Chapter 16

Nonparametric Tests

The statistical tests we have examined so far are called **parametric tests**, because they assume the data have a known distribution, such as the normal, and test hypotheses about the parameters of this distribution. Examples of such tests are the F test in ANOVA, and one- or two-sample t tests. Parametric tests can also be constructed for other distributions, such as the Poisson and binomial.

While ANOVA and other procedures are derived assuming the data are normal, they can also be validly applied to non-normal data provided sample sizes are large, due to the central limit theorem (Glass et al. 1972). For example, the means used in the ANOVA F tests are assumed to have a normal distribution, which will be true for normal data. This will also hold for non-normal data, provided the sample sizes are sufficiently large for the central limit theorem to operate (Chapter 7). Thus, the tests used in ANOVA will still be valid for large sample sizes, regardless of the distribution of the data. Valid in this context means the tests have the correct Type I error rate (such as $\alpha = 0.05$) and power levels.

There are conditions where parametric procedures are less than ideal, such as non-normal data and relatively small sample sizes. We cannot rely on the central limit theorem here, and so parametric tests based on the normal distribution might be invalid. **Nonparametric tests** are often useful in this situation, because they do not assume a particular probability distribution for the data. For this reason they are also known as **distribution-free** methods. Nonparametric tests can be more powerful than parametric tests for non-normal data (Conover 1999; Hollander et al. 2014). The increase in power can be substantial for distributions with heavy tails compared to

the normal distribution, which implies that extreme observations are more common. While nonparametric tests are less powerful than parametric ones for normal data, the loss of power is often quite minimal.

We will examine three types of nonparametric tests for one-way designs. The first are tests based on ranks. These replace the data values with their rank values, obtained by ordering the data from smallest to largest. They then utilize test statistics that are functions of these ranks rather than the original data values. We will cover rank tests for two or more groups, in particular the Wilcoxon and Kruskal-Wallis tests (Conover 1999; Hollander et al. 2014). They are used to test whether the distributions for each group differ in location, and serve a function similar to parametric tests like one-way ANOVA. We will also examine the two-sample Kolmogorov-Smirnov test, which can detect differences in both the shape and location of two distributions (Conover 1999; Hollander et al. 2014). It makes use of the empirical distribution function for each group, the empirical counterpart of the cumulative distribution function for continuous random variables (Chapter 6). The last type of nonparametric test we will consider are randomization tests. These tests examine whether the data are consistent with a null hypothesis of randomness (Hinkelmann & Kempthorne 1994; Manly 1997). The behavior of a test statistic (often a parametric one like an F statistic) is examined under this null hypothesis, in a process that involves randomly permuting or rearranging observations across the groups many times.

We will use data from a study of chitons (a kind of mollusk) in the intertidal zone (Flores-Campaña et al. 2012) to illustrate the use of nonparametric tests. Populations of *Chiton albolineatus* were sampled from three islands in Mazatlan Bay, Mexico. For each island, samples were taken from sites that were exposed or sheltered from wave action, and the body length of the chitons measured. The authors found that the distribution of chiton length was non-normal, and so used the nonparametric Kruskal-Wallis test to compare the lengths of chitons across islands and sites. They found significant differences in length among various combinations of island and site, and a tendency for chiton to be larger in exposed sites. We will use a small subset of these data in our calculations, shown in Tables 16.1 and 16.2.

Table 16.1: Example 1 - Body lengths of *Chiton albolineatus* in the intertidal zone of the island of Venados (Flores-Campaña et al. 2012). Chitons were sampled from sites sheltered or exposed to wave action. Also shown are the rank values (R_{ij}) for each observation, and the sum of the ranks for each groups ($\sum_{j=1}^{n_i} R_{ij}$, where n_i is the sample size for each group.)

Site	Y_{ij} = Length (mm)	R_{ij}	i	j	$\sum_{j=1}^{n_i} R_{ij}$
Sheltered	44.39	20	1	1	
Sheltered	22.30	3	1	2	
Sheltered	21.31	2	1	3	
Sheltered	23.80	5	1	4	
Sheltered	26.23	8	1	5	70
Sheltered	27.98	10	1	6	
Sheltered	28.10	11	1	7	
Sheltered	24.39	6	1	8	
Sheltered	22.32	4	1	9	
Sheltered	15.16	1	1	10	
Exposed	30.20	16	2	1	
Exposed	29.36	14	2	2	
Exposed	28.88	12	2	3	
Exposed	32.23	19	2	4	
Exposed	26.54	9	2	5	140
Exposed	24.85	7	2	6	
Exposed	30.54	17	2	7	
Exposed	31.36	18	2	8	
Exposed	28.98	13	2	9	
Exposed	29.49	15	2	10	

Table 16.2: Example 2 - Body length of *C. albolineatus* on the sheltered side of three islands, located in Mazatlan Bay, Mexico (Flores-Campaña et al. 2012). Also shown are the rank values (R_{ij}) for each observation, and the sum of the ranks for each group ($\sum_{j=1}^{n_i} R_{ij}$)

Site	Y_{ij} = Length (mm)	R_{ij}	i	j	$\sum_{j=1}^{n_i} R_{ij}$
Lobos	23.86	16	1	1	
Lobos	20.20	6	1	2	
Lobos	29.32	27	1	3	
Lobos	23.56	13	1	4	
Lobos	24.32	17	1	5	157
Lobos	22.33	12	1	6	
Lobos	23.69	14	1	7	
Lobos	26.78	21	1	8	
Lobos	27.32	23	1	9	
Lobos	21.22	8	1	10	
Pajaros	32.90	29	2	1	
Pajaros	32.73	28	2	2	
Pajaros	26.94	22	2	3	
Pajaros	29.09	26	2	4	
Pajaros	12.32	1	2	5	142
Pajaros	15.25	5	2	6	
Pajaros	25.87	19	2	7	
Pajaros	20.21	7	2	8	
Pajaros	13.96	3	2	9	
Pajaros	12.48	2	2	10	
Venados	44.39	30	3	1	
Venados	22.30	10	3	2	
Venados	21.31	9	3	3	
Venados	23.80	15	3	4	
Venados	26.23	20	3	5	166
Venados	27.98	24	3	6	
Venados	28.10	25	3	7	
Venados	24.39	18	3	8	
Venados	22.32	11	3	9	
Venados	15.16	4	3	10	

16.1 Wilcoxon two-sample test

The Wilcoxon test provides a nonparametric alternative to a two-sample t test or a one-way ANOVA for two groups (see Chapter 11). It does not assume any particular distribution of the data, except that it is a continuous one (see Chapter 6). The null and alternative hypotheses for the Wilcoxon test are expressed in terms of the cumulative distribution for the two groups, say $F_1(y)$ and $F_2(y)$. Under the null hypothesis the two distribution are supposed to be identical, which can be expressed as $H_0 : F_2(y) = F_1(y)$ for all y (Fig. 16.1). Under the alternative, one distribution is shifted from the other by a distance Δ, but they otherwise have the same shape (Conover 1999; Hollander et al. 2014). This can be expressed as $H_1 : F_2(y) = F_1(y-\Delta)$ (Fig. 16.2).

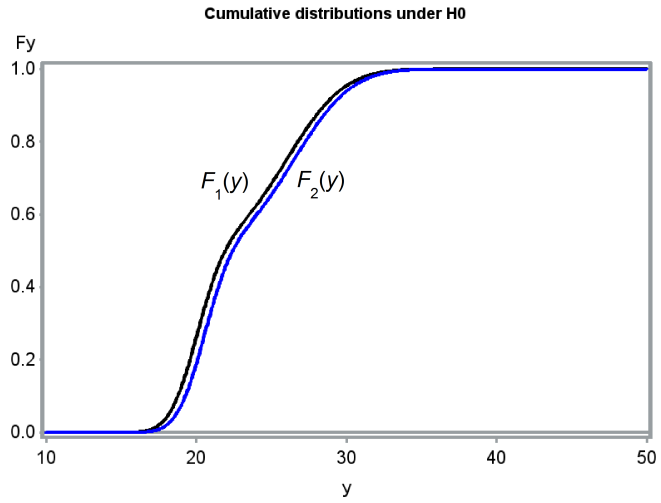


Figure 16.1: Cumulative distributions for two groups under $H_0 : \Delta = 0$.

The Wilcoxon test statistic W is based on the ranks of the observations. The observations are first assigned ranks from the smallest to the largest across the two groups. The test statistic is then the sum of the ranks for one of the groups. Typically the one with the smallest sample size is chosen, or if the sample sizes are equal, one is arbitrarily selected (SAS uses group order). For the Example 1 data the sample sizes are equal, so we could use

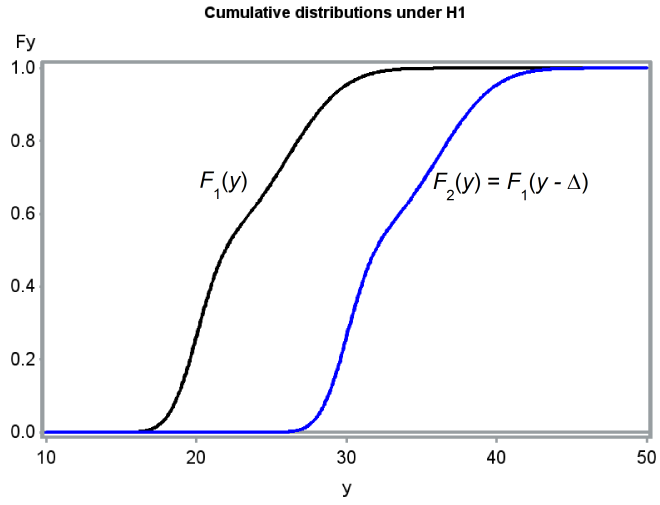


Figure 16.2: Cumulative distributions for two groups under $H_1 : \Delta = 10$.

the summed ranks for the Sheltered chiton group, namely

$$W = \sum_{j=1}^{n_1} R_{1j} = 70 \tag{16.1}$$

(Conover 1999; Hollander et al. 2014). We would expect small values of this statistic when F_1 is located to the left of F_2 ($\Delta > 0$), because this implies that values of Y_{1j} are more likely to be small relative to Y_{2j} ones. Conversely, large values of the statistic would occur when F_1 is to the right of F_2 ($\Delta < 0$). W is also sensitive to differences in the expected values (means) of the two distributions, because of the relationship between expected values and distributions. For a two-tailed test, we would reject H_0 if W is sufficiently large, or sufficiently small. An exact P value for both one- and two-tailed tests can be calculated using the distribution of W. We will let SAS handle the calculations for exact tests.

For large sample sizes, the distribution of W under H_0 approaches the normal distribution with mean and variance given by

$$E_{H_0}[W] = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{16.2}$$

and

$$Var_{H_0}[W] = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}. \tag{16.3}$$

The expected value formula assumes W is calculated using the first group. We then have

$$Z = \frac{W - E_{H_0}[W]}{\sqrt{Var_{H_0}[W]}} \sim N(0,1) \tag{16.4}$$

for large sample sizes. We can use this approximation to find P values for both one- and two-tailed tests (Hollander et al. 2014).

The Wilcoxon statistic W can be derived starting with a two-sample t test (see Chapter 11), and simply replacing the observations with their rank values (Bickel & Doksum 1977). It is also equivalent to the Mann-Whitney U test, another common nonparametric test. Modifications of the Wilcoxon test are also available to deal with the problem of tied observations. The tied observations are assigned the average of the tied ranks, and the variance equation is modified to account for the number of ties (Hollander et al. 2014).

Wilcoxon test - sample calculation

For the Example 1 data, we see that $W = 70$ for the Sheltered chitons (see Table 16.1). We will use the normal approximation for this statistic to obtain a two-tailed P value for the test. We have $E_{H_0}[W] = 10(10+10+1)/2 = 105$ and $Var_{H_0}[W] = 10 \cdot 10(10+10+1)/12 = 175$, and so

$$Z = \frac{70 - 105}{\sqrt{175}} = -2.646. \tag{16.5}$$

From Table Z, we find that $P[Z < -2.646] = 1 - P[Z < 2.646] \approx 1 - 0.9960 = 0.0040$. The two-tailed P value is then twice this value, or $P = 2(0.0040) = 0.0080$.

16.1.1 Wilcoxon test for Example 1 - SAS demo

We now conduct the Wilcoxon test using the Example 1 data and the SAS procedure `npar1way`, which implements a number of nonparametric procedures for one-way (single factor) designs (SAS Institute Inc. 2018). See program listing below. The Wilcoxon test is invoked by adding the `wilcoxon` option in the `proc npar1way` statement. The `class` statement identifies the group variable, while `var` selects the dependent variable. The `exact wilcoxon` line generates exact P values for the test. The program also includes `proc gplot` code to plot the group means (SAS Institute Inc. 2016a). For purposes

of comparison, a one-way ANOVA is also conducted using `proc glm`. See program and output below (Fig. 16.3-16.6).

We see that the Wilcoxon two-tailed test was highly significant, for both the exact test ($W = 70, P = 0.0068$) and the normal approximation ($Z = -2.6080, P = 0.0091$). The value of Z calculated by SAS differs slightly from our earlier result, because it includes a correction that improves the normal approximation. From the summed ranks for each group, as well as the graph, it appears that the Sheltered chitons were smaller than the Exposed ones. Note that the parametric one-way ANOVA for these data was non-significant ($F_{1,18} = 2.13, P = 0.1619$). This likely occurred because of one very large and one small chiton at the Sheltered site, which would be outliers in the ANOVA. In the analysis using ranks, these are simply the largest and smallest rank values, only one step away from the next ones.

SAS Program

```
* WKWtest_chitons_Venados.sas;
title 'Wilcoxon and Kruskal-Wallis tests for chiton length';
data chitons;
    input site :$10. length;
    datalines;
Sheltered   44.39
Sheltered   22.30
Sheltered   21.31
Sheltered   23.80
Sheltered   26.23

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Plot means, standard error, and observations;
proc gplot data=chitons;
    plot length*site / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Kruskal-Wallis/Wilcoxon tests;
proc npar1way wilcoxon data=chitons;
    class site;
    var length;
    exact wilcoxon;
run;
* One-way ANOVA for comparison;
proc glm data=chitons;
    class site;
    model length = site;
run;
quit;
```

Wilcoxon and Kruskal-Wallis tests for chiton length

Obs	site	length
1	Sheltered	44.39
2	Sheltered	22.30
3	Sheltered	21.31
4	Sheltered	23.80
5	Sheltered	26.23
6	Sheltered	27.98
7	Sheltered	28.10
8	Sheltered	24.39
9	Sheltered	22.32
10	Sheltered	15.16
11	Exposed	30.20
12	Exposed	29.36
13	Exposed	28.88
14	Exposed	32.23
15	Exposed	26.54
16	Exposed	24.85
17	Exposed	30.54
18	Exposed	31.36
19	Exposed	28.98
20	Exposed	29.49

Figure 16.3: `WKWtest_chitons_Venados.sas - proc print`

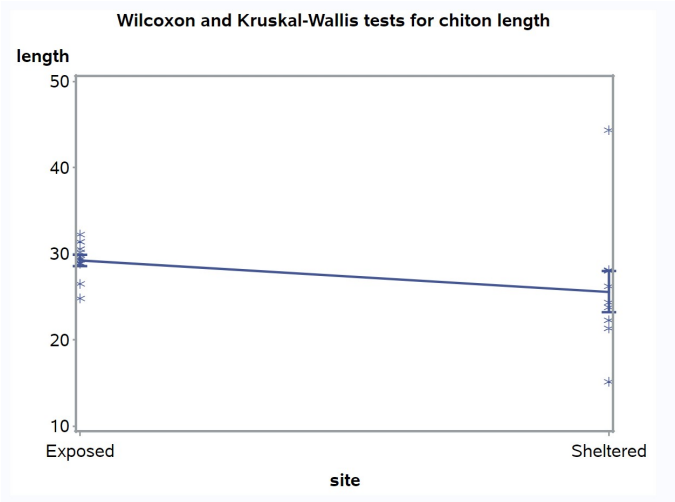
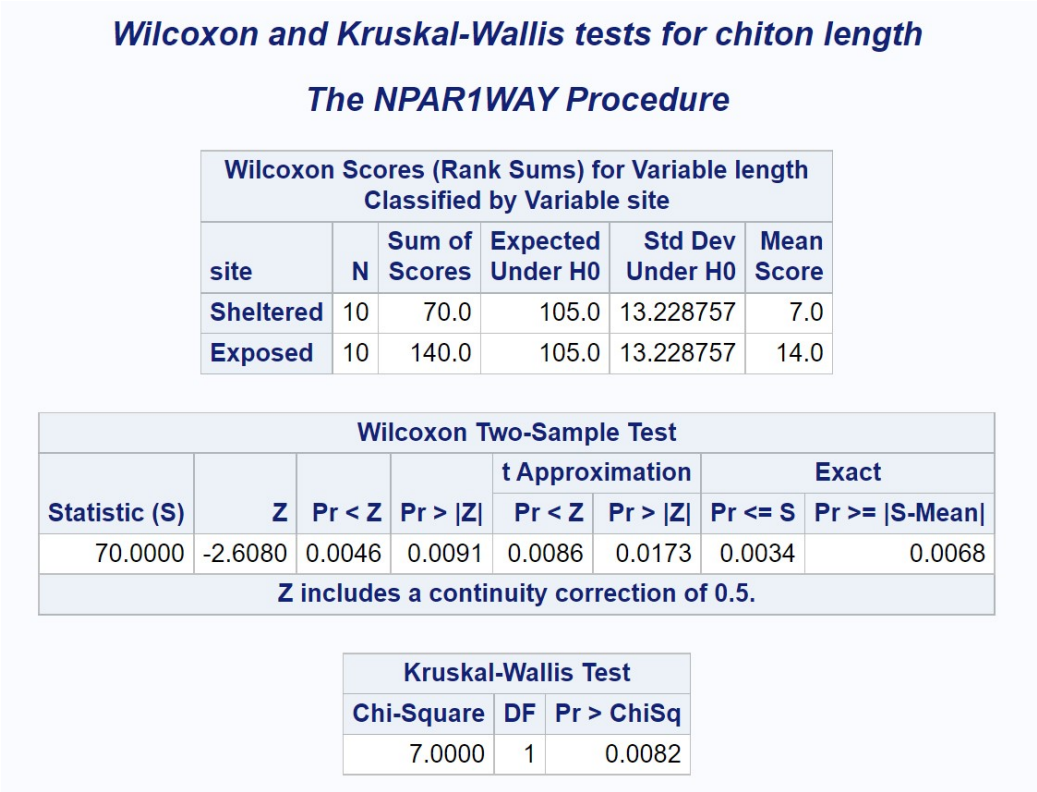


Figure 16.4: `WKWtest_chitons_Venados.sas - proc gplot`



Wilcoxon and Kruskal-Wallis tests for chiton length

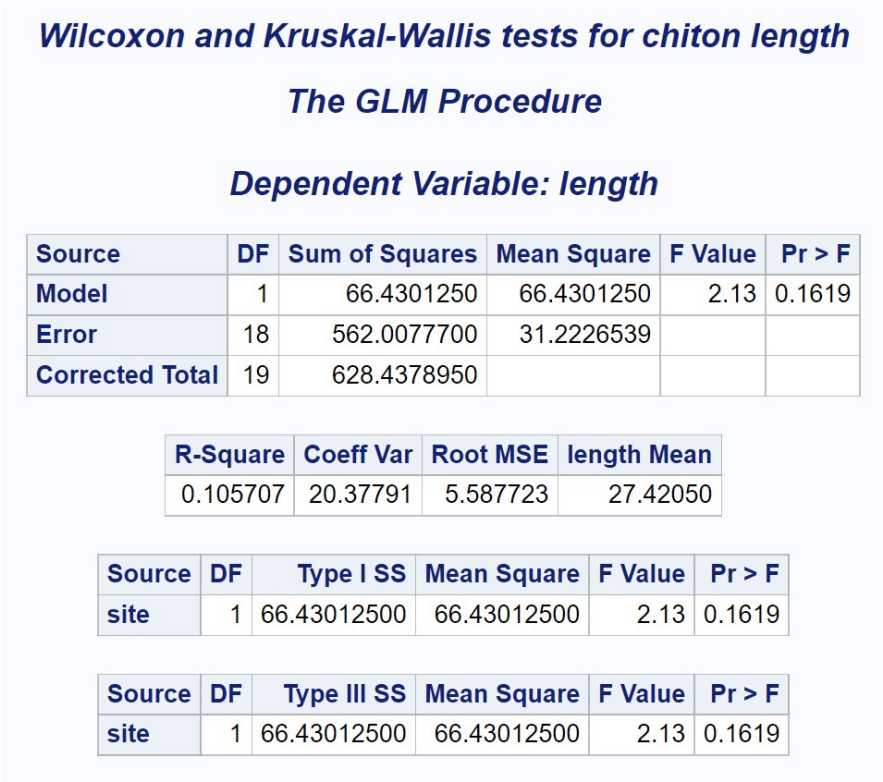
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable length Classified by Variable site					
site	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Sheltered	10	70.0	105.0	13.228757	7.0
Exposed	10	140.0	105.0	13.228757	14.0

Wilcoxon Two-Sample Test							
				t Approximation		Exact	
Statistic (S)	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|	Pr <= S	Pr >= |S-Mean|
70.0000	-2.6080	0.0046	0.0091	0.0086	0.0173	0.0034	0.0068
Z includes a continuity correction of 0.5.							

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
7.0000	1	0.0082

Figure 16.5: `WKWtest_chitons_Venados.sas - proc npar1way`



Wilcoxon and Kruskal-Wallis tests for chiton length

The GLM Procedure

Dependent Variable: length

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	66.4301250	66.4301250	2.13	0.1619
Error	18	562.0077700	31.2226539		
Corrected Total	19	628.4378950			

R-Square	Coeff Var	Root MSE	length Mean
0.105707	20.37791	5.587723	27.42050

Source	DF	Type I SS	Mean Square	F Value	Pr > F
site	1	66.43012500	66.43012500	2.13	0.1619

Source	DF	Type III SS	Mean Square	F Value	Pr > F
site	1	66.43012500	66.43012500	2.13	0.1619

Figure 16.6: `WKWtest_chitons_Venados.sas - proc glm`

16.2 Kruskal-Wallis test

The Kruskal-Wallis test is an extension of rank methods to one-way designs with three or more groups. The null and alternative hypotheses are similar to the Wilcoxon test, with the cumulative distributions for the different groups the same under H_0, and differing by shift parameters under H_1. The Kruskal-Wallis test is sensitive to these shifts as well as differences among the means of the groups.

The Kruskal-Wallis test statistic H is calculated using the ranks of the observations across all groups. Suppose we have a different groups, and for simplicity assume the same sample size n for each group. The Kruskal-Wallis test statistic is

$$H = \frac{12n}{an(an+1)} \sum_{i=1}^{a} \left(\frac{\sum_{j=1}^{n} R_{ij}}{n} - \frac{an+1}{2} \right)^2 \tag{16.6}$$

(Conover 1999; Hollander et al. 2014). Note that the left term in parentheses is the mean rank for each group, while the right one is the mean rank across all the groups. This implies that H will become large when the mean rank differs among groups, similar to the way differences in the group means affect the F statistic for one-way ANOVA. In fact, the Kruskal-Wallis statistic can be derived from the F test by substituting ranks for the observations (Bickel & Doksum 1977). A more complex form of H is used when sample sizes are unequal, or when there are ties in the data. Under H_0, H has approximately a χ^2 distribution with $a-1$ degrees of freedom.

Kruskal-Wallis test - sample calculation

We will illustrate the Kruskal-Wallis test using both the Example 1 and 2 data sets. For Example 1, we have two groups with ten observations each, so $a = 2$ and $n = 10$. The summed ranks for the two groups are 70 (Sheltered)

and 140 (Exposed). It follows that

$$\begin{aligned} H &= \frac{12 \cdot 10}{2 \cdot 10(2 \cdot 10 + 1)} \left[\left(\frac{70}{10} - \frac{2 \cdot 10 + 1}{2} \right)^2 + \left(\frac{140}{10} - \frac{2 \cdot 10 + 1}{2} \right)^2 \right] \\ &= \frac{120}{420} \left[(7 - 10.5)^2 + (14 - 10.5)^2 \right] \\ &= 0.2857 \left[12.25 + 12.25 \right] \\ &= 7.00. \end{aligned}$$

The degrees of freedom are $a - 1 = 2 - 1 = 1$. From Table C, we find that $P < 0.01$, and so the Exposed and Sheltered chitons were significantly different in length ($H = 7.00, df = 1, P < 0.01$).

The Example 2 data involves chitons collected from three different islands ($a = 3$), with ten chitons sampled per island ($n = 10$). The summed ranks for the three islands are 157, 142, and 166. From this information, we calculate that

$$\begin{aligned} H &= \frac{12 \cdot 10}{3 \cdot 10(3 \cdot 10 + 1)} \\ &\quad \cdot \left[\left(\frac{157}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 + \left(\frac{142}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 + \left(\frac{166}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 \right] \\ &= \frac{120}{930} \left[(15.7 - 15.5)^2 + (14.2 - 15.5)^2 + (16.6 - 15.5)^2 \right] \\ &= 0.129 \left[0.04 + 1.69 + 1.21 \right] \\ &= 0.38. \end{aligned}$$

The degrees of freedom are $a - 1 = 3 - 1 = 2$. From Table C, we find that $P < 0.9$. There was no significant difference in length among the three islands ($H = 0.38, df = 2, P < 0.9$).

16.2.1 Kruskal-Wallis test for Example 1 - SAS demo

The Kruskal-Wallis test is automatically calculated when the `wilcoxon` option for `proc npar1way` is used (see previous output). We see there was a highly significant difference in length betwee the Sheltered and Exposed sites ($H = 7.00, df = 1, P = 0.0082$).

16.2.2 Kruskal-Wallis test for Example 2 - SAS demo

The Kruskal-Wallis test for the Example 2 data is shown below (Fig. 16.9). There was no significant difference in length among the three islands ($H = 0.38, df = 2, P = 0.8272$). Note that an exact version of this test is also provided ($P = 0.8386$).

SAS Program

```
* KWtest_chitons_3islands.sas;
title 'Kruskal-Wallis test for chiton length';
data chitons;
    input island $ length;
    datalines;
Lobos       23.86
Lobos       20.20
Lobos       29.32
Lobos       23.56
Lobos       24.32

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Plot means, standard error, and observations;
proc gplot data=chitons;
    plot length*island / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Kruskal-Wallis/Wilcoxon tests;
proc npar1way wilcoxon data=chitons;
    class island;
    var length;
    exact wilcoxon;
run;
quit;
```

Kruskal-Wallis test for chiton length

Obs	island	length	lengthRank
1	Lobos	23.86	16
2	Lobos	20.20	6
3	Lobos	29.32	27
4	Lobos	23.56	13
5	Lobos	24.32	17
6	Lobos	22.33	12
7	Lobos	23.69	14
8	Lobos	26.78	21
9	Lobos	27.32	23
10	Lobos	21.22	8
11	Pajaros	32.90	29
12	Pajaros	32.73	28
13	Pajaros	26.94	22
14	Pajaros	29.09	26
15	Pajaros	12.32	1

etc.

Figure 16.7: `KWtest_chitons_3islands.sas - proc print`

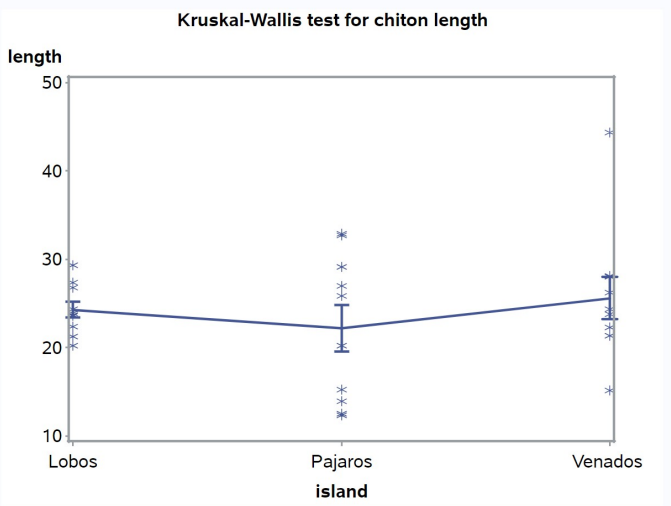


Figure 16.8: `KWtest_chitons_3islands.sas - proc gplot`

Kruskal-Wallis test for chiton length

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable length Classified by Variable island

island	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Lobos	10	157.0	155.0	22.730303	15.70
Pajaros	10	142.0	155.0	22.730303	14.20
Venados	10	166.0	155.0	22.730303	16.60

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq	Exact Pr >= ChiSq
0.3794	2	0.8272	0.8386

Figure 16.9: `KWtest_chitons_3islands.sas - proc npar1way`

16.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a nonparametric procedure used to compare the probability densities of two groups or samples, using their cumulative distributions (see Chapter 6). Let $F_1(y)$ be the cumulative distribution function for the first group, while $F_2(y)$ is the second. The null hypothesis for the Kolmogorov-Smirnov test is $H_0 : F_2(y) = F_1(y)$, which means that the two groups have the same distribution. The alternative hypothesis is $H_1 : F_2(y) \neq F_1(y)$ for some y, implying there is some difference in the distributions, which could involve their location, general shape, variance, and so forth. This is a broader alternative hypothesis than the rank tests we examined earlier, where the distributions had the same shape but differed by location.

The Kolmogorov-Smirnov test statistic is calculated using the empirical distribution functions of the two groups, which estimates the underlying cumulative distribution function. For a sample with n_i observations, the empirical distribution function is defined as

$$G_i(y) = \frac{\text{Number of } Y_{ij} \text{ values} \leq y}{n_i}. \tag{16.7}$$

$G_i(y)$ increases in a step-like fashion as y increases, with a jump occurring at every value of Y_{ij} (Conover 1999; Hollander et al. 2014). Fig. 16.10 shows these functions for the two sites in Example 1. The Kolmogorov-Smirnov test uses the maximum vertical distance between the two functions as the test statistic. The distance is defined using the formula

$$D = \max_y |G_1(y) - G_2(y)| \tag{16.8}$$

(Conover 1999; Hollander et al. 2014). D is the largest distance between $G_1(y)$ and $G_2(y)$ over all values of y, with the absolute value making it a positive quantity. We would then reject H_0 for sufficiently large values of D. The P value for the test can be calculated exactly for small sample sizes, and there is also a large sample approximation for the test. We will let SAS handle the details. This test can also be used when there are ties in the observations, in which case it is conservative, meaning it is less likely to reject H_0 (Hollander et al. 2014).

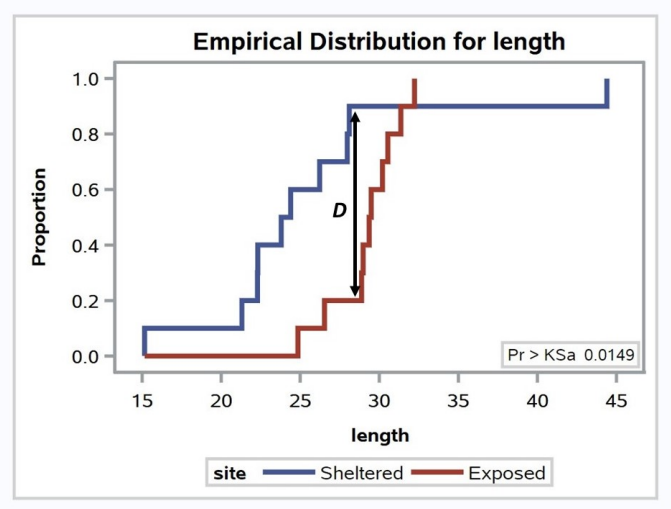


Figure 16.10: Empirical distribution functions for the Example 1 data. Also shown is the maximum value of D for the two samples.

16.3.1 Kolmogorov-Smirnov test for Example 1 - SAS demo

The SAS procedure `npar1way` can also be used for the Kolmogorov-Smirnov test (SAS Institute Inc. 2018). It is invoked by adding the `edf` option in the `proc npar1way` statement (see program below). This option also generates a graph of the empirical distribution function for the two groups (Fig. 16.10). An exact version of test can be calculated using the line `exact ks`. The program also includes `proc gchart` code to generate histograms of the two groups (SAS Institute Inc. 2016a). This seems more appropriate for the Kolmogorov-Smirnov test than plotting the means, because this test can detect differences in both shape and location. Examining the SAS output, we see that $D = 0.7$ (Fig. 16.12). The P value for the exact version of the test was significant ($P = 0.0123$), implying there was some difference in the distributions of the two sites. The graph generated by `proc gchart` suggests they differed in both location and variance (Fig. 16.11).

SAS Program

```
* KStest_chitons_Venados.sas;
title 'Kolmogorov-Smirnov test for chiton length';
data chitons;
    input site :$10. length;
    datalines;
Sheltered   44.39
Sheltered   22.30
Sheltered   21.31
Sheltered   23.80
Sheltered   26.23

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Histograms for the two groups;
proc gchart data=chitons;
    vbar length / group=site axis=axis1 gaxis=axis1 maxis=axis2;
    axis1 label=(height=2) value=(height=2) width=3 minor=none;
    axis2 label=(height=1.5) value=(height=1.5) width=1.5;
run;
* Kolmogorov-Smirnov test;
proc npar1way edf data=chitons;
    class site;
    var length;
    exact ks;
run;
quit;
```

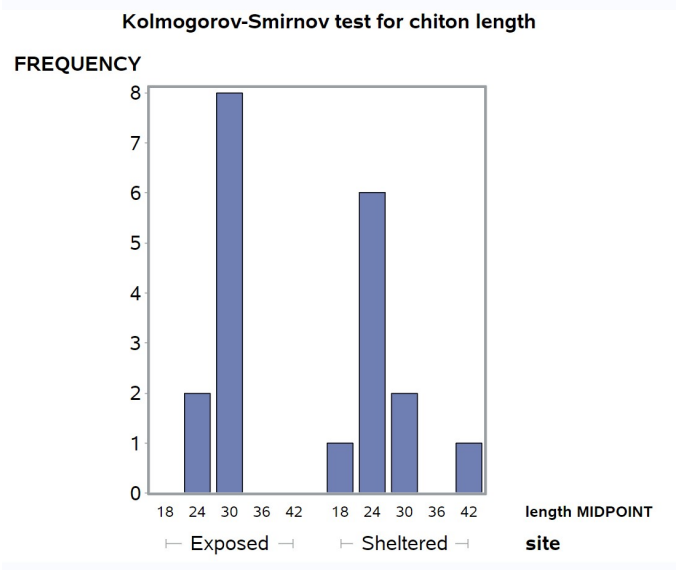
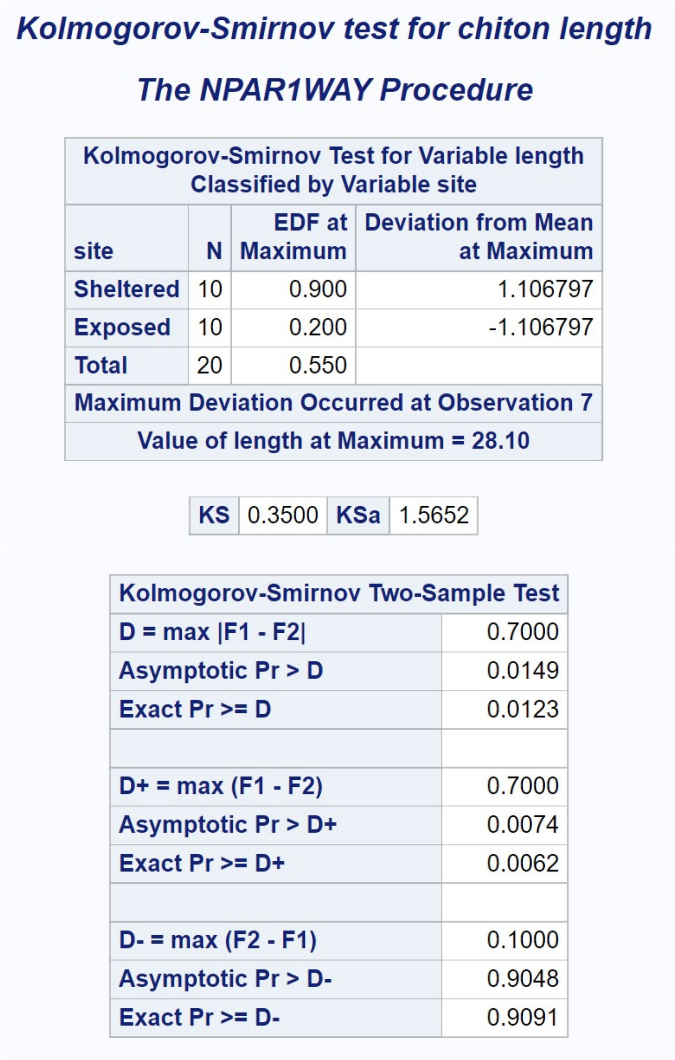


Figure 16.11: `KStest_chitons_Venados.sas - proc gchart`



Kolmogorov-Smirnov test for chiton length

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable length Classified by Variable site			
site	N	EDF at Maximum	Deviation from Mean at Maximum
Sheltered	10	0.900	1.106797
Exposed	10	0.200	-1.106797
Total	20	0.550	
Maximum Deviation Occurred at Observation 7			
Value of length at Maximum = 28.10			

KS	0.3500	KSa	1.5652

Kolmogorov-Smirnov Two-Sample Test	
D = max |F1 - F2|	0.7000
Asymptotic Pr > D	0.0149
Exact Pr >= D	0.0123
D+ = max (F1 - F2)	0.7000
Asymptotic Pr > D+	0.0074
Exact Pr >= D+	0.0062
D- = max (F2 - F1)	0.1000
Asymptotic Pr > D-	0.9048
Exact Pr >= D-	0.9091

Figure 16.12: `KStest_chitons_Venados.sas - proc npar1way`

16.4 Randomization tests

Randomization tests are another common kind of nonparametric test used for one-way designs, as well as more complex ones (Hinkelmann & Kempthorne 1994; Manly 1997). The null hypothesis for these tests is different from other tests we have considered, which involved statements about probability distributions and their parameters. For randomization tests, the null hypothesis is that all possible permutations (rearrangements) of the data among groups are equally likely, given no treatment or group effects, with the observed data being one such arrangement (Hinkelmann & Kempthorne 1994; Manly 1997). These tests commonly employ a parametric test statistic to examine the null hypothesis, one that is sensitive to potential differences among groups. For one-way designs, the F_s statistic from one-way ANOVA (Chapter 11) is often used to detect differences in the group means. To conduct a randomization test using this statistic, we first calculate the value of $F_s(obs)$ for the observed data. Similar to one-way ANOVA, we then need to determine if $F_s(obs)$ is sufficiently large to consider rejecting H_0. This is accomplished by permuting or rearranging the observations many times across groups, and calculating the value of F_s for each permutation. The justification for this procedure follows directly from the definition of H_0. The P value for the test is defined as the proportion of the F_s values greater than or equal to $F_s(obs)$, including $F_s(obs)$ as one of the values.

For small data sets it may be possible to carry out all possible permutations, but for larger data sets this may be impractical. Instead, the observations are randomly rearranged across groups a large number of times, in effect drawing a random sample from all possible permutations. The collection of F_s values obtained by this process is called the **randomization distribution**. How many of these randomizations are needed to generate an accurate P value for the test? Some guidance is provided by Manly (1997), who suggests that 1000 randomizations should be sufficient for $P \approx 0.05$, and 5000 for $P \approx 0.01$.

An interesting feature of randomization tests is that the randomization distribution of F_s under H_0 can be approximated by the parametric F distribution (Hinkelmann & Kempthorne 1974) under some conditions. This provides another justification for the use of F tests when the normality assumption of these tests is violated.

We will use data on nematode intensities for male vs. female bobcats (*Lynx rufus*) to illustrate randomization tests. The sampled bobcats

were recent roadkill collected from the Southern Illinois region (Francisco A. Jimenez-Ruiz and Eliot A. Zieman, unpublished data). The guts were examined for nematodes as well as other parasites, and the total number counted (Table 16.3). These data have many zeroes as well as large values, as is common for parasite intensity data. The data are clearly non-normal and so a nonparametric test seems warranted.

Table 16.3: Example 3 - Number of nematode parasites found in the gut of male and female bobcats collected from Southern Illinois .

Sex	Nematodes	Sex	Nematodes	Sex	Nematodes	Sex	Nematodes
F	0	F	0	M	6	M	8
F	8	F	5	M	10	M	0
F	0	F	0	M	1	M	60
F	0	F	0	M	0	M	25
F	0	F	0	M	5	M	1
F	0	F	11	M	59	M	0
F	0	F	0	M	2	M	74
F	1	F	5	M	3	M	3
F	2	F	11	M	0	M	1
F	1	F	0	M	44	M	15
F	1	F	24	M	1	M	0
F	6	F	13	M	1	M	7
F	1	F	2	M	0	M	0
F	6			M	2	M	0
F	2			M	17		
F	1			M	5		
F	13			M	3		
F	0			M	26		
F	0			M	20		
F	7			M	3		

16.4.1 Randomization test for Example 3 - SAS demo

We will analyze the bobcat data using both one-way ANOVA and the analogous randomization test, comparing the parasite intensities for male vs. female cats. The SAS program below first generates a graph showing the mean intensities for both sexes, then conducts a standard one-way ANOVA (Fig. 16.14, 16.15). We see that the mean intensity for male bobcats was higher than females, and the ANOVA showed this difference was significant ($F_{1,65} = 5.50, P = 0.0221$).

The program then uses two SAS macro programs to conduct the randomization test (Cassell 2002). SAS macros are chunks of code that are used to carry out custom calculations, ones not available in standard SAS procedures (SAS Institute Inc. 2016b). They are inserted into a main program through the use of `%include` statements, which point to the file locations of the macros on the user's computer. Note that the percent sign (`%`) tells SAS that a particular line contains macro code. The first macro, `%rand_gen.sas`, is used to generate the desired number of random permutations of the data. Once the macro is included in the program, it can be called using the following arguments. The input data set is specified using the `indata=parasites` statement, while the output data set specified by `outdata=outrand` contains all the randomizations. The statement `numreps=5000` sets the number of randomizations, with the dependent variable specified by `depvar=nematodes`.

The next step in the randomization test is to conduct a one-way ANOVA for each one of the randomizations, as well as the original data set. This is accomplished using `proc glm` with a `by replicate` statement. The variable `replicate` is generated by the `rand_gen` macro to number the different randomizations. In addition, a data file containing the statistical output of the ANOVA is specified using the statement `outstat=outstat1`. The ANOVA for the original data corresponds to a `replicate = 0` in this output file. The `noprint` option is used to suppress the printing of each ANOVA.

The last step in the randomization test uses the second macro, `%rand_anl.sas`, to determine the P value for the test. The data file containing the statistical output from `proc glm` is specified using a `randdata=outstat1` argument. The `where=_source_='sex' and _type_='SS3'` argument tells the macro which part of the statistical output to use, in particular the test associated with the sex effect and Type III sum of squares. The `testprob=prob` statement tells the macro to use the P value for this F test in calculating the P value for the randomization test. The macro uses the P rather than F_s value to provide some

additional flexibility for other kinds of tests (Cassell 2002). As the F_s and P value for the ANOVA are related, it yields the same result. *The P value for the randomization test is provided in the SAS log.* The `testlabel=Model F test` argument provides some labeling for this output. Examining the SAS log, we find that the randomization test was significant ($P = 0.0182$). The P value for this test was smaller than the one found using one-way ANOVA, and makes no assumptions about the distribution of the data.

The remaining portion of the program generates a graph of the randomization distribution of F_s, and displays the value of this statistic for the original distribution (Fig. 16.16). We see that the original value of F_s lies far above most of the randomizations. This illustrates the pattern for a significant randomization test. For a non-significant test, we would see an F_s value that is more central within the randomization distribution.

SAS Program

```
* Randtest_bobcat_parasites.sas;
title 'Randomization test for bobcat parasites';
data parasites;
    input nematodes sex $;
    datalines;
0  F
8  F
0  F
0  F
0  F

etc.

;
run;
* Print data set;
proc print data=parasites;
run;
* Plot means, standard error, and observations;
proc gplot data=parasites;
    plot nematodes*sex / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way ANOVA;
proc glm data=parasites;
    class sex;
    model nematodes = sex;
run;
* Include two macros for randomization test;
%include "/home/u49852288/sasuser.v94/Statistics Book 2/Chapter 16/rand_gen.sas";
%include "/home/u49852288/sasuser.v94/Statistics Book 2/Chapter 16/rand_anl.sas";
* Sampled randomization test;
%rand_gen(indata=parasites,outdata=outrand,depvar=nematodes,numreps=5000)
proc glm data=outrand noprint outstat=outstat1;
    by replicate;
    class sex;
    model nematodes = sex;
run;
%rand_anl(randdata=outstat1,where=_source_='sex' and _type_='SS3',testprob=prob,testlabel=Mo
* Extract F values from outstat1 for null distribution graph;
data nulldist;
    set outstat1;
```

```
    if _type_="SS3";
    * Assign original F value to macro variable;
    if replicate=0 then call symput('F',F);
run;
* Null distribution;
title2 "Null distribution";
proc univariate data=nulldist noprint;
    var F;
    histogram F / vscale=count href=&F hreflabel="F";
run;
quit;
```

Randomization test for bobcat parasites

Obs	nematodes	sex
1	0	F
2	8	F
3	0	F
4	0	F
5	0	F
6	0	F
7	0	F
8	1	F
9	2	F
10	1	F

etc.

Figure 16.13: `Randtest_bobcat_parasites.sas - proc print`

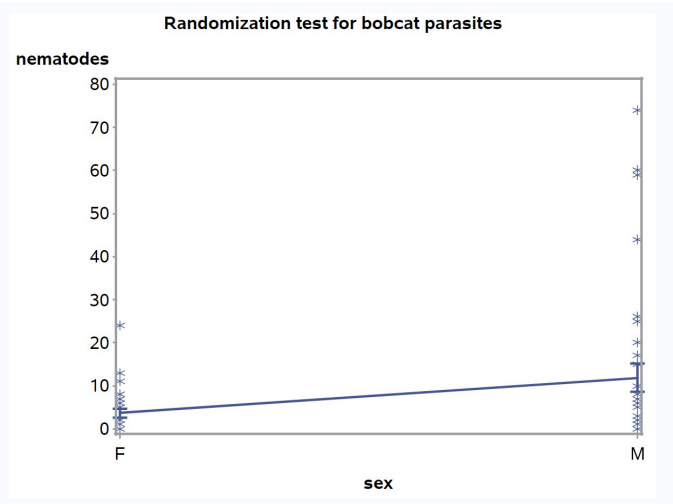


Figure 16.14: `Randtest_bobcat_parasites.sas - proc gplot`

Randomization test for bobcat parasites

The GLM Procedure

Class Level Information		
Class	Levels	Values
sex	2	F M

Number of Observations Read	67
Number of Observations Used	67

Randomization test for bobcat parasites

The GLM Procedure

Dependent Variable: nematodes

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1122.49709	1122.49709	5.50	0.0221
Error	65	13274.57754	204.22427		
Corrected Total	66	14397.07463			

R-Square	Coeff Var	Root MSE	nematodes Mean
0.077967	183.4248	14.29071	7.791045

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	1122.497087	1122.497087	5.50	0.0221

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	1122.497087	1122.497087	5.50	0.0221

Figure 16.15: `Randtest_bobcat_parasites.sas - proc glm`

SAS Log

```
Randomization test for Model F test where _source_='sex' and _type_='SS3'
has significance level of 0.0182
```

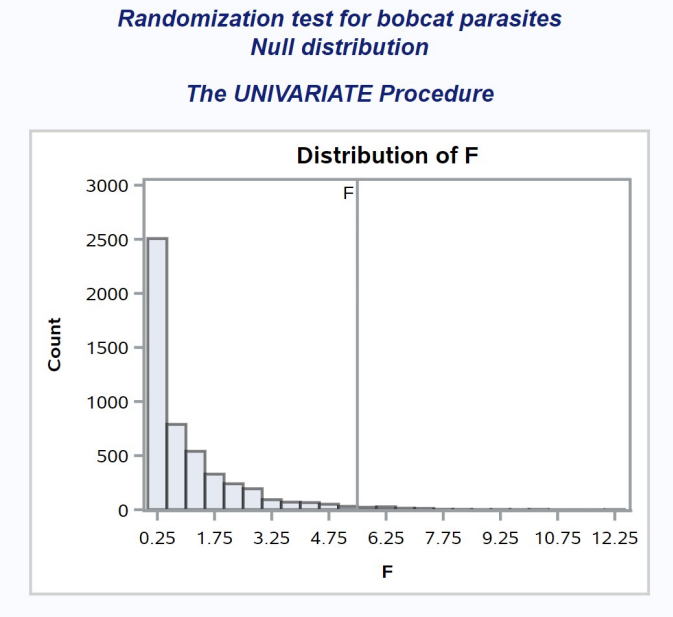


Figure 16.16: `Randtest_bobcat_parasites.sas - proc univariate`

16.5 Limitations of nonparametric tests

While nonparametric tests can be useful for non-normal data, they do have some drawbacks. One is that the number of designs that have nonparametric tests are fairly limited. We have seen nonparametric tests analogous to one-way ANOVA and two-sample t tests. There is also a rank test for randomized block designs called Friedman's test, as well as procedures for multiple comparisons (Hollander et al. 2014). Unfortunately, for more complex designs there are few available procedures.

Although nonparametric tests are not based on a particular distribution, they do make some assumptions. Consider the null and alternative hypotheses for the Wilcoxon test. The two groups are assumed to have the same cumulative distribution function, differing only by a shift parameter Δ. This implies the two groups have the same variance under both hypotheses, similar to parametric tests. When the variances are unequal as well as the sample sizes, both parametric and nonparametric tests may not be valid (Stewart-Oaten 1995). In particular, they may not have the correct Type I error rate.

Table 16.4 illustrates how unequal variances and sample sizes can affect the Type I error rate. It summarizes a simulation study comparing the validity of several different methods of comparing samples from two groups, including parametric and nonparametric methods. The first six columns give the theoretical mean, variance, and the sample sizes for the two groups. The simulated data were normally distributed with these parameters. Each data set was then analyzed using a two-sample t test, a Welch t test that implements a correction for unequal variances, the Wilcoxon test, and a randomization test. Any significant differences detected by these tests are Type I errors, because the two groups have the same mean. A total of 5000 simulated data sets were generated and analyzed. The proportion of simulated data sets showing significant results is an estimate of the Type I error rate (α) for each test. If the test is conducted using $\alpha = 0.05$, for example, we would expect this proportion of the simulations to be significant.

Regardless of differences in the variance between the two groups, when the sample sizes are equal all methods yielded a Type I error rate near the nominal $\alpha = 0.05$ level. When sample sizes are unequal, the t test, Wilcoxon test, and the randomization test all yielded Type I error rates higher or lower than $\alpha = 0.05$. Note that the pattern depends on which group (high or low variance) has the smaller sample size. Thus, the validity of these procedures

depends on equal variances, especially when sample sizes are unequal across groups. This assumption needs to be carefully examined within applying both parametric and nonparametric tests.

The only valid test in this scenario was the Welch t test, which employs a correction for unequal variances. The correction alters the degrees of freedom for the test, based on the sample sizes and variances of the two groups (Stuart et al. 1999). It is conducted automatically by `proc ttest` in SAS, with the output labeled `Satterthwaite` (see Chapter 11). There is also a similar procedure for one-way designs called Welch ANOVA. It can conducted under `proc glm` using the `welch` option for the `means` statement.

Table 16.4: Effect of unequal variances and sample sizes on the estimated Type I error rate for common parametric and nonparametric tests, using $\alpha = 0.05$ for all tests. See text for further details.

μ_1	σ_1^2	n_1	μ_2	σ_2^2	n_2	t	Welch	Wilcoxon	Randomization
10	1	10	10	1	10	0.0474	0.0454	0.0422	0.0484
10	1	10	10	2	10	0.0516	0.0504	0.0514	0.0524
10	1	5	10	2	15	0.0208	0.0510	0.0236	0.0214
10	1	15	10	2	5	0.0956	0.0578	0.0662	0.0954
10	1	10	10	4	10	0.0510	0.0452	0.0464	0.0510
10	1	5	10	4	15	0.0104	0.0494	0.0170	0.0108
10	1	15	10	4	5	0.1588	0.0574	0.0836	0.1598

16.6 References

Cassell, D. L. (2002) A randomization-test wrapper for SAS PROCs. SUGI 27: Paper 251-27.

Conover, W. J. (1999) *Practical Nonparametric Statistics.* John Wiley & Sons, Inc., New York, NY.

Flores-Campaña, L. M., Arzola-González, J. F., & León-Herrera, R. (2012) Body size structure, biometric relationships and density of *Chiton albolineatus* (Mollusca: Polyplacophora) on the intertidal rocky zone of three islands of Mazatlan Bay, SE of the Gulf of California. *Revista de Biologia Marina y Oceanographía* 47: 203-211.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972) Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research* 42: 237-288.

Hinkelmann, K., & Kempthorne, O. (1994) *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design.* John Wiley & Sons, Inc., New York, NY.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014) *Nonparametric Statistical Methods, Third Edition.* John Wiley & Sons, Inc., Hoboken, NJ.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

SAS Institute Inc. (2016a) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016b) *SAS 9.4 Macro Language: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

Stuart, A., Ord, J. K., & Arnold, S. (1999) *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model.* Oxford University Press Inc., New York, NY.

Manly, B. F. J. (1997) *Randomization, Bootstrap, and Monte Carlo Methods in Biology.* Chapman & Hall, New York, NY.

16.7 Problems

1. Using the Example 3 data, conduct a Wilcoxon test comparing parasite intensity in male vs. female bobcats. How do the results compare to the randomization test for these data in the text?

2. Data were also collected on the number of cestode parasites found in the bobcats from Example 3 (see below). Cestodes are another common type of gut parasite. Conduct a randomization test comparing the cestode intensity for male vs. female bobcats.

Sex	Cestodes	Sex	Cestodes	Sex	Cestodes	Sex	Cestodes
F	1	F	0	M	9	M	3
F	7	F	7	M	31	M	2
F	9	F	6	M	5	M	2
F	0	F	33	M	0	M	0
F	1	F	2	M	10	M	3
F	1	F	1	M	6	M	7
F	8	F	18	M	0	M	2
F	0	F	6	M	0	M	5
F	0	F	1	M	6	M	1
F	32	F	14	M	9	M	1
F	11	F	12	M	6	M	4
F	4	F	6	M	18	M	0
F	3	F	0	M	4	M	3
F	13			M	9	M	1
F	2			M	6		
F	2			M	5		
F	12			M	17		
F	4			M	4		
F	1			M	8		
F	3			M	11		