STA 291
Lecture 10, Chap. 6

• Describing Quantitative Data
  – Measures of Central Location
  – Measures of Variability (spread)

• First Midterm Exam a week from today,
  • Feb. 23 5-7pm
  • Cover up to mean and median of a
    sample (begin of chapter 6). But not any
    measure of spread (i.e. standard
deviation, inter-quartile range etc)

Summarizing Data Numerically

• Center of the data
  – Mean (average)
  – Median
  – Mode (…will not cover)

• Spread of the data
  – Variance, Standard deviation
  – Inter-quartile range
  – Range
Mathematical Notation: Sample Mean

- Sample size $n$
- Observations $x_1, x_2, \ldots, x_n$
- Sample Mean “x-bar” — a statistic

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

Mathematical Notation: Population Mean for a finite population of size $N$

- Population size (finite) $N$
- Observations $x_1, x_2, \ldots, x_N$
- Population Mean “mu” — a Parameter

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i$$

Infinite populations

- Imagine the population mean for an infinite population.
- Also denoted by $\mu$ or $\mu$
- Cannot compute it (since infinite population size) but such a number exist in the limit.
- Carry the same information.
Infinite population

• When the population consists of values that can be ordered
• Median for a population also make sense: it is the number in the middle… half of the population values will be below, half will be above.

Mean

• If the distribution is highly skewed, then the mean is not representative of a typical observation
• Example:
  Monthly income for five persons
  1,000  2,000  3,000  4,000  100,000
  • Average monthly income:  = 22,000
  • Not representative of a typical observation.

• Median = 3000
Median

• The median is the measurement that falls in the middle of the *ordered* sample
• When the sample size *n* is odd, there is a middle value
• It has the *ordered index* \((n+1)/2\)

Example: 1.1, 2.3, 4.6, 7.9, 8.1

\(n=5\), \((n+1)/2=6/2=3\), so index = 3,

Median = \(3^{rd}\) smallest observation = 4.6

Median

• When the sample size *n* is even, average the two middle values

Example: 3, 7, 8, 9, \(n=4\),

\((n+1)/2=5/2=2.5\), index = 2.5

Median = midpoint between 2\(^{nd}\) and 3\(^{rd}\) smallest observation

= \((7+8)/2 = 7.5\)

Summary: Measures of Location

<table>
<thead>
<tr>
<th>Mean</th>
<th>Arithmetic Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean of a Sample - (\bar{x})</td>
</tr>
<tr>
<td></td>
<td>Mean of a Population - (\mu)</td>
</tr>
</tbody>
</table>

Median – Midpoint of the observations when they are arranged in increasing order

Mode...

Notation: Subscripted variables

\(n = \# \) of units in the sample

\(N = \# \) of units in the population

\(x = \) Variable to be measured

\(x_i = \) Measurement of the \(i\)th unit
## Mean vs. Median

<table>
<thead>
<tr>
<th>Observations</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1, 2, 3, 4, 100</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>3, 3, 3, 3, 3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1, 2, 3, 100, 100</td>
<td>3</td>
<td>41.2</td>
</tr>
</tbody>
</table>

### Mean vs. Median

- If the distribution is symmetric, then $\text{Mean} = \text{Median}$
- If the distribution is skewed, then the mean lies more toward the direction of skew
- [Mean and Median Online Applet](#)

## Why not always Median?

- Disadvantage: Insensitive to changes within the lower or upper half of the data
- Example: $1, 2, 3, 4, 5, 6, 7$ vs. $1, 2, 3, 4, 100, 100, 100$
- For symmetric, bell shaped distributions, mean is more informative.
- Mean is easy to work with. Ordering can take a long time
- Sometimes, the mean is more informative even when the distribution is slightly skewed
Given a histogram, find approximately the mean and median.
Percentiles

- The $p$th percentile is a number such that $p\%$ of the observations take values below it, and $(100-p)\%$ take values above it
- 50th percentile = median
- 25th percentile = lower quartile
- 75th percentile = upper quartile

Quartiles

- 25th percentile = lower quartile = $Q_1$
- 75th percentile = upper quartile = $Q_3$

Interquartile range = $Q_3 - Q_1$
(a measurement of variability in the data)

SAT Math scores

- Nationally (min = 210  max = 800 )
  - $Q_1 = 440$
  - Median = $Q_2 = 520$
  - $Q_3 = 610$  (→ you are better than 75% of all test takers)
- Mean = 518  (SD = 115  what is that?)
Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Statistical Software SAS output (Murder Rate Data)

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% Max</td>
<td>20.30</td>
</tr>
<tr>
<td>75% Q3</td>
<td>10.30</td>
</tr>
<tr>
<td>50% Median</td>
<td>6.70</td>
</tr>
<tr>
<td>25% Q1</td>
<td>3.90</td>
</tr>
<tr>
<td>0% Min</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Example: The five-number summary for a data set is min=4, Q1=256, median=530, Q3=1105, max=320,000.

What does this suggest about the shape of the distribution?
Box plot

• A box plot is a graphic representation of the five number summary --- provided the max is within 1.5 IQR of Q3 (min is within 1.5 IQR of Q1)

• Otherwise the max (min) is suspected as an outlier and treated differently.
• Box plot is most useful when compare several populations

Measures of Variation
• Mean and Median only describe the central location, but not the spread of the data
• Two distributions may have the same mean, but different variability
• Statistics that describe variability are called measures of spread/variation

Measures of Variation
• Range: $= \text{max} - \text{min}$
  Difference between maximum and minimum value
• Variance: $s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$
• Standard Deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$
• Inter-quartile Range: $= Q3 - Q1$
  Difference between upper and lower quartile of the data
Deviations: Example

- Data: 1, 7, 4, 3, 10
- Mean: (1+7+4+3+10)/5 = 25/5 = 5

<table>
<thead>
<tr>
<th>Data</th>
<th>Deviation</th>
<th>Dev. square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1 - 5) = -4</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>(3 - 5) = -2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>(4 - 5) = -1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>(7 - 5) = 2</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>(10 - 5) = 5</td>
<td>25</td>
</tr>
<tr>
<td>Sum</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

Sample Variance

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

The variance of \( n \) observations is the sum of the squared deviations, divided by \( n-1 \).

Variance: Example

<table>
<thead>
<tr>
<th>Observation</th>
<th>Mean</th>
<th>Deviation</th>
<th>Squared Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Sum of the Squared Deviations</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n - 1 )</td>
<td>5-1=4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum of the Squared Deviations / ( (n-1) )</td>
<td>50/4=12.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• So, sample variance of the data is 12.5
• Sample standard deviation is 3.53

\[ \sqrt{12.5} = 3.53 \]

Attendance Survey Question
• On a 4”x6” index card
  – write down your name and section number
  – Question:
    – Lexington Average temperature in Feb. Is about ________?

Example: Mean and Median
• Example: Weights of forty-year old men
  158, 154, 148, 160, 161, 182, 166, 170, 236, 195, 162
• Mean =
• Ordered weights: (order a large dataset can take a long time)
  148, 154, 158, 160, 161, 162, 166, 170, 182, 195, 236
• Median =