Peer Assessments (https://class.coursera.org/introstats-001/human_grading/) / Assignment 2

Submission Phase

1. Do assignment ☑ (/introstats-001/human_grading/view/courses/970433/assessments/7/submissions)

Evaluation Phase

2. Evaluate peers ☑ (/introstats-001/human_grading/view/courses/970433/assessments/7/peerGradingSets)

Results Phase

3. See results ☑ (/introstats-001/human_grading/view/courses/970433/assessments/7/results/mine)

Your grade is 2 , which is simply the grade you received from your peers.

See below for details.

Introduction to the Assignment

Hello everyone and welcome to your final assignment for the course! This assignment asks you to apply several of the methods and concepts that you've learned in the course, to the analysis of some data collected in a scientifc study. You'll have to look through various tables and figures about these data, and answer six different statistical questions about them, using the statistical ideas you have learned

Show more assignment details

Question 1
We're interested in the prevalence of the A variant of the TNF-alpha gene at position 308. It is thought that this variant is less common, that is, it occurs in less than half of the population. Carry out an hypothesis test to test this. In your answer, state:

- Whether you're carrying out a one-sided or two-sided test.
- The value of the test statistic that you calculate.

- Based on the test statistic, whether the P-value is large or small.
- Your conclusion.

- I would carry an one-sided test, namely I would use a null-hypethesis **p(A308) = 0.5** versus an alternative hypothesis **p(A308) < 0.5**
- Our observed variable is **p_hat = 0.3** (table 3)
- The test statistic is **(p_hat - p) / sqrt(p * (1 - p) / n)**, which, given **p = 0.5** and **n = 110** (table 3), works out to approx **-4.195**
- Based on the test statistics, we can calculate **p-value = P(N(0, 1) <= -4.195) = 1.362942e-05**
- The p-value is too small so we reject the null-hypothesis and conclude that the *A308* variant occurs less than in the half of the population

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → excellent

peer 2 → excellent

peer 3 → Awesome answer, covering every part of the question perfectly!

peer 4 → Excellent - all aspects addressed.

peer 5 → Excellent answer. You addressed all required components of the test and made correct conclusion. Excellent.

Question 2
From our data, we estimate that 28.2% of the population has the A variant at position 238, and 30.0% has the A variant at position 308, where the "population" here is the population that is represented by our subjects. We're interested in whether the prevalence of the A variant is the same for position 238 and position 308 in this population. That is, is the proportion of people with the A variant at position 238 the same as the proportion of people with the A variant at position 308? Using only the information you are given in the pdf document of plots and summary statistics, this test cannot be carried out. Why not?

This would be a matching pair test, but for carrying it we would need to know the corresponding values of each pair - and we don't know that.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → excellent

peer 2 → good

peer 3 → Good. Saying that we need a matched pair test means that you know that in this case the two proportion groups are not independent.

peer 4 → Excellent - I think from your response you're clear that these are not independent groups.

peer 5 → This is incorrect answer. It cannot be carried out because data is not independent. Poor.

Question 3

Construct a 95% confidence interval for the difference in the proportion of males with the A variant at position 308, and the proportion of females with the A variant at position 308. State the confidence interval that you calculate. Based on your confidence interval, is there evidence that the proportion of the A variant at 308 differs between males and females? You may use the fact that the critical value for a 95% confidence interval from a standard normal distribution is 1.96.

- From table 5 for males we have **pm_hat = 0.228** and **nm = 58**, and for females **pf_hat = 0.310** and **nf = 52**
- The critical value is **1.96 * sqrt(pm_hat * (1 - pm_hat) / nm + pf_hat * (1 - pf_hat) / nf),** which yields **cv = 0.1657**
- The 95% confidence interval is **(pm_hat - pf_hat) +- cv**, which works out to **[0.0837 -0.2477]**
- Based on the confidence interval we may conclude that there is no evidence that the proportion of A308 differs between males and females as the interval includes **0**

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → excellent

peer 2 → good

peer 3 → Good job. Wrong confidence interval but correct interpretation. (next time for interval, list the lower number on the left and bigger one on the right)

peer 4 → Good - the confidence interval calculation is slightly off, but you've made the correct conclusion based on the fact that the CI included zero. The CI equation works out to 0.31-0.288 +/- 1.96*SQRT(((0.31*(1-0.31))/58)+((0.288*(1-0.288))/52)) = [0.193, -0.149] (or you could choose to reverse the order of males and females to get [-0.193, 0.149]).

peer 5 → You took correct values for the proportions, but there is mistake in your confidence interval calculations. You draw correct conclusions. Good.

Question 4

Carry out a statistical test to determine if the mean of HDL is the same for males and females for the population represented by this study. You should assume that the variance is the same for males and females. In your answer, state:

- Whether you're carrying out a one-sided or two-sided test.
- The value of the test statistic that you calculate.
- Based on the test statistic, whether the P-value is large or small.
- Your conclusion.

- Let **X_m** denote the observed mean for males and **X_f** - for females, and **mu_m** and **mu_f** - the actual means of the population
- Also, let **s^2_m** and **s^2_f** denote variance for males and females respectively, and **nm** and **nf** - the number of males and females
- The null hypothesis is that **mu_m** *does not equal* **mu_2** versus **mu_m** *equals* **mu_f**

- Assuming that the variance is the same, we can calculate pooled variance **s^2 = [(nm - 1) s^2_m + (nf - 1) s^2_f] / [nm + nf - 2]**, which, using table 4, yields **0.06028**
- Next, we calculate test statistics **[(x_m - x_f) - (mu_m - mu_f)] / sqrt(s^2 / nm + s^2 / nf) = -3.7536**
- The p-value for the test would be **2 * P(|t_df| <= 3.7536)**
- As t-distribution becomes closer to N(0, 1) as the number of degrees of freedom grows, we can assume that for **df = 108** the z-value is **1.96**
- **3.7536** is far from **1.96** and that means that the p-value is small enough to reject the null hypothesis
- And conclude that the true mean of HDL is the same for both males and females

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → good, mean is not the same

peer 2 → good but the conclusion was false.

peer 3 → The null hypothesis should always assume that both means are equal, whereas the alternative would say they are not. Having the p-value this small, we can reject the null hypothesis. There is enough evidence to suggest that true means of HDL differ for both males and females. Wrong interpretation of the answer because you have wrong null hypothesis statement. You're Okay since other parts are good. So, I give you GOOD.

peer 4 → Fair. In this instance the null hypothesis would actually be that there is no difference between the means (i.e., that the mean HDL for males and females is the same). The test statistic calculation is correct and the statement is accurate that the p-value is small enough to reject the null, but the conclusion is inaccurate. Since the null hypothesis should be that there is no difference in the mean HDL between the male and female groups, we'd conclude that there is a difference between the two groups.

peer 5 → Another excellent answer. You addressed all required components of the test and made correct conclusion. Excellent.

Question 5

It is believed that higher consumption of polyunsaturated fats (PUFA) increases HDL cholesterol levels, but this relationship may be affected by whether or not a subject has the A genetic variant at position 308. Does this seem to be the case based on the plots and summary statistics you are given? Investigate this question by examining whether or not there is a relationship between HDL and PUFA, and whether or not that relationship differs between subjects who have and do not have the A variant at position 308. Support your answer by mentioning relevant plots or summary statistics. Indicate whether or not you have any concerns about the appropriateness of the analysis of the HDL - PUFA relationship; that is, might a different type of analysis, or analysis on modified data be more appropriate?

According to figures 3, 4 and 5, there is no linear relationship between *HDL* and *PUFA*, for all subjects, for subject with *A308* and for subjects without it. That is supported by the correlation coefficients (table 6), which are near zero for all cases as well as by the results of the linear regression (table 7).

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → excellent

peer 2 → excellent

peer 3 → Fair. Including analysis of residual plots and the p-value of beta would be nice also. The key says there seems to be weak negative relationship between subjects with A variant at position 308 and PUFA, which I disagree.

peer 4 → Good - the description is not extensive, but the conclusion for the first portion of the question is correct. It seems that the 2nd portion of the question was overlooked.

peer 5 → Correct answer. You analysed relevant parts of the tables and figures. You made correct conclusions. Good.

Question 6

In addition to polyunsaturated fats, many different components of diet could be investigated (for example, total calories consumed, and percent of dietary intake from protein, carbohydrates, alcohol, and other types of fat), and many different risk factors of heart disease could be investigated (for example, LDL cholesterol, triglycerides, total cholesterol, and C-reactive protein). So this assignment reflects only one small part of the analysis that was carried out. How does knowing this affect the implications of the conclusions you made?

We should pay careful attention to extraneous/confounding variables as they affect the regressions and introduce bias.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Please refer to the answers document (https://spark-public.s3.amazonaws.com/introstats/Assignment2/XA2/yyz_A2ans.pdf). Rate the answer as Excellent, Good, Fair, or Poor as outlined in the answers document and provide some written feedback for why you gave the answer the rating you did.

peer 1 → good

peer 2 → fair

peer 3 → Good. Looking for keywords: Type I error, observational studies, confounding variables

peer 4 → Fair - the point made regarding confounding is very valid (I think the point you're making here is that it would be worthwhile to run an analysis that includes controls for additional factors such as sex). The main risk addressed by the question is that of Type I error due to performing a large number of analyses.

peer 5 → Unfortunately incorrect answer. You didn't noted that there is possibility of making Type I errors. Also you didn't commented on the observational character of the study. Poor.

Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Please give the assignment you have just read an overall rating of 0, 1, or 2. Give a 2 if your peer made a good effort to answer each of the questions. Give a 1 if your peer's effort was somewhat lacking, because their answers did not reflect many of the features of good answers, or perhaps because they did not answer every question. Give a 0 ONLY in the most EXTREME cases in which your peer did not make any effort at all on this assignment. We expect that most students will earn a 2.

Score from your peers: **2**