


# Stat 401 B – Lecture 33

 **Categorical Variables**

- Response: Highway MPG
- Explanatory: Type of drive
  - All Wheel
  - Rear Wheel
  - Front Wheel

1

---

---

---


---

---

---

---

---

 **Indicator Variables**

- We have used indicator variables so that we can trick JMP into analyzing the data using multiple regression.

2

---

---

---


---

---

---

---

---

 **Categorical Variables**

- There is a more straight forward analysis that can be done with categorical explanatory variables.

3

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Categorical Variables

- The analysis is an extension of the two independent sample analysis we did at the beginning of the semester.
- Body mass index for men and women (Lectures 4 and 5).

4

---

---

---

---

---

---

---

---

## Analysis of Variance

- Response: numerical, Y
- Explanatory: categorical, X
- Total Sum of Squares

$$SS_{Total} = \sum (y - \bar{y})^2$$

5

---

---

---

---

---

---

---

---

## Sum of Squares Total

$$\bar{y} = 27.7$$

$$SS_{Total} = \sum (y - \bar{y})^2 = 3669.0$$

$$df = 99$$

6

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Analysis of Variance

- Partition the Total Sum of Squares into two parts.
  - Due to differences among the sample means for the categories.
  - Due to variation within categories, i.e error variation.

7

---

---

---

---

---

---

---

---

## Sum of Squares Factor

$$SS_{Factor} = \sum n_i (\bar{y}_i - \bar{y})^2$$

$n_i$  = number of observations in category  $i$   
 $\bar{y}_i$  = sample mean for category  $i$

8

---

---

---

---

---

---

---

---

## Category Sample Means

	Mean	Sample Size
All Wheel	22.608	23
Rear Wheel	26.529	17
Front Wheel	29.983	60

9

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Sum of Squares Drive

$$SS_{Factor} = \sum n_i(\bar{y}_i - \bar{y})^2$$

$$SS_{Drive} = 23(22.608 - 27.7)^2$$

$$+ 17(26.529 - 27.7)^2$$

$$+ 60(29.983 - 27.7)^2$$

$$SS_{Drive} = 932.3$$

10

---

---

---

---

---

---

---

---

## Sum of Squares Error

$$SS_{Error} = \sum (n_i - 1)s_i^2$$

$n_i$  = number of observations in category  $i$

$s_i^2$  = sample variance for category  $i$

11

---

---

---

---

---

---

---

---

## Category Sample Variances

	Variance	Sample Size
All Wheel	15.613	23
Rear Wheel	9.890	17
Front Wheel	37.881	60

12

---

---

---

---


---

---

---

---

# Stat 401 B – Lecture 33



## Sum of Squares Error

$$SS_{Error} = \sum (n_i - 1)s_i^2$$
$$SS_{Error} = 22(15.613)$$
$$+ 16(9.89)$$
$$+ 59(37.881)$$
$$SS_{Error} = 2736.7$$

13

---

---

---


---

---

---

---

---



## Mean Square

- A mean square is the sum of squares divided by its associated degrees of freedom.
- A mean square is an estimate of variability.

14

---

---

---


---

---

---

---

---



## Mean Square Factor

- The mean square factor estimates the variability due to differences among category sample means.
- If the mean square factor is large, this indicates the category sample means are quite different.

15

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Mean Square Error

- The mean square error estimates the naturally occurring variability, i.e. the error variance,  $\sigma^2$ .
- This is the ruler against which the variability among sample means is measured.

16

---

---

---

---

---

---

---

---

## Test of Hypothesis

- $H_0$ : all the category population means are equal.
- $H_A$ : some of the category population means are not equal.
- Similar to the test of model utility.

17

---

---

---

---

---

---

---

---

## Test Statistic

- $F = MS_{\text{Factor}} / MS_{\text{Error}}$
- P-value = Prob  $> F$
- If the P-value is small, reject  $H_0$  and declare that at least two of the categories have different population means.

18

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Analysis of Variance

Source	df	SS	MS	F
Factor (Model)	$k - 1$	$SS_{\text{Factor}}$	$MS_{\text{Factor}}$	$MS_{\text{Factor}} / MS_{\text{Error}}$
Error	$N - k$	$SS_{\text{Error}}$	$MS_{\text{Error}}$	
Total	$N - 1$	$SS_{\text{Total}}$		

19

---

---

---

---

---

---

---

---

## Analysis of Variance

Source	df	SS	MS	F
Factor (Model)	2	932.3	466.15	16.52
Error	97	2736.7	28.213	
Total	99	3669.0		

20

---

---

---

---

---

---

---

---

## Test of Hypothesis

- $F = 16.52$ ,  $P\text{-value} < 0.0001$
- Because the P-value is so small, there are some categories that have different population means.

21

---

---

---

---


---

---

---

---

# Stat 401 B – Lecture 33

 **JMP**

- Response, Y: Highway MPG (numerical)
- Explanatory, X: Drive (categorical)
- Fit Y by X

22

---

---

---

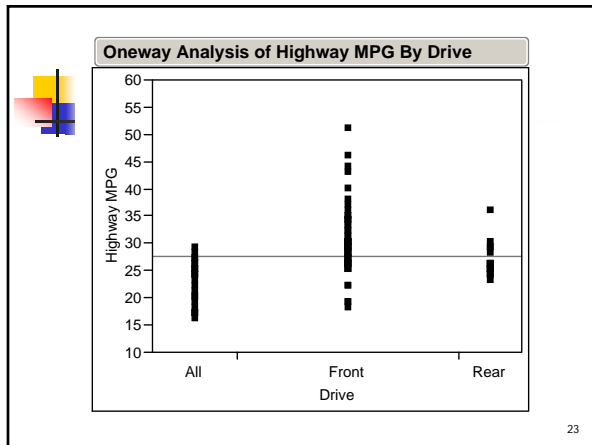
---

---

---

---

---



---

---

---


---

---

---

---

---

 **JMP Fit Y by X**

- From the red triangle pull down menu next to Oneway select Means/Anova
- Display options
  - Uncheck Mean Diamonds
  - Check Mean Lines

24

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Test of Significance

- The test of significance, like the test of model utility, is very general. We know there are some categories with different population means but which categories are they?

25

---

---

---

---

---

---

---

---

## Multiple Comparisons

- In ANOVA, a statistically significant F test is often followed up by a procedure for comparing the sample means of the categories.

26

---

---

---

---

---

---

---

---

## Least Significant Difference

- One multiple comparison method is called Fisher's Least Significant Difference, LSD.
- This is the smallest difference in sample means that would be declared statistically significant.

27

---

---

---

---

---

---

---

---

# Stat 401 B – Lecture 33

## Least Significant Difference

$$LSD = t^* RMSE \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where  $t^*$  is a value for 95% confidence  
and degrees of freedom =  $df_{Error}$   
 $n_i$  and  $n_j$  are the sample sizes for  
categories  $i$  and  $j$

28

---

---

---

---

---

---

---

---

## Least Significant Difference

- When the number of observations in each category is the same, there is one value of LSD for all comparisons.
- When the numbers of observations in each category are different, there is one value of LSD for each comparison.

29

---

---

---

---

---

---

---

---

## Compare All to Rear Wheel

- All Wheel:  $n_i = 23$
- Rear Wheel:  $n_j = 17$
- $t^* = 1.9847$
- $RMSE = 5.3116$

30

---

---

---

---


---

---

---

---

# Stat 401 B – Lecture 33

 **Compare All to Rear Wheel**

$$LSD = 1.9847(5.3116)\sqrt{\left(\frac{1}{23} + \frac{1}{17}\right)}$$
$$LSD = 3.372$$

31

---

---

---


---

---

---

---

---

 **Compare All to Rear Wheel**

- All Wheel: mean = 22.609
- Rear Wheel: mean = 26.529
- Difference in means = 3.92
- 3.92 is bigger than the LSD = 3.37, therefore the difference between All Wheel and Rear Wheel is statistically significant.

32

---

---

---


---

---

---

---

---

 **JMP – Fit Y by X**

- From the red triangle pull down menu next to Oneway select Compare Means – Each Pair Student's t.

33

---

---

---

---

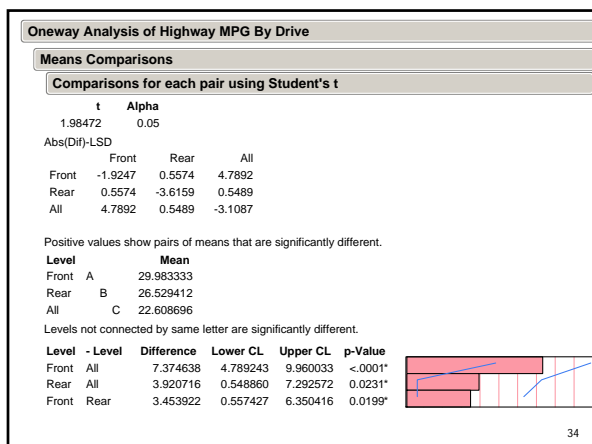
---

---

---

---

# Stat 401 B – Lecture 33



---

---

---

---

---

---

---

---

## Regression vs ANOVA

- Note that the P-values for the comparisons are the same as the P-values for the slope estimates in the regression on indicator variables.

35

---

---

---

---

---

---

---

---

## Regression vs ANOVA

- Multiple regression with indicator variables and ANOVA give you exactly the same analysis.

36

---

---

---

---

---

---

---

---