Medical Statistics 2^{nd} Semester Take Home Final Exam

Due Date: Dec 19 2022 (16:00) to Dec 23 2022 (23:59)

Name: ________________________________

Notice

- Please **DO SOLVE ANSWERS BY YOURSELVES!!**
- You can use materials from other textbooks, lecture notes, and websites but you have to provide proper **CITATIONS**.
- Write down your answers in the **MS Word Document** and save your word filename like [`student number-name.docx`, i.e. `202015015-boncho-ku.docx`]. it is admittable to submit your answers by converting your word file to pdf.
- If you make up your mind to submit your answers, send an E-mail to Dr. Boncho Ku and Dr. Mimi Ko with the attachment of your answer file.

Questions

1. Which one of the following statements is False?

   a) The probability of a type II error is the probability that you reject the null hypothesis when it is true. then are followed to document occurrence of disease. (**FALSE**)

   b) Subjects are enrolled or grouped on the basis of their exposure, then are followed to document occurrence of disease in prospective cohort study. (**TRUE**)

   c) Especially when more than 20% of cells have expected frequencies < 5, we need to use Fisher's exact test to determine if there are associations between two categorical variables. (**TRUE**)

   d) To use the two-sample t-test, we need to assume that the data from both samples are normally distributed and they have the same variances. (**TRUE**)

   e) In test for heterogeneity of meta-analysis, if Higgins $I^2 \leq 25\%$, studies are regarded homogeneous and the fixed effect model of meta-analysis can generally be used.bles. (**TRUE**)

2. There are five urns, and they are numbered 1 to 5. Each urn contains 10 balls. Urn i has i defective balls and $10 - i$ non-defective balls, $i = 1, 2, \ldots, 5$. For example, urn 3 has three defective balls and seven non-defective balls. Consider the following random experiment: First an urn is selected at random, and then a ball is selected at random from the selected urn. Suppose that the experimenter does not know which urn was selected. Let's ask two questions.

   a) What is the probability that a defective ball will be selected?

   b) If we have already selected the ball and noted that it is defecvive, what is the probability that it came from urn 5?

Answers

Let A denote the event that a defective ball is selected and B_i denote the event that urn i is selected, $i = 1, \ldots, 5$. Then $P(B_i) = 1/5$ and $P(A|B_i) = i/10$, $i = 1, \ldots, 5$. Using the **theorem of total probabilities**, The solution of a) is

$$P(A) = \sum_{i=1}^{5} = P(A|B_i)P(B_i) = \sum_{i=1}^{5} \frac{i}{10} \cdot \frac{1}{5} = \frac{1}{50} \sum_{i=1}^{5} i = \frac{15}{50} = \frac{3}{10}$$

Employing Bayes' formula, the solution of b) is

$$P(B_5|A) = \frac{P(A|B_5)|P(B_5)}{\sum_{i=1}^{5} P(A|B_i)P(B_i)} = \frac{1/2 \cdot 1/5}{3/10} = \frac{1}{3}$$

3. The distribution of grades in a large statistics course is as follows:

Grade:	A	B	C	D	F
Probability	0.1	0.4	0.3	0.1	0.1

To calculate student grade point averages, grades are expressed in a numerical scale with A=4, B=3, and so on down to F=0.

a) Find the expected value.

b) Describe your strategy to **simulate** choosing students at random and recording their grades.

c) Based on your strategy described in b), perform the simulation with a sample size of size 30 and calculate the mean of their 30 grades **using R**.

d) Repeat c) 10,000 times and calculate the average of 10,000 means.

e) Make a histogram of 10,000 means.

f) Describe your conclusion based on the results of a) to e).

Answers

Let X be a random variable representing numerical points for each grade. The solution of a) is $E(X) = 0.1 \cdot 4 + 0.4 \cdot 3 + 0.3 \cdot 2 + 0.1 \cdot 1 + 0.1 \cdot 0 = 2.3$

Using `sample()` function implemented in R, we can generate synthetic samples with the assignment of the given probabilities and scores for grades. For example, we assume that 30 students are randomly selected from the given probability distribution of grades with replacement. Then we simply write down scripts as follows (solution b) to c)).

```r
set.seed(20221225)  # for the reproducibility
p <- c(0.1, 0.4, 0.3, 0.1, 0.1)
x <- 4:0

xi <- sample(x, size = 30, replace = TRUE, prob = p)
mean(xi)
```

```
## [1] 2.5
```

For the solution d) to e),

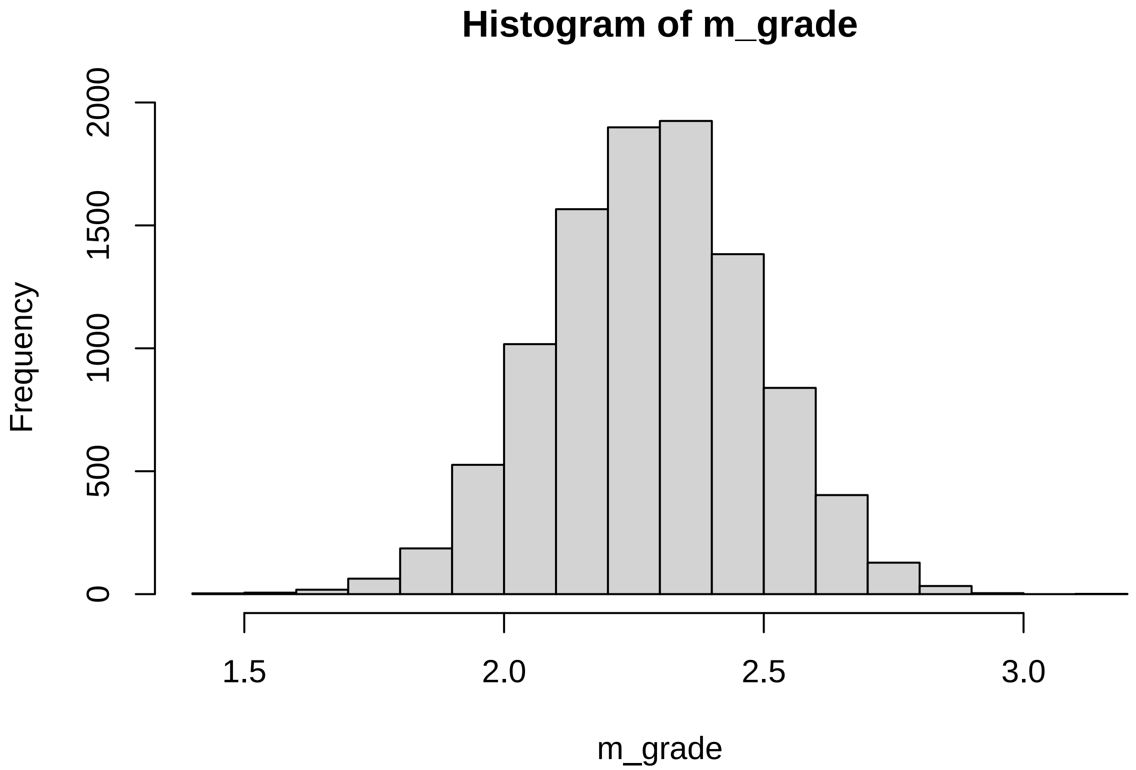
```r
N <- 10000
n <- 30
p <- c(0.1, 0.4, 0.3, 0.1, 0.1)
x <- 4:0

m_grade <- integer(N)
set.seed(20221225)
for (i in 1:N) {
    s <- sample(x, size = n, replace = TRUE, prob = p)
    m_grade[i] <- mean(s)
}

mean(m_grade)
```

```
## [1] 2.299533
```

```r
hist(m_grade)
```



We generate 10,000 independent samples of grade scores with a size of 30 and calculate sample means for each sample. Therefore we obtain 10,000 sample means and their distribution is approximately normal. Furthermore, the empirical mean of sample means is almost close to the theoretical expectation of grade scores with the given distribution.

4. Describe the relationship between Pearson's correlation coefficients and the regression coefficient (slope) of univariate regression analysis. You're definitely able to refer to the Internet or other textbooks but you have to give descriptions or formulas in your own words and cite appropriately.

Answer

The simple (univariate) regression model for n observation can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where $E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$ and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. To estimate unknown parameter β_0 and β_1 that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, the ***least squares***[1] approach can help find the solution. The solution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is

[1] Please check it

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Recall that the Pearson's correlation coefficients ($\hat{\rho}_{xy}$) between two variables x and y is

$$\hat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

The slope of the regression line $\hat{\beta}_1$ can be rewritten as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \frac{\sqrt{(y_i - \bar{y})^2/(n-1)}}{\sqrt{(x_i - \bar{x})^2/(n-1)}} = \hat{\rho}_{xy} \frac{s_y}{s_x}$$

where s_x and s_y are standard deviation of x and y, respectively.

When we developed the regression model, we need a statistical measurement to examine how the explanatory variable x is well predictable to the response variable. In this context, the coefficient of determination R^2 plays an important role to assess the regression model performance. R^2 is defined as the proportion of the variance explained by the independent variables, relative to the total variance in the data. This can be quantified as the ratio of explaned sum of squares to total sum of squares.

$$R^2 = \frac{SSR}{SST}, \quad SSR = \sum_{i=1}^n (\hat{y}_i - y_i)^2, \quad SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

Using the solution of the least squares for regression coefficients,

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \hat{\beta}_1^2 \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2}{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{s_x^2}{s_y^2} = \left(\frac{s_x}{s_y}\hat{\beta}_1\right)^2 \end{aligned}$$

Using the relationship between the correlation coefficient and the regression slope estimate,

$$R^2 = \left(\frac{s_x}{s_y}\hat{\beta}_1\right)^2 = \hat{\rho}_{xy}^2$$

5. The attached data file(`P5-passive-avoidance.xlsx`) represents the results of the passive-avoidance test to evaluate the protective effects of the MC treatment on scopolamine-induced memory impairments in mice. The experiment considers four levels of the treatment; `Control`, `Scopolamine` (memory impairment), `MC-420` (scopolamine impairment + MC 420 mg/kg), and `MC-840` (scopolamine impairment + MC 840 mg/kg). The values in the data represent the ratio of retention and acquisition (retention/acquisition * 100 (%)). Perform your own data analysis **using R** (***Hint***: for the analysis, you may need to transform the original data).

Answer

```
## read dataset
library(tidyverse)
library(readxl)

pa_dat <- read_xlsx("final-exam/P5-passive-avoidance.xlsx")
pa_dat2 <- pa_dat %>%
    pivot_longer(cols = -case_num, names_to = "treatment") %>%
    drop_na %>%
    mutate(treatment = factor(treatment, levels = c("Control",
        "Scopolamine", "MC-420", "MC-840")))

## Check data
head(pa_dat2)
```

```
## # A tibble: 6 x 3
##   case_num treatment   value
##      <dbl> <fct>       <dbl>
## 1        1 Control     140
## 2        1 Scopolamine  31.2
## 3        1 MC-420      283.
## 4        1 MC-840      227.
## 5        2 Control     240
## 6        2 Scopolamine 242.
```

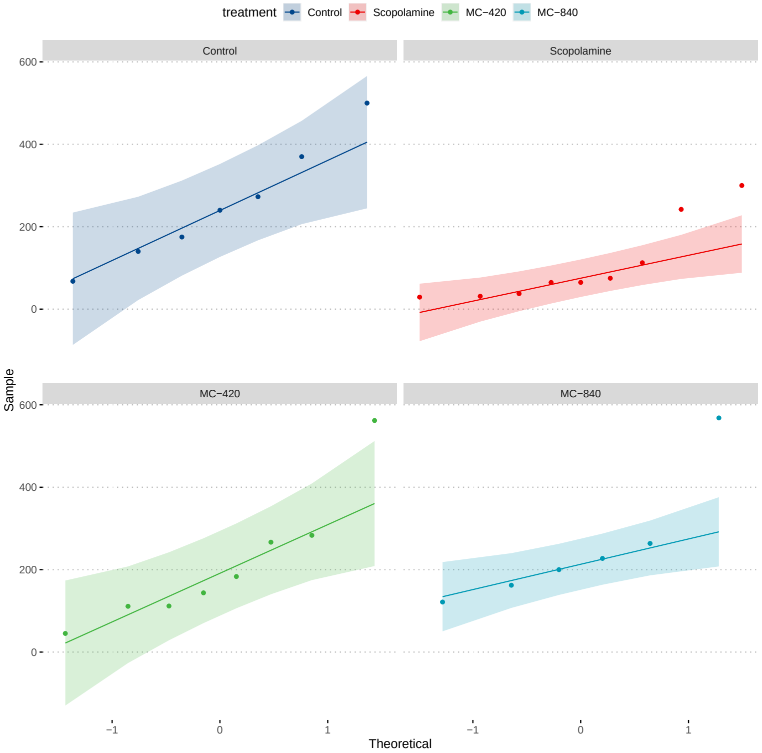
Descriptive Statistics

```
pa_dat2 %>%
    group_by(treatment) %>%
    summarise(N = n(), Mean = mean(value), SD = sd(value), Min = min(value),
        Median = median(value), IQR = IQR(value), Max = max(value))
```

```
## # A tibble: 4 x 8
##   treatment       N  Mean    SD   Min Median   IQR   Max
##   <fct>       <int> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Control         7  252. 146.   67.6  240   164.   500
## 2 Scopolamine     9  106.  97.9  29.2   64.9  75    300
## 3 MC-420          8  213. 162.   45.2  164.  159.   562.
## 4 MC-840          6  257. 160.  121.   214.   82.9  568.
```

Check Normality: Q-Q plot

```
library(broom)
library(ggpubr)  # ggplot for the publication
ggqqplot(pa_dat2, x = "value", color = "treatment", group = "treatment",
    facet.by = "treatment", palette = "lancet", ggtheme = theme_pubclean(base_size = 12))
```



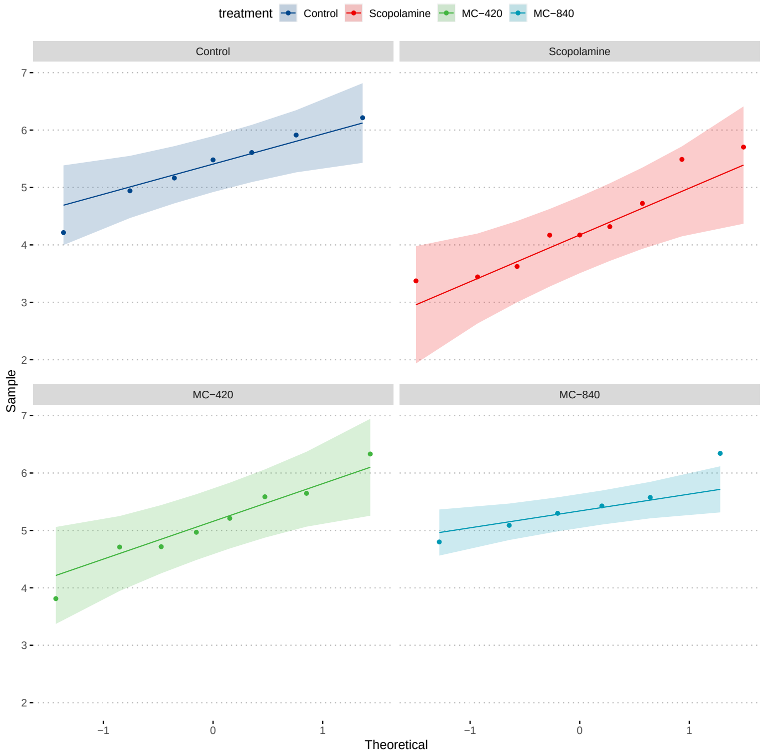
Check Normality: Shapiro-Wilk's Test

```
# Shapiro-Wilk's Normality test for each group
pa_dat2 %>%
    group_by(treatment) %>%
    nest %>%
    mutate(norm_test = map(data, ~shapiro.test(.x$value))) %>%
    mutate(norm_test = map(norm_test, tidy)) %>%
    select(norm_test) %>%
    unnest(norm_test)
```

```
## # A tibble: 4 x 4
## # Groups:   treatment [4]
##   treatment   statistic p.value method
##   <fct>           <dbl>   <dbl> <chr>
## 1 Control         0.969  0.894  Shapiro-Wilk normality test
## 2 Scopolamine     0.774  0.0101 Shapiro-Wilk normality test
## 3 MC-420          0.854  0.105  Shapiro-Wilk normality test
## 4 MC-840          0.788  0.0461 Shapiro-Wilk normality test
```

The data of Scopolamine and MC-840 do not satisfy the normality assumption. Consider the log transformation and check the normality again.

```
pa_dat2 <- pa_dat2 %>%
    mutate(log_value = log(value))
ggqqplot(pa_dat2, x = "log_value", color = "treatment", group = "treatment",
    facet.by = "treatment", palette = "lancet", ggtheme = theme_pubclean(base_size = 12))
```



```
# Shapiro-Wilk's Normality test for each group
pa_dat2 %>%
    group_by(treatment) %>%
    nest %>%
    mutate(norm_test = map(data, ~shapiro.test(.x$log_value))) %>%
    mutate(norm_test = map(norm_test, tidy)) %>%
    select(norm_test) %>%
    unnest(norm_test)
```

```
## # A tibble: 4 x 4
## # Groups:   treatment [4]
##   treatment   statistic p.value method
##   <fct>           <dbl>   <dbl> <chr>
## 1 Control         0.974   0.928 Shapiro-Wilk normality test
## 2 Scopolamine     0.912   0.330 Shapiro-Wilk normality test
## 3 MC-420          0.978   0.953 Shapiro-Wilk normality test
## 4 MC-840          0.938   0.646 Shapiro-Wilk normality test
```

6. A total of 160 men of different ethnic backgrounds were included in a cross-sectional study of factors related to blood clotting. We compared mean platelet levels in the four groups using a one-way ANOVA. It was reasonable to assume Normality and constant variance.

Group	N (%)	Mean($\times 10^9$)	SD ($\times 10^9$)
Caucasian	100 (62.5)	268.1	77.08
Afro-Caribbean	18 (11.3)	254.3	67.50
Mediterranean	23 (14.4)	281.1	71.09
Other	19 (11.9)	273.3	63.42

Fill the following ANOVA table

Source	SS	DF	MS	F-ratio	P-value
Between Group	9333.0	(1)	(3)	(5)	0.423
Within Group	966108.0	(2)	(4)		
Total	975441.0				

7. Calculate the sample size for the following questions.

a) An active-controlled randomized trial proposes to assess the effectiveness of Herbal medicine A in reducing pain. A previous study showed that Herbal medicine A can reduce pain score by 5 points from baseline to week 24 with a standard deviation (σ) of 1.195. A clinically important difference of 0.5 as compared to active drug is considered to be acceptable. (Level of significance = 5%, Power = 80%, Type of test =two-sided)

b) A placebo-controlled randomized trial proposes to assess the effectiveness of Drug A in curing infants suffering from sepsis. A previous study showed that proportion of subjects cured by Drug A is 50% and a clinically important difference of 16% as compared to placebo is acceptable. (Level of significance = 5%, Power = 80%, Type of test =two-sided)