R for Data Analysis

Marcus W. Beck   Sergey Berg

Department of Fisheries, Wildlife, and Conservation Biology
University of Minnesota, Twin Cities

May 21, 2013

What you'll learn about

- Data organization
- Data exploration and visualization
  - Common functions
  - Graphing tools
- Data analysis and hypothesis testing
  - Common functions
  - Evaluation of output
  - Graphing tools

Interactive! Interrupt me!

Data organization

We'll use the same dataset we used in Excel, replicating the analyses

First we have to import the data into our R 'workspace'

The workspace is a group of R objects that are loaded for our current session

Data are loaded into the workspace by importing (or making within R) and assigning them to a variable (object) with a name of our choosing

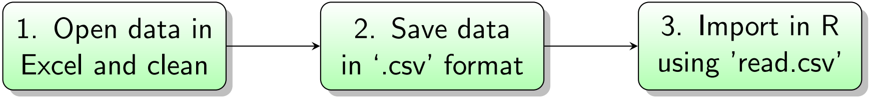
We can see what's loaded in our workspace:

```
> a<-c(1,2)
> ls()
```

```
[1] "a"
```

Data organization

Import the data following this workflow:



1. Open data in Excel and clean

2. Save data in ‘.csv’ format

3. Import in R using ’read.csv’

- Column names should be simple
- Ensure all data will be easy to read

- File, Save as .csv
- Creates a comma separated file that looks like a spreadsheet
- One spreadsheet at a time

- header = T
- See ?read.csv for list of function options
- Remember to assign a name

Data organization

If the data are a text file... open the text file, how are the columns separated?

- comma
- tabs
- space
- arbitrary character

Use the read.table function and identify the column delimiter:

```
> setwd('C:/Documents/monitoring_workshop')
> ls()
```

```
[1] "a"
```

```
> dat<-read.table('RWBB Survey.txt',sep='\t',header=T)
> ls()
```

```
[1] "a"   "dat"
```

Data organization

Now that the data are in our workspace, let's explore!

Did the data import correctly (rarely a problem)?

```
> head(dat) #or tail(dat)
```

```
  SiteName Year Restoration Reference   ObserverNames Precipitation Temperature
1      IGH 2005           3         3    Tyler_Amanda             0          48
2    Kelly 2005           4         2 Patrick_Chelsea             0          48
3  Carlton 2005           2         3     David_Megan             0          48
4      IGH 2006           9         6    Tyler_Amanda             0          52
5    Kelly 2006           9         1     David_Megan             0          52
6  Carlton 2006           7         3 Patrick_Chelsea             0          52
```

Data exploration

What object class is the data?

```
> class(dat)
```

```
[1] "data.frame"
```

What are the dimensions of the data frame?

```
> dim(dat)
```

```
[1] 18  7
```

```
> nrow(dat)
```

```
[1] 18
```

```
> ncol(dat)
```

```
[1] 7
```

The data contain 18 rows and 7 columns, is this correct?

Data exploration

Can we get a summary of the data frame?

```
> summary(dat)
```

```
    SiteName      Year       Restoration      Reference
 Carlton:6   Min.   :2005   Min.   : 2.00   Min.   : 1.000
 IGH    :6   1st Qu.:2006   1st Qu.: 7.50   1st Qu.: 2.250
 Kelly  :6   Median :2008   Median :11.50   Median : 3.000
             Mean   :2008   Mean   :11.11   Mean   : 4.389
             3rd Qu.:2009   3rd Qu.:14.75   3rd Qu.: 5.000
             Max.   :2010   Max.   :24.00   Max.   :18.000
         ObserverNames Precipitation  Temperature
 David_Megan    :6     Min.   : 0    Min.   :41.00
 Jeremy_Lucy    :1     1st Qu.: 0    1st Qu.:48.00
 Patrick_Chelsea:6     Median : 0    Median :53.00
 Tyler_Amanda   :5     Mean   : 2    Mean   :51.83
                       3rd Qu.: 0    3rd Qu.:55.00
                       Max.   :12    Max.   :61.00
```

Data exploration

Individual summmaries of variables are also possible

How do we obtain variables of interest?

```
> names(dat)
```

```
[1] "SiteName"      "Year"          "Restoration"   "Reference"
[5] "ObserverNames" "Precipitation" "Temperature"
```

We can get a variable directly using $ or via indexing with [,]

```
> dat$Temperature
```

```
 [1] 48 48 48 52 52 52 41 41 41 54 54 54 55 55 55 61 61 61
```

```
> dat[,'Temperature'] #same as dat[,7]
```

```
 [1] 48 48 48 52 52 52 41 41 41 54 54 54 55 55 55 61 61 61
```

Data exploration

Just as we had summaries of the data frame, we can get summaries of individual variables

```
> summary(dat$Temperature)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.00   48.00   53.00   51.83   55.00   61.00
```

Or more simplistically...

```
> mean(dat$Temperature)
```

```
[1] 51.83333
```

```
> range(dat$Temperature)
```

```
[1] 41 61
```

```
> unique(dat$Temperature)
```

```
[1] 48 52 41 54 55 61
```

Data exploration

Note that the classes of our variables affect how R functions interpet them

For example, the summary function returns different information...

```
> class(dat$Temperature)

[1] "integer"

> summary(dat$Temperature)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.00   48.00   53.00   51.83   55.00   61.00

> class(dat$SiteName)

[1] "factor"

> summary(dat$SiteName)

Carlton     IGH   Kelly
      6       6       6
```

Data exploration

What about site-specific evaluations? What if we want to look at the temperature only at Kelly?

```
> Kelly<-subset(dat, dat$SiteName=='Kelly')
```

We've created a new object in our workspace that is our original data frame with sites only from Kelly

```
> dim(Kelly)

[1] 6 7

> Kelly$SiteName

[1] Kelly Kelly Kelly Kelly Kelly Kelly
Levels: Carlton IGH Kelly
```

Data exploration

What about site-specific evaluations? What if we want to look at the temperature only at Kelly?

```
> Kelly<-subset(dat, dat$SiteName=='Kelly')
```

Now we can evaluate the temperature, for example, only at Kelly

```
> mean(Kelly$Temperature) #this is the same as all sites
[1] 51.83333
```

Data exploration

What abour our restoration project? Aren't we comparing the abundances of breeding birds between restored and reference sites?

Let's start with our reference sites...

```
> ref<-dat$Reference
> summary(ref) #or summary(dat$Reference)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   2.250   3.000   4.389   5.000  18.000 
```

Now the restored sites...

```
> rest<-dat$Restoration
> summary(rest)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   2.00    7.50   11.50   11.11   14.75   24.00 
```

Data visualization

Textual summaries of our data are nice, but we should also visualize:

- How are our data distributed?
- Are there any outliers or extreme observations?
- How do our variables compare (to a reference, to one another, over time, etc.)?

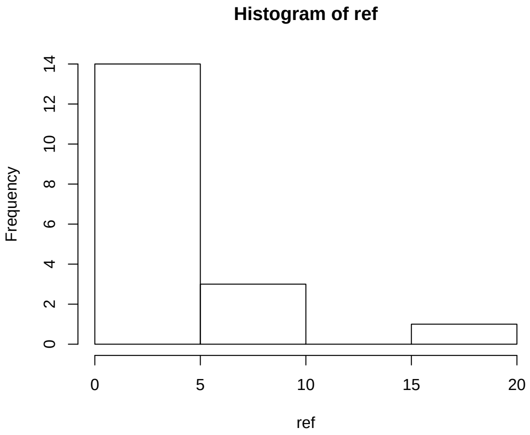
R has many built in functions for data exploration and plotting

- hist - plots a histogram (binned densities of continuous values)
- qqplot - comparison of a variable to a normal distribution
- barplot - for bar plots...
- plot - bivariate comparison of two variables
- Much, much more...

Data visualization

Let's examine the distribution of abundances for the breeding birds at our reference site

```
> hist(ref) #or hist(dat$Reference)
```

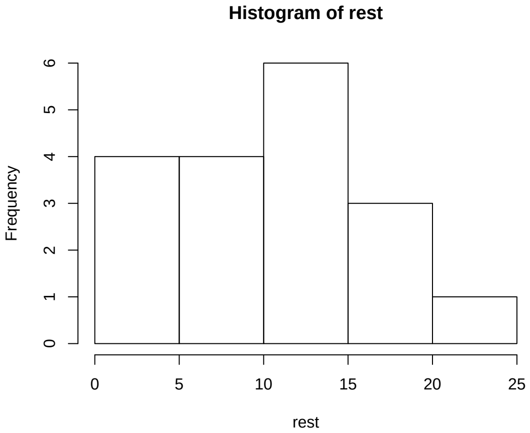


14 of our reference sites have abundances between 0–5 breeding birds

Data visualization

How does it compare to our restoration site?

```
> hist(rest) #or hist(dat$Restoration)
```

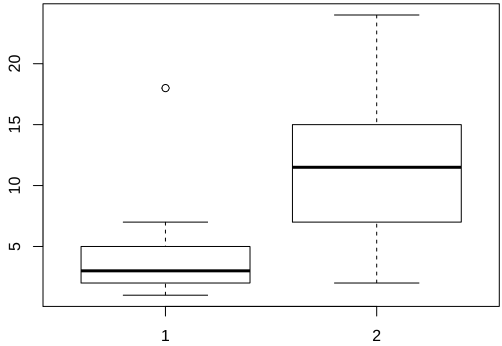


Six of our reference sites have abundances between 10–15 breeding birds

Data visualization

Now that we've seen the distribution, how can we compare directly?

```
> boxplot(ref,rest)
```

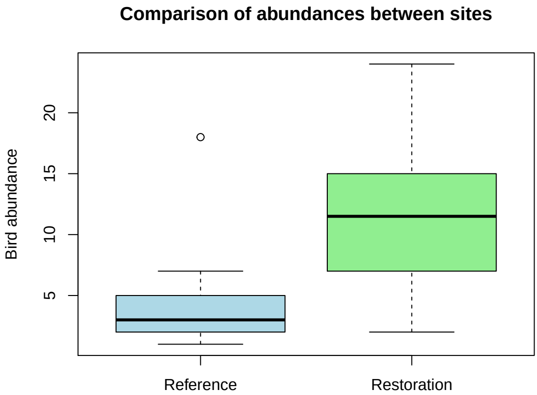


Let's make it look better...

Data visualization

Now that we've seen the distribution, how can we compare directly?

```
> boxplot(ref,rest,names=c('Reference','Restoration'),
+  ylab='Bird abundance',col=c('lightblue','lightgreen'),
+  main='Comparison of abundances between sites')
```



Dark line is median, box is 25^{th} to 75^{th} quartile (or IQR), whiskers are 1.5 × IQR

Beyond can be considered outliers...

Data visualization

What's going on with the outlier at our reference site? How can we identify it?

We can use the boxplot function for the dirty work...

```
> myplot<-boxplot(ref,rest)
> myplot$out
[1] 18
```

This gives us the actual value, now we need to find it in our data frame

```
> outlier<-myplot$out
> out.row<-which(ref==outlier)
> out.row #this is the row number
[1] 8
```

Data visualization

```
> dat[out.row,] #same as dat[8,]
```

```
  SiteName Year Restoration Reference ObserverNames Precipitation Temperature
8    Kelly 2007           2        18   Jeremy_Lucy            12          41
```

Now we know that our outlier was from Kelly in 2007...

What's odd about this record?

Let's look at our records from Kelly...

Data visualization

```
> Kelly
   SiteName Year Restoration Reference   ObserverNames Precipitation
2     Kelly 2005           4         2 Patrick_Chelsea             0
5     Kelly 2006           9         1     David_Megan             0
8     Kelly 2007           2        18     Jeremy_Lucy            12
11    Kelly 2008          14         5     David_Megan             0
14    Kelly 2009          16         5     David_Megan             0
17    Kelly 2010          15         3 Patrick_Chelsea             0
   Temperature
2           48
5           52
8           41
11          54
14          55
17          61
```

2007 was cold and rainy, could that have been the reason?

Let's look at 2007 for all sites...

Data visualization

```
> subset(dat,dat$Year=='2007')
```

```
  SiteName Year Restoration Reference   ObserverNames Precipitation Temperature
7      IGH 2007          12         7     David_Megan            12          41
8    Kelly 2007           2        18     Jeremy_Lucy            12          41
9  Carlton 2007          11         2 Patrick_Chelsea            12          41
```

IGH and Carlton don't have high abundances at their reference sites during 2007 even though the weather was the same

What else could have caused this outlier?

```
> summary(dat$ObserverNames)
```

```
    David_Megan     Jeremy_Lucy Patrick_Chelsea    Tyler_Amanda
              6               1               6               5
```

Data visualization

```
> summary(dat$ObserverNames)
```

```
    David_Megan      Jeremy_Lucy Patrick_Chelsea     Tyler_Amanda 
              6                1               6                5 
```

This is probably Jeremy and/or Lucy's fault, most likely switched the restoration and reference records

What to change?

```
> dat[out.row,'Restoration']<-18
> dat[out.row,'Reference']<-2
```

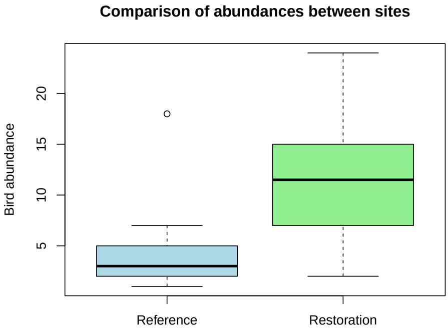
Or...

```
> dat<-dat[-out.row,] #do this one
```

Or... fire Jeremy and Lucy.

Data analysis and hypothesis testing

Now we need to evaluate the statistical certainty of our data, i.e., are our results due to random chance and how can we quantify this?



We want to determine if the abundance of birds or variation among sites is actual or random

What is an appropriate hypothesis?

Data analysis and hypothesis testing

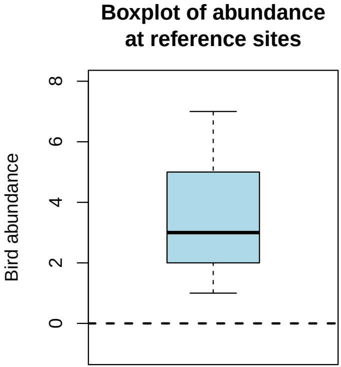
What is an appropriate hypothesis? Let's start simple...

Null hypothesis

The mean abundance of breeding birds at our reference site is zero.

Alternative hypothesis

The mean abundance of breeding birds at our reference site is not zero.



Data analysis and hypothesis testing

The t.test function lets us test this hypothesis, very simple...

```
> t.test(dat$Reference)

One Sample t-test

data:  dat$Reference
t = 7.8998, df = 16, p-value = 6.528e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.625334 4.551137
sample estimates:
mean of x
 3.588235
```

What does this mean? What are default arguments?

Data analysis and hypothesis testing

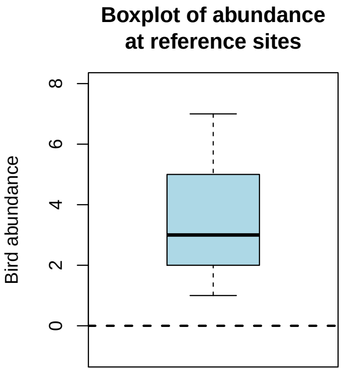
Perhaps a one-tailed alternative hypothesis is better, we have prior assumptions about the data...

Null hypothesis

The mean abundance of breeding birds at our reference site is zero.

Alternative hypothesis

The mean abundance of breeding birds at our reference site is greater than zero.



Data analysis and hypothesis testing

Slight modification of alternative argument for one-tailed test, default is two-tailed

```
> t.test(dat$Reference, alternative='greater')

One Sample t-test

data:  dat$Reference
t = 7.8998, df = 16, p-value = 3.264e-07
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 2.795222      Inf
sample estimates:
mean of x
 3.588235
```

What does this mean?

Data analysis and hypothesis testing

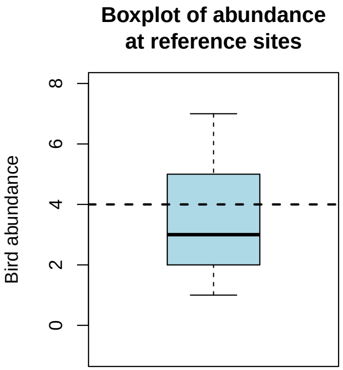
Let's explore more flexibility of the t.test function by changing our basis of comparison for the alternative hypothesis

Null hypothesis

The mean abundance of breeding birds at our reference site is four.

Alternative hypothesis

The mean abundance of breeding birds at our reference site is greater than four.



Data analysis and hypothesis testing

Test a different alternative hypothesis by changing the mu argument

```
> t.test(dat$Reference, mu=4, alternative='greater')

One Sample t-test

data:  dat$Reference
t = -0.9065, df = 16, p-value = 0.8109
alternative hypothesis: true mean is greater than 4
95 percent confidence interval:
 2.795222      Inf
sample estimates:
mean of x
 3.588235
```

What does this mean?

Data analysis and hypothesis testing

Now the real question, let's compare our sites to one another...

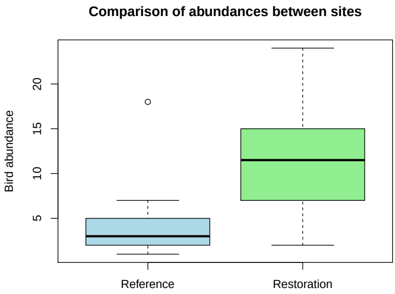
What are our hypotheses?

Null hypothesis

Differences in the mean abundance between restoration and reference sites is zero.

Alternative hypothesis

Differences in the mean abundance between restoration and reference sites will be greater than zero.



Data analysis and hypothesis testing

Use the t.test function again as a two-sample test, order matters as do arguments

```
> t.test(dat$Restoration,dat$Reference,
+  alternative='greater',var.equal=T)

Two Sample t-test

data:  dat$Restoration and dat$Reference
t = 5.3121, df = 32, p-value = 4.006e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.489093      Inf
sample estimates:
mean of x mean of y
11.647059  3.588235
```

What does this mean?

Data analysis and hypothesis testing

Order of arguments matters...

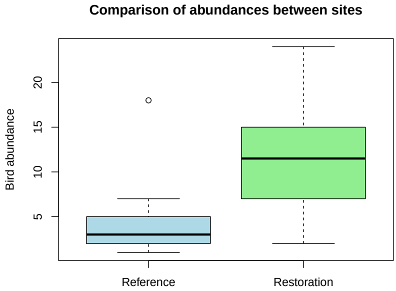
```
> t.test(dat$Reference,dat$Restoration,
+  alternative='greater',var.equal=T)
```

```
Two Sample t-test

data:  dat$Reference and dat$Restoration
t = -5.3121, df = 32, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -10.62855       Inf
sample estimates:
mean of x mean of y
 3.588235 11.647059
```

What does this mean? What happens if we change the alternative argument?

Data analysis and hypothesis testing

Our results suggest that the abundance of breeding birds at the restoration site is significantly greater than at the reference site



Our p-value is 4.006e-06, what does this mean?

There is a 0.0004006% chance that our results were observed due to randomness (within the constraints of our test).

Data analysis and hypothesis testing

Other common tests:

- χ^2 test of independence - chisq.test
- analysis of variance - anova or aov
- correlations - cor.test or cor
- regression - lm or glm
- Much, much more....

Data analysis and hypothesis testing

One last example... we've used common tests to compare our data to a standard or reference (e.g., mean is zero, differences in means is greater than zero)

What about a more interesting analysis, such as comparison of data over time or relationships between variables?

We'll close by illustrating use of linear regression with our data

This is an evaluation of the mean response of a variable conditional on another, i.e., a predictor

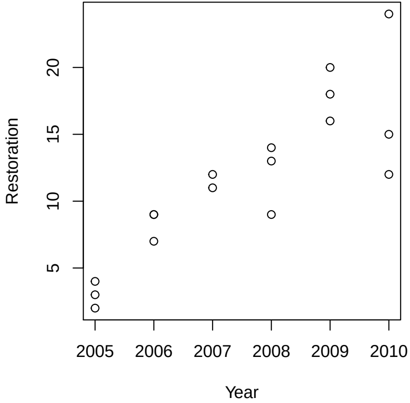
Data analysis and hypothesis testing

Perhaps we expect the abundance of breeding birds to increase at our restoration site over time, let's plot it:

```
> plot(Restoration~Year,data=dat)
```

The first argument is entered as a 'formula' specifying the variables

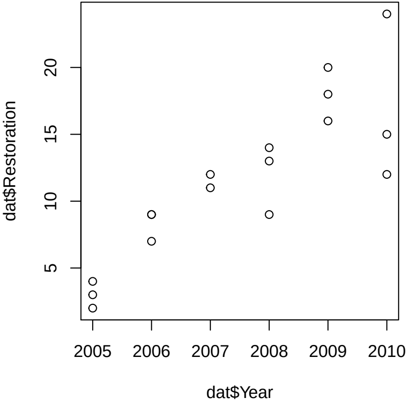
The data argument specifies location of the variables in the workspace



Data analysis and hypothesis testing

We can also call the variables directly in the plot function, x variable first, y second:

```
> plot(dat$Year,dat$Restoration)
```



Note the change of the x and y labels, we can modify these using the xlab and ylab arguments in the plot function

Notice the clear trend...

Data analysis and hypothesis testing

How do we quantify this trend across time? Use the lm function for regression...

```
> lm(Restoration~Year,data=dat)

Call:
lm(formula = Restoration ~ Year, data = dat)

Coefficients:
(Intercept)         Year
  -5721.568        2.856
```

The abundance increases, on average, by 2.856 birds per year.

Data analysis and hypothesis testing

We can get more information using the summary command

```
> mod<-lm(Restoration~Year,data=dat)
> summary(mod)
```

```
Call:
lm(formula = Restoration ~ Year, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7027 -1.4234  0.1532  1.7207  5.2973 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5721.5676   860.1976  -6.651 7.72e-06 ***
Year            2.8559     0.4285   6.665 7.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.097 on 15 degrees of freedom
Multiple R-squared:  0.7476, Adjusted R-squared:  0.7307 
F-statistic: 44.42 on 1 and 15 DF,  p-value: 7.544e-06
```

Data analysis and hypothesis testing

What does this mean?

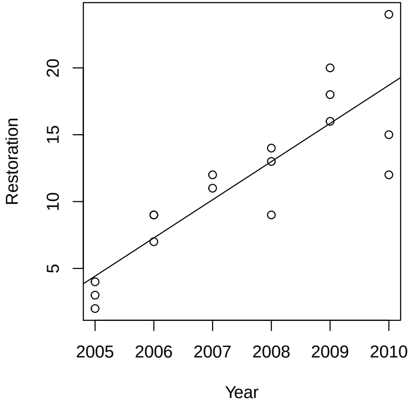
```
Call:
lm(formula = Restoration ~ Year, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7027 -1.4234  0.1532  1.7207  5.2973

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5721.5676   860.1976  -6.651 7.72e-06 ***
Year            2.8559     0.4285   6.665 7.54e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.097 on 15 degrees of freedom
Multiple R-squared:  0.7476, Adjusted R-squared:  0.7307
F-statistic: 44.42 on 1 and 15 DF,  p-value: 7.544e-06
```

Data analysis and hypothesis testing

How do we plot the model?

```
> plot(Restoration~Year, data=dat)
> abline(reg=mod)
```

We tell the abline function to plot our model, named ‘mod’



Data analysis and hypothesis testing

Can we use our model for prediction?

What are the predicted data for our observation years?

```
> predict(mod)
        1         2         3         4         5         6         7         9
 4.423423  4.423423  4.423423  7.279279  7.279279  7.279279 10.135135 10.135135
       10        11        12        13        14        15        16        17
12.990991 12.990991 12.990991 15.846847 15.846847 15.846847 18.702703 18.702703
       18
18.702703
```

What about other years not in our dataset?

Data analysis and hypothesis testing

Can we use our model for prediction?

What about predicted abundance for 2011?

```
> predict(mod,newdata=data.frame(Year=2011))
       1
21.55856
```

We can expect, on average, 21.56 birds at our restoration sites in 2011 (within the constraints of our model)

Conclusion

What we've learned:

- Data organization - read.csv, read.table
- Data exploration - head, dim, nrow, ncol, summary, [,], $, names, subset, mean, range, unique
- Data visualization - hist, boxplot, plot, abline
- Data analysis and hypothesis testing - t.test, lm, predict

Questions?