1. Chi-squared Goodness of Fit Testing for a Gaussian Distribution

Recall that so far, we have applied the χ^2 to test for discrete distributions only. In the problems on this page, we will further extend the χ^2 goodness of fit test to determine whether or not a sample has a continuous distribution, and will use the family of Gaussian distribution as an example (which one of the most common).

Chi-squared Goodness of Fit Testing for a Gaussian Distribution I

3.0/3 points (graded)

Note: The solution to this part along with remarks will be available to you once you answer correctly or used all your attempts.

Let $X_1, \ldots, X_n \overset{iid}{\sim} X \sim \mathbf{P}$ for some unknown distribution $\mathbf{P}$ with continuous cdf F. Below we describe a χ^2 test for the null and alternative hypotheses

$$\begin{aligned} H_0 : \mathbf{P} &\in \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2 > 0} \\ H_1 : \mathbf{P} &\notin \{N(\mu, \sigma^2)\}_{\mu \in \mathbb{R}, \sigma^2 > 0}. \end{aligned}$$

We divide the sample space into 5 disjoint subsets refered to as **bins** :

$$\begin{aligned} A_1 &= (-\infty, -2), \quad A_2 = (-2, -0.5), \\ A_3 &= (-0.5, 0.5), \quad A_4 = (0.5, 2) \\ A_5 &= (2, \infty). \end{aligned}$$

Now, define **discrete** random variables Y_i as functions of X_i by

$$Y_i = k \qquad \text{if } X_i \in A_k.$$

For example, if $X_i = 0.1,$ then $X_i \in A_3$ and so $Y_i = 3.$ In other words, Y_i is the label of the bin that contains X_i.

By the definition above,

$$Y_1, \ldots, Y_n \overset{iid}{\sim} Y$$

and Y follows the multinomial distribution on $\{1,2,3,4,5\}$ with (vector) parameter $\mathbf{p} = (p_1 \quad p_2 \quad p_3 \quad p_4 \quad p_5) \in \Delta_5$ where p_j denote the probability that $Y = j$.

Assume the following special case of the null hypothesis holds:

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1).$$

What is the vector parameter $\mathbf{p} \in \Delta_5$ of the multinomial distribution followed by Y_i? Fill in the first three entries p_1, p_2, p_3 below.

(Enter **Phi(x)** for the cdf $\Phi(x)$ of a standard normal distribution, e.g. type **Phi(1)** for $\Phi(1)$, or enter your answers accurate to 3 decimal places)

$\mathbf{p}_1 =$ Phi(-2) ✔ **Answer:** Phi(-2)

$\mathbf{p}_2 =$ Phi(-0.5)-Phi(-2) ✔ **Answer:** Phi(-0.5)-Phi(-2)

$\mathbf{p_3} =$ Phi(0.5)-Phi(-0.5) ✔ **Answer:** Phi(0.5)-Phi(-0.5)

(What is p_4 and p_5 in terms of p_1, p_2, p_3?)

STANDARD NOTATION

Solution:

By the assumption in the problem statement, we have $X_1 \sim N(0,1)$. Therefore,

$$P(Y_1 = A_1) = P(X_1 \in (-\infty, -2)) = \Phi(-2) \approx 0.0228.$$

Hence $\mathbf{p_1} = 0.0228$. Similarly,

$$P(Y_1 = A_2) = P(X_1 \in (-2, -0.5)) = \Phi(-0.5) - \Phi(-2) \approx 0.2858$$

and

$$P(Y_1 = A_3) = P(X_1 \in (-0.5, 0.5)) = \Phi(0.5) - \Phi(-0.5) \approx 0.3829,$$

so $\mathbf{p_2} = 0.2858$ and $\mathbf{p_3} = 0.3829$.

Remark 1: By symmetry, under the assumption that $X_1, \ldots, X_n \stackrel{iid}{\sim} N(0,1)$, we have that $Y_1, \ldots, Y_n \stackrel{iid}{\sim} \mathbb{P}_{\mathbf{p}}$ where

$$\mathbf{p} = (0.0228, 0.2858, 0.3829, 0.2858, 0.0228).$$

Remark 2: In general, if the null hypothesis holds, we will not know the distribution of $X_1, \ldots, X_n$, but we will know that it is Gaussian with some unknown mean μ and unknown variance $\sigma^2 > 0$. Then we see that, for example,

$$\begin{aligned}
P(X_1 \in A_1) &= P(X_1 \in A_5) = \Phi_{\mu,\sigma^2}(-2) \\
P(X_1 \in A_2) &= P(X_1 \in A_4) = \Phi_{\mu,\sigma^2}(-0.5) - \Phi_{\mu,\sigma^2}(-2) \\
P(X_1 \in A_3) &= \Phi_{\mu,\sigma^2}(0.5) - \Phi_{\mu,\sigma^2}(-0.5).
\end{aligned}$$

If n is very large, then we may approximate these unknown quantities with the consistent estimators

$$\begin{aligned}
\Phi_{\widehat{\mu},\widehat{\sigma}^2}(-2) &\approx \Phi_{\mu,\sigma^2}(-2) \\
\Phi_{\widehat{\mu},\widehat{\sigma}^2}(-0.5) - \Phi_{\widehat{\mu},\widehat{\sigma}^2}(-2) &\approx \Phi_{\mu,\sigma^2}(-0.5) - \Phi_{\mu,\sigma^2}(-2) \\
\Phi_{\widehat{\mu},\widehat{\sigma}^2}(0.5) - \Phi_{\widehat{\mu},\widehat{\sigma}^2}(-0.5) &\approx \Phi_{\mu,\sigma^2}(0.5) - \Phi_{\mu,\sigma^2}(-0.5)
\end{aligned}$$

where $(\widehat{\mu}, \widehat{\sigma}^2)$ is the MLE for the statistical model $(\mathbb{R}, \{N(\mu,\sigma^2)\}_{\mu,\sigma^2})$, Gaussian with unknown mean and unknown variance. These estimators will be used to design our χ^2 test statistic in the next problem.

Submit You have used 2 of 3 attempts

ℹ Answers are displayed within the problem

Chi-squared Goodness of Fit Testing for a Gaussian Distribution II

1/1 point (graded)

Recall the statistical set-up above. Recall that $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathbf{P}$ are iid from an unknown distribution $\mathbf{P}$. For all $1 \le i \le n$, Y_i is a discrete random variable supported on $\{1, \ldots, 5\}$ that denotes which bin contains the realization of X_i.

Let $\mathbf{P}_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$ and let $(\widehat{\mu}, \widehat{\sigma}^2)$ denote the MLE for the statistical model $(\mathbb{R}, \{P_{\mu,\sigma^2}\}_{\mu\in\mathbb{R},\sigma^2\in(0,\infty)})$, i.e. Gaussian with unknown mean and unknown variance. For $1 \leq j \leq 5$, let N_j denote the **frequency** of j (i.e. number of times that j appears) in the data set $Y_1, \ldots, Y_n$.

Define the χ^2 test statistic

$$T_n = n \sum_{j=1}^{5} \frac{\left(\frac{N_j}{n} - P_{\widehat{\mu},\widehat{\sigma}^2}(Z \in A_j)\right)^2}{P_{\widehat{\mu},\widehat{\sigma}^2}(Z \in A_j)}.$$

where $Z \sim \mathcal{N}(\widehat{\mu}, \widehat{\sigma}^2)$. Then it holds that

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_\ell$$

for some constant $\ell > 0$.

What is ℓ?

Hint: Use the result on the very last page of Lecture 15.

$l =$ 2 ✔ **Answer:** 2

Solution:

Consider the finite case of the χ^2 goodness of fit test. In this case, we are trying to figure out if an iid sample $Z_1, \ldots, Z_n \overset{iid}{\sim} Q$ is generated from some member of a family of distributions $\{Q_\theta\}_{\theta\in\mathbb{R}^d}$ with support $\{1, \ldots, K\}$ and pmf f_θ. If indeed $Q \in \{Q_\theta\}_{\theta\in\mathbb{R}^d}$ and some additional technical assumptions hold, then

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

where $\hat{\theta}$ is the MLE under the statistical model $(\{1, \ldots, N\}, \{Q_\theta\}_{\theta\in\mathbb{R}^d})$ and for $1 \leq j \leq K$, N_j denotes the number of times that j appears in the data set $Z_1, \ldots, Z_n$.

We apply this convergence result to the discrete random variables $Y_1, \ldots, Y_n$. Under the null hypothesis that $X_1, \ldots, X_n$ have a Gaussian distribution $N(\mu, \sigma^2)$ for some unknown mean μ and variance $\sigma^2 > 0$, then from the remark in the solution to the previous problem, we know that

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \mathbf{P}_{\mathbf{p}}$$

where for $1 \leq j \leq 5$,

$$\mathbf{p}_j = P_{\mu,\sigma^2}(X_1 \in A_j).$$

Then we have that

$$\widehat{\mathbf{p}}_j := P_{\widehat{\mu},\widehat{\sigma}^2}(W \in A_j), \quad W \sim N(\widehat{\mu}, \widehat{\sigma}^2)$$

plays the role of $f_{\hat{\theta}}(j)$ above, N_j denotes the number of times A_j appears in the data set $Y_1, \ldots, Y_n$, the support size is $K = 5$, and the dimension of the MLE is $d = 2$. Therefore, we have that

$$n \sum_{j=1}^{5} \frac{\left(\frac{N_j}{n} - P_{\widehat{\mu},\widehat{\sigma}^2}(Z \in A_j)\right)^2}{P_{\widehat{\mu},\widehat{\sigma}^2}(Z \in A_j)} \xrightarrow[n\to\infty]{(d)} \chi^2_\ell$$

where

$$\ell = K - d - 1 = 5 - 2 - 1 = 2$$

.

Remark: As in the other χ^2 tests we have seen, the distribution χ^2_{K-d-1} is **pivotal**, so its quantiles can be consulted using a table. We can use this to design a test of asymptotic level η of the form

$$\psi_n = \mathbf{1}\left(T_n > q_\eta\right)$$

where q_η is the η quantile of χ^2_{K-d-1}. Note however, for any fixed n, the distribution of T_n is **not pivotal**. Hence, a test designed in this form is inherently **non-asymptotic**.

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Asymptotic versus Non-asymptotic Normality Tests

0/1 point (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathbf{P}$ for some distribution with continuous cdf. A **normality test** is a hypothesis test where the null and alternative hypothesis are specified by

$$H_0 : P \in \mathcal{F}$$
$$H_1 : P \notin \mathcal{F}$$

where $\mathcal{F} \subset \{N(\mu, \sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0}$, *i.e.* $\mathcal{F}$ is a **subset** of the family of all Gaussian distributions.

For example, the Kolmogorov-Smirnov test is a normality test for $\mathcal{F} = \{\mathcal{N}(0,1)\}$– that is, when $\mathcal{F}$ consists of a single Gaussian distribution. The Kolmogorov-Lilliefors test is a normality test with $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0}$– that is, when $\mathcal{F}$ consists of all Gaussian distributions. The χ^2 test studied on this page is also a normality test with $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2)\}_{\mu\in\mathbb{R},\sigma^2>0}$.

Which of these tests mentioned above are non-asymptotic in the sense that, $\boxed{\text{for any fixed } n,}$ the distribution of the test statistic under the null can be consulted via tables? (Hence, it is possible to specify the *non-asymptotic level* of the test and not just the asymptotic level.) (Choose all that apply.)

对于所有的n

- [ ] Kolmogorov-Smirnov Test ✔
- [x] Kolmogorov-Lilliefors Test ✔
- [x] χ^2 test

✖

Solution:

We examine the choices in order.

- The first choice "Kolmogorov-Smirnov test" is correct. In Lecture 4.4, we discussed that the Kolmogorov-Smirnov test statistic

$$\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-\Phi|$$

is pivotal for all $n \in \mathbb{R}$ under the null hypothesis that the cdf of our data is given by Φ, the cdf of a standard normal. We also saw tables of the distribution of the test statistic for several values of n.

- The second choice "Kolmogorov-Lilliefors test" is correct. Previously in lecture 16, we discussed that the Kolmogorov-Lilliefors test statistic

$$\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t)-\Phi_{\widehat{\mu},\widehat{\sigma}^2}|$$

is pivotal for all $n \in \mathbb{R}$ under the null hypothesis that the cdf of our data is the cdf of *some* Gaussian distribution. The distribution of this test statistic is different form that of the Kolmogorov-Smirnov test (in particular, the Kolmogorov-Lilliefors test statistic has smaller quantiles). However, its distribution may still be referred to in tables.

- The third choice "χ^2 test" is incorrect. As emphasized in both the finite and infinite cases, the χ^2 test is only asymptotic. For small n, the test statistic

$$n\sum_{j=1}^{K}\frac{\left(\frac{N_i}{n}-p_j(\hat{\theta})\right)^2}{p_j(\hat{\theta})}$$

can depend heavily on the distribution of $X_1,\ldots,X_n$ (even under the null hypothesis). Therefore, this test is **not** asymptotic, because we do not know the distribution of the test statistic.

this means, we do not know the exact distribution when n is small

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 4 Hypothesis testing:Homework 8 / 1. Chi-squared Goodness of Fit Testing for a Gaussian Distribution