Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

April 3, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

Analysis of Variance (ANOVA)

The Trouble with Multiple Comparisons

- Often you have more than two groups you want to compare and contrast
- Why not use the two-sample $t-test$ and compare two groups at a time?
- In any single trial we have a certain probability of a significant result by chance alone (α) and hence $1-\alpha$ is the probability of no significant result
- If my experiment has 5 groups (A, B, C, D, and E) and I compare two groups at a time, I am simultaneously testing AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE
- What is the probability that at least one of these pairs throws up a significant result by chance alone? ... this is P(Type I) error $= \alpha = 0.05$

$$
\begin{aligned}
P(\text{no Type I error in 1 comparison}) &= 0.95 \\
P(\text{no Type I error in 2 comparisons}) &= 0.95 \times 0.95 = 0.9025 \\
\text{Note: P(Type I error in 2 comparisons) is} &= 1 - 0.9025 = 0.0975 \\
P(\text{no Type I error in 3 comparisons}) &= 0.95 \times 0.95 \times 0.95 = 0.8573 \\
\text{Note: P(Type I error in 3 comparisons) is} &= 1 - 0.857375 = 0.1426
\end{aligned}
$$

... the probability of making a Type I error is exploding!

Correcting for Multiple Comparisons: Bonferroni

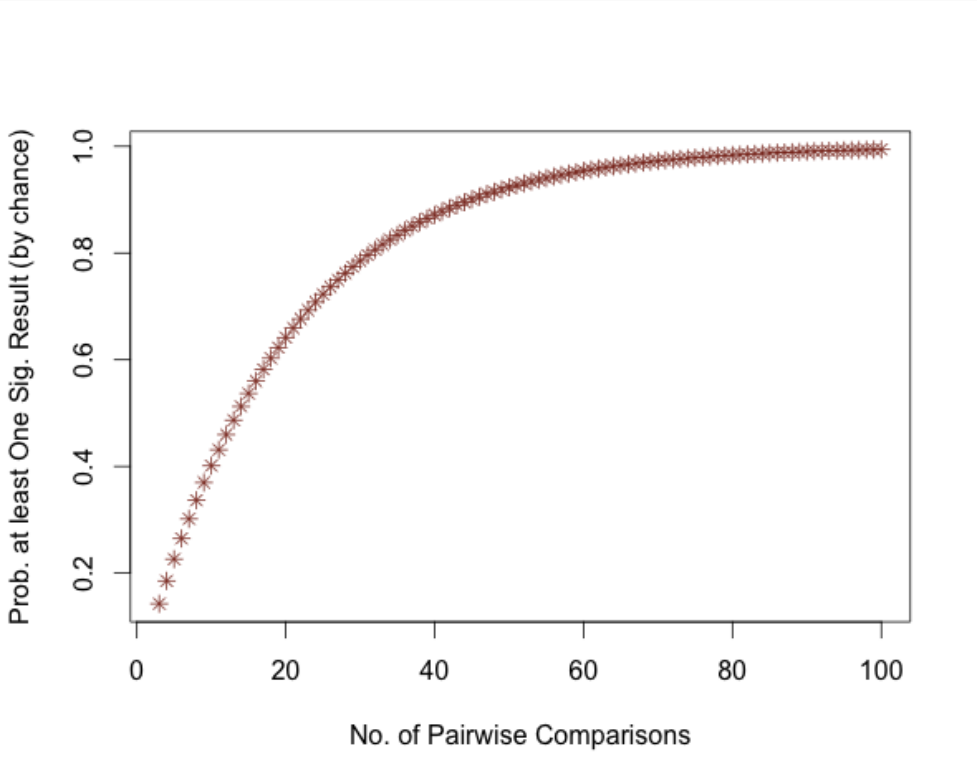


Table 1: $\alpha = 0.05$

Trials	α^*
3	0.0167
4	0.0125
5	0.0100
6	0.0083
7	0.0071
8	0.0063
9	0.0056

Use $\alpha^* = \dfrac{\alpha}{\text{no. of trials}}$ and Reject each H_0 only if $P - value \leq \alpha^*$

So if I have 3 pairwise comparisons (A - B, A - C, B - C) and I am using $\alpha = 0.05$, then use $\alpha^* = \dfrac{0.05}{3} = 0.0166$ for each pairwise comparison

The Logic of ANOVA

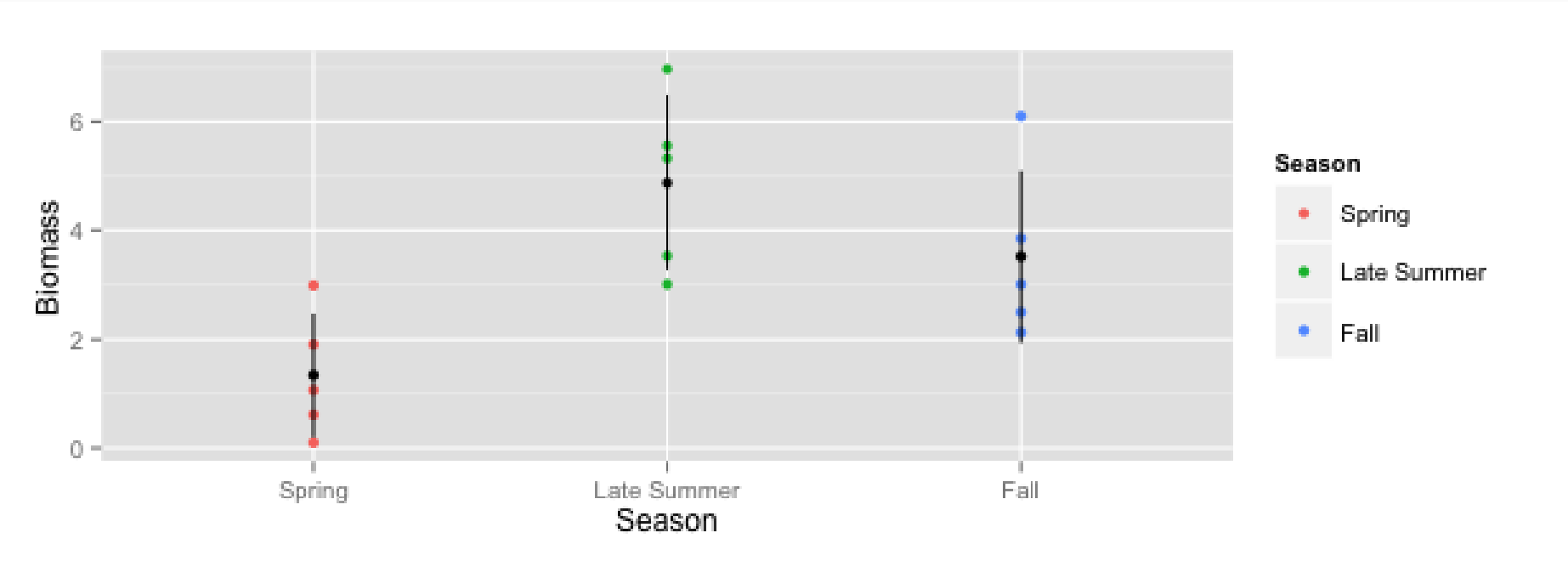
Example

One particularly contentious issue among restoration ecologists is the timing of prairie burns. Although natural fires may primarily have been sparked by late-summer lightning strikes, most controlled burns are done during the spring or fall. The timing of burning may strongly influence the outcome of prairie restorations because burns done at different times of year can favor dramatically different plant species. You could collect data to answer the following question: How does the timing of controlled burns influence the biomass of desirable prairie plant species?

Total biomass (g/m^2) of *Rudbeckia hirta* (Black-eyed Susan) growing in each of 15 $0.5m^2$ plots.

One third of the plots were burned in the spring, late summer, and fall of 1998, respectively. All plots were sampled during summer 1999.

Season of Controlled Burn		
Spring	Late Summer	Fall
0.10	5.56	3.85
0.61	6.97	3.01
1.91	3.01	2.13
2.99	5.33	2.50
1.06	3.53	6.10
$\bar{y}_1 = 1.33$	$\bar{y}_2 = 4.88$	$\bar{y}_3 = 3.52$
$s_1^2 = 1.30$	$s_2^2 = 2.59$	$s_3^2 = 2.50$



Season of Controlled Burn		
Spring	Late Summer	Fall
0.10	5.56	3.85
0.61	6.97	3.01
1.91	3.01	2.13
2.99	5.33	2.50
1.06	3.53	6.10
$\bar{y}_1 = 1.33$	$\bar{y}_2 = 4.88$	$\bar{y}_3 = 3.52$
$s_1^2 = 1.30$	$s_2^2 = 2.59$	$s_3^2 = 2.50$

- Biomass varies in each plot and in each season, this could be something to do with the season or sheer chance (each plot is unique after all)
- Biomass also varies most in Late Summer and least in Spring
- Average biomass is highest in Late Summer and lowest in Spring
- How then do we figure out if the treatment (the season when the controlled burn occurred) really matters or whether what we see is strictly due to chance?

Analysis of Variance (ANOVA)

- ANOVA is a hypothesis testing procedure that allows us to simultaneously compare more than two groups and determine if there are statistically significant differences between the groups
- ANOVA also lets us test the influence of two or more *factors* (i.e., independent variables) on the *response* (i.e., dependent variable)
- The basic test statistic is a ratio ... $\frac{\text{Difference between groups}}{\text{Difference within groups}}$
- If difference between groups $>$ difference within each group, it must be because something makes at least one group different from the other groups. This "something" has to be the treatment since we are assuming everything else has been controlled for (or doesn't influence the response)

How shall we measure and analyze “difference”?

- We could ask: How much does each observation differ from the overall Mean? This would give us total variability in the full sample
- We could also ask: How do the groups (i.e., the treatments) differ on average from the overall mean? This would give us variability between each treatment group.
- We could also ask: In each treatment group, how much does each group member differ from his/her group Mean? This would give us variability within each treatment group
- What would be a good measure of variability? The variance of course!
- So we look at a few variances
  1. Total variation of all scores
  2. Variation between-groups
  3. Variation within-groups

Between-groups and Within-groups Variance

- Within-groups, differences must be due to chance because the treatment is a constant for each group.
- Between-groups, differences may be due to (i) chance and/or (ii) treatments. In other words, if the treatment has an effect it must influence each measurement in more or less the same way
- Test statistic: Ratio of between-groups variance to within-groups variance

$$F = \frac{\text{Variance Between-groups}}{\text{Variance Within-groups}}$$

$$\therefore F = \frac{\text{Variance due to Chance} + \text{Variance due to Treatments}}{\text{Variance due to Chance}}$$

- If variance due to treatments $= 0$, what will F be?
- If variance due to treatments is large, what will F be?

The Elements of ANOVA

- We refer to the independent variable(s) as the factor(s)
  - With just one factor we speak of a single-factor design ... for example, the season of the burn (one factor with three categories, spring, late summer, fall)
  - With two or more factors we speak of a factorial design ... for example, (1) the temperature in a chamber and (2) the fertilizer being used to see impact on plant growth
- The values (categories) of a factor we refer to as the treatments
- The outcome is referred to as the response variable
- Assumptions of ANOVA
  1. The response variable is $\sim N(.)$ (i.e., Normally distributed) ... how would you test this?
  2. The variance (σ^2) of the response variable is the same for all groups ... how would you test this?
  3. The observations are independent within each group (i.e., random sampling was not violated) ... assumedly true since otherwise we would not be doing any statistical test at all

- Recall that the variance is calculated as follows: $\frac{\sum (y_i - \bar{y})^2}{N}$. Here, the numerator ... $\sum (y_i - \bar{y})^2$ is the sum of squares
- We need a common anchor for all the data, something to serve as the basis for however we wish to answer our question. A good starting point thus becomes the overall mean, which we call the grand mean and denote this by $\bar{\bar{y}}$. For the burn data $\bar{\bar{y}} = 3.244$
- How does each plot differ from this overall mean? ... A good measure would be some estimate of the difference between each biomass and the grand mean. This could be done by calculating $y_i - \bar{\bar{y}}$. If we want total variation around the grand mean we would have to sum the square of these differences (since if we don't square the differences their sum will be zero)
- The resulting quantity will be the total sum of squares ... $\sum (y_i - \bar{\bar{y}})^2 = \left((0.10 - 3.244)^2 + (0.61 - 3.244)^2 + \ldots + (2.50 - 3.244)^2 + (6.10 - 3.244)^2\right) = 57.54476$.
- This is all the variation of the full sample around the grand mean, either because individual plots differ (by chance) or because the seasons (the treatments) really creates more/less biomass
- We know we have two parts to total sum of squares , (a) that part due to chance, and (b) that part due to the treatments

- How could we measure how much each Season's average differs from the grand mean? By subtracting each Season's mean from the grand mean, squaring this difference, and then weighting the squared difference by the number of observations in each group ... $\sum \left(\bar{y}_j - \bar{\bar{y}}\right)^2$
- ... $5 \times (1.334 - 3.244)^2 + 5 \times (4.880 - 3.244)^2 + 5 \times (3.518 - 3.244)^2 = 18.2405 + 13.38248 + 0.37538 = 31.99836$ ... This is the Between Groups sum of squares
- Note: We are multiplying the difference between each group mean and the grand mean by the number of observations in each sample to "weight" the calculation otherwise the assumption is that each sample is providing equal information but it may be that $n_1 \neq n_2 \neq n_3$

- How about the variability within each Season (which should be by chance since it is the same Season for all plots)? By subtracting (and squaring) the distance of each plot's biomass from its seasonal average ... We call this the Within Groups sum of squares
- $\sum \left(y_{ij} - \bar{y}_j\right)^2 = (0.10 - 1.334)^2 + (0.61 - 1.334)^2 + \ldots + (1.06 - 1.334)^2$ for Spring, and similarly for Late Summer and for Fall. The resulting sums will be 5.19612, 10.3524 and 9.99788, resulting in the Within Groups sum of squares of 25.5464
- Total sum of squares = Between Groups SS + Within Groups SS ... $57.54476 = 31.99836 + 25.5464$

Mean Square Between Groups

Thus far, we have only looked at the sums of squares and while this is useful, we need to calculate the (a) variance between groups and the (b) variance within groups

When we calculated the variance, what did we do? Well, we did the following for a sample: $s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$

So how can we calculate the variance Between Groups? By dividing the total sum of squares between groups by the number of groups - 1.

This would give us: $\frac{31.99836}{3-1} = 15.99918$

Mean Square Within Groups

The variance Within Groups can be estimated likewise, except by dividing the total sum of squares within groups by the total sample size - number of groups, and this would yield: $\frac{25.5464}{15-3} = 2.128867$

Note: If the denominator in the variance Within Groups does not make sense think of the calculation as follows:

- variance for Spring: $\frac{\sum (y_i - \bar{y_1})^2}{n_1 - 1}$ +
- variance for Late Summer: $\frac{\sum (y_i - \bar{y_2})^2}{n_2 - 1}$ +
- variance for Fall: $\frac{\sum (y_i - \bar{y_3})^2}{n_3 - 1}$
- ... $n_1 - 1 + n_2 - 1 + n_3 - 1 = n_1 + n_2 + n_3 - 3 = n - 3$

The Mechanics of ANOVA

- The Hypotheses: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
  H_a : Not all population means are equal
- i indexes observations; j indexes groups; μ_j is mean of the j^{th} group; n_j is sample size of group j; k is the total number of groups; y_{ij} is score for observation i for group j
- $\bar{y}_j$ is mean for group j

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

- s_j^2 and s_j are variance and standard deviation for group j

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

- $\bar{\bar{y}}$ is the overall sample mean

$$\bar{\bar{y}} = \frac{1}{n_T} \sum_{j=1}^{k} \sum_{i=1}^{n_j} y_{ij}$$

- Note that $n_T = n_1 + n_2 + \cdots + n_k$

Season of Controlled Burn		
Spring	Late Summer	Fall
0.10	5.56	3.85
0.61	6.97	3.01
1.91	3.01	2.13
2.99	5.33	2.50
1.06	3.53	6.10
$\bar{y} = 1.33$	$\bar{y} = 4.88$	$\bar{y} = 3.52$
$s^2 = 1.30$	$s^2 = 2.59$	$s^2 = 2.50$
The overall mean is $\bar{\bar{y}} = 3.24$		

Between-Groups and Within-Groups

- Mean Square due to Groups (MS_{Groups}) is given by $\frac{SS_{Groups}}{k-1}$

$$SS_{Groups} = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{\bar{y}})^2 = 5(1.33-3.24)^2 + 5(4.88-3.24)^2 + 5(3.52-3.24)^2 = 32.00$$

$$MS_{Groups} = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{y}_j - \bar{\bar{y}})^2 = \frac{32.00}{3-1} = \frac{32.00}{2} = 15.99$$

- Mean Square due to Error (MS_{Error}) is given by $\frac{SS_{Error}}{n_T - k}$

$$SS_{Error} = \sum_{j=1}^{k} (n_j - 1)s_j^2 = 4(1.30) + 4(2.59) + 4(2.50) = 25.55$$

$$MS_{Error} = \frac{1}{n_T - k}\sum_{j=1}^{k} (n_j - 1)s_j^2 = \frac{25.55}{15-3} = \frac{25.55}{12} = 2.129$$

- Note also that $SS_{Total} = SS_{Groups} + SS_{Error} = 32.00 + 25.55 = 57.55$

$$SS_{Total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (y_{ij} - \bar{\bar{y}})^2$$

Calculating the F-ratio

- $F = \frac{MS_{Groups}}{MS_{Error}} = \frac{15.99}{2.129} = 7.515$
- $F \sim F_{df_{Numerator};df_{Denominator}} = F_{2,12}$
  1. $df_{Numerator} = k - 1 \cdots$ No. of groups - 1
  2. $df_{Denominator} = n_T - k \cdots$ Total sample size - No. of groups
- Reject H_0 if $P - value \leq \alpha$; Do not reject H_0 otherwise

  Assume we set $\alpha = 0.05$

  For $F = 7.515$ with $df_{Numerator} = 2$ and $df_{Denominator} = 12$ the $P - value = 0.00765$

  Hence we reject H_0; The timing of the controlled burn season does seem to influence biomass

The ANOVA Table

The ANOVA table will look like the following when you run ANOVA in R

```
> lm.ex1 <- lm(Biomass ~ Season, data=ex1)
> anova(lm.ex1)
Analysis of Variance Table
Response: Biomass
          Df Sum Sq Mean Sq F value Pr(>F)
Season     2 31.998 15.9992 7.5154 0.007655 **
Residuals 12 25.546 2.1289
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

Sum Sq is the Sum of Squares; Mean Sq is Mean Squares

Mean Sq for Season is 31.998 divided by 2; Mean Sq for Error (Residuals) is 25.546 divided by 12

F value is Mean Sq Season divided by Mean Sq Error (Residuals)

Pr(>F) is the $P-value$ for the calculated F

** indicates the $P-value$ is < 0.01 and * indicates the $P-value$ is < 0.05

$R^2 = \frac{SS_{Groups}}{SS_{Total}}$ and indicates what proportion of the variation in y is explained by group differences

$0 \leq R^2 \leq 1$; The closers is R^2 to 1 the more the proportion of variation in y explained by group differences

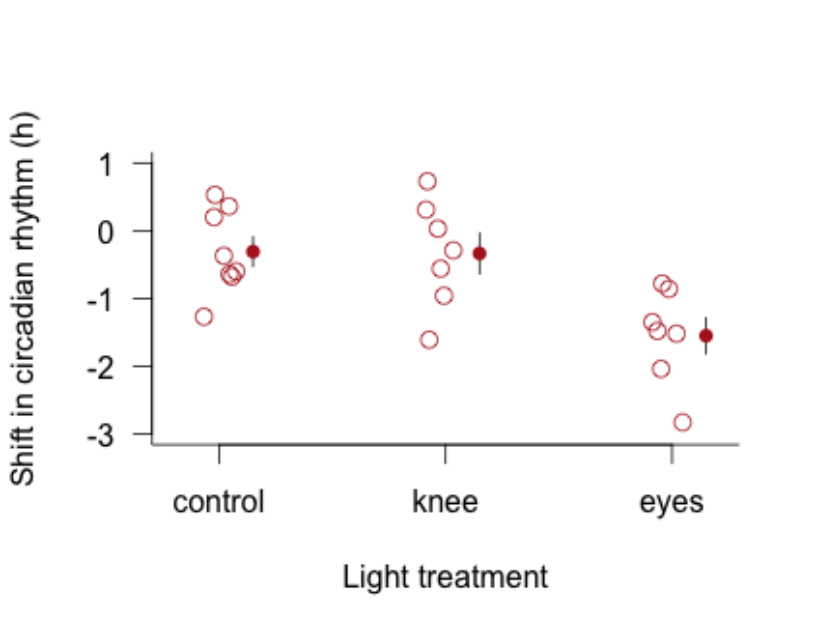
The JetLagKnees Example

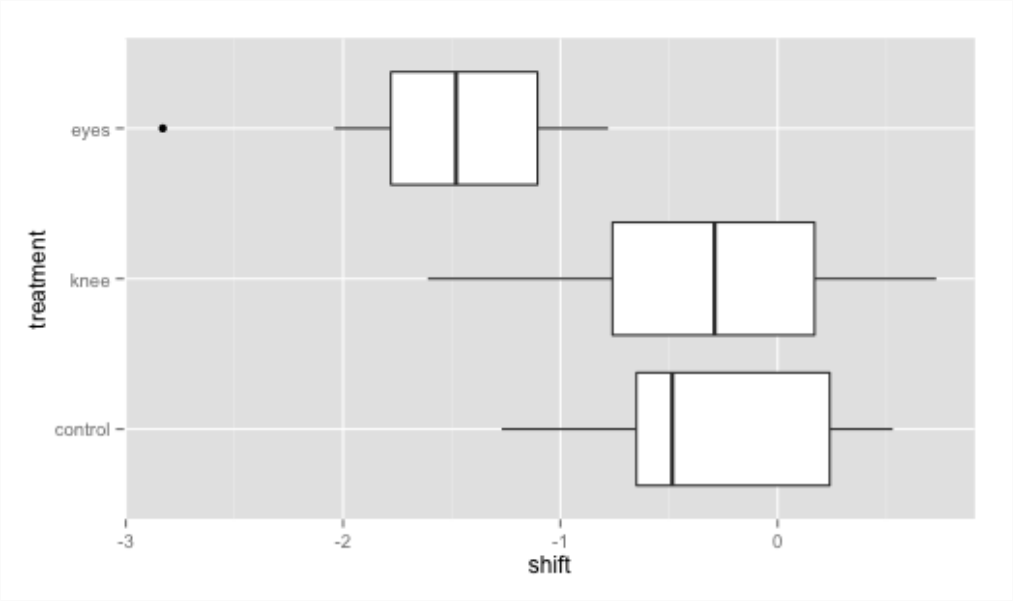
Traveling to a different time zone can cause jet lag but people adjust as the schedule of light to their eyes in the new time zone resets their internal circadian clock. This change in the internal clock is called a phase shift. Researchers suggested that the human circadian clock can also be reset by exposing the back of the knee to a light. This claim met with skepticism and a new experiment was conducted.

This new experiment measured circadian rhythm by the daily cycle of melatonin production in 22 people randomly assigned to one of three light treatments. Participants were awakened from sleep and subjected to a single three-hour episode of bright lights applied (a) to the eyes only, (b) to the knees only, or (c) to neither (the Control group).

The effects of the light treatment were measured two days later by the magnitude of the phase shift in each participant's daily cycle of melatonin production. A negative value indicates a delay in melatonin production (which is what the light treatment should do as hypothesized). Positive values indicate an advance. Does light treatment affect phase shift?

The JetLagKnees Example





Running an ANOVA Test on JetLagKnees

H_0: Average phase shift is equal in all three groups, i.e., $\mu_1 = \mu_2 = \mu_3$

H_A: Average phase shift is not equal in all three groups (i.e., At least one μ_j is different)

Set $\alpha = 0.05$

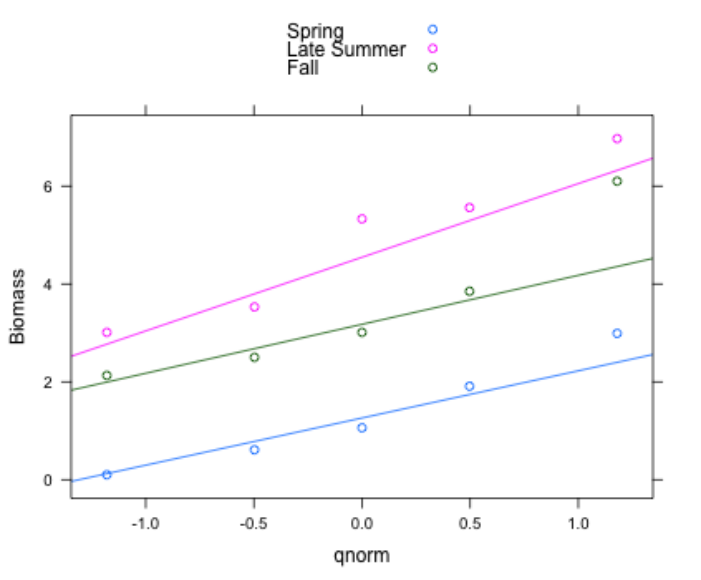
```
> lm.knees = lm(shift ~ treatment, data=circadian)
> anova(lm.knees)
Analysis of Variance Table
Response: shift
          Df Sum Sq Mean Sq F value Pr(>F)
treatment 2 7.2245 3.6122 7.2894 0.004472 **
Residuals 19 9.4153 0.4955
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1        1
```

Since the $P - value = 0.004472$ is less than 0.05 we reject H_0; the data suggest that at least one of the groups has a different mean phase shift as compared to the other two groups

ANOVA Redux

- ANOVA allows us to compare if three or more groups differ by looking at the variance between groups (where each group corresponds to a treatment) versus the variance within groups
- It is built on some assumptions (see below) but is fairly robust to modest violations of assumptions
  1. The response variable is Normally distributed ... minor deviations will not inflate Type I error rates so long as the skew is roughly similar for each group and so long as the samples are large enough (say at least 30 in each group)
  2. The variance (σ^2) of the response variable is the same for all groups ... This is tricky because if this assumption is violated the p-value for the test will be incorrect. You may get away with it if the samples are large, equally sized, and no group's variance is at least 10-times that of another group (or so the text tells you, but use this 10-times rule with caution)
  3. The observations are independent within each group (i.e., random sampling was not violated) ... Don't do any simple test if you don't have a random sample
- ANOVA can be used to compare two-groups but the t-test is more efficient here

Testing Assumptions: Normality



H_0: Biomass is normally distributed in each group (Season)

H_A: Biomass is NOT normally distributed in each group (Season)

```
> with(ex1, tapply(Biomass, Season, shapiro.test))
$Spring
W = 0.961, p-value = 0.8147
$‘Late Summer‘
W = 0.9415, p-value = 0.6763
$Fall
W = 0.879, p-value = 0.3047
```

Testing Assumptions: Variances

The F-test will not work here (because we have more than 2 groups). Levene's would work. Recall the H_0 here: The samples are drawn from populations with homogeneity of variances (i.e., equal variances)

```
> leveneTest(ex1$Biomass ~ ex1$Season, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2  0.3794 0.6922
      12
> leveneTest(ex1$Biomass ~ ex1$Season, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  0.1672 0.8479
      12
```

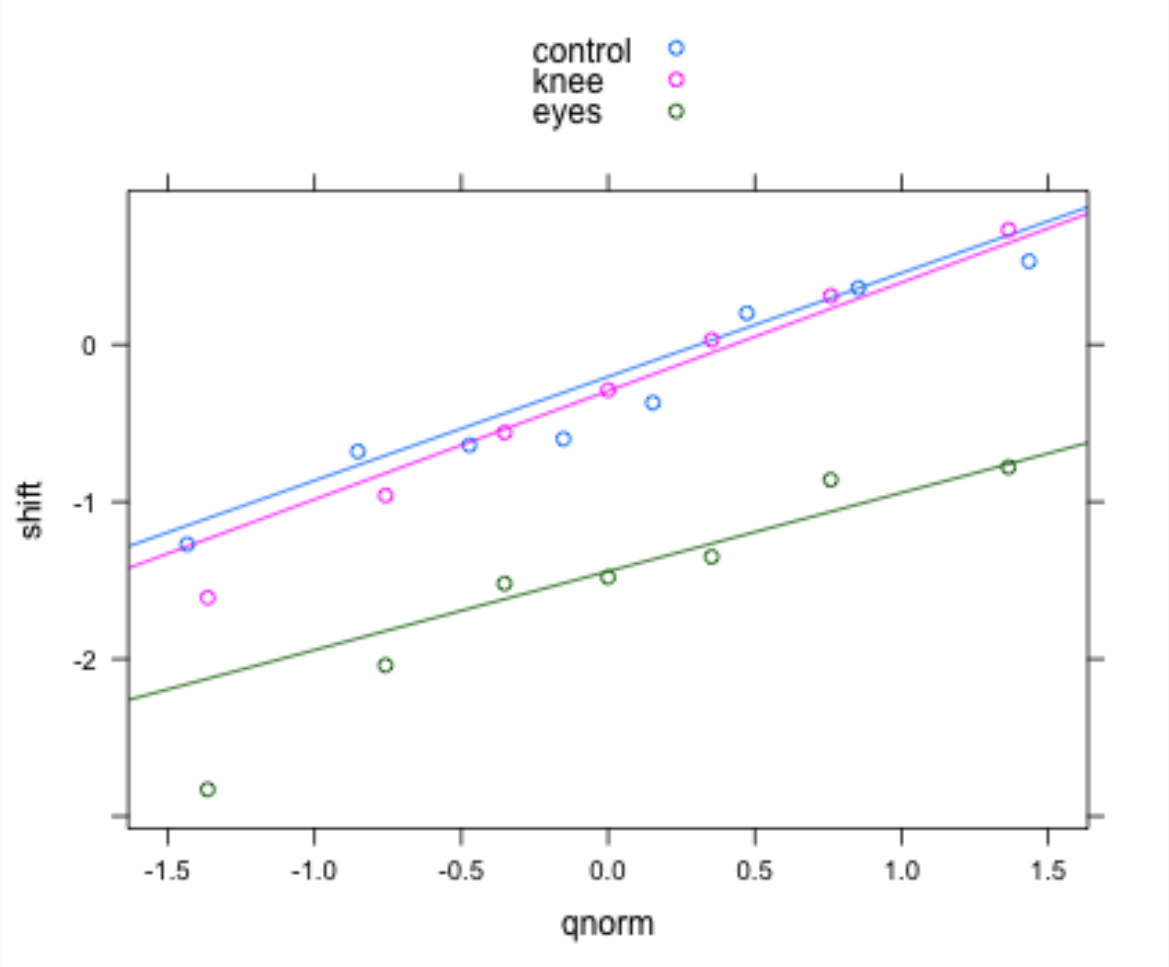
In a nutshell, then, Normality could be assumed and so could equal variances, which mean the ANOVA test results should be okay

If normality was rejected we would have to find a transformation, test for normality, test for equal variances, and then proceed with ANOVA if both normality and equal variances could be assumed

Welch's one-way ANOVA would be appropriate if the data were normal but the variances were unequal

If you have non-normal data no transformation whips into shape, you may have to use a non-parametric test ... Kruskal-Wallis

Testing Assumptions of the Jet Lag ANOVA



```
> with(circadian, tapply(shift, treatment, shapiro.test))
$control
       Shapiro-Wilk normality test
data:  X[[1L]]
W = 0.9329, p-value = 0.5428
$knee
       Shapiro-Wilk normality test
data:  X[[2L]]
W = 0.9887, p-value = 0.9907
$eyes
       Shapiro-Wilk normality test
data:  X[[3L]]
W = 0.9194, p-value = 0.4648

> leveneTest(circadian$shift ~ circadian$treatment, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2  0.1563 0.8564
      19
> leveneTest(circadian$shift ~ circadian$treatment, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  0.1586 0.8545
      19
```

Planned Comparisons

Planned Comparisons

Planned comparisons are comparisons between groups, and identified before the data were gathered (i.e., this was planned during the research design stage)

You have to pick ONE or at most TWO pairs to compare

Essentially no different from the two-sample $t-test$ except in that the standard error is calculated differently: $SE = \sqrt{MS_{Error}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

So for the *knee* versus *control* planned comparison we have

$$SE = \sqrt{0.4955\left(\frac{1}{7}+\frac{1}{8}\right)} = \sqrt{0.1327232} = 0.364312$$

The test involves calculating the difference in the means of the two groups and dividing this by SE to get a t-value. We can then calculate the p-value for this t with $df = N - k = 22 - 3 = 19$

If the resulting p-value is less than 0.05 then we can reject the NULL of no difference between these two groups. Note: R lists t but it is really q

```
Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
knee - control == 0 -0.02696 0.36433 -0.074   0.942
(Adjusted p values reported -- single-step method)
```

in detail now ...

Let us look at the mean shift in each of the three groups ...

```
> with(circadian, tapply(shift, treatment, mean))
   control      knee      eyes
-0.3087500 -0.3357143 -1.5514286
```

Notice that $\bar{y}_{knee} - \bar{y}_{control} = -0.3357143 - (-0.3087500) = -0.0269643$. This is what was reported on the previous slide. Dividing this difference by the $SE = 0.36433$ yields $q = -0.07401065$. Calculating the p-value of this q (see below) shows it to be

```
> 2 * (1 - pt(-0.07401065, df=19, lower.tail=FALSE))
[1] 0.9417756
```

Since this is > 0.05 we cannot reject H_0; there is no statistically significant difference in mean phase shift between the knee and the control group.

Planned Comparison for the Controlled Burn data

Planned ... (assume our interest was in Spring versus Fall)

```
> burnPlanned <- glht(lm.ex1, linfct = mcp(Season = c("Spring - Fall = 0")))
> summary(burnPlanned)

        Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: User-defined Contrasts
Fit: lm(formula = Biomass ~ Season, data = ex1)
Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
Spring - Fall == 0 -2.1840   0.9228 -2.367  0.0356 *
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
(Adjusted p values reported -- single-step method)

> confint(burnPlanned)

        Simultaneous Confidence Intervals
Multiple Comparisons of Means: User-defined Contrasts
Fit: lm(formula = Biomass ~ Season, data = ex1)
Quantile = 2.1788
95% family-wise confidence level
 Linear Hypotheses:
                 Estimate lwr    upr
Spring - Fall == 0 -2.1840 -4.1946 -0.1734
```

Unplanned Comparisons

Unplanned Comparisons: Tukey-Kramer Method

Also known as Tukey's Honestly Significant Differences (HSD) Test or the Tukey-Kramer Test

These are *post hoc* tests because during the research design phase we had no idea which groups to focus on. So this test is akin to saying "ANOVA results show at least one group is different; let us go hunt for which groups are different"

1. Arrange all means in ascending order, then compare two groups one at a time
2. Test statistic: $q = \frac{\bar{y}_1 - \bar{y}_2}{SE}$; SE is as calculated before; and $\bar{y}_1$ is the higher mean of the two group means. Note: R lists t but it is really q
3. p-values are adjusted for the multiple comparisons by multiplying the raw p-value by the number of comparisons being made

```
> circadianTukey <- glht(lm.knees, linfct = mcp(treatment = "Tukey"))
> summary(circadianTukey)
       Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = shift ~ treatment, data = circadian)
Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
knee - control == 0 -0.02696 0.36433 -0.074 0.99699
eyes - control == 0 -1.24268 0.36433 -3.411 0.00776 **
eyes - knee == 0  -1.21571   0.37628 -3.231 0.01165 *
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
(Adjusted p values reported -- single-step method)
```

Unplanned Comparisons for the Controlled Burn data

Time to hunt ...

```
> burnTukey <- glht(lm.ex1, linfct = mcp(Season = "Tukey"))
> summary(burnTukey)

        Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: lm(formula = Biomass ~ Season, data = ex1)
Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Late Summer - Spring == 0 3.5460   0.9228  3.843 0.00586
Fall - Spring == 0        2.1840   0.9228  2.367 0.08438
Fall - Late Summer == 0  -1.3620   0.9228 -1.476 0.33595

Late Summer - Spring == 0 **
Fall - Spring == 0      .
Fall - Late Summer == 0
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1        1
(Adjusted p values reported -- single-step method)
```

Note: Compare the results of this unplanned comparisons to that of the planned comparison. Why the different outcomes?

Welch’s One-Way ANOVA

What if Normality holds but Variances are Unequal?

Should normality hold but variances are unequal, Welch's one-way ANOVA comes in handy

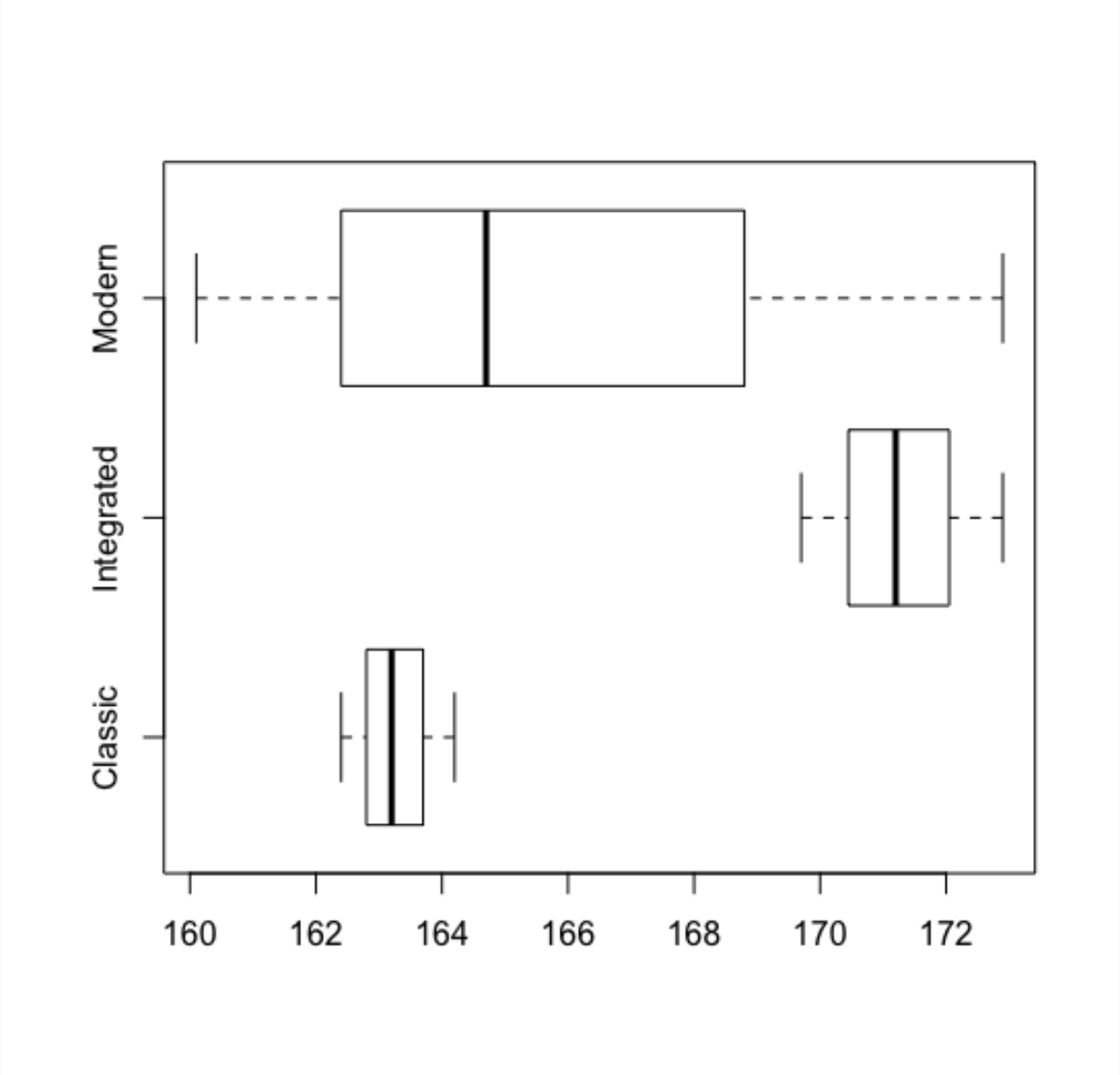
$$W^* = \frac{\sum w_j \left(\bar{y}_j - \hat{\mu}\right)^2 / (k-1)}{1 + \left[2(k-2) / \left(k^2-1\right)\right] \sum h_j}$$

where $w_j = \frac{n_j}{s_j^2}$; $\hat{\mu} = \frac{\sum w_j y_j}{W}$; $W = \sum w_j$; $h_j = \frac{\left(1 - w_j/W^2\right)}{\left(n_j - 1\right)}$, and; $f = \frac{\left(k^2-1\right)}{3\sum h_j}$

$W^* \sim F\left(df_1 = k-1, df_2 = f\right)$

H_0 and H_A are the usual ones for ANOVA

An Example: Skiing Performance



Checking Normality & Equal Variances

```
> with(ski, tapply(score, grips, FUN = shapiro.test))

$Classic
       Shapiro-Wilk normality test
data:  X[[i]]
W = 0.9959, p-value = 0.8777

$Integrated
       Shapiro-Wilk normality test
data:  X[[i]]
W = 0.9987, p-value = 0.9311

$Modern
       Shapiro-Wilk normality test
data:  X[[i]]
W = 0.97406, p-value = 0.8665

> with(ski, leveneTest(score, grips, center = median))

Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  9.5746 0.009925 **
       7
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

The test & Multiple Comparisons

```
> oneway.test(score ~ grips, data = ski, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)

data:  score and grips
F = 24.975, num df = 2.0000, denom df = 4.1069, p-value = 0.005029
```

We can reject Ho; at least one grip-type is different. Given that we have unequal variances and unequal sample sizes, the Games-Howell pairwise tests would be perfectly suited to this situation:

```
> library(userfriendlyscience)
> posthocTGH(y = ski$score, x = ski$grips)
          n means variances
Classic   3   163     0.81
Integrated 3  171     2.56
Modern    4   157   209.98

                    t  df     p
Classic:Integrated 7.54 3.2 0.0084
Classic:Modern    0.83 3.0 0.7138
Integrated:Modern 1.92 3.1 0.2746
```

Only significant difference appears to be between Classic and Integrated.

Nonparametric Alternatives to ANOVA

The Kruskal-Wallis Test

The Kruskal-Wallis test is the nonparametric alternative to the ANOVA

1. Rank all the scores *regardless of which group they belong to*
2. Tied scores get average ranks
3. Add the ranked scores in each group
4. Calculate $H = \left[\frac{12}{N(N+1)} \times \sum \frac{Tc^2}{n_c} \right] - 3(N+1)$

   N is the total number of scores in the study; Tc is the rank total for each group; n_c is the number of units in each group
5. Test statistic is $H \sim \chi^2$ with $df = k - 1$
6. Reject H_0 that the distribution of scores is the same across the groups if p-value is $\leq \alpha$

...

H_0: The populations represented by the k groups have the same distribution of scores on the response variable
H_A: The populations represented by the k groups do not have the same distribution of scores on the response variable

Note: There is no use of the word "normal" in H_0 and H_A; the test is simply asking whether they do or don't come from an identical distribution (whatever that distribution might be)

The test will be weak if you have very differently skewed distributions (say one left the others right) or the variances are very different

Best use may thus be if you have ranked measurements rather than actual scores, and even then only to test whether the mean or median ranks differ

Ranking Wines

To assess the effects of expectation on the perception of aesthetic quality, an investigator randomly sorts 24 amateur wine aficionados into three groups, A, B, and C, of 8 subjects each. Each subject is scheduled for an individual interview. Unfortunately, one of the subjects of group B and two of group C fail to show up for their interviews, so the investigator must make do with samples of unequal size: $n_a = 8$, $n_b = 7$, and $n_c = 6$, for a total of $N = 21$. The subjects who do show up for their interviews are each asked to rate the overall quality of each of three wines on a 10-point scale, with "1" standing at the bottom of the scale and "10" at the top.

As it happens, the three wines are the same for all subjects. The only difference is in the texture of the interview, which is designed to induce a relatively high expectation of quality in the members of group A; a relatively low expectation in the members of group C; and a merely neutral state, tending in neither the one direction nor the other, for the members of group B. At the end of the study, each subject's ratings are averaged across all three wines, and this average is then taken as the raw measure for that particular subject.

Wine Ratings			Wine Ranks		
Group A	Group B	Group C	Group A	Group B	Group C
6.4	2.5	1.3	11	2	1
6.8	3.7	4.1	12	3	4
7.2	4.9	4.9	13	5.5	5.5
8.3	5.4	5.2	17	8	7
8.4	5.9	5.5	18	10	9
9.1	8.1	8.2	19	14	15.5
9.4	8.2		20	15.2	
9.7			21		
			$n_A = 8$	$n_B = 7$	$n_C = 6$
			$T_A = 131$	$T_B = 58$	$T_C = 42$
			$M_A = 16.4$	$M_B = 8.3$	$M_C = 7.0$

$$H = \left[\frac{12}{N(N+1)} \times \sum \frac{Tc^2}{n_c}\right] - 3(N+1)$$

$$H = \left[\frac{12}{21(21+1)} \times \left(\frac{131^2}{8} + \frac{58^2}{7} + \frac{42^2}{6}\right)\right] - 3(21+1)$$

$$H = [0.02597403 \times 2919.696] - 66 = 9.836271$$

$H \sim \chi^2_{df=k-1}$ and here the p-value is 0.007312753 so we reject H_0; the observed distribution differs across the groups.

Kruskal-Wallis on Ranked Data

Cafazzo et al. (2010) observed a group of free-ranging domestic dogs in the outskirts of Rome. Based on the direction of 1815 observations of submissive behavior, they were able to place the dogs in a dominance hierarchy, from most dominant (Merlino) to most submissive (Pisola). Do male and female dogs differ in submissiveness?

```
> head(dogs)
         Dog   Sex Rank
1    Merlino  Male    1
2    Gastone  Male    2
3      Pippo  Male    3
4       Leon  Male    4
5      Golia  Male    5
6 Lancillotto  Male    6

> kruskal.test(Rank ~ Sex, data=dogs)

        Kruskal-Wallis rank sum test

data:  Rank by Sex
Kruskal-Wallis chi-squared = 4.6095, df = 1, p-value = 0.03179
```

Given the low p-value we can conclude that male and female dogs differ in terms of submissive behavior.

Air Quality

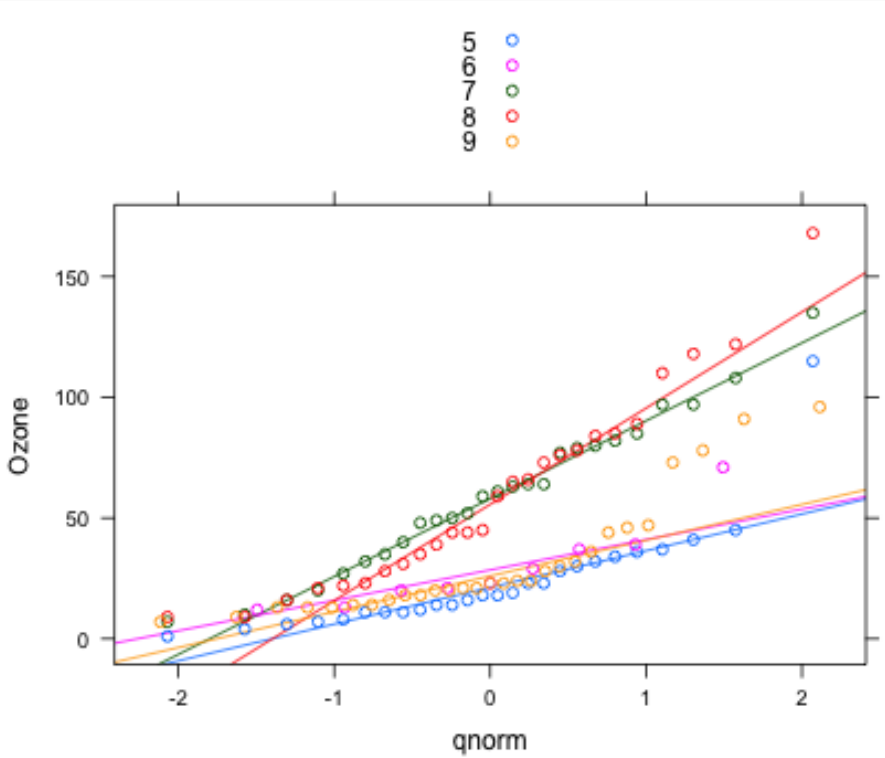
Does airquality vary by month? We have daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973: Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island, NY.

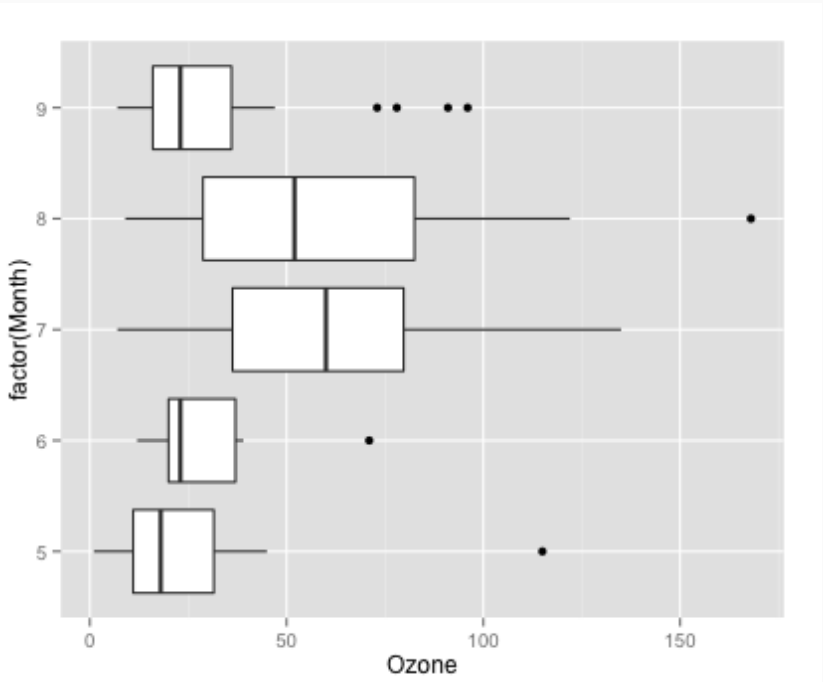
shapiro.test() shows May and September data to be non-normal. Levene's test shows unequal variances.

```
> leveneTest(airquality$Ozone ~ factor(airquality$Month), center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value  Pr(>F)
group  4  4.1019 0.003874 **
     111
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1

> leveneTest(airquality$Ozone ~ factor(airquality$Month), center="median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value  Pr(>F)
group  4  3.9558 0.004863 **
     111
---
Signif. codes: 0   ***   0.001   **   0.01   *   0.05   .   0.1       1
```

The log, square-root, inverse, etc. don't help. So perhaps we need to run the Kruskal-Wallis test






```
> kruskal.test(Ozone ~ Month, data=airquality)
        Kruskal-Wallis rank sum test
data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 6.901e-06
```

Reject H_0; Mean ranks of ozone levels differ across the months.