MTH5129 Probability & Statistics II
Coursework 8

1. A computer scientist has developed an algorithm for generating pseudo-random integers $0, 1, \ldots, 9$. He codes the algorithm and generates 1000 pseudo-random digits. The data are as follows

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	94	93	112	101	104	95	100	99	108	94

Test whether there is evidence against the hypothesis that the digits are all equally likely.

Solution:

Null hypothesis H_0 : All digits equally likely

Alternative hypothesis $H_1 : \neg H_0$

The expected frequencies if H_0 is true are all 100.

Test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

degrees of freedom $\nu = 10 - 1 = 9$. $X^2 \sim \chi^2_9$ if H_0 is true.

Observed value of $X^2 = \frac{(94-100)^2}{100} + \cdots + \frac{(94-100)^2}{100} = 3.72$.

P value is $P(X^2 > 3.72)$, which can be computed in R as follows

```
> 1- pchisq(3.72, 9)
[1] 0.9288566
```

Thus there is no evidence against the hypothesis that all digits are equally likely.

2. The lifetime (in hours) of 500 batteries was recorded and is shown in the following frequency table.

Time	0-50	50-100	100-150	150-200	200-250
Frequency	218	117	70	35	25
Time	250-300	300-350	350-400	400+	
Frequency	18	11	6	0	

Test the hypothesis that the distribution of lifetimes follows an exponential distribution at the 5% significance level.

Solution:

Taking the mid-points of the intervals we find $\bar{y} = 91$. So $\hat{\lambda} = 1/91$.

So

$$P(0 < Y < 50) = \int_0^{50} \frac{1}{\hat{\lambda}} \exp(-\hat{\lambda} y) dy = 1 - \exp(-50\hat{\lambda}) = 0.4227$$

.

The other probabilities are found similarly so we have

Time	Probability	$E_i =$ Prob $\times 500$	O_i
0-50	0.4227	211.35	218
50-100	0.2440	122.00	117
100-150	0.1409	70.45	70
150-200	0.0813	40.65	35
200-250	0.0469	23.45	25
250-300	0.0271	13.55	18
300-350	0.0156	7.80	11
350-400	0.0090	4.50	6
400+	0.0123	6.15	0

To make all the expected frequencies larger than 5 we merge the last two classes to give 350+ with an $E_i = 10.65$ and $O_i = 6$.

Null hypothesis H_0 : Data are from an exponential distribution

Alternative hypothesis $H_1 : \neg H_0$

Test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

degrees of freedom $\nu = 8 - 1 - 1 = 6$. $X^2 \sim \chi^2_6$ if H_0 is true.

Observed value of $X^2 = \frac{(218-211.35)^2}{211.35} + \cdots + \frac{(6-10.65)^2}{10.65} = 6.109$.

The rejection region is $\{X^2 : X^2 > 12.59\}$.

Thus we can't reject H_0 at the 5% significance level. The data are compatible with coming from an exponential distribution.

3. 100 observations on a continuous random variable Y gave the following frequency table

Interval	$0-\pi/4$	$\pi/4-\pi/2$	$\pi/2-3\pi/4$	$3\pi/4-\pi$
Frequency	10	38	41	11

Test the hypothesis that Y has the pdf

$$f(y) = \begin{cases} \frac{1}{2}\sin y & 0 \leq y \leq \pi \\ 0 & \text{otherwise,} \end{cases}$$

using the 5% significance level.

Solution: We see that

$$\begin{aligned} P(0 < X \leq \frac{\pi}{4}) &= \int_0^{\pi/4} \frac{1}{2}\sin y dy \\ &= \left[-\frac{1}{2}\cos y\right]_0^{\pi/4} \\ &= \frac{1}{2}(1 - \frac{1}{\sqrt{2}}) \\ &= 0.146. \end{aligned}$$

Similarly $P(\frac{\pi}{4} < Y \leq \frac{\pi}{2}) = 0.354, P(\frac{\pi}{2} < Y \leq \frac{3\pi}{4}) = 0.354$ and $P(\frac{3\pi}{4} < Y \leq \pi) = 0.146$. So we have the following table

x	O_i	E_i	$\frac{(O-E)^2}{E}$
$0-\pi/4$	10	14.6	1.45
$\pi/4-\pi/2$	38	35.4	0.19
$\pi/2-3\pi/4$	41	35.4	0.89
$3\pi/4-\pi$	11	14.6	0.89
Total			3.42

Null hypothesis H_0 : The data are from this distribution

Alternative hypothesis $H_1 : \neg H_0$

Test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

degrees of freedom $\nu = 4 - 1 = 3$. $X^2 \sim \chi_3^2$ if H_0 is true.

Observed value of $X^2 = 3.42$.

We reject H_0 if $X^2 > 7.815$. Thus we cannot reject H_0 at the 5% significance level. The data are compatible with coming from this idstribution.

4. Five dice were thrown 150 times and the number of sixes was recorded. The data are given in the following table.

No. of sixes	0	1	2	3	4	5
Frequency	46	63	23	12	5	1

We want to know if there is any evidence that the dice are not fair. Compute the p-value.

Solution: If the dice are fair then the number of sixes will have a binomial distribution with $n = 5$ and $p = 1/6$. We test the null hypothesis H_0 that the data come from this distribution against an alternative that they don't. The probabilities and expected frequencies are shown below where I have merged the expected and observed frequency for scores three to five so that the expected number is more than 5.

No. of sixes	0	1	2	3	4	5
Probability	0.4019	0.4019	0.1608	0.0322	0.0032	0.0001
E_i	60.3	60.3	24.1		5.3	
O_i	46	63	23		18	

Test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

degrees of freedom $\nu = 4 - 0 - 1 = 3$. $X^2 \sim \chi^2_3$ if H_0 is true.

The observed value of $X^2 = 3.39 + 0.12 + 0.05 + 30.43 = 33.99$ so the P value is $P(X^2 > 33.99)$. This p- value is 1.99×10^{-7} so there is overwhelming evidence against the null hypothesis.

5. The masses measured on a population of 100 animals were grouped in the following table, after being recorded to the nearest gram

Mass	≤ 89	90-109	110-129	130-149	150-169	170-189	≥ 190
Frequency	3	7	34	43	10	2	1

You are given that the sample mean of the data is 131.5 and the sample standard deviation is 20.0. Test the hypothesis that the distribution of masses follows a normal distribution at the 5% significance level.

Solution: We draw up a table of probabilities. Note that y_i is the upper end point of the class.

y_i	$z_i = \frac{y_i - 131.5}{20}$	$\Phi(z_i)$	$\Phi(z_i) - \Phi(z_{i-1}$	E_i	O_i
89.5	-2.1	0.0179	0.0179	1.8	3
109.5	-1.1	0.1357	0.1178	11.8	7
129.5	-0.1	0.4602	0.3245	32.4	34
149.5	0.9	0.8159	0.3557	35.6	43
169.5	1.9	0.9713	0.1554	15.5	10
189.5	2.9	0.9981	0.0268	2.7	2
∞	∞	1.0000	0.0019	0.2	1

Now to ensure that $E_i > 5$ we merge the first two classes and the last three classes and have the following values of O_i and E_i.

O_i	10	34	43	13
E_i	13.6	32.4	35.6	18.4

Null hypothesis H_0 : The data are normally distributed.

Alternative hypothesis $H_1 : \neg H_0$

Test statistic

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

degrees of freedom $\nu = 4 - 2 - 1 = 1$. $X^2 \sim \chi_1^2$ if H_0 is true.

Observed value of $X^2 = 4.15$.

$\chi_1^2(0.05) = 3.841$ so we reject H_0 if $X^2 > 3.841$. Thus we can reject H_0 at the 5% significance level. The data are not compatible with a normal distribution.

Note that with only 100 observations we had to merge a lot of classes and ended up with only one degree of freedom. Ideally we would have a bigger sample size and more degrees of freedom.

You are given the following values from R:

```
> pchisq(3.72, 9)
[1] 0.07114341
> qchisq(0.95,5)
[1] 11.0705
> qchisq(0.95,6)
[1] 12.59159
> qchisq(0.95,7)
[1] 14.06714
```

```
> qchisq(0.95,8)
[1] 15.50731
> qchisq(0.95,9)
[1] 16.91898
> qchisq(0.95,4)
[1] 9.487729
> qchisq(0.95,3)
[1] 7.814728
> qchisq(0.95,2)
[1] 5.991465
> pchisq(33.99,3)
[1] 0.9999998
> pnorm(-2.1)
[1] 0.01786442
> pnorm(-1.1)
[1] 0.1356661
> pnorm(-0.1)
[1] 0.4601722
> pnorm(0.9)
[1] 0.8159399
> pnorm(1.9)
[1] 0.9712834
> pnorm(2.9)
[1] 0.9981342
> qchisq(0.95,1)
[1] 3.841459
```