STAT 215
A One-Predictor ANOVA Model of Means

Colin Reimer Dawson

Oberlin College

April 20, 2018

Notes

Outline

Notes

Outline

Notes

Quantitative Vs. Categorical Predictor and Response

		Response: Quantitative	Response: Categorical
Predictor	Quantitative	Linear Reg.	Logistic Reg.
	Categorical	ANOVA	

Notes

Do Light Patterns Affect Metabolism in Mice?

In this study[1], 27 mice were randomly split into three groups. One group was on a normal light/dark cycle (LD), one group had bright light on all the time (LL), and one group had light during the day and dim light at night (DM). The dim light was equivalent to having a television set on in a room. The mice in darkness ate most of their food during their active (nighttime) period, matching the behavior of mice in the wild. The mice in both dim light and bright light, however, consumed more than half of their food during the well-lit rest period, when most mice are sleeping.

[1] Fonken, L., et. al., "Light at night increases body mass by shifting time of food intake," *Proceedings of the National Academy of Sciences*, October 26, 2010; 107(43): 18664-18669.

Notes

Research Questions

- Do the differences in light conditions lead to differences in
  - Amount of food consumed
  - Activity levels
  - Stress levels
  - Glucose levels after sugar intake

Notes

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	Categorical
Predictor	Quantitative	Linear Reg.	Logistic Reg.
	Categorical	ANOVA	

Notes

ANOVA: Test vs. Model

- We have already seen "Analysis of Variance" (ANOVA) in the context of a test to compare models.
- ANOVA is used to refer both to this test, and to a means model for which the same test is used.
- Note: You have likely seen the test of this model in Stat 1, but it may not have been presented as a model

Notes

The One-Way (One-predictor) ANOVA (Means) Model

DATA = PATTERN + IDIOSYNCRACIES

The One-way ANOVA Population Model (X categorical)

$$Y_i = f(X_i) + \varepsilon_i$$
$$Y_i = \mu_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One μ for each *level* of X

Notes

Example: Light and Food Intake in Mice

$$Y_i = \mu_{X_i} + \varepsilon_i$$
$$\texttt{Consumption}_i = \mu_{\texttt{Light}_i} + \varepsilon_i$$

$$X_i = \texttt{Light}_i = \begin{cases} \texttt{LD} & \text{normal light/dark cycle} \\ \texttt{DM} & \text{dim light at night} \\ \texttt{LL} & \text{bright light at night} \end{cases}$$

$Y_i = \texttt{Consumption}_i$ (daily food intake in grams)

$$\mu_{X_i} = \begin{cases} \mu_{\texttt{LD}} & \text{if } X_i = \texttt{LD} \\ \mu_{\texttt{DM}} & \text{if } X_i = \texttt{DM} \\ \mu_{\texttt{LL}} & \text{if } X_i = \texttt{LL} \end{cases}$$

Notes

Parameter Estimates

The One-way ANOVA Fitted Model (X categorical)

$$Y_i = \hat{\mu}_{X_i} + \hat{\varepsilon}_i \qquad \hat{\varepsilon}_i \sim \mathcal{N}(0, \hat{\sigma}^2_\varepsilon)$$
$$Y_i = \bar{Y}_{X_i} + (Y_i - \bar{Y}_{X_i})$$

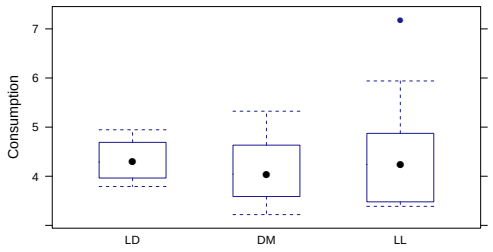
One $\bar{Y}_{group}$ for each level of X

$$\texttt{Consumption}_i = \hat{\mu}_{\texttt{Light}_i} + \hat{\varepsilon}_i \qquad \hat{\varepsilon}_i \sim \mathcal{N}(0, \hat{\sigma}^2_\varepsilon)$$
$$\texttt{Consumption}_i = \overline{\texttt{Consumption}}_{\texttt{Light}_i} + (\texttt{Consumption}_i - \overline{\texttt{Consumption}}_{\texttt{Light}_i})$$

Notes

Four Plots (1)

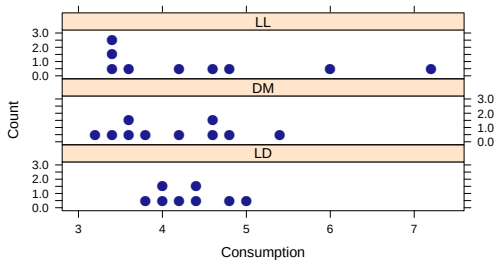
```
library("mosaic")
LaN <- read.file("http://lock5stat.com/datasets/LightatNight4Weeks.csv")
LaN <-   # This line is just to reorder the categories for the plot
    mutate(LaN,
           Light = factor(Light, levels = c("LD", "DM", "LL")))
bwplot(Consumption ~ Light, data = LaN)
```



Notes

Four Plots (2)

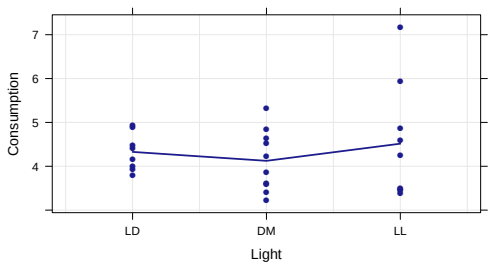
```
dotPlot(~Consumption | Light, data = LaN,
        width = 0.2, layout = c(1,3)) #controls bin width and arrangement
```



Notes

Four Plots (3)

```
xyplot(Consumption ~ Light, data = LaN,
       type = c("p","a"), grid = TRUE)
```

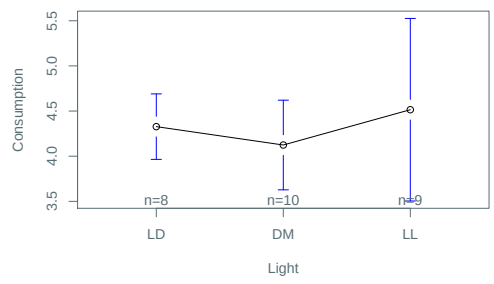



```
## type = c("p", "a") plots the points and a line through the averages
## grid = TRUE draws a grid in the background
```


Notes

Four Plots (4)

```
library("gplots")  # May need to install it first
plotmeans(Consumption ~ Light, data = LaN)
```




```
## Plots means and individual confidence intervals for the means
```


Notes

Descriptive Stats

```
favstats(Consumption ~ Light, data = LaN)
```

```
  Light   min      Q1 median      Q3   max     mean        sd  n missing
1    LD 3.791 3.98375 4.2885 4.58925 4.946 4.327500 0.4337033  8       0
2    DM 3.219 3.59350 4.0440 4.60325 5.324 4.124100 0.6937946 10       0
3    LL 3.387 3.47900 4.2400 4.87300 7.177 4.514889 1.3148988  9       0
```


Notes

The One-Way (One-predictor) ANOVA (Means) Model

The One-way ANOVA Population Model (X categorical)

$$Y_i = f(X_i) + \varepsilon_i$$
$$Y_i = \mu_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One μ for each *level* of X

Question: Is there evidence that the μs are different?

Notes

Outline

Notes

The One-Way ANOVA Model: Effect Coded

The One-way ANOVA Population Model (X categorical)

$$Y_i = f(X_i) + \varepsilon_i$$
$$Y_i = \mu_{overall} + \alpha_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

One α_{group} for each level of X: group deviation from overall mean

Question: Is there evidence that the αs are not all zero?

Notes

Mice: Effect Coded

$$\texttt{Consumption}_i = \mu_{Overall} + \alpha_{\texttt{Light}_i} + \hat{\varepsilon}_i \qquad \hat{\varepsilon}_i \sim \mathcal{N}(0, \hat{\sigma}^2_\varepsilon)$$

where α_{group} represents the difference between the group-specific mean consumption and the overall mean consumption

$$\texttt{Consumption}_i = \begin{cases} \mu_{Overall} + \alpha_{\texttt{LD}} + \varepsilon_i & \text{if } \texttt{Light}_i = \texttt{LD} \\ \mu_{Overall} + \alpha_{\texttt{DM}} + \varepsilon_i & \text{if } \texttt{Light}_i = \texttt{DM} \\ \mu_{Overall} + \alpha_{\texttt{LL}} + \varepsilon_i & \text{if } \texttt{Light}_i = \texttt{LL} \end{cases}$$

Question: Is there evidence that mean consumption is modified (i.e., at least one $\alpha \neq 0$) based on light conditions?

Notes

Parameter Estimates

The One-way ANOVA Fitted: (Alternative Formulation)

$$Y_i = \hat{f}(X_i) + \hat{\varepsilon}_i$$
$$Y_i = \hat{\mu}_{overall} + \hat{\alpha}_{X_i} + \hat{\varepsilon}_i$$
$$Y_i = \bar{Y}_{Overall} + (\bar{Y}_{X_i} - \bar{Y}_{Overall}) + (Y_i - \bar{Y}_{X_i})$$

One $\bar{Y}_{X_i} - \bar{Y}_{Overall}$ for each level of X: Group deviation from overall mean

Notes

Parameter Estimates: Mice

```
mu.hat.overall <- mean(~Consumption, data = LaN) # overall sample mean
mu.hat.overall %>% round(2)

     [1] 4.31

mu.hat.xis <- mean(Consumption ~ Light, data = LaN) # group means
mu.hat.xis %>% round(2)

       LD   DM   LL
     4.33 4.12 4.51

alpha.hats <- mu.hat.xis - mu.hat.overall
## group deviations from overall mean
alpha.hats %>% round(2)

       LD    DM    LL
     0.01 -0.19  0.20
```


Notes

Outline

Notes

Partitioning Variability

The One-way ANOVA Fitted (Effect Coded)

$$Y_i = \bar{Y}_{overall} + (\bar{Y}_{X_i} - \bar{Y}_{overall}) + (Y_i - \bar{Y}_{X_i})$$

One $\bar{Y}_X - \bar{Y}_{overall}$ for each level of X: Group deviation from overall mean

Partitioning Variability

$$(Y_i - \bar{Y}) = (\bar{Y}_k - \bar{Y}) + (Y_i - \bar{Y}_k)$$

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\bar{Y}_k - \bar{Y})^2 + 0 + \sum_i (Y_i - \bar{Y}_k)^2$$

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Notes

Parameter Estimates: Mice

```
mean(~Consumption, data = LaN) %>% round(2)

     [1] 4.31
```

```
mean(Consumption ~ Light, data = LaN) %>% round(2)

      LD   DM   LL
    4.33 4.12 4.51
```

Overall mean, $\bar{Y} \approx 4.31$
Individual means $\bar{Y}_X$s $= 4.33, 4.12, 4.51$

$$Y_i = \begin{cases} 4.31 + (4.33 - 4.31) + \hat{\varepsilon} & X_i = \texttt{LD} \\ 4.31 + (4.12 - 4.31) + \hat{\varepsilon} & X_i = \texttt{DM} \\ 4.31 + (4.51 - 4.31) + \hat{\varepsilon} & X_i = \texttt{LL} \end{cases}$$

Notes

Exercise: Find the SS components

Overall mean, $\bar{Y} \approx 4.31$
Individual means $\bar{Y}_{LD} = 4.33$, $Y_{DM} = 4.12$, $Y_{LL} = 4.51$

Light	Consumption	$(\bar{Y}_{Light_i} - \bar{Y}_{Overall})^2$	$(Y_i - \bar{Y}_{Light_i})^2$
LD	3.79		
LD	3.92		
DM	3.41		
DM	3.22		
LL	3.39		
LL	3.45		
		$SS_{Model} =$	$SS_{Error} =$

Notes

How Much Variability is Explained?

- Still have R^2

$$R^2 = \frac{SS_{Model}}{SS_{Total}}$$

Notes

Outline

Notes

Model Comparison (ASSESS)

Null Population Model:

$$Y_i = \mu_{Overall} + \varepsilon$$

Groups Population Model:

$$Y_i = \mu_{Overall} + \alpha_{X_i} + \varepsilon$$

H_0 : $\alpha_X \equiv 0$ for all $X \Rightarrow$ high SS_{Model} due to chance
H_1 : some $\alpha_X \neq 0 \Rightarrow$ high SS_{Model} due to grouping

Notes

Conditions for Test of ANOVA Model

Can view these as regression models and compute P-value assuming the same conditions as for SLR (except linearity):

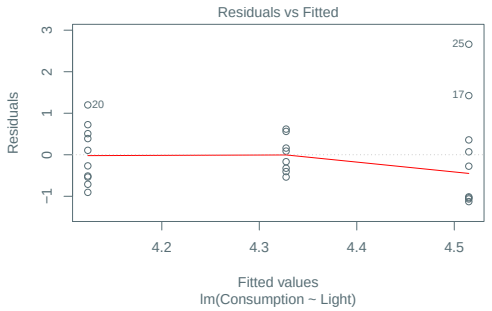
Conditions for Test of ANOVA Model

1. Zero mean: Residuals centered at 0
2. Constant variance: Same variability at all X (Homoskedasticity)
3. Independence: No relationship among errors
4. Normality (for standard form): At each X, Ys are Normally distributed

Notes

Checking Conditions: Residual Plots

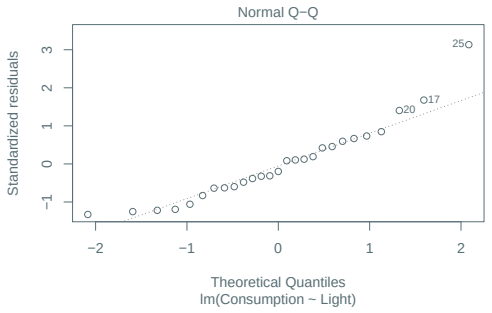
```
consumption.model <- lm(Consumption ~ Light, data = LaN)
plot(consumption.model, which = 1)
```



Notes

Checking Conditions: Residual Plots

```
plot(consumption.model, which = 2)
```



Notes

Checking Conditions: Homoskedasticity

Homoskedasticity: $=$ Equal variances

```
sd(Consumption ~ Light, data = LaN)

        LD        DM        LL
0.4337033 0.6937946 1.3148988
```

Rough rule: largest $s \leq 2 \cdot$smallest s

Notes

Conditions: A Caveat

If sample sizes are (nearly) equal within groups, the ANOVA test is fairly *robust* to violations of normality and homoskedasticity.

Notes

The ANOVA Table

```
consumption.model <- aov(Consumption ~ Light, data = LaN)
summary(consumption.model)

             Df Sum Sq Mean Sq F value Pr(>F)
Light         2  0.725  0.3626   0.447  0.645
Residuals    24 19.481  0.8117
```

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{SS_{Model}/df_{Model}}{SS_{Error}/df_{Error}}$$

has an F distribution if H_0 is true and conditions are met.

Notes

Conclusion

- Little evidence that mean consumption would differ at 4 weeks for mice under these conditions in general

Notes

2 hr Post-Meal Glucose Levels

```
mean(GTT120 ~ Light, data = LaN) %>% round(1)

       LD    DM    LL
    173.5 258.7 321.4

sd(GTT120 ~ Light, data = LaN) %>% round(1)

       LD    DM    LL
     41.9 113.0 109.0

model2 <- aov(GTT120 ~ Light, data = LaN)
summary(model2)

              Df Sum Sq Mean Sq F value Pr(>F)
    Light      2  92958   46479   5.018 0.0151 *
    Residuals 24 222299    9262
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Notes

Conclusion

- Significant evidence that 2 hour post-meal glucose 4 weeks after the intervention would differ across the three conditions
- Can we attribute the differences to the intervention? (I.e., can we make a causal conclusion?)

Notes

Notes