IEP Workshop Training: Analysis of Variance (ANOVA), Kruskal-Wallis, and Multiple Comparisons

Emily Ryznar

3/23/2022

1 Foundational background

1.1 General

1. Analysis of Variance (ANOVA) and Kruskal-Wallis are statistical analyses suitable for data with a continuous response variable and categorical predictor variables (often called "factors") with >2 predictor categories (often called "groups"; Fig. 1).

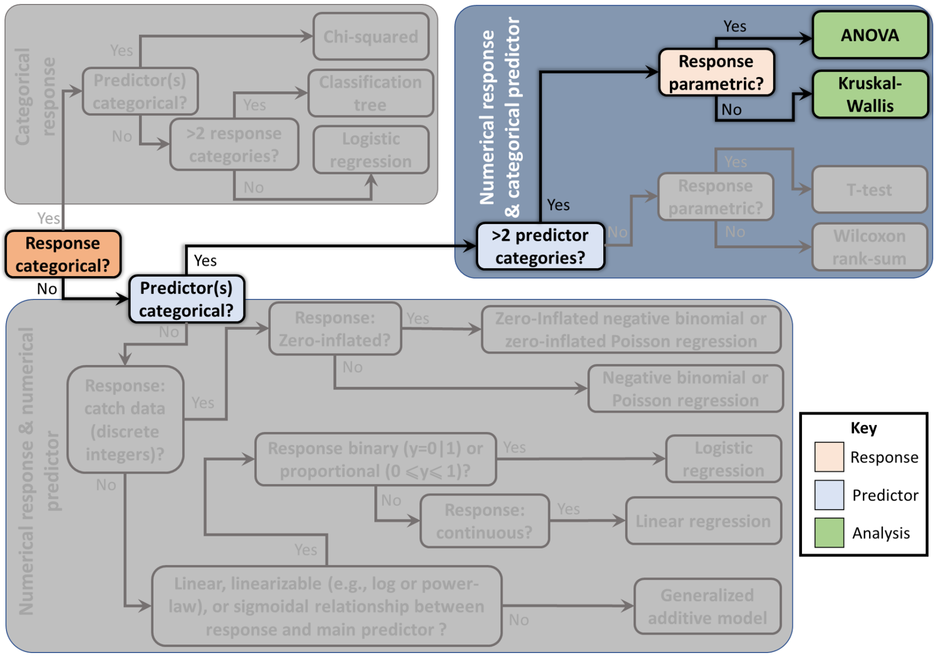
Categorical response
Chi-squared
Yes
Predictor(s) categorical?
Classification tree
Yes
No
>2 response categories?
Logistic regression
No
Numerical response & categorical predictor
Yes
ANOVA
Response parametric?
No
Kruskal-Wallis
Yes
T-test
Response parametric?
Yes
No
No
Wilcoxon rank-sum
>2 predictor categories?
Yes
Response categorical?
Yes
No
Predictor(s) categorical?
No
Numerical response & numerical predictor
Response: Zero-inflated?
Yes
Zero-Inflated negative binomial or zero-inflated Poisson regression
No
Negative binomial or Poisson regression
Response: catch data (discrete integers)?
Yes
No
Response binary (y=0|1) or proportional (0 ⩽y⩽ 1)?
Yes
Logistic regression
No
Response: continuous?
Yes
Linear regression
Yes
Linear, linearizable (e.g., log or power-law), or sigmoidal relationship between response and main predictor ?
No
Generalized additive model
Key
Response
Predictor
Analysis

Figure 1: Univariate statistical analysis flowchart with pathways to Analysis of Variance (ANOVA) and Kruskal-Wallis highlighted.

2. A simple example of a analytical goal that would be well suited to this type of analysis would be to test whether mean fish length ("fork length") varies among Chinook salmon life stages (e.g., groups such as fry, parr, and smolt; Fig. 2).

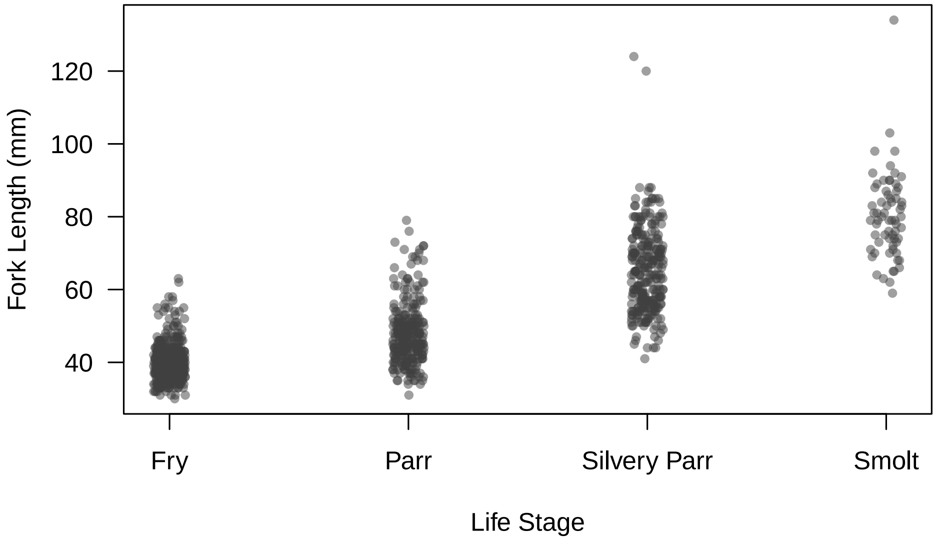

Figure 2: Raw fork length data by Chinook salmon life stage (Gilbert et al. 2021)

1.1.1 Terminology

When referencing ANOVAs and Kruskal-Wallis, you may come across terminology that is common for these kinds of analyses. There is the dependent response variable and independent categorical predictor variable that is often called a *factor*. Each *factor* can have multiple *groups* or *predictor categories* within a factor (also frequently referred to as *levels*). Within each group, there are multiple independent values, which can be observations, samples, replicates, etc., depending on your experimental design. Finally, each *factor* x *group* combination is frequently referred to as a *treatment*. For the purpose of this exercise, we will be using "response variable" for our dependent variable, "predictor variable" for our categorical independent variable (unless when referring to a 1-Factor ANOVA), and "group" for the different categories within our predictor variable, and "observations" for independent values within each group.

1.1.2 ANOVA

ANOVA models require a numerical response and categorical predictor variables (i.e., factors; Fig. 1) and evaluate whether the mean responses are equal among groups when there are more than two groups in a predictor variable. By comparing means between groups, an ANOVA takes into account the variance around the mean (hence the "Analysis of Variance"). An ANOVA evaluates whether variability between groups is greater than or equal to variability within groups and whether between-group variability is due to the predictor variable(s) and not random.

1.1.3 Kruskal-Wallis

Kruskal-Wallis (also called the Kruskal-Wallis H test because a "H-statistic" is calculated) models can be used as a nonparametric ANOVA equivalent when assumptions for parametric statistics are not met, there is only one categorical predictor variable (Fig. 1; Hollander and Wolfe 1973), and the response is numeric. Broadly, a Kruskal-Wallis analysis functions on data value ranks instead of the data points themselves and evaluates whether the median responses of two or more groups are different.

1.2 Model assumptions

1.2.1 ANOVA

To utilize a parametric model such as an ANOVA, our data needs to meet three assumptions. First, the residuals (experimental errors, or all of the variation that is not explained by our predictor variables) of our response variable should be normally distributed. Second, residual variance of the response variable should be equal between groups. Finally, an ANOVA assumes that observations in one group are independent from observations in another group.

1.2.2 Kruskal-Wallis

A Kruskal-Wallis also requires a numerical response and categorical predictor variables with more than two groups. While Kruskal-Wallis doesn't require that response data follow a normal distribution, it does assume that observations within each group of the predictor variable follow a similar distribution. If not, a Kruskal-Wallis can compare mean instead of median ranks (Hollander and Wolfe 1973). As with an ANOVA, Kruskal-Wallis assumes independent observations.

1.3 Model notation

1.3.1 ANOVA

An ANOVA is a specific type of linear regression without a continuous term. Therefore, following a linear regression framework, an ANOVA can be modeled as follows:

$$y_i = \alpha + \beta X + \varepsilon_i \tag{1}$$

where α is the coefficient of our first group, βX is a matrix of potential coefficients (β) and predictors (X) for the remaining groups, ε_i is error or uncertainty (residuals) not captured by the model, and i indicates the observation number for $i = 1, 2, ..., n$. Depending on which group is being predicted, X is coded as one or zero. For example, if we were predicting the mean response of group one, $X = 0$ and $y_i = \alpha + \varepsilon_i$. Further, if we were predicting the mean of group two, $X = 1$ and $y_i = \alpha + \beta * 1 + \varepsilon_i$. It is assumed that $\varepsilon_i \sim\ N(0, \sigma^2)$, meaning that the residuals ε_i are normally distributed with a mean of zero and exhibit an equal variance of σ^2.

Further, various summary statistics can be calculated for an ANOVA such as degrees of freedom, sum of squares, mean squares, and importantly, the F-statistic and F-critical value, which represent how much variability among group means exceeds what is expected by chance. For an in-depth description for how to calculate ANOVA summary statistics, see Quinn and Keough (2002).

1.3.2 Kruskal-Wallis

As a general overview, the first step in a Kruskal-Wallis is to pool the data across all groups and order the values from smallest to largest. From that ordering, you can then assign ranks to those values (e.g., smallest value=1, second smallest=2, etc.) and add up the different ranks for each group of your predictor variable. As such, Kruskal-Wallis does not produce coefficients and instead ranks the values and outputs an answer. Due to this, typical notation for a Kruskal-Wallis is for calculating the H-statistic (similar to an ANOVA F-statistic) and is not a model notation (a shortcoming of this approach). The equation for calculating the H-statistic is as follows:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \tag{2}$$

where N is the total number of observations across all predictor variables and groups, R_i is the sum of ranks in group i, and n_i is the number of observations in group i.

We can take the degrees of freedom for our data, calculated as one less than the total number of groups and our designated alpha value (i.e. the threshold of significance) to find the critical chi-square value in a chi-square probability table. If the chi-square value is less than the H-statistic calculated above, there is a significant difference between group medians.

1.4 Recommended reference material

If you wish to dive deeper into ANOVAs, nonparametric alternatives such as the Kruskal-Wallis test, and multiple comparisons, I recommend the following references:

- "Experimental Design" by Roger Kirk, 3rd Edition
- "Designing Experiments and Analyzing Data: A Model Comparison Perspective" by Scott Maxwell and Harold Delaney, 2nd Edition
- "Experimental Design and Data Analysis for Biologists" by Gerry Quinn and Michael Keough, 1st Edition
- "Nonparametric Statistical Methods" by Myles Hollander and Douglas A. Wolfe

These provide great overviews of experimental design (as ANOVAs and Kruskal-Wallis tests are frequently used to analyze experimentally-derived data) and corresponding statistical methods. All can be found on Amazon, and I had success finding most in my university library.

2 Analysis

Throughout this exercise (and throughout the 2022 IEP statistics micro-training), we will be referring to and utilizing the general analysis framework outlined in Fig. 3. Ideally, one would start with a question and figure out what type(s) of data should be collected to best address that question. Once you've collected your data, you should explore it for missing values, outliers, etc., apply your model of choice, evaluate and validate model results, and interpret and present model findings, iterating through earlier steps where necessary.

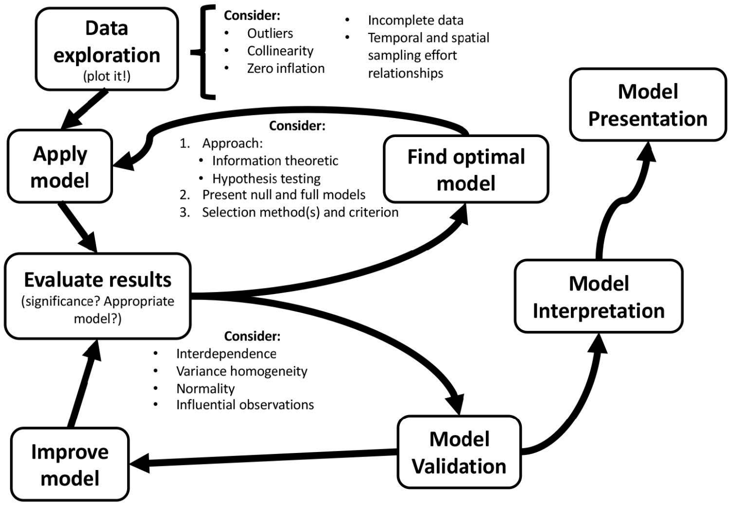

Figure 3: A general statistical modeling framework

2.1 Case Study Data

We will be using a publicly available dataset collected as part of the US Fish and Wildlife Service and IEP's Delta Juvenile Fish Monitoring Program (DJFMP) at Liberty Island. Specifically, we will be analyzing beach seine data for young-of-year fishes collected at sampling locations throughout Liberty Island from 2002-2004 and 2013-2019 (Gilbert et al. 2021).

Let's take a quick look at the data:

```r
##Import data
data<- read.csv(file="LI_seine.csv", header=T)

##View data
head(data)
```

```
##   StationCode SampleDate SampleTime GearConditionCode WeatherCode DO
## 1      LI001W  9/17/2003   11:53:00                 1         CLR NA
## 2      LI001W  9/17/2003   11:53:00                 1         CLR NA
## 3      LI001W  9/17/2003   11:53:00                 1         CLR NA
## 4      LI001W  9/17/2003   11:53:00                 1         CLR NA
## 5      LI001W  9/17/2003   11:53:00                 1         CLR NA
## 6      LI001W  9/17/2003   11:53:00                 1         CLR NA
##   WaterTemperature Turbidity Conductivity SiteDisturbance AlternateSite
## 1             18.9        NA        194.7            <NA>             N
## 2             18.9        NA        194.7            <NA>             N
## 3             18.9        NA        194.7            <NA>             N
## 4             18.9        NA        194.7            <NA>             N
## 5             18.9        NA        194.7            <NA>             N
## 6             18.9        NA        194.7            <NA>             N
##   SeineLength SeineWidth SeineDepth Volume OrganismCode        CommonName
## 1          12         15        0.9     81           LP Bigscale Logperch
## 2          12         15        0.9     81          MSS Inland Silverside
## 3          12         15        0.9     81          MSS Inland Silverside
## 4          12         15        0.9     81          MSS Inland Silverside
## 5          12         15        0.9     81          MSS Inland Silverside
## 6          12         15        0.9     81          STB      Striped Bass
##   MarkCode StageCode Maturation ForkLength Count
## 1     None        NA       <NA>         70     1
## 2     None        NA       <NA>         27     2
## 3     None        NA       <NA>         28     2
## 4     None        NA       <NA>         29     1
## 5     None        NA       <NA>         45     1
## 6     None        NA       <NA>         79     1
```

There are numerous variables in the dataset, but for purposes of our analysis, we will only be focusing on the variables below:

- *StationCode*: sampling station code
- *SampleDate*: date in MM/DD/YYYY format
- *CommonName*: common name of species sampled
- *StageCode*: salmonid development index (life stage)
- *ForkLength*: length of fish (mm)

Full metadata, the salmonid development index key (for life stages), and other datasets from the DJFMP at Liberty Island can be found here.

2.2 Question

Let's say there was a disease outbreak in the early 2000s that stunted the growth of older life history stages of Chinook salmon, resulting in similar lengths among young Chinook stages and older stages. However, it is unclear whether this disease impacted Chinook migrating through the SF Estuary. By subsetting our data to only include Chinook salmon and a year in the early 2000s (in our case, 2004), we can use a 1-Factor ANOVA (or Kruskal-Wallis) to evaluate the following question:

1) Are fork lengths similar among Chinook salmon life history stages in 2004?

Therefore, based on our hypothetical scenario, if mean fork length is similar among life stages in 2004, the disease is present. In this case, "life stage" would be our only predictor variable (our single "factor") with four groups ("fry", "parr", "silvery parr", and "smolt").

2.3 Preparing data for analysis

Before we dive into our analysis, lets prepare our data and explore it for outliers and/or missing values.

First, notice how the current date format combines the month, day, and year. Let's pull out the years using the "mdy()" and "year()" functions in the package *lubridate* and store them in a new column called "SampleYear" which we will use later.

```r
##Load "lubridate" package
library(lubridate)

##Change SampleDate format so recognizable by "lubridate"
data$SampleDate<-mdy(data$SampleDate)

##Pull out sample year from SampleDate, storing in a new column
data$SampleYear<-year(data$SampleDate)
```

Since we are only interested in Chinook salmon, let's subset the data to omit all other species (since there are multiple species represented by the data) and only include data collected in 2004. Also, for the purposes of the analysis, let's only include Chinook life stages 2-5 (stages 1-5 are represented in the data). Once we have the data subset, we can add more informative labels corresponding to each life stage to aid our analysis.

```r
##Subset data for only Chinook salmon stages 2:5 in 2004
ch_2004<-subset(data,data$StageCode!="1" & data$CommonName=="Chinook Salmon"
                & data$SampleYear=="2004")

##Generate data frame of labels for corresponding stage codes
labels = data.frame(StageCode = 2:5,
                    stage = c("Fry", "Parr", "Silvery parr", "Smolt"))

##Merge labels data frame with salmon data frame
ch_2004<-merge(ch_2004, labels, all.x=TRUE)
```

Now that we have our data subset for Chinook salmon stages 2:5 in 2004 and labeled the life stages, let's take a look at our raw data.

```r
##Plot raw data by stage
plot(ForkLength ~ jitter(StageCode, factor=1/3), pch=20,
     col=gray(0.25,0.5), data=ch_2004, xaxt="n", las=1,
     ylab="Fork Length (mm)", xlab="Life Stage")
axis(side = 1, at=2:5,
     labels = c("Fry", "Parr", "Silvery Parr", "Smolt"))
```

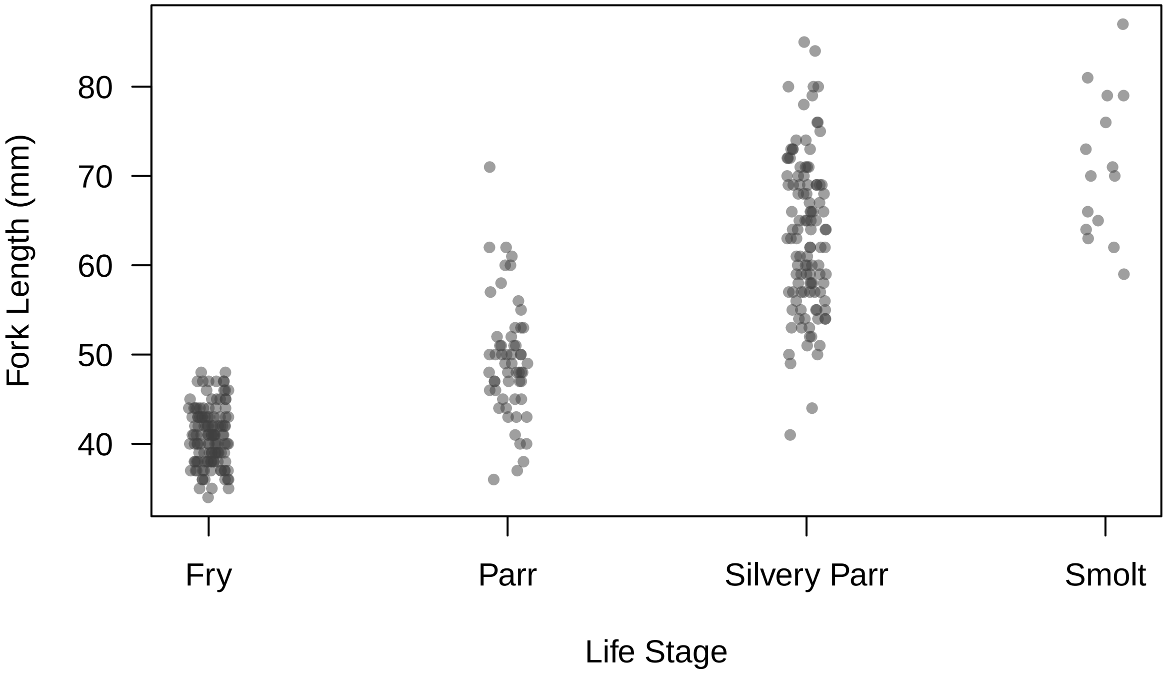

Figure 4: Raw fork length data by Chinook salmon stage in 2004 (from Gilbert et al. 2021)

In Fig. 4, notice that there is one point for parr fork length and one to two points for silvery parr fork length that don't necessary group with the rest. These may be considered data outliers, and if we had the raw datasheets (and good data collection notes), we might be able to justify their removal, particularly if our analytical results don't pass visual inspection. But, since we don't have that luxury, let's proceed.

Finally, let's check for missing values.

```
##Check for missing values
unique(is.na(ch_2004$stage)) #false=none!
```

```
## [1] FALSE
```

Great! No missing values. We now have data ready for analysis.

2.3 1-Factor ANOVA

2.3.1 Analysis

Let’s make sure “stage” in our subset data frame is recognized as a factor (otherwise the analysis won’t work). Then we can perform our 1-factor ANOVA as a linear model using the “lm()” function in base R. Note: an ANOVA can also be run using the “aov()” function in base R.

```
##Make sure "stage" is recognized as a factor
ch_2004$stage <- as.factor(ch_2004$stage)

##Running 1-factor ANOVA
un_test<-lm(ForkLength~stage, ch_2004) #untransformed

test<-lm((1/ForkLength)~stage, ch_2004) #1/x transformed
```

You can see above that there are two, 1-Factor ANOVAs, one on untransformed fork length data (“un_test”) and another on $\frac{1}{x}$ transformed fork length data (“test”). The transformation helped fitted model residuals meet parametric assumptions (see below). If your response does not meet ANOVA assumptions untransformed, you will want to try various transformations (e.g., $\sqrt{x}$, x^2, $log(x)$, and $\frac{1}{x}$) as we did. If no transformations help your data meet parametric assumptions necessary to run a valid ANOVA, a nonparametric equivalent such as Kruskal-Wallis should be utilized (see section 2.4 below).

2.3.2 Model validation

Let’s validate how well our data meets the ANOVA assumptions of normality and variance homogeneity, using our untransformed and $\frac{1}{x}$ transformed data for comparison. Both assumptions can be assessed by visual inspection and/or statistical analysis of the model’s residuals. For visual inspection, histograms and box plots are useful tools for assessing normality and variance assumptions, respectively.

```
##Specify plotting setup
par(mfrow=c(2,2), mar=c(4.1,4.1,0.5,0.5))

##Histogram for untransformed normality
hist(un_test$residuals, main=NULL, cex.axis=0.85, cex.lab=0.85, ylab="", xlab="")
title(ylab="Frequency", xlab="un_test$residuals",line=2, cex.lab=0.85)
box(which = "plot")
text(-24, 90, "(a)")

##Boxplot for untransformed variance
plot(un_test$residuals~ch_2004$stage, cex.axis=0.85, cex.lab=0.85, ylab="", xlab="")
title(ylab="un_test$residuals", xlab="Stage",line=2, cex.lab=0.85)
text(0.6, 18, "(b)")

##Histogram for 1/x transformed normality
hist(test$residuals, main=NULL, cex.axis=0.85, cex.lab=0.85, ylab="", xlab="")
title(ylab="Frequency", xlab="test$residuals",line=2, cex.lab=0.85)
box(which = "plot")
text(-0.0075, 103, "(c)")

##Boxplot for 1/x transformed variance
plot(test$residuals~ch_2004$stage, cex.axis=0.85, cex.lab=0.85, ylab="", xlab="")
```

```
title(ylab="test$residuals", xlab="Stage",line=2, cex.lab=0.85)
text(0.6, 0.0073, "(d)")
```

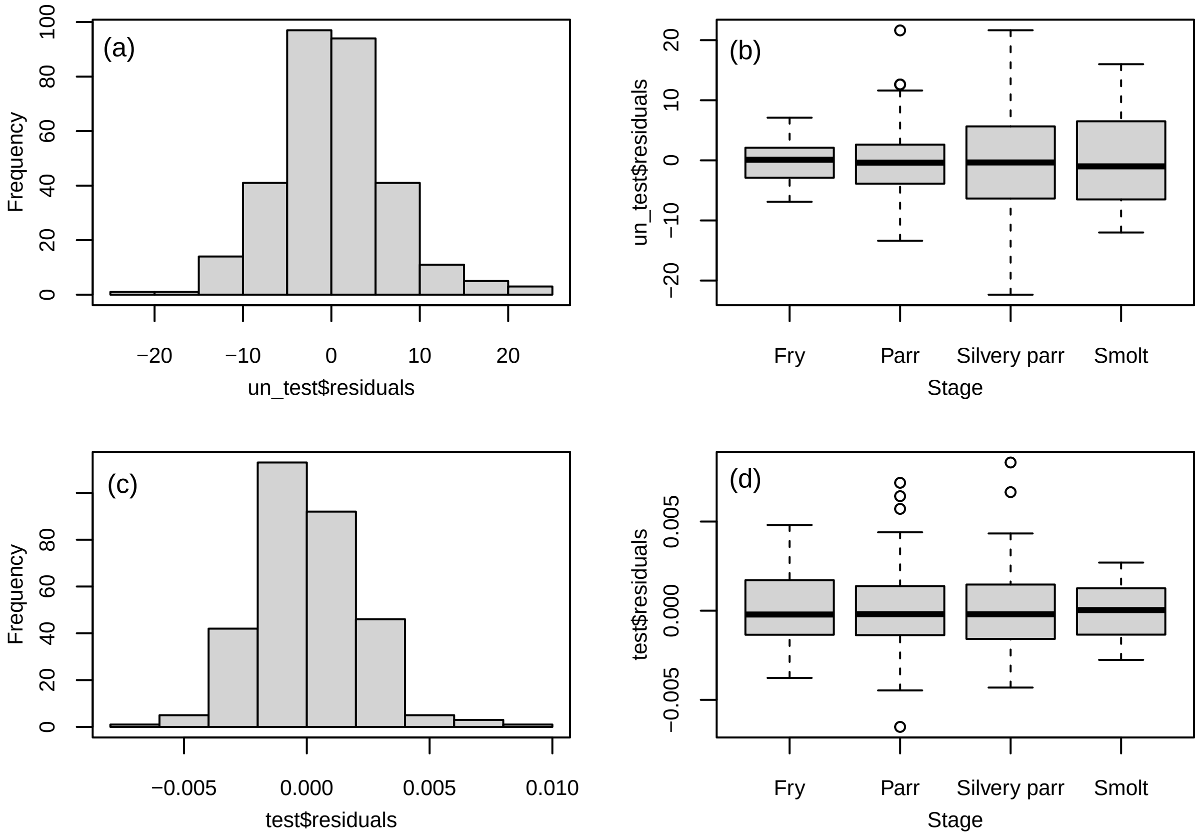

Figure 5: Checking ANOVA assumptions, with histograms (a) and (c) assessing normality and box plots (b) and (d) assessing homogeneity of variance. Untransformed data used for (a) and (b), 1/x transformed data used for (c) and (d).

As we can see in Fig. 5a, our residuals (using untransformed data) appear to be normally distributed as they follow the characteristic “bell shape” curve. However, in Fig. 5b, we can see that our variances (using untransformed data) may be unequal among groups, as evidenced by differing box plot “lengths”. For example, you can see that silvery parr variance is much greater than fry variance. Ideally, we would want all of the variance “lengths” to be relatively equal among groups, so we’d want to try transforming our response data. Once we apply a $\frac{1}{x}$ transformation, we can see that model residuals still generally follow a normal distribution (Fig. 5c) and now the variance among groups appears relatively equal (Fig. 5d).

If we didn’t trust our visual inspection skills, we could also use statistical methods to evaluate whether our data meets ANOVA assumptions. The Shapiro-Wilk’s (Shapiro and Wilk 1965) test is generally the recommended statistical method to evaluate normality and provides better power than other tests such as the Kolmogorov-Smirnov normality test (Ghasemi and Zahediasl 2012).

To perform a Shapiro-Wilk’s test, we will use the “shapiro.test()” base R function to statistically test whether our model residuals are significantly different than the normal distribution for our untransformed and transformed data.

```
##Perform Shapiro-Wilk's test
shapiro.test(un_test$residuals) #untransformed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  un_test$residuals
## W = 0.98243, p-value = 0.0007962
```

```
shapiro.test(test$residuals) #transformed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  test$residuals
## W = 0.98752, p-value = 0.009322
```

You can see that the p-values for both untransformed and transformed data are significant from the Shapiro-Wilk's test, meaning model residuals are not normally distributed. However, transforming our data gets us closer to non-significance, and ANOVAs are generally robust to slight variations in normality (Quinn and Keough 2002).

Now, let's statistically assess whether our model residuals exhibit equal variance. Two common statistical tests are Bartlett's test (Bartlett 1937) and Levene's test (Levene 1960). Both are used to compare variances of two or more groups (in our case, "stage"), whereas Levene's test is more robust if the data is not normally distributed. To perform these tests, we will use the "bartlett.test()" function in base R and the "leveneTest()" function in the **car** package and again compare between our untransformed and transformed data.

```
##Load "car" package
library(car)

##Untransformed
bartlett.test(un_test$residuals~stage, ch_2004)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  un_test$residuals by stage
## Bartlett's K-squared = 99.124, df = 3, p-value < 2.2e-16
```

```
leveneTest(un_test$residuals~stage, ch_2004)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   3  25.902 5.997e-15 ***
##       304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Transformed
bartlett.test(test$residuals~stage, ch_2004)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  test$residuals by stage
## Bartlett’s K-squared = 11.274, df = 3, p-value = 0.01033
```

```
leveneTest(test$residuals~stage, ch_2004)
```

```
## Levene’s Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  1.6716 0.1731
##       304
```

For our untransformed data, we can see that p-values from both Bartlett’s and Levene’s test are significant, meaning residual variance is not equal among stages, a violation of the assumption. However, residual variance from an ANOVA using $\frac{1}{x}$ transformed data is close to equal according to Bartlett’s test (p=0.01), and equal according to a non-significant p-value from Levene’s test. A non-significant p-value (and therefore equal variance) result for Bartlett’s or Levene’s test is generally satisfactory.

2.3.3 Analysis interpretation

So, we’ve concluded that $\frac{1}{x}$ transformed data generally meets assumptions for parametric statistics and therefore, our 1-Factor ANOVA using this data is the most valid. Now, we can dive a bit deeper into our 1-Factor ANOVA that uses transformed data, starting with viewing a summary of our model output.

```
##View results
summary(test)
```

```
##
## Call:
## lm(formula = (1/ForkLength) ~ stage, data = ch_2004)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0065204 -0.0015857 -0.0002105  0.0014654  0.0083118
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.0246058  0.0001965  125.21   <2e-16 ***
## stageParr         -0.0040008  0.0003533  -11.32   <2e-16 ***
## stageSilvery parr -0.0085273  0.0002859  -29.83   <2e-16 ***
## stageSmolt        -0.0103537  0.0006004  -17.25   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.002197 on 304 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.766
## F-statistic: 336.1 on 3 and 304 DF,  p-value: < 2.2e-16
```

The model output provides estimates of each of our coefficients ("fry"=(Intercept); not true means because we used transformed data) and whether each coefficient is significant (all are significant). The p-value for each of our coefficients basically tells us whether the difference between the coefficients of the intercept ("fry" in our case) and each subsequent group ("parr", "silvery parr", "smolt") are significantly different from zero. The model also provides our residual variance σ ("residual standard error"), the R-squared value, or how much of the total variation is explained by the model (higher the better; 77% is pretty good!), the F-statistic on 3 (group df: total # of groups-1), and 304 (residual df: total # of observations - total number of groups), and the p-value of the model. Our model p-value is <0.05, indicating that yes, fork length did differ among Chinook salmon life history stages in 2004 (no hypothetical disease!).

2.3.4 Multiple comparisons and interpretation

We can explore pairwise differences in fork length among our different stages using **multiple comparisons**. When conducting multiple comparisons, one should control for increasing experimentwise error rates that result from testing multiple hypotheses simultaneously (i.e., conducting pairwise comparisons of group means). Adjusting for multiple comparisons basically makes your level of significance more stringent, and there are different adjustment methods based on your data and how conservative you want to be. Chen et al. (2017) provides a great overview of adjusting for multiple comparisons and methods to do so.

A common test (and adjustment method) for all pairwise comparisons following an ANOVA is the Tukey Honest Significant Difference (Tukey HSD) test. This can be performed using the "TukeyHSD()" function in the **stats** package, or using the "ghlt()" function in the **multcomp** package. I prefer the latter because it outputs a compact letter display, which helps visualize differences.

```
##Load "multcomp" package
library(multcomp)

##Assess pairwise differences
tuk <- glht(test, linfct=mcp(stage="Tukey"))

##View pairwise summary
summary(tuk)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = (1/ForkLength) ~ stage, data = ch_2004)
##
## Linear Hypotheses:
##                              Estimate Std. Error t value Pr(>|t|)
## Parr - Fry == 0            -0.0040008  0.0003533 -11.324   <0.001 ***
## Silvery parr - Fry == 0    -0.0085273  0.0002859 -29.830   <0.001 ***
## Smolt - Fry == 0           -0.0103537  0.0006004 -17.245   <0.001 ***
## Silvery parr - Parr == 0   -0.0045265  0.0003596 -12.588   <0.001 ***
## Smolt - Parr == 0          -0.0063528  0.0006388  -9.945   <0.001 ***
## Smolt - Silvery parr == 0 -0.0018263  0.0006041  -3.023   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
#View compact letter display of pairwise summary
cld(tuk)
```

```
##          Fry         Parr Silvery parr        Smolt
##          "d"          "c"          "b"          "a"
```

From our pairwise difference test, we can see that all pairwise combinations are significantly different from each other. This is further visualized in our compact letter display where no group shares a letter, meaning all groups are significantly different in their fork lengths.

2.4 Kruskal-Wallis

What happens if your response data still doesn't meet parametric assumptions for an ANOVA, despite transformation?

For simplification purposes, let's use our untransformed fork length data from above and assume we weren't able to eventually transform it to meet parametric assumptions. We can use a Kruskal-Wallis to evaluate the same question we investigated with our 1-Factor ANOVA: "Are fork lengths similar among Chinook salmon life history stages in 2004?"

2.4.1 Pre-analysis assumption checking

First, let's check to see if our fork length data follow a similar distribution among groups, a Kruskal-Wallis assumption.

From Fig. 6, it seems like the fork length distributions among stages are *roughly* similar (though the parr distribution (Fig. 6b) may be pushing it). However, let's proceed with our analysis.

2.4.2 Analysis and interpretation

There may be alternatives, but here we will use the function "kruskal.test()" from the package **FSA**, which follows the same syntax as the "aov()" function above.

```
##Load "FSA" package
library(FSA)

##Run Kruskal-Wallis test
kw_test<-kruskal.test(ForkLength~stage, ch_2004)

##View results
kw_test
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ForkLength by stage
## Kruskal-Wallis chi-squared = 231.52, df = 3, p-value < 2.2e-16
```

Instead of the F-statistic in our ANOVA, we instead have the H-statistic ("Kruskal-Wallis chi-squared") and degrees of freedom as before. We can see that our p-value is significant, indicating fork lengths differed among Chinook life history stages in 2004. However, as with a 1-Factor ANOVA, our output does not show *where* the differences are, which calls for a multiple comparisons test.

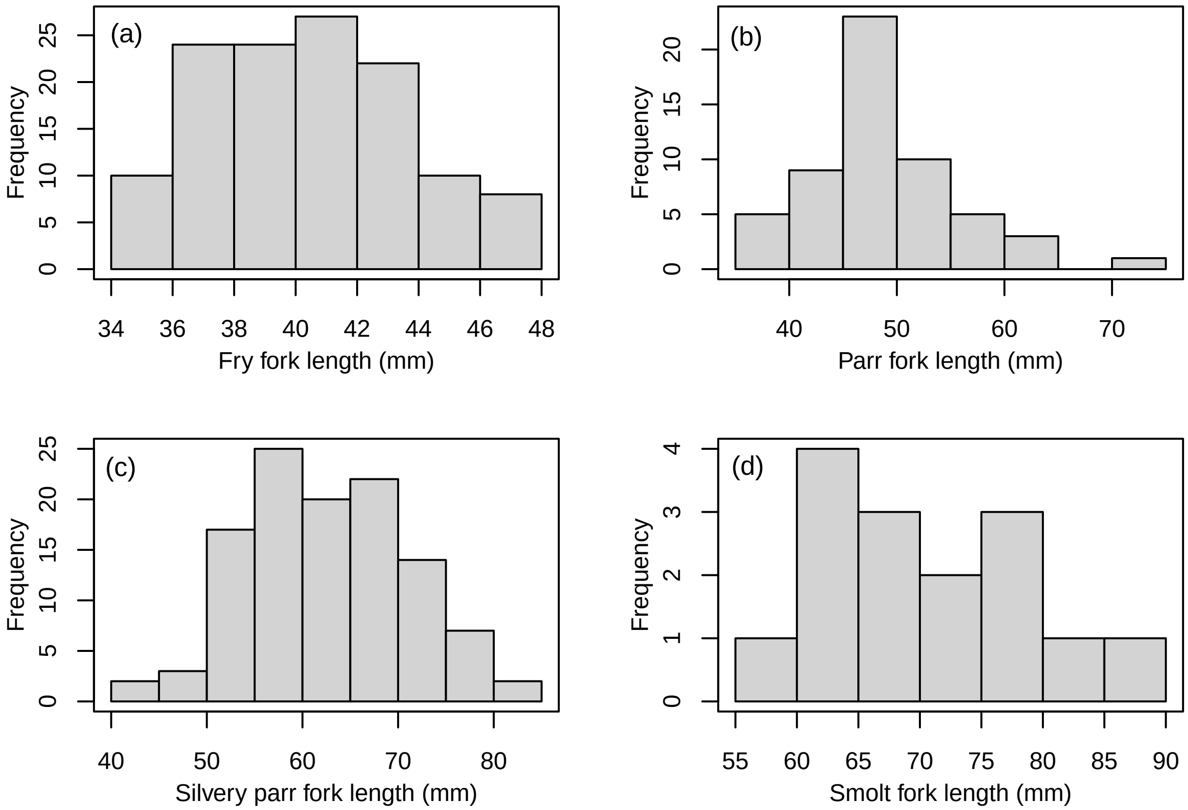

Figure 6: Evaluating Kruskal-Wallis assumptions of similar fork length (mm) distributions among Chinook stages, with (a)=fry, (b)=parr, (c)=silvery parr, and (d)=smolt.

2.4.3 Multiple comparisons and interpretation

A common test to assess pairwise differences following a Kruskal-Wallis is the Wilcoxon test (Hollander and Wolfe 1973), controlling for multiple comparisons. To perform this, we will use the "pairwise.wilcox.test()" function in the **stats** package. Note: this function has various options to control for multiple comparisons. We will use the Benjamini and Hochberg (1994) method ("BH").

```
##Load "stats" package
library(stats)

##Run pairwise Wilcox test with BH correction
pairwise.wilcox.test(ch_2004$ForkLength, ch_2004$stage, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  ch_2004$ForkLength and ch_2004$stage
##
##              Fry     Parr    Silvery parr
## Parr         < 2e-16 -       -
## Silvery parr < 2e-16 < 2e-16 -
## Smolt        3.7e-10 1.3e-08 0.0026
##
## P value adjustment method: BH
```

From our pairwise comparisons, we can see that all stages had fork lengths that were significantly different from one another in 2004.

3 Final model presentation

How do we present the results of our ANOVA and Kruskal-Wallis? Depending on the complexity and number of models you wish to report, it can be useful to report them both textually and visually.

3.1 ANOVAs

ANOVAs can be reported in-text by including the factor and residual degrees of freedom, the F-value, and the p-value. For example, to report the results of our 1-Factor ANOVA, we could say:

"Fork length was significantly different among Chinook salmon life history stages in 2004 (ANOVA, F(3,304)=336.1, p<0.001), and multiple comparisons revealed significant differences among all stages"

You can also represent your ANOVAs graphically. Boxplots are useful to visualize and compare variance among groups (Fig. 7). Adding the mean fork length (blue points) for each stage illustrate even more information related to our ANOVA. If the mean value for one stage falls outside the interquartile range (shaded region) of another stage, it is assumed those groups are significantly different. As we can see in Fig. 7, none of the means (blue points) for each stage overlap with the interquartile range of other stages, meaning they are all significantly different. This corresponds to our ANOVA. If applicable, you can add your multiple comparison post-hoc letters to further help the reader visualize pairwise differences, as in Fig. 7.

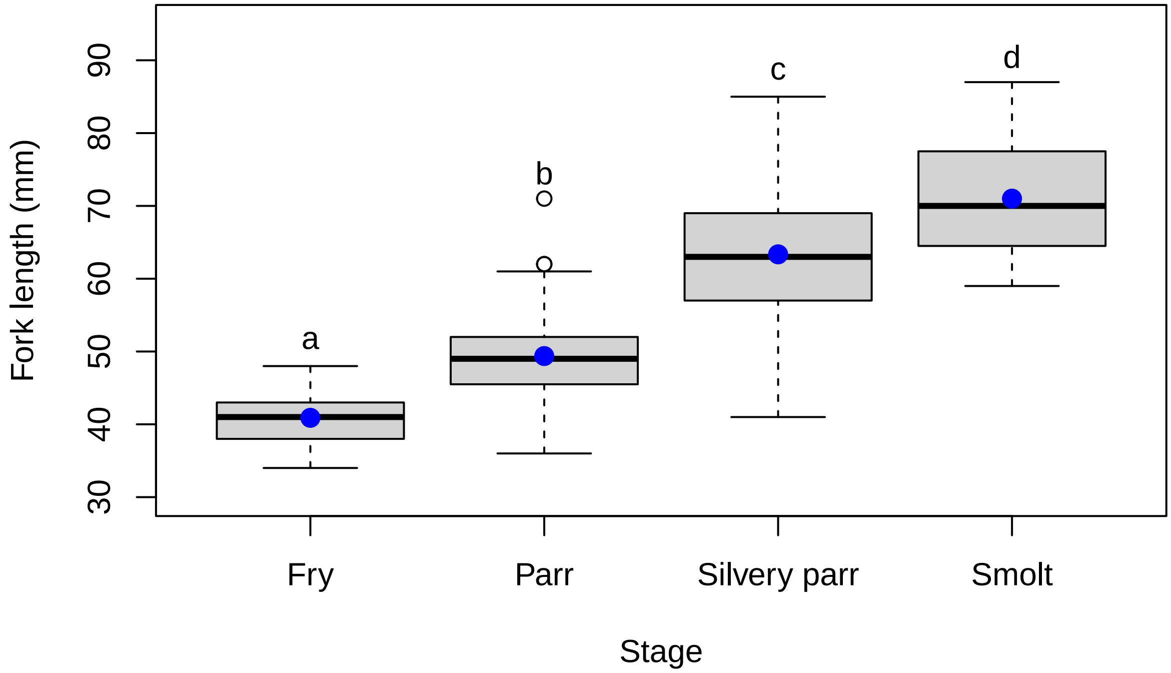

Figure 7: Chinook salmon fork length (mm) by life history stage, with mean fork length for each stage denoted as blue points. Stages not sharing a letter (via Tukey post-hoc) are significantly different.

3.2 Kruskal-Wallis

We can report our Kruskal-Wallis results the same way as our 1-Factor ANOVAs, both textually and visually. However, in addition to the degrees of freedom and p-value, we want to report the H-statistic ("chi-square value") from the Kruskal-Wallis output. Textually, we can say:

"Fork length was significantly different among Chinook salmon life history stages in 2004 (Kruskal-Wallis, H(3)=231.52, p<0.001), and multiple comparisons revealed significant differences among all stages"

Kruskal-Wallis model output can also be represented and graphically as in Fig. 7, adding letters for multiple comparisons.

4 Other Considerations

We covered 1-Factor ANOVAs, Kruskal-Wallis tests, and multiple comparisons. However, in the ANOVA family, there are other extensions such as 2-Factor ANOVAs, repeated measures ANOVA, MANOVA, and ANCOVAs. A 2-Factor ANOVA is just like a 1-Factor ANOVA, but with two factors that can have individual effects and interactive effects on the response when in a fully-crossed design. A repeated measures ANOVA compares means across one or more variables that are based on repeated observations. For example, you could measure the fork lengths of the same individual Chinook salmon over five years and use a repeated measures ANOVA to analyze whether mean fork length changes over time. A MANOVA is a multivariate ANOVA, which can be used when you have multiple, continuous dependent variables (e.g., fork length and weight) and one or more categorical independent variables. Finally, an ANCOVA, or Analysis of Covariance, controls for the effects of other continuous variables that may co-vary with the dependent variable but aren't the main focus of your study (e.g., fork length and temperature, with temperature added as the co-variate).

5 Literature Cited

Bartlett, M.S. 1937. Properties of sufficiency and statistical tests. *Proc Roy Soc Lond A* 160: 268-282.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J of the Roy Stat Soc B* 57: 289-300.

Chen, S., Feng, Z., and Yi, X. 2017. A general introduction to adjustment for multiple comparisons. *J Thorac Dis* 9(6): 1725-1729.

Ghasemi, A., and Zahediasl, S. 2012. Normality tests for Statistical Analysis: A Guide for Non-Statisticians. *Int J Endocrinol Metab* 10(2): 486-89.

Gilbert, M.D., Smith, L., Steinhart, G., and IEP. 2021. Interagency Ecological Program and US Fish and Wildlife Service: Juvenile/Larval Fish and Zooplankton collections at Liberty Island, California 2002-2005 & 2013-2019 ver 1. Environmental Data Initiative.

Hollander, M., and Wolfe, D.A. 1973. Nonparametric Statistical Methods. John Wiley and Sons, New York, NY.

Quinn, G.P., and Keough, M.J. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK. Print.

Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press, Stanford, CA.

Shapiro, S.S., and Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3-4): 591-611.