Lecture 23

Analysis of Variance (ANOVA)

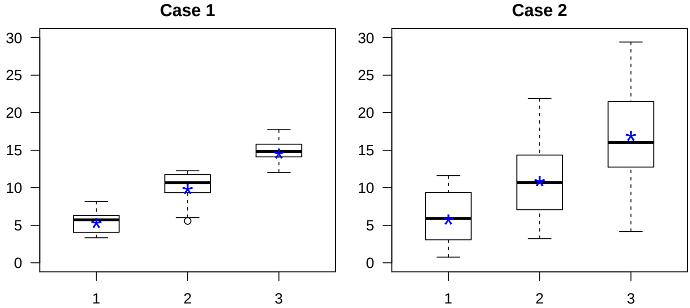
STAT 8010 Statistical Methods I
October 16, 2019

Whitney Huang
Clemson University

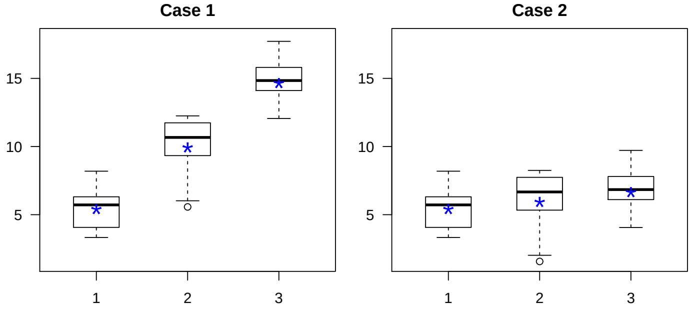
Testing for a Difference in More Than Two Means

- In the last few lectures we have seen how to test a difference in two means, using two sample t-test
- **Question**: what if we want to test if there are differences in a set of **more than two means**?
- The statistical tool for doing this is called analysis of variance (ANOVA)

A Quick Quiz: To Detect Differences in Means

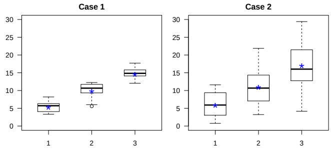


Another Quiz: To Detect Differences in Means

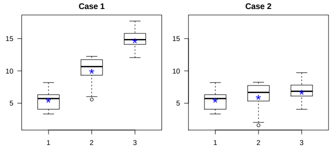


Decomposing Variance to Test for a Difference in Means

- In the first quiz, the data within each group is not very spread out for Case 1, while in Case 2 it is



- In the second quiz, the group means are quite different for Case 1, while they are not in Case 2



- In ANOVA, we compare average between group variance (“signal”) to average within group variance (“noise”) to detect a difference in means

Notation

$$X_{ij} = \mu_j + \varepsilon_{ij},\ \varepsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2),\ i = 1, \cdots, n_j, 1 \le j \le J$$

- J: number of groups
- $\mu_j, j = 1, \cdots, J$: population mean for j_{th} group
- $\bar{X}_j, j = 1, \cdots, J$: sample mean for j_{th} group
- $s_j^2, j = 1, \cdots, J$: sample variance for j_{th} group
- $N = \sum_{j=1}^{J} n_j$: overall sample size
- $\bar{X} = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ij}}{N}$: overall sample mean

Partition of Sums of Squares

"Sums of squares" refers to sums of squared deviations from some mean. ANOVA decomposes the total sum of squares into treatment sum of squares and error sum of squares:

- **Total sum of square**: SSTo $= \sum_{j=1}^{J} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$
- **Treatment sum of square**: SSTr $= \sum_{j=1}^{J} n_j (\bar{X}_j - \bar{X})^2$
- **Error sum of square**: SSE $= \sum_{j=1}^{J} (n_j - 1) s_j^2$

We can show that $\boxed{\text{SSTo} = \text{SSTr} + \text{SSE}}$

Mean squares

A mean square is a sum of squares divided by its associated degrees of freedom

- **Mean square of treatments**: $\text{MSTr} = \frac{\text{SSTr}}{J-1}$
- **Mean square of error**: $\text{MSE} = \frac{\text{SSE}}{N-J}$

Think of MSTr as the “signal”, and MSE as the “noise” when detecting a difference in means $(\mu_1, \cdots, \mu_J)$. A nature test statistic is the signal-to-noise ratio i.e.,

$$F^* = \frac{\text{MSTr}}{\text{MSE}}$$

ANOVA Table and F Test

Source	df	SS	MS	F statistic
Treatment	$J-1$	SSTr	$\text{MSTr} = \frac{\text{SSTr}}{J-1}$	$F = \frac{\text{MSTr}}{\text{MSE}}$
Error	$N-J$	SSE	$\text{MSE} = \frac{\text{SSE}}{N-J}$	
Total	$N-1$	SSTo		

F-Test

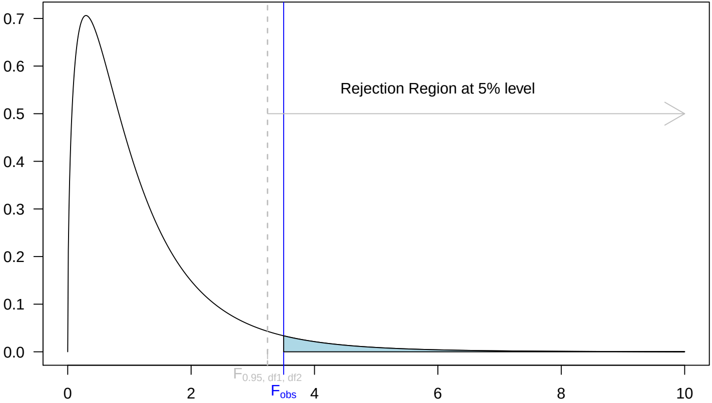
- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_J$
  H_a : At least one mean is different
- Test Statistic: $F^* = \frac{\text{MSTr}}{\text{MSE}}$. Under H_0, $F^* \sim F_{df_1=J-1, df_2=N-J}$
- **Assumptions:**
  - The distribution of each group is normal with equal variance (i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2$)
  - Responses for a given group are independent to each other

F Distribution and the Overall F-Test

Consider the observed F test statistic: $F_{obs} = \frac{\text{MSTr}}{\text{MSE}}$

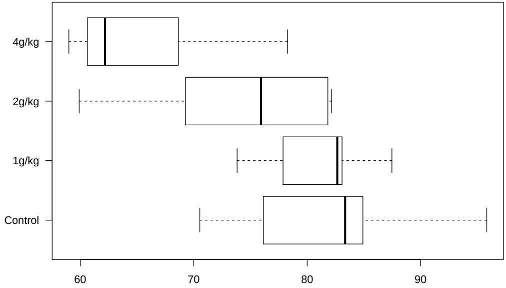
- Should be "near" 1 if the means are equal
- Should be "larger than" 1 if means are not equal

$\Rightarrow$ We use the null distribution of $F^* \sim F_{df_1=J-1, df_2=N-J}$ to quantify if F_{obs} is large enough to reject H_0



Example

A researcher who studies sleep is interested in the effects of ethanol on sleep time. She gets a sample of 20 rats and gives each an injection having a particular concentration of ethanol per body weight. There are 4 treatment groups, with 5 rats per treatment. She records Rapid eye movement (REM) sleep time for each rat over a 24-period. The results are plotted below:



Set Up Hypotheses and Compute Sums of Squares

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs.
  H_a : At least one mean is different
- Sample statistics:

Treatment	Control	1g/kg	2g/kg	4g/kg
Mean	82.2	81.0	73.8	65.7
Std	9.6	5.3	9.4	7.9

- Overall Mean $\bar{X} = \frac{\sum_{j=1}^{4} \sum_{i=1}^{5} X_{ij}}{20} = 75.67$
- SSTo $= \sum_{j=1}^{4} \sum_{i=1}^{5} (X_{ij} - \bar{X})^2 = 1940.69$
- SSTr $= \sum_{j=1}^{4} 5 \times (X_j - \bar{X})^2 = 861.13$
- SSE $= \sum_{j=1}^{4} (5-1) \times s_j^2 = 1079.56$

ANOVA Table and F-Test

Source	df	SS	MS	F statistic
Treatment	$4-1=3$	861.13	$\frac{861.13}{3}=287.04$	$\frac{287.04}{67.47}=4.25$
Error	$20-4=16$	1079.56	$\frac{1079.56}{16}=67.47$	
Total	19	1940.69		

Suppose we use $\alpha = 0.05$

- Rejection Region Method:
  $F_{obs} = 4.25 > F_{0.95,df_1=3,df_2=16} = 3.24$
- P-value Method:
  $\mathbb{P}(F^* > F_{obs}) = \mathbb{P}(F^* > 4.25) = 0.022 < 0.05$

Reject $H_0 \Rightarrow$ We do have enough evidence that not all of population means are equal at 5% level.

R Output

```
Analysis of Variance Table

Response: Response
          Df  Sum Sq Mean Sq
Treatment  3  861.13 287.044
Residuals 16 1079.56  67.472
          F value  Pr(>F)
Treatment  4.2542 0.02173 *
Residuals
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
  0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary

In this lecture, we learned

- Analysis of Variance (ANOVA)

In next lecture we will learn

- Multiple Comparisons