**CFA 1 Quantitative Methods**

**1.1 Rates and Returns**

**Interest rates**

3 interpretations of interest rates

* Required rates of return – min rate of return to accept the investment
* Discount rates – rate used to calculate the PV
* Opportunity cost

Formula for interest rate :

Real risk free rate: Return for no risk in an inflation free environment

Inflation premium: Compensation for expected inflation in a risk free environment

Default risk premium: Compensation for possible failure to pay

Liquidity premium: Compensation for difficulty to convert to cash quickly

Maturity premium: Compensation for increased uncertainty of longer dated debt

Nominal risk free rate is given by:

To be accurate:

**Holding period return**

Return generated over the period you held the asset

Where is the income generated over the period

**Arithmetic mean**

Just the average:

**Geometric mean AKA CAGR**

Includes the compounding effect:

Always less than or equal to the arithmetic mean return

* Arithmetic is biased upwards compared to the geometric

**Harmonic mean**

A weighted mean where an observation’s weight is inversely proportional to its magnitude

You sum up the reciprocals, take the average by dividing by (number of observations), then you take the reciprocal of the whole thing by flipping it over

Appropriate when the variable is a rate or ratio (e.g., P/Es)

Appropriate for averaging ratios (amount per unit), where the ratios are repeatedly applied to a fixed quantity to yield a variable number of units

* Can be used for cost averaging, e.g., finding the average entry price

Less affected by extreme values than the arithmetic mean

* Useful for dealing with outliers

Harmonic mean doesn’t work for negative numbers, so need to convert to format

Relationship between the three:

**Dealing with outliers**

Trimmed mean: Removes a small defined % of the largest and smallest values before calculating the mean

Winsorized mean: Replace the extreme values at both ends with values of their nearest observations

**Which mean to use:**

To include outliers: Arithmetic

Compounding: Geometric

Extreme outliers: Harmonic, Trimmed, Winsorized

**Money weighted return**

Essentially the same as the IRR (rate that makes NPV=0):

This only considers cash inflows and outflows, it doesn’t consider investment gains/losses

When the money is deposited into the account, that occurs at (cash outflow)

Assume the investment is sold at the end (cash inflow)

Can be appropriate if the investor manager exercises control over additions and withdrawals to the portfolio (e.g., if they tell the investor to invest more)

* If it is the investor who is in control, it can be a bit unfair

**Time weighted returns**

Time weighted returns are not sensitive to the additions and withdrawals of funds

GIPS Preferred measure for portfolios as it neutralises the effect of cash withdrawals or additions to the portfolio

* If a client adds funds to invest at an unfavourable time, the money-weighted rate of return will tend to be depressed
* If a client adds funds at a favourable time, the money-weighted return will tend to be elevated
* Time weighted cleans all this out

Steps for calculation:

1. Price the portfolio prior to any significant additions or withdrawals of funds
2. Break the period into sub-periods based on cash inflow and outflow dates
3. Calculate the Holding period return for each sub-period
4. Compound the period returns for an annual rate for the year, and find the geometric mean of annual returns for multiple years

**Annualised return**

Investments may pay interest more than once a year

* However, you may only be given the quoted annual rate (AKA Stated annual interest rate, Nominal rate)

Annualising returns makes comparison easier

* Can convert daily, weekly, monthly, etc into annual

is the number of compounding periods per year

is the quoted annual interest rate

is the number of years

To annualise something shorter than one year, put it to the power of a proportion of one year inversed

* A week need needs to be compounded 52 times, a month 12 times, etc

Example: The weekly return in 0.2%. That means the compound annual return is 10.95%

To annualise something longer than 1 year, just put it to the power of a proportion of one year inversed (same thing)

Example: 18 month return of 20% can be annualised as:

**Continuously compounded returns**

Continuously compounded return from to :

From 0 to :

**Return measures**

Gross return: Return on assets minus any trading expenses and commissions

* Reflects the skill of manager
* Management fees, taxes, etc are not included

Net return: Return earned for the investor

* Includes management and admin fees
* Smaller funds at a disadvantage compared to larger funds due to fixed admin costs

Gross and Net returns are both pre-tax nominal returns

* No adjustments for taxes or inflation

After-tax nominal return: Total return minus taxes

* Bonds issued at a discount to par may be taxed based on accrued gains instead of realised

After tax real return

* Essentially what the investor receives for postponing consumption and assuming risk after paying taxes
* Useful for comparison

**Real returns (use these for calculations, not the addition ones)**

Nominal risk free rate:

Real return:

**Leveraged return**

Two ways of creating a claim on asset returns that are greater than the investment of one’s own money

1. Trade future contracts where the money to take a position is a fraction of the notional value of the asset
2. Borrowing money to purchase the asset

The size of the leveraged position increases by the additional borrowed capital

is the unleveraged return of the portfolio

is portfolio equity

is debt (borrowed funds)

is the borrowing cost on debt

If , then leverage decreases

**1.2 The Value of Money in Finance**

**Time value of money**

Future value is given by:

If the number of compounding periods tends to infinity:

**1) Discount**

Discount: Pay initial price and receive a single principal cash flow at maturity

* represents the interest earned

**2) Periodic**

Periodic interest: Pay initial price and receive interest cash flows at pre-determined intervals, and a final interest payment and principal at maturity

If the coupon payments are semi-annual, divide the yearly discount rate by 2 and the yearly coupon by 2

Perpetual bond: No stated maturity date

**3) Annuity**

Level payments: Pay initial price and receive uniform cash flows at predetermined intervals through the maturity which presents both interest and principal repayment

* E.g., Mortgages

The periodic annuity cash flow occurs at the end of each period is given by:

is the periodic cash flow

is the market interest rate

is the present value or principal amount of bond

is the number of payment periods

When doing calculations make sure to check if it is monthly!!

Can also work it out with the standard Texas formula

* Just need to make FV = 0 as loan will be entirely paid off

**Equity instruments**

Equity investments have no maturity date and are assumed to remain outstanding indefinitely

* Can assume dividends are constant, growing steadily, or changing

here is required rate of return

1) Equity with constant dividends (Preferred stock):

* Pays a fixed dividend as a stated % of par value (like a bond)

2) Equity with constant dividend growth:

* Where dividends grow at rate of per period and are paid at the end of each period

Where

3) Equity with changing dividend growth (e.g., 2 stage, first stage , second stage )

**Implied return for fixed income**

Having the current price and future cash flows means you can calculate the implied returns

For a discount bond (one payment at the end), its implied return is given by:

For a coupon bond (use Texas):

**Implied return for equity**

When calculating :

When calculating :

**Implied P/E ratio**

Implied forward price to earnings ratio is given by:

is the dividend payout ratio

If you don’t want the forward P/E, switch the to

**Cash flow additivity**

Cash flow additivity: The PV of any stream of cash flows equals the sum of the PVs of the cash flows

Illustrates the no arbitrage principle

Forward rates are set so investors cannot earn riskless arbitrage profits

Works with the FX market

* Any combination of spot and forward rates should have the same cost

**Forward exchange rates**

% difference between forward and spot rates is approx. the difference between two countries interest rates

**1.3 Statistical measures of asset returns**

**Measures of central tendency**

Shows where data is centred

1) Arithmetic mean: Most commonly used

2) Median: position

3) Mode (can be unimodal, bimodal, trimodal – if more than one value occurs most frequently, or modal interval for continuous data)

Options to deal with outliers

1. Transform the dataset
2. Do nothing
3. Delete outliers – e.g., trimmed mean (delete % of extremes)
4. Replace outliers – e.g., winsorized mean (replace % of extremes)

For box and whisker plots, the fences don’t have to be the range, but can be 1.5x the interquartile range

**Measures of dispersion**

Absolute dispersion shows the variability without comparison to any reference point

* Range, Mean Absolute Deviations, Standard Deviations

1) Mean absolute deviations is:

2) Sample variance is:

We divide by as after you calculate the sample mean, there are only independent pieces of information left ( would be biased estimator)

3) Standard deviation is the square root of the variance

* Unlike sample variance, it is expressed in the same unit as the data itself (easier to interpret)

**Coefficient of variation**

Relative dispersion should be used for meaningful comparisons

* Amount of variability around a benchmark

Coefficient of variation is the ratio of the SD to the sample mean:

If the observations are returns, CV gives the risk per unit of reward

**Downside deviation**

Looks at the downside risk, the other measures look at upside and downside

Target semideviation measures the dispersion of observations below a target:

Where is the target

**Shapes of distribution**

Normal distribution

* Has the same mean, median, and mode
* Is completely described by 2 parameters – mean and variance

**Skewness**

Skewness – distributions that aren’t symmetrical

For a positively skewed distribution: Mean > Median > Mode

* Skewed to right

For a negatively skewed distribution: Mean < Median < Mode

* Skewed to left

When is large, sample skewness is approximated by (positive if positive, negative if negative):

**Kurtosis**

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution

E.g., The proportion of the total probability that is outside of 2.5 standard deviations of the mean

Leptokurtic: Fat-tailed, kurtosis > 3

Platykurtic: Thin-tailed, kurtosis < 3

Mesokurtic: Tails similar to normal distribution, kurtosis = 3

* A normal distribution has a kurtosis of 3

Excess kurtosis is the kurtosis relative to the normal distribution

Sample excess kurtosis is given by:

**Covariance**

Covariance shows how two variables in a sample move together:

It is a measure of joint variability of two variables

* If they vary in the same direction, then covariance is positive

**Correlation**

The sample correlation coefficient normalises the covariance (covariance hard to interpret)

It is given by the ratio of the sample covariance to the product of the two variable’s standard deviations:

It shows the strength of the linear relationship between 2 variables

Correlation is always between -1 and 1

* 0 means they are uncorrelated

**Limits of correlation analysis**

Two variables can have a strong nonlinear relation and still have a very low correlation

Correlation may also be sensitive to outliers

Spurious correlation:

* Can be due to chance
* Can be induced if both variables are mixed with a third variable
* Can arise from not direct relation, but from their relation to a third variable

**1.4 Probability trees and conditional expectations**

**Expected value and variance**

Expected value is for the population, the sample mean is for the sample

* The expected value is our forecast

Variance: The expected value of squared deviations from the expected value

Variance is in the squared units of

Standard deviation is in the same units as the random variable

**Probability trees and conditional expectations**

Conditional expected values: Contingent on outcome of another event

The expectation of given scenario :

Bayes formula:

**1.5 Portfolio Mathematics**

**Probability models for portfolio return and risk**

Expected return of portfolio:

Covariance of returns between 2 assets:

Sample covariance calculated the way mentioned in previous section

Covariance of a random variable with itself is variance

Covariance matrix: Shows all the covariances between assets in a table

* The diagonal line is variance cos it is covariance with itself

Portfolio variance for 2 assets:

For 3 assets:

Remember that

**Shortfall risk and Roy’s safety first criterion**

Shortfall risk: Probability portfolio value/return will fall below a target value in a given period

Roy’s safety first criterion: Optimal portfolio minimises probability the return falls below min acceptable level (threshold level)

* Minimise

If portfolio is normally distributed, Roy’s safety first criterion is:

Larger Safety first ratio is better

Remember to convert to return % before calculating ratio

**1.6 Simulation Methods**

**Lognormal distribution**

Lognormal distribution: Generated by

* Normal distribution generated by
* The logarithm of lognormal is normal

Always 0 or above

Useful for modelling continuously compounded returns

asset price at time 0

asset price at time

is continuously compounded return from time 0 to

is approximately normally distributed

is lognormally distributed

We assume returns are independently and identically distribution

* Independent: Past returns can’t predict future ones
* Identical: Mean and variance don’t change over time

**Monte Carlo**

Monte Carlo simulation: Repeated generation of risk factors the generate distribution of security values

* Risk factors have probability distributions, and computer generates random values for each risk factor to plug in
* Repeated many times
* Can draw inferences about expected value and variance

Statistical, not analytic method

Uses

* Value securities
* Simulate trading strategy
* Estimate value at risk

Advantage: Can test scenarios that haven’t happened in the past

Disadvantage: Complex and depends on assumptions

**Resampling**

Resampling: Repeatedly draw subsamples from observed sample each with same number of observations

* Generates data to use in simulation
* Can inter population parameters

Bootstrap resampling: Repeatedly draw samples of size from dataset, replacing them each time

**1.7 Estimation and Inference**

**Sampling methods**

Probability sampling: Selecting a sample when we know the probability of each sample member in the overall population

* Random sampling: Each item assumed to have the same probability of being selected

Non-probability sampling: Not everyone has a known chance of being picked

* Can be low cost, or easy access data, or judgement on data selection

**Probability sampling**

Simple random sampling: Randomly selecting a data point each time

Stratified random sampling: Uses classification system to separate population into smaller groups, then random samples the groups and pooled to form combined sample

* Size of samples from each group is based on size of group relative to population

Stratified random sampling used for bond indexing as hard to recreate population of bonds (e.g., classify by maturity, duration, coupon rate)

Cluster sampling: Based on subsets of population, assume each subset is representative for the population

One-stage cluster sampling: Random cluster selected, all data in it comprises sample

Two-stage cluster sampling: Random samples from each selected cluster comprises sample

* Has more sampling error than 1 stage

Cluster sampling advantage: Low cost and quick

**Non probability sampling methods**

Convenience sampling: Selecting sample data based on ease of access

* Greater sampling error as it is not random

Judgemental sampling: Each observation is selected by researcher using judgement

* Researcher bias or bad judgement can lead to sampling error
* Can be more representative if done well

Need to consider if data distribution is the same over time

**Central limit theorem**

Central limit theorem: For simple random samples of size from a population with mean and variance , the sampling distribution of sample mean approaches normal distribution with mean and variance when sample size is large enough

Inferences can be made about population mean regardless of population distribution

Large enough sample size usually >30

Characteristics of CLT:

* Large enough sample size means sampling distribution of sample means is approximately normal
* Mean of population and mean of distribution of samples means are equal
* Variance of distribution of sample means is

Standard error of sample mean: Standard deviation of the distribution of sample means

If standard deviation of population is known, the SE of sample mean is:

is standard error of sample mean

is standard deviation of population

Since cannot be known:

is standard deviation of sample

**Resampling to estimate sampling distribution**

2 ways to estimate SE of sample mean with resampling data

Jackknife: Calculates multiple sample means, each with one observation removed from sample

* SD of sample means can estimate SE of sample means
* Low cost tool

Bootstrap: Repeatedly draw samples of size with replacement

* SD of sample means can estimate SE of sample means
* Can improve accuracy but more computer power needed
* Can construct confidence intervals

**1.8 Hypothesis testing**

**Hypothesis testing basics**

Hypothesis: Statement about population parameter for testing

Procedure

1. State hypothesis
2. Select appropriate test statistic
3. Specify significance level
4. State decision rule regarding the hypothesis
5. Collect sample and calculate sample statistics
6. Make a decision regarding hypothesis
7. Make a decision based on test results

Null hypothesis : The hypothesis the researcher wants to reject

* Stated as a statement about a population parameter (e.g., )

Alternative hypothesis : What is concluded if there is sufficient evidence to reject null and is what you are trying to assess

* If null is discredited, the alternative hypothesis is valid

Null and Alternative must be mutually exclusive and exhaustive

Decision rule: Reject if test statistic is out of range of critical values for the significance level

Common critical values

* 1.65: 2 tailed at 10% significance, or 1 tailed at 5% significance
* 1.96: 2 tailed at 5% significance
* 2.33: 1 tailed at 1% significance
* 2.58: 2 tailed at 1% significance

Test statistic: Difference between sample statistic and hypothesised value, scaled by SE of sample statistic

Critical value for appropriate test statistic depends on distribution

* 4 distributions: T-distribution, Z-distribution, Chi-square, F-distribution

**Type 1 and 2 errors**

Type 1 error: Rejection of null when it is actually true

* False rejection

Type 2 error: Failure to reject null when it is actually false

* Missed detection

Significance level is probability of making type 1 error

Power of a test: Probability of correctly rejecting the null when it is false

* It is
* I.e., 1 minus probability of not rejecting null when it is false

|  |  |  |
| --- | --- | --- |
|  | is true | is false |
| Do not reject | Correct decision | Incorrect decision  Type 2 error |
| Reject | Incorrect decision  Type 1 error  Significance level = P(Type 1 error) | Correct decision  Power of the test  1-P(Type 2 error) |

Decreasing significance level will reduce chance of Type 1 error, but increase chance of Type 2 error

Can decrease probability of Type 2 error at a given significance level by increasing sample size

Cannot say accept null hypothesis, it can only be supported or rejected

Decision rule for rejecting or failing to reject null hypothesis is based on distribution of test statistic

* E.g., use z-distribution if it is normal

p-value: Probability of obtaining test statistic that would lead to rejection of null, assuming null is true

* It is smallest significance level for which null can be rejected

**Types of hypothesis tests**

For value of population mean, use t-test, or z-test if sample is large enough

For equality of 2 population means, use t-test

* If samples are independent use difference in means test
* If samples are dependent use paired comparisons test

For value of population variance, use chi-squared test

For equality of 2 population variances, use F-test

**Value of population mean**

See if test statistic is out of the critical value range

* If out, then reject

**Difference between means (independent samples)**

T-test for differences between means of 2 populations requires:

* Independent samples and normally distributed populations

Testing to see if means are equal:

If sample means are close, then we do not reject equality

If sample means are far, we reject equality

**Paired comparisons (means of dependent samples)**

Paired comparisons test used for difference between means of 2 dependent samples

* E.g., If both samples were affected by an economic event

is mean of population of paired differences

is hypothesised mean of paired differences, commonly 0

Key difference with Difference between means and Paired comparisons

* One about the significance of the difference between the means of 2 populations
* Other about significance of the mean of the differences between pairs of observations in the same group
* One is for independent, one is for dependent

**Value of population variance**

Chi-square test used for variance of normally distributed

is true population variance, is hypothesised variance

Chi-square is denoted as

Chi-square distribution is asymmetrical, and approaches normal distribution in share as degrees of freedom increases

Chi-square distribution can’t have values below 0

For Chi-square, the critical values aren’t symmetrical

* So for a 5% 2 tailed significance level, look for the value for 0.975 and 0.025

**Comparing 2 population variances**

F-distributed test used for equality of the variances of 2 populations

F-test assumes populations are normally distributed and samples are independent

and are variances of population 1 and 2

F-distribution is right skewed and can’t be negative

Shape is determined by 2 separate degrees of freedom – the numerator and denominator

Lower critical value is the reciprocal of upper critical value

* Upper critical always >1, lower critical always <1

**Degrees of freedom**

|  |  |  |
| --- | --- | --- |
|  |  | **Degrees of freedom** |
| One population mean | t-statistic |  |
| Two population means | t-statistic |  |
| One population variance | Chi-square statistic |  |
| Two population variances | F-Statistic |  |

**Parametric and non-parametric tests**

Parametric tests: Rely on assumptions about the distribution of the population and are specific to population parameters

* E.g., Z-test relies on mean, SD and large sample

Nonparametric tests: Do not consider a particular population parameter or have few assumptions about the population being sampled

* Used when there are concerns about quantities other than parameters (e.g., median, mode, ranks), or when assumptions of parametric tests can’t be supported, or when data is not suitable for parametric tests

Situations for nonparametric tests:

* Assumptions about distribution of random variable that support a parametric test are not met (e.g., not normal distribution and small size sample)
* When data are ranks, not values
* Hypothesis doesn’t involve parameters of the distribution (e.g., hypothesis of is data random)

**1.9 Parametric and Non-parametric tests of independence**

**Tests hypothesis for correlation**

Test statistic for hypothesis the population correlation equals 0, assuming both variables are normally distributed:

is sample correlation

is sample size

Follows t distribution with degrees of freedom

Spearman rank correlation test (non-parametric test) tests if 2 sets of ranks are correlated

* Ranks are ordered values

Rank correlation is given as:

is the difference between 2 ranks

Then test it using this

* Needs to have sample size >30 to have t-distribution and degrees of freedom

**Contingency data table**

Contingency data table: Shows number of observations from sample that have a combination of 2 characteristics

* E.g., Earnings growth and Div yield

Test statistic for contingency table:

is the number of observations in cell (row and column )

is the expected number of observations in cell

is the number of row categories

is the number of column categories

Follows chi-square distribution

Degrees of freedom:

is given by:

must be above critical value (found with degrees of freedom and significance level) to reject null hypothesis

**1.10 Simple Linear Regression**

Linear regression: Explains variation in a dependent variable in terms of the variation in an independent variable

Dependent variable: Variable whose variation is explained by the independent variable

* AKA Explained variable, endogenous variable, predicted variable

Independent variable: Variable used to explain the variation of dependent variable

* AKA Explanatory variable, exogenous variable, predicting variable

Linear regression model:

is the intercept

* I.e., when is 0

is the slope

* I.e., the increase in for a one unit increase in

is the residual

Regression line (line of best fit) is given by:

The hat shows it is a predicted value

Regression line minimises the sum of squared errors (differences between predicted vales and actual values)

* AKA OLS regression

is calculated as:

can be calculated as:

is mean of

is mean of

Conclusions regarding importance of independent variable explaining dependent variable are based on statistical significance of the slope coefficient

* Need to conduct hypothesis test

**Linear regression assumptions**

1) Linear relationship between dependent and independent variables

* Plot the observed dependent and independent variables on a graph to get a sense

2) Homoskedasticity: Variance of residual term is constant for all observations

* Heteroskedasticity is when this is violated: Residual variance has a trend over time or with independent variable

3) Residual term is independently distributed

* Residual for one observation is not correlated with another observation
* Plot residuals by themselves to see if they have a pattern (e.g., seasonality)

4) Normally distributed residual term

* This means we can conduct hypothesis testing for evaluating goodness of fit of model

**Linear Regression Terminology**

Total sum of squares SST: Gives the total variation in the dependent variable

* Sum of squared differences between actual and mean

Sum of squares regression SSR: Gives the variation in the dependent variable that is explained by the independent variable

* Sum of squared differences between predicted and mean

Mean square regression MSR: SSR divided by number of independent variables

* Simple linear regression only has 1, so MSR=SSR

Sum of squared errors SSE: Gives the unexplained variation in the dependent variable

* AKA Sum of squared residuals

Mean squared error MSE: SSE divided by the degrees of freedom (which is minus number of independent variables)

* So for simple linear regression

**Analysis of variance (ANOVA)**

ANOVA: Analyses totally variability of the dependent variable

Gives an ANOVA table

* Has source of variation (regression, error, total), degrees of freedom for each, etc

For a simple linear regression:

|  |  |  |  |
| --- | --- | --- | --- |
| Source of variation | Degrees of freedom | Sum of squares | Mean Sum of Squares |
| Regression (explained) |  |  |  |
| Error (unexplained) |  |  |  |
| Total |  |  |  |

**Standard error of estimate SEE**

SEE: Standard deviation of residuals

* Lower is better

**Coefficient of determination**

Coefficient of determination : Percentage of total variation in the dependent variable explained by the independent variable

For a simple linear regression, is found by squaring the correlation coefficient

**F statistic**

F-Statistic: Assesses how well a set of independent variables as a group explains the variation in the dependent variable

Calculated as:

For a simple linear regression:

ALWAYS a one-tailed test

For simple linear regression, since there is only one independent variable, the F-test is equivalent to a t-test of the statistical significance of the slope coefficient

Reject if

**Hypothesis test of a regression coefficient**

T-test can be used to test the hypothesis that the true slope coefficient

is the estimate for

is the standard error

There are degrees of freedom

Then do t-test that was the same as the one for the correlation coefficient

**Predicted values**

Predicted value: Values of the dependent variable based on estimated regression coefficients and a prediction about the value of the independent variable

They are values predicted by regression, given an estimate of the independent variable

is the predicted value of dependent variable

is the forecasted value of independent variable

**Confidence intervals for predicted values**

The confidence interval for a predicted value is:

is the two-tailed critical t-value at the desired significance level with dof

* So for 95% confidence level, use the 2.5% t-value

If sample is large, use z-statistic

is the standard error of the forecast

**Different functional forms**

If variables aren’t linear, can transform them so they are

3 common ones:

* Log-lin model: Dependent variable log, independent variable linear
* Lin-log model: Dependent variable linear, independent variable log
* Log-log model: Both log

Log lin model:

* Relative change for dependent variable for absolute change in independent variable

Lin log model:

* Absolute change for dependent variable for relative change in independent variable

Log log model:

* Relative change for dependent variable for relative change in independent variable

**1.11 Big data techniques**

**Fintech**

Fintech: Tech for financial services

Where it is developing:

* Handling and analysis of large datasets

**Big data**

Big data: All potentially useful info generated in an economy

* Includes financial data, alternative data (e.g., social media), corporate exhaust (bank records, retail scanner data from businesses), IoT

Big data characteristics

* Volume: Huge growth
* Velocity: How quick it is communicated
* Variety: Different data structures, structured vs unstructured

Data science: How to extract info from Big Data

* Capture: Collecting and transforming data
* Curation: Adjusting for bad or missing data
* Storage: Archiving and accessing data
* Search: Looking through stored data for info
* Transfer: Moving data

AI can help process qualitative and unstructured data

* Can use Neural networks

Machine learning

* Given inputs and outputs, and learns to model output data on input data
* Requires training dataset to find relationships, validation dataset to refine relationship models, and test dataset to see predictive ability

Supervised learning: Input and output data are labelled

Unsupervised learning: Input data not labelled

Deep learning: Uses layers of neural networks to identify patterns, can use supervised or unsupervised

Overfitting: Learns input and output data too closely, models noise as true parameters

* Too complex a model

Underfitting: Fails to identify actual patterns, treating true parameters as noise

* Model not complex enough

**Applications to investment management**

Text analytics: Analysis of unstructured text or voice data

* E.g., Analysing frequency of words/phrases

Natural language processing: Interpret human language

* E.g., used for compliance, detect sentiment

Useful for risk modelling

Algo trading: Trading based on predetermined rules

* E.g., executing trade based on real time data, splitting up large orders, high frequency trading to take advantage of inter-day mispricing