The χ^2 Distribution and Goodness-of-Fit Tests

Yi J Zhu

June 30, 2021

1 The χ^2 Distribution

Suppose $X_1, X_2, \ldots, X_k$ are k i.i.d.[1] standard normal distributions $X_i \sim N(0,1)$. We define the chi-squared distribution with k degrees of freedom as the sum of the squared X_i distributions,
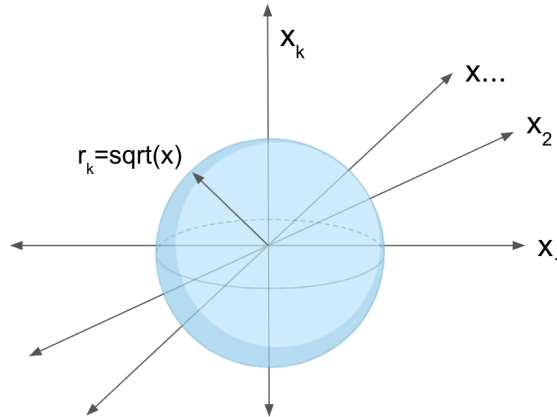
$$\chi_k^2 = \sum_{i=1}^{k} X_i^2 \tag{1.1}$$

1.1 Probability Density Function

We can determine the chi-squared PDF $f_{\chi_k^2}(x)$ by first writing out its CDF $F_{\chi_k^2}(x)$. If we define,

$$r_k = \sqrt{\sum_k x_i^2} = \sqrt{x} \tag{1.2}$$

Then CDF is the volume integral of the probability over an n-sphere of radius $r_k = \sqrt{x}$,

---

[1]Independent, identically distributed.

$$F(x) = \int_0^{\sqrt{x}} \left(\prod_k \frac{e^{-x_i^2/2}}{\sqrt{2\pi}} \right) A_k(r_k) dr_k \tag{1.3}$$

$$= \int_0^{\sqrt{x}} \frac{e^{-r_k^2/2}}{(2\pi)^{k/2}} A_k(r_k) dr_k \tag{1.4}$$

$$= \int_0^x \frac{e^{-x'/2}}{(2\pi)^{k/2}} A_k\left(\sqrt{x'}\right) \left(\frac{dx'}{2\sqrt{x'}} \right) \tag{1.5}$$

Where $A_k(r)$ is the surface area of the k-dimensional sphere,

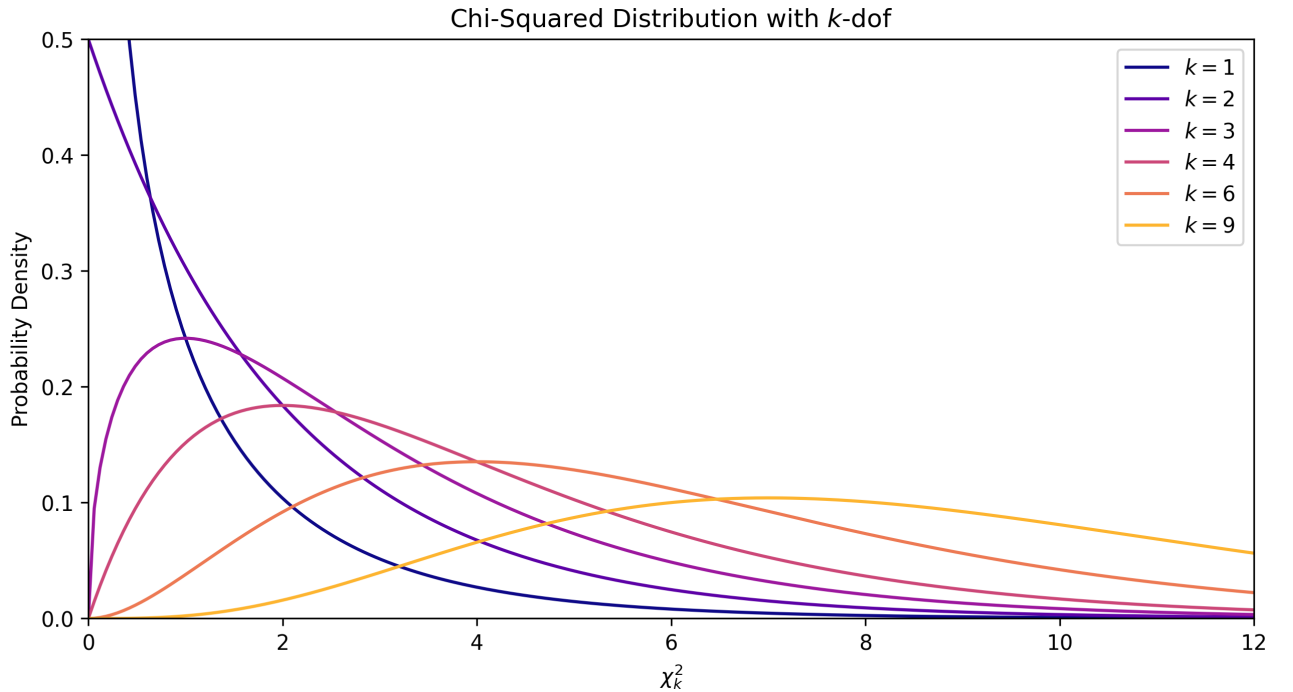$$A_k = \frac{2r^{k-1}\pi^{k/2}}{\Gamma(k/2)} \tag{1.6}$$

Thus,

$$F(x) = \int_0^x \frac{e^{-x'/2}}{(2\pi)^{k/2}} \frac{2x'^{(k-1)/2}\pi^{k/2}}{\Gamma(k/2)} \frac{dx'}{2\sqrt{x'}} \tag{1.7}$$

$$F(x) = \int_0^x \frac{e^{-x'/2}}{2^{k/2}} \frac{x'^{(k/2-1)}}{\Gamma(k/2)} dx' \tag{1.8}$$

Now, we obtain the PDF:

$$f(x) = \frac{dF}{dx} \tag{1.9}$$

$$f_{\chi_k^2}(x) = \frac{e^{-x/2} \; x^{(k/2-1)}}{2^{k/2} \; \Gamma(k/2)} \tag{1.10}$$

Chi-Squared Distribution with k-dof

1.2 MGF, Mean, and Variance

We can calculate the moment generating function of χ_k^2 via brute force,

$$E(e^{t\chi^2}) = \frac{1}{2^{k/2}\ \Gamma(k/2)} \int_0^\infty \left(e^{-x/2}\ x^{(k/2-1)}dx\right) e^{tx} \tag{1.11}$$

$$= \frac{1}{2^{k/2}\ \Gamma(k/2)} \int_0^\infty x^{(k/2-1)} e^{-x(1/2-t)}dx \tag{1.12}$$

for, $u = x(1/2 - t)$,

$$E(e^{t\chi^2}) = \frac{1}{2^{k/2}\ \Gamma(k/2)} \int_0^\infty \left(\frac{u}{1/2-t}\right)^{k/2-1} e^{-u} \frac{du}{1/2-t} \tag{1.13}$$

$$= \frac{1}{2^{k/2}\ \Gamma(k/2)} \left(1/2-t\right)^{-k/2} \int_0^\infty u^{k/2-1} e^{-u}du \tag{1.14}$$

$$= \frac{1}{2^{k/2}\ \Gamma(k/2)} \left(1/2-t\right)^{-k/2} \Gamma(k/2) \tag{1.15}$$

Thus,

$$M_{\chi_k^2}(t) = E(e^{t\chi_k^2}) = (1-2t)^{-k/2} \tag{1.16}$$

Now let's calculate the mean,

$$E(\chi_k^2) = \left.\frac{\mathrm{d}M_{\chi_k^2}}{\mathrm{d}t}\right|_{t=0} = \left.\frac{k}{(1-2t)^{k/2+1}}\right|_{t=0} \tag{1.17}$$

$$E(\chi_k^2) = k \tag{1.18}$$

The fact that expectation value is proportional to the degrees of freedom should make sense. We know that $E[\sum_k X_i^2] = \sum_k E(X_i^2) = kE(X_1^2)$. It happens that $E(X_1^2) = 1$, so the mean is exactly equal to the number of degrees of freedom.

In other words, the expectation value per degree of freedom is 1. We'll make use of this fact in the next section!

Now, let's calculate the variance. First,

$$E\left((\chi_k^2)^2\right) = \left.\frac{\mathrm{d}^2 M_{\chi_k^2}}{\mathrm{d}t^2}\right|_{t=0} = \left.\frac{k(k+2)}{(1-2t)^{k/2+2}}\right|_{t=0} = k^2 + 2k \tag{1.19}$$

Now,

$$Var(\chi_k^2) = E\left[(\chi_k^2)^2\right] - \left[E(\chi_k^2)\right]^2 = k^2 + 2k - k^2 \tag{1.20}$$

$$Var(\chi_k^2) = 2k \tag{1.21}$$

2 Goodness-of-Fit Testing

There are two types of statistical tests that are commonly performed with the Chi-squared distribution: (1) the reduced Chi-squared test and (2) Pearson's Chi-squared test. The reduced Chi-squared test is used to determine how well a model fits observations assuming Gaussian noise. Pearson's Chi-squared test is a p-value test for a multinomial distribution.

2.1 Reduced Chi-Squared Test

Suppose we collect a set of data $\{x_i, y_i\}$ relating how some variable y relates to variable x. Furthermore, let's assume that the sources of error for our measurements are Gaussian, and that the uncertainties for each data point $\{\sigma_i\}$ is known.

We can fit out data to a model $\hat{y}(\vec{a}; x)$ where $\vec{a} = (a_1, a_2, \ldots, a_m)$ are the fit parameters and $f_i(x)$ are arbitrary functions of x that do not involve $\vec{a}$,

$$\hat{y}(\vec{a}; x) = \sum_{j=1}^{m} a_j f_j(x) \tag{2.1}$$

If we define the variable χ^2 (which we'll justify in a bit),

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}(\vec{a}; x_i)}{\sigma_i} \right)^2 \tag{2.2}$$

Then we can determine the optimal fit parameters $\vec{a}$ for our model $\hat{y}(\vec{a}, x)$ by minimizing χ^2. This is known as the method of least squares![2]

But how do know how well our fit *fits* our data? We might expect that the smaller the value of χ^2, the smaller the residuals, and the better our fit. This reasoning is correct, but misleading. Instead, we often used the reduced Chi-squares statistic χ_r^2 as a heuristic for how well our model fits the data.
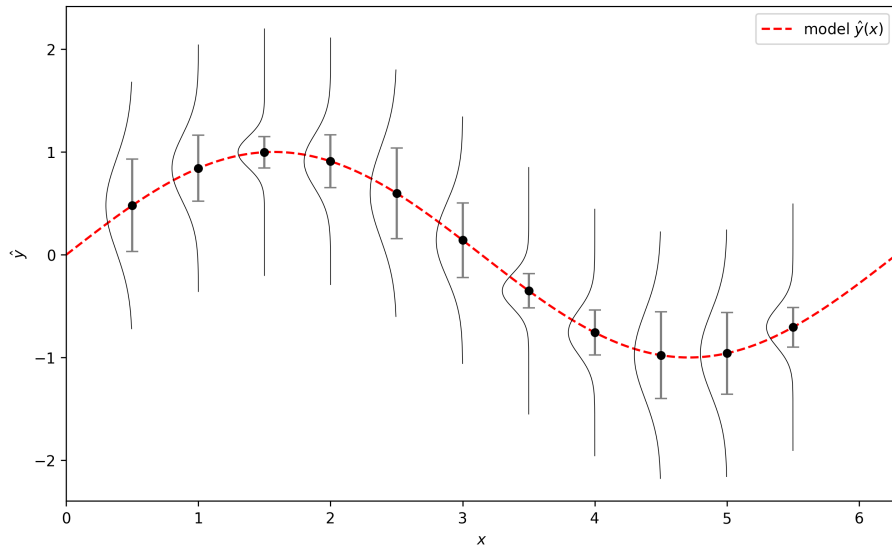
Suppose our model is correct. Then, for every point x_i, we would expect the distribution of our observed y_i to be Gaussian with mean $\hat{y}(\vec{a}, x_i)$ and variance σ_i^2. In other words,

$$y_i \sim N\left(\hat{y}(\vec{a}, x_i), \sigma_i^2\right) \tag{2.3}$$

---

[2]This is how Scipy's curve fit function operates.

If we "normalize" the random variable y_i,

$$\frac{y_i - \hat{y}(\vec{a}, x_i)}{\sigma_i} = N(0, 1) = X_i \tag{2.4}$$

Defining X_i to be i.i.d standard normal like in the previous section,

$$\sum_n \left(\frac{y_i - \hat{y}(\vec{a}, x_i)}{\sigma_i} \right)^2 = \sum_i X_i^2 = \chi_k^2 \tag{2.5}$$

Thus we were justified in calling 2.2 χ^2. But what is the number of degrees of freedom k?

Each data point represents a distinct degree of freedom, so with n data points, we have n degrees of freedom. However, our model with m fit parameters uniquely determines m degrees of freedom.

In other words, given m fit parameters: if we know the first $n - m$ values of y_i, then without loss of generality, we must know the last m values of y_i because these are the values such that χ^2 is minimized with respect to our model. Therefore, the Chi-squared distribution has $k = n - m$ degrees of freedom.[3]

Now, we define the reduced Chi-squared value to be,

$$\chi_r^2 = \frac{\chi^2}{\nu}; \quad \nu = \text{degrees of freedom} \tag{2.6}$$

Recall that we found (1.18) the expectation value of χ_r^2 to be 1. This means that if our model correctly describes the observed data, we would not expect the value of χ_r^2 to be as small as possible. Instead we expect $\chi_r^2 = 1$[4].

If $\chi_r^2 > 1$, it is likely that the model is incorrect or that the fit parameters do not accurately describe the data.

---
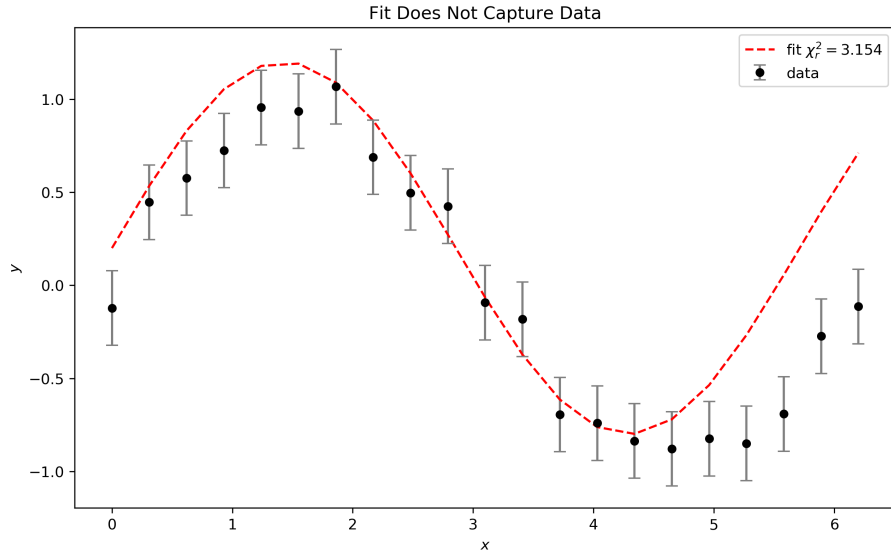
[3]The assumption that our model is a linear combination of functions (2.1) is critical! If we use an arbitrary model, we can not generally say that there are $n - m$ dof.

[4]Notice that the variance per degree of freedom is also constant (1.21).

If $\chi_r^2 < 1$, it is possible that the data is over fitted: there are too many degrees of freedom in the model and we are fitting noise.
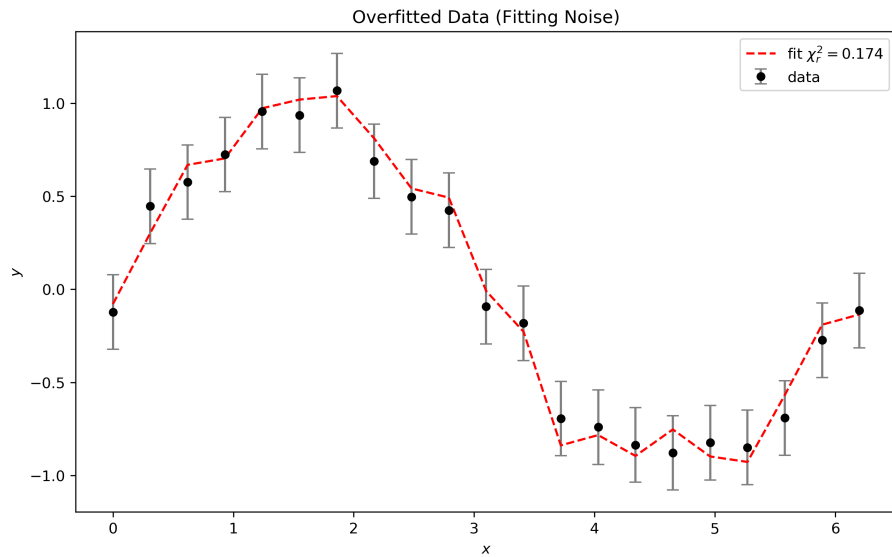
Generally, over fitting is not so dramatic. More subtly, the fit parameters will deviate from their true values in order to account for noise.

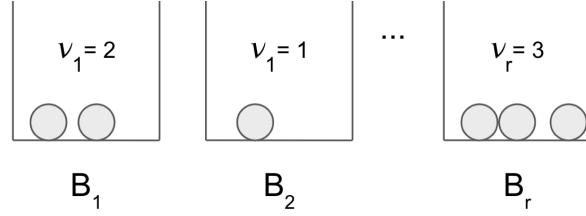Another cause for $\chi_r^2 < 1$ is if the data uncertainties are overestimated.
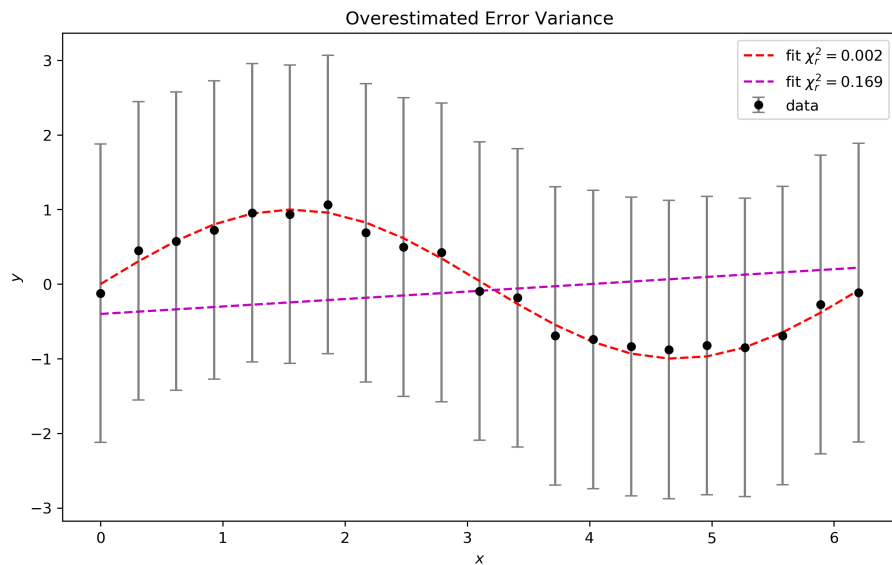
Here, the measurement uncertainties are so large that the fit is not longer meaningful.

A rule of thumb for the reduced Chi-squared test is that for a good fit,

$$0.8 < \chi_r^2 < 1.5 \tag{2.7}$$

2.2 Pearson's Chi-Squared Test

Suppose we throw n balls at r boxes B_i. The probability of any single ball landing in box B_i is p_i and independent of any other ball. After all of the balls have been thrown, we observe each box contains ν_i balls so that,

$$\sum_{i=1}^{r} \nu_i = n \tag{2.8}$$

Given that we observe ν_i, we want to determine if p_i is a good model. The probability of observing ν_i is given by the multinomial distribution. While we can make use of this fact, another approach is to use Pearson's theorem.

Theorem 2.1 (Pearson).

$$\sum_{i=1}^{r} \frac{(\nu_i - np_j)^2}{np_j} \xrightarrow{n \text{ large}} \chi^2_{r-1} \tag{2.9}$$

The random variable on the LHS is called the Pearson's Chi-squared variable and converges in distribution (large n) to the Chi-squared distribution with one less degree of freedom.

A rigorous treatment of Pearson's theorem is a bit involved, but here I shall present some intuition (this is not a proof!).

We can define the indicator variable,

$$I(X_i \in B_j) = \begin{cases} 1, & \text{the i-th ball lands in the j-th box} \\ 0, & \text{else} \end{cases} \tag{2.10}$$

Then, we will find that $\nu_j = \sum_{i=1}^{n} I(X_i \in B_j)$[5] and[6],

$$E(\nu_j) = E\left(\sum_{i=1}^{n} I(X_i \in B_j)\right) = \sum_{n} E\left[I(X_i \in B_j)\right] = np_j \tag{2.11}$$

$$Var(\nu_j) = Var\left(\sum_{i=1}^{n} I(X_i \in B_j)\right) = np_j(1 - p_j) \tag{2.12}$$

Thus, by the central limit theorem,

$$\frac{\nu_j - nE(\nu_j)}{\sqrt{nVar(\nu_j)}} = \frac{\nu_j - np_j}{\sqrt{np_j(1 - p_j)}} \to N(0, 1) \tag{2.13}$$

---

[5]The sum of i i.i.d. distributions.
[6]Skipping some work to show that for an indicator I, $E(I) = p$ and $Var(I) = p(1 - p)$

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to \sqrt{1 - p_j}N(0,1) = N(0, 1 - p_j) \approx N(0,1) \qquad (2.14)$$

Thus,

$$\frac{(\nu_j - np_j)^2}{np_j} \to N(0,1)^2 \qquad (2.15)$$

We would expect summing over the r boxes to obtain χ_r^2. However, notice that ν_i are not independent. Because $\sum_r \nu_i = n$, the system has one less degree of freedom (i.e. if we know the first $r-1$ values of ν_i, then we must know ν_r). The fact that ν_i are not independent is a major oversight and determining $\sum_r X_i^2$ is not so simple!

Regardless, we have now shown some intuition for why Pearson's Chi-squared value should converge to χ_{r-1}^2. Finally, we can use this relation to perform a p-test for our hypothesis.

For example, suppose a system has 6 degrees of freedom and we find that (where O_i are the observed values for each category and E_i are the expected values for each category),

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 12.44 \qquad (2.16)$$
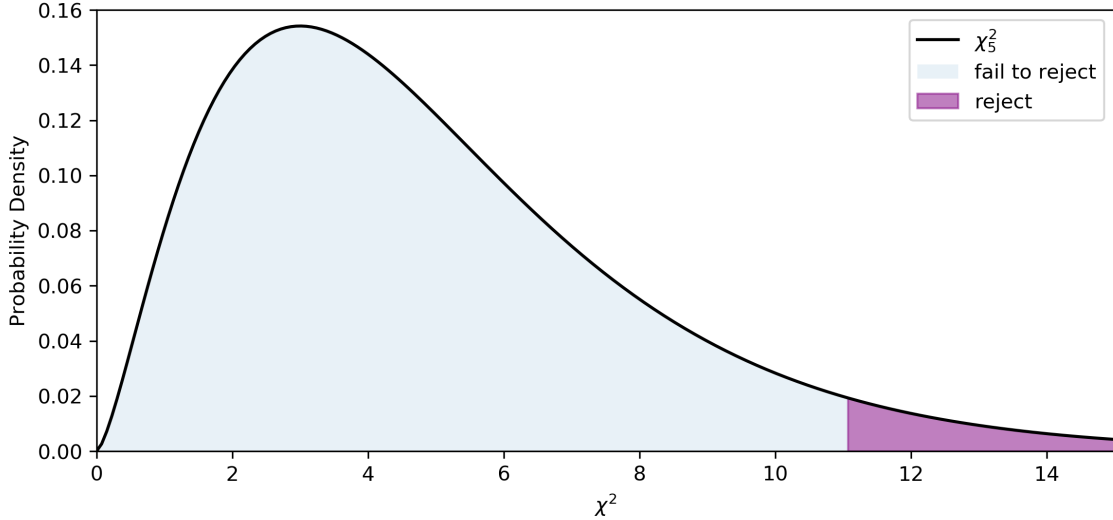
For a $\chi_{k=5}^2$ distribution with a confidence interval of 0.95, the critical value is 11.07. This means that there is a 5% chance of observing a χ^2 greater than 11.07.

In this case, $12.44 > 11.07$, we reject our null hypothesis and claim our model is incorrect. If instead, $\chi^2 < 11.07$, we fail to reject the null hypothesis.