12. Nonparametric Methods

Readings: Chihara and Hesterberg: 3.3-3.4
Rosner: 9.1-6

R: expand.table (epitools package), sample function, exactRankTests (package)

Homework: Homework 5 due by 11:59 pm on October 8
Homework 6 due by 11:59 pm on October 22

Overview

A) Permutation tests
B) Wilcoxon rank sum
C) Wilcoxon signed rank (paired quantitative data – see Rosner for details)
D) Sign test (paired quantitative data)

A. **Permutation tests** (also known as randomization tests and exact tests)

Up until now we have mostly considered parametric procedures, i.e. those that, based on the central limit theorem, depend on the *normality of the statistics* being computed.

Nonparametric methods are useful when the assumption of normality does not hold (e.g. small samples, heavily skewed data, ordinal data).

In bootstrap sampling we randomly resample *with* replacement while maintaining the membership of an observation with its group (i.e., for two groups, an observation from Group 1 may appear multiple times in a bootstrap sample for Group 1 [the *with* replacement piece], but will <u>never</u> appear as a member of the Group 2 bootstrap sample).

For permutation testing we will randomly "exchange labels" on data points to derive a p-value. In this case, we will sample *without* replacement from a pooled sample of all data, and each observation will only be represented once in each permutation resample. However, now the group membership (or other features) will vary based on resampling, so our previous observation from Group 1 in our original sample may be in either Group 1 or Group 2 for any given permutation sample.

Two-Sample Permutation Test

Pool the *m* values from Sample 1 and *n* values from Sample 2.

Repeat the following steps

1. Draw a resample of size *m* without replacement to represent Sample 1
2. Use the remaining *n* observations to represent Sample 2
3. Calculate the difference in means (or any other statistic to compare the samples).

until we have "enough" samples

Calculate the p-value as the fraction of times the random statistics exceed or are equal to the original statistic. Multiply by 2 for a two-sided test.

Optionally, plot a histogram of the random statistic values.

Source: Chihara and Hesterberg (pg. 40)

Required assumption: under the null hypothesis, the distribution for the two groups is equal.

The distribution of your difference (or other statistic) across all permutation resamples is the *permutation distribution*. It can be exact (exhaustively calculate all permutations) or approximate (implemented with sampling).

Note! For permutation testing, the distribution doesn't have to be normal. It can be anything, so long as the two populations (groups) have the same distribution under the null hypothesis. Thus, group labels are said to be *exchangeable*. Differences in spread under H_0 can yield misleading results.

The summary statistic for each permutation can be a difference in means, medians, proportions, etc. so the approach can be easily generalized. For example, permutation tests have been described for regression models.

Example: Contingency Tables and Hypothesis Tests

Survey data from 2001 on support for marijuana for medicinal purposes: Does support for medical marijuana depend on age? (H_0: no association between age and favoring use of medical marijuana)

Age group	Response	
	For	Against
18-29 yo	172	52
30-49 yo	313	103
≥50 yo	258	119

We need to transform this summary data into individual level data:

Person	Age group	Response
1	18-29 yo	For
2	18-29 yo	For
. . .	. . .	. . .
1017	≥50 yo	Against

How do we generate a null distribution for this statistic?

- Use a chi-square distribution with (r-1)x(c-1) degrees of freedom (Lecture 10, Section C)
- Permutation (Lecture 7, Section B1 for Fisher's approach)

Permutation Test for Independence of Two Variables

Store the data in a table with one row per observation and one column per variable.
Calculate a test statistic for the original data. Normally large values of the test statistic suggest dependence.

Repeat the following steps

1. Randomly permute the rows in one of the columns (i.e., hold the other fixed).
2. Calculate the test statistic for the permuted data.

until we have "enough" samples

Calculate the p-value as the fraction of times the random statistics exceed or are equal to the original statistic.
Optionally, plot a histogram of the resampled statistic values.

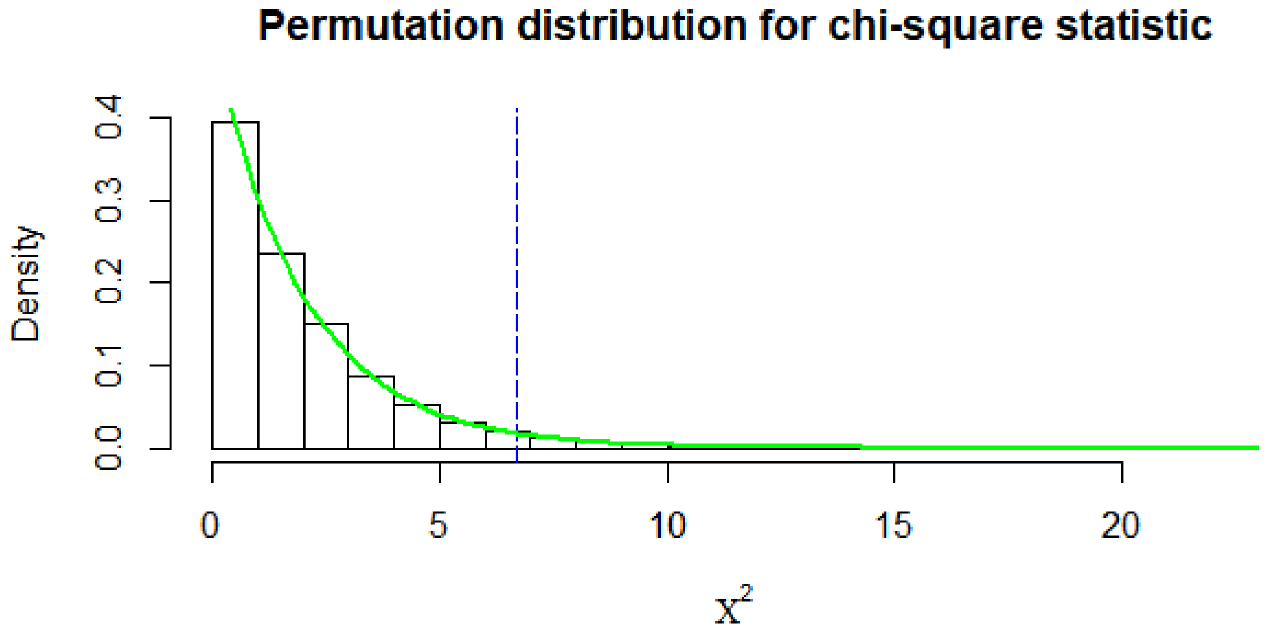
Source: Chihara and Hesterberg (pg. 55)

R code

```
X <- matrix(nrow=3, byrow=T, c(172,52,313,103,258,119),
+               dimnames=list(c("18-29","30-49",">50"),c("for","against")))

# Obtain row proportions in the table
prop.table(X, margin=1)
            for   against
18-29 0.7678571 0.2321429
30-49 0.7524038 0.2475962
>50   0.6843501 0.3156499

library(epitools)

dat <- expand.table(X) # transform data from 3x2 table to 1 obs per row

# define a function to do the chisquare test

chisq <- function(Obs){
+   #Obs is the observed contingency table
+   Expected <- outer(rowSums(Obs),colSums(Obs))/sum(Obs)
+   sum((Obs-Expected)^2/Expected)
}

# do a permutation test
# first compute observed statistic

agegrp <- dat[,1]
response <- dat[,2]

observed <- chisq(table(agegrp, response)) #2 degrees of freedom: (r-1)(c-1)
observed
[1] 6.681429
```

R code cont.

```
B <- 10^5-1  #set number of times to repeat this process result
result <- numeric(B) # space to save the random differences
set.seed(515)

for(i in 1:B){
+   agegrp.permuted <- sample(agegrp) #sample group labels without replacement
+   perm.table <- table(agegrp.permuted, response)
+   result[i] <- chisq(perm.table)
}

##Plot
hist(result, freq=FALSE, xlab = expression(Chi^2), main="Permutation distribut
ion for chi-square statistic")
abline(v = observed, col = "blue", lty=5)
curve(dchisq(x, 2), add=TRUE, col="green", lwd=2)
```



R code cont.

```
#Compute p-value from the permutation distribution
(sum(result >= observed)+1)/(B + 1)  #P-value
[1] 0.03591

# compute p-value from chi-square distribution
1-pchisq(observed, df=2)
[1] 0.03541165
```

Conclusion:

Practical notes on permutation testing:

- Adding 1 to the numerator and denominator corresponds to using the observed sample as part of the null distribution (i.e., `(sum(result >= observed)+1)/(B + 1) #P-value`)
- Various statistics can be used: e.g. mean, median, proportions, etc.
- For more precision, a larger number of permutations should be used.
- Sampling (entire) permutations without replacement is most appropriate but with replacement is acceptable and faster (i.e., we don't necessarily need to check that a given permutation sample is unique amongst all our previous permutation samples).
- Any strictly increasing function of the statistic will yield the same p-value.
- For two-sided alternative hypotheses, conduct both one-sided tests and multiply the smaller p-value by 2. (In our case the chi-squared test takes the squared part into account, so we don't need to multiply by 2 even though it is a two-sided test.)
- Normality of underlying distributions not assumed. Robust to skewness and imbalance as long as the underlying distributions are equal under the null hypothesis.
- No random sampling assumption is required. If random sampling is not assumed, inference to a population can't then be made but a conclusion about the sample can be drawn. However, treatment (exposure) assignments are assumed to be random.

B. Wilcoxon rank sum test (*aka* Mann Whitney U test; for two independent samples, 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.

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.). When using the tables caution should be exercised 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. https://www.graphpad.com/guides/prism/7/statistics/index.htm?stat_nonparametric_tests_dont_compa.htm

Procedure:

1. Pool the 2 samples

2. Rank the observations (while keeping track of the sample each observation is from)
   a. If there are ties in the ranks (i.e., multiple observations have the same value), compute the average rank (e.g., ties for the 10th and 11th rank would result in a value of 10.5 for both)

3. Calculate the test statistic, 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.

For the asymptotic test:

R_1 = sum of the ranks in one sample (choice is arbitrary---some tables require choosing the smaller of the two sums)

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

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

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:

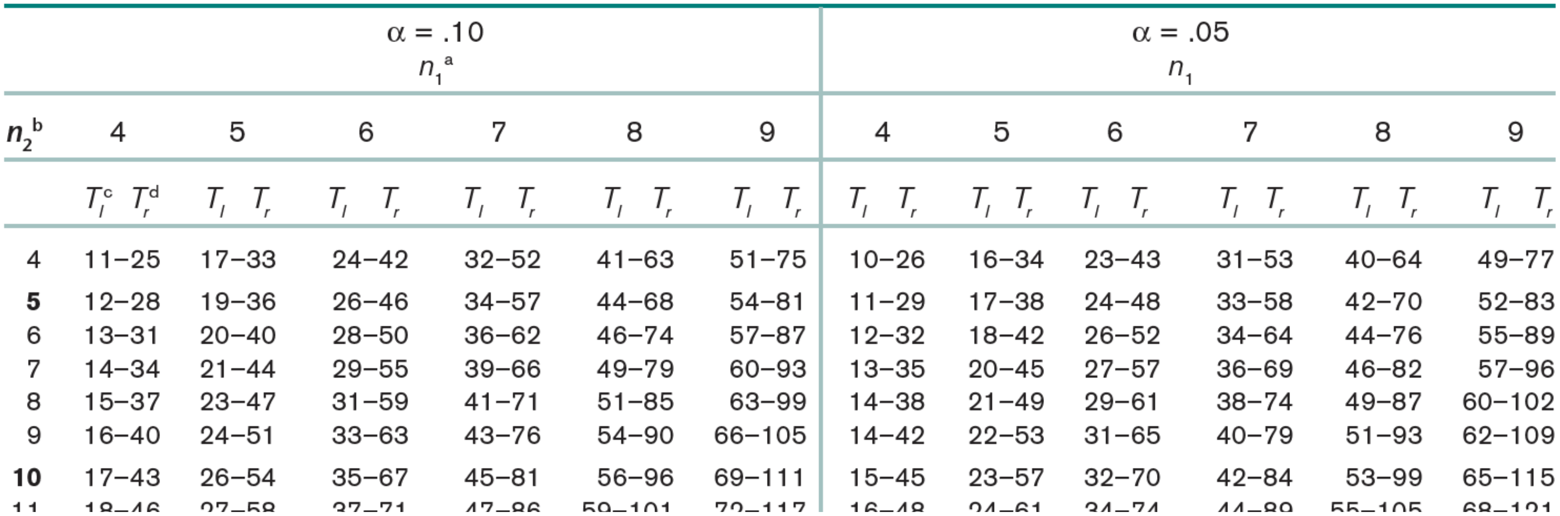
$$Z = \frac{|R_1 - E[R_1]| - 0.5}{\sqrt{V[R_1]}}$$

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 = R₁** 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.

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

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



	$\alpha = .10$ n_1[a]						$\alpha = .05$ 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

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 below. How can these data be used to test if the **distribution** of visual acuity is different between the two groups?

Visual Acuity	Dominant	Sex-linked	Combined Sample	Range of Ranks	Average 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
	25	30	55		

Manual Calculation:

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

R_1 = sum of the ranks in one sample = 5(3.5) + 9(13.5) + 6(25.5) + 3(34) + 2(42.5) = 479

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

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

$$= \left(\frac{(25)30}{12}\right)\left[56 - \frac{(6^3 - 6) + (14^3 - 14) + \cdots + (2^3 - 2)}{55(54)}\right] = 3386.74$$

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

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

R code

```
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)
colnames(eye.test) <- c('acuity','grp')
eye.test$acuity <- as.numeric(eye.test$acuity)
eye.test$d_sl <- as.numeric(eye.test$grp)

library(exactRankTests)

# Two-sided exact test
wilcox.exact(acuity~d_sl,eye.test)

     Exact Wilcoxon rank sum test

data:  acuity by d_sl
W = 154, p-value = 8.496e-05 # W is the Mann-Whitney U statistic – see below
alternative hypothesis: true mu is not equal to 0 # mu is the location shift
parameter for the two distributions, not really the best statement, we're comp
aring the distributions of the two samples/groups NOT mu=mean or median

# Two-sided asymptotic test
wilcox.exact(acuity~d_sl,eye.test ,exact=F)

     Asymptotic Wilcoxon rank sum test

data:  acuity by d_sl
W = 154, p-value = 0.0001461
alternative hypothesis: true mu is not equal to 0
```

(Note: R calculates W as $R_1 - n_1(n_1+1)/2$.) Conclusion:

Alternative way of computing the test

- The Mann-Whitney U test. Test statistic is a U statistic: the number of times an observation from distribution 2 is less than an observation from distribution 1. Here, "less than" means "worse acuity".

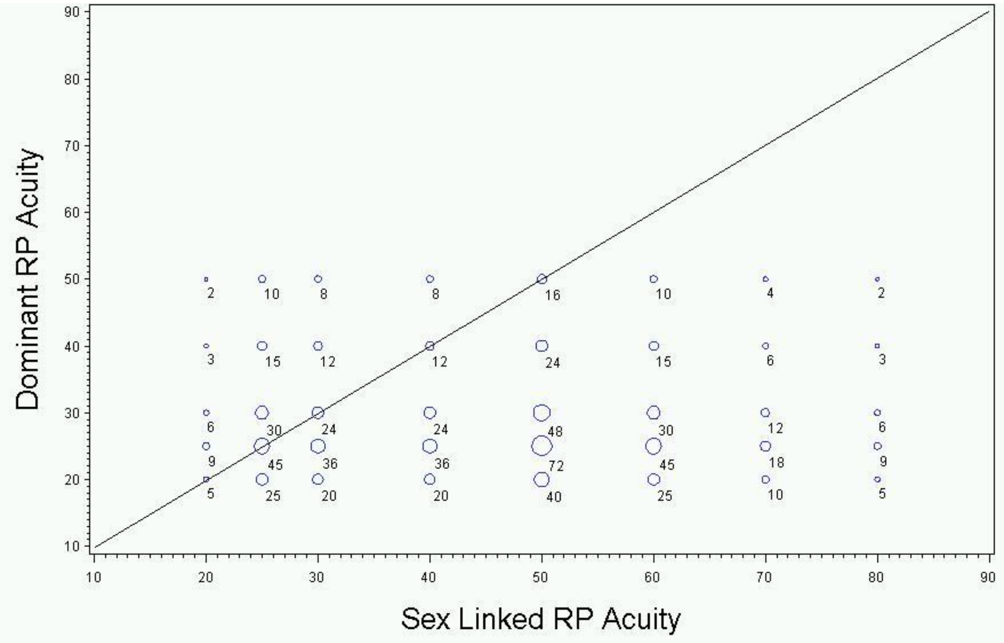
- The test is based on the P(an observation in sample 1 > an observation in sample 2) – H_0: $[P(X_1 > X_2) + P(X_1 = X_2)] = 0.5$ vs. H_1: $[P(X_1 > X_2) + P(X_1 = X_2)] \neq 0.5$.

- $\frac{U}{mn} = p'' = P(X_1 > X_2)$

- In this case *mn* = 25*30=750 – the number of pairs where one member is in the dominant and one is in the sex linked group, with acuity for dominant, X_2, and X_1 for sex linked:
  - U = 2+10+8+8+3+15+12+6+30+9 + [(5+45+24+12+16)/2] = 103 + 102/2 [these are the pairs in which the acuity is tied; they receive a weight of ½]
  - U = 103+51 = 154
  - p" = P(Sex Linked RP VA > Dominant RP VA) = 154/750 = 0.187

Visualization of the Wilcoxon-Mann-Whitney test

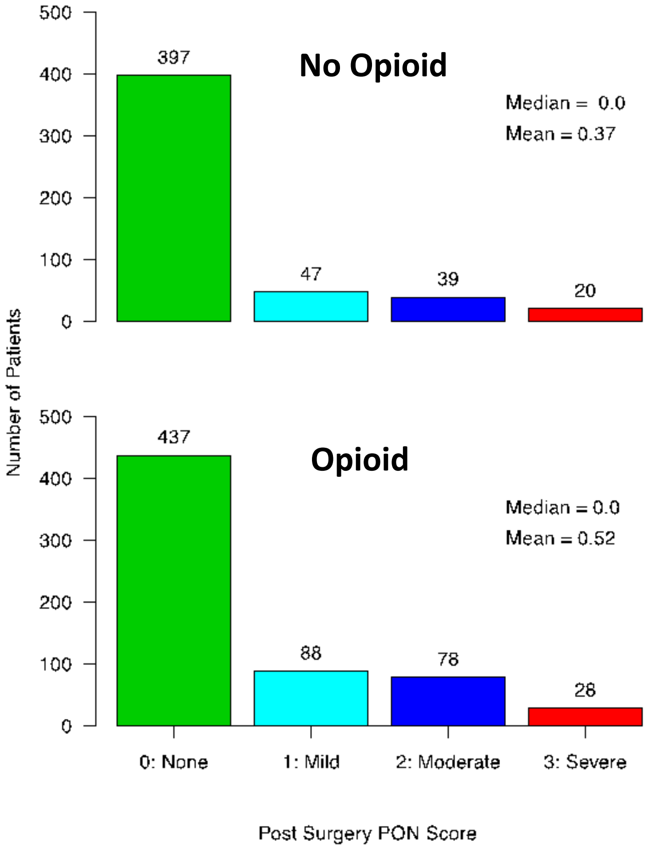
H_0: P(Dominant RP VA < Sex linked RP VA) = 0.5, i.e. the proportion of bubble *areas* below the identity line = 0.5, where each bubble represents the number of Sex Linked/Dominant pairs with that combination of VA.



Be aware that several authors (including Rosner) mistakenly refer to the WMW test as a **test of the medians**! ☹

Counterexample: Equal Medians, but a significant difference using WMW test - Aromatherapy Study

- Primary outcome measure: a four level verbal descriptive scale (VDS) for nausea (0: none; 1: mild; 2: moderate; 3: severe)
- Opioid medication for reduction of post-operative pain
- WMW test provides a p-value of 0.001.



Notes

- Efficiency
  - WMW is ~95% efficient against the 2-sample t-test for normally distributed samples;
  - more efficient than 2-sample t-test for many heavy tailed distributions;
  - compared to the t-test the efficiency of the rank sum test is never less than 0.864.

- The Rank Sum Test has a very unintuitive feature: it can lack transitivity with multiple groups! This means that it's possible to see the following contradictory results, as an example – $P(X_1 > X_2) > 0.5$, $P(X_2 > X_3) > 0.5$, but $P(X_3 > X_1) > 0.5$. For more on this, see http://www.emersonstatistics.com/courses/formal/b517_2010/b514hw6key.pdf, section 8.

- A remedy for this is to use a test that compares the multiple groups all at once instead of as pairwise tests. It's called the Kruskal-Wallis test and it's analogous to one-way analysis of variance (ANOVA) which we will start talking about in a few lectures.

- Nevertheless, you will see the WMW test used widely in the literature in situations where there are small samples and/or heavily skewed/kurtotic data.

- Two relevant papers on the WMW test are posted in the Canvas Paper Repository:
  - Fagerland, 2012
  - Divine et al., 2017

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 you 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 (section D) for one sample contexts where you have a proposed null value for the median

C. Wilcoxon Signed Rank Test (quantitative paired data)

A related test, the Wilcoxon Signed Rank test 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: $[P(X_1 + X_2 < 0)] = 0.5$

Most texts treat this as a test of the median difference which is not correct, but the quantity in H_0 above is not as interpretable as $[P(X_1 > X_2) + P(X_1 = X_2)] = 0.5$ is for the WMW test.

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.

D. Sign Test (quantitative paired data or test to median)

The sign test is most useful for paired observations (x,y) that are expressed as x>y, x=y, or x<y. If we have numeric values or ranks that are of interest, then there are methods that are statistically more power (e.g., t-tests for means or the Wilcoxon signed rank test).

For the sign test we define our test statistic as $p = \Pr\,(X > Y)$ and test $H_0: p = 0.50$ (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 the binom.test with the default p=0.5 for our sign test for paired data.

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

Example (matched pairs two-sided test) comparing deer hind leg length and foreleg length:

Deer	Hind leg length (cm)	Foreleg length (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	+

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

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.

Notice our WSR and paired t-test are significant and may be more appropriate given the numeric values.

R code

```
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) )

binom.test(x=8,n=10,p=0.5) #test number H>F against expe
cted proportion of 0.5 with SIGN TEST

        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

wilcox.exact(x=deer.data$hind,y=deer.data$fore,paired=T)
        Exact Wilcoxon signed rank test

data:  deer.data$hind and deer.data$fore
V = 51, p-value = 0.01172
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 diff in means is not equal to 0
95 percent confidence interval:
 1.113248 5.486752
sample estimates:
mean of the differences
                    3.3
```

Example (one-sample median): 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.

R code

```
los_vec <- c(4,4,5,7,8,12.5,14,14,15,18)

SIGN.test(x=los_vec, m=14) #sign test for one
sample to compare to expected median of 14 days

        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
```

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.

***Frank Wilcoxon* (1882 -1965)**

Frank Wilcoxon was born in Glengarriffe Castle, near Cork, Ireland, to wealthy American parents. He was soon brought to the United States where he attended Pennsylvania Military College, Rutgers, and Cornell. After receiving his doctorate as a physical chemist, Wilcoxon joined the Boyce Thompson Institute for Plant Research and began to study the use of copper compounds as fungicides. While doing so, he became part of a group, along with W. J. Youden (Biography 5.1), that studied the newly published *Statistical Methods for Research Workers* by Ronald A. Fisher (Biography 13.1). Through these achievements (and either in spite or because of his lifelong preoccupation with biochemistry, plant pathology, and entomology), he became a significant member of that small group of twentieth-century pioneers who developed new statistical methodology. In a now famous 1945 paper, he presented the *rank-sum test* and the *signed-rank test* now named after him. The basic idea of replacing actual sample data by their ranks, which seems so utterly simple in retrospect, proved to be inspirational to the further development of the entire field of nonparametric statistics. The elegant simplicity of these tests led to their widespread adoption and the fact that Wilcoxon in 1945 was not even aware of all the advantages of his new methods does not dim the luster of his contribution.

These advantages, not all of them discussed in text Chapter 21, include the ease and rapidity of calculation, the availability of exact significance levels without the restrictive normality assumption, the relative insensitivity to outlying sample observations, the invariance under certain monotonic transformations of the data, the applicability to situations where the data are ordinal, the excellent power properties for wide classes of alternative distributions, and the availability of distribution-free confidence intervals for the location parameters of interest. In addition, Wilcoxon contributed mightily to other aspects of statistics, in particular biological assay methods and sequential analysis (discussed in text Chapter 23). *Source:* Adapted from *International Encyclopedia of Statistics*, vol. 2 (New York: The Free Press, 1978), pp.1245-1250.