ANOVA and regression review

KEY

Wed Mar 10 13:46:25 2021

Analysis of variance (ANOVA)

To review ANOVA, we have data on weight gain in Daphnia fed on cyanobacteria. Import the data into a data frame called "anova"

```
library(readxl)

daphnia <- read_excel("review_data.xls","anova")
```

To graph the group means and error bars we need to summarize the data. Calculate the means, standard deviations, and sample sizes, assemble them into a data frame, and then calculate the standard errors:

```
daphnia.sumstats <- do.call("data.frame", aggregate(resistance ~ cyandensity,
data = daphnia, FUN = function(x) c(mean = mean(x), s = sd(x), n = length(x),
se = sd(x)/sqrt(length(x)))))
daphnia.sumstats$cyandensity <- ordered(daphnia.sumstats$cyandensity, levels
= c("low","med","high"))
```

We can also create a function called summarySE() that will do this same set of calculations:

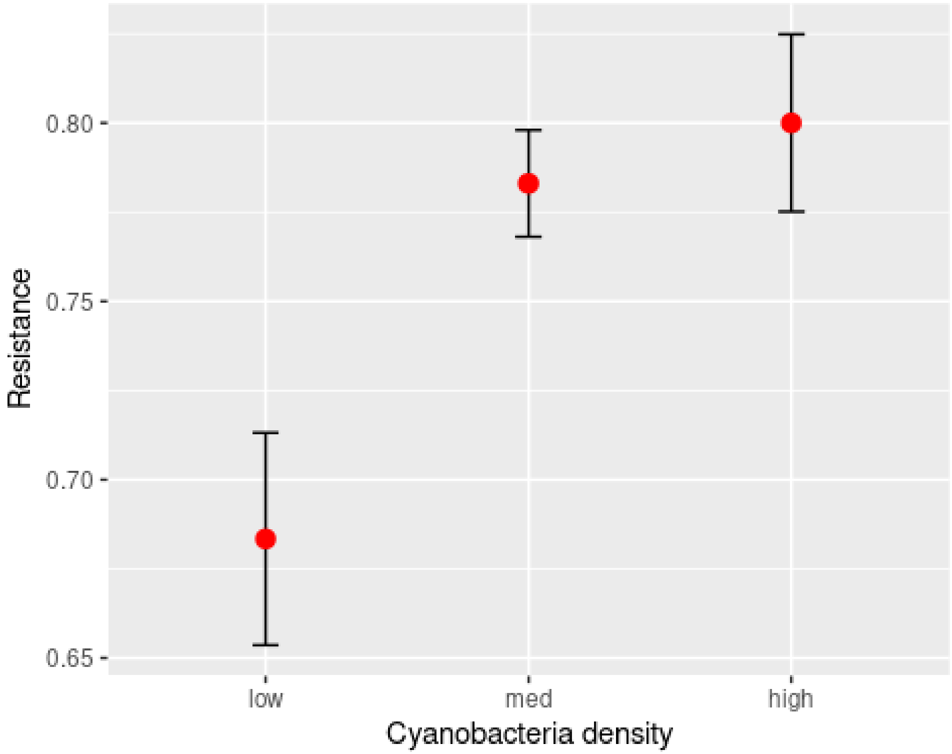
```
summarySE <- function(dataset, measure.vars, groups) {
  grps <- aggregate(as.formula(paste(measure.vars, "~", groups)), data =
dataset, FUN = mean)[,groups]
  m <- aggregate(as.formula(paste(measure.vars, "~", groups)), data =
dataset, FUN = mean)[,measure.vars]
  s <- aggregate(as.formula(paste(measure.vars, "~", groups)), data =
dataset, FUN = sd)[,measure.vars]
  n <- aggregate(as.formula(paste(measure.vars, "~", groups)), data =
dataset, FUN = length)[,measure.vars]
  se <- s/sqrt(n)
  sumstats <- data.frame(grps, m, s, n, se)
  colnames(sumstats) <- c(groups, "mean","sd","n","se")
  return(sumstats)
}
```

Question: what is the null hypothesis for an ANOVA?

> Ho: mu(high) = mu(med) = mu(low), or in other words, the population means are no different for the three cyanobacteria densities

Plot the means and error bars:

```
library(ggplot2)
ggplot(daphnia.sumstats, aes(x = cyandensity, y = resistance.mean)) +
geom_errorbar(aes(ymax = resistance.mean + resistance.se, ymin =
resistance.mean - resistance.se), width = 0.1) + geom_point(size = 3, color =
"red") + labs(x = "Cyanobacteria density", y = "Resistance")
```



Question: based on the graph does it appear that adaptation to cyanobacteria toxins is having the expected effect? How do you know?

Yes, the resistance levels go up from low to high cyanobacteria levels.

Test for normality:

```
with(daphnia, tapply(resistance, cyandensity, shapiro.test))

## $high
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95101, p-value = 0.6805
##
##
## $low
##
##  Shapiro-Wilk normality test
##
```

```
## data:  X[[i]]
## W = 0.91379, p-value = 0.2385
##
##
## $med
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.99377, p-value = 0.9995
```

Question: do we meet the normality assumption for all three groups? How do you know?

Yes, the p-values are all greater than 0.05, and since normality is the null hypothesis we meet the normality assumption.

Test for HOV:

```
with(daphnia, bartlett.test(resistance, cyandensity))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  resistance and cyandensity
## Bartlett's K-squared = 5.0532, df = 2, p-value = 0.07993
```

Question: do we meet the HOV assumption? How do you know?

Yes, p is greater than 0.05.

Run the ANOVA:

```
daphnia.aov <- aov(resistance ~ cyandensity, data = daphnia)
anova(daphnia.aov)
```

```
## Analysis of Variance Table
##
## Response: resistance
##             Df   Sum Sq  Mean Sq F value   Pr(>F)
## cyandensity  2 0.089195 0.044598  6.6916 0.004078 **
## Residuals   29 0.193277 0.006665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: do we reject the null hypothesis? How do you know?

Yes, p is less than 0.05

Question: what do you know about differences between the three densities at this stage?

There is at least one pair of means that is different, but we don’t know which ones.

Run a Tukey HSD procedure to evaluate which means are different from which:

```
TukeyHSD(daphnia.aov)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = resistance ~ cyandensity, data = daphnia)
##
## $cyandensity
##                 diff         lwr         upr     p adj
## low-high -0.11666667 -0.20299362 -0.03033972 0.0063918
## med-high -0.01700000 -0.10716556  0.07316556 0.8878183
## med-low   0.09966667  0.01333972  0.18599362 0.0210326
```

Question: which means are differnet from which?

Low is different from high, medium is different from low, but high and medium are not different.

Simple linear regression

These data represent an experiment on the effects of environmental temperature on body temperature, measured in the thorax or in the abdomen of winter moths. We will use simple linear regression to evaluate how body temperature is affected by environmental temperature.

Question: what is the null hypothesis for a simple linear regression?

Ho: beta = 0, where beta is the slope of the line (thus, the null is that the line is flat)

Question: what should the y-intercept be if the null hypothesis is true?

It should be equal to the mean of the body temperature variable.

Import the regression worksheet:

```
library(readxl)

moth <- read_excel("review_data.xls","regression")
```

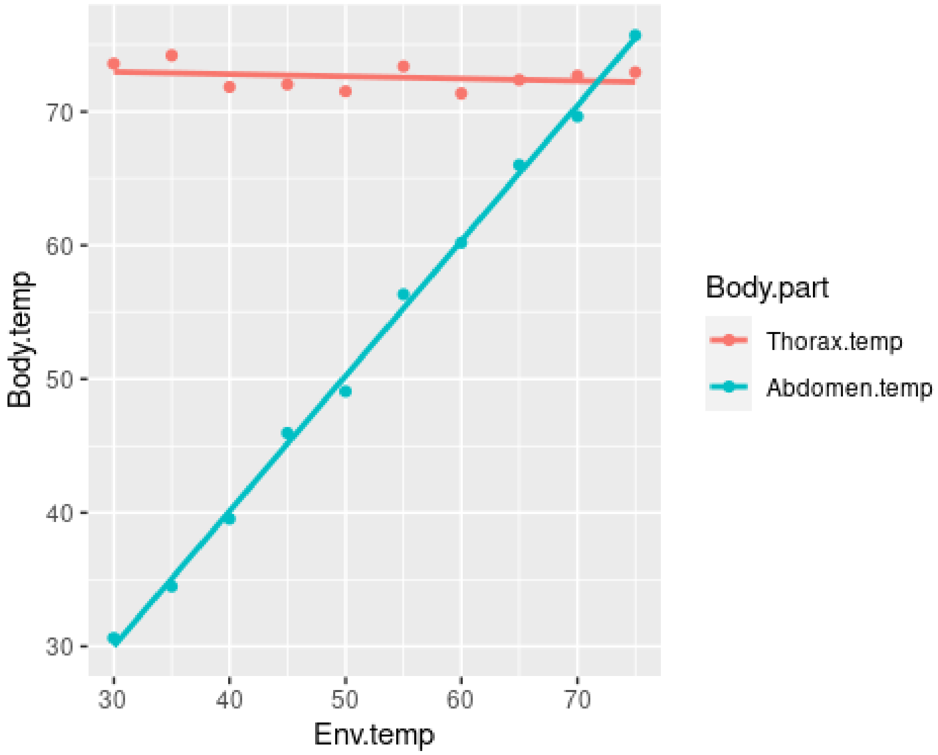
Stack the data for plotting:

```
stack(moth, c(2,3)) -> moth.stacked
data.frame(Env.temp = moth$Env.temp, moth.stacked) -> moth.stacked
names(moth.stacked) <- c("Env.temp","Body.temp","Body.part")
```

Plot the data:

```
ggplot(moth.stacked, aes(x = Env.temp, y = Body.temp, group = Body.part,
color = Body.part)) + geom_point() + geom_smooth(method = "lm", se = F)

## `geom_smooth()` using formula 'y ~ x'
```



Question: which body temperature measurement appears to change when environmental temperature changes?

> Abdomen temperature. Thorax temperature is flat across the range of environmental temperatures tested.

Run the regressions to get the fitted models:

```
lm(Abdomen.temp ~ Env.temp, data = moth) -> abdomen.lm
lm(Thorax.temp ~ Env.temp, data = moth) -> thorax.lm
```

Get the ANOVA tables:

```
anova(abdomen.lm)

## Analysis of Variance Table
##
## Response: Abdomen.temp
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Env.temp   1 2107.3 2107.28  3241.2 1.006e-11 ***
## Residuals  8    5.2    0.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(thorax.lm)

## Analysis of Variance Table
##
## Response: Thorax.temp
##           Df Sum Sq Mean Sq F value Pr(>F)
## Env.temp   1 0.5909 0.59085  0.6355 0.4484
## Residuals  8 7.4382 0.92977
```

Question: which body temperature variable(s) are related to environmental temperature?

Abdoment temperature is, but thorax temperature is not.

Get the coefficients and r2:

```
summary(abdomen.lm)

##
## Call:
## lm(formula = Abdomen.temp ~ Env.temp, data = moth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16297 -0.57980  0.03181  0.61661  1.04760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.31194    0.96637  -0.323    0.755
## Env.temp     1.01080    0.01775  56.931 1.01e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8063 on 8 degrees of freedom
## Multiple R-squared:  0.9975, Adjusted R-squared:  0.9972
## F-statistic:  3241 on 1 and 8 DF,  p-value: 1.006e-11

summary(thorax.lm)

##
## Call:
## lm(formula = Thorax.temp ~ Env.temp, data = moth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1303 -0.8986  0.1900  0.7062  1.3125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.47539    1.15563  63.580 4.16e-12 ***
## Env.temp    -0.01693    0.02123  -0.797    0.448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9642 on 8 degrees of freedom
## Multiple R-squared:  0.07359,    Adjusted R-squared:  -0.04221
## F-statistic: 0.6355 on 1 and 8 DF,  p-value: 0.4484
```

Question: what is the slope estimate for thorax temperature? If the null is true what should it be? Is it a problem that it isn't exactly equal to the null hypothetical value?

> The slope is -0.01693, and the null hypothesis is that it's equal to 0. It's fine that the estimate isn't equal 0, because the null value is what we expect at the population level, and the estimate is taken from our sample of data. We don't expect sample estimates to be exactly equal to their population values.

Question: give the regression equation for each model.

> Thorax: Thorax.temp = 73.47 -0.01693 Env.temp
>
> Abdomen: Abdomen.temp = -0.312 + 1.01 Env.temp

Question: do your regression results support your interpretation of the graphs? In other words, did you get the results you expected given that the graph showed a flat line for thorax temperature and sloped line for abdomen temperature?

> Yes, the thorax temperature is very flat on the graph, and has a non-significant slope that is close to 0. The abdomen temperature is sloped, and has a significant slope that's near 1.