

*Introduction to  
Statistics in R  
Presented by:*



BROWN  
*Center for*  
Biomedical Informatics



# Introduction to Statistics in R

## Day 3 - ANOVA

Adam J Sullivan

**What is ANOVA?**

**Recap on our progress:**

**Enter ANOVA**

**Enter ANOVA**

# Enter ANOVA

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

At least one group is different

# What is our test then?

$K$

$t$

$$k = \frac{\text{Measure of Between-Group Variability}}{\text{Measure of Within-Group Variability}}$$



# The math

$$SS_B = \sum_{i=1}^k n_i (\bar{y})_i - \bar{y})^2$$

$$SS_W = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})_i)^2$$

$$SS = SS_B + SS_W$$

$i$

$k$

$j$

$n_i$

# ANOVA Variances

# ANOVA Table

	DF	SUM SQ.	MEAN SQ	F VALUE	PR(>F)
Between (treatment)	$k - 1$	$SS_B$	$MS_B = \frac{SS_B}{k - 1}$	$\frac{MS_{trt}}{MS_{err}}$	p-value
Within (error)	$N - k$	$SS_W$	$MS_W = \frac{SS_W}{N - k}$		
Total	$N - 1$	$SS_T$			

# Calculating ANOVA

# The Data for Class

---

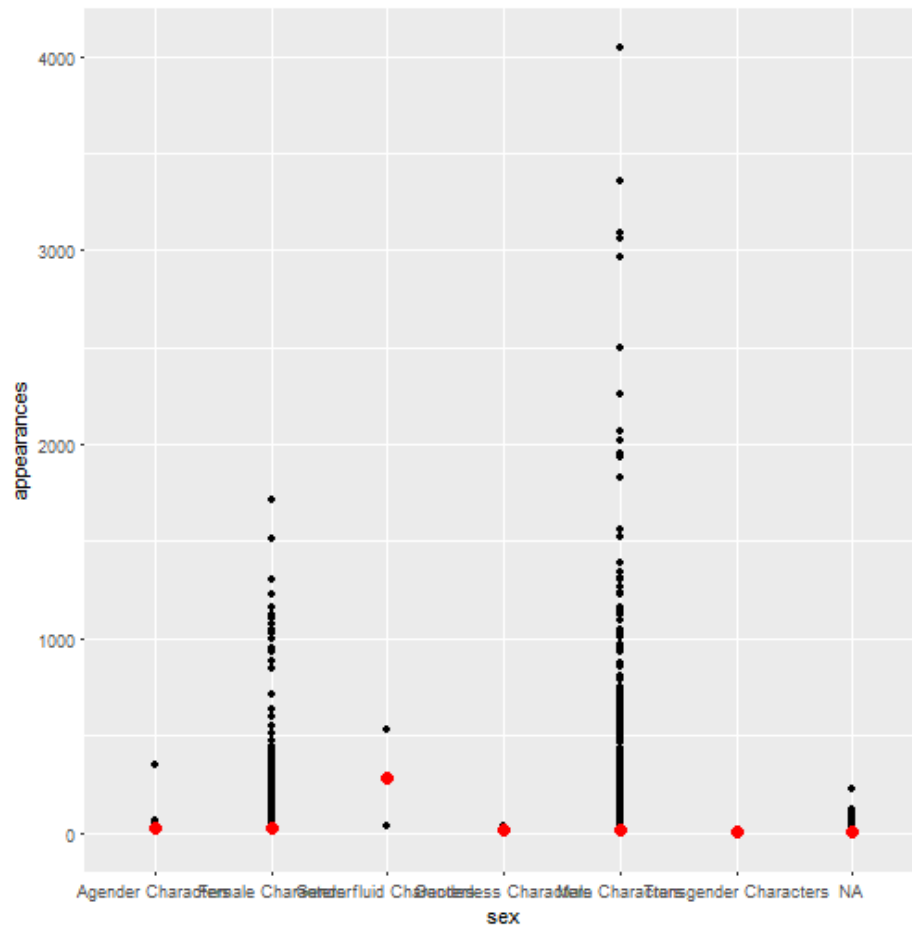
fivethirtyeight

```
library(fivethirtyeight)
?comic_characters
```

# Difference in Appearances by Gender

```
ggplot(comic_characters, aes(x = sex, y = appearances)) +  
  geom_point() +  
  geom_point(stat = "summary", fun.y = "mean", color = "red", size = 3)
```

# Difference in Appearances by Gender



# Table of Gender

```
comic_characters %>%  
  group_by(sex) %>%  
  tally(sort = TRUE)
```



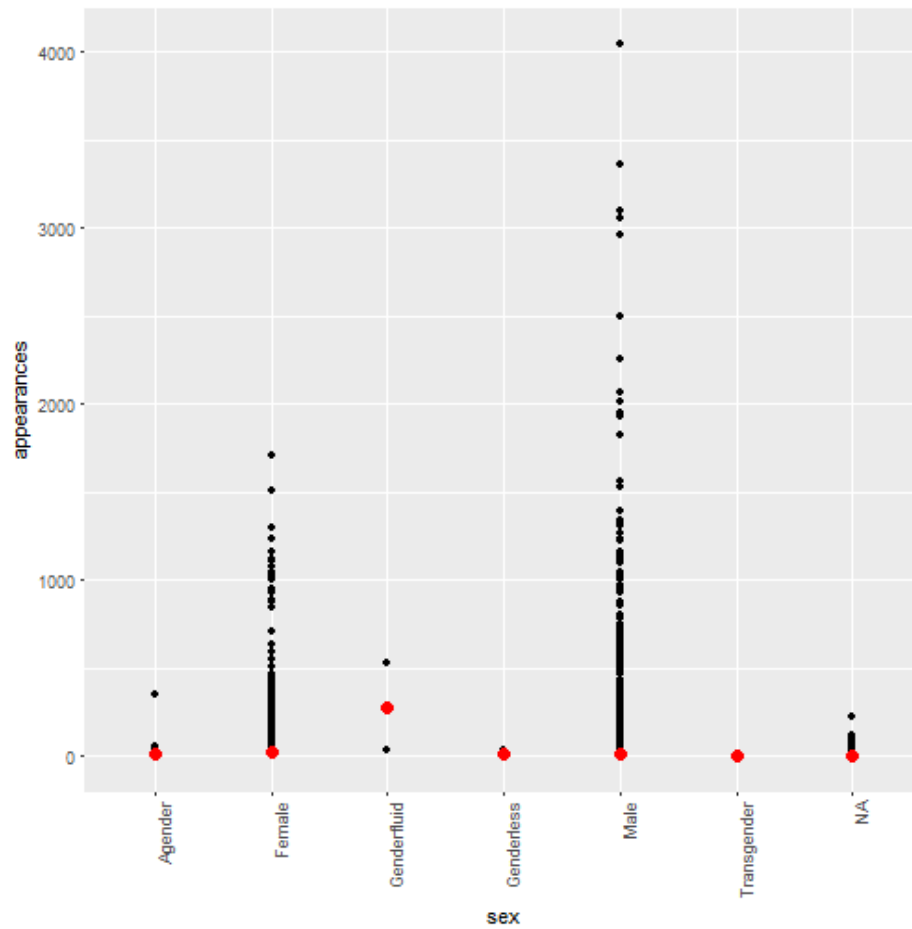
# Table of Gender

```
## # A tibble: 7 x 2
##   sex          n
##   <chr>      <int>
## 1 Male Characters 16421
## 2 Female Characters 5804
## 3 <NA>          979
## 4 Agender Characters 45
## 5 Genderless Characters 20
## 6 Genderfluid Characters 2
## 7 Transgender Characters 1
```

# Data Cleaning

```
comic <- comic_characters %>%  
  mutate(sex = fct_recode(sex,  
    "Agender" = "Agender Characters",  
    "Female" = "Female Characters",  
    "Genderfluid" = "Genderfluid Characters",  
    "Genderless" = "Genderless Characters",  
    "Male" = "Male Characters",  
    "Transgender" = "Transgender Characters"  
  ))
```

# Data Cleaning



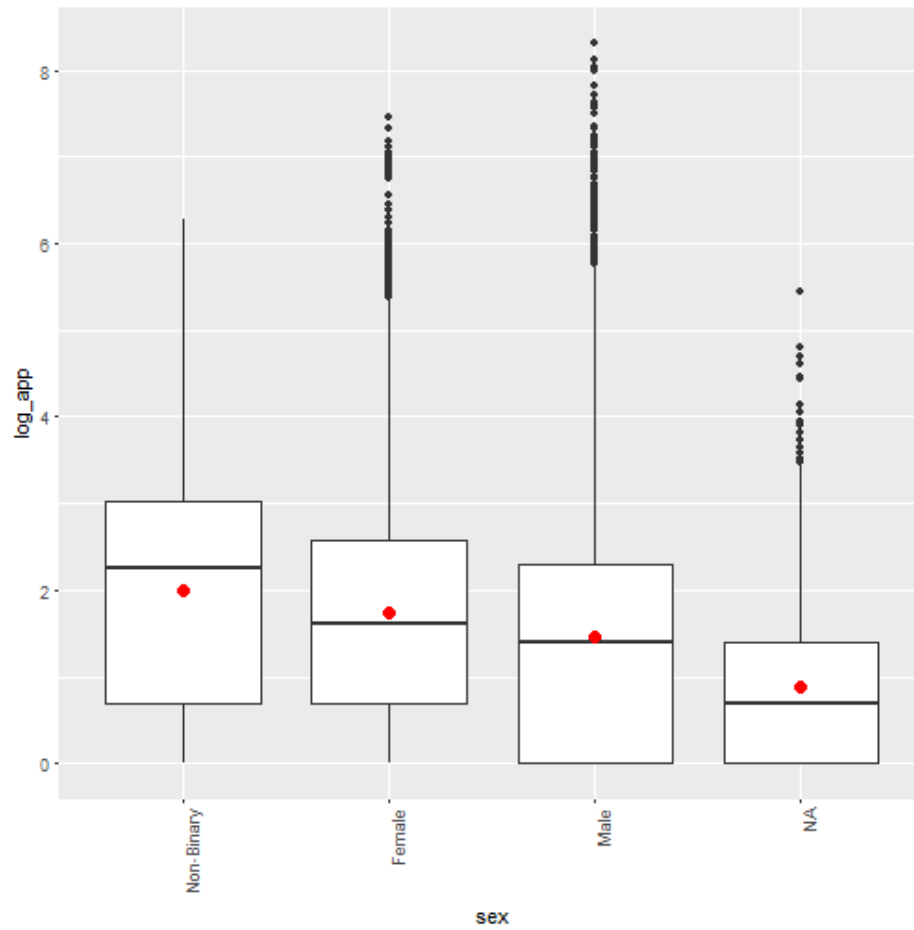
# Cleaning Data

```
comic <- comic_characters %>%  
  mutate(sex = fct_recode(sex,  
    "Non-Binary" = "Agender Characters",  
    "Female" = "Female Characters",  
    "Non-Binary" = "Genderfluid Characters",  
    "Non-Binary" = "Genderless Characters",  
    "Male" = "Male Characters",  
    "Non-Binary" = "Transgender Characters"  
  ))
```

# Cleaning Data

```
comic <- comic %>%  
  mutate(log_app = log(appearances))
```

# Boxplots



# Finally ANOVA

```
aov(log_app~sex, data=comic)
```

```
## Call:
##   aov(formula = log_app ~ sex, data = comic)
##
## Terms:
##              sex Residuals
## Sum of Squares   296.09 40225.14
## Deg. of Freedom     2    20966
##
## Residual standard error: 1.385132
## Estimated effects may be unbalanced
## 2303 observations deleted due to missingness
```

# What can we do to get more information

```
my_anova <- aov(log_app~sex, data=comic)
names(my_anova)
```

```
## [1] "coefficients" "residuals"      "effects"        "rank"
## [5] "fitted.values" "assign"         "qr"            "df.residual"
## [9] "na.action"     "contrasts"      "xlevels"       "call"
## [13] "terms"         "model"
```



# Summary

```
summary(my_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sex              2     296   148.05    77.16 <2e-16 ***
## Residuals    20966   40225     1.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2303 observations deleted due to missingness
```

# What were we testing again?

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

At least one group is different

**What is Next?**

# What about Multiple Testing

$$\begin{aligned}\Pr(\text{At least 1 Significant Result}) &= 1 - \Pr(\text{No Significant Results}) \\ &= 1 - (1 - 0.05)^{20} \\ &= 0.6415141\end{aligned}$$

# What Type of Multiple Tests for ANOVA

$$\text{FWER} \leq 0.05$$

# The Bonferroni Correction

$$n$$

$$\alpha^* = \frac{\alpha}{n}$$

$$\min \left[ 2 \times \binom{k}{2} \times \Pr(|t| < t_{n-k}), 1 \right]$$

# The Bonferroni Correction

$$n = 20$$

$$\alpha = 0.05$$

$$\alpha^* = \frac{\alpha}{n} = \frac{0.05}{20} = 0.0025$$

$$\begin{aligned}\Pr(\text{At least 1 Significant Result}) &= 1 - \Pr(\text{No Significant Results}) \\ &= 1 - (1 - 0.0025)^{20} \\ &= 0.04883012\end{aligned}$$

# Bonferonni in R

```
pairwise.t.test(x,g,p.adjust.method,...)
```

x

g

p.adjust.method

...



# Bonferonni in R

```
attach(comic)
pairwise.t.test(log_app,sex, p.adjust="none")
detach()
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: log_app and sex
##
##      Non-Binary Female
## Female 0.1283      -
## Male   0.0023      <2e-16
##
## P value adjustment method: none
```

$$\alpha = 0.0025$$

# Bonferroni in R

```
attach(comic)
pairwise.t.test(log_app,sex, p.adjust="bonferroni")
detach()
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: log_app and sex
##
##          Non-Binary Female
## Female 0.3850      -
## Male   0.0069    <2e-16
##
## P value adjustment method: bonferroni
```

# Tukey HSD Test

# Tukey HSD in R

```
TukeyHSD(my_anova, conf.level=0.95)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = log_app ~ sex, data = comic)
##
## $sex
##              diff          lwr          upr      p adj
## Female-Non-Binary -0.2648371 -0.6730209  0.1433466 0.2811377
## Male-Non-Binary   -0.5292183 -0.9358797 -0.1225569 0.0064760
## Male-Female       -0.2643812 -0.3154401 -0.2133222 0.0000000
```

# Results

# Assumptions of ANOVA

# Testing Assumptions of ANOVA

# Testing Assumptions of ANOVA

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

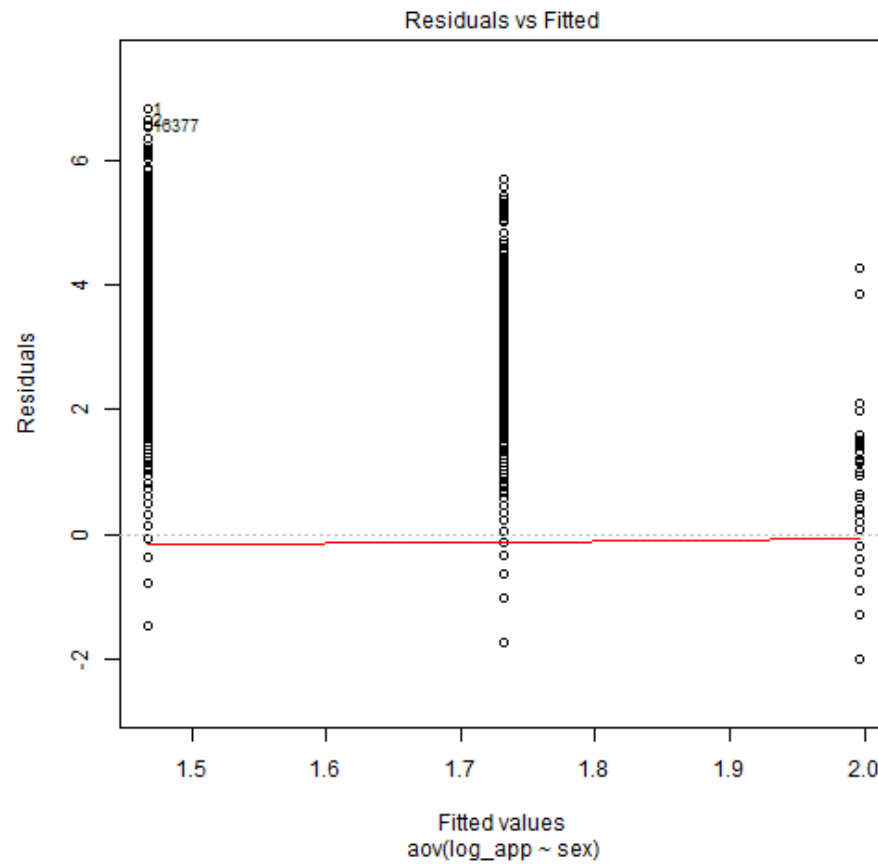
at least one variance is different

```
plot(my_anova, 1)

library(car)
leveneTest(log_app~sex, data = comic)
```



# Testing Assumptions of ANOVA



# Testing Assumptions of ANOVA

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value Pr(>F)
## group      2  0.1827 0.8331
##           20966
```

# What if we do not have Homoscedastic Variances?

```
attach(comic)
pairwise.t.test(log_app,sex, p.adjust="bonferroni", pool.sd=FALSE)
detach()
```

# Testing Assumptions of ANOVA

Population is Normally Distributed

Population is not Normally Distributed

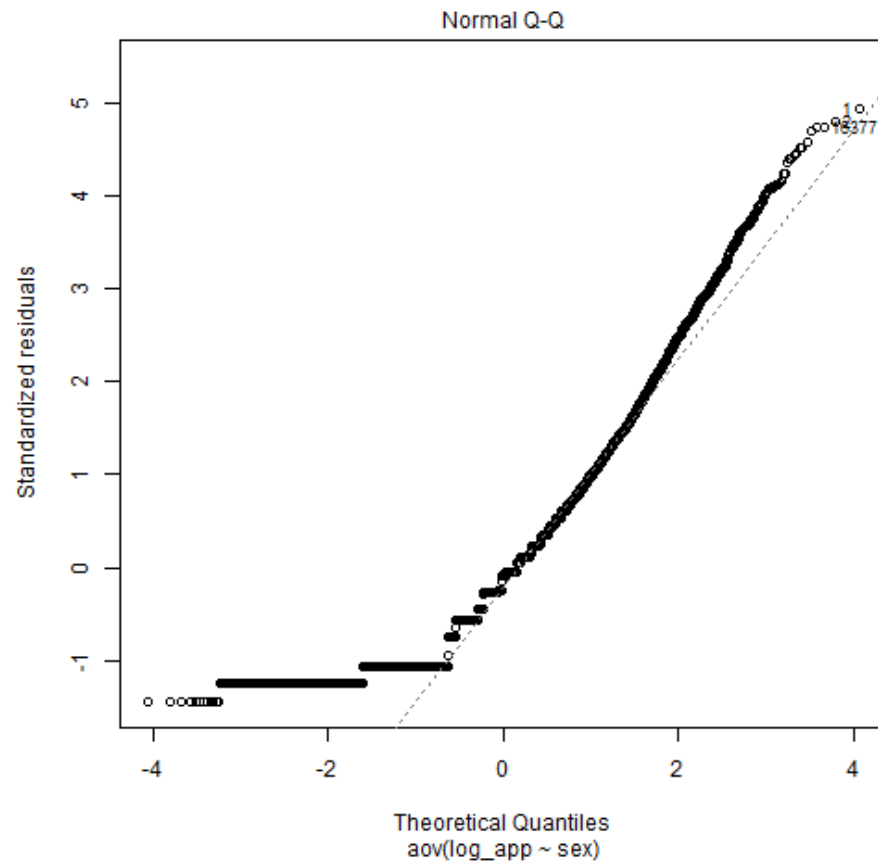
```
plot(my_anova, 2)
```

```
#install.packages("nortest")
```

```
library(nortest)
```

```
lillie.test(my_anova_resid)
```

# Testing Assumptions of ANOVA



# Testing Assumptions of ANOVA

```
## Error in sort(x[complete.cases(x)]): object 'my_anova_resid' not found
```

**What is Normality is not met?**

**Questions**



**Lab Time**