Statistics Lecture 2: Quantitative Data

Manuel

04/12/2015

We'll work with our own earthquake data, from earthquake.usgs.gov. I downloaded a Comma-Separated Value table with all the earthquakes for March 2015.

Preparation

Let's read the file:

```
earthquakes=read.csv("earthquakes.csv", header=TRUE) # reads a CSV file into a data frame
summary(earthquakes) # Gives summary statistics for all the values in the data frame
```

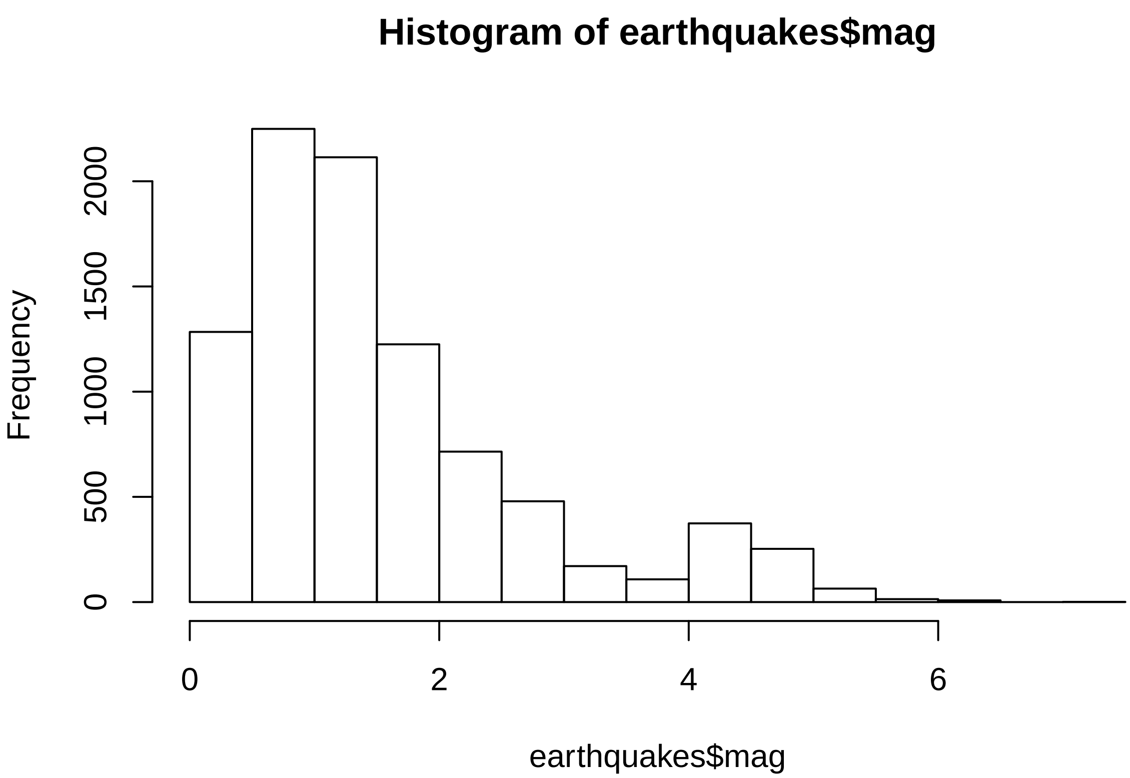
```
##                         time          latitude        longitude
##  2015-03-06T14:22:59.500Z:   2   Min.   :-71.73   Min.   :-180.0
##  2015-03-08T10:36:59.340Z:   2   1st Qu.: 35.70   1st Qu.:-148.7
##  2015-03-22T20:59:54.000Z:   2   Median : 38.82   Median :-122.3
##  2015-03-27T21:13:31.750Z:   2   Mean   : 41.15   Mean   :-112.2
##  2015-03-01T00:00:19.000Z:   1   3rd Qu.: 59.47   3rd Qu.:-116.7
##  2015-03-01T00:02:50.000Z:   1   Max.   : 80.42   Max.   : 180.0
##  (Other)                 :9049
##      depth             mag            magType          nst
##  Min.   : -3.42   Min.   :0.000   ml     :5318   Min.   :  0.00
##  1st Qu.:  3.52   1st Qu.:0.750   md     :2404   1st Qu.:  9.00
##  Median :  9.12   Median :1.230   mb     : 706   Median : 14.00
##  Mean   : 26.17   Mean   :1.564   Md     : 330   Mean   : 18.05
##  3rd Qu.: 20.25   3rd Qu.:2.000   mb_lg  : 108   3rd Qu.: 22.00
##  Max.   :649.74   Max.   :7.500   mc     :  63   Max.   :171.00
##  NA's   :1                        (Other): 130   NA's   :3151
##       gap            dmin              rms              net
##  Min.   : 13    Min.   : 0.0000   Min.   :0.0000   ak     :2737
##  1st Qu.: 69    1st Qu.: 0.0180   1st Qu.:0.0900   nc     :2097
##  Median : 99    Median : 0.0601   Median :0.2100   ci     :1495
##  Mean   :118    Mean   : 0.5443   Mean   :0.3175   us     :1159
##  3rd Qu.:148    3rd Qu.: 0.1662   3rd Qu.:0.5100   nn     : 688
##  Max.   :358    Max.   :43.7960   Max.   :1.9800   uw     : 243
##  NA's   :1885   NA's   :2966      NA's   :17       (Other): 640
##           id                             updated
##  ak11520339:   1   2015-03-01T01:25:04.702Z:   1
##  ak11520340:   1   2015-03-01T01:36:05.299Z:   1
##  ak11520343:   1   2015-03-01T01:48:05.053Z:   1
##  ak11520344:   1   2015-03-01T01:58:05.461Z:   1
##  ak11520348:   1   2015-03-01T05:32:03.368Z:   1
##  ak11520349:   1   2015-03-01T06:07:02.261Z:   1
##  (Other)   :9053   (Other)                 :9053
##                                 place              type
##  13km SE of Anza, California      : 149   earthquake  :8937
##  6km NW of The Geysers, California: 137   explosion   :  47
```

```
##  68km ESE of Lakeview, Oregon     :  87   quarry blast:  74
##  14km SE of Anza, California      :  74   sonicboom   :   1
##  69km ESE of Lakeview, Oregon     :  74
##  3km W of Cobb, California        :  68
##  (Other)                          :8470
```

Graphs

Histograms

```r
hist(earthquakes$mag)
```



Stem and Leaf

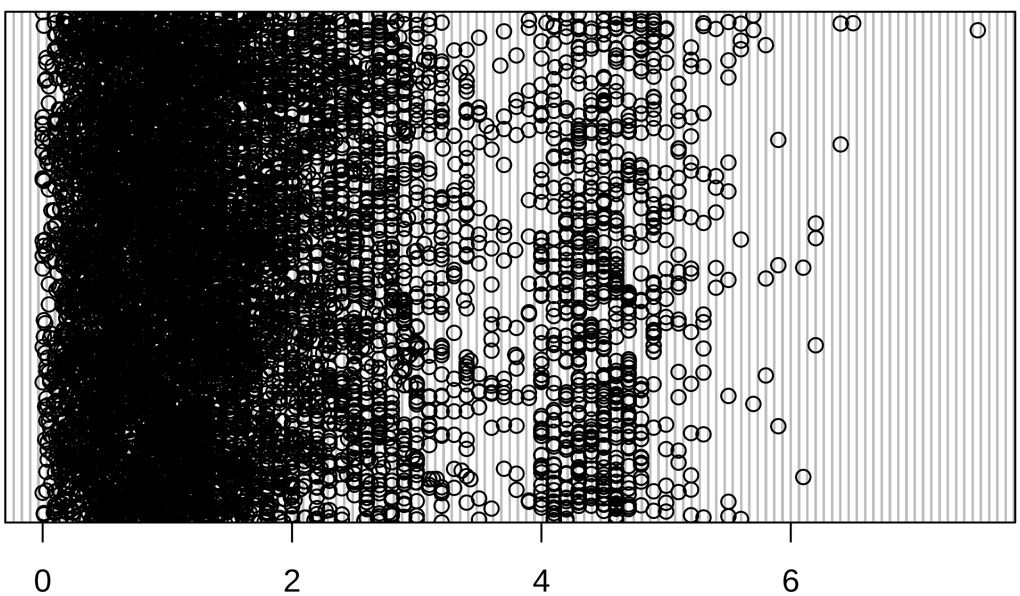
```r
stem(earthquakes$mag)
```

```
##
##   The decimal point is at the |
##
##   0 | 000000000000000000000000000000000000000000000000000011111111111111111+974
##   0 | 555555555555555555555555555555555555555555555555555555555555555555555+2056
##   1 | 000000000000000000000000000000000000000000000000000000000000000000000+2100
##   1 | 555555555555555555555555555555555555555555555555555555555555555555555+1274
```

```
##    2 | 00000000000000000000000000000000000000000000000000000000000000000000+649
##    2 | 55555555555555555555555555555555555555555555555555555555555555555555+469
##    3 | 00000000000000000000000000000000000000000000000000000000000000000111+140
##    3 | 55555555555555566666666666666667777777777777777888888888888889999999
##    4 | 00000000000000000000000000000000000000000000000000111111111111111111+252
##    4 | 55555555555555555555555555555555555555555555555555555555555555555555+243
##    5 | 00000000000000000000011111111111111111111111222222222222222233333333 3
##    5 | 55555555566666777888999
##    6 | 1122244
##    6 | 5
##    7 |
##    7 | 5
```

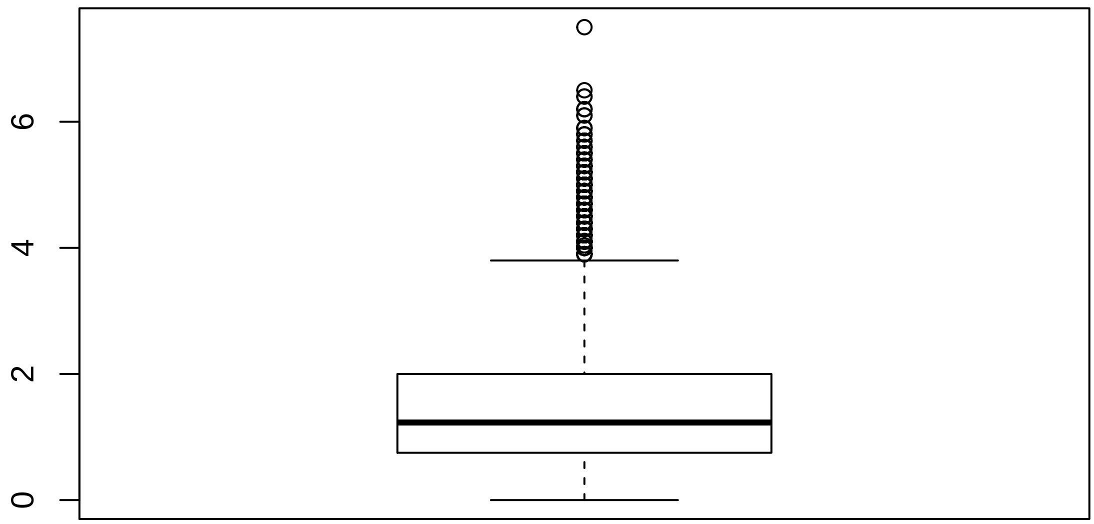
Dotplots

```r
dotchart(earthquakes$mag)
```



Boxplots

```r
boxplot(earthquakes$mag)
```



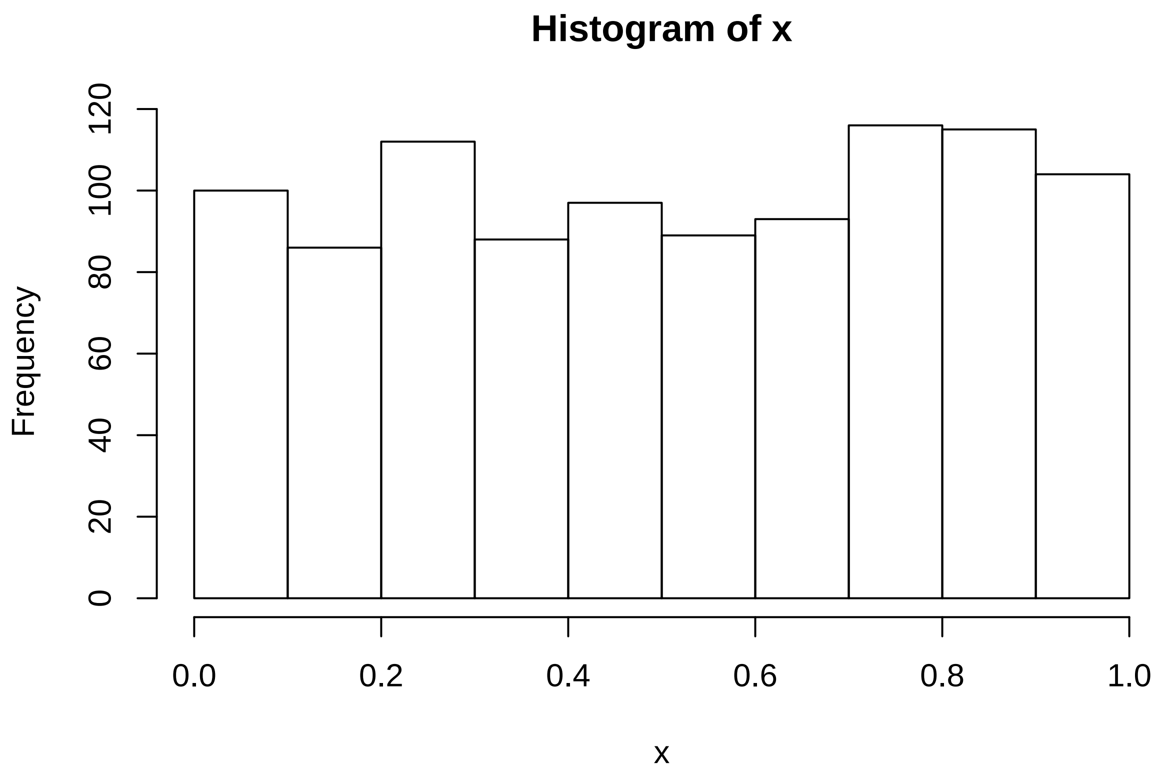
Shapes

Modes

Uniform

We can generate data that is uniform:

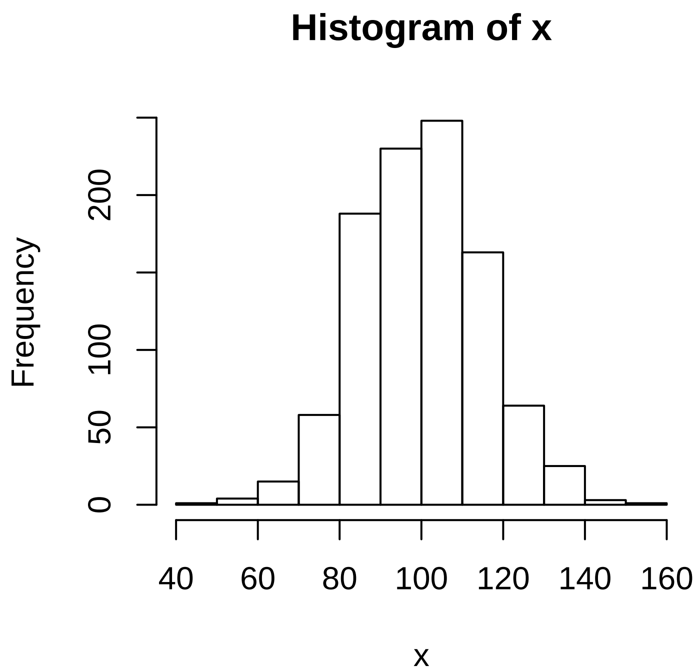
```
x = runif(1000) # Generate 1000 [R]andom [Unif]orm numbers
hist(x) # Plot a histogram
```



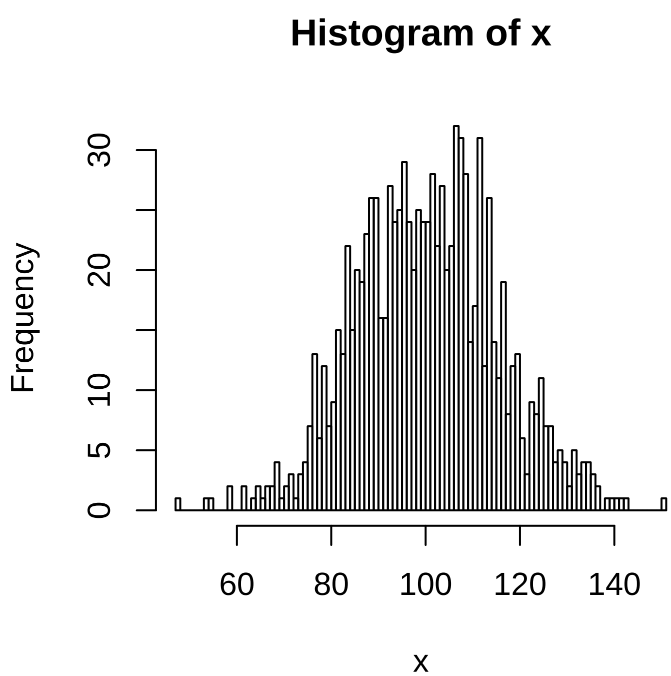
Unimodal

The most common unimodal distribution is the *Normal* distribution:

```
x = rnorm(1000,100,15) # Generate 1000 [R]andom [Norm]al numbers
# with mean 100 and SD 15 (Does that remind you of something?)
hist(x) # Default Histogram
```



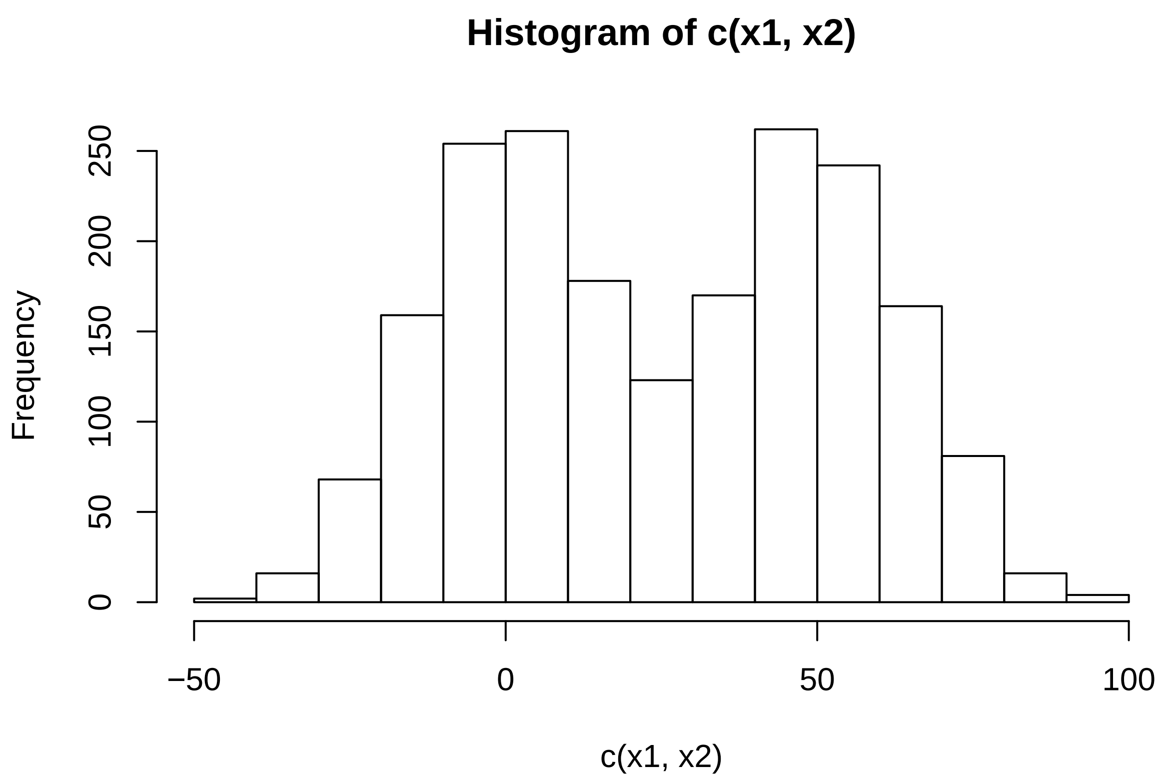

```
hist(x, breaks=100) # This time with more bins
```



Bimodal

Is generated by, for example, adding two normal distributions with different means:

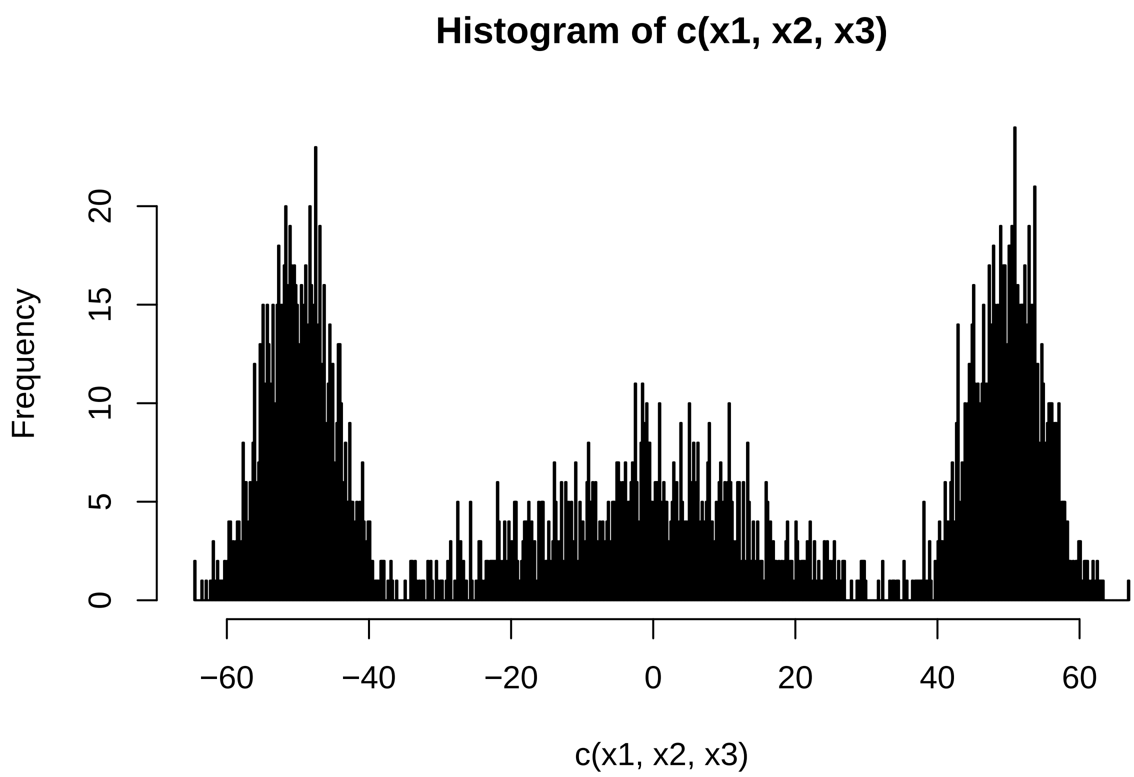
```
x1 = rnorm(1000,0,15)
x2 = rnorm(1000,50,15)
hist(c(x1,x2))
```



Multimodal

Same thing as Bimodal, just with more distributions added:

```
x1 = rnorm(1000,0,15)
x2 = rnorm(1000,50,5)
x3 = rnorm(1000,-50,5)
hist(c(x1,x2,x3),breaks = 500)
```



Numerical Description

Median

```r
median(earthquakes$mag)
```

```
## [1] 1.23
```

Spread

For any data:

- Range

```r
range(earthquakes$mag)
```

```
## [1] 0.0 7.5
```

- IQR

```r
IQR(earthquakes$mag)
```

```
## [1] 1.25
```

- 5 Numbers

```r
fivenum(earthquakes$mag)
```

```
## [1] 0.00 0.75 1.23 2.00 7.50
```

For symmetrical data

```
- Mean
```

```r
mean(earthquakes$mag)
```

```
## [1] 1.56394
```

```
- Standard Deviation
```

```r
sd(earthquakes$mag)
```

```
## [1] 1.184144
```

Calculating by hand

Mean and Standard Deviation

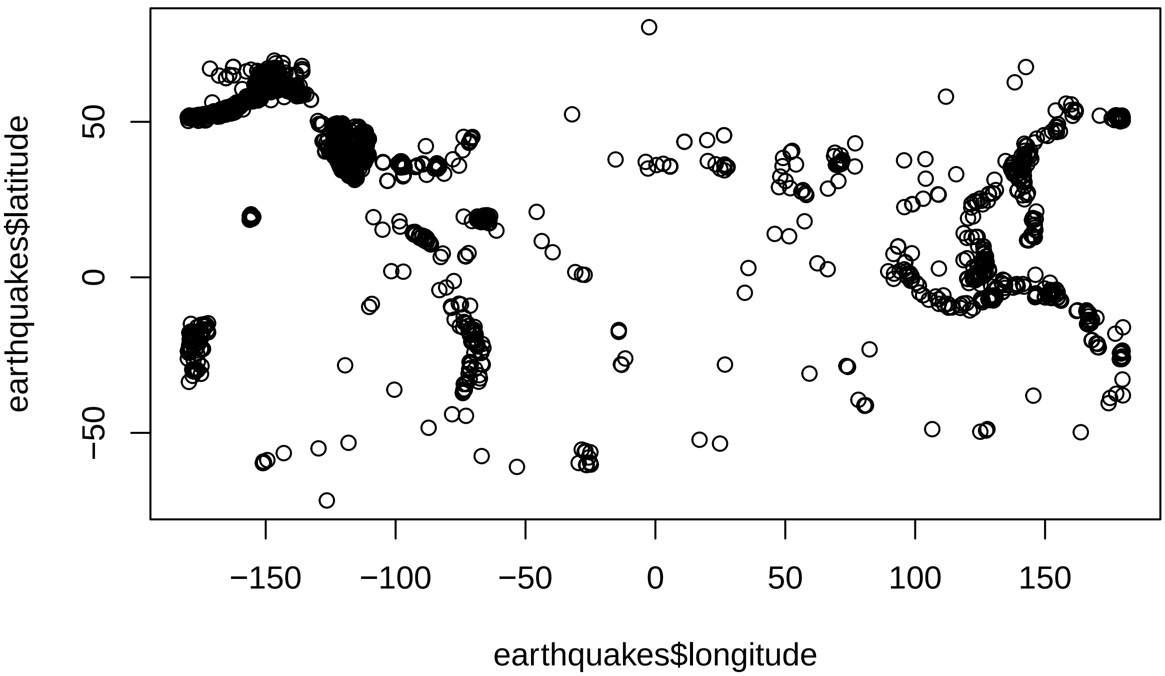
For the following sets, calculate the mean and the standard deviation:

1	2	3	4	5	6
152	83	40	58	72	128
84	76	105	81	60	74
84	78	27	35	57	88
-14	-16	-1	21	6	-8
32	45	59	20	8	4
-8	20	109	39	111	29
61	54	63	81	69	135
85	74	91	109	129	91
-4	6	-38	69	48	26
-4	-3	31	8	-8	1

For fun

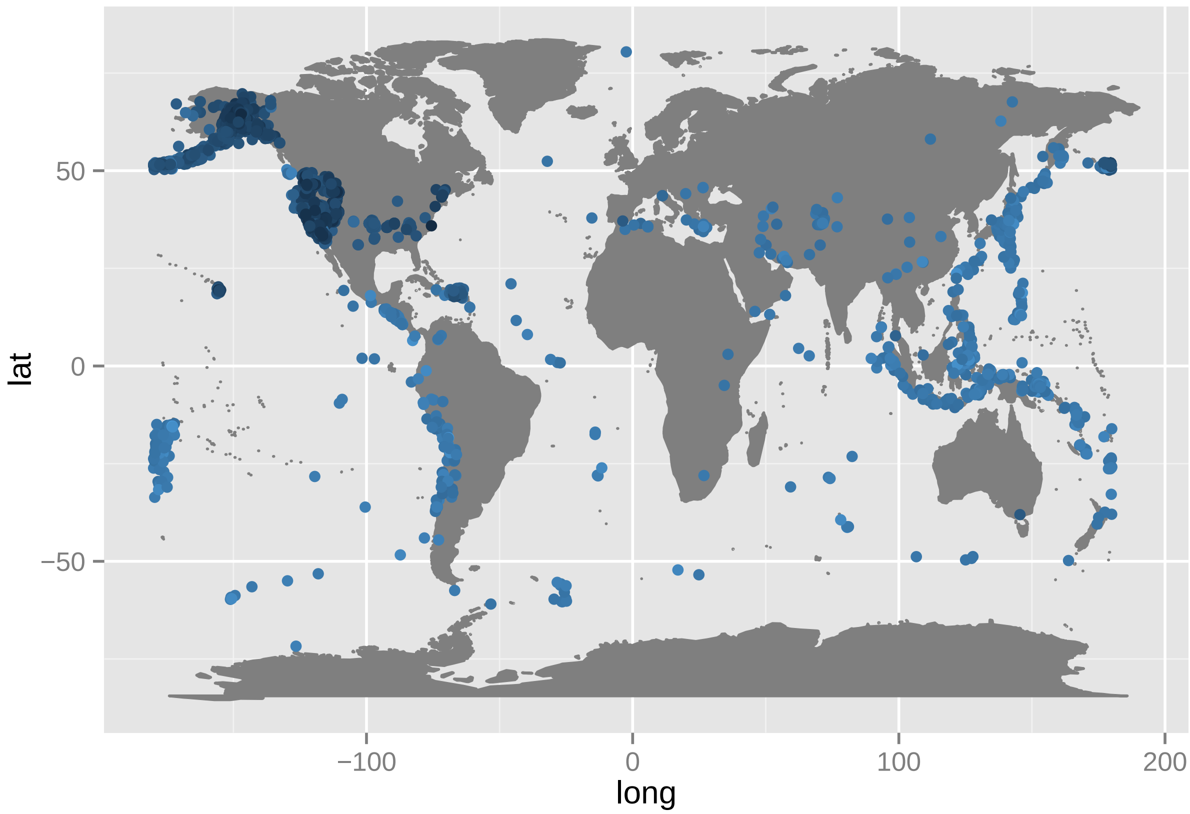
Let's do a plot of the latitude and longitude of the earthquake epicenters:

```r
plot(earthquakes$longitude, earthquakes$latitude)
```



Here is a prettier version of the same thing. Eventually, I want you to be able to use Google to arrive at solutions for similar problems!

```
library(maps)
library(maptools)
library(ggmap)
mp = NULL
mapWorld = borders("world", colour="gray50", fill="gray50")
mp = ggplot() + mapWorld
mp = mp + geom_point(aes(
  x = earthquakes$longitude,
  y = earthquakes$latitude,
  color = earthquakes$mag))
mp + scale_color_continuous(guide=FALSE)
```



Solutions

```r
pander(round(solutions,2))
```

Mean	SD
88.83	42.84
80	14.79
61.5	26.05
-2	13.93
28	21.48
50	49.05
77.17	29.75
96.5	19.55
17.83	38.29
4.17	14.22