LECTURE 10: ANOVA

ENVS475: Exp. Design and Analysis

Spring 2023

outline

1) Overview

2) ANOVA as a linear model

3) ANOVA table

4) Multiple Comparisons

general idea

Extension of the t-test for comparing > 2 populations

motivating example

Ecologists are interested in whether or not tree density changes across elevations. Sample 5 plots (replicates) at 3 elevations (levels).

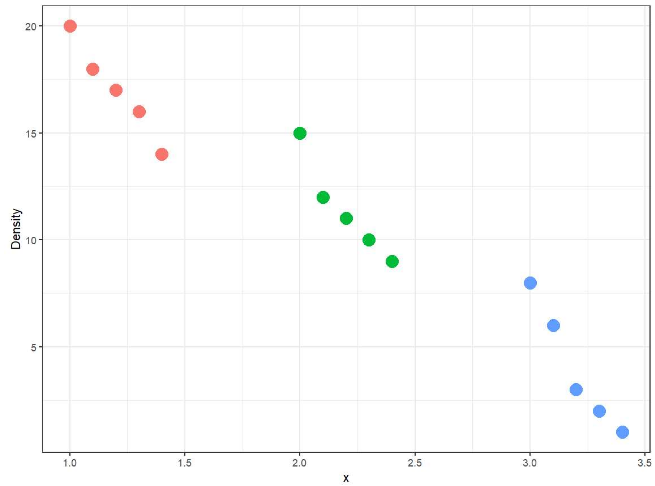
	Elevation		
Replicate	low	medium	high
1	16	10	2
2	14	11	6
3	18	15	8
4	17	9	1
5	20	12	3

Notation

- There is a single factor, elevation.
- The number of groups (AKA treatments, levels) is $k = 3$ (high, medium, low)
- The number of observations within each group (replicates) is $n = 5$
- y_{ij} denotes the jth observation from the ith group

motivating example

Are there differences in tree densities at different elevations?



Hypotheses

- $H_0 : \mu_{low} = \mu_{medium} = \mu_{high}$
- H_a : At least one inequality

How should we test the null?

We could do this using 3 *t*-tests

But this would alter the overall (experiment-wise) α level because each individual test has a chance (usually $\alpha = 0.05$) of incorrectly rejecting a true null hypothesis, and this is multiplied when multiple tests are used

An alternative procedure involves comparing the variation among the groups with the variation within the groups. If H_0 is false, then the variance among is greater than the variance within groups.

Analysis of Variance: ANOVA

As the name implies, this is a method for partitioning the variance into different components; the *signal* and the *noise*.

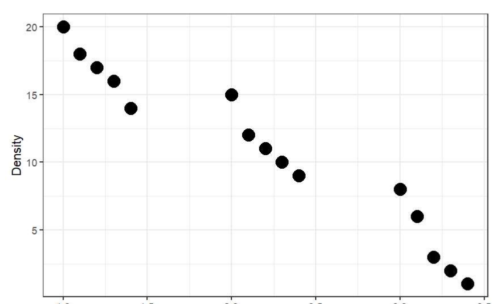
$$\frac{\text{signal}}{\text{noise}}$$

If the treatment (signal) is greater than the variation (noise), we can conclude that there is at least one difference between the groups.

To calculate the signal and the noise, we need to calculate the total variation, the among-group variation, and the within-group variation.

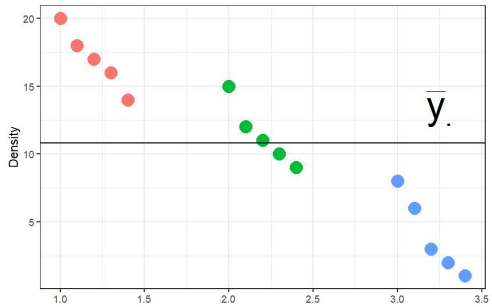
partitioning the variance

Let's look at all of the observations, ignoring the groups



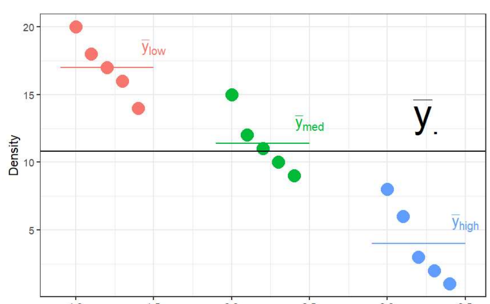
partitioning the variance

Now, let's plot the groups by color, and put a reference line at the global mean ($\bar{y}.$):



partitioning the variance

Now, let's add the group means:



the sum of squares

Now that we have our individual observation (y_{ij}), our global mean ($\bar{y}.$), and the group means ($\bar{y}_i$), we can estimate the variance using modified sum of squares equations.

General formula:

$$SS = \sum_i (\text{observation} - \text{mean})^2$$

- Recall that the Sum of Squares is also how we calculate variance using the `var()` function

Sums of Squares

Variation among groups (treatment effect, or *signal*).

- Group mean - global mean

$$SS_{treatment} = n \sum_i (\bar{y}_i - \bar{y}.)^2$$

Variation within groups (*noise*).

- group observation - group mean (AKA SS_{error})

$$SS_{residual} = \sum_j \sum_i (y_{ij} - \bar{y}_i)^2$$

Total Variation

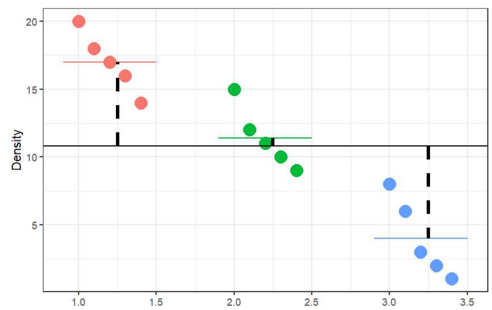
- observations - global mean

$$SS_{total} = SS_{treat} + SS_{resid} = \sum_j \sum_i (y_{ij} - \bar{y}.)^2$$

Sums of Squares

Variation among groups: *signal*.

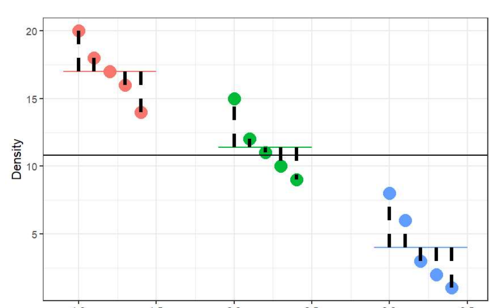
$$SS_{treatment} = n\sum_i (y_i - \bar{y}.)^2$$



Sums of Squares

Variation within groups: *noise*.

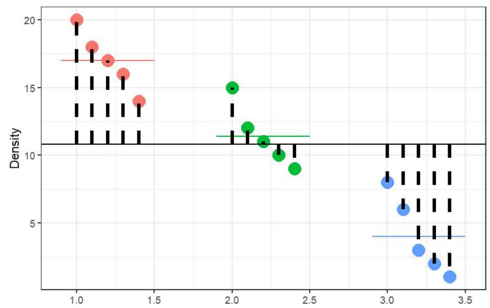
$$SS_{residual} = \sum_j \sum_i (y_{ij} - \bar{y}_i)^2$$



Sums of Squares

Total Variation

$$SS_{total} = \sum_j \sum_i (y_{ij} - \bar{y}.)^2$$



mean squares

To covert the sums of squares to variances, divide by the degrees of freedom

Mean squares among

$$MS_{treat} = \frac{SS_{treat}}{k-1}$$

Mean squares within

$$MS_{resid} = \frac{SS_{resid}}{k(n-1)}$$

F-statistic

$$F_{value} = \frac{MS_{treat}}{MS_{resid}}$$

To test the null hypothesis

- Calculate p-value of F-value
- F-distribution described by two df values
- `pf(f_val, df1, df2, lower.tail = FALSE)`

anova table

anova table

Source	df	SS	MS	F
Among groups	$k-1$	$n\sum_i(\bar{y}_i - \bar{y}.)^2$	$\frac{SS_{treat}}{k-1}$	$\frac{MS_{treat}}{MS_{resid}}$
Within groups	$k(n-1)$	$\sum_i\sum_j(y_{ij} - \bar{y}_i)^2$	$\frac{SS_{treat}}{k(n-1)}$	
Total	$kn-1$	$\sum_i\sum_j(y_{ij} - \bar{y}.)^2$		

ANOVA table from `lm()` in R

We can fit a linear model in R and use the `anova()` function

```
pine_lm <- lm(value ~ Elevation, data = pine_long)
anova(pine_lm)
```

```
## Analysis of Variance Table
##
## Response: value
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Elevation  2  425.2 212.600  33.925 1.152e-05 ***
## Residuals 12   75.2   6.267
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Calculate F value:
  - $MS_{treat}/MS_{resid} = 212.6/6.267 = 33.9$
- Calculate p-value:
  - `pf(33.925, 2, 12, lower.tail = FALSE)`

calculate ANOVA table results from `lm()` summary

Residuals

- `lm()` also returns residuals (e.g., $y_i - E[y_i]$)

```
pine_lm$residual[1:5]
```

```
##             1             2             3             4             5
##  3.000000e+00  1.000000e+00  2.997602e-15 -1.000000e+00 -3.000000e+00
```

```
sum(pine_lm$residuals^2)
```

```
## [1] 75.2
```

- This is the $SS_{residual}$ in the ANOVA table

calculate ANOVA table results from `lm()` summary

Residuals

What about among group variation?

```
pine_lm$fitted.values
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 17.0 17.0 17.0 17.0 17.0 11.4 11.4 11.4 11.4 11.4  4.0  4.0  4.0  4.0  4.0
```

```
sum((pine_lm$fitted.values - mean(pine_lm$fitted.values))^2)
```

```
## [1] 425.2
```

- So the model is the same, the only difference is *how* we present the results

Interpret ANOVA table

```
##        term df sumsq     meansq statistic      p.value
## 1 Elevation  2 425.2 212.600000  33.92553 1.151869e-05
## 2 Residuals 12  75.2   6.266667        NA           NA
```

> Based on the data, we can reject the null hypothesis and conclude that there is at least one difference in the mean tree density across elevations (one-way ANOVA: $F_{2,12} = 33.925, p < 0.001$)

But how do we know which groups are different?

- linear model summary
- Multiple Comparisons

ANOVA as a linear model

General form

$$y_j = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \epsilon_j$$

R model Output

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

Named coefficients

$$y_j = \beta_0 + \beta_{med}\, x_{med} + \beta_{high}\, x_{high} + \epsilon_j$$

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

Before we can interpret this output, we need to understand how R fits this model

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

The model matrix

```
model.matrix(pine_lm)[c(1:2, 6:7),]
```

```
##   (Intercept) Elevationmedium Elevationhigh
## 1           1               0             0
## 2           1               0             0
## 6           1               1             0
## 7           1               1             0
```

- One row for each observation
- Intercept = reference level (alphabetical order by default)
- medium and high treated as *dummy variables* (0/1)

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

The model matrix

```
##   (Intercept) Elevationmedium Elevationhigh
## 1           1               0             0
## 6           1               1             0
```

- Multiplied by the vector of model coefficients β_0, β_1, β_2 to get $E[y_i]$
- `R` names the coefficients `Intercept`, `Elevationmedium`, `Elevationhigh`
- e.g., row 1 = $E[y_1] = Intercept \times 1 + Elevationmedium \times 0 + Elevationhigh \times 0$
- e.g., row 6 = $E[y_1] = Intercept \times 1 + Elevationmedium \times 1 + Elevationhigh \times 0$

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

How do we interpret the coefficients?

- `Intercept` is the expected count at a low elevation site
  - **Note** I set "low" to be the reference value
  - By default R would set reference value alphabetically ("high")
- `Elevationmedium` is the *difference* between medium and low elevation
- `Elevationhigh` is the *difference* between high and low elevation

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

How do we interpret the Intercept p-value?

- Null hypothesis is that $\beta_0 = 0$
- Essentially a one-sample t-test for the average of our reference group. In this case, reference is the "low" elevation group
- Conclusion: the average density at low elevations is not equal to 0 (t-stat = 15.186, p < 0.001).
- What about the other coefficients?

anova as a linear model

```
##              term estimate std.error statistic      p.value
## 1     (Intercept)     17.0  1.119524 15.185029 3.377764e-09
## 2 Elevationmedium     -5.6  1.583246 -3.537038 4.093067e-03
## 3   Elevationhigh    -13.0  1.583246 -8.210981 2.877430e-06
```

How do we interpret the p-values?

- Null hypothesis is that $\beta_i = 0$
- Essentially a t-test for differences between reference (low) level and pairwise combinations of other levels (medium, high)
- Conclusion: the average density at both medium and high elevations is significantly different from average tree density at low elevations (t-stat = -3.54 and -8.21, respectively, p < 0.001).
- What about the difference between medium and high elevations?

Testing for significant pairwise differences

> Following a significant *F*-test (ANOVA), the next step is to determine which means differ

> If all group means are to be compared, then we should correct for multiple testing

> Conducting many (~>10) tests increases the probability of having a false positive

Correcting for Multiple Comparisons

- Fisher's Least Significant Difference
  - Wider 95% CI bars
- Pairwise t-test p-value corrections
  - Bonferroni adjustment: multiply p-value by number of tests
- Tukey's Honestly Significantly Different Test
  - This is what we will do in class

tukey's hsd test

tukey's hsd test

According to Tukey's Honestly Significant Difference test, two means ($\bar{y}_i$ and $\bar{y}_j$) are different if:

$$|\bar{y}_i - \bar{y}_j| \geq q_{1-\alpha,k,k(n-1)}\sqrt{\frac{MSW}{n}}$$

where q comes from the "Studentized Range Distribution"(see `qtukey` in `R`). MSW comes from the ANOVA table

example

example

Is there a difference between tree density at different elevations?

Process

1) Fit an `lm()` model

- `pine_lm <- lm(value ~ Elevation, data = pine_long)`

2) Save lm_model as an `aov()` object

- `pine_aov <- aov(pine_lm)`

3) Perform multiple comparison with `TukeyHSD()`

- `TukeyHSD(pine_aov)`

`TukeyHSD()` in R

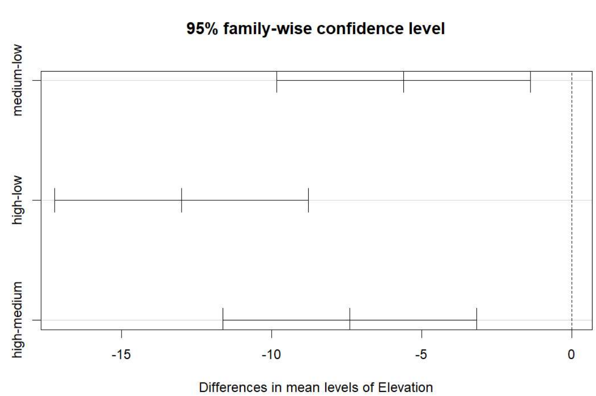
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = pine_lm)
##
## $Elevation
##              diff        lwr       upr     p adj
## medium-low   -5.6  -9.823883 -1.376117 0.0105710
## high-low    -13.0 -17.223883 -8.776117 0.0000080
## high-medium  -7.4 -11.623883 -3.176117 0.0014411
```

- Output has a row for each pairwise comparison
- Estimated difference and 95%CI
- adjusted p-value
  - Adjustment is already accounted for, so compare with standard $\alpha = 0.05$

Plot TukeyHSD intervals

We can also plot the estimates and 95% CI

```
plot(TukeyHSD(pine_aov), xlim = c(-17, 0))
```



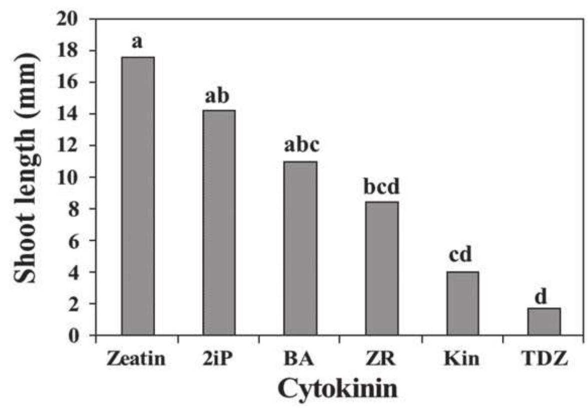
- Since the intervals do not cross 0, we can conclude that all of the differences are significant

Include Tukey results on a plot

You will often see letters on grpahs indicating which groups are different.

Groups with the same letter --> Not Significantly different

Unfortunately the letters are only easy to interpret when the differences are obvious, and can be very confusing if many comparisons are being made.



summary

- One-way ANOVA (*F*-test) can only tell you *IF* at least one group is different
- Depending on question of interest, you may be able to set up your `lm()` analysis to answer your question directly
  - i.e., control versus all other treatment levels
- Multiple comparisons may be required or desired
  - Only do multiple comparison tests after a significant *F*-test
- There are many types of multiple comparison tests
- Tukey's HSD test is probably the method of choice these days. However,
  - It is so conservative that sometimes you won't see any pairwise differences even after a significant *F*-test

Looking Forward

- One-way ANOVA lab and homework assignment
- Reading: Hector Chapters 11 and 12
- Next Week: Factorial analysis and two-way ANOVA