STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 16-18, 2018

STA 303/1002: Week 2 Outline

- The General Linear Model
- One-way ANOVA
  - With G=2
  - With $G > 2$
- Case Study 1 continued
- Diagnostics- checking model assumptions
  - Normality of errors
  - Constant variance
  - Uncorrelated errors
- Multiple comparisons: Bonferroni and Tukey's

Week 1 Review

- Review:
  - One sample t-test $H_0: \mu = \mu_0$
  - Two sample t-tests (t.test() or summary(lm()) or anova() in R) $H_0: \mu_1 = \mu_2$
  - Testing equal variances
  - Assessing normality
  - Case Study 1: Question 1

The General Linear Model

X

cat cts

Y-cts
X-cts

- Response, Y is continuous
- Explanatory variable(s), X is(are) categorical and/or continuous
- Y is linear in the β's-i.e. no predictor is a linear function or combination of other predictors
- In R: `lm()`

Review of Regression in Matrix Terms

Model:

$$Y_i = \boxed{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}} + \epsilon_i$$

for $i = 1, \ldots, N$

Matrix Form: $Y = X\beta + \epsilon$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Np} \end{pmatrix}_{N\times(p+1)}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}_{N\times 1}$$

Least-squares Estimates for β

- $\hat{\beta} = (X'X)^{-1}X'Y$
- $X'X$ has dimension $(p+1) \times (p+1)$
- Need $X'X$ to be of full rank to be invertible:
  - $rank(X'X) = rank(X)$
  - Need X to be rank $p+1$
  - The columns of X must be linearly independent

Gen LM: Hypothesis and Assumptions

- Null Hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

- Assumptions:
  - Linear Model is appropriate: Errors have zero expectation, $E[\epsilon_i] = 0$
  - Homoscedasticity of variances: Errors have constant variance, $Var(\epsilon_i) = \sigma^2$
  - Errors are uncorrelated
  - Errors are jointly normally distributed

Indep. observations

$G=2$

(GLM) $H_0: \beta_1 = 0$

$\Updownarrow$

(One-way ANOVA) $H_0: \mu_A = \mu_{A^c}$

GLM: Sum of Squares Decomposition

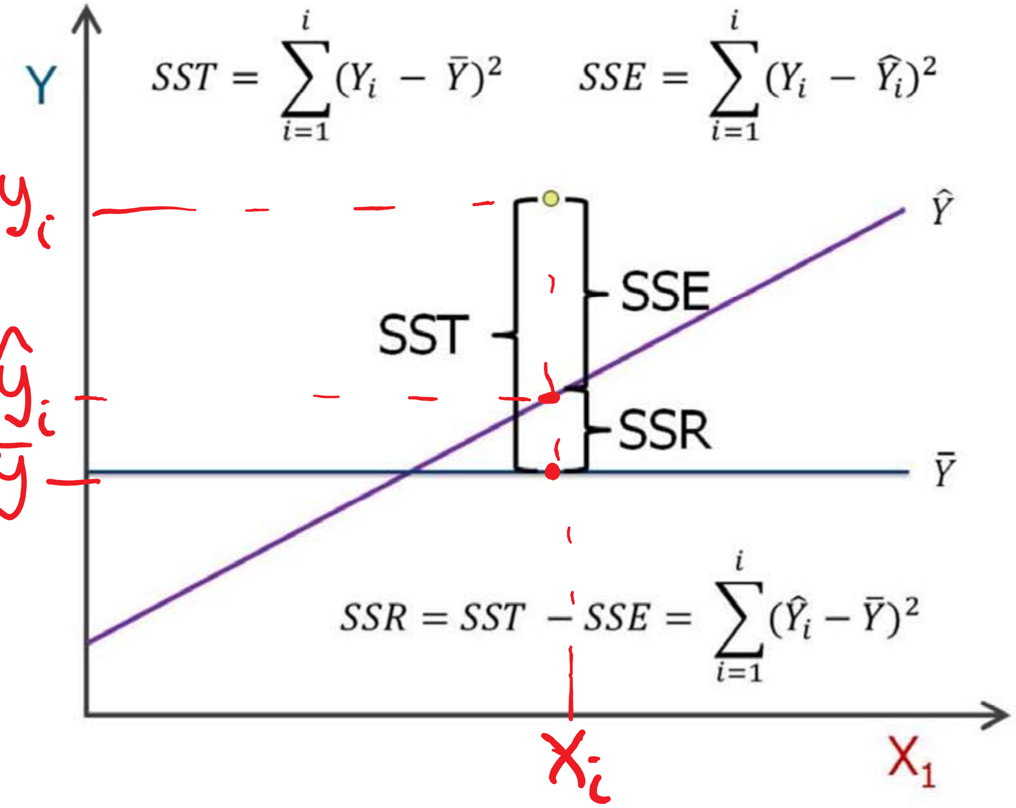
$\boxed{\text{Aim, } E(y_i) = \mu_y}$

$\widehat{E[y_i|X_i]} = \hat{y}_i$

observed, y_i

estimate, $\hat{y}_i$

sample mean, $\bar{y}$



$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$

$\hat{y}_i = X_i\hat{\beta}$
(model)

$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$

SS – sum of squares

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}), \quad \boxed{\hat{Y}_i = X_i\hat{\beta}}$$

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{SSTotal} = \underbrace{\sum_i (Y_i - \hat{Y}_i)^2}_{SSE=RSS} + \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{SSReg}$$

One-way ANOVA

One-Way ANOVA

- Response/Outcome is continuous → y
- One factor (categorical/grouping variable) with at least 2 levels ($G \geq 2$)
- Aim: Compare G group means

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_G \cancel{=0}, \quad H_a : \exists i \neq j \ \ s.t. \ \ \mu_i \neq \mu_j$$

- Predictors are indicator variables that classify the observations one way (into G groups)
  - Special case of a general linear model (GLM)
  - Equivalent to GLM with one-way classification (one factor)
  - GLM uses $\underline{G-1}$ dummy variables.
- ANOVA: compare means by analyzing variability

X 1 0

Eg

$$\begin{array}{cccc} & & x_{i1} & x_{i2} \\ X = & \left(\begin{matrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & 0 & 1 \\ \vdots & 0 & 1 \\ 1 & 0 & 1 \end{matrix}\right) & & \end{array}$$

$\Rightarrow x_{i1} + x_{i2} = \underline{\underline{1}}$

lin. dep. colms.

(Q1) $\underline{G=2}$, $y_i = \beta_0 + \beta_1 X_{iA}$ where $X_{iA} = \begin{cases} 1 & \text{if ith obs in group A} \\ 0 & \text{o-w} \end{cases}$

Brief History of ANOVA

- Dates back to early work by R. A. Fisher in 1918 on mathematical genetics
- Further developed by Fisher in 1920
- The convenient acronym - ANOVA was coined much later by John W. Tukey (1915-2000), the pioneer of exploratory data analysis (EDA)
- The test developed was named the F in his honour

Data layout and Notation

	Treatment or factor levels				
	1	2	$\cdots$	G	
	Y_{11}	Y_{21}	$\cdots$	Y_{G1}	
	Y_{12}	Y_{22}	$\cdots$	Y_{G2}	
	$\vdots$	$\vdots$		$\vdots$	
	Y_{1n_1}	Y_{2n_2}	$\cdots$	Y_{Gn_G}	
Group means — Sample mean	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	$\cdots$	$\bar{Y}_{G.}$	$\boxed{\bar{Y}_{..}}$
Group — Sample variance	S_1^2	S_2^2	$\cdots$	S_G^2	

n_g = group sizes

$$\frac{\sum_{i=1}^{N} y_i}{N}$$

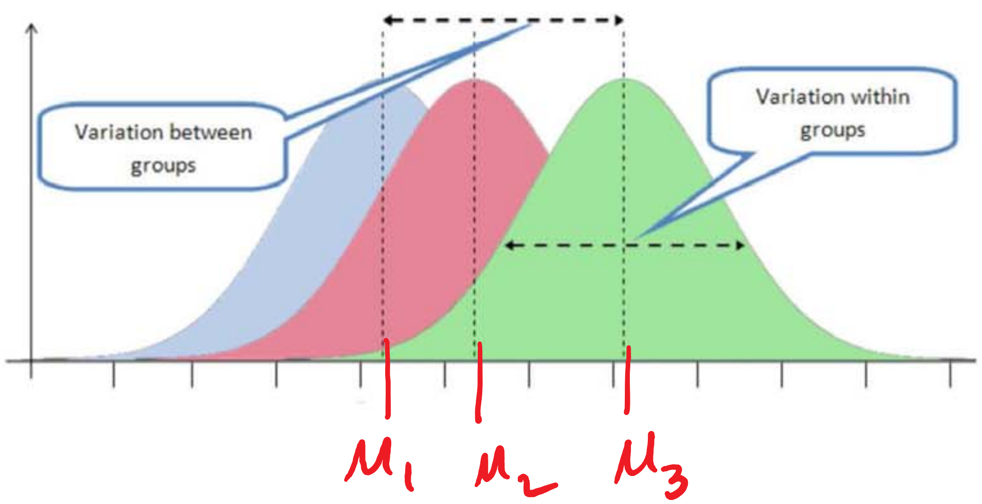
$$\bar{y}_{g.} = \frac{\sum_{j=1}^{n_g} y_{gj}}{n_g} \qquad S_g^2 = \frac{1}{n_g - 1} \sum_{j=1}^{n_g} (y_{gj} - \bar{y}_{g.})^2$$

One-way ANOVA Assumptions

- The G samples are independently drawn from G specific populations with unknown means $\mu_1, \mu_2, \ldots, \mu_G$.
- Each population is normally distributed.
- Each population has the same variance, σ^2.

Compactly written:

$$E_i \sim Normal(\mathbf{0}_G, \sigma^2\mathbf{I})$$



One-way: Expectations and Estimates

- Expected values of Y, $\cancel{\mu_i} = E(Y_i)$ parameter
- Predicted values of Y, $\hat{Y}_i$
- Estimates of coefficients, $\hat{\beta}$

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1 \\ \vdots \\ \beta_0 + \beta_{G-1} \\ \beta_0 \end{cases}, \hat{Y}_i = \begin{cases} b_0 + b_1 \\ \vdots \\ b_0 + b_{G-1} \\ b_0 \end{cases}, \hat{\beta} = \begin{cases} b_0 = \bar{y}_G \\ b_1 = \bar{y}_1 - \bar{y}_G \\ \vdots \\ b_{G-1} = \bar{y}_{G-1} - \bar{y}_G \end{cases}$$

$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{G-1} X_{iG-1} + \epsilon_i$

$E(y_i / X_{i1} = 1) = \beta_0 + \beta_1$

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$H_0: \beta_1 = 0$ $\Leftrightarrow$ $\mu_A = \mu_{A^c}$

$b_1 = \bar{y}_A - \bar{y}_{A^c}$

by l.s. estimation ($G = 2$, $\hat{\beta}_0$, $\hat{\beta}_1$

$\sum_{i=1}^{N} x_{i1} = n_1$

$\sum_{i=1}^{N} x_{i1}^2 = n_1$

Decomposition of SST

Model

Error

$$SST = \sum_{i}^{N}(Y_i - \bar{Y})^2 = SS_{Reg} + RSS$$

$$= \sum_{i}^{N}(\hat{Y}_i - \bar{Y})^2 + \sum_{i}^{N}(Y_i - \hat{Y}_i)^2$$

- $N = n_1 + \ldots + n_G$
- $\hat{Y}_i$ = mean of observations for group g from which the i*th* observation belongs
- $\hat{Y}_i$ is one of $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_G$
- $\bar{Y} = \bar{Y}_{..}$ is the grand mean

One-way ANOVA Table

$SS/_{DF}$ $\simeq$ Variances.

SOURCE	DF	SS	MS	F
Model	G-1	SSReg	MSReg=SSReg/G-1	MSReg/MSE
Error	N-G	RSS	MSE= RSS/N-G	
TOTAL	N-1	SST		

- SSReg: "between groups" SS
- RSS: "within groups" SS
- Overall idea: If between groups SS is larger than within groups SS, there is evidence that means are different

Which group means differ? Which is bigger- SSReg or RSS?

within

SSReg < RSS

SSReg > RSS

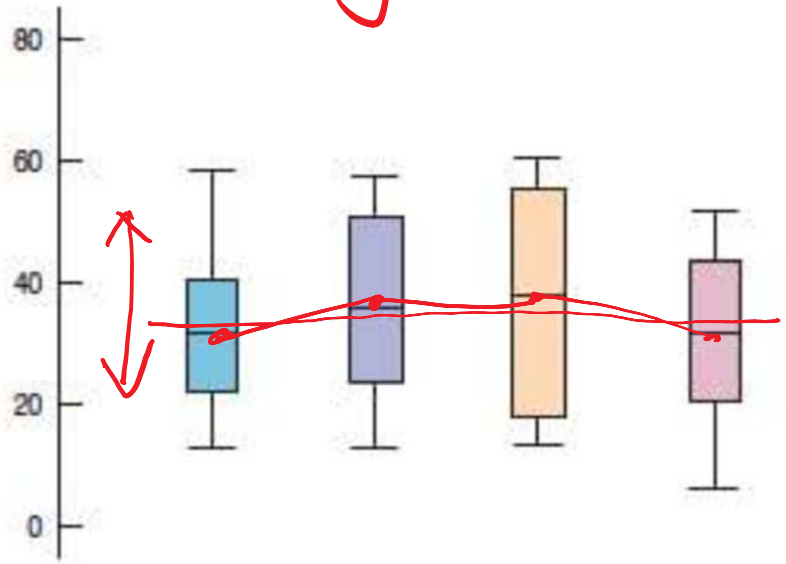


Figure 25.2
It's hard to see the difference in the means in these boxplots because the spreads are large relative to the differences in the means.

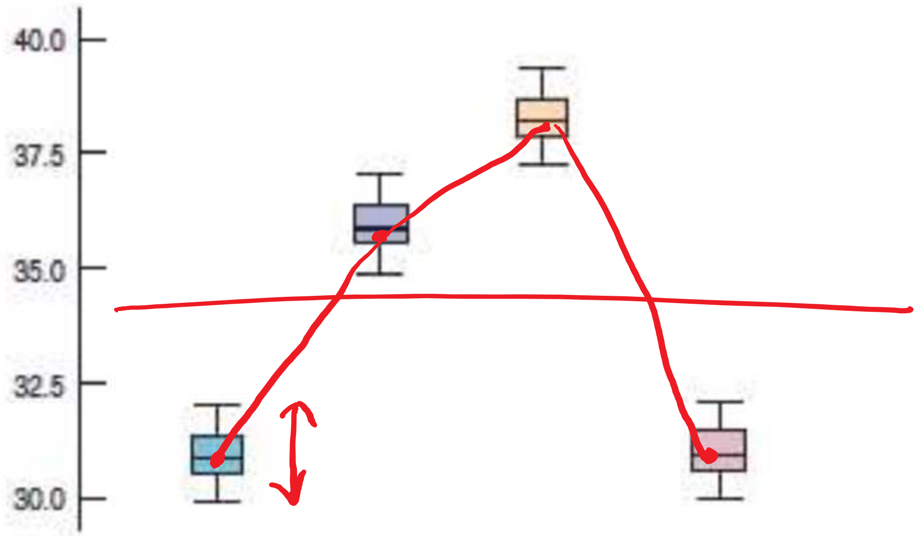


Figure 25.3
In contrast with Figure 25.2, the smaller variation makes it much easier to see the differences among the group means.

(*SDM, 2nd Canadian ed. by De Veaux et. al.*)

3rd.

Derivation of SS's: SSReg and RSS

$$SSreg = \sum_i^N (\hat{Y}_i - \bar{Y})^2 = \sum_g^G n_g (\bar{Y}_g - \bar{Y})^2$$

$$RSS = \sum_i^N (Y_i - \hat{Y}_i)^2 = \sum_{g=1}^G \sum_{(g)} (Y_i - \bar{Y}_g)^2$$

- g is the group index
- $\hat{Y}_i$ is one of $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_G$
- $\sum_{(g)}$ -summation over observations in group g

Case Study 1 continued:
The Spock Conspiracy Trial

Case Study 1: The Spock Conspiracy Trial

Recall the 2 main questions:

(Q1) Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges?

(Q2) Is there a difference among the 6 other judges?

(A1): Two-sample t-test/ Simple linear regression model with 1 dummy predictor variable/ One-way ANOVA with G=2
(A2): Multiple linear regression model with 5 dummy predictor variables/ One-way ANOVA with G=6

Overall task: Compare the percent of women on venires of all 7 judges

Case Study 1: The Spock Conspiracy Trial Data

Get the data (from desktop):

```
#Juries data
juries<-read.csv(
  "/Users/Shivon/STA303_1002/LectureNotes/Lec1/juries.csv", header=T)
attach(juries)
head(juries)
```

```
##   PERCENT  JUDGE
## 1     6.4 SPOCKS
## 2     8.7 SPOCKS
## 3    13.3 SPOCKS
## 4    13.6 SPOCKS
## 5    15.0 SPOCKS
## 6    15.2 SPOCKS
```

Case Study 1: The Spock Conspiracy Trial Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case0502
library(Sleuth3)
#Juries data
jury = case0502
attach(jury)
head(jury)
```

Ramsey & Schafer
The Statistical Sleuth
3rd ed.

```
##   Percent   Judge
## 1     6.4 Spock's
## 2     8.7 Spock's
## 3    13.3 Spock's
## 4    13.6 Spock's
## 5    15.0 Spock's
## 6    15.2 Spock's
```

Case Study 1: How many venires for each Judge?

```
table(Judge)
```

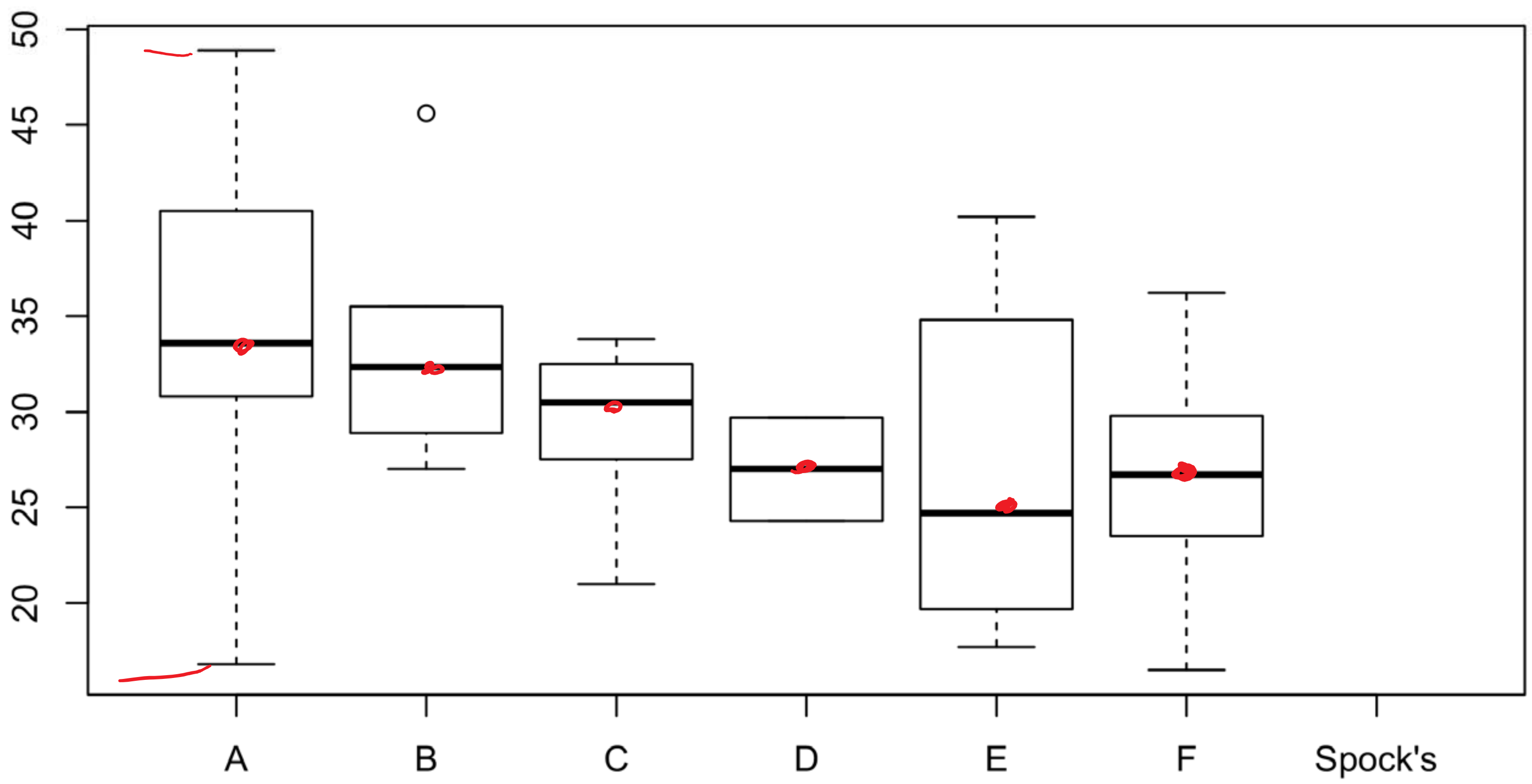
```
## Judge
##        A        B        C        D        E        F Spock's
##        5        6        9        2        6        9       9
```

```
with(jury, tapply(Percent,Judge,mean))
```

```
##        A        B        C        D        E        F  Spock's
## 34.12000 33.61667 29.10000 27.00000 26.96667 26.80000 14.62222
```

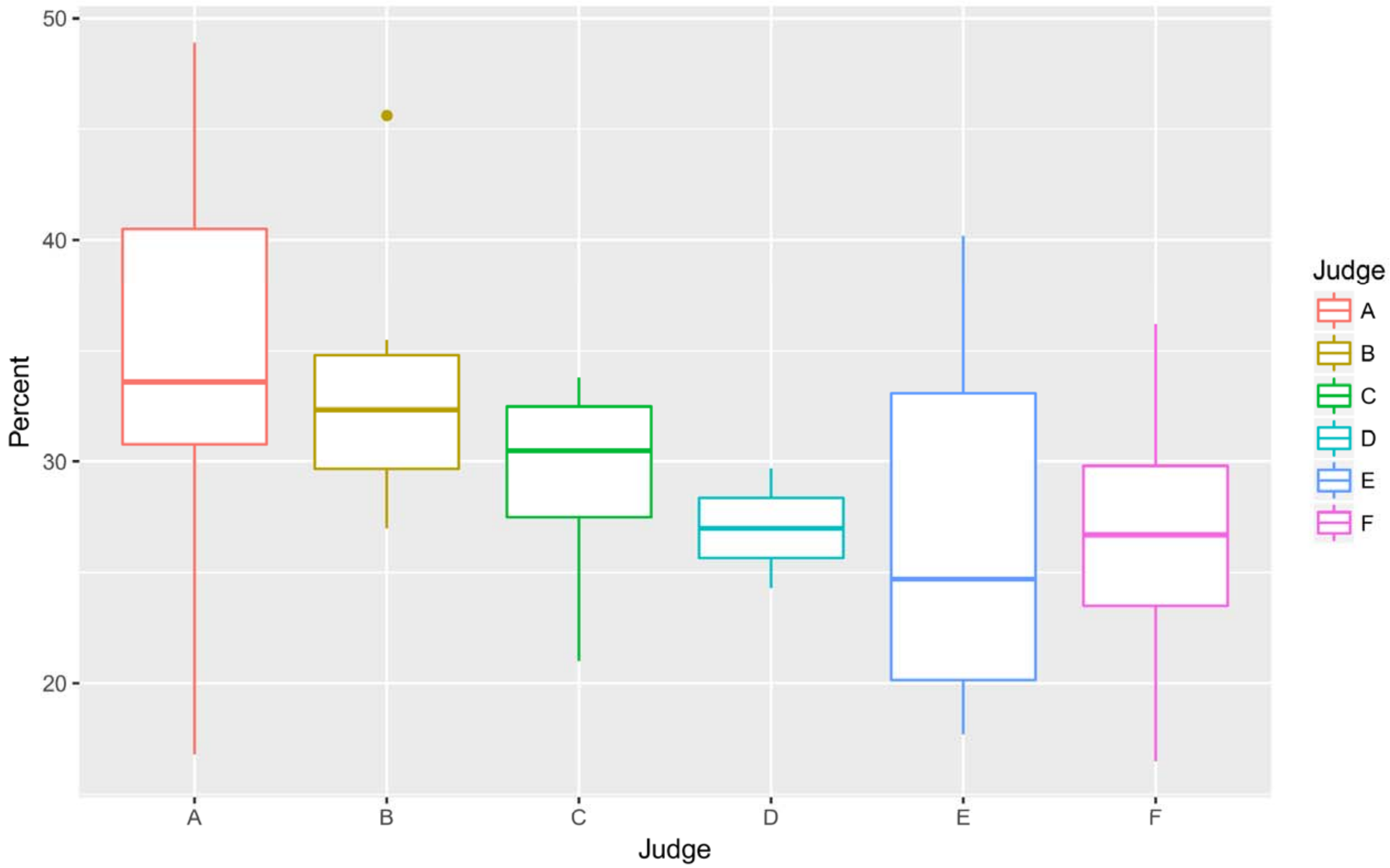
Case Study 1: Boxplot of Judges

```
# Get data subset of other judges
Others <- subset(jury, Judge != "Spock's")
boxplot(Percent~Judge, data=Others)
```



Case Study 1: Boxplot of Judges

```
#install.packages("ggplot2")
library(ggplot2)
ggplot(Others, aes(x=Judge,y=Percent, color=Judge))+geom_boxplot()
```



Case Study 1: Q2-Compare the 6 other judges

G = 6.

```
summary(aov(Percent~Judge,data=Others))
```

MS Reg.

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Judge        5  326.5   65.29   1.218  0.324
## Residuals   31 1661.3   53.59
```

MSE

$H_0: \mu_A = \mu_B = \cdots = \mu_F$

P-value = 0.324

- P-value is not small.
- We do not have evidence against the null hypoth.
- Data supports the idea that the venires of the other judges do not differ.

Case Study 1 Partial Summary

(Q1) Data provides evidence that Spock's judge's venires underrepresent women.

- Homoscedasticity satisfied
- Normal errors hold

(Q2) Data supports the hypothesis that the venires of the other six judges do ~~not~~ have similar percentages of women.

- Where does the difference lie?
- Are the model assumptions satisfied?