ITCS 121 Statistics

Lecture 2

Describing Data With Numerical Measures

Parameters

- In the last class we called a collection of measurements we are interested in a *population*.
- Measurements can be numbers (quantitative variables), or not (qualitative variables).
- We saw how we can visualize the data from a population in different ways (dot plots, pie charts, histograms etc.)
- For quantitative data we can also try to summarize it using numbers.
- A **parameter** is number calculated from a population.
- Parameters are supposed to tell us something important about the population.

Some important parameters

- Measures of centre (what is the ‘average’ value?):
  - Mean
  - Median
  - Mode
- Measures of variability (how ‘spread out’ is the data?):
  - Range
  - Variance
  - Standard deviation

Mean

- The **mean** is the most commonly used measure of centre.
- When people say ‘average’ they are usually talking about the mean.
- Very easy to calculate (in theory): add up every value in the population, and divide by the size of the population.

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

- We use the Greek letter mu, μ, to represent the mean of a population.
- Here the population is $\{x_1, x_2, \ldots, x_n\}$.

Example: mean

- Population is 2, 9, 11, 5, 6.

- Population mean is $\frac{2+9+11+5+6}{5} = \frac{33}{5} = 6.6.$

- So $\mu = 6.6.$

Median

- If Jeff Bezos (Founder of the multi-national technology company Amazon) walked into this room, the average person would become a billionaire.
- Does this reflect how rich we are, or just how rich Jeff Bezos is?
- The problem with the mean is that if extreme values are much bigger or smaller than the others, the value of the mean does not give us good information about the typical values.
- To avoid this problem, statisticians often use the **median.**

Calculating the median

- To calculate the median, we follow these steps:
  1. Order the population from smallest to largest.
  2. If n is the size of the population, calculate $\frac{n+1}{2}$.
  3. If $\frac{n+1}{2}$ is a whole number, the median is the value at position $\frac{n+1}{2}$ (remember we ordered the values).
  4. If $\frac{n+1}{2}$ is *not* a whole number, the median is the value that lies half way between position $\left\lfloor\frac{n+1}{2}\right\rfloor$ and $\left\lceil\frac{n+1}{2}\right\rceil$.

Example: median

- The population is 2, 4, 9, 8, 6, 5, 3.
  - Reorder: 2, 3, 4, 5, 6, 8, 9.
  - $\frac{n+1}{2} = \frac{8}{2} = 4.$
  - Median is in fourth place. So median is 5.
- The population is 2, 4, 9, 8, 6, 5.
  - Reorder: 2, 4, 5, 6, 8, 9.
  - $\frac{n+1}{2} = \frac{7}{2} = 3.5.$
  - Median is in position 3.5. This is half way between position 3 and position 4.
  - So median is $\frac{5+6}{2} = 5.5.$

Mode

- The **mode** is just the value the occurs most often in the data.
- If there is one value that occurs more than any other then the data is **unimodal**.
- If there are two values that occur more than the others the data is **bimodal**.
- There could be more than two values that share the top position. E.g. we could have trimodal data.
- If there are ‘too many’ modes, e.g. if every value occurs just once, we would say there is no mode.
- There’s no fixed rule for how many modes is ‘too many’ though.

Example: mode

- Population 2, 4, 9, 8, 8, 5, 3.
  - Mode is 8 (unimodal).
- Population 2,2, 9, 8, 8, 5, 3.
  - Modes are 2 and 8 (bimodal).
- Population 2, 4, 9, 8, 5, 3.
  - No mode.
- Population 2, 2, 4, 4, 8, 8.
  - Should it be three modes or no modes?
  - A matter or opinion.

Class activity 1

In a psychological experiment, the time taken to complete a task was recorded for 10 subjects. These measurements are in seconds:

175	190	250	230	240	200	185	190	225	265

a) Find the mean time taken.

b) Find the median time taken.

c) Does this population have a mode?

Range

- We move on to measures of variability.
- The **range** is the simplest of these measures.
- The range is just the difference between the largest value and the smallest value in the population.
- $range = max - min.$
- Example:
  - Population is 2, 4, 9, 8, 8, 5, 3.
  - $range = 9 - 2 = 7.$

Variance

- Variance is a measure of the average difference between the values in the population and the mean of the population.
- We use the formula

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

- We use the Greek letter sigma, σ, and we use σ^2 for the variance for a reason that will become clear on the next slide.
- Why not use $\frac{\sum_{i=1}^{n}(x_i - \mu)}{n}$? Because this is always zero!
  - To see why this is true think about the definition of μ.

Standard deviation

- Standard deviation is just the square root of the variance.

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

Example: Variance and standard deviation

Population is 5, 12, 6, 8, 14.

$$\mu = \frac{5 + 12 + 6 + 8 + 14}{5} = 9$$

$$\sigma^2 = \frac{(5-9)^2+(12-9)^2+(6-9)^2+(8-9)^2+(14-9)^2}{5} = \frac{16+9+9+1+25}{5} = \frac{60}{5} = 12$$

$$\sigma = \sqrt{12} = 3.46$$

Another way to calculate the variance

- We can also calculate the population variance using:

$$\sigma^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}$$

- This can be easier to calculate, but it always gives the same answer as the original formula.
- It’s not too hard to check that this formula is equivalent to the original one, but we won’t do that here.

Example: alternative variance formula

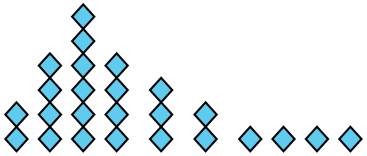
Population is 5, 12, 6, 8, 14.

x	x^2
5	25
12	144
6	36
8	64
14	196
$\Sigma = 45$	$\Sigma = 465$

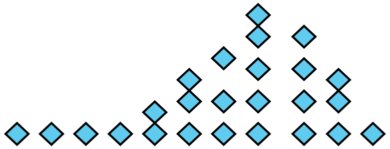
$$\sigma^2 = \frac{465 - \frac{45^2}{5}}{5} = 12$$

Mean, median, skew

- In the last class, we talked about skewness using our intuition:



Skewed right



Skewed left

- We can formalize this intuitive picture.
- Let μ be the population mean, let σ be the population standard deviation, and let υ (nu) be the median.
- Define the **skew** to be $\frac{(\mu-\upsilon)}{\sigma}$.
- If this is positive we have right skew, if it's negative we have left skew. I.e:
- mean > median = right skew,
- mean < median = left skew.

Warning! Skew

- The way we defined skew on the previous slide is slightly old fashioned.
- There are other definitions for skew that do not depend on the relationship between median and mean.
- We won't discuss these here, but you should know that they exist.

Samples

- In the real world, we usually cannot take measurements for every object we are interested in.
  - E.g. could we weigh every person in Thailand?
- To get information about populations where we do not have all the data, we use samples.
- A **sample** is a subset of a population
- We will talk more about how to take samples later in the course.
- We can use samples to estimate the values of population parameters (e.g. mean, standard deviation).
- Numerical estimates of population parameters calculated from samples are called **statistics**.

Sample means

- If we want to use a sample $\{x_1, \ldots, x_n\}$ to estimate the mean of a population, we use the formula

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- This is the same formula as the one for calculating the mean of a population.
- The only difference is we use $\bar{x}$ in place of μ.
- If we were using, say, $\{y_1, \ldots, y_n\}$ for the sample we would use $\bar{y}$ for the sample mean.
- In other words, the notation depends on the letter we are using for the values.

Sample variances

- If we want to use a sample $\{x_1, \dots, x_n\}$ to estimate the variance of a population, we use the formulas

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \text{or} \qquad s_x^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

- Again the notation depends on the choice of variable symbol.
- Notice that we divide by $n-1$, and not n, like we do in the formulas for population variance.
- This is because the variance of a sample usually underestimates the population variance.
- We correct for this by dividing by $n-1$.
- The technical reason is that by using sample data to estimate the population mean in the formula we lose one ‘degree of freedom’.
- We don’t need to understand exactly what this means.

Example: sample standard deviation

Sample is 5, 12, 6, 8, 14.

x	x^2
5	25
12	144
6	36
8	64
14	196
$\Sigma = 45$	$\Sigma = 465$

$$s_x^2 = \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s_x = \sqrt{15} = 3.87$$

Class activity 2

You are given n = 8 measurements:

3	1	5	6	4	4	3	5

a) Calculate the range.

b) Calculate the sample mean.

c) Calculate the sample variance and standard deviation.

d) Compare the range and the standard deviation. The range is approximately how many standard deviations?

Chebychev’s theorem

In any set of numbers, the proportion of values that lie within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$.

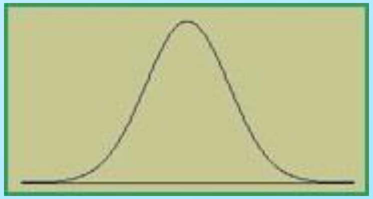
- In other words, suppose we have a set of numbers, and suppose this set has mean μ and standard deviation σ.
- Then a proportion of at least $1 - \frac{1}{k^2}$ of these numbers lie in the interval $[\mu - k\sigma, \mu + k\sigma]$.
- E.g. using $k = 2$, at least $\frac{3}{4}$ of the values lie within 2 standard deviations of the mean.
- Note that Chebychev is a Russian name, so there are some alternative English spellings.

The empirical rule

- Chebychev's theorem can be useful, but it is quite weak.
- For an important kind of data distribution we have something better:

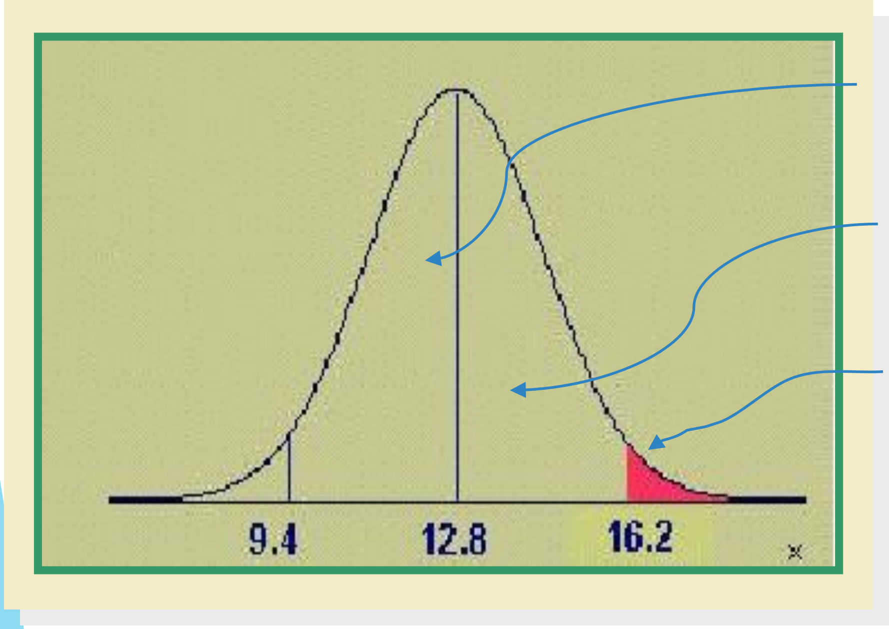
Given a distribution of values that is approximately mound-shaped:

- Approximately 68% of the values occur within 1 standard deviation of the mean.
- Approximately 95% of the values occur within 2 standard deviations of the mean.
- Approximately 99.7% of the values occur within 3 standard deviations of the mean.



Example: empirical rule

- The length of time for a worker to complete a task averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?
- Because the distribution is mound shaped we can use the empirical rule.



95% between 9.4 and 16.2

47.5% between 12.8 and 16.2

(50-47.5)% = 2.5% above 16.2

Example: Faculty

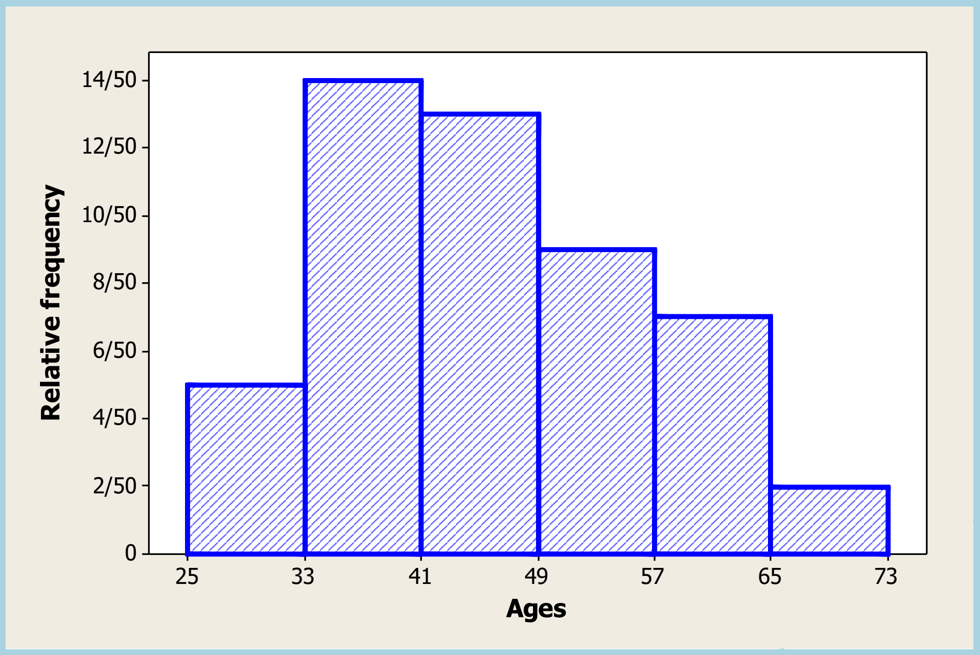
- The ages of 50 faculty members at a university.

34 48 **70** 63 52 52 35 50 37 43 53 43 52 44

42 31 36 48 43 **26** 58 62 49 34 48 53 39 45

34 59 34 66 40 59 36 41 35 36 62 34 38 28

43 50 30 43 32 44 58 53



Example: Faculty

- The population mean $\mu = 44.9$.
- The population standard deviation $\sigma = 10.62$.

k	$\mu + k\sigma$	Interval	Proportion in Interval	Chebychev	Empirical Rule
1	44.9 ± 10.62	34.28 to 55.52	31/50 (.62)	At least 0	≈ .68
2	44.9 ± 21.24	23.66 to 66.14	49/50 (.98)	At least .75	≈ .95
3	44.9 ± 31.86	13.04 to 76.76	50/50 (1.00)	At least .89	≈ .997

- We notice that this data agrees with Chebychev’s theorem (which it must do!).
- Not great agreement with empirical rule, but this is because distribution is not perfectly mound shaped.

Class activity 3

A distribution of measurements is relatively mound-shaped with mean 50 and standard deviation 10.

a) Approximately what proportion of the measurements will fall between 40 and 60?

b) Approximately what proportion of the measurements will fall between 30 and 70?

c) Approximately what proportion of the measurements will fall between 30 and 60?

d) If a measurement is chosen at random from this distribution, what is the probability that it will be greater than 60?

z-scores

- Suppose a value x from a population is 10.5 bigger than the population mean (i.e. $x = \mu + 10.5$).
- Is this an unusual event or not?
- It depends on the standard deviation of the population, σ.
- To talk about how extreme values from a population are, we use a normalized measure called the **z-score**.

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

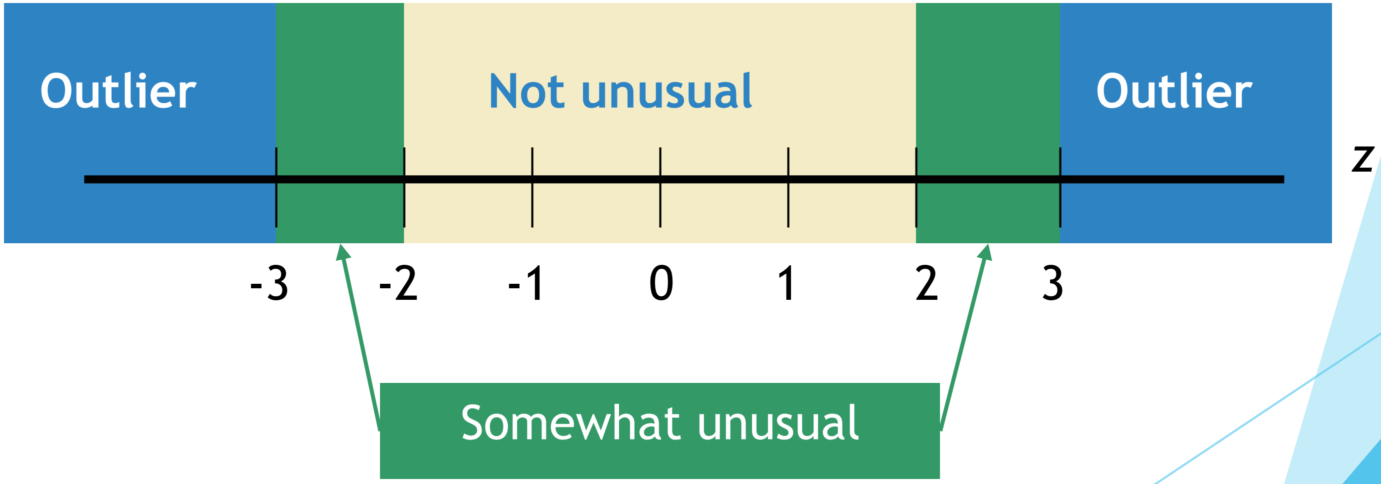
- The z-score tells us how many standard deviations x is from the population mean.

Which values are extreme?

- Chebychev’s theorem says a proportion of at least $1 - \frac{1}{k^2}$ values in a population have z-score in the range $[-k, k]$.
- The empirical rule says that, for mound shaped distributions, we have, approximately:
  - 68% of values with z-score in range $[-1,1]$
  - 95% of values with z-score in range $[-2,2]$
  - 99.7% of values with z-score in range $[-3,3]$

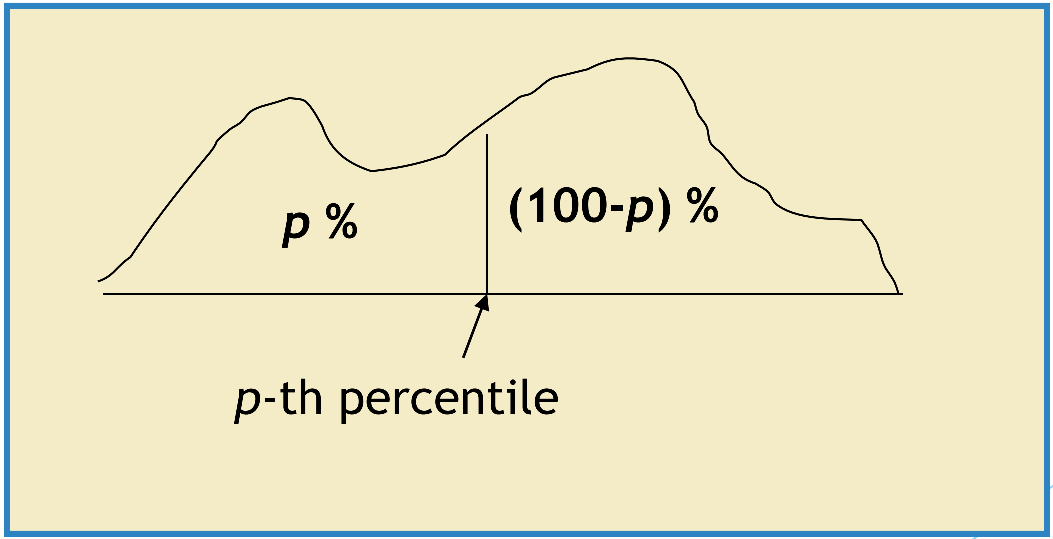
Outliers

- As a rule of thumb, we sometimes call a value an outlier if its z-score is less than -3 or more than 3.



Percentiles

- When we want to express where a value lies in a population, we often use its **percentile**.
- This is what percentage of values lie below the value of interest.



Quartiles

- Some percentiles have special names
- The 25th percentile is called the **first quartile** (Q1).
- The 50th percentile is the **second quartile** (Q2), AKA the median.
- The 75th percentile is the **third quartile** (Q3).
- We define the **interquartile range** (IQR) by

$$IQR = Q3 - Q1$$

Calculating quartiles

- We calculate Q1 and Q3 in a similar way to how we calculate the median.
  1. First arrange the data in order of size.
  2. Q1 is at position $0.25(n + 1)$.
  3. Q2 is at position $0.5(n + 1)$.
  4. Q3 is at position $0.75(n + 1)$.
- Usually $0.25(n + 1)$ and $0.75(n + 1)$ are not whole numbers.
- When this happens the quartile lies between $\lfloor 0.25(n + 1) \rfloor$ and $\lceil 0.25(n + 1) \rceil$, for Q1, and between $\lfloor 0.75(n + 1) \rfloor$ and $\lceil 0.75(n + 1) \rceil$ for Q3.
- Exactly how this calculation is done is easiest to understand with an example.

Example: Quartiles

The prices ($) of 18 brands of walking shoes:

40 60 65 65 | 65 68 68 70 70
70 70 70 70 74 | 75 75 90 95

Position of Q1 = .25(18 + 1) = 4.75

Position of Q3 = .75(18 + 1) = 14.25

Q1 is three quarters of the way between 65 and 65.
Q3 is a quarter of the way between 74 and 75.

$$Q1 = 65 + 0.75(65 - 65) = 65$$

$$Q3 = 74 + 0.25(75 - 74) = 74.25$$

The five number summary

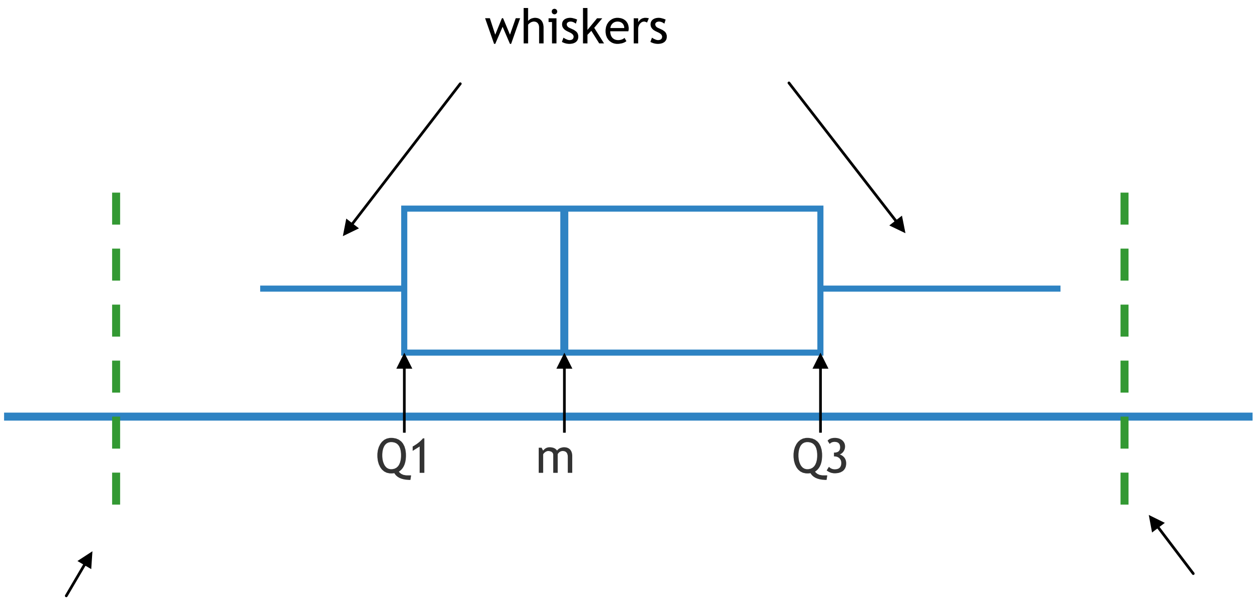
- We can summarize a population using the numbers:

min   Q1   median   Q3   max

- This is called the **five number summary.**

Box plots

- We often illustrate information about a population using a **box plot**, AKA **box and whiskers plot.**



$Lower\ fence = Q1 - 1.5 \times IQR$

$Upper\ fence = Q3 + 1.5 \times IQR$

Left whisker starts at smallest value inside fences.
Right whisker ends at largest value inside fences.
Values outside fences are considered outliers.

Example: Box plot

- Amount of sodium in eight brands of cheese:

  260 290 300 320 330 340 340 520

Q3 = 340

m = 325

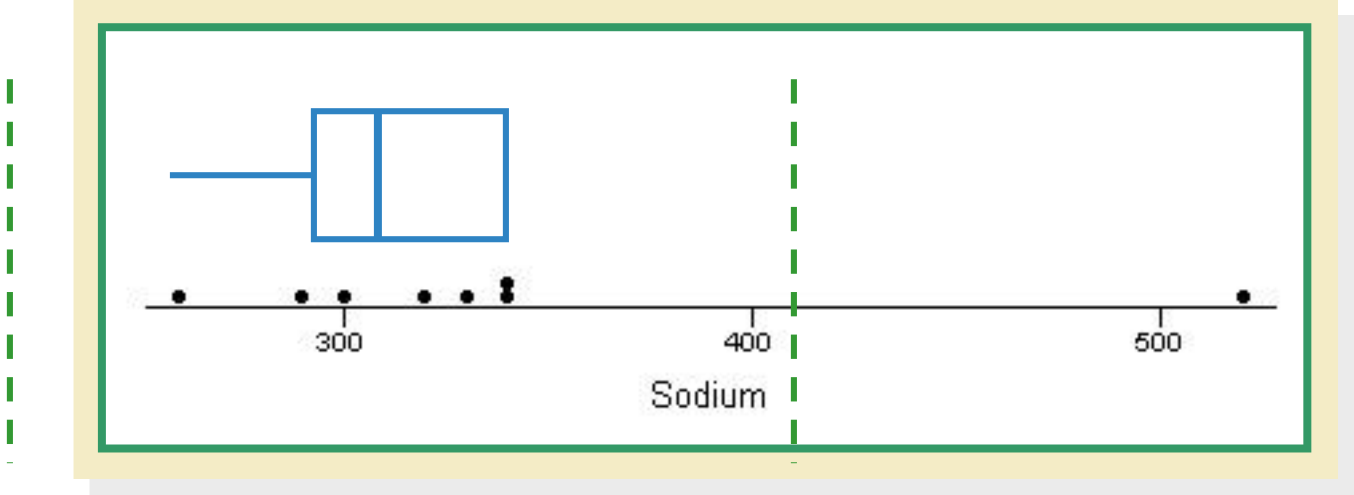
Q1 = 292.5

IQR = 340-292.5 = 47.5

Lower fence = 292.5-1.5(47.5) = 221.25

Upper fence = 340 + 1.5(47.5) = 411.25

Outlier: *x* = 520



Warning! Whiskers

- Some people don’t use fences, and draw whiskers using the min and max value of the data.
- It comes down to personal choice, but on this course we do it as we described in these slides (i.e. with fences).

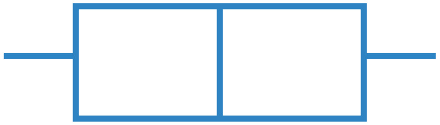
Outliers: Box plots vs z-score

- We've seen two ways to detect outliers.
  - Z-scores: $z > 3$ or $z < -3$.
  - Box plot: value above $Q3 + 1.5IQR$ of below $Q1 - 1.5IQR$.
- Not obvious that these are the same, and actually they are slightly different.
- Assuming a mound-shaped distribution, the area to the left of Q3 accounts for around 75% of the data.
- This corresponds to a z-score of roughly 0.675, assuming a distribution centered on Q2
  - We need to know about normal distributions for this, which we cover later.
- So $IQR = 2(0.675)$, and $Q3 + 1.5IQR$ corresponds to $4(0.675) = 2.7$.
- So the box plot outlier test using IQR roughly corresponds to counting $z > 2.7$ or $z < -2.7$.
- This is not exactly the same as using ± 3, but using ± 3 counts 0.3% of the data as outliers (by empirical rule), while using ± 2.7 counts 0.7% of the data as outliers, so quite similar.

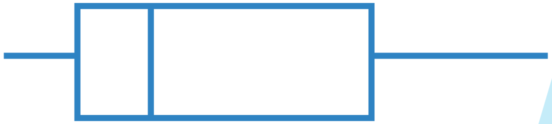
Box plots and skew

- Box plots can give us information about the skew of the data.
- It isn't always correct, but it can be a useful guide:

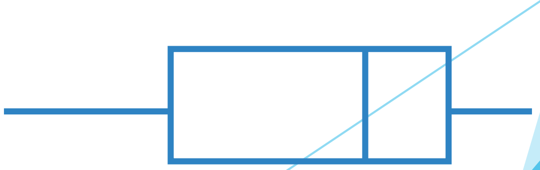
Median line in center of box and whiskers of equal length—symmetric distribution



Median line left of center and long right whisker—skewed right



Median line right of center and long left whisker—skewed left



Class activity 4

Construct a box plot for this data and identify any outliers.

3, 9, 10, 2, 6, 7, 5, 8, 6, 6, 4, 9, 22

Class activity 5

A population of long-stemmed roses has an approximately mound shaped distribution with a mean stem length of 15 inches and standard deviation of 2.5 inches.

a) Approximately what percentage of these roses have a stem length less than 12.5 inches?

b) Approximately what percentage of these roses have a stem length between 12.5 and 20 inches?