Applied Analysis

## Question 1

* What is my question and hypothesis?
* Difference between two levels?
* Difference across many different levels?
* Trends?
* Correlation?

## Question 2

* What kind of data do I have?
* Discrete
  + Success/Failure
  + Count
  + Are there a lot of zeros?
* Continuous
  + Positive only?
  + Whole real line?
  + Is it a proportion?
* Is it approximately Normal?

## Sea Cucumbers

* Want to know if they move more when there is less food, and how that varies with density.
* Positively restricted continuous variable. Might be approximately normal unless there are some zeros…

## Mean Difference:

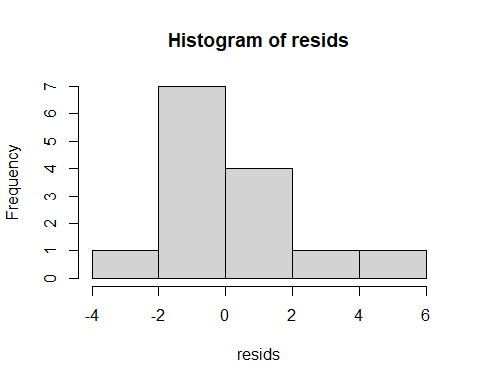
* Big or small data set (can I rely on the central limit theorem?)
* Are the observations independent and do they have equal variance?

high <- c(9.57, 16.7, 12.9, 8.95, 10.3, 9.88, 11.0)  
low <- c(5.27, 5.22, 6.81, 5.62, 2.83, 4.45, 3.42)  
  
fit.t <- t.test(high, low, alternative = "two.sided", mu = 0, paired = FALSE, var.equal = TRUE)  
fit.t

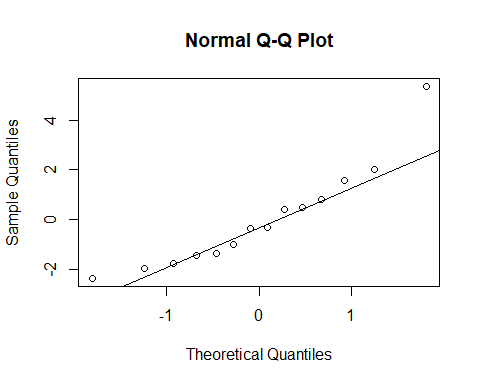
##   
## Two Sample t-test  
##   
## data: high and low  
## t = 5.7364, df = 12, p-value = 9.361e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 4.047097 9.004332  
## sample estimates:  
## mean of x mean of y   
## 11.328571 4.802857

## Check Assumptions

resids <- c(low - mean(low), high - mean(high))  
 hist(resids)



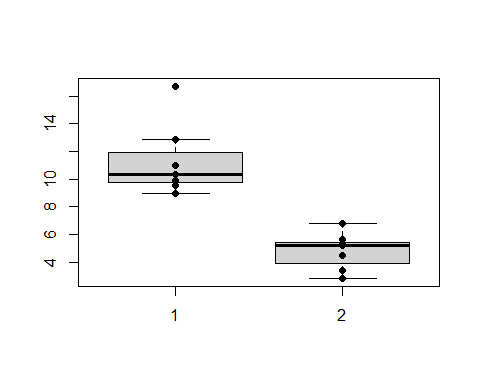
qqnorm(resids)  
 qqline(resids)



## Reporting

* A two sample t-test was used to test the mean difference between the two groups assuming equal variances based on an F-test. We validated the test by visually inspecting the two groups for normality.
* We found that in higher concentrations of total organic material (TOM), sea cucumbers on average moved more (Mean, 95% confidence interval) 6.5 m (4.0, 9.0).

## Visualize it

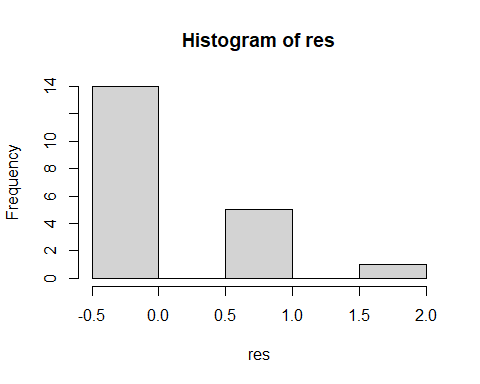


## Bad t-test

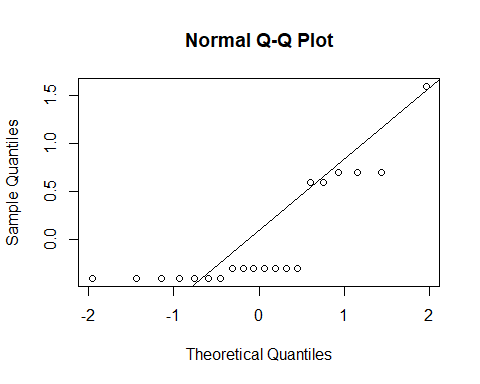
grp1 <- rpois(10, 0.5)  
grp2 <- rpois(10, 0.9)  
t.test(grp1, grp2)

##   
## Welch Two Sample t-test  
##   
## data: grp1 and grp2  
## t = -0.3721, df = 15.997, p-value = 0.7147  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.669716 0.469716  
## sample estimates:  
## mean of x mean of y   
## 0.3 0.4

res <- c(grp1 - mean(grp1), grp2- mean(grp2))  
 hist(res)



qqnorm(res)  
 qqline(res)



## Count data difference

* Let’s assume we count the total and as a Bionimal distribution, assume that each group is equally likely.

n <- sum(c(grp1, grp2)); x1 <- sum(grp1); x2 <- sum(grp2)  
 poisson.test(c(x1, x2), r = 1)

##   
## Comparison of Poisson rates  
##   
## data: c(x1, x2) time base: 1  
## count1 = 3, expected count1 = 3.5, p-value = 1  
## alternative hypothesis: true rate ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.1098635 4.4332599  
## sample estimates:  
## rate ratio   
## 0.75

binom.test(x1, n, p = 0.5)

##   
## Exact binomial test  
##   
## data: x1 and n  
## number of successes = 3, number of trials = 7, p-value = 1  
## alternative hypothesis: true probability of success is not equal to 0.5  
## 95 percent confidence interval:  
## 0.09898828 0.81594843  
## sample estimates:  
## probability of success   
## 0.4285714

* Manually compute the p-value and CI

probs <- dbinom(0:n, prob = 0.5, size = n)   
 sum(probs[probs <= dbinom(x1, prob = 0.5, size = n)])

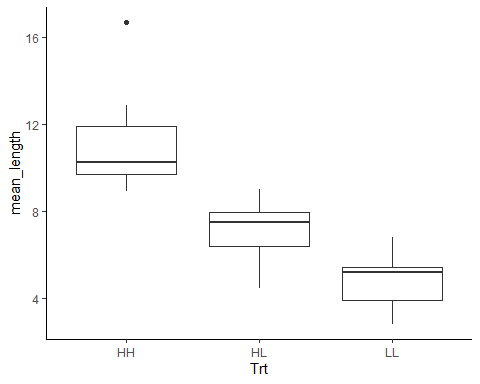
## [1] 1

## More than one Mean to compare:

* Question 1: Is there a significant treatment effect?
* Question 2: Which treatment is having an impact?
* Example data: Sea Cucumber distance traveled for high density animals, comparing H-L, H-H, L-L.

## Analysis of Variance

dat.cuke <- cuke.length.example %>% filter(DensHL == "H")  
 ggplot(data = dat.cuke, aes(x = Trt, y = mean\_length)) +   
 geom\_boxplot() + theme\_classic()



## Question 1: Anything significant?

* Response Variable: for .
* One-Way ANOVA
* Main idea: Consider Total Sum of Squares ()
* Remember that sum of squares of normals end up distributed.
* Ratios of distributions end up F distributed.
* If there is a significant treatment effect,
  + the treatment variance should be LARGER than the Sum of Squared Error (residual variance).
* Assumptions
  + The observations are made from each mean group independently and with constant `white noise’ error.

## Analysis (1-way ANOVA)

fit.aov <- aov(mean\_length ~ Trt, data = dat.cuke)  
 summary(fit.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Trt 2 152.96 76.48 20.29 2.44e-05 \*\*\*  
## Residuals 18 67.85 3.77   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Sum of Squared Error  
 dat.cuke %>% group\_by(Trt) %>%   
 summarize(SS = sum((mean\_length - mean(mean\_length))^2)) %>%  
 ungroup() %>% summarize(SE = sum(SS)) %>% as.numeric()

## [1] 67.84978

## Sum of Squared Treatment  
 dat.cuke %>% mutate(avg = mean(mean\_length)) %>%  
 group\_by(Trt) %>%   
 summarize(SS = sum((mean(mean\_length) - avg)^2)) %>%  
 ungroup() %>% summarize(SS = sum(SS)) %>% as.numeric()

## [1] 152.9569

## Is this useful?

* We just get at least one significant effect… Better to analyze the treatment effects!

em.cuke <- emmeans(fit.aov, specs = ~Trt)  
 em.cuke

## Trt emmean SE df lower.CL upper.CL  
## HH 11.32 0.734 18 9.78 12.86  
## HL 7.11 0.734 18 5.56 8.65  
## LL 4.80 0.734 18 3.26 6.35  
##   
## Confidence level used: 0.95

## Is this useful?

diff.cuke <- contrast(em.cuke, "pairwise")  
diff.cuke

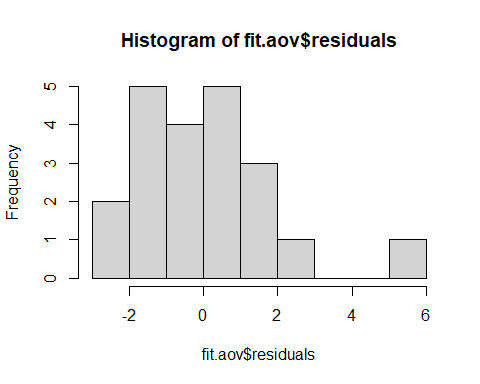
## contrast estimate SE df t.ratio p.value  
## HH - HL 4.22 1.04 18 4.062 0.0020   
## HH - LL 6.52 1.04 18 6.281 <.0001   
## HL - LL 2.30 1.04 18 2.218 0.0950   
##   
## P value adjustment: tukey method for comparing a family of 3 estimates

confint(diff.cuke)

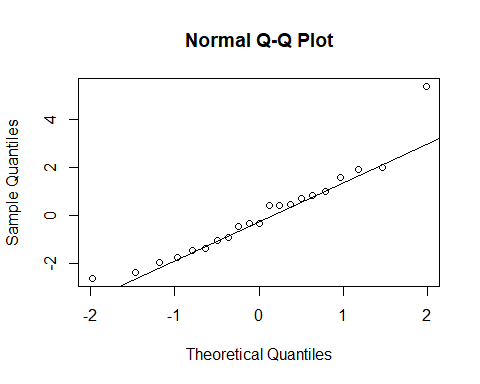
## contrast estimate SE df lower.CL upper.CL  
## HH - HL 4.22 1.04 18 1.567 6.86  
## HH - LL 6.52 1.04 18 3.869 9.17  
## HL - LL 2.30 1.04 18 -0.346 4.95  
##   
## Confidence level used: 0.95   
## Conf-level adjustment: tukey method for comparing a family of 3 estimates

## Is this valid?

hist(fit.aov$residuals)



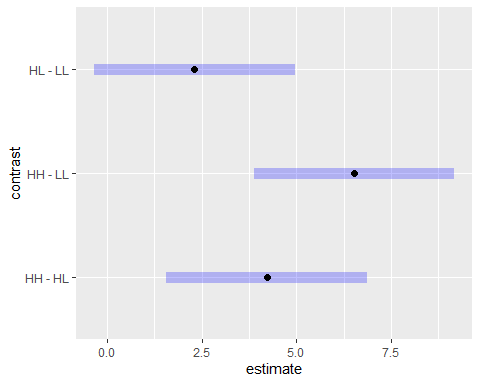
qqnorm(fit.aov$residuals)  
 qqline(fit.aov$residuals)



## Reporting

* In analysis section: The data were analyzed for a treatment effect using an analysis of variance. We performed a residual analysis to ensure that the model assumptions were met.
* The experiments showed a very strong high TOM effect when compared with low only, animals moved 6.5 m (3.9-9.2) more on average. There was also a strong High to mixed (High-Low) TOM effect, (mean difference: 4.22 m, 1.6-6.9). There was not as much evidence for a difference between mixed and Low TOM. See Figure 1 for a comparison of the different treatments.

plot(diff.cuke)



## Analysis (2-way ANOVA)

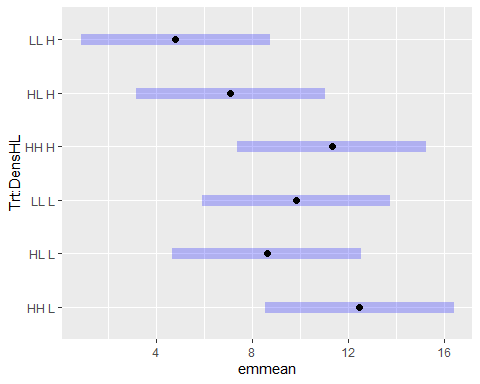
mod.aov <- aov(mean\_length ~ Trt\*DensHL, data = cuke.length.example)  
summary(mod.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Trt 2 174.8 87.40 3.327 0.0472 \*  
## DensHL 1 68.5 68.51 2.608 0.1151   
## Trt:DensHL 2 32.3 16.17 0.616 0.5460   
## Residuals 36 945.8 26.27   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

mod.aov$coefficients

## (Intercept) TrtHL TrtLL DensHLH TrtHL:DensHLH   
## 12.4567643 -3.8505965 -2.6256535 -1.1353947 -0.3650299   
## TrtLL:DensHLH   
## -3.8921308

em.cuke <- emmeans(mod.aov, specs = ~Trt\*DensHL)  
plot(em.cuke)



contrast(em.cuke, "pairwise")

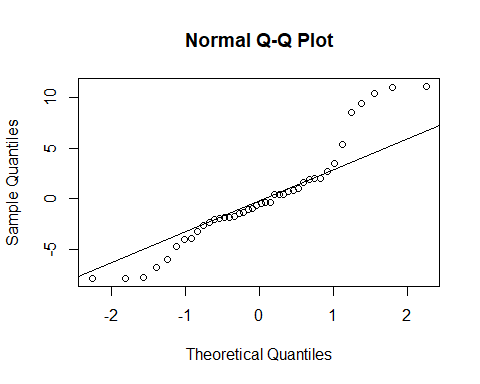
## contrast estimate SE df t.ratio p.value  
## HH,L - HL,L 3.85 2.74 36 1.405 0.7235   
## HH,L - LL,L 2.63 2.74 36 0.958 0.9279   
## HH,L - HH,H 1.14 2.74 36 0.414 0.9983   
## HH,L - HL,H 5.35 2.74 36 1.953 0.3882   
## HH,L - LL,H 7.65 2.74 36 2.793 0.0817   
## HL,L - LL,L -1.22 2.74 36 -0.447 0.9976   
## HL,L - HH,H -2.72 2.74 36 -0.991 0.9178   
## HL,L - HL,H 1.50 2.74 36 0.548 0.9937   
## HL,L - LL,H 3.80 2.74 36 1.388 0.7338   
## LL,L - HH,H -1.49 2.74 36 -0.544 0.9939   
## LL,L - HL,H 2.73 2.74 36 0.995 0.9166   
## LL,L - LL,H 5.03 2.74 36 1.835 0.4572   
## HH,H - HL,H 4.22 2.74 36 1.539 0.6424   
## HH,H - LL,H 6.52 2.74 36 2.379 0.1907   
## HL,H - LL,H 2.30 2.74 36 0.840 0.9578   
##   
## P value adjustment: tukey method for comparing a family of 6 estimates

TukeyHSD(mod.aov)

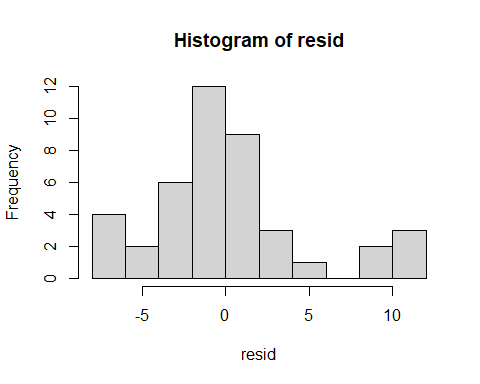
## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = mean\_length ~ Trt \* DensHL, data = cuke.length.example)  
##   
## $Trt  
## diff lwr upr p adj  
## HL-HH -4.0331114 -8.768568 0.7023454 0.1079099  
## LL-HH -4.5717188 -9.307176 0.1637379 0.0602841  
## LL-HL -0.5386074 -5.274064 4.1968494 0.9583377  
##   
## $DensHL  
## diff lwr upr p adj  
## H-L -2.554448 -5.762566 0.6536697 0.115073  
##   
## $`Trt:DensHL`  
## diff lwr upr p adj  
## HL:L-HH:L -3.850596 -12.093566 4.3923729 0.7235473  
## LL:L-HH:L -2.625653 -10.868623 5.6173159 0.9279458  
## HH:H-HH:L -1.135395 -9.378364 7.1075746 0.9983008  
## HL:H-HH:L -5.351021 -13.593990 2.8919482 0.3881938  
## LL:H-HH:L -7.653179 -15.896148 0.5897904 0.0816781  
## LL:L-HL:L 1.224943 -7.018026 9.4679124 0.9975594  
## HH:H-HL:L 2.715202 -5.527768 10.9581711 0.9178248  
## HL:H-HL:L -1.500425 -9.743394 6.7425447 0.9936718  
## LL:H-HL:L -3.802582 -12.045552 4.4403868 0.7337902  
## HH:H-LL:L 1.490259 -6.752711 9.7332281 0.9938681  
## HL:H-LL:L -2.725368 -10.968337 5.5176017 0.9166238  
## LL:H-LL:L -5.027526 -13.270495 3.2154438 0.4571920  
## HL:H-HH:H -4.215626 -12.458596 4.0273430 0.6424237  
## LL:H-HH:H -6.517784 -14.760754 1.7251851 0.1907206  
## LL:H-HL:H -2.302158 -10.545127 5.9408115 0.9577891

## Is it valid?

resid <- mod.aov$residuals  
 qqnorm(resid)  
 qqline(resid)



hist(resid)

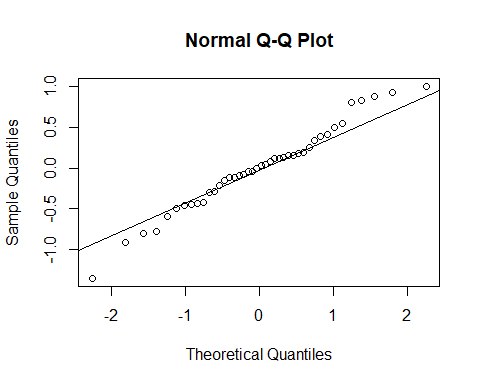


## Maybe a transformation is better?

mod.aov2 <- aov(log(mean\_length) ~ Trt\*DensHL, data = cuke.length.example)  
summary(mod.aov2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Trt 2 2.506 1.2532 4.192 0.0231 \*  
## DensHL 1 0.227 0.2270 0.759 0.3893   
## Trt:DensHL 2 0.682 0.3412 1.141 0.3307   
## Residuals 36 10.763 0.2990   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qqnorm(mod.aov2$residuals)  
 qqline(mod.aov2$residuals)



## Interpreting new results?

emmeans(mod.aov2, specs = ~Trt\*DensHL)

## Trt DensHL emmean SE df lower.CL upper.CL  
## HH L 2.32 0.207 36 1.90 2.74  
## HL L 1.96 0.207 36 1.54 2.38  
## LL L 2.03 0.207 36 1.62 2.45  
## HH H 2.41 0.207 36 1.99 2.82  
## HL H 1.94 0.207 36 1.52 2.36  
## LL H 1.53 0.207 36 1.11 1.95  
##   
## Results are given on the log (not the response) scale.   
## Confidence level used: 0.95

emmeans(mod.aov2, specs = ~Trt\*DensHL, type = 'response')

## Trt DensHL response SE df lower.CL upper.CL  
## HH L 10.21 2.109 36 6.71 15.52  
## HL L 7.10 1.468 36 4.67 10.80  
## LL L 7.65 1.580 36 5.03 11.63  
## HH H 11.08 2.291 36 7.29 16.85  
## HL H 6.96 1.438 36 4.57 10.58  
## LL H 4.63 0.956 36 3.04 7.03  
##   
## Confidence level used: 0.95   
## Intervals are back-transformed from the log scale

## ANOVA as a LM

* Factor 1: TOM HH, HL, LL
* Factor 2: Dens H, L

model.matrix(mean\_length ~ Trt\*DensHL, data = cuke.length.example)

## (Intercept) TrtHL TrtLL DensHLH TrtHL:DensHLH TrtLL:DensHLH  
## 1 1 1 0 0 0 0  
## 2 1 0 1 0 0 0  
## 3 1 1 0 1 1 0  
## 4 1 0 1 1 0 1  
## 5 1 0 0 1 0 0  
## 6 1 0 0 0 0 0  
## 7 1 0 0 1 0 0  
## 8 1 1 0 1 1 0  
## 9 1 0 0 0 0 0  
## 10 1 0 1 0 0 0  
## 11 1 1 0 0 0 0  
## 12 1 0 1 1 0 1  
## 13 1 0 0 1 0 0  
## 14 1 1 0 0 0 0  
## 15 1 0 1 0 0 0  
## 16 1 0 0 0 0 0  
## 17 1 0 1 1 0 1  
## 18 1 1 0 1 1 0  
## 19 1 0 0 0 0 0  
## 20 1 1 0 0 0 0  
## 21 1 1 0 1 1 0  
## 22 1 0 1 1 0 1  
## 23 1 0 1 0 0 0  
## 24 1 0 0 1 0 0  
## 25 1 0 1 1 0 1  
## 26 1 0 0 1 0 0  
## 27 1 0 0 0 0 0  
## 28 1 1 0 1 1 0  
## 29 1 0 1 0 0 0  
## 30 1 1 0 0 0 0  
## 31 1 0 0 0 0 0  
## 32 1 0 1 0 0 0  
## 33 1 1 0 1 1 0  
## 34 1 0 1 1 0 1  
## 35 1 1 0 0 0 0  
## 36 1 0 0 1 0 0  
## 37 1 0 0 0 0 0  
## 38 1 0 1 0 0 0  
## 39 1 1 0 1 1 0  
## 40 1 1 0 0 0 0  
## 41 1 0 1 1 0 1  
## 42 1 0 0 1 0 0  
## attr(,"assign")  
## [1] 0 1 1 2 3 3  
## attr(,"contrasts")  
## attr(,"contrasts")$Trt  
## [1] "contr.treatment"  
##   
## attr(,"contrasts")$DensHL  
## [1] "contr.treatment"

## ANOVA as LM

fit.lm <- lm(log(mean\_length) ~ Trt\*DensHL, data = cuke.length.example)

##   
## Call:  
## lm(formula = log(mean\_length) ~ Trt \* DensHL, data = cuke.length.example)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.35215 -0.29724 0.01347 0.24240 1.00336   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.32289 0.20666 11.240 2.5e-13 \*\*\*  
## TrtHL -0.36218 0.29226 -1.239 0.223   
## TrtLL -0.28858 0.29226 -0.987 0.330   
## DensHLH 0.08258 0.29226 0.283 0.779   
## TrtHL:DensHLH -0.10369 0.41332 -0.251 0.803   
## TrtLL:DensHLH -0.58514 0.41332 -1.416 0.165   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5468 on 36 degrees of freedom  
## Multiple R-squared: 0.2409, Adjusted R-squared: 0.1355   
## F-statistic: 2.285 on 5 and 36 DF, p-value: 0.06664

## Summarizing in an ANOVA way.

anova(fit.lm)

## Analysis of Variance Table  
##   
## Response: log(mean\_length)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Trt 2 2.5064 1.25320 4.1918 0.02309 \*  
## DensHL 1 0.2270 0.22699 0.7592 0.38934   
## Trt:DensHL 2 0.6824 0.34122 1.1413 0.33068   
## Residuals 36 10.7627 0.29896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Basic T Test as an LM

datHL <- data.frame(trt = rep(c("H", "L"), each = 7), length = c(high, low))  
 t.test(high, low, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: high and low  
## t = 5.7364, df = 12, p-value = 9.361e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 4.047097 9.004332  
## sample estimates:  
## mean of x mean of y   
## 11.328571 4.802857

## Basic T Test as an LM

fit.t <- lm(length ~ trt, data = datHL)  
 summary(fit.t)

##   
## Call:  
## lm(formula = length ~ trt, data = datHL)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3786 -1.4321 -0.3407 0.7296 5.3714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.3286 0.8044 14.083 7.98e-09 \*\*\*  
## trtL -6.5257 1.1376 -5.736 9.36e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.128 on 12 degrees of freedom  
## Multiple R-squared: 0.7328, Adjusted R-squared: 0.7105   
## F-statistic: 32.91 on 1 and 12 DF, p-value: 9.361e-05

## Next Steps

* Nearly everything is a ‘linear’ model.
* We will not use the aov function anymore, but instead us lm() or later glm()
* We will then distinguish between continuous covariates that fit an actual ‘slope’ vs categorical variables that are coded as a ‘factor’ in R.