Homework 4

Sonya Eason

```
if (!require(pacman)) install.packages("pacman")
```

```
Loading required package: pacman
```

```
pacman::p_load(knitr, broom, tidyverse)
```

```
library(tidyverse)
library(knitr)
library(broom)
```

```
crab <- read_csv("data/crab.csv")
```

```
Rows: 173 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (5): Color, Spine, Width, Satellite, Weight

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ambiguity <- read_csv("data/ambiguity.csv")
```

```
Rows: 870 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): name
dbl (10): ambiguity, distID, ideology, totalIssuePages, democrat, mismatch, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exercise 1

a) The response is the mean value of fisher caught per week by each visitor.

b) The possible values are 0, 1, 2, 3, up to the maximum number fish there are in the state wildlife park.

c) Lambda represents the mean number of fish caught per wekk.

d) A zero-inflated model could be considered here since there are likely a lot of zeroes representing people who caught zero fish during their stay. These zeroes can be separated into two subgroups the people who never fish, i.e. the true zeroes., and those who just didn't fish on their trips this time.

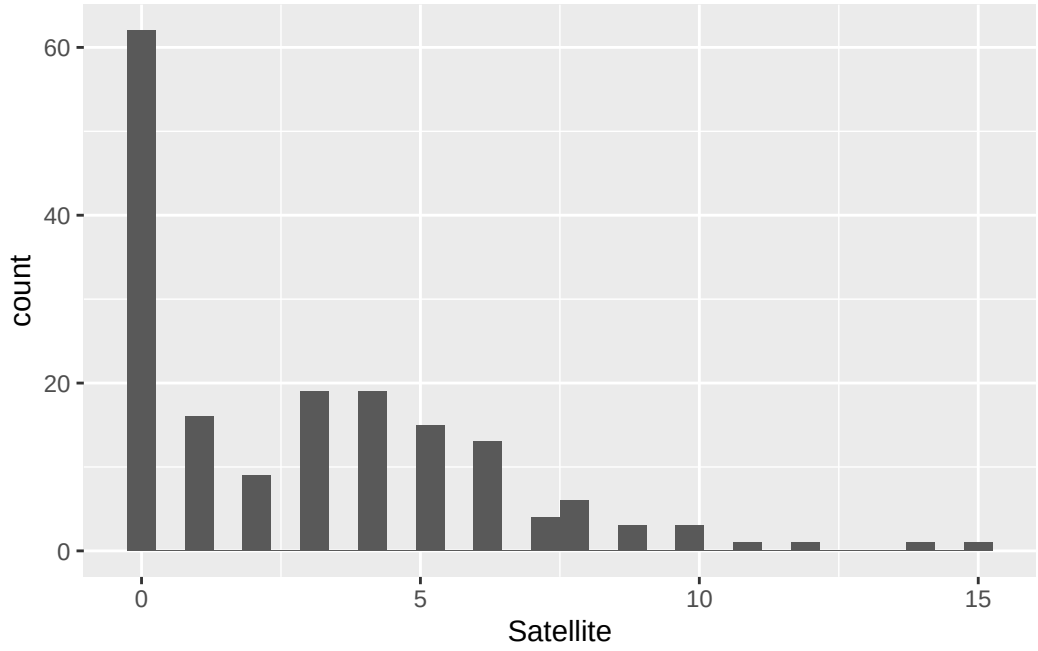
Exercise 2

a)

```
crab <- crab |>
  mutate(
    Color = factor(Color),
    Spine = factor(Spine)
  )
```

```
ggplot(crab, aes(x = Satellite)) +
  geom_histogram()
```

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



Is there preliminary evidence the number of satellites could be modeled as a Poisson response? Briefly explain

In a Poisson response, we do expe…

b)

```
sat_model <- glm(Satellite ~ Width + Weight + Spine, family = poisson, data = crab)


sat_model |>
  tidy(conf.int = T) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.062	0.928	-1.144	0.253	-2.875	0.763
Width	0.039	0.048	0.816	0.415	-0.055	0.132
Weight	0.000	0.000	2.771	0.006	0.000	0.001
Spine2	-0.214	0.211	-1.017	0.309	-0.644	0.185
Spine3	-0.049	0.108	-0.458	0.647	-0.257	0.165

c) When a female crab has one worn or broken spine, the number of satellites is expected to change by a multiplicative factor of 0.8073484compared to when the female crab has

two spines that are both in good condition, holding all else constant. When a female crab has two worn or broken spines, the number of satellites is expected to change by a multiplicative factor of 0.9521811compared to when the female crab has two spines that are both in good condition, holding all else constant.

Exercise 3

```
crab |>
  group_by(Spine) |>
  summarize(mean = mean(Satellite),
            var = var(Satellite))
```

```
# A tibble: 3 x 3
  Spine  mean   var
  <fct> <dbl> <dbl>
1 1      3.65 11.5
2 2      2     5.57
3 3      2.81  9.82
```

A quasi-Poisson regression is suitable because there is evidence of overdispersion seen by variances that are larger than means at each level. In a typical Poisson model, we'd expect mean = variance, so we'd use quasi-Poison to account for the fact that that does not hold here.

b)

```
sat_model_2 <- glm(Satellite ~ Width + Weight + Spine, family = quasipoisson, data = crab)

sat_model_2 |>
tidy(conf.int = TRUE) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.062	1.652	-0.643	0.521	-4.281	2.195
Width	0.039	0.085	0.458	0.647	-0.130	0.203
Weight	0.000	0.000	1.556	0.122	0.000	0.001
Spine2	-0.214	0.375	-0.571	0.568	-1.006	0.480
Spine3	-0.049	0.192	-0.257	0.797	-0.416	0.337

c)

```r
se <- tidy(sat_model)$std.error

se_overdis <- tidy(sat_model_2)$std.error

dispersion_param <- (se_overdis/se)^2

dispersion_param
```

```
[1] 3.169448 3.169448 3.169448 3.169448 3.169448
```

The estimated dispersion parameter is 3.169448.

d) The estimated coefficients do not change at all between this model and the previous one while the standard errors increase by a multiplicative factor of 23.7943463

Exercise 4