# Stats

## Contents

# Probability Distributions

## *Binomial*

The binomial distribution is used to model a situation with a fixed number of independent trials each with a constant probability of success.

You can model $X$ as a binomial distribution if:

- There a fixed number of trials, $n$
- Each trial must succeed or fail
- There is a fixed probability of success, $p$
- Each trial is independent

Formula Book

If $X \sim \mathrm{B}(n, p)$, then $\boxed{P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}}$ $(0 \leq x \leq n)$

## *Normal*

The normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ is symmetrical, meaning the mean and median are equal.

When doing questions that involve the normal distribution, sketching the bell curve on the right is always a good idea.

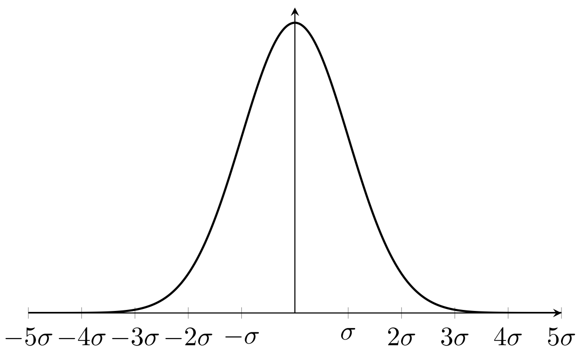
The standard normal distribution $Z \sim \mathrm{N}(0, 1^2)$ is very useful, since it allows you to find values for $\mu$ and $\sigma$ when they're unknown. The normal random variable $X$ can be coded using

Remember

$\boxed{Z = \dfrac{X - \mu}{\sigma}}$

and you can use this equation to find the unknown parameters of $X$.

The 'Normal CD' function on your calculator will calculate the area between an upper and lower bound on the bell curve. The 'Inverse Normal' function will find a value for which the area to the left of that value is the area you specify.



# Hypothesis Testing

Every hypothesis test has two hypotheses:

$$H_0 : \text{The null hypothesis - this is what you assume to be true by default}$$
$$H_1 : \text{The alternative hypothesis}$$

The hypotheses are written in different forms depending on whether the test is one- or two-tailed.

| One-tailed: | Two-tailed: |
|---|---|
| $H_0 : p = k$ | $H_0 : p = k$ |
| $H_1 : p \lesseqgtr k$ | $H_1 : p \neq k$ |

If the question says that someone measured and got a value, then you plug that value into a probability calculation with the parameters from the null hypothesis, with the inequality sign in the same direction as the alternative hypothesis. If the probability of the event happening when assuming $H_0$ is less than the level of significance, then we reject $H_0$ and accept the alternative hypothesis.

A critical value is the smallest or largest value (depending on the direction of the inequality) obtained by a random variable such that $H_0$ would be rejected. Finding a critical value is often best done with the tables in the back of the book.

# Correlation

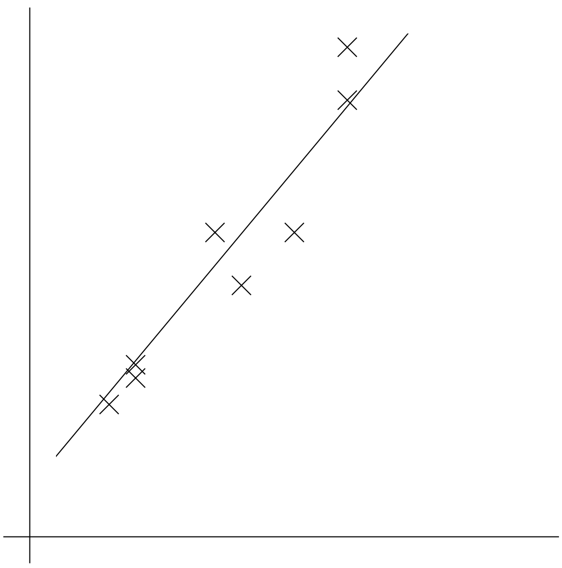


Figure 1: A strong positive correlation

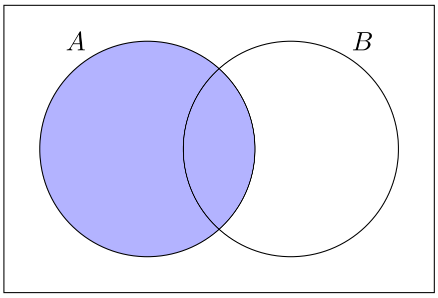
The **product moment correlation coefficient** or "Pearson's correlation coefficient" is a measure of correlation between two variables. It is often called $r$ and is measured in the range $[-1, 1]$. $r = \pm 1$ means the data perfectly follows a positive or negative correlation respectively.

## *Hypothesis testing*

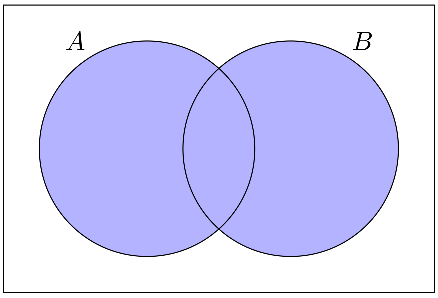
Use $H_0 : \rho = 0$ and $H_1 : \rho \lesseqgtr 0$ or $H_1 : \rho \neq 0$. Get $r$ from your calculator (use section 6 and option 4 regression calc) and $\rho$ from the table in the formula book by finding the sample size and significance level. Remember to halve the significance level if the test is two-tailed.

# Conditional probability

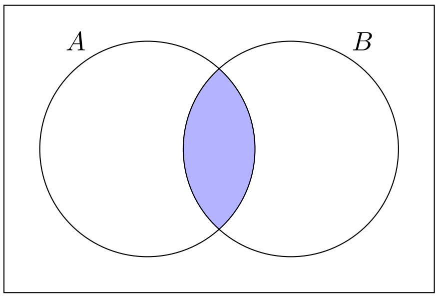
$P(A)$



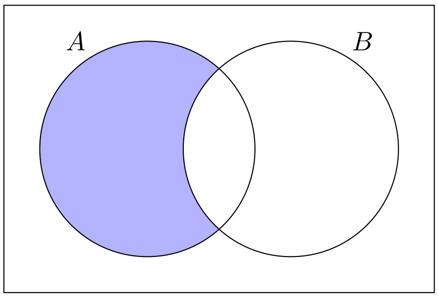

$P(A \cup B)$



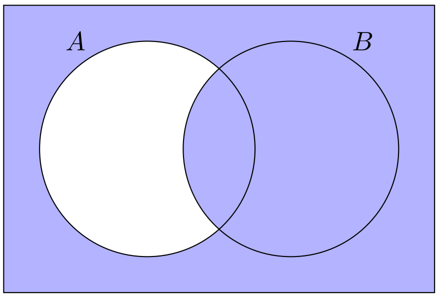

$P(A \cap B)$



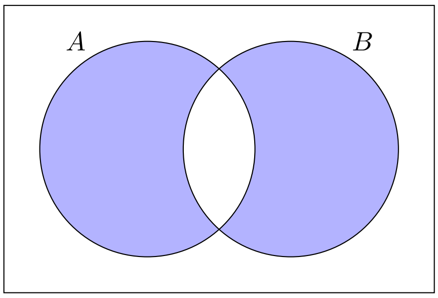

$P(A \cap B')$




$P(A' \cup B)$




$P\big((A \cap B') \cup (B \cap A')\big)$




**Remember**

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

For independent events $A$ and $B$, we know that $P(A \cap B) = P(A) \times P(B)$. This means that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times \cancel{P(B)}}{\cancel{P(B)}} = P(A)$$

# Approximations

## *Binomial with normal*

To approximate a binomial distribution with a normal distribution, use the mean and variance of the of the binomial distribution as the parameters of the normal distribution. To do so, we need to have a large $n$ and a $p$ which is close to 0.5.

To find a probability in this normal approximation, you need to apply a continuity correction. This consists of making the inequalities non-strict (making them $\leq$ or $\geq$) and then extending the range by 0.5 in either direction. Take the following examples:

$$X \sim B(n, p) \qquad\qquad Y \sim N(np, npq)$$

$$P(2 < X \leq 8) = P(3 \leq X \leq 8) = P(2.5 < Y < 8.5)$$

$$P(5 \leq X < 12) = P(5 \leq X \leq 11) = P(4.5 < Y < 11.5)$$

$$P(X = 4) = P(3.5 < Y < 4.5)$$