Homework 11: ANOVA

XDASI Fall 2021

ANSWER KEY

November 13, 2021

Q0: Load the data set and examine it

For this exercise, we will be looking at brood size and fertilization defects in a *C. elegans* gene called *mel-28*, alone and in combination with a mutation in a second gene that suppresses these phenotypes. There are 30 biological replicates for each condition. The measurements taken were:

- **Brood size** - number of eggs laid by one worm in one day
- **Fertilized** - number of eggs laid that were fertilized embryos

First, load the data and take a look at it. Make sure that the `genotype` column is a factor and that you set "WT" as the reference level.

Hint: you can use either the `factor()` or `relevel()` commands to set/reorder the levels; what weird thing happens if you use `level()` instead?

```
anova_data <- read.csv("Ce_mel28_Emb_Ste.csv")
anova_data$genotype = factor(anova_data$genotype, levels=c("WT", "A", "A;B"))

# anova_data <- read.csv("Ce_mel28_Emb_Ste.csv", stringsAsFactors = TRUE)
# relevel(anova_data$genotype,"WT")

str(anova_data)
```

```
## 'data.frame':    90 obs. of  4 variables:
##  $ genotype  : Factor w/ 3 levels "WT","A","A;B": 1 1 1 1 1 1 1 1 1 1 ...
##  $ replicate : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ brood_size: int  160 189 148 174 132 184 171 169 144 165 ...
##  $ fertilized: int  160 187 148 174 132 184 171 169 144 165 ...
```

```
head(anova_data)
```

```
##   genotype replicate brood_size fertilized
## 1       WT         1        160        160
## 2       WT         2        189        187
## 3       WT         3        148        148
## 4       WT         4        174        174
## 5       WT         5        132        132
## 6       WT         6        184        184
```

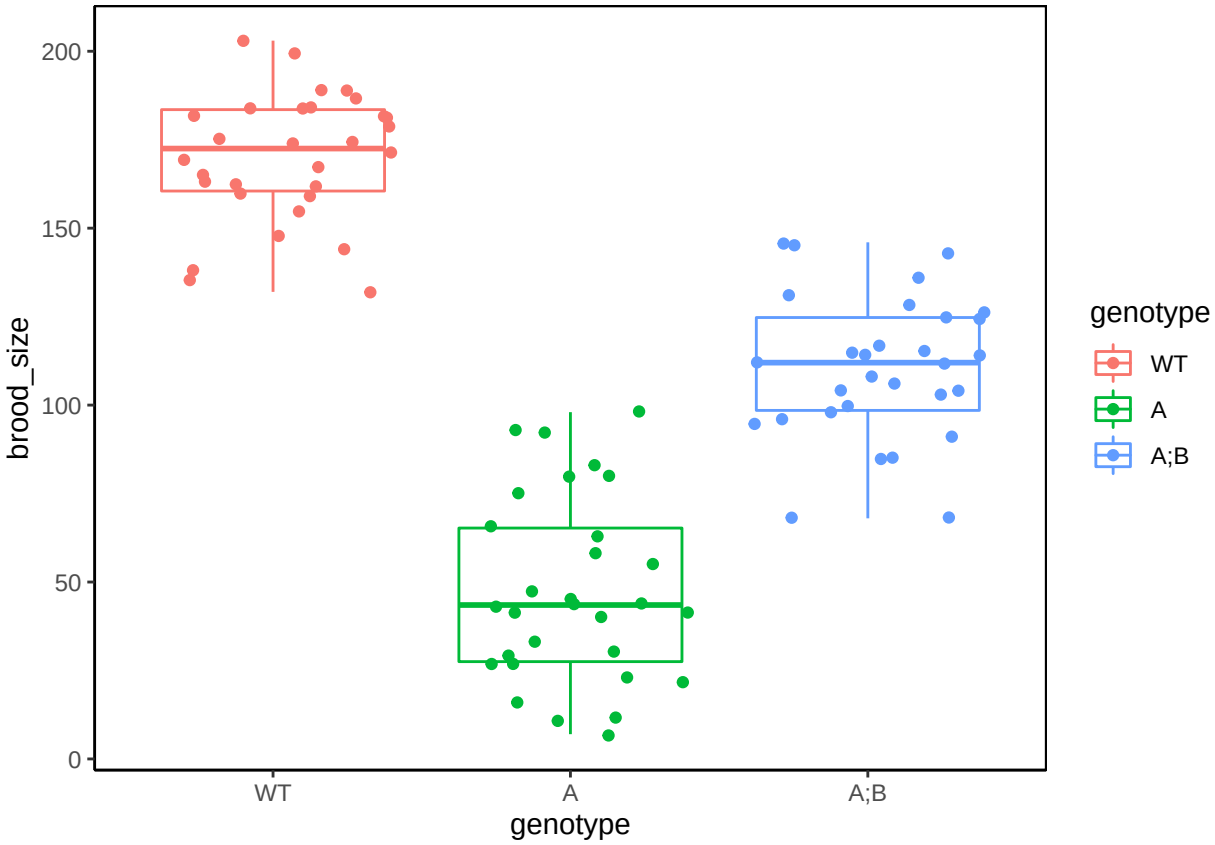
Q1: Brood size

For this first part, we will be looking just at the ***brood sizes*** in the wild-type N2 strain (WT), in single mutants (A), and in double mutants (A;B).

a) Strip chart / boxplot

Plot the brood size data, including the individual data points. For readability, make sure that the WT is the leftmost boxplot, followed by the single mutant and then the double mutant.

```
ggplot(anova_data, aes(x= genotype, y= brood_size, color= genotype)) +
  geom_boxplot() +
  geom_jitter() +
  theme_classic() +
  theme(panel.border=element_rect(colour="black",fill=NA))
```



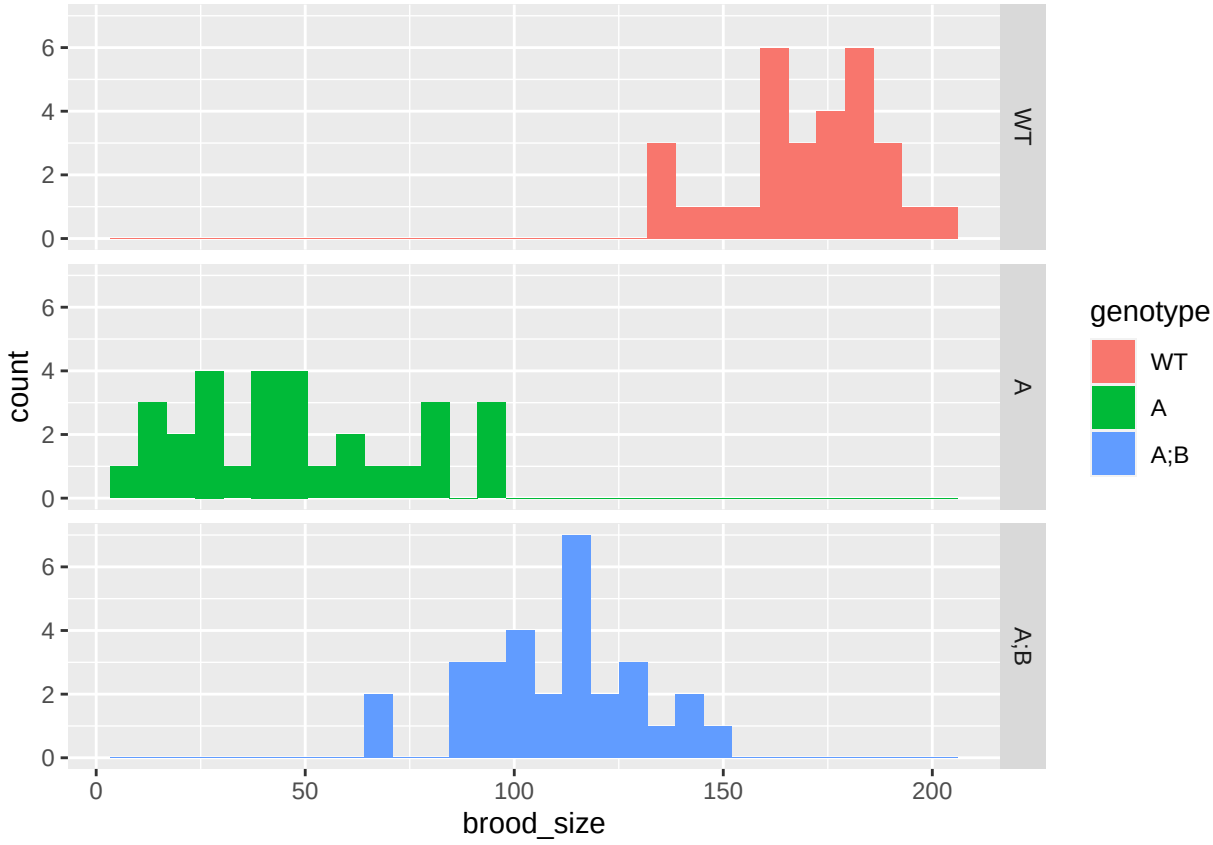
b) Histograms

Use histogram plots to visualize the distributions. and perform an appropriate statistical test to determine if the measurements from the different groups are normally distributed.

```
ggplot(anova_data, aes(x = brood_size, fill = genotype)) +
  geom_histogram() +
```

```
  # theme_classic() +
  # theme(panel.border=element_rect(colour="black",fill=NA)) +
  facet_grid(genotype ~ .)
```

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



c) ANOVA assumptions

In order to proceed with ANOVA, what two requirements should the data fulfill?

```
# your answer here
1) the data must be normally distributed.
2) the variation must be the same between groups.
```

d) Check for normality

Perform statistical tests to determine whether the data are normally distributed. (You may find it easier to subset the data at this point.)

```
anova_WT = anova_data[anova_data$genotype == "WT","brood_size"]
anova_A  = anova_data[anova_data$genotype == "A","brood_size"]
anova_AB = anova_data[anova_data$genotype == "A;B","brood_size"]

shapiro.test(anova_WT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_WT
## W = 0.96664, p-value = 0.4517
```

```
shapiro.test(anova_A)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_A
## W = 0.9485, p-value = 0.1541
```

```
shapiro.test(anova_AB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_AB
## W = 0.97547, p-value = 0.6965
```

e) Check for homoscedasticity

Determine if the variation is the same between the groups. This is called ***"homoscedasticity"***.

- First, use the `var.test()` function to make pairwise comparisons between groups.
- Then, use `bartlett.test()` to perform Bartlett's test on all of the groups at once.

```
var.test(anova_WT, anova_A)
```

```
##
##  F test to compare two variances
##
## data:  anova_WT and anova_A
## F = 0.48311, num df = 29, denom df = 29, p-value = 0.05467
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2299411 1.0150023
## sample estimates:
## ratio of variances
##          0.4831053
```

```
var.test(anova_WT, anova_AB)
```

```
##
##  F test to compare two variances
##
## data:  anova_WT and anova_AB
## F = 0.81212, num df = 29, denom df = 29, p-value = 0.5789
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.386541 1.706263
## sample estimates:
## ratio of variances
##          0.8121209
```

```
var.test(anova_A, anova_AB)
```

```
##
##  F test to compare two variances
##
## data:  anova_A and anova_AB
## F = 1.681, num df = 29, denom df = 29, p-value = 0.1679
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8001174 3.5318648
## sample estimates:
## ratio of variances
##           1.681043
```

```
bartlett.test(brood_size ~ genotype, anova_data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  brood_size by genotype
## Bartlett's K-squared = 4.1495, df = 2, p-value = 0.1256
```

What is the test statistic used by `var.test()`? What is the value of the statistic when the null hypothesis is true?

```
# your answer here
The function performs an F test that compares the variances of two samples from normal populations. F=1
```

What distribution does Bartlett's test statistic follow?

```
# your answer here
The K statistic follows a Chi-square distribution.
```

Are the assumptions for performing ANOVA met for this dataset?

```
# your answer here
Yes, the data are normally distributed and the variances are similar.
```

Q2: Sum of Squares and F-statistic

a) SSE and SSG

Calculate the sum of squares for SS_{error} and SS_{group} (you may use either base R methods or dplyr, as you prefer). Recall the formulas:

$$SS_{total} = SS_{error} + SS_{group}$$
$$SS_{error} = \sum_i \sum_j (Y - \overline{Y}_i)^2$$

$$SS_{group} = \sum_i n_i (\overline{Y}_i - \overline{Y})^2$$

```
## Base R methods ============================================================ ##
# SS_error
anova_WT_sum <- sum((anova_WT - mean(anova_WT))^2)
anova_A_sum  <- sum((anova_A  - mean(anova_A))^2)
anova_AB_sum <- sum((anova_AB - mean(anova_AB))^2)

SS_error <- anova_WT_sum + anova_A_sum + anova_AB_sum

## SS_group
mean_all = mean(anova_data$brood_size) # total mean for all data points

SS_group <- length(anova_WT) * (mean(anova_WT) - mean_all)^2 +
  length(anova_A) * (mean(anova_A) - mean_all)^2 +
  length(anova_AB) * (mean(anova_AB) - mean_all)^2

## dplyr methods ============================================================= ##

# SS_error: first get individual values, then sum across them
#   (could do this in one step, but a bit less confusing this way)
SSE = anova_data %>%
  group_by(genotype) %>%
  summarize(ss = (brood_size - mean(brood_size))^2)
```

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

```
SSE = SSE %>%
  group_by(genotype) %>% summarize(sum = sum(ss)) %>%
  select(sum) %>% sum

## SS_group
mean_all = mean(anova_data$brood_size) # total mean for all data points

SSG = anova_data %>%
  group_by(genotype) %>%
  summarize(ss = length(brood_size)*(mean(brood_size) - mean_all)^2) %>%
  select(ss) %>% sum
```

b) F-statistic and p-value

Calculate the F-statistic and the p-value using the `pf()` function. Recall that:

$$F = \frac{MS_{groups}}{MS_{error}} = \frac{SS_{groups}/df_{groups}}{SS_{error}/df_{error}}$$

```r
## general method to get df for any number of groups from long data ========= ##
# degrees of freedom for SS_error (30*3 - 3 = 87)
df_error = anova_data %>%
  group_by(genotype) %>%
  summarize(df = (length(brood_size) - 1)) %>%
  select(df) %>%
  sum %>%
  unlist
df_error
```

```
## [1] 87
```

```r
# degrees of freedom for SS_groups (3 - 1 = 2)
df_group = anova_data %>%
  select(genotype) %>%
  unique %>% nrow - 1
df_group
```

```
## [1] 2
```

```r
# F-statistic
MS_error <- SS_error/df_error # same as sum(per-group variance)
MS_error
```

```
## [1] 474.005
```

```r
MS_group <- SS_group/df_group
MS_group
```

```
## [1] 112333.8
```

```r
Fstat <- MS_group/MS_error
Fstat
```

```
## [1] 236.9887
```

```r
# p-value using CDF for F-statistic
pf(Fstat, df_group, df_error, lower.tail = F )
```

```
## [1] 6.165467e-36
```

Q3: Planned and Unplanned ANOVA

a) Planned vs. Unplanned

Explain whether this experiment should be treated as a PLANNED or an UNPLANNED dataset. How would the analysis differ between these two scenarios?

```
# your answer here

Datasets that contain a "control" treatment are treated as "planned" experiments, in which different tr

Analysis for an "unplanned" dataset looks for total variation, followed by a Tukey Honest Significant D
```

b) Planned experiment

Assume the data is a planned experiment. Use the `lm()` function to perform the analysis.

First, use the `anova()` function on the results of of the `lm()` function to look at the overall result. *Note: you can use the special notation `$"Pr(>F)"` to get the exact p-value from the `anova()` function.*

```
# anova for lm of brood_size ~ genotype
anova_lm = lm(anova_data$brood_size ~ anova_data$genotype)
anova(anova_lm)
```

```
## Analysis of Variance Table
##
## Response: anova_data$brood_size
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## anova_data$genotype  2 224668  112334  236.99 < 2.2e-16 ***
## Residuals           87  41238     474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(anova_lm)$"Pr(>F)"
```

```
## [1] 6.165467e-36           NA
```

How do the SS_error, SS_group, MS_error, MS_group, $Fstat$, and p-value compare to the ones you calculated by hand?

```
# your answer here
These should be exactly the same.
```

Now, use the `summary()` function on the results of the the linear model.

```
summary(anova_lm)
```

```
##
## Call:
## lm(formula = anova_data$brood_size ~ anova_data$genotype)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.467 -14.492  -1.667  14.133  50.500
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             169.867      3.975   42.73   <2e-16 ***
```

```
## anova_data$genotypeA   -122.367      5.621  -21.77   <2e-16 ***
## anova_data$genotypeA;B  -59.400      5.621  -10.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.77 on 87 degrees of freedom
## Multiple R-squared:  0.8449, Adjusted R-squared:  0.8413
## F-statistic:   237 on 2 and 87 DF,  p-value: < 2.2e-16
```

What do the different parts of the `summary()` output tell you? What is R-squared and what does it mean?

```
# your answer here
```

Is the model significant overall? How about the different groups: are they significantly different from the control?

```
# your answer here
The model is significant overall, and all the groups are significantly different from each other.
```

c) Unplanned experiment

Now assume the data are from an unplanned experiment. Use the `aov()` function to perform the analysis, and then use the `tukeyHSD()` function on the results of `aov()`.

```
# aov for brood_size ~ genotype
aov(anova_data$brood_size ~ anova_data$genotype)
```

```
## Call:
##    aov(formula = anova_data$brood_size ~ anova_data$genotype)
##
## Terms:
##                 anova_data$genotype Residuals
## Sum of Squares            224667.62  41238.43
## Deg. of Freedom                   2        87
##
## Residual standard error: 21.77166
## Estimated effects may be unbalanced
```

Use the `tukeyHSD()` function on the results of `aov()` to perform the Tukey-Kramer test to look at all pairwise comparisons.

```
# Tukey's honest significant difference test
TukeyHSD(aov(anova_data$brood_size ~ anova_data$genotype))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = anova_data$brood_size ~ anova_data$genotype)
##
## $`anova_data$genotype`
##              diff        lwr        upr p adj
```

```
## A-WT   -122.36667 -135.77082 -108.96251     0
## A;B-WT  -59.40000  -72.80415  -45.99585     0
## A;B-A    62.96667   49.56251   76.37082     0
```

What does Tukey's HSD test do? What do these results tell you? What is the biological interpretation of these results?

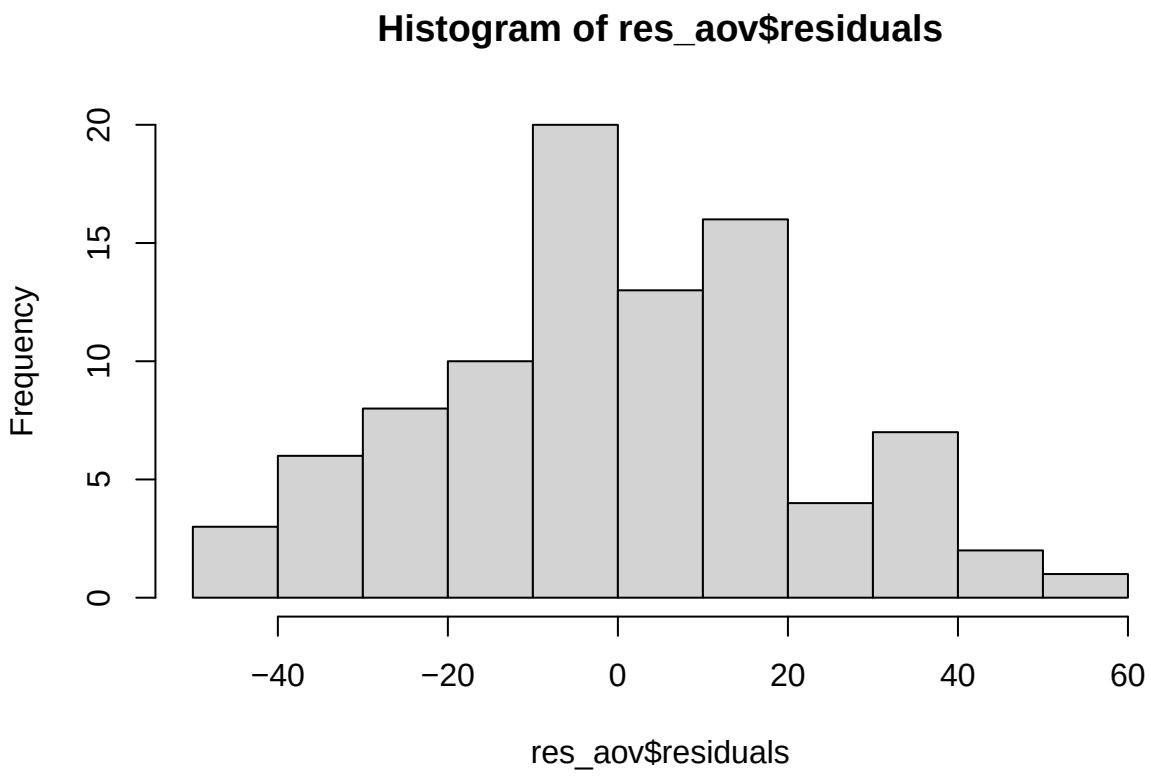
```
# your answer here
Tukey's HSD tests all pairwise combinations for possible significant differences. The results here indi

The single mutant has a reduced brood size phenotype that is significantly different from WT. The double
```

d) Residuals

Once you have performed an ANOVA, you can also check your assumption of normality by checking to see if the ANOVA residuals are normally distributed.

- Do this by first assigning the results of your `aov()` function to a variable (e.g. `res_aov`.
- Then, make a histogram of the residuals, which can be found in `res_aov$residuals`.
- Finally, perform a test for normality on the residuals.

```
res_aov <- aov(brood_size ~ genotype, data = anova_data)
hist(res_aov$residuals)
```



```
shapiro.test(res_aov$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res_aov$residuals
## W = 0.98715, p-value = 0.5251
```

e) ANOVA for unequal variances

If the data look normal, but the variances are unequal, what kind of ANOVA could you do instead?

```
# your answer here
Welch's ANOVA: This can be used when the data are normal, but the variances are unequal.
```

Perform this test below. Also extract the exact p-value from the test.

```
oneway.test(brood_size ~ genotype, data = anova_data, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  brood_size and genotype
## F = 227.72, num df = 2.000, denom df = 56.891, p-value < 2.2e-16
```

```
oneway.test(brood_size ~ genotype, data = anova_data, var.equal = FALSE)$p.value
```

```
## [1] 7.053586e-28
```

How does the p-value compare with the ANOVA you performed above? Explain.

```
# your answer here
The p-value is highly significant but slightly larger than for the ANOVA assuming
equal variances. This is because a little bit of power is lost when the data are
not treated as homoscedastic (in which case they are called "heteroscedastic").
```

Note: If you want to add annotations to your boxplots to show p-values or significance levels for various statistical tests, the ggpubr() package offers a variety of options that you can read about on the R-bloggers website.

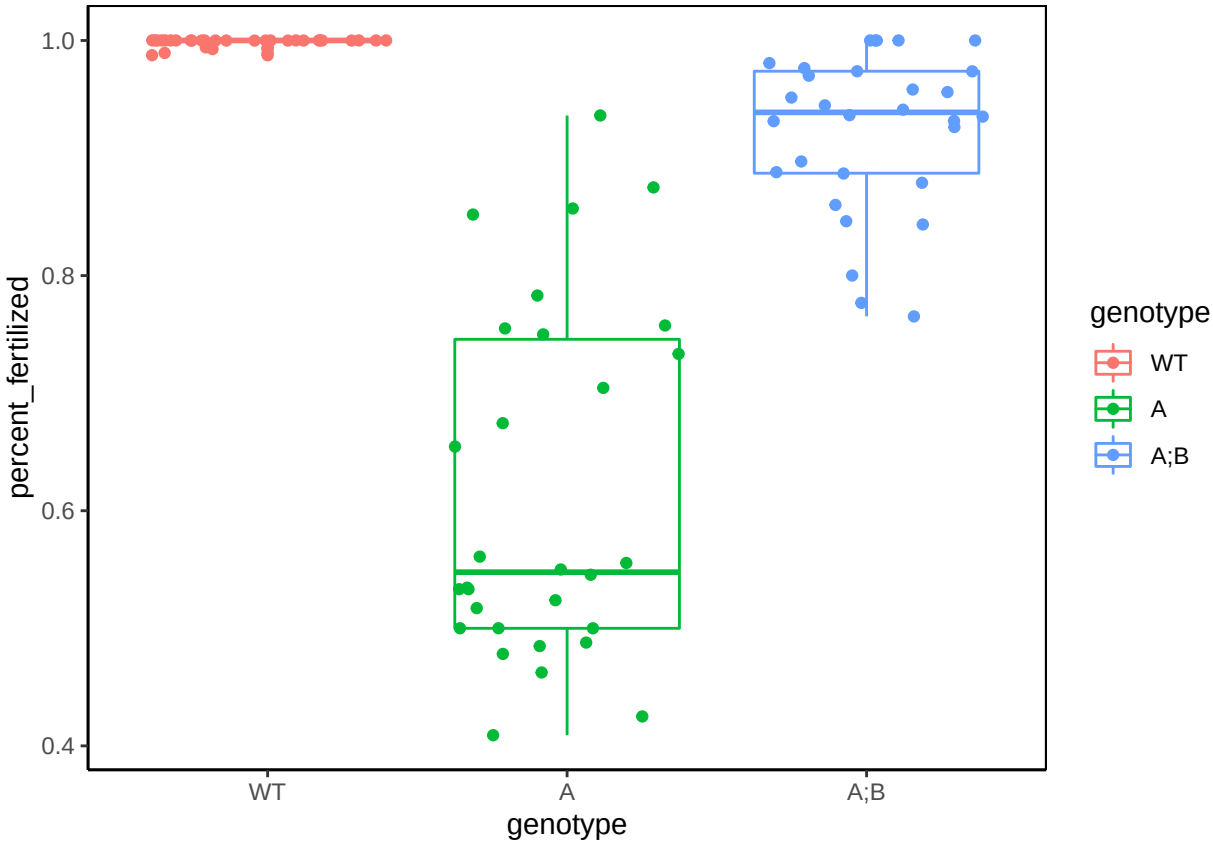
Q4: Fertilization phenotype

Now we will look at the fertilization phenotype.

a) Percent fertilization

- Add a column to your original data frame containing the ***percent fertilization***.
- Take a look at the percent fertilization using a stip chart / boxplot (as you did for Q1).

```
anova_data = anova_data %>% mutate(percent_fertilized = fertilized/brood_size)

ggplot(anova_data, aes(x= genotype, y= percent_fertilized,
                       color= genotype)) +
  geom_boxplot() +
  geom_jitter() +
  theme_classic() +
  theme(panel.border=element_rect(colour="black",fill=NA))
```



b) Check assumptions for ANOVA

Perform the appropriate tests to check whether you can perform an ANOVA on these data and draw histograms to visualize the distributions. (You may again find it useful at this point to subset the data by group.)

```
anova_WT = anova_data[anova_data$genotype == "WT","percent_fertilized"]
anova_A  = anova_data[anova_data$genotype == "A","percent_fertilized"]
anova_AB = anova_data[anova_data$genotype == "A;B","percent_fertilized"]

shapiro.test(anova_WT)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  anova_WT
## W = 0.42289, p-value = 8.965e-10
```

```
shapiro.test(anova_A)
```

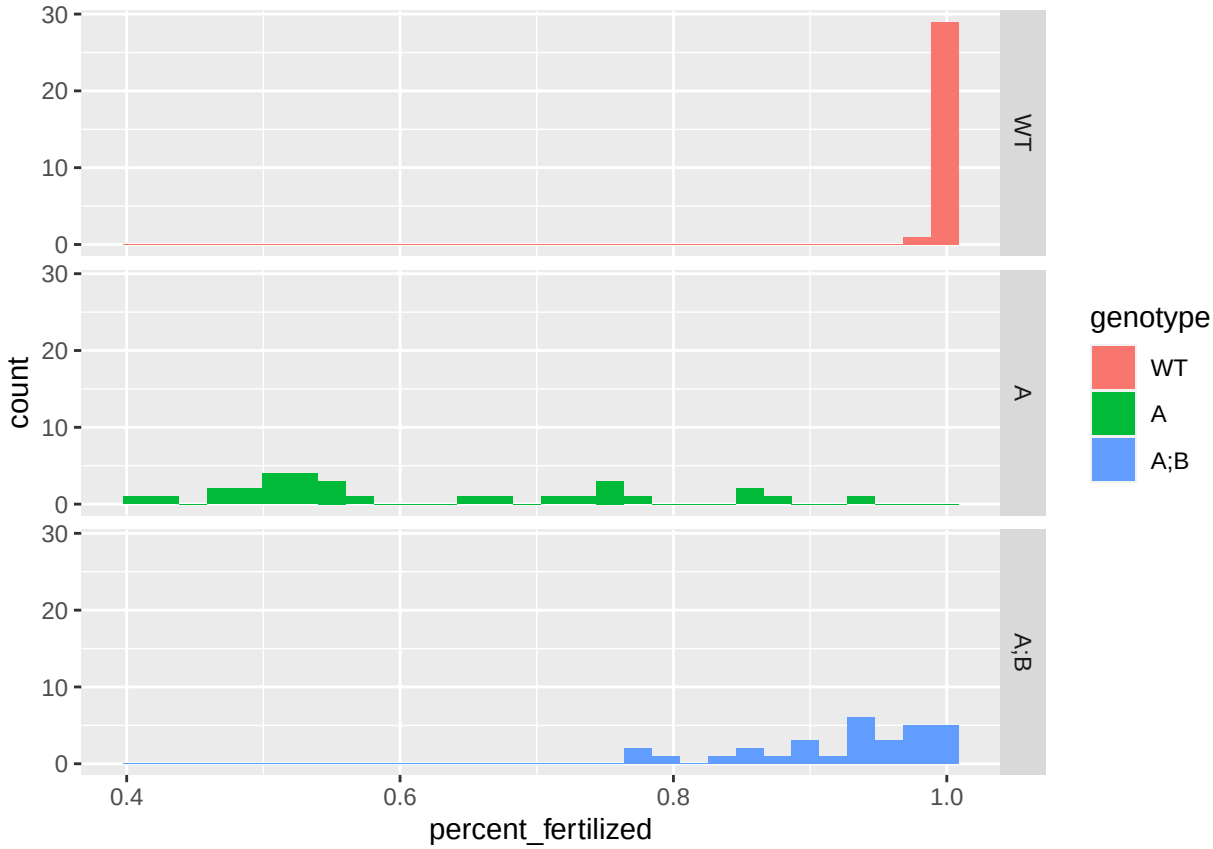
```
##
##  Shapiro-Wilk normality test
##
## data:  anova_A
## W = 0.89891, p-value = 0.0079
```

```
shapiro.test(anova_AB)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anova_AB
## W = 0.89877, p-value = 0.007837
```

```
ggplot(anova_data, aes(x = percent_fertilized, fill = genotype)) +
  geom_histogram() +
  facet_grid(genotype ~ .)
```

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



Do these data meet the assumptions required for ANOVA? If not, what alternative tests could you perform? Explain the conditions under which either test would be appropriate.

```
# your answer here
If the data are not normal, a non-parametric version of ANOVA can be performed instead.

1) Data are normal but the variances are unequal: Welch's ANOVA
2) Data are not normal, and variances may not be unequal:
  a) Kruskal-Wallace test
  b) Pairwise Wilcoxon Rank Sum test, with adjustment for multiple hypothesis testing (e.g. Benjamini-H
```

c) Statistical test

Perform an appropriate alternative to ANOVA for these data.

```
kruskal.test(percent_fertilized ~ genotype, data = anova_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  percent_fertilized by genotype
## Kruskal-Wallis chi-squared = 72.331, df = 2, p-value < 2.2e-16
```

```
kruskal.test(percent_fertilized ~ genotype, data = anova_data)$p.value
```

```
## [1] 1.965366e-16
```

```
pairwise.wilcox.test(anova_data$percent_fertilized, anova_data$genotype,
                     p.adjust.method = "BH")
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot compute
## exact p-value with ties

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  anova_data$percent_fertilized and anova_data$genotype
##
##     WT      A
## A   1.2e-11 -
## A;B 6.5e-09 9.9e-10
##
## P value adjustment method: BH
```

Q5: Error, Power, and Sample size

Recall that the power is $1 - \beta$, where β is the Type II error. We can calculate the power of the ANOVA test using the command `power.anova.test()`, which is contained in the base `stats` package. The arguments for the command are:

- **groups** = the number of groups in the dataset
- **n** = the number of observations per group
- **between.var** = between-group variance, SS_{group}
- **within.var** = within-group variance, SS_{error}
- **sig.level** = significance cutoff α (Type I error)
- **power** = power of the test ($1 - \beta = 1$ - Type II error)

Given all of the arguments except for one, this function will return a value for the missing argument.

a) Power of ANOVA for this dataset

Given the data that we have been provided and the p-value we obtained for the `brood_size`, what is the power of our dataset, given a Type I error rate (significance cutoff) of $\alpha = 0.05$?

```
# first, check the documentation
help(power.anova.test)

# compute the power
power.anova.test(groups=3,
                 n=30,
                 within.var=SS_error,
                 between.var=SS_group,
                 power=NULL,
                 sig.level=0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 30
##     between.var = 224667.6
##      within.var = 41238.43
##       sig.level = 0.05
##           power = 1
##
## NOTE: n is number in each group
```

b) Sample size

If we would like to achieve a power of 95% with a Type I error rate of $\alpha = 0.05$, how many observations for each group would we need?

```
# compute the required sample size
power.anova.test(groups=3,
                 n=NULL,
                 within.var=SS_error,
                 between.var=SS_group,
                 power=0.95,
                 sig.level=0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 2.747844
##     between.var = 224667.6
##      within.var = 41238.43
##       sig.level = 0.05
##           power = 0.95
##
## NOTE: n is number in each group
```

```
# sample size needed = 3 (round up to the nearest integer)
```

How do you feel about this result?

```
# your answer here
This seems like a pretty small sample size to rely on, especially if it's
not that hard to collect the data ... but if it is, then we could make do with
quite a bit fewer observations and still do really well with this effect size.

Personally, I'd feel more comfortable with at least 10-12 biological replicates
for this experiment.
```

c) Effect size

The ratio of the between-groups variation to the within-groups variation (SS_{group}/SS_{error}) gives us an approximate effect size.

The `brood_size` data here has a pretty big effect size, and we have a lot of power because of this.

If the effect size were 10 times smaller, how many animals would we need to get a power of 95% at a Type I error threshold of 5%?

```
power.anova.test(groups=3,
                 n=NULL,
                 within.var=SS_error,
                 between.var=SS_group/10,
                 power=0.95,
                 sig.level=0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
```

```
##          groups = 3
##               n = 15.22547
##     between.var = 22466.76
##      within.var = 41238.43
##       sig.level = 0.05
##           power = 0.95
##
## NOTE: n is number in each group
```

```
power.anova.test(groups=3,
                 n=NULL,
                 within.var=SS_error*10,
                 between.var=SS_group,
                 power=0.95,
                 sig.level=0.05)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 15.22547
##     between.var = 224667.6
##      within.var = 412384.3
##       sig.level = 0.05
##           power = 0.95
##
## NOTE: n is number in each group
```

```
# we would need ~16 animals per group (round up to nearest integer)
```