CFA I - Quantitative Methods

April 25, 2022

Contents

1 The Time Value of Money

1.1 Interest Rates

$r = real\ risk\ free\ interest\ rate +$
$inflation\ premium +$
$default\ risk\ premium +$
$liquidity\ premium +$
$maturity\ premium$

1.2 Future Value of a Single Cash Flow (Lump Sum)

$FV_N = PV(1+r)^N$, where:

1. r - stated interest rate per period
2. N - number of compounding periods

1.3 Non-annual Compounding (Future Value)

$FV_N = PV(1+\frac{r_s}{m})^{mN}$, where:

1. r_s - stated annual interest rate
2. m - number of compounding periods per year
3. N - number of years

1.4 Continuous Compounding, Stated and Effective Rates

$FV_N = PV * e^{r_s N}$, where:

1. r_s - stated annual interest rate
2. N - number of years
3. e - transcendental number $e = 2.7182818$

Effective annual rate : $EAR = (1 + periodic\ interest\ rate)^m - 1$
with continuous compounding: $EAR = e^{r_s} - 1$

1.5 Future Value of a Series of Cash Flows, Future Value Annuities

$FV_N = A(\frac{(1+r)^N - 1}{r})$, where:

1. A - annuity amount
2. N - number of time periods
3. r - interest rate per period

1.6 Present Value of a Single Cash Flow (Lump Sum)

$PV = FV_N(1+r)^{-N}$

1.7 Non-Annual Compounding (Present Value)

$PV = FV_N(1 + \frac{r_s}{m})^{-mN}$, where:

1. m - number of compounding periods per year
2. r_s - quoted annual interest rate
3. N - number of years

1.8 Present Value of a Series of Equal Cash Flows (Annuities) and Unequal Cash Flows

The present value of a series of equal cash flows:
$PV = A(\frac{1-\frac{1}{(1+r)^N}}{r})$, where:

1. A - annuity amount
2. N - number of annuity payments
3. r - interest rate per period corresponding to the frequency of annuity payments (annual, quarterly, monthly)

The present value of a series of unequal cash flows:

$PV = \frac{A_1}{(1+r)} + \frac{A_2}{(1+r)^2} + \frac{A_3}{(1+r)^3} + ... + \frac{A_{N-1}}{(1+r)^{N-1}} + \frac{A_N}{(1+r)^N}$

1.9 Present Value of a Perpetuity and Present Values Indexed at Times other than $T = 0$

$PV = A * \sum_{t=1}^{\infty}(\frac{1}{(1+r)^t}) = \frac{A}{r}$

1.10 Solving for Interest Rates, Growth Rates, and Number of Periods

$(r =)g = (\frac{FV_N}{PV})^{1/N} - 1$

2 Organizing, Visualizing, and Describing Data

2.1 Measures of Central Tendency

Sample mean/average: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- the sum of the deviations around the mean always equals 0
- very sensitive to outliers

Weighted mean : $\bar{X}_w = \sum_{i=1}^{n} w_i X_i$, where $\sum_i w_i = 1$

Geometric mean: $\bar{X}_G = \sqrt[n]{X_1 X_2 X_3 ... X_4}$, with $X_i \geq 0$ for $i = 1, 2, ..., n$

Geometric mean return: $R_G = [\Pi_{t=1}^{T}(1 + R_t)]^{\frac{1}{T}} - 1$

Harmonic mean: $\bar{X}_H = \frac{n}{\sum_{i=1}^{n}(1/X_i)}$, with $X_i > 0$ for $i = 1, 2, ..., n$

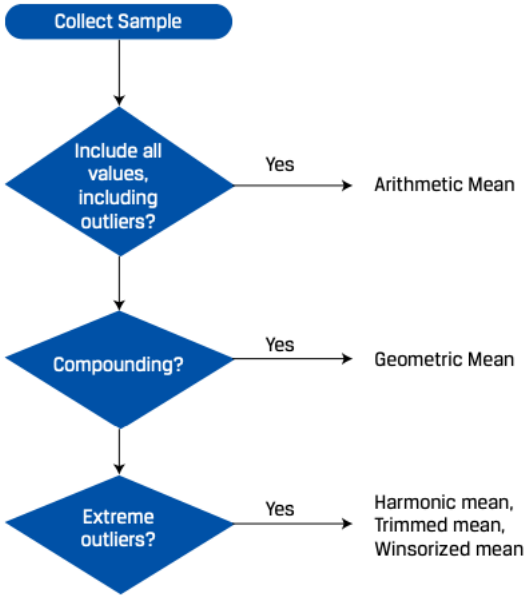

Figure 1: Central Tendency Measures

2.2 Quantiles

Interquartile range: $IQR = Q_3 - Q_1$

Percentile location (position) in an array with n entries sorted in ascending order: $L_y = (n + 1)\frac{y}{100}$, where y is the percentage point at which we are dividing the distribution.

2.3 Measures of Dispersion

Range = Maximum value - Minimum value

Mean absolute deviation: $MAD = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$, where n is the number of observations in the sample

Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

Sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

Dispersion in the relationship between the arithmetic and geometric means: $\bar{X}_G \approx \bar{X} - \frac{s^2}{2}$

2.4 Downside Deviation and Coefficient of Variation

Target downside deviation (target semideviaiton): $s_{Target} = \sqrt{\sum_{\forall X_i \leq B}^{n} \frac{(X_i - B)^2}{n-1}}$, where B is the target

Coefficient of variation (relative dispersion): $CV = s/\bar{X}$, where s is the sample standard deviation, and $\bar{X}$ is the sample mean

2.5 The Shape of the Distributions

Sample skewness: $sk \approx (\frac{1}{n}) \frac{\sum_{i=1}^{n} (X_i - \bar{(X)})^3}{s^3}$ for samples with $n \geq 100$

Kurtosis - measure of the combined weight of the tails of a distribution relative to the rest of the distribution (fat-tailed distribution = *leptokurtic*, thin-tailed distribution = *platykurtic*, normal distribution = *mesokurtic*). A normal distribution has kurtosis of 3.0, so a fat-tailed distribution has a kurtosis of above 3 and a thin-tailed distribution of below 3.0.

Kurtosis: $K \approx [(\frac{1}{n}) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{s^4}]$

Excess kurtosis: $K_E = K - 3$

2.6 Correlation Between Two Variables

Sample covariance: $s_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

Sample correlation coefficient: $r_{XY} = \frac{s_{XY}}{\sigma_X \sigma_Y}$

3 Probability Concepts

3.1 Introduction, Probability Concepts, and Odd Ratios

Odds for E = $P(E)/[1 - P(E)]$. Given odds for E of a to b, the implied probability of E is $a/(a+b)$

Odds against E = $[1 - P(E)]/P(E)$. Given odds against E of a to b, the implied probability of E is $b/(a+b)$

3.2 Conditional and Joint Probability

$P(A|B) = P(AB)/P(B), P(B) \neq 0$

$P(AB) = P(A|B) * P(B)$

Addition rule for probabilities: $P(A\,or\,B) = P(A) + P(B) - P(AB)$. Only if two events A and B were mutually exclusive, so that $P(AB) = 0$, would it be correct to state that $P(A\,or\,B) = P(A) + P(B)$

Two events are *independent* if the occurrence of one event does not affect the probability of occurrence of the other event. Two events are *independent* iif $P(A|B) = P(A)$, or, equivalently, $P(B|A) = P(B)$

Multiplication rule for independent events: $P(AB) = P(A)P(B)$ (multiplication rule generalized to more than two independent events)

Total probability rule: $P(A) = P(AS_1) + P(AS_2) + ... + P(AS_n) = P(A|S_1)P(S_1) + P(A|S_2)P(S_2) + ... + P(A|S_n)P(S_n)$, where $S_1, S_2, ..., S_n$ are mutually exclusive and exhaustive scenarios or events.

3.3 Expected Values (Mean), Variance, and Conditional Measures of Expected Value and Variance

The *expected value* of a random variable is the probability-weighted average of the possible outcomes of the RV: $E(X) = P(X_1)X_1 + P(X_2)X_2 + ... + P(X_n)X_n = \sum_{i=1}^{n} P(X_i)X_i$

The *variance* of a RV is the expected value of squared deviations from the TV's expected value: $\sigma^2(X) = E\{[X - E(X)]^2\} = \sum_{i=1}^{n} P(X_i)[X_i - E(X)]^2$

Conditional expected values: $E(X|S) = P(X_1|S) * X_1 + P(X_2|S) * X_2 + ... + P(X_n|S) * X_n$

Total probability rule for expected value: $E(X) = E(X|S_1) * P(S_1) + E(X|S_2) * P(S_2) + ... + E(X|S_n) * P(S_n)$, where $S_1, S_2, ..., S_n$ are mutually exclusive and exhaustive scenarios.

3.4 Expected Value, Variance, Standard Deviation, Covariances, and Correlations of Portfolio Returns

Expected return on the portfolio: $E(R_p) = E(w_1R_1 + w_2R_2 + ... + w_nR_n) = w_1E(R_1) + w_2E(R_2) + ... + w_nE(R_n)$

Covariance between two RVs: $Cov(R_i, R_j) = E[(R_i - ER_i)(R_j - ER_j)]$

Sample covariance between two RVs: $Cov(R_i, R_j) = \sum_{n=1}^{n}(R_{i,t} - \bar{R}_i)(R_{j,t} - \bar{R}_j)/(n-1)$

Portfolio variance: $\sigma^2(R_p) = E\{[R_p - E(R_p)]^2\}$; $\sigma^2(R_p) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_iw_jCov(R_i, Rj)$ *(the diversification benefit increases with decreasing covariance)*

3.5 Covariance Given a Joint Probability Function

The joint probability function of two random variables X and Y, denoted $P(X, Y)$, gives the probability of joint occurrences of values of X and Y.

Covariance between random variables R_A and R_B: $Cov(R_A, R_B) = \sigma_i\sigma_j P(R_{A,i}, R_{B,i})(R_{A,i} - ER_A)(R_{B,i} - ER_B)$ (sum all possible deviation cross-products weighted by the appropriate joint probability)

Definition of independence for random variables: Two random variables X and Y are independent if and only if $P(X, Y) = P(X)P(Y)$.

Multiplication rule for expected value of the product of uncorrelated random variables: $E(XY) = E(X)E(Y)$, if X and Y are uncorrelated.

3.6 Bayes' Formula

Bayes' formula is a rational method for adjusting our viewpoints as we confront new information. Bayes' formula uses the occurrence of the event to infer the probability of the scenario generating it (it is sometimes called an inverse probability).

Updated probability of event given new information =

$\frac{Probability\ of\ the\ new\ information\ given\ event}{Unconditional\ probability\ of\ the\ new\ information} * Prior\ probability\ of\ event$

$$P(Event|Information) = \frac{P(Information|Event)}{P(Information)} * P(Event)$$

3.7 Principles of Counting

Multiplication Rule for Counting: If one task can be done in n_1 ways, and a second task, given the first, can be done in n_2 ways, and a third task, given the first two tasks, can be done in n_3 ways, and so on for k tasks, then the number of ways the k tasks can be done is $(n_1)(n_2)(n_3)...(n_k)$

Multinomial Formula (General Formula for Labeling Problems): The number of ways that n objects can be labeled with k different labels, with n_1 of the first type, n_2 of the second type, and so on, with $n_1 + n_2 + ... + n_k = n$, is given by: $\frac{n!}{n_1!n_2!...n_k!}$

Combination Formula (Binomial Formula): The number of ways that we can choose r objects from a total of n objects, when the order in which the r objects are listed does not matter, is $\binom{n}{x} = \frac{n!}{(n-r)!r!}$

An ordered listing is known as a permutation (a permutation is an ordered subset of n distinct objects).
Permutation Formula: The number of ways that we can choose r objects from a total of n objects, when the order in which the r objects are listed does matter, is $\frac{n!}{(n-r)!}$

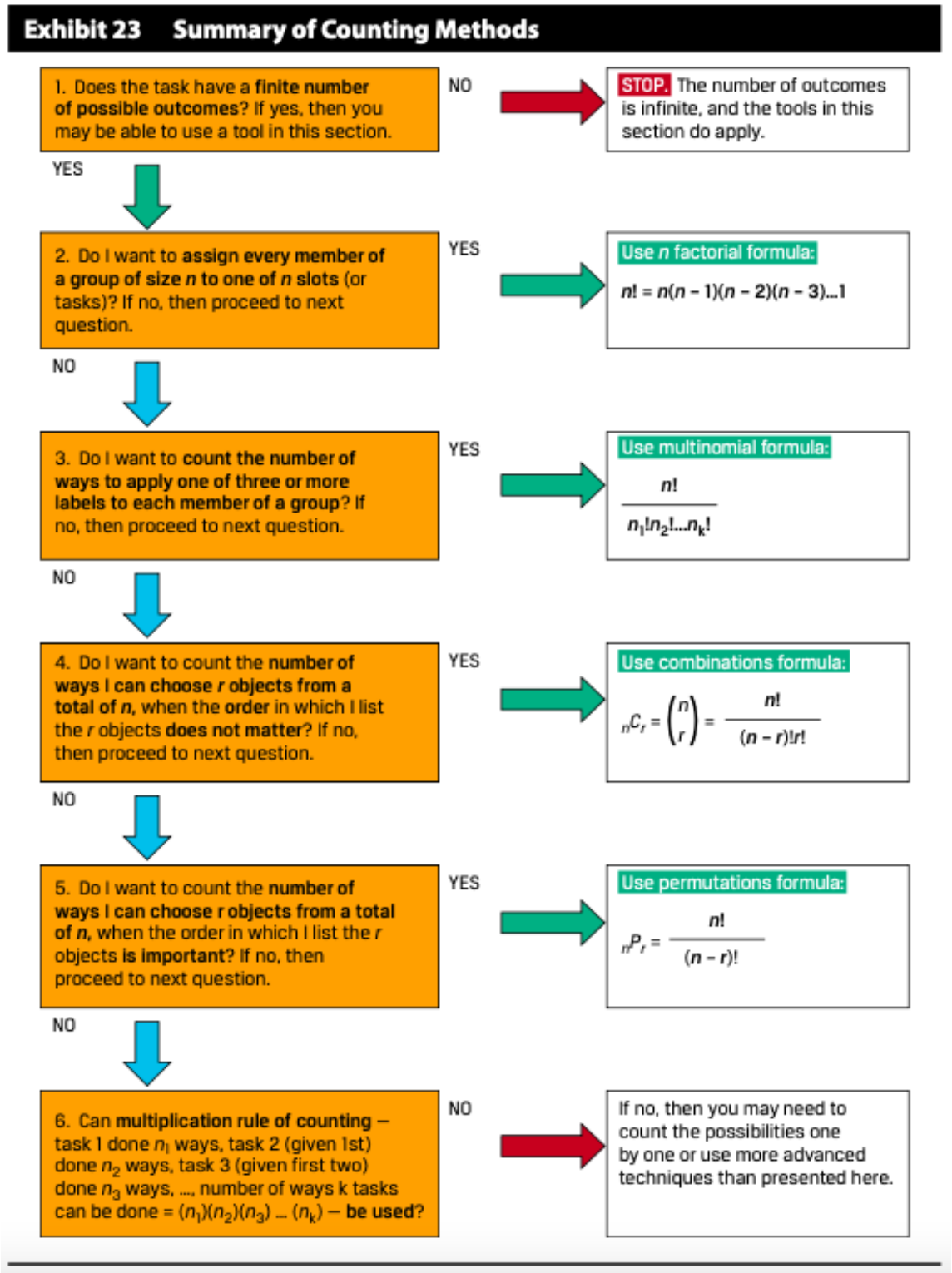

Figure 2: Counting Methods - Summary

4 Common Probability Distributions

4.1 Introduction and Discrete Random Variables

Probability function specifies the probability that the RV takes on a specific value $p(x) = P(X = x)$ (discrete RV), $f(x) = P(X = x)$ (continuous RV, probability density function).

Cumulative distribution function gives the probability that a RV X is less than or equal to a particular value x, $P(X \leq x)$. For both discrete and continuous random variables, the shorthand notation is $F(x) = P(X \leq x)$

4.2 Discrete and Continuous Uniform Distribution

The pdf for a uniform RV:

$$f(x) = \begin{cases} \frac{1}{b-a} & for a \leq x \leq b \\ 0 & otherwise \end{cases}$$

The cdf for a uniform RV:

$$f(x) = \begin{cases} 0 & for x < a \\ \frac{x-a}{b-a} & for a \leq x \leq b \\ 1 & for x < b \end{cases}$$

The mathematical operation that corresponds to finding the area under the curve of a pdf $f(x)$ from a to b is the definite integral of $f(x)$ from a to b: $P(a \leq X \leq b) = \int_a^b f(x)dx$

The probability of a continuous RV equaling any fixed point is 0. For any continuous random variable X, $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$, because the probabilities at the endpoints a and b are 0.

4.3 Binomial Distribution

Bernoulli RV: $p(1) = P(Y = 1) = p$, $p(0) = p(Y = 0) = 1 - p$

A binomial RV X is defined as the number of successes in n Bernoulli trials. A binomial RV is the sum of Bernoulli RVs Y_i, where $i = 1, 2, ..., n$: $X = Y_1 + Y_2 + ... + Y_n$, where Y_i is the outcome of the i-th trial.

Binomial distribution assumptions:

1. The probability p of success is constant for all trials
2. The trials are independent

Under these two assumptions, a binomial RV is completely described by two parameters, n and p: $X \sim B(n, p)$. Bernoulli RV is a binomial RV with $n = 1 : Y \sim B(1, p)$.

For X distributed $B(n, p)$, the probability of x successes in n trials is given by:

$$p(x) = P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

4.4 Normal Distribution

Central limit theorem: the sum (and mean) of a large number of iid RVs (with finite variance) is approximately normally distributed (whatever distribution the RVs follow). (A linear combination of two or more RVs is also normally distributed.)

A multivariate distribution specifies the probabilities for a group of related random variables.

PDF of the normal distribution for the one RV (for $-\infty < x < +\infty$):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(x-\mu)^2}{2\sigma^2})$$

Probabilities using the normal distribution:

- approx. 50% of all observations fall in the interval $\mu \pm (2/3)\sigma$
- approx. 68% of all observations fall in the interval $\mu \pm \sigma$
- approx. 95% of all observations fall in the interval $\mu \pm 2\sigma$
- approx. 99% of all observations fall in the interval $\mu \pm /3\sigma$

Standardizing a RV: $Z = (X - \mu)/\sigma$

4.5 Applications of the Normal Distribution

Safety-first optimal portfolio (focus on shortfall risk): in returns are normally distributed, the SF optimal portfolio maximizes the safety-first ratio (SFRatio): $SFRatio = [E(R_p) - R_L]/\sigma_p$, where R_L is the lowest acceptable portfolio return. For a portfolio with a given safety-first ratio, the probability that its return will be less than R_L is $N(-SFRatio) = 1 - N(SFRatio)$

Value at risk (VaR) is a money measure of the minimum value of losses expected over a specified time period at a given level of probability.

4.6 Lognormal Distribution and Continuous Compounding

A random variable Y follows a lognormal distribution if its natural logarithm, lnY, is normally distributed (if the natural logarithm of random variable Y, lnY, is normally distributed, then Y follows a lognormal distribution).

The two parameters of a lognormal distribution are the mean and standard deviation (or variance) of its associated normal distribution: the mean and variance of lnY, given that Y is lognormal.

The expressions for the mean and variance of a lognormal variable (where μ and σ^2 are the mean and variance of the associated normal distribution:

- mean (μ_L) of a lognormal RV = $exp(\mu + 0.5\sigma^2)$
- variance (σ_L^2) of a lognormal RV = $exp(2\mu + \sigma^2) * [exp(\sigma^2) - 1]$

If a stock's continuously compounded return is normally distributed, then future stock price is necessarily lognormally distributed.

$S_T = S_0 * exp(r_{0,T})$, where:

- S_T - stock price at time T
- S_0 - current stock price

- $r_{0,T}$ - contunuously compounded return from 0 to T

The continuously compounded return associated with a holding period return is the natural logarithm of 1 plus that holding period return, or equivalently, the natural logarithm of the ending price over the beginning price (the price relative): $r_{t,t+1} = ln(S_{t+1}/S_t) = ln(1 + R_{t,t+1})$

Continuously compounded return to time T is the sum of the one-period continuously compounded returns: $r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + ... + r_{0,1}$

Volatility measures the standard deviation of the continuously compounded returns on the underlying asset.

5 Stundet's T-, Chi-Square, and F-Distributions

Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and the population's standard deviation is unknown.

Let $X_1, ..., X_n$ be independently and identically drawn from the distribution $N(\mu, \sigma^2)$, i.e. this is a sample of size n from a normally distributed population with expected mean value μ and variance σ^2.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean and let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be the sample variance.

Then the RV $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution and the RV $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student's t-distribution with $n-1$ degrees of freedom.

The χ^2 distribution with k degrees of freedom is the distribution of the sum of the squares of k independent standard normally distributed RVs.

The relationship between the χ^2 distribution and F-distribution is as follows: if χ_1^2 is one chi-square RV with m degrees of freedom and χ_2^2 is another chi-square RV with n degrees of freedom, then $F = (\chi_1^2/m)/(\chi_2^2/n)$ follows an F-distribution with m numerator and n denominator degrees of freedom.

TODO: append additional information of these three distributions

5.1 Monte Carlo Simulation

A characteristic feature of Monte Carlo simulation is the generation of a large number of random samples from a specified probability distribution or distributions to represent the role of risk in the system.

Random observations from any distribution can be generated using a uniform random variable.

6 Sampling and Estimation

6.1 Sampling Methods

Simple random sample - subset of a larger population created in such a way that each element of the population has an equal probability of being selected to the subset (particularly useful when data in

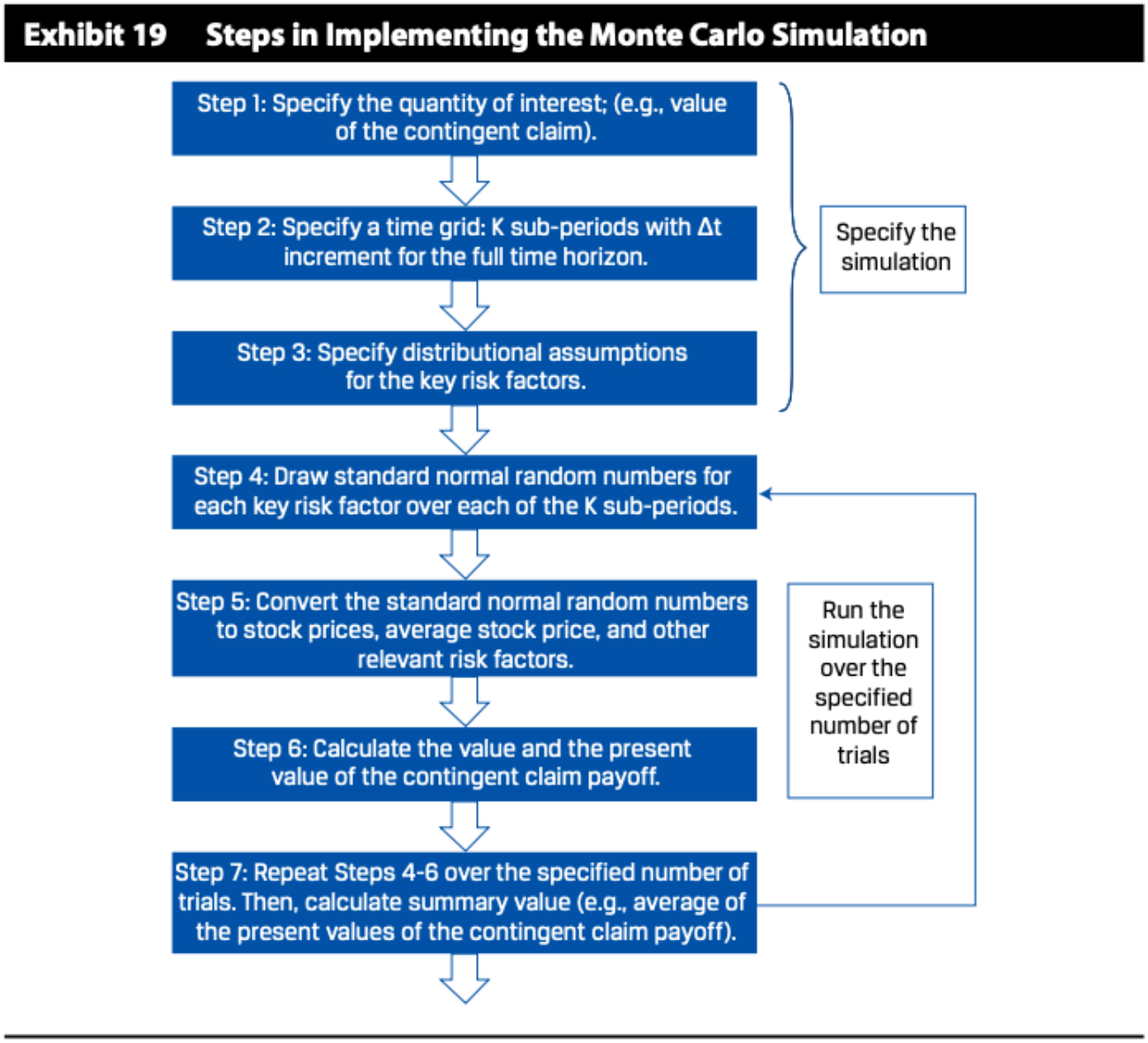

Figure 3: Monte Carlo Simulation Scheme

the population is homogeneous).

Sampling error is the difference between the observed value of a statistic and the quantity it is intended to estimate as a result of using subsets of the population.

The sampling distribution of a statistic - distribution of all the distinct possible values that the statistic can assume when computed from samples of the same size randomly drawn from the same population. A sample statistic is a random variable (has its own distribution).

Stratified random sampling - the population is divided into subpopulations (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled to form a stratified random sample.

Culster sampling - population is divided into clusters, each of which is essentially a mini-representation of the entire populations. Then certain clusters are chosen as a whole using simple random sampling.

A major difference between cluster and stratified random samples is that in cluster sampling, a whole cluster is regarded as a sampling unit and only sampled clusters are included. In stratified random sampling, however, all the strata are included and only specific elements within each stratum are then selected as sampling units.

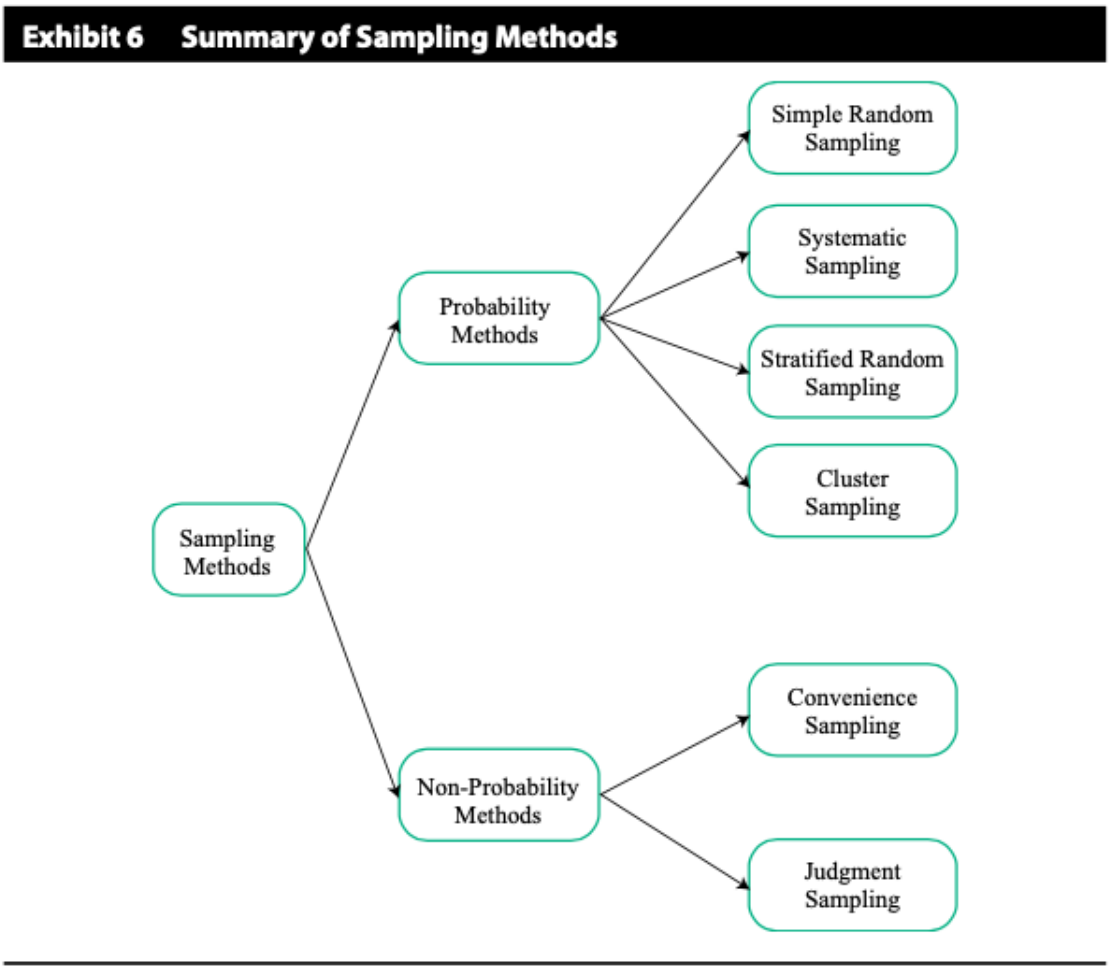

Figure 4: Summary of the Sampling Methods

6.2 Distribution of the Sample Mean and the Central Limit Theorem

The central limit theorem: given a population described by any probability distribution having mean μ and finite variance σ^2, the sampling distribution of the sample mean $\bar{X}$ computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance σ^2/n (the population variance divided by n) when the sample size n is large.

Standard error of the sample mean: For sample mean $\bar{X}$ calculated from a sample generated by a population with standard deviation σ, the standard error of the sample mean is given by one of two expressions:

- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (when we know σ, the population standard deviation)
- $s_{\bar{X}} = \frac{s}{\sqrt{n}}$ (when we do not know the population standard deviation and need to use the sample standard deviation, s, to estimate it)

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

6.3 Point Estimates of the Population Mean

Three desirable properties of estimators:

- Unbiasedness: an unbiased estimator is one whose expected value (the mean of its sampling distribution) equals the parameter it is intended to estimate.
- Efficiency: an unbiased estimator is efficient if no other unbiased estimator of the same parameter has a sampling distribution with smaller variance.
- Consistency: a consistent estimator is one for which the probability of estimates close to the value of the population parameter increases as sample size increases.

6.4 Confidence Intervals for the Population Mean and Selection of Sample Size

Confidence interval: a range for which one can assert with a given probability $1-\alpha$, called the degree of confidence, that it will contain the parameter it is intended to estimate. This interval is often referred to as the $100(1-\alpha)\%$ confidence interval for the parameter.

Construction of confidence intervals: point estimate $\pm$ reliability factor * standard error

Reliability factor = a number based on the assumed distribution of the point estimate and the degree of confidence $(1-\alpha)$ for the confidence interval.

The quantity "Reliability factor $\times$ Standard error" is sometimes called the precision of the estimator; larger values of the product imply lower precision in estimating the population parameter.

The notation z_α denotes the point of the standard normal distribution such that α of the probability remains in the right tail.

Confidence intervals for the population mean (normally distributed population with known variance): a $100(1-\alpha)\%$ confidence interval for population mean μ when we are sampling from a normal distribution with known variance σ^2 is given by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Reliability factors for confidence intervals based on the standard normal distribution: we use the following reliability factors when we construct confidence intervals based on the standard normal distribution:

- 90% ci : use $z_{0.05} = 1.65$
- 95% ci : use $z_{0.025} = 1.96$
- 99% ci : use $z_{0.005} = 2.58$

A $100(1-\alpha)\%$ confidence interval for population mean μ when sampling from any distribution with unknown variance and when sample size is large is given by:

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

Confidence intervals for the population mean (population variance unknown) — t-distribution. If we are sampling from a population with unknown variance and either of the conditions below holds:

- the sample is large, or

- the sample is small, but the population is normally distributed, or approximately normally distributed,

then a $100(1-\alpha)\%$ confidence interval for population mean μ is given by:

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

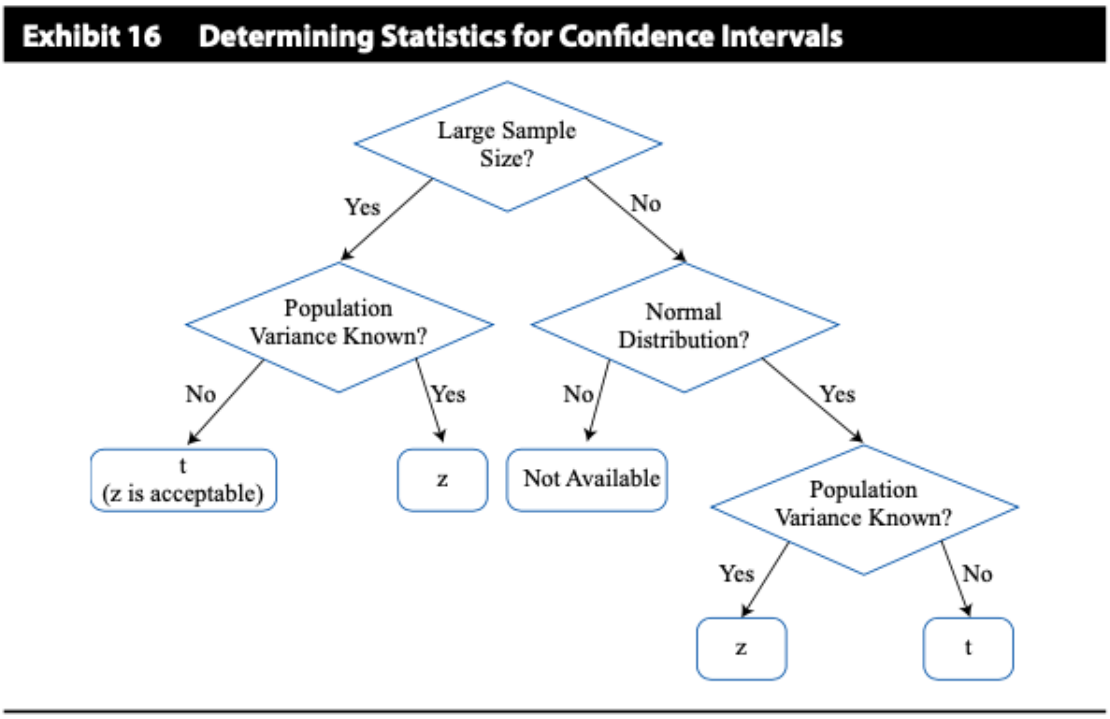

Figure 5: Determining Statistics for Confidence Intervals

The sample size to obtain a desired value of E at a given degree of confidence $(1-\alpha)$ can be derived as $n = [(t * s)/E]^2$.

6.5 Resampling

Bootstrap: we have no knowledge of what the population looks like, except for a sample with size n drawn from the population. Bootstrap mimics the process by treating the randomly drawn sample as if it were the population.

Standard error of the sample mean (bootstrap): $s_{\bar{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\Theta}_b - \bar{\Theta})^2}$, where:

- $s_{\bar{X}}$ is the estimate of the standard error of the sample mean
- B denotes the number of resamples drawn from the original sample
- $\hat{\Theta}_b$ denotes the mean of the sample
- $\bar{\Theta}$ denotes the mean across all the resample means

Jackknife: unlike bootstrap, which repeatedly draws samples with replacement, jackknife samples are selected by taking the original observed data sample and leaving out one observation at a time from the set (and not replacing it).

6.6 Data and Biases

Data snooping: practice of determining a model by extensive searching through a dataset for statistically significant patterns.

Sample selection bias: when data availability leads to certain assets being excluded from the analysis.

Time-period bias: a test design is based on a period that may make the results period specific.

7 Hypothesis Testing

In hypothesis testing, we test to see whether a sample statistic is likely to come from a population with the hypothesized value of the population parameter.

7.1 The Process of Hypothesis Testing

1. State the hypothesis
2. Identify the appropriate test statistic
3. Specify the level of significance
4. State the decision rule
5. Collect data and calculate test statistic
6. Make a decision

The null hypothesis is a statement concerning a population parameter or parameters considered to be true unless the sample we use to conduct the hypothesis test gives convincing evidence that the null hypothesis is false. In fact, the null hypothesis is what we want to reject. The null and alternative hypotheses are stated in terms of population parameters, and we use sample statistics to test these hypotheses.

7.2 Identification of Appropriate Test Statistics

A test statistic is a value calculated on the basis of a sample that, when used in conjunction with a decision rule, is the basis for deciding whether to reject the null hypothesis.

7.3 Level of Significance Specification

The probability of a Type I error in testing a hypothesis (α). This probability is also known as the level of significance of the test, and its complement, $(1 - \alpha)$, is the confidence level.

The only way to reduce the probabilities of both types of errors simultaneously is to increase the sample size, n.

The power of a test is the probability of correctly rejecting the null—that is, the probability of rejecting the null when it is false.

The standard approach to hypothesis testing involves choosing the test statistic with the most power and then specifying a level of significance.

Exhibit 4 Test Statistics and Their Distributions

What We Want to Test	Test Statistic	Probability Distribution of the Statistic	Degrees of Freedom
Test of a single mean	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	*t*-Distributed	$n - 1$
Test of the difference in means	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$	*t*-Distributed	$n_1 + n_2 - 2$
Test of the mean of differences	$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$	*t*-Distributed	$n - 1$
Test of a single variance	$\chi^2 = \frac{s^2(n-1)}{\sigma_0^2}$	Chi-square distributed	$n - 1$
Test of the difference in variances	$F = \frac{s_1^2}{s_2^2}$	*F*-distributed	$n_1 - 1, n_2 - 1$
Test of a correlation	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	*t*-Distributed	$n - 2$
Test of independence (categorical data)	$\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	Chi-square distributed	$(r - 1)(c - 1)$

Note: μ_0, μ_{d0}, and σ_0^2 denote hypothesized values of the mean, mean difference, and variance, respectively. The $\bar{x}$, $\bar{d}$, s^2, s, and r denote for a sample the mean, mean of the differences, variance, standard deviation, and correlation, respectively, with subscripts indicating the sample, if appropriate. The sample size is indicated as n, and the subscript indicates the sample, if appropriate. O_{ij} and E_{ij} are observed and expected frequencies, respectively, with r indicating the number of rows and c indicating the number of columns in the contingency table.

Figure 6: Test Statistics and Their Distributions

7.4 Decision Rule

If we find that the calculated value of the test statistic is more extreme than the critical value or values, then we reject the null hypothesis; we say the result is statistically significant.

Making a decision based on critical values and confidence intervals for a two-sided alternative hypothesis:

- Compare the calculated test statistic with the critical values: if the calculated test statistic is less than the lower critical value or greater than the upper critical value, reject the null hypothesis.
- Compare the calculated test statistic with the bounds of the confidence interval: if the hypothesized value of the population parameter under the null is outside the corresponding confidence interval, the null hypothesis is rejected.

7.5 The Role of p-Values

The p-value is the area in the probability distribution outside the calculated test statistic. Stated another way, the p-value is the smallest level of significance at which the null hypothesis can be rejected. The smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis; if the p-value is less than the level of significance, we reject the null hypothesis.

Exhibit 5 Correct and Incorrect Decisions in Hypothesis Testing

	True Situation	
Decision	**H_0 True**	**H_0 False**
Fail to reject H_0	Correct decision: Do not reject a true null hypothesis.	Type II error: Fail to reject a false null hypothesis. False negative
Reject H_0	Type I error: Reject a true null hypothesis. False positive	Correct decision: Reject a false null hypothesis.

Figure 7: Correct and Incorrect Decisions in Hypothesis Testing

Exhibit 6 Probabilities Associated with Hypothesis Testing Decisions

	True Situation	
Decision	**H_0 True**	**H_0 False**
Fail to reject H_0	Confidence level $(1 - \alpha)$	β
Reject H_0	Level of significance α	Power of the test $(1 - \beta)$

Figure 8: Probabilities Associated with Hypothesis Testing Decisions

7.6 Multiple Tests and Interpreting Significance

A Type I error is the risk of rejection of a true null hypothesis. Another way of phrasing this is that it is a false positive result; that is, the null is rejected (the positive), yet the null is true (hence, a false positive). The expected portion of false positives is the false discovery rate (FDR).

TODO: multiple testing problem and p-value adjusting

7.7 Tests Concerning a Single Mean

The sampling distribution of the mean when the population standard deviation is unknown is t-distributed, and when the population standard deviation is known, it is normally distributed, or z-distributed.

If the population sampled has unknown variance, then the test statistic for hypothesis tests concerning a single population mean, μ, is: $t_{n-1} = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ (this test statistic is t-distributed with n - 1 degrees of freedom), where:

- $\bar{X}$ - sample mean
- μ_0 - hypothesized value of the population mean
- s - sample standard deviation
- n - sample size
- $s_{\bar{X}} = s/\sqrt{n}$ - estimate of the sample mean standard error

7.8 Tests Concerning Differences Between Means with Independent Samples

When it is reasonable to believe that the samples are from populations at least approximately normally distributed and that the samples are also independent of each other, we use the test of the differences in the means. We may assume that population variances are equal or unequal. However, our focus in discussing the test of the differences of means is using the assumption that the population variances are equal.

When we can assume that the two populations are normally distributed and that the unknown population variances are equal, we use a t-distributed test statistic based on independent random samples:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}},$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is a pooled estimator of the common variance. The number of degrees of freedom for this t-distributed test statistic is $n_1 + n_2 - 2$.

7.9 Tests Concerning Differences Between Means with Dependent Samples

When we want to conduct tests on two means based on samples that we believe are dependent, we use the test of the mean of the differences (test itself is sometimes referred to as the paired comparisons test). We begin by calculating $\bar{d}$ the sample mean difference, or the mean of the differences, d_i: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, where n is the number of pairs of observations. The sample standard deviation, s_d, is the standard deviation of the differences, and the standard error of the mean differences is $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$.

When we have data consisting of paired observations from samples generated by normally distributed populations with unknown variances, the t-distributed test statistic is: $t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$, with $n - 1$ degrees of freedom, where n is the number of paired observations.

Importantly, if we think of the differences between the two samples as a single sample, then the test of the mean of differences is identical to the test of a single sample mean.

7.10 Testing Concerning Tests of Variances (χ^2 Test)

Tests of a single variance

In tests concerning the variance of a single normally distributed population, we make use of a chi-square test statistic. Unlike the t-distribution, the chi-square distribution is bounded below by zero; it does not take on negative values. If we have n independent observations from a normally distributed population, the appropriate test statistic is: $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$ with $n - 1$ degrees of freedom. The sample variance (s^2) is in the numerator, and the hypothesized variance (σ_0^2) is in the denominator.

In contrast to the t-test, for example, the chi-square test is sensitive to violations of its assumptions. If the sample is not random or if it does not come from a normally distributed population, inferences based on a chi-square test are likely to be faulty.

Test concerning the equality of two variances (F-test)

Suppose we have a hypothesis about the relative values of the variances of two normally distributed populations with variances of σ_1^2 and σ_2^2. Given independent random samples from these populations, tests related to these hypotheses are based on an F-test, which is the ratio of sample variances.

Suppose we have two samples, the first with n_1 observations and a sample variance s_1^2 and the second with n_2 observations and a sample variance s_2^2 . The samples are random, independent of each other, and generated by normally distributed populations. A test concerning differences between the variances of the two populations is based on the ratio of sample variances, as follows: $F = \frac{s_1^2}{s_2^2}$ with $df_1 = (n_1 - 1)$ numerator degrees of freedom and $df_2 = (n_2 - 1)$ denominator degrees of freedom. When we rely on tables to arrive at critical values, a convention is to use the larger of the two sample variances in the numerator.

7.11 Parametric vs Non-parametric Tests

We primarily use nonparametric procedures in four situations: (1) when the data we use do not meet distributional assumptions, (2) when there are outliers, (3) when the data are given in ranks or use an ordinal scale, or (4) when the hypotheses we are addressing do not concern a parameter.

7.12 Tests Concerning Correlation

Parametric test of a correlation

If the two variables are normally distributed, we can test to determine whether the null hypothesis (H_0: $\rho = 0$) should be rejected using the sample correlation, r. The formula for the t-test is: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. This test statistic is t-distributed with $n - 2$ degrees of freedom.

The Spearman rank correlation coefficient

When we believe that the population under consideration meaningfully departs from normality, we can use a test based on the Spearman rank correlation coefficient. The calculation of r_S requires the following steps:

1. Rank the observations on X from largest to smallest. Assign the number 1 to the observation with the largest value, the number 2 to the observation with second largest value, and so on. In case of ties, assign to each tied observation the average of the ranks that they jointly occupy. For example, if the third and fourth largest values are tied, we assign both observations the rank of 3.5 (the average of 3 and 4). Perform the same procedure for the observations on Y.

2. Calculate the difference, d_i, between the ranks for each pair of observations on X and Y, and then calculate d_i^2 (the squared difference in ranks).

3. With n as the sample size, the Spearman rank correlation is given by: $r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

7.13 Test of Independence Using Contingency Table Data

When faced with categorical or discrete data, we cannot use the methods that we have discussed up to this point to test whether the classifications of such data are independent. Test of independence using a nonparametric test statistic that is chi-square distributed (having a contingency/frequency table):

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where:

- m - the number of cells in the table, which is the number of groups in the first class multiplied by the number of groups in the second class
- O_{ij} - the number of observations in each cell of row i and column j (i.e., observed frequency)
- E_{ij} - the expected number of observations in each cell of row i and column j, assuming independence (i.e., expected frequency)

This test statistic has $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns.

$$E_{ij} = \frac{(Total\ row\ i) * (Total\ column\ j)}{Overall\ total}$$

For the test of independence using a contingency table, there is only one rejection region, on the right side.

8 Introduction to Linear Regression

Sum of squares total (SST) = total sum of squares = variation of $Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

8.1 Estimating the Parameters of a Simple Linear Regression

A way that we often describe this simple linear regression relation is that Y is regressed on X.

Sum of squares error (SSE) = residual sum of squares = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (\hat{b}_0 + \hat{b}_1 X_i)]^2$

$\hat{b_1} = \frac{Cov(Y,X)}{Var(X)} = \frac{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})}{n-1}}{\frac{\sum_{i=1}^{n}(X_i-\bar{X})}{n-1}} = \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})}$

$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X}$

The slope is the change in the dependent variable for a one-unit change in the independent variable.

8.2 Assumptions of the Simple Linear Regression Model

1. Linearity: The relationship between the dependent variable, Y, and the independent variable, X, is linear.
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence: The observations, pairs of Ys and Xs, are independent of one another. This implies the regression residuals are uncorrelated across observations.
4. Normality: The regression residuals are normally distributed.

8.3 Analysis of Variance

Coefficient of determination = $R^2 = \frac{SSR}{SST}$

To see if our regression model is likely to be statistically meaningful, we will need to construct an F-distributed test statistic:

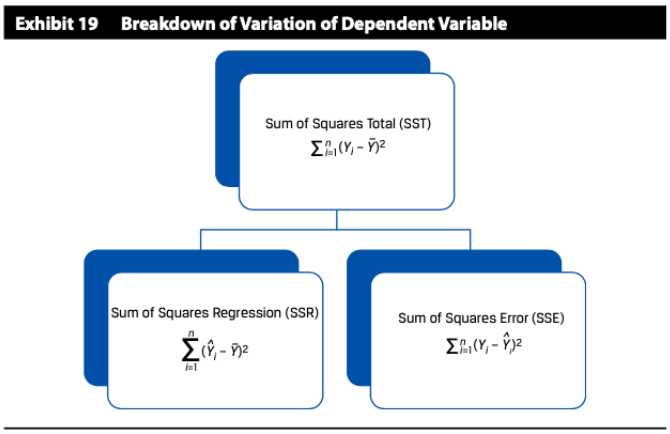

Figure 9: Breakdown of Variation of Dependent Variable

- $H_0 : b_1 = b_2 = ... = b_k = 0$
- H_a: at least one b_k is not equal to zero

We divide the sum of squares regression by the number of independent variables, represented by k. In the case of a simple linear regression, $k = 1$, so we arrive at the mean square regression (MSR), which is the same as the sum of squares regression:$MSR = \frac{SSR}{k}$.

Next, we calculate the mean square error (MSE), which is the sum of squares error divided by the degrees of freedom, which are $nk1$: $MSE = \frac{SSE}{n-k-1}$.

$F = \frac{MSR}{MSE}$

which is distributed with 1 and n - 2 degrees of freedom in simple linear regression. The F-statistic in regression analysis is one sided, with the rejection region on the right side, because we are interested in whether the variation in Y explained (the numerator) is larger than the variation in Y unexplained (the denominator).

The standard error of the estimate is an absolute measure of the distance of the observed dependent variable from the regression line:

$$s_e = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

8.4 Hypothesis Testing of Linear Regression Coefficients

To test a hypothesis about a slope: $t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$ (statistic is t-distributed with n - k - 1 or n - 2 degrees of freedom).

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

A feature of simple linear regression is that the t-statistic used to test whether the slope coefficient is equal to zero and the t-statistic to test whether the pairwise correlation is zero are the same value. Another interesting feature of simple linear regression is that the test-statistic used to test the fit of the model (that is, the F-distributed test statistic) is related to the calculated t-statistic used to test whether the slope coefficient is equal to zero: $t^2 = F$.

8.5 Prediction Using Simple Linear Regression and Prediction Intervals

The estimated variance of the prediction error:

$$s_f^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{\left(X_f - \bar{X}\right)^2}{(n-1)s_X^2} \right] = s_e^2 \left[1 + \frac{1}{n} + \frac{\left(X_f - \bar{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \right]$$

Standard error of the forecast:

$$s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{\left(X_f - \bar{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}}$$

Prediction interval: $\hat{Y}_f \pm t_{critical\ for\ \alpha/2} s_f$

8.6 Functional Forms for Simple Linear Regression

1. Log-Lin model: $lnY_i = b_0 + b_1 X_i$ (The slope coefficient in this model is the relative change in the dependent variable for an absolute change in the independent variable)

2. Lin-Log model: $Y_i = b_0 + b_1 lnX_i$ (The slope coefficient in this regression model provides the absolute change in the dependent variable for a relative change in the independent variable)

3. Log-Log model: $lnY_i = b_0 + b_1 lnX_i$ (This model is useful in calculating elasticities because the slope coefficient is the relative change in the dependent variable for a relative change in the independent variable)