Testing The Mean For Business Data: Should One Use The Z-Test, T-Test, F-Test, The Chi-Square Test, Or The P-Value Method?

Jiajuan Liang, (Email: JLiang@newhaven.edu), University of New Haven
William S. Pan, (Email: wpan@newhaven.edu), University of New Haven

Abstract

In testing the mean of a population or comparing the means from two populations. There are several statistics available: the t-test, z-test, F-test and the chi-square test. Both the t-test and the z-test are usually used for continuous populations, and the chi-square test is used for categorical data. The F-test is used for comparing more than two means. In this paper we will discuss: 1) the conditions on using these tests; 2) the relationship among these test; and 3) illustration of the p-values of these tests by graphs. Some concluding remarks will be provided.

INTRODUCTION

An important topic in the elementary business statistics course is testing hypothesis about the mean of a population or comparing the means from two populations. In some higher-level applications, it is required to find out whether difference exists among more than two means, such as in the topic of one-way ANOVA (analysis of variance). Some other topics in the business statistics course, such as testing the proportion for one population or comparing two proportions from two populations, are usually taught in a separate chapter (or chapters) other than in the same chapter as for testing the mean of a population or comparing the means from two populations. These topics actually can be reduced to testing the mean of a population or comparing the means from two populations. We will address this equivalence in next section. In some higher-level applications, we may want to compare more than two proportions from several populations. This can also be reduced to comparing the means from more than two populations. Therefore the topic of testing the mean for business data covers quite a few of small topics and it constitutes a substantial part in the business statistics course, see, for example, Anderson, Sweeney and Williams (2005); Keller and Warrack (2003); and Levine, Stephan, Krehbiel and Berenson (2005).

Testing the mean for business data can be carried out according to different cases. For example, based on the types of data (continuous or categorical), we can determine whether a t-test, z-test or a chi-square test should be used; based on the sample size (large or small) or normal data, we can determine whether the t-test or the z-test should be used; based on how many means to be compared for continuous data, we can determine whether the two-sample t-test or the F-test in one-way ANOVA should be used; based on how many proportions to be compared, we can determine whether the approximate z-test (two-sample case) or the chi-square test should be used. Therefore these points should be carefully addressed in teaching the topic of testing the mean for business data.

After learning the course of business statistics, based on our experience, many students often get confused on which test (the t-test, z-test, F-test or the chi-square test) should be used for a given data set, or what is the relationship among these tests. For some given data, maybe more than one test can be used. In this case, most students have a difficult time in understanding which test may be more effective (or powerful). Questions like these can be made clear to students if we give a good summary on the topic of testing the mean after we finish teaching the separate chapters

on testing the mean, comparing proportions and one-way ANOVA in the business statistics course. In this paper we will summarize our experience in teaching these topics.

TESTING THE MEAN: ONE-SAMPLE CASE

This is the beginning topic in testing hypothesis in the business statistics course. We usually assume the data under study are collected from a population that is characterized by a continuous random variable, most of the time, a normal random variable, or we call the normal population. Testing the mean in one-sample case is to test the hypothesis

$$H_0 : \mu = \mu_0 \qquad (1)$$

versus the alternative hypothesis $H_1 : \mu \neq \mu_0$ in the case of two-tailed test, where μ_0 is a known number. One-tailed alternative hypotheses can be defined similarly. It is known that both the t-test and the z-test are may be used for testing the hypothesis in (1). They are defined by

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_n} \sim t(n-1), \quad \text{for small } n, \quad \text{and} \quad z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s_n} \to N(0,1), \quad \text{for large } n, \qquad (2)$$

where $\bar{x}$ is the sample mean and s_n is the sample standard deviation. The t-test is for the case of normal samples with small sample size. It is usually suggested to plot the histogram of the data before using the t-test. If the histogram looks like bell-shaped, the t-test can be used with confidence. Compare to the t-test, the condition on using the z-test for (1) is less restrictive in distributional assumption. That is, the data may be non-normal but the sample size should be large enough. It is noted that

$$z^2 = \left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{s_n} \right)^2 \to \chi^2(1), \quad \text{the chi-square distribution with 1 degree of freedom}, \quad (n \to \infty) \qquad (3)$$

and

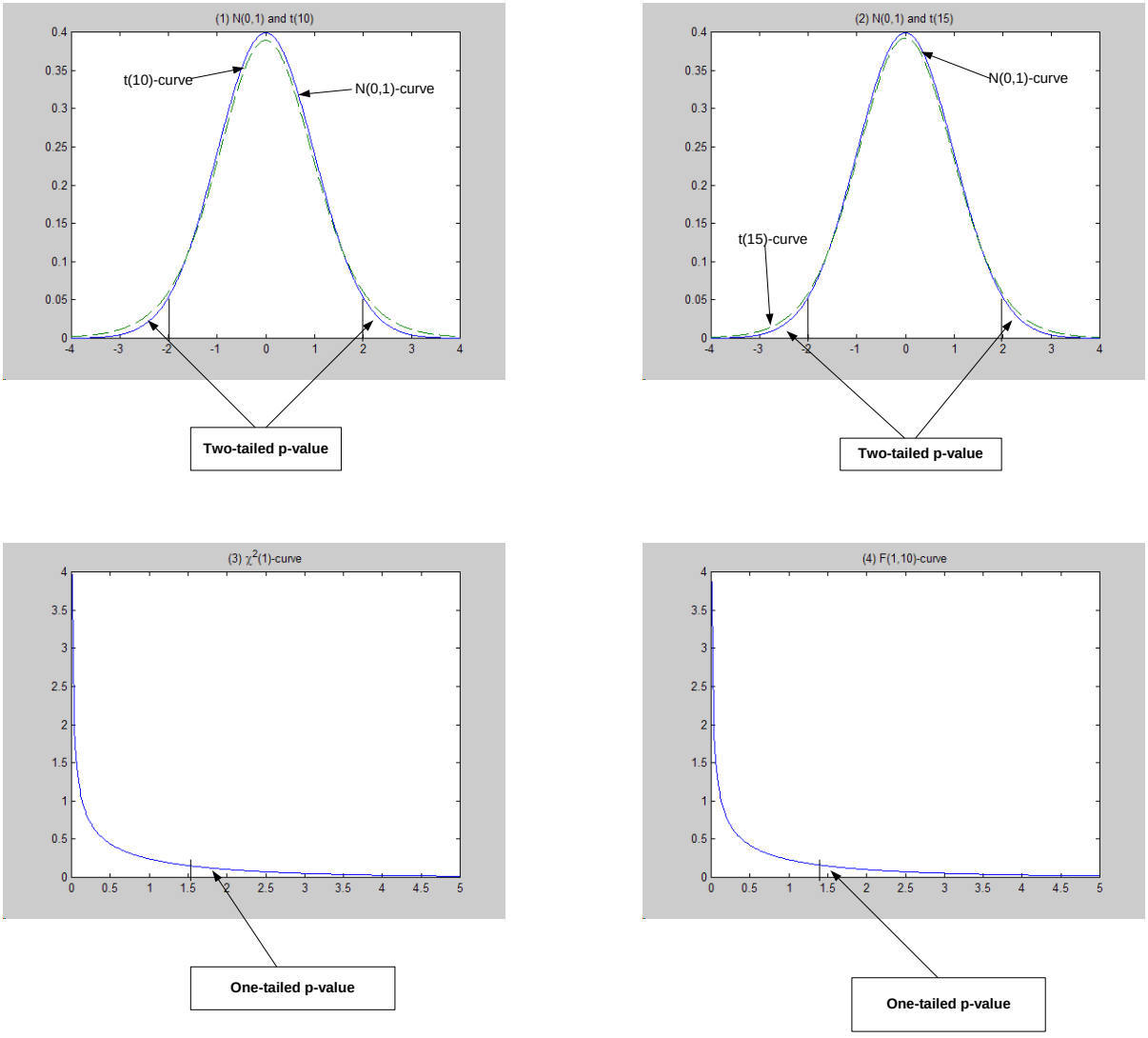
$$t^2 = \left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{s_n} \right)^2 \sim F(1, n-1), \quad \text{for small } n, \qquad (4)$$

therefore in testing hypothesis (1), the chi-square test $\chi^2(1)$ is equivalent to the z-test for large sample size, and the t-test is equivalent to the $F(1, n-1)$ -test for small sample size. Because the t-test $t(n-1)$ with (n-1) degrees of freedom has the property

$$t(n-1) \to N(0,1) \quad \text{in distribution}, \quad (n \to \infty) \qquad (5)$$

this means that the difference between the t-test and the z-test becomes smaller and smaller if the sample size becomes larger and larger and the data are normal. An illustration for the p-value of the two-tailed t-test and that of the two-tailed z-test, the p-value of the $F(1, n-1)$ -test and that of the $\chi^2(1)$ -test is given by Figure 1.

Figure 1. Comparison of p-values among the t-test, z-test, chi-square test and the F-test.



From Figure 1, we can see that the t-curve always has a thicker tail than that of the normal *N(0,1)*, this implies that the p-value of the t-test is usually larger than that of the z-test for the same data with small sample size. For a given significance level α (e.g., 1%, 5% or 10%), if p-value of the t-test is less than α (the hypothesis in (1) is rejected by the t-test at level α), we must have p-value of the z-test is also less than α (the hypothesis in (1) is also rejected by the z-test at level α). But the converse is not necessary true. This can be summarized as

$$\text{Rejection from the t-test at a given level} \Rightarrow \text{Rejection from the z-test at the same level.}$$

$$\chi^2(1)-\text{test} \Leftrightarrow \text{z-test}, \qquad F(1,n-1) \Leftrightarrow \text{t-test}, \tag{6}$$

where the sign "$\Rightarrow$" means "result in", and "$\Leftrightarrow$" means "equivalent to". It is also noted that the $\chi^2(1)$-test and the $F(1,n-1)$-test are always one-tailed tests, and there is no real difference between the t-test by *t(n-1)* and the z-test for

normal data when the sample size *n* is large enough, for example, $n \geq 15$. This is illustrated by the *t(15)*-curve and the $N(0,1)$-curve in Figure 1.

The problem of testing proportion arises from analysis of survey data characterized by "Yes" and "No" questions. For example, products can be divided into "good" (Yes) and "defective" (No); opinions on some specific questions in a sample survey may be "agree" (Yes) and "disagree" (No). Data collected from such surveys are usually quantified as 1=Yes and 0=No. Let *p*=proportion of "Yes" in the data set. A test of *p* is a test of proportion with the null hypothesis like

$$H_0 : p = p_0, \qquad (7)$$

versus the alternative hypothesis $H_1 : p \neq p_0$ in the case of two-tailed test, where p_0 is a known proportion (percentage). One-tailed alternative hypotheses can be defined similarly. For example, a manufacturer may claim that his products have a defective rate of no more than 1%. Then the hypothesis is: $H_0 : p = 1\%$ versus $H_1 : p > 1\%$. Note that a population with a proportion can be characterized by a random variable like

$$X = \begin{cases} 1, & \text{if the response is "Yes"}, \\ 0, & \text{if the response is "No"}. \end{cases} \qquad (8)$$

If we assume *p*=percentage of "Yes" in the responses, then *1-p*=percentage of "No" in the responses. The expected (mean) value of the population characterized by *X* is $E(X) = 1 \times p + 0 \times (1-p) = p$. Therefore a test about the proportion *p* in (7) is a test about the population mean. It is known that the approximate z-test is used for testing (7), which is defined by

$$z = \frac{p_s - p_0}{\sqrt{p_0(1-p_0)/n}} \to N(0,1), \qquad (n \to \infty) \qquad (9)$$

where p_s stands for the sample proportion. It is noted that

$$z^2 = \frac{(p_s - p_0)^2}{p_0(1-p_0)/n} \to \chi^2(1), \qquad (n \to \infty) \qquad (10)$$

Therefore we can summarize

$$\text{The z-test for (7) by (9)} \Leftrightarrow \text{the } \chi^2(1)\text{-test for (7) by (10).} \qquad (11)$$

The p-value of the z-test by (9) and the p-value of the χ^2-test by (10) can be referred to the $N(0,1)$-curve and the $\chi^2(1)$-curve in Figure 1.

COMPARING TWO MEANS: TWO-SAMPLE CASE

The hypothesis for comparing two means from two populations can be stated as

$$H_0 : \mu_1 = \mu_2, \qquad (12)$$

versus the alternative hypothesis $H_1: \mu_1 \neq \mu_2$ in the case of two-tailed test. One-tailed alternative hypotheses can be defined similarly. In the case of small sample size and assuming normal data from the two populations with equal variances, the two-sample t-test is given by

$$T = \sqrt{\frac{nm(n+m-2)}{n+m}} \bullet \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_{1n}^2 + (m-1)s_{2m}^2}} \sim t(n+m-2), \quad n, m \text{ are small,} \qquad (13)$$

where *m* and *n* are sample sizes for the two samples, $\bar{x}$ and $\bar{y}$ are the two sample means, and s_{1n}^2 and s_{2m}^2 are the two sample variances. The z-test for testing (12) does not require the equal variance assumption for the two populations but it requires large sample sizes for the two samples. The two-sample z-test is given by

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_{1n}^2}{n} + \frac{s_{2m}^2}{m}}} \to N(0,1), \qquad n \to \infty, \ \ m \to \infty. \qquad (14)$$

According to the similar properties of the t-test and z-test to those mentioned in (3), (4) and (5), we have

$$T^2 = [t(n+m-2)]^2 \sim F(1, n+m-2), \quad Z^2 \sim \chi^2(1), \quad t(n+m-2) \to N(0,1), \ \ n, m \text{ are large.} \qquad (15)$$

Therefore, in testing hypothesis (12), we have the equivalence:

$$\begin{aligned} &\text{The t-test for (12) by } t(n+m-2) \text{ in (13)} \Leftrightarrow \text{the F-test for (12) by } T^2 \sim F(1, n+m-2) \text{ in (15),} \\ &\text{The z-test for (12) by } Z \sim N(0,1) \text{ in (14)} \Leftrightarrow \text{the chi-square test for (12) by } Z^2 \sim \chi^2(1). \\ &\text{The t-test for (12) by } t(n+m-2) \text{ in (13)} \Leftrightarrow \text{the z-test for (12) by } Z \sim N(0,1) \text{ for large } n, m. \end{aligned} \qquad (16)$$

The p-values of these tests can be referred to the t-curve, the $N(0,1)$-curve, the χ^2-curve and the F-curve in Figure 1.

The problem of comparing two proportions is also a test of comparing two population means. For example, we can define the two populations associated with the two random variables as

$$X = \begin{cases} 1, & \text{if the response for population 1 is "Yes",} \\ 0, & \text{if the response for population 1 is "No",} \end{cases}$$

$$Y = \begin{cases} 1, & \text{if the response for population 2 is "Yes",} \\ 0, & \text{if the response for population 2 is "No".} \end{cases}$$

Assuming the two proportions as $P(X=1) = p_1$ and $P(Y=1) = p_2$, we have $E(X) = p_1$ and $E(Y) = p_2$. Therefore the following hypothesis for comparing the two proportions is one for comparing two population means:

$$H_0: p_1 = p_2 \qquad (17)$$

versus the alternative hypothesis $H_1: p_1 \neq p_2$ in the case of two-tailed test. One-tailed alternative hypotheses can be defined similarly. It is known that the approximate z-test for testing (17) is given by

$$Z_a = \frac{p_{1s} - p_{2s}}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)p_s(1-p_s)}} \to N(0,1), \quad n \to \infty,\ m \to \infty, \tag{18}$$

where $p_{1s} = x/n$ and $p_{2s} = y/m$ are the two sample proportions, *x*=the total number of responses to "Yes" in sample 1, and *y*=the total number of responses to "Yes" in sample 2, $p_s = (x+y)/(n+m)$ is called the "pooled sample proportion". Similar to (3), (10) and (15), we have

$$Z_a^2 \to \chi^2(1), \quad n \to \infty,\ m \to \infty. \tag{19}$$

Therefore we have the equivalence

$$\text{The z-test by } Z_a \text{ in (18) for testing (17)} \Leftrightarrow \text{the chi-sqaure test (19) for testing (17).} \tag{20}$$

The p-values for the above tests can be referred to Figure 1, the $N(0,1)$-curve and the $\chi^2(1)$-curve.

COMPARING THE MEANS: MULTI-SAMPLE CASE

The problem of comparing more than two means arises from the necessity of comparing the mean effects among different treatments on a subject. For example, if there is r advertising media (such newspaper, TV, magazine, etc.) that are claimed to be effective in increasing the amount of sales of some product, a case study may want to find out whether there is real difference in using these media after some time. Let μ_i (*i=1, .., r*) be the average (mean) amount of sales from using medium *i*. Then the problem of comparing the *r* media is reduced to testing the hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r, \tag{21}$$

versus H_1: not all means are equal. It is well-known that hypothesis (21) is a one-way ANOVA problem. A statistic for testing (21) is the F-test with an F-distribution $F(r-1, N-r)$ with degrees of freedom $(r-1, N-r)$, where *N* is the total sample size from all treatments (populations). Note that when *r=2*, that is, only two means are compared, hypothesis (21) reduces to hypothesis (12) and the F-test reduces to $F(1, N-2)$. This is exactly the statistic $T^2 \sim F(1, n+m-2)$ in (15) with $N = n+m$. According to the equivalence in (16), when *r=2*, the F-test by $F(1, N-2)$ is equivalent to a t-test by $t(N-2)$. When $r \geq 3$, there is no t-test that is equivalent to the F-test $F(r-1, N-r)$ for testing (21).

The problem of comparing more than two proportions can be easily transferred to comparing more than two means by the same method as in Sections 2 and 3 by defining suitable random variables taking only values 1 and 0. The hypothesis can be stated as

$$H_0 : p_1 = p_2 = \cdots = p_r, \tag{22}$$

versus H_1: not all proportions are equal. When *r=2*, hypothesis (2) reduces to hypothesis (17) and an approximate z-test or an approximate chi-square test can be used. When $r \geq 3$, there is no equivalent z-test. The problem becomes a higher-level application of chi-square test. Wilks (1935) derived a chi-square test with an approximate distribution $\chi^2(r-1)$ for testing (22). It is obvious that when *r=2*, Wilks' (1935) statistic $\chi^2(r-1)$ reduces to $\chi^2(1)$, the same distribution as pointed out in (19).

Note that all of the above-mentioned tests are parametric tests. That is, they are used to test some parameter(s) from one or more populations. According to Wilks (1935, 1938), all of the above-mentioned hypotheses can be tested by a unified approach: the likelihood ratio (LR) approach that results in some approximate chi-square tests. Therefore all of the above-mentioned tests are related by some LR-type chi-square tests. Because the likelihood ratio (LR) approach is not covered in an elementary business statistics course, we only summarize the following conclusions for students' higher-level reference:

(1) Hypothesis (1) (testing the mean from one population) can be tested by an LR-statistic with distribution $\chi^2(1)$;

(2) Hypothesis (7) (testing the proportion from one population) can be tested by an LR-statistic with distribution $\chi^2(1)$;

(3) Hypothesis (12) (comparing two means) can be tested by an LR-statistic with distribution $\chi^2(1)$;

(4) Hypothesis (17) (comparing two proportions) can be tested by an LR-statistic with distribution $\chi^2(1)$;

(5) Hypothesis (21) (comparing any finite number of means) can be tested by an LR-statistic with distribution $\chi^2(r-1)$;

(6) Hypothesis (22) (comparing any finite number of proportions) can be tested by an LR-statistic with distribution $\chi^2(r-1)$.

Therefore, all tests for testing the mean in the one-sample case or for comparing two or more means for the multi-sample case are related to some chi-square tests in the point of view of likelihood ratio. It is our experience that a summary on the relationships among the tests for population mean(s) helps students in better understanding this big topic in the business statistics course. We have received many positive comments on such a summary. It is our belief that a textbook with such a summary as a supplement or appendix will receive much welcome from many business students.

AN EXAMPLE

In marketing children's products, it is extremely important to produce television commercials that hold the attention of the children who view them. A psychologist hired by a marketing research firm wants to determine whether differences in attention span exist among children watching advertisements for different types of products. One hundred fifty children under 10 years of age were recruited for an experiment. One third watched a 60-second commercial for a new computer game, one third watched a commercial for a breakfast cereal, and another one third watched a commercial for children's clothes. Their attention spans were measured. The results (in seconds) are stored in a data set. The marketing manager wants to know if the data provide enough evidence to conclude that there are differences in attention span among the three products advertised.

Note that there are three groups of children in the experiment. Let

Population A= {children watching a 60-second commercial for a new computer game}
Population B= {children watching a commercial for a breakfast cereal}
Population C={children watching a commercial for children's clothes}

The marketing manager's question can be answered by comparing three means. That is, the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad (23)$$

versus the alternative hypothesis H_1: at least two means differ. This is a multi-sample (more than 2) case. According to Section 4, we have $r = 3$, $N = 150$, and the F-test is $F(r-1, N-r) = F(2,147)$ for the one-way ANOVA. In order to double check the conclusion from the analysis of one-way ANOVA, we also carried out the test for a comparison of two means. There are a total of two-sample comparisons. Let AB denote the two-sample comparison between the

means from populations A and B as defined above. The other notations such as AC and BC have a similar meaning. ABC means the comparison among all three means (hypothesis (23)). We employ the two-sample t-test given by (13), the two-sample z-test given by (14), the F-test and the chi-square test given by (15) for all two-sample comparisons. The observed statistics and their associated p-values are summarized in Table 1.

Table 1: Summary information for the tests

Statistics	3-sample comparison ABC: $F(2,147)=1.8167$, p-value=0.1662					
	2-sample comparisons					
	AB		AC		BC	
	Observed	p-value	Observed	p-value	Observed	p-value
t-test: t(98)	T=0.0631	0.9498	T=1.6562	0.1009	T=1.7342	0.0860
z-test: N(0,1)	Z=0.0631	0.9497	Z=1.6562	0.0977	Z=1.7342	0.0829
F-test: F(1,98)	$F=T^2$ =0.0040	0.9498	$F=T^2$ =2.7429	0.1009	$F=T^2$ =3.0074	0.0860
χ^2-test: $\chi^2(1)$	$\chi^2=Z^2$ =0.0040	0.9496	$\chi^2=Z^2$ =2.7429	0.0977	$\chi^2=Z^2$ =3.0074	0.0829

From Table 1 we can summarize the following conclusions:

(1) Hypothesis (23) is not rejected at all three levels (1%, 5% and 10%) based on p-value=0.1662. This means that the data give no enough evidence to conclude that there are differences in attention span among the three products advertised;

(2) The information on the 2-sample comparison for AB strongly support the fact that there is no difference in attention span between group A and group B;

(3) The information on the 2-sample comparison for AC implies a possible difference in attention span between group A and group C at the significance level 10% because some of the p-values tend to be less than 10%;

(4) The information on the 2-sample comparison for BC implies a slight difference in attention span between group B and group C at the significance level 10% because all of the p-values are less than 10% but close to 10%;

(5) The slight difference in attention span between group B and group C at the significance level 10% may be worth special attention because all 2-sample tests give the smallest p-values among the other 2-sample comparisons. While the 3-sample comparison ABC gives a p-value=0.1662>10%, which shows no difference in attention span among the three products advertised at level 10%, the 2-sample comparisons can provide more information on which two groups may have potential difference.

Based on our experience, an example like the above one greatly helps the students in understanding the relationships among different tests for population mean(s). They feel that the topic of testing the mean in the business statistics course is really useful in their anticipating career and different methods can enhance their understanding in this topic.

Acknowledgment: This work was supported by a University of New Haven 2005 Summer Faculty Fellowship.

REFERENCES

1. David R. Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2005). *Statistics for Business and Economics*, 9th Edition, Thomson, South-Western.
2. Keller, G. and Warrack, B. (2003). *Statistics for Management and Economics*, 6th Edition, Thomson Learning, Inc., USA.

3. Levine, D. M., Stephan, D., Krehbiel, T. C., and Berenson, M. L. (2005). *Statistics for Managers: Using Microsoft Excel*, 4th Edition, Prentice Hall, Upper Saddle River, New Jersey.
4. Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *The Annals of Mathematical Statistics*, 6 (No. 4), 190-196.
5. Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9 (No. 1), 60-62.

NOTES