SAS for Data Management, Analysis, and Reporting

Lecture 3
Introduction to Nonparametric Methods
Analysis of Variance and Wilcoxon Rank-Sum Test
June 23, 2004

Kate Cowies
374 SH
kcowies@stat.uiowa.edu

"Nonparametric" or "distribution-free" statistical methods

- allow for testing hypotheses that are not statements about population parameter values
- may be used when the form of the distribution of the sampled population is unknown
- can be used when data being analyzed consist merely of rankings or classifications
  - i.e. when arithmetic operations required for parametric procedures cannot be done
  - example: data on patient conditions reported as "better," "same," or "worse"

Parametric methods

- based on the assumption that the population(s) from which our samples are drawn follow a distribution, the general form of which is known
  - e.g. normal or binomial
- research interest is in estimating, or testing a hypothesis about, one or more population parameters
- examples: z tests, t tests, and ANOVA for making inference about means of populations assumed to be normal

Example for the Sign Test

- We wish to compare the effectiveness of two ointments (A, B) in reducing sunburn in people whose skin is sensitive to sunlight.
- For each person in the study, we randomly select either the left arm or the right arm and apply ointment A. We then apply ointment B to the same area of the other arm.
- We then expose the person to 1 hour of sunlight and compare the two arms with respect to degree of redness.
- We can make only the following qualitative assessments:
  1. "A" arm is not as red as "B" arm.
  2. "A" arm is redder than "B" arm.
  3. Arms are equally red.
How might we compare the effectiveness of the two ointments if we were able to measure redness on a quantitative scale?

In the situation described here, we cannot observe the actual values of within-person differences in redness between the A arm and the B arm.

What we can observe are the signs of the differences:

1. “A” arm is not as red as “B” arm (+)
2. “A” arm is redder than “B” arm (-)
3. Arms are equally red (0)

To carry out the sign test:

- Ignore the pairs (or observations) with difference of 0.
- Denote the number of remaining pairs as $n$.
- Count the number of plus signs, and denote it $D$.
- Note that under the null hypothesis, we would expect approximately equal numbers of plus and minus signs.
  - more precisely, under the null hypothesis, $D$ follows a binomial distribution with success probability $p = 1/2$ and number of trials $n$
  - This binomial distribution has
    \[ \text{mean} = np = \frac{n}{2} \]
    \[ \text{standard deviation} = \sqrt{np(1-p)} = \sqrt{\frac{n}{4}} \]

### The Sign Test

The null hypothesis of the sign test is that in the underlying population of differences, the median difference $M$ is 0.

\[ H_0 : M = 0. \]

The alternative hypothesis may be either one-sided or two-sided.

\[ H_0 : M > 0 \]
\[ H_0 : M < 0 \]
\[ H_0 : M \neq 0 \]

- We must evaluate how likely we would have been to obtain a value of $D$ as extreme as what we got, or more extreme, if the null is true.
- Your textbook gives the test statistic for use with a normal approximation to the binomial distribution. This is appropriate for use if $n \geq 20$. The value is compared to the standard normal distribution.
- Otherwise, we will use the binomial distribution directly.
The sign test for the skin ointment data

We wish to do a two-sided test, i.e.

\[ H_0 : M = 0 \]

at the \( \alpha = .05 \) significance level.

The results for 45 subjects are:

1. 22 people had the “A” arm less red (+)
2. 18 people had the “B” arm less red (-)
3. 5 people had no difference (0)

- \( n = 45 - 5 = 40 \)
- \( D = 22 \)
- normal approximation is valid because \( n \geq 20 \).

\[
Z_+ = \frac{D - (n/2)}{\sqrt{n/4}}
\]

So again, we cannot reject \( H_0 \). We conclude that the data do not provide evidence that one ointment is better than the other.

\[
Z_{+} = \frac{22 - 20}{\sqrt{10}} = 0.632
\]

For a 2-sided test, we must compare this value to the .025 cutoff for the standard normal distribution, which is 1.96.

Because 0.632 < 1.96, we cannot reject \( H_0 \).

Equivalently, we can determine the p-value of our test by finding \( P(z > 0.632) = \approx .264 \).

- This would be the p-value for a 1-sided test.
- To find the p-value for our 2-sided test, we multiply by 2.

\[
p = 2(.264) = .528 > \alpha = .05
\]

The sign test with small sample size

Suppose that instead of 40 patients with non-zero differences, we had had

1. 5 people had the “A” arm less red (+)
2. 3 people had the “B” arm less red (-)
3. 37 people had no difference (0)

Then

- \( n = 45 - 37 = 8 \)
- \( D = 5 \)
- normal approximation is inappropriate because \( n < 20 \).

- we will do exact calculation of the p-value using the binomial distribution
Because $D > n/2 = 4$, we will compute

\[
P(D \geq 5|H_0) = P(D = 5) + P(D = 6) \\
+ P(D = 7) + P(D = 8) \\
= .2188 + .1094 + .0313 + .0039 \\
= 0.3634
\]

This is a one-sided p-value. We must multiply by 2 to get the approximate 2-sided p-value.

\[
2(0.3634) = 0.7268 > .05
\]

So again we would not reject $H_0$.

**More on the sign test**

- Can be used with single-sample or paired-sample problems
- Frees us from having to make any assumptions about the underlying distribution of differences
- If we have any information about the magnitude of the individual differences, the sign test wastes it.

**The Wilcoxon Signed-Rank Test**

- for single sample or paired samples
- useful when the population distribution is not normal and the sample size is not large
  - of the within-pair differences in paired sample case or of individual values in single sample case
- makes use of the magnitudes of the differences as well as their signs

**Example**


As part of the study, 17 families participated in a training program. Before and after the training program, the primary parent took the Behavioral Vignettes test, which assesses knowledge of behavioral modification principles. A higher score indicates greater knowledge.
The following are the pre- and post-test training scores for 12 of their families:

<table>
<thead>
<tr>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
</tr>
</tbody>
</table>

May we conclude from these data that the training program increases knowledge of behavior modification principles? (We will test at the \( \alpha = .01 \) level.)

**Steps in the Wilcoxon signed-rank procedure**

1. Select a random sample of \( n \) pairs of observations.
2. Compute the difference \( d_i \) in each pair of observations. Delete all pairs in which \( d_i = 0 \), and reduce \( n \) accordingly.
3. Ignoring the signs of the \( d_i \)'s, rank their absolute values from smallest to largest. When there are ties in absolute values, assign each tied value the mean of the rank positions the tied values occupy.
4. Assign to each rank the sign of the \( d_i \) that yields that rank.

5. Find \( T_+ \), the sum of the ranks with positive signs, and \( T_- \), the sum of the ranks with negative signs.
6. Let the test statistic \( T \) equal the smaller of \( T_+ \) and \( T_- \).

**Hypotheses of the Wilcoxon Signed Rank Test**

The null hypothesis is that, in the underlying population of differences among pairs, the median difference is equal to 0.

\[
H_0 : M_d = 0
\]

The alternative hypothesis may be one- or two-sided.

\[
H_a : M_d > 0 \quad H_a : M_d < 0 \quad H_a : M_d \neq 0
\]

If we define our differences as post - pre, then our alternative would be:

\[
H_a : M_d > 0
\]
Pre- and Post-Test example

<table>
<thead>
<tr>
<th>Pre</th>
<th>Post</th>
<th>d_i</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>11</td>
<td>4</td>
<td>9.5</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>-1</td>
<td>-2.5</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>4</td>
<td>9.5</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>15</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

The sum of the negative ranks is $T = 2.5$.

SAS for the Wilcoxon Signed Rank Test

- carried out automatically by `proc univariate`
- SAS computes a slightly different form of the test statistic
  \[ S = \sum(\text{positive ranks}) - \frac{n(n + 1)}{4} \]
  recalling that $n$ is the number of differences whose value is not equal to 0.
- computes $p$-value in two different ways depending on sample size
  - if $n \leq 20$, $p$-value is computed from each distribution of $S$, which can be enumerated under null hypothesis that distribution is symmetric around 0
  - when $n > 20$ approximate $S$ is compared to approximate $t$ distribution

```sas
data whatever ;
input pre post ;
diff = post - pre ;
datalines ;
  7  11
  6  14
 10  16
 16  17
  8  9
 13  15
  8  9
 14  17
 16  20
 11  12
 12  14
 13  15
;
run ;
```

```
proc univariate ;
var diff ;
run ;
```

The UNIVARIATE Procedure

Variable: diff

Tests for Location: Mu=0

<table>
<thead>
<tr>
<th>Test</th>
<th>-Statistic-</th>
<th>----p Value----</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student's t</td>
<td>t 4.521908</td>
<td>Pr &gt;</td>
</tr>
<tr>
<td>Sign</td>
<td>M 6</td>
<td>Pr &gt;</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S 39</td>
<td>Pr &gt;</td>
</tr>
</tbody>
</table>

Note: For a 1-sided $p$-value, we would divide the 2-sided $p$-value by 2.
Interpreting the results

- Recall that we wanted to determine whether the audiovisual instruction improved parent’s test scores.
- The null and alternative hypotheses regarding the median difference (that is, the median of post - pre), are

\[ H_0 : M_d = 0 \]
\[ H_a : M_d > 0 \]

- Can we reject \( H_0 \) at the .01 significance level?

- What does this mean with respect to the research question?

The Wilcoxon Rank Sum Test

- used to compare nonparametrically two samples that have been drawn from independent populations
  - nonparametric analog of two-independent-sample t-test
- also called Mann-Whitney test, Mann-Whitney U test, and Mann-Whitney-Wilcoxon test

Sign Test in SAS

- Note that `proc univariate` also automatically carries out the sign test
- its version of sign test statistic is

\[ M = \frac{n^+ - n^-}{2} \]

- use sign test if sample size is small and it is unreasonable to assume that population distribution is symmetric
- sign test p-value will often be a little larger than that of the Wilcoxon signed rank test (not so in this case)

Assumptions of the Wilcoxon Rank Sum Test

- Two samples, of sizes \( n \) and \( m \), have been drawn independently and randomly from their respective populations
- The measurement scale is at least ordinal
- The variable of interest is continuous
- If the populations differ, they differ only with respect to their medians
  - i.e., otherwise their shapes are approximately the same
Example: a question in pharmacokinetics

- Is total plasma clearance of cefpiramide different in healthy people vs. patients with alcoholic cirrhosis? Demotes-Mainard et al. (1991) measured total plasma clearance (ml/min)
- following a single 1-gram intravenous injection of cefpiramide in 10 healthy volunteers and 10 patients with alcoholic cirrhosis.
- They chose to use a non-parameteric method of analysis because their medical knowledge indicated that plasma clearance of drugs tends to have a skewed (not normal!) distribution

<table>
<thead>
<tr>
<th>Case number</th>
<th>CRR CLEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.700</td>
</tr>
<tr>
<td>2</td>
<td>29.300</td>
</tr>
<tr>
<td>3</td>
<td>25.300</td>
</tr>
<tr>
<td>4</td>
<td>22.800</td>
</tr>
<tr>
<td>5</td>
<td>21.300</td>
</tr>
<tr>
<td>6</td>
<td>31.200</td>
</tr>
<tr>
<td>7</td>
<td>29.200</td>
</tr>
<tr>
<td>8</td>
<td>29.700</td>
</tr>
<tr>
<td>9</td>
<td>17.200</td>
</tr>
<tr>
<td>10</td>
<td>26.700</td>
</tr>
<tr>
<td>11</td>
<td>14.600</td>
</tr>
<tr>
<td>12</td>
<td>18.100</td>
</tr>
<tr>
<td>13</td>
<td>12.300</td>
</tr>
<tr>
<td>14</td>
<td>8.800</td>
</tr>
<tr>
<td>15</td>
<td>10.300</td>
</tr>
<tr>
<td>16</td>
<td>8.500</td>
</tr>
<tr>
<td>17</td>
<td>29.300</td>
</tr>
<tr>
<td>18</td>
<td>8.100</td>
</tr>
<tr>
<td>19</td>
<td>6.900</td>
</tr>
<tr>
<td>20</td>
<td>7.900</td>
</tr>
</tbody>
</table>

Can we conclude at the $\alpha = .01$ significance level that median clearance rate is different in healthy patients vs. those with alcoholic cirrhosis?

Hypotheses for the Wilcoxon Rank Sum Test

$$H_0 : M_1 = M_2$$

The alternative hypothesis may be one- or two-sided.

$$H_a : M_1 > M_2$$

$$H_a : M_1 < M_2$$

$$H_a : M_1 \neq M_2$$

Procedure for the Wilcoxon Rank Sum Test

- Combine the two samples into one large group, and sort values from smallest to largest.
- Rank the values. When there are ties in absolute values, assign each tied value the mean of the rank positions the tied values occupy.
- Sum the ranks within each original sample
- The test statistic is $W$, the smaller of the two sums.
Ranked values for clearance example

<table>
<thead>
<tr>
<th>Case number</th>
<th>CIRR</th>
<th>CLEAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>21.700</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>29.300</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>25.300</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>22.800</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>21.300</td>
</tr>
<tr>
<td>6</td>
<td>0.000</td>
<td>31.200</td>
</tr>
<tr>
<td>7</td>
<td>0.000</td>
<td>29.200</td>
</tr>
<tr>
<td>8</td>
<td>0.000</td>
<td>28.700</td>
</tr>
<tr>
<td>9</td>
<td>0.000</td>
<td>17.200</td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>25.700</td>
</tr>
<tr>
<td>11</td>
<td>1.000</td>
<td>14.600</td>
</tr>
<tr>
<td>12</td>
<td>1.000</td>
<td>18.100</td>
</tr>
<tr>
<td>13</td>
<td>1.000</td>
<td>12.300</td>
</tr>
<tr>
<td>14</td>
<td>1.000</td>
<td>8.800</td>
</tr>
<tr>
<td>15</td>
<td>1.000</td>
<td>10.300</td>
</tr>
<tr>
<td>16</td>
<td>1.000</td>
<td>8.500</td>
</tr>
<tr>
<td>17</td>
<td>1.000</td>
<td>29.300</td>
</tr>
<tr>
<td>18</td>
<td>1.000</td>
<td>8.100</td>
</tr>
<tr>
<td>19</td>
<td>1.000</td>
<td>6.900</td>
</tr>
<tr>
<td>20</td>
<td>1.000</td>
<td>7.900</td>
</tr>
</tbody>
</table>

The Wilcoxon Rank Sum test in SAS

```sas
proc npar1way; 
    class cirr; 
    var clear; 
    run; 
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable clear 
Classified by Variable cirr

<table>
<thead>
<tr>
<th>cirr</th>
<th>N</th>
<th>sum of scores</th>
<th>expected under ho</th>
<th>std dev</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>145.50</td>
<td>105.0</td>
<td>13.22382</td>
<td>14.50</td>
</tr>
</tbody>
</table>

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic 145.500

Normal Approximation

Z 3.0249

One-Sided Pr > Z  0.0012

Two-Sided Pr > |Z|  0.0025

t Approximation

One-Sided Pr > Z  0.0035

Two-Sided Pr > |Z|  0.0070

Z includes a continuity correction of 0.5.

Comparing more than two population means

Example: Does the presence of pets or friends affect responses to stress?

- subjects: 45 women who described themselves as dog lovers
- randomly assigned to three groups: to do a stressful task
  1. alone
  2. with a good friend present
  3. with their dog present
- Subjects’ mean heart rate during the task was one measure of the effect of stress.
Goal: to compare population means under three different “treatments”

- a three-independent-sample problem
- Call the population mean heart rates $\mu_1$ for when pets are present, $\mu_2$ for when friends are present, and $\mu_3$ for when women perform task alone; then

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \mu_1 \neq \mu_2 \text{ or } \mu_1 \neq \mu_3 \text{ or } \mu_2 \neq \mu_3$$
* not one-sided or 2-sided

**Multiple comparisons procedures in statistics**

- issue: how to do many comparisons at once with some overall measure of confidence in all our conclusions
- two steps
  - overall test of whether there is good evidence of any differences among parameters we wish to compare
  - follow-up analysis to decide which of parameters differ and to estimate size of differences

**SAS descriptive statistics:**

Analysis Variable: BEATS

<table>
<thead>
<tr>
<th>group</th>
<th>Obs</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>15</td>
<td>15</td>
<td>82.5240667</td>
<td>9.2415747</td>
<td>62.6460000</td>
<td>99.04600</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>15</td>
<td>91.3251333</td>
<td>8.3411341</td>
<td>76.9080000</td>
<td>102.15400</td>
</tr>
<tr>
<td>P</td>
<td>15</td>
<td>15</td>
<td>73.4830667</td>
<td>9.9068202</td>
<td>58.6920000</td>
<td>97.53000</td>
</tr>
</tbody>
</table>

**Step one: One-Way Analysis of Variance (ANOVA)**

- step one (overall test) for *some* difference among 3 or more population means
- uses an $F$ test to compute a p-value
Main idea of ANOVA

What matters is how far apart sample means are relative to variability of individual observations.

• F statistic
  \[ F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}} \]

• compare to a cutoff value in an F distribution

Notation:

• \( I \) = number of different populations whose means we are studying

• \( n_i \) = number of observations in sample from \( i \)th population

• \( N \) = total number of observations in all samples combined

F distributions

• many different F distributions, identified by two parameters
  – numerator degrees of freedom = \( I - 1 \)
  – denominator degrees of freedom = \( N - I \)

Assumptions for One-Way ANOVA

• We have \( I \) independent simple random samples, one from each of \( I \) populations.

• Each population \( i \) has a normal distribution with unknown mean \( \mu_i \).
  – As with \( t \)-tests, if sample sizes are large enough in each sample, Central Limit Theorem says inference based on sample means is OK even if population distributions are not exactly normal.

• All of the populations have the same standard deviation \( \sigma \) (unknown)
  – unlike \( t \)-tests, there is no general procedure when population standard deviations are not assumed to be equal
  – rough rule of thumb: if largest sample standard deviation is no more than twice the smallest sample standard deviation, then population standard deviations probably are close enough to equal that ANOVA procedure is OK
We can check assumptions using proc means and proc univariate

```plaintext
Optimal linesize = 79;

data pet; 
infile '/temp/pet.dat';
input group $ beats;
run;

proc sort data = pet; 
by group; 
run;

proc means data = pet; 
by group; 
var beats;
run;

proc univariate normal plot data = pet; 
by group; 
var beats;
run;
```

---

The MEANS Procedure

Analysis Variable : beats

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Obs</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>15</td>
<td>15</td>
<td>82.5240667</td>
<td>9.241547</td>
<td>62.6860000</td>
<td>99.0460000</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>15</td>
<td>91.3291333</td>
<td>8.3411341</td>
<td>76.9090000</td>
<td>102.1540000</td>
</tr>
<tr>
<td>P</td>
<td>15</td>
<td>15</td>
<td>73.4830667</td>
<td>9.9698202</td>
<td>58.0920000</td>
<td>97.5380000</td>
</tr>
</tbody>
</table>

---

The UNIVARIATE Procedure

Variable: beats

Moments

<table>
<thead>
<tr>
<th>N</th>
<th>15 Sum Weights 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>82.5240667 Sum Observations 1237.861</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>9.24157468 Variance 88.4067026</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.4776032 Kurtosis 0.3251807</td>
</tr>
<tr>
<td>Uncorrected SS</td>
<td>103349.018 Corrected SS 1195.69844</td>
</tr>
<tr>
<td>Coeff Variation</td>
<td>11.1986414 Std Error Mean 2.36616432</td>
</tr>
</tbody>
</table>

Basic Statistical Measures

Location

Variability

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>---Statistic---</th>
<th>-----p Value-----</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.970236</td>
<td>Pr &lt; W 0.8615</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.185622</td>
<td>Pr &gt; D &gt; 0.1500</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.052736</td>
<td>Pr &gt; W-Sq &gt; 0.2500</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.285759</td>
<td>Pr &gt; A-Sq &gt; 0.2500</td>
</tr>
</tbody>
</table>

Stem Leaf # Boxplot

```
Stem Leaf

```

Multiply Stem Leaf by 10^1

Proc univariate provides side-by-side boxplots at the end of the output when a by statement is used. (It does not do this when a class statement is used to get separate output for different groups.)
So go ahead and do ANOVA

Dogs, friends, and stress example:

```
proc anova data = pet;
  class group;
  model beats = group;
  run;
```

### Analysis of Variance Procedure

**Class Levels Values**

| GROUP | 3 | C | F | P |

**Number of Observations in data set = 45**

**Analysis of Variance Procedure**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of</th>
<th>Mean</th>
<th>C. V.</th>
<th>Root MSE</th>
<th>BEATS Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>0.401360</td>
<td>11.16915</td>
<td>9.2083030</td>
<td>82.444069</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GROUP</td>
<td>2387.6889920</td>
<td>1193.8444960</td>
<td>14.08</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step two: individual t-tests with correction for multiple comparisons**

This is the follow-up test.

- should be carried out only if the F test from one-way ANOVA is significant at the chosen significance level.

Goal: to set the overall probability of committing a type I error at $\alpha$ when doing pairwise comparisons of $k$ different means

- we will perform $\binom{k}{2}$ two-independent-sample t-tests
- we will conduct each one at the significance level $\alpha^*$

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$
This is called the *Bonferroni correction*  
— very conservative

— Equivalently, we could multiply the p-value from each t-test by 3.

* If the result was less than .05, we would consider the difference between two population means to be significantly different from zero at the .05 level

Dogs, friends, and stress example

- There are $k = 3$ samples, so there are $\binom{k}{2} = 3$ different pairs to compare.
- To get an overall significance level $\alpha = .05$ on all 3 tests considered together, we conduct each one at
  \[
  \alpha^* = \frac{.05}{3} = .0167
  \]
  - That is, we would consider the difference between two population means to be significantly different from zero at the .05 level only if the p-value for the the t-test for that pair was less than .0167.

SAS does the adjusting and prints a grouped list of the classes. Means with the same letter are not significantly different at the specified alpha level.

```sas
proc anova data = pet ;
class group ;
model beats = group ;
means group / bon alpha = .05 ;
run ;
```

### Analysis of Variance Procedure

*Bonferroni (Bon) T tests for variable: BEATS

**NOTE**: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05  df= 42  MSE= 84.79285  
Critical Value of T= 2.49  
Minimum Significant Difference= 8.3847

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Bon Grouping</th>
<th>Mean</th>
<th>N GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>91.326</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>82.524</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>73.483</td>
<td>15</td>
</tr>
</tbody>
</table>
The Kruskal-Wallis test

- ANOVA extends idea of 2-independent-sample t-test to more than 2 independent samples
- Wilcoxon Rank-Sum test is non-parametric way of comparing centers of population distributions using two independent samples
- Kruskal-Wallis test extends idea of Wilcoxon Rank-Sum test to more than 2 independent samples
- like Wilcoxon Rank-Sum, Kruskal-Wallis has less power than ANOVA, but it does not depend on such strong distributional assumptions
- SAS proc npar1way is used for the Kruskal-Wallis test

Kruskal Wallis test for pets and stress data

Note: the Kruskal Wallis test is not the best choice for this dataset, because both subject-matter knowledge and inspection of sample data suggest that assumption of population normality probably is reasonable.

```
proc npar1way wilcoxon;
class group;
var beats;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable beats
Classified by Variable group

<table>
<thead>
<tr>
<th>group</th>
<th>N</th>
<th>Sum of Scores</th>
<th>Expected Under HO</th>
<th>Std Dev Under HO</th>
<th>Mean Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>15</td>
<td>350.0</td>
<td>345.0</td>
<td>41.53319</td>
<td>23.33333</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>495.0</td>
<td>345.0</td>
<td>41.53319</td>
<td>33.000000</td>
</tr>
<tr>
<td>P</td>
<td>15</td>
<td>190.0</td>
<td>345.0</td>
<td>41.53319</td>
<td>12.666667</td>
</tr>
</tbody>
</table>

Kruskal-Wallis Test

- Chi-Square 17.9003
- DF 2
- Pr > Chi-Square 0.0001