Distribution-Free Procedures
(Devore Chapter Fifteen)

MATH-252-01: Probability and Statistics II*

Fall 2016

Contents

Tuesday 29 November 2016

1 Nonparametric Hypothesis Tests

Most of the hypothesis tests we've considered so far have assumed samples drawn from some underlying distribution or distributions, and made statements about the parameters of those distributions. Now we consider a couple of techniques designed to make more general statements with fewer assumptions about the distributions in question.

1.1 The Wilcoxon Rank Sum Test

Note: This test is in Section 15.2 of Devore

Consider the pooled t-test, which we covered in our study of two-sample inference (Chapter 9). It was designed to handle two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ which were assumed to be drawn from normal distributions $N(\mu_1.\sigma^2)$, respectively $N(\mu_2.\sigma^2)$, and make statements about $\mu_1 - \mu_2$. One choice for the null hypothesis was $H_0 : \mu_1 = \mu_2$; given the assumptions of the test, that null hypothesis states that the two samples are drawn from the *same* normal distribution $N(\mu, \sigma^2)$, without any assumption about the value of μ or σ. But what if we want to simply test whether the two samples were drawn from the same distribution, without making any assumptions at all about that distribution? That is the aim of the Wilcoxon rank sum test, also known as the Mann-Whitney U-test.[1]

[1] Mann and Whitney's original paper, *Annals of Mathematical Statistics* **18**, 50 (1947), `https://dx.doi.org/10.1214/aoms/1177730491`, is quite clearly written.

Consider the following data

$$\{x_i\} = 8.56, 5.03, 48.1, 1.31, 4.82 \tag{1.1a}$$

$$\{y_j\} = 15.0, 12.3, 28.0, 13.9 \tag{1.1b}$$

We might like to test whether they come from the same distribution. To construct a test, we should have some idea what the alternative hypothesis is, beyond just "not the same distribution". In the case of the pooled t test, we had three-choices of alternative hypothesis: 1. $\mu_1 > \mu_2$ (one-sided), 2. $\mu_1 < \mu_2$ (one-sided) or 3. $\mu_1 \neq \mu_2$ (two-sided). Similarly, the Mann-Whitney test considers as its alternative hypotheses:

1. the first distribution generally gives larger values than the second (one-sided)
2. the first distribution generally gives smaller values than the second (one-sided)
3. the first distribution either generally gives larger values than the second, or generally gives smaller values (two-sided)

In this case, let's consider alternative hypothesis (2), that the second set of values is "stochastically larger" than the first.

Now, we can't just pair them off and see if $y_1 > x_1$, $y_2 > x_2$, etc, because the sample sizes are not the same, and there's no meaning the order of the samples anyway. But we can sort the full set of $m + n = 9$ data values and see whether we tend to find the ys later in the list than the xs:

rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Data	1.31	4.82	5.03	8.56	12.3	13.9	15.0	28.0	48.1
Set	x	x	x	x	**y**	**y**	**y**	**y**	x

We'll define our test statistic as the sum of all the ranks of the xs:

$$W = 1 + 2 + 3 + 4 + 9 = 19 \tag{1.2}$$

Note that if the samples really were from the same distribution, each of the x ranks would be equally likely to appear anywhere from 1 to 9, so each must have an expectation value of 5, and as a random variable

$$E(W) = 5 + 5 + 5 + 5 + 5 = 25 \qquad \text{if } H_0 \text{ true} \tag{1.3}$$

In general the ranks go from 1 to $m + n$, so the "middle" rank is $(m + n + 1)/2$, and

$$E(W) = \frac{m(m + n + 1)}{2} \qquad \text{if } H_0 \text{ true} \tag{1.4}$$

As an aside, the Mann-Whitney U statistic, which leads to equivalent testing procedures, is defined as

$$U = W - \frac{m(m + 1)}{2} \tag{1.5}$$

It's the the total over the xs, of how many ys each of them is greater than. I.e., the sum W of the ranks will always include a contribution $\frac{m(m+1)}{2}$, because the lowest x will have a rank of at least 1, the second lowest at least 2, etc. The U statistic subtracts off this minimum so, so the lowest possible U score is zero.

Returning to our example, since $W = 19$ is less than $\frac{m(m+n+1)}{2} = 25$, this statistic is on the low side. That means that the xs seem to appear earlier in the list than the ys do, so we are in the direction indicated by the alternative hypothesis, which says the $\{Y_j\}$ distribution are "stochastically larger" than the $\{X_i\}$.

But how unlikely is so low a rank score given the null hypothesis? If the samples really are drawn from the same distribution, then any set of 5 out of the 9 possible ranks is equally likely. There are

$$\binom{9}{5} = \frac{9!}{5!4!} = \frac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2 \times 1} = 126 \tag{1.6}$$

possibilities. If we tabulate the possible sets of ranks corresponding to each statistic value, we find

W	Ranks	Prob
15	12345	1/126
16	12346	1/126
17	12347, 12356	2/126
18	12348, 12357, 12456	3/126
19	12349, 12358, 12367, 12457, 13456	5/126

So there's a $12/126 \approx 9.5\%$ chance of getting a W statistic this low by chance if the null hypothesis is true and the samples are actually drawn from the same distribution, and a one-tailed test would reject H_0 at the 10% level but not the 5% level.

As you can imagine, this can get quite tedious, so as usual there are tabulated values of thresholds for the test. It's not hard to see that the distribution for W is symmetric, so for example, there is the same chance it will have its maximum possible value of $35 = 5+6+7+8+9$ as its minimum value of 15, the same chance that $W = 34$ as $W = 16$, etc. Thus

$$P(W \leq 19) = P(W \geq 31) \tag{1.7}$$

In this case, though, there's another complication, that the table in the book only lists percentiles where $m \leq n$ and in our case ($m = 5$, $n = 4$) that's not true. However, we could switch the roles of the x and y variables and construct a statistic W' which is the sum of the ranks of all the ys. Since the ranks go from 1 to $m+n$, the sum of *all* the ranks must be

$$W + W' = \frac{(m+n)(m+n+1)}{2} \tag{1.8}$$

In our case this is 45, and thus

$$P(W \leq 19) = P(W' \geq 45 - 19) = P(W' \geq 26) \tag{1.9}$$

Now, if we look up the $m = 4$, $n = 5$ section of Table A.14, we'll find that $P(W' \geq 27) \approx .056$ but in fact $W' \geq 26$ is too common to be listed in the table. (We can check that $P(W \leq 18) = 7/126 \approx 0.056$ as advertized in the table.)

Practice Problems

15.11, 15.13

Thursday 1 December 2016

1.2 Normal Approximation

The tables only go up to $m = n = 8$. Beyond that, one can treat the rank-sum statistic as approximately normal and convert it into an approximately standard normal statistic

$$Z = \frac{W - E(W)}{\sqrt{V(W)}} \tag{1.10}$$

As noted above, $E(W) = \frac{m(m+n+1)}{2}$ in general. With a little more work (see Devore for details), we find

$$V(W) = \frac{mn(m+n+1)}{12} \tag{1.11}$$

Note that since the Mann-Whitney U differs from W by a constant,

$$E(U) = E(W) - \frac{m(m+1)}{2} \qquad \text{and} \qquad V(U) = V(W) \tag{1.12}$$

One more modification is necessary if it happens that any of the x and/or y values are exactly equal. First, in the calculation of the rank-sum, all the "tied" values are assigned the average value of the ranks spanned by the tie. For example, if the second,

third, fourth and fifth-smallest values in the combined list are all the same, we assign them all the rank 3.5. This reduces the variance of the statistic, so that

$$V(W) = \frac{mn}{12}\left((m+n+1) - \sum_i \frac{\tau_i^3 - \tau_i}{(m+n)(m+n-1)}\right) \tag{1.13}$$

where τ_i is the number of values included in the ith tie.

1.3 Rank-Sum Test for Nonzero Difference

The null hypothesis of the $\{X_i\}$ and $\{Y_j\}$ being drawn from identical distributions can be generalized to a hypothesis that the two distributions differ only by an offset $\Delta_0 = \mu_1 - \mu_2 = E(X_i) - E(Y_j)$. We can repeat the test with each x_i replaced by Δ_0. For instance, recalling the data from last time

rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Data	1.31	4.82	5.03	8.56	12.3	13.9	15.0	28.0	48.1
Set	x	x	x	x	**y**	**y**	**y**	**y**	x

suppose we have $\Delta_0 = -5$. Then after adding 5 to each x value and sorting the results, we get

rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Data	6.31	9.82	10.03	12.3	13.56	13.9	15.0	28.0	53.1
Set	$x+5$	$x+5$	$x+5$	**y**	$x+5$	**y**	**y**	**y**	$x+5$

Now the sum of x-ranks has become

$$1+2+3+5+9 = 20 \tag{1.14}$$

This is actually less extreme than we had before, and so we would pass the one-sided test at the 10% level. On the other hand, if $\Delta_0 = 21$, we'd have

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
−19.69	−16.18	−15.97	−12.44	12.3	13.9	15.0	27.1	28.0
$x-21$	$x-21$	$x-21$	$x-21$	**y**	**y**	**y**	$x-21$	**y**

and the rank-sum statistic would become

$$1+2+3+4+8 = 18 \tag{1.15}$$

If we recall that $P(W \le 18) = 0.056$ from last time, we see that the hypothesis $\mu_1 - \mu_2 = 21$ is rejected in a one-sided test at e.g., confidence level 6%, which the original null hypothesis $\mu_1 - \mu_2 = 0$ survived.

2 Nonparametric Confidence Intervals

2.1 Rank-Sum Intervals

We can use this property, of different hypotheses for $\mu_1 - \mu_2$ passing or failing the rank-sum test at a given confidence level, to create a distribution-free confidence interval for that quantity.[2] We should work with the two-sided test to get a confidence band rather than an upper or lower bound, so for example the fact that, for $m = 5$ and $n = 4$, $P(W \le 18) = P(W \ge 32) \approx 0.056$ means that a 88.8% confidence-level test will reject $H_0 : \mu_1 - \mu_2 = \Delta_0$ if the resulting W is less than 18 or greater than 32. It will be consistent with the hypothesis if W is between 18 and 32. We can try sliding the x values back and forth with respect to the y values until that happens, but it's more efficient to phrase things a different way.

If we recall the Mann-Whitney U statistic

$$U = W - \frac{m(m+1)}{2} \tag{2.1}$$

[2]The definition can be stated more generally, to allow for distributions without well-defined means, ad the value Δ_0 such that $f_X(x) = f_Y(x - \Delta_0)$.

this is the number of times an $x - \Delta_0$ value appears later in the list than a y value. It can also be described as the number of (i, j) pairs for which $x_i - \Delta_0 > y_j$, i.e., $x_1 - y_j > \Delta_0$. There are mn possible pairs, so the minimum value is 0 and the maximum is mn. In the example in question, W reaches 18 and $U = W - 15$ reaches 3 when Δ_0 goes below the fourth smallest $x_i - y_j$ value. On the other hand, W reaches 32 and U reaches $17 = 20 - 3$ when Δ_0 goes above the 17th smallest $x_i - y_j$ value. So what we really need to do is construct the list (grid) of $x_i - y_j$ values:

	12.3	**13.9**	**15.0**	**28.0**
1.31	−10.99	−12.59	−13.69	−26.69
4.82	−7.48	−9.08	−10.18	−23.18
5.03	−7.27	−8.87	−9.97	−22.97
8.56	−3.74	−5.34	−6.44	−19.44
48.1	35.8	34.2	33.1	20.1

The smallest (most negative) numbers are in the top right: −26.69, −23.18, −22.97, and −19.44, while the largest (most positive) are in the bottom left: 35.8, 34.2, 33.1, and 20.1. So if $\Delta_0 < -19.44$, U is 3, which is at the edge of the 88.8% range. If $\Delta_0 > 20.1$, U is 18, which again is at the edge of the most likely 88.8%. so the 88.8% confidence interval on $\mu_1 - \mu_2$ is from −19.44 to 20.1. Again, the critical values for these thresholds can be found in the back of Devore (although only for m and n at least 5). Note that these are actually critical values on the Mann-Whitney U as it turns out. They are close to mn, and the indices in the ordered list of differences we're looking for are $mn - c + 1$ and c. (In this example $c = 17$ and $mn - c + 1 = 20 - 17 + 1 = 4$.)

Again, if m and n are large, we can use the normal approximation, and find

$$c \approx \frac{mn}{2} + z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}} \tag{2.2}$$

while

$$mn - c + 1 \approx \frac{mn}{2} - z_{\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}} + 1 \tag{2.3}$$

Practice Problems

15.21, 15.35