IMGD 2905

Fundamentals of Statistics

Chapter 1

Why Do We Need Statistics?

Aggregate data into meaningful information.

\[ \bar{x} = \ldots \]

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

Ok, but what are statistics?
Key Words

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

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 in first row in IMGD 2905

• Often hope *sample* is representative of *population*. ...
  – (e.g., poll: “did you finish chart for Project 1, Part 3?”)
• But Is it? → method to obtain sample is 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 *Starcraft 2*
  – e.g., vehicle choice in *Grand Theft Auto* (GTA)

• **Independent variable** is inherent in population, versus **dependent variable** that want to assess

\[ Y = -5X - 4 \]

Key Words

• **Observation** – all variable values for 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 (Champions)

• Often, data in grid
  – **Observation** in rows
  – **Variables** in columns

<table>
<thead>
<tr>
<th>Player</th>
<th>Hours</th>
<th>Champ</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>Leona</td>
</tr>
<tr>
<td>B</td>
<td>7.5</td>
<td>Teemo</td>
</tr>
</tbody>
</table>

• Consider our project 1 → *HOTS* data!
Key Words

- **Parameter** – measure of dependent variable for population
  - e.g., average crashes in *Mario Cart* level for everyone
  - 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

“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’s *League of Legends* data
  - e.g., Metacritic’s reviews and ratings
  - e.g., HOTS Logs dataset on *Heroes of the Storm*
- **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
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

• 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
  \[\rightarrow\text{That’s where statistics come in!}\]

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