Biostatistics

Descriptive Statistics

Michael Luu, MPH

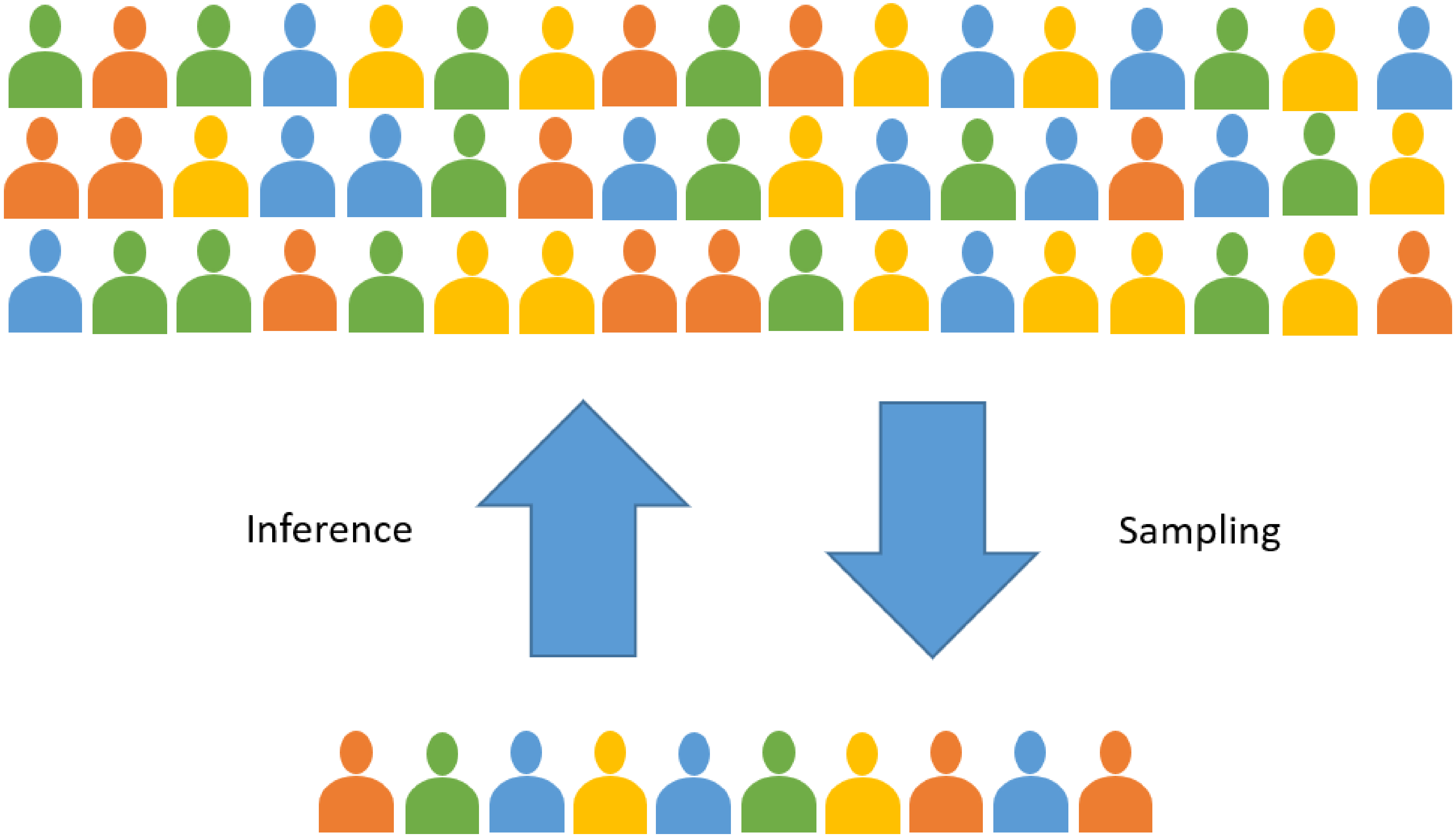
Biostatistics and Bioinformatics Research Center | Cedars Sinai Medical Center

June 21, 2022

Introduction

What is descriptive statistics?

- **Descriptive statistics** is a collection of statistical measures and tools used to give us a better sense of the sampled data
- Not to be confused with **inferential statistics** where we are trying to reach conclusions that extend beyond the sampled data



Study Population

Inference

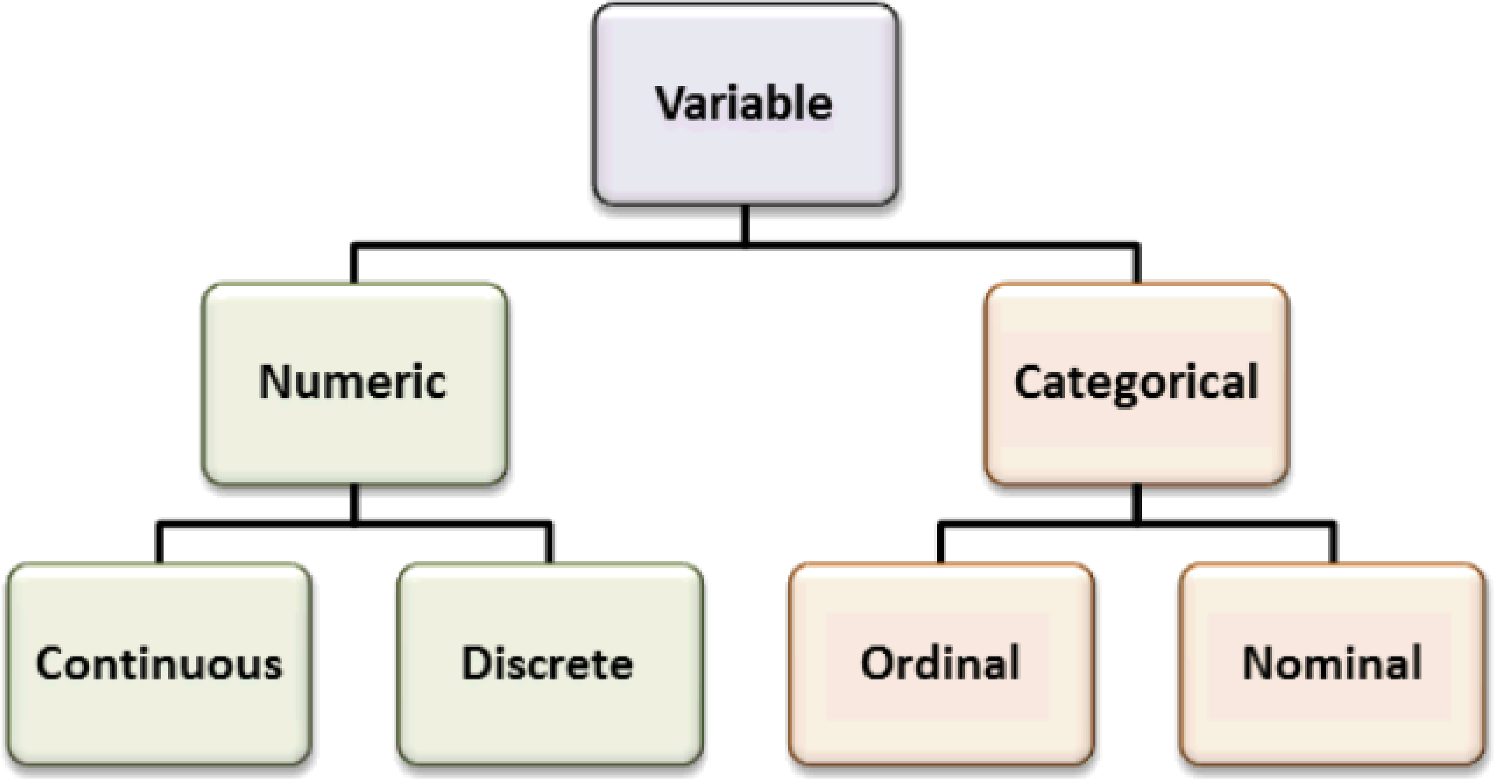
Sampling

Sample

Why descriptive statistics ?

- Provides an understanding of the underlying sample population
- Simplifies large amounts of data to a simpler summary
- Identifies potential measurement errors or mistakes

Types of Data


Variable

Numeric

Categorical

Continuous

Discrete

Ordinal

Nominal

Types of Data

Quantitative (Numeric)

Variable that has been measured on a numeric or quantitative scale

Continuous

- Can theoretically take on an infinite number of values - accuracy is limited only by the measuring instrument
    - e.g. age, BMI, BSA, height, weight, etc..

Discrete

- Numerical variables that are measured and can only be whole numbers
    - e.g. age, heart rate, number of medication taken, number of relapses, etc..

Types of Data

Qualitative (Categorical)

Variables that are typically not directly measured by an instrument, and are based on observations

Ordinal

- Variables that have an inherent hierarchical order to the relationship among the different categories
    - e.g. pain scores, stage of cancer, education level, etc..

Nominal

- Variables that are "named" or classified into one or more qualitative groups
- Do not have a sense of ordering between the different categories
    - e.g. risk factors, types of medications consumed, types of symptoms experienced, surgical outcomes, blood type, gender, etc..

Types of Data

Why do we need to identify the types of data ?

- In statistics we have specialized tools or measures to handle different type of data
- You will **NEED** to understand what kind of data you have in order to correctly summarize your data

Quantitative Data

Numerical Summarization

Measures of Location

- Mean
- Median
- Mode

Measures of Variability or Dispersion

- Minimum and Maximum
- Percentiles / Interquartile Range (IQR)
- Standard Deviation

Measures of Location

Mean

- The sample mean is the most commonly used and readily understood measure of central tendency.
- The sample mean can be defined as:

$$\bar{x} = \frac{\sum x_i}{n}$$

Example Data

patient_id	sex	age
1	M	32
2	F	35
3	F	35
4	F	34
5	M	36
6	F	36
7	F	30
8	F	36
9	F	33
10	F	32

Example

We have a collection of **age** that was collected from the sample population

```
[1] 32 35 35 34 36 36 30 36 33 32
```

- The **total sum** of all of the sampled age is 339
- The **total number of measurements** collected is 10
- The **mean** or average age among the sample population is 33.9

Measures of Location

Median

- The median is the midpoint of the values
  - The midpoint value is the point at which half the observations are above the value and half the observations are below the value *(50th percentile*).
  - If there are two 'middle' values then the median is the average of the two mid values

Example

Let's recall the collection of **age** that we have collected from the sample population

```
 [1] 32 35 35 34 36 36 30 36 33 32
```

- We start by ranking the data from smallest to largest

```
 [1] 30 32 32 33 34 35 35 36 36 36
```

- We identify the **middle** value from the data

```
[1] 34 35
```

- We then take the average of the two middle value to obtain the **median**

```
[1] 34.5
```

Measures of Location

Mode

- The mode is the value that appears most often in a set of values.
- Not always a measure of central tendency
- The mode is only useful for discrete values or continuous values with limited digits of accuracy
- It's possible to have more than 1 mode

Example

```
 [1] 32 35 35 34 36 36 30 36 33 32
```

- Let's tabulate the occurences of each of the sampled age

```
30 32 33 34 35 36 
 1  2  1  1  2  3 
```

- The mode is the value that occured the most often.

- In our example, the mode is **36**

Measures of Variability or Dispersion

Minimum and Maximum

A measure of dispersion and is defined as the smallest and largest value

Example

```
[1] 32 35 35 34 36 36 30 36 33 32
```

- We will rank the data from smallest to largest

```
[1] 30 32 32 33 34 35 35 36 36 36
```

- The min and max corresponds to the smallest and largest values of our sample.
- In our example, the min would be **30**, and the max would be **36**

Measures of Variability or Dispersion

Percentiles / Interquartile Range (IQR)

- The interquartile range is defined as the range between the 25th and 75th percentiles

$$IQR = Q_3 - Q_1$$

- It is commonly denoted after presenting the median
- The interquartile range can be used to describe the spread of the data. As the spread of the data increases, the IQR becomes larger.
- It is also used to build box plots.
- Depending on the statistical software you are using, there are multiple ways of calculating quantile

Example

We determine the interquartile range by calculating quantiles at the 25th and the 75th percentile.

```
 [1] 32 35 35 34 36 36 30 36 33 32
```

- We begin by sorting the values

```
 [1] 30 32 32 33 34 35 35 36 36 36
```

- The 25th percentile is the value that divides the data where 25 percent falls below this value and 75 percent falls above this value.

- The 75th percentile is the value that divides the data where 75 percent falls below this value and 25 percent falls above this value.

```
  0%  25%  50%  75% 100% 
30.0 32.0 34.5 36.0 36.0 
```

- The interquantile range is the distance between the value of the 75th percentile and the 25th percentile.

- In this example, the IQR is **4**

Measures of Variability or Dispersion

Standard Deviation

The standard deviation is a measure of how spread out the data are about the mean.

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

Example

```
[1] 32 35 35 34 36 36 30 36 33 32
```

- The mean sample age is **33.9**
- There is a total of **10** measurements

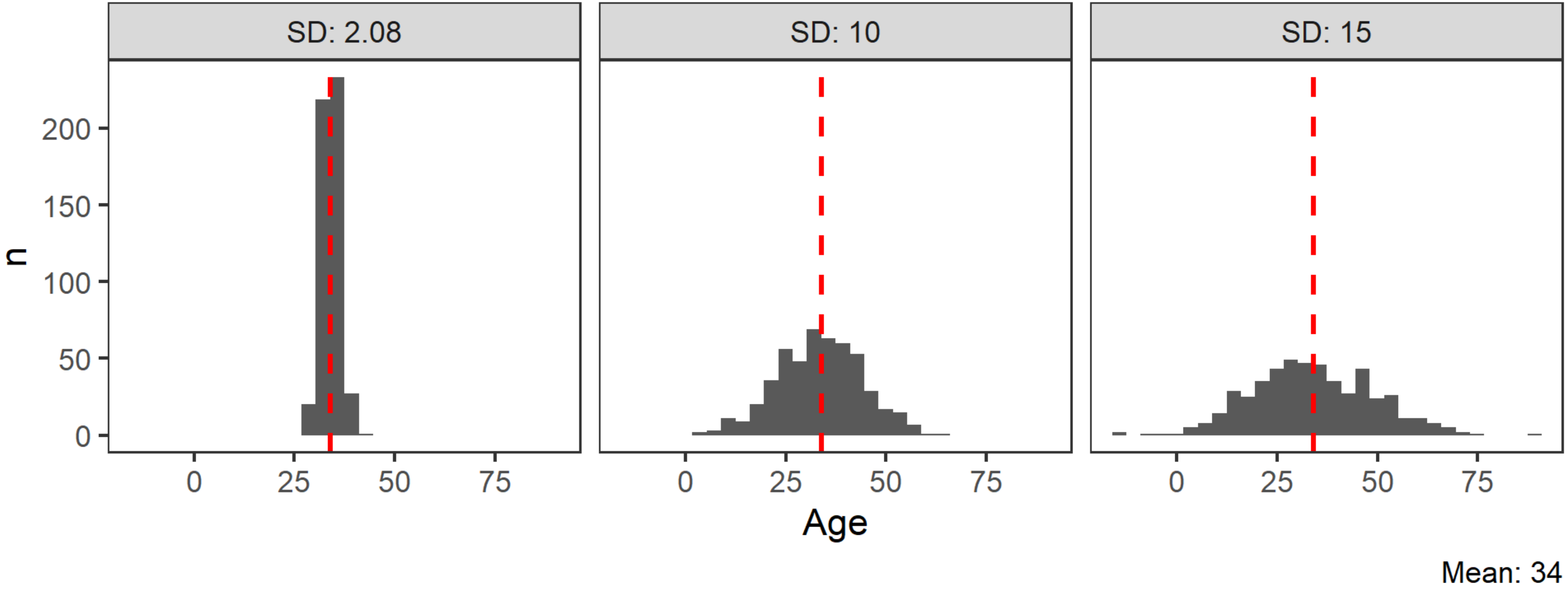
```
# A tibble: 10 × 4
     age mean_age   diff   diff2
   <int>    <dbl>  <dbl>   <dbl>
 1    32     33.9 -1.90   3.61
 2    35     33.9  1.10   1.21
 3    35     33.9  1.10   1.21
 4    34     33.9  0.100  0.0100
 5    36     33.9  2.10   4.41
 6    36     33.9  2.10   4.41
 7    30     33.9 -3.9   15.2
 8    36     33.9  2.10   4.41
 9    33     33.9 -0.900  0.810
10    32     33.9 -1.90   3.61
```

Example

```
# A tibble: 1 × 4
  summ_diff2 total_length_minus_one   var    sd
       <dbl>                  <dbl> <dbl> <dbl>
1       38.9                      9  4.32  2.08
```

The sample standard deviation is **2.08**

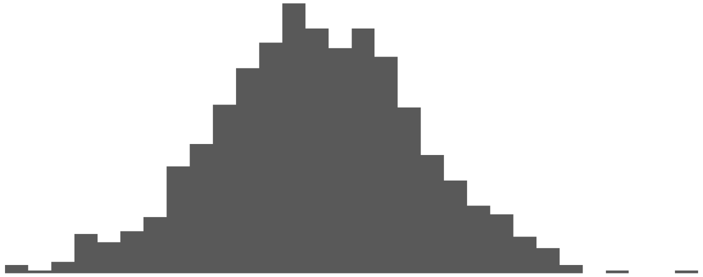
Example



Presenting / Describing your data

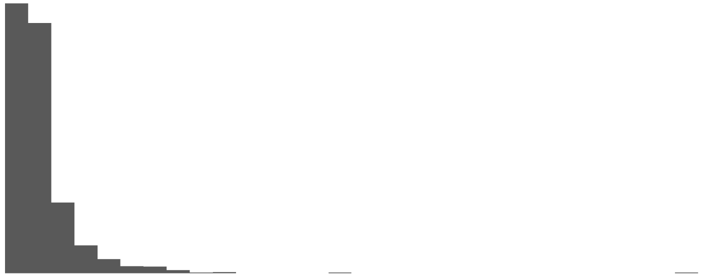
- You should always include **BOTH** a *measure of central tendency* and *measure of dispersion* when presenting your data
- The choice is dependent on the distribution of your data

Symmetric Distribution



- mean ± SD
- median (25%, 75% quantiles)
- median (IQR)
- median (min - max)

Asymmetric Distribution



- median (25%, 75% quantiles)
- median (IQR)
- median (min - max)

Qualitative Data

Summarizing Qualitative Data

- The primary method of summarizing qualitative data is frequency **counts** and **percentages**

```
 [1] "M" "F" "F" "F" "M" "F" "F" "F" "F" "F"
```

- Let's tabulate the total occurrences of M and F in our data

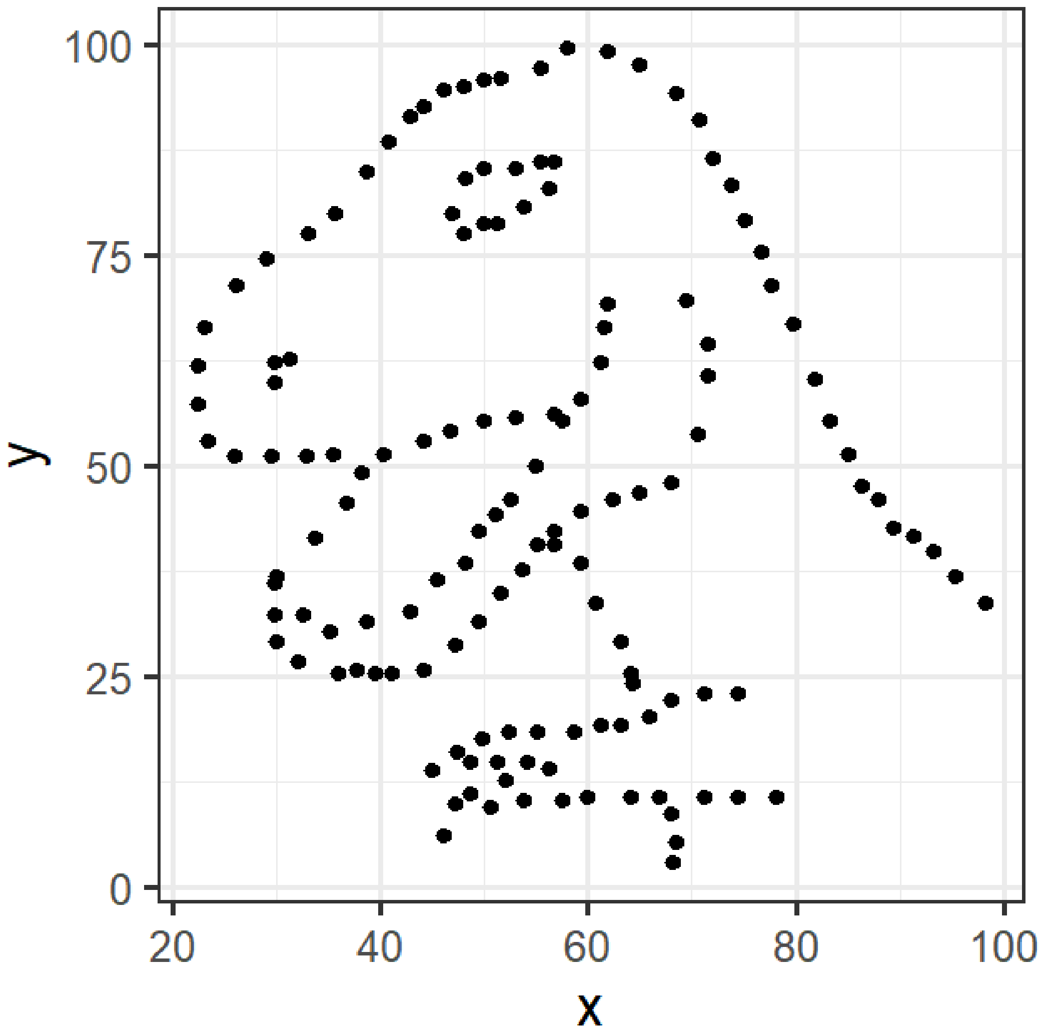
```
# A tibble: 2 × 2
  sex       n
  <chr> <int>
1 F         8
2 M         2
```

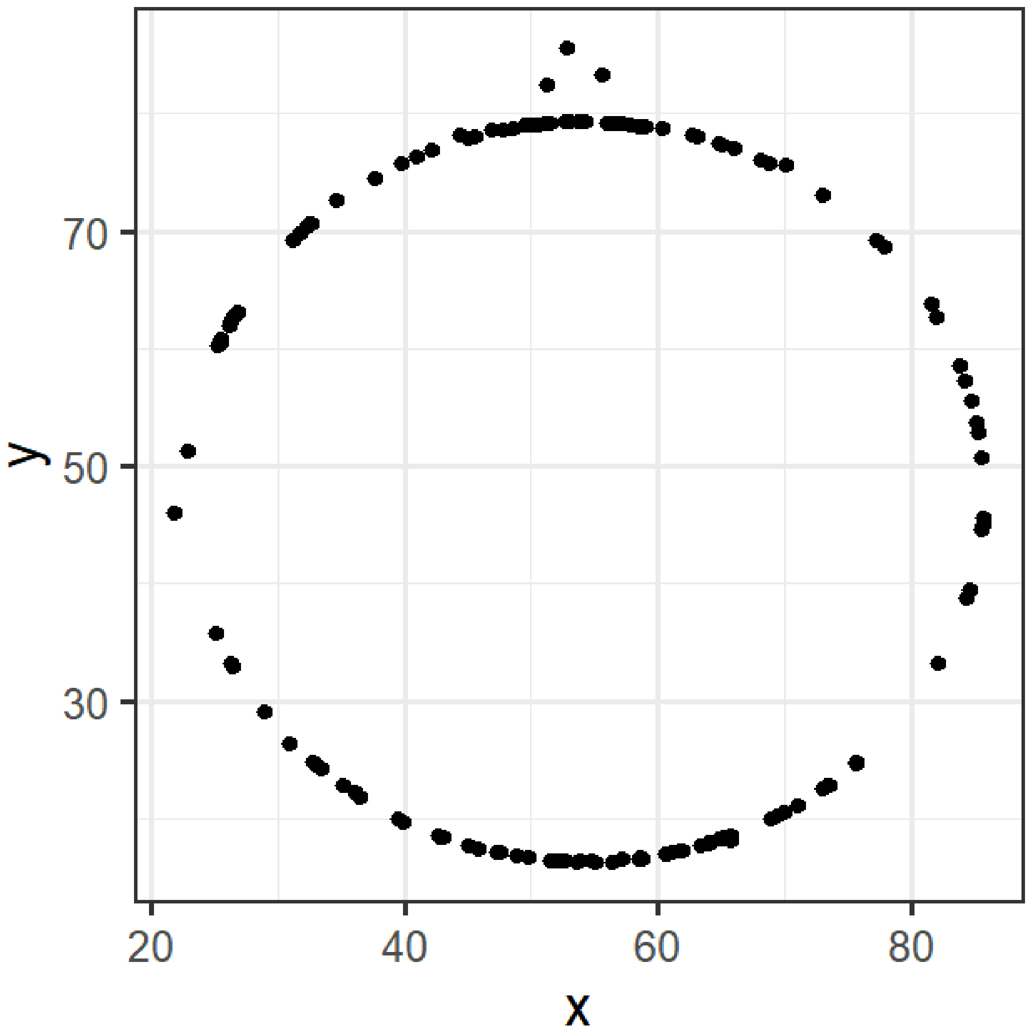
- Next we take the total and divide by the total number of patients

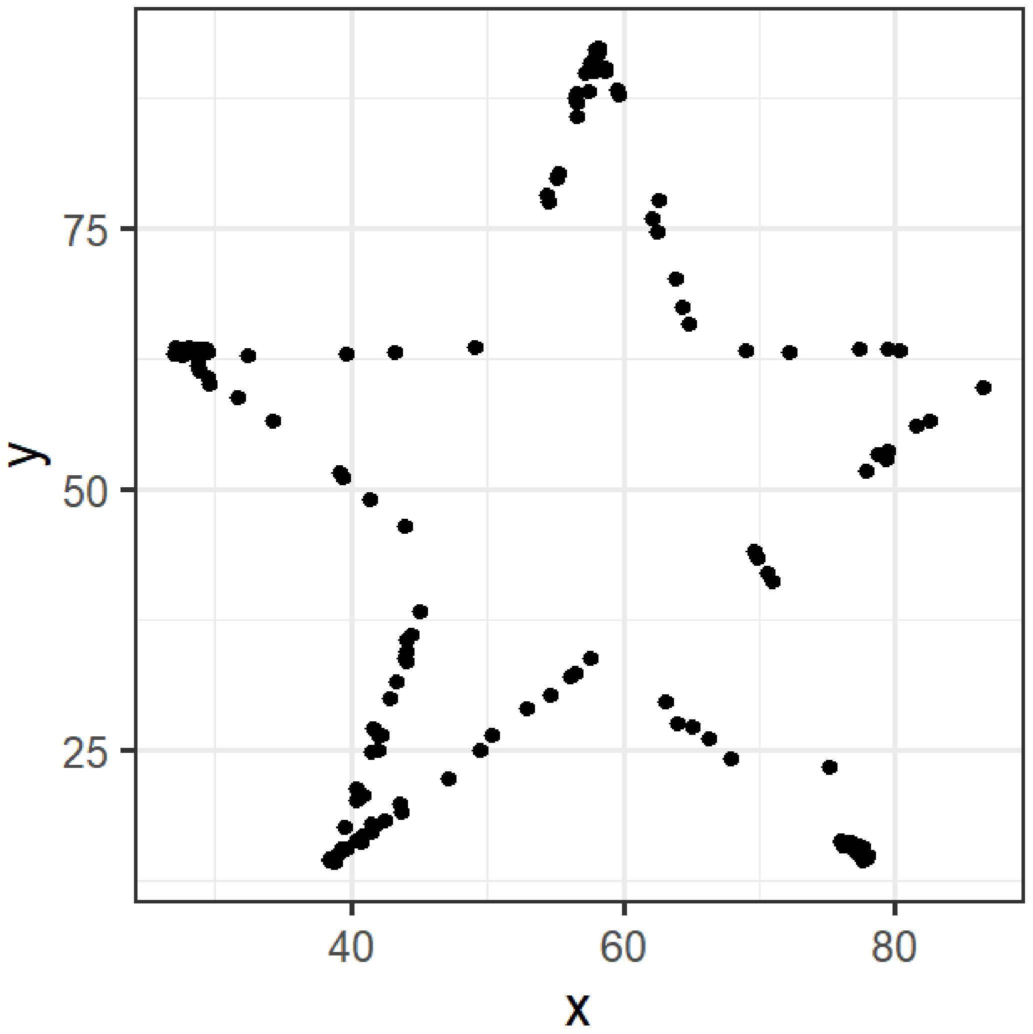
```
# A tibble: 2 × 3
  sex       n proportion
  <chr> <int> <chr>
1 F         8 80.0%
2 M         2 20.0%
```

Graphical Summarizations

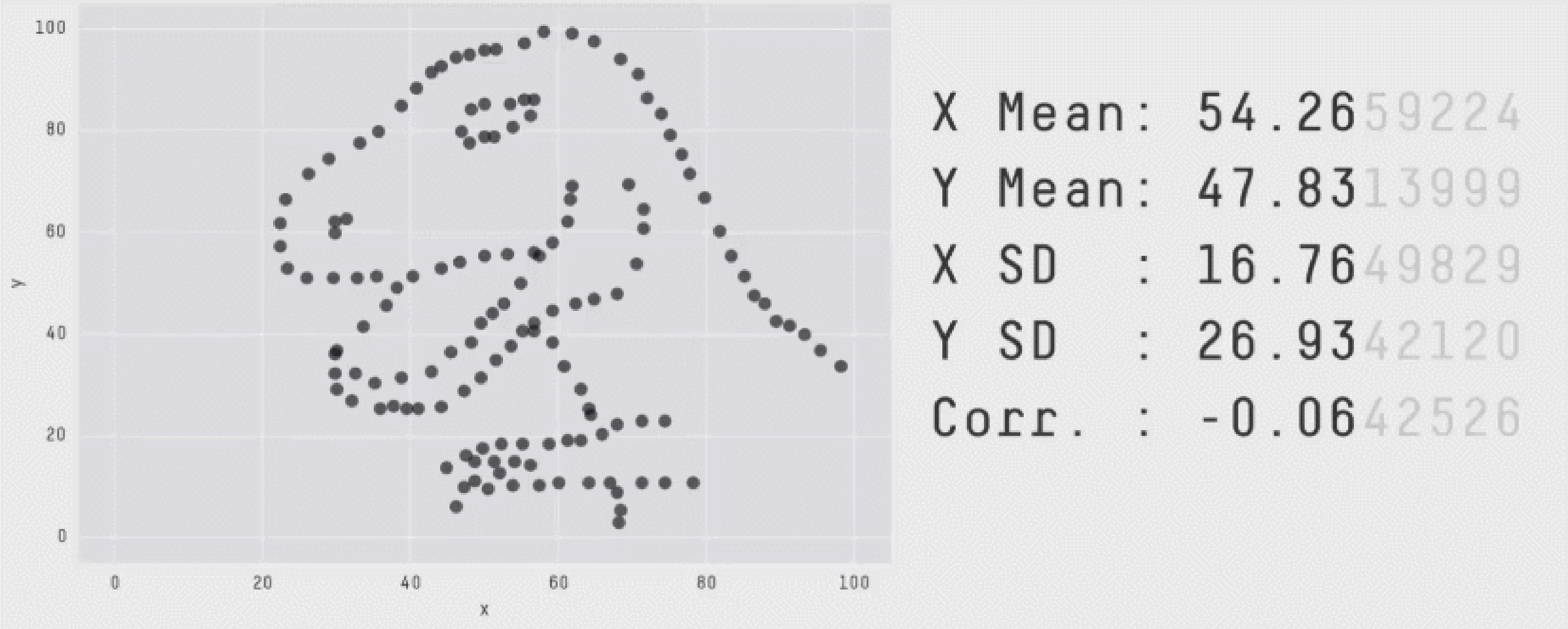
What do the following figures have in common?


100
75
50
25
0
y
20
40
60
80
100
x


70
50
30
y
20
40
60
80
x


75
50
25
y
40
60
80
x

They all have the same summary statistic …



```
X Mean: 54.2659224
Y Mean: 47.8313999
X SD  : 16.7649829
Y SD  : 26.9342120
Corr. : -0.0642526
```

https://www.autodesk.com/research/publications/same-stats-different-graphs

Quantitative summary measures are useful …

Graphical summarizations provides an additional perspective

Graphical Summarization

Quantitative Data

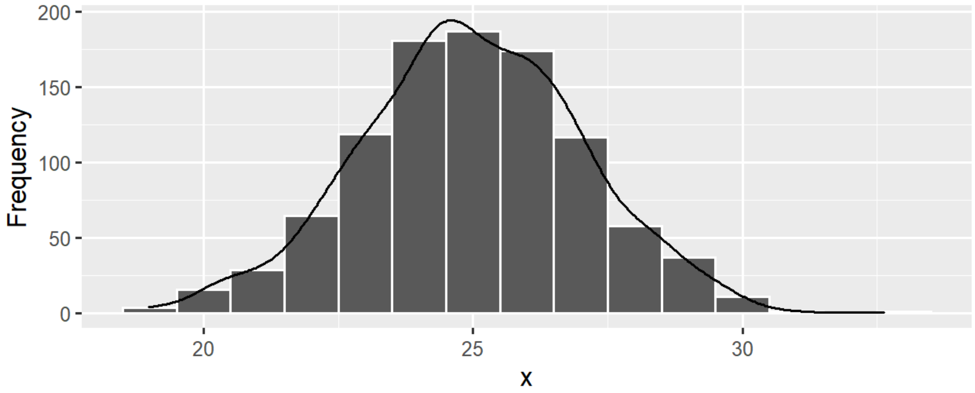
- Histograms
- Dot plots
- Box plots

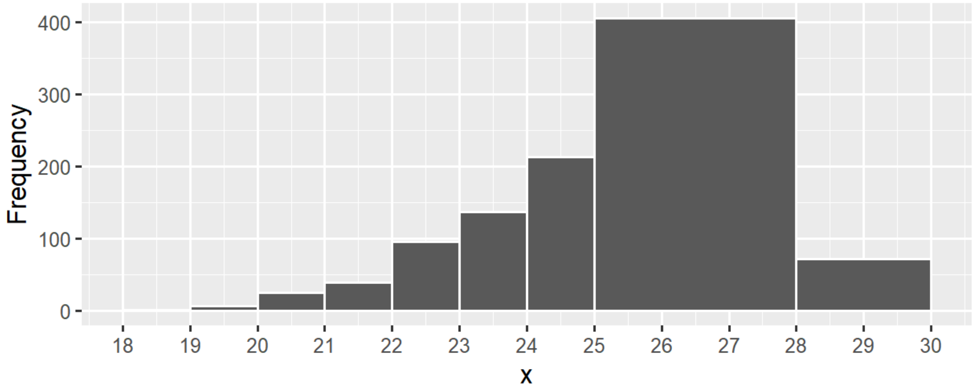
Qualitative Data

- Bar plots

Histogram

- Useful for all sized data (small and large)
- Allows us to visualize the spread and distribution of continuous variables
- Each bar represents a 'bin' or a defined interval of values
- Although not as common, the width of the bins does NOT have to be equal!
- The y axis or the height of the bar represents the count of the number of values that fall into each bin
- The y axis is also commonly normalized to 'relative' frequencies to show the proportion of cases or density that falls into each bin.

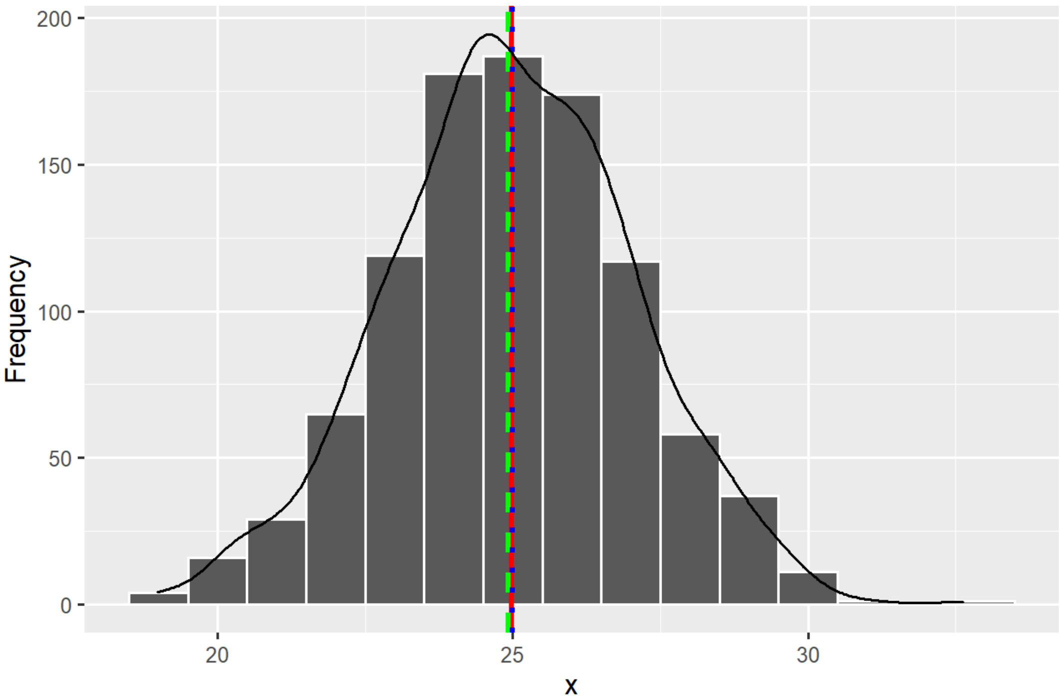




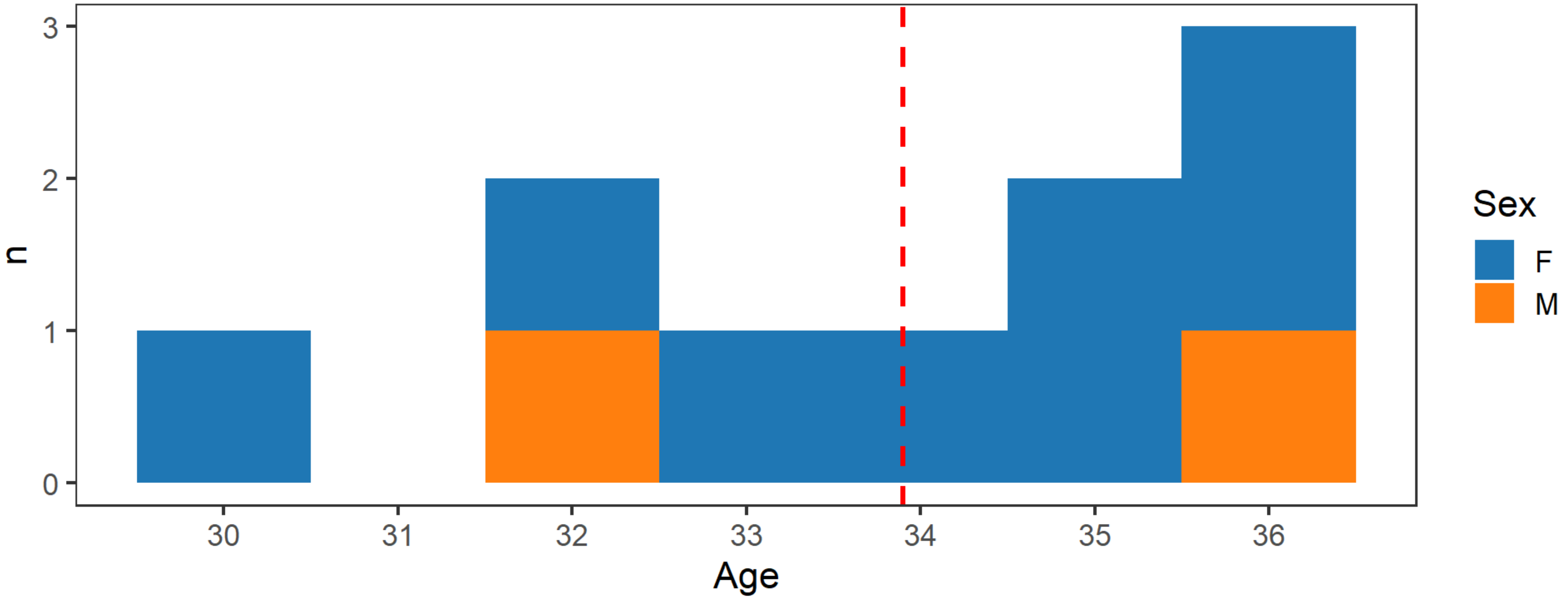
Distribution

> "A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically."

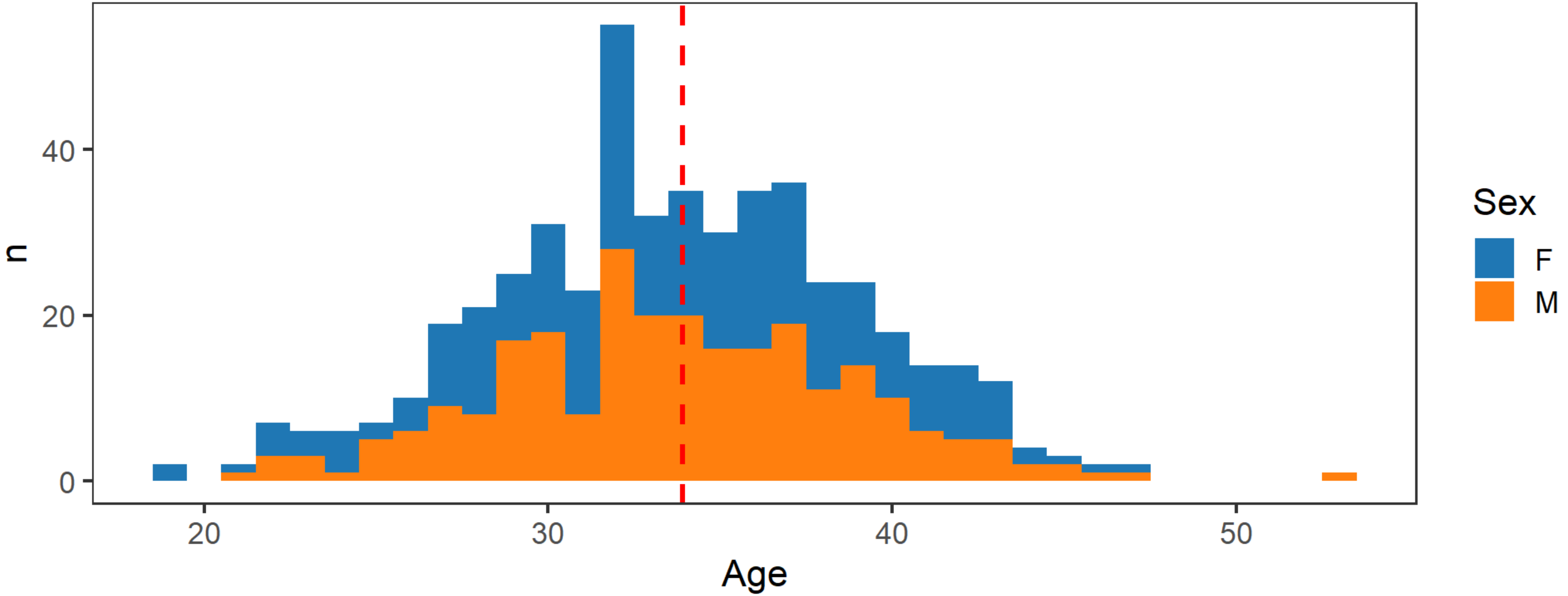
— Page 6, Statistics in Plain English, Third Edition, 2010.



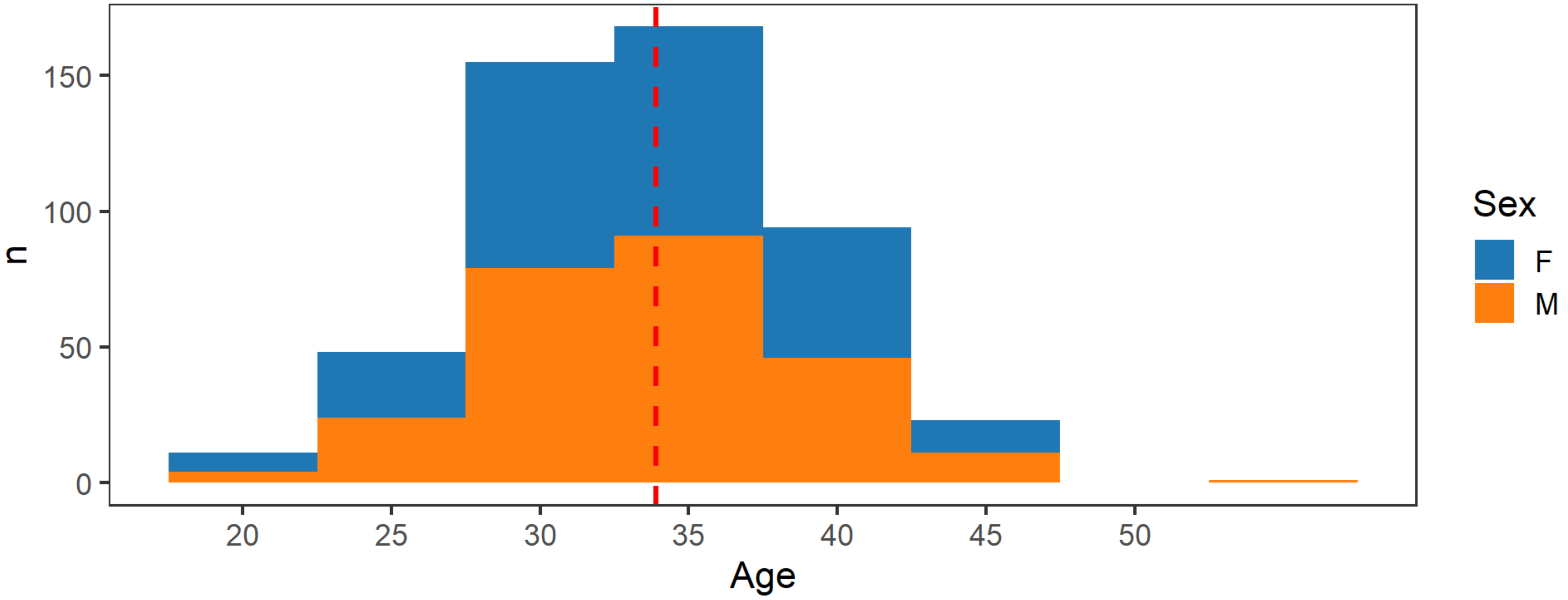
Example (n = 10, binwidth = 1)



Example (n = 500, binwidth = 1)

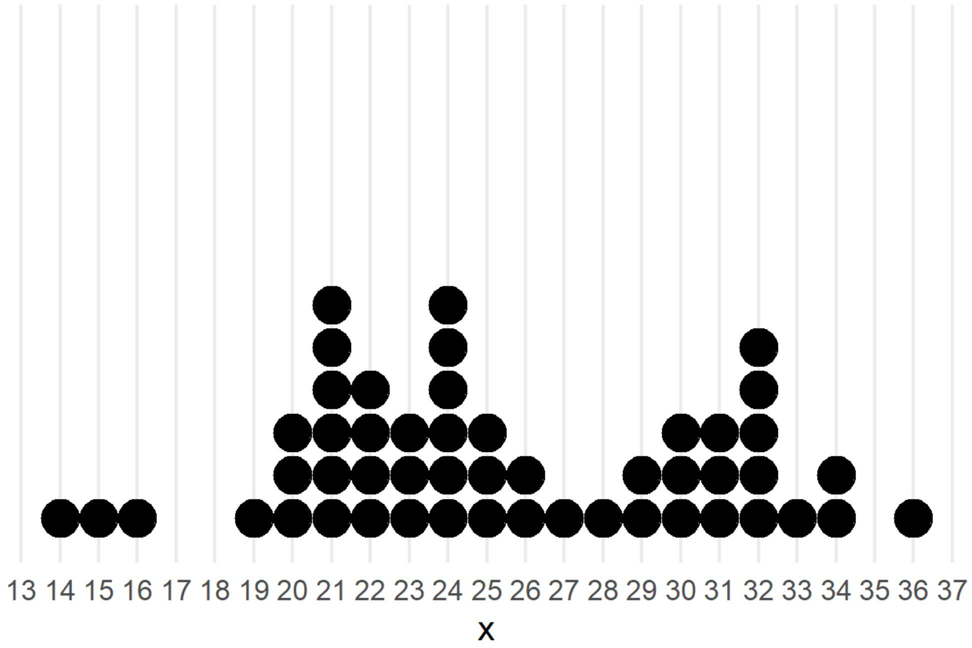


Example (n = 500, binwidth = 5)

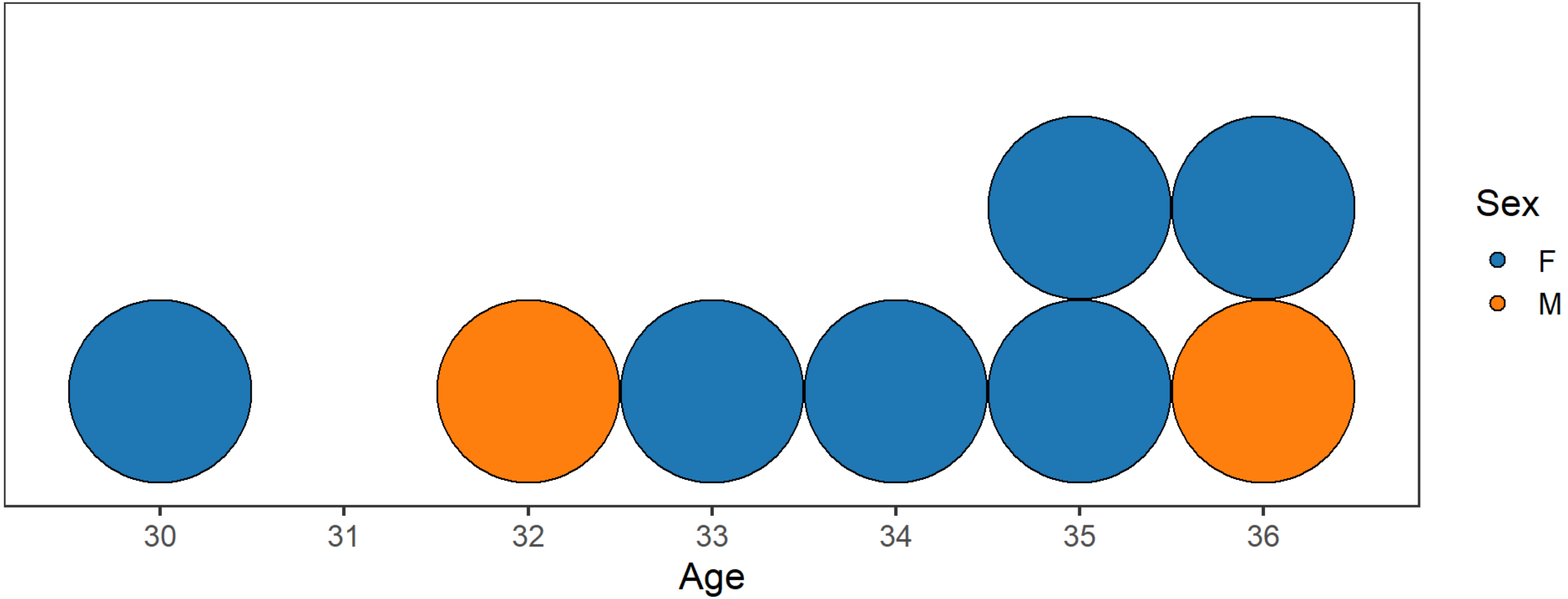


Dot plot

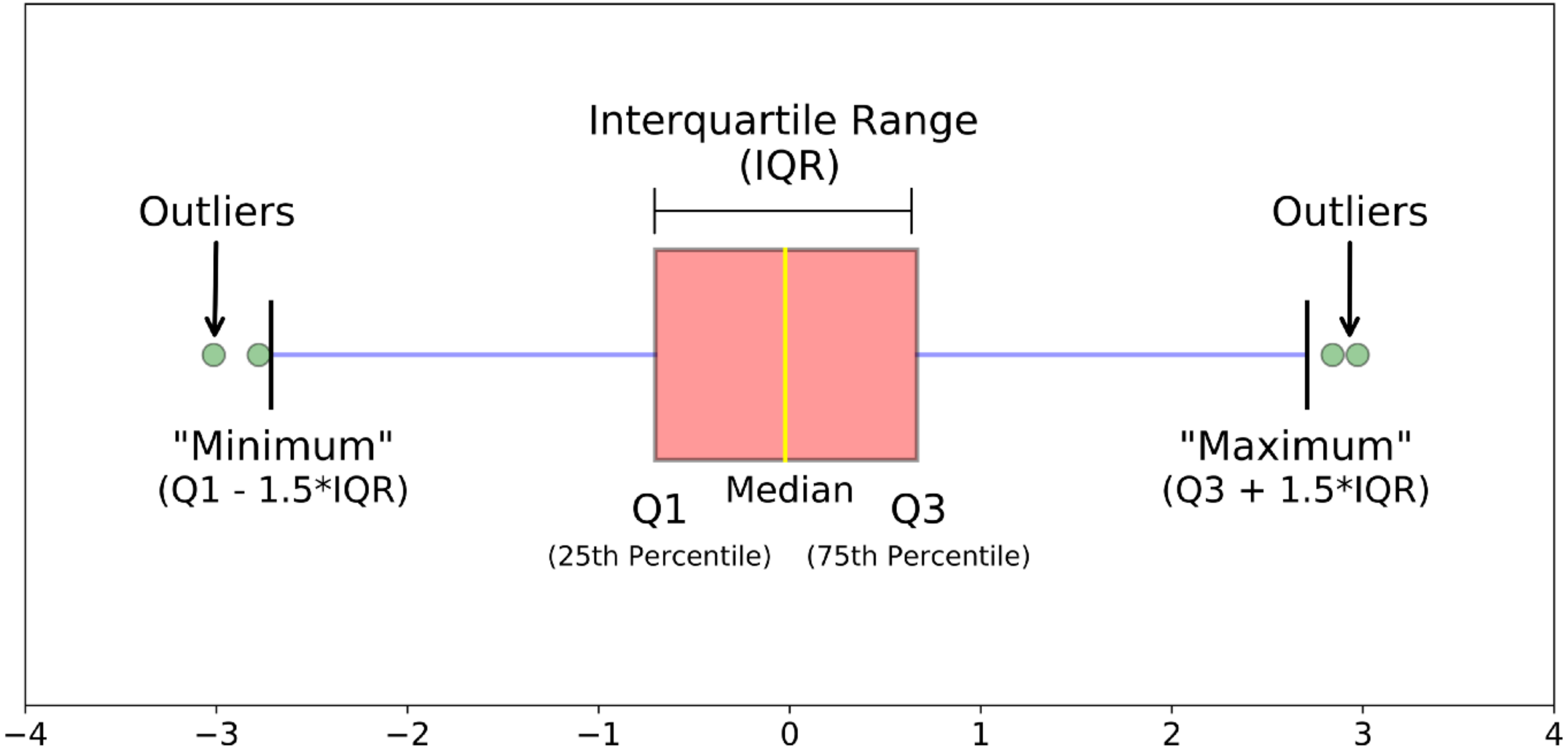
- Useful for small to moderate sized data
- Allows us to visualize the spread and distribution of one continuous discrete variables
    - e.g. length of stay
- The X axis is the variable of interest and each dot represents a single observation
- Easy to identify the mode
- Highlights clusters, gaps, and outliers
- Intuitive and easy to understand



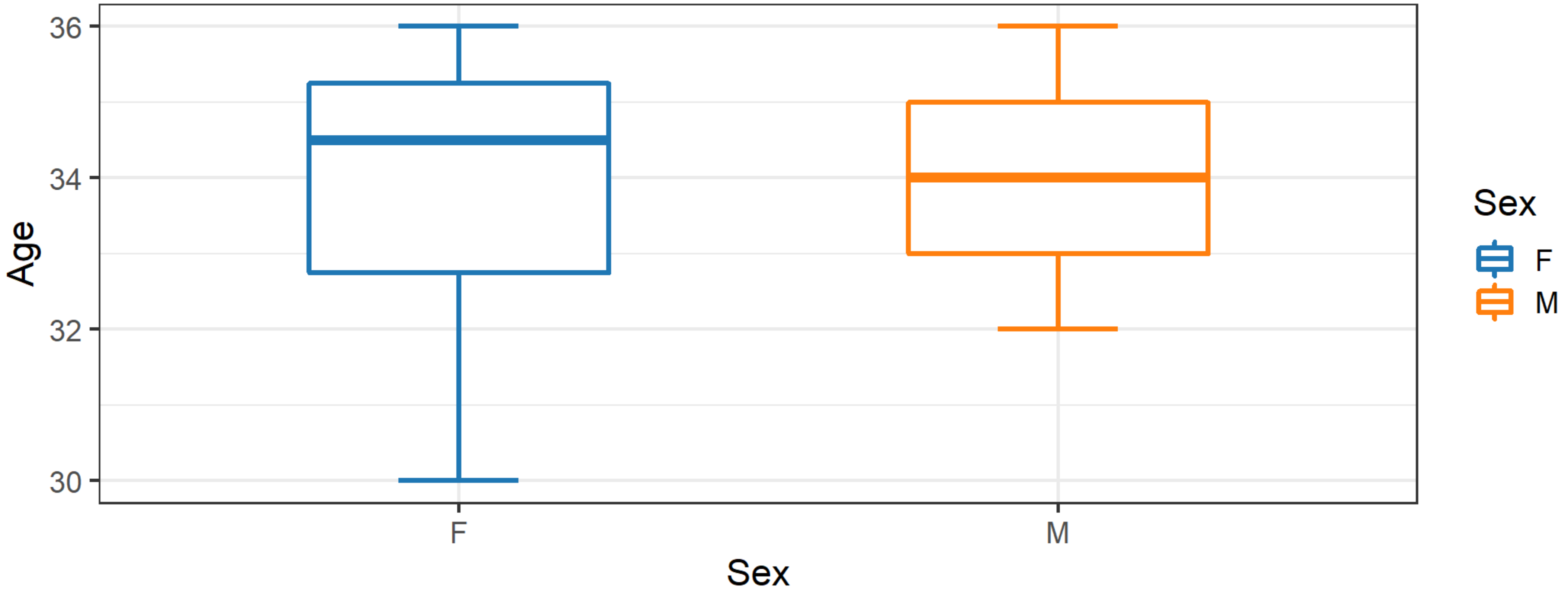
Example



Box plots

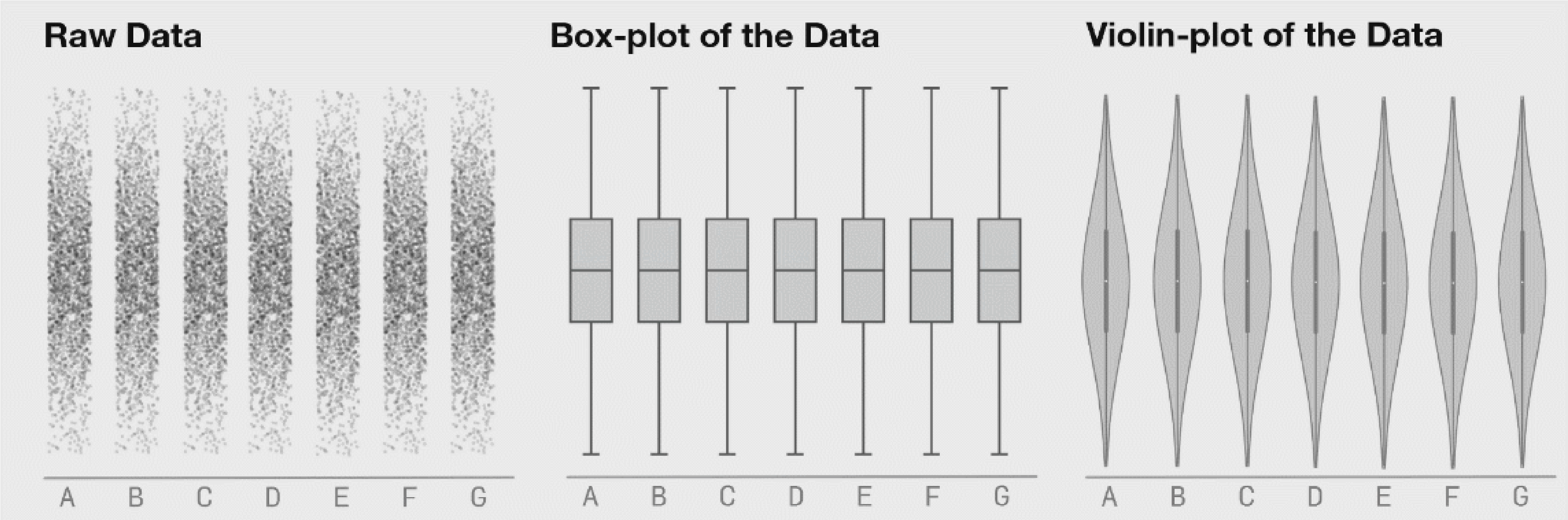

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
"Maximum"
(Q3 + 1.5*IQR)
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

Example



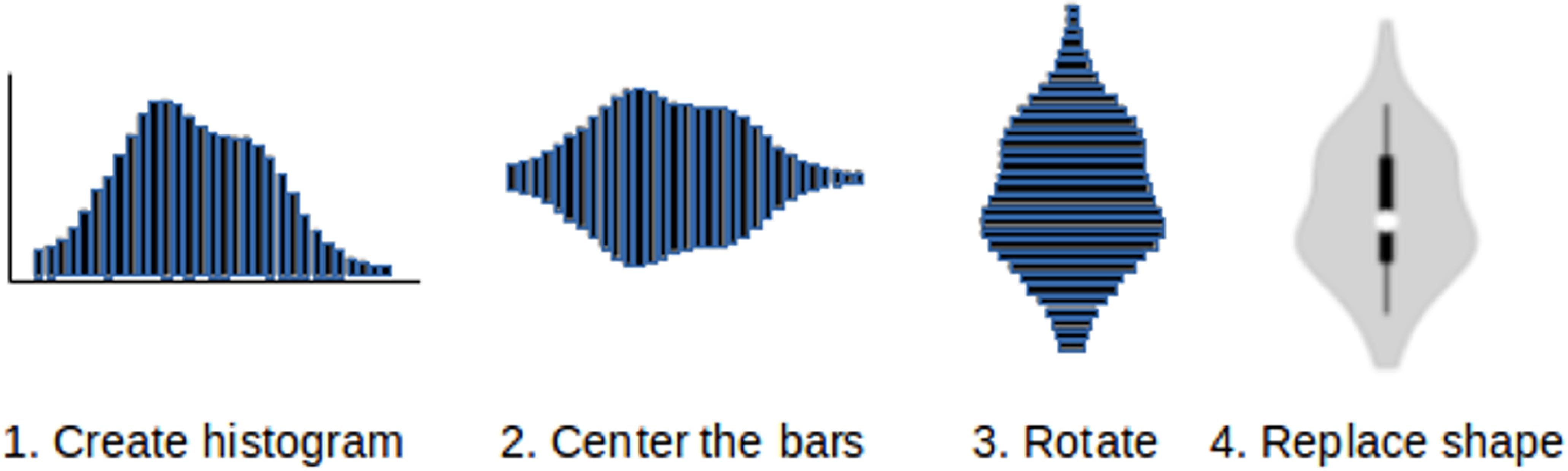
Boxplots are not perfect …

Raw Data vs Box Plot vs Violin Plot



https://www.autodesk.com/research/publications/same-stats-different-graphs

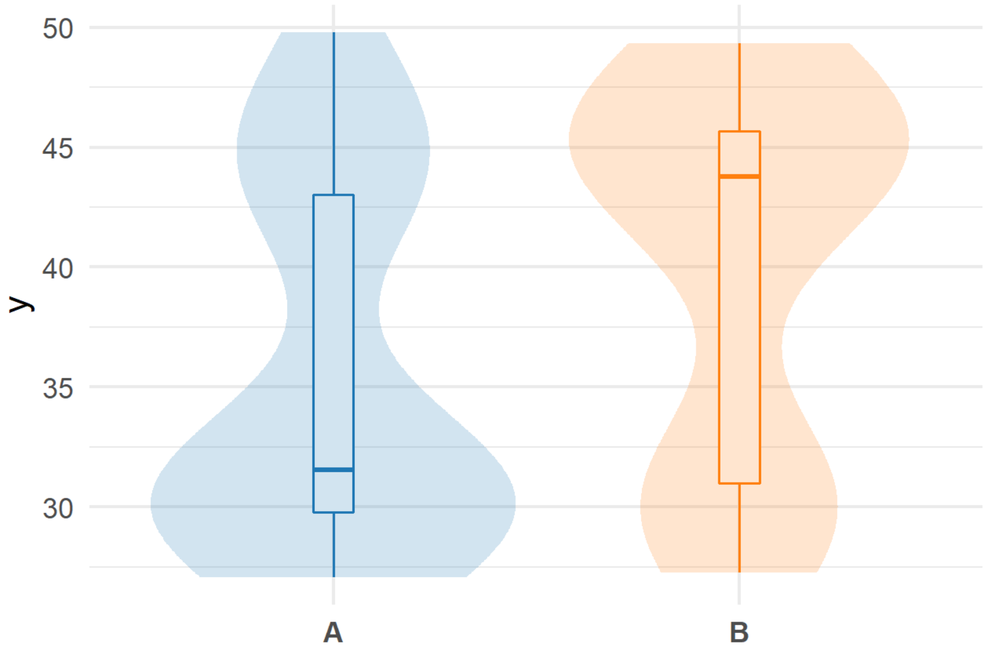
How are violin plots made?



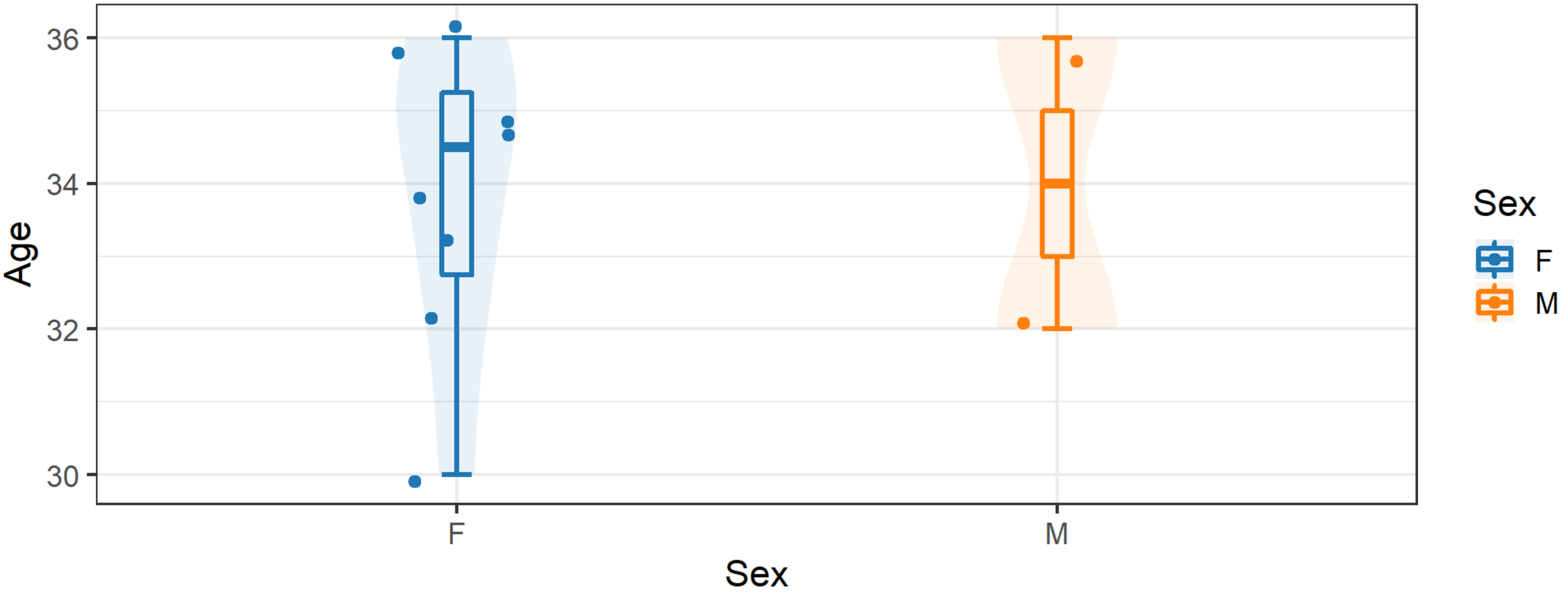
- Hintze, J. L., & Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. The American Statistician, 52(2), 181-184.

Violin plot

- Violin plots are box plots, with an overlay of the density distribution (histogram) of the data
- More informative than a simple box plot
- Visualizes the full distribution of the data
- Especially useful for bimodal or multimodal distribution
    - e.g. distribution of data with multiple peaks



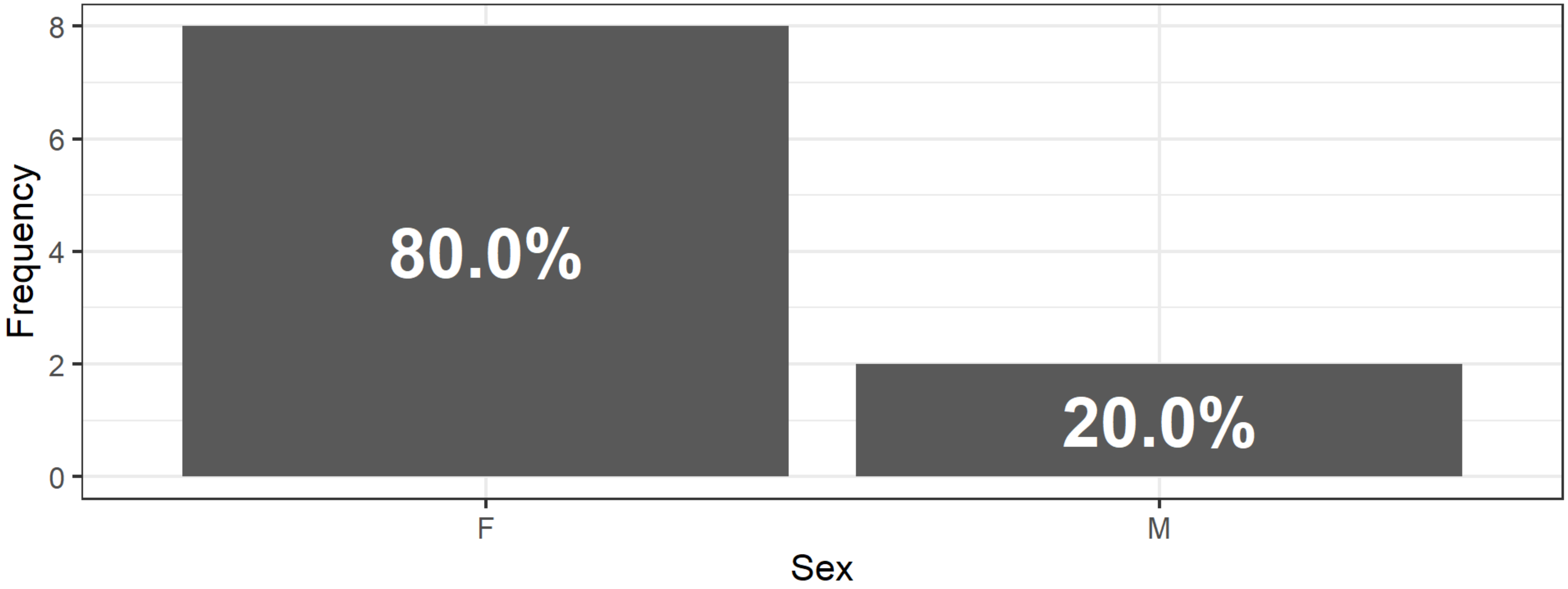
Example



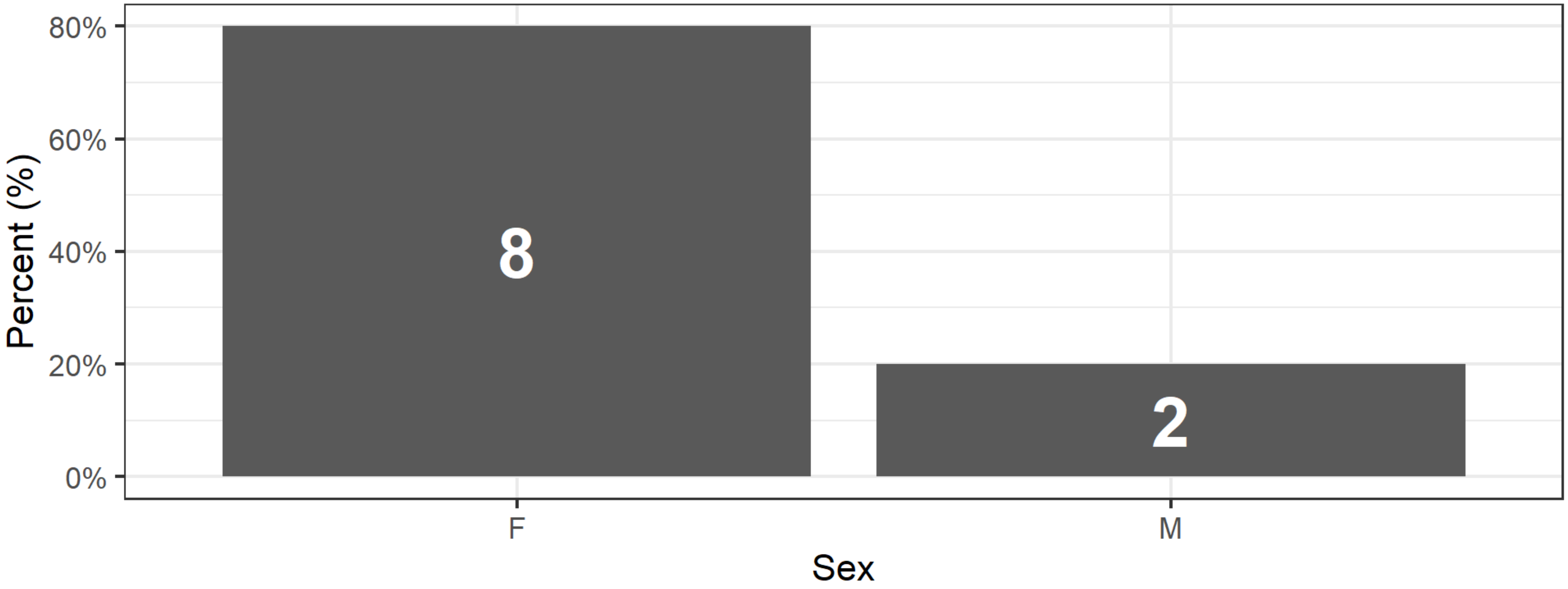
Bar plot

- Useful for visualizing **categorical** data
- Commonly used to present counts and proportion of each level
- Allows us to quickly observe the difference in magnitude of each level based on the height of each bar

Bar plot

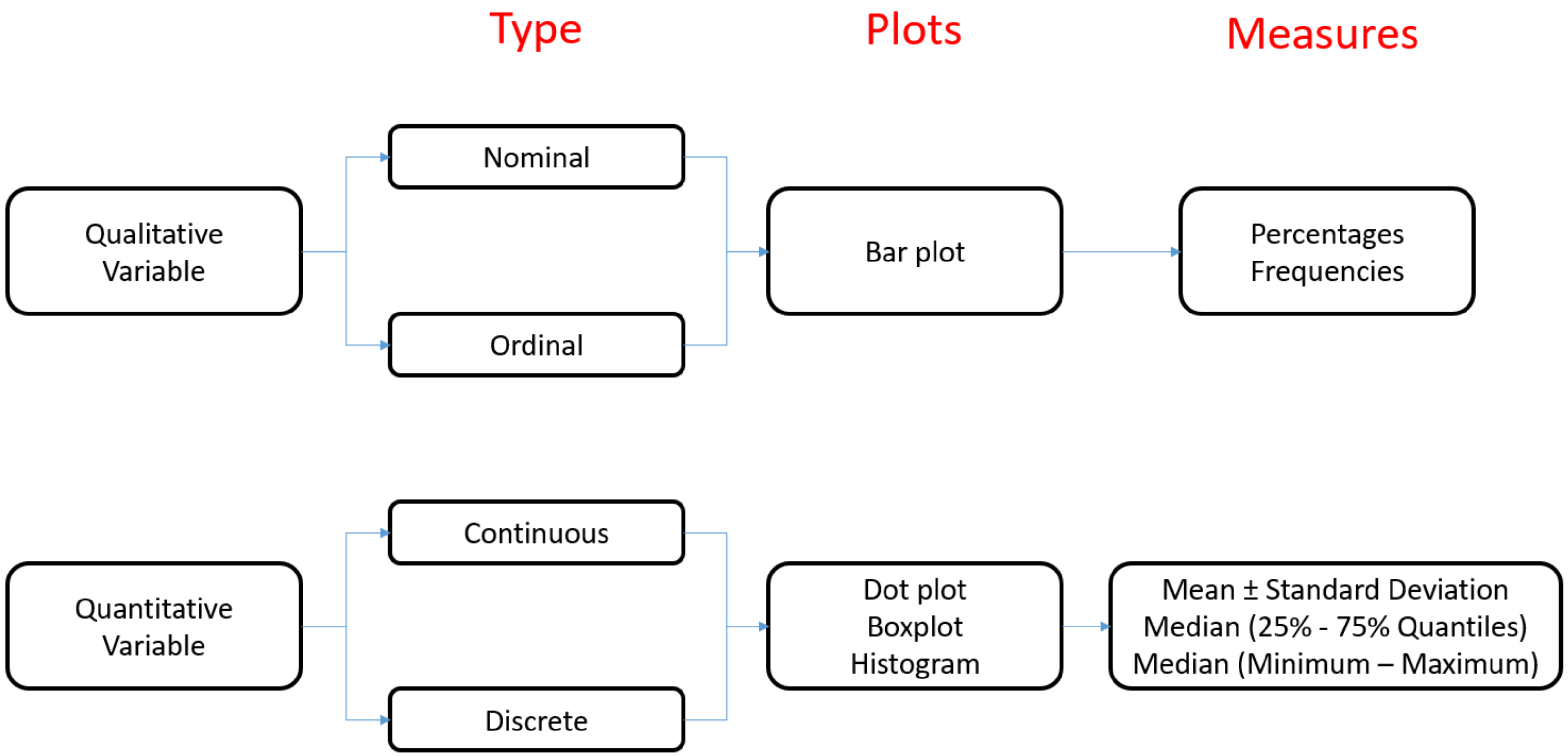


Bar plot



Summary

Summary


Type
Plots
Measures
Qualitative Variable
Nominal
Ordinal
Bar plot
Percentages Frequencies
Quantitative Variable
Continuous
Discrete
Dot plot Boxplot Histogram
Mean ± Standard Deviation
Median (25% - 75% Quantiles)
Median (Minimum – Maximum)

Questions