Mini Project #5
Name: Niveditha Varadha Chandrasekaran

Exercise 1

Problem Statement:

The following table shows the Myers-Briggs personality preferences for a random sample of 389 past computer science graduates (from a University) in the listed professions

Occupation	Personality preference Type	
	E	I
Faculty	62	45
Data Scientist	56	81
Entrepreneur	94	51

Determine if there is any association between the listed occupations and the personality preferences at 5% level of significance.

Solution:

Test: The test to be applied is Chi square test of Independence.

Null and Alternative hypothesis:

- Null hypothesis H_0: Personality preferences and occupations are independent
- Alternative hypothesis H_1: Personality preferences and occupations are not independent

Test Statistic:

Chi-square test for independence

Test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left\{Obs(i,j) - \widehat{Exp}(i,j)\right\}^2}{\widehat{Exp}(i,j)},$$

where

$Obs(i,j) = n_{ij}$ are observed counts,

$\widehat{Exp}(i,j) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$ are estimated expected counts,

and χ^2_{obs} has $(k-1)(m-1)$ d.f.

Occupation	Personality preference Type		Total
	E	I	
Faculty	62 (Expected value:58.3136)	45 (Expected value:48.6864)	107
Data Scientist	56 (Expected value:74.6632)	81 (Expected value:62.3368)	137
Entrepreneur	94 (Expected value:79.0231)	51 (Expected value:65.9769)	145
Total	212	177	**389**

Therefore,
Chi square Obs = 17.003

Null distribution: $\chi^2_{(3-1)(2-1)}$ => : χ^2_2

Result:

```
        Pearson's Chi-squared test

data:  xmat
X-squared = 17.003, df = 2, p-value = 0.0002031
```

From the test it is evident that p-value is lesser than α (0.05). Therefore we reject the Null hypothesis and we can say that at 5% level of significance, there is sufficient evidence to conclude that personality preferences and occupations are not independent.

Exercise 2

Problem Statement:

Makers of generic drugs are required by the FDA to show that the extent of absorption of their drug in blood does not differ significantly from the "brand-name" drug that they imitate. To show this, 20 healthy nonsmoking male subjects are selected. For each subject, one of the two drugs is randomly chosen and given first. Then after a washout period, the other drug is given. In both the cases, the absorption of the drug in the blood is measured. The dataset stored in the file medicine.txt gives the measurements taken on 20 subjects. Do the drugs differ significantly in absorption?

Solution:

Test: The test to be applied is Signed rank test.

Null and Alternative hypothesis:

- Null hypothesis H_0: The drugs do not differ significantly in absorption
- Alternative hypothesis H_1: The drugs differ significantly in absorption

Test Statistic:

Wilcoxon signed rank test

Test of the median, $H_0: M = m$.

Test statistic $W = \sum_{i: X_i > m} R_i$, where R_i is the rank of $d_i = |X_i - m|$.

Null distribution: Table A8 or recursive formula (10.6).

For $n \geq 15$, $W \approx Normal\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$

Assumptions: the distribution of X_i is continuous and symmetric

Subject	Brand-name	Generic	Difference
1	4108	1755	2353
2	2526	1138	1388
3	2779	1613	1166
4	3852	2254	1598
5	1833	1310	523
6	2463	2120	343
7	2059	1851	208
8	1709	1878	-169
9	1829	1682	147
10	2594	2613	-19
11	2344	2738	-394
12	1864	2302	-438
13	1022	1284	-262
14	2256	3052	-796
15	938	1287	-349
16	1339	1930	-591
17	1262	1964	-702
18	1438	2549	-1111
19	1735	3340	-1605
20	1020	3050	-2030

Median = -215.5
Therefore, p-value = 0.5958

Result:

```
        Wilcoxon signed rank test

data:  x and y
V = 90, p-value = 0.5958
alternative hypothesis: true location shift is not equal to 0
```

The P Value is 0. 5958 (large) so there is no evidence to reject the NULL hypothesis; hence we accept the null hypothesis. Therefore there is no significant difference in the absorption of both the drugs in blood.

Exercise 3

Problem Statement:

This landmark experiment in genetics investigated whether, for a certain kind of sweet pea plant, the traits “flower color" and “pollen grain types" are inherited independently or not. Flower color can be purple (P) or red (R), but the purple color is dominant) and Grain type can be Long (L) or Round (R), but long grain type is dominant. According to Mendel's law of independent segregation, the genes for these two traits segregate independently and the “flower color" and “pollen grain type" combinations-P&L, P&R, R&L and R&R are expected in 9:3:3:1 ratio. The following are the observed frequencies for each combination when the experiment was carried out on 256 sweet pea plant.

“flower color" and “pollen grain type" combinations	Observed Frequencies
P&L	177
P&R	15
R&L	15
R&R	49

Write your conclusion at 5% level of significance.

Solution:

Test: We have to find whether the traits “flower color" and “pollen grain type" are inherited independently or not using chi square test for the goodness of fit.

Null and Alternative hypothesis:

- Null hypothesis H_0: The traits “flower color" and “pollen grain type" are inherited independently
- Alternative hypothesis H_1: The traits “flower color" and “pollen grain type" are not inherited independently

We can use chi-square statistic for the goodness of fit to show that the probabilities of the combination of the gene traits are equal to the ones in Mendel’s Law.

Test Statistic:

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

N=256

Result:

```
        Chi-squared test for given probabilities

data:  freq
X-squared = 121, df = 3, p-value < 2.2e-16
```

From the test it is evident that p-value is lesser than α (0.05). Therefore we reject the Null hypothesis and we can say that the traits “flower color" and “pollen grain type" are not inherited independently.

R Code:

Exercise 1:

```
> x <- c(62,45,56,81,94,51)
>
> #Format the given Data in a table format
> xmat <- matrix(x, byrow=T, ncol=2)
>
> #Compute the chi square statistics
> result <- chisq.test(xmat)
> result
```

Pearson's Chi-squared test

data: xmat
X-squared = 17.003, df = 2, p-value = 0.0002031

Exercise 2:

```
> x = c(4108,2526,2779,3852,1833,2463,2059,1709,1829,2594,2344,1864,1022,2256,938,1339,
1262,1438,1735,1020)
> y = c(1755,1138,1613,2254,1310,2120,1851,1878,1682,2613,2738,2302,1284,3052,1287,193
0,1964,2549,3340,3050)
>
> #Wilcox Test statistics to find the P Value
> wilcox.test(x,y,paired=TRUE)
```

Wilcoxon signed rank test

data: x and y
V = 90, p-value = 0.5958
alternative hypothesis: true location shift is not equal to 0

Exercise 3:

```
> freq = c(177,15,15,49)
> # specify probabilities
> probs = c(9,3,3,1)/16
> chisq.test(freq,p=probs)
```

Chi-squared test for given probabilities

data: freq
X-squared = 121, df = 3, p-value < 2.2e-16