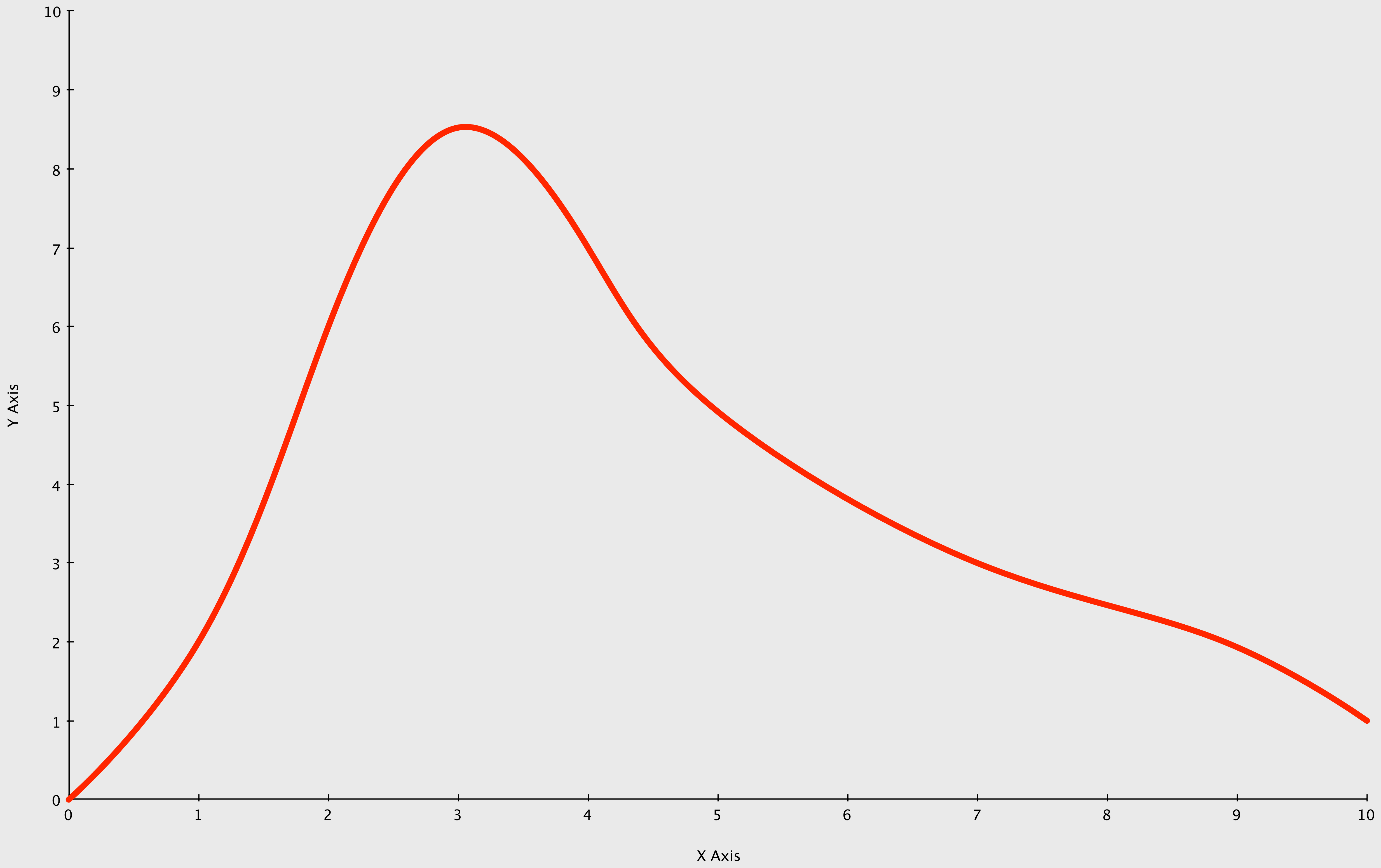
DEFINITION

THE MOST COMMON VALUE

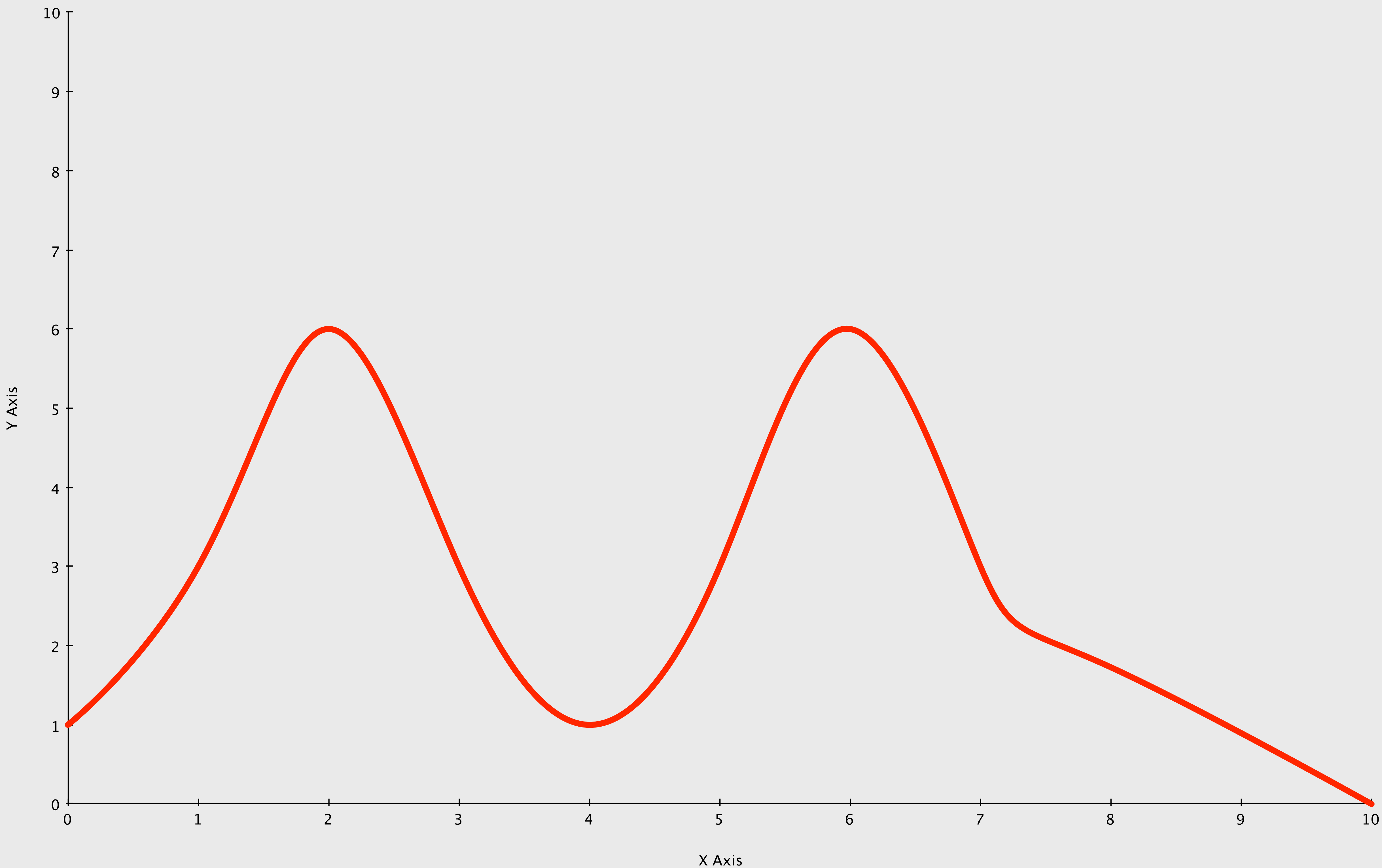
MODE

```
> library(tidyverse)

> autoData <- mpg

> table(mpg$cyl)

 4  5  6  8
81  4 79 70

> prop.table(table(mpg$cyl))

         4          5          6          8
0.34615385 0.01709402 0.33760684 0.29914530
```

MODE



MODE



MEDIAN

DEFINITION

THE MIDDLE-MOST VALUE WHEN VALUES ARE ORDERED FROM LOWEST TO HIGHEST

MEDIAN (ODD)

1, 3, 4, 16, 18, $\boxed{19,}$ 22, 36, 52, 64, 81

$$m = \left(\frac{n+1}{2}\right)^{th}$$

$$m = \left(\frac{11+1}{2}\right)^{th} = \left(\frac{12}{2}\right)^{th} = 6^{th}$$

MEDIAN (EVEN)

1, 3, 4, 16, 18, 19, 22, 36, 52, 64

Let m_a = the middlemost position:

$$m_a = \left(\frac{n+1}{2}\right)^{th}$$

$$m_a = \left(\frac{10+1}{2}\right)^{th} = \left(\frac{11}{2}\right)^{th} = (5.5)^{th}$$

MEDIAN (EVEN)

1, 3, 4, 16, 18, 19, 22, 36, 52, 64

Let x_a = the next lower value before m_a:

$$x_a = 18$$

Let x_b = the next higher value after m_a:

$$x_b = 19$$

MEDIAN (EVEN)

1, 3, 4, 16, 18, 19, 22, 36, 52, 64

Let m_b = the median:

$$m_b = \left(\frac{x_a + x_b}{2}\right)$$

$$m_b = \left(\frac{18 + 19}{2}\right) = \left(\frac{37}{2}\right) = 18.5$$

MEDIAN

```
> library(tidyverse)

> autoData <- mpg

> median(mpg$cyl)

[1] 6
```

SIGMA NOTATION

$$\sum_{i=m}^{n} x_i = x_m + x_{m+1} + x_{m+2} + \cdots + x_{n-1} + x_n$$

upper bound

sigma (sum)

$$\sum_{i=m}^{n} x_i$$

operation

index (lower bound)

SIGMA NOTATION

$$\sum_{i=1}^{100} 2x = 2(1) + 2(2) + \cdots + 2(99) + 2(100)$$

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n$$

MEAN

FIRST MOMENT

$$\overline{x} = \frac{\sum_{i=1}^{n} x}{n}$$

Refers to the sample mean; Greek letter mu (μ) used for population

DEFINITION

A MEASURE OF THE "MIDDLE" OF THE DISTRIBUTION

MEAN

1, 3, 4, 16, 18, 19, 22, 36, 52, 64, 81

$$\overline{x} = \frac{\sum_{i=1}^{n} x}{n}$$

$$\overline{x} = \frac{1 + 3 + 4 + 16 + 18 + 19 + 22 + 36 + 52 + 64 + 81}{11}$$

$$\overline{x} = \frac{316}{11} = 32.82$$

ISSUES WITH THE MEAN

11, 31, 36, 41, 42, 52, 65, 72, 82

$$\overline{x} = 48$$

10, 11, 11, 36, 42, 78, 78, 82, 84

$$\overline{x} = 48$$

ISSUES WITH THE MEAN

2, 6, 8, 12, 24, 36, 38, 40

$$\overline{x} = 20.75$$

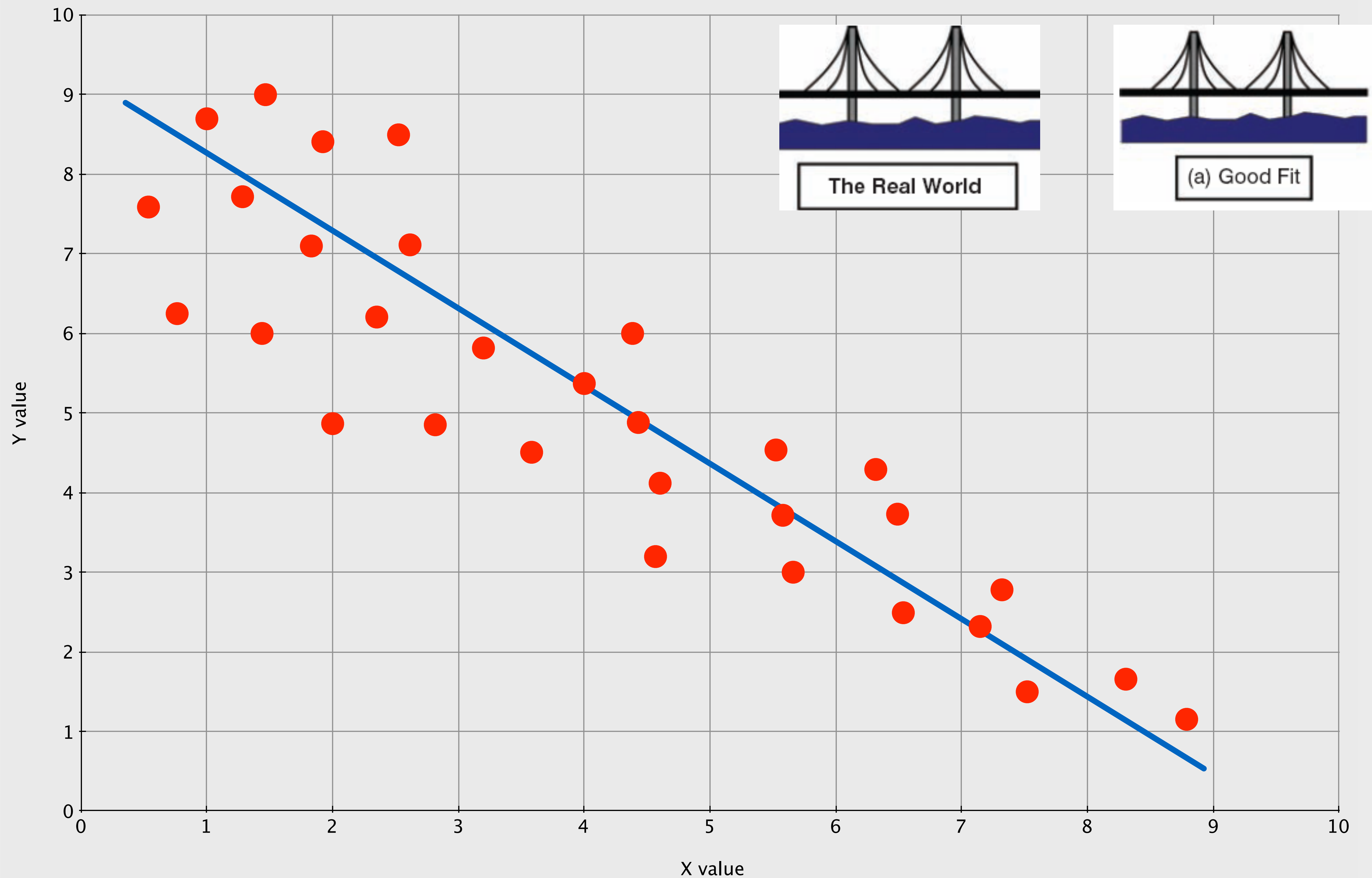
2, 6, 8, 12, 24, 36, 38, 44

$$\overline{x} = 22.25$$

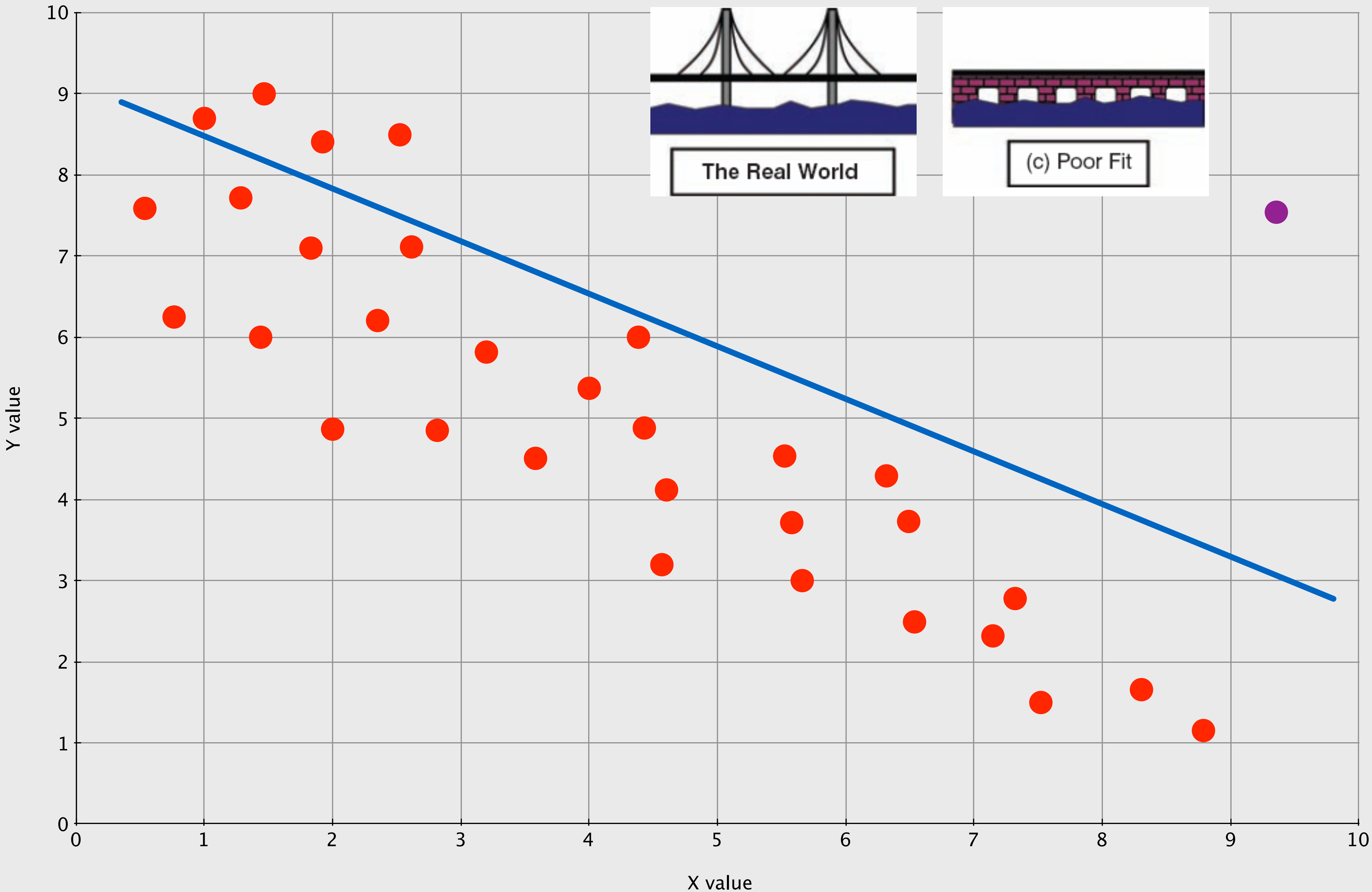
2, 6, 8, 12, 24, 36, 38, 2000

$$\overline{x} = 265.75$$

ISSUES WITH THE MEAN



ISSUES WITH THE MEAN



MEAN

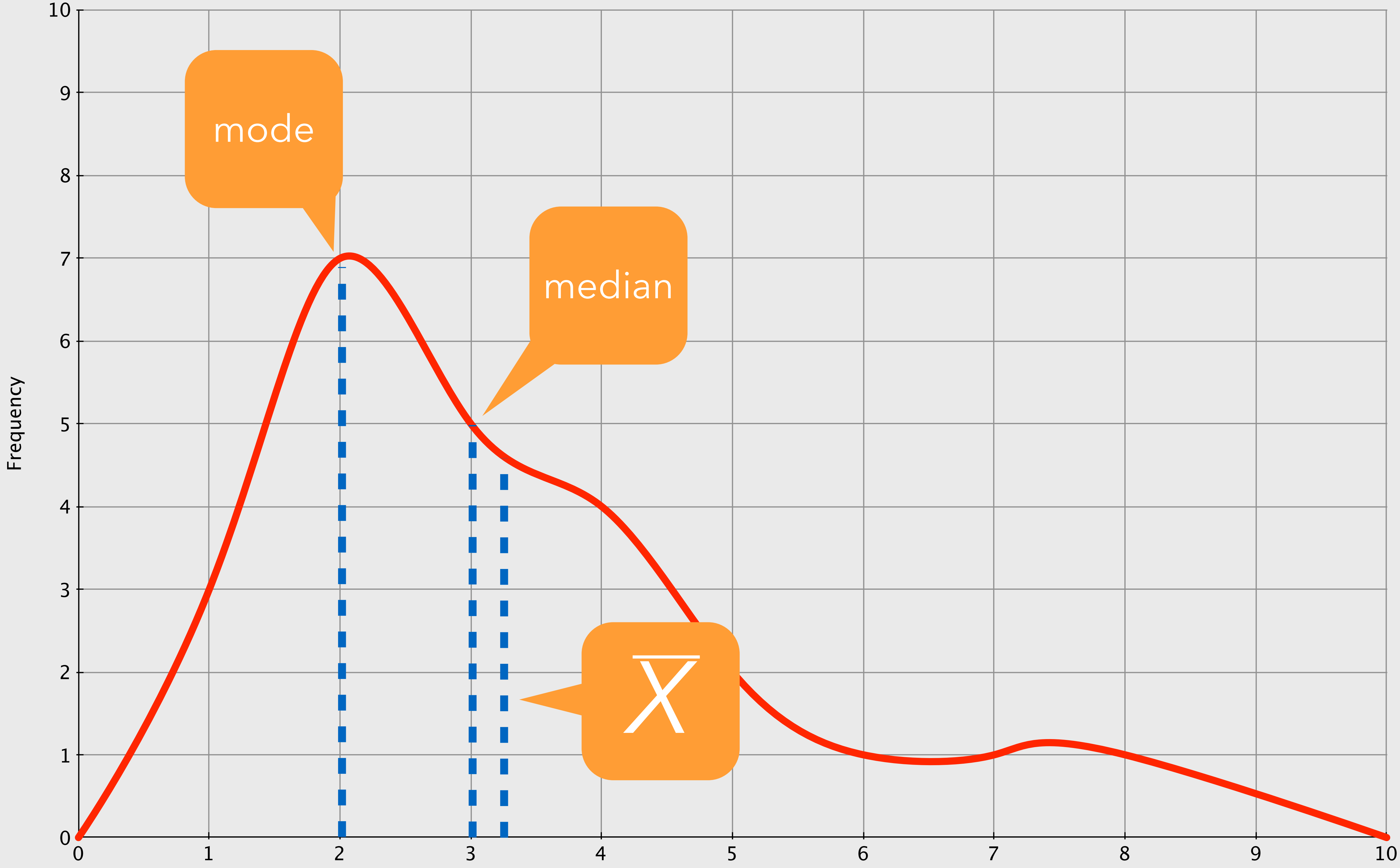
```
> library(tidyverse)

> autoData <- mpg

> mean(autoData$hwy)

[1] 23.44017
```

MEAN

```
> library(tidyverse)

> autoData <- mpg

> autoData <- mutate(autoData,
                     subaru = ifelse(manufacturer == “subaru",
                     TRUE, FALSE))

> prop.table(table(autoData$subaru))

     FALSE       TRUE
0.94017094 0.05982906

> mean(autoData$subaru)

[1] 0.05982906
```

SPECIAL PROPERTY
THE MEAN OF A BINARY VARIABLE IS THE PROPORTION OF VALUES THAT EQUAL '1'

DESCRIPTIVE STATISTICS



DESCRIPTIVE STATISTICS

```
> library(tidyverse)

> autoData <- mpg

> summary(autoData$hwy)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   18.00   24.00   23.44   27.00   44.00
```

DESCRIPTIVE STATISTICS

CENTRAL TENDENCY
HOW VALUES CONGREGATE AROUND THE CENTER OF THE DISTRIBUTION

MEASURES OF VARIABILITY

- VARIANCE
- STANDARD DEVIATION

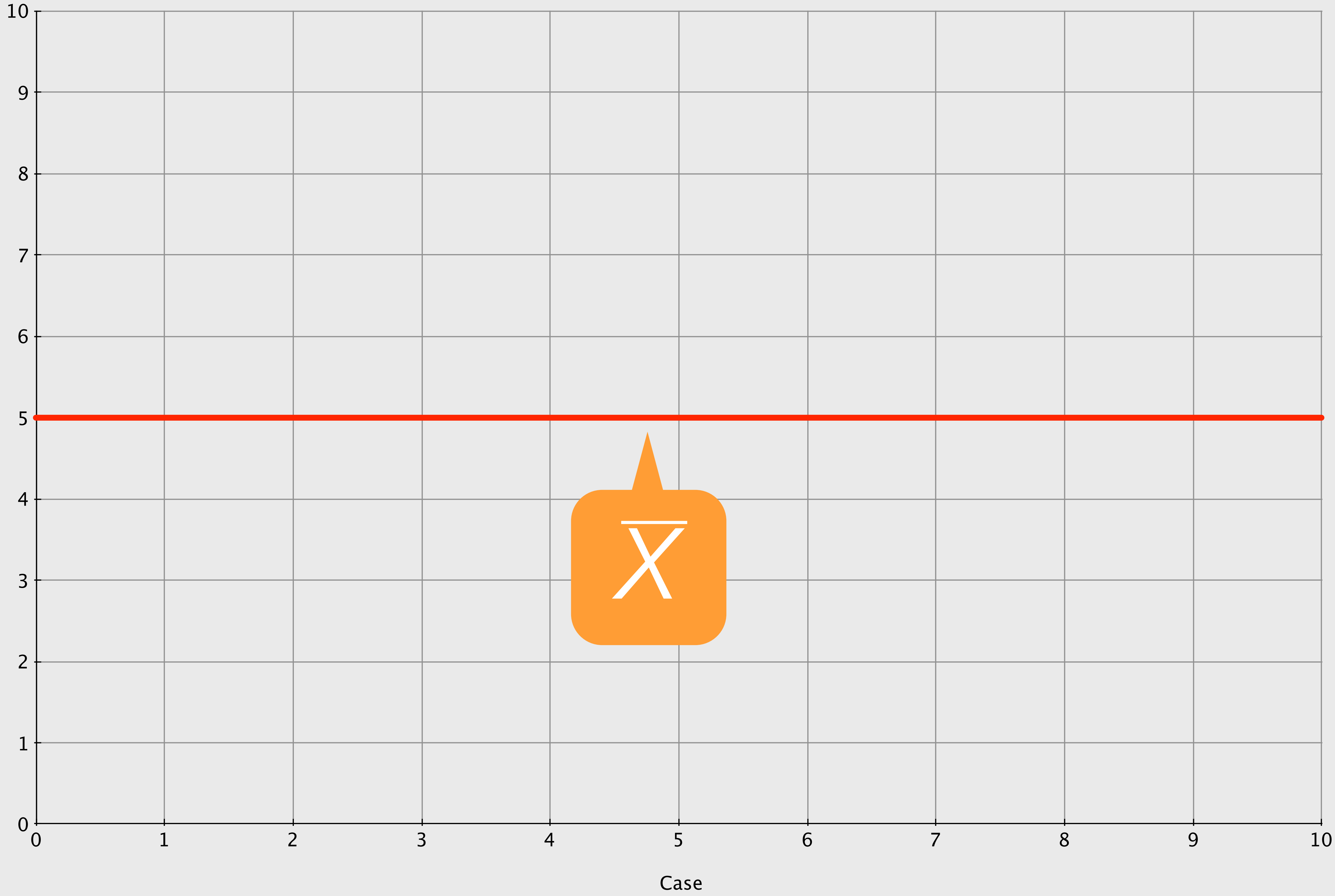
DEVIANCE

$$D = (x - \overline{x})$$

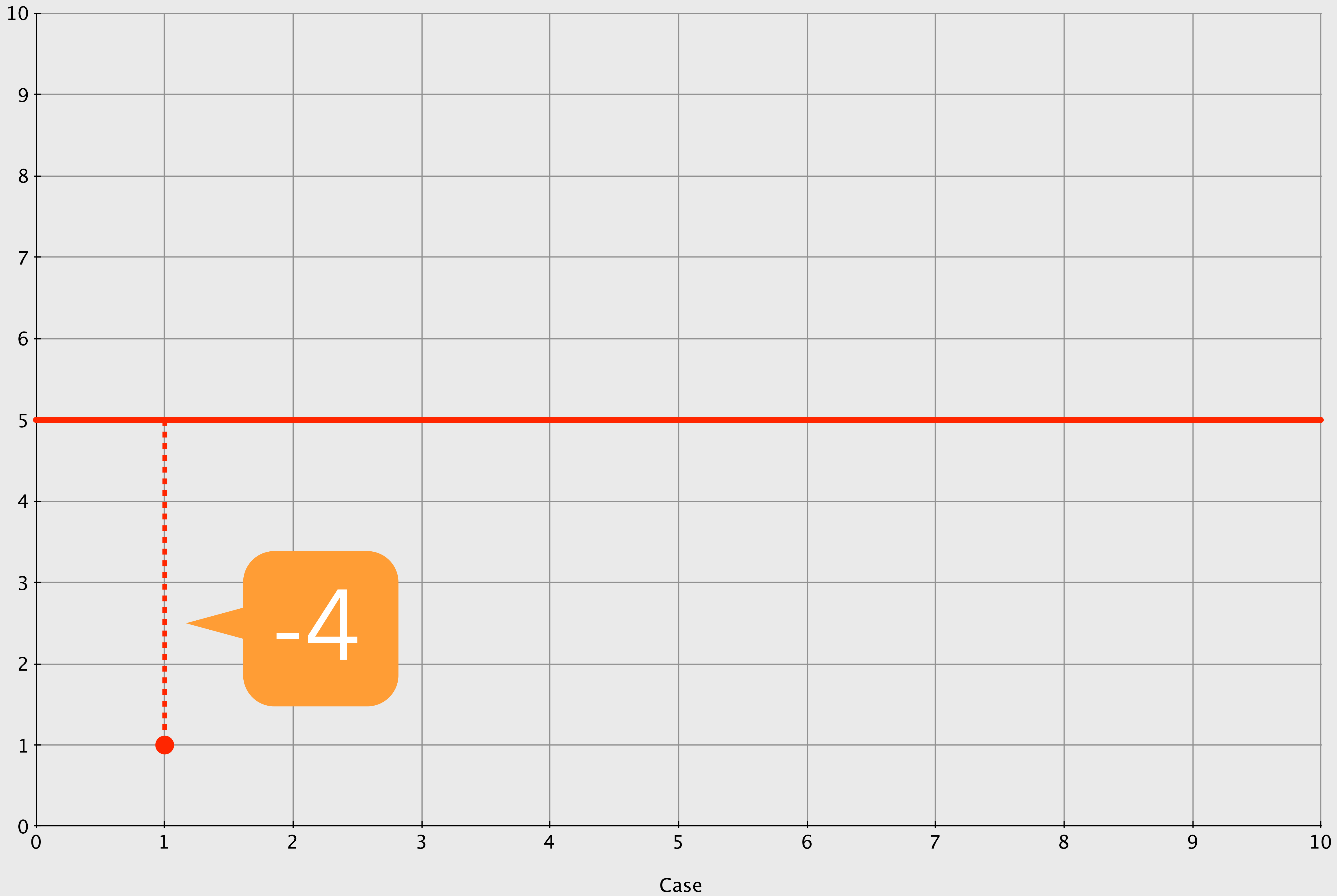
DEFINITION

DIFFERENCE BETWEEN AN OBSERVED VALUE AND THE MEAN VALUE OF THE VARIABLE

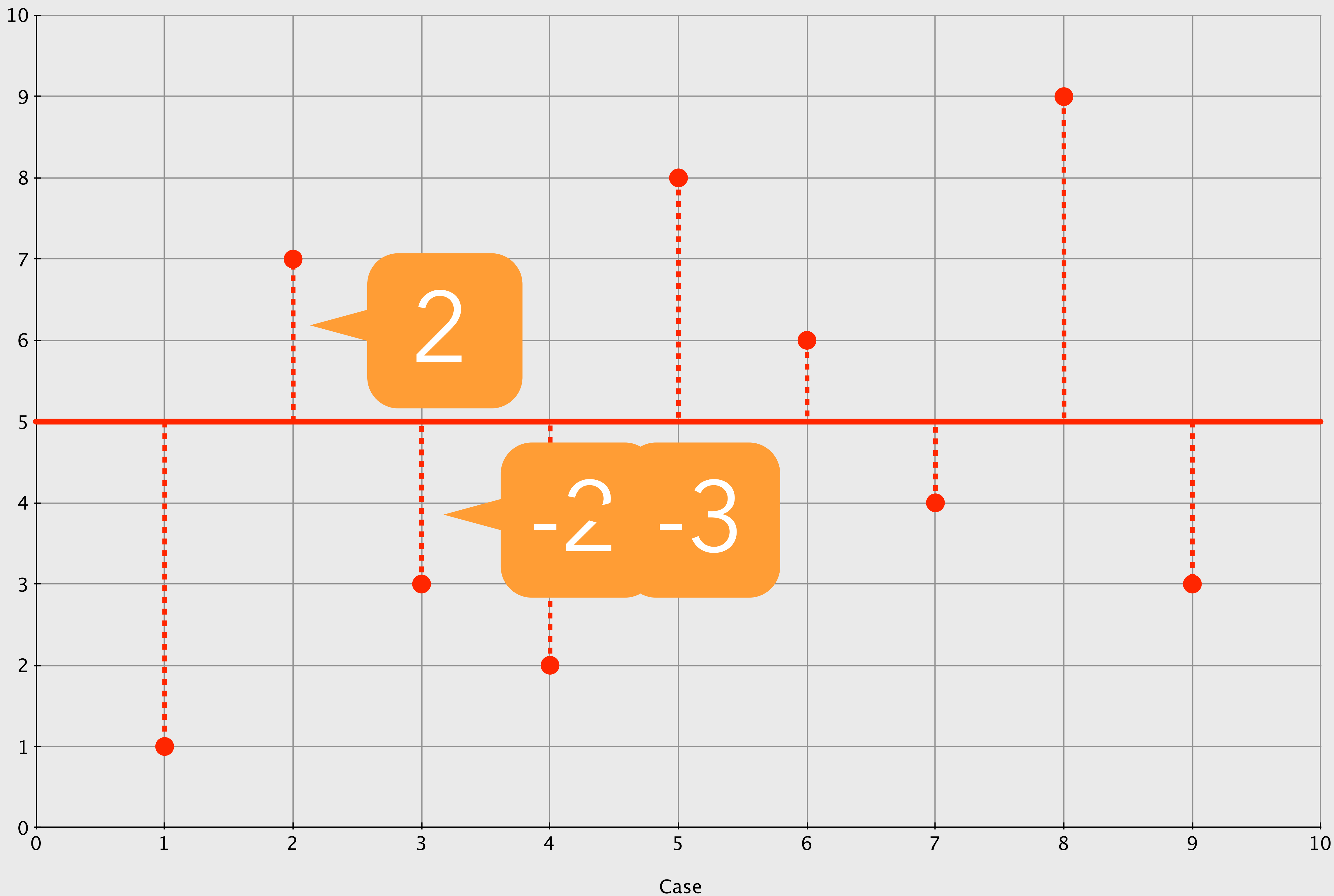
DEVIANCE



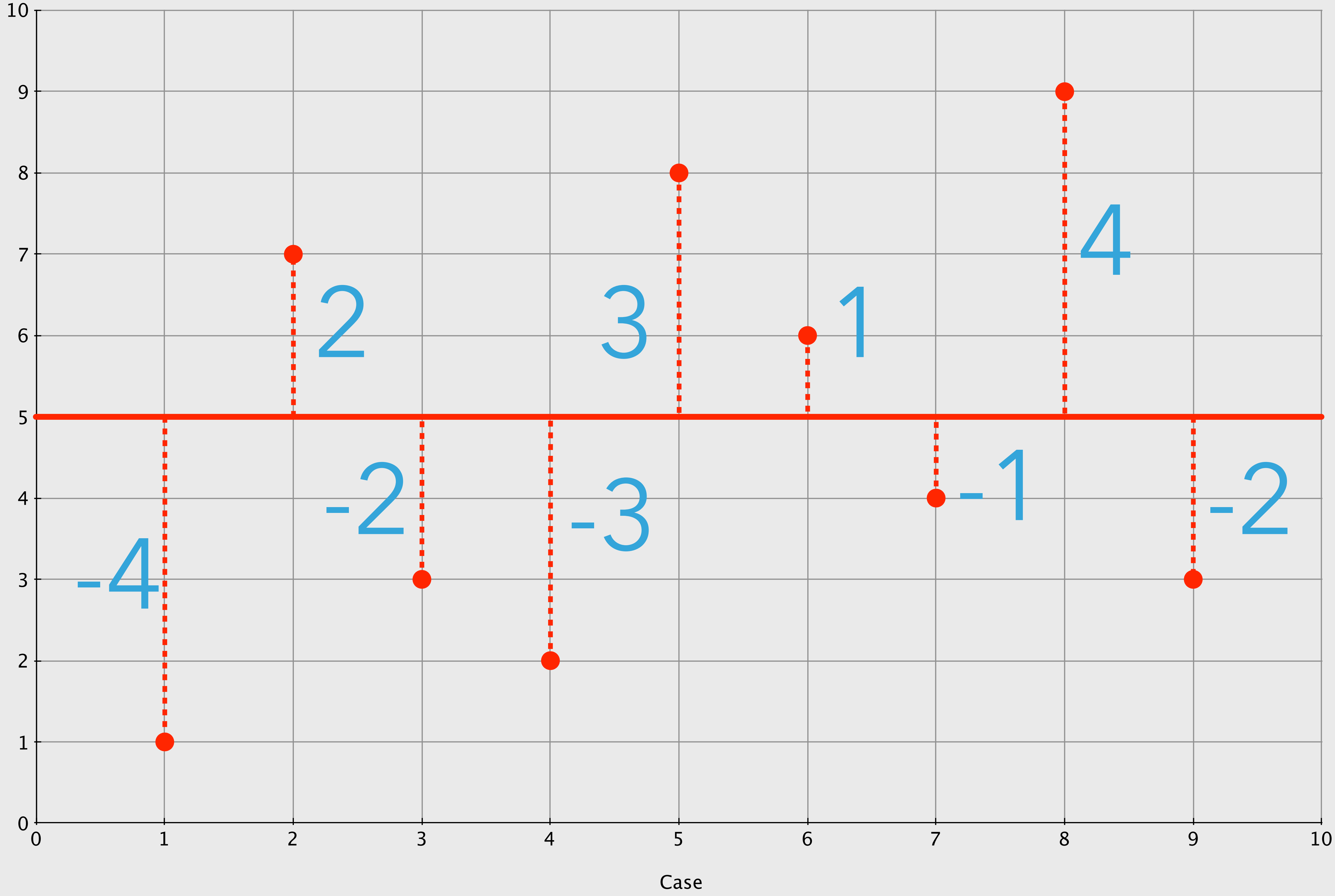
DEVIANCE



DEVIANCE



DEVIANCE



TOTAL ERROR

$$TE = \sum_{i=1}^{n} (x - \overline{x})$$

DEFINITION

SUM OF ALL DEVIANCES

TOTAL ERROR

$$TE = \sum_{i=1}^{n} (x - \overline{x}) = 0$$

DEFINITION

SUM OF ALL DEVIANCES;
ALWAYS EQUAL TO ZERO
IF CALCULATED CORRECTLY

SUM OF SQUARED ERROR

$$SS = \sum_{i=1}^{n} (x - \overline{x})^2$$

DEFINITION

SUM OF ALL DEVIANCES, SQUARED

VARIANCE

$$s^2 = \frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n - 1}$$

Refers to the sample variance; Greek letter sigma (σ^2) used for population

DEFINITION

SUM OF ALL DEVIANCES, SQUARED AND DIVIDED BY ONE DEGREE OF FREEDOM; EXPECTATION OF HOW DISTRIBUTION DEVIATES FROM THE MEAN

VARIANCE

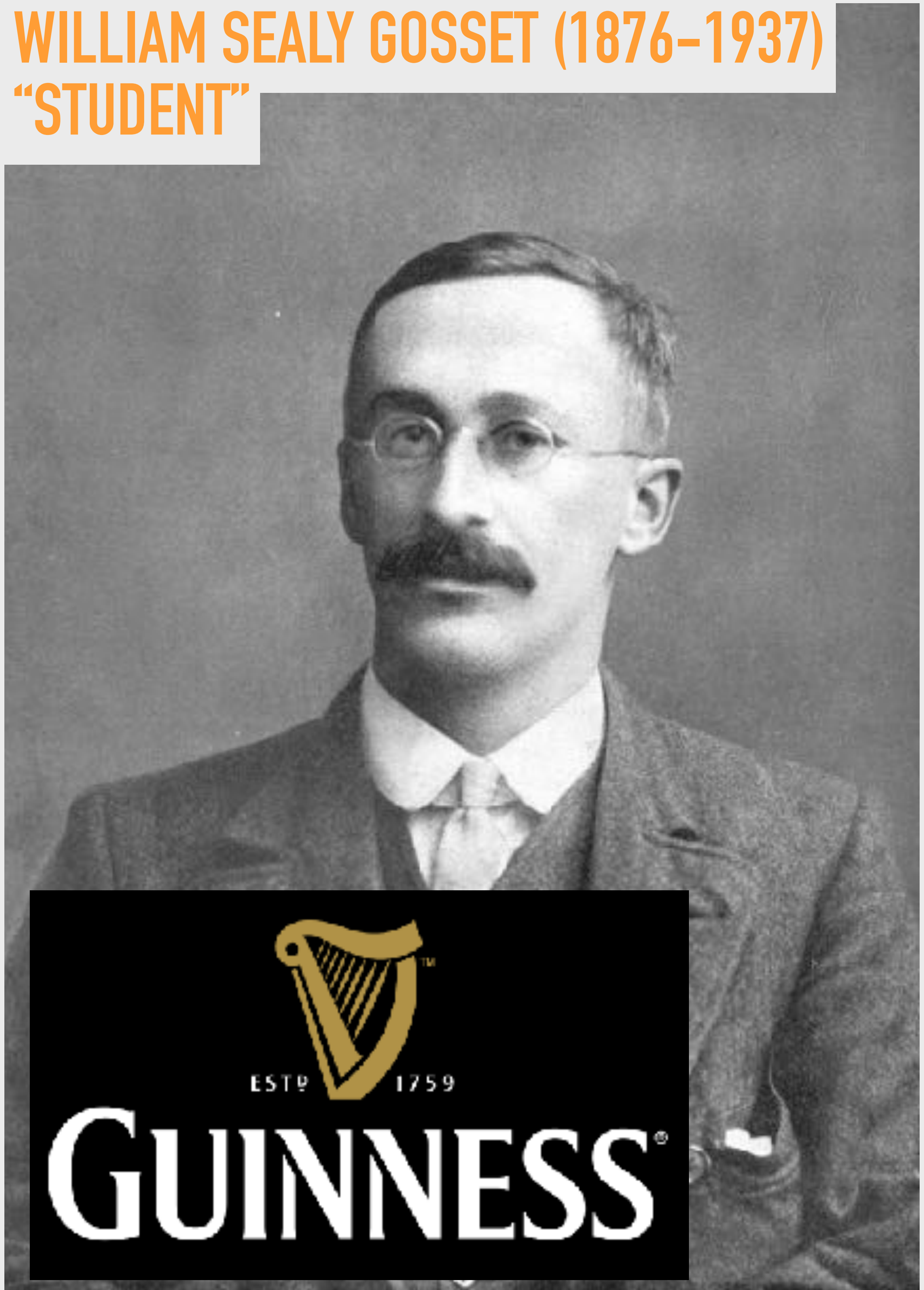
SECOND MOMENT

$$s^2 = \frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n - 1}$$

PROPERTIES

WITH LARGE SAMPLES, THE SAMPLE VARIANCE (s^2) APPROACHES THE POPULATION VARIANCE (σ^2)

WILLIAM SEALY GOSSET (1876-1937)
"STUDENT"



DEGREES OF FREEDOM

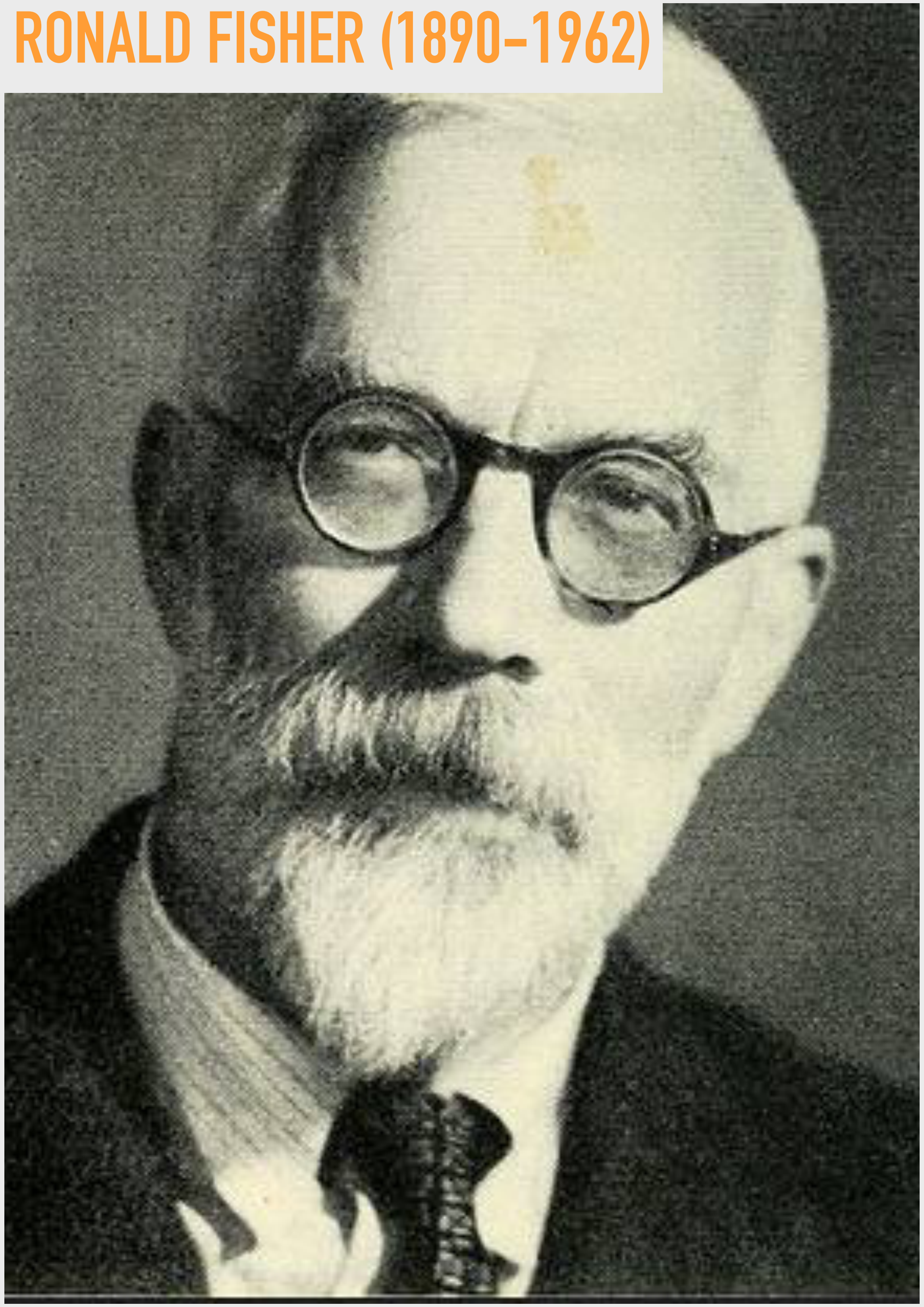
WILLIAM SEALY GOSSET (1876-1937)
"STUDENT"

DEGREES OF FREEDOM

DEFINITION

THE NUMBER OF OBSERVATIONS THAT ARE FREE TO VARY WHEN ESTIMATING A PARTICULAR STATISTICAL PARAMETER

RONALD FISHER (1890-1962)



DEGREES OF FREEDOM

DEFINITION

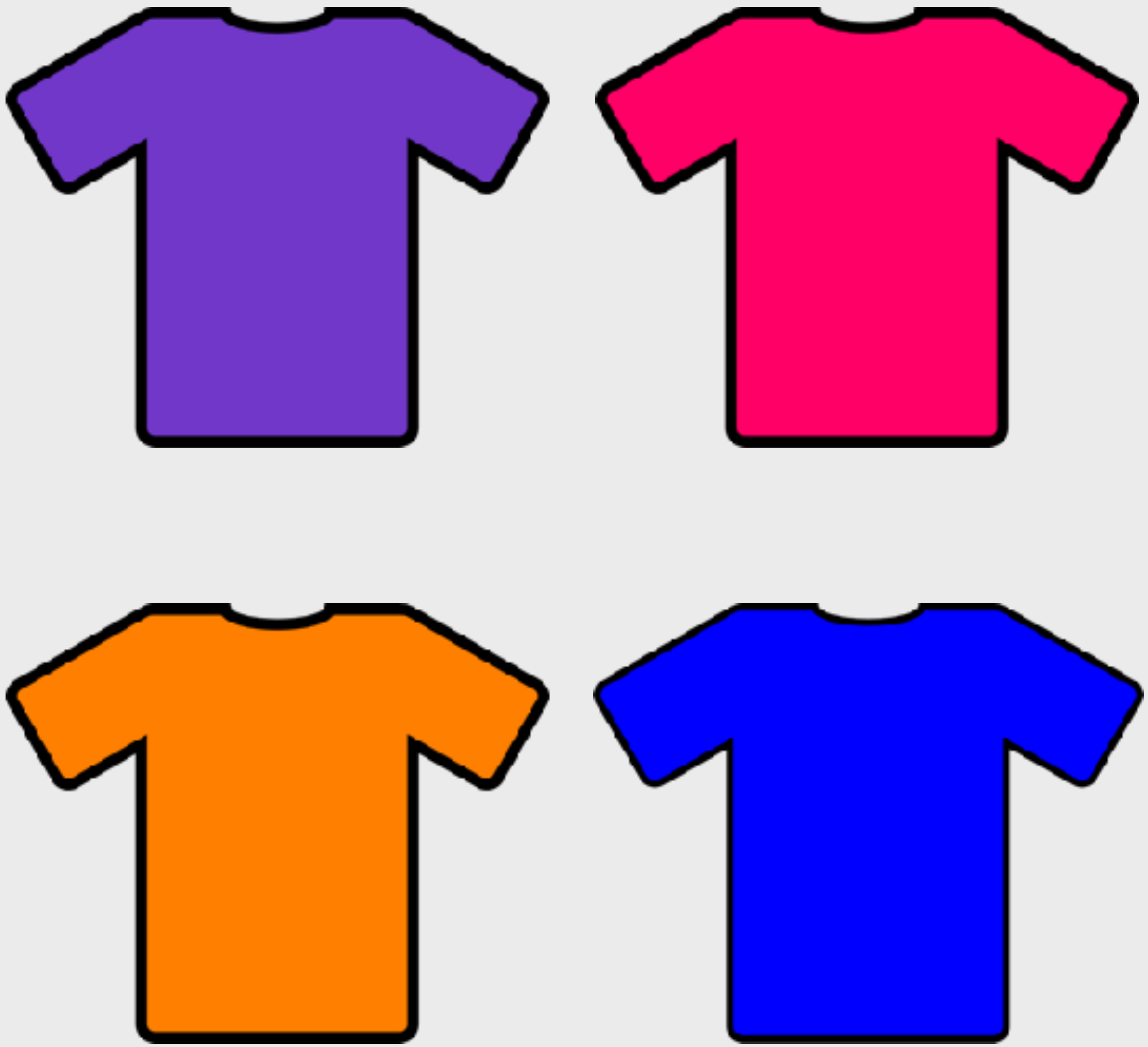
THE NUMBER OF OBSERVATIONS THAT ARE FREE TO VARY WHEN ESTIMATING A PARTICULAR STATISTICAL PARAMETER

DEGREES OF FREEDOM (EXAMPLE 1)

You are on a four day vacation, and have one shirt for each of the four days.

By the time you reach the fourth and final day of your trip, you have no choices left - you must wear the orange shirt.

You had *n*-1 (4-1=3) days in which you had freedom to over what you wore.

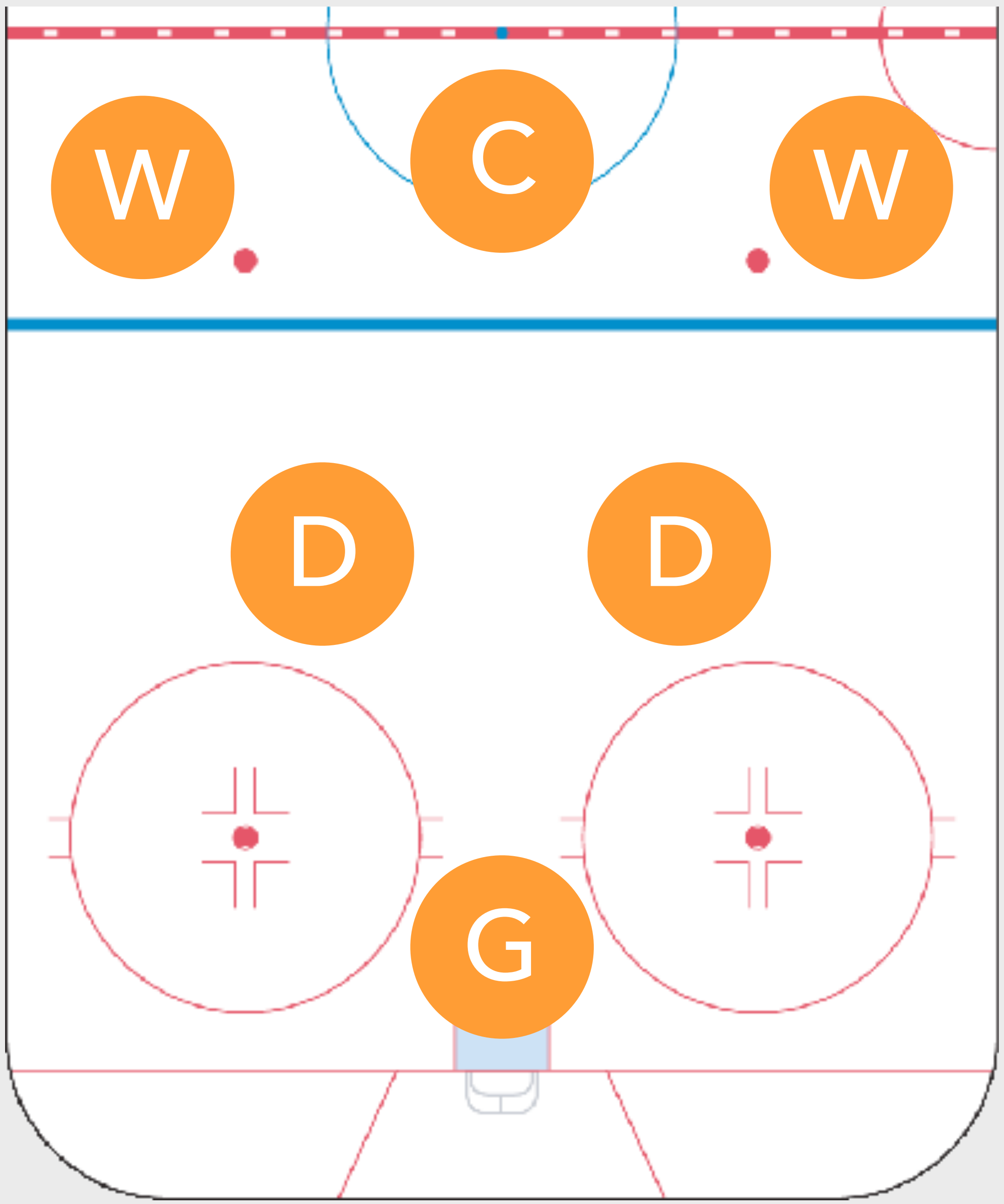


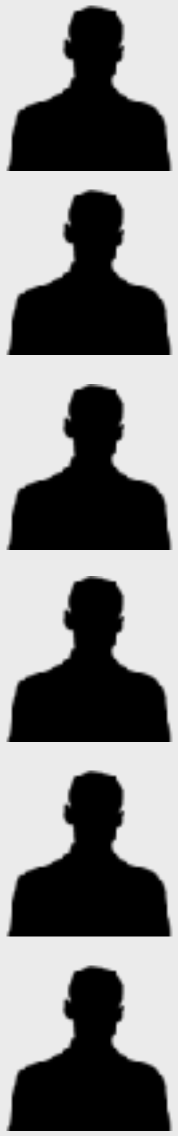
DEGREES OF FREEDOM (EXAMPLE 2)

A hockey team has a total of six men (intentionally!) on the ice at any one time. You have six players on your roster.

By the time you reach the goalie, your sixth and final player must play that position.

You had *n*-1 (6-1=5) positions where you had freedom to decide who played where.





DEGREES OF FREEDOM (EXAMPLE 3)

In the mathematical equation on the right, you can select whatever values you want for *x* and *y* so long as they total to 7.

$$x + y = 7$$

However, if I tell you that *x* is equal to 4, your choice becomes limited.

$$4 + y = 7$$

DEGREES OF FREEDOM (EXAMPLE 4)

In the mathematical equation on the right, you can select whatever values you want for *x*, *y*, and *z* so long as they equal 0.

$$x + y + z = 0$$

However, if I tell you that *x* is equal to 4, your choice becomes constrained.

$$4 + y + z = 0$$

If I tell you that *y* is equal to -2, your choice becomes further constrained.

$$4 + -2 + z = 0$$

DEGREES OF FREEDOM

SAMPLE VARIANCE

$$s^2 = \frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n - 1}$$

POPULATION VARIANCE

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

Degrees of freedom (v) is always calculated by subtracting the number of constraints (relationships) from the total number of observations. When you calculate deviance, you impose a limiting relationship. Thus $n - 1$ is included in calculations.

FRIEDRICH BESSEL (1784–1846)

DEGREES OF FREEDOM

BESSEL'S CORRECTION

IF WE DO NOT USE DEGREES OF FREEDOM, OUR ESTIMATE OF THE POPULATION VARIANCE IS BIASED DOWNWARDS IN THE TYPICAL SAMPLE.

STANDARD DEVIATION

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x - \overline{x})^2}{n - 1}}$$

DEFINITION

SQUARE ROOT OF VARIANCE; PLACES DEVIATION FROM THE MEAN IN EASY-TO-USE (I.E. STANDARDIZED) UNITS

VARIANCE & STANDARD DEVIATION

```
> library(tidyverse)

> autoData <- mpg

> var(autoData$hwy)

[1] 35.45778


> sd(autoData$hwy)

[1] 5.954643
```

DESCRIPTIVE STATISTICS

CENTRAL TENDENCY
HOW VALUES CONGREGATE AROUND THE CENTER OF THE DISTRIBUTION

DISPERSION
HOW VALUES ARE "SPREAD"

DESCRIPTIVE STATISTICS

RANGE

INTER-
QUARTILE
RANGE

DISPERSION
HOW VALUES ARE "SPREAD"

DESCRIPTIVE STATISTICS

RANGE
DISTANCE BETWEEN THE LARGEST SCORE AND THE SMALLEST

INTER-QUARTILE RANGE
DISTANCE BETWEEN THE 25TH AND 75TH PERCENTILES OF THE DATA

RANGE & IQR

1, 3, $\boxed{4,}$ 16, 18, $\boxed{19,}$ 22, 36, $\boxed{52,}$ 64, 81

$$range = 81 - 1 = 80$$

$$iqr = 52 - 4 = 48$$

RANGE & IQR

```
> library(tidyverse)

> autoData <- mpg

> range(autoData$hwy)

[1] 12 44


> iqr(autoData$hwy)

[1] 9
```

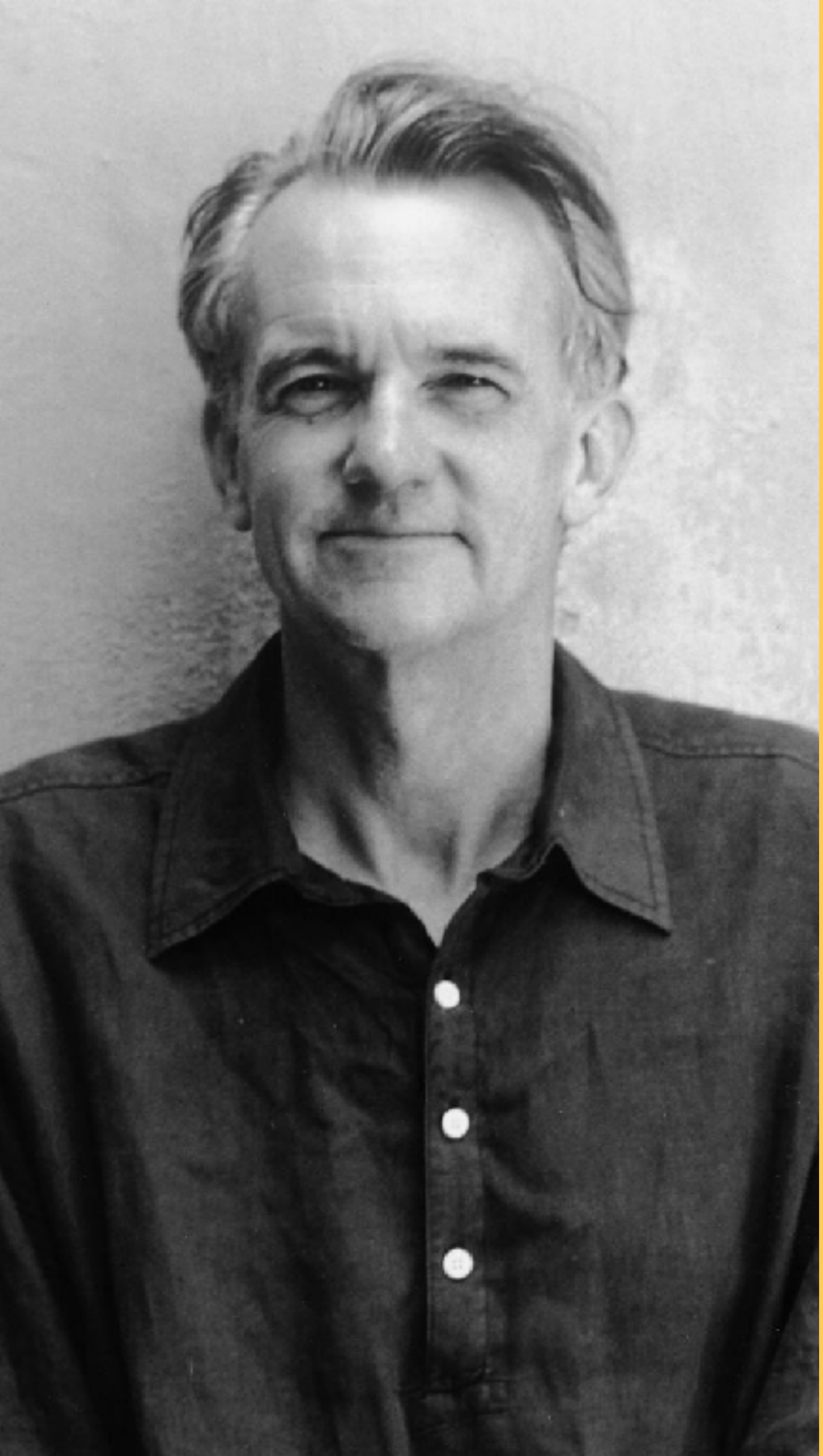
WHAT TO USE WHEN*

Level of Measurement	Mode	Median	Mean
Categorical	Yes	No	No
Ordinal	Yes	Yes	Yes**
Continuous	Yes	Yes	Yes

* General advice - not gospel!

** The mean of an ordinal variable is based on your coding scheme - use caution!

4 VISUALIZING DISTRIBUTIONS

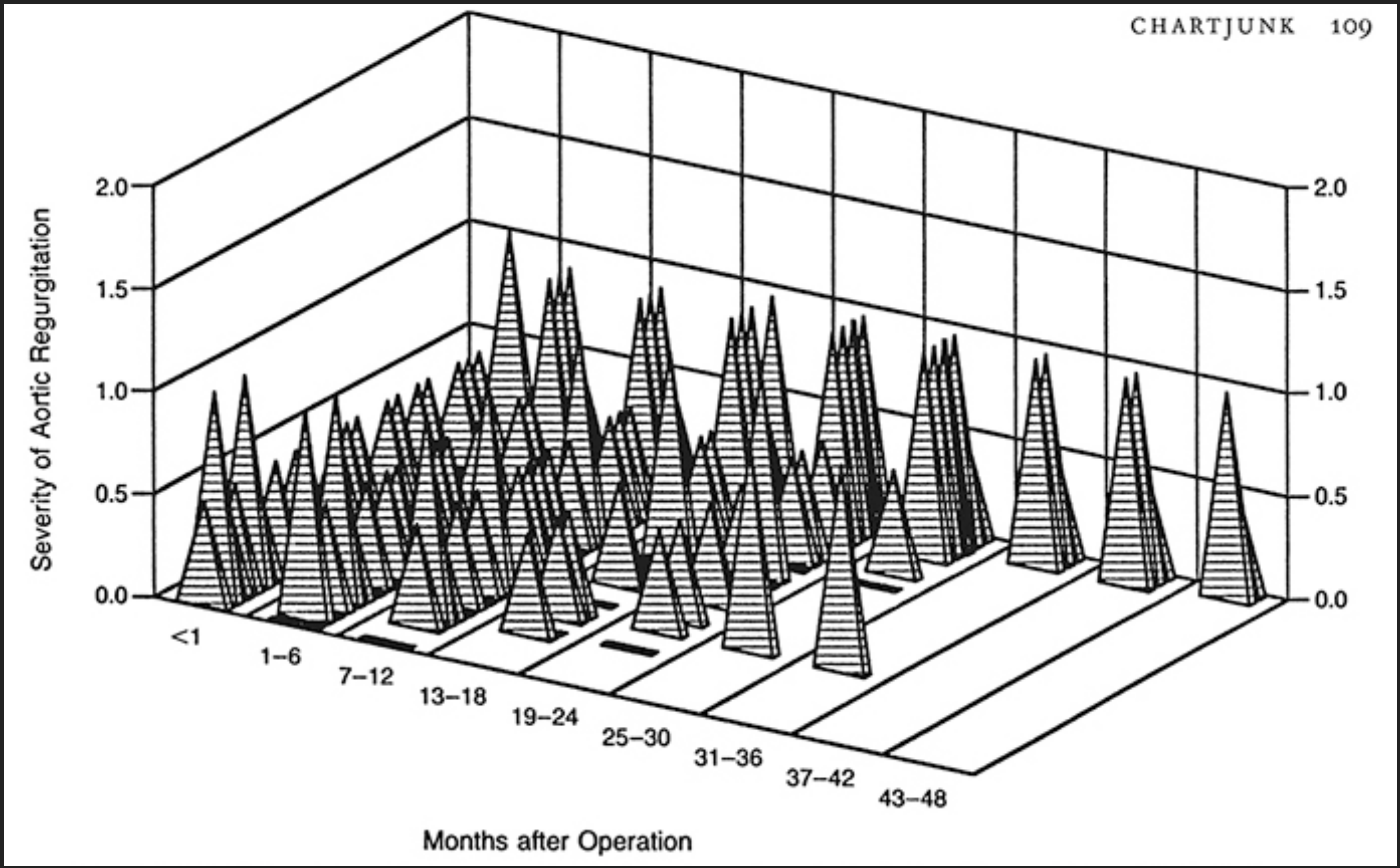


THE INTERIOR DECORATION OF GRAPHICS GENERATES A LOT OF INK THAT DOES NOT TELL THE VIEWER ANYTHING NEW. THE PURPOSE OF DECORATION VARIES — TO MAKE THE GRAPHIC APPEAR MORE SCIENTIFIC AND PRECISE, TO ENLIVEN THE DISPLAY, TO GIVE THE DESIGNER AN OPPORTUNITY TO EXERCISE ARTISTIC SKILLS. REGARDLESS OF ITS CAUSE, IT IS ALL NON-DATA-INK OR REDUNDANT DATA-INK, AND IT IS OFTEN CHARTJUNK.

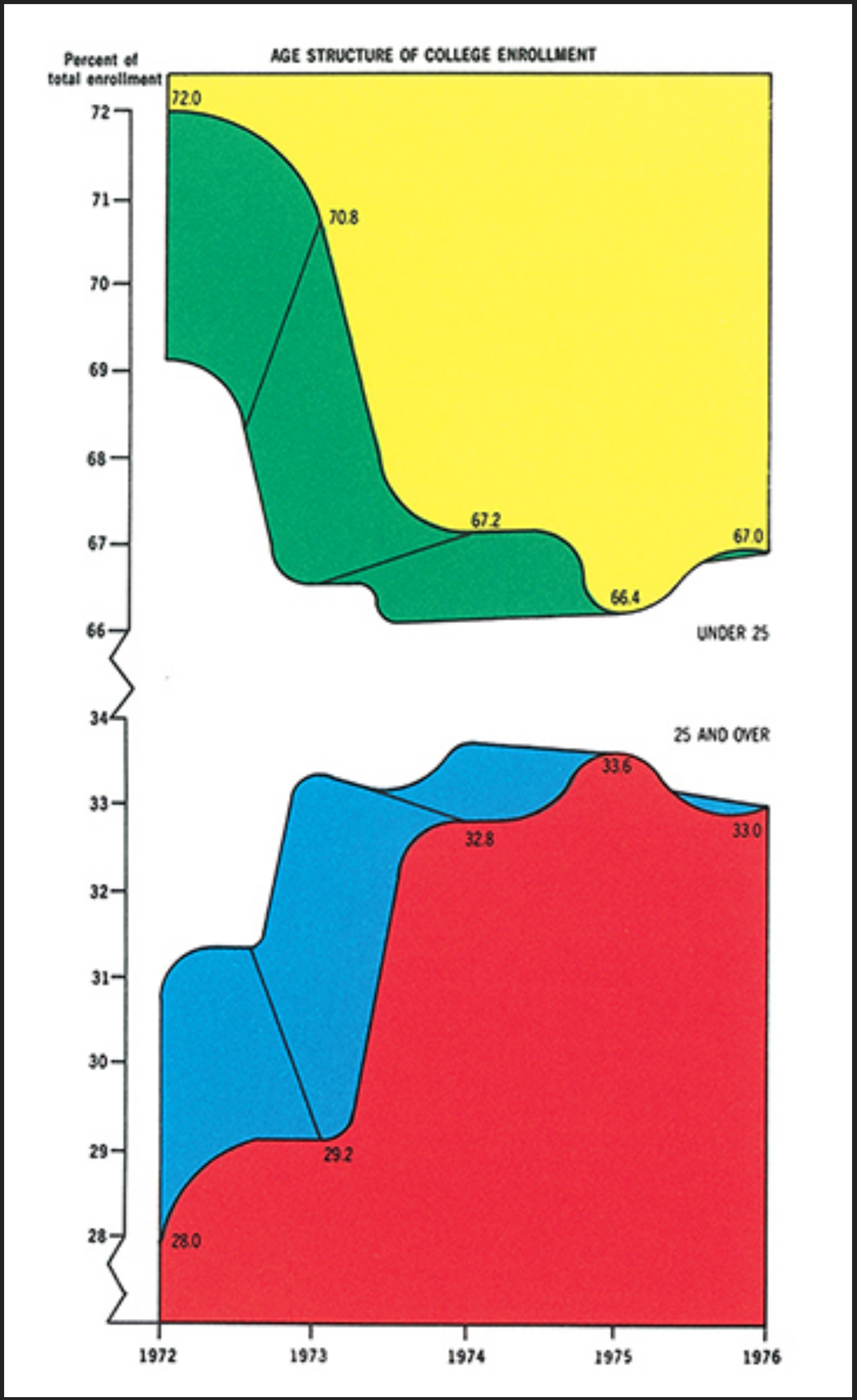
Edward Tufte, Ph.D.

Professor Emeritus, Yale University

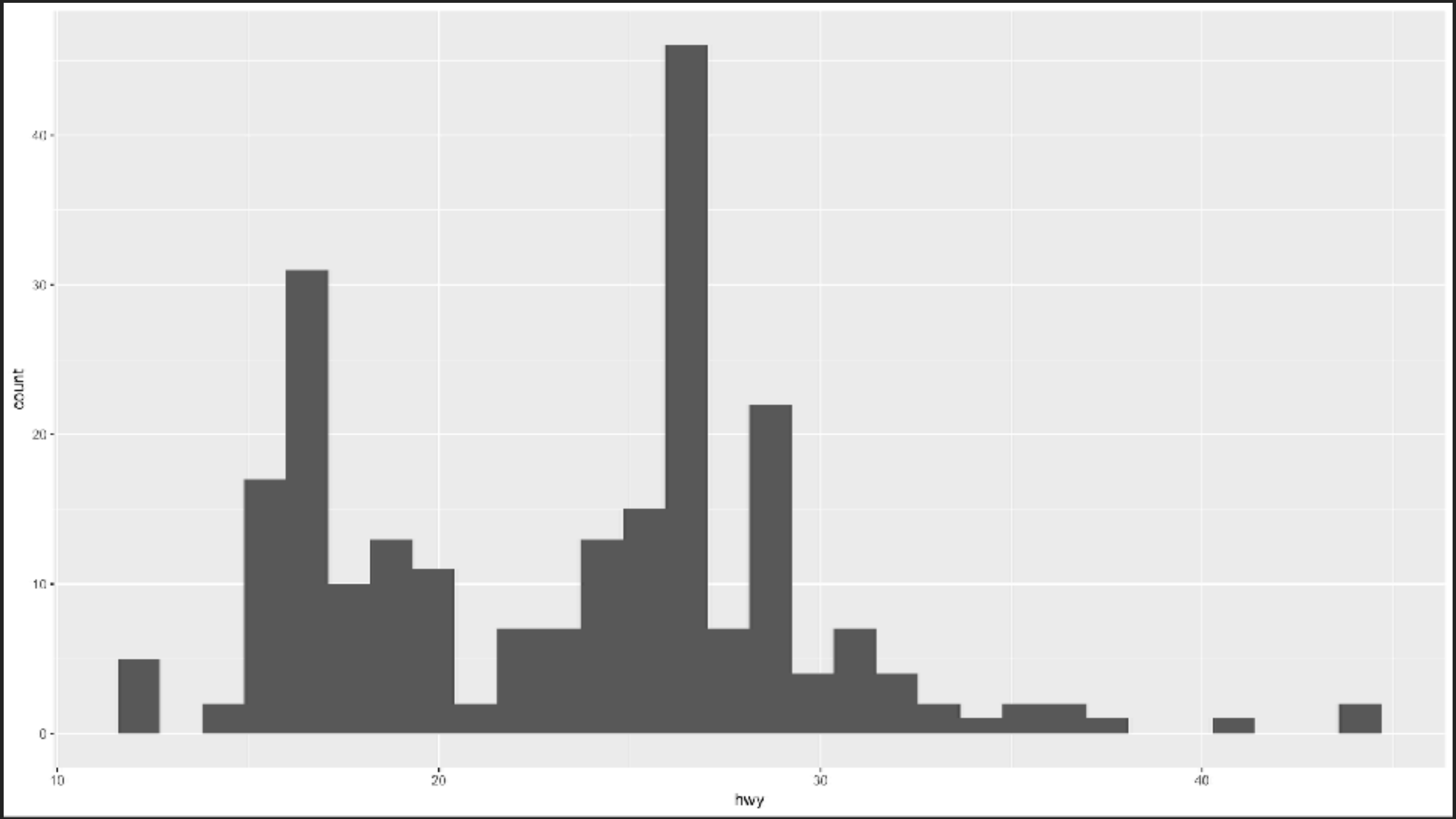
CHARTJUNK


CHARTJUNK 109
2.0
1.5
1.0
0.5
0.0
Severity of Aortic Regurgitation
<1
1–6
7–12
13–18
19–24
25–30
31–36
37–42
43–48
Months after Operation

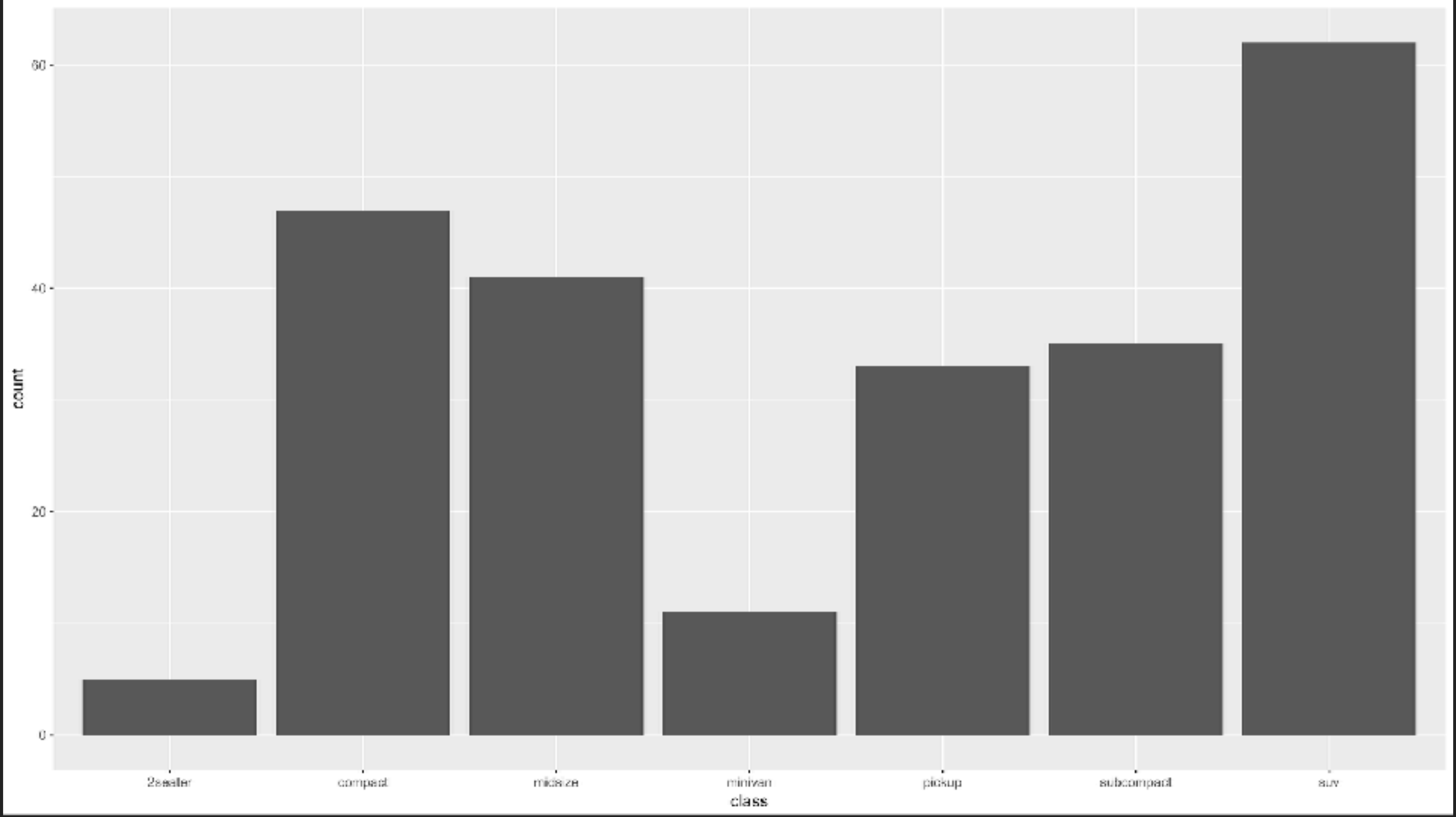
CHARTJUNK


AGE STRUCTURE OF COLLEGE ENROLLMENT
Percent of total enrollment
72
71
70
69
68
67
66
34
33
32
31
30
29
28
72.0
70.8
67.2
66.4
67.0
UNDER 25
25 AND OVER
33.6
32.8
33.0
29.2
28.0
1972
1973
1974
1975
1976

HISTOGRAMS



BAR PLOTS


count
0
20
40
60
2seater
compact
midsize
minivan
pickup
subcompact
suv
class

5 ANSCOMBE'S QUARTET

FRANK ANSCOMBE

- English mathematician who spent his career at Princeton and Yale
- Founded Yale's Department of Statistics in 1963
- Early proponent of statistical computing and the important of graphing distributions

ANSCOMBE'S QUARTET