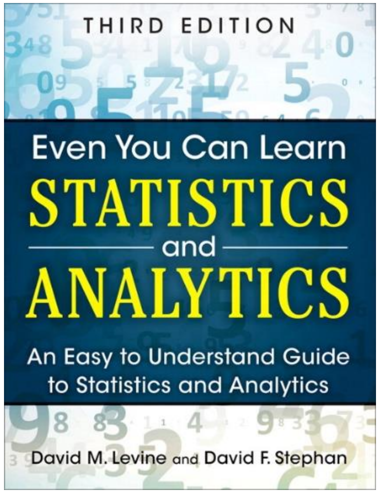
IMGD 2905

Fundamentals of Statistics

Chapter 1



THIRD EDITION

Even You Can Learn
STATISTICS
and
ANALYTICS

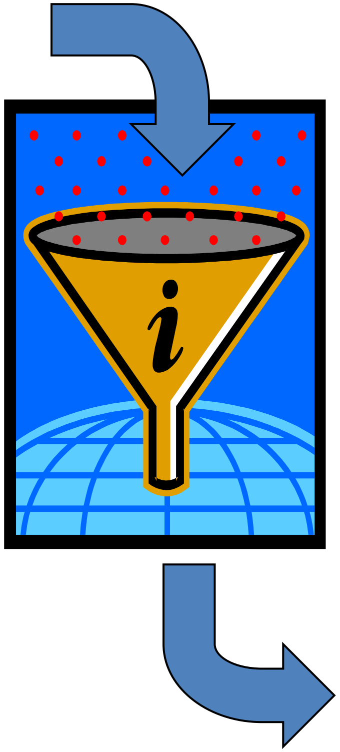
An Easy to Understand Guide
to Statistics and Analytics

David M. Levine and David F. Stephan

Why Do We Need Statistics?

Game

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?

→ First, some key words

Key Words

- Population – all members of group pertaining to a study

Q: examples?

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

Key Words

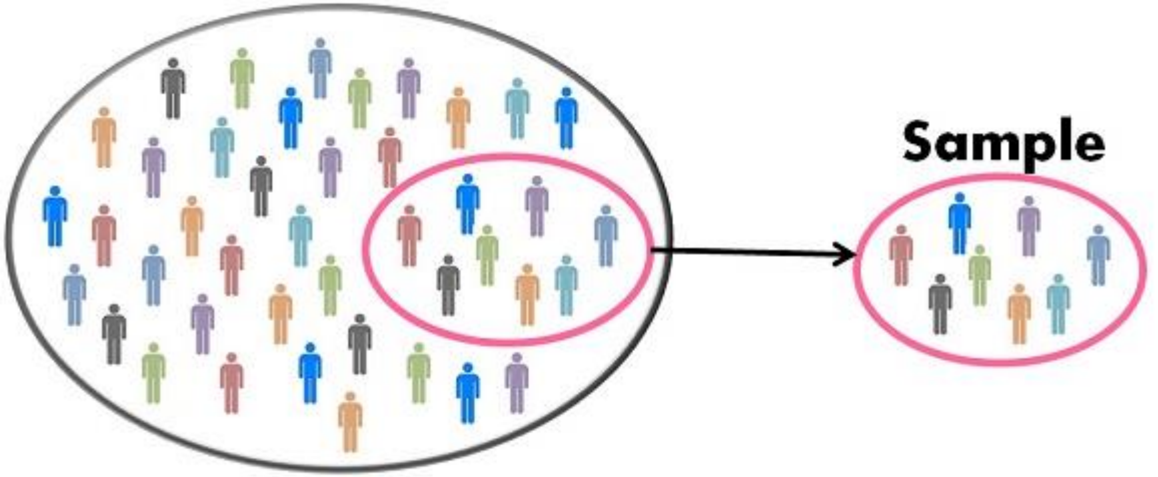
- Population – all members of group pertaining to a study
  - e.g., every person in IMGD 2905 in D-term
  - e.g., every *League of Legends* player in North America
- In many cases, impossible to survey a population!
  - Typical for game analytics → want to understand/improve game for all

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

Q: So … what to do?

Key Words

- Sample – part of population selected for analysis
  - e.g., all *League of Legends* players at WPI
  - e.g., students at one table in IMGD 2905



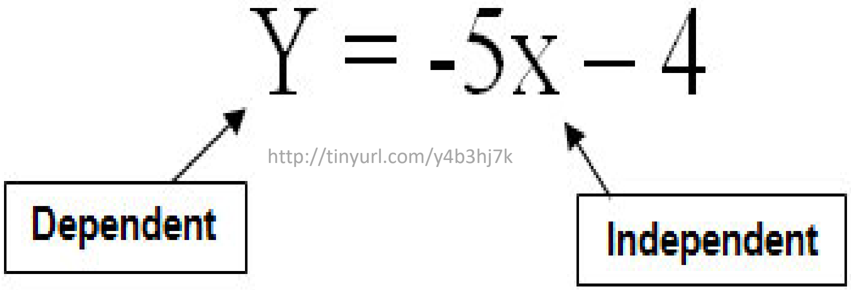
http://keydifferences.com/wp-content/uploads/2016/04/census-vs-sample.jpg

Q: Is sample same as population?
Is it *representative*?

- Often want *sample* to be representative of *population.* …
  - (e.g., poll: “did you finish chart for Project 2, Part 1?”)
- But Is it? → method to obtain sample is important! (We won’t talk much about this right now, however.)

More Key Words

- Variable – characteristic of individuals in population analyzing
  - e.g., time spent in competitive mode in *Starcraft 2*
  - e.g., vehicle choice in *Grand Theft Auto* (GTA)
- Independent variable is inherent in population, versus dependent variable that want to assess



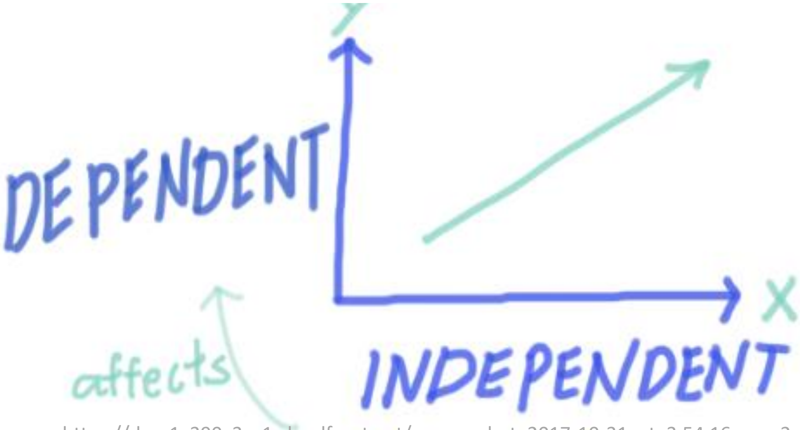
$Y = -5x - 4$

Dependent

Independent

http://tinyurl.com/y4b3hj7k

https://www.coursepics.com/wp-content/uploads/2016/11/Independent-and-Dependent-Variable.jpg



https://dqm1v390v3ac1.cloudfront.net/screen_shot_2017-10-31_at_3.54.16_pm_2.png

More Key Words

- Observation – all variable values for sample
  - e.g., *PUBG* hours/week and Best Rank (that week). Two observations could be:

    “Player A: Rank #2, 2 hours”

    “Player B: Rank #30, 7.5 hours”

  - Can be continuous (time) or discrete (rank)
- Often, data in grid
  - Observation in rows
  - Variables in columns
  - Format works well for spreadsheet
  - Consider our project 1 → *PUBG* data!

Player	Hours	Rank
A	2	#2
B	7.5	#30

Putting It Together

- Designing *Super Mario World* levels
- What are some dependent variables?
- What are some independent variables?
- Other variables of interest?
- What are some observations?

https://tinyurl.com/trb4h7v

https://tinyurl.com/s8tcprt

Putting It Together

https://tinyurl.com/s8tcprt

https://tinyurl.com/trb4h7v

- Designing *Super Mario World* levels
- What are some dependent variables?
- What are some independent variables?
- Other variables of interest?
- What are some observations?

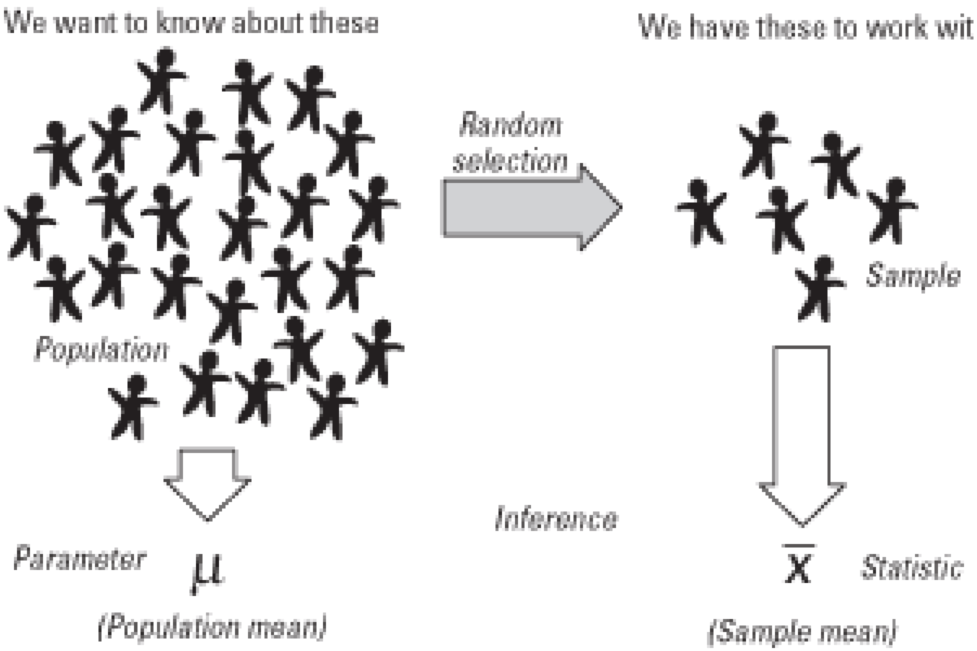
- Time, Deaths/fails, Fun …
- Koopas, power ups, gap lengths …
- Time spent getting coins, Number of jumps …

A, 10s, 12 jumps

Even More Key Words

- Parameter – measure of dependent variable for population
  - e.g., average crashes in *Mario Cart* level for every player
  - Usually what we want to know, but can't get easily
- Statistic – measure of dependent variable in sample
  - e.g., average crashes in Mario Cart level for IMGD 2905 class

- Statistics – set of numerical methods for getting information about population based on data from sample, usually to get information about population parameters



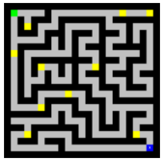
https://qph.ec.quoracdn.net/main-qimg-058791361f10bc9a0339823e1e01d3ec

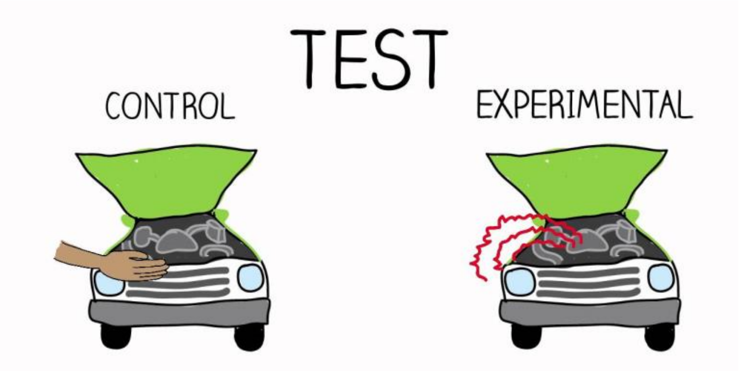
"Statistics - a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."
-Merriam-Webster dictionary

Sources of Data

- Published – generally made available from those that collected it
  - e.g., *PUBG* Developer API data
  - e.g., Metacritic's reviews and ratings
  - e.g., HOTS Logs dataset on *Heroes of the Storm*
- Experiments – multiple trials to collect data from sample
  - 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

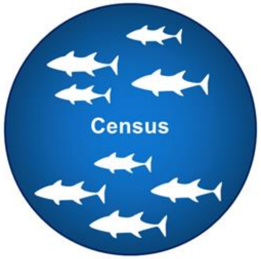
  → Institute Review Board (IRB) for approval with human subjects





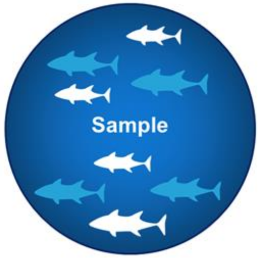
https://i.ytimg.com/vi/qtLnBz6lbRQ/maxresdefault.jpg

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



https://tinyurl.com/y4nu9ckf

Sampling Concepts



https://tinyurl.com/y4nu9ckf

- Sampling – process by which members of population selected for sample
  - e.g., choose ½ class based on seat, 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

https://tinyurl.com/y3ndyrom

Using Sample Data

- 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!

Statistics – set of numerical methods for getting information about population based on data from sample, usually to get information about population parameters