Statistical Inference HW#3

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Thursday 26th January, 2023

Dept. of Computer Engineering

University of Tehran

Contents

1 Problem 1

1.1 Wilcoxon signed rand test

In Wilcoxon signed rank test we have null hypothesis which in here is :

$$H_0 : m_0 = 50,000 \tag{1}$$

So the alternative hypothesis would be:

$$H_1 : m > 50,000 \tag{2}$$

Now let start the procedure of Wilcoxon test as follow:

- Calculate each $(x_i - m_0)$
- Sort absolute values of $(x_i - m_0)$
- W^+ is the sum of the ranks assosiated with positive values of $(x_i - m_0)$
- W^- is the sum of the ranks assosiated with negative values of $(x_i - m_0)$

Calculate each $(x_i - m_0)$:(Note that for m_0 we have 50 thousand of dollors so for simplicity we can assume this number is 50 due to given data is thousand dollors numbers).

$$\begin{aligned} x_i &= (43, 47, 52, 68, 72, 55, 61, 44, 58, 63, 54, 59, 77, 36, 80, 53, 60) \\ x_i - m_0 &= (-7, -3, 2, 18, 22, 5, 11, -6, 8, 13, 4, 9, 27, -14, 30, 3, 10) \end{aligned} \tag{3}$$

Sorted values:

$$x_{(i)} - m_0 = (-14, -7, -6, -3, 2, 3, 4, 5, 8, 9, 10, 11, 13, 18, 22, 27, 30,) \tag{4}$$

Sorted absolute numbers are:

$$sort(|x_i - m_0|) = (2, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 18, 22, 27, 30) \tag{5}$$

Rank of numbers are:

$$Ranks = (1, 2.5, 2.5, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17) \tag{6}$$

Attention: cause we have 2 same numbers for 3 so we must use median rank for both thats why we write two ranks as 2.5.

$$\begin{aligned} W^+ &= 1 + 2.5 + 4 + 5 + 8 + 9 + 10 + 11 + 12 + 14 + 15 + 16 + 17 \\ &= 124.5 \\ W^- &= 13 + 7 + 6 + 2.5 \\ &= 28.5 \end{aligned} \tag{7}$$

As discussed in the class we can use approximation with Normal distribution as follow:

$$E[W_*] = \frac{n(n+1)}{4}$$
$$Var[W_*] = \frac{n(n+1)(2n+1)}{24} \tag{8}$$

Where W_* is for test statistic so it is either W^+ or W^-, So we shall say that:

$$Z = \frac{W_* - E[W_*]}{\sqrt{var(W_*)}} \tag{9}$$

Cause we have 17 samples and its bigger than 12 we can use Z approximation and say that:

$$\begin{aligned} P(T \leq t) &= P\left(\frac{T-\mu_t}{\sigma_t} \leq \frac{t-\mu_t}{\sigma_t}\right) \\ &= P\left(Z \leq \frac{t-\mu_t}{\sigma_t}\right) \end{aligned} \tag{10}$$

The probability of $P(Z \leq \frac{t-\mu_t}{\sigma_t}) = 0.95$ would be $Z = 2$ so we can say that:

$$\begin{aligned} \frac{t-\mu_t}{\sigma_t} &= 2 \\ t &= 76.5 + 2 * 21.124630174 \\ &= 118.74926034855523 \end{aligned} \tag{11}$$

So cuase to $124.5 > 118$ null hypothesis would be rejected.

1.2 Mann-Whitney-Wilcoxon

Since Mann-Whitney is for two samples and here we have just one sample therefore two methods are considered to solve:

1.2.1 Approach1

For this approach we divide the 17 subjects into two group with 9 subjects in first group and 8 subjects into second group as follow:

$$\begin{aligned} Group_1 &= (36, 44, 52, 54, 58, 60, 63, 72, 80) \\ Group_2 &= (43, 47, 53, 55, 59, 61, 68, 77) \end{aligned} \tag{12}$$

Now lets merge data and calculate the U but at first calculate W:

$$\begin{aligned} W_1 &= 1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 \\ &= 81 \\ W_2 &= 2 + 4 + 6 + 8 + 10 + 12 + 14 + 16 \\ &= 72 \end{aligned} \tag{13}$$

Obtain U from equation below:

$$
\begin{aligned}
U_1 &= nm + \frac{n(n+1)}{2} - W_1 \\
&= 9*8 + \frac{9*10}{2} - 81 \\
&= 36 \\
U_2 &= nm + \frac{m(m+1)}{2} - W_2 \\
&= 9*8 + \frac{8*9}{2} - 72 \\
&= 36 \\
U &= min(U_1, U_2) \\
&= 36
\end{aligned} \tag{14}
$$

So from table of U we can say that:

$$U(\alpha = 0.05, n_1 = 9, n_2 = 8) = 15 \tag{15}$$

But we obtain 36 which is bigger thatn 15 so null hypothesis would be rejected.

1.2.2 Approach2

consider another sample with 17 software developers with 50 thousand salary.Now we have 2 sample and we can calculate the W_i:

$$
\begin{aligned}
Group_1 &= (36, 43, 44, 47, 52, 53, 54, 55, 58, 59, 60, 61, 63, 68, 72, 77, 80) \\
Group_2 &= (50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50)
\end{aligned} \tag{16}
$$

- Merging data and sort them:

$$
\begin{aligned}
\text{Sort data} = [&36, 43, 44, 47, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, 50, \\
&50, 50, 50, 50, 50, 50, 52, 53, 54, \\
&55, 58, 59, 60, 61, 63, 68, 72, 77, 80]
\end{aligned} \tag{17}
$$

- Assign Ranks:

$$
\begin{aligned}
\text{Rank} = [&1, 2, 3, 4, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13, \\
&13, 13, 13, 13, 13, 13, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]
\end{aligned} \tag{18}
$$

- Calculate W_i:

$$
\begin{aligned}
W_1 &= 1 + 2 + 3 + 4 + 22 + 23 + 24 + 25 + 26 + 27 + 28 + 29 + 30 + 31 + 32 + 33 + 34 \\
&= 374 \\
W_2 &= 17 * 13 \\
&= 221
\end{aligned} \tag{19}
$$

- Statistics of test:

$$\begin{aligned} U_1 &= nm + \frac{n(n+1)}{2} - W_1 \\ &= 68 \\ U_2 &= nm + \frac{m(m+1)}{2} - W_2 \\ &= 221 \end{aligned} \tag{20}$$

- Approximation with z-score:

$$\begin{aligned} \mu_{U_i} &= \frac{17 * 17}{2} \\ &= 144.5 \\ \sigma_{U_i} &= \sqrt{\frac{nm(n+m+1)}{12}} \\ &= \sqrt{\frac{10115}{12}} \\ &= \sqrt{842.916666667} \\ &= 29.03302717 \end{aligned} \tag{21}$$

- Calculate z-score:

$$\begin{aligned} z &= \frac{U_i - \mu_{U_i}}{\sigma_{U_i}} \\ z_1 &= \frac{68 - 144.5}{29.03302717} \\ &= -2.634930197 \\ z_2 &= \frac{221 - 144.5}{29.03302717} \\ &= 2.634930197 \end{aligned} \tag{22}$$

- Critical region for 0.05 significant level:

$$z_\alpha = \pm 1.96 \tag{23}$$

Since the 2.63 is in the rejection area therfore the null hypothesis will be rejected.(Also for one sided null hypothesis will be rejected since 2.63 is above 1.96 and it is indeed in rejection area)

2 Problem 2

2.1 Do participants ..

We can assume the question said is median of first column is bigger of less than median of column 2 or not.
So we can use signed pair test to do this experminet to see the if median is zero or not.
Attention:each row is independent due to differentiation between participants and left ear and right ear is dependent cause it's for a same person therfore we can use signed pair test as follow:

$$\begin{aligned} d_i &= x_{i1} - x_{i2} \\ &= (-7, -1, 3, -5, 7, -8, 1, -4, -2, 0, -10, -27) \end{aligned} \tag{24}$$

Where x_{i1} and x_{i2} are for left and right ears respectively therefore we can form null hypothesis as follow:

$$H_0 : median = 0 \tag{25}$$

S will be $S = 4$ (note that S is the positive numbers):

$$\begin{aligned} p - value &= 2min \left\{ P\left[S \leq 4\right], P\left[S \geq 4\right] \right\} \\ &= 2P\left[S \leq 4\right] \\ &= 2 * 0.19385 \\ &= 0.3877 \end{aligned} \tag{26}$$

So for $\alpha = 0.05$, p-value is bigger than α therefore the null hypothesis would not rejected. And we can not say that participants report more words from one ear than the other.
Attention: We have no idea about distribution of our data so we can assume if median is less than or greater than zero and define our null hypothesis where median is zero then if we reject null hypothesis so we can report one ear of participants hear more words than the other. But as we saw we could not reject null hypothesis therefore we can not apply this report.

2.2 Mann-Whitney-Wilcoxon

In here we have paired samples and Mann-Whitney is for 2 different samples for 2 different populations we must use paired Wilcoxon Signed Rank Test as follow:

$$d_i = x_{i1} - x_{i2} \tag{27}$$

Now we can use same procedure for one sample wilcoxon signed rank test:

$$d_i = (-7, -1, 3, -5, 7, -8, 1, -4, -2, 0, -10, -27) \tag{28}$$

Obtain ranks:

$$r_i = (8.5, 2.5, 5, 7, 8.5, 10, 2.5, 6, 4, 1, 11, 12) \tag{29}$$

Calculate W^+ and W^-:

$$\begin{aligned} w^+ &= 5 + 8.5 + 2.5 + 1 \\ &= 17 \\ w^- &= 8.5 + 2.5 + 7 + 10 + 6 + 4 + 11 + 12 \\ &= 61 \end{aligned} \tag{30}$$

For W^+ and $n = 12$ we can use table therefore for two sided $\alpha = 0.05$ since we have greater than 14 we can say it would not reject our null hypothesis and the $m = 0$ can't be rejected.

3 Problem 3

3.1 Make a histogram

For this question we plot the histogram for each men and female data which you can see in figure 1 and 2.

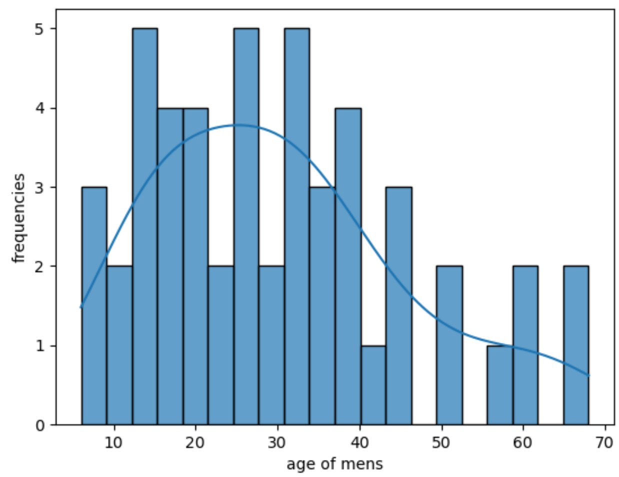

Figure 1: Histogram for men and bins= 20

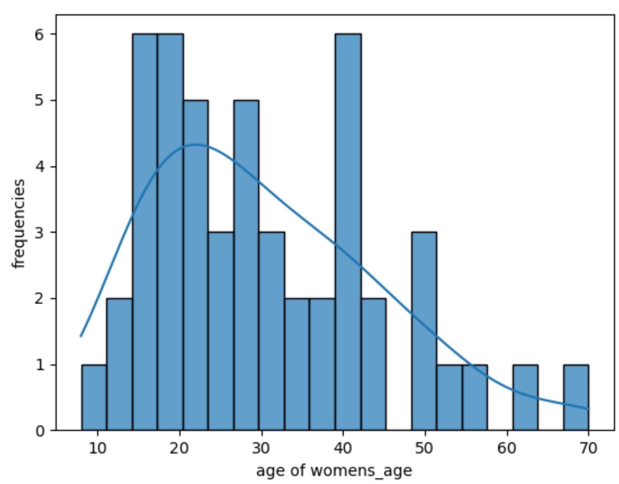

Figure 2: Histogram for female and bins= 20

3.2 Is it normal ?

From maximum likelyhood we can find that the parameters that fit data is:

$$\mu = \overline{X}$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{31}$$

So we can calculate the mean and variance of men's ages:

$$\mu = 30.24$$
$$\sigma = 15.353905040738008 \tag{32}$$

For mean and variance of women's ages:

$$\mu = 30.36$$
$$\sigma = 13.774991833028432 \tag{33}$$

So we can use goodness of fit to see if we can reject our null hypothesis which is being normality distribution of mens and womens ages.At first we are going to calculate the frequecny of each bins and probability of each edges and exepected means:

Table 1: Table for men

interval	probability	frequecny	exepected mean
$(6.0 - 9.1)$	0.027083019378079574	3	1.3541509689039788
$(9.1 - 12.2)$	0.03572872408381289	2	1.7864362041906445
$(12.2 - 15.3)$	0.045257843054860025	5	2.2628921527430013
$(15.3 - 18.4)$	0.05504605154003095	4	2.7523025770015472
$(18.4 - 21.5)$	0.06428572057391502	4	3.214286028695751
$(21.5 - 24.6)$	0.07208732687899128	2	3.6043663439495637
$(24.6 - 27.7)$	0.07761745813046306	5	3.8808729065231526
$(27.7 - 30.8)$	0.08024463726301195	2	4.012231863150598
$(30.8 - 33.90)$	0.07965787854103024	5	3.982893927051512
$(33.9 - 37.0)$	0.0759272320359966	3	3.7963616017998305
$(37.0 - 40.1)$	0.06949003038751578	4	3.4745015193757887
$(40.1 - 43.2)$	0.061066568416416445	1	3.0533284208208222
$(43.2 - 46.3)$	0.05152767880139342	3	2.576383940069671
$(46.3 - 49.4)$	0.04174780530007127	0	2.0873902650035636
$(49.4 - 52.5)$	0.03247750629956103	2	1.6238753149780516
$(52.5 - 55.6)$	0.024259821254999814	0	1.2129910627499907
$(55.6 - 58.7)$	0.017399963665185147	1	0.8699981832592574
$(58.7 - 61.8)$	0.011982979850865827	2	0.5991489925432913
$(61.8 - 64.9)$	0.007923863203512571	0	0.39619316017562856
$(64.9 - 68.0)$	0.005031119839876919	2	0.25155599199384593

Table 2: Table for women

interval	probability	frequecny	exepected mean
$(8.0 - 11.1)$	0.028758649530104277	1	1.437932476505214
$(11.1 - 14.2)$	0.0393411277344875	2	1.967056386724375
$(14.2 - 17.3)$	0.05117081743626367	6	2.5585408718131837
$(17.3 - 20.4)$	0.063284195011936	6	3.1642097505968
$(20.4 - 23.5)$	0.07441587981647096	5	3.720793990823548
$(23.5 - 26.6)$	0.08320195145393966	3	4.160097572696983
$(26.6 - 29.7)$	0.08845023433285037	5	4.422511716642519
$(29.7 - 32.8)$	0.08940505514324654	3	4.470252757162327
$(32.8 - 35.9)$	0.0859256409207576	2	4.296282046037881
$(35.9 - 39.0)$	0.07852014333875068	2	3.9260071669375343
$(39.0 - 42.1)$	0.06822396306857936	6	3.411198153428968
$(42.1 - 45.2)$	0.05636250590795577	2	2.8181252953977887
$(45.2 - 48.3)$	0.04427321343434143	0	2.2136606717170713
$(48.3 - 51.4)$	0.03306656516445827	3	1.6533282582229136
$(51.4 - 54.5)$	0.023481956298045947	1	1.1740978149022974
$(54.5 - 57.6)$	0.01585537284842553	1	0.7927686424212765
$(57.6 - 60.7)$	0.010179242325147886	0	0.5089621162573943
$(60.7 - 63.8)$	0.006213711033099534	1	0.3106855516549767
$(63.8 - 66.9)$	0.0036064758636353833	0	0.18032379318176917
$(66.9 - 70.0)$	0.0019902660538866357	1	0.09951330269433178

Now from 1 and 2 we can calculate the chi-square as follow:

$$\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i} \tag{34}$$

Note that we have 2 parameters and 50 data so we can use $\chi^2_{\alpha,m-t-1}$ where:

$$\begin{aligned} m &= 50 \\ t &= 2 \\ \alpha &= 0.05 \\ \chi^2_{0.05,47} &= 61 \end{aligned} \tag{35}$$

For man and women we have $\chi^2 = 29.32808366794874$ and $\chi^2 = 26.758692311946277$ so cause both of this values are less than 61 therefore we can not reject normal distribution and we can reject it if and only if $\chi^2 \geq \chi^2_{\alpha,m-t-1}$ so null hypothesis can not be rejected and we can say they are normal distributions. So for begin more ccaution with our test we will consider shapiro where statistics and p-values derive from python code is in the table 29:

Table 3: Statistics and p-values for men and womens ages

	Statistics	p-values
mens	0.9460382461547852	0.023510854691267014
womens	0.9431541562080383	0.01799035631120205

Shapiro test will reject null hypothesis with 0.05 significant level.

3.3 Can we use parametric tests?

Generally speaking of paramateric tests we can use mean,variance and proportion but we can not use proportion tests cause X is not the number of success so we can not use mean and variance tests cause assumtion for nomrality population were rejected due to previous parts.

3.4 Transform the data

For transformation we have several ways to do as follows:

- Logarithmic Transformation: Applied to positive data, can help reduce right-skewed data.
- Square Root Transformation: This method is similar to lograithmic transformation but less strong, and this method is useful for count data.
- Box-Cox Transformation: A family of power transformations that includes logarithmic, square root, and inverse transformations.
- Yeo-Johnson Transformation:Similar to Box-Cox but can handle zero and negative values.

So we assign multiple transform as you can see in the figures 3, 4, 5, 6, 7, 8, 9 and 10. We recheak the shapiro

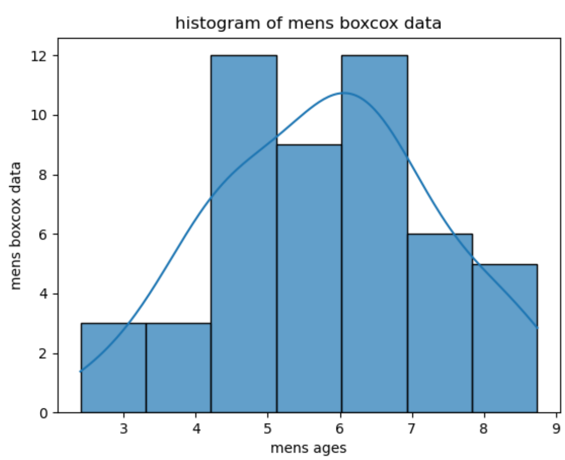

Figure 3: BoxCox transformation for mens ages

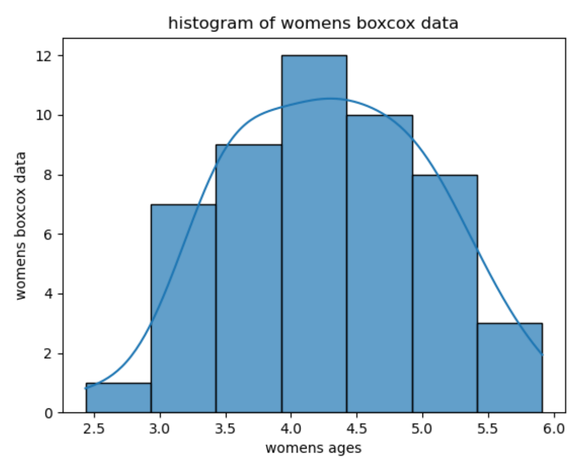

Figure 4: BoxCox transformation for womens ages

test for being normality assumtion which you can see in the mens data table 4 and womem's data table 5:

Table 4: Recheck each transformations with shapiro test for men

men's data	Logarithmic	Square Root	BoxCox	Yeo-Johnson
p-value of shapiro	0.44	0.68	0.84	0.83

From each tabel either men of women we observed that boxcox is the best transformation so we will use it since null hypothesis would not be rejected since p-value is large enough.

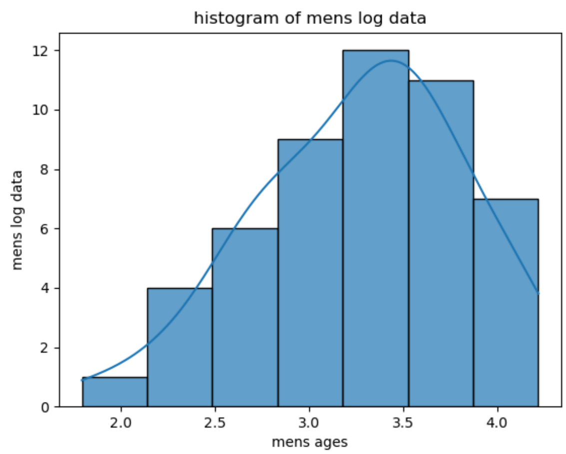

Figure 5: Logarithmic transformation for mens ages

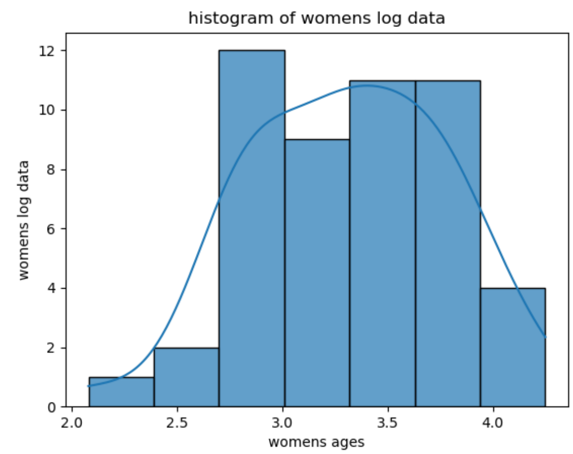

Figure 6: Logarithmic transformation for womens ages

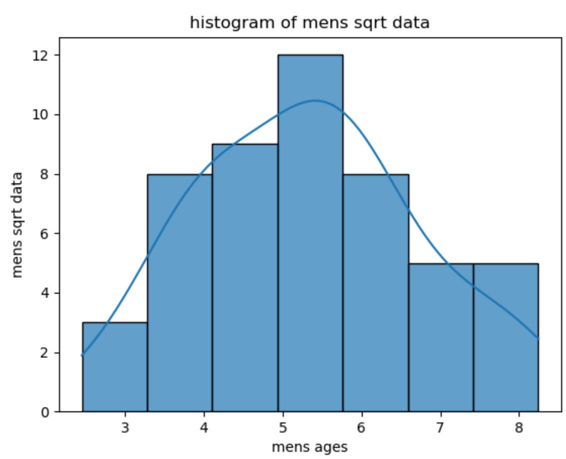

Figure 7: Square Root transformation for mens ages

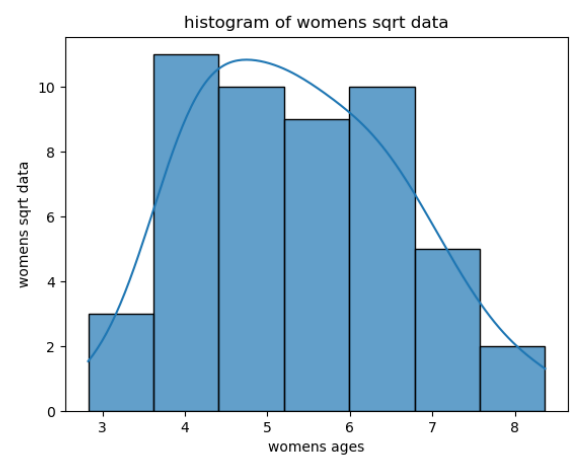

Figure 8: Square Root transformation for womens ages

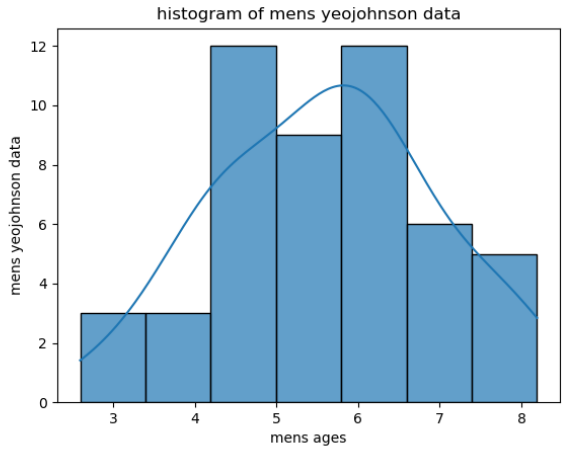

Figure 9: Yeo-Johnson transformation for mens ages

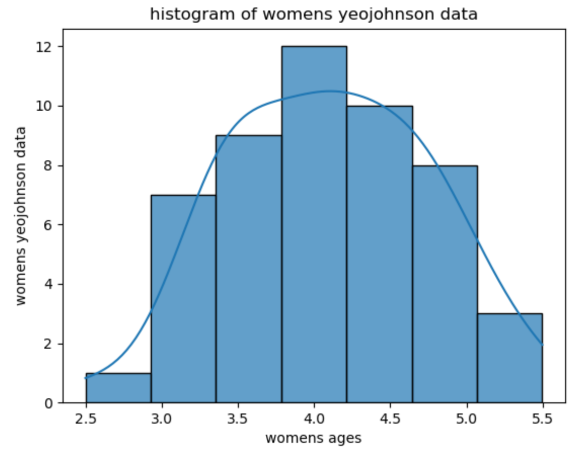

Figure 10: Yeo-Johnson transformation for womens ages

Table 5: Recheck each transformations with shapiro test for women

women's data	Logarithmic	Square Root	BoxCox	Yeo-Johnson
p-value of shapiro	0.70	0.45	0.78	0.77

3.5 Parametric test

We use differences between means as follow:(using the box cox transformation data)

- null hypothesis: $\mu_1 - \mu_2 = \delta = 0$
- alternative: $\mu_1 - \mu_2 > 0$
- From likelyhood ratio technique we can use $z = \frac{\overline{x}_1 - \overline{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

where μ_1 and μ_2 are for men and women means and δ is zero.
For independent random samples from populations with unknown variances that may not even be normal:

- as long as both samples are large enough to invoke the central limit theorem
- just s_1 substituted for σ_1 and s_2 substituted for σ_2.

Calculating the mean and variances of each gender using .mean() and .var() in tabel 6: Now lets calculate the

Table 6: mean and variacne of each gender's boxcox transformation data

	Men	Women
Mean	5.81	4.30
Variacne	2.24	0.57

z-score:

$$z = \frac{5.81 - 4.30}{\sqrt{0.056101817}} = 6.391342207 \tag{36}$$

Now for $\alpha = 0.05$ we z-score is 1.64 therefore the null hypothesis will be rejected since $6.39 > 1.64$, and male fans are older than female fans.

3.6 Non-parametric tests

In non-paramateric tests we just have Mann-Whitney test that is useful since we have 2 independent samples so we can use it. The procedure is:

- $H_0 : m = m_{men} - m_{women} = 0$
- $H_1 : m > 0$
- Arrange the observations from the two samples combined in increasing order (retaining sample membership) and assign ranks to the observations
- Let $W_1 =$ the sum of the ranks for sample 1.
- Let $W_2 =$ the sum of the ranks for sample 2.
- Obtain U_1 and U_2:

$$U_1 = nm + \frac{n(n+1)}{2} - W_1$$
$$U_2 = nm + \frac{m(m+1)}{2} - W_2$$

- For $n_1 > 10$ and $m_1 > 10$ we can use z approximation:

$$\mu_{U_i} = \frac{nm}{2}$$

$$\sigma_{U_i} = \sqrt{\frac{nm(n+m+1)}{12}}$$

Calculate W_i:

$$W_1 = 2488$$
$$W_2 = 2562.0 \tag{37}$$

Calculate μ_{U_i}:

$$\mu_{U_1} = \mu_{U_2} = \frac{50 * 50}{2} = 1250 \tag{38}$$

Obtain σ_{U_i}:

$$\sigma_{U_1} = \sigma_{U_2} = \sqrt{\frac{(50 * 50)(50 + 50 + 1)}{12}} \tag{39}$$
$$= 145.057459879 \tag{40}$$

$$\begin{aligned} U_1 &= 2500 + 1275 - 2488 \\ &= 1287 \\ U_2 &= 2500 + 1275 - 2562 \\ &= 1213 \end{aligned} \tag{41}$$

$$\begin{aligned} z_1 &= \frac{1287 - 1250}{145.057459879} \\ &= 0.255071335 \\ z_2 &= \frac{1213 - 1250}{145.057459879} \\ &= -0.255071335 \end{aligned} \tag{42}$$

since both z_i are less than $z_{alpha/2}$ which is 1.64 therfore we can not reject null hypothesis and $m_{men} = m_{women}$ But compare to paramateric test, it will be rejected while non-paramateric test will not.Since Neyman-Pearson Lemma says among all such possible partitions, that based on the likelihood ratio maximizes the power therefore power of our paramateric test is more stronger than non-paramateric one.

3.7 which test is more appropriate?

Since Neyman-Pearson Lemma says among all such possible partitions, that based on the likelihood ratio maximizes the power. They are more powerful (i.e., you can detect smaller effect sizes with smaller samples) than comparable non-parametric tests when the parametric assumptions are correct (or approximately correct).
So we can say paramateric tests are more powerful than non-parametric, and paramateric test are more sutiable than non-parametric cuase of it’s power and strength of test.

4 Problem 4

4.1 Show that rows and cols are independent...

Assuming $p_{ij} = p_i p_j$ for null hypothesis so in a table we calculate $p_i p_j$ for first table 7:

Table 7: Calculate the $p_i * p_j$

$0.3 * 0.5$	$0.3 * 0.3$	$0.3 * 0.2$
$0.3 * 0.5$	$0.3 * 0.3$	$0.3 * 0.2$
$0.4 * 0.5$	$0.4 * 0.3$	$0.4 * 0.2$

The value calculated in table 7 is exactly same in real given table so they are independent.

4.2 Generate numbers

We generated the random numbers in python and the numbers are in the table 8:

Table 8: Generated numbers in python

49	22	17
51	25	26
52	42	16

4.3 Consider table and conclude null hypothesis

Our null hypothesis will be independency of rows and cols so we can use contigency table as follow:

- Calculate the $\theta_{i.} = \frac{\sum_{j=1}^{c} f_{ij}}{f_{\text{total}}}$

$$\begin{aligned}\theta_{1.} &= 0.293333333 \\ \theta_{2.} &= 0.34 \\ \theta_{3.} &= 0.366666667\end{aligned} \tag{43}$$

- Calcuate the $\theta.j = \frac{\sum_{i=1}^{r} f_{ij}}{f_{\text{total}}}$

$$\begin{aligned}\theta_{.1} &= 0.506666667 \\ \theta_{.2} &= 0.296666667 \\ \theta_{.3} &= 0.196666667\end{aligned} \tag{44}$$

- Calculate the $e_{ij} = \theta_{i.}\theta_{.j} f_{\text{total}}$

Table 9: Expected frequenies

44.586666645	26.106666666	17.306666676
51.680000034	30.260000034	20.060000034
55.733333421	32.6333334	21.63333339

- Calculate the χ^2:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\ &= 0.436846097 + 0.645992509 + 0.005434001 + 0.008947369 + 0.914329158 + 1.758903267 \\ &+ 0.250079756 + 2.688491614 + 1.466923498 \\ &= 8.175947269 \end{aligned} \tag{45}$$

for $\chi^2_{df=(r-1)(c-1),\alpha=0.05} = 9.488$ and since the test gives us 8.17 and it is less than 9.448 then null hypothesis will not be rejected.
But if we assume that expected frequenies is $p_{ij} * 300$ as you can see in tabale 10:

Table 10: Expected frequenies based on p_{ij}

45	27	18
45	27	18
60	48	24

Then calculate the χ^2 statistics:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\ &= 0.35555556 + 0.92592593 + 0.05555556 \\ &+ 0.8 + 0.14814815 + 3.55555556 \\ &+ 1.06666667 + 1 + 2.66666667 \\ &= 10.574074074074073 \end{aligned} \tag{46}$$

And the p-value is 0.032 where null hypothesis will be rejected.

4.4 Optional

For following question we consider 10000 simulation where in each we generate numbers like section b and form the 3*3 table of observed values and then calculate the χ^2 statistics the same way before. (Assume the expected frequenies are $p_{ij} * 300$) then we plot histogram for statistic as you can see in the figure 11:
From the obtained stats we can calculate the p-value by np.mean(chi-stats $>=$ chi-score) where p-value will be:

$$\text{p-value} = 0.2236 \tag{47}$$

So the null hypothesis with this simulations will not be rejected since p-value is bigger than significant level of 5%.

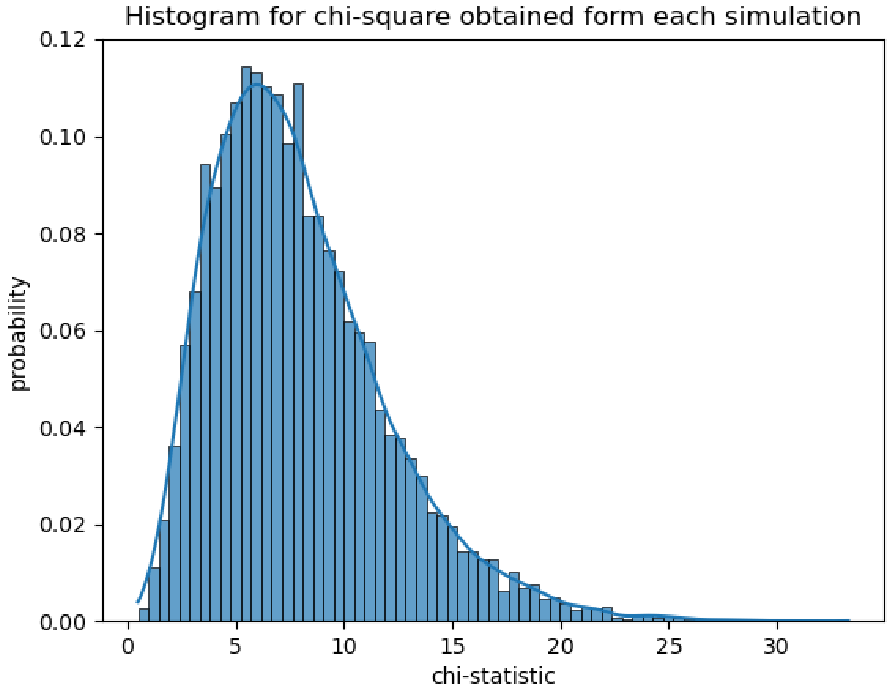

Figure 11: Histogram of statistics for 10000 simulations

5 Problem 5

5.1 Independency test

As we know it is considered as contigency table and from lecture 9 we have:

$$\begin{aligned} \hat{\theta}_{i.} &= \frac{f_{i.}}{f} \\ \hat{\theta}_{.j} &= \frac{f_{.j}}{f} \\ e_{ij} &= \frac{f_{i.}f_{.j}}{f} \end{aligned} \tag{48}$$

We can calculate χ^2 afterward:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{49}$$

$$
\begin{aligned}
\hat{\theta}_{1.} &= 0.4 \\
\hat{\theta}_{2.} &= 0.6 \\
\hat{\theta}_{.1} &= 0.8 \\
\hat{\theta}_{.2} &= 0.2
\end{aligned}
\tag{50}
$$

Calculate the e_{ij} in the table 11: Now lets calculate the χ^2 as follow:

Table 11: Table containing the e_{ij} for each row and col

e_{ij}	1	2
1	24	8
2	48	12

$$
\begin{aligned}
\chi^2 &= \frac{(28-24)^2}{24} + \frac{(12-8)^2}{8} + \frac{(52-48)^2}{48} + \frac{(8-12)^2}{12} \\
&= 0.666666667 + 2 + 0.333333333 + 1.333333333 \\
&= 4.333333333
\end{aligned}
\tag{51}
$$

what is the degree of freedom ?

$$
\begin{aligned}
(r, c) &= (2, 2) \\
\text{degree of freedom} &= (r-1)(c-1) = 1
\end{aligned}
\tag{52}
$$

The value for $\chi^2_{\alpha,1}$ is:

$$
\chi^2_{\alpha,1} = \begin{cases} 3.838 & \text{when } \alpha \text{ is } 0.05 \\ 6.617 & \text{when } \alpha \text{ is } 0.01 \end{cases}
\tag{53}
$$

So for $\alpha = 0.01$ the null hypothesis would not be rejected and for $\alpha = 0.05$ the null hypothesis would be rejected since $\chi^2 \geq 3.838$.

6 Problem 6

6.1 Independecy test

- Calculate $\hat{\theta_{i.}}$ and $\hat{\theta_{.j}}$
- Obtain e_{ij} from $\hat{\theta_{i.}}\hat{\theta_{.j}} f_{total}$
- What is the value of χ^2.

Step 1:

$$\begin{aligned}\hat{\theta}_{1.} &= \frac{244}{300} \\ \hat{\theta}_{2.} &= \frac{56}{300} \\ \hat{\theta}_{.1} &= \frac{95}{300} \\ \hat{\theta}_{.2} &= \frac{116}{300} \\ \hat{\theta}_{.3} &= \frac{61}{300} \\ \hat{\theta}_{.4} &= \frac{28}{300}\end{aligned} \tag{54}$$

Step 2: Step 3:

Table 12: Calculate e_{ij}

e_{ij}	1	2	3	4
1	77.266666667	94.346666667	49.613333333	22.773333333
2	17.733333333	21.653333333	11.386666667	5.226666667

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\ &= 0.289962611 + 0.302997927 + 0.387856311 + 0.62520687 \\ &+ 1.263408521 + 1.320205255 + 1.689945355 + 2.724115646 \\ &= 8.603698496\end{aligned} \tag{55}$$

Degree of freedom is:

$$\text{degree of freedom} = (r-1)(c-1) = 3 * 1 = 3 \tag{56}$$

Calculate χ^2:

$$\chi^2_{3,\alpha} = \begin{cases} 11.35 & \text{where } \alpha \text{ is } 0.01 \\ 7.815 & \text{where } \alpha \text{ is } 0.05 \end{cases} \tag{57}$$

When α is 0.01 the null hypothesis would be rejected while 0.05 would be rejected since $\chi^2_{3,0.05} \geq 7.815$

7 Problem 7

7.1 Prove that the minimum of ...

At first lets see the CDF of $F_n(x)$ where n is five sample in this quesiton:

$$F_n(x) = \begin{cases} 0 & -\infty < x < y_1 \\ 0.2 & y_1 \leq x < y_2 \\ 0.4 & y_2 \leq x < y_3 \\ 0.6 & y_3 \leq x < y_4 \\ 0.8 & y_4 \leq x < y_5 \\ 1 & y_5 \leq x < \infty \end{cases} \tag{58}$$

So considering $F(y_1)$ will be helpful to prove this. Assume that $F(y_1) \geq 0.1$ so F will be close to 0.1 when x approaching from the left to y_1 but F_n is zero in that area,therfore superimum over this interval would be at least as large as 0.1.
But what if $F(y_1) < 0.1$?
We consider the right part of y_1 when x is greater or equal to y_1 then F_n is 0.2 while $F(y_1) < 0.1$ so:

$$|F(y_1) - 0.2| \geq 0.1 \tag{59}$$

So the superimum would be larger than 0.1. For next part if $D_n = 1$ we have only and only if condition so at first we assume that $D_n = 0.1$ then proving $F(y_1) = 0.1$ and so on.

$$D_n = \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = \begin{cases} |F(x)| & -\infty < x < y_1 \\ |0.2 - F(x)| & y_1 \leq x < y_2 \\ |0.4 - F(x)| & y_2 \leq x < y_3 \\ |0.6 - F(x)| & y_3 \leq x < y_4 \\ |0.8 - F(x)| & y_4 \leq x < y_5 \\ |1 - F(x)| & y_5 \leq x < \infty \end{cases} \tag{60}$$

Now we must see the values for y_1, y_2, y_3, y_4, y_5:

$$\begin{aligned} \left|F(y_1^-)\right| &\leq 0.1 \\ -0.1 \leq F(y_1^-) &\leq 0.1 \\ \left|0.2 - F(y_1^+)\right| &\leq 0.1 \\ 0.1 \leq F(y_1^+) &\leq 0.3 \end{aligned} \tag{61}$$

As question said $F(x)$ is continuous as we know for continuous distribution $F(y_1^-) = F(y_1^+)$ therefore with two conditions above we will reach $F(y_1) = 0.1$.Like for y_1 we can say same to y_2,y_3,y_4 and y_5.

$$\begin{aligned} \left|0.2 - F(y_2^-)\right| &\leq 0.1 \\ 0.1 \leq F(y_2^-) &\leq 0.3 \\ \left|0.4 - F(y_2^+)\right| &\leq 0.1 \\ 0.3 \leq F(y_2^+) &\leq 0.5 \rightarrow \\ F(y_2) &= 0.3 \end{aligned} \tag{62}$$

$$
\begin{aligned}
\left|0.4 - F(y_3^-)\right| &\leq 0.1 \\
0.3 \leq F(y_3^-) &\leq 0.5 \\
\left|0.6 - F(y_3^+)\right| &\leq 0.1 \\
0.5 \leq F(y_3^+) &\leq 0.7 \rightarrow \\
F(y_3) &= 0.5
\end{aligned} \tag{63}
$$

$$
\begin{aligned}
\left|0.6 - F(y_3^-)\right| &\leq 0.1 \\
0.5 \leq F(y_3^-) &\leq 0.7 \\
\left|0.8 - F(y_3^+)\right| &\leq 0.1 \\
0.7 \leq F(y_3^+) &\leq 0.9 \rightarrow \\
F(y_4) &= 0.7
\end{aligned} \tag{64}
$$

$$
\begin{aligned}
\left|0.8 - F(y_3^-)\right| &\leq 0.1 \\
0.7 \leq F(y_3^-) &\leq 0.9 \\
\left|1 - F(y_3^+)\right| &\leq 0.1 \\
0.9 \leq F(y_3^+) &\leq 1.1 \rightarrow \\
F(y_5) &= 0.9
\end{aligned} \tag{65}
$$

Now we must prove that if $F(y_1) = 0.1$, $F(y_2) = 0.3$, $F(y_3) = 0.5$, $F(y_4) = 0.7$, and $F(y_5) = 0.9$ then $D_n = 0.1$.

$$
D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)| = \begin{cases}
|0 - 0.1| = 0.1 & x \rightarrow y_1^- \\
|0.2 - 0.1| = 0.1 & x \rightarrow y_1^+ \\
|0.2 - 0.3| = 0.1 & x \rightarrow y_2^- \\
|0.4 - 0.3| = 0.1 & x \rightarrow y_2^+ \\
|0.4 - 0.5| = 0.1 & x \rightarrow y_3^- \\
|0.6 - 0.5| = 0.1 & x \rightarrow y_3^+ \\
|0.6 - 0.7| = 0.1 & x \rightarrow y_4^- \\
|0.8 - 0.7| = 0.1 & x \rightarrow y_4^+ \\
|0.8 - 0.9| = 0.1 & x \rightarrow y_5^- \\
|1 - 0.9| = 0.1 & x \rightarrow y_5^+
\end{cases} \tag{66}
$$

As you can see in each situation the value for D_n will be 0.1.

7.2 Under the conditions of the question...

We must prove if $D_n \leq 0.2$ then $F(y_1) \leq 0.2 \leq F(y_2) \leq 0.4 \leq F(y_3) \leq 0.6 \leq F(y_4) \leq 0.8 \leq F(y_5)$.From previous part we must show following facts:

$$
\begin{aligned}
\left|F(y_1^-)\right| &\leq 0.2 \\
-0.2 \leq F(y_1^-) &\leq 0.2 \\
\left|0.2 - F(y_1^+)\right| &\leq 0.2 \qquad (67) \\
0 \leq F(y_1^+) &\leq 0.4 \qquad (68)
\end{aligned}
$$

So if we intersect between these two intervals we have $F(y_1) \in [0, 0.2]$.Similary we have:

$$
\begin{aligned}
\left|0.2 - F(y_2^-)\right| &\leq 0.2 \\
0 \leq F(y_2^-) &\leq 0.4 \\
\left|0.4 - F(y_2^+)\right| &\leq 0.2 \quad (69) \\
0.2 \leq F(y_2^+) &\leq 0.6 \quad (70)
\end{aligned}
$$

Intersect between two intervals we have $F(y_2) \in [0.2, 0.4]$.

$$
\begin{aligned}
\left|0.4 - F(y_3^-)\right| &\leq 0.2 \\
0.2 \leq F(y_3^-) &\leq 0.6 \\
\left|0.6 - F(y_3^+)\right| &\leq 0.2 \quad (71) \\
0.4 \leq F(y_3^+) &\leq 0.8 \quad (72)
\end{aligned}
$$

Intersect between two intervals we have $F(y_3) \in [0.4, 0.6]$.

$$
\begin{aligned}
\left|0.6 - F(y_4^-)\right| &\leq 0.2 \\
0.4 \leq F(y_4^-) &\leq 0.8 \\
\left|0.8 - F(y_4^+)\right| &\leq 0.2 \quad (73) \\
0.6 \leq F(y_4^+) &\leq 1 \quad (74)
\end{aligned}
$$

Intersect between two intervals we have $F(y_4) \in [0.6, 0.8]$.

$$
\begin{aligned}
\left|0.8 - F(y_5^-)\right| &\leq 0.2 \\
0.6 \leq F(y_5^-) &\leq 1 \\
\left|1 - F(y_5^+)\right| &\leq 0.2 \quad (75) \\
0.8 \leq F(y_5^+) &\leq 1.2 \quad (76)
\end{aligned}
$$

Intersect between two intervals we have $F(y_5) \in [0.8, 1]$. So we prove if $D_n \leq 0.2$ then $F(y_1) \leq 0.2 \leq F(y_2) \leq 0.4 \leq F(y_3) \leq 0.6 \leq F(y_4) \leq 0.8 \leq F(y_5)$.But we are not finished if $F(y_1) \leq 0.2 \leq F(y_2) \leq 0.4 \leq F(y_3) \leq 0.6 \leq F(y_4) \leq 0.8 \leq F(y_5)$ then D_n must be less or equal to 0.2:

$$
D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)| = \begin{cases}
0 \leq F(y_1) \leq 0.2 \rightarrow |0 - F(y_1)| \leq 0.2 & \text{if } x \rightarrow y_1^- \\
0 \leq F(y_1) \leq 0.2 \rightarrow |0.2 - F(y_1)| \leq 0.2 & \text{if } x \rightarrow y_1^+ \\
0.2 \leq F(y_2) \leq 0.4 \rightarrow |0.2 - F(y_2)| \leq 0.2 & \text{if } x \rightarrow y_2^- \\
0.2 \leq F(y_2) \leq 0.4 \rightarrow |0.4 - F(y_2)| \leq 0.2 & \text{if } x \rightarrow y_2^+ \\
0.4 \leq F(y_3) \leq 0.6 \rightarrow |0.4 - F(y_3)| \leq 0.2 & \text{if } x \rightarrow y_3^- \\
0.4 \leq F(y_3) \leq 0.6 \rightarrow |0.6 - F(y_3)| \leq 0.2 & \text{if } x \rightarrow y_3^+ \\
0.6 \leq F(y_4) \leq 0.8 \rightarrow |0.6 - F(y_4)| \leq 0.2 & \text{if } x \rightarrow y_4^- \\
0.6 \leq F(y_4) \leq 0.8 \rightarrow |0.8 - F(y_4)| \leq 0.2 & \text{if } x \rightarrow y_4^+ \\
0.8 \leq F(y_5) \leq 1 \rightarrow |0.8 - F(y_5)| \leq 0.2 & \text{if } x \rightarrow y_5^- \\
0.8 \leq F(y_5) \leq 1 \rightarrow |1 - F(y_5)| \leq 0.2 & \text{if } x \rightarrow y_5^+
\end{cases} \quad (77)
$$

Now we prove in each situation D_n is less or equal to 0.2.

8 Problem 8

8.1 Uniform Distribution Hypothesis

In order to use Kolmogorov-Smirnov test we shall sort the data in increasing order.

Table 13: Real data

0.42	0.06	0.88	0.40	0.90
0.38	0.78	0.71	0.57	0.66
0.48	0.35	0.16	0.22	0.08
0.11	0.29	0.79	0.75	0.82
0.30	0.23	0.01	0.41	0.09

Sort the 25 data in table 14:

Table 14: Sorted data

0.01	0.06	0.08	0.09	0.11
0.16	0.22	0.23	0.29	0.3
0.35	0.38	0.4	0.41	0.42
0.48	0.57	0.66	0.71	0.75
0.78	0.79	0.82	0.88	0.9

From Uniform distribution we must calculate each x_i in table above:

Table 15: Uniform distribution for each data

0.01	0.06	0.08	0.09	0.11
0.16	0.22	0.23	0.29	0.3
0.35	0.38	0.4	0.41	0.42
0.48	0.57	0.66	0.71	0.75
0.78	0.79	0.82	0.88	0.9

Now lets calculate the D_n where:

$$D_n = \max_{1 \le i \le n} \left| \frac{i}{n} - F(x_i) \right| \tag{78}$$

Derive the absolute value from tables 14 and 15.

Table 16: Calculate absolute values

0.03	0.02	0.04	0.07	0.09
0.08	0.06	0.09	0.07	0.1
0.09	0.1	0.12	0.15	0.18
0.16	0.11	0.06	0.05	0.05
0.06	0.09	0.1	0.08	0.1

Maximum valu of for D_n is 0.18 therefore as we know critical value for 0.05 rejection area is $D_{\text{crit},0.05} = \frac{1.36}{\sqrt{n}}$ so we have:

$$D_{\text{crit},0.05} = \frac{1.36}{5} = 0.272 \tag{79}$$

Cause of 0.18 is less than $D_{\text{crit},0.05}$ so the null hypothesis will not rejected.

8.2 Continuous Distribution Hypothesis

As we know in table 14 we have sorted data and we must know the CDF[1] of given PDF[2].

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{3}{2}x & \text{for } 0 < x \leq \frac{1}{2} \\ \frac{2}{4} + \frac{1}{2}x & \text{for } \frac{1}{2} < x < 1 \\ 1 & \text{for } 1 \leq x \end{cases} \tag{80}$$

Now lets calculate real CDF for each x_i in sorted data in table 14:

Table 17: CDF of each data

0.015	0.09	0.12	0.135	0.165
0.24	0.33	0.345	0.435	0.45
0.525	0.57	0.6	0.615	0.63
0.72	0.785	0.83	0.855	0.875
0.89	0.895	0.91	0.94	0.95

Now lets fill the table of absolute value of $F_n(x) - F(x)$:

Table 18: Absolute values

0.025	0.01	0.	0.025	0.035
0.	0.05	0.025	0.075	0.05
0.085	0.09	0.08	0.055	0.03
0.08	0.105	0.11	0.095	0.075
0.05	0.015	0.01	0.02	0.05

Maximum value for D_n is 0.11 and for $D_{\text{crit},0.05} = 0.272$ so it wont be rejected.

[1]Cumulative Density Function
[2]Probability Density Function

8.3 Posterior Probability Assessment

Firstly we are going to calculate the likelyhood of each given pdf as follows:

$$\begin{aligned} L(\text{given pdf}) &= \prod_{i=1}^{25} f(x_i) \\ &= 1.282892256975174 \\ L(\text{Uniform}) &= 1 \end{aligned} \tag{81}$$

The marginal likelihood of the data is a sum of the likelihoods weighted by their prior probabilities:

$$\begin{aligned} P(\text{data}) &= L(\text{Uniform})P(\text{Unifrom}) + L(\text{given pdf})P(givenpdf) \\ &= 1.282892256975174 * 0.5 + 1 * 0.5 \\ &= 1.141446128 \end{aligned} \tag{82}$$

The posterior probability that they were obtained from a uniform distribution:

$$\begin{aligned} P(\text{Uniform}|\text{data}) &= \frac{L(\text{Uniform})P(\text{Uniform})}{P(data)} \\ &= \frac{1 * \frac{1}{2}}{1.141446128} \\ &= 0.438040822 \end{aligned} \tag{83}$$

9 Problem 9

9.1 Comparison of Two Distributions

As we know we merged data and sorting them like table below 19:

Table 19: Calculate the CDF for each observations

data	-2.47	-1.73	-1.28	-0.82	-0.74	-0.71	-0.56	-0.4	-0.39
$F(x)$	0.04	0.08	0.12	0.16	0.2	0.2	0.24	0.28	0.32
$F(y)$	0	0	0	0	0	0.05	0.05	0.05	0.05
data	-0.37	-0.32	-0.30	-0.27	-0.06	0.00	0.05	0.06	0.20
$F(x)$	0.32	0.36	0.36	0.36	0.40	0.40	0.44	0.48	0.48
$F(y)$	0.1	0.1	0.15	0.2	0.2	0.25	0.25	0.25	0.3
data	0.29	0.31	0.36	0.38	0.44	0.51	0.52	0.59	0.61
$F(x)$	0.52	0.56	0.56	0.56	0.56	0.6	0.6	0.64	0.68
$F(y)$	0.3	0.3	0.35	0.4	0.45	0.45	0.50	0.50	0.50
data	0.64	0.66	0.7	0.96	1.05	1.06	1.09	1.31	1.38
$F(x)$	0.72	0.72	0.72	0.72	0.76	0.80	0.84	0.88	0.88
$F(y)$	0.50	0.55	0.6	0.65	0.65	0.65	0.65	0.65	0.70
data	1.5	1.56	1.64	1.66	1.77	2.2	2.31	2.36	3.29
$F(x)$	0.88	0.88	0.92	0.92	0.96	0.96	0.96	1.00	1.00
$F(y)$	0.75	0.80	0.80	0.85	0.85	0.90	0.95	0.95	1.00

Now lets calculate the absolute value of differences between $F(X)$ and $F(Y)$ in table 20:

Table 20: Calculate the CDF for each observations

data	-2.47	-1.73	-1.28	-0.82	-0.74	-0.71	-0.56	-0.4	-0.39
$F(X) - F(Y)$	0.04	0.08	0.12	0.16	0.2	0.15	0.19	0.23	0.27
data	-0.37	-0.32	-0.30	-0.27	-0.06	0.00	0.05	0.06	0.20
$F(X) - F(Y)$	0.22	0.26	0.21	0.16	0.20	0.15	0.19	0.23	0.18
data	0.29	0.31	0.36	0.38	0.44	0.51	0.52	0.59	0.61
$F(X) - F(Y)$	0.22	0.26	0.21	0.16	0.11	0.15	0.10	0.14	0.18
data	0.64	0.66	0.7	0.96	1.05	1.06	1.09	1.31	1.38
$F(X) - F(Y)$	0.22	0.17	0.12	0.07	0.11	0.15	0.19	0.23	0.18
data	1.5	1.56	1.64	1.66	1.77	2.2	2.31	2.36	3.29
$F(X) - F(Y)$	0.13	0.08	0.12	0.07	0.11	0.06	0.01	0.05	0.0

The maximum absolute value for $F(X) - F(Y)$ is 0.27 and the critical region for $D_{\text{crit},\alpha=0.05}$ is

$$\begin{aligned} D_{\text{crit},0.05} &= 1.36\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \\ &= 1.36\sqrt{\frac{1}{25} + \frac{1}{20}} \\ &= 0.408 \end{aligned} \tag{84}$$

As you can see if significant level was $\alpha = 0.05$ then null hypothesis won't be rejected since $0.27 < 0.408$.

9.2 Shifted Distribution Comparison

After shift X by 2 and merged the data with Y we can reach table 21:

Table 21: Calculate the CDF for each observations

data	-0.71	-0.47	-0.37	-0.3	-0.27	0.0	0.2	0.27	0.36
$F(X)$	0	0.04	0.04	0.04	0.04	0.04	0.04	0.08	0.08
$F(Y)$	0.05	0.05	0.1	0.15	0.2	0.25	0.3	0.3	0.35
data	0.38	0.44	0.52	0.66	0.7	0.72	0.96	1.18	1.26
$F(X)$	0.08	0.08	0.08	0.08	0.08	0.12	0.12	0.16	0.2
$F(Y)$	0.40	0.45	0.50	0.55	0.6	0.6	0.65	0.65	0.65
data	1.38	1.44	1.5	1.56	1.6	1.61	1.66	1.68	1.94
$F(X)$	0.2	0.24	0.24	0.24	0.28	0.32	0.32	0.36	0.40
$F(Y)$	0.70	0.70	0.75	0.80	0.80	0.80	0.85	0.85	0.85
data	2.05	2.06	2.2	2.29	2.31	2.31	2.51	2.59	2.61
$F(X)$	0.44	0.48	0.48	0.52	0.56	0.56	0.6	0.64	0.68
$F(Y)$	0.85	0.85	0.90	0.90	0.90	0.95	0.95	0.95	0.95
data	2.64	3.05	3.06	3.09	3.29	3.31	3.64	3.77	4.36
$F(X)$	0.72	0.76	0.80	0.84	0.84	0.88	0.92	0.96	1.00
$F(Y)$	0.95	0.95	0.95	0.95	1.00	1.00	1.00	1.00	1.00

From table 21 calculate the $F(X) - F(Y)$:

Table 22: Calculate the CDF for each observations

data	-0.71	-0.47	-0.37	-0.3	-0.27	0.0	0.2	0.27	0.36
$F(X) - F(Y)$	0.05	0.01	0.06	0.11	0.16	0.21	0.26	0.22	0.27
data	0.38	0.44	0.52	0.66	0.7	0.72	0.96	1.18	1.26
$F(X) - F(Y)$	0.32	0.37	0.42	0.47	0.52	0.48	0.53	0.49	0.45
data	1.38	1.44	1.5	1.56	1.6	1.61	1.66	1.68	1.94
$F(X) - F(Y)$	0.5	0.46	0.51	0.56	0.52	0.48	0.53	0.49	0.45
data	2.05	2.06	2.2	2.29	2.31	2.31	2.51	2.59	2.61
$F(X) - F(Y)$	0.41	0.37	0.42	0.38	0.34	0.39	0.35	0.31	0.27
data	2.64	3.05	3.06	3.09	3.29	3.31	3.64	3.77	4.36
$F(X) - F(Y)$	0.23	0.19	0.15	0.11	0.16	0.12	0.08	0.04	0.0

The maximum value for $F(X) - F(Y)$ is 0.56 and for $D_{\text{crit},0.05}$ we have:

$$\begin{aligned} D_{\text{crit},0.05} &= 1.36\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \\ &= 1.36\sqrt{\frac{1}{25} + \frac{1}{20}} \\ &= 0.408 \end{aligned} \tag{85}$$

Cause to 0.56 is greater than 0.408 the null hypothesis will be rejected.

9.3 Scaled Distribution Comparison

We must consider the conditions from earlier questions so we must see if $X+2$ and $3Y$ has same distributions or not.First step is calculate each $F(X)$ and $F(Y)$ as you can see in the table 23:

Table 23: Calculate the CDF for each observations

data	-2.13	-1.12	-0.90	-0.81	-0.47	0.0	0.27	0.60	0.72
$F(X)$	0	0	0	0	0.04	0.04	0.08	0.08	0.12
$F(Y)$	0.05	0.1	0.15	0.2	0.2	0.25	0.25	0.3	0.3
data	1.08	1.14	1.18	1.26	1.32	1.44	1.56	1.6	1.61
$F(X)$	0.12	0.12	0.16	0.2	0.2	0.24	0.24	0.28	0.32
$F(Y)$	0.35	0.40	0.40	0.40	0.45	0.45	0.50	0.50	0.50
data	1.68	1.94	1.98	2.05	2.06	2.10	2.29	2.31	2.51
$F(X)$	0.36	0.40	0.40	0.44	0.48	0.48	0.52	0.56	0.6
$F(Y)$	0.50	0.50	0.55	0.55	0.55	0.6	0.6	0.6	0.6
data	2.59	2.6better1	2.64	2.88	3.05	3.06	3.09	3.31	3.64
$F(X)$	0.64	0.68	0.72	0.72	0.76	0.80	0.84	0.88	0.92
$F(Y)$	0.6	0.6	0.6	0.65	0.65	0.65	0.65	0.65	0.65
data	3.77	4.14	4.36	4.5	4.68	4.98	6.60	6.93	9.87
$F(X)$	0.96	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$F(Y)$	0.65	0.70	0.70	0.75	0.80	0.85	0.90	0.95	1.00

Afterward we must calculate the maximum absolute value by differentiate the $F(X)-F(Y)$ which it is calculated in table 24:

Table 24: Calculate the CDF for each observations

data	-2.13	-1.12	-0.90	-0.81	-0.47	0.0	0.27	0.60	0.72
$F(X)-F(Y)$	0.05	0.1	0.15	0.2	0.16	0.21	0.17	0.22	0.18
data	1.08	1.14	1.18	1.26	1.32	1.44	1.56	1.6	1.61
$F(X)-F(Y)$	0.23	0.28	0.24	0.20	0.25	0.21	0.26	0.22	0.18
data	1.68	1.94	1.98	2.05	2.06	2.10	2.29	2.31	2.51
$F(X)-F(Y)$	0.14	0.10	0.15	0.11	0.07	0.12	0.08	0.04	0.0
data	2.59	2.61	2.64	2.88	3.05	3.06	3.09	3.31	3.64
$F(X)-F(Y)$	0.04	0.08	0.12	0.07	0.11	0.15	0.19	0.23	0.27
data	3.77	4.14	4.36	4.5	4.68	4.98	6.60	6.93	9.87
$F(X)-F(Y)$	0.31	0.26	0.30	0.25	0.20	0.15	0.10	0.05	0.0

Eventually maximum value for $F(X)-F(Y)$ is 0.31 which is less than $D_{\text{crit},0.05}=0.408$ so the null hypothesis will not be rejected.

10 Problem 10

10.1 Creating Frequency Table and Visualization

a) Read CSV[3] with pandas library.

b) Created contingency table by method of crosstab in pandas library which can be see in table 25.

c) Save the data with contingency table to csv using pandas mehtods in sex_survive_table.

d) For mosaic plot we add statsmodels.graphics library and as you can see in the figure 12.

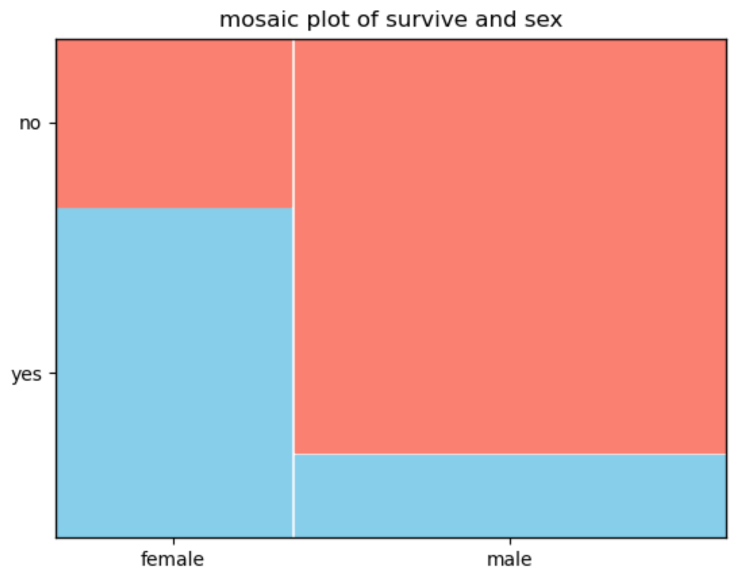

Figure 12: Mosaic plot for sex and survive

e) Contingency table show in table 25.

Table 25: Contingency Table for sex and survive

		Survive		All
		no	yes	
Sex	male	156	307	463
	female	708	142	850
All		864	449	1313

f) Genereating a better mosaic plot with more properties in figure 13:

[3]Comma-separated values

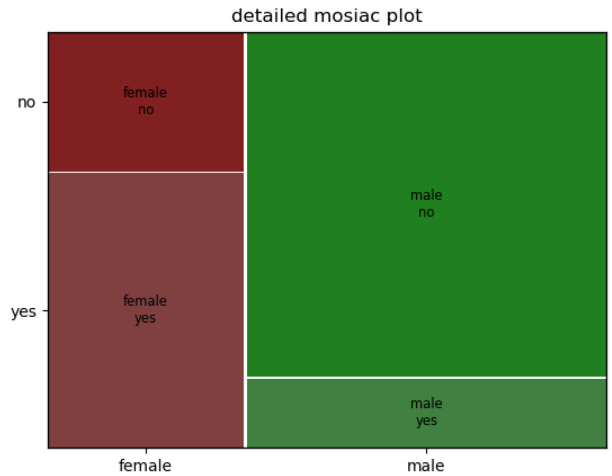

Figure 13: Detailed mosaic plot

10.2 χ^2 Contingency Test and Fisher's Exact Test

a) For conducting the χ^2 test, the null hypothesis is sex and survived are independent variables and alternative hypothesis is they are correlated.Therefore we use $\chi^2\_contingency$ function in python to calculate wanted things.

b) If significant level is 0.01 then calculated p-value for χ^2 test is:

$$p - value = 9.164113332735093e^{-73} \tag{86}$$

Therefore for even harder significant level of 0.01 the null hypothesis will be rejected, since the null hypothesis is rejected sex and survive variables are not independent and indeed they are correlated. For verifying the assumtions of χ^2 test is when we calculated the expected mean, each of them must be greater than 5.(For prove we can refer you to the table 28.)

c) we use fisher exact function from stats in scipy library:
Lets first explain the fisher exact test a little bit.Mainly fisher's exact test used for 2*2 contingency tables however χ^2 test's approximations might not be accurate for smaller sample sizes. Fisher's exact test calculating the probability of observing the particular of counts under the null hypothesis(probability is calculated by hypergeometric distribuitons).For example consider the table 26: In here probability is:

Table 26: Example for fisher's exact test

	Men	Women	Row Total
Studying	$\boldsymbol{a}$	$\boldsymbol{b}$	$a+b$
Non-studying	$\boldsymbol{c}$	$\boldsymbol{d}$	$c+d$
Column Total	$a+c$	$b+d$	$a+b+c+d(=n)$

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} \tag{87}$$

The advantages of using fisher's exact test:

- Regardless of the size of data is more accurate than χ^2 test.
- It is not affected by sample size.

d) The conclusion of χ^2 test is in the table 27

Table 27: Result of χ^2 test

χ^2 statistics	p-value	degree of freedom
325.50	$9.16e^{-73}$	1

Expected frequencies are in table 28: Since p-value is less than significant level 0.01 then null hypothesis

Table 28: Expected frequecnies

Expected		
	304.67	158.32
	559.32	290.67

would be rejected.

e) P-value of fisher's exact test is $5.187445473452701e-73$ where is less than significant level 0.01 therefore also fisher's exact test will reject null hypothesis, and sex and survive are correleted.

11 Problem 11

1. True of false:

   a) True.Eventhough a sample can be in the rejection area but it has a chance of being true under null hypothesis.

   b) False,Since the large and noteworthy number still can be not one that is statisticaly significant.More presicely statisticaly significance is determined by hypothesis testing

   c) True, Based on 5% significance level p-value of 4.7 would lead to the rejection of the null hypothesis,however 5.2% will not.

2. Which of the following questions does a test of significance deal with?

   a) True.Like previous section.

   b) This relates to the practical significance or effect size, which is different from statistical significance.

   c) Statistical tests do not prove anything with certainty; they only provide evidence against the null hypothesis.

   d) The design of the experiment is assessed before the data collection and analysis; it's not something that a statistical test can determine.

3. False.since if we assume that z-score will calculate the following formula:

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

   Then for the investigator with 900 sample size, the z-score might be higher since he can obtain more evidence against the null hypothesis.So the p-value is related to null hypothesis mean(50 in here) and deviation and aslo sample size.

4. Even if all companies hire without regard to race, some companies will, by chance, end up with proportions that deviate from the population proportion of 10%. If the sample size is large enough, these deviations could be statistically significant according to the z-test. This is because statistical tests like the z-test are based on probabilities and are influenced by sampling variability. A statistically significant result does not imply intent or causation; it only indicates that the observed result is unlikely to have occurred by random chance alone, given the assumptions of the test.

12 Problem 12

12.1 What is the likelyhood ...

We must calculate the type I error as follow:

$$\begin{aligned} P_{H_0}(X \leq 40 \cup 60 \leq X) &= P_{H_0}(X \leq 40) + 1 - P_{H_0}(X \leq 59) \\ &= 0.02844 + 1 - 0.97156 \\ &= 0.05688 \end{aligned} \tag{88}$$

12.2 Equivalent

From previous section we obtained the probability of rejection null hypothesis eventhough it is valid either less than 40 or bigger than 60 heads is 56.88%.So the equivalent α for decision rule is 56.88%.

12.3 Programming Part

We plot the two binomial distributions as you can see in the figure 14:

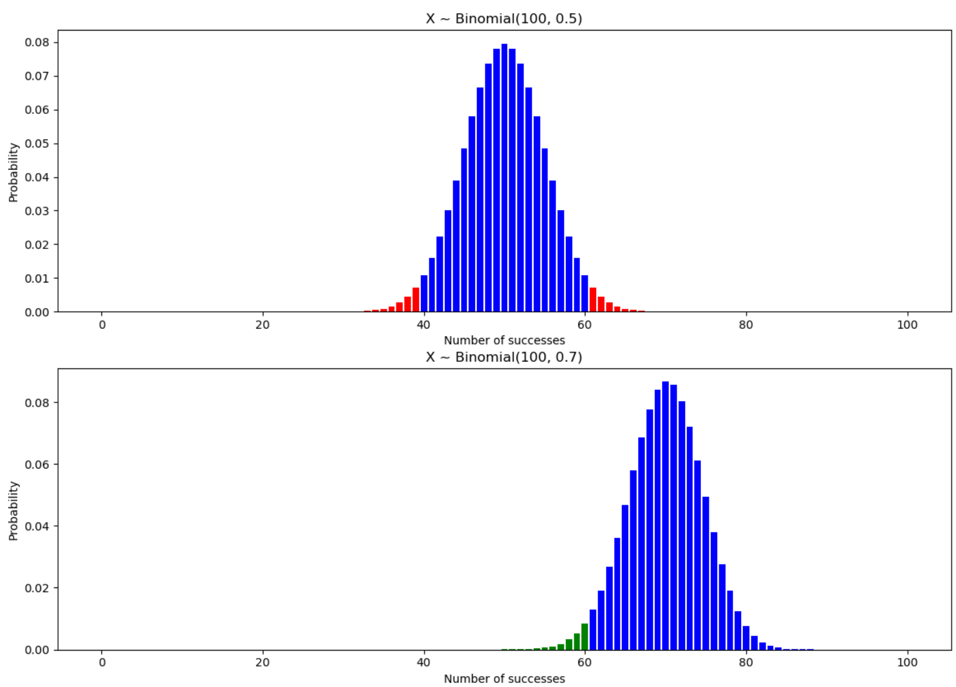

Figure 14: Binomial distribution for $p = 0.5$ and $p = 0.7$

As you can see in figure 14 red region is for type I error and green area is for type II error.

The likelyhood of adopting the null hypothesis is:

$$\begin{aligned} P_{H_0}(41 \leq X \leq 59) &= P_{H_0}(X \leq 59) - P_{H_0}(X \leq 40) \\ &= 0.9431120663590191 \end{aligned} \tag{89}$$

And strength of the test is:

$$P_{H_1}(61 \leq X) = 0.9790114239960752 \tag{90}$$

13 Problem 13

13.1 Which statitical test

Since we have 2 samples from 2 different populations and we want to determine if the mean scores at the two schools differ significantly so the test which suit this is differecne between means.

13.2 Test whether the difference is significant, with $\alpha = 0.05$

Applying the likelyhood ratio technique, we will arrive at a test based on $\overline{x}_1 - \overline{x}_2$ and since our n_i is greater than 30 we can use Z distribution for null hypothesis. Our null hypothesis is:

$$H_0 : \mu_1 - \mu_2 = \delta = 0$$

$$H_1 : \mu_1 - \mu_2 \neq \delta$$

$$\begin{aligned} z &= \frac{\overline{x}_1 - \overline{x}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{70 - 74}{\sqrt{\frac{16}{50} + \frac{36}{65}}} \\ &= \frac{-4}{0.934797387} \\ &= -4.279002119 \end{aligned} \tag{91}$$

So if $Z_{\alpha=0.025} = 1.96$ therefore if $z \leq -1.96$ or $z \geq 1.96$ then it would be rejected and cause of $z = -4.279002119$ then our null hypothesis will be rejected.

14 Problem 14

14.1 Observational or controlled

We did not control the pupils to do such a thing but we observe players and non-players and did an experiment on them.More precisely researcher did not mainpulate any variables and just observed subjects in their enviroment.Also in a controlled experminet the researcher would randomly assign participant to groups and control for other variables that might influnce the outcomes. So if we wrap all things up this is a observational study.

14.2 Mean Differences

We know standard deviation for each groups which is 20, 50 respectively. The average score for those play vidio games is 120 while for non-players gropus is 100.

$$\begin{aligned}(\mu, \sigma)_{\text{players}} &= (120, 20) \\ (\mu, \sigma)_{\text{non-players}} &= (100, 50)\end{aligned} \tag{92}$$

Form null hypothesis:

$$\begin{aligned}H_0 &: \mu_{\text{player}} = \mu_{\text{non-players}} \\ H_1 &: \mu_{\text{player}} \neq \mu_{\text{non-players}}\end{aligned} \tag{93}$$

Since the number of each samples are not big enough to use CLT[4] to use z score we can use t-test for this:

$$t = \frac{\overline{x}_1 - \overline{x}_2 - \delta}{sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{94}$$

where sp is s pool:

$$sp^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{95}$$

Lets calculate the sp at first step:

$$\begin{aligned}sp &= \sqrt{\frac{19 * 50^2 + 14 * 20^2}{20 + 15 - 2}} \\ &= \sqrt{\frac{53100}{33}} \\ &= 40.113475405\end{aligned} \tag{96}$$

Obtain t from equation 94:

$$\begin{aligned}t &= \frac{120 - 100}{40.113475405 * \sqrt{\frac{1}{20} + \frac{1}{15}}} \\ &= \frac{20}{13.701360251} \\ &= 1.459709082\end{aligned} \tag{97}$$

[4] Centeral Limit Theorem

since degree of freedom for t is $n_1 + n_2 - 2 = 33$ and type I error is 0.05 and 0.01 t score will be ± 2.034515 and ± 2.733277 then null hypothesis is not rejected cause in each type I errors 1.46 is less than $+t_{\frac{\alpha}{2},33}$ and greater than $-t_{\frac{\alpha}{2},33}$.

14.3 Would a substantial difference ...?

For this particular experiment we can not say this indication will be true.Cause it is observational study. Becuase other factors (which is confunding variables) might influnce both vidio games and spatial aptitude, for example age, education, or inherent ability. Furthermore selection bias could be presenet.So a controlled experiment would be required to indicate the results.

15 Problem 15

15.1 How strong is a research ...

First lets calculate the d:

$$\begin{aligned} d &= \frac{n^{\frac{1}{2}}(\mu_A - \mu_0)}{\sigma} \\ &= \frac{7 * (0.5)}{7} \\ &= 0.5 \end{aligned} \tag{98}$$

So we can calculate the power of this research:

$$\begin{aligned} \text{power} &= \Phi(Z_\alpha + d) \\ &= \Phi(1.645 + 0.5) \\ &= \Phi(2.145) \\ &= 0.98402 \end{aligned} \tag{99}$$

Now lets calculate the typical p-value to determine this test would be good or not:

$$\begin{aligned} \text{Typical p-value} &= 1 - \Phi(d) \\ &= 1 - \Phi(0.5) \\ &= 1 - 0.69146 \\ &= 0.30854 \end{aligned} \tag{100}$$

It means that we have 30.85% to reject H_0.Speaking of strength of this research we can say that if we have more than 80% power then it is strong and calculated power is 98.4% so it is strong research.

15.2 Type II error

The type II error denoted by β is:

$$\begin{aligned} \beta &= 1 - \text{power} \\ &= 1 - 0.98402 \\ &= 0.015976 \end{aligned} \tag{101}$$

15.3 Find a sample size

For obtaining sample size for power of 0.99 we can use this formula:

$$
\begin{aligned}
\text{power} &= \Phi(Z_\alpha + d) \\
0.99 &= \Phi(1.645 + \frac{n^{\frac{1}{2}} * 0.5}{7}) \\
2.326 &= 1.645 + \frac{n^{\frac{1}{2}} * 0.5}{7} \\
0.681 * 14 &= \sqrt{n} \\
n &\approx 91
\end{aligned} \tag{102}
$$

So for getting 99% power sample size must be 91.

16 Problem 16

16.1 Find α

Given rejection area we can calulate the α as follow:

$$
\begin{aligned}
\alpha &= P(X < 40) + P(X > 60) \\
&= P(X \leq 39) + P(X \geq 61) \\
&= 0.0176 + 0.0176 \\
&= 0.0352
\end{aligned} \tag{103}
$$

16.2 Draw an approximation

For purpose of drawing an approximation of power digaram as a function of p, we assume that p for alternative hypothesis is a list of $[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$.
Now lets calculate the power of each test:

$$
\begin{aligned}
Power[i] &= 1 - F(60|p[i]) + F(39|p[i]) \\
&= [1, 0.97901, 0.4621, 0.4621, 0.97901, 1]
\end{aligned} \tag{104}
$$

So for $p = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]$, the power is $[1, 0.97901, 0.4621, 0.0352, 0.4621, 0.97901, 1]$ respectively and as you can see in figure 15, we draw and approximation of power as a function of p:

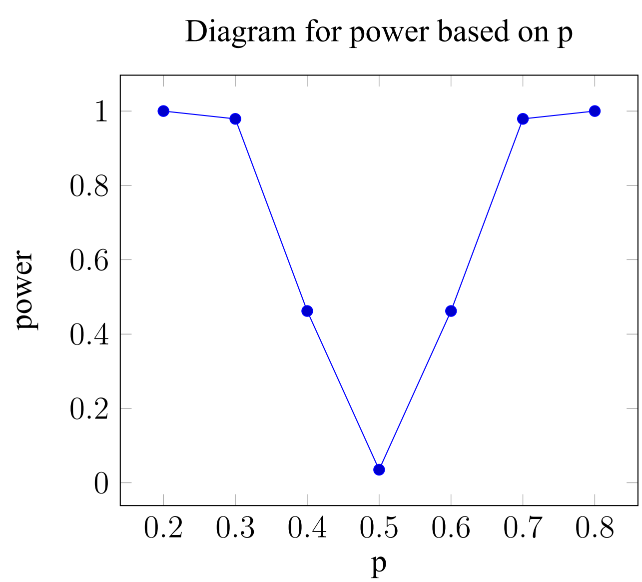

Figure 15: An approximation of power diagram

17 Problem 17

17.1 Likelyhood ratio

If our null hypothesis and alternative hypothesis was simple we could use Neyman-Pearson lemma and use likelyhood ratio but our alternative hypothesis is complex so we must use generalized likelyhood ratio as follow:

$$\Lambda^* = \frac{\max_{\theta \in \Theta_0} l(\theta)}{\max_{\theta \in \Theta} l(\theta)} \tag{105}$$

$$\Lambda = \min \{\Lambda^*, 1\} \tag{106}$$

Cause our null hypothesis is simple then the generalized likelyhood would be changed to:

$$\Lambda^* = \frac{l(\theta_0)}{\max_{\theta \in \Theta} l(\theta)} \tag{107}$$

We can use maximum likelyhood as follow:

$$\Lambda^* = \frac{l(\theta_0)}{l(\theta_{mle})} \tag{108}$$

Now lets calculate the maximum likelyhood for the exponential distribution:

$$\begin{aligned} l(\theta_{mle}) &= \prod_{i=1}^{n} \theta_{mle} e^{-\theta_{mle} x_i} \\ &= \theta_{mle}^n e^{-\theta_{mle} \sum_{i=1}^{n} x_i} \end{aligned} \tag{109}$$

Using log of l for simplicity:

$$\begin{aligned} \log l(\theta_{mle}) &= n \log(\theta_{mle}) - \theta_{mle} \sum_{i=1}^{n} x_i \\ &= n \log(\theta_{mle}) - n\theta_{mle}\overline{x} \\ \frac{\partial \log l(\theta_{mle})}{\partial \theta} &= \frac{n}{\theta_{mle}} - n\overline{x} = 0 \\ \theta_{mle} &= \frac{1}{\overline{x}} \end{aligned} \tag{110}$$

Obtain Λ^* from MLE[5]

$$\begin{aligned} \Lambda^* &= \frac{\prod_{i=1}^{n} \theta_0 e^{-\theta_0 x_i}}{\prod_{i=1}^{n} (\frac{1}{\overline{x}}) e^{-\frac{x_i}{\overline{x}}}} \\ &= \overline{x}^n \theta_0^n e^{n - \theta_0 \sum_{i=1}^{n} x_i} \\ &= \overline{x}^n \theta_0^n e^{n(1-\theta_0 \overline{x})} \end{aligned} \tag{111}$$

Assume that $\overline{x}\theta_0 = R$ then we have:

$$\Lambda^* = R^n e^{n(1-R)} \tag{112}$$

[5] Maximum Likelyhood Estimator

17.2 Critical Region

From previous part we know if R grows then Λ^* will gets lower and inversely if R gets lower then Λ will gets higer so the test statistic will be $\overline{x}$ and it would be a normal distribution cause of CLT so we can use z score to calculate the critical region as you can see in figure 16.

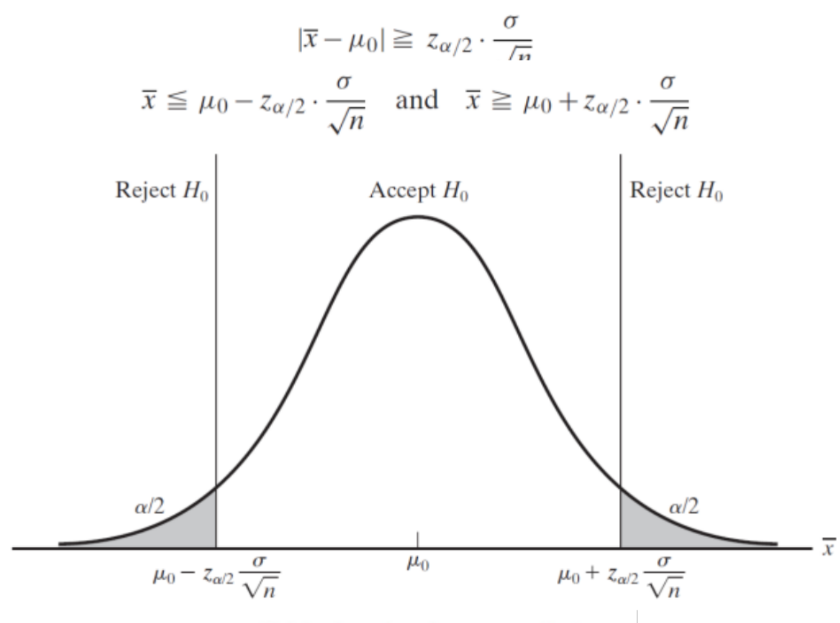

Figure 16: Critical Region for $\overline{x}$

where $\mu_0 = \frac{1}{\theta}$ and $\sigma = \sqrt{\frac{1}{\theta}}$.

18 Problem 18

18.1 Assuming $\alpha = 0.05$, what conclusions can you draw from this sample?

Since the mean and variance are unknown we can use estimated $\overline{x}$ and S for using the t-distribution however cuase our sample is higher than 30 we can use standard normal distribution as follow:(Note that we assumed this sample are given by nomral distribution eventhough σ^2 is unknow we can approximate its value with S^2)

$$Z \sim \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{113}$$

From 0.05 for type I error and alternative hypothesis we can use one tail to reject the null hypothesis if $z - score < z_{(\alpha)}$.

$$\begin{aligned} z &= \frac{25.9 - 28}{\frac{5.6}{\sqrt{50}}} \\ &= \frac{-2.1}{0.791959595} \\ &= -2.651650429 \end{aligned} \tag{114}$$

Now must see if this value is less than $-Z_\alpha$ then it will be rejected otherwise we can not reject it.For $\alpha = 0.05$ for left tail we have -1.645 and cause $-2.65 < -1.645$ then our null hypothesis would be rejected.
So under the normality assumption, the null hypothesis would be rejected due to our test statistic and critical region.

18.2 Type II error

We need to caluclulate the β, at first lets illustrate where the β is when one tail(left tail) is considered:

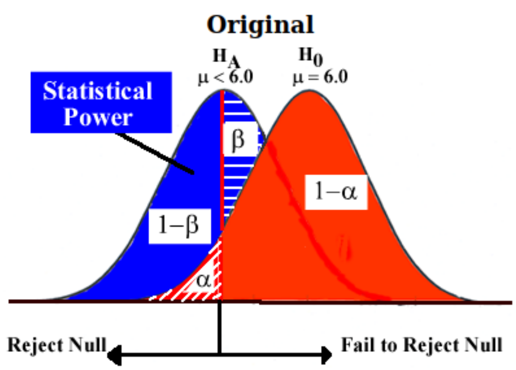

Figure 17: Illustration of type I and II errors

As you can see in the figure 17 for a simple null hypothesis and alterantive H_A which is one tailed we have power on left blue part so we can say that:

$$\begin{aligned} \beta &= P_{H_1}(Z > z_\alpha) \\ &= P_{H_1}(Z - d > z_\alpha - d) \\ &= 1 - \Phi(z_\alpha - d) \end{aligned} \tag{115}$$

Under H_0, $Z \sim N(0,1)$ and under H_1 $Z \sim N(d,1)$ where d is:

$$\begin{aligned} d &= \frac{\mu - \mu_0}{\frac{S}{\sqrt{n}}} \\ &= \frac{27 - 28}{\frac{5.6}{\sqrt{50}}} \\ &= \frac{-1}{0.791959595} \\ &= -1.262690681 \end{aligned} \tag{116}$$

Now we can calculate the type II error as follow:

$$\begin{aligned} \beta &= 1 - \Phi(z_\alpha - d) \\ &= 1 - \Phi(-1.646 - (-1.262690681)) \\ &= 1 - \Phi(-0.382309319) \\ &= 1 - 0.35112 \\ &= 0.64888 \end{aligned} \tag{117}$$

So the β is 0.64888(careful maximum of β is 0.64888 cause we just know real mean is not greater than 27).

19 Problem 19

19.1 What conclution can be drawn?

The study suggests that the average duration for the surgical procedure in question is approximately 130 minutes, with a relatively small variation (standard deviation of 5 minutes). The 95% confidence interval obtained via bootstrap suggests that we can be 95% confident that the population mean of the surgery duration lies between 128.5 and 131.5 minutes

19.2 How can the bootstrap method be used to construct a confidence interval for the median?

To construct a confidence interval for the mean using the bootstrap method, you would repeatedly sample from the observed data with replacement (each sample is called a bootstrap sample), calculate the median for each bootstrap sample, and then determine the distribution of these means. The 95% confidence interval for the mean would be the range between the 2.5th and 97.5th percentiles of these bootstrap medians.

19.3 What is the advantage of using the bootstrap method over parametric methods in this study?

The bootstrap method does not assume that the data follows any specific parametric distribution (like normal distribution), which is an advantage in situations where the underlying distribution of the data is unknown or not well-modeled by parametric assumptions. It can be particularly useful when the sample size is small, or the data are skewed.

19.4 Is the bootstrap method always the best method for estimating parameters? Explain.

No, the bootstrap method is not always the best option. It is computationally intensive and may not perform well with very small sample sizes or with data that have complex dependencies. Parametric methods can be more efficient and can provide more accurate estimates when their assumptions are met.

19.5 What are bootstrap statistics and how are they calculated?

Bootstrap statistics are estimates of a population parameter (like mean, median, variance) that are obtained by repeatedly resampling the original data with replacement and calculating the statistic of interest. The resampling process creates many simulated samples (bootstrap samples), and the distribution of the statistic across these samples is used to estimate the true population parameter and its variability.
Conclution:
The bootstrap method is a powerful non-parametric tool used widely in statistics to assess the uncertainty of an estimator. The basic idea is to mimic the process of obtaining new sample data by resampling with replacement from the original dataset and to draw conclusions about the population from these simulated samples.

20 Problem 20

20.1 Generate sample size of 50

We generated a sample of size 50 from beta distribution with $\alpha = 2, \beta = 5$ with np.random.beta as you can see in the figure 18:

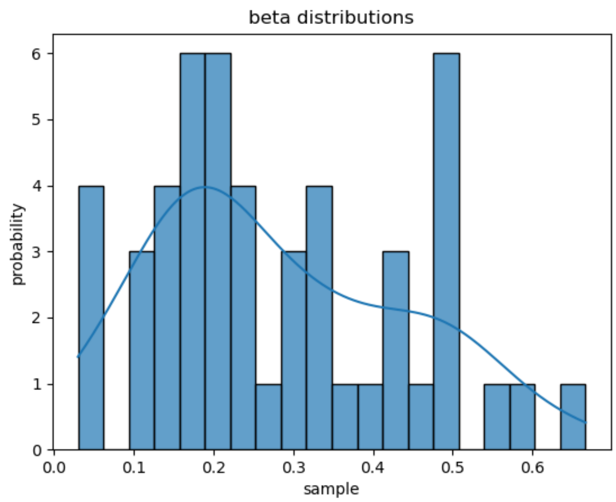

Figure 18: Generated 50 sample from beta distribution

20.2 Median test

Now we want to consider the median of the sample above as M. for test we consider the null hypothesis as follows:

$$\begin{aligned} H_0 &: M = 0.2 \\ H_1 &: M > 0.2 \end{aligned} \tag{118}$$

We can get help from sign test as follows:

- Calculate the differences $d_i = x_i - m_0$
- Test statistic S is the number of cases where $d_i > 0$
- If the null hypothesis H_0 is true, the test statistic follows binomial distribution with paramter n and 1/2.

Calculate the d_i and pass number of positive values to the binomtest in python and it returns the p-value for this test which is 0.016419568782134242 where for $\alpha = 0.05$ the null hypothesis will be rejected.

20.3 Calculate power

For calculating the power we can use 1000 simulation and for each sample calculate the p-value for that particular sample and see if it will be rejected or not. So the number of rejected samples over all of simulations is our power.Where for 1000 simulations the power is 0.678.

20.4 M_0 is 0.3

From changing the value of 0.2 to 0.3 we can obtain the p-value of 0.9835804312178658. cause we can not reject null hypothesis cause we can not accept alternative hypothesis so we retain null hypothesis 0.3 due to it is closer to real median. and the power test is close to zero(0.001).

20.5 Compare the results

We know the median of beta distribution $\alpha = 2, \beta = 5$ is close to 0.22 then null hypothesis M=0.2 is where powerful since the power of test was around 0.678 but M=0.3 is very week since power of test was close to zero(0.01).

21 Problem 21

21.1 What method do you suggest for testing the hypothesis

They are several methods to test following null hypothesis:

$$H_0 : \sigma_x^2 = \sigma_y^2$$
$$H_1 : \sigma_x^2 \neq \sigma_y^2 \tag{119}$$

Appropriate mehtods to test variance:

- F-Test for Equal Variances: This test is based on the ratio of the variances of the two samples. Since variance is the square of the standard deviation, this test can also be used to compare standard deviations. It assumes that the samples come from normally distributed populations.

- Levene's Test: Levene's test is used to assess the equality of variances for two or more groups. Although it is typically used for variances, it can also be adapted for standard deviations, as it tests for equal spreads.

- Bartlett's Test: Like the F-test, Bartlett's test compares variances and assumes that the data comes from normally distributed populations. It is sensitive to departures from normality.

21.2 Two normal distributions

We generated 50 sample from each normal distributions with $\sigma = 3$ and $\sigma = 10$ and the mean of each distributions is set to zero Which you can see in the figures 19 and 3:

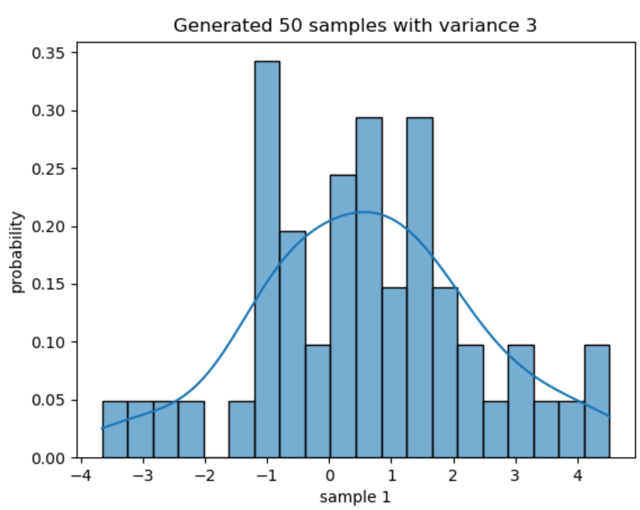

Figure 19: Generated samples from variance 3

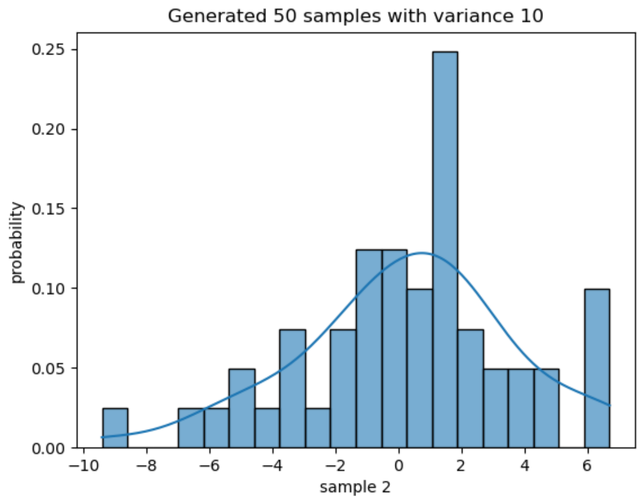

Figure 20: Generated samples from variance 10

Now lets conduct fisher and levene and bartlett test and see the p-values for each test in table 29:

Table 29: Table of analyse the tests

	fisher	levene	bartlett
p-value	8.285340600361369e-06	0.0009734143505761916	1.6615356894278033e-05
statistics	0.2790620843852304	11.565706537288632	18.54254765106779

Conclusion of each test is rejecting null hypothesis with 0.05 rejection area and type I error since their p-values is less than significant level.

22 Problem 22

22.1 Introduce statistics ...

We have observed frequencies from number of sons from 6115 families and for testing these data comes from binomial distribution we can consider goodness of fit as follow:

$$\begin{aligned} H_0 &: \text{Data comes from Binomial}(12, 0.5) \\ H_1 &: \text{Data does not come from Binomial}(12, 0.5) \end{aligned} \tag{120}$$

Steps:

- Calculate number of sons probability from binomial with paramters $(n, p) = (12, 0.5)$
- Obtain expected frequencies with following formula

$$\text{expected frequencies} = \text{probability} * \text{number of whole families}$$

- The statistic for this test is $\chi^2 = \sum_{i=1}^{m} \frac{(f_i - e_i)^2}{e_i}$
- Degree of freedom is:

$$df = m - t - 1$$

where m(13) is number of terms in the summation and t(2 paramters in binomial distribution) is the number of independent parameters estimated on the basis of the sample data

Now we create 2 functions which one of them calculate the statistic while other one calculate the expected frequencies which can be seen in table 30.

Table 30: Exepected frequencies

Number of sons	Exepected frequencies
0	1.49291992
1	17.91503906
2	98.53271484
3	328.44238281
4	738.99536133
5	1182.39257813
6	1379.45800781
7	1182.39257813
8	738.99536133
9	328.44238281
10	98.53271484
11	17.91503906
12	1.49291992

The χ^2 is 249.19 and critical region is $\chi^2_{\alpha, m-t-1} = \chi^2_{0.05,10}$ which is 18.30 so the null hypothesis will be rejected since $249.19 > 18.30$.
From built-in fucntions in python, we obtain same statistics and the p-value were around $2 * 10^{-46}$ where the null hypothesis will be rejected.

22.2 Using simulation, obtain the distribution of the introduced statistic and draw its histogram.

Now consider 5000 simulation which in each we will gets 6115 random numbers from binomial distribution and assume the expected frequncies is table ??.Next step for each sample we will calculate the χ^2 score and then plot its histogram which you can see in the figure 21:

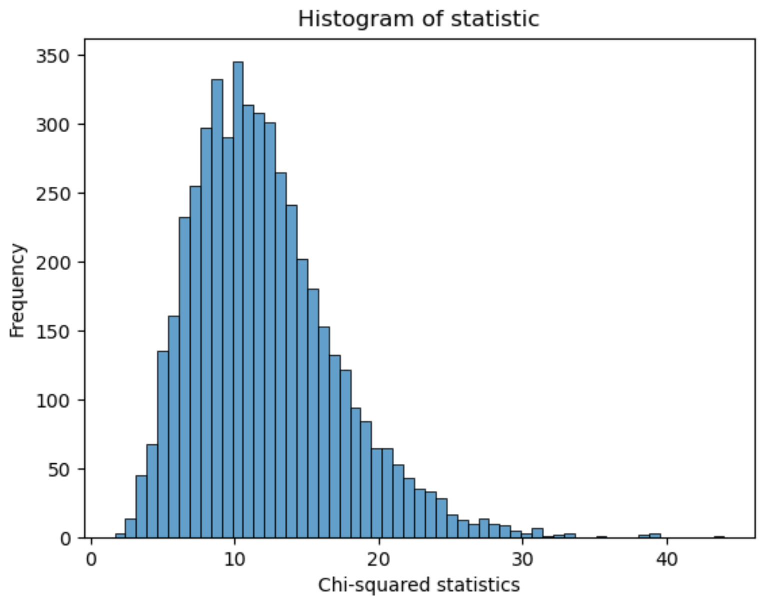

Figure 21: Histogram of the statistic

22.3 Based on the obtained distribution, calculate the p-value for the statistic calculated in part a. With $\alpha = 0.05$, is H0 rejected?

Based on obtain distribution of figure 21 we can easily calculate the p-value with numpy which we can get a summation if our simulatied statistic is greater equal to chi-score obtained in part $a)$ so the p-value is:

$$\text{p-value} = 0 \tag{121}$$

As you can see the p-value is 0 where definitely is in the rejection area with $\alpha = 0.05$ therefore the null hypothesis H_0 will be rejected.