STAT 234 Lecture 14
Graphical and Numerical Summary of Data
Section 1.2-1.4

Yibi Huang
Department of Statistics
University of Chicago

Outline

- Data Matrix & Types of Variables
- Histograms (1.2)
- Measure of Location: Mean and Median (1.3)
- Measure of Variability (1.4)
    - Standard Deviation
    - Five-Number Summary and Box Plots (1.6.5)

Data Basics

Data Matrix

Example: Bird Strike data

A collection of collisions between aircraft and wildlife reported to the US Federal Aviation Administration in 1990-1997. See https://www.openintro.org/data/index.php?data=birds

	num.engs	height	sky	birds.struc
1	2	7000	No Cloud	1
2	2	10	No Cloud	2-10
3	3	400	Some Cloud	1
4	2	100	Overcast	1
⋮	⋮	⋮	⋮	⋮
19302	4	50	No Cloud	1

Cases

Each **row** of a 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.
- What's a case in the bird strike data?

	num.engs	height	sky	birds.struc	
1	2	7000	No Cloud	1	
2	2	10	No Cloud	2-10	
3	3	400	Some Cloud	1	← *case*
4	2	100	Overcast	1	
⋮	⋮	⋮	⋮	⋮	
19302	4	50	No Cloud	1	

Cases

Each **row** of a 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.
- What's a case in the bird strike data?
  - A case is a reported aircraft-wildlife collision

	num.engs	height	sky	birds.struc	
1	2	7000	No Cloud	1	
2	2	10	No Cloud	2-10	
3	3	400	Some Cloud	1	← *case*
4	2	100	Overcast	1	
⋮	⋮	⋮	⋮	⋮	
19302	4	50	No Cloud	1	

Variables

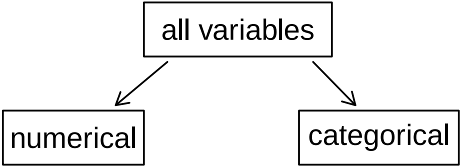
Each **column** of the data matrix contains the values of one variable of all cases.

- A variable is a characteristic of a case, which varies among cases

variable ↓

	num.engs	height	sky	birds.struc
1	2	7000	No Cloud	1
2	2	10	No Cloud	2-10
3	3	400	Some Cloud	1
4	2	100	Overcast	1
⋮	⋮	⋮	⋮	⋮
19302	4	50	No Cloud	1

Type 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**.

- `age`, `body weight`, `annual income` are numerical variables
- `gender`, `blood type`, `nationality` are categorical
- `Zip codes`, `area codes` are *categorical* even though they are numbers since it makes no sense to take average of zip codes

Variables of the Bird-Strike Data

- `num.engs`: Number of engines on the aircraft
- `height`: Feet above ground level
- `sky`: cloud cover, classified as: `No Cloud`, `Some Cloud`, `Overcast`.
- `birds.struc`: Number of birds/wildlife struck: 0, 1, 2-10, 11-100, Over 100.

Variables of the Bird-Strike Data

- `num.engs`: Number of engines on the aircraft
  .... numerical
- `height`: Feet above ground level

- `sky`: cloud cover, classified as: `No Cloud`, `Some Cloud`, `Overcast`.
- `birds.struc`: Number of birds/wildlife struck: 0, 1, 2-10, 11-100, Over 100.

Variables of the Bird-Strike Data

- `num.engs`: Number of engines on the aircraft

  ....numerical
- `height`: Feet above ground level

  ....numerical
- `sky`: cloud cover, classified as: `No Cloud`, `Some Cloud`, `Overcast`.
- `birds.struc`: Number of birds/wildlife struck: 0, 1, 2-10, 11-100, Over 100.

Variables of the Bird-Strike Data

- `num.engs`: Number of engines on the aircraft .... numerical
- `height`: Feet above ground level .... numerical
- `sky`: cloud cover, classified as: `No Cloud`, `Some Cloud`, `Overcast`. .... categorical
- `birds.struc`: Number of birds/wildlife struck: 0, 1, 2-10, 11-100, Over 100.

Variables of the Bird-Strike Data

- `num.engs`: Number of engines on the aircraft .... numerical
- `height`: Feet above ground level .... numerical
- `sky`: cloud cover, classified as: `No Cloud`, `Some Cloud`, `Overcast`. .... categorical
- `birds.struc`: Number of birds/wildlife struck: 0, 1, 2-10, 11-100, Over 100. .... categorical

Histograms

Example: FEV Data

The FEV data are data of a sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970's. The variables include

- `age`: subject's age in years
- `fev`: lung capacity of subject, measured by **forced expiratory volume** (abbreviated as **FEV**), the amount of air an individual can exhale in the first second of forceful breath in liters
- `ht`: subject's height in inches
- `sex`: gender (0 = Female, 1 = Male)
- `smoke`: smoking status (0 = Nonsmoker, 1 = Smoker)

```
  age   fev   ht sex smoke
    9 1.708 57.0   0     0
    8 1.724 67.5   0     0
    7 1.720 54.5   0     0
...(omitted)...
   15 3.211 66.5   0     0
```

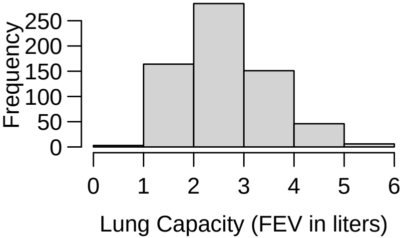
How to Make a Histogram in Frequency Scale?

How to make a histogram for `fev`, a measure of people's lung capacities in the `fevdata`?

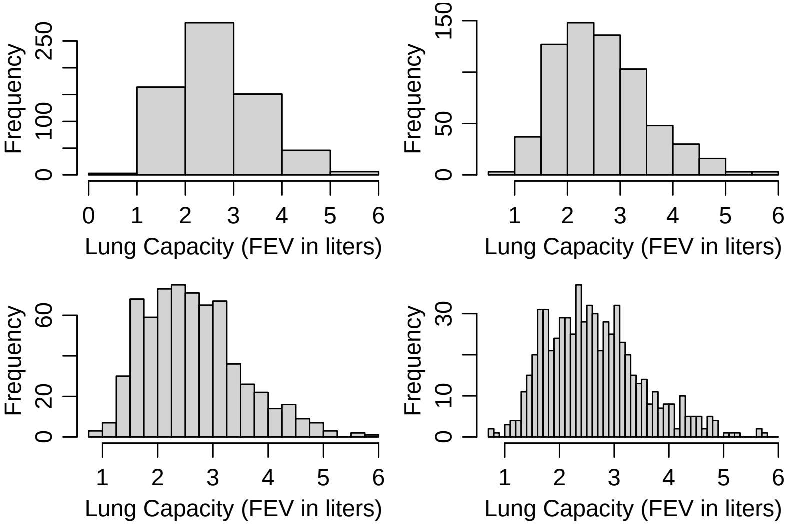
1. Divide the range of values into *class intervals* or *bins*.
2. Count the number of values in each class interval

Interval	0-1	1-2	2-3	3-4	4-5	5-6
Count	3	164	284	151	46	6

3. Draw the histogram and label the axes (with units)



Changing Binwidths Can Alter the Shape of a Histogram



Selection of Binwidth

It is an iterative process — try and try again.

What binwidth should you use?

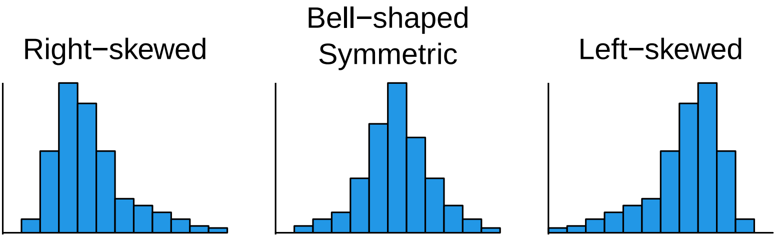
- Not too small that most bins have either 0 or 1 counts
- Not too big that you lose the details in a bin
- (There may not be a unique "perfect" bin size)

General rule: **the more observations, the more bins**.

A Rule of Thumb:

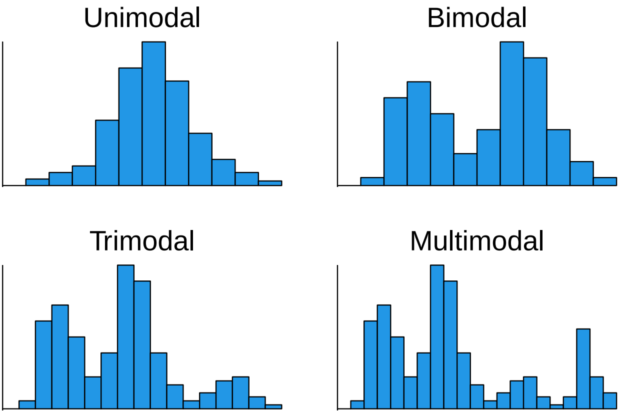
$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

Skewness of Histograms



Direction of Skewness = Direction of the Longer Tail

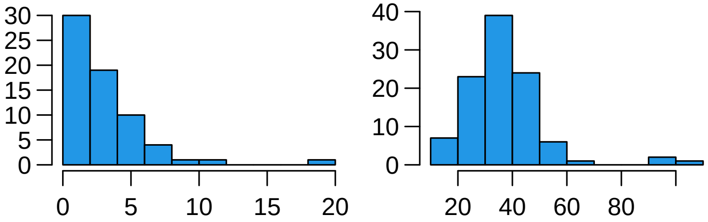
Mode of Histograms (= Number of Peaks)



Two or more modes indicate data might come from two or more distinct populations.

Outliers

We can also check if there are any unusual observations or potential *outliers* from a histogram.

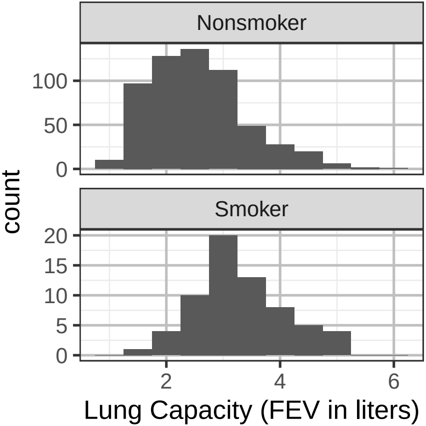


Outliers could be unusual observations or mistakes. Check them!

Stacking Histograms Vertically for Comparison

Stacking histograms vertically makes it easier to **compare the horizontal scale**.

Did smokers or nonsmokers have greater lung capacities in general?



Histogram in Density Scale

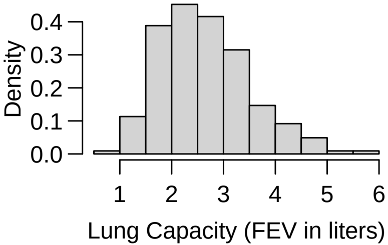
So far we only introduced histograms in *frequency* scale, of which the bar height represents the **count** of **frequency** of observations in the class interval (bin).

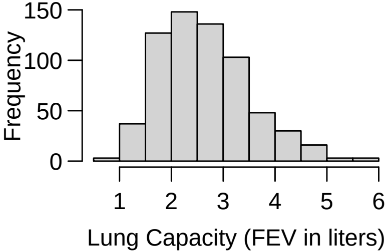
For histogram in *density* scale,

bar area = **proportion** of observations in the bin.

$$\text{Hence bar height} = \frac{\text{\# of observations in the bin}}{\text{(total \# of observations)(bin width)}}$$

Interval	Count	Proportion = $\frac{\text{Count}}{\text{Total}}$	Bar Height = $\frac{\text{Proportion}}{\text{Bin Width}}$
0.5-1	3	$\frac{3}{654} \approx 0.0046$	$\frac{0.0046}{0.5} = 0.0092$
1-1.5	37	$\frac{37}{654} \approx 0.0566$	$\frac{0.0566}{0.5} = 0.1131$
1.5-2	127	$\frac{127}{654} \approx 0.1942$	$\frac{0.1942}{0.5} = 0.3884$
2-2.5	148	$\frac{148}{654} \approx 0.2263$	$\frac{0.2263}{0.5} = 0.4526$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
5.5-6	3	$\frac{3}{654} \approx 0.0046$	$\frac{0.0045}{0.5} = 0.0092$
Total	654	1	



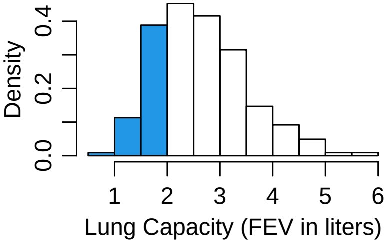


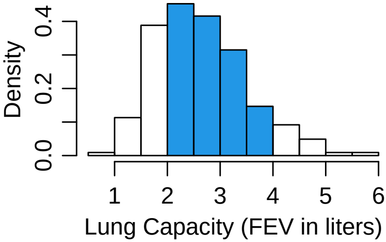
Area Under a Histogram = Proportion

In density scale,

Blue *area* = *proportion* of subjects with FEV between 0.5 to 2.0

Blue *area* = *proportion* of subjects with FEV between 2.0 and 4.0

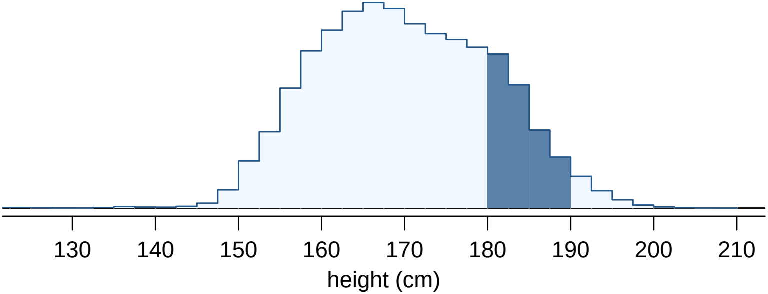




In the density scale, the total area under a histogram is 1 (why?).

Exercise

Below is a histogram of the heights of US adults.



About what percentage of US adults are between 180 to 190 cm tall? Choose the closest percentage.

5% 20% 50% 75%

Summary of Histograms

What to Look in a Histogram?

- **Shape**
  - symmetric or skewed (lopsided)
  - number of modes (peaks)
- **Outliers**
- **Center**: Where is the "middle" of the histogram?
  - typically represented by mean and median
- **Variability**: What big the range the data spread?
  - typically represented by SD and IQR (will introduce shortly)

Keep in mind that

$$\boxed{\text{Area Under a Histogram} = \text{Proportion}}$$

Mean and Median

Mean

The *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, 12.

The mean of the variable is given by:

$$\overline{x} = \frac{4 + 8 + 3 + 5 + 12}{5} = \frac{32}{5} = 6.4.$$

Median

The *median* of a numerical variable 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, 12.

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

Ex 2: If the variable has one more observation: 4, 8, 3, 5, 12, *100*.

data	$\rightarrow$	4	8	3	5	12	100
sorted	$\rightarrow$	3	4	5	8	12	100

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

Median

The *median* of a numerical variable 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, 12.

$$\begin{array}{rcccccc} \text{data} & \rightarrow & 4 & 8 & 3 & 5 & 12 \\ \text{sorted} & \rightarrow & 3 & 4 & \mathit{5} & 8 & 12 \\ & & & & \downarrow & & \\ & & & & \textit{Median} & & \end{array}$$

Ex 2: If the variable has one more observation: 4, 8, 3, 5, 12, *100*.

$$\begin{array}{rccccccc} \text{data} & \rightarrow & 4 & 8 & 3 & 5 & 12 & 100 \\ \text{sorted} & \rightarrow & 3 & 4 & 5 & 8 & 12 & 100 \end{array}$$

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

Median

The *median* of a numerical variable 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, 12.

$$
\begin{array}{rccccccc}
\text{data} & \rightarrow & 4 & 8 & 3 & 5 & 12 \\
\text{sorted} & \rightarrow & 3 & 4 & \mathit{5} & 8 & 12 \\
 & & & & \downarrow & & \\
 & & & & \textit{Median} & &
\end{array}
$$

Ex 2: If the variable has one more observation: 4, 8, 3, 5, 12, *100*.

$$
\begin{array}{rccccccc}
\text{data} & \rightarrow & 4 & 8 & 3 & 5 & 12 & 100 \\
\text{sorted} & \rightarrow & 3 & 4 & \mathit{5} & \mathit{8} & 12 & 100
\end{array}
$$

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

Median is More Robust to Extreme Values Than Mean

Example. If the variable has a new 6th observation of value 100,

$$4, 8, 3, 5, 12, \textit{100}$$

The new mean of the variable becomes

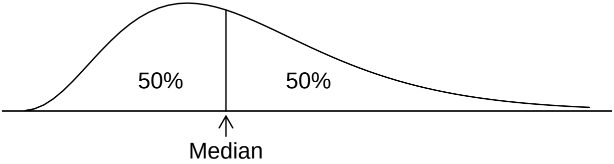
$$\bar{x} = \frac{4 + 8 + 3 + 5 + 13 + 100}{6} = \frac{132}{6} = 22.$$

Data	Mean	Median
4, 8, 3, 5, 12	6.4	5
4, 8, 3, 5, 12, 100	22	6.5

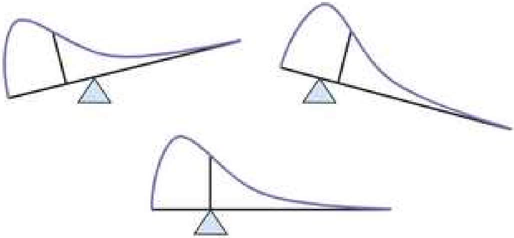
The median is less affected by the extreme value (100) than the mean. We say the median is more *robust* than the mean.

Median and Mean of a Histogram

The **median** divides the area of a histogram evenly.



The **mean** is the **balance point** of the distribution/histogram, if it were a solid mass.



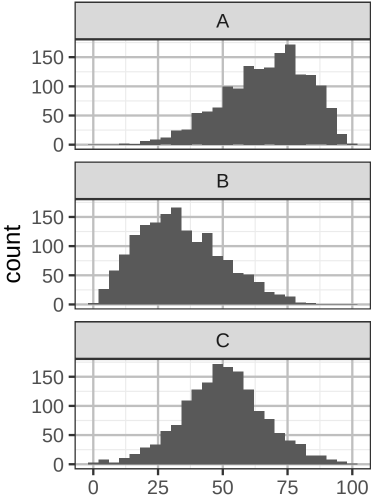
It's always true that

$$\sum\nolimits_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$\begin{aligned}
\sum\nolimits_{i=1}^{n} (x_i - \bar{x}) &= \sum\nolimits_{i=1}^{n} x_i - \sum\nolimits_{i=1}^{n} \bar{x} \\
&= \sum\nolimits_{i=1}^{n} x_i - n\bar{x} && \text{(adding up } n \text{ identical things)} \\
&= \sum\nolimits_{i=1}^{n} x_i - n\frac{1}{n} \sum\nolimits_{i=1}^{n} x_i && \text{(definition of } \bar{x}\text{)} \\
&= \sum\nolimits_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i && \text{(}n\text{'s cancel)} \\
&= 0.
\end{aligned}$$

Exercise (Comparing the Centers of Histograms)

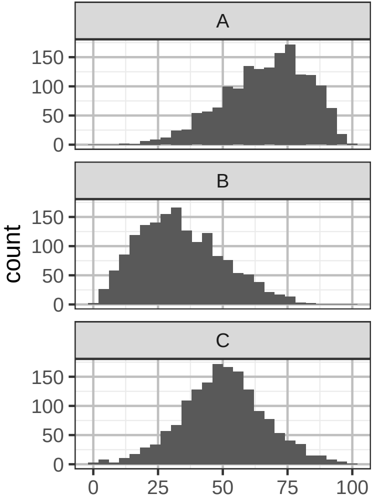
Please order the **means** of the 3 histograms from low to high.



Exercise (Comparing the Centers of Histograms)

Please order the **means** of the 3 histograms from low to high.

$B < C < A$

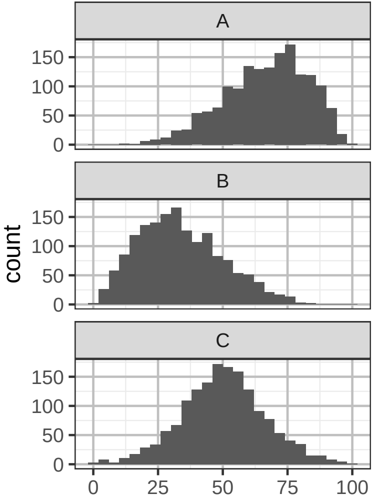


Exercise (Comparing the Centers of Histograms)

Please order the **means** of the 3 histograms from low to high.

$B < C < A$

How about the **medians**?



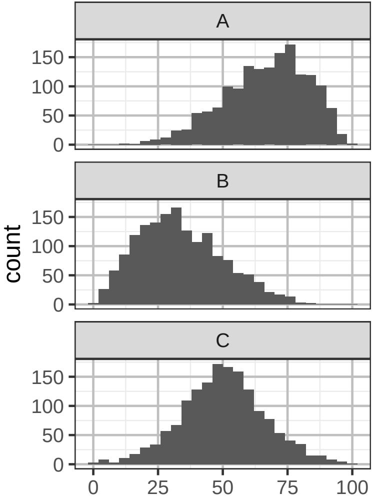
Exercise (Comparing the Centers of Histograms)

Please order the **means** of the 3 histograms from low to high.

$B < C < A$

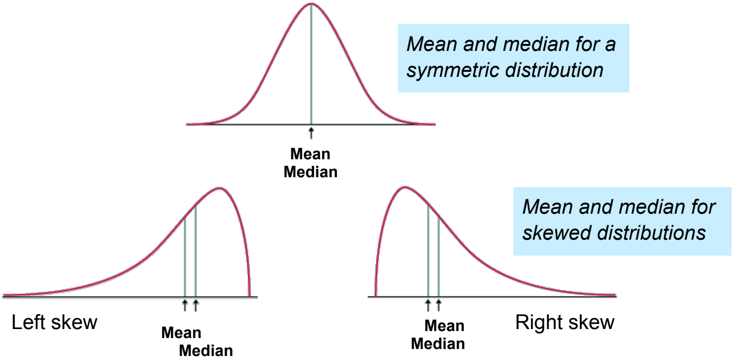
How about the **medians**?

$B < C < A$

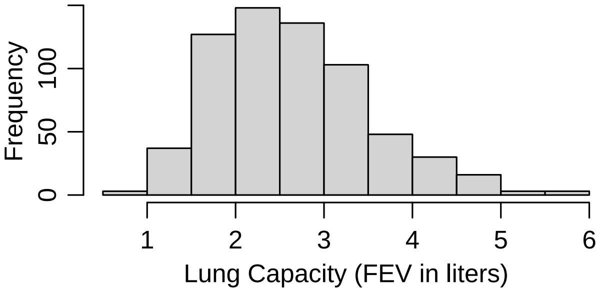


Mean vs. Median and Skewness

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



Example: Mean and Median of FEV



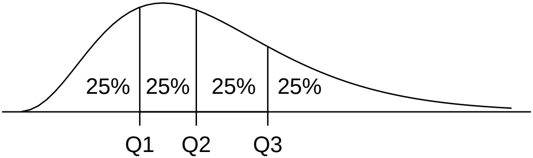
Mean = 2.6368 > Median = 2.5475

The mean is slightly higher than the median as the histogram is right-skewed.

Five-Number Summary & Boxplots

Quartiles, IQR, Five-Number Summary

- **Quartiles** divide data into 4 even parts
  - **first quartile $\mathbf{Q_1}$** = 25th percentile:
    25% of data fall below it and 75% above it
  - **second quartile $\mathbf{Q_2}$** = median = 50th percentile
  - **third quartile $\mathbf{Q_3}$** = 75th percentile
    75% of data fall below it and 25% above it



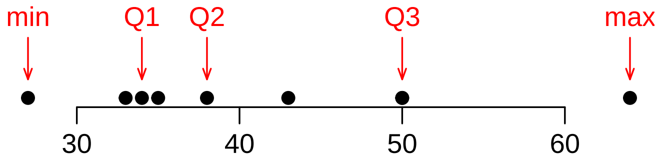
- **Interquartile Range** $(IQR) = Q_3 - Q_1$
- **Five-Number Summary**:

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

Example 1 — Five Number Summary

For the 9 numbers: 43, 35, 50, 33, 38, 53, 64, 27, 34

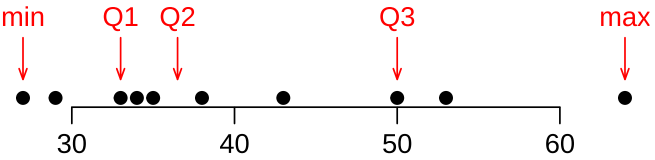
```
x = c(43, 35, 50, 33, 38, 50, 64, 27, 34)
sort(x)
## [1] 27 33 34 35 38 43 50 50 64
summary(x)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   34.00   38.00   41.56   50.00   64.00
fivenum(x)
## [1] 27 34 38 50 64
```



Example 2

For the 10 numbers: 43, 35, 50, 33, 38, 53, 64, 27, 34, 29

```r
x = c(43, 35, 50, 33, 38, 53, 64, 27, 34, 29)
sort(x)
##  [1] 27 29 33 34 35 38 43 50 53 64
summary(x)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   33.25   36.50   40.60   48.25   64.00
fivenum(x)
## [1] 27.0 33.0 36.5 50.0 64.0
```



In fact, statisticians have no consensus on the calculation of quartiles. There are several formulas for quartiles, varying from book to book, software to software.

Different commands in R sometimes give different quartiles.

E.g., for the 10 numbers in Example 2,

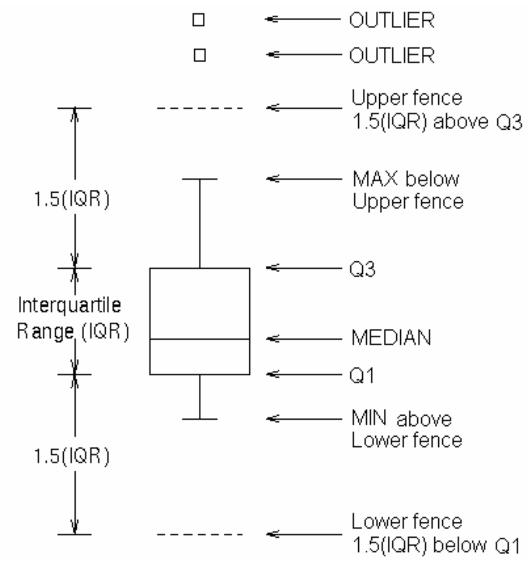
```
x = c(43, 35, 50, 33, 38, 53, 64, 27, 34, 29)
summary(x)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.00   33.25   36.50   40.60   48.25   64.00
fivenum(x)
## [1] 27.0 33.0 36.5 50.0 64.0
```

Don't worry about the formula. Just keep in mind that

$$\boxed{\text{quartiles divide data into 4 even parts}}$$

In HWs, just report whatever values your software gives.

Box-and-Whiskers Plot (also called Boxplot)

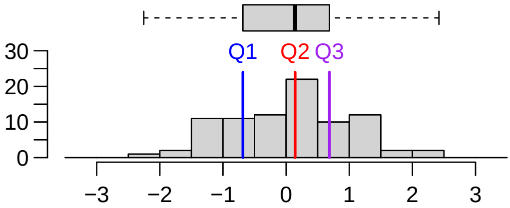


Boxplot v.s. Histogram — Skewness

What does a boxplot look like if the histogram is symmetric?

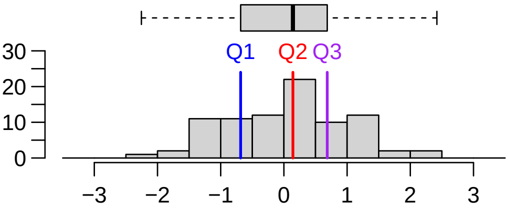
Boxplot v.s. Histogram — Skewness

What does a boxplot look like if the histogram is symmetric?



Boxplot v.s. Histogram — Skewness

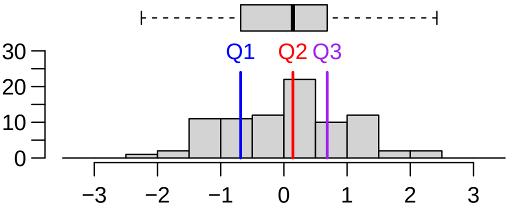
What does a boxplot look like if the histogram is symmetric?



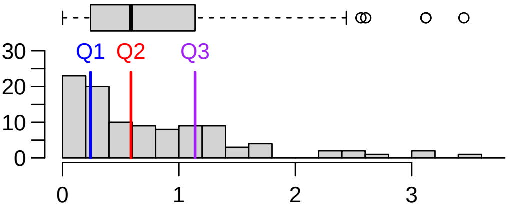
Ditto, if right-skewed?

Boxplot v.s. Histogram — Skewness

What does a boxplot look like if the histogram is symmetric?



Ditto, if right-skewed?

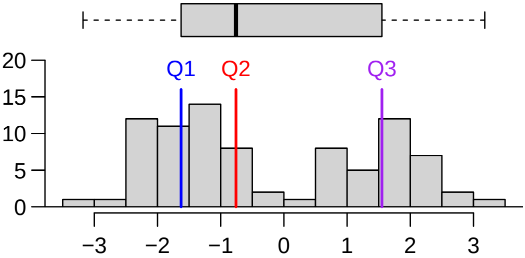


Boxplot v.s. Histogram — Modality

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

Boxplot v.s. Histogram — Modality

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



Side-by-Side Boxplots

Just like histograms, boxplots of related distributions are often placed side-by-side for comparison.