International Journal of Statistics and Probability; Vol. 13, No. 4; December 2024
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On Invariance of Chi-squared Tests Under Different Probability Models

Khairul Islam[1] & Tanweer J. Shapla[2]

[1,2] Department of Mathematics and Statistics, Eastern Michigan University, USA.

Correspondence: Khairul Islam, Department of Mathematics and Statistics, Eastern Michigan University, Ypsilanti, MI 48197, USA

Received: October 10, 2024 Accepted: November 26, 2024 Online Published: November 28, 2024

doi:10.5539/ijsp.v13n4p18 URL: https://doi.org/10.5539/ijsp.v13n4p18

Abstract

A chi-squared test is a popular test for assessing relationship between two factors or categorical variables summarized in the form of a contingency table. In this study, we establish the invariance of a chi-squared test under three different study designs, namely, cohort, case-control and cross-sectional studies involving distinct probabilistic models. By the invariance of the chi-squared test**,** we refer to the fact that the form of a chi-squared test remains unchanged under different probabilistic models. The theoretical derivation of expected cell frequencies carried out in this study, under different study designs and probability models, will be exemplary and invaluable to researchers to understand as to why they can use an identical form of the chi-squared test for a contingency table resulting from case-control, cohort or cross-sectional study design for testing independence. This study is also useful in academia to demonstrate why contingency table resulting under different study designs is subject to identical form of a chi-squared test, which has not been well documented in existing literature. The examples and applications utilized in this study provide directions as to how differently formulated studies are implemented via a chi-squared test.

Keywords: Chi-squared test, Invariance property, Cohort study, Case-control study, Cross-sectional study

1. Introduction

The chi-squared test, also known as the Pearson Chi-squared test, is a popular test for independence of two categorical variables summarized in the form of a contingency table (Agresti, 2012; Casella & Berger, 2024; Mann, 2003, 2012; Perez-Guerrero et al., 2024). As a matter of fact, a contingency table may result from any of the three study designs, namely, case-control, cohort and cross-sectional studies. Given a contingency table, a chi-squared test is carried out by comparing observed cell frequencies with the expected frequency distributions. Due to simplicity of designs, a 2×2 contingency table is widely employed in clinical trials for testing independence of a risk factor and disease outcome. A 2×2 contingency table is generally presented via a table of observed frequencies or unknown probabilities as follows:

Table 1. Frequency distribution of subjects and probabilities in a 2×2 contingency table

Table of observed frequencies

Factor X	Outcome Y		Total
	Label 1	Label 2	
Label 1	n_{11}	n_{12}	n_{1+}
Label 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Table of unknown probabilities

Factor X	Outcome Y		Total
	Label 1	Label 2	
Label 1	π_{11}	π_{12}	π_{1+}
Label 2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	π_{++}

where

(1) n_{ij} refers to the random frequency in ith row and jth column; $i, j = 1, 2$

(2) π_{ij} is the unknown probability that a subject belongs to cell (i, j), for which the observed frequency is n_{ij}

(3) It appears that $n_{+1} + n_{+2} = n_{1+} + n_{2+} = n$, the total number of subjects in the study

Under above notations, the test of independence is carried out by employing the Pearson's chi-squared test given by

$$Q_P = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi_1^2 \quad (1)$$

where $e_{ij} = E(n_{ij})$ is the expected frequency of the random frequency n_{ij}, which depends on the π_{ij}, the probability with which n_{ij} is being observed. Of course, π_{ij} is unknown and needs to be estimated given the observed frequency n_{ij}. In existing literature, it is noted that e_{ij} is estimated by the relation

$$e_{ij} = \frac{n_{i+}n_{+j}}{n} \tag{2}$$

It is important to note that a contingency table may result from a case-control, cohort or cross-sectional study design and as such an estimation of π_{ij} and hence e_{ij} need to be investigated thoroughly to evaluate if the relation (2) holds for all three designs, namely, case-control, cohort and cross-sectional study designs, which is absent in existing literature.

In this study, we establish chi-squared tests under three different study designs, which involves distinct forms of probabilistic models. It appears that e_{ij} follows an invariance property under different probability models and therefore, the form of the chi-squared test remains the same under different probabilistic models, which we term as the invariance of the chi-squared test. Therefore, a contingency table resulting from case-control, cohort and cross-sectional studies enjoys the same form of the chi-squared test. This study is significant in that it provides theoretical justification of identical chi-squared test resulting from varying probability models.

2. Methods

Case-control, cohort and cross-sectional studies are collectively referred to as observational studies and are very popular in epidemiological research (McMugh, 2013; Omair, 2016; Robert & Berman, 2015; Song & Chung, 2010; Taur, 2022). These studies are particularly useful to study the relationship between exposure to a risk factor and disease outcome. In this section, we provide a brief description of various study designs and investigate forms of the corresponding chi-squared tests in reference to the test of independence.

2.1 Case-control Study

A case-control study (McMugh, 2013; Robert & Berman, 2015; Song & Chung, 2010; Taur, 2022), also called the retrospective study, is an epidemiological study which starts with two samples categorized as diseased (also called a case) and non-diseased (also called a control), and then look back to their past to see who in diseased and non-diseased samples had exposure and non-exposure to a specified factor. The objective is to assess if disease outcome and exposure to the factor is related or not. If the rate of exposure among diseased and non-diseased samples are equal, then the diseased outcome and exposure to the factor is said to be independent.

Under design of a case-control study, suppose two samples of subjects, one from case (diseased, D=1) of size n_{+1} and the other from control (without disease, D=2) of size n_{+2}, are asked about their past history (←) of being exposed (E=1) or non-exposed (E=2) to a factor. The contingency table resulting from the design may be presented as follows:

Table 2. Frequency distribution of subjects and probabilities in 2×2 contingency table for a case-control study

Exposed status	← Disease Status		Total
	D=1	D=2	
E=1	n_{11}	n_{12}	n_{1+}
E=2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Exposed status	← Disease Status		Total
	D=1	D=2	
E=1	π_{11}	π_{12}	π_{1+}
E=2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}=1	π_{+2}=1	

In above table, n_{ii} refers to the observed frequency of subjects with disease status $D = j$ and exposure status $E = i$, and π_{ij} is the unknown probability of observing $n_{ij}; i, j = 1, 2$. In a case control study, $n_{11} \sim B(n_{+1}, \pi_{11})$ and $n_{12} \sim B(n_{+2}, \pi_{12})$ are independent with $\pi_{+1} = \pi_{+2} = 1$ and corresponding expected frequency $e_{11} = n_{+1}\pi_{11}$ and $e_{12} = n_{+2}\pi_{12}$. By the property of the binomial distribution, it also follows that $e_{21} = n_{+1}\pi_{21} = n_{+1}(1 - \pi_{11})$ and $e_{22} = n_{+2}\pi_{22} = n_{+2}(1 - \pi_{12})$. Therefore, to employ Pearson's chi-squared test of independence, we need to estimate unknown parameters π_{ij}, which we wish to achieve via maximum likelihood method (Casella & Berger, 2024; Agresti, 2012).

Under this study design, an exposure to the factor is said to be independent of disease outcome if the rate of exposure in disease and non-disease are identical, i.e., $P(E = 1|D = 1) = P(E = 1|D = 2)$. In other words, for the test of independence, we test the null hypothesis

$$H_0: P(E = 1|D = 1) = P(E = 1|D = 2)$$

$$\Rightarrow H_0: \frac{P(E=1,D=1)}{P(D=1)} = \frac{P(E=1,D=2)}{P(D=2)} \Rightarrow H_0: \frac{\pi_{11}}{\pi_{+1}} = \frac{\pi_{12}}{\pi_{+2}} \Rightarrow H_0: \pi_{11} = \pi_{12} \quad (3)$$

Now, to find the maximum likelihood estimates (MLEs) of unknown parameters π_{ij} for a case-control study design, we maximize the joint likelihood function of n_{11} and n_{12} given by

$$L = \binom{n_{+1}}{n_{11}} {\pi_{11}}^{n_{11}} (1-\pi_{11})^{n_{21}} \times \binom{n_{+2}}{n_{12}} {\pi_{12}}^{n_{12}} (1-\pi_{12})^{n_{22}}$$

$\Rightarrow logL = k_1 + n_{11}\log(\pi_{11}) + n_{21}\log(1-\pi_{11}) + n_{12}\log(\pi_{12}) + n_{22}\log(1-\pi_{12})$, where k_1 is a constant, independent of π_{ij}, and hence is irrelevant for estimating π_{ij}.

To find the MLEs, we solve $\frac{\partial logL}{\partial \pi_{11}} = 0$ and $\frac{\partial logL}{\partial \pi_{12}} = 0$. Note that

$$\frac{\partial logL}{\partial \pi_{11}} = \frac{n_{11}}{\pi_{11}} + \frac{n_{21}}{1-\pi_{11}}(-1) = 0 \Rightarrow \frac{\pi_{11}}{1-\pi_{11}} = \frac{n_{11}}{n_{21}} \Rightarrow \hat{\pi}_{11} = \frac{n_{11}}{n_{+1}}$$

Therefore, $\hat{\pi}_{21} = 1 - \hat{\pi}_{11} = 1 - \frac{n_{11}}{n_{+1}} = \frac{n_{21}}{n_{+1}}$

Similarly,

$$\frac{\partial logL}{\partial \pi_{12}} = \frac{n_{12}}{\pi_{12}} + \frac{n_{22}}{1-\pi_{12}}(-1) = 0 \Rightarrow \frac{\pi_{12}}{1-\pi_{12}} = \frac{n_{12}}{n_{22}} \Rightarrow \hat{\pi}_{12} = \frac{n_{12}}{n_{+2}}$$

and $\hat{\pi}_{22} = 1 - \hat{\pi}_{12} = 1 - \frac{n_{12}}{n_{+2}} = \frac{n_{22}}{n_{+2}}$

Putting all estimates together, for a case-control study we have $\hat{\pi}_{ij} = \frac{n_{ij}}{n_{+j}},\ i,j = 1,2.$

Therefore, under the null hypothesis H_0 of equation (3), the common value of π_{11} and π_{12} are estimated by the pooled estimate:

$\hat{\pi}_{11} = \hat{\pi}_{12} = \frac{n_{11}+n_{12}}{n_{+1}+n_{+2}} = \frac{n_{1+}}{n}$, which implies that $e_{11} = \frac{n_{1+}n_{+1}}{n}$ and $e_{12} = \frac{n_{1+}n_{+2}}{n}$

In a similar manner, using the fact that $\pi_{21} = 1 - \pi_{11}$ and $\pi_{22} = 1 - \pi_{12}$, it is easy to verify that

$$e_{21} = \frac{n_{2+}n_{+1}}{n} \text{ and } e_{22} = \frac{n_{2+}n_{+2}}{n}$$

Therefore, given a case-control study, the Pearson Chi-square test is carried out by the test statistic

$Q_P = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij}-e_{ij})^2}{e_{ij}} \sim \chi_1^2$, where $e_{ij} = \frac{n_{i+}\times n_{+j}}{n}; i,j = 1,2.$

2.2 Cohort Study

A cohort study is prospective (McMugh, 2013; Robert & Berman, 2015; Song & Chung, 2010; Taur, 2022) in that it starts with two samples categorized as exposed (E=1) and non-exposed (E=2), and then look forward to observing how many in the exposed and non-exposed samples develop an outcome of interest, say disease (D=1) and no-disease (D=2). The objective is to assess if rate of disease in exposed and non-exposed samples are same or not. If the rate of disease among exposed and non-exposed samples are equal, then the diseased outcome and exposure to the factor is said to be independent.

Under design of a cohort study, suppose two samples of subjects, one exposed to a factor (E=1) of size n_{1+} and the other unexposed to it (E=2) of size n_{2+}, are followed up ($\rightarrow$) for a certain period of time to see how many of them develop an outcome (D=1) and how many does not develop the outcome (D=2). Under the notation of a cohort study, a contingency table looks like as appears in Table 3.

Table 3. Frequency distribution of subjects and probabilities in 2×2 contingency table for a cohort study

Exposed status→	Disease Status D=1	D=2	Total
E=1	n_{11}	n_{12}	n_{1+}
E=2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Exposed status→	Disease Status D=1	D=2	Total
E=1	π_{11}	π_{12}	π_{1+}=1
E=2	π_{21}	π_{22}	π_{2+}=1
Total	π_{+1}	π_{+2}	

In above table, n_{ij} refers to the observed frequency of subjects with exposure status $E = i$ and disease status $D = j$, and π_{ij} is the unknown probability of observing $n_{ij}; i, j = 1, 2$. In a cohort study, $n_{11} \sim B(n_{1+}, \pi_{11})$ and $n_{21} \sim B(n_{2+}, \pi_{21})$ are independent with $\pi_{1+} = \pi_{2+} = 1$ and the expected frequencies $e_{11} = n_{1+}\pi_{11}$ and $e_{21} = n_{2+}\pi_{21}$. By the property of the binomial distribution, it also follows that $e_{12} = n_{1+}\pi_{12} = n_{1+}(1 - \pi_{11})$ and $e_{22} = n_{2+}\pi_{22} = n_{2+}(1 - \pi_{21})$. Therefore, all we need to employ the Pearson's chi-squared test of independence is to estimate the unknown parameters π_{ij}.

For a cohort study, disease and exposure to the factor is said to be independent if the prevalence of disease in exposed and non-exposed samples are identical, i.e., $P(D = 1|E = 1) = P(D = 1|E = 2)$. In other words, for the test of independence in a cohort study, we wish to test the null hypothesis

$$H_0: P(D = 1|E = 1) = P(D = 1|E = 2)$$

$$\Rightarrow H_0: \frac{P(E=1,D=1)}{P(E=1)} = \frac{P(E=2,D=1)}{P(E=2)} \Rightarrow H_0: \frac{\pi_{11}}{\pi_{1+}} = \frac{\pi_{21}}{\pi_{2+}} \Rightarrow H_0: \pi_{11} = \pi_{21} \quad (4)$$

In order to estimate π_{ij}, under the cohort study design, we employ an MLE method applied to the joint likelihood function of n_{11} and n_{21} given by

$$\text{L}=\binom{n_{1+}}{n_{11}} {\pi_{11}}^{n_{11}}(1 - \pi_{11})^{n_{12}} \times \binom{n_{2+}}{n_{21}} {\pi_{21}}^{n_{21}}(1 - \pi_{21})^{n_{22}}$$

$$\Rightarrow logL = k_2 + n_{11}\log(\pi_{11}) + n_{12}\log(1 - \pi_{11}) + n_{21}\log(\pi_{21}) + n_{22}\log(1 - \pi_{21})$$

where k_2 is a constant, independent of π_{ij}, and hence is irrelevant for estimating π_{ij}.

To find the MLEs, we solve $\frac{\partial logL}{\partial \pi_{11}} = 0$ and $\frac{\partial logL}{\partial \pi_{21}} = 0$.

Note that $\frac{\partial logL}{\partial \pi_{11}} = 0 \Rightarrow \frac{n_{11}}{\pi_{11}} + \frac{n_{12}}{1-\pi_{11}}(-1) = 0 \Rightarrow \frac{\pi_{11}}{1-\pi_{11}} = \frac{n_{11}}{n_{12}} \Rightarrow \hat{\pi}_{11} = \frac{n_{11}}{n_{1+}}$

Therefore, $\hat{\pi}_{12} = 1 - \hat{\pi}_{11} = 1 - \frac{n_{11}}{n_{1+}} = \frac{n_{12}}{n_{1+}}$

Similarly,

$\frac{\partial logL}{\partial \pi_{21}} = \frac{n_{21}}{\pi_{21}} + \frac{n_{22}}{1-\pi_{21}}(-1) = 0 \Rightarrow \frac{\pi_{21}}{1-\pi_{21}} = \frac{n_{21}}{n_{22}} \Rightarrow \hat{\pi}_{21} = \frac{n_{21}}{n_{2+}}$

and $\hat{\pi}_{22} = 1 - \hat{\pi}_{21} = 1 - \frac{n_{21}}{n_{2+}} = \frac{n_{22}}{n_{2+}}$

Putting all estimates together, for a cohort study, we have $\hat{\pi}_{ij} = \frac{n_{ij}}{n_{+j}}, \; i, j = 1, 2.$

Therefore, under the null hypothesis H_0 of equation (4), the common value of π_{11} and π_{21} are estimated by the pooled estimate:

$\hat{\pi}_{11} = \hat{\pi}_{21} = \frac{n_{11}+n_{21}}{n_{1+}+n_{2+}} = \frac{n_{+1}}{n}$, which implies that $e_{11} = \frac{n_{1+}n_{+1}}{n}$ and $e_{21} = \frac{n_{2+}n_{+1}}{n}$

In a similar manner, using the fact that $\hat{\pi}_{12} = 1 - \hat{\pi}_{11}$ and $\hat{\pi}_{22} = 1 - \hat{\pi}_{21}$, it is easy to verify that

$e_{12} = \frac{n_{1+}n_{+2}}{n}$ and $e_{22} = \frac{n_{2+}n_{+2}}{n}$.

Therefore, given a cohort study, the Pearson Chi-square test is carried out by the test statistic

$Q_P = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \sim \chi_1^2$, where $e_{ij} = \frac{n_{i+}\times n_{+j}}{n}; i,j = 1,2.$

2.3 Cross-sectional Study

A cross sectional study is the quickest way to determine the prevalence of infectious disease (Capili, 2021; McMugh, 2013; Robert & Berman, 2015; Song & Chung, 2010; Taur, 2022). In this study, researchers collect information of disease outcome (D=1 and D=2) and exposure status (E=1 and E=2) from a given sample of subjects at a single point in time. In other words, in a cross-sectional study, an evaluation of exposure and disease outcome on a sample of subjects takes place simultaneously. Given a sample of n subjects under a cross-sectional study, the resulting 2×2 contingency table of observed frequency distribution is as follows:

Table 4. Frequency distribution of subjects in 2×2 contingency table for a cross-sectional study

Exposure status	Disease Status		Total
	$D = 1$	$D = 2$	
Exposed $E = 1$	n_{11}	n_{12}	n_{1+}
Unexposed $E = 2$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

In above table, n_{ij} refers to the observed frequency of subjects with disease status $D = j$ and exposure status $E = i$, observed simultaneously, with unknown probability π_{ij}; $i,j = 1,2$. In a cross-sectional study,

$N = (n_{11}, n_{12}, n_{21}, n_{22}) \sim Multinomial(n, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$

with the constraint $\sum_{i,j=1}^{2}\pi_{ij} = 1$ and the expected frequencies $e_{ij} = n\pi_{ij}$, where we have to estimate the unknown probabilities $\pi_{ij}; i,j = 1,2$.

Also, for a cross-sectional study, the disease outcome and exposure to the factor is said to be independent if joint probability is the product of marginal probabilities. In other words, for the test of independence of disease outcome and exposure, the null hypothesis is

$$H_0: P(E = i \cap D = j) = P(E = i)P(D = j); i,j = 1,2$$
$$\Rightarrow H_0: \pi_{ij} = \pi_{i+}\pi_{+j} \quad (5)$$

To estimate π_{ij} and hence π_{i+} and π_{+j}, we use MLE applied to the likelihood function

$$L(\pi_{ij}) = \frac{n!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}\prod_{i,j=1}^{2}{\pi_{ij}}^{n_{ij}}$$

By taking the log, the log-likelihood function is

$$\log\left(L(\pi_{ij})\right) = k_3 + \sum_{i,j=1}^{2} n_{ij}\log(\pi_{ij})$$

where k_3 is a constant, independent of π_{ij} and hence is irrelevant in finding the MLE of π_{ij}. Note that to find the MLE with the constraint $\sum_{i,j=1}^{2}\pi_{ij} = 1$, we have to take the derivative of the Lagrangian function (Kalman, 2009).

$$\mathcal{L}(\lambda, \pi_{ij}) = \log\left(L(\pi_{ij})\right) + \lambda\left(1 - \sum_{i,j=1}^{2}\pi_{ij}\right)$$

By taking the derivative of $\mathcal{L}(\lambda, \pi_{ij})$ with respect to π_{ij}, setting derivative equal to zero and solving we get

$$\frac{\partial}{\partial\pi_{ij}}\mathcal{L}(\lambda,\pi) = \frac{n_{ij}}{\pi_{ij}} - \lambda = 0 \Rightarrow \hat{\pi}_{ij} = \frac{n_{ij}}{\lambda}$$

Then, by making use of the constraint $\sum_{i,j=1}^{2}\pi_{ij} = 1$, it follows that $1 = \sum_{i,j=1}^{2}\frac{n_{ij}}{\lambda} \Rightarrow \lambda = n$. Therefore, the MLE of π_{ij} is given by

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}$$

By the properties of the MLE, the MLEs of π_{i+} and π_{+j} are

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n} \text{ and } \hat{\pi}_{+j} = \frac{n_{+j}}{n}; i, j = 1, 2$$

Under the null hypothesis H_0 of equation (5), the estimate of the expected frequency of n_{ij} is then given by

$$e_{ij} = n\hat{\pi}_{ij} = n(\hat{\pi}_{i+})(\hat{\pi}_{+j}) = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+} \times n_{+j}}{n}$$

Therefore, the test of independence of disease outcome and exposure under the cross-sectional study is carried out by implementing the Pearson Chi-squared statistic

$$Q_P = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \sim \chi_1^2, \text{ where } e_{ij} = \frac{n_{i+}\times n_{+j}}{n}; i, j = 1, 2.$$

Therefore, the chi-squared test is invariant under different study designs (e.g., case-control, cohort and cross-sectional studies) and probability models (e.g., binomial and multinomial distributions).

3. Examples and Applications

In this section, we provide three examples, one from each of the underlying study designs and test the independence by employing Pearson's chi-squared test implemented by applying statistical software R.

Example 3.1: Is lung cancer related to smoking?

To investigate the link between lung cancer and smoking, 709 lung cancer patients (case) and 709 patients without lung cancer (control) admitted at the same hospital were queried about their smoking behavior. A smoker in the study was defined as a person who had smoked at least one cigarette a day for at least a year. For all patients in two samples, the hospital recorded the smoking behavior of lung cancer and non-lung cancer patients. The summary of patients with respect to lung cancer and smoking has been reported in Table 5. The 709 patients in the first column of Table 5 are those having the lung cancer (cases) and the 709 patients in the second column are those not having the lung cancer (controls). The information of Table 5 is available in Agresti (2012).

This study looks into the past of lung cancer patients (case) and those without lung cancer (control) for their smoking history and therefore, is a case-control study. The objective is to investigate if lung cancer and smoking is related.

Table 5. Distribution of patients by smoking and lung cancer under a case-control study

Smoking	Lung cancer		Total
	Yes (D=1)	No (D=2)	
Yes (E=1)	688	650	1338
No (E=2)	21	59	80
Total	709	709	1418

For employing a chi-squared test of independence between smoking and lung cancer patients, we computed expected frequency of each cell. We also computed value of the chi-squared statistic along with the p-value for the null hypotheses that smoking and the incidence of lung cancer are independent. The value of the chi-squared statistic is $Q_P = 18.136$ and p-value is 0.00002 with degrees of freedom equal to 1. The given result suggests that we reject the null hypothesis at 5% level of significance and conclude that smoking and lung cancer are related.

Example 3.2: Is vaccination and incidence of pneumonia related?

To investigate if being vaccinated makes any difference in relation to the development of pneumonia, two samples of subjects, each of size 92, one sample receiving vaccination (E=1) and the other not receiving vaccination (E=2) are observed for pneumonia outcome. The study recorded those who developed pneumonia (D=1) and those who did not develop pneumonia (D=2) in each of exposed (E=1, received vaccination) and unexposed (E=2, not received vaccination) samples. The summary of the results with exposure and disease status are summarized in Table 6. The information of Table 6 appears in McHugh (2013).

Table 6. Distribution of subjects by vaccination and pneumonia under a cohort study

Vaccination	Pneumonia		Total
	Yes (D=1)	No (D=2)	
Yes (E=1)	15	77	92
No (E=2)	31	61	92
Total	46	134	184

The outcome of the study of having pneumonia (D=1) and no pneumonia (D=2) develops after the follow-up of exposed (E=1) and unexposed (E=2) and hence the study is a cohort study. The objective of this study is to investigate any association between vaccination and pneumonia.

For the chi-squared test of independence between vaccination and pneumonia, we computed expected frequency of each cell, and the value of the chi-squared statistic along with the p-value for the null hypotheses that vaccination and the incidence of pneumonia are independent. The value of the chi-squared statistic is $Q_P = 6.522$ and p-value is 0.01066 with degrees of freedom equal to 1. The given result suggests that we reject the null hypothesis at 5% level of significance and conclude that vaccination and pneumonia are related.

Example 3.3: Does obesity affect activity?

A sample of $n = 550$ subjects in a study are cross-classified by obesity (factor exposed to) and disease status of being sedentary to assess if obesity and disease are related. The exposure status—being obese (E=1) and not obese (E=2) and disease status—being sedentary (D=1, low activity level) and not sedentary (D=2, moderate to high activity level)) are reported in Table 7.

Table 7. Distribution of patients by obesity and sedentarity under a cross-sectional study

Obese	Disease		Total
	Sedentary (D=1)	Not sedentary (D=2)	
Yes (E=1)	75	25	100
No (E=2)	250	200	450
Total	325	225	550

This study design belongs to a cross-sectional study because it is based on a single sample with disease and exposure status evaluated simultaneously. The information of Table 7 is available in Capili (2021). The objective of this study is to establish if there is any relation between obesity and disease. In order to do so, we perform a chi-squared test of independence with the null hypothesis that there is no association or dependence between obesity and sedentarity. The value of the chi-squared statistic is $Q_P = 12.005$ and p-value is 0.00053 with degrees of freedom equal to 1. The given result suggests that we reject the null hypothesis at 5% level of significance and conclude that obesity and sedentarity are significantly related.

4. Concluding Remarks

This study establishes that the chi-squared test is invariant with respect to different study designs resulting to different probability models. As such, the chi-squared test is a robust statistical tool for testing independence or non-association of two factors varying in designs or probabilistic models. The property of invariance makes it appealing to researchers interested in testing hypothesis of independence of observed two-categorical data in the form of a contingency table, whether it results from case-control, cohort or cross-sectional study designs. The probability distributions of observed cell frequencies and thereby, the computation of expected cell frequencies will be exemplary to users of chi-squared test as to justify why an identical form of the chi-squared test applies to case-control, cohort and cross-sectional study designs. The examples and applications provide directions to how different formulated studies are implemented via a chi-squared test.

Acknowledgement

The authors would like to thank the editorial team and referees for useful comments and suggestions.

Authors contributions

Khairul Islam and Tanweer J. Shapla were responsible for study design and methodology, literature review, and

preparation of entire manuscript.

Funding

Not applicable.

Competing interests

The authors declare no known competing financial interests or personal relationships that influenced the work reported in this paper.

References

Agresti, A. (2012). *Categorical Data Analysis* (3rd ed.). Wiley.

Capili, B. (2021). Overview: Cross-Sectional Studies. *American Journal of Nursing*, *121*(10), 59–62. https://doi:10.1097/01.NAJ.0000794280.73744.fe.

Casella, G., & Berger, R. (Author). (2024). *Statistical Inference* (2nd ed.). Chapman and Hall/CRC.

Kalman, D. (2009). Leveling with Lagrange: An alternate view of constrained optimization. *Mathematics Magazine*, *82*(3),*186-196.* https://doi:*10.1080/0025570X.2009.11953617*

Mann, C. J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medical Journal*, *20*(1), 54–60. https://doi: 10.1136/emj.20.1.54

Mann, C. J. (2012). Observational research methods—Cohort studies, cross sectional studies, and case–control studies. *African Journal of Emergency Medicine*, *2*(1), 38–46. https://doi:10.1016/j.afjem.2011.12.004

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, *23*(2), 143–9.

Omair, A. (2016). Selecting the appropriate study design: Case–control and cohort study designs. *Journal of Health Specialties*, *4*, 37-41.

Perez-Guerrero, E. E., Guillén-Medina, M. R., Márquez-Sandoval, F. M., & Vera-Cruz, J. M., et al. (2024). Methodological and Statistical Considerations for Cross-Sectional, Case–Control, and Cohort Studies. *Journal of Clinical Medicine*, *13*(14), 4005. https://doi:10.3390/jcm13144005

Robert, A. P., & G. Berman, N. (2015). *Planning Clinical Research*. 1st ed. New York: Cambridge. University Press.

Song, J. W., & Chung, K. C. (2010). Observational Studies: Cohort and Case-Control Studies. *Plastic & Reconstructive surgery, 126*(6), 2234–2242. https://doi:10.1097/PRS.0b013e3181f44abc

Taur, S. R. (2022). Observational designs for real-world evidence studies. *Perspectives in Clinical Research*, *13*(1), 12-6.