Data Collection

Do you write "the data is..." or "the data are..."?

Is it singular or plural?

That is a deep philosophical question, actually.

It depends on whether you consider data to be facts (plural) or information (which is singular).

It's a fascinating grammatical conundrum.

What if I only have one data point?

Then you have bigger problems than grammar.
Topics for today

• What is data collection?
• What we must consider when choosing a dataset
• Comprehensive vs Sampled Data
• How data is represented in different formats on your computer
• In class activity
What is Data Collection

• Data collection is the process of collecting information
• Either you collect, process and format the data yourself
• Or you can find a dataset someone else already produced
Considerations about Data Collection

- What data is necessary to answer our question?
- How difficult is it to analyze a dataset?
- Comprehensive data vs sampled data?
- What is the allowed usage of data under it's license?
- Who collected the data?
- When was the data collected?
- How was the data collected?
- How is the data formatted?
- Does your data collection procedures need to be approved by an IRB?
- Confidentiality Concerns
Difficulty of analyzing data

Easy for people

Difficult for people

Confusion at Palm Beach County polls
Some AI Gops supporters may have mistakenly voted for Buchanan because of the ballot's design.

Although PA Democrats are listed first, they are the Ball side on the ballot.
Difficulty of analyzing data

We want data that’s easy for the computer to analyze, but often the data we are interested in is easy for humans to analyze.
Difficulty of analyzing data

Often we can suitably answer our questions from data pulled from these items:

- Speech transcripts
- Scripts
- Transcribed Lyrics
Comprehensive Data

• We have access to all the data points that exist, which is usually a lot

• Collected and digitized as part of general procedures of an institution

The New York Times

13 million articles

~500 Million tweets per day

CONGRESS.GOV

100,000s votes per year
Sampled Data

• When collecting individual data is relatively expensive

• Only a portion of individuals of a population are sampled

• Not just restricted to polling or surveys
Error vs Bias

- Error is the difference between the sampled population and the population as a whole.
- Error can be systematic and non-systematic.
- Systematic Error is also known as bias.
- Non-systematic error is also referred to as the variance.
Margin of Error

- Communicates the amount of random sampling error in a survey's results

- Margin of Error is always tied to a confidence. If it's not specified, it's probably 95%

- Asserts that with a given confidence, if you surveyed the entire population your result would be within the margin of error

- Scales with the square root of the number of samples

- Quadrupling your sample size reduces your margin of error by half
Biases in Sampled Data

• A bias in sampled data occurs when a procedure causes the sample to overrepresent a subpopulation

• They may not necessarily be intentional

• Even if you don’t think overrepresentation of a subpopulation will bias the dataset with regard to your question, it’s still a bias

• Always strive to minimize any biases in your data collection procedures
Gallup

- Randomly calls two groups of ~500 people a day by sampling among all possible phone numbers
- For landlines, asks for household member who has the next birthday
- Calls people living in all 50 states
- Tries to assure 70% cellphone, 30% landlines
- Weights data to reflect the demographics of the general population
IMDB Ratings

- Registered users rate films and TV shows on a 1-10 star rating
- Registered users are an overrepresented subpopulation relative to the general population
- Registered users who rate movies in their free time is further over represents a specific segment of the general population
- “Men Are Sabotaging The Online Reviews Of TV Shows Aimed At Women” - fivethirtyeight.com
  - 60% who rated Sex in the City were women. Women gave it a 8.1, men gave it 5.8.
Data representation
Bits and Bytes

- Data is stored to the computer in **binary**
- The smallest unit in binary is a **bit** (binary + digit)
- Bits can be either 1 or 0
- 8 consecutive bits are known as a **byte**
ASCII and UNICODE

- ASCII and UNICODE are representations for characters (‘h’, ‘3’, ‘!', ‘I’, ‘Ø’)

- ASCII (1963) was not the first the first representation, these days its fairly ubiquitous

- Originally ASCII only contained 128 different characters, but expanded to 256

- UNICODE exists in three different formats, UTF-8, UTF-16, UTF-32 but can represent the same values

- UTF-8 is commonly used as it is most space efficient

- UNICODE allows 1,112,064 possible characters but only 10% are currently used. Many are added each year (.*, *, #)

---

### ASCII TABLE

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>\0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>\1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>\2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>\3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>\4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>\5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>\6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>\7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>\8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>\9</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>\A</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>\B</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>\C</td>
</tr>
<tr>
<td>13</td>
<td>D</td>
<td>\D</td>
</tr>
<tr>
<td>14</td>
<td>E</td>
<td>\E</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>\F</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>\0</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>\1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>\2</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>\3</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>\4</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>\5</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>\6</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>\7</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>\8</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>\9</td>
</tr>
<tr>
<td>26</td>
<td>1A</td>
<td>\A</td>
</tr>
<tr>
<td>27</td>
<td>1B</td>
<td>\B</td>
</tr>
<tr>
<td>28</td>
<td>1C</td>
<td>\C</td>
</tr>
<tr>
<td>29</td>
<td>1D</td>
<td>\D</td>
</tr>
<tr>
<td>30</td>
<td>1E</td>
<td>\E</td>
</tr>
<tr>
<td>31</td>
<td>1F</td>
<td>\F</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decimal</th>
<th>Hex</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>20</td>
<td>\0</td>
</tr>
<tr>
<td>33</td>
<td>21</td>
<td>\1</td>
</tr>
<tr>
<td>34</td>
<td>22</td>
<td>\2</td>
</tr>
<tr>
<td>35</td>
<td>23</td>
<td>\3</td>
</tr>
<tr>
<td>36</td>
<td>24</td>
<td>\4</td>
</tr>
<tr>
<td>37</td>
<td>25</td>
<td>\5</td>
</tr>
<tr>
<td>38</td>
<td>26</td>
<td>\6</td>
</tr>
<tr>
<td>39</td>
<td>27</td>
<td>\7</td>
</tr>
<tr>
<td>40</td>
<td>28</td>
<td>\8</td>
</tr>
<tr>
<td>41</td>
<td>29</td>
<td>\9</td>
</tr>
<tr>
<td>42</td>
<td>2A</td>
<td>\A</td>
</tr>
<tr>
<td>43</td>
<td>2B</td>
<td>\B</td>
</tr>
<tr>
<td>44</td>
<td>2C</td>
<td>\C</td>
</tr>
<tr>
<td>45</td>
<td>2D</td>
<td>\D</td>
</tr>
<tr>
<td>46</td>
<td>2E</td>
<td>\E</td>
</tr>
<tr>
<td>47</td>
<td>2F</td>
<td>\F</td>
</tr>
<tr>
<td>48</td>
<td>30</td>
<td>\0</td>
</tr>
<tr>
<td>49</td>
<td>31</td>
<td>\1</td>
</tr>
<tr>
<td>50</td>
<td>32</td>
<td>\2</td>
</tr>
<tr>
<td>51</td>
<td>33</td>
<td>\3</td>
</tr>
<tr>
<td>52</td>
<td>34</td>
<td>\4</td>
</tr>
<tr>
<td>53</td>
<td>35</td>
<td>\5</td>
</tr>
<tr>
<td>54</td>
<td>36</td>
<td>\6</td>
</tr>
<tr>
<td>55</td>
<td>37</td>
<td>\7</td>
</tr>
<tr>
<td>56</td>
<td>38</td>
<td>\8</td>
</tr>
<tr>
<td>57</td>
<td>39</td>
<td>\9</td>
</tr>
<tr>
<td>58</td>
<td>3A</td>
<td>\A</td>
</tr>
<tr>
<td>59</td>
<td>3B</td>
<td>\B</td>
</tr>
<tr>
<td>60</td>
<td>3C</td>
<td>\C</td>
</tr>
<tr>
<td>61</td>
<td>3D</td>
<td>\D</td>
</tr>
<tr>
<td>62</td>
<td>3E</td>
<td>\E</td>
</tr>
<tr>
<td>63</td>
<td>3F</td>
<td>\F</td>
</tr>
</tbody>
</table>
Binary vs Plain Text

• All data stored to the computer is fundamentally represented in binary

• Data stored to a file that can be completely interpreted as ASCII or UNICODE is called **plain text**

• Binary data is simply data that cannot be interpreted as plain text

• Generally speaking, if your data file needs to be opened by a specific program it is probably stored as binary data

• If you can read your data as is on any text editor, it is likely plain text
Some examples of Plain Text Data

• Plain text, what people really mean when they say plain text

• Ends in .txt (generally)

• No formatting, font type, font size, color, etc.

• Text position is provided by whitespace characters (space, tab, return)

ALICE’S ADVENTURES IN WONDERLAND

Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversations?’
Some examples of Plain Text Data

• XML (.xml)

```
<roll_call_vote>
  <congress>115</congress>
  <session>1</session>
  ...
  <members>
    <member>
      <member_full>Alexander (R-TN)</member_full>
      <last_name>Alexander</last_name>
      <first_name>Lamar</first_name>
      <party>Р</party>
      <state>TN</state>
      <vote_cast>Yea</vote_cast>
      ...
    </member>
  </members>
</roll_call_vote>
```
Some examples of Plain Text Data

- CSV (.csv)
- Tab-separated (.tsv)
- **Delimiter**: The character that separates each value

```
stephen,brawner,instructor,stephen_brawner@brown.edu
linda,chang,hta,linda_chang@brown.edu
emma,sloan,ta,emma_sloan@brown.edu
palak,goel,ta,palak_goel@brown.edu
monica,mendoza,ta,monica-ann_mendoza@brown.edu
```
Some examples of Plain Text Data

- **JSON (.json)**
- **JavaScript Object Notation**
- Like XML, data is annotated
- A nesting of key-value pairs
- When whitespace is removed, can be more space efficient than XML

```json
{"menu": {
   "id": "file",
   "value": "File",
   "popup": {
      "menuitem": [
         {"value": "New", "onclick": "CreateNewDoc()"},
         {"value": "Open", "onclick": "OpenDoc()"},
         {"value": "Close", "onclick": "CloseDoc()"}
      ]
   }
}}
```
Some examples of Plain Text Data

- YAML (.yaml)
- YAML Ain’t Markup Language
- Annotated similarly to JSON
- Programs that can read YAML can read JSON

```yaml
invoice: 34843
date   : 2001-01-23
bill-to: &id001
  given : Chris
  family: Dumars
address:
  lines: |
  458 Walkman Dr.
  Suite #292
  city   : Royal Oak
  state  : MI
  postal : 48046
ship-to: *id001
```
File Extensions

• File extensions are the characters that follow the ‘.’ in a filename (hello.txt)

• Sometimes your computer will only allow programs to open files with a specific file extension

• Even if many programs can read a plain text data file, file extensions are used to communicate the type of data inside (.xml, .csv, .yaml, .json)
About Activity/Homework

• Todays homework is open-ended and asks you to formulate questions and find a suitable dataset to answer them

• It’s kind of a miniature version of your first project

• Take sometime to look through the datasets available on Kaggle and choose one that is meaningful to you

• If there are functions you would like to use and but unfamiliar with, check the online documentation or ask TAs for some guidance

• Likewise, if you’re unsure how to answer a question with the data you have, check with the TAs
Overview of Kaggle

https://www.kaggle.com/datasets