# FAS6337C - Lab 1

Marcel Gietzmann-Sanders

## Testing for Differences in Population Size Structure

Data for this laboratory are from Largemouth Bass Micropterus salmoides and Florida Bass Micropterus floridanus collected at Lakes Annie, Crooked, and Disston, Florida in 2002. Fish were collected with 20-min electrofishing transects.

```r
setwd("/workspaces/schooling/population_dynamics/lab_1/")
lfdata <- read.table("data/LMB_size_structure.txt", header=T, sep="")
head(lfdata)
```

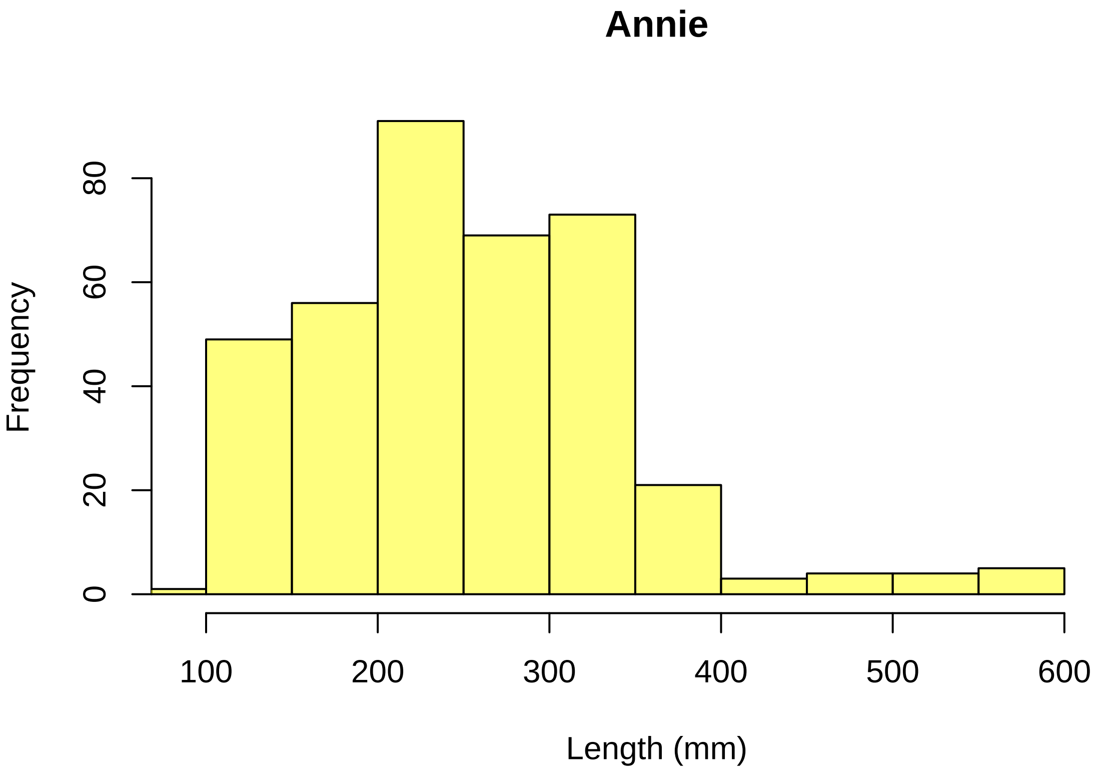
```
##   Lake   Date Rep Effort Species  TL      TLe        Wt        Wte
## 1  ANN 1/4/02   1     20     LMB  90 3.543307  6.069182 0.01338024
## 2  ANN 1/4/02   1     20     LMB 109 4.291339 13.057132 0.02878601
## 3  ANN 1/4/02   1     20     LMB 114 4.488189 21.667118 0.04776776
## 4  ANN 1/4/02   1     20     LMB 142 5.590551 35.648739 0.07859192
## 5  ANN 1/4/02   1     20     LMB 143 5.629921 43.149283 0.09512777
## 6  ANN 1/4/02   1     20     LMB 147 5.787402 32.304074 0.07121821
```

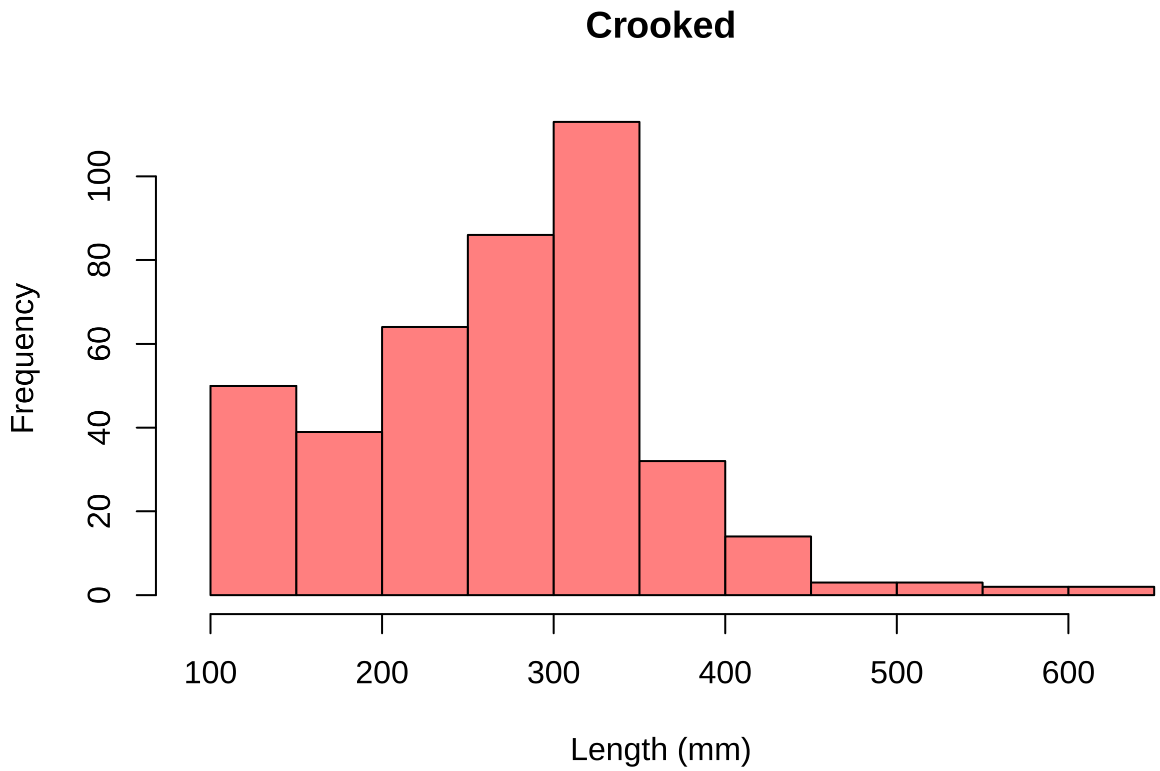
```r
# a nice function from Prof. Siders
col2rgbA<-function(color,transparency)
{
  rgb(t(col2rgb(color))/255,alpha=transparency)
}

# get colors for our lakes (Annie, Crooked, and Disston)
col.lk <- col2rgbA(c('yellow','red','blue'), 0.5)
```

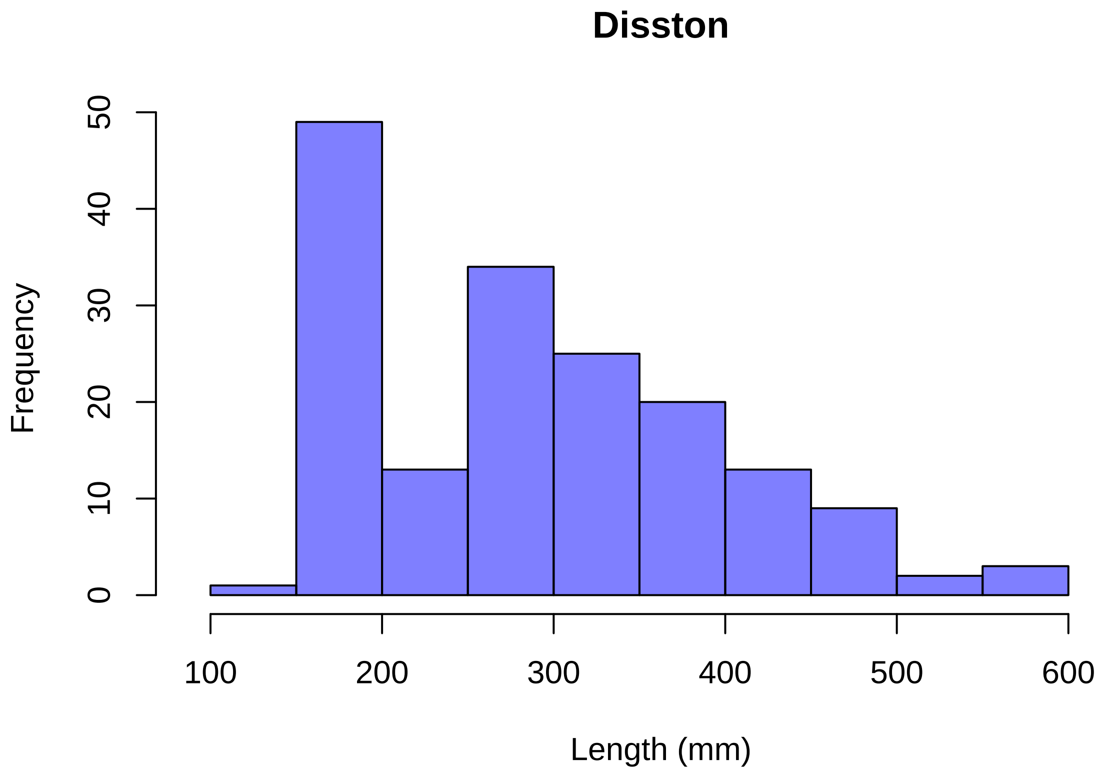
### 1. Using the length frequency for each lake:

#### Make a histogram of the length frequency for each lake

```r
with(
    lfdata, {
        xlim=range(TL)
        hist(
            TL[Lake=="ANN"], main="Annie", col=col.lk[1],
            xlim=xlim, xlab="Length (mm)"
        )
        hist(
            TL[Lake=="CRO"], main="Crooked", col=col.lk[2],
            xlim=xlim, xlab="Length (mm)"
        )
        hist(
            TL[Lake=="DIS"], main="Disston", col=col.lk[3],
            xlim=xlim, xlab="Length (mm)"
        )
    }
)
```

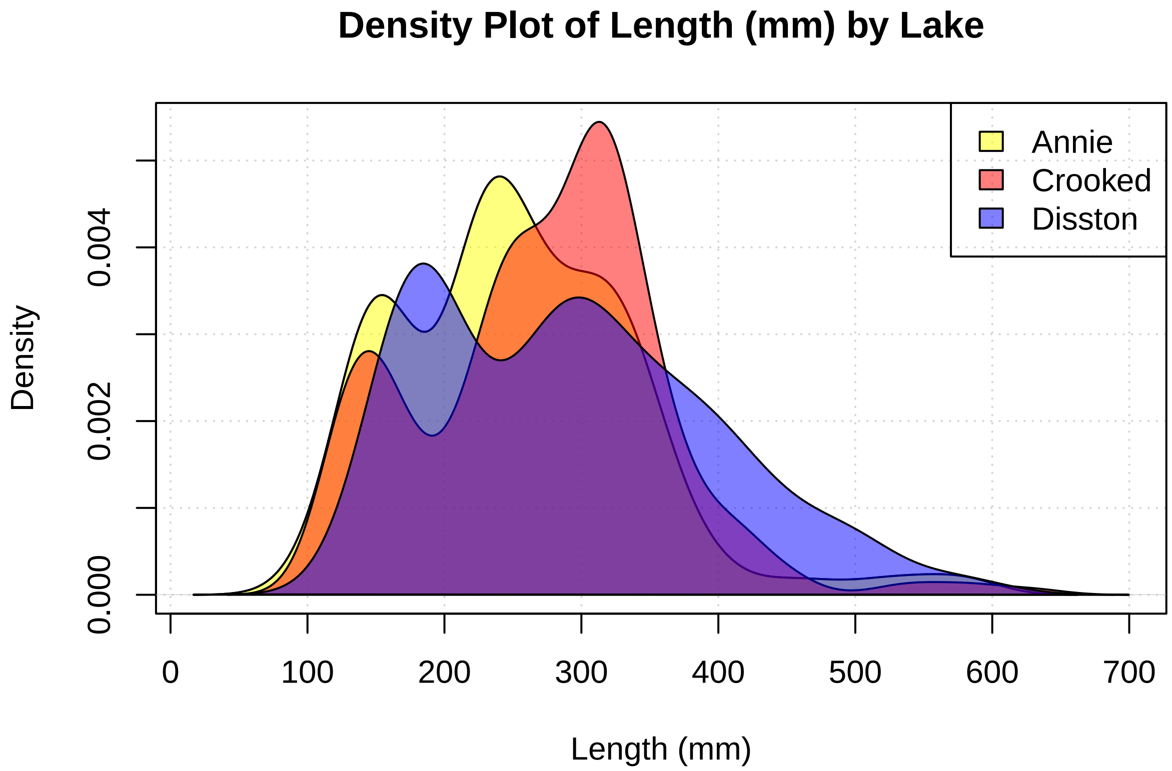

Annie
Frequency
0
20
40
60
80
100
200
300
400
500
600
Length (mm)


Crooked
Frequency
0
20
40
60
80
100
100
200
300
400
500
600
Length (mm)



**Make a density plot of length frequency for each lake**

```r
with(
    lfdata, {
        densAnn <- density(TL[Lake=="ANN"])
        densCro <- density(TL[Lake=="CRO"])
        densDis <- density(TL[Lake=="DIS"])
        xlim <- range(densAnn$x, densCro$x, densDis$x)
        ylim <- range(densAnn$y, densCro$y, densDis$y)
        plot(
            densAnn, xlim=xlim, ylim=ylim, xlab="Length (mm)",
            main="Density Plot of Length (mm) by Lake",
            panel.first = grid()
        )
        polygon(densAnn, density=-1, col=col.lk[1])
        polygon(densCro, density=-1, col=col.lk[2])
        polygon(densDis, density=-1, col=col.lk[3])
    }
)
legend("topright", legend=c("Annie","Crooked","Disston"),
       fill=col.lk)
```



**Do there appear to be differences in size structure?**

Absolutely! The density plot provides the nicest summary of this in my opinion. Couple of interesting observations:

1. While all three lakes seem to have fish getting up to 600-700 mm, Disston takes the long tails from the other distributions to another level.
2. All three lakes are roughly bimodal but the locations of the peaks are different in quite interesting ways. It seems each lake shares a peak with another lake but not both peaks. Very odd.
3. Annie and Crooked seem to have inverted relationship around 200 - 400 mm. They peak on opposite sides of the range and then trough off in opposite directions as well.

## 2. Create a new column of class *factor* that groups TL into the following groupings with the appropriate labels: (nothing needs to be reported for this question).

- Small: [0, 200)
- Quality: [200, 300)
- Preferred: [300, 380)
- Memorable: [380, 450)
- Trophy: [450, )

```
lfdata$sizegrp <- cut(
    lfdata$TL, breaks=c(0, 200, 300, 380, 450, Inf),
    labels=c("S","Q","P","M","T"), right=FALSE
)
head(lfdata)
```

```
##   Lake   Date Rep Effort Species  TL      TLe        Wt        Wte sizegrp
## 1  ANN 1/4/02   1     20     LMB  90 3.543307  6.069182 0.01338024       S
## 2  ANN 1/4/02   1     20     LMB 109 4.291339 13.057132 0.02878601       S
## 3  ANN 1/4/02   1     20     LMB 114 4.488189 21.667118 0.04776776       S
## 4  ANN 1/4/02   1     20     LMB 142 5.590551 35.648739 0.07859192       S
## 5  ANN 1/4/02   1     20     LMB 143 5.629921 43.149283 0.09512777       S
## 6  ANN 1/4/02   1     20     LMB 147 5.787402 32.304074 0.07121821       S
```

## 3. Use a $\chi^2$ test to test if the proportion of fish in each length group varies among the three lakes.

### Report the observed proportion and the expected proportion of fish in each length group by lake.

```
size_groups_by_lake <- xtabs(~sizegrp+Lake, data=lfdata)
size_groups_by_lake
```

```
##        Lake
## sizegrp ANN CRO DIS
##       S 104  89  50
##       Q 160 147  45
##       P  91 141  37
##       M   8  21  23
##       T  13  10  14
```

We find the observed proportion of fish in each length by lake as:

```
prop_by_lake <- prop.table(size_groups_by_lake, margin=2)
prop_by_lake
```

```
##        Lake
## sizegrp        ANN        CRO        DIS
##       S 0.27659574 0.21813725 0.29585799
##       Q 0.42553191 0.36029412 0.26627219
##       P 0.24202128 0.34558824 0.21893491
##       M 0.02127660 0.05147059 0.13609467
##       T 0.03457447 0.02450980 0.08284024
```

The expected proportion is just the total proportion per size group across all three lakes.

```
size_groups <- xtabs(~sizegrp, data=lfdata)
size_groups
```

```
## sizegrp
##   S   Q   P   M   T
## 243 352 269  52  37
```

```
prop_overall <- prop.table(size_groups)
prop_overall
```

```
## sizegrp
##          S          Q          P          M          T
## 0.25498426 0.36935992 0.28226653 0.05456453 0.03882476
```

And we can see the differences between observed and expected with:

```
sweep(prop_by_lake, 1, prop_overall, "-")
```

```
##        Lake
## sizegrp          ANN          CRO          DIS
##       S  0.021611484 -0.036847005  0.040873728
##       Q  0.056171999 -0.009065798 -0.103087727
##       P -0.040245250  0.063321709 -0.063331616
##       M -0.033287937 -0.003093945  0.081530142
##       T -0.004250296 -0.014314960  0.044015473
```

### Report the $\chi^2$ statistic, the degrees of freedom, and the $p$-value

```
chisq.test(lfdata$Lake, lfdata$sizegrp)
```

```
##
##  Pearson's Chi-squared test
##
## data:  lfdata$Lake and lfdata$sizegrp
## X-squared = 61.545, df = 8, p-value = 2.317e-10
```

### Do the size structures differ between all the lakes?

Given the $p$-value is extremely low («0.05) I think it's safe to reject the null hypothesis that these lakes are the same.

### If so, use pairwise $\chi^2$ tests to test if each lake's length frequency differs from one another (three $\chi^2$ tests need to be reported).

```
do_chisq_test <- function(lakes) {
    mask <- lfdata$Lake %in% lakes
    pairwise <- lfdata[mask,]
    chisq.test(
        pairwise$Lake,
        pairwise$sizegrp
    )
}

do_chisq_test(c("ANN", "CRO"))
```

### Annie vs Crooked

```
##
##  Pearson's Chi-squared test
##
## data:  pairwise$Lake and pairwise$sizegrp
## X-squared = 17.434, df = 4, p-value = 0.001591
```

```
do_chisq_test(c("ANN", "DIS"))
```

### Annie vs Disston

```
##
##  Pearson's Chi-squared test
##
## data:  pairwise$Lake and pairwise$sizegrp
## X-squared = 40.785, df = 4, p-value = 2.977e-08
```

```r
do_chisq_test(c("CRO", "DIS"))
```

### Crooked vs Disston

```
##
##  Pearson's Chi-squared test
##
## data:  pairwise$Lake and pairwise$sizegrp
## X-squared = 33.383, df = 4, p-value = 9.973e-07
```

While we would with a $p$ « 0.05 reject all three null hypotheses it is interesting to note just how much Disston is dislike the others.

## 4. Use a Kolmogorov-Smirnov test between pairs of lakes.

```r
do_ks_test <- function(lakes) {
    mask <- lfdata$Lake %in% lakes
    pairwise <- lfdata[mask,]
    ks.test(
        as.integer(factor(pairwise$Lake)),
        as.integer(pairwise$sizegrp)
    )
}

do_ks_test(c("ANN", "CRO"))
```

### Annie vs Crooked

```
## Warning in ks.test(as.integer(factor(pairwise$Lake)),
## as.integer(pairwise$sizegrp)): p-value will be approximate in the presence of
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  as.integer(factor(pairwise$Lake)) and as.integer(pairwise$sizegrp)
## D = 0.36224, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
do_ks_test(c("ANN", "DIS"))
```

### Annie vs Disston

```
## Warning in ks.test(as.integer(factor(pairwise$Lake)),
## as.integer(pairwise$sizegrp)): p-value will be approximate in the presence of
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  as.integer(factor(pairwise$Lake)) and as.integer(pairwise$sizegrp)
## D = 0.40734, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
do_ks_test(c("CRO", "DIS"))
```

**Crooked vs Disston**

```
## Warning in ks.test(as.integer(factor(pairwise$Lake)),
## as.integer(pairwise$sizegrp)): p-value will be approximate in the presence of
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  as.integer(factor(pairwise$Lake)) and as.integer(pairwise$sizegrp)
## D = 0.4662, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

### Does this analysis give the same results as you found in question #3?

This test is really rather sensitive. It considers each test statistic to have a near zero *p*-value. For the comparisons against Disston this follows what we found in #3 but for the Annie vs Crooked comparison this test is clearly far too sensitive (although the ultimate conclusion is the same).

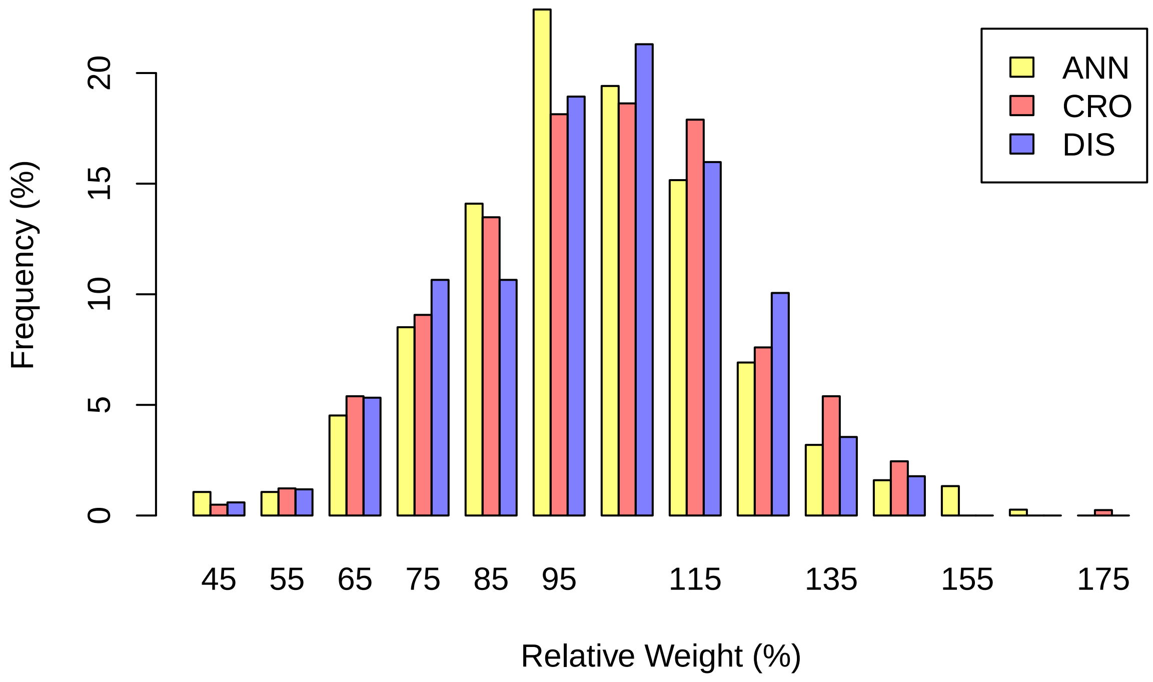
## 5. Calculate the weight standard for each Largemouth Bass using a length-weight relationship from the Structural Indices PDF on Canvas.

### Determine the relative weight using the weight standard and plot a histogram of the relative weight by lake.

```r
a <- 10^-5.316
b <- 3.191
lfdata$WS <- a*lfdata$TL^b
lfdata$WR <- lfdata$Wt / lfdata$WS * 100
head(lfdata)
```

```
##   Lake   Date Rep Effort Species  TL      TLe        Wt        Wte sizegrp
## 1  ANN 1/4/02   1     20     LMB  90 3.543307  6.069182 0.01338024       S
## 2  ANN 1/4/02   1     20     LMB 109 4.291339 13.057132 0.02878601       S
## 3  ANN 1/4/02   1     20     LMB 114 4.488189 21.667118 0.04776776       S
## 4  ANN 1/4/02   1     20     LMB 142 5.590551 35.648739 0.07859192       S
## 5  ANN 1/4/02   1     20     LMB 143 5.629921 43.149283 0.09512777       S
## 6  ANN 1/4/02   1     20     LMB 147 5.787402 32.304074 0.07121821       S
##          WS        WR
## 1  8.317405  72.96966
## 2 15.325969  85.19612
## 3 17.684117 122.52304
## 4 35.641183 100.02120
## 5 36.448302 118.38489
## 6 39.802433  81.16105
```

```r
breaks <- seq(40,180,by=10)
wthA <- hist(
    lfdata$WR[lfdata$Lake=='ANN'],
    breaks=breaks,plot=FALSE
)
wthC <- hist(
    lfdata$WR[lfdata$Lake=='CRO'],
    breaks=breaks,plot=FALSE
)
wthD <- hist(
    lfdata$WR[lfdata$Lake=='DIS'],
    breaks=breaks,plot=FALSE
)
WtRelmat <- matrix(
    c(
        wthA$counts/sum(wthA$counts) * 100,
        wthC$counts/sum(wthC$counts) * 100,
        wthD$counts/sum(wthD$counts) * 100
    ),
    nrow=3, ncol=length(wthA$counts), byrow=TRUE
)
rownames(WtRelmat) <- c('ANN','CRO','DIS')
colnames(WtRelmat) <- paste0("br.", wthA$mids)

barplot(
    WtRelmat, beside=TRUE,
    names.arg=wthA$mids,
    xlab='Relative Weight (%)',
    ylab='Frequency (%)',
    col=col.lk, legend.text=c('ANN','CRO','DIS')
)
```



**Which lake has the heaviest trophy-sized fish on average?**

```
aggregate(WR~Lake, data=lfdata[lfdata$sizegrp == "T",], FUN=mean)
```

```
##   Lake        WR
## 1  ANN  96.61110
## 2  CRO 103.35779
## 3  DIS  98.81896
```

Looks like Crooked has the heavist (by relative weight) fish in the trophy category.

## 6. By adjusting the width of the length bin (shiny app), reassess your comparison of length frequency across the lakes.

### Which length bin do you feel is the most appropriate for comparing length frequencies and why?

I found the 6mm bin to be the most helpful. As you increase the length bin you start to lose some of the finer grained splits in each of the more aggregate modes. For example the lower sizes in Disston actually split into two peaks which becomes practically invisible at larger length bins. However going below 6mm just makes every single bin look like some kind of pattern. I.e. there's no real way to tell the difference between pattern and noise.

### Which of the statistical tests ($\chi^2$ test, Kolmogorov-Smirnov test) are affected by the choice of length bin?

Varying the length bin in the app I don't actually see any differences in either of the tests across any of the comparisons. However given KS is comparing distributions I suspsect that as the bin size becomes larger the test statistic becomes more course and the KS test would become less sensitive. I believe this was also mentioned in lecture - binning is the only way to make KS make sense.

## 7. By showing the density estimate and adjusting the bandwidth, compare the density plots to the histograms.

### Which visualization method do you feel does a better job of showing the differences between the lakes?

While I find the detail in the histograms really useful for understanding the specifics about *each* lake I find the density plot to be better at helping me seeing the difference *between* lakes. The general patterns in modes and skews around those modes is easier to see with the density plot.

## 8. Use an ANOVA to assess differences in the mean length among the three lakes

For the overal comparison:

```
with(
    lfdata, {
        summary(aov(TL~Lake))
    }
)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)    
## Lake          2  187807   93904   11.19 1.57e-05 ***
## Residuals   950 7971492    8391                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Disston v Crooked

```
do_aov <- function(lakes) {
    pairwise <- lfdata[lfdata$Lake %in% lakes,]
    with(
        pairwise, {
            summary(aov(TL~Lake))
        }
    )
}

do_aov(c("CRO", "DIS"))
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)  
## Lake           1   51545   51545   5.915 0.0153 *
## Residuals    575 5010980    8715                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Crooked v Annie

```
do_aov(c("CRO", "ANN"))
```

```
##               Df  Sum Sq Mean Sq F value  Pr(>F)   
## Lake           1   66121   66121    8.39 0.00388 **
## Residuals    782 6162905    7881                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Disston v Annie

```
do_aov(c("DIS", "ANN"))
```

```
##               Df  Sum Sq Mean Sq F value   Pr(>F)    
## Lake           1  178716  178716   20.35 7.92e-06 ***
## Residuals    543 4769099    8783                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### How do these results compare to the $\chi^2$ test?

Overall these results are similar to our $\chi^2$ test. However the likelihoods are quite different. In general the $p$-values are higher with Disston v Crooked really standing out.

### What reservations do you have about the results of this test?

An ANOVA assumes that the statistic being tested is normally distributed. We have some pretty extreme right hand skew here and obviously fish cannot actually be less than 0 (mm) in length. So the assumption that length is normally distributed is definitely violated here.

## 9. Using the lengths and weights of each fish in each lake:

### Fit a length weight relationship for each lake and report the length-weight relationship parameters.

First we're going to move to log space to create a linear relationship - $\log(W) = \log(a) + b * \log(TL)$

```r
lfdata$lWt <- log(lfdata$Wt, 10)
lfdata$lTL <- log(lfdata$TL, 10)
head(lfdata)
```

```
##   Lake   Date Rep Effort Species  TL      TLe        Wt        Wte sizegrp
## 1  ANN 1/4/02   1     20     LMB  90 3.543307  6.069182 0.01338024       S
## 2  ANN 1/4/02   1     20     LMB 109 4.291339 13.057132 0.02878601       S
## 3  ANN 1/4/02   1     20     LMB 114 4.488189 21.667118 0.04776776       S
## 4  ANN 1/4/02   1     20     LMB 142 5.590551 35.648739 0.07859192       S
## 5  ANN 1/4/02   1     20     LMB 143 5.629921 43.149283 0.09512777       S
## 6  ANN 1/4/02   1     20     LMB 147 5.787402 32.304074 0.07121821       S
##          WS        WR       lWt      lTL
## 1  8.317405  72.96966 0.7831302 1.954243
## 2 15.325969  85.19612 1.1158478 2.037426
## 3 17.684117 122.52304 1.3358012 2.056905
## 4 35.641183 100.02120 1.5520442 2.152288
## 5 36.448302 118.38489 1.6349736 2.155336
## 6 39.802433  81.16105 1.5092573 2.167317
```

Next for each lake we can fit our linear model. The intercept and *lTL* coefficient correspond to our intercept and slope as in the resource from Canvas.

Annie:

```r
df <- lfdata[lfdata$Lake=="ANN",]
lm(df$lWt~df$lTL)
```

```
##
## Call:
## lm(formula = df$lWt ~ df$lTL)
##
## Coefficients:
## (Intercept)       df$lTL
##      -5.392        3.218
```

Crooked:

```r
df <- lfdata[lfdata$Lake=="CRO",]
lm(df$lWt~df$lTL)
```

```
##
## Call:
## lm(formula = df$lWt ~ df$lTL)
##
## Coefficients:
## (Intercept)       df$lTL
##      -5.215        3.147
```

Disston:

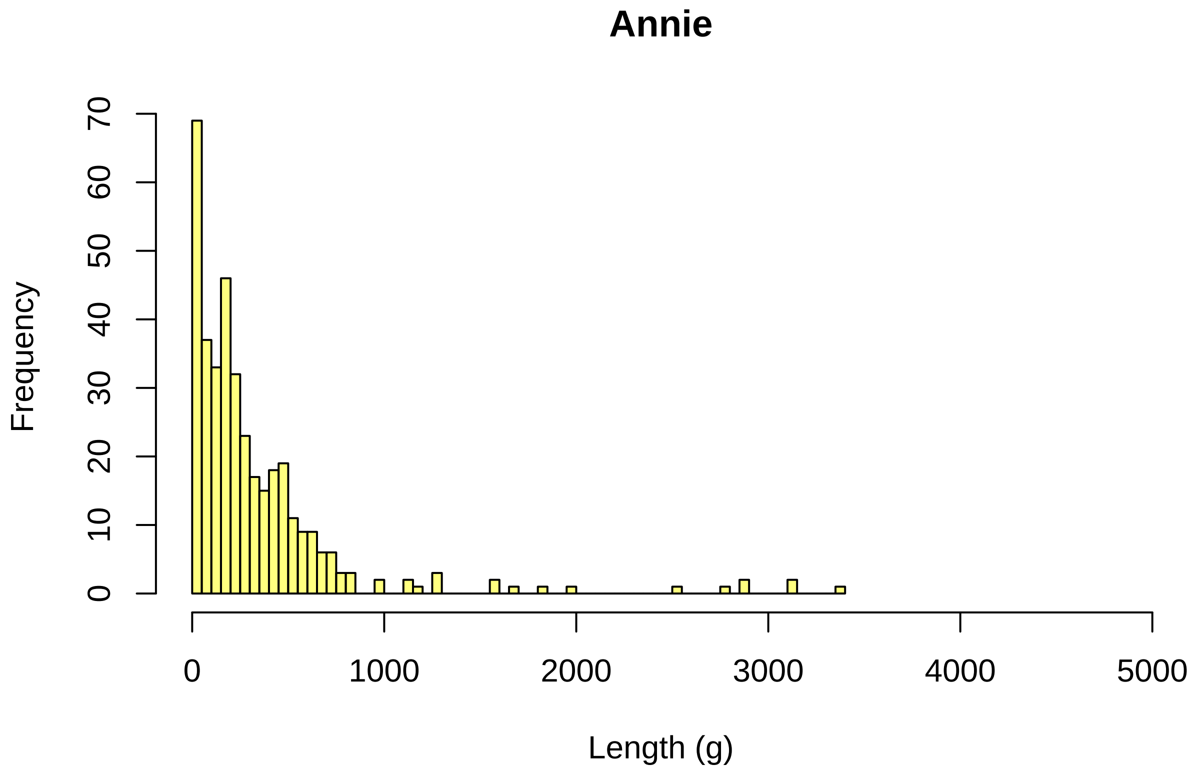
```r
df <- lfdata[lfdata$Lake=="DIS",]
lm(df$lWt~df$lTL)
```

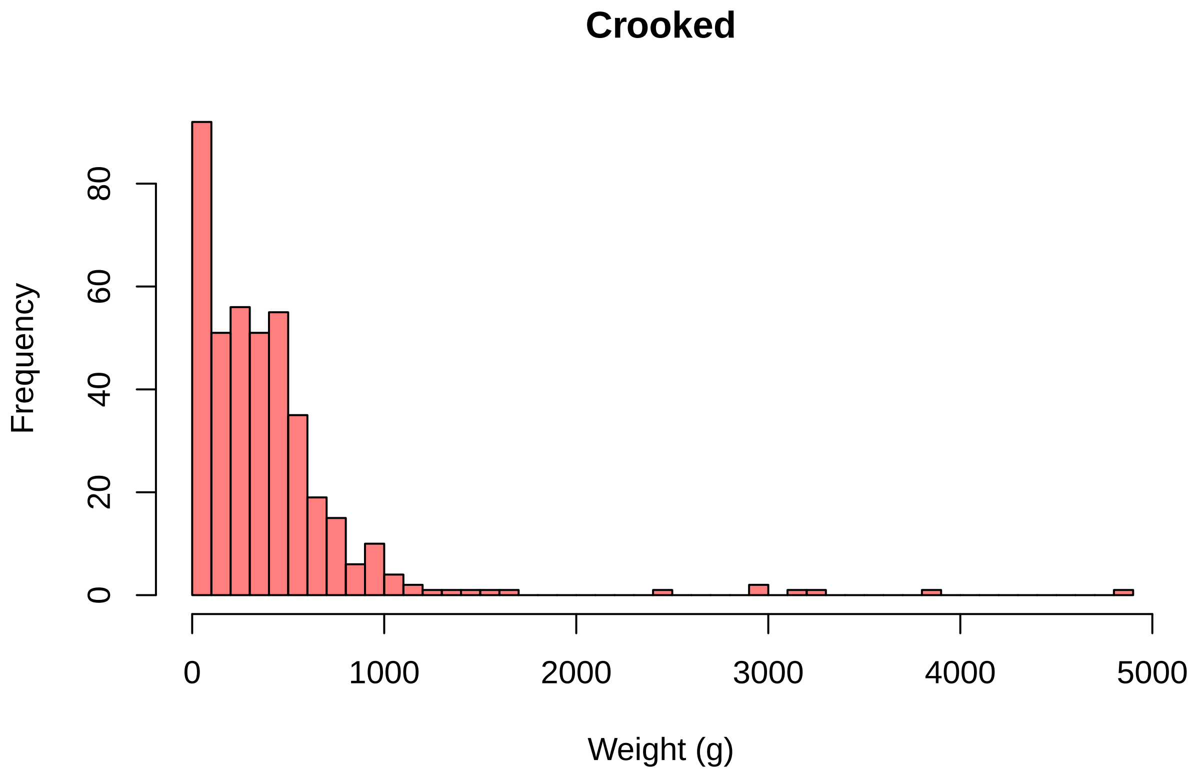
```
##
## Call:
## lm(formula = df$lWt ~ df$lTL)
##
## Coefficients:
```

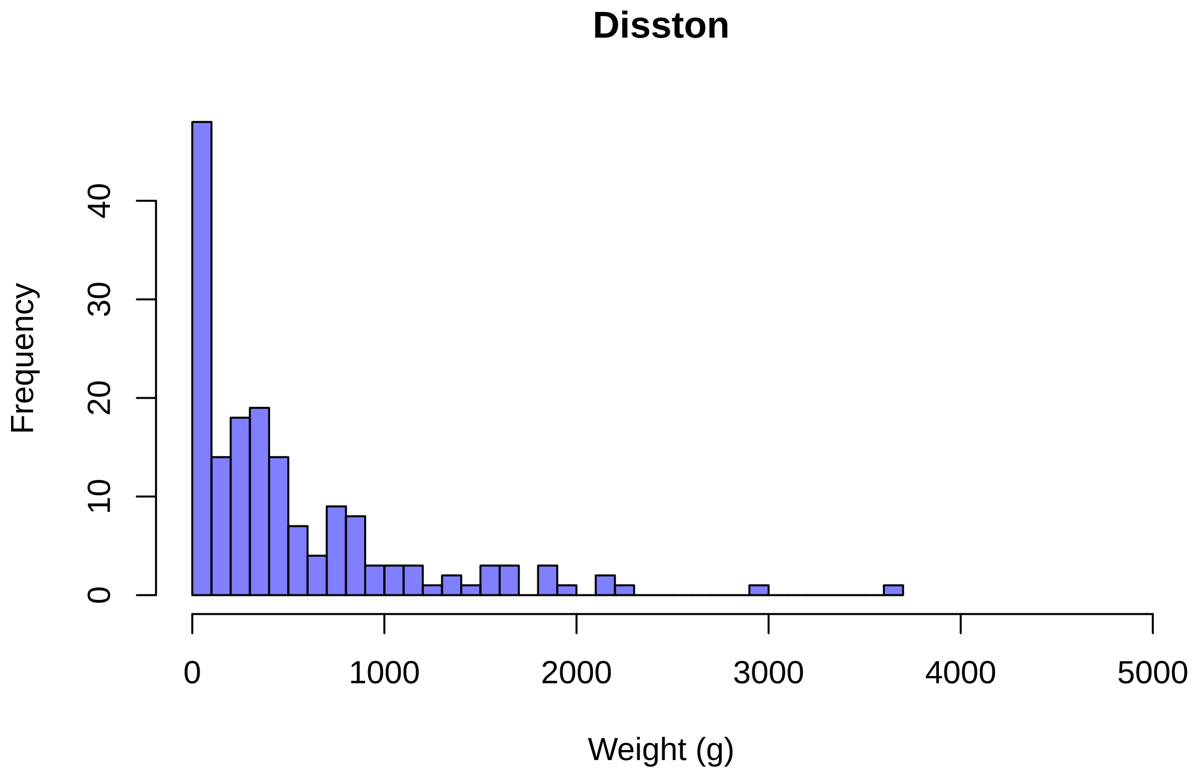
```
## (Intercept)       df$lTL
##      -5.144        3.116
```

**Plot the observed weight on a histogram for each lake.**

```
with(
    lfdata, {
        breaks <- 50
        xlim=range(Wt)
        hist(
            Wt[Lake=="ANN"], main="Annie", col=col.lk[1],
            xlim=xlim, xlab="Length (g)", breaks=breaks
        )
        hist(
            Wt[Lake=="CRO"], main="Crooked", col=col.lk[2],
            xlim=xlim, xlab="Weight (g)", breaks=breaks
        )
        hist(
            Wt[Lake=="DIS"], main="Disston", col=col.lk[3],
            xlim=xlim, xlab="Weight (g)", breaks=breaks
        )
    }
)
```




Crooked
Frequency
0
20
40
60
80
0
1000
2000
3000
4000
5000
Weight (g)



**What might the dispersion of the histogram tell us about the population dynamics**

- Disston seems to have a higher relative frequency of heavier fish. This seems to imply that they aren't getting crowded out by loads of smaller fish or that they are in fact predating on the smaller population (these are opportunistic predators after all).
- In constrast both Annie and Crooked largely top out at 1000 (g) which would seem to imply that due to competitive pressure amongst the younger age group none are really able to get the nutrition to get to the larger size classes?
- It's also possible that the difference here is in some kind of predation selectivity difference between the lakes?

## 10. Using the R-Shiny application, interpret the GLM results.

### What are the lengths that the GLM predicted CDF does a good job of predicting?

The GLM does a pretty good job of predicting lengths above 300mm. For Annie and Disston it also does relatively well for 250-300mm.

### What are the lengths that the GLM predicted CDF does a bad job of predicting?

The GLM does quite poorly on everything between 250mm. This makes sense as there's a lot of steep slope to plateu patterns in the lower lengths (quick accumulation of length in small fish). In addition none of the lengths really get below 100 which the GLM has a hard time predicting.

**Does your answer change as a function of the chosen width of the length bin? If so, how?**

In the extreme where we set the length bin as high as possible (40mm) the points get smoothed out to such an extent that that steep slope to plateu pattern dissapears allowing the GLM to predict well below 300mm. However the GLM still cannot get down to 0% at ~100mm and so still fails to predict below 150mm or so. However at this level of binning, all the interesting relationships are gone!

**Using the GLM coefficient table, which parameters differ between lakes?**

Between Annie and Crooked the main difference seems to be in the Intercept (-4.9 vs -5.3) rather than the slope (0.02 v 0.019). For Disston both the slope and intercept are quite different from the other lakes.

## 11. Using the R-Shiny application or your previous analysis, which statistical test would you choose to assess differences in length frequencies between the lakes and why?

From all we've done here it's pretty clear to me that the only test with assumptions that match up to our data is the $\chi^2$ test.

- The ANOVA assumes normality that clearly isn't there.
- The GLM is too simple to capture the interesting relationships in length frequency (which is what distinguishes one lake from the next).
- The KS is just far too sensitive to tell us anything meaningful.

The $\chi^2$ not only has good assumptions built in but also passes the sniff test after all of this plotting.