Lecture 26: Distribution-free Tests

Relevant textbook passages:

Larsen–Marx [8]: Chapter 14

26.1 Distribution-free tests

All of the significance testing we have discussed so far has been based on likelihood functions. That is we assume we know the function $f(x;\theta)$. There are hypothesis test that do not rely such knowledge. Instead they rely on the fact (Glivenko–Cantelli Theorem 7.10.3) that the empirical cdf is approximately the same as the cdf of the data generating process. Since the empirical cdf is a step function with jumps at the sample values, the jumps occur at the order statistics of the sample, and the next set of tests rely heavily on order statistics. This material is covered in Larsen–Marx [8, Chapter 14], but Breiman [1, Chapters 8–9], Hogg and Craig [3, § 9.6], and van der Waerden [13, § 63–64] provide more detail.

26.2 A test for the median

Larsen–Marx [8]: 14.2

If we have a sample from a continuous pdf, there is a simple test for the null hypothesis

$$H_0\colon \text{median } f = \theta_0, \quad \text{against the alternative} \quad H_1\colon \text{median } f \neq \theta_0.$$

By definition the probability of exceeding the median is 1/2, so for an independent sample $X_1, \dots, X_n$, the statistic

$$T = |\{i : x_i > \theta_0\}|$$

has a Binomial$(n, 1/2)$ distribution, which has mean $n/2$ and variance $n/4$.

For the two sided alternative the test takes the form: choose a critical value k^* and reject the null hypothesis if

$$T \leqslant k^* \text{ or } T \geqslant n - k^*.$$

Unless you want to use a randomized rule, you are unlikely to find to find a k^* that will give you a significance level, of say 0.05, but you can compute the Binomial$(n, 1/2)$ probability of $P(T \leqslant k^* \text{ or } T \geqslant n - k^*)$ to get the size of the test.

Or if n is large enough ($\geqslant$ 10) you could use the Normal approximation and treat $z = (T - n/2)/\sqrt{n/4}$ as a standard Normal and reject the null if $|z| \geqslant z_{\alpha/2}$, to get a test of size α.

One-sided tests are analogous.

26.3 Testing the equality of two distributions

Suppose we have two random samples $x_1, \dots, x_n$ and $y_1, \dots, y_m$, and want to know if they came from the same continuous distribution. The null hypothesis and the alternative are

$$H_0\colon f_X = f_Y \qquad H_1\colon f_X \neq f_Y.$$

This exposition is based on Breiman [1, pp. 290–298].

26.3.1 A simple test

If $n+m$ is even, a simple idea to test the null hypothesis is this: Let $\bar{m}$ be the sample median. If the null hypothesis is true, then the number N of the x_is that are less than $\bar{m}$ should be about half of the x_is, namely $n/2$. In fact, the exact distribution of N is given by

$$P(N=k) = \frac{\binom{n}{j}\binom{m}{\frac{n+m}{2}-j}}{\binom{n+m}{\frac{nm}{2}}}.$$

This is known as the **hypergeometric distribution**. It is equivalent to the following experiment. An urn contains n red balls and m white balls. A sample of size $(n+m)/2$ is drawn at random without replacement. Then $P(N=k)$ is the probability that k of them are red. See, e.g., Pitman [11, p. 125].

For large n, m this is approximately normal with mean $n/2$ and variance $\big(nm/(n+m-1)\big)/4$. This can be used as a basis for a test in the usual fashion. However this test is very inefficient.

To be fair, this is really only a test of the hypotheses

$$H_0\colon \text{median } f_X = \text{median } f_Y, \quad \text{against the alternative} \quad H_1\colon \text{median } f_X \neq \text{median } f_Y.$$

26.4 The Wilcoxon–Mann–Whitney test

A better test, known as the **Wilcoxon rank test** or the **Mann–Whitney test**, is based on the order statistics. (According to William Kruskal [5], this test was independently discovered at least seven times, dating back to Deuchler [2] in 1914.) Array the combined sample from smallest to largest, sort of like this:

$$x_3 < x_7 < y_4 < \cdots y_2 < x_5,$$

and assign them ranks from 1 to $n+m$ starting with the smallest. (In the event of ties, assign each tied spot the average of the ranks.) Let

$$s_i = |\{j : y_j > x_i\}|, \quad r_i = |\{j : x_j > x_i\}|.$$

With no ties, the rank t_i of x_i is just $n+m-s_i-r_i$. The Wilcoxon test statistic is

$$T = \sum_{i=1}^{n} t_i.$$

If the null hypothesis is true, then T has the same distribution as the sum of n numbers drawn at random without replacement from the set $1, \dots, n+m$. The Mann–Whitney test statistic[1] is

$$U = \sum_{i=1}^{n} s_i.$$

The relationship with T is straightforward:

$$T = \sum_{i=1}^{n}(n+m-s_i-r_i), \quad U = \sum_{i=1}^{n} s_i,$$

so

$$T + U = n(n+m) - \sum_{i=1}^{n} r_i,$$

[1] Van der Waerden [13, p. 275] defines the Mann–Whitney statistic by $U = \sum_{i=1}^{n}(m - s_i)$.

but a moment's reflection should tell you that

$$\sum_{i=1}^{n} r_i = (n-1) + (n-2) + \cdots + 1 + 0 = n(n-1)/2.$$

Therefore

$$U + T = n(n+m) - \frac{n(n-1)}{2} = nm + \frac{n(n+1)}{2},$$

so knowing one statistic tells us the other.

It turns out the Mann–Whitney statistic U is more convenient to work with. Let $h(x, y)$ be the indicator of $x < y$. That is

$$h(x,y) = \begin{cases} 1 & \text{if } x < y \\ 0 & \text{otherwise.} \end{cases}$$

Then, if the null hypothesis is true, $P(X < Y) = 1/2$, so

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} h(x_i, y_j), \quad \text{so} \quad \boldsymbol{E}\, U = \sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{E}\, h(x_i, y_j) = nmP(X < Y) = nm/2.$$

A tedious computation along the same lines shows that

$$\boldsymbol{Var} U = \frac{mn(n+m+1)}{12}.$$

Moreover, it can be shown that the standardized U,

$$\frac{U - \frac{mn}{2}}{\sqrt{\frac{mn(n+m+1)}{12}}}$$

is approximately Normal$(0, 1)$ when m and n are both large. [13, p. 277]. (The proof uses the Second Limit Theorem 11.8.1.)

It is possible to get a recursive formula for the exact distribution of U and it can be used when n and m are small. Here is the argument [9]: Let $\varphi(u; m, n)$ denote the probability that $U = u$ under the null hypothesis. There are two ways $U = u$ can occur. Case 1: The largest value is an x, so it contributes nothing to the sum that is U. The probability of this is just $n/(n+m)$. Case 2: The largest value is a y, so that if we were to drop it, each s_i would decrease by 1. This happens with probability $m/(n+m)$. Thus

$$\varphi(u; m, n) = \frac{n}{n+m}\varphi(u, n-1, m) + \frac{m}{n+m}\varphi(u-n, n, m-1).$$

There are some simple to compute boundary cases: $\varphi(u; 0, k) = \varphi(u; k, 0) = 1$ if $u = 0$ and $= 0$ if $u > 0$ and $k \geqslant 1$. And $\varphi(u; n, m) = 0$ if $u < 0$.

Note that U/mn is an unbiased estimate of $P(X < Y)$. Mann and Whitney [9] proposed it as a test of the following hypotheses:

$$H_0 \colon f_X = f_Y \qquad H_1 \colon f_Y \text{ stochastically dominates } f_X,$$

where, as you may recall, Y dominates X if for all t, $P(Y > t) \geqslant P(X > t)$. For such hypotheses, a one-sided test is appropriate, and the null should be rejected for large values of U.

26.5 The Kruskal–Wallis test

The Wilcoxon rank test can be extended to several samples. It is best suited to testing location of distributions.

Start with a continuous density f. Define

$$f_\theta(x) = f(x - \theta).$$

Then f_θ and $f_{\theta'}$ differ only in their **location parameter** θ.

For example, the mean μ of a normal distribution is a location parameter (for σ^2 fixed).

Consider $k \geqslant 2$ independent samples of sizes $n_1, \dots, n_k$ from distributions $f_{\theta_1}, \dots, f_{\theta_k}$.

How can we test the hypothesis

$$H_0 \colon \theta_1 = \cdots = \theta_k$$

against the alternative

$$H_1 \colon \text{not all } \theta_j\text{s are equal.}$$

The idea behind the Kruskal–Wallis test is that if the location parameters are all the same, then the values of from each sample ought to be "evenly distributed" among the set of values. So arrange the $n = n_1 + \cdots + n_k$ values in order from smallest to largest, and assign each its rank in the list (average out ties). Let R_{ij} denote the rank of x_{ij}, the i^{th} observation of the j^{th} group, in the overall list. Define

$$R_{\bullet j} = \sum_{i=1}^{n_j} R_{ij}.$$

Under the null hypothesis, we would expect the average rank $R_{\bullet j}/n_j$ to be about the same for each j. In fact, under H_0 the test statistic

$$B = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_{\bullet j}^2}{n_j} - 3(n+1)$$

is approximately χ^2 with $k-1$ degrees of freedom. The null hypothesis H_0 should be rejected at the α-level of significance if $B > {\chi^2}_{1-\alpha,k-1}$. [8, Theorem 14.4.1]

26.5.1 Example (Case study 14.4.1, pp. 678, [8]) The first Vietnam war draft lottery was held on Dec 1, 1969. The lottery was conducted by putting capsules with birthdays in an urn (actually a plexiglass cylinder), mixing it up, and then drawing them out. The first birthday drawn got rank 1, the next 2, on down to rank 366. Men who were born from 1944 through 1950 were subject to the draft in order of their rank. (Low ranks first.) The last birthday called for service was the 195^{th}.

An unusual pattern emerged. The later months in the in the year had much lower average ranks than the early months. Only five birthdays in December had ranks greater than 195. It turns out the urn had been loaded with January on the bottom, then February, etc., and not very thoroughly mixed.

A Kruskal–Wallis test of the hypothesis that the month averages are equal yielded a $\chi^2(11)$ statistic of 25.95, which has a p-value of 0.006602.

The lotteries conducted in later years, for later birth cohorts were better designed. □

26.6 Testing for trends

26.6.1 Kendall's τ

Suppose we have an ordered sample of random variables $X_1, \dots, X_n$ and we wish to test whether there is a noisy trend. The are a coupl of ways to formalize this. We could, as with Kruskal–Wallis problem assume that

$$X_k \sim f(x + \theta_k),$$

and test the null hypothesis

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_n$$

against the one-sided alternative

$$H_1 : \theta_1 < \theta_2 < \cdots < \theta_n$$

or

$$H_1' : \theta_1 > \theta_2 > \cdots > \theta_n.$$

Or we might have the less specific null hypothesis that

$$H_0 : f_{X_1} = f_{X_2} = \cdots = f_{X_n}$$

versus a vague alternative such as

$$H_1'' : i > j \implies X_i \text{ stochastically dominates } X_j.$$

Either way, given a sample $x_1, \ldots, x_n$, define

$$u_{kj} = \begin{cases} 1 & z_k > z_j \\ 0 & z_k = z_j \\ -1 & z_k < z_j \end{cases}$$

and, for each $k = 1, \ldots, n$ define

$$s_k = u_{k1} + u_{k2} + \cdots + u_{k,k-1}.$$

That is, s_k is the count of the number of predecessors of x_k that are less than x_k minus the number of predecessors that are greater than x_k. If there is an upward trend then we intuitively expect each s_i to be positive. With no trend we expect each s_i to about zero. (If k is odd and there are no ties, then of course, $s_k \neq 0$, but the average ought to be zero.) Finally, set

$$\hat{S} = \sum_{k=1}^{n} s_k = \sum_{j<k} u_{kj}.$$

If $z_1 < z_2 < \cdots < z_n$, then $\hat{S}$ is maximized and

$$\hat{S}_{\max} = \frac{n(n-1)}{2}.$$

If $z_1 > z_2 > \cdots > z_n$, then $\hat{S}$ is minimized and

$$\hat{S}_{\min} = \frac{-n(n-1)}{2}.$$

This suggest the following normalization. Define

$$\hat{\tau} = \frac{\hat{S}}{\frac{n(n-1)}{2}},$$

which is named **Kendall's τ**, or **Kendall's rank correlation coefficient**. Clearly

$$-1 \leqslant \hat{\tau} \leqslant 1.$$

So to test the null hypothesis $H_0 : f_{X_1} = f_{X_2} = \cdots = f_{X_n}$ versus the one-sided alternative $H_1 :$ upward trend, we would reject the null if $\hat{\tau}$ is "close" to one.

How close is close? That of course depends on the distribution of $\hat{\tau}$. According to Breiman [1, p. 278], under the null hypothesis $H_0 : f_{X_1} = f_{X_2} = \cdots = f_{X_n}$, for $n > 10$, the statistic $\hat{\tau}$ is approximately distributed as a Normal distribution with mean zero and variance $\sigma^2 = (2/9)(2n+5)/\big(n(n-1)\big)$. So

$$Z = \frac{\hat{\tau}}{\sqrt{(2n+5)/\big(n(n-1)\big)}} \sim N(0,1).$$

So to test H_0 versus the one-sided alternative of an increasing trend at the α-level of significance, you reject the null if $Z > z_\alpha$, where, as you recall, $\Phi(z_\alpha) = 1 - \alpha$.

The other one-sided alternative or the two-sided re tested in a simlar fashion: reject if $Z < -z_\alpha$ of $|Z| > z_{\alpha/2}$.

26.6.2 Spearman Rank Correlation

Spearman [12] proposed a measure of association based on ranks. It can be used to test for trends: Given an ordered sample $x_1, \ldots, x_n$, compute the correlation between the index k in the sample and the rank t_k of x_k in the set of order statistics. If there is no trend there should be no correlation.

Define

$$\hat{R} = \sum_{k=1}^{n} (k - \bar{k})(t_k - \bar{t}) = \sum_{k=1}^{n} k(t_k - \bar{t}),$$

where $\bar{t} = \bar{k} = (n+1)/2$.

If $x_1 < \cdots < x_n$, then $t_k = k$, so

$$\hat{R}_{\max} = \sum_{k=1}^{n} k(k - \bar{t}) = \sum_{k=1}^{n} k^2 - \bar{t} \sum_{k=1}^{n} k = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4} = \frac{n(n^2-1)}{12}.$$

If $x_1 > \cdots > x_n$, then $t_k = n + 1 - k$, so

$$\hat{R}_{\min} = \sum_{k=1}^{n} k(n + 1 - k - \bar{t}) = -\frac{n(n^2-1)}{12}.$$

So normalizing leads us to **Spearman's ρ** or **Spearman's Rank Correlation Coefficient**,

$$\hat{\rho}_S = \frac{12\hat{R}}{n(n-1)}$$

Clearly $-1 \leqslant \hat{\rho}_S \leqslant 0$. This leads to a test of the form: Reject the null hypothesis of identical distributions if $|\hat{\rho}_S|$ is "too large." For $n \leqslant 30$, the critical values of the test have been tabulated, and you can find them for instance in Breiman [1, Table 9, p. 387], which is based on Olds [10]. For $n \geqslant 30$, the distribution under the null hypothesis is approximately Normal with mean zero, and variance $1/(n-1)$.

Bibliography

[1] L. Breiman. 1973. *Statistics: With a view toward applications*. Boston: Houghton Mifflin Co.

[2] G. Deuchler. 1914. Über die Methoden der Korrelationsrechnuning der Pädagogik und Psychologie. *Zeitschrift für Pädagogische Psychologie und Experimentelle Pädagogik* 15:114–131, 145–159, 229–242.

[3] R. V. Hogg and A. T. Craig. 1978. *Introduction to mathematical statistics*, 4th. ed. New York: Macmillan.

[4] H. Hotelling and M. R. Pabst. 1936. Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics* 7(1):29–43. http://www.jstor.org/stable/2957508

[5] W. H. Kruskal. 1957. Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association* 52(279):356–360. http://www.jstor.org/stable/2280906

[6] W. H. Kruskal and W. A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260):583–621. http://www.jstor.org/stable/2280779

[7] ——— . 1953. Errata: Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 48(264):907–911. http://www.jstor.org/stable/2281082

[8] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[9] H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1):50–60. http://www.jstor.org/stable/2236101.pdf

[10] E. G. Olds. 1938. Distributions of sums of squares of rank differences for small numbers of individuals. *Annals of Mathematical Statistics* 9(2):133–148. DOI: 10.2307/2957608

[11] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[12] C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15(1):72–101. http://www.jstor.org/stable/1412159

[13] B. L. van der Waerden. 1969. *Mathematical statistics*. Number 156 in Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete. New York, Berlin, and Heidelberg: Springer–Verlag. Translated by Virginia Thompson and Ellen Sherman from *Mathematische Statistik*, published by Springer-Verlag in 1965, as volume 87 in the series Grundlerhen der mathematischen Wissenschaften.

[14] F. Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6):80–83. http://www.jstor.org/stable/3001968