Construction of the χ^2 Goodness-of-Fit Test

By JunYang (Michael) Ma

Author Bio

Michael Ma is a senior from Pinehurst School in Auckland, New Zealand. While participating in the 2023 International Maths Olympiad and various other maths and physics competitions, Michael has developed his interest in pure mathematics and statistics and the development of mathematical models related to theoretical physics and quantitative biology. At school Michael acts as the Head Boy, as well as the founder of the Pinehurst Maths club and lead saxophone in the school Jazz Band.

Abstract

The goal of statistical tests is to use data to learn about the nature of the system from which it is collected. It is an important bridge between raw data and interpreting its meaning/ drawing conclusions from it. One recent use of statistical tests is in COVID-19 tests, where statistics is used to determine the false positive rate of such tests to assess their accuracy.

In this paper we will focus on one of the most used statistical tests, the Chi-square test, first published by Karl Pearson in 1900 to investigate the ‘deviations from the probable’.[6] We will first discuss the general procedure of hypothesis testing, followed by the construction of an appropriate estimator for statistical models before finally, finding the limiting distribution of the estimator. Using these properties, we will construct the Chi-square distribution and discuss its relevance to hypothesis testing. Python simulations using the Chi-square test will be generated to investigate the effect of multiple variables on its statistical power.

Keywords: Chi-square, Goodness-of-Fit test, Hypothesis testing, Multinomial distribution, Maximum likelihood estimation, Fisher information, Central limit theorem, Monte Carlo Simulation, Mendelian inheritance, Overparametrisation

Table of Notations

Symbol	Meaning
θ_*/θ	True value of parameter θ
$\hat{\theta}$	MLE of parameter θ
p and $\hat{p}$	True probabilities and estimated probabilities respectively
Ω	Sample space
Θ	Parameter space
X_i	Random variable with index i
Y_j	Number of observations of category j

Hypothesis Testing Framework

The goal of this paper is to rigorously conduct statistical tests. To do this, we need to have clear definitions for each part of the test, from collecting the data to extracting information from the data. The collected data points are realisations of random variables that represent the outcome of an experiment. In statistical terms, the possible outcomes of the random variable come from the sample space, which is the set of all possible outcomes. Since random variables are randomly drawn, we can only calculate the probability of a specific outcome. The data collected are then used to construct a statistical model: a collection of probability distributions, which try to describe the true distribution that the data originate from.

The procedure of statistical tests is structured around H_0 (the null hypothesis) and H_1 (the alternative hypothesis). H_0 proposes that the data collected have no statistical significance. However, if the statistical test shows that the data collected have enough evidence against H_0, then we can reject it, and that is when a significant discovery is made. Due to the test's random nature, there will almost always be a chance of errors. Let α denote the probability of type I error, which is when the test rejects H_0 given it is true, while β is the probability of type II error, which is when H_0 is retained (not rejected) given H_1 is true. If H_0 is rejected when a test has a small type I error, there is a higher chance that the information is part of a general trend and not a fluke. The power of a test is denoted by $1 – \beta$ -- good test would have not only a small type I error but also a high power. A low type I is preferred since scientists are comfortable making a discovery only when they are fairly certain it is true.

A test statistic is needed to reject H_0, which will be compared to a critical value depending on how much error is tolerated.

Since we are working with probabilities, the error tolerated is in sense that chance that we are wrong. Sometimes it is more valuable to know the strength of the evidence against H_0. Towards this end, a p-value could be calculated, and the smaller the p-value (smaller than 0.05 is generally considered strong), the stronger the evidence against H_0.

One type of statistical tests are the goodness-of-fit tests. They are useful in investigating whether the distributions in the statistical model matches the collected data. We will focus on the Chi-square test, with the following test statistic:

$$C_n = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}.$$

The intuition behind this test statistic is to find the size of the difference between the observed, Y_j, and expected, p_j, values in each category that the data is recorded in (divided by np_j to adjust for the size of the category). Hence when C_n is larger, the statistic implies that the observed data deviates from the expected distribution more, making it more likely that H_0 is rejected.

Origin of the MLE; Derivation from the KL Divergence

The purpose of this section is to motivate the derivation of the test-statistics. For our statistical model we will use parametric distributions. This is when the p.d.f (probability distribution function) can be defined by a finite number of parameters. E.g. The parameter for a weighted coin flip is the probability of it coming out heads. To find the parameters would be equivalent to knowing the entire distribution.

The parametric model we will focus on is the multinomial distribution, where there are k categories that the outcome could be, with each outcome having a fixed probability. Taking n samples from such distribution would yield observations $X_1=(X_{11},\ldots,X_{1k})$, …, $X_n=(X_{n1},\ldots,X_{nk})$, where each X_i is a unit vector resulting from a specific category. The multinomial distribution captures the probability distribution for the all the possible outcomes:

$$f(\vec{Y};\vec{p}) = \binom{n}{Y_1 \cdots Y_k} \prod_{j=1}^{k} p_j^{Y_j}$$

where Y_j, equal to the sum of X_{ij} over all i's, is the total number of observations for each of the k categories, and $p=(p_1,p_2,\ldots,p_k)$ is the defining parameter or the probability of getting each of the k different colours in one draw. Note that the distribution itself is governed by each Y_j. The multinomial distributions are a good fit for many models where the outcomes are a fixed number of discrete categories, with the simplest being the binomial distribution.

In order to construct a statistical test, we first need a good estimator of the parameters from our sample. First, we need to develop a notion that can assess the "closeness of two distributions.

Define the Kullback-Leibler divergence as [3]

$$\mathrm{KL}(f(x),g(x)) = \int_x f(x)\log\frac{f(x)}{g(x)}\,\mathrm{dx}$$

where f and g denote two distributions. For convenience we denote $KL(\theta_*, \hat{\theta})$ to mean $KL(f(x;\theta_*), f(x;\hat{\theta}))$. It is useful because of the crucial property that the KL divergence is 0/minimised if and only if the two input distributions are identical. Therefore, finding a good estimator, becomes the problem of minimising the KL divergence. From these properties we can derive the maximum likelihood estimator, the function, which is maximised when good estimates of the parameters are inputted [4,p.122]

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i;\theta)$$

Proof is in Section 8.1.

Here f represents the probability distribution of the model, and the X_i's represents a specific set of observations. As the name suggests, the goal is to find the parameters θ that maximises the likelihood function. By maximising the MLE we will have found estimates for the parameters which makes the given observations most probable. Specifically for the multinomial distribution the likelihood function is

$$\mathcal{L}(\vec{p}) = \prod_{j=1}^{k} p_j^{Y_j}$$

So how do we find the parameter which gives the maximum value of L(p)? we can use calculus and differentiate the function with respect to each parameter. This gives us the following estimations:

$$\hat{p}_j = \frac{Y_j}{n}$$

for each category j, hence giving us our MLE estimation, which is simply the sample average

Variance Matrix: Convergent Distribution of the Multinomial Distribution

Even though we have a good estimator for the parameter that converges to the true parameter, we still don't know the rate of convergence. To construct the Chi-square test it is crucial to find the variance, or the spread of the MLE around * to construct the Chi-square test.

Asymptotic Normality with Central Limit Theorem

Here's a key result from statistics that will help us.

Proposition 1 Central Limit Theorem (CLT) [4,p.77] Let $X_1,X_2,\ldots$ be IID (independent and identically distributed random variables) with finite mean and variance and let $\bar{X}$ be the sample average. Then

$$\frac{\bar{X}_n-\mu}{\sqrt{\mathbb{V}(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \rightsquigarrow N(0,1) \text{ as } n\to\infty$$

where $V(\bar{X})$ is the variance of the sample average and μ is the expectation. The squiggly arrow shows convergence in distribution, meaning that the limit of the *cumulative* distribution function of the L.H.S of the expression approaches that of the R.H.S if n is sufficiently large.

The theorem says that the sample average of a random variable will have the recognisable normal distribution with increasing sample size, as well as telling us the rate of convergence. For the multinomial distribution with several parameters, we need another version of the CLT.

Proposition 2 Multinomial CLT [4,p.53]
Let Y and μ be vectors such that

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}, \vec{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} = \begin{bmatrix} \mathbb{E}(Y_1) \\ \vdots \\ \mathbb{E}(Y_k) \end{bmatrix}$$

and the variance-covariance matrix is

$$\mathbb{V}(Y) = \begin{bmatrix} \mathbb{V}(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_k) \\ Cov(Y_2, Y_1) & \mathbb{V}(Y_2) & \cdots & Cov(Y_2, Y_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_k, Y_1) & Cov(Y_k, Y_2) & \cdots & \mathbb{V}(Y_k) \end{bmatrix}$$

The vectors p̂ and p are the vectors Y and μ divided by n respectively, and the asymptotic normality of vector p̂ is

$$\sqrt{n}(\hat{\vec{p}} - \vec{p}) \rightsquigarrow N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{bmatrix}$$

Thus, we have Σ, which tells us the spread of the asymptotically normal distribution. Note that the CLT applies here because the MLE estimation happens to be the sample averages for the multinomial distribution. Otherwise, this would not work, and the variance would need to be computed through the Fisher information matrix. Overall, the limiting distribution of the MLE vector is

$$\sqrt{n}(\hat{\vec{p}} - \vec{p}) \rightsquigarrow N(0, \Sigma)$$

We will need to use the Σ^{-1} in section 5, which is calculated in section 8.2.

$$W - \mu \sim N(0, \sigma^2)$$
$$\frac{W - \mu}{\sigma} \sim N(0, 1) = Z$$

We will use this transformation in section 4 as part of our proof.

Properties of Chi-square test

Convergence of the Chi-Square Statistic

With the consistency and asymptotic normality of the MLE, we will be able to find the distribution of the Chi-square statistic. First, we need to define the Chi-square distribution.

If $Z=(Z_1,\ldots,Z_k)^T$ is a vector of k independent standard normal distributions, then

$$\chi^2_k = \sum_{i=1}^{k} Z_i^2 = \vec{Z}^T\vec{Z}$$

There is only one parameter, k, for the Chi-square distribution, where k is a positive integer.
So far, we concluded that the MLE estimators are $\hat{p}=(\hat{p}_1,\ldots, \hat{p}_{k-1})^T=(Y_1/n,\ldots,Y_{k-1}/n)^T$, and it has the variance matrix . To compute the distribution of the Chi-square statistic, we will rewrite the following expression:

$$\sqrt{n}(\hat{\vec{p}} - \vec{p})^T\Sigma^{-1}\sqrt{n}(\hat{\vec{p}} - \vec{p})$$

in two different ways. The motivation is to change the Chi-square distribution into a form which includes the MLE, and hence can be expanded algebraically.

Firstly, by expanding the expression (see 8.3) using matrix operations,

$$\sqrt{n}(\hat{\vec{p}} - \vec{p})^T\Sigma^{-1}\sqrt{n}(\hat{\vec{p}} - \vec{p}) = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} = C_n$$

gives us the Chi-square statistic. Proof is in 8.4. We recognise Σ^{-1} as I(p).

Secondly, we observe that the expression can be split up into two objects, which is given by (13) and its transpose. Since √n(p̂-p) converges to N(0, Σ) by (7), it can be transformed to the standard normal by (8) and (9).

Σ is a positive semi-definite matrix, so it can have 'square roots.' Hence, we have

$$\begin{aligned}&\sqrt{n}(\hat{p}-p)^T\Sigma^{-1}\sqrt{n}(\hat{p}-p)\\&=(\Sigma^{-1/2}\sqrt{n}(\hat{p}-p))^T(\Sigma^{-1/2}\sqrt{n}(\hat{p}-p))\\&\rightsquigarrow \vec{Z}^T\vec{Z}\\&=\chi^2_{k-1}\end{aligned}$$

since p has k-1 rows. The second line is because the positive semidefinite property of Σ implies it is symmetric, so $\Sigma^T = \Sigma$. Hence, we can move one of the $\Sigma^{-1/2}$ into the transpose with the identity $(AB)^T=B^TA^T$.

$$\therefore C_n=\sum_{j=1}^{k}\frac{(Y_j-np_j)^2}{np_j}\sim\chi^2_{k-1}$$

As there is one redundant parameter, the convergence to is expected. Q.E.D

Using the Chi-Square Test

To conduct the Chi-Square test, first we set up our null and alternative hypotheses:

H_0: the multinomial distribution has the parameter p.

H_1: the multinomial distribution does not have the parameter p. and set the significance level α of the test, which is the allowance on type I error.

After finding Y_j for each category, we calculate the Chi-square statistic C_n by (1). We also need to find the critical value of the Chi-square test based on the significance level and the degrees of freedom, k-1.

Simulations

With the Chi-square test's convergence proven theoretically, we shall explore some conditions under which the Chi-square test is applicable; specifically, we will investigate how several variables can affect the power of the test. In the first simulation we will look at a practical scenario involving the colour of camellia flowers and how we can test a basic result in gene theory with the Chi-square test. In the second simulation we will see the idea of overparametrisation (when there is an excess of parameters) and how it can affect the Chi-square test.

Simulation 1: Colours of Camellias

In a simple model of inheritance, genes are lengths of DNA that determine one trait of an organism. Different versions of one gene are called alleles, and they are always in the same location on the same chromosome.

In most animals and plants, the cells are diploid, meaning that each type of gene contains two alleles, which might be different or the same. The nice property about alleles, which makes them fit for the Chi-square test, is that they only give finitely many expressions. The genotypes (specific alleles for an organism's genes) give a discrete set of phenotypes (traits given by the phenotype).

During reproduction, meiosis scrambles the alleles in the resulting cells to create genetic diversity. Hence the occurrence of a phenotype in an offspring is a random variable, and the multinomial distribution is a good model for the phenotypic ratio of the offspring plants that can be tested by the Chi-square test.

The specific example we will look at is the colour of the petals of camellia flowers. They are usually red or white, but when red and white camellias are bred together, they produce a third type with red and white patterns. In total there are three discrete outcomes.[2] According to gene theory, if a field of only the red-and-white variant of camellias are bred, then the offspring will have a colour ratio of red:red-and-white:white = 1:2:1. The Chi-square test can be used to investigate whether this theory holds, and how effectively the test rejects it.

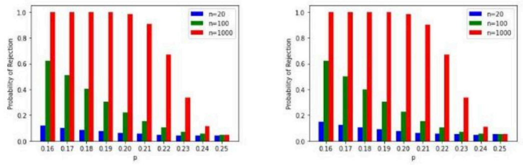
Method

We use Monte Carlo simulations, which entails generating random data that will create converging numerical values. With the multinomial distribution in the code, we will generate random data to arrive at the power of each test. Let p_1,p_2 and p_3 represent the probabilities for red, red-and-white, and white respectively. We will set up the following hypotheses, and carry the test at significance level of 5%: H_0 -- the multinomial distribution for the colours have parameters $(p_1,p_2,p_3) = (1/4,1/2,1/4)$ / H1: the multinomial distribution is not described by H_0.

To carry out the simulation, we will use python to generate multinomial distributions with $(p_1,p_2,p_3) = (p,1/2,1/2\text{-}p)$, where p is from (0.16,0.17,…,0.25). We will also test sample sizes of 20, 100, and 1000. Each experiment will be repeated 10,000 times and the probability of rejection is calculated by the number of times the test rejected H0 divided by 10000. For simulation 1, we will also directly compute the theoretical probability of rejection for comparison.

Interpreting results

Both theoretical and experimental powers for varying p are displayed side by side:



(a) Theoretical probability of rejection (b) Simulated probability of rejection

Figure 4: Independent variables p and n are represented by the x-axis and colour of bars respectively. Probability of rejection is the dependent variable and is represented by the y-axis.

The two plots are almost identical upon initial inspection. For small values of n, such as 20, there is a lower probability of rejection in the simulations. This could just be due to the simulation's random nature, as smaller sample sizes have smaller information to converge to theoretical values. The overwhelming similarity suggests that running 10000 Monte Carlo trials is sufficient to generate an accurate representation of power in this simulation.

When p=0.25 (when H_0 is true), all three sample sizes produce a probability of rejection of around 0.05. This is expected since the test is designed to have a significance level of 0.05, so the probability of type I error will be 0.05. Further comparing the p=0.25 data, the n=20 bar is slightly lower than the n=100 and n=1000 bars, which are closer to 0.05. The difference demonstrates the LLN; as n increases, the convergence to the expected value also strengthens.

From right to left, the models deviate more from H_0. Note that all values of p except for 0.25 are different from H_0 and so should be rejected. It can be seen that from all three sample sizes, the power increases from right to left, suggesting that the Chi-square test is able to reject more of the tests, when p is smaller. We can conclude that the power, the probability of rejecting the H_0 when it is false, of the Chi-square test also increases as the discrepancy between H_0 and the generating distribution increases.

Comparing the trend between all bars of different sample sizes, there is a clear increase in power from n of 20 to 100 to 1000, across all p (except for p=0.24 and 0.25). Whereas the n=20 bars never exceed 0.2, the n=1000 bars already reached 100% rejection by p=0.19. In addition, as discrepancy with H0 increases, the rate at which the power increases is also faster for higher sample sizes. This means that the power of the Chi-square test increases significantly when sample size is increased. An explanation for this is that the variance of the MLE (which is the sample average) is inversely proportional to n. Therefore, with higher sample sizes the variance decreases, so it is easier to tell when the data does not come from H_0.

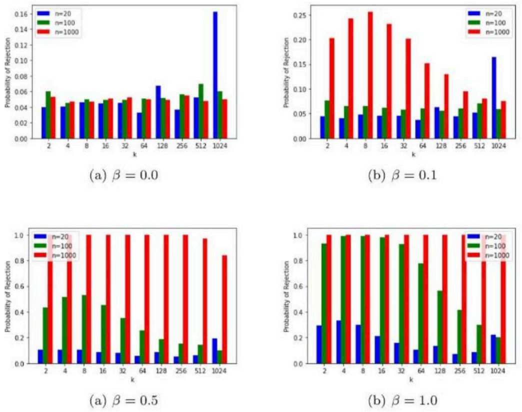
Hence from varying the two parameters n and p, the bar graph shows that the power of the Chi-square test has a positive correlation with both the sample size and the discrepancy between H_0 and the true distribution to H_0.

Overparametrisation

Overparametrisation is an active topic of discussion. As we try to understand the world with models, like in a complex ecosystem, we can often end up with too many parameters and not enough observations. In this simulation we repeat a similar procedure to the previous one, but instead of varying the value of p, we will vary the number of categories in the model then introduce a variable β which will deviate the true distribution based on the category number. For the true distribution we will use $p=(1^{\beta}c,2^{\beta}c,\ldots,k^{\beta}c)$, where c is a constant that will adjust all the probabilities to sum to 1.

The hypotheses are as follows -- H0: the multinomial distribution have k parameters and each parameter is the same / H_1: the multinomial distribution is not described by H_0

The bigger the value of β, the more it deviates from H_0. When it is 0, the true distribution will be the same as H_0. Like experiment 1, when the true distribution agrees with H_0, the power of the test fluctuates around 0.05. Large values of n tend to show less fluctuation, but not by a great extent. A bigger number of iterations may yield more converging results.



(a) $\beta = 0.0$

(b) $\beta = 0.1$

(a) $\beta = 0.5$

(b) $\beta = 1.0$

Figure 5: Probability of rejection against number of categories k for four values of β.

The bigger the parameter β, the less uniform the true distribution, hence increasing discrepancy to H_0. The graphs for β=0.1, 0.5, 1 display an eventually decreasing power with increasing k. This is caused by an increase in the parameter to sample ratio as k increases. The more parameters there are compared to sample size, the less accurately the observations will be able to capture the true distribution.

However, there is not a directly negative correlation between k and power. Notably, for bigger values of β, the power is unimodal and peaking at around k=4 or k=8. There are likely two competing effects going on as k is increased; increasing k likely decreases the power but also increases the distance from H_0.

There are also significant anomalies in the trends for k=1024, especially for lower values of n and β. There could be two possible sources of this anomaly. Either there is an error with the random seed generation in drawing the random variables, or the number of categories is too large to generate a converging power. The first explanation is unlikely, since repeats of this experiment in a different Python code yielded similar but not identical results. The latter explanation is more likely as the CLT can converge slower for more parameters.

Conclusion

In the paper we have shown the key parts which are needed to construct the Chi-square test statistic. From the multinomial distribution to estimating the mean and variance through the MLE and CLT. With these two tools it is possible to link the Chi-square test statistic with the Chi-square distribution through its definition.

Although the multinomial distribution may seem limiting, it is versatile as it can model many distributions with discrete categories. This gives the Chi-square test a range of applications in real life scenarios.

Through the computer simulations, we have shown that the Chi-square test is very sensitive to reject the null hypothesis when the tested distribution differs slightly from the null. The sensitivity is further magnified with an increasing number of trials, which confirms the central limit theorem in practice.

Appendix

From KL to MLE

Here we derive an expression for the MLE function from KL divergence. Using the definition of the KL divergence, we can express it as the difference of two expectations.

$$\begin{aligned}\mathrm{KL}(\theta_*,\theta) &= \int_x \log\frac{f_{\theta_*}(x)}{f_\theta(x)} f_{\theta_*}(x)dx \\ &= -\left(\int_x \log(f_\theta(x)) f_{\theta_*}(x)dx - \int_x \log(f_{\theta_*}(x)) f_{\theta_*}(x)dx\right) \\ &= -\mathbb{E}_{f_{\theta_*}}[\log f_\theta(x)] + \mathbb{E}_{f_{\theta_*}}[\log f_{\theta_*}(x)]\end{aligned}$$

which is fit for applying the LLN, allowing us to replace the expectations with averages

$$\hat{\mathrm{KL}}(\theta_*,\theta) = -\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) + c(\theta_*)$$

$c(\theta_*)$ simply denotes a constant independent of θ. Using the identification property of KL, we know that when it is at the absolute minimum, then the estimator will converge to the truth. Hence, we want to minimise KL.

$$\begin{aligned}\hat{\theta}^{\mathsf{MLE}} &= \underset{\theta\in\Theta}{\arg\min}\, \hat{\mathrm{KL}}(\theta_*,\theta) \\ &= \underset{\theta\in\Theta}{\arg\min} -\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) + c(\theta_*) \\ &= \underset{\theta\in\Theta}{\arg\min} -\frac{1}{n}\sum_{i=1}^{n} f_\theta(X_i) \\ &= \underset{\theta\in\Theta}{\arg\max} \sum_{i=1}^{n} f_\theta(X_i)\end{aligned}$$

Thus, we obtain the likelihood function.

Inverting (through row reduction)

$$\left[\begin{array}{cccc|cccc} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} & 1 & 0 & \cdots & 0 \\ -p_2p_1 & p_2(1-p_2) & \cdots & \vdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & -p_{k-2}p_{k-1} & \vdots & \vdots & \ddots & \vdots \\ -p_{k-1}p_1 & \cdots & -p_{k-1}p_{k-2} & p_{k-1}(1-p_{k-1}) & 0 & 0 & \cdots & 1 \end{array}\right]$$

$r_i \leftarrow r_i - \frac{p_i}{p_1} r_1 \text{ for } i \in \{2,\ldots,k-1\}$

$$\left[\begin{array}{cccc|cccc} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_{k-1} & 1 & 0 & \cdots & 0 \\ -p_2 & p_2 & \cdots & \vdots & -\frac{p_2}{p_1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ -p_{k-1} & \cdots & 0 & p_{k-1} & -\frac{p_{k-1}}{p_1} & 0 & \cdots & 1 \end{array}\right]$$

$r_1 \leftarrow \frac{r_1}{p_1}$

$$\left[\begin{array}{cccc|cccc} 1-p_1 & -p_2 & \cdots & -p_{k-1} & \frac{1}{p_1} & 0 & \cdots & 0 \\ -p_2 & p_2 & \cdots & \vdots & -\frac{p_2}{p_1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ -p_{k-1} & \cdots & 0 & p_{k-1} & -\frac{p_{k-1}}{p_1} & 0 & \cdots & 1 \end{array}\right]$$

$r_1 \leftarrow r_1 + \sum_{i=2}^{k-1} r_i \left(\sum_{i=2}^{k-1} r_i = \left[\begin{array}{cccc|cccc} p_1+p_k-1 & p_2 & \cdots & p_{k-1} & 1+\frac{p_k}{p_1}-\frac{1}{p_1} & 1 & \cdots & 1 \end{array}\right]\right)$

$$\left[\begin{array}{cccc|cccc} p_k & 0 & \cdots & 0 & 1+\frac{p_k}{p_1} & 1 & \cdots & 1 \\ -p_2 & p_2 & \cdots & \vdots & -\frac{p_2}{p_1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ -p_{k-1} & \cdots & 0 & p_{k-1} & -\frac{p_{k-1}}{p_1} & 0 & \cdots & 1 \end{array}\right]$$

$r_1 \leftarrow \frac{r_1}{p_k}, r_i \leftarrow \frac{r_i}{p_i} \text{ for } i \in \{2,\ldots,k-1\}$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & \frac{1}{p_1}+\frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ -1 & 1 & \cdots & \vdots & -\frac{1}{p_1} & \frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ -1 & \cdots & 0 & 1 & -\frac{1}{p_1} & 0 & \cdots & \frac{1}{p_{k-1}} \end{array}\right]$$

$r_i \leftarrow r_i + r_1 \text{ for } i \in \{2,\ldots,k-1\}$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & \cdots & 0 & \frac{1}{p_1}+\frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ 0 & 1 & \cdots & \vdots & \frac{1}{p_k} & \frac{1}{p_2}+\frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & 0 & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}}+\frac{1}{p_k} \end{array}\right]$$

giving us Σ^{-1} on the right.

Chi-Square Statistic from Matrix Expansions

when it is at the absolute minimum, then the estimator will converge to the truth. Hence, we want to m

$$
\begin{aligned}
&\sqrt{n}(\hat{p}-p)^T\Sigma^{-1}\sqrt{n}(\hat{p}-p)\\
&= n\begin{bmatrix} \frac{Y_1}{n}-p_1 \\ \cdots \\ \frac{Y_{k-1}}{n}-p_{k-1}\end{bmatrix}^T \begin{bmatrix} \frac{1}{p_1}+\frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2}+\frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}}+\frac{1}{p_k}\end{bmatrix} \begin{bmatrix} \frac{Y_1}{n}-p_1 \\ \vdots \\ \frac{Y_{k-1}}{n}-p_{k-1}\end{bmatrix}\\
&= \frac{1}{n}\begin{bmatrix} Y_1-np_1 \\ \vdots \\ Y_{k-1}-np_{k-1}\end{bmatrix}^T \begin{bmatrix} \frac{1}{p_1}+\frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \frac{1}{p_k} & \frac{1}{p_2}+\frac{1}{p_k} & \cdots & \frac{1}{p_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_k} & \frac{1}{p_k} & \cdots & \frac{1}{p_{k-1}}+\frac{1}{p_k}\end{bmatrix} \begin{bmatrix} Y_1-np_1 \\ \vdots \\ Y_{k-1}-np_{k-1}\end{bmatrix}\\
&= \frac{1}{n}\begin{bmatrix} Y_1-np_1 \\ \vdots \\ Y_{k-1}-np_{k-1}\end{bmatrix}^T \begin{bmatrix} \frac{Y_1-np_1}{p_1}+\frac{1}{p_k}(\sum_{j=1}^{k-1}Y_j-n\sum_{j=1}^{k-1}p_j) \\ \vdots \\ \frac{Y_{k-1}-np_{k-1}}{p_{k-1}}+\frac{1}{p_k}(\sum_{j=1}^{k-1}Y_j-n\sum_{j=1}^{k-1}p_j)\end{bmatrix}\\
&= \frac{1}{n}\begin{bmatrix} Y_1-np_1 \\ \vdots \\ Y_{k-1}-np_{k-1}\end{bmatrix}^T \begin{bmatrix} \frac{Y_1-np_1}{p_1}+\frac{1}{p_k}(n-Y_k-n(1-p_k)) \\ \vdots \\ \frac{Y_{k-1}-np_{k-1}}{p_{k-1}}+\frac{1}{p_k}(n-Y_k-n(1-p_k))\end{bmatrix}\\
&= \frac{1}{n}\begin{bmatrix} Y_1-np_1 \\ \vdots \\ Y_{k-1}-np_{k-1}\end{bmatrix}^T \begin{bmatrix} \frac{Y_1-np_1}{p_1}+\frac{np_k-Y_k}{p_k} \\ \vdots \\ \frac{Y_{k-1}-np_{k-1}}{p_{k-1}}+\frac{np_k-Y_k}{p_k}\end{bmatrix}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{j=1}^{k-1}(Y_j-np_j)\left(\frac{Y_j-np_j}{p_j}+\frac{np_k-Y_k}{p_k}\right)\\
&= \sum_{j=1}^{k-1}\frac{(Y_j-np_j)^2}{np_j}+\left(\frac{np_k-Y_k}{np_k}\right)\sum_{j=1}^{k-1}(Y_j-np_j)\\
&= \sum_{j=1}^{k-1}\frac{(Y_j-np_j)^2}{np_j}+\frac{np_k-Y_k}{np_k}(n-Y_k-n(1-p_k))\\
&= \sum_{j=1}^{k-1}\frac{(Y_j-np_j)^2}{np_j}+\frac{(np_k-Y_k)^2}{np_k}\\
&= \sum_{j=1}^{k}\frac{(Y_j-np_j)^2}{np_j}=C_n
\end{aligned}
$$

References

[1] Victor Chernozhukov, course materials for 14.385 Nonlinear Econometric Analysis, Fall 2007. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Downloaded on [20 April 2022].

[2] (n.d.). Codominance Explained with Examples. BiologyWise. Retrieved May 12, 2022, from https://biologywise.com/codominance-explained-with-examples

[3] MIT OpenCourseWare. (2017, August 18). 4. Parametric Inference (cont.) and Maximum Likelihood Estimation [Video]. YouTube. url:https://youtu.be/rLlZpnT02ZU

[4] Wasserman, L. (2004). All of statistics: a concise course in statistical inference (Vol. 26, p. 86). New York: Springer.

[5] Plackett, R. L. (1971). The Application of the Chi-Squared Test. The Mathematical Gazette, 55(394), 363–366. https://doi.org/10.2307/3612358

[6] Plackett, R. L. (1983). Karl Pearson and the Chi-Squared Test. International Statistical Review / Revue Internationale de Statistique, 51(1), 59–72. https://doi.org/10.2307/1402731