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

The General Linear Model with Dummy Variables

One-way ANOVA

Simple Linear Model with 1 dummy variable:

$(G=2)$

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

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

$$Y_i = \beta_0 + \beta_1 X_{i,A} + \epsilon_i$$

$X_{i,A} = \begin{cases} 1 & \text{if } i\text{th obs. is from group } A \\ 0 & \text{o.w.} \end{cases}$

Multiple Linear Model with G-1 dummy variables:

$(G>2)$

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_{\cancel{p}\,G-1} X_{i,G-1} + \epsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{G-1} = 0$

$\beta_1 = \mu_1 - \mu_G \quad \beta_2 = \mu_2 - \mu_G \quad \cdots \quad \beta_{G-1} = \mu_{G-1} - \mu_G = 0$

One-way ANOVA

$\Leftrightarrow \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_G$

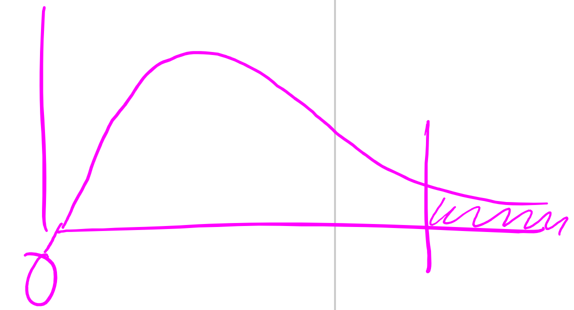
One-way ANOVA Table

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**		

- Test statistic: $F = \frac{MSReg}{MSE}$

$MSE = \hat{\sigma}^2 = s^2$

- Distribution of test statistic: $\sim F_{G-1, N-G}$



- P-value: $P(F_{G-1,N-G} > F)$

General Linear Model vs One-way ANOVA

- General Linear Model:
  - Response/Outcome, Y is continuous
  - $X's$ are categorical and/or continuous
  - Assumptions stated in terms of the errors, ie., $E_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
  - Assumptions are equivalent to One-way ANOVA
  - In R: `lm()`

- One-way ANOVA
  - Response/Outcome, Y is continuous
  - One factor/categorical variable ($G \geq 2$)
  - Assumptions are equivalent to General LM
  - In R: `aov()`

Multiple Comparisons

$H_0: \mu_1 = \mu_2 = \mu_3$

- Post hoc procedure: further comparisons after significant result from overall One-way ANOVA
- 'Post hoc' means 'after this' in Latin
- Max of ${}^{G}C_2$ pairwise comparisons
- Major issue: There is an increased chance of making at least one Type I error when carrying out many tests.
- Two common solutions: based on controlling family Type I error rate
  - Bonferroni — controls α at pairwise level
  - Tukey's — controls α at family level

$G = 3$

$k = {}^{3}C_2 = \binom{3}{2}$

$= \frac{3!}{2!\,1!} = 3$

Pairs

$$\begin{bmatrix} 1,2 \\ 2,3 \\ 1,3 \end{bmatrix}$$

Multiple Comparisons

$= P(\text{Type I Error})$.

Q: If $\alpha = 0.10$, what is the chance of committing 'at least 1' Type 1 Error...

- in 2 independent tests?
- in 10 independent tests?
- in k independent tests?

T1 T2

$1-(1-\alpha)(1-\alpha)$

$1-(1-\alpha)^k$

$G=3,\ k=3$

$G=5,\ k=10$

$G=10,\ k=45$

As $k \uparrow$, $(1-\alpha)^k \downarrow$, $1-(1-\alpha)^k \uparrow$

Multiple Comparisons: Bonferroni's Method

$(k=3)(\alpha=0.10) = 0.3.$

$k\alpha$

- Based on the Bonferroni's inequality:
  $P(AUB) \leq P(A) + P(B)$
- Let A_i be the event that the ith test results in a Type I error. Then $P(UA_i) \leq \sum P(A_i)$
- Denote $k =^G C_2 = \binom{G}{2}$, total number of pairwise comparisons of G means.

$k=45, \alpha=0.01$

$k\alpha = 0.45$

- Method: Conduct each of k pairwise tests at level α/k.

$\boxed{\alpha^* = \alpha/k}$

- Then the overall family Type I error rate of the k tests is at most α, i.e., the chance that at least 1 test results in a Type I error error is at most α.

$\leq \alpha.$

$\alpha/k + \alpha/k + \cdots + \alpha/k = \alpha.$

Multiple Comparisons: Bonferroni's Method

- For CIs: If each CI has confidence level $(1 - \alpha/k)100\%$, then CI coverage rate is at most α
- Bonferroni CI: $|\bar{y}_i - \bar{y}_j| \pm \boxed{t^{*}_{\alpha/2k}}\, se(\bar{y}_i - \bar{y}_j)$
- Conservative: overall Type I error rate (chance of making at least one Type I error) is usually much less than α if tests are not mutually independent. $\leq \alpha$
- Type II error inflation. Not as powerful.

$\begin{bmatrix} 1,2 \\ 2,3 \\ 1,3 \end{bmatrix}$

1,2
3,4
5,6

Multiple Comparisons: Tukey's Approach

- Based on Tukey's Honestly Significant Difference (HSD)

- Requires Tukey's "Studentized Range Distribution" of $\max_{a,b\in\{1,\ldots,G\}}\{\bar{y}_a - \bar{y}_b\}$
- Usually less conservative than Bonferroni's method, particularly if group sample sizes are similar.
- Adv Precisely controls the overall Type I error rate at α; simultaneous CI coverage rate is $(1-\alpha)100\%$

Tukey's Approach: The Studentized Range distribution

- Consider n realizations- $\{x_1, \ldots, x_n\}$ of a Normally distributed random variable, $X \sim N(\mu, \sigma^2)$. Determine the distribution of the largest and smallest value of $\{x_1, \ldots, x_n\}$.
- Denote $\max\{X_1, \ldots, X_n\} = X_{(n)}$ and $\min\{X_1, \ldots, X_n\} = X_{(1)}$. $Range = X_{(n)} - X_{(1)}$.
- Based on n observations from X, the Studentized Range statistic is:
  $Q_{stat} = \frac{x_{(n)} - x_{(1)}}{s}$, s is the sample standard deviation
- Based on G group means, with n observations per group:

$$\bar{Q}_g = \frac{\sqrt{n}\left(\bar{y}_{(g)} - \bar{y}_{(1)}\right)}{s_\nu},$$

  where s_ν is the estimator of the pooled standard deviation, based on $\nu = N - G = G(n-1)df$.

Significant Differences in 1-way ANOVA setting

- If there are G groups, then there is a maximum of $k = {}^{G}C2$ pairwise differences.
- Controlling Overall/ Family/ Batch/ Experimentwise/ Simultaneaous Type I error rate versus Individual/ Comparisonwise/ Pairwise Type I error rate
- Finding a pairwise significant difference:
  - Compare method-wise Significant Difference, $c(\alpha)$ with $|\bar{y}_i - \bar{y}_j|$ OR
  - Determine whether confidence interval contains 0 OR
  - Compare P-value with α
- $s = \sqrt{MSE}$ with $df = \nu =$dfERROR

Sig diff $\boxed{H_0 : \mu_i = \mu_j}$

1. $c(\alpha)$ vs $|\bar{y}_i - \bar{y}_j|$
2. P(2-sided) vs α : $P < \alpha$
3. CI does not contain 0

Tukey's Honestly Significant Difference (HSD)

- Denote critical values from the Studentized Range distribution as $q(G, \nu, \alpha)$ or t^*.
- $\boxed{\text{Family rate} = \alpha}$
- Pairwise rate $\leq \alpha$
- Tukey's HSD $= \boxed{\dfrac{q(\underbrace{G}_{\#\text{ of means}}, \underbrace{\nu}_{df\ \text{ERROR}}, \alpha)\underbrace{s}_{\sqrt{MSE}}}{\sqrt{n}}}$

$c(\alpha)$

Similar to:

$$\boxed{t^* \frac{s}{\sqrt{n}}}$$

Critical value from Tukey's distribution $q(\quad)$

Bonferroni's significant differences

- Conduct each test at level α/k
- $\boxed{\text{Family rate} \leq \alpha}$
- Pairwise rate$=\alpha/k$
- $c(\alpha)$ Significant difference $= t_{\alpha/2k,\nu} s\sqrt{\frac{2}{n}}$ width of C.I.

Bartlett's Test for Homogeneity of Variances

- Extension of F-test for equality of 2 variances
- Hypotheses:

$$H_0 : \sigma_1^2 = \sigma_1^2 = \cdots \sigma_G^2$$

$$H_a : \text{At least one } \sigma_g^2 \text{is different from the others}$$

- Test statistic:

$$T = \frac{(N-G)\ln S_p^2 - \sum_{g=1}^{G}(n_g - 1)\ln S_g^2}{1 + \frac{1}{3(G-1)}\left(\sum_{g=1}^{G}\frac{1}{n_g-1} - \frac{1}{N-G}\right)} \sim_{H_0} \chi^2_{G-1}$$

  where $S_p^2 = \sum_g^G (n_g - 1)S_g^2/(N-G)$ is the pooled variance
- In R: `bartlett()`
- A robust alternative test: Levene's, `levene.test()`

Model diagnostics: Any problems with model assumptions?

- Homoscedasticity: look at residuals in the diagnostic plots, use Bartlett's test, use rule of thumb
- Normality: use residual plots, or normal qq-plots
- Results: One slightly unusual observation but not influential value (large negative residual)

Model diagnostics: Constant variance?

2 ▶ Constant variance: Rule of Thumb for variances

$$\boxed{\text{If } \frac{\textit{largest} \;\; s_g}{\textit{smallest} \;\; s_g} < 2, \text{ assume variances are equal.}}$$

▶ For Spock's example, ignoring judge D since $n_D = 2$:

$$\frac{\textit{largest} \;\; s_g}{\textit{smallest} \;\; s_g} = \frac{11.9}{4.6} > 2$$

Hence, we may have a problem. Consider all inferences as only approximate.

3 ▶ Uncorrelated errors: This is satisfied if venires are chosen independently.

Case Study I Conclusion

Evidence of differences between
Spocks & A
" & B
" & C
" & E
" & F

We have evidence that mean % women on venires is different between Spock's judge and all other judges except judge D ($n_D = 2$) and no evidence of difference among other judges.

$H_0: \mu_S = \mu_D$ Not Rejected

Week 2 R functions

- One-way ANOVA: aov()
- Multiple Linear Regression Model: summary(lm())

- Barlett's Test of Equal Variance: bartlett.test()

- Bonferroni's: pairwise.t.test(), confint()
- Tukey's HSD: TukeyHSD(), confint()

STA303/1004 - Week 2 R Markdown

January 16-18, 2018

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)
```

```
##   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
```

```
Percent=jury$Percent
Judge=jury$Judge
```

Compare variances of 6 other judges: Rule of thumb

```
others <- jury[Judge != "Spock's",]
sss<-with(others, tapply(Percent,Judge,sd))
sss
```

```
##         A         B         C         D         E         F    Spock'
## 11.941817  6.582224  4.592929  3.818377  9.010142  5.968878        N
```

```
dim(sss)
```

```
## [1] 7
```

```
max(sss, na.rm=T)
```

```
## [1] 11.94182
```

```
min(sss, na.rm=T)
```

```
## [1] 3.818377
```

```
isTRUE((max(sss, na.rm=T)/min(sss, na.rm=T))>2)
```

```
## [1] TRUE
```

Compare variances of 6 other judges: Bartlett's

```
bartlett.test(Percent~Judge, data=others)
```

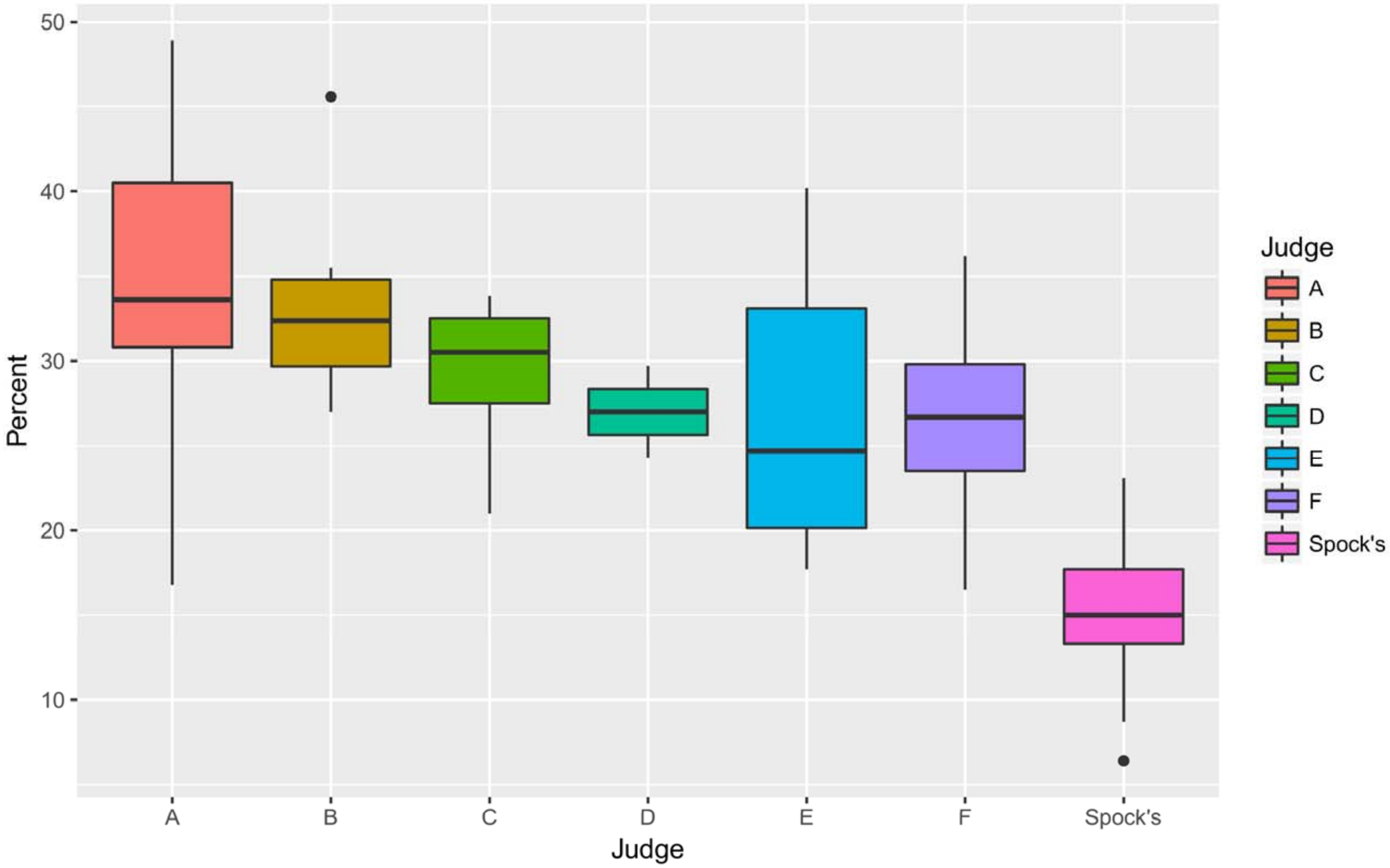
```
##
##  Bartlett test of homogeneity of variances
##
## data:  Percent by Judge
## Bartlett's K-squared = 6.3125, df = 5, p-value = 0.277
```

Note: Group sizes are uneven and some are very small

Do not reject H_0
H_0 assumed equal variances.

Compare all 7 judges

```
#boxplot(Percent~Judge)
library(ggplot2)
ggplot(jury, aes(x=Judge,y=Percent, fill=Judge))+geom_boxplot()
```



Compare means of all 7 judges: One-way ANOVA

$g=7$

```
summary(aov(Percent~Judge))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Judge        6   1927   321.2   6.718 6.1e-05 ***
## Residuals   39   1864    47.8                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Small P-value

Sig. result.

Compare means of all 7 judges: Gen Linear Model

```r
summary(lm(Percent~Judge))
```

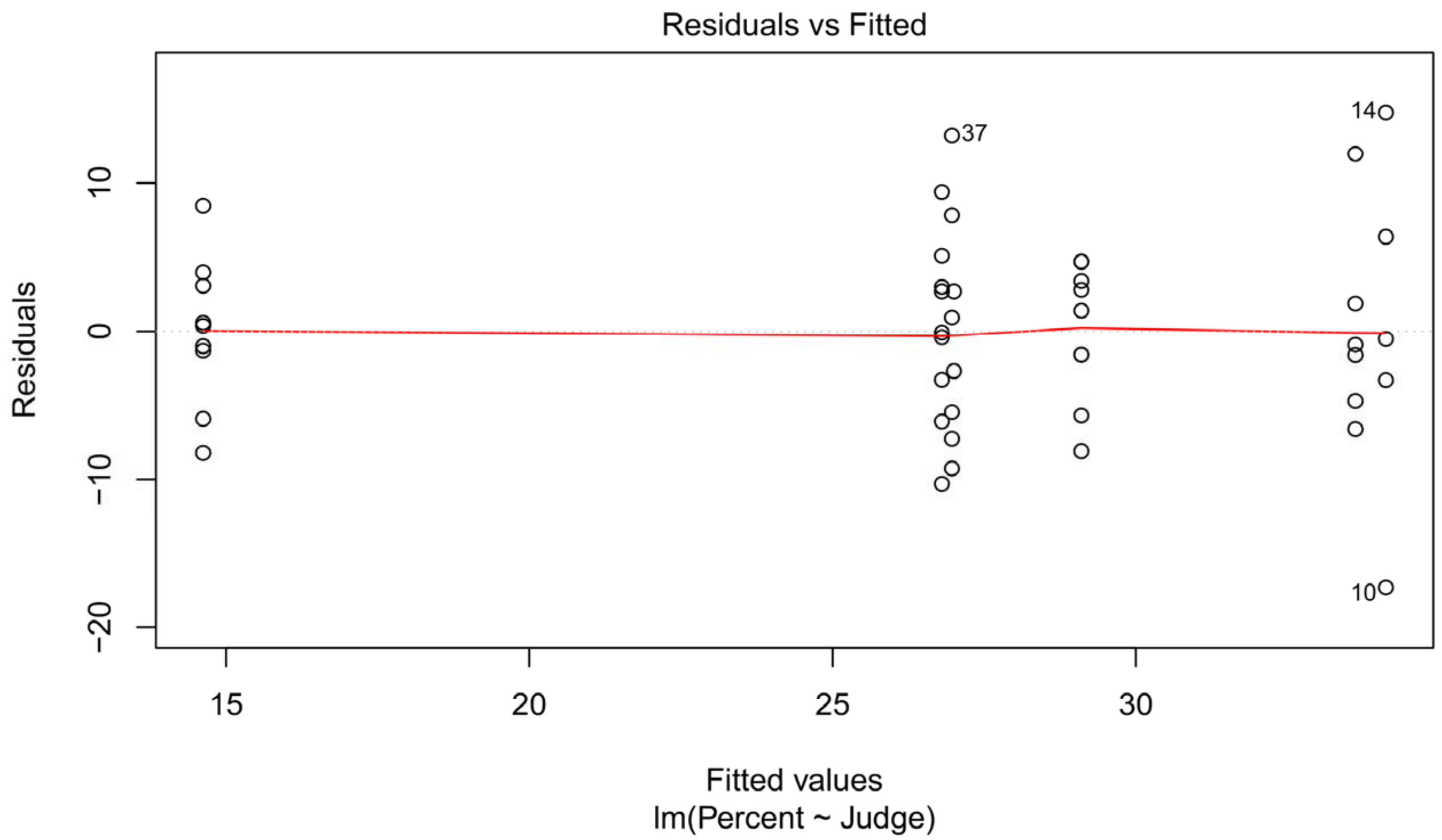
```
##
## Call:
## lm(formula = Percent ~ Judge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.320   -4.367   -0.250    3.319   14.780
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.1200     3.0921  11.034 1.47e-13 ***
## JudgeB          -0.5033     4.1868  -0.120   0.9049
## JudgeC          -5.0200     3.8566  -1.302   0.2007
## JudgeD          -7.1200     5.7848  -1.231   0.2258
## JudgeE          -7.1533     4.1868  -1.709   0.0955 .
## JudgeF          -7.3200     3.8566  -1.898   0.0651 .
## JudgeSpock's  -19.4978     3.8566  -5.056 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.914 on 39 degrees of freedom
```

β

Same P-value as in anova table on previous page

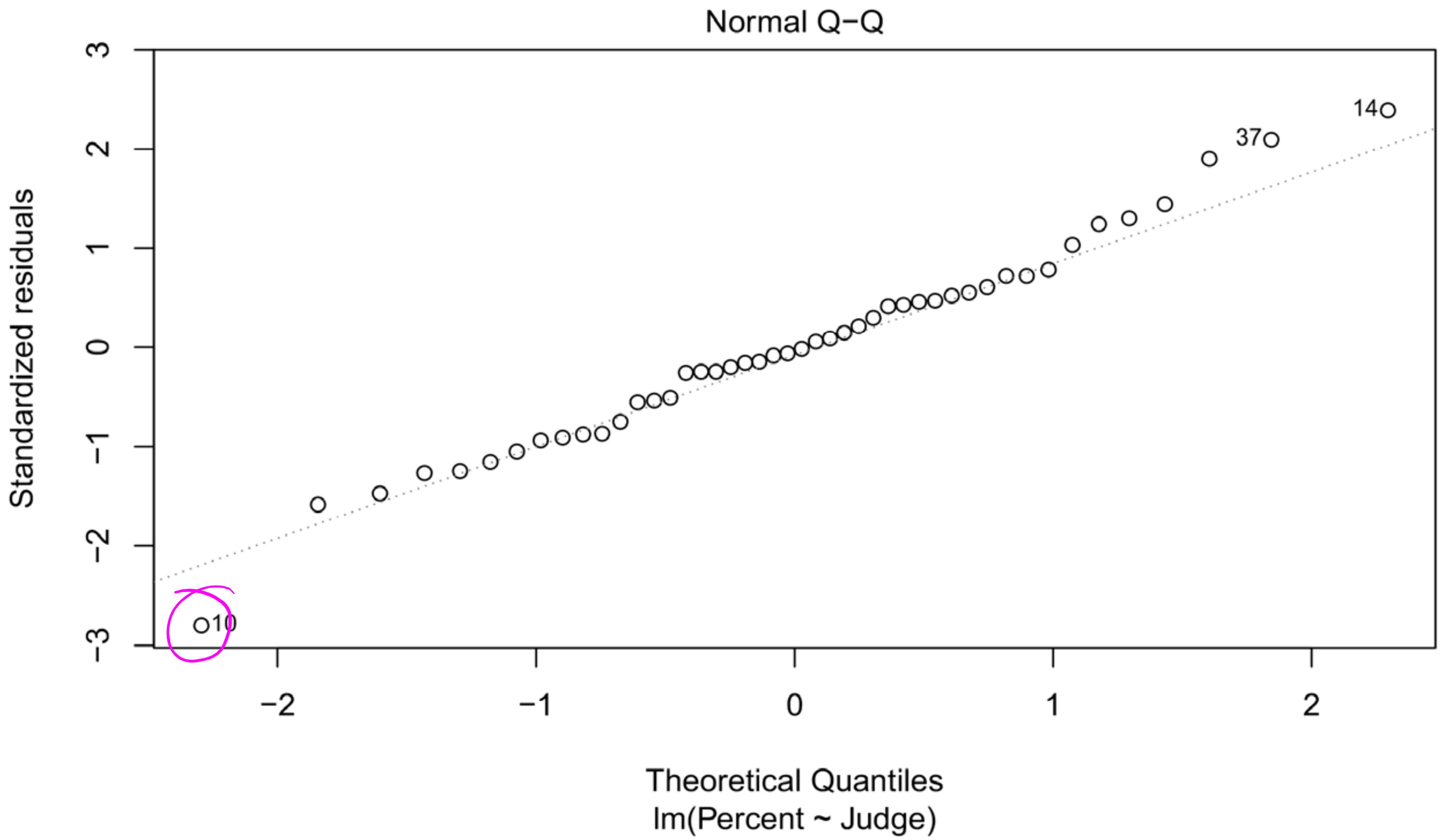
Check Normality: Linear Model

```
plot(lm(Percent~Judge), which=1)
```



Check Normality: Linear Model

```
plot(lm(Percent~Judge), which=2)
```



Compare variances of all 7 judges: RoT

```
ssa=with(jury, tapply(Percent, Judge, sd))
ssa
```

```
##         A        B        C        D        E        F   Spock'
## 11.941817 6.582224 4.592929 3.818377 9.010142 5.968878 5.03879
```

```
isTRUE((max(ssa, na.rm=T)/min(ssa, na.rm=T))>2)
```

```
## [1] TRUE
```

Compare variances of all 7 judges: Bartlett's

```
bartlett.test(Percent~Judge, data=jury)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Percent by Judge
## Bartlett's K-squared = 7.7582, df = 6, p-value = 0.2564
```

Case Study 1: Bonferroni's

```r
Judge=relevel(Judge, ref="Spock's")
pairwise.t.test(Percent, Judge, p.adj="bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Percent and Judge
##
##   Spock's A       B       C       D       E
## A 0.00022 -       -       -       -       -
## B 0.00013 1.00000 -       -       -       -
## C 0.00150 1.00000 1.00000 -       -       -
## D 0.57777 1.00000 1.00000 1.00000 -       -
## E 0.03408 1.00000 1.00000 1.00000 1.00000 -
## F 0.01254 1.00000 1.00000 1.00000 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

$\mu_A - \mu_S$
$\mu_B - \mu_S$
$\vdots$

Large P-value (0.5777) for the test of $H_0: \mu_D - \mu_{spock} = 0$
$\Rightarrow$ Evidence that μ_D is similar to μ_{spock}

Case Study 1: Bonferroni's CIs

```
lmod=lm(Percent~Judge)
nlevels(jury$Judge)
```

```
## [1] 7
```

```
confint(lmod, level=1-0.05/nlevels(jury$Judge))
```

```
##                  0.357 % 99.643 %
## (Intercept)   8.078085 21.16636
## JudgeA        8.547341 30.44821
## JudgeB        8.647255 29.34163
## JudgeC        5.222970 23.73259
## JudgeD       -2.969585 27.72514
## JudgeE        1.997255 22.69163
## JudgeF        2.922970 21.43259
```

Includes 0

Case Study 1: Bonferroni's CIs

```r
summary(lmod)
```

```
##
## Call:
## lm(formula = Percent ~ Judge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.320  -4.367  -0.250   3.319  14.780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.622      2.305   6.344 1.72e-07 ***
## JudgeA        19.498      3.857   5.056 1.05e-05 ***
## JudgeB        18.994      3.644   5.212 6.39e-06 ***
## JudgeC        14.478      3.259   4.442 7.15e-05 ***
## JudgeD        12.378      5.405   2.290 0.027513 *
## JudgeE        12.344      3.644   3.388 0.001623 **
## JudgeF        12.178      3.259   3.736 0.000597 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.914 on 39 degrees of freedom
```

Case Study 1: Tukey's CIs

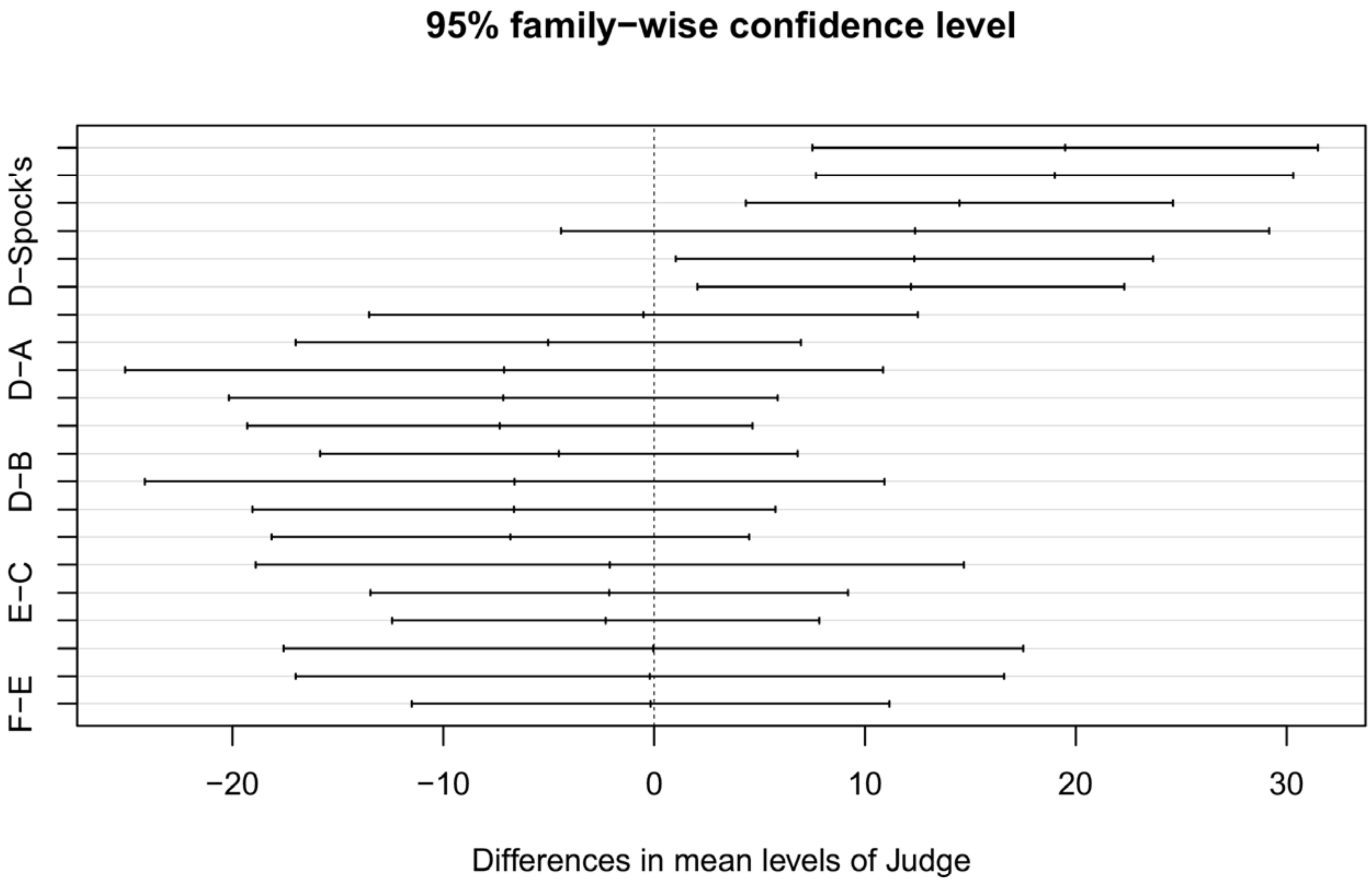
```r
amod=aov(Percent~Judge)
TukeyHSD(amod,"Judge")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Percent ~ Judge)
##
## $Judge
##                  diff        lwr       upr     p adj
## A-Spock's 19.49777778   7.514686 31.480870 0.0001992
## B-Spock's 18.99444444   7.671487 30.317402 0.0001224
## C-Spock's 14.47777778   4.350216 24.605339 0.0012936
## D-Spock's 12.37777778  -4.416883 29.172438 0.2744263
## E-Spock's 12.34444444   1.021487 23.667402 0.0248789
## F-Spock's 12.17777778   2.050216 22.305339 0.0098340
## B-A       -0.50333333 -13.512422 12.505755 0.9999997
## C-A       -5.02000000 -17.003092  6.963092 0.8470097
## D-A       -7.12000000 -25.094638 10.854638 0.8777485
## E-A       -7.15333333 -20.162422  5.855755 0.6146238
## F-A       -7.32000000 -19.303092  4.663092 0.4936379
## C-B       -4.51666667 -15.839625  6.806291 0.8742030
## D-B       -6.61666667 -24.158118 10.924784 0.9003280
```

k=21

Case Study 1: Tukey's CI's

```
plot(TukeyHSD(amod,"Judge"))
```



A, S
B, S
C, S
E, S
F, S