Adjusting Standard ANOVA Methods to Account for Heterogeneous Variances With an Application to Turfgrass Management

Michael Dumelle

November 30, 2020

Section 1

Introduction

Background

- Research statistician at the Environmental Protection Agency
- PhD in Statistics from Oregon State University (2020)
- Met Alec Kowalewski and Clint Mattox through OSU Statistics Consulting Practicum
  - Encourage you to sign up!
  - Long format vs drop-in
  - Faculty are encouraged too - separate process
  - https://stat.oregonstate.edu/content/consulting-services

Accessing Slides

- I will interweave R code to illustrate ideas (I will also provide SAS code!)

```
# this is a comment
this_is_an_object <- this_is_a_function(this_is_an_argument)
this_is_an_object
#> [1] "this is output"

# mean of 1, 2, 3
x <- c(1, 2, 3)
mean(x)
#> [1] 2
```

- Slides and code available on my GitHub
    - https://github.com/michaeldumelle/OSUHort_11302020
    - https://michaeldumelle.github.io/ → CV → Presentations
- Slide numbers in bottom left

Background

- Use Analysis of Variance (ANOVA) to study designed experiments
  - Are there *statistically significant differences* among treatment effects?
- One common problem: unequal variance within treatment groups
  - Non constant variance, heterogeneous variance
  - Focus of the talk!
- Will be an **initial introduction** to addressing this problem using GVANOVA
  - An **Illustration of concepts**, not an exhaustive comparison of ANOVA and GVANOVA

Section 2

ANOVA

ANOVA Overview

- Often use ANOVA to analyze data from a designed experiment
  - Focus on one-way ANOVA with categorical (group) structure
  - Response = True Mean + Treatment Effect + Random Error
  - $Y_i = \mu + \alpha_i + \epsilon_i, i = 1, ..., n,$ Var$(\epsilon_i) = \sigma^2)$
- Several attractive propreties **when assumptions are satisfied** (accurate, precise, p-values reliable)
- One important assumption **constant variance (homogeneous variance)**
  - All ϵ_i have the same variance (standard deviation)
  - standard deviation = $\sqrt{\text{variance}}$
- Explore ANOVA on percent green cover data having non constant variance (heterogeneous variance)

Percent Green Cover Data

Figure 1: Healthy vs. Non-Healthy Turfgrass. Percent green cover is the proportion of healthy turfgrass.

Percent Green Cover Data

- Use simulated data to study analysis methods
  - So helpful because we know the truth!
  - Study several scenarios without having to design an experiment, collect data, etc.

Table 1: Treatment Means, Standard Deviations (StDev), and Replicates

Treatment	Mean	StDev	Replicates
Trt1	50	5.0	8
Trt2	50	2.0	8
Trt3	58	1.0	8
Trt4	60	0.5	8

Percent Green Cover Data

```
set.seed(1130)
data <- create_data(treatments = c("Trt1", "Trt2", "Trt3", "Trt4"),
                    means = c(50, 50, 58, 60),
                    stdevs = c(5, 2, 1, 0.5),
                    replicates = c(8, 8, 8, 8))
head(data, n = 9)
#>   treatments pct_green
#> 1       Trt1  43.98231
#> 2       Trt1  54.94049
#> 3       Trt1  45.64911
#> 4       Trt1  50.33370
#> 5       Trt1  45.03723
#> 6       Trt1  54.45938
#> 7       Trt1  47.62064
#> 8       Trt1  40.34577
#> 9       Trt2  50.07621
```

Visualizing the Data

- Visualization always a good first step – notice the difference in spread!

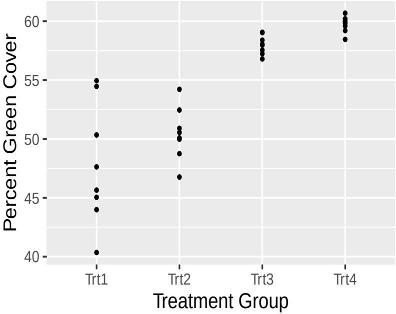

Figure 2: Percent Green Cover by Treatment Group.

ANOVA Code

- Let's perform an ANOVA assuming constant variance – pretend we don't know the truth!

```
# Perform the ANOVA
anova_model <- gls(pct_green ~ treatments, data = data) # nlme package
# Pairwise comparisons among treatments
anova_trtmeans <- emmeans(anova_model, "treatments") # emmeans package
pairs(anova_trtmeans, adjust = "bonferroni") # emmeans package
```

```
# SAS Code
proc mixed data=data;
  class treatments;
  model pct_green = treatments;
  lsmeans treatments / diff adjust=BON;
run;
```

ANOVA Results

Table 2: ANOVA Pairwise Comparison Results

contrast	estimate	SE	df
Trt1 - Trt2	-2.660	1.424	28
Trt1 - Trt3	-10.209	1.424	28
Trt1 - Trt4	-11.946	1.424	28
Trt2 - Trt3	-7.548	1.424	28
Trt2 - Trt4	-9.286	1.424	28
Trt3 - Trt4	-1.737	1.424	28

- Next we need to check assumptions!

ANOVA Residuals

- Commonly use ANOVA residuals to check assumptions
- Recall the ANOVA model: $Y_i = \mu + \alpha_i + \epsilon_i$
  - Fitted values: $\hat{\mu} + \hat{\alpha}_i$ (group mean)
  - Residual: $Y_i - (\hat{\mu} + \hat{\alpha}_i)$ (observed value minus fitted value)
- Use residuals (normalized) to check constant variance assumptions!
  - Residuals divided by their estimated standard deviation are normalized residuals
- Fitted vs residuals plot should show even spread around zero if variance is constant

ANOVA Residuals

- That variance does NOT look constant!

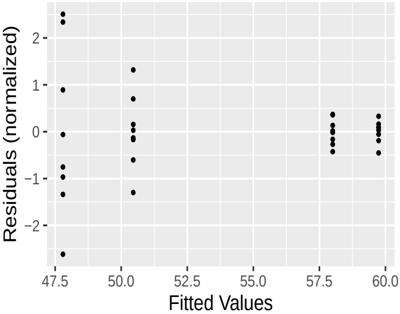

Figure 3: Fitted Values vs Normalized residuals Using ANOVA for Percent Green Cover

Can We Trust Our Analysis?

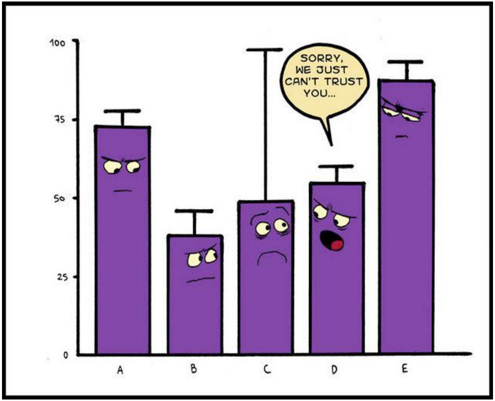

Figure 4: Can we trust our ANOVA results when the variance is not constant?

Section 3

Warning Signs?

Warning Signs?

- In addition to the fitted vs residuals plot, were there any other warning signs?
  - YES
- What were they?

1. Graphics of the Data
2. Ratio of largest variance and smallest variance
3. Statistical hypothesis tests for constant variance

Graphics

- It it looks off, it probably is!
- Similar to the spread we saw in fitted vs residuals plot

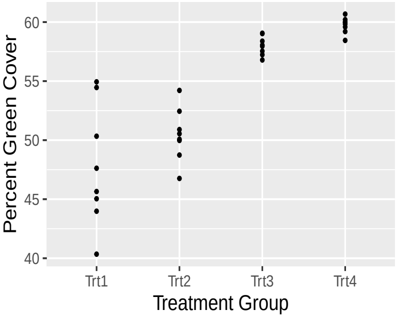

Figure 5: Percent Green Cover by Treatment Group.

Variance Ratios

- Rule of thumb: ANOVA problems when variance ratios larger than 1.5 to 9 (seen cutoff suggestions within this range)
    - Standard deviation range of 1.22 to 3

```
(trt_stdevs <- data %>%
  group_by(treatments) %>%
  summarize(grp_stdev = sd(pct_green)))
#> # A tibble: 4 x 2
#>   treatments grp_stdev
#>   <fct>          <dbl>
#> 1 Trt1           5.13
#> 2 Trt2           2.25
#> 3 Trt3           0.809
#> 4 Trt4           0.678
(stdev_ratio <- max(trt_stdevs$grp_stdev) /
    min(trt_stdevs$grp_stdev)) # much higher than 3!
#> [1] 7.563561
```

Statistical Tests for Constant Varince

- Hypothesis test constant variance assumption is questioned
  - Levene's test, Brown-Forsythe test are two examples – there are many others
  - Come with their own assumptions
  - Low p-value → evidence the variances are NOT equal

```
leveneTest(pct_green ~ treatments,
           center = "mean", data = data) # car package
#> Levene's Test for Homogeneity of Variance (center = "mean")
#>       Df F value    Pr(>F)
#> group  3  9.0112 0.0002445 ***
#>       28
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What Now?

- We know constant variance assumption is invalid – what now?
- We could transform the response variable
  - Hope the transformed data has constant variance
- This approach can be very useful!
- But there are some drawbacks!

What Now?

1. Generally require relationship between mean and variance to be successful
   - Example: Log transformations successful when mean increases $\rightarrow$ variance increases
2. Analysis on transformed scale – NOT original scale
   - Statistically significant difference on transformed scale does not necessarily imply a statistically significant difference on the original scale

Transformations

- Most common is the log transformation – lets hope this works!

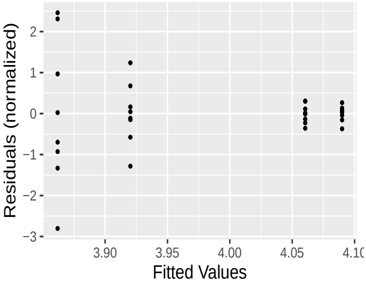

Figure 6: Fitted Values vs Normalized Residuals Using ANOVA for Log$_e$ of Percent Green Cover

- Square root, cube root, reciprocal transformations don't work either – we need another approach!

Section 4

GVANOVA

What is GVANOVA?

- Can use Generalized Variance ANOVA (GVANOVA) to directly model variances within groups
  - Separate variance for each group, Var($\epsilon_i = \sigma_g^2$)
  - No mean / variance relationship required
  - Analysis on original scale
  - More variance parameters require estimation
- Goal here is to introduce an alternative approach to transformations
  - Important to be aware of both – transformations are still a useful tool in the toolbox!

GVANOVA Code

- Let's perform an GVANOVA

```
# Perform the GVANOVA
gvanova_mod <- gls(pct_green ~ treatments,
                   weights = varIdent(form = ~ 1 | treatments),
                   data = data) # nlme package
# Pairwise comparisons among treatments
gvanova_trtmeans <- emmeans(gvanova_mod, "treatments") # emmeans package
pairs(gvanova_trtmeans, adjust = "bonferroni") #emmeans package
```

```
# SAS Code
proc mixed data=data;
  class treatments;
  model pct_green = treatments / ddfm=SAT; # this is different
  repeated / group = treatments; # this is different
  lsmeans treatments / diff adjust=BON;
run;
```

GVANOVA Analysis

Table 3: GVANOVA Pairwise Comparison Output

contrast	estimate	SE	df
Trt1 - Trt2	-2.660	1.980	9.589
Trt1 - Trt3	-10.209	1.836	7.349
Trt1 - Trt4	-11.946	1.829	7.246
Trt2 - Trt3	-7.548	0.844	8.784
Trt2 - Trt4	-9.286	0.829	8.263
Trt3 - Trt4	-1.737	0.373	13.586

GVANOVA Residuals

- Use residuals (normalized) to check assumptions!
- Even spread yields evidence the GVANOVA assumptions are satisfied

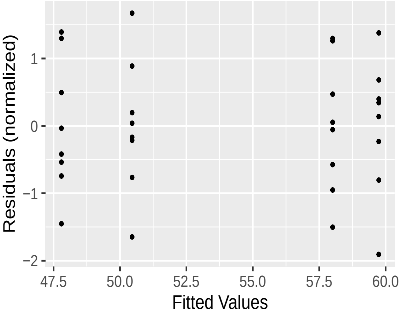

Figure 7: Fitted Values vs Normalized Residuals Using GVANOVA for Percent Green Cover

ANOVA vs GVANOVA

Table 4: Standard Errors and P-values of ANOVA (*.a) and GVANOVA (*.gva). True contrast standard errors (c.se) and differences (c.d) are provided for context.

contrast	c.se	SE.a	SE.gva	c.d	p.a	p.gva
Trt1 - Trt2	1.904	1.424	1.980	0	0.434	1.000
Trt1 - Trt3	1.803	1.424	1.836	-8	0.000	0.004
Trt1 - Trt4	1.777	1.424	1.829	-10	0.000	0.002
Trt2 - Trt3	0.791	1.424	0.844	-8	0.000	0.000
Trt2 - Trt4	0.729	1.424	0.829	-10	0.000	0.000
Trt3 - Trt4	0.395	1.424	0.373	-2	1.000	0.002

- More uncertainty refleted in `Trt1 - Trt2`
- Less uncertainty reflected in `Trt3 - Trt4`

Section 5

Takeaways

Takeaways

- ANOVA is the best tool we have when assumptions are satisfied
- Constant variance assumption should not be overlooked
    - Remember the warning signs!
- Two approaches: transformations and GVANOVA
- When true variance is not constant, using an analysis approach accomodating this will generally yield **a more accurate representation of the truth**

Additional Resources

- R: *Mixed Effects Models and Extensions in Ecology with R* by Alain Zuur Et al. 2009.
  - Chapter 4
- SAS: *SAS for Mixed Models* by Ramon C. Littell Et al. 2006.
  - Chapter 9

Acknowledgements

Thank you to

- Everyone here!
- Horticulture Department at Oregon State University
- Special thanks to Alec Kowalewski and Clint Mattox