IMGD 2905

Descriptive Statistics

Chapter 3
Summarizing Data

• With lots of playtesting, there is a lot of data
  – This is a good thing!
• But raw data is often just a pile of numbers
  – Rarely of interest
  – Or even sensible
• Q: How to summarize all this information?
Summarizing Data

• With lots of playtesting, there is a lot of data
  – This is a good thing!
• But raw data is often just a pile of numbers
  – Rarely of interest
  – Or even sensible
• **Q:** How to summarize all this information?

Measures of central tendency

Examples? Pros and Cons?
Breakout 2

4 3 7 8 3 4 22 3 5 3 2 3

- Different for *central tendency* with **one** number?
- What are *pros* and *cons* of each?
- Icebreaker, Groupwork, Questions

https://web.cs.wpi.edu/~imgd2905/d20/breakout/breakout-2.html
Measure of Central Tendency: **Mean**

The sum of the measurements divided by the number of measurements gives you the mean.

\[
\frac{(6 + 4 + 5 + 4 + 8 + 3)}{6} = 5.
\]

- Also called the “*arithmetic mean*” or “*average*”
- In Excel, `=AVERAGE(range)`
  - `=AVERAGEIF()` – averages if numbers meet certain condition
Measure of Central Tendency: Median

• Sort values low to high and take middle value

Median

• In Excel, =MEDIAN(range)
Measure of Central Tendency: Mode

• Number which occurs most frequently
• Not too useful in many cases
→ Best use for categorical data
  – e.g., most popular Hero group in Heroes of the Storm

• In Excel, =MODE( )

Depiction: **Mean, Median, Mode?**

(a) Triangular distribution

(b) Uniform distribution

(c) Rectangular distribution

(d) Normal distribution

(e) Lorentzian distribution
Depiction: **Mean, Median, Mode?**

(a) Mean, median, mode

(b) No mode

(c) Mean, median

(d) Mode

(e) Mode
Which to Use, Mean, Median, Mode?
Which to Use, **Mean**, **Median**, **Mode**?

- **Mean** many statistical tests with sample
  - Estimator of population mean
  - Uses all data

- **Median** is useful for skewed data
  - e.g., income data (US Census) or housing prices (Zillo)
  - e.g., *Overwatch* team (6 players): 5 people level 5, 1 person level 275
    - Mean is 50 - not so useful since no one at this level
    - Median is 5 - more representative
  - Does not use all data. “Resistant” to extremes (e.g., 275)
  - But what if were exam scores? Hard to “bring up” grade

- **Mode** is useful primarily for categorical data only
  - Most played League champion, most popular maze, ...
Other Measures of Position

• May not always want center
  – e.g., want to know best LoL Champions

• What other positions may be desired?
Other Measures of Position

• Maximum / Minimum
  – Not discussed more

• Trimmed Mean
• Quartiles
• Percentiles

• May not always want center
  – e.g., want to know best LoL Champions
Trimmed Mean

• Take “trimming” off top and bottom (typically 5% or 10%)
  – Reduces effects of extreme values, like median

• In Excel, =TRIMMEAN(array, percent)

[Image: Trimmed Mean graph with blue and red bars, showing the original mean and trimmed mean.]

Blue – original mean
Red – trimmed mean

http://support.minitab.com/en-us/minitab/17/histogram_mean_vs_trimmed_mean.png
Quartiles

• Sort values
• First quartile (Q1) is 25% from bottom
• Third quartile (Q3) is 75% from bottom
• (What is second quartile?)
• In Excel, =QUARTILE(array,n)
Percentiles

- Generalization of quartiles
- $N^{th}$ percentile is data point $n\%$ from bottom of data
- Interpolate as for first quartile
- In Excel, =PERCENTILE(array,k) (k: 0 to 1)
Summarizing Data, Part 2

• Ok, pile of numbers can now be summarized as *one* number
  – Mean, median, mode
• But is that enough?
• Q: What other major aspect of numbers haven’t we summarized?
Summarizing Data, Part 2

- Ok, pile of numbers can now be summarized as one number
  - Mean, median, mode
- But is that enough?
- Q: What other major aspect of numbers haven’t we summarized?

Measures of variation (aka measures of dispersion, or measures of spread)
Summarizing Data, Part 2

“Then there is the man who drowned crossing a stream with an average depth of six inches.” – W.I.E. Gates

- Summarizing by single number rarely enough \(\Rightarrow\) need statement about **dispersion** (aka variation)

Above: does single number (**mean**) tell you enough about data?
Dispersion Overview (1 of 3)

- Is data clumped or spread out?
Dispersion Overview (2 of 3)

- Is data clumped or spread out?

![Histogram of Age of Best Actress Award Winners 1928–2009 (n = 83)]
Dispersion Overview (3 of 3)

- Is data clumped or spread out?

“Motion and Scene Complexity for Streaming Video Games”
What are Some Measures of Dispersion?
Breakout 3

Set A: 2 4 6 8 10
Set B: 2 9 9 10 10

• Different ways to report dispersion with one number?
• What are pros and cons of each?
• Icebreaker, Groupwork, Questions

Range

- Difference between smallest and largest value
- Somewhat obvious, but doesn’t tell you much about “clumping”
  - Minimum may be zero
  - Maximum can be from outlier
    - Event not related to phenomena studied (e.g., 0 on project)
    - Maximum gets larger with # samples, so no “stable” point

In Excel, `=MAX(array) - MIN(array)`
Variance

• Compute mean of sample
• Compute how far each value in sample is from mean
  – Some can be less than mean, some greater
  → So square this difference (why square?)
• Divide by number of sample values – 1
  – The “-1” corrects “bias” when trying to estimate population variance using sample variance

Sample Variance = \( s^2 = \frac{\sum (X - \overline{X})^2}{n - 1} \)
Variance Example

- Sample kills in *League of Legends* match
  - 12, 20, 16, 18, 19
  - What is sample variance?

- First, mean = 85 / 5 = 17

<table>
<thead>
<tr>
<th>Kills</th>
<th>X – mean</th>
<th>(X – mean)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

\[
s^2 = \frac{(25 + 9 + 1 + 1 + 4)}{(5 - 1)} = 40 / 4 = 10 \text{ kills squared}
\]

- In Excel, =VAR(array)

"Larger" means "more spread" ...
but units odd
Standard Deviation

• Square-root of variance

• Usually, use standard deviation instead of variance
  – Why? → Same *units* as data (e.g., “kills” in previous example)

• Can compare standard deviation to mean (*coefficient of variation*, next)

• But first:
  – Mendenhall’s Empirical Rule
  – Z-score
Mendenhall’s Empirical Rule

1. About 68% data within one standard deviation of mean
   - interval between mean-s and mean+s contains about 68% of data
2. About 95% within 2 standard deviations of mean
3. Almost all data within 3 standard deviations of mean

Rule assumes normal ("Bell curve") distribution
Z-Score

• Measure of how “far” from center (mean) single data point is
  – *Not* measure of dispersion for whole data set

\[ Z = \frac{X - \bar{X}}{S} \]

Example

<table>
<thead>
<tr>
<th>Mean</th>
<th>469</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std dev</td>
<td>119</td>
</tr>
<tr>
<td>X</td>
<td>650</td>
</tr>
</tbody>
</table>

Z-score for X?

\[
\frac{(650 - 469)}{119} = 1.52
\]

https://www.animatedsoftware.com/pics/stats/sgzscore2.gif