Dr Arjun Anil Kumar 4425 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 uncertainity in a random experiment. Understanding basic concepts of probability helps in better decision making. For example, you may predict the chance that a customer is going to pay back the loan based on features such age, gender, income.

Definition 1 *If at all an experiment results in different outcomes, despite being repeated in the same manner every time, then it is called a random experiment.*

Definition 2 *We define the set of all possible outcomes of a random experiment as sample space (S) of the experiment. For example, if we toss a coin, the sample space S = H,T, where H,T refers to the event of obtaining head and tail respectively.*

Definition 3 *A sample space (S) is discrete if it possesses a finite or countable infinite set of outcomes. A sample space is continuous if it contains an interval (either finite or infinite) of real numbers.*

Definition 4 *An event (E) is a subset of the sample space (S) of a random experiment.*

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.*
- *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.*
- *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.*

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

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, and 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 *Permuation 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 selectiong 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*

- *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*

- *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*

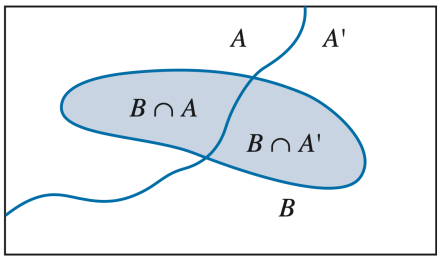

Figure 1.1: Partitioning an event B into two mutually exclusive events ($A \cap B$ and $A^c \cap B$

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 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 methods are used to make decisions and draw conclusions about populations. This aspect of statistics is generally called statistical inference. Statistical Inference is divided into Two Major Areas, namely, Parameter Estimation & Hypothesis Testing.

Definition 21 *Hypothesis Testing & Sampling*

A hypothesis refers to a statement made about a parameter of the population from which the random sample has been drawn and samples are drawn from a population to check the veracity of the hypothesis. For testing the hypothesis, parameter estimation is followed by computation of appropriate statistic defined for the hypothesis test.

Definition 22 *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 23 *Central Limit Theorem*

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* $n- > \infty$.

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{2.1}$$

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*

$$P\bigg(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}\bigg) = 1 - \alpha \tag{2.2}$$

where α is the significance level that ranges between 0 and 1.

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 24 *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).

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

Definition 25 *Confidence Interval - Mean*

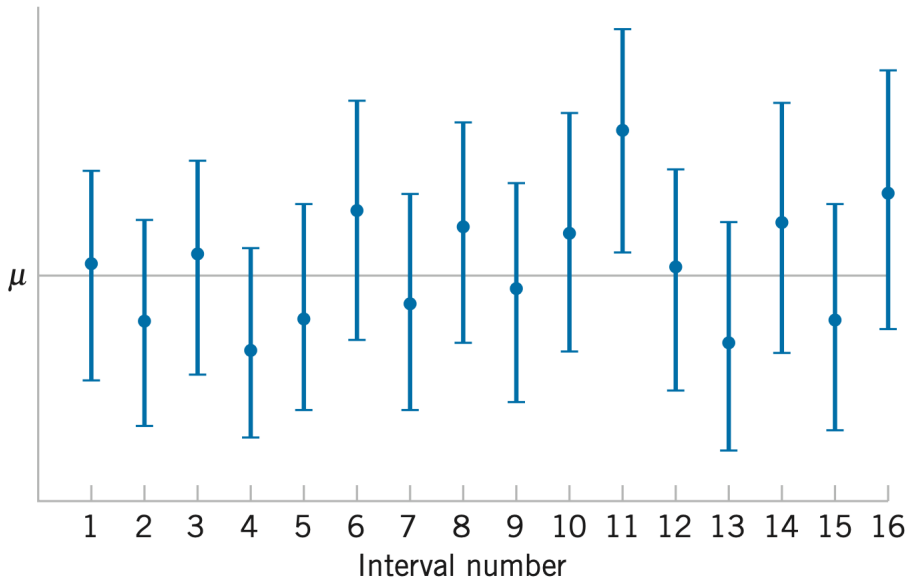

Figure 2.1: Confidence Interval - Frequentist Interpretation

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

$$P\left[\bar{X} - K \leq \mu \leq \bar{X} + K\right] = 1 - \alpha \tag{2.3}$$

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

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 26 *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* $E[\hat{\theta}]$.

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.

# # # # #