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

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:

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

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 $(\{0, \ldots, K\}, \{\mathrm{Bin}\,(K, \theta)\}_{\theta \in (0,1)})$ denote a binomial statistical model. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Bin}\,(K, \theta^*)$ 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 $\sum_{i=1}^n X_i$.

Sigma ✔ **Answer:** Sigma

Σ

What is B?

theta ✔ **Answer:** theta

θ

STANDARD NOTATION

Solution:

The pmf of $\mathrm{Bin}\,(K, \theta)$ is

$$j \mapsto \binom{K}{j} \theta^j (1-\theta)^{K-j}$$

for $j \in \{1, \ldots, K\}$.

Therefore, the likelihood is given by

$$\begin{aligned} L_n(X_1, \ldots, X_n, \theta) &= \prod_{i=1}^n \left(\binom{K}{X_i} \theta^{X_i} (1-\theta)^{K-X_i} \right) \\ &= \left(\prod_{i=1}^n \binom{K}{X_i} \right) \theta^{\sum_{i=1}^n X_i} (1-\theta)^{nK - \sum_{i=1}^n X_i}. \end{aligned}$$

Taking the logarithm, we have

$$\log L_n(X_1, \ldots, X_n, \theta) = \log \left(\prod_{i=1}^n \binom{K}{X_i} \right) + \left(\sum_{i=1}^n X_i \right) \log \theta + \left(nK - \sum_{i=1}^n X_i \right) \log(1-\theta).$$

Therefore, $A = \sum_{i=1}^n X_i$ and $B = \theta$.

Submit You have used 1 of 4 attempts

Answers are displayed within the problem

Review: MLE for a Binomial Distribution

1/1 point (graded)

As above, 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 $\theta^* \in (0,1)$.

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

- $\sum_{i=1}^n X_i$
- $\frac{1}{n} \sum_{i=1}^n X_i$
- $\frac{1}{K} \sum_{i=1}^n X_i$
- $\frac{1}{nK} \sum_{i=1}^n X_i$ ✔

Solution:

Recall from the previous problem that

$$\log L_n(X_1, \ldots, X_n, \theta) = C + \left(\sum_{i=1}^n X_i \right) \log \theta + \left(nK - \sum_{i=1}^n X_i \right) \log(1-\theta)$$

where C does not depend on θ.

To compute the MLE, we need to maximize the above with respect to the parameter θ. We set the derivative to be 0:

$$0 = \frac{\sum_{i=1}^{n} X_i}{\theta} - \frac{nK - \sum_{i=1}^{n} X_i}{1-\theta}.$$

The above holds when

$$p = \frac{1}{nK}\sum_{i=1}^{n} X_i.$$

Therefore, the right-hand side is the MLE for this statistical model.

Submit   You have used 1 of 2 attempts

Answers are displayed within the problem

χ^2-Test for a Family of Distributions :

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

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

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

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:

$$\begin{aligned} H_0: &\quad \mathbf{P} \in \{\mathbf{P}_\theta\}_{\theta\in\Theta} \\ H_1: &\quad \mathbf{P} \notin \{\mathbf{P}_\theta\}_{\theta\in\Theta}, \end{aligned}$$

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

1/1 point (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.)

$\tau =$ 5.991 ✔ **Answer:** 5.991

Solution:

Since $K=3$ and $d=1$, we know that the limiting distribution of T_n is χ^2_2. Thus, the asymptotic level is the value τ such that

$$\lim_{n\to\infty} P(T_n > \tau) = P(Z > \tau) = 0.05$$

where $Z \sim \chi^2_2$. Hence, τ should be chosen to be 5.991 (from the given table).

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

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

3/3 points (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.)

$T_n =$ 0.88286 ✔ **Answer:** 0.8829

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: 0.64312 ✔ **Answer:** 0.6431

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 ✔

Solution:

Observe that the MLE is given by

$$\hat{p} = \frac{1}{3 \cdot 1000}(455 + 2 \cdot 180 + 3 \cdot 26) \approx 0.29767.$$

Thus for this data set,

$$\begin{aligned} T_n = 1000 \ &\cdot\left(\frac{\left(\frac{339}{1000} - \binom{3}{0}\left(0.2977^0\right)(0.7023)^{3-0}\right)^2}{\binom{3}{0}\left(0.2977^0\right)(0.7023)^{3-0}} + \frac{\left(\frac{455}{1000} - \binom{3}{1}\left(0.2977^1\right)(0.7023)^{3-1}\right)^2}{\binom{3}{1}\left(0.2977^1\right)(0.7023)^{3-1}} + \right. \\ &\left. \frac{\left(\frac{180}{1000} - \binom{3}{2}\left(0.2977^2\right)(0.7023)^{3-2}\right)^2}{\binom{3}{2}\left(0.2977^2\right)(0.7023)^{3-2}} + \frac{\left(\frac{26}{1000} - \binom{3}{3}\left(0.2977^3\right)(0.7023)^{3-3}\right)^2}{\binom{3}{3}\left(0.2977^3\right)(0.7023)^{3-3}} \right) \\ &\approx 0.8829 \end{aligned}$$

The asymptotic p-value for this data set is given by

$$\lim_{n \to \infty} P\left(T_n > 0.8829\right) = P\left(Z > 0.8829\right).$$

where $Z \sim \chi^2_2$. Consulting the suggested link, we see that $P\left(Z > 0.8829\right) \approx 0.6431$.

According to the golden rule of p-values, since $0.6431 > 0.05$, we should **fail to reject** the null hypothesis that $X_1, \ldots, X_{1000}$ are distributed as $\mathrm{Bin}(3, \mathrm{p})$ for some value of the parameter p.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Discussion

Show 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