Contents

2 Two-Sample Methods

Data. Suppose we observe two independent random samples:

- $X_1, \cdots, X_m$ from distribution $F(x)$ (treatment 1)
- $Y_1, \cdots, Y_n$ from distribution $G(y)$ (treatment 2)
- both $F(\cdot)$ and $G(\cdot)$ are continuous distributions

Problem of Interest: make inference about the difference between two location parameters of F and G.

- Null hypothesis $H_0 : F(x) = G(x)$
- Alternative hypothesis
  - $H_a : F(x) < G(x)$: observations from treatment 1 tend to be **larger** than those for treatment 2
  - $H_a : F(x) > G(x)$: observations from treatment 1 tend to be **smaller** than those for treatment 2

- $H_a : F(x) \neq G(x)$: two distributions differ (either in mean/median or variance or other aspects)

Location-shift Model: we assume that the two distributions differ only with respect to the location parameter, if they differ at all. Mathematically, we express this as

$$F(x) = G(x - \Delta). \tag{2.1}$$

That is, X_i and $Y_i + \Delta$ have the same distribution. Here Δ is the difference of the centers of X and Y, i.e. the treatment effect in comparison of two treatment groups.

The assumption (2.1) is equivalent to

$$P(X \leq x) = P(Y \leq x - \Delta) \text{ for any } x.$$

- $\Delta = 0$: no location shift between two populations
- $\Delta > 0$: X tends to be larger than Y, the center of X is Δ larger than that of Y

- $\Delta < 0$: X tends to be smaller than Y
- $\Delta \neq 0$: X and Y differ in mean/median

In other words, we can write (2.1) as:

$$X_i = \Delta + Y_i, \quad Y_i \ i.i.d \ \sim G(\cdot).$$

Goals:

- Test $H_0 : \Delta = 0$ versus
$$\begin{cases} H_a : \Delta > 0 & \text{upper-tailed test (X is shifted to the right);} \\ H_a : \Delta < 0 & \text{lower-tailed test (X is shifted to the left);} \\ H_a : \Delta \neq 0 & \text{two-tailed test.} \end{cases}$$
- Point estimation of Δ
- Confidence interval of Δ

For example, consider the two populations in the following figure. The red (dashed) population generally has larger values (mean 5) than the black (solid) population (mean 1). So for the same x, $F(x) \geq G(x)$.

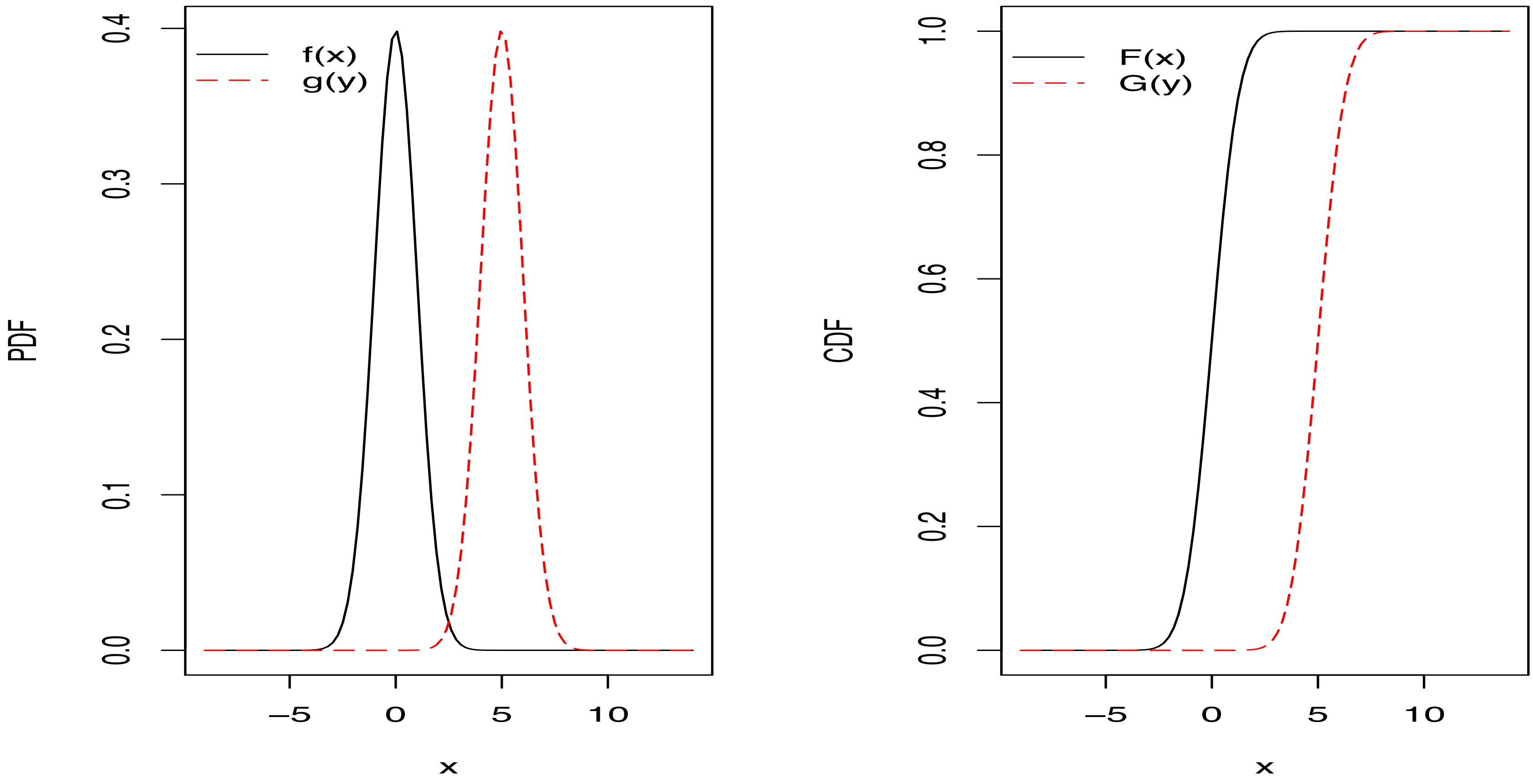

Figure 1: Two distributions with a location shift $\Delta = -4$.

2.1 Classic Method

- Assumptions: $X_1, \cdots, X_m$ *i.i.d.* $\sim N(\mu_X, \sigma^2)$ independent of $Y_1, \cdots, Y_n$, *i.i.d.* $N(\mu_Y, \sigma^2)$, σ unknown
- Parameter of interest: $\Delta = \mu_X - \mu_Y$
- Two-sample t-test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

  where $S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$, S_X^2 and S_Y^2 are the sample variances of X and Y groups, respectively. The S_p^2 is an estimator of the common variance (σ^2) of X and Y populations.

- Under H_0, $T \sim t_{m+n-2}$.
- **Upper-tailed test**: reject H_0 when $T > t_{\alpha, m+n-2}$
  p-value $= P(T_{m+n-2} > t_{obs})$

where $t_{\alpha,m+n-2}$ is the αth percentage point of t_{m+n-2} distribution, T_{m+n-2} is a r.v. from the t distribution with $m+n-2$ d.f.

- **Lower-tailed test**: reject H_0 when $T < -t_{\alpha,m+n-2}$
  p-value $= P(T_{m+n-2} < t_{obs})$
- **Two-tailed test**: reject H_0 when $|T| > t_{\alpha/2,m+n-2}$
  p-value $= 2P(T_{m+n-2} > |t_{obs}|)$
- Estimator $\widehat{\Delta} = \bar{X} - \bar{Y}$
- $(1-\alpha)$ CI for Δ:

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2,m+n-2} S_p \sqrt{1/m + 1/n}$$

For example, a random sample of test scores with two instruction methods:

Method	Score			Mean	Median
New X_i	37	55	57	49.7	55
Traditional Y_i	23	31	70	41.3	31

Test if the mean score of the new method is significantly larger than that of the traditional method.

R code:

```
x=c(37, 55, 57); y=c(23, 31, 70)
# manual calculation
num = mean(x)-mean(y)
m=length(x); n=length(y)
Sp2 = ((m-1)*var(x)+(n-1)*var(y))/(m+n-2)
denom = sqrt(Sp2) * sqrt(1/m+1/n)
tobs = num/denom
(pval = 1-pt(tobs, m+n-2))
alpha=0.05
(critical.val = qt(1-alpha, m+n-2))
##95% confidence interval
c(num-qt(1-0.025,m+n-2)*denom, num+qt(1-0.025,m+n-2)*denom)
# or use the existing function
t.test(x, y, var.equal=T, alternative="greater")#one-sided confi. bound
t.test(x, y, var.equal=T, alternative="two.sided")
##together with 95% two-sided confidence interval
```

2.2 A Two-sample Permutation Test

2.2.1 Permutation test

Example 2.2.1 *For example, a random sample of test scores with two instruction methods:*

Method	Score			Mean	Median
New X_i	37	55	57	49.7	55
Traditional Y_i	23	31	70	41.3	31

- The mean and median of the new method are larger than those of the traditional method. But are the differences significant?
- Test $H_0 : \mu_X = \mu_Y$ versus $H_a : \mu_X > \mu_Y$ or equivalently

$$H_0 : \Delta = 0 \text{ versus } H_a : \Delta > 0, \quad \Delta = \mu_X - \mu_Y .$$

- Two candidate estimators (test statistics)

- $\widehat{\Delta}_1 = \bar{X} - \bar{Y} = 8.3$
- $\widehat{\Delta}_2 = median(X) - median(Y) = 21$ (more robust to outliers)

- How to determine the p-values without distributional assumptions of X_i and Y_i?

Intuition of permutation test:

- If no difference between TM (traditional method) and NM (new method), partitioning of the 6 scores into two groups of size 3 (i.e. any shuffling of the 6 scores) will be equally likely.
- Total $J = \binom{6}{3} = 20$ partitions, each giving one test statistic value.

j	Permuted NM			Permuted TM			$\widehat{\Delta}^*_{1j}$(mean diff)	$\widehat{\Delta}^*_{2j}$(med. diff)
1	37	55	57	23	31	70	8.3	24.0
2	37	55	23	57	31	70	-14.3	-20.0
3	37	55	31	57	23	70	-9.0	-20.0
4	37	55	70	57	23	31	17.0	24.0
5	37	57	23	55	31	70	-13.0	-18.0
6	37	57	31	55	23	70	-7.7	-18.0
7	37	57	70	55	23	31	18.3	26.0
8	37	23	31	55	57	70	-30.3	-26.0
9	37	23	70	55	57	31	-4.3	-18.0
10	37	31	70	55	57	23	1.0	-18.0
11	55	57	23	37	31	70	-1.0	18.0
12	55	57	31	37	23	70	4.3	18.0
13	55	57	70	37	23	31	30.3	26.0
14	55	23	31	37	57	70	-18.3	-26.0
15	55	23	70	37	57	31	7.7	18.0
16	55	31	70	37	57	23	13.0	18.0
17	57	23	31	37	55	70	-17.0	-24.0
18	57	23	70	37	55	31	9.0	20.0
19	57	31	70	37	55	23	14.3	20.0
20	23	31	70	37	55	57	-8.3	-24.0

- The $\{\widehat{\Delta}^*_{1j}\}$ (mean differences based on the permutation samples) provide a **permutation distribution** for the difference in means $\widehat{\Delta}_1$
- The $\{\widehat{\Delta}^*_{2j}\}$ provide a permutation distribution for $\widehat{\Delta}_2$.

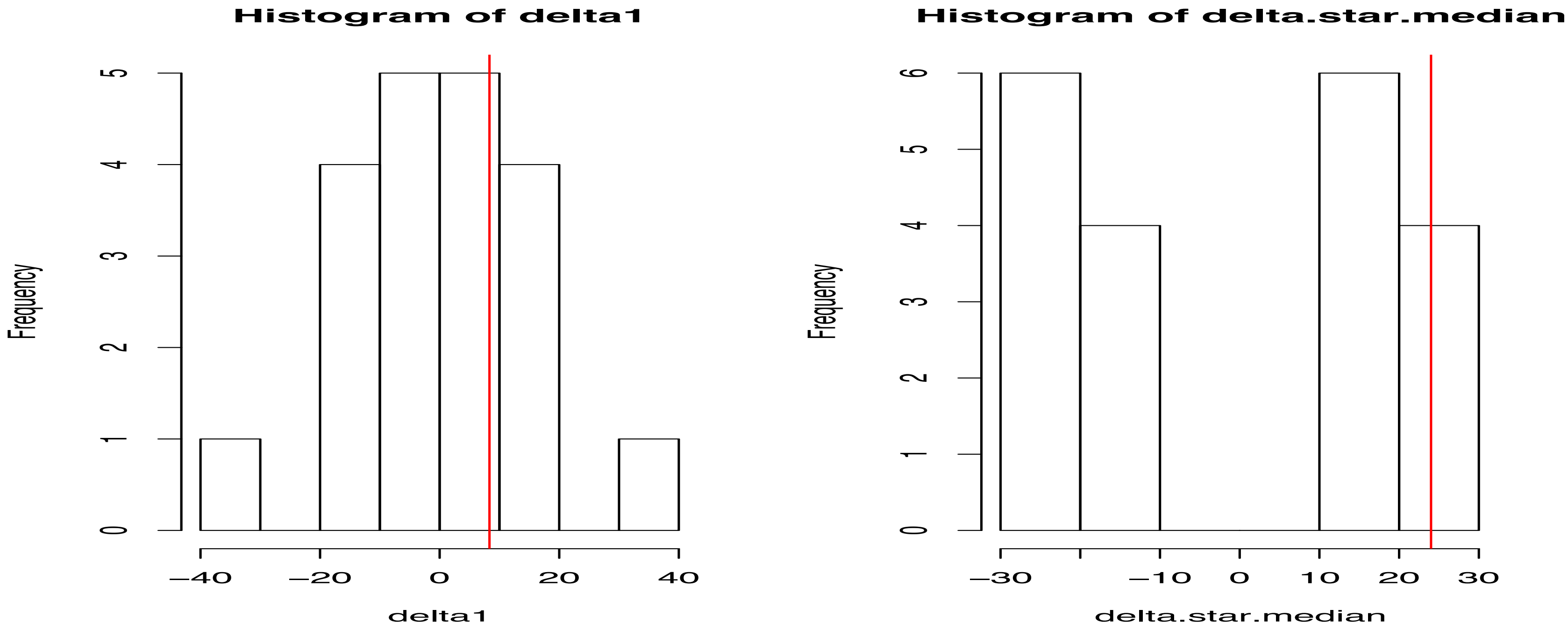

Figure 2: Permutation distributions.

- The permutation distribution is an appropriate reference distribution for determining the p-value for the test.
- Recall p-value is the probability that the test statistic is more extreme than the observed one under the null hypothesis.
- Permutation p-value can be calculated as
  - two-tailed test: p-value $= \#\{|\widehat{\Delta}_j^*| \geq |\widehat{\Delta}|\}/J$
  - upper-tailed test: p-value $= \#\{\widehat{\Delta}_j^* \geq \widehat{\Delta}\}/J$
  - lower-tailed test: p-value $= \#\{\widehat{\Delta}_j^* \leq \widehat{\Delta}\}/J$
- For this data, p-value=7/20=0.35 (mean), =4/20=0.2 (median)

2.2.2 Steps for a two-sample permutation test

Suppose we have m observations from group 1 and n observations from group 2. The value of the statistic of interest (e.g. difference in means) is D_{obs}.

- Permute the $m+n$ observations so that m are assigned to group 1 and n are assigned to group 2. In R, function combinations() in package “gtools” can generate all possible partitions.
- Calculate the statistic of interest D for all $\binom{m+n}{m} = \frac{N!}{m!n!}$ possible permutations, $N = m+n$.
- Calculate the p-value accordingly.
  - two-tailed test: p-value $= \frac{\#|D|'s \geq |D_{obs}|}{\binom{m+n}{m}}$
  - upper-tailed test: p-value $= \frac{\#D's \geq D_{obs}}{\binom{m+n}{m}}$
  - lower-tailed test: p-value $= \frac{\#D's \leq D_{obs}}{\binom{m+n}{m}}$

R code for analyzing the previous test data set using permutation based on the difference of means:

```
idx = combinations(n=6, r=3)
x=c(37,55,57)
y=c(23,31,70)
xy = c(x,y) # the combined data set
permut = NULL # the permuted data set (a 20*6 matrix)
for(i in 1:20){

    permut = rbind(permut, c(xy[idx[i,]], xy[-idx[i,]]))

    }
permut.x = permut[, 1:3] # the permuted X matrix (20*3)
permut.y = permut[, 4:6] # the permuted Y matrix (20*3)

delta1 = apply(permut.x, 1, mean) - apply(permut.y, 1, mean)
delta2 = apply(permut.x, 1, median) - apply(permut.y, 1, median)

delta1.obs = mean(x)-mean(y)
delta2.obs = median(x) - median(y)

#pvalue for permutation of sample mean
pval1.upper = mean(delta1 >= delta1.obs)  #upper-tailed
pval1.2sided = mean(abs(delta1) >= abs(delta1.obs))  #two-tailed
```

```
#pvalue for permutation of sample median
pval2.upper = mean(delta2 >= delta2.obs)  #upper-tailed
pval2.2sided = mean(abs(delta2) >= abs(delta2.obs))  #two-tailed
```

2.2.3 The choice of statistic in the permutation test

Different statistics can be considered in the two-sample permutation:

- **Difference of means** $\bar{X}-\bar{Y}$: commonly used, but sensitive to unusually large or small values (outliers)
- **The sum of observations from one group**. e.g. Let T_Y and T_X be the sum within group Y and X, and let $T = T_X + T_Y$. Then

  $$\bar{X}-\bar{Y} = \frac{T_X}{m} - \frac{T_Y}{n} = \frac{T_X}{m} - \frac{T - T_X}{n} = T_X\left(\frac{1}{m}+\frac{1}{n}\right) - \frac{T}{n}.$$

  Since T is the same for all permutations, the test statistics T_X, T_Y and $\bar{X}-\bar{Y}$ are equivalent and lead to the same permutation p-value.
- **Two-sample t-test statistic**: similar to difference of means, but with different denominator (standard error) for each

permutation sample.

- **Difference of medians**: robust to outliers
- **Difference of trimmed means**: can reduce the influence of extreme observations; requires symmetry of the distribution. E.g. 10% trimmed mean is the mean of all observations after deleting the top 5% largest and the bottom 5% smallest values.

Some Recommendations. If data are

- approximately normal ⇒ use difference in means
- symmetric but have outliers ⇒ use difference in trimmed means
- asymmetric ⇒ use difference in medians

2.2.4 Random sampling permutation

As m and n increase, the number of permutations could be too large. For example, $\binom{16}{8} = 12,870$, $\binom{20}{10} = 184,756$. It is not feasible to exhaust all of those permutations.

An approximate permutation p-value:

- take a random sample of say, $R = 1000$ permutations
- calculate the statistic of interest D for each permutation sample
- approximate the p-value using the reference distribution formed by the R permutation statistics in the same manner as the exact permutation test

The approximation improves as R increases. The standard error for approximate p-value: $\sqrt{p(1-p)/R}$, where p is the exact p-value. Therefore, the approximate p-value will have about a $100(1-\alpha)\%$

chance of falling into the interval

$$\left(p - z_{\alpha/2}\sqrt{p(1-p)/R},\ p + z_{\alpha/2}\sqrt{p(1-p)/R}\right).$$

For instance, if true p=0.05, $R = 1000$, with about a 95% chance, the approximate p-value will be within $\pm 1.96\sqrt{0.05 * 0.95/1000} = 0.0135$ of the true p-value.

Example 2.2.2 *(Tolerance of Violence) Table 4.4 of Hollander and Wolfe (1999). Toleration of violence was measured by time (in seconds) each child stayed in the room after he or she witnessed the two younger children's first act of violence. Do the data indicate that the children who viewed the violent TV tend to take longer to seek help (were more tolerant) than the children who viewed the nonviolent sports-action TV?*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Olym X watcher	12	44	34	14	9	19	156	23	13	11	47	26	14	33	15	62	5	8	0	154	146
Karate Y watcher	37	39	30	7	13	139	45	25	16	146	94	16	23	1	290	169	62	145	36	20	13

$H_0 : \Delta = 0$ versus $H_a : \Delta < 0$.

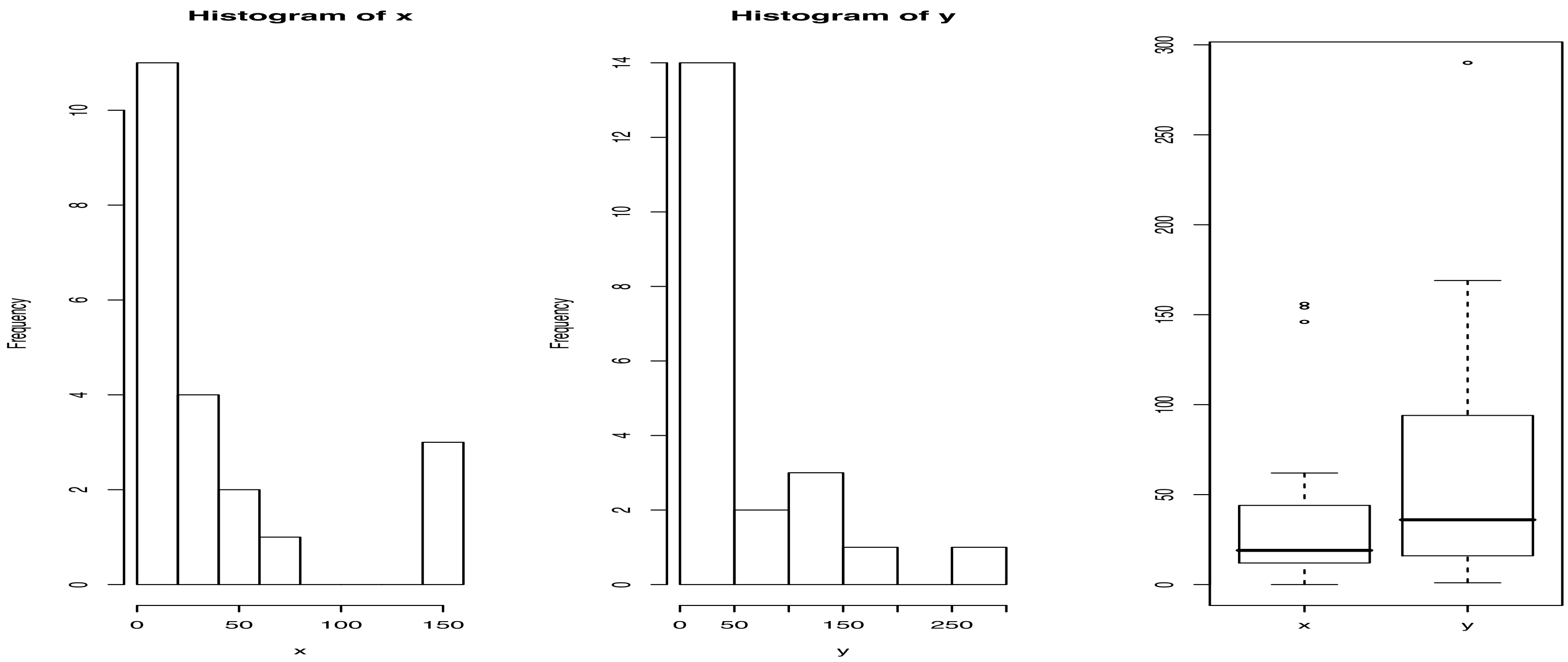

Figure 3: Histogram of tolerance of violence data.

Both distributions are clearly skewed to the right, so t-test is not a good choice.

R code:

```
source("functions-Ch2.R")
dat = read.csv("toleration-violence.csv")
x = dat[,1]
y = dat[,2]
# take a look at the data
par(mfrow=c(1,3))
hist(x)
hist(y)
boxplot(x, y, names=c("x","y"))

#random permutation using function rand.perm
rand.perm(x, y, R=1000, alternative = "less", stat= "mediandiff")
$pval
[1] 0.352
$Dobs
[1] -17

rand.perm(x, y, R=1000, alternative = "less", stat= "meandiff")
$pval
[1] 0.626
$Dobs
[1] -24.80952

# compare to the two-sample t-test
t.test(x, y, "less")
```

```
Welch Two Sample t-test
data:  x and y
t = -1.2869, df = 34.952, p-value = 0.1033
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 7.764467
sample estimates:
mean of x mean of y
 40.23810  65.04762
```

2.3 Wilcoxon Rank-Sum Test

2.3.1 Rank-Sum Test Statistic

- For distribution with heavy tailes, there may be extremely large/small values, and thus make t-test invalid or lack of power.
- Instead of comparing the means from two groups, we compare the ranks.
- Combine $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ to get a combined sample of $N = m + n$ observations.
- Let $R_i, i = 1, \cdots, m$ be the rank of X_i, and $S_j, j = 1, \cdots, n$ be the rank of Y_j in the combined sample. Use midranks for ties.

Refer to Example 2.2.1:

Method	Score			Mean	Rank Sums
New X_i	37	55	57	49.7	
rank R_i	3	4	5		12
Traditional Y_j	23	31	70	41.3	
rank S_j	1	2	6		9

- Then we can conduct the test using the same approach as permutation test, where the statistic of interest is

$$W = \text{sum of ranks from group 1 (or 2).}$$

Define

$$\mathbf{W_y = \sum_{j=1}^{n} S_j, W_x = \sum_{i=1}^{m} R_i.}$$

Note that

$$W_x + W_y = 1 + \cdots + N = \frac{N(N+1)}{2}, N = m + n.$$

- $H_a : \Delta > 0$: X is shifted to the right. Reject H_0 when W_x is large, or equivalently when W_y is small.
- $H_a : \Delta < 0$: X is shifted to the left. Reject H_0 when W_x is small, or equivalently when W_y is large.
- $H_a : \Delta \neq 0$. Reject H_0 when W_x is either too small or too large.
- Benefit of conducting test based on ranks:
  - the test will not be affected by extreme values (ranks but not magnitudes matter)
  - Under H_0, $E(W)$ and $V(W)$ are only functions of m and n, so the critical values of W for upper-tailed and lower-tailed tests can be tabulated

2.3.2 The null distribution of W

The null distribution of W is needed to obtain the critical values and the p-value. Under H_0, as in the permutation test, the $\binom{m+n}{n}$ partitionings have the equal chance to occur, each yielding one value of W. So we can obtain the null distribution of W by looking at all permutations when m and n are small. Such calculation is tedious and not feasible for larger sample sizes.

Refer to Example 2.2.1. **Calculation of p-value**:

- The observed test statistic value $W_{obs} = 12$ (sum of ranks for group X).
- Among 20 permutations, total 7 of W_k^* are ≥ 12.
- Therefore, the exact p-value for the upper-tailed test is $7/20 = 0.35$.
- The p-value for two-tailed test is $2 \times 0.35 = 0.7$.

- Since $\#\{W_k^* \leq 12\} = 16$, the exact p-value for the lower-tailed test is $16/20 = 0.8$.

In general, for

- upper-tailed test: p-value $= \#\{W_k^* \geq W_{obs}\}/\binom{m+n}{m}$
- lower-tailed test: p-value $= \#\{W_k^* \leq W_{obs}\}/\binom{m+n}{m}$
- two-tailed test: double of the p-value for the one-tailed test

k	Rank(permuted NM)			Rank(permuted TM)			W_k^*
1	3	4	5	1	2	6	12
2	3	4	1	5	2	6	8
3	3	4	2	5	1	6	9
4	3	4	6	5	1	2	13
5	3	5	1	4	2	6	9
6	3	5	2	4	1	6	10
7	3	5	6	4	1	2	14
8	3	1	2	4	5	6	6
9	3	1	6	4	5	2	10
10	3	2	6	4	5	1	11
11	4	5	1	3	2	6	10
12	4	5	2	3	1	6	11
13	4	5	6	3	1	2	15
14	4	1	2	3	5	6	7
15	4	1	6	3	5	2	11
16	4	2	6	3	5	1	12
17	5	1	2	3	4	6	8
18	5	1	6	3	4	2	12
19	5	2	6	3	4	1	13
20	1	2	6	3	4	5	9

The critical values are chosen to make the Type I error probability equal α.

For small m and n, the critical values of W are tabulated.

- the critical values are for one-sided tests
- sum is taken for treatment with n observations. For instance, if you use test statistic W as the rank sum from treatment 1 with 5 observations, and treatment 2 has 3 observations, then you should look for critical values with $n = 5$ and $m = 3$.

Rejection Region: For notational convenience, denote $w_{\alpha,U}$ as the upper critical value, and $w_{\alpha,L}$ as the lower critical value.

In general, with significance level of α,

- $H_a : \Delta > 0$: reject H_0 if $W_x \geq w_{\alpha,U}$ (X's shifted to the right)
- $H_a : \Delta < 0$: reject H_0 if $W_x \leq w_{\alpha,L}$
- $H_a : \Delta \neq 0$: reject H_0 when $W_x \geq w_{\alpha/2,U}$ or $W \leq w_{\alpha/2,L}$

Refer to Example 2.2.1. **Rejection region approach**:

The observed test statistic $W_x = 12$. Using $\alpha = 0.05$ and $m = n = 4$ (approximate),

- $H_a : \Delta > 0$: upper critical value is 25. Since $12 < 25$, we fail to reject H_0.
- $H_a : \Delta < 0$: lower critical value is 11. Since 12 is not smaller than 11, we fail to reject H_0.
- $H_a : \Delta \neq 0$: $w_{0.025,U} = 26$ and $w_{0.025,L} = 10$, we fail to reject H_0.

2.3.3 Large sample approximation

The null distribution of W can be found by looking at the rank sums under all possible partitions, but it's tedious and not feasible for large sample sizes.

As $n \to \infty$ and $m \to \infty$, under H_0,

$$\frac{W_x - E_0(W_x)}{\sqrt{V_0(W_x)}} \sim N(0,1) \text{ approximately,} \tag{2.2}$$

where

$$E_0(W_x) = m(N+1)/2,\ V_0(W_x) = mn(N+1)/12$$

are the mean and variance of W_x under H_0, $N = m + n$.

If $W = W_y$ is used, then

$$E_0(W_y) = n(N+1)/2, \quad V_0(W_x) = V_0(W_y).$$

Ties: when there are ties, we use the midranks. The variance should be replaced by

$$V_0(W_x) = V_0(W_y) = \frac{mn}{12}\left[N - 1 - \frac{\sum_{j=1}^{g}(t_j - 1)t_j(t_j + 1)}{N(N-1)}\right],$$

where g is the $\#$ of tied groups, t_j is the size of tied group j.

The symmetry of W_x under H_0: since W_x is symmetric around $m(N+1)/2$, for any w, $P(W_x \leq w) = P\{W_x \geq m(N+1) - w\}$ under H_0.

Example 2.3.1 *Dry weights of strawberry plants (Table 2.4.3 in Higgins). Data were obtained on 7 untreated and 9 treated with herbicide. It's expected that the untreated plants will have larger dry weights than the treated ones. Carry out the Wilcoxon's rank sum test. Use* $\alpha = 0.1$.

treated X	0.65 0.59 0.44 0.60 0.47 0.58 0.66 0.52 0.51
untreated Y	0.55 0.67 0.63 0.79 0.81 0.85 0.68

Solution:

Step 1: construct the appropriate null and alternative hypotheses. For this problem, $H_0 : \Delta = 0$ versus $H_0 : \Delta < 0$ (X is shifted to the left)

Step 2: estimate the rank sum test statistic

I used R for the tedious calculation

```
x=c(0.65,0.59,0.44,0.60,0.47,0.58,0.66,0.52,0.51)
y = c(0.55,0.67,0.63,0.79,0.81,0.85,0.68)
# sort the combined data
```

```
xy =c(x,y)
sort(xy)
#obtain the ranks of x
(m=length(x))
(n=length(y))
(rank.x = rank(xy)[1:m])
#rank sum statistic
(W=sum(rank.x))
```

Step 3: obtain the critical values and draw conclusion.

REMARK: note that the critical value is for the rank-sum statistic where the sum if taken for the group with n observations.

Here our W is defined by the sum of ranks for the treatment 1 with 9 observations. So we should look for the critical value corresponding to $n = 9$ and $m = 7$, that is, $w_{0.05,U} = 93$ and $w_{0.05,L} = 9 \times 17 - 93 = 60$. Since $52 < 60$, we reject H_0. The data provides enough evidence to support the expectation.

Using the approximate normal distribution.

For this example, we can also calculate the p-value by using the approximate normal distribution.
$E_0(W) = m(n + m + 1)/2 = 9 * 17/2 = 76.5$,
$V_0(W) = mn(m + n + 1)/12 = 89.25$. The standardized statistic $Z = (52 - 76.5)/\sqrt{89.25} = -2.59$. So for the lower-tailed test, p-value$=P(Z \leq -2.59) \approx \Phi(-2.59) = 0.005 < 0.1$, we reject H_0.

2.4 Mann-Whitney Test

Let $X_1, \cdots, X_m$ be observations from group 1, $Y_1, \cdots, Y_n$ be observations from group 2.

Mann-Whitney's Statistic:

$$U = \# \text{ of pairs } (X_i, Y_j) \text{ for which } X_i < Y_j.$$

Equivalence of U and rank-sum statistic W:

Suppose that all Y_j are less than all the X_i. So the sorted combined data is

	Y_1	Y_2	Y_3	Y_4	Y_5	X_1	X_2	X_3	X_4
rank	1	2	3	4	5	6	7	8	9

Then the rank-sum test statistic based on the ranks of Y_j is

$$W_Y = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}.$$

Every time when a Y_j exceeds an X_i (X_i becomes smaller than Y_j), the rank of that Y_i increases 1. e.g.

	Y_1	Y_2	Y_3	Y_4	X_1	X_2	X_3	Y_5	X_4
rank	1	2	3	4	5	6	7	8	9

Note that $U = 3$, and the rank of Y_5 increases 3. Therefore W_Y increases U. That is,

$$\boxed{W_Y = \frac{n(n+1)}{2} + U,\ W_X = \frac{N(N+1)}{2} - W_Y}$$

Ties: when there is an X_i and a Y_j such that $X_i = Y_j$, we add 1/2 to U for this pair.

Example 2.4.1 *The hours untill recharge of batteries of two brands are observed. Brand 1:* $X_i = \{3.6, 3.9, 4.0, 4.3\}$*. Brand 2:* $Y_j = \{3.8, 4.1, 4.5, 4.8\}$*. Then* $U = ?$

Solution: $U = 4 + 3 + 3 + 2 = 12.$

What if $Y_j = \{3.8, 4.0, 4.5, 4.8\}$? Then $U = 4 + 3 + 2.5 + 2 = 11.5.$

REMARK:

- U uniquely determines W and vice versa
- Note the equivalency of U and W.
  each critical value wrt to U equals the corresponding
  entry wrt to W minus $n(n+1)/2$,
  i.e $u_{\alpha,U} = w_{\alpha,U} - n(n+1)/2$
- The upper- and lower-critical values are related by the equation
  $u_{\alpha,U} = mn - u_{\alpha,L}$
- However, for U, it does not matter whether m is for group X or Y , as the table is symmetric for (m, n) and for (n, m). This is true because U counts the total number of pairs among $m \times n$ such that $X_i < Y_j$, but W counts the sum of ranks from one group so it should be distinguished which group the rank sum is associated with.
- The Wilcoxon's rank-sum test / Mann-Whitney test is

implemented in R function "wilcox.test". For wilcox.test(x, y, paired=FALSE): the output test statistic is defined as "the number of pairs (x_i, y_j) for which $y_j \leq x_i$. Therefore, if we use wilcox.test(y, x, paired=FALSE), the returned test statistic will just be the U statistic by our definition.

For the previous example, R code:

```
x = c(3.6, 3.9, 4.0, 4.3)
y = c(3.8, 4.1, 4.5, 4.8)
wilcox.test(y, x, paired=FALSE)
y = c(3.8, 4.0, 4.5, 4.8)
wilcox.test(y, x, paired=FALSE)
```

2.5 Confidence Interval and Hodges-Lehmann Estimate for the Shift Parameter Δ

Recall the location-shift model assumption:

- $X \sim F(\cdot)$, $Y \sim G(\cdot)$
- Location shift model: $F(x) = G(x - \Delta)$
- That is, X_i and $Y_i + \Delta$ have the same distribution
- Δ is the difference of locations of X and Y

Confidence interval for Δ:

- Arrange the mn **pairwise differences** $X_i - Y_j$ (pwd) from the smallest to the largest
- For the CI for Δ using

$$pwd(k_l) < \Delta \leq pwd(k_u), \tag{2.3}$$

where k_l and k_u are integers, $pwd(k)$ is the kth smallest pwd

- The inequality (2.3) holds if at least k_l and no more than $k_u - 1$ of the pairs (X_i, Y_j) satisfy

$$X_i - Y_j < \Delta \Leftrightarrow X_i < Y_j + \Delta$$

- X_i and $Y_j + \Delta$ have the same distribution, so probabilities involving $X_i < Y_j + \Delta$ can be obtained from the U statistic. That is

$$P(\text{at least } k_l \text{ and no more than } k_u - 1 \text{ pairs satisfy } X_i < Y_j + \Delta) \\ = P(k_l \leq U \leq k_u - 1)$$

- We want to find k_l and k_u such that

$$P(k_l \leq U \leq k_u - 1) = 1 - \alpha$$

- So let

$$k_l = u_{\alpha/2,L} + 1, \quad k_u = u_{\alpha/2,U},$$

where $u_{\alpha/2,U}$ and $u_{\alpha/2,L}$ are the upper- and lower-critical values of the U distribution.

Hodges-Lehmann Estimate for Δ: the median of the mn pwd's.

Example 2.5.1 *Suppose a verbal comprehension test is given to independent samples of educationally handicapped (EH) (population 1) and educable mentally retarded (EMR) children (population 2). The scores from the test are given in the table below.*

Educationally Handicapped (EH) 77 78 70 72 65 74 ($m = 6$)
Educable Mentally Ret. (EMR) 60 62 70 76 68 72 70 ($n = 7$)
Give Hodges-Lehmann estimate and 95% CI for Δ, where X and $Y + \Delta$ have the same distribution.

Solution:

The sorted $mn = 42$ pairwise differences are:
-11 -7 -6 -5 -5 -4 -3 -2 -2 0 0 0 1 2 2 2 2 2 3 4 4 4 5 5 6 6 7
7 8 8 8 9 10 10 10 12 12 14 15 16 17 18

The H-L estimate $\Delta\hat{} = .$__________

The lower- and upper-critical values of U are

$u_{0.025,L} =$ _______ and $u_{0.025,U} =$ __________. So $k_l =$ __________

and $k_u =$ __________. Thus, the 95% CI for Δ is

__________ $< \Delta \leq$ __________.

R code:

```
x = c(77, 78, 70, 72, 65, 74)
y = c(60, 62, 70, 76, 68, 72, 70)
pwd = outer(x,y,"-")
pwd = sort(pwd)
```

2.6 Comparison of 2-sample Tests

2-sample t-test

- If both populations are normal with unknown but equal variances, t-test has correct Type I error rate and the highest power among unbiased tests (whose Type I error rates are smaller than the significance level)
- When data are non-normal and sample sizes are large, then the Type I error rate of t-test is ok, but t-test may have low power

t-test versus Wilcoxon Rank-Sum

Distribution	Small sample sizes	Moderate large sizes
light tail (e.g. normal, uniform)	choose t-test	choose t-test
heavier tail (e.g. exponential Laplace, t_2)	t-test may have inflated Type I error power is comparable for many heavy-tailed dists	Wilcoxon often better

Power of Permutation Test v.s. Wilcoxon

- Ordinary 2-sample t-test and permutation t-test (or equivalently permutation based on the difference of means) have the same asymptotic power
- For light-tailed distributions (normal, uniform etc): permutation test on difference of means tends to have better power than the Wilcoxon test and the permutation test based on difference of medians
- For heavier-tailed distributions (Cauchy, Laplace etc), permutation test of medians performs the best, followed by Wilcoxon test, and the permutation test of means (t-test) is the worst.

2.7 Other Types of Tests

2.7.1 Tests for Equality of Scales

Sometimes we are interested in comparing the scales or spreads of two populations.

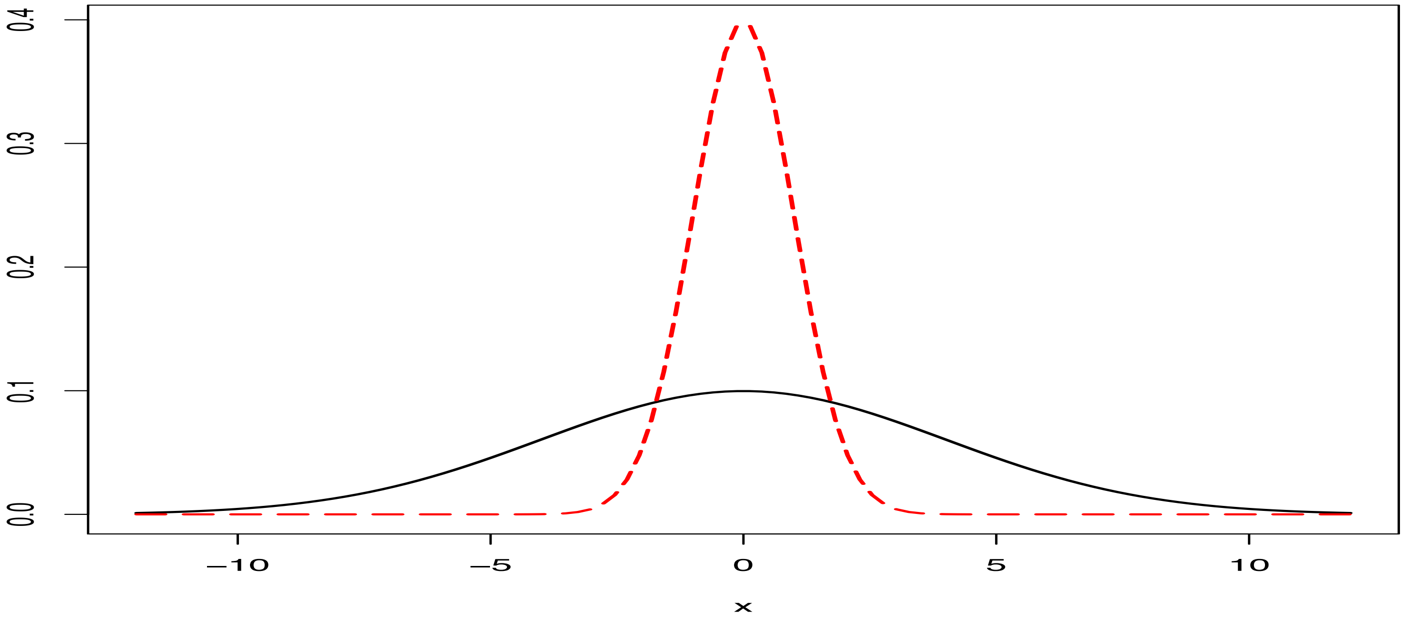

Figure 4: Two distributions with different scales.

Assume

$$X_i = \mu + \sigma_x \epsilon_i, \quad i = 1, \cdots, m,$$

$$Y_j = \mu + \sigma_y \epsilon_j, \quad j = 1, \cdots, n,$$

where ϵ_i are $i.i.d.$ with median 0. Note the same mean μ is assumed for each group. We wish to test the hypothesis

$$H_0 : \sigma_x = \sigma_y.$$

Siegel-Tukey Test (Steps)

- Arrange the combined data from smallest to largest
- Assign rank 1 to the smallest, rank 2 to the largest, then rank 3 to the next largest, and rank 4 to the next smallest, and so on. For example:

ordered data	x(min)	x	y	y	$\cdots$	y	y	x	x (max)
Siegel rank	1	4	5	8	$\cdots$	7	6	3	2

Note that if X has larger variability, then the X_i's will have more of the smaller ranks, (eg. 1,2,3,4...).

- Apply the Wilcoxon rank-sum test. The group with larger variability (σ) tend to have smaller ranks and thus rank-sum.

Example 2.7.1 *Data on the amount of liquid in randomly selected beverage containers before and after the filling system is fixed have been observed as follows. Test the equality of the scales of two groups.*

Before Repair X	16.55 15.36 15.94 16.43 16.01
After Repair Y	16.05 15.98 16.1 15.88 15.91

Solution:

- Construct the appropriate hypotheses $H_0 : \sigma_x = \sigma_y$ versus $H_a : \sigma_x \neq \sigma_y$ (two-tailed test).
- The sorted combined data:

group	x	y	y	x	y	x	y	y	x	x
sorted	15.36	15.88	15.91	15.94	15.98	16.01	16.05	16.10	16.43	16.55
siegel										

- Rank-sum statistic (for group X): $W = ?$
- Find critical values:
- Conclusion:
- The sample standard deviations: sd(x)=0.47, sd(y)=0.09. What if we want to test $H_a : \sigma_x > \sigma_y$?
- p-value calculation using permutation.

2.7.2 Tests on Deviances

Now suppose that the mean of X_i is not necessarily the same as the mean of Y_j. Assume

$$X_i = \mu_x + \sigma_x \epsilon_i, \quad i = 1, \cdots, m(\text{group } 1),$$

$$Y_j = \mu_y + \sigma_y \epsilon_j, \quad j = 1, \cdots, n(\text{group } 2).$$

So here μ_x and μ_y may differ, ϵ's are $i.i.d.$ with median 0.

We want to test the hypothesis $H_0 : \sigma_x = \sigma_y$.

If the ϵ's are normally distributed, the best test is F-test based on

$$F = \frac{S_x^2}{S_y^2},$$

where S_x^2 and S_y^2 are the sample variances of X_i and Y_j.

When ϵ's are not normally distributed, the Type I error rate of the F-test can be inflated (larger than the nominal level α).

Now assume that the location parameters μ_x and μ_y are known. If not, we can estimate them by the sample medians, denoted by med_x and med_y, respectively.

- Calculate deviances

$$dev_{ix} = X_i - \mu_x \ \ (\text{or } X_i - med_x)$$

$$dev_{iy} = Y_i - \mu_y \ \ (\text{or } Y_j - med_y)$$

- The test statistic is the ratio of absolute mean deviances

$$RMD = \frac{\sum_{i=1}^{m} |dev_{ix}|/m}{\sum_{j=1}^{n} |dev_{iy}|/n}.$$

- The p-value is calculated using the standard permutation method:
  - Find all (or a sample) of the $\binom{m+n}{m}$ permutations of the $m+n$ deviances, and calculate RMD^* for each permutation
  - For $H_a : \sigma_x > \sigma_y$, p-value is the proportion of RMD^*'s that are $\geq RMD_{obs}$

- For $H_a : \sigma_x < \sigma_y$, p-value is the proportion of RMD^*'s that are $\leq RMD_{obs}$
- For $H_a : \sigma_x \neq \sigma_y$, calculate

$$RMD_{2sided} = \frac{\max\{\sum_{i=1}^m |dev_{ix}|/m, \sum_{j=1}^n |dev_{iy}|/n\}}{\min\{\sum_{i=1}^m |dev_{ix}|/m, \sum_{j=1}^n |dev_{iy}|/n\}}$$

  for both the observed sample and all the permutations. Then p-value is the proportion of RMD^*_{2sided}'s $\geq RMD_{obs,2sided}$

2.7.3 Kolmogorov-Smirnov Test

- Designed to detect differences of two distributions in either location, scale (variability) or shape.
- Considered as an "omnibus test".
- Test statistic:

  $$KS = \max_w |\hat{F}(w) - \hat{G}(w)|,$$

  where $\hat{F}(w)$ and $\hat{G}(w)$ are the empirical CDF of X and Y distributions.
- p-value is calculated based on KS^* from all (or a sample of) permutations.

Example 2.7.2 *Refer to Example 2.7.1. Test if there is any difference in the distributions of two groups.*

R code:

```
x = c(16.55, 15.36, 15.94, 16.43, 16.01)
y = c(16.05, 15.98, 16.1, 15.88, 15.91)
##plot the ECDF curves
plot(ecdf(x), verticals=TRUE)
lines(ecdf(y), verticals=TRUE, col="red", lty="dashed")
ks.test(x,y)
```