Question 1: Observational Studies and Experiments (12 Points)

According to a study done at Kaiser Permanente in Walnut Creek, California, users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status. Investigators concluded that the pill causes cervical cancer. Answer the following three questions:

1. Is this a controlled experiment or an observational study? Circle your answer.
2. Why did the investigators adjust for age, education, and marital status?

(b) Age, education, and marital status are variables that are known to have associations with cervical cancer and pill use. They adjusted for these variables to try to eliminate problems due to these confounding factors.

3. Were the conclusions of the study justified by the data? Answer yes or no and explain briefly.

(d) The conclusions are not justified by the study. There are several possible confounding factors that were not adjusted for. For example, the number of sexual partners is likely to be higher of pill-users than non-pill users, and it could be that cervical cancer is a sexually transmitted disease.

Question 2: Histograms (8 Points)

The following three histograms are based on the Old Faithful data set. The observations are the duration (in minutes) for eruptions of the Old Faithful geyser in Yellowstone National Park. There exist two types of eruptions: shorter ones (about 2 minutes) and longer ones (about 4 minutes). Which of the three histograms best describes the underlying data. Shortly explain your answer and indicate why you think the other 2 histograms don't represent the data as well as the one you have selected.

- 1 if "explains the description of the data from above, i.e., the 2 types of eruptions are clearly visible"
- 2 if "spiky [too many classes - not clear, not the 2 types]"
- 3 if "spiky [too many classes - not clear, not the 2 types]"
- 4 if "spiky [too many classes - not clear, not the 2 types]"

Please turn over!
Statistics 1040, Section 006, Quiz 2 (20 Points)
September 13, 2002

Your Name: ______________________

Question 1: Measures of Center and Spread (20 Points)

The table below, published in USA Today on Friday, May 15, 1998, lists the 15 most widely held stocks and their change year-to-date (YTD). Suppose we hold one share each of AT&T (-6.5), Bell Atlantic (+0.1), Coca-Cola (+16.6), Merck (+11.4), and SBC Comm. (+3.4).

1. Find the average change (YTD) for the 5 stocks we own. Show your work!

\[ \text{avg} = \frac{-6.5 + 0.1 + 16.6 + 11.4 + 3.4}{5} \]

\[ = \frac{25}{5} \]

\[ = 5 \]

-3 if only final result (correct)
-2 for incorrect final result (Wrong)
-2 for incorrect exam
-2 for missing value (in sum)

2. Find the median change (YTD) for the 5 stocks we own.

\text{sorted list: } -6.5 \ 0.1 \ 3.4 \ 11.4 \ 16.6

\text{Center value: } 3.4

\text{median: } 3.4

-3 for incorrect median

3. Find the standard deviation of the changes (YTD) for the 5 stocks we own. Show your work!

\[ \text{avg} = 5 \]

\[ -6.5 - 5 = -11.5 \]

\[ 0.1 - 5 = -4.9 \]

\[ 16.6 - 5 = 11.6 \]

\[ 11.4 - 5 = 6.4 \]

\[ 3.4 - 5 = -1.6 \]

\[ (-11.5)^2 = 132.25 \]

\[ (-4.9)^2 = 24.01 \]

\[ (11.6)^2 = 134.56 \]

\[ (6.4)^2 = 40.96 \]

\[ (11.6)^2 = 2.56 \]

\[ \text{sum of deviations from avg} = 334.39 \]

\[ \text{average of deviations from avg} = 66.878 \]

\[ \sqrt{66.878} = 8.177 \approx 8.2 \]

Formulas:

\[ \text{avg} = \frac{\text{sum of all numbers}}{\text{how many numbers}} \]

\[ \text{SD} = \sqrt{\frac{\text{average of (deviations from avg)^2}}{}} \]
Statistics 1040, Section 006, Quiz 3 (20 Points)
Sep 20, 2002

Your Name: __________________________

Question 1: Normal Approximation for Data (20 Points)

Car drivers in the United States average 12,400 miles a year, nearly 50 percent more than European drivers (The Economist, June 22, 1996). Assume that the number of yearly miles by U.S. drivers approximately follows a normal curve with a standard deviation of 3,200 miles.

1. Determine the percentage of drivers who travel between 10,000 and 15,000 miles in a year.

\[ z = \frac{X - \mu}{\sigma} = \frac{10000 - 12400}{3200} = -0.75 \text{ z.u.} \]
\[ \frac{15000 - 12400}{3200} = 0.81 \approx 0.80 \text{ z.u.} \]

2. Find the area (percentage) related to the numbers from (1) (use Table)

Area from -0.75 to 0.80: 0.64%
Area from -0.75 to 0: 0.54%
Area from 0 to 0.80: 0.57%
Area from -0.75 to 0.80: 3.09%
Area from 0.80 to 1.70: 54.11%
Area from 1.70 to 3.09: 65.61%

3. Find the area (percentage) of interest:

Area from 0.75 to 2.00: 2.33%
Area from 0.75 to 0.60: 0.38%
Area from 0.60 to 2.00: 2.33% + 0.38% = 2.71%

4. What percentage of all U.S. drivers travel between 0,000 and 15,000 miles a year (in %)?

5. And what percentage of drivers travels more than 30,000 miles in a year?

\[ z = \frac{X - \mu}{\sigma} = \frac{30000 - 12400}{3200} = 3.56 \text{ z.u.} \]

Note: 3.56 is not in table! But notice area from -3.5 to 3.5 is even closer to 100%

Show your work!
Statistics 1040, Section 006, Quiz 4 (20 Points)

September 27, 2002

Your Name: ______________________

Question 1: Percentiles and the Normal Curve (12 Points)

The Graduate Record Examination (GRE) is a test taken by college students who intend to pursue a graduate degree in the United States. For all college seniors and graduates who took the exam in the past few years, the mean score for the verbal ability portion of the exam was 497 with a standard deviation of 115. Assuming the scores are bell-shaped, fill in the blanks below. Show your work!

1. A student who received a score of 650 on the verbal ability portion of the GRE exam was at the __________ th percentile of the score distribution.

2. A graduate school program in English will admit only students with GRE verbal ability scores in the top 30%. Therefore, the lowest GRE score they will accept is __________. Choose one of the options below, and explain.

   a. 500
   b. 550
   c. 600
   d. 650

3. The correlation between the ages of husbands and wives in the U.S. is __________. Choose one option, and explain.

   a. 0.45
   b. 0.55
   c. 0.65
   d. 0.75

4. Options: (a) exactly -1 (b) close to -1 (c) close to 0 (d) close to 1

   a. 0.80
   b. 0.90
   c. 1.00
   d. 1.20

5. The correlation between the ages of husbands and wives is about __________; the husbands were, on average, __________ years older than their wives.

   a. 0.95
   b. 0.90
   c. 0.85
   d. 0.75

6. A normal table

   a. 0.00
   b. 0.05
   c. 0.10
   d. 0.15

7. Close to 1; this is like part (a), with some noise thrown into the data.

   a. 0.25
   b. 0.30
   c. 0.35
   d. 0.40

8. Choose one option, and explain.

   a. 0.50
   b. 0.55
   c. 0.60
   d. 0.65

9. Close to 0; smaller than part (a), with some noise thrown into the data.

   a. 0.75
   b. 0.80
   c. 0.85
   d. 0.90

10. Choose one option, and explain.

    a. 0.95
    b. 1.00
    c. 1.20
    d. 1.35

11. Close to 0; smaller than part (a), with some noise thrown into the data.

    a. 0.90
    b. 0.95
    c. 1.00
    d. 1.05

12. Options: (a) exactly -1 (b) close to -1 (c) close to 0 (d) close to 1

    a. 0.75
    b. 0.80
    c. 0.85
    d. 0.90

13. Choose one option, and explain.

    a. 0.95
    b. 1.00
    c. 1.05
    d. 1.10

14. Close to 1; this is like part (a), with some noise thrown into the data.

    a. 0.85
    b. 0.90
    c. 0.95
    d. 1.00

15. Choose one option, and explain.

    a. 0.95
    b. 1.00
    c. 1.05
    d. 1.10

16. Options: (a) exactly -1 (b) close to -1 (c) close to 0 (d) close to 1

    a. 0.85
    b. 0.90
    c. 0.95
    d. 1.00

17. Choose one option, and explain.

    a. 0.95
    b. 1.00
    c. 1.05
    d. 1.10
Statistics 1040, Section 006, Quiz 5 (20 Points)
October 11, 2002

Your Name: ____________________

Question 1: The Regression Line (20 Points)

In one study, the correlation between the educational level of husbands and wives in a certain town was about 0.50; both averaged 12 years of schooling completed, with an SD of 3 years.

1. Find the regression equation for predicting the educational level of a wife from the educational level of her husband. (10 Points)
   \[ \text{Slope} = r \times \frac{SD_y}{SD_x} = 0.50 \times \frac{3}{3} = 0.50 \]
   \[ \text{Intercept} = \text{avg}_y - \text{Slope} \times \text{avg}_x = 12 - 0.50 \times 12 = 6 \]
   \[ \text{Regression equation:} \quad \text{educ level of wife} = 6 + 0.50 \times \text{educ level of husband} \]
   or \[ y = 6 + 0.50 \times x \]

2. Use the regression equation from part 1. to predict the educational level of a woman whose husband has completed 18 years of schooling. (5 Points)
   \[ \text{educ level of wife} = 6 + 0.50 \times 18 = 6 + 9 = 15 \]
   (i.e., 15 years of education in the predicted average education level for a woman whose husband has completed 18 years of schooling)

3. Find the r.m.s. error for predicting the wife's educational level from the educational level of her husband. (5 Points)
   \[ \text{r.m.s. error} = \sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.25} \times 3 = \sqrt{0.75} \times 3 = 0.866 \times 3 = 2.598 \approx 2.6 \]

Please turn over!

(i.e., on average, the predicted education level is 2.6 years off the observed education level)

Formulas:

- r.m.s. error = \( \sqrt{1 - r^2} \times SD_y \)
- slope = \( r \times \frac{SD_y}{SD_x} \)
- intercept = \( \text{avg}_y - \text{slope} \times \text{avg}_x \)

Grading Criteria:

- 2 for each calculation error
- 2 for each incorrect value used
- 2 for missing x and y
- 2 for incorrect formula for slope
- 3 for incorrect formula for intercept
- 2 if no final equation stated
- 1 if only one of the equation stated \( y = 6 + 0.50 x \)
- 3 for incorrect formula for prediction
- 1 if correct results, but according to old method
- 3 for incorrect formula for r.m.s. error
Statistics 1040, Section 006, Quiz 6 (20 Points)
October 18, 2002

Your Name: ____________________

Question 1: Chance/Probability (20 Points)

1. A deck of 52 cards is shuffled and two cards are drawn without replacement.

   (a) (3 Points) What is the chance that the first card is a ♠ or a ♦?
   
   \[
   \text{Chance that 1st is ♠ or ♦} = \frac{13}{52} \times \frac{12}{51} = \frac{26}{52} \times \frac{1}{2} = 0.5 \times 0.5 = 0.25.
   \]

   (b) (4 Points) What is the chance that the first card is a ♠ and the second card is a ♦?
   
   \[
   \text{Chance that 1st is ♠ and 2nd is ♦} = \frac{13}{52} \times \frac{12}{51} = \frac{1}{4} \times \frac{1}{13} \times \frac{12}{51} = 0.0637 = 0.637\%.
   \]

   (c) (4 Points) What is the chance that both cards are ♠?
   
   \[
   \text{Chance that both cards are ♠} = \frac{13}{52} \times \frac{12}{51} = \frac{1}{4} \times \frac{1}{13} \times \frac{12}{51} = 0.0637 = 0.637\%.
   \]

   (d) (4 Points) What is the chance that neither card is a ♠?
   
   \[
   \text{Chance that neither card is ♠} = \frac{39}{52} \times \frac{38}{51} = \frac{3}{4} \times \frac{38}{51} \times \frac{1}{2} = 0.3846 = 38.46\%.
   \]

2. (5 Points) There are two options:

   (a) You toss a coin 100 times; on each toss, if it lands heads you win $1, if it lands tails you lose $1.

   (b) You draw 100 times at random with replacement from the box \[1, 0\].

   (i) (a) You draw 100 times at random with replacement from the box \[1, 0\].

   On each draw, you are paid (in dollars) the number on the ticket.

   Which option is better? Or are they the same? Explain briefly.

Chapter 13, Review question 11; page 235:

Workbook answer:

"Option (a) is better: you lose 50% chance of winning 40 dollars, but for option (b) you have no way to lose money."
Statistics 1040, Section 006, Quiz 7 (20 Points)
October 25, 2002

Your Name: ____________________________

Angel 17, Period 4, p. 305

Question 1: Box Models, EV, and SE (15 Points)

A large group of people get together. Each one rolls a die 180 times and counts the number of ones. Show your work!

1. (4 Points) Find the box model.

```
1
2
3
```

number of draws: 180

```
1
2
3
```

-2 for minor error
-3 for major error (e.g., box 1 has only 1)

2. (5 Points) Find the expected value for the number of ones.

```
EV_{sum} = 180 \cdot \frac{1}{6} = 30
```

3. (6 Points) Find the standard error.

```
box SD = \sqrt{\frac{1^2 + 5 \cdot \frac{-1}{6}^2}{6}}
```

```
= \sqrt{\frac{\frac{25}{36} + \frac{5}{36}}{6}}
```

Please turn over!

```
= \frac{\sqrt{\frac{30}{36}}}{2} \approx 0.3727
```

```
SE_{sum} = \sqrt{\frac{180^2 \cdot 0.3727}{13.416 \cdot 0.3727}} = 5.00
```

Question 2: Law of Averages (5 Points)

The meaning of "The probability of a Head is 1/2" in tossing a coin is best expressed by saying:

1. The coin has only two sides, so the chance of each is 1/2.
2. The coin will come up Heads exactly half the time: 50 Heads in 100 tosses, 500 Heads in 1000 tosses, and so on.
3. The odds against a Head are 2 to 1.
4. The fraction of tosses that come up Head will get ever closer to 1/2 as more tosses are made.

Explain your answer!

1. Correct, but not the best choice
2. False - this is very unlikely to happen
3. False - actually the odds against a Head are 1 to 1
4. Correct, but not best; another way how to phrase the law of averages:

Formulas:

\[ box\ average = \frac{\text{sum of all numbers in box}}{\text{how many numbers in box}} \]

\[ box\ SD = \sqrt{\frac{\text{average of [(deviations from box average)]}^2}{\text{number of draws}}} \]

\[ EV_{sum} = \text{number of draws} \times \text{box average} \]

\[ SE_{sum} = \sqrt{\text{number of draws} \times \text{box SD}} \]
Statistics 1040, Section 006, Quiz 8 (20 Points)
November 1, 2002

Your Name: _______________________

Question 1: EV, SE, and Normal Curve (20 Points)

According to the U.S. Census Bureau, 68% of Utah residents are 18 years of age or older. Suppose that 200 Utah residents have been randomly chosen to participate in a survey.

1. (4 Points) Find the box model.

\[
\begin{array}{c|c|c|c}
68 & 32 & 0 \\
\hline
\end{array}
\]

\text{number of draws = 200}

2. (8 Points) Find the expected number of Utah residents in this sample of 200 who are 18 years of age or older. What is the corresponding SE?

\[
\text{box average} = \frac{68}{100} \cdot 200 = 0.68 \cdot 200 = 136
\]

\[
\text{box SD} = \sqrt{\frac{68}{100} \cdot \frac{32}{100} - \frac{0.68 \cdot 0.32}{100}} = \sqrt{0.2176} = 0.466
\]

\[
\text{EV}_{\text{sum}} = 200 \cdot 0.68 = 136
\]

\[
\text{SE}_{\text{sum}} = \sqrt{200} \cdot 0.466 = 14.14 \cdot 0.466 = 6.59
\]

3. (8 Points) Using the normal curve, find the chance that at least 130 of the Utah residents in the sample are 18 years of age or older.

\[
\text{s.u.: } \frac{130 - 136}{6.59} = -0.91
\]

\[
\text{area between -0.90 and 0.90} = 63.18\%
\]

\[
\text{area above -0.90} = 50\% + \frac{63.18\%}{2} = 81.59\% \approx 81.6\%
\]

There is a chance of about 81.6% that at least 130 of the Utah residents in the sample are 18 years of age or older.

Please turn over!
Statistics 1040, Section 6, Quiz 9 (20 Points)
November 15, 2002

Your Name: ______________________

Chapter 20, Review question 3 (c), p. 371

Question 1: EV%, SE%, and Normal Curve (20 Points)

A group of 50,000 tax forms has an average gross income of $37,000, with an SD of $20,000. Furthermore, 20% of the forms have a gross income over $50,000. A group of 900 forms is chosen at random for audit. Estimate the chance that between 19% and 21% of the forms chosen for audit have gross income over $50,000. Show your work!

\[
\text{1: income over } \geq 50,000 \\
\text{0: income under } < 50,000 \\
\text{Note: avg = 37,000 & SD = 20,000 are not used to answer this question!}
\]

\[
\text{Note: } \frac{20 \times 1}{20 \times 1} = \frac{80 \times 0}{100} \quad \text{or} \quad \frac{10,000 \times 1}{10,000 \times 0} = \frac{40,000 \times 0}{900}
\]

\[
\text{number of draws: } 900
\]

\[
\text{box var} = \frac{20}{100} = \frac{10,000}{50,000} = 0.2
\]

\[
\text{box SD} = \sqrt{\frac{20}{100}} = \sqrt{\frac{10,000}{50,000}} = \sqrt{\frac{40,000}{50,000}} = 0.2 \cdot 0.8 = \sqrt{0.16} = 0.4
\]

\[
\text{EV}_\% = 20\% \\
\text{SE}_{\text{sum}} = \sqrt{900} \cdot 0.4 = 30 \cdot 0.4 = 12
\]

\[
\text{SE}_\% = \frac{12}{900} \cdot 100\% = 1.33\% \\
\text{S.A.}: \frac{19\% - 20\%}{1.33\%} = -0.75
\]

\[
\frac{21\% - 20\%}{1.33\%} = 0.75
\]

\[
\text{area between } -0.75 \text{ to } 0.75: 54.67\% \times 55\%
\]

Please turn over!
Question 1: Confidence Intervals (20 Points)

A real estate office wants to make a survey in a certain town, which has 50,000 households, to determine how far the head of household has to commute to work. A simple random sample of 1,000 households is chosen, the occupants are interviewed, and it is found that on average, the head of the sample households commuted 8.7 miles to work; the SD of the distances was 9.0 miles. (All distances are one-way; if someone isn't working, or is working at home, the commute distance is defined to be 0.)

Fill the blanks in the statements below and show your work!

1. (10 Points) The average commute distance of all 50,000 heads of households in the town is estimated as \( \frac{8.7}{1} \) mile, and this estimate is likely to be off by \( \frac{0.3}{1} \) mile or so.

   \[
   \text{lose: unknown} \\
   \text{lose \( \bar{x} \) = sample \( \bar{x} \) = 8.7} \\
   \text{lose SD = sample SD = 9.0} \\
   \text{SE_{sum} = \sqrt{100} \cdot 9.0 = 31.6 \cdot 9.0 = 284.4} \\
   \text{SE_{avg} = \frac{284.4}{1000} = 0.2844 \cong 0.3}
   \]

2. (10 Points) If possible, find a 95%-confidence interval for the average commute distance of all heads of households in the town. If this isn't possible, explain why not.

   Note that the data does not follow the normal curve, otherwise:

   for possible, any calculation based on any
   But:
   the average of 1000 draws will follow the normal curve.
   So we can obtain confidence intervals:

   \( 8.7 \pm 2 \cdot 0.3 = 8.7 \pm 0.6 \)

   \( 8.7 \pm 2 \cdot 0.3 = 8.1 \) to \( 9.3 \) miles

   Please turn over!
Chapter 26, Review question 8, p. 493

Question 1: Tests of Significance (20 Points)

Bookstores like education, because national data show that 71% of college graduates have read a book in the past year, compared to 54% of the general population age 18 and over. The data also show the nationwide average educational level to be 13 years of schooling completed, with an SD of about 3 years, for persons age 18 and over. A bookstore is doing a market survey in a certain county, and takes a sample of 1,000 people age 18 and over. They find the average educational level to be 14 years, and the SD is 5 years. Can the difference in average educational level between the sample and the nation be explained by chance variation? If not, what other explanations can you give? Please follow the steps below in answering these questions.

1. (5 points) State the null and the alternative hypothesis for this problem, in words and in terms of the box model.
   
   Null: County's avg education level matches nationwide level, i.e., box avg = 13 years
   
   Alternative: County's avg education level is different from nation, i.e., box avg ≠ 13 years

2. (5 points) Calculate the appropriate test statistic.

   $\bar{x}_{\text{observed}} = 14$
   
   $\bar{x}_{\text{expected}} = 13$
   
   $SE_{\text{sum}} = \sqrt{\frac{1}{1000} \cdot 5} = 158.1$
   
   $SE_{\text{avg}} = \frac{158.1}{1000} = 0.158$
   
   $z = \frac{14 - 13}{0.158} = 6.33$

3. (5 points) Obtain the P-value (use the normal table on the back).

   P-value is area on both sides, but this is about 0%.

4. (5 points) State conclusions in terms of rejecting the null hypothesis and in your own words.

   reject null hypothesis; result is highly statistically significant (P-value < 0.01);
   
   the education level for this county is different from the national average, more
   
   specifically it is above the national average; this could be a rich suburban county.
Grading Criteria:

1) wrong null and alternative hypothesis, e.g.:
   - wrong null and alternative
   - "14" in hypothesis instead of "13"
   - hypothesis stated in words only or in numbers only

   -3
   -3
   -1 each

2) incorrect t, e.g. \( \frac{\text{exp.-obs}}{SE} \) or \( \frac{\text{obs.-exp.}}{SD} \)
   - incorrect SE avg
   - calculation error

   -2
   -2
   -1 each

3) incorrect even
   - incorrect tibble value
   - calculation error

   -2
   -2
   -1

4) reject null if p-value > 5%
   (or do not reject null if p-value < 5%)
   - no explanation (e.g., reject, but no conclusion)
   -2
   - correctly rejecting, but explanation mixed up
   -2
   - if not speaking of rejecting/not rejecting
   -2
   - if not speaking of (highly) statistically significant
   -1
Statistics 1040, Section 006, Quiz 12 (20+ Points)
Due on or before December 11, 2002

Your Name: ____________________________

This is a take-home quiz. You should work on it on your own and bring it to me on or before the final examination day. Please work on this quiz independently, getting as little help as possible from your friends, books, and notes.

Question 1:

(20 Points) A thermostat used in an electrical device is to be checked for the accuracy of its design setting of 200 degrees Fahrenheit. Ten thermistats were tested to determine their actual setting, resulting in the following data:

202.2  203.4  200.4  202.5  206.3  198.0  203.7  200.8  201.3  199.0

Is the mean setting of these thermometers different from 200 degrees Fahrenheit? State the null and the alternative hypothesis, calculate test statistic (after finding the average and SD of the sample), obtain the P-value, and clearly state your conclusions. Assume that the thermometer settings follow the normal curve.

\[
\bar{x} = \frac{202.2 + \ldots + 199.0}{10} = 201.76
\]

\[
SD = \sqrt{\frac{(202.2 - 201.76)^2 + \ldots + (199.0 - 201.76)^2}{10}} = 2.29
\]

1. Null: Mean setting of thermometers does not differ, i.e., \( \bar{x} = 200 \) F
2. Alternative: Mean setting of thermometers differs, i.e., \( \bar{x} \neq 200 \) F

3. Sample size is large: \( n \geq 30 \)

4. SD unknown, data follows normal curve.

5. \( SD^t = 2.29 \cdot \sqrt{\frac{10}{9}} = 2.41 \)

6. \( SE_{\bar{x}} = \sqrt{10} \cdot 2.41 = 7.62 \)

7. \( SE_{\bar{x}} = 7.62 \cdot \frac{10}{10} = 0.76 \)

8. \( t = \frac{201.76 - 200}{0.76} = 2.32 \)

4. Reject null, result is statistically significant (P-value between 2% and 5%).

There is some evidence that the mean setting of thermometers differs from 200 F.
The following questions are extra-credit questions. You may obtain a maximum of 20 extra-points if you complete both questions.

**Question 2:**

(10 Points) In an experiment to study the dependence of hypertension on smoking habits, the following data were taken on 180 individuals:

<table>
<thead>
<tr>
<th>obs</th>
<th>exp</th>
<th>Nonsmokers</th>
<th>Moderate Smokers</th>
<th>Heavy Smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>21</td>
<td>24</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>No hypertension</td>
<td>21</td>
<td>18</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>62</td>
<td>62</td>
<td>62</td>
</tr>
</tbody>
</table>

Is the presence or absence of hypertension independent of smoking habits? Conduct an appropriate statistical test to answer this question.

1. null: hypertension and smoking are independent, i.e., boxes are the same
2. alternative: hypertension and smoking are dependent, i.e., at least 1 box differs

1. **test for independence**

    Expected: \( \frac{42.87}{155} = 22.6 \approx 24 \)
    
    \( \frac{62.87}{155} = 34.8 \approx 35 \)
    
    \( \frac{51.87}{155} = 28.6 \approx 29 \)

    \( \chi^2 \approx 18 \)

2. **test for independence**

    \( \chi^2 = \frac{(21-24)^2}{24} + \frac{(36-35)^2}{35} + \frac{(30-29)^2}{29} + \frac{(21-18)^2}{18} + \frac{(26-27)^2}{27} + \frac{(21-22)^2}{22} \)

    \( \chi^2 \approx 10.2 \)

3. df = \((2-1) \cdot (3-1) = 2 \)

4. **test for independence**

   \( \chi^2 = 1.02 \) between 0.71 and 1.39

   \( P \)-value is between 70% and 50%.

**Question 3:**

(10 Points) A study was made to estimate the difference in salaries of college professors in private and state colleges of North Carolina. A random sample of 100 professors in private colleges showed an average 9-month salary of \$32,000 with a standard deviation of \$1300. A random sample of 200 professors in state colleges showed an average salary of \$32,900 with a standard deviation of \$1400. Is there any statistical evidence that professors in state colleges have higher average salaries than professors in private colleges? Conduct an appropriate statistical test to answer this question.

1. do not reject null based on the particular study

   hypertension and smoking are independent, i.e., this study doesn't provide enough evidence that smoking is associated with hypertension (although another study might show...
Question 3

Box A: Private
Avg A: 32,000
SD A: 1,300
Sample size A: 100

Box B: State
Avg B: 32,900
SD B: 1,400
Sample size B: 200

1. Null: any salaries are the same, i.e., \( \text{Avg}_B - \text{Avg}_A = 0 \)
Alternative: any salaries in state colleges are higher, i.e., \( \text{Avg}_B - \text{Avg}_A > 0 \)

2. 2-sample z-test:

Obs. difference: \( 32,900 - 32,000 = 900 \)
Exp. difference: 0

\[
\begin{align*}
SE_{\text{Avg}_A} &= \sqrt{\frac{100 \cdot 1,300}{100}} = 130 \\
SE_{\text{Avg}_B} &= \sqrt{\frac{200 \cdot 1,400}{200}} = 19.799 \\
SE_{\text{diff}} &= \sqrt{130^2 + 19.799^2} = 163.4 \\
2 &= \frac{900 - 0}{163.4} = 5.5
\end{align*}
\]

3.

4. reject null;
result is highly statistically significant (P-value < 1%)
there is considerable evidence that professors have higher average salaries in state colleges than in private colleges (in North Carolina)