3190 Week 2

Describing data
Ordered arrays

• The simplest organizational tool for working with data is to order it

• An ordered array is a list of numerical values associated with a variable in rank order from the smallest value to the largest value

• So, are the unemployment data an ordered array?
Frequency Distributions

• These are tables of your grouped data that show the frequency of cases in each group

• The groups are in the left column

• The frequencies are in the adjacent column to the right

• Percentages are in a third column to the right
Central Tendency

- These are calculations that represent the central or typical value in a distribution of values for a variable
  - Three types
    - Mean
    - Mode
    - Median
Measures of Dispersion

• Calculations that depict the amount of spread or variability in a set of data values of a variable
  • There are several
    • The range
    • The interquartile range
    • The variance
    • The standard deviation
    • The coefficient of variation
The Mean

- The arithmetic average of a sample calculated as summation of the scores divided by the number of cases for a variable

- Notation
  - $\sum = \text{“sum of”}$
  - $x_i = \text{“a particular score for a case”}$
  - $n = \text{“total number of cases” or “sample size”}$
  - $\mu = \text{“population mean”}$
  - $\bar{X} = \text{“sample mean”}$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$
Why different symbols...?

• It is important to note that a sample is a sub-set of values from a population

• Thus, any statistic calculated from a sample is an estimate of the population parameter.

• So, we use two symbols to keep them straight; we rarely have population level data.
Unweighted vs. weighted mean

data from Table 2.6, Table 3.2 McGrew & Monroe

• When we use the raw data for a distribution, as the formula indicates, the mean is unweighted (the normal way of calculating)
  •  = 1598/40 = 39.95 inches

• When we use grouped data, we weight the mean.

\[
\bar{X}_w = \frac{\sum X_j f_j}{n}
\]

• Simply multiply the class midpoint by the frequency of cases in each class
  •  = 1610/40 = 40.25 inches
<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Class midpoint $X_j$</th>
<th>Class frequency $f_j$</th>
<th>$X_j f_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25–29.99</td>
<td>27.5</td>
<td>4</td>
<td>110.0</td>
</tr>
<tr>
<td>30–34.99</td>
<td>32.5</td>
<td>5</td>
<td>162.5</td>
</tr>
<tr>
<td>35–39.99</td>
<td>37.5</td>
<td>12</td>
<td>450.0</td>
</tr>
<tr>
<td>40–44.99</td>
<td>42.5</td>
<td>9</td>
<td>382.5</td>
</tr>
<tr>
<td>45–49.99</td>
<td>47.5</td>
<td>5</td>
<td>237.5</td>
</tr>
<tr>
<td>50–54.99</td>
<td>52.5</td>
<td>4</td>
<td>210.0</td>
</tr>
<tr>
<td>55–59.99</td>
<td>57.5</td>
<td>1</td>
<td>57.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
<td></td>
<td><strong>1610.0</strong></td>
</tr>
</tbody>
</table>

$$\bar{X}_w = \frac{\sum X_j f_j}{n} = \frac{1610.0}{40} = 40.25$$

Source: National Climatic Data Center, U.S. Dept. of Commerce.
The Weighted Mean

• Why is it weighted? Because it is not based on the original data distribution.

• Comes in handy when all we have to work with is a frequency distribution, a histogram, or a frequency polygon

That is, when we do not have access to the original data
The mode

• The mode is the most frequently occurring score in a data distribution

• Often there is no mode, or there is more than one mode in a distribution

• There is no mode in Table 2.6 because each score is different
The crude mode

• This is the “weighted” mode, based on grouped data; it is the midpoint of the most frequently represented group

• Look at Table 3.2, which group has the most members?
  • Although there was no mode; there is a crude mode for the grouped data
  • It is 37.5 inches
The median

• An ordered array’s middle value; the value with an equal number of cases above and below it

• When there is an odd number of cases, it is the middle value, aka the 50th percentile

• When there is an even number of cases it is half way between the two middle scores
  • Easily, efficiently calculated as

\[
\text{Median} = \frac{(n + 1)}{2} \text{ ordered observation}
\]
An example

• For Table 2.6
  
  • $(40 + 1) / 2 = \text{position } 20.5$
  
  • So the median \textit{score} is halfway between the scores at 20 & 21
  
  • \textbf{Score at P20 = 39.62; at P21 = 39.86}
  
  • Sum the two scores, and divide by 2 to get the score of the median

  \[ \text{= 39.74 inches} \]
Which one should I use...?

- The mean, median, & mode are the same in a perfectly symmetrical distribution
- Mode is not useful if the distribution is multimodal

- The mean is the preferred measure for symmetrical (or nearly symmetrical) distributions because there are many statistical inference tools designed specifically for analyzing means later in the course (see syllabus)
Pearson’s skewness

- Based on the logic that the mean is more affected by skewness than the median
  - So we can use the difference between them to assess severity and direction of skewness
  - Pear. Skw. varies between $\pm 3.0$
  - $> \pm 0.6$ is substantial skewness
  - If skewness is substantial use the median not the mean

\[
\text{Pearson’s skewness} = \frac{3(\bar{X} - \text{median})}{s}
\]
Important terminology

• Mean, standard deviation (based on mean), all referred to as “parametric,” “interval-scale” or “absolute” statistics in this course
  – Based on original data scores in data distributions
  – Related inferential tests use more assumptions (e.g., symmetry, normality)

• Median, interquartile range, range, ordered array, Pearson’s skewness all referred to as “non-parametric,” “ordinal-scale” or “relative” statistics in this course
  – Relies on positions or scores at positions in ordered arrays
  – Related inferential tests use few assumptions (does not assume normality)
Assuming normality

• To assume normality is to assume that the shape of the distribution of a variable for a population is unimodal & symmetrical
  – remember “parametric” = “about population”

• We would like to be able to assume normality
  – Then we can use parametric statistics, which are more powerful
  – More powerful because we can use the normal probability distribution to make predictions

• If our sample is random, we can assume normality at samples $n \geq 30$, why?
  – We will discuss normality in more detail later on, but learn these basics
Dispersion

Non-parametric & parametric alternatives
The five points of data summary

- With 5 non-parametric stats we can learn a lot about a dataset
  - The **minimum** (score at the lowest position)
  
  - The **25th percentile** (the score at the position with 25% of the cases below it; 75% above)
    \[ 25^{\text{th}} \text{ percentile} = \text{score at position } \frac{n+1}{4} \]

  - The **median** (50th percentile; the score at the position with 50% of the cases above & below it)
    \[ \text{Median} = \text{score at position } \frac{n+1}{2} \]

  - The **75th percentile** (the score at the position with 75% of the cases below it; 25% above)
    \[ 75^{\text{th}} \text{ percentile} = \text{score at position } \frac{3(n+1)}{4} \]

  - The **maximum** (score at the highest position)
### Table 2.6

**Annual Precipitation for Washington, D.C.: A Ranked 40-Year Record (in Inches)**

<table>
<thead>
<tr>
<th>Class Interval</th>
<th>Class Midpoint $X_j$</th>
<th>Frequency $f_j$</th>
<th>Product $X_jf_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>26.87-26.94</td>
<td>35.38</td>
<td>40.21</td>
<td>46.02</td>
</tr>
<tr>
<td>28.28</td>
<td>35.96</td>
<td>40.54</td>
<td>47.73</td>
</tr>
<tr>
<td>29.48</td>
<td>36.02</td>
<td>41.11</td>
<td>47.90</td>
</tr>
<tr>
<td>31.56</td>
<td>36.65</td>
<td>41.34</td>
<td>48.02</td>
</tr>
<tr>
<td>32.78</td>
<td>36.83</td>
<td>41.44</td>
<td>50.50</td>
</tr>
<tr>
<td>33.07</td>
<td>36.99</td>
<td>41.46</td>
<td>51.17</td>
</tr>
<tr>
<td>33.62</td>
<td>38.15</td>
<td>41.94</td>
<td>51.97</td>
</tr>
<tr>
<td>34.98</td>
<td>39.34</td>
<td>43.30</td>
<td>54.29</td>
</tr>
<tr>
<td>35.99</td>
<td>39.62</td>
<td>43.53</td>
<td>57.54</td>
</tr>
</tbody>
</table>

*Source: National Climatic Data Center, U.S. Dept. of Commerce.*

### Table 3.2

**Worktable for Calculating Weighted Mean of Washington, D.C., Precipitation Data**

<table>
<thead>
<tr>
<th>Class Interval $j$</th>
<th>Class Midpoint $X_j$</th>
<th>Frequency $f_j$</th>
<th>Product $X_jf_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25–29.99</td>
<td>27.5</td>
<td>4</td>
<td>110.0</td>
</tr>
<tr>
<td>30–34.99</td>
<td>32.5</td>
<td>5</td>
<td>162.5</td>
</tr>
<tr>
<td>35–39.99</td>
<td>37.5</td>
<td>12</td>
<td>450.0</td>
</tr>
<tr>
<td>40–44.99</td>
<td>42.5</td>
<td>9</td>
<td>382.5</td>
</tr>
<tr>
<td>45–49.99</td>
<td>47.5</td>
<td>5</td>
<td>237.5</td>
</tr>
<tr>
<td>50–54.99</td>
<td>52.5</td>
<td>4</td>
<td>210.0</td>
</tr>
<tr>
<td>55–59.99</td>
<td>57.5</td>
<td>1</td>
<td>57.5</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
<td>1610.0</td>
</tr>
</tbody>
</table>

$$\bar{X}_w = \frac{\sum X_jf_j}{n} = \frac{1610.0}{40} = 40.25$$
Measures of Dispersion

- Calculations that depict the amount of spread or variability in a set of data values of a variable
  - There are several
    - The range
    - The interquartile range
    - The variance
    - The standard deviation
    - The coefficient of variation
The Range

• The simplest measure of variability or dispersion
  – Simply subtract the minimum score from the maximum score
  – Precipitation data = 57.54 – 26.87 = 30.67

• Weakness is that extreme scores can be misleading.
  – Let’s say we add year 41 & its precipitation is 72.13 inches
  – The range is now 72.13 – 26.87 = 45.26
  – But most of the scores are not different than w/o year 41
Interquartile Range (IQR)

• The difference between the 75\textsuperscript{th} and the 25\textsuperscript{th} percentile... the “middle half of the data”
  – Precip. data 75\textsuperscript{th} = P30.75 = 45.10 inches
  – 25\textsuperscript{th} = P10.25 = 35.12 inches
  – IQR = 9.98 inches

• Add a 41\textsuperscript{st} year to the dataset at 72.13 and
  – 75\textsuperscript{th} = 45.82; 25\textsuperscript{th} = 35.145; IQR = 10.675 inches
  – So the IQR did not change as much as the range
When to use...

- Use the range and IQR as non-parametric descriptions *when you do not want to assume the data are symmetrical*

- Very handy to use with the 5 points of data summary & boxplots
Parametric Measures of Dispersion

- These rely on **deviation** from the mean of scores for cases in a sample
  - Notation \( X_i - \overline{X} \) where \( X_i \) is a score for a case
  - To calculate the “deviation” of a score from the mean you simply subtract the mean from it
    - If the score is > the mean it will be +
    - If the score is < the mean it will be –

- We are interested in a **summary** of how all scores for cases together vary about the mean
  - How can it be determined?
The Variance

• The average **squared deviation** of scores around \( \bar{X} \)

• What is an average (mean)?
  
  – The sum of scores \( \div \) the \( n \) of cases
  
  – But we are not interested in *average scores* (here), but *average deviation of scores* from the mean

\[
S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}
\]
These equations are similar

\[ \bar{X} = \frac{\sum X_i}{n} \]

Sum of scores for all cases

Number of cases (sample size)

\[ S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1} \]

Sum of deviations for all cases

Number of cases

Two questions:
1) Why n-1?
2) Why do we square S?
The Variance

• The sum of raw deviations is always 0
  – Of no use to us because – outweigh +
  – Must square to get rid of –

<table>
<thead>
<tr>
<th>Pencil</th>
<th>Length (inches)</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10 - 3.7 = 6.3</td>
<td>39.69</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4 - 3.7 = 0.3</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2 - 3.7 = -1.7</td>
<td>2.89</td>
</tr>
<tr>
<td>4</td>
<td>1.5</td>
<td>1.5 - 3.7 = -2.2</td>
<td>4.84</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1 - 3.7 = -2.7</td>
<td>7.29</td>
</tr>
</tbody>
</table>

Mean: 3.7
Sum: 0

\[ S^2 = \frac{54.8}{4} = 13.7 \]

– Then we divide by sample size (n-1)

UNT Geog 3190, Wolverton
The Variance

• So for two samples with the same mean...
• The higher the variance, the greater the dispersion around the mean
  – It is always a positive number

• But, the variance is in “units$^2$”
  – It is average squared deviation
  – Inches$^2$ in our pencil example

• It is of greater use to us if we change that
The Standard Deviation

• We can take the square root of the variance to get *back to the original units*
  • We still have a good measure of dispersion because we summed *before* we took the square root

• Called the **Standard Deviation**
  • $S^2 = 13.7$, $S = 3.701$ inches
    • A pencil of 7.401 inches is 1 $S$ above the mean (= 3.7)
    • This is a better measure of average deviation (not squared)
An example

• Let’s say you sample two counties (A & B) for chlorine-contaminated water wells
  - Same mean of 1250 mg/L
  - So, it should require the same effort to clean up both counties, correct?
    • Not necessarily... $S_A = 125$ mg/L but for $S_B = 410$ mg/L
    • Because $S_B$ is quite a bit higher, there are higher (and lower) deviants in the sample (the sample is more dispersed)
      - Those high concentration sites may be very difficult to clean up

• But this direct comparison with $S$ only works if the means are similar...
### Test 1

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores A</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>80</td>
<td>30</td>
<td>900</td>
</tr>
<tr>
<td>Achilles</td>
<td>65</td>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>Zues</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aphrodite</td>
<td>35</td>
<td>-15</td>
<td>225</td>
</tr>
<tr>
<td>Hercules</td>
<td>20</td>
<td>-30</td>
<td>900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>2250</td>
</tr>
</tbody>
</table>

\[
\bar{s}^2 = 562.5, \quad \bar{s} = 23.71708245
\]

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores B</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>51</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Apollo</td>
<td>50.5</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>Hera</td>
<td>50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Socrates</td>
<td>49.5</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>Neptune</td>
<td>49</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>2.5</td>
</tr>
</tbody>
</table>

\[
\bar{s}^2 = 0.625, \quad \bar{s} = 0.79056942
\]

### Test 2

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores A</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>90</td>
<td>37.4</td>
<td>1398.76</td>
</tr>
<tr>
<td>Achilles</td>
<td>65</td>
<td>12.4</td>
<td>153.76</td>
</tr>
<tr>
<td>Zues</td>
<td>50</td>
<td>-2.6</td>
<td>6.76</td>
</tr>
<tr>
<td>Aphrodite</td>
<td>35</td>
<td>-17.6</td>
<td>309.76</td>
</tr>
<tr>
<td>Hercules</td>
<td>23</td>
<td>-29.6</td>
<td>876.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.6</td>
<td>0</td>
<td>2745.2</td>
</tr>
</tbody>
</table>

\[
\bar{s}^2 = 686.3, \quad \bar{s} = 26.19732811
\]

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores B</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>100</td>
<td>47.4</td>
<td>2246.76</td>
</tr>
<tr>
<td>Apollo</td>
<td>92.5</td>
<td>39.9</td>
<td>1592.01</td>
</tr>
<tr>
<td>Hera</td>
<td>88</td>
<td>35.4</td>
<td>1253.16</td>
</tr>
<tr>
<td>Socrates</td>
<td>85</td>
<td>32.4</td>
<td>1049.76</td>
</tr>
<tr>
<td>Neptune</td>
<td>49</td>
<td>-3.6</td>
<td>12.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.9</td>
<td>0</td>
<td>6154.65</td>
</tr>
</tbody>
</table>

\[
\bar{s}^2 = 391.05, \quad \bar{s} = 19.7749842
\]
The Coefficient of Variation

- We can make $S$ comparative by using it to calculate the CV

$$CV = \frac{S}{X} \cdot 100\%$$

- The coefficient of variation expresses sample standard deviation as a percentage of the sample mean
  - Answers, relative to the mean, how large is $S$?
  - Because it is a relative measure, samples with unequal means can be compared in terms of dispersion
## Test 2

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores A</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>90</td>
<td>37.4</td>
<td>1398.76</td>
</tr>
<tr>
<td>Achilles</td>
<td>65</td>
<td>12.4</td>
<td>153.76</td>
</tr>
<tr>
<td>Zues</td>
<td>50</td>
<td>-2.6</td>
<td>6.76</td>
</tr>
<tr>
<td>Aphrodite</td>
<td>35</td>
<td>-17.6</td>
<td>309.76</td>
</tr>
<tr>
<td>Hercules</td>
<td>23</td>
<td>-29.6</td>
<td>876.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.6</td>
<td>0</td>
<td>2745.2</td>
</tr>
</tbody>
</table>

\[ S^2 \text{ CV} \]
\[
\begin{array}{ccc}
S^2 & S & CV \\
686.3 & 26.19732811 & 49.8048063
\end{array}
\]

<table>
<thead>
<tr>
<th>Student</th>
<th>Test Scores B</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>100</td>
<td>47.4</td>
<td>2246.76</td>
</tr>
<tr>
<td>Apollo</td>
<td>92.5</td>
<td>39.9</td>
<td>1592.01</td>
</tr>
<tr>
<td>Hera</td>
<td>88</td>
<td>35.4</td>
<td>1253.16</td>
</tr>
<tr>
<td>Socrates</td>
<td>85</td>
<td>32.4</td>
<td>1049.76</td>
</tr>
<tr>
<td>Neptune</td>
<td>49</td>
<td>-3.6</td>
<td>12.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>Sum</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>82.9</td>
<td>0</td>
<td>6154.65</td>
</tr>
</tbody>
</table>

\[ S^2 \text{ CV} \]
\[
\begin{array}{ccc}
S^2 & S & CV \\
391.05 & 19.7749842 & 23.85402195
\end{array}
\]

UNT Geog 3190, Wolverton
Why describe dispersion (variability)

• Allows us to determine between two or more samples, which one has a broader dispersion of scores
  – In grading, if one lab is highly variable and one is not there might be many reasons
    • Attendance is poor in the early lab, but those who come to class get more attention (fewer students) = broad variability in scores
    • More people come to the second lab, but there is less attention = narrow variability
  • Or, the professor is tired early on, and only mathematically gifted students understand him/her, but others do poorly = broad variability
  • The second lab gets the professor when he/she is more awake, and all students do better with out the broad dispersion of scores = narrow variability

• Allows us to target reasons for differences between samples (deer size)
Basic Descriptives in SPSS
5 pts of data summary (& range)

1. SPSS Data Editor screen showing variables and data entries.
2. Frequencies: Statistics dialog box selecting statistics to compute.
3. Confirming selection with the ok button.
Now, calculate the IQR

Then, redo the descriptives using parametric statistics

<table>
<thead>
<tr>
<th>N</th>
<th>Valid</th>
<th>Missing</th>
<th>Median</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200.00</td>
<td>280.00</td>
<td>145.00</td>
<td>425.00</td>
<td>175.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200.00</td>
<td></td>
<td></td>
<td></td>
<td>240.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>75</td>
</tr>
</tbody>
</table>