Module E: Advanced Inference

Slide Deck E6:

Beyond the ANOVA Procedure

The section in which we cover the requirements, the alternative, and the extensions to the the Analysis of Variance procedure. If the requirements of ANOVA are not met, one should use the Kruskal-Wallis test. If a difference is detected, Tukey's HSD test (or the Kruskal multiple comparisons test) should be used to determine which is different.

Today's Objectives

By the end of this slidedeck, you should

1. understand the theory behind testing. . .
   - the means of more than two populations
   - whether a categorical variable helps understand a numeric
   - independence between a numeric and a categorical variable
2. determine if ANOVA or the Kruskal-Wallis test should be used
3. determine *which* level is different using Tukey's HSD or the Kruskal multiple comparisons test
4. explain the p-value and how to test hypotheses

Review: Procedure Requirements

Always, we have to make assumptions in determining the distribution of the test statistic. In many cases, you saw the assumptions. In some, they were hidden.

Test	Assumption
z-test	Normality
t-test	Normality
Variance test	Normality
Wilcoxon test	Symmetry
Mann-Whitney test	none*
Binomial test	none*
Proportions test	Normality**
Chi-Square Goodness-of-Fit test	Normality**

ANOVA Requirements

For the analysis of variance procedure, there are two assumptions that we needed to make:

- Normality in each group
- Equal variances across the groups

These assumptions can be tested using the following tests:

- Shapiro-Wilk test `shapiroTest`
- Fligner-Killeen test `fligner.test`

The null hypothesis for each test is

- the data are from a Normal distribution
- the variances are equal

Thus, small p-values indicate the requirement is not met by the data.

Rice Example

Example

Does rice variety influence average yield amongst these four varieties?

First, because we are testing for independence between a numeric (yield) and a categorical (variety) variable, we would like to use the analysis of variance procedure.

The hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \text{At least one mean differs from the others}$$

Rice Example

We would like to use the analysis of variance procedure, because it is the most powerful test for this type of hypothesis. However, this test has two requirement:

- Normality in each group
- Equal variances across the groups

The following code loads the data and tests these two assumptions

```
source("http://rfs.kvasaheim.com/stat200.R")
rice=read.csv("http://rfs.kvasaheim.com/data/rice.csv")
attach(rice)

shapiroTest(yield ~ variety)
fligner.test(yield ~ variety)
```

Rice Example

Here is the resulting output:

```
$adjustment
[1] "Bonferroni (4)"

$results
  Level    p.value
1      A 0.3148640
2      B 1.0000000
3      C 0.5065811
4      D 0.8381207
```

... and ...

```
        Fligner-Killeen test of homogeneity of variances

data:  yield by variety
Fligner-Killeen:med chi-squared=1.0026, df=3, p-value=0.8006
```


Rice Example

Incomplete Conclusion:

We are asked to determine if the rice yield and the rice variety are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. Neither assumption is violated. The minimum p-value from the Shapiro-Wilk test is 0.0787, which is greater than $\alpha = 0.05$. The p-value from the Fligner-Killeen test, 0.8006, is also greater than our $\alpha = 0.05$. Because neither assumption is violated, we can use the analysis of variance procedure.

The small p-value of 0.00503 of the ANOVA procedure indicates that the two variables, yield and variety, are dependent. Not all varieties have the same average yield.

Beyond ANOVA

Note that the conclusion was

- at least one mean differed from the others

That is hardly helpful.

What we really want to know is "*which* variety is different?"

ANOVA cannot answer that question.

However, there is a procedure that *can* answer it.

- Tukey's Honestly Significant Difference test

Its assumptions are the same as for ANOVA. So, if you use ANOVA, then you can use Tukey's HSD test.

Beyond ANOVA

The code to perform Tukey's HSD test on our model is

```
TukeyHSD(ricemod)
```

The resulting output is

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ variety)

$variety
      diff         lwr      upr     p adj
B-A  -56.25 -191.592699  79.0927 0.6185496
C-A  -46.00 -181.342699  89.3427 0.7473470
D-A  132.00   -3.342699 267.3427 0.0567296
C-B   10.25 -125.092699 145.5927 0.9957690
D-B  188.25   52.907301 323.5927 0.0066015
D-C  178.00   42.657301 313.3427 0.0097522
```

Beyond ANOVA

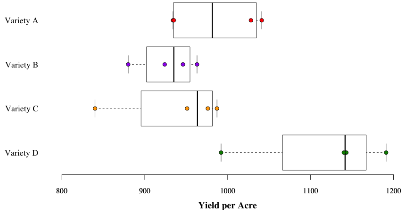
```
      diff         lwr      upr     p adj
B-A -56.25 -191.592699  79.0927 0.6185496
C-A -46.00 -181.342699  89.3427 0.7473470
D-A 132.00   -3.342699 267.3427 0.0567296
C-B  10.25 -125.092699 145.5927 0.9957690
D-B 188.25   52.907301 323.5927 0.0066015
D-C 178.00   42.657301 313.3427 0.0097522
```

Interpreting this abbreviated output:

- Find the p-values and interpret them for the null hypothesis "the difference in averages between these two levels is 0."
- At our usual level of significance, we were only able to detect differences in average yield between Varieties B and D and between Varieties C and D.
- Variety D has a higher average yield than Variety B by between 53 and 324.
- Variety D has a higher average yield than Variety C by between 43 and 313.
- We did not detect a difference between average yield between any other comparisons.

Beyond ANOVA

These conclusions are hardly surprising in light of the data:



Shouldn't Do ANOVA?

So, what if any requirement (assumption) is violated?

- Unsurprisingly: You should *not* use ANOVA

Like the t-test, there is a non-parametric alternative to ANOVA. It is called the **Kruskal-Wallis test**.

- It has the same requirements as the Mann-Whitney test
- It is a part of the `agricolae` package, which you have to install
- Its output is simpler than that of ANOVA
- It has its own multiple comparisons test, the Kruskal (multiple comparisons) test
- The Kruskal test output is also simpler than that of Tukey's HSD

Extension Example: Citizen's Initiative

Example

The citizen's initiative allows the people of the state to force a vote on a given issue. Do the different political cultures use the initiative at the same rates?

The hypotheses are

$$H_0 : \mu_m = \mu_i = \mu_t$$
$$H_a : \text{At least one average differs}$$

Extension Example: Citizen's Initiative

Because we are comparing multiple means, we would like to use the ANOVA procedure. It has two assumptions:

- The data come from a Normal distribution in each group
- The data have the same variance across the groups

To check this using R, we run the following code

```
shapiroTest(constituse ~ domPolCulture)
fligner.test(inituse ~ domPolCulture)
```


Extension Example: Citizen's Initiative

Here are the results:

```
$adjustment
[1] "Bonferroni (3)"

$results
            Level      p.value
1  Individualistic 1.357651e-03
2       Moralistic 2.048182e-03
3 Traditionalistic 9.940607e-06
```

... and ...

```
        Fligner-Killeen test of homogeneity of variances

data:  inituse by domPolCulture
Fligner-Killeen:med chi-squared=10.007, df=2, p-value=0.0067
```

Extension Example: Citizen's Initiative

Because *at least one* requirement was not met, we should not use ANOVA. We will use the Kruskal-Wallis test:

```
kruskal.test(inituse, domPolCulture)
```

The resulting output is

```
        Kruskal-Wallis rank sum test

data:  inituse and domPolCulture
Kruskal-Wallis chi-squared=6.3238, df=2, p-value=0.04235
```

Because the p-value is less than our usual $\alpha = 0.05$, we reject the null hypothesis. We can conclude that at least one mean differs from the others.

- Which one?

Extension Example: Citizen's Initiative

To determine *which is different*, we use Kruskal's multiple comparisons test:

```
print(kruskal(inituse, domPolCulture))
```

The partial output is

```
$groups
                 inituse groups
Moralistic       32.32353      a
Individualistic  24.91176     ab
Traditionalistic 20.76471      b
```

From this, we know we detected a difference in average initiative use between the moralistic states and the traditionalistic states. No other differences were detected.

Extension Example: Citizen's Initiative

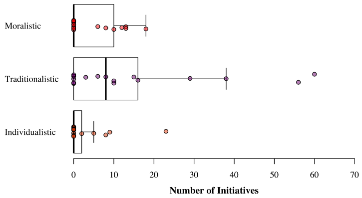
Conclusion:

We are asked to determine if the initiative use and the political culture are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. Both assumptions are violated. The maximum p-value from the Shapiro-Wilk test is 0.00204, which is much less than $\alpha = 0.05$. The p-value from the Fligner-Killeen test, 0.0067, is also less than our $\alpha = 0.05$. Because both assumptions are violated, we should not use the analysis of variance procedure. We must use the Kruskal-Wallis test.

The small p-value of 0.0424 of the Kruskal-Wallis test indicates that the two variables, initiative use and political culture are dependent. Not all political cultures use the initiative process equally. According to the Kruskal multiple comparisons test, we can conclude that states with a moralistic political culture tend to use the initiative more frequently than states with a traditionalistic culture. No other comparisons were significant.

Extension Example: Citizen's Initiative

Again, these conclusions make sense when seeing the graphic:



Example 1: Violent Crime

Example

Does the 2000 violent crime rate significantly vary across the four census regions?

The two hypotheses are

$$H_0 : \mu_N = \mu_S = \mu_M = \mu_W$$
$$H_a : \text{At least one mean differs}$$

Example 1: Violent Crime

Because we are comparing multiple means, we would like to use the ANOVA procedure. It has two assumptions:

- The data come from a Normal distribution in each group
- The data have the same variance across the groups

To check this using R, we run the following

```
shapiroTest(vcrime00 ~ census4)
fligner.test(vcrime00 ~ census4)
```

Example 1: Violent Crime

Here are the results:

```
$adjustment
[1] "Bonferroni (4)"

$results
      Level   p.value
1   Midwest 1.0000000
2 Northeast 1.0000000
3     South 0.2251474
4      West 1.0000000
```

... and ...

```
  Fligner-Killeen test of homogeneity of variances

data:  vcrime00 by census4
Fligner-Killeen:med chi-squared=6.0913, df=3, p-value=0.1073
```


Example 1: Violent Crime

Because no requirement was violated, we can use ANOVA. Here is the code:

```
cmod1 = aov(vcrime00~census4)
summary(cmod1)
```

and the results

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
census4      3  689569  229856   4.855 0.00505 **
Residuals   47 2225268   47346
```

Because the p-value of 0.00505 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. We can conclude that at least one mean differs from the others.

Again, which one?

Example 1: Violent Crime

To determine *which* is different, we use the Tukey HSD test:

```
TukeyHSD(cmod1)
```

The partial output is

```
$census4
                        diff        lwr       upr     p adj
Northeast-Midwest  -55.21111 -310.76041 200.33819 0.9389452
South-Midwest      196.87647  -21.62826 415.38120 0.0910331
West-Midwest       -74.66154 -306.65974 157.33667 0.8266966
South-Northeast    252.08758   13.18663 490.98853 0.0350137
West-Northeast     -19.45043 -270.75207 231.85122 0.9968573
West-South        -271.53801 -485.05941 -58.01661 0.0075817
```

From this, we now we detected a difference in average violent crime rate between the South and Northwest states and the West and South states, where the South is significantly higher than the Northeast and the South is significantly higher that the West.

Example 1: Violent Crime

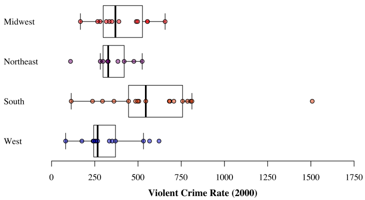
Conclusion:

We are asked to determine if the violent crime rate and the census region are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. Neither assumption is violated. The minimum p-value from the Shapiro-Wilk test is 0.2251, which is greater than $\alpha = 0.05$. The p-value from the Fligner-Killeen test, 0.1073, is also greater than our $\alpha = 0.05$. Because neither assumption is violated, we can — and should — use the analysis of variance procedure.

The small p-value of 0.00505 of the ANOVA test indicates that the two variables, violent crime rate and census region are dependent. Not all census regions have the same average violent crime rate. According to Tukey's HSD test, we can conclude that southern states have a significantly higher average violent crime rate than the Northeastern states and the Western states. No other comparisons were significant.

Example 1: Violent Crime

Again, these conclusions make sense when seeing the graphic:



Example 2: Education

Example

Does the 2000 weighted average educational attainment (WAEA) significantly vary across the three political cultures?

The two statistical hypotheses are

$$H_0 : \text{All means are the same}$$
$$H_a : \text{At least one mean differs}$$

Example 2: Education

Because we are comparing multiple means, we would like to use the ANOVA procedure. It has two assumptions:

- The data come from a Normal distribution in each group
- The data have the same variance across the groups

To check this using R, we run the following

```
shapiroTest(waea00 ~ domPolCulture)
fligner.test(waea00 ~ domPolCulture)
```


Example 2: Education

Here are the results:

```
$adjustment
[1] "Bonferroni (3)"

$results
            Level   p.value
1  Individualistic 1.0000000
2       Moralistic 0.6071567
3 Traditionalistic 1.0000000
```

... and ...

```
  Fligner-Killeen test of homogeneity of variances

data:  waea00 by domPolCulture
Fligner-Killeen:med chi-squared=2.8698, df=2, p-value=0.2381
```

Example 2: Education

Because no requirement was violated, we can use ANOVA. Here is the code:

```
waeamod = aov(waea00~domPolCulture)
summary(waeamod)
```

and the results

```
               Df  Sum Sq  Mean Sq  F value    Pr(>F)
domPolCulture   2   276.0   138.02    17.19 2.35e-06 ***
Residuals      48   385.4     8.03
```

Because the p-value is much less than our usual $\alpha = 0.05$, we reject the null hypothesis. We can conclude that at least one mean differs from the others.

Which one?

Start of Lecture Material
Procedure Requirements
Beyond ANOVA
Four Examples
End of Section Material
Example 1: Violent Crime
Example 2: Education
Example 3: Average Wealth
Example 4: Legislature Professionalism

Example 2: Education

To determine *which* is different, we use Tukey's HSD test:

```
TukeyHSD(waeamod)
```

The re-formatted output is

```
$domPolCulture
                 diff        lwr        upr     p adj
Moral-Indiv  1.725882 -0.6247889  4.076554 0.1885858
Tradt-Indiv -3.840588 -6.1912595 -1.489917 0.0007316
Tradt-Moral -5.566471 -7.9171418 -3.215799 0.0000019
```

From this, we now we detected a difference in average WAEA rate between the traditionalist states and both of the other two types. In both cases, the traditionalistic states tended to have lower WAEA. No other comparisons were statistically significant.

Example 2: Education

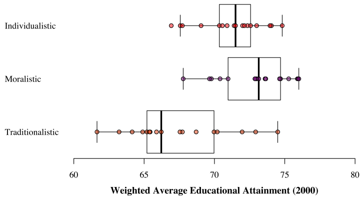
Conclusion:

We are asked to determine if the weighted average educational attainment and the dominant political culture are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. Neither assumption is violated. The minimum p-value from the Shapiro-Wilk test is 0.6072, which is greater than $\alpha = 0.05$. The p-value from the Fligner-Killeen test, 0.2381, is also greater than our $\alpha = 0.05$. Because neither assumption is violated, we can — and should — use the analysis of variance procedure.

The small p-value of 2.35×10^{-6} of the ANOVA test indicates that the two variables, WAEA and dominant political culture are dependent. According to Tukey's HSD test, we can conclude that traditionalistic states have a significantly lower average WAEA than the moralistic and individualistic states. No other comparisons were significant.

Example 2: Education

Again, these conclusions make sense when seeing the graphic:



Example 3: Average Wealth

Example

Does the 2000 GSP per capita significantly vary across the three political cultures?

The two hypotheses are

$$H_0 : \mu_I = \mu_M = \mu_T$$
$$H_a : \text{At least one mean differs}$$

Example 3: Average Wealth

Because we are comparing multiple means, we would like to use the ANOVA procedure. It has two assumptions:

- The data come from a Normal distribution in each group
- The data have the same variance across the groups

To check this using R, we run the following

```
shapiroTest(gspcap00 ~ domPolCulture)
fligner.test(gspcap00 ~ domPolCulture)
```

Example 3: Average Wealth

Here are the results:

```
$adjustment
[1] "Bonferroni (3)"

$results
            Level      p.value
1  Individualistic 7.609321e-03
2       Moralistic 7.771441e-01
3 Traditionalistic 3.239349e-05
```

... and ...

```
  Fligner-Killeen test of homogeneity of variances

data:  gspcap00 by domPolCulture
Fligner-Killeen:med chi-squared=0.7711, df=2, p-value=0.6801
```


Example 3: Average Wealth

Because at least one requirement was violated, we should not use ANOVA. We need to use the Kruskal-Wallis test. Here is the code:

```
kruskal.test(gspcap00 ~ domPolCulture)
```

and the results

```
  Kruskal-Wallis rank sum test

data:  gspcap00 by domPolCulture
Kruskal-Wallis chi-squared=13.752, df=2, p-value=0.0010
```

Because the p-value of 0.0010 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. We can conclude that at least one mean differs from the others.

- Which one?

Example 3: Average Wealth

To determine *which* is different, we use the Kruskal multiple comparisons test:

```
print( kruskal(gspcap00, domPolCulture) )
```

The partial output is

```
$groups
                  gspcap00  groups
Individualistic   36.35294       a
Moralistic        23.82353       b
Traditionalistic  17.82353       b
```

From this, we now we detected a difference in average GSP per capita between the individualistic states and both of the other two types. In both cases, the individualistic states tended to have higher GSP per capita. No other comparisons were statistically significant.

Example 3: Average Wealth

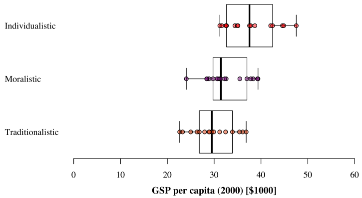
Conclusion:

We are asked to determine if the average GSP per capita and the dominant political culture are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. The Normality assumption is violated. The minimum p-value from the Shapiro-Wilk test is 0.000 033, which is much less than $\alpha = 0.05$. Because an assumption is violated, we need to use the Kruskal-Wallis test.

The small p-value of 0.0010 of the Kruskal-Wallis test indicates that the two variables, GSP per capita and dominant political culture are dependent. According to the Kruskal multiple comparison test, we can conclude that individualist states have a significantly higher GSP per capita than the moralistic and traditionalistic states. No other comparisons were significant.

Example 3: Average Wealth

Again, these conclusions make sense when seeing the graphic:



Example 4: Legislature Professionalism

Example

Does the professional level of the state's legislature significantly vary across the three political cultures?

The two hypotheses are

$$H_0 : \mu_I = \mu_M = \mu_T$$
$$H_a : \text{At least one mean differs}$$

Example 4: Legislature Professionalism

Because we are comparing multiple means, we would like to use the ANOVA procedure. It has two assumptions:

- The data come from a Normal distribution in each group
- The data have the same variance across the groups

To check this using R, we run the following lines

```
shapiroTest(profleg ~ domPolCulture)
fligner.test(profleg ~ domPolCulture)
```


Example 4: Legislature Professionalism

Here are the results:

```
$adjustment
[1] "Bonferroni (3)"

$results
             Level     p.value
1  Individualistic 0.033975358
2       Moralistic 0.002017848
3 Traditionalistic 1.000000000
```

... and ...

```
  Fligner-Killeen test of homogeneity of variances

data:  profleg by domPolCulture
Fligner-Killeen:med chi-squared=5.1794, df=2, p-value=0.0750
```

Example 4: Legislature Professionalism

Because at least one requirement was violated, we should not use ANOVA. We need to use the Kruskal-Wallis test. Here is the code:

```
kruskal.test(profleg ~ domPolCulture)
```

and the results

```
  Kruskal-Wallis rank sum test

data:  profleg by domPolCulture
Kruskal-Wallis chi-squared=6.5777, df=2, p-value=0.0373
```

Because the p-value of 0.0373 is less than our usual $\alpha = 0.05$, we reject the null hypothesis. We can conclude that at least one mean differs from the others.

- Which one?

Example 4: Legislature Professionalism

To determine *which* is different, we use Kruskal's multiple comparisons test:

```
print( kruskal(profleg, domPolCulture) )
```

The partial output is

```
$groups
                  profleg groups
Individualistic  32.82353      a
Moralistic       22.35294      b
Traditionalistic 21.06250      b
```

From this, we now we detected a difference in average level of legislative professionalism between the individualistic states and both of the other two types. In both cases, the individualistic states tended to have higher levels of legislative professionalism. No other comparisons were statistically significant.

Example 4: Legislature Professionalism

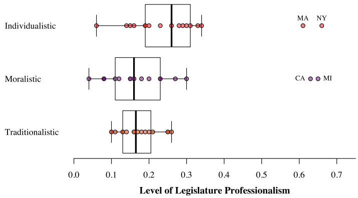
Conclusion:

We are asked to determine if the level of legislative professionalism and the dominant political culture are independent. To do this, we would prefer to use the analysis of variance procedure, because it is the most powerful of the available tests. It has two requirements: The data are from a Normal distribution in each group; and the variances are the same across the groups. The Normality assumption is violated. The minimum p-value from the Shapiro-Wilk test is 0.0020, which is much less than $\alpha = 0.05$. Because an assumption is violated, we need to use the Kruskal-Wallis test.

The small p-value of 0.0373 of the Kruskal-Wallis test indicates that the two variables, legislative professionalism and dominant political culture, are dependent. According to Kruskal's multiple comparisons test, we can conclude that individualistic states tend to have a significantly higher level of legislative professionalism than the moralistic and traditionalistic states. No other comparisons were significant.

Example 4: Legislature Professionalism

Again, these conclusions make sense when seeing the graphic:



Today's Objectives

Now that we have concluded this lecture, you should be able to

1. understand the theory behind testing. . .
   - the means of more than two populations
   - whether a categorical variable helps understand a numeric
   - independence between a numeric and a categorical variable
2. determine if ANOVA or the Kruskal-Wallis test should be used
3. determine *which* level is different using Tukey's HSD or the Kruskal multiple comparisons test

Since we used R to perform the calculations, we were better able to focus on the interpretation than on the tedious calculations.
As always: Please do not forget to be familiar with the `allProcedures` document that lists all of the statistical procedures we will use in R.

Today's R Functions

Here is what we used the following R functions:

- `shapiroTest(x ~ g)`
- `fligner.test(x ~ g)`

- `aov(x ~ g)`
- `TukeyHSD(x ~ g)`

- `kruskal.test(x,g)`
- `print(kruskal(x,g))`

Supplemental Activities

The following activities are currently available from the STAT 200 website to give you some practice in performing hypothesis tests concerning the ANOVA procedure and its extensions.

- SCA 42a
- SCA 42b

Source: `https://www.kvasaheim.com/courses/stat200/sca/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding how to fully test the equality of means:

• Hawkes Learning:	None
• Intro to Modern Statistics:	None
• R for Starters:	Chapter 7
• Wikipedia:	ANOVA Tukey's range test Kruskal-Wallis test Multiple comparisons problem