Week 6 Assignment

Jessalyn Chuang

```r
# Load necessary packages.
library(here)
```

```
## here() starts at /home/guest/Statistical_Modeling_Sp25
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

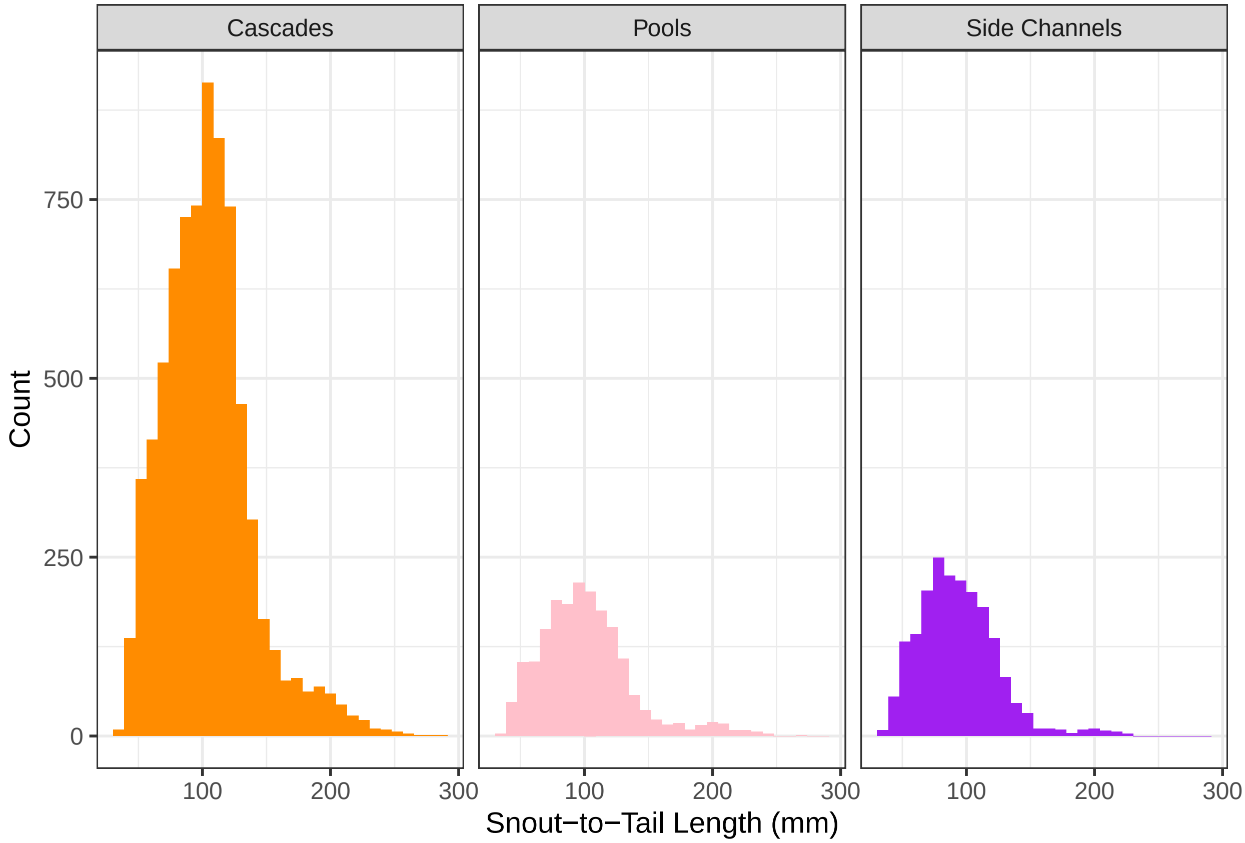
```r
library(moments)
library(dplyr)
```

#(1) Coastal Giant Salamander- a. Filter the dataset only for coastal giant salamanders in cascades (C), pools (P), and side channels (SC). Create a figure that helps you to evaluate whether their snout-to-tail length (length_2_mm) by habitat type (unittype) is normally distributed. You may also calculate skew and kurtosis values to help with your decision-making. Are these data normally distributed? Why or why not? If they are not, apply a log-transform (log10()) to the length data and re-evaluate if the data now meets the criteria to be considered normally-distributed.

```r
vertebrates_data <- read.csv(file = here("./Assignments/week6/and_vertebrates.csv"),
                             stringsAsFactors = TRUE)

salamander_data <- filter(vertebrates_data,
                              species == "Coastal giant salamander",
                              unittype %in% c('C', 'P', 'SC'))

salamander_fig1 <- ggplot(salamander_data, aes(x = length_2_mm, fill = unittype)) +
  geom_histogram() +
  scale_fill_manual(values = c("darkorange", "pink", "purple")) +
  labs(x = "Snout-to-Tail Length (mm)", y = "Count") +
  facet_grid(. ~ unittype, labeller = labeller(unittype = c("C" = "Cascades",
                                                            "P" = "Pools",
                                                          "SC" = "Side Channels"))) +
  theme_bw() +
  theme(legend.position = "none")


salamander_fig1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 220 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
# Separate data by unittype
C_dat <- salamander_data %>%
filter(unittype == "C")

P_dat <- salamander_data %>%
filter(unittype == "P")

SC_dat <- salamander_data %>%
filter(unittype == "SC")

# Calculate skew.
skewness(C_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 0.9428883
```

```r
skewness(P_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 1.106651
```

```r
skewness(SC_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 0.9912964
```

```r
# Calculate kurtosis.
kurtosis(C_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 4.691787
```

```r
kurtosis(P_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 4.894066
```

```r
kurtosis(SC_dat$length_2_mm, na.rm = TRUE)
```

```
## [1] 4.937146
```

Skew results: Cascades: 0.94 Pool: 1.11 Side Channels: 0.99

Kurtosis results: Cascades: 4.69 Pool: 4.89 Side Channels: 4.94

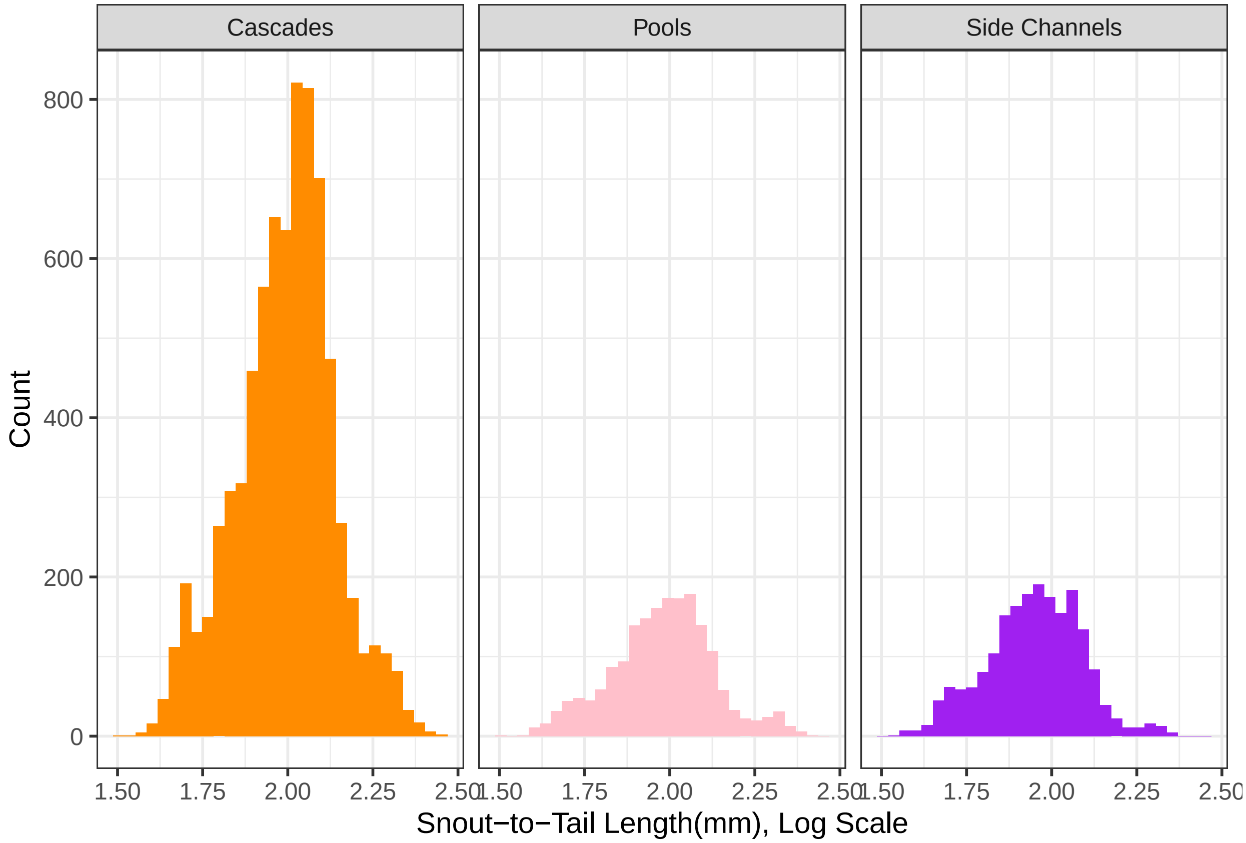
Visually, the three data from the three habitats all seem to be positive skewed, with the majority of points concentrated at lower values. From the skew results, all three have values that suggest positive skewing as they are all greater than 0. Lastly for kurtosis, all three have values greater than 3, suggesting that the data is leptokurtic. Therefore, this data is not normally distributed!

```r
#log-transformation of data
salamander_data_transformed <- mutate(salamander_data, length_2_mm = log10(length_2_mm))
```

```r
salamander_fig2 <- ggplot(salamander_data_transformed, aes(x = length_2_mm, fill = unittype)) +
  geom_histogram() +
  scale_fill_manual(values = c("darkorange", "pink", "purple")) +
  labs(x = "Snout-to-Tail Length(mm), Log Scale", y = "Count") +
  facet_grid(.~unittype, labeller = labeller(unittype = c("C" = "Cascades",
                                                          "P" = "Pools",
                                                          "SC" = "Side Channels"))) +
  theme_bw() +
  theme(legend.position = "none")

salamander_fig2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 220 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
# Separate data by unittype
C_dat_trans <- salamander_data_transformed %>%
filter(unittype == "C")

P_dat_trans <- salamander_data_transformed %>%
filter(unittype == "P")

SC_dat_trans <- salamander_data_transformed %>%
filter(unittype == "SC")

# Calculate skew.
skewness(C_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] -0.1713156
```

```
skewness(P_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] -0.04432009
```

```
skewness(SC_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] -0.07959806
```

```
# Calculate kurtosis.
kurtosis(C_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] 3.08717
```

```
kurtosis(P_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] 3.083047
```

```
kurtosis(SC_dat_trans$length_2_mm, na.rm = TRUE)
```

```
## [1] 2.981829
```

Now, all three habitat types appear to be normally distributed from the visual inspection. Their skewness values are now all around 0 and kurtosis values are all around 3 (mesokurtic).

b. If you found the data from part a were not normally distributed and a log-transform did not change this finding, you may stop here and proceed to question 2. If you found the data from part a were normally distributed, conduct a Bartlett's test for equal variance to determine if these data also satisfy the need for homogeneity of variances across groups. Do these data have approximately equal variances? Why or why not? Remember, the data may not pass the Bartlett's test, but if they adhere to the rule of thumb mentioned above, you may proceed with a one-way ANOVA.

```
# Perform Bartlett test.
salamander_var <- bartlett.test(salamander_data_transformed$length_2_mm,
                                salamander_data_transformed$unittype)
# Examine results.
salamander_var
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  salamander_data_transformed$length_2_mm and salamander_data_transformed$unittype
## Bartlett's K-squared = 11.213, df = 2, p-value = 0.003674
```

Since p-value is less than 0.05, I reject the null hypothesis stating that the variance between all groups included in the data are the same in favor of the alternate hypothesis that states that the variance between all groups included in the data are not the same. Thus, this data does not pass the Bartlett's test. However, if the largest sample variance is less than 4 times the smallest sample variance, we may still assume variances are equal across samples and conduct an ANOVA. Checking this:

```
C_var <- var(C_dat_trans$length_2_mm, na.rm = TRUE)
P_var <- var(P_dat_trans$length_2_mm, na.rm = TRUE)
SC_var <- var(SC_dat_trans$length_2_mm, na.rm = TRUE)

C_var
```

```
## [1] 0.0208925
```

```
P_var
```

```
## [1] 0.02270146
```

```
SC_var
```

```
## [1] 0.01949329
```

Variance results: Cascades: 0.02 Pool: 0.02 Side Channels: 0.02

When finding their variances, they actually all turned out equal, so we may assume variances are approximately equal across samples and conduct an ANOVA.

c. If you found the data from part b did not display approximately equal variances, you may stop here. If you found the data from part b did display approximately equal variances, conduct a one-way ANOVA to see if there is a significant difference between coastal giant salamander lengths across cascades, pools, and side channels of Mack Creek. If you find evidence of significant differences, perform a post-hoc Tukey's HSD test to determine which are significant from which habitats. Communicate your findings as a figure (with an appropriate caption) and in a sentence, as it might appear in a final report.

```
salamander_ANOVA <- aov(length_2_mm ~ unittype, data = salamander_data_transformed)
summary(salamander_ANOVA)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)    
## unittype        2    3.2  1.5993   76.35 <2e-16 ***
## Residuals   11411  239.0  0.0209                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 220 observations deleted due to missingness
```

```
#Perform a post-hoc Tukey's HSD test since p < 0.001
salamander_Tukey <- TukeyHSD(salamander_ANOVA)
salamander_Tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = length_2_mm ~ unittype, data = salamander_data_transformed)
## 
## $unittype
##             diff         lwr          upr     p adj
## P-C  -0.01310767 -0.02187368 -0.004341656 0.0013319
## SC-C -0.04502145 -0.05359132 -0.036451579 0.0000000
## SC-P -0.03191378 -0.04286293 -0.020964629 0.0000000
```

```
salamander_summary <- salamander_data %>%
  group_by(unittype) %>%
  summarize(mean = mean(length_2_mm, na.rm = TRUE),
  sd = sd(length_2_mm, na.rm = TRUE)) %>%
  ungroup()

salamander_summary
```

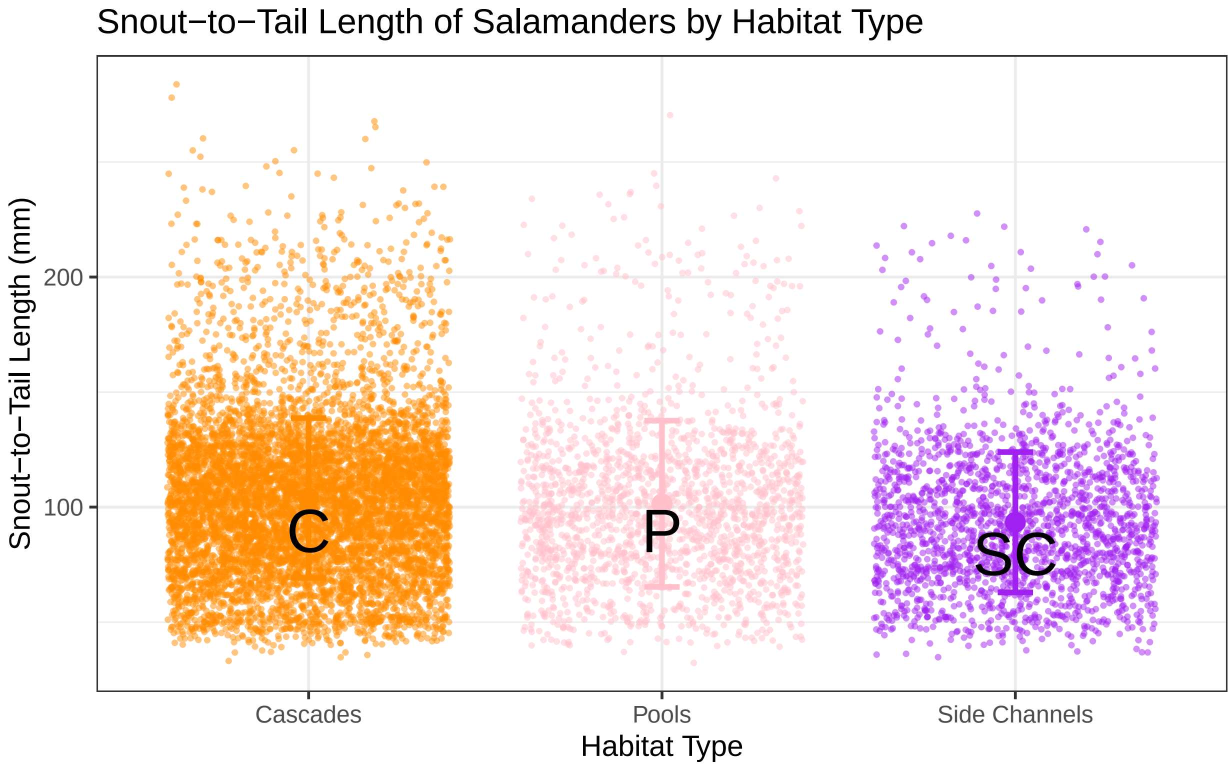
```
## # A tibble: 3 x 3
##   unittype  mean    sd
##   <fct>    <dbl> <dbl>
## 1 C        104.   34.7
## 2 P        101.   36.1
## 3 SC        93.4  30.5
```

```
salamander_summary <- salamander_summary %>%
rename(length_2_mm = mean)

salamander_data$unittype <- factor(salamander_data$unittype,
                                   levels = c("C", "P", "SC"),
                                   labels = c("Cascades", "Pools", "Side Channels"))

salamander_summary$unittype <- factor(salamander_summary$unittype,
                                      levels = c("C", "P", "SC"),
                                      labels = c("Cascades", "Pools", "Side Channels"))

salamander_fig3 <- ggplot() +
  # Add raw data points
  geom_jitter(data = salamander_data, aes(x = unittype, y = length_2_mm,
                                          color = unittype), alpha = 0.5, size = 0.5) +
  # Add summary statistics
  geom_point(data = salamander_summary, aes(x = unittype, y = length_2_mm,
                                            color = unittype), size = 3) +
  # Add error bars
  geom_errorbar(data = salamander_summary, aes(x = unittype,
                                               ymin = length_2_mm - sd,
                                               ymax = length_2_mm + sd,
                                               color = unittype), width = 0.10,
                size = 1) +
  # Add text annotations
  annotate("text", x = "Cascades", y = 90, label = "C", size = 8) +
  annotate("text", x = "Pools", y = 90, label = "P", size = 8) +
  annotate("text", x = "Side Channels", y = 80, label = "SC", size = 8) +
  # Edit colors
  scale_color_manual(values = c("darkorange", "pink", "purple")) +
  # Label axes and add caption
  labs(
    x = "Habitat Type",
    y = "Snout-to-Tail Length (mm)",
    title = "Snout-to-Tail Length of Salamanders by Habitat Type",
    caption ="This plot shows the distribution of snout-to-tail lengths for salamanders across different
    Side Channels). Each point represents an individual salamander, and summary statistics with standard
    bars are overlaid."
  ) +
  # Adjust caption text size
  theme_bw() +
  theme(
    legend.position = "none",
    plot.caption = element_text(size = 8, hjust = 0)
    )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

```
# View figure.
salamander_fig3
```

```
## Warning: Removed 220 rows containing missing values or values outside the scale range
## ('geom_point()').
```



This plot shows the distribution of snout−to−tail lengths for salamanders across different habitat types (Cascades, Pools, Side Channels). Each point represents an individual salamander, and summary statistics with standard deviation error bars are overlaid.

Salamander species displayed significant differences in Snout-to-Tail length as determined by one-way ANOVA (F(2, 11411) = 76.35, p < 0.001). Post-hoc testing by Tukey's HSD revealed that mean Snout-to-Tail length for salamanders in the Cascades habitat (mean = 104 mm, s.d. = 34.7 mm), Pool habitat (mean = 101 mm, s.d. = 36.1 mm), and Side Channels habitat (mean = 93.4 mm, s.d. = 30.5 mm) all differed significantly from each other.

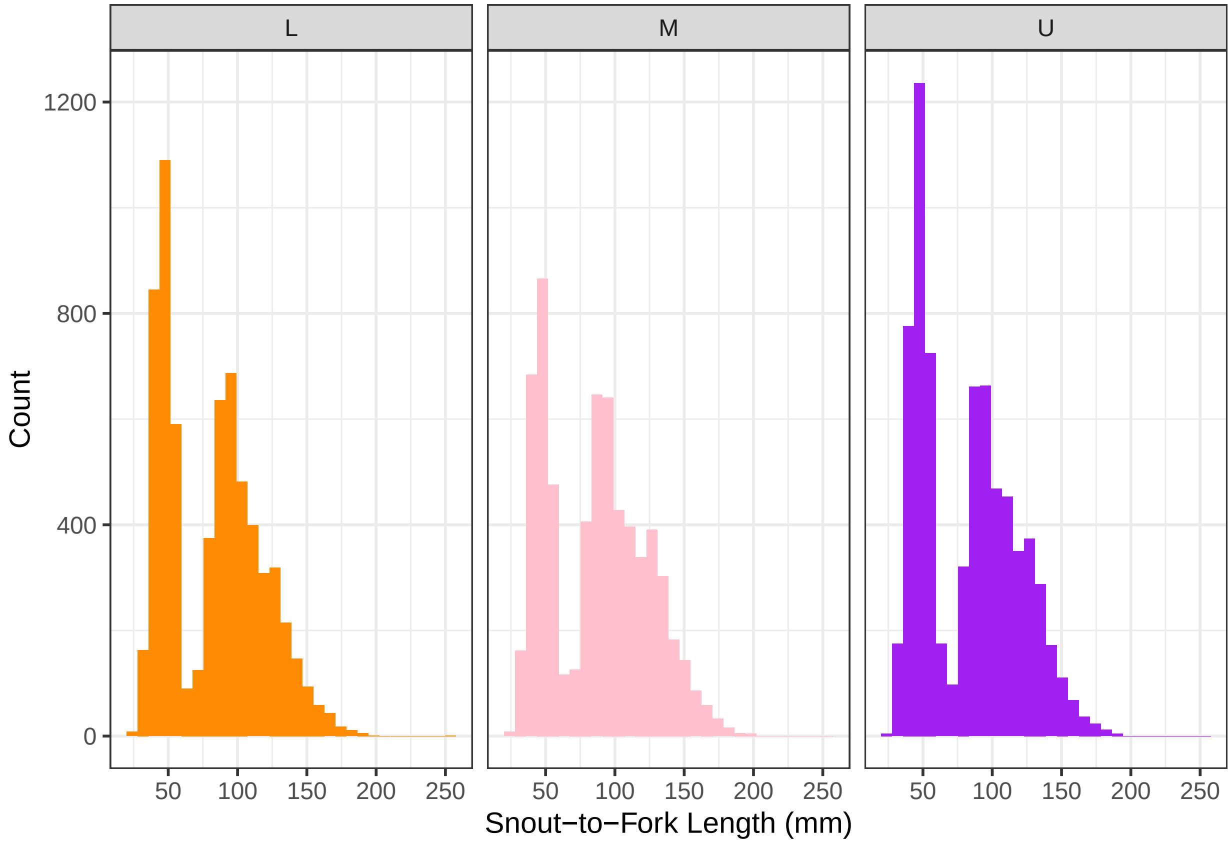
#(2) Cutthroat Trout-

a. Filter the dataset only for cutthroat trout and create a figure that helps you to evaluate whether their snout-to-fork length (length_1_mm) by reach (reach) is normally distributed. You may also calculate skew and kurtosis values to help with your decision-making. Are these data normally distributed? Why or why not? If they are not, apply a log-transform (log10()) to the length data and re-evaluate if the data now meets the criteria to be considered normally- distributed.

```
trout_data <- filter(vertebrates_data,
                              species == "Cutthroat trout")

trout_fig1 <- ggplot(trout_data, aes(x = length_1_mm, fill = reach)) +
  geom_histogram() +
  scale_fill_manual(values = c("darkorange", "pink", "purple")) +
  labs(x = "Snout-to-Fork Length (mm)", y = "Count") +
  facet_grid(.~reach) +
  theme_bw() +
  theme(legend.position = "none")

trout_fig1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```
# Separate data by unittype
L_dat <- trout_data %>%
filter(reach == "L")

M_dat <- trout_data %>%
filter(reach == "M")
```

```
U_dat <- trout_data %>%
filter(reach == "U")

# Calculate skew.
skewness(L_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 0.4042481
```

```
skewness(M_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 0.2626566
```

```
skewness(U_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 0.3832519
```

```
# Calculate kurtosis.
kurtosis(L_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 2.286412
```

```
kurtosis(M_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 2.158395
```

```
kurtosis(U_dat$length_1_mm, na.rm = TRUE)
```

```
## [1] 2.104812
```

Skew results: L: 0.40 M: 0.26 U: 0.38

Kurtosis results: L: 2.28 M: 2.16 U: 2.10

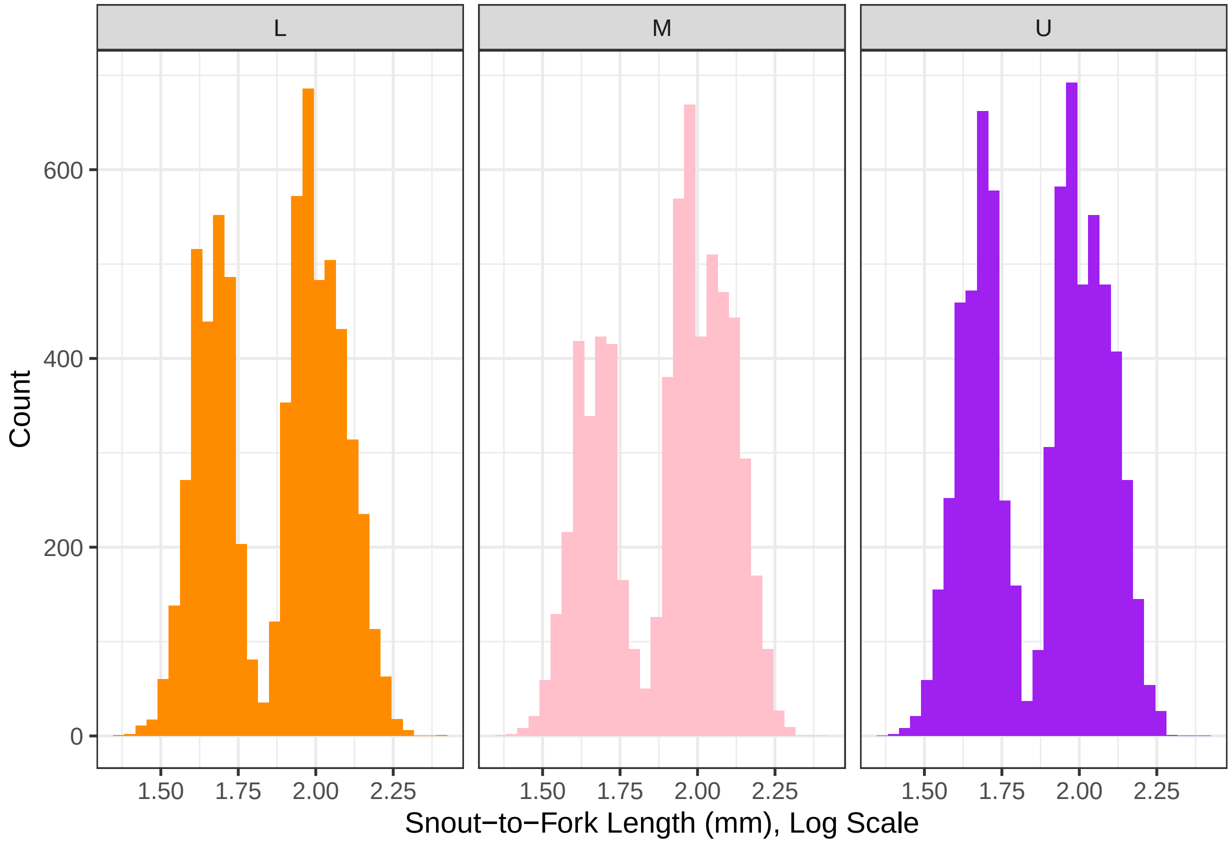
Visually, the three data sets appear bi-modal. From the skew results, all three have values slightly above 0 indicating slight positive skewing. All three have kurtosis values less than 3, indicating that they may be platykurtic. Overall, they do not seem normally distributed.

```
#log-transformation of data
trout_data_transformed <- mutate(trout_data, length_1_mm = log10(length_1_mm))
```

```
trout_fig2 <- ggplot(trout_data_transformed, aes(x = length_1_mm, fill = reach)) +
  geom_histogram() +
  scale_fill_manual(values = c("darkorange", "pink", "purple")) +
  labs(x = "Snout-to-Fork Length (mm), Log Scale", y = "Count") +
  facet_grid(.~reach) +
  theme_bw() +
  theme(legend.position = "none")

trout_fig2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 5 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



```r
# Separate data by unittype
L_dat_trans <- trout_data_transformed %>%
filter(reach == "L")

M_dat_trans <- trout_data_transformed %>%
filter(reach == "M")

U_dat_trans <- trout_data_transformed %>%
filter(reach == "U")

# Calculate skew.
skewness(L_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] -0.1397749
```

```r
skewness(M_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] -0.3168654
```

```r
skewness(U_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] -0.1169573
```

```r
# Calculate kurtosis.
kurtosis(L_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] 1.73097
```

```r
kurtosis(M_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] 1.882288
```

```r
kurtosis(U_dat_trans$length_1_mm, na.rm = TRUE)
```

```
## [1] 1.697403
```

After applying the log 10 transformation, the data is still bi-modal as apparent from a visual inspection. While skewness was brought closer to 0, kurtosis values are still less than 3, and so these distributions are considered platykurtic. The data are still not normally distributed.

b. If you found the data from part a were not normally distributed and a log-transform did not change this finding, you may stop here. If you found the data from part a were normally distributed, conduct a Bartlett's test for equal variance to determine if these data also satisfy the need for homogeneity of variances across groups. Do these data have approximately similar variances? Why or why not? Remember, the data may not pass the Bartlett's test, but if they adhere to the rule of thumb mentioned above, you may proceed with a one-way ANOVA.

Since the data was not normally distributed, I did not conduct a Bartlett's test for equal variance.

c. If you found the data from part b did not display equal variances, you may stop here. If you found the data from part b did display equal variances, conduct a one-way ANOVA to see if there is a significant difference between cutthroat trout lengths across lower, middle, and upper reaches of Mack Creek. If you find evidence of significant differences, perform a post- hoc Tukey's HSD test to determine which are significant from which reaches. Communicate your findings as a figure (with an appropriate caption) and a sentence, as it might appear in a final report.

Since the data was not normally distributed, I did not conduct a one-way ANOVA.