HW1: Solutions

Blain Morin

February 15, 2019

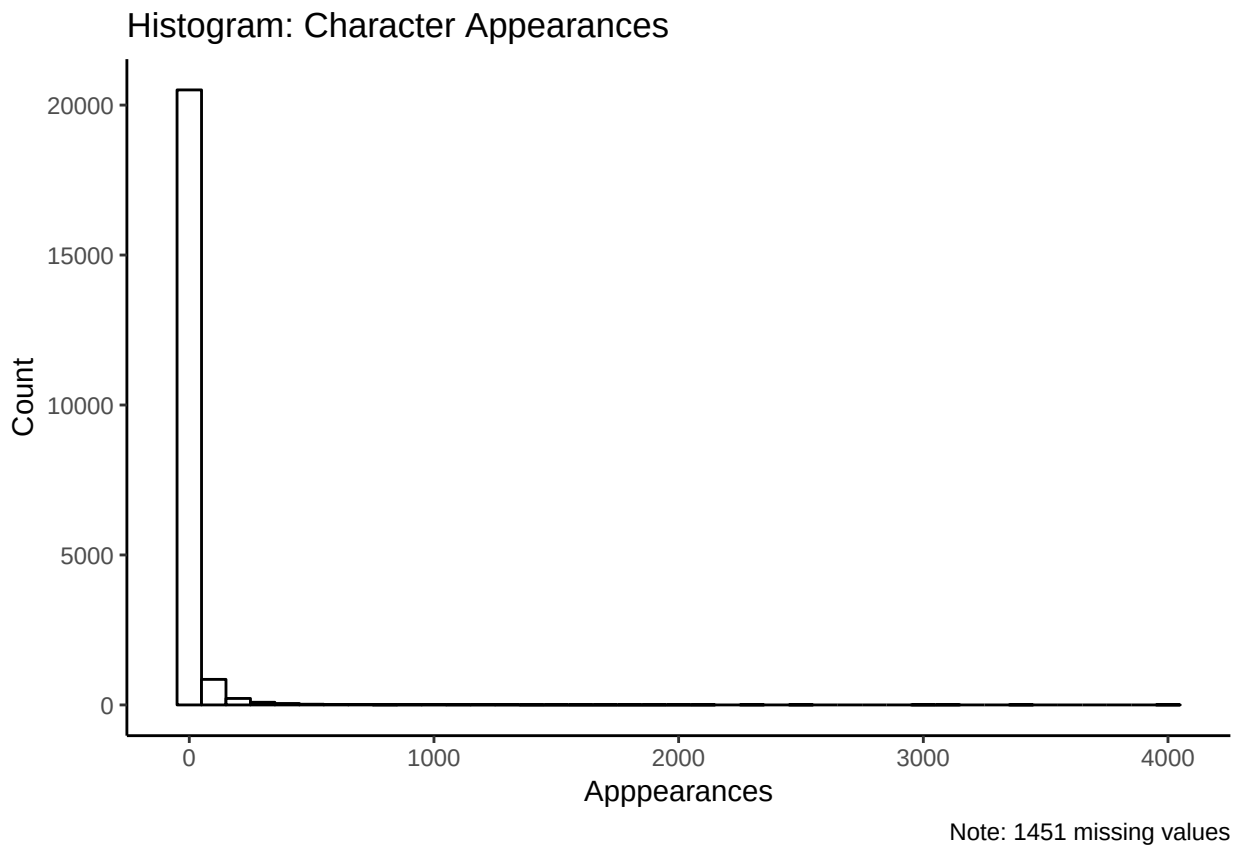
```
### Set knitr options
knitr::opts_chunk$set(warning = FALSE, message = FALSE, cache = TRUE)

### Load Required Libraries
require(fivethirtyeight) == T || install.packages("fivethirtyeight")
require(dplyr) == T || install.packages("dplyr")
require(ggplot2) == T || install.packages("ggplot2")
require(forcats) == T || install.packages("forcats")
require(stargazer) == T || install.packages("stargazer")
require(xtable) == T || install.packages("xtable")
require(knitr) == T || install.packages("knitr")

### Load data
comic = comic_characters
```

1.) We will evaluate the number of appearances as the outcome. Create a plot to display the distribution of this variable. Then interpret the graph and note the normality, skewness and anything else you notice about this.

```
### Make a histogram
comic %>%
  ggplot(aes(x = appearances)) +
  geom_histogram(binwidth = 100, color = 'black', fill = 'white') +
  ylab("Count") +
  xlab("Apppearances") +
  ggtitle("Histogram: Character Appearances") +
  theme_classic() +
  labs(caption = "Note: 1451 missing values")
```

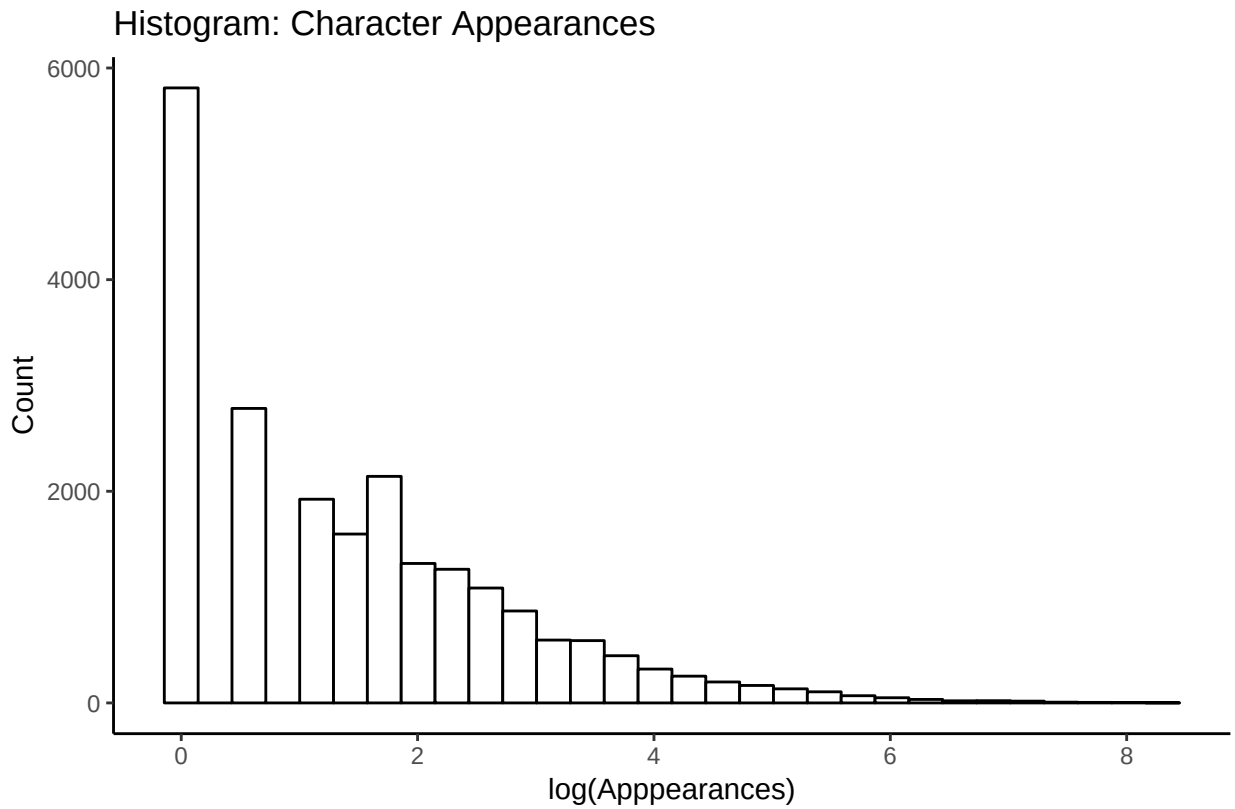


The histogram of character appearances is heavily skewed right. The appearance variable is also bounded by 0. The histogram does not look normally distributed.

2.) Hopefully you noted that this was not normally distributed so the next step would be to consider a log transform of this variable. Use the mutate() function to add a variable log_app which is the log of the appearances. Then create the same distribution plot above for log_app.

```
### Create log_app
comic = comic %>%
  mutate(log_app = log(appearances))

### Make a histogram
comic %>%
  ggplot(aes(x = log_app)) +
  geom_histogram(color = 'black', fill = 'white') +
  ylab("Count") +
  xlab("log(Apppearances)") +
  ggtitle("Histogram: Character Appearances") +
  theme_classic() +
  labs(caption = "Note: 1451 missing values")
```



3.) We will now consider the variable sex in the dataset. Make a table of counts of how many characters are in each group. You can use tally() or count() for this. Comment on what you notice about these groups.

```
### Tally by sex
sex_count = comic %>%
  group_by(sex) %>%
  tally() %>%
  rename(Sex = sex)

### Change NA to a string for the table
sex_count[7,1] = "NA"

### Make a table using stargazer
stargazer(sex_count,
          header = FALSE,
          title = "Number of Characters in Each Gender Group",
          summary = FALSE,
          rownames = FALSE,
          table.placement = 'H')
```

Table 1: Number of Characters in Each Gender Group

Sex	n
Agender Characters	45
Female Characters	5804
Genderfluid Characters	2
Genderless Characters	20
Male Characters	16421
Transgender Characters	1
NA	979

We see that of the 23272 characters, more than half are male. About 25% are female. 979 are missing gender information. There are relatively few observations in the rest of the categories. Moreover, some of the categories seem like they can be combined (agender and gender neutral for example).

4.) Run an ANOVA considering log_app over the different genders (sex). Are the appearances different by gender (sex)? Do you think ANOVA is appropriate across all these categories? Why or Why not?

```
### Regress log_app on sex
model = lm(log_app ~ sex, data = comic)

### Run ANOVA
anova.4 = aov(model)

### Make a table
anova.4.table = xtable(summary(anova.4),
       caption = "ANOVA: Appearances by Gender")

print(anova.4.table,
      caption.placement = 'top',
      comment = FALSE)
```

Table 2: ANOVA: Appearances by Gender

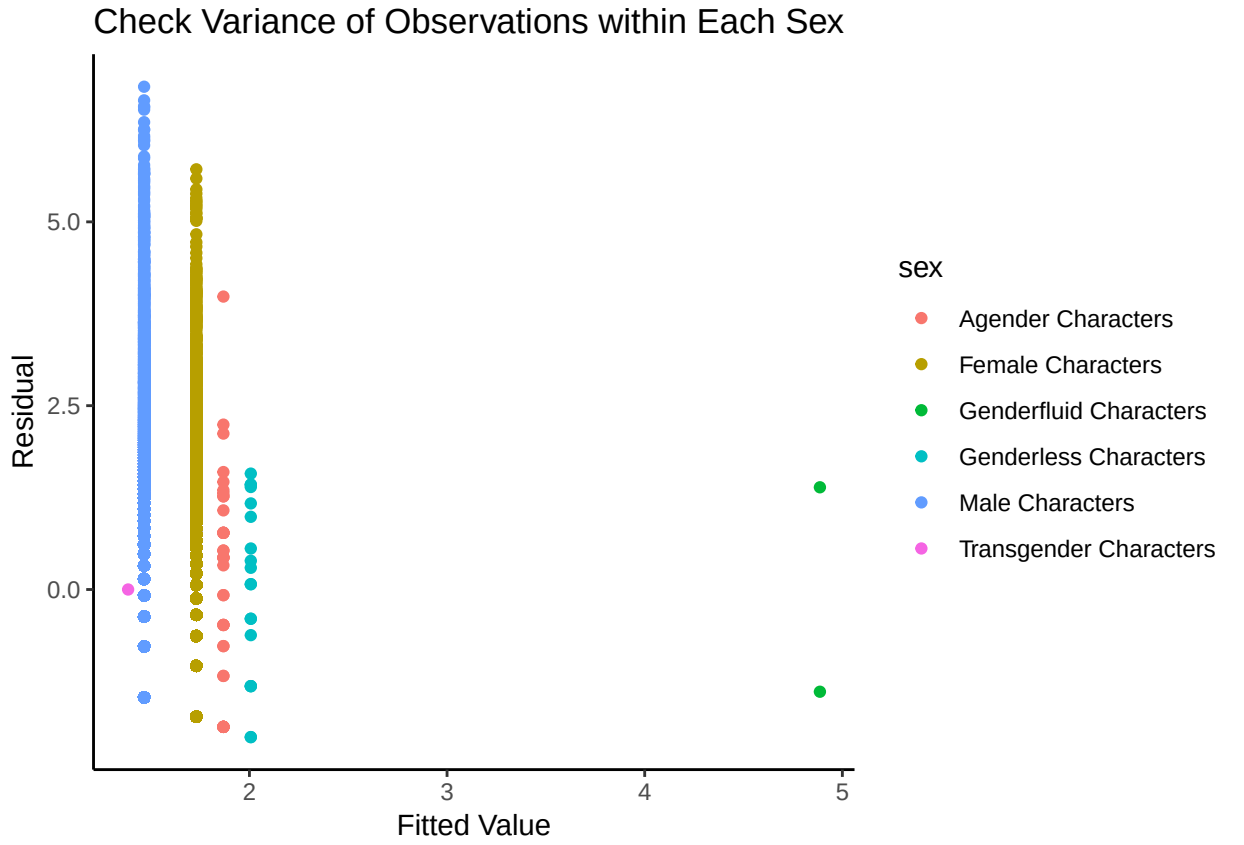
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	5	313.86	62.77	32.73	0.0000
Residuals	20963	40207.37	1.92		

The F-test is highly significant. This means that there is a significant difference in appearances in at least one of the groups. ANOVA is not appropriate here because some of the groups do not have enough observations. For example, the transgender category only has one observation. There is not enough information to test whether or not this category is statistically different.

5.) What are the assumptions of ANOVA? Which if any would be a problem with log_app and sex?

- The observations are obtained independently and randomly from the population defined by the factor levels
- The data of each factor level are normally distributed
- These normal populations have a common variance

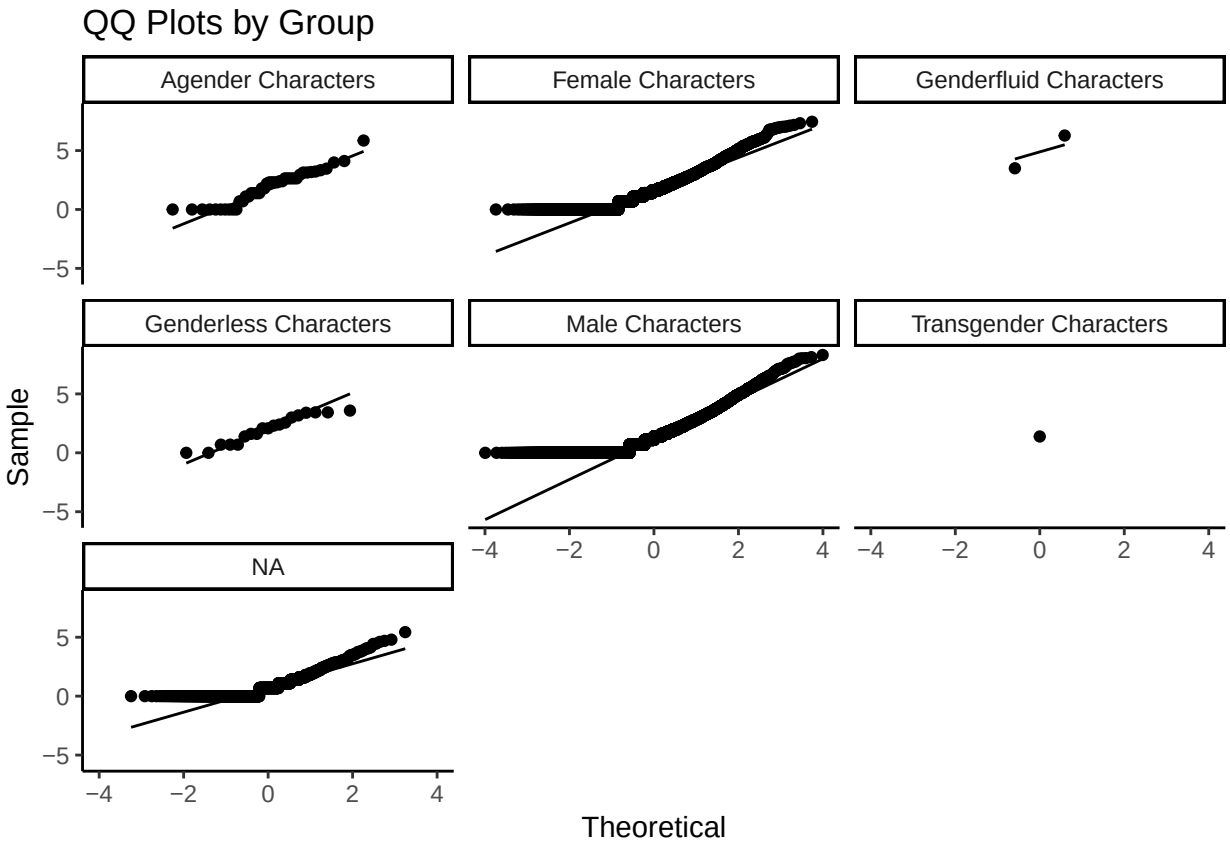
```
### Check common variance assumption
model %>%
  ggplot() +
  geom_point(aes(x = .fitted, y = .resid, color = sex)) +
  ylab("Residual") +
  xlab("Fitted Value") +
  ggtitle("Check Variance of Observations within Each Sex") +
  theme_classic()
```



The variance for the male group appears to be greater than the variance for the other groups. This violates our ANOVA assumptions.

```
### Check normality of the observations within each group
comic %>%
  ggplot(aes(sample = log_app)) +
  facet_wrap(~sex) +
  geom_qq() +
  geom_qq_line() +
  xlab("Theoretical") +
```

```
  ylab("Sample") +
  ggtitle("QQ Plots by Group") +
  theme_classic()
```



If the distribution was normal, the points would be close to the line. We see that the points diverge from the line at the lower ends. This is evidence that their distribution is not normal. Also, we do not have enough points in the gender fluid and transgender factors to make any conclusions about their distributions.

6.) Using mutate() and fct_relevel() create a new variable called gender with three categories: male, female, and non-binary

```
### Refactor sex
comic = comic %>%
  filter(!is.na(sex)) %>%
  mutate(gender = ifelse(sex == "Male Characters",
                         "male", ifelse(sex == "Female Characters",
                                "female", "non-binary"))) %>%
  mutate(gender = as.factor(gender))
```

7.) Run an ANOVA considering log_app over the new gender variable. Are there differences between your three gender categories?

```
### Regress log_app on sex
model7 = lm(log_app ~ gender, data = comic)

### Run ANOVA
anova.7 = aov(model7)

### Make a table
anova.7.table = xtable(summary(anova.7),
       caption = "ANOVA: Appearances by Gender")

print(anova.7.table,
      caption.placement = 'top',
      comment = FALSE)
```

Table 3: ANOVA: Appearances by Gender

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	2	296.09	148.05	77.16	0.0000
Residuals	20966	40225.14	1.92		

The F-test is again highly significant. This suggests that there is a significant difference in appearances for at least on of the gender groups.

8.) Using TukeyHSD() perform p-value adjusted t-tests to compare each category of gender. What groups are different than the others?

```
### Use TukeyHSD
tukey.8 = TukeyHSD(anova.7)

### Make a table
stargazer(tukey.8$gender,
          header = FALSE,
          summary = FALSE,
          table.placement = 'H',
          title = "Tukey Pairwise Comparisons for Gender Appearances")
```

Table 4: Tukey Pairwise Comparisons for Gender Appearances

	diff	lwr	upr	p adj
male-female	-0.264	-0.315	-0.213	0
non-binary-female	0.265	-0.143	0.673	0.281
non-binary-male	0.529	0.123	0.936	0.006

There is a significant difference between male and female appearances. There is also a significant difference between male and non-binary appearances.

9.) Interpret the overall results of this ANOVA. What does this tell you about appearances and gender.

Overall, we see that male characters on average have a significantly higher number of appearances.

10.) Does your interpretation make sense for what you think is going on? Why might the one particular group be lower than the others?

The data shows that females and non-binary characters are underrepresented in the current comic universe. Males may be over-represented because the industry perceives males to be the target audience.