Intro to R: Week 8

Topics Covered: Random sampling, generating synthetic datasets, t-tests, ANOVA, linear models

```
setwd("~/Desktop/IntroR/Week 8/")

library("ggplot2")
library("dplyr")
library("lubridate")
library("reshape2")
```

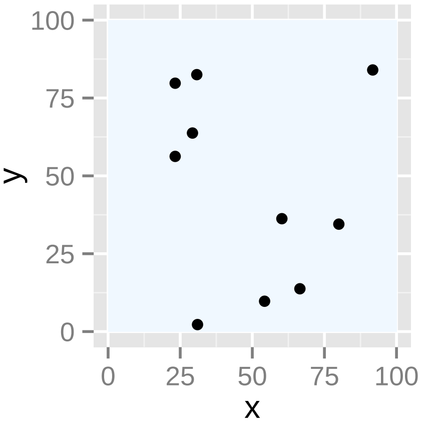
Task 1: Designing Random Surveys

Step 1.1 Imagine that you are running an experiment in an aquarium and you want to randomly assign individual corals to grid positions within the aquarium. Write an algorithm to assign 20 individuals to grid positions A1 - D5.

```
samples <- paste("N", 1:20, sep="")

grid <- paste(rep(LETTERS[1:4], each=5), 1:5, sep="")

positions <- data.frame(grid, "sample"=sample(samples))
```

Step 1.2 Imagine that you have a study area bounded by the coordinates 0,0, 0,100, 100,100, 100,0 and you want to randomly choose 10 locations to sample within that area. Find the coordinates of those random samples and plot them using ggplot.

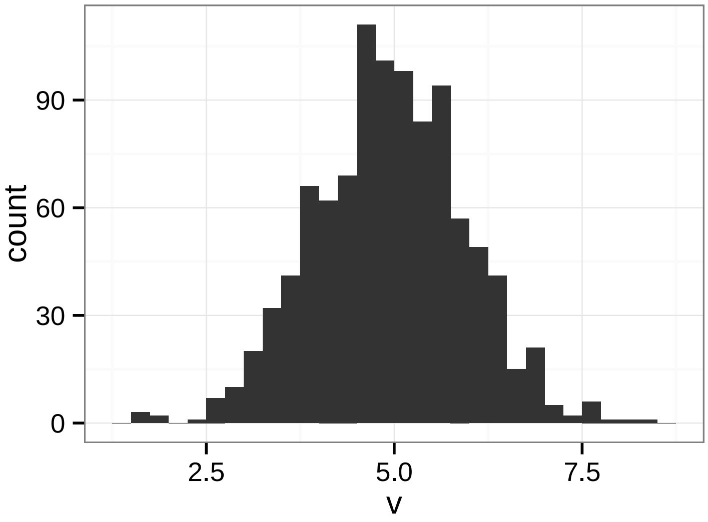
```
area <- data.frame("x"=c(0,0,100,100), "y"=c(0,100,100,0))

x.coords <- seq(0, 100, by=0.25)
y.coords <- seq(0, 100, by=0.25)

samples <- data.frame("x"=sample(x.coords, 10, replace=TRUE),
                      "y"=sample(y.coords, 10, replace=TRUE))

# This is probably fine, but has a very small chance of returning the same
# coordinates more than once.

all.coords <- expand.grid(x.coords, y.coords) # all possible combinations

samples <- all.coords[sample(1:nrow(all.coords), 10, replace=FALSE), ]

ggplot()+
  geom_polygon(data=area, aes(x, y), fill="aliceblue")+
  geom_point(data=samples, aes(Var1, Var2))
```



Task 2: Generating a Synthetic Dataset

Step 2.1 Create a vector of 1000 observations with a mean of 5 and a standard deviation of 1. Plot the observations as a histogram. Is the mean of these observations significantly different from 4.95?

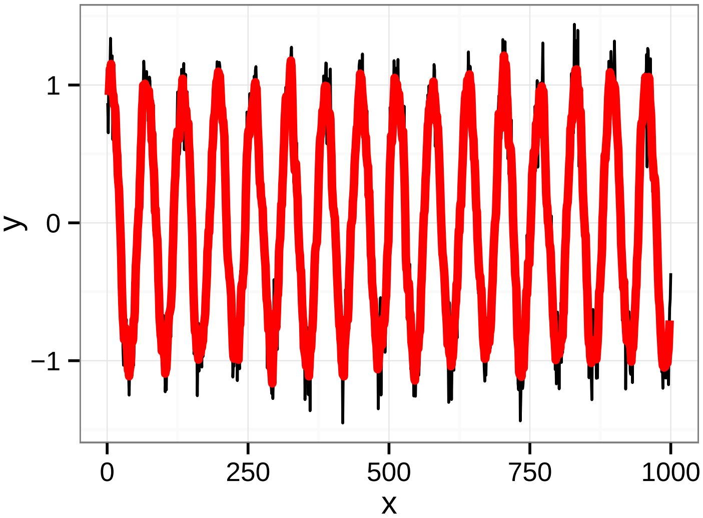
```r
set.seed(2015)

v <- rnorm(1000, mean=5, sd=1)

v.df <- data.frame(v)

ggplot(data=v.df, aes(v))+
  geom_bar(binwidth=0.25)+
  theme_bw()
```



```r
t.test(v, mu=4.95)
```

Step 2.2 Create a sine wave that is 1000 observations long and add error with a standard deviation of 0.2. Plot your observations. Now filter the observations in equally weighted sets of 5 and add the filtered observations as a red line.

```
y <- sin(seq(1, 100, length.out=1000)) + rnorm(1000, sd=0.2)

x <- 1:1000

f <- stats::filter(y, rep(.2, 5))

s.df <- data.frame(x,y,f)

ggplot(s.df)+
  geom_line(aes(x, y))+
  geom_line(aes(x, f), color="red", size=1.5)+
  theme_bw()
```

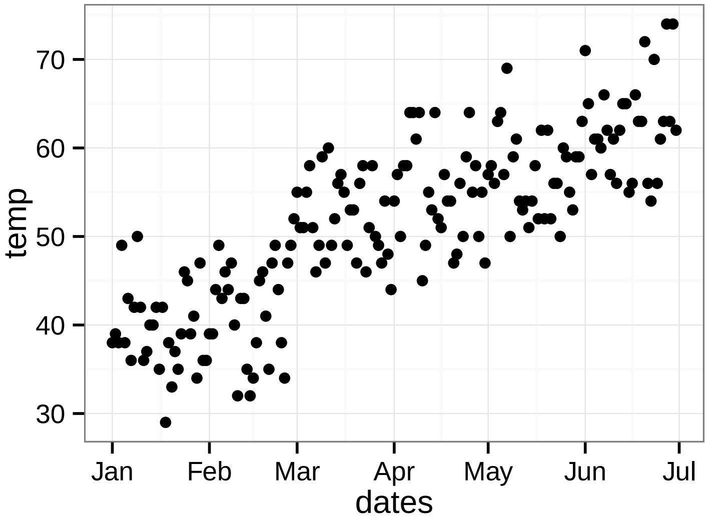


Note: I stole this question from Trevor Branch

Step 2.3 Create a data frame with a column for dates between January 1st, 2010 and June 30th, 2010 and a column for temperature. Randomly generate a temperature for each day with means of 40, 42, 51, 55, 58, 62 and a standard deviation of 5, then round to the nearest whole number. Plot your simulated temperature dataset and calculate the "observed" mean temperature for each month.

```
dates <- seq(ymd(20100101), ymd(20100630), by="day")

temp.data <- data.frame(dates, temp=rep(NA, length(dates)))

month <- months(dates)

month <- factor(month, levels = unique(month), ordered = TRUE)

month.days <- table(month)

means <- c(40, 42, 51, 55, 58, 62)

mean.vector <- rep(means, month.days)

temp.data$temp <- round(rnorm(n=length(mean.vector),
                              mean=mean.vector,
                              sd=5))
```

```
actualMeans <- tapply(temp.data$temp, month, mean)

ggplot(temp.data, aes(dates, temp))+
  geom_point()+
  theme_bw()
```



Task 3: ANOVA

Assuming that the `iris` dataset meets the criteria for ANOVA, test whether the sepal length of irises differs by species.

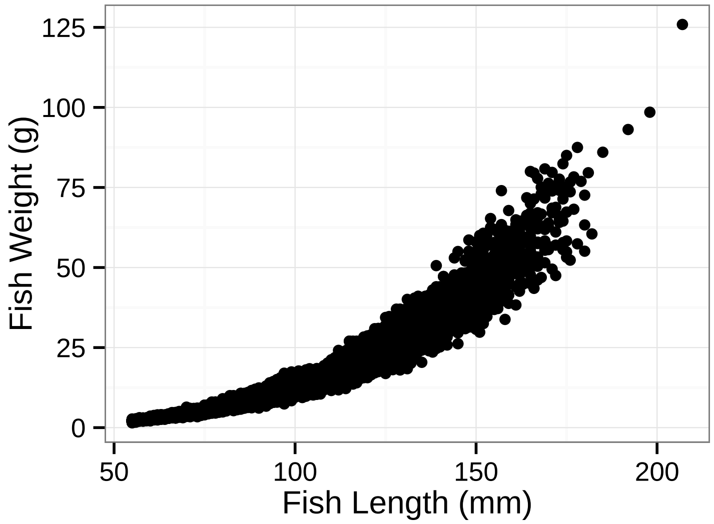
```
iris.df <- melt(iris, id.var="Species")

a.sl <- aov(value ~ Species, data = iris.df[iris.df$variable=="Sepal.Length",])

summary(a.sl)
```

Task 4: Analyzing Fish Lengths and Weights

Note: I stole this question from Derek Ogle

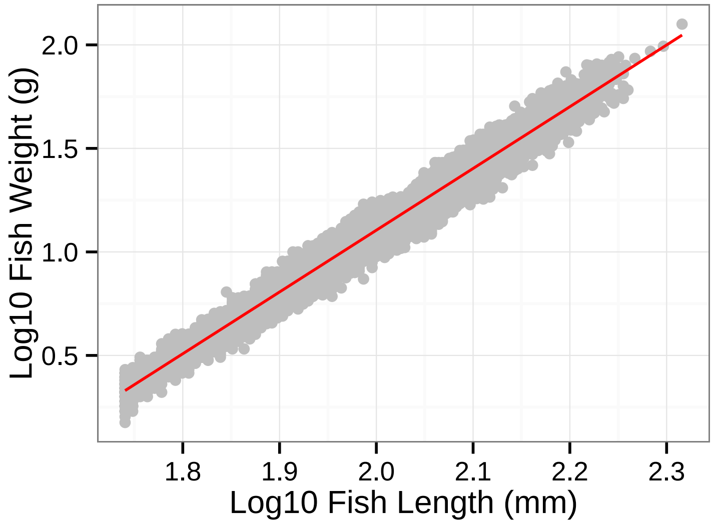
Step 4.1 Read in the data from `FishData.csv` and plot fish length v. weight. What do you notice about the plot? What do you hypothesize is the relationship between fish length and weight?

```
fish <- read.csv("FishData.csv", header=TRUE, stringsAsFactors=FALSE)

ggplot(fish, aes(tl, wt))+
  xlab("Fish Length (mm)")+
  ylab("Fish Weight (g)")+
#  scale_x_log10()+ # uncomment these lines to plot log-log relationship
#  scale_y_log10()+
  geom_point()+
  theme_bw()
```



Step 4.2 Use a linear model to describe the relationship between the log- transformed length and weight data. What are the coefficients of the model? What is the predicted weight of fish that are 125 mm long? Plot the model-predicted fish weight as a function of fish length.

```
fish <- mutate(fish, Log10L=log10(tl), Log10W=log10(wt))

lw.fit <- lm(Log10W~Log10L, data=fish)

summary(lw.fit)

new.data <- data.frame(Log10L=log10(125))

10^(predict(lw.fit, new.data, interval="confidence"))

ggplot(fish, aes(Log10L, Log10W))+
  xlab("Log10 Fish Length (mm)")+
  ylab("Log10 Fish Weight (g)")+
  geom_point(color="gray")+
  geom_smooth(method="lm", color="red")+
  theme_bw()
```


Log10 Fish Weight (g)
Log10 Fish Length (mm)