36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 13: χ^2 Tests and p-Values

Lecturer: Jing Lei

13.1 Review and Outline

Last time we discussed hypothesis testing:

1. Simple versus simple hypothesis tests and the Neyman-Pearson Lemma.
2. Wald's test.

In this note we will discuss the χ^2 test and p-values. We will follow Chapter 10 of the Wasserman book.

13.2 A test for normal means

Suppose we observe $X_1, \ldots, X_n \sim N(\theta, I_d)$, where $\theta \in \mathbb{R}^d$, and I_d is the $d \times d$ identity matrix. We would like to test the hypotheses:

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta \neq \theta_0. \end{aligned}$$

In this case, we could first form:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and observe that under the null:

$$\sqrt{n}(\widehat{\theta} - \theta_0) \sim N(0, I_d).$$

This motivates the use of the test statistic:

$$T = n \sum_{j=1}^{d} (\widehat{\theta}_j - \theta_{0j})^2,$$

which under the null has a distribution that equals the sum of squares of d independent standard normals.

13.3 The χ^2 distribution

Suppose $Z_1, \ldots, Z_k$ are independent standard Gaussian RVs. Then the variable:

$$V_k = \sum_{i=1}^{k} Z_i^2,$$

has a χ^2 distribution with k degrees of freedom. It is easy to check that the mean $\mathbb{E}[V_k] = k$, with some effort you can also show that the variance $\mathrm{Var}(V_k) = 2k$.

The density function of χ_k^2 distribution is

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-x/2}}{2^{k/2}\Gamma(k/2)} \mathbf{1}(x > 0)\,,$$

where $\Gamma(\cdot)$ is the Gamma function. The exact form of $\Gamma(\cdot)$ does not matter because we can just view it as a normalization constant to make the density function integrate to 1.

We will use $\chi_{k,\alpha}^2$ to denote the upper α quantile of this distribution, i.e.

$$\mathbb{P}(V_k \geq \chi_{k,\alpha}^2) = \alpha.$$

Back to the normal mean test. We know that $T = n\sum_{j=1}^{d}(\hat{\theta}_j - \theta_{0j})^2$ has the χ_d^2 distribution. So we would reject the null hypothesis when:

$$T \geq \chi_{d,\alpha}^2\,.$$

13.4 Pearson's χ^2 test

Pearson's χ^2 test was originally used to test hypotheses about multinomials. More generally, one can use it to test hypotheses about multivariate parameters that have a Gaussian distribution (at least asymptotically) under the null.

Suppose that we observe counts for n samples drawn from a multinomial on k categories with probabilities $(p_1, \ldots, p_k)$: $(X_1, \ldots, X_k)$. The MLE is given by:

$$\widehat{p}_i = X_i/n.$$

For some fixed vector of probabilities p_0, we want to test the hypotheses:

$$\begin{aligned} H_0 &: \quad p = p_0 \\ H_1 &: \quad p \neq p_0. \end{aligned}$$

Pearson suggested the test statistic:

$$T = \sum_{i=1}^{k} \frac{(X_i - np_{0i})^2}{np_{0i}}.$$

Here X_i is the observed count, and np_{0j} expected count so the statistic makes intuitive sense: we expect it to be large under the alternate and small under the null. The precise reason for this form actually comes from the multivariate central limit theorem. The test statistic T then just has a χ^2 distribution with $(k-1)$ degrees of freedom under the null. We would reject the null hypothesis if:

$$T \geq \chi^2_{k-1,\alpha}.$$

The Pearson's test is generally used as an analog of Wald's test when you are estimating multiple parameters. It can be used in any case where you have an appropriate central limit theorem for the parameters.

Example: Mendel's experiment. Mendel's theory predicts that the probability of a pea plant in a certain hybrid generation falling in one of four categories is $\vec{p}_0 = (9/16, 3/16, 3/16, 1/16)$. The observed count vector is $X = (315, 101, 108, 32)$ with a total count $n = 556$. The predicted counts are $\vec{E}_0 = n\vec{p}_0 = (312.75, 104.25, 104.25, 34.75)$. The χ^2 statistic is

$$T = \frac{(315-312.75)^2}{312.75} + \frac{(101-104.25)^2}{104.25} + \frac{(108-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.47\,.$$

The 0.05 upper quantile of χ^2_3 is 7.815. So we retain the null.

13.5 p-values

In the Mendel's example above, the test statistic $T = 0.47$ seems to be quite small compared to the rejection threshold value 7.815 for $\alpha = 0.05$. Therefore, one may wonder if we can say something about how far our test statistic is from the rejection threshold. This leads to the concept of p-values.

A first question to answer here is: How far is far? In Figure 13.1 we see that the observed test statistic is rather small, because a majority of probability mass of χ^2_3 is on the right hand side of the point 0.47.

Suppose we have a hypothesis testing problem (that is, the model $f_\theta(x)$, the null and alternative hypotheses), an observed data set $(X_1, ..., X_n)$, a test statistic $T_n = T_n(X_1, ..., X_n)$, and rejection region $R = \{\vec{x} : T_n(\vec{x}) \geq t\}$. Suppose we have a simple null hypothesis $H_0 : \theta = \theta_9$. The p-value can be defined in one of the following two equivalent ways.

1. p-value is the answer to the question: "what is the smallest α for which the test would reject the null hypothesis?"

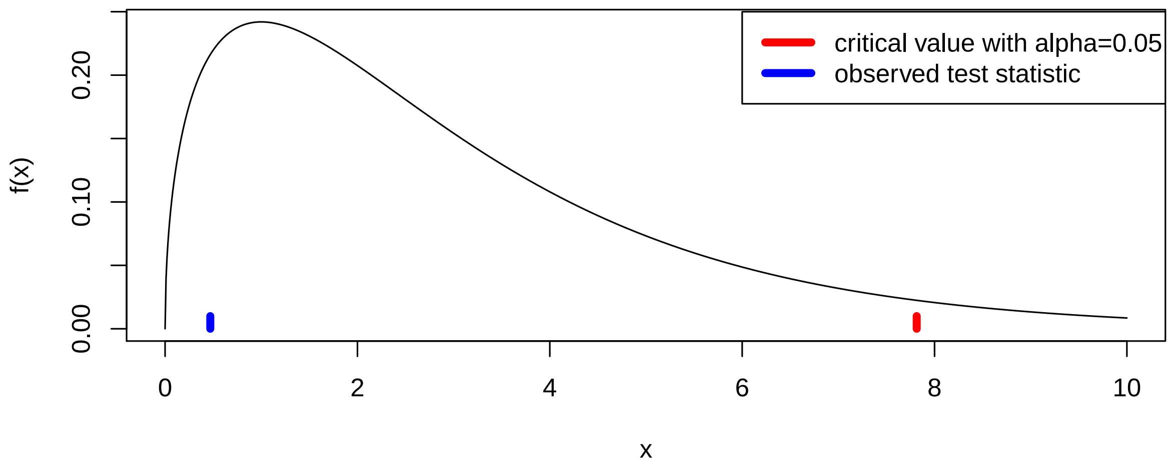

Figure 13.1: Density function of χ^2_3. Highlighted locations are the rejection threshold with $\alpha = 0.05$ (red), and the observed test statistic in Mendel's experiment (blue).

2. Let $\vec{X}'$ be an independent draw from $f_{\theta_0}(\cdot)$, the p-value is the conditional probability

$$\mathbb{P}_{\theta_0}\left[T_n(\vec{X}') \geq T_n(\vec{X})\middle|\vec{X}\right].$$

Remark: The most important thing to know about p-value is

$$\boxed{p\text{-value} \;\; \text{IS NOT} \;\; \mathbb{P}(H_0 \text{ is true})\,.}$$

The latter probability is not well-defined since θ is not random (p-values are not used in a Bayesian framework).

Remark: The second most important thing to remember about p-value is that

$$\boxed{p\text{-value is a random variable, because it is a function of the data.}}$$

Intuitively, a smaller p-value is stronger evidence against the null. Scientists often report p-values, and informally a p-value of < 0.01 is considered strong evidence against the null, and < 0.05 is moderate evidence against the null.

In the Mendel's experiment, the p-value for the χ^2 test is

$$P(V_3 \geq 0.47) \approx 0.95.$$

The second definition gives a common and useful interpretation of a p-value: A p-value is just the probability under the null of seeing a more (or equally) extreme test statistic than the one you actually observed. Intuitively, if it is very unlikely to see such an extreme test statistic under the null then we should reject the null. Recall that the "extremeness" is defined by larger values of T_n.

In the second definition if we have a composite null then it changes to

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left[T_n(\vec{X}') \geq T_n(\vec{X}) \middle| \vec{X} \right].$$

Example: p-values for the Wald test: In the two-sided Wald test, we have a simple null $\theta = \theta_0$, and our test rejects at level α if $|T_n| \geq z_{\alpha/2}$, where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

Suppose that T_n is positive, then the probability of seeing a test statistic larger that T_n is just $1 - \Phi(T_n)$ and the probability of seeing a test statistic smaller than $-T_n$ is $\Phi(-T_n) = 1 - \Phi(T_n)$. In this case the p-value:

$$\text{p-value} = 2(1 - \Phi(T_n)) = 2\Phi(-T_n).$$

More generally, when we do not assume T_n is positive this is given by:

$$\text{p-value} = 2\Phi(-|T_n|).$$

Fact (proof is an exercise): When the null hypothesis is simple and T_n has a continuous distribution then the p-values under the null will have a $U[0, 1]$ distribution.