Dr Arjun Anil Kumar 7375 words

APPLIED BUSINESS STATISTICS

by Dr Arjun Anil Kumar

SYNOPSIS: This book covers the syllabus of course MS6107E (Monsoon) - MBA I year. Book has been enriched by the feedback given by students in the previous years.

#

Chapter 1

Introduction to Probability [1]

Probability is the branch of study that quantifies the degree of uncertainty in a random experiment. Understanding basic probability concepts is key to better decision-making, as it assesses risks and predicts outcomes in uncertain situations. This chapter covers the basic building blocks of probability theory.

Definition 1 *Random Experiment*

If an experiment results in different outcomes, despite being repeated in the same manner every time, it is called a random experiment.

AB Testing for Marketing Campaigns : A company wants to understand which e-mail marketing strategy results in better customer engagement. The company randomly allots customers to two groups and sends e-mail version-1 to group-1 and e-mail version-2 to group-2. This is a random experiment as it has two outcomes. While the first e-mail version could be emphasizing on discount, the second e-mail version could be emphasizing on product benefits. For both the groups, there are two outcomes.

Outcome 1 - Customer clicks the advertisement

Outcome 2 - Customer does not click the advertisement.

Definition 2 *Sample Space*

The set of all possible outcomes of a random experiment is called the sample space (S) of the experiment.

AB Testing for Marketing Campaigns : The sample space for the AB Testing for Marketing campaigns is denoted by S = {Customer clicks the advertisement, Customer does not click the advertisement}. The instance of customer clicking the advertisement refers to outcome 1 and the instance of customer not clicking the advertisement refers to outcome 2.

Definition 3 *Discrete & Continuous Sample Space*

A sample space (S) is discrete if it possesses a finite or countable infinite set of outcomes.

The AB Testing of Marketing Campaigns is an example of finite discrete sample space S = {Customer clicks the advertisement, Customer does not click the advertisement}. The number of customers arriving at a store every hour, the sample space S = {0,1,3,2,...} is an example of discrete sample space with countable infinite set of outcomes.

A sample space is continuous if it contains an interval (either finite or infinite) of real numbers.

*The speed of a car on a highway, limited between 40 km/h and 100 km/h is an example of finite continuous sample space S = [40,100]. The time taken by a customer to complete a transaction is an example of infinite continuous sample space S = [0,*inf*).*

Definition 4 *Event*

An event (E) is a subset of the sample space (S) of a random experiment. For the AB Testing of Marketing Campaigns, event (E) refers to the instance where the customer responds to the e-mail by clicking the *advertisement.*

E = {Customer clicks the advertisement}

Definition 5 *Basic Set Operations*

- *The union of two events ($E_1 \cup E_2$) is the event that consists of all outcomes that are contained in either of the two events.*

  AB Testing for Marketing Campaigns : Assume a company is conducting an A/B test to measure the customer responses to two different e-mails with different marketing campaigns.

  Event E_1 : Customer in group 1 clicks the advertisement as a response to email-version 1.

  Event E_2 : Customer in group 2 clicks the advertisement as a response to email-version 2.

  The union of events ($E_1 \cup E_2$) represents the event where a customer clicks an advertisement link as a response to email-version 1 or email-version 2.

- *The intersection of two events ($E_1 \cap E_2$) is the event that consists of all outcomes that are contained in both of the two events.*

  AB Testing for Marketing Campaigns : There is no intersection between Event E_1 and Event E_2 as the AB testing for marketing campaigns is designed as mutually exclusive i.e. A customer in a group is exposed to only one version of the advertisement.

- *The complement of an event (E^c) in a sample space is the set of outcomes in the sample space that are not in the event.*

  Event E_1 : Customer in group 1 clicks the advertisement as a response to email-version 1.

Event E_1^c : Customer in group 1 does not click the advertisement as a response to email-version 1.

Event E_2 : Customer in group 2 clicks the advertisement as a response to email-version 2.

Event E_2^c : Customer in group 2 does not click the advertisement as a response to email-version 2.

Definition 6 *Two events, denoted as E_1 and E_2, such that are said to be mutually exclusive if and only if $E_1 \cap E_2 = \emptyset$*

Definition 7 *Counting Technique*

Assume an operation is defined as a sequence of k steps, and the number of ways of completing step 1 is n_1, the number of ways of completing step 2 is n_2 for each way of completing step 1, and the number of ways of completing step 3 is n_3 for each way of completing step 2, and so forth, the the total number of ways of completing the operation is $n_1.n_2..n_k$.

Definition 8 *Permutation*

*A permutation of the elements is an ordered sequence of the elements. The number of permutations of n different elements is n! where n! = n * (n-1) * (n-2) * (n-3) .. 2 * 1*

Definition 9 *Permutation of Similar Objects*

The number of permutations of n objects where $n = n_1 + n_2 + ..n_r$ n_1 are of one type, n_2 are of a second type and n_r are of an r^{th} type is $\frac{n!}{n_1!n_2!..n_r!}$.

Definition 10 *Combination*

The number of ways of selecting subsets of size r from a set of n elements, is denoted as $\binom{n}{r}$ = $\frac{n!}{(r!)(n-r)!}$

Definition 11 *Interpretations and Axioms of Probability [1]*

- *Probability quantifies the likelihood or the chance that an outcome of a random experiment will occur.*
- *The likelihood of an outcome is quantified by assigning a number from the interval [0, 1] to the outcome.*
- *Higher numbers or higher probability indicate that the outcome is more likely than other outcomes.*
- *The probability of an outcome is interpreted as the limiting value of the proportion of times the outcome occurs in n repetitions of the random experiment, as n increases beyond all bounds - Relative Frequency Interpretation of Probability.*
- *Whenever a sample space consists of N possible outcomes that are equally likely, the probability of each outcome is* $\frac{1}{N}$.

Definition 12 *Axioms of Probability [1]*

- *Probability is a number that is assigned to each event E of a random experiment that satisfies the following properties. If S is the sample space and E is any event in a random experiment.*

  1. $P(S) = 1$

  2. $0 \leq P(E) \leq 1$

  3. *For two events* E_1 *and* E_2 *with* $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

  4. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$

  5. $P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_2 \cap E_3) - P(E_1 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$

6.A collection of events, E_1, E_2, E_k, is said to be mutually exclusive if for all pairs, then P($E_1 \cup E_2 \cup ..E_k$) = $P(E_1) + P(E_2) + ...P(E_k)$

Definition 13 *Total Probability Rule - Two Events*

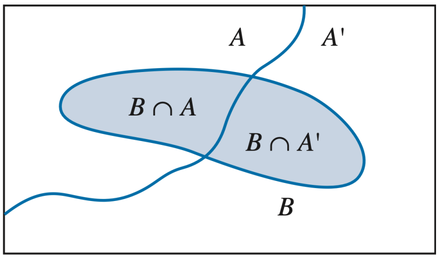

Figure 1.1: Partitioning an event B into two mutually exclusive events $A \cap B$ and $A^c \cap B$[1]

The event B can be represented as the union of two mutually exclusive events $A \cap B$ and $A^c \cap B$.

$$P(B) = P((A \cap B) \cup (A^c \cap B)) = P(B/A)P(A) + P(B/A^c)P(A^c) \tag{1.1}$$

Definition 14 *Conditional Probability*

The conditional probability of an event A given an event B, denoted as P($\frac{B}{A}$), is

$$P(\frac{A}{B}) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A).P(A)}{P(B)} \tag{1.2}$$

where $P(A) > 0$ and $P(B) = P(B/A)P(A) + P(B/A^c)P(A^c)$.

Definition 15 *Independent Events*

If the probability of an event A is not affected by the knowledge that another event B has already occured ($P(A/B) = P(A)$), than events A and B are said to be independent. If two events A and B are independent, then

$$P(A \cap B) = P(A).P(B) \tag{1.3}$$

Definition 16 *Bayes Theorem Bayes theorem updates the probability of a hypothesis (A) based on new evidence (B).*

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B/A).P(A)}{P(B)} \tag{1.4}$$

1. *Posterior probability P($\frac{A}{B}$) is the probability of hypothesis A given that B (evidence) has occurred.*
2. *Likelihood P($\frac{B}{A}$) is the probability of evidence B given that hypothesis A is true.*
3. *Prior probability P(A) is the initial probability of hypothesis A, prior to considering the evidence B.*
4. *Marginal probability P(B) is the total probability of evidence and* $P(B) > 0$ *and can be computed using Equation 1.1.*

Definition 17 *Random Variable*

A random variable is a function that assigns a real number to each outcome in the sample space of a random experiment. It could be discrete or continuous.

Definition 18 *Discrete Probability Distribution*

For a discrete random variable X with possible values x_1, x_2, .. x_n *, a probability mass function is a function such that*

$f(x_i) >= 0$

$\sum_1^n f(x_i) = 1$

$f(x_i) = P(X = x_i)$

Definition 19 *Continuous Probability Distribution*

For a continuous random variable X, a probability density function is a function such that

$f(x) >= 0$

$\int_{-\infty}^{\infty} f(x)dx = 1$

$P(a \leq X \leq b) = \int_a^b f(x)dx$ => *Area under f(x) from a to b*

Chapter 2

Hypothesis Testing

Definition 20 *Statistical Inference*

Statistical Inference helps to draw conclusions about population parameters from samples. Statistical Inference is divided into Two Major Areas, namely, Parameter Estimation & Hypothesis Testing.

Definition 21 *Parameter Estimation*

Parameter estimation helps to estimate population parameters such as mean μ, and variance σ^2 from samples. Parameter Estimation is done is two ways, namely, Point Estimation and Interval Estimation. Point estimate gets us a single point estimate of the population parameter and interval estimate gets us an interval that captures the population parameter.

Definition 22 *Point Estimation*

A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic T . The statistic T is called the point estimator (Table 2.1).

Definition 23 *Interval Estimation*

Parameter	Point Estimator
Mean μ	$\hat{\mu} = \frac{X_1+X_2...+X_n}{n}$
Two Group Means	$\hat{\mu}_1 - \hat{\mu}_2 = \frac{X_1+X_2...+X_{n_1}}{n_1} - \frac{Y_1+Y_2...+Y_{n_2}}{n_2}$
One Group Proportion	$\hat{p} = \frac{\sum_{i=1}^{n}[X_i]}{n}$
Two Group Proportions	$\hat{p_1} - \hat{p_2} = \frac{\sum_{i=1}^{n_1}[X_i]}{n_1} - \frac{\sum_{i=1}^{n_2}[Y_i]}{n_2}$
Variance σ^2	$\hat{\sigma}^2 = \frac{\sum_1^n[X_i - \hat{\mu}]^2}{n-1}$
Two Group Variance $\frac{\sigma_1{}^2}{\sigma_2{}^2}$	$\frac{\hat{\sigma_1}^2}{\hat{\sigma_2}^2} = \frac{\frac{\sum_1^{n_1}[X_i - \hat{\mu_1}]^2}{n_1-1}}{\frac{\sum_1^{n_2}[Y_i - \hat{\mu_2}]^2}{n_2-1}}$

Table 2.1: Point Estimators

The interval estimates for population mean μ, population variance σ^2 and population proportion p is shown in Table 2.2.

Parameter	CI	Remarks
μ	$P\left[\hat{\mu} - z_{\frac{\alpha}{2}}.\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\frac{\alpha}{2}}.\frac{\hat{\sigma}}{\sqrt{n}}\right] = 1 - \alpha$	$n > 40$
μ	$P\left[\hat{\mu} - t_{\frac{\alpha}{2},n-1}.\frac{s}{\sqrt{n}} \leq \mu \leq \hat{\mu} + t_{\frac{\alpha}{2},n-1}.\frac{s}{\sqrt{n}}\right] = 1 - \alpha$	$n < 40$, X $\sim \mathcal{N}(\mu, \sigma^2)$
p	$P[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} \leq p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}] = 1 - \alpha$	X $\sim$ Ber(p)
σ^2	$P\left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)\hat{\sigma}^2}{\chi^2_{\frac{-\alpha}{2},n-1}}\right] = 1 - \alpha$	X $\sim \mathcal{N}(\mu, \sigma^2)$

Table 2.2: Confidence Interval

Definition 24 *Concept of Confidence Interval of Population Mean*

For a sample of size n, the confidence interval (Figure 2.1) for population mean μ and population standard deviation σ with a confidence of 1 - α, where α is the level of significance, is given by

$$P\left[\bar{X} - K \leq \mu \leq \bar{X} + K\right] = 1 - \alpha \tag{2.1}$$

where $K = z_{\frac{\alpha}{2}} * SE$ and standard error SE $= \frac{\sigma}{\sqrt{n}}$

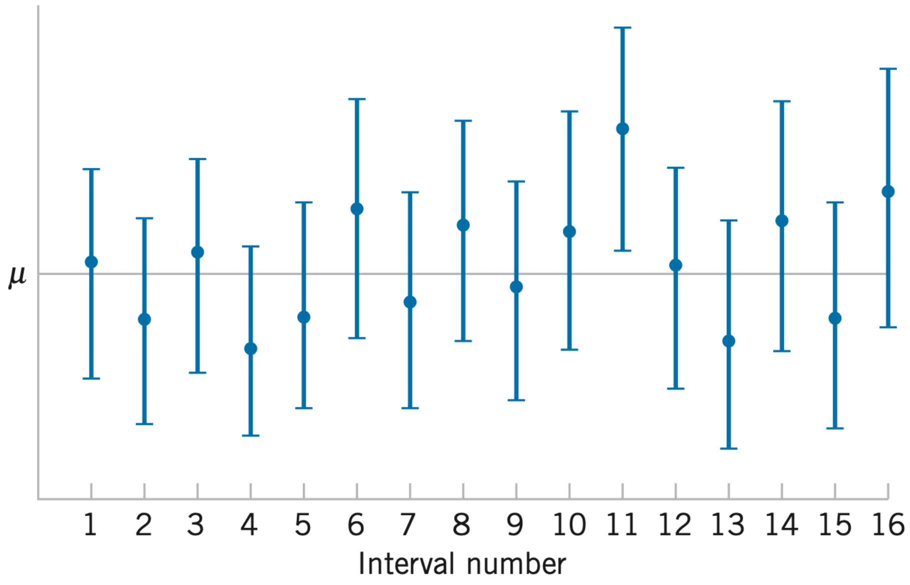

Figure 2.1: Confidence Interval - Frequentist Interpretation[1]

1. *A high confidence* $(1-\alpha)$ *results in high confidence interval (2K) and low precision.*

2. *Precision refers to how close the repeated measurements of point estimates are to each other. It represents the consistency of a point estimator in measuring the same result, when sampling is done repeatedly.*

3. *Precision is inversely related with Standard Error* $SE(\hat{\theta})$ *of the point estimate* $\hat{\theta}$*.*

4. *For a given confidence interval, we can improve the precision by increasing the sample size n.*

Definition 25 *Standard Error*

The Standard Error (SE) measures the variability of a sample statistic $T=\hat{\theta}$ *(e.g. sample mean, sample proportion, sample variance) from its expected value* $E[\hat{\theta}]$*.*

Definition 26 *Choice of Sample Size and Error The precision of an estimate (E) represents the difference between the estimate $\hat{\theta} - E[\hat{\theta}]$.*

1. *Precision of Estimate $\hat{\mu}$ of Population Mean μ*

$$P\left[\hat{\mu} - z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[- z_{\frac{\alpha}{2}} \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$P\left[|E| < z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

$$P\left[|E| < E_{max}\right] = 1 - \alpha \tag{2.2}$$

If $\hat{\mu}$ is the sample mean and μ is the population mean, we can be $100(1-\alpha)$% confidence that the error |$\hat{\mu} - \mu$| will not exceed an error E_{max}, if the minimum sample size n_{min} = $\left[\frac{z_{\frac{\alpha}{2}} * \sigma}{E_{max}}\right]^2$

Definition 27 *Hypothesis Testing & Sampling*

1. *A null hypothesis refers to a claim about a population parameter.*

2. *Hypothesis testing involves collecting data from randomly drawn samples to check for evidence that rejects the claim made about the population parameter.*

3. *In hypothesis testing, the pre-defined statistic has to be computed and checked whether it falls in the rejection region.*

4. *Samples are drawn randomly from a population.*

5. *Samples are independent of each other or the probability of occurence of the first sample does not have an impact on the probability of occurence of any other sample.*

6. *The rejection region has been set for a pre-defined level of significance (α).*

7. *If the pre-defined statistic does not fall into the rejection region, the data does not provide enough evidence to reject the claim regarding the null hypothesis.*

8. *The alternate hypothesis is the complement of the null hypothesis in the case of a two-tailed test.*

9. *The alternate hypothesis has to be carefully framed based on the problem statement in the case of one-tailed test.*

Definition 28 *Hypothesis Testing - Examples*

1. *Eg : The average marks scored in Statistics exam is 54 or μ=54 (One sample - Means Test)*

2. *Eg : Proportion of students who get placed in two companies is 20% or p = .2 (One sample - Proportions Test)*

3. *Eg : The difference of means between salary of employees with 5 years and employees with 7 years experience post MBA is 45000 Rs a month or $\mu_{G1} - \mu_{G2}$ =45000 (Two sample - Means Test)*

4. *Eg : The variance of salary of men in Kinfosys is 250000 or σ^2 = 250000 (One Sample - Variance Test)*

5. *Eg : The variance of salary of men and variance of salary of women in Kinfosys are equal or ${\sigma_{G1}}^2$ = ${\sigma_{G2}}^2$ (Two Samples - Variance Test)*

Definition 29 *Central Limit Theorem[1]*

If X_1, X_2, .., X_n is a random sample of size n taken from a population (either finite or infinite) with parameters mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean or estimate of the mean, the limiting form of the distribution of random variable Z follows a standard normal distribution Z $\sim$ $\mathcal{N}(0,1)$, as $n->\infty$.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.3}$$

The Central Limit Theorem (CLT) tells us that the sampling distribution of the sample mean is, at least approximately, normally distributed, regardless of the distribution of the underlying random sample, with a mean (μ) and standard deviation $\frac{\sigma}{\sqrt{n}}$.

- *Sample Mean $\bar{X} = \sum_1^n X_i$ is a random variable and follows a normal distribution with mean equal to population mean and standard variance equal to population standard variance divided by value of n or $\bar{X}$ $\sim$ $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.*

- *If n >= 30, the distribution of the sample mean would be always normal irrespective of the distribution of the parent distribution from which X_1, X_2, .., X_n are sampled.*

- *If n < 30, the central limit theorem will work only if the distribution of the parent distribution from which X_1, X_2, .., X_n are sampled are not severly non-normal. Ideally, they should have been normal.*

- *If the distribution of the X_i, i from 1 to n, is symmetric, unimodal or continuous, then a sample size n as small as 4 or 5 yields an adequate approximation.*

- *If the distribution of the X_i, i from 1 to n, is skewed, then a sample size of at least 25 or 30 yields an adequate approximation.*
- *If the distribution of the X_i, i from 1 to n, is extremely skewed, then you may need an even larger.*
- *For the standardized normal variable $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ with mean 0 and variance 1, the probability that the variable Z takes a value between $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ is given by*

$$P\Bigg(- z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\Bigg) = 1 - \alpha \tag{2.4}$$

where α is the significance level that ranges between 0 and 1. Note that (1-α) is called the confidence level.

The statistical test broadly fall into two categories, parametric and non-parametric tests. While parametric test make lot of assumption about the underlying population distribution, non-parametric tests make minimal assumptions about the underlying population distribution.

Definition 30 *Parametric Tests - One Group*

1. *Z-Test for large samples ($n > 40$) is used for hypothesis testing for population mean with a known population standard deviation σ(Unrealistic). The null H_o and alternate hypothesis H_a for the-tailed and two-tailed tests are provided in Table 2.3.*

   The test statistic Z is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.5}$$

Left Tailed Test	Right Tailed Test	Two Tailed Test
$H_O : \mu = \mu_0$	H_O: $\mu = \mu_0$	$H_O : \mu = \mu_0$
$H_a : \mu < \mu_0$	H_a: $\mu > \mu_0$	$H_a : \mu \neq \mu_0$
A : $Z > Z_{-\alpha}$	A : $Z < Z_{\alpha}$	A : $Z_{-\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$
R : $Z < Z_{-\alpha}$	A : $Z > Z_{\alpha}$	A : $Z < Z_{-\frac{\alpha}{2}}$ or $Z > Z_{\frac{\alpha}{2}}$
A : $T > T_{-\alpha,n-1}$	A : $T < T_{\alpha,n-1}$	A : $T_{-\frac{\alpha}{2},n-1} < T < T_{\frac{\alpha}{2},n-1}$
R : $T < T_{-\alpha,n-1}$	A : $T > T_{\alpha,n-1}$	A : $T < T_{-\frac{\alpha}{2},n-1}$ or $T > T_{\frac{\alpha}{2},n-1}$

Table 2.3: Z-Test and T-Test - One Group

where σ is population standard deviation and n is the sampling size. Z is computed assuming that the null hypothesis H_o is true. The critical region is selected based on the Z-Distribution with confidence level of 100*α, α is the level of significance. If Z falls in the rejection region (R), we reject the null hypothesis. If Z falls in the acceptance region (A), we fail to reject the null hypothesis.

2. T-Test for small samples ($n < 40$) is used for hypothesis testing for population mean with unknown population standard deviation (Realistic). The null H_o and alternate hypothesis H_a for one-tailed and two-tailed tests are provided in Table 2.3.

   The test statistic T is given by

$$T = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{2.6}$$

   where $\hat{\sigma}$ is the sample standard deviation, $\hat{\sigma}$ is the point estimate of the population standard deviation (σ) and n is sampling size. T is computed assuming that the null hypothesis H_o is true. The critical region is selected based on the T-Distribution with n-1 degrees of freedom and confidence level 100*α, α is the level of significance. If T falls in the rejection region, we reject the null hypothesis.

If T falls in the acceptance region, we fail to reject the null hypothesis.

3. Proportion Test is for testing the hypothesis on proportion of samples satisfying a criteria. The hypothesis for proportion test is given in Table 2.4. The test statistic Z is given by

Left Tail Test	Right Tail Test	Two Tail Test
$H_O : p = p_0$	H_O: $p = p_0$	$H_O : p = p_0$
$H_a : p < p_0$	H_a: $p > p_0$	$H_a : p \neq p_0$
A : $Z > Z_{-\alpha}$	A : $Z < Z_\alpha$	A : $Z_{-\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$
R : $Z < Z_{-\alpha}$	A : $Z > Z_\alpha$	A : $Z < Z_{-\frac{\alpha}{2}}$ or $Z > Z_{\frac{\alpha}{2}}$

Table 2.4: Hypothesis Testing for Population Proportion - One Group

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}} \tag{2.7}$$

where $\hat{p}$ is the sample proportion estimate, p_o is the proportion of samples if the null hypothesis H_o is true and n is the number of samples. When the statistic Z falls in the rejection region (R), there is enough evidence in the data to reject the null hypothesis. When the statistic Z falls in the acceptance region (A), there is not enough evidence to reject the null hypothesis.

4. χ^2 Test for Population Variance is used for checking the claim made about the population variance. The statistic χ^2 is a weighted representation of the point estimate of the population variance where the weight is $\frac{n-1}{\sigma_o^2}$, where n is the number of samples and σ_o^2 is the variance claimed in the null-hypothesis, as shown in Table 2.5. The χ^2 statistic is computed from the samples using

$$\chi^2 = \frac{(n-1)\hat{\sigma}^2}{{\sigma_0}^2} \tag{2.8}$$

Left Tailed Test	Right Tailed Test	Two Tailed Test
$H_O : \sigma^2 = \sigma_o^2$	H_O: $\sigma^2 = \sigma_o^2$	$H_O : \sigma^2 = \sigma_o^2$
$H_a : \sigma^2 < \sigma_o^2$	H_a: $\sigma^2 > \sigma_o^2$	$H_a : \sigma^2 \neq \sigma_o^2$
A : $\chi^2 > \chi^2_{-\alpha,n-1}$	A : $\chi^2 < \chi^2_{\alpha,n-1}$	A : $\chi^2_{-\frac{\alpha}{2},n-1} < \chi^2 < \chi^2_{\frac{\alpha}{2},n-1}$
R : $\chi^2 < \chi^2_{-\alpha,n-1}$	A : $\chi^2 > \chi^2_{\alpha,n-1}$	A : $\chi^2 < \chi^2_{-\frac{\alpha}{2},n-1}$ or $\chi^2 > \chi^2_{\frac{\alpha}{2},n-1}$

Table 2.5: Hypothesis Testing for Population Variance

assuming that the null hypothesis H_O is true. If χ^2 falls in the rejection region (R), we reject the null hypothesis. If χ^2 falls in the acceptance region (A), we fail to reject the null hypothesis.

Definition 31 *Parametric Tests - Two Group*

1. *Z-Test for large samples ($n1 > 40, n2 > 40$) is used for hypothesis testing for difference of means with variances known. We define $X_1, X_2, , X_{n_1}$ as a random sample of size n_1 drawn from population 1. We define $Y_1, Y_2, , Y_{n_2}$ as a random sample of size n_2 drawn from population 2. Z-Test assumes that the samples are drawn independently from two normal populations represented by $X \sim \mathcal{N}(\mu_1, \sigma^2)$and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$. The purpose of the Z-test is to draw inferences on the difference of population means $\mu_1 - \mu_2$, where μ_1 and μ_2 are the unknown population means of the two populations. The point estimator of difference of population means is $\hat{\mu_1} - \hat{\mu_2}$, that has a mean of $\mu_1 - \mu_2$ and variance $\frac{{\sigma_1}^2}{n_1}$ +$\frac{{\sigma_2}^2}{n_2}$ or*
   $\hat{\mu_1} - \hat{\mu_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$.
   Statistic Z for the difference of means test has a normal distribution with mean 0 and variance 1.

$$Z = \frac{\hat{\mu_1} - \hat{\mu_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{2.9}$$

Table 2.6 shows the criteria for rejecting and accepting the null

Left Tailed Test	Right Tailed Test	Two Tailed Test
$H_O : \mu_1 = \mu_2$	H_O: $\mu_1 = \mu_2$	$H_O : \mu_1 = \mu_2$
$H_a : \mu_1 < \mu_2$	H_a: $\mu_1 > \mu_2$	$H_a : \mu_1 \neq \mu_2$
A : $Z > Z_{-\alpha}$	A : $Z < Z_{\alpha}$	A : $Z_{-\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$
R : $Z < Z_{-\alpha}$	A : $Z > Z_{\alpha}$	A : $Z < Z_{-\frac{\alpha}{2}}$ or $Z > Z_{\frac{\alpha}{2}}$
A : $T > T_{-\alpha,n-1}$	A : $T < T_{\alpha,n-1}$	A : $T_{-\frac{\alpha}{2},n-1} < T < T_{\frac{\alpha}{2},n-1}$
R : $T < T_{-\alpha,n-1}$	A : $T > T_{\alpha,n-1}$	A : $T < T_{-\frac{\alpha}{2},n-1}$ or $T > T_{\frac{\alpha}{2},n-1}$

Table 2.6: Z-Test and T-Test - Two Group

hypothesis of two sample mean tests.

2. *T-Test for small samples (*$n1 > 40, n2 > 40$*) is used for hypothesis testing for difference of means with variances unknown. We define* $X_1, X_2, , X_{n_1}$ *as a random sample of size* n_1 *drawn from population 1. We define* $Y_1, Y_2, , Y_{n_2}$ *as a random sample of size* n_2 *drawn from population 2. The samples are drawn independently from two normal populations represented by* $X \sim \mathcal{N}(\mu_1, \sigma^2)$*and* $Y \sim \mathcal{N}(\mu_2, \sigma^2)$*.* $\hat{\mu_1} - \hat{\mu_2}$ *has a mean of* $\mu_1 - \mu_2$ *and variance* $\frac{\sigma_1{}^2}{n_1}$ *+*$\frac{\sigma_2{}^2}{n_2}$*.*

   For pooled T-test, we assume that the unknown population variances are equal or $\sigma_1{}^2 = \sigma_2{}^2$ *or* $\hat{\sigma_1}^2 = \hat{\sigma_2}^2$*. We define a pooled estimator of* σ^2*,* $\hat{\sigma}^2_{pooled}$

$$\hat{\sigma}^2_{pooled} = \frac{(n_1 - 1)\hat{\sigma_1}^2 + (n_2 - 1)\hat{\sigma_2}^2}{n_1 + n_2 - 2} \tag{2.10}$$

   The pooled variance represents a weighted sum of the estimated variances of the two groups. The test statistic for the pooled-T test follows a T distribution with n1 + n2 - 2 degrees of freedom.

$$T = \frac{\hat{\mu_1} - \hat{\mu_2} - (\mu_1 - \mu_2)}{\hat{\sigma}_{pooled} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.11}$$

   For an unpooled T-test, the test statistic would be T-distributed with v degrees of freedom.

$$T = \frac{\hat{\mu_1} - \hat{\mu_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{\sigma_1}^2}{n_1} + \frac{\hat{\sigma_2}^2}{n_2}}} \tag{2.12}$$

$$v = \frac{\left(\frac{\hat{\sigma_1}^2}{n_1} + \frac{\hat{\sigma_2}^2}{n_2}\right)^2}{\frac{\left[\frac{\hat{\sigma_1}^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{\hat{\sigma_2}^2}{n_2}\right]^2}{n_2 - 1}} \tag{2.13}$$

3. Proportion Test for Two Groups is used to check whether there is a significant difference of proportion between two groups. The conditions for rejecting and accepting the null-hypothesis in two group proportion testing is given in Table 2.7. The statistic Z

$$Z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\frac{p_1.(1-p_1)}{n_1} + \frac{p_2.(1-p_2)}{n_2}}} \tag{2.14}$$

where $\hat{p_1}$ and $\hat{p_2}$ are estimates of proportions of the two groups respectively and p_1 and p_2 are the proportions in the null-hypothesis.

Left Tail Test	Right Tail Test	Two Tailed Test
$H_O : p_1 = p_2$	H_O: $p_1 = p_2$	$H_O : p_1 = p_2$
$H_a : p_1 < p_2$	H_a: $p_1 > p_2$	$H_a : p_1 \neq p_2$
A : $Z > Z_{-\alpha}$	A : $Z < Z_\alpha$	A : $Z_{-\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$
R : $Z < Z_{-\alpha}$	A : $Z > Z_\alpha$	A : $Z < Z_{-\frac{\alpha}{2}}$ or $Z > Z_{\frac{\alpha}{2}}$

Table 2.7: Proportion Test - Two Group

4. F-Test is used for checking whether the difference of population variance between two groups is significant. We define $X_1, X_2, , X_{n_1}$ as a random sample of size n_1 drawn from population 1. We define $Y_1, Y_2, , Y_{n_2}$ as a random sample of size n_2 drawn from population 2. F-Test assumes that the samples are drawn independently from two

normal populations represented by $X \sim \mathcal{N}(\mu_1, \sigma^2)$and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$. The F-statistic is the ratio of point estimates of variances of the groups. The conditions for rejecting and accepting the null hypothesis of population variances is provided in Table 2.8.

$$F = \frac{\hat{\sigma_1}^2}{\hat{\sigma_2}^2} = \frac{\frac{\sum_1^{n_1}[X_i - \hat{\mu_1}]^2}{n_1 - 1}}{\frac{\sum_1^{n_2}[Y_i - \hat{\mu_2}]^2}{n_2 - 1}} \tag{2.15}$$

Left Tailed Test	Right Tailed Test	Two Tailed Test
$H_O : \sigma_1^2 = \sigma_2^2$	H_O: $\sigma_1^2 = \sigma_2^2$	$H_O : \sigma_1^2 = \sigma_2^2$
$H_a : \sigma_1^2 < \sigma_2^2$	H_a: $\sigma_1^2 > \sigma_2^2$	$H_a : \sigma_1^2 \neq \sigma_2^2$
A : $F > F_{-\alpha, n_1-1, n_2-1}$	A : $F < F_{\alpha, n_1-1, n_2-1}$	A : $F_{-\frac{\alpha}{2}, n_1-1, n_2-1} < F < F_{\frac{\alpha}{2}, n_1-1, n_2-1}$
R : $F < F_{-\alpha, n_1-1, n_2-1}$	A : $F > F_{\alpha, n_1-1, n_2-1}$	A : $F < F_{-\frac{\alpha}{2}, n_1-1, n_2-1}$ or $F > F_{\frac{\alpha}{2}, n_1-1, n_2-1}$

Table 2.8: F-Test - Population Variance of Two Groups

Chapter 3

Practice Problems

1. Suppose there is a medical test for a rare disease D of the brain. We are given the following information on the outcome of the test T and the disease D.

   - Prevalence of disease or the probability of having disease D ($P(D)$): 1% ($P(D) = 0.01$)
   - Test sensitivity or the probability of getting a positive test result given that you have the disease ($P(T^+|D)$): 95% ($P(T^+|D) = 0.95$)
   - Test specificity or the probability of getting a negative test result given that you dont have the disease ($P(T^-|D^c)$): 90% ($P(T^-|D^c) = 0.90$)

Given a positive test result (T^+), report the probability that the patient actually has the disease ($P(D|T^+)$).

Bayes' Theorem is given by:

$$P(D|T^+) = \frac{P(T^+|D) \cdot P(D)}{P(T^+)}$$

The total probability of a positive test result ($P(T^+)$) is calculated as:

$$P(T^+) = P(T^+|D) \cdot P(D) + P(T^+|D^c) \cdot P(D^c)$$

Substituting the values:

$$P(T^+) = (0.95)(0.01) + (0.10)(0.99)$$

$$P(T^+) = 0.0095 + 0.099 = 0.1085$$

Substitute $P(T^+)$ and the other known values into Bayes' Theorem:

$$P(D|T^+) = \frac{P(T^+|D) \cdot P(D)}{P(T^+)}$$

$$P(D|T^+) = \frac{(0.95)(0.01)}{0.1085}$$

$$P(D|T^+) \approx 0.0876 \text{ (8.76\%)}$$

Hence, the probability that you have the disease given that the test result is positive is only 8.76%.

Bibliography

[1] Douglas C Montgomery and George C Runger. Applied statistics and probability for engineers. John wiley & sons, 2010.

# # # # #