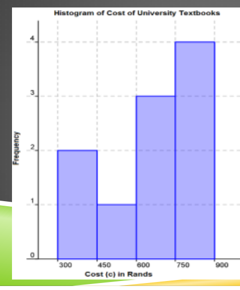
10. STATISTICS

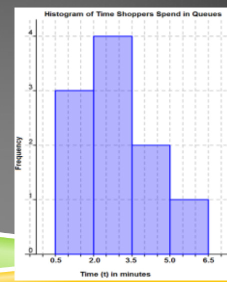
- Histograms
- Frequency Polygons
- Cumulative Frequency Graphs (Ogives)
- Measures of Central Tendency
  - * Mean
  - * Median
  - * Mode
- Measures of Dispersion
  - * Range
  - * Interquartile Range
  - * Semi-interquartile
  - * Range
  - * Quartiles
  - * Outliers
  - * Standard Deviation
- Box-and-Whisker Plots
- Distribution Curves
- Scatter Plots

1

HISTOGRAMS

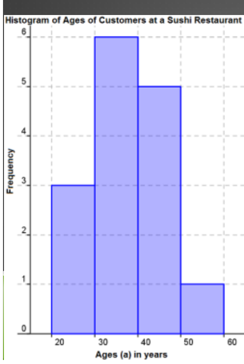
- Usually used to display grouped data
- NO spaces between the bars





2

EXAMPLE: READING-OFF HISTOGRAMS



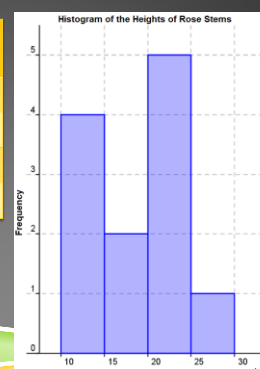
1. How many 40 – 50 year old's were eating at the Sushi restaurant?
   5 people
2. Which age group was most represented in the restaurant?
   30 – 40 year old's

3

EXAMPLE: DRAWING HISTOGRAMS

Height of Rose Stems (cm)	Frequency
$10 < h \leq 15$	4
$15 < h \leq 20$	2
$20 < h \leq 25$	5
$25 < h \leq 30$	1

NB: Be sure to label your graph & axes!!



4

EXERCISE: HISTOGRAMS

Draw a histogram of the length of time an average battery lasts.

Time a battery lasts (in hours)	Frequency
$1 < t \leq 3$	5
$3 < t \leq 6$	9
$6 < t \leq 9$	12
$9 < t \leq 12$	3

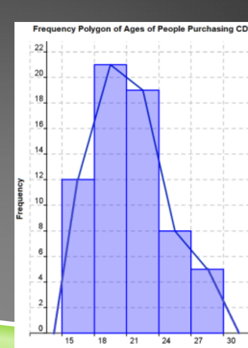
Now answer the following questions:

1. How many hours does the longest battery last?
2. What percentage of batteries last 4 hours?

5

FREQUENCY POLYGONS

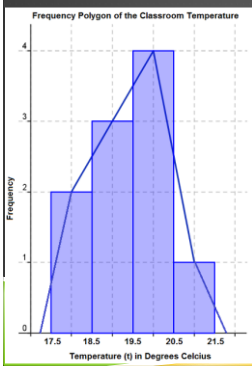
- Can be thought of a histogram with a line graph drawn over it
- The points of the line graph are plotted on top of each bar of the histogram, in the position of the midpoint of each bar



Drawing a frequency polygon

6

EXAMPLE: READING-OFF FREQUENCY POLYGONS



1. How often was the class-room between 18,5-19,5 degrees Celcius?
   3 occasions
2. What was the lowest temperature recorded in the classroom?
   17,5 degrees Celcius

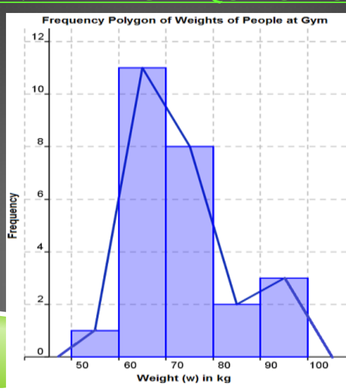
EXAMPLE: DRAWING FREQUENCY POLYGONS

Weight (kg)	Frequency	Midpoint
$50 < w \leq 60$	1	55
$60 < w \leq 70$	11	65
$70 < w \leq 80$	8	75
$80 < w \leq 90$	2	85
$90 < w \leq 100$	3	95

Midpoint $= \frac{50 + 60}{2}$
$= 55$

NB: The straight line must be extended down to the x-axis, on either side of the bars!

EXAMPLE: DRAWING FREQUENCY POLYGONS



EXERCISE: FREQUENCY POLYGONS

Draw a frequency polygon of volume of water in a swimming pool over time.

Volume (litres)	Frequency
$1000 < v \leq 3000$	5
$3000 < v \leq 5000$	9
$5000 < v \leq 7000$	12
$7000 < v \leq 9000$	3

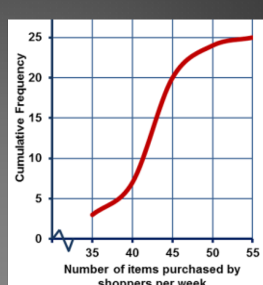
First calculate the midpoints!

Now answer the following questions:

1. How much water do most swimming pools hold?
2. What percentage of pools hold 8000 litres of water?

CUMULATIVE FREQUENCY GRAPHS

- Has a typical "S"- shaped curve
- Also known as OGIVES
- Reflects the cumulative frequency (i.e. a running total of frequencies)



EXAMPLE: DRAWING OGIVES

Mass (in kg)	20 -	30 -	40 -	50 -	60-70
Frequency	3	5	10	8	4
Cumulative Frequency	3	3+5 = 8	8+10 = 18	18+8 = 26	26+4 = 30

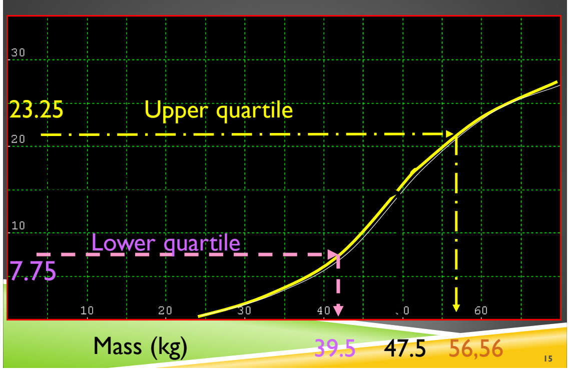
Cumulative frequency = previous frequency + current frequency

- Cumulative frequency is plotted against the upper class boundary for each class interval

EXAMPLE: READING-OFF OGIVES

- Since ogives indicate the cumulative frequency, we can determine the position of the quartiles
- Recap! Quartiles divide the data into quarters
  * Lower quartile (Q1): 25% of the data lies below Q1 and 75% of the data above
  * Median (Q2): 50% of the data lies below Q2 and 50% of the data above
  * Upper quartile (Q3): 75% of the data lies below Q3 and 25% of the data above

EXAMPLE: READING-OFF OGIVES



EXAMPLE: READING-OFF OGIVES

- For data values smaller than 50 (i.e. n < 50), the position of the quartiles can be found as follows:
- In our example: n = 30:

The lower quartile $= \frac{1}{4}(n+1) = \frac{1}{4}(31)$
$= 7.75^{th}$ position

The median $= \frac{1}{2}(n+1) = \frac{1}{2}(31)$
$= 15.5^{th}$ position

The upper quartile $= \frac{3}{4}(n+1) = \frac{3}{4}(31)$
$= 23.25^{th}$ position

Cumulative Frequency Graphs

EXERCISE: OGIVES

Draw a cumulative frequency graph of the distance workers travel to work per day:

Distance (km)	Frequency
$0 < d \leq 20$	12
$20 < d \leq 40$	26
$40 < d \leq 60$	8
$60 < d \leq 80$	4

NB! Plot cumulative frequency vs the upper class boundary!

Now answer the following questions:

1. How many workers were surveyed?
2. Determine the upper quartile.
3. Determine the lower quartile.

MEASURES OF CENTRAL TENDENCY

1. MEAN

- Also known as the average
- Easy to calculate:

$$\bar{x} = \frac{\sum x}{n} = \frac{\text{sum of all the } x \text{ values}}{\text{number of values in the data set}}$$

- Use it if all the actual values are relevant
- Do not use it if it is distorted by outliers

MEASURES OF CENTRAL TENDENCY

2. MEDIAN

Recap: Mean and Median

- Also known as Q2 (from quartiles), so 50% of the values lie above and 50% of the values lie below it
- Data values must be organized into order
- Can only be found for numbers
- If there are an even number of entries, it may not be one of the values.

19

MEASURES OF CENTRAL TENDENCY

3. MODE

Measures of Central Tendency Example

- Mode is the value that occurs most frequently
- Easy to find from diagram, frequency table or bar graph.
- No calculations are necessary to find it.
- Is always one of the data values

20

MEASURES OF DISPERSION

1. RANGE

- Measures how far spread out the data is
- Easy to calculate:

  Range = highest value – lowest data value

21

MEASURES OF DISPERSION

2. QUARTILES

Finding Quartiles

- Divides the data into quarters
- For data values < 50, we can calculate
  - Lower quartile (Q1): $Q_1 = \frac{1}{4}(n+1)$
  - Median (Q2): $Q_2 = \frac{1}{2}(n+1)$
  - Upper quartile (Q3): $Q_3 = \frac{3}{4}(n+1)$

22

MEASURES OF DISPERSION

- The 3 quartiles, together with the smallest and largest data value makes up the "5-Number Summary"

Data Representation of USA Statistics

23

MEASURES OF DISPERSION

3. INTERQUARTILE RANGE (IQR)

- Measures the spread of the "core" 50% of the data
- Far more useful than range
- Need to first calculate the upper (Q1)and lower quartiles (Q3)
- IQR = Q3 – Q1

Working with quartiles and IQR

24

MEASURES OF DISPERSION

4. SEMI-QUARTILE RANGE (Q_s)

- Is half the inter-quartile range
- Q_s = ½ (Q3 – Q1)

MEASURES OF DISPERSION

5. OUTLIERS

- Any value that is considered outside the group of data
- Calculated as:

  Outlier > Q3 + 1,5 IQR or

  Outlier < Q1 – 1,5 IQR

MEASURES OF DISPERSION

6. STANDARD DEVIATION

Understanding Standard Deviation

- Measure of spread around the mean
- Commonly used in statistical investigations
- Takes all data values into account
- Can be calculated using a table or your calculator …

Steps to calculate the standard deviation:

E.g. The temperatures for a week in December were recorded as: 27, 31, 27, 27, 28, 30, 27

- 1. Calculate the mean

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{197}{7}$$

$= 28,14$

[From Step 1: Mean = 28,14]

- 2. Take each data value and subtract it from the mean …

So, we are finding the deviation from the mean

x	$(x - \bar{x})$
27	-1,14
31	2,86
27	-1,14
27	-1,14
28	-0,14
30	1,86
27	-1,14

- 3. Square the difference between the data value and the mean

x	$(x - \bar{x})$	$(x - \bar{x})^2$
27	-1,14	1,2996
31	2,86	8,1796
27	-1,14	1,2996
27	-1,14	1,2996
28	-0,14	0,0196
30	1,86	3,4596
27	-1,14	1,2996

- 4. Add the sum of all the squared differences …

x	$(x-\bar{x})$	$(x-\bar{x})^2$
27	-1,14	1,2996
31	2,86	8,1796
27	-1,14	1,2996
27	-1,14	1,2996
28	-0,14	0,0196
30	1,86	3,4596
27	-1,14	1,2996
	$\sum(x-\bar{x})^2$	16,8572

31

- [From Step 4: $\sum(x-\bar{x})^2 = 16,8572$]
- 5. To calculate the variance (σ^2), divide the sum of all the squared differences by the total number of data values

$$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$$
$$= \frac{16,8572}{7}$$
$$= 2,408 \ldots$$

Calculating the mean and variance

32

- [From Step 5: $\sigma^2 = 2.408$ …]
- 6. To calculate the standard deviation (σ), take the square root of the variance

$$Standard\ deviation = \sqrt{Variance}$$
$$\sigma = \sqrt{2,408 \ldots}$$
$$= 1,55$$

- The smaller the number the narrower the data spread (and vice versa)

33

Using a calculator to find the standard deviation:

Using a CASIO fx-82ES PLUS calculator:

- 1. Get the calculator into Stats Mode:
  [MODE] [2:STAT]
- 2. Specify that you are working with univariate data:
  [1: 1-VAR]

34

E.g. Data set: 2, 5, 6, 8, 9

- 3. Enter data values and press = after each one
  [2 =] [5 =] [6 =] [8 =] [9 =]
- 4. Now clear your screen after the data values have been inputted
  [AC]

35

- 5. To find the mean
  [SHIFT STAT]
  [4: Var]
  [2: $\bar{x}$] [=]
  Answer = 6
- 6. To find the standard deviation
  [SHIFT STAT]
  [4: Var]
  [3: σx] [=]
  Answer = 2,44948…

36

STANDARD DEVIATION

- Given the following set of data: 12, 4, 11, 26, 8

a) Identify the outlier.
26

b) Determine the standard deviation.
$\bar{x} = 12{,}2$
$\sigma = 7{,}44$

STANDARD DEVIATION

Given the following set of data: 12, 4, 11, 26, 8
[From previous question: $\bar{x} = 12{,}2$ and $\sigma = 7{,}44$]

c) How many data values fall within one standard deviation of the mean?

$\bar{x} + \sigma = 12{,}2 + 7{,}44 = 19{,}64$

$\bar{x} - \sigma = 12{,}2 - 7{,}44 = 4{,}66$

Data set that falls within 4,66 – 19,64 is:
2 ✓, 4 ×, 11 ✓, 26 ×, 8 ✓ ∴3 data values

BOX-AND-WHISKER PLOTS

- Show the distribution of the "5-Number Summary" i.e. Lowest no; Q1; Q2; Q3; and Highest no

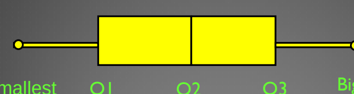
Playing with Box-and-Whisker Plots

Histograms & Box-and-Whisker Plots

- The distribution or spread of the data can then be classified as symmetrical or skewed

BOX-AND-WHISKER PLOTS

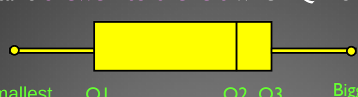
- Symmetrical distribution can be seen when Q2 is in the middle of Q1 and Q3



Interpreting Box-and-Whisker Plots

BOX-AND-WHISKER PLOTS

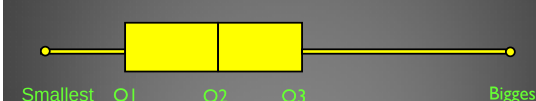
- Data is skewed to the left when Q2 is closer to Q3

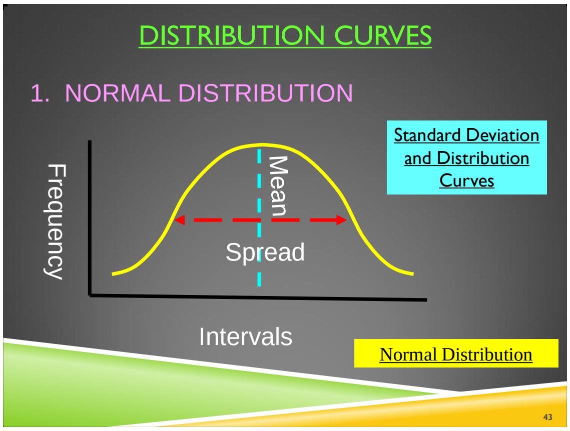


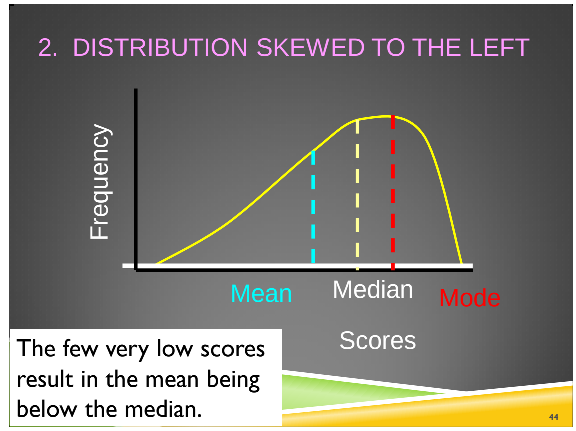
- Data is skewed to the right when Q2 is closer to Q1

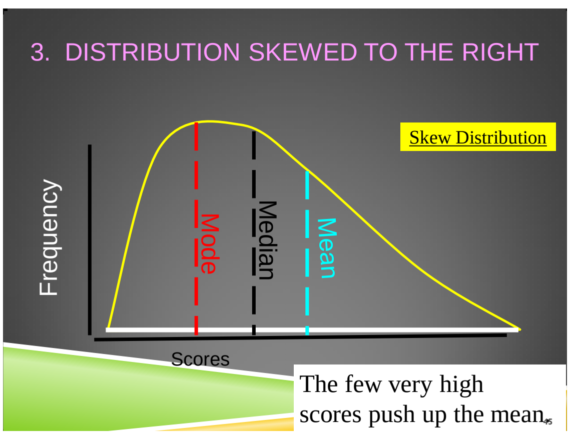
BOX-AND-WHISKER PLOTS

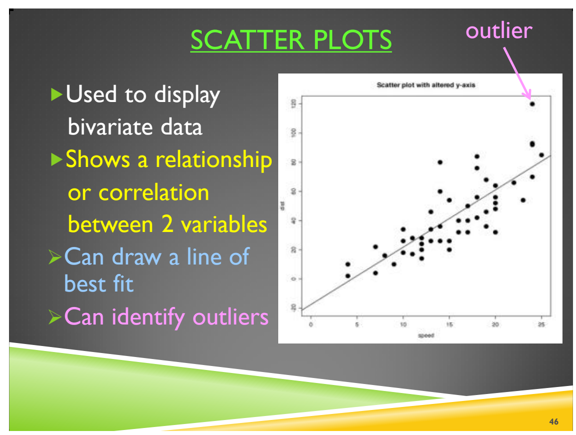
- If the biggest number in the data set is far removed from the bulk of the data, then it is an outlier
- The outlier will result in a long "whisker"

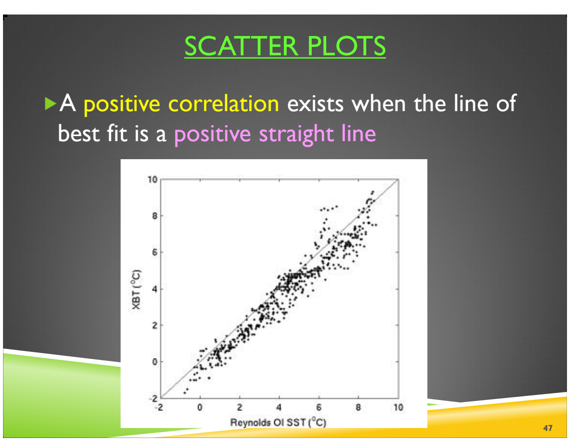


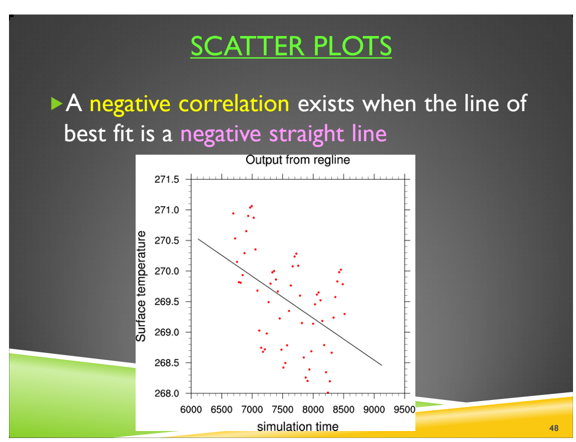

DISTRIBUTION CURVES
1. NORMAL DISTRIBUTION
Standard Deviation and Distribution Curves
Frequency
Mean
Spread
Intervals
Normal Distribution
43


2. DISTRIBUTION SKEWED TO THE LEFT
Frequency
Mean
Median
Mode
Scores
The few very low scores result in the mean being below the median.
44


3. DISTRIBUTION SKEWED TO THE RIGHT
Skew Distribution
Frequency
Mode
Median
Mean
Scores
The few very high scores push up the mean.


SCATTER PLOTS
outlier
Used to display bivariate data
Shows a relationship or correlation between 2 variables
Can draw a line of best fit
Can identify outliers
46

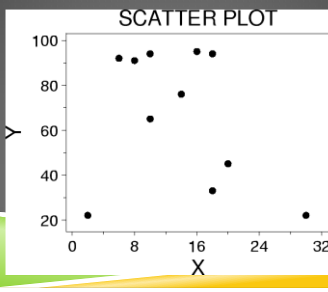

SCATTER PLOTS
A positive correlation exists when the line of best fit is a positive straight line
XBT (°C)
Reynolds OI SST (°C)
47


SCATTER PLOTS
A negative correlation exists when the line of best fit is a negative straight line
Output from regline
Surface temperature
simulation time
48

SCATTER PLOTS

Scatter plots and correlations

- No correlation exists when one cannot draw a line of best fit



49