China-USA Business Review, November 2014, Vol. 13, No. 11, 688-703
doi: 10.17265/1537-1514/2014.11.002

Chi-square Distribution as a Research Tool, and Some of Its Most Important Parametric and Non-parametric Applications

Athanasios Vasilopoulos
St. John's University, New York, USA

When researchers are testing the validity of claims during their research, they may use either parametric methods (if they exist) or non-parametric methods if appropriate parametric methods do not exist. The Chi-square (x^2) distribution plays an important role in both parametric and non-parametric methods and many of its most important applications are explored in this paper. This paper provides an excellent summation of the properties and capabilities of the very versatile x^2 distribution, and many specific applications and suggestions for additional future applications.

Keywords: Chi-square (x^2) distribution and its most important properties, use of the x^2 distribution as a test statistic, applying the x^2 distribution to perform parametric tests on the population parameters σ^2 and σ, testing the equality of three population variances, applying the x^2 distribution to perform non-parametric tests on frequencies, goodness of fit, independence, homogeneity

Introduction

Most of the statistical methods that people are familiar with are referred to as parametric statistics and the term is used to indicate the nature of the population from which the sample data set: $x_1, x_2, x_3, \ldots, x_n$, which this paper is about to analyze, came from, for example when the empirical rule is used, this paper makes the assumption that the sample data came from a normal distribution. However, when this paper uses CHEBYCHEV's inequality, namely:

$$P[\bar{x} - k\hat{s} \leq X \leq \bar{x} + k\hat{s}] \geq 1 - \frac{1}{k^2}, \text{ for } k > 1 \quad (1)$$

which states that the probability that a random variable X is between k standard deviations of the mean $(\bar{x})$ is at least $1 - \frac{1}{k^2}$, the result is valid for all possible distributions of the random variable x.

Such "distribution-free" results, are called non-parametric statistics. In these nonparametric tests, the parameters of the distribution continue to be important. What is not important is the nature of the distribution of the population from which the sample came from. What is important is that these tests are valid whether the population distribution is normal, binomial, uniform, exponential, etc..

Athanasios Vasilopoulos, Ph.D., professor, St. John's University, New York, USA.
Correspondence concerning this article should be addressed to Athanasios Vasilopoulos, The Peter J. Tobin College of Business, St. John's University, 8000 Utopia Parkway, Jamaica, N.Y. 11439. E-mail: vasilopa@stjohns.edu.

When both parametric and non-parametric methods exist for the same application, it is natural to ask which test is preferable, the parametric, or nonparametric one. The answer is the parametric test, because a much larger sample (i.e. nonparametric methods are less efficient than parametric methods) is required for the non-parametric tests to achieve the same results (i.e. the same power).

But there are many situations in which the form (or nature) of the population distribution is not well known and the nonparametric method is the only meaningful alternative. This is also the case when no parametric alternative exists.

However, even though the computations on nonparametric statistics are usually less complicated than those for parametric statistics, the calculations for many nonparametric statistics can become very tedious, when the samples are large.

Another disadvantage for most nonparametric methods is the fact that the null hypothesis (H_0) being tested is less precise than that in the parametric methods, and the conclusions drawn may be somewhat vague. But, even with these drawbacks, nonparametric statistics are very useful, and it is important to know when and where they can be used, and the conclusions which can be drawn from their application.

The Chi-square (x^2) distribution can and is used as both a parametric and non-parametric tool, as shown below:

(1) The parametric applications of the Chi-square (x^2) distribution include the following tests:

a. testing the hypothesis that the variance of a population is equal to a claimed value (e.g., H_0: $\sigma^2 = 25$ vs. H_1: $\sigma^2 \neq 25$);

b. testing the hypothesis that the standard deviation of a population is equal to a claimed value (e.g., H_0: $\sigma = 10$ vs. H_1: $\sigma \neq 10$);

c. testing the hypothesis that the variances of three or more populations are equal (e.g., H_0: ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2 = \ldots = {\sigma_k}^2$ vs. H_1: The k variances are not all equal).

(2) The most important non-parametric applications of the Chi-square (x^2) include the following tests:

a. tests on frequencies—are the frequencies of classes consistent with expectations?

- tests on the frequencies of two classes;
- tests on the frequencies of more than two classes.

b. tests on the independence of two or more characteristics of the same population (contingency tables)—Are two characteristics of the elements of a population independent of each other?

c. tests on homogeneity—are two or more Independent random samples drawn from the same population?

d. tests on goodness of fit—does a population under investigation follow a specific probability model (i.e. uniform, normal, exponential, etc.)?

- Is it a uniform distribution?
- Is it a normal distribution?

Before proceeding with the discussion of these applications, a brief literature review, the research methods used and the research results obtained are stated.

Literature Review

Because the objective of this paper is to establish the x^2 as a well understood and properly used research tool, an exhaustive search was made to determine how other authors applied the x^2 distribution to problem solving.

Canal and Micciolo (2014) attempted to explain certain limitations of the goodness of fit test. Hwang and

Wang (2008) proposed a Chi-square test for testing the hypothesis that a truncation distribution follows a parametric family. Lang and Iannario (2013) discussed a new approach for improving statistical tests of independence between two categorical variables *R* and *C*, where *C* is ordinal and *R* may or may not be ordinal. McHugh (2013) analyzed the Chi-square test of independence and its limitations. Xie (2014) used it in more complex parametric multiple testing methods. Berenson, Levine, and Krehbiel (2004) used the x^2 for an independence test. Black (2004) used it for a goodness of fit and for some novel applications. Canavos (1984) used the moment generating function of the distribution in his discussion. Carlson and Thorne (1997) used it for a homogeneity test. Freund and Williams (1982) used it for a test of independence. McClave, Benson, and Sincich (2001) used it in contingency table analysis. Salvatore (1982) used it for tests of goodness of fit and independence. Vasilopoulos (2007) used it to test the equality of three or more variances.

Research Methods

A comparison was made of the way the cited authors applied the Chi-square distribution, a list was constructed of the many applications that have already been made, and ideas were sought for future applications.

Reasearch Results

After major applications of the x^2 research tool were discussed and the procedures of their analysis were identified, five examples of the most important applications are solved completely, to make sure the methods are clearly understood. These examples are found in the following section, which follows.

Discussion

Before starting discussing both the parametric and non-parametric applications of the Chi-square (x^2) distribution, this paper first discusses briefly the "nature" of the x^2 distribution and some of its most important characteristics, such as: the density function of the distribution, the shape, expected value and variance of the distribution, and the method of calculating probabilities using the chi-square (x^2) table (because the density function is too complicated to allow the use of the closed-form (or analytical) method of integration).

Chi-square (x^2) Distribution and Some of Its Most Important Properties

$$\text{If } Y = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} Z_i^2, \tag{2}$$

where, Z_1, Z_2, Z_3, …, Z_n are all standard normal variables (i.e. $\mu = 0$ and $\sigma = 1$), then, Y is said to be Chi-square distributed with degrees of freedom:

$$Y = \text{DOF} = n \tag{3}$$

$$\text{If } Y = x^2, \tag{4}$$

then Y has a density function; $f(y) = f(x^2)$ given by:

$$f(y = x^2) = \frac{(x^2)^{\frac{n}{2}-1} e^{\frac{-x^2}{2}}}{2^{\frac{n}{2}}[\Gamma(\frac{n}{2})}, \quad 0 \leq x^2 < \infty, \tag{5}$$

where, $\Gamma(n/2)$ = Tabulated gamma function,

$$= \int_0^{\infty} x^{\frac{n}{2}-1} e^{-x} dx, \tag{6}$$

whose role is to keep the area, under the f(x^2) function, from 0 to ∞, equal to 1 (as required by an axiom of probability) for all values of *n* (Canavos, 1984),

$$\Gamma\ (½)\ =\ \sqrt{\pi} \tag{7}$$

$$\Gamma\ (2) = 1, \Gamma(1) = 1 \text{ and } \Gamma(n + 1) = n\Gamma(n), \text{ for } n > 0.$$

Equation (5) represents an entire family of curves, one for each value of *n*. When plotted, for a given value of *n*, f(x^2) is a positively skewed distribution, starting at $x^2 = 0$ and going all the way to +∞, as shown below in Figure 1.

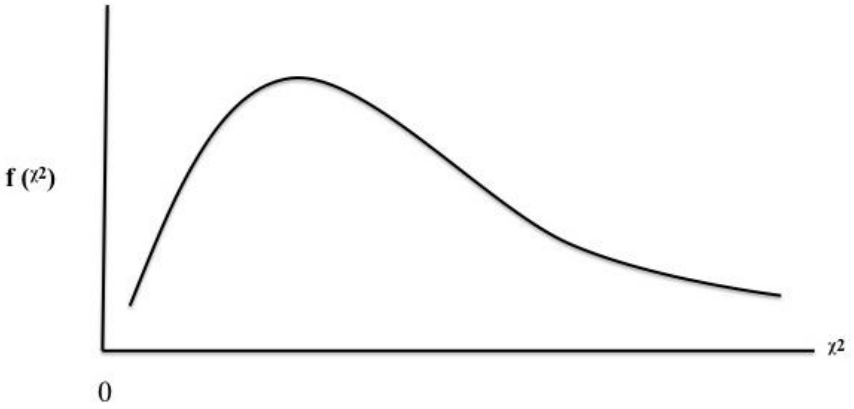

Figure 1. Shape of density of function of x^2.

The expected value = E(x^2) = E(*y*), E(y^2), and V(*y*) = V(x^2) are given respectively by,

$$\mathrm{E}(y) = \mathrm{E}(\chi^2) = \int_0^{\infty} y\mathrm{f}(y)dy = \mathrm{DOF} = n \tag{8}$$

$$\mathrm{E}(y^2) = \mathrm{E}[(\chi^2)^2] = \int_0^{\infty} y^2\mathrm{f}(y)dy = 2n + n^2 \tag{9}$$

$$\mathrm{V}(y) = \mathrm{V}(\chi^2) = \mathrm{E}(y^2) - [\mathrm{E}(y)]^2 = 2 \times \mathrm{DOF} = 2n \tag{10}$$

It is important to emphasize that the expected value of a x^2 variable is equal to its degrees of freedom (DOF = *n* here), while its variance is equal to 2 × DOF = 2*n*. To find the probability that a x^2 variable is between two values, $x^2 = a$ and $x^2 = b$, it needs to integrate the x^2 density function between these two values, as shown below (Figure 2).

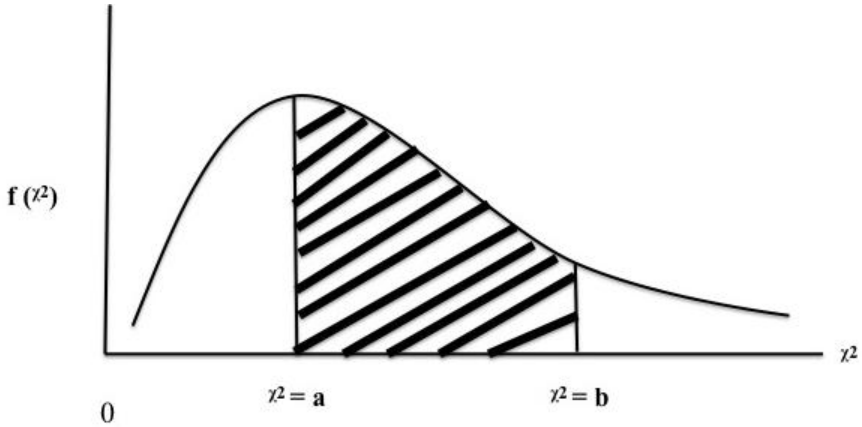

Figure 2. Probability calculations of x^2.

$$P[\, a \leq \chi^2 \leq b \,] = \int_a^b f(\chi^2) d\chi^2 \tag{11}$$

$$= [g(\chi^2)]_a^b = g(b) - g(a) \tag{12}$$

if a function $g(x^2)$ can be found such that

$$g'(x^2) = f(x^2) \tag{13}$$

i.e. if the derivative of $g(x^2)$ is equal to $f(x^2)$, then the probability can be found by using equation (12). This "closed-form" or "analytical" integration is possible only if $f(x^2)$ is a relatively "simple" function (Black, 2004).

But, since the density function of the x^2 distribution, as given by equation (5) is a complicated equation, and the corresponding $g(x^2)$ function can not be found, the required integration will be performed using a x^2 table.

As can be seen from such a table, there is a different x^2 density function for each DOF = *df*, and the shaded area represents the value of "α" shown (α = 0.995, α = 0.990, α = 0.975, α = 0.950, α = 0.900, α = 0.100, α = 0.050, α = 0.025, α = 0.010, α = 0.005).

The table gives, at the intersection of the row (*df*) and column (α value), the value of the x^2 variable (with the given DOF) which will make the area, under the $f(x^2)$ function, from this value to $+\infty$, equal to the α of the column, for example, when *df* = 10 and α = 0.975, $\chi^2_{10,\ 0.975} = 3.247$ while,

$$\text{when } df = 10 \text{ and } \alpha = 0.025,\ \chi^2_{10,\ 0.025} = 20.483 \tag{14}$$

therefore, $P[3.247 \leq \chi^2_{df=10} \leq 20.483] = 0.95$ (0.975-0.025).

Note 1: If DOF and/or α change, the results of the integration will also change.

Note 2: When $n \geq 30$, probabilities can be calculated using the normal distribution with

$$\text{Expected value} = n, \tag{15}$$

$$\text{Standard deviation} = \sqrt{2n} = \sqrt{\text{variance of } \chi^2} \tag{16}$$

Note 3: For the parametric applications of the x^2 distribution, this paper makes use of the fact that

$$\hat{s}^2 = \text{unbiased sample variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{17}$$

χ^2_{n-1} = Chi-square with DOF = $n - 1$.

Specifically, it can show that

$$\frac{(n-1)\hat{s}^2}{\sigma^2} = \chi^2_{n-1} \tag{18}$$

from which it can easily show that

$$E(\hat{s}^2) = \sigma^2 \text{ (i.e. } \hat{s}^2 \text{ is an unbiased estimator of } \sigma^2\text{)} \tag{19}$$

$$V(\hat{s}^2) = 2\sigma^4/(n-1) \tag{20}$$

Some of the Most Important Parametric Applications of the Chi-square Distribution

As previously mentioned, these include:

(1) testing the hypothesis that the variance of a population (σ^2) is equal to a specified value (σ_0^2) (Berenson et al., 2004);

(2) testing the hypothesis that the standard deviation of a population (σ) is equal to a specified value (σ_0);

(3) testing the hypothesis that the variances of three or more populations are equal to each other.

Now this paper can show how these three hypothesis tests are carried out and, for problems (1) and (2), it

can also construct confidence intervals (CIs) and show their equivalency to the hypothesis test solutions:

(1) Testing that $\sigma^2 = \sigma_0^2$, the following steps will be used:

(a) $H_0: \sigma^2 = \sigma_0^2$ vs. $H_1: \sigma^2 \neq \sigma_0^2$;

(b) selecting the value of α (usually $\alpha = 0.05$ or $\alpha = 0.01$, or both);

(c) the estimator for the parameter σ^2 is $\hat{s}^2$ and, according to equation (18) $\frac{(n-1)\hat{s}^2}{\sigma^2} = \chi^2_{n-1}$;

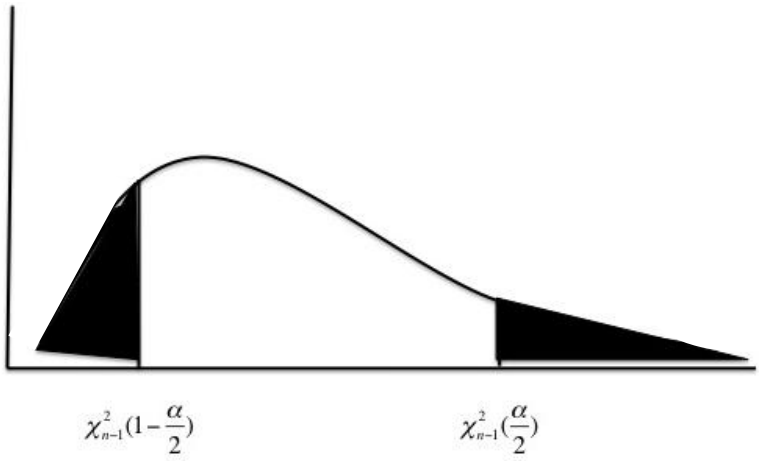

Figure 3. Construction of Rejection Region of χ^2.

Note: The shaded area is the Rejection Region and the non-shaded area is the Acceptance Region.

(d) constructing a rejection region (RR), using the x^2 table with DOF = $n - 1$; in particular, obtain from the x^2 table, the values: $\chi^2_{n-1}\left(\frac{\alpha}{2}\right)$ and $\chi^2_{n-1}\left(1-\frac{\alpha}{2}\right)$ and construct the RR\Acceptance Region (AR) as shown above in Figure 3;

(e) calculating the value of the test statistic

$$x^{2*} = \frac{(n-1)\hat{s}^2}{\sigma^2} \tag{21}$$

(f) comparing x^{2*} to the RR (Salvatore, 1982):

- If x^{2*} falls inside the RR, reject the validity of H_0;
- If x^{2*} falls outside the RR, do not reject the validity of H_0;
- If $H_0: \sigma^2 = \sigma_0^2$ is rejected, it concludes that $\sigma^2 \neq \sigma_0^2$;
- If $H_0: \sigma^2 = \sigma_0^2$ is not rejected, it concludes that $\sigma^2 = \sigma_0^2$.

(2) Testing that $\sigma = \sigma_0$, the following steps will be used.

(a) $H_0: \sigma = \sigma_0$ vs. $H_1: \sigma \neq \sigma_0$;

(b) selecting the value of α (usually $\alpha = 0.05$ or $\alpha = 0.01$, or both);

(c) the estimator of σ is $\hat{s}$ but authors do not know the sampling distribution of $\hat{s}$.

To solve the problem, H_0 and H_1 are reformulated in terms of σ^2 are reformulated ($\sigma^2 = (\sigma)^2$, or $\sigma = \sqrt{\sigma^2}$), because its estimator $\hat{s}^2$ is known to χ^2_{n-1}.

That is, authors change $H_0: \sigma = \sigma_0$ vs. $H_1: \sigma \neq \sigma_0$ to $H_0: \sigma^2 = \sigma_0^2$ vs. $H_1: \sigma^2 \neq \sigma_0^2$, and then follow the procedure shown in problem (1) above.

A$(1 - \alpha)$ confidence interval for σ^2 is obtained from

$$\mathrm{P}\left[\frac{(n-1)\hat{s}^2}{\chi^2_{n-1,\ \alpha/2}} \leq \sigma^2 \leq \frac{(n-1)\hat{s}^2}{\chi^2_{n-1,\ 1-\frac{\alpha}{2}}}\right] = 1-\alpha \tag{22}$$

while a $1 - \alpha$ confidence interval for σ is obtained from:

$$\mathrm{P}\left[\sqrt{\frac{(n-1)\hat{s}^2}{\chi^2_{n-1,\ \alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)\hat{s}^2}{\chi^2_{n-1,\ 1-\frac{\alpha}{2}}}}\right] = 1-\alpha \tag{23}$$

by taking square roots of the quantities inside the brackets of equation (22).

The equivalency between the hypothesis test solution and confidence interval solution is as follows (Chou, 1992):

If the hypothesized value σ_0^2 falls inside the limits of equation (22), the hypothesis $\mathrm{H}_0: \sigma^2 = \sigma_0^2$ is not rejected. Also, if the hypothesized value σ_0 falls inside the limits of equation (23), the hypothesis $\mathrm{H}_0: \sigma = \sigma_0$ is not rejected.

If the hypothesized value σ_0^2 falls outside of the limits of equation (22), the hypothesis $\mathrm{H}_0: \sigma^2 = \sigma_0^2$ is rejected. Also, if the hypothesized value σ_0 falls outside the limits of equation (23), the hypothesis $\mathrm{H}_0: \sigma = \sigma_0$ is rejected.

For this equivalency to exist, the two tests must be similar; i.e. a two-sided hypothesis test solution must be compared only to a two-sided confidence interval solution.

It is well known that to test the equality of two population variances (i.e. $\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$ vs. $\mathrm{H}_1: \sigma_1^2 \neq \sigma_2^2$), the following will be used.

The normal distribution, if $n_1 \geq 30$ and $n_2 \geq 30$ (and preferably $n_1 \geq 100$ and $n_2 \geq 100$) by testing

$$\mathrm{H}_0: \Delta\sigma^2 = \sigma_1^2 - \sigma_2^2 = 0 \ vs. \ \mathrm{H}_1: \Delta\sigma^2 \neq 0 \tag{24}$$

whose estimator,

$$\Delta\hat{s}^2 = \hat{s}_1^2 - \hat{s}_2^2 \tag{25}$$

is normally distributed with

$$E(\Delta\hat{s}^2) = \Delta\sigma^2 = 0 \tag{26}$$

$$\sigma(\Delta\hat{s}^2) = \sqrt{\frac{2\hat{s}_1^4}{n_1-1} + \frac{2\hat{s}_2^4}{n_2-1}} \tag{27}$$

in which the value of the test statistic $Z^* = \frac{\Delta\hat{s}^2 - 0}{\sigma(\Delta\hat{s}^2)}$ (28) is compared to the RR of $\pm Z_{\alpha/2}$ (29).

For all other values of n_1 and n_2, H_0 and H_1 are formulated as:

$$\mathrm{H}_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \ vs. \ \mathrm{H}_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1. \tag{30}$$

The estimator of $\frac{\sigma_1^2}{\sigma_2^2}$ is $\frac{\hat{s}_1^2}{\hat{s}_2^2}$ which is distributed as $F_{n_2-1}^{n_1-1}$.

The test is implemented by comparing the value of the test statistic

$$F^* = \frac{\hat{s}_1^2}{\hat{s}_2^2} \tag{31}$$

against the rejection/acceptance regions which are defined by

$$F_{n_2-1}^{n_1-1}\left(\frac{\alpha}{2}\right) \tag{32}$$

$$F_{n_2-1}^{n_1-1}\left(1-\frac{\alpha}{2}\right) = \frac{1}{F_{n_1-1}^{n_2-1}\left(\frac{\alpha}{2}\right)}. \tag{33}$$

To test the more general case that the variances of k ($k \geq 3$) populations are equal, the Barlett test for homogeneity of variance is used and the pairs of hypotheses are tested: H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$ $vs.$ H_1 (34), at least two variances differ (Vasilopoulos, 2007).

To complete testing this pair of hypotheses, the following will be done: obtaining independent samples of sizes n_1, n_2,…, n_k from the k populations; calculating the k unbiased sample variances $\hat{s}_1^2, \hat{s}_2^2, \hat{s}_3^2, \ldots, \hat{s}_k^2$, from the k independent samples; calculating the Barlett test statistic B from

$$\mathrm{B} = \frac{[\sum_{i=1}^{k}(n_i-1)]ln\left[\frac{\sum_{i=1}^{k}(n_i-1)\hat{s}_i^2}{\sum_{i=1}^{k}(n_i-1)}\right] - [\sum_{i=1}^{k}(n_i-1)ln\hat{s}_i^2]}{1+\frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{(n_i-1)} - \frac{1}{\sum_{i=1}^{k}(n_1-1)}\right]} \tag{35}$$

$$\mathrm{B} = \frac{[(n_1-1)+(n_2-1)+(n_3-1)+\cdots+(n_k-1)]ln\left[\frac{(n_1-1)\hat{s}_1^2+(n_2-1)\hat{s}_2^2+\cdots+(n_k-1)\hat{s}_k^2}{(n_1-1)+(n_2-1)+(n_3-1)+\cdots+(n_k-1)}\right]}{1+\frac{1}{3(k-1)}\left[\frac{1}{n_1-1}+\frac{1}{n_2-1}+\cdots+\frac{1}{n_k-1} - \frac{1}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}\right]}$$

$$-\frac{[(n_1-1)ln\hat{s}_1^2+(n_2-1)ln\hat{s}_2^2+(n_3-1)ln\hat{s}_3^2+\cdots+(n_k-1)ln\hat{s}_k^2]}{1+\frac{1}{3(k-1)}\left[\frac{1}{n_1-1}+\frac{1}{n_2-1}+\cdots+\frac{1}{n_k-1} - \frac{1}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)}\right]}. \tag{36}$$

The test statistic B under the H_0 hypothesis of equal variances has a sampling distribution which is Chi-square with DOF = $k - 1$, or $\mathrm{B} = \chi_{k-1}^2$.

The rejection region, for a given "α" value, consists of the upper tail of the x^2 distribution (i.e. $\chi_{k-1}^2(\alpha)$).

The decision rule is: Do not reject H_0 if $\mathrm{B} \leq \chi_{k-1}^2(\alpha)$ and reject H_0 if $\mathrm{B} > \chi_{k-1}^2(\alpha)$.

Suppose this paper wants to test the equality of the variances of three populations, it means that it wants to test the hypotheses H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$ $vs.$ H_1: The three variances are not all equal.

Assume that random sampling from the three populations produced the results:

Population 1: $n_1 = 16$ and $\hat{s}_1^2 = 10$;

Population 2: $n_2 = 21$ and $\hat{s}_2^2 = 7$;

Population 3: $n_3 = 26$ and $\hat{s}_3^2 = 4$;

Then is H_0 rejected or not rejected, based on this sample information, when α = 0.1, α = 0.05, α = 0.025, α = 0.01, α = 0.005.

The solutions will be got in Table 1.

Table 1

Partial Calculation of the Barlett Test Statistic B

Population	n_i	$n_i - 1$	$\hat{s}_i^2$	$(n_i - 1)\hat{s}_i^2$	$ln\hat{s}_i^2$	$(n_i - 1)ln\hat{s}_i^2$
1	16	15	10	150	2.302585	34.538776
2	21	20	7	140	1.945910	38.918203
3	26	25	4	100	1.386294	34.657359
		$\sum_{i=1}^{3}(n_i - 1) = 60$		$\sum_{i=1}^{3}(n_i - 1)\hat{s}_i^2 = 390$		$\sum_{i=1}^{3}(n_i - 1)ln\hat{s}_i^2 = 108.114$

$$\sum_{i=1}^{k}\frac{1}{n_i-1} = \frac{1}{15} + \frac{1}{20} + \frac{1}{25} = \frac{20+15+12}{300} = \frac{47}{300} = 0.156667$$

$$ln\left[\frac{\sum_{i=1}^{k}(n_i-1)\hat{s}_i^2}{\sum_{i=1}^{k}(n_i-1)}\right] = ln\left[\frac{\sum_{i=1}^{3}(n_i-1)\hat{s}_1^2}{\sum_{i=1}^{3}(n_i-1)}\right] = ln\left[\frac{390}{60}\right] = \ln(6.5) = 1.87180.$$

Then, substituting the above quantities into equation (35), it can be obtained

$$B = \frac{60(1.87180)-108.114}{1+\frac{1}{3(3-1)}\left[\frac{47}{300}-\frac{1}{60}\right]} = \frac{112.308-108.114}{1+\frac{1}{6}\left[\frac{47}{300}-\frac{5}{300}\right]} = \frac{4.19413}{1+\frac{1}{6}\left(\frac{42}{300}\right)} = \frac{4.19413}{1+0.02333} = \frac{4.19413}{1.02333} = 4.0985$$

since, $k = 3$, $\chi^2_{k-1}(\alpha) = \chi^2_2(\alpha) = 4.605$, if $\alpha = 0.10$; $\chi^2_{k-1}(\alpha) = 5.991$ if $\alpha = 0.05$; $\chi^2_{k-1}(\alpha) = 7.378$ if $\alpha = 0.025$; $\chi^2_{k-1}(\alpha) = 9.210$ if $\alpha = 0.01$; $\chi^2_{k-1}(\alpha) = 10.597$ if $\alpha = 0.005$.

Since, $B = 4.0985 < \chi^2_2(\alpha)$, for all of these alpha values, H_0 is not rejected and it concludes that the three population variances are equal, or $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$.

Some of the Most Important Non-parametric Applications of the Chi-square Distribution

The most important non-parametric applications of the Chi-Square (x^2) include the following tests: Tests on frequencies, tests on the frequency of two classes, tests on the frequency of more than two classes, tests on goodness of fit (Hwang & Wang, 2008), for the uniform distribution, for the normal distribution, tests on independence—contingency tables (McHugh, 2013), and tests on homogeneity—are two or more independent random samples drawn from the same population?

This paper proceeds to briefly discuss how each of these tests is implemented, using the Chi-square distribution, after some general comments which apply to all of these tests.

To perform any one of these test, the general procedure below (Canal & Micciolo, 2014) is as follows: formulate H_0 under which expected (or theoretical) frequencies are determined, analyze sample data to establish observed frequencies, and compare the two sets of frequencies by forming corresponding differences.

Specify, on the basis of these differences, a decision criterion to determine whether the observed frequencies differ (or do not differ) significantly from the expected frequencies is given.

If the differences are small and can be attributed to chance variation in random sampling, H_0 is not rejected. But if the differences are large and cannot be attributed to chance variation in random sampling, H_0 is rejected.

Let o_i represent observed frequencies, e_i represent expected frequencies, $o_i - e_i$ represent the difference between corresponding observed and expected frequencies, then

$$Y = \frac{(o_1-e_1)^2}{e_1} + \frac{(o_2-e_2)^2}{e_2} + \cdots + \frac{(o_k-e_k)^2}{e_k} = \sum_{i=1}^{k} \frac{(o_i-e_i)^2}{e_i}. \tag{37}$$

It is a Chi-square variable (χ^2_δ) with degrees of freedom

$$\delta = k - 1 - m \tag{38}$$

where, k = number of classes, m = number of estimators calculated before expected frequencies are calculated ($m = 0, 1, 2, \ldots$)

Each of the k theoretical classes must have at least five items in it, for the Chi-square approximation to be valid. If one or more classes have expected frequencies of fewer than five items, the classes will need to be combined before forming the differences $o_i - e_i$, and determine the degrees of freedom δ, after the regrouping of classes. When $\delta = 1$, equation (37) is modified by introducing a continuity correction factor of 1/2 in computing the value of the variable Y or

$$Y = \chi_s^2 = \begin{cases} \sum_{i=1}^{k} \frac{\left(|o_1-e_1|-\frac{1}{2}\right)^2}{e_1} & if\ |o_i - e_1| \geq \frac{1}{2} \\ 0 & if\ |o_i - e_1| < \frac{1}{2} \end{cases}. \tag{39}$$

They consist of problems, which test the frequencies of two classes (in which the continuity correction factor of 1/2 must be used) and tests on the frequencies of more than two classes (in which the continuity correction factor of 1/2 is not used). The use of this test will be illustrated using the following example.

Three coins are tossed 100 times to determine whether the three coins are fair and the following data were obtained by counting the number of times obtained (0, 1, 2, 3 heads, i.e. these are the observed frequencies), as shown in Table 2 below.

Table 2

Observed Frequencies

X	0	1	2	3
$O_i(X) = n(X)$	14	34	36	16

The expected frequencies are obtained by using the binomial law or

$$100b\left(x, n=3, p=\frac{1}{2}\right) = 100\left[\frac{n!}{x!n-x)!}\mathrm{P}^x(1-p)^{n-x}\right] \tag{40}$$

where $b\left(x, n=3, p=\frac{1}{2}\right)$ is the binomial distribution with $n = 3$ and $p = 1/2$.

When $n = 3$, $x = 0, 1, 2, 3$, from equation (40), it can be obtained that P(x = 0) = 1/8, P(x= 1) = 3/8, and $P(x$ = 2) = 3/8, and P(x = 3) = 1/8, and 100P(x = 0) = 100(1/8) = 12.5, 100P(x = 1) = 100(3/8) = 37.5, 100P(x= 2) = 100(3/8) = 37.5, 100P(x = 3) = 100(1/8) = 12.5.

Table 3

Expected Frequencies

x_i	0	1	2	3
E(x_i)	12.5	37.5	37.5	12.5

Therefore, the expected frequencies are as shown in (Table 3).

Then this paper performs the following test to check the validity of the claim:

H_0: The distribution of heads is the same as that produced by three fair coins.

H_1: The distribution of heads is different from that produced by three fair coins.

Let the level of significance $\alpha = 0.05$, define the test statistic $Y = x_\delta^2 = x_3^2 = \sum_{i=1}^{k=4}\frac{\left(o_i - e_i\right)^2}{e_i}$, with $k = 4$ and $\delta = k - 1 = 3$ or

$$Y = x_\delta^2 = x_3^2 = \frac{\left(o_1 - e_1\right)^2}{e_1} + \frac{\left(o_2 - e_2\right)^2}{e_2} + \frac{\left(o_3 - e_3\right)^2}{e_3} + \frac{\left(o_4 - e_4\right)^2}{e_4}$$

$$= \frac{\left(14 - 12.5\right)^2}{12.5} + \frac{\left(34 - 37.5\right)^2}{37.5} + \frac{\left(36 - 37.5\right)^2}{37.5} + \frac{\left(16 - 12.5\right)^2}{12.5}$$

$$= 0.18+0.32667+0.06+0.98=1.54667.$$

The RR must be the upper end of the Chi-square distribution because any departure from the expected frequencies will result in an increased value for the Chi-square value. Then the RR is defined by

$$x_{k-1}^2(\alpha) = x_3^2(0.05) \geq 7.815$$

Do not reject H_0, because $Y = 1.54667 < x_3^2(0.05) \geq 7.815$. Therefore the three coins are fair.

On many occasions, it appears that the population under investigation follows a specific probability model, such as uniform, normal, exponential, etc.. A statistical procedure exists which can be used to verify the validity of such preliminary conclusions, called a "goodness of fit" test and consists of the following steps (Carlson & Thorne, 1997).

(a) formulating the null hypothesis that a given population has a specific probability model (such as: uniform, normal, poisson, exponential, etc.);

(b) obtaining a random sample from the population and analyze it to derive the observed frequencies;

(c) using the theoretical distribution, specified in H_0 to generate expected frequencies, by multiplying the probability values for the classes by the sample size;

(d) after these preliminary steps, the Chi-square test for goodness-of-fit becomes similar to the procedure used in the test for frequencies;

(e) illustrating the use of this test using the following example (Freund & Williams, 1982):

Supposing a die is tossed 120 times, the following results are obtained, as shown in Table 4 below.

Table 4

Observed Occurrences

Number showing	1	2	3	4	5	6	Total
Observed occurrences	10	19	30	29	21	11	120

Are these results consistent with the hypothesis that the die is fair at $\alpha = 0.01$?

Following the general testing procedure, the goodness-of-fit test becomes:

H_0: The number showing is uniformly distributed (or the die is fair);

H_1: The number showing is not uniformly distributed (or the die is not fair).

Let the level of significance $\alpha = 0.01$, The test statistic is $Y = \chi^2_{\delta=6-1} = \chi^2_5 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i}$. The observed data o_i comes from the table above, while the expected data comes from the assumed uniform distribution which implies that P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6, and the expected number of occurrences for each number are equal to 120P(1) = 120P(2) = 120P(3) = 120P(4) = 120P(5) = 120P(6) = 20.

Then

$$Y = x_5^2 = \sum_{i=1}^{k=6} \frac{\left(o_i - e_i\right)^2}{e_i} = \frac{\left(o_1 - e_1\right)^2}{e_1} + \frac{\left(o_2 - e_2\right)^2}{e_2} + \frac{\left(o_3 - e_3\right)^2}{e_3} + \frac{\left(o_4 - e_4\right)^2}{e_4} + \frac{\left(o_5 - e_5\right)^2}{e_5} + \frac{\left(o_6 - e_6\right)^2}{e_6}$$

$$= \frac{(10-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(29-20)^2}{20} + \frac{(21-20)^2}{20} + \frac{(11-20)^2}{20}$$

$$= \frac{1}{20}\left[(-10)^2 + (-1)^2 + (10)^2 + (9)^2 + (1)^2 + (-9)^2\right] = \frac{1}{20}(364) = 18.2 .$$

The rejection region is the upper tail of the Chi-square distribution and is given by:

$$x_5^2 (0.01) \geq 15.806.$$

Reject H_0, because Y =18.2 > x_5^2 (0.01) = 15.806, and conclude that the die is not fair (or that the numbers showing are not uniformly distributed).

In the test on frequencies discussed above populations, samples are classified by a single characteristic. When populations or samples are classified by two (or more) characteristics, this paper uses "tests of independence" to determine whether the characteristics are statistically independent or not. Tests for independence are also called "contingency-table tests", and the types of hypotheses being tested here are (McClave et al., 2001):

H_0: The two characteristics are independent;

H_1: The two characteristics are dependent.

In such tests, the observed frequencies may, in general, occupy r rows and c columns, while the smallest possible "contingency table test" consists of two rows and two columns. For each observed frequency in an $r \times c$ contingency table, there is a corresponding expected frequency which is defined by the null hypothesis (H_0) being tested. The total frequencies in each row or column are called "marginal frequencies", while the observed and/or expected frequencies of each cell of the contingency table are called "cell frequencies".

To test the hypotheses above, the following is used

$$x_{\delta}^{2} = \sum_{i=1}^{rc} \frac{\left(o_i - e_i\right)^2}{e_i} \tag{41}$$

which is the same x^2 test used in the frequency tests, except that

$$\delta = \text{degrees of freedom} = (r - 1) \times (c - 1) \tag{42}$$

when $r = 2$ and $c = 2$ (smallest possible contingency table), $\delta = 1$, for large samples the continuity correction factor can be ignored. However, for small samples, the continuity correction factor of 1/2 should be used in equation (39).

To test the hypothesis that high-income families choose to send their children to private universities and low-income families to state universities, 2,000 families were selected at random, nationwide, and the following results were obtained, shown in Table 5 below.

Table 5

Observed Data

Income level	University type		Totals
	Private	Public	
Low	632	618	1,250
High	548	202	750
Totals	1,180	820	2,000

From this table, it is obvious that a greater proportion of high-income families 548/750 = 0.73 send their children to private universities than low-income families 632/1250 = 0.51. To determine whether this proportional difference is statistically significant or not, this paper uses the Chi-square test for independence. But this paper first constructs a table of expected data based on assumption that income level and type of university are independent. Under this assumption expect the proportion of all families that send their children to private universities, it will be equal to 1,180/2,000 = 0.59. Then, the expected number of low-income families that send their children to private universities is = (1,180/2,000) × 1,250 = 738. The other cell values could be calculated in a similar manner. But, since the "marginal frequencies" of the observed data are known, once one of the cell values has been found, the other cell values can be calculated by inspection. This statement

explains why the degrees of freedom = DOF = $\delta = 1$ for a 2 × 2 contingency table. Then the expected data are shown in Table 6 below.

Table 6

Expected Data

Income level	University type		Totals
	Private	Public	
Low	738	512	1,250
High	442	308	750
Totals	1,180	820	2,000

The tests of independence proceeds are as follows:

H_0: Income level and university choice are independent;

H_1: Income level and university choice are dependent.

Level of significance α = 0.05, test statistic, *Y*, with $\delta = (r - 1) \times (c - 1) = (2 - 1) \times (2 - 1) = 1$.

RR is defined by $\chi^2_\delta(\alpha) = \chi^2_1(\alpha) = \begin{Bmatrix} 2.706 & if & \alpha = 0.1 \\ 3.801 & if & \alpha = 0.05 \\ 5.024 & if & \alpha = 0.025 \\ 6.635 & if & \alpha = 0.01 \end{Bmatrix}$

$$Y = \chi^{2*} = \frac{(632-738)^2}{738} + \frac{(618-512)^2}{512} + \frac{(548-442)^2}{442} + \frac{(202-308)^2}{308} = 99.072.$$

Since x^{2*} falls in the rejection region (for all α values shown), H_0 is rejected and it concludes that family income level and type of university selection are not independent.

The continuity correction factor (CCF) of 1/2 was not used (in the calculation of x^{2*}), because the sample size is large (see equation 39).

When it is said that "things" are homogeneous, it means that they have something in common, that they are the same, or that they are equal. Here is the question: Are two or more independent random samples drawn from the same population or from different populations? Tests on homogeneity can be considered as an extension of the Chi-square test for independence. Both of these tests are concerned with the analysis of cross-sectional data, and both use the same test statistic $x^2_\delta = \sum_{i=1}^{rc} \frac{\left(o_i - e_i\right)^2}{e_i}$.

But these tests also have their differences, which mainly are due to the types of problems solved in each case (Vasilopoulos, 2007).

In tests of independence, a single sample is obtained from one population and the problem is to determine whether two characteristics of the elements of the population, from which the sample came from, are independent of each other (Lang & Iannario, 2013).

In tests of homogeneity, however, two or more independent samples have been obtained and the problem is to determine whether these samples come from the same population or from different populations.

Three random samples of students are taken at a university. The first sample consists of 100 graduate students, the second of 100 seniors, and the third of 100 sophomore students. Each group is asked to grade the course instruction they are receiving at the university as excellent, good, or average. The following results were

obtained (observed data), as shown in Table 7 below.

Table 7

Observed Data for Example Above

Student classification	Instruction quality			
	Excellent	Good	Average	Totals
Graduate	77	12	11	100
Senior	73	7	20	100
Sophomore	85	10	5	100
Totals	235	29	36	300

The null hypothesis being tested here is: H_0: The three samples come from the same population (i.e. the three classifications of students are homogeneous in their opinion about quality of instruction). If this hypothesis is true, then the best estimates for the proportions specifying.

"Excellent instruction", "good instruction", and "average instruction", respectively, should be 235/300, 29/300, and 36/300. Therefore, for the 100 graduate students, the expected frequencies for the three categories become 235/300 = 78.33, 29/300 = 9.67, and 36/300 = 12 and similarly for the 100 senior and 100 sophomore students.

Therefore, the expected data are shown in Table 8 below.

Table 8

Expected Data for Example Above

Student classification	Instruction quality			
	Excellent	Good	Average	Totals
Graduate	78.33	9.67	12.00	100
Senior	78.33	9.67	12.00	100
Sophomore	78.33	9.67	12.00	100
Totals	234.99	29.01	36.00	300

The test for homogeneity becomes:

H_0: The three samples are drawn from the same population;

H_1: The three samples are drawn from different populations.

Level of significance $\alpha = 0.05$, test statistic

$$x_{\delta}^{2} = \sum_{i=1}^{rc} \frac{\left(o_i - e_i\right)^2}{e_i} = \sum_{i=1}^{rc} \frac{o_i^{\,2}}{e_i} - n = \sum_{i=1}^{9} \frac{o_i^{\,2}}{e_i} - 300 \tag{43}$$

with $\delta = (r - 1) \times (c - 1) = (3 - 1) \times (3 - 1) = 4$.

RR is defined by $x_{\delta(\alpha)}^{2} = x_4^2$ $(\alpha = 0.05) \geq 9.488$.

Value of test statistic is:

$$\Upsilon = x^{2\phi} = \left(\frac{77^2}{78.33} + \frac{12^2}{9.67} + \frac{11^2}{120}\right) + \left(\frac{73^2}{78.33} + \frac{7^2}{9.67} + \frac{20^2}{120}\right) + \left(\frac{85^2}{78.33} + \frac{10^2}{9.67} + \frac{5^2}{120}\right) = 311.75 - 300 = 11.75$$

Since $\chi_4^{2*} = 11.75 > \chi_4^2(0.05) = 9.488$, H_0 is rejected and it concludes that "different level students"

have "different opinions" concerning the quality of instruction at the university.

The definition of $x_{\delta}^{2} = \sum_{i=1}^{rc} \frac{\left(o_i - e_i\right)^2}{e_i}$ can be shown to be equal to $\sum_{i=1}^{rc} \frac{o_i^{\,2}}{e_i} - n$. This equivalent formula was used in the calculation of x^{2*}, and it appears to be somewhat easier, because it avoids having to take differences between observed and expected frequencies before squaring.

If the level of significance was changed from $\alpha = 0.05$ to $\alpha = 0.01$, then $x_4^2(0.01) = 13.277$, and H_0 is not rejected and it would conclude that the three samples came from the same population (or that the three classifications of students are homogeneous in their opinion about the quality of instruction).

Conclusions

The Chi-square (x^2) distribution is very versatile and has many applications, some of them parametric and some of them nonparametric. It is used, as a parametric test, to solve hypothesis test problems and construct confidence intervals for the population parameter σ^2 (and also σ), because the sampling distribution of $\hat{s}^2$, which is the estimator if σ^2 is distributed as a Chi-square variable with $n - 1$ degrees of freedom (χ^2_{n-1}). But the Chi-square (x^2) distribution can also be used as a non-parametric test (or distribution-free statistic) to perform the following tests: tests on frequencies, goodness-of-fit tests, test on independence or contingency-table tests, tests on homogeneity (two or more independent random samples drawn from the same population of from different populations).

Most of the parametric tests discussed (test on σ^2 and σ for example) are well known, but the test on the equality of three or more population variances, using the Barlett B statistic, is not.

This paper has discussed many nonparametric tests in which the parameters of the distribution continue to be important, but the nature of the distribution, from which the sample data used in the analysis came, is not important and is not needed to perform these tests. This is in contrast to the parametric tests, which depend very much on the nature of the population from which the data set came from.

Some of the non-parametric tests have corresponding parametric tests; but the majority of them do not.

Examples were included, for all the tests discussed, to make their understanding and applications easier.

The non-parametric methods can solve the same type of problems that parametric methods can solve (but with reduced efficiency) and can also solve additional problems when no parametric methods are available.

The use of a statistic software tool, like minitab, simplifies the application of these tests considerably. Unfortunately, not every non-parametric test is supported by minitab, but most parametric tests are.

Minitab uses the *p*-value (i.e. observed level of significance), instead of α (the *a*-priori level of significance), to reject or not to reject a hypothesis.

References

Berenson, M., Levine, D., & Krehbiel, T. (2004). *Basic business statistics* (9th ed.). Upper Saddle River: Prentice Hall.

Black, K. (2004). *Business statistics* (4th ed.). Hoboken*:* Wiley.

Canal, L., & Micciolo, R. (2014). The chi-square controversy: What if Pearson had R? *Journal Of Statistical Computation & Simulation*, *84*(5), 1015-1021.

Canavos, G. C. (1984). *Applied probability and statistical methods*. Boston: Little Brown.

Carlson, W., & Thorne, B. (1997). *Applied statistical methods.* Upper Saddle River: Prentice Hall.

Chou, Y. (1992). *Statistical analysis for business and economics*. New York: Elsevier.

Freund, J., & Williams, F. (1982). *Elementary business statistics: The modern approach*. Upper Saddle River: Prentice Hall.

Hwang, Y., & Wang, C. (2008). A goodness of fit test for left-truncated and right-censored data. *Statistics & Probability Letters, 78*(15), 2420-2425.

Lang, J., & Iannario, M. (2013). Improved tests of independence in singly-ordered two-way contingency tables. *Journal of Computational Statistics & Data Analysis, 68,* 339-351.

McClave, T., Benson, P., & Sincich, T. (2001). *Statistics for business and economics* (8th ed.). Upper Saddle River: Prentice Hall.

McHugh, M. (2013). The Chi-square test of independence. *Biochemia Medica, 23*(2), 143-149.

Salvatore, D. (1982). *Theory and problems of statistics and econometrics.* SCHAUM'S OUTLINE SERIES. New York: McGraw-Hill.

Vasilopoulos, A. (2007). *Business statistics—A logical approach.* Boston: Pearson Custom Publishing.

Xie, C. (2014). Relations among three parametric multiple testing methods for correlated tests. *Journal Of Statistical Computation & Simulation, 84*(4), 812-818.