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

Solutions

Name: ________________________________

Notice

- Please DO SOLVE ANSWERS BY YOURSELVES!!
- You can use materials from other textbooks, lecture notes, and websites but you have to cite materials in every answer.
- Write down your answers to each question in this document file (make a space for answers below the question)

Questions

1. Which one of the following statements is False?

   a. In (prospective) cohort study, subjects are enrolled or grouped on the basis of their exposure, then are followed to document occurrence of disease. **TRUE**
   b. The probability of a type 1 error is the probability that you reject the null hypothesis. **TRUE**
   c. In test for heteregeneity of meta-analysis, if Higgins I^2<=25%, studies are regarded homogeneous and the fixed effect model of meta-analysis can generally be used. **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. 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. **FALSE**

2. Suppose all students who officially enrolled in Medical Statistics are playing a game called "The Prisoners and Warder". Joel, Minh, and Roggers played roles as prisoners and they all had been sentenced to death (I'm really sorry to give you guys such a role!!). And Noel has a role as a warder. Noel has selected one of the prisoners randomly to be pardoned. Noel has already received the name which one is pardoned from the governor, but she is not allowed to tell to them. Roggers asks to Noel: "If Minh

is going to be pardoned, give me the name of Joel. If Joel is pardoned, then give me Minh's name. If I'm the one to be pardoned, just flip a coin to decide whether to name Minh or Joel." Noel reckons for a while and decides to tell Roggers that Joel to be executed. Roggers 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 Minh to be pardoned. Roggers secretly whispered to Minh to tell the brand new information. When Minh has heard this news, he reasons that the chance of Roggers to be pardoned is not changed at 1/3, but he is pleased since Minh's own chance has gone up to 2/3. Which prisoner is correct? Please give a detailed explanation of your reasoning.

ANSWER

Exactly an identical structure to the "Monty Hall Problem". Let's define the events that Joel, Minh, and Roggers become pardoned before hearing from the warden are A, B, and C, respectively. Then $P(A) = P(B) = P(C) = 1/3$. Let b be the event that the warden tells Roggers (A) that Joel (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, Noel did not provide any information on whether Roggers to be pardoned or not. However, Minh's chance to be pardoned becomes double after hearing Joel (B) is not pardoned.

3. Consider a sample of size 2 drawn without replacement from an urn containing three balls, numbered 1, 2, and 3. Let X be the number on the first ball drawn and Y the larger of the two numbers drawn

a. Find the joint discrete distribution of X and Y

b. Find the marginal distribution of Y

c. Find $P(X = 1|Y = 3)$

d. Find the $\text{Cov}(X, Y)$

ANSWER

a. Joint Distribution of X abd Y

	$X = 1$	$X = 2$	$X = 3$	$f_Y(y)$
$Y = 1$	0	0	0	0
$Y = 2$	1/6	1/6	0	1/3
$Y = 3$	1/6	1/6	1/3	2/3
$f_X(x)$	1/3	1/3	1/3	1

b. Marginal distribution of Y

	$Y = 1$	$Y = 2$	$Y = 3$	$\sum$
$f_Y(y)$	0	1/3	2/3	1

c. $P(X = 1|Y = 3)$

$$P(X = 1|Y = 3) = \frac{P(X = 1, X = 3)}{P(Y = 3)} = \frac{1/6}{2/3} = \frac{1}{4}$$

d.

$$E(XY) = 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 6\frac{1}{6} + 9\frac{1}{3} = \frac{11}{2} = 5.5$$
$$E(X) = 1\frac{1}{3} + 2\frac{1}{3} + 3\frac{1}{3} = 2$$
$$E(Y) = 2\frac{1}{3} + 3\frac{2}{3} = \frac{8}{3}$$

$\therefore \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 11/2 - 16/3 = 1/6$

4. 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 and X_1.

ANSWER

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

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

5. 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, 25^{th} and 75^{th} 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?

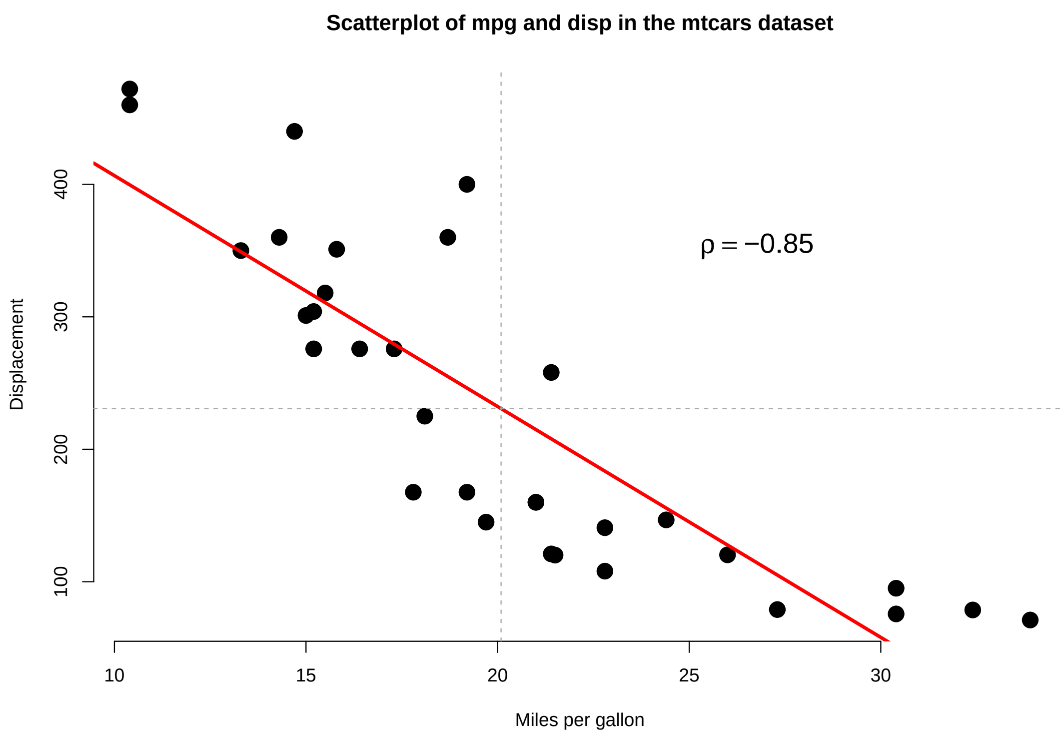
```r
x <- mtcars$mpg
y <- mtcars$disp  # a
# b
summ_vec <- function(x, ...) {
    m <- mean(x, ...)
    s <- sd(x, ...)  # mean and sd
    cv <- s/m
    iqr <- IQR(x, ...)
    out <- c(mean = m, sd = s, cv = cv, min = min(x, ...), q25 = quantile(x, 0.25,
        ...), median = median(x, ...), q75 = quantile(x,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

```r
summ_vec(y)
```

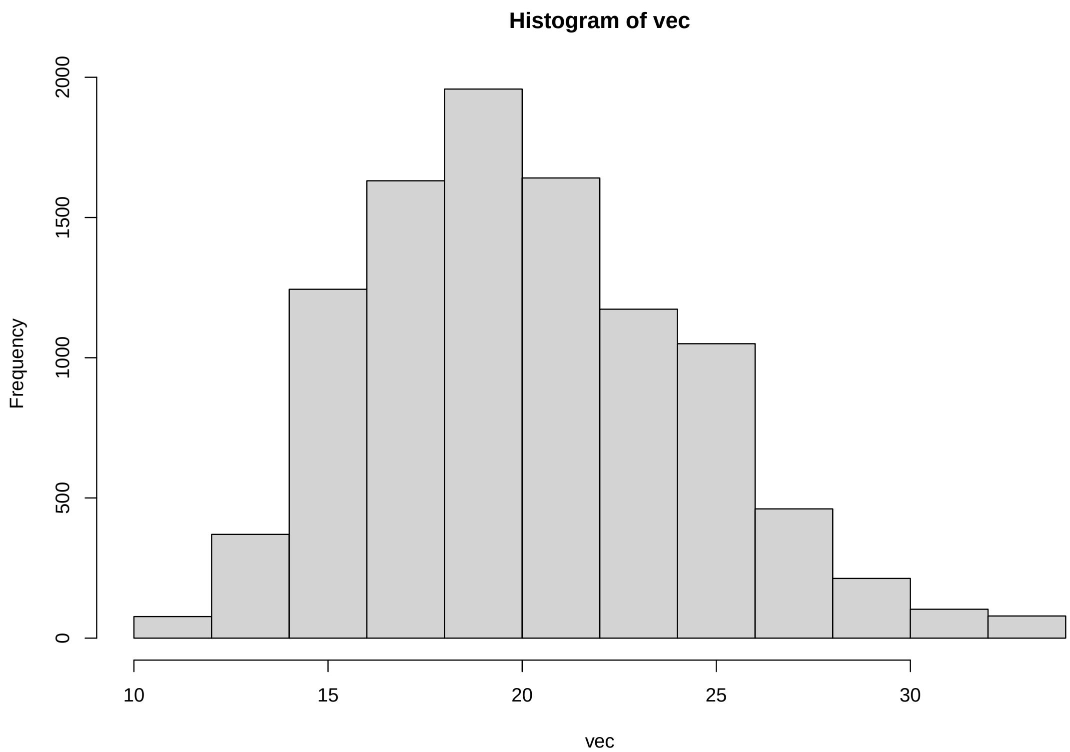
```
##        mean          sd          cv         min    q25.25%      median 
## 230.7218750 123.9386938   0.5371779  71.1000000 120.8250000 196.3000000 
##     q75.75%         max         iqr 
## 326.0000000 472.0000000 205.1750000
```

```r
# 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)
```




```r
# 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
```

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

```
[1] 4.261696
```

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

6. A total of 144 women 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. Fill the following ANOVA table.

Group	N (%)	Mean ($\times 10^9$)	SD ($\times 10^9$)
Caucasian	90 (62.5)	268.1	77.08
Afro-Caribbean	21 (14.6)	254.3	67.50
Mediterranean	19 (13.2)	281.1	71.09
Other	14 (9.7)	273.3	63.42

Fill the following ANOVA table

Source	SS	DF	MS	F-ratio	P-value
Between Group	7500.0	3	2500	(3)	0.670
Within Group	840000.0	(1)	(2)		
Total	847500.0				

ANSWER

- $df_{\mathrm{B}} = g - 1 = 4 - 1 = 3$
- $df_{\mathrm{W}} = 144 - 4 = 140$ (1)
- $MSW = SSW/df_B = 840000/140 = 6000$ (2)
- $F_0 = MSB/MSW = 2500/6000 = 0.417$

7. Calculate the sample size for the following questions

   a. Suppose the response rate of the patient population under study after treatment is expected to be around 55% (i.e., $\theta = 0.55$). At $\alpha = 0.05$, the required sample size for achieving an 80% power ($\beta = 0.2$) correctly detecting a difference between the post-treatment response rate and the reference value say, 35% (i.e., $\theta_0 = 0.35$) is $N =$? (Hint: Test for Equality, One sample design)

b. Suppose a low density lipidproteins (LDLs) is considered of clinically meaningful difference. Assuming that the standard deviation is 15% (i.e., population variance is 0.15), the required sample size of each group to achieve an 80% power ($\beta = 0.2$) at $\alpha = 0.05$ for correctly detecting such difference of $\mu_2 - \mu_1 = 0.07$ change obtained by normal approximation as $N_1, N_2 =$? (Hint: Test for Equality, Two sample parallel Design)

ANSWER

a. Large sample test for proportions (test for equality, one sample design)

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \theta(1-\theta)}{(\theta - \theta_0)^2} = \frac{0.55(1-0.55)(1.96+0.84)^2}{(0.55-0.35)^2} = 48.51 \approx 49$$

b. Comparing two means (two-sample parallel design)

$$\begin{aligned} n &= \frac{2(z_{\alpha/2} + z_\beta)^2}{\eta^2}, \quad \eta = \frac{\mu_2 - \mu_1}{\hat{\sigma}} \\ &= \frac{2(1.96+0.84)^2}{(0.07/0.15)^2} = 72 \end{aligned}$$

The total sample size is $72 \times 2 = 144$