36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 24: October 28

Lecturer: Siva Balakrishnan

24.1 Review and Outline

In the last class we discussed hypothesis testing:

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

Today we will continue our discussion of the Wald test, and then discuss p-values and the χ^2 test. We will follow Chapter 10 of the Wasserman book.

24.2 Wald Test Example

We discussed the Wald test for the parameter of a Bernoulli. Now lets consider the Wald test for comparing two Bernoullis via a paired comparison test.

Example 1: A typical scenario is that we have two prediction algorithms that we would like to compare. Suppose first, that we test the two algorithms by evaluating them on two different test sets: Algorithm 1 on a test set of size m and Algorithm 2 on a test set of size n. Let X be the number of correct predictions made by Algorithm 1 and Y be the number of correct predictions made by Algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$ where p_1 and p_2 are the error rates of the first and second Algorithm respectively. Defining $\delta = p_1 - p_2$, our hypotheses are:

$$\begin{aligned} H_0 : &\quad \delta = 0 \\ H_1 : &\quad \delta \neq 0. \end{aligned}$$

To do a Wald test, we compute the MLEs $\widehat{p}_1$ and $\widehat{p}_2$, and compute $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$. We can also estimate the variance of our estimator in the usual way:

$$\text{Var}(\widehat{\delta}) = \frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}.$$

So our Wald test would reject the null hypothesis if

$$\frac{|\widehat{\delta}|}{\sqrt{\mathrm{Var}(\widehat{\delta})}} \geq \Phi^{-1}(1-\alpha/2).$$

The more typical setting is when we have a single data set on which we evaluate both algorithms. In this case the two predictions are not independent so we cannot directly compute the variance of $\widehat{\delta}$. You can try to do this and observe that it will depend on the covariance between the two algorithms which you will have to estimate.

A convenient way to test this hypothesis is via a paired comparison test. The high-level principle is to try to pair outcomes in a certain way. Typically this is done in randomized studies, for instance via paired t-tests, in order to eliminate certain biases. We will revisit this idea more generally when we discuss causal inference. In this case, one can imagine testing both Algorithms on the data but recording, the difference $D_i = X_i - Y_i$, i.e.

1. $D_i = 0$ if both algorithms have the same prediction.
2. $D_i = 1$ if Algorithm 1 is correct while 2 is wrong.
3. $D_i = -1$ if Algorithm 2 is correct while 1 is wrong.

In this case you can observe that:

$$\delta = \mathbb{E}(X) - \mathbb{E}(Y) = \mathbb{E}(X - Y) = \mathbb{E}(D).$$

Our test statistic is simply:

$$T_n = \frac{\frac{1}{n}\sum_{i=1}^n D_i}{\sqrt{\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n D_i)}},$$

and the variance is:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n D_i\right) = \frac{1}{n^2}\sum_{i=1}^n \left(D_i - \frac{1}{n}\sum_{i=1}^n D_i\right)^2.$$

As usual the Wald test would reject the null hypothesis if:

$$|T_n| \geq \Phi^{-1}(1-\alpha/2).$$

24.2.1 Wald and confidence intervals

The Wald test for $H_0 : \theta = \theta_0$, versus $H_1 : \theta \neq \theta_0$, can also be viewed as just computing an asymptotic $1 - \alpha$ interval for θ and rejecting the null if the interval does not contain θ_0.

Formally, we compute the interval:

$$C_n = \left(\frac{\widehat{\theta}}{\sqrt{\text{Var}(\widehat{\theta})}} - z_\alpha/2, \frac{\widehat{\theta}}{\sqrt{\text{Var}(\widehat{\theta})}} + z_\alpha/2 \right)$$

which is an asymptotic $1 - \alpha$ confidence interval, i.e.

$$\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha.$$

We reject the null if $\theta_0 \notin C_n$.

It is also worth understanding the distinction between statistical significance, and scientific significance. Suppose that we are measuring the effect of a drug and the null is false and $\theta = \theta_1 \neq \theta_0$ but $\theta_1 - \theta_0$ is tiny. Then typically, a statistical test will reject the null provided you collect enough data but this does not mean that the drug is very effective.

This is something that is more clear from a confidence interval, which in some sense contains more information, particularly it conveys the typical effect size which might be of more practical significance in some cases. On the other hand a test just says "reject" or "fail to reject".

24.3 p-values

A p-value is a way to try to extract more information than "reject" or "fail to reject" in the hypothesis testing framework.

We could ask: "what is the smallest α for which the test would reject the null hypothesis?"

This value is known as a p-value.

Formally, if for each α our test has a rejection region R_α then:

$$\text{p-value} = \inf \{\alpha : T_n \in R_\alpha\}.$$

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.

It is tricky to turn the above definition into something that we can actually analytically compute. The way this is done is by reducing calculating the p-value to computing a probability. Lets consider a simple example:

Suppose that the test is one-sided and the size α test is of the form: reject H_0 if

$$T_n \geq t_\alpha$$

for some threshold t_α. Then the p-value is given by:

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_{X_n \sim \theta}(T(X_n) \geq T_n).$$

If we have a simple null this reduces to:

$$\text{p-value} = \mathbb{P}_{X_n \sim \theta_0}(T(X_n) \geq T_n),$$

which is often easy to compute.

The above expression also gives a common 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.

It is easy to extend this reasoning to two-sided tests.

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

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:

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

You will prove this in your Homework: when the null hypothesis is simple then the p-values under the null will have a $U[0,1]$ distribution. Under the alternative one might expect that the p-values will be closer to zero on average.

24.4 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.

24.4.1 The χ^2 distribution

Suppose we have $Z_1, \ldots, Z_k$ which 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 $\mathrm{Var}(V_k) = 2k$.

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

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

24.4.2 Testing multinomials

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 central limit theorem. It is not difficult to show that:

$$\frac{(X_i - np_{0i})}{\sqrt{np_{0i}}} \to N(0, 1),$$

in distribution under the null hypothesis. This is nice because the test statistic T then just has a χ^2 distribution with k degrees of freedom under the null. In this case, we would reject the null hypothesis if:

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

The Pearson 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.