ANOVA Homework

Luke Ghallahorne

2025-05-21

## Fish

We are interested in understanding salmon growth as fish migrate downstream. Our scientific hypothesis is that mean fish size increases with distance downstream because as salmon migrate from their natal location, they eat and grow, and thus, position in the river system would determine size. We are curious about where they seem to grow the most.

So, last April, we took our beach seine out and collected juvenile salmon in 3 habitats along a gradient in the Skagit River (on the same day): Tidal Freshwater (tf, most upstream), Delta (delta, middle), and Estuary (estuary, farthest downstream). We measured lengths of all the fish caught (fork length, in mm) in two seine hauls at each site. We would like to test our hypothesis of equal growth in fish at habitats along a river gradient.

### a. State your statistical hypotheses.

H0:

Null: Mean fork length will be equal at all 3 test habitats.

Ha1: and/or and/or

Alternative 1: Mean fork length in at least one habitat is not equal to the others.

Ha2:

Alternative 2: Mean fork length will be greater at sites further downstream than those upstream. (This requires analysis beyond an ANOVA, which can only determine if one or more of the means is different, not which or how much.)

### b. Generate a subsample of 25 fish from our fish data.

set.seed(50)  
 # set.seed creates reproducible RNG; code after the set.seed() will create the same array of random numbers the next time it is run. The seed number itself is arbitrary.  
  
#Generate subsample of Fish Data   
T<-round(rnorm(25, 29.7, 0.6), 1)  
D<-round(rnorm(25, 32.7, 0.8), 1)  
E<-round(rnorm(25, 33.5, 1.2), 1)  
 # rnorm(n, mean, sd)  
 # round(x, digits = 1)  
  
#Generate habitat identifiers  
t<-rep("tf", 25)  
d<-rep("delta", 25)  
e<-rep("estuary", 25)  
 # rep() repeats an input x a number of times  
 # rep(x, times)  
  
#Make a data frame  
habitat<-c(t, d, e)  
fl<-c(T,D,E)  
fish<-data.frame(habitat,fl)  
  
head(fish)

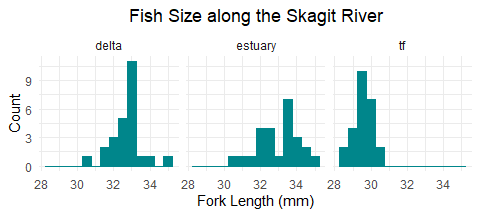
## habitat fl  
## 1 tf 30.0  
## 2 tf 29.2  
## 3 tf 29.7  
## 4 tf 30.0  
## 5 tf 28.7  
## 6 tf 29.5

dim(fish)

## [1] 75 2

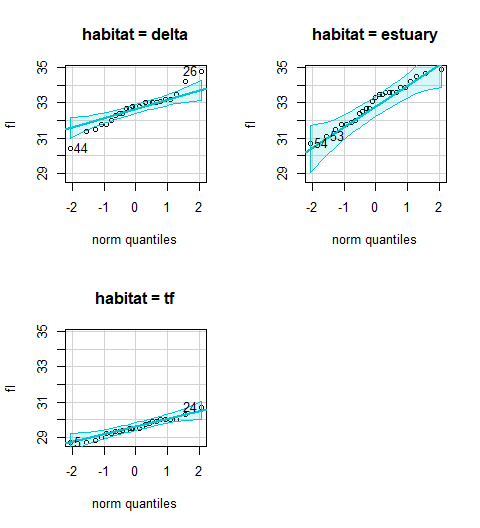
### c. Assessment of assumptions.

#### Testing normality: Histograms.



From looking at the histograms, fork length appears normally distributed.

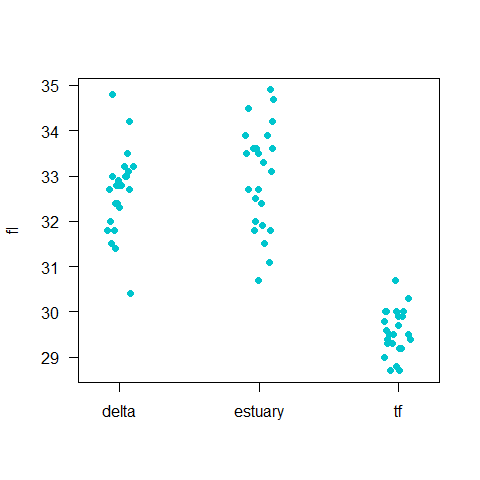
#### Testing normality: qqplots.



The qqplots of each habitat further indicate that the fork length data are normally distributed.

#### Testing for homoscedasticity.

I first created a stripchart of fork length by habitat to visualize the spread of each group.



Variances between the delta and estuary habitats are very similar; tidal freshwater has about half the variance as delta or estuary, visually.

I then calculated standard deviation for each habitat to compare variances.

tapply(fish$fl, fish$habitat, sd)

## delta estuary tf   
## 0.8952281 1.1343133 0.4899660

tf

delta

estuary

The standard deviations of each habitat are close enough to assume equal variances for analysis. I could conduct a Levene’s Test for Homogeneity of Variance but our sample size is rather small, and a test does not really feel necessary.

### d. Conduct an ANOVA.

Fish Data ANOVA Table:

## Df Sum Sq Mean Sq F value Pr(>F)   
## fish$habitat 2 180.47 90.23 116.3 <2e-16 \*\*\*  
## Residuals 72 55.88 0.78   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Degrees of freedom: 2, 72 F value: 116.3 p < 0.001

At least one mean is significantly different from the rest.

I compared the means of each habitat to see which differed from the others.

## delta estuary tf   
## 32.668 33.000 29.556

tf

delta

estuary

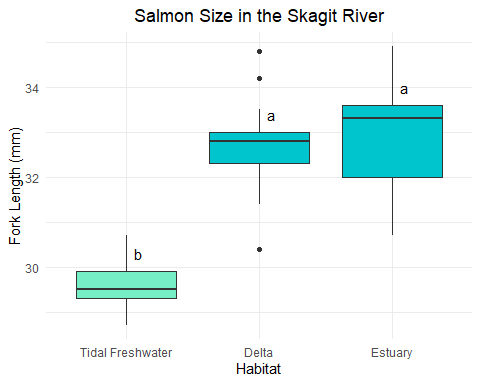
Tidal freshwater looks like the best candidate for the different mean, but I performed a Tukey HSD test to verify that.

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = fish$fl ~ fish$habitat)  
##   
## $`fish$habitat`  
## diff lwr upr p adj  
## estuary-delta 0.332 -0.2642887 0.9282887 0.3820299  
## tf-delta -3.112 -3.7082887 -2.5157113 0.0000000  
## tf-estuary -3.444 -4.0402887 -2.8477113 0.0000000

Fork lengths in the estuary and delta habitats both differed significantly from tidal freshwater (p < 0.001 for each), but do not differ from each other (p < 0.382).

### e. Conclusions.

With an , the ANOVA shows that mean fork length in at least one habitat of the Skagit River differs significantly from the others (2,72 = 116.3, p < 0.001). This provides evidence that we can reject the null hypothesis that mean lengths are equal across habitats.



Mean fork length at the tidal freshwater habitat (tf = 29.6 mm) was significantly lower than at the estuary (estuary = 33.0 mm, p < 0.001) or delta (delta = 32.7 mm, p < 0.001) habitats. However, fork length did not differ between the delta and the estuary (p < 0.382).

Since tidal freshwater is the most upstream habitat, most growth in salmon in the Skagit River occurs between there and the delta, but the fish experience little to no growth while migrating from the delta to the estuary.

## Fish with a Covariate

Refit the fish data, but add temperature as a covariate.

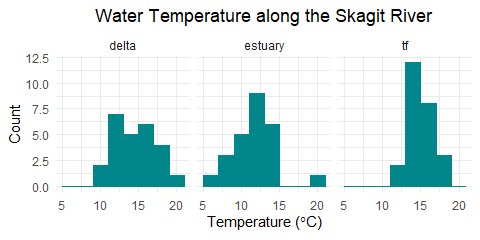
temp<-c(rnorm(25, 15, 1.5), rnorm(25, 14, 2.5), rnorm(25, 12, 2.5))   
fish2<-cbind(fish, temp)  
head(fish2)

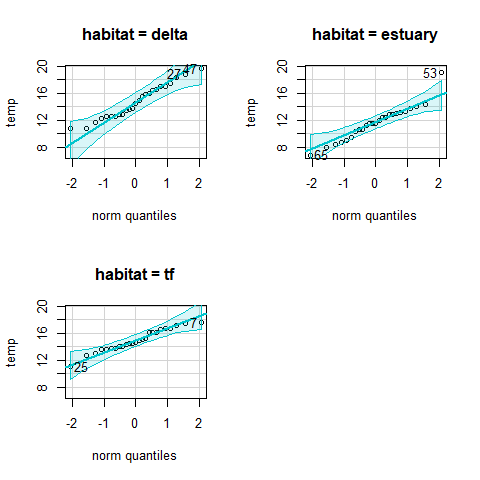
## habitat fl temp  
## 1 tf 30.0 14.02047  
## 2 tf 29.2 12.65899  
## 3 tf 29.7 13.67603  
## 4 tf 30.0 14.02871  
## 5 tf 28.7 13.57616  
## 6 tf 29.5 16.69032

### a. Carry through the analysis.

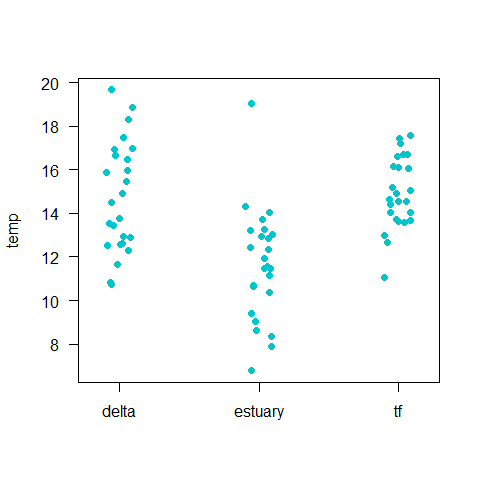
*I realized after going about the next section that I should have grouped temperature into categories for the 2-way ANOVA. I left in the first analysis to show my process - the final one with two temperature groups is at the end of this section.*

#### Check for normality and homoscedasticity of temperature.





From the histograms and qqplots, temperature is normally distributed enough for analysis with ANOVA.



tapply(fish2$temp, fish2$habitat, sd)

## delta estuary tf   
## 2.529255 2.555118 1.644046

tf

delta

estuary

Spread and standard deviations show that variances are sufficiently equal.

#### ANOVA.

str(fish2)

## 'data.frame': 75 obs. of 3 variables:  
## $ habitat: chr "tf" "tf" "tf" "tf" ...  
## $ fl : num 30 29.2 29.7 30 28.7 29.5 29.9 29.3 30.3 28.8 ...  
## $ temp : num 14 12.7 13.7 14 13.6 ...

fish2.aov <- aov(fl ~ habitat\*temp, data = fish2)  
summary(fish2.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## habitat 2 180.47 90.23 115.242 <2e-16 \*\*\*  
## temp 1 0.38 0.38 0.480 0.491   
## habitat:temp 2 1.47 0.74 0.942 0.395   
## Residuals 69 54.03 0.78   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Fork length is significantly affected by habitat (2,69 = 117.893, p < 0.001) but not by temperature (1,69 = 2.504, p < 0.118), and there is no significant interaction between habitat and temperature (2,69 = 0.750, p < 0.476). Adding temperature to the analysis had no effect on the conclusions that fish grow from one habitat to the next heading downstream.

I realized after conducting this 2-way ANOVA that because temperature is a continuous variable, an ANOVA like this is not an appropriate analysis. I separated the temperatures into two categories: warm (above the mean temperature for all habitats) and cold (below that mean).

summary(fish2$temp)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.77 12.32 13.68 13.76 15.94 19.67

fish3 <- fish2   
fish3$temp.cat <- cut(fish3$temp, breaks = c(6,13.98,21), labels = c("cold", "warm"))  
head(fish3)

## habitat fl temp temp.cat  
## 1 tf 30.0 14.02047 warm  
## 2 tf 29.2 12.65899 cold  
## 3 tf 29.7 13.67603 cold  
## 4 tf 30.0 14.02871 warm  
## 5 tf 28.7 13.57616 cold  
## 6 tf 29.5 16.69032 warm

fish3.aov <- aov(fl ~ habitat\*temp.cat, data = fish3)  
summary(fish3.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## habitat 2 180.47 90.23 117.596 <2e-16 \*\*\*  
## temp.cat 1 1.73 1.73 2.255 0.138   
## habitat:temp.cat 2 1.20 0.60 0.782 0.461   
## Residuals 69 52.94 0.77   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

With two categories of temperature, cold (below mean) and warm (above mean), there is no significant effect on fork length (1,69 = 2.780, p < 0.100) and no significant interaction between temperature and habitat (2,69 = 1.082, p < 0.345). There remains a significant effect on fork length by habitat (2,69 = 114.9, p < 0.001). Adding temperature as a covariate changes the degrees of freedom but not the outcome of the test.

After our discussion of ANCOVA and linear models in class, I decided to run the analysis with temperature as a covariate through a linear model to see if anything changed.

fish.lm<-lm(fl~habitat\*temp, data=fish2)   
summary(fish.lm)

##   
## Call:  
## lm(formula = fl ~ habitat \* temp, data = fish2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2900 -0.4365 0.0005 0.4768 2.2650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.13665 1.06586 29.213 <2e-16 \*\*\*  
## habitatestuary 2.24684 1.35736 1.655 0.102   
## habitattf -1.82594 1.96404 -0.930 0.356   
## temp 0.10405 0.07141 1.457 0.150   
## habitatestuary:temp -0.13704 0.10048 -1.364 0.177   
## habitattf:temp -0.08762 0.13103 -0.669 0.506   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8849 on 69 degrees of freedom  
## Multiple R-squared: 0.7714, Adjusted R-squared: 0.7548   
## F-statistic: 46.57 on 5 and 69 DF, p-value: < 2.2e-16

anova(fish.lm)

## Analysis of Variance Table  
##   
## Response: fl  
## Df Sum Sq Mean Sq F value Pr(>F)   
## habitat 2 180.466 90.233 115.2425 <2e-16 \*\*\*  
## temp 1 0.376 0.376 0.4798 0.4908   
## habitat:temp 2 1.474 0.737 0.9416 0.3950   
## Residuals 69 54.026 0.783   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The linear model brings me to the same conclusion as the ANOVA above - we fail to reject the null that temperature does not have an interacting effect on salmon fork length.

## ANOVA in the Wild.

### Article using ANOVA.

Zacherl DC, Morgan SG, Swearer SE, Warner RR. 2009. A shell of its former self: can Ostrea lurida Carpenter 1864 larval shells reveal information about a recruit’s birth location? Journal of Shellfish Research. 28(1): 23-32. <http://dx.doi.org/10.2983/035.028.0107> Retrieved from <https://escholarship.org/uc/item/4qd191nt>

This paper investigates the potential of using chemical tags in larval shells of Ostrea lurida as an indicator of birth location. Zacherl et al. compared the calcium-element ratios for seven metals after larvae were reared in differing concentrations, as well as changing ratios throughout larval development.

The researchers performed a two-way ANOVA for each element they tested, comparing seawater chemistry, ontogeny, and the interaction between seawater and ontogeny. For magnesium, seawater had 2,6 = 1.72, p < 0.257; ontogeny had 1,6 = 100.83, p < 0.001; and seawater-ontogeny interactions had 2,6 = 0.86, p < 0.468. Seawater had 3 groups, ontogeny 2. Samples were 7 across the experiment.

The question was to determine if there is a difference between the elements taken in from seawater when *O. lurida* larvae are being brooded versus when they are pelagic in the plankton. They found that seawater concentration changes around the larvae do not significantly alter shell composition (2,6 = 1.72, p < 0.257), but shell chemistry does change as development progresses in the planktonic phase (1,6 = 100.83, p < 0.001). There was no significant effect from the interaction between seawater and ontogeny (2,6 = 0.86, p < 0.468).