11. Chi-Squared Test for a Family of Discrete Distributions

Bookmark this page

Exercises due Aug 3, 2021 19:59 EDT
In the problems on this page, you will apply the χ^2 goodness of fit test to determine whether or not a sample has a binomial distribution.

So far, we have used the χ^2 test to determine if our data had a categorical distribution with specific parameters (e.g. uniform on an N element set).

For the problems on this page, we extend the discussion on χ^2 tests **beyond** what was discussed in lecture to the following more general statistical set-up.

Let $X_1, \ldots, X_n \overset{iid}{\sim} X \sim \mathbf{P}$ denote iid discrete random variables supported on $\{0, \ldots, K\}$. We will decide between the following null and alternative hypotheses:

$$\begin{aligned} H_0: &\quad \mathbf{P} \in \{\text{Bin}(K, \theta)\}_{\theta \in (0,1)} \\ H_1: &\quad \mathbf{P} \notin \{\text{Bin}(K, \theta)\}_{\theta \in (0,1)}, \end{aligned}$$

where the null hypothesis can be rephrased as:

$$H_0: \quad \text{there exists } \theta \in (0,1) \text{ such that for all } j = 0, \ldots, K, \text{ we have } P(X = j) = \binom{K}{j} \theta^j (1-\theta)^{K-j}.$$

Review: Log-likelihood for a Binomial Distribution

2/2 points (graded)
Let $(\{0, \ldots, K\}, \{\text{Bin}(K, \theta)\}_{\theta \in (0,1)})$ denote a binomial statistical model. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bin}(K, \theta^*)$ for some unknown parameter $\theta^* \in (0,1)$.

The log-likelihood of this statistical model can be written

$$C + A \log B + (nK - A) \log(1 - B)$$

where C is independent of θ, A depends on $\sum_{i=1}^n X_i$, and B depends on θ.

What is A?

Use **Sigma** to stand for $\sum_{i=1}^n X_i$.

What is B?

STANDARD NOTATION

Submit You have used 1 of 4 attempts

✔ Correct (2/2 points)

Review: MLE for a Binomial Distribution

1/1 point (graded)

As above, let $\left(\{0,\ldots,K\},\{\text{Bin}\,(K,\theta)\}_{\theta\in(0,1)}\right)$ denote a binomial statistical model. Let $X_1,\ldots,X_n \overset{iid}{\sim} \text{Bin}\,(K,\theta^*)$ for some unknown parameter $\theta^* \in (0,1)$.

Which of the following denotes the MLE for θ^*?

- (selected) $\sum_{i=1}^n X_i$
- $\frac{1}{n}\sum_{i=1}^n X_i$
- $\frac{1}{K}\sum_{i=1}^n X_i$
- $\frac{1}{nK}\sum_{i=1}^n X_i$

Submit

You have used 1 of 2 attempts

✔ Correct (1/1 point)

χ^2 **-Test for a Family of Distributions** :

Now, we return to the following more general statistical set-up.

Let $X_1,\ldots,X_n \overset{iid}{\sim} \mathbf{P}$ denote iid discrete random variables supported on $\{0,\ldots,K\}$. We will decide between the following null and alternative hypotheses.

$$H_0:\quad \mathbf{P}\in\{\text{Bin}\,(K,\theta)\}_{\theta\in(0,1)}.$$

$$H_1:\quad \mathbf{P}\notin\{\text{Bin}\,(K,\theta)\}_{\theta\in(0,1)}.$$

Let f_θ denote the pmf of the distribution $\text{Bin}\,(K,\theta)$, and let $\hat{\theta}$ denote the MLE of the parameter θ from the previous problem.

Further, let N_j denote the number of times that j ($j\in\{0,1,\ldots,K\}$) appears in the data set $X_1,\ldots,X_n$ (so that $\sum_{j=0}^{K} N_j = n$.) The χ^2 **test statistic for this hypothesis test** is defined to be

$$T_n := n\sum_{j=0}^{K}\frac{\left(\frac{N_j}{n}-f_{\hat{\theta}}\,(j)\right)^2}{f_{\hat{\theta}}\,(j)}.$$

This statistic is different from before. Previously, under the null hypothesis, $\mathbf{P}\,(X=j)=p_j$ for some fixed p_j. Here, instead, we use $f_{\hat{\theta}}\,(j)$ to estimate $\mathbf{P}\,(X=j)$. This statistic still converges in distribution to a χ^2 distribution, but the number of degrees of freedom is smaller.

Degrees of Freedom for χ^2 Test for a Family of Distribution

More generally, to test if a distribution $\mathbf{P}$ is described by some member of a family of discrete distributions $\{\mathbf{P}_\theta\}_{\theta\in\Theta\subset\mathbb{R}^d}$ where $\Theta\subset\mathbb{R}^d$ is d-dimensional, with support $\{0,1,2,\ldots,K\}$ and pmf f_θ, i.e. to test the hypotheses:

$$H_0:\quad \mathbf{P}\in\{\mathbf{P}_\theta\}_{\theta\in\Theta}$$

$$H_1:\quad \mathbf{P}\notin\{\mathbf{P}_\theta\}_{\theta\in\Theta},$$

then if indeed $\mathbf{P} \in \{\mathbf{P}_\theta\}_{\theta \in \Theta \subset \mathbb{R}^d}$ (*i.e.*, the null hypothesis H_0 holds), and if in addition some technical assumptions hold, then we have that

$$T_n := n \sum_{j=0}^{K} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)} \xrightarrow[n \to \infty]{(d)} \chi^2_{(K+1)-d-1}.$$

Note that $K+1$ is the support size of $\mathbf{P}_\theta$ (for all θ.)

In our example testing for a binomial distribution, the parameter θ is one-dimensional, i.e. $d = 1$. Therefore, under the null hypothesis H_0, it holds that

$$T_n \xrightarrow[n \to \infty]{(d)} \chi^2_{(K+1)-1-1} = \chi^2_{K-1}.$$

Chi-squared Test for a Binomial Distribution on a Sample Data Set I

1 point possible (graded)

Consider the same statistical set-up as above. In particular, we have the test statistic

$$T_n := n \sum_{j=0}^{K} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)}.$$

where $\hat{\theta}$ is the MLE for the binomial statistical model $(\{0, 1, \ldots, K\}, \{\text{Bin}(K, \theta)\}_{\theta \in (0,1)})$.

We define our test to be

$$\psi_n = \mathbf{1}(T_n > \tau),$$

where τ is a threshold that you will specify. For the remainder of this page, we will assume that $K = 3$ (the sample space is $\{0, 1, 2, 3\}$).

What value of τ should be chosen so that ψ_n is a test of asymptotic level 5%? Give a numerical value with at least 3 decimals.

(Use this table or software to find the quantiles of a chi-squared distribution.)

$\tau =$

Submit You have used 0 of 2 attempts

Chi-squared Test for a Binomial Distribution on a Sample Data Set II

3 points possible (graded)

Consider the same statistical set-up as above. Suppose we observe a data set consisting of 1000 observations as described in the following (format: i, number of observations of i):

i	N_i
0	339
1	455
2	180
3	26

What is the value of the test statistic T_n for this data set? Give a numerical value with at least 4 decimals. (You are encouraged to use computational software.)

$T_n =$

What is the p-value of this data set with respect to the test ψ_{1000}? Give a numerical value with at least 4 decimals.

Use this tool to find the tail probabilities of a χ^2 distribution (you may also use any other software). If you are using this tool, note that you need to set "Choose Type of Control" to "Adjust X-axis quantile (Chi square) value" to find the tail probability associated with an x-axis value for a chi-squared distribution with degrees of freedom set in the "Degrees of Freedom" box.

p-value:

If ψ_n is designed to have level 5%, would you **reject** or **fail to reject** on the given data set?

- Reject
- Fail to reject

Submit You have used 0 of 3 attempts

Discussion

Hide Discussion

Topic: Unit 4 Hypothesis testing:Lecture 15: Goodness of Fit Test for Discrete Distributions / 11. Chi-Squared Test for a Family of Discrete Distributions

Add a Post

Show all posts by recent activity

? Tn Calculation
I feel like I'm the only one who is lost. What is sigma(Xi) when replacing theta with MLE for f(theta)(j) calculation? 2

? Confusion on the last question. Test Statistic much too high? 3

? Any hint for last question?
How does the function f(j) looks like? 13

? Any hint for Tn?
Specifically What is f_hattheta(j) ? 4

? ?? A little help with the PMF
Just want to ensure I'm interpreting the PMF (and MLE) correctly. I have the MLE, which suggests I should be using [redacted] as my denominat... 2

? question about degrees of freedom (Chi-squared Test for a Binomial Distribution on a Sample Data Set I)
Why is it not K=4 but K=3 otherwise the sample space have four numbers(0, 1, 2, 3)? I can't understand how to judge the degrees of freedom of ... 2

? Testing for a Binomial with a specific probability
In this question we tested to see if our data could come from a Binomial distribution. Can you also set up a Chi-Square test to see if your data ca... 2

? Why the zero isn't calculated as part of K degrees of freedom?
I know it is probably a dumb question but I was tripped to it. 4

? Range of j (Review: Log-likelihood for a Binomial Distribution)
In the solution of exercise Review: Log-likelihood for a Binomial Distribution...shouldn't the range of j, in the definition of a binomial distribution, b... 4

? Binomial K-1 degrees of freedom gives 0 degrees of freedom for Bernoulli (as an edge case) 4

? "Standard Notation" page is not accessible
The "standard notation" page is not accessible anymore; any reason? 1