Final (100 pts)

Due 12/4 at 11:59 PM ET

The final is an open book format, but must be completed on your own. The final must be submitted by the deadline. **Late submissions will not be accepted**. The questions are a mix of short answer, programming, and interpreting output. For questions that involve programming, please include the corresponding code and output.

Where relevant, use an alpha (α) threshold of 0.05 and a two-sided test.

Use the *heart.csv* dataset to answer the following questions.

Data dictionary:

Heart\_disease: 1=Yes, 2=No

Sex: 0=Female, 1=Male

Chol, Total Cholesterol (mg/dl)

Age, age in years

1. We are interested in examining the predictors of heart disease among adults. We enroll a sample of 1012 adults upon being admitted to the hospital. At the time of their admission, participants are measured on multiple potential predictors (including age, sex, and cholesterol) and whether or not they have heart disease.
   1. What type of sampling plan does this describe (cross-sectional, prospective, or retrospective) and why? (5 pts)

This study describes Cross sectional sampling plan

I choose cross sectional sampling plan in this study because we are assessing patients for exposure (predictors of heart disease) and disease (heart disease) at the same time

Here we are identifying participants irrespective of their treatment/exposure or outcome

* 1. Based on the sampling plan, what measures of association (risk difference, risk ratio, odds ratio) could be calculated from this data? (5 pts)

We can calculate all the three measures of association for this sampling plan (cross-sectional)

* Risk difference
* Risk ratio
* Odds ratio
  1. We are particularly interested in the association between sex (exposure) and heart disease (outcome). Run a chi-squared test to assess whether there is a statistically significant association between sex and heart disease. Be sure to write out your null and alternative hypotheses, p-value, decision, and conclusion. (10 pts)

Table

Description automatically generatedGraphical user interface, text, application

Description automatically generated

H0: Heart disease and sex are **independent**

HA: Heart disease and sex are **associated**

P-value = <0.0001 (<0.05)

**Decision:** Since p<0.05, we reject the null hypothesis

**Conclusion:**  We reject the null hypothesis. We have sufficient evidence to suggest that there is an association between sex and heart disease among heart.csv patients. There is a statistically significant association between sex and heart disease (p=<0.0001)

* 1. Calculate and interpret the odds ratio for heart disease in females (sex=0) compared to males (sex=1). (*Hint: the table orientation is the same as the examples we did in class and on your previous homeworks, so if you calculate the OR in SAS, it will be set up to interpret females vs. males).* (5 pts)

Table

Description automatically generatedText

Description automatically generated

**Odds ratio = 0.2608, 95% CI = 0.1938 – 0.3510**

OR<1: Negative Association

* Odds of disease is lower in the females than in males
* Odds of disease in females is 0.7392 or 73.92% lower than in males
* We are 95% confidence that the true OR falls between 0.1938 and 0.3510
  1. Report and interpret the 95% confidence interval for the odds ratio from part d. How does it relate to the results/decision of your chi-squared test in part c? (5 pts)

We are 95% confidence that the true OR falls between 0.1938 and 0.3510

Based on the 95% confidence interval for the OR, I would conclude that there was a statistically significant association between sex and heart disease. The 95% CI for the OR does not contain 1, which would be the value we would expect if there were no association (the value under the null hypothesis).

* 1. We suspect that increased cholesterol may also be associated with heart disease. We check first to see whether it is associated with sex. If cholesterol were associated with both sex and heart disease, what type of variable might it be? (2 pts)

If cholesterol were associated with both sex and heart disease it might be a confounder

* 1. Plot a histogram of cholesterol (chol) by sex. Provide the appropriate measures of center and spread (**Make sure to include the numbers for your selected measures in your answer**). What hypothesis test would you use to assess whether cholesterol differs between males and females and why? (8 pts)

Chart, histogram

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

Sex (0): female

Table

Description automatically generated

Shape of distribution for variable chol (cholesterol) for sex (0) female is **Symmetric.** The measure of center/spread is mean/Standard Deviation

My data seems to be roughly symmetric. Therefore, I choose mean/Standard deviation (center/spread)

Mean: 253.4314

Standard Deviation: 51.03070

Sex (1): male

Table

Description automatically generated

Shape of distribution for variable chol (cholesterol) for sex (1) men is **Symmetric.** The measure of center/spread is mean/Standard Deviation

My data seems to be roughly symmetric. Therefore, I choose mean/Standard deviation (center/spread)

Mean: 239.2370

Standard Deviation: 43.15553

I will conduct a parametric Two Sample T-Test

* The total number of observations in sex 0 are 229 and sex 1 are 713 where n>30
* We can apply Central Limit Theorem, use the properties of normal distribution and t-test.
* In this scenario we have two groups of unknown mean and testing the hypothesis that they are equal to each other, with n>30, where Central Limit Theorem applies. Hence, we will conduct a Parametric Two Sample T-Test.
  1. Conduct the hypothesis test from part g. Make sure to write out your hypotheses, p-value, decision, and conclusion (10 pts)

I will conduct a **Two Sample T test**

**Null Hypothesis**: Mean cholesterol for female population of heart.csv patients is equal to Mean cholesterol for male population of heart.csv patients

OR

The difference in the mean cholesterol in females and males is equal to zero

**Alternate Hypothesis**: Mean cholesterol for female population of heart.csv patients is not equal to Mean cholesterol for male population of heart.csv patients OR

The difference in the mean cholesterol in females and males is not equal to zero

H0: μ1= μ2 (μ1-μ2 =0)

H1: μ1 ≠μ2 (μ1-μ2 ≠ 0)

H0: μ1= μ2 (μ1-μ2 =0)

Graphical user interface

Description automatically generated with medium confidenceH1: μ1 ≠μ2 (μ1-μ2 ≠ 0)

Table

Description automatically generated

* P-Value for equality of Variances= 0.0004

Here our P-Value is <0.05 (P-value< α)

* If P-Value is less than 0.05 we can assume that our Variance are not equal, and we will use the Satterthwaite method for P value of t test
* Therefore P-Value=<0.0001 (P<0.05 or P-value<α)

**Interpretation**

There is 0.01% chance of observing our sample means of 253.4 and 239.2 if the population of heart.csv in females and male’s group has the same cholesterol level

**Decision**

Since this result is very unlikely if the null hypothesis were true (P-value <0.05 or P-value<α), we reject the null hypothesis. We conclude that the mean cholesterol for males and females are significantly different

* 1. Based on the output below from a multiple logistic regression, predicting heart disease from sex, age and cholesterol, interpret the p-value, OR and 95% confidence interval for the relationship between sex (females compared to males) and heart disease (*Hint:*  *use the interpretation slides from Week 6 lecture 4 as a guide)* (10 pts)

Table

Description automatically generated

Interpretation for **sex**:   
Βsex=-1.7010, P-value= <0.0001

OR (female vs male) = 0.183

95% CI = 0.131 – 0.254

The change in log-odds of heart disease for sex is significantly different than 0 (p=0.0001). There is a significant association between sex and heart disease in heart.csv patients, ***after adjusting for cholesterol and age.*** Compared to males, females have 81.7% ((1-0.183)%) lower odds of heart disease. We are 95% confident the true odds ratio is between 0.131 and 0.254

Interpretation for **cholesterol**

Βchol= 0.00639, P-value= <0.0001

OR (95% CI) = 1.006 (1.003 – 1.009)

The change in log-odds of heart disease for cholesterol is significantly different than 0 (p=<0.0001). There is a significant association between cholesterol and heart disease in heart.csv patients, ***after adjusting for sex and age.*** For every 1-year increase in cholesterol, odds of heart disease increase 1.006-fold. We are 95% confident the true odds ratio is between 1.003 and 1.009.

Interpretation for **age**

Βage= 0.0646, P-value= <0.0001

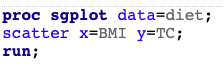
OR (95% CI) = 1.067 (1.050 – 1.084)

The change in log-odds of heart disease for age is significantly different than 0 (p=<0.0001). There is a significant association between age and heart disease in heart.csv patients, ***after adjusting for sex and cholesterol.*** For every 1-year increase in age, odds of heart disease increase 1.067-fold. We are 95% confident the true odds ratio is between 1.050 and 1.084.

1. Two groups of researchers are separately investigating the effect of a new therapy on outcomes. Both studies recruit patients from the same population, randomize patients to treatment and placebo arms, and measure the same outcome. The only difference in how they conduct their studies is that Research Group A enrolls 50 patients, and Research Group B enrolls 150. Group A is unable to detect a difference in the treatment groups, while Group B is able to conclude that the patients receiving the therapy did better. In 1-2 sentence, describe why the studies reached different conclusions. (5 pts)

Studies reached different conclusions because of Sample size. Two groups have two different sample sizes (A-50, B-150). Greater the sample size greater differences we can detect and greater precision of a study. Therefore, small sample size in group A failed to detect difference in treatment group.

1. We are running a dietary study to investigate predictors of total cholesterol (TC), including BMI (BMI), and dietary intake of animal fat (animfat), and alcohol (alcohol).
   1. Based on the scatterplot below, describe the relationship between BMI and total cholesterol (5 pts)



Chart, scatter chart

Description automatically generated

There appears to be a slight positive association between BMI and cholesterol.

* 1. Table

     Description automatically generatedReport the correlation coefficient (r) between BMI and total cholesterol and interpret. Comment on whether the correlation is significantly different from 0. (5 pts)

Text

Description automatically generated with medium confidence

r= 0.33506

P-value= 0.0162 (P<0.05)

There is a positive association between BMI and total cholesterol (r=0.33506). The correlation is significantly different from 0 (p=0.0162).

* 1. Based on the output below of the simple linear regression predicting total cholesterol from BMI, report and interpret the slope and 95% confidence interval. Comment on whether the slope differs significantly from 0. (5 pts)

Table

Description automatically generated

Bchol = 5.0445

95% CI = 0.97210 – 9.11679

P-value = 0.0162

There is a positive linear relationship between total cholesterol and BMI. The slope is significantly different from 0 (p=0.0162). For every 1 unit increase in BMI, total cholesterol increases by 5.0445 on average. We are 95% confident the population slope is between 0.97210 – 9.11679.

* 1. Using the output from part c, Identify and interpret the coefficient of determination (R2). (5 pts)

R2= 0.1123

R^2 is 0.1123. This means that 11.23% of the variability in cholesterol is explained by BMI. Although there is a linear relationship, BMI only explains a 11.23% of the change in cholesterol.

* 1. Compare the following two models for predicting total cholesterol (Model 1: BMI+Alcohol; Model 2: BMI+Alcohol+Animal Fat) for parts e-g. Which model would you choose and why? (5 pts)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Model 1** | |  | **Model 2** | |  |
|  | Estimate  (95% CI) | p-value | VIF | Estimate  (95% CI) | p-value | VIF |
| BMI | 4.8 (0.8-8.7) | 0.0193 | 1.00 | 4.5 (0.6-8.4) | 0.0247 | 1.01 |
| Alcohol | 0.35 (0.02-0.67) | 0.0392 | 1.00 | 0.34 (0.01-0.66) | 0.0435 | 1.01 |
| Animal Fat |  |  |  | 0.43 (-0.19- 1.06) | 0.1686 | 1.01 |
| R2 | 0.1544 | | | 0.1709 | | |

I will select Model-1

I choose Model-1 because in Model-2 the P-Vale for Animal Fat is 0.1686 (P>0.05), which says there is no relationship between Animal Fat and Cholesterol, which made me reject model-2 and select Model-1.

* 1. Interpret the R2 for the final model you picked in part e. (2 pts)

R2= 0.1544

R^2 is 0.1544. 15.44% of the variability in total cholesterol can be explained by BMI and alcohol.

* 1. Report and interpret the slopes and 95% confidence intervals for each variable in the final model you selected in part e (8 pts)

Bvv gb bnm

**BMI**

P-value= 0.0193

BBMI= 4.8

95% CI = 0.8-8.7

For every 1 unit increase in BMI, total cholesterol increases by 4.8 units, on average, adjusting for alcohol. We are 95% confident the true slope is between -0.8-8.7

**Alcohol**

P-value= 0.0392

Balcohol= 0.35

95% CI = 0.02-0.67

VIF=1.00

For every 1 unit increase in alcohol, total cholesterol increases by 0.35 units, on average, adjusting for BMI. We are 95% confident the true slope is between -0.02-0.67