DS 401: Fall 2021: Final exam - Part I solutions

(1) [6 pts.] A low-calorie snack company claims that, on average, there are 95 calories in their small bag of crackers. A consumer watchdog group is concerned that the company may be lying and that the population average is not actually 95 calories. An employee of the group samples 45 bags of chips and computes a 95% confidence interval of $(92, 118)$. When the employee reports her results, the head of the group says that he wanted to do inference using a 99% confidence interval.
(a) What conclusion/inference can we expect from a 99% confidence interval in the context of the watchdog group's concern? Why?

Solution: Since the original 95% confidence interval contains 95 calories as a possible value for the true mean, a 99% confidence interval will also lead to the same conclusion. This is because the 99% CI will be wider. The conclusion is that the sample does not provide strong enough evidence to refute the snack company's claim that the true mean is 95 calories.

(b) They decide to do a hypothesis test to check whether the true mean calories are higher than what the company claims ($H_0 : \mu = 95$ calories, $H_A : \mu > 95$). What significance level α would match the 95% confidence interval?

Solution: $\alpha = 0.025$ would match the 95% confidence interval, since our hypothesis test is 1-tailed and there is 2.5% in each tail of a 95% CI.

Grade: (a)=4 point, (b)=2 pt.
For (a): 2 pt = correct answer; 2 pt = reason.
For (b): Correct answer is sufficient.

(2) [6 pts.] The alumni association of a college has gathered a large dataset on graduating seniors who have an accepted job offer. Some of the variables in the data set are gender (F, M, nonbinary), major (STEM, other), GPA, age, and starting salary. They are interested in exploring a linear model to predict starting salary from the other variables. Write the equation of the regression model and clearly describe each variable in the model. Assume GPA, age, and starting salary are quantitative, and the other variables contain the indicated categories.

Solution: I will use the following variables in my model:

$\widehat{\text{salary}}$ = predicted starting salary
gender:F = 1 if the person's gender is F, and 0 otherwise
gender:M = 1 if the person's gender is M, and 0 otherwise
(if gender:F = 0 and gender:M = 0, then by default gender is "nonbinary")
major = 1 if the person is a STEM major, and 0 otherwise
GPA = the numerical value of the person's GPA
age = the numerical value of the person's age

The regression model would look like

$$\widehat{\text{salary}} = b_0 + b_1 \text{ gender:F} + b_2 \text{ gender:M} + b_3 \text{ major} + b_4 \text{ GPA} + b_5 \text{ age}$$

where $b_0, b_1, b_2, b_3, b_4, b_5$ are constant parameters to be determined.

Grade: 1pt=any correct MLR model containing LHS and RHS terms;
1+1 pt = correct treatment of gender + major in final model;
0.5 + 0.5 pt = correct treatment of GPA + age term in final model
1 pt = correct LHS; 1 pt = variables used are clear.

(3) [3 pts.] State the null and the alternative hypotheses and explain under what situation a Type II error would be committed. Be sure to clearly define any parameters used.

(a) The average age of customers who shop at a clothing store location in a city is 34 years. The company wants to know if customers who shop on their website are younger on average.

Solution: Let μ = true mean age in years of customers who shop on the company's website.
Null hypothesis $H_0 : \mu = 34$
Alt hypothesis $H_A : \mu < 34$

A Type II error would be committed if we conclude the true mean age of the website customers is not younger than 34 years, when in fact it is.

(b) Are a majority of students who attend Earlham College politically liberal? We want to test a hypothesis to find out.

Solution: Let p = true proportion of Earlham students who are politically liberal.

Null hypothesis $H_0 : p = 0.5$
Alt hypothesis $H_A : p > 0.5$

A Type II error would be committed if we conclude a majority of Earlham students are not politically liberal, when in fact they are.

Grade: 3 points each. For each: 1 pt = correct hypotheses; 1 pt = clarify parameter in hypotheses; 1 pt = explain consequence of Type II error.

(4) [6 pts.] Most regression methods – including simple linear, multilinear, logistic, and curvilinear – are designed based on a least squares principle. That means, the constant parameters in the model are computed to minimize the sum of squares of the errors (or residuals), as we did at various points in class. In this exercise we want to use least squares to fit a regression model of the form

$$y = mx + b + \frac{n}{x}$$

with x = predictor variable, y = response variable, and m, n, b are constant parameters.

(a) Use the x, y data given in the table and find the function we want to minimize. This function should not contain any unknowns, except m, n and b.

(b) Write a couple of sentences explaining how to find the numerical values of m, n and b.

x	y
1	10
2	3
4	6

Solution: Assume regression model of the form: $y = mx + b + \frac{n}{x}$

The table below lists the quantities I will need to construct the least squares function:

x_i	y_i	$\hat{y}_i$	$e_i = (y_i - \hat{y}_i)$
1	10	$m + b + n$	$10 - m - b - n$
2	3	$2m + b + 0.5n$	$3 - 2m - b - 0.5n$
4	6	$4m + b + 0.25n$	$6 - 4m - b - 0.25n$

The function to be minimized is:

$$f(m, n, b) = (10 - m - b - n)^2 + (3 - 2m - b - 0.5n)^2 + (6 - 4m - b - 0.25n)^2$$

It is okay to leave this function un-simplified for the next step in least squares.

(b) The numerical values of m, n and b are found by minimizing $f(m, n, b)$ using calculus or other numerical minimization methods. If using calculus, we get 3 algebraic equations in the 3 unknowns by setting $\frac{\partial f}{\partial m} = 0$, $\frac{\partial f}{\partial n} = 0$ and $\frac{\partial f}{\partial b} = 0$.

Grade: (a)=4 points, (b)=2 points.
For (a): 1pt = attempt to find $e_i = y_i - \hat{y}_i$ in a valid way;
1.5pt = get all 3 e_i's correctly; 1.5pt = square and add them.
For (b): 1pt = mention that we must minimize f;
1pt = mention m, n, b are to be found using calculus or other method.

(5) [6 pts.] In 2015 the average credit card debt per US household was about \$15,355. The distribution of household debt was right skewed with mean \$15,355 and standard deviation \$10,000.

(a) If it is possible, find the probability that a randomly sampled household has more than \$16,000 in credit card debt. If it is not possible, explain why.

Solution: It is not possible to find this probability for one household sampled from this population, since the distribution shape is unknown and skewed.

(b) If possible, find the probability that a random sample of 1,000 households has more than \$16,000 of average credit card debt. If it is not possible, explain why.

Solution: Since the sample is random and large, the central limit theorem applies. Thus, the sampling distribution will follow the model

$$N(\mu, \frac{\sigma}{\sqrt{n}}) = N(15355, \frac{10000}{\sqrt{1000}}) = N(15355, 316.23)$$

The z-score $= \dfrac{16000 - 15355}{316.23} = 2.04.$ From z-table, $\Pr(z > 2.04)$=0.0207

The probability that the mean credit card debt $> \$16,000$ is 0.0207

Grade: (a)=2 points, (b)=4 points.
For (a): 1+1 pt = correct answer + correct reason.
For (b): 2 pt = correct shape, mean and SD of sampling distribution model;
1+1 pt = correct z score + correct probability lookup.

(6) [6 pts.] Give brief answers to each of the following, as instructed;

(a) In a recent classwork exercise, we looked at how to construct polynomial regression models, and how to implement them in R. In one of the examples, the response variable was the odor in a chemical process, and there were three predictor variables – temperature, ratio, and height. A multilinear model for these data turned out to give a poor fit. Results were much better using a quadratic model for all 3 predictor variables. Write the general form of the equation of this regression model and clarify what each variable represents.

Solution: A quadratic regression model would look like

$$\begin{aligned}\widehat{\text{odor}} = b_0 &+ b_1(\text{temperature}) + b_2(\text{temperature})^2 \\ &+ b_3(\text{ratio}) + b_4(\text{ratio})^2 + b_5(\text{height}) + b_6(\text{height})^2\end{aligned}$$

where the variables correspond to those given in the problem, together with the squares of their numerical values, and $b_0, \ldots, b_6$ are constant parameters.

(b) A researcher tests whether the mean cholesterol level among those who frequently eat frozen pizza exceeds the value considered to be a health risk. She gets a P-value of 0.09. Explain the meaning of this P-value in context of this study.

Solution: The P-value means: If the mean cholesterol level among those who eat frozen pizza does not exceed the health risk limit, then there is a 0.09 chance of observing the mean the researchers found in their sample (or even more extreme results).

(c) Which, if any, of the following scenarios would result in the same p-value?

i. H_0: $p = 0.3$ vs H_a: $p > 0.3$, $z = 1.65$
ii. H_0: $p = 0.25$ vs H_a: $p \neq 0.25$, $z = 1.65$
iii. H_0: $p = 0.5$ vs H_a: $p \neq 0.5$, $z = 2.65$
iv. H_0: $p = 0.75$ vs H_a: $p < 0.75$, $z = -1.65$
v. H_0: $p = 0.5$ vs H_a: $p > 0.5$, $z = 2.65$

Solution:

Only i. and iv. will have the same p-value.

Grade: 6 points. Distribution: (a)=3 pt, (b)=2 pt, (c)=1 pt.
For (a): 1 pt = correct linear terms; 1.5 pt = correct quadratic terms; 0.5 pt = constant term.
For (b): 0.5 pt = answer reflects understanding that P-value is a probability; 1.5 pt = rest of it is correct.
For (c): correct answer is sufficient.

DS 401: Fall 2021: Final exam - Part II solutions

(7) [8 pts.] In this problem we examine the median income and education level (percent of population with at least Bachelor's degree) for 50 U.S. cities. Shown below are some graphs and summary statistics for constructing a linear regression model to predict median household income (in dollars) from education level. The correlation is $r = 0.65$.

Income	mean= \$46,520, standard deviation= \$15,500
Education	mean= 28.4 %, standard deviation= 12 %

(a) Find the regression model to predict income from education level. Show all steps.

(b) Carefully check the conditions and discuss the appropriateness of your model.

(c) An economist who has been studying similar questions in other regions claims the true slope for this relationship is \$1020 per %. Carry out a hypothesis test to determine whether your result is consistent with this claim. From software, the standard error for the slope estimate is 144.4.

Solution:

(a) Since we want to predict median income from education:
x-variable = education level (% Bachelor's degree)
y-variable = median household income (in dollars)
Given summary statistics are:

$\bar{x} = 28.4\%, \quad s_x = 12\% \ , \quad r = 0.65$
$\bar{y} = \$46,520, \quad s_y = \$15,500$

Model looks like: $\hat{y} = b_0 + b_1 x$.

$$b_1 = r\frac{s_y}{s_x} = 0.65\left(\frac{15500}{12}\right) = 839.58\ \frac{\$}{\%}$$

Then we have: $\hat{y} = b_0 + 839.58\,x$.

Find b_0 by plugging in $(\bar{x}, \bar{y})$:
$46520 = b_0 + 839.58(28.4) \ \Rightarrow \ b_0 = 46520 - 839.58 \times 28.4 = 22675.93\ \$$

Equation of regression line is:

$\hat{y} = 22675.93 + 839.58\,x$ OR $\boxed{\widehat{\text{Income}} = 22675.93 + 839.58\ (\text{Education level})}$

(b) Condition I: Independent observations? It is not clear whether the sample is random, or at least representative. No relevant information has been provided to verify independence. Thus, we cannot assume this condition is met.

Condition II: Linear relationship? The scatter plot of income vs education level (the first graph) looks approximately linear. So this condition is met.

Condition III: Residuals have constant variability? The scatter plot of residuals vs predicted (2nd graph) is not perfect, but may be close enough to say this condition is met.

Condition IV: Residuals normally distributed? The normal probability plot of the residuals is somewhat nonlinear on both ends. But the bulk of the central portion looks linear enough to assume this condition is met.

(c) Let β_1 denote the true slope of the relationship.

Null hypothesis $H_0 : \beta_1 = 1020$ (\$ per percent)
Alt hypothesis $H_A : \beta_1 \neq 1020$

From the sample we have: $b_1 = 839.58$, $SE = 144.4$, $df = n - 2 = 48$.

Therefore: $t_{48} = \dfrac{839.58 - 1020}{144.4} = 1.249.$ P-value > 0.20

Since the P-value is high, we retain H_0 and conclude that our sampled estimate of the slope is consistent with the researcher's claim that its value is \$1020 per %. We note, however, that the conditions needed for inference may not have all been met.

Grade: 8 points. Distribution: (a)=3 pt, (b)=2 pt, (c)=3 pt.
For (a): 0.5 pt = correctly pick x, y; 1 pt = compute slope; 1.5 pt = find intercept + inculde units at least somewhere + correct regression equation.
For (b): 0.5 pt for checking each of 4 conditions.
For (c): 1 pt for correct/clear hypotheses; 1 pt for t-score + P-value computation; 1 pt for correct conclusion.

(8) [8 pts.] Supermarkets often place similar types of cereal on the same shelf. Researchers collect data on the shelf placement, as well as sugar, sodium, and calorie content of 77 cereals. They test whether the mean sugar content varies by shelf.

(a) Write the null and the alternative hypotheses and explain what your parameters mean.

(b) Complete the ANOVA table and state the conclusion of your hypothesis test.

(c) Are the conditions for ANOVA met?

(d) Suppose we want to carry out a Tukey's HSD test for pairwise comparisons of mean values. What are the pairwise combinations we must test?

Solution:

(a) Null hypothesis: The mean sugar content in the cereals is the same on all 3 shelves.
That is, $\mu_1 = \mu_2 = \mu_3$

Alt hypothesis: At least one of the mean values differs from at least one other.

The parameters μ_1, μ_2, μ_3 denote the true mean amount of sugar in cereals placed on shelf 1, 2, and 3, respectively.

(b) The completed table is shown below

	SS	df	MS	F-ratio	P-value
Factor	248.4	$3-1=2$	$\frac{248.4}{2}=124.2$	7.334	0.0012
Residuals	1253.1	$77-3=74$	$\frac{1253.1}{74}=16.934$		
Total	1501.5	76			

I will assume a significance level of 10%. Since the P-value is below that, we reject the null hypothesis and conclude that at least one of the mean values differs from at least one other. We note, however, that the conditions check in part (c) of this question suggests not all the needed conditions were met.

(c) Condition I: Independent observations? It is not clear whether the sample is random or at least representative. No other information has been provided to verify independence. Thus, we cannot assume this condition is met.

Condition II: Approximately equal variance across groups? The boxplots show similar spread, suggesting this condition is close enough to be met.

Condition III: Data within each group approximately normal? The boxplots suggest some skew, especially for shelf 1 and 2. This condition may not be met.

(d) There are 3 pairwise combinations we must test:

Shelf 1 vs Shelf 2

Shelf 1 vs Shelf 3

Shelf 2 vs Shelf 3

Grade: 8 points. Distribution: (a)=1.5 pt, (b)=3.5, pt, (c)=1.5, (d)=1.5 pt.
For (a): -0.5 pt if parameters or statement of hypotheses not 100% clear;
For (b): 0.5 pt each for (i) SS, (ii) df_B, (iii), df_w, (iv) MS_B, (v) MS_W, (vi) F, (vii) conclusion.
For (c): 0.5 pt each for check of 3 conditions.
For (d): 0.5 pt each for 3 correct pairwise combinations.

(9) [8 pts.] A few years back, a group of Earlham students taking a stats class collected data on smoking patterns among EC students. The table below shows frequency counts of smokers vs non-smokers based on nationality (Domestic, or International).

	Domestic	International
Smoker	25	8
Nonsmoker	53	14

(a) Construct a logistic regression model to predict the probability of smoking based on student nationality (Domestic, or International).

(b) Interpret the meaning of the slope coefficient in your model in the context of odds ratios and student nationality.

(c) Another variable in the smoking data set is number of hours of exercise per week. A logistic regression to predict smoking probability using both nationality and hours of exercise was carried out. The software output is given below. Compute and interpret a 95% confidence interval for the slope of the hours of exercise.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.9204     0.5868   1.568   0.1168
NationalityInternational  -0.9132     0.8002  -1.141   0.2538
Hrs_exercise              -0.2591     0.1025  -2.528   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution:

(a) The conditions for logistic regression require: (i) independent observations; and (ii) linear relationship between each predictor variable and the logit. While independence of observations is unclear in this problem, the linearity condition is met because the predictor is categorical, with two categories.

Let $\hat{y}$ denote the probability of being a smoker. The logistic model has the form

$$\ln\left[\frac{\hat{y}}{1-\hat{y}}\right] = b_0 + b_1(\texttt{nationality})$$

where `nationality` $= 0$ for domestic and 1 for international.
Assuming the sample is random/representative, we can use the data in the above 2-way table to estimate the needed probabilities. Accordingly, for domestic we have

$$\ln\left[\frac{25}{53}\right] = b_0 + b_1(0) \;\Rightarrow\; b_0 = -0.7514$$

Using the data for international students, we get

$$\ln\left[\frac{8}{14}\right] = b_0 + b_1(1) \;\Rightarrow\; b_1 = 0.1918$$

Thus, the logistic model for the probability of being a smoker is

$$\ln\left[\frac{\hat{y}}{1-\hat{y}}\right] = -0.7514 + 0.1918(\texttt{nationality})$$

(b) Interpretation of b_1: The log odds of the probability of being a smoker is on average 0.1918 higher for international students compared with domestic. Equivalently, the odds of an international student being a smoker is higher by a factor of $e^{0.1918} (= 1.21)$ on average compared with those same odds for a domestic student.

(c) The slope coefficient for hours of exercise is -0.2591, and the standard error is $SE = 0.1025$. Thus, the confidence interval is

$$CI = b \pm z^* \times SE = -0.2591 \pm 1.96 \times 0.1025 = (-0.460, -0.058)$$

With 95% confidence, for every additional hour of exercise per week, the predicted log odds of the probability of being a smoker decrease between -0.460 and -0.058 on average, when all other variables are held constant.

Grade: 8 points. Distribution: (a)=3 points, (b)=2 points, (c)=3 points.
For (a): 0.5+0.5 pt=correct form of model + correct treatment of nationality;
1 + 1 pt=find correct b_0 and b_1.
For (b): correct interpretation of either log odds, or plain odds is sufficient.
For (c): 1.5pt = cumpote correct CI; 1.5pt=interpret it correctly.

Bonus question (3 points. No partial credit!)

A data set consists of n pairs of (x, y) values. Let (x_i, y_i) for $i = 1, 2, \ldots, n$ represent these pairs of data. If you want to think in terms of a table, here is what it would look like

x_1	x_2	x_3	$\ldots$	x_n
y_1	y_2	y_3	$\ldots$	y_n

We want to use least squares to find the best regression model of the form $\hat{y} = mx^2$ for these data. Find the model. That means, find an expression for m in terms of the (x_i, y_i) values.

Solution outline:

Key steps that must be shown

$$\begin{aligned} e_i &= y_i - mx_i^2 \\ f(m) &= \sum (e_i)^2 = \sum (y_i - mx_i^2)^2 \\ f'(m) &= 0 \quad \Rightarrow \quad \sum -2x_i^2(y_i - mx_i^2) = 0 \\ &\Rightarrow \quad m = \frac{\sum x_i^2 y_i}{\sum x_i^4} \end{aligned}$$

This minimizes f since $f''(m) > 0$.