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

Due Date: Dec 19 2023 (13:00) to Dec 23 2023 (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 I error is the probability that you reject the null hypothesis when it is true.

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

   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.

   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.

   e) In test for heterogeneity of meta-analysis, if Higgins $I^2 > 75\%$, studies are regarded homogeneous and the fixed effect model of meta-analysis can generally be used.

Answer: e

2. Suppose all students who officially enrolled in Medical Statistics are playing a game called “The Prisoners and Warder”. Ari, Umar, and Happy played roles as prisoners and they all had been sentenced to death (I’m really sorry to give you guys such a role!!). And Joyce has a role as a warder. Joyce has selected one of the prisoners randomly to be pardoned. Joyce has already received the name which one is pardoned from the governor, but Joyce is not allowed to tell to them. Umar asks to Joyce: “If Ari is going to be pardoned, give me the name of Happy. If Happy is pardoned, then give me Ari’s name. If I’m the one to be pardoned, just flip a coin to decide whether to name Ari or Happy.” Joyce reckons for a while and decides to tell Umar that Happy to be executed. Umar is so pleased because he believes that his probability of surviving has gone up from 1/3 to 1/2, as it is now between him and Ari to be pardoned. Umar secretly whispered to Ari to tell the brand new information. When Ari has heard this news, he reasons that the chance of Umar to be pardoned is not changed at 1/3, but he is pleased since Ari’s own chance has gone up to 2/3. Which prisoner is correct? Please give a detailed explanation of your reasoning.

ANSWER

Let’s define the events that Umar, Happy, and Ari become pardoned before hearing from the warder (Joyce) are A, B, and C, respectively. Then $P(A) = P(B) = P(C) = 1/3$. Let b be the event that the warder tells Umar (A) that Happy (B) is to be executed. Using Bayes’ theorem,

$$\begin{aligned} P(A|b) &= \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)} \\ P(b|A) &= \frac{P(b \cap A)}{P(A)} = \frac{1/2 \cdot 1/3}{1/3} = \frac{1}{2} \\ P(b|B) &= \frac{P(b \cap B)}{P(B)} = \frac{0}{1/3} = 0 \\ P(b|C) &= \frac{P(b \cap C)}{P(C)} = \frac{1/3}{1/3} = 1 \end{aligned}$$

Plugging the above to the equation for $P(A|b)$, then

$$P(A|b) = \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} = \frac{1}{3}$$

Similarly,

$$\begin{aligned} P(C|b) &= \frac{P(b|C)P(C)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)} \\ &= \frac{1 \cdot 1/3}{1/2} = \frac{2}{3} \end{aligned}$$

Therefore, Joyce did not provide any information on whether Umar to be pardoned or not. However, Ari’s chance to be pardoned becomes double after hearing Happy (B) is not pardoned.

3. Solve following problems.

   a. Let X be a random variable having binomial distribution with parameters $n = 25$ and $p = 0.2$. Evaluate $P(X < \mu_x - 2\sigma_X)$.

   b. If X is a random variable with Poisson distribution satisfying $P(X = 0) = P(X = 1)$, what is $E(X)$?

   c. If X is normally distributed mean 2 and variance 1, find $P(|X - 2| < 1)$.

   d. Suppose X has a binomial distribution with parameters n and p. For what p is $\text{var}(X)$ maximized if we assumed n is fixed?

ANSWER

a. $X \sim \text{binomial}(n = 25, p = 0.2)$, then $\mu_X = np = 5$, $\sigma_X^2 = np(1 - p) = 4$. $P(X < 5 - 4 = 1) = P(X = 0) = (0.8)^{25}$ (There was mistypo in the problem a. All students will get the proper score.)

b. Since $X \sim \text{Pois}(\lambda)$, then

$$f_X(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

$P(X = 0) = \exp(-\lambda)$, $P(X = 1) = \lambda \exp(-\lambda)$, $\exp(-\lambda) = \lambda \exp(-\lambda)$, $\therefore \lambda = 1$ (There was mistypo in the problem b. All students will get the proper score.)

c. $X \sim N(2, 1)$, $P(|X - 2| < 1 = P(1 < X < 3) = 0.6826$

d. X binom(n, p), $\sigma_X^2 = np(1 - p)$, to find the maximum for p,

$$\frac{d}{dp}\sigma_X^2 = n - 2pn = 0, \quad p = \frac{1}{2}$$

When $p = 1/2$, we can obtain the maximum of σ_X^2.

4. Let X be the maximum usage time (hour) of an A-manufactured cellular phone after a full charge. Assume that X is normally distributed with mean 75 (hours) and variance σ^2. If a purchaser of such cellular phones requires that at least 90 percent of the cellular phones are avaiable to use exceeding 60 hours, what is the largest value of σ can be and still have the purchaser satisfied?

ANSWER

Since $X \sim N(75, \sigma^2)$ and $P(X > 60) \geq 0.9$, Let $\Phi(X = x)$ be the cumulative distribution function of standardized normal distribution.

$$\begin{aligned} \Phi\left(Z = \frac{X - 75}{\sigma}\right) &>= 0.9 \\ \Phi\left(\frac{60 - 75}{\sigma}\right) &>= 0.9 \\ 1 - \Phi\left(\frac{60 - 75}{\sigma}\right) &>= 0.9 \end{aligned}$$

Therefore,

$$\frac{60 - 75}{\sigma} <= -1.282, \quad \sigma = 15/1.282 = 11.70$$

5. Solve the following problems:

   a. Let X_1, X_2 and X_3 be uncorrelated random variables with common variance σ^2. Find the correlation coefficient between $X_1 + X_2$ and $X_2 + X_3$

   b. Let X_1 and X_2 be uncorrelated random variables. Find the correlation coefficient between $X_1 + X_2$ and $X_2 - X_1$ in terms of $\text{Var}(X_1)$ and $\text{Var}(X_2)$.

   c. Let X_1, X_2, and X_3 be independently distributed random variables with common mean μ and common variance σ^2. Find the correlation coefficient between $X_2 - X_1$ and $X_3 - X_1$.

a. $\text{Cor}(Y = X_1 + X_2, Z = X_2 + X_3)$

$$\begin{aligned}
\rho_{YZ} = \frac{\text{Cov}(Y,Z)}{\sigma_Y \sigma_Z} &= \frac{E[(X_1+X_2)(X_2+X_3)] - E[X_1+X_2]E[X_2+X_3]}{\sqrt{\text{Var}[X_1+X_2]\text{Var}[X_2+X_3]}} \\
&= \frac{E[X_1X_2 + X_2^2 + X_1X_3 + X_2X_3] - \left\{E[X_1]E[X_2] + E[X_1]E[X_3] + [E[X_2]]^2 + E[X_2]E[X_3]\right\}}{\sqrt{2\sigma^2 \cdot 2\sigma^2}} \\
&= \frac{\text{Cov}(X_1,X_2) + \text{Cov}(X_1,X_3) + \text{Cov}(X_2,X_3) + E[X_2^2] - [E[X_2]]^2}{2\sigma^2} \\
&= \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}
\end{aligned}$$

b. $\text{Cor}(X_1 + X_2, X_2 - X_1)$

$$\begin{aligned}
\rho_{X_1+X2,X_2-X_1} &= \frac{E[(X_1+X_2)(X_2-X_1)] - E[X_1+X_2]E[X_2-X_1]}{\sqrt{\text{Var}(X_1+X_2)\text{Var}(X_2-X_1)}} \\
&= \frac{E[X_2^2 - X_1^2] - \{E[X_2]^2 - E[X_1]^2\}}{\sqrt{(\text{Var}[X_1] + \text{Var}[X_2])^2}} \\
&= \frac{\text{Var}[X_2] - \text{Var}[X_1]}{\text{Var}[X_1] + \text{Var}[X_2]}
\end{aligned}$$

c. $\text{Cor}(X_2 - X_1, X_3 - X_1)$

$$\begin{aligned}
\rho_{X_2-X_1,X_3-X_1} &= \frac{E[(X_2-X_1)(X_3-X_1)] - E[X_2-X_1]E[X_3-X_1]}{\sqrt{\text{Var}(X_2-X_1)\text{Var}(X_3-X_1)}} \\
&= \frac{E[X_2X_3 - X_1X_2 - X_1X_3 + X_1^2] - \{E(X_2)E(X_3) - E(X_1)E(X_2) - E(X_1)E(X_3) + [E(X_1)^2]\}}{\sqrt{2\sigma^2 2\sigma^2}} \\
&= \frac{\text{Cor}(X_2,X_3) - \text{Cor}(X_1,X_2) - \text{Cor}(X_1,X_3) + E(X_1^2) - [E(X_1)]^2}{2\sigma^2} \\
&= \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}
\end{aligned}$$

6. When you start R with Rstudio, there is an example dataset named with `mtcars`. The `mtcars` dataset was extracted from the 1974 Motor Trend US magazine, and comprise fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 74 models). The detailed description of variables in `mtcars` dataset can be checked by typing `help(mtcars)` in the prompt of R console window. Write R scipts and confirm the results for the following questions.

   a. Extract `mpg` and `disp` variables from `mtcars` dataset and restore it in an object `x` and `y`, respectively.
   b. Calculate mean, standard deviation, coefficient of variation, minimum, maximum, median, 25th and 75th quantiles, and interquartile range of `x` and `y`.
   c. Make scatterplot of `x` and `y` and interpret in terms of the correlation coefficient between `x` and `y`.
   d. Assume that `x` is the population of a mile per gallon of all automobiles of US from 1973 to 1974. Suppose a sample of size 2 automobiles are drawn from the population with replacement and calculate sample mean. Then repeat the same procedure 10,000 times (Hint: check the function `sample()`).
      - Make histogram of 10,000 sample means.
      - Calculate the mean and standard deviation of 10,000 sample means.
      - Compare the above results to the population in terms of mean and standard deviation: is the mean of 10,000 sample mean is approximate to the population mean? In what proportion did the standard deviation of the sample mean decrease compared to the standard deviation of the population?

```r
x <- mtcars$mpg
y <- mtcars$disp  # a
# b
summ_vec <- function(x, ...) {
    m <- mean(x, ...)
    s <- sd(x, ...)  # mean and sd
    cv <- s/m
    iqr <- IQR(x, ...)
    out <- c(mean = m, sd = s, cv = cv, min = min(x, ...), q25 = quantile(x, 0.25,
        ...), median = median(x, ...), q75 = quantile(x, 0.75, ...), max = max(x,
        ...), iqr = iqr)
    out
}
summ_vec(x)
```

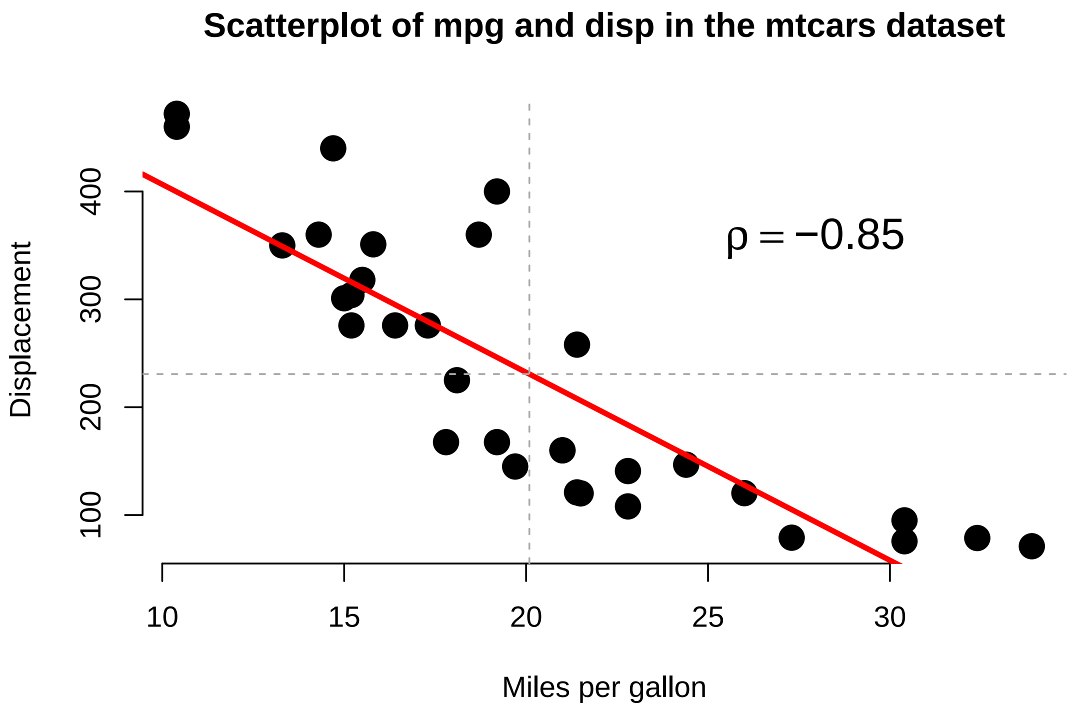
```
      mean         sd         cv        min    q25.25%     median    q75.75%
20.0906250  6.0269481  0.2999881 10.4000000 15.4250000 19.2000000 22.8000000
```

```
       max        iqr 
33.9000000  7.3750000 
```

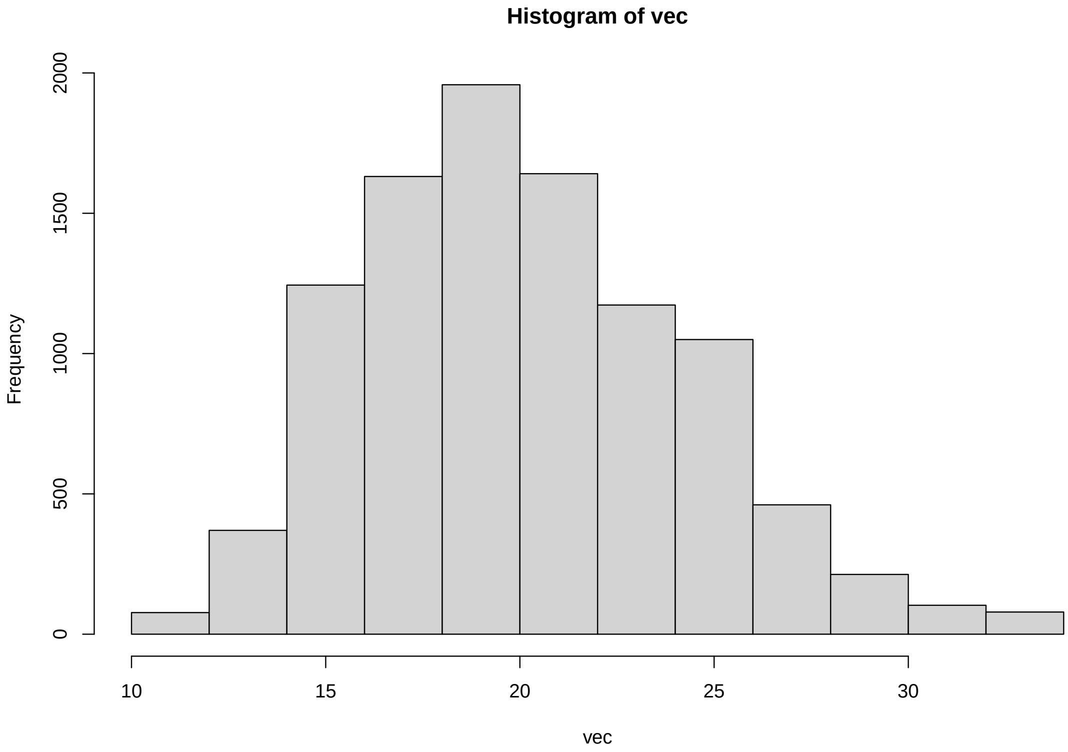
```
summ_vec(y)
```

```
       mean          sd          cv         min     q25.25%      median 
230.7218750 123.9386938   0.5371779  71.1000000 120.8250000 196.3000000 
    q75.75%         max         iqr 
326.0000000 472.0000000 205.1750000 
```

```
# c
plot(x, y, type = "n", bty = "n", main = "Scatterplot of mpg and disp in the mtcars dataset",
    xlab = "Miles per gallon", ylab = "Displacement")
points(x, y, pch = 16, cex = 2)
abline(lm(y ~ x), lty = 1, lwd = 3, col = "red")
abline(h = mean(y), lty = 2, col = "darkgray")
abline(v = mean(x), lty = 2, col = "darkgray")
text(25, 350, bquote(paste(rho == .(sprintf("%.2f", cor(x, y))))), adj = 0, cex = 1.5,
    pos = 4)
```



```
# d-1
vec <- numeric(10000)
set.seed(20211218)
for (i in 1:10000) {
    vec[i] <- sample(x, 2, replace = TRUE) |>
        mean()
}
hist(vec)
```



```
# d-2
n <- 10000
mean(vec)
```

```
[1] 20.11549
```

```
sd(vec) * ((n - 1)/n)
```

```
[1] 4.183035
```

```
# d-3
sd(x)
```

```
[1] 6.026948
```

```
sd(x)/sqrt(2)
```

```
[1] 4.261696
```

```
# The standard deviation of 10,000 sample means with the sample size of 2
# decreased in the proportion of sqrt(2).
```

7. 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($\times 10^9$)	Standard deviation ($\times 10^9$)
Caucasian	100 (62.5)	262.1	70.08
Afro-Caribbean	18 (11.3)	245.3	69.50
Mediterranean	23 (14.4)	254.1	69.09
Other	19 (11.9)	254.3	71.42

Fill the following ANOVA table

Source	Sum of square	df	Mean square	F-ratio	P-value
Between Group	15369.0	(1)	(3)	0.515	0.325
Within Group	1552824.0	(2)			
Total	1568193.0				

Note: df, degree of freedom

ANSWER

(1) 3 (2) 156 (3) 5123

8. Calculate the sample size for the following questions (Hint. Test for equality, two sample parallel design).

a) An active-controlled randomized trial proposes to assess the effectiveness of Drug A in reducing pain. A previous study showed that Drug A can reduce pain score by 5 points from baseline to week 24 with a standard deviation (σ) of 1.5. A clinically important difference of 0.4 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 60% and a clinically important difference of 15% as compared to placebo is acceptable. (Level of significance = 5%, Power = 80%, Type of test =two-sided)

ANSWER

a. $\delta = 0.4$, $\sigma = 1.5$, $\alpha = 0.05$, $1 - \beta = 0.8$

$$n_1 = n_2 = \frac{2(Z_{1-\alpha/2} + Z_\beta)^2\sigma^2}{\delta^2} = \frac{2(1.96 + 0.84)^2(1.5)^2}{0.4^2} = 220.75 \approx 221$$

The estimated sample size in this study is 221 per group.

b. $\theta_1 = 0.6$ and $\theta_2 = 0.45$ since $\delta = 0.15$

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[\theta_1(1-\theta_1) + \theta_2(1-\theta_2)]}{(\theta_1 - \theta_2)^2} = \frac{(1.96 + 0.84)^2[0.6(1-0.6) + 0.45(1-0.45)]}{(0.15)^2} = 170.06 \approx 171$$

The estimated sample size in this study is 171 per group (note that ceiling is generally applied when calculating the sample size.