Statistics is the science of data, and is used to evaluate claims.

Ex: I make 80% of free throws I shoot.

## Chapter 1: Picturing Distributions with Graphs

Def: An individual is an object described by data.

Ex: Person, city, animal, company.

Def: A variable is a characteristic of an individual.

Ex: Age, population, species, profit.

Ex: We randomly select 4 people in the US and ask them to report their age and gender. We also ask them what state they're living in.

| State | Age | Reported Gender |
|---|---|---|
| Kentucky | 61 | Female |
| Florida | 27 | Female |
| Wisconsin | 27 | Male |
| California | 33 | Female |

catagorical quantative catagorical

4 individuals and 3 variables measured for each individual

Def: A variable is <u>quantitative</u> if it takes numerical values and arithmetic makes sense.

<u>Def</u>: A variable is <u>catagorical</u> if it is not quantative.

Now we ask for zip codes

| State | Age | Reported Gender | Zip |
|---|---|---|---|
| Kentucky | 61 | Female | 41375 |
| Florida | 27 | Female | 93402 |
| Wisconsin | 27 | Male | 97403 |
| California | 33 | Female | 49102 |

↗ catagorical!

<u>Ex</u>: A study classifies bison in Yellowstone as young or adult. State the

individuals, variables, and the type of variable.

Bison, age, catagorical

Def: The distribution of a variable is the information of both its possible values and how often they occur.

Ex:

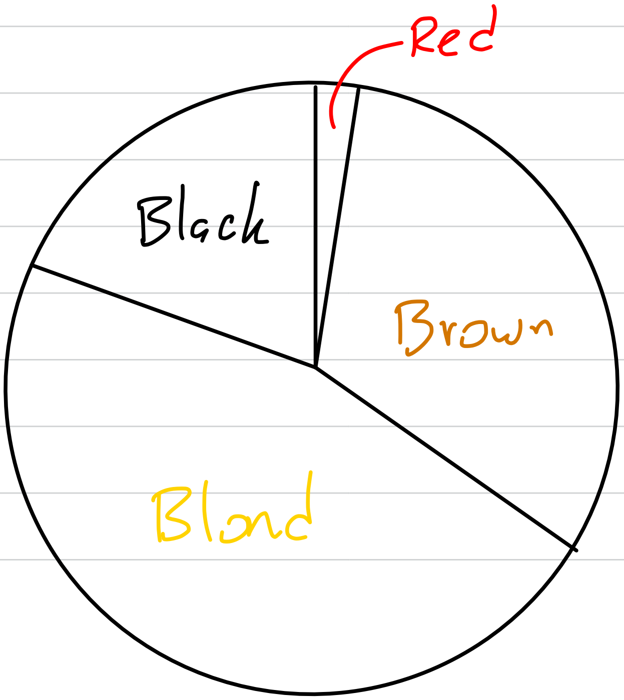
| Student ID | Hair color |
|---|---|
| 003 | Red |
| 005 | Brown |
| 035 | Brown |
| 089 | Black |

not a distribution

| Hair color | % of students w/ this color |
|---|---|
| Red | 2 % |
| Brown | 35 % |
| Blond | 43 % |
| Black | 20 % |

distribution

Pie chart



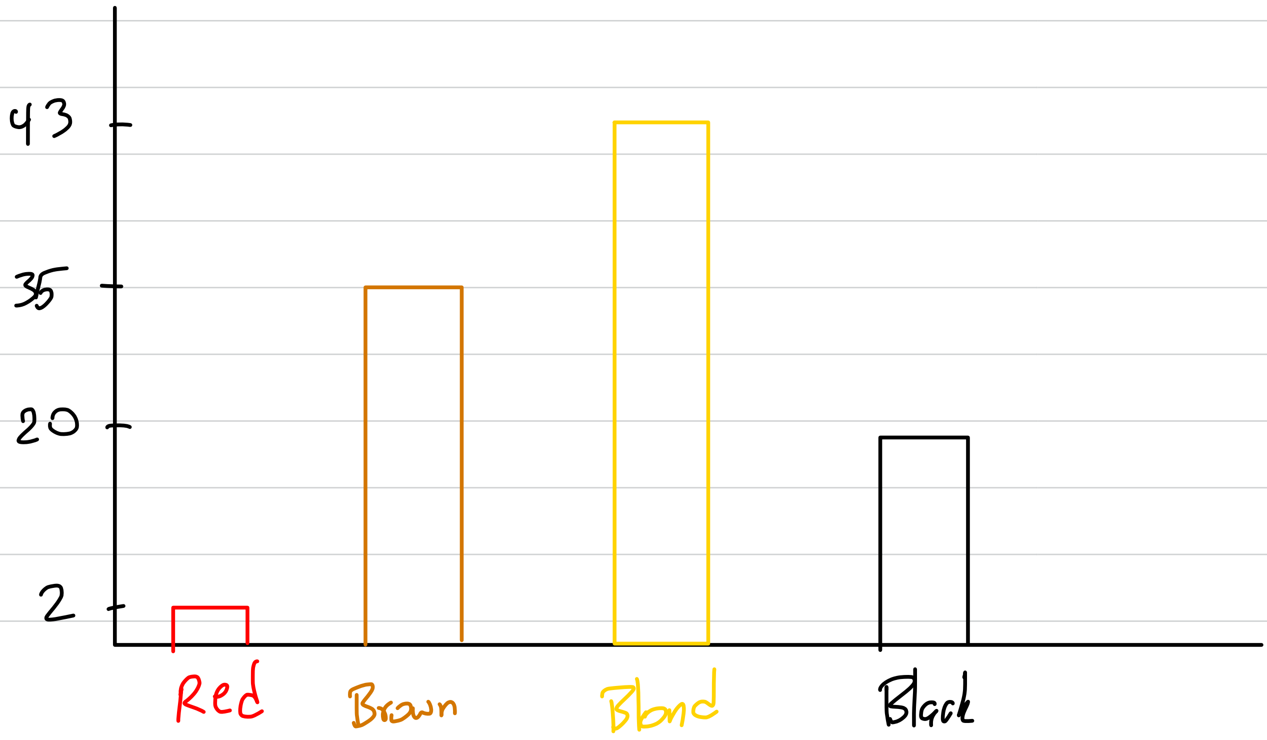
**Comment:** Only use pie charts when the values the variable can take are **mutually exclusive** — i.e. every individual has at most one value. Hair color is mutually exclusive since you can have at most one. A survey asking which types of soda you'd had in the past month would not be mutually exclusive since you could have had more than one type.

| | % pop |
|---|---|
| Sprite | 30 % |
| Dr. Pepper | 25 % |

this doesn't reflect the people who have had both

| Hair color | % of students w/ this color |
|---|---|
| Red | 2 % |
| Brown | 35 % |
| Blond | 43 % |
| Black | 20 % |

Bar graph:



Ex

| Music source | % of 12-24 year olds who have used it |
|---|---|
| Radio | 72 |
| YouTube | 77 |
| iTunes | 47 |

Don't use a pie chart, because the different music sources aren't mutually exclusive!

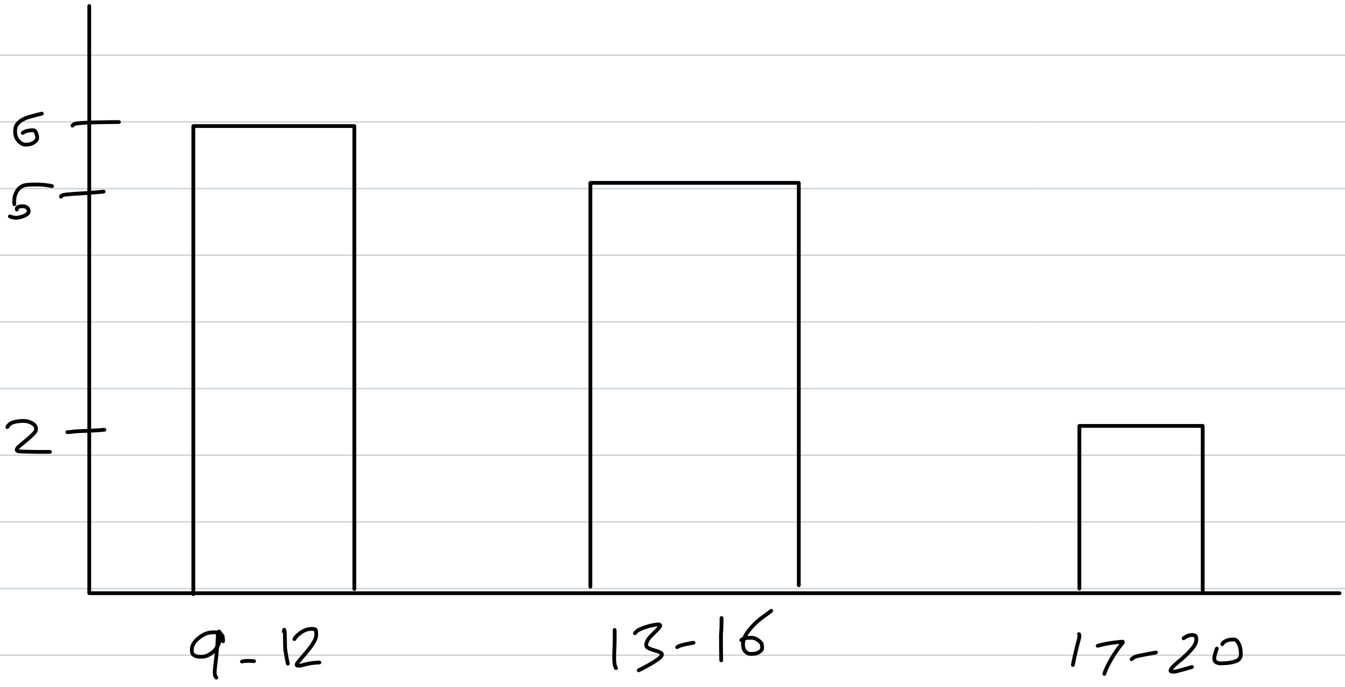
Histograms: when given a sample of individuals, you can make a histogram by dividing the data into ranges (called classes) and counting the number of individuals in each class. Then we make a bar graph of the result. This roughly approximates the distribution.

Ex: We get a set of ages:

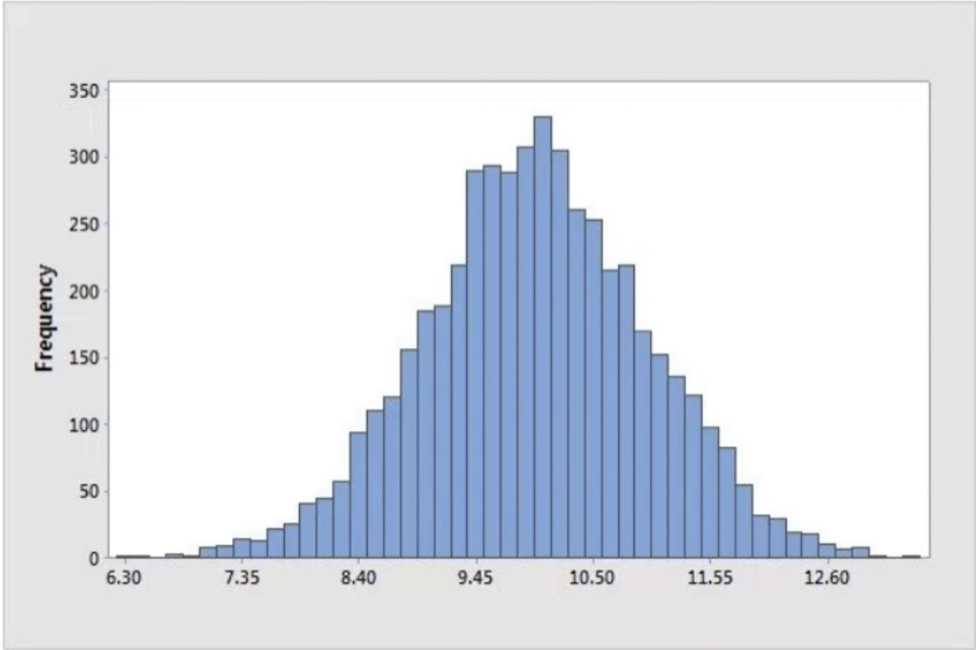
9 10 10 11 12 12 14 15 15

16 16 18 20

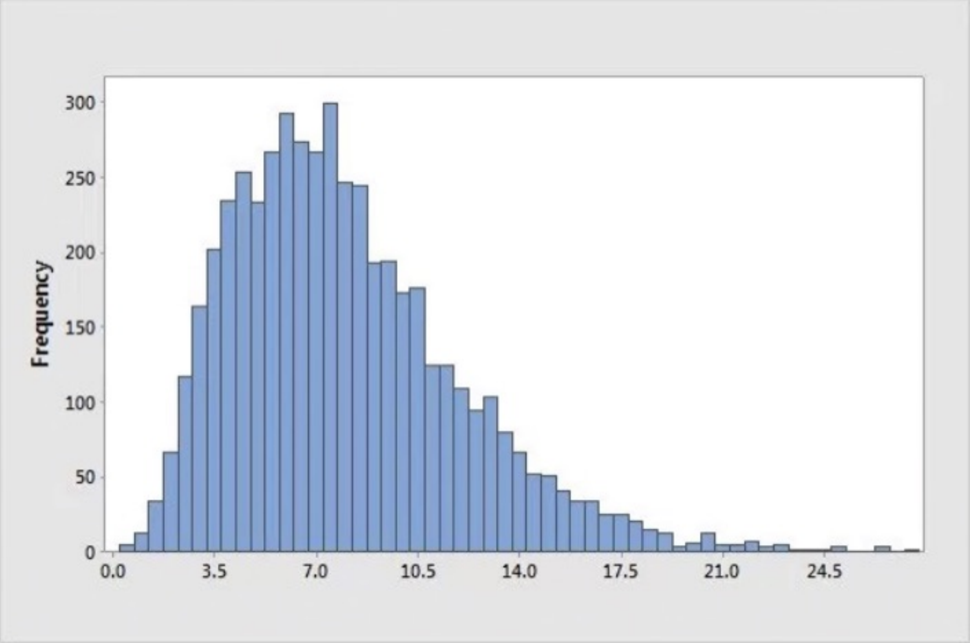
Classes: $\underbrace{9-12}_{6}, \underbrace{13-16}_{5}, \underbrace{17-20}_{2}$



A **symmetric** distribution.
Ex: Heights of young women, Lengths of bird bills

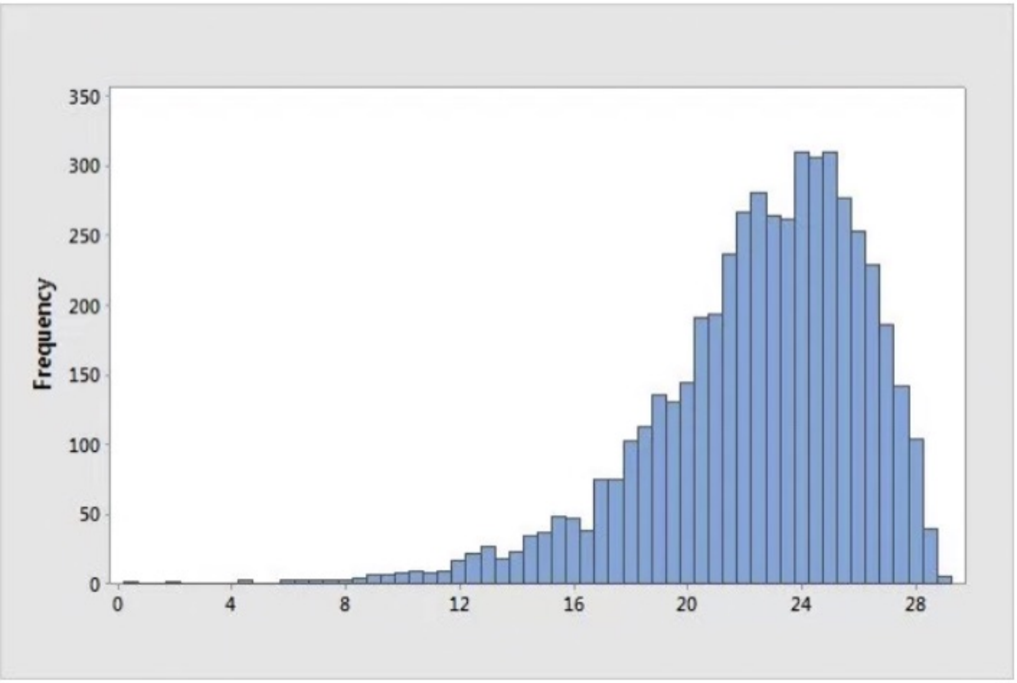


Right-skewed

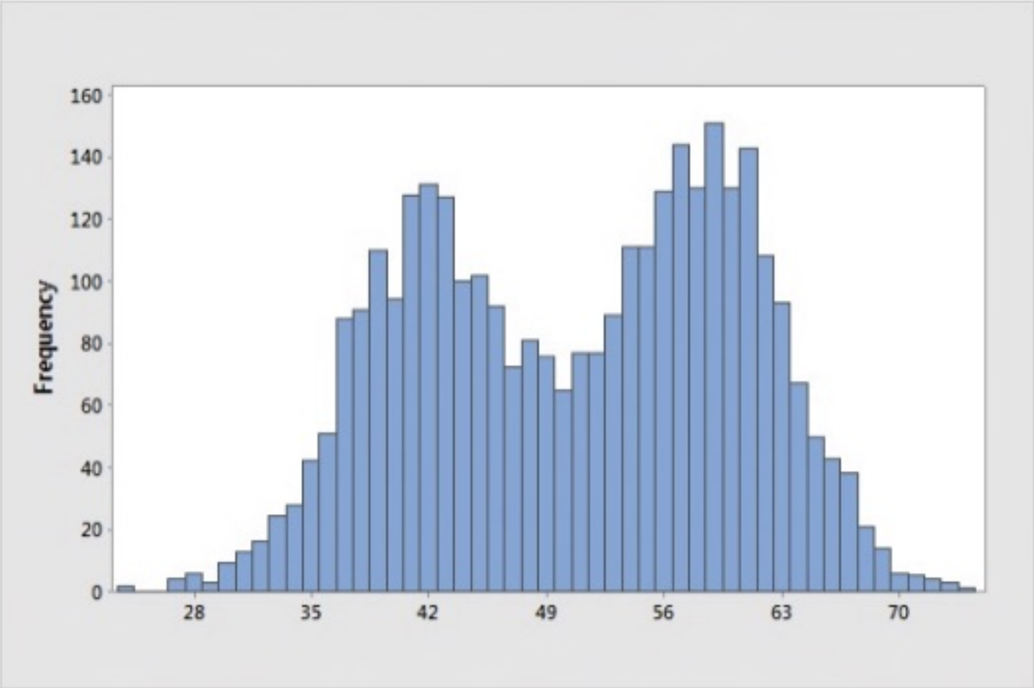


Ex: incomes

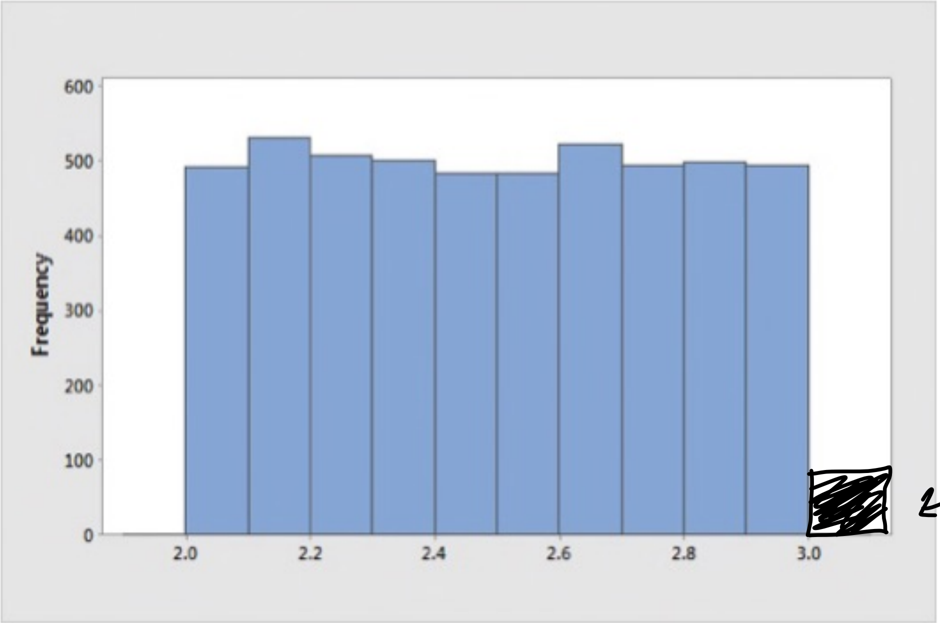
A **left-skewed** distribution.
Ex: Grades on an easy test



A **bimodal** distribution.
Ex: Exam scores when one group studied and another didn't



An **approximately uniform** distribution.
Ex: Rolling a die



outlier

Def: The center of the distribution is the mean or median. The variability is roughly how spread out the distribution is. Outliers are individuals who don't fit the pattern.

Def: Given a set $\overbrace{\text{of data}}^{\text{quantatire}}$, we can form a stem-and-leaf plot: take all of the numbers and split them into the last digit and all the other digits. Then write the second piece (i.e. the prefix) and all the final digits with that prefix.

Ex: 9 10 10 11 12 12 14 15 15 16 16 18 20

| | |
|---|---|
| 0 | 9 |
| 1 | 0 0 1 2 2 4 5 5 6 6 8 |
| 2 | 0 |

Ex: 5, 13, 18, 32, 91  40, 45, 19, 60

| | |
|---|---|
| 0 | 5 |
| 1 | 3 8 |
| 3 | 2 |
| 9 | 1 |

correct

Webwork + Textbook:

| | | |
|---|---|---|
| | 0 | 5 |
| 9 | 1 | 38 |
| | 2 | |
| | 3 | 2 |
| 05 | 4 | |
| | 5 | |
| 0 | 6 | |
| | 7 | |
| | 8 | |
| | 9 | 1 |

Comment: we can also split the stems:

| | |
|---|---|
| 0 | 9 |
| 1 | 0 0 1 2 2 4 5 5 6 6 8 |
| 2 | 0 |

||

| | |
|---|---|
| 0 | 9 |
| 0 | |
| 1 | 0 0 1 2 2 |
| 1 | 4 5 5 6 6 8 |
| 2 | 0 |
| 2 | |

## Chapter 2 : Describing Distributions with Numbers

Ex: A list of travel times to work in North Carolina:

30, 20, 10, 40, 25, 20, 10, 60,
15, 40, 5, 30, 12, 10, 10

How to calculate center? One way is taking the average.

Def Given a set of data $x_1, \dots, x_n$, the mean of the data is

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Ex: $$\bar{x} = \frac{30 + 20 + 10 + \cdots + 12 + 10 + 10}{\underset{\uparrow}{15}} = 22.5$$

15 samples

Ex: 5, 10, 15, 200 ← Right-skewed

$$\bar{x} = \frac{5+10+15+200}{4} = \frac{230}{4} = 57.5$$

Comment: In a skewed distribution, the mean is drawn toward the skew (i.e. the tail). We say the mean is not a resistant measure of center.

Def: Let $x_1, \dots, x_n$ be a set of data. The median is $M$, defined by:

① if $n$ is odd, then $M$ is the data point such that as many $x_i$ are greater than $M$ as are less than $M$

② if $n$ is even, $M$ is the average of the two numbers with as many $x_i$ greater than them as there are $x_i$ less than them

Ex: 30, 20, 10, 40, 25, 20, 10, 60,
15, 40, 5, 30, 12, 10, 10

First arrange from smallest to largest

$\overbrace{5, 10, 10, 10, 10, 12, 15}^{7}, \boxed{20,}\ \overbrace{20, 25, 30, 30,}^{7}$

$\underbrace{40, 40, 60}$

median

15 data points, which is odd, so we want the number "in the middle"

Ex: $5, \boxed{10, 15}, 200$

average is $\frac{10+15}{2} = 12.5$

median : 12.5

**Comment**: The median *is* a resistant measure of center.

*Ex*: you roll a die. If you roll a 1-5, you get nothing. If you roll a 6, you get $100. What should you expect to get on average from rolling 6 times?

$$0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 100$$

median: $0$

mean: $\frac{100}{6}$ ← this is better for our purposes!

How do we measure variability?

Start small: min and max

Ex. $\boxed{5}$, 10, 10, 10, 10, 12, 15, $\boxed{20}$, 20, 25, 30, 30,

40, 40, $\boxed{60}$

5, 60

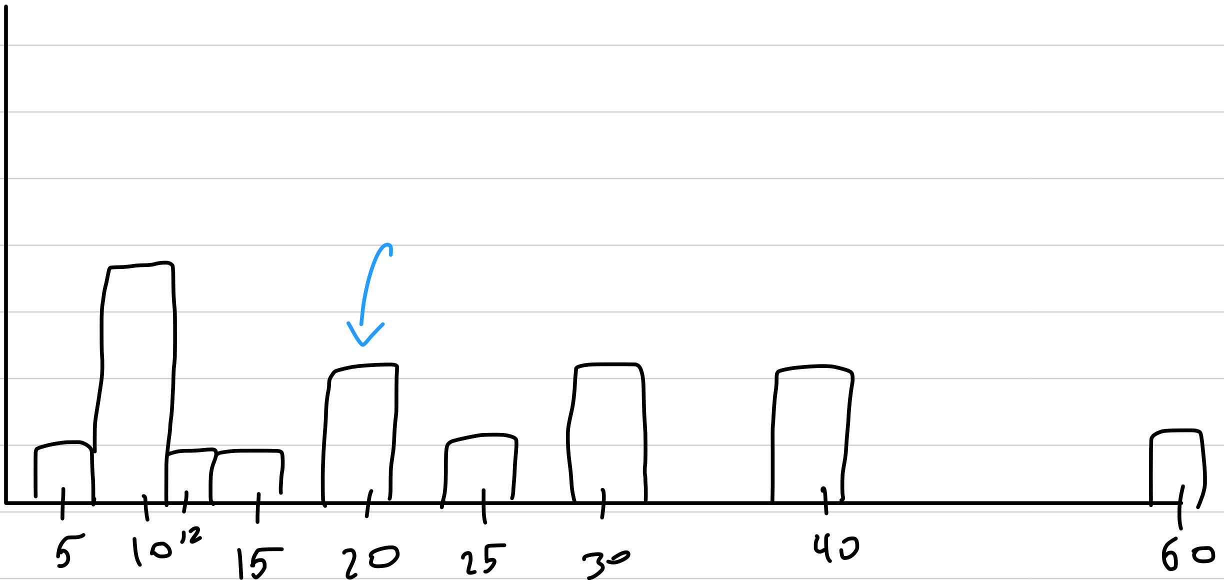
Better: min, median, max

5, 20, 60

↑

gap indicates that

this is a right-skewed distribution



Def: The first and third quartiles, $Q_1$ and $Q_3$, are the medians of the two halves of the data, not including the median of the whole data.

$$\boxed{\boxed{5}, 10, 10, \boxed{10}, 10, 12, 15,}\ \boxed{20}, \boxed{20, 25, 30, \boxed{30}, 40, 40, \boxed{60}}$$

(you could say that $Q_2 = 20$)

$Q_1 = 10$

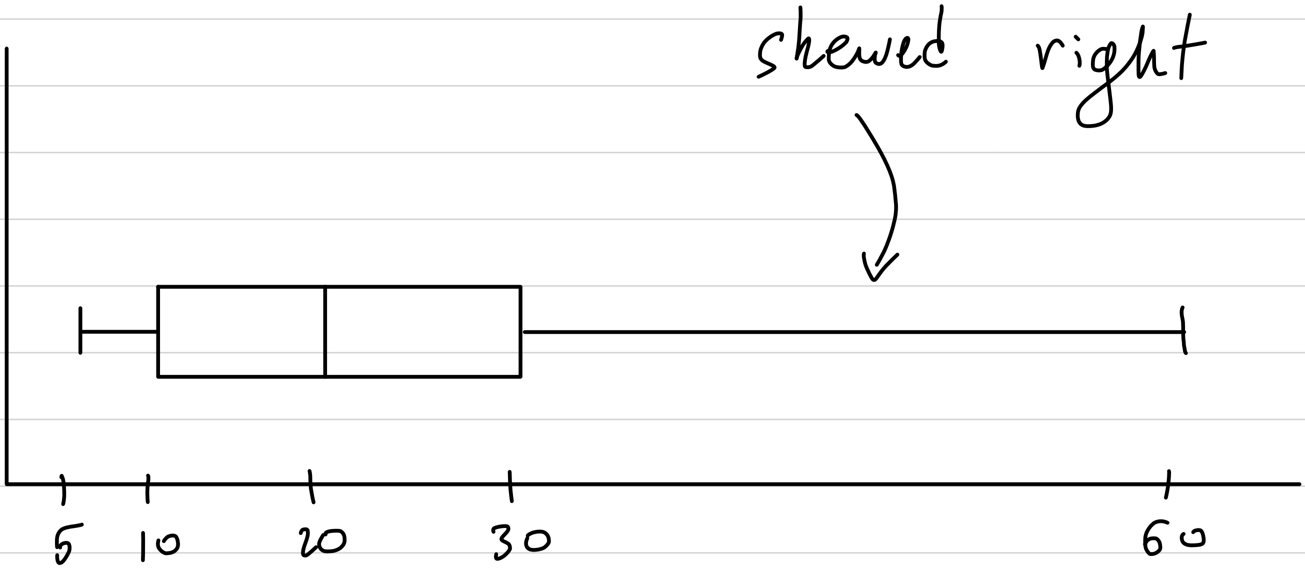
$Q_3 = 30$

Def: The 5-number summary of a set of data is min, $Q_1$, median, $Q_3$, max

Ex: 5, 10, 20, 30, 60

All 4 gaps have the same # of data points.

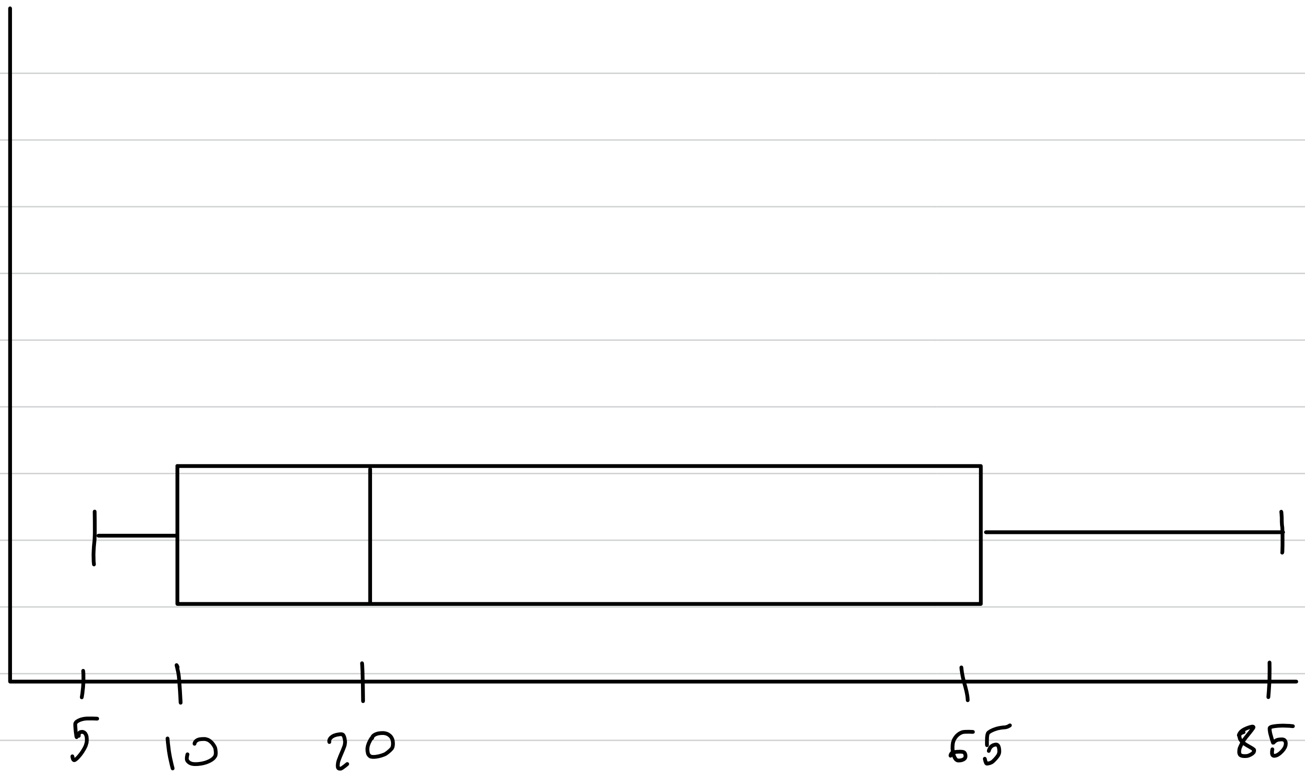
Box plots:



Ex: Draw a box plot of

10, 30, 5, 85, 65, 20, 10.

$\boxed{\boxed{5}, \boxed{10}, 10}, \boxed{20}, \boxed{30, \boxed{65}, \boxed{85}}$



Def: The interquartile range, or IQR, is given by $IQR = Q_3 - Q_1$

Def: An outlier in a data set is any point more than 1.5 IQR above $Q_3$ or below $Q_1$.

Ex: 10, 30, 5, 1000, 65, 20, 10.

5-num: 5, 10, 20, 65, 1000

$Q_1 = 10$
$Q_3 = 65$

$IQR = 65 - 10 = 55$

$1.5\, IQR = 82.5$

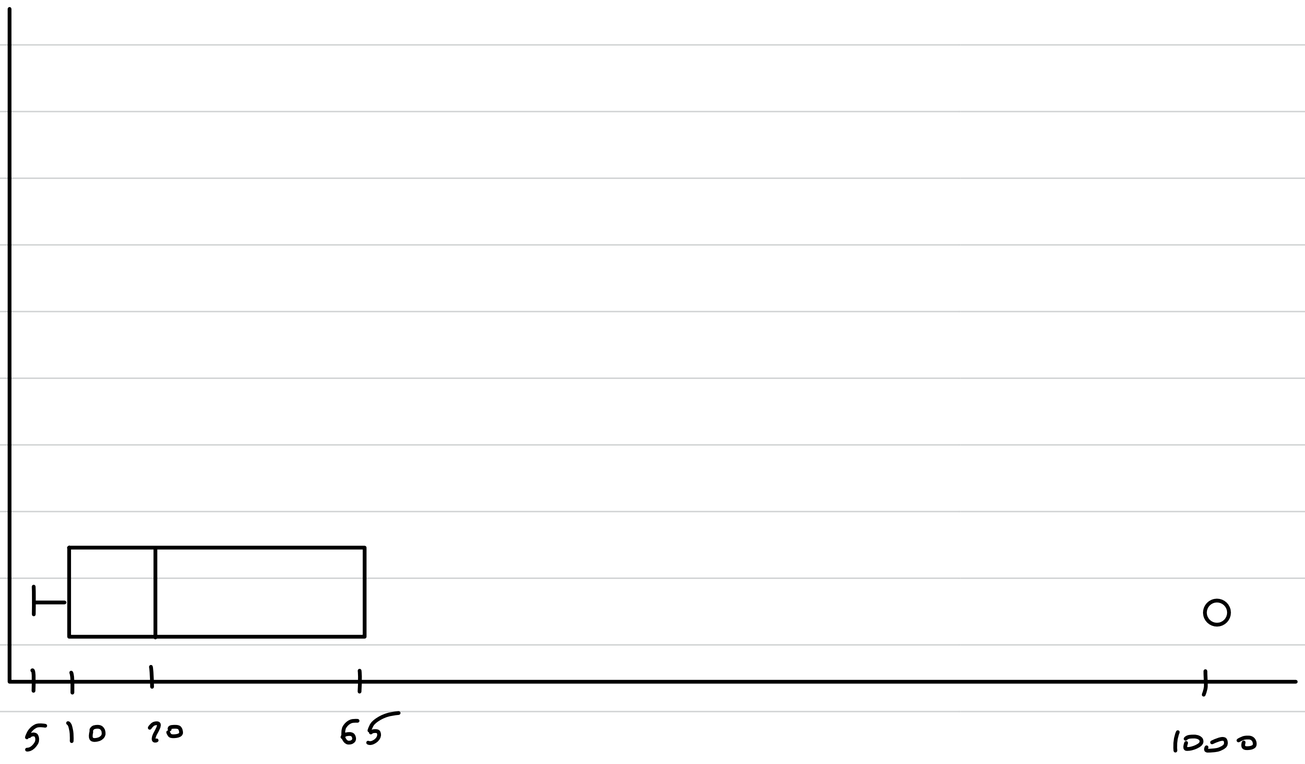
$Q_3 + 82.5 = 147.25$

$Q_1 - 82.5 = -72.5$

outliers are anything not between $-72.5$ and $147.25$. So 1000 is an outlier.

Represent outliers by modifying the

box-plot : make the whiskers only reach the non-outliers.



The 5-num summary is a resistent measure of variability (but it's a little lacking)