Lecture 2
Numerical Summaries

Population distributions
Summarising Numerical Data
- Measures of centre and spread

Summarising Categorical data
- Proportions and Percentages

In the Last Lecture….

We considered various ways of graphing data.

- **Bar charts** and **pie charts** are used to display categorical data.
- **Histograms** are used to display numerical data.
- **Box plots** are also used for displaying numerical data and are particularly useful if we want to compare two or more groups.
- We needed some numerical information, the **median** and **quartiles**, for constructing box plots.

In this lecture we shall continue with summarising data numerically.

From the Last Lecture - Displaying Data: Graphical Displays/Graphical Summaries

DATA	categorical	numerical	
categorical	clustered bar charts	comparative box plots	bar charts or pie charts
numerical	comparative box plots	scatter plots	histograms
	bar charts or pie charts	histograms	

Quiz 1

What visual display would be appropriate to:

a. compare the proportions of male and female students enrolled in STAT170?

b. identify the association between the major a student is enrolled in and the final grade obtained in STAT170.

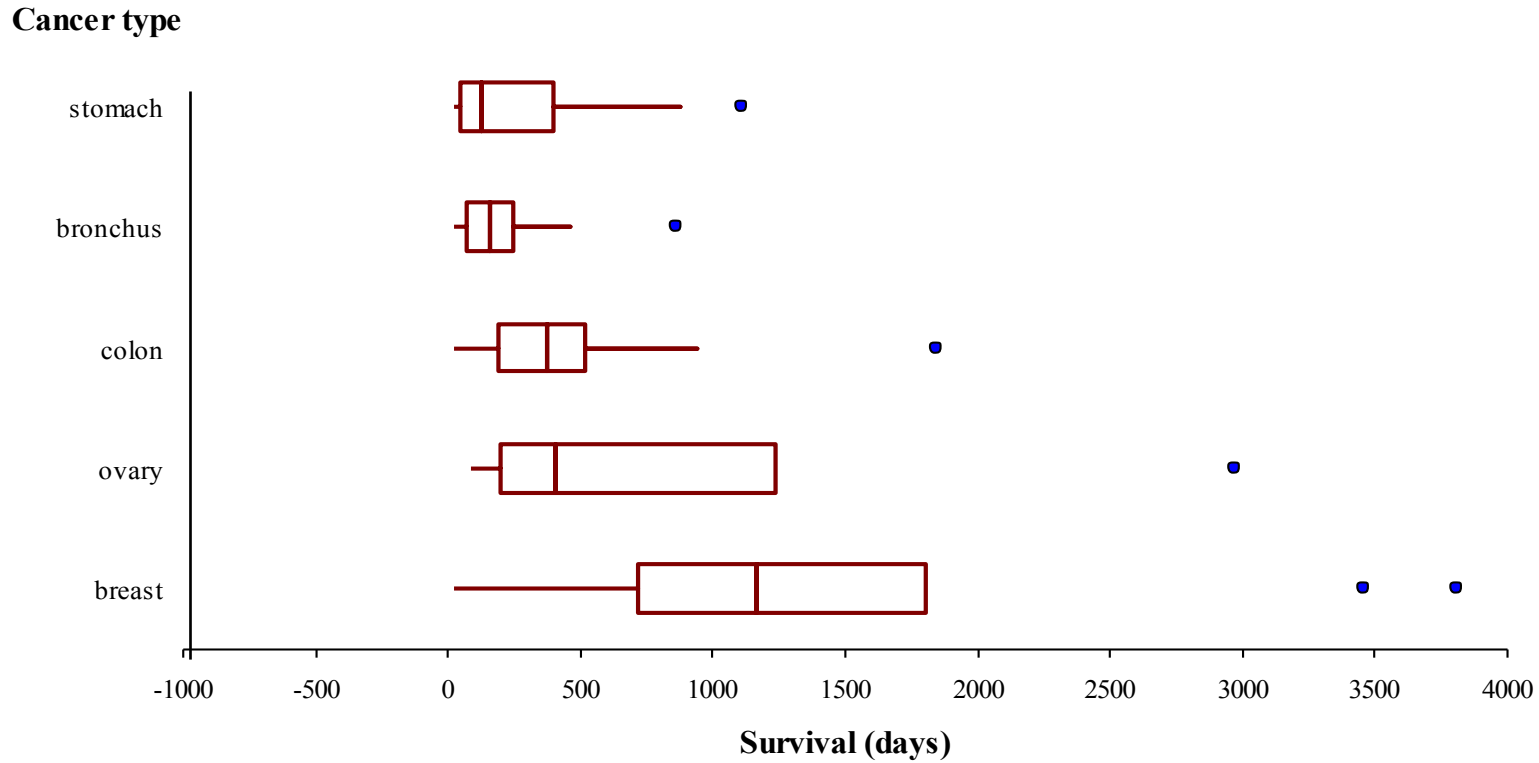
c. identify the relationship between assignment marks and exam marks?

d. compare exam marks for students majoring in Computing and students majoring in Marketing.

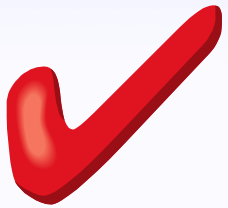
e. display mid-semester test marks for students.

Quiz 2

Comment on the following display, which compares survival times (in days) for people with five different types of cancer.



Source: *Proceedings of the National Academy of Science* (1975)

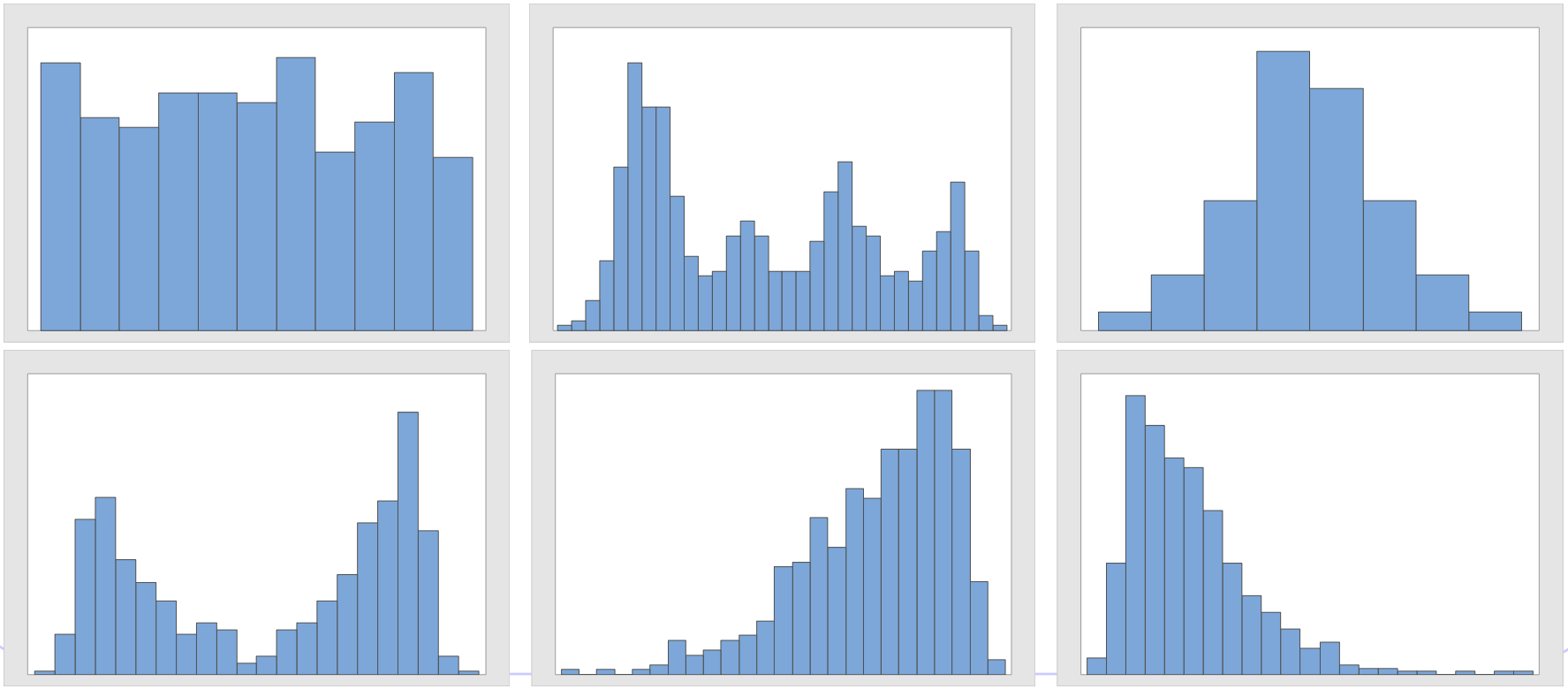


Solution to Quiz 2

Population distributions

Samples

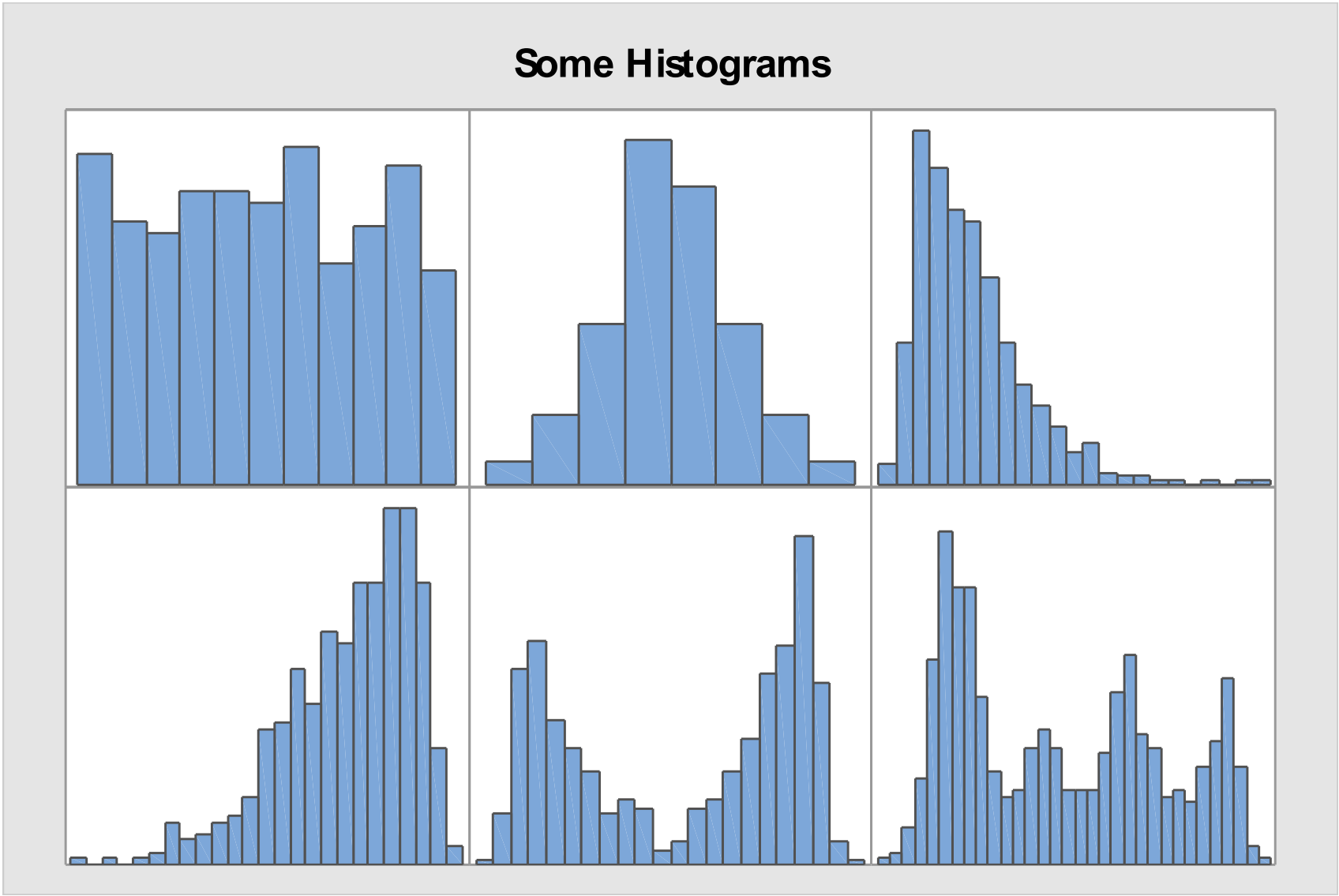
Consider a sample of numerical data which has been randomly selected from a population. We can use a histogram to display the sample. What do these histograms suggest about the populations from which they have been drawn?



Centre, Spread and Shape of a Histogram

- A ***measure of centre*** tells us where we can find the 'middle' of the data – that is, roughly, the 'middle' of the histogram.
- A ***measure of spread*** tells us how spread out the histogram is.
- What these measures don't tell us about is the ***shape*** of the data.

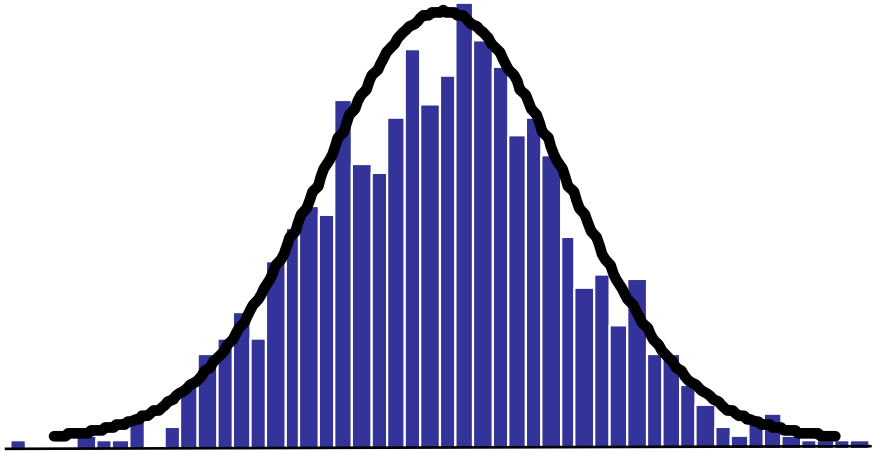
Centre, Spread and Shape of a Histogram

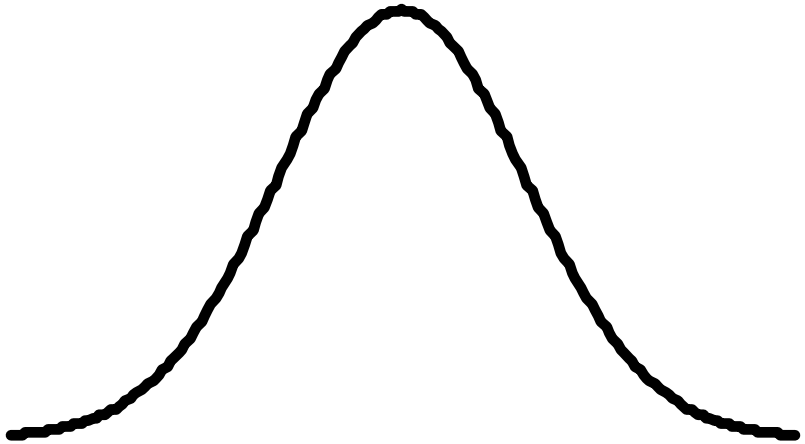


What is a Distribution?

- What do we mean by '*distribution*'? For numerical data, we can think of this as the centre, spread, and shape of its histogram.
- In the case of a *continuous population distribution*, we represent it as a *smooth curve* rather than a histogram.
- This is because our population distribution is often a theoretical model with an infinite number of observations!

Histogram of a Sample → Population Distribution





We draw a smooth curve over the top of the bars of the histogram. It should fit as closely as possible, but not outline every bump.

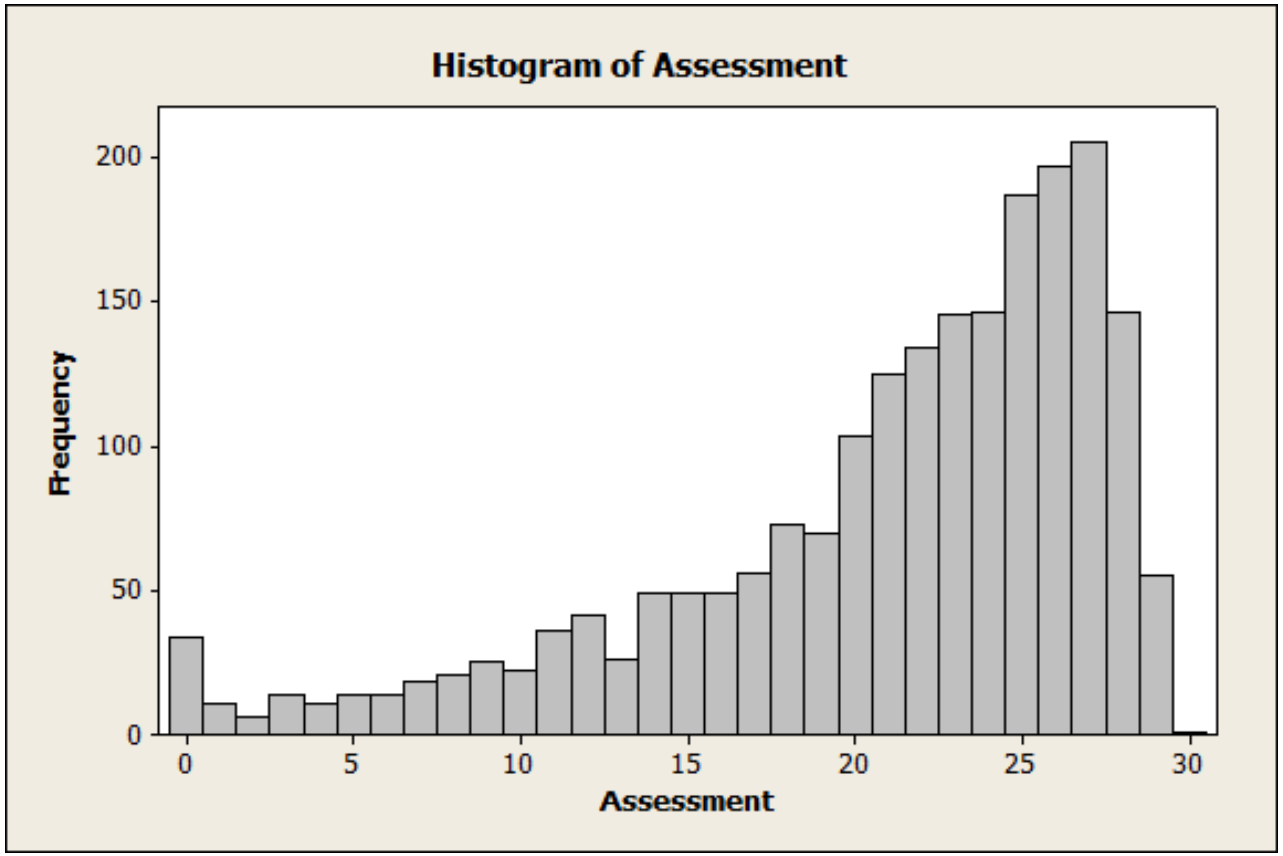
Now, to show a *population distribution*, we would just draw the curve.

Population Distributions

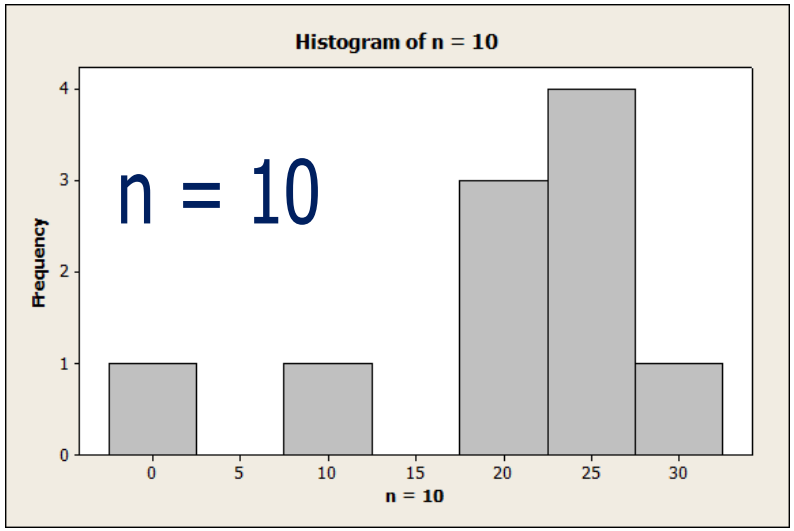
So - we usually think of samples of data as coming from some theoretical *population distribution*.

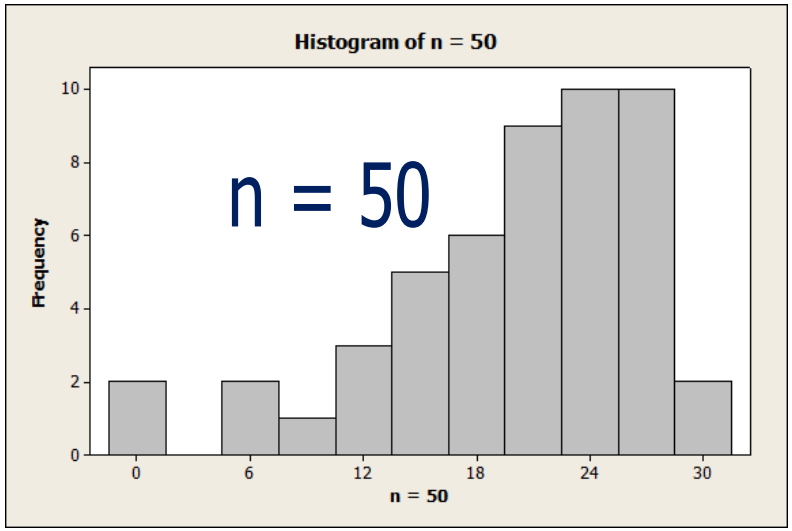
- If this is the case, then the histogram of our samples should roughly 'fit' the smooth population distribution curve.
- We cannot expect that samples will always perfectly match the populations from which they are drawn. Small samples may not always 'fit' as well as larger samples.
- We would generally expect that histograms of large samples would more closely resemble the parent population distribution than histograms of small samples.

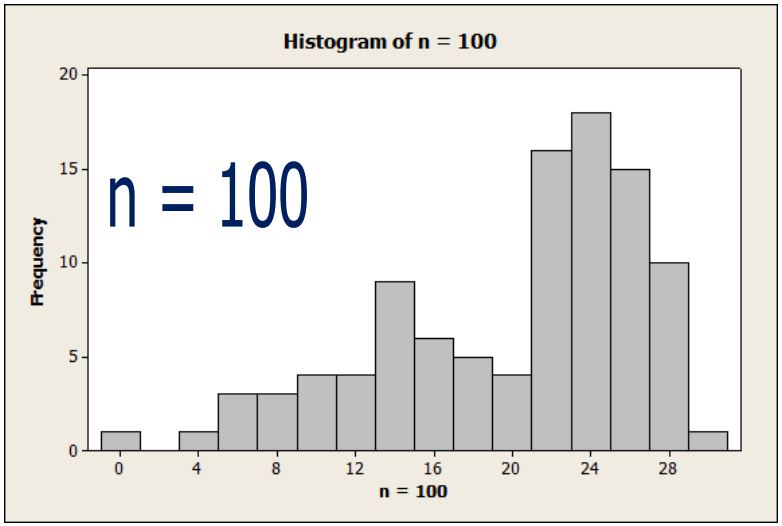
STAT170 Students' Assessment Marks

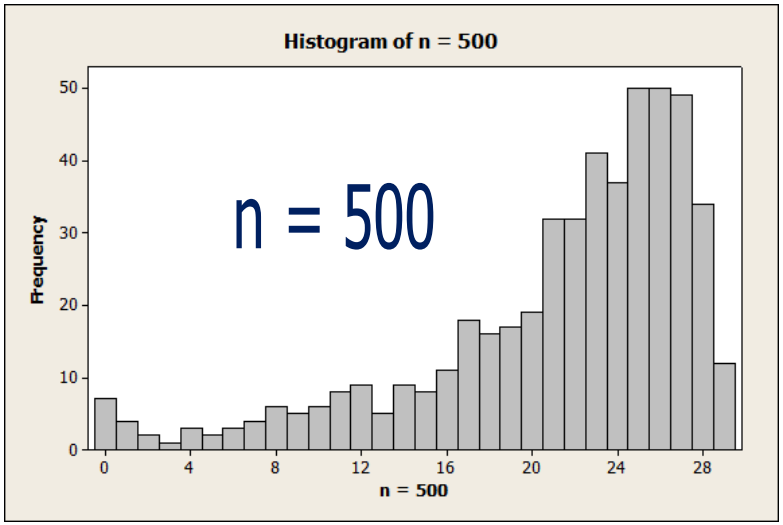


Increasing the Sample Size







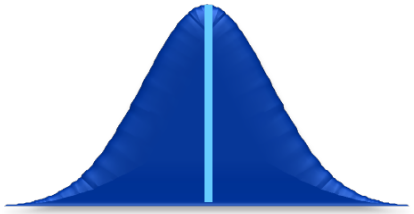


Population Distributions

- Population distributions come in many different shapes!
  - Some are symmetric.
  - Some are skewed.
  - Some are bimodal or multimodal.

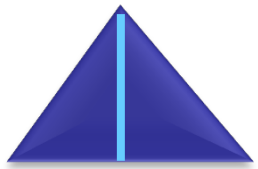
Different Population Distributions

Symmetric



Normal

Uniform



Triangular

Skewed

tail left

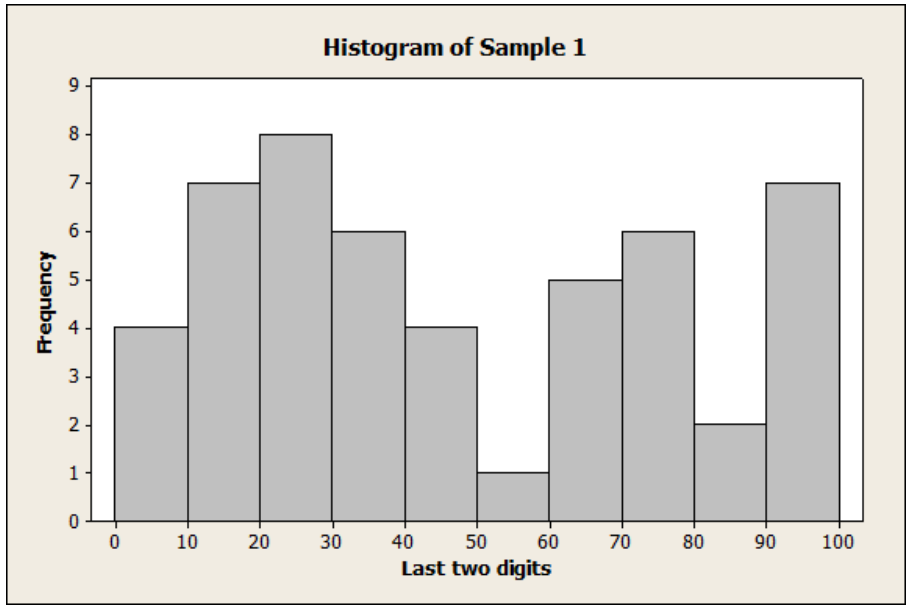
Skewed Left

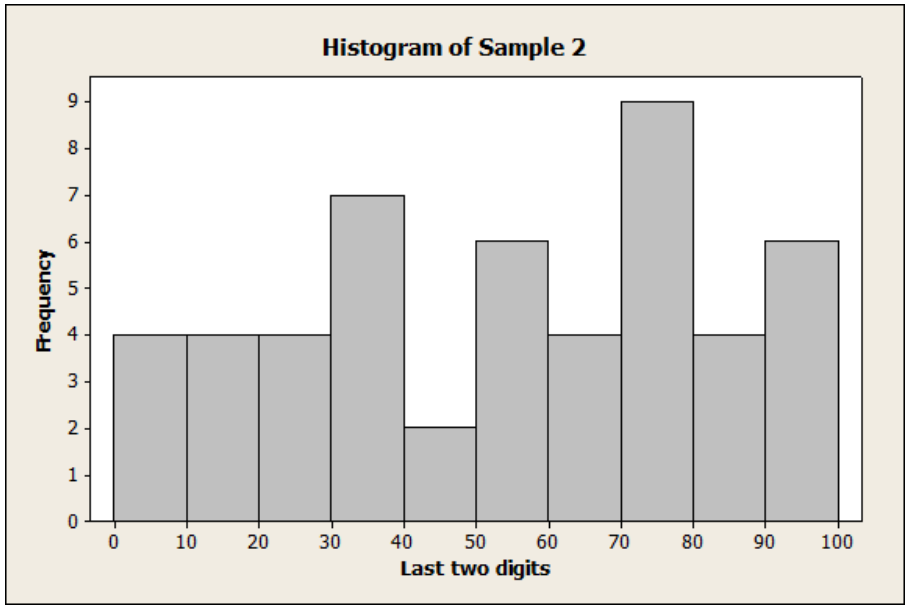
tail right

Skewed Right

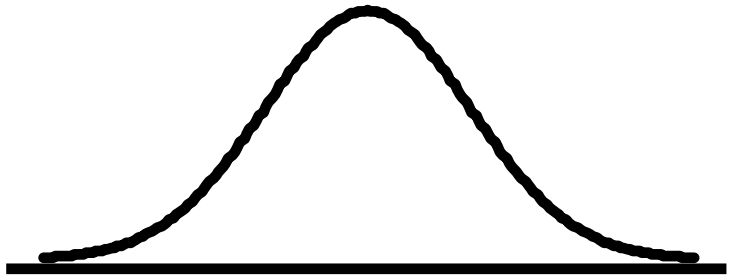
A Uniform Distribution

Histograms of two samples
(last two digits of 50 telephone numbers)

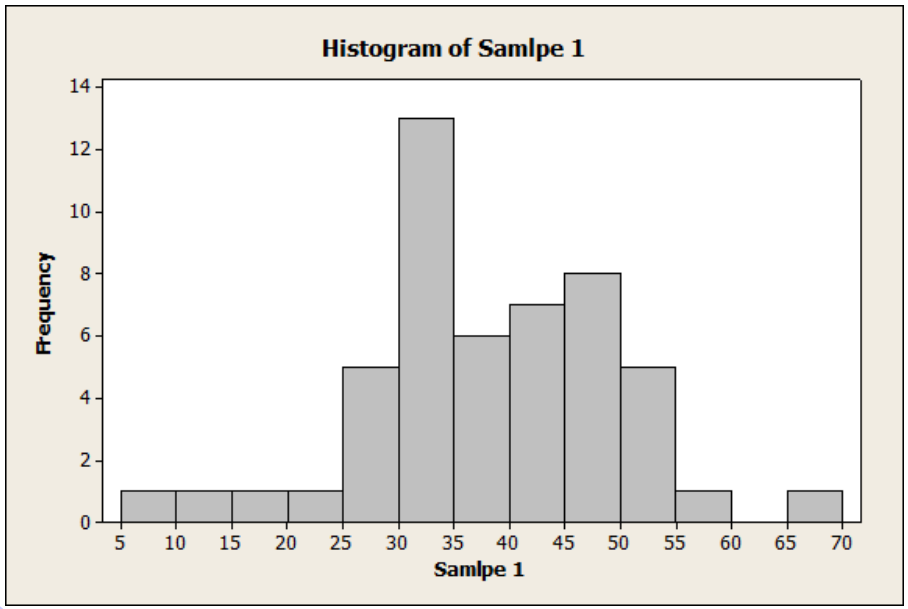


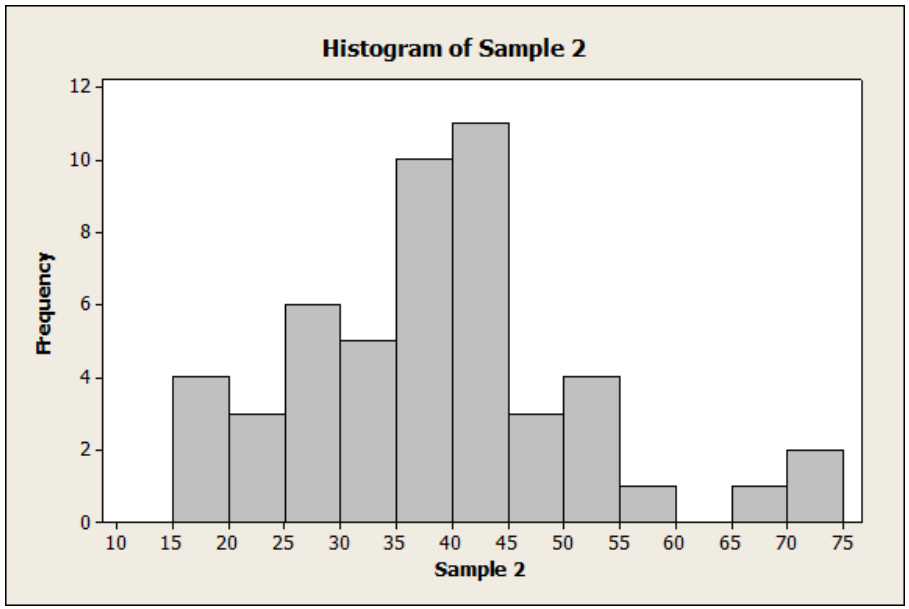


A Normal Distribution



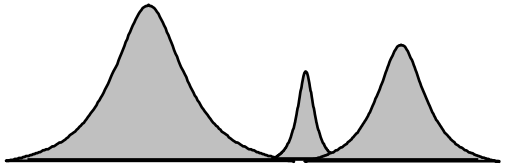
Histograms of two samples of size 50
(Heights of daffodils)



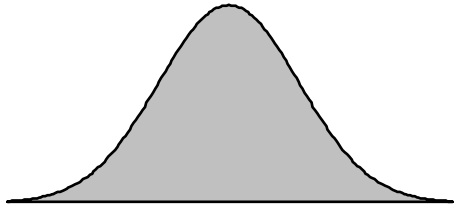


Some other Population Distributions

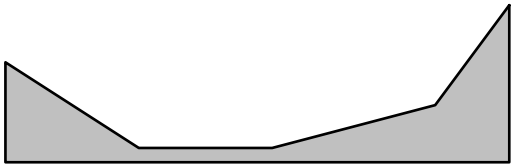
*speed of wind (*multimodal*)*



*GPAs of students (*normal*)*

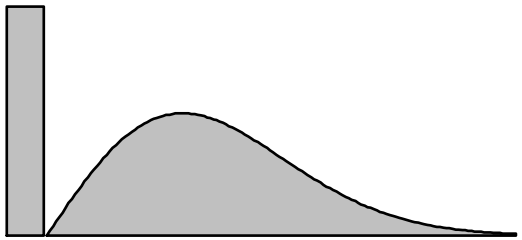


daily hours sunshine (U-shaped)

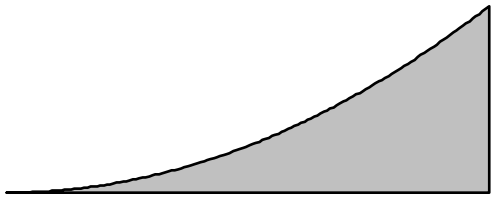


salaries of workers (skewed right)

*workers' income taxes (*zero component*)*



marks on an easy test (skewed left)



Sample Statistics and Population Parameters

Answering Research Questions

We use **sample statistics** to estimate **population parameters**

Sample Statistics and Population Parameters

- Statistical studies are concerned with using samples to learn about the population from which they have been drawn.
- We display/summarise samples graphically in order to understand what the underlying populations look like.
- We also use numerical summaries to learn about the underlying population. ie. ***we calculate sample statistics to estimate population parameters***.
  - We will start by considering some sample statistics for summarising numerical variables.
  - We will also consider the population parameters that these statistics are used to estimate.

Summarising numerical data

- Measures of centre
- Measures of spread

Measures of Centre

If we denote the variable of interest as Y then:

- The *median* (denoted $\tilde{y}$ in a sample, or $\tilde{\mu}$ in a population) is a measure of the *centre* of a set of *numerical* data defined on an interval. It is the *middle value* (or the average of the middle two values, if the sample size n is even) of a data set which has been sorted.
- The *mean* (denoted $\bar{y}$ in a sample, or μ in a population) is another measure of centre. It is defined as the *arithmetic average*, so for data $y_1, y_2, \ldots, y_n$, the *sample mean* is

$$\bar{y} = \frac{y_1 + y_2 + \ldots\ldots\ldots + y_n}{n}$$

Notation

	Sample Statistics		Population Parameters
		estimate	
Mean	$\bar{y}$	→	μ
Median	$\tilde{y}$	→	$\tilde{\mu}$

Some Sample Means and Medians

- Countries (a sample of n = 27)

  life expectancy: mean = 71.7, median = 77.1

  unemployment rate: mean = 9.9, median = 6.1

- Mammals (a sample of n = 30)

  lifespan: mean = 23.2, median = 20.0

  gestation: mean = 134.5, median = 117.5

- Students (a sample of n = 30)

  test: mean = 19.00, median = 19.50

  exam: mean = 61.78, median = 62.25

Quiz 3

Calculate the median and the mean of the following samples:

a. 1 2 3 4 5

b. 1 4 9 16 25

c. 1 2 3 4 90

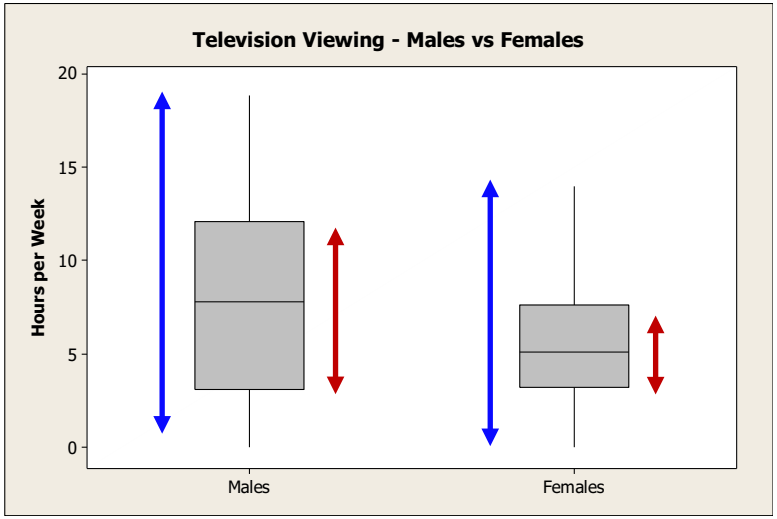
d. 1 2 3 4 990

Measures of Spread

100 male students and 100 females students took part in a study of television viewing habits. The number of hours typically spent watching television per week was recorded for each student and the following statistics were obtained:

Males: median = 7.81 hours, mean = 8.03 hours

Females: median = 5.12 hours, mean = 5.45 hours



However, we can also see that the spread of viewing times is greater among males than among females. *How should we measure spread/variation?*

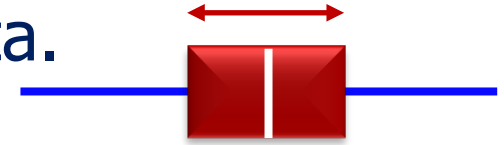
Spread/Variation

- The median or mean summarises a sample/distribution using a single measure of its *centre*. The box plots on the previous slide show that the average time males typically watch television each week appears to be about two or three hours longer than the average time for females.

- To summarise data and to make useful comparisons we also need to consider the *spread* or *variation in the data*. We can see from the comparative boxplots on the previous slide that the variation in times spent watching television each week is also greater among males than among females. So how can we measure spread/variation?

Inter-Quartile Range

- The *inter-quartile range (iqr)* is the difference between the upper and lower quartiles. It therefore gives the range of the middle 50% of a set of data.



- So the inter-quartile range is one way of summarising the *spread* of data. For the television viewing times:

Males: lq = 3.1, uq = 12.1, so iqr = 9 hours

Females: lq = 3.2, uq = 7.6, so iqr = 4.4 hours

IQR is also referred to as the midspread.

Range

- *The range is difference between the maximum value and the minimum value in the data.*

- The ***range***, unlike the *inter-quartile range*, will be influenced by large or small outliers in the data.
- The range of television viewing is 19 hours for males and 14 hours for females.
- Both the inter quartile range and the range indicate that variability in television viewing times is greater among males than among females.

Quiz 4

Find the range and the inter-quartile range of each of the following samples:

a. Maximum temperature (°C) for one week in December:

23.7 24.3 26.9 29.9 30.4 30.7 36.5

b. Number of days to row solo across the Atlantic Ocean for sample of 8 people:

40 67 70 78 81 87 106 153

c. Winning number of hot dogs eaten in 10 minutes in 4th of July New York City hot dog eating contest over the past 10 years:

45 49 50 54 54 54 59 62 66 68

Source: *Statistics, Unlocking the Power of Data, Lock et al, 2013*

Standard Deviation

The ***standard deviation*** is another measure of spread. It is defined in terms of the deviations of the data from the mean (called *residuals*). It is the square root of the average squared residual.

Residual = $y_i - \bar{y}$, **ie. observed value – sample mean.**

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots\ldots + (y_n - \bar{y})^2}{n - 1}}$$

To get the average squared residual, you divide the sum of squared residuals by n – 1 (not n).

The sample standard deviation, s, estimates the population standard deviation, σ. The variance, σ^2, is the square of the standard deviation and is estimated by s^2.

Calculating the Standard Deviation

Number of days to row solo across the Atlantic Ocean:

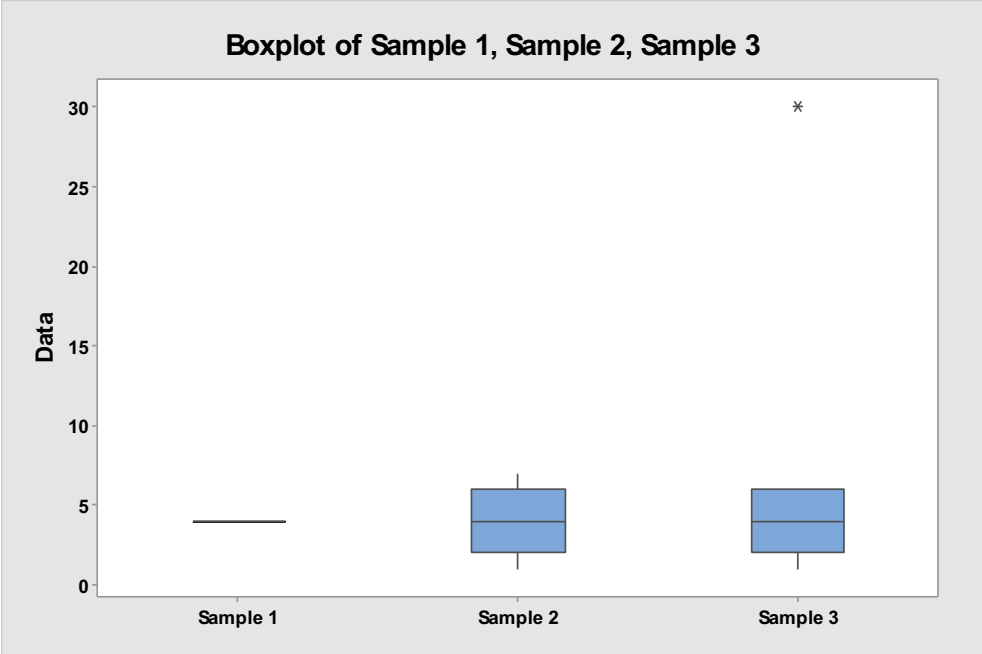
Number of days with mean = 85.25 days		
y	$(y-\bar{y})$	$(y-\bar{y})^2$
40	-45.25	2048
67	-18.25	333
70	-15.25	233
78	-7.25	53
81	-4.25	18
87	1.75	3
106	20.75	431
153	67.75	4590
	$\sum(y-\bar{y})=0$	$\sum(y-\bar{y})^2=7708$

$$s=\sqrt{\frac{\sum_i(y_i-\bar{y})^2}{n-1}}$$

$$=\sqrt{\frac{7708}{7}}=33.2$$

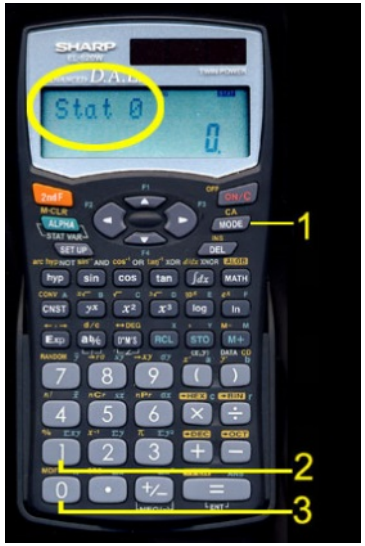
Standard Deviation

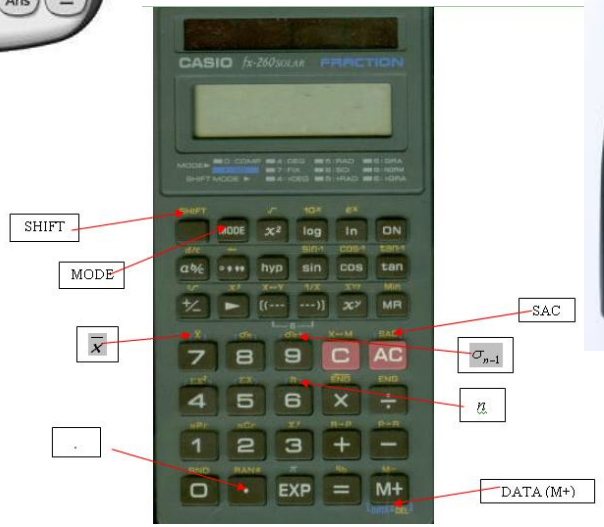
								Median	Mean	StDev
Sample 1	4	4	4	4	4	4	4	4	4	0
Sample 2	1	2	3	4	5	6	7	4	4	2.16
Sample 3	1	2	3	4	5	6	30	4	7.29	10.16



The standard deviation gives a measure of how much the data are spread out around the mean

Using Calculator in STAT/SD Mode





Quiz 5

Using your calculator in stat (sd) mode, compute the standard deviation of each of the following samples:

a. Sale price ($'000s) for six 2-bedroom apartments in Glebe.

520 539 650 660 780 805

b. Waiting times (minutes) for eight customers in a bank.

8 8 9 9 10 10 10 11 14 14

Notation

	Samples Statistics	estimate	Populations Parameters
Mean	$\bar{y}$	→	μ
Median	$\tilde{y}$	→	$\tilde{\mu}$
Std.dev	s	→	σ
Variance	s^2	→	σ^2

Presentation – Summary Statistics for Numerical Data

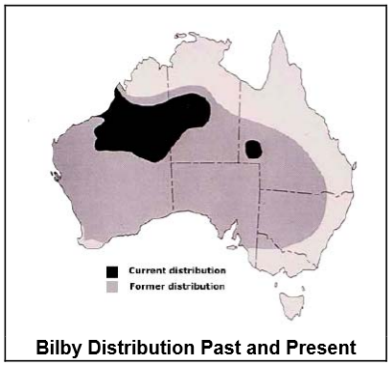
The best way to ensure that data summaries are easy to read is to display them in a table.

For numerical data the table should show the variable name, the sample size (n), the mean, median and standard deviation. You could also include the lower and upper quartiles, the maximum and minimum values.

Variable	Mean	Median	StDev	Q1	Q3	Maximum	Minimum

Table of Descriptive Statistics

Bilbies are desert-dwelling marsupial omnivores. The Greater Bilby is on the endangered list and the Lesser Bilby is believed to be extinct. A study of the Greater Bilby involved recording the sex and weight (kg) of a random sample of 40 of these marsupials. Descriptive statistics comparing weights of males and females are presented in a table:



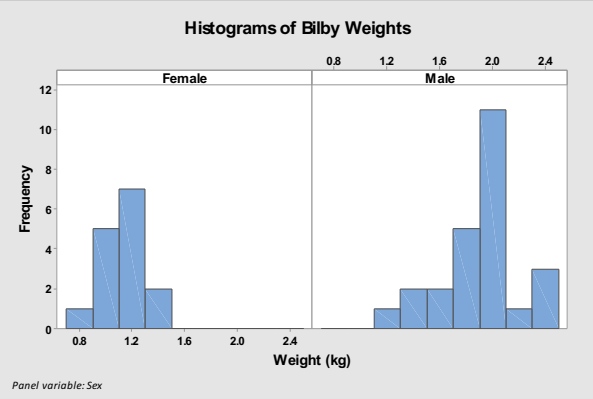
Bilby Distribution Past and Present

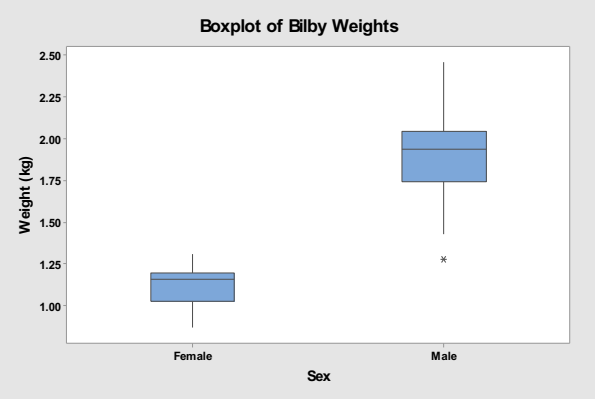
Descriptive Statistics: Weight

Variable	Sex	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Weight	Female	15	1.12	0.12	0.87	1.02	1.16	1.19	1.31
	Male	25	1.90	0.29	1.27	1.74	1.93	2.04	2.46

Quiz 6

Use the numerical summary and the graphical summaries provided to comment on the weights of bilbies.





Descriptive Statistics: Weight

Variable	Sex	N	Mean	StDev	Min	Q1	Median	Q3	Max
Weight	Female	15	1.12	0.12	0.87	1.02	1.16	1.19	1.31
	Male	25	1.90	0.29	1.27	1.74	1.93	2.04	2.46

Describing the Centre and the Spread of Population Distributions

- Population distributions are usually defined by their mean μ (*mu*), standard deviation σ (*sigma*) and shape.
  - Should we always use the mean rather than the median to describe the centre of a population distribution?
  - Should we always use the standard deviation rather than the range or the interquartile range to describe the spread of a population distribution?
- That all depends on the shape of the distribution.

Mean vs Median

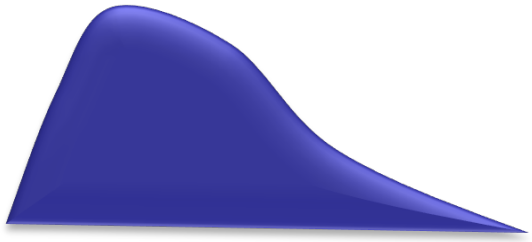
The *median* divides the area in the distribution into two equal parts.

The *mean* is the centre of gravity (point of balance) of the data.

Which is Better - Mean or Median?

- Both the median and the mean summarise numerical data, using a single value to give an indication of the overall location of the measurements.
- The mean and median are **not the same**, unless the distribution is symmetric eg:

$\mu = \tilde{\mu}$



$\mu \neq \tilde{\mu}$

- The median is not as sensitive to outliers as the mean.
- So the median is a ***more robust*** measure of centre than the mean (less easily influenced by outliers).

Which is Better – Range, Inter-quartile Range or Standard Deviation?

- In previous examples, we saw that the range may be greatly inflated by even one outlier.
- We also saw that the standard deviation may be greatly affected by outliers.
- We saw that unlike the range and the standard deviation, the inter-quartile range is not sensitive to outliers. It is not always necessarily affected by outliers.
- So the inter-quartile range, like the median, is a ***robust*** summary measure. The inter-quartile range is a more robust measure of spread than the standard deviation.

Summarising categorical data

- Proportions and Percentages

Summarising Categorical Data

- The median and mean are measures of *centre* for *numerical* data.
- For *categorical* data, the mean and median are *not* appropriate.
- To summarise a categorical variable, construct a table showing the *variable name, the name of each category, the count, proportion* and/or *percentage* of observations in each category. The table should also show the total sample size (n).

Quiz 7

Consider the categorical variable 'Region' in the Countries data set we looked at in Week 1.

Country	Region	Country	Region	Country	Region	Country	Region
Australia	Oceania	Germany	Europe	Malaysia	Asia	Sri Lanka	Asia
Cameroon	Africa	Greece	Europe	Mozambique	Africa	Sweden	Europe
Chad	Africa	Hong Kong	Asia	Namibia	Africa	Tanzania	Africa
China	Asia	India	Asia	New Zealand	Oceania	Tunisia	Africa
Ethiopia	Africa	Indonesia	Asia	Singapore	Asia	UK	Europe
Fiji	Oceania	Italy	Europe	Solomon Is.	Oceania	Zambia	Africa
France	Europe	Japan	Asia	Spain	Europe		

Produce a table showing the number and the proportion of observations occurring in each category and indicate the mode, the category which occurs most frequently.

Source: *World Fact Book* (2013)

Solution to Quiz 7

Summary of Region:

	Number	Proportion
Africa		
Asia		
Europe		
Oceania		
Total		

Homework Questions

Homework Question 1

Consider the variables outlined on the next slide.

For each variable measured, draw a rough diagram of what you think the underlying population distribution might look like.

Solution to Homework Question 1

Diagrams of the population distributions:

Variable	Distribution
Heights of mature spotted gum trees.	
Ages of people surfing at Manly beach.	
Winning numbers in Lotto.	
Heights of emus.	

Homework Question 2

The following data represents the weights in kg of 6 new born babies:

2.5	2.7	2.9	3.3	3.8	4.0

a. Compute the mean, median, range and standard deviation of the babies' weights.

b. Each of these babies puts on 500g in the first two weeks. What was the mean, median, range and standard deviation of their weights after two weeks?

c. Each of these babies doubles their weight in their first six months. What is the mean, median, range and standard deviation of their weights at six months of age?

Solution to Homework Question 2

Lecture 3 Summary

- The median & mean summarise *centres* of data sets.
- The range, inter-quartile range (iqr) & standard deviation (s) summarise *spreads* of data sets.
- For categorical data we use the proportions in each category to summarise the data.
- Population distributions can take many shapes.
- A population distribution is often specified by its mean, standard deviation and shape.
- We expect histograms of large samples to more closely resemble the parent population distribution than histograms of small samples.

Textbook References

Further information on the topics discussed in this lecture can be found in:

Modern Statistics: An Introduction
by Don McNeil and Jenny Middledorp
(ISBN 9781486007011).

- Chapter 3: Pages 52 – 68