How to Use R Studio

HATS

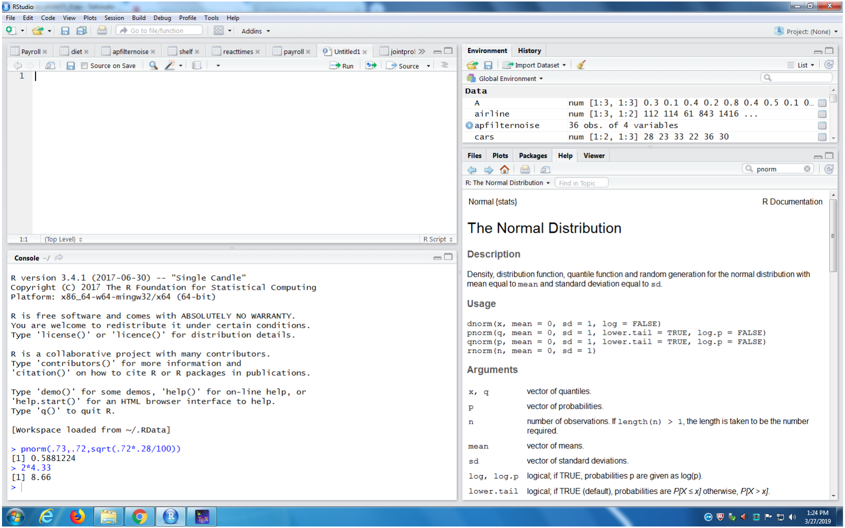
Cathy Poliak, Ph.D.
cathy@math.uh.edu
Office Fleming 11c

Department of Mathematics
University of Houston

Outline

R and R-studio

- Open source software (free) for statistical analysis.
- R download: `https://cran.cnr.berkeley.edu/`
- R-studio download: `https://www.rstudio.com/products/rstudio/download/`
- Help in R-studio: Right hand bottom panel.
- Today's R script: `https://www.math.uh.edu/~cathy/Math3339/HATS.R`


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Payroll diet apfilternoise shelf reacttimes payroll Untitled1 jointprol
Source on Save
Run
Source
1:1 (Top Level) R Script
Environment History
Import Dataset
List
Global Environment
Data
A num [1:3, 1:3] 0.3 0.1 0.4 0.2 0.8 0.4 0.5 0.1 0...
airline num [1:3, 1:2] 112 114 61 843 1416 ...
apfilternoise 36 obs. of 4 variables
cars num [1:2, 1:3] 28 23 33 22 36 30
Files Plots Packages Help Viewer
pnorm
R: The Normal Distribution
Find in Topic
Normal {stats}
R Documentation
The Normal Distribution
Description
Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.
Usage
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
Arguments
x, q vector of quantiles.
p vector of probabilities.
n number of observations. If length(n) > 1, the length is taken to be the number required.
mean vector of means.
sd vector of standard deviations.
log, log.p logical; if TRUE, probabilities p are given as log(p).
lower.tail logical; if TRUE (default), probabilities are P[X ≤ x] otherwise, P[X > x].
Console ~/
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
> pnorm(.73,.72,sqrt(.72*.28/100))
[1] 0.5881224
> 2*4.33
[1] 8.66
>
1:24 PM
3/27/2019

How To Input Data

- Preloaded data
- Packages: e.g. Mosaic data
- Excel file: Save an Excel file select "Import Data" » "From Excel" » input the file you want to import.
- Directly into R: x=c(1,5,6,10)
- Examples to download
  - Grades: `https://www.math.uh.edu/~cathy/Math3339/data/grades.txt`
  - ERA: `https://www.math.uh.edu/~cathy/Math3339/data/Era.txt`
  - Stress: `https://www.math.uh.edu/~cathy/Math3339/data/Stress.txt`

Example for Basic Statistics

```
> summary(grades)
 Student           Score              Grade               Tests       
 Min.   : 1.00   Min.   :  8.194   Length:30          Min.   : 14.67  
 1st Qu.: 8.25   1st Qu.: 54.853   Class :character   1st Qu.: 65.17  
 Median :15.50   Median : 82.305   Mode  :character   Median : 77.35  
 Mean   :15.50   Mean   : 71.094                      Mean   : 71.33  
 3rd Qu.:22.75   3rd Qu.: 90.639                      3rd Qu.: 91.88  
 Max.   :30.00   Max.   :103.955                      Max.   :103.32  

 Quiz            HW                 Opt-out              Session         
 Min.   :10.38   Min.   :  3.175   Length:30          Length:30         
 1st Qu.:67.85   1st Qu.: 58.174   Class :character   Class :character  
 Median :79.00   Median : 76.667   Mode  :character   Mode  :character  
 Mean   :71.18   Mean   : 69.837                                        
 3rd Qu.:87.27   3rd Qu.: 84.841                                        
 Max.   :99.23   Max.   :101.905                                        

> mean(grades$Tests)
[1] 71.32833
> sd(grades$Tests)
[1] 27.12584
> fivenum(grades$Tests)
[1]  14.66667  64.96667  77.35000  92.00000 103.31667
```

Gas Prices in Houston

I took a "Random" Sample of 30 stations from
`http://www.houstongasprices.com/GasPriceSearch.aspx`

```
> gasprice = c(2.25,2.27,2.32,2.35,2.35,2.35,2.39,2.39,2.39,2.39,2.39,
+              2.39,2.39,2.39,2.39,2.39,2.39,2.39,2.39,2.4,2.42,2.44,
+              2.45,2.45,2.45,2.49,2.51,2.54,2.59,2.59)
> mean(gasprice)
[1] 2.409667
> median(gasprice)
[1] 2.39
> sd(gasprice)
[1] 0.07730296
> fivenum(gasprice)
[1] 2.25 2.39 2.39 2.45 2.59
```

Quantiles or Percentiles

- Let $p \in (0, 1)$ be a number between 0 and 1. The p^{th} quantile of x is more commonly known as the $100p^{th}$ percentile; e.g., the 0.8 quantile is the same as the 80th percentile.
- The p**th percentile** of data is the value such that p percent of the observations fall at or below it.
- If you are looking for the measurement that has a desired percentile rank, the $100P^{th}$ percentile, is the measurement with rank (or position in the list) of $nP + 0.5$, where n represents the number of data values in the sample.

Determining Percentiles (Quantiles)

- In R-studio there are several different ways to determine quantiles in R studio. For more information you can type ?quantile in the console.
- The type that is describe previously is type 5.
- getting the 95th percentile.
-
```
> quantile(gasprice,0.95,type = 5)
95%
2.59
```

Stem-and-Leaf Plot

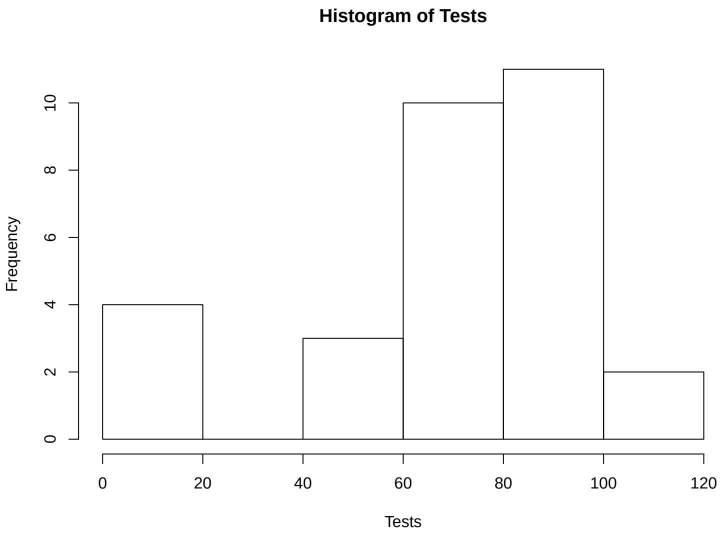
```
> stem(grades$Tests)

The decimal point is 1 digit(s) to the right of the |

1 | 5677
2 |
3 |
4 | 446
5 |
6 | 56
7 | 02345789
8 | 346
9 | 02245699
10 | 13
```

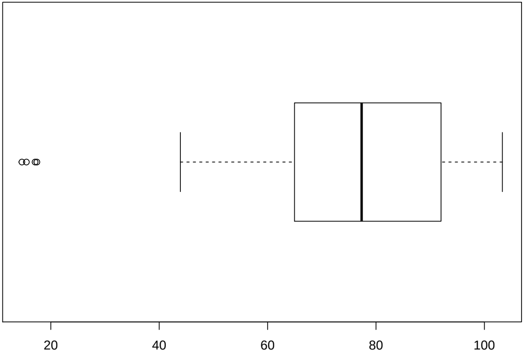
Histogram

```
hist(grades$Tests,main = "Histogram of Tests", xlab = "Tests")
```

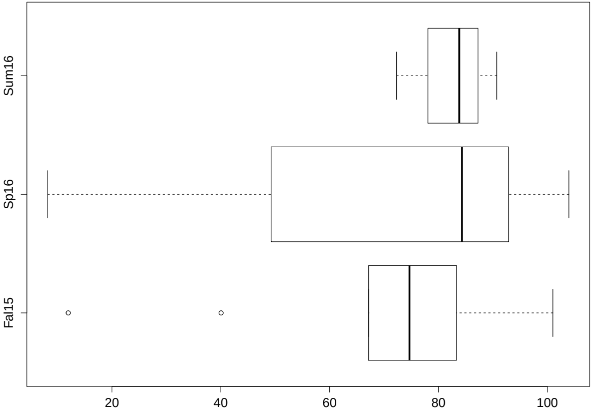


Boxplot of Test Scores

boxplot(grades$Tests,horizontal = T)



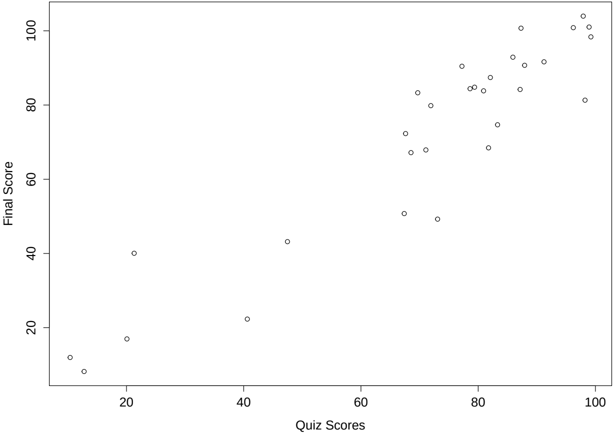
Boxplot of Course Scores by Session




```
boxplot(grades$Score~grades$Session,horizontal=TRUE)
```

Scatterplot

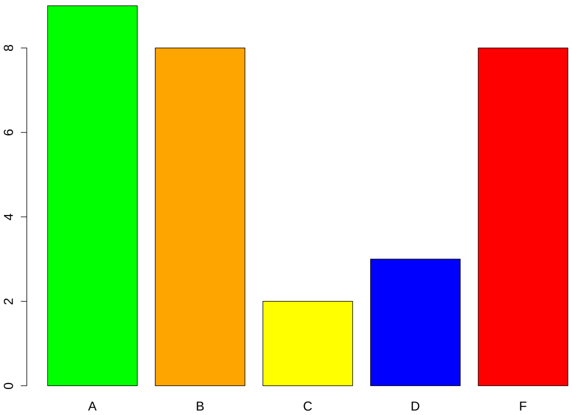
```
plot(grades$Quiz,grades$Score,xlab="Quiz Scores",ylab="Final
Score")
```



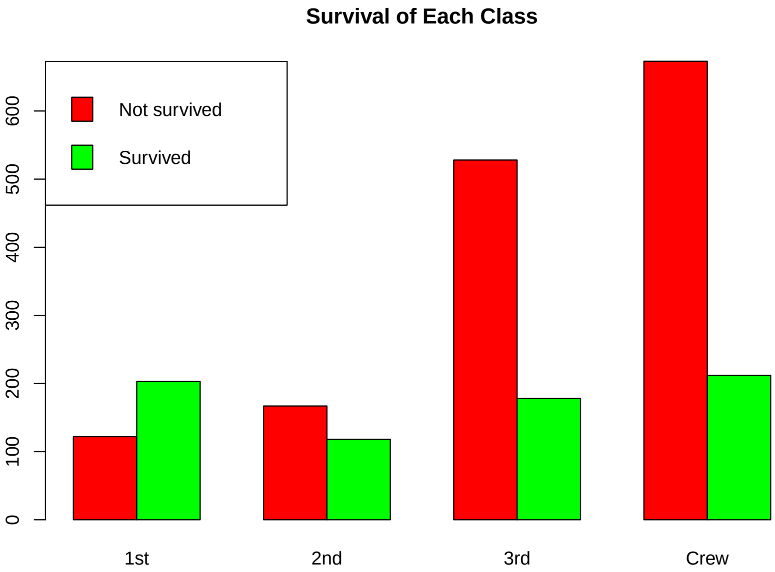
Bar Graphs

To do a bar graph you have to put the data into a table

```
counts=table(grades$Grade)
barplot(counts,col=c("green","orange","yellow","blue","red"))
```



Side by Side Bar Graph



Code

```
titanic.data=margin.table(Titanic,c(4,1))
#cross table of survival by class
barplot(titanic.data,
main = "Survival of Each Class",
xlab = "Class",
col = c("red","green"),
beside=T
)
legend("topleft",
c("Not survived","Survived"),
fill = c("red","green")
)
```

Finding Probabilities for Popular Distributions

- For any "named" distribution we can use R to find the probabilities and the quantiles.
- To find $P(X = x) = d\ldots$(x,list of parameters).
- To find $P(X \leq x) = p\ldots$(x,list of parameters).
- To find c such that $P(X \leq c) = p$, $c = q\ldots$(p,list of parameters).

Binomial Distribution

Suppose that in a large metropolitan area, 80% of all households have a flat screen television. Suppose you are interested in selecting a group of six households from this area. Let X be the number of households in a group of six households from this area that have a flat screen television.

1. For what proportion of groups will exactly four of the six households have a flat screen television?

```
> dbinom(4,6,0.8)
[1] 0.24576
```

2. For what proportion of groups will at most two of the households have a flat screen television?

```
 pbinom(2,6,0.8)
[1] 0.01696
```

3. What is the probability that between 2 and 4 inclusive will have a flat screen television?

Normal Distribution

The length of time needed to complete a certain test is normally distributed with mean 77 minutes and standard deviation 11 minutes. Find the probability that it will take between 74 and 80 minutes to complete the test.

More Normal Distribution

Part a: Let Z be the standard normal random variable. Calculate the following.

1. $P(Z < 2.4) =$
```
pnorm(2.4)
[1] 0.9918025
```

2. $P(Z > -1.9) =$
```
1-pnorm(-1.9)
[1] 0.9712834
```

3. Find c such that $P(Z > c) = 0.98$

```
qnorm(1-0.98)
[1] -2.053749
```

More Normal Distribution

Part b: Let X be a normal random variable with a mean of 47 and a standard deviation of 3. Calculate the following.

1. $P(X < 50.4) =$
2. $P(43.5 < X < 50.4) =$
3. Find x such that $P(X < x) = 0.74$

```
> pnorm(50.4,47,3)
[1] 0.8714629
> pnorm(50.4,47,3) - pnorm(43.5,47,3)
[1] 0.7497903
> qnorm(.74,47,3)
[1] 48.93004
>
```

Sampling Distribution of $\bar{X}$

A random sample of 1024 12-ounce cans of fruit nectar is drawn from among all cans produced in a run. Prior experience has shown that the distribution of the contents has a mean of 12 ounces and a standard deviation of 0.12 ounce. What is the probability that the mean contents of the 1024 sample cans is less than 11.994 ounces?

```
 pnorm(11.994,12,0.12/sqrt(1024))
[1] 0.05479929
```

Sampling Distribution of $\hat{p}$

In a large population, 67% of the households have cable tv. A simple random sample of 256 households is to be contacted and the sample proportion computed. What is the mean and standard deviation (standard error) of the sampling distribution of the sample proportions? What is the probability that the sampling distribution of sample proportions is less than 73%?

```
> pnorm(.73,.67,sqrt(.67*.33/256))
[1] 0.9794058
```

Confidence Intervals

1. A random sample of 64 observations produced a mean value of 73 and standard deviation of 6.5. Determine a 90% confidence interval for the population mean μ.

```
> #Confidence Intervals
> 73+c(1,-1)*qt(0.05,63)*6.5/sqrt(64)
[1] 71.64361 74.35639
```

2. A random sample of 121 observations produced a sample proportion 35%. Determine an approximate 95% confidence interval for the population proportion.

```
> 0.35+c(1,-1)*qnorm(0.025)*sqrt(.35*.65/121)
[1] 0.2650143 0.4349857
```

How good is a Pitcher for MLB?

- In MLB is the number of wins is attributed to the starting pitcher. Also, the ERA (earned run average) is calculated for the pitcher. Can we use ERA to predict the number of wins that is attributed to a pitcher?
- The following data is from the 2015 baseball season: `https://www.math.uh.edu/~cathy/Math3339/data/Era.txt`
- We will use R to:
  - Construct a scatterplot.
  - Find the LSRL and fit it to the scatterplot.
  - Find r and r^2.
  - Does there appear to be a linear relationship between the two variables? Based on what you found, would you characterized the relationship as positive or negative? Strong or weak?
  - Draw the residual plot.
  - What does the residual plot reveal?
  - `http://insider.espn.com/mlb/insider/story/_/id/13752413/atlanta-braves-pitcher-shelby-miller-terrible-luck-`

One-Sample T-test

Quart cartons of milk should contain at least 32 ounces. A sample of 22 cartons contained the following amounts in ounces. Does sufficient evidence exist to conclude the mean amount of milk in cartons is less than 32 ounces? The data is: (31.5, 32.2, 31.9, 31.8, 31.7, 32.1, 31.5, 31.6, 32.4, 31.6, 31.8, 32.2, 32.1, 31.8, 31.6, 32.0, 31.6, 31.7, 32.0, 31.9, 31.8, 31.6)

```
> t.test(milk,mu = 32, alternative =
"less",conf.level = 0.95)

 One Sample t-test

data:  milk
t = -3.1677, df = 19, p-value = 0.002534
altemative hypothesis: true mean is less
than 32
95 percent confidence interval:
    -Inf 31.35284
sample estimates:
mean of x
   30.575
```

Two-sample T-test

Is there a difference in the mean miles per gallon of a Honda Civic and a Toyota Prius? The following is data from 5 Honda's and 6 Toyota's:

Honda	32.2	29.8	29.7	29.7	28.1	
Toyota	36.5	33	33	31.7	31	28.8

```
> honda = c(32.2,29.8,29.7,29.7,28.1)
> toyota = c(36.5,33,33,31.7,31,28.8)
> t.test(honda,toyota,alternative = "two.sided",conf.level = 0.95)

 Welch Two Sample t-test

data:  honda and toyota
t = -1.9684, df = 8.1315, p-value = 0.08396
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.2759386  0.4092719
sample estimates:
mean of x mean of y
 29.90000  32.33333
```

Matched Pair Test

In a experiment on relaxation techniques, subject's brain signals were measured before and after the relaxation exercises with the following results:

Person	1	2	3	4	5
Before	32	38	65	50	30
After	25	35	56	52	24

Is there sufficient evidence to suggest that the relaxation exercise slowed the brain waves? Assume the population is normally distributed.

```
> before = c(32,38,65,50,30)
> after = c(25,35,56,52,24)
> t.test(before,after,alternative = "greater",conf.level = 0.9,paired = T)

 Paired t-test

data:  before and after
t = 2.4045, df = 4, p-value = 0.037
alternative hypothesis: true difference in means is greater than 0
90 percent confidence interval:
 1.666804      Inf
sample estimates:
mean of the differences
                    4.6
```

Stress

A study was conducted to examine the effect of pets in stressful situations. Fifteen subjects were randomly assigned to each of three groups to do a stressful task alone (the control group), with a good friend present, or with their dog present. The subject's mean heart rate (in beats per minutes) during the task is one measure of the effect of stress. The data has is the mean heart rates during stress with a pet (P), with a friend (F) and for the control group (C).

- Make a side by side box plot of the heart rates by the three groups. To do this in R use: boxplot(Rate Group,data=Stress)
- Does the data suggest that there is a difference among the three groups?
- If there seems to be a difference, complete a Bonferroni pairwise test to determine which or if all the means are different from each other.

```
> boxplot(Rate~Group,data = Stress)
> stress.lm = lm(Rate~Group,data = Stress)
> anova(stress.lm)
Analysis of Variance Table

Response: Rate
          Df Sum Sq Mean Sq F value    Pr(>F)
Group      2 2387.7 1193.84  14.079 2.092e-05 ***
Residuals 42 3561.3   84.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> pairwise.t.test(Stress$Rate,Stress$Group,"bon")

 Pairwise comparisons using t tests with pooled SD

data:  Stress$Rate and Stress$Group

  C     F
F 0.037 -
P 0.031 1.2e-05

P value adjustment method: bonferroni
```

Chi-Square Test

The Blue Diamond Company advertises that their nut mix contains (by weight) 40% cashews, 15% Brazil nuts, 20% almonds and only 25% peanuts. The truth-in-advertising investigators took a random sample (of size 20 lbs) of the nut mix and found the distribution to be as follows: 6 lbs of Cashews, 3 lbs of Brazil nuts, 5 lbs of Almonds and 6 lbs of Peanuts. At the 0.01 level of significance, is the claim made by Blue Diamond true?

1. Calculate the test statistic for this test.

2. Determine the p-value.

3. Give the decision to Reject H_0 or Fail to Reject H_0.

```
> pnuts = c(.4,.15,.2,.25)
> nuts=c(6,3,5,6)
> chisq.test(nuts,p=pnuts)

 Chi-squared test for given
 probabilities

data:  nuts
X-squared = 0.95, df = 3, p-value = 0.8133

Warning message:
In chisq.test(nuts, p = pnuts) : Chi-squared approximation may be incorrect
```

Fair Die

A six-sided die is thrown 50 times. The numbers of occurrences of each face are shown below.

Face	1	2	3	4	5	6
Count	12	5	9	11	6	7

Can you conclude that the die is not fair?

```
> chisq.test(count)

 Chi-squared test for given probabilities

data:  count
X-squared = 4.72, df = 5, p-value = 0.451

> chisq.test(c(12,5,9,11,6,7))

 Chi-squared test for given probabilities

data:  c(12, 5, 9, 11, 6, 7)
X-squared = 4.72, df = 5, p-value = 0.451
```

Example

The following table shows three different airlines **row variable** and the number of delayed or on-time flights **column variable** from flightstats.com.

	Delayed	On-time	Total
American	112	843	955
Southwest	114	1416	1530
United	61	896	957
Total	287	3155	3442

- Does on-time performance depend on airline?
- We will use a significance test to answer this question.

Chi-square Test Using R

1. Input the data as a matrix.
2. R-code: chisq.test(matrix name,correction=FALSE)

```
> airline<-matrix(c(112,114,61,843,1416,896),nrow=3,ncol=2)
> chisq.test(airline)

Pearson's Chi-squared test

data:  airline
X-squared = 20.762, df = 2, p-value =3.102e-05
```

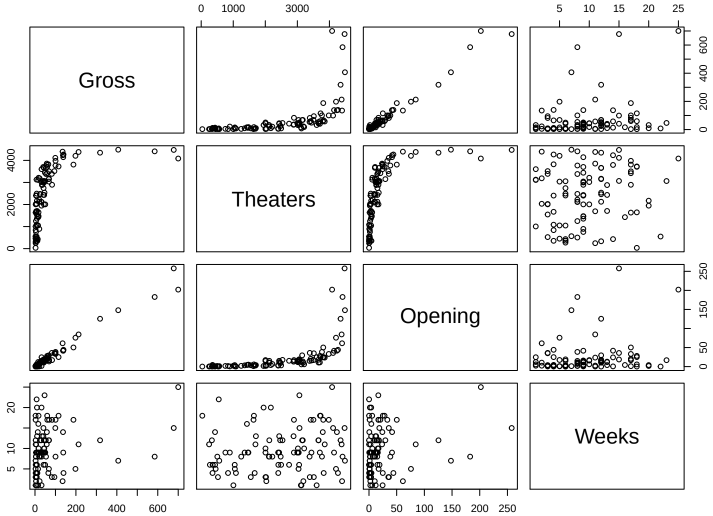
Can we predict total gross for a movie

`https://www.math.uh.edu/~cathy/data/movies.csv`

- response variable - Total Gross, in million dollars
- predictor 1 - Opening Weekend Gross, in million dollars
- predictor 2 - Theaters
- predictor 3 - Number of weeks open
- Top 100 gross movies of 2018 as of August 8.

Scatterplots of Movie Variables

```
pairs(movies[,3:6])
```



Movie Data

```
movieall.lm=lm(Gross~Theaters+Opening+Weeks)
summary(movieall.lm)

Call:
lm(formula = Gross ~ Theaters + Opening + Weeks)

Residuals:
Min       1Q  Median      3Q     Max
-73.513  -7.733   0.363   4.634  95.983

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.101133   5.403947  -1.314 0.191956
Theaters    -0.002171   0.001850  -1.173 0.243576
Opening      2.904524   0.057292  50.697  < 2e-16 ***
Weeks        1.331971   0.364575   3.653 0.000422 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.15 on 96 degrees of freedom
Multiple R-squared:  0.9762,Adjusted R-squared:  0.9754
F-statistic:  1310 on 3 and 96 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Gross}} = -7.10113 - 0.002171 \times \text{Theaters} + 2.904524 \times \text{Opening} + 1.331971 \times \text{Weeks}$$

What If We have Several Predictors?

The **stepwise** regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.
There are three strategies of stepwise regression (James et al. 2014,P. Bruce and Bruce (2017)):

1. **Forward** selection, which starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.

2. **Backward** selection (or backward elimination), which starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.

3. **Stepwise** selection (or sequential replacement), which is a combination of forward and backward selections. You start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection).

We can use the function step() in R to select the predictors.

R Output

```
step(movieall.lm)
Start:  AIC=594.4
Gross ~ Theaters + Opening + Weeks

Df Sum of Sq    RSS    AIC
- Theaters  1       505  35716 593.82
<none>                   35212 594.40
- Weeks     1      4896  40107 605.41
- Opening   1    942720 977931 924.80
```

Step 2

```
Step:  AIC=593.82
Gross ~ Opening + Weeks

Df Sum of Sq     RSS    AIC
<none>                    35716 593.82
- Weeks     1       4791   40507 604.41
- Opening  1   1350583 1386300 957.70

Call:
lm(formula = Gross ~ Opening + Weeks)

Coefficients:
(Intercept)  Opening        Weeks
-11.436        2.866        1.317
```