Data

• **Experimental unit**: A person, animal, object, on which we collect data. Ex.: houses.

• **Variable**: A characteristic of the experimental unit with outcomes that vary across units. Ex.: number of bedrooms, assessed price, location (neighborhood).

• **Quantitative variables**: Can be measured on a numerical scale. Ex.: number of bedrooms, assessed price.

• **Qualitative variables**: Can only be classified into groups or categories. Ex.: location.

• Example: model the sale price of houses as a function of assessed price, number of bedrooms and neighborhood.
Populations and samples

- **Population:** Collection of data measured on all experimental units of interest. Ex.: the population of registered voters in the US, or the population of salaries paid to all graduating MBAs in 2004.

- We are typically interested in making inferences about some population.

- Except in trivial cases, populations are too large to measure. Instead, we collect a *sample* from the population of interest.

- **Sample:** A subset of data selected from the population.

- $n$ is sample size, $N$ is population size (if known).

- **Inference:** An estimate, prediction, or other generalization about a population that is based on sample information.
Sampling and reliability

• Because we generalize from a sample to a population, inferences are uncertain.

• Must report **reliability** of each inference. Ex.: $43\% \pm 2\%$ of consumers prefer brand X.

• Reliability increases when
  1. Sample is *representative*: exhibits characteristics typical of the population.
  2. Sample size increases: less uncertainty with larger samples, with caveats.

• Easiest sampling method to obtain representative sample is *simple random sampling*: each unit in the population has equal probability of selection.

• Other sampling approaches: unequal probability of selection, stratified sampling, others.
Describing qualitative data

- Examples of qualitative variables: product brand, types of cancer, gender, zip code, tax bracket, ethnic group, country, industrial sector, age category.

- Qualitative data can be classified into classes or categories.

- **Class frequency**: Number of observations in the sample in each category.

- **Relative frequency**: proportion of observations in each category:

\[
\text{Class relative freq.} = \frac{\text{Class frequency}}{n}.
\]
Example: Bladder cancer experiment

• Clinical trial for bladder cancer drugs.

• 307 patients randomized to three treatments: 135 to Placebo, 89 to Pyridoxine, and 83 to Thiotepa.

• Three causes of death during treatment recorded: bladder cancer, other causes, or patient alive.

• Data summary:

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Bladder Cancer</th>
<th>Other</th>
<th>Alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>135</td>
<td>4</td>
<td>24</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3%)</td>
<td>(18%)</td>
<td>(79%)</td>
</tr>
<tr>
<td>Pyridoxine</td>
<td>89</td>
<td>0</td>
<td>15</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0%)</td>
<td>(17%)</td>
<td>(83%)</td>
</tr>
<tr>
<td>Thiotepa</td>
<td>83</td>
<td>2</td>
<td>27</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2%)</td>
<td>(33%)</td>
<td>(65%)</td>
</tr>
</tbody>
</table>
Quantitative data - graphics

- We summarize graphically quantitative data via stem-and-leaf plots and histograms.
- Stem-and-leaf plots used for small datasets.
- Example: 20 final grades in Stat 328 are 67, 75, 98, 89, 86, 94, 92, 93, 76, 79, 90, 68, 42 (yikes), 68, 89, 74, 87, 67, 71, 78.
- Stems are 4, 5, 6, 7, 8, 9 and leaves are 0,...,9:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0,2,3,4,8</td>
</tr>
<tr>
<td>8</td>
<td>6,7,9,9</td>
</tr>
<tr>
<td>7</td>
<td>1,4,5,6,8,9</td>
</tr>
<tr>
<td>6</td>
<td>7,7,8,8</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Stem 7 has the most leaves, so that grades in the 70s were the most common.
Quantitative data - graphics (cont’d)

• For larger datasets, we use *histograms*.

• As in stem-and-leaf plots, histograms display the frequency of observations in specified classes (or categories).

• Example: Sale prices (in $1000) of over 1,000 homes sold in Tampa, FL, in 1999.
Quantitative data - numerics

• **Mean:** Measures the *central tendency* in the data.

• Let $y_1, y_2, \ldots, y_n$ denote the $n$ sample observations. Then

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = n^{-1} \sum_{i=1}^{n} y_i$$

is the *sample mean*.

• The sample mean is an estimator of the *population mean* or *expected value of $y$* $E(y)$, which we denote by $\mu$.

• Baby example: sample observations are $(y_1 = 5, y_2 = 8, y_3 = 4, y_4 = 3, y_5 = 5)$:

$$\bar{y} = \frac{5 + 8 + 4 + 3 + 5}{5} = 6.0$$
Quantitative data - numerics (cont’d)

• **Median:** Another measure of central tendency. In a population, the median is the value of the variable such that half of the population is below and half is above that value.

• The *sample median* is computed as:
  1. Order observations: $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$
  2. For $n$ odd, the median is the $(\text{Int}(n/2) + 1)$th value of $y$.
  3. For $n$ even, median is the average of two middle values.

• Baby example again: $(y_1 = 5, y_2 = 8, y_3 = 4, y_4 = 3, y_5 = 5)$:
  1. Order observations: $(3, 4, 5, 5, 8)$
  2. Median is middle value: 5.

• If sample had been $(y_1 = 5, y_2 = 8, y_3 = 4, y_4 = 3)$, then median is

$$\text{Median} = \frac{4 + 5}{2} = 4.5.$$
• Measures of the *spread* or *variability* in the data are range, variance, and standard deviation.

• **Range:** Difference between largest and smallest sample observation.

• **Variance:** Defined as the average of the squared deviations of observations from the mean.

• For a population, variance is defined as:

\[ \sigma^2 = E[(y - \mu)^2]. \]

• For a sample \( y_1, y_2, \ldots, y_n \), we compute the sample variance as

\[ S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}. \]

• Sample variance is sometimes denoted \( \hat{\sigma}^2 \).

• Variance is *always* positive (0 only when all observations are identical), but *never* negative.

• **Standard deviation:** Is the positive square root of the variance.

  – Population standard deviation: \( \sigma = +\sqrt{\sigma^2} \).
  – Sample standard deviation: \( S = +\sqrt{S^2} \).
Range and variance (con’t)

• Note that:

\[ \frac{\sum_i (y_i - \bar{y})^2}{n - 1} = \frac{\sum_i (y_i^2 - 2y_i\bar{y} + \bar{y}^2)}{n - 1} = \frac{\sum_i y_i^2 - 2\bar{y}\sum_i y_i + n\bar{y}}{n - 1}. \]

But

\[ \sum_i y_i = n\bar{y}, \]

so that

\[ \frac{\sum_i y_i^2 - 2\bar{y}\sum_i y_i + n\bar{y}}{n - 1} = \frac{\sum_i y_i^2 - 2n\bar{y}^2 + n\bar{y}^2}{n - 1} = \frac{\sum_i y_i^2 - n\bar{y}^2}{n - 1}. \]

• Baby example with four observations \((y_1 = 5, y_2 = 8, y_3 = 4, y_4 = 3)\).

Range = 8 - 3 = 5.

\[ S^2 = \frac{(5 - 5)^2 + (8 - 5)^2 + (4 - 5)^2 + (3 - 5)^2}{4 - 1} = \frac{0 + 9 + 1 + 4}{3} = 4.67. \]

• Can also compute as: \(S^2 = [(25 + 64 + 16 + 9) - 4 \times 25]/3.\)

• Finally, \(S = +\sqrt{S^2} = 2.16.\)
Standard deviation

- In any dataset, at least 75% of the observations will lie within two standard deviations of the mean. More formally:

\[ \text{Prob. obs. in } (\bar{y} - 2 \times S, \bar{y} + 2 \times S) \geq 0.75. \]

- If data are distributed symmetrically about the mean, then about 95% of the observations are within two standard deviations of the mean.

Hotel no-shows example:

- Question: How many rooms can the hotel over-book and still be confident that all reservations can be honored?

- Data: Number of no-shows over 30 days at large hotel (Table 1.6 in book).

- Sample mean: \( \bar{y} = 15.133 \) and sample standard deviation \( S = 2.945 \). Distribution of values approximately symmetric around mean.

- About 95% of number of no-shows lie in \( \bar{y} + / - 2S = (9.24, 21.02) \).

- The hotel can safely overbook 10 rooms per day and be highly confident that all reservations can be honored.