Stat 23400 Lecture 11: Introduction to Statistical Inference, Exploratory Data Analysis

Outline for Today

- Topics:
  - Introduction to Statistical Inference (MMSA 7.1).
  - Exploratory data analysis.
    - Data Matrix, Cases, Variables (Section 1.2 in OIS);
    - Histogram (Section 2.1.3 in OIS);
    - Numerical Summary: Mean and Median (Section 2.1 in OIS);
    - Measuring Variability (Section 2.1.4 in OIS);
    - Five Number Summary and Boxplot (Section 2.1.5 in OIS);
    - Density Curves (Section 3.5.1 in OIS);
    - Checking for normality and Q-Q plot;
- Lab06: Descriptive Statistics and Exploratory Data Analysis
- Hw6 will be posted on Canvas.

Introduction to Statistical Inference

Population and Sample

Statisticians are often called upon to study characteristics of a large group of people/countries/stuff, a population.

However, it is always either too expensive or simply impossible to measure every observational unit. For this reason, we usually study the characteristics of a sample.

Definition

- **Population** is the complete group (often infinite) of observational units, the entities on which you may collect measurements and features.
- **Sample** is a finite subgroup of the population on which you actually collect measurements and features.

Example: Population and Sample

In the Road Data Analysis, we studied *Loads resulting from stressful events on a vehicle suspension system over time*.

- **Population** :10,000 GM cars with the new suspension system.
- **Sample** : 65 selected cars.

If we can investigate the whole population, we can then simply calculate the mean, variance, etc.

In many applications, however, we can not observe the whole population. Instead, we take a sample from the population.

Random Sample

A **random sample** from the distribution of X is a collection, $X_1, \cdots, X_n$ of independent random variables, all with the same distribution of X.

- $X_1, \cdots, X_n$ are called i.i.d. samples (independent and identically distributed).
- For a finite population, this assumes sampling *with replacement*.
- Population is large.

Simple Random Samples

Suppose $X_1, X_2, \ldots, X_n$ are n draws at random **without replacement** from the finite population $\mathcal{P} = \{1, \ldots, N\}$. That is,

- In the first draw, every unit has $1/N$ chance to be selected
- In the second draw, each of the remaining $N-1$ has $1/(N-1)$ chance to be selected
- $\vdots$
- In the nth draw, each of the remaining $N-n+1$ has $1/(N-n+1)$ chance to be selected

Then $\{X_1, X_2, \ldots, X_n\}$ is called a **simple random sample (SRS)** of size n.

Properties of Simple Random Samples = i.i.d.

Suppose the population has distribution F.

- The X_i's are **identically distributed** since every $X_i \sim F$.
- The X_i's are **(nearly) independent**
  - Since we usually sample **without** replacement, draws are not independent.
  - As long as the sample size n is less than 10% of the population size N, the dependencies among sampled values are small and are generally ignored.
  - When sampling from an infinite population ($N = \infty$), the X_i's are independent. (e.g., sampling from a production line.)

For these reasons, when **simple random sample**, we often assume

$$X_1, X_2, \ldots, X_n \ \textbf{i.i.d.} \sim F\,.$$

Statistical Inference

- Suppose $X_1, \cdots, X_n$ is a random sample from the distribution of X. Because they result from random choices, we think of X_i's as random variables and apply probability results.
- After the items are selected and measured, we can think of them as numbers $\{x_1, \cdots, x_n\}$ (lower cases), which is a realization of $\{X_1, X_2, \ldots, X_n\}$.
- **Descriptive statistics**: summarize the information in the data by numbers, graphs, etc.
- **Inferential statistics**: draw conclusions about the population X.
  - Estimation of the mean.
  - Confidence interval for the mean.
  - Calculation of the sample size.
  - Testing of hypotheses.

Example: Sample and Statistic

In general, a **statistic** is a random variable $T = T(X_1, X_2, \ldots, X_n)$, which is a function of $X_1, X_2, \ldots, X_n$.

- A statistic cannot involve any *unknown* parameter, for example, $\overline{X} - \mu$ is not a statistic if the population mean μ is unknown.
- Being a random variable, T has its own distribution, mean, variance, etc...This distribution of T allows us to determine the accuracy and reliability of our estimate.
  - We use the **sample mean** $\overline{X}$ to estimate the population mean μ,
  - the **sample variance** S^2 to estimate the population variance σ^2,

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{(n-1)}.$$

Summary of Statistical Inference

- In probability, we assume $X_1, \ldots X_n \sim F(x \mid \theta)$, where θ is the known parameter.
- In statistical inference, we collect data $x_1, \ldots x_n$ as realizations of the random variables $X_1, \ldots, X_N$, and we try to get information about θ using the data.

	Probability	Statistical Inference
θ	Fixed and known	Unknown
X	Random and unknown	Observed Random Realizations

Data Matrix, Cases, Variables

Data Matrix

Example: Data Set `email50`

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Cases

	spam	num_char	line_breaks	format	number	
1	no	21,705	551	html	small	
2	no	7,011	183	html	big	
3	yes	631	28	text	none	← case
⋮	⋮	⋮	⋮	⋮	⋮	
50	no	15,829	242	html	small	

Each row of data matrix corresponds to a **case**.

- In a study, we collect information — data — from **cases**. **Cases** can be individuals, corporations, animals, or any objects of interest.
- In the data matrix above, a case is an email.

Variables

variable ↓

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Each column of the data matrix contains the values of one **variable** of all cases.

- A **variable** is a characteristic of a case. A variable varies among cases.
  E.g., age, blood pressure, leaf length, first language

Example: Data set (email)

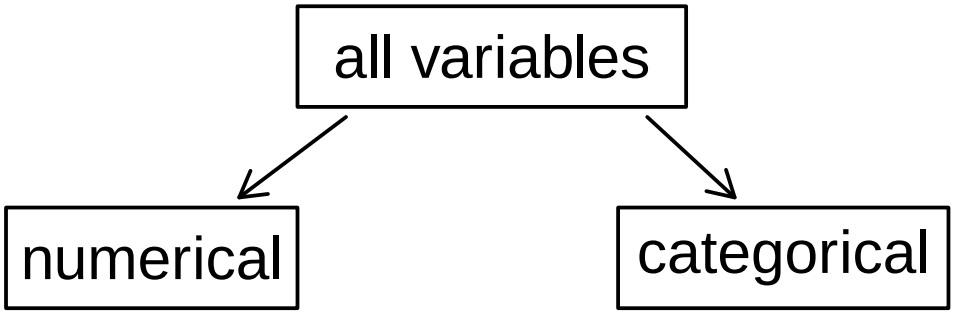
These data represent incoming emails for the first three months of 2012 for an email account.
Some variables:

- `spam` Indicator for whether the email was spam.
- `to_multiple` Indicator for whether the email was addressed to more than one recipient.
- `num_car` The number of characters in the email, in thousands.

Example: Data set (email)

```
data("email", package = "openintro")
glimpse(select(email, spam, to_multiple,
               viagra, num_char))

Observations: 3,921
Variables: 4
$ spam        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ to_multiple <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,...
$ viagra      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ num_char    <dbl> 11.370, 10.504, 7.773, 13.256, 1.23...
```

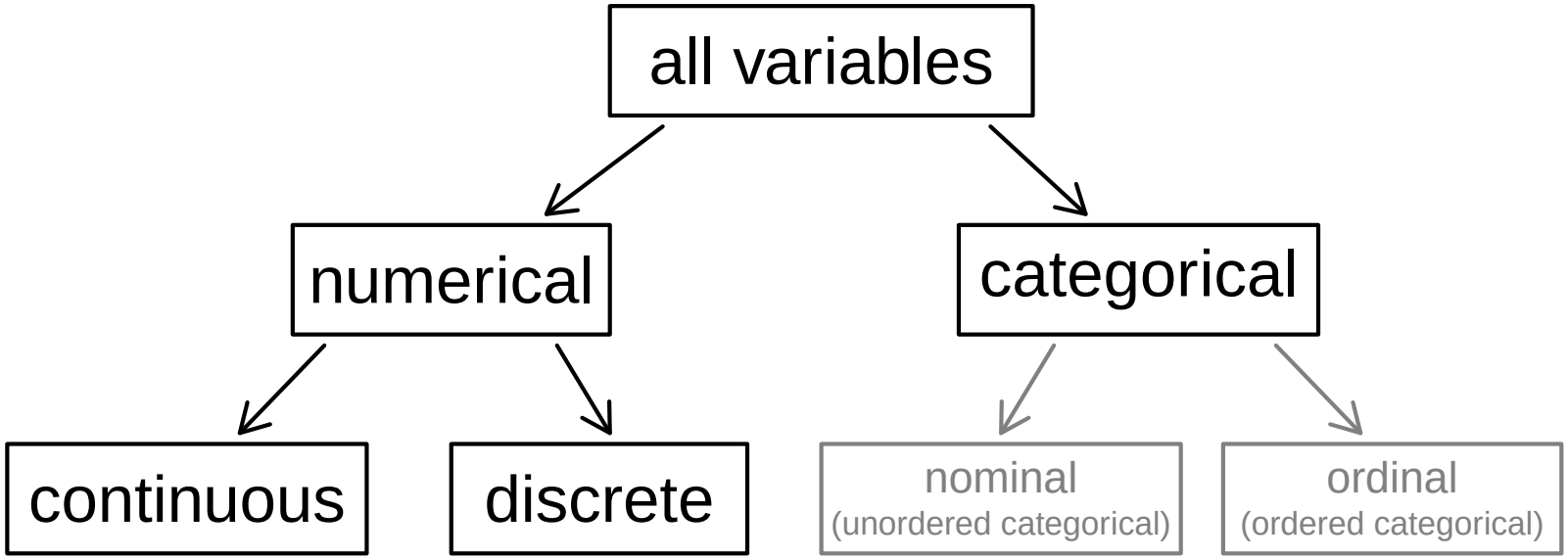
Types of variables



A variable is numerical when it can take a wide range of numerical values, and it is sensible to take arithmetic operations (addition, subtraction, average) with those values. Otherwise, it is categorical.

- e.g., Zip codes, area codes are NOT numerical variables

Types of variables



A numerical variable is

- discrete if its possible values form a set of separate numbers, such as 0, 1, 2, 3, ...
- continuous if its possible values form an interval.

A categorical variable with ordered categories is ordinal.

Types of variables (cont.)

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

- `spam` ..... categorical
- `num_char` ..... numerical
- `line_breaks` ..... numerical
- `format` ..... categorical, nominal
- `number` ..... categorical, ordinal

Sometimes an ordinal categorical variable can also be regarded as a numerical variable, e.g.,

- rating of a movie from 1 star to 5 stars
- stage of cancer from 0 to 4

Histogram

Distribution

Distribution

The distribution of a variable tells us what values it takes and how often it takes these values.

- How do we describe variables?
- How do we summarize their characteristics?
- What we are interested in is a variable's *distribution*.
- There are two main ways we describe the distribution of a variable: *graphically* or *numerically*.

Histogram

Histograms plot the frequencies (counts), percents, or proportions of equal-width classes of values.

E.g.

```
x <- c(1, 1.2, 2, 3, 3.5, 3.9)
```

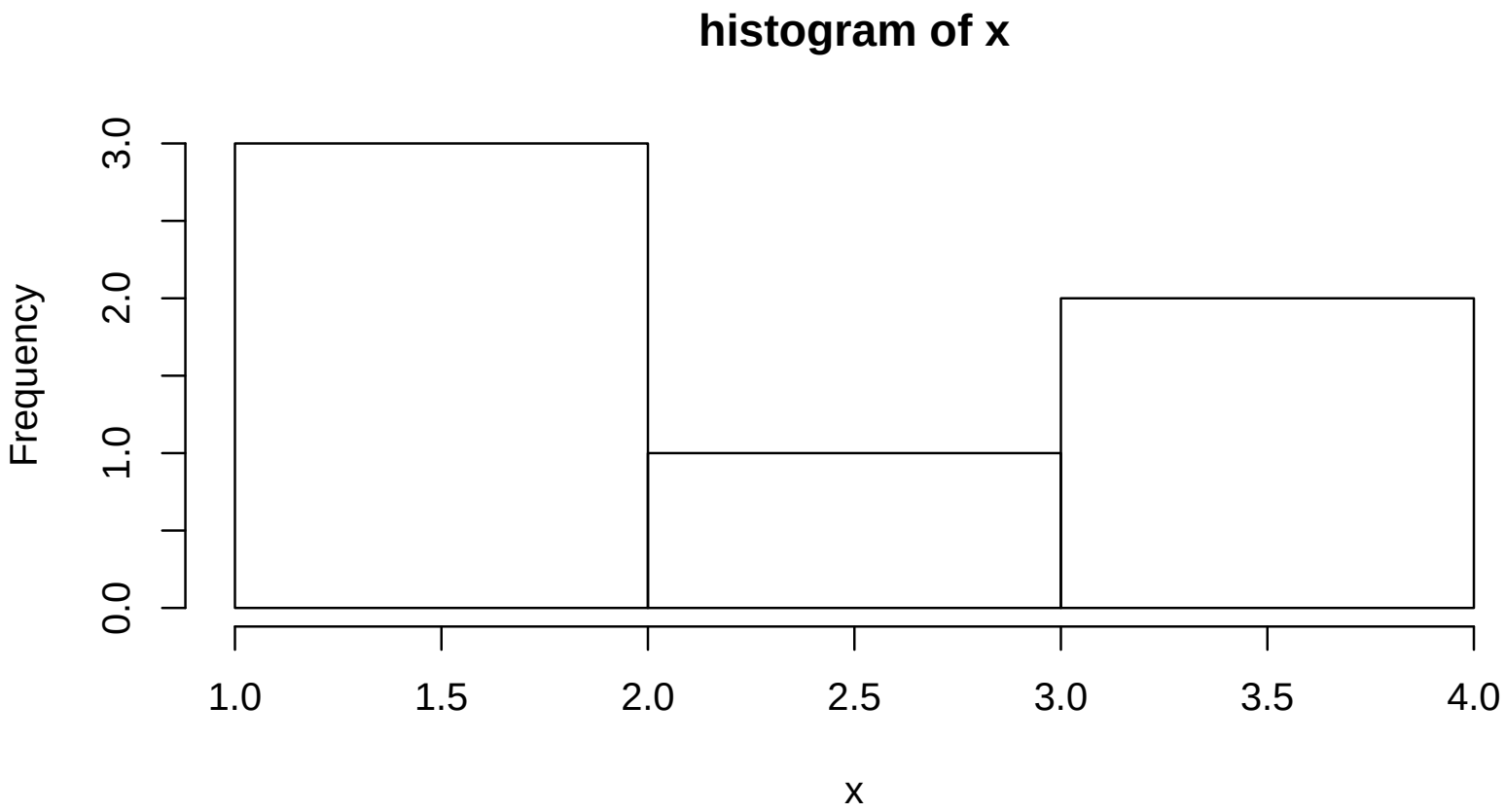
Bin the observations into one of three groups:

- group1 $= x : x \leq 2$
- group2 $= x : 2 < x \leq 3$
- group3 $= x : 3 < x \leq 4$

Then make a plot with bars where the height of each bar is proportional to the counts within each group.

histogram continued

```
hist(x, main = "histogram of x")
```



Example: Infant Mortality Rates Data

Data: Infant mortality rates (number of deaths under one year of age per 1000 live births) of 201 countries/regions in 2010-2015:

	Country.Region	Continent	X2010.2015
1	Burundi	Africa	77.9
2	Comoros	Africa	58.1
3	Djibouti	Africa	55.3
...			
200	Samoa	Oceania	19.7
201	Tonga	Oceania	20.4

https://en.wikipedia.org/wiki/List_of_countries_by_infant_mortality_rate

How to Make Histograms

Data file: infmort2015.txt

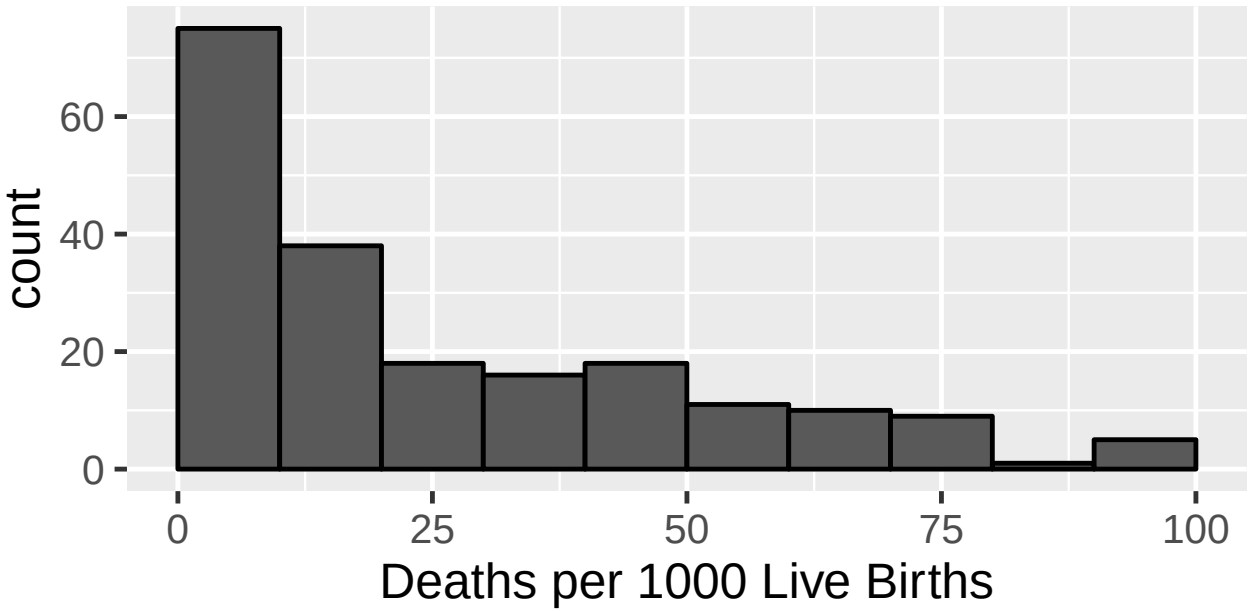
- Step 1: Divide the range of values into class intervals.
- Step 2: Count the number of values in each class interval.

Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Count	75	38	18	16	18	11	10	9	1	5

- Step 3: Draw the histogram
  - No space between bars.
  - Label the horizontal axes (with units)!

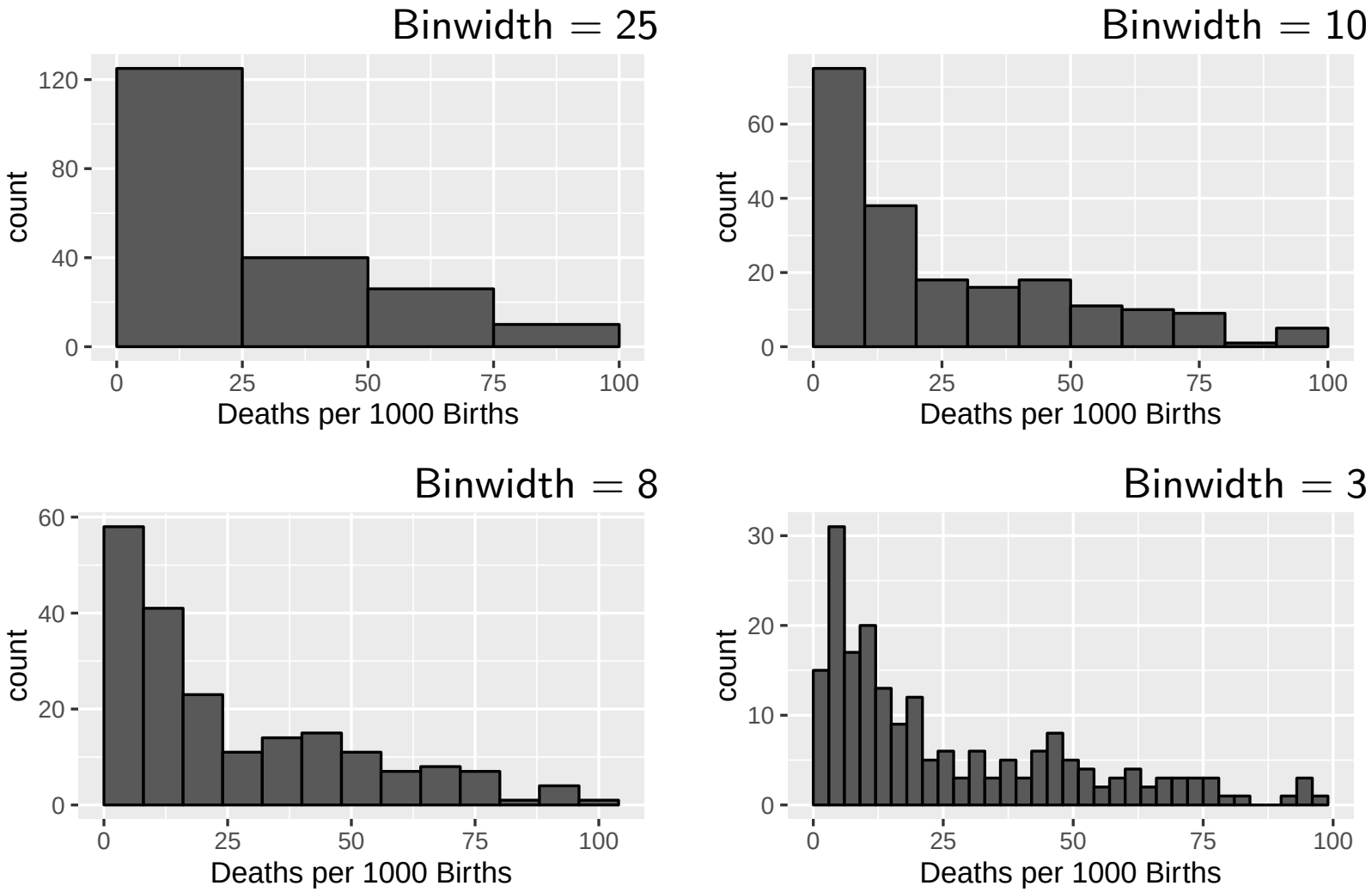
How to Make Histograms (Cont'd)

```
infmort <- read.table("../../data/infmort2015.txt")
hist(infmort$X2010.2015, breaks = 10, xlab =
  "Deaths per 1000 Births",   ylab ="count")
```



Try Different Binwidth

Which one(s) of these histograms reveal too much about the data? Which hide too much?



Selection of Binwidth

It is an iterative process — try and try again.

What bin with should you use?
General rule: **the more observations, the more bins.**

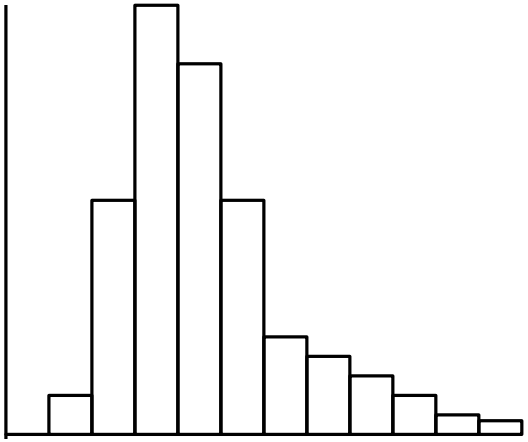
- Not too small that most bins have either 0 or 1 counts
- Not too big that you lose the details in a bin
- (There may not be a unique "perfect" bin size)

What to Look in a Histogram?

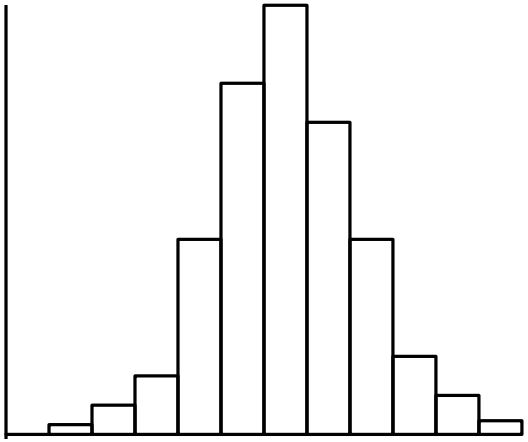
- **Shape**
  - symmetric or skewed (lopsided)
  - number of modes (peaks)
- **Outliers**: Are there any observations that lie outside the overall pattern? They could be unusual observations, or they could be mistakes. Check them!
- **Center**: Where is the "middle" of the histogram?
  - typically represented by mean and median
- **Spread**: What is the range of data?
  - typically represented by SD and IQR (will introduce soon)

Skewness of Histograms

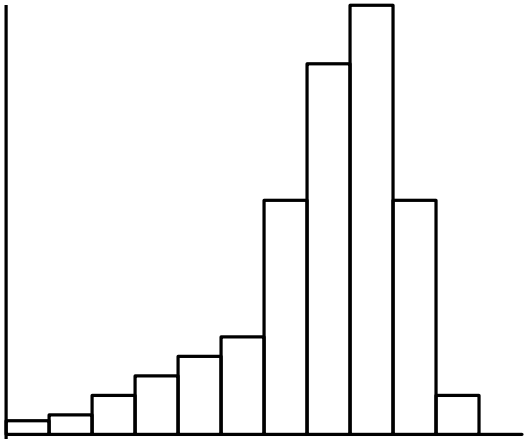
Right−skewed



Symmetric/Bell−shaped



Left−skewed



On Skewness and Symmetry

- Many physical measurements follow symmetric distributions: e.g. height or weight.
- Many variables are specifically designed to follow symmetric distributions: IQ test scores, SAT scores.
- Variables with boundaries tend to be skewed: e.g. income cannot be below zero so tends to be skewed right. Tweets have a max length of 140 characters, so tends to be skewed left.

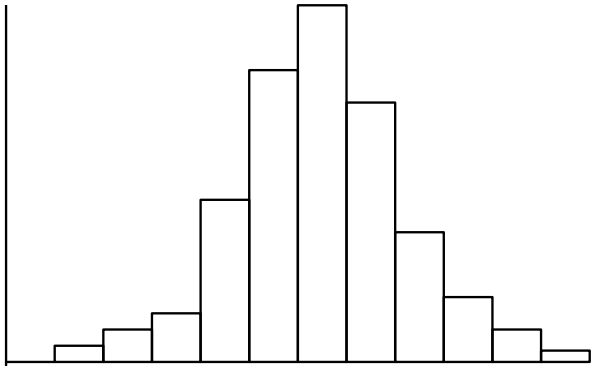
Mode

Mode

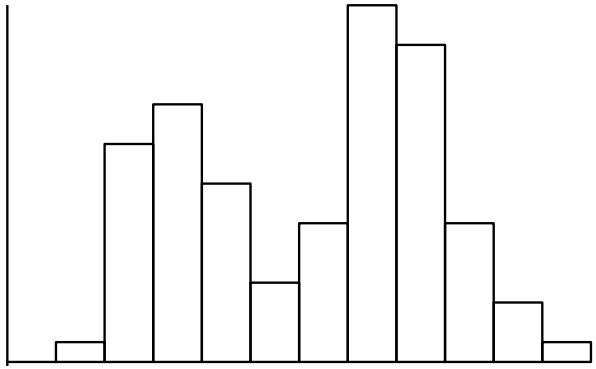
A mode is a prominent peak in a distribution. A distribution with one mode is unimodal. A distribution with two modes is bimodal. A distribution with more than one mode is multimodal.

Mode of Histograms

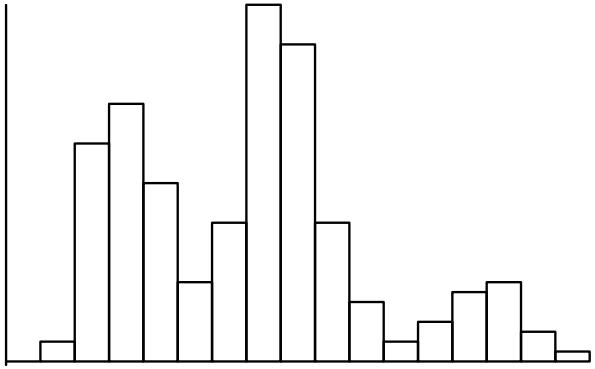
Unimodal



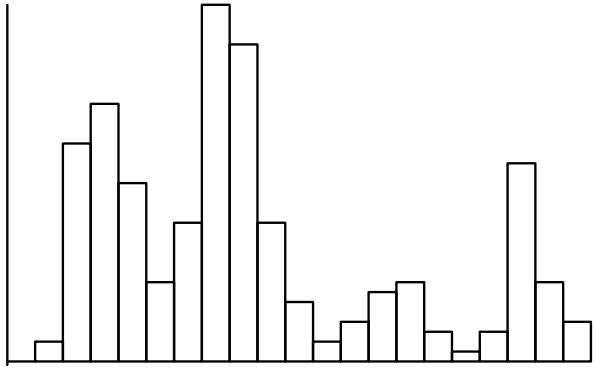
Bimodal



Trimodal



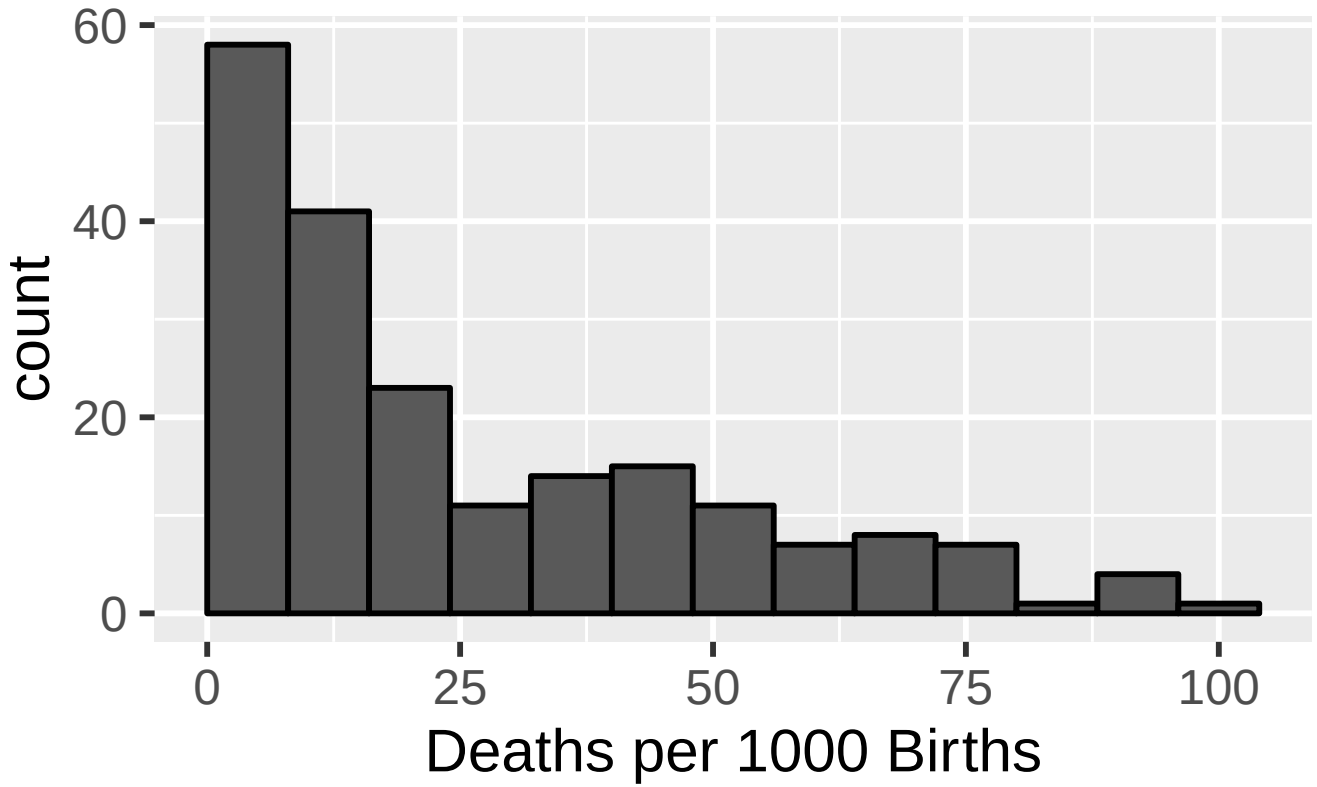
Multimodal



A histogram with two or more modes may indicate that the data is a mixture of two or more distinct populations.

Example (Infant Mortality Rates)

In addition to the major peak near 0, there appears to be a secondary peak around 40-50.

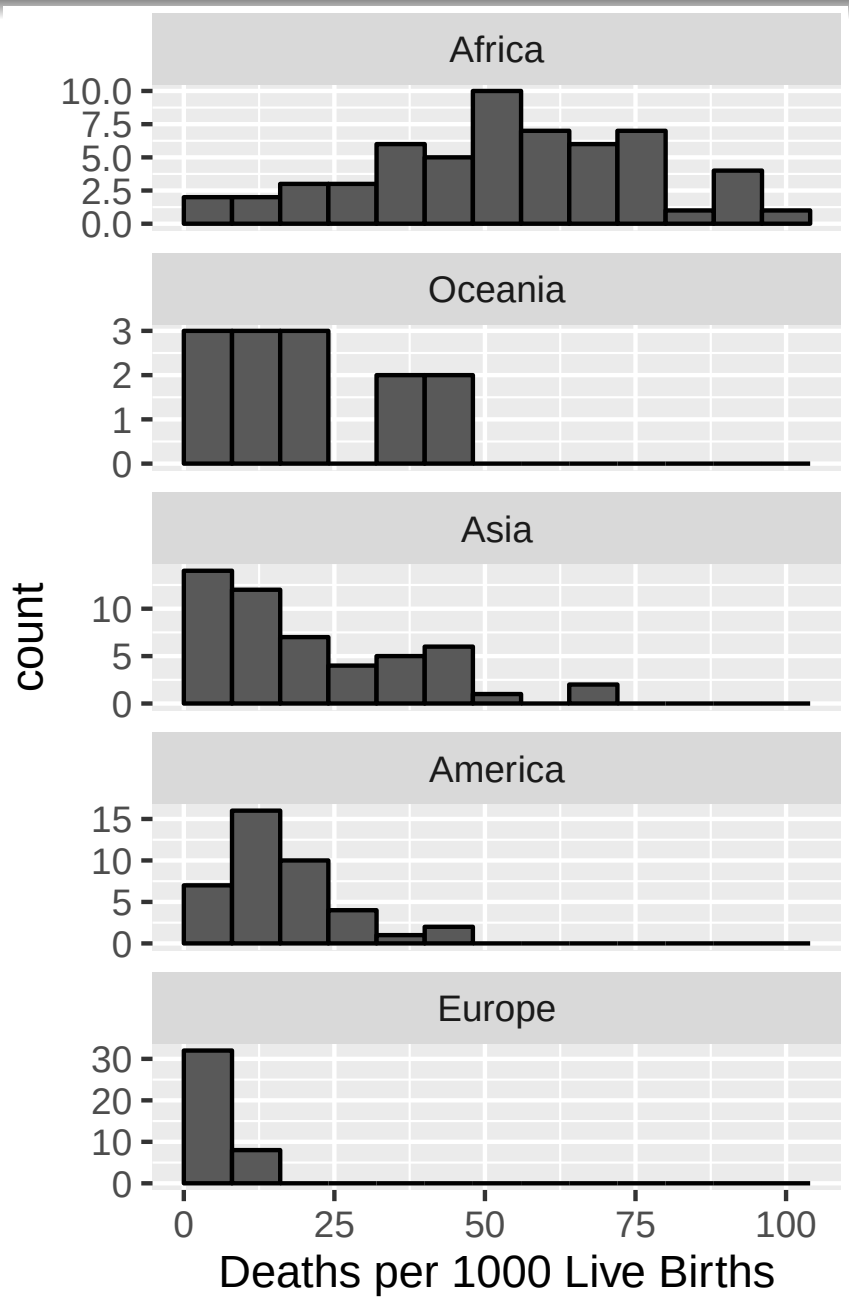


Side-by-Side Histograms — Infant Mortality Rate Data

If countries are grouped by continent and histograms are made separately on the same horizontal axis, we can compare the infant mortality rates of countries in the 5 continents by the location of the histograms, which were

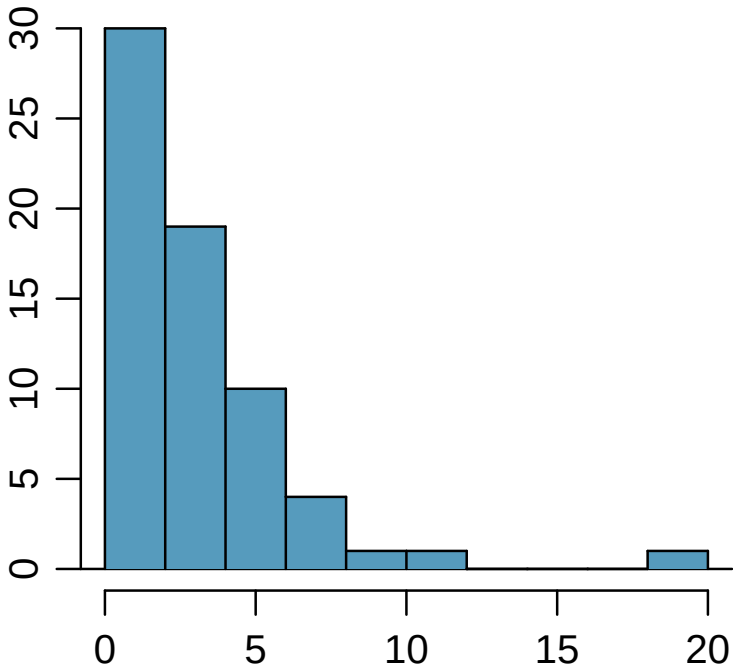
- uniformly low in Europe
- much higher and with greater variability in Africa.

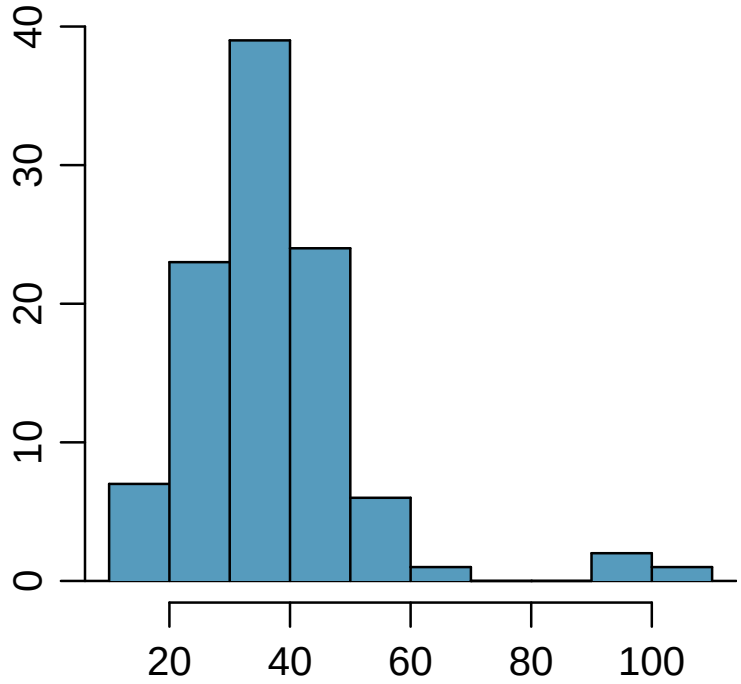
This explains why the histogram for the whole world to be bimodal.



Outliers

Another thing we look at a histogram is whether there are any unusual observations or potential outliers?





More Data Examples

- SAT scores:

```
data(satGPA, package = "openintro")
hist(satGPA$SATV, breaks = 15, xlab = "SATV")
```

- Email Length:

```
data("email", package = "openintro")
hist(email$num_char)
```

- Trump's Tweet Length

```
trump <- read.csv("../../data/trump.csv")
hist(trump$length)
```

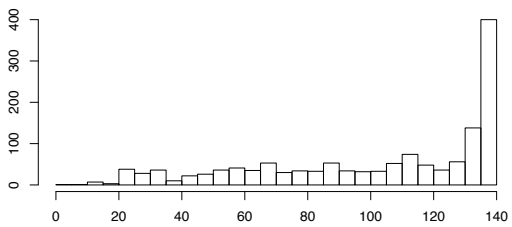
Practice: Shapes of Distributions

Match the following variables with the histograms and bar graphs given below. [Hint: Think about how each variable should behave.]

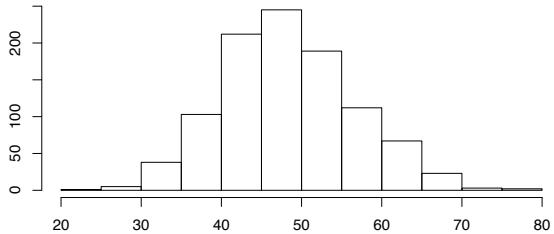
(a) SAT Verbal Scores.

(b) Email Length.

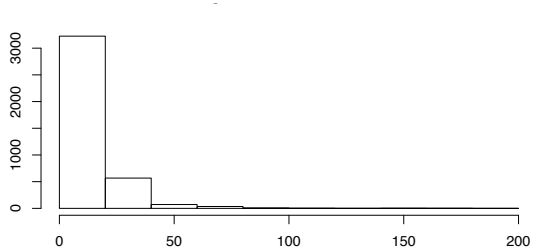
(c) Trump's tweet length.



(c)



(a)



(b)

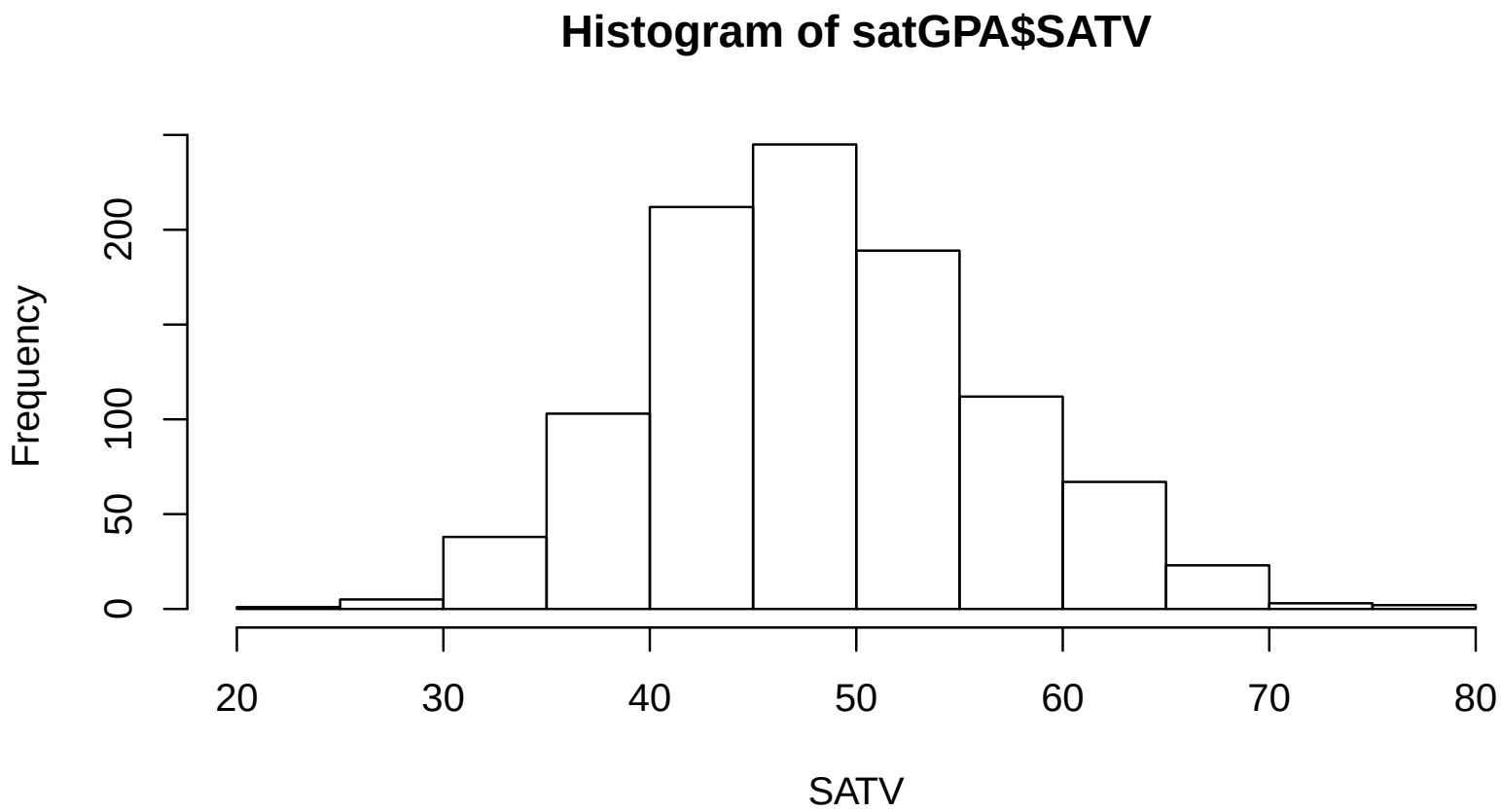
Numerical Summary: Mean and Median

Numerical Summaries

- Sometimes it is inconvenient to provide a graphical summary of your data.
- An alternative is to provide *numerical* summaries of data.
- Summarizing the data numerically can also provide insights into distributions.

Where is the distribution's "center"?

```
library(tidyverse)
data(satGPA, package = "openintro")
hist(satGPA$SATV, breaks = 15, xlab = "SATV")
```



Mean

One measure of center is the (sample) mean. The (sample) mean of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

Example. Suppose a variable has 5 observed values:

$$4,\ 8,\ 3,\ 5,\ 13.$$

The mean of the variable is given by:

$$\overline{x} = \frac{4 + 8 + 3 + 5 + 13}{5} = \frac{33}{5} = 6.6.$$

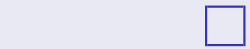
The mean is the balancing point

The average is the 'balancing point' of the data, the 'center of mass'.

prove: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for **any** sample.

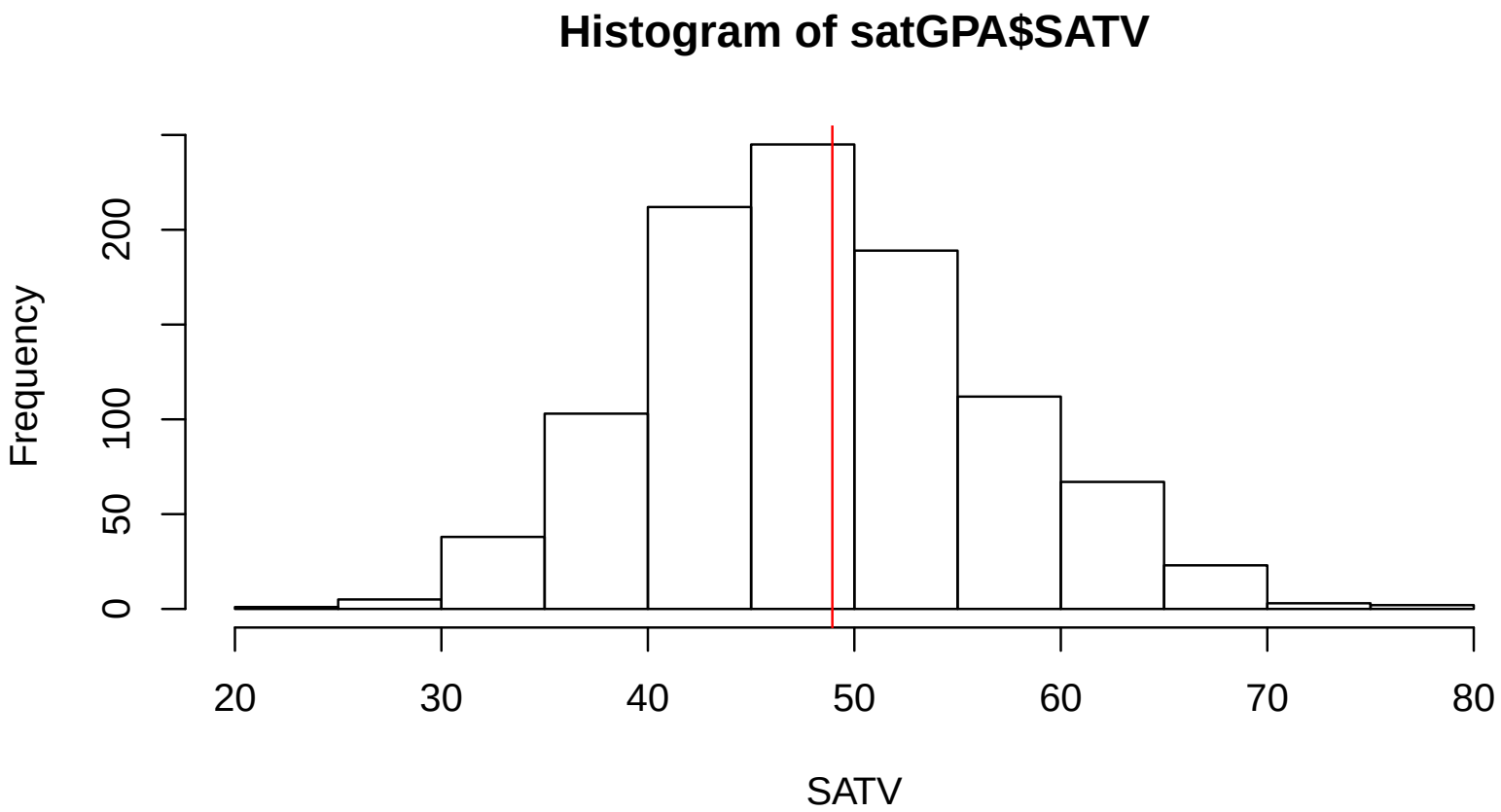
Proof.

$$\begin{aligned}
\sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \text{ (associative property)} \\
&= \sum x_i - n\bar{x} \text{ (summing up } n \text{ identical things)} \\
&= \sum x_i - n\frac{1}{n}\sum x_i \text{ (definition of } \bar{x}) \\
&= \sum x_i - \sum x_i \text{ (}n\text{'s cancel)} \\
&= 0.
\end{aligned}$$



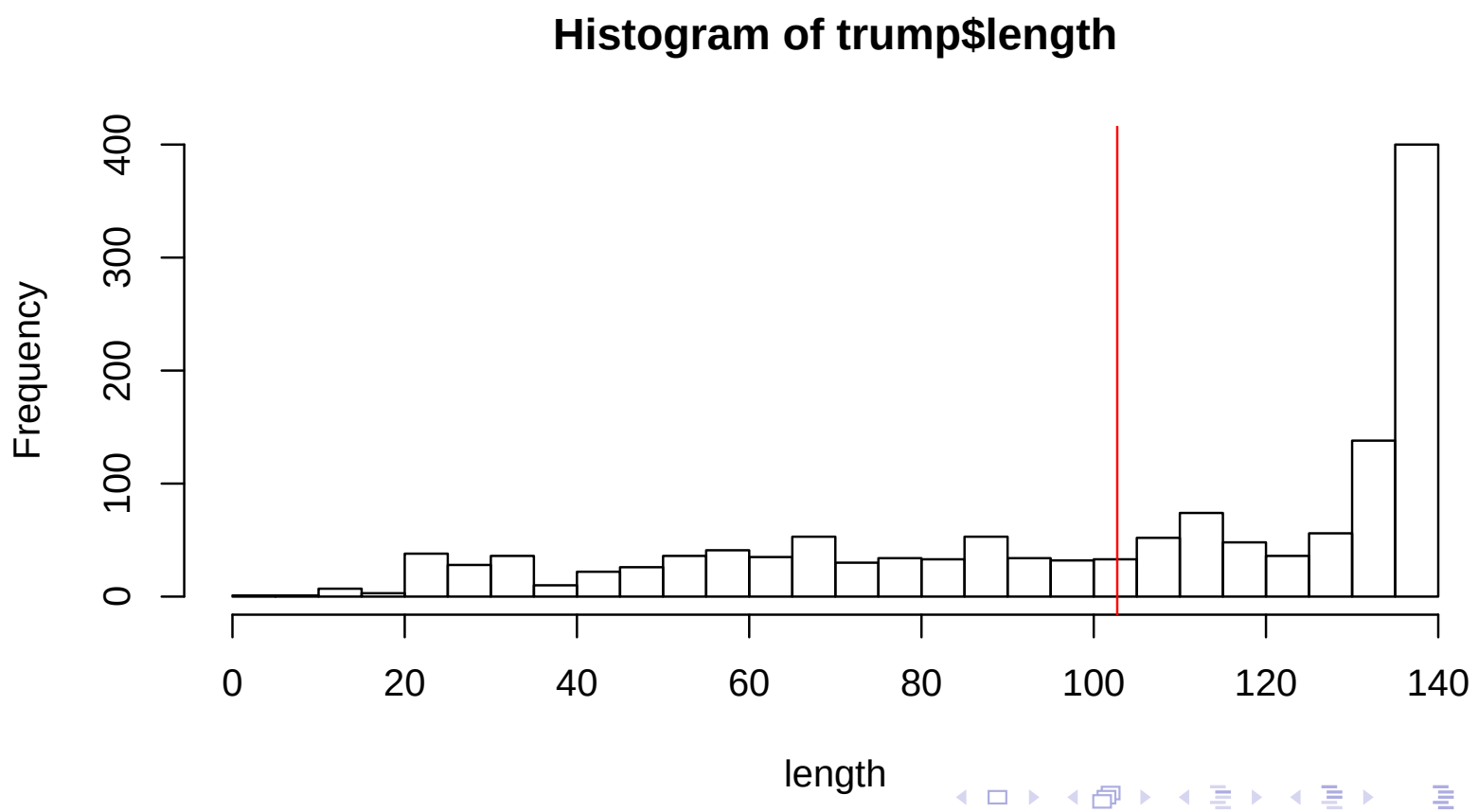
Mean makes sense here

```
xbar <- mean(satGPA$SATV)
hist(satGPA$SATV, breaks = 15, xlab = "SATV")
abline(v = xbar, col = "red")
```



But what about here?

```
trump <- read.csv("../../data/trump.csv")
xbar <- mean(trump$length)
hist(trump$length, breaks = 30, xlab = "length")
abline(v = xbar, col = "red")
```



Why does this happen?

- The skewness is pulling the mean to the left.
- This is because the mean can be interpreted as the "center of mass" of the distribution.
- The mean is not a "typical" value of the length of a tweet.

Example: Mean

The mean is not robust to extreme observations.

```
mean(c(1, 2, 2, 3, 3))

[1] 2.2

mean(c(1, 2, 2, 3, 10))

[1] 3.6

mean(c(1, 2, 2, 3, 20))

[1] 5.6

mean(c(1, 2, 2, 3, 100))

[1] 21.6
```

Another measure of center: The Median

The **median** is the midpoint of a distribution. For a numerical variable, it is a number such that half of the observed value are smaller than it and half are larger than it.

Ex 1: Suppose a variable has 5 observed values: 4, 8, 3, 5, 13.

data	$\rightarrow$	4	8	3	5	13
sorted	$\rightarrow$	3	4	5	8	13
				$\downarrow$		
				Median		

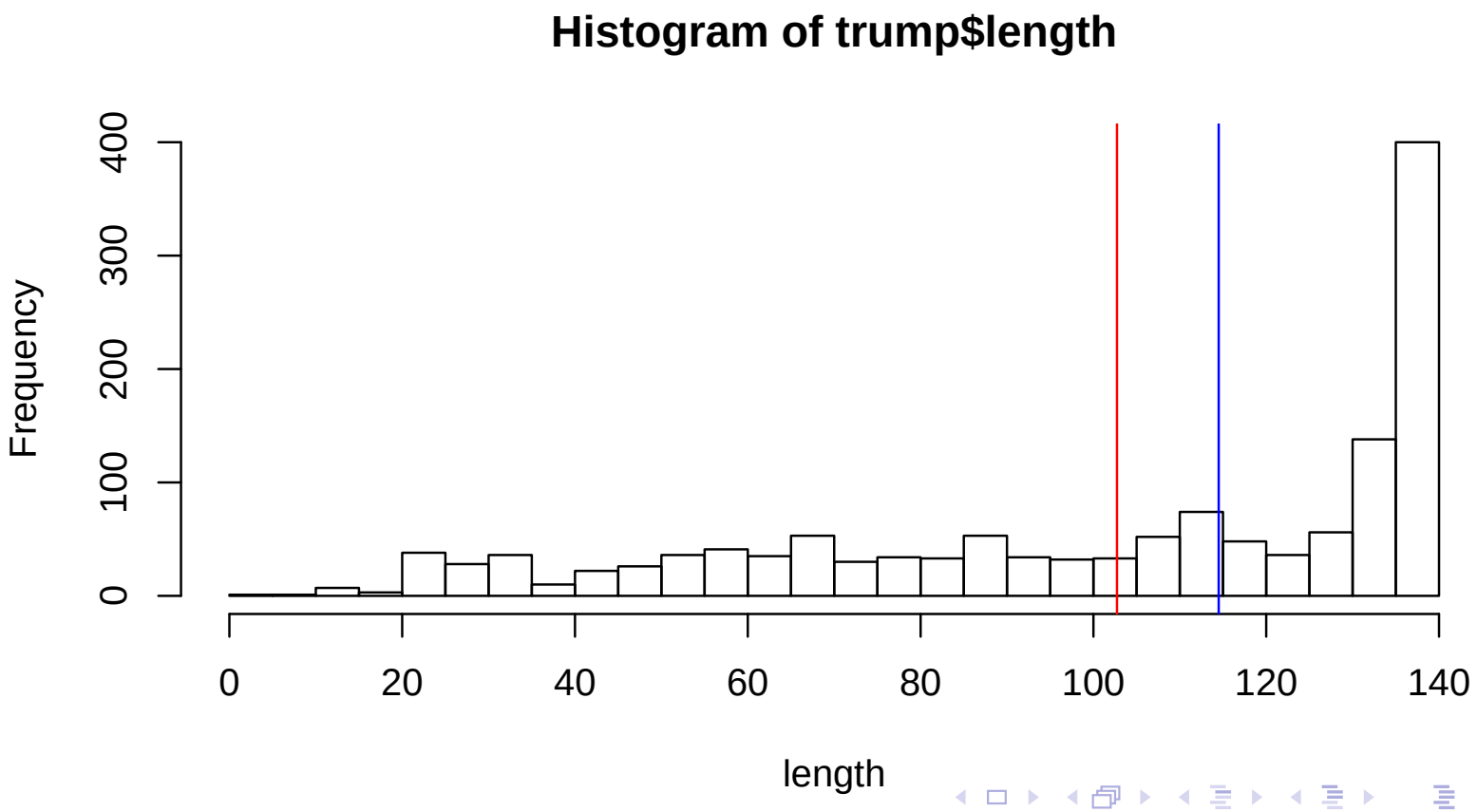
Ex 2: Suppose a variable has 6 observed values: 4, 8, 3, 5, 13, 12.

data	$\rightarrow$	4	8	3	5	13	12
sorted	$\rightarrow$	3	4	5	8	12	13

The median is thus $= \dfrac{5+8}{2} = 6.5$.

Trump's Tweets

```
M <- median(trump$length)
hist(trump$length, breaks = 30, xlab = "length")
abline(v = xbar, col = "red")
abline(v = M, col = "blue")
```



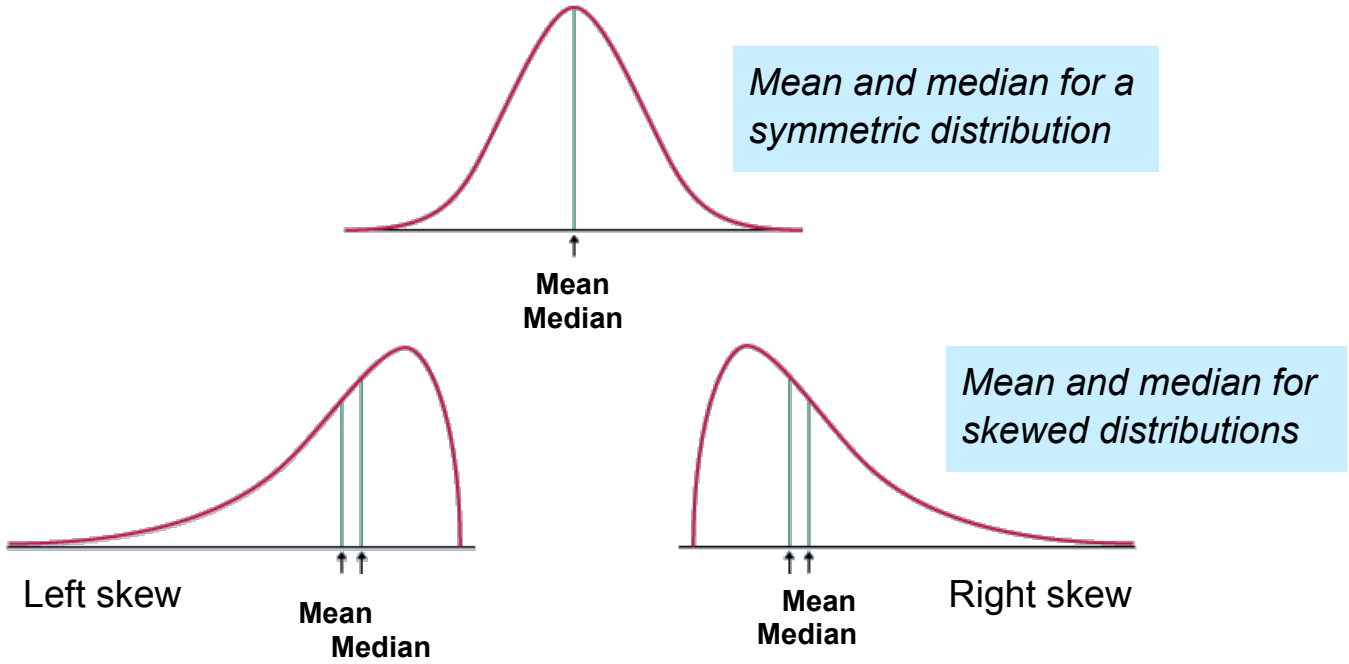
Example: Median

The median is robust to extreme observations.

```
median(c(1, 2, 2, 3, 3))

[1] 2

median(c(1, 2, 2, 3, 10))

[1] 2

median(c(1, 2, 2, 3, 20))

[1] 2

median(c(1, 2, 2, 3, 100))

[1] 2
```

Mean vs. Median

- In a symmetric distribution, mean $\approx$ median.
- In a skewed distribution, the mean is pulled toward the longer tail.



- Median is more resistent, i.e., less sensitive to extreme values or outliers than the mean. We say the median is more **robust**.

Standard Deviation: Measuring Variability

- How do we describe variability of the points from the center? Can we use the average of the deviations from the mean $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$?

- A popular way to describe how the observations spread out is the (sample) standard deviation (SD).

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

Example: Standard Deviation

The standard deviation of $\{1, 2, 2, 7\}$ is

$$\sqrt{\frac{(-2)^2 + (-1)^2 + (-1)^2 + 4^2}{4 - 1}} \approx 2.71$$

Meaning of the Standard Deviation

Recall the formula for the (sample) standard deviation (SD) is

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

- The standard deviation (SD) describes how far away numbers in a list are from their average.
- The SD is often used as a "plus-or-minus" number, as in "Adult women tend to be about 5'4, plus or minus 3 inches".
- Why divided by $n-1$? Not n?

Sample Variance

The square of the sample standard deviation is called the sample variance, denoted as

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

which is roughly the average squared deviation from the mean.

The 68% and 95% Rule (Section 4.1.5 in Text)

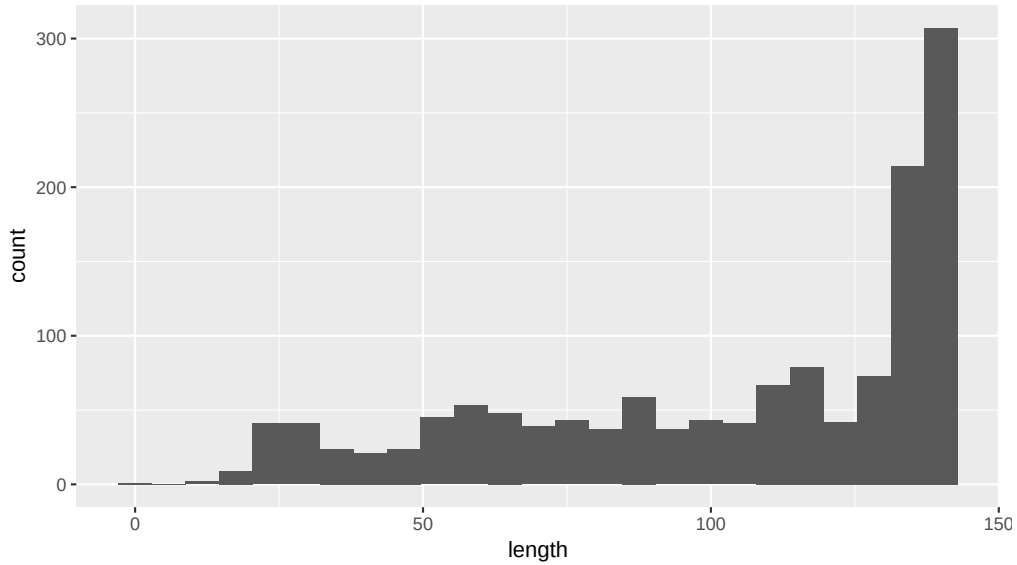
- Roughly 68% of the observations will be within 1 SD away from the mean
- Roughly 95% will be with 2 SD away from the mean

The 68% and 95% rules work very well for bell-shaped data, and reasonably well for unimodal and not seriously skewed data, but not for all data.

Numerical Summary: Five Number Summary and Boxplot

Are these sufficient summaries?

- Mean = 102.73, median = 114.5, Standard deviation = 37.47



- Tells us nothing about the left skew.
- Doesn't tell us that 25% of all tweets are greater than 138 characters.
- Doesn't tell us that small tweets are quite rare.

Percentiles

percentile

The pth percentile of a distribution is the value that has p percent of the observations fall at or below it. To calculate the percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list.

- If we know a few percentiles, that gives us an idea of the shape of a distribution.
- Knowing the **same** percentiles of two distributions makes it easy to quickly compare them.
- It's usual to return the 0th (= minimum), 25th, 50th (= median), 75th, and 100th (= maximum) percentiles.
- The 25th and 75th percentiles are called the first quartile $\mathbf{Q_1}$ and the third quartile $\mathbf{Q_3}$, respectively.

Quartiles, IQR, Five-Number Summary

- **Quartiles** divide data into 4 even parts
  - **first quartile Q_1** = 25th percentile:
    25% of data fall below it and 75% above it
  - **second quartile Q_2** = median = 50th percentile
  - **third quartile Q_3** = 75th percentile
    75% of data fall below it and 25% above it
- **Interquartile Range** $(IQR) = Q_3 - Q_1$
- **Five-Number Summary**:

$$\text{min}, \quad Q_1, \quad \text{Median}, \quad Q_3, \quad \text{max}$$

Example

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

$$\begin{array}{cccl} 43 & & 27 & \\ 35 & & 33 & \\ 43 & & 34 & \leftarrow \text{median of this half} = \frac{33+34}{2} = 33.5 = Q_1 \\ 33 & & 35 & \\ 38 & \xrightarrow{\text{sort}} & 38 & \leftarrow \text{overall median} = Q_2 \\ 53 & & 43 & \\ 64 & & 43 & \\ 27 & & 53 & \leftarrow \text{median of this half} = \frac{43+53}{2} = 48 = Q_3 \\ 34 & & 64 & \end{array}$$

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Calculation of Quartiles

In fact, statisticians don't have a consensus on the calculation of quartiles.
E.g., for the 9 numbers example, our computation gives $Q_1 = 33.5$, $Q_3 = 48$, but R gives $Q_1 = 34$, $Q_3 = 43$.

```
> x = c(43,35,43,33,38,53,64,27,34)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   34.00   38.00   41.11   43.00   64.00
> fivenum(x)
[1] 27 34 38 43 64
> IQR(x)
[1] 9
```

Don't worry about the formula. Just keep in mind that **quartiles divide data into 4 even parts**. In HWs, just report whatever values your software gives.

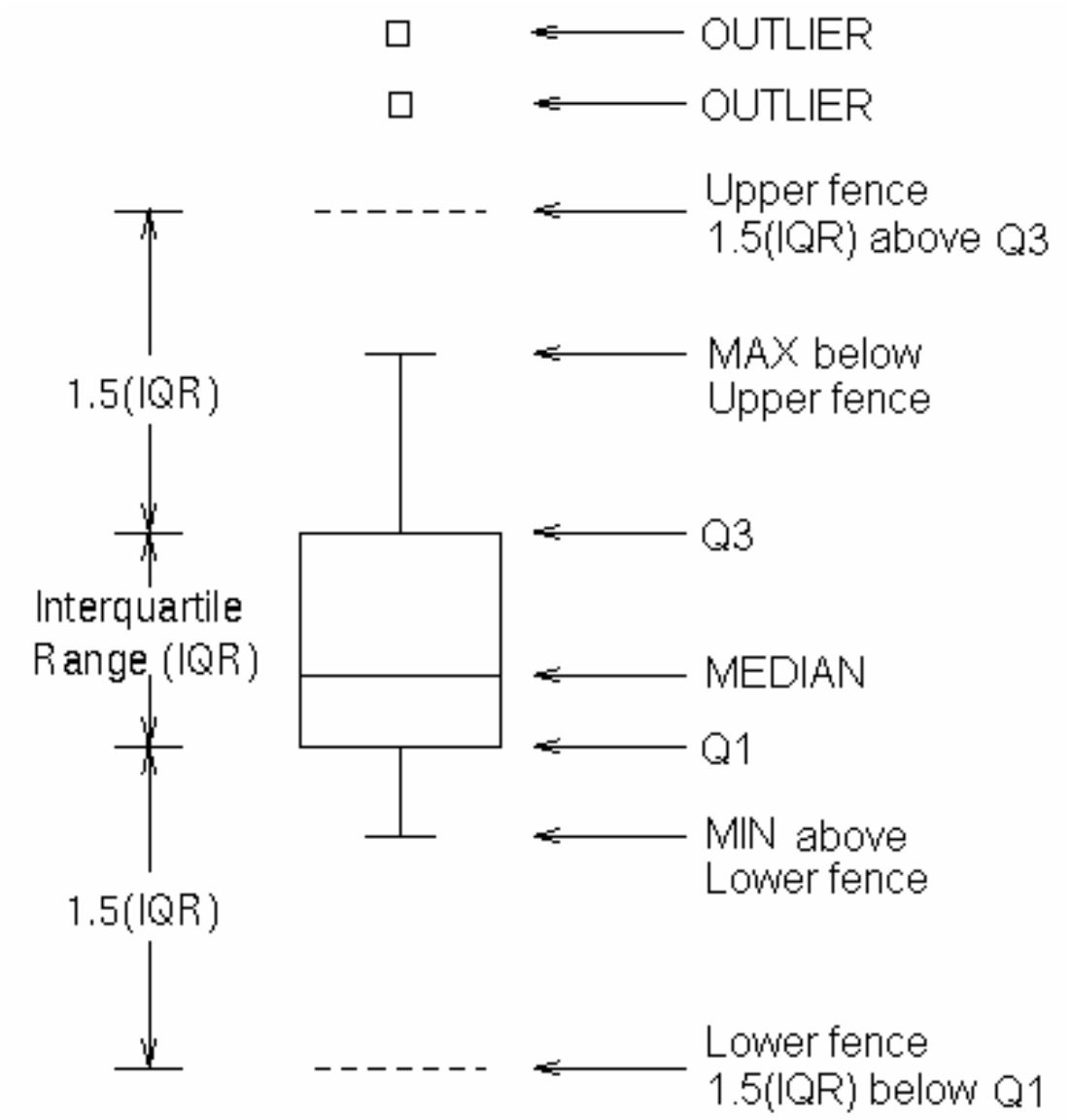
Boxplot

- It's *very* useful to plot these quantiles in what is called a boxplot.

boxplot

A boxplot is a graph of the five number summary. A central box spans the quartiles Q_1 and Q_3. A line in the box marks the median M. Lines (the "whiskers") extend from the box out to the smallest and largest reachable observations.

Box-and-Whiskers Plot (also called Boxplot)



$1.5 \times IQR$ Rule

People will often call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

- In most boxplots, the upper whisker extends to the largest observation within $1.5 \times IQR$ of Q_3.
- In most boxplots, the lower whisker extends to the smallest observation within $1.5 \times IQR$ of Q_1.
- Points outside of $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ are labelled "suspsected outliers" and are plotted individually.

Think About It ...

- What does a boxplot look like if the distribution is symmetric?

  See examples to follow.

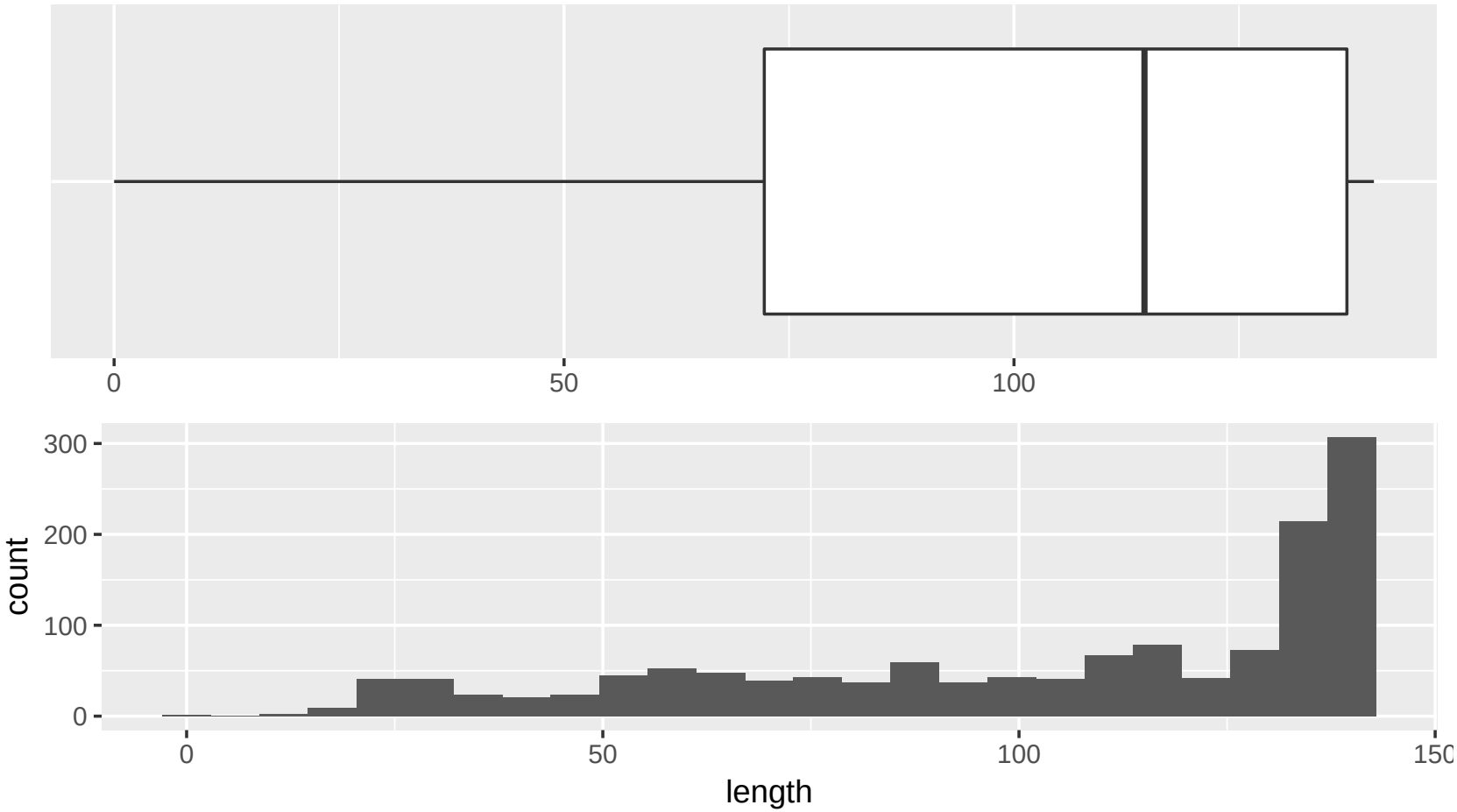
- if right-skewed?

  See examples to follow.

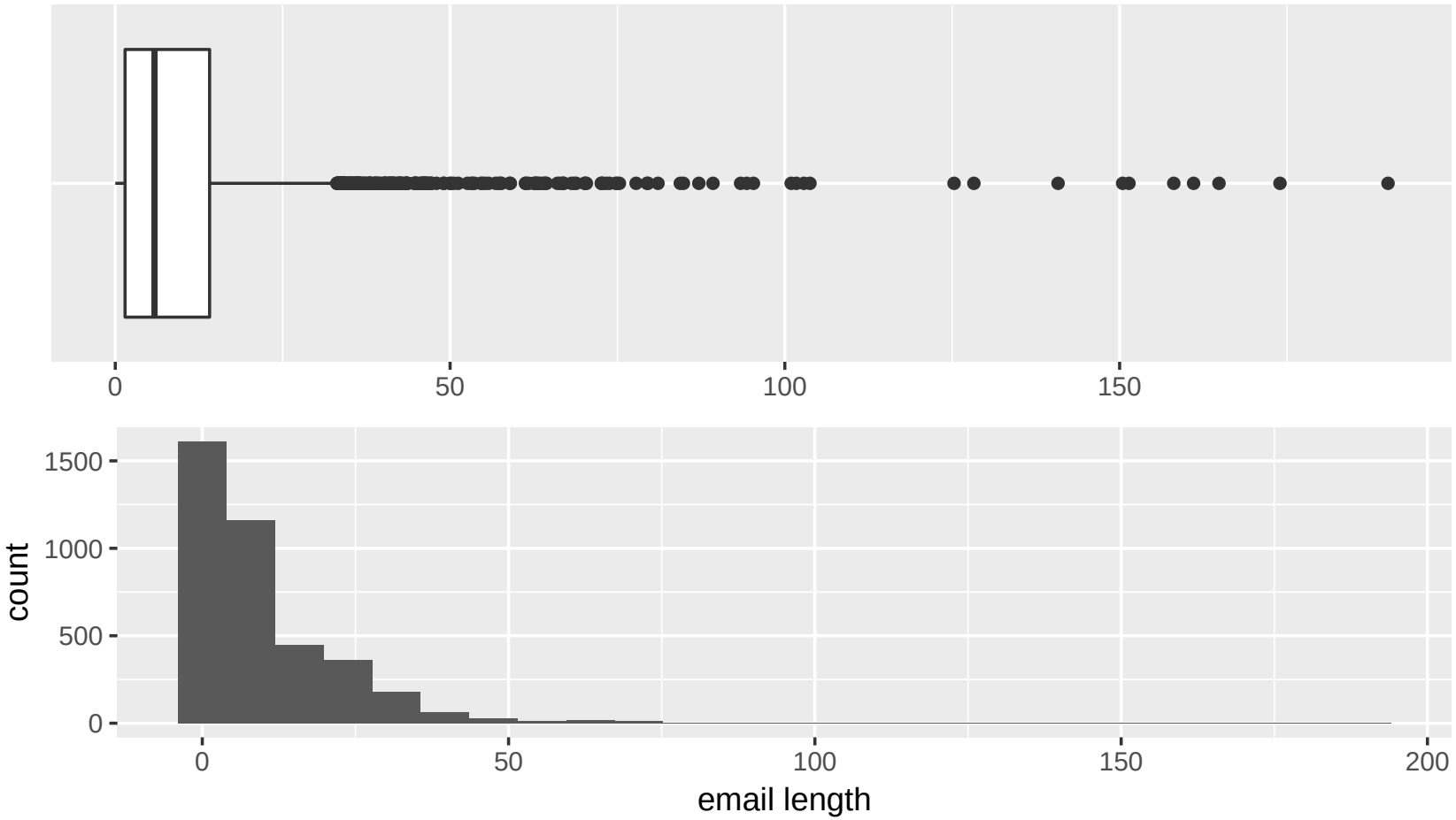
- Can you tell from a boxplot whether the distribution is unimodal or bimodal?

  Unfortunately not, as boxplots only plot the numerical summary of the data.

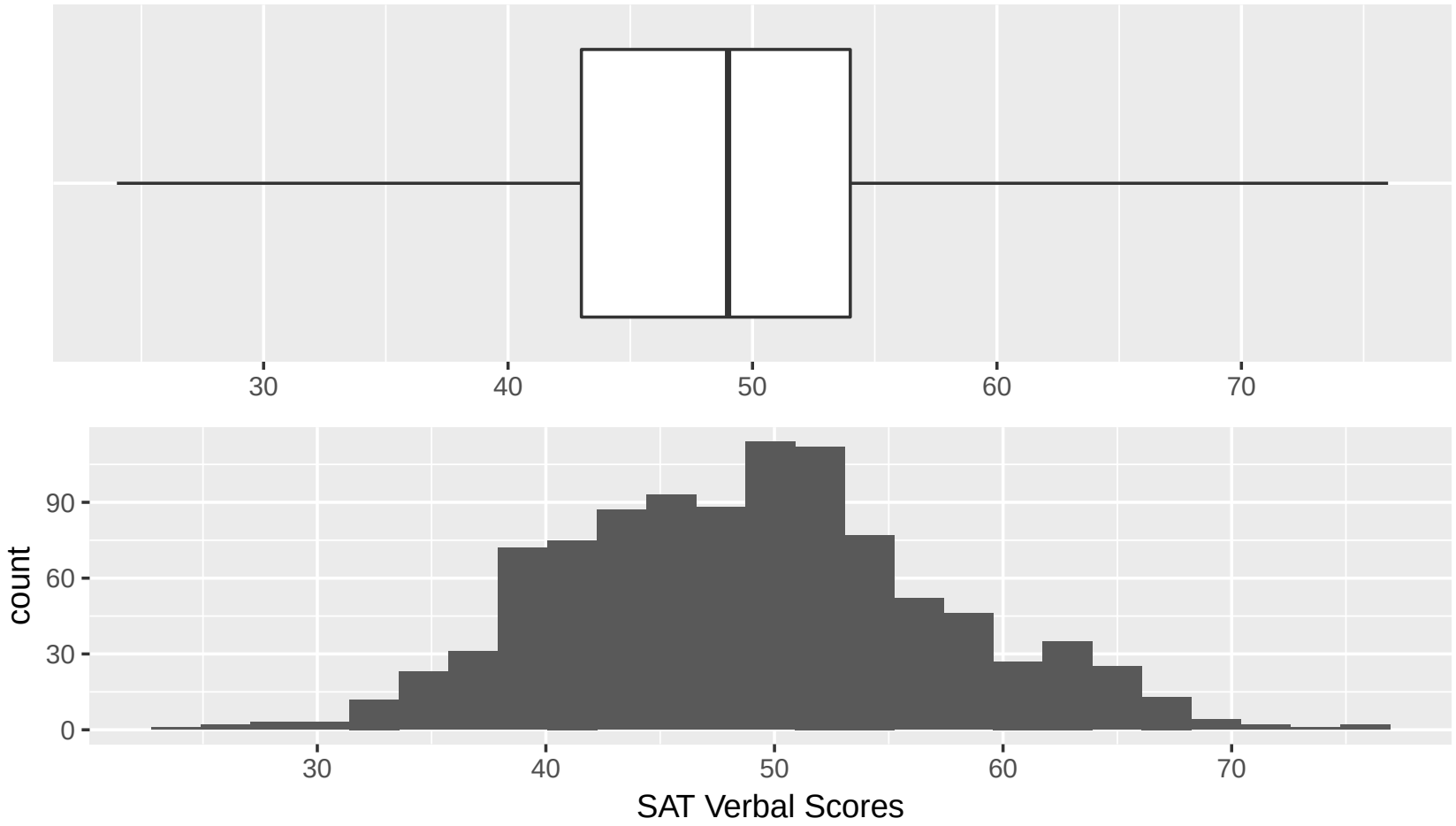
Boxplots tell us about skewness: trump



Boxplots tell us about skewness: email

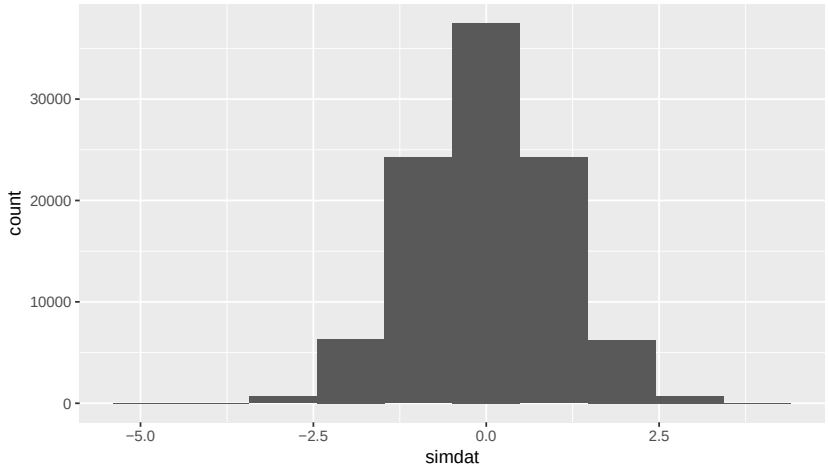


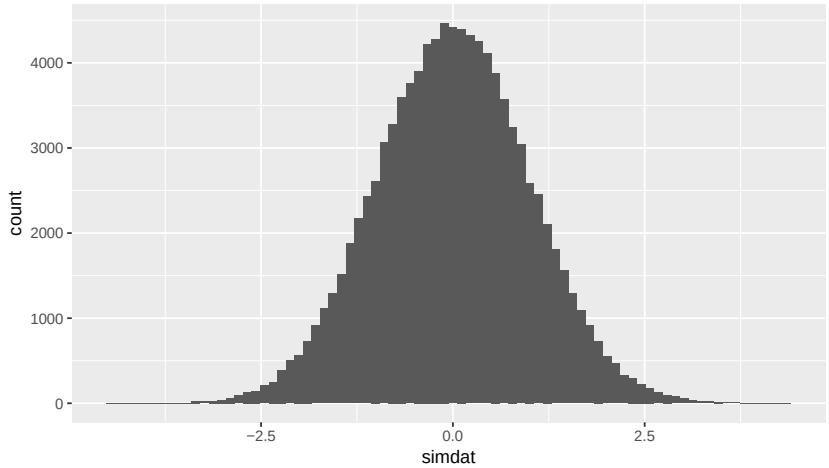
Boxplots tell us about skewness: satGPA

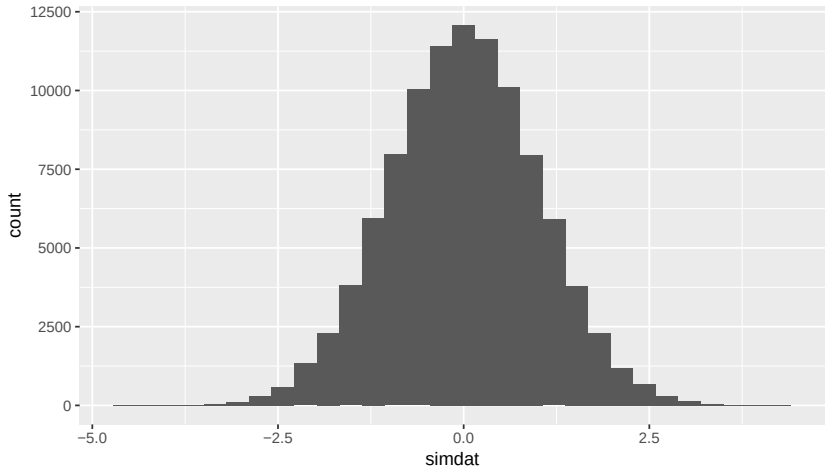


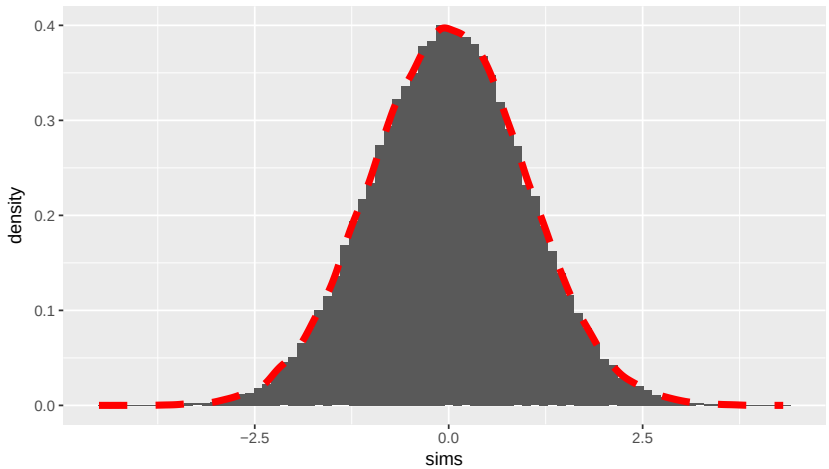
Density Curves

From histogram to density









Density curve

- The distributions of many quantitative variables can be approximated by a density curve

density curve

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range. A density cuve is a curve that

- Is always on or above the horizontal axis.
- Has area exactly 1 underneath it.

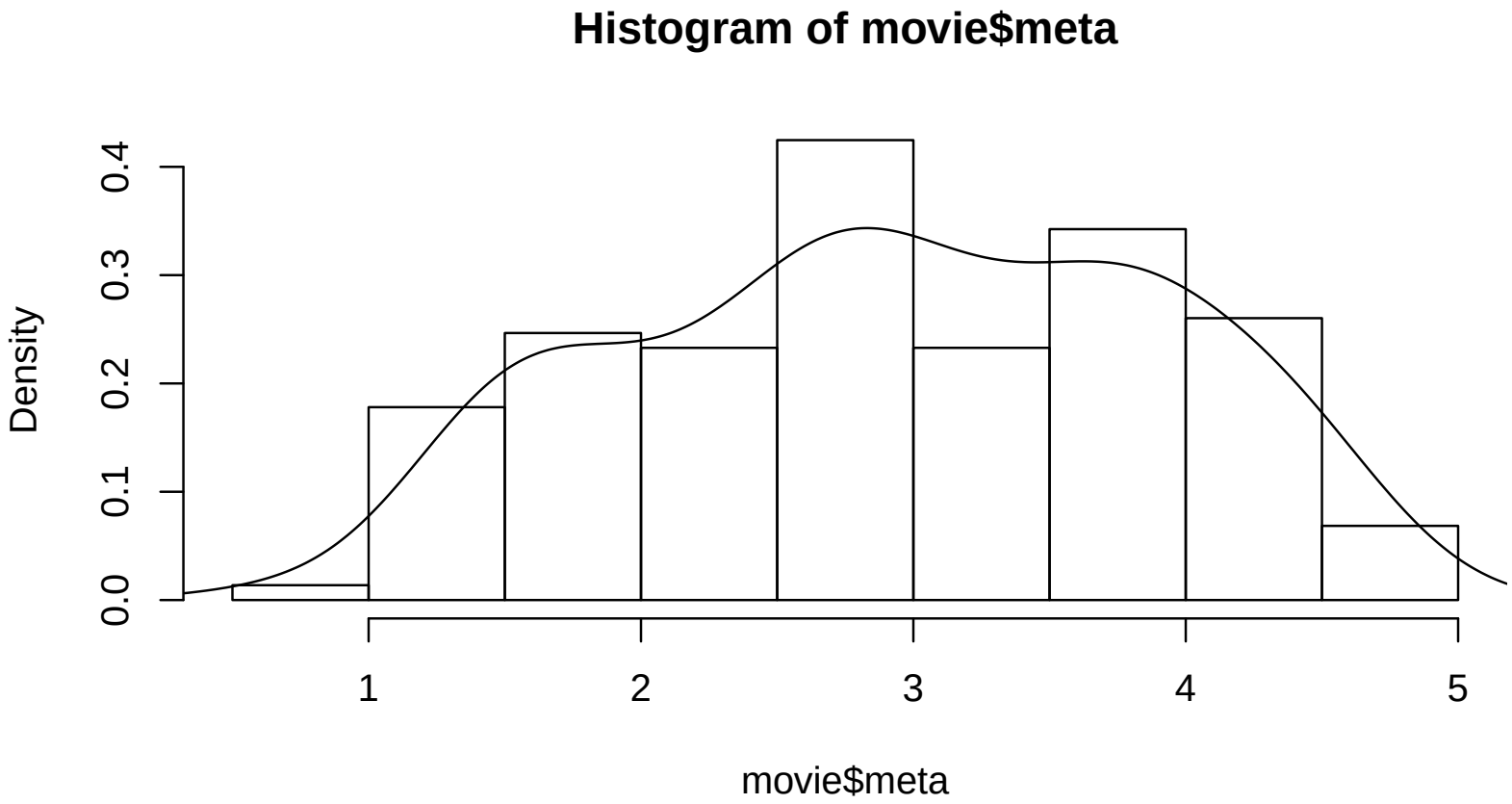
Data: Movie Scores

Observational units: Movies that sold tickets in 2015.
Variables:

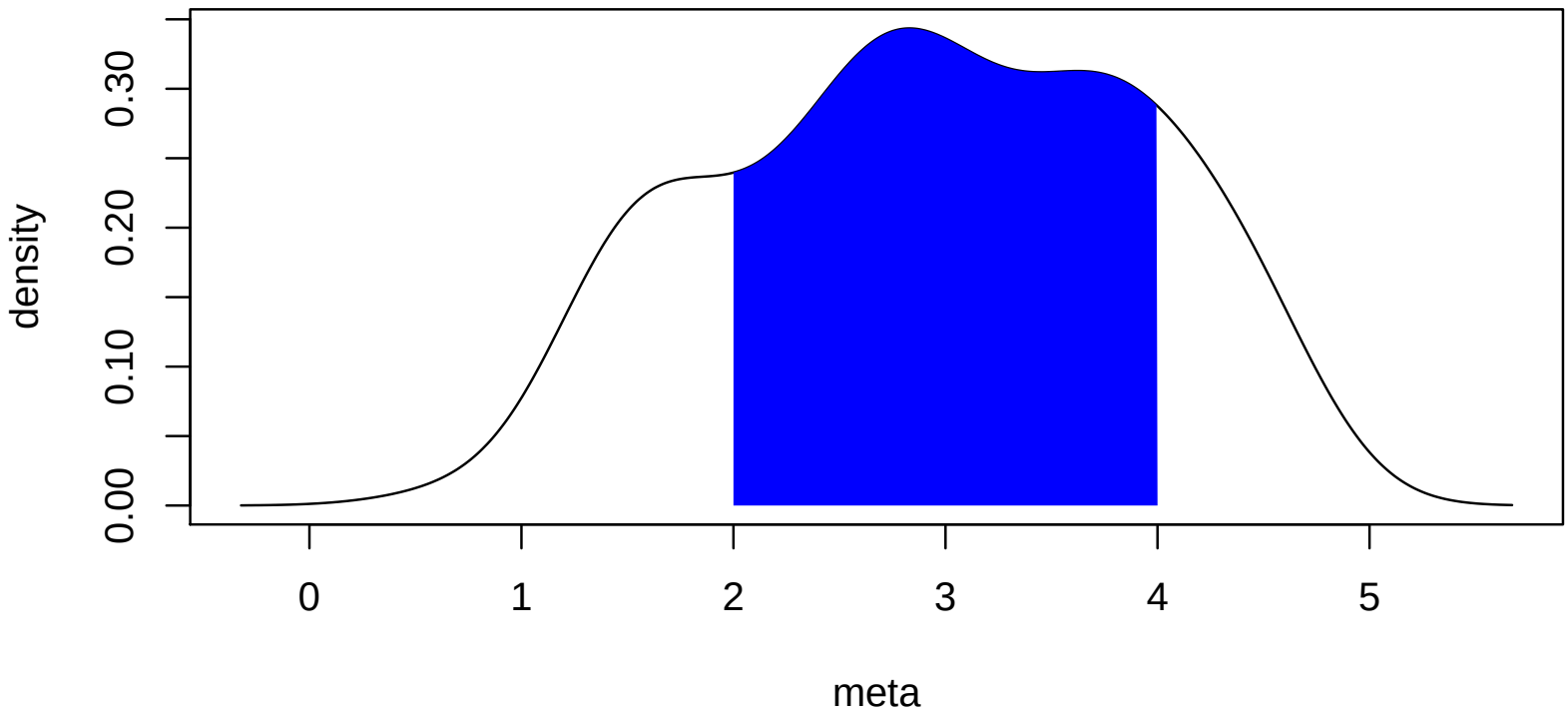
- `rt` Rotten tomatoes score normalized to a 5 point scale.
- `meta` Metacritic score normalized to a 5 point scale.
- `imdb` IMDB score normalized to a 5 point scale.
- `fan` Fandango score.

Density of Metacritic scores

```
md <- density(movie$meta)
hist(movie$meta, freq = FALSE)
lines(md$x, md$y)
```

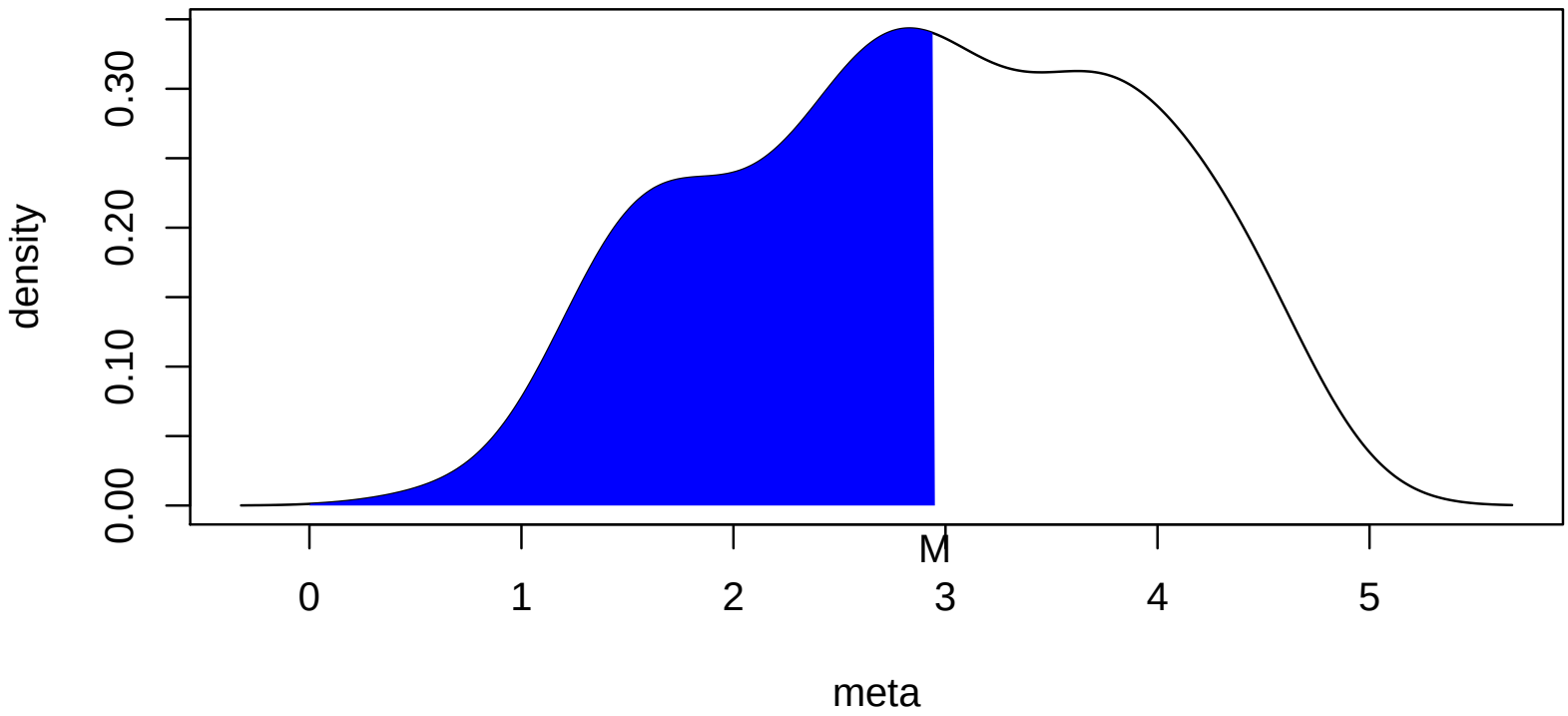


Density example



E.g. Area of shaded region is approximately the proportion of metacritic scores that falls between 2 and 4.

Median



Median M is where half of the area is to the left and to the right of M.

Checking for normality and Q-Q plot

It's sometimes important to check if normality is a valid approximation.

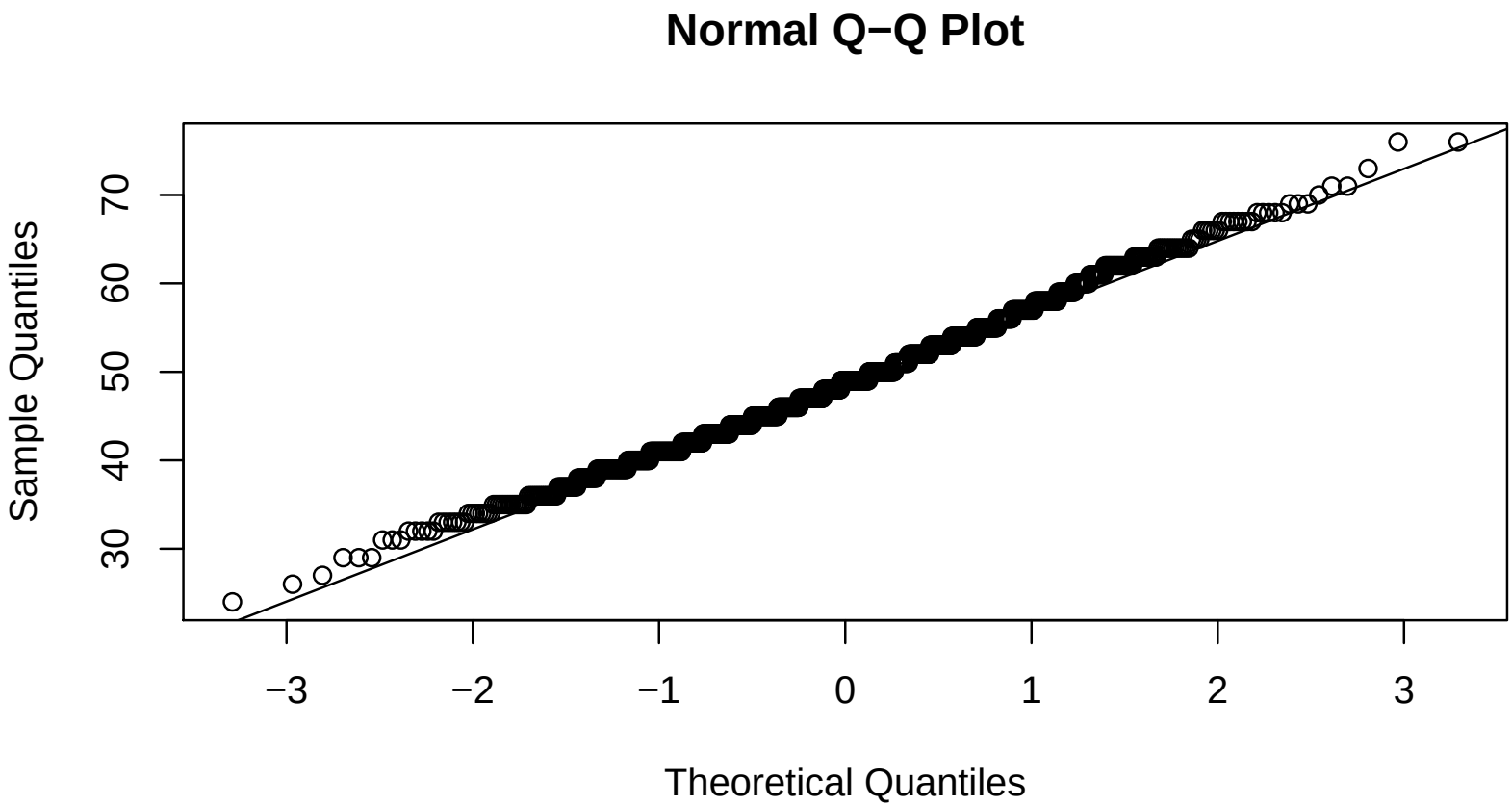
- Idea: Is the 68-95-99.7 rule approximately correct for the `satGPA` data?
- More generally, do the percentiles (quantiles) of the data match with the percentiles (quantiles) of the theoretical normal distribution?
- Compare the pth percentile (quantile) of the data and the pth percentile (quantile) of a $N(\bar{x}, s^2)$ distribution. If they are pretty close, then normality is a good approximation.

Quantile-quantile plot

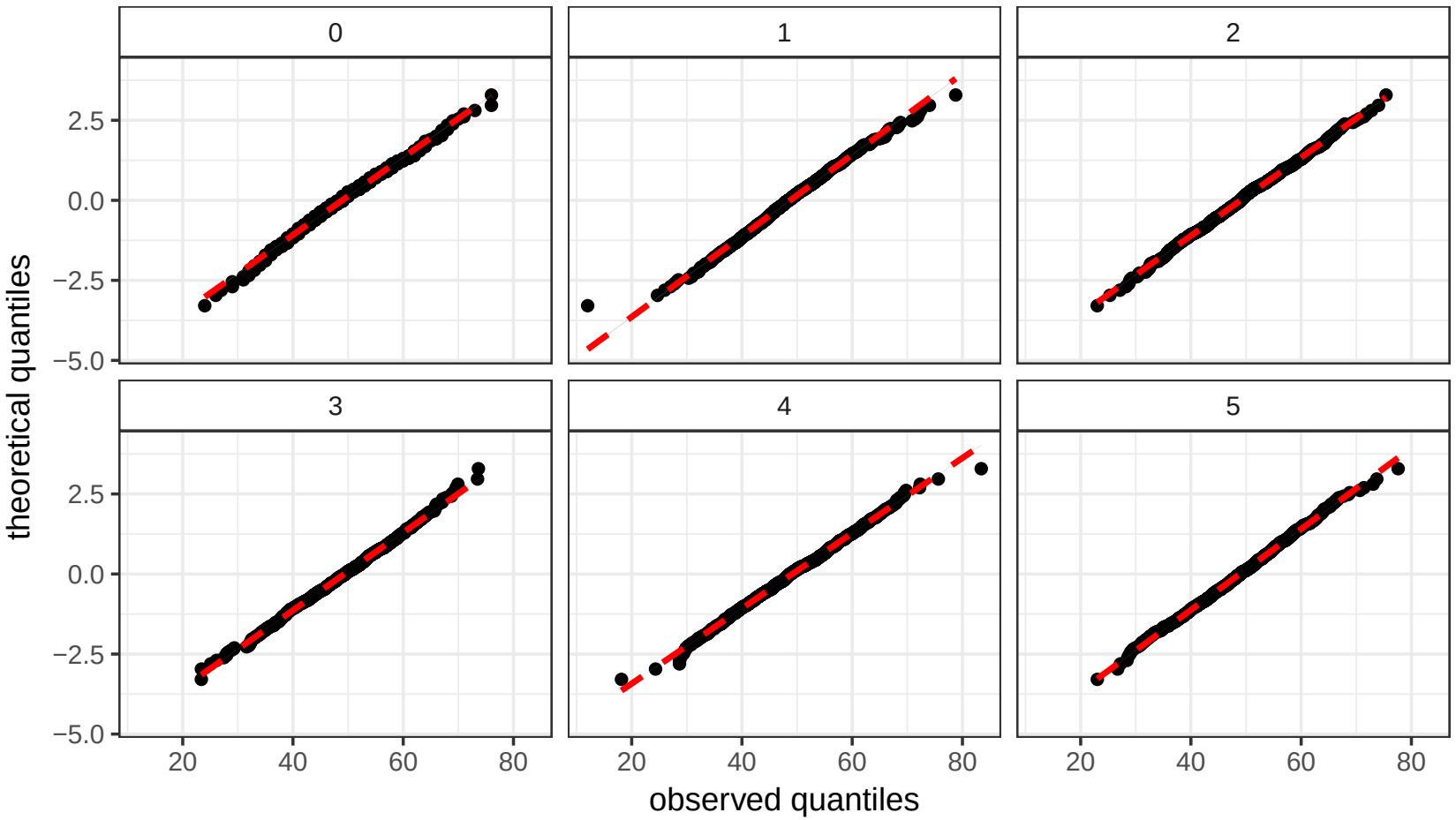
- Plot the observed quantiles against the quantiles of a $N(\bar{x}, s^2)$ density.
- If the points lie close to a line, then the normal approximation is approximately correct.
- Can just plot the observed quantiles against $N(0, 1)$ and look for a straight line (more on why later).

QQplot

```
qqnorm(satGPA$SATV)
qqline(satGPA$SATV)
```

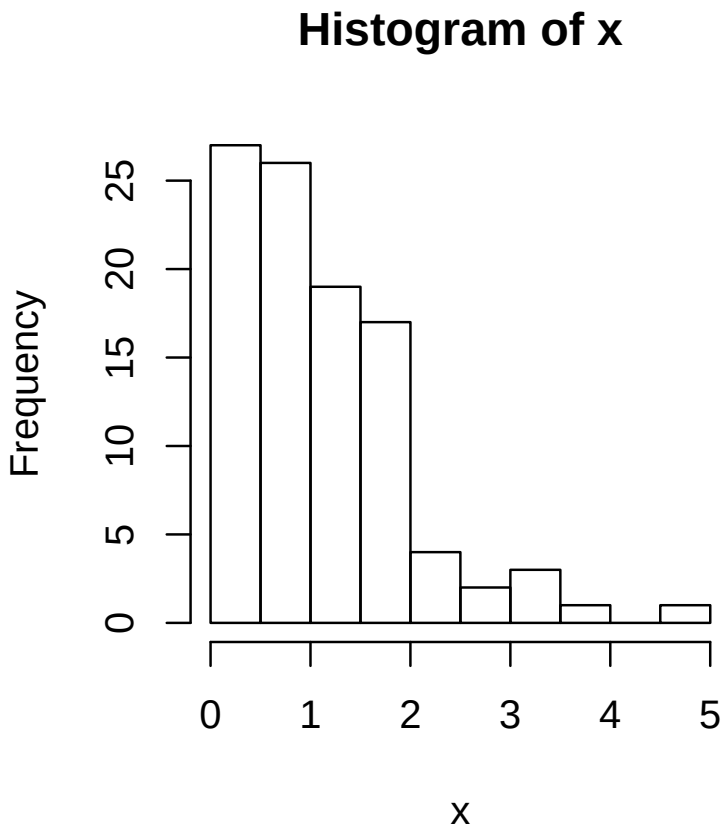


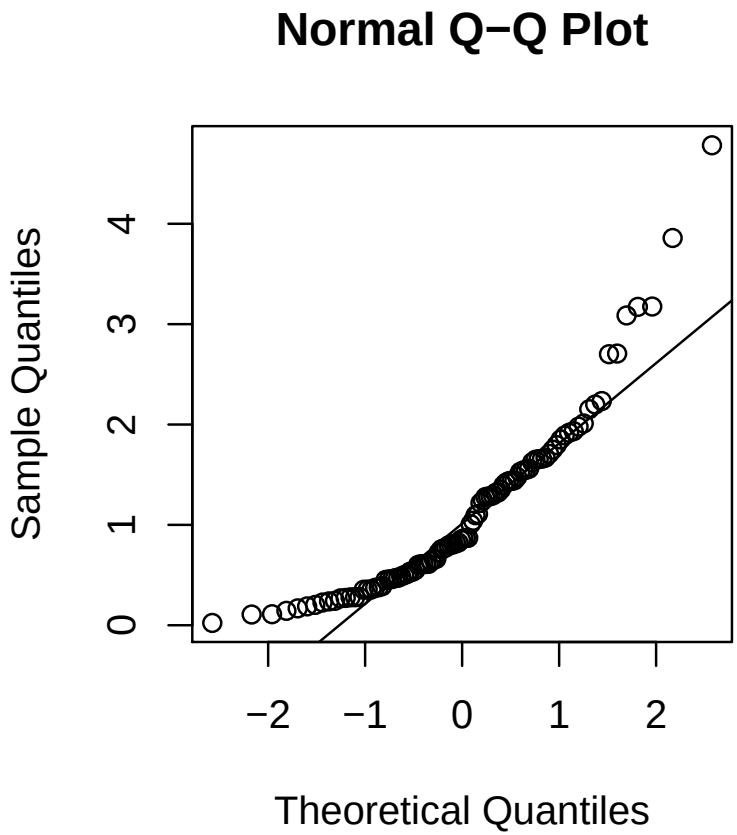
But what does a "good" qqplot look like?



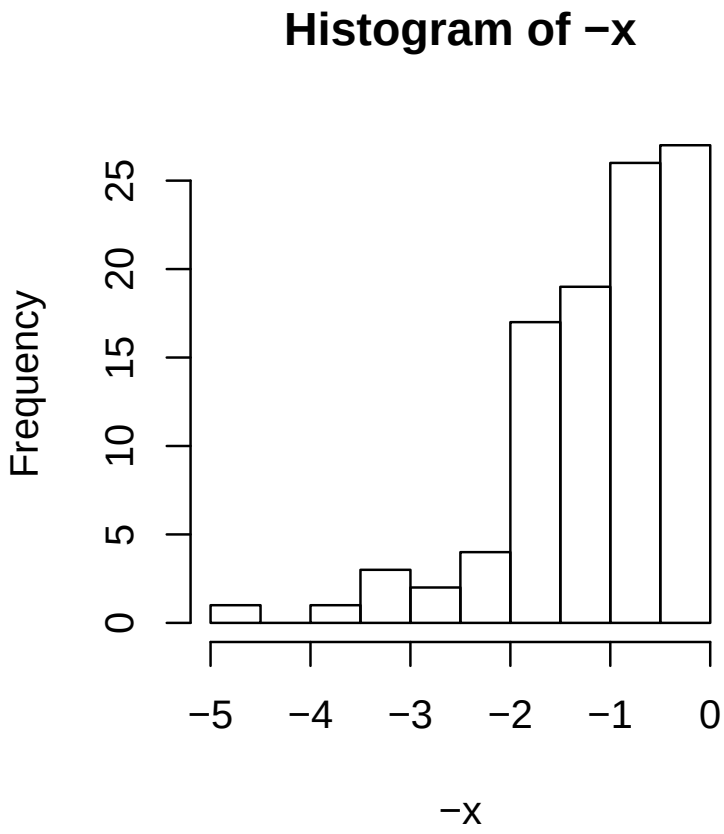
Top left is real data, rest are simulated from $N(\bar{x}, s^2)$ — looks good to me!

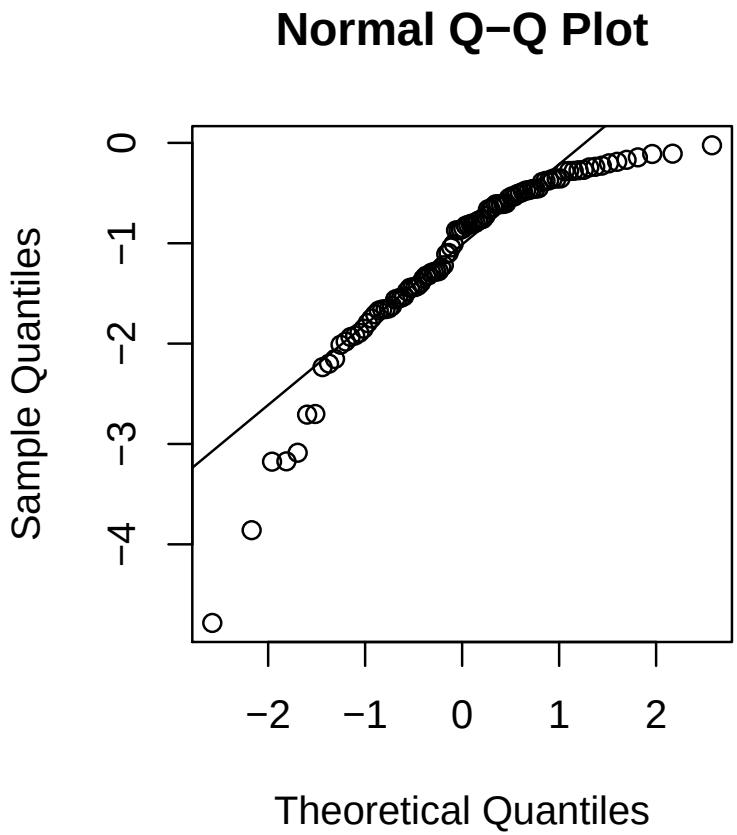
Problem: Skewed right



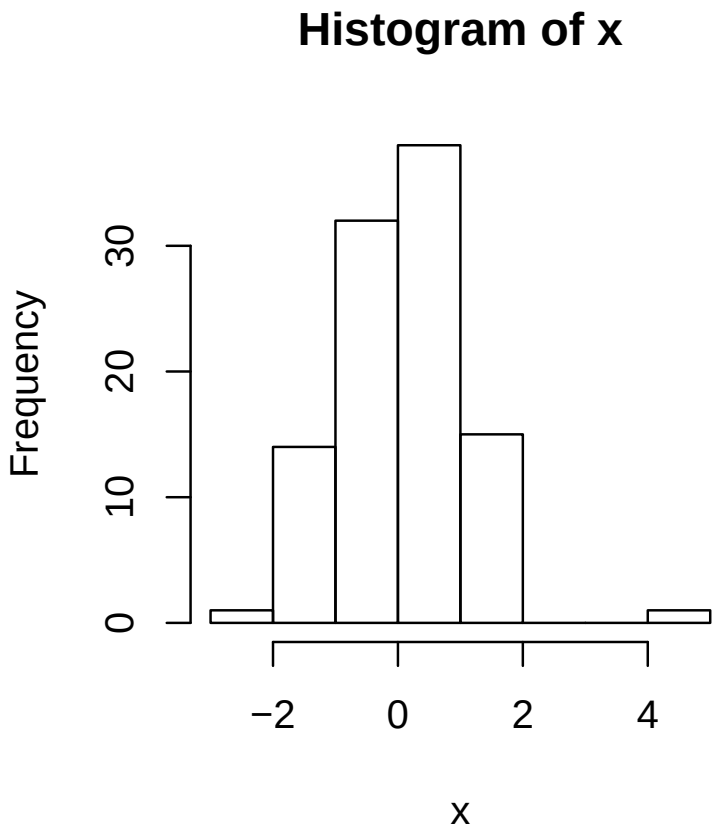


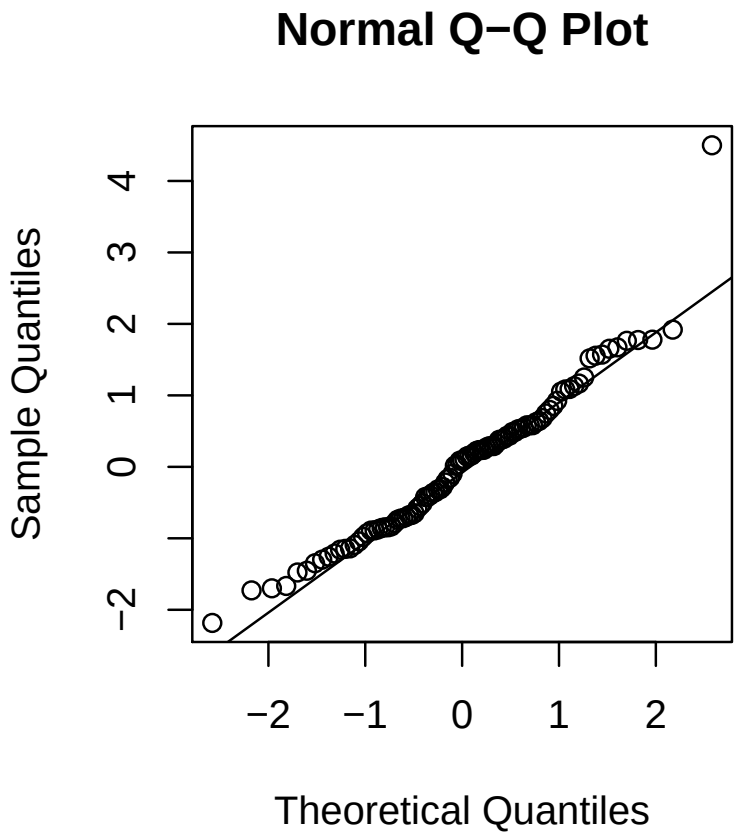
Problem: Skewed left



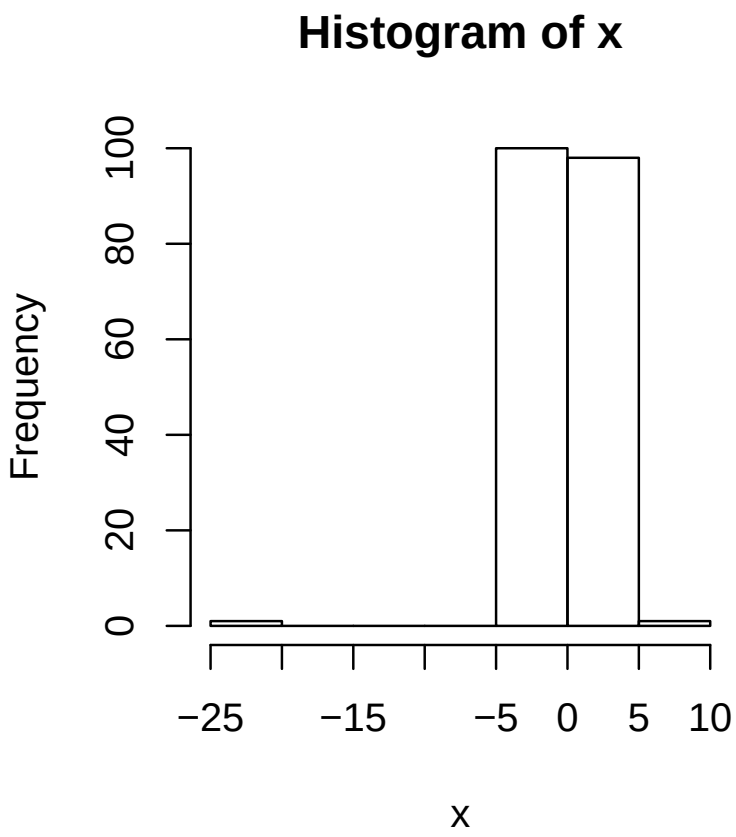


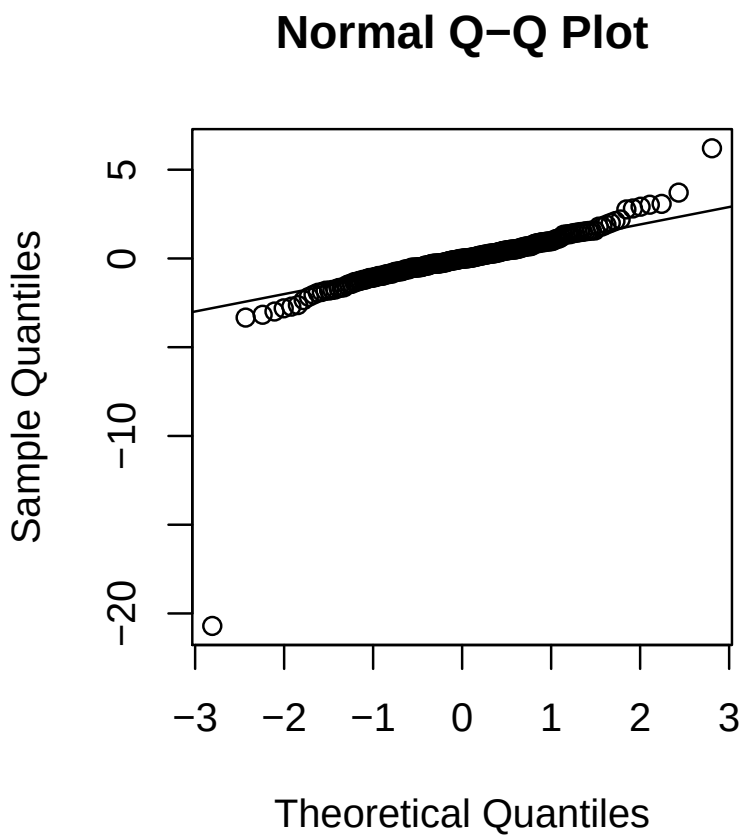
Problem: Outliers



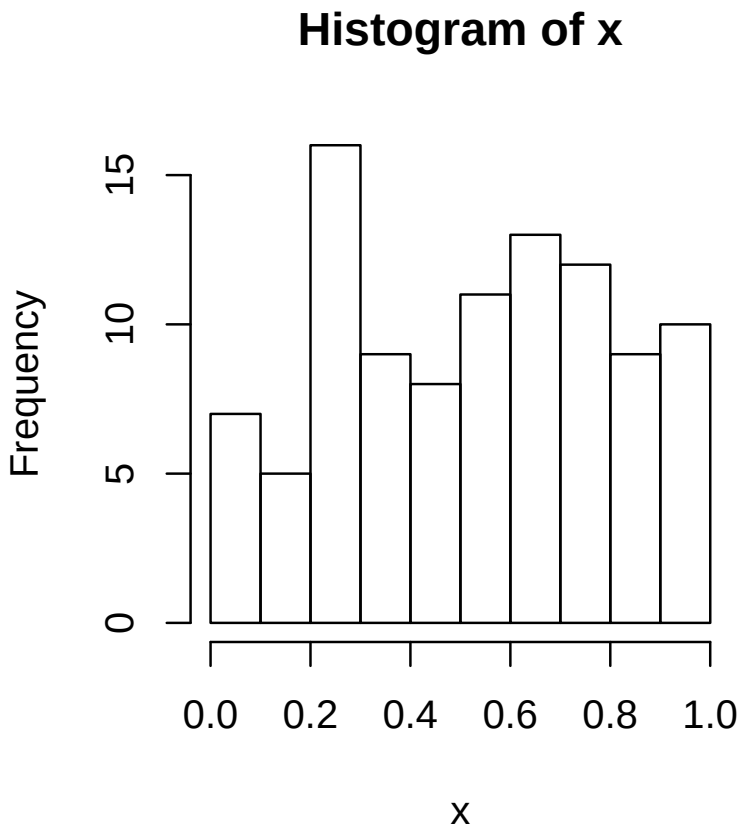


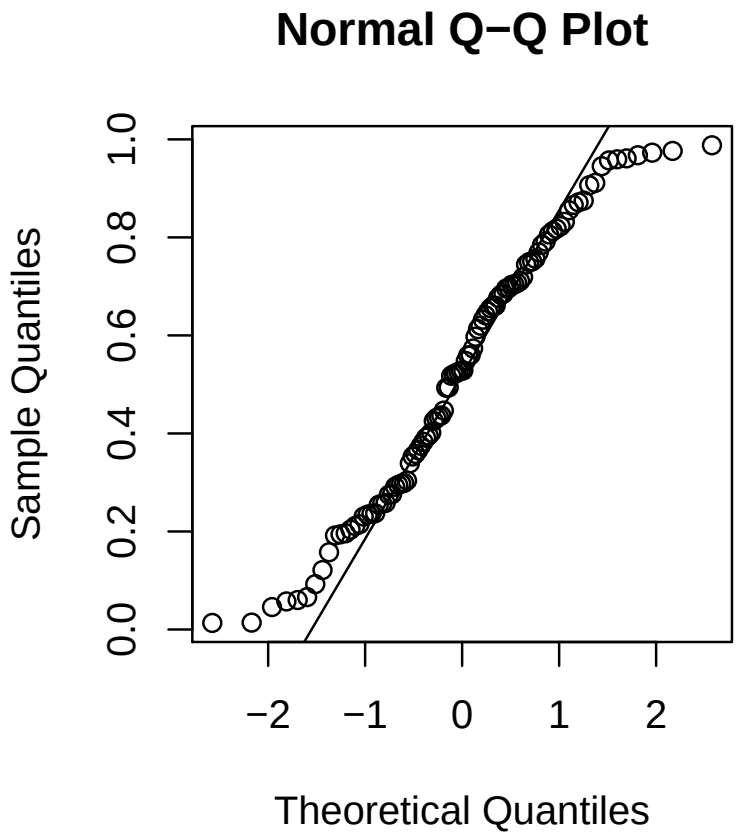
Problem: Heavy tails





Problem: Light tails





Linear Transformations

Linear Transformations

Sometimes, we want to analyze data in different units.

- Celsius $= \frac{5}{9}$(Fahrenheit - 32)
- Curved score = score + (0.25)(100-score)(This curve adds back 25% of exam points missed.)
- Standardized score $z_i = \frac{(x_i - \bar{x})}{s}$

All three are examples of linear transformations: $y = a + bx$

Relationships

Let $y_i = a + bx_i$ for $i = 1, 2, \ldots, n$.

- **Mean** $\bar{y} = a + b\bar{x}$
- **Median** $\text{median}(y_1, \ldots, y_n) = a + b\ \text{median}(x_1, \ldots, x_n)$
- **Standard deviation** $\text{SD}(y) = |b|\text{SD}(x)$

Standardizing and z-scores

Let $y_i = a + bx_i$ for $i = 1, 2, \ldots, n$.

standardizing and z-scores

If x is an observation from a distribution with mean μ and standard deviation σ, the standardized value of x is

$$z = \frac{x - \mu}{\sigma}.$$

A standardized value is often called a z-score.

The z-score is in units of standard deviations above the mean.