3. Is there a correlation between test anxiety and exam score performance? Data on \( x \) = score on a measure of test anxiety and \( y \) = exam score are given in the table below.

<table>
<thead>
<tr>
<th>X = test anxiety</th>
<th>23</th>
<th>14</th>
<th>14</th>
<th>0</th>
<th>7</th>
<th>20</th>
<th>20</th>
<th>15</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y = score on exam</td>
<td>43</td>
<td>59</td>
<td>48</td>
<td>77</td>
<td>50</td>
<td>52</td>
<td>46</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

a. Which one of the variables is the explanatory and which is the response variable?

Explanatory variable = \( X \) = test anxiety
Response variable = \( Y \) = score on exam

b. Construct a scatter plot and comment on the features of the plot. (Overall direction, form, strength)

Direction - negative
Form - linear
Strength - weak

![Scatter plot](image)

\[
r = -0.788
\]
\[
r^2 = 0.621
\]
\[
y = 68.838 - 1.064x
\]

c. Find the correlation coefficient, the coefficient of determination and the LSRL.

\[
x \quad y \quad y - \hat{y}
\]
| 0 | 8.1618 |
| 7 | -11.39 |
| 14 | 5.0544 |
| 14 | -5.946 |
| 15 | -1.882 |
| 20 | 4.437 |
| 20 | -1.563 |
| 21 | 4.5008 |
| 23 | -1.372 |

d. Construct a residual table and the residual plot.

![Residual plot](image)

e. Comment on the relationship between test anxiety and test scores based upon the analysis you performed.

The relationship between test anxiety and exam score is not best modelled by the least-squares regression line. The coefficient of determination \( r^2 = 0.621 \) is weak and the residual plot is not scattered.

f. If we were to add the data point (5, 100) how would it affect the LSRL? What is this point called?

- Slope would decrease (that is \( b < -1.064 \))
- Outlier \( (y\text{-direction}) \)
4. The sample correlation coefficient between annual raises and teaching evaluations for a sample of 353 college faculty was found to be \( r = .11 \).

a. Interpret this value.

\[ r = 0.11 \to \text{weak correlation between annual raises and teaching evaluations} \]

b. If a straight line were fit to the data using least squares regression, what proportion of variation in evaluations could be attributed to the approximate linear relationship between raises and evaluation?

\[ r^2 = 0.11^2 = 0.0121 \]

5. In physics class, the intensity of a 100-watt light bulb was measured by a sensing device at various distances from the light source, and the following data was collected. Note that a candela (cd) is an international unit of luminous intensity.

<table>
<thead>
<tr>
<th>Distance</th>
<th>1</th>
<th>1.1</th>
<th>1.2</th>
<th>1.3</th>
<th>1.4</th>
<th>1.5</th>
<th>1.6</th>
<th>1.7</th>
<th>1.8</th>
<th>1.9</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity (candela)</td>
<td>.2965</td>
<td>.2522</td>
<td>.2055</td>
<td>.1746</td>
<td>.1534</td>
<td>.1352</td>
<td>.1145</td>
<td>.1024</td>
<td>.0923</td>
<td>.0832</td>
<td>.0734</td>
</tr>
</tbody>
</table>

a. Plot the data. Based on the pattern of the points, propose a model for the data. Then use a transformation followed by a linear regression and then an inverse transformation to construct a model.

\[
\begin{align*}
\ln \hat{y} &= a + b \ln x \\
L_1 &= x \\
L_2 &= y \\
L_3 &= \ln x = \ln (L_1) \\
L_4 &= \ln y = \ln (L_2) \\
\ln \hat{y} &= -1.205 - 2.013 \ln x \\
\hat{y} &= e^{-1.205} e^{-2.013 \ln x} \\
\hat{y} &= 0.29969 x^{-2.013}
\end{align*}
\]

b. Describe the relationship between the intensity and the distance from the light source.

\[ r^2 = 0.999 \]

There is a strong, negative, curved relationship between distance from light source and light intensity.
6. The following table reports Census Bureau data on undergraduate students in U.S. colleges and universities in the fall of 1991.

**UnderGraduate College enrollment by age of students – Fall 1991 (thousands of students)**

<table>
<thead>
<tr>
<th>Age</th>
<th>2-yr Full-time</th>
<th>2-yr part-time</th>
<th>4-yr full time</th>
<th>4-yr part-time</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-17</td>
<td>44</td>
<td>4</td>
<td>79</td>
<td>0</td>
<td>127</td>
</tr>
<tr>
<td>18-21</td>
<td>1345</td>
<td>456</td>
<td>3869</td>
<td>159</td>
<td>5829</td>
</tr>
<tr>
<td>22-29</td>
<td>489</td>
<td>690</td>
<td>1358</td>
<td>494</td>
<td>3031</td>
</tr>
<tr>
<td>30-44</td>
<td>287</td>
<td>704</td>
<td>289</td>
<td>627</td>
<td>1907</td>
</tr>
<tr>
<td>&gt;=45</td>
<td>49</td>
<td>209</td>
<td>62</td>
<td>160</td>
<td>480</td>
</tr>
<tr>
<td>Totals</td>
<td>2214</td>
<td>2063</td>
<td>5657</td>
<td>1440</td>
<td>GT (11,374)</td>
</tr>
</tbody>
</table>

a. Fill in the "totals" in the table above. What is the grand total (GT) of students who were enrolled in colleges and universities in the fall of 1991?

\[ \text{11,374,000 students} \]

b. What percent of all **undergraduate students** were 18-21 years old in the fall of the 1991?

\[ \frac{5829}{11,374} = 51.25\% \]

c. Find the percent of the undergraduates enrolled in each of the four types of programs who were 18-21 years old. Make a bar chart to compare these percents.

\[ \frac{1345}{2214} = 60.75\% \]
\[ \frac{456}{2063} = 22.10\% \]
\[ \frac{3869}{5657} = 68.39\% \]
\[ \frac{159}{1440} = 11.04\% \]

d. The 18-21 group is the "traditional" age group for college students. Briefly summarize what you have learned from the data about the extent to which this group predominates in different kinds of college programs.

The data shows that 60.75% of 18-21 year-olds are enrolled in a 4-year full time program. Further, 18-21 year-olds constitute 68.39% of all students enrolled in a 4-year full time program. Further, students aged 18-21 make up 51.25% of all undergraduates enrolled in U.S. colleges.
7. Define these terms:

a. Census - a study that contacts every individual in the population.

b. Population - the entire group of individuals we seek information about.

c. Sample - a group of people extracted from the population meant to represent the population.

d. Survey - an observational study in which individuals are asked questions that lend insight into characteristics of the population.

e. Simple Random Sample (SRS) - an SRS of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

f. Bias in a sample - a sample is biased if it systematically favors certain outcomes; sample is not representative of the population.

g. Variability in a sample - the natural tendency for individuals in a sample to vary from one another (with respect to their responses).

h. Confounding Variable - a variable that is unaccounted for in the experimental design but may be influencing the results.

i. Stratified random sample - the population is divided into strata based on certain characteristics and an SRS is taken from each stratum to form the whole sample.

j. Cluster Sample - the population is divided into clusters, often based on geography, and an SRS of the clusters is taken. All the individuals in the chosen clusters then form the sample.

k. Block design - a group of experimental units are known before the experiment to be similar in some way that is expected to systematically affect the response to the treatments.

l. BLock design - a form of control in which the random assignment of units to treatments is carried out separately within each block.

m. Experiment - a study in which treatments are imposed on the experimental units/subjects.

n. Observational study - a study in which individuals are observed or asked questions but are not manipulated (through treatments) in any way.
8. The Ministry of Health in the Canadian Province of Ontario wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records but that source doesn’t allow us to compare people who use health services with those who don’t. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who live in the Province of Ontario.

a. What is the population for this sample survey? What is the sample?

The population consists of all individuals who live in the Province of Ontario, while the sample consists of 61,239 randomly selected individuals who live in the Province of Ontario.

b. The survey found the 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why or why not?

Since the sample was taken randomly and N = 61,239, the estimates are likely to be close to the truth about the entire population.

c. Is this an experiment or an observation study? How can you tell?

This is an observational study because no treatment was imposed on the surveyed individuals.

9. What are the characteristics of a well-designed and well-conducted study?

* Individuals are representative of the population of interest
* The experiment is controlled
* Replication is performed
* Randomization is used

10. Elaine is enrolled in a self-paced course that allows three attempts to pass an examination on the material. She does not study and has 2 out of 10 chances of passing on any one attempt by pure luck. What is Elaine’s likelihood of passing on at least one of the three attempts? (Assume the attempts are independent because she takes a different exam at each attempt.)

\[ P(\text{passing}) = 0.2 \]
\[ 1 - P(\text{passing none}) = 1 - (0.8)^3 = 0.4889 \]

a. Explain how you would use random digits to simulate one attempt at the exam. Elaine will of course stop taking the exam as soon as she passes.

Assign the digits 0 and 1 to a pass and the digits 2, 3, 4, 5, 6, 7, 8, and 9 to a fail. Using a random number table, simulate Elaine’s exam attempt and correctly identify each attempt with a pass or a fail.

b. Simulate 50 repetitions. What is your estimate of Elaine’s likelihood of passing the course?

Enter Table B at line 138

<table>
<thead>
<tr>
<th>46831</th>
<th>68908</th>
<th>40772</th>
<th>25582</th>
<th>47781</th>
<th>33586</th>
<th>79667</th>
<th>06928</th>
</tr>
</thead>
<tbody>
<tr>
<td>69588</td>
<td>99404</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \frac{8}{50} = 16\% \]
11. Can aspirin help prevent heart attacks? The Physicians' Health Study, a large medical experiment involving 22,000 male physicians, attempted to answer this question. One group of about 11,000 physicians took an aspirin every second day, while the rest took a placebo. After several years the study found that subjects in the aspirin group had significantly fewer heart attacks than the subjects in the placebo group.

a. Identify the experimental subjects, the factor and its levels, and the response variable in the health study.

- Experimental subject: 22,000 male physicians
- Factor: use of medication
- Levels: aspirin, placebo
- Response Variable: number of heart attacks

b. Use a diagram to outline a completely randomized design for the health study.

12. A mortgage lender routinely places advertisements in a local newspaper. The advertisements are of three different types: one focusing on low interest rates, one featuring low fees for first-time buyers, and one appealing to people who may want to refinance their homes. The lender would like to determine which advertisement format is most successful in attracting customers to call for more information. Describe be an experiment that would provide the information needed to make this determination. Be sure to consider extraneous factors such as the day of the week that the advertisement appears in the paper, the section of the paper in which the advertisement appear, daily fluctuations of the interest rate and so forth. What role does randomization play in your design? Diagram the design.

- Block on weekday vs. weekend
- Random Allocation
- Weekday: Front, Middle, Back
- Weekend: Front, Middle, Back
- \( T_1 \) = low interest rates
- \( T_2 \) = low fees for first-time buyers
- \( T_3 \) = refinancing
Topic III  Anticipating Patterns:  Exploring Random Phenomena using Probability and Simulation

13. Probability is a measure of how likely an event is to occur. Match one of the probabilities that follow with each statement about an event.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.01</th>
<th>0.3</th>
<th>0.6</th>
<th>0.99</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>The sun will rise in the west in the morning.</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.</td>
<td>Thanksgiving will be on Thursday next year.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c.</td>
<td>An event is very unlikely, but it will occur rarely.</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d.</td>
<td>The event will occur most of the time. Very rarely will it not occur.</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| e. | Give an example where the other 2 probabilities may occur.
   There are 15 yellow marbles, 3 red marbles, 1 blue marble, and 1 orange marble in a bag.
   \[ P(\text{red}) = 0.13 \]
   \[ P(\text{yellow or blue}) = 0.6 \]

14. What is the formula used for each of the following probabilities:

a. Addition Rule
   \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

b. Multiplication Rule
   \[ P(A \cap B) = P(A) \times P(B|A) \]

  \[ P(A|B) = \frac{P(A \cap B)}{P(B)} \]

c. Conditional Probability

15. The type of medical care a patient receives may vary with the age of the patient. A large study of women who had a breast lump investigated whether or not each woman received a mammogram and a biopsy when the lump was discovered. Here are some probabilities estimated by the study. The entries in the table are the probabilities that both of two events occur. For example: 0.321 is the probability that a patient is under 65 years of age and the tests were done.

<table>
<thead>
<tr>
<th></th>
<th>Tests</th>
<th>Done</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Under 65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.321</td>
<td>0.124</td>
</tr>
<tr>
<td>No</td>
<td>0.624</td>
<td>0.876</td>
</tr>
<tr>
<td>Age 65 and Over</td>
<td>0.365</td>
<td>0.635</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0.635</td>
<td>0.365</td>
</tr>
</tbody>
</table>

a. What is the probability that a patient in this study is under 65?
   \[ P(\text{under 65}) = 0.321 + 0.124 = 0.445 \]

b. Is 65 or over?
   \[ P(\text{65 or over}) = 1 - 0.445 = 0.555 \]

c. What is the probability that the tests were done for a patient? That they were not done?
   \[ P(\text{yes}) = 0.321 + 0.365 = 0.686 \]
   \[ P(\text{no}) = 1 - 0.686 = 0.314 \]

d. Are the events A = (patient was 65 or older) and B = (the tests were done) independent? Were the tests omitted on older patients more or less frequently that would be the case if testing were independent of age?

The tests were done more frequently than if testing were independent of age, meaning test were omitted less frequently on older patients.

The tests were done more frequently than if testing were independent of age, meaning test were omitted less frequently on older patients.
16. Here are the counts (in thousands) of earned degrees in the United States in a recent year, classified by level and by the sex of the degree recipient:

<table>
<thead>
<tr>
<th></th>
<th>Bachelor's</th>
<th>Master's</th>
<th>Professional</th>
<th>Doctorate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>616</td>
<td>194</td>
<td>30</td>
<td>16</td>
<td>856</td>
</tr>
<tr>
<td>Male</td>
<td>529</td>
<td>171</td>
<td>44</td>
<td>26</td>
<td>770</td>
</tr>
<tr>
<td>Total</td>
<td>1145</td>
<td>365</td>
<td>74</td>
<td>42</td>
<td>1620</td>
</tr>
</tbody>
</table>

a. If you choose a degree recipient at random, what is the probability that the person you choose is a woman?

\[
\frac{856}{1620} = 52.6490
\]

b. What is the conditional probability that you choose a woman, given that that person chosen received a professional degree?

\[
P(\text{woman}|\text{prof. degree}) = \frac{P(\text{woman} \cap \text{prof. degree})}{P(\text{prof. degree})} = \frac{30}{1626} = \frac{74}{1626} = \frac{40.5490}{52.6490}
\]

c. Are the events "choose a woman" and "choose a professional degree recipient" independent? How do you know?

Let A = choose a woman, let B = choose a prof. degree recipient.

If events A and B are independent, then

\[
P(A) \times P(B) = P(A \cap B)
\]

\[
\frac{856}{1620} \neq \frac{30}{1626} \neq \frac{74}{1626}
\]

The events are not independent.

17. Consolidated Builders has bid on two large construction projects. The company president believes that the probability of winning the first contract (event A) is 0.6, that the probability of winning the second (event B) is 0.4 and the joint probability of winning both jobs (event A and B) is 0.2.

a. Draw the Venn diagram that illustrates the relationship between events A and B.

```
A
0.6
0.4
0.2
```

b. Find the following probabilities:

\[
P(A \text{ or } B) = 0.8 \quad P(A \text{ and } B) = 0.2 \quad P(A, \text{ and Not B}) = 0.4
\]

\[
P(\text{Not A, and B}) = 0.2 \quad P(\text{not A and not B}) = 0.2
\]

18. What is the difference between discrete and continuous random variables?

A discrete random variable has a countable number of possible values; that is, the set of numbers for which the random variable is defined is the set of integers. A continuous random variable, on the other hand, takes all values in an interval of numbers. The values of the random variable are not restricted to integers.
19. Let $x$ be the number of courses for which a randomly selected student at a certain university is registered. The probability distribution of $x$ appears in the accompanying table.

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(x)$</td>
<td>0.02</td>
<td>0.03</td>
<td>0.09</td>
<td>0.25</td>
<td>0.40</td>
<td>0.16</td>
<td>0.05</td>
</tr>
</tbody>
</table>

$\sum P(x) = 1$

a. What is $P(x = 4)$? 0.25
b. What is $P(x \leq 4)$? 0.39

c. What is the probability that the selected student is taking at most five courses?

$$1 - 0.21 = 0.79$$

d. What is the probability that the selected students is taking at least five courses?

$$0.40 + 0.16 + 0.05 = 0.61$$

e. Calculate $P(3 \leq x \leq 6)$ and $P(3 \leq x < 6)$. Explain why the two probabilities are different.

The first probability is inclusive on its lower and upper bounds and the second probability is exclusive on its lower and upper bounds.

$$P(3 \leq x \leq 6) = 0.09 + 0.25 + 0.40 + 0.16 = 0.9$$
$$P(3 \leq x < 6) = 0.25 + 0.40 = 0.65$$

f. Find the mean, standard deviation and variance of the random variable $x$.

$$\mu_x = \sum x_i p_i = 4.16$$
$$\sigma_x^2 = \sum (x_i - \mu_x)^2 p_i = 1.88$$
$$\sigma_x = 1.37$$

20. You have two scales for measuring weights in a chemistry lab. Both scales give answers that vary a bit in repeated weightings of the same item. If the true weight of a compound is 2.00 grams, the first scale produces readings $X$ that have mean 2.00 grams and standard deviations 0.002 grams. The second scale’s readings $Y$ have mean 2.001 grams and standard deviation of 0.001 grams.

a. What are the mean and standard deviation of the difference $y - x$ between the readings? (The readings $X$ and $Y$ are independent.)

$$\mu_{y-x} = \mu_Y - \mu_X = 2.001 - 2.000 = 0.001$$
$$\sigma_{y-x} = \sqrt{\sigma_Y^2 + \sigma_X^2} = \sqrt{0.002^2 + 0.001^2} = 0.00224$$

b. You measure once with each scale and average the readings. Your result is $Z = (X + Y)/2$. What are the mean and standard deviation of $Z$?

$$\mu_Z = \frac{2.001 + 2.000}{2} = 2.0005$$
$$\sigma_Z = \frac{\sigma_{Y-x}^2}{2} = 0.00112$$

21. Among employed women, 25% have never been married. You select 10 employed women at random.

a. The number in your sample who have never been married has a binomial distribution. What are $n$ and $p$?

$$n = 10$$
$$p = 0.25$$

b. Create a binomial distribution table, a probability distribution table and a cumulative distribution table for this data. Let $x$ be the number of women who have never been married.
c. What is the probability that exactly 2 of the 10 women in your sample has never been married?

0.2816

d. What is the probability that 2 or fewer have never been married?

0.5256

e. What is the mean and standard deviation for this binomial distribution?

\[
\text{mean: } np = 10(0.25) = 2.5 \\
\text{standard deviation: } \sqrt{np(1-p)} = \sqrt{2.5(0.75)} = 1.369
\]

22. A basketball players makes 80% of his free throws. We put him on the free throw line and ask him to shoot free throws until he misses one. Let X = number of free throws the player takes until he misses.

a. What assumptions do you need to make in order for the geometric model to apply? With these assumptions, verify that X has a geometric distribution. What actions constitutes “success” in this context?

- Each observation falls into one of two categories: success (a successful free throw) or failure (a missed free throw).
- The free throws are independent.
- The probability of success, \( p = 0.2 \), is the same for all observations.
- The variable of interest is the number of free throws required to obtain the first miss.

b. Create a geometric distribution table for \( x \) values from 1 to 10. Create a pdf and a cdf.
c. What is the probability that the player will make 5 shots before he misses?

\[ P(X = 5) = 0.0655 \] 16 shots required for a miss.

d. What is the probability that he will make at most 5 shots before he misses?

\[ P(Y \leq 5) = 0.7278 \]

e. What is the mean of this geometric distribution?

\[ \text{mean: } \frac{1}{p} = \frac{1}{0.2} = 5 \]

23. The area under the curve for a normal distribution is represented by a bell-shaped curve.

a. What are the properties of a normal distribution? Sketch a normal curve.

\[ \text{Properties:} \]
\[ \begin{align*}
\text{- Symmetric / bell-shaped} \\
\text{- 68-95-99.7 Rule}
\end{align*} \]

24. A certain population of whooping cranes that migrate between Wisconsin and Florida every year has a SRS taken. The sample of 15 male cranes were weighed before they left Wisconsin to begin their trip. The mean weight of the 15 males was found to be 22.7 pounds with a standard deviation of 2.3 pounds. Why is this population considered a normal distribution?

a. What is the probability that a random selected male crane weights less than 20 pounds? Sketch the curve and put in all the appropriate values. Write the probability statement.

\[ P(X < 20) = 0.1202 \]

b. What is the probability that a random selected male crane weights more than 25 pounds?

\[ P(X > 25) = 0.1587 \]

c. What is the probability that a random selected male crane weights between 21 and 26 pounds?

\[ P(21 < X < 26) = 0.1944 \]
25. The Helsinki Heart Study asks whether the anti-cholesterol drug gemfibrozil will reduce heart attacks. In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study plans to give gemfibrozil to 2000 men and a placebo to another 2000 men. The probability of a heart attack during the 5-year period of the study for men this age is about 0.04. We can think of the study participants as an SRS from a large population, of which the proportion $p = 0.04$ will have heart attacks.

   a. What is the mean number of heart attacks that the study will find in one group of 2000 men if the treatment doesn't change the probability of 0.04?

   $0.04 \times 2000 = 80$ heart attacks

   b. What is the probability that the group will suffer at least 75 heart attacks? Sketch the curve, show all the work and write the probability statement.

   $P(p > 0.0315) = \frac{0.07158}{1}$

   Sample distribution with $\mu = 0.4$ and $\sigma = \sqrt{\frac{0.04(0.96)}{2000}}$

26. Children in kindergarten are sometimes given the Raven Progressive Matrices Test (RPMT) to assess their readiness for learning. Experience at Southward Elementary School suggests that the RPMT scores for its kindergarten pupils have a mean of 13.6 with a standard deviation of 3.1. The distribution is close to normal. Mr. Brown has 22 children in his kindergarten class this year.

   a. What is the probability that class's mean score will be less than 12.0?

   $P(X \leq 12) = 0.3029$

   *Use norminv

   b. Mr. Brown suspects that the class RPMT scores will be unusually low because the test was interrupted by a fire drill. He wants to find the level $L$ such that there is only a probability of 0.05 that the mean score of his class fall below $L$. What is this value of $L$.

   $L - 13.6 = \frac{-1.645}{3.1}$

   $L = 13.501$

27. Explain what is meant by the Law of Large Numbers. How does this law apply to sampling distributions?

   The Law of Large Numbers states that as the sample size increases, the experimental probability of a certain phenomenon approaches the theoretical probability. In regards to sampling distributions, the Law of Large Numbers implies that the statistic $\bar{X}$ approaches the parameter $\mu$ as the sample size increases.

28. What is the Central Limit Theorem? How is the CLT used in sampling distributions?

   The Central Limit Theorem states that as the sample size increases, the sampling distribution of the parameter of interest approaches a Normal distribution. The Normality condition is satisfied by the CLT if $n \geq 30$.\
   a. What is mean by “unbiased” estimator?

   A statistic is an unbiased estimator of a parameter if the mean of its sampling distribution is equal to the true value of the parameter; that is, \( \bar{x} \) and \( \hat{p} \) are unbiased estimators of \( \mu \) and \( \pi \) respectively.

   b. Does unbiasedness alone guarantee that the estimate will be close to the true value? Explain.

   No, the conditions for confidence intervals must be met and the experimental design must be sound in order for legitimate conclusions to be made. Moreover, larger sample sizes have less variability, so the statistic will be closer to the parameter.

30. What is meant by the standard error of a population parameter? What are the standard errors for the following:

   - Population Mean
     \[ \frac{s}{\sqrt{n}} \]
   - Population Proportion
     \[ \sqrt{\hat{p}(1-\hat{p})} \]

31. What is the general form of all confidence intervals?

   \[ \text{estimate} \pm \text{margin of error} = \text{estimate} \pm (\text{critical value})(\text{standard error}) \]

32. Suppose that a random sample of 50 bottles of a particular brand of cough medicine is selected and the alcohol content of each bottle is determined. Let \( \mu \) denote the average alcohol content for the population of all bottles of the brand under the study. Suppose that the sample mean is 8.2 grams with a standard deviation of 1.5 grams.

   a. Find the 95% confidence interval for the mean alcohol content of the cough medicine. Report the margin of error and show all the work.

   \[ 8.2 \pm 2.010 \left( \frac{1.5}{\sqrt{50}} \right) \]
   Margin of error = 0.920 (7.774, 8.626)

   b. Explain in words any layman can understand what the 95% confidence interval means.

   We are 95% confident that the true population mean falls between 7.774 g and 8.626 g.

   c. Would the 90% confidence interval be narrower or wider? Explain why.

   Narrower: we are less confident so the range of values of \( \mu \) is smaller.

   d. The manufacturer claims that the alcohol content is 8.0 grams per bottle. Make a brief statement about the manufacturer’s claim.

   This is a reasonable claim, as it falls within the 95% confidence interval.
33. Explain the relationship between the t-test and the z-test in hypothesis testing.

A z-test, based on the standard normal distribution, is used when the population standard deviation is known in a particular setting, whereas a t-test, based on a t-distribution determined by the degrees of freedom, is used when only the sample standard deviation is known. As the degrees of freedom increase, the t-distribution approaches the normal distribution.

The margin of error is the amount by which the estimate is expected to differ from the parameter, in either direction, at a stated confidence interval.

34. What is meant by margin of error in a confidence interval?

35. Are girls less inclined to enroll in science courses than boys? One recent study asked randomly selected 4th, 5th, and 6th graders how many science courses they intend to take in high school. The following data was obtained:

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>203</td>
<td>3.42</td>
<td>1.49</td>
</tr>
<tr>
<td>Females</td>
<td>224</td>
<td>2.42</td>
<td>1.35</td>
</tr>
</tbody>
</table>

a. Calculate a 99% confidence interval for the difference between males and females in mean number of science courses planned. Interpret your interval.

\[
(t^* = 2.60) \\
1 \pm 2.6 \sqrt{\frac{1.49^2}{203} + \frac{1.35^2}{224}}
\]

CI = (0.611, 1.359)

We are 99% confident that the true difference between males and females in mean number of science courses planned lies between 0.611 and 1.359.

b. The science teacher at the high school these students plan on attending claims that there is no difference in the number of courses boys and girls take. Test the science teacher’s claim.

1. Hypothesis: We aim to test the claim that there is no difference in the mean number of science courses boys and girls take.
   \[H_0: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2\]
   Let \( \mu_1 \) be the mean number of science courses males enroll in and let \( \mu_2 \) be the mean number of science courses females enroll in.

2. Conditions: * Two-sample t-test
   ① Assume the male and female samples are SRS’s.
   ② Normality - For each sample, the Central Limit Theorem is satisfied since \( n_1 = 203 \geq 30 \) and \( n_2 = 224 \geq 30 \).
   ③ The samples are independent.

3. Calculations
   \[
   t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\
   = \frac{3.42 - 2.42}{1.49 \sqrt{\frac{1}{203} + \frac{1}{224}}} \\
   = \frac{1}{0.0746} \\
   t = 13.366
   \]
   The two-sample t-test has a test statistic of \( t = 13.366 \) and a p-value of \( 2.225 \times 10^{-11} \) for df = 409.

4. Interpretation - The small p-value \( p = 2.225 \times 10^{-11} \) gives strong evidence to reject \( H_0 \) and conclude that there is a difference in the mean number of science classes taken by males and females.
36. Techniques for processing poultry were examined by a manufacturer of canned chicken. Whole chickens were chilled 0, 2, 8 and 24 hours before being cooked and canned. To determine whether the chilling time affected the texture of the canned chicken, samples were evaluated by trained testers. One characteristic of interest was hardness. Each mean is based on 36 ratings.

<table>
<thead>
<tr>
<th>Chilling Time</th>
<th>0 hour</th>
<th>2 hour</th>
<th>8 hour</th>
<th>24 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Hardness</td>
<td>7.52</td>
<td>6.55</td>
<td>5.70</td>
<td>5.65</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>.96</td>
<td>1.74</td>
<td>1.32</td>
<td>1.50</td>
</tr>
</tbody>
</table>

a. Does the data suggest that there is a difference in mean hardness for chicken chilled 0 hours before cooking and chicken chilled 2 hours before cooking? Use a significance level of 0.05.

Hypothesis:
- \( H_0: \mu_{0 hr} = \mu_{2 hr} \)
- \( H_a: \mu_{0 hr} \neq \mu_{2 hr} \)

Conditions:
1. Two sample t-test
2. The ratings are representative of all test takers' evaluations.
3. Normality - by the Central Limit Theorem, 36 is a sufficient sample size.
4. The ratings for both levels of chilling time are independent.

Calculations:
- \( t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \) where \( s_p \) is the pooled standard deviation.
- Compute \( t \) and compare to the critical value from the t-distribution.

Interpretation:
- Since \( p \) is less than 0.05, we can reject \( H_0 \) and conclude that there is a significant difference in mean hardness between the 0 hour and 2 hour chilling times.

b. Does the data suggest that there is a difference in mean hardness for chicken chilled 8 hours before cooking and chicken chilled 24 hours before cooking?

Hypothesis:
- \( H_0: \mu_{8 hr} = \mu_{24 hr} \)
- \( H_a: \mu_{8 hr} \neq \mu_{24 hr} \)

Conditions:
- See part (a) conditions for two sample t-test
- Normality
- Independence

Calculations:
- Use the same formula as part (a)
- Compute \( t \) and compare to the critical value.

Interpretation:
- Since \( p \) is less than 0.05, we can reject \( H_0 \) and conclude that there is a significant difference in mean hardness between the 8 hour and 24 hour chilling times.

Use a 90% confidence interval to estimate the difference in mean hardness for chicken chilled 2 hours before cooking and chicken chilled 8 hours before cooking.

\[
\left( \bar{X}_1 - \bar{X}_2 \right) \pm t_{0.95} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

\[
\left( 5.55 - 5.70 \right) \pm 1.690 \sqrt{\frac{1.74^2}{36} + \frac{1.32^2}{36}}
\]

Confidence Interval: \( (0.235, 1.465) \)

If a Type I error were made in part a, what would this mean? What are the consequences?

If a Type II error were made in part b, what would this mean? What are the consequences?
37. The discharge of industrial wastewater into rivers affects water quality. To assess the effect of a particular power plant on water quality, 24 water specimens were taken 16 km upstream and 4 km downstream of the plant. Alkalinity (mg/L) was determined for each specimen, resulting in the summary quantities in the table below.

<table>
<thead>
<tr>
<th>Location</th>
<th>n</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream</td>
<td>24</td>
<td>75.9</td>
<td>1.83</td>
</tr>
<tr>
<td>Downstream</td>
<td>24</td>
<td>183.6</td>
<td>1.70</td>
</tr>
</tbody>
</table>

a. Does the data suggest that the true mean alkalinity is higher downstream than upstream by more than 50 mg/L? Perform a hypothesis test.

1. **Hypotheses:**
   - $H_0: \mu_1 = \mu_2$
   - $H_a: \mu_1 < \mu_2$

2. **Conditions:**
   - Two sample t-test
   - Assume both samples of specimens are SRS's
   - $n_1 = n_2 = 24$
   - Assume the distributions are Normal
   - Assume the samples are independent

3. **Calculations**

   $t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{d}}} = \frac{75.9 - 183.6}{1.70} = -211.29$

4. **Interpretation:**
   - The two sample t-test yields a test statistic of $t = -211.29$ and a p-value of $p \approx 0$
   - For $df = 45.752$.

   The small p-value of $p \approx 0$ gives strong evidence to reject $H_0$ and conclude that mean alkalinity is higher downstream.

b. Find the 90% confidence interval for the mean difference. Does this confirm your conclusion in the hypothesis test?

The confidence interval is calculated as:

$$CI = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, df} \left( \frac{s_{\bar{d}}}{\sqrt{n}} \right)$$

For $90\%$ CI and $df = 45.752$, $t_{0.05, df} = 1.674$,

$$CI = (75.9 - 183.6) \pm 1.674 \left( \frac{1.70}{\sqrt{24}} \right)$$

The confidence interval is approximately $(-107.7, 105.9)$.

Yes. The interval does not fall within the interval, so the conclusion from the hypothesis test is confirmed.

8. The article “Softball Sliding Injuries” provided a comparison of breakaway bases (designed to reduce injuries) and stationary bases. Consider the accompanying data. Does the use of breakaway bases reduce the proportion of games in which a player suffers a sliding injury? Perform the test at a 1% significance test.

<table>
<thead>
<tr>
<th></th>
<th>Number of Games Played</th>
<th>Number of Games Where a Player Suffered a Sliding Injury</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stationary Bases</td>
<td>1250</td>
<td>90</td>
</tr>
<tr>
<td>Breakaway Bases</td>
<td>1250</td>
<td>20</td>
</tr>
</tbody>
</table>

1. **Hypotheses:**
   - $H_0: p_1 = p_2$
   - $H_a: p_1 < p_2$

2. **Conditions:**
   - Two proportion z-test
   - Assume the games are SRS’s
   - Normality: $n_1(\hat{p}_1) = 1250(0.044) > 5$
   - $n_2(\hat{p}_2) = 1250(0.044) > 5$

3. **Calculations**

   $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{0.044 - 0.044}{\sqrt{\frac{0.044(1-0.044)}{1250} + \frac{0.044(1-0.044)}{1250}}} = 0$$

4. **Interpretation:**
   - We can reject $H_0$ since $z$ is small at $\alpha = 0.01$. The proportion of games in which a player suffers a sliding injury is reduced by the use of breakaway bases.
39. The color vision of birds plays a role in their foraging behavior. Birds use color to select and avoid certain types of food. The authors of the article “Color Avoidance in Northern Bobwhites” studied the pecking behavior of 1-day-old bobwhites. In an area painted white, they inserted four pins with different colored heads. The color of the pin chosen on the birds first peck for each of 33 bobwhites, resulting in the accompanying table. Does this data provide evidence of color preference? Test at the .1 significance level.

<table>
<thead>
<tr>
<th>Color</th>
<th>Blue</th>
<th>Green</th>
<th>Yellow</th>
<th>Red</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Peck Frequency</td>
<td>16</td>
<td>8</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

\[H_0: P_{\text{blue}} = P_{\text{green}} = P_{\text{yellow}} = P_{\text{red}} = \frac{1}{4}\]

We want to compare the observed counts for first peck frequency with expected counts.

\[H_1: \text{at least 1 of the proportions differ from the expected value}\]

\[\chi^2 = 6.05\]

**Calculation:**

\[\chi^2 = \frac{(O - E)^2}{E}\]

\[O = \text{observed count}, E = \text{expected count}\]

**Interpretation:** Since \(p = 0.0105 < 0.05 = 0.01\), we can reject \(H_0\) and conclude that the distribution of first peck frequency is not uniform.

40. Do women have different patterns of work behavior than men? The article “Workaholism in Organizations: Gender Differences” attempts to answer this question. Each person in a random sample of 423 graduates of a business school in Canada were polled and classified by gender and workaholism type, resulting in the accompanying table:

<table>
<thead>
<tr>
<th>Workaholism</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Enthusiasts</td>
<td>20</td>
<td>41</td>
</tr>
<tr>
<td>Workaholics</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>Enthusiastic Workaholics</td>
<td>34</td>
<td>46</td>
</tr>
<tr>
<td>Unengaged Workers</td>
<td>43</td>
<td>52</td>
</tr>
<tr>
<td>Relaxed Workers</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>Disenchanted Workers</td>
<td>37</td>
<td>30</td>
</tr>
</tbody>
</table>

a. Test the hypothesis that gender and workaholism type are independent.

\[H_0: \text{Gender and workaholism type are independent}\]

\[H_1: \text{Gender and workaholism type are dependent}\]

**Conditions:** \(\chi^2\) for independence

\[\chi^2 = \frac{\sum(O - E)^2}{E}\]

\[O = \text{observed count}, E = \text{expected count}\]

**Interpretation:** The large p-value gives strong evidence to fail to reject \(H_0\) and conclude that gender and workaholism type are independent.

b. The author writes “women and men fell into each of the six workaholism types to a similar degree.” Does the outcome of the test you performed in part a support this conclusion? Explain.

Yes, the distribution of workaholism for females and males are the same according to the statistical test.
It is certainly plausible that workers are less likely to quit their jobs when wages are high than when they are low. The paper "Investigating the Causal Relationship Between Quits and Wages" presented the accompanying data on $x =$ average hourly wages and $y =$ quit rate (number of employees per 100 who left jobs during 1986). Each observation is for a different industry.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1.4</td>
<td>.7</td>
<td>2.6</td>
<td>3.4</td>
<td>1.7</td>
<td>1.7</td>
<td>1.0</td>
<td>.5</td>
<td>2.0</td>
<td>3.8</td>
<td>2.3</td>
<td>1.9</td>
<td>1.4</td>
<td>1.8</td>
<td>2.0</td>
</tr>
</tbody>
</table>

The following is the Minitab output:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>Std Dev</th>
<th>$t$-ratio</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>@ 4.8615</td>
<td>0.5201</td>
<td>9.35</td>
<td>0.0000</td>
</tr>
<tr>
<td>Wage</td>
<td>@ -0.34655</td>
<td>0.3605866</td>
<td>-5.91</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

$s = 0.4862$  \quad R-$sq = 72.9\%$  \quad R-$sq(adj) = 70.8\%$

Analysis of Variance

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>8.2507</td>
<td>8.2507</td>
<td>34.90</td>
<td>0.0000</td>
</tr>
<tr>
<td>Error</td>
<td>13</td>
<td>3.0733</td>
<td>0.2364</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>11.3240</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Identify the slope and y-intercept for the LSRL for average hourly wages and quit rate.

slope = $b = -0.34655$

$y -$intercept = $a = 4.8615$

b. What is the LSRL?

$\hat{y} = 4.8615 - 0.34655x$

df = 13

$t* = 2.160$

$-0.34655 \pm 2.160 (0.05866)$

Confidence Interval: $(-.4733, .2198)$

c. Find the 95\% confidence interval for the slope of the line.

d. Test the hypothesis that there is a linear relationship between average hourly wages and quit rate.

1) Hypotheses: $H_0: \beta = 0$  \quad $H_a: \beta \neq 0$

2) Conditions:

1) Linear relationship between $x$ and $y$

2) Residuals are scattered and have same vertical spread

3) Calculations: The linear regression $t$-test yields a test statistic $t = -5.908$ and a $p$-value of $p = 5.171 \times 10^{-5}$ for df = 13.

4) Interpretation: The small $p$-value gives strong evidence to reject $H_0$ and conclude that there is a linear relationship between average hourly wages and quit rate.