STAT 502 - Homework 3

Due date: Thursday, October 28th, 23:59PM. Submit your homework solutions to the course Canvas page. Total points: 20. **Late homework will not be accepted.**

1. **(7 Points)** In this exercise, we will be considering the `sleep` data set (use `?sleep` for more information on the data set). As a response you are given the increase in hours of sleep when taking a certian medication (medication: A, or B). Consider testing $H_0 : \mu_A = \mu_B$ vs. $H_1 : \mu_A \neq \mu_B$ and assume that your samples are i.i.d. and drawn from $Y_{1,A}, \ldots, Y_{n_A,A} \sim N(\mu_A, \sigma^2)$ and $Y_{1,B}, \ldots, Y_{n_B,B} \sim N(\mu_B, \sigma^2)$. As a test statistic consider the two-sample t-statistic $t(\mathbf{Y}_A, \mathbf{Y}_B)$ (for equal variances).

   (a) **(1pt)** What is the 95% two-sided confidence interval for $\mu_A - \mu_B$?

   (b) **(3pt)** A non-central t-distributed random variable can be represented as:

   $$T = \frac{Z + \gamma}{\sqrt{X/\nu}}, \tag{1}$$

   where Z is a standard normal random variable, γ is a constant and X is a χ^2 distributed random variable with ν degrees of freedom, independent of Z (see slide 13 from the "Confidence intervals and power" lecture).

   If the true difference $\mu_A - \mu_B$ is δ, prove that the t-statistic from 1a(a) follows a non-central t-distribution with a non-centrality parameter γ **(1pt)**. You can use that $\overline{Y}_A - \overline{Y}_B$ is independent of s_p^2 and that $\frac{n_A+n_B-2}{\sigma^2} s_p^2$ has a chi-squared distribution with $n_A + n_B - 2$ degrees of freedom.

   i. **(0.5pt)** What part of the t-test statistic $t(\mathbf{Y}_A, \mathbf{Y}_B)$ corresponds to Z in equation (1)?

   ii. **(0.5pt)** What part of the t-test statistic $t(\mathbf{Y}_A, \mathbf{Y}_B)$ corresponds to X in equation (1)?

   iii. **(1pt)** What is the non-centrality parameter γ of the t-test statistic?

   (c) *(Balance)* **(2pt)** Suppose we are going to run a new two-group completely randomized design to compare another 2 sleep medications and suppose we have a total of N people participating in our experiment. Let n_A and n_B the respective sample sizes of the two groups. Suppose that the population variance is equal to σ^2 in both groups and that the true mean difference $\mu_A - \mu_B = \delta$ and σ^2 are known. What values of n_A and n_B with $n_A + n_B = N$ will maximize the power of this two sample t-test? To do this, you may use the fact that the power is an increasing function of the of the absolute value of the non-centrality parameter. Provide a "rough proof" of your answer, by treating the sample sizes n_A and $n_B = N - n_A$ as continuous variables.

2. **(2 Points)** Assume that the data comes from the treatment mean model, $Y_{ij} = \mu_i + \epsilon_{ij}$, $1 \leq i \leq m, 1 \leq j \leq n$, where ϵ_{ij} are i.i.d. random variables with mean $E[\epsilon_{ij}] = 0$ and variance $Var[\epsilon_{ij}] = \sigma^2$ for all $i = 1 \ldots m$, $j = 1 \ldots n$. Let $\boldsymbol{\mu} = (\mu_1, \ldots \mu_m)^T$ and $\bar{\mu} = \frac{1}{m} \sum_i \mu_i$. Prove that

   $$E[MST] = E\left[\frac{\sum_{i=1}^m n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{m-1} | \boldsymbol{\mu}\right] = \sigma^2 + \frac{n \sum_{i=1}^m (\mu_i - \bar{\mu})^2}{m-1}$$

   (slide 16 in "Introduction to ANOVA slides").

3. **(7 Points)** 24 animals were randomly assigned to 4 different diets to study the effect of diet on blood coagulation time.

Treatment							
A	62	60	63	59	64		
B	65	67	73	65	66		
C	69	66	71	67	67	68	62
D	66	62	65	61	64	65	63

   (a) **(2pt)** Write out the treatment variation (effects) model for the experiment. Explain the meaning of the mean parameters $\mu, \tau_A, \ldots, \tau_D$. State the assumptions of the treatment effects model.

   (b) **(1pt)** Plot the data, compute the group means and the overall mean of the data. Do these indicate a difference in coagulation time for the 4 diets?

   (c) **(1pt)** Compute the group sample variances s_i^2 and the pooled estimate of variance MSE.

   (d) **(1pt)** Compute MST and compare it with MSE (without formal test).

   (e) **(1pt)** Compute the analysis of variance table (ANOVA) table with the p-values. Would you say that there is a difference in coagulation times for these four diets?

   (f) **(1pt)** Compute the residuals $\hat{\epsilon}_{ij}$ as $\hat{\epsilon}_{ij} = y_{ij} - \hat{\mu}_i$. Do the residuals appear to come from a normal distribution? Plot the `qqnorm()` plot and analyze the output.

4. **(4 Points)** The dataset in `zinc.RDS` (available on Canvas) contains Zinc levels (variable `ZINC`) with levels background, low, medium, high of different rivers and the corresponding biodiversity (variable `DIVERSITY`). We aim to investigate, whether biodiversity is the same regardless of the Zinc levels.

   (a) **(1pt)** Plot the data, compute the group means and the overall mean of the data. Do these indicate a difference in biodiversity of rivers with different Zinc levels?

   (b) **(1pt)** Compute the group sample variances s_i^2 and the pooled estimate of variance MSE.

   (c) **(1pt)** Compute MST and the F-ratio. Is the F-ratio value what you would expect under the null?

   (d) **(1pt)** Perform a randomization test using a Monte-Carlo sample of size 1000 (sample with replacement) and calculate the p-value for your observed F-ratio. Do different Zinc levels appear to affect biodiversity?