**Homework 5** **Name:** Ivan Wang
(Part 9; 60 pts)

1. The following data were obtained from a 2021 study investigating the effectiveness of vitamin D3 on the death rate of hospitalized COVID patients (*Calcifediol Treatment and Hospital Mortality Due to COVID-19: A Cohort Study*). (13 pts)

| D3 Use | Death | |
|---|---|---|
| | Yes | No |
| Yes | 4 | 75 |
| No | 92 | 366 |

a) Create a matrix containing the count data, and apply appropriate row and column labels.

```
> data <- matrix(c(4, 75, 92, 366), nrow = 2, byrow = TRUE)
> rownames(data) <- c("D3_Yes", "D3_No")
> colnames(data) <- c("Death_Yes", "Death_No")
> data
        Death_Yes Death_No
D3_Yes          4       75
D3_No          92      366
```

b) Researchers are interested in the probability of a COVID fatality. Give estimates of this probability for both treatment groups.

```
> prob_d3_yes <- data[1,1] / (data[1,1] + data[1,2])
> prob_d3_no <- data[2,1] / (data[2,1] + data[2,2])
> prob_d3_yes
[1] 0.05063291
> prob_d3_no
[1] 0.2008734
```

c) Test ($\alpha = 0.01$) whether the difference between the death probabilities is different from 0. Give the hypotheses, test statistic, p-value, and conclusion in context.

```
> prop.test(c(4, 92), c(4+75, 92+366), correct = FALSE)

	2-sample test for equality of proportions without continuity correction

data:  c(4, 92) out of c(4 + 75, 92 + 366)
X-squared = 10.359, df = 1, p-value = 0.001288
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.21093477 -0.08954614
sample estimates:
    prop 1     prop 2 
0.05063291 0.20087336 
```

Ho: p1 - p2 = 0 (no difference in death probabilities between the treatment groups)
Ha: p1 - p2 != 0
X-squared = 10.359, df = 1, p-value = 0.001288
With p-value = 0.001288 < 0.01, we reject Ho. There is significant difference in death probabilities between the treatment groups(with D3 and without).

d) Report a 99% confidence interval for the difference between death probabilities.

99 percent confidence interval: -0.2300063 -0.0704746

e) With this data set, is it appropriate to use the normal-based test, which relies on the Central Limit Theorem?

```
> d3prob1
[1] 4
> d3prob2
[1] 75
> non_d3_prob1
[1] 92
> non_d3_prob2
[1] 366
> 
```

Since one of the groups D3 group deaths has expected count below 5, it's not appropriate to use normal approximation. In addition, Fisher’s Exact Test might be used here because its valid regardless of sample size or expected count.

2. Continue to use the contingency table given in question 1. (10 pts)

a) Give a point estimate of the relative risk of death, and interpret your estimate.

```
> prob_d3_yes
[1] 0.05063291
> prob_d3_no
[1] 0.2008734
>
> RR <- prob_d3_yes / prob_d3_no
> RR
[1] 0.2520638
```

The risk of death for patients treated with D3 is about 25.20% of the risk for untreated patients. This suggest that the treatment using D3 is lower risk.

b) Test whether the relative risk is different from 1 ($\alpha = 0.01$). Give the hypotheses, p-value, and conclusion.

Ho: RR = 1(no difference in risk between patients treated with D3 and patients that did not treat with D3)
Ha: RR != 1 (The risk between the groups differ)

```
$p.value
         NA
two-sided   midp.exact fisher.exact  chi.square
   D3_No            NA           NA          NA
   D3_Yes 0.0004328222 0.0006860971 0.001288234
```

Since all the p-values (from the chi-square test, Fisher's exact test, and the exact midpoint p-value) are less than 0.01, we reject the null hypothesis ($H_0$).There is significant evidence that vitamin D3treatment affects the risk of death.

c) Give a 99% confidence interval for $\pi_1/\pi_2$.

```
                          NA
risk ratio with 99% C.I. estimate    lower    upper
                  D3_No  1.000000       NA       NA
                  D3_Yes 1.188006 1.085631 1.300035
```

3. Again, use the contingency table given in question 1. (10 pts)

a) Give a point estimate of the odds ratio of death, and interpret your estimate.

```
> odd_D3 <- 4/75
> odd_nonD3 <- 92/366
>
> oddRatio <- odd_D3/odd_nonD3
> oddRatio
[1] 0.2121739
>
```

The odds ratio of death for patients treated with D3 are approximately 21.21% of the odds for untreated patients.

b) Test whether the odds ratio is different from 1 ($\alpha = 0.01$). Provide the hypotheses, p-value, and conclusion in context.

H0: OR = 1(no association between D3 use and death)
Ha: OR != 1 (there is a association between D3 use and death)

```
$p.value
         NA
two-sided   midp.exact fisher.exact  chi.square
   D3_No            NA           NA          NA
   D3_Yes 0.0004328222 0.0006860971 0.001288234
```

With all the pvalues(midp.exact, fisher.exact, chi.square) < 0.01, we reject H0. There is a significant association between D3 use and death.

c) Give a 99% confidence interval for the odds ratio.

```
> #c)   Give a 99% confidence interval for the odds ratio.
> oddsratio(data, rev = "rows", conf.level = 0.99)$measure
                           NA
odds ratio with 99% C.I. estimate    lower    upper
                  D3_No  1.000000       NA       NA
                  D3_Yes 4.540674 1.422144 25.13127
>
```

4. A binary happiness metric was recorded on a collection of subjects with a digestive condition. The condition has three levels of severity, and the researcher hypothesizes that the probability of being categorized as happy ($Y = 1$) should decrease as the condition becomes more severe. The data appear below. (15 pts)

| | 0 | 1 |
|---|---|---|
| stage I | 68 | 83 |
| stage II | 24 | 22 |

| | | |
|---|---|---|
| stage III | 69 | 38 |

a) Create a matrix containing the count data, and apply appropriate row and column labels.

```
> data2 <- matrix(c(68, 83, 24, 22, 69, 38), nrow = 3, byrow = TRUE)
> rownames(data2) <- c("stage1", "stage2", "stage3")
> colnames(data2) <- c("0", "1")
> data2
        0  1
stage1 68 83
stage2 24 22
stage3 69 38
```

b) Explain whether the researcher believes disease stage is independent of happiness classification.

The researchers does not believe that happiness is independent of disease.

Stage 1 shows more patients were happy as compared to not happy and stage 3 shows that more people were unhappy compared to happy.

The researchers believe that happiness decrease as the conditions become more severe.

c) Perform the chi-square test for association. Give the hypotheses, test statistic, p-value, and conclusion in context.

H0: Happiness is independent of disease stage (no association)
Ha: Happiness is not independent of disease stage (there is an association)

```
        Pearson's Chi-squared test

data:  data2
X-squared = 9.5259, df = 2, p-value = 0.00854
```

With X-squared = 9.5259, p-value = 0.00854 < 0.05, we reject H0. There is significant evidence that Happiness is not independent of disease stage (there is an association)

d) Present the sample proportions $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$.

```
> row_totals
stage1 stage2 stage3
   151     46    107
> pi_hat_1
   stage1
0.5496689
> pi_hat_2
   stage2
0.4782609
> pi_hat_3
   stage3
0.3551402
```

e) We will now perform a *trend test*, for which the hypotheses are:

$$H_0: \pi_1 = \pi_2 = \pi_3$$
$$H_a: \pi_1 > \pi_2 > \pi_3$$

The CochranArmitageTest(.) function in the "DescTools" package will perform the test. Note that by default, the function will give a two-sided p-value, and we are specifically interested in a decreasing trend. Give the p-value and conclusion in context.

```
> CochranArmitageTest(data2, alternative = c("two.sided", "one.sided"))

        Cochran-Armitage test for trend

data:  data2
Z = 3.0695, dim = 3, p-value = 0.002144
alternative hypothesis: two.sided
```

Z = 3.0695, dim = 3, p-value = 0.002144

With Z = 3.0695, p-value = 0.002144 < 0.05, we reject Ho.
There is decreasing trend for the sample proportions of happy individuals as the severity of the digestive condition increases.

5. (12 pts)

| Marital Status | Education Level | | |
|---|---|---|---|
| | high school | bachelors | graduate |
| never married | 36 | 21 | 15 |

| | | | |
|---|---|---|---|
| married | 36 | 45 | 57 |
| divorced/widowed | 18 | 18 | 15 |

a) Create a matrix containing the count data, and apply appropriate row and column labels.

```
> data3 <- matrix(c(36,21,15,36,45,57,18,18,15), nrow = 3, byrow = TRUE)
> rownames(data3) <- c("never married", "married", "divorced/widowed")
> colnames(data3) <- c("high school", "bachelors", "graduate")
> data3
                 high school bachelors graduate
never married             36        21       15
married                   36        45       57
divorced/widowed          18        18       15
>
```

b) Perform the chi-square test for association using $\alpha = 0.05$. Give the hypotheses, test statistic, p-value, and conclusion in context.

H0: Marital status and education level are independent.
Ha: Marital status and education level are not independent.

```
        Pearson's Chi-squared test

data:  data3
X-squared = 14.464, df = 4, p-value = 0.005953
```

With X-squared = 14.464, p-value = 0.005953 < 0.05, we reject Ho.
There is significant evidence that martial status and education level is not independent.

c) Provide a table of observed cell counts (same as part (a)), a table of expected cell counts under independence, and a table of standardized residuals.

```
> chitest2$expected
                 high school bachelors graduate
never married       24.82759  23.17241       24
married             47.58621  44.41379       46
divorced/widowed    17.58621  16.41379       17
> chitest2$stdres
                 high school  bachelors   graduate
never married      3.2553022 -0.6439911 -2.6440634
married           -3.0226693  0.1555942  2.8935241
divorced/widowed   0.1359033  0.5300289 -0.6623107
```

d) Use the tables from part (c) to write a summary statement about the nature of the association between marital status and education level.

Never Married/Graduate group has large negative residual (-2.6440634), this means that there are fewer people who were never married and have graduate degree than expected under independence.

Never Married/ High School group has a positive residual of 3.2553022, which means that there's a slightly high number of never-married individuals with only high school degree than expected.

Married/ High School group has a negative residual of -3.0226693, which means that there are fewer people that are married with only high school degree than expected.

Married/Graduate group has large positive residual 2.8935241), this means that there are slightly more people who were married and have graduate degree than expected under independence.

Code:

```
#HW 5

"""
1. The following data were obtained from a 2021 study investigating the effectiveness of
vitamin D3 on the death rate of hospitalized COVID patients (Calcifediol Treatment and
Hospital Mortality Due to COVID-19: A Cohort Study). (13 pts)
"""
#a)	Create a matrix containing the count data, and apply appropriate row and column labels.

data <- matrix(c(4, 75, 92, 366), nrow = 2, byrow = TRUE)
rownames(data) <- c("D3_Yes", "D3_No")
colnames(data) <- c("Death_Yes", "Death_No")
data

#b)	Researchers are interested in the probability of a COVID fatality.
#Give estimates of this probability for both treatment groups.

prob_d3_yes <- data[1,1] / (data[1,1] + data[1,2])
prob_d3_no <- data[2,1] / (data[2,1] + data[2,2])
prob_d3_yes
prob_d3_no

#c)	Test (α=0.01) whether the difference between the death probabilities is different from 0.
#Give the hypotheses, test statistic, p-value, and conclusion in context.

"""
Ho: p1 - p2 = 0 (no difference in death probabilities between the treatment groups)
Ha: p1 - p2 != 0
"""

prop.test(c(4, 92), c(4+75, 92+366), correct = FALSE)

#X-squared = 10.359, df = 1, p-value = 0.001288
#With p-value = 0.001288 < 0.05, we reject Ho. There is significant difference in
#death probabilities between the treatment groups(with D3 and without).


#d)	Report a 99% confidence interval for the difference between death probabilities.
prop.test(c(4, 92), c(4+75, 92+366), conf.level = 0.99, correct = FALSE)


#e)	With this data set, is it appropriate to use the normal-based test, which relies
#on the Central Limit Theorem?


#check if np > 5, n(1-p) > 5
p1 <- 4 / 79
np1 <- 1 - (4 / 79)
p2 <- 92 / (92 + 366)
np2 <- 1 - (92 / (92 + 366))

d3prob1 <- p1 * 79
d3prob2 <- np1 * 79
non_d3_prob1 <- p2 * 458
non_d3_prob2 <- np2 * 458

d3prob1
d3prob2
non_d3_prob1
non_d3_prob2


"""
since one of the groups D3 group deaths has expected count below 5,
it's not appropriate to use normal approximation.
In addition, Fisher’s Exact Test might be used here because its valid regardless of sample size or expected count.
"""
```

```
#2. Continue to use the contingency table given in question 1. (10 pts)
install.packages("epitools")
library(epitools)

#a)          Give a point estimate of the relative risk of death, and interpret your estimate.

prob_d3_yes
prob_d3_no

RR <- prob_d3_yes / prob_d3_no
RR

#The risk of death for patients treated with D3 is about 25.20% of the risk for untreated patients
#This suggest that the treatment using D3 is lower risk.


#b)          Test whether the relative risk is different from 1 (α=0.01). Give the hypotheses, p-value, and conclusion.

"""
Ho: RR = 1(no difference in risk between patients treated with D3 and patients that did not treat with D3)
Ha: RR != 1 (The risk between the groups differ)
"""
rr <- riskratio(data, rev = "rows", conf.level = 0.99)
rr
"""
$p.value
       NA
two-sided   midp.exact fisher.exact  chi.square
  D3_No            NA           NA          NA
  D3_Yes 0.0004328222 0.0006860971 0.001288234

Since all the p-values (from the chi-square test, Fisher's exact test, and the exact midpoint p-value)
are less than 0.01, we reject the null hypothesis (H0).There is significant evidence that vitamin D3
treatment affects the risk of death.
"""

#c)          Give a 99% confidence interval for π_1/π_2.


"""
                 NA
risk ratio with 99% C.I. estimate    lower    upper
             D3_No  1.000000       NA       NA
             D3_Yes 1.188006 1.085631 1.300035
"""
rr$measure


#3. Again, use the contingency table given in question 1. (10 pts)

#a)          Give a point estimate of the odds ratio of death, and interpret your estimate.
odd_D3 <- 4/75
odd_nonD3 <- 92/366

oddRatio <- odd_D3/odd_nonD3
oddRatio

"""
The odds ratio of death for patients treated with D3 are approximately 21.21% of the odds for untreated patients.
"""

#b)          Test whether the odds ratio is different from 1 (α=0.01). Provide the hypotheses, p-value, and conclusion in context.
"""
H0: OR = 1(no association between D3 use and death)
Ha: OR != 1 (there is a association between D3 use and death)
"""
oddsratio(data, rev = "rows", conf.level = 0.99)
```

```
"""
$p.value
        NA
two-sided   midp.exact fisher.exact  chi.square
   D3_No           NA           NA          NA
   D3_Yes 0.0004328222 0.0006860971 0.001288234

With all the pvalues(midp.exact, fisher.exact, chi.square) < 0.01, we reject H0.
There is a significant association between D3 use and death.
"""


#c)	Give a 99% confidence interval for the odds ratio.
oddsratio(data, rev = "rows", conf.level = 0.99)$measure


"""
4. A binary happiness metric was recorded on a collection of subjects with a digestive condition.
The condition has three levels of severity, and the researcher hypothesizes that the
probability of being categorized as happy (Y=1) should decrease as the condition
becomes more severe. The data appear below. (15 pts)
"""

#a)	Create a matrix containing the count data, and apply appropriate row and column labels.
data2 <- matrix(c(68, 83, 24, 22, 69, 38), nrow = 3, byrow = TRUE)
rownames(data2) <- c("stage1", "stage2", "stage3")
colnames(data2) <- c("not happy", "happy")
data2


#b)	Explain whether the researcher believes disease stage is independent of happiness classification.

"""
The researchers does not believe that happiness is independent of disease.

Stage 1 shows more patients were happy as compared to not happy and stage 3 shows that more people were unhappy compared to happy.

The researchers believe that happiness decrease as the conditions become more severe.
"""


#c)	Perform the chi-square test for association. Give the hypotheses, test statistic, p-value, and conclusion in context.

"""
H0: Happiness is independent of disease stage (no association)
Ha: Happiness is not independent of disease stage (there is an association)
"""
chitest <- chisq.test(data2)
chitest
#X-squared = 9.5259, df = 2, p-value = 0.00854

"""
With X-squared = 9.5259, p-value = 0.00854 < 0.05, we reject H0.
There is significant evidence that Happiness is not independent of disease stage (there is an association)
"""


#d)	Present the sample proportions  π^_1, π^_2, and π^_3.
row_totals <- rowSums(data2)

pi_hat_1 <- data2[1,2] / row_totals[1]
pi_hat_2 <- data2[2,2] / row_totals[2]
pi_hat_3 <- data2[3,2] / row_totals[3]
row_totals
pi_hat_1
pi_hat_2
pi_hat_3

"""
e)	We will now perform a trend test, for which the hypotheses are:
```

$H_0: \pi_1=\pi_2=\pi_3$
$H_a: \pi_1>\pi_2>\pi_3$

The CochranArmitageTest(.) function in the “DescTools” package will perform the test.
Note that by default, the function will give a two-sided p-value, and we are
specifically interested in a decreasing trend. Give the p-value and conclusion in context.
"""

```
install.packages("DescTools")
library(DescTools)

CochranArmitageTest(data2, alternative = c("two.sided", "one.sided"))
```

"""
Z = 3.0695, dim = 3, p-value = 0.002144

With Z = 3.0695, p-value = 0.002144 < 0.05, we reject Ho.
There is decreasing trend for the sample proportions of happy individuals as the severity of the digestive condition increases.

"""

"""
5. (12 pts)
Marital Status Education Level
high school bachelors graduate
never married 36 21 15
married 36 45 57
divorced/widowed 18 18 15

"""

#a) Create a matrix containing the count data, and apply appropriate row and column labels.

```
data3 <- matrix(c(36,21,15,36,45,57,18,18,15), nrow = 3, byrow = TRUE)
rownames(data3) <- c("never married", "married", "divorced/widowed")
colnames(data3) <- c("high school", "bachelors", "graduate")
data3
```

#b) Perform the chi-square test for association using α=0.05. Give the hypotheses,
#test statistic, p-value, and conclusion in context.

```
chitest2 <- chisq.test(data3)
chitest2
```

"""
H0: Marital status and education level are independent.

Ha: Marital status and education level are not independent

Pearson's Chi-squared test

data: data3
X-squared = 14.464, df = 4, p-value = 0.005953

With X-squared = 14.464, p-value = 0.005953 < 0.05, we reject Ho.
There is significant evidence that martial status and education level is not independent.
"""

#c) Provide a table of observed cell counts (same as part (a)), a table of expected cell
#counts under independence, and a table of standardized residuals.

```
chitest2$expected
chitest2$stdres
```

#d) Use the tables from part (c) to write a summary statement about the nature of the association between marital status and education level.

"""
Never Married/Graduate group has large negative residual (-2.6440634), this means that there are fewer people who were never married and have graduate degree than expected under independence.

Never Married/ High School group has a positive residual of 3.2553022, which means that there’s a slightly high number of never-married individuals with only high school degree than expected.

Married/ High School group has a negative residual of -3.0226693, which means that there are fewer people that are married with only high school degree than expected.

Married/Graduate group has large positive residual 2.8935241), this means that there are slightly more people who were married and have graduate degree than expected under independence.

"""