STA 291
Lecture 13, Chap. 6

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

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 “\( \bar{x} \)” --- a statistic

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

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[\sum = \text{SUM}\]
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

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

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 )
  - Q1 = 440
  - Median = Q2 = 520
  - Q3 = 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>
Five-Number Summary
• Maximum, Upper Quartile, Median, Lower Quartile, Minimum

• 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: \( s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \)
- Standard Deviation: \( s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \)
- Inter-quartile Range: \( Q3 - Q1 \)
  Difference between upper and lower quartile of the data

Deviations: Example

- Sample Data: 1, 7, 4, 3, 10
- Mean (\( \bar{x} \)): \( \frac{1+7+4+3+10}{5} = 5 \)

<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=25</td>
<td>Sum = 0</td>
<td>sum = 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

$\sqrt{12.5} = 3.53$

• Variance/standard deviation is also more susceptible to extreme valued observations.

• We are using x-bar and variance/standard deviation mostly in the rest of this course.
Population variance/standard deviation

- Notation for Population variance/standard deviation (usually obtain only after a census)

- Sigma-square / sigma

\[ \sigma^2 / \sigma \]

standardization

- Describe a value in a sample by
- “how much standard deviation above/below the average”

- The value 6 is one standard deviation above mean -- the value 6 corresponds to a z-score of 1
- May be negative (for below average)

Attendance Survey Question

- On a 4”x6” index card
  - write down your name and section number
  - Question: Independent or not?
  - Gender of first child and second child from same couple.