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?
Measure of Central Tendency: **Mean**

The sum of the measurements

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

gives you the mean.

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

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

- Number which occurs most frequently
- Not so useful in many cases
  - Best use for *categorical data*
    - e.g., most popular Champion group in League of Legends

- In Excel, `=MODE()`
Which to Use, Mean, Median, Mode?

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

- **Median** can be 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** can be useful primarily for categorical data
  - Most played League champion, most popular maze, ...
Other Measures of Position?

• May not always want center
  – e.g., what weapon that gets most kills in PUBG
• What other positions may be desired?
Other Measures of Position

- May not always want center
  - e.g., want to know *best* LoL Champions

- Maximum / Minimum
  - Not discussed more

- Trimmed Mean
- Quartiles
- Percentiles
Trimmed Mean

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

• In Excel, =TRIMMEAN(array, percent)  

Blue – original mean
Red – trimmed mean
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
- $N^\text{th}$ percentile is data point $n\%$ from bottom of data
- Interpolate as for first quartile
- In Excel, $=\text{PERCENTILE}(\text{array}, k)$ ($k: 0$ to $1$)

![Diagram](https://www.mathsisfun.com/data/images/percentile-80.svg)

![Graph](http://www.isical.ac.in/~jeexiiscore_normal/PercentilesAdvantages.htm)
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 → 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 showing the age distribution of Best Actress Award Winners from 1928 to 2009](chart.png)
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? → Groupwork

Age of Best Actress Award Winners 1928–2009 ($n = 83$)

More Spread

Scene Complexity (lbs) vs. Motion (pffm)
Groupwork

Group A: 0 6 12 18 26
Group B: 0 18 20 22 26

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

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/3-dispersion/handout.html
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, \( \text{=MAX(array)} - \text{MIN(array)} \)

![Diagram showing cumulative distribution with range calculation](http://idolosol.com/images/range-3.jpg)
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 (what does squaring do?)
• Divide by number of sample values – 1
  – The “-1” corrects “bias” when trying to estimate *population variance* using *sample variance*

\[
\text{Sample Variance} = s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}
\]
Variance Example

• Sample kills in PUBG matches
  – 12, 20, 16, 18, 19
  – What is sample variance?
• First, mean = \( \frac{85}{5} = 17 \) kills

<table>
<thead>
<tr>
<th>Kills</th>
<th>X – mean</th>
<th>((X – \text{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>

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

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

\[ s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n - 1}} \]

Average “distance” of points from mean
\[ c = \sqrt{(a^2 + b^2)} \]
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 holds for 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?

(650 – 469)/119 1.52
Coefficient of Variation (CV)

- Size of **standard deviation relative to mean**
  - e.g., large sd & large mean, not so spread
  - but large sd & small mean, more spread

- **Standard deviation divided by mean**
  - Can do this since same units!

- CV is “unit-less”, so measure of spread independent of quantity
  - E.g. seconds, clicks, spaces

Shown as percent (multiply by 100)

\[ CV = \frac{s}{\bar{x}} \times 100 \]

What is the relative CV for each curve?
Semi-Interquartile Range

• ½ distance between Q3 (75th percentile) and Q1 (25th percentile)

\[
\text{Q3} - \text{Q1} \quad \frac{}{2}
\]

• Guideline: use semi-interquartile (SIQR) for index of dispersion whenever using median as index of central tendency
Index of Dispersion Example

- First, sort. Then, compute:
  - Mean = 4.4
  - Min = 1.9, Max = 5.9
  - Median = [16 / 2] = 8\text{th} = 4.5
  - Q1 = 16 / 4 = 8\text{th} = 4.1
  - Q3 = 3 * 16 / 4 = 12\text{th} = 5.1

- \( SIQR = (Q3 - Q1) / 2 \) = 0.5
- \( Variance \) = 0.96
- \( Stddev \) = 0.98
- \( CV = stddev / mean \) = 0.22
- \( Range = \text{max} - \text{min} \) = 4
Groupwork

• Rank *measures of dispersion* by sensitivity to outliers
  – CoV
  – Range
  – Std Dev
  – Semi-interquartile Range

https://web.cs.wpi.edu/~imgd2905/d23/groupwork/4-outlier-effect/handout.html
Ranking of Affect by Outliers?

Measure of Dispersion

- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

Most to Least

http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg
Ranking of Affect by Outliers?

Measure of Dispersion
• Range
• Standard Deviation
• Coefficient of Variation
• Semi-interquartile Range

Most to Least
• Range susceptible
• Variance
  – Standard Deviation
  – Coefficient of Variation
• SIQR resistant

Only for quantitative data! Categorical can’t quantify spread since no ‘distance’
Instead, give categories for given percentile of samples
  e.g., “90% of samples are in 3 categories” (Pareto chart)
Depicting Dispersion in Charts

- Histogram
- Cumulative distribution
- Box-and-Whiskers
- Error Bars
Box-and-Whiskers Chart

• Way of showing variation
• Highlight middle 50% (interquartile range, IQR)
  – “Box”
• Lines go to smallest non-outlier
  – “Whiskers”
• Points indicate outliers
• Middle line shows median
• Sometimes with mean
• Outlier? → Data value “way out there”, “far” from the rest
  – Formally, 1.5+ IQRs away from quartile
• Available in Excel

Also called “boxplot”
Cumulative Distribution

- Cumulative amount of data with value or less
- Easy to see min, max, median
- Compare shapes of distributions

Demo: lol-patches.xlsx

Select column R (Bug Fixes)
Sort low to high
New column S for percent [=ROW()/164]
Select column → paste down all
Select both column R and S
Insert → Scatter plot with lines

“Nerfs, Buffs and Bugs - Analysis of the Impact of Patching on League of Legends”
http://www.cs.wpi.edu/~claypool/papers/lol-crawler/
Error Bars for Columns and Points

- Line through graph point parallel to axis with “caps”
- Denotes uncertainty (variation) in value

Excel: click “+” → “Error Bars” → “type”

- Often:
  - 1 standard deviation
- Can be (discuss later):
  - 1 standard error
  - 1 confidence interval

State clearly!