Select Statistically

Srilakshminarayana, G.

Introduction

Over the past few years, the world has seen changes in our buying behavior with respect to different products. Most of us are busy in our day-to-day routine and find no time to leisurely go to a shop and purchase the items we want. We look for an easy mode to purchase the items we need. E-commerce sites have provided us an opportunity to select the items we need online and place the order which will be delivered at our door steps. Among various items, one important item that most of us prefer to purchase online is a good book. When we think of a good, we remember that a good book is a good friend which guides us in our daily routine. The first e-commerce site that we hear now-a-days, atleast in the Indian context is Flipkart.

Flipkart provides service in various categories starting from books to kitchen appliances. They have developed a system that helps us to select the item we want and book the order by paying it online or cash on delivery. They clearly mention number of business days taken to deliver an order, based on the item we choose. They provide discounts on the items and also provide warranty to the items purchased. They have good delivery system, which takes care of the safety of the items and delivers exactly on time. Confirmation both through email and SMS reaches us before and after the delivery of the items.

Among various categories, it has got good name with respect to books. This is one reason that we have considered only books. This category has several other categories and we have considered only academic and professional category for the study.

The case writer Dr. Srilakshminarayana, Assistant Professor, SDMIMD, Mysore may be reached at lakshminarayana@sdmimd.ac.in. This case is fictionalized and any resemblance to actual person or entities is coincidental. This publication may not be digitized, photocopied, or otherwise reproduced, posted, or transmitted, without the permission of SDM RCMS, SDMIMD, Mysore. For Teaching Notes please contact sdmrcms@sdmimd.ac.in.

Objective

The objective of the case is to introduce Statistical estimation and testing tools which are commonly used both at corporate and academic levels. It gives steps to use a particular tool along with its limitations. We show how a method fails if the assumptions are not satisfied and provide an alternative tool. The data used to explain the statistical tools is the secondary data collected from Flipkart's website.

The following are the techniques that are discussed in this case study:

1. Estimation: Point and interval.
2. Testing of hypothesis:
   a. Parametric tests: t-test for single mean, Paired t-test, ANOVA.

   b. Non-Parametric tests: Wilcoxon-Signed rank test, Kruskal-Wallis test.

   c. Chi-square test for independent of attributes.

Note that the emphasis is more on the methods used and care has been taken to collect sufficient sample from Flipkart's website.

What Should I Do Now?

Mr.LN looked at his watch and felt that it is time for him to visit a book shop and purchase the book he is looking for. But the constraint is that he can't leave his office at that time and learnt that the books he is looking for is not available in the shop and has to wait for a week to get the book. At that time he found Mr.JJ walking to his cabin to call for a cup of tea. Mr.JJ understood that his friend is need of the book very badly and suggested him to get it by placing an order online. Upon hearing this Mr.LN is surprised and looked at Mr.JJ and asked him about placing the order online. Here starts the actual story of Flipkart. Interestingly after learning from Mr.JJ, about flipkart Mr.LN used his statistical knowledge to analyze the services of Flipkart. Go ahead and read.

Flipkart

Let us know briefly about Flipkart.

Flipkart.com is an Indian e-commerce company headquartered in Bangalore, Karnataka. Flipkart was founded in 2007 by Sachin Bansal and Binny Bansal, both alumni of the Indian Institute of Technology Delhi. They worked for Amazon.com before quitting and founding their own company. Initially they used word of mouth marketing to popularize their company. A few months later, the company sold its first book on flipkart.com—John Woods' Leaving Microsoft to Change the World. Today, as per Alexa traffic rankings, Flipkart is among the top 30 Indian Web sites and has been credited with being India's largest online bookseller with over 11 million titles on offer. Flipkart broke even in March 2010 and claims to have had at least 100% growth every quarter since its founding. The store started with selling books and in 2010 branched out to selling CDs, DVDs, mobile phones and accessories, cameras, computers, computer accessories and peripherals, and in 2011 Pens & stationery, other electronic items such as home appliances, kitchen appliances, personal care gadgets, health care products etc. Further in 2012, Flipkart added A.C, Air coolers, School supplies, Office supplies, Art Supplies & life style products to its product portfolio. As of today, Flipkart employs over 4500 people. Source: Wikipedia

Services of Flipkart

Flipkart's service is a very well established system, which caters the needs of the customers systematically. Flipkart caters needs of the customers in different categories. They include books, home appliances, computers, toys etc. A customer can choose the product online and can make the payment by choosing one of the options. Once an order is placed, the customer receives an SMS and an email confirming the order. The customer receives a confirmation mail after the shipment of the order and sends an SMS before delivery. The product will be delivered before 7.00 pm and an email will be sent after delivering the order.

A Good Book is a Good Friend

In this case, we have considered only books. Why did we consider only books? The answer is simple. According to us a good book is a good friend. Many would choose an e-commerce company like Flipkart, most of the times, to purchase a book. We have considered the category- books and tried to explore some of the important aspects related to the same. It is interesting to note that there are several sub categories under books. To know more, one can visit the http://www.flipkart.com/all-categories-books. Since we are into academics, we have focused more on sub-category-academic and professional. Among these, we have considered only few categories for comparison, as other categories can be studied on similar lines. This is because our emphasis is on using statistical tools rather than the categories or sub-categories.

Statistical Analysis of Academic and Professional Books

What are we going to do in this section? We introduce the statistical concepts and use them appropriately to compare various sub-categories in books category. We mainly concentrate on cost of the books before discount, cost of the books after discount, discount offered. They form the random variables in the case. We also provide justification to each method used.

Methodology

As we all know that it is difficult to study the entire population, we have considered a random sample for our study. We have adopted stratified random sampling technique to collect our sample from Flipkart's website. We first start with sample size determination i.e. estimating the sample size required for the study. To do this we start with a pilot study. Before explaining the pilot study, we explain the stratified random sampling technique in brief.

Stratified Random Sampling

The method starts with dividing the heterogeneous population into homogeneous subgroups called as strata. After this, we distribute the sample size *'n',* using an allocation procedure, across the strata and select the sample. Here *'n'* is divided into $n_1, n_2, \ldots, n_k$ where, n_i represent the size of *i-th* stratum. Using simple random sampling, we draw units from each stratum, such that total sample size is equal to *n.*

In this case study, we have considered each discipline as a stratum and used proportional allocation to allocate the units to each stratum. To understand this sampling design better, one can refer to Cochran (2007).

The Pilot Study

Before the actually survey is conducted, a pilot study is conducted to understand the dynamics of the population characteristics under study. There are many authors who discussed the importance of a pilot study before the actual study. Sample size determination usually starts with deciding the variable that needs to be focused. In our study, the average cost of a book is the key population characteristic. Hence we consider the sample average cost $(\overline{X})$ of a book as the estimator of the population average cost (μ) of a book. We fix the level of significance (α) as 5% and the desired level of precision i.e. the absolute difference between $\overline{X}$ and μ as Rs.10.

The sample size for pilot study is taken as 50. Since there are several categories under academic and professional category, we have considered few categories for pilot study. The categories are selected based on the number of books. Preference has been given to the category with more number of books. For example, Psychology, Political Science, Social Science, Medical, and General are given preference, as they have more number of books as compared with others. This is only for convenience. Again under each category there are several sub-categories. The sub-categories with more number of books have been considered for drawing a pilot sample. This is one way of drawing a pilot sample.

When compared to the actual size of each category, the allotted size seems to be small. Since this is only a pilot study, we restrict the pilot sample size to 50. This is in turn used to estimate the mean and standard deviation, which are used to determine the sample size. We assume that, unit cost has incurred to collect each sample point.

Determination of the Sample Size

In order to determine the sample size, a grouped mean and standard deviation are calculated as an unbiased estimatorsof the population mean and standard deviation, respectively. Recall that the quantitative variable under study is the average cost of a book.

While collecting the data, care has been taken to avoid outliers or extreme observations. This is because they will have an impact on the estimators and on subsequent analysis, results. For example, books under each category are collected such that the cost of the book is in between 700 and 1000. This is called as trimming, which is applied below and above. One can question the Unbiasedness of the sample. Omission is done after collecting the sample first at random and then observations are replaced by another set of observations.

Now, recall that we have fixed the level of significance as 5% and the desired level of precision as **Rs.10.** Hence the required sample size is 300. The formula used to estimate the sample size is

$$n = \frac{Z^2_{\frac{\alpha}{2}} \hat{\sigma}^2}{E^2}$$

where,

$Z_{\frac{\alpha}{2}}$: Value of the standard normal variable at chosen confidence level (1-α).

$\hat{\sigma}^2$: Sample variance.

E : Desired the level precision.

Note that the sample size estimated using this formula gives an idea about the sample size required for the actual survey. The sample size obtained should be checked for feasibility. That it, we need to check whether we can collect the sample from the source defined. Most of the times, it is difficult to identify the source from where the data can be collected. Also, note that the sample drawn is used to study the population parameters, which are unknown. Hence, the sample drawn should be taken, systematically by adopting a random sampling design. This design should be clearly defined before drawing the sample. It is interesting to note that the precision used to estimate the sample size indicates the distance between the sample mean and the population mean. In this case, the precision assumed is **Rs.10**. This is chosen by taking into consideration, the feasibility of colleting the sample from the Flipkart's website.

This section is introduced mainly to discuss the procedure used to estimate the sample size required for estimating the population parameters with desired precision. The sample size of 300 is distributed to the strata (categories) using proportional allocation.

Proportional Allocation

The category Psychology has 65883 books. The sample size allotted is (65883/570043)*300=35. Similarly, other sample sizes are allotted based on stratum size.

Sl. No.	Category (Stratum)	Stratum Size	Sample size allocated
1	Psychology	65883	35
2	Political Science	111832	59
3	Social Science	162230	85
4	Medical	134917	71
5	General	95181	50
	Total	**570043**	**300**

Psychology

Under this category, there are 33 sub-categories. We have considered the four sub-categories based on the number of books. Now, the sample size 35 has been distributed to each sub-category using proportional allocation as under.

Sl. No.	Category (Stratum)	Stratum Size	Sample size allocated
1	General	17260	16
2	Psychotherapy	11119	10
3	Movements	4357	4
4	Cognitive Psychology	5165	5
	Total	37901	35

Political Science

Under this, there are 23 sub-categories. Based on the number of books, we have selected again 4 subcategories and used proportional allocation to distribute the sample size 294.

Sl. No.	Category (Stratum)	Stratum Size	Sample size allocated
1	General	29445	21
2	International relations	21504	16
3	Political ideologies	13105	10
4	Public Policy	16218	12
	Total	**80722**	**59**

Social Science

Based on the number of books in each sub-category, seven categories are considered from 52 sub-categories.

Sl.	Category (Stratum)	Stratum Size	Sample size allocated
1	Anthropology	22642	21
2	Emigration and Immigration	4156	4
3	Ethnic Studies	13675	13
4	Folklore & Mythology	6168	6
5	Media Studies	9429	9
6	Sociology	30062	28
7	Homosexuality	4024	4
	Total	**90156**	**85**

Medical

From available 24 categories, 4 sub-categories are considered. The allocation is shown in the following table:

Sl. No.	Category (Stratum)	Stratum Size	Sample size allocated
1	Surgery	11259	11
2	Basic Sciences	14157	13
3	Medicine	32388	30
4	Para Clinic	18726	17
	Total	**76530**	**71**

General

This category does not have sub-categories and hence the sample of size 50 is collected directly.

Descriptive statistics of the sample data collected

Category	Sample Size	Sample (Before	Standard Deviation	Sample (After)	Standard deviation	Average Discount (%)	Standard deviation
Psychology	35	882.37	77.8423	726.2286	82.4372	18.1765	5.3906
Political Science	60	894.1167	91.9855	720.1667	90.1848	19.7797	5.1996
Social Science	85	895.4235	79.3225	716.8353	74.2923	19.9647	3.8311
Medical	70	919.7857	73.8651	739.1857	77.4631	19.6571	3.9851
General	50	811.2	86.1181	666.68	84.0771	17.74	5.9618
Total	**300**	**885.2867**	**88.5616**	**715.4533**	**83.7837**	**19.2785**	**4.7926**

Analysis of the Sample Drawn

In this section, we discuss the statistical tools used to analyze the sample drawn. We first start with estimation.

Point Estimation

We provide estimates to the average cost of a book in the category academic and professional and in between **Rs.700** and **Rs.900,** before and after providing the discount. Also, we provide estimate of the average discount that Flipkart is offering in the same category. These are the three parameters we have estimated and the results are as follows:

- The average cost of a book is **Rs.885.2867** with standard deviation **Rs.88.5616**
- The average discount offered by Flipkart is **Rs.19.2785** with standard deviation **Rs.4.7926**

- The average cost of a book after discount is **Rs.715.4533** with standard deviation **Rs.83.7837**

Interval Estimation

The 95% confidence intervals are given by

- Population average cost of a book before discount: **(875.2245, 895.3489)**
- Population average discount of a book: **(18.7339, 19.8230)**
- Population average cost of a book after discount: **(705.9339, 724.9727)**

For calculations, one can refer to Levin and Rubin (2002).

Interpretation of these Intervals is as Follows

From the above confidence intervals, we conclude that if repeated samples are drawn from the population with respect to the average cost of a book, average discount offered by Flipkart, and average cost after discount, from Flipkart's website, 95% of these sample estimates will be in the above intervals, respectively. The chance that the estimates fall outside the interval is 0.05.

Hypothesis Testing

In this section, we will be using t-test for single mean, Paired t-test, ANOVA to test the hypotheses constructed. The major objective is to introduce these tools along with the assumptions associated. We also provide tools to test the assumptions associated with each of these methods and alternative methods when the assumptions are not satisfied. At the end we introduce chi-square test for independence of attributes.

Motivation to Construct the Hypotheses

Let us ask a question. What is the motivation to construct the hypothesis? Why can't we stop at estimation and why do we need hypothesis testing? Note that, there is 5% percent chance that the estimates fall outside the intervals. Also, if we take a follow up sample, there is every chance that the situations change and they demand that we have to change the strategies. In such cases, we can use hypothesis testing to test the claims made as per the situations, based on the sample drawn.

Test for One Sample (t-test/Wilcoxon Signed Rank Test)

This test is used to test the following hypotheses:

Hypothesis-1

H_0: the average cost of a book before discount is greater than or equal to Rs.896 i.e. $\mu \geq 896$.

H_1: the average cost of a book before discount is less than Rs.896 i.e., $\mu < 896$.

Hypothesis-2

H_0: the average discount of a book is greater than or equal to Rs.20i.e., $\mu \geq 20$.

H_1: the average cost of a book before discount is less than Rs.20 i.e., $\mu < 20$.

Hypothesis-3

H_0: the average cost of a book before discount is greater than or equal to Rs.726 i.e., $\mu \geq 726$.

H_1: the average cost of a book before discount is less than Rs.726 i.e., $\mu < 726$.

Assumptions of a t-test

1. The population from where the data has been drawn follows a normal distribution.
2. The responses are collected independently.
3. The population variance is unknown.
4. The sample should be a random sample.

Note that the most important assumption is the first one. If the sample does not satisfy this assumption then, we cannot use a t-test.

The Population from Where the Data has Been Drawn Follows a Normal Distribution

This assumption is tested using PP-plot and non-parametric test Kolmogorov-Smirnov test (K-S test)

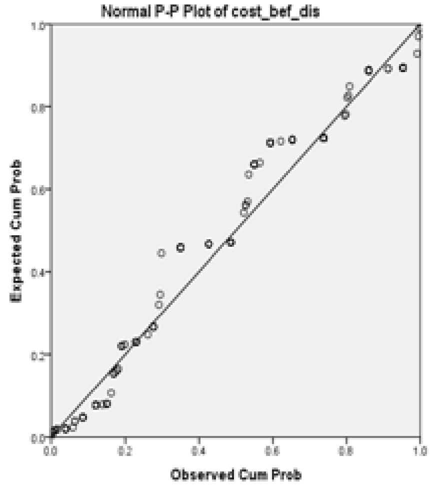


Figure 1 : PP-Plot Cost before discount

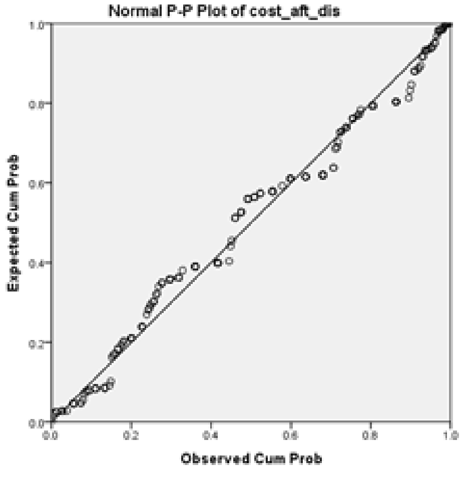
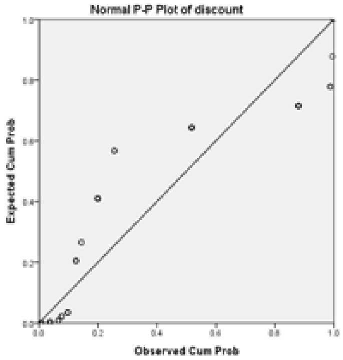
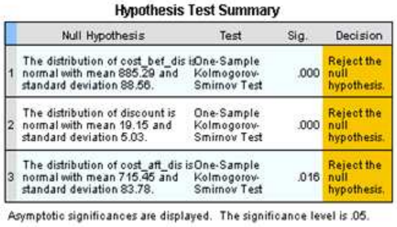


Figure 2 : PP-Plot Cost after discount





Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of cost_bef_dis is normal with mean 885.29 and standard deviation 88.56.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
2	The distribution of discount is normal with mean 19.15 and standard deviation 5.03.	One-Sample Kolmogorov-Smirnov Test	.000	Reject the null hypothesis.
3	The distribution of cost_aft_dis is normal with mean 715.45 and standard deviation 83.78.	One-Sample Kolmogorov-Smirnov Test	.016	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Figure 3 : PP-Plot Discount offered

- From the above plots, it is apparent that the data do not follow normal distribution.
- From the K-S test, we conclude that the data do not follow normal distribution.

Now suppose that we use a t-test, in spite of the failure of normality assumption.

The results are summarized in the following tables:

Random variable	t-value	d.f.	p-value
Cost of book before discount	-2.095	299	0.037
Cost of book after discount	-2.928	299	0.004
Discount	-2.180	299	0.030

From the above results, we conclude that the null hypotheses are rejected at 5% level of significance. The powers of the tests are as follows: (Powers are calculated using the website http://www.statisticalsolutions.net/pss_calc.php)

- Hypothesis-1: Power is 0.67 at 5% level of significance
- Hypothesis-2: Power is 0.70 at 5% level of significance
- Hypothesis-3: Power is 0.83 at 5% level of significance

It seems to be alright but it is not. We will be happy if the power of the test is close to 1. In the third hypothesis the power is 0.83. This is because the level of significance is 5%. If level of significance is 1% then the power is 0.61. This indicates that alternative procedures should be used to test the hypothesis. This is due to the fact that the data do not satisfy the assumption of normality.

Now one can use Wilcoxon signed rank test to test the significance of Median, which is an alternative test for a t-test for single mean. But the only difference is that the parameter is Median instead of Mean.

Wilcoxon Signed Rank Test

The Wilcoxon signed rank test is another example of a non-parametric or distribution free test. The Wilcoxon signed rank test is used to test the null hypothesis that the median of a distribution is equal to some value. It can be used a) in place of a one-sample t-test b) in place of a paired t-test or c) for ordered categorical data where a numerical scale is inappropriate but where it is possible to rank the observations.

For the sample collected, median cost of a book before discount is 879, median discount is 21 and median cost of a book after discount is 728. Using this information, we test the hypotheses defined above by replacing population means with population medians. Hence we test the hypotheses that population medians are equal to 881, 730, 23 respectively against they are not equal. The following table gives the summary of the results.

Random variable	Population Median	Sample Median	p-value	Decision
Cost of a book before discount	881	879	0.562	Null hypothesis not rejected
Cost of a book after discount	730	728	0.012	Reject the null hypothesis
Discount offered	23	21	0.0001	Reject the null hypothesis

Test for Dependent Samples (Paired t-test/ Wilcoxon Signed Rank Test)

This test has been used to test the hypothesis

H_0: There is no significant difference between the average cost of a book before and after the discount.

H_1: There is a significant difference between the average cost of a book before and after the discount.

Assumptions

1. The sample is a random sample.
2. The difference follows normal distribution.
3. The samples are dependent.

The results are as follows: t-calculated= 62.742, d.f. =299 and p-value=0.0001.

Since the p-value (0.0001) is less than 0.05, we reject the null hypothesis and conclude that the there is significant difference between the average cost of a book before and after discount.

But, the assumption of normal distribution is not satisfied by the data. Hence, the alternative method should be used before taking a decision about the hypothesis. The alternative non-parametric method is Wilcoxon-Signed Rank test.

When we use Wilcoxon Signed rank test for paired sample, the hypothesis is that the median difference is zero. The p-value is 0.0001 and the null hypothesis is rejected at 5% level of significance and this indicates that there is significant difference between the medians of cost before and after discount. Now based on this we can a decision and give the conclusion about the difference between the two medians.

Paired t-test also gave the similar conclusion about means. But we cannot take this as the data do not follow normal distribution.

Test for Differences of k-population Means

This test is used to test when we have more than two populations and are interested in testing the significant difference between population means. When there are only two means, we can use a t-test for difference of means. In this case we test the significance of means across different categories.

Assumptions

1. The samples are random samples.
2. The samples satisfy the assumption normality across the populations.
3. Equality of variances.
4. Errors are uncorrelated.

Testing the Assumptions

The first assumption that one has to test is assumption of normality. For the sample considered, the normality assumption is not satisfied. Hence ANOVA cannot be applied. The alternate non-parametric method is Kruskal-Wallis test should be applied.

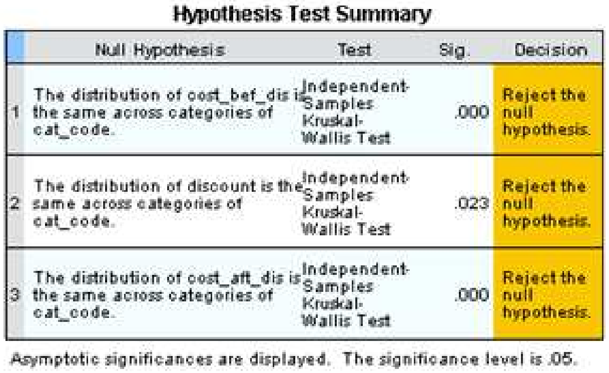
Hypotheses

H_0: There is no significant difference between means across different categories with respect to cost of a book, discount offered, cost of books after discount.

H_1: There is significant difference between means across different categories with respect to cost of a book, discount offered, cost of books after discount.

To test the above hypothesis, Kruskal-Wallis test should be applied as the data do not follow normal distribution.

The independent k-sample Kruskal-Wallis test results are summarized in the following table:



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of cost_bef_dis is the same across categories of cat_code.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The distribution of discount is the same across categories of cat_code.	Independent-Samples Kruskal-Wallis Test	.023	Reject the null hypothesis.
3	The distribution of cost_aft_dis is the same across categories of cat_code.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Note: All calculations and tables recorded in this case study are obtained from SPSS.

Chi-Square Test for Independence of Attributes

This test has been used to test whether category has an impact on cost of the book. This test is also used to check whether category effects discount offered, cost of the book after discount.

Before applying the test, we have to divide the data points corresponding to each random variable in to categories such that there will be atleast 5 expected frequencies under each category. If there are less than 5, then those groups should be clubbed with other groups and accordingly the degrees of freedom should be adjusted. For more details, one can refer to Levin and Rubin (2002).

Null Hypothesis:

The category of the book does not influence the cost of the book before discount, after discount and discount offered.

Alternative Hypothesis:

The category of the book influences the cost of the book before discount, after discount and discount offered.

The results are as follows:

Attribute-1	Attribute-2	p-Value	Decision
Book Category	Cost of a book before discount	0.0001	Null hypothesis is rejected
Book Category	Cost of a book after discount	0.0001	Null hypothesis is rejected
Book Category	Discount	0.005	Null hypothesis is rejected

From the above table, we conclude that the category of the book has an impact on the cost of the book, before and after the discount. Also that category has an influence on the discount offered.

Conclusion

From the above discussion, we conclude that the cost of a book and discount offered by Flipkart depends on the category of the book and also that on average, the cost of the book before discount and after discount are different. In general, one can observe that statistical study of this data helps us to confirm some of the general perceptions of customers with respect to cost of a book and discount offered. In this case we only justify these perceptions using statistical tools.

Finally, we conclude by saying that this case study introduces some of the important statistical tools and their alternatives. Using these, we conclude that all the hypotheses constructed based on Flipkart's data are tested appropriately and conclusions are provided as per the results obtained using SPSS.

References

Levin, R.I., Rubin, D.S. (2002): Statistics for Management. Seventh edition. Pearson.

Cochran, W.G. (2007): Sampling techniques. Third edition. Wiley India Pvt. Ltd.

Web references

http://www.flipkart.com/

Wipro's Acquisition of Promax Application Group

Ullas Rao & Malathi Sriram

On 30th April 2012, Wipro announced acquisition of Promax Application Group (PAG) based out in Australia. The deal value was pegged at AUD 35 million (Australian dollar). Given the Promax's significant penetration in the growing analytics space, the deal was expected to give Wipro a major headwind in this space dominated by few players. An acquisition of this nature is certainly expected by the management to contribute positively by boosting both top-line as well as bottom-line over several years.

The present case seeks to critically look at the acquisition undertaken by a major player in the Information Technology (IT) sector like Wipro Ltd., from the perspective of strategy as well as financial synergy. This would necessarily enable the participants to remain sensitive to the broader issues encountered while undertaking acquisitions.

It is also not uncommon to find companies flush with cash engaging in buyouts to justify judicious employment of cash that seeks to maximize the interests of shareholders. Otherwise, it is common for firms flush with cash resorting to distributing cash dividends or engaging in buyback programs to reward the shareholders.

From the financial synergy point of view, it will be interesting to observe the impact of acquisition announcement made by Wipro Ltd., on April 30th, 2012 on the wealth status of shareholders. The overarching question that needs to be addressed is whether this acquisition announcement led to the maximization of returns for the

The case writers Dr. Ullas Rao, Assistant Professor and Prof. Malathi Sriram Associate Professor SDMIMD Mysore. may be reached at ullasrao@sdmimd.ac.in and malathi@sdmimd.ac.in. This case is fictionalized and any resemblance to actual person or entities is coincidental. For Teaching Notes please contact sdmrcms@sdmimd.ac.in.