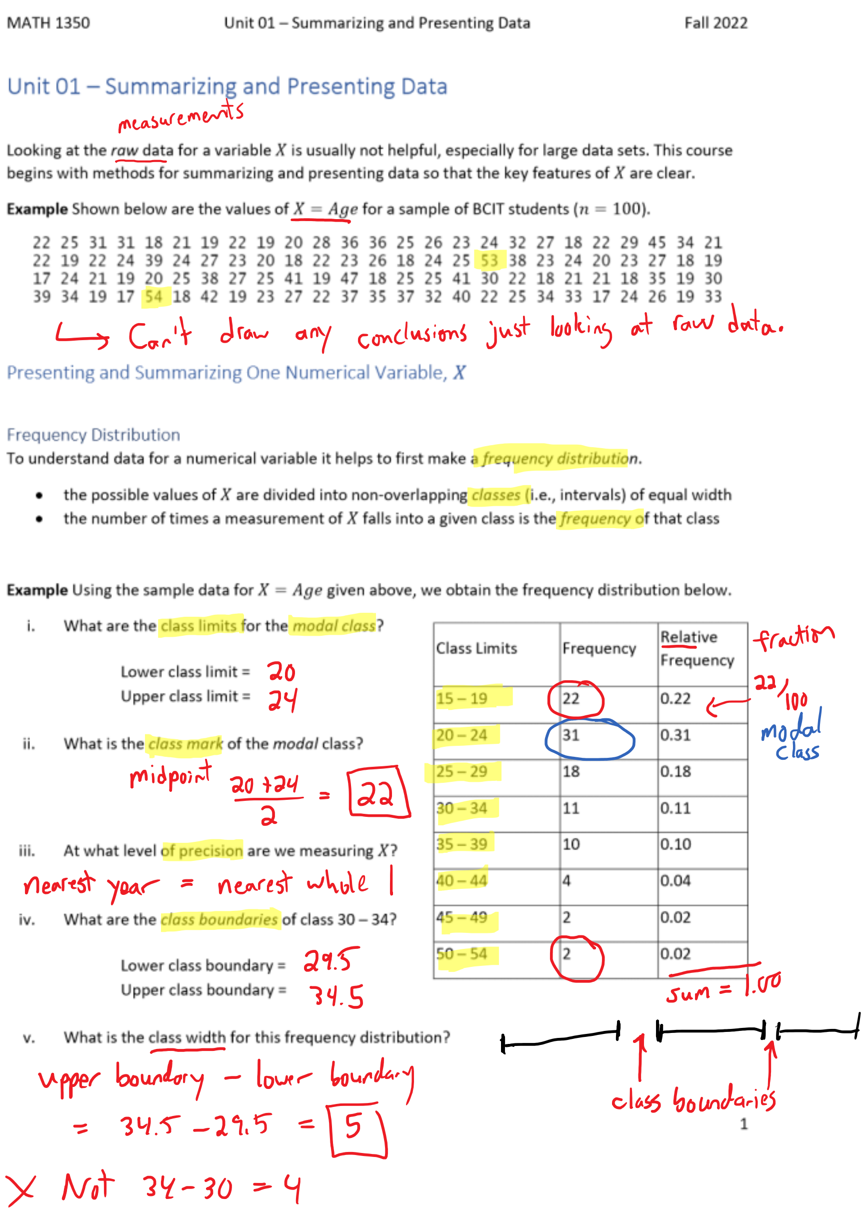
Unit 1 - B - Summarizing and Presenting Data

July 5, 2022 11:11 AM



Unit 01 – Summarizing and Presenting Data

measurements

Looking at the *raw* data for a variable X is usually not helpful, especially for large data sets. This course begins with methods for summarizing and presenting data so that the key features of X are clear.

Example Shown below are the values of $X = Age$ for a sample of BCIT students ($n = 100$).

22 25 31 31 18 21 19 22 19 20 28 36 36 25 26 23 24 32 27 18 22 29 45 34 21
22 19 22 24 39 24 27 23 20 18 22 23 26 18 24 25 53 38 23 24 20 23 27 18 19
17 24 21 19 20 25 38 27 25 41 19 47 18 25 25 41 30 22 18 21 21 18 35 19 30
39 34 19 17 54 18 42 19 23 27 22 37 35 37 32 40 22 25 34 33 17 24 26 19 33

↳ Can't draw any conclusions just looking at raw data.

Presenting and Summarizing One Numerical Variable, X

Frequency Distribution

To understand data for a numerical variable it helps to first make a *frequency distribution*.

- the possible values of X are divided into non-overlapping *classes* (i.e., intervals) of equal width
- the number of times a measurement of X falls into a given class is the *frequency* of that class

Example Using the sample data for $X = Age$ given above, we obtain the frequency distribution below.

Class Limits	Frequency	Relative Frequency
15 – 19	22	0.22
20 – 24	31	0.31
25 – 29	18	0.18
30 – 34	11	0.11
35 – 39	10	0.10
40 – 44	4	0.04
45 – 49	2	0.02
50 – 54	2	0.02

fraction

$22/100$

modal class

sum = 1.00

i. What are the class limits for the *modal class*?

Lower class limit = 20
Upper class limit = 24

ii. What is the *class mark* of the *modal* class?

midpoint $\frac{20+24}{2} = \boxed{22}$

iii. At what level of precision are we measuring X?

nearest year = nearest whole 1

iv. What are the *class boundaries* of class 30 – 34?

Lower class boundary = 29.5
Upper class boundary = 34.5

v. What is the class width for this frequency distribution?

upper boundary – lower boundary

$= 34.5 - 29.5 = \boxed{5}$

✗ Not $34 - 30 = 4$

class boundaries

Terminology for Classes

The terminology here can be confusing!

Lower Class Limit – the smallest value of X in a class (based on the level of precision)

Upper Class Limit – the largest value of X in a class (based on the level of precision)

Class Mark of a class = the average of its lower class limit and its upper class limit

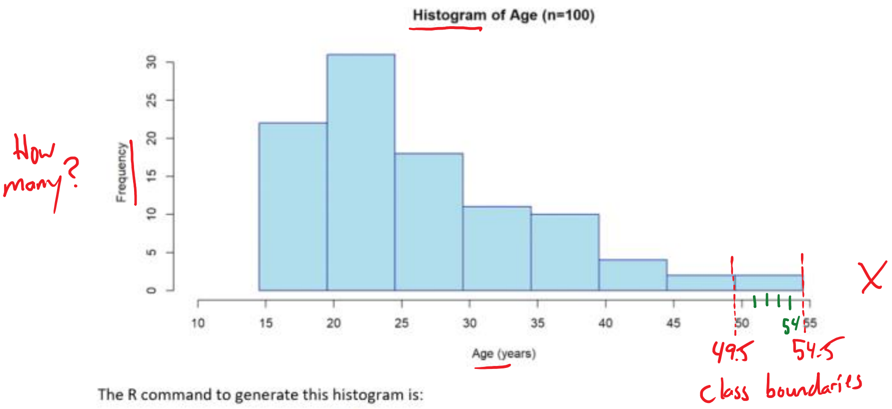
Class Boundaries – the numbers in *between* the upper class limit for one class and the lower class limit for the next class.

Class Width - is the distance from one class boundary to the next class boundary

Histograms

A *histogram* is a visual representation of the corresponding frequency distribution. You need to determine the *class boundaries* to correctly position the left and right sides of the rectangles.

Example The frequency distribution in the previous example gives the following histogram.



The R command to generate this histogram is:

```
hist(data$Age, freq=TRUE,
     breaks=seq(14.5, 54.5, 5),
     col="lightblue", border="darkblue",
     xlim=c(10, 55),xlab="Age (years)",xaxp=c(10,55,9),
     ylab="Frequency",
     main="Histogram of Age (n=100)")
```

More on Classes

- each rectangle goes from one *class boundary* to the next *class boundary*
    - in R, specify classes precisely using the `breaks` option
- there must be no gaps between the rectangles (unless there is an empty class)
- the classes should be of equal width and must cover all observed values of X.
- the number of classes should be approximately $\sqrt{n}$, where n is the sample size.

Statistics and Parameters (Numerical Summaries of One Variable, X)

We just covered how to summarize a variable X *graphically*. To summarize a variable *numerically*, we have two primary types of statistics:

- measures of *center* – "What is a typical or central value of X?"
- measures of *spread* – "How much does X vary around its center? How spread out is X?"

❖ Note: We call these measures "statistics" when we are summarizing *sample* data. They are called "parameters" when we are summarizing *population* data.

Measures of Center

The three measures of center we will use are:

1. Mean – average
2. Median – middle value
3. Mode – most frequent

Mean - the *arithmetic mean* of a numerical data set is the familiar "average" you are used to. In R, use the function `mean()`.

Sample mean: $\bar{X} = \dfrac{X_1+X_2+\ldots+X_n}{n}$ (statistic) n is the sample size

"X-bar"

Population mean: $\mu = \dfrac{X_1+X_2+\ldots+X_N}{N}$ (parameter) N is the population size

"mu"

Note N is probably much larger than n.

Note also: μ is a fixed number even if we don't know it.
$\bar{X}$ is a random Variable since it depends on the Sample.

Median - the *median* of a set of numerical values is the *middle* when the values are sorted *in increasing order*. The median is denoted by $\tilde{X}$ (pronounced "x tilde") or Q_2. In R, use the function `median()`.

- For an *odd* number of data values, the median is in the exact middle of the data set.

"2nd Quartile"
"Median"

20 40 53 64 75 82 99

$Q_2 = 64$

$n = 7$
$\frac{1}{2} \times n = \frac{1}{2} \times 7 = 3.5$ ↗ 4

- For an *even* number of data values, the median is the mean of the two middle numbers.

12 23 27 30 70 80 90 91

$Q_2 = \dfrac{30+70}{2} = 50$

$n = 8$
$\frac{1}{2} \times n = \frac{1}{2} \times 8 = 4$

Mode - the *mode* of a data set is the value that occurs *most frequently*. There is no built-in R function!

- the *mode* is also called the *modal value of X*
- If two values of X occur with the same greatest frequency, the data set is said to be *bimodal*.
- If there are more than two modal values, the data set is said to be *multi-modal*.
- If no value is repeated, we say that there is no mode.

Example Calculate the mean, median, and mode based on the sample data for $X = Age$.

$X =$

22	25	31	31	18	21	19	22	19	20	28	36	36	25	26	23	24	32	27	18	22	29	45	34	21
22	19	22	24	39	24	27	23	20	18	22	23	26	18	24	25	53	38	23	24	20	23	27	18	19
17	24	21	19	20	25	38	27	25	41	19	47	18	25	25	41	30	22	18	21	21	18	35	19	30
39	34	19	17	54	18	42	19	23	27	22	37	35	37	32	40	22	25	34	33	17	24	26	19	33

22 25 31 31 18 21 19 22 19 20 28 36 36 25 26 23 24 32 27 18 22 29 45 34 21
22 19 22 24 39 24 27 23 20 18 22 23 26 18 24 25 53 38 23 24 20 23 27 18 19
17 24 21 19 20 25 38 27 25 41 19 47 18 25 25 41 30 22 18 21 21 18 35 19 30
39 34 19 17 54 18 42 19 23 27 22 37 35 37 32 40 22 25 34 33 17 24 26 19 33

mean = $\overline{X} = (22+25+31+31+18+\ldots+33)/100 = 26.49 = 26.5$

Using R: mean(X) = 26.5

median = $Q_2 = \frac{X_{50}+X_{51}}{2} = \frac{24+24}{2} = 24$ Using R: median(X) = 24.0

mode = most frequent = 19 (frequency 10)

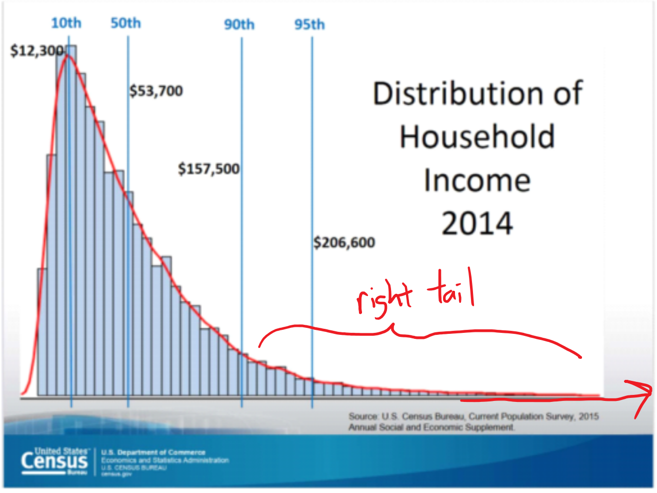
Using R: table(X)

Sensitivity to Extreme Values

If a data set contains *extreme* values, which measure of center best describes the *central value*?

Generally, the *median* is more representative of the center than the *mean* when there are extreme values since those values "pull up" or "pull down" the mean but don't affect the median.

Example Shown below is the income distribution for USA households in 2014. The *median* income is $53700 per year. The *mean* income will be *larger* than the median due to the long "right tail".



Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement

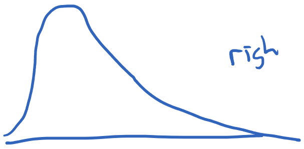
rel. freq

X (income)

P_{50} = median
= \$53,700
(middle household)

μ = mean
≙ \$100,000

mode ≙ \$13,000



Measures of Spread

There are four ways that we will measure the spread or variability of a data set:

"How different are the X valus from each other"?

1. Range
2. Standard Deviation ✮ important ✮
3. Variance
4. Inter-quartile Range IQR → boxplots

Range – the range is the *distance* between the maximum and minimum value.

Range = max(X) − min(X)

5 ~~to~~ 30

Range – the range is the *distance* between the maximum and minimum value.

Range = max(X) − min(X)

5 ~~to~~ 30

- In R use: `max(data$X) - min(data$X)`
- Note that range is a *single* number (because you subtract), not a pair of numbers.
- The range is simple to compute, but it is determined by the extreme values only.

Standard Deviation – the standard deviation of a data set is calculated using *all* of the data points. In R use the function `sd()` for sample standard deviation.

(sample standard deviation)

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sqrt{\frac{1}{n-1}\left[(X_1-\bar{X})^2+(X_2-\bar{X})^2+\ldots+(X_n-\bar{X})^2\right]}$$

Using R: sd(X)

(population standard deviation)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

"sigma"

σ uses μ instead of $\bar{X}$ and N instead of $n-1$

- Use the formula for s when you only know X for a random sample of the entire population
- Use the formula for σ when you know X for the entire population (which is rare).
- In either case, the value of s or σ is the *typical distance* between X and the mean value of X.

Example Calculate the sample standard deviation given the following sample values of X

$$18, 20, 21, 22, 24 \qquad n=5$$

find $\bar{X} = \frac{18+20+21+22+24}{5} = \frac{105}{5} = 21.0$

find $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} = \sqrt{\frac{1}{4}\left[(18-21)^2+(20-21)^2+(21-21)^2 + (22-21)^2+(24-21)^2\right]}$

$= \sqrt{\frac{1}{4}\left[9+1+0+1+9\right]} = \sqrt{5} = 2.236\ldots$

$= \boxed{2.2}$

Variance – the *variance* of a data set is directly related to standard deviation; specifically, variance is the square of the standard deviation. We therefore represent variance using the symbols s^2 and σ^2.

Sample Variance = $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (no square root)

pop. variance = $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$ (no square root)

Variance is used in deeper theory of probability | statistics

Inter-Quartile Range (IQR) – the distance between the 25th percentile and the 75th percentile of a data set. We will come back to this.

$IQR = P_{75} - P_{25}$

"range of middle half".

Percentiles and Quartiles

The k^{th}-percentile of a numerical variable X is denoted P_k. It divides the data into two parts: the lower k% of values and the upper $(100 - k)$% of values.

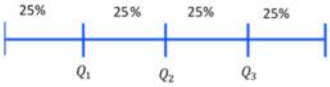
Example Look at the age data that was given above, but now sorted.

```
> sort(data$Age)
  [1] 17 17 17 18 18 18 18 18 18 18 18 18 19 19 19 19 19 19 19 19 19 19 20 20 20
 [26] 20 21 21 21 21 21 22 22 22 22 22 22 22 22 22 23 23 23 23 23 23 24 24 24 24
 [51] 24 24 24 25 25 25 25 25 25 25 25 26 26 26 27 27 27 27 27 28 29 30 30 31 31
 [76] 32 32 33 33 34 34 34 35 35 36 36 37 37 38 38 39 39 40 41 41 42 45 47 53 54
```

Determine $P_{25} =$ Value greater than 25% of sample $= \frac{20+20}{2} = 20$

$P_{53} =$ Value greater than 75% of sample $= \frac{31+32}{2} = 31.5$

Quartiles are percentiles in multiples of 25.



$Q_1 = P_{25}$

$Q_2 = P_{50}$ $Q_3 = P_{75}$

Example On the Math 12 Provincial Exam, Bill scored in the 75th percentile.

Interpretation: Not that Bill's score was 75%.

Bill's score was greater than 75% of students and less than 25% of students.

How to Calculate Percentiles "By Hand"

To calculate the k^{th} percentile of a set of n data values, proceed as follows:

1. Sort the data from smallest to greatest.
2. Calculate the location L of the theoretical position of P_k in the sorted list: $L = \left(\frac{k}{100}\right) \times n$
3. If L is a whole number, then
$$P_k = \frac{X_L + X_{L+1}}{2} \quad \text{(average of } X_L \text{ and } X_{L+1}\text{)}$$
4. If L is not a whole number, then round L *up* to the next whole number and use
$$P_k = X_L$$

Example Using the sample data for $X = Age$, calculate the following quantiles "by hand" and using R.

```
  [1] 17 17 17 18 18 18 18 18 18 18 18 18 19 19 19 19 19 19 19 19 19 19 20 20 20
 [26] 20 21 21 21 21 21 22 22 22 22 22 22 22 22 22 23 23 23 23 23 23 24 24 24 24
 [51] 24 24 24 25 25 25 25 25 25 25 25 26 26 26 27 27 27 27 27 28 29 30 30 31 31
 [76] 32 32 33 33 34 34 34 35 35 36 36 37 37 38 38 39 39 40 41 41 42 45 47 53 54
```

P_{89} $L = \frac{89}{100} \times 100 = 89$ $P_{89} = \frac{X_{89} + X_{90}}{2} = \frac{38+38}{2} = 38$

[using R: quantile(data$Age, 0.89) = 38]

$Q_1 = P_{25}$ $L = \frac{25}{100} \times 100 = 25$ $Q_1 = (X_{25} + X_{26})/2 = (20+20)/2 = 20$

[quantile(data$Age, 0.25) = 20]

$Q_1 = P_{25}$ $L = \frac{25}{100} \times 100 = 25$ $Q_1 = (X_{25} + X_{26})/2 = (20+20)/2 = 20$

[quantile(data$Age, 0.25) = 20]

"X-tilde" $\tilde{X} = Q_2 = \frac{X_{50}+X_{51}}{2} = 24$

$Q_3 = P_{75} = \frac{X_{75}+X_{76}}{2} = \frac{31+32}{2} = 31.5$ → 75% of data is < 31.5
25% of data is > 31.5

[quantile(data$Age, 0.75) = 31.25]

$IQR = Q_3 - Q_1$
$= P_{75} - P_{25} = 31.5 - 20 = \boxed{11.5}$

IQR(data$Age)
= 11.25

Note: R calculates quantiles using an interpolation method. The two answers will differ somewhat.

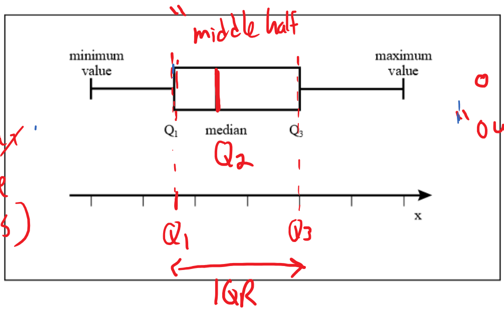
Box plots (Box-and-Whisker Plots)

If X is a numerical variable, then the following five statistics make up the *five-number summary*.

- Min, Max
- Q_1, Q_2, Q_3

A *boxplot* is a simple way to visualize the five-number summary.

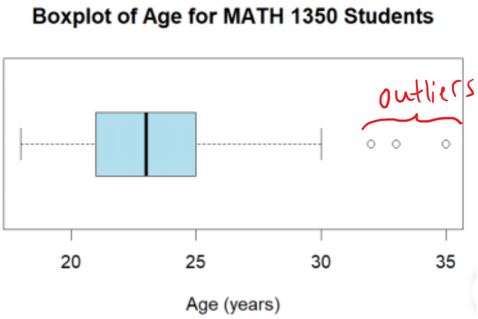
The whiskers go out to min and max' (unless there are outliers)



o o * +
"outliers"

- A boxplot shows the full *range* (distance between min and max values) and also the range of the middle 50% of the observations (the box), which is called the *inter-quartile range (IQR)*.
- If the right half of the box is wider, then *X is skewed right*.
- If the left half of the box is wider, then *X is skewed left*.

Example Using the data for students in MATH 1350 in September 2021, make a boxplot of Age. Identify any outliers based on the plot.



$X = 32, 33, 35$

Outliers

It is always a difficult question to decide what to do with *outliers* (extremely small or large values of X).

Should we delete outliers?

Yes, only if we have some evidence they arise from human error.

We can never be certain when an extreme value is due to a mistake or not. We need an objective rule to identify possible outliers so that the decision does not come down to subjective preference.

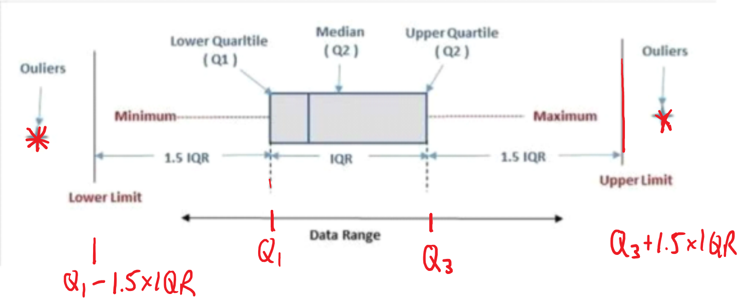
The rule we will use for identifying outliers is based on the Inter Quartile Range (IQR).

$$IQR = Q_3 - Q_1 = \text{width of box.}$$

$$\text{lower limit} = \underline{Q_1} - 1.5 \times IQR$$

$$\text{upper limit} = \underline{Q_3} + 1.5 \times IQR$$

Any values *below* the lower limit or *above* the upper limit are marked as *outliers*.



Example Using the *Age* data for students in MATH 1350 in 2021, calculate the lower and upper limit for outliers and identify any outliers.

$Q_1 = 21$

$Q_3 = 25$

$IQR = Q_3 - Q_1 = 4$

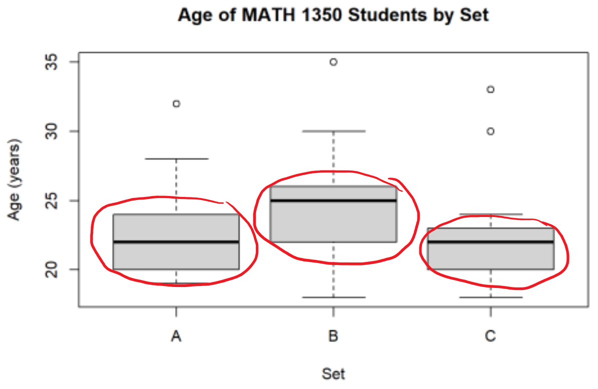
lower limit $= Q_1 - 1.5 \times IQR = 21 - 1.5 \times 4 = 15$

upper limit $= Q_3 + 1.5 \times IQR = 25 + 1.5 \times 4 = 31$

outliers : $X = 32, 33,$ and 35.

Another important use of boxplots is to *compare* a variable X for two different groups.

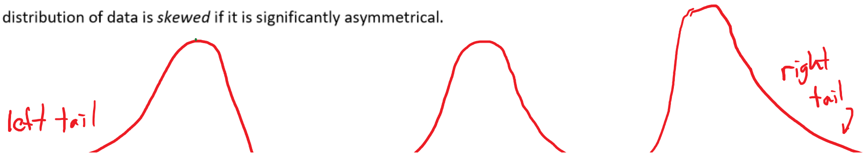
Example Make side-by-side boxplots for $X = Age$ using the data for sets A, B, and C for MATH 1350 students in 2021. Draw a conclusion based on what you see.

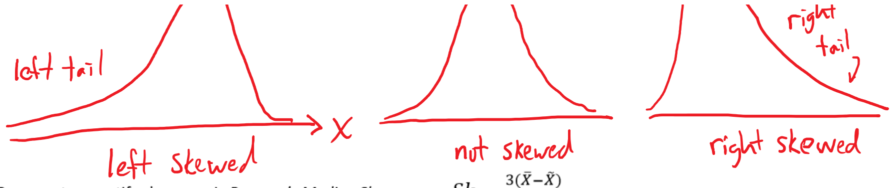


Set B tends to be older by one or two years.

Skewness

A distribution of data is *skewed* if it is significantly asymmetrical.





One way to quantify skewness is *Pearson's Median Skewness*: $Sk = \frac{3(\bar{X}-\tilde{X})}{s}$

- If $Sk < -1.0$, then the data is skewed left.
- If $Sk > +1.0$, then the data is skewed right.
- If $-1.0 < Sk < 1.0$, then the data is not significantly skewed.

$$Sk = \frac{3(\bar{X}-Q_2)}{S}$$

Example Calculate Sk for the variable $X = Age$ of MATH 1350 students. Is Age significantly skewed?

$\bar{X} = 23.37$

$Q_2 = 23$

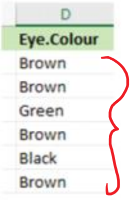
$S = 3,85$

$Sk = 0.29$

Sk positive by not > 1.0.
Age is not significantly skewed.

Presenting and Summarizing One Categorical Variable, X

First, let's consider one specific categorical variable, $X = Eye.Colour$.



D
Eye.Colour
Brown
Brown
Green
Brown
Black
Brown

raw data

The first thing we would likely do is create a *frequency table*, which simply *counts* how many times each value of X occurred in our sample.

To get the *relative frequencies*, divide each frequency by the total size of the sample.

Example Using the data for MATH 1350 students in 2021, we get the frequency table:

Eye Colour	Frequency	Relative Frequency
Brown	43	43/59 = 0.729
Black	10	10/59 = 0.169
Blue	3	3/59 = 0.051
Green	3	3/59 = 0.051

R command:
```
>table(data$Eye.Colour)
>prop.table(table(data$Eye.Colour))
```

Graphical Presentation of one Categorical Variable

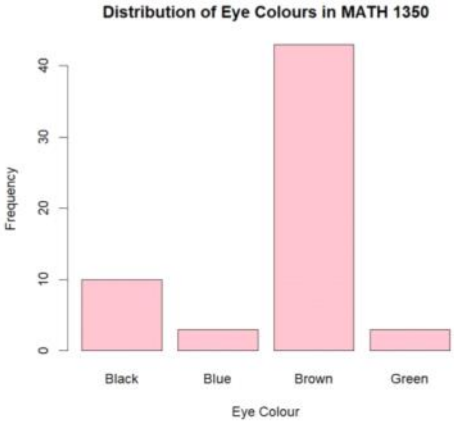
The best way to present one categorical variable is using a *bar chart*.

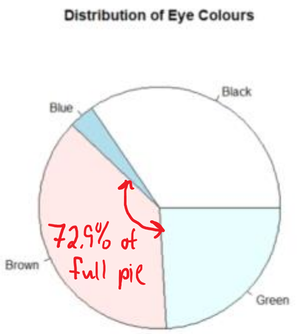
The possible values of X go along the horizontal axis.

The frequencies (counts) go along the vertical axis.

R command:

```
>barplot(data$Eye.Colour,
        xlab="Eye Colour",
       ylab="Frequency",
        main="Distribution…",
       col="pink")
```





An alternative to a bar chart is a *pie* chart, although these can make it harder to compare different categories.

R command:

```
> pie(table(data$Eye.Colour),
main="Distribution of Eye Colours")
```

- Only see relative frequency.
- Hard to compare categories.

Numerical Summaries for Categorical Variable

The most important way to summarize categorical variables is to calculate a *proportion*.

A *proportion* is the fraction of individuals who fall into some specified category. We distinguish between *sample proportion* and *population proportion*.

☆ (sample) $\hat{p} = \frac{x}{n} = \frac{\text{number of units in a certain category}}{\text{Sample size}}.$

p-hat

☆ (population) $p = \frac{x}{N} = \frac{\text{num in Category}}{\text{pop. size}}$

Example Calculate the sample proportion for students with brown eyes in MATH 1350 in 2021.

~~R commands~~:

$$\hat{p} = \frac{x}{n} = \frac{46}{59} = 0.729$$

Note that $\hat{p}$ changes depending on the same.
But p is not random.

Presenting and Summarizing *two* Categorical Variables: Two-Way Tables

"Contingency"

When we are dealing with *two* categorical variables, the main way to present the data is using a *two-way* table (also called a *contingency table*).

Example A two-way table for `Eye.Colour` and `Wears.Glasses` would look like:

		Wears.Glasses			
		Yes	No	Sometimes	Totals
Eye.Colour	Black	4	4	2	10
	Blue	1	2	0	3
	Brown	24	16	3	43
	Green	0	3	0	3
	Totals	29	25	5	Grand Total = 59

R command:

```
    > table(data[c("Eye.Colour","Wears.Glasses")])
```

rows columns

We can also calculate various proportions based on a two-way table.

From R:

```
           Wears.Glasses
Eye.Colour No Sometimes Yes
     Black  4         2   4
     Blue   2         0   1
     Brown 16         3  24
     Green  3         0   0
```

Example Calculate the following:

i. What proportion of students have Black eyes?

$\hat{p}$ (Black eyes) $= \frac{10}{59}$

ii. What proportion of students said they "Sometimes" wear glasses?

$\hat{p}$ (Sometimes) = 5/

p (Black eyes) = $\frac{}{59}$

```
Green   3           0   0
```

ii. What proportion of students said they "Sometimes" wear glasses?

$\hat{p}$ (Sometimes) = $5/59$

iii. What proportion of students have Brown eyes and said "Yes" they wear glasses?

$\hat{p}$ (Brown and Yes) = $\frac{24}{59}$

Conditional Probability

iv. Of the students who said "Yes" they wear glasses, what proportion have Brown eyes?

$\hat{p}$ (Brown given "Yes") = $\frac{24}{29}$ = 0.828

v. Of the students who have Brown eyes, what proportion said "Yes" they wear glasses?

$\hat{p}$ ("Yes" given Brown) = $\frac{24}{43}$ = 0.558