Medical Statistics 2nd Semester Take Home Final Exam

Due Date: Dec 17 2024 (10:00) to Dec 20 2024 (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**.
- Please avoid using LLM based chatbots (i.e. chatGPT, Claude, Gemini, and so on) to make your answers.
- 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 **TRUE**? **ANSWER: 1**

   a) The probability of a type I error is the probability that you fail to reject the null hypothesis when it is false.

   b) In a prospective cohort study, subjects are randomly assigned to groups based on their exposure, and then the occurrence of disease is followed.

   c) When more than 20% of cells have expected frequencies < 5, the chi-square test should always be used, regardless of sample size.

   d) To use the two-sample t-test, it is not necessary for the data to be normally distributed or for the variances to be equal.

   e) In a 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.

2. 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 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 $\text{binom}(n, p)$, $\sigma_X^2 = np(1-p)$, 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.

3. 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 70 (hours) and variance σ^2. If a purchaser of such cellular phones requires that at least 90 percent of the cellular phones are available to use exceeding 60 hours, what is the largest value of σ can be and still have the purchaser satisfied?

ANSWER

Since $X \sim N(70, \sigma^2)$ and $P(X > 60) \geq 0.9$, Let $\Phi(X = x)$ be the cumulative distribution function of standardized normal distribution.

$$\Phi\left(Z = \frac{X-70}{\sigma}\right) >= 0.9$$
$$\Phi\left(\frac{60-70}{\sigma}\right) >= 0.9$$
$$1-\Phi\left(\frac{60-70}{\sigma}\right) >= 0.9$$

Therefore,

$$\frac{60-70}{\sigma} <= -1.282, \quad \sigma = 10/1.282 = 7.801$$

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 $X_3 - X_1$.

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

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

5. 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?

ANSWER

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

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

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, 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?

ANSWER

```
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)
```

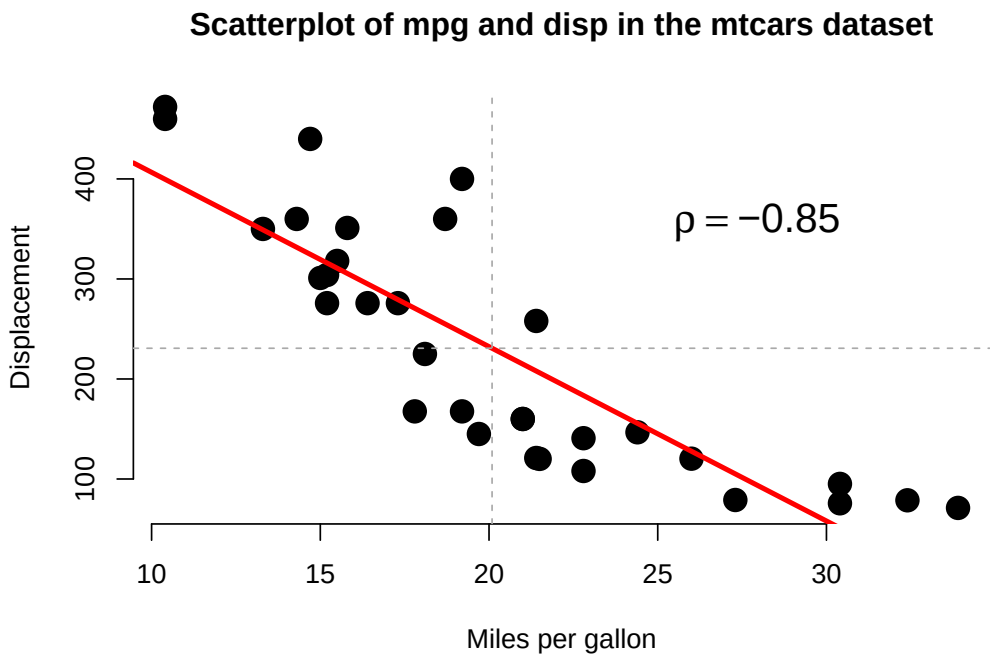
```
      mean         sd         cv        min    q25.25%     median    q75.75%
20.0906250  6.0269481  0.2999881 10.4000000 15.4250000 19.2000000 22.8000000
       max        iqr
33.9000000  7.3750000
```

```
summ_vec(y)
```

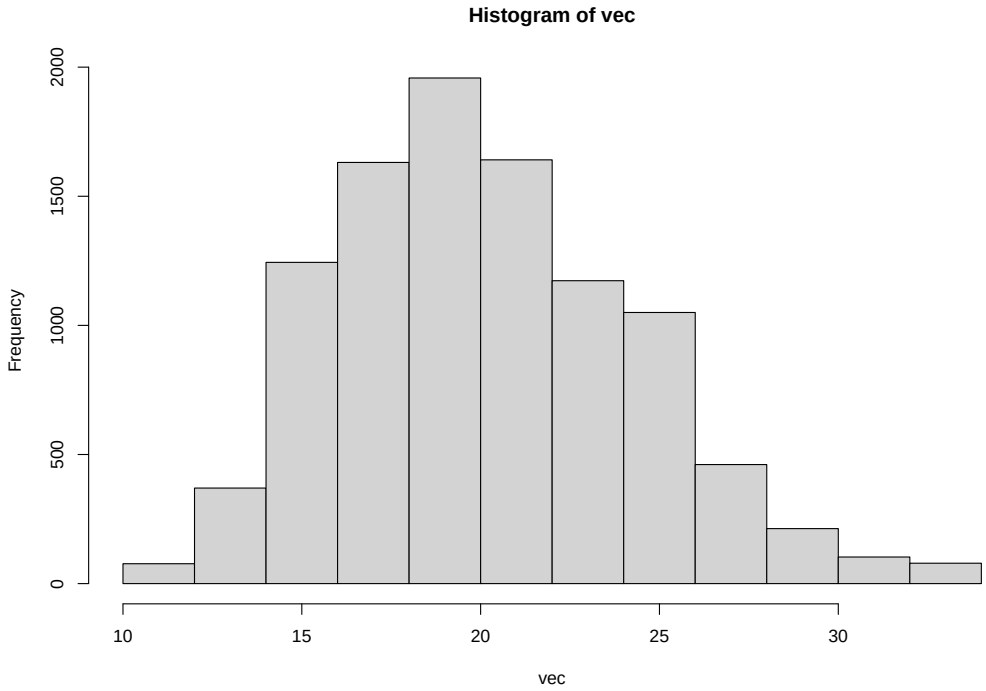
```
       mean          sd          cv         min     q25.25%      median
230.7218750 123.9386938   0.5371779  71.1000000 120.8250000 196.3000000
    q75.75%         max         iqr
```

```
326.0000000 472.0000000 205.1750000
```

```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
```

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

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.

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

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).

ANSWER

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)).

```
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
```

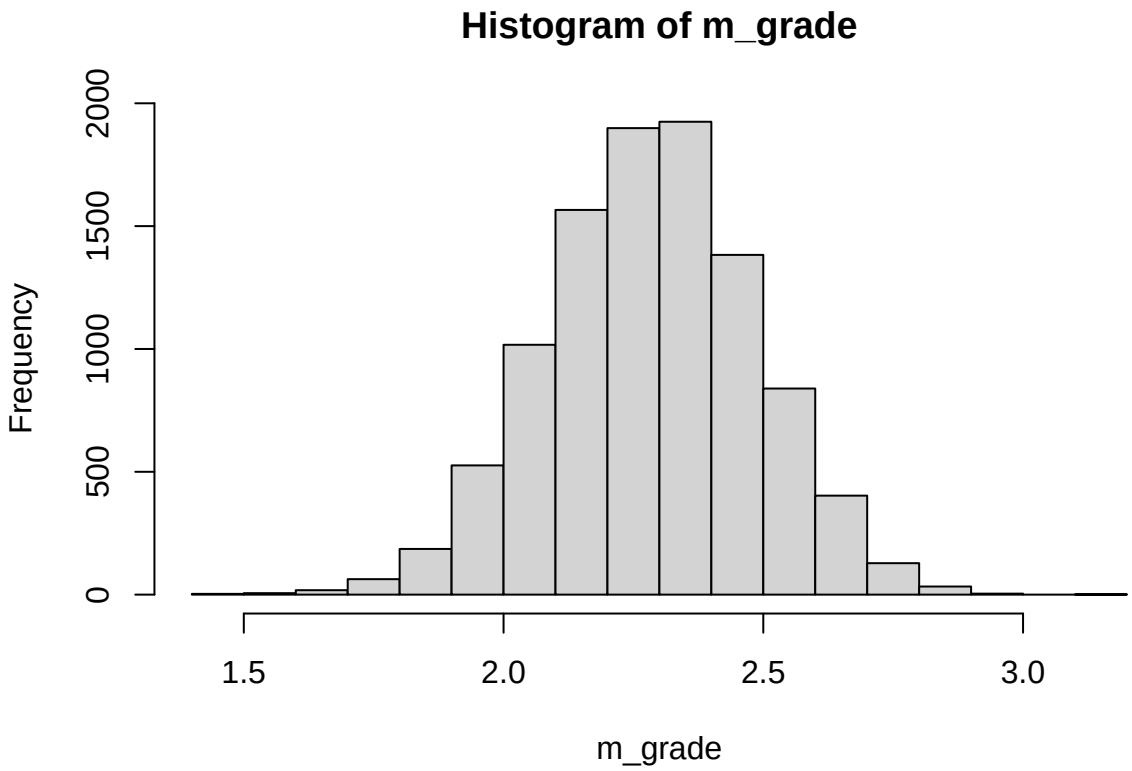
For the solution d) to e),

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

m_grade <- integer(N)
```

```r
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.

8. A pharmaceutical company has developed two types of sleeping pills, namely A and B. Seven healthy adults were randomly selected, and the time it took for them to fall asleep after taking each sleeping pill was measured (assume that time to fall asleep follows a normal distribution with unknown variance

(σ^2 is unknown). The data collected for sleeping pill A and sleeping pill B are as follows:

a) Conduct an appropriate hypothesis test to determine if there is a difference in the effectiveness of sleeping pills A and B. Use a significance level of 0.05.

b) Calculate the 95% confidence interval for the difference in the mean time to fall asleep between sleeping pills A and B.

c) Calcuate the p-value.

ID	A	B
1	15	13
2	19	17
3	16	18
4	14	15
5	17	16
6	14	16
7	17	19

ANSWER

Set null and alternative hypotheses as follows:

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

Calculate the test statistic for the two-sample t-test as follows:

$\bar{X}_A = 16$, $\bar{X}_B = 16.29$, $S_A = 1.83$, $S_B = 1.98$, $n_A = n_B = 7$

Pooled Variance:

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} = 3.06$$

Test statistic:

$$t = \frac{\bar{X}_A - \bar{X}_B}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = -0.28$$

Since the degrees of freedom is $n_A + n_B - 2 = 12$, the critical value for the two-tailed test at $\alpha = 0.05$ is $t_{\alpha/2,12} = 2.179$. Since $-0.28 > -2.179$, we fail to reject the null hypothesis.

The 95% confidence interval for the difference in the mean time to fall asleep between sleeping pills A and B is calculated as follows:

$$\bar{X}_A - \bar{X}_B \pm t_{\alpha/2,12} \cdot S_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

Therefore, the 95% confidence interval is $(-2.50, 1.93)$

The p-value for the two-sided test is calculated as $p = 0.79$

CHECK

```
## Manual Calculation
A = c(15, 19, 16, 14, 17, 14, 17)
B = c(13, 17, 18, 15, 16, 16, 19)


ma = mean(A)
mb = mean(B)
sa = sd(A)
sb = sd(B)
na <- nb <- 7
sp2 <- ((na - 1) * sa^2 + (nb - 1) * sb^2)/(na + nb - 2)
stat <- (ma - mb)/sqrt(sp2 * (1/na + 1/nb))
cval <- qt(0.975, df = na + nb - 2)
ci <- c(ma - mb - cval * sqrt(sp2 * (1/na + 1/nb)), ma - mb +
    cval * sqrt(sp2 * (1/na + 1/nb)))
pval <- 2 * pt(-abs(stat), df = na + nb - 2)
stat
```

```
## [1] -0.2809757
```

```
cval
```

```
## [1] 2.178813
```

```
ci
```

```
## [1] -2.501272  1.929843
```

```
pval
```

```
## [1] 0.7835149
```

```r
## Using t.test
t.test(A, B, var.equal = TRUE, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  A and B
## t = -0.28098, df = 12, p-value = 0.7835
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.501272  1.929843
## sample estimates:
## mean of x mean of y
##  16.00000  16.28571
```

9. 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

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

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

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

[1] Please check it

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 explained 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,

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

10. A total of 180 men of different ethnic backgrounds were included in a cross-sectional study investigating factors related to blood clotting. We compared the mean platelet levels across 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	110 (61.1)	265.0	72.15
Afro-Caribbean	22 (12.2)	248.4	68.30
Mediterranean	28 (15.6)	259.2	70.42
Other	20 (11.1)	257.8	69.12

Fill the following ANOVA table

Source	SS	DF	MS	F-ratio	P-value
Between Group	18540.0	(1) 3	(3) 6180	(5) 0.702	0.5520
Within Group	1549120.0	(2) 176	(4) 8801.8		
Total	1567660.0				

11. Calculate the sample size for the following questions (*Hint*: Test for Equality).

a) An active-controlled randomized trial proposes to assess the effectiveness of Drug A in reducing blood pressure. A previous study showed that Drug A can reduce blood pressure by 10mmHg from baseline to week 12 with a standard deviation (σ) of 2.0 mmHg. A clinically important difference of 1.0 mmHg 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 patients with acute respiratory distress syndrome. A previous study showed that proportion of subjects cured by Drug A is 55% and a clinically important difference of 10% as compared to placebo is acceptable (Level of significance = 5%, Power = 80%, Type of test =two-sided)

ANSWER

a. $\delta = 1.0$, $\sigma = 2.0$, $\alpha = 0.05$, $1 - \beta = 0.8$

$$n_1 = n_2 = \frac{2(Z_{1-\alpha/2} + Z_\beta)^2\sigma^2}{\delta^2} = \frac{2(1.96 + 0.84)^2(1.5)^2}{0.4^2} = 62.79 \approx 63$$

The estimated sample size in this study is 63 per group.

b. $\theta_1 = 0.55$ and $\theta_2 = 0.45$ since $\delta = 0.10$

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[\theta_1(1-\theta_1) + \theta_2(1-\theta_2)]}{(\theta_1 - \theta_2)^2} = \frac{(1.96 + 0.84)^2[0.55(1-0.55) + 0.45(1-0.45)]}{(0.10)^2} = 388.52 \approx 389$$

The estimated sample size in this study is 389 per group (note that ceiling is generally applied when calculating the sample size.