Clinical statistics for non-statisticians: Day one

Steve Simon

One warning

- Lots of real world analogies, but
  - May be too specific to U.S.A.
  - Please ask about anything obscure

Start with a bad joke

Two statistics are sitting in a bar. One turns to the other and asks, “So, how do you like married life?”

The other statistic responds …

Put your reaction (“Ha ha”, “Groan”, etc.) in the chat box.

Introduction

- Tell us one interesting number about yourself
- Examples
  - 8: I have traveled to eight countries outside the United States
    - (Canada, Italy, China, France, Russia, England, Holland, and Iceland)
  - 29: I did not learn how to drive until I was 29 years old
  - 1802: My highest chess rating was 1802, but I am not that good any more.

Your turn

A bit more about myself

- PhD in Statistics in 1982 from the University of Iowa
- Currently full professor
- Part-time statistical consultant
- Funded on 18 research grants
- Over 100 peer-reviewed publications
- Website with over 2,000 pages
- Many invitations to talk at conferences

Outline of the three day course

- Day one: Numerical summaries and data visualization
- Day two: Hypothesis testing and sampling
- Day three: Statistical tests to compare treatment to a control and regression models

My goal: help you to become a better consumer of statistics

Day one topics

- Numerical summaries
  - When should you present the mean versus the median
  - When should you present the range versus standard deviation
  - How should you display percentages
  - Why should you round liberally

Day one topics (continued)

- Data visualization
  - How should you display continuous data
  - Why is the normal bell-shaped curve important
  - How should you display categorical data
  - How do you illustrate trends and patterns
  - What are some common mistakes in the choice of colors

Quiz question 1

```
          No             Yes           Total
Female   154 (33.3%)    308 (66.7%)    462 (100%)
Male     709 (83.3%)    142 (16.7%)    851 (100%)
Total    863 (65.7%)    450 (34.3%)   1313 (100%)
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer

Counting and proportions

- Counts are the most common statistic
  - Counts are error prone
  - Counts require a solid operational definition

Student exercise

Count the number of occurrences of the letter “e”.

```
A quality control  program is easiest
to implement from the top down.
Make sure that you understand the
the commitment of time and money
that is involved. Every workplace is
different, but think about allocating
10% of your time and 10% of the
time of all your employees to
quality control.
```

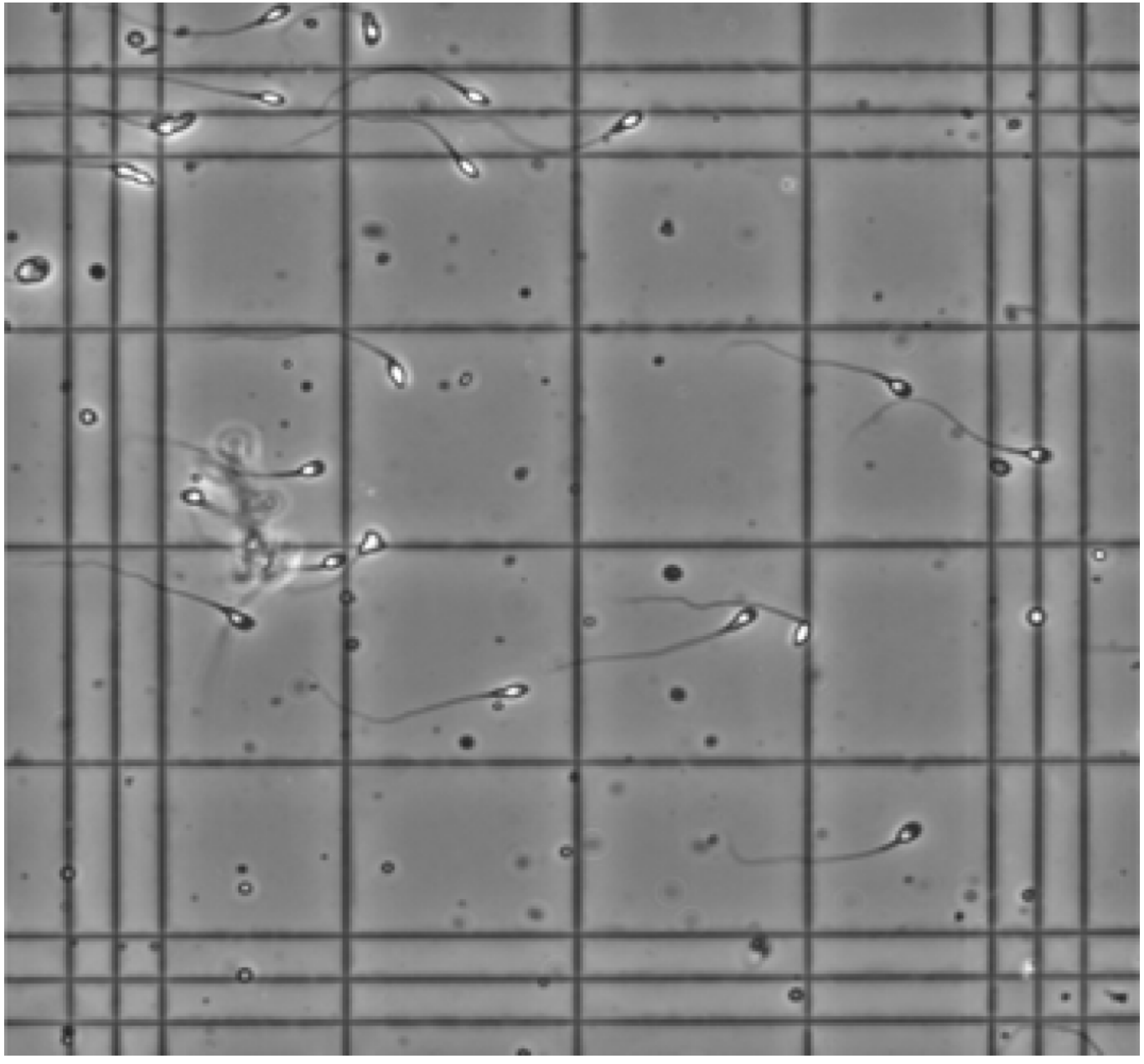


Figure 1: Image of a haemocytometer

Sex * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
Sex	female	154	308	462
	male	709	142	851
Total		863	450	1313

Figure 2: Titanic data: counts of survival by gender

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% within Survived	17.8%	68.4%	35.2%
	male	Count	709	142	851
		% within Survived	82.2%	31.6%	64.8%
Total		Count	863	450	1313
		% within Survived	100.0%	100.0%	100.0%

Figure 3: Titanic data with column percentages

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% within Sex	33.3%	66.7%	100.0%
	male	Count	709	142	851
		% within Sex	83.3%	16.7%	100.0%
Total		Count	863	450	1313
		% within Sex	65.7%	34.3%	100.0%

Figure 4: Titanic data with row percentages

Percentages divided by grand total

Sex * Survived Crosstabulation

			Survived		
			No	Yes	Total
Sex	female	Count	154	308	462
		% of Total	11.7%	23.5%	35.2%
	male	Count	709	142	851
		% of Total	54.0%	10.8%	64.8%
Total		Count	863	450	1313
		% of Total	65.7%	34.3%	100.0%

Figure 5: Titanic data with cell percentages

My recommendations

- Treatment or exposure as rows
- Outcome as columns
- Usually report row percentages
  - Female survival rate: 67%
  - Male survival rate: 17%
- But sometimes column percentages
  - Survivors: 68% female, 32% male

Some rationale for these choices

My way

		Survived	
		No	Yes
Sex	Female	33% (154)	67% (308)
	Male	83% (863)	17% (142)

Not my way

		Sex	
		Female	Male
Survived	No	33% (154)	83% (863)
	Yes	67% (308)	17% (142)

Break

- What have you just learned?
  - Displaying percentages
- What is coming next?
  - Practice exercise
  - Calculation of the mean and median

On your own

Calculate row and column percentages for the following tables. Interpret your results.

PClass * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
PClass	1st	129	193	322
	2nd	161	119	280
	3rd	573	138	711
Total		863	450	1313

Figure 6: Titanic passenger class counts

Child * Survived Crosstabulation

Count

		Survived		
		No	Yes	Total
Child	No	386	244	630
	Yes	57	69	126
Total		443	313	756

Figure 7: Titanic child counts

Figure 8: Cartoon image of Professor Mean

Figure 9: Road with a median strip

Calculation of the mean and median

- Mean
  - Add up all the values, divide by the sample size
- Median
  - Sort the data
    - Select the middle value if n is odd
    - go halfway between the two middle values if n is even

Formal mathematical definitions

- Mean
  - $\bar{X} = \frac{1}{n}\Sigma X_i$
- Median
  - Sorted values $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$
    - $X_{[(n+1)/2]}$ if n is odd,
    - $(X_{[n/2]} + X_{[n/2+1]})/2$ if n is even

Bacteria before and after A/C upgrade

Room	Before	After	Change
121	11.8	10.1	-1.7
125	7.1	3.8	-3.3
163	8.2	7.2	-1.0
218	10.1	10.5	0.4
233	10.8	8.3	-2.5
264	14	12	-2.0
324	14.6	12.1	-2.5
325	14	13.7	-0.3

Before remediation mean

11.8 + 7.1 + 8.2 + 10.1 + 10.8 + 14 + 14.6 + 14 = 90.6

90.6 / 8 = 11.325

Round to 11.3

After remediation mean

```
10.1 + 3.8 + 7.2 + 10.5 + 8.3 + 12 + 12.1 + 13.7 = 77.7

77.7 / 8 = 9.7125

Round to 9.7
```

Before remediation median (1/4)

```
121  11.8

125   7.1

163   8.2

218  10.1

233  10.8

264  14.0

324  14.6

325  14.0
```

Before remediation median (2/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0

325  14.0

324  14.6
```

Before remediation median (3/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8

121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

Before remediation median (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8  10.8
                   (10.8 + 11.8) / 2 = 11.3
121  11.8  11.8

264  14.0

325  14.0

324  14.6
```

After remediation median (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

After remediation median (2/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0

324  12.1

325  13.7
```

After remediation median (3/4)

```
125   3.8

163   7.2

233   8.3

121  10.1  10.1

218  10.5  10.5

264  12.0

324  12.1

325  13.7
```

After remediation median (4/4)

```
125   3.8

163   7.2

233   8.3

121  10.1  10.1
                    (10.1 + 10.5) / 2 = 10.3
218  10.5  10.5

264  12.0

324  12.1

325  13.7
```

Break

- What have you just learned?
  - Calculation of the mean and median
- What is coming next?
  - Criticisms of the mean and median

Criticisms of the mean and median

- Are you combining apples and onions?
- Are you ignoring minorities?

Use of the mean for ordinal data

- Stevens scales of measurement (controversial!)
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Addition/subtraction not allowed for ordinal data
  - Mean of ordinal data is meaningless

An example of ordinal data.

- “Do you agree or disagree with the following statements”
  - “I believe that knowledge of Statistics is important for my job.”
    - 1 = Strongly disagree,
    - 2 = Disagree
    - 3 = Neutral
    - 4 = Agree
    - 5 = Strongly agree

Another example of ordinal data, course grades

- A = 4
- B = 3
- C = 2
- D = 1
- F = 0

Phoenix5 To help men and their companions overcome the effects of prostate cancer

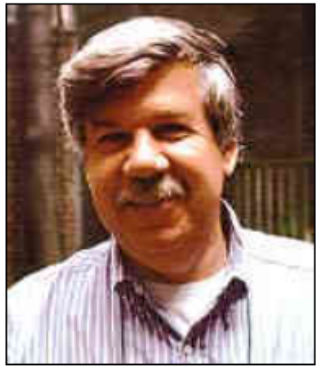
main menu - articles - prostate - stories - sexuality - resources - glossary - search

from
The Articles Menu

"This is a personal story of statistics..."

THE MEDIAN ISN'T THE MESSAGE

by Stephen Jay Gould



Born in 1941, Stephen Jay Gould was a geologist, zoologist, paleontologist and evolutionary biologist at Harvard. He was also one of the most noted, prolific and best-selling scientific writers of our day. He was diagnosed in 1982 with abdominal mesothelioma, a rare and very deadly form of cancer associated with exposure to asbestos. This is his story. It was first published in Discover magazine in June 1985 and was reprinted here at Phoenix5 with his kind permission. He beat the cancer for 20 years, finally passing on May 20, 2002, giving all of us a valuable lesson in beating the odds.

Figure 10: Excerpt from Gould 1985 publication

Choosing between the mean and median

- Often a source of controversy
- When do you use the mean?
  - When totals are important
- When do you use the median
  - When outliers/skewness might distort your conclusions
- Often, either is fine

Eur Respir J 2001; 17: 495–498
Printed in UK – all rights reserved

European Respiratory Journal
ISSN 0903-1936

Airway resistance measured by the interrupter technique: expiration or inspiration, mean or median?

P.D. Bridge, S.A. McKenzie

Figure 11: Exceprt from Bridge and McKenzie 2001, PMID: 11405531

Bridge 2001, PMID: 11405531 (continued)

The measurement of airway resistance by the interrupter technique (Rint) needs standardization. Should measurements be made be during the expiratory or inspiratory phase of tidal breathing? **In reported studies, the measurement of Rint has been calculated as the median or mean of a small number of values, is there an important difference?**

Bridge 2001, PMID: 11405531 (continued)

In the present data the mean of a set of values contributing to a measurement was not significantly different from the median. **However, the use of the median has been recommended since it is less affected by possible outlying values such as might be included by fully automated equipment.**

Published in final edited form as:

Value Health. 2019 December ; 22(12): 1387–1395. doi:10.1016/j.jval.2019.08.005.

Trends in the Price per Median and Mean Life-Year Gained Among Newly Approved Cancer Therapies 1995 to 2017

Alice J. Chen, PhD[1,2,*], **Xiaohan Hu, MPH**[2], **Rena M. Conti, PhD**[3], **Anupam B. Jena, MD, PhD**[4,5], **Dana P. Goldman, PhD**[1,2,5]

Figure 12: Chen et al 2019

Chen 2019, PMID: 31806195 (continued)

Background: The prices of newly approved cancer drugs have risen over the past decades. **A key policy question is whether the clinical gains offered by these drugs in treating specific cancer indications justify the price increases.**

Chen 2019, PMID: 31806195 (continued)

Results: We found that between 1995 and 2012, price increases outstripped median survival gains, a finding consistent with previous literature. **Nevertheless, price per mean life-year gained increased at a considerably slower rate, suggesting that new drugs have been more effective in achieving longer-term survival.** Between 2013 and 2017, price increases reflected equally large gains in median and mean survival, resulting in a flat profile for benefit-adjusted launch prices in recent years.

Break

- What have you just learned?
  - Criticisms of the mean and median
- What is coming next?
  - Computing percentiles

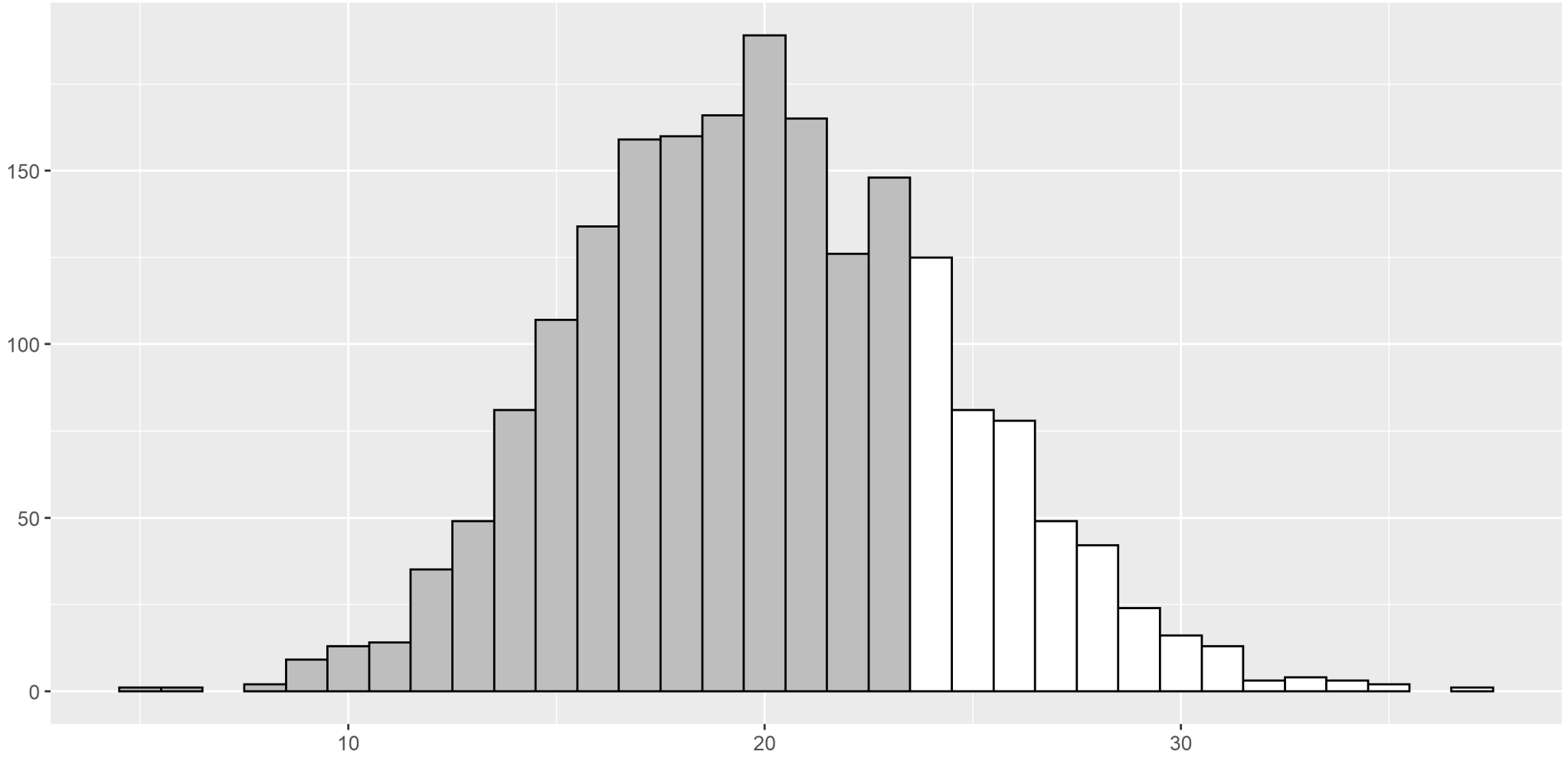


Figure 13: Illustration of the 75th percentile

Computing percentiles

- Many formulas
  - Differences are not worth fighting over
- My preference (pth quantile)
  - Sort the data
  - Calculate p*(n+1)
  - Is it a whole number?
    - Yes: Select that value, otherwise
    - No: Go halfway between
    - Special cases: p(n+1) < 1 or > n

Some examples of percentile calculations

- Example for n=39
  - For 5th percentile, p(n+1)=2 -> 2nd smallest value
  - For 4th percentile, p(n+1)=1.6 -> halfway between two smallest values
  - For 2nd percentile, p(n+1)=0.8 -> smallest value

Some terminology

- Percentile: goes from 0% to 100%
- Quantile: goes from 0.0 to 1.0
  - 90th percentile = 0.9 quantile
- 25th, 50th, and 75th percentiles: quartiles
  - 25th percentile: Q_1, $X_{0.25}$ or lower quartile
  - Median/50th percentiles: Q_2 or $X_{0.5}$
  - 75th percentile: Q_3, $X_{0.75}$ or upper quartile

Before remediation upper quartile (1/4)

121	11.8
125	7.1
163	8.2
218	10.1
233	10.8
264	14.0
324	14.6
325	14.0

Before remediation upper quartile (2/4)

125	7.1
163	8.2
218	10.1
233	10.8
121	11.8
264	14.0
325	14.0
324	14.6

Before remediation upper quartile (3/4)

125	7.1	
163	8.2	
218	10.1	
233	10.8	
121	11.8	
264	14.0	14
325	14.0	14
324	14.6	

Before remediation upper quartile (4/4)

```
125   7.1

163   8.2

218  10.1

233  10.8

121  11.8

264  14.0  14
                   (14 + 14) / 2 = 14
325  14.0  14

324  14.6
```

After remediation upper quartile (1/4)

```
121  10.1

125   3.8

163   7.2

218  10.5

233   8.3

264  12.0

324  12.1

325  13.7
```

After remediation upper quartile (2/4)

125	3.8
163	7.2
233	8.3
121	10.1
218	10.5
264	12.0
324	12.1
325	13.7

After remediation upper quartile (3/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0  12

324  12.1  12.1

325  13.7
```

After remediation upper quartile (4/4)

```
125   3.8

163   7.2

233   8.3

121  10.1

218  10.5

264  12.0  12
                   (12 + 12.1) / 2 = 12.05
324  12.1  12.1

325  13.7
```

When you should use percentiles

- Characterize variation
  - Middle 50% of the data
- Exposure issues
  - Not enough to control median exposure level
- Quantify extremes
  - What does “upper class” mean?
- Quality control
  - Almost all products must meet a minimum standard

Break

- What have you just learned?
  - Computing percentiles
- What is coming next?
  - Computing the standard deviation

Standard deviation

$$S = \sqrt{\frac{1}{n-1}\Sigma(X_i - \bar{X})^2}$$

At least one alternative formulas.

Why is variation important

- Variation = Noise
  - Too much noise can hide signals
- Variation = Heterogeneity
  - Too little heterogeneity, hard to generalize
  - Too much heterogeneity, mixing apples and oranges
- Variation = Unpredictability
  - Too much unpredictability, hard to prepare for the future
- Variation = Risk
  - Too much risk can create a financial burden

Should you try to minimize variation?

- Yes, for early studies
  - Easier to detect signals
  - Proof of concept trials
- No, for later studies
  - Easier to generalize results
  - Pragmatic trials

Standard deviation

$$S = \sqrt{\frac{1}{n-1}\Sigma(X_i - \bar{X})^2}$$

At least one alternative formulas.

The bell shaped curve

- Does your variation follow a bell shaped curve?
- Synonyms: normality, normal distribution
  - Values in the middle are most common
  - Frequencies taper off away from the center
  - Symmetry on either side
- A bell shaped curve = better characterization of variation

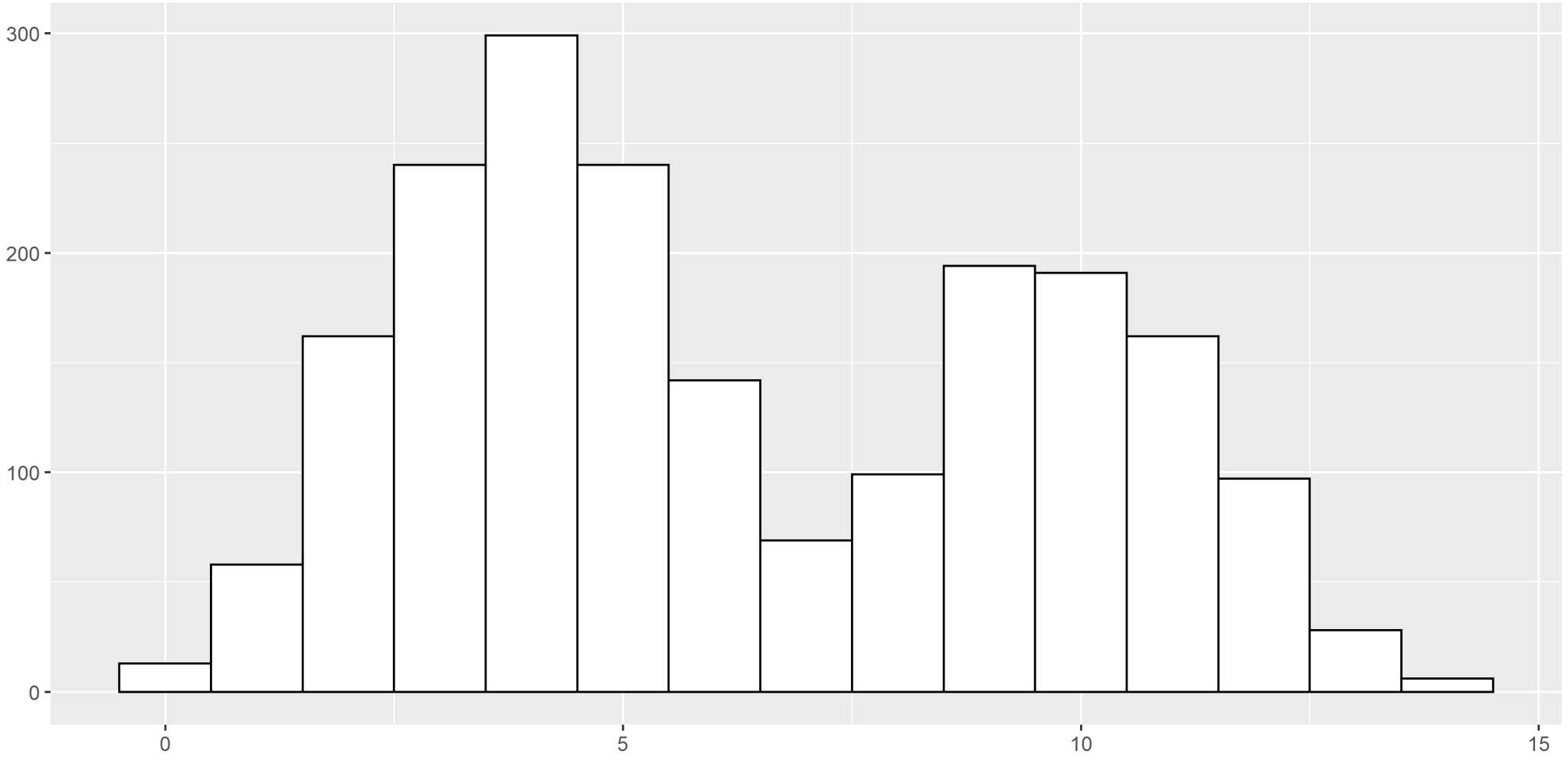


Figure 14: Bimodal histogram, not a bell shaped curve

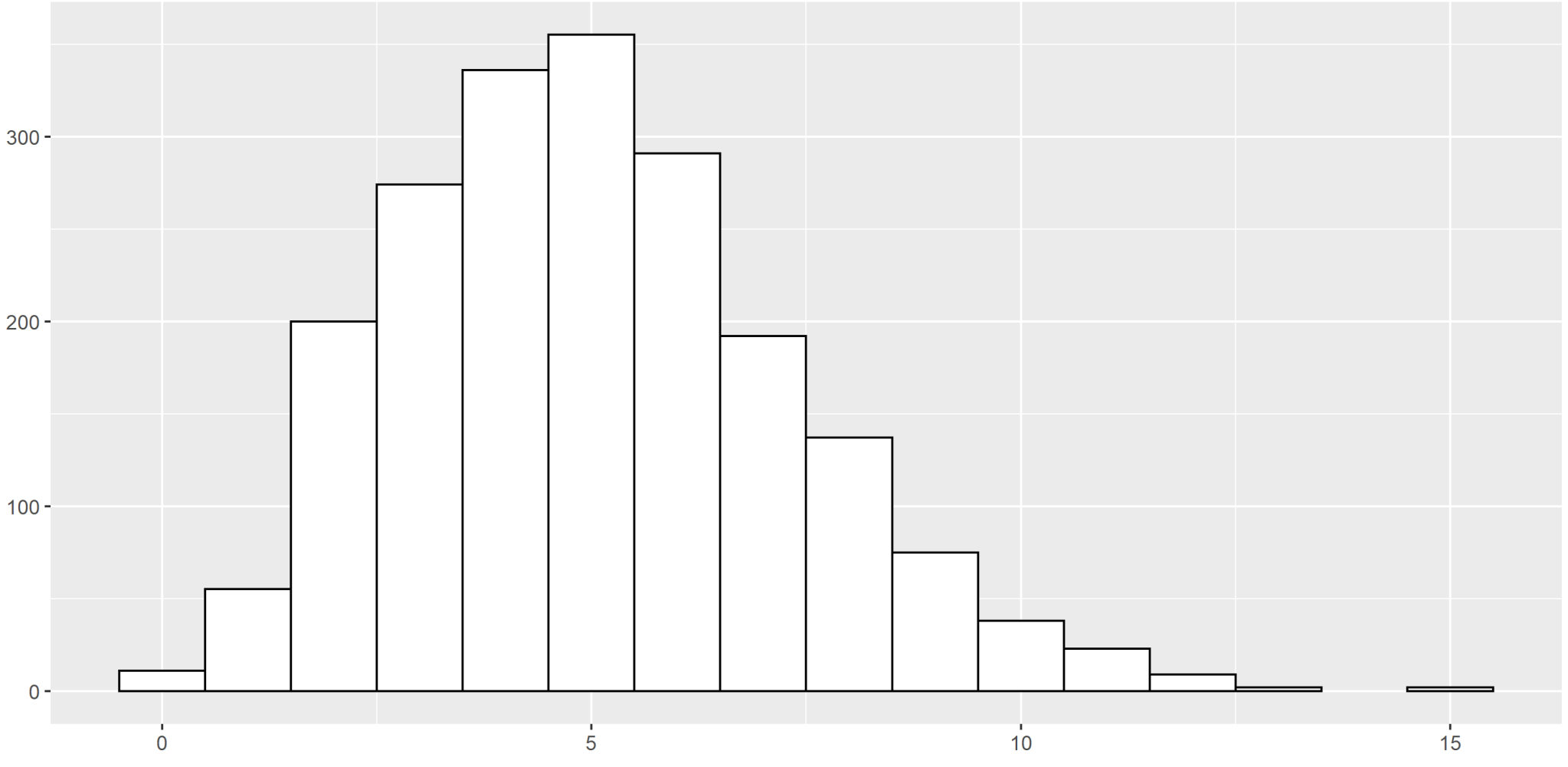


Figure 15: Skewed histogram, not a bell shaped curve

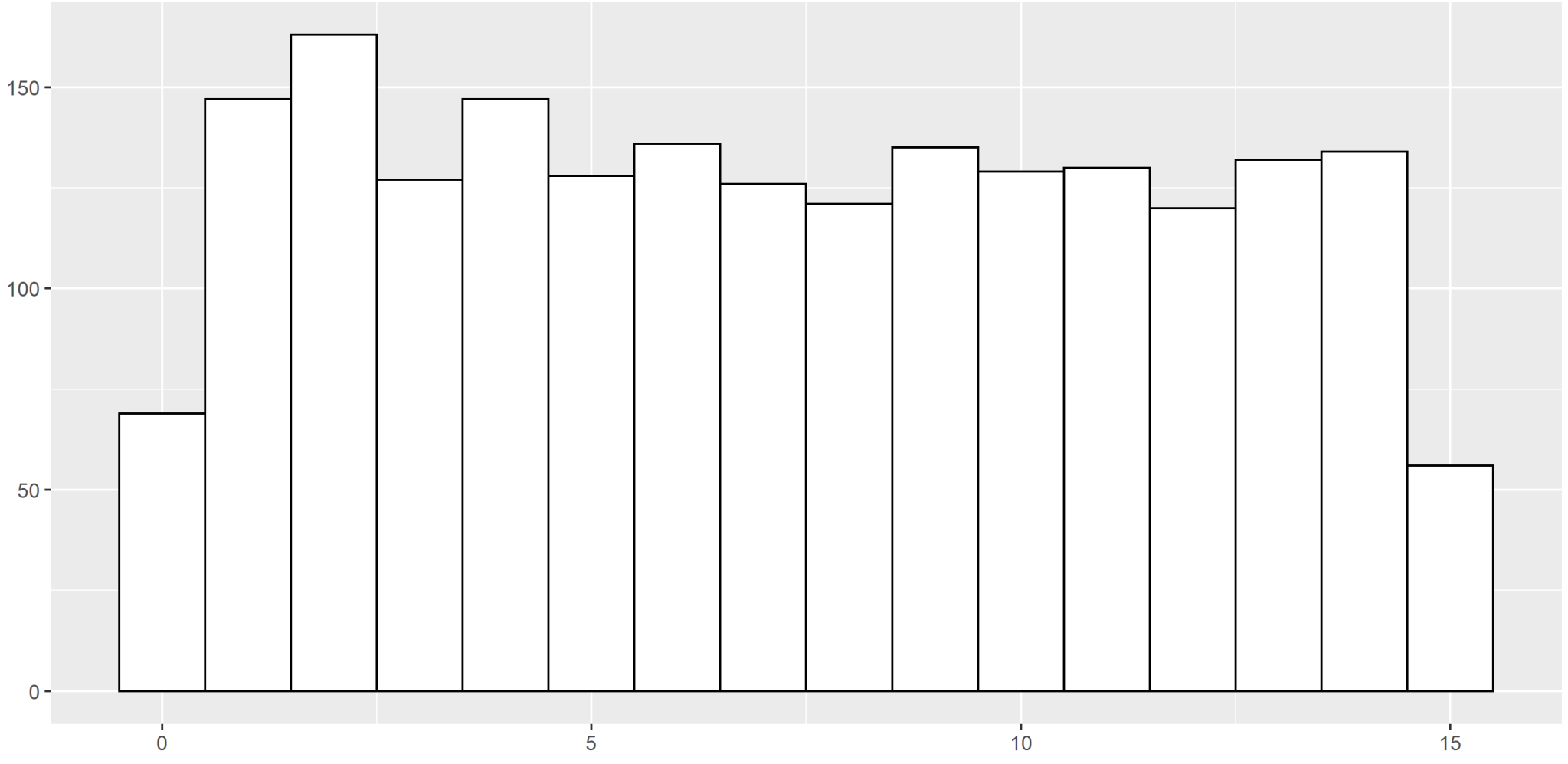


Figure 16: Uniform histogram, not a bell shaped curve

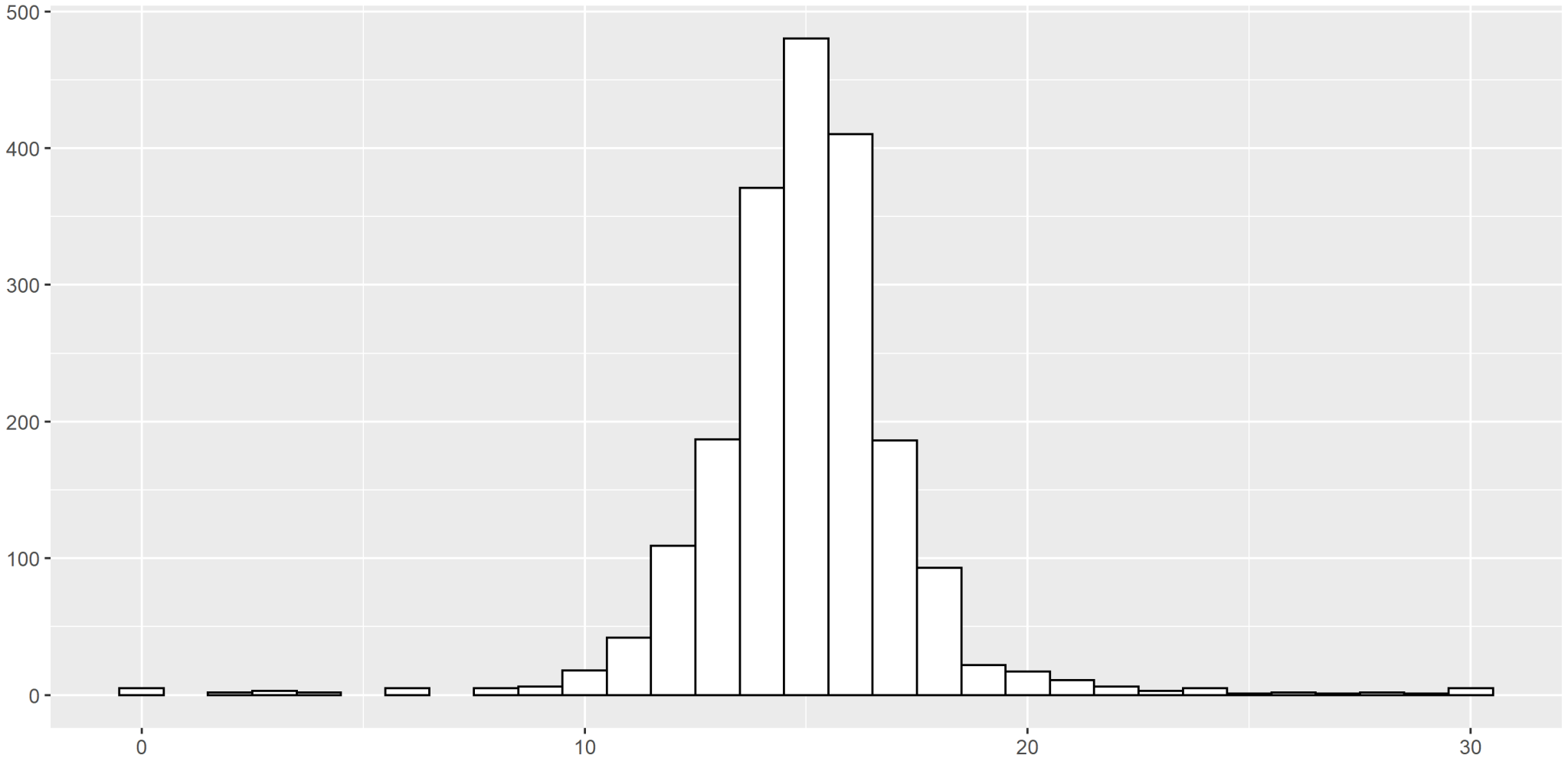


Figure 17: Heavy-tailed histogram, not a bell shaped curve

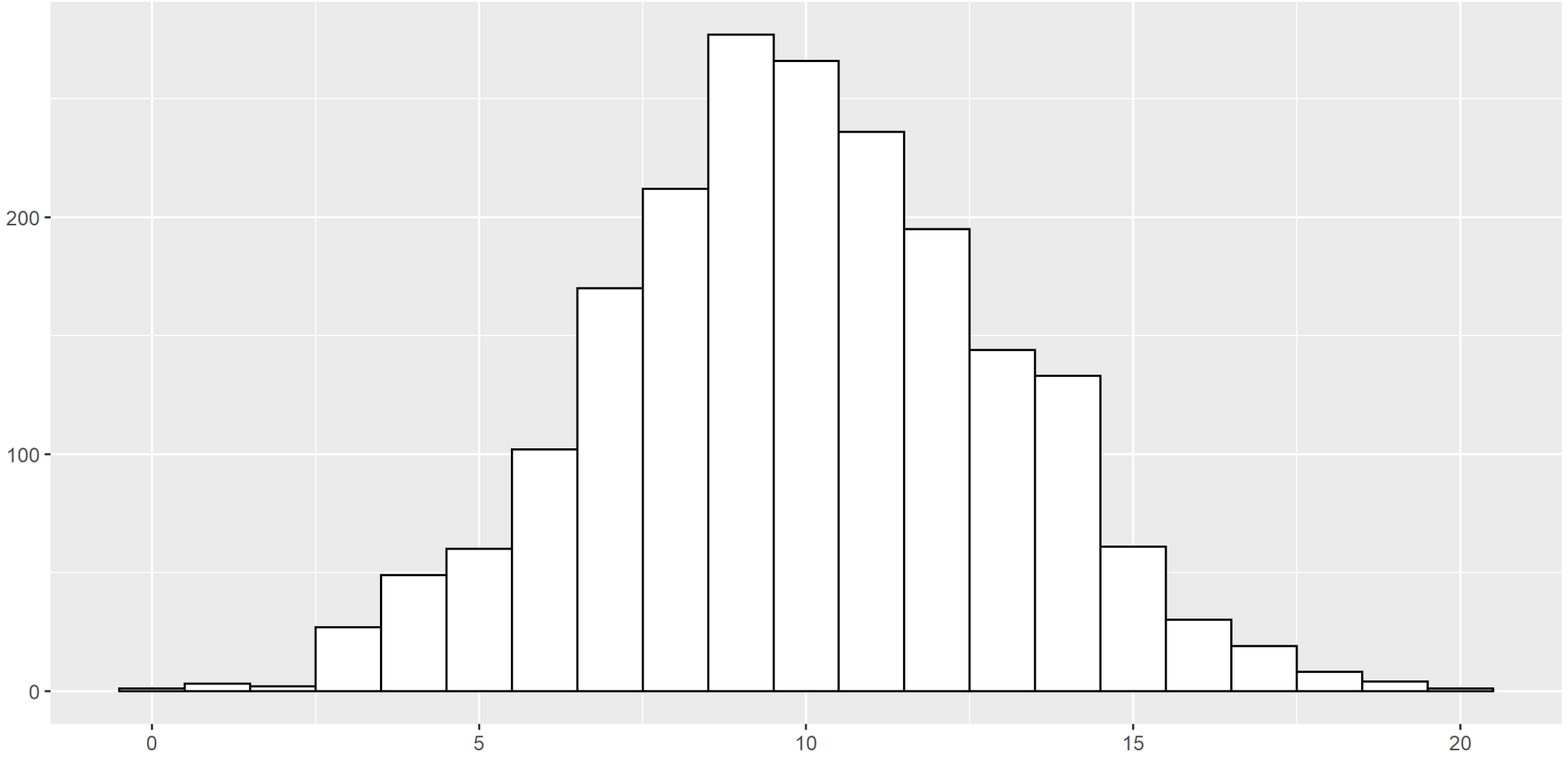


Figure 18: Bell-shaped histogram, finally!

Why concern yourself with the bell shaped curve?

- You can characterize individual observations
- You can characterize summary measures

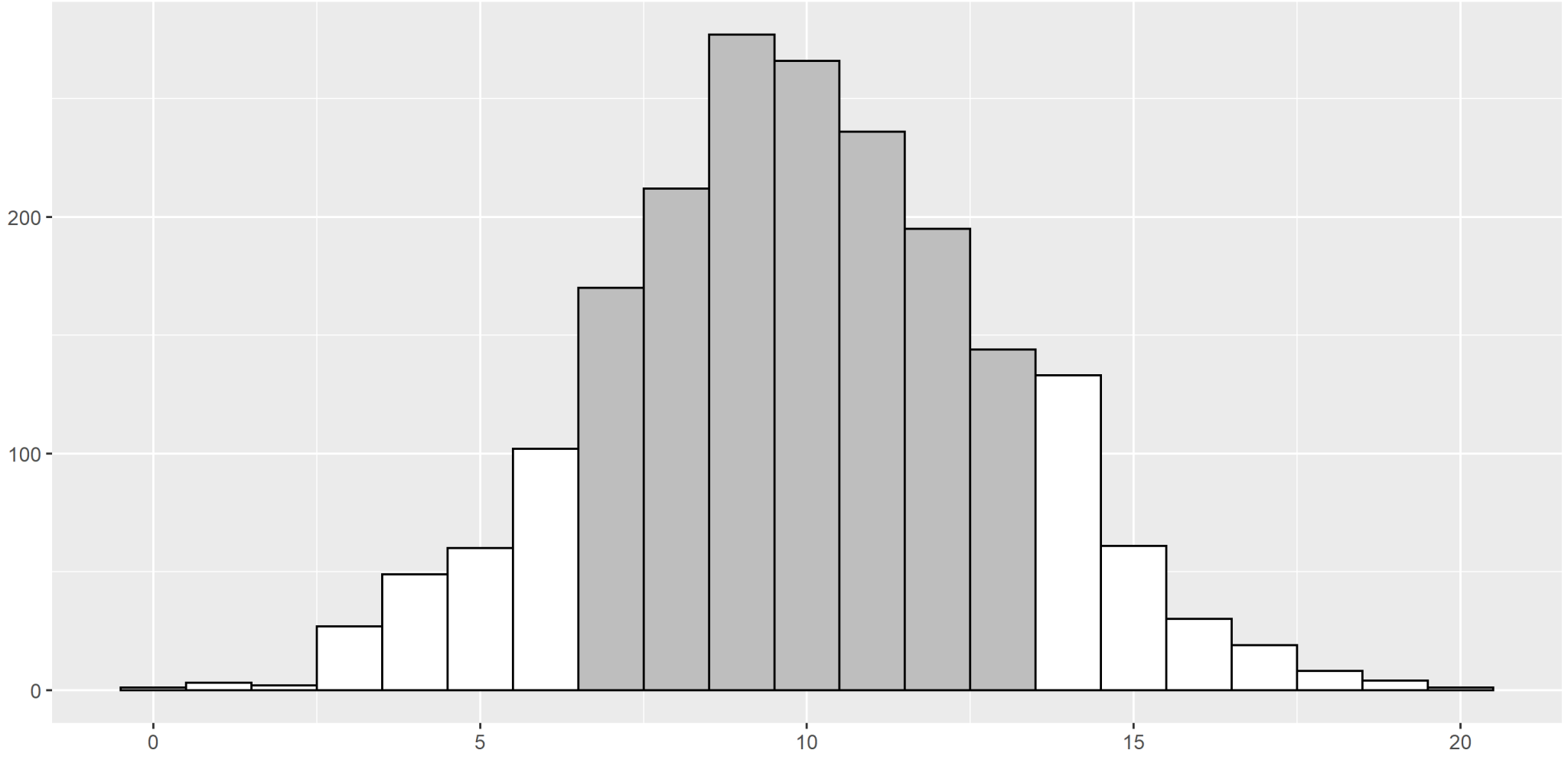


Figure 19: Percentage within one standard deviation

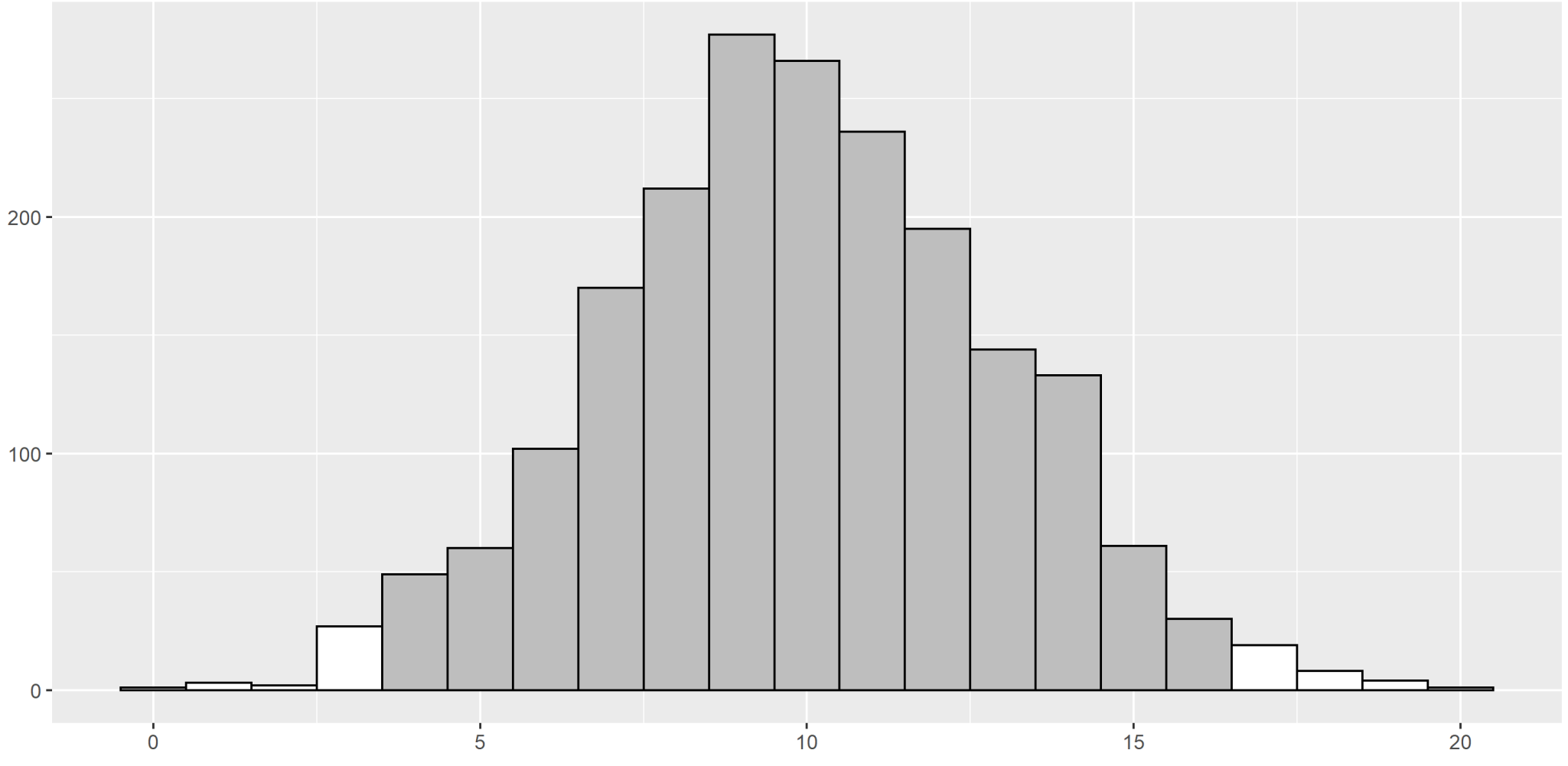


Figure 20: Percentage within two standard deviations

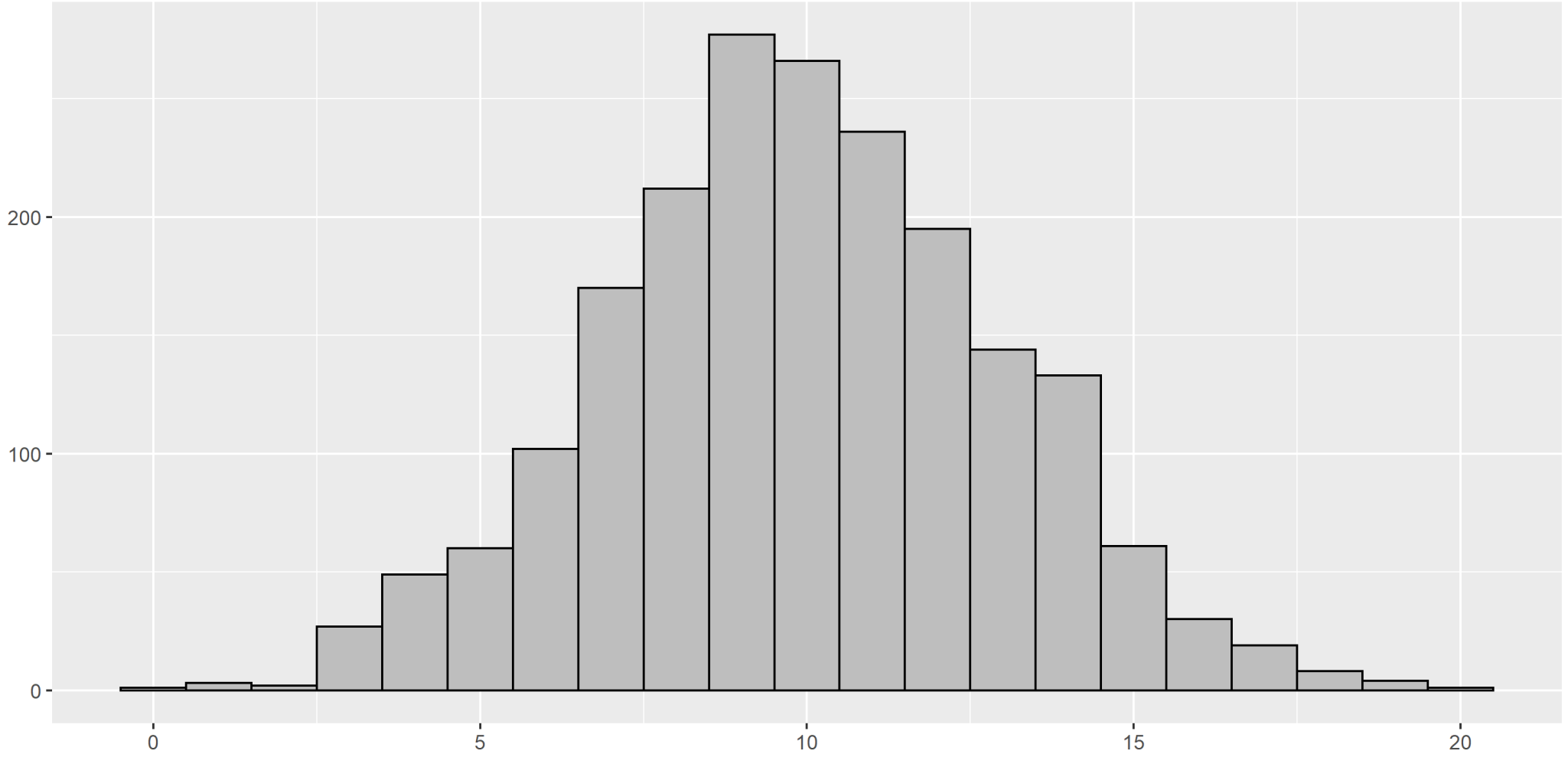


Figure 21: Percentage within three standard deviations

Behavior of the mean versus an individual

- Central Limit Theorem
  - Sample mean is approximately normal
  - Even if individual observations are not
- Standard error: $S/\sqrt{n}$

Diagnosing distributional issues (1/2)

- For all data
  - $\bar{X} \gg X_{0.5}$
  - $\bar{X}\ and/or\ X_{0.5}$ not midway between Q_1 and Q_3
  - $\bar{X}\ and/or\ X_{0.5}$ not midway between min and max

Diagnosing distributional issues (2/2)

- For non-negative data
  - $S > 0.5 \times \bar{X}$
- For data with an lower and/or upper bound
  - $Q1 = lower bound
  - $Q3 = upper bound
- Don’t overdiagnose, especially with small sample sizes!

› Prim Care Companion CNS Disord. 2022 Sep 13;24(5):21m03173. doi: 10.4088/PCC.21m03173.

Unemployment, Homelessness, and Other Societal Outcomes Among US Veterans With Schizophrenia Relapse: A Retrospective Cohort Study

Dee Lin [1 2], Hyunchung Kim [3], Keiko Wada [3], Maya Aboumrad [4], Ethan Powell [4], Gabrielle Zwain [4], Carmela Benson [1], Aimee M Near [3]

Affiliations + expand

PMID: 36126916 DOI: 10.4088/PCC.21m03173

Figure 22: Lin et al 2022, PMID: 36126916

Variable	Relapse Cohort (n = 16,862)	Nonrelapse Cohort (n = 16,862)
Age at index, y		
Mean (SD)	56.4 (13.3)	56.6 (13.0)
Median (Q1, Q3)	58 (51, 64)	59 (51, 65)
Minimum	20	20
Maximum	99	98

Figure 23: Excerpt from Table 1 of Lin et al 2022: ages

CCI		
Mean (SD)	2.5 (2.7)	2.4 (2.4)
Median (Q1, Q3)	2 (0, 4)	2 (0, 4)
Minimum	0	0
Maximum	20	18

Figure 24: Excerpt from Table 1 of Lin et al 2022: CCI

PHQ-2 score, n (%)[f]	9,472 (56.1)	10,848 (64.4)
Mean (SD)	1.1 (1.7)	1.0 (1.6)
Median (Q1, Q3)	0 (0, 2)	0 (0, 2)
Minimum	0	0
Maximum	6	6

Figure 25: Excerpt from Table 1 of Lin et al 2022: PHQ-2

J Am Med Dir Assoc. 2021 Sep;22(9):1840-1844. doi: 10.1016/j.jamda.2021.07.003. Epub 2021 Jul 19.

Prevalence and Predictors of Persistence of COVID-19 Symptoms in Older Adults: A Single-Center Study

Matteo Tosato [1], Angelo Carfì [1], Ilaria Martis [1], Cristina Pais [1], Francesca Ciciarello [1], Elisabetta Rota [1], Marcello Tritto [1], Andrea Salerno [1], Maria Beatrice Zazzara [1], Anna Maria Martone [1], Annamaria Paglionico [1], Luca Petricca [1], Vincenzo Brandi [1], Gennaro Capalbo [1], Anna Picca [1], Riccardo Calvani [2], Emanuele Marzetti [3], Francesco Landi [3]; Gemelli Against COVID-19 Post-Acute Care Team

Affiliations + expand

PMID: 34352201 PMCID: PMC8286874 DOI: 10.1016/j.jamda.2021.07.003

Figure 26: Tosato et al 2021, PMID: 34352201

Tosato 2021, PMID: 34352201 (continued)

Symptom persistence weeks after laboratory-confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) clearance is a relatively common long-term complication of Coronavirus disease 2019 (COVID-19). Little is known about this phenomenon in older adults. The present study aimed at determining the prevalence of persistent symptoms among older COVID-19 survivors and identifying symptom patterns.

Tosato 2021, PMID: 34352201 (continued)

The mean age was 73.1 ± 6.2 years (median 72, interquartile range 27), and 63 (38.4%) were women. The average time elapsed from hospital discharge was 76.8 ± 20.3 days (range 25-109 days).

Ielapi 2021, PMID: 34968328

Nurs Rep. 2021 Jul 12;11(3):530-535. doi: 10.3390/nursrep11030050.

Insomnia Prevalence among Italian Night-Shift Nurses

Nicola Ielapi [1 2], Michele Andreucci [3], Umberto Marcello Bracale [4], Davide Costa [2 5], Egidio Bevacqua [2 6], Andrea Bitonti [7], Sabrina Mellace [8], Gianluca Buffone [9], Stefano Candido [10], Michele Provenzano [6], Raffaele Serra [2 6]

Affiliations + expand

PMID: 34968328 PMCID: PMC8608071 DOI: 10.3390/nursrep11030050

Figure 27: Ielapi et al 2021, PMID: 34968328

Ielapi 2021, PMID: 34968328 (continued)

Background. Insomnia is one of the major health problems related with a decrease in quality of life (QOL) and also in poor functioning in night-shift nurses, that also may negatively affect patients' care. The aim of this study is to evaluate the prevalence of insomnia in night shift nurses.

Ielapi 2021, PMID: 34968328 (continued)

Excerpt from Table 1.

Data reported as mean ± standard deviation or median [Q1-Q3]

Overall (n = 2′355)	
Age, years	40.4 ± 10.3
Months of work	168 [72–300]
Night shifts per month, number	6.3 ± 1.4
Time to reach workplace, minutes	45 [45–65]
Rest time, minutes	180 [4–240]
Rest in the afternoon, minutes	30 [0–120]
Number of coffees, mean	2.5 ± 1.5
Number of coffees during night shift, mean	1.4 ± 1.1

Break

- What have you just learned?
  - Computing the standard deviation
- What is coming next?
  - Visualization

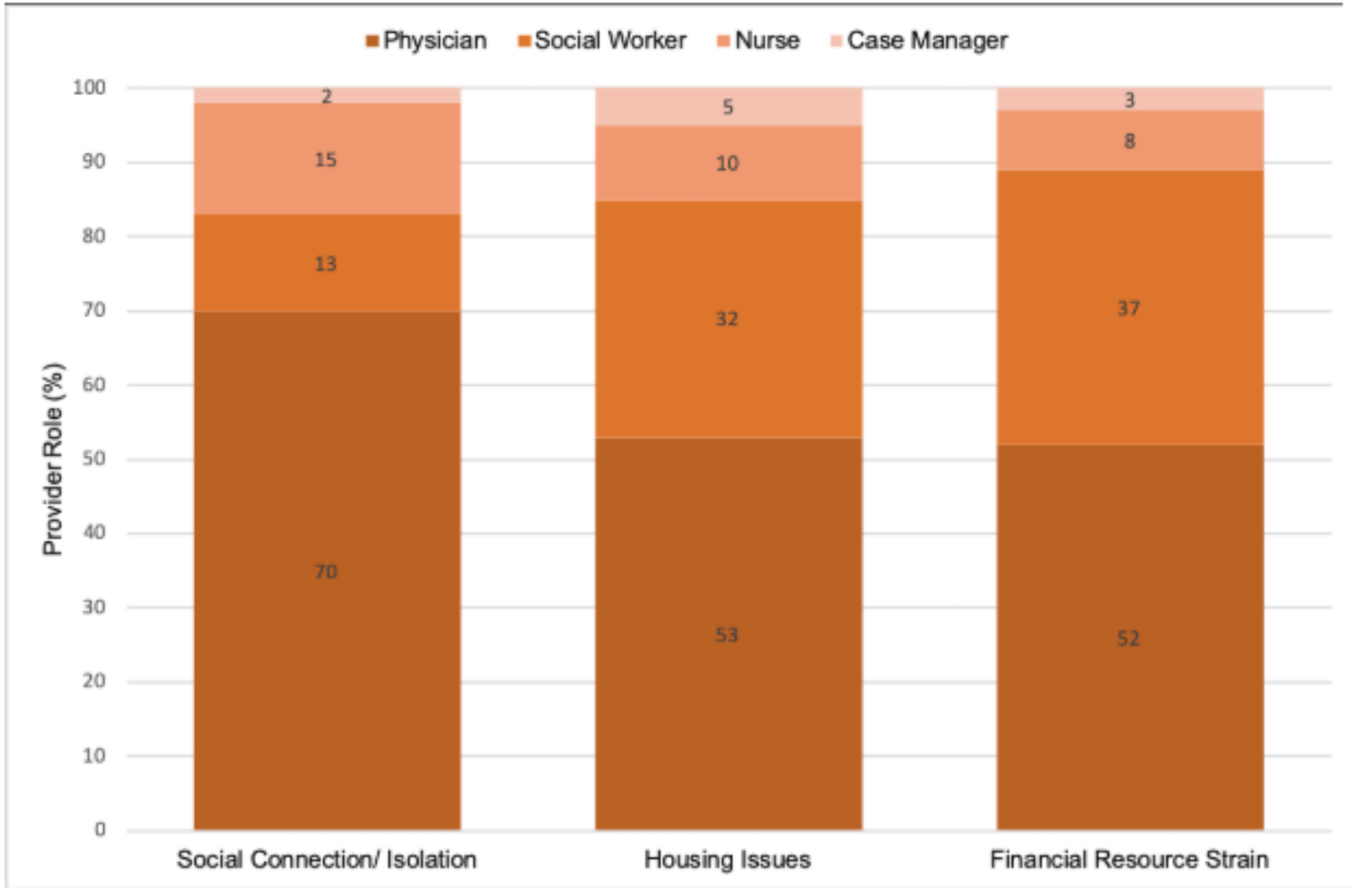


Figure 1. Characteristics of the electronic health record's unstructured data containing social and behavioral determinants of health, stratified by provider role.

Entries in the electronic health record by job title

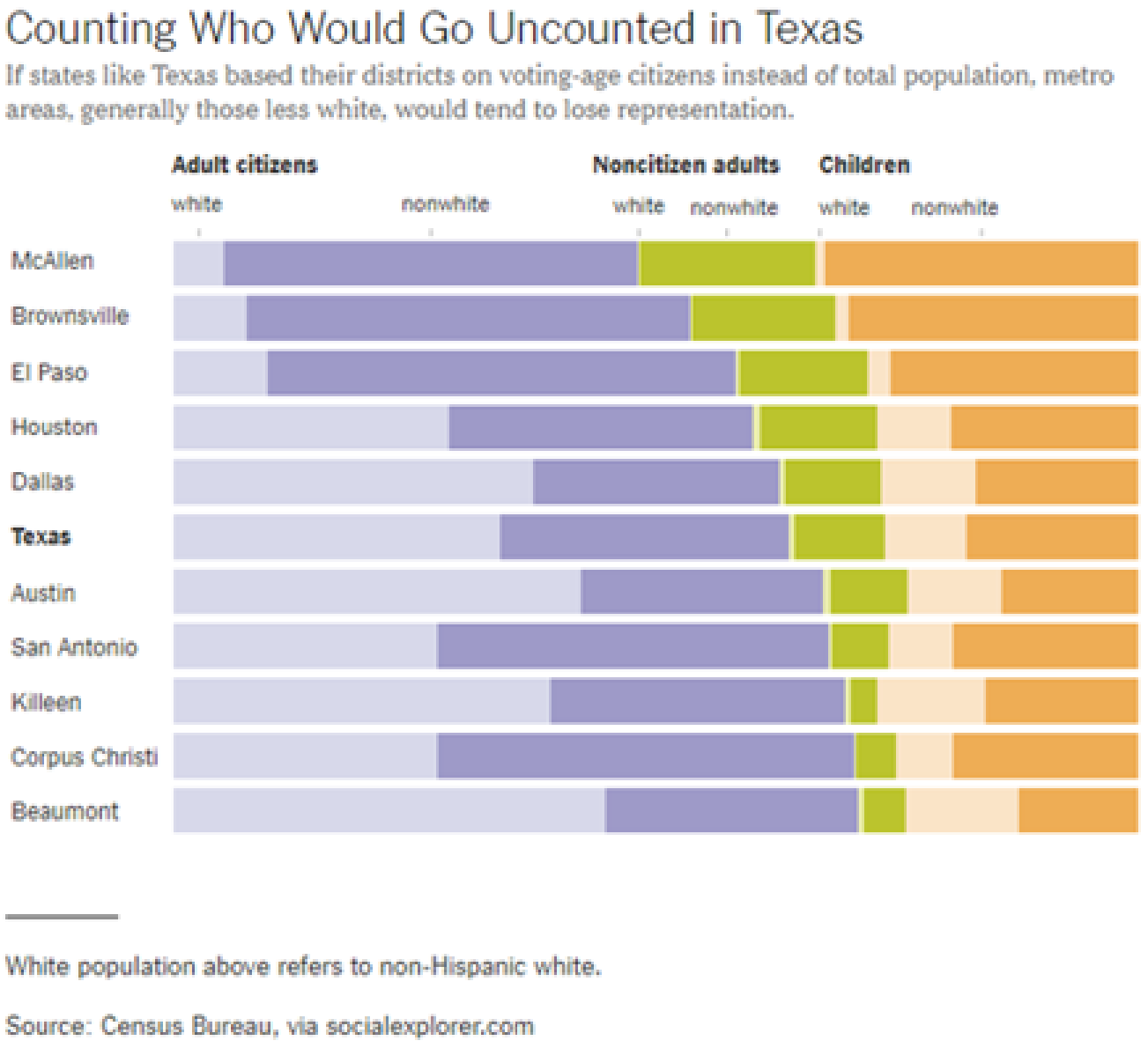


Figure 28: Demographic distribution of voters and non-voters in Texas

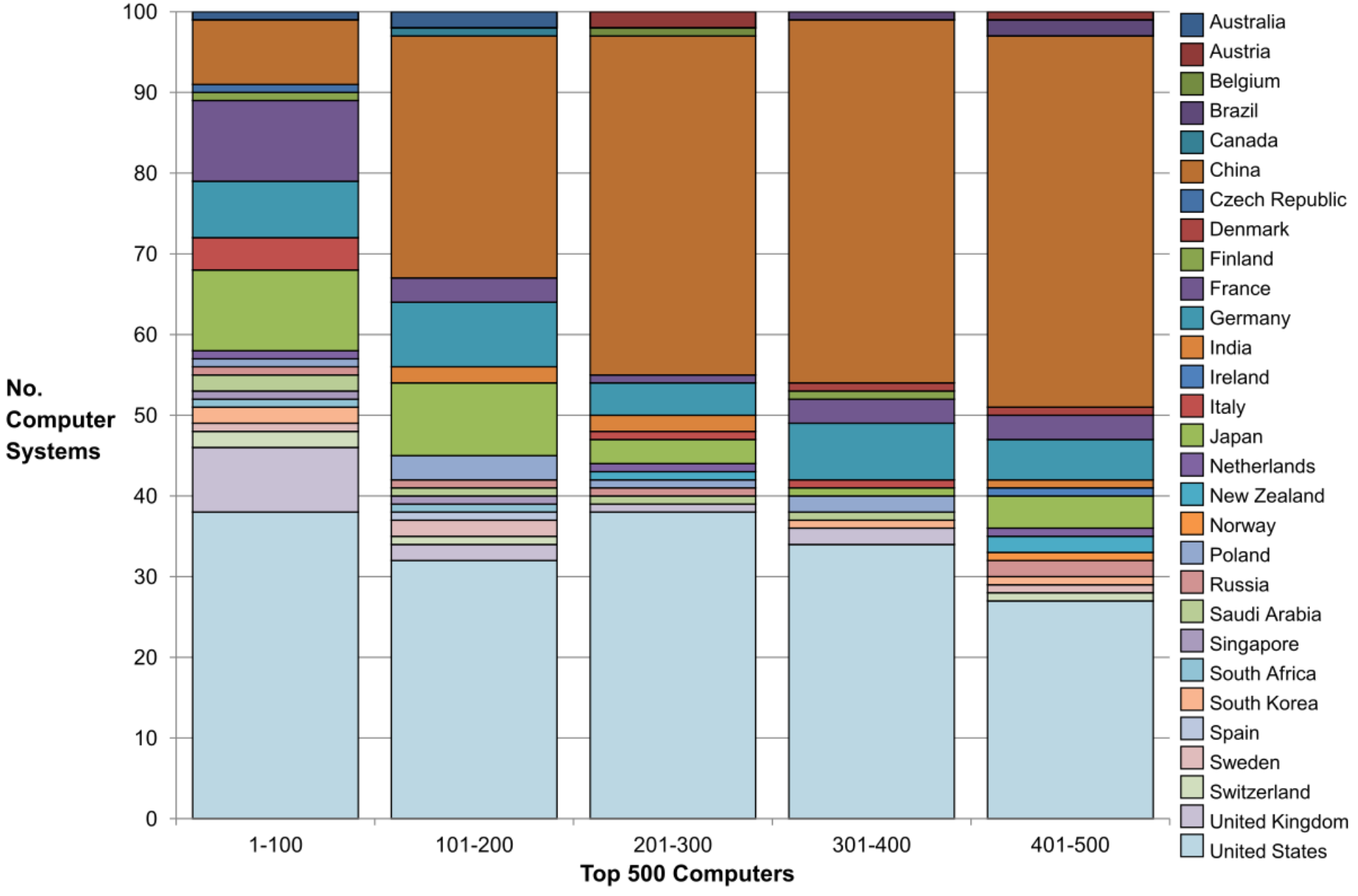


Figure 29: Fastest computers by country

Visualization

- Categorical data
  - Pie charts
  - Bar charts
- Continuous data
  - Bar charts
  - Error bars
  - Boxplot
  - Histogram
  - Plot all the data

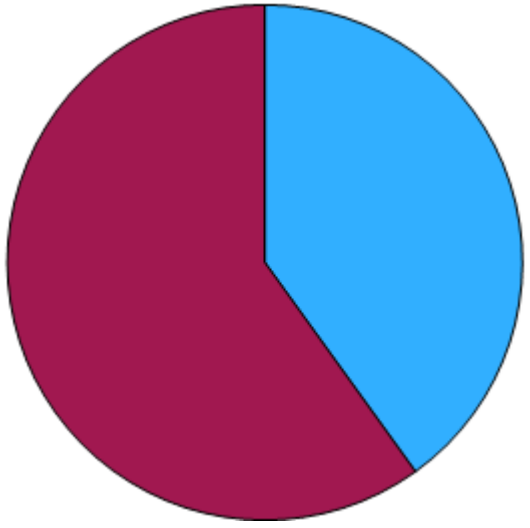


Figure 30: Survivors among first class passengers

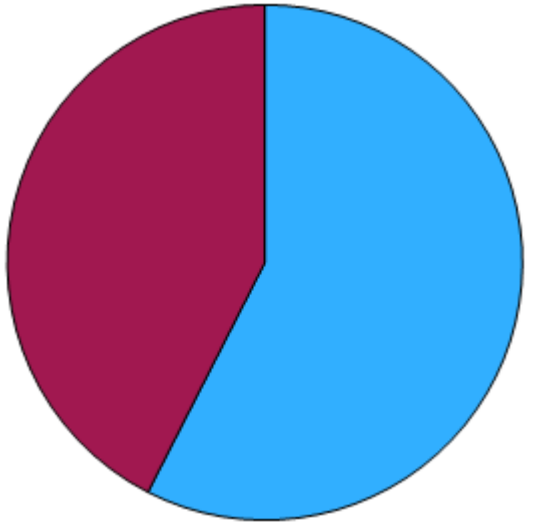


Figure 31: Survivors among second class passengers

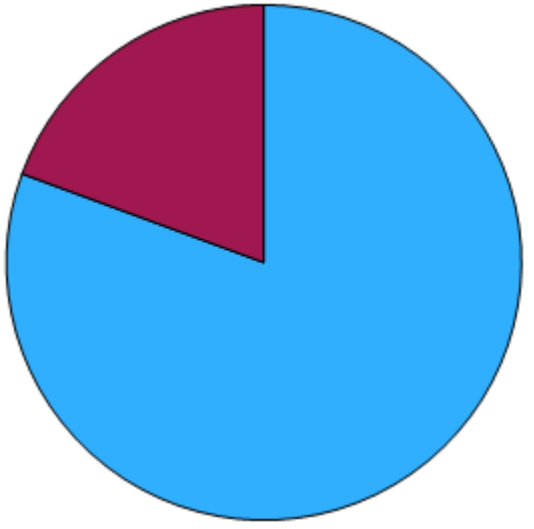


Figure 32: Survivors among third class passengers

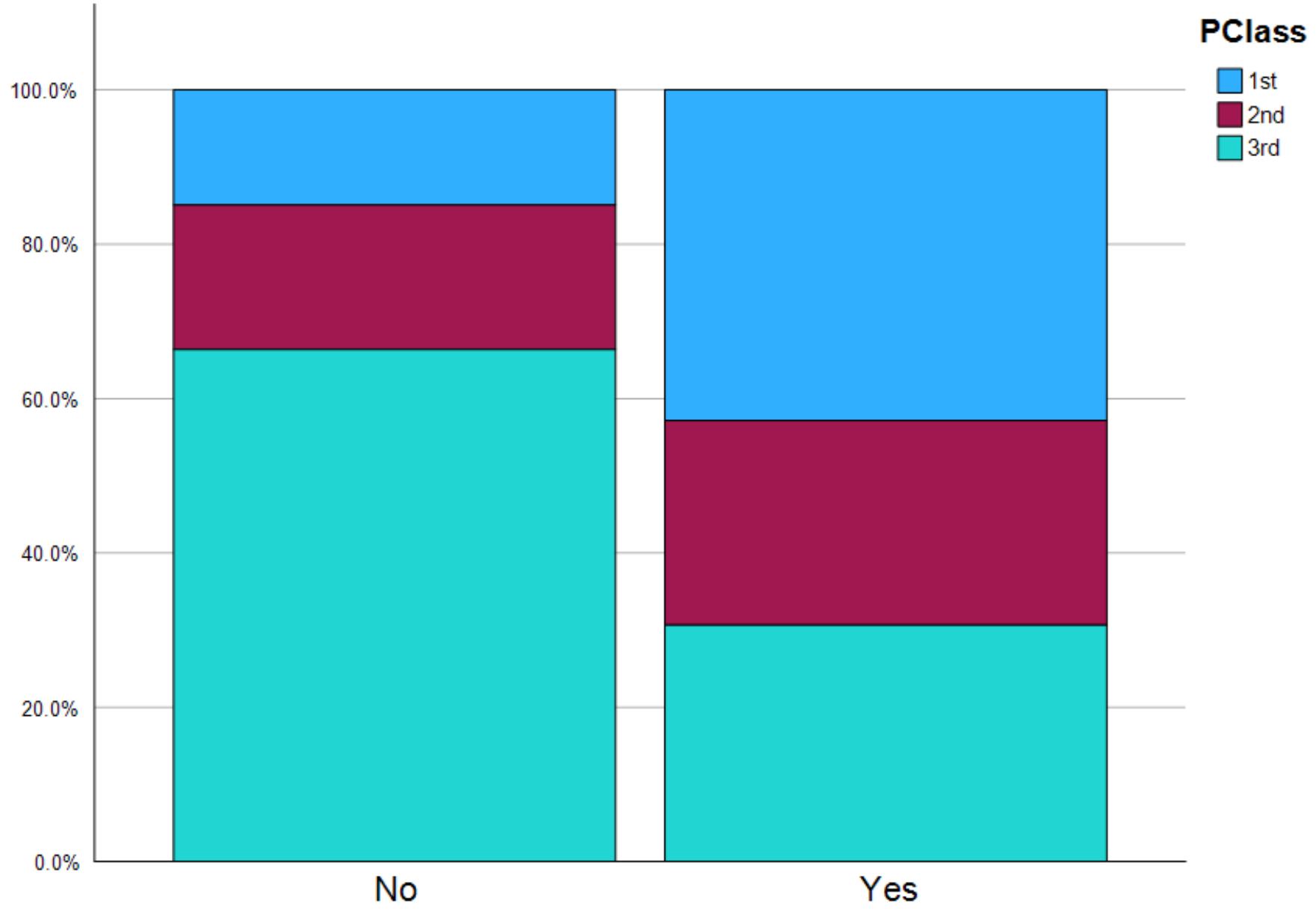
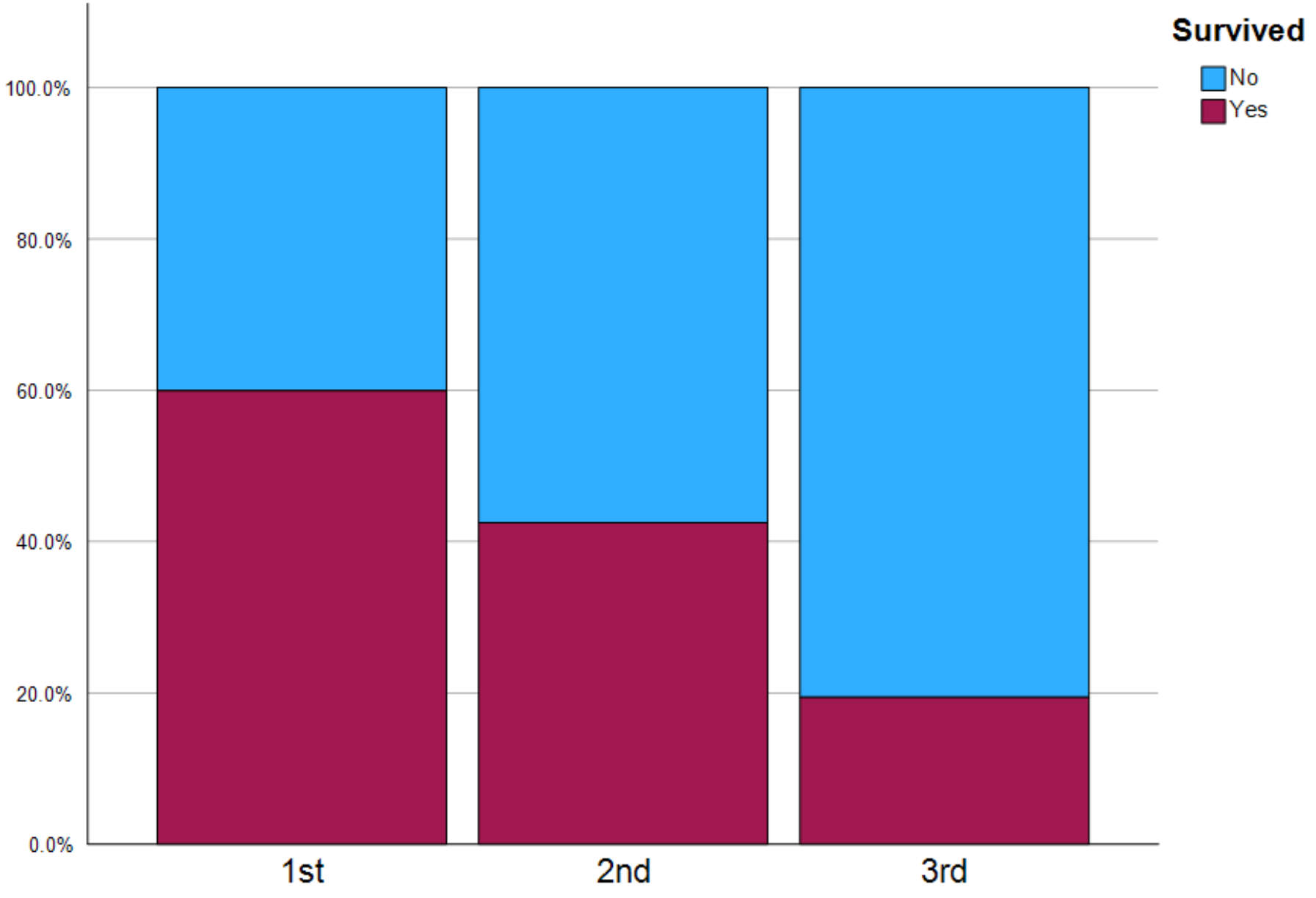
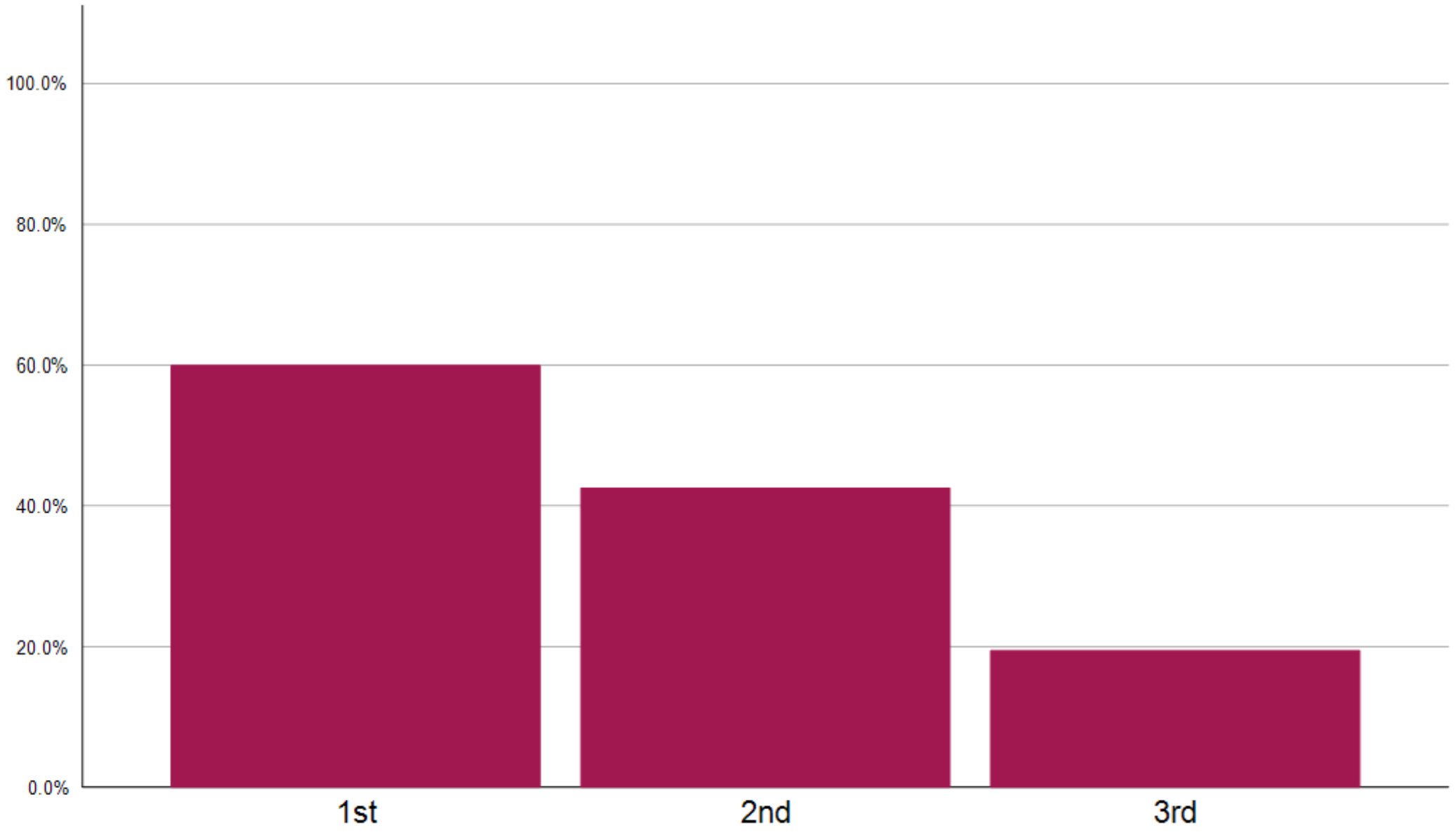


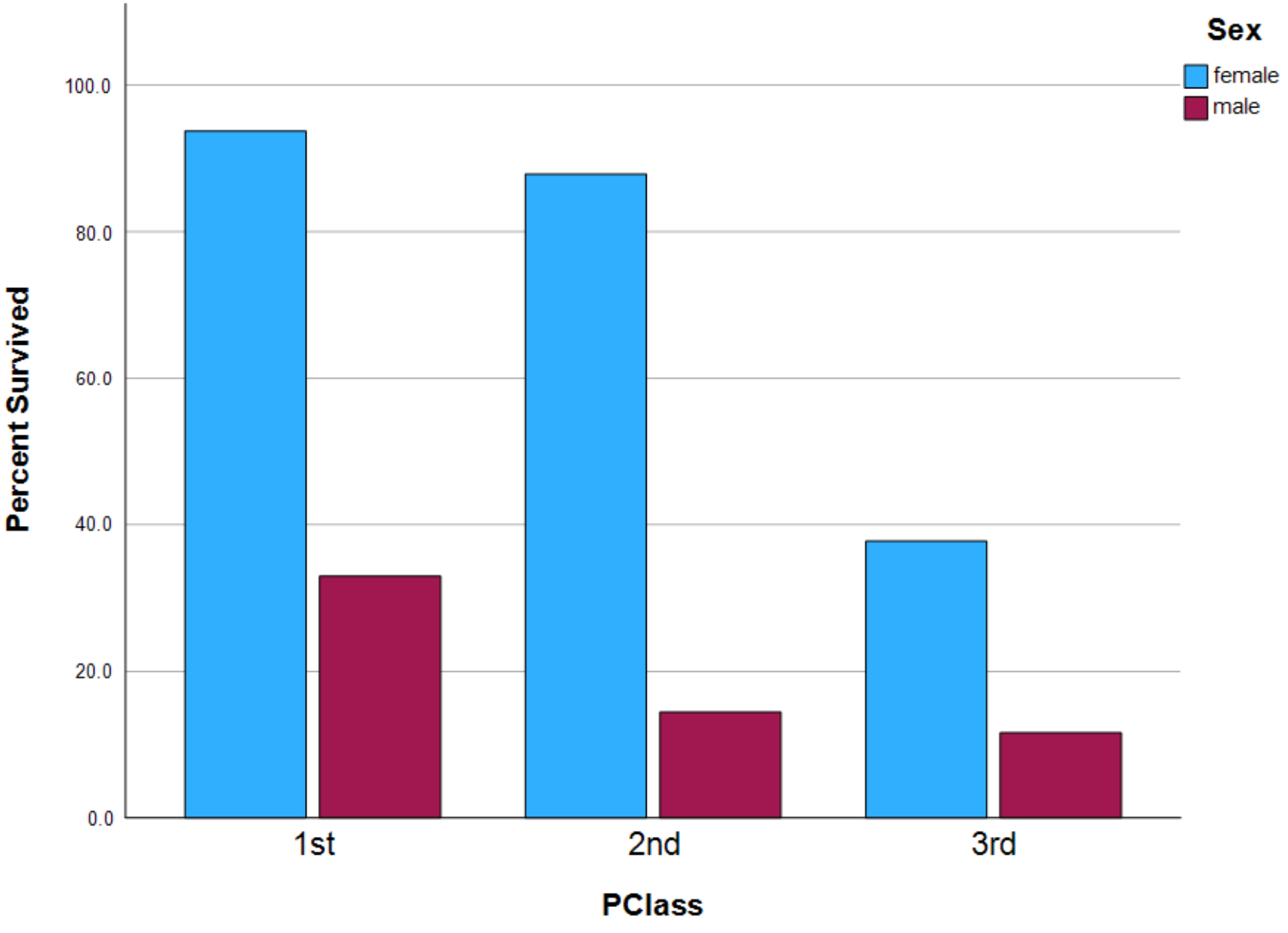
Figure 33: Bar chart showing proportion of passenger classes among deaths and survivors



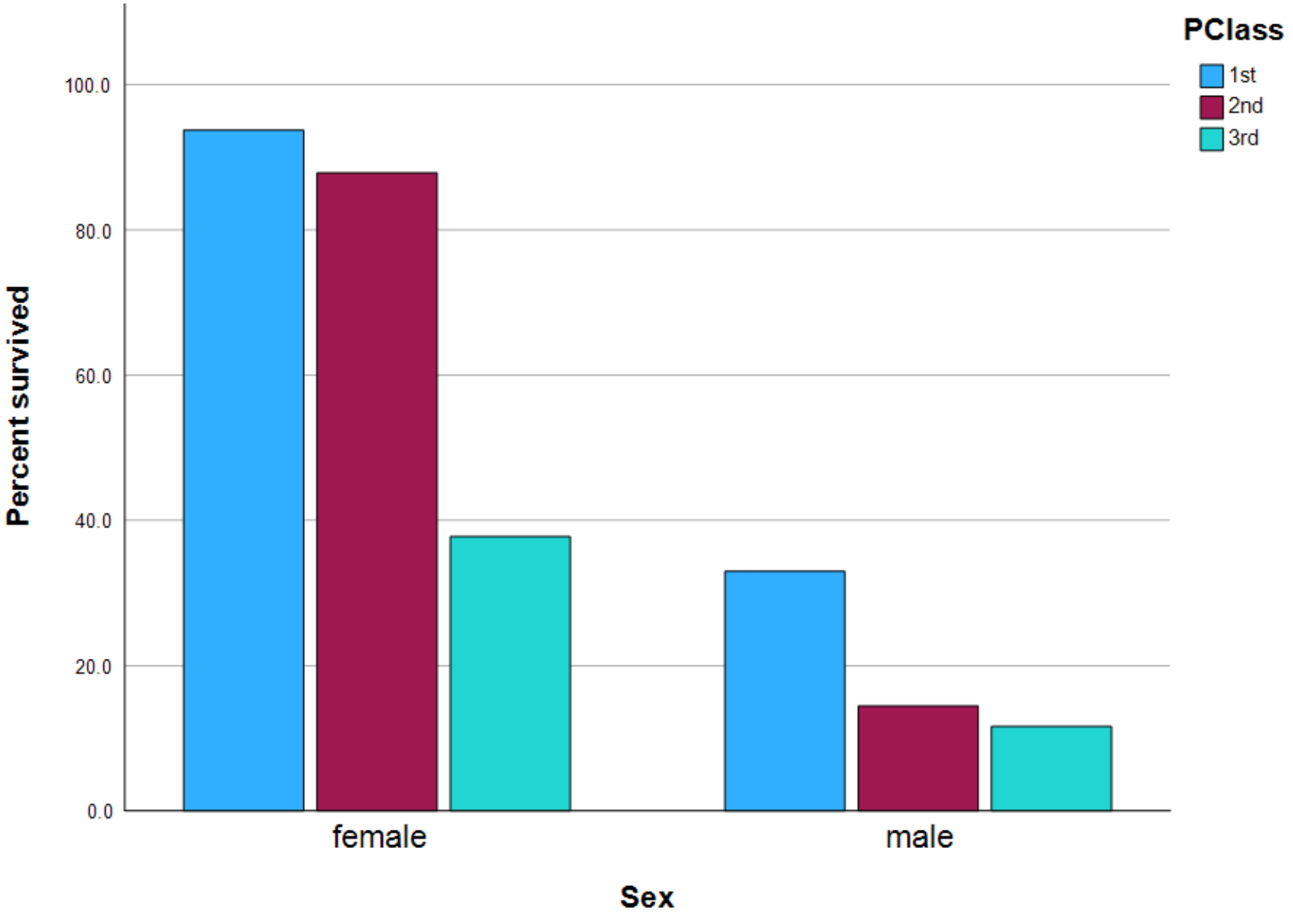
Bar chart showing proportion of deaths and survivors among passenger classes



Bar chart showing only proportion of survivors among passenger classes



Bar chart showing proportion of survivors among passenger classes and sex



Bar chart showing proportion of survivors among sex and passenger classes

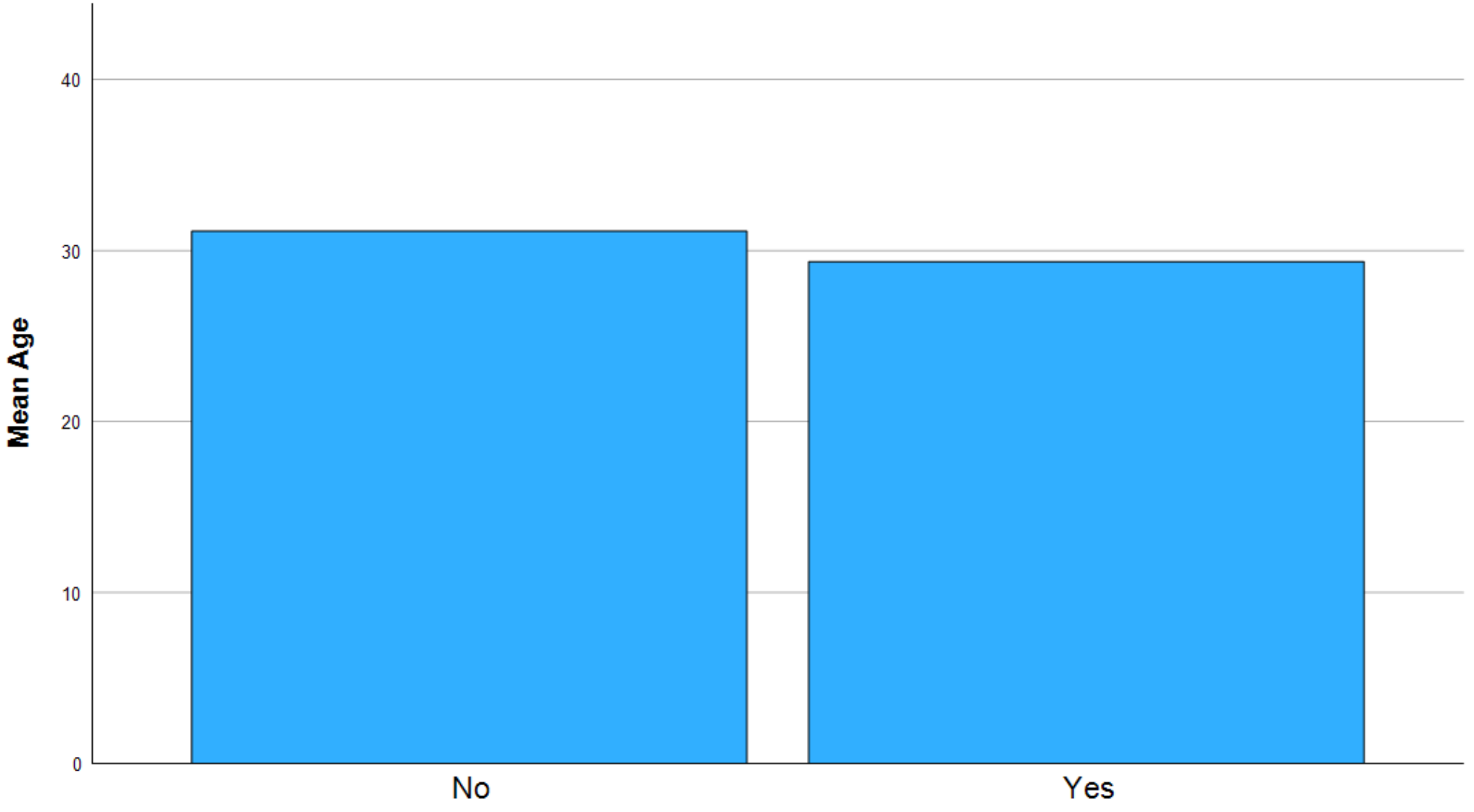
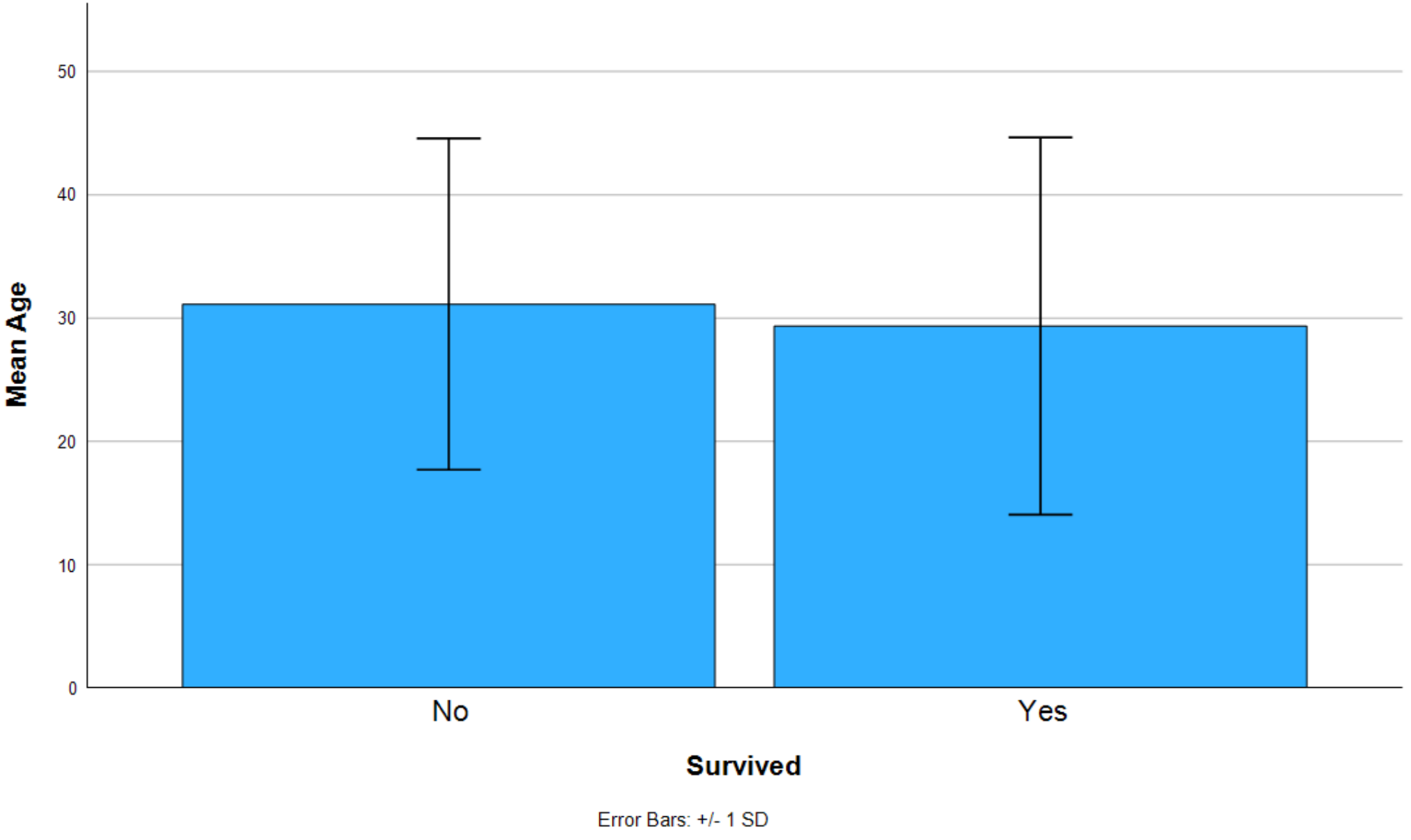


Figure 34: Bar chart showing average age among deaths and survivors



Bar chart with error bars showing proportion of survivors among sex and passenger classes

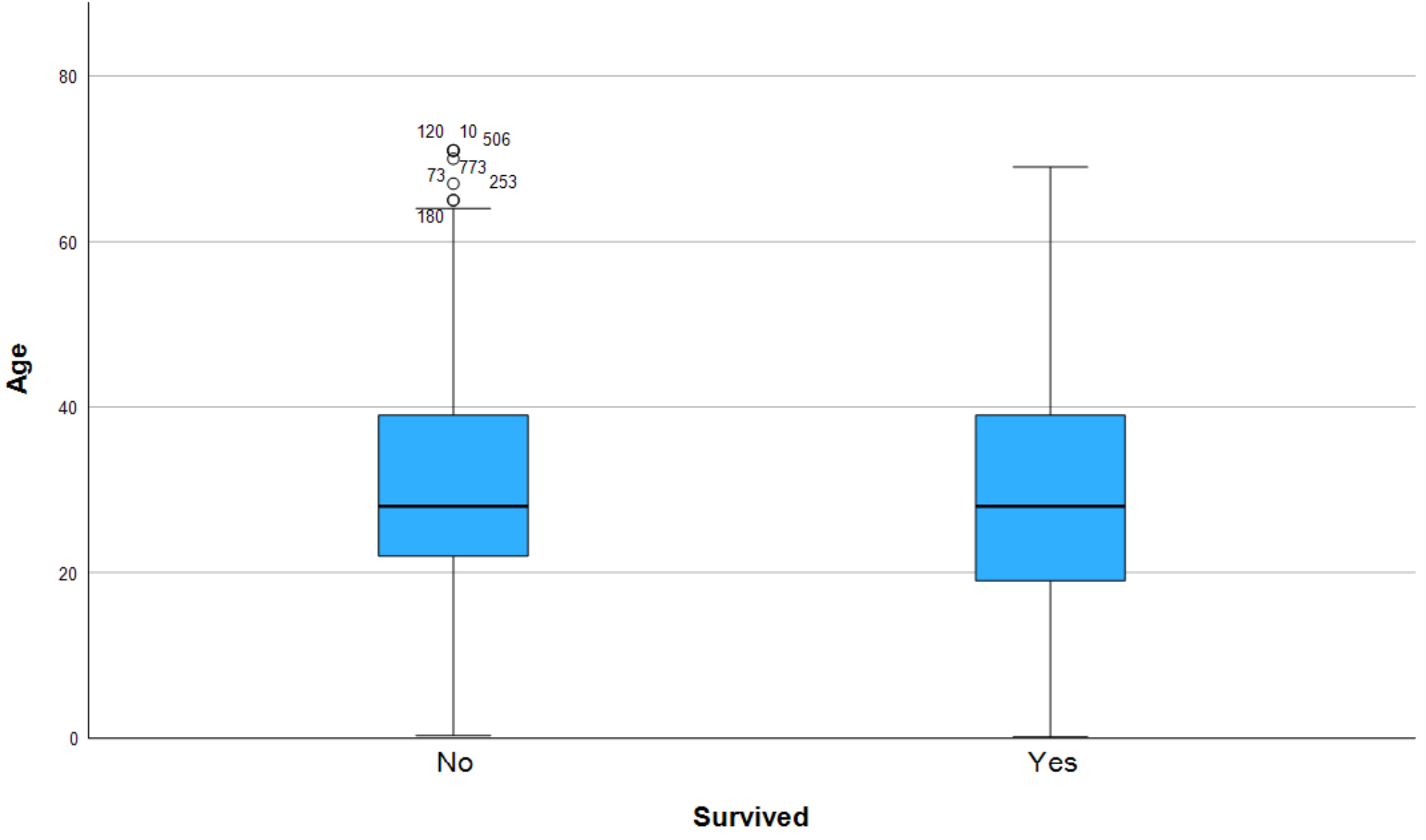


Figure 35: Boxplot showing ages of deaths and survivors

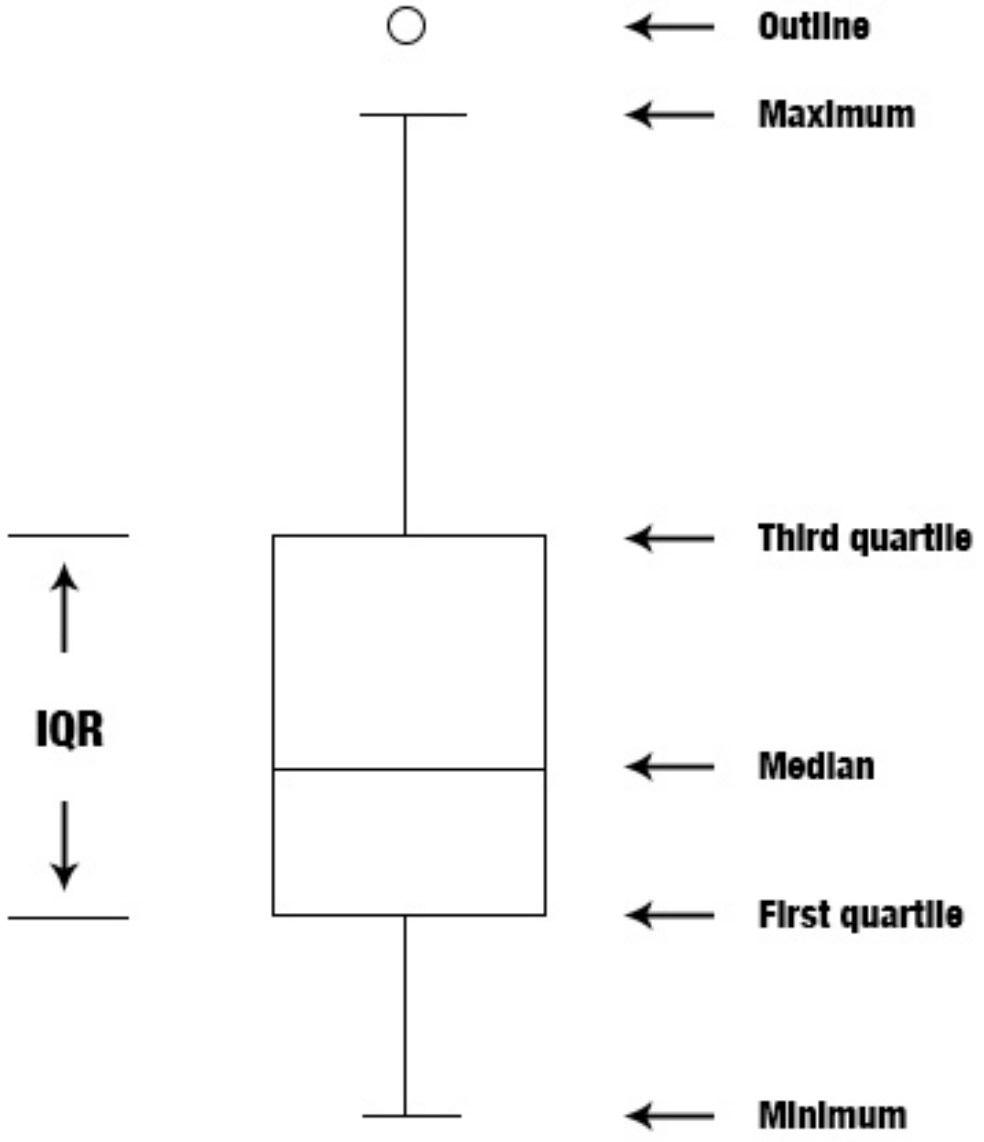


Figure 36: Annotated boxplot

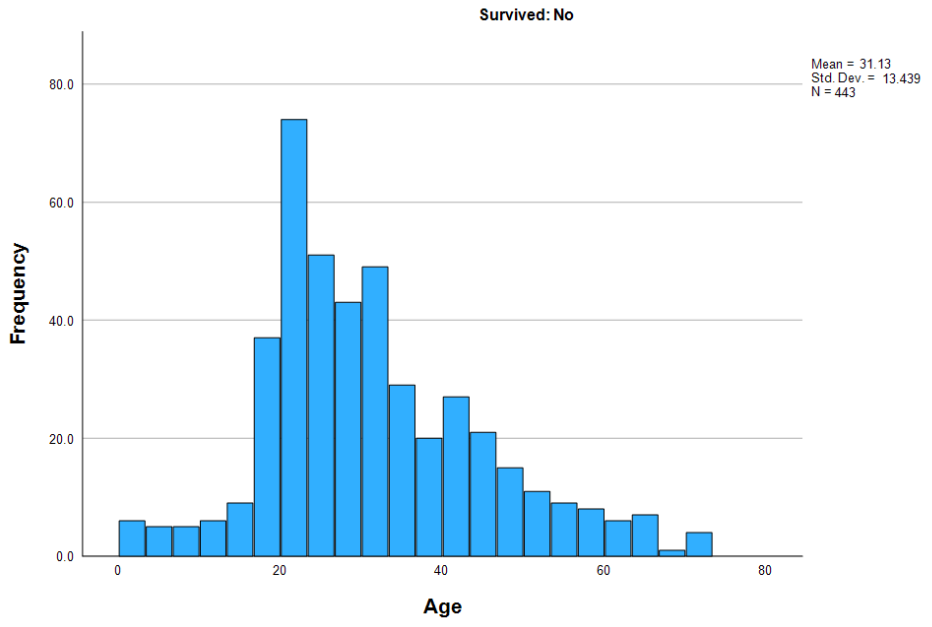


Figure 37: Histogram of ages of passengers that died

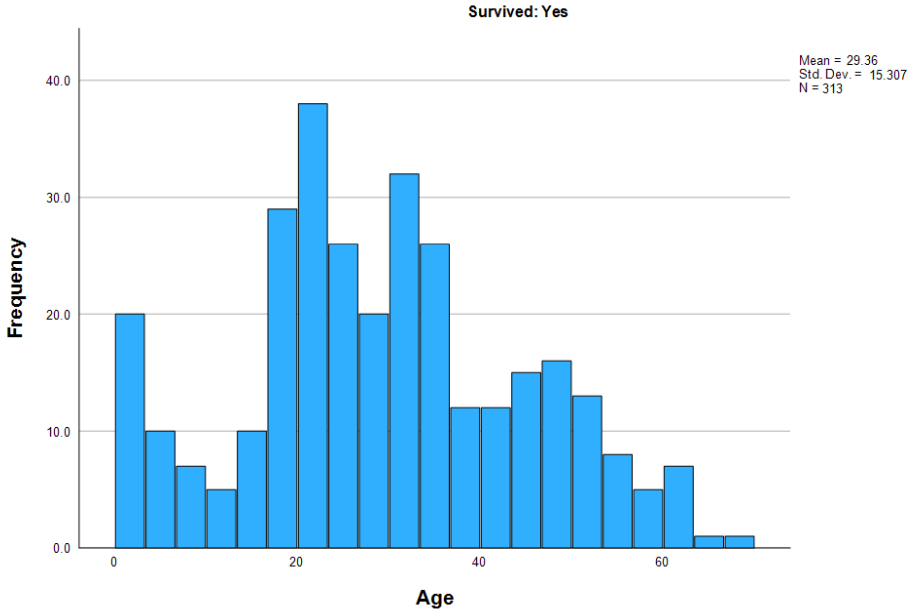


Figure 38: Histogram of ages of passengers that survived

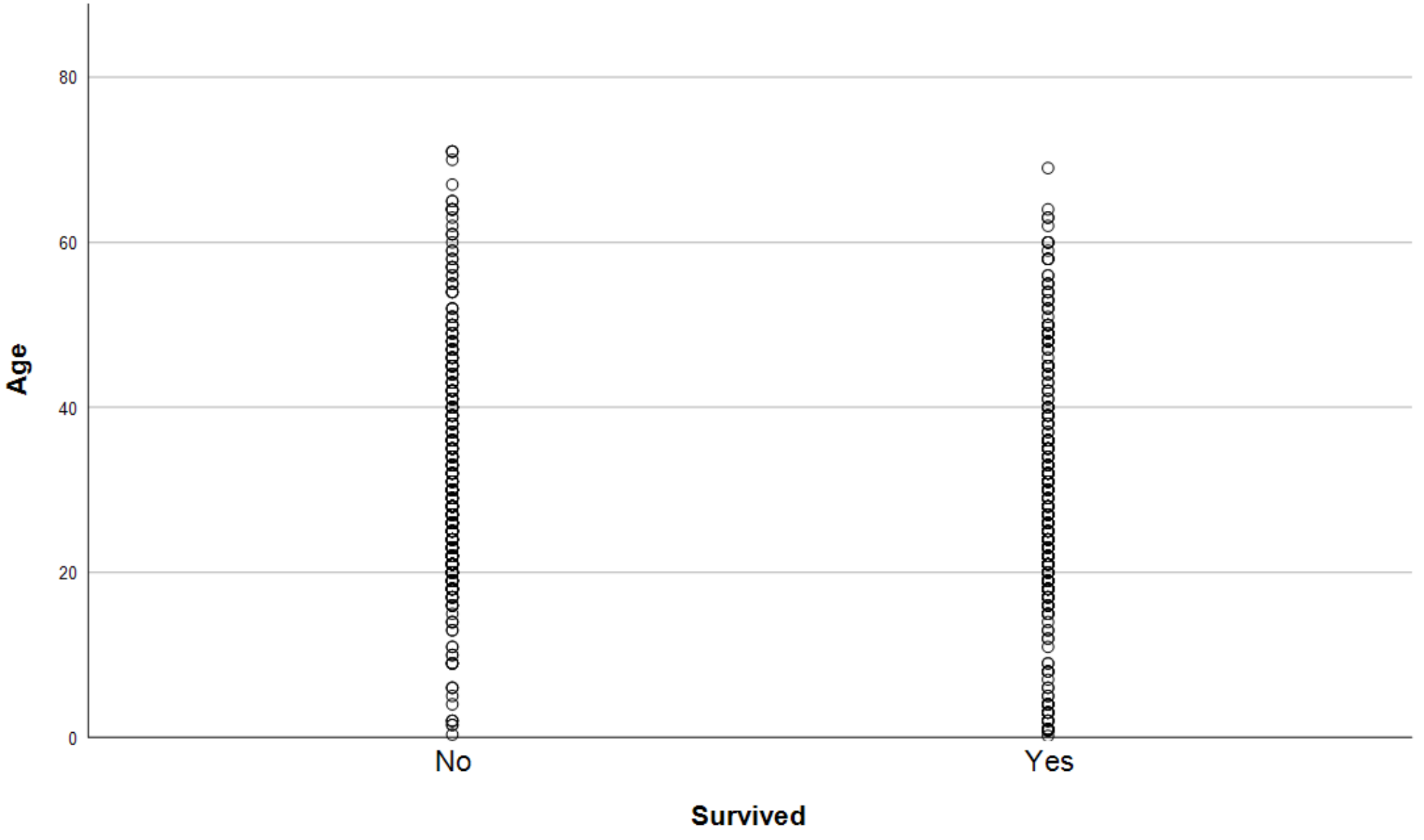


Figure 39: SCatter plot of age vs survived

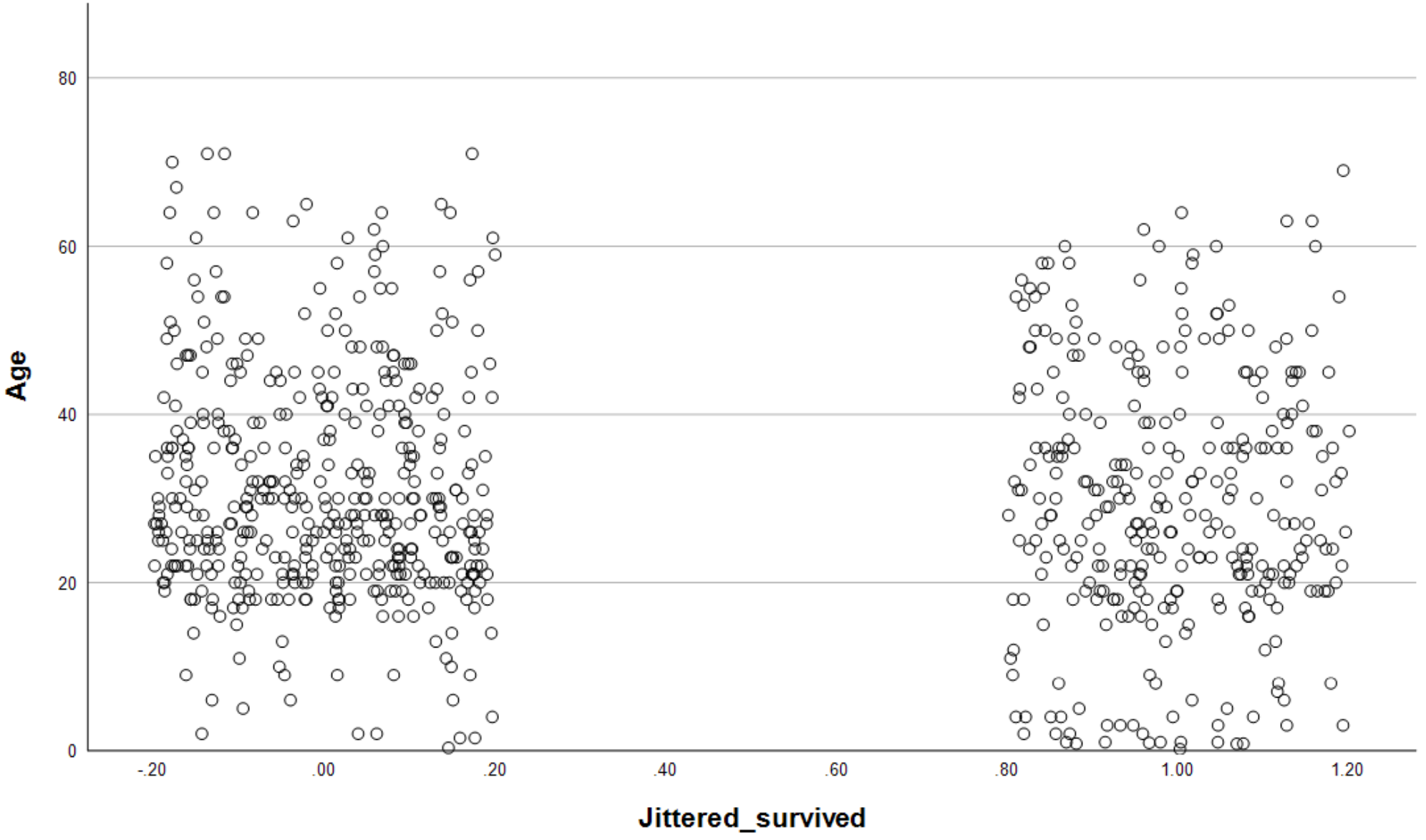


Figure 40: Scatter plot of age vs survived with jittering

Article feedback by type of article

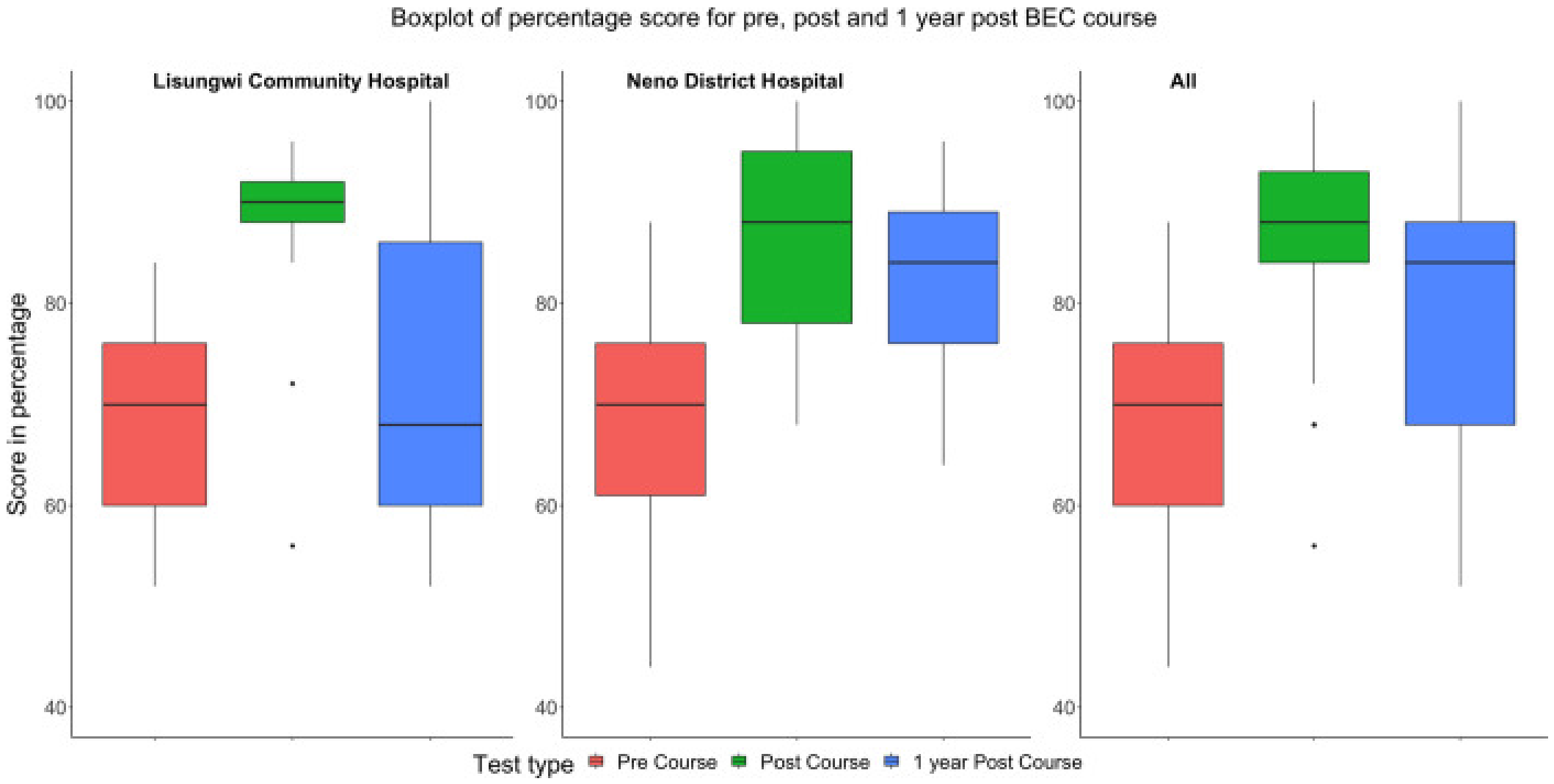
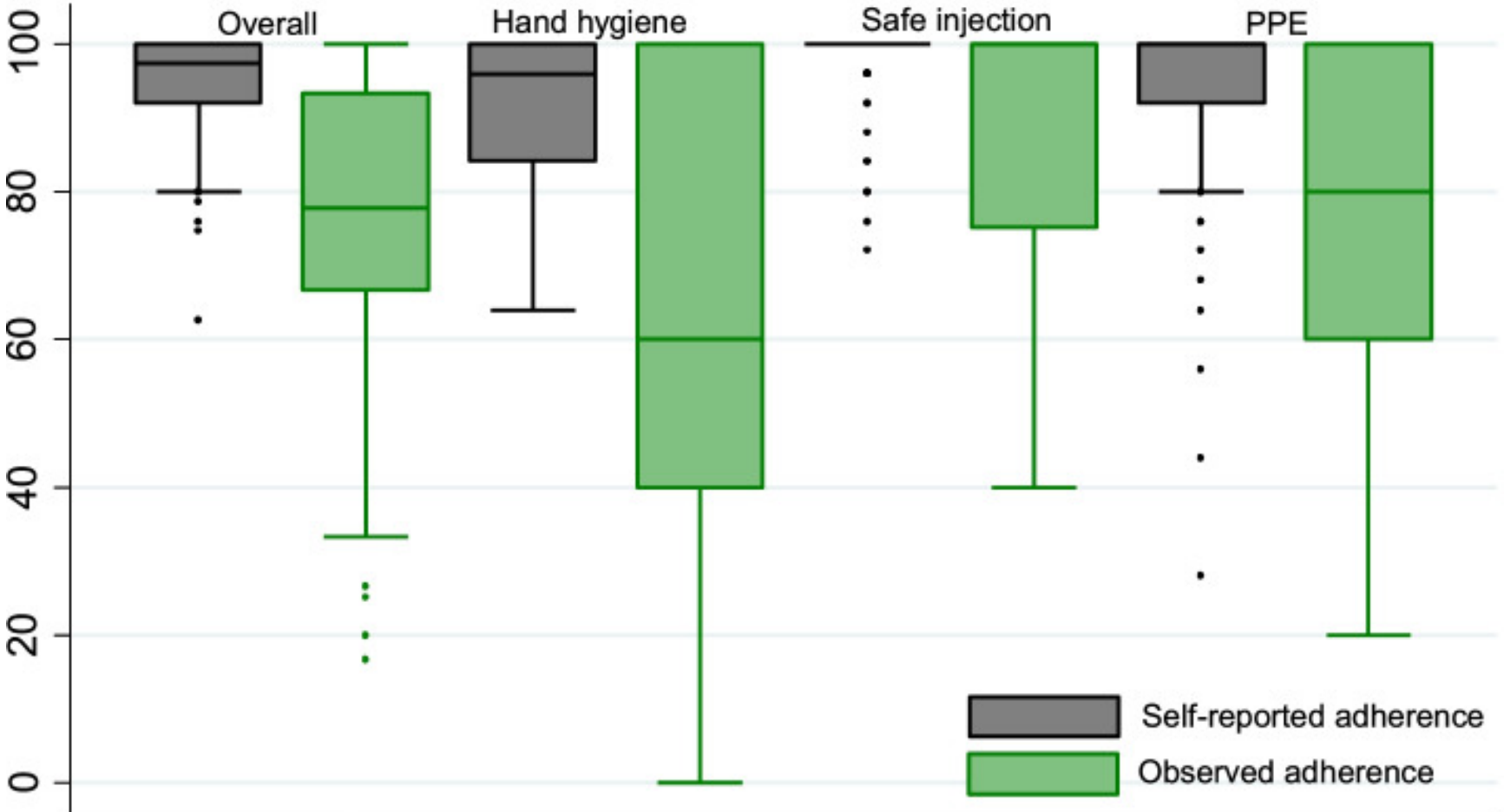


Figure 41: Emergency care course test scores



	Overall (Mean±S.D)	Subcategory (Mean±S.D)		
		Hand hygiene	Safe injection	PPE
Self-report	94.9±6.5	92.3±8.6	98.1±5.1	94.3±10.8
Observation	76.8±19.6	68.2±30.5	90.2±15.7	70.0±24.5

Figure 42: Self reported versus observed adherence

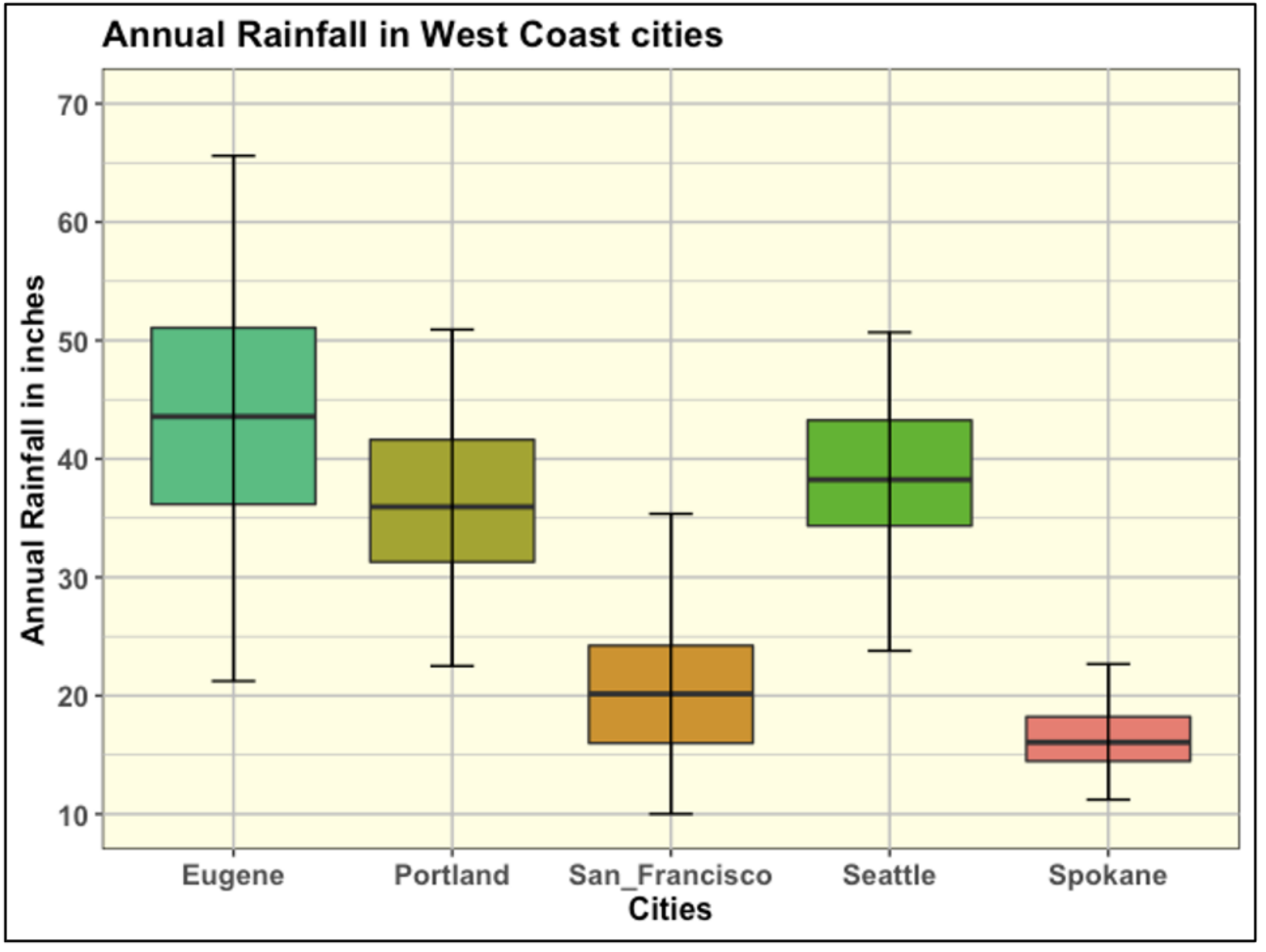


Figure 43: Annual rainfall in selected cities

Which visualization to choose?

- Categorical data
  - Use a table
  - Minimize distance of key comparisons
- Continuous data
  - Small datasets: plot all the data
  - Large datasets: boxplots

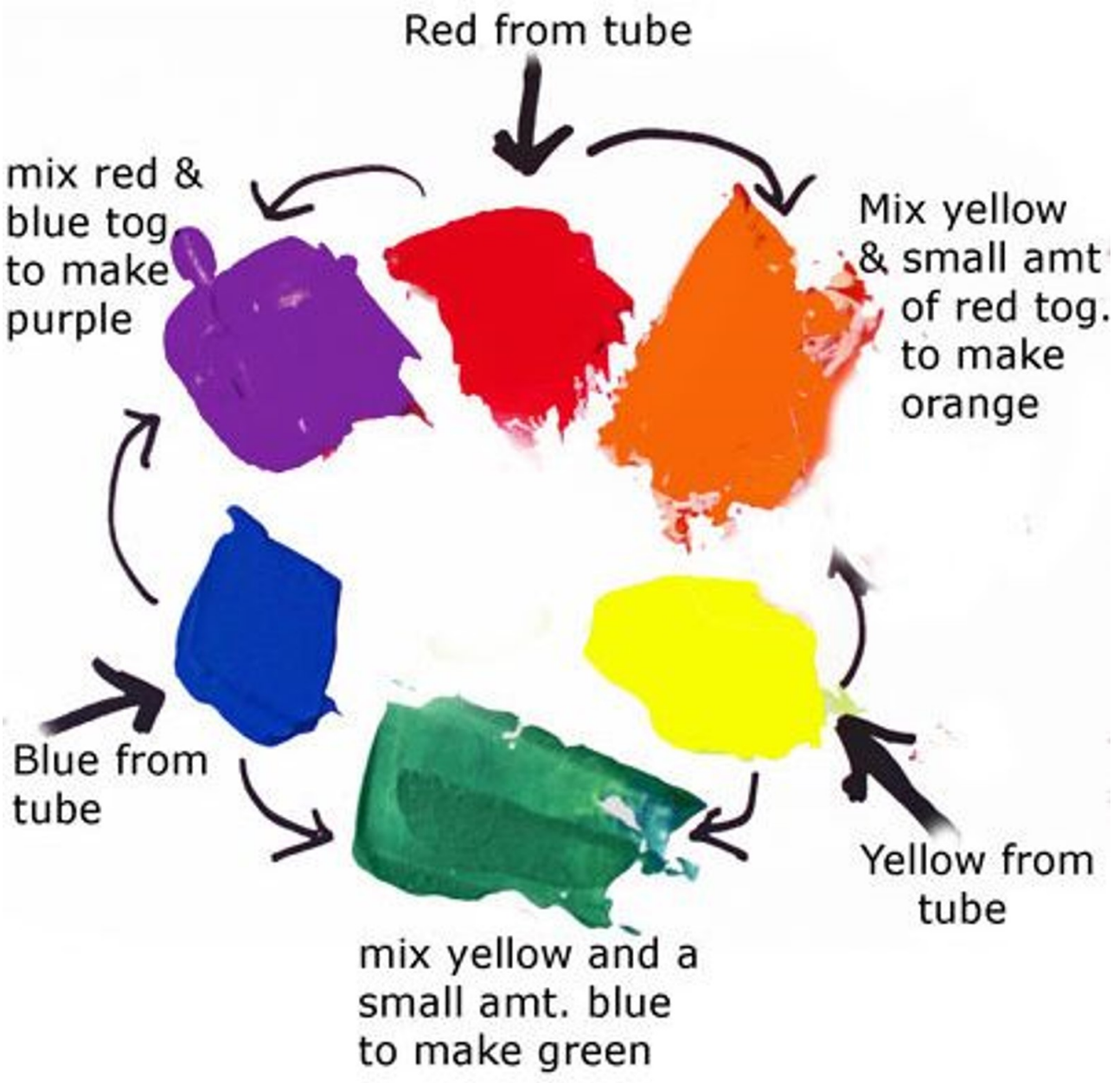


Figure 44: Color combinations

The RGB color system

- #rrggbb format
  - #000000 is pure black
  - #FFFFFF is pure white
  - #FF0000 is pure red
  - #00FF00 is pure green
  - #0000FF is pure blue
- You can mix and match to get 16,777,216 colors
  - #800080 is purple, #FF69B4 is pink, #40E0D0 is turquoise

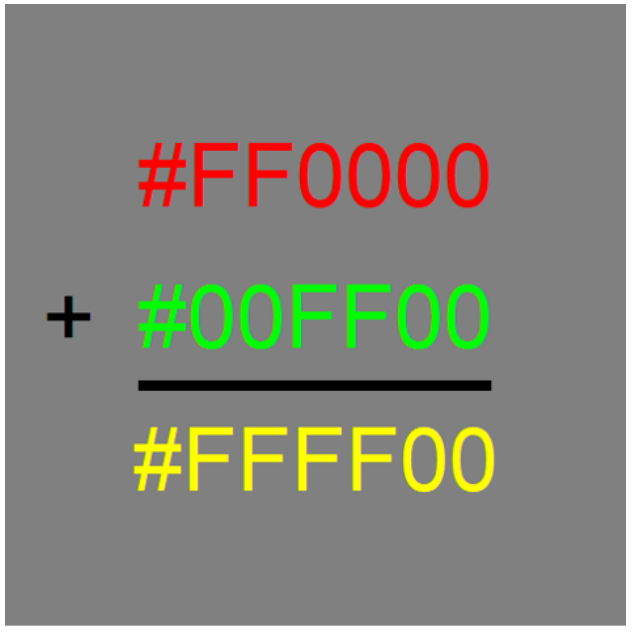


Figure 45: Red plus green equals yellow

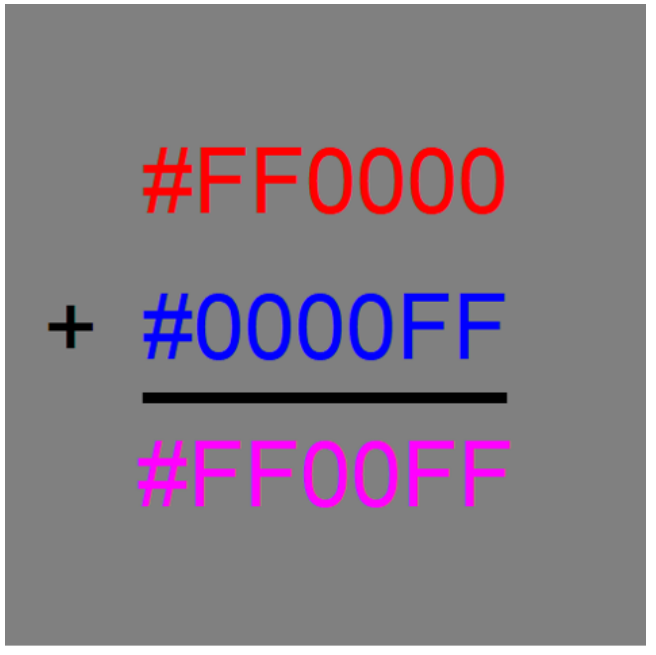


Figure 46: Red plus blue equals magenta

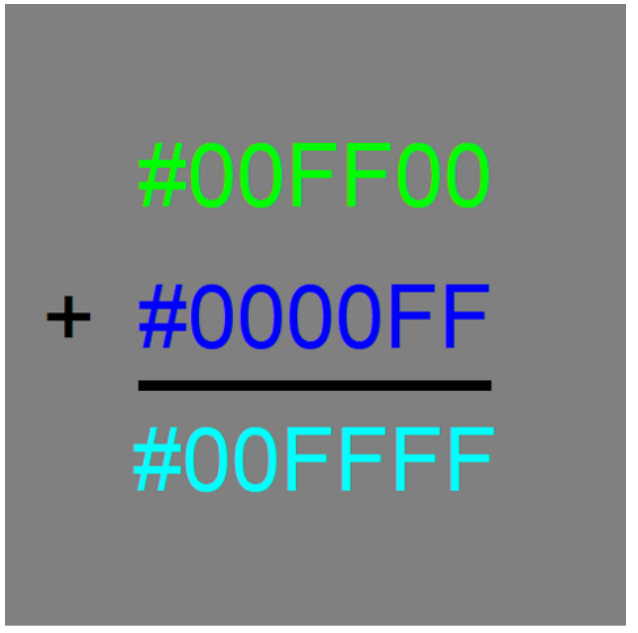


Figure 47: Green plus blue equals cyan

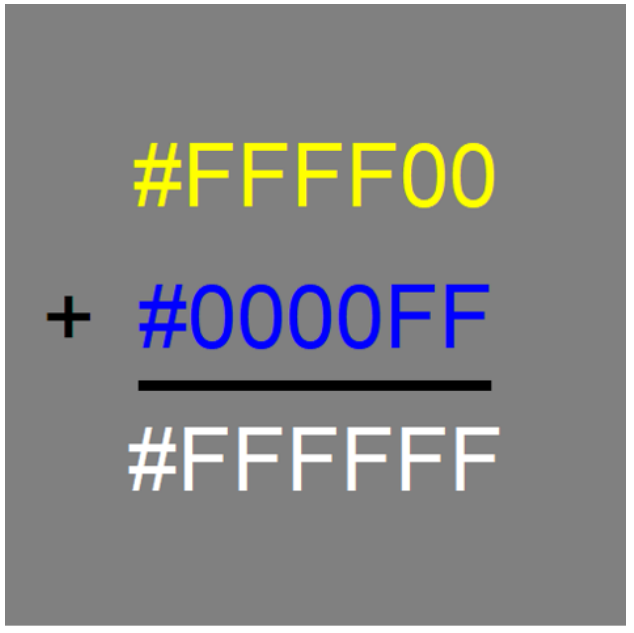


Figure 48: Yellow plus blue equals white

Figure 49: Magenta plus green equals white

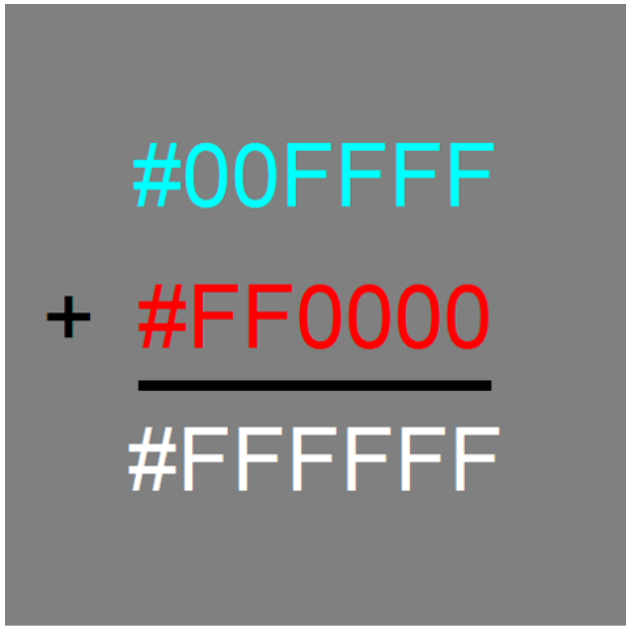


Figure 50: Cyan plus red equals white

The color cube

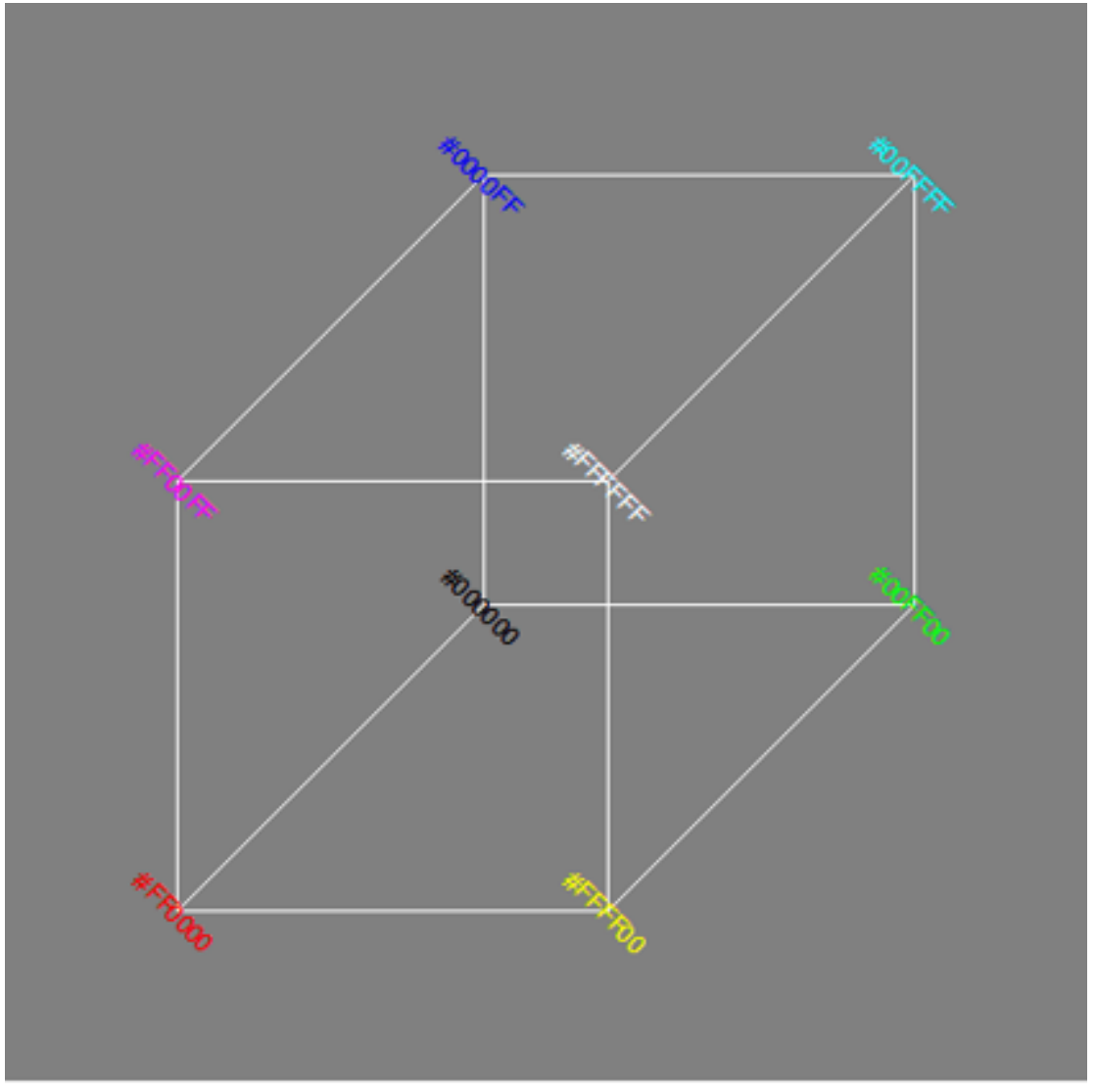


Figure 51: Illustration of the color cube

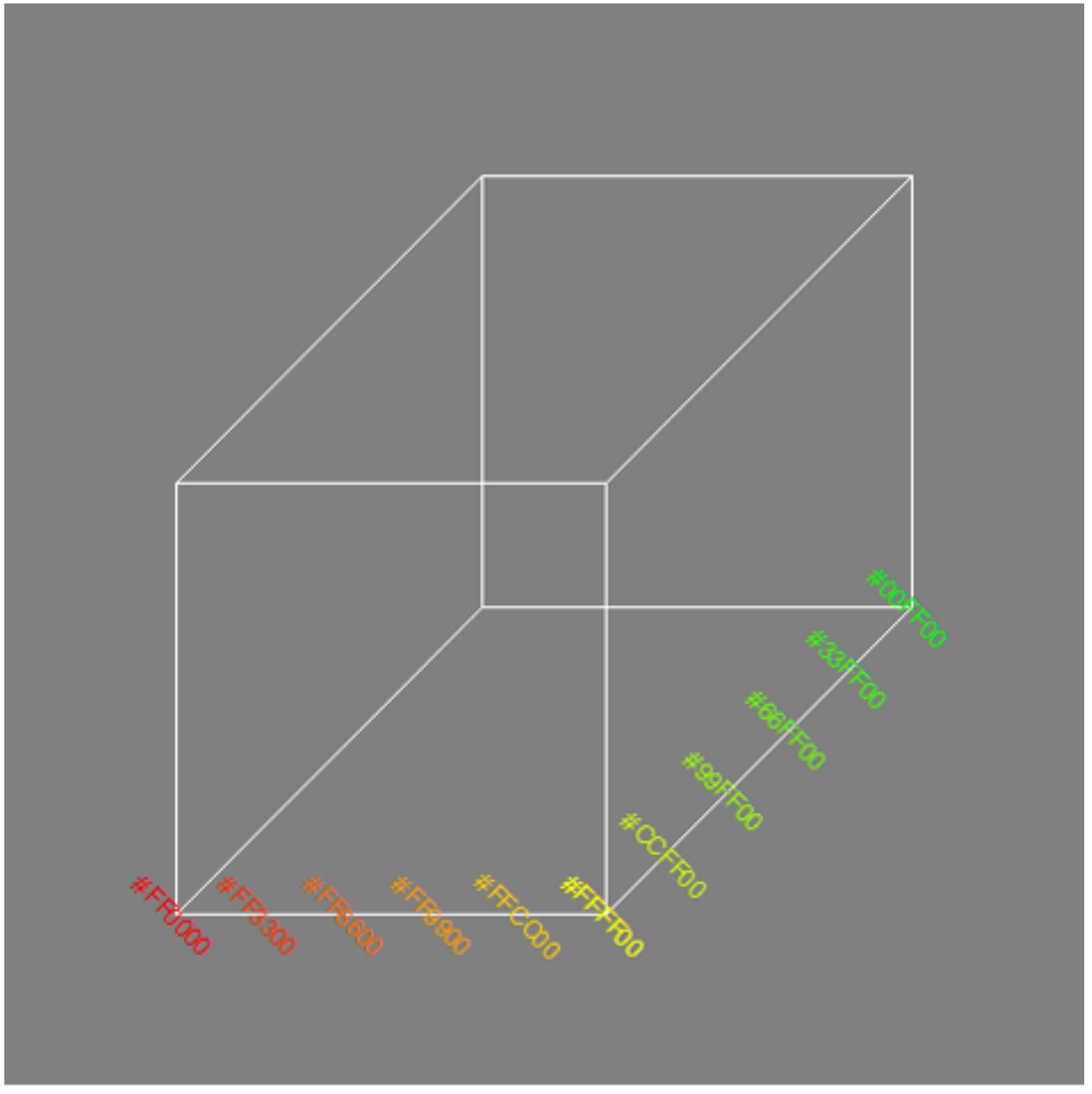
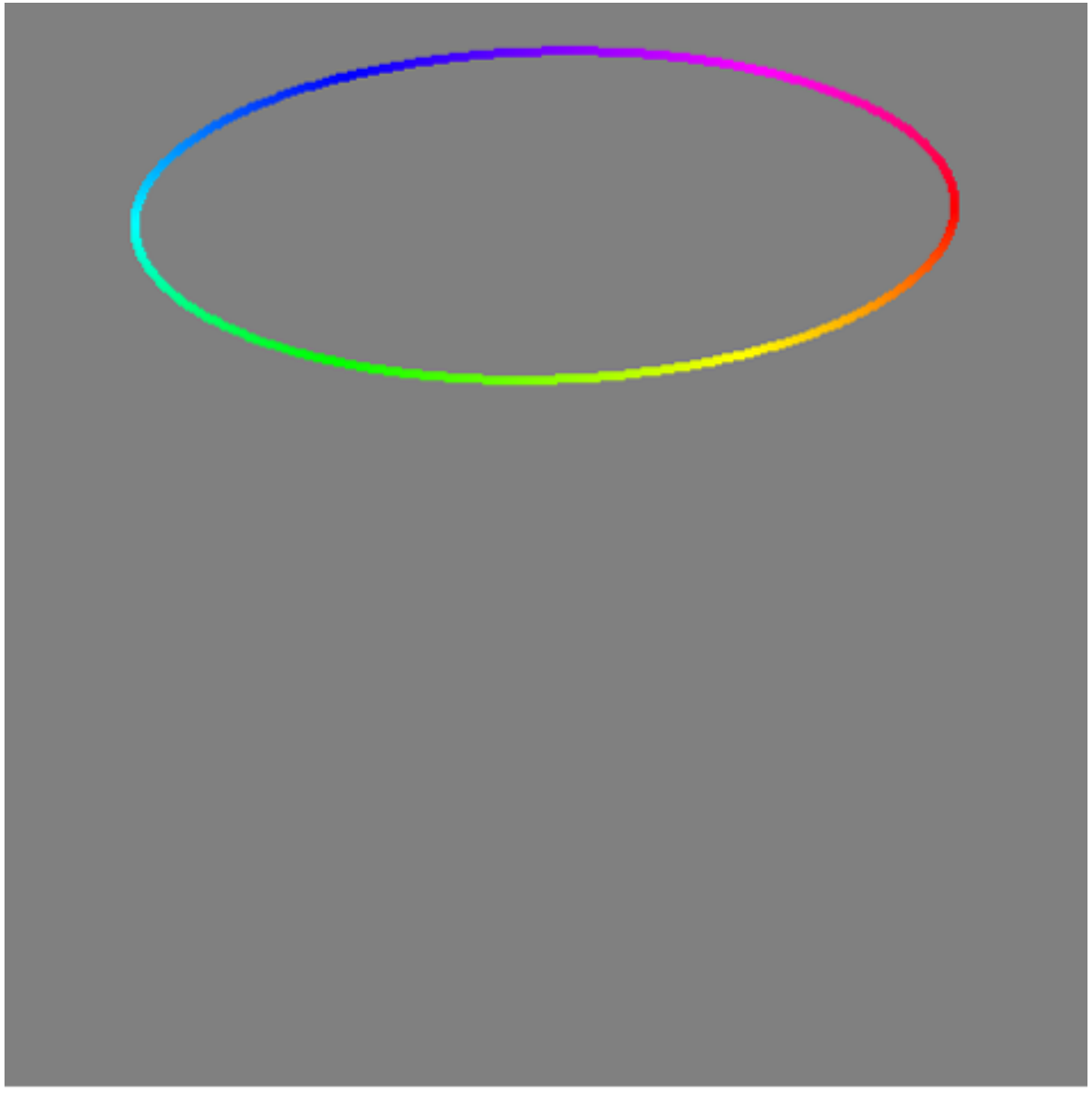


Figure 52: The red to green gradient on the color cube

Rainbow



The color cylinder

Figure 53: Color cylinder

Figure 54: Various foreground and background color combinations

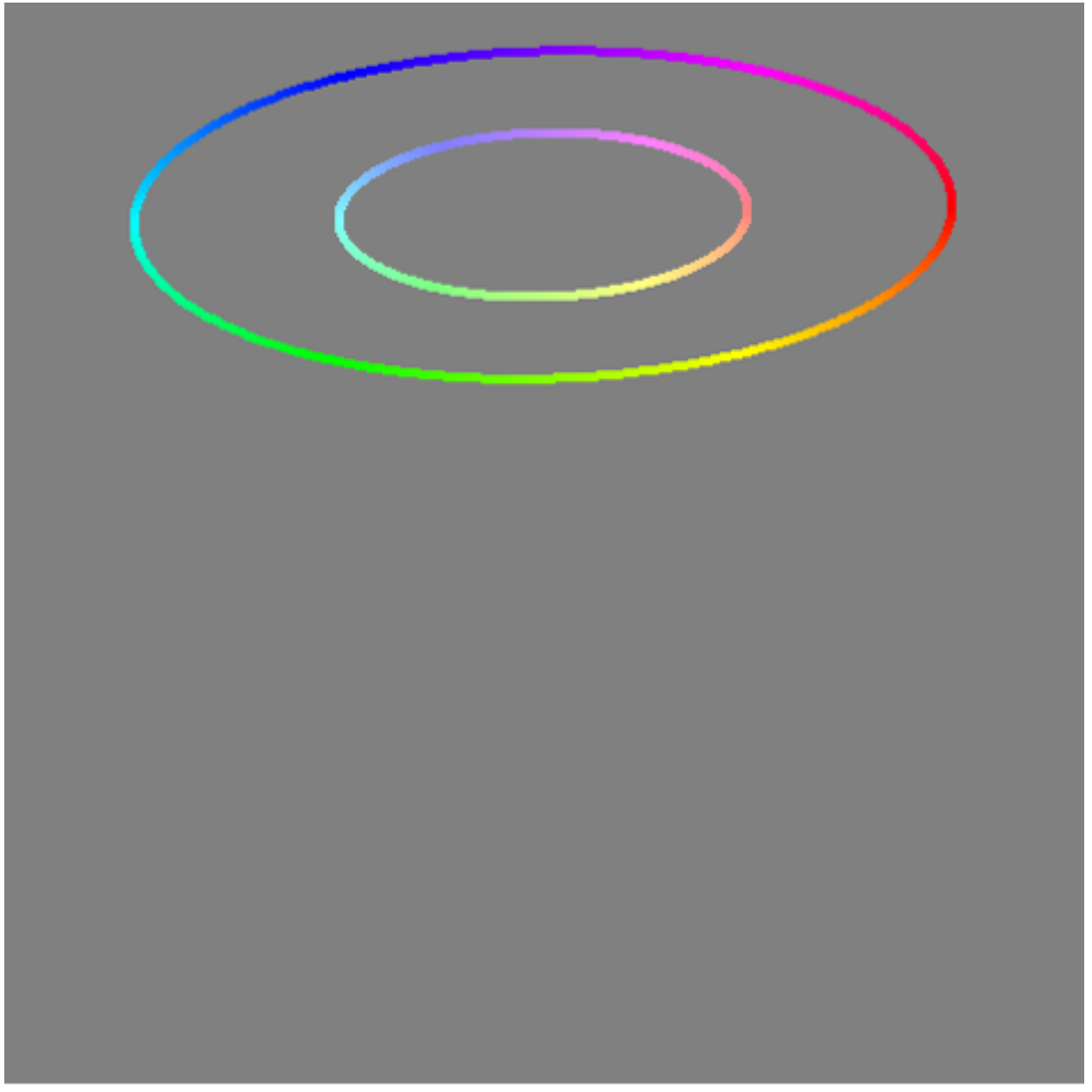


Figure 55: A brighter version of the rainbow

Darker rainbow

Figure 56: A darker version of the rainbow

Figure 57: Color combinations using darker foregrounds and lighter backgrounds

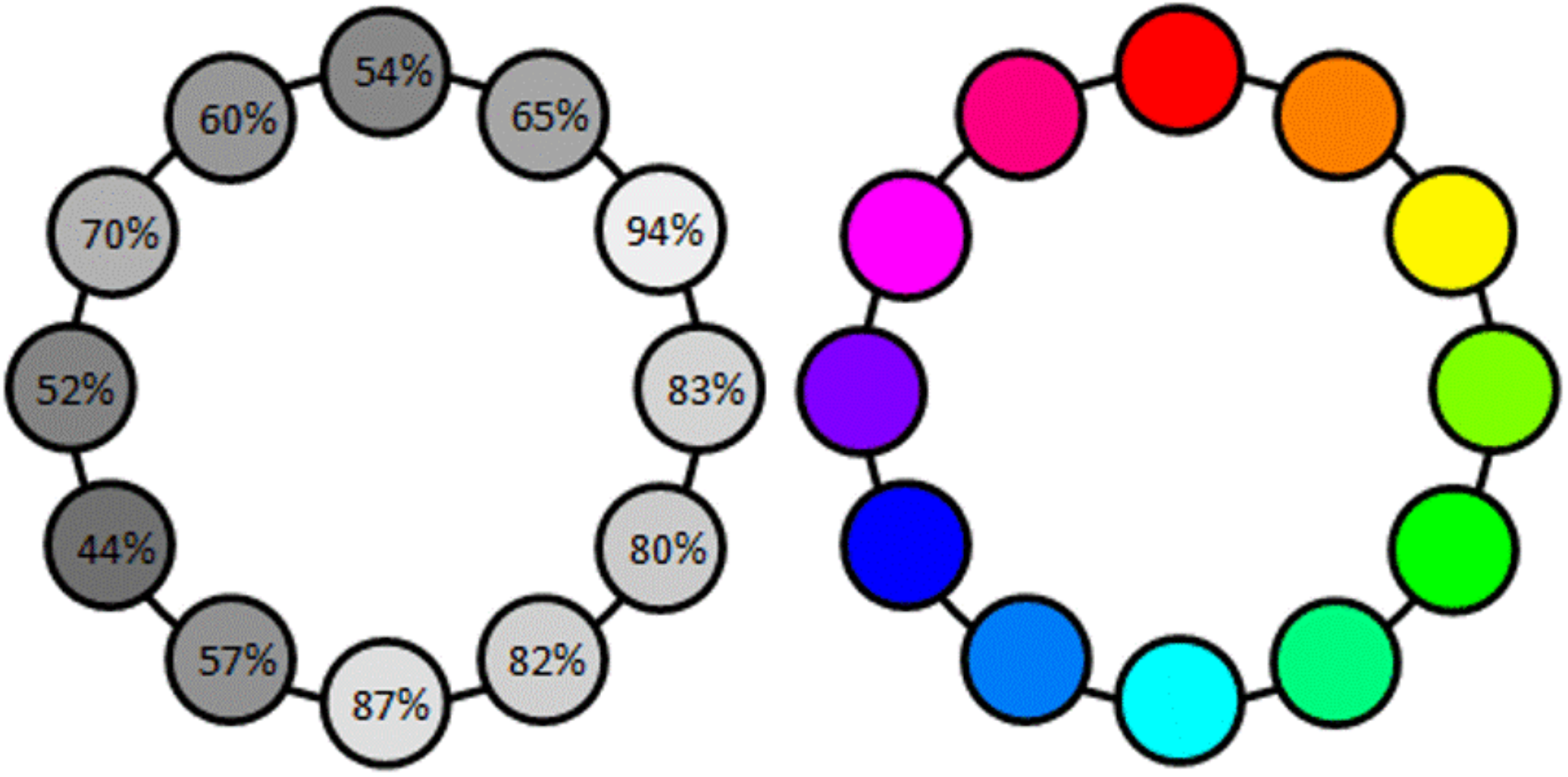


Figure 58: DIffering luminance values of the rainbow

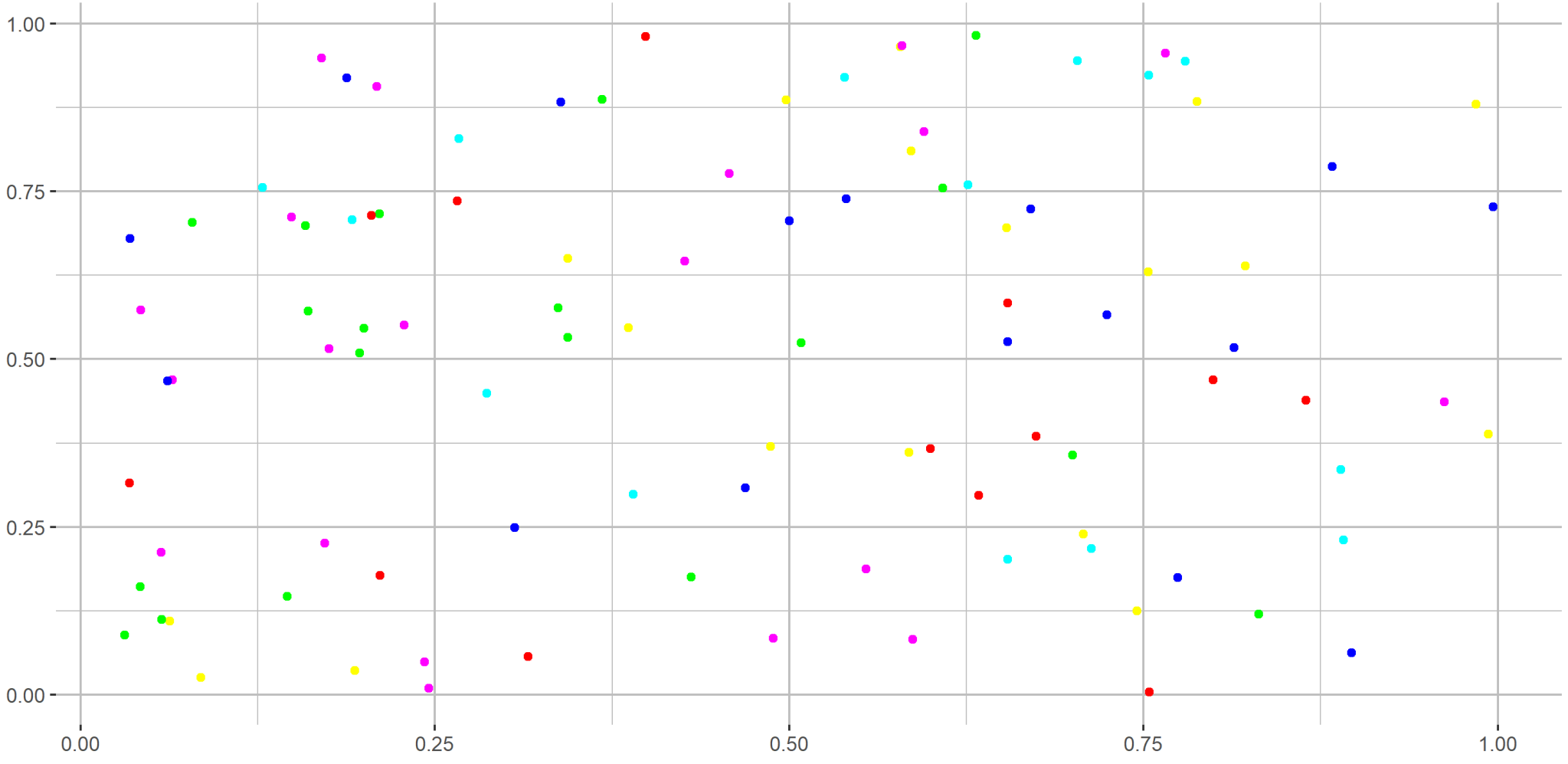


Figure 59: Rainbow colors on a white background

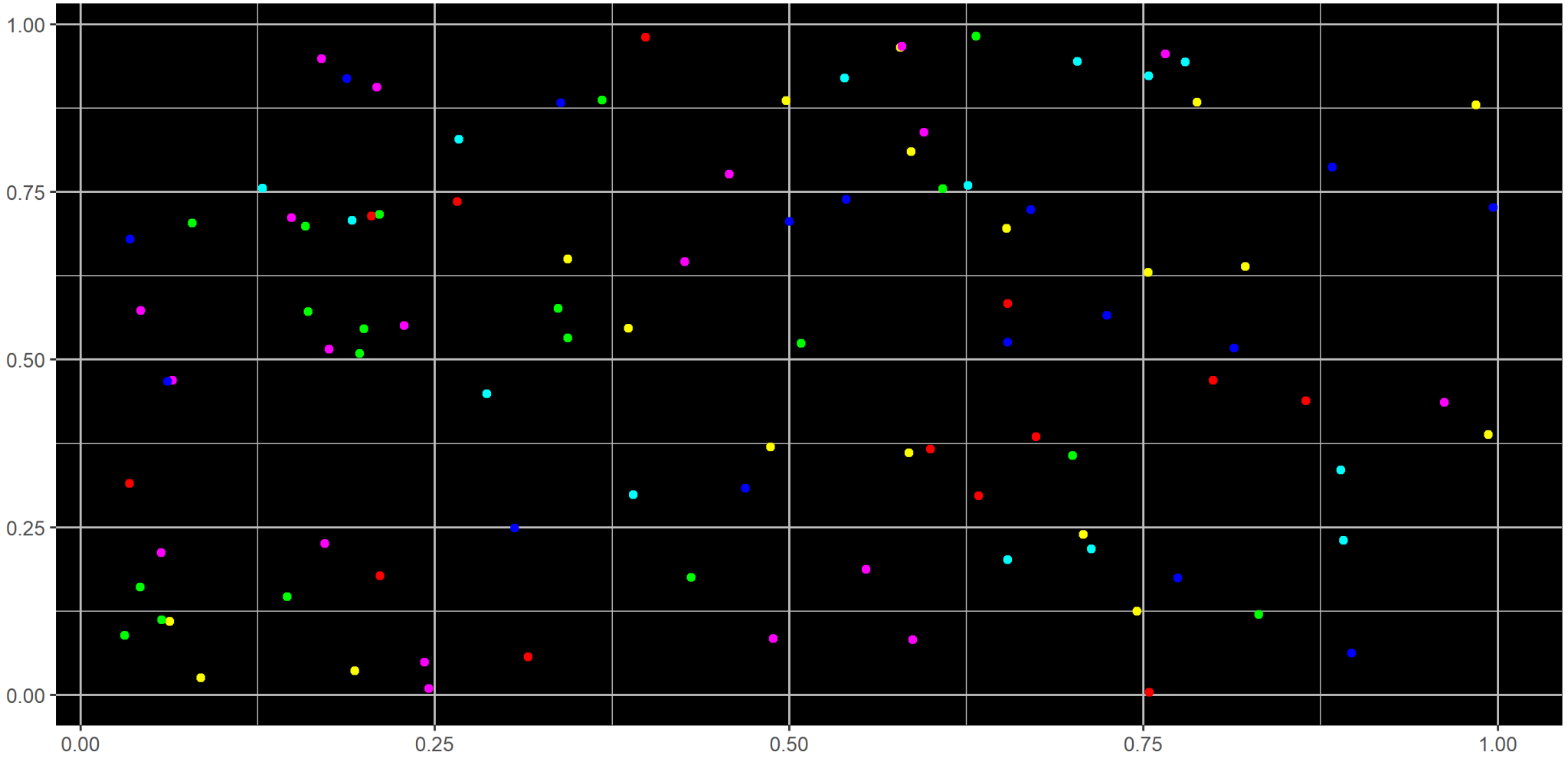


Figure 60: Rainbow colors on a black background

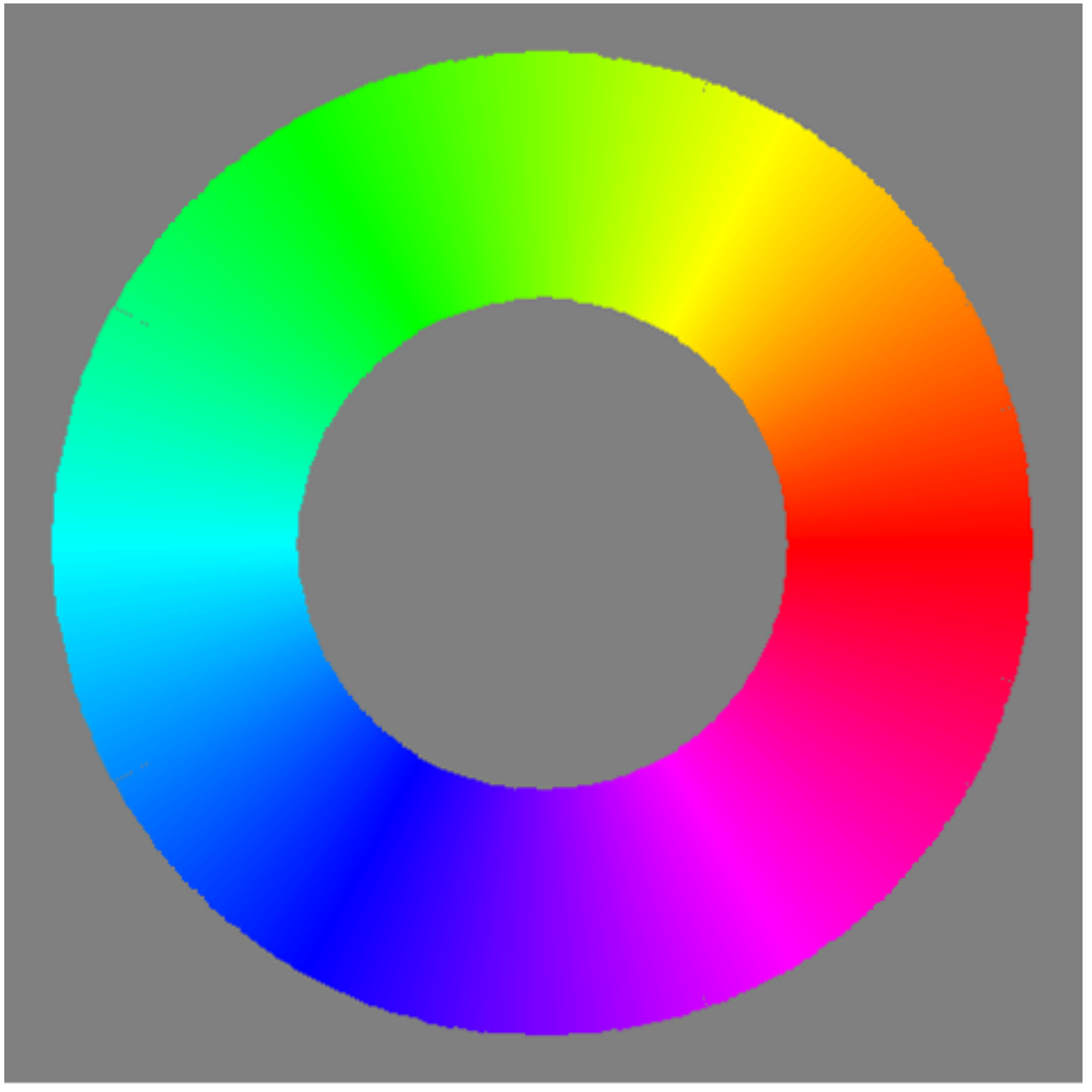


Figure 61: Rainbow colors showing a banding effect

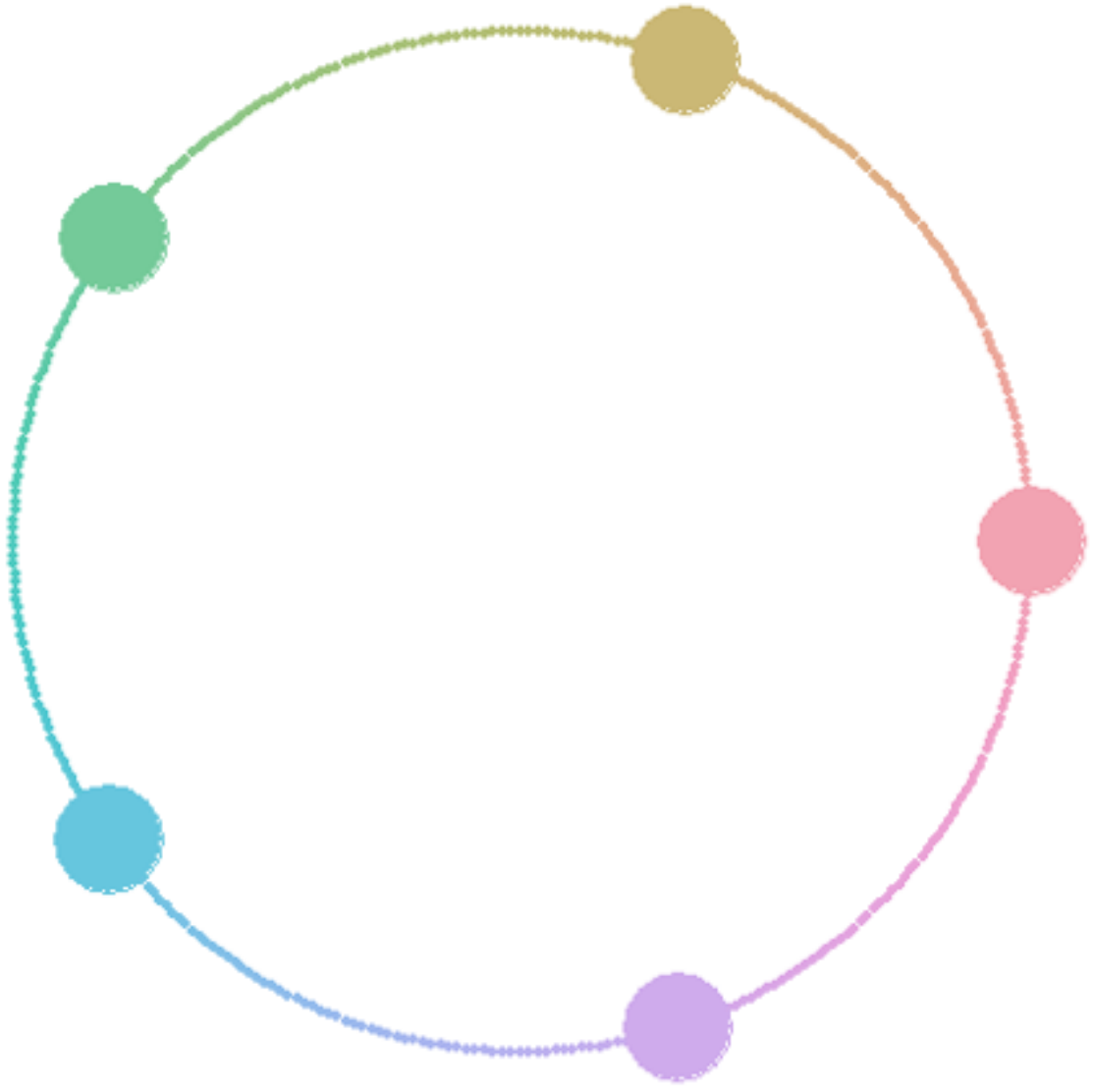
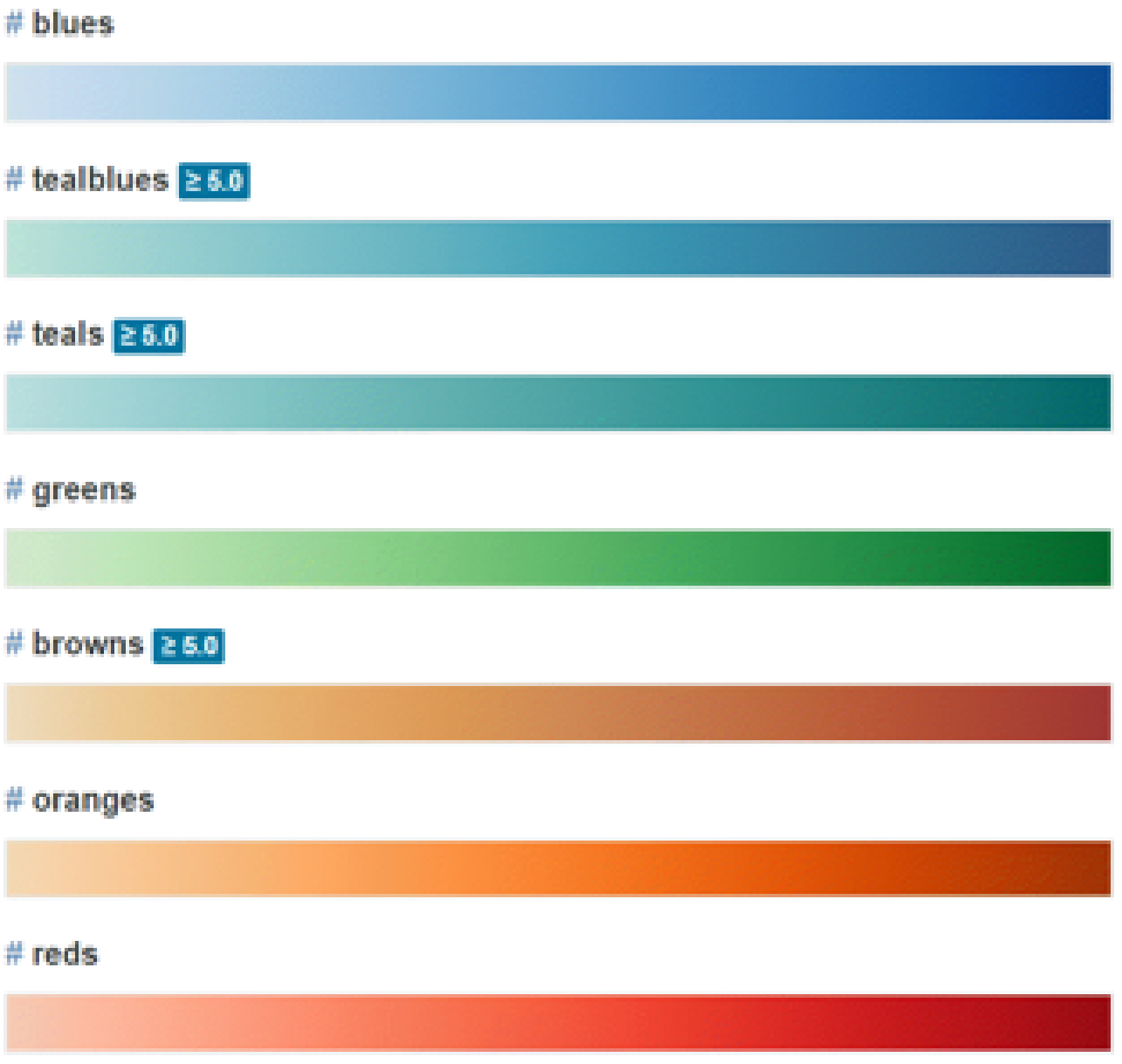


Figure 62: One good set of color choices for nominal data



Examples of light to dark gradients

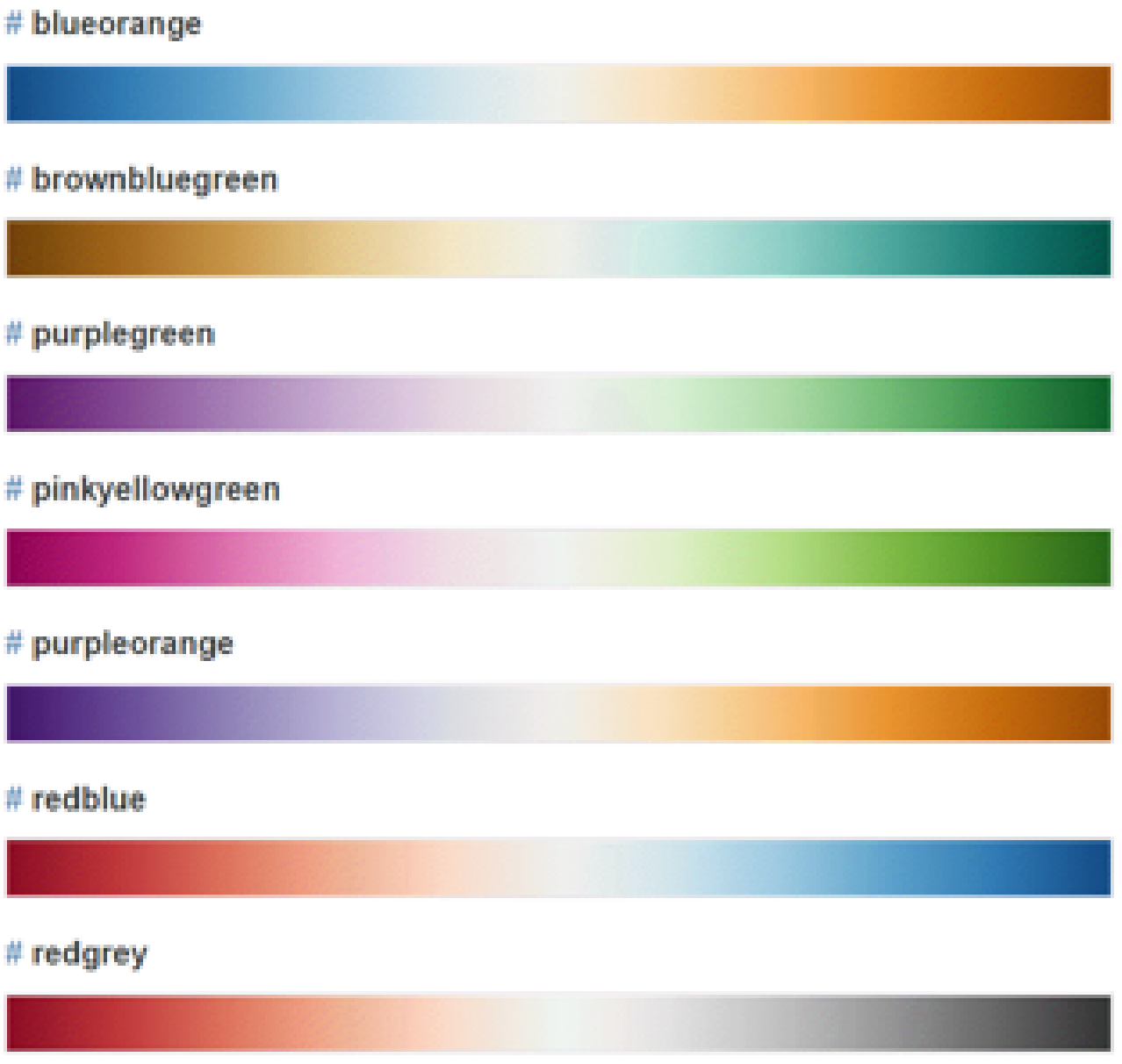


Figure 63: Examples of diverging gradients

Color blindness

- Up to 10% of your audience is color blind
  - Most common: red-green
- Suggestions
  - Use alternate cues (shape, shading)
  - Test your image
  - Find color blind friendly palettes.

#4 TOO MANY COLORS

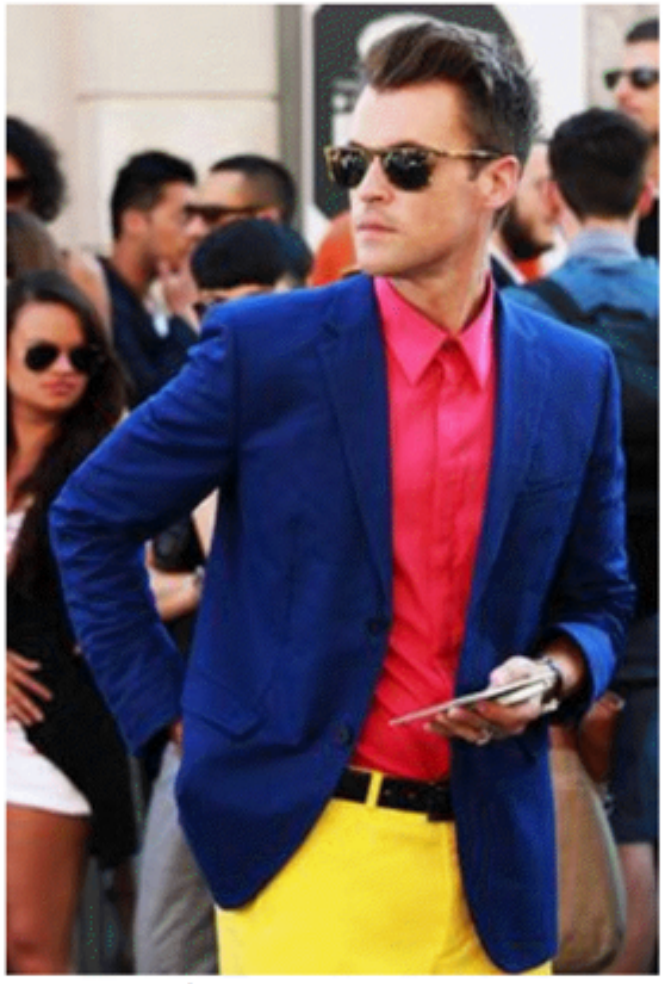


photo source: pinterest.com

Honestly, we find the best application for this quote in fashion: *"simplicity is the ultimate form of sophistication"*. Keeping it simple might be very difficult for several men, and, looking to the picture above, it really is.

Figure 64: Clothing mistake: using too many colors

ADvertisement with a single red umbrella

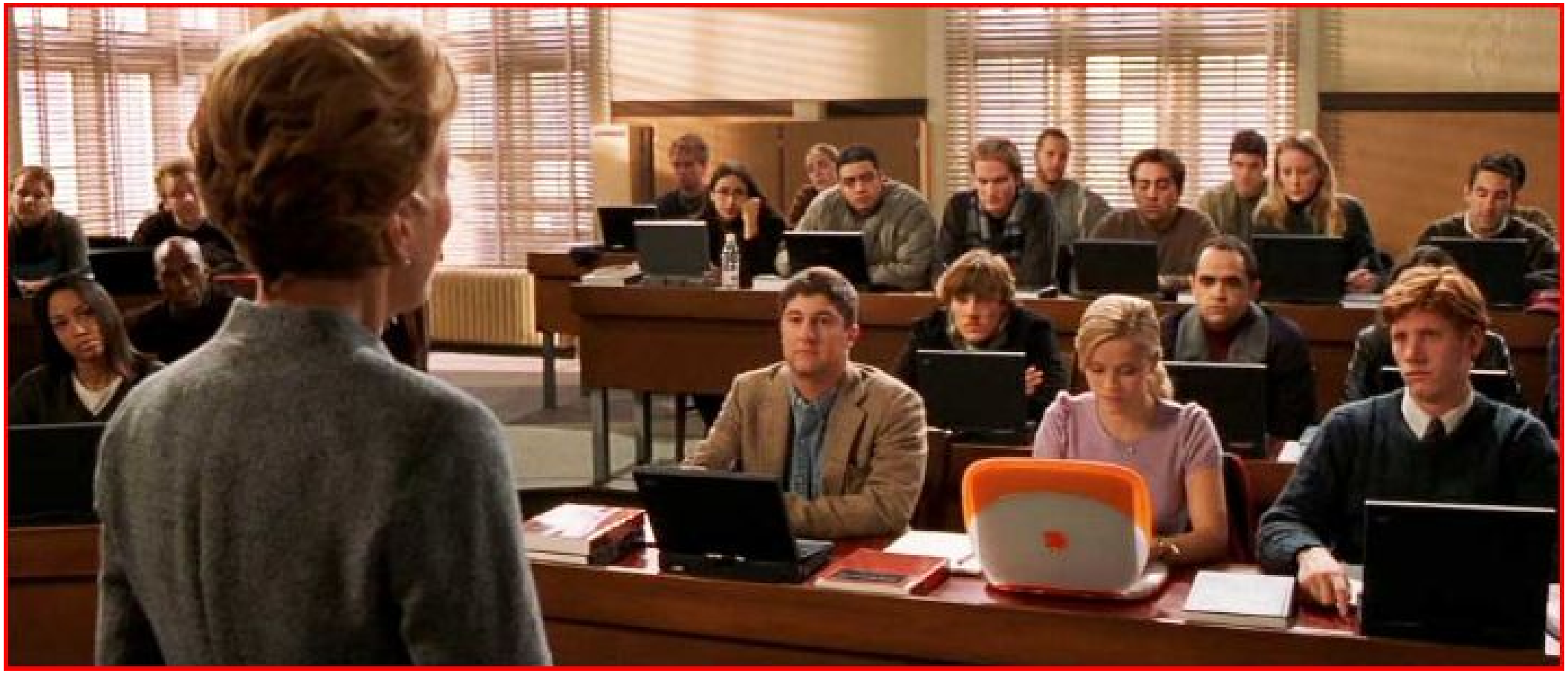


Figure 65: Use of color to highlight a single individual

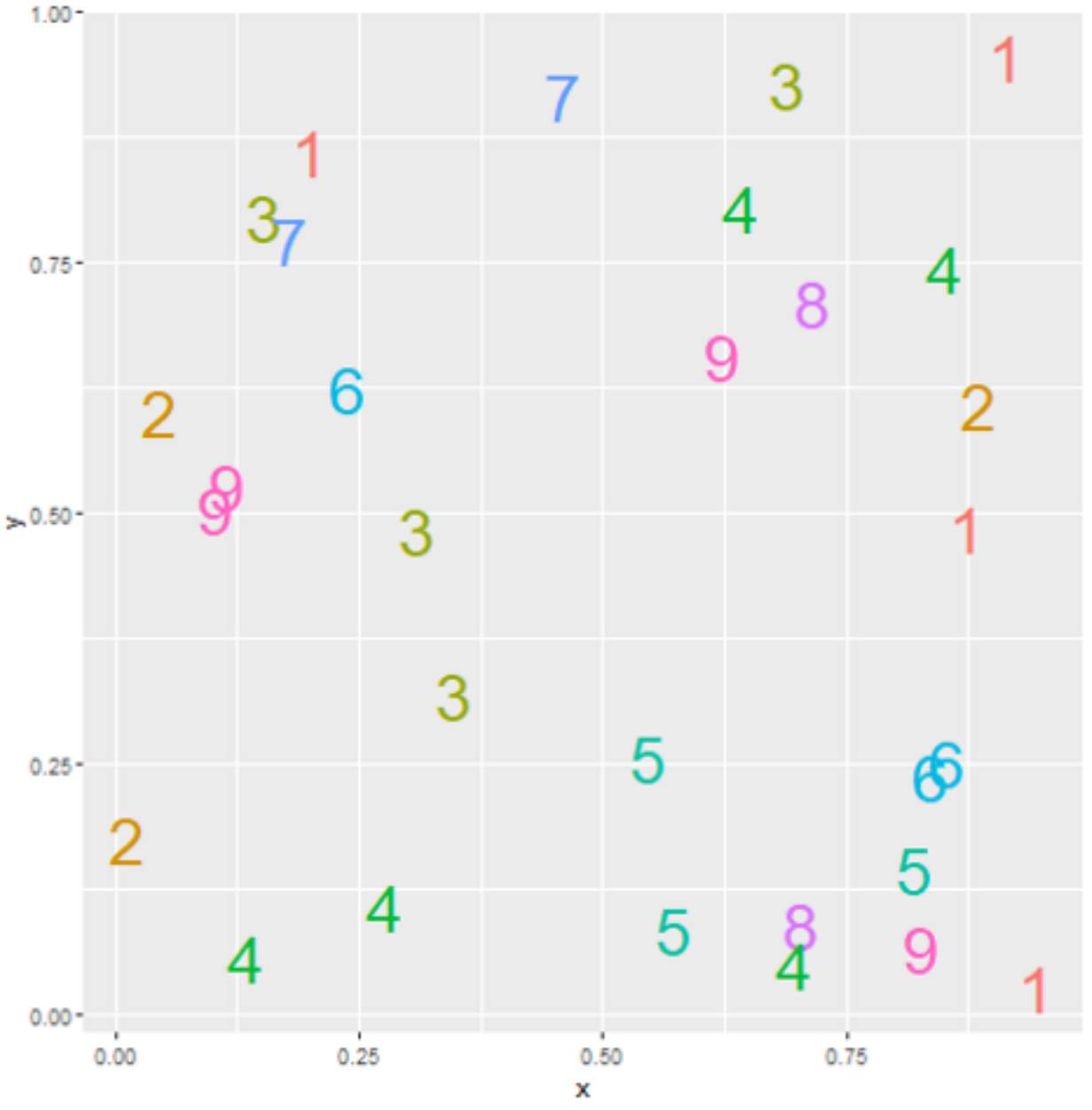


Figure 66: How many “5’s” are in this figure?

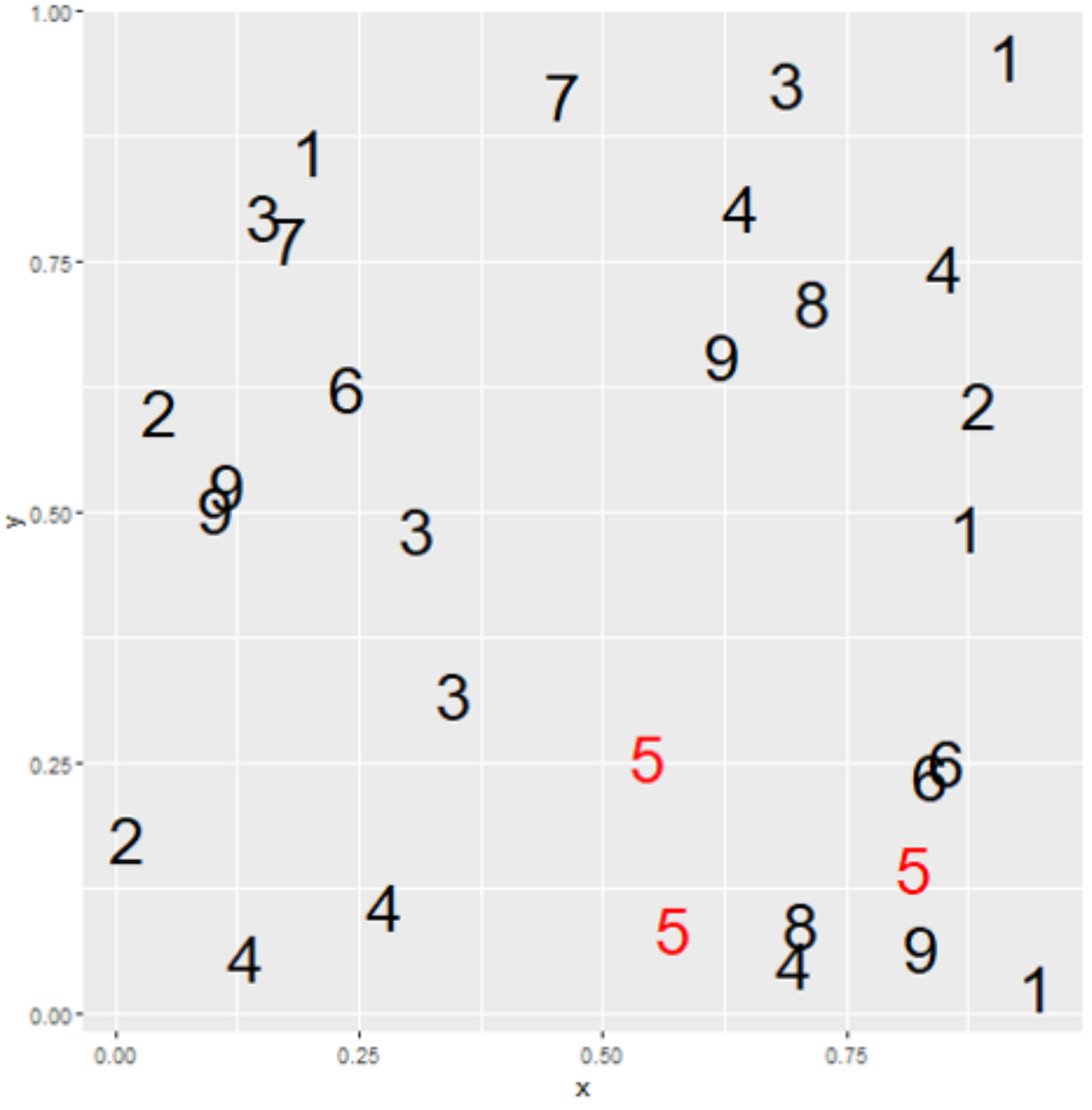


Figure 67: Repeat question. How many “5’s” are in this figure?

Repeat quiz question 1

```
          No             Yes           Total
Female   154 (33.3%)    308 (66.7%)    462 (100%)
Male     709 (83.3%)    142 (16.7%)    851 (100%)
Total    863 (65.7%)    450 (34.3%)   1313 (100%)
```

This data table shows counts and …

1. cell percents
2. column percents
3. row percents
4. I do not know the answer

Repeat quiz question 2

The median might be preferred to the mean if

1. a single extreme value distorts the mean
2. the data follows a bell shaped curve
3. there is very little variation in the data
4. you have a biased sample
5. I do not know the answer

Repeat quiz question 3

The problem with error bars is that they

1. fail to show if the data is skewed
2. have several competing definitions
3. use only two numbers to characterize your data
4. all of the above are correct
5. none of the above are correct
6. I do not know the answer