Statistics for Data Science - 2

Week 11 Graded Assignment Solution

1. The IQs (intelligence quotients) of 25 students from one batch of IITM students showed a mean of 110 with a standard deviation of 8, while the IQs of 25 students from another batch of IITM students showed a mean of 115 with a standard deviation of 7. Is there a significant difference between the IQs of the two groups at a 0.05 level of significance?

   a) Yes

   b) No

   Hint: Use $F_Z^{-1}(0.025) = -1.96$

   Solution:

   Let X_i and Y_i represent the IQ's of both batch of students.

   $X_1, X_2, \ldots, X_{25} \sim \mathrm{N}(\mu_1, 8^2)$ and $Y_1, Y_2, \ldots, Y_{25} \sim \mathrm{N}(\mu_2, 7^2)$

   $\overline{X} = 110$ and $\overline{Y} = 115$

   Consider, $H_0 : \mu_1 = \mu_2, H_A : \mu_1 \neq \mu_2$

   $T = \overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \frac{64}{25} + \frac{49}{25})$ i.e. $\mathrm{N}(\mu_1 - \mu_2, \frac{113}{25})$

   Test: Reject H_0 if $|T| > c$.

$$\alpha = P(|T| > c \mid H_0) = P\left(|\frac{T}{\sqrt{113/25}}| > \frac{c}{\sqrt{113/25}}\right)$$
$$= P\left(|Z| > \frac{c}{\sqrt{113/25}}\right) = 2F_Z\left(\frac{-c}{\sqrt{113/25}}\right)$$

   $\Rightarrow c = -\sqrt{\frac{113}{25}} F_Z^{-1}(\alpha/2)$

   $\Rightarrow c = -\sqrt{\frac{113}{25}} F_Z^{-1}(0.025)$

   $\Rightarrow c = -\sqrt{\frac{113}{25}} \times (-1.96) = 4.167$

   Since, $|\overline{X} - \overline{Y}| = |110 - 115| = 5 > 4.167$

   Therefore, we will reject H_0.

   This implies that there a significant difference between the IQs of the two groups at a 0.05 level of significance.

2. A sociologist focusing on popular culture and media believes that the average number of hours per week (hrs/week) spent on social media is different for men and women. The researcher knows that the standard deviations of amount of time spent on social media are 5 hrs/week and 6 hrs/week for men and women, respectively. Examining two independent random samples of 64 individuals each, if the average number of hrs/week

spent on social media for the sample of men is 1.5 hours greater than that for the sample of women, what conclusion can be made from a hypothesis test where, $H_0 : \mu_M = \mu_W$ and $H_A : \mu_M \neq \mu_W$? Take $\alpha = 0.05$.

a) Reject H_0

b) Accept H_0

Solution:
Let X_i and Y_i represent the average number of hrs/week spent on social media by men and women respectively.
$X_1, X_2, \ldots, X_{64} \sim \mathrm{N}(\mu_1, 5^2)$ and $Y_1, Y_2, \ldots, Y_{64} \sim \mathrm{N}(\mu_2, 6^2)$
$|\overline{X} - \overline{Y}| = 1.5$
Consider, $H_0 : \mu_1 = \mu_2, H_A : \mu_1 \neq \mu_2$
$T = \overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \frac{25}{64} + \frac{36}{64})$ i.e. $\mathrm{N}(\mu_1 - \mu_2, \frac{61}{64})$
Test: Reject H_0 if $|T| > c$.

$$\alpha = P(|T| > c \mid H_0) = P\left(|\frac{T}{\sqrt{61/64}}| > \frac{c}{\sqrt{61/64}}\right)$$
$$= P\left(|Z| > \frac{c}{\sqrt{61/64}}\right) = 2F_Z\left(\frac{-c}{\sqrt{61/64}}\right)$$

$\Rightarrow c = -\sqrt{\frac{61}{64}} F_Z^{-1}(\alpha/2)$
$\Rightarrow c = -\sqrt{\frac{61}{64}} F_Z^{-1}(0.025)$
$\Rightarrow c = -\sqrt{\frac{61}{64}} \times (-1.96) = 1.913$
Since, $|\overline{X} - \overline{Y}| = 1.5 < 1.913$
Therefore, we will accept H_0.

3. An IITM instructor conducts two live sessions for two different classes, call it A and B, in Statistics. Session A had 25 students attending while session B had 36 students. The instructor conducted a test for the two sessions. Although there was no significant difference in mean grades, session A had a standard deviation of 10 while session B had a standard deviation of 14. Can we conclude at the 0.01 level of significance that the variability in marks of class B is greater than that of A?

a) Yes

b) No

Hint: Use $F^{-1}_{F_{(35,24)}}(0.99) = 2.529$
Solution:
$H_0 : \sigma_1 = \sigma_2, H_A : \sigma_1 < \sigma_2$

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

We know that, $\frac{S_B^2}{S_A^2} \sim F(n_2 - 1, n_1 - 1)$

$n_1 = 25, n_2 = 36$

$\Rightarrow \frac{S_B^2}{S_A^2} \sim F(35, 24)$

Therefore,

$$\begin{aligned} \alpha &= 1 - F_{F(35,24)}(1 + c_R) \\ \Rightarrow 1 + c_R &= F^{-1}_{F(35,24)}(1 - \alpha) = F^{-1}_{F(35,24)}(0.99) \\ \Rightarrow 1 + c_R &= 2.529 \end{aligned}$$

Since, $\frac{S_B^2}{S_A^2} = \frac{14^2}{10^2} = 1.96 < 2.529$

Therefore, we will accept H_0.

This implies that at the 0.01 level of significance the variability in marks of class B is not greater than that of A.

4. The manufacturer of a new car claims that a typical car gets a mileage of 40 kilometres per litre. We think that the mileage is less. To test our suspicion, we perform the hypothesis test with $H_0 : \mu = 40$ and $H_A : \mu < 40$. Suppose we take a random sample of 900 new cars and find that their average mileage is 39.8 kilometres per litre and sample standard deviation is 2, what does a t-test say about a null hypothesis with a significance level of 0.05?

   a) Reject H_0

   b) Accept H_0

   Hint: Use $F^{-1}_{t_{899}}(0.05) = -1.646$

   Solution:

   Null hypothesis, $H_0 : \mu = 40$

   Alternate hypothesis, $H_A : \mu < 40$

   Test: Reject H_0 if $\overline{X} < c$

   Given, $\alpha = 0.05$ and $\overline{X} = 39.8$

   In this problem, we do not know the population variance, σ^2.

   The sample variance $S^2 = 2^2$

$$\alpha = P(\overline{X} < c | \mu = 40)$$

$$\alpha = P\left(\frac{\overline{X} - 40}{\sqrt{S^2/n}} < \frac{c - 40}{\sqrt{S^2/n}}\right)$$

$$\alpha = P\left(\frac{\overline{X} - 40}{\sqrt{4/900}} < \frac{c - 40}{\sqrt{4/900}}\right)$$

$$\alpha = F_{t_{899}}\left(\frac{c - 40}{\sqrt{4/900}}\right)$$

$$0.05 = F_{t_{899}}\left(\frac{c - 40}{\sqrt{4/900}}\right)$$

$$c = 40 + \sqrt{\frac{4}{900}} F^{-1}_{t_{899}}(0.05)$$

$$c = 39.89$$

Since, $\overline{X} < c$, reject H_0.

5. The standard deviation of weights of 70 gram bags of white cheddar popcorn is expected to be 2.5 grams. A random sample of 20 packages showed a standard deviation of 3 grams. Is the apparent increase in variability significant at the 0.05 level.?

   a) Yes

   b) No

   Hint: Use $F^{-1}_{\chi^2_{19}}(0.95) = 30.14$

   Solution:

   As per given information, the null and alternative hypothesis are given by

   $$H_0 : \sigma = 2.5, \qquad H_A : \sigma > 2.5$$

   Define a test statistic T as $T = S^2$.

   We know that $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{19S^2}{2.5^2} \sim \chi^2_{19}$.

   Test: reject the null hypothesis if $S^2 > c^2$.

   If the significance level of the test is 0.05, then

$$
\begin{aligned}
&P(S^2 > c^2) = 0.05 \\
\Rightarrow &P\left(\frac{19S^2}{2.5^2} > \frac{19c^2}{2.5^2}\right) = 0.05 \\
\Rightarrow &P\left(\chi^2_{19} > \frac{19c^2}{2.5^2}\right) = 0.05 \\
\Rightarrow &1 - P\left(\chi^2_{19} < \frac{19c^2}{2.5^2}\right) = 0.05 \\
\Rightarrow &\frac{19c^2}{2.5^2} = 30.14 \\
\Rightarrow &c^2 = \frac{6.25 \times 30.14}{19} = 9.91
\end{aligned}
$$

Since $S^2 = 9 < 9.91$, we will not reject the null hypothesis.
Therefore, the apparent increase in variability is not significant at the 0.05 level.

6. Independent random samples of ceramic produced by two different processes were tested for hardness. The results are:

Process 1	Process 2
8.5	9.0
9.5	9.5
8.0	10.5
9.0	9.5
10.0	10.0
9.5	9.0
10.5	9.0
10.0	9.5

Table 11.1.G

Can we conclude at 5% level of significance that the variances in hardness are equal?

a) Yes

b) No

Hint: Use $F^{-1}_{F_{(7,7)}}(0.025) = 0.2$

Solution:

Let Process 1 and Process 2 values denoted by X_i and Y_i respectively.

$H_0 : \sigma_1 = \sigma_2, H_A : \sigma_1 \neq \sigma_2$

Test: Reject H_0 if $\frac{S_X^2}{S_Y^2} > 1 + c_R$ or $\frac{S_X^2}{S_Y^2} < 1 - c_L$

We know that, $\frac{S_X^2}{S_Y^2} \sim F(n_1 - 1, n_2 - 1)$

$n_1 = 8, n_2 = 8$

$\Rightarrow \frac{S_X^2}{S_Y^2} \sim F(7, 7)$

Therefore,

$$\alpha/2 = F_{F(7,7)}(1 - c_L)$$
$$\Rightarrow 1 - c_L = F^{-1}_{F(7,7)}(\alpha/2) = F^{-1}_{F(7,7)}(0.025)$$

$\Rightarrow 1 - c_L = 0.2$

Since, $\frac{S_X^2}{S_Y^2} = \frac{0.6964}{0.2857} = 2.437 > 0.2$

Similarly we can check for other condition.

$$\alpha/2 = 1 - F_{F(7,7)}(1 + c_R)$$
$$\Rightarrow 1 + c_R = F^{-1}_{F(7,7)}(1 - \alpha/2) = F^{-1}_{F(7,7)}(0.975)$$

$\Rightarrow 1 + c_R = 4.99$

Since, $\frac{S_X^2}{S_Y^2} = \frac{0.6964}{0.2857} = 2.437 < 4.99$

Therefore, we will accept H_0.

7. Let $X \sim \text{Normal}(0, \sigma^2)$. Consider the test $H_0 : \sigma = 4$ against $H_A : \sigma = 5$. A sample $X_1, X_2, \ldots, X_{10}$ is observed. What is the likelihood ratio function for the observed sample?

   a) $\left(\frac{5}{4}\right)^{10} \exp\left(\frac{9}{800}\sum_{i=1}^{10} X_i^2\right)$

   b) $\left(\frac{4}{5}\right)^{10} \exp\left(-\frac{9}{800}\sum_{i=1}^{10} X_i^2\right)$

   c) $\left(\frac{5}{4}\right)^{10} \exp\left(-\frac{9}{800}\sum_{i=1}^{10} X_i^2\right)$

   d) $\left(\frac{4}{5}\right)^{10} \exp\left(\frac{9}{800}\sum_{i=1}^{10} X_i^2\right)$

   Solution:

   Given $X \sim \text{Normal}(0, \sigma)$

   The null and alternative hypothesis are -

   $$H_0 : \sigma = 4$$

   $$H_A : \sigma = 5$$

Samples $X_1, \ldots, X_{10}$ are observed.
Now, likelihood ratio,

$$\begin{aligned}
L =& \frac{\prod_{i=1}^{10} \text{Normal}(0, 5^2)}{\prod_{i=1}^{10} \text{Normal}(0, 4^2)} \\
=& \frac{\left(\frac{1}{5}\right)^{10} \exp\left(-\frac{\sum_{i=1}^{i=10} X_i^2}{50}\right)}{\left(\frac{1}{4}\right)^{50} \exp\left(-\frac{\sum_{i=1}^{i=10} X_i^2}{32}\right)} \\
=& \left(\frac{4}{5}\right)^{10} \exp\left(\sum_{i=1}^{i=10} X_i^2 \left(\frac{-1}{50} + \frac{1}{32}\right)\right) \\
=& \left(\frac{4}{5}\right)^{10} \exp\left(\frac{9}{800} \sum_{i=1}^{i=10} X_i^2\right)
\end{aligned}$$

8. You have a coin and you would like to check whether it is fair or biased. Let p be the probability of heads, $p = P(H)$. Let the null and alternative hypothesis be $H_0 : p = 0.5$ and $H_A : p = 0.45$. You toss the coin 100 times and observe 49 heads. Can you reject H_0 using likelihood ratio test at significance level $\alpha = 0.05$?

   a) Yes

   b) No

   Solution: The null and the alternative hypothesis are:

$$H_0 : p = 0.5$$

$$H_A : p = 0.45$$

   Given, $w = 49, n = 100$
   Test statistic: $T = \dfrac{(0.45)^w (0.55)^{n-w}}{(0.5)^n}$
   Test: Reject H_0, if $T > c$.

$$\ln T = w \ln(0.45) + n \ln(0.55) - w \ln(0.55) - n \ln(0.5)$$
$$\Rightarrow \ln T = w \ln\left(\frac{0.45}{0.55}\right) + n \ln\left(\frac{0.55}{0.5}\right)$$
$$\Rightarrow \ln T = Y(\text{let}) = (X_1 + \ldots + X_{100}) \ln\left(\frac{9}{11}\right) + n \ln\left(\frac{11}{10}\right)$$

$$\alpha = P(T > c \mid p = 0.50)$$
$$\implies 0.05 = P(\ln T > \ln c \mid p = 0.50)$$
$$\implies 0.05 = P\left(\frac{Y - E[Y]}{SD(Y)} > \frac{\ln c - E[Y]}{SD(Y)}\right)$$
$$\implies 0.05 = P\left(Z > \frac{\ln c - E[Y]}{SD(Y)}\right)$$
$$\implies 1.64 = \frac{\ln c - E[Y]}{SD(Y)}$$

And $E[Y] = E[(X_1 + \ldots + X_{100}) \ln\left(\frac{9}{11}\right) + n \ln\left(\frac{11}{10}\right)]$
$\implies E[Y] = 100 \times 0.5 \times \ln\left(\frac{9}{11}\right) + 100 \ln\left(\frac{11}{10}\right) = -0.50$

Similarly $\text{Var}(Y) = 100 \times 0.5 \times 0.5 \times \ln\left(\frac{9}{11}\right)^2 = 1.006 \implies SD(Y) = 1.003$
$\Rightarrow ln\ c = 1.64 \times 1.003 - 0.50 = 1.14$
$\Rightarrow c = 3.13$
And
$T = \frac{(0.45)^w (0.55)^{n-w}}{(0.5)^n} = \frac{(0.45)^{49}(0.55)^{100-49}}{(0.5)^{100}} = 0.739$

Since, $T = 0.739$ and $c = 3.13$, i.e. $T < c$, Accept H_0.

9. A random number generator is expected to produce digits $0, 1, 2, \ldots, 9$ uniformly at random. In a sample of 250 digits, the observed frequencies are given below.
Is the above a good-enough fit at a significance level of 0.05?

   a) Yes

   b) No

Hint: Use $F^{-1}_{\chi^2_9}(0.95) = 16.9$
Solution:

Digit	Observed frequency	Expected frequency
0	17	25
1	31	25
2	29	25
3	18	25
4	14	25
5	20	25
6	35	25
7	30	25
8	20	25
9	36	25

Table 11.2.G

Value of the test statistic T is given by

$$\begin{aligned} T &= \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \frac{(14-25)^2}{25} \\ &+ \frac{(20-25)^2}{25} + \frac{(35-25)^2}{25} + \frac{(30-25)^2}{25} + \frac{(20-25)^2}{25} + \frac{(36-25)^2}{25} \\ &= \frac{582}{25} \\ &= 23.28 \end{aligned}$$

Test: Reject H_0 if $T > c$.
We know that $\alpha = P(T > c \mid H_0) \approx 1 - F_{\chi^2_{k-1}}(c)$
$\Rightarrow c = F^{-1}_{\chi^2_{k-1}}(1-\alpha)$
Here $k = 10, \alpha = 0.05$
$\Rightarrow c = F^{-1}_{\chi^2_9}(1-0.05) = F^{-1}_{\chi^2_9}(0.95) = 16.9$
Since $T > c$, we will reject null hypothesis.
This implies that the above data is not good-enough fit at a significance level of 0.05.

10. On two major e-commerce websites A and B the sales on a particular day is given as a contingency table [Table 11.3.G].

	A	B	Total
Bought item	1000	1500	2500
Didn't buy item	1200	2000	3200
Total	2200	3500	5700

Table 11.3.G: E-commerce sales data

Can we say that the sales is independent of websites at a significance level of 0.05?

a) Yes

b) No

Hint: Use $F^{-1}_{\chi^2_1}(0.95) = 3.84$
H_0 : Joint PMF is product of marginals
H_A: Joint PMF is not the product of marginals

$$\text{Number of people bought items via website A} = \frac{2500 \times 2200}{5700} = 964.91$$

$$\text{Number of people bought items via website B} = \frac{2500 \times 3500}{5700} = 1535.08$$

$$\text{Number of people did not buy items via website A} = \frac{3200 \times 2200}{5700} = 1235.08$$

$$\text{Number of people did not buy items via website B} = \frac{3200 \times 3500}{5700} = 1964.91$$

Therefore, value of the test statistic T is given by

$$\begin{aligned} T &= \frac{(1000 - 964.91)^2}{964.91} + \frac{(1500 - 1535.08)^2}{1535.08} + \frac{(1200 - 1235.08)^2}{1235.08} + \frac{(2000 - 1964.91)^2}{1964.91} \\ &= 1.276 + 0.8016 + 0.9963 + 0.6266 \\ &= 3.7005 \end{aligned}$$

We will reject the null hypothesis if $T > c$.

At a significance level of 0.05, we have

$$\begin{aligned} 0.01 &= P(T > c) \\ \Rightarrow 0.01 &= 1 - P(T \leq c) \\ \Rightarrow P(T \leq c) &= 0.95 \\ \Rightarrow F_{\chi^2_1}(c) &= 0.95 \\ \Rightarrow c &= 3.84 \end{aligned}$$

Since $T = 3.7005 < 3.84$, we will accept the null hypothesis.
This implies that sales is independent of websites at a significance level of 0.05.