MATH2859 Concise Notes:

- *Random experiment* — an experiment where the outcome cannot be known before conducting the experiment
- *Outcome* — a result of a random experiment
- *Event* — a *set* of outcomes, ie. a subset of the *sample space*
- *Sample space* — the *universal set* of outcomes
- *Probability* of event E — $P(E)$ is the probability
- *Population* — a set of individuals that we could use to obtain a measure of a *variable*
    - *Sample* — a subset of the population (should be 'representative' of the whole population)
    - *Sampling* — selecting the *sample* for the experiment
        - *Random sampling* — choosing individuals randomly from the population, given that every individual has a known probability of being selected
- *Variables*
    - *Categorical* — variable values can be organised into categories (eg. gender)
    - *Quantitative* — variable values are measured as a number (eg. height)
        - *Location* — the value which most variable values are centered around
            - *Mean* — $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
            - *Median* — value which equally partitions the ordered sample data in two. Also known as the *second quartile* q_2
                - *First quartile* q_1 — value that partitions from 0% to 25%
                - *Second quartile* q_2 — value that partitions from 0% to 50%
                - *Third quartile* q_3 — value that partitions from 0% to 75%
                - *Percentile* — value that partitions from 0% to p%
        - *Variability* — how spread out the variable values are from such center
            - *Variance* — $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{x} - x_i\right)^2$ → 'mean' of sum of squared deviations from the mean value.
            - *Standard deviation* — $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{x} - x_i\right)^2}$ → roughly means 'the size of a typical deviation from the mean'
            - *Interquartile range* — $iqr = q_3 - q_1$. Tends to be more resistant to *outliers* than *variance* and *standard deviation*
                - *Outlier* — any datapoint further than $1.5iqr$ from the closest quartile (as a rough rule)
- Conditional probability — $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Bayes' rule — $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ or $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$
    - $P(Cause \mid Effect) = \frac{P(Effect \mid Cause) \cdot P(Cause)}{P(Effect)}$

- Random variable — a real-valued function: $X : S \to \mathbb{R}$, where S is the sample space
  - $(X = x) = \{\omega \in S : X(\omega) = x\}$ — a set of outcomes ω that satisfy $X = x$
  - $(X \le x) = \{\omega \in S : X(\omega) \le x\}$ — a set of outcomes ω that satisfy $X \le x$
  - Discrete random variable — the random variable can only take on a *countable* number of possible values
  - Continuous random variable — the random variable can take on infinitely many possible values
- Probability distribution — a function of a random variable that gives the probabilities of the occurrence of the possible outcomes for an experiment
  - Cumulative distribution function — $F(x) = P(X \le x)$. If $X \sim F$, we say that "X follows the distribution F"
    - Probability density function — the function f in $F(x) = \int_{-\infty}^{x} f(y)dy$
  - Probability mass function — $p(x) = P(X = x)$, the probability that X takes on the value x
  - Types of distributions:
    - Binomial (discrete): $Bin(n, \pi)$
      - $p(x) = \binom{n}{k}\pi^x(1-\pi)^{n-x}$
      - $E(X) = n\pi$
      - $Var(X) = n\pi(1-\pi)$
    - Poisson (discrete): $Poisson(\lambda)$
      - $p(x) = \frac{(e^{-\lambda}\lambda^x)}{x!}$
      - $E(X) = \lambda$
      - $Var(X) = \lambda$
    - Exponential (continuous): $Exp(\mu)$
      - $f(x) = \frac{1}{\mu}e^{\frac{-x}{\mu}}$
      - $E(X) = \mu$
      - $Var(X) = \mu^2$
    - Normal/Gaussian (continuous): $Norm(\mu, \sigma)$
      - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
      - $E(X) = \mu$
      - $Var(X) = \sigma^2$
    - Student's t-distribution (continuous): t_{n-1}
      - $f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}$, with $\nu > 2$
      - $E(X) = 0$
      - $Var(X) = \frac{\nu}{\nu - 2}$
- Expected value — the weighted sum of all outcomes and their probabilities
  - Discrete — $\mu = E(X) = \sum_{x \in S} xp(x)$

- Continuous — $\mu = E(X) = \int_S x f(x) dx$
- Linear transformation — $E(aX + b) = aE(X) + b$

- Variance — $\sigma^2 = Var(X) = E\left((X - \mu)^2\right)$ or $Var(X) = E(X^2) - \mu^2$ — a measure of the random variable's 'spread' around the expected value
  - Discrete — $Var(X) = \sum_{x \in S} (x - \mu)^2 p(x)$
  - Continuous — $Var(X) = \int_S (x - \mu)^2 f(x) dx$
  - Linear transformation — $Var(aX + b) = a^2 Var(X)$
- Sampling — the selection of a subset of a general population, or a set of observations $X_1, X_2, ..., X_n$
  - Random sample — the random selection of a subpopulation where:
    - Any two observations X_1 and X_2 are independent
    - All random variables have the same probability distribution
- Estimator — an *estimator* of a parameter of interest, θ, is a function of random variables: $\widehat{\Theta} = h(X_1, X_2, ..., X_n)$, where each X_i is a separate observation in a sample
- Estimate — the output of an estimator: $\widehat{\theta} = h(X_1, X_2, ..., X_n)$ which represents an estimate of the true value of θ
- Central Limit Theorem — for any X, regardless of its distribution, we have approximately have $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, where $\overline{X}$ is the sample mean estimator
  - $\frac{\sigma}{\sqrt{n}}$ the standard error of the sample mean
- Confidence Intervals — Eg. a 95% confidence interval for a parameter θ is an interval $[L, U]$ where $P(L \leq \theta \leq U) = 0.95$
- Hypothesis Testing

  Procedure:
  1. State the <u>null</u> and <u>alternative hypothesis</u>
  2. Calculate the <u>test statistic</u> and its <u>null distribution</u>
  3. Calculate the <u>P-value</u> or <u>rejection region</u>
  4. Conclusion
     - Null hypothesis H_0: the statement being tested
     - Alternative hypothesis H_a: describes the type of evidence to look for
     - Test statistic: a calculated statistic that lets us observe the difference with what would be expected *under the null hypothesis*
     - P-value: a calculated measure of how much evidence there is against H_0 *in the direction of the alternative hypothesis* H_a
       - The probability the test statistic would take a value as or more extreme than the value we actually observed, if H_0 were true
       - Small P-value: evidence against H_0
       - Large P-value: data is plausible under H_0 - data doesn't provide evidence against H_0

        - Conclusion: a judgement on the magnitude of the p-value and whether it provides weak/strong evidence against the null hypothesis
- Linear regression — quantifies the relationship between one or more predictor variables and one outcome variable
    - Simple linear regression:
        - $Y = \beta_0 + \beta_1 X + \varepsilon$ — where Y is the random response variable conditional on the predictor X taking value x. Also, $\varepsilon \sim N(0, \sigma)$ with values of ε from different observations are assumed to be independent of each other
        - In regression, we look at how to predict Y at a given value of X, that is, we 'condition on X', and then treat them as fixed values.
        - X does not have to be a random variable. It can take on predetermined values - this is called *fixed design*
        - $Y|(X = x) \sim N(\beta_0 + \beta_1 x, \sigma)$
          σ — tells us how much the values of Y vary around the regression line
    - Least squares — method of fitting a regression line by minimising the squared vertical distance between each datapoint and the regression line

- Additive Law —
- Conditional Probability —
- Bayes’ Rule —
- Law of Total Probability —
-
-
-
-
-
-
-

Week 0:

Set Theory:

$A \subseteq B$ — the occurrence of A implies the occurrence of B
$A \bigcap B$ — events A and B both happen
$A \bigcap B = \oslash$ — A and B are mutually exclusive events
$A \bigcup B$ — one or both or A and B happen

Probability Axioms:

1. $0 \leq P(E) \leq 1$ — for any event E
2. $P(S) = 1$ — one of the possible outcomes is guaranteed to result
3. For any sequence of mutually exclusive events, $E_1,\ E_2,\ ...,\ E_n$:
   $P\left(\sum_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$

Theorems that can be derived from these 3 axioms:

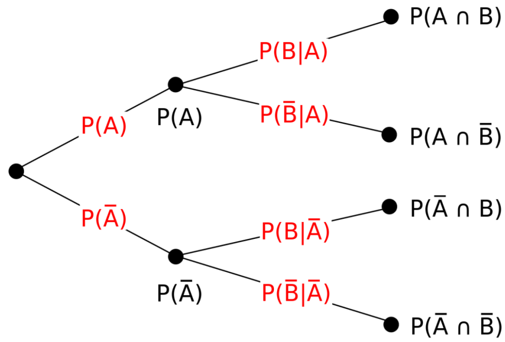
- *Additive law* — $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$

Equally Likely Outcomes:

If all possible outcomes have equal chance of occurring, then $P(E) = \frac{|E|}{|S|}$ where S is the set of all possible outcomes.

Conditional Probability:

$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$ — probability of A occurring, given that B has occurred
$P(A \bigcap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$ — multiplication rule



Theorems:

- If $B \subseteq A$, then $P(A|B) = 1$ — "if B occurred, then A must have occurred"
- *Law of total probability:* $P(A) = \big(P(A|B) \times P(B)\big) + \big(P(A|B^c) \times P(B^c)\big)$
- Bayes' rule: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$ — useful for calculating the converse conditional

Independence:

A and B are independent iff $P(A|B) = P(A)$ and $P(B|A) = P(B)$. That is, whether or not B happens doesn't impact whether or not A happens, *and vice versa*.

- A and B are independent iff $P(A \cap B) = P(A) \times P(B)$ — derived from above
- Generalises for more than two events

Week 1:

Data Analysis:

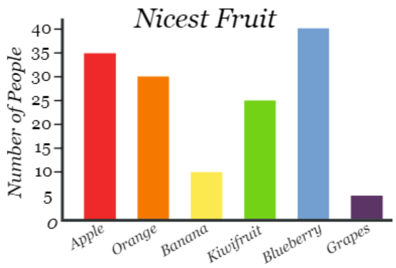
Two key considerations:

1. What are the properties — ie. the variables and whether they are categorical/quantitative
2. What is the question
   a. Is the research question *descriptive* — specific to the current sample
   b. Is the research question *inferential* — trying to generalise the current sample to the population

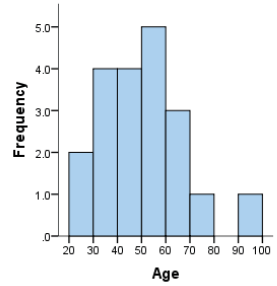
The graphical plot of the data and its interpretation is done with respect to the properties and the question of the experiment.

Graph Representations:

Bar chart — for discrete variables

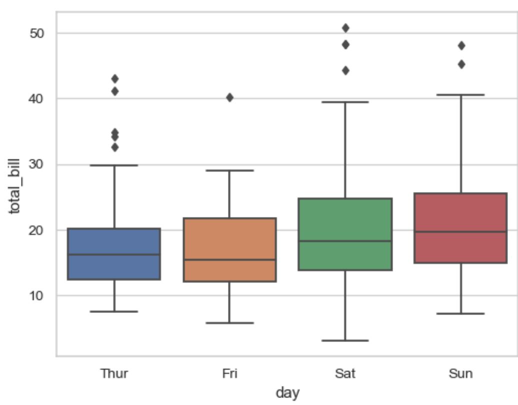


Histogram — frequency distribution of continuous variables

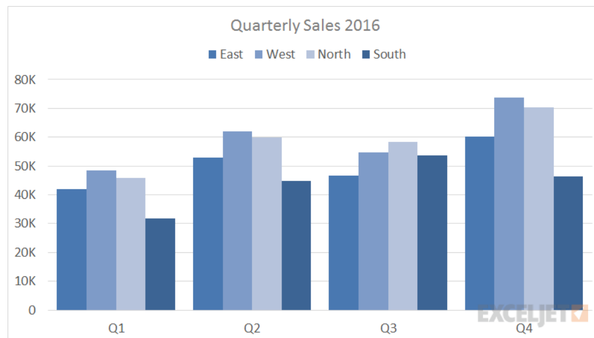


Rough rule for deciding the histogram intervals: $num\ intervals = \sqrt{num\ of\ observations}$

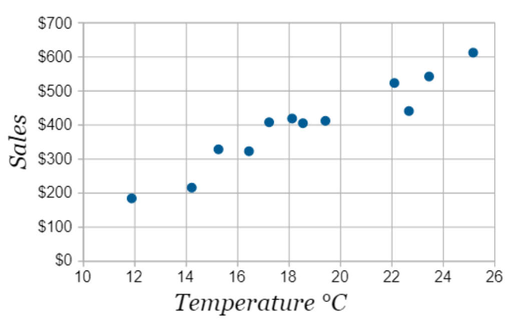
Boxplot — a graph of five-number summaries and outliers



Clustered bar chart

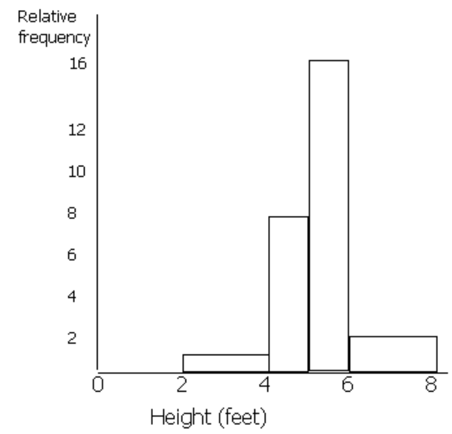


Scatter plot



Density histogram:
The rectangle heights correspond to the densities of each class rather than frequency.

$$relative\ frequency = density \times class\ width$$
$$= height \times width$$



	one variable		two variables		
variable type	**categorical**	**quantitative**	**both categorical**	**one categorical, one quantitative**	**both quantitative**
useful graphs:	bar chart	boxplot or histogram	clustered bar chart	comparative boxplots	scatterplot

Numerical Representation:

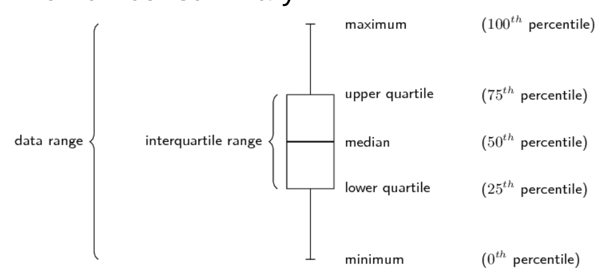
Frequency table:

Boeing plane:	B-707	B-737	B-747	B-757	B-767
Frequency	5	140	35	40	20

Two-way frequency table:

Gender	Smoker	Non-smoker
Female	40	13
Male	26	21

Five-number summary:



Where $interquartile\ range = q_3 - q_1$

Five-number summaries for categories:

	min	q_1	m	q_3	max
Seeded	4	92	221	430	2746
Unseeded	1	24	44	463	1203

Week 2:

Random Variables:

A random variable is a real-valued function: $X : S \to \mathbb{R}$, where S is the sample space

- Typically, capital letters denote random variables and lowercase letters denote observations
- $(X = x) = \{\omega \in S : X(\omega) = x\}$ — a set of outcomes ω that satisfy $X = x$
- $(X \leq x) = \{\omega \in S : X(\omega) \leq x\}$ — a set of outcomes ω that satisfy $X \leq x$

Cumulative distribution function (CDF)

The cumulative distributive function of a random variable X is given as: $F(x) = P(X \leq x)$, for any $x \in \mathbb{R}$. If $X \sim F$, we say that "X follows the distribution F"

- $P(X \in S) = 1$
- $P((X = x_1) \bigcup (X = x_2)) = P(X = x_1) + P(X = x_2)$, if $x_1 \neq x_2$ and x_1 are mutual exclusive
- $P(X < x) = 1 - P(X \geq x)$ — complement
- For all $a \leq b$, we have $P(a < X \leq b) = F(b) - F(a)$
- F is non-decreasing
- $\lim\limits_{x \to \infty} F(x) = 1$ and $\lim\limits_{x \to -\infty} F(x) = 0$

Discrete random variables

A random variable is *discrete* if it can only take on a countable number of possible values.

Probability mass function (PMF):

The probability mass function of a discrete random variable X is given as: $p(x) = P(X = x)$

- $P(X \in A) = \sum\limits_{x \in A} p(x)$
- $P(X \in S) = 1$

Continuous random variables:

A random variable is *continuous* if there exists a non-negative function f such that for any set

$B \subseteq \mathbb{R}$, $P(X \in B) = \int\limits_{B} f(x)dx$

Probability density function (PDF):

$f(x)$ is the probability density function.

- $F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)dy$, or alternatively, $f(x) = F'(x)$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $f(x) \ge 0$ — non-negative function
- $P(a \le x \le b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$
    - $P(X = x) = \int_{x}^{x} f(x)dx = 0$

Expected Value and Variance:

The PMF of a discrete random variable and the PDF of a continuous random variable lets us determine some key values of interest:

- *Expected value* — mean value of the random variable
    - Discrete: $\mu = E(X) = \sum_{x \in S} xp(x)$ — weighted average
    - Continuous: $\mu = E(X) = \int_{S} xf(x)dx$
    - In general, for functions of X:
        - Discrete: $E(g(X)) = \sum_{x \in S} g(x)p(x)$
        - Continuous: $E(g(X)) = \int_{S} g(x)f(x)dx$
    - Linear transformation:
      $E(aX + b) = aE(X) + b$
- *Variance* — $\sigma^2 = Var(X) = E\left((X - \mu)^2\right)$ — a measure of the random variable's 'spread'
    - Alternatively: $Var(X) = E(X^2) - \mu^2$ — sometimes easier to compute
    - Discrete: $Var(X) = \sum_{x \in S} (x - \mu)^2 p(x)$
    - Continuous: $Var(X) = \int_{S} (x - \mu)^2 f(x)dx$
    - Linear transformation:
      $Var(aX + b) = a^2 Var(X)$ — independent of b
- Standardisation — the standardised random variable, Z, is given by $Z = \frac{X - \mu}{\sigma}$
    - Properties of standardised random variable Z:
        - $E(Z) = 0$
        - $Var(Z) = 1$

Covariance:

Good measures of the association between two random variables.

Covariance — $Cov(X, Y) = E\Big((X - E(X)) \cdot (Y - E(Y))\Big)$ — "the product of X and Y's deviation from the mean". A positive covariance indicates that X and Y are positively related and a negative covariance indicates that X and Y are negatively correlated

- Alternatively, $Cov(X, Y) = E(XY) - E(X) \cdot E(Y)$
- Linear transformation: $Cov(aX + b, cY + d) = acCov(X, Y)$
- $Cov(X_1 + X_2, Y_1 + Y_2) = Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2)$

The units for covariance is the unit of X times the unit of Y. Since sometimes that's a garbage unit (like $kg \cdot m$), we can use the dimensionless *correlation coefficient* ρ instead

- $\rho = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$ — always takes values between -1 to 1 indicating the measure of negative/positive correlation

Variance of a sum of random variables

$Var(aX + bY) = Cov(aX + bY, aX + bY)$

$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$ — if X and Y are independent random variables

$Var(X \pm Y) = Var(X) + Var(Y)$ — if X and Y are independent random variables

Week 3:

Distributions:

Random variables can have a particular distribution

Binomial Distribution (Discrete):

- Outcome is eithe   r a success (with probability π) or a failure
- X is the random variable counting the number of successes. We call X a *binomial random variable*

  $X \sim Bin(n, \pi)$, $X = \sum_{i=1}^{n} X_i$ where X_i's are n independent Bernoulli random variables
  - $X_1 + X_2 \sim Bin(n_1 + n_2, \pi)$
  - $E(X_i) = \pi$

    $Var(X_i) = \pi(1 - \pi)$
  - $E(X) = n\pi$

    $Var(X) = n\pi(1 - \pi)$
- $p(x) = \binom{n}{x}\pi^x(1 - \pi)^{n-x}$

Poisson Distribution (Discrete):

Number of occurrences of some random phenomenon in a fixed period of time. The Poisson distribution was derived as a of the Binomial distribution for large n and small π

$X = number\ of\ occurrences$

$X \sim P(\lambda)$, if $p(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$

- $E(X) = \lambda$
  $E(X^2) = \lambda^2 + \lambda$
  $Var(X) = \lambda$

Uniform Distribution (Continuous):

A random variable is uniformly distributed, $X \sim U_{[\alpha,\ \beta]}$, if its pdf is given by $f(x) = \frac{1}{\beta-\alpha}$ for $x \in [\alpha,\ \beta]$

- $E(X) = \frac{\alpha+\beta}{2}$
  $Var(X) = \frac{(\beta-\alpha)^2}{12}$
- $P(a < X < b) = \frac{b-a}{\beta-\alpha}$
- Linear transformations like $Y = c + dX$ also have $Y \sim U_{[c+ad,\ c+bd]}$

Exponential Distribution (Continuous):

A random variable X is an exponential random variable if $X \sim Exp(\mu)$.

PDF: $f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$ for $x \geq 0$

CDF: $F(x) = 1 - e^{-\frac{x}{\mu}}$ for $x \geq 0$

- $E(X) = \mu$
- $Var(X) = \mu^2$

Normal Distribution (Continuous):

Most widely used statistical distribution. Also called Gaussian distribution

If $X \sim N(\mu,\ \sigma)$,

- $E(X) = \mu$
- $Var(X) = \sigma^2$

Standard normal distribution

When $\mu = 0$ and $\sigma = 1$, we have a standard normal distribution. We denote this with the standard normal random variable $Z \sim N(0,\ 1)$

- PMF: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- PDF: $F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$ — no closed form exists. Trapezoidal rule is often used
- $E(Z) = 0$
- $Var(Z) = 1$

Standardisation:

Standardisation is the linear transformation $Z = \frac{X-\mu}{\sigma}$ in:

If $X \sim N(\mu,\ \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,\ 1)$

- 68-95-99 rule — the probability $P(Z = 1) = 0.68$, $P(Z = 2) = 0.95$, $P(Z = 3) = 0.95$. Basically, the value of Z counts how many standard deviations it is from the mean
  - Eg. suppose the average women's height in Australia is normally distributed with $\mu = 163.83cm$ and $\sigma = 6.35cm$. We expect 68% of women to be between $\mu \pm \sigma$ heights, so $157.48cm$ and $170.18cm$
- Normal quantiles — eg. the 95th percentile is given by $P(Z \le z) = 0.95$

Further Properties:

- If $X \sim N(\mu,\ \sigma)$, then $aX + b$ is also normally distributed
- If $X_1 \sim N(\mu_1,\ \sigma_1)$ and $X_2 \sim N(\mu_2,\ \sigma_2)$, then $aX_1 + bX_2$ is also normally distributed, with $aX_1 + bX_2 \sim N(a\mu_1 + b\mu_2,\ \sqrt{a^2\sigma_1{}^2 + b^2\sigma_2{}^2})$
- From the previous two, the linear combination of a random variable/s is also a normal random variable
- The mean of a sum is the the sum of the means
- The variance of a sum is the sum of the variances, IF the random variables are independent

Normal Quantile Plots:

Normal quantile plots (or 'qqplots') are the best way to check how close your data is to a normal distribution.

n observations $\{x_1,\ x_2,\ ...,\ x_n\}$

Cumulative probabilities: $\alpha_i = \frac{i-0.5}{n}$ for i in $[0..n]$

Standard normal quantiles of level α_i: z_{α_i} such that $P(Z \le z_{\alpha_i}) = \alpha_i$ where $Z \sim N(0,\ 1)$

Plot the pairs $(x_i,\ z_{\alpha_i})$ and see if it roughly follows the line $y = x$

Normal Transformations:

If a dataset is not normal, we can transform it so that there is less skew to the right or left. For example, applying the log function to a right skewed dataset will make it more closely resemble a normal distribution

Week 4:

Statistical Inference:

Statistical inference is when we have a subset of a population to experiment on and we want to make an inference about the population as a whole.

Statistical inference focuses on approximating a particular parameter of the population, such as:

- The expected value of some random variable for the whole population
- The proportion of individuals in the population belonging to a certain class
- Difference in expected value between two subsets
- Difference in proportion of individuals in a certain class between two subsets

Random Sample:
Random sampling is very important. If the sample of the population is not random, then statistical methods will not work well.

The set of observations $X_1, X_2, ..., X_n$ is said to be a random sample if:

- X_1 and X_2 are independent random variables
- X_i and X_j have the same probability distribution for all i and j in $[0..n]$

"Independent and identically distributed"

Estimators and Estimates:

Suppose θ is some parameter of interest (eg. sample mean, standard deviation, etc.)

- An *estimator* of θ is a function of random variables: $\widehat{\Theta} = h(X_1, X_2, ..., X_n)$, where each of X_i is a different observation across different samples.
    - $\widehat{\Theta}$ itself is a random variable that takes different values across different samples and has a *sampling distribution* that depends on the distribution of X and the sample size. The sampling distribution tells us how well the estimator $\widehat{\Theta}$ estimates θ
        - The greater the random sample size, the lower the variance, ie. the better the estimate (the less it will vary sample-to-sample)
    - Standard error of Θ — the standard deviation of Θ. Just a clear term because σ is the standard deviation of the population, not the sample
- An *estimate* of θ is a calculation on the sample values: $\widehat{\theta} = h(x_1, x_2, ..., x_n)$

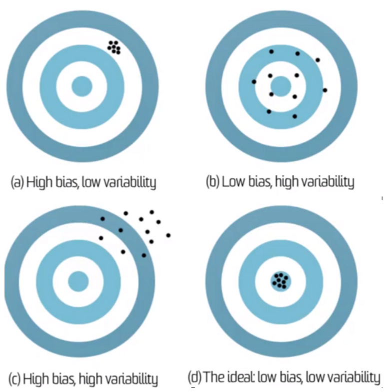
<u>Hat notation:</u>
Estimators $\widehat{\Theta}$ and estimates $\widehat{\theta}$ for the parameter θ have a hat to distinguish it from the true parameter value of θ in the whole population - since estimators and estimates are only on a subset of the population.

- Sample mean is denoted with a bar: $\overline{X}$ as the estimator, $\overline{x}$ as the estimate
  - It can be determined that $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- Standard deviation is denoted as: S as the estimator, and s as the estimate

Picking the estimator function:
Previously, an estimator for the population mean was just the sample mean. We could instead have:
- $\widehat{\Theta}_1 = X_1$ — the first observed value
- $\widehat{\Theta}_2 = \frac{X_1 + X_n}{2}$ — the average of the first and last value
- $\widehat{\Theta}_3 = \frac{2X_1 + X_n}{2}$ — this one is biased

Criteria:
- We want *Unbiased* estimators Θ, where $E(\Theta)$ centred around the true value of θ.
- We want lower variance.
- We want the estimator to be *consistent*, ie. bigger samples should give better estimations
  - RTP $Var(\Theta) \to 0$ as $n \to \infty$. For the $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$, the variance does not tend to 0 as n gets larger

Ultimately, we'll still pick the default sample mean estimator $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$



Central Limit Theorem:

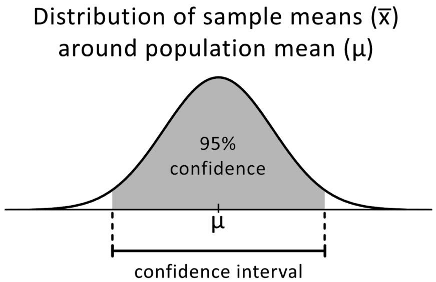
For any X, regardless of its distribution, we have $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, approximately.
We call $\frac{\sigma}{\sqrt{n}}$ the standard error of the sample mean

Week 5:

Confidence Intervals:

A 95% confidence interval for a parameter θ is an interval $[L,\ U]$ where $P(L \le \theta \le U) = 0.95$.

- Eg. A 95% confidence interval for μ means that if the sample mean is normally distributed, then 95% of the time, the sample mean is within 1.96 standard deviations from the true mean
- An approximate confidence interval of $100 \times (1-\alpha)\%$ for μ is given by $[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$, provided that $\bar{X}$ is approximately normal and σ is known
    - $z_\lambda = norminv(\lambda)$
    - Note how smaller standard deviation σ and larger sample size n shrinks the confidence interval



Margin of error:
Suppose we want to be 95% sure that our random sample mean will be 15 minutes within the true population μ, and we know σ. How large should the sample size n be?
Just solve $e = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ for n.

Student's T-Distribution:

Estimating Standard Deviation in a Random Sample:
Since we often do not know σ, the true standard deviation in the population, we usually estimate it based on just the sample. $\sigma \approx S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$

But with this, $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim N(0,\ 1)$ is no longer a good approximation of the standard normal for smaller sample sizes.
Instead, $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$ — the exact distribution is a T-distribution with $n-1$ degrees of freedom

Student's T-distribution:
If $X \sim t_v$, with v degrees of freedom, we have

- $E(X) = 0$
- $Var(X) = \frac{v}{v-2}$ for $v > 2$

As $v \to \infty$, the t-distribution tends towards becoming a standard normal distribution.

T-confidence interval:
Just like how we get the z-score at the 97.5th percentile, we use $tinv$ in matlab to get the t-score instead.

$$\left[\overline{X} - t_{n-1,\,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}},\ \overline{X} + t_{n-1,\,1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$$

Confidence Interval for Proportions:

The estimator of proportion π of a population belonging to a class of interest is $\widehat{P} = \frac{Y}{n}$ where Y is a random variable counting the number of occurrences and $Y \sim Bin(n, \pi)$.

- $\widehat{P} = \frac{Y}{n}$ is unbiased since $E(\widehat{P}) = \frac{1}{n}E(Y) = \pi$
- $Var(\widehat{P}) = \frac{1}{n^2}Var(Y) = \frac{\pi(1-\pi)}{n}$

Sampling distribution of $\hat{P}$

The sample proportion $\hat{P}$ can be written as a sample mean, $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are independent Bernoulli(π) variables.

So the Central Limit Theorem applies to $\hat{P}$:

$$\boxed{\frac{\hat{P} - \pi}{\sqrt{\pi(1-\pi)/n}} \overset{a}{\sim} \mathcal{N}(0,1)}$$

if n is 'large' ($\overset{a}{\sim}$ again stands for "approximately follows")

It turns out that we can also say that

$$\frac{\hat{P} - \pi}{\sqrt{\hat{P}(1-\hat{P})/n}} \overset{a}{\sim} \mathcal{N}(0,1)$$

(this result is more easily used to construct a confidence interval for π.)

CLT works better when n is larger and when the X_i are more symmetric
$\to \pi$ should not be too close to 0 or 1
$\to$ empirical rule: $n\hat{p}(1-\hat{p}) > 5$

Confidence Interval:

Deriving a confidence interval for π

Step 1 – a range of values for Z.

$$\mathbb{P}(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

where $Z \sim \mathcal{N}(0,1)$ and $z_{1-\alpha/2}$ is its quantile at level $1 - \alpha/2$

Step 2 – apply the CLT result

$$\mathbb{P}\left(-z_{1-\alpha/2} \leq \frac{\hat{P} - \pi}{\sqrt{\hat{P}(1-\hat{P})/n}} \leq z_{1-\alpha/2}\right) \simeq 1 - \alpha$$

Step 3 – solve for π

$$\mathbb{P}\left(\hat{P} - z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \leq \pi \leq \hat{P} + z_{1-\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \simeq 1 - \alpha$$

so an approximate $100(1-\alpha)\%$ confidence interval for π is

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

If roughly $n\widehat{p}(1-\widehat{p}) > 5$, then $\widehat{p}$ is an observation from an approximately normal distribution.

$$\left[\widehat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right]$$

Assumptions:

- Empirical rule: $n\widehat{p}(1-\widehat{p}) > 5$
- Independent observations with the same probability π for each observation

Week 7:

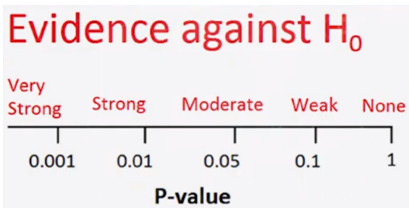
Hypothesis Testing:

Procedure:

1. State the <u>null</u> and <u>alternative hypothesis</u>
2. Calculate the <u>test statistic</u> and its <u>null distribution</u>
3. Calculate the <u>P-value</u> or <u>rejection region</u>
4. Conclusion

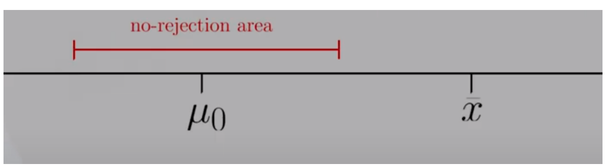
- Null hypothesis H_0: the statement being tested
- Alternative hypothesis H_a: describes the type of evidence to look for
- Test statistic: a calculated statistic that lets us observe the difference with what would be expected *under the null hypothesis*
- P-value: a calculated measure of how much evidence there is against H_0, *in the direction of the alternative hypothesis* H_a
  - The probability the test statistic would take a value as or more extreme than the value we actually observed, if H_0 were true

- Small P-value: evidence against H_0
    - Large P-value: data is plausible under H_0 - data doesn't provide evidence against H_0
- Conclusion:
  "There is ___ evidence against H_0"
  "There is ___ evidence that "



Rejection region
Significance level α is used to define the rejection region.
Eg. if $\alpha = 0.05$, then if we get a test statistic that lies out of norminv(0.05), we will reject the null hypothesis
One-sided alternative hypotheses:
- If $H_a:\ \mu < \mu_0$ — rejection region: $\bar{x} < \mu_0 - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$
- If $H_a:\ \mu > \mu_0$ — rejection region: $\bar{x} > \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$

Two-sided alternative hypothesis:
- If $H_a:\ \mu \neq \mu_0$ — rejection region: $\bar{x} \in [\mu_0 - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}},\ \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$



If $[l,\ u]$ is a $100 \times (1-\alpha)\%$ confidence interval for a parameter θ, then the hypothesis test for $H_0:\ \theta = \theta_0$ against $H_a:\ \theta \neq \theta_0$
Will reject H_0 at significance level α iff θ_0 is not in $[l,\ u]$

Hypothesis tests and confidence intervals are more or less equivalent
- Confidence intervals give a range of likely values for θ
- Hypothesis tests display the risk levels, such as p-values associated with a specific decision

Errors:
- Type I error: we reject the null hypothesis when it is true — $P(reject\ H_0|H_0\ is\ true) = \alpha$
    - The probability of a type I error is equal to the significance level α
- Type II error: we don't reject the null hypothesis when it is false —
  $P(not\ rejecting\ H_0|H_0\ is\ false) = \beta$

- $1 - \beta = P(rejecting\ H_0\ when\ it\ is\ false)$ is called the power of the test
  β depends on the true value of the parameter

Determining power of the test:

Power is the probability that we calculate a value within the rejection region
Determine the rejection region

Week 8:

Mean Difference:

If:

1. We have independent random samples of variables X_1 and X_2
2. X_1 and X_2 are normally distributed (although it works well even if they are not - as long as $\overline{X_1}$ and $\overline{X_2}$ are normally distributed)
3. The standard deviations of X_1 and X_2 are equal ($\sigma_1 = \sigma_2$)

Then $T = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ with $S_p = \sqrt{\frac{(n_1-1)S_1^{\ 2} + (n_2-1)S_2^{\ 2}}{n_1+n_2-2}}$

For the hypothesis: $H_0:\ \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$, we can use the test statistic: $t = \frac{\overline{X_1} - \overline{X_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ (if we can make the 3 assumptions above)

Two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\left[(\overline{x_1} - \overline{x_2}) - t_{1-\frac{\alpha}{2},\, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\overline{x_1} - \overline{x_2}) + t_{1-\frac{\alpha}{2},\, n_1+n_2-2} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

Paired Data (Paired T-Test):

When sets of data occur in pairs, we cannot apply the above procedure (since $\overline{X_1}$ and $\overline{X_2}$ are *not independent*)

The test statistic is $t = \frac{\overline{D}}{S_d/\sqrt{n}}$ where $\overline{D} = \overline{X_1} - \overline{X_2}$.

t comes from $T \sim t_{n-1}$ if:

1. We have a random sample of differences
2. The differences are normally distributed

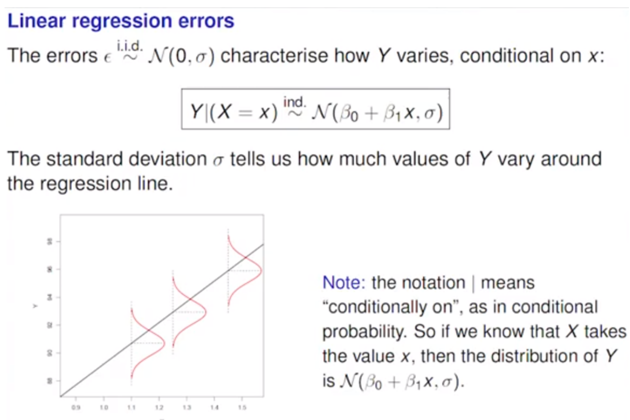
Week 9:

Linear Regression:

Simple linear regression:
$Y = \beta_0 + \beta_1 X + \varepsilon$ — where Y is the random response variable conditional on the predictor X taking value x. Also, $\varepsilon \sim N(0,\ \sigma)$ with values of ε from different observations are assumed to be independent of each other

- In regression, we look at how to predict Y at a given value of X, that is, we ‘condition on X’, and then treat them as fixed values.
- X does not have to be a random variable. It can take on predetermined values - this is called *fixed design*

$Y|(X = x) \sim N(\beta_0 + \beta_1 x,\ \sigma)$
σ — tells us how much the values of Y vary around the regression line



Least Squares:

Least squares regression is a method of fitting a regression line by minimising the squared vertical distance between each datapoint and the regression line.

Least squares estimators of β_0 and β_1:

$\widehat{\beta_0} = \frac{S_{XY}}{S_{XX}}$

$\widehat{\beta_1} = \overline{Y} - \frac{S_{XY}}{S_{XX}}\overline{X}$

$S_{XX} = \sum_{i-1}^{n} \left(X_i - \overline{X}\right)^2$

$S_{XY} = \sum_{i-1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$

Regression Assumptions:

$Y_i | (X_i = x_i) \overset{\text{ind.}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$ for $i = 1, 2, \ldots, n,$

1. Linearity: the conditional mean is a *linear* function of x

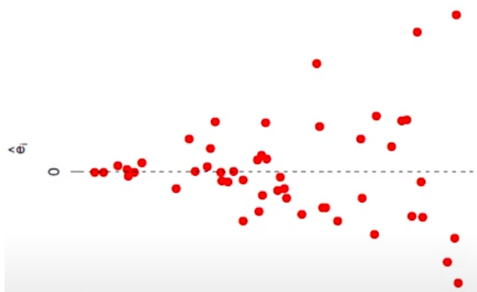
Assumptions on the error terms: $e_i = y_i - (\beta_0 + \beta_1 x_i)$ for $i = 1..n$

2. Each e_i has been drawn independently of each other
3. Each e_i has the same variance
4. Each e_i has been drawn from a normal distribution. Ie. errors need to have a normal distribution
    - CLT gives robustness to violations of normality, so it tends to not matter too much if the errors are not 'that' normally distributed

Checking Regression Assumptions:

We don't know the true e_i terms as we don't know β_0 and β_1, but we can estimate them using the observed *residuals* from the fitted model: $\widehat{e_i} = y_i - \widehat{y}(x_i) = y_i - (\widehat{b_0} + \widehat{b_1} x_i)$

- Residuals vs fits plot to check assumptions 1 and 3 - we want them to look like random noise, no pattern
    - Eg. same variance is violated here:



- Normal quantile plot of residuals to check assumption 4

Standard error of the slope

Sampling distribution $\widehat{\beta_1} \sim N\left(\beta_1, \frac{\sigma}{\sqrt{S_{XX}}}\right)$

So the standard error of $\widehat{\beta_1}$ is $\frac{\sigma}{\sqrt{S_{XX}}}$

Where σ can be estimated from the estimator: $s = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \widehat{e}_i^2}$, with $n-2$ degrees of freedom since we needed to first estimate two parameters (slope and intercept) before we can estimate σ (previously we used $n-1$ degrees of freedom because we only needed to estimate one parameter μ in order to estimate σ)

- Degrees of freedom = $n - (\#\ \textit{parameters needed to get mean})$

Testing hypotheses about the slope

$H_0 : \beta_1 = 0$ — if X and Y are independent of each other

$H_a : \beta_1 \neq 0$
We care about testing this hypothesis so that we can say whether X and Y are related (by testing against the hypothesis that the true slope β_1 is 0)

$\frac{\widehat{\beta}_1-\beta_1}{\frac{\sigma}{\sqrt{s_{XX}}}} \sim N(0,\ 1)$
$\frac{\widehat{\beta}_1-\beta_1}{\frac{S}{\sqrt{s_{XX}}}} \sim t_{n-2}$

Confidence interval for β_1:
As $\frac{\widehat{\beta}_1-\beta_1}{\frac{S}{\sqrt{s_{XX}}}} \sim t_{n-2}$, we can write,
$P(-t_{n-2,\ 1-\frac{\alpha}{2}} \leq \frac{\widehat{\beta}_1-\beta_1}{S/\sqrt{s_{XX}}} \leq t_{n-2,\ 1-\frac{\alpha}{2}}) = 1-\alpha$
$\left[\widehat{b_1} - t_{n-2,\ 1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{s_{XX}}},\ \widehat{b_1} + t_{n-2,\ 1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{s_{XX}}}\right]$ — confidence interval for the true slope β_1

Inference about the Intercept:
$Var(\widehat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{s_{XX}}\right)$

Predicting new datapoints:
For X taking some value, what is the expected response Y and what is the confidence interval?
Estimated mean for when $X = x$:

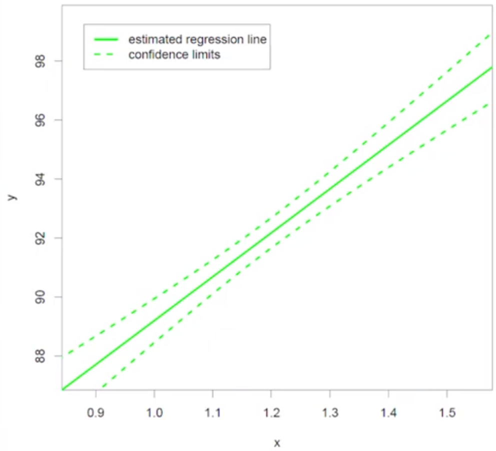
$$\hat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\mu_{Y|X=x}, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right)$$

For predicting new datapoints:

$$\hat{y}(x_0) \pm s t_{n-2,1-\alpha/2}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}}$$

$\widehat{\mu}_{Y|X=1} = \widehat{b}_0 + \widehat{b}_1 x$
Confidence interval of the true mean for $X = x$ is given by:

$$\left[\hat{y}(x) - t_{n-2;1-\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \hat{y}(x) + t_{n-2;1-\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right]$$



Variability Decomposition:
The total amount of variability in response values can be measured by

- Total sum of squares: $ss_t = s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

The variability in observed values y_i arises from two factors:

1. Because the x_i values are different, all Y_i have different means. This variability is quantified by the *regression sum of squares*: $ss_r = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2$
2. Each value Y_i has variance σ^2 around its mean. This variability is quantified by the *error sum of squares*: $ss_e = \sum_{i=1}^{n} (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{n} \hat{e}_i^2$

$ss_t = ss_r + ss_e$
Correlation coefficient r has equation: $r^2 = \frac{ss_r}{ss_t}$ (r^2 is also known as the proportion of variability in the response, explained by the predictor)

- A value close to 1 is good, a value close to 0 is bad TODO

Week 10:

ANOVA Test:

ANOVA — analysis of variance
Basically a generalisation of a 2-sample t-test to more than 2 samples.

Eg. four groups of students had a different teaching technique applied to them

We need independent random samples. Example way of satisfying this: get each student to randomly draw a label out of a hat to assign them to a particular teaching method

To make inferences from an experiment about the effect of a factor, we need:

- The factor of interest to be the only thing that varies across samples (eg. teaching technique)
- The treatment to be applied independently to each subject (not once in bulk)

In ANOVA testing, we have k different groups to compare. Each group is often called a *treatment*
The key to ANOVA testing is to compare the between-group variance with the within-group variance.
A low in-group variance and high between-group variance indicates that a difference was caused by particular treatments

ANOVA Model:
$X_{ij} = \mu_i + \varepsilon_{ij}$

- μ_i is the true mean response for the ith treatment
- ε_{ij} is an individual random error component
    - As usual for errors, we assume $\varepsilon_{ij} \sim N(0, \sigma)$. This is the central assumption of the ANOVA model

For each treatment: $X_{ij} \sim N(\mu_i, \sigma)$

ANOVA Hypotheses:
$H_0 : \mu_1 = \mu_2 = ... = \mu_k$
$H_a : not\ all\ means\ are\ equal$ — ie. that at least two of the means differ, not necessarily all of them

Variability decomposition

- Total sum of squares: $ss_{Tot} = \sum_{i=1}^{k}\left(\sum_{j=1}^{n_j}\left(X_{ij} - \overline{\overline{X}}\right)^2\right)$
- Treatment sum of squares: $ss_{Tr} = \sum_{i=1}^{k}\left(n_i\left(\overline{X}_i - \overline{\overline{X}}\right)^2\right)$
- Error sum of squares: $ss_{Er} = \sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}\left(X_{ij} - \overline{X}_i\right)^2\right)$
- Sum of squares identity: $ss_{Tot} = ss_{Tr} + ss_{Er}$

Mean Squares:

- Mean Squared Error: (Within-group variance)
  $MS_{Er} = \frac{SS_{Er}}{n-k}$ — pooled estimator of error variance σ^2 of the ANOVA model
- Treatment mean square: (Between-group variance)
  $MS_{Tr} = \frac{SS_{Tr}}{k-1}$ — estimator of the error variance σ^2 *if the null hypothesis is true*
- If H_0 were true, these two mean squares would be similar.
  If H_0 were false, we expect $MS_{Tr} > MN_{Er}$

Fisher's Distribution:

It H_0 is true, then the ratio of treatment mean square to mean squared error $F = \frac{MS_{Tr}}{MS_{Er}} = \frac{\frac{SS_{Tr}}{k-1}}{\frac{SS_{Er}}{n-k}}$ has distribution $F \sim F_{k-1,\, n-k}$ with $k-1$ and $n-k$ degrees of freedom

To test the hypothesis $H_0 :\ \mu_1 = \mu_2 = ... = \mu_k$ against $H_a :\ \textit{not all means are equal}$, we can use the test statistic: $f = \frac{MS_{Tr}}{MS_{Er}}$ where $f \sim F_{k-1,\, n-k}$ IF:

1. $X_1,\ X_2,\ ...,\ X_k$ are independent random samples
2. The variables are normally distributed
3. The standard deviations of $X_1,\ X_2,\ ...,\ X_k$ are all equal

Note that MS_{Er} is a good estimate of error variance, irrespective of whether the null hypothesis is true or not

To calculate the p-value, $p = P(X > f)$, where X has the F distribution.
A rejection region for the test statistic is: reject H_0 if $\frac{ms_{Tr}}{ms_{Er}} > f_{k-1,\, n-k;\, 1-\alpha}$

ANOVA Model Assumptions:

- Random variables: $\varepsilon_{ij} = X_{ij} - \mu_i$
  Independent AND normally distributed: $\varepsilon_{ij} \sim N(0,\ \sigma)$
  With the same variance in each group

We don't have access to values for ε_{ij}, since μ_i s are unknown, but we can estimate them using the *residuals*: $\widehat{e_{ij}} = x_{ij} - \overline{x_i}$ - do a normal quantile plot of residuals

Determining differences between treatment groups:

Treatment mean CI: $\left[\overline{x_i} - t_{n-k,\, 1-\alpha/2}\sqrt{\frac{MS_{Er}}{n_i}},\ \overline{x_i} + t_{n-k,\, 1-\alpha/2}\sqrt{\frac{MS_{Er}}{n_i}}\right]$

Bonferroni-adjusted t-test:

To compare treatment i with treatment j, the null hypothesis is $H_0: \mu_i = \mu_j$ and $H_a: \mu_i \neq \mu_j$

Test statistic: $t_0 = \frac{\overline{x}_i - \overline{x}_j}{\sqrt{MS_{Er}(\frac{1}{n_i} + \frac{1}{n_j})}}$

P-value: $p = 2P(T > |t_0|)$ where $T \sim t_{n-k}$

Reject H_0 if p-value is less than $\frac{\alpha}{K}$, where $K = comb(k, 2)$