2 Two-Sample Methods

Review of the two-sample *t*-test

Problem #2.1.1 (two diets), problem2.1.1.txt: Nieman, Groot, and Jansen (1952), "The nutritive value of butter fat compared with that of vegetable fats, I," *Koninkl. Ned, Akad. Van. Wetenschap, Ser. C* **55:** 588–604.

In a comparison of the effect on growth of two diets A and B, a number of growing rats were placed on these two diets, and the following growth figures were observed after 7 weeks.

A	156	183	120	113	138	145	142			
B	109	107	119	162	121	123	76	111	130	115

Use the two-sample t-test to test if the mean growth rate differs for the two diets, at level $\alpha = 0.05$.

Do not assume that the two populations have equal variances.

```
> z =
   scan("http://educ.jmu.edu/~garrenst/math324.dir/datasets/problem2.1.1.txt",
   comment.char="#")
> z = scan2( "problem2.1.1.txt" )
> x = z[ 1 :  7 ]
> y = z[ 8 :  length( z ) ]
```

$H_0 : \ \mu_x = \mu_y, \ \ H_a : \ \mu_x \neq \mu_y$

$$T = \frac{\bar{X} - \bar{Y} - (\mu_{x,0} - \mu_{y,0})}{\sqrt{s_x^2/m + s_y^2/n}}$$

The estimated number of degrees of freedom is

$$\frac{(s_x^2/m + s_y^2/n)^2}{\frac{s_x^4/m^2}{m-1} + \frac{s_y^4/n^2}{n-1}}.$$

You do NOT need to memorize this formula for estimated degrees of freedom.

```
> mean( x )

> mean( y )

> sd( x )

> sd( y )

> ?t.test
```

What assumptions were necessary for performing this two-sample t-test?

```
> stripchart( x, "stack" ) # For a dotplot of x.
> stripchart( y, "stack" ) # For a dotplot of y.
> lineGraph( x ) # For a line graph of x.
> lineGraph( y ) # For a line graph of y.
```

□

2.1.1 The Permutation Test

Revisit example (two diets), problem2.1.1.txt: Test for inequality of *means* at level 0.05.

A	156	183	120	113	138	145	142			
B	109	107	119	162	121	123	76	111	130	115

Diet A has 7 observations, and diet B has 10 observations.

If $\bar{X} - \bar{Y}$ is sufficiently large or small, then reject H_0 in favor of H_a.

First, compute $\bar{X} - \bar{Y}$.

Next, permute the 17 observations, and assign 7 of them to diet A and the remaining 10 of them to diet B.

Compute the *permuted* value of $\bar{X} - \bar{Y}$.

How many *combinations* exist for the 17 observations, with 7 of them assigned to diet A and 10 of them assigned to diet B?

Compute $\bar{X} - \bar{Y}$ for each *permutation*.

Determine the proportion of times that the *permuted* values of $|\bar{X} - \bar{Y}|$ are at least as large as the *observed* value of $|\bar{X} - \bar{Y}|$.

That will be your p-value!

□

Problem #2.1.2; simpler example (smaller sample sizes):

Consider the following mutually independent observations from two different continuous populations.

A	35	25	
B	50	30	70

(a) Test $H_0: \ \mu_x = \mu_y$ versus $H_a: \ \mu_x < \mu_y$ at level 0.05, using the difference in means as the test statistic.

Compute the *observed* value of $\bar{X} - \bar{Y}$.

```
> mean( c(35, 25) ) - mean( c(50, 30, 70) )
```

How many *combinations* exist for the 5 observations, with 2 of them assigned to diet A and 3 of them assigned to diet B?

List all possible groupings of the 5 observations, and compute the *permuted* value of $\bar{X} - \bar{Y}$, along with $\sum_{i=1}^{m} X_i$.

Permuted Samples	Sample A		Sample B			$\bar{X}-\bar{Y}$	$\sum_{i=1}^{m} X_i$
1	25	30	35	50	70	-24.17	55
2^*	25	35	30	50	70	-20	60
3	25	50	30	35	70	-7.5	75
4	25	70	30	35	50	9.17	95
5	30	35	25	50	70	-15.83	65
6	30	50	25	35	70	-3.33	80
7	30	70	25	35	50	13.33	100
8	35	50	25	30	70	0.83	85
9	35	70	25	30	50	17.5	105
10	50	70	25	30	35	30	120

List the **permutation distribution** of $\bar{X}-\bar{Y}$.

$\bar{X}-\bar{Y}$	-24.17	-20^*	-15.83	-7.5	-3.33	0.83	9.17	13.33	17.5	30	sum
probability	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1

Determine the proportion of times that the *permuted* values of $\bar{X}-\bar{Y}$ are at least as **small** as the *observed* value of $\bar{X}-\bar{Y}$.

(b) Test $H_0:\ \mu_x=\mu_y$ versus $H_a:\ \mu_x<\mu_y$ at level 0.05, using the **sample sum** from treatment A as the test statistic.

(c) Test $H_0:\ \mu_x = \mu_y$ versus $H_a:\ \mu_x \neq \mu_y$ at level 0.05, using the difference in means as the test statistic.

Is the p-value $= 2 \times 0.2 = 0.4$?

```
> lineGraph( c(-24.17, -20, -7.5, 9.17, -15.83, -3.33, 13.33, 0.83,
  17.5, 30) )
```

(d) Test $H_0:\ \mu_x = \mu_y$ versus $H_a:\ \mu_x < \mu_y$ at level 0.05, using the difference in means as the test statistic, and using the function `perm.test` from *R*-package `jmuOutlier`.

```
> perm.test( c(25, 35), c(30, 50, 70), "less" )
```

```
> x = c(25, 35) ;  y = c(30, 50, 70)
> perm.test( x, y, "less" )
```

(e) Test $H_0:\ \mu_x = \mu_y$ versus $H_a:\ \mu_x \neq \mu_y$ at level 0.05, using the difference in means as the test statistic, and using the function `perm.test`.

□

Revisit problem #2.1.1 (two diets), problem2.1.1.txt: Test for inequality of *means* at level 0.05.

$H_0:\ \mu_x = \mu_y,\ \ H_a:\ \mu_x \neq \mu_y$

A	156	183	120	113	138	145	142			
B	109	107	119	162	121	123	76	111	130	115

How many *combinations* are required to calculate this exact p-value?

When using `perm.test`, the parameter `num.sim` is an upper limit on the number of *combinations* to be computed.

```
> perm.test( x, y )
```

□

Example: Consider the nonparametric test based on the difference between two *means*, and also consider the nonparametric test based on the sample sum from treatement A. Are these two tests equivalent? Justify your answer mathematically.

□

Homework p. 73: Exercises 2.1* (based on means; hand-calculations and

`perm.test` by *R*), 2.2a (hand-calculations and *R*), 2.2b (hand-calculations)

Hints for homework exercise 2.1*: Assume that the permutation test is based on the difference between two means. **EXPLAIN** why your numerical answer is the correct *p*-value. Solve this exercise **TWICE,** once using hand-calculations (i.e., definitely use *R*, but withOUT any *R*-functions containing the word `test`) and once using `perm.test`. When using `perm.test`, just create a data set according to the specified conditions. Either retype your *p*-value as a comment using "#", or highlight the *p*-value in yellow. This should be a RIGHT-sided test. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.1 </span>`.

2.2 Permutation Tests Based on the Median and Trimmed Means

Recall: Heavy-tailed distributions tend to produce outliers, which have large impacts on *means* but not on *medians*.

Hence, when sampling from heavy-tailed distributions, a nonparametric test based on the difference between the two *medians* might be more appropriate than one based on the difference between the two *means*.

2.2.1 A Permutation Test Based on Medians

Revisit problem #2.1.2, section 2.1: Consider the following mutually independent observations from two different populations.

A	35	25	
B	50	30	70

(a) Test $H_0: \widetilde{\mu}_x = \widetilde{\mu}_y$ versus $H_a: \widetilde{\mu}_x < \widetilde{\mu}_y$ at level 0.05, using the difference in *medians* as the test statistic.

Compute the *observed* value of difference in sample medians.

Permuted Samples	Sample A		Sample B			$\text{med}(X_i) - \text{med}(Y_i)$
1	25	30	35	50	70	-22.5
2*	25	35	30	50	70	-20
3	25	50	30	35	70	2.5
4	25	70	30	35	50	12.5
5	30	35	25	50	70	-17.5
6	30	50	25	35	70	5
7	30	70	25	35	50	15
8	35	50	25	30	70	12.5
9	35	70	25	30	50	22.5
10	50	70	25	30	35	30

```
> x = c(25, 35) ;  y = c(30, 50, 70)
> perm.test( x, y, "less", stat=median )
```

(b) Test $H_0: \widetilde{\mu}_x = \widetilde{\mu}_y$ versus $H_a: \widetilde{\mu}_x \neq \widetilde{\mu}_y$ at level 0.05, using the difference between *medians* as the test statistic.

□

Revisit problem #2.1.1 (two diets), problem2.1.1.txt: Test for the inequality of *medians* at level 0.05, using the difference in medians as the test statistic.

$H_0: \widetilde{\mu}_x = \widetilde{\mu}_y, \ H_a: \widetilde{\mu}_x \neq \widetilde{\mu}_y$

A	156	183	120	113	138	145	142			
B	109	107	119	162	121	123	76	111	130	115

2.2.2 Trimmed Means

A **trimmed mean** is a compromise between a sample **mean** and a sample **median**.

For example, a 20% trimmed mean is the mean of the middle 80% of the observations.

In other words, we delete the smallest 10% and the largest 10% of the observations, and then average the remaining 80% of the observations to obtain our 20% **trimmed mean**.

Technically, when *trimmed* sample means are used, the hypothesis test is based on a *trimmed* population mean. However, if the population is symmetric (and has finite mean), then the population mean μ is equal to the *trimmed* population mean.

Problem #2.2.1 (hypnosis), problem2.2.1.txt: Agosti and Camerota (1965), "Some effects of hypnotic suggestion on respiratory function," *Intern. J.*

Clin. Exptl. Hypnosis, **13**: 149–156.

At the beginning of a study of the effect of hypnotism, the following measurements of ventilation were taken on eight treatment subjects (to be hypnotized) and eight controls.

Control:	4.69	4.19	3.99	4.21	4.84	4.54	5.48	4.64
Treatment:	5.52	4.36	5.08	5.20	4.78	5.74	4.67	5.16

(a) Construct a line graph and a Q-Q plot for each of the two data sets.

```
> z = scan2( "problem2.2.1.txt" )
> x = z[ 1 :  8 ]
> y = z[ 9 :  16 ]
```

(b) Using the (two-sample) t-test, test $H_0 :\ \mu_x = \mu_y$ versus $H_a :\ \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means of the control and treatment groups, respectively.

(c) Using the permutation test based on **means**, test $H_0 :\ \mu_x = \mu_y$ versus $H_a :\ \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means of the control and treatment groups, respectively. Compute the **exact** p-value.

```
> perm.test( x, y, "less" )
```

```
> perm.test( x, y, "less", plot=TRUE ) # To plot the permutation
  distribution.
```

(d) Using the permutation test based on **medians**, test $H_0: \widetilde{\mu}_x = \widetilde{\mu}_y$ versus $H_a: \widetilde{\mu}_x < \widetilde{\mu}_y$ at level 0.05, where $\widetilde{\mu}_x$ and $\widetilde{\mu}_y$ are the population medians of the control and treatment groups, respectively. Compute the **exact** p-value.

(e) Using the permutation test based on a 25% **trimmed** mean, test $H_0: \mu_x = \mu_y$ versus $H_a: \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means of the control and treatment groups, respectively. Compute the **exact** p-value.

Below is the *sorted* table.

Control:	3.99	4.19	4.21	4.54	4.64	4.69	4.84	5.48
Treatment:	4.36	4.67	4.78	5.08	5.16	5.20	5.52	5.74

```
> mean( sort(x)[ 2 :  7 ] )
```

```
> ?mean
```

Value of test statistic is:

```
> ?perm.test
```

```
> perm.test( x, y, "less", trim=0.125 )
```

□

Homework p. 73: Exercise 2.3* (hand-calculations)

Hints for homework exercise 2.3*: The permutation distribution consists of all possible values of the difference in medians, without duplication, in order from smallest to largest, along with the associated probabilities, which must sum to unity. Since $m = n = 3$, then the permutation distribution must be symmetric. The homework exercise does **NOT** ask you to compute a p-value. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.3 </span>`.

2.3 Random Sampling the Permutations

In the previous example (problem #2.2.1; hypnosis), the permutation tests involved computations based on $\binom{16}{8} = 12,870$ groupings (or combinations) of the 16 subjects.

```
> choose( 16, 8 )
```

```
> q0 = Sys.time( ) ; perm.test( x, y, "less" ) ; Sys.time( ) - q0
```

Suppose we had 10 subjects in each group.

Suppose we had 20 subjects in each group.

Suppose we had 25 subjects in each group.

□

2.3.1 An Approximate *p*-Value Based on Random Sampling the Permutations

Recall section 0.6, regarding permutation methods.

Revisit problem #2.2.1 (hypnosis), problem2.2.1.txt: Using the permutation test based on **means**, test $H_0: \mu_x = \mu_y$ versus $H_a: \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means of the control and treatment groups, respectively. Approximate the p-value based on 10,000 randomly sampled permutations.

Control:	4.69	4.19	3.99	4.21	4.84	4.54	5.48	4.64
Treatment:	5.52	4.36	5.08	5.20	4.78	5.74	4.67	5.16

Steps for approximating the p-value

Use the permutation test based on **means** to test $H_0: \ \mu_x = \mu_y$ versus $H_a: \ \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means.

Approximate the p-value based on 10,000 randomly sampled permutations.

(1) Compute $\bar{X} - \bar{Y}$; i.e., the difference in sample means for the observed data.

(2) Permute (shuffle, at random) the 16 observations to reassign 8 observations to the control group and the remaining 8 observations to the treatment group.

(3) Compute $\bar{X} - \bar{Y}$ for this **permuted** data set.

(4) Repeat steps (2) and (3) many times, such that all selected permutations of 16 observations are independent, until 10,000 independent **permuted** values of $\bar{X} - \bar{Y}$ are generated.

(5) Determine the proportion of the **permuted** values of $(\bar{X} - \bar{Y})$ which are at least as small as the **observed** value of $(\bar{X} - \bar{Y})$. That is your approximated p-value!

```
> ?perm.test
> perm.test( x, y, "less", num.sim=1e4 )
```

□

Accuracy of the Procedure

Suppose p is the true (exact) p-value based on $(\bar{X} - \bar{Y})$, when testing $H_0: \ \mu_x = \mu_y$ versus $H_a: \ \mu_x < \mu_y$.

What is the probability that a randomly selected **permutation** of the data produces a value of $(\bar{X} - \bar{Y})$ at least as small as the **observed** value of $(\bar{X} - \bar{Y})$ (conditional

on the original data set)?

Let $\hat{p}$ be the sample proportion of times that randomly selected **permutations** of the data produce values of $(\bar{X} - \bar{Y})$ at least as small as the **observed** value of $(\bar{X} - \bar{Y})$.

What is the approximate distribution of $\hat{p}$ (conditional on the original data set)?

What is an approximate 95% confidence interval on p (conditional on the original data set)?

Revisit problem #2.2.1 (hypnosis), problem2.2.1.txt: Using the permutation test based on **means**, test $H_0: \mu_x = \mu_y$ versus $H_a: \mu_x < \mu_y$ at level 0.05, where μ_x and μ_y are the population means of the control and treatment groups, respectively. Construct a **95% confidence interval** on the (true) p-value based on 10,000 randomly sampled permutations.

```
> p.value = perm.test( x, y, "less", num.sim=1e4 )$p.value
```

```
> J = 1e4
> p.value - qnorm(0.975) * sqrt( p.value * (1-p.value) / J )
> p.value + qnorm(0.975) * sqrt( p.value * (1-p.value) / J )
```

How can we decrease the width of the confidence interval?

Homework p. 74: Exercise 2.8 (first question)

2.4 Wilcoxon Rank-Sum Test

For a data set $X_1,\ X_2, \ldots, X_m$ with no ties, the **rank** of an observation X_i is

$$R(X_i) = \text{ number of } X_j's \leq X_i.$$

Revisit problem #2.1.2, section 2.1. Consider the following mutually independent observations from two different populations.

A	35	25	
B	50	30	70

```
> z = c( 35, 25, 50, 30, 70 )
> rank( z )
```

Let W_i be the sum of the ranks of the observations in treatment $\#i$, $i = 1, 2$.

Compare W_1 to all **permuted** values of W_1.

The p-value is the proportion of **permuted** values of W_1 which are at least as extreme as the **observed** value of W_1.

How many ways can we divide the five observations into two groups, where group A has 2 observations and group B has 3 observations?

(a) Use *hand-calculations* (i.e., calculations may use R but not any functions containing the word `test`) and the Wilcoxon rank-sum test to test at level 0.05, H_0: the populations A and B are the same, versus H_a: the values in population A tend to be **smaller** than the values in population B.

Permuted Samples	Sample A		Sample B			$R(X_1)$	$R(X_2)$	W_1
1	25	30	35	50	70	1	2	3
2*	25	35	30	50	70	1	3	4
3	25	50	30	35	70	1	4	5
4	25	70	30	35	50	1	5	6
5	30	35	25	50	70	2	3	5
6	30	50	25	35	70	2	4	6
7	30	70	25	35	50	2	5	7
8	35	50	25	30	70	3	4	7
9	35	70	25	30	50	3	5	8
10	50	70	25	30	35	4	5	9

(b) Repeat part (a) using the function `wilcox.test`.

```
> ?wilcox.test

> wilcox.test( x, y, "less" )
```

(c) Use the Wilcoxon rank-sum test to test at level 0.05, H_0: the populations A and B are the same, versus H_a: the populations A and B are different.

Under H_0, is W_1 symmetric?

(d) Repeat part (c) using the function `wilcox.test`.

```
> ?wilcox.test
```

□

2.4.2 Comments on the Use of the Wilcoxon Rank-Sum Test

Let W_1 be the sum of the ranks of the m observations in treatment #1, and let W_2 be the sum of the ranks of the n observations in treatment #2.

What does $(W_1/m - W_2/n)$ represent?

Example: Is a test based on $(W_1/m - W_2/n)$ equivalent to a test based on W_1? Justify your answer mathematically.

2.4.3 A Statistical Table for the Wilcoxon Rank-Sum Test *AND*

2.4.4 Computer Analysis

Since R provides high precision for p-values and confidence intervals for virtually any practical sample size, we will disregard Table A3 (whose values of m and n cannot exceed 10, and whose list of α values is limited to 0.01, 0.025 and 0.05).

Homework C2.4.1*: Perform the two-sided Wilcoxon rank-sum test at level $\alpha = 0.05$ for the data below. Clearly state the null and alternative hypotheses. Also, state the conclusion in both statistical terms and regular English. Show how you obtain the p-value. Either retype your p-value as a comment using "#", or highlight the p-value in yellow. Introduce the question number and letter as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise C2.4.1(a) </span>`.

A	97	64	51
B	45	73	32

(a) using **ONLY** hand-calculations (i.e., you may use R, but not `wilcox.test` or `perm.test`),

(b) using **ONLY** the R-function `wilcox.test`.

End of Homework C2.4.1*. □

Homework p. 73: Exercises 2.4, 2.5, 2.6* (such that one p-value exceeds 0.2 and the other p-value is smaller than 0.03, using $m = n = 4$, where all eight observations differ from each other), 2.8 (2nd and 3rd questions)

Hints for homework exercise 2.6*: In general, when the textbook does not specify, assume a **TWO**-sided test, as in this exercise. Show that your data set meets the specified conditions using `wilcox.test` and `t.test`. State the null and alternative hypotheses. These hypotheses **differ** for the two testing procedures. Either retype your p-values as a comment using "#", or highlight the p-values in yellow. State the conclusion regarding whether you reject or fail to reject the null hypothesis. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.6 </span>`.

2.5 Wilcoxon Rank-Sum Test Adjusted for Ties

Example: Modify the data from problem #2.1.2, section 2.1. Consider the following mutually independent observations from two different populations.

A	35	25	
B	50	35	70

The **adjusted rank** is the average rank of the **tied** observations.

What rank should be assigned to the two values of 35?

```
> z = c( 35, 25, 50, 35, 70 )
```

```
> rank( z )
```

Use *hand-calculations* and the Wilcoxon rank-sum test to test at level 0.05, H_0: the populations A and B are the same versus H_a: the values in population A tend to be *smaller* than the values in population B.

Permuted Samples	Sample A		Sample B			$R(X_1)$	$R(X_2)$	W_1
1	25	35	35	50	70	1	2.5	3.5
2*	25	35	35	50	70	1	2.5	3.5
3	25	50	35	35	70	1	4	5
4	25	70	35	35	50	1	5	6
5	35	35	25	50	70	2.5	2.5	5
6	35	50	25	35	70	2.5	4	6.5
7	35	70	25	35	50	2.5	5	7.5
8	35	50	25	35	70	2.5	4	6.5
9	35	70	25	35	50	2.5	5	7.5
10	50	70	25	35	35	4	5	9

□

Homework p. 74: Exercise 2.7

2.6 Mann-Whitney Test and a Confidence Interval

2.6.1 The Mann-Whitney Statistic

Consider observations $X_1, X_2, \ldots, X_m$ from treatment A, and consider observations $Y_1, Y_2, \ldots, Y_n$ from treatment B.

The **Mann-Whitney statistic** is defined to be:

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

The null distribution of U may be determined, under the null hypothesis that the two populations are the same.

Compute the empirical probability that the *observed* value of U falls in the appropriate tail of the null distribution of U to determine what?

In other words, compute the proportion of the **permuted** values of U which are at least as extreme as the **observed** value of U to determine what?

What is the total number of pairings of (X_i, Y_j)?

Revisit problem #2.1.2, section 2.1. Consider the following mutually independent observations from two different populations.

A	35	25	
B	50	30	70

Use *hand-calculations* and the Mann-Whitney test to test at level 0.05, H_0: the populations A and B are the same versus H_a: the values in population A tend to be *smaller* than the values in population B.

Should we reject H_0 for large or small values of U?

For how many values of Y_j do we have $35 < Y_j$?

For how many values of Y_j do we have $25 < Y_j$?

What is the *observed* value of U?

List all combinations of the five observations such that two observations are in group A, and compute U for each combination.

Permuted Samples	Sample A		Sample B			U
1	25	30	35	50	70	$3+3=6$
2*	25	35	30	50	70	$3+2=5$
3	25	50	30	35	70	$3+1=4$
4	25	70	30	35	50	$3+0=3$
5	30	35	25	50	70	$2+2=4$
6	30	50	25	35	70	$2+1=3$
7	30	70	25	35	50	$2+0=2$
8	35	50	25	30	70	$1+1=2$
9	35	70	25	30	50	$1+0=1$
10	50	70	25	30	35	$0+0=0$

What proportion of the **permuted** values of U are at least as large as our **observed** value of $U = 5$?

□

How should **ties** be handled with the Mann-Whitney test?

2.6.2 Equivalence of Mann-Whitney and Wilcoxon Rank-Sum Statistics

A monotone increasing (in fact, linear) relationship between the Mann-Whitney statistic and the Wilcoxon rank-sum statistic exists.

What must be true regarding the p-value obtained from the Mann-Whitney statistic compared to the p-value obtained from the Wilcoxon rank-sum statistic?

Revisit problem #2.2.1 (hypnosis), problem2.2.1.txt: Use the **Mann-Whitney** test to test at level 0.05 H_0: the control and treatment groups are the same versus H_a: the values in the control group tend to be *smaller* than the values in the treatment group.

Control:	4.69	4.19	3.99	4.21	4.84	4.54	5.48	4.64
Treatment:	5.52	4.36	5.08	5.20	4.78	5.74	4.67	5.16

```
> z = scan2( "problem2.2.1.txt" )
> x = z[ 1 :  8 ]
```

```
> y = z[ 9 :  16 ]
```

□

2.6.3 A Confidence Interval for a Shift Parameter, Δ, and the Hodges-Lehmann Estimate

Suppose we sample mutually independent observations from two distributions, which may differ by merely the location parameter.

Draw probability density functions of $N(2, 1)$ (for X) and $N(0, 1)$ (for Y) on the same graph.

Draw probability density functions of Cauchy$(2, 1)$ (for X) and Cauchy$(0, 1)$ (for Y) on the same graph.

However, the only assumption is that the distributions are identical except for the location parameters.

Estimating the shift parameter, Δ

Let X be an observation from a distribution with location μ.

Let Y be an observation (independent of X) from the same distribution but with location $\mu - \Delta$.

What is $P(X - \Delta < Y)$, for continuous X and Y?

What is $P(X - Y < \Delta)$, for continuous X and Y?

What is the population median of $X - Y$, for continuous X and Y?

How should we estimate the population median of $X - Y$ based on our sample of mutually independent observations $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, for continuous X and Y?

This estimate of Δ is called the **Hodges-Lehmann estimate**.

Theoretical Basis for the Confidence Interval

Consider the new data set based on all possible values of $X_i - Y_j$.

Idea: Construct a 95% confidence interval on the population median, based on this new DEPENDENT data set.

The confidence interval on Δ is based on the **Mann-Whitney statistic**, which was defined (in section 2.6.1) to be

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

The textbook suggests using Table A4 (Lower and Upper Critical Values for Mann-Whitney Statistic) to construct the confidence interval, whereas we will use *R*.

Problem #2.6.1 (shocking rats): Solomon and Coles (1954), "A case of failure of generalization of imitation across drives and across situations," *J. Abnorm. Soc. Psychol.*, **49**: 7–13.

From a group of nine rats available for a study of the transfer of learning, five were selected at random and were trained to imitate leader rats in a maze. They were then placed together with four untrained control rats in a situation where imitation of the leaders enabled them to avoid receiving an electrical shock. The results (the number of trials required to obtain ten correct responses in ten consecutive trials) were as follows:

Controls (X):	110	70	53	51	
Trained rats (Y):	78	64	75	45	82

(a) Determine the **Hodges-Lehmann estimate** of Δ, without using the function `wilcox.test`.

Assume that the two above populations are identical in distribution except for the population medians, and define the shift parameter Δ to be the median of population A minus the median of population B.

pairings		
X	Y	$X - Y$
110	78	$110 - 78 = 32$
110	64	$110 - 64 = 46$
110	75	$110 - 75 = 35$
110	45	$110 - 45 = 65$
110	82	$110 - 82 = 28$
70	78	$70 - 78 = -8$
70	64	$70 - 64 = 6$
70	75	$70 - 75 = -5$
70	45	$70 - 45 = 25$
70	82	$70 - 82 = -12$
53	78	$53 - 78 = -25$
53	64	$53 - 64 = -11$
53	75	$53 - 75 = -22$
53	45	$53 - 45 = 8$
53	82	$53 - 82 = -29$
51	78	$51 - 78 = -27$
51	64	$51 - 64 = -13$
51	75	$51 - 75 = -24$
51	45	$51 - 45 = 6$
51	82	$51 - 82 = -31$

```
> x = c( 110, 70, 53, 51 )
> y = c( 78, 64, 75, 45, 82 )
```

```
> u1 = x[1] - y

> u2 = x[2] - y

> u3 = x[3] - y

> u4 = x[4] - y

> u = c( u1, u2, u3, u4 )


> median( u )
```

(b) Construct the 90% confidence interval on Δ, **using** the function `wilcox.test`.

Again, assume that the two above populations are identical in distribution except for the location parameters.

```
> wilcox.test( x, y, conf.int=TRUE, conf.level=0.9 )
```

(c) Without assuming normality, test at level 0.05, $H_0 : \Delta = 0$ versus $H_a : \Delta \neq 0$, where the two populations are identical under H_0.

□

Two-sample *t*-test with pooled standard deviation

What assumptions are needed to construct a two-sample t-test or two-sample t-confidence interval?

When constructing the Wilcoxon rank-sum test, the two populations are identical under H_0, and the two populations may differ only by the location parameter when constructing confidence intervals.

To appropriately compare the Wilcoxon rank-sum test with the t-test (or confidence intervals), what assumptions are needed on the populations?

The *pooled* standard deviation is

$$s_\rho = \sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}}.$$

The appropriate confidence interval on $\mu_1 - \mu_2$ is

$$\bar{X} - \bar{Y} \pm t \; s_\rho \sqrt{1/m + 1/n},$$

where the t critical value is based on $(m+n-2)$ degrees of freedom.

Revisit problem #2.6.1 (shocking rats):

Controls (X):	110	70	53	51	
Trained rats (Y):	78	64	75	45	82

Assume that the data are mutually independent observations from two normal populations $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$.

(a) Are there any outliers?

(b) Compute the difference in sample means.

(c) Construct the 90% confidence interval on $\Delta = \mu_1 - \mu_2$.

```
> t.test( x, y, var.equal=TRUE, conf.level=0.9 ) # For pooled standard
  deviation.
```

(d) Test at level 0.05, $H_0 : \Delta = 0$ versus $H_a : \Delta \neq 0$.

```
> t.test( x, y, var.equal=TRUE )
```

□

Homework p. 74: Exercise 2.12* (Estimate Δ by hand and by `wilcox.test`; compute C.I.)

Hints for homework exercise 2.12*: Make sure that you define what you are estimating; i.e, clearly define Δ. Either retype your answers as a comment using "#", or highlight the answers in yellow. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.12 </span>`.

2.7 Scoring Systems

Example: Suppose we have observations {35, 47, 15, 21, 96, 34, 52}. Then, their respective **ranks** are {4, 5, 1, 2, 7, 3, 6}.

These **ranks** may be viewed as **scores** of the original data set.

□

Suppose the data are believed to be from some particular distribution.

How should the scores be selected?

⊙ Suppose a data set of size N is believed to be from a Uniform$(0, N+1)$ distribution.

```
> plotDist( "dunif", 0, 11 ) # For a sample of size 10.
```

On average, what is the expected value (mean) of the smallest observation?

```
> x = replicate( 1e5, min( runif( 10, 0, 11 ) ) ) ;   mean( x )
```

On average, what is the expected value (mean) of the 2nd smallest observation?

On average, what is the expected value (mean) of the 3rd smallest observation?

Hence, if we **score** the original observations under the assumption of a Uniform$(0,\ N+1)$ distribution, then we are effectively **ranking** the data.

Therefore, a **permutation** test (for two populations), based on **uniform scores** and the difference between means or the sample sum of the scores from group 1, is equivalent to a **Wilcoxon rank-sum** test.

Thus, when using the Wilcoxon rank-sum test, the **uniform** distribution is the most appropriate distribution.

2.7.1 Three Common Scoring Systems

Normal Scores

Instead of **scoring** the observations according to **ranks** or the expected value (mean) of the ordered statistics from a **uniform** distribution, we could **score** the observations according to the expected value (mean) of the ordered statistics from a **normal** distribution, to obtain the **normal scores**.

Normal scores are often used for constructing Q-Q plots.

Using **normal scores** with the **permutation test** is reasonable for populations which are approximately **normal**.

For 10 observations (with no ties), the normal scores (based on $N(0,1)$) are -1.539, -1.001, -0.656, -0.376, -0.123, 0.123, 0.376, 0.656, 1.001, 1.539.

```
> mean( replicate( 1e5, min( rnorm( 10 ) ) ) ) # N = 10
```

Van der Waerden Scores

When the data are believed to be approximately normal, an alternative to using **normal scores** is using **van der Waerden scores**, which are based on the *quantiles* of a $N(0,1)$ distribution.

Specifically, for sample size N, these quantiles (with no ties) correspond to the situation where the standard normal cdf is equal to $1/(N+1)$, $2/(N+1)$, $3/(N+1)$, ..., $N/(N+1)$.

General example (without specific observations): Suppose we have *nine* ordered observations (with no ties), where we believe that the observations are from a normal population, perhaps under a null hypothesis.

(a) Determine the **van der Waerden scores**.

```
> N = 9
```

```
> p = 1 :  N  / ( N + 1 ) # These are the values of the cdf.
```

```
> q = qnorm( p )
```

(b) Observe some of the **van der Waerden scores** graphically.

```
> shadeDist( q[ 1 ] ) # The default distribution is "dnorm".
```

□

The textbook lists the **van der Waerden scores** for 12 ordered observations (with no ties) in table 2.7.1 on p. 51.

> qnorm(1:12 / 13) # These results should match table 2.7.1.

Exponential or Savage Scores

Recall: If the lifetime of something is *memoryless* and continuous, then the lifetime has an **exponential** distribution.

When the data are believed to be approximately **exponential**, we could **score** the observations according to the expected value (mean) of the ordered statistics from an **exponential** distribution, to obtain the **exponential scores**.

Letting N be the sample size, these **exponential scores** are $1/N$, $1/N + 1/(N-1)$, $1/N + 1/(N-1) + 1/(N-2)$, ...

```
> mean( replicate( 1e5, min( rexp( 10 ) ) ) ) # N = 10
```

The **Savage scores** are the **exponential scores** minus one, and have mean zero.

Hence, tests based on **Savage scores** are equivalent to tests based on **exponential scores**.

General example (without specific observations): Suppose we have *twelve* ordered observations (with no ties), where we believe that the observations are from an exponential population, perhaps under a null hypothesis. Obtain the **exponential scores**.

```
> N = 12
> 1 / N ;  1 / N + 1 / (N-1) ;  1 / N + 1 / (N-1) + 1 / (N-2) # And so
  on.
> cumsum( 1 / N:1 )
```

□

◇

Scoring and permutation tests

Problem #2.7.1 (live vs. TV), problem2.7.1.txt: MacLachlan (1965), "Variations in learning behavior in two social situations according to personality type," unpublished master's thesis at University of California at Berkeley.

In a business administration course, a set of lectures was given televised to one group and live to another. In each case an examination was given prior to the lectures and immediately following them. The differences between the two examination scores for the *women* in the two groups were as follows:

Live: 20.3, 23.5, 4.7, 21.9, 15.6, 20.3, 26.6, 21.9, −9.4, 4.7, −1.6, 25.0

TV: 6.2, 15.6, 25.0, 4.7, 28.1, 17.2, 14.1, 31.2, 12.6, 9.4, 17.2, 23.4

Test at level 0.05 whether or not the two groups differ in their (change in) scores.

Let Δ be the population median for the *live* group minus the population median for the *TV* group.

(a) Construct line graphs to view the two data sets.

```
> z = scan2( "problem2.7.1.txt" )
```

(b) Use the **Wilcoxon rank-sum test**.

(c) Use the permutation test based on **means**.

(d) Use the permutation test based on **van der Waerden scores**.

```
> z = score( x, y )
```

```
> perm.test( z$x, z$y )
```

```
> # Alternative method for generating scores.
> z = score( c(x, y) )

> perm.test( z[1:12], z[13:24] )
```

(e) Use the permutation test based on **exponential scores**.

```
> z = score( x, y, expon = TRUE )



> perm.test( z$x, z$y )




> # Alternative method for generating scores.
> z = score( c(x, y), expon = TRUE )

> perm.test( z[1:12], z[13:24] )
```

(f) Use the *t*-test with *unequal* variances.

```
> t.test( x, y )
```

(g) Use the *t*-test with *equal* variances.

```
> t.test( x, y, var.equal = TRUE )
```

□

Homework p. 74: Exercises 2.9, 2.10*, 2.11

Hints for homework exercise 2.10*: State H_0 and H_a. Either retype your *p*-value as a comment using "#", or highlight the *p*-value in yellow. State the conclusion both in statistical terms and in regular English. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.10 </span>`.

2.8 Tests for Equality of Scale Parameters and an Omnibus Test

Idea: Based on samples from two populations, we compare the **spreads** of the two populations.

Example: According to the rules of the United States Tennis Association: "The [tennis] ball shall ... have a mass of more than 56.0 grams and less than 59.4 grams."

Suppose a manufacturer produces tennis balls whose median mass is 58 grams, but variability in individual masses is quite large.

Would these tennis balls conform to the standards of the United States Tennis Association?

□

Sample mutually independent observations X_i, $i = 1, \ldots, m$, and Y_j, $j = 1, \ldots, n$, from the models,

$$X_i = \widetilde{\mu} + \widetilde{\sigma}_1 \varepsilon_{ix} \quad \text{and} \quad Y_j = \widetilde{\mu} + \widetilde{\sigma}_2 \varepsilon_{jy},$$

such that the εs are identically distributed with median 0.

Note that $\widetilde{\mu}$ is the same for the two distributions.

What does $\widetilde{\mu}$ represent?

What do $\widetilde{\sigma}_1$ and $\widetilde{\sigma}_2$ represent?

Plot two Cauchy distributions with common median 50 but different scales, 5 and 10.

```
> plotDist( "dcauchy", 50, 5, "dcauchy", 50, 10 )
```

GOAL: Test $H_0 : \widetilde{\sigma}_1 = \widetilde{\sigma}_2$ versus a one-sided or two-sided alternative.

Siegel-Tukey test (section 2.8.1)

(1) Order all $m + n$ observations from smallest to largest.

(2) Assign rank #1 to the **smallest** observation, rank #2 to the **largest** observation, rank #3 to the **2nd largest** observation, rank #4 to the **2nd smallest** observation, rank #5 to the **3rd smallest** observation, rank #6 to the **3rd largest** observation, rank #7 to the **4th largest** observation, and so on.

Note: The sample with the **smallest** ranks tends to have **larger** variability than the sample with the **largest** ranks.

(3) Apply the Wilcoxon rank-sum test, replacing the original values of X and Y by their assigned ranks. □

Example: Assume the above models for X_i and Y_j. Use the Siegel-Tukey test to test at level $\alpha = 0.1$ for equality of the two **scale** parameters versus the alternative that the **scale** parameter of population A is **smaller** than the **scale** parameter of population B, based on the following data.

Sample A	57	45		
Sample B	60	20	80	40

(a) Perform calculations using neither the function `wilcox.test` nor the function `siegel.test`.

```
> x = c( 57, 45 )
> y = c( 60, 20, 80, 40 )
```

```
> # Step 1:  Order the observations from smallest to largest.
> sort( c( x, y ) )

> # Step 2:  Assign the ranks.

> # Apply the Wilcoxon rank-sum test (using hand-calculations),
  replacing the original values of X and Y by their assigned ranks.
```

How many combinations will be used in the Wilcoxon rank-sum test?

Permuted samples	Sample A		Sample B				W_1
1	1	2	3	4	5	6	3
2	1	3	2	4	5	6	4
3	1	4	2	3	5	6	5
4	1	5	2	3	4	6	6
5	1	6	2	3	4	5	7
6	2	3	1	4	5	6	5
7	2	4	1	3	5	6	6
8	2	5	1	3	4	6	7
9	2	6	1	3	4	5	8
10	3	4	1	2	5	6	7
11	3	5	1	2	4	6	8
12	3	6	1	2	4	5	9
13	4	5	1	2	3	6	9
14	4	6	1	2	3	5	10
15	5	6	1	2	3	4	11

Which permuted sample corresponds to the original data set?

Determine the p-value for this one-sided test.

Perform the two-sided test at level 0.1.

(b) Perform the **one**-sided test using the function `wilcox.test` but not the function `siegel.test`.

```
> wilcox.test( c(5, 6), c(1, 2, 3, 4), "greater" )
```

(c) Perform the **one**-sided test using the function `siegel.test`.

```
> ?siegel.test
```

```
> siegel.test( x, y, "less" )
```

□

Suppose that with the **Siegel-Tukey** test, the ranks were assigned beginning with the **largest** observation rather than the **smallest** observations.

Would we obtain the same p-value?

Example 2.8.1, p. 53, table2.8.1.txt: The amount of soda dispensed into 16-ounce bottles might be correctly centered at 16 ounces. However, if variability is large, then some bottles would be overfilled while others would be underfilled.

Table 2.8.1 (below) contains data on the amounts of liquid in randomly selected 16-ounce beverage containers before and after the filling process has been repaired. Use the Siegel-Tukey test to test at level $\alpha = 0.05$ whether or not the repairs were successful. Assume the above models for X_i and Y_j. In other words,

$$X_i = \widetilde{\mu} + \widetilde{\sigma}_1 \varepsilon_{ix} \quad \text{and} \quad Y_j = \widetilde{\mu} + \widetilde{\sigma}_2 \varepsilon_{jy},$$

such that the εs are identically distributed with median 0, where all observations are mutually independent.

Treatment 1 (before process repair)				
16.55	15.36	15.94	16.43	16.01
Treatment 2 (after process repair)				
16.05	15.98	16.10	15.88	15.91

(a) State the null and alternative hypotheses.

(b) Perform the hypothesis test, without using the function `siegel.test`.

```
> z = scan2( "table2.8.1.txt" )
> x = z[ 1 :  5 ]
> y = z[ 6 :  10 ]
> sort( z )
```

[1]	15.36	15.88	15.91	15.94	15.98	16.01	16.05	16.10	16.43	16.55
rank	1	4	5	8	9	10	7	6	3	2

What are the ranks associated with treatment #1?

What are the ranks associated with treatment #2?

```
> wilcox.test( c( 1, 2, 3, 8, 10 ), c( 4, 5, 6, 7, 9 ), "less" )
```

(c) Perform the hypothesis test, using the function `siegel.test`.

```
> siegel.test( x, y, "greater" )
```

(d) Perform the hypothesis test, reversing the rankings (e.g., assigning rank #1 to the largest observation, and so on), without using the function `siegel.test`.

[1]	15.36	15.88	15.91	15.94	15.98	16.01	16.05	16.10	16.43	16.55
rank	2	3	6	7	10	9	8	5	4	1

What are the ranks associated with treatment #1?

What are the ranks associated with treatment #2?

```
> wilcox.test( c( 1, 2, 4, 7, 9 ), c( 3, 5, 6, 8, 10 ), "less" )
```

(e) Perform the hypothesis test, reversing the rankings (e.g., assigning rank #1 to the largest observation, and so on), using the function `siegel.test`.

```
> siegel.test( x, y, "greater", TRUE )
```

□

Which method is preferred, assigning rank #1 to the smallest observation or to the largest observation?

To overcome this ambiguity, one may use the **Ansari-Bradley** test below.

Ansari-Bradley test (section 2.8.1)

The **Ansari-Bradley** ranks are the averages the Siegel-Tukey ranks from the **forward** direction (assigning rank #1 to the **smallest** observation) with the Siegel-Tukey ranks from the **reverse** direction (assigning rank #1 to the **largest** observation).

However, the new rank-sum does not follow the Wilcoxon distribution.

Instead of deriving this distribution, we will simply use the function in *R*.

```
> ?ansari.test
```

```
> ansari.test( x, y, "greater" )
```

Homework p. 74: Exercise 2.15a

2.8.2 Tests for Deviances

This model allows for different *location* parameters, unlike the model from section 2.8.1 where the Siegel-Tukey and Ansari-Bradley tests were used.

Let

$$X_i = \widetilde{\mu}_1 + \widetilde{\sigma}_1 \varepsilon_{ix} \quad \text{and} \quad Y_j = \widetilde{\mu}_2 + \widetilde{\sigma}_2 \varepsilon_{jy},$$

such that the εs are identically distributed with median 0, where all observations are mutually independent.

The deviances for the data set are $X_i - \widetilde{\mu}_1$ and $Y_j - \widetilde{\mu}_2$.

Since $\widetilde{\mu}_1$ and $\widetilde{\mu}_2$ typically are unknown, how should we estimate them?

First, define $X_i^* = X_i - \hat{\widetilde{\mu}}_1$ and $Y_j^* = Y_j - \hat{\widetilde{\mu}}_2$, where $\hat{\widetilde{\mu}}_1$ and $\hat{\widetilde{\mu}}_2$ are the sample medians of the **original** data sets.

The test statistic is based on the *ratio of mean (absolute value of) deviances* (**RMD**), and is estimated by the following:

$$\widehat{RMD} = \frac{\sum_{i=1}^{m} |X_i^*|/m}{\sum_{j=1}^{n} |Y_j^*|/n}$$

To determine **one**-sided p-values, the observed value of $\widehat{RMD}$ is compared to the

values of $\widehat{RMD}$ under either *all* permutations or a *large number* of *simulated* permutations of X_i^* and Y_j^*.

To determine **two**-sided p-values, the test statistic is defined as the following:

$$\widehat{RMD}_{2-sided} = \frac{\max\left(\sum_{i=1}^m |X_i^*|/m,\ \sum_{j=1}^n |Y_j^*|/n\right)}{\min\left(\sum_{i=1}^m |X_i^*|/m,\ \sum_{j=1}^n |Y_j^*|/n\right)},$$

and p-values are determined by the **right** tail only.

Problem 2.8.1 (diabetes), problem2.8.1.txt: In a study designed to determine whether middle-aged and old subjects with maturity-onset diabetes respond to exercise by producing high levels of fasting serum growth hormone, A. P. Hansen (1973, *Diabetes*) collected the following data regarding hormone level (in nanograms per milliliter).

Assume the model:

$$X_i = \widetilde{\mu}_1 + \widetilde{\sigma}_1\varepsilon_{ix} \quad \text{and} \quad Y_j = \widetilde{\mu}_2 + \widetilde{\sigma}_2\varepsilon_{jy},$$

such that the εs are identically distributed with median 0, where all observations are mutually independent.

We are interested in testing at level 0.05 whether or not the scale parameters are equal.

Controls (X)	1.4	1.6	1.4	4.1	2.6	1.1	0.4	1.8
	2.2	0.3	1.3	1.7	1.0	1.2	1.4	0.5
	1.1	1.5	1.1	3.3	2.6	0.7	0.1	1.6
	2.5	0.7	1.7	0.3	1.9	0.0	0.5	
Diabetics (Y)	1.2	0.2	0.3	0.9	4.2	0.9	0.3	0.7
	0.9	1.1	3.0	0.9	2.3	1.3	0.2	1.2
	1.5	2.1	7.7	20.0	1.2	3.4	2.2	0.1
	4.3	0.7	0.7	1.3	1.3	9.8	0.9	4.7
	0.0	0.4	21.0	12.0	4.2	2.7	1.7	0.5
	1.0	0.9	2.1	0.1	1.7	1.0	3.9	1.0
	0.5	0.7	0.2	0.9	0.9	0.8	0.5	1.5
	1.1	1.1	1.6	1.5	4.0	4.7	0.9	

(a) Construct a line graph of X and a line graph of Y.

```
> z = scan2( "problem2.8.1.txt" )
> x = z[ 1 :  31 ]
> y = z[ 32 :  94 ]
```

(b) Determine the value of the $\widehat{RMD}$ test statistic (not the p-value), without using the function `rmd.test`.

```
> rmd.one.sided = mean( abs( x - median(x) ) ) / mean( abs( y -
  median(y) ) )
```

```
> rmd.two.sided = 1 / rmd.one.sided
```

(c) Describe verbally how the p-value is obtained, without using the function `rmd.test`.

(d) Obtain the p-value.

```
> ?rmd.test
```

```
> rmd.test( x, y )
```

□

Suppose the εs are approximately normally distributed.

Which **parametric** test typically is used on the **scale** parameters?

Is such a test valid for our *diabetes* example?

Homework p. 74: Exercise 2.15b

2.8.3 Kolmogorov-Smirnov Test

The **Kolmogorov-Smirnov** test is an **omnibus** test; i.e., the null hypothesis is that the two distributions of interest are the *same* versus the alternative hypothesis that the two distributions are *different*.

Plot the probability density functions of $N(90,\ 20)$ and Cauchy(150, 30) random variables.

```
> plotDist( "dnorm", 90, 20, "dcauchy", 150, 30 )
```

Example: Suppose a huge sample is obtained from a $N(0, 1)$ distribution, and another huge sample is obtained from a Laplace$(0, 1)$ distribution, such that all observations are mutually independent.

Would the two-sided t-test for equality of means or the F-test for variances be powerful in detecting that the two distributions differ; i.e., will these two tests typically produce small p-values?

Plot the *probability density functions* (pdfs) of the $N(0, 1)$ and Laplace$(0, 1)$ random variables.

```
> plotDist( "dnorm", 0, 1, "dlaplace" )
```

Plot the *cumulative distribution functions* (cdf*s*) of the $N(0,1)$ and Laplace$(0,1)$ random variables.

```
> plotDist( "pnorm", 0, 1, "plaplace" )
```

The largest (vertical) difference between the two cdf*s* will be estimated by the **Kolmogorov-Smirnov** *test statistic*.

□

How may we estimate the *cdf* of a population, when given a data set?

The **Kolmogorov-Smirnov** *test statistic* is the maximum vertical difference between the two *empirical* cdf*s*.

Hence, the **Kolmogorov-Smirnov** test statistic is

$$\text{K-S} = \max_{w} |\hat{F}_1(w) - \hat{F}_2(w)|,$$

where $\hat{F}_1(w)$ and $\hat{F}_2(w)$ are the empirical cdf*s* of the two populations of interest.

To determine the p-value, the (original) Kolmogorov-Smirnov test statistic is compared to the values of K-S under either *all* permutations or a *large number* of *simulated* permutations.

Example: Mutually independent observations are sampled from two populations,

such that the first sample is {42, 60, 12, 23} and the second sample is {31, 56, 4, 85, 77}.

(a) Construct the empirical cdf*s* on the same graph.

```
> x = c( 42, 60, 12, 23 )
> y = c( 31, 56, 4, 85, 77 )
```

(b) Determine the Kolmogorov-Smirnov **test statistic**, without using the function `ks.test`.

□

(c) Determine the Kolmogorov-Smirnov **test statistic** and the *p*-value, *using* the function 'ks.test'.

```
> ?ks.test
> ks.test( x, y )
```

□

Revisit problem 2.8.1 (diabetes), problem2.8.1.txt: Test at level 0.05 whether or not the *diabetics* and *control* populations differ in terms of fasting serum growth hormone levels after exercise.

```
> z = scan2( "problem2.8.1.txt" )
```

```
> x = z[ 1 :  31 ]

> y = z[ 32 :  94 ]
```

```
> ks.test( x, y ) # Keep 'alternative' set to 'two.sided'.
```

□

Example: *Normal(0,1) vs. Laplace(0,1) distributions*

□

Homework p. 74: Exercises 2.13, 2.14* (use $m = 3$, $n = 2$, and no ties)

Hints for homework exercise 2.14*: The permutation distribution consists of all possible values of the Kolmogorov-Smirnov test statistic, without duplication, in order from smallest to largest, along with the associated probabilities, which must sum to unity. Clearly show how you are calculating the Kolmogorov-Smirnov test statistic for each of the ten permutations. Hence, show **all** of your *R*-code and output for each of the ten permutations. The homework exercise does **NOT** ask you to compute a *p*-value. Introduce the question number as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise 2.14 </span>`.

2.9 Selecting Among Two-Sample Tests

Recall: **Power** is defined to be the probability of rejecting H_0 given that H_a is true.

For which distributions do we prefer the t-test over nonparametric tests?

For which distributions do we prefer nonparametric tests over the t-test?

Recall from section 1.3.2 an example where the binomial test was **more** powerful than the t-test for Laplace alternatives, but **less** powerful for normal alternatives.

In this section we assume two populations which are identical except for possibly the location parameter; i.e., the cdf*s* satisfy $F_1(x) = F_2(x - \Delta)$.

Test $H_0 : \Delta = 0$ versus a one-sided or two-sided alternative.

2.9.1 The *t*-Test

When distributions F_1 and F_2 are allowed to differ only by the location parameter (but have equal positive finite variances), the t-test we consider is based on the *pooled* sample standard deviation.

When the two populations are **normal** and have equal variances, the pooled t-test (for level α) is the **most powerful** among all tests with level no larger than α, for *all* sample sizes (for one-sided tests).

Is the pooled t-test valid for **nonnormal** populations with equal positive finite variances and large sample sizes?

When the two populations are **nonnormal** but have equal positive finite variances, is the pooled t-test (for level α) the **most powerful** among all tests with level no larger than α, for *large* sample sizes (for one-sided tests)?

2.9.2 The Wilcoxon Rank-Sum Test versus the *t*-Test

The textbook compares the powers of the Wilcoxon rank-sum test and t-test for different sample sizes and alternative distributions.

The *small* sample sizes are for $m + n = 12$.

The *moderate* sample sizes are from $m + n = 36$ to $m + n = 108$.

Construct graphs to show how the *normal* pdf compares with the *uniform*, *Laplace*, and *exponential* pdfs, using identical means and identical standard deviations.

```
> plotDist( "dunif", -1, 1, "dnorm", 0, 1/sqrt(3) )
```

```
> plotDist( "dlaplace", 0, 1, "dnorm" )
```

```
> plotDist( "dexp", 1, , "dnorm", 1 )
```

Recall when using `qqnorm` the difficulty in distinguishing between normal and exponential data for small sample sizes, specifically for $n = 7$.

Textbook shows comparisons of powers in table 2.9.1.

Conclusions:

⊙ When the alternative distribution is **uniform**, the t-test tends to be **more** powerful than the Wilcoxon rank-sum test for **small** and **moderate** sample sizes.

⊙ When the alternative distribution is **Laplace**, the t-test tends to be **more** powerful than the Wilcoxon rank-sum test for **small** sample sizes but **less** powerful for **moderate** sample sizes.

⊙ When the alternative distribution is **exponential**, the t-test tends to be somewhat **equal** in power to the Wilcoxon rank-sum test for **small** sample sizes but **much less** powerful for **moderate** sample sizes.

When the alternative distribution is Cauchy, which test is better, the t-test or the Wilcoxon rank-sum test?

2.9.3 Relative Efficiency

Instead of *small* or *moderate* sample sizes, we now consider *large* sample sizes.

The textbook again discusses hypothesis testing on Δ, and defines **asymptotic efficiency** to compare two tests with certain sample sizes.

Definition: Let $(m_t + n_t)$ be the sample size required for the two-sample t test to achieve the same power as the two-sample Wilcoxon rank-sum test with a sample size of $(m_W + n_W)$, for large sample sizes. The **asymptotic efficiency** of the Wilcoxon rank-sum test to the t-test is $(m_t + n_t)/(m_W + n_W)$.

Table 2.9.2, p. 63

Distribution	*Efficiency*
Normal	0.955
Uniform	1.0
Laplace	1.5
Exponential	3.0
Cauchy	∞

Conclusions:

⊙ When the alternative distribution is **normal**, a t-test with sample size 955 has approximately the same power as a Wilcoxon rank-sum test with sample size 1,000.

⊙ When the alternative distribution is **uniform**, a t-test with sample size 1,000 has approximately the same power as a Wilcoxon rank-sum test also with sample size 1,000.

⊙ When the alternative distribution is **Laplace**, a t-test with sample size 1,500 has approximately the same power as a Wilcoxon rank-sum test with sample size 1,000.

⊙ When the alternative distribution is **exponential**, a t-test with sample size 3,000 has approximately the same power as a Wilcoxon rank-sum test with sample size 1,000.

⊙ When the alternative distribution is **Cauchy**, a t-test with any (arbitrarily large)

sample size has less power than a Wilcoxon rank-sum test with sample size 1,000.

2.9.4 Power of Permutation Tests

Analysis of table 2.9.3: permutation test vs. *t*-test

Here we compare the **permutation test** based on the difference between two **means** with the t-test for **normal** alternatives.

What is the test statistic associated with the **permutation test**?

What is the test statistic associated with the t-test with *pooled* sample standard deviation?

Which test statistic is heavily influenced by outliers?

Which test statistic should perform well when the alternative distribution is *normal*.

When the two populations are **normal** and have equal variances, the pooled t-test (for level α) is the **most powerful** among all tests with level no larger than α, for *all* sample sizes for one-sided tests (as already mentioned in section 2.9.1).

For large sample sizes (and finite population variances), what is the approximate distribution of the **permutation** statistic?

For large sample sizes, what is the approximate distribution of the t-statistic?

Table 2.9.3 compares the power of the **permutation test** with the pooled t-test under **normal** alternatives with $m = n = 10$ (or 20).

Even for these small sample sizes, the t-test is only slightly more powerful than the **permutation test**.

Analysis of table 2.9.4: permutation testS vs. Wilcoxon-rank sum test

Here we compare the **permutation test** based on the difference between two **means** (or two **medians**) with the Wilcoxon rank-sum test, under **normal**, **Laplace**, and **Cauchy** alternatives, for small sample sizes $m = n = 10$.

Recall from section 2.9.2: When comparing the Wilcoxon rank-sum test to the t-test, which is more powerful?

First, consider means.

⊙ A permutation test based on the difference between two **means** is similar to which famous test?

⊙ For **Laplace** and **Cauchy** alternatives with $m = n = 10$ (or 20), which typically is more powerful, the **permutation test** based on the difference between two **means** or the **Wilcoxon** rank-sum test?

⊙ For **normal** alternatives with $m = n = 10$ (or 20), which is more powerful, the **permutation test** based on the difference between two **means** or the **Wilcoxon** rank-sum test?

Next, consider medians.

⊙ Which *nonparametric* test statistic is more susceptible to outliers, the **permutation test** based on the difference between two **medians** or the **Wilcoxon** rank-sum test?

⊙ For **Laplace** and **Cauchy** alternatives with $m = n = 10$ (or 20), which typically is more powerful, the **permutation test** based on the difference between two **medians** or the **Wilcoxon** rank-sum test?

⊙ For **normal** alternatives with $m = n = 10$ (or 20), which typically is more powerful, the **permutation test** based on the difference between two **medians** or the **Wilcoxon** rank-sum test?

2.10 Large-Sample Approximations

2.10.1 Sampling Formulas

When sampling a large number of independent observations from a population with positive finite variance, what is the *approximate* distribution of the **sample mean**?

When sampling a large number of independent observations from a population with positive finite variance, what is the *approximate* distribution of the **sample sum**?

2.10.2 Application to the Wilcoxon Rank-Sum Test

When performing the Wilcoxon rank-sum test, what is the test statistic?

Is the variance (or standard deviation) of these ranks finite?

Are these ranks independent?

Letting $N = m + n$, what is the average rank?

What is the *mean* of the Wilcoxon rank-sum statistic, W (under the null hypothesis that the two *continuous* populations are the same)?

The *variance* of the Wilcoxon rank-sum statistic (under the null hypothesis that the two *continuous* populations are the same) is $\sigma_W^2 = mn(N+1)/12$ (need not memorize), as derived in the textbook.

How do we obtain a p-value based on the asymptotic distribution of W under H_0?

Improvement: Use a continuity correction since W, an integer, is being approximated by a continuous (in fact, normal) distribution.

This normal approximation is fairly accurate even for the small sample sizes of $m = n = 6$, as shown in table 2.10.1, p. 67.

Exercise 2.18, p. 75, exercise2.18.txt: A biologist examined the effect of a fungal infection on the eating behavior of rodents. Infected apples were offered to a group of eight rodents, and sterile apples were offered to a group of four. The amounts consumed (grams of apple/kilogram of body weight) are listed in the table. Assume that the two populations of eating behavior may differ only by a location parameter. Using the Wilcoxon rank sum test, we wish to test at level 0.05 whether or not these two location parameters are equal.

Experimental Group	11	33	48	34	112	369	64	44
Control Group	177	80	141	332				

(a) Compute the exact p-value.

```
> v = scan2( "exercise2.18.txt" )
> x = v[ 1 :  8 ]
> y = v[ 9 :  12 ]
```

(b) Using hand-calculations (i.e., you may use R, but not `wilcox.test`), determine the asymptotic p-value based on the normal approximation with continuity correction.

```
> rank( c( x, y ) )
> W = sum( rank( c( x, y ) )[1:8] )

> m = length( x )

> n = length( y )

> N = m + n
> mu = m * (N+1) / 2 # population mean of W under Ho

> sigma = sqrt( m * n * (N+1) / 12 ) # population sd of W under Ho

> # Is W (= 41) at the left tail or the right tail?
```

```
> z = (41.5 - mu) / sigma

> p.value = 2 * pnorm( z )



> shadeDist( c(z, -z) ) # Graph in terms of Z.
> shadeDist( c(41.5, 62.5), "dnorm", mu, sigma ) # Graph in terms of W.
```

(c) Using the function `wilcox.test`, determine the asymptotic p-value based on the normal approximation with continuity correction.

□

Homework C2.10.1*: Use the data from exercise 2.4 (i.e., chapter 2, exercise #4), p. 73. Introduce the question number and letter as a comment using "#" or in red using .html code; e.g., `<span style="color: red"> Exercise C2.10.1(a) </span>`.

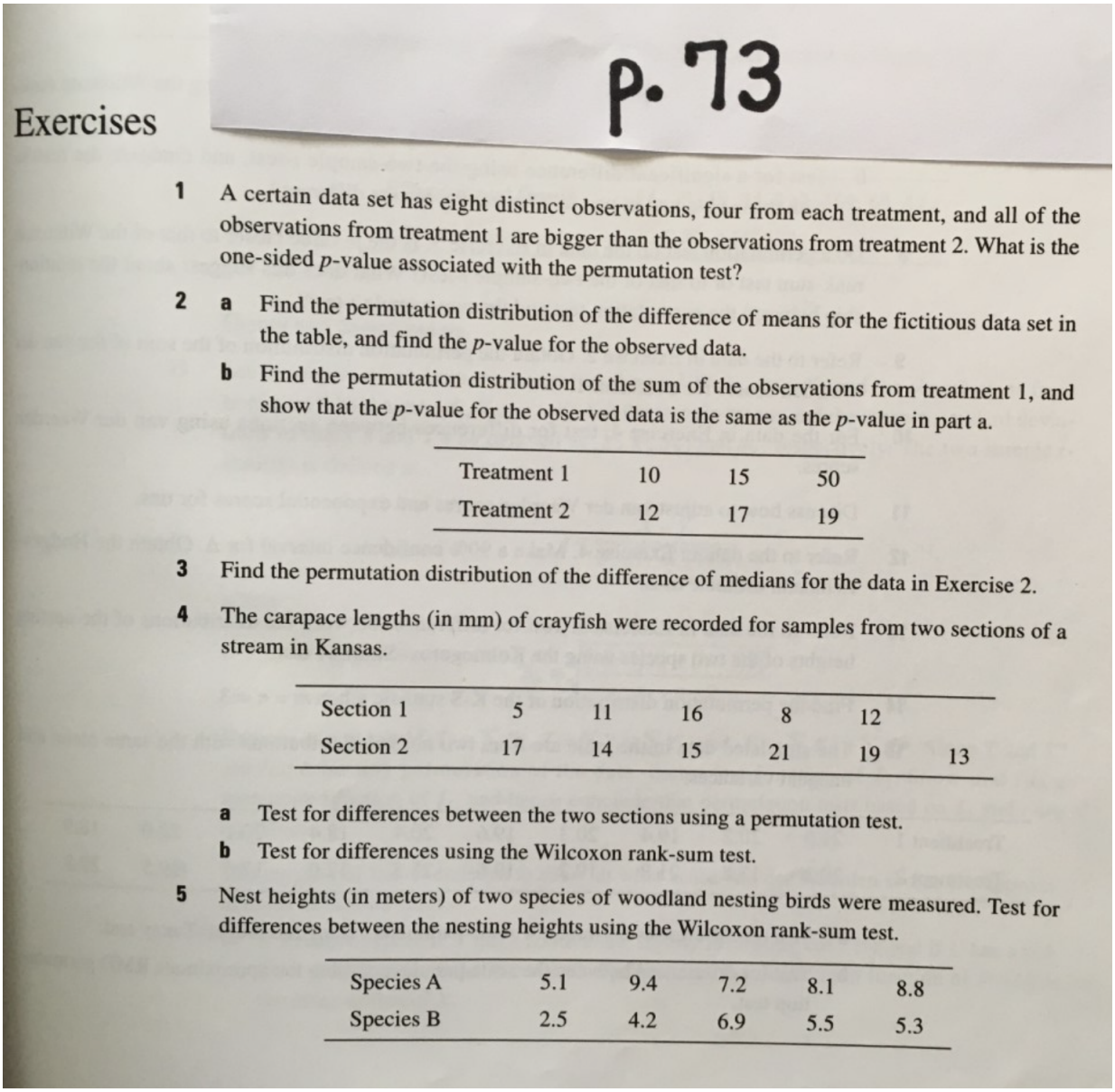
(a) Test for unequal medians using the Wilcoxon rank-sum test, based on the normal approximation with continuity correction, using **ONLY** hand-calculations (i.e., without using the function `wilcox.test`). State H_0, H_a, and conclusion in statistical terms and in regular English, and **define** any notation used. Either retype your p-value as a comment using "#", or highlight the p-value in yellow.

(b) Test for unequal medians using the Wilcoxon rank-sum test, based on the normal approximation with continuity correction, using the function `wilcox.test`. State H_0, H_a, and conclusion in statistical terms and in regular English, and **define**

any notation used. Either retype your p-value as a comment using "#", or highlight the p-value in yellow.

(c) Compare your *approximated* p-value (from parts (a) and (b)) with the *exact* p-value (from exercise 2.4b).

End of Homework C2.10.1*. □



p. 73

Exercises

1 A certain data set has eight distinct observations, four from each treatment, and all of the observations from treatment 1 are bigger than the observations from treatment 2. What is the one-sided p-value associated with the permutation test?

2 **a** Find the permutation distribution of the difference of means for the fictitious data set in the table, and find the p-value for the observed data.

b Find the permutation distribution of the sum of the observations from treatment 1, and show that the p-value for the observed data is the same as the p-value in part a.

Treatment 1	10	15	50
Treatment 2	12	17	19

3 Find the permutation distribution of the difference of medians for the data in Exercise 2.

4 The carapace lengths (in mm) of crayfish were recorded for samples from two sections of a stream in Kansas.

Section 1	5	11	16	8	12	
Section 2	17	14	15	21	19	13

a Test for differences between the two sections using a permutation test.

b Test for differences using the Wilcoxon rank-sum test.

5 Nest heights (in meters) of two species of woodland nesting birds were measured. Test for differences between the nesting heights using the Wilcoxon rank-sum test.

Species A	5.1	9.4	7.2	8.1	8.8
Species B	2.5	4.2	6.9	5.5	5.3

6 Create a fictitious data set where the Wilcoxon rank-sum test and the two-sample t-test lead to different conclusions at the 5% level of significance. (*Hint:* Try a data set in which one treatment has a few very large observations in comparison with all other observations in either treatment.)

7 Students in an introductory statistics class were asked how many brothers and sisters they have and whether their hometown is urban or rural.

Number of Siblings in Rural versus Urban Areas

Rural	3	2	1	1	2	1	3	2	2	2	2	5	1	4	1	1	1	1	6	2	2	2	1	1
Urban	1	0	1	1	0	0	1	1	1	8	1	1	1	0	1	1	2							

a Test for a significant difference between rural and urban areas using the Wilcoxon rank-sum test.

b Test for a significant difference using the two-sample t-test, and compare the results with those obtained in part a. Why are the results different?

8 Do a permutation test on the data in Exercise 7. Is the p-value closer to that of the Wilcoxon rank-sum test or to that of the two-sample t-test? What does this suggest about the relationship between the permutation test and the two-sample t-test?

9 Refer to the data in Exercise 2. Obtain the permutation distribution of the sum of the van der Waerden scores for treatment 1.

10 For the data in Exercise 4, test for differences between sections using van der Waerden scores.

11 Discuss how to adjust van der Waerden scores and exponential scores for ties.

12 Refer to the data in Exercise 4. Make a 90% confidence interval for Δ. Obtain the Hodges–Lehmann estimate of Δ.

13 Refer to the data in Exercise 5. Test for differences between the distributions of the nesting heights of the two species using the Kolmogorov–Smirnov test.

14 Find the permutation distribution of the K-S statistic when $m = n = 3$.

15 The simulated data in the table are from two normal distributions with the same mean and unequal variances.

Treatment 1	21.9	20.2	19.4	20.3	19.6	20.4	18.4	20.1	22.0	18.9
Treatment 2	20.2	13.8	21.8	19.2	19.6	25.5	17.0	17.6	19.5	22.2

a Test for differences between the scale parameters using the Siegel–Tukey test.

b Test for differences between the scale parameters using the approximate RMD permutation test.