A Gentle Introduction to R

James Church

February 12, 2013

A Gentle Introduction to R

Outline

- Why R?
- Introduction
- Random Number Generation
- Computing Pi
- Data Analysis with Data Frames
- Multiple Linear Regression

Why R?

- You wish to analyze.
- You wish to create beautiful plots.

Tools Used in this Presentation

- R
- RStudio

Data Types

The (IMO) Important Data Types

- Numerical
- Character
- Logical (TRUE, T, FALSE, F)
- Vectors
- Data Frames
- Factors

Getting Started

Which of the three statements is the correct way to assign a variable in R?

```
> a <- 1
> 2 -> b
> c = 3
```

Answer: All three. The "R way"^{TM} of assigning variables is to use the left arrow.

Vectors :: 1

Like arrays, but better.

```
> my.vector <- c(1, 3, 7, 15, 31)
> my.vector + 1
[1]  2  4  8 16 32
> my.vector * 2
[1]  2  6 14 30 62
> my.vector ^ 2
[1]   1   9  49 225 961
> my.vector > 10
[1] FALSE FALSE FALSE  TRUE  TRUE
```

Vectors :: 2

Like arrays, but better.

```
> my.vector <- c(1, 3, 7, 15, 31)
> sum(my.vector)
[1] 57
> sum(my.vector>10)
[1] 2
> mean(my.vector)
[1] 11.4
> median(my.vector)
[1] 7
> sd(my.vector)
[1] 12.19836
```

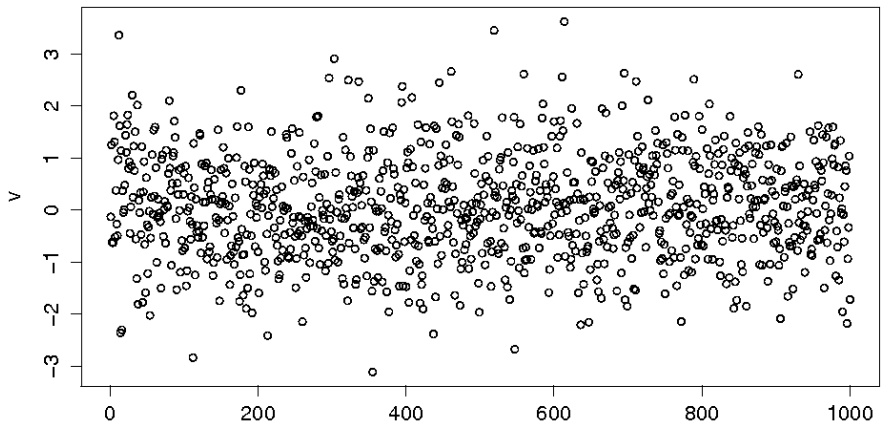
Random Numbers

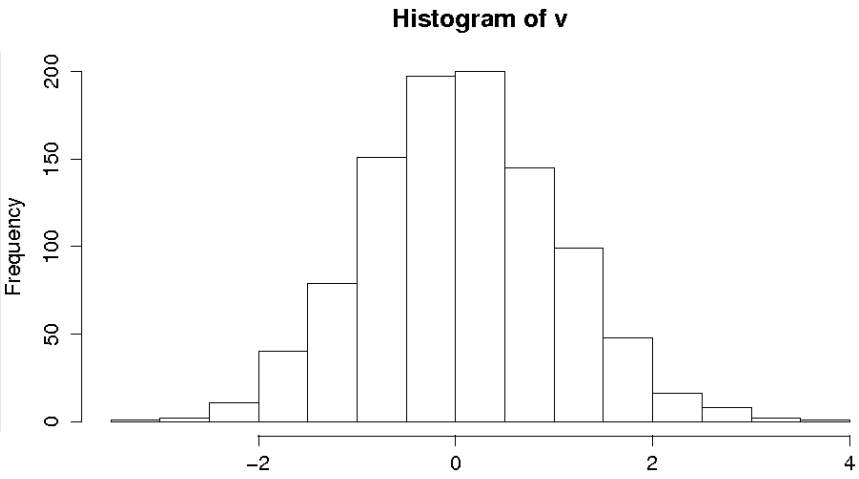
- Normal: rnorm(n, mean=0, sd=1)
- Uniform: runif(n, low=0, high=1)

Quick Test for Normal

Is it true that 'rnorm' produces a normal curve with a mean of 0 and a standard deviation of 1?

```
> v <- rnorm(1000)
> mean(v)
[1] 0.06681055
> sd(v)
[1] 0.9872857
> plot(v)
> hist(v)
```

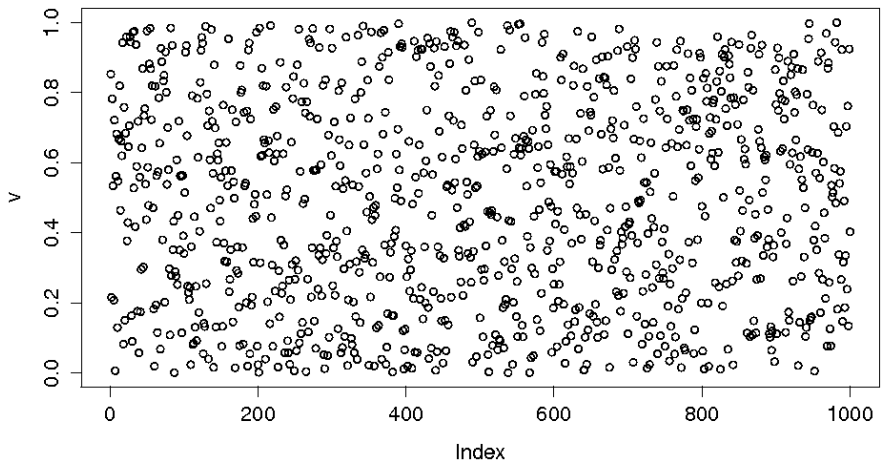


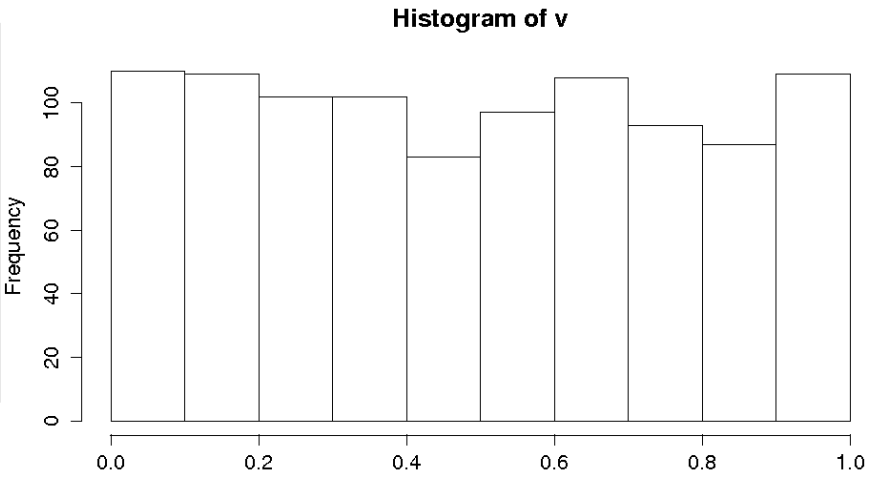


Quick Test for Uniform

Is it true that 'runif' produces a uniform curve with a low of 0 and a high of 1?

```
> v <- runif(1000)
> min(v)
[1] 0.0007571692
> max(v)
[1] 0.9999074
> plot(v)
> hist(v)
```





Computing Pi using Random Numbers :: Theory

- A square with sides of length 2 has its center at the origin.
- A circle with a radius of length 1 has its center at the origin.
- The length of the side of the square is 2 times the radius of the circle.
- Area of the square: $(2r)^2$ or $4r^2$ or 4.
- Area of the circle: πr^2 or π.
- The ratio of the area of the two shapes is
  $\frac{\pi}{4} = \frac{Area\ of\ the\ Circle}{Area\ of\ the\ Square}$
- This means we can write π as
  $\pi = \frac{4 \times Area\ of\ the\ Circle}{Area\ of\ the\ Square}$

Computing Pi using Random Numbers :: Practice

- Imagine our square is now a dartboard.
- We have 100,000 darts to throw at the dartboard.
- All 100,000 darts must land in the square.
- We count the darts landing in the circle.
- We then compute π:
  $\pi = \frac{4 \times \textit{Area of the Circle}}{\textit{Area of the Square}} = \frac{4 \times \textit{Darts in the Circle}}{100{,}000}$

Computing Pi using Random Numbers :: Code

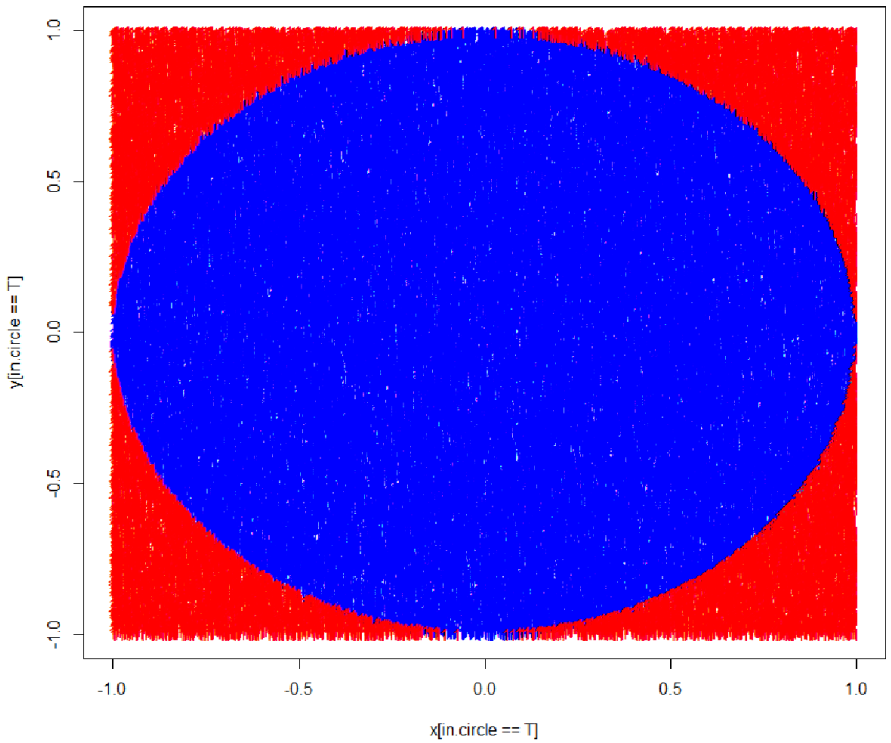
```
> # Start throwing darts
> n <- 100000
> x <- runif(n, -1, 1)
> y <- runif(n, -1, 1)

> # Determine which darts are in the circle
> in.circle <- sqrt(x^2 + y^2)<=1

> # Estimate pi and calcualte the error.
> estimated.pi <- 4 * sum(in.circle) / n
> estimated.pi
[1] 3.14512
> estimated.pi.error <- 100*abs(estimated.pi - pi)/pi
> estimated.pi.error
[1] 0.1122789
```

Computing Pi using Random Numbers :: Plotting

```
> matplot(x[in.circle==T], y[in.circle==T],
          col='blue');
> matplot(x[in.circle==F], y[in.circle==F],
          col='red', add=T)
```



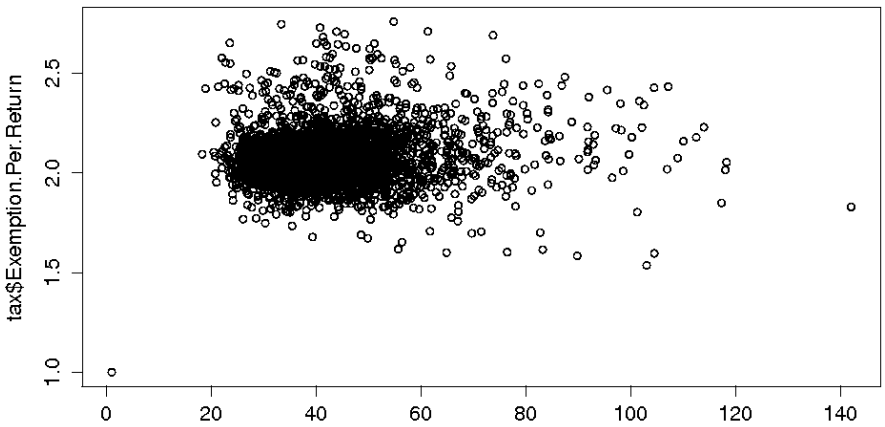
Data Analysis :: Reading CSV Files

```
> setwd("~/code/RIntro") # Your working directory
> tax <- read.csv('Tax_Year_2007_County_Income_Data.csv')
> summary(tax)
     Wages                Dividend            Interest
 Min.   :       -1   Min.   :      -1   Min.   :      -1
 1st Qu.:   125193   1st Qu.:    2434   1st Qu.:    6200
 Median :   320627   Median :    7234   Median :   14626
 Mean   :  3327009   Mean   :   98482   Mean   :  155972
 3rd Qu.:   999196   3rd Qu.:   25503   3rd Qu.:   41676
 Max.   :669494988   Max.   :19742493   Max.   :34132623
> names(tax)
 [1] "State.Code"       "County.Code"      "State.Abbr"
     "County"           "Num.Tax.Returns" "Num.Exemptions"
     "Adjusted.Gross"
 [8] "Wages"            "Dividend"         "Interest"
```

Data Analysis :: Preparing Our Data

Does wealth influence the number of exemptions?

```
> tax$Adjusted.Gross.Per.Return <- tax$Adjusted.Gross / tax$Num.Tax.Ret
> tax$Exemption.Per.Return <- tax$Num.Exemptions / tax$Num.Tax.Returns
> median(tax$Adjusted.Gross.Per.Return)
[1] 40.77554
> median(tax$Exemption.Per.Return)
[1] 2.066597
> plot(tax$Adjusted.Gross.Per.Return, tax$Exemption.Per.Return)
```



Data Analysis :: Performing the Test

Does the number of exemptions **depend** on the adjusted gross income?

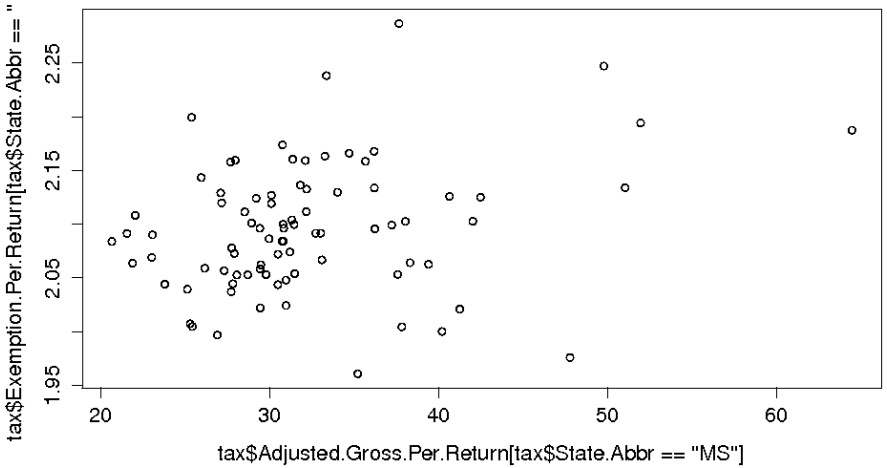
```
> summary(lm(tax$Exemption.Per.Return ~ tax$Adjusted.Gross.Per.Return))
Coefficients:
                Estimate Std.       Error t value Pr(>|t|)
(Intercept)        2.0408565  0.0084952 240.237  < 2e-16 ***
tax$Adj...         0.0008869  0.0001891   4.691 2.83e-06 ***

Residual standard error: 0.1372 on 3191 degrees of freedom
Multiple R-squared: 0.006849,   Adjusted R-squared: 0.006538
F-statistic: 22.01 on 1 and 3191 DF,  p-value: 2.832e-06
```

Data Analysis :: Preparing Our Data

Does wealth influence the number of exemptions in Mississippi?

```
> plot(tax$Adjusted.Gross.Per.Return[tax$State.Abbr=="MS"],
       tax$Exemption.Per.Return[tax$State.Abbr=="MS"])
```



Data Analysis :: Performing the Test

Does the number of exemptions **depend** on the adjusted gross income in Mississippi?

```
> summary(lm(tax$Exemption.Per.Return[tax$State.Abbr=="MS"] ~
             tax$Adjusted.Gross.Per.Return[tax$State.Abbr=="MS"]))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.0261054  0.0291375  69.536   <2e-16 ***
tax$Adj...  0.0021531  0.0008805   2.445   0.0166 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.05834 on 81 degrees of freedom
Multiple R-squared: 0.06875,    Adjusted R-squared: 0.05725
F-statistic: 5.979 on 1 and 81 DF,  p-value: 0.01664
```

A Story About A Restaurant

To Bus Station (m)	Floor Space (m^2)	Manager Age	Sales
80	10	42	469
0	8	29	366
200	8	33	371
200	5	41	208
300	7	33	246
230	8	35	297
40	7	40	363
0	9	46	436
330	6	44	198
180	9	34	346

The Code

```
toBusStation <- c(80,0,200,200,300,230,40,0,330,180)
floorSpace <- c(10,8,8,5,7,8,7,9,6,9)
shopManagerAge <- c(42,29,33,41,33,35,40,46,44,34)
monthlySales <- c(469,366,371,208,246,297,363,436,198,364)
```

The Report :: All 3 Variables

```
> summary(lm(monthlySales ~ toBusStation + floorSpace
                          + shopManagerAge))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    17.69976   88.27609   0.201 0.847710
toBusStation   -0.33271    0.08184  -4.066 0.006609 **
floorSpace     42.20883    6.56029   6.434 0.000667 ***
shopManagerAge  1.08740    1.52345   0.714 0.502175
---
Residual standard error: 25.32 on 6 degrees of freedom
Multiple R-squared: 0.9495, Adjusted R-squared: 0.9243
F-statistic: 37.62 on 3 and 6 DF,  p-value: 0.000276
```

The Report :: Best 2 Variables

```
> summary(lm(monthlySales ~ toBusStation + floorSpace))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.32392   55.73834   1.172 0.279546
toBusStation -0.34088    0.07814  -4.362 0.003304 **
floorSpace   41.51348    6.25612   6.636 0.000294 ***

Residual standard error: 24.42 on 7 degrees of freedom
Multiple R-squared: 0.9452, Adjusted R-squared: 0.9296
F-statistic: 60.41 on 2 and 7 DF,  p-value: 3.844e-05
```

The Report :: Best Variable

```
> summary(lm(monthlySales ~ floorSpace))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -91.279     76.914  -1.187 0.269366
floorSpace    54.945      9.824   5.593 0.000514 ***

Residual standard error: 44.04 on 8 degrees of freedom
Multiple R-squared: 0.7964, Adjusted R-squared: 0.7709
F-statistic: 31.28 on 1 and 8 DF,  p-value: 0.0005144
```

The Conclusion to our Restaurant Story

Using the 'lm' (Linear Model) function in R, we were able to rank the three independent variables by order of importance:

- Floor Space
- Distance from Store to Bus Stop
- Manager Age

We discovered that both the size of the restaurant and the distance to the bus stop had an impact on the monthly sales of a restaurant. The age of the manager only had a noisy impact on the monthly sales.

Conclusions to this Presentation

- The Good: Plotting data
- The Good: Analyzing data
- The Good: Working with large amounts of data
- The Good: R+RStudio provide a clean, relaxed environment for working with data.
- The Bad: The R syntax is non-intuitive.

Resources

- R: `http://www.r-project.org/`
- RStudio: `http://www.rstudio.com/`
- This Presentation: `https://github.com/jcchurch/RIntro`
- Manga Guide to Statistics: `http://nostarch.com/mg_statistics.htm`