Statistics for Data Science - 2

Practice assignment week 11

1. 46 people are divided into two groups - experimental and control. The experimental group is inocculated against a disease while the control group is not. Both the groups are then exposed to the disease and the data obtained is recorded in the contingency table given below:

	Control group	Experimental group	Total
Contracted	13	8	21
Not Contracted	10	15	25
Total	23	23	46

Use Chi-squared test to check if the inocculation and the contract of disease are independent at a significance level of 5%.

(a) Yes

(b) No

Solution: If independent, expected counts will be

Number of people who contracted the disease in the control group $= \frac{23 \times 21}{46} = 10.5$

Number of people who did not contract the disease in the control group $= \frac{23 \times 25}{46} = 12.5$

Number of people who contracted the disease in the experimental group $= \frac{23 \times 21}{46} = 10.5$

Number of people who did not contract the disease in the experimental group $= \frac{23 \times 25}{46} = 12.5$

H_0 : Joint PMF is the product of marginals.
H_A : Joint PMF is not the product of the marginals.

Test : Reject H_0, if $T > c$, where

$$T = \frac{(13-10.5)^2}{10.5} + \frac{(10-12.5)^2}{12.5} + \frac{(8-10.5)^2}{10.5} + \frac{(15-12.5)^2}{12.5}$$

$$= 2.19$$

We know that $T \approx \chi_1^2$ & $\alpha = 0.05$

Now, $\alpha = P(T > c \mid H_0 \text{ is true})$

$$= P(\chi_1^2 > c)$$

$$\Rightarrow 0.95 = F_{\chi_1^2}(c)$$

$$\Rightarrow c = F_{\chi_1^2}^{-1}(0.95) = 3.84$$

Since $T = 2.19 < 3.84 = c$, we will accept H_0.

∴ Inoculation and contraction of the disease are independent.

2. Consider the following cross tabulation of status of completion of courses by learners across three different websites:

	Website A	Website B	Website C	
Completed	182	213	203	598
Not completed	154	138	110	402
	336	351	313	

Are status of completion of course independent of the websites? (Use $\alpha = 0.05$)

(a) Yes

(b) No

Solution: Let H_0 : Joint PMF is the product of marginals.
H_A : Joint PMF is not the product of marginals.

If the status of course is independent of the websites, then the expected count will be

	Website A	Website B	Website C
Completed	200.928	209.898	181.174
Not completed	135.072	141.102	125.826

Test statistic, $T = \frac{(182-200.9)^2}{200.9} + \frac{(213-209.9)^2}{209.9} + \frac{(203-181.2)^2}{181.2} + \frac{(154-135.1)^2}{135.1} + \frac{(138-141.1)^2}{141.1} + \frac{(110-125.8)^2}{125.8}$

$= 9.14$

We know that $T \approx \chi^2_2$, $\alpha = 0.05$

$\therefore 0.05 = 1 - F_{\chi^2_2}(c)$

$\Rightarrow c = F^{-1}_{\chi^2_2}(0.95) = 5.99$

$\because T = 9.14 > 5.99 = c$
$\Rightarrow$ Reject H_0
$\therefore$ Not independent

3. Suppose that a sample of 150 electron tubes are tested and the following summary of their life length (in hours), T is reported:

Life length	$0 \leq T < 100$	$100 \leq T < 200$	$200 \leq T < 300$	$T > 300$
Number of electron tubes	47	40	35	28

The sample mean is recorded to be 200 hours. Test the hypothesis that T is exponentially distributed at a significance level of 0.01 using Chi-squared test.

(a) Accept the hypothesis that T is exponentially distributed.

(b) Reject the hypothesis that T is exponentially distributed.

Solution: Let the random variable T represent the life length of electron tubes.

Given, $T \sim Exp(\lambda)$, $n = 150$, $\bar{T} = 200$ hrs

ML estimate of λ is

$$\hat{\lambda}_{ML} = \frac{150}{X_1 + \cdots + X_{150}} = \frac{1}{\bar{T}} = 0.005$$

Let $p_1 = P(0 \leq T < 100) = F_T(100) - F_T(0)$

$$= 1 - e^{-0.005(100)} = 0.39$$

$$p_2 = P(100 \leq T < 200) = F_T(200) - F_T(100)$$

$$= 1 - e^{-0.005(200)} - 0.39 = 0.24$$

$$p_3 = P(200 \leq T < 300) = F_T(300) - F_T(200)$$

$$= 1 - e^{-0.005(300)} - 0.63 = 0.15$$

$$p_4 = P(T \geq 300) = 1 - F_T(300)$$

$$= e^{-0.005(300)} = 0.22$$

H_0 : T is exponentially distributed.
H_A : T is not exponentially distributed.

Now, the expected and observed counts are —

Life length	$0 \le T < 100$	$100 \le T < 200$	$200 \le T < 300$	$T \ge 300$
Observed	47	40	35	28
Expected	$0.39 \times 150 = 58.5$	$0.24 \times 150 = 36$	$0.15 \times 150 = 22.5$	$0.22 \times 150 = 33$

Test : Reject H_0, if $T > c$, where

$$T = \frac{(47-58.5)^2}{58.5} + \frac{(40-36)^2}{36} + \frac{(35-22.5)^2}{22.5} + \frac{(28-33)^2}{33}$$

$$= 11.56$$

Also $T \approx \chi^2_3$

Now, $\alpha = P(T > c \mid H_0 \text{ true})$

$\Rightarrow 0.01 = 1 - F_{\chi^2_3}(c)$

$\Rightarrow F^{-1}_{\chi^2_3}(0.99) = c \Rightarrow \boxed{c = 11.34}$

Since $T = 11.56 > c = 11.34$, reject H_0.

$\therefore$ At 1% level of significance, data doesn't fit the exponential distribution.

4. The number of accidental deaths in a country are tabulated each day for a specified period of 400 days, along with expected frequencies according to a Poisson fit.

Number of deaths	0	1	2	3	4+
Observed frequency	1448	805	206	34	7
Expected frequency	1450	775	200	25	50

Use Chi-squared test to check if the above fit is acceptable at a significance level of 5%.

(a) Yes

(b) No

Solution: Let H_0 : The data fits the Poisson distribution

H_A : The data does not fit the Poisson distribution.

Test : Reject H_0, if $T > c$, where

$$T = \frac{(1448-1450)^2}{1450} + \frac{(805-775)^2}{775} + \frac{(206-200)^2}{200} + \frac{(34-25)^2}{25} + \frac{(7-50)^2}{50}$$

$$= 41.56$$

Also $T \approx \chi^2_4$, $\alpha = 0.05$

$$\alpha = P(T > c \mid H_0 \text{ true})$$

$$\Rightarrow 0.05 = 1 - F_{\chi^2_4}(c)$$

$$\Rightarrow F^{-1}_{\chi^2_4}(0.95) = c \Rightarrow \boxed{c = 9.49}$$

$\because T = 41.56 > 9.49 = c$, reject H_0.

$\therefore$ The given data does not fit Poisson distribution.

5. Let $X_1, \ldots, X_n \sim \text{Normal}(0, \sigma^2)$, and consider testing $H_0 : \sigma = 1$ versus $H_A : \sigma = 2$. Find the likelihood ratio of the observed samples 2, 3, 1, 4.4, 5, 5, 3.6, 6, 4, 6.

(a) 1.02×10^{-27}

(b) 1.02×10^{27}

(c) 0.02×10^{27}

(d) 0.02×10^{-27}

Solution: Given $X_1, \ldots, X_n \sim \text{Normal}(0, \sigma^2)$

$H_0 : \sigma = 1$

$H_A : \sigma = 2$

Likelihood ratio is given by

$$L(X_1, \ldots, X_n) = \frac{\prod_{i=1}^{n} \left(\frac{1}{2\sqrt{2\pi}}\right) e^{\frac{-x_i^2}{2 \cdot 2^2}}}{\prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}}\right) e^{\frac{-x_i^2}{2 \cdot 1^2}}}$$

$$= \frac{\left(\frac{1}{2}\right)^n e^{-\frac{1}{8}\left(\sum_{i=1}^{n} x_i^2\right)}}{e^{-\frac{1}{2}\left(\sum x_i^2\right)}} = \left(\frac{1}{2}\right)^n e^{\frac{3}{8}\left(\sum_{i=1}^{n} x_i^2\right)}$$

The samples are 2, 3, 1, 4.4, 5, 5, 3.6, 6, 4, 6

$\therefore L(X_1, \ldots, X_{10}) = \left(\frac{1}{2}\right)^{10} e^{\frac{3}{8}\left(\sum_{i=1}^{10} x_i^2\right)} = 1.02 \times 10^{27}$

6. A random sample of 40 fibres manufactured using process A has a mean length of 16.7 cm, and standard deviation of 0.5 cm. A random sample of 60 fibres manufactured using process B has mean length of 16.4 cm and standard deviation of 0.6 cm. Test the hypothesis that the mean length of fibres manufactured using process A and B are the same.

   (a) Identify the null and alternative hypothesis.

      i. $H_0 : \mu_1 = \mu_2,\ H_A : \mu_1 < \mu_2$
      ii. $H_0 : \mu_1 = \mu_2,\ H_A : \mu_1 > \mu_2$
      iii. $H_0 : \mu_1 = \mu_2,\ H_A : \mu_1 \neq \mu_2$
      iv. $\mu_1 - \mu_2 = 0.3$ cm, $H_A : \mu_1 - \mu_2 \neq 0.3$ cm

   (b) Choose the correct options from the following:

      i. Reject H_0 at a significance level of 0.05.
      ii. Accept H_0 at a significance level of 0.05.

Solution: Let the random variable X represent the fibres manufactured using process A.

Let the random variable Y represent the fibres manufactured using process B.

Given, $n_X = 40$ $\quad n_Y = 60$

$\bar{X} = 16.7$ $\quad \bar{Y} = 16.4$

$\sigma_X = 0.5$ $\quad \sigma_Y = 0.6$

(a) $H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

, where $\mu_1 = \mu_X$ & $\mu_2 = \mu_Y$

(b) Test statistic : $\bar{X} - \bar{Y} = T$

Test : Reject H_0, if $|T| > c$. at $\alpha = 0.05$

$$\mu_T = \mu_1 - \mu_2 = 0$$

$$\sigma_T^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \quad \sigma_1^2 = \sigma_X^2, \ \sigma_2^2 = \sigma_Y^2, \quad n_1 = n_X, \ n_2 = n_Y$$

$$\alpha = P(|T| > c \mid H_0 \text{ true})$$

$$= P\left(\left| \frac{T}{\sqrt{\frac{0.5^2}{40} + \frac{0.6^2}{60}}} \right| > \frac{c}{\sqrt{\frac{0.5^2}{40} + \frac{0.6^2}{60}}} \right)$$

$$= 2P\left(Z < \frac{-c}{\sqrt{0.012}}\right)$$

$$\Rightarrow \frac{0.05}{2} = F_Z\left(\frac{-c}{\sqrt{0.012}}\right)$$

$$\Rightarrow -1.96 = \frac{-c}{\sqrt{0.012}} \Rightarrow \boxed{c = 0.215}$$

Now, $|T| = |\bar{X} - \bar{Y}| = |16.7 - 16.4| = 0.3 > c.$

$\Rightarrow$ Reject H_0.

7. One wants to check if the average IQ of girls and boys are the same. It is known that the IQ's of both boys are girls have a standard deviation of 10. Mean IQ of 200 randomly selected boys is 99 and mean 1Q of 300 randomly selected girls is 97. Using 1% level of significance, comment on the IQ's of girls and boys.
Hint: $F_Z^{-1}(-2.58) = 0.005$

   (a) The average IQ of both boys and girls are the same.

   (b) The average IQ of boys are more as compared to the IQ's of girls.

   (c) The average IQ of boys are less as compared to the IQ's of girls.

Solution : Let X represents the IQ of girls.
Let Y represents the IQ of boys.

Let $n_1 = 300$, $n_2 = 200$
$\bar{X} = 97$, $\bar{Y} = 99$
$\sigma_1 = \sigma_2 = 10$

Let $H_0 : \mu_1 = \mu_2$
$H_A : \mu_1 \neq \mu_2$

Test Statistics : $T = \bar{X} - \bar{Y}$
Test : Reject H_0, if $|T| > c$ at $\alpha = 0.01$.

$\mu_T = 0$
$\sigma_T^2 = \frac{10^2}{300} + \frac{10^2}{200} = \frac{5}{6}$

Now, $\alpha = P(|T| > c \mid \mu_1 - \mu_2 = 0)$

$= P\left(\left|\frac{T}{\sqrt{5/6}}\right| > \frac{c}{\sqrt{5/6}}\right)$

$\Rightarrow 0.01 = 2P\left(Z < \frac{-c}{\sqrt{5/6}}\right)$

$\Rightarrow 0.005 = F_Z\left(\frac{-c}{\sqrt{5/6}}\right)$

$\Rightarrow -2.58 = \frac{-c}{\sqrt{5/6}} \Rightarrow \boxed{c = 2.355}$

Since, $|T| = |\bar{X} - \bar{Y}| = |97 - 99| = 2 < 2.355$

$\Rightarrow$ Accept H_0.

We conclude the average IQ of both boys and girls are same.

8. The amount of saturated fat present in 100 grams of cheese of two different brands are measured. The data are as follows:

Brand A	Brand B
21	20
19	39
20	24
23	33
22	30
28	28
32	30
19	22
13	33
18	24

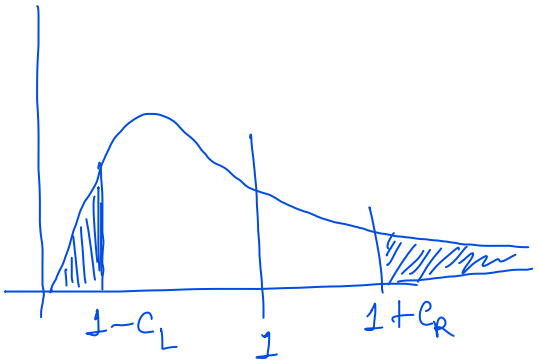
Can we conclude at 5% level of significance that the two variances are equal?
Hint: $s_A = 5.3177, s_B = 5.8699, F_{F(9,9)}(0.248) = 0.025$

(a) Yes

(b) No

Solution : Let $H_0 : \sigma_A^2 = \sigma_B^2$

$H_A : \sigma_A^2 \neq \sigma_B^2$



Test Statistics : $\frac{S_A^2}{S_B^2}$

Test : Reject H_0, if either $\frac{S_A^2}{S_B^2} > 1 + c_R$

or $\frac{S_A^2}{S_B^2} < 1 - c_L$

Also, $\frac{S_A^2}{S_B^2} \sim F(10-1, 10-1) = F(9,9)$

Consider $\frac{\alpha}{2} = P\left(\frac{S_A^2}{S_B^2} < 1 - C_L\right)$

$\Rightarrow \frac{0.05}{2} = F_{F(9,9)}(1 - C_L)$

$\Rightarrow F^{-1}_{F(9,9)}(0.025) = 1 - C_L$

$\Rightarrow \boxed{0.248 = 1 - C_L}$

Now, $\frac{S_A^2}{S_B^2} = \left(\frac{5.3177}{5.8699}\right)^2 = 0.820 > 1 - C_L$

$\therefore$ $\boxed{\text{Accept } H_0.}$

9. A study is conducted to compare the duration of time a certain dose of pain reliever works when administered to men and women. Standard deviation for a random sample of 11 men is found to be 6.1 and, for a random sample of 14 women, it is found to be 5.3. Use a significance level of $\alpha = 0.05$ to check the hypothesis that the variation in time of relief is equal for both genders against the alternative that it is larger for men. Hint: Use $F^{-1}_{F(10,13)}(0.95) = 2.67$.

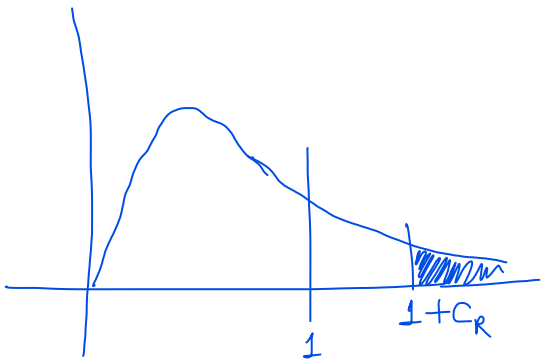
   (a) Reject H_0.

   (b) Fail to reject H_0.

Solution : Let X & Y represent the duration of time the pain reliever works when administered to men & women, respectively.

Let $n_1 = 11$ $\quad n_2 = 14$

$\sigma_1 = 6.1$ $\quad \sigma_2 = 5.3$



Let $H_0 : \sigma_1^2 = \sigma_2^2$

$H_A : \sigma_1^2 > \sigma_2^2$

Test Statistic : $\dfrac{S_1^2}{S_2^2}$

Test : Reject H_0, if $\dfrac{S_1^2}{S_2^2} > 1 + c_R$ (c say)

Also, $\dfrac{S_1^2}{S_2^2} \sim F(11-1, 14-1) = F(10, 13)$

Now, $\alpha = P\left(\frac{S_1^2}{S_2^2} > 1+c_R\right)$

$\Rightarrow 0.05 = 1 - F_{F(10,13)}^{(1+c_R)}$

$\Rightarrow F^{-1}_{F(10,13)}(0.95) = 1+c_R$

$\Rightarrow \boxed{2.67 = 1+c_R}$

$\frac{S_1^2}{S_2^2} = \frac{6.1^2}{5.3^2} = 1.32 < 1+c_R \Rightarrow \boxed{\text{Accept } H_0}$

10. Past experience indicates that the time required for athletes to complete a 200 m race is a normal random variable with a mean $\mu = 35$ seconds. If a random sample of 20 athletes took an average of 33.1 seconds to complete the race with a standard deviation of 4.3 seconds, test the hypothesis, at the 0.05 level of significance, that $\mu = 35$ seconds against the alternative that $\mu < 35$ seconds.

(a) Accept the null hypothesis.

(b) Reject the null hypothesis.

Solution: Let the random variable X represent the time required by athletes to complete a 200 m race.

Given, $X \sim \text{Normal}(35, \sigma)$

For $n = 20$, $\bar{X} = 33.1$, $S = 4.3$

$H_0 : \mu = 35$, $H_A : \mu < 35$

Test statistics : $\bar{X}$

Test : Reject H_0 if $\bar{X} < c$ at $\alpha = 0.05$

$$\alpha = P(\bar{X} < c \mid \mu = 35)$$

$$= P\left(\frac{\bar{X} - 35}{S/\sqrt{n}} < \frac{c - 35}{S/\sqrt{n}}\right)$$

$$\Rightarrow 0.05 = F_{t_{19}}\left(\frac{c - 35}{4.3/\sqrt{20}}\right) \quad \left\{\because \frac{\bar{X} - 35}{S/\sqrt{n}} \sim t_{19}\right\}$$

$$\Rightarrow F^{-1}_{t_{19}}(0.05) = \frac{c - 35}{4.3/\sqrt{20}}$$

$$\Rightarrow -1.729 = \frac{c - 35}{4.3/\sqrt{20}} \Rightarrow \boxed{c = 33.33}$$

$\because \bar{X} = 33.1 < 33.33 = c \Rightarrow$ **Reject H_0** Ans

11. A company manufactures mobiles chargers with an output voltage of 5V and variance $0.5V^2$. The company wants to test the variance. They take a random sample of 12 chargers and the following voltages are obtained:
5.34, 5.65, 4.76, 5.00, 5.55, 5.54, 5.07, 5.35, 5.44, 5.25, 5.35, 4.61

Test the hypothesis that $\sigma^2 = 0.5$ at 0.05 level of significance.

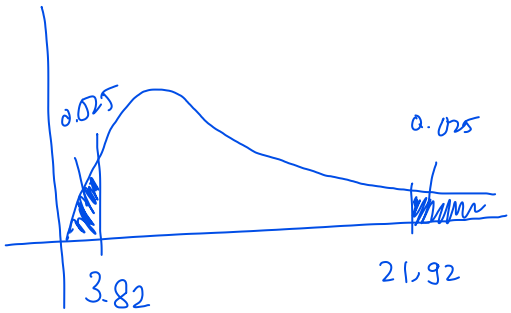
(a) Accept H_0

(b) Reject H_0

Solution:

$H_0 : \sigma^2 = 0.5$

$H_A : \sigma^2 \neq 0.5$



Sample variance, $s^2 = 0.10388$

$n = 12$

Test Statistics $= \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

Observed test statistic value is $\frac{(12-1)(0.10388)}{0.5}$ (Assuming H_0 true)

$= 2.28$

Also, $F^{-1}_{\chi^2_{11}}(0.025) = 3.82$ and $F^{-1}_{\chi^2_{11}}(0.975) = 21.92$ { Since, it is a two-sided test.

∴ The observed test statistic value lies in the rejection region, we will reject H_0.

12. The standard deviation of a component in a drug is expected to be 0.00002 kg. A pharmacist suspecting the variability to be higher obtains a sample of 8 drugs and found the sample standard deviation to be 0.00005 kg.

(a) Identify the null and alternative hypothesis:

i. $H_0 : \sigma = 0.00002,\ H_A : \sigma > 0.00002$

ii. $H_0 : \sigma = 0.00005,\ H_A : \sigma \neq 0.00005$

iii. $H_0 : \sigma = 0.00002,\ H_A : \sigma < 0.00002$

iv. $H_0 : \sigma = 0.00002,\ H_A : \sigma \neq 0.00002$

(b) What conclusion would a χ^2 test reach at a significance level of 0.01?

i. Accept the null hypothesis.

ii. Accept the alternative hypothesis.

Solution : (a) $H_0 : \sigma = 0.00002$

$H_A : \sigma > 0.00002$

Given $n = 8$, $s = 0.00005$ kg

Test Statistic : S^2

Test : Reject H_0, if $S^2 > c^2$ at $\alpha = 0.01$

$$\alpha = P(S^2 > c^2 \mid \sigma = 0.00002)$$

$$= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)c^2}{\sigma^2} \mid \sigma = 0.00002\right)$$

$$\Rightarrow 0.01 = 1 - F_{\chi^2_7}\left(\frac{7c^2}{0.00002^2}\right)$$

$$\Rightarrow F^{-1}_{\chi^2_7}(0.99) = \frac{7c^2}{0.00002^2}$$

$$\Rightarrow 18.47 = \frac{7c^2}{0.00002^2} \Rightarrow c^2 = 1.05 \times 10^{-9}$$

Now, $s^2 = 0.00005^2 = 2.5 \times 10^{-9} > c^2$

$\Rightarrow$ **Reject H_0.** Ans