2020 Summer STAT 22000 Practice Final Exam Solutions

1. [Roulette] [16 points]

The game of roulette involves spinning a wheel with 38 slots: 18 red, 18 black, and 2 green. A ball is spun onto the wheel and will eventually land in a slot, where each slot has an equal chance of capturing the ball. One popular bet is that it will stop on a red slot; such a bet has an 18/38 chance of winning.

(a) (2pts) Suppose a gambler bets on red in 6 different spins. What is the probability that he wins the first 3 spins but loses the next 3 spins?

Answer:

$$\left(\frac{18}{38}\right)^3\left(\frac{20}{38}\right)^3 \approx 0.0155$$

(b) (2pts) Let X be the number of times the gambler wins in the 6 spins. Explain why X has a binomial distribution.

Answer:

- A Bernoulli trial is whether the gambler wins in a spin, only two possible outcomes: win or loss
- There are $n = 6$ trials in total (he bets on 6 different spins), which is fixed in advance
- The probability of getting a success is $\mathrm{P}(\text{wins in a spin}) = p = 18/38$ is identical for all trials.
- The trials (spins) are independent

So $X \sim Bin(n = 6, p = 18/38)$.

(c) (3pts) Suppose a gambler bets on red in 6 different spins. What is the probability that the gambler wins exactly 3 of the 6 spins, i.e., $P(X = 3)$?

Answer: The number of times X the gambler wings the binomial distribution $Bin(n = 6, p = 18/38)$. The chance that $X = 3$ is

$$\begin{aligned} P(X=3) &= \binom{6}{3}(18/38)^3 \cdot (20/38)^3 \\ &= 20 \cdot (18/38)^3 \cdot (20/38)^2 \approx \boxed{0.3099} \end{aligned}$$

Here $\binom{6}{3}$ equals 20 because

$$\binom{6}{3} = \frac{6!}{3!\,3!} = \frac{6\cdot5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(3\cdot2\cdot1)} = \frac{6\cdot5\cdot4}{3\cdot2\cdot1} = 20$$

For part (d-e) below, suppose the gambler bets on red in 50 different spins.

(d) (4pts) How many times do you expect the gambler to win in the 50 spins? And with what standard deviation?

Answer: Let Y be the number of time he wins in the 40 spins. Now $Y \sim Bin(n = 50, p = 18/38)$.

So the expected value is $np = 50 \times (18/38) = 360/19 \approx \boxed{23.684}$. .... (2pts)

The SD is $\sqrt{np(1-p)} = \sqrt{50(18/38)(20/38)} \approx \boxed{3.531}$.. (2pts)

(e) (5pts) Find an approximate value for the probability that the gambler wins at least 26 times in the 50 spins. Please calculate using normal approximation to Binomial WITH continuity correction. If you don't know how to do the continuity correction, you can also use normal approximation WITHOUT continuity correction and get 4 points if it is done correctly.

Answer: By normal approximation of binomial, we know

$$Y \sim Bin(n = 50, p = 18/38) \approx N(\mu = np = 23.684, \sigma = \sqrt{np(1-p)} = 3.531)$$

With continuity correction, the end point 26 is changed to 25.5 since $\{Y \geq 26\}$ and $\{Y \geq 25.5\}$ contains the same set of integers.

$$P(Y \geq 26) = P(Y \geq 25.5) = P\left(Z > \frac{25.5 - 23.684}{3.531}\right) \approx P(Z \geq 0.514) = 1 - P(Z < 0.514) \approx 1 - 0.696 = \boxed{0.304}.$$

(Values between $1 - P(Z < 0.51) = 0.3050$ and $1 - P(Z < 0.52) = 0.3015$ are acceptable.)

Without continuity correction, the approximated probability would be

$$P(X \geq 26) = P(Z > \frac{26 - 23.684}{3.531}) \approx P(Z \geq 0.656) = 1 - P(Z < 0.656) \approx 1 - 0.744 = \boxed{0.256}.$$

(Values between $1 - P(Z < 0.65) = 0.2578$ and $1 - P(Z < 0.66) = 0.2546$ are acceptable.)

Remark:

- A few students rounded the expected value $np = 50(18/38) = 23.684$ to an whole number 24 approximates Bin$(n = 40, p = 18/38)$ with $N(\mu = 24, \sigma = 3.531)$. This way will give a z score 0.425 and the answer will be around 0.3354. For normal approximation of binomial, the normal distribution is centered at np, not the rounded value of np. With such rounding, the normal approximation won't be as accurate. I took 0.5 point off for such answer.
- The exact probability based on binomial formula is $\sum_{k=26}^{50} \binom{50}{k} (18/38)^k (20/38)^{50-k} \approx 0.3031$, which is closer to the value 0.304 calculated with the correction than the one 0.256 without the correction.
- 5pts = 1pt for changing the endpoints to 25.5 + 2pts for the z-score + 2pts for the prob.

2. [Housing Prices in CA] [10 points]

A housing survey was conducted to determine the price of a typical home in large town in CA. The distribution of housing prices has a population mean of \$1.3 million with a population standard deviation of \$0.3 million. There were no homes listed below \$0.6 million but a few are above \$3 million. There are more than 1000 homes in the town.

(a) (3pts) Can we calculation (an approximate) probability that a randomly chosen home in this town costs more than \$1.33 million using the normal distribution? If yes, please (i) explain why you can and (ii) calculate the probability. If no, explain why not.

Answer: We can not since the population distribution is clearly right-skewed, (length of lower tail is at most $1.3 - 0.6 = 0.7$; length of upper tail is at least $3 - 1.3 = 1.7$). As the population distribution is not normal, we cannot use the normal curve to find the probability about the distribution of a single observation.

(b) (4pts) Can we calculation an approximate probability that the mean price of 100 randomly chosen homes in the town is more than \$1.33 million using the normal distribution? If yes, please (i) explain why you can and (ii) calculate the probability. If no, explain why not.

Answer: Despite of the non-normal population distribution, the Central Limit Theorem guarantees the sampling distribution of the mean price of 100 randomly chosen houses to be approximately $N(\mu = 1.3, \sigma = 0.3/\sqrt{100} = 0.03)$ as the sample size 100 is fairly large. So

$$P(\overline{X} > 1.33) = P\left(Z > \frac{1.33 - 1.3}{0.03}\right) = P(Z > 1) = 1 - 0.8413 = 0.1587.$$

(c) (3pts) True or False and explain: the histogram of the prices of 100 randomly chosen homes in this town is roughly normal.

Answer: False. The histogram of the sample will be close to the distribution of the population, which is right-skewed.

Remark: It is the sampling distribution of the sample mean that will have an approximately normal distribution.

3. [How often Read a Newspaper] [9 points]

In a survey of a random sample of 50 students in a certain University, it is found that on average the subjects in the sample read a newspaper 4.1 times in the week prior to the survey, with a standard deviation of 3.0 times.

(a) [4 points] Find a 95% confidence interval for the mean number of times students in this University read a newspaper the week prior to the survey.

Answer: The degrees of freedom is $50 - 1 = 49$. The critical value t^* for a 95% CI is 2.01.

$$\begin{aligned}\overline{x} \pm t^* \frac{s}{\sqrt{n}} = 4.1 \pm 2.01 \times \frac{3.0}{\sqrt{50}} &= 4.1 \pm 0.85 \\ &= 3.25 \text{ to } 4.95.\end{aligned}$$

(b) [2 points] About the number of times students read a newspaper in the week prior to the survey, which of the following statement is true?

(i) The population distribution is normal so it's legitimate to construct the confidence interval in (a).

(ii) The population distribution is not normal since this variable only take integer values. Hence we cannot construct a confidence interval based on a t distribution.

(iii) The population distribution is not normal, but we can construct the confidence interval in (a) as long as the sample contains no outlier and is not severely skewed, and the observations are independent.

Answer: (iii)

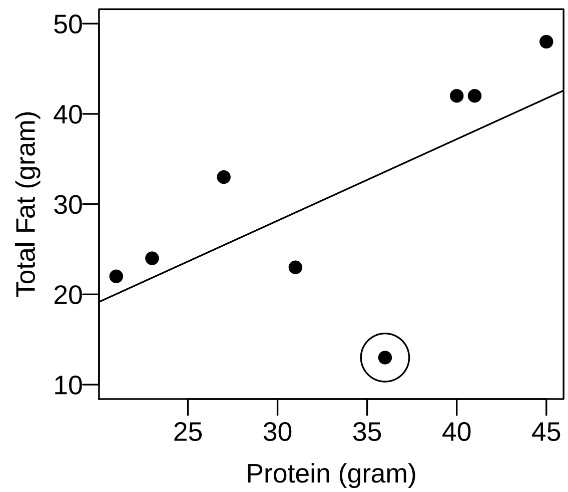
(c) [3 points] Explain to someone who knows no statistics what a 95% confidence interval in part (a) means. More specifically, what is the thing that has a 95% probability to happen?

Answer: If one repeats the same process — taking a SRS of size 50 from the University, and then compute an interval: sample mean $\pm\, 2.01 \times$ (sample SD)$/\sqrt{50}$, a huge number of times, then for 95% of the time the interval computed will cover the true population mean.

4. [Fast Food] [9 points]

Data were obtained from the A&W Web site for the total fat in grams and the protein content in grams for various items on their menu. Some summary statistics and a scatter plot are also provided:

Item	Total fat (grams)	Protein (grams)
Kid' Cheeseburger	24	23
Kid' Hamburger	22	21
Original Bacon Cheeseburger	33	27
Original Bacon Double Cheeseburger	48	45
Original Double Cheeseburger	42	40
Crispy Chicken Sandwich	23	31
Grilled Chicken Sandwich	13	36
Papa Burger	42	41
Mean	30.875	33.000
SD	12.264	8.864
Correlation	$r = 0.653$	



(a) [4 points] Find the equation of the least-squares regression line for predicting total fat from protein.

Answer: Predicted total fat $= 1.06 + 0.903$ (Protein.)

$y =$ total fat, $x =$ protein.

$$\text{slope} = r\frac{s_y}{s_x} = (0.653)\frac{12.264}{8.864} \approx 0.90347 \qquad \text{(2 pts)}$$
$$\text{intercept} = \overline{y} - (\text{slope})\overline{x} = 30.875 - (0.9035)(33) \approx 1.06 \qquad \text{(2 pts)}$$

[Grading]

- 2pts for the slope, 2pts for the intercept
- If the student obtained a wrong value for the slope, and use it to compute the intercept in the right way, please give the 2pts for the intercept even if the value for the intercept is wrong
- Due to difference in rounding, the intercept can range from 1.04 to 1.08.

(b) [5 points] For each of the following statements about the circled data point on the scatterplot, determine whether it is TRUE or FALSE. No explanation is required.

(i) $\boxed{\text{TRUE}}$ or FALSE: The circled data point on the scatterplot is for the Grilled Chicken Sandwich.

(ii) $\boxed{\text{TRUE}}$ or FALSE: The residual associated with this data point will have a negative value.

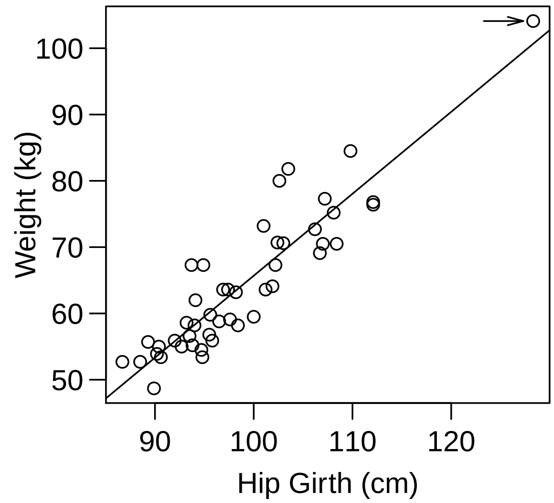
(iii) $\boxed{\text{TRUE}}$ or FALSE: This point would likely be considered an outlier.

(iv) TRUE or $\boxed{\text{FALSE}}$: This point has a high leverage.

(v) $\boxed{\text{TRUE}}$ or FALSE: Without this point, the correlation between total fat and protein would be higher.

5. [Hip Girth & Weight] [14 points]

The scatterplot on the right shows the weights (in kg) and hip girths (in cm) of 46 physically active women age 35-44. The following regression output is for predicting women's body weight from their hip girth. Part of the output is blurred.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  XXXXXXXX    8.46196  XXXXXX 1.94e-08
hip.girth    XXXXXXXX    0.08523  XXXXXX  < 2e-16
```

Here is a summary of the data:

	Body weight (kg)	Hip girth (cm)
Mean	64.41	98.97
SD	10.80	7.95

Correlation $r \approx 0.91$

(a) (4pts) Write down the equation of the least square regression line for predicting women's body weights (in kg) from their hip girths (in cm).

Answer: y = weight, x = hip girth.

$$\text{slope} = r\frac{s_y}{s_x} \approx 0.91\frac{10.80}{7.95} \approx 1.236 \qquad \text{(2 pts)}$$
$$\text{intercept} = \overline{y} - (\text{slope})\overline{x} = 64.41 - (1.236)(98.97) \approx -57.9 \qquad \text{(2 pts)}$$

The equation of the least square regression line is

$$\text{Predicted weight in kg } = -57.9 + 1.236 \text{ (hip girth in cm)}.$$

[Grading]

- 2pts for the slope, 2pts for the intercept
- If the student obtained a wrong value for the slope, and use it to compute the intercept in the right way, please give the 2pts for the intercept even if the value for the intercept is wrong

(b) (4pts) Calculate a 95% confidence interval for the slope of the regression line for predicting women's body weights (in kg) from their hip girths (in cm).

Answer: With $n - 2 = 46 - 2 = 44$ degrees of freedom, the critical value is $t^* = 2.02$. The 95% CI for the slope is

$$\text{estimate} \pm t^*\,\text{SE} = 1.236 \pm 2.02 \times 0.08523 = 1.236 \pm 0.1721646 \approx (1.06, 1.41)$$

- 1 pt for the estimated slope 1.236
- 2 pts for the SE 0.08523
- 1 pt for df = 44 and $t^* = 2.02$

(c) (2pts) If body weight is measured in pounds and hip girth is measured in inches, what will be the correlation between body weight and hip girth? (1 pound = 0.454 kg, 1 inch = 2.54 cm.)

Answer: r is not affected because correlation is not affected by a change of unit.

(d) (4pts) Predict the hip girth of a woman that weighs 60 kg using linear regression.

Answer: y = hip girth, x = weight

$$\text{slope} = r\frac{s_y}{s_x} \approx 0.91\frac{7.95}{10.80} \approx 0.670$$
$$\text{intercept} = \overline{y} - (\text{slope})\overline{x} = 98.97 - (0.670)(64.41) \approx 55.8$$

$$\text{Predicted hip girth in cm } = 55.8 + 0.67 \text{ (weight in kg)}.$$

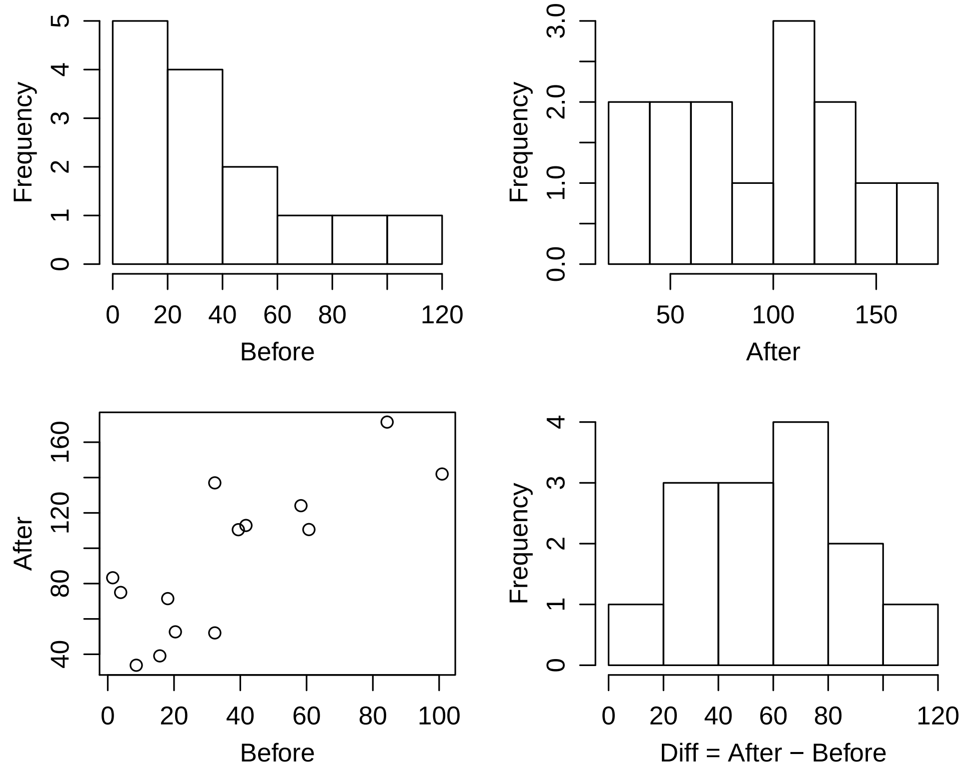
Predicted hip girth of a women that weighs 60 kg is $\approx 55.8 + 0.670 \times 60 = \boxed{96 \text{ cm}}$.

6. [Fortified Orange Juice] [9 points]

V. Tangpricha et al. conducted a study to determine whether fortifying orange juice with vitamin D would increase serum 25-hydroxyvitamin D [25(OH)D] concentration in the blood[1]. In medicine, the blood concentration of 25(OH)D is considered the best indicator of how much vitamin D is in the body. Most experts consider a serum 25(OH)D level of less than 30 nmo/L as indicative of vitamin D deficiency.

In the study, 14 subjects drank 240 mL per day of orange juice fortified with 1000 IU of vitamin D. Concentration levels were recorded at the beginning of the experiment and again at the end of 12 weeks. The before and after serum 25(OH)D concentrations in the blood, in nanomoles per liter (nmo/L), of the 14 subjects and their differences. The plots below are the histograms of the before and after serum 25(OH)D level and their scatter plot, as well as the histogram of their difference.

Subject	Before	After	Difference =After−Before
1	8.6	33.8	25.2
2	3.9	75.0	71.1
3	32.3	137.0	104.7
4	1.5	83.3	81.8
5	60.7	110.6	49.9
6	18.1	71.5	53.4
7	20.4	52.7	32.3
8	100.9	142.0	41.1
9	39.4	110.5	71.1
10	84.3	171.4	87.1
11	15.7	39.1	23.4
12	32.3	52.1	19.8
13	58.3	124.1	65.8
14	41.7	112.9	71.2
Mean	37.01	94.00	56.99
SD	29.94	42.10	26.20



(a) [5 points] Construct a 99% confidence interval for the mean increase of the serum 25(OH)D concentration after 12 weeks of drinking fortified orange juice.

Answer: With $df = 14 - 1 = 13$, the multiplier t^* is 3.01

$$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}} = 56.99 \pm 3.01 \times \frac{26.2}{\sqrt{14}} = 56.99 \pm 21.08 = (35.91, 78.07) \text{ (nmo/L)}$$

[1] V. Tangpricha et al. (2003) Fortification of Orange Juice with Vitamin D: A Novel Approach for Enhancing Vitamin D Nutritional Health. *American Journal of Clinical Nutrition*, Vol. 77, pp. 1478-1483

(b) [2 points] Which of the following statement is true? No explanation is required.

(i) The normality assumption for constructing the confidence interval in part (a) is violated because the distribution of the "Before" blood 25(OH)D level appears to be skewed.

(ii) The independence assumption for constructing the confidence interval in part (a) is violated because from the scatter plot, the "Before" and "After" blood 25(OH)D level appear to be highly correlated.

(iii) Both the above are true.

(iv) $\boxed{\text{All the above are false.}}$

Answer: For paired data, we just only required the differences to be independent and normally distributed. We expect the "Before" and "After" values within a pair to be dependent, and neither the "Before" nor "After" values are required to be normally distributed.

(c) [2 points] From the scatter plot, the correlation coefficient between the "Before" and "After" blood 25(OH)D level is closest to which of the following?

(i) -0.7 (ii) -0.2 (iii) 0.3 $\boxed{\text{(v) } 0.8}$

No explanation is required.

7. [T or F & Multiple Choice] [14 points]

(a) (6pts) Determine whether the following statements are true or false. No explanation is required.

(i) $\boxed{\text{TRUE}}$ or FALSE: The significance level of a test is the probability of making a Type 1 error when the null hypothesis is true.

(ii) TRUE or $\boxed{\text{FALSE}}$: Increasing the significance level will increase the probability of making a Type 2 error. $\Rightarrow$ Increased significance level $\Rightarrow$ more likely to reject H_0 $\Rightarrow$ more likely to make a Type 1 error

(iii) TRUE or $\boxed{\text{FALSE}}$: A Type 2 error is made if a correct null hypothesis is rejected. $\Rightarrow$ Type 1 error

(iv) TRUE or $\boxed{\text{FALSE}}$: We usually try to avoid making a Type 2 error more than to a Type 1 error $\Rightarrow$ the other way around

(v) $\boxed{\text{TRUE}}$ or FALSE: The t-distribution has heavier tails than the normal distribution.

(vi) TRUE or $\boxed{\text{FALSE}}$: Confidence intervals based on a t-distribution will be shorter than confidence intervals based on the Normal distribution. $\Rightarrow$ t interval is wider

(b) (2pts) A certain brand of cigarettes advertises that the mean nicotine content of their cigarettes is $\mu = 1.5$ milligrams (mg). To test this, a random sample of 100 cigarettes of this brand were examined and the p-value for testing $H_0 : \mu = 1.5$ mg versus H_a: $\mu \neq 1.5$ mg was found to be $= 3.2\%$. Determine whether the following statements are true or false. No explanation is required.

(i) TRUE or $\boxed{\text{FALSE}}$: A p-value of 3.2% means the probability that H_0 is true is 3.2%. So the evidence supporting H_0 is weak. $\Rightarrow$ p-value is not the probability that H_0 is true.

(ii) TRUE or $\boxed{\text{FALSE}}$: The value 1.5 mg is in the 95% confidence interval for the actual mean nicotine content of cigarettes of this brand. $\Rightarrow$ As p-value $= 3.2\% < 5\%$, H_0: $\mu = 1.5$ mg is rejected. 95% CI for μ does not contain 1.5 mg.

(c) (4pts) A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false. No explanation is required.

(i) TRUE or $\boxed{\text{FALSE}}$: About 95% patients at this hospital's emergency wait between 128 and 147 minutes

(ii) $\boxed{\text{TRUE}}$ or FALSE: The margin of error is 9.5 minutes and the sample mean is 137.5 minutes

(iii) TRUE or $\boxed{\text{FALSE}}$: Doubling the sample size would cut the margin of error by half.

(iv) $\boxed{\text{TRUE}}$ or FALSE: If we used a different confidence level, the interval would be symmetric about the sample mean

(d) (2pts) The World Bank reports that 1.7% of the US population lives on less than \$2 per day. A policy maker claims that this number is misleading because of variation from state to state and rural to urban. To investigate this, she takes a random sample of 100 households in Atlanta to compare with the national average and finds that 2.1% of the Atlanta population live on less than \$2/day. Select the null and alternative hypothesis to test whether Atlanta differs significantly from the national percentage.

(i) H_0: $p = 2.1$, H_a: $p \neq 2.1$

(ii) H_0: $\mu = \$2$ per day, H_a: $\mu > \$2$ per day

(iii) H_0: $p = 0.017$, H_a: $p = 0.021$

(iv) H_0: $p = 0.021$, H_a: $p \neq 0.021$

$\boxed{\text{(v) } H_0\text{: } p = 0.017,\ H_a\text{: } p \neq 0.017}$

No explanation is required.

(e) (2pts) The alumni association for U of Chicago has gathered a large dataset on graduating seniors. Some of the variables in the data set are gender, major, GPA, and starting salary. They are interested in looking at relationships among these variables. For which of the following pairs of variables would a two sample t-test be appropriate to examine whether there is a relationship between the two variables?

(i) gender and major

$\boxed{\text{(ii) GPA and gender}}$

(iii) major and salary

(iv) GPA and salary

(v) none of the above

No explanation is required. ⇒ One can use two-sample t-test two compare the mean GPAs of men and women as GPA is numerical and gender has only two categories. For (i), as both gender and major are categorical and major has more than 2 categories, we would have to use a Chi-square test in Section 6.4 which we didn't have time to introduce. For (iii), as salary is numerical and major has more than 2 categories, one has to use an ANOVA test in Section 5.5, which we didn't cover. For (iv), as both GPA and salary are numerical, we can use regression methods to investigate their relations.

8. [Offshore drilling] [20 points]

A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent is a college graduate.

	College Grad	
	Yes	No
Support	154	132
Oppose	180	126
Do not know	104	131
Total	438	389

(a) (4 points) What proportion of respondents **opposed** offshore drilling among those with a college degree? What is the corresponding proportion among those without a college degree?

Answer: Among those with a college degree, the proportion that opposed offshore drilling is $\widehat{p}_c = \frac{180}{438} \approx 0.411$.

Among those without a college degree, the proportion is $\widehat{p}_n = \frac{126}{389} \approx 0.324$.

(b) (8 points) Test whether or not those with a college degree had a different tendency to oppose offshore drilling from those without a college degree. Please state the null and alternative hypotheses, give an appropriate test statistic, report the p-value, and state the conclusion in the context using a 0.05 significance level.

Answer: Let $p_{c,o}$ be the proportion of respondents with a college degree that **opposed** offshore drilling, and $p_{n,o}$ be the corresponding proportion for those without a college degree. The hypotheses are

$$H_0 : p_{c,o} = p_{n,o} \quad \text{v.s.} \quad H_a : p_{c,o} \neq p_{n,o}.$$

Under H_0, the estimate of the common p is the pooled sample proportion,

$$\widehat{p} = \frac{180+126}{438+389} = \frac{306}{827} \approx 0.370$$

The standard error of the difference under H_0 is then

$$\mathrm{SE} = \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_c}+\frac{1}{n_n}\right)} = \sqrt{0.370 \times (1-0.370)\left(\frac{1}{438}+\frac{1}{389}\right)} \approx 0.03364$$

The test statistic is

$$z = \frac{\widehat{p}_m - \widehat{p}_f}{\mathrm{SE}} \approx \frac{0.411-0.324}{0.03364} = \frac{0.087}{0.03364} \approx 2.59$$

The two-sided P-value is $2P(Z > 2.59) = 2(1-0.9952) = 0.0096$.

Conclusion: As the P-value 0.0096 is below the 0.05 significance level, we can reject H_0 and conclude that those with a college degree had a significantly higher tendency to oppose offshore drilling from those without a college degree.

- 1 pt for the hypotheses
- 4pts for the test statistic. Deduct 2pts if the z-statistic is computed as

$$z = \frac{0.411-0.324}{\sqrt{\frac{0.411(1-0.411)}{438}+\frac{0.324(1-0.324)}{389}}} \approx \frac{0.087}{0.0334} \approx 2.60$$

- 2 pts for the P-value
- 1 pt for the conclusion

(c) (5 points) Construct a 95% confidence level for the difference $p_{c,s} - p_{n,s}$ where $p_{c,s}$ the proportion **supported** offshore drilling among those with a college degree, and $p_{n,s}$ is the corresponding proportion among those without a college degree.

Answer: Let $p_{c,s}$ and $p_{n,s}$ to be the proportion of college graduates and non-college graduates supporting offshore drilling. Their estimates are

$$\widehat{p}_c = \frac{154}{438} \approx 0.3516, \quad \widehat{p}_n = \frac{132}{389} \approx 0.3393$$

The 95% CI for $p_{c,s} - p_{n,s}$ is

$$\begin{aligned}
&(\widehat{p}_c - \widehat{p}_n) \pm 1.96\sqrt{\frac{\widehat{p}_c(1-\widehat{p}_c)}{n_c} + \frac{\widehat{p}_n(1-\widehat{p}_n)}{n_n}} \\
=&\, 0.3516 - 0.3393 \pm 1.96\sqrt{\frac{0.3516(1-0.3516)}{438} + \frac{0.3393(1-0.3393)}{389}} \\
=&\, 0.0123 \pm 1.96 \times 0.0331 \\
=&\, 0.0123 \pm 0.0649 = (-0.0526, 0.0772)
\end{aligned}$$

(d) (3 points) Does the confidence interval in part (c) agree or contradict with the conclusion of the test in part (b)? Is it surprising or possible or is there anything wrong? Explain.

Answer: The CI in part (c) contains 0, meaning there is no significant difference in the percentages of college graduates and non-college graduates supported offshore drilling. However, this does not contradict the conclusion of the test in (b) because the test in (b) we test if college graduates had a different tendency to **oppose** offshore drilling from those without a college degree. Those who didn't oppose may not necessarily support offshore drilling as there is a third answer "Do Not Know." There were about the same percentage of people **supporting** offshore drilling in the two populations, but there is a higher percentage ($180/438 \approx 41\%$) of college graduates that **oppose** compare to $126/389 \approx 32\%$ among non-college graduates.