Lab 5 – Assumptions of ANOVA

September 17 & 18, 2018
FANR 6750

Richard Chandler and Bob Cooper

Today's Topics

1. Assumptions of ANOVA
2. Transformations
3. Non-parametrics

Assumptions of ANOVA

A common misconception is that the response variable must be normally distributed when conducting an ANOVA.

This is incorrect because the normality assumptions pertain to the *residuals*, not the response variable. The key assumption of ANOVA is that the residuals are independent and come from a normal distribution with mean 0 and variance σ^2.

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$$

We can assess this assumption by looking at the residuals themselves or the data within each treatment

Normality diagnostics

Consider the data:

```
infectionRates <- read.csv("infectionRates.csv")
str(infectionRates)

## 'data.frame': 90 obs. of  2 variables:
##  $ percentInfected: num  0.21 0.25 0.17 0.26 0.21 0.21 0.22 0.27 0.23 0.14 ...
##  $ landscape      : Factor w/ 3 levels "Park","Suburban",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(infectionRates)

##  percentInfected    landscape
##  Min.   :0.010   Park    :30
##  1st Qu.:0.040   Suburban:30
##  Median :0.090   Urban   :30
##  Mean   :0.121
##  3rd Qu.:0.210
##  Max.   :0.330
```

These data are made-up, but imagine they come from a study in which 100 crows are placed in $n = 30$ enclosures in each of 3 landscapes. The response variable is the proportion of crows infected with West Nile virus at the end of the study.

ANOVA diagnostics

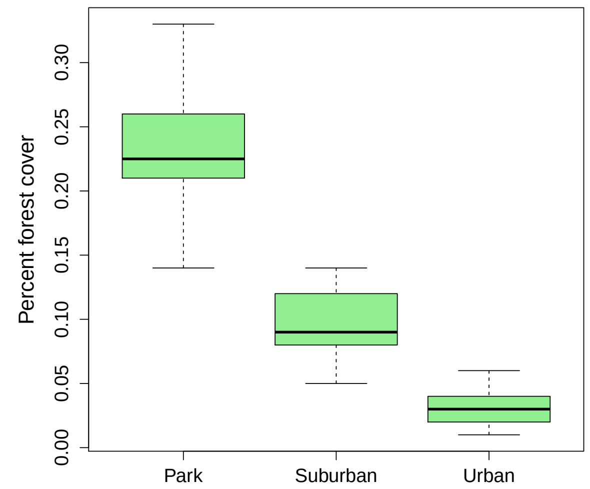
```r
anova1 <- aov(percentInfected ~ landscape,
              data=infectionRates)
summary(anova1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## landscape    2 0.6384  0.3192     306 <2e-16 ***
## Residuals   87 0.0908  0.0010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant, but did we meet the assumptions?

Boxplots

```r
boxplot(percentInfected~landscape, infectionRates,
        col="lightgreen", cex.lab=1.5, cex.axis=1.3,
        ylab="Percent forest cover")
```



Are group variances equal?

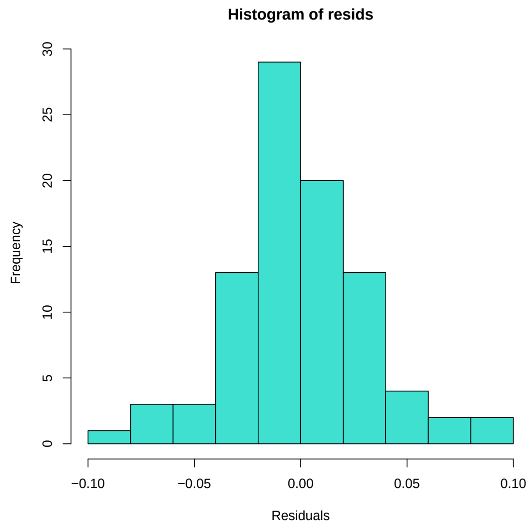
```r
bartlett.test(percentInfected~landscape, data=infectionRates)

##
##  Bartlett test of homogeneity of variances
##
## data:  percentInfected by landscape
## Bartlett's K-squared = 42.926, df = 2, p-value = 4.773e-10
```

We reject the null hypothesis that the group variances are equal

Histogram of residuals

```r
resids <- resid(anova1)
hist(resids, col="turquoise", breaks=10, xlab="Residuals")
```



Normality test on residuals

```
shapiro.test(resids)

##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.95528, p-value = 0.003596
```

We reject the null hypothesis that the residuals come from a normal distribution. Time to consider transformations and/or nonparametric tests.

Logarithmic Transformation

$$y = \log(u + C)$$

- The constant C is often 1, or 0 if there are no zeros in the data (u)
- Useful when group variances are proportional to the means

Square Root Transformation

$$y = \sqrt{u + C}$$

- C is often 0.5 or some other small number
- Useful when group variances are proportional to the means

Arcsine-square root Transformation

$$y = \arcsin(\sqrt{u})$$

- Used on proportions.
- logit transformation is an alternative: $y = \log(\frac{u}{1-u})$

Reciprocal Transformation

$$y = \frac{1}{u + C}$$

- C is often 1 but could be 0 if there are no zeros in u
- Useful when group SDs are proportional to the squared group means

ANOVA on transformed data

Tranformation can be done in the `aov` formula

```
anova2 <- aov(log(percentInfected)~landscape,
              data=infectionRates)
summary(anova2)

##              Df Sum Sq Mean Sq F value Pr(>F)
## landscape     2  60.93   30.46   303.5 <2e-16 ***
## Residuals    87   8.73    0.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we fail to reject the normality assumption – good news

```
shapiro.test(resid(anova2))

##
##  Shapiro-Wilk normality test
##
## data:  resid(anova2)
## W = 0.97092, p-value = 0.04106
```


Non-parametric Tests

Wilcoxan rank sum test

- For 2 group comparisons
- a.k.a. the Mann-Whitney U test
- `wilcox.test`

Kruskal-Wallis One-Way ANOVA

- For testing differences in > 2 groups
- `kruskal.test`

These two functions can be used in almost the exact same way as `t.test` and `aov`, respectively.

Assignment

(1) Decide which transformation is best for the `infectionRates` data by conducting an ANOVA on the untransformed and transformed data. Use graphical assessments, Bartlett's test, and Shapiro's test to evaluate each of the following tranformations:
   - log
   - square-root
   - acrsine square-root
   - reciprocal

(2) Does transformation alter the conclusion about the null hypothesis of no difference in means? If not, were the transformations necessary?

(3) Test the hypothesis that infection rates are equal between suburban and urban landscapes using a Wilcoxan rank sum test. What is the conclusion?

(4) Conduct a Kruskal-Wallis test on the data. What is the conclusion?

Use comments in your R script to explain your answers. Upload your results to ELC at least one day before your next lab.