Summary Statistics (numeric data)

Statistics is the name of an academic discipline, but also a numerical value that summarizes a dataset in some way. The first statistics we will discuss are summarizing the center and spread.

Measures of Center

1) The sample mean, average, or expected value.

Let our sample be made of data points $y_1, y_2, \dots, y_n$ (n total data). The sample mean is denoted $\bar{y}$, and is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = y_1/n + y_2/n + \cdots + y_n/n$$

$$= (\text{sum of all data})/n.$$

Notice the mean uses $1/n^{Th}$ of all data to create a "typical value", which means every piece of data is weighted equally.

2) Median / Percentiles / Quartiles

The k^{th} percentile of a dataset is the value for which k% of the data lies below it, and (100-k)% of the data lies above it (for k between 0 to 100).

To calculate the k^{th} percentile (denoted $y^{(k)}$):

1) Order the data from smallest to largest

2) Calculate $(\frac{k}{100})(n+1)$ [n = sample size]

2) Use "the rounding rule": If 2) resulted in a whole number, the k^{th} percentile is the ordered number in the $[(k/100)(n+1)]^{Th}$ location.

If 2) resulted in a decimal, round $(\frac{k}{100})(n+1)$ up and down, and average the ordered numbers in those two locations.

The median is denoted $\tilde{y}$, and is the 50^{th} percentile. It is the literal center of the data.

Notice the median uses locations, and is calculated

based off of a maximum of two values (unlike $\bar{y}$).

Quartiles split the data into quarters, and
$Q_1 = 25^{th}$ percentile = first quartile
$Q_2 = 50^{th}$ percentile = second quartile = median
$Q_3 = 75^{th}$ percentile = third quartile

Ex: The survival time of patients with severe chronic heart disease (in months) was:
5, 7, 10, 15, 16, 17, 19, 20, 29, 32, 35 $(n = 11)$

a) Calculate the mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{11}(5+7+10+15+\cdots+35) = \frac{1}{11}(205) = 18.636$$

b) Calculate the median

$\tilde{y}$ is the 50^{th} percentile, so
1) Data is already ordered (yay)
2) $50/100(n+1) = 0.50(12) = 6^{th}$
3) 6 is a whole #, so $\tilde{y} = 6^{th}$ value $= 17$

c) Calculate Q_1, Q_2, Q_3, and the 90^{th} percentile
$Q_1 = 25^{th}$ percentile, $0.25(n+1) = 0.25(12) = 3^{rd}$ value
so $Q_1 = 10$
$Q_2 = \tilde{y} = 17$.
$Q_3 = 75^{th}$ percentile, $0.75(n+1) = 0.75(12) = 9^{th}$ value,
so $Q_3 = 29$
For $y^{(90)}$, $0.90(12) = 10.8^{th}$ value, so average 10^{th} and 11^{th}:
$y^{(90)} = (32+35)/2 = 33.5$

* Outliers

Outliers are unusually small or large observations. There are multiple definitions, but a common one uses a "boxplots" definition.

Boxplot outlier: Any observation that is

i) larger than $Q_3 + 1.5(Q_3 - Q_1)$ (upper cutoff)
ii) Smaller than $Q_1 - 1.5(Q_3 - Q_1)$ (lower cutoff)

Ex: From our previous example,
lower cutoff $= Q_1 - 1.5(Q_3 - Q_1) = 10 - 1.5(29-10) = -18.5$
upper cutoff $= Q_3 + 1.5(Q_3 - Q_1) = 29 + 1.5(29-10) = 57.5$

So there are no outliers.

✳ Measures of Spread (numeric data)
We may also be interested in the spread of the data - either overall or from its center.

1) The range of a data set: This is simply the maximum difference in values of the dataset.
I.e: range $= \max\{y_i\} - \min\{y_i\}$
= (maximum of dataset) - (minimum of dataset)

2) Sample Variance
The variance is the "typical squared deviation from the mean", and denoted s^2.
A deviation from the mean is: $(y_i - \bar{y})$, so a squared deviation is $(y_i - \bar{y})^2$. The total squared deviations are $\sum_{i=1}^{n}(y_i - \bar{y})^2$, and the variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]$$

Fact: We don't use simply $\frac{1}{n-1}\Sigma(y_i - \bar{y})$ because $\Sigma(y_i - \bar{y}) = 0$ (this is also why $\bar{y}$ is considered the "center" of the data).

Fact: The denominator is $(n-1)$ because then s^2 more closely estimates the overall population variance.

Most people do not interpret the variance, since its units are (units of y)^2

3) Sample Standard Deviation, denoted s

$s = \sqrt{s^2}$ (deviation)

Which is interpreted as "The typical distance of a data point to its mean"

Ex: Continuing our example:

d) Calculate s and interpret it.

$\Sigma(y_i^2) = 5^2 + 7^2 + 10^2 + \cdots + 35^2 = 4795$

$s^2 = \frac{1}{11-1}(4795 - (11)[18.636)^2) = 97.469$

So

$s = \sqrt{97.469} = 9.873$

A typical deviation of survival time in months from the mean is 9.873

Graphs for numerical data

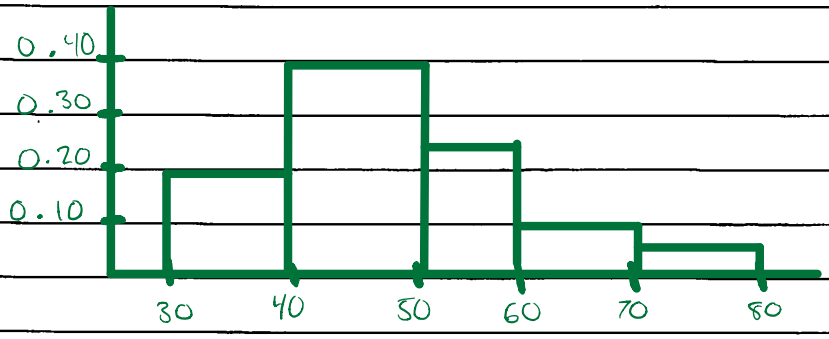
I will only go over the most common graphs, which are the ones you will learn to do in R.

Definition: The frequency of a datapoint is the number of times the data point occured.

Definition: The relative frequency is the frequency divided by n (always between 0 and 1).

1) **Histograms** - Histograms take numerical data and selects groups of bins (ranges, typically equi-width). These give the widths of the boxes. For each interval, we also calculate the frequency / relative frequency of data points in the interval. This is the height of the boxes.

Ex: Weights of border collies (a type of dog) in lbs.



From here we can tell the interval that is most/least common, and the shape of the data (the distribution).

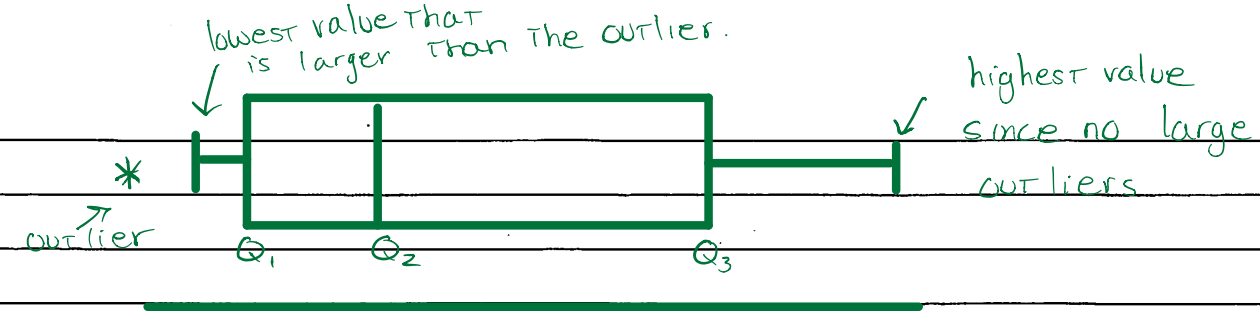
Definition: The Five Number Summary

This is simply five numbers that tell use a lot about a data set: Min, Q_1, Q_2, Q_3, Max

They are also found in our next graph:

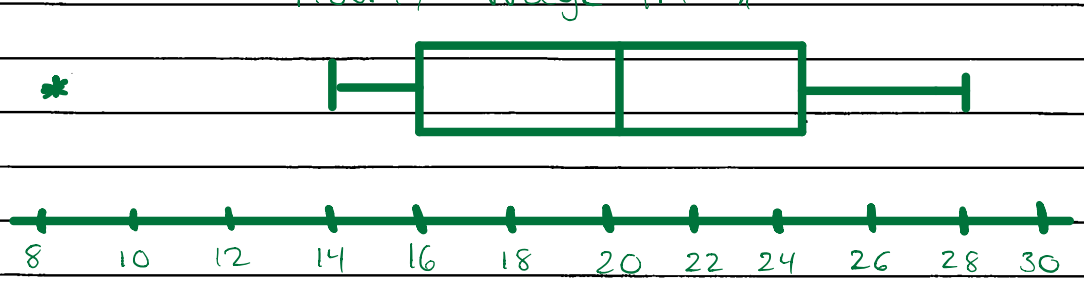
2) **Boxplot** - We draw a box around lines at Q_1, Q_3, and put another line at Q_2. Then, we identify any outliers, and draw "whiskers" from the box to the highest and lowest value that are not outliers. Lastly, we mark outliers with *'s or o's.

Ex:



Boxplots can also be vertical instead.

Ex:



a) What value has 25% of the data above it? Q_3, or ~24

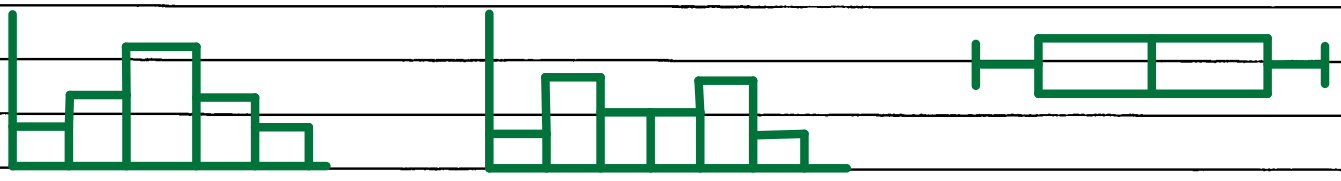
b) What is the range of the data? (max-min) = 28-8 = 20

c) What values contain the middle 50% of the data (with 25% on either side)? (16, 24)

* Distributions of Data (Shapes)

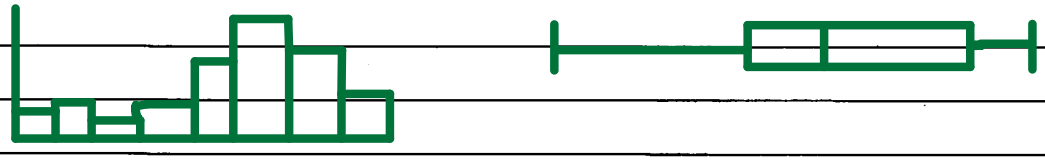
There are three main shapes of data.

1) Symmetric data: One half is a mirror image of the other.



Examples of ≈ symmetric distributions: IQ, weight, height

2) Left Skewed: Most of the data is to the right, with a long left "tail"



3) Right Skewed: Most of the data is to the left, with a long right "tail"

Categorical Data

For summarizing categorical data, we really only have freq. and relative freq.

Ex:

Grade:	A	B	C	D
Freq:	15	20	10	5
Rel. Freq:	0.30 $= 15/50$	.40 $20/50$	.20 $10/50$	.10 $5/50$

If we had a second categorical variable, we could also calculate relative freq. by group

Ex:

	A	B	C	D
(24) Male	8	10	2	4
(26) Female	7	10	8	1

Relative Freq. Overall

	A	B	C	D
M	$8/50 = .16$	$10/50 = .20$	$2/50 = .04$	$4/50 = .08$
F	.14	.20	.16	.02

Relative Freq for M only:

A	B	C	D
$8/24 \approx .33$	$10/24 \approx .42$	$2/24 = 0.08$	$4/24 \approx 0.17$

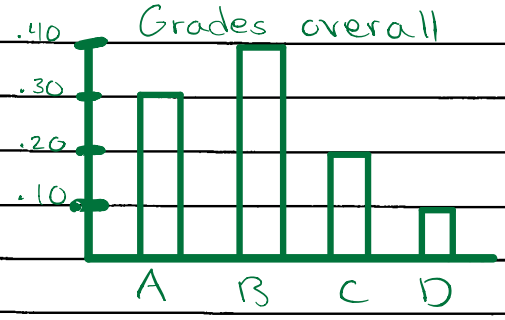
Relative Freq for F only

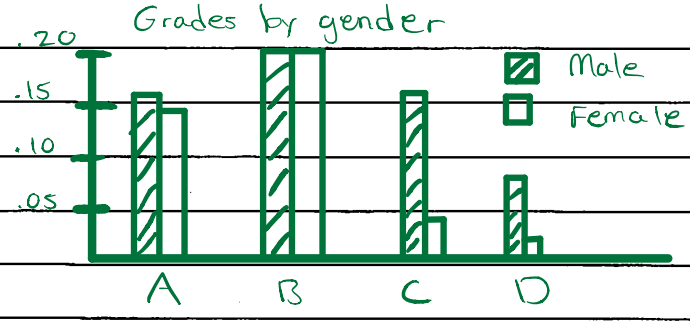
A	B	C	D
$7/26 = 0.27$	$10/26 = .38$	$8/26 = .31$	$1/26 = .04$

Graphs for categorical data

I will go over two main graphs, that can be used for one or two categorical variables.

1) Bar Plots - Essentially histograms for categorical data. The main difference is the bars are separated





Notice that: i) The y-axis changed between the two graphs
ii) The second plot uses the "overall" relative freq, not the "by group" ones.

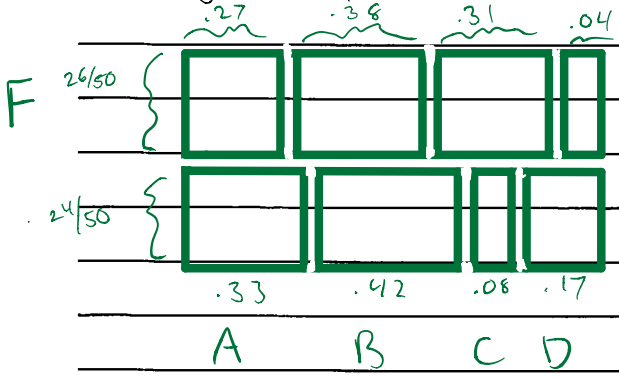
It is relatively easy to quickly compare grades by gender, or the different categories.

2) Mosaic Plots: A plot of boxes, where the height and width of boxes represent relative freq by group.

One variable:

The C box is 1/2 of the B, and the D is 1/4 of B etc.

Two variables: Two rows of the above, but the height of each row represents the proportion of either grades / gender, and boxes are based on rel freq by group (M/F)



It could also be flipped, so the "by group" is by grade.

You do not have to draw any plots by hand, but may have to interpret them.

Now we move on to starting probability Theory. First, we should distinguish between a population value, or a sample value.

Definition: A parameter is the value of a statistic based on population data. Almost always not measurable.

Definition: A sample statistic is the value of a statistic based on sample data.

We use sample statistics to estimate population parameters. We also use different notation for each:

	Parameter	Sample Statistic
Mean	μ	$\bar{y}$
Std. Dev	σ	s
Varrance	σ^2	s^2
Proportion	p	$\hat{p}$

* Probability Theory

The probability of an event is a number between 0 and 1 which represents how likely the event is to occur.
Notation: $Pr\{A\}$ = The prob. of an event A.

Three basic prob. facts:
i) $0 \leq Pr\{A\} \leq 1$ for any event A
ii) If $Pr\{A\} = 0$, the event A does not occur
iii) If $Pr\{A\} = 1$, the event A always occurs

There are two main ways to calculate probabilities

1) Logic Based Calculations.
This is where we assume all possible events are either equally likely, or their probabilities are known, based on a logical or natural assumption. We also assume

we know all possible outcomes in advance

Ex: A coin. A coin has two sides, and we assume both sides are equally likely when we flip a coin. Thus, $Pr\{Head\} = Pr\{Tail\} = 1/2$

In general, a probability is calculate by

$$Pr\{A\} = \frac{\#\text{ of events with A}}{\#\text{ of all possible events}}$$

using the logic-based definition. This requires a lot of assumptions

2) Relative Frequency Calculations

These are usually based on samples, and

$$\hat{Pr}\{A\} = \frac{(\#\text{ of times A occured in the sample})}{n}$$

Notice I used $\hat{Pr}\{A\}$. This is because $\hat{Pr}\{A\}$ estimates $Pr\{A\}$ (which would be the "True" probability). We can show that as you sample more people, or as $n \to \infty$, $\hat{Pr}\{A\} \longrightarrow Pr\{A\}$.

✳ Probability Rules

There are several fundamental rules of probability that we will use throughout the course:

1) $0 \leq Pr\{A\} \leq 1$
2) If we have n possible outcomes of an experimen, $A_1, A_2 \ldots, A_n$, then $\sum_{i=1}^{n} Pr\{A_i\} = 1$ (all probabilities add to 1)
3) The probability that A does not occur is denoted $Pr\{A^c\}$, where $Pr\{A^c\} = 1 - Pr\{A\}$ ⓐ

Definition: The union of two events A and B is denoted $\{A \text{ or } B\}$, and is the event where either A occurs, B occurs, or both occur. "One or the other or both".

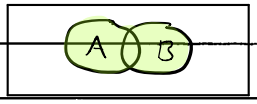
Definition: The intersection of two events A and B is denoted $\{A \text{ and } B\}$, and is the event where both

A and B occur at the same time. Note: $Pr\{A \text{ and } B\} = Pr\{B \text{ and } A\}$

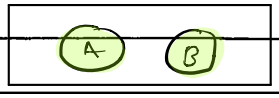
Definition: A and B are disjoint if $Pr\{A \text{ and } B\} = 0$, i.e they cannot occur at the same time.

Rules continued...

4) For any two events A and B,
$Pr\{A \text{ or } B\} = Pr\{A\} + Pr\{B\} - Pr\{A \text{ and } B\}$



5) For disjoint events A and B,
$Pr\{A \text{ or } B\} = Pr\{A\} + Pr\{B\}$



Ex: Plants were given three treatments, and their survival after one year was recorded.

	A	B	C	
Survived	13	18	10	41
Died	17	12	30	59
	30	30	40	100

Let S = plant survival. Find the following:

a) The probability a plant survives.
$\hat{Pr}\{S\} = (\#\text{ of plants who S})/n = 41/100 = 0.41$

b) The probability a plant is not in treatment A.
$\hat{Pr}\{A^c\} = 1 - Pr\{A\} = 1 - {}^{30}/_{100} = 70/100 = 0.70$

c) The probability they did not survive, or they had treatment B, or both.
$\hat{Pr}\{S^c \text{ or } B\} = \hat{Pr}\{S^c\} + \hat{Pr}\{B\} - \hat{Pr}\{S^c \text{ and } B\}$
$= {}^{59}/_{100} + {}^{30}/_{100} - {}^{12}/_{100} = {}^{77}/_{100} = 0.77$

We may also be interested in the prob. of survival by group. Or, "out of the A plants, what was the prob. of survival?" This is a conditional probability because it subsets the data into only one group.

Definition: The conditional probability of A given B is the prob that A occured, out of only the events where B occured. It is denoted $\{A \mid B\}$.

Rules Continued...

6) $Pr\{A \mid B\} = Pr\{A \text{ and } B\} / Pr\{B\}$ or
$Pr\{B \mid A\} = Pr\{B \text{ and } A\} / Pr\{A\}$

Which gives (using algebra)

7) $Pr\{A \text{ and } B\} = Pr\{A \mid B\} Pr\{B\}$ or
$Pr\{A \text{ and } B\} = Pr\{B \mid A\} Pr\{A\}$

Ex Cont.

d) Out of only the A plants, the prob of survivial.
$\hat{Pr}\{A \text{ and } S\} = 13/100$, $\hat{Pr}\{A\} = 30/100$, so that
$\hat{Pr}\{S \mid A\} = \hat{Pr}\{S \text{ and } A\} / \hat{Pr}\{A\} = (13/100)/(30/100) = 13/30 = 0.433$

e) If a plant survived, what is the prob. it was an A?
$\hat{Pr}\{A \mid S\} = \hat{Pr}\{A \text{ and } S\} / \hat{Pr}\{S\} = (13/100)/(41/100) = 13/41 = 0.317$

Notice that $\hat{Pr}\{A \mid S\} \neq \hat{Pr}\{S \mid A\}$!!