Introduction to Data Science

Live Session 06 - Unit 05 & 06

Future Plan

- Feb 21 : Live session 06 – unit 05 & unit 06
- Feb 28 : Live session 07 – unit 07
- Mar 07 : Live session 08 – Unit 08 (Case Study 1)
- Mar 14: Live session 09 – Unit 09 (API presentations)

Case study 01

- Feb 28 –
- Mar 07 – Discussion (Case study 01)
- Mar 14 – Submit before the live session

API Presentations (9.3)

- Install and load one of the packages given in the list for downloading APIs on this link: https://github.com/ropensci/opendata . The video gives a different URL, but I think this one is easier to navigate.
- The link is to a GitHub page. Examine the README file to find a list of R functions that interface with various APIs.
- Choose your favorite subject and select ONE R package from the list.
- Find/create an example to download some data using that library.
- Change an argument to the function/example to see what it does.
- Create a PPT presentation to show in the live session.
- And above all - have fun with it!

Future Plan

- Mar 14 – Unit 9 Videos (API presentations)
- Mar 21 – Unit 10
- Mar 28 – No live session
- Apr 04 – Unit 11
- Apr 11 – Unit 12 Videos (Python)
- April 17 – Python Presentations
- April 24 – Case Study II

Office hours

- Raunak : Friday 6.30p.m.-7.30p.m. CT
- Chen Mo :

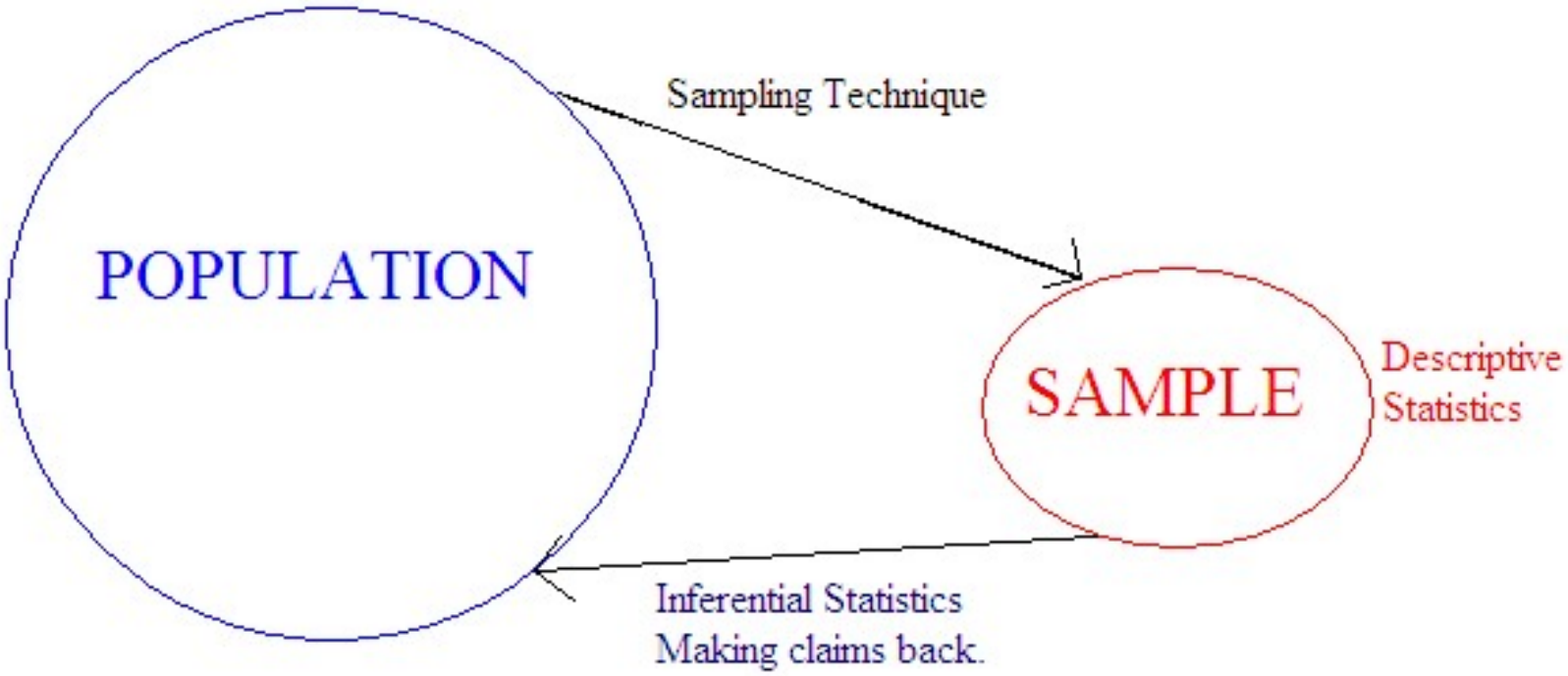
Write your own function

- Specify the length (n) as the argument.
- Produce n data points from standard normal distribution.
- Output the percentage of how many data points in between -2 to +2.

```
test<-function(n){
  x<-rnorm(n)
  count<-0
  for(i in 1:n){
    if (x[i]>= -2 & x[i] <=2){
      count=count+1}
    }
perc<-(count/n)*100
return(perc)
}

test(1000)
```

Parameters and Statistics



Parameters and Statistics

A *statistic* is a characteristic or measure which uses the data values from a *sample.*

A *parameter* is a characteristic or measure which uses all of the data values from a specific *population*.

Parameters and Statistics (Mean)

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots\ldots\ldots + X_n}{n} = \frac{\sum X}{n}$$

Statistic

$$\mu = \frac{X_1 + X_2 + X_3 + \ldots\ldots\ldots + X_N}{N} = \frac{\sum X}{N}$$

Parameter

Sampling Distribution

- Remember:
  - A *parameter* is a numerical index that describes some feature of a population (universe)
  - A *statistic* is a numerical index that describes some feature of the sample. It is a function of the data collected in a sample.
- The sampling distributions tells us how the statistic is related to the parameter.

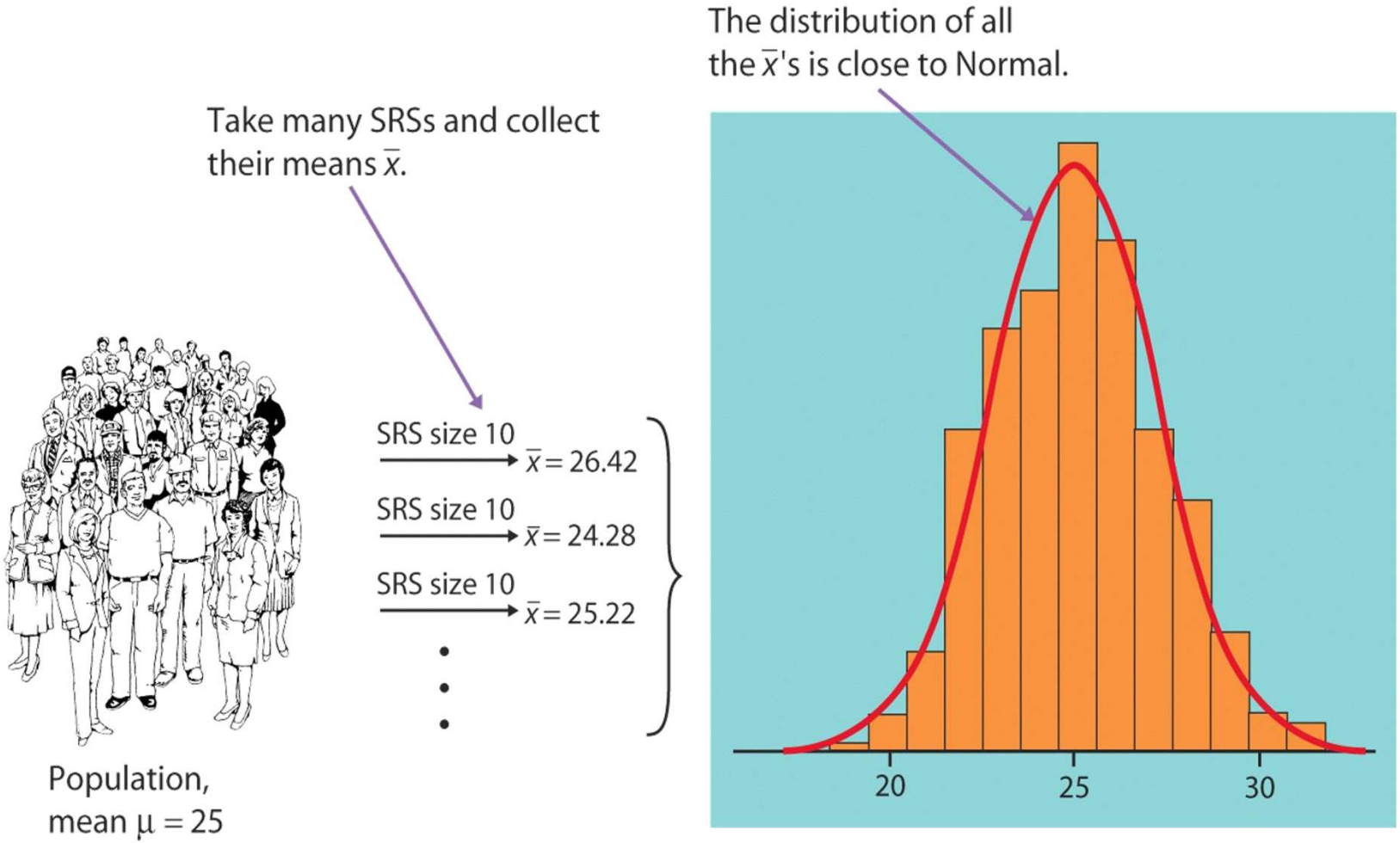
Sampling Distributions

Let's say we're interested in the mean of a population.

- We take a sample and find $\bar{x}$ and s
- What do this tell us about the population mean? Can we make a guess based on this?
- We need to know how sample means are distributed (in relation to the population distribution) to answer this.
- The distribution of $\bar{x}$ (all the sample means) is a sampling distribution

Sampling Distribution

- what would happen in many samples?



Sampling Distributions

This is a helpful site in understanding sampling distributions: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Central Limit Theorem

What we have just shown is the Central Limit Theorem!

Central Limit Theorem (CLT)

Say we take a **SRS** of size n from any population with mean μ and standard deviation σ.

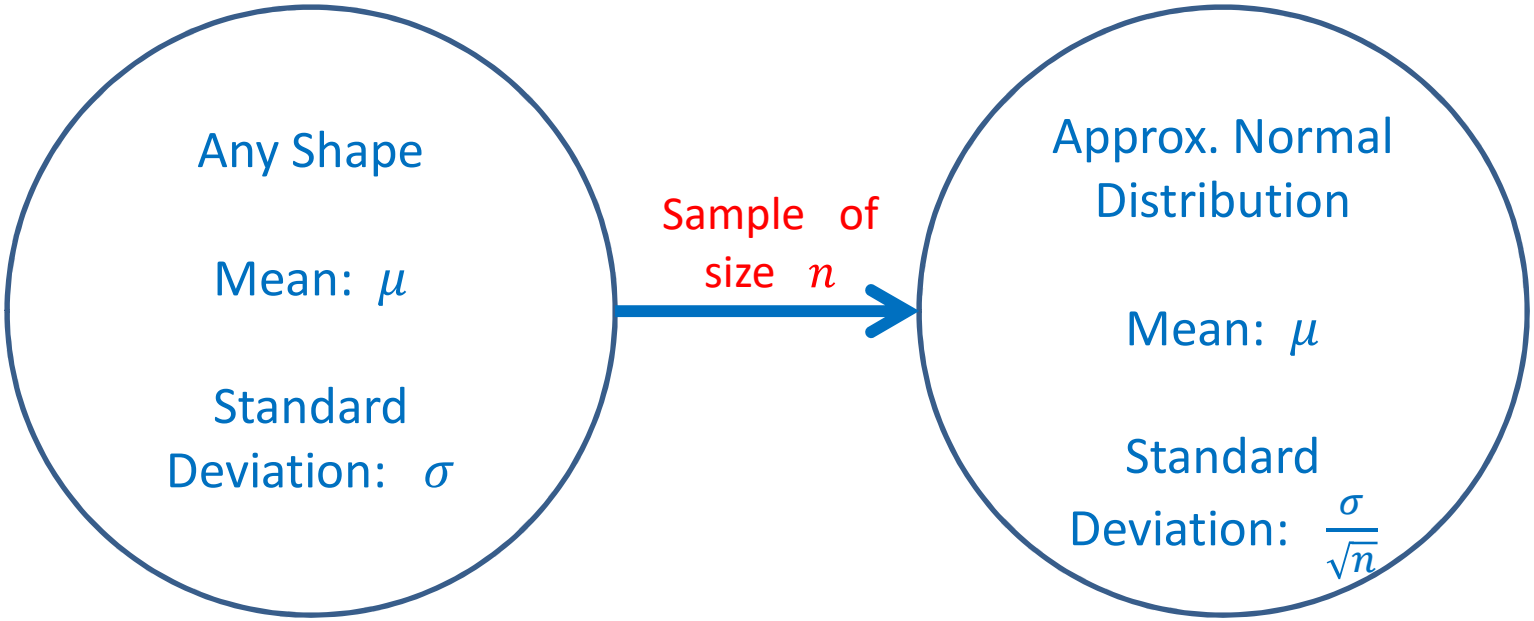
If n is large enough, the sampling distribution of the sample mean is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

This is why the normal distribution is so important.

Central Limit Theorem

Original Population: x

Population of means, $\bar{x}$, of **all** samples



Estimation

Goal: How can we use sample data to estimate values of population parameters?

Point estimate: A single statistic value that is the "best guess" for the parameter value

Interval estimate: An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. Called a ***confidence interval***.

(Based on sampling distribution and the point estimate)

Confidence Interval

- A **confidence interval** (CI) is an interval of numbers believed to contain the parameter value.

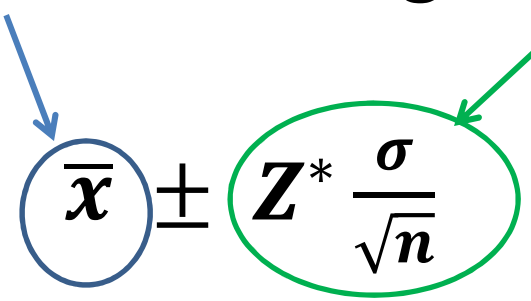
- The probability the method produces an interval that contains the parameter is called the **confidence level.** Most studies use a confidence level such as 0.95 or 0.99.

- Most CIs have the form

point estimate ± margin of error

Confidence Interval

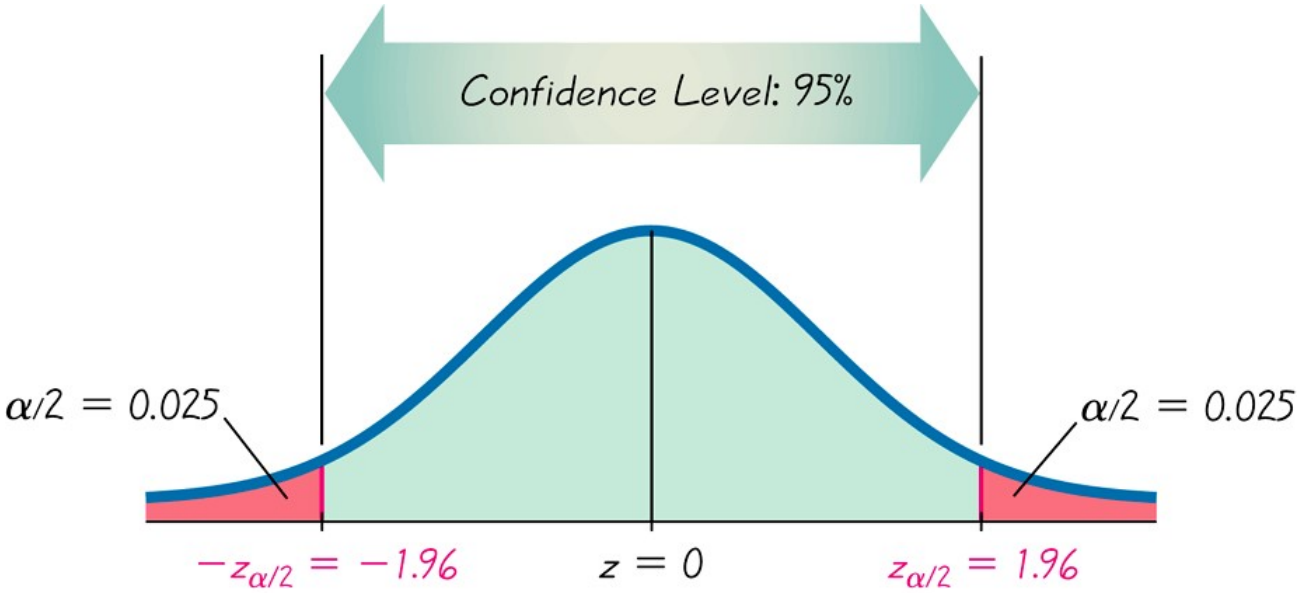
point estimate ± margin of error



$$\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

Confidence Interval $= \left(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + Z^* \frac{\sigma}{\sqrt{n}} \right)$

Finding $z_{\alpha/2}$ for a 95% Confidence Level



$-z_{\alpha/2}$ $z_{\alpha/2}$

Critical Values

α = 5%

$\alpha/2$ = 2.5% = .025

```
> qnorm(c(0.025,0.975))
[1] -1.959964  1.959964
```

z^* Values

- Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	*z^* value*
90%	**1.64**
95%	**1.96**
99%	**2.58**

Confidence Interval

- A large school gives a standardized test every year. In the past, the standard deviation is 200. This year, the average of the first group of students taking the test was 510. Assume that the first group of students is a simple random sample with size 50 and the scores of the test follow a normal curve. What is the point estimate for the population average score? Calculate 95% confidence interval?

point estimate? = $\bar{x}$ = 510

Confidence Interval

- A large school gives a standardized test every year. In the past, the standard deviation is 200. This year, the average of the first group of students taking the test was 510. Assume that the first group of students is a simple random sample with size 50 and the scores of the test follow a normal curve. Calculate 95% confidence interval?

$$\bar{x} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

$$\mathbf{510 \pm 1.96 \frac{200}{\sqrt{50}}}$$

$$510 \pm 55.44$$

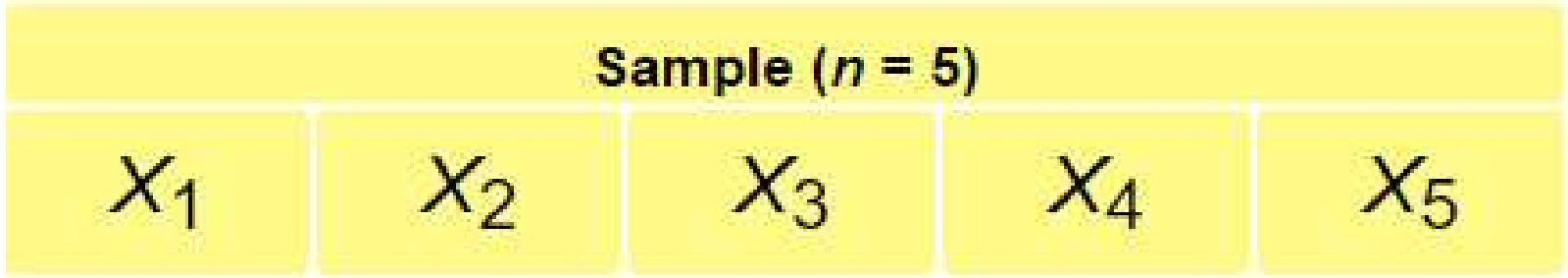
95% confidence interval = (510 – 55.44 , 510 + 55.44)

= **(454.56 , 565.44)**

Bootstrap Definition

- Let $X_1, X_2, \ldots, X_n$ be a sample of n IID observations from a distribution F.
  - IID = independent, identically distributed
- Let θ be a parameter of distribution F.
- Let $\theta(X_1, X_2, \ldots, X_n)$ be a real valued estimator of θ.
- Replace F with F_n
  - Where F_n places probability mass $\frac{1}{n}$ at each observation X_i
  - Weighting all the observations equally

Bootstrap Sampling with Replacement



Bootstrap Sample

X_1^* X_2^* X_3^* X_4^* X_5^*

Big idea!

- Two sample designs can be compared by comparing the behavior of the estimators that produce, over many realizations of the sample.
- Two important features of the sampling distribution: its mean and variance.
  - Small (or no!) bias is good.
  - Small variance is good.
- When comparing two designs producing unbiased estimators, the better one is the one with smaller variance
- When comparing two designs producing estimators that are not unbiased, the better one is the one with the smaller mean squared error (MSE), defined as:

$$\text{MSE} = \text{Variance of estimator} + \text{Bias}^2.$$

Assessing Accuracy of Estimates

- Mean square error (MSE)

$$E\left(\hat{\theta}-\theta\right)^2$$

Bootstrap MSE for General Statistic

1. Compute statistic (θ^*) from the original data.
2. Draw a sample size of n from the original data, with replacement.
3. Repeat REP times. (REP = 1000, 5000)
4. Compute statistic ($\theta_{b,i}^*$), the estimate of θ^* from the ith bootstrap sample.
5. Obtain bootstrap MSE.

$$\hat{MSE} = \frac{1}{reps}\sum_{i=1}^{reps}\left(\theta_{b,i}^* - \theta^*\right)^2$$

Bootstrap Bias and Variance

- **Statistical bias:** difference between expected value of an estimate and quantity being estimated

$$B = E\left(\theta^*\right) - \theta$$

- MSE = variance + bias^2

Bootstrap estimates of bias and variance:

$$E^* = \frac{1}{REP}\sum_{i=1}^{REP}\theta_{bj}^* \qquad B^* = E^* - \theta^* \qquad \text{var} = \frac{1}{REP}\sum_{i=1}^{REP}\left(\theta_{bj}^* - E^*\right)^2$$

Question 1

- The bootstrap is most useful in the following situation (only one answer).
  a. When we want to recreate the sampling distribution of a statistic from observed data.
  b. When the sample size is small.
  c. When we know the population distribution.
  d. When we want to sample with replacement.

Question 2

- When we calculate a bootstrap mean squared error (MSE), what do we use to estimate the value of the parameter of interest?
  a. The value of the parameter in the original population.
  b. The parameter value is estimated from one bootstrap sample.
  c. The bias of a bootstrap sample
  d. The parameter value is estimated from the original sample.

Question 3

- What are the two sources of variation that occur from employing the bootstrap?
  a. Variation from drawing a sample from the original population and from obtaining random samples from the original sample
  b. Variation from drawing the original sample from the population and mistakes in coding.
  c. Variation from drawing the original sample from the population and variation in calculating statistics from each sample
  d. Variation from drawing the original sample from the population and variation in the structure of each problem.

Git Hub and RStudio

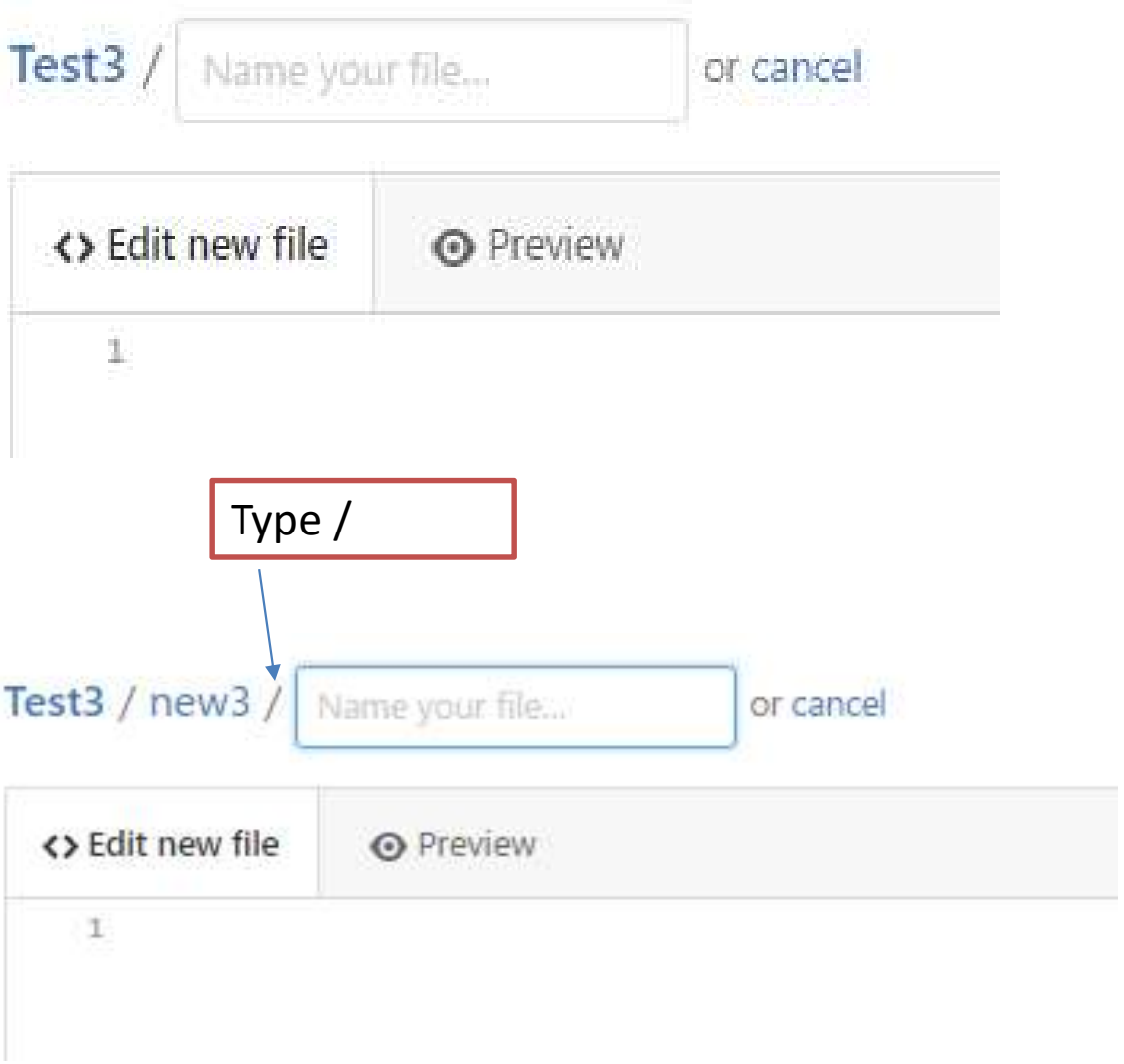
- Creating RStudio projects from GitHub Repositories

https://www.youtube.com/watch?v=YxZ8J2rqhEM

- GitHub and Git Bash: Create Folders in a Repository

https://www.youtube.com/watch?v=rVNFPj9jtb0

New Folder in GitHub



Live Session Assignment

Case Study Practice - due on Tuesday, February 28.

- Rolling Housing Sales for NYC
- I have code for an analysis of Brooklyn housing data on the course website (from pages 49 and 50 in the O'Neil and Schutt text). Using the Rolling Data Sales website, download and examine another housing sales data set, which will be given to you in the breakout room.
- Goal: Create an RStudio project for the analysis of this data set. Your file structure within the project should include the following:
  - A README file in the project root directory that includes an explanation of the purpose of the project and the other files
  - A data directory containing files to load in and clean up the data. The clean up should include finding out where there are outliers or missing values, deciding how you will treat them, making sure values you think are numerical are being treated as such (correct R class), etc.
  - An Analysis directory containing a file (or files) for exploratory data analysis on the clean data to visualize the relationship between square footage and sales price.
  - A Paper directory containing a file (plain text or Markdown) that explains any meaningful patterns in this dataset.
- Deliverable: A link to a repository (test-repo is fine) on GitHub containing the above. I need only one link per group. Since this is a group project, I expect you to divide the labor.