Descriptive statistics

Measure trunk diameter of 30 trees in your neighbourhood

Read data

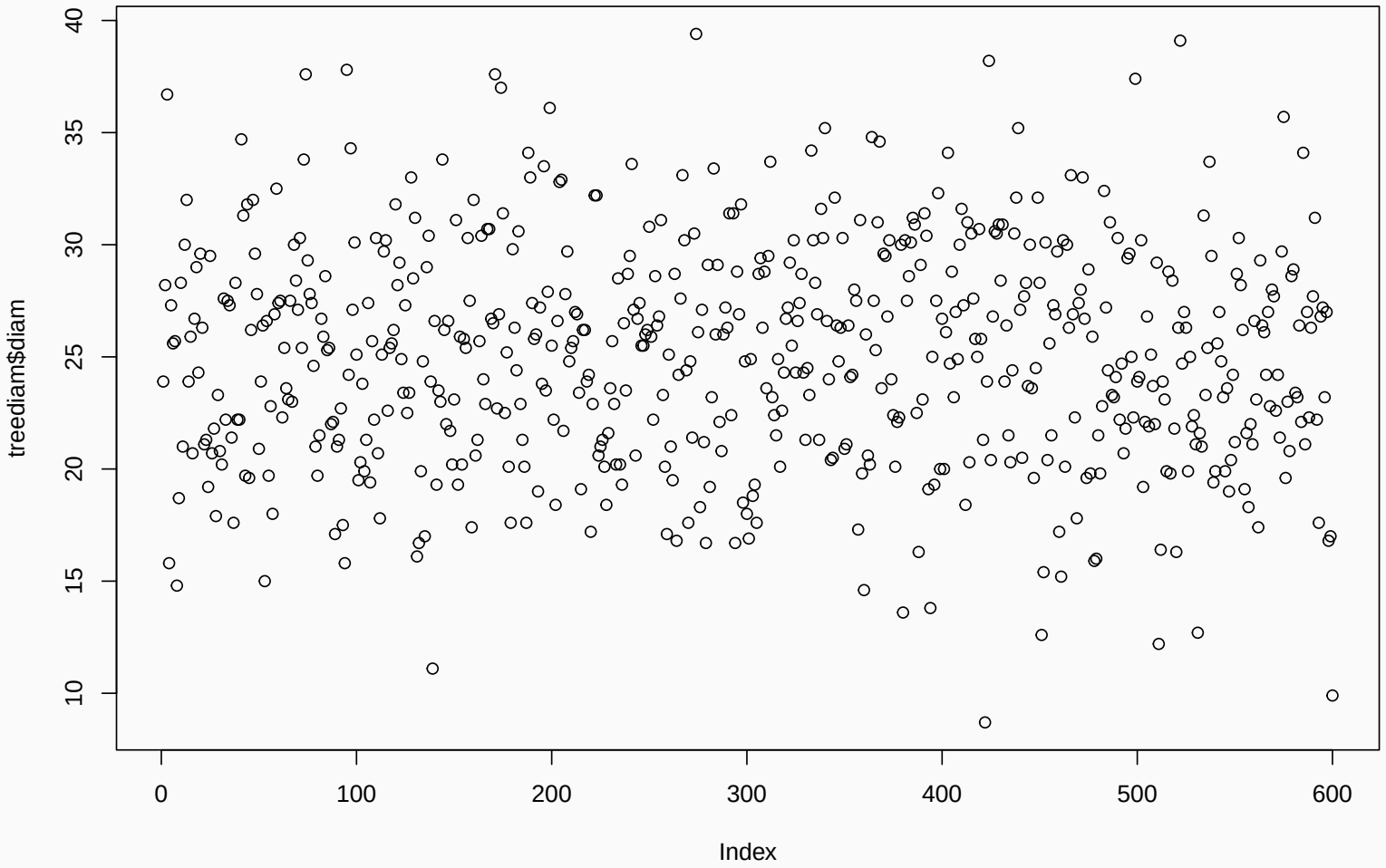
```
treediam <- read.csv("treediam.csv")
```

```
summary(treediam)
```

```
      site            tree            diam      
 Min.   : 1.00   Min.   : 1.0   Min.   : 8.70  
 1st Qu.: 5.75   1st Qu.: 8.0   1st Qu.:21.40  
 Median :10.50   Median :15.5   Median :25.25  
 Mean   :10.50   Mean   :15.5   Mean   :25.04  
 3rd Qu.:15.25   3rd Qu.:23.0   3rd Qu.:28.40  
 Max.   :20.00   Max.   :30.0   Max.   :39.40  
```

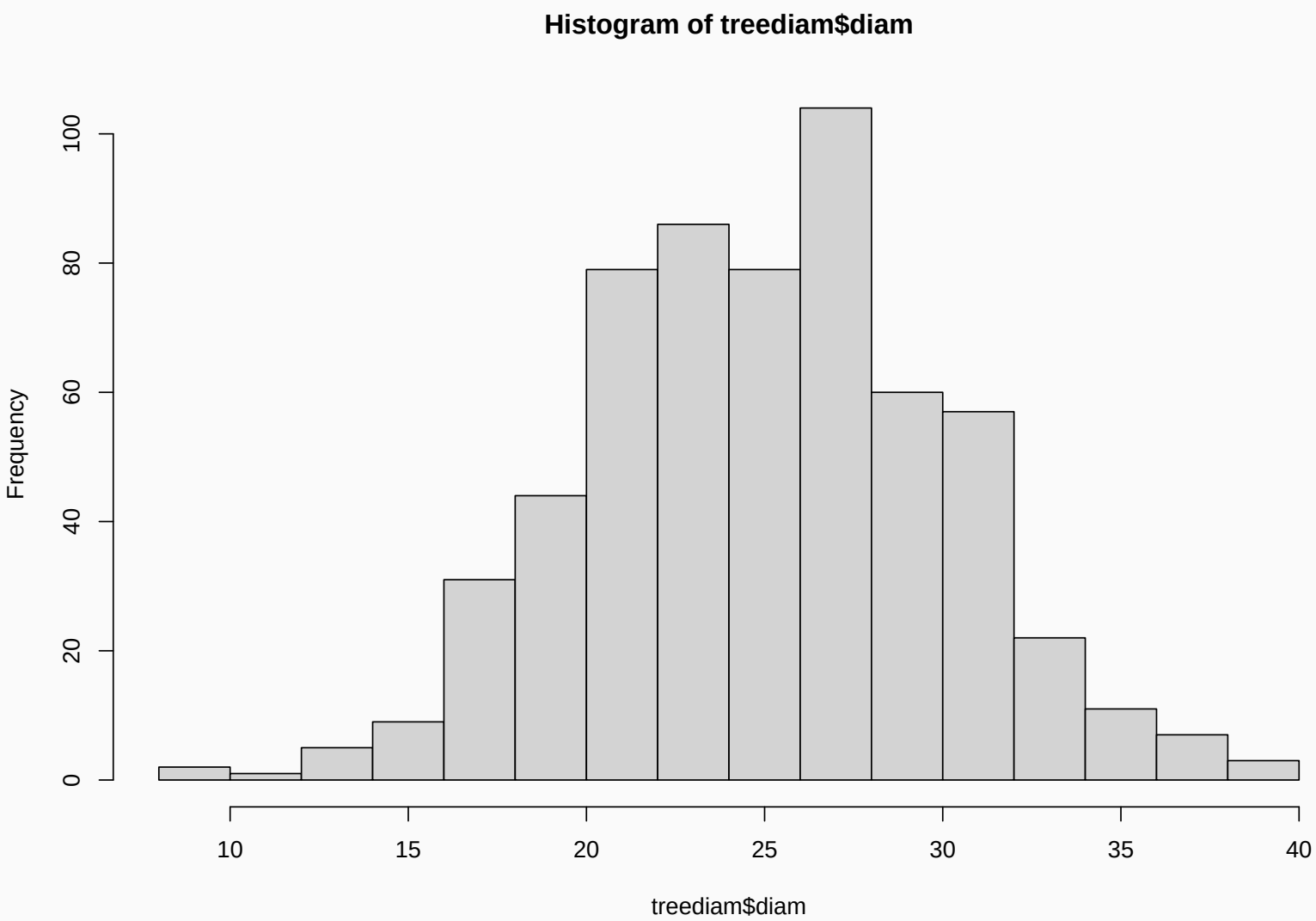
Visualisation of tree diameters

```
plot(treediam$diam)
```



Visualisation of tree diameters

```
hist(treediam$diam)
```



How well do these values

represent actual tree diameters

in your neighbourhood?

https://pollev.com/franciscorod726

- At what height did you measure?

- At what height did you measure?
- Did you include bark?

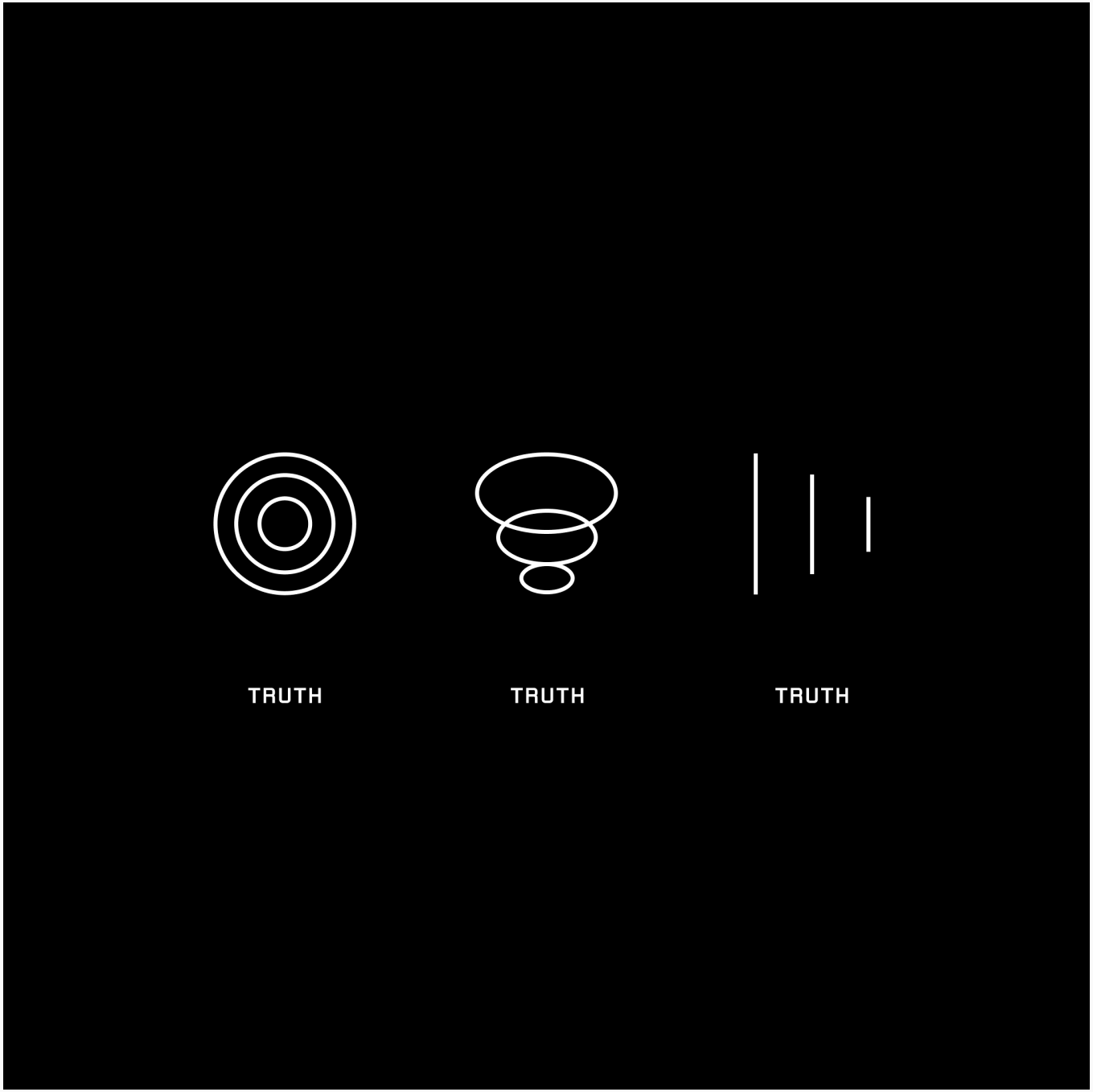
- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
    - (trees may get thinner w/ high evapotranspiration)

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
    - (trees may get thinner w/ high evapotranspiration)
- Where did you measure?

- At what height did you measure?
- Did you include bark?
- Did you measure with tape, caliper, by eye?
- When did you measure: dawn, midday, night?
    - (trees may get thinner w/ high evapotranspiration)
- Where did you measure?
    - (differences among streets, species, etc)



https://visualizevalue.com/

Data are hardly ever objective.

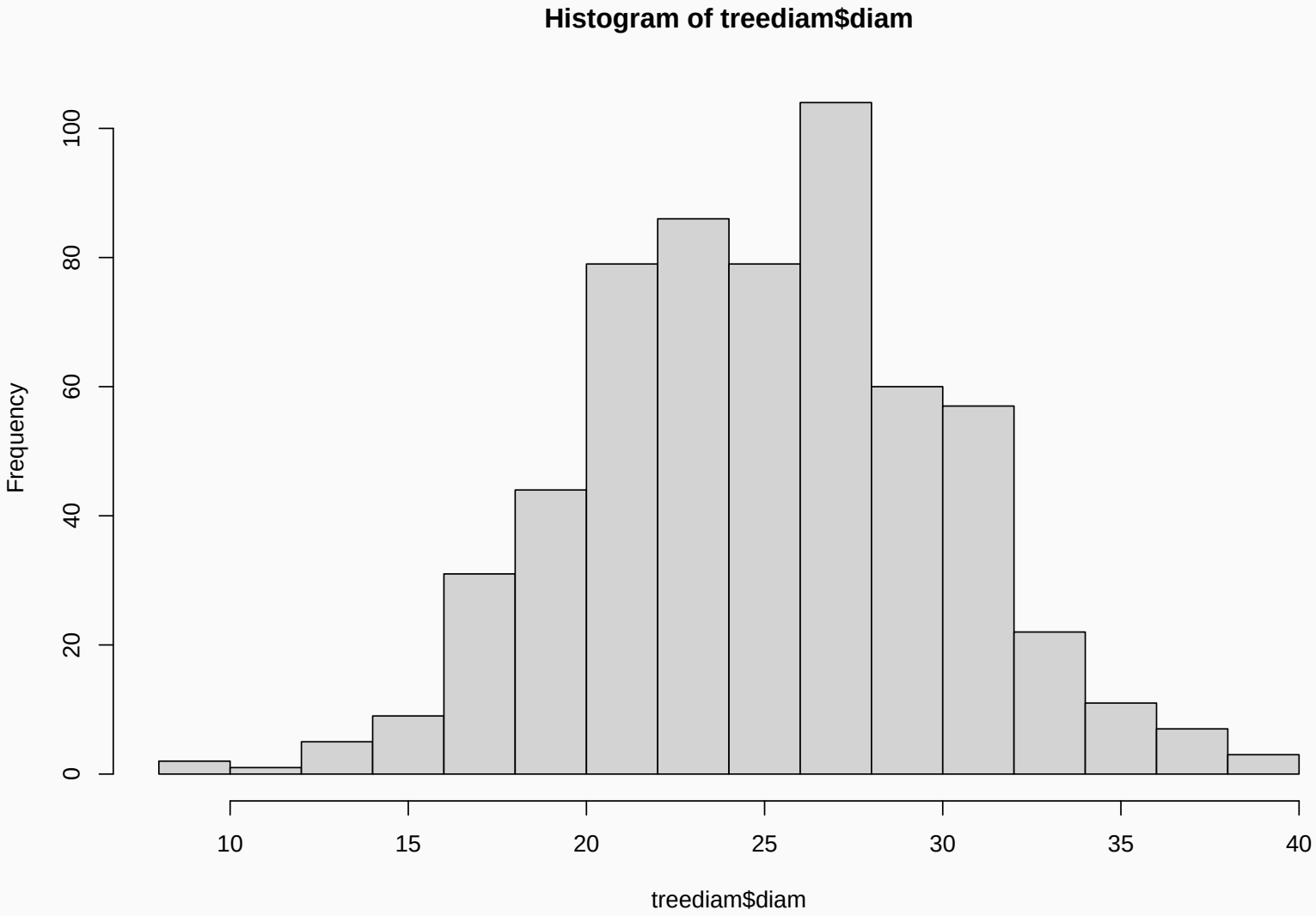
We decide **what to measure, when, where, and how**.

Always consider:

How well do data reflect what we are trying to measure?

Describing your data

How would you describe this distribution?

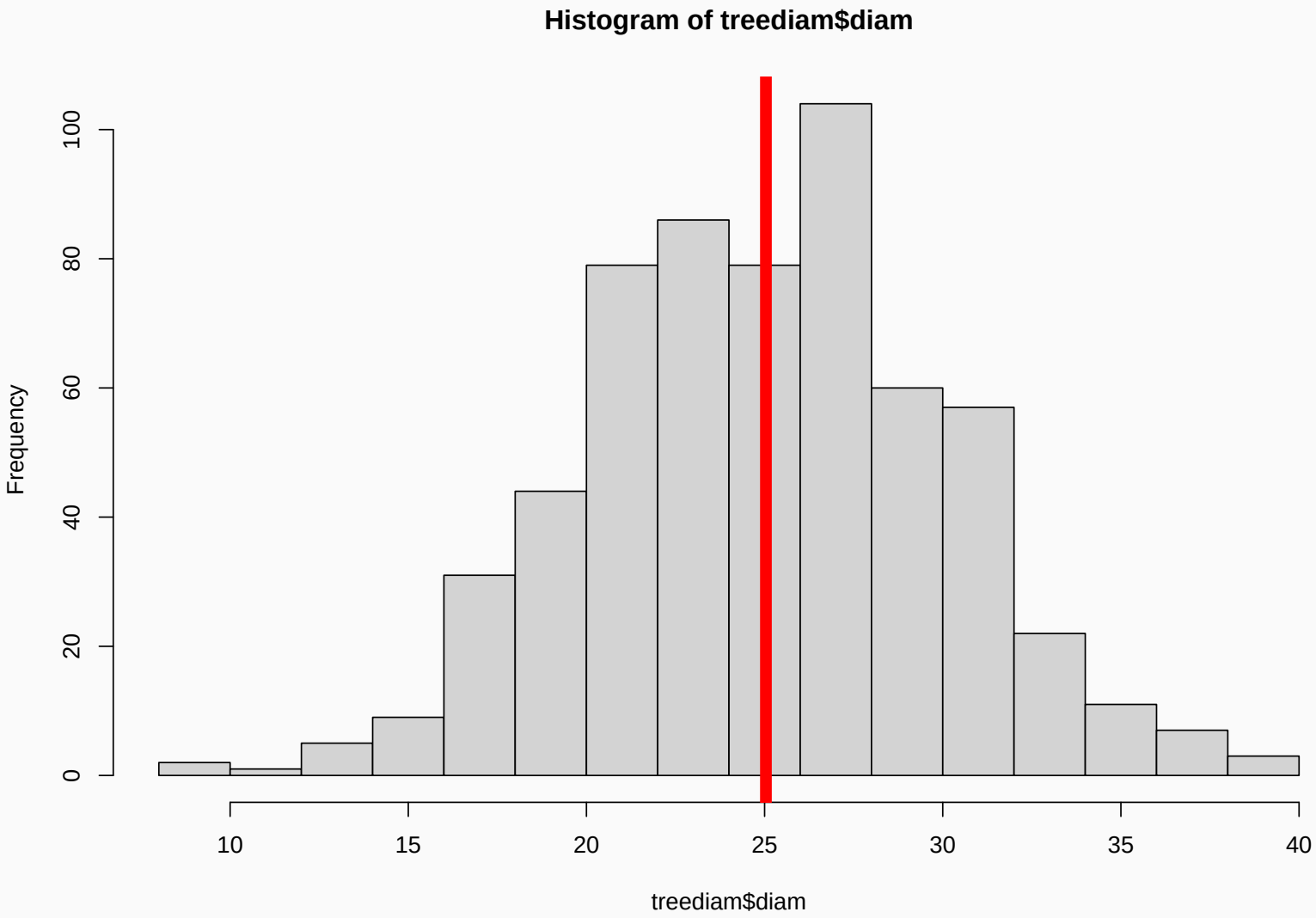


(Discuss with your partner)

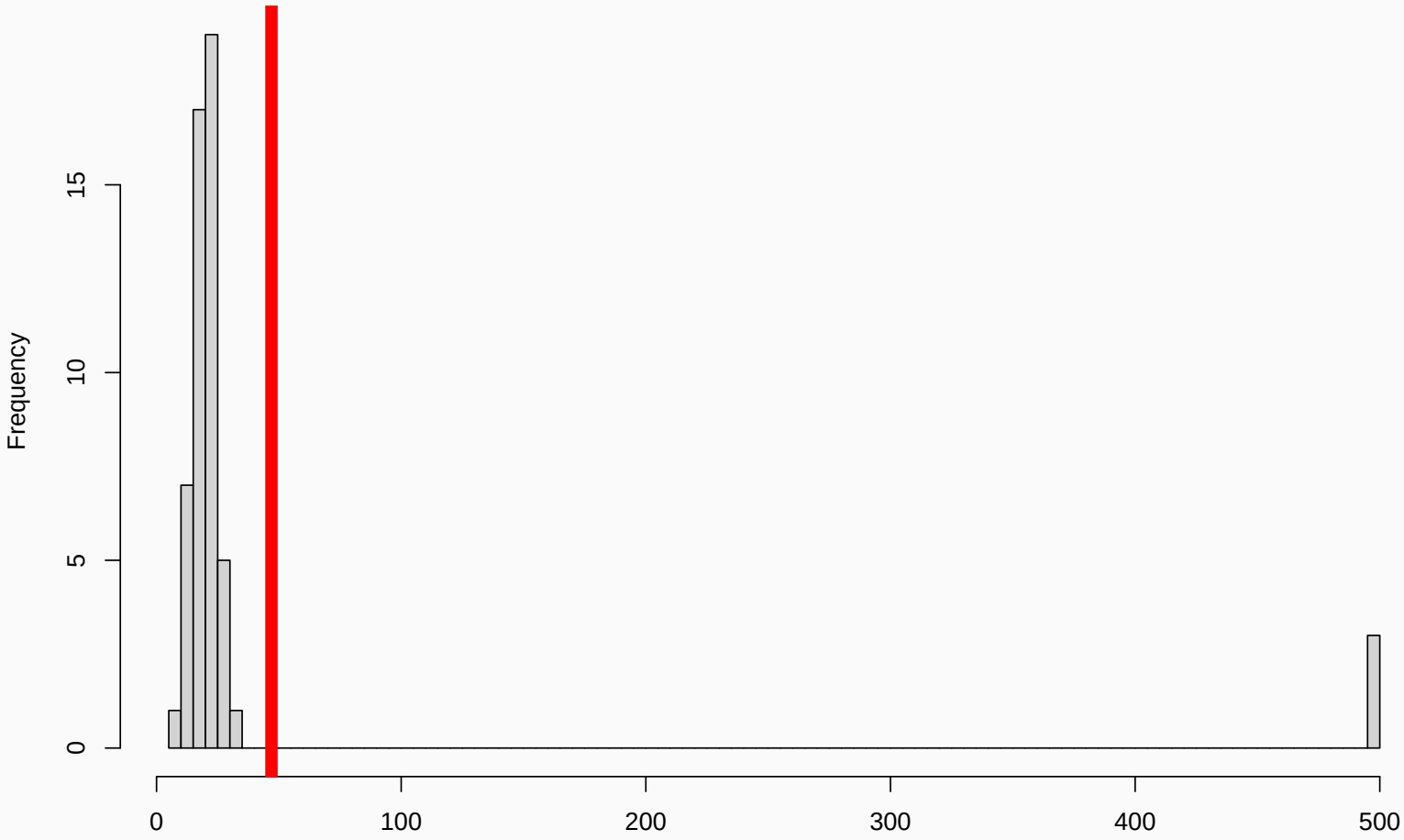
Location / Central tendency

Mean / Average

$$\text{mean} = \frac{d_1 + d_2 + d_3}{n}$$

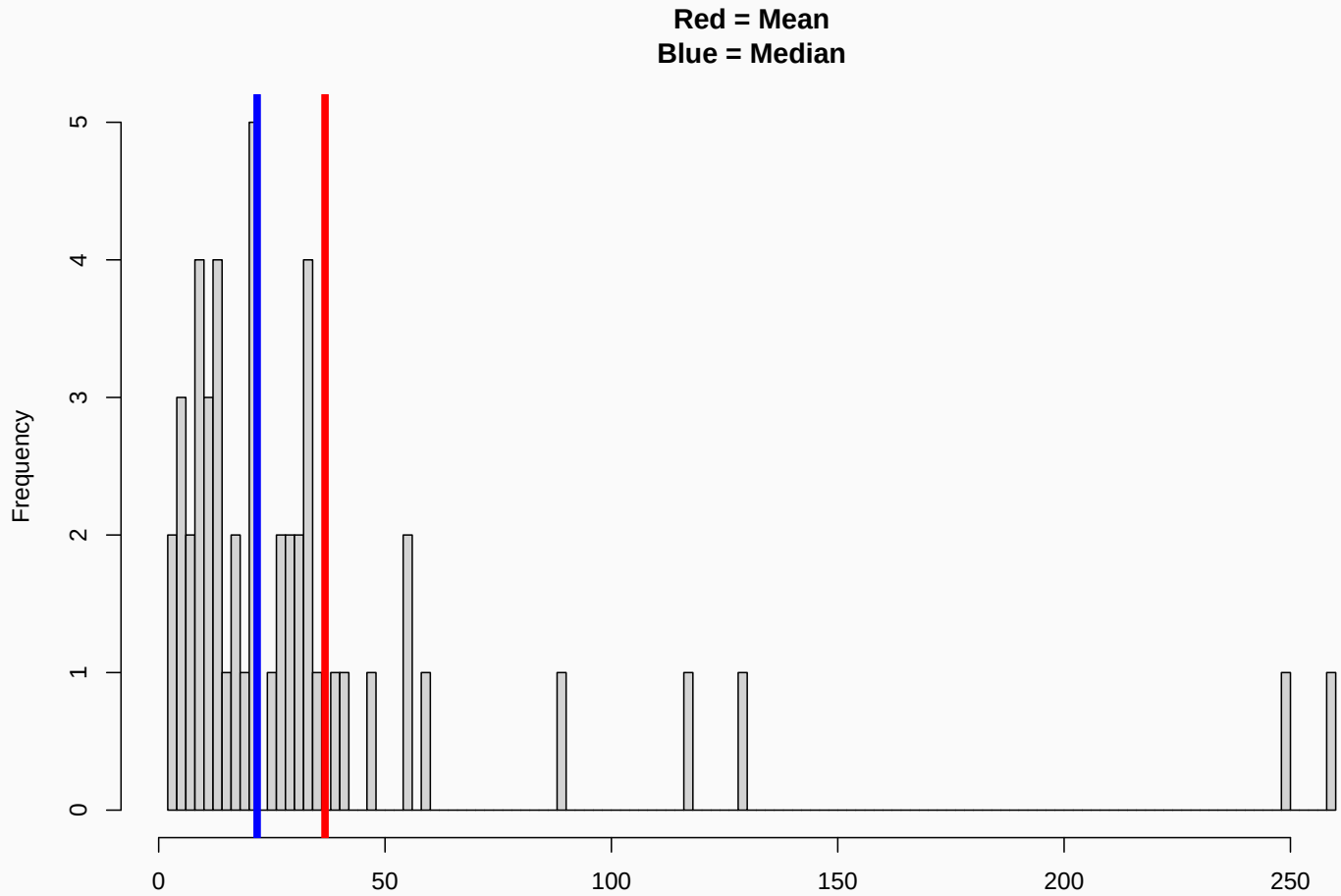


Mean is sensitive to skew/outliers



Median

50% percentile. Leaves half of the data values on each side

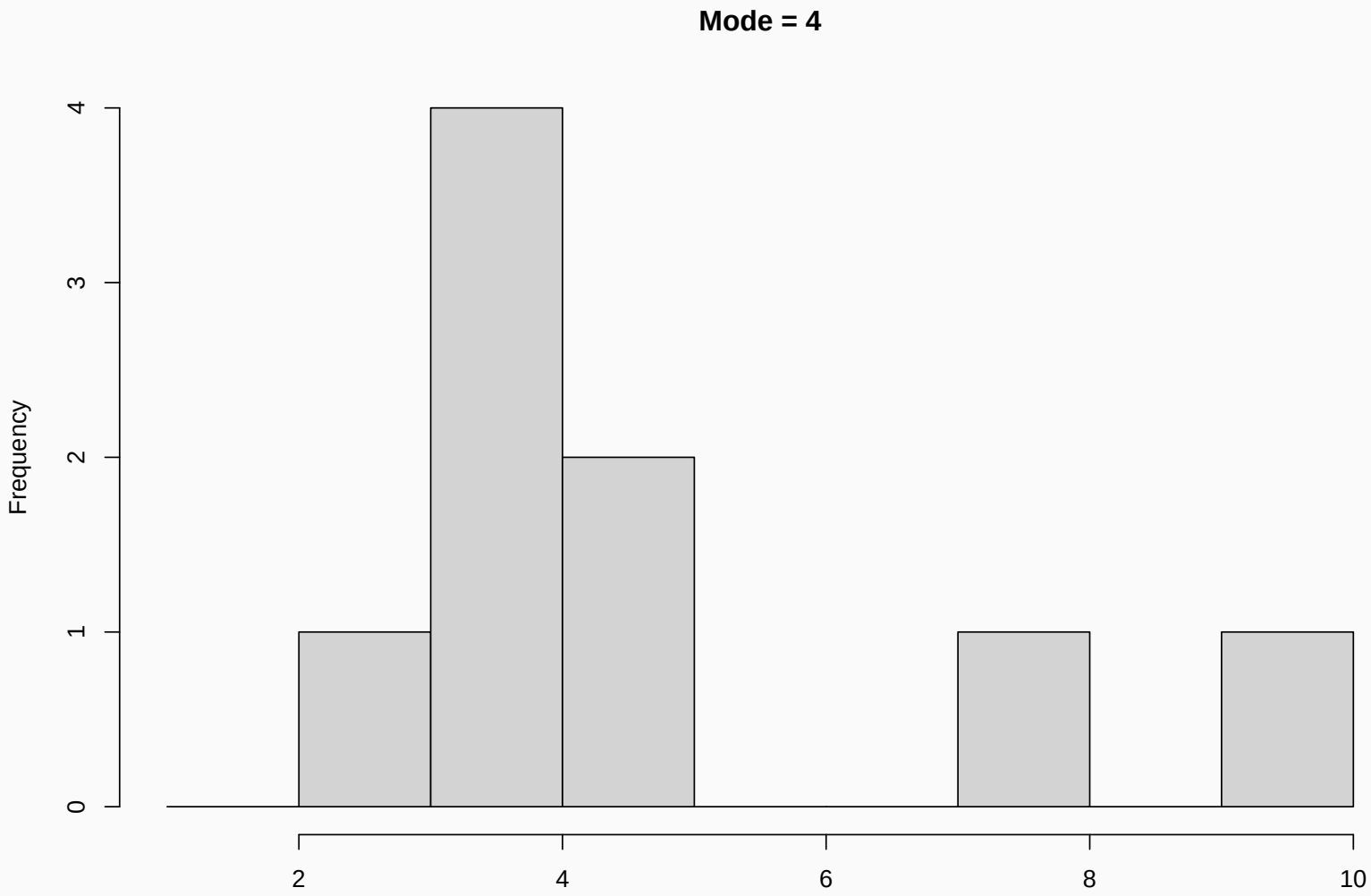


Median of c(2, 4, **6**, 8, 10) = 6

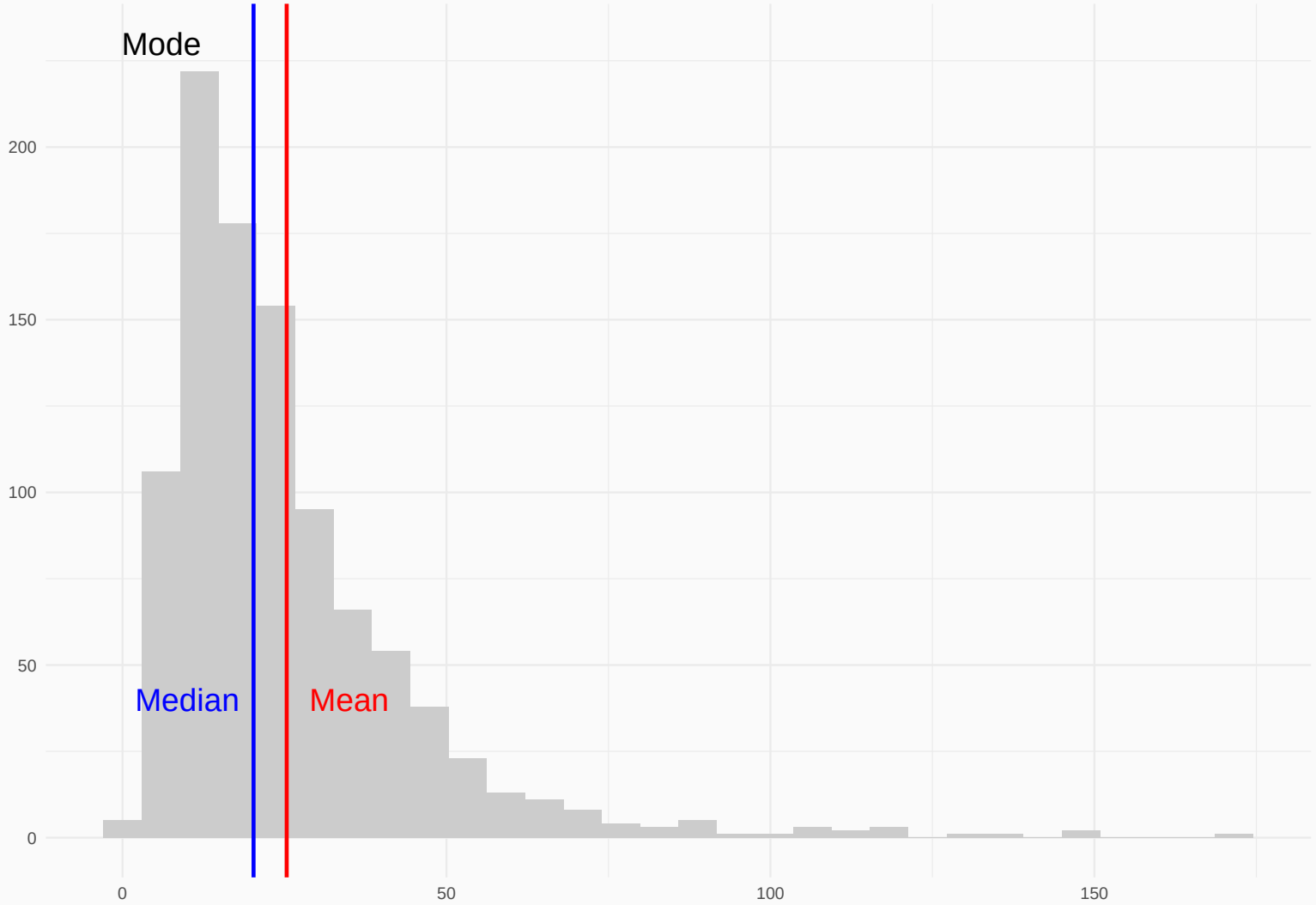
Median of c(2, 4, 6, 8) = (4 + 6) / 2 = 5

Mode

Most frequent value

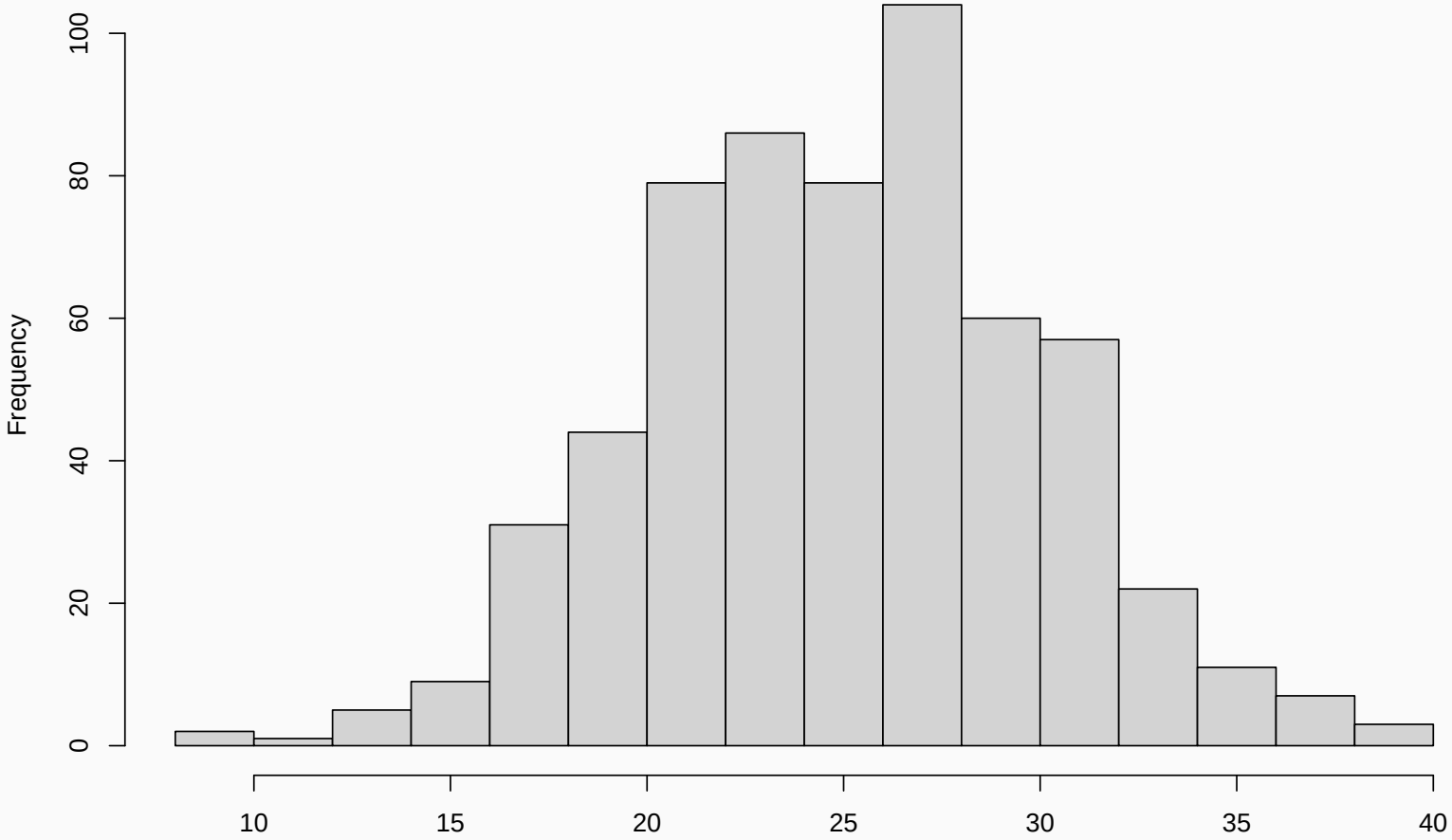


Describing the location / central tendency



Describing Variation / Spread

Minimum, Maximum, Range



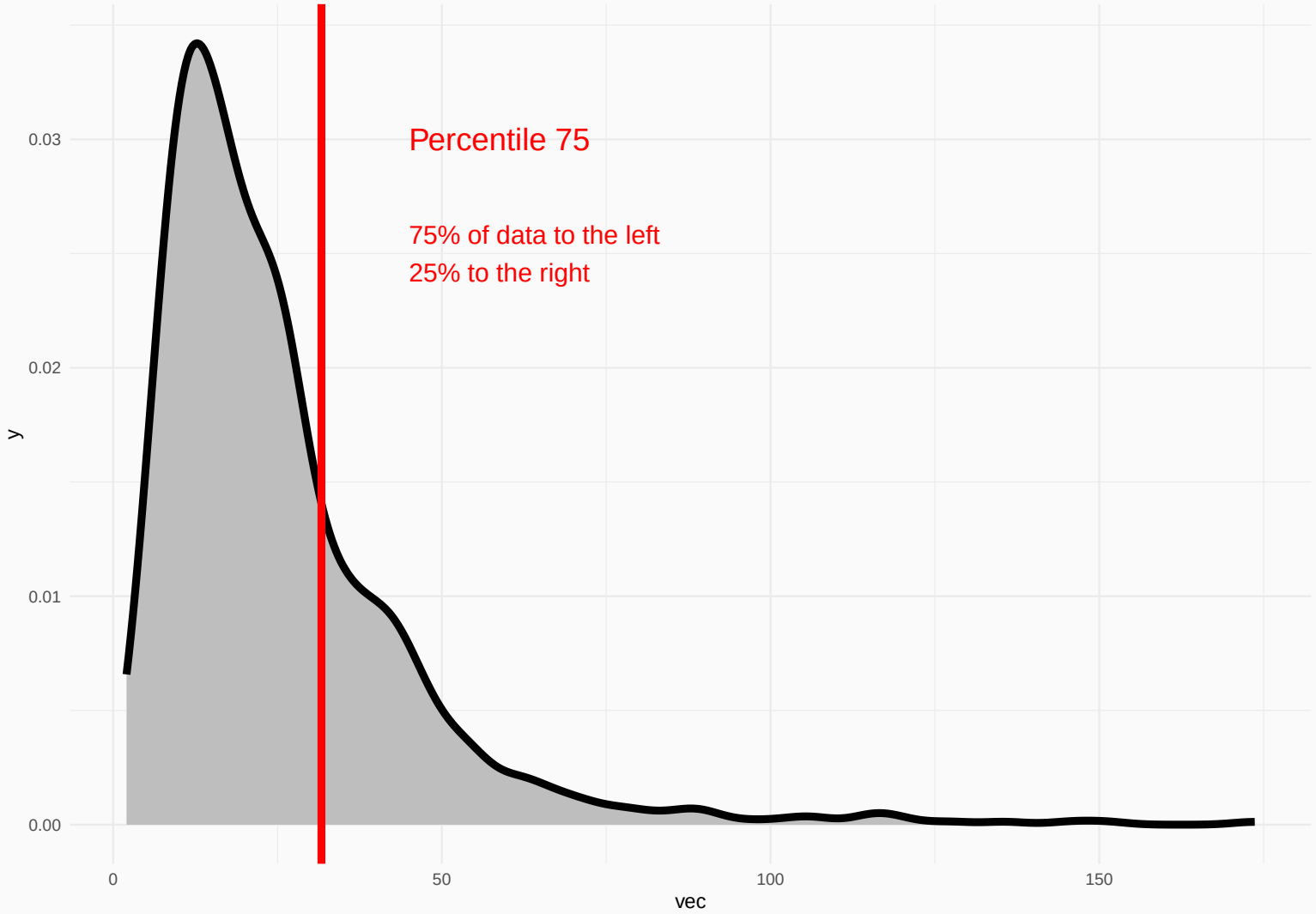
Minimum = 9.2

Maximum = 41.9

Range = 9.2, 41.9

Quantiles

Quartiles, Percentiles...

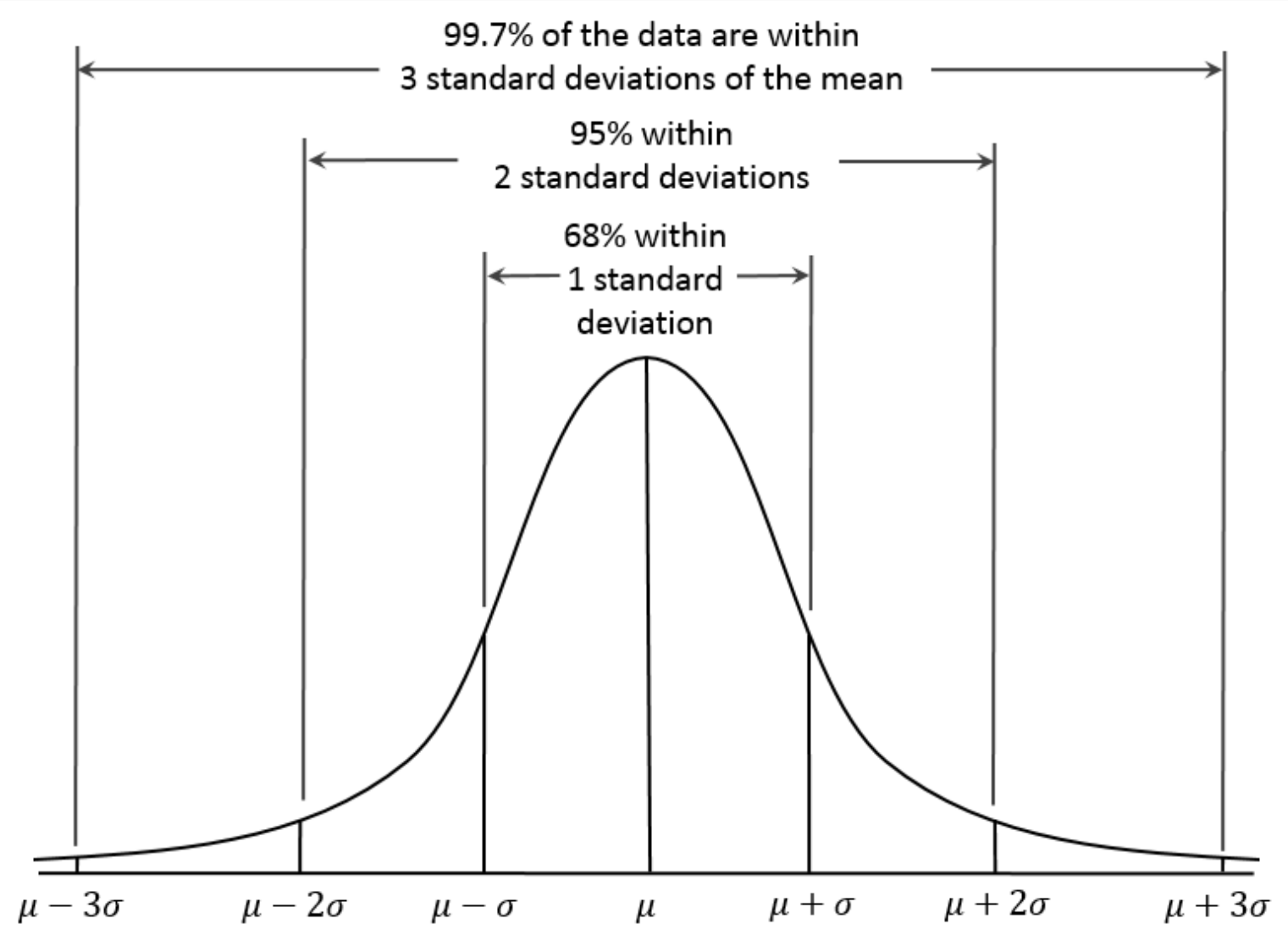


Standard Deviation

Average distance between data points and the mean

$$SD = \sqrt{\frac{\sum (x - \mu)^2}{n - 1}}$$

In a Normal distribution

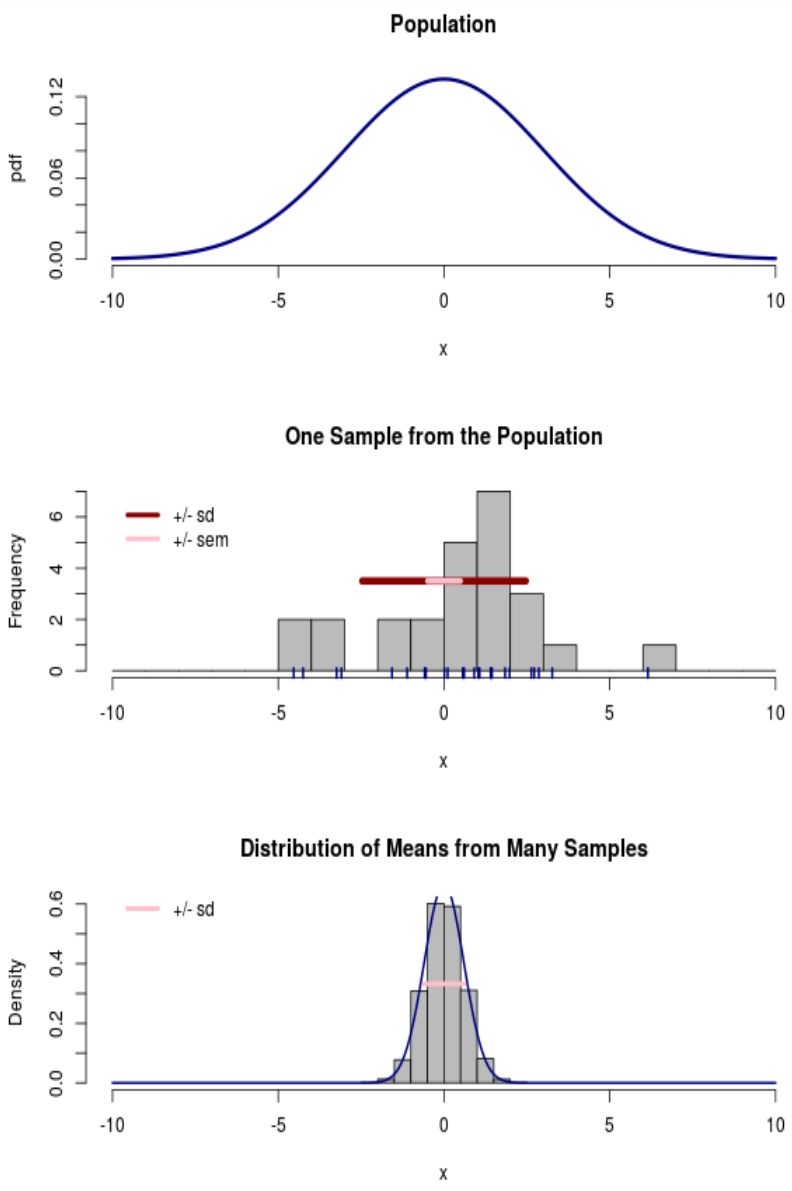


Standard Error of the Mean

$$SEM = \frac{SD}{\sqrt{n}}$$

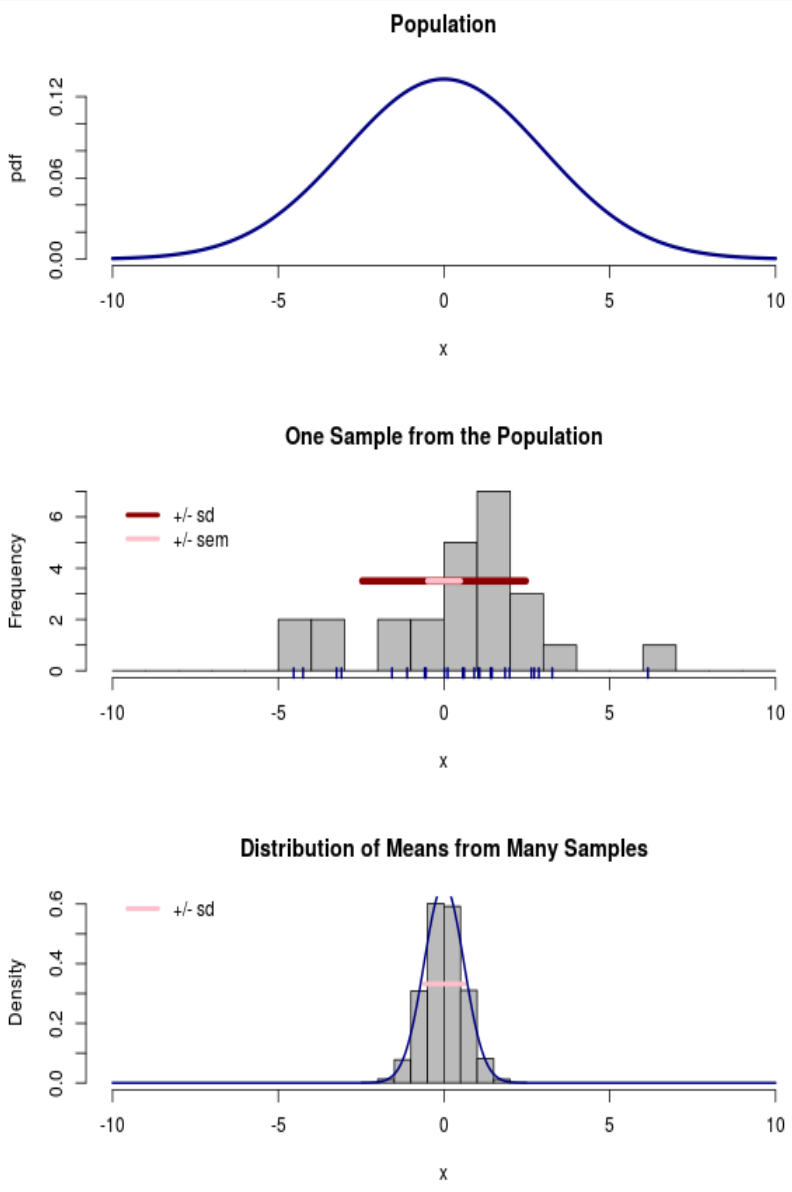
Estimates uncertainty (spread) of the parameter ‘mean’

Relationship between SD and SEM



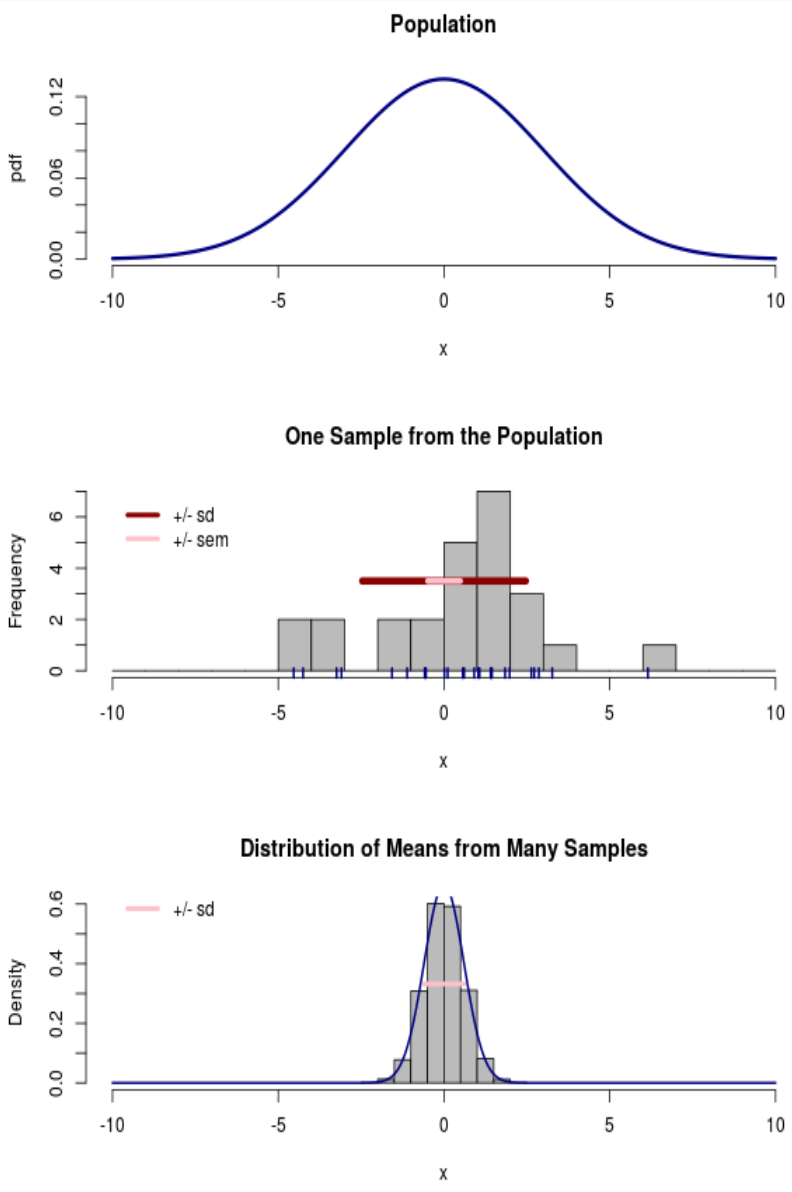
- SD quantifies scatter in population

Relationship between SD and SEM



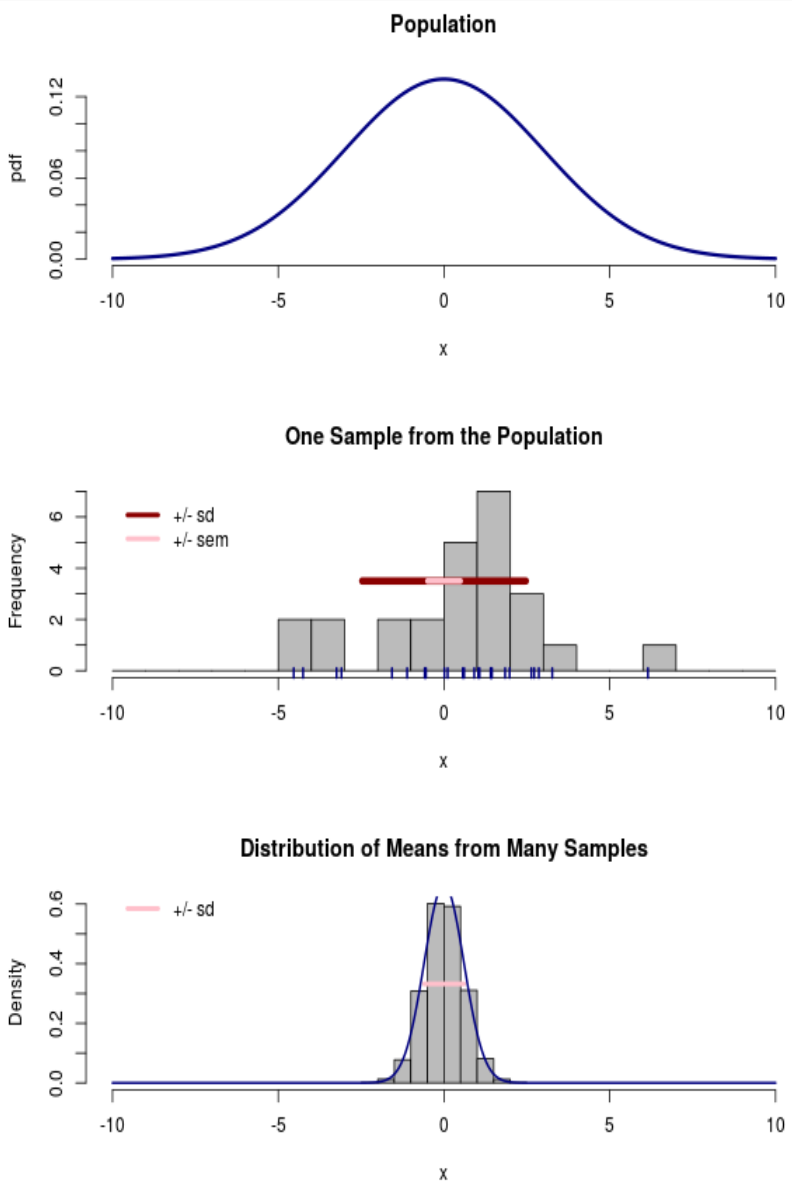
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)

Relationship between SD and SEM



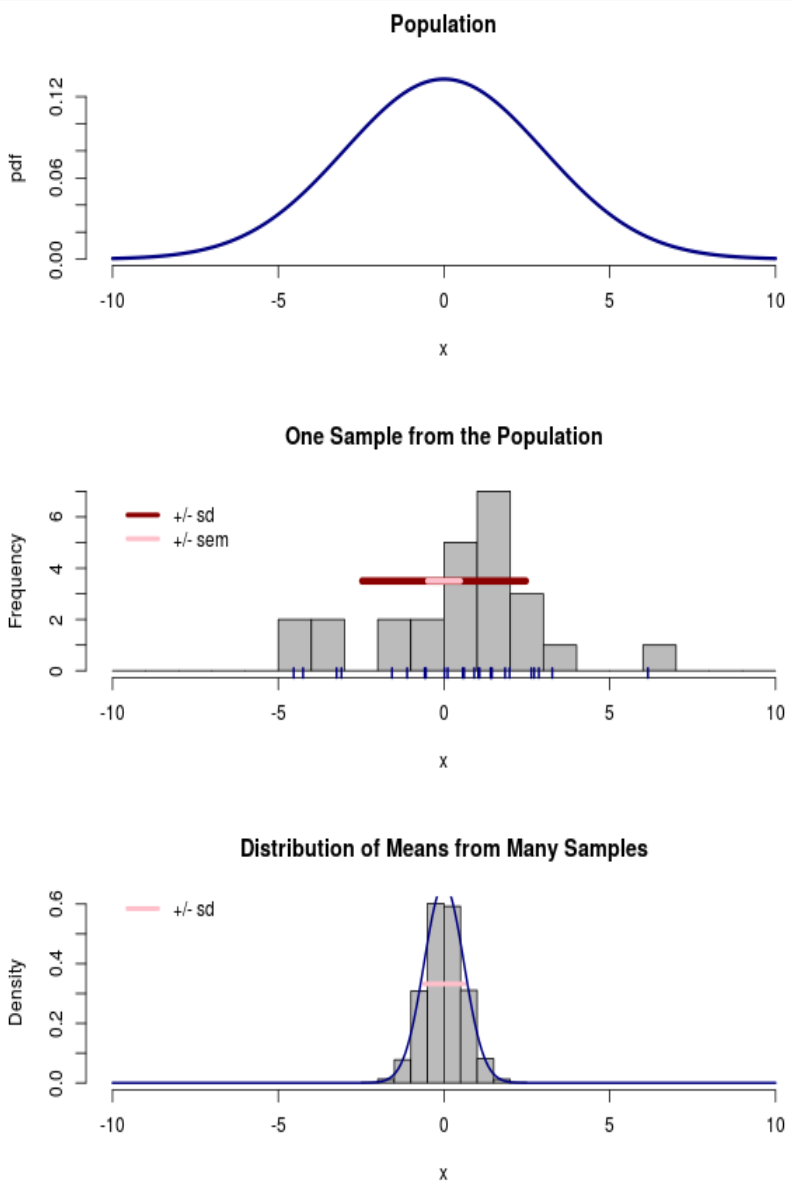
- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)

Relationship between SD and SEM



- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)
- SEM decreases with sample size (mean better known), SD does not.

Relationship between SD and SEM

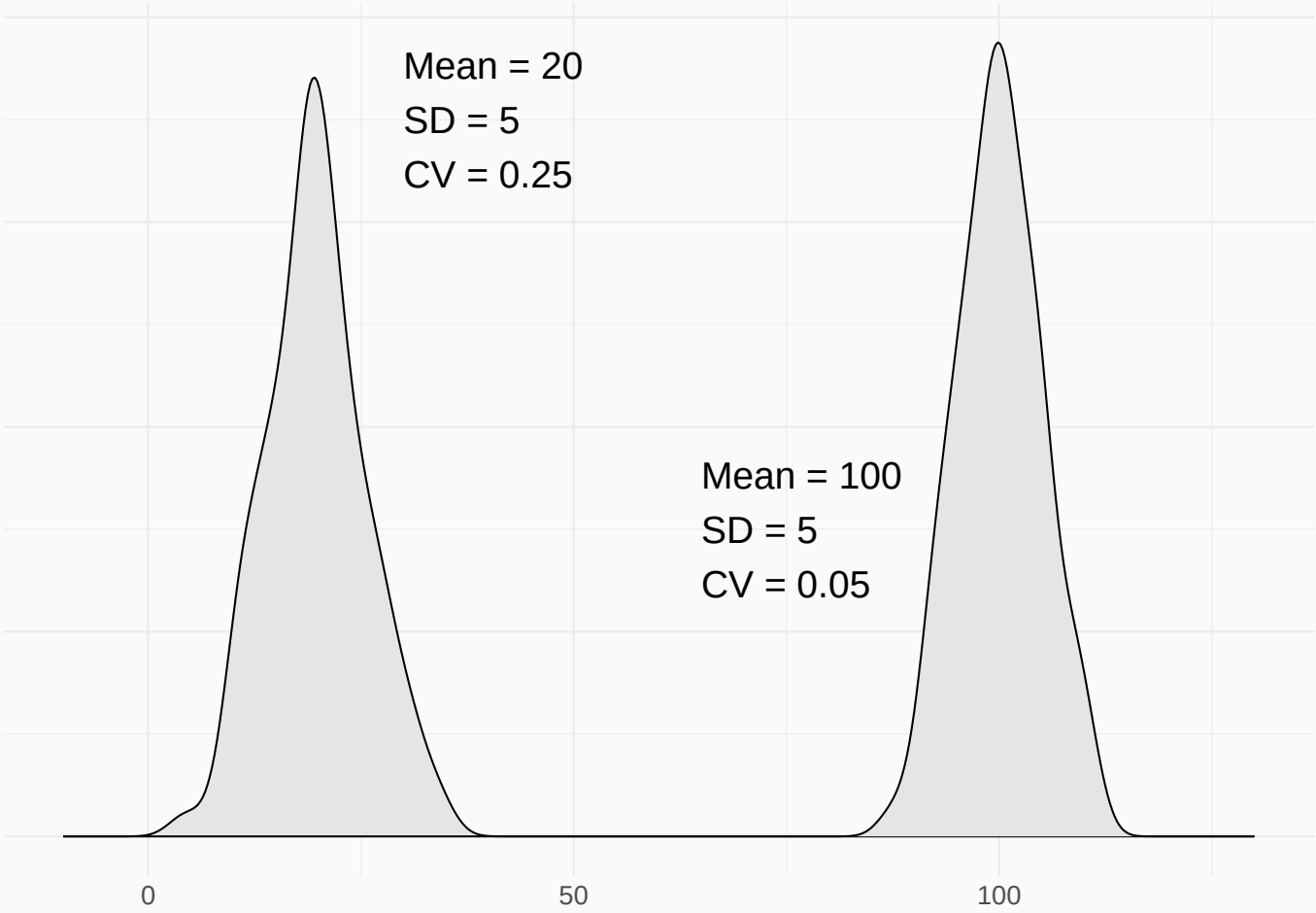


- SD quantifies scatter in population
- SEM quantifies uncertainty in parameter estimate (population mean)
- SEM = SD/sqrt(n)
- SEM decreases with sample size (mean better known), SD does not.
- https://gallery.shinyapps.io/sampling_and_stderr/

Coefficient of Variation

Facilitates comparing spread of distributions with different means

$$CV = \frac{SD}{mean}$$

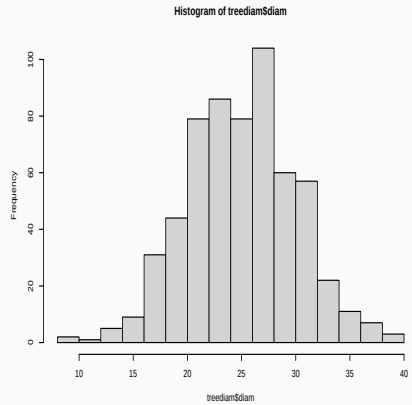


Summarise a distribution

Central tendency / location

- mean (average)

Variation / Spread

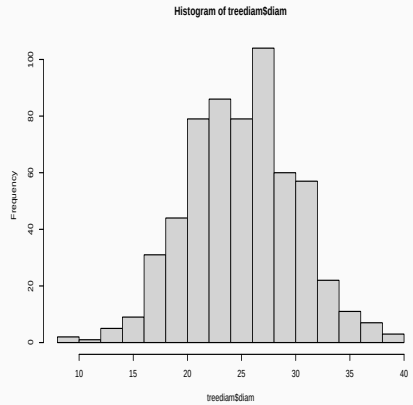


Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)

Variation / Spread

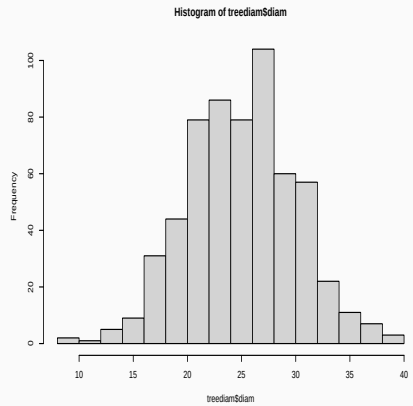


Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread



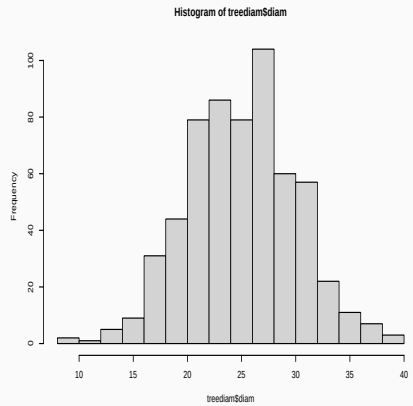
Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread

- min, max, range



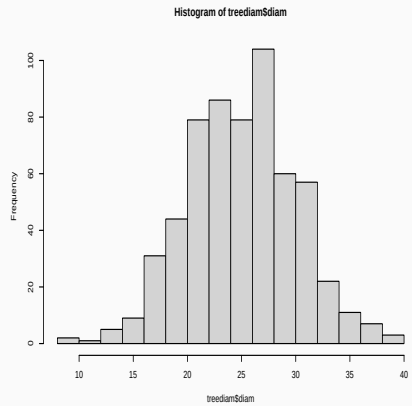
Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)



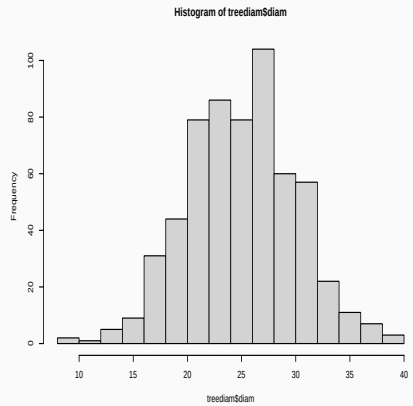
Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation



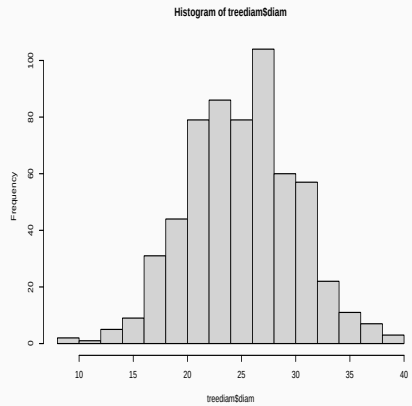
Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation
- standard error of the mean



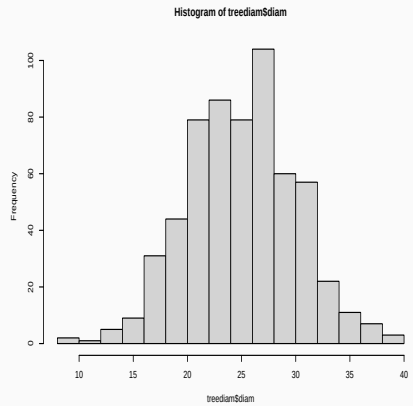
Summarise a distribution

Central tendency / location

- mean (average)
- median (50% percentile)
- mode (most frequent value)

Variation / Spread

- min, max, range
- quantiles (quartiles, percentiles...)
- standard deviation
- standard error of the mean
- coefficient of variation



What statistical descriptors are best? (and why)

https://pollev.com/franciscorod726