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 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 just a pile of numbers
  – Rarely of interest
  – Or even sensible
• Q: How to summarize all this information?

Measures of central tendency

Examples?

Measure of Central Tendency: 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

  ![Median Diagram](https://betterexplained.com/wp-content/uploads/average/median.png)

- 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)**: A graph showing a peak with mean, median, and mode labeled.
- **(b)**: A graph showing two peaks with mean and median labeled, and no mode.
- **(c)**: A graph showing a flat line with median labeled, and no mean or mode.
- **(d)**: A graph showing a peak with mean, median, and mode labeled.
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 League Champions

• What other positions may be desired?

Other Measures of Position

• May not always want center
  – e.g., want to know best League 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)`

![Trimmed Mean Diagram](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)`

![Quartiles Diagram](https://www.hackmath.net/images/quartiles.png)
Percentiles

- Generalization of quartiles
- $N^{th}$ percentile is data point $n\%$ from bottom of data
- Interpolate as for first quartile
- In Excel, \(=\text{PERCENTILE(array,k)}\) (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)

“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 variation

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

- Is data clumped or spread out?

![Histogram of Less Spread](https://mathbitsnotebook.com/Algebra1/StatisticsData/STSpread.html)

![Histogram of More Spread](https://mathbitsnotebook.com/Algebra1/StatisticsData/STSpread.html)

---

**Variation Overview (2 of 3)**

- Is data clumped or spread out?

![Histogram of Best Actress Award Winners](https://mathbitsnotebook.com/Algebra1/StatisticsData/STSpread.html)
**Variation Overview (3 of 3)**

- Is data clumped or spread out?

  "Motion and Scene Complexity for Streaming Video Games"

---

**What are Some Measures of Variation?**
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
  – 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

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

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

  First, mean = \( \frac{85}{5} = 17 \)

<table>
<thead>
<tr>
<th>Kills</th>
<th>( X - \text{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? \( \Rightarrow \) 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

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

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?

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

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

Semi-Interquartile Range

- \(\frac{1}{2}\) distance between **Q3** (75\textsuperscript{th} percentile) and **Q1** (25\textsuperscript{th} percentile)

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

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

<table>
<thead>
<tr>
<th>Lap Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
</tr>
<tr>
<td>2.7</td>
</tr>
<tr>
<td>3.9</td>
</tr>
<tr>
<td>4.1</td>
</tr>
<tr>
<td>4.2</td>
</tr>
<tr>
<td>4.2</td>
</tr>
<tr>
<td>4.4</td>
</tr>
<tr>
<td>4.5</td>
</tr>
<tr>
<td>4.5</td>
</tr>
<tr>
<td>4.8</td>
</tr>
<tr>
<td>4.9</td>
</tr>
<tr>
<td>5.1</td>
</tr>
<tr>
<td>5.1</td>
</tr>
<tr>
<td>5.3</td>
</tr>
<tr>
<td>5.6</td>
</tr>
<tr>
<td>5.9</td>
</tr>
</tbody>
</table>

- First, sort. Then, compute:
  - **Mean** = 4.4
  - **Min** = 1.9, **Max** = 5.9
  - **Median** = \[16 / 2\] = 8\textsuperscript{th} = 4.5
  - **Q1** = 16 / 4 = 8\textsuperscript{th} = 4.1
  - **Q3** = 3 * 16 / 4 = 12\textsuperscript{th} = 5.1
  - **SIQR** = (Q3 - Q1) / 2 = 0.5
  - **Variance** = 0.96
  - **Stddev** = 0.98
  - **CV** = stddev/mean = 0.22
  - **Range** = max – min = 4

Ranking of Affect by Outliers?

<table>
<thead>
<tr>
<th>Measure of Variation</th>
<th>Most to Least</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td></td>
</tr>
<tr>
<td>Range</td>
<td></td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
</tr>
<tr>
<td>Coefficient of Variation</td>
<td></td>
</tr>
<tr>
<td>Semi-interquartile Range</td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing outliers and data points](http://www.a-levelmathstutor.com/images/statistics/outliers-graph01.jpg)
Ranking of Affect by Outliers?

**Measure of Variation**
- Variance
- Range
- Standard Deviation
- Coefficient of Variation
- Semi-interquartile Range

**Most to Least**
- Range
- Variance
- Standard Deviation
- Coefficient of Variation
- SIQR

Index of Variation Summary

- Ranking of affect by outliers
  - Range
  - Variance
    - Standard deviation
    - Coefficient of variation
  - Semi-interquartile range

- Note, all only applied to **quantitative** data!
  - For **categorical** data, can’t quantify spread since no ‘distance’ between
  - Instead, give number of categories for given percentile of samples
    - e.g., “90% of samples are in 3 categories”
Depicting Variation in Charts

- Histogram (done)
- Cumulative distribution (done)
- Box-and-Whiskers (new)
- Error Bars (new)

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 2016

Sometimes called “boxplot”
Error Bars

- 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!