Statistics 312: Introduction To Mathematical Statistics Lecture 14

Professor Michael R. Kosorok
Department of Statistics

Outline

1. More on the Mann-Whitney-Wilcoxon test
2. The paired t-test
3. The signed rank test

1 More on the Mann-Whitney-Wilcoxon test

The Mann-Whitney-Wilcoxon can be derived using a different approach. Let $X_1, \ldots, X_n$ have distribution F and $Y_1, \ldots, Y_m$ have distribution G.

Consider the following measure of the difference:

$$\pi = \mathrm{P}\left[X < Y\right],$$

where X and Y are independent with distributions F and G.

If the X's are smaller than the Y's, one would expect π to be greater than 1/2. If X's are greater than Y's, π should be smaller than 1/2.

π can be estimated with

$$\hat{\pi} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij},$$

where

$$Z_{ij} = \begin{cases} 1, & \text{if } X_i < Y_j, \\ 0, & \text{otherwise.} \end{cases}$$

The statistic we use is

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} Z_{ij}.$$

We will now show how this compares with R equal to the sum of the ranks of the Y sample. First, define

$$V_{ij} = \begin{cases} 1, & X_{(i)} < Y_{(j)}, \\ 0, & \text{otherwise.} \end{cases},$$

and note that

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij}.$$

$$\begin{aligned}\sum_{j=1}^{m}\sum_{i=1}^{n} V_{ij} = {} & (\text{number of } X\text{'s that are less than } Y_{(1)}) \\ & +(\text{number of } X\text{'s that are less than } Y_{(2)}) \\ & +\cdots+(\text{number of } X\text{'s that are less than } Y_{(m)}).\end{aligned}$$

If the rank of $Y_{(k)}$ in the combined sample is denoted by R_{yk}, then the number of X's less than $Y_{(1)}$ is $R_{y1} - 1$ (why?), and the number of X's less than $Y_{(k)}$ is $R_{yk} - k$ in general (why?).

Thus

$$\begin{aligned} U = \sum_{i=1}^{n}\sum_{j=1}^{m} V_{ij} &= (R_{y1} - 1) + (R_{y2} - 2) \\ & \quad + \cdots + (R_{ym} - m) \\ &= \sum_{j=1}^{m} R_{yj} - \sum_{j=1}^{m} j \\ &= R - \frac{m(m+1)}{2}. \end{aligned}$$

Thus, under the null hypothesis,

$$\begin{aligned} \mathrm{E}[U] &= \mathrm{E}[R] - \frac{m(m+1)}{2} \\ &= \frac{m(m+n+1)}{2} - \frac{m(m+1)}{2} \\ &= \frac{mn}{2}, \end{aligned}$$

and

$$\mathrm{var}[U] = \mathrm{var}[R] = \frac{mn(m+n+1)}{12}.$$

The R function wilcox.test computes U, but the p-values are the same as they would be if R were used instead.

Why use the t-test if it requires stronger assumptions than the MWW test? Because, if the data are truly normal with common variance in both groups, the t-test is more powerful than the MWW test.

Assume that the alternative hypothesis is H_A: $G(x) = F(x-\Delta)$ (the "shift" model). How do we get a confidence interval for the size Δ of the shift?

Compute

$$K = \frac{mn}{2} - z_{1-\alpha/2}\sqrt{\frac{mn(m+n+1)}{12}},$$

and round down to the nearest integer k.

Let $D_{(1)}, \ldots, D_{(mn)}$ be the ordered values of all differences (from smallest to largest) of the form $X_i - Y_j$, for $i = 1 \ldots n$ and $j = 1 \ldots m$.

Then an approximate $100(1-\alpha)\%$ confidence interval for Δ is $[D_{(k)}, D_{(mn-k+1)}]$.

Consider the iron absorption example from lecture 13. Here, $m = n = 18$. Thus, for $\alpha = 0.05$,

$$\begin{aligned} K &= \frac{mn}{2} - z_{1-\alpha/2}\sqrt{\frac{mn(m+n+2)}{12}} \\ &= 18 \times 18/2 - 1.96 \times \sqrt{18 \times 18 \times 37/12} \\ &= 100.05, \end{aligned}$$

which means that $k = 100$ and $mn - k + 1 = 225$.

If we sort all 324 differences of the form $X_i - Y_j$, the confidence interval consists of the 100'th and the 225'th values, which gives the interval $[-3.66, -0.53]$.

The 95% confidence interval for $\mu_x - \mu_y$ based on the two-sample t statistic is $[-3.90, -0.58]$.

2 The paired t-test

Frequently, treatment and control observations are paired rather than independent. For example, treatment may be given to a randomly selected eye while the remaining eye receives the standard treatment, or two different laboratory techniques may be applied to the same samples.

In these settings, the data is of the form (X_i, Y_i), $i = 1 \ldots n$, where the X data involves one treatment while the Y data another. The pairs are independent of each other, but the values within pairs may be quite correlated.

If X_1 and Y_1 are independent, the two-sample methods we have just discussed are applicable. If they are not independent, two-sample methods are not valid and paired methods are needed.

Let $D_i = X_i - Y_i$, $i = 1 \dots n$. Define $\bar{D} = \bar{X} - \bar{Y}$, and note that

$$\mathrm{E}\left[\bar{D}\right] = \mu_x - \mu_y$$

and

$$\begin{aligned}\mathrm{var}\left[\bar{D}\right] &= \frac{1}{n}(\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y) \\ &= \frac{1}{n}\sigma_D^2.\end{aligned}$$

Paired statistics are based on $D_1, \dots, D_n$. Let s_D^2 be the sample variance of the D's and let $\mu_D = \mu_x - \mu_y$. If the D's are normal with mean μ_D and variance σ_D^2, then inference for μ_D is based on the fact that

$$t = \sqrt{n}\frac{\bar{D} - \mu_D}{s_D}$$

has a t distribution with $n - 1$ degrees of freedom.

In the Rice book, $s_{\bar{D}} = s_D/\sqrt{n}$, and thus

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}.$$

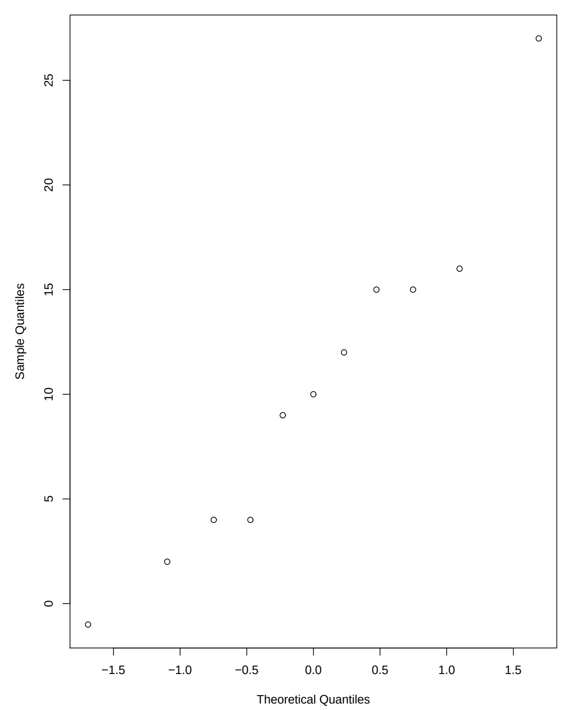
A $100(1-\alpha)$ two-sided confidence interval for μ_D is $\bar{D} \pm s_{\bar{D}} t_{1-\alpha/2}(n-1)$, while a test of $H_0 : \mu_D = 0$ is $t_0 = \sqrt{n}\bar{D}/s_D$, where t_0 has a t distribution with $n-1$ degrees of freedom under H_0.

For example, Levine (1973) studied the effect of cigarette smoking on platelet aggregation (tendency for the blood to clot) by drawing blood from 11 individuals before and after smoking and measured percentage of platelet aggregation following a stimulus:

Before	After	Difference
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	-1
28	43	15

$\bar{D} = 10.27$ and $s_{\bar{D}} = 2.40$. Thus $t_0 = 4.27$, and the p-value is 0.0016. The 95% confidence interval for μ_D is $10.27 \pm 2.40 * 2.23 = [4.91, 15.63]$.

Figure 1: Normal QQ plot for platelet data.



3 The signed rank test

In some settings, the normality assumption may not be appropriate, and a nonparametric approach would be more appealing. One such approach is the signed rank statistic.

The Wilcoxon signed rank statistic is computed as follows:

- Rank the absolute values of the differences $|D_i|$, $i = 1 \ldots n$, and denote these ranks $D_1, \ldots, D_n$.
- Restore the signs of the D_i to the ranks, obtaining the *signed ranks*.
- Calculate W_+, the sum of those ranks that have positive sign.

If the D's are symmetric about zero, we would expect about half of the signs to be negative and half to be positive. If μ_D is larger than zero, we would expect W_+ to be larger.

More specifically, the null hypothesis is H_0, the D's are symmetrically distributed about zero (the mean doesn't even have to exist). The two-sided alternative is that the data are not centered around zero.

Assuming there are no ties, the sign of each rank should be positive with probability 1/2. Thus the null distribution of W_+ is obtained by generating n Bernoulli's $B_1, \ldots, B_n$ (with probability of success 1/2) and then calculating

$$\tilde{W} = \sum_{j=1}^{n} jB_j.$$

The distribution of $\tilde{W}$ computed this way is the same as the null distribution of W_+.

Thus we have

$$\begin{aligned} \mathrm{E}\left[W_+\right] &= \mathrm{E}\left[\sum_{j=1}^{n} jB_j\right] \\ &= \sum_{j=1}^{n} \frac{j}{2} = \frac{n(n+1)}{4}, \end{aligned}$$

and

$$\begin{aligned} \mathrm{var}\left[W_+\right] &= \mathrm{var}\left[\sum_{j=1}^{n} jB_j\right] \\ &= \sum_{j=1}^{n} \frac{j^2}{4} \\ &= \frac{n(n+1)(2n+1)}{24}. \end{aligned}$$

Thus one can obtain p-values using a table, or, when the sample size is large, inference can be based on

$$Z = \frac{W_+ - \mathrm{E}\left[W_+\right]}{\sqrt{\mathrm{var}\left[W_+\right]}}.$$

Values of D_i that are zero are thrown out of the sample and the sample size n is reduced accordingly. Ranks for tied values are replaced by the averages ranks as with the MWW test. The p-values for the statistic assuming no ties is conservative for when there are ties (as with the MWW statistic).

Consider again the platelet example:

Difference	Rank of $\lvert D_i \rvert$
2	2
4	3.5
10	6
12	7
16	10
15	8.5
4	3.5
27	11
9	5
-1	1
15	8.5

Here, $W_{+1} = 65$, $\mathrm{E}\left[W_+\right] = 10 \times 11/4 = 27.5$, $\mathrm{var}\left[W_+\right] = 96.25$, $Z = 3.82$, and $p = 0.0001$ (two-sided).

Here is how the example is analyzed with the signed rank test in R (assuming the differences are in the vector dd):

```
> wilcox.test(dd)

        Wilcoxon signed rank test with
          continuity correction

data:  dd
V = 65, p-value = 0.005056
alternative hypothesis: true mu is not
  equal to 0

Warning message:
Cannot compute exact p-value with ties
  in: wilcox.test.default(dd)
```

Why use the paired t-test when we could use the signed rank test? (There is an important difference in small samples.)