

Introduction to Biostatistics and R

[Eliana Ibrahimi](#)

Biostatistician at the Department of Biology
University of Tirana, Albania

eliana.ibrahimi@fshn.edu.al

Online Workshop
Hands-on Biostatistics
July, 2020

Overview

- What is Biostatistics and why we need it?
- Variables in biostatistics
- Data analysis steps
- Inference statistics
 - Estimation
 - Hypotheses testing

How knowledgeable are you in biostatistics and R?

- Check the link in the chat and let me know by answering the questions.

<https://www.wooclap.com/HOB3>



What is Biostatistics?

Definition

- Biostatistics is a branch of applied statistics that applies statistical methods to collection, analyze, and interpret data related to biology, health, and medicine.

--- *Much more statistics than biology, however biostatisticians must learn the biology also.*



Why do I need to learn biostatistics?

Three main reasons

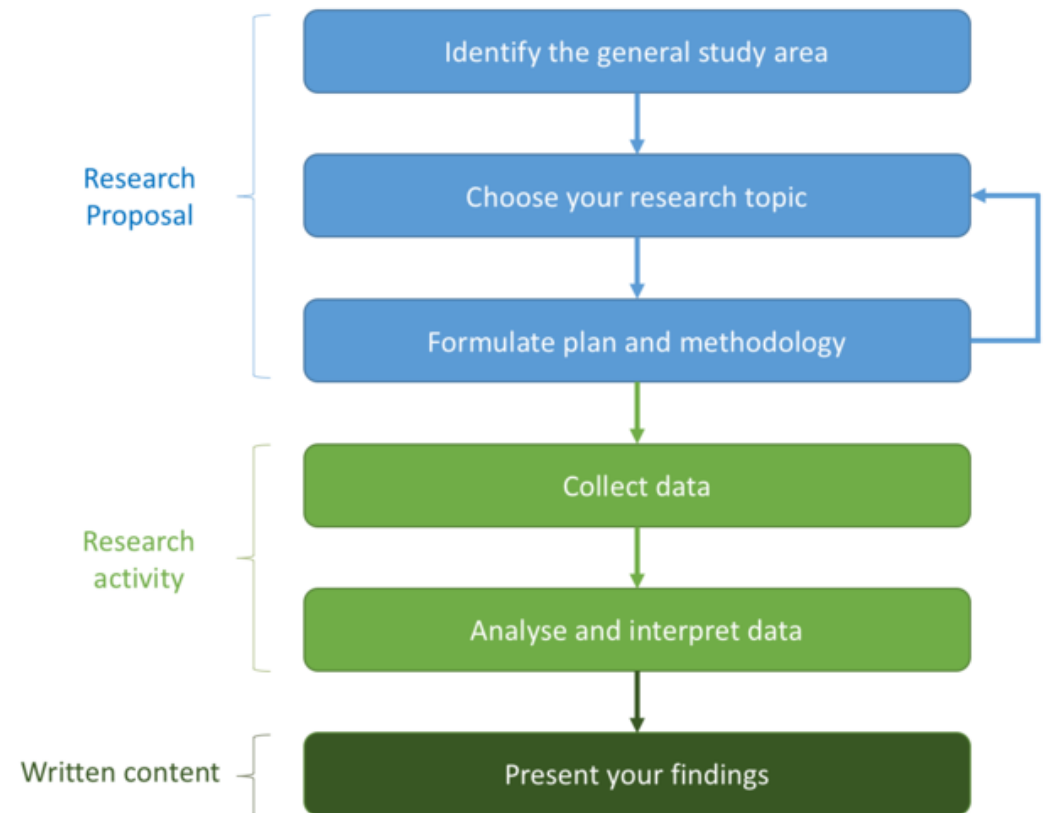
1. To effectively conduct research
2. To be able to read journals
3. To further develop critical and analytical thinking

Biostatistics Ranked #1

Forbes recently named Biostatistics as the #1 best Master's degree for jobs

What is the role of biostatistics in research?

- A good way to learn about biostatistics and its role in the research process is to follow a research study from study design to its publication.
- The biostatistician should be present in each step.



Variables in biostatistics

- A random variable is a characteristic that can take on different values for different individuals, places or things.

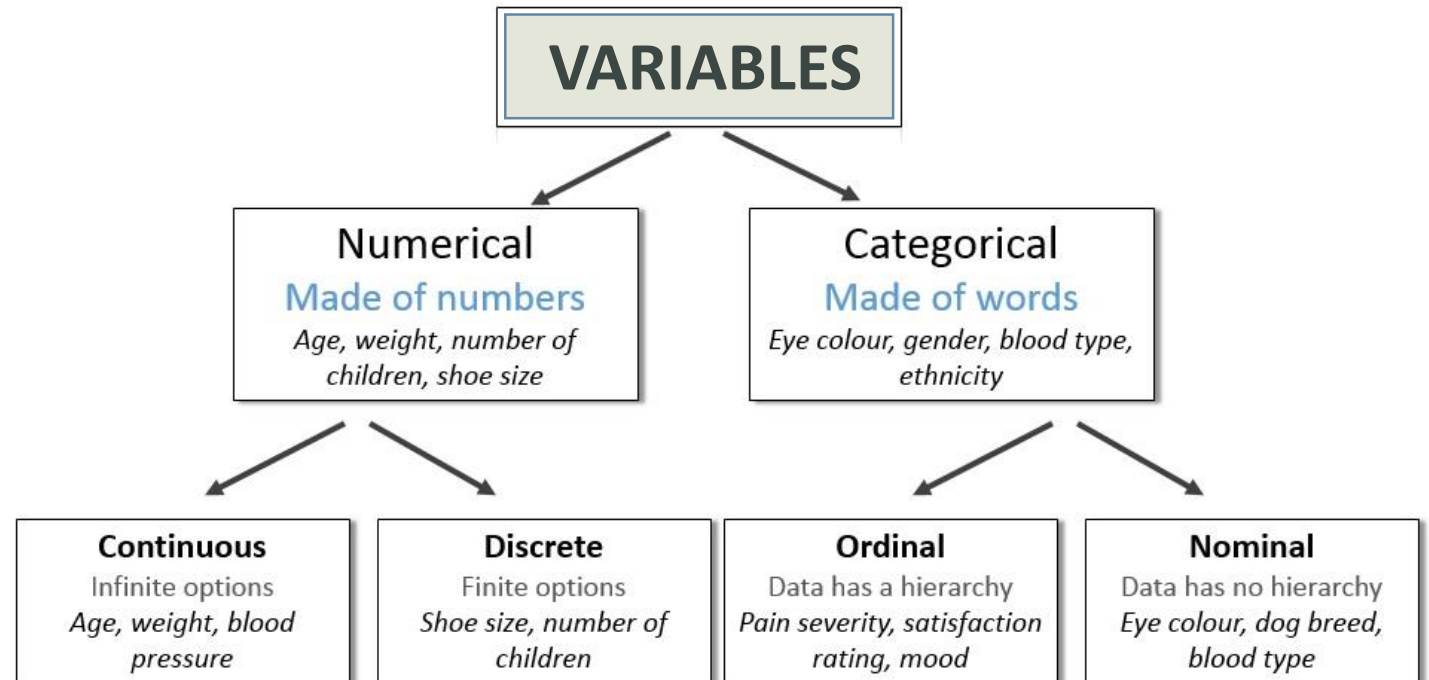


Image source: gauraw.tech

Data analysis

- According to Wikipedia, Data analysis is a process of inspecting, cleansing, transforming and modeling data to discover useful information and supporting decision-making.



Image source: vippng.com

- ✓ There are 5 steps in the process of data analysis

1. Data preparation (organize, transform, clean)
2. Data exploration
3. Data modeling
4. Draw conclusions
5. Communicating results

Data preparation

The collected data is generally not in the form to be analyzed directly.

- Data preparation includes editing, coding, data entry and is the process that ensures data accuracy and their transformation from raw to reduced and classified forms that are appropriate for analysis.
- Use excel to edit, code and tabulate your data if you are not expert in R or another programming software.
- Check this [material](#) for details.

Data preparation: Raw data

- An example of raw data

RAW DATA NP I		Number of colonies					
Treatment	Concentration	dish 1	dish 2	dish 3	dish 4	dish 5	dish 6
Positive Control	100μM	0	0	0	0	0	0
Control	0	122	132	120	134	123	154
Solvent Control	0.04%	152	139	132	118	148	142
I	1	145	134	144	149	138	129
I	5	137	133	143	155	141	135
I	10	129	124	135	138	146	143
I	12.5	146	113	131	138	130	145
I	15	72	75	75	82	96	101
I	20	55	28	17	77	41	10
I	25	0	0	0	0	0	0

Data preparation: Well organized data

- An example of well organized data

Month	Name	Gender	Diagnosis	Treatment
May	Jessica	F	Allergy	Eye Drops
May	Sam	M	Allergy	Eye Drops
May	Wes	M	Cataract	Cataract Surgery
May	Rachel	F	Pterygium	Eye Drops
May	Lily	F	Allergy	Eye Drops
May	Hannah	F	Cataract	Cataract Surgery
May	Denise	F	Allergy	Eye Drops
May	Sharon	F	Allergy	Eye Drops
May	Robin	F	Allergy	Eye Drops
May	Lianna	F	Pterygium	Eye Drops
May	Thomas	M	Presbyopia	Reading Glasses
May	Kimberly	F	Refractive Error	Distance Glasses
May	Michael	M	Refractive Error	Distance Glasses
May	Jacob	M	Conjunctivitis	Eye Drops
June	John	M	Presbyopia	Reading Glasses
June	Tim	M	Refractive Error	Distance Glasses
June	Allison	F	Cataract	Cataract Surgery
June	Laura	F	Pterygium	Eye Drops
June	Scott	M	Cataract	Cataract Surgery
June	Sarah	F	Pterygium	Eye Drops
June	Alex	M	Pterygium	Eye Drops
June	Robert	M	Cataract	Cataract Surgery

Data exploration

Nominal variables

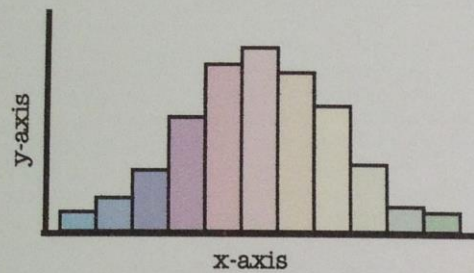
- Frequency
 - Count (How frequent values occur)
 - Relative (The % of observations with a specific value)
- Graphs (simple and clustered)
 - Bar
 - Pie
 - Area
 - ...

Numerical variables

- Descriptive statistics
 - Mean, Mode, Median
 - Variance, Standard deviation, Standard error
 - Range, Percentiles
- Graphs (simple and clustered)
 - Histogram
 - Boxplot
 - Error bars
 - Q&Q plots
 - Scatterplot
 - ...

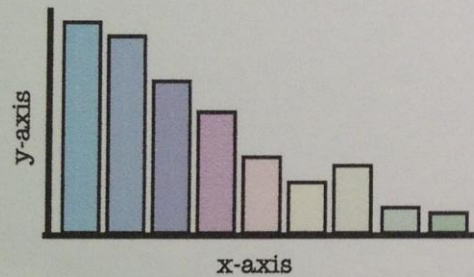
Data exploration: Graphs

Histogram:



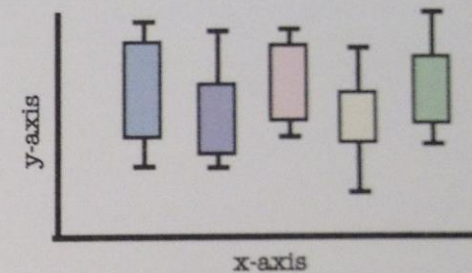
Shows distribution of values.

Bar Chart:



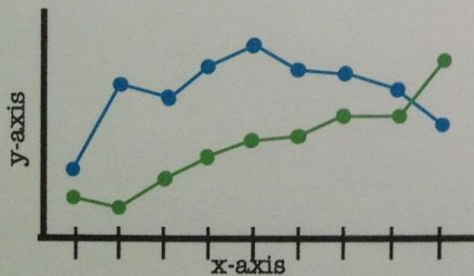
Compares categories.

Box Plot:



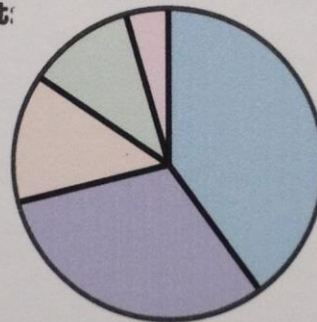
Shows range, median, and standard deviation of each category.

Line Chart:



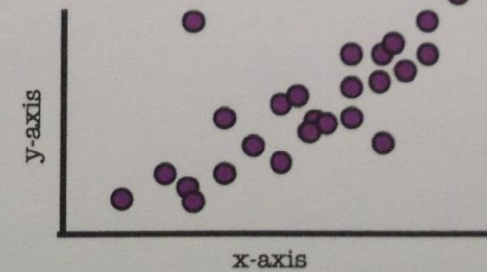
Shows trends, usually over time.

Pie Chart:



Compares percentages of a whole.

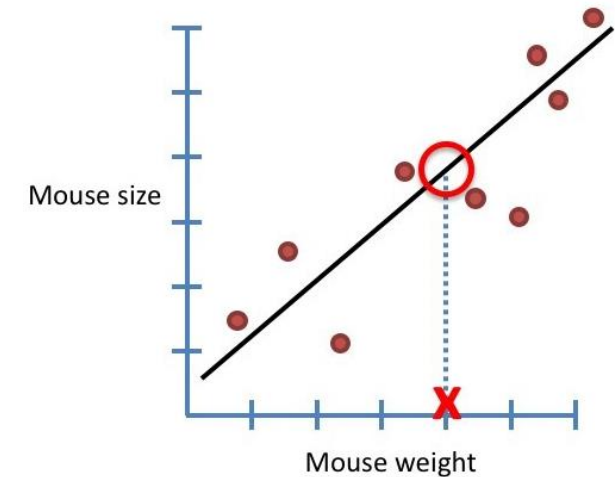
Scatterplot:



Uses large collections of data to find correlations.

Statistical modeling

- A statistical model is usually specified as a mathematical relationship between one or more random variables and other non-random variables. As such, a statistical model is "a formal representation of a theory" and represents, often in considerably idealized form, the data-generating process (Wikipedia).
- Includes
 - T-tests
 - ANOVA
 - Linear Regression
 - General linear regression
 - And many more...



https://www.youtube.com/watch?v=yQhTtdq_y9M

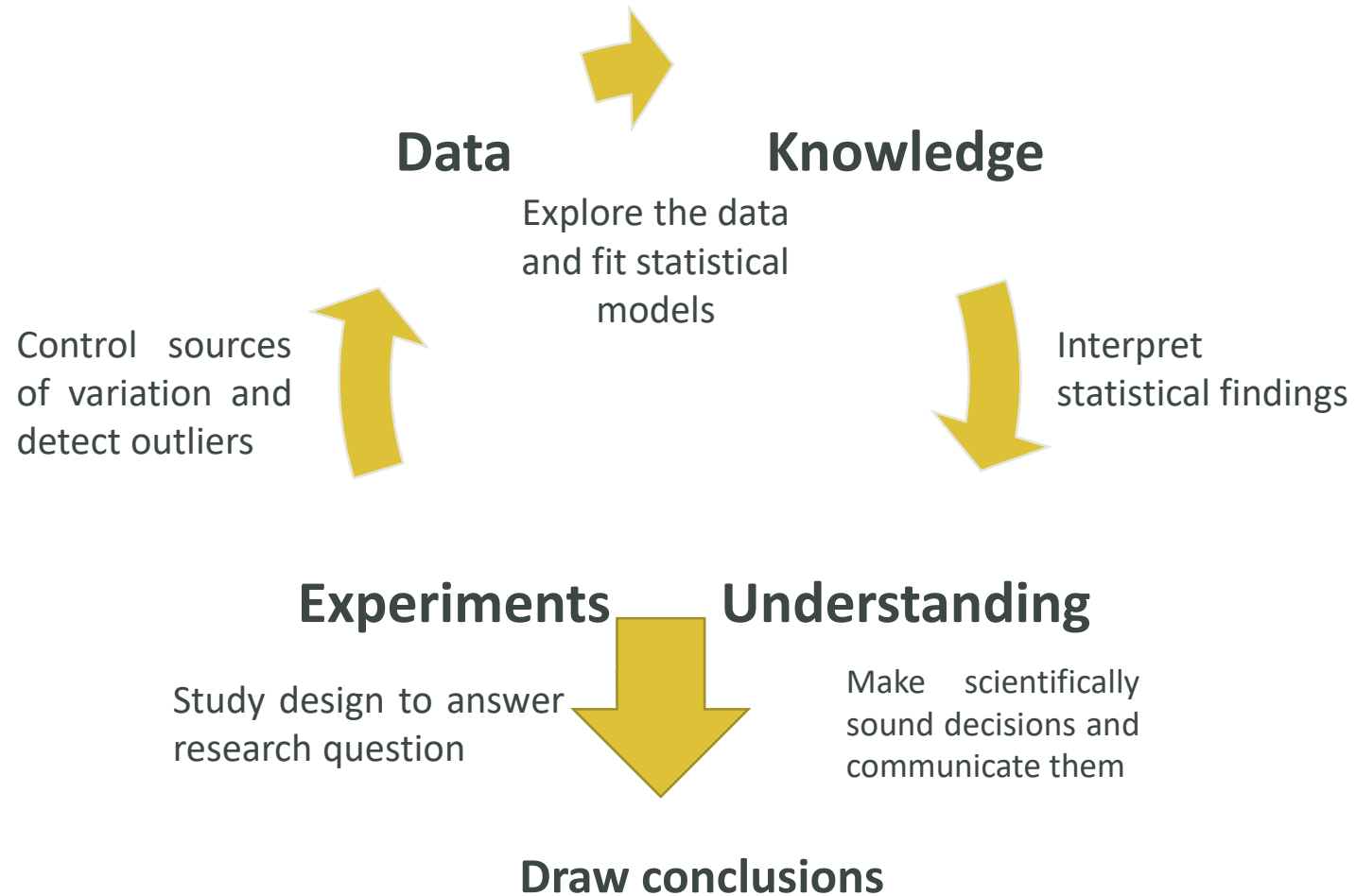
Draw conclusions

- While interpreting the results ask yourself:
 - Did the analysis answer my research question?
 - Was there any limitation in my analysis which would affect my conclusions?
 - Was the analysis sufficient enough to help decision making?

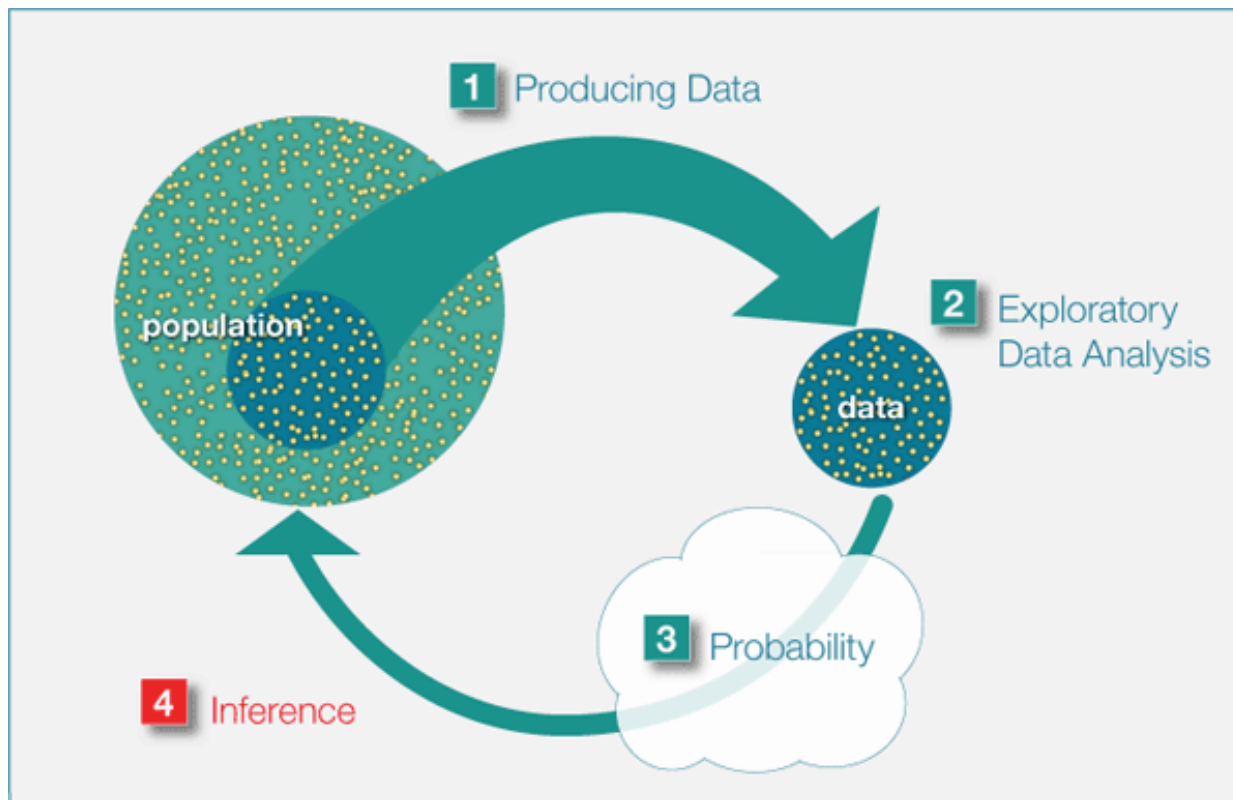
Communicate the results

- Now it's time to communicate your findings
 - Documentation of statistics
 - Making presentations
 - Writing reports, or blogs
 - Writing manuscripts
- Extra skills needed at this stage
 - Writing
 - Presenting
 - Communication

Let's summarize



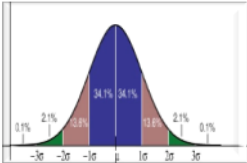
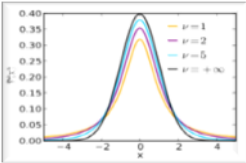
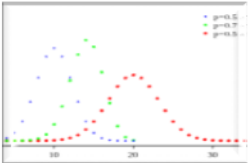
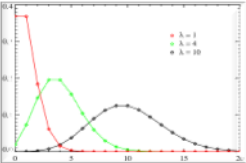
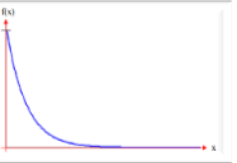
Statistical inference



- Statistical inference: draw conclusions about a population based on the data obtained from a sample chosen from it.

- Point Estimation
- Interval estimation
- Hypothesis testing

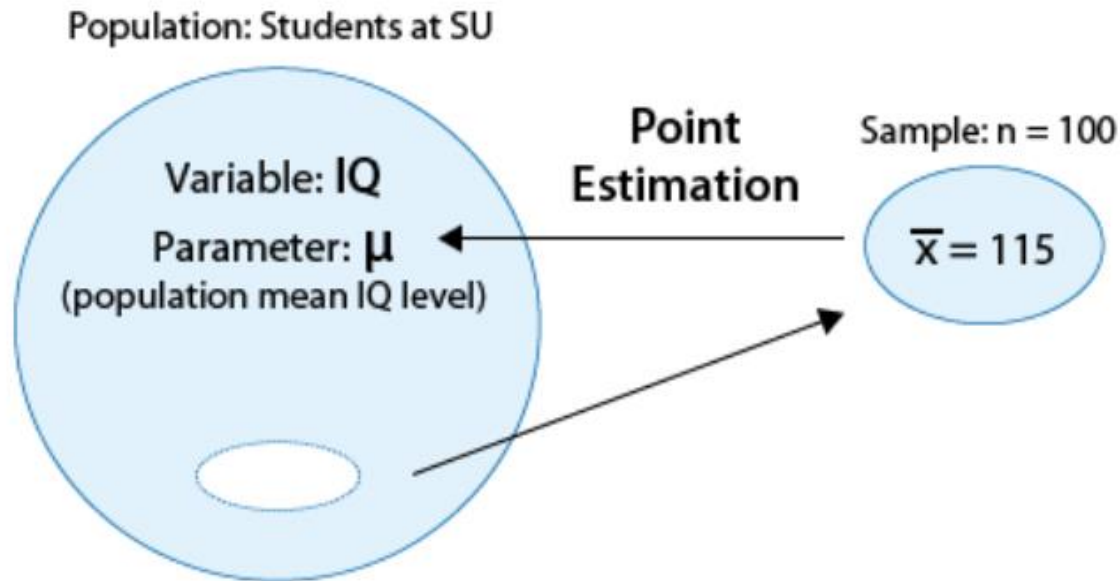
Statistical inference is based on probability distributions

	Normal Distribution	Student's T-Distribution	Binomial Distribution	Poisson Distribution	Exponential Distribution
What does it look like?					
Defining Characteristics	Distinctive Bell Shape	Shorter, fatter than the normal distribution.	Two outcomes: Success/Failure	Various shapes, but valid only for integers on the x-axis.	Models Time Between Events
Example of When to Use It	Modeling natural phenomena (height, weight, IQ, test scores etc.)	When you have small samples or don't know the population variance (σ^2).	Coin Toss Probability (Heads, Tails)	Gives probability of number of events in a fixed interval.	"How much time will go by before a major hurricane hits the Atlantic Seaboard?"
Example of DS Application	Least squares fitting or propagation of uncertainty.	Unknown σ^2 is common in real life data, you you'll have to use the T instead of the normal in that case.	Anywhere where binary (yes/no, black/white, vote/don't vote) data is used.	Anywhere there is a waiting time between events.	Building continuous-time Markov chains.

Source: Data Science Central

- To fully understand the theory behind statistical inference you will need some concepts related to probability distributions.
- We will not focus on this, but:
 - If you haven't taken any course in statistical theory during your studies, I recommend you check this great [book](#) by Rosner (2010).

Point Estimation



- In **point estimation**, we estimate an unknown parameter using a single number that is calculated from the sample data.
- Point estimates are totally unbiased estimates for the population parameter only if the sample is random and the study design is not flawed.

Interval estimation

- In **interval estimation**, we estimate an unknown parameter using an interval of values that is likely to contain the true value of that parameter (and how confident we are that this interval captures the true value of the parameter).
- For example: We are 95% confident that μ for IQ in the previous sample is covered by the interval (112, 118).
 - How we got these numbers? We'll see that later in R. For now make sure you understand the idea.

The diagram illustrates the formula for interval estimation: $\bar{x} \pm T_c \cdot s/\sqrt{n}$. Annotations include:

- Sample Mean and center of interval**: Points to \bar{x} .
- Critical T-value (depends on confidence level)**: Points to T_c .
- Standard Error**: Points to s/\sqrt{n} .
- Margin of Error**: A bracket under $T_c \cdot s/\sqrt{n}$ points to this label.

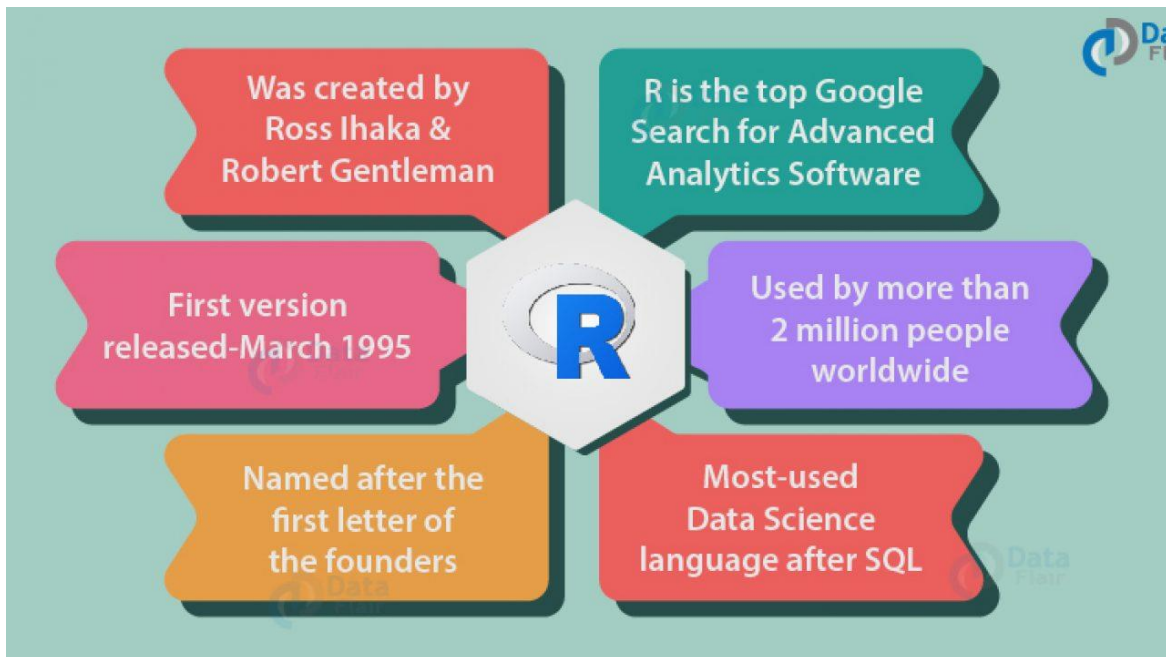
Available statistical software?



The battle is ...



Why R?



<https://data-flair.training/blogs/why-learn-r/>

References/Useful links

1. Rosner, Bernard. Fundamentals Of Biostatistics. Cengage Learning, 2011.
2. Pezzullo, John. Biostatistics For Dummies. Wiley, 2013.
3. <https://bolt.mph.ufl.edu/6050-6052>
4. <http://www.biostat handbook.com/HandbookBioStatThird.pdf>
5. <https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/estimating-a-population-mean-3-of-3/>
6. <https://data-flair.training/blogs/why-learn-r/>