AP Statistics Review

Part I & Part II & Part III:
Exploring and Understanding Data
Exploring Relationships Between Variables
Gathering Data
Part I:
Chapters 2 - 6

EXPLORING AND UNDERSTANDING DATA
GRAPHICAL DISPLAYS
Frequency Distributions

- Tabular display of data.
- Both qualitative and quantitative data.
- Summarize the data.
- Help ID distinctive features.
- Used to graph data.
- Categories/Classes – non-overlapping, each datum falls into only one.
- Frequency – number of counts in each category/class.
- Relative Frequency – fraction or ratio of category/class frequency to total frequency.
Example: Final grades for 42 students in a Basic Algebra class were as follows:

73, 81, 70, 71, 74, 74, 72, 72, 87, 90, 87, 84, 70, 73, 68, 74, 84, 80, 73, 58, 89, 71, 79, 67, 81, 78, 82, 78, 88, 80, 75, 85, 72, 83, 72, 90, 83, 72, 91, 82, 53, 84

The range is 91 − 53 or 38. The table is based on 8 classes. The class width is \( \frac{38}{8} \) or 4.75, which rounds to 5.

**Final Grades for 42 Basic Algebra Students**

<table>
<thead>
<tr>
<th>Class</th>
<th>Scores</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Frequency</th>
<th>Relative Cum. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52–56</td>
<td>1</td>
<td>( \frac{1}{42} = 0.024 )</td>
<td>1</td>
<td>0.024</td>
</tr>
<tr>
<td>2</td>
<td>57–61</td>
<td>1</td>
<td>0.024</td>
<td>2</td>
<td>0.048</td>
</tr>
<tr>
<td>3</td>
<td>62–66</td>
<td>0</td>
<td>0.000</td>
<td>2</td>
<td>0.048</td>
</tr>
<tr>
<td>4</td>
<td>67–71</td>
<td>6</td>
<td>0.143</td>
<td>8</td>
<td>0.190</td>
</tr>
<tr>
<td>5</td>
<td>72–76</td>
<td>12</td>
<td>0.286</td>
<td>20</td>
<td>0.476</td>
</tr>
<tr>
<td>6</td>
<td>77–81</td>
<td>7</td>
<td>0.167</td>
<td>27</td>
<td>0.643</td>
</tr>
<tr>
<td>7</td>
<td>82–86</td>
<td>8</td>
<td>0.190</td>
<td>35</td>
<td>0.833</td>
</tr>
<tr>
<td>8</td>
<td>87–91</td>
<td>7</td>
<td>0.167</td>
<td>42</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Total** | **42** | **1.000** |
Graphs of Categorical Data

- Bar Chart/Graph
- Pie Chart/Graph
- Describe the distribution in the CONTEXT of the data.
- Not appropriate to describe the shape of the distribution. Descriptions such as “symmetric” or “skewed” would not make sense, since the ordering of the categories is arbitrary.
Bar Graph

- Have spaces between each category.
- Order of the categories not important.
- Either frequency (counts) or relative frequency (proportions) can be shown on the y-axis.
- Title, both axes labeled and have appropriate scales.
Pie Chart

- Commonly used for presenting relative frequency distributions for qualitative data.
- Slice the circle into pieces whose size is proportional to the fraction of the whole in each category.
- Title, label sectors (included proportion).
Two-Variable Categorical Data

- Contingency Table (2-way table)
  - Conditional Distributions
  - Marginal Distributions
- Segmented Bar Graphs
  - Display association between variables
Contingency Table

<table>
<thead>
<tr>
<th>Party</th>
<th>Class level</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freshman</td>
<td>Sophomore</td>
<td>Junior</td>
<td>Senior</td>
<td></td>
</tr>
<tr>
<td>Democratic</td>
<td>0.167</td>
<td>0.267</td>
<td>0.417</td>
<td>0.429</td>
<td>0.325</td>
</tr>
<tr>
<td>Republican</td>
<td>0.667</td>
<td>0.533</td>
<td>0.333</td>
<td>0.286</td>
<td>0.450</td>
</tr>
<tr>
<td>Other</td>
<td>0.167</td>
<td>0.200</td>
<td>0.250</td>
<td>0.286</td>
<td>0.225</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
<td><strong>1.000</strong></td>
</tr>
</tbody>
</table>

### Association

- The variables “political party affiliation” and “class level” are associated because knowing the value of the variable “class level” imparts information about the value of the variable “political party affiliation. If we do not know the class level of a student in the course, there is a 32.5% chance that the student is a Democrat. But, if we know that the student is a junior, there is a 41.7% chance that the student is a Democrat.

- If the variables “political party affiliation” and “class level” were not associated, the four conditional distributions of political party affiliation would be the same as each other and as the marginal distribution of political party affiliation; in other words, all five columns would be identical.
Segmented Bar Graphs

- **Association**
  - A segmented bar graph lets us visualize the concept of association. The first four bars of the segmented bar graph show the conditional distributions and the fifth bar gives the marginal distribution of political party affiliation.
  - If political party affiliation and class level were not associated, the four bars displaying the conditional distributions of political party affiliation would be the same as each other and as the bar displaying the marginal distribution of political party affiliation; in other words, all five bars would be identical. That political party affiliation and class level are in fact associated is illustrated by the nonidentical bars.
Quantitative Data

• One-Variable
• Graphs
  • Histogram
  • Ogive
  • Stem-and-Leaf Plot
  • Dotplot
• Group data into classes of equal width.
• The counts in each class is the height of the bar.
• Describe distribution by; shape, center, spread.
• Unusual features should also be noted; gaps, clusters, outliers.
• Relative freq. and freq. histograms look the same except the vertical axis scale.
• Remember to describe the shape, center, and spread in the CONTEXT of the problem.
Common Distribution Shapes

(a) Bell-shaped
(b) Triangular
(c) Uniform (or rectangular)
(d) Reverse J-shaped
(e) J-shaped
(f) Right skewed
(g) Left skewed
(h) Bimodal
(i) Multimodal
Stem-and-Leaf Plot

- Contains all the information of histograms.
- Advantage – individual data values are preserved.
- Used for small data sets.
- The leading digit(s) are the “stems,” and the trailing digits (rounded to one digit) are the “leaves.”
- Back-to-Back Stem-and-leaf Plots are used to compare related data sets.
Stem-and-Leaf Plot

Example: 5.26, 5.78, 5.82, 6.03, 6.22, 6.47, 6.49, 6.77, 7.10, 7.35, 7.37, 7.46, 8.20, 8.89, 10.66, 12.42

Title

<table>
<thead>
<tr>
<th>Stems</th>
<th>Lengths of Widgets (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Machine A</td>
</tr>
<tr>
<td>5</td>
<td>3 8 8</td>
</tr>
<tr>
<td>6</td>
<td>0 2 5 5 8</td>
</tr>
<tr>
<td>7</td>
<td>1 4 4 5</td>
</tr>
<tr>
<td>8</td>
<td>2 9</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
</tbody>
</table>

Leaves are ordered (rounded to the nearest tenth).

Key (Don't forget your units.)

5 | 3 = 5.3 cm (rounded)
Back-to-Back Stem-and-Leaf Plot

<table>
<thead>
<tr>
<th>Lengths of Widgets (CM)</th>
<th>Machine B</th>
<th>Machine A</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 3 2</td>
<td>5</td>
<td>3 8 8</td>
</tr>
<tr>
<td>9 9 8 4 3 3</td>
<td>6</td>
<td>0 2 5 5 8</td>
</tr>
<tr>
<td>2 1</td>
<td>7</td>
<td>1 4 4 5</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>2 9</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

5 | 3 = 5.3 cm (rounded)
Dataplot

- Quick and easy display of distribution.
- Good for displaying small data sets.
- Individual data values are preserved.
- Construct a dotplot by drawing a horizontal axis and scale. Then record each data value by placing a dot over the appropriate value on the horizontal axis.
DESCRIBING DISTRIBUTIONS NUMERICALLY
Five-Number Summary

• Minimum value, Quartile 1 (Q1) (25\textsuperscript{th} percentile), median, Quartile 3 (Q3) (75\textsuperscript{th} percentile), maximum. In that order.

• Boxplot is a visual display of the five-number summary.

• Interquartile Range (IQR) – difference between the quartiles, IQR = Q3 – Q1. Used as a measure of spread.
Checking for Outliers

- Values that are more than 1.5 times the IQR below Q1 or Q3 are outliers.
- Calculate upper fence: Q3 + 1.5(IQR)
- Calculate lower fence: Q1 – 1.5(IQR)
- Any value outside the fences is an outlier.
Boxplot

- A box goes from the Q1 to Q3.
- A line is drawn inside the box at the median.
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier and from the upper end of the box to the largest observation that is not a potential outlier.
- The potential outliers are shown separately.
- Title and number line scale.
Side-by-Side Boxplot

- Box Plots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions.
- Allows us to see which distribution has the higher median, which has the greater IQR and which has the greater overall range.
Measures of Center

• **Mean:** \( \bar{x} = \frac{\text{sum of values}}{\text{number of values}} \)

  • Good measure of center when the shape of the distribution is approximately unimodal and symmetric.
  • Non-resistant.

• **Median:** The middle of a *ordered* set of data.

  • Resistant.

• **Note:** While the median and the mean are approximately equal for unimodal and symmetric distributions, there is more that we can do and say with the mean than with the median. The mean is important in inferential statistics.
Relationship Measures of Center

- **Left-Skewed**
  - Mean
  - Median
  - Mode

- **Symmetric**
  - $\text{Mean} = \text{Median} = \text{Mode}$

- **Right-Skewed**
  - Mode
  - Median
  - Mean
Measures of Spread (Variation)

- **Standard Deviation**: $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$
  - The square root of the average of the deviations from the mean. Contains the mean, so is non-resistant.
  - The square root of the variance.
  - Used for unimodal, symmetric data.
  - Use when using the mean as the measure of center.
  - Will equal zero only if all data values are equal.

- **Interquartile Range (IQR)**: $Q3 - Q1$
  - Gives the spread of the middle 50%.
  - B/C it doesn’t use extreme values, it is resistant.
  - Used when outliers are present or with skewed data.
  - Use when using the median as the measure of center.

- **Range**: Max. value – Min. value.
  - Single number and very sensitive to extreme values.
  - Supplementary piece of info, not a stand alone measure of spread.
<table>
<thead>
<tr>
<th>Measure of Center</th>
<th>Corresponding Measure of Spread</th>
<th>Resistant to Extreme Values?</th>
<th>When Is It Appropriate?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Standard Deviation</td>
<td>NO</td>
<td>Data are approximately unimodal and symmetric (bell-shaped).</td>
</tr>
<tr>
<td>Median</td>
<td>IQR</td>
<td>YES</td>
<td>Data are strongly skewed or contain outliers.</td>
</tr>
</tbody>
</table>
Standard Deviation as a Ruler: Z-scores

- **Z-score**: Standardized value, using units of standard deviation.

  \[ z = \frac{x - \mu}{\sigma} \]

- Standardizing – shifts the data by subtracting the mean and rescales the values by dividing by the SD.

- Adding (or subtracting) a constant to each value of a data set adds (or subtracts) the same constant to the mean or median. Measures of spread (SD and IQR) remain unchanged.

- Multiplying (or dividing) a constant to each value of a data set changes both the measures of center (mean and median) and spread (SD and IQR). These measures are multiplied (or divided) by the same constant.
Normal Models

• Distributions whose shapes are unimodal and roughly symmetric (bell-shaped).
• Described by 2 parameters, mean and SD. Notation: $N(\mu, \sigma)$.
• 68-95-99.7 (Empirical) Rule: Thumb rule for normal distributions.
• Standard Normal Distribution: $N(0, 1)$.
• 2 types of problems.
  • Finding normal percentiles.
  • Finding a value given a proportion.
• Assessing Normality
  • Picture – histogram, stem-and-leaf, boxplot, dotplot.
  • Normal Probability Plot on the graphing calculator.
Empirical Rule (68-95-99.7%)
What You need to Know

- Categorical vs. quantitative variables
- How to read, interpret, describe, and compare graphs
- How to compare distributions, like with a segmented bar graph (marginal/conditional distributions)
- Know that the median is the 50% mark, Q1 is 25% and Q3 is 75%
- Know how outliers affect the summary statistics
- Know the properties of the mean and standard deviation
- How to find the 1-variable stats using the calculator
- How center and spread are affected by changes in the dataset (adding 50, multiply by 10%) – shifting/scaling
What You need to Know

• How to use a freq table to estimate center and spread
• How variance and standard deviation are related
• What a standard deviation of 0 represents
• How to test for outliers and create a mod boxplot
• Which summary is best (skewed is median/IQR; symmetric is mean/st. dev)
• The Empirical Rule and how to use it
• Standard Normal curve is N(0,1)
• What a z-score means
• How to find a z-score and use it to find cutoff points and percentiles
• How to use z-scores to compare items
PRACTICE PROBLEMS
Given the first type of plot indicated in each pair, which of the second plots could not always be generated from it?

A. dot plot -> histogram  
B. stem and leaf -> dot plot  
C. dot plot -> box plot  
D. histogram -> stem and leaf plot  
E. All of these can always be generated
If the test scores of a class of 30 students have a mean of 75.6 and the test scores of another class of 24 students have a mean of 68.4, then the mean of the combined group is

a. 72
b. 72.4
c. 72.8
d. 74.2
e. None of these
If a distribution is relatively symmetric and bell-shaped, order (from least to greatest) the following positions:

1. a z-score of 1
2. the value of Q3
3. a value in the 70th percentile

a. 1, 2, 3
b. 1, 3, 2
c. 3, 2, 1
d. 3, 1, 2
e. 2, 3, 1
If each value of a data set is increased by 10%, the effects on the mean and standard deviation can be summarized as

A. mean increases by 10%; st. dev remains unchanged
B. mean remains unchanged; st. dev increases by 10%
C. mean increases by 10%; st. dev increases by 10%
D. mean remains unchanged; st. dev remains unchanged
E. the effect depends on the type of distribution
#6

If all values in a data set are converted into standard scores (z-scores) then which of the following statements is not true?

A. Conversion to standard scores is not possible for some data sets.
B. The mean and st. dev of the transformed data are 0 and 1 respectively only for symmetric and bell-shaped distributions.
C. The empirical rule applies consistently to both the original and transformed data sets.
D. The z-scores represent how many standard deviations each value is from the mean.
E. All of these are true statements.
In skewed right distributions, what is most frequently the relationship of the mean, median, and mode?

A. mean > median > mode  
B. median > mean > mode  
C. mode > median > mean  
D. mode > mean > median  
E. mean > mode > median
A random survey was conducted to determine the cost of residential gas heat. Analysis of the survey results indicated that the mean monthly cost of gas was $125, with a standard deviation of $10. If the distribution is approximately normal, what percent of homes will have a monthly bill of more than $115?

a. 34%
b. 50%
c. 68%
d. 84%
e. 97.5%
The average life expectancy of males in a particular town is 75 years, with a standard deviation of 5 years. Assuming that the distribution is approximately normal, the approximate 15th percentile in the age distribution is: (Hint: percentile rank is “at or below” that value)

a. 60
b. 65
c. 70
d. 75
e. 80
Part II
Chapters 7 - 10
EXPLORING RELATIONSHIPS BETWEEN VARIABLES
SCATTERPLOTS
Scatterplots

• Used to display the relationship between two quantitative variables.
• Explanatory or predictor variable on the x-axis.
• Response variable (the variable you hope to predict or explain) on the y-axis.
• When analyzing a scatterplot, you want to discuss:
  • Direction
  • Form
  • Strength
Direction

Positive

Negative
Form

Straight (linear)

Curved
Strength

- Association does not imply causation. The only way to assess causation is through a randomized, controlled experiment.
Correlation

• Describes a linear relationship between two quantitative variables.
• Direction (sign) and strength (value).
• Correlation Coefficient ($r$):

$$
r = \frac{1}{n - 1} \sum \left( \frac{x - \bar{x}}{s_x} \right) \left( \frac{y - \bar{y}}{s_y} \right)
$$
Facts About the Correlation Coefficient (r)

• Formula uses standardized observations, so it has no units.
• Makes no distinction between explanatory and response variables – correlation (x, y) = correlation (y, x).
• Correlation does require both variables be quantitative.
• The sign of r indicates the direction of association.
• -1≤r≤1: The magnitude of r reflects the strength of the linear association as viewed in a scatterplot. (0≤r<.25 no correlation, .25≤r<.5 weak correlation, .5≤r<.75 moderate correlation, .75≤r<1 strong correlation).
• r measures only the strength of a linear relationship. It does not describe a curved relationship.
• r is not resistant to outliers since it is calculated using the mean and SD.
• r is not affected by changes in scale or center (uses standardized values).
• A scatterplot or correlation alone cannot demonstrate causation.
Least Squares Regression Line (LSRL)

- LSRL is the line that minimizes the sum of the squared residuals.
- It is a linear model of the form:

\[
\hat{y} = b_0 + b_1 x
\]
Facts About the LSRL

- The slope is: \( b_1 = r \frac{S_y}{S_x} \)

- Every LSRL goes through the point \((\bar{x}, \bar{y})\). Substituting into the equation of the LSRL the y-intercept is: \( b_0 = \bar{y} - b_1 \bar{x} \)

- \( R^2 \), the coefficient of determination, indicates how well the model fits the data.

- \( R^2 \) gives the fraction of the variability of \( y \) that is explained or accounted for by the least squares linear regression line is in relating \( y \) to \( x \).

- Causation cannot be demonstrated by the coefficient of determination.

- Residuals are what are left over after fitting the model. They are the difference between the observed values and the corresponding predicted values.

- The sum of the residuals is always equal to zero.
Residuals

\[ \text{Res} = \text{Obs} - \text{Pred} = (y - \hat{y}) \]
Residual Plot

- The residual is the directed distance between the observed and predicted value.
- A residual plot graphs these directed distances against either the explanatory or the predicted variable.

- No regression analysis is complete without a residual plot to check that the model is reasonable.
- A reasonable model is one whose residual plot shows no discernible pattern.
- Any function is linear if plotted over a small enough interval. A residual plot will help you see patterns in the data that may not be apparent in the original graph.
Extrapolation

• Making predictions for x-values that lie far from the data we used to build the regression model is highly dangerous. There are no guarantees that the pattern we see in the model will continue.
Outliers and Influential Points

- Outliers can strongly influence regression.
- Can have outliers in the x-value, the y-value, or from the overall pattern (x and y values).
- A point has leverage and is called an influential point if its removal causes a dramatic change in the slope of the regression line.
The indicated outlier lies outside the overall pattern of the data, its removal has little effect on the slope of the regression line. It would not be considered an influential point.
Outliers and Influential Points

• The outlier in the x direction, if removed causes a dramatic change in the slope of the regression line. This point has leverage and is an influential point.
Creating and Using a LSRL

• Conditions for regression.
  • Data follow a straight-line pattern.
  • No outliers.
  • Residual plot shows no obvious patterns.
Computer Outputs

• It is necessary to be able to read computer outputs to be successful on the AP exam.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>16.04891</td>
<td>1</td>
<td>16.04895</td>
<td>52.1</td>
</tr>
<tr>
<td>Residual</td>
<td>1.539675</td>
<td></td>
<td>0.307934</td>
<td></td>
</tr>
</tbody>
</table>

- $R^2 = 91.2\%$, Adjusted $R^2 = 89.5\%$
- $s = 0.5549$ with $7 - 2 = 5$ degrees of freedom

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>s.e. of Coeff</th>
<th>t-ratio</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.132238</td>
<td>0.4266</td>
<td>0.310</td>
<td>0.7691</td>
</tr>
<tr>
<td>Estimated...</td>
<td>0.920811</td>
<td>0.1275</td>
<td>7.22</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

• There will be things on the printout that you might not be familiar with. Don’t worry about those values. Focus on finding the information you need to write the equation of the LSRL and describe the strength of the relationship.
Typical Questions Regarding the LSRL

- State the equation of the LSRL. Define any variables used.
- Interpret the slope and the y-intercept of the LSRL.
- State and interpret the correlation coefficient.
- State and interpret the coefficient of determination.
- Predict a response value using the LSRL.
- Calculate a residual.
Re-expressing Data:
Strengthening Relationships

- Used to create a graph that is more linear.
- The process is often one of trial and error.
- Get a “feel” for a model, try it, then check the residual plot and the coefficient of determination for appropriateness of the model.
Why Re-express Data?

• To make the form of a scatterplot more nearly linear. Take the log of the x or y or both.

• To make a scatterplot have a more constant spread throughout rather than follow a fan shape. Take the log of both the x and y.
Why Re-express Data?

• Correlation and regression are used only to describe linear relationships. Transformations provide us with a method for straightening curved data so that we can use the tools of linear regression to summarize and analyze curved data.

• If the data changes direction (curve downward then upward or vice versa), it cannot be transformed to make it linear.
Using logarithms to Transformation Data

• Remember, after making a transformation, reexamine a residual plot to check for the desired effect.

• When you use transformed data to create a linear model, your regression equation is not in terms of \((x, y)\) but in terms of the transformed variable.

• After finding a LSRL on the transformed data, conduct an inverse transformation of the LSRL to obtain a model for the original data.
What You Need to Know

• How to make a scatterplot. Don’t forget to label axes and mark scales.
• How to describe a relationship in terms of direction, form, and strength.
• The difference between explanatory and response variables.
• Know that r is the correlation coefficient and what it measures.
• The properties of r.
• That the LSRL is the regression line that minimizes the sum of the squared residuals.
• How to find the r and the LSRL using the TI84.
• How to find the LSRL using the slope and intercept formulas when given summary statistics.
• How to use the LSRL to make predictions.
What You Need to Know

• Know how to interpret the slope of the LSRL in the context of the problem (it is the approximate change in the y-variable as the x-variable increases by 1).
• Know how to interpret the intercept of the LSRL in the context of the problem (it is the predicted value of y when x=0).
• How to find r-squared using the TI84 and what it is.
• How to interpret r-squared in the context of the problem.
• How to find a residual (error) for a point … residual=actual-predicted.
• Positive residuals are above the line and indicate the line underestimated the true value.
• Negative residuals are below the line and indicate the line overestimated the true value.
• How to interpret a residual plot to determine the fit of the line.
PRACTICE PROBLEMS
#1

Given a set of ordered pairs \((x, y)\) so that \(s_x = 1.6\), \(s_y = 0.75\), and \(r = 0.55\), what is the slope of the LSRL?

a) 1.82  
b) 1.17  
c) 2.18  
d) 0.26  
e) 0.78
A study found a correlation of \( r = -0.58 \) between hours spent watching television and hours per week spent exercising. Which of the following statements is most accurate?

a) About 1/3 of the variation in hours spent exercising can be explained by hours spent watching TV.

b) A person who watches less television will exercise more.

c) For each hour spent watching television, the predicted decrease in hours spent exercising is 0.58 hours.

d) There is a cause and effect relationship between hours spent watching TV and a decline in hours spent exercising.

e) 58% of the hours spent exercising can be explained by the number of hours watching TV.
#3

There is an approximate linear relationship between the height of females and their age (from 5 to 18 years) described by: \( \text{height} = 50.3 + 6.01(\text{age}) \) where height is measured in cm and age in years. Which of the following is not correct?

a) The estimated slope is 6.01 which implies that children increase by about 6 cm for each year they grow older.

b) The estimated height of a child who is 10 years old is about 110 cm.

c) The estimated intercept is 50.3 cm which implies that children reach this height when they are \( \frac{50.3}{6.01} \approx 8.4 \) years old.

d) The average height of children when they are 5 years old is about 50% of the average height when they are 18 years old.

e) My niece is about 8 years old and is about 115 cm tall. She is taller than average.
#4

A correlation between college entrance exam grades and scholastic achievement was found to be -1.08. On the basis of this you would tell the university that:

a. the entrance exam is a good predictor of success.
b. they should hire a new statistician.
c. the exam is a poor predictor of success.
d. students who do best on this exam will make the worst students.
e. students at this school are underachieving.
Under a "scatter diagram" there is a notation that the coefficient of correlation is .10. What does this mean?

a. plus and minus 10% from the means includes about 68% of the cases
b. one-tenth of the variance of one variable is shared with the other variable
c. one-tenth of one variable is caused by the other variable
d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10
The correlation coefficient for X and Y is known to be zero. We then can conclude that:

a. X and Y have standard distributions
b. the variances of X and Y are equal
c. there exists no relationship between X and Y
d. there exists no linear relationship between X and Y
e. none of these
Suppose the correlation coefficient between height as measured in feet versus weight as measured in pounds is 0.40. What is the correlation coefficient of height measured in inches versus weight measured in ounces? [12 inches = one foot; 16 ounces = one pound]

a. .4
b. .3
c. .533
d. cannot be determined from information given
e. none of these
A coefficient of correlation of \(-.80\)

a. is lower than \(r=+.80\)
b. is the same degree of relationship as \(r=+.80\)
c. is higher than \(r=+.80\)
d. no comparison can be made between \(r=+.80\) and \(r=+.80\)
#9

A random sample of 35 world-ranked chess players provides the following:

Hours of study: avg=6.2, s=1.3  
Winnings: avg=$208,000, s=42,000  
Correlation=0.15

Find the equation of the LSRL.

a. Winnings=178,000+4850(Hours)
b. Winnings=169,000+6300(Hours)
c. Winnings=14,550+31,200(Hours)
d. Winnings=7750+32,300(Hours)
e. Winnings=-52,400+42,000(Hours)
Part III:
Chapters 11 - 13

GATHERING DATA
UNDERSTANDING RANDOMNESS
Random Outcomes

• A random event is one whose outcome we cannot predict.
• This may suggest that random events are totally chaotic and therefore not useful in modeling real-world situations – not so.
• Although the outcomes of individual trials of a random event are unknown, over the long run there is a pattern.
• It is this long-run predictability that makes randomness a useful tool in reaching conclusions.
Simulation

• Is a powerful tool for gaining insight into events whose outcomes are random.

• Preforming a Simulation
  1) Identify the event to be repeated.
  2) Outcomes, state how you will model the random occurrence of an outcome (assign digits to outcomes).
  3) Trial, explain how you will simulate a trial and what the response variable is.
  4) Run several trials and tabulate the results.
  5) Conclusion, summarize your results and draw your conclusion in the context of the problem.
Example

• Your school decided to hold a raffle to defray the cost of tickets to the senior prom. The breakdown of ticket sales was; Students: 650 and Faculty: 325. At an assembly, the principal reached into a jar and drew three winning tickets. To everyone’s dismay, all three winners were members of the faculty. The students cried foul. Their argument was that, given the breakdown of sales between the two groups, it would be highly unlikely for all three winners to be faculty members.

• Conduct a simulation, using 10 trials and starting on line 130 of the random digit table, to determine if the outcome of the drawing was fair.
Solution

- ID event being repeated – selecting a ticket from the jar.
- Outcomes
  - 000 – 649 student ticket
  - 650 – 974 faculty ticket
  - 975 – 999 skip
  - If a number appears more than once in a trial it is ignored. Can’t select the same ticket twice.
Solution

- Trial
  - Select 3 tickets and determine if student or faculty.
  - Response variable – whether are not all 3 tickets drawn belong to a member of the faculty or not (yes/no).
- Conclusion: In our simulation all 3 winners were faculty members only 10% of the time. While this result is unlikely, we might suspicious, but would need to run many more trials and a smaller percent of all faculty winners before we make an accusation of unfairness.
Producing Data

• To draw meaningful conclusions from measured or observed data, it is essential that we understand proper data-collection methods.
• Bad sample designs yield worthless data.
• There is *no* way to correct for a bad sample.
Basic Concepts of Sampling

• **Population** – the entire group of individuals whom we hope to learn about. The population is determined by what we want to know.

• **Sample** – a smaller group of individuals selected from the population. The sample size is determined by what is practical and representative of the population we are interested in learning about.
Terminology of Sampling

- **Sampling Frame** – a list of individuals from the population of interest from which the sample is drawn.
- **Census** – a sample that consists of the entire population.
- **Sampling Variability** – the natural tendency of randomly drawn samples to differ, one from the another. Sampling variability is not an error, just the natural result of random sampling. Although samples vary, they do not vary haphazardly but rather according to the laws of probability.
Parameters and Statistics

- **Parameter:**
  - A number that characterizes some aspect of the population such as the mean or standard deviation of some variable of the population.
  - We rarely know the true value of a population parameter.
  - Denote with Greek letters.

- **Statistic:**
  - Values calculated from sample data.
  - Use statistics to estimate values in the population (parameters).
  - Denote statistics with standard letters.
Parameters and Statistics

<table>
<thead>
<tr>
<th>Name</th>
<th>Statistic</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{x}$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Correlation</td>
<td>$r$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>Regression Coefficient (slope)</td>
<td>$b_1$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Proportion</td>
<td>$\hat{p}$</td>
<td>$p(\theta \ or \ \pi)$</td>
</tr>
</tbody>
</table>
Sample Size

• The number of individuals selected from our sampling frame.
• The size of the population does not dictate the size of a sample.
• The general rule is that the sample size should be no more than 10% of the population size (10n<N).
Sample Designs

- The method used to choose the sample.
- Incorporate the idea that chance, rather choice, is used to select the sample.
  - **Probability Sample** – chosen using a random mechanism in such a way that each individual or group of individuals has the same chance of being selected.
  - **Random Sample** – chosen using a random mechanism in such a way that the probability of each sample being selected can be computed.
    - May be drawn with or without replacements.
Sample Designs (cont.)

• **Simple Random Sample (SRS)** – a random sample chosen without replacement and meets the following rules for a SRS of size \( n \);
  • Each individual has an equal chance of selection.
  • Each possible set of \( n \) individuals has an equal chance of selection.

• **Stratified Random Sampling** – divides the population into *homogeneous* groups called **strata**.
  • Strata are made up of individuals similar in a way that may effect the response variable.
  • SRS is applied within each stratum before the results are combined.
Sample Designs (cont.)

• **Cluster Sample** – when the population exists in readily defined heterogeneous groups or clusters, a cluster sample is an SRS of the clusters.
  
  • This method of sampling uses the data from all of the individuals from the selected clusters.
  
  • Often used to reduce the cost of obtaining a sample.
Sample Designs (cont.)

• **Systematic Sample** – selected according to a predetermined scheme.
  - Can be random if the starting point of the scheme is randomly selected.
  - Can never produce a SRS because each sample of size n does not have an equal chance of being chosen.
  - Often used to simplify the sampling process.
  - When the order of the list is not associated with the responses sought, this method gives a representative sample.
Sample Designs (cont.)

- **Multistage Sampling** – produces a final sample in stages, taking each sample from the one before it.
  - May combine several methods of sampling.
  - Can be random but will not produce a SRS.
- **Convenience Sample** – obtained exactly as its name suggests, by sampling individuals who are conveniently available.
  - Unlikely to represent the population of interest because it is unlikely that every member of this population is conveniently available.
  - Are not probability samples nor are they random.
  - May lead to bias.
Bias

• **Bias** is any systematic failure of a sample to represent its population of interest.
  • Very important to reduce bias.
  • Best defense against bias is randomization.
  • There is no way to recover from a biased sample.
  • Remember, you can reduce bias, but you can never completely eliminate it.
Sources of Bias

- **Undercoverage Bias** – excluding or underrepresenting some part of the population.
- **Response Bias** – anything that influences responses.
  - Examples: question bias and interviewer bias.
- **Nonresponse Bias** – occurs when individuals selected for the sample fail to respond, cannot be contacted, or decline to participate.
- **Voluntary Response Bias** – when choice rather than randomization is used to obtain a sample.
  - People with strong opinions tend to be overrepresented.
Observational Study

- Researchers observe individuals and record variables of interest but do not impose a treatment.
- It is not possible to prove a cause-and-effect relationship with an observational study.
Experiment

• An experiment differs from an observational study in that the researcher deliberately *imposes a treatment*.

• An experiment must have at least one *explanatory variable* to manipulate and at least one *response variable* to measure.

• In an experiment, it is possible to determine a cause-and-effect relationship between the explanatory and response variables.
Completely Randomized Experiment

- Subjects are randomly assigned to a treatment group.
- The researcher then compares the subject groups’ responses to each treatment.

- It is not necessary to start with a random selection of subjects.
- The randomization occurs in the random assignment to treatment groups.
Block Design

- If our experimental units differ in some characteristic that may affect the results of our experiment, we should separate the groups into blocks based on that characteristic and then randomly assign the subjects within each block.
- In effect, we are conducting parallel experiments.

- Blocks reduce variability so that the effects of the treatments can be seen.
- Blocks themselves are not treatments.
- **Blocking is to experimental design as stratifying is to sampling design.**
Matched-Pairs Design

- Is a form of block design.
- Two types:
  - **One Subject**: Uses just one subject, who receives both treatments. The order in which the subject receives the treatments is randomized.
  - **Two Subjects**: Two subjects are paired based on common characteristics that might affect the response variable. One subject from each pair is randomly assigned to each of the treatments. The response variable is then the difference in the response to the two treatments for each pair.
Four Principles of Experimental Design

1. **Control** – Reduces variability by controlling the sources of variation.
   - **Comparison** is an important form of control.
   - Every experiment must have *at least two* groups so that the effect of a new treatment can be compared with either the effect of a traditional treatment or the effect of no treatment at all.
   - The control group is the group given the traditional treatment, no treatment, or a **placebo** (a treatment known to have no effect).
Four Principles of Experimental Design (cont.)

2. **Randomize** – randomization to treatment groups reduces bias by equalizing the effects of **lurking variables**.
   - Lurking Variables are variables that we did not think to measure but can affect the response variable.
   - Does not eliminate unknown or uncontrollable sources of variation but spreads them out across the treatment levels and makes it easier to detect differences caused by the treatments.
Four Principles of Experimental Design (cont.)

3. **Replicate** – One or two subjects do not constitute an experiment. We should include many subjects in a comparative experiment. Experiments should be designed in such a way that other researchers can replicate our results.

4. **Block** – Although blocking is not required in an experimental design, it may improve the design. If the experimental units are different in some way that may affect the results of the experiment, the groups should be separated into blocks based on that characteristic.
Other Considerations in the Design of Experiments

• **Blinding**
  • **Single-Blind**: The subjects of the experiment do not know which treatment group they have been assigned or those who evaluate the results of the experiment do not know how the subjects have been assigned to the groups.
  • **Double-Blind**: Neither the subjects nor the evaluators know how the subjects have been allocated to treatment groups.

• **Confounding** – An experiment is said to be confounded if we cannot separate the effect of a treatment (explanatory variable) from the effects of other influences (confounding variables) on the response variable.
Other Considerations in the Design of Experiments (cont.)

- **Statistical Significance** – When an observed difference is too large for us to believe that it is likely to have occurred by chance alone, we consider the difference to be statistically significant.

- **Placebo Effect** – The tendency in humans to show a response whenever they think a treatment is in effect.
  - Well designed experiments use a control group so that the placebo effect operates equally on both the treatment group and the control group, thus allowing us to attribute changes in the response variable to the explanatory variable.
What You need to Know

• How to explain and conduct a simulation, including assigning digits.
• Know the types of sampling design.
• Know the types of bias.
• What is a sampling frame.
• Observational studies vs experiments.
• Language of experiments (experimental units, factors, levels, treatments, response).
• How to use the random table to assign subjects to treatments.
What You need to Know

• Major principles of experimental design (control, randomization, replication, and blocking**).
• Know why and when to use a blocked design.
• Know the difference between a completely randomized design and a blocked design.
• Know how to diagram an experiment.
• Know the idea of “significance”.
• Know what is meant by “confounding”.
• Know the idea of a “matched pairs” design.
In one study on the effect of niacin on cholesterol level, 100 subjects who acknowledged being long-time niacin takers had their cholesterol levels compared with those of 100 people who had never taken niacin. In a second study, 50 subjects were randomly chosen to receive niacin and 50 were chosen to receive a placebo.

a) The first study was a controlled experiment, while the second was an observational study.

b) The first study was an observational study, while the second was a controlled experiment.

c) Both studies were controlled experiments

d) Both studies were observational studies.
Each of the 29 NBA teams has 12 players. A sample of 58 players is to be chosen as follows. Each team will be asked to place 12 cards with their players names into a hat and randomly draw out two names. The two names from each team will be combined to make up the sample. Will this method result in a SRS of the players?

a) Yes, because each player has the same chance of being selected.

b) Yes, because each team is equally represented.

c) Yes, because this is an example of stratified sampling, which is a special case of SRS.

d) No, because the teams are not chosen randomly.

e) No, because not each group of players has the same chance of being selected.
A consumer product agency tests miles per gallon for a sample of automobiles using each of four different octane of gasoline. Which of the following is true?

a) There are four explanatory variables and one response variable.
b) There is one explanatory variable with four levels of response.
c) Miles per gallon is the only explanatory variable, but there are four response variables.
d) There are four levels of a single explanatory variable.
e) Each explanatory level has an associated level of response.
Your company has developed a new treatment for acne. You think men and women might react differently to the medication, so you separate them into two groups. Then the men are randomly assigned into two groups and the women are randomly assigned into two groups. One of the two groups is given the medicine, the other is given a placebo. The basic design of this study is:

a) completely randomized  
b) randomized block, blocked by gender  
c) completely randomized, stratified by gender  
d) randomized block, blocked by gender and type of medication.  
e) a matched pairs design
A double-blind design is important in an experiment because:

a) There is a natural tendency for subjects in an experiment to want to please the researcher.
b) It helps control for the placebo effect.
c) Evaluators of the responses in a study can influence the outcomes if they know which treatment the subject received.
d) Subjects in a study might react different if they knew which treatment they were receiving.
e) All of the above reasons are valid.
A school committee member is lobbying for an increase in the gasoline tax to support the county school system. The local newspaper conducted a survey of country residents to assess their support for such an increase. What is the population of interest here?

a) All school-aged children.
b) All county residents
c) All county residents with school-aged children
d) All county residents with children in the school system.
e) All county school system teachers.
An experiment was designed to test the effect of 3 different types of paints on the durability of wooden toys. Since boys and girls tend to play differently with toys, a randomly selected group of children was divided into 2 groups by gender. Which of the following statements about this experiment is true.

a) Type of paint is a blocking factor  
b) Gender is a blocking factor  
c) This is a completely randomized design  
d) This is a matched-pairs design in which one boy and one girl are matched to form a pair
#8

Which of the following is not a source of bias in a survey?

a) non-response
b) wording of the question
c) voluntary response
d) use of a telephone survey
e) all are sources of bias
#9

Which of the following is not a valid sampling design

a) Number every member of the population and select 100 randomly chosen members

b) Divide a population by gender and select 50 individuals randomly from each group

c) Select every 20th person, starting at a random point.

d) Select five homerooms at random from all the homerooms in a large high school

e) All of these are valid.