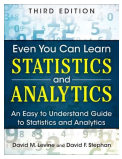
IMGD 2905

Fundamentals of Statistics

Chapter 1



Why Do We Need Statistics?

445 446 397 226
388 3445 188 1002
47762 432 54 12
98 345 2245 8839
77492 472 565 999
1 34 882 545 4022
827 572 597 364

Aggregate data into meaningful information.

$\bar{x} = \ldots$

Ok, but what *are* statistics?

Key Words

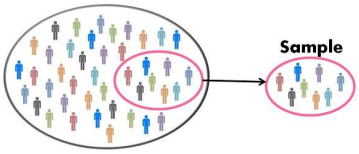
- Population – all members of group pertaining to study
  - e.g., every League of Legends player in the world
  - e.g., every person in IMGD 2905 in D-term
- In many cases, impossible to survey a population!
  - Typical for game analytics → want to understand/improve game for all

So … what to do?

http://www.mycariboonow.com/wp-content/uploads/2016/02/Population.jpg

Key Words

- Sample – part of population selected for analysis
  - e.g., all League of Legends players at WPI
  - e.g., students in first row in IMGD 2905 (e.g., poll: finish part 3?)



http://keydifferences.com/wp-content/uploads/2016/04/census-vs-sample.jpg

- Often hope *sample* is representative of *population*. …
- But Is it? → sampling important! (We won't talk much about this right now, however.)

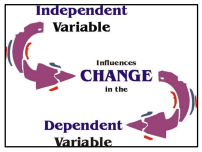
Key Words

- Variable – characteristic of individuals in population analyzing
  - e.g., time spent in competitive mode in League of Legends
  - e.g., vehicle choice in GTA
- *Independent* inherent in population, versus *dependent* that want to assess

http://study.com/cimages/videopreview/true-experimental-design_102058.jpg

$$Y = -5x - 4$$

Dependent Independent



http://9be7037176ebed526de6-851b178c8c16a40a353f5f093bed3d01d5.r37.cf2.rackcdn.com/067B2FAB-622E-4735-B750-12FB6022F61.jpg

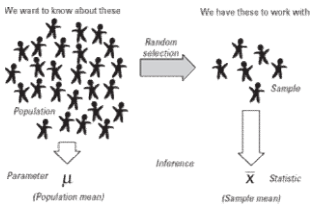
Key Words

- Observation – all variable values for individual sample
  - e.g., League of Legends competitive hours/week and Champion most played could be (2 observations)
    - "Player A: Leona, 2 hours"
    - "Player B: Teemo, 7.5 hours"
  - Can be continuous (time) or discrete (Champs)
- Often, data in grid
  - Observation rows
  - Variable columns

Player	Hours	Champ
A	2	Leona
B	7.5	Teemo

Key Words

- Parameter – measure of dependent variable in population
  - e.g., average crashes in Mario Cart Level
  - This is usually what we really want to know, but can't get
- Statistic – measure of dependent variable in sample
- Statistics is set of numerical methods for getting information about a population based on data from a sample, usually to get information about population parameters

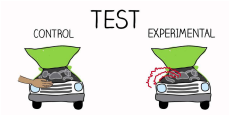


https://qph.ec.quoracdn.net/main-qimg-05879136f10bc9a0339823e1e01d9ec

"Statistics - a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."
-- Merriam-Webster dictionary

Sources of Data

- Published – generally made available from those that collected it
  - e.g., Riot League of Legends data
  - e.g., Metacritic reviews and ratings
- Experiments – multiple trials to collect data
  - Can be in laboratory or "real world" setting
  - e.g., play shooter, add lag and play again
- Survey – ask people to answer questions
  - e.g., self-rating as gamer, difficulty with level, …
  - Ethical issues with stress and use of data
  - → Institute Review Board (IRB) for approval with human subjects



https://i.ytimg.com/vi/qtLnBz6lbRQ/maxresdefault.jpg

http://www.mayersmemorial.com/pictures/content/122253.jpg

Sampling Concepts

- Sampling –process by which members of population are selected for sample
  - e.g., choose ½ class based on spacing, or choose ½ class based on alphabet
- Probability sampling – sampling considering likelihood of selection
  - e.g., survey for intended Champ, ask ½ class, but when tournament starts, result different. Why? → sample didn't consider League players! (e.g., often similar analogy for voter polls)
  - e.g., voluntary polls/surveys
  - Use probability sampling whenever possible, but sometimes it is not (cost) or not known
- Sampling with replacement – once sample, put back in pool
  - e.g., die roll to see which attack boss makes
- Sampling without replacement – once sample, won't sample again
  - e.g., user survey – don't allow to submit twice
  - E.g., deck of 52 cards for blackjack

Using Sample Data

- The word "sample" comes from same root word as "example"
  - Similarly, one sample does not prove a theory, but rather is an example
- Basically, in general, definite statement *cannot* be made about characteristics of all systems
- Instead, make **probabilistic statement** about range of most systems

→ That's where statistics come in!