Class Exercise - Answer Key

Guy F. Sutton

Centre for Biological Control, Rhodes University

The case-study - tick abundance on grasses and between seasons

For this exercise, you were analysing a dataset containing the recorded abundances of a tick on different grass species and from surveys performed across seasons.
- You have collected 30 stems from each of 3 different grass species, namely: grass_sp1, grass_sp2 and grass_sp3, in both summer and winter.

```
## # A tibble: 6 x 3
##   season grass_sp  tick_abundance
##   <chr>  <chr>              <dbl>
## 1 Summer grass_sp1              3
## 2 Summer grass_sp1              9
## 3 Summer grass_sp1              5
## 4 Summer grass_sp1              1
## 5 Summer grass_sp1              6
## 6 Summer grass_sp1              8
```

Ultimately, you wanted to know whether:
(1) The number of ticks recorded differs between the different grass species?
(2) The number of ticks recorded differs between seasons?, and
(3) Whether season impacted tick abundance differently for the different grass species?

Questions

(1) What is the response variable?

The response variable is `tick_abudance`. Remember, the response variable is the focus of the study.

(2) What are the predictor variables?

The predictor variables are `grass_sp` and `season`. Remember, a predictor variable is something that we measure or record to explain our response variable. Typically, we want to know how our response variable changes (`tick_abundance`) changes with a change in the predictor variables (`grass_sp` and `season`).

(3) Do we need to specify an interaction term to answer our research question? Explain.

Yes. To answer question (3) above, we needed an interaction term. Remember, an interaction term allows the effect of one of our predictor variables to vary between levels of another predictor variable. For example, the interaction term would allow us to test whether the effect of `season` impacted ticks differently for each of the different `grass_sp`. Without the interaction term, we are fitting a model where we assume that tick abundances change with `season` with the same magnitude across the different `grass_sp`.

(4) Which GLM provided the best fit (e.g. Gaussian, Poisson, Negative Binomial, Binomial, Multivariate)? Explain, and make use of at least three figures to support your argument.

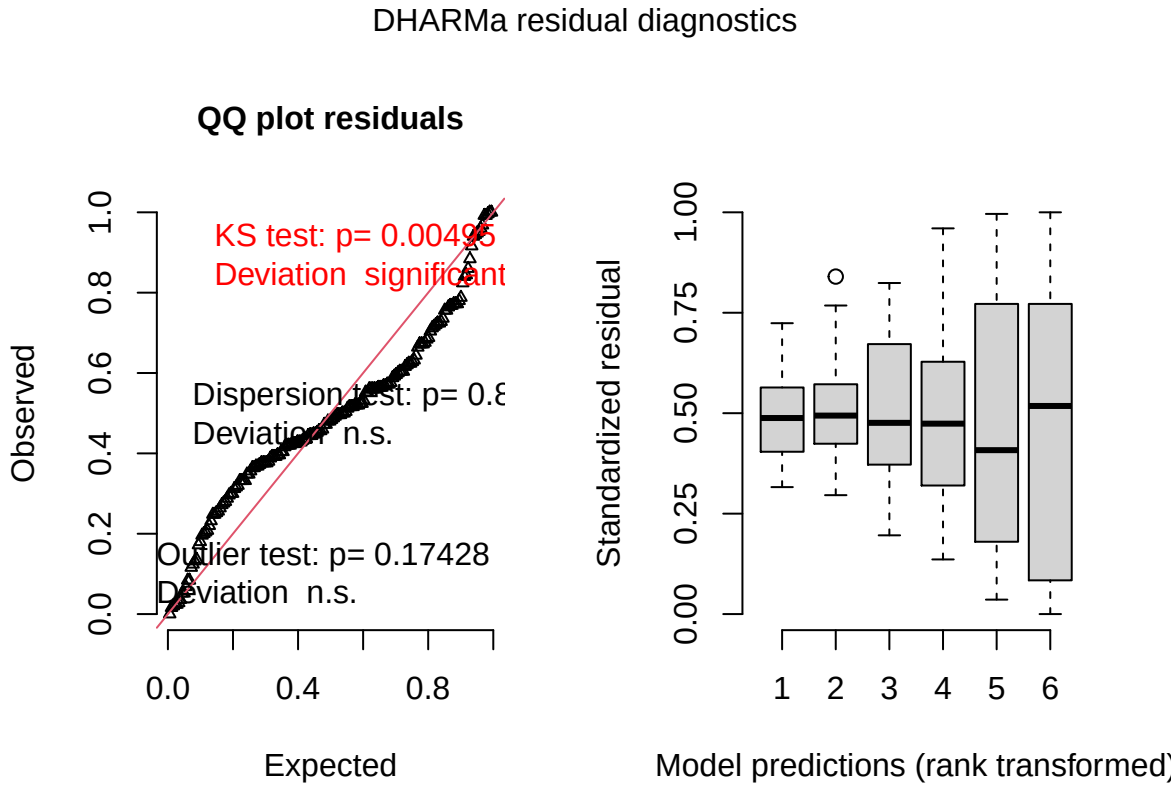
The GLM with the best fit was the negative binomial GLM. You cannot fit a multivariate or binomial GLM to the tick abundance dataset, because you only have the abundance of a single species (`multivariate`), and because the response variable is not a proportion or percentage (`binomial`).

Because the response variable (`tick_abundance`) is a count, you definitely needed to fit a Poisson and Negative binomial GLM, because these two statistical distributions are specifically designed for count data. However, you were asked to provide three figures, so ultimately, I would have liked to have seen you fit a Gaussian GLM (the same as an ANOVA) to the data, just so that you could see how bad a choice the ANOVA would have been.

4.1. Gaussian GLM (ANOVA)

```
# Fit model
mod_gaussian <- glm(tick_abundance ~ season * grass_sp,
                    data = data)
```

```
# Check residuals
DHARMa::simulateResiduals(mod_gaussian, plot = TRUE)
```




```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.48 0.724 0.568 0.332 0.564 0.664 0.564 0.496 0.42 0.432 0.56 0.376 0.404 0
```

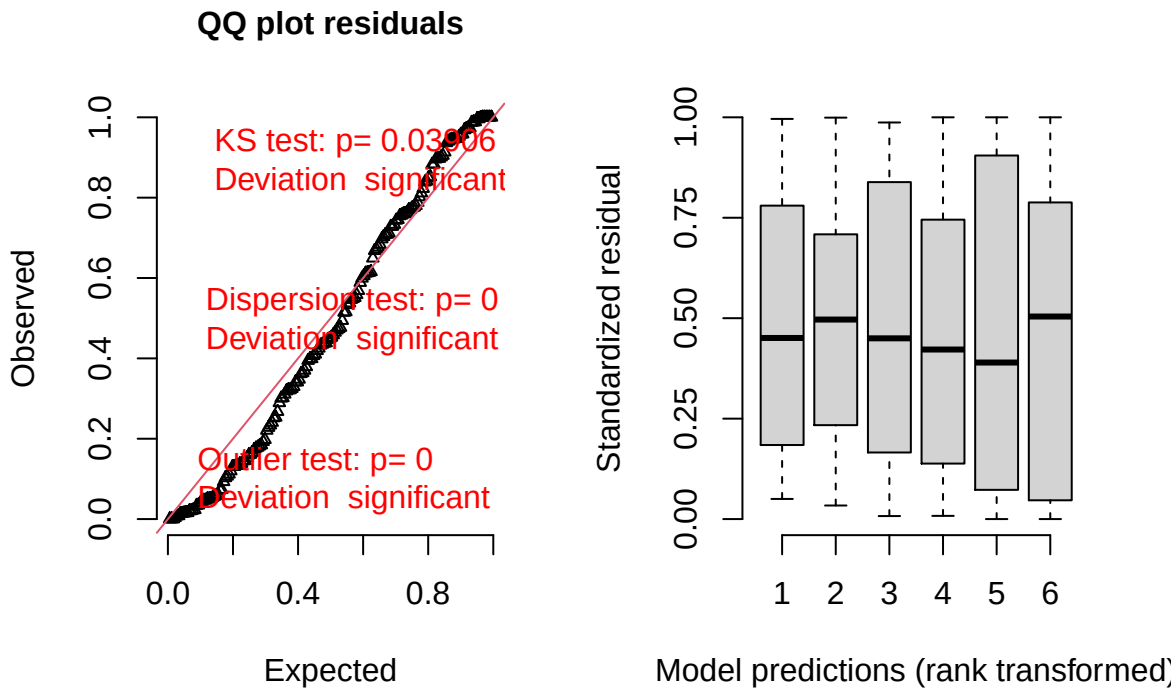
The Gaussian GLM was a poor fit. The QQplot shows significant departures from expectations of normality. The KS test is significant and points are far awar from the red 1:1 line. There is strong evidence for unequal variances looking at the right-hand side plot - some boxplots have a much greater range than others.

4.2. Poisson GLM

As above, because we are analysing counts, the Poisson GLM is a natural starting point.

```
# Fit model
mod_poisson <- glm(tick_abundance ~ season * grass_sp,
                   data = data,
                   family = poisson(link = "log"))
```

```
# Check residuals
DHARMa::simulateResiduals(mod_poisson, plot = TRUE)
```

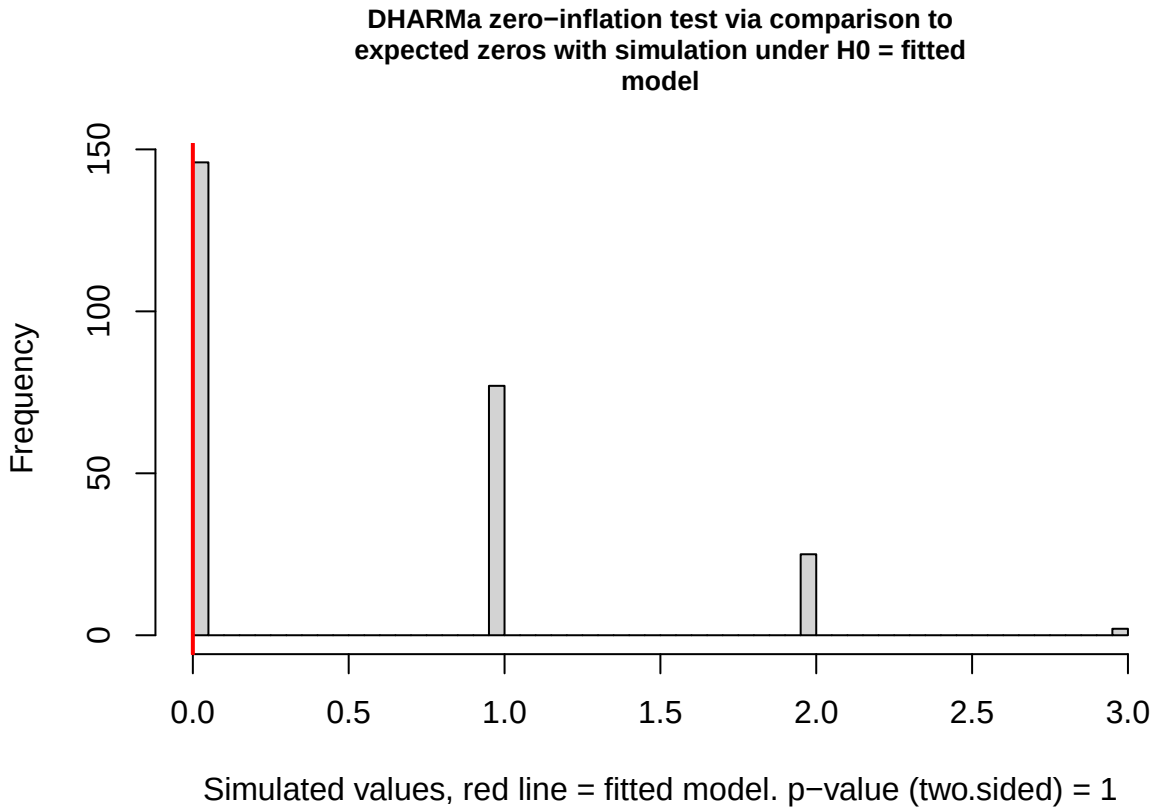



```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.398653 0.996 0.7546701 0.05852327 0.8113905 0.9471283 0.7801621 0.5511754 0
```

The Poisson GLM was a poor fit, but much better than the Gaussian GLM. The QQplot shows significant departures from expectations of normality. The KS test is significant and points are far awar from the red 1:1 line. There is no evidence for unequal variances looking at the right-hand side plot - all boxplots have a similar range of y-values.

When we fit count models, we always need to check for zero-inflation (too much zeroes than expected under the statistical distribution we are fitting, here: *Poisson*).

```
# Test zero inflation
DHARMa::testZeroInflation(mod_poisson)
```




```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0, p-value = 1
## alternative hypothesis: two.sided
```

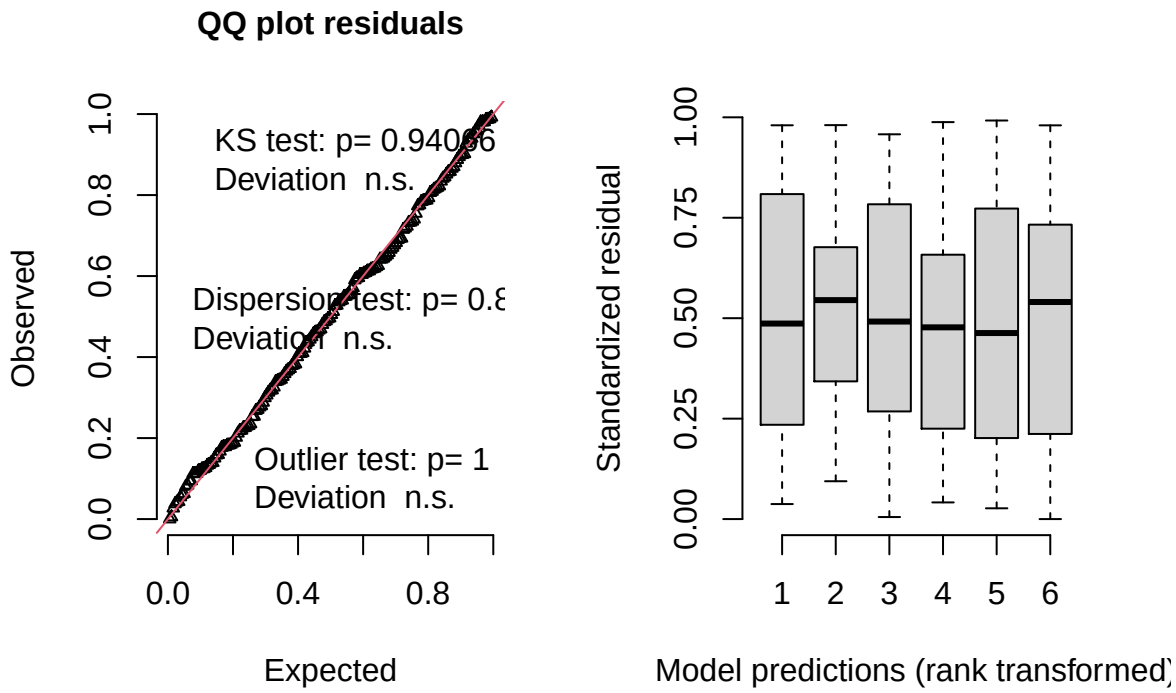
No evidence for zero inflation ($P > 0.05$).

4.3. Negative binomial GLM

The Poisson GLM above improved the model fit significantly versus the Gaussian GLM/ANOVA we initially ran. This shows us that fitting a statistical distribution designed for counts/abundance data was a good option. Because the Poisson GLM wasn't a great fit, our next stop is the Negative Binomial GLM. The Negative Binomial (variance > mean) expects more variance than the Poisson (variance = mean).

```
# Fit model
mod_nb <- glmmTMB::glmmTMB(tick_abundance ~ season * grass_sp,
                           data = data,
                           family = "nbinom2")
```

```
# Check residuals
DHARMa::simulateResiduals(mod_nb, plot = TRUE)
```

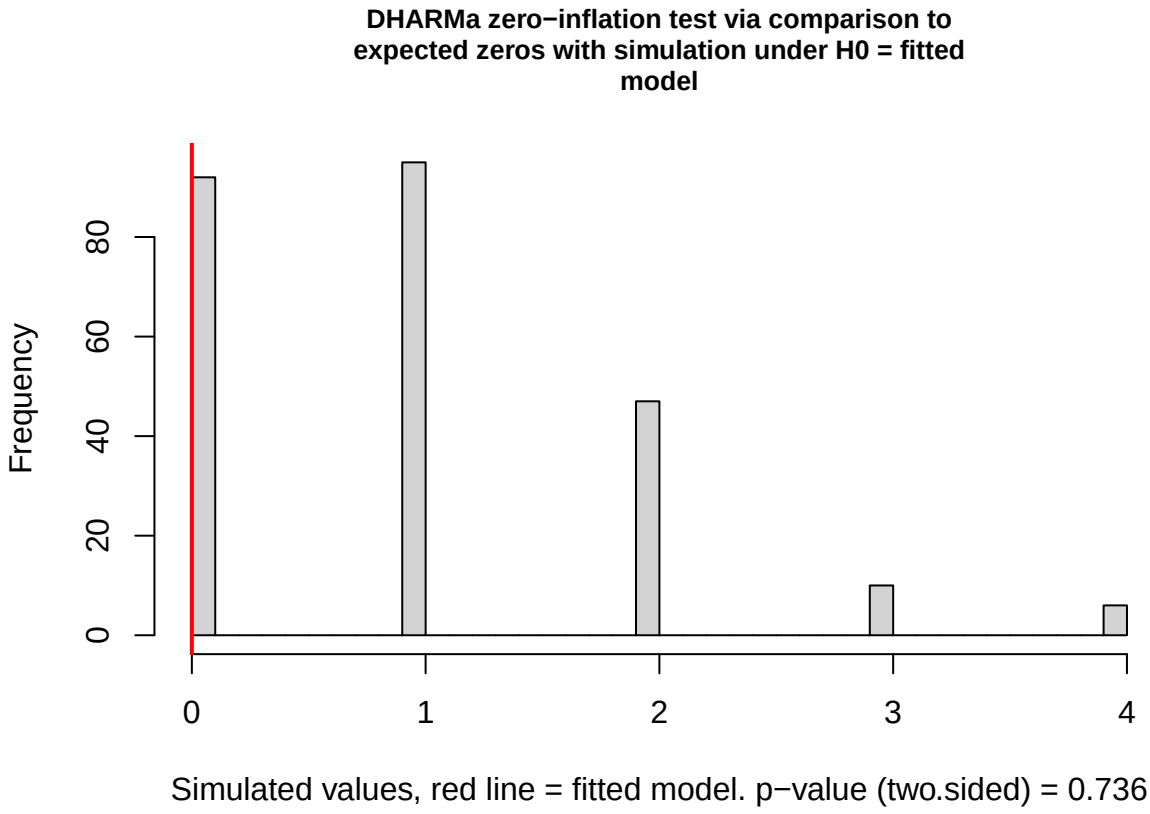



```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.3129105 0.9801323 0.6824131 0.06074607 0.7863508 0.9596677 0.8513814 0.5123
```

The Negative Binomial GLM was a good fit. The QQplot shows no evidence for a departure from expectations of normality. The KS test is non-significant and points fall approximately along the red 1:1 line. There is no evidence for unequal variances looking at the right-hand side plot - all boxplots have a similar range of y-values.

When we fit count models, we always need to check for zero-inflation (too much zeroes than expected under the statistical distribution we are fitting, here: *Negative binomial*).

```
# Test zero inflation
DHARMa::testZeroInflation(mod_nb)
```




```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 0, p-value = 0.736
## alternative hypothesis: two.sided
```

No evidence for zero inflation ($P > 0.05$). We are now happy that we have a good fitting model.

(5) Should you use type I, II or III sum-of-squares to test for parameter significance? Explain.

We need to use type III (3) sum-of-squares (SOS) when assessing the statistical significance of our predictor variables.

```
# NB GLM - type III sum-of-squares
car::Anova(mod_nb,
           type = "III",
           test = "Chisq")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: tick_abundance
```

```
##                   Chisq Df Pr(>Chisq)
## (Intercept)     185.557  1  < 2.2e-16 ***
## season           34.853  1  3.556e-09 ***
## grass_sp        117.838  2  < 2.2e-16 ***
## season:grass_sp  12.088  2   0.002372 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Type I (1) SOS: Only appropriate when you have a single (1) predictor variable because type I SOS calculates p-values based on the order in which you specify the predictors. E.g. If you type `tick_abundance ~ season * grass_sp` and `tick_abundance ~ grass_sp * season`, you will get different p-values...
- Type II (2) SOS: Appropriate for when you have 2 or more variables AND no interaction term. They estimate the contribution of each predictor variable AFTER accounting for the variation explained by the other variable.
- Type III (3) SOS: Appropriate for when you have 2 or more predictor variables AND an interaction term. They estimate the contribution of each predictor variable AFTER accounting for the variation explained by the other variable AND the interaction term.

*** Note: When you fit type III SOS, and the interaction term is not significant ($P > 0.05$), you should recalculate the p-values for the two predictor variables using type II SOS (type II is more powerful than type III).

(6) Write a short summary paragraph (6-12 lines) summarizing the results of your statistical analysis

Every person has their own writing style. The example below is just one way to do this, however, there are definitely key components that you need to include in your results section that I will highlight.

You always need to report the following things:
- Test statistic(s)
- Degrees of freedom
- P-values
- Summary statistics (mean/median/standard deviation/confidence intervals)
- Explain results using statistics AND using human words

Let's build a paragraph step-by-step:

Firstly, let's report the main finding of our study in terms of our statistical analysis.

- There was a statistically significant difference in tick abundances recorded on the three grass species ($\chi^2 = 117.84$, d.f. $= 2$, $P < 0.001$) and between surveys performed in different seasons ($\chi^2 = 34.853$, d.f. $= 1$, $P < 0.001$).

I usually like to be explicit about what statistical test is being reported whenever I report test statistics, but this is not very common in ecology journals (more common in statistical journals). In the methods, we would have mentioned that to calculate statistical significance of our predictor variables, we would use Likelihood Ratio Tests (LRT) with type III sum-of-squares, so let's just use the acronym.

- There was a statistically significant difference in tick abundances recorded on the three grass species (LRT: $\chi^2 = 117.84$, d.f. $= 2$, $P < 0.001$) and between surveys performed in different seasons (LRT: $\chi^2 = 34.853$, d.f. $= 1$, $P < 0.001$).

Now let's explain our results so that a human can read and understand our results. This provides a lot more context for HOW our response variable changed, not just telling the reader that it changed. Usually, we would want to report mean ± sd/se changes between grass species or seasons, or even % changes in abundance.

```
# Calculate mean +- standard error change in tick abundance across grass_sp.
data %>%
  dplyr::group_by(grass_sp) %>%
  dplyr::summarise(
    tick_mean = mean(tick_abundance),
    tick_sd = sd(tick_abundance),
    n = n(),
    tick_se = tick_sd/sqrt(n)
  )
```

```
## # A tibble: 3 x 5
##   grass_sp  tick_mean tick_sd     n tick_se
## * <chr>         <dbl>   <dbl> <int>   <dbl>
## 1 grass_sp1      6.25    3.37    60   0.435
## 2 grass_sp2     16.6    10.6     60   1.36
## 3 grass_sp3     33.6    22.1     60   2.85
```

- Averaged across seasons, tick abundance was 73% higher on grass_spB (16.6 ± 1.36) than grass_spA (6.25 ± 0.44), but ticks were less abundant on both grass_spA (81% lower) and grass_spB (51% lower) than grass_sp3 (33.6 ± 2.85 ticks per grass stem).

```
# Calculate mean +- standard error change in tick abundance across seasons.
data %>%
  dplyr::group_by(season) %>%
  dplyr::summarise(
    tick_mean = mean(tick_abundance),
    tick_sd = sd(tick_abundance),
    n = n(),
    tick_se = tick_sd/sqrt(n)
  )
```

```
## # A tibble: 2 x 5
##   season tick_mean tick_sd     n tick_se
## * <chr>      <dbl>   <dbl> <int>   <dbl>
## 1 Summer      9.14    5.87    90   0.619
## 2 Winter     28.5    20.9     90   2.20
```

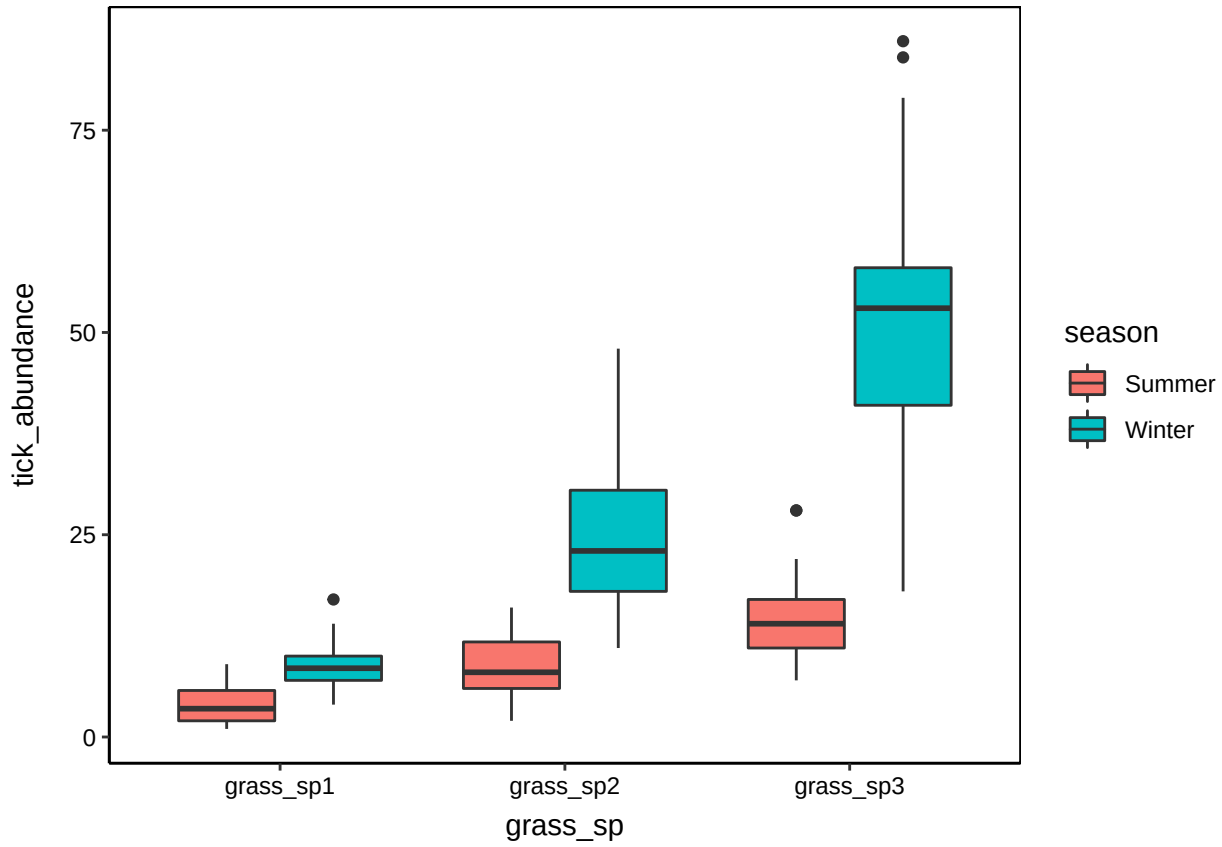
- Averaged across the three grass species, tick abundances were lower in summer (9.14 ± 0.62) than winter (28.5 ± 2.20).

Remember, we have another key result to present though: our significant interaction term.

- However, the seasonal effect on tick abundances differed across the three grass species (LRT: $\chi^2 = 12.09$, d.f. $= 2$, $P = 0.002$).

And, explain it like a human.

```
# Using a plot makes this easier to explain
data %>%
  ggplot(data = ., aes(x = grass_sp,
                       y = tick_abundance,
                       fill = season)) +
  geom_boxplot() +
  theme(legend.position = "right")
```




```
# Calculate mean +- standard error change in tick abundance across seasons and grasses.
data %>%
  dplyr::group_by(grass_sp, season) %>%
  dplyr::summarise(
    tick_mean = mean(tick_abundance),
    tick_sd = sd(tick_abundance),
    n = n(),
    tick_se = tick_sd/sqrt(n)
  )
```

```
## `summarise()` has grouped output by 'grass_sp'. You can override using the `.groups` argument.

## # A tibble: 6 x 6
## # Groups:   grass_sp [3]
##   grass_sp  season tick_mean tick_sd     n tick_se
##   <chr>     <chr>      <dbl>   <dbl> <int>   <dbl>
## 1 grass_sp1 Summer         4    2.18    30   0.398
```

```
## 2 grass_sp1 Winter      8.5     2.81    30   0.514
## 3 grass_sp2 Summer      8.73    3.73    30   0.681
## 4 grass_sp2 Winter     24.6     9.11    30   1.66
## 5 grass_sp3 Summer     14.7     5.27    30   0.962
## 6 grass_sp3 Winter     52.5    15.0     30   2.75
```

- The significant interaction term was driven by tick abundances being 73% higher in winter (52.5 ± 2.75) than summer (14.7 ± 0.96) for grass_sp3, and 64% higher in winter (24.6 ± 1.66) than summer (8.73 ± 0.68) for grass_sp2, but only 53% higher in winter (8.5 ± 0.51) than summer (4.00 ± 0.40) for grass_sp1.

Let's put it all together:

There was a statistically significant difference in tick abundances recorded on the three grass species (LRT: $\chi^2 = 117.84$, d.f. $= 2$, $P < 0.001$) and between surveys performed in different seasons (LRT: $\chi^2 = 34.853$, d.f. $= 1$, $P < 0.001$). Averaged across seasons, tick abundance was 73% higher on grass_spB (16.6 ± 1.36) than grass_spA (6.25 ± 0.44), but ticks were less abundant on both grass_spA (81% lower) and grass_spB (51% lower) than grass_sp3 (33.6 ± 2.85 ticks per grass stem). Averaged across the three grass species, tick abundances were lower in summer (9.14 ± 0.62) than winter (28.5 ± 2.20). However, the seasonal effect on tick abundances differed across the three grass species (LRT: $\chi^2 = 12.09$, d.f. $= 2$, $P = 0.002$). The significant interaction term was driven by tick abundances being 73% higher in winter (52.5 ± 2.75) than summer (14.7 ± 0.96) for grass_sp3, and 64% higher in winter (24.6 ± 1.66) than summer (8.73 ± 0.68) for grass_sp2, but only 53% higher in winter (8.5 ± 0.51) than summer (4.00 ± 0.40) for grass_sp1.