**IST 686 Midterm Exam**

**Name: Yunkai Yao**

**SUID (last 4 digits): 3618**

**General Instructions**: **This is an honor system exam that is open book and open notes.** **You must not confer or collaborate with any human besides me and you must not include any AI generated material.** **Any indication otherwise will result in a grade of zero (0) and be reported to the University.**

**You may access all course materials including any of the feedback I have provided to you. Submit this word document answering the exam questions. For #3, include the R code used and the output from R. Submit the exam in Blackboard by 5:15 PM on Wednesday, March 6, 2024. Exams submitted after this time will incur a late penalty.**

**^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^**

**1. (10 pts)** A sampling distribution of many samples of size 75 has a mean of 215 and a standard error of 10. **What, if anything, do you know about the shape, mean, and standard deviation of the population distribution these samples came from?**

According to the central limit theorem, if the sample size is large enough, the sample distribution of the sample mean approximates a normal distribution. The mean of the sample distribution is 215, which is also the mean of the population distribution. The standard deviation is calculated to be approximately 86.6. Due to the large sample size, the central limit theorem ensures that the shape of the sample mean distribution is approximately normal. However, if the original population distribution itself is normal, then the sample distribution of the mean will also be normal, no matter how large the sample size is.

**2. (15 pts)** The effectiveness of a diet designed to lower cholesterol was tested on 45 patients identified as at risk for cardiovascular disease. Cholesterol levels for these patients were measured before and after being on this diet for 8 weeks.

The following 95% confidence interval for the difference in mean cholesterol levels was found based on the data collected.

**-1.2 < before - after cholesterol levels < 20.3**

**What would you conclude regarding the effectiveness of this diet based on this confidence interval? Be specific in terms of the context of this research.**

Based on the 95% confidence intervals provided in this study, the confidence intervals ranged from -1.2 to 20.3. After 8 weeks on this diet, the mean cholesterol level of the patients decreased. The lower limit of the confidence interval (-1.2) suggests that cholesterol levels may have increased slightly in some cases, but this result is not significant because it is close to zero and not representative of the majority of cases. The upper limit of the confidence interval (20.3) indicates that cholesterol levels may have decreased by 20.3 units in some patients. Overall, the confidence intervals do not include zero and are mostly above zero. Confidence intervals around 0 are also asymmetric, further indicating that most of the effect is a decrease rather than an increase in cholesterol levels. This suggests that this diet is generally effective and can reduce cholesterol levels to some degree in patients at high risk for cardiovascular disease.

**3.** **A data set contains data from Instagram users regarding time spent on Instagram, their profession, and whether the user owns a home.**

The variables are as follows.

time: Time spent on Instagram

profession: Profession of the user

home: Whether or not the user owns a home (TRUE or FALSE)

Based on the data collected, your first task is to answer the question **“Is there a difference** **in** **time spent on Instagram based on whether or not the user owns a home?”**

Import the data file and run the following code in R, **placing the last 4 digits of your SUID for the seed.**

**Make sure you are using myData below, a sample of 100, for all analyses and not the entire data set!**

***set.seed(####)***

***myData<-Time[sample(1:nrow(Time), 100, replace=FALSE),]***

**For a – d below, provide a brief summary as well.**

1. **(5 pts)** Examine your data with str() and summary statistics.

图形用户界面, 文本, 应用程序

描述已自动生成

文本, 信件

描述已自动生成

**b. (5 pts)** Compare these distributions with boxplots graphed together.



图表, 箱线图

描述已自动生成

**c. (10 pts)** Run a frequentist analysis to determine if there is a difference in time spent on Instagram based on whether or not the user owns a home. State the null and alternative hypotheses, report on the results of the t-test and the confidence interval.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

The null hypothesis is that there is no difference in the average time spent on Instagram by owners and non-owners, while the alternative hypothesis is that there is a difference in the average time spent on Instagram by owners and non-owners. According to the results of the t-test, the t-value is 4.9876 which is very high. The test has 98 degrees of freedom. The p-value is 2.645e-06 which is much less than the typical alpha level of 0.05. This indicates that the observed mean difference is highly statistically significant. Thus, supporting the alternative hypothesis that there is indeed a mean difference between the two groups. The 95% confidence intervals for the mean time spent on Instagram ranged from 5.346225 to 6.726502 for owners and 3.111536 to 4.310687 for non-owners. These confidence intervals do not overlap, which confirms the results of the t-test that there is a significant difference in the mean time spent on Instagram between the two groups. Owners spend more time on Instagram than non-owners.

**d. (10 pts)** Run the corresponding Bayesian analysis. Interpret the HDI for the difference of means. Also report the percentage of values in the posterior distribution that are above zero and below zero. Use either the BEST package in R or the website that runs the same analysis.

<https://www.sumsar.net/best_online/>

文本

描述已自动生成

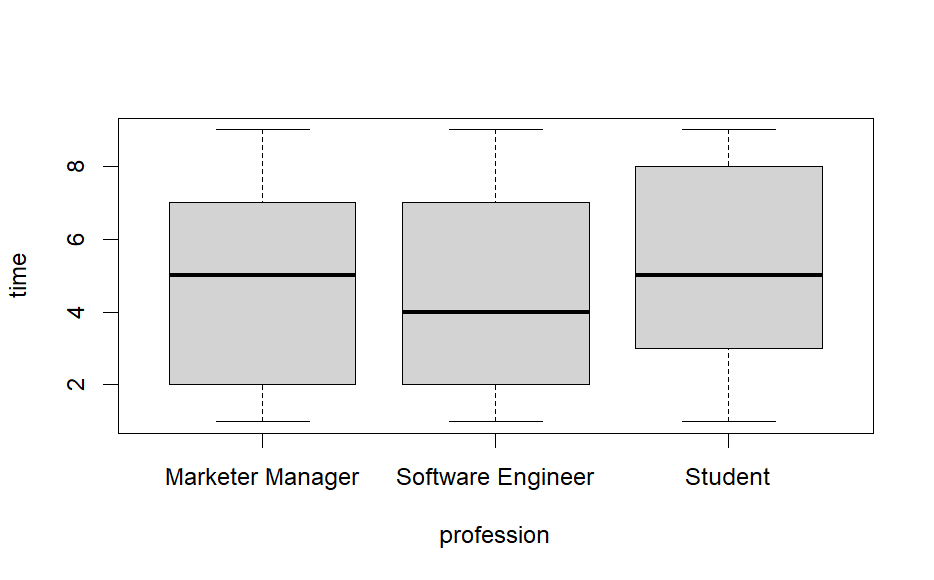
In the Bayesian analysis, owners were set to 1 and non-owners were set to 2. Based on the HDI values, we can 95% determine that the true mean difference between owners and non-owners is between 3.45 and 4.00. This interval does not contain zero. This suggests that the two means are quite different. In the posterior distribution, we can infer that 100% of the mean difference of the posterior distribution is greater than zero because the entire 95% of the HDI is greater than zero. This suggests that almost certainly owners in the analysis spend more time on Instagram on average than non-owners in the analysis.

**e. (10 pts)** **Is there a difference in time spent on Instagram based on whether or not the user owns a home? Write a paragraph to answer this question providing support from both analyses.**

According to the results of frequentist theory and Bayesian analysis, there is a difference in the time spent on Instagram between users who own houses and users who do not own houses. The frequency theory approach using a t-test yielded statistically significant results with P-values ​​well below 0.05. The mean difference between the two groups of users was also significant, indicating that owners spent more time on Instagram. This is also confirmed by Bayesian analysis, where the 95% HDI of the mean difference does not include zero and is significantly above zero, indicating that owners are likely to spend more time on Instagram than non-owners. In addition, the range of HDI is narrow (3.45-4.00), indicating that the estimate of the mean difference is accurate. Bayesian analysis also shows that the posterior distribution of the mean difference is 100% above zero, which is strong evidence that a real difference exists. The combined evidence from these two analyzes strengthens the conclusion that Instagram usage differs significantly by the presence or absence of housing.

**4. (35 pts) Your next task will be to determine if there is a difference in time spent on Instagram based on the users profession.**

**The following is the R output for both the frequentist and the Bayesian analyses. Provide a complete interpretation of these results in a paragraph at the end.**



aovOut<-aov(time~profession,data=myData)

> summary(aovOut)

Df Sum Sq Mean Sq F value Pr(>F)

profession 2 9.3 4.673 0.632 0.534

Residuals 97 716.8 7.390

aovBayesOut<-anovaBF(time~profession,data=myData)

|==================================================================================| 100%

Warning message:

data coerced from tibble to data frame

> aovBayesOut

Bayes factor analysis

--------------

[1] profession : 0.1532161 ±0.03%

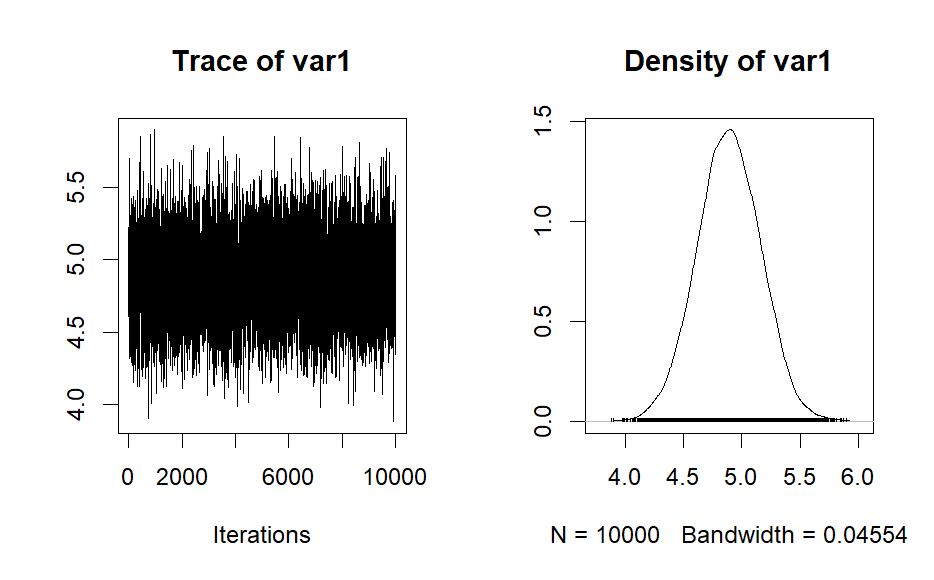
Against denominator:

Intercept only

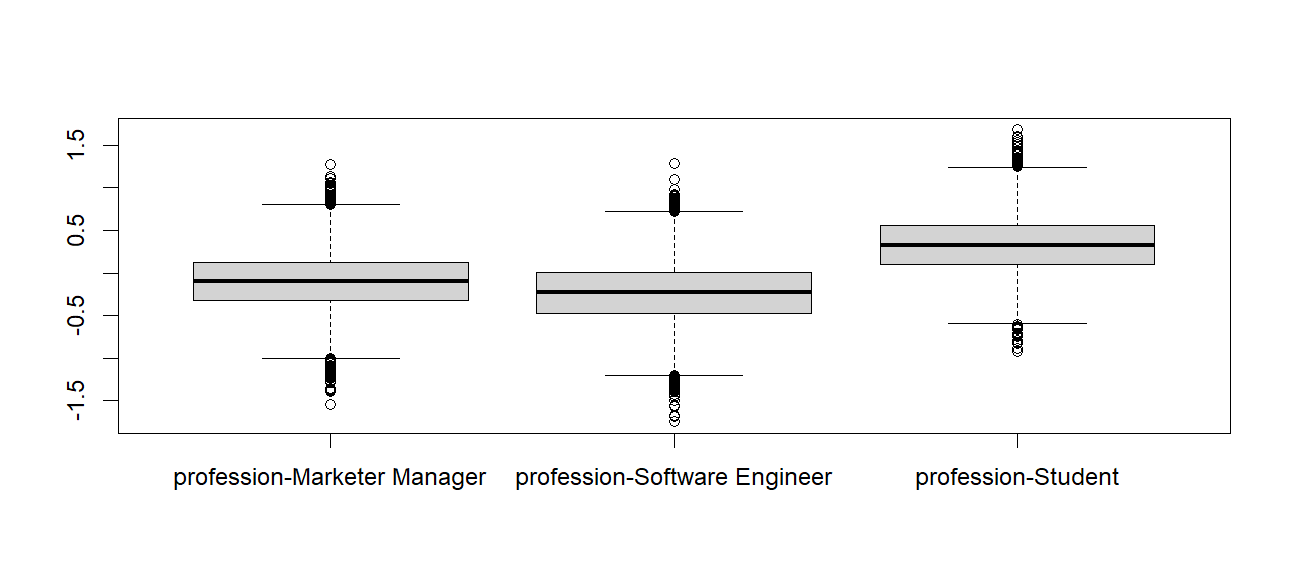
---

Bayes factor type: BFlinearModel, JZS

plot(mcmcOut[,"mu"])



boxplot(as.matrix(mcmcOut[,2:4]))



summary(mcmcOut)

Iterations = 1:10000

Thinning interval = 1

Number of chains = 1

Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,

plus standard error of the mean:

Mean SD Naive SE Time-series SE

mu 4.89012 0.2714 0.002714 0.002714

profession-Marketer Manager -0.09804 0.3431 0.003431 0.003431

profession-Software Engineer -0.23454 0.3606 0.003606 0.003606

profession-Student 0.33258 0.3437 0.003437 0.003437

sig2 7.41436 1.0770 0.010770 0.010664

g\_profession 0.31966 1.6510 0.016510 0.017078

2. Quantiles for each variable:

2.5% 25% 50% 75% 97.5%

mu 4.35916 4.71081 4.88841 5.074027 5.4125

profession-Marketer Manager -0.78063 -0.32394 -0.09914 0.126805 0.5829

profession-Software Engineer -0.95135 -0.47819 -0.22766 0.003469 0.4659

profession-Student -0.32503 0.10037 0.32750 0.558412 1.0263

sig2 5.59153 6.65023 7.30972 8.072854 9.7768

g\_profession 0.03279 0.07792 0.13511 0.263210 1.4466

**Write your complete interpretation of these results below.**

Frequency analysis (ANOVA) showed a p-value of 0.534, indicating that there is no significant difference in the amount of time spent on Instagram across occupations. This is further supported by the Bayesian analysis. The Bayesian analysis resulted in a p-value of 0.1532161 suggesting that the data is more likely to be based on the null hypothesis that there is no difference between occupations rather than the alternative hypothesis. Means and standard deviations for the various occupations, as well as quartiles for each variable, are also provided to highlight the statistical characteristics of the time spent on Instagram for the various occupations. Overall, the results indicate that occupation does not have a significant effect on time spent on Instagram.