CORRELATION & REGRESSION

KEY WORDS & DEFINITIONS

1. **Correlation** A description of the linear relationship between two variables.
2. **Bivariate data** Pairs of values for two variables
3. **Causal relationship** Where a change in a variable causes a change in another. Not always true.
4. **Least squares regression line** A type of line of best fit which is a straight line in the form $y = a + bx$
5. **'b' of a regression line** The gradient of the line; indicating positive correlation if it is positive and negative correlation if it is negative..
6. **Independent or Explanatory variable** The variable which occurs regardless of the other variable (e.g. time passing). Plotted on the x axis.
7. **Dependent or Response variable** The variable whose value depends on the independent variable's data points.
8. **Interpolation** Estimating a value within the range of the data. Reliable.
9. **Extrapolation** Estimating a value outside of the range of the data. NOT reliable.
10. **Product Moment Correlation Coefficient** A measure of the strength and type of correlation.

WHAT DO I NEED TO KNOW

Interpreting 'b' of a regression line:
Refer to the change in the variable y for each unit change of the variable x IN CONTEXT

PMCC, r is the PMCC for a population sample

PMCC, ρ is the PMCC for the entire population

Range of PMCC, r: $-1 \leq r \leq 1$

Hypotheses for one tailed test on PMCC:
$H_0: \rho = 0$
$H_1: \rho > 0$ or $H_1: \rho < 0$

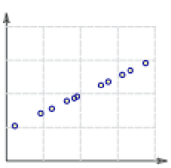
Hypotheses for two tailed test on PMCC:
$H_0: \rho = 0$
$H_1: \rho \neq 0$

Check sample size is big enough to draw a valid conclusion and comment on it if not.

A regression line is only a valid model when the data shows linear correlation.

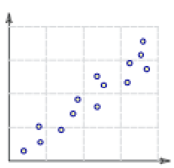
Only make predictions for the dependent variable using the regression line of y on x within the range of the original data

Perfect positive correlation



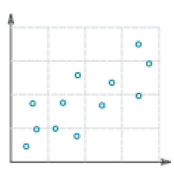
$r = 1$

Strong positive correlation



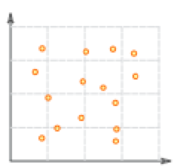
$r = 0.8$

Weak positive correlation



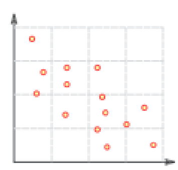
$r = 0.3$

No correlation



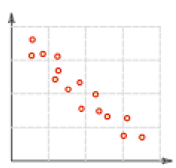
$r = 0$

Weak negative correlation



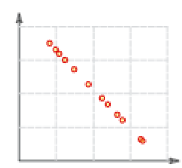
$r = -0.3$

Strong negative correlation



$r = -0.8$

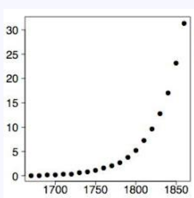
Perfect negative correlation

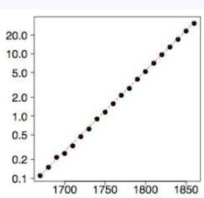


$r = -1$

EXPONENTIAL MODELS

You can use logarithms and coding to transform graphs and examine trends in non-linear data





If $y = ax^n$ then $\log y = \log a + n\log x$

If $y = kb^x$ then $\log y = \log k + x\log b$

@Lucyjc1612

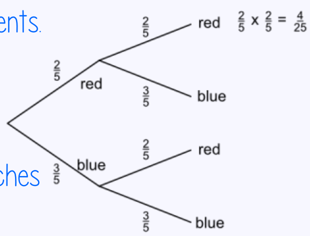
PROBABILITY

KEY WORDS & DEFINITIONS

1. **Experiment** A repeatable process that results in a number of outcomes.
2. **Event** A collection of one or more outcomes.
3. **Sample Space** The set of all possible outcomes. ξ is the universal set.
4. **Mutually Exclusive** Events that have no outcomes in common.
5. **Independent** When events have no effect on another.
6. **Intersection** When two or more events all happen.
7. **Union** When one or both events happen.
8. **Complement** When an event does not happen.

TREE DIAGRAMS

You can use tree diagrams to show the outcome of 2 or more successive events.



Multiply ALONG the branches

Add all the favourable final probabilities.

WHAT DO I NEED TO KNOW

Probabilities of all possible outcomes add to 1

Probability values must be between 0 and 1

Intersection $A \cap B \Rightarrow$ A AND B happen

Union $A \cup B \Rightarrow$ A OR B OR BOTH happen

Complement of A is A′ $\Rightarrow$ NOT A

$P(A') = 1 - P(A)$

Mutually Exclusive events:

$P(A \cup B) = P(A) + P(B)$

Independent Events:

$P(A \cap B) = P(A) \times P(B)$

Probability of B, given A has occurred:

$P(B \mid A)$

For independent events:

$P(A \mid B) = P(A \mid B') = P(A)$

In formulae book:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

VENN DIAGRAMS

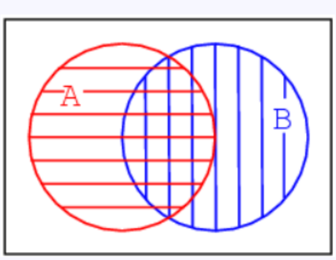
Venn diagrams can be used to show either probabilities or the number of outcomes.

n(A) is the number of outcomes while P(A) is the probability of an outcome

e.g. n(Aces) = 4   P(Ace) = 4/52

Use cross hatch shading to help you work out probabilities.

Focus on one condition at a time, ignoring the other condition completely when you shade.



If P(A) = // and P(B) = \\

$P(A \cap B)$ = #

$P(A \cup B)$ = // + \\ + #

@Lucyjc1612

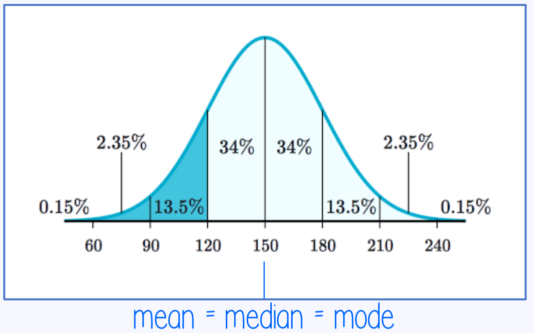
THE NORMAL DISTRIBUTION

KEY WORDS & DEFINITIONS

The Normal Distribution
A continuous probability distribution that can be used to model variables that are more likely to be grouped around a central value than at extremities.

THE NORMAL DISTRIBUTION CURVE

Symmetrically bell-shaped, with asymptotes at each end.
68% percent of data is within one s.d. of μ
95% percent of data is within two s.d. of μ
99.7% percent of data is within three s.d. of μ



mean = median = mode

THE NORMAL DISTRIBUTION TABLE

To find z-values that correspond to given probabilities, i.e. $P(Z > z) = p$ use this table:

p	z	p	z
0.5000	0.0000	0.0500	1.6449
0.4000	0.2533	0.0250	1.9600
0.3000	0.5244	0.0100	2.3263
0.2000	0.8416	0.0050	2.5758
0.1500	1.0364	0.0010	3.0902
0.1000	1.2816	0.0005	3.2905

CALCULATORS FOR NORMAL DISTRIBUTION

Casio fx-991EX:
Menu 7 – Normal PD, Normal CD or Inverse Normal

Casio CG50:
Menu 2 - F5 Dist – F1 Normal – Npd, Ncd or InvN

Choose extremely large or small values for upper or lower limits as appropriate

WHAT DO I NEED TO KNOW

1. The area under a continuous probability distribution curve = 1

2. If X is a normally distributed random variable, with population mean, μ, and population variance, σ^2 we say $X \sim N(\mu, \sigma^2)$

3. To find an unknown value that is a limit for a given probability value, use the inverse normal distribution function on the calculator.

4. The notation of the standard normal variable Z is $Z \sim N(0, 1^2)$

5. The formula to standardise X is $z = \dfrac{x-\mu}{\sigma}$

6. The notation for the probability $P(Z < a)$ is $\phi(a)$

7. To find an unknown mean or standard deviation use coding and the standard normal variable, Z.

8. Conditions for a Binomial distribution to be approximated by a Normal distribution:
n must be large
p must be close to 0.5

9. The mean calculated from an approximated Binomial distribution is $\mu = np$

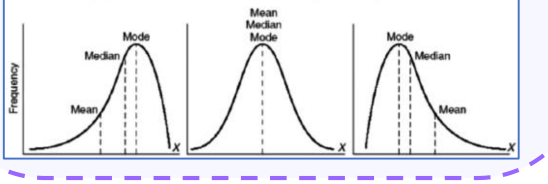
10. The variance calculated from an approximated Binomial distribution is $\sigma^2 = np(1-p)$

11. Apply a continuity correction when calculating probabilities from an approximated Binomial distribution using limits so that the integers are completely included or excluded, as required.

12. The mean of a sample from normally distributed population, is distributed as:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{then} \quad z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

13. Skewed data is NOT 'Normal'



@Lucyjc1612