Variables, distributions, and samples
(cont.)

Phil 12: Logic and Decision Making
Fall 2010
UC San Diego
10/18/2010
Review

• Recording observations
  – Must extract that which is to be analyzed: coding systems, etc.

• Analyze observations in terms of **variables**: characteristic or feature that varies and takes on different values

• Four types of variables
  – Nominal ordinal interval ratio

• Values of variables are distributed
  – Important goal: characterizing the distribution
Nominal & ordinal variables and bar graphs

- Example: Profile of pet ownership in San Diego County

- Value of graphs: provide an intuitive appreciation of the data

- **Bar graphs** and **pie charts** work well with nominal and ordinal variables
Score variables and histograms

- Since score variables are continuous, **histograms** rather than bar graphs are used.
- This is done by creating **bins** and tabulating the number of items in each bin.
- The size of bins can create radically different pictures of the distribution!
Normal and non-normal distributions

- **Normal distributions**
  - Have a single peak
  - Scores equally distributed around the peak
  - Fewer scores further from the peak

- **Non-normal distributions:**
  - Skewed
  - Bimodal
Clicker question 1

The distribution below is

<table>
<thead>
<tr>
<th></th>
<th>0-99</th>
<th>100-199</th>
<th>200-299</th>
<th>300-399</th>
<th>400-499</th>
<th>500-599</th>
<th>600-699</th>
<th>700-799</th>
<th>800-899</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>63</td>
<td>45</td>
<td>35</td>
<td>37</td>
<td>82</td>
<td>35</td>
<td>39</td>
<td>53</td>
<td>53</td>
</tr>
</tbody>
</table>

A. Normal since it has one peak
B. Normal since scores are equally distributed around the peak
C. Not normal since because there are not fewer scores further from the peak
D. Not normal because scores are not equally distributed around the peak
Describing distributions

• Two principal measures:

1. Central tendency

Two comparable distributions differing in central tendency

2. Variability

Two distributions with same central tendency but differing in variability
Three measures of central tendency

**Mean:** the arithmetic average--sum of all the scores divided by the number of instances

**Median:** the score of which half are higher and half are lower

**Mode:** the most frequent score

Consider this distribution of values:

2, 6, 9, 7, 9, 9, 10, 8, 6, 7

mean = 73 / 10 = 7.3

median = 7.5

mode = 9
Which measure to use?

• If the distribution is normal, all three measures of central tendency give the same result
  - The mean is the easiest to calculate and the most frequently reported

• If there are extreme outliers in one direction, the mean may be distorted
  - Exam scores: 21, 72, 76, 79, 82, 84, 87, 88, 90, 91, 95
    • Mean: 78.6
    • Median: 84
  - In such a case, the median gives a better picture of the central tendency of the class
Measures of variability

- Variability concerns: How much do the scores vary?
- **Range**: the lowest value to the highest value
Measures of variability

• Variability concerns: How much do the scores vary?

• **Range**: the lowest value to the highest value

• **Variance**: \[ \frac{\sum (X-\text{mean})^2}{N} \]

• **Standard deviation**: \[ \sqrt{\text{Variance}} \]
Measures of variability

- Variability concerns: How much do the scores vary?
- **Range**: the lowest value to the highest value
- **Variance**: \[\frac{\sum (X - \text{mean})^2}{N}\]
- **Standard deviation**: \[\sqrt{\text{Variance}}\]

Intuitive interpretation:
- One standard deviation: the part of the range in which 68% of the scores fall
- Two standard deviations: the part of the range in which 95% of the scores fall
- Three standard deviations: the part of the range in which 99% of the scores fall
Variance

• Consider a distribution:

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>6</th>
<th>7</th>
<th>7</th>
<th>8</th>
<th>Mean = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-μ</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>X - mean</td>
</tr>
<tr>
<td>(X-μ)^2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>(X-mean)^2</td>
</tr>
</tbody>
</table>

\[
\text{Variance} = \frac{\sum (X-\text{mean})^2}{N} = \frac{12}{9} = 1.33
\]

\[
\text{SD} = \sqrt{\text{Variance}} = \sqrt{1.33} = 1.15
\]

Range of 1 SD = 6 ± 1.15 = 4.85 to 7.15
Range of 2 SD = 6 ± 2.30 = 3.70 to 8.30
Range and Standard Deviation

range

-2 SD  -1 SD  Mean  1 SD  2 SD

Monday, October 18, 2010
Clicker question 2

On an exam on which scores were distributed normally and the mean was 86 and the SD was 4,

A. 68% of the scores were between 82 and 90
B. 95% of the scores were between 82 and 90
C. 99% of the scores were between 78 and 94
D. 68% of the scores were between 78 and 94
Populations

• The phenomena about which we seek to draw conclusions in a study are known as the population.

• Sometimes one can study each member of the population of interest

• But if the population is large:
  - it may be impossible to study the whole population
  - there may be no need to study the whole population
Samples

- A **sample** is a subset of the population chosen for study.

- From studying the distribution of a variable in a sample, one makes an **estimate** of the distribution in the actual population.

- Sometimes the estimate from a sample may be more accurate than trying to study the population itself.

  - U.S. Census
Is the sample biased?

• If information about the sample is to be informative about the actual population, the sample must be representative

- **Randomization**: attempt to insure that the sample is representative by avoiding bias in selecting the sample

• Risk: inadvertently developing a misrepresentative sample

  - E.g., using telephone numbers in the phonebook to sample electorate
Does the sample reflect the population?

- Does the mean of the sample reflect the mean of the actual population?
  - Sampling distribution simulation
  - Very unlikely that the mean of the sample will exactly equal the mean of the population
  - Key question: how much does the mean of the sample vary from the mean of the actual population?

- Given the mean of a sample, what is the range within which the mean of the actual population lies?
  - To determine this, the standard deviation measure is very useful
Standard deviation and mean

In $\approx 68\%$ of samples, the **mean of the population** will fall within 1 standard deviation of the **mean of the sample**

In $\approx 95\%$ of samples, the **mean of the population** will fall within 2 standard deviations from the **mean of the sample**
What happens as sample size gets larger?

- As sample size grows, the SD of the sample shrinks
- So with larger samples, the range of 2 standard deviations shrinks

Assume sample mean is 50:

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Range of 2 SD (95% confidence interval)</th>
<th>Range of 3 SD (99% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>34.5-65.5</td>
<td>29.5-70.5</td>
</tr>
<tr>
<td>20</td>
<td>39-61</td>
<td>35.6-64.4</td>
</tr>
<tr>
<td>50</td>
<td>43-57</td>
<td>40.9-59.1</td>
</tr>
<tr>
<td>100</td>
<td>45-55</td>
<td>43.5-56.5</td>
</tr>
<tr>
<td>500</td>
<td>47.8-52.2</td>
<td>47.1-52.9</td>
</tr>
<tr>
<td>1000</td>
<td>48.4-51.6</td>
<td>48-52</td>
</tr>
</tbody>
</table>
Example of estimating population mean from sample mean

- Example: age of person eating at the Food Court

  - Draw a sample to make inference of average age of person eating at the Food Court

<table>
<thead>
<tr>
<th></th>
<th>&lt;17</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>&gt;25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>6</td>
<td>18</td>
<td>23</td>
<td>34</td>
<td>32</td>
<td>18</td>
<td>26</td>
<td>29</td>
<td>14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Sample</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Estimating real distribution

<table>
<thead>
<tr>
<th></th>
<th>&lt;17</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>&gt;25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>6</td>
<td>18</td>
<td>23</td>
<td>34</td>
<td>32</td>
<td>18</td>
<td>26</td>
<td>29</td>
<td>14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Sample 1 (n = 10)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 2 (n=20)</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean of the actual population: 20.63

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of the sample:</td>
<td>19.4</td>
<td>20.1</td>
</tr>
<tr>
<td>SD of the sample:</td>
<td>1.9</td>
<td>1.6</td>
</tr>
<tr>
<td>Range of 1 SD:</td>
<td>17.5-22.3</td>
<td>18.5-21.7</td>
</tr>
<tr>
<td>Range of 2 SD:</td>
<td>15.9-24.2</td>
<td>16.9-23.3</td>
</tr>
</tbody>
</table>

Want to predict more accurately?

Use a larger sample size
Review

• Four types of variables:
  - Nominal  ordinal  interval  ratio
• Values of variables are distributed
  - Important goal: characterizing the distribution
• Graphs
  - Bar graphs for nominal and ordinal variables
  - Histograms for score variables
• Normal versus non-normal distributions
  - Skewed, bimodal, etc
Review

• Two principal measures of distributions
  - Central tendency
    • Mean, median, mode
  - Variability
    • Range, variance, SD
      - 1 SD includes approx. 68% of scores
      - 2 SD includes approx. 95% of scores
      - 3 SD includes approx. 99% of scores
Review

- Population and samples
  - From studying the distribution in sample, estimate the distribution in the actual population
  - Mean of actual population will
    - Fall within one SD of mean of sample for 68% of samples
    - Fall within two SD of mean of sample for 95% of samples
    - Fall within three SD of mean of sample for 99% of samples
  - Larger sample yields smaller SD and hence more precise estimate
  - Hence, to improve the precision of an estimate, use a larger sample