Homework 8: Squirrel Sizes

Comparing a Numerical Variable Across More Than Two Groups

Researchers set out to explore how squirrel size is impacted by latitude. Specifically, focusing on the lengths (in mm) observed in four distinct California locations: Big Bear, Hemet, Susanville, and Loop Hill. As these agile creatures navigate different environments, their lengths become a fascinating avenue for investigation, offering insights into potential ecological variations and adaptation strategies. Understanding the nuances in squirrel lengths across diverse locations not only contributes to our knowledge of local fauna but also has broader implications for ecosystem dynamics and the impact of environmental factors on wildlife morphology.

```
head(squirrels)
```

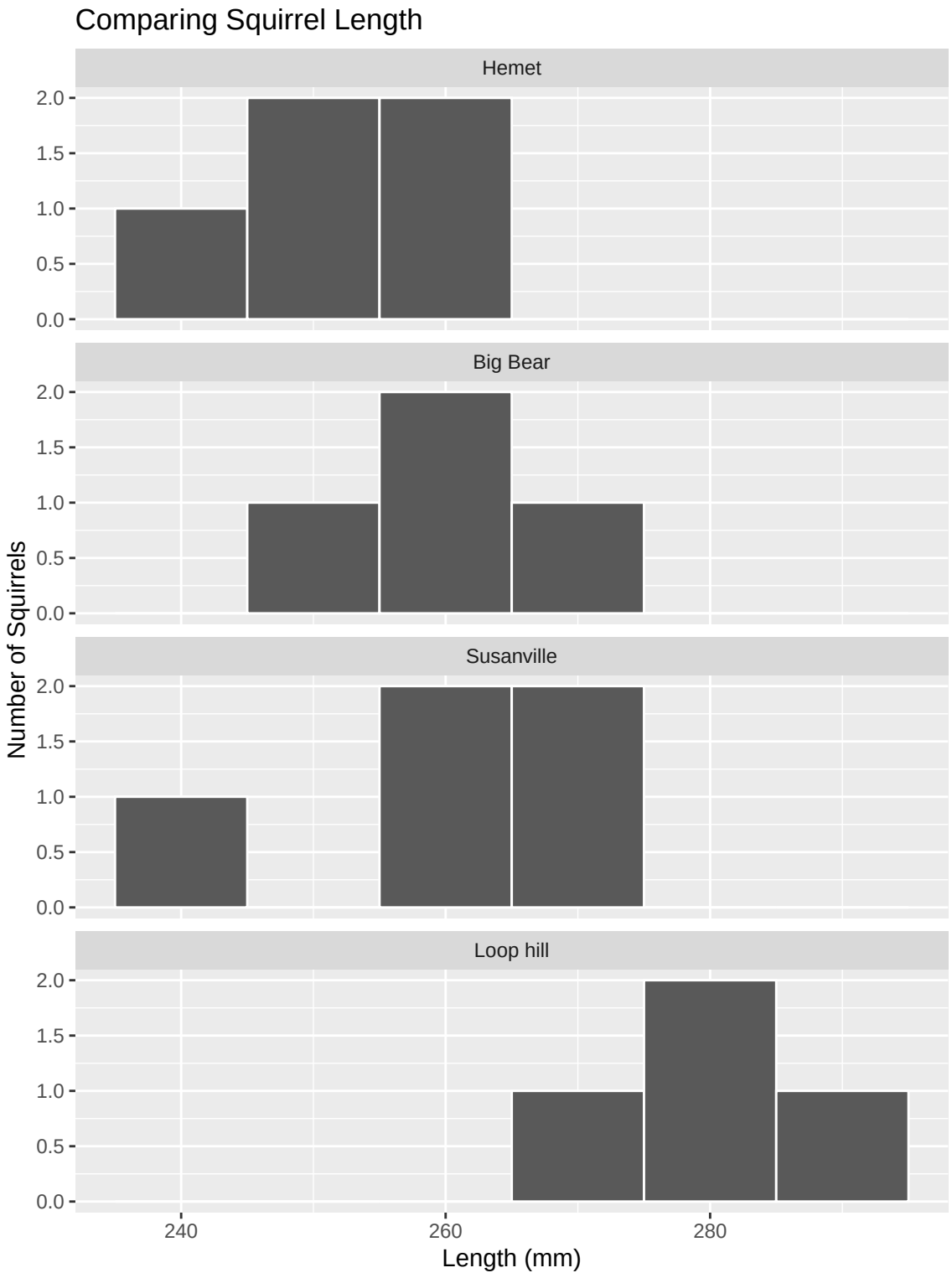
```
# A tibble: 6 x 2
  Location  Length
  <fct>      <dbl>
1 Big Bear     249
2 Hemet        248
3 Hemet        242
4 Big Bear     256
5 Hemet        251
6 Loop hill    291
```

Research question Does the mean squirrel differ between the four selected California locations?

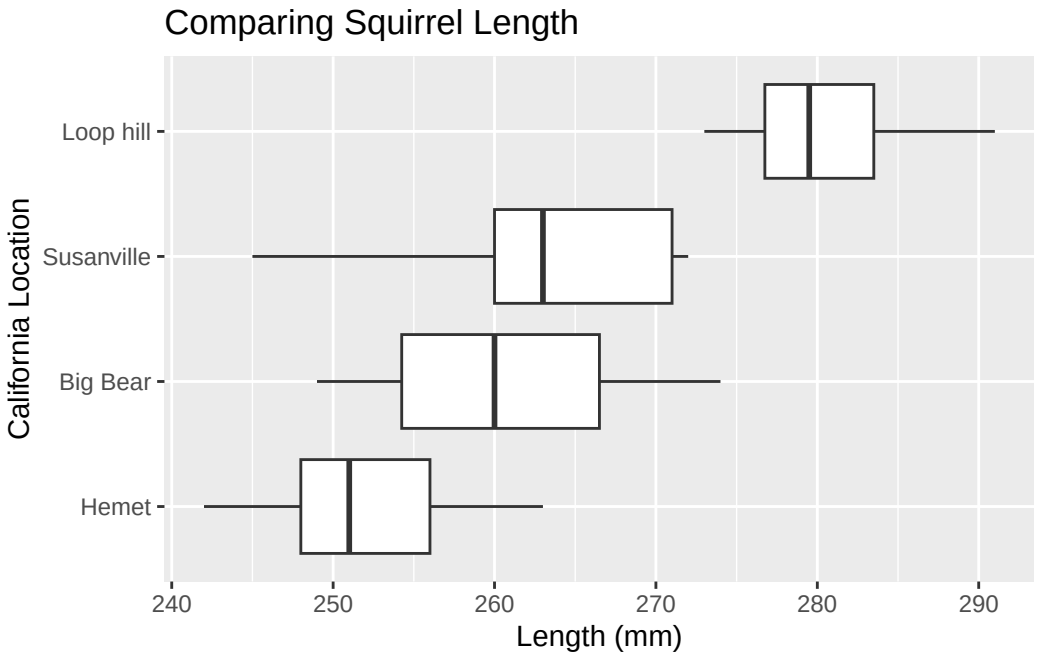
```
favstats(Length ~ Location, data = squirrels)
```

```
    Location min     Q1 median    Q3 max   mean        sd n missing
1      Hemet 242 248.00  251.0 256.0 263 252.00  7.968689 5       0
2   Big Bear 249 254.25  260.0 266.5 274 260.75 10.750969 4       0
3 Susanville 245 260.00  263.0 271.0 272 262.20 10.894953 5       0
4  Loop hill 273 276.75  279.5 283.5 291 280.75  7.588368 4       0
```

```r
ggplot(data = squirrels,
       mapping = aes(x = Length)) +
  geom_histogram(binwidth = 10, color = "white") +
  facet_wrap(~Location, ncol = 1) +
  labs(title = "Comparing Squirrel Length",
       x = "Length (mm)",
       y = "Number of Squirrels")
```


Comparing Squirrel Length
Hemet
Big Bear
Susanville
Loop hill
Number of Squirrels
0.0
0.5
1.0
1.5
2.0
240
260
280
Length (mm)

```
ggplot(data = squirrels,
       mapping = aes(x = Length,
                     y = Location)
       ) +
  geom_boxplot() +
  labs(title = "Comparing Squirrel Length",
       x = "Length (mm)",
       y = "California Location")
```



1. Identify the variables (levels/units) and data type.

- Explanatory:

- Response:

2. Is this an experimental study or observational study? Explain.

3. Discuss the visual comparison of the the four California locations. Which location appears to have a higher average squirrel length? The smallest? Do you think you will find evidence of a discernible difference in length between locations?

4. Identify the observed mean, standard deviation, and sample size of each Location in the study. make sure to assign appropriate symbols.

5. State the parameters in words with symbols (hint: there should be 4 of them).

7. Write the null and alternative hypotheses using the appropriate mathematical symbols.

8. Check the assumptions necessary to conduct an ANOVA F-test.

9. The following R code conducts an ANOVA F-test on the data. I have intentionally removed the degrees of freedom, mean square values, and statistic. Show how to determine / calculate these values using the information provided. Make sure to indicate what each value represents (e.g., df1, MSE, F-statistic, etc.).

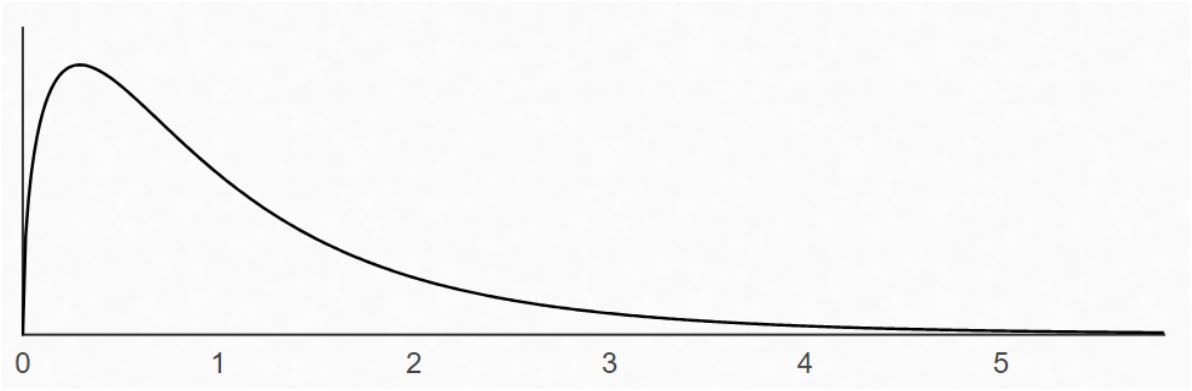
```
squirrel_model <- aov(Length ~ Location,
    data = squirrels
    )

squirrel_model |>
  tidy()
```

```
# A tibble: 2 x 6
  term      df    sumsq meansq statistic  p.value
  <chr>     <chr> <dbl> <chr>  <chr>        <dbl>
1 Location  A     1888. C      E           0.00401
2 Residuals B     1248. D      <NA>       NA
```

- A:
- B:
- C:
- D:
- E

10. Using the F-distribution below, show how you would calculate/estimate the p-value.

11. Using the output above and an $\alpha = 0.01$ significance level, write a conclusion in the context of the problem.



12. What type of error could have been made? Explain.

    Type I Type II

12. The code below shows the pairwise comparisons with *no multiplicity* adjustment. At an $\alpha = 0.01$ significance level, which locations indicate a discernible difference in mean squirrel length?

```
emmeans(squirrel_model,
        specs = ~ Location) |>
  pairs(adjust = "none")
```

```
contrast               estimate   SE df t.ratio p.value
Hemet - Big Bear          -8.75 6.33 14 -1.381  0.1888
Hemet - Susanville       -10.20 5.97 14 -1.708  0.1097
Hemet - Loop hill        -28.75 6.33 14 -4.539  0.0005
Big Bear - Susanville     -1.45 6.33 14 -0.229  0.8222
Big Bear - Loop hill     -20.00 6.68 14 -2.995  0.0096
Susanville - Loop hill   -18.55 6.33 14 -2.928  0.0110
```

13. The code below shows the pairwise comparisons with a *tukey* multiplicity adjustment. At an $\alpha = 0.01$ significance level, which locations indicate a discernible difference in mean squirrel length?

```
emmeans(squirrel_model,
        specs = ~ Location) |>
  pairs(adjust = "tukey")
```

```
 contrast               estimate   SE df t.ratio p.value
 Hemet - Big Bear          -8.75 6.33 14 -1.381  0.5303
 Hemet - Susanville       -10.20 5.97 14 -1.708  0.3560
 Hemet - Loop hill        -28.75 6.33 14 -4.539  0.0023
 Big Bear - Susanville     -1.45 6.33 14 -0.229  0.9956
 Big Bear - Loop hill     -20.00 6.68 14 -2.995  0.0425
 Susanville - Loop hill   -18.55 6.33 14 -2.928  0.0480

P value adjustment: tukey method for comparing a family of 4 estimates
```

14. Compare your decisions/results in the previous two questions. How does including a multiplicity adjustment (e.g., Tukey's) impact the results?