Biostatistics Final Exam Prep

Problem 1

In some city the only type of precipitation it gets is rain; it rains in this city about 35% of the time. When it rains the buses arrive late about 30% of the time, and when it's not raining the buses arrive on time about 90% of the time.

a) Write probability statements for the given information:

b) Create a tree diagram for the conditional probabilities. If a bus is late, what is the probability that it is raining?

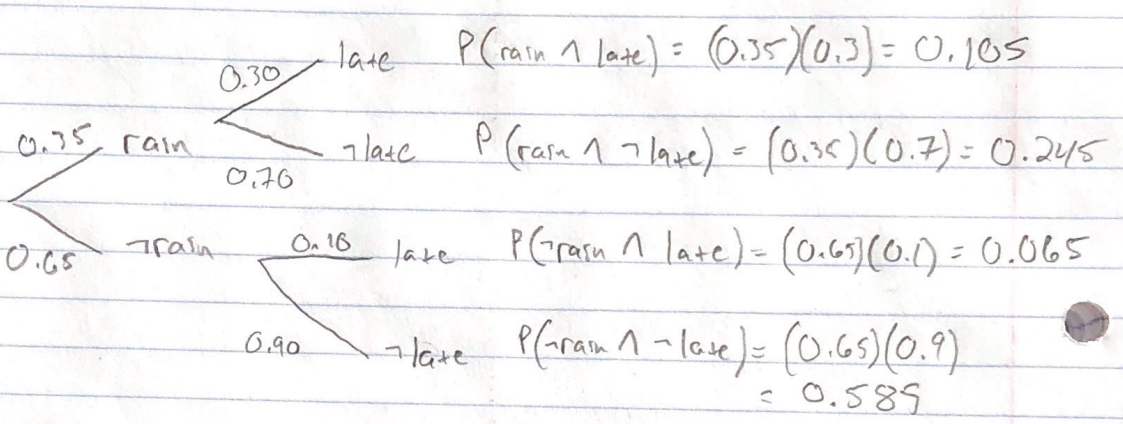
Problem 1

a)

$P(\text{rain}) = 0.35$ $\quad P(\text{late} \mid \text{rain}) = 0.30$

$P(\neg\text{late} \mid \neg\text{rain}) = 0.90$

b)



We want to determine $P(\text{rain} \mid \text{late})$

$$P(\text{rain} \mid \text{late}) = \frac{P(\text{rain} \wedge \text{late})}{P(\text{late})} = \frac{0.105}{0.105 + 0.065}$$

because $P(\text{late}) = P(\text{rain} \wedge \text{late}) + P(\neg\text{rain} \wedge \text{late})$

$= 0.105 + 0.065$

so $P(\text{rain} \mid \text{late}) \approx 0.62$

Problem 2

From this list:

- Confidence interval for a population mean
- Confidence interval for proportion
- Confidence interval for paired data
- Confidence interval for the difference between 2 proportions
- Confidence interval for the difference between 2 means
- Hypothesis test for a population mean
- Hypothesis test for a population proportion
- Hypothesis test for paired data
- Hypothesis test for the difference between 2 proportions
- Hypothesis test for the difference between 2 means

classify the following statistical problems by matching the appropriate analysis procedure with each situation:

a) How much of the food that a person buys ends up being thrown out? A refrigerator was monitored that had 58 perishable items.

Confidence interval for proportion

b) Before the the recent increase to a city's businesses' taxes, 68% of all new businesses closed down within one year of opening. 48 businesses that have opened since the tax increase are being tracked. Has the new business rate of failure increased since the tax increase?

Hypothesis test for population proportion

c) 500 drivers using cell phones and 700 drivers not using cell phones were observed to compare the accident rates. How much more likely are drivers that use their cell phones while driving to get into an accident?

Confidence interval for the difference between 2 proportions

d) 60 college students and 60 high school students were asked what percent of their free time they spent listening to podcasts. Estimate the difference in time that high school students spend listening to podcasts vs college students.

Confidence interval for the difference between 2 means

e) 78 University of Iowa students were asked how many cups of coffee they drank each week and how many glasses of water they drank each week. Estimate the mean difference between the average number of cups of coffee and the average number of glasses of water that University of Iowa students drink each week.

Confidence interval for paired data

f) Will plants grow larger in soil treated with a phosphorus fertilizer vs a nitrogen fertilizer? 62 plants were grown in phosphorus fertilized soil and 62 plants in nitrogen fertilized soil.

Hypothesis test for the difference between 2 means

g) At ISU, the average student takes 14 semester hours. 47 University of Iowa students were asked how many semester hours they were planning to take next semester. Is there evidence to suggest that University of Iowa students take fewer semester hours on average than ISU college students?

Hypothesis test for a population mean

h) Are people more likely to die from a stroke or respiratory disease? 8,000 death certificates were examined.

Hypothesis test for the difference between 2 proportions

i) Do more car accidents happen during the daylight hours compared to night? Police records from 180 consecutive days were considered.

Hypothesis test for paired data

j) You want to estimate the average number of hours that college students spend watching Netflix each week. You survey 200 college students.

Confidence interval for a population mean

Problem 3

We want to see if whether a person is introverted or extroverted is related to if they are left handed or right handed. Two hundred-eighty Americans were randomly chosen for the study. Subjects reported whether they were right handed or left handed and then were given a psychological exam to determine if they were introverted or extroverted. The data is presented in the following contingency table:

Table 1: Observed data

	Introvert	Extrovert	Total
Left Handed	20	15	35
Right Handed	116	129	245
Total	136	144	280

Perform a Chi-Square test of independence to see if the categorical variables are related, and find the expected cell counts assuming the variables are independent.

Table 2: Expected cell counts under assumption of independence

	Introvert	Extrovert	Total
Left Handed	17	18	35
Right Handed	119	126	245
Total	136	144	280

- List hypotheses:
  - H_0: Being left-handed vs right-handed is independent of being introverted vs extroverted
  - H_A: Being left-handed vs right-handed is not independent of being introverted vs extroverted
- Check conditions:
  - Count data - data are counts categorized on 2 categorical variables
  - Randomization condition - observed cell counts are based on random sample
  - Large enough - cell counts ≥ 5 for each cell
- Name the test: Chi-Square Test of Independence with 1 degree of freedom

$$(2 \text{ rows } - 1) * (2 \text{ columns } - 1) = 1$$

- Test statistic from calculator: $\chi^2 = 1.17647$
- p-value from calculator: $p = 0.278$
- This p-value is larger than any reasonable α so we fail to reject H_0
- Conclusion: Based on these sample results, there is insufficient evidence to reject the hypothesis that being left-handed vs right-handed is independent of whether a person is introverted or extroverted.

Problem 4

A civil engineer randomly chooses 90 days from the previous two years to look at and see how many traffic accidents there were per day in Iowa City. They observe the following data:

Traffic accidents on a given day	Days observed
0	52
1	22
2	9
3	5
4	2

The civil engineer wants to see if they can model the rate of traffic accidents as a Poisson distribution. Using the observed data, they estimate an average rate of 0.7 accidents per day and the following probability distribution:

P(X=0)	P(X=1)	P(X=2)	P(X=3)	P(X=4)
0.497	0.348	0.122	0.028	0.005

Find the expected number of days out of 90 days that one would observe these number of traffic accidents under the assumption of the civil engineer's proposed model and perform a Chi-Square goodness of fit test.

Traffic accidents on a given day	Expected days observed
0	44.73
1	31.32
2	10.98
3	2.52
4	0.45

- List hypotheses:
  - H_0: The number of traffic accidents per day can be modeled as proposed with the Poisson distribution
  - H_A: The number of traffic accidents per day is not approximated by a Poisson distribution
- Check conditions:
  - Count data - data are counts of a categorical variable
  - Randomization condition - observed cell counts are based on random sample
  - Large enough - cell counts ≥ 5 for each cell
- Name the test: Chi-Square Goodness of Fit Test with 4 degrees of freedom

$$5 \text{ categories } - 1 = 4$$

- Test statistic from calculator: $\chi^2 = 12.0916$
- p-value from calculator: $p = 0.017$
- This p-value is smaller than an α level of 0.05 so we reject H_0
- Conclusion: Based on these sample results, there is sufficient evidence to reject the hypothesis that the number of car accidents follows the civil engineer's proposed Poisson distribution.

Problem 5

We want to see if we can linearly model the relationship between miles-per-gallon and vehicle weight. We sample 30 random vehicles and obtain their weights and respective miles per gallon. We get a standard deviation of mpg observations of 5.62 and a standard deviation of weight observations of 1.03. We find a correlation between the variables of -0.862. The average mpg for our sample is 20.61, and the average weight is 3.97 tons.

Determine the linear regression model for our data.

$$\hat{b} = r\frac{s_y}{s_x} = (-0.862)\frac{5.62}{1.03} = -4.70334$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 20.61 - (-4.70334)(3.97) = 39.28226$$

$$\hat{mpg} = 39.28226 - (4.70334)weight$$

What is the coefficient of determination and what does it mean?

$R^2 = r^2 = (-0.862)^2 = 0.743044$

The coefficient of determination is the fraction of the data's variation that is accounted for by the model. In this case, our R^2 means that in this model, about 74% of the variation in mpg is explained by the least squares regression of mpg on vehicle weight.

We observe a vehicle that weighs 2.2 tons and gets about 34 miles per gallon. What is the residual for this observation based on our model?

Based on our model we expect a vehicle weighing 2.2 tons to get about:

$$\hat{mpg} = 39.28226 - (4.70334)(2.2) = 28.93491$$

The actual observed mpg is 34. The residual for this observation is:

$$\text{actual - predicted } = 34 - 28.93491 = 5.06509$$

Interpret the slope of our linear regression model.

The model predicts that for every 1 ton increase in vehicle weight the miles-per-gallon decreases by about 4.7 mpg