Data Analytics

Yong Zheng
Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Schedule

- Quick Reviews
- Numerical Data
  - Descriptive Statistics
  - Probability Distribution
- Intro: R

Schedule

- Quick Reviews
  - Statistical Applications
  - Data: Population and Sample
  - Data Types
  - Descriptive Statistics
    - For nominal variables
      - By metrics
      - By visualizations (note: be able to interpret plots)
    - For numerical variables
      - By metrics

Schedule

- Quick Reviews
- Numerical Data
  - Descriptive Statistics
  - Probability Distribution
- Intro: R

Describe Quantitative Data

- Describe quantitative data Numerically
  - By range, min, max, mean, median, mode
  - By variance, standard deviation
  - By q1, q2, q3
- Describe quantitative data by visualizations
  - By stem-and-leaf
  - By histogram
  - By box plot
  - By probability distribution

Describe Quantitative Data

- Describe quantitative data by visualizations
  - ~~By stem-and-leaf [Optional]~~
  - By histogram
  - By box plot
  - By probability distribution

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By histogram

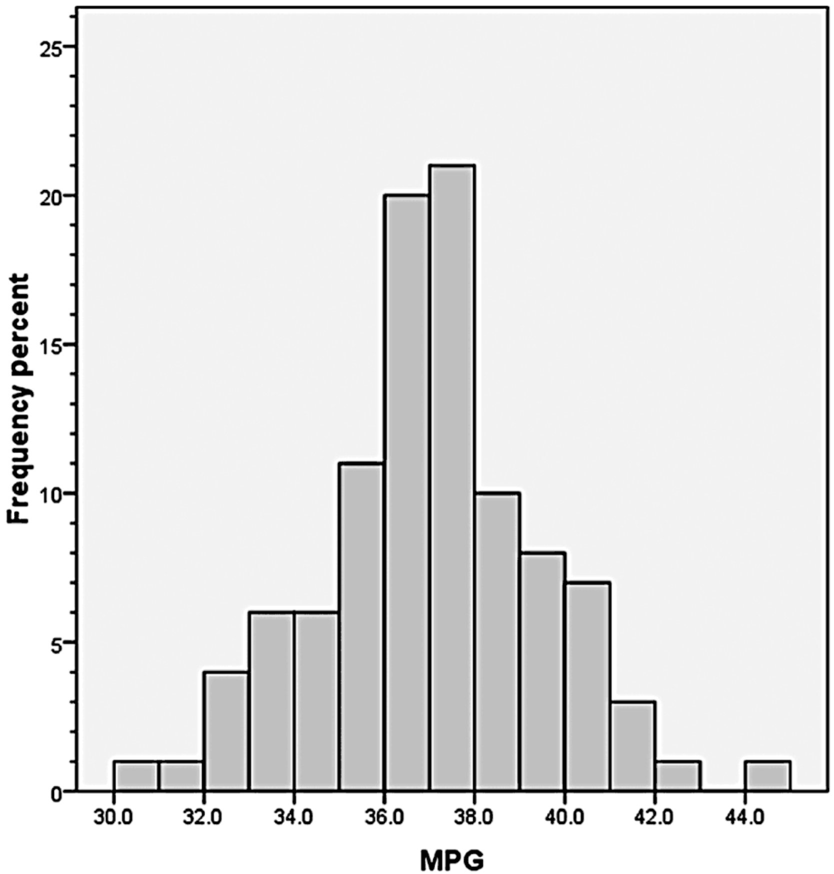


Table 2.2 EPA Mileage Ratings on 100 Cars

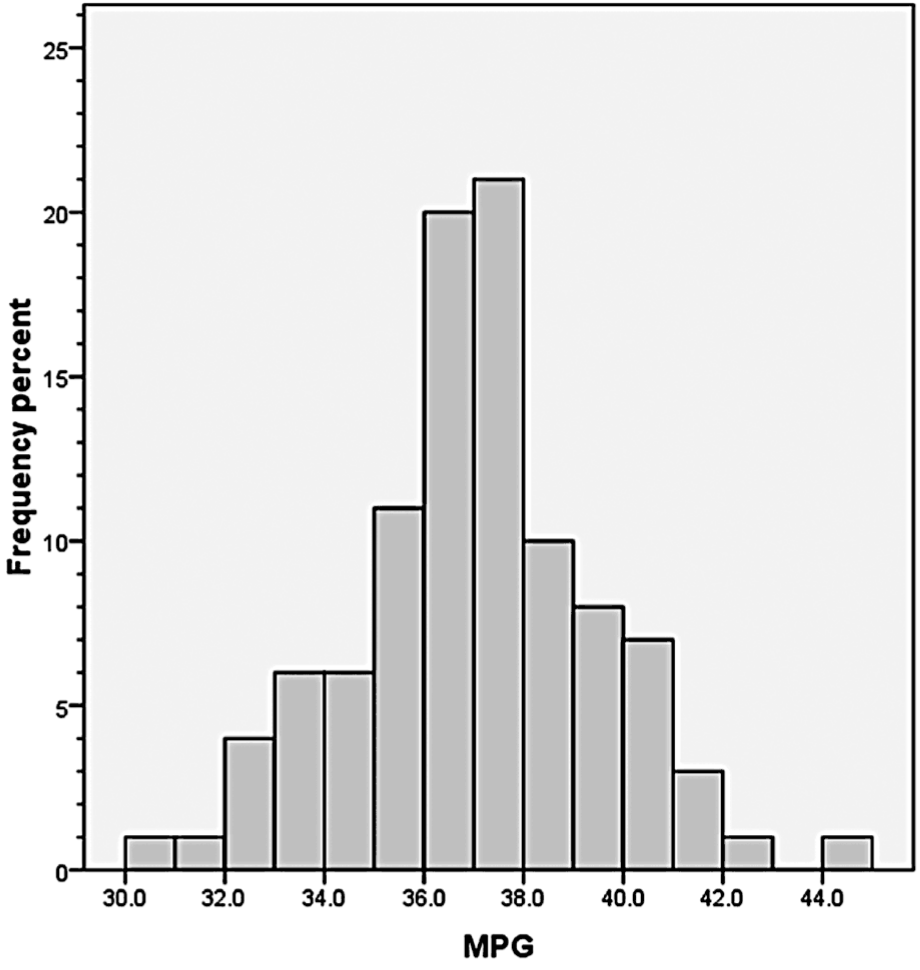
36.3	41.0	36.9	37.1	44.9
32.7	37.3	41.2	36.6	32.9
40.5	36.5	37.6	33.9	40.2
36.2	37.9	36.0	37.9	35.9
38.5	39.0	35.5	34.8	38.6
36.3	36.8	32.5	36.4	40.5
41.0	31.8	37.3	33.1	37.0
37.0	37.2	40.7	37.4	37.1
37.1	40.3	36.7	37.0	33.9
39.9	36.9	32.9	33.8	39.8
36.8	30.0	37.2	42.1	36.7
36.5	33.2	37.4	37.5	33.6
36.4	37.7	37.7	40.0	34.2
38.2	38.3	35.7	35.6	35.1
39.4	35.3	34.4	38.8	39.7
36.6	36.1	38.2	38.4	39.3
37.6	37.0	38.7	39.0	35.8
37.8	35.9	35.6	36.7	34.5
40.1	38.0	35.2	34.8	39.5
34.0	36.8	35.0	38.1	36.9

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By histogram

    It is similar to the bar graph used to describe categorical data.

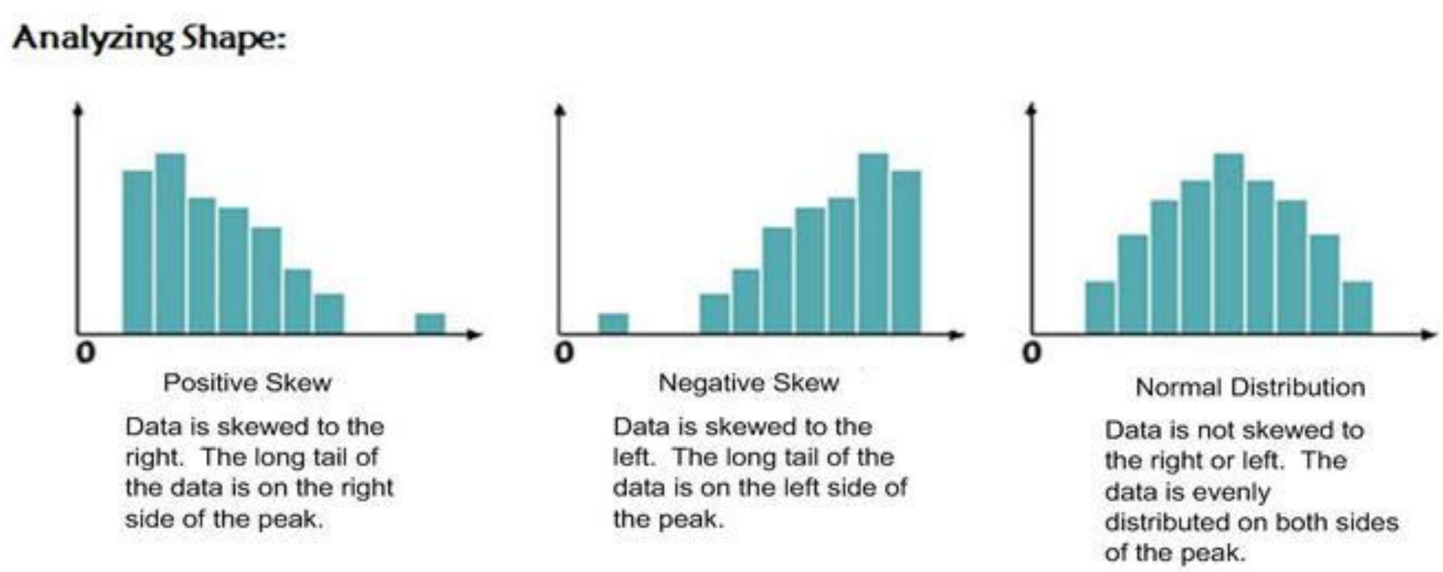
    Here, we present class frequency for a range of values, e.g., [30, 32]



Describe Quantitative Data

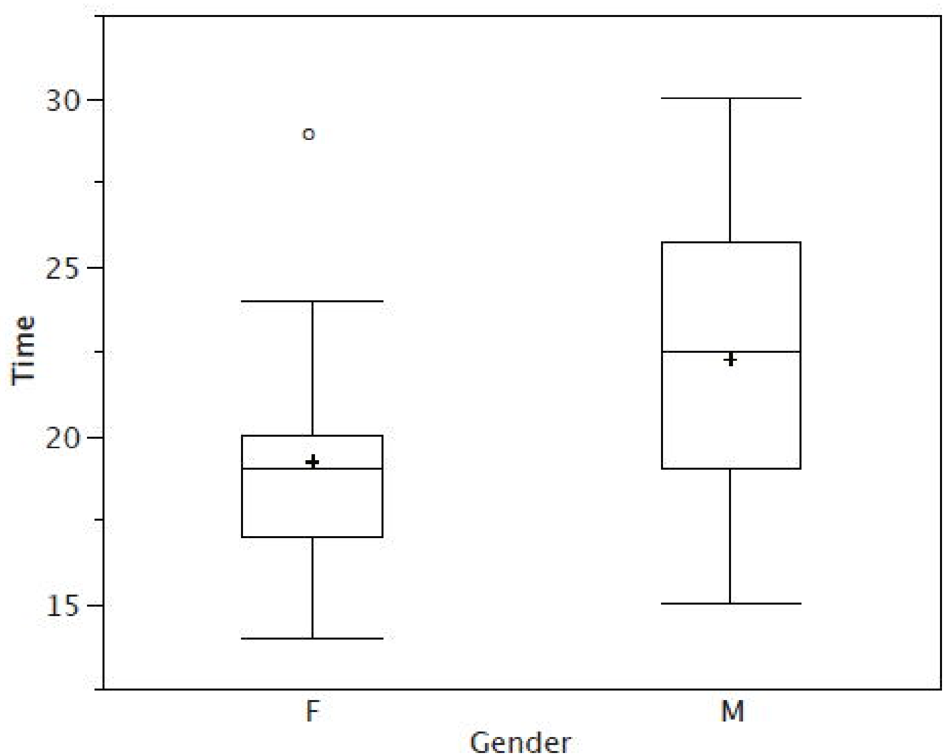
- Describe quantitative data by visualizations
  - By histogram

  How to interpret histogram? (skewness and outlier)



Describe Quantitative Data

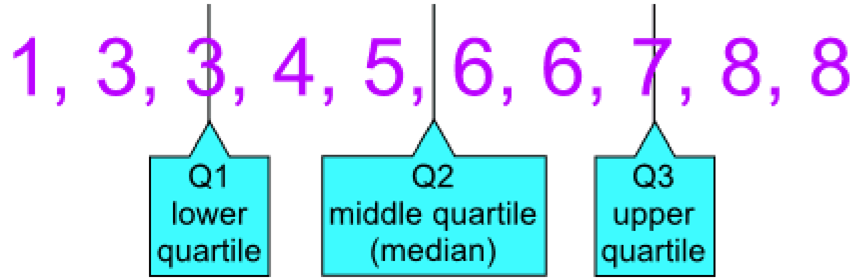
- Describe quantitative data by visualizations
  - By box plot

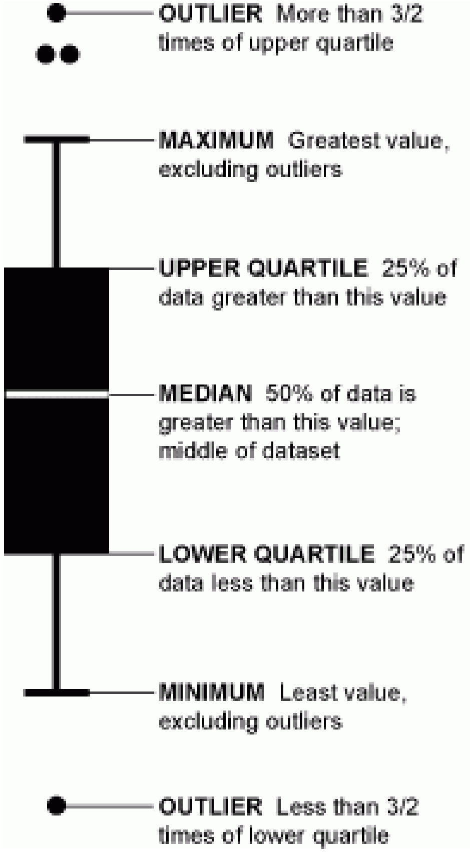


Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

1). Quartile

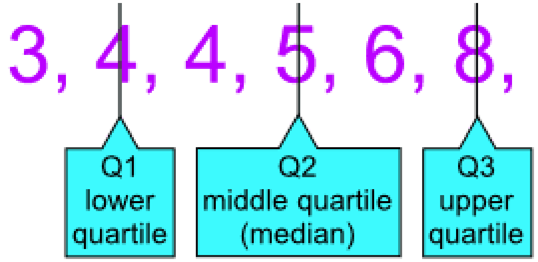


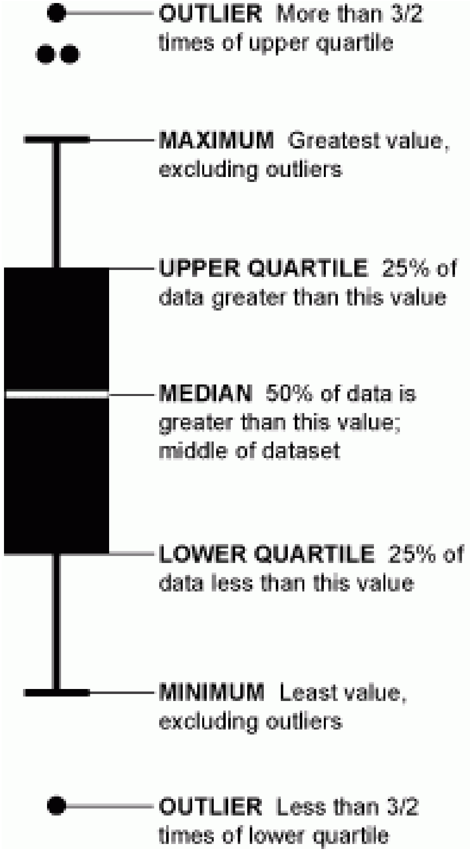


Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

1). Quartile





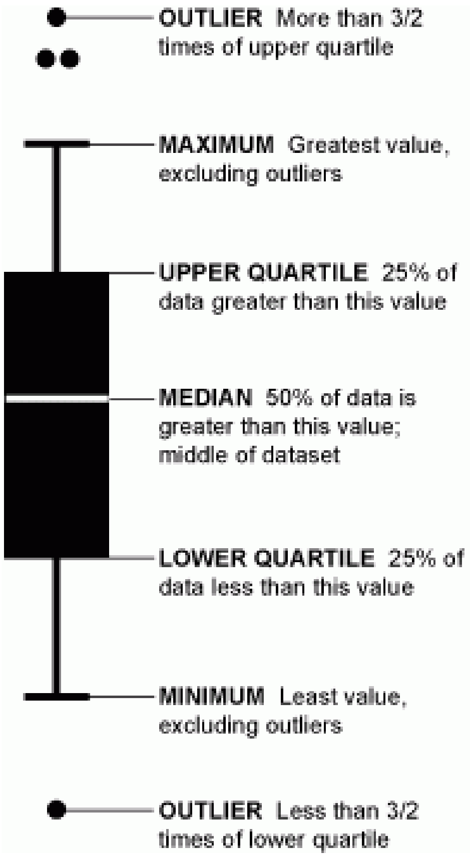
Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

2). Median

Median = 2nd quartile = q2

Note: we usually use either mean or median to represent a set of quantitative data



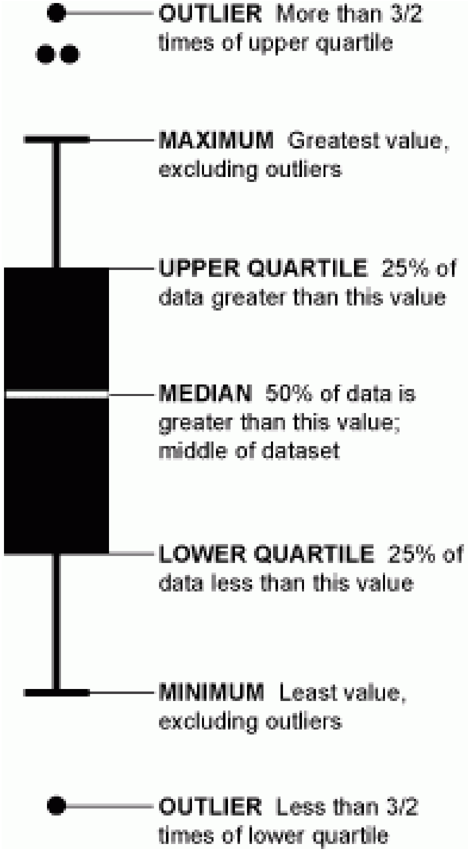
Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

3). Min, Max, Outlier

Here, the min and max values are the ones without considering outliers.

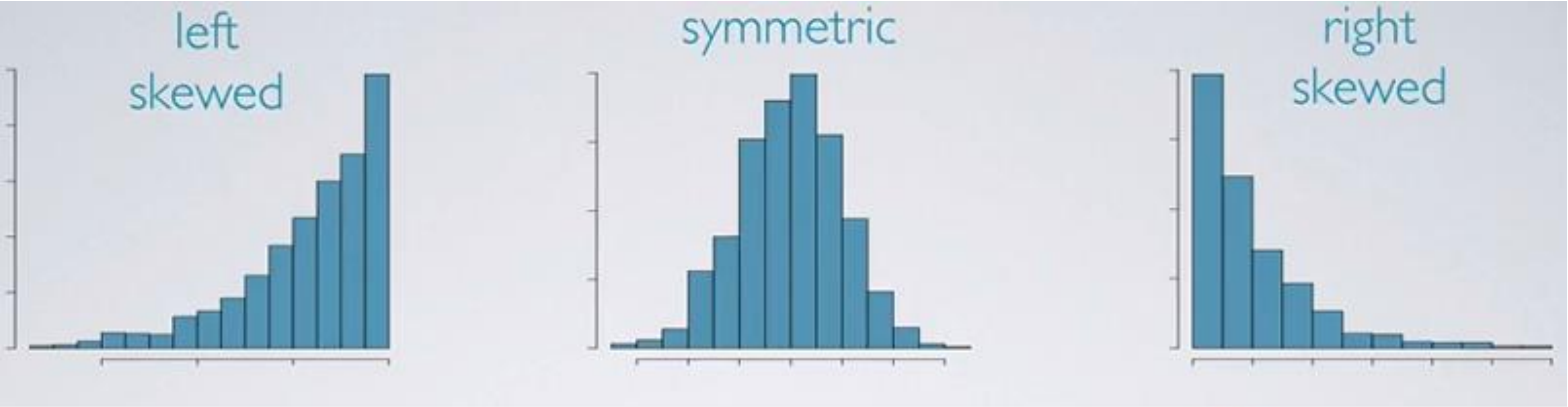
So, range ≠ Max-Min <u>from the box plot</u>!!!!!!!



Describe Quantitative Data

- Describe quantitative data by visualizations
  - By box plot: Interpretations

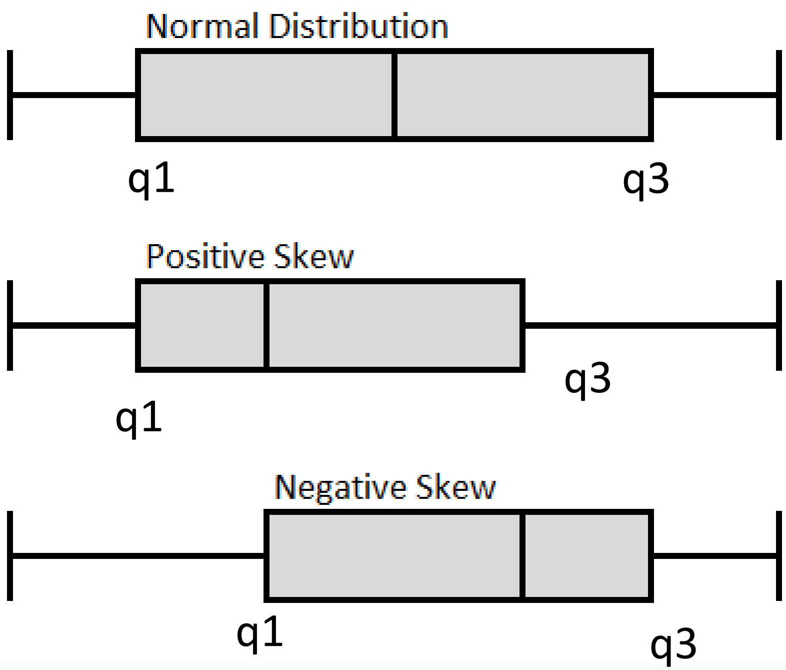
4). Skewness



Describe Quantitative Data

- Describe quantitative data by visualizations

  How to make a decision about skewness from the box plot? We focus on the median and box only



Median is exactly in the middle

Median is closer to the q1

Median is closer to q3

Describe Quantitative Data

- Describe quantitative data by visualizations
  - By probability distribution

Schedule

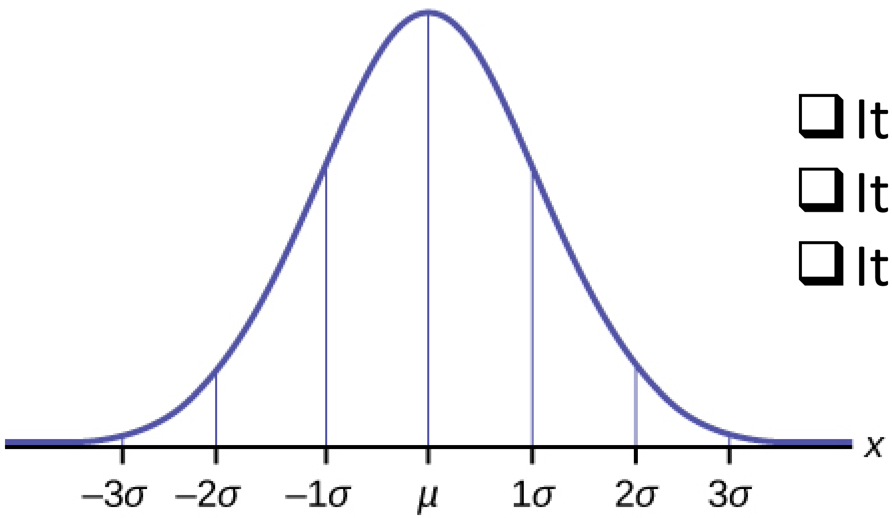
- Quick Reviews
- Numerical Data
  - Descriptive Statistics
  - Probability Distribution
- Intro: R

Week 2 - 3

- Probability Distributions
- Sampling Distributions
- Central Limit Theorem

Probability Distribution

- In general, probability distribution refers to the mathematical way to model the relative frequency distribution for a quantitative variable.
- For example: Normal Probability Distribution



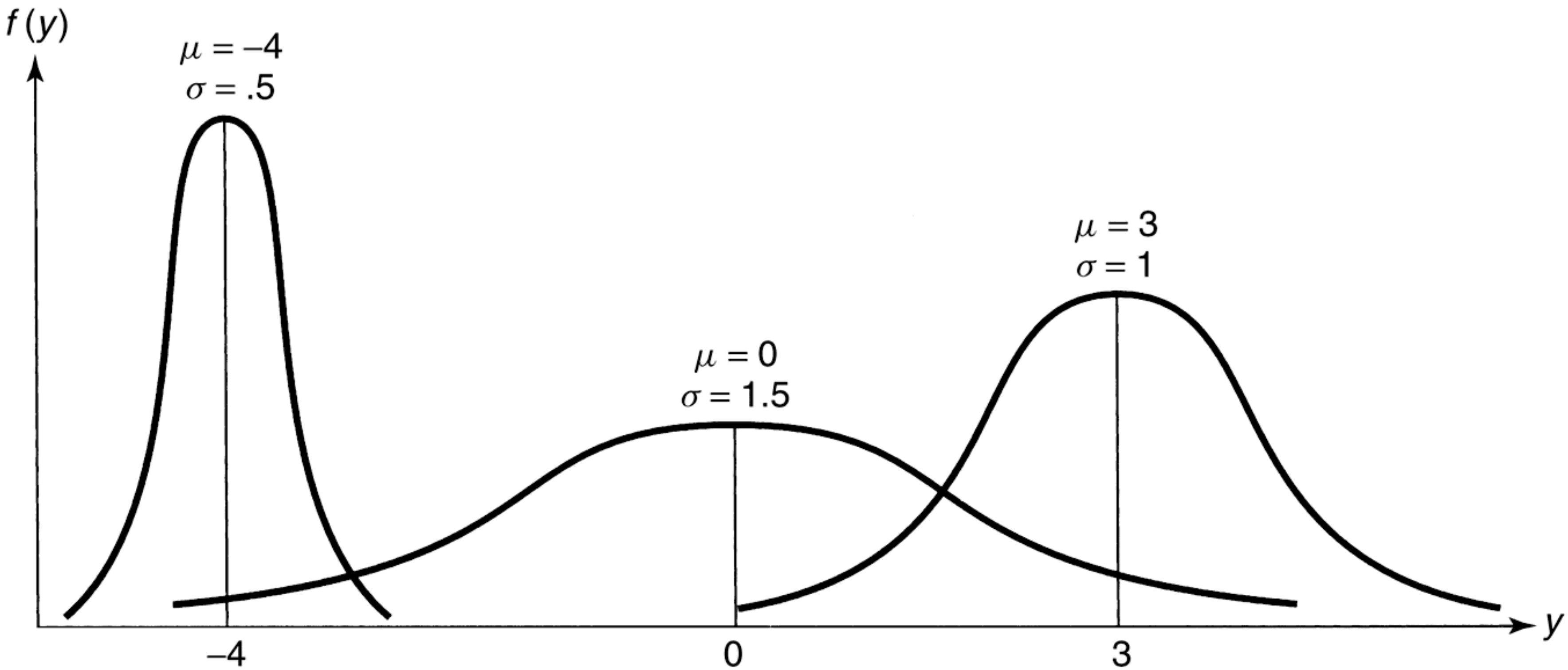
- ❑ It is a symmetric distribution.
- ❑ It is centered by mean μ
- ❑ Its spread is determined by STD σ

Data Types and Distributions

- There are two types of numerical variable: Discrete and Continuous
- Distribution for Discrete Variables
  - Binominal Distribution
  - Poisson Distribution
- Distribution for Continuous Variables
  - Normal Distribution

Normal Distribution

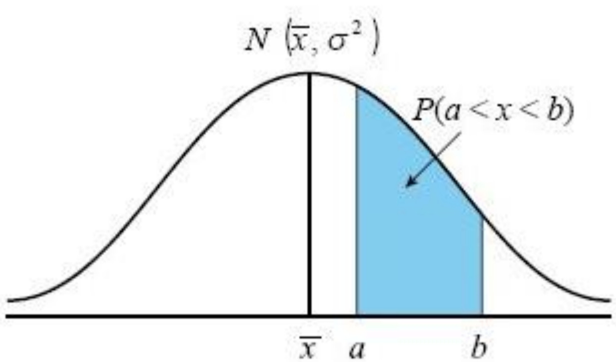
- Normal Probability Distribution: More Examples



Normal Distribution

- Normal Probability Distribution
  - ❑ It is a symmetric distribution.
  - ❑ It is centered by mean μ
  - ❑ Its spread is determined by STD σ

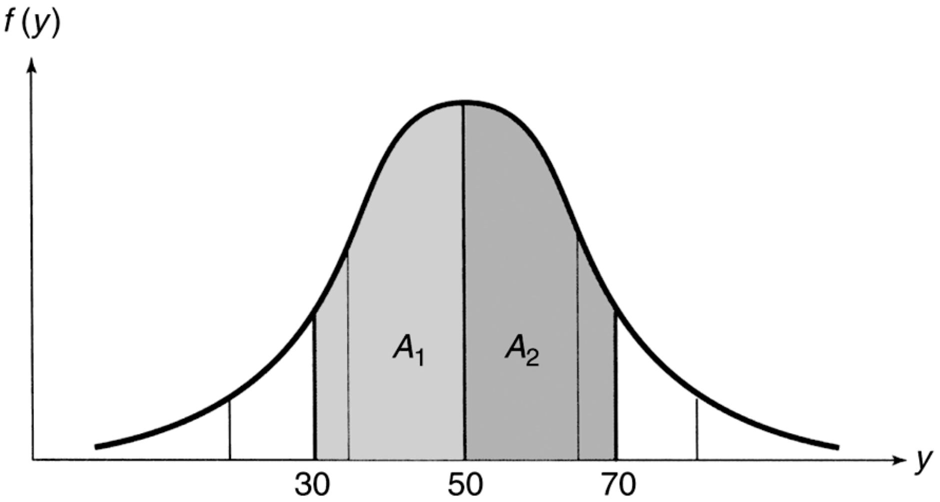
- Notes
  - – Variable X follows normal distribution, $X \sim N(\mu, \sigma^2)$



The normal curve area between a and b is the area under the normal distribution curve, and it is equal to the probability that x falls into the range [a, b]

Normal Distribution

- Example: Normal Distribution with $\mu = 50, \sigma = 15$



What is P (30 < y < 70)?

= Area of A1 + Area of A2

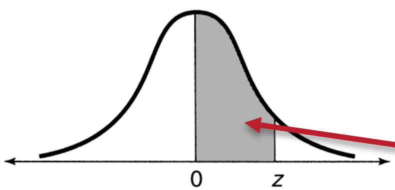
= 2 × Area of A2

= 2 × P (50 < y < 70)

Normal Distribution

- z score = # STDs from the data point to the mean

Table 1.7 Reproduction of part of Table 1 of Appendix D



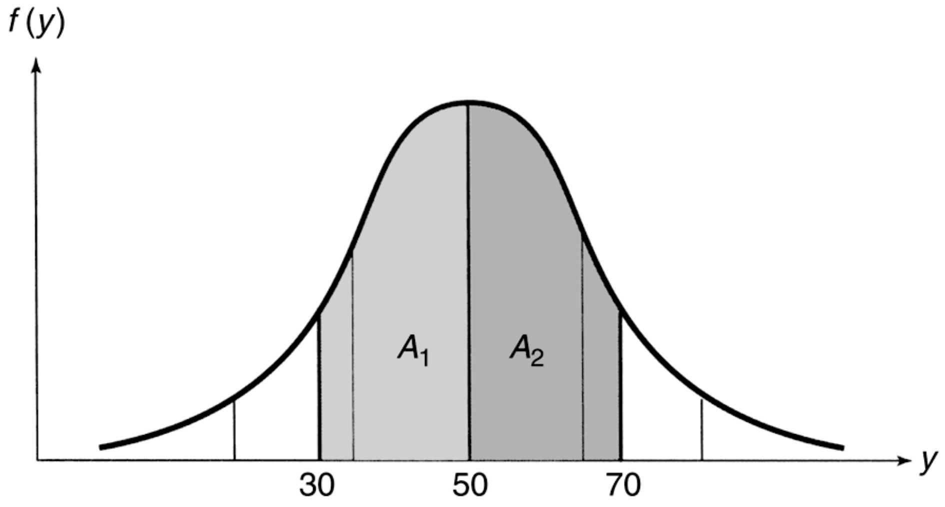
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

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

The area or the probability can be Inferred from the table on the left by assigning the specific z score

Normal Distribution

- Example: Normal Distribution with μ = 50, σ = 15



What is P (30 < y < 70)?

= Area of A1 + Area of A2

= 2 × Area of A2

= 2 × P (50 < y < 70)

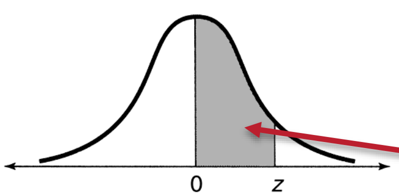
$z = \frac{y-\mu}{\sigma} = \frac{70-50}{15} = 1.33$, the area or probability P (50 < y < 70) = 0.4082

P (30 < y < 70) = 2 × 0.4082 = 0.8164

Normal Distribution

- z score = # STDs from the data point to the mean

Table 1.7 Reproduction of part of Table 1 of Appendix D



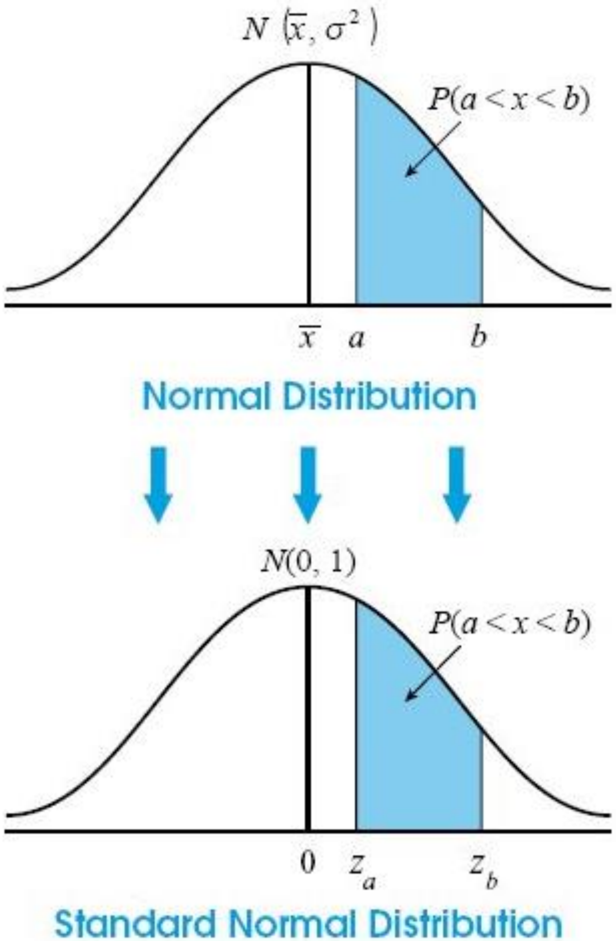
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

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

The area or the probability can be Inferred from the table on the left by assigning the specific z score

A z score refers to the number of STDs from the mean a data point is. Note: in z distribution, we assume we know population STD σ. Usually we do not know population mean, while we use sample mean.

Standard Normal Distribution



- For convenience, we usually transform normal distribution to a standard normal distribution, i.e., z distribution
  - ❑ It is a symmetric distribution.
  - ❑ It is centered by mean μ
  - ❑ Its spread is determined by STD σ

  - ❑ $\mu = 0$
  - ❑ $\sigma = 1$
  - ❑ X-axis represents z score
  - ❑ $z = (x - \mu)/\sigma$

Week 2-3

- Probability Distributions
- Sampling Distributions
- Central Limit Theorem

Sampling Distribution

- For example
  Population: average age of people in Illinois (13M)
  Population statistics: $\mu = 32$, $\sigma = 5$
  We get a sample of 20 people, mean = 28
  We get a sample of 20 people, mean = 29
  We get a sample of 20 people, mean = 31
  We repeat it again and again to get a list of sample means → We describe them by sampling distribution of the sample mean, i.e., the class frequency distribution of sample means through a large number of samples [Independent samples!!!!]

Sampling Distribution

- Mean

  The mean of sampling distribution of the sample means is equal to population mean, μ

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$

$$= \left(\frac{1}{n}\right)E(X_1 + X_2 + \ldots + X_n)$$

$$= \left(\frac{1}{n}\right)(E(X_1) + E(X_2) + \ldots + E(X_n))$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu)$$

$$= \frac{1}{n} \cdot n\mu = \mu$$

Sampling Distribution

❑Variance

The variance of sampling distribution of the sample means is equal to σ^2/n

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$

$$= \left(\frac{1}{n}\right)^2 Var(X_1 + X_2 + \ldots + X_n)$$

$$= \left(\frac{1}{n}\right)^2 (Var(X_1) + Var(X_2) + \ldots + Var(X_n))$$

$$= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \ldots + \sigma^2)$$

$$= \frac{1}{n^2}\cancel{n}\sigma^2 = \frac{\sigma^2}{n}$$

Sampling Distribution

- **Standard Deviation**

  The STD of sampling distribution of the sample means is equal to $SE_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

- **Standard Error of the estimate of sample mean**

  The STD above is also known as the standard error of the estimate of sample mean. It measures how accurate our estimation is. We expect this standard error to be as small as possible

Standard Deviation vs Standard Error

- The standard deviation of a variable X
  - It is used to measure of the data variation in X
- The standard error of the estimate of sample mean
  - It is used to measure how accurate our estimate is

Terminologies: Sampling Distribution

- If we are going to perform multiple independent experiments, we can collect multiple samples with same sample size: X1, X2, X3, X4, X5, …
- We calculate their means: $\overline{x1}, \overline{x2}, \overline{x3}, \overline{x4}, \overline{x5},$
- We focus on the distribution of these means: sampling distribution of sample means
- We found that, if n is large enough
  - mean of sample means = population mean
  - Standard deviation of sample means = $\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$ = Standard Error of the estimate
  - $\bar{x}$ ~ N(μ, $\frac{\sigma^2}{n}$), $\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, z $= \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

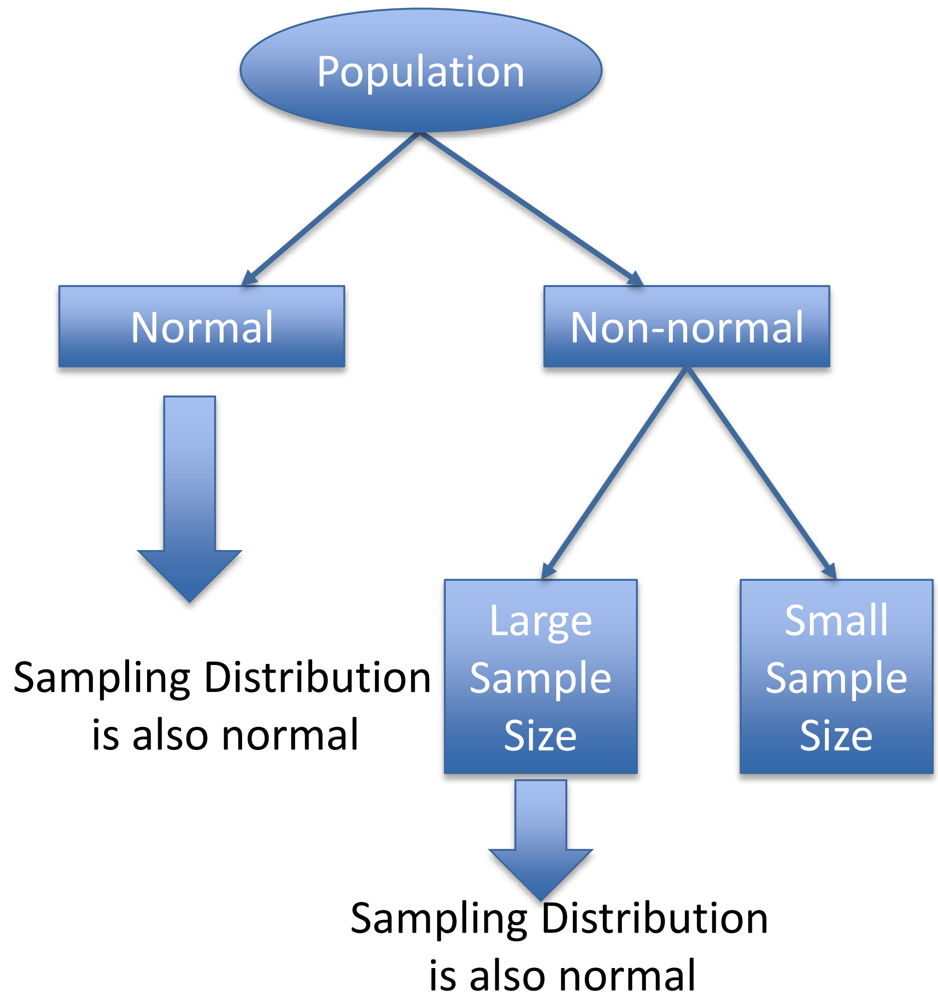
Week 2-3

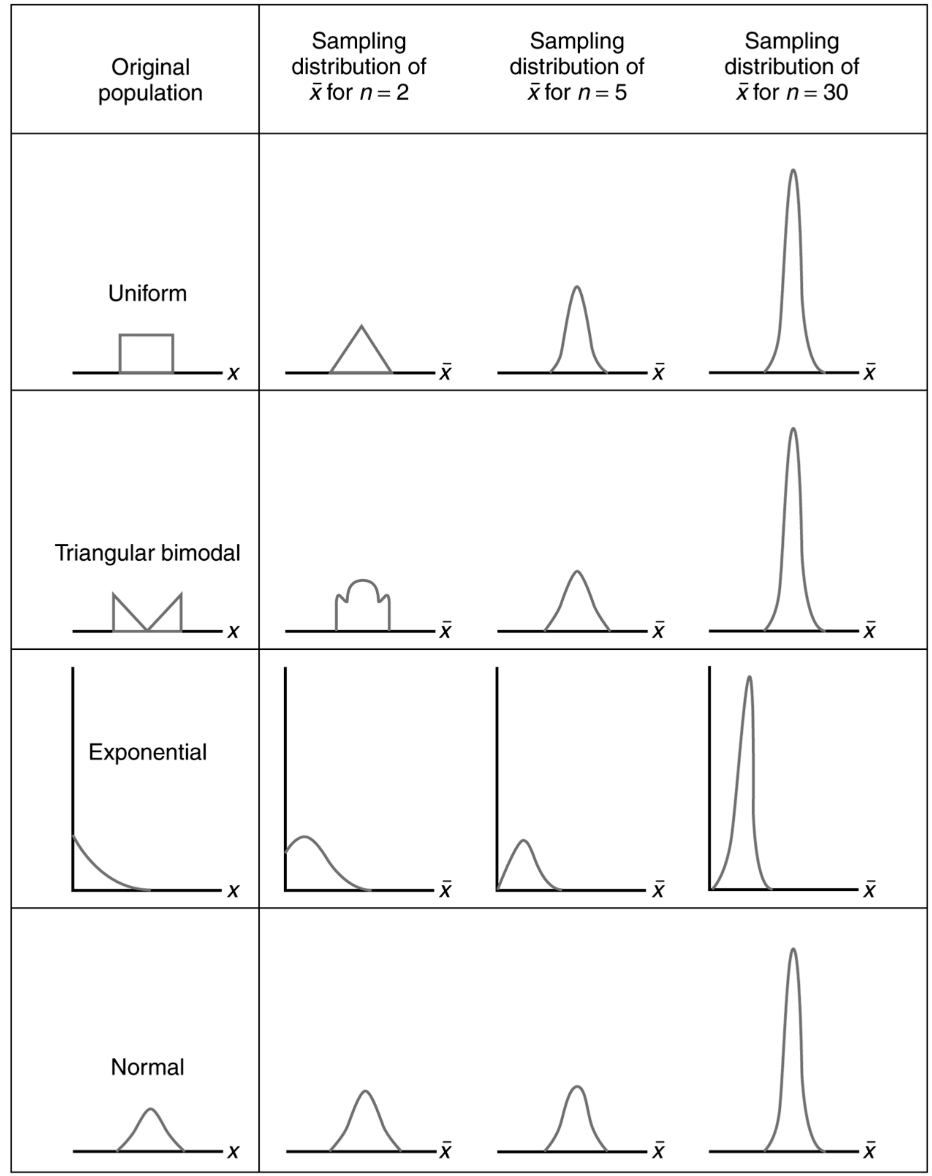
- Probability Distributions
- Sampling Distributions
- Central Limit Theorem

The Central Limit Theorem

- For large sample size (n>=30), the mean of a sample from a population with mean μ and STD σ has a sampling distribution (mean is μ, standard error is $\frac{\sigma}{\sqrt{n}}$) that is approximately normal, regardless of the probability distribution of the sampled population. It's better if the sample size is larger

The Central Limit Theorem





The Central Limit Theorem

- It is related to two important questions
  1) Why and how we can use sample statistics to estimate the population?
     The sample mean will follow normal distribution, while mean of sample means is population mean. We can use sample mean and SE to estimate the population mean
     It makes the *confidence interval*, *statistical inference* and *hypothesis testing* possible in data analytics
  2) Why we need normal distribution?
     It is easy for inference. We can describe distribution by mean and deviation. Based on CLT, we can assume it follows normal distribution as long as the number of samples is large enough, no matter what distribution it looks like.

Schedule

- Quick Reviews
- Numerical Data
  - Descriptive Statistics
  - Probability Distribution
- Intro: R

Introduction to R

- R, https://www.r-project.org/
- Open source, free, light weight
- With supports by many plugins/packages/libraries
- It is available for both Windows/Mac platforms
- R programming: R scripts/commands
- You can download and install either R or R Studio (https://www.rstudio.com/).

R: Snapshot

Console | R Markdown

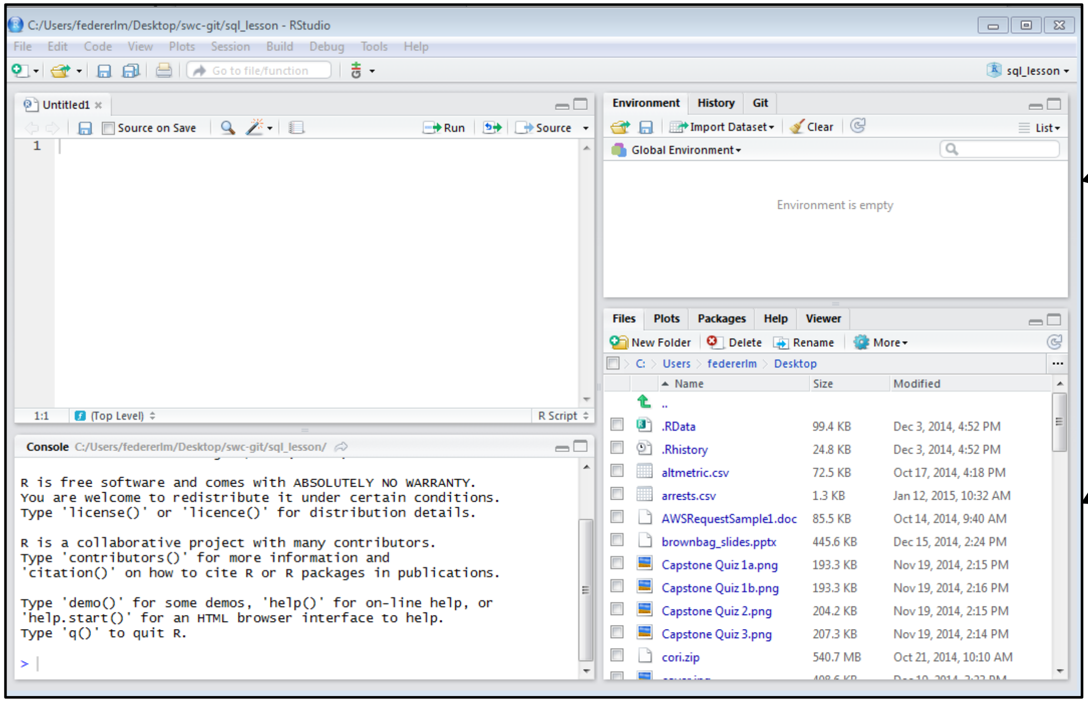
~/

```
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]

> full_data <- merge(batting2014, master, by = "playerID")
```

RStudio: Snapshot

R Script pane

Console



Environment pane

Navigation pane

Important Notes About R

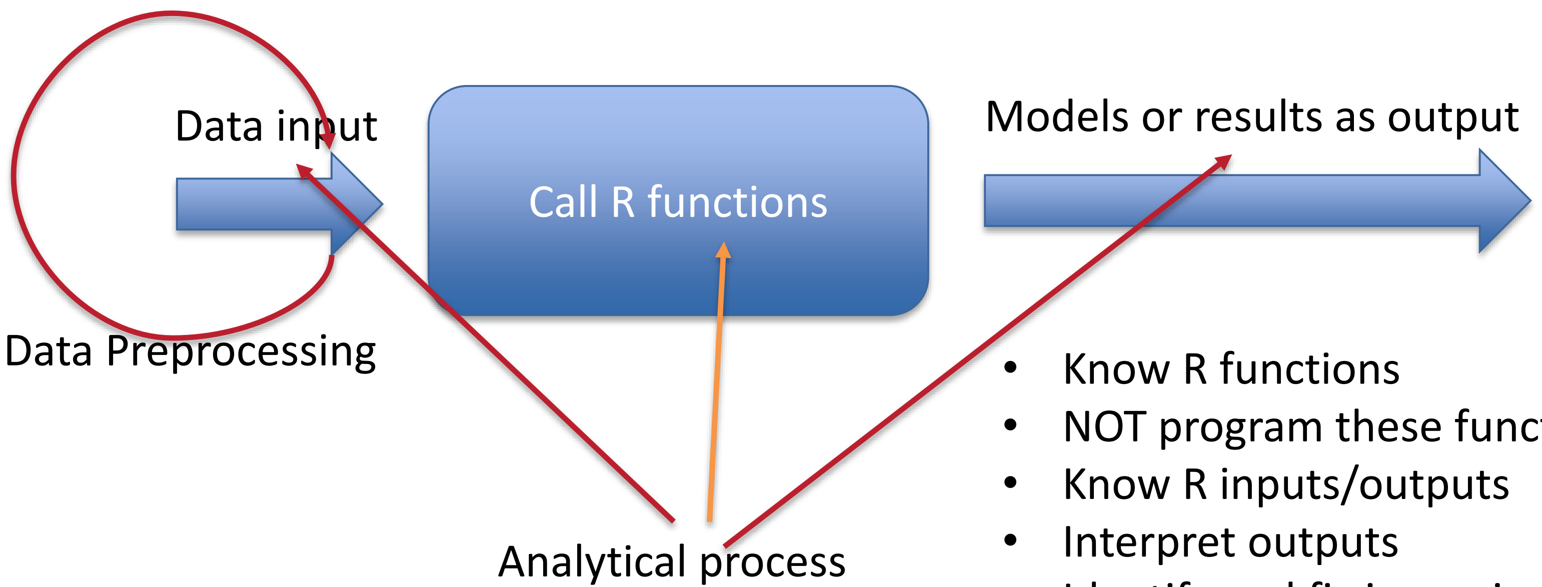
- Free R Manuals, https://cran.r-project.org/manuals.html
  - An Introduction to R
  - R Data Import/Export
- Find helps in R

R Help: help() and ?

The help() function and ? help operator in R provide access to the documentation pages for R functions, data sets, and other objects, both for packages in the standard R distribution and for contributed packages. To access documentation for the standard lm (linear model) function, for example, enter the command help(lm) or help("lm"), or ?lm or ?"lm" (i.e., the quotes are optional).

Important Notes About R

- R is considered as a scripting language, not a programming language

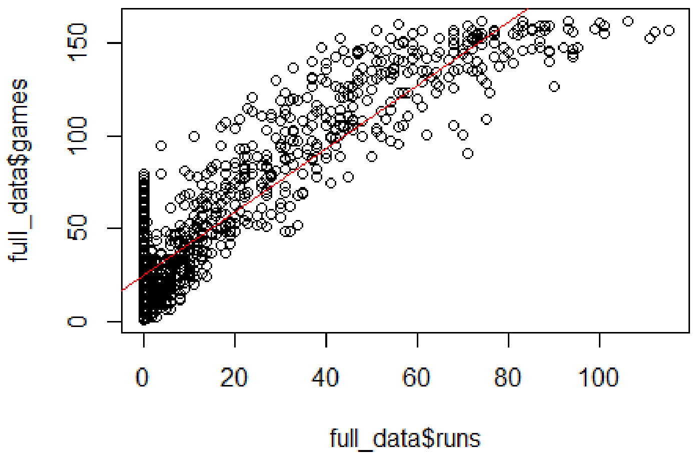


- Know R functions
- NOT program these functions
- Know R inputs/outputs
- Interpret outputs
- Identify and fix issues in outputs

Example of R Outputs

```r
stat.desc(batting_figures) #gives us a table of descriptive stats about each variable
```

```
##                      runs        hits      doubles          X3B
## nbr.val      1.435000e+03  1435.000000 1435.0000000 1.435000e+03
## nbr.null     6.680000e+02   609.000000  774.0000000 1.107000e+03
## nbr.na       0.000000e+00     0.000000    0.0000000 0.000000e+00
## min          0.000000e+00     0.000000    0.0000000 0.000000e+00
## max          1.150000e+02   225.000000   53.0000000 1.200000e+01
## range        1.150000e+02   225.000000   53.0000000 1.200000e+01
## sum          1.976100e+04 41595.000000 8137.0000000 8.490000e+02
## median       1.000000e+00     2.000000    0.0000000 0.000000e+00
## mean         1.377073e+01    28.986063    5.6703833 5.916376e-01
## SE.mean      6.159246e-01     1.261722    0.2574403 3.911696e-02
## CI.mean.0.95 1.208210e+00     2.475020    0.5050000 7.673260e-02
## var          5.443860e+02  2284.439136   95.1053635 2.195746e+00
## std.dev      2.333208e+01    47.795807    9.7521979 1.481805e+00
## coef.var     1.694324e+00     1.648924    1.7198481 2.504582e+00
```



R: Learning Style in this class

- We are not going to learn R programming step by step in the class
- We learn R for data analytics
  - Data inputs
  - Call R functions for descriptive & inferential statistics
  - Call R functions for data preprocessing
  - Learn how to interpret outputs, identify & fix issues
- R examples are provided in the class
- We do have in-class practices. I will provide one or two practice in which you learn from demos step-by-step

Schedule

- Next class: Using R for descriptive statistics
  – Install R or R studio by yourself in advance
  – Bring your laptop to the class
  – Learn R for descriptive statistics step-by-step