STAT234: Lecture 1 - Basics of Data Analysis !!

Kushal K. Dey

Requisites

Main software for this course: R and RStudio

Requisites

Main software for this course: R and RStudio

Rstudio Installation:
`http://statistics.uchicago.edu/~collins/Rinstall/`

Requisites

Main software for this course: R and RStudio

Rstudio Installation:
`http://statistics.uchicago.edu/~collins/Rinstall/`

You would need to load the package *mosaic* every time you start R.

Requisites

Main software for this course: R and RStudio

Rstudio Installation:
`http://statistics.uchicago.edu/~collins/Rinstall/`

You would need to load the package *mosaic* every time you start R.

Office hour time (7 sessions) + Problem session (4 hrs) each week:
Timings to be decided later

Let's Start!

What is Data?

Let's Start!

What is Data?

Simply put, a collection of facts that can be analyzed !

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples:

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples: (Rather questions !)

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples: (Rather questions !)

- How many apps do you have on your phone?

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples: (Rather questions !)

- How many apps do you have on your phone?
- How much money did you spend on lunch?

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples: (Rather questions !)

- How many apps do you have on your phone?
- How much money did you spend on lunch?
- How What grade did you get in STAT 234?

Let's Start!

What is Data?
Simply put, a collection of facts that can be analyzed !

Collection and analysis of data is called *statistics*.

Where can I get data?
Everywhere! ....well almost!

Examples: (Rather questions !)

- How many apps do you have on your phone?
- How much money did you spend on lunch?
- How What grade did you get in STAT 234?

Backdrop

Beginning of Statistics?

Backdrop

Beginning of Statistics?

1532 $\rightarrow$ First weekly data on deaths in London (Sir W. Petty)

1539 $\rightarrow$ Data collection on marriages, baptism and death in France

1654 $\rightarrow$ Correspondence with gambling and probability (Fermat and Pascal)

Backdrop

Beginning of Statistics?
1532 $\rightarrow$ First weekly data on deaths in London (Sir W. Petty)
1539 $\rightarrow$ Data collection on marriages, baptism and death in France
1654 $\rightarrow$ Correspondence with gambling and probability (Fermat and Pascal)

Topics covered in this course will focus on the period between 1890s-1960s.

Backdrop

Beginning of Statistics?

1532 $\rightarrow$ First weekly data on deaths in London (Sir W. Petty)

1539 $\rightarrow$ Data collection on marriages, baptism and death in France

1654 $\rightarrow$ Correspondence with gambling and probability (Fermat and Pascal)

Topics covered in this course will focus on the period between 1890s-1960s.

Recently there is a lot of interest in Data Science and Big Data Analysis, which are essentially applying statistics on data of large size (many GBs or TBs). For example, Facebook, Twitter and Google generates massive data of user activity for billions of users daily!

Lets look at some Data!!

As a toy exercise, lets analyze twitter feed of.....

Lets look at some Data!!

As a toy exercise, lets analyze twitter feed of.....

Load the data

```
#library(devtools); install_github('kkdey/TrumpTwitterFeed')
library(TrumpTwitterFeed)
data("trump.data.frame")
dim(trump.data.frame)
```

```
[1] 1336    6
```

Snapshot of the Data-1

```
head(trump.data.frame[,1:5], 3)

  tweet_month tweet_year tweet_day retweets favorites
1    2016-Mar       2016        26     7625     24147
2    2016-Mar       2016        26     6412     19867
3    2016-Mar       2016        26     4773     15029

tail(trump.data.frame[,1:5], 3)

     tweet_month tweet_year tweet_day retweets favorites
1334    2015-Oct       2015        13      732      1640
1335    2015-Oct       2015        13      974      2254
1336    2015-Oct       2015        13     4578      8393
```

Snapshot of the Data- 2

```
glimpse(trump.data.frame)

Observations: 1,336
Variables: 6
$ tweet_month (fctr) 2016-Mar, 2016-Mar, 2016-Mar, 2016...
$ tweet_year  (fctr) 2016, 2016, 2016, 2016, 2016, 2016...
$ tweet_day   (fctr) 26, 26, 26, 26, 26, 26, 26, 26, 26...
$ retweets    (dbl) 7625, 6412, 4773, 7079, 5143, 5374,...
$ favorites   (dbl) 24147, 19867, 15029, 20798, 15922, ...
$ tweet_text  (fctr) Remember, I am the only candidate ...
```

```
summary(trump.data.frame)

   tweet_month  tweet_year    tweet_day       retweets    
 2015-Oct: 93   2015:434   03      : 93   Min.   :  322  
 2015-Nov:138   2016:902   28      : 89   1st Qu.: 1270  
 2015-Dec:203              12      : 75   Median : 2356  
 2016-Jan:226              24      : 70   Mean   : 3113  
 2016-Feb:324              23      : 69   3rd Qu.: 4130  
 2016-Mar:352              15      : 68   Max.   :25524  
                           (Other):872                   
   favorites    
 Min.   :  713  
 1st Qu.: 3729  
 Median : 6905  
```

Sorting number of retweets

```
sorted_retweet_counts <- sort(trump.data.frame$retweets)

tail(sorted_retweet_counts)

[1] 18287 18303 18638 19252 25323 25524

head(sorted_retweet_counts)

[1] 322 375 412 418 439 455
```

Sorting number of retweets

```
quantile(~retweets, data=trump.data.frame)
```

```
   0%   25%   50%   75%  100%
  322  1270  2356  4130 25524
```

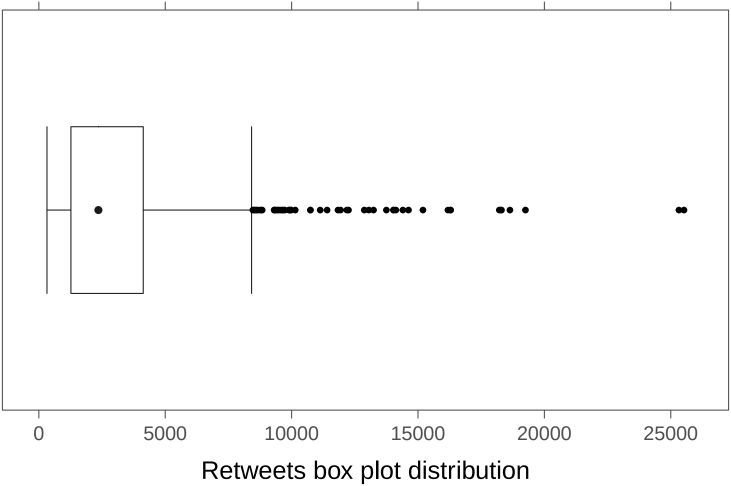
IQR or inter-quartile range is the difference between the 75 th quantile and the 25 tquantile.

```
IQR(~retweets, data=trump.data.frame)
```

```
[1] 2860
```

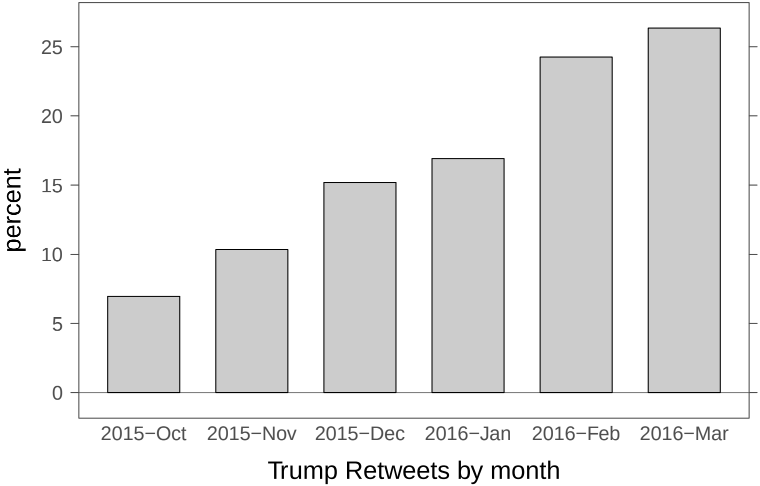
Box plot of retweets

```
bwplot(~ retweets, data=trump.data.frame,
       xlab="Retweets box plot distribution")
```



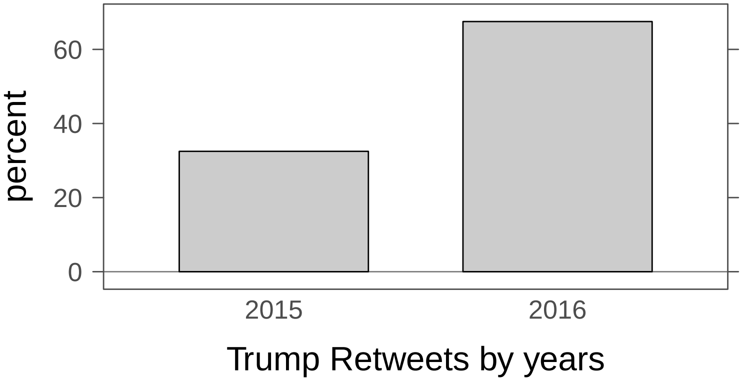
Bar graph

```
bargraph(retweets ~ tweet_month, data=trump.data.frame,
   type="percent", xlab="Trump Retweets by month", cex=0.5)
```



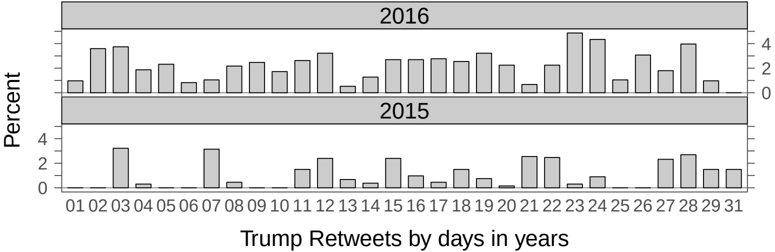
Bar graph

```
bargraph(retweets ~ tweet_year,
         data=trump.data.frame, type="percent",
          xlab="Trump Retweets by years", cex=0.5)
```



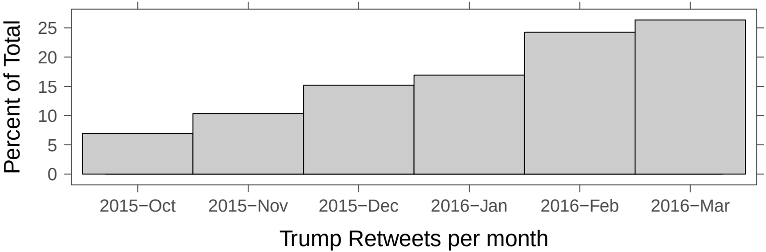
Bar graph

```
bargraph(retweets~ tweet_day | tweet_year,
         data=trump.data.frame, type="percent",
         xlab="Trump Retweets by days in years",
         ylab="Percent", layout=c(1,2))
```



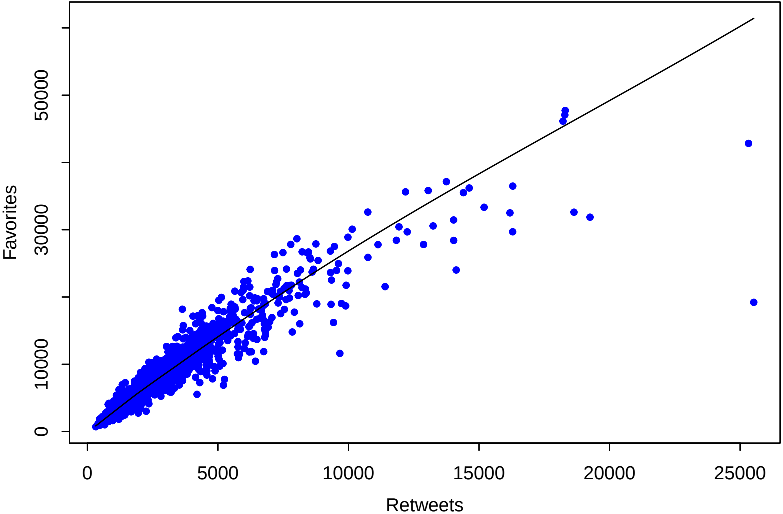
Histogram

```
histogram(retweets ~ tweet_month, data=trump.data.frame,
          type="percent", xlab="Trump Retweets per month")
```



Scatter Plot

```
scatter.smooth(trump.data.frame$retweets,
      trump.data.frame$favorites, lwd=1, pch=20,
      col="blue",xlab="Retweets", ylab="Favorites")
```



The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

Talk a moment with your neighbor. See if you can come up an equation to express this "balancing point" property of the average.

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

Talk a moment with your neighbor. See if you can come up an equation to express this "balancing point" property of the average.

Proof: Show that for **any** sample of size n, $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$

How to Prove the Math Stuff I

- A proof is a “paragraph” of mathematical “sentences”,
- written in order to make logical sense to the reader.
  ...just like you do in the Core all the time!
- It’s your personal argument as to why a claim must be true.
- Justify each step (”sentence”) using statistics
  (and using results already proven in the course).

How to Prove the Math Stuff II

OK. Our first proof is to confirm an equation.

Proof: Show that for **any** sample of size n, $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

Start on the left side: $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})$

$=$ rewrite

$=$ and rewrite

$=$ and rewrite again

$=$ until arriving at the right side $= 0$

In groups: Write down a first step.

Our First Proof!

Four common starting points.

Three are great, but one is incorrect. Which one? Why?

1. $$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$
2. $$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}\left[x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right]$$
3. $$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$
4. $$\sum_{i=1}^{n}(x_i - \overline{x}) = \left[\sum_{i=1}^{n}x_i\right] - \left[\sum_{i=1}^{n}\overline{x}\right]$$

Is "Σ" confusing you? Read Chapter 0 (Math Supplement).

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$=$$

$$=$$

$$=$$

$$=$$

Our First Proof!

Starting with the first option:

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\
&= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}} \\
&= \\
&= \\
&=
\end{aligned}
$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$=$$

$$=$$

Our First Proof!

Starting with the first option:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\ &= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}} \\ &= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x} \\ &= \\ &= \end{aligned}$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$= n\overline{x} - n\overline{x}$$

$$=$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$= n\overline{x} - n\overline{x} \quad \text{since } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Justification required!)}$$

$$=$$

Our First Proof!

Starting with the first option:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\
&= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}} \\
&= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x} \\
&= n\overline{x} - n\overline{x} \quad \text{since } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Justification required!)} \\
&= 0
\end{aligned}$$

Let's agree that $b - b = 0$ for any real number b. :)

Measuring Spread of Data Distribution I

The average devation $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$ **always** $= 0$!

Need a different measure for "typical size of deviations" (spread)

There are many measures of spread:

- mean squared deviation (*MSD* or "variance"),
- mean absolute deviation (*MAD*),
- standard deviation (*SD*) = root $MSD = RMSD = \sqrt{MSD}$,
- interquartile range (*IQR*= range of middle 50% of data)
- range,
- ...and more (not covered in this course).

Measuring Spread of Data Distribution II

Let's consider two common loss functions (measures of spread)

- The mean of absolute deviations:

$$MAD(w) = \frac{1}{n}\sum_{i=1}^{n} |x_i - w|$$

- The mean of squared deviations:

$$MSD(w) = \frac{1}{n}\sum_{i=1}^{n} (x_i - w)^2$$

What value of w should we choose using MAD? Using MSD?

It seems reasonable that w should be in the "center" of the data for each measure. But which value in the middle would be best?

One optimality criteria: Choose w that minimizes MAD or MSD.

A more objective qualification

- Let's say w is a good candidate for the center. Then in some sense $x_1 - w$, $x_2 - w$, $x_n - w$ should be small in a collective fashion.
- How about we combine these quantities?
- $$MSD(w) = \frac{1}{n}\{(x_1-w)^2+(x_2-w)^2+\cdots(x_n-w)^2\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-w)^2$$

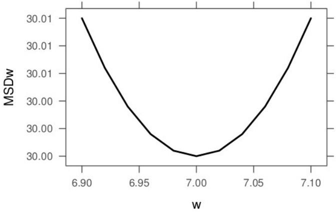
Behavior of MSD function

Let's take the very simple dataset comprising of only 4 points 1, 3, 15 and 9.

```
       w MSDw
 [1,]  1   66
 [2,]  2   55
 [3,]  3   46
 [4,]  4   39
 [5,]  5   34
 [6,]  6   31
 [7,]  7   30
 [8,]  8   31
 [9,]  9   34
[10,] 10   39
[11,] 11   46
[12,] 12   55
[13,] 13   66
[14,] 14   79
[15,] 15   94
```

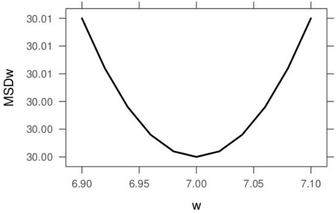
Behavior of MSD function

Let's take the very simple dataset comprising of only 4 points 1, 3, 15 and 9.



Behavior of MSD function

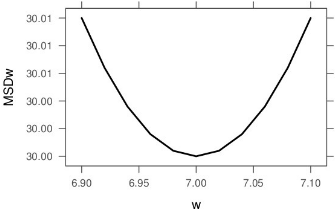
Let's take the very simple dataset comprising of only 4 points 1, 3, 15 and 9.



7 is the mean !!

Behavior of MSD function

Let's take the very simple dataset comprising of only 4 points 1, 3, 15 and 9.



7 is the mean !! Can we prove it analytically as well?

We call MSD as this function evaluated at $\bar{x}$, i.e

$$MSD = L_1(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Loss function

- MSD function:

$$L_1(w) = \frac{1}{n}\sum_{i=1}^{n}(x_i - w)^2$$

- MAD function:

$$L_2(w) = \frac{1}{n}\sum_{i=1}^{n}|x_i - w|$$

- MAD stands for Mean Absolute Deviation
- Minimize MSD function $\rightarrow$ Mean

Loss function

- MSD function:

$$L_1(w) = \frac{1}{n}\sum_{i=1}^{n}(x_i - w)^2$$

- MAD function:

$$L_2(w) = \frac{1}{n}\sum_{i=1}^{n}|x_i - w|$$

- MAD stands for Mean Absolute Deviation
- Minimize MSD function $\rightarrow$ Mean
- Minimize MAD function $\rightarrow$ Median

Loss function

- MSD function:

$$L_1(w) = \frac{1}{n}\sum_{i=1}^{n}(x_i - w)^2$$

- MAD function:

$$L_2(w) = \frac{1}{n}\sum_{i=1}^{n}|x_i - w|$$

- MAD stands for Mean Absolute Deviation
- Minimize MSD function $\rightarrow$ Mean
- Minimize MAD function $\rightarrow$ Median
- Median is another measure of central tendency

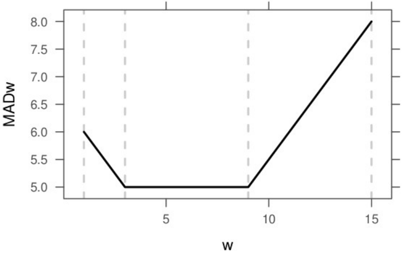
Behavior of MAD

Let's take the same dataset again 1, 3, 15 and 9.

```
       w MADw
 [1,]  1  6.0
 [2,]  2  5.5
 [3,]  3  5.0
 [4,]  4  5.0
 [5,]  5  5.0
 [6,]  6  5.0
 [7,]  7  5.0
 [8,]  8  5.0
 [9,]  9  5.0
[10,] 10  5.5
[11,] 11  6.0
[12,] 12  6.5
[13,] 13  7.0
[14,] 14  7.5
[15,] 15  8.0
```

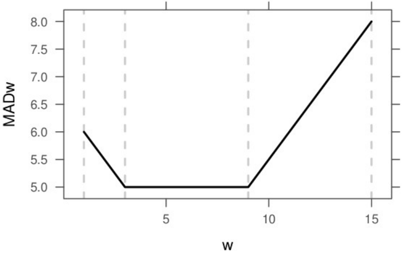
Behavior of MAD

Let's take the same dataset again 1, 3, 15 and 9.



Behavior of MAD

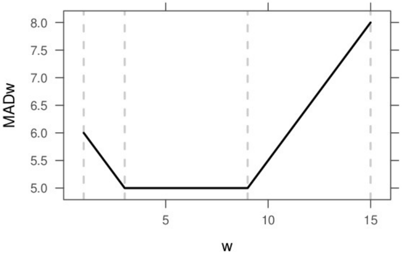
Let's take the same dataset again 1, 3, 15 and 9.



Minimum attained at any point between 3 to 9.

Behavior of MAD

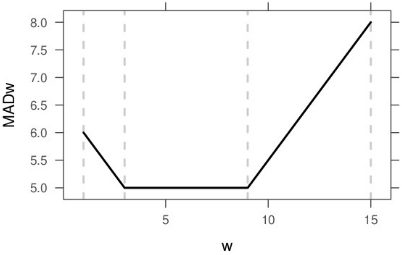
Let's take the same dataset again 1, 3, 15 and 9.



Minimum attained at any point between 3 to 9. All the points in [3,9] qualify as the median.

Behavior of MAD

Let's take the same dataset again 1, 3, 15 and 9.



Minimum attained at any point between 3 to 9. All the points in [3,9] qualify as the median. To keep it definite we will take $(3+9)/2=6$ as our median here.

Measures of Center: Median

Definition First order the data points $x_1, x_2, \cdots x_n$ in ascending order (including repetitions). Then **sample median** $\tilde{x}$ of the *sample* $x_1, x_2, \cdots x_n$ is the single middle value of the *ordered set* is n is odd and the average of two middle values if n is even.

Let's take an example.

Measures of Center: Median

Example Consider the following data set:

$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?

Measures of Center: Median

Example Consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?
  Ans. 7. Odd.

Measures of Center: Median

Example Consider the following data set:

$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?
  Ans. 7. Odd.
- What's the next step?

Measures of Center: Median

Example Consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?
  Ans. 7. Odd.
- What's the next step?
  Ans. Order them. $3.8, 5.5, 6.3, 6.8, 7.9, 8.0, 10.2$

Measures of Center: Median

Example Consider the following data set:

$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?

  Ans. 7. Odd.
- What's the next step?

  Ans. Order them. $3.8, 5.5, 6.3, 6.8, 7.9, 8.0, 10.2$
- What is the middle position? What is the median?

Measures of Center: Median

Example Consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8$

- Number of observations? Odd or even?
  Ans. 7. Odd.
- What's the next step?
  Ans. Order them. $3.8, 5.5, 6.3, 6.8, 7.9, 8.0, 10.2$
- What is the middle position? What is the median?
  Ans. Middle position is 4. Median is 6.8

Measures of Center: Median

Example Now consider the following data set:

$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8, 7.3$

- Number of observations 8 which is even

Measures of Center: Median

Example Now consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8, 7.3$

- Number of observations 8 which is even
- Order them from smallest to largest:
  $3.8, 5.5, 6.3, 6.8, 7.3, 7.9, 8.0, 10.2$

Measures of Center: Median

Example Now consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8, 7.3$

- Number of observations 8 which is even
- Order them from smallest to largest:
  $3.8, 5.5, 6.3, 6.8, 7.3, 7.9, 8.0, 10.2$
- What are the 2 middle positions? What is the median?

Measures of Center: Median

Example Now consider the following data set:
$6.3, 10.2, 3.8, 7.9, 8.0, 5.5, 6.8, 7.3$

- Number of observations 8 which is even
- Order them from smallest to largest:
  $3.8, 5.5, 6.3, 6.8, 7.3, 7.9, 8.0, 10.2$
- What are the 2 middle positions? What is the median?
  Ans. Middle positions are 4 and 5. Median is $\frac{6.8+7.3}{2} = 7.05$.

Measures of Center: Median

So we can formulate the sample median $\tilde{x}$ as:

- The $\left(\frac{n+1}{2}\right)$th ordered valued in the ordered list obtained from the sample when n is odd.
- The average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th ordered values in the ordered list when n is even.

Formulas for Sample Average, Variance, SD

$$\text{sample average } = \overline{x} = \text{ ``x-bar'' } = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{sample variance } = s^2 = \text{ ``s-squared'' } = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\text{sample standard deviation } = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$= \text{ ``typical'' distance from the average}$$

Why divide by $(n-1)$ instead of n for sample variance and SD?

Why divide by $(n-1)$ for sample variance and SD? I

Variance has a particular meaning in statistics:
mean squared distance from the average

$$MSD_n(\,\overline{x}\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Why collect data (**statistics**)?
To learn about the population (**parameters**).

$$\text{population mean} = \mu = \text{“myoo”} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{popn variance} = \sigma^2 = \text{“sigma squared”} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Why divide by $(n-1)$ for sample variance and SD? II

$$\text{truth} = \text{popn variance} = \sigma^2 = MSD_N(\,\mu\,) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

If we know the true popn mean (μ) and had a sample of n, use

$$\text{estimate} = \widehat{\sigma}^2_{\mu} = MSD_n(\,\mu\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad (1)$$

But, we almost never know μ! That's why we sample!

$$\text{realistic estimate} = \widehat{\sigma}^2_{\overline{x}} = MSD_n(\,\overline{x}\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (2)$$

The problem: (2) $\leq$ (1). Why? ...and why is this a problem?
How does dividing by $(n-1)$ for (2) help? solve the problem?

Why divide by $(n-1)$ for sample variance and SD? III

OK. So, we should divide by a number smaller than n.
But, why $(n-1)$ in particular?

Claim: Just $(n-1)$ observations and $\overline{x}$ are sufficient to determine the one remaining observation.

Proof: We know $n\overline{x} = x_1 + x_2 + \cdots + x_n$, since $\overline{x} = \frac{1}{n}\sum x_i$
So, $x_n = n\overline{x} - (x_1 + x_2 + \cdots + x_{n-1})$.

In a sense, $\sum(x_i - \overline{x})^2$ adds up $(n-1)$ "independent" values.
We say that the sum $\sum(x_i - \overline{x})^2$ has $(n-1)$ **degrees of freedom** .
So, the sample average squared deviation (variance) is defined as

$$s^2 = \text{ “s-squared” } = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Linear Transformation of Data I

Sometimes we want to analyze data in different units

Temperature unit in USA : degree Fahrenheit
Temperature unit in India: degree Celsius

Temperature: $\text{Celsius} = \frac{5}{9}(\text{Fahrenheit} - 32)$
Temperature: $\text{Celsius} = -\left(\frac{160}{9}\right) + \left(\frac{5}{9}\right)\text{Fahrenheit}$

Linear Transformation of Data II

High temperature in Chicago last 5 days of December

```
Fahrenheit <- c(39, 39, 29, 28, 31)
OLD <- options(digits=3)
```

```
Fahrenheit
```

```
[1] 39 39 29 28 31
```

```
mean(Fahrenheit)
```

```
[1] 33.2
```

Linear Transformation of Data III

```
Celsius = -(160/9) + (5/9)*Fahrenheit

rbind(Fahrenheit, Celsius)

            [,1]  [,2]  [,3]  [,4]   [,5]
Fahrenheit 39.00 39.00 29.00 28.00 31.000
Celsius     3.89  3.89 -1.67 -2.22 -0.556

mean(Celsius)

[1] 0.667

mean(Celsius)

[1] 0.667

-(160/9) + (5/9) * mean(Fahrenheit)

[1] 0.667
```

Linear Transformation of Data IV

Claim: If data $x_1, x_2, \ldots, x_n$

are linearly transformed to $y_i = a + bx_i$

Then, $\overline{y} = a + b\overline{x}$.

Proof:

A proof appears in Section 1.4 (Math Supplement).

Linear Transformation of Data V

```
sd(Celsius)
```

```
[1] 3
```

```
(5/9) * sd(Fahrenheit)
```

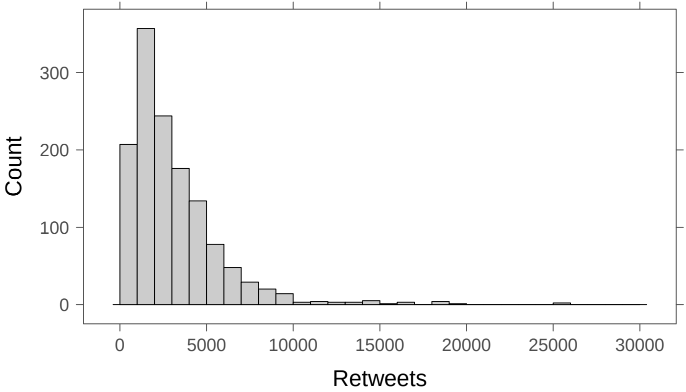
```
[1] 3
```

Claim: If data $x_1, x_2, \ldots, x_n$
are linearly transformed to $y_i = a + bx_i$

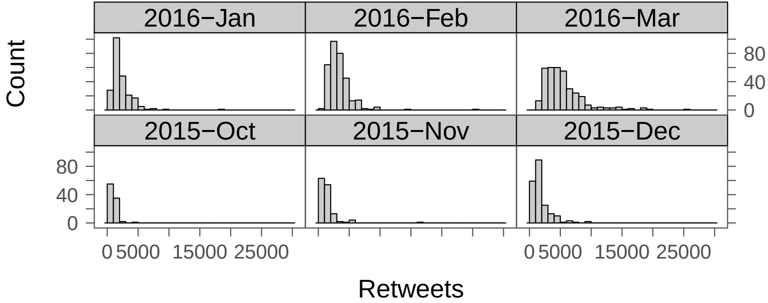
Then, $SD(y) = s_y = |b|s_x = |b|SD(x)$.

Proof: On your own for HW $\#2$.

The distribution of retweets I

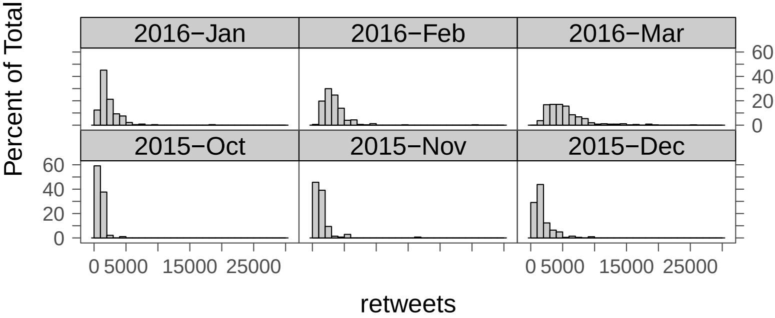


The distribution of retweets II

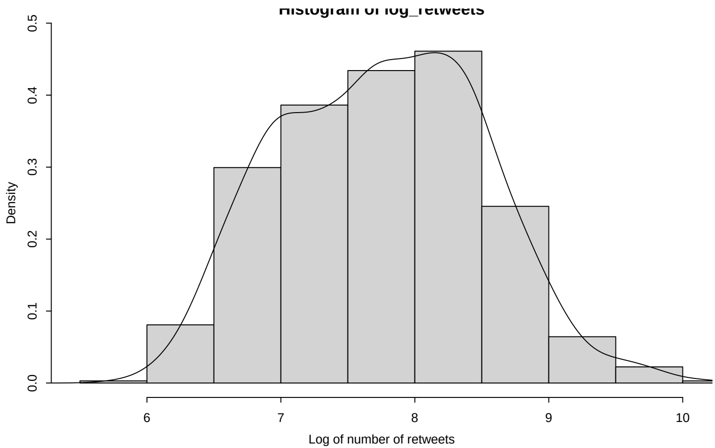


The distribution of retweets III

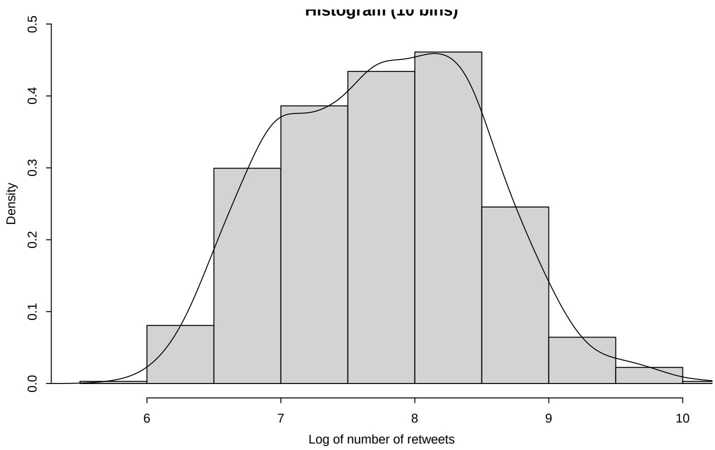
Let's make the comparison based on percentages, not counts



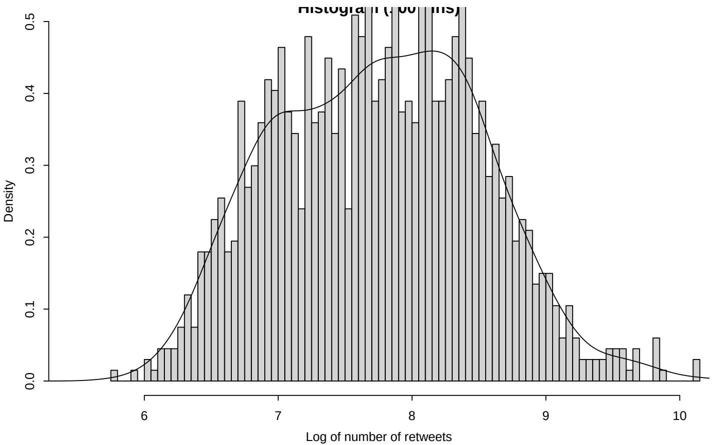
The distribution of logarithm of retweets I



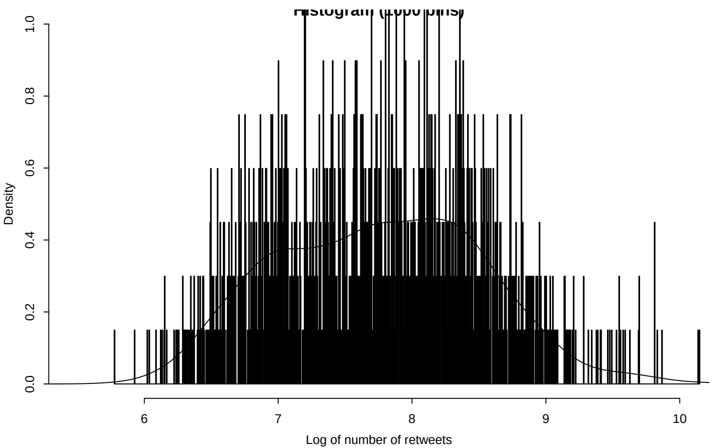
The distribution of logarithm of retweets II



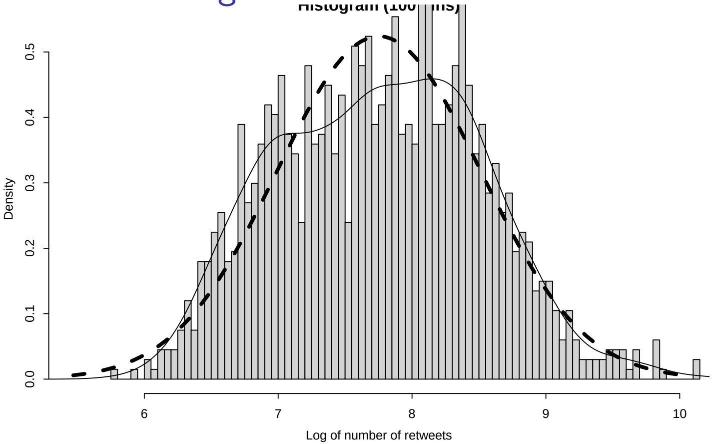
The distribution of logarithm of retweets III



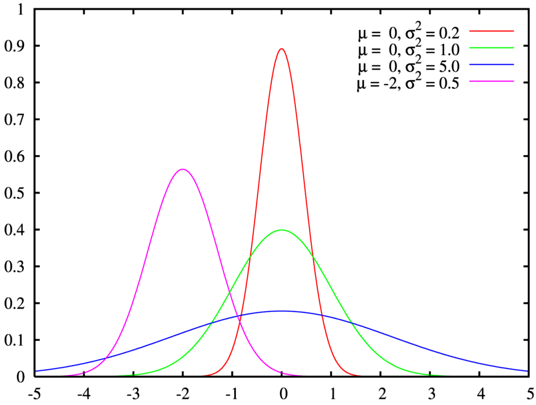
The distribution of logarithm of retweets IV



The distribution of logarithm of retweets V



Normal Distribution



Facts about Normal Distribution

- Two parameters. Mean (μ) and standard deviation (σ)

Facts about Normal Distribution

- Two parameters. Mean (μ) and standard deviation (σ)
- μ and σ are population parameters.

Facts about Normal Distribution

- Two parameters. Mean (μ) and standard deviation (σ)
- μ and σ are population parameters.
- $\mu = 0$ and $\sigma = 1$ refers to the standard normal distribution.

Facts about Normal Distribution

- Two parameters. Mean (μ) and standard deviation (σ)
- μ and σ are population parameters.
- $\mu = 0$ and $\sigma = 1$ refers to the standard normal distribution.
- Centred around μ.
- For standard normal

$$P(X > 1) = P(X < -1)$$

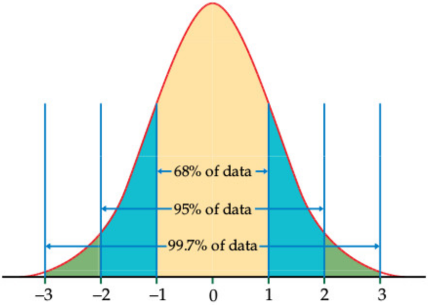
  Why?

- The formula for $N(\mu, \sigma)$ is (don't be scared - just FYI!)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Properties of Normal distribution

For a Normal Distribution (mean $\mu = 0$ and sd $\sigma = 1$)

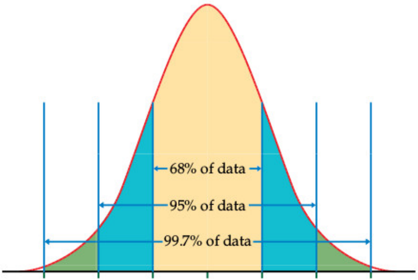


This is called the $68\% - 95\% - 99.7\%$ rule. This often simplifies our calculations.

What happens for a general μ and σ?

Properties of Normal distribution

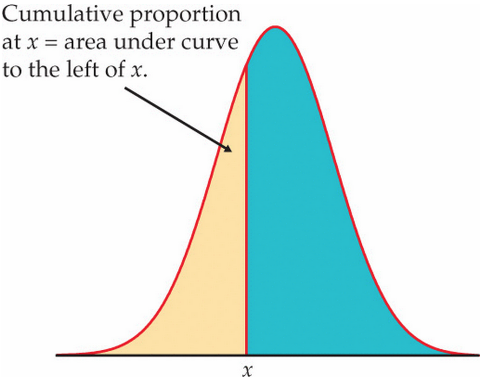
For a Normal Distribution ($\mu = 1000$ and $\sigma = 100$)



700	800	900	1000	1100	1200	1300
$\mu - 3\sigma$	$\mu - 2\sigma$	$\mu - \sigma$	μ	$\mu + \sigma$	$\mu + 2\sigma$	$\mu + 3\sigma$

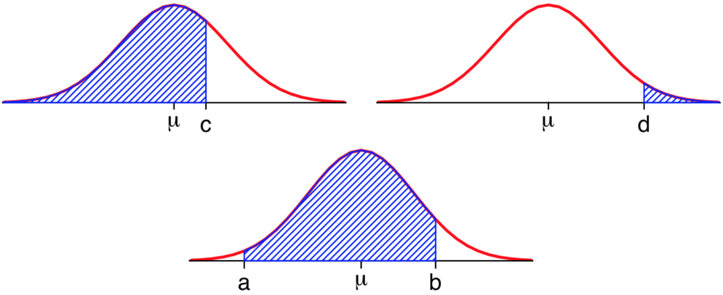
Cumulative Proportions and Standard Normal Table

- **Cumulative proportions**: The proportion of observations in a distribution that lie at or below a given value x.
- For $N(0, 1)$, use the standard normal table (Table A in textbook) to calculate cumulative proportions for x.



Cumulative Proportions and Standard Normal Table

For $N(0,1)$, we can calculate all of the following areas using cumulative proportions:



Normal Distribution

- Many sets of data follow normal distribution.

Normal Distribution

- Many sets of data follow normal distribution.
- It is a symmetric, bell-curved distribution and has many nice properties !

Normal Distribution

- Many sets of data follow normal distribution.
- It is a symmetric, bell-curved distribution and has many nice properties !
- A lot of statistical methodologies are based on the properties of this distribution.

Normal Distribution

- Many sets of data follow normal distribution.
- It is a symmetric, bell-curved distribution and has many nice properties !
- A lot of statistical methodologies are based on the properties of this distribution.
- Naturally, there is a tendency to use it everywhere we can even though the data is not actually coming from a normal distribution.

Normal Distribution

- Many sets of data follow normal distribution.
- It is a symmetric, bell-curved distribution and has many nice properties !
- A lot of statistical methodologies are based on the properties of this distribution.
- Naturally, there is a tendency to use it everywhere we can even though the data is not actually coming from a normal distribution.
- We should start by comparing the data with normal distribution.

Normal Distribution

- Many sets of data follow normal distribution.
- It is a symmetric, bell-curved distribution and has many nice properties !
- A lot of statistical methodologies are based on the properties of this distribution.
- Naturally, there is a tendency to use it everywhere we can even though the data is not actually coming from a normal distribution.
- We should start by comparing the data with normal distribution.

Standardization and z-score

- If x is an observation from a distribution with mean (Population mean) μ and standard deviation σ then , the standardized value is
$$z = \frac{x - \mu}{\sigma}$$

Standardization and z-score

- If x is an observation from a distribution with mean (Population mean) μ and standard deviation σ then , the standardized value is

$$z = \frac{x - \mu}{\sigma}$$

- This is also called the $z - score$.

Standardization and z-score

- If x is an observation from a distribution with mean (Population mean) μ and standard deviation σ then , the standardized value is
$$z = \frac{x - \mu}{\sigma}$$
- This is also called the $z - score$.
- Example?

Standardization and z-score

- If x is an observation from a distribution with mean (Population mean) μ and standard deviation σ then , the standardized value is

$$z = \frac{x - \mu}{\sigma}$$

- This is also called the $z - score$.
- Example?Assume that, time spent on the calls (The data from Lecture 1) follows approximately a normal distribution with mean $\mu = 1000$ and $\sigma = 100$.

Standardization and z-score

- If x is an observation from a distribution with mean (Population mean) μ and standard deviation σ then , the standardized value is

$$z = \frac{x - \mu}{\sigma}$$

- This is also called the $z - score$.
- Example?Assume that, time spent on the calls (The data from Lecture 1) follows approximately a normal distribution with mean $\mu = 1000$ and $\sigma = 100$.
- Suppose on particular observation is 870,

$$z - \text{score} = \frac{670 - 1000}{100} = -1.3$$

- We will use this z-score later to make some conclusions.

The normal density model in R I

The "normal density" model.

A good fit for these data distributions?

A numerical look

What percent of area under standard normal density is above/below 1?

The normal density model in R II

```
pnorm(-1, m=0, s=1)

[1] 0.1587

pnorm(-1)

[1] 0.1587

1 - pnorm(1)

[1] 0.1587
```

The normal density model in R III

What percent of area under *any* normal density is above/below 1 sd from mean?

```
mtweets <- mean(log_retweets)
mtweets
```

```
[1] 7.753
```

```
stweets <- sd(log_retweets)
stweets
```

```
[1] 0.7613
```

```
1 - pnorm(mtweets + stweets, m=mtweets, s=stweets)
```

```
[1] 0.1587
```

```
pnorm(mtweets - stweets, m=mtweets, s=stweets)
```

```
[1] 0.1587
```

The normal density model in R IV

What percent of the observed data are right/left of 1 sd from mean?

```
sum(log_retweets >= mtweets + stweets, na.rm=TRUE) / n
```

```
[1] 0.1639
```

```
sum(log_retweets <= mtweets - stweets, na.rm=TRUE) / n
```

```
[1] 0.1886
```

The normal density model in R V

What percent of the model/data are 2 sd to the right mean?

```
1 - pnorm(2)
```

```
[1] 0.02275
```

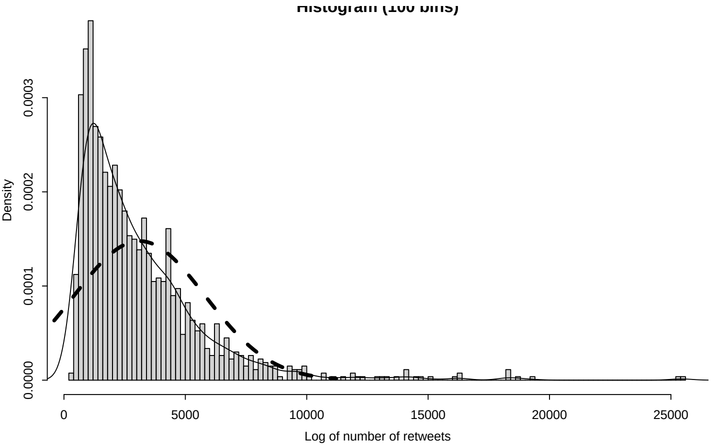
```
sum(log_retweets >= mtweets + 2*stweets, na.rm=TRUE) / n
```

```
[1] 0.02096
```

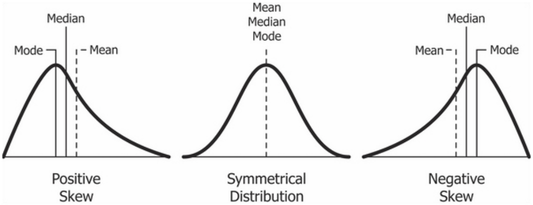
What percent of the model/data are 2 sd to the left of mean?

```
pnorm(-2)
sum(log_retweets <= mtweets - 2*stweets, na.rm=TRUE) / n
```

When things are not normal!!



Skewness


Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

A Step back: Quantiles

- The k th quantile of a set of values divides them so that $100 * k$ of the values lie below and $100 * (1 - k)$ of the values lie above.

A Step back: Quantiles

- The k th quantile of a set of values divides them so that $100 * k$ of the values lie below and $100 * (1 - k)$ of the values lie above.
- the 0.25th quantile is known first/lower quartile (Q_1)

A Step back: Quantiles

- The k th quantile of a set of values divides them so that $100 * k$ of the values lie below and $100 * (1 - k)$ of the values lie above.

- the 0.25th quantile is known first/lower quartile (Q_1)

- the 0.50th quantile is known as median (Q_2)

A Step back: Quantiles

- The k th quantile of a set of values divides them so that $100 * k$ of the values lie below and $100 * (1 - k)$ of the values lie above.
- the 0.25th quantile is known first/lower quartile (Q_1)
- the 0.50th quantile is known as median (Q_2)
- the 0.75th quantile is known as third/upper quartile (Q_3)

A Step back: Quantiles

- The k th quantile of a set of values divides them so that $100 * k$ of the values lie below and $100 * (1 - k)$ of the values lie above.
- the 0.25th quantile is known first/lower quartile (Q_1)
- the 0.50th quantile is known as median (Q_2)
- the 0.75th quantile is known as third/upper quartile (Q_3)

A Step back: Quantiles

Lets look at the quantiles of a set of values - $3.4, 2.3, 6.7, 2.1, 5.0$

A Step back: Quantiles

Lets look at the quantiles of a set of values - $3.4, 2.3, 6.7, 2.1, 5.0$

First sort the values in order

$$2.1, \quad 2.3 \quad 3.4 \quad 5.0 \quad 6.7$$

Then the quantiles for this data is given as follows

Sample fraction	0	0.25	0.50	0.75	1
Quantiles	2.1	2.3	3.4	5.0	6.7

Quantiles in R

```
  0%  25%  50%  75% 100%
 2.1  2.3  3.4  5.0  6.7
```

Consider the Trump twitter feed data

```
favstats(retweets | tweet_year, data=trump.data.frame)
```

```
  tweet_year min     Q1 median   Q3   max mean   sd   n
1       2015 322  818.2   1108 1674 16289 1527 1397 434
2       2016 581 2065.2   3239 4753 25524 3876 2841 902
  missing
1       0
2       0
```

IQR, Boxplot and outliers

- Another measure is $IQR = Q_3 - Q_1$

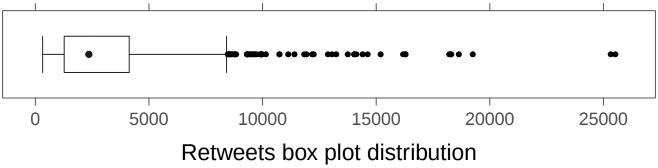
IQR, Boxplot and outliers

- Another measure is $IQR = Q_3 - Q_1$
- $1.5\times$ IQR rule : If an observation falls more than $1.5 \times IQR$ above the third quartile or below the first quartile, call it a *suspected* outlier (Caution: Not always!!).

IQR, Boxplot and outliers

- Another measure is $IQR = Q_3 - Q_1$
- $1.5\times$ IQR rule : If an observation falls more than $1.5 \times IQR$ above the third quartile or below the first quartile, call it a *suspected* outlier (Caution: Not always!!).

```
bwplot(  retweets, data=trump.data.frame, xlab="Retweets box
plot distribution")
```



Normal Quantile Plot

How can I tell whether my data is sufficiently close to normal?

- Given data $x = (x_1, x_2, \cdots, x_n)$, compute

$$y_i = \frac{x_i - \bar{x}}{s}$$

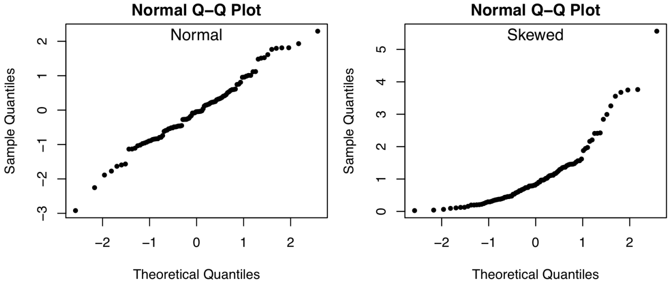
where $\bar{x} = mean(x)$ and $s = sd(x)$.

- Arrange the y data in increasing order:

$$y_{[1]} \leq y_{[2]} \leq ... \leq y_{[n]}$$

- Find the z-scores for all the percentiles $(\frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n})$ from standard normal table
- Plot $y_{[i]}$ values on the vertical axis against z-scores on the horizontal axis from $i = 1, ..., n$.

How Normal Quantile plot looks!

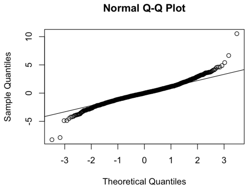


If the data are approximately normal,the Q-Q plot will be **close to a straight line**.

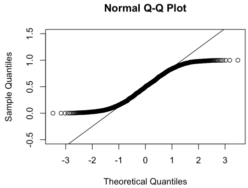
Normal Quantile Plot

Systematic deviations from a straight line indicate a non-normal distribution

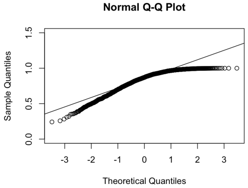
Heavy tails at both end



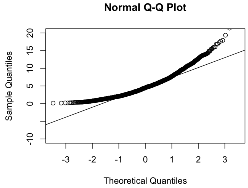
Light tails at both end



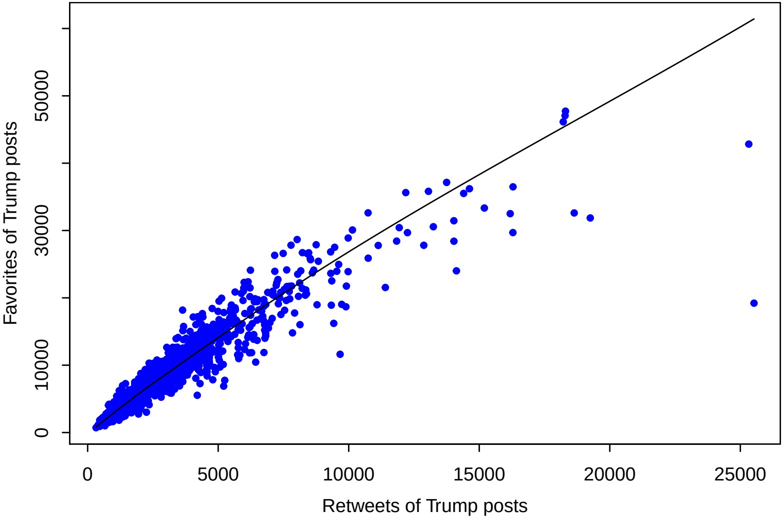
Left skewed



Right skewed



Association between Variables



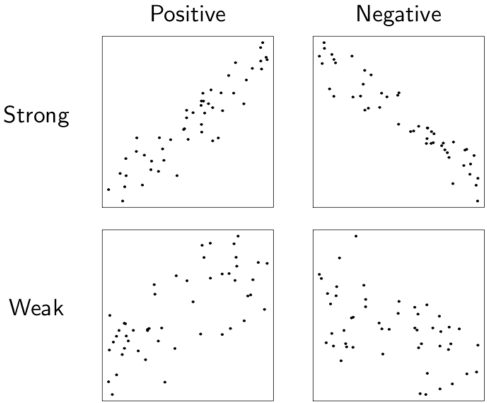
Describing Scatterplots

You can describe patterns in a scatterplot in three aspects:

- Form: Linear, curved...
- Direction: Positive, negative
- Strength: Strong, weak

Scatterplot

Examples of linear relationships:



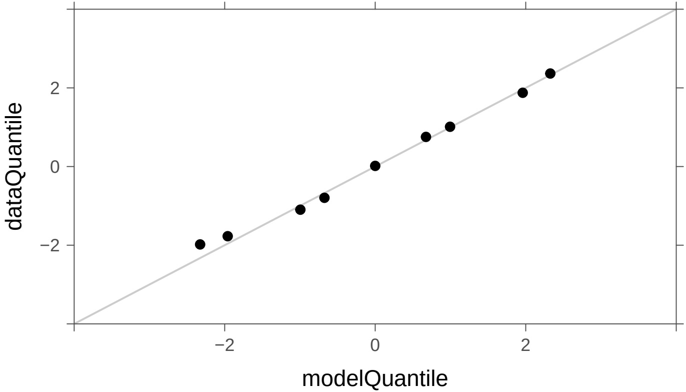
Applying quantile plots to log Retweets data I

```
mtweets <- mean(log(trump.data.frame$retweets+1))

stweets <- sd(log(trump.data.frame$retweets+1))

stdtweets = (log(trump.data.frame$retweets+1) - mtweets) / stwee

head(sort(stdtweets),3)

[1] -2.594 -2.395 -2.271

tail(sort(stdtweets),3)

[1] 2.775 3.135 3.145
```

Applying quantile plots to log Retweets data II

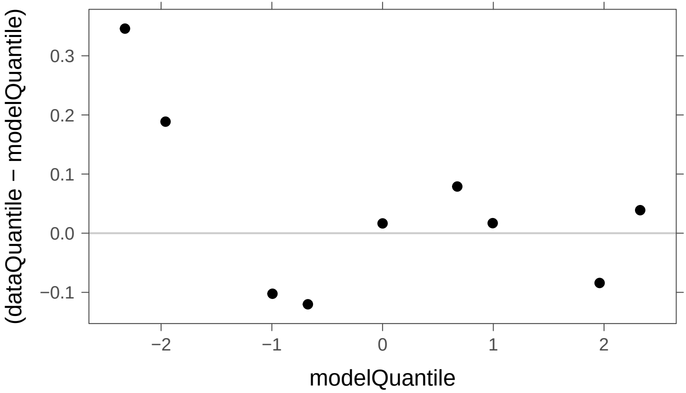
```
p = c(0.01, 0.025, 0.16, 0.25, 0.50, 0.75, 0.84, 0.975, 0.99)
modelQuantile = qnorm(p)
modelQuantile

[1] -2.3263 -1.9600 -0.9945 -0.6745  0.0000  0.6745  0.9945
[8]  1.9600  2.3263

dataQuantile = quantile(stdtweets, p, na.rm=TRUE)
dataQuantile

      1%     2.5%      16%      25%      50%      75%
-1.98042 -1.77134 -1.09683 -0.79466  0.01657  0.75343
     84%    97.5%      99%
 1.01146  1.87581  2.36527

rbind(dataQuantile, modelQuantile)

                  1%   2.5%     16%     25%     50%    75%
dataQuantile  -1.980 -1.771 -1.0968 -0.7947 0.01657 0.7534
modelQuantile -2.326 -1.960 -0.9945 -0.6745 0.00000 0.6745
                 84% 97.5%   99%
dataQuantile  1.0115 1.876 2.365
modelQuantile 0.9945 1.960 2.326
```

Quantile plots I



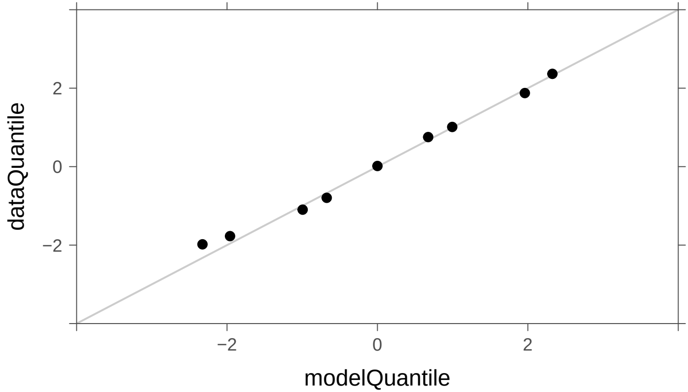
Quantile plots II

I wish the "normal probability plot" was actually plotted like this (much easier to read)

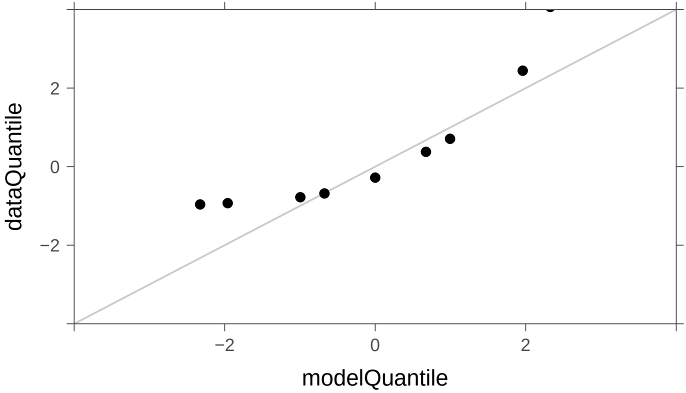


Quantile plots III

But here is the style of plot traditionally called the "normal probability plot" or "normal quantile plot"



Quantile plot for Retweets Data I



Questions?