Chapter 1

Statistics and Data Science

Professor Jung Jin Lee

1.1 Statistics and Data Science

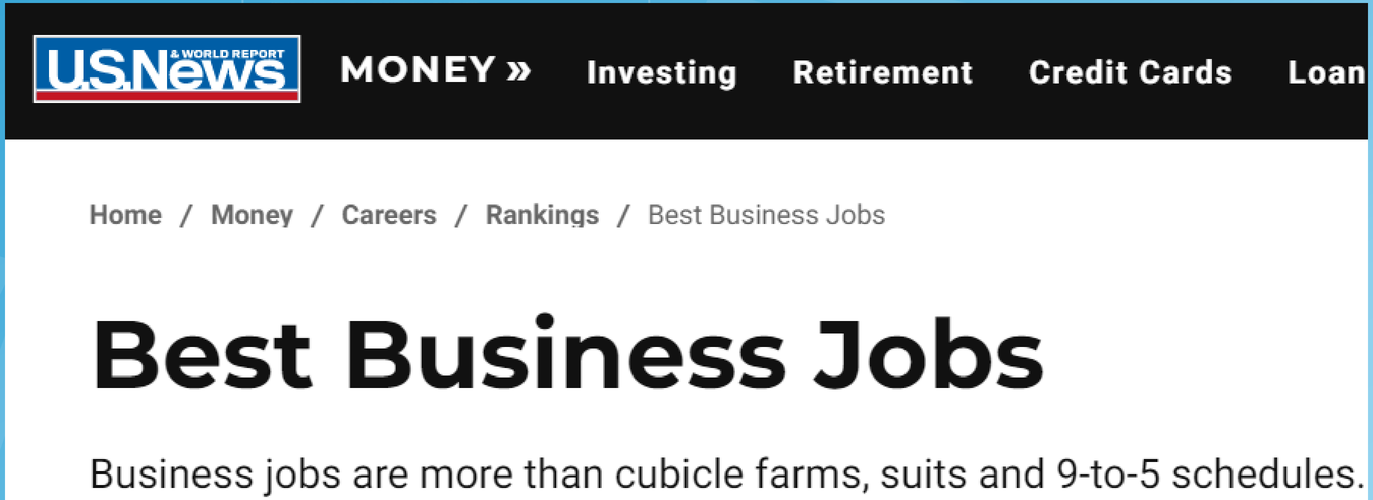
1.2 Population and Sample

1.3 Variables and Data

1.4 Software for Statistical Analysis

1.1 Statistics and Data Science

https://money.usnews.com/careers/best-jobs/rankings/best-business-jobs



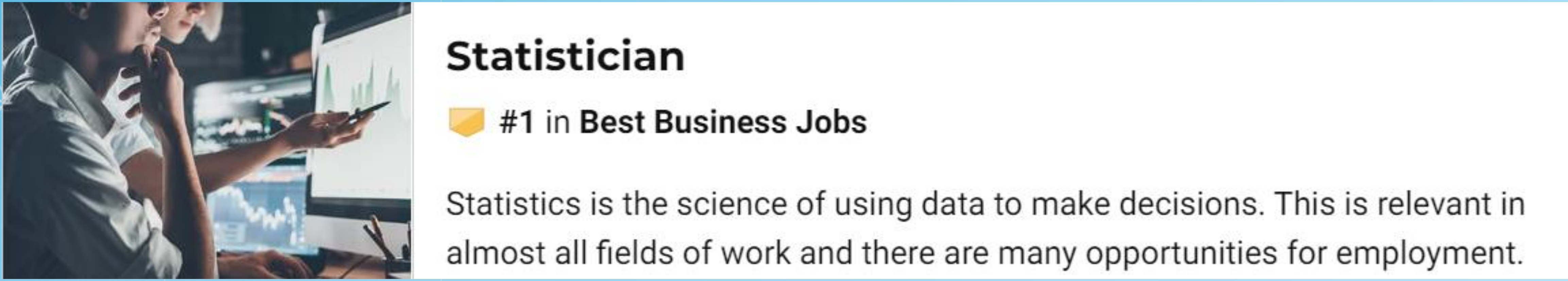
Home / Money / Careers / Rankings / Best Business Jobs

Best Business Jobs

Business jobs are more than cubicle farms, suits and 9-to-5 schedules.

In Google, type

‘Best Business Jobs US News Careers’

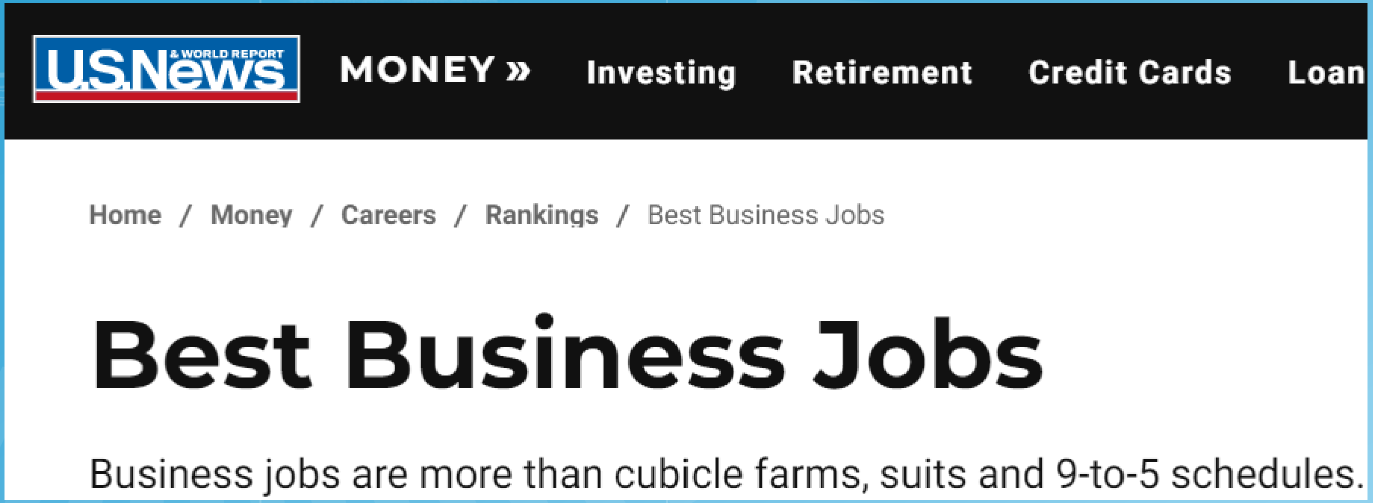


Statistician

#1 in Best Business Jobs

Statistics is the science of using data to make decisions. This is relevant in almost all fields of work and there are many opportunities for employment.

1.1 Statistics and Data Science

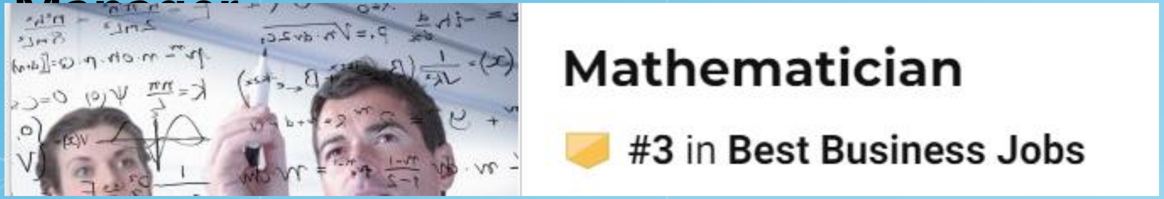


Home / Money / Careers / Rankings / Best Business Jobs

Best Business Jobs

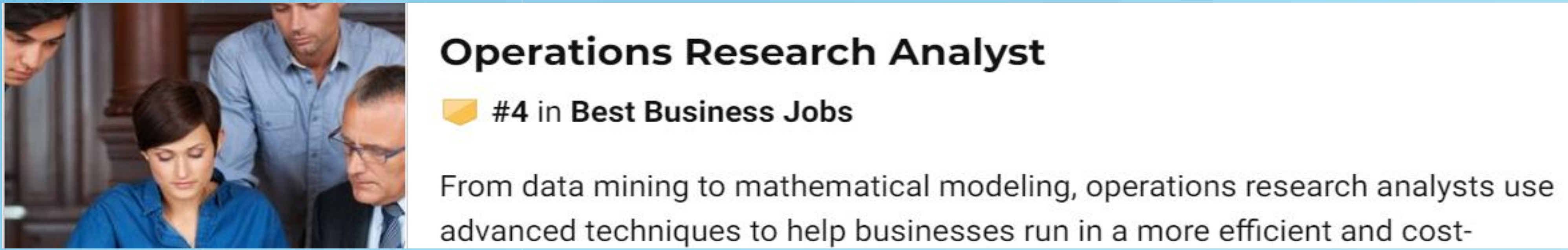
Business jobs are more than cubicle farms, suits and 9-to-5 schedules.

#2 Medical and Health Service Manager



Mathematician

#3 in Best Business Jobs



Operations Research Analyst

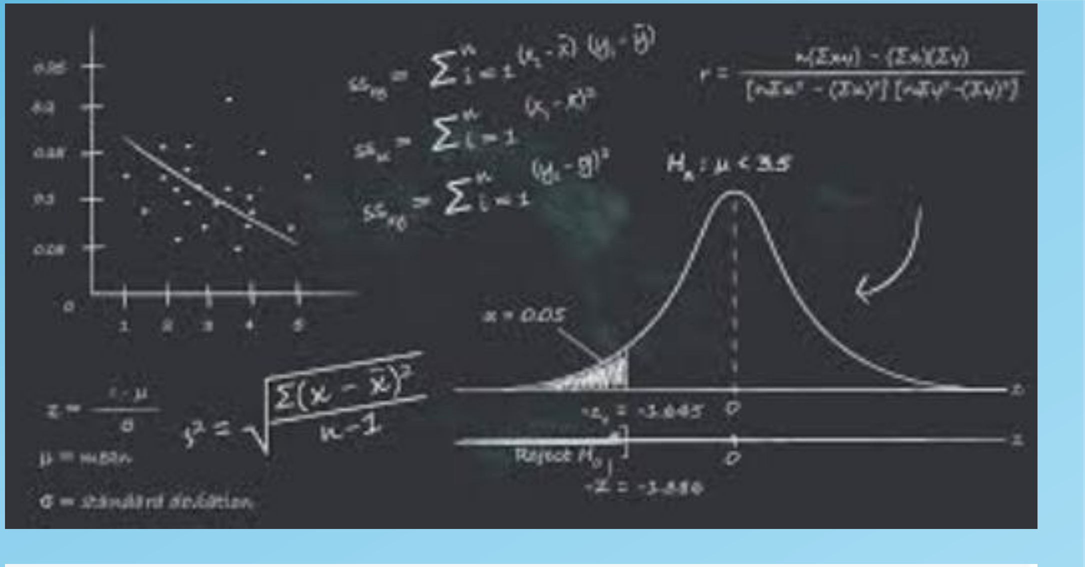
#4 in Best Business Jobs

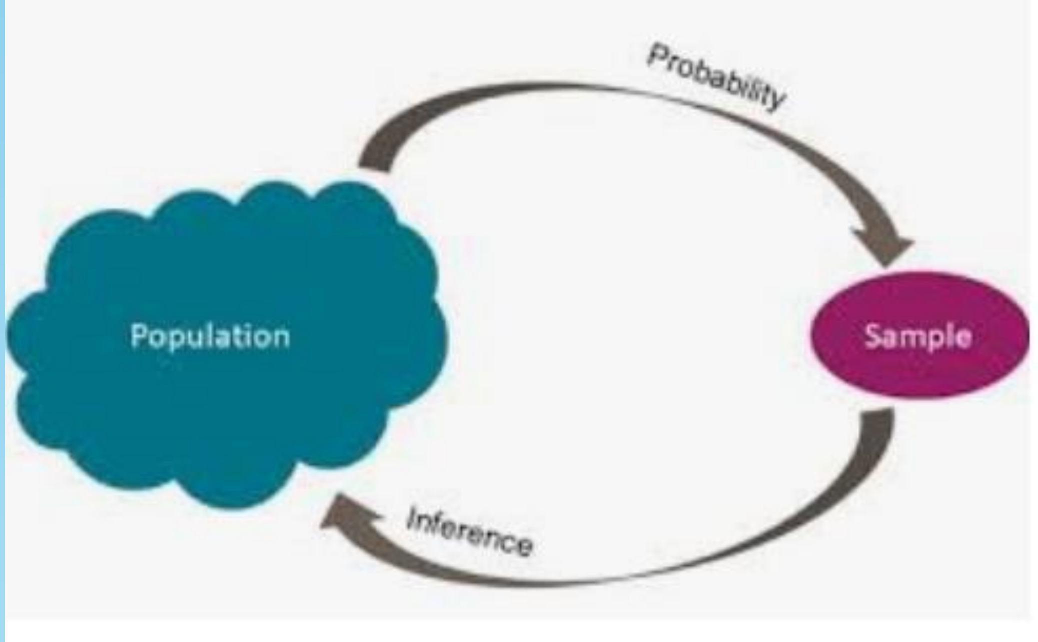
From data mining to mathematical modeling, operations research analysts use advanced techniques to help businesses run in a more efficient and cost-

#5 Financial Manager
#6 Financial Advisor
#7 Accountant
#8 Market Research Analyst
#9 Business Operation Manager
#10 Social and Community Service Manager
#11 Actuary

1.1 Statistics and Data Science

- **Statistics**





Statistics = ‘State ‘ + ‘istics’

1.1 Statistics and Data Science

▪ Statistics

- History tells which country appeared where, when, how large its territory, how much population and how many households
  - In Egypt, Greece and Rome, population and farmland area were used for the management of their country.
- 8th to 13th century, concept of probability and inference
  Al-Khalil (717–786), Al-Kindi (801–873), Ibn Adlan (1187–1268)
- 17th to early 19th century, mathematical foundations of statistics
  Gerolamo Cardano, Blaise Pascal and Pierre de Fermat.
- late 19th century, Francis Galton and Karl Pearson
- early 20th century, Ronald Fisher

1.1 Statistics and Data Science

▪ Statistics

- Today, statistical methods are applied in all fields
  => decision making, accurate inferences from data
  => management, economics, politics, social science, education
  physics, chemistry, biology, computer science
  medical science, pharmacy, agricultural science
  electrical, electronical, chemical, civil engineering

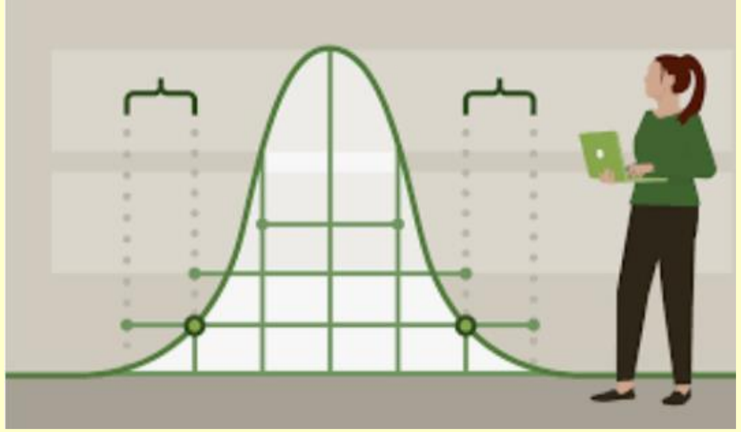
- Modern computers has expedited large-scale statistical computations.

1.1 Statistics and Data Science

▪ Statistics

- **Modern statistics is the discipline**
  => efficiently collect data, summarize data
  => analyze data to make scientific decisions using various probabilistic models in uncertain situations.

  - **Company predicts sales, government establish economic development plan**



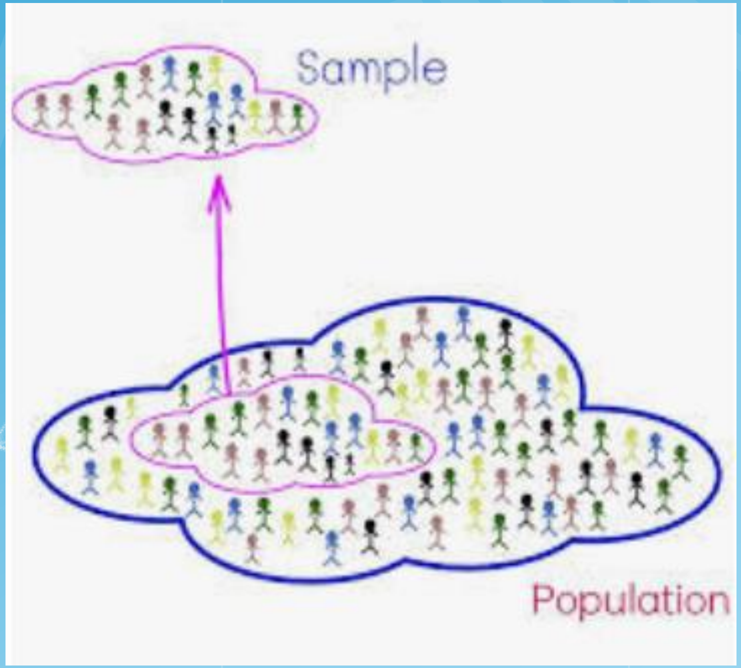
1.1 Statistics and Data Science

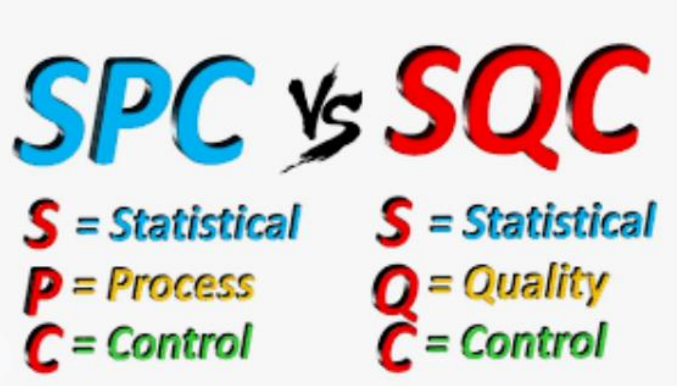
▪ Application of Statistics

- sample surveys to predict the winners of the election.

- Test new drug by a pharmaceutical company.

- Quality Control



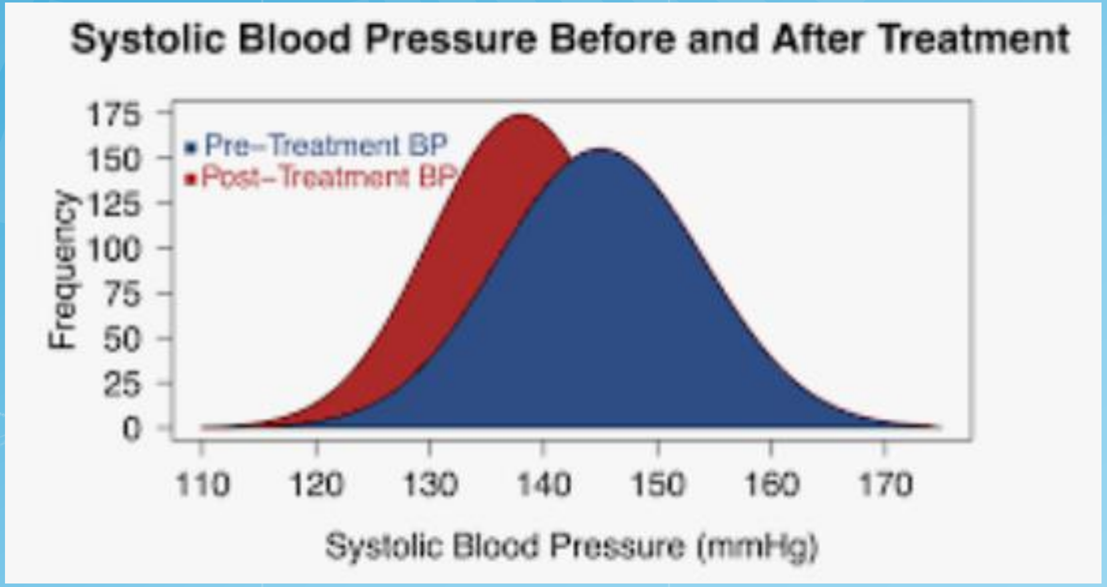


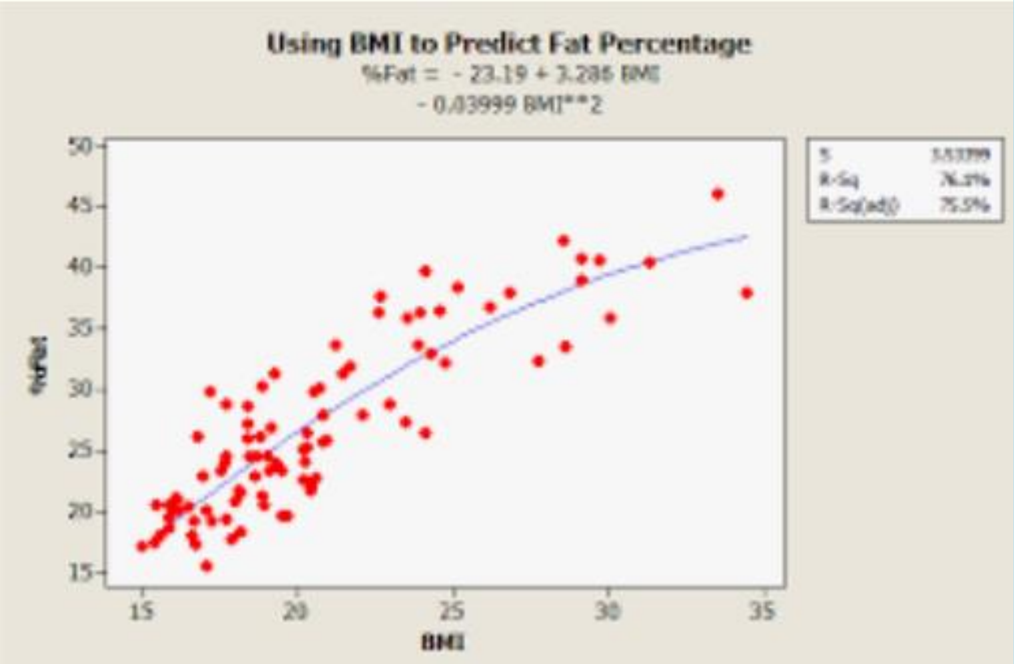
1.1 Statistics and Data Science

▪ Application of Statistics

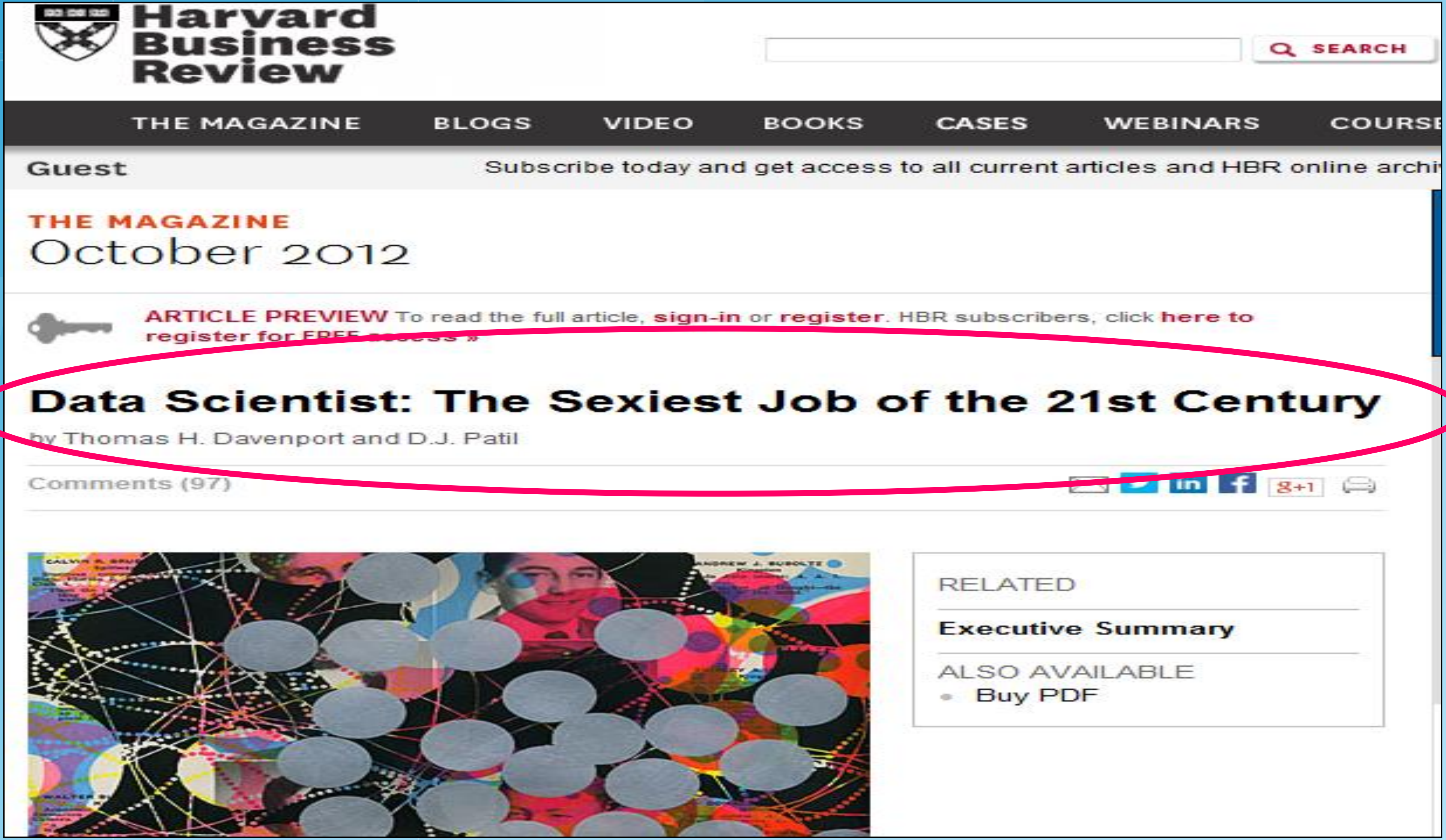
- Examine blood pressure before and after treatment

- Using BMI to predict FAT Percentage





1.1 Statistics and Data Science



Harvard Business Review

SEARCH

THE MAGAZINE BLOGS VIDEO BOOKS CASES WEBINARS COURSE

Guest

Subscribe today and get access to all current articles and HBR online archi

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (97)

RELATED

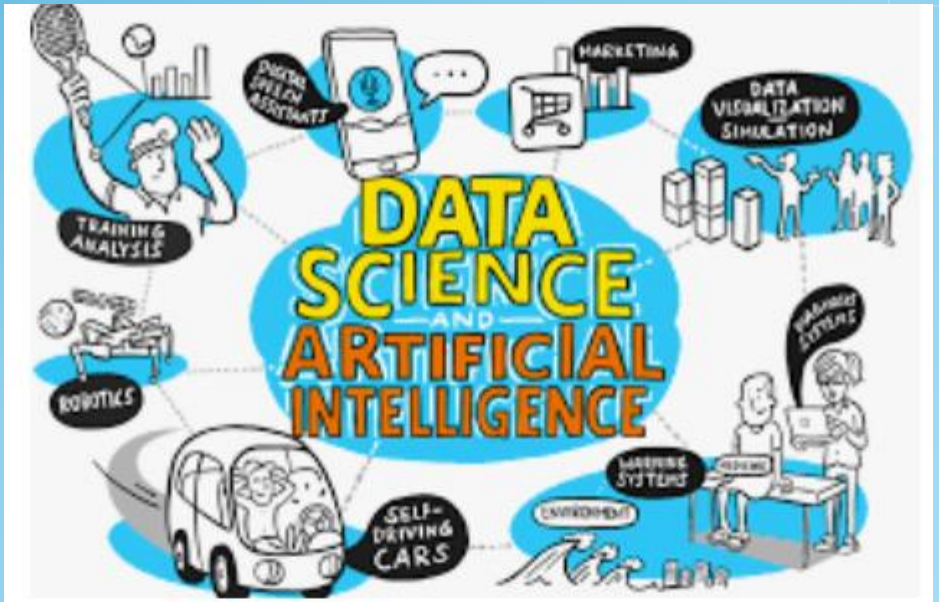
Executive Summary

ALSO AVAILABLE

- Buy PDF

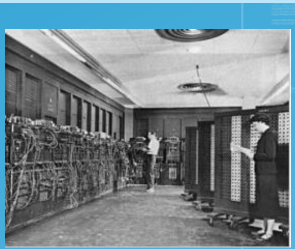
1.1 Statistics and Data Science

Data Science


DATA SCIENCE AND ARTIFICIAL INTELLIGENCE
MARKETING
DIGITAL SPEECH ASSISTANTS
DATA VISUALIZATION SIMULATION
TRAINING ANALYSIS
ROBOTICS
SELF-DRIVING CARS
WARNING SYSTEMS
ENVIRONMENT

1.1 Statistics and Data Science

▪ Rapid advance of computer and tele-communication technology



- 1946 Modern Digital Computer(ENIAC) by Eckert and Mouchly of Univ of Pennsylvania

- 1981 IBM Personal Computer
- Microsoft Operating System by Bill Gates

The CERN data centre in 2010 housing some WWW servers

- 1990s Networking of computers in the world
- World Wide Web by Berners-Lee
- Google search engine
- Yahoo, MSN web portal

Two smartphones: a Samsung Galaxy J5 (left) and an iPhone 6S (right)

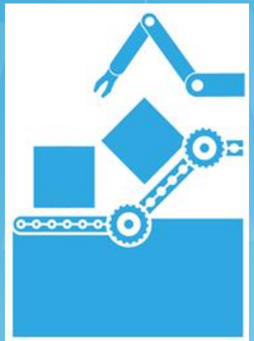
- 2000s Smartphone = PC + Phone
- www + wireless connection of Smartphone
- YouTube, Facebook, Twitter, LinkedIn

Big Data

- SNS Data
- web log data
- Bank data
- credit card data
- Health data

1.1 Statistics and Data Science

18th Century
1st Industrial Revolution



19th –Early 20th Century
2nd Industrial Revolution

Late 20th Century -
3rd Industrial Revolution

4th Industrial Revolution is under going by using Big Data

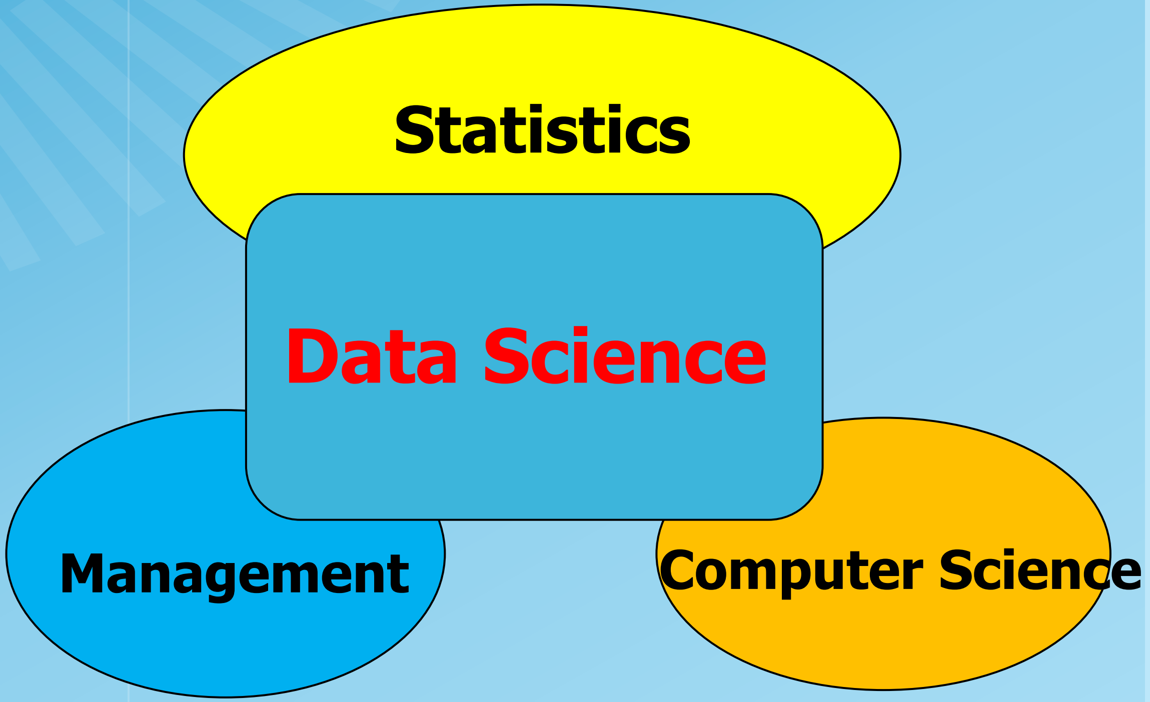
- Artificial Intelligence (AI)
- Internet of Things (IoT)
- Hyper-forecasting

Automatic driving car
3D printing
Virtual Reality
Alpha Go
....

1.1 Statistics and Data Science

- Data Science is to collect big data, analyze and apply it in real life

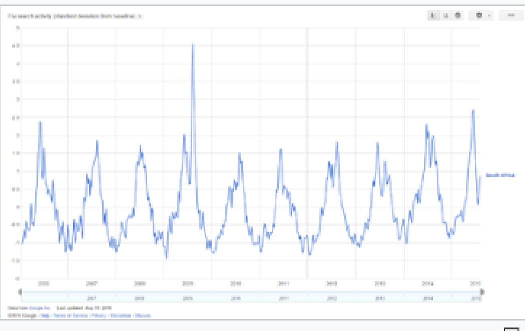
⇨ Data Science is a fusion of several science



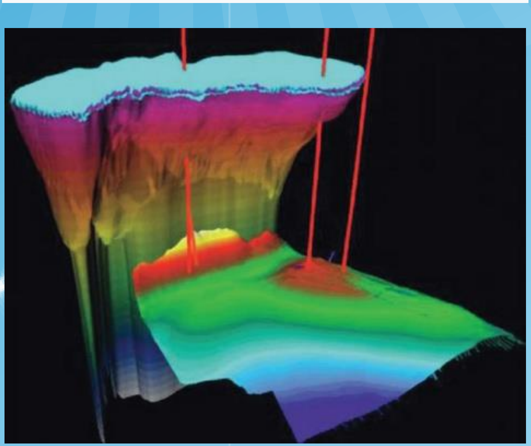
- Probability
- Estimation
- Testing
- Sampling
- Multivariate Stat Anal
- Database
- Information Retrieval
- Distributed Computing
- Artificial Intellignece
- Pattern Recognition
- Machine Learning
- Optimization
- MIS
- Marketing

1.1 Statistics and Data Science

❖ Example of Data Science


Google Flu Trends data, South Africa

- Google Flu Trend to estimate influenza activity



- Crude oil exploration

- Market basket analysis

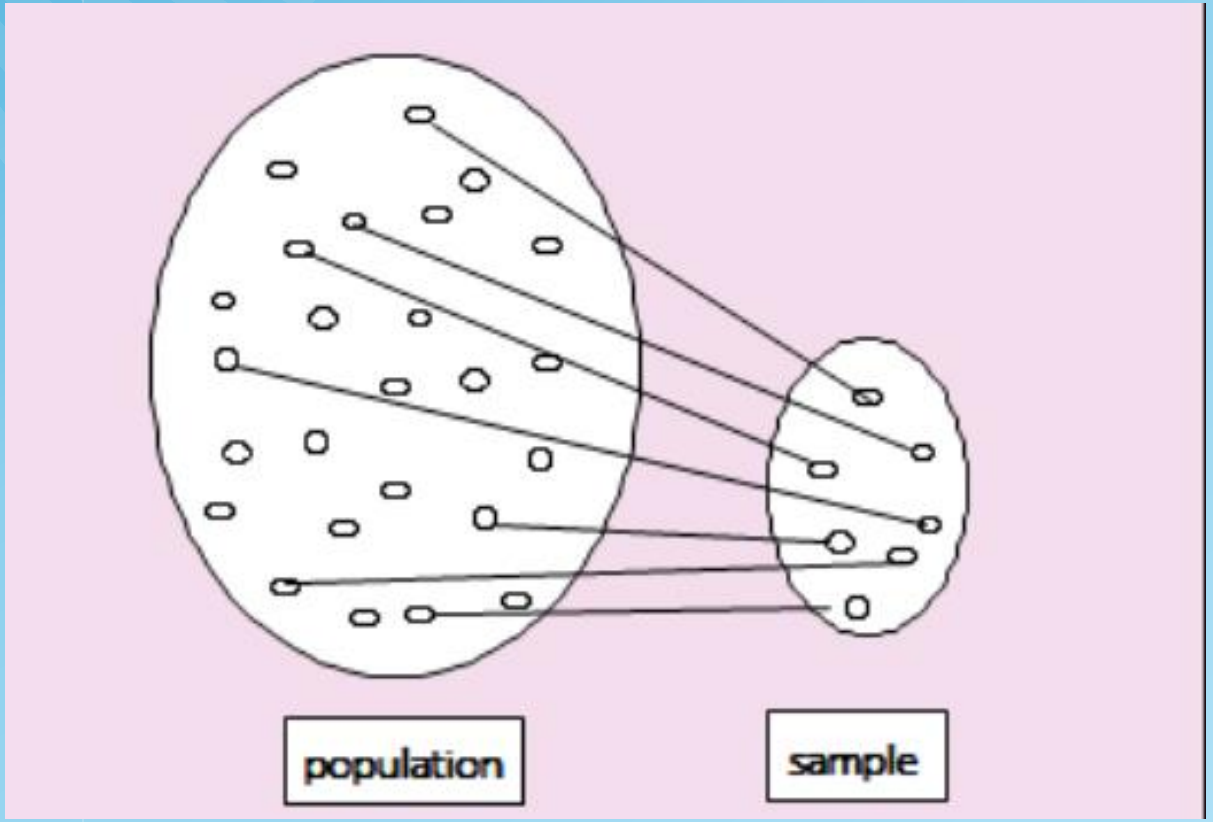
- Car insurance fraud detection

=> Introductory Statistics toward Data Science

1.2 Population and Sample

Population Sample

- **Population** is a whole set of data which we are interested in.
- **Sample** is some number of data extracted from the population.



1.2 Population and Sample

- Descriptive statistics with Excel, eStat
  - data visualization using graphs
  - data summary using frequency table and measures
  - probability, distributions
- Inferential Statistics by using statistical software, eStat
  - sampling distribution
  - estimation
  - testing hypothesis
  - simple linear regression

1.2 Population and Sample

Example 1.2.1 The voting result for the 2016 United States Presidential Election are summarized as the following table. What field of statistics is this?

Candidate	Votes	%	Electral vote
Donald John Trump (Republican)	62,984,828	46.09%	304
Hillary Diane Clinton (Democratic)	65,853,514	48.18%	227
Gary Earl Johnson (Libertarian)	4,489,341	3.28%	0
Jill Ellen Stein (Green)	1,457,218	1.07%	0
David Evan McMullin (Independent)	731,991	0.54%	0
Darrell Lane Castle (Constitution)	203,090	0.15%	0
Gloria Estela La Riva (Socialism)	74,401	0.05%	0

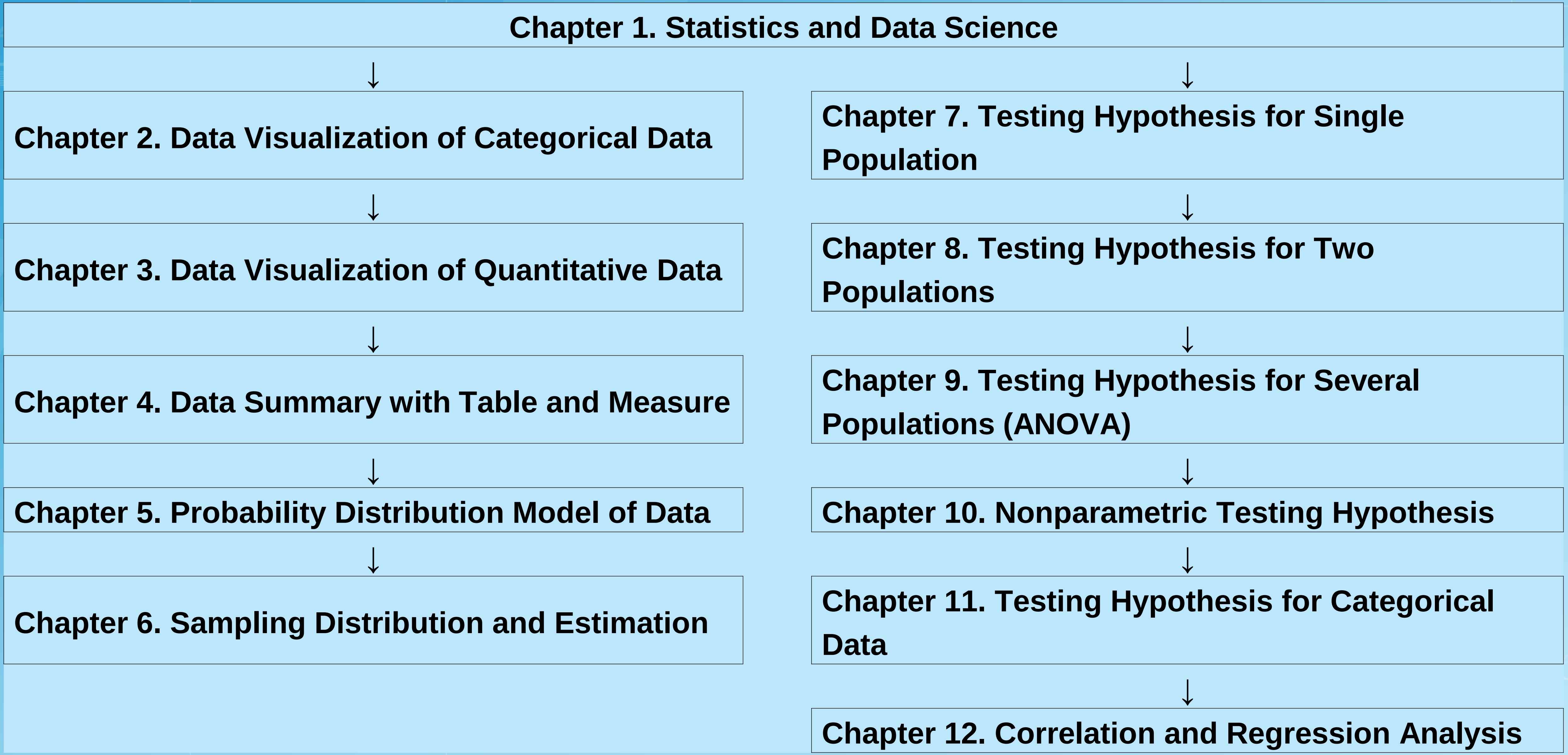
1.2 Population and Sample

Example 1.2.2

The CNN poll was conducted from May 7, 2020 to May 10, 2020 for the 2020 United States Presidential Election by using a sample of 1001 registered voters. The result of poll was as follows. Margin of error = +/-4 percentage points. What field of statistics is this?

Candidate	%
Donald John Trump (Republican)	46%
Joe Biden (Democratic)	51%

1.2 Population and Sample



Chapter 1. Statistics and Data Science

Chapter 2. Data Visualization of Categorical Data

Chapter 3. Data Visualization of Quantitative Data

Chapter 4. Data Summary with Table and Measure

Chapter 5. Probability Distribution Model of Data

Chapter 6. Sampling Distribution and Estimation

Chapter 7. Testing Hypothesis for Single Population

Chapter 8. Testing Hypothesis for Two Populations

Chapter 9. Testing Hypothesis for Several Populations (ANOVA)

Chapter 10. Nonparametric Testing Hypothesis

Chapter 11. Testing Hypothesis for Categorical Data

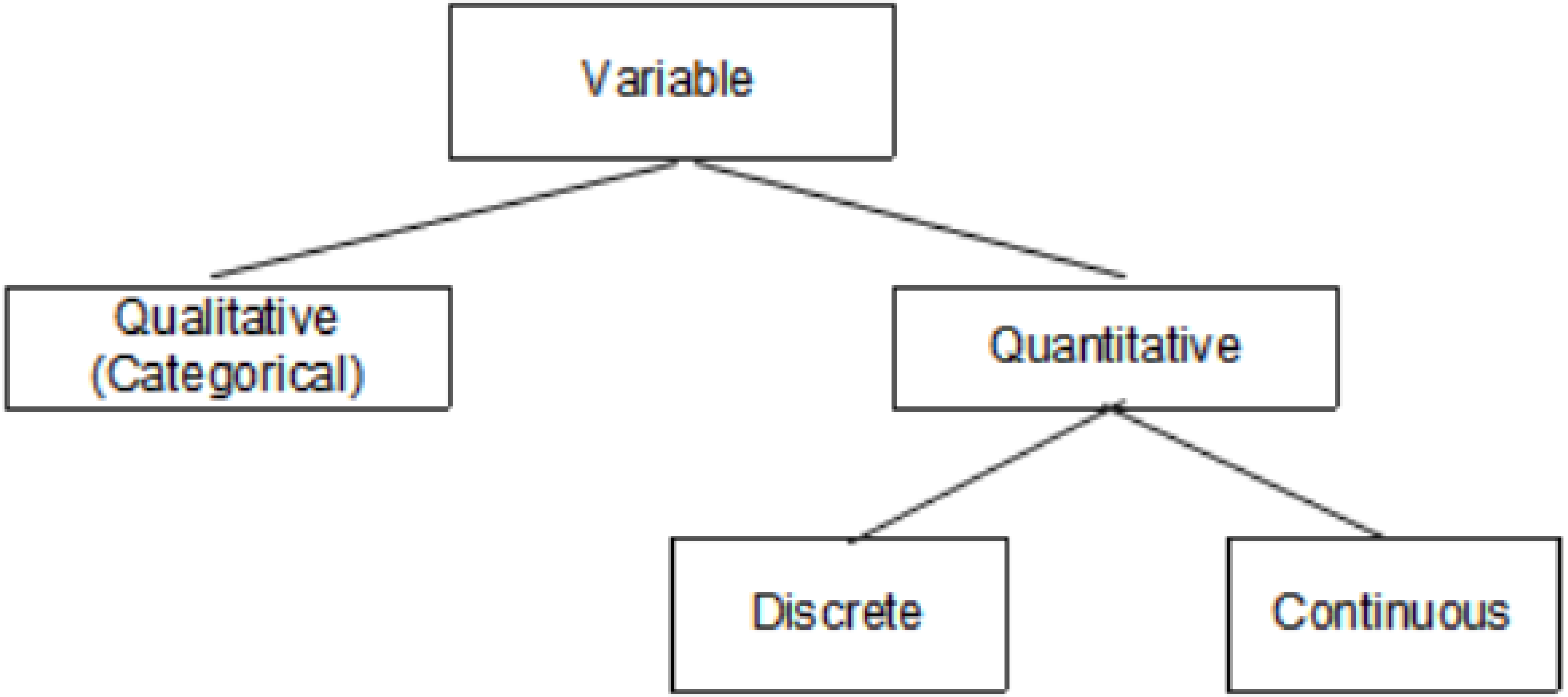
Chapter 12. Correlation and Regression Analysis

1.3 Variables and Data

- Data are values that observe or measure the properties of an object or event that are of interest by a given rule.
- The property of these objects or events is called a variable.
  - if you measured the gender and height of a college student, there are two variables (gender and height).

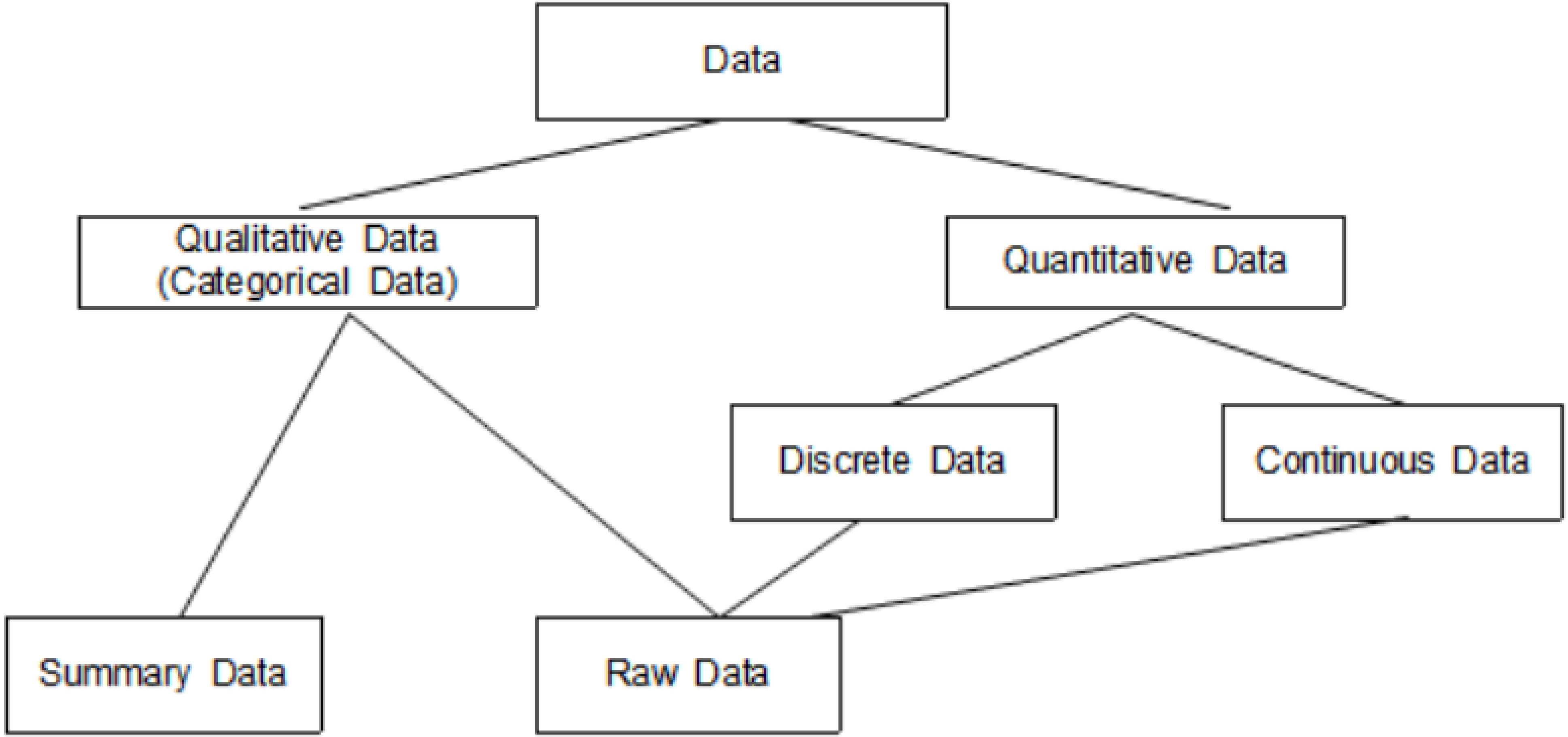
- Qualitative variable: Attributes of gender are 'Male', 'Female',… (Categorical variable)
  => frequency analysis
- Quantitative variable: Height are 180cm, 165cm, 175cm, …
  => Discrete(countable) vs Continuous(uncountable) variable
  => statistical analysis

1.3 Variables and Data



<Figure 1.3.1> Types of variables

1.3 Variables and Data



<Figure 1.3.2> Types of data

1.3 Variables and Data

- Categorical data are classified either raw data or summary data

  - Raw Data of Gender

row	Gender
1	Male
2	Female
3	Male
4	Female
5	Male
6	Male
7	Male
8	Female
9	Female
10	Male

  - Summary Data of Gender

Gender	Students
Male	6
Female	4

- ‘Gender’ variable name
- , ‘Male’ or ‘Female’ variable value

1.4 Software for Statistical Analysis

- Computer software is essential for Statistics & Data Science

  - Elementary: Excel
  - Advanced: statistical packages such as SAS, SPSS, R, Stata
    for advanced user
    no educational module
    expensive except R
    not an web/mobile

1.4 Software for Statistical Analysis

- *eStat* Development Project (2015 ~ 2018)
  - by J.J.Lee and others

- freeware
- web/mobile ready - anytime and anywhere
- easy user interface
- dynamic graphs
- from elementary to university users

1.4 Software for Statistical Analysis - *eStat*

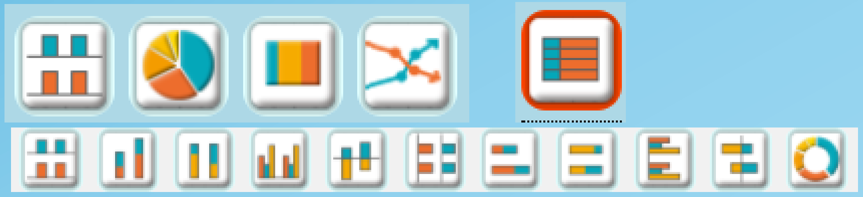
◎ Technology & Manpower for *eStat*

- HTML5, CSS3, JavaScript
- D3.js for dynamic graphs
- Handson table sheet
- Statistical distribution library
  - include nonparametric distributions
- Professors in statistics, statistical computing
  Professors in mathematics education
  Elementary, middle, high school teachers

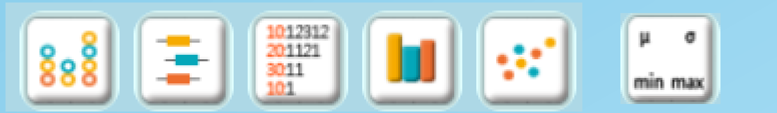
1.4 Software for Statistical Analysis - *eStat*

◎ *eStat* modules

- Elementary School



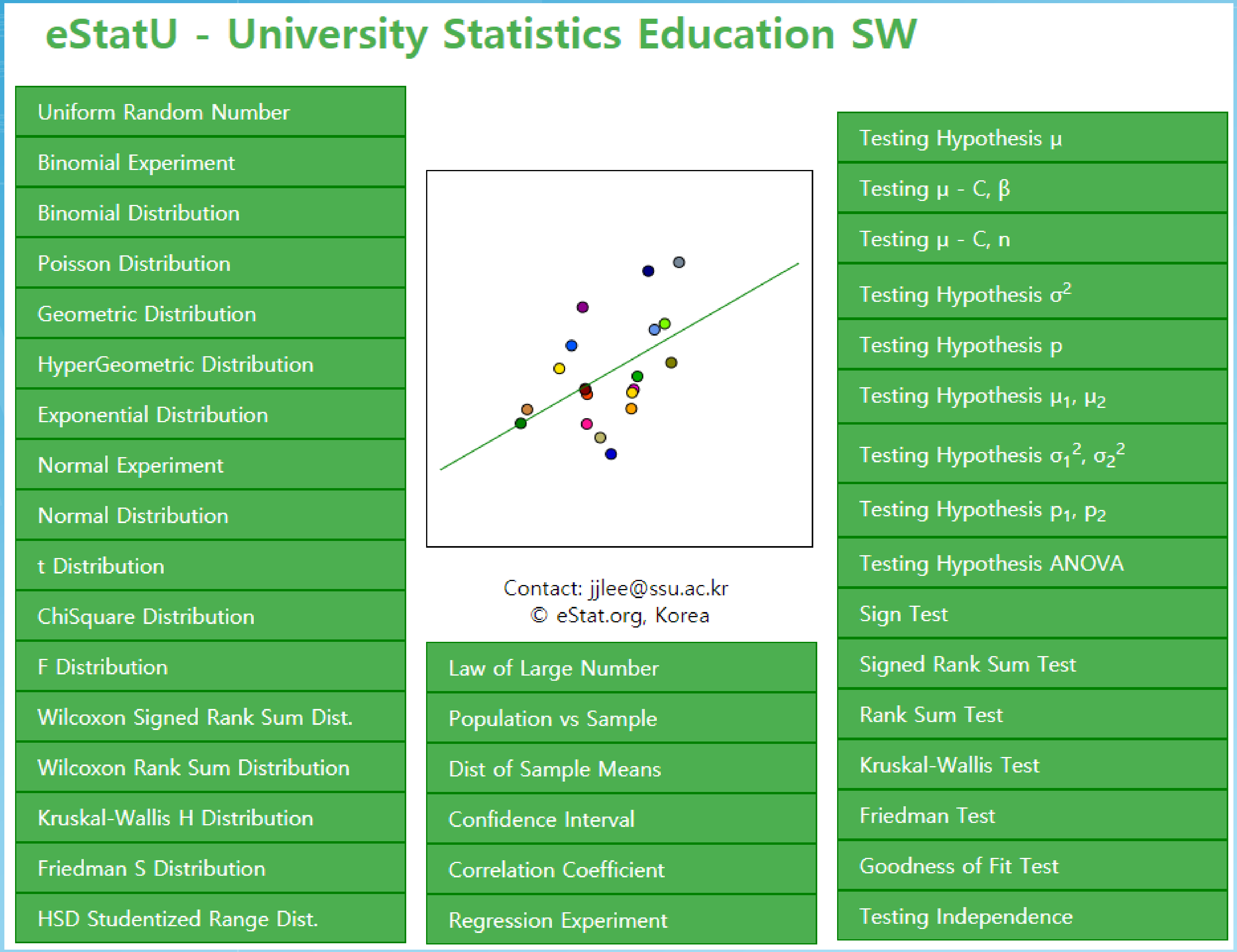
- Middle School



- High School

Binomial, Normal, Sampling Distribution, Law of Large Number, Confidence Interval

1.4 Software for Statistical Analysis - *eStat*



eStatU - University Statistics Education SW

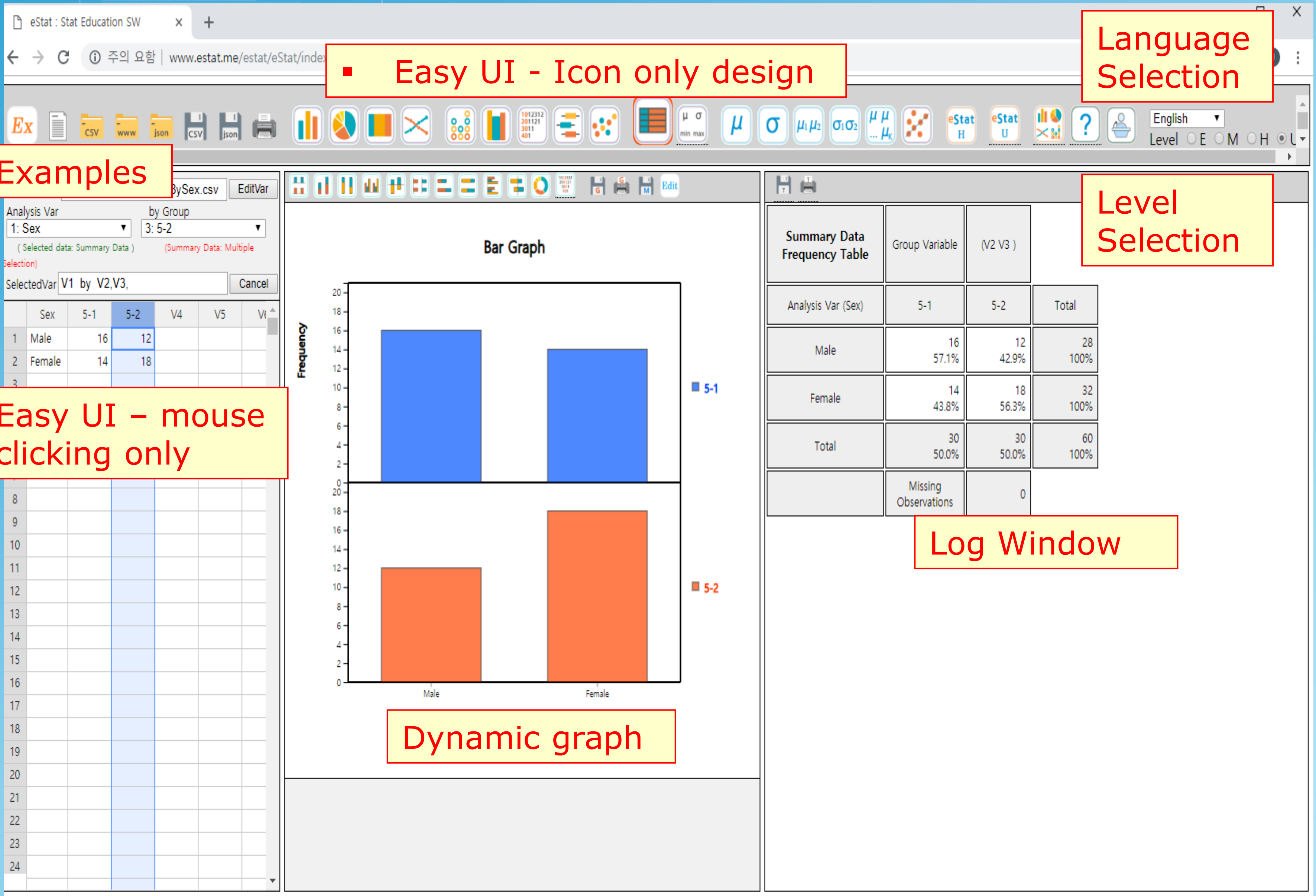
- Uniform Random Number
- Binomial Experiment
- Binomial Distribution
- Poisson Distribution
- Geometric Distribution
- HyperGeometric Distribution
- Exponential Distribution
- Normal Experiment
- Normal Distribution
- t Distribution
- ChiSquare Distribution
- F Distribution
- Wilcoxon Signed Rank Sum Dist.
- Wilcoxon Rank Sum Distribution
- Kruskal-Wallis H Distribution
- Friedman S Distribution
- HSD Studentized Range Dist.

Contact: jjlee@ssu.ac.kr
© eStat.org, Korea

- Law of Large Number
- Population vs Sample
- Dist of Sample Means
- Confidence Interval
- Correlation Coefficient
- Regression Experiment

- Testing Hypothesis μ
- Testing μ - C, β
- Testing μ - C, n
- Testing Hypothesis σ^2
- Testing Hypothesis p
- Testing Hypothesis μ_1, μ_2
- Testing Hypothesis σ_1^2, σ_2^2
- Testing Hypothesis p_1, p_2
- Testing Hypothesis ANOVA
- Sign Test
- Signed Rank Sum Test
- Rank Sum Test
- Kruskal-Wallis Test
- Friedman Test
- Goodness of Fit Test
- Testing Independence

1.4 Software for Statistical Analysis - *eStat*



Analysis Var: 1: Sex

by Group: 3: 5-2

(Selected data: Summary Data) (Summary Data: Multiple Selection)

SelectedVar: V1 by V2,V3,

Cancel

	Sex	5-1	5-2	V4	V5
1	Male	16	12		
2	Female	14	18		

Summary Data Frequency Table	Group Variable	(V2 V3)	
Analysis Var (Sex)	5-1	5-2	Total
Male	16 57.1%	12 42.9%	28 100%
Female	14 43.8%	18 56.3%	32 100%
Total	30 50.0%	30 50.0%	60 100%
	Missing Observations	0	

- Easy UI - Icon only design

Language Selection

Examples

Level Selection

Easy UI – mouse clicking only

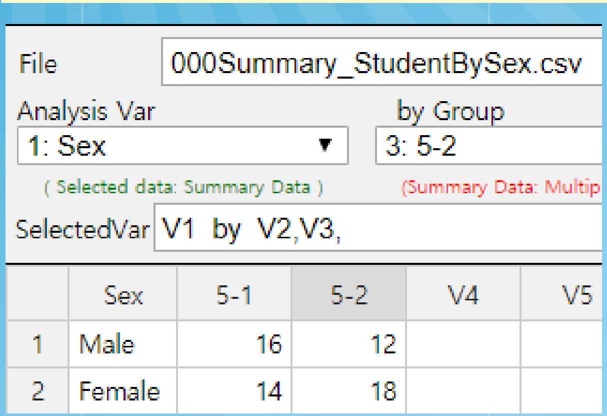
Log Window

Dynamic graph

1.4 Software for Statistical Analysis - *eStat*

◎ Data and Dynamic Graph

- **Support csv and json format**
- **Support summary and raw data for data processing**
- **Dynamic graph**



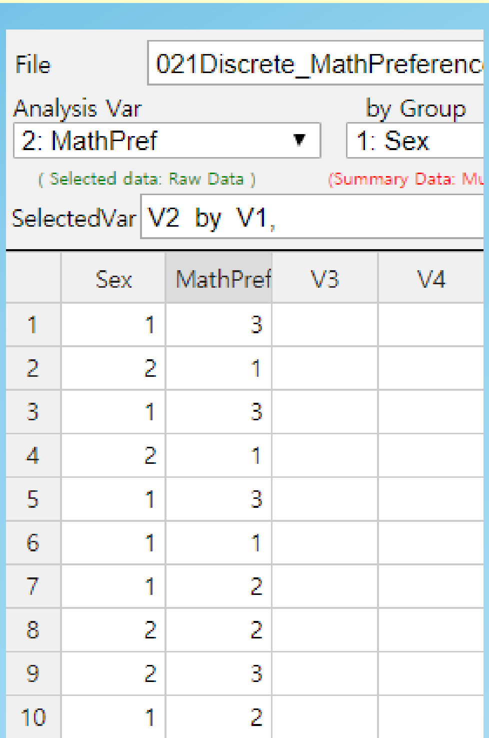
File 000Summary_StudentBySex.csv

Analysis Var 1: Sex ▼ by Group 3: 5-2

(Selected data: Summary Data) (Summary Data: Multip

SelectedVar V1 by V2,V3,

	Sex	5-1	5-2	V4	V5
1	Male	16	12		
2	Female	14	18		



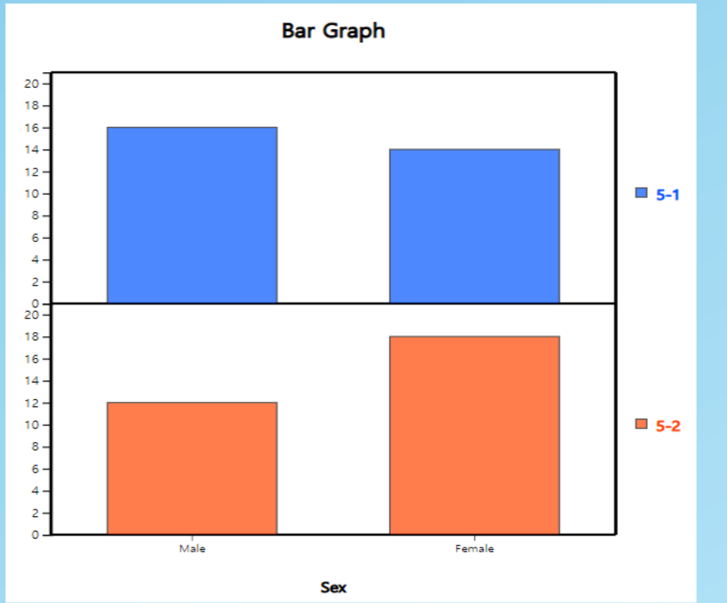
File 021Discrete_MathPreferenc

Analysis Var 2: MathPref ▼ by Group 1: Sex

(Selected data: Raw Data) (Summary Data: Mu

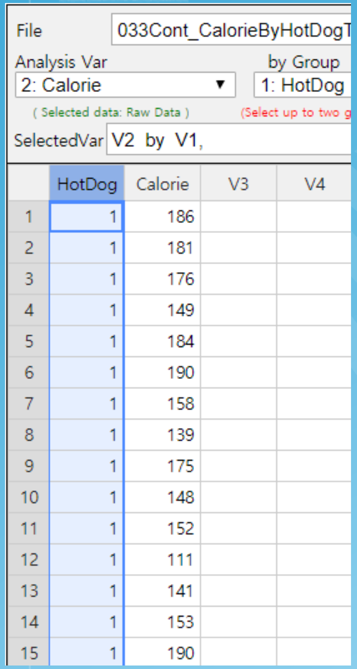
SelectedVar V2 by V1,

	Sex	MathPref	V3	V4
1	1	3		
2	2	1		
3	1	3		
4	2	1		
5	1	3		
6	1	1		
7	1	2		
8	2	2		
9	2	3		
10	1	2		



1.4 Software for Statistical Analysis - *eStat*

◎ Graphical Result of Statistical Analysis - ANOVA



File 033Cont_CalorieByHotDogT

Analysis Var 2: Calorie

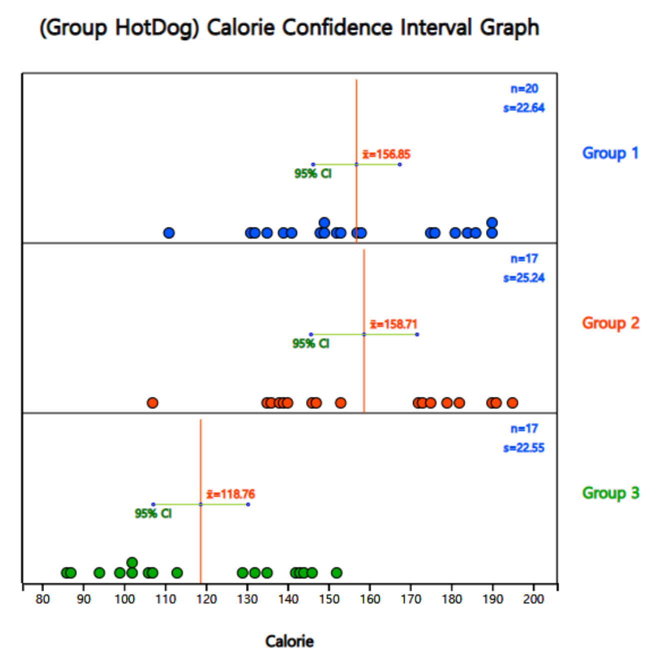
by Group 1: HotDog

(Selected data: Raw Data)

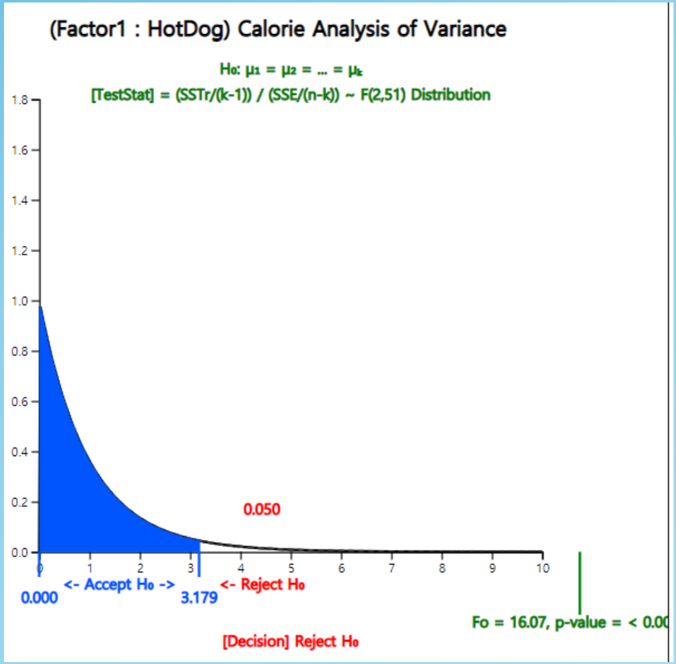
(Select up to two g

SelectedVar V2 by V1,

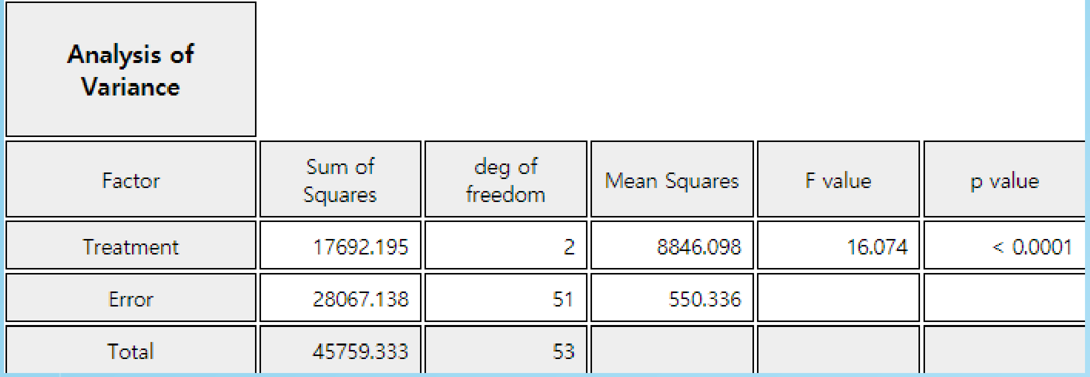
	HotDog	Calorie	V3	V4
1	1	186		
2	1	181		
3	1	176		
4	1	149		
5	1	184		
6	1	190		
7	1	158		
8	1	139		
9	1	175		
10	1	148		
11	1	152		
12	1	111		
13	1	141		
14	1	153		
15	1	190		



(Group HotDog) Calorie Confidence Interval Graph



(Factor1 : HotDog) Calorie Analysis of Variance

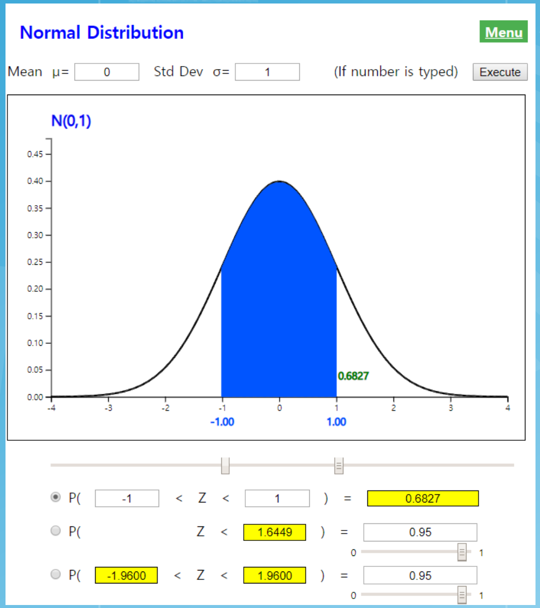


Analysis of Variance

Factor	Sum of Squares	deg of freedom	Mean Squares	F value	p value
Treatment	17692.195	2	8846.098	16.074	< 0.0001
Error	28067.138	51	550.336		
Total	45759.333	53			

1.4 Software for Statistical Analysis - *eStat*

◎ All tables of statistical distributions are on smart-phone



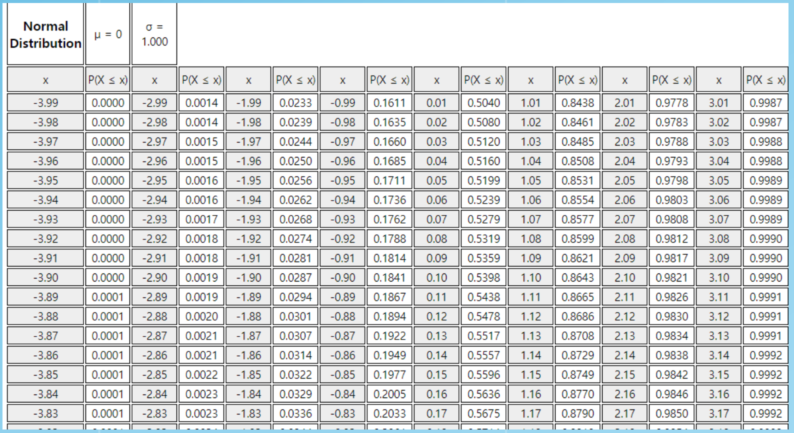
Normal Distribution **Menu**

Mean μ= 0 Std Dev σ= 1 (If number is typed) Execute

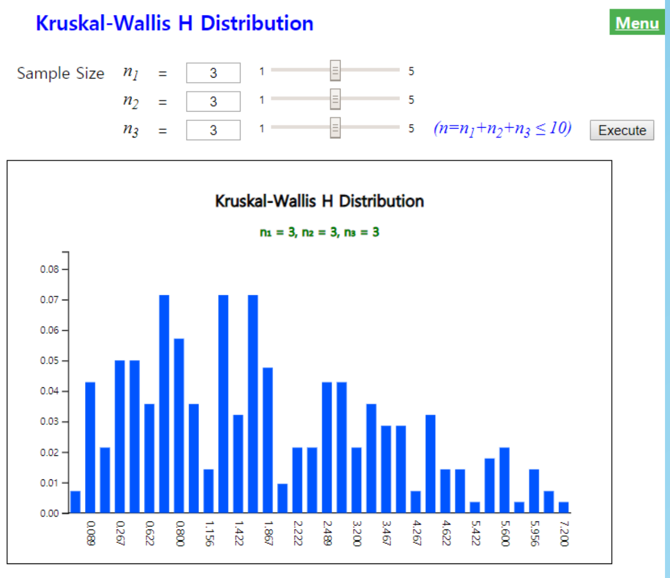
● P(-1 < Z < 1) = 0.6827

○ P(Z < 1.6449) = 0.95

○ P(-1.9600 < Z < 1.9600) = 0.95



Normal Distribution	μ = 0	σ = 1.000											
x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)	x	P(X ≤ x)
-3.99	0.0000	-2.99	0.0014	-1.99	0.0233	-0.99	0.1611	0.01	0.5040	1.01	0.8438	2.01	0.9778	3.01	0.9987
-3.98	0.0000	-2.98	0.0014	-1.98	0.0239	-0.98	0.1635	0.02	0.5080	1.02	0.8461	2.02	0.9783	3.02	0.9987
-3.97	0.0000	-2.97	0.0015	-1.97	0.0244	-0.97	0.1660	0.03	0.5120	1.03	0.8485	2.03	0.9788	3.03	0.9988
-3.96	0.0000	-2.96	0.0015	-1.96	0.0250	-0.96	0.1685	0.04	0.5160	1.04	0.8508	2.04	0.9793	3.04	0.9988
-3.95	0.0000	-2.95	0.0016	-1.95	0.0256	-0.95	0.1711	0.05	0.5199	1.05	0.8531	2.05	0.9798	3.05	0.9989
-3.94	0.0000	-2.94	0.0016	-1.94	0.0262	-0.94	0.1736	0.06	0.5239	1.06	0.8554	2.06	0.9803	3.06	0.9989
-3.93	0.0000	-2.93	0.0017	-1.93	0.0268	-0.93	0.1762	0.07	0.5279	1.07	0.8577	2.07	0.9808	3.07	0.9989
-3.92	0.0000	-2.92	0.0018	-1.92	0.0274	-0.92	0.1788	0.08	0.5319	1.08	0.8599	2.08	0.9812	3.08	0.9990
-3.91	0.0000	-2.91	0.0018	-1.91	0.0281	-0.91	0.1814	0.09	0.5359	1.09	0.8621	2.09	0.9817	3.09	0.9990
-3.90	0.0000	-2.90	0.0019	-1.90	0.0287	-0.90	0.1841	0.10	0.5398	1.10	0.8643	2.10	0.9821	3.10	0.9990
-3.89	0.0001	-2.89	0.0019	-1.89	0.0294	-0.89	0.1867	0.11	0.5438	1.11	0.8665	2.11	0.9826	3.11	0.9991
-3.88	0.0001	-2.88	0.0020	-1.88	0.0301	-0.88	0.1894	0.12	0.5478	1.12	0.8686	2.12	0.9830	3.12	0.9991
-3.87	0.0001	-2.87	0.0021	-1.87	0.0307	-0.87	0.1922	0.13	0.5517	1.13	0.8708	2.13	0.9834	3.13	0.9991
-3.86	0.0001	-2.86	0.0021	-1.86	0.0314	-0.86	0.1949	0.14	0.5557	1.14	0.8729	2.14	0.9838	3.14	0.9992
-3.85	0.0001	-2.85	0.0022	-1.85	0.0322	-0.85	0.1977	0.15	0.5596	1.15	0.8749	2.15	0.9842	3.15	0.9992
-3.84	0.0001	-2.84	0.0023	-1.84	0.0329	-0.84	0.2005	0.16	0.5636	1.16	0.8770	2.16	0.9846	3.16	0.9992
-3.83	0.0001	-2.83	0.0023	-1.83	0.0336	-0.83	0.2033	0.17	0.5675	1.17	0.8790	2.17	0.9850	3.17	0.9992

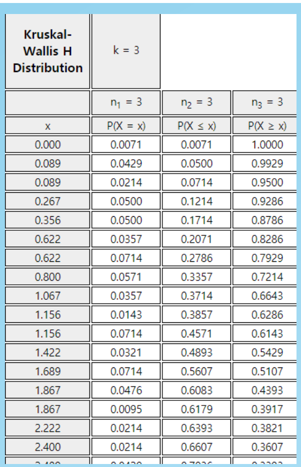


Kruskal-Wallis H Distribution **Menu**

Sample Size n_1 = 3

n_2 = 3

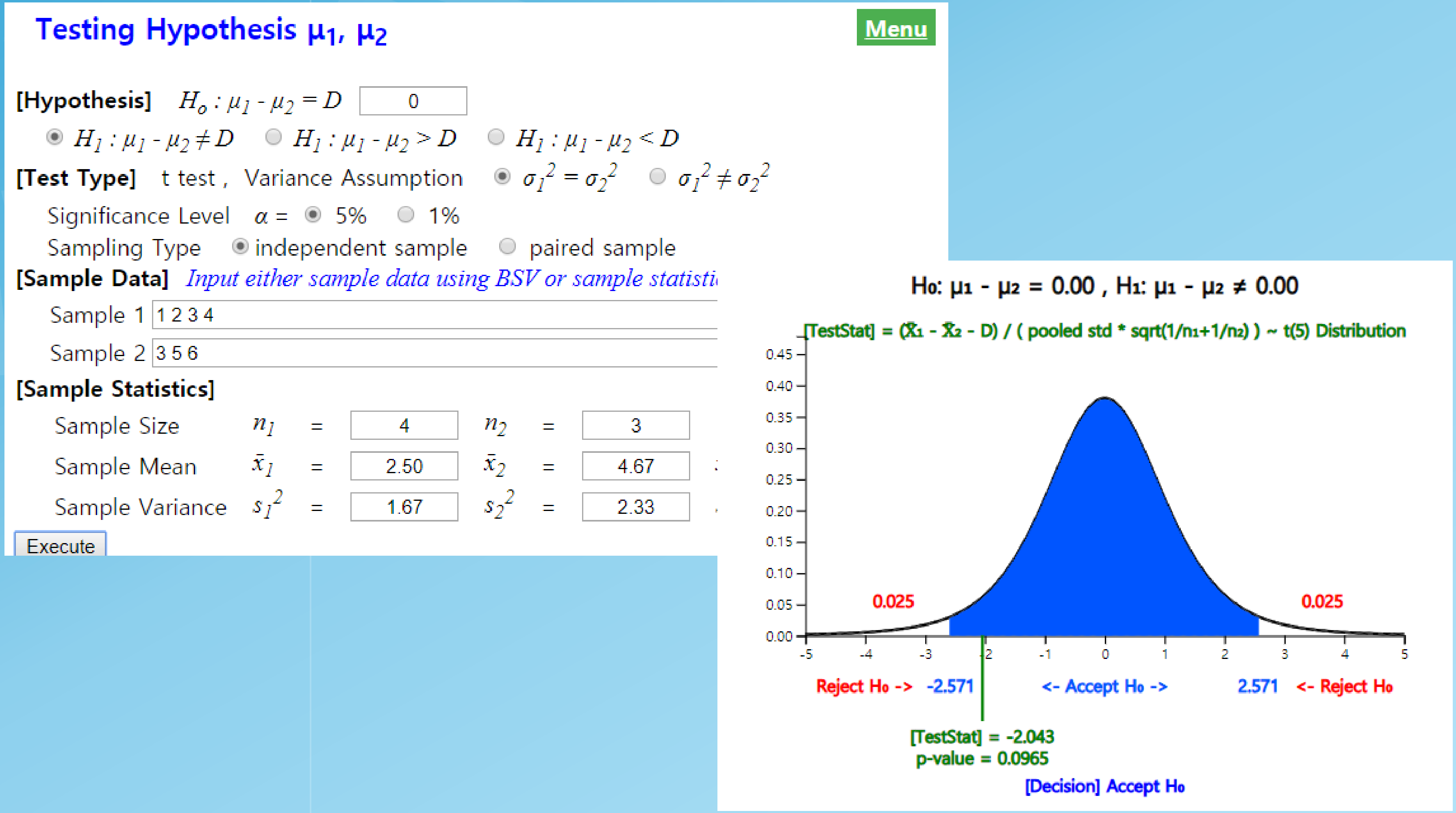
n_3 = 3 $(n = n_1 + n_2 + n_3 \le 10)$ Execute



Kruskal-Wallis H Distribution	k = 3		
	n_1 = 3	n_2 = 3	n_3 = 3
x	P(X = x)	P(X ≤ x)	P(X ≥ x)
0.000	0.0071	0.0071	1.0000
0.089	0.0429	0.0500	0.9929
0.089	0.0214	0.0714	0.9500
0.267	0.0500	0.1214	0.9286
0.356	0.0500	0.1714	0.8786
0.622	0.0357	0.2071	0.8286
0.622	0.0714	0.2786	0.7929
0.800	0.0571	0.3357	0.7214
1.067	0.0357	0.3714	0.6643
1.156	0.0143	0.3857	0.6286
1.156	0.0714	0.4571	0.6143
1.422	0.0321	0.4893	0.5429
1.689	0.0714	0.5607	0.5107
1.867	0.0476	0.6083	0.4393
1.867	0.0095	0.6179	0.3917
2.222	0.0214	0.6393	0.3821
2.400	0.0214	0.6607	0.3607

1.4 Software for Statistical Analysis - *eStat*

◎ Modules for Home Work Assignment - eStatU



Testing Hypothesis μ1, μ2

[Hypothesis] $H_o : \mu_1 - \mu_2 = D$ 0

● $H_1 : \mu_1 - \mu_2 \neq D$ ○ $H_1 : \mu_1 - \mu_2 > D$ ○ $H_1 : \mu_1 - \mu_2 < D$

[Test Type] t test , Variance Assumption ● $\sigma_1^2 = \sigma_2^2$ ○ $\sigma_1^2 \neq \sigma_2^2$

Significance Level α = ● 5% ○ 1%

Sampling Type ● independent sample ○ paired sample

[Sample Data] *Input either sample data using BSV or sample statisti*

Sample 1 1 2 3 4

Sample 2 3 5 6

[Sample Statistics]

Sample Size	n_1	=	4	n_2	=	3
Sample Mean	$\bar{x}_1$	=	2.50	$\bar{x}_2$	=	4.67
Sample Variance	s_1^2	=	1.67	s_2^2	=	2.33

H₀: μ₁ - μ₂ = 0.00 , H₁: μ₁ - μ₂ ≠ 0.00

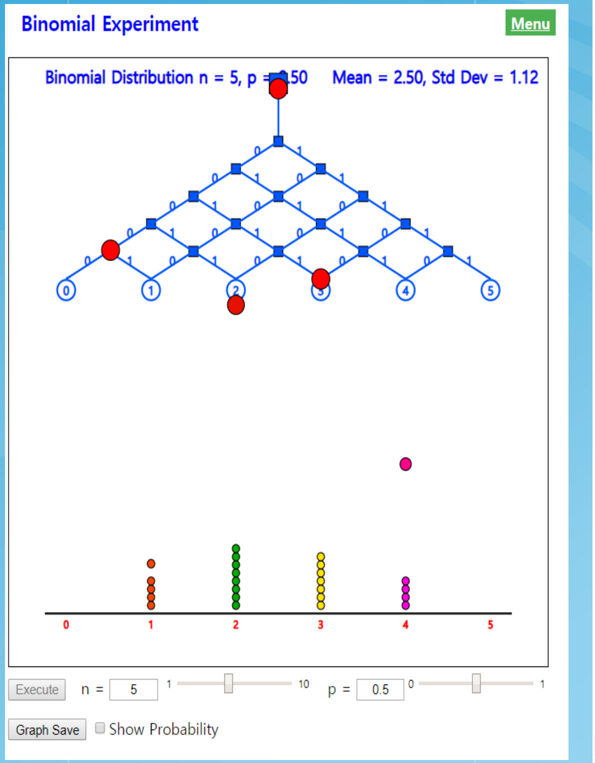
[TestStat] = (X̄₁ - X̄₂ - D) / (pooled std * sqrt(1/n₁+1/n₂)) ~ t(5) Distribution

[TestStat] = -2.043
p-value = 0.0965

[Decision] Accept H₀

1.4 Software for Statistical Analysis - *eStat*

◎ Simulation Experiments

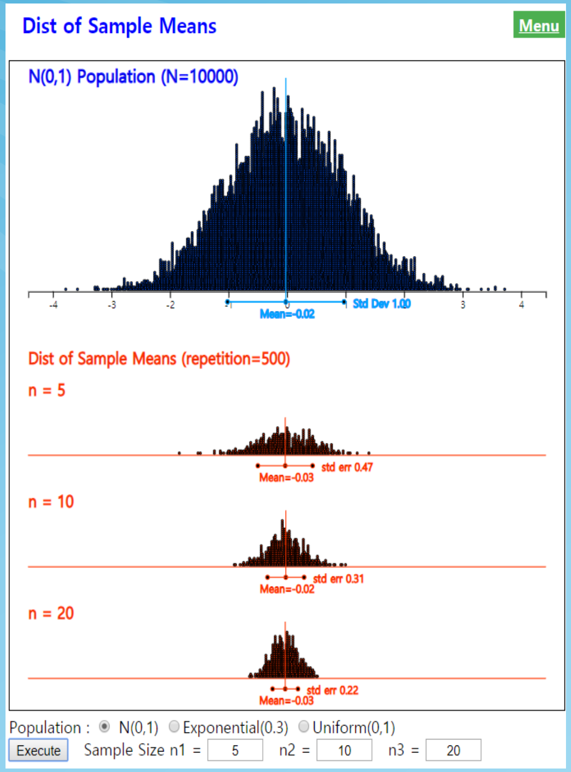


Binomial Experiment

Binomial Distribution n = 5, p = 0.50 Mean = 2.50, Std Dev = 1.12

Execute n = 5 p = 0.5

Graph Save □ Show Probability



Dist of Sample Means

N(0,1) Population (N=10000)

Mean=-0.02 Std Dev 1.00

Dist of Sample Means (repetition=500)

n = 5

Mean=-0.03 std err 0.47

n = 10

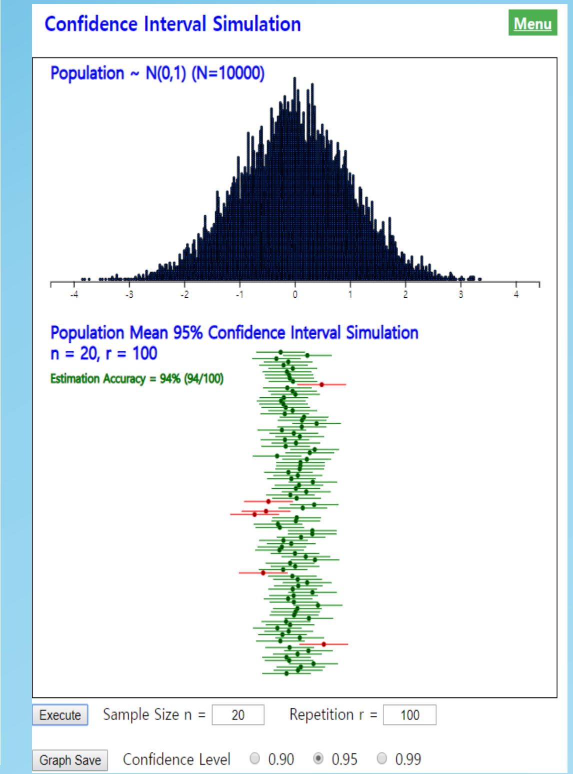
Mean=-0.02 std err 0.31

n = 20

Mean=-0.03 std err 0.22

Population : ● N(0,1) ○ Exponential(0.3) ○ Uniform(0,1)

Execute Sample Size n1 = 5 n2 = 10 n3 = 20



Confidence Interval Simulation

Population ~ N(0,1) (N=10000)

Population Mean 95% Confidence Interval Simulation

n = 20, r = 100

Estimation Accuracy = 94% (94/100)

Execute Sample Size n = 20 Repetition r = 100

Graph Save Confidence Level ○ 0.90 ● 0.95 ○ 0.99

1.4 Software for Statistical Analysis - *eStat*

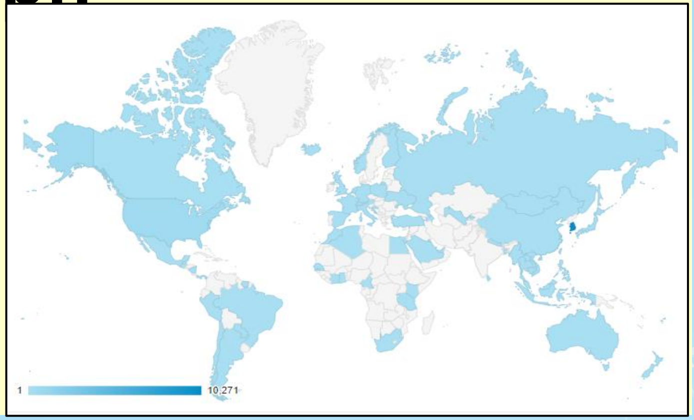
http://www.estat.me

eStat works 100% with Chrome

1) Enter system
2) Data input/save/open
3) Draw graph and data analysis
4) Save results / print results
5) Log out the system
6) Educational modules
7) Others

1.5 Summary

- Statistics : long history to manage an organization
- Introductory Statistics toward Data Science Using *eStat*
  - data visualization, data summary
  - probability, distribution function, estimation
  - testing hypothesis, regression
- *eStat* is an educational statistical software
  - web/mobile based, easy UI, dynamic graph
  - from elementary school to university
  - freeware, multilingual
- *eStat* is widely used in the world
  - USCOTS, JINSE(Japan) recommended



Çox sağ ol !