Measures of Central Tendency & Variability

Week 2

Prof. Patrick Bennett

Describing Data Distributions

- Often we wish to summarize distributions of data, rather than showing histograms.
- Two basic descriptions of a distribution include its "middle" (central tendency) and "how spread out it is" (variability).

Part 1 - Central Tendency

- Often want to describe "typical" score in data
- How to define a "typical" score?
- Common measures:
  - mode (most common)
  - median (middle score; 50th percentile)
  - mean (average)

Example – Binocular Rivalry Times

<table>
<thead>
<tr>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>300</td>
</tr>
<tr>
<td>400</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>600</td>
</tr>
</tbody>
</table>

Binocular Rivalry (Young Adults)

Mode = 0.875

Mode is most common score
Mode - Most Common Score

- The mode is the most common/frequent score
- If 2 adjacent scores occur with equal frequency:
  - mode = average of those two scores
- If 2 non-adjacent scores occur with equal frequency:
  - distribution is bi-modal
  - report both numbers
- If range of number occur with nearly equal frequency:
  - mode is ill-defined
  - report as: “the mode fell within the range of X to Y”

Mode may be poorly defined

In some cases, small fluctuations in frequencies can produce BIG changes in mode

Median = Middle Score

Odd number of scores (N=11)

> scores:
  92  95 103 108  79 111 106  99 100 108  85

> sorted scores:
  79  85  92  95  99 100 103 106 108 108 111

> median(scores):
  100

Median Location = (N+1)/2 = (11+1)/2 = 6

Median = Middle Score

Even number of scores (N=10)

> scores:
  92  95 108  79 111 106  99 100 108  85

> sorted scores:
  79  85  92  95  99 100 106 108 108 111

> median(scores):
  (99+100)/2 = 99.5

Median Location = (N+1)/2 = (10+1)/2 = 5.5
Cumulative Frequency Plot & Percentiles

Binocular Rivalry (Young Adults)

Cumulative Frequency

- 25% cutoff = 0.8 s
- 50% cutoff = 1.23 s
- 75% cutoff = 1.9 s
- 95% cutoff = 5 s

Cumulative Frequency of x is the frequency/count of scores ≤ x

Upper asymptote total count = 5285

Binocular Rivalry (Young Adults) Duration (s)

Cumulative Frequency

Median = 1.23

50% line

Median is the middle score (50th percentile)

50% cutoff = 1.23 s

Median is Robust to Outliers

Odd number of scores (N=11) replace 108 with 99,999

> scores:
92  95 103 108  79 111 106  99 100 99,999 85

> sorted scores:
79  85  92  95  99 100 103 106 108 111 99,999

> median(scores):
100

Median Location = (N+1)/2 = (11+1)/2 = 6
Mode vs. Median

Binocular Rivalry (Young Adults)

- Frequency
- Percept duration (sec)

Mode = 0.875
Median = 1.23

Interlude: Summation Notation

- Suppose you have a set of \( n \) scores, \( X \):
  - \( X_1, X_2, X_3, \ldots, X_n \)
  - Represent an arbitrary score as \( X_i \)
- Following notation represents the summation of all \( n \) scores:

\[
\sum_{i=1}^{n} X_i = (X_1 + X_2 + X_3 + \cdots + X_n)
\]

Summation (continued)

- We often use a compact form to represent summing over all scores:

\[
\sum X
\]

\[
\sum_{i=1}^{n} X_i = (X_1 + X_2 + X_3 + \cdots + X_n)
\]

Summation (continued)

- Operations to the right of Sigma are performed on individual \( Y \)'s before summation

\[
\sum X^2 = (X_1^2 + X_2^2 + X_3^2 + \cdots + X_n^2)
\]
Summation (continued)

• Brackets indicate operations performed after summation

\[ (\sum X)^2 = (X_1 + X_2 + X_3 + \cdots + X_n)^2 \]

Mean

• most commonly used measure of central tendency
• the **average** score:
  - (sum of all scores) divided by (number of scores)

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}
\]

early illustration of calculating a mean

Why the mean?

• consider the following game:
  - on each trial, I randomly pick a score from my data
  - you guess the value
  - repeat for many trials
  - define “error” as difference between guess & number
  - your goal: minimize the sum of squared errors

• what is the best strategy?
  - Answer: guess the mean on every trial
  - the mean minimizes the sum of squared errors
  - mean is “least squares” estimate of a typical score
Mean is NOT Robust to Outliers

Replacing score of 108 with 99,999 had no effect on median:

> scores: 92 95 103 108 79 111 106 99 100 99,999 85
> sorted scores: 79 85 92 95 99 100 103 106 108 111 99,999
> median(scores): 100

but has a BIG effect on the mean:
with 108, mean = 98.7; with 99,999, mean = 9179.7

Trimmed Means

- To make the mean less sensitive to extreme values, we can trim a certain percentage of values off of the tails
  - Example: scores = {2, 2, 3, 5, 6, 7, 8, 8, 9, 501}
    - mean = 55.1
  - Now trim 10% off tails at both ends of sorted scores:
    - trimmed scores = {2, 2, 3, 5, 6, 7, 8, 8, 9, 501}
      - 10% trimmed mean = 6
  - Amount of trimming varies across applications/situations:
    - e.g., 20% trimmed mean removes upper & lower 20% of scores

Mode, Median, & Mean

Binocular Rivalry (Young Adults)

Mode = 0.875
Median = 1.23
Mean = 1.71
Mode, Median, & Mean

**Survey of Consumer Finances (N=1000)**

<table>
<thead>
<tr>
<th>Value of All Vehicles ($)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50000</td>
<td>150</td>
</tr>
<tr>
<td>100000</td>
<td>200</td>
</tr>
<tr>
<td>150000</td>
<td>150</td>
</tr>
<tr>
<td>200000</td>
<td>100</td>
</tr>
</tbody>
</table>

**Mode** = $0

**Median** = $11,000

**Mean** = $15,398

**Mode – Advantages & Disadvantages**

- **Advantages**
  - robust to outliers (extreme scores)
  - value actually appears in the data
  - represents the greatest probability of subjects having a score
  - can be found for nominal data
    ‣ e.g., mode of household pet type “dog”; no analogous mean or median
- **Disadvantages**
  - depends on how we bin scores
  - can be poorly defined & unstable for flat distributions

**Median – Advantages & Disadvantages**

- **Advantages**
  - robust to outliers (extreme scores)
  - can be calculated even with flat distributions
  - good index of “typical” score in skewed distributions
- **Disadvantages**
  - no mathematical formula for the median
    ‣ difficult to use median in mathematical derivations/equations
  - in some situations, between-sample variability is greater for median than mean (i.e., median less stable than mean)

**Mean – Advantages & Disadvantages**

- **Advantages:**
  - in some situations, between-sample variation is less for sample means than sample modes & medians
    ‣ i.e., means are more stable across samples
  - enters readily into algebraic equations
- **Disadvantages:**
  - value may not actually exist in the data
  - less robust than median to extreme values
    ‣ use trimmed means instead?
City’s income growth is tops for Ontario cities
Steve Buist

Income growth in Hamilton over the past decade was the highest of Ontario’s five largest metropolitan areas, according to new census data released Wednesday by Statistics Canada.

That, along with a noticeable drop in the city’s poverty rate, is the good news.

The bad news is that Ontario’s income performance between 2005 and 2015 lagged far behind every other province and territory in Canada.

Median total household income in the Hamilton census metropolitan area, which includes Burlington and Grimsby, grew by 5.3 per cent over the past decade, substantially better than the 3.8 per cent rise in income across Ontario as a whole.

Hamilton household income increase, 2005 to 2015: $65,440 to $69,020

Part 1 - Central Tendency (summary)

- Mode, Median, Mean
  - methods of calculation
  - advantages & disadvantages
- Summation Notation
- Trimmed Means