Descriptive Statistics

Chapter 3

Summarizing Data

• With lots of playtesting, there will be 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

• Also called the “arithmetic mean” or “average”
• In Excel, =AVERAGE(range)

Measure of Central Tendency: Mean

• Number which occurs most frequently
• Not too useful in many cases
  ➔ Best use for categorical data
  – e.g., most popular position in FIFA 18
• In Excel, =MODE()
Depiction: 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 (Zillow)
  - 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

- Maximum / Minimum
  - Not discussed more
- Trimmed Mean
- Quartiles
- Percentiles

Other Measures of Position

- Take “trimming” off top and bottom (typically 5% or 10%)
  - Reduces effects of extreme values, like median
- In Excel, =TRIMMEAN(array, percent)
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
- Nth 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?

Variation Overview (1 of 3)
- Is data clumped or spread out?
- Summarizing by single number rarely enough \(\Rightarrow\) need statement about variation

Above: does single number (mean) tell you enough about data?
Variation Overview (2 of 3)

• Is data clumped or spread out?

Variation Overview (3 of 3)

• Is data clumped or spread out?

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

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, \[ \text{mean} = \frac{12 + 20 + 16 + 18 + 19}{5} = 17 \]
\[ \begin{array}{c|c|c}
\text{Kills} & X - \text{mean} & (X - \text{mean})^2 \\
12 & -5 & 25 \\
20 & 3 & 9 \\
16 & -1 & 1 \\
18 & 1 & 1 \\
19 & 2 & 4 \\
\end{array} \]
\[ s^2 = \frac{(25 + 9 + 1 + 1 + 4)}{(5 - 1)} = 10 \text{ kills squared} \]

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

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

\([\text{Low Standard Deviation}]
\hspace{2cm} [\text{High Standard Deviation}]\]

**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

(Rules assume normal 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>X</th>
<th>X-bar</th>
<th>S</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>469</td>
<td>469</td>
<td>119</td>
<td>1.52</td>
</tr>
</tbody>
</table>

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

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

\[ \text{SIQR} = \frac{Q3 - Q1}{2} \]

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>(sorted) Lap Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>First, sort. Then, compute:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean = 4.4</td>
</tr>
<tr>
<td>Min = 1.9, Max = 5.9</td>
</tr>
<tr>
<td>Median = [16 / 2] = 8th = 4.5</td>
</tr>
<tr>
<td>Q1 = 16 / 4 = 4th = 4.1</td>
</tr>
<tr>
<td>Q3 = 3 * 16 / 4 = 12th = 5.1</td>
</tr>
<tr>
<td>SIQR = (Q3 - Q1) / 2 = 0.5</td>
</tr>
<tr>
<td>Variance = 0.96</td>
</tr>
<tr>
<td>Stdddev = 0.98</td>
</tr>
<tr>
<td>CV = stddev/mean = 0.22</td>
</tr>
<tr>
<td>Range = max – min = 4</td>
</tr>
</tbody>
</table>
Ranking of Affect by Outliers?

Measure of Variation Most to Least
• Variance
• Range
• Standard Deviation
• Coefficient of Variation
• Semi-interquartile Range

Index of Variation Summary
• Ranking of affect by outliers
  – Range susceptible
  – Variance
    • Standard deviation
    • Coefficient of variation
  – Semi-interquartile range resistant
• 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”

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