Nonparametric Methods

BIOS 6611

CU Anschutz

Week 6

1. Wilcoxon Rank Sum Test

2. Wilcoxon Signed Rank Test

3. Sign Test

Wilcoxon Rank Sum Test

Wilcoxon Rank Sum Test

The Wilcoxon rank sum test (also known as the Mann-Whitney U test or Wilcoxon-Mann-Whitney test) is for two independent samples of quantitative data.

The nonparametric two independent sample test is "analogous" to the parametric independent sample t-test, but it has nothing to do with comparing means (or medians or even distributions!). **The Wilcoxon rank sum test compares the mean ranks between groups.**

H_0 is that the mean ranks are equal between groups.

The Mann-Whitney U test is computed differently but is completely equivalent to the Wilcoxon rank sum test.

Wilcoxon Rank Sum Test Assumptions

- Independent observations (random sampling)
- The test is based on the P(an observation in sample 1 $>$ an observation in sample 2), H_0: $P(X_1 > X_2) + P(X_1 = X_2) = 0.5$ vs. H_1: $P(X_1 > X_2) + P(X_1 = X_2) \neq 0.5$.
- Does not require normality, even for small n
- For large enough samples ($n_1 \geq 10$ and $n_2 \geq 10$) we can use the normal approximation form of the test; for small n use Table 12 in the Rosner text (or R, SAS, Stata, etc.). Caution should be exercised with tables when there are a lot of ties in the data.
- If we assume the two populations have the same shape (even if shifted, i.e., different medians), then it can be considered a test of medians (or even means). However, this is a strong assumption.

General Procedure

For the Wilcoxon Rank Sum test we

1. Pool the 2 samples
2. Rank the observations (while keeping track of the sample each observation is from). If there are ties assign the average rank (e.g., ties for the 10th and 11th rank result in a value of 10.5 for both)
3. Calculate the test statistic and p-value

If $n_1 \geq 10$ and $n_2 \geq 10$ we can use a normal approximation, otherwise we need to use the tabled critical values which are derived from exact distributions of the sum of the ranks based on permutation theory with ranks of the data measurements used, not the measurements themselves.

The Asymptotic Version

For the asymptotic test (where $n_1 \geq 10$ and $n_2 \geq 10$), let $R_1 =$ sum of the ranks in one sample (choice is arbitrary, although some tables require choosing the smaller of the two sums).

$$E\left[R_1\right] = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$V\left[R_1\right] = \left(\frac{n_1 n_2}{12}\right)\left[n_1 + n_2 + 1 - \frac{\sum_{i=1}^{g}\left(t_i^3 - t_i\right)}{(n_1 + n_2)(n_1 + n_2 - 1)}\right]$$

If there are ties, we need to correct the variance for the ties occurring between samples where $g =$ number of distinct tied values and $t_i =$ number of ties at a specific value (the portion of $V[R_1]$ after the minus sign). Finally, we calculate our Z statistic to use for estimating the p-value:

$$Z = \frac{|R_1 - E[R_1]| - 0.5}{\sqrt{V[R_1]}}$$

The Exact Version

If either sample size is less than 10, a small-sample table of exact significance levels must be used. These are based on enumeration of all possible permutations of the data and the resulting possible rank sums.

Table 12 in the Rosner text gives upper and lower critical values for the rank sum statistic $T = R_1$ for a two-sided test. In general, the results are statistically significant at a particular α-level if $T \leq T_l =$ the lower critical value or $T \geq T_r =$ the upper critical value.

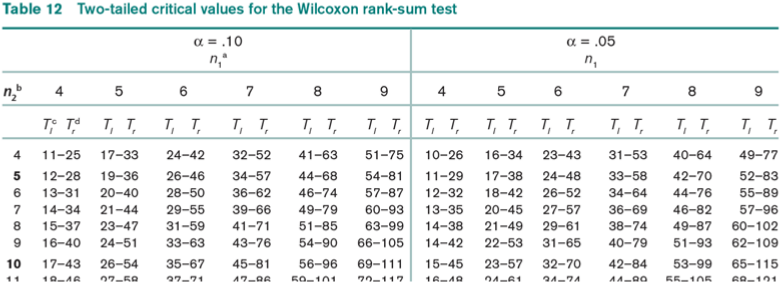


Table 12 Two-tailed critical values for the Wilcoxon rank-sum test

	$\alpha = .10$						$\alpha = .05$					
	n_1[a]						n_1					
n_2[b]	4	5	6	7	8	9	4	5	6	7	8	9
	T_l[c] T_r[d]	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r	T_l T_r
4	11–25	17–33	24–42	32–52	41–63	51–75	10–26	16–34	23–43	31–53	40–64	49–77
5	12–28	19–36	26–46	34–57	44–68	54–81	11–29	17–38	24–48	33–58	42–70	52–83
6	13–31	20–40	28–50	36–62	46–74	57–87	12–32	18–42	26–52	34–64	44–76	55–89
7	14–34	21–44	29–55	39–66	49–79	60–93	13–35	20–45	27–57	36–69	46–82	57–96
8	15–37	23–47	31–59	41–71	51–85	63–99	14–38	21–49	29–61	38–74	49–87	60–102
9	16–40	24–51	33–63	43–76	54–90	66–105	14–42	22–53	31–65	40–79	51–93	62–109
10	17–43	26–54	35–67	45–81	56–96	69–111	15–45	23–57	32–70	42–84	53–99	65–115
11	18–46	27–58	37–71	47–86	59–101	72–117	16–48	24–61	34–74	44–89	55–105	68–121

WRS Example - Ophthalmology

Different genetic types of the disease retinitis pigmentosa (RP) are thought to have different rates of progression with the dominant form of the disease progressing the most slowly, the recessive form of the disease the next most slowly, and the sex-linked form of the disease progressing most quickly.

This hypothesis can be tested by comparing the visual acuity of people ages 10-19 who have different genetic types of RP. Suppose there are 25 people with dominant disease and 30 people with sex-linked disease. The best corrected visual acuities (i.e. with appropriate glasses) in the better eye of these people are presented on the next slide. How can these data be used to test if the **distribution** of visual acuity is different between the two groups?

WRS Example - Ophthalmology

Acuity	Dominant	Sex-Linked	Combined	Rank Range	Avg. Rank
20/20	5	1	6	1-6	3.5
20/25	9	5	14	7-20	13.5
20/30	6	4	10	21-30	25.5
20/40	3	4	7	31-37	34
20/50	2	8	10	38-47	42.5
20/60	0	5	5	48-52	50
20/70	0	2	2	53-54	53.5
20/80	0	1	1	55	55

WRS Example - Manual Calculation

Since $n_1 \geq 10$ and $n_2 \geq 10$ we can use the normal approximation:

$$R_1 = 5(3.5) + 9(13.5) + 6(25.5) + 3(34) + 2(42.5) = 479$$

$$E\left[R_1\right] = \frac{n_1(n_1+n_2+1)}{2} = \frac{25(25+30=1)}{2} = 700$$

$$\begin{aligned} V\left[R_1\right] &= \left(\frac{n_1 n_2}{12}\right)\left[n_1 + n_2 + 1 - \frac{\sum_{i=1}^{g}\left(t_i^3 - t_i\right)}{\left(n_1+n_2\right)\left(n_1+n_2-1\right)}\right] \\ &= \left(\frac{(25)\,30}{12}\right)\left[56 - \frac{(6^3-6)+(14^3-14)+\ldots+(2^3-2)}{55(54)}\right] \\ &= 3386.74 \end{aligned}$$

$$Z = \frac{|R_1 - E[R_1]| - 0.5}{\sqrt{V[R_1]}} = \frac{|479-700]|-0.5}{\sqrt{3386.74}} = 3.7887$$

$$p = 2 \times \left(1 - \Phi\left(3.79\right)\right) = 2 \times \left(1 - 0.9999247\right) = 0.00015$$

WRS Example - with R

```
library(coin) # implement WRS
library(epitools) # convert table to DF

Y <- matrix(nrow=8, byrow=T, c(5,1,9,5,6,4,3,4,2,8,0,5,0,2,0,1),
            dimnames=list(c(20,25,seq(30,80,10)),c("dom","sexlink")))

eye.test <- expand.table(Y) #from epitools package used in previous example
colnames(eye.test) <- c('acuity','grp')
eye.test$acuity <- as.numeric(eye.test$acuity)
eye.test$d_sl <- as.factor(eye.test$grp)
```

WRS Example - with R

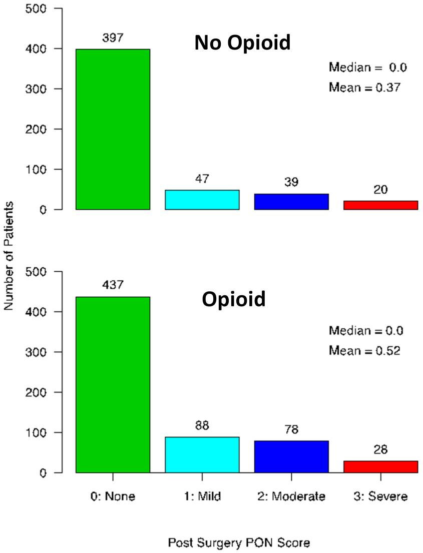
```
# Two-sided exact test
wilcox_test(acuity ~ d_sl, eye.test, distribution='exact')
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  acuity by d_sl (dom, sexlink)
## Z = -3.7975, p-value = 8.496e-05
## alternative hypothesis: true mu is not equal to 0

# Two-sided asymptotic test
wilcox_test(acuity ~ d_sl, eye.test)
##
##  Asymptotic Wilcoxon-Mann-Whitney Test
##
## data:  acuity by d_sl (dom, sexlink)
## Z = -3.7975, p-value = 0.0001461
## alternative hypothesis: true mu is not equal to 0
```

Conclusion: $p = 0.00015 < 0.05$, so we reject H_0. The mean ranks are not the same between groups.

WRS Example - Aromatherapy

- Primary outcome measure: a four-level verbal descriptive scale (VDS) for nausea.
- Groups: Opioids vs. no opioids for postoperative pain.
- The WRS test provides a p-value of 0.001.
- **In other words, the WRS is NOT, by default, a test of medians!**



WRS Notes

- The WRS is ∼95% efficient against the 2-sample t-test for normally distributed samples, is more efficient for many heavy tailed distributions, and is never less than 0.864. *However, these aren't testing the same H_0!*
- The WRS can lack transitivity if comparing multiple groups, i.e., $P(X_1 > X_2) > 0.5$, $P(X_2 > X_3) > 0.5$, but $P(X_3 > X_1) > 0.5$. One remedy to this is to use the *Kruskal-Wallis test* which compares the mean ranks of multiple groups simultaneously.
- The WRS/MW/WMW test is used widely in the literature in situations with small samples and/or heavily skewed/kurtotic data.
- The Fagerland (2012) and Divine et al. (2017) papers in our Canvas Paper Repository discuss some additional WRS details.

Alternative Tests for Medians

Wait a minute, am I even able to compare medians without making the assumption of identical shapes???

If you are truly interested in comparing the medians between samples without assuming identical shapes of the distributions, we have some alternatives to consider:

- Mood's Median Test (https://rcompanion.org/handbook/F_09.html) using the `coin` package in `R`, it is a special case of Pearson's chi-squared test
- Quantile Regression using the `quantreg` package in `R` can evaluate any quantile of interest, including the median, and adjust for other covariates (however, this is beyond the scope of our material this semester) (https://data.library.virginia.edu/getting-started-with-quantile-regression/)
- The Sign Test for one sample contexts where you have a proposed null value for the median

Wilcoxon Signed Rank Test

Wilcoxon Signed Rank Test

Related to the Wilcoxon Rank Sum test is the **Wilcoxon Signed Rank test** for quantitative paired data.

It incorporates the sign and magnitude of the differences in the paired data setting. It is also useful when the outcome variable describes ordering but not necessarily physical distance or difference (unequal magnitude/distance between points – ordinal vs. discrete scale).

e.g. a Likert scale: Patient is $1 =$ much improved, $2 =$ slightly improved, $3 =$ same, $4 =$ slightly worse, $5 =$ much worse

The test is based on the paired differences: $H_0 : P(X_1 + X_2 < 0) = 0.5$. This is not a test of the median difference.

For more details, see Divine, G., Norton, H., Hunt, R., & Dienemann, J. (2013). A Review of Analysis and Sample Size Calculation Considerations for Wilcoxon Tests. *Anesthesia & Analgesia*, 699-710.

Sign Test

Sign Test

The **sign test** is most useful for quantitative for paired data (x,y) it is most useful if it can be expressed as x>y, x=y, or x<y. If we have numeric values or ranks that are of interest, there are methods that have greater power (e.g., t-tests for means, Wilcoxon signed rank test).

For the sign test we define our test statistic as $p = P(X > Y)$ and test $H_0 : p = 0.5$ (i.e., for a random pair of measurements (x_i, y_i), it is equally likely for either to be larger than the other).

Special functions exist in `R`, but we can simply use `binom.test` with the default `p=0.5` for our sign test for paired data.

The sign test can also be used with one-sample to *compare the observed median to some null median value*. We can use either `SIGN.test()` from the `BSDA` package or `SignTest()` from the `DescTools` package.

Sign Test Example - Deer Hind Leg Length

A study was completed to compare deer hind and foreleg length.[1]

Deer	Hind leg (cm)	Foreleg (cm)	Difference (H>F)
1	142	138	+
2	140	136	+
3	144	147	-
4	144	139	+
5	142	143	-
6	146	141	+
7	149	143	+
8	150	145	+
9	142	136	+
10	148	146	+

[1] Zar, Jerold H. (1999), "Chapter 24: More on Dichotomous Variables", *Biostatistical Analysis (Fourth ed.)*, Prentice-Hall, pp. 516–570

Sign Test Example - with R

```r
library(BSDA)
library(coin)
deer.data <- data.frame( hind = c(142,140,144,144,142,146,149,150,142,148),
    fore = c(138,136,147,139,143,141,143,145,136,146) )

#Sign test:
SIGN.test(x=deer.data$hind, y=deer.data$fore)
```

```
##
##  Dependent-samples Sign-Test
##
## data:  deer.data$hind and deer.data$fore
## S = 8, p-value = 0.1094
## alternative hypothesis: true median difference is not equal to 0
## 95 percent confidence interval:
##  -0.02666667  5.67555556
## sample estimates:
## median of x-y
##           4.5
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.8906  2.0000 5.0000
## Interpolated CI       0.9500 -0.0267 5.6756
## Upper Achieved CI     0.9785 -1.0000 6.0000
```

Conclusion: For our sign test, we fail to reject H_0 that $p = 0.5$, therefore we cannot conclude that the hind length and foreleg length in our sample of 10 deer are different.

Sign Test Example - Compare to Binomial Test

```r
#test number H>F against expected proportion of 0.5
binom.test(x=8,n=10,p=0.5)
```

```
##
##  Exact binomial test
##
## data:  8 and 10
## number of successes = 8, number of trials = 10, p-value = 0.1094
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4439045 0.9747893
## sample estimates:
## probability of success
##                    0.8
```

The p-value for our (exact) binomial test matches the p-value for the sign test on the previous slide.

Sign Test Example - Compare to Wilcoxon Signed Rank and Paired t-test

We can also compare our results to the Wilcoxon Signed Rank test or paired t-test, which are more powerful:

```
wilcoxsign_test( hind ~ fore, data=deer.data)
##
##  Asymptotic Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = 2.4047, p-value = 0.01618
## alternative hypothesis: true mu is not equal to 0

t.test( x=deer.data$hind, y=deer.data$fore, paired=T)
##
##  Paired t-test
##
## data:  deer.data$hind and deer.data$fore
## t = 3.4138, df = 9, p-value = 0.007703
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.113248 5.486752
## sample estimates:
## mean of the differences
##                     3.3
```

One-Sample Sign Test Example - LOS

The length of stay for patients undergoing a procedure is observed to be

$$4, 4, 5, 7, 8, 12.5, 14, 14, 15, 18$$

The hospital wishes to compare this to a national median of 14 days. Use the one-sample sign test to evaluate this hypothesis.

One-Sample Sign Test Example - with R

```r
library(BSDA)
los_vec <- c(4,4,5,7,8,12.5,14,14,15,18)
#sign test for one sample to compare to expected median of 14 days
SIGN.test(x=los_vec, m=14)
```

```
##
##  One-sample Sign-Test
##
## data:  los_vec
## s = 2, p-value = 0.2891
## alternative hypothesis: true median is not equal to 14
## 95 percent confidence interval:
##   4.324444 14.675556
## sample estimates:
## median of x
##       10.25
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt  U.E.pt
## Lower Achieved CI     0.8906 5.0000 14.0000
## Interpolated CI       0.9500 4.3244 14.6756
## Upper Achieved CI     0.9785 4.0000 15.0000
```

Conclusion: For our one-sample sign test for medians, we fail to reject H_0 that median=14, therefore we cannot conclude that our sample's median of 10.25 days is significantly different than the national median of 14 days.

Nonparametric Conclusion

- Nonparametric methods are useful when we are concerned about distributional assumptions.
- However, they are frequently misused and misinterpreted. Caution should be taken when using them and consuming them in the literature.
- Many parametric methods are fairly robust to departures from their assumptions.

Frank Wilcoxon cruising along to some rank based comparisons.