CH5020

ASSIGNMENT 2

1) Mayur Vikas Joshi – **ME16B148**
2) Sushant Uttam Wadavkar – **ME16B172**

Problem 1

It is claimed that sports-car owners drive on the average 18,000 miles per year. A consumer firm believes that the average mileage is probably lower. To check, the consumer firm obtained information from 40 randomly selected sports-car owners that resulted in a sample mean of 17,463 miles with a sample standard deviation of 1348 miles. What can we conclude about this claim? Use α = 0.01.

Solution:

The null and the alternative hypothesis are H0 : μ = 18000 vs. Ha : μ < 18000.

The observed value of the z−statistic is (under the null hypothesis), in our usual notation,

$z = (\bar{x} - \mu0)/ \sigma/\surd n$

$= (17463 - 18000) / (1348/ \surd 40)$

$= -2.52.$

On the other hand, the z−score corresponding to the left-sided test at the 0.01 significance level is −2.33.

Because the observed value of the z statistic −2.52 is less than −2.33, the null hypothesis is rejected at the significance level of 0.01.

There is sufficient evidence to conclude that the mean mileage on sport cars is less than 18,000 miles per year.

Problem 2

A plant is suspected of discharging harmful effluents above the stipulated limit of 200 mg/L into a nearby river. The plant denies this claim and shows results from sampling of the river carried out by them. These are given in the table below.

However, the Court orders an independent testing agency to sample the effluent concentrations. The results are also tabulated below. The plant lawyer further argues that his client's results are more accurate as his sample shows less standard deviation. The Court appoints a neutral expert to give his recommendation.

- **a.** State the claims of the Plant and the Neutral Agency
- **b.** What conclusions will be drawn by the expert hypothesis testing? State the hypotheses clearly.
- **c.** What conclusion will be drawn if the Plant tries to confuse the judge by invoking the ≠ alternate hypothesis?
- **d.** What conclusions will be drawn by the expert using 95% confidence interval approaches?

Detail	**Plant**	**Neutral Agency**
Claim	?	?
Sample size	3	20
Mean concentration (mg/L)	195	205
Sample standard deviation (mg/L)	4	6
Sampling location(s)	points near mixing of river with the sea	random locations near the plant discharge

Solution:

a. Plant $\mu < 200$; Agency $\mu > 200$

b. Plant

$$H_o : \mu_1 = 200$$

$$H_1 : \mu_1 < 200$$

$$T = \frac{(195 - 200)}{\frac{4}{\sqrt{3}}}$$ = -2.165

```
tcdf(-2.165, 2)
```

z-value = 0.0814

z-value > 0.05

Fail to reject the null hypothesis.

Agency:

$$H_o : \mu_1 = 200$$

$$H_1 : \mu_1 > 200$$

$$T = \frac{(205 - 200)}{\frac{6}{\sqrt{20}}}$$ = 3.728

```
tcdf(3.728, 19, 'upper')
```

t-value = 7.1288e-04

t-value < 0.05

Reject the null hypothesis.

c. Two Tailed Test

$$H_o : \mu_1 = 200$$

$$H_1 : \mu_1 \neq 200$$

$$T = \frac{(195 - 200)}{\frac{4}{\sqrt{3}}}$$ = -2.165

```
tcdf(-2.165, 2)
```

z-value = 0.0814

z-value > 0.025 (alpha/2)

Fail to reject the null hypothesis.

d.

95% CI Plant:

```
tinv(0.025,2)
```

$L = 195 - (4.3027* \frac{4}{\sqrt{3}}) = 185.06$

$U = 195 + (4.3027* \frac{4}{\sqrt{3}}) = 204.93$

95% CI Agency:

```
tinv(0.025,19)
```

$L = 205 - (2.093* \frac{6}{\sqrt{20}}) = 202.192$

$U = 205 + (2.093* \frac{6}{\sqrt{20}}) = 207.808$

Problem 3

The electrical resistances of components are measured as they are produced. A sample of six items gives a sample mean of 2.62 ohms and a sample standard deviation of 0.121 ohms. At what observed level of significance is this sample mean significantly different from a population mean of 2.80 ohms? In other words, is there less than 2% probability of getting a sample mean this far away from 2.80 ohms or farther purely by chance when the population mean is 2.80 ohms? Also indicate whether you may get the same conclusion using a 95% confidence interval test.

Solution:

H0: $\mu = 2.80$

H1: $\mu \neq 2.80$

$\sigma = 0.121$, $\bar{x}$ = 2.62, n=6

The test static t = $\frac{\left(\bar{x} - \mu_0\right)}{\sigma/\sqrt{n}}$

t = $\frac{(2.62 - 2.80)}{0.121/\sqrt{6}}$ = -3.64

From t-distribution table calculating probability for |t|=3.64

matlab code

```
tpdf(3.64,5)
ans = 0.0078
```

Hence, $\alpha = 0.0078$

Hence there is less than 2% probability of getting a sample mean far away from 2.80 ohms

Problem 4

A sample of 15 concrete cylinders was taken randomly during production from a plant. The strength of each specimen was determined, giving a sample standard deviation of 215 kN/m2. Find the 95% confidence interval (with equal probabilities in the two tails) for *standard deviation* of the strengths. Assume the strengths of the concrete cylinders follow a normal distribution.

Solution:

s=215, n=15

Number of degrees of freedom = n-1 =14

α = 0.05

For 95% confidence interval from chi square distribution

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

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

The critical values of χ^2 at $\frac{\alpha}{2}$ = 0.025 and $1-\frac{\alpha}{2}$ = 0.975 for 14 degrees of freedom are

The matlab code to find χ^2 at 0.025 and 0.975

```
chi2inv(0.025,14)
ans = 5.6287
chi2inv(0.975,14)
ans = 26.1189
```

The corresponding limits on σ^2 for χ^2=5.6287 is $\sigma^2 = (215)^2 * \frac{14}{5.6287}$ = 11500 and for χ^2 = 26.1189 is $\sigma^2 = \frac{215^2(14)}{26.1189}$ = 24800

The values of $\sigma = 339.12$ and 157.48

L = 157kN/m2 ; U = 339 kN/m2

Hence, for the 95% confidence interval standard deviation is from 157 kN/m2 to 339 kN/m2

Problem 5

After an Olympiad test is given to all the students in the country, the average is estimated from random samples, with one sample taken from each of the five zones. All these samples have a sample size of 5. The average of each sample was calculated. Let the actual (true) average performance be 35%.

Part I

Which of the following is/are TRUE

1. It is highly likely that all the 5 sample means will be identically equal to 35%-**False**
2. If a larger sample size had been chosen, the distribution of the sample means around the true mean would have been narrower-**True**
3. If a smaller sample size had been chosen, the distribution of the sample means would have been narrower-**False**
4. The average of the distribution of the sample means will be 35%-**True**

Statements 2 and 4 are True

Part II

Which of the following is/are TRUE

If the distribution of marks in the Olympiad's population is **not** normally distributed then

1. The distribution of sample means with sample size equal to 5, will be normal-**True**
2. If the sample size had exceeded 30 for all samples, the distribution of the sample means would have been normal-**True**
3. Higher the sample size, lower will be the precision of the estimated population mean-**False**
4. Higher the sample size, higher will be the precision of the estimated population mean-**True**

Statements 1,2, and 4 are True

Problem 6

From historical data, the yields of power from a nuclear reactor supplied by XYZ Company are normally distributed. This reactor supplied by this company is operated in several plants around the world. The mean daily output of power from a random sample of 6 measurements carried out over different days taken at an Indian plant is 27.33 GW and sample standard deviation is 9 GW.

a. Can the Indian plant accept this yield to be possible if XYZ Company guarantees (or warranties?) an average daily power output of 30 GW from its reactors for a given set of operating conditions?

b. If the same power output and variance are obtained from a sample size of 41, can the observed power output in the Indian plant be still considered to be acceptable?

If you were to make a scientific and unbiased report to the plant management, state your conclusions in both cases clearly.

Solution:

Part a.

$$\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

```
tinv(1-0.025, 5)
tinv(0.025, 5)
```

$$L = 27.33 - 2.5706 * \left(\frac{9}{\sqrt{6}}\right) = 17.89$$

$$U = 27.33 + 2.5706 * \left(\frac{9}{\sqrt{6}}\right) = 36.77$$

30GW lies within the 95% confidence interval.

Part b.

```
norminv(1-0.025)
norminv(0.025)
```

$$L = 27.33 - 1.96 * \left(\frac{9}{\sqrt{41}}\right) = 24.57$$

$$U = 27.33 + 1.96 * \left(\frac{9}{\sqrt{41}}\right) = 30.08$$

30GW lies within the 95% confidence interval.

Problem 7

A manufacturer of car batteries guarantees that his batteries will last, on the average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5 and 4.2 years, is the manufacturer still convinced that his batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

Solution:

Calculating variance using $s^2 = \frac{n\sum_{i=1}^{i=n} x_i^2 - \left(\sum_{i=1}^{i=n} x_i\right)^2}{n(n-1)}$

$$\sum_{i=1}^{5} x_i = 1.9 + 2.4 + 3.0 + 3.5 + 4.2 = 15$$

$$\sum_{i=1}^{5} x_i^2 = 1.9^2 + 2.4^2 + 3.0^2 + 3.5^2 + 4.2^2 = 48.26$$

n=5

$s^2 = \frac{5 * 48.26 - 225}{5 * (5-1)}$ =0.815

For 95% confidence interval from chi square distribution

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

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

The critical values of χ^2 at $\frac{\alpha}{2}$ = 0.025 and $1 - \frac{\alpha}{2}$ = 0.975 for 4 degrees of freedom are

The matlab code to find χ^2 at 0.025 and 0.975

```
chi2inv(0.025,4)
ans = 0.4844
chi2inv(0.975,4)
ans = 11.1433
```

The corresponding limits on σ^2 for $\chi^2_{0.025}$ =0.4844 is $\sigma^2 = \frac{(0.815) * (4)}{0.4844}$ = 6.735 and for $\chi^2_{0.975}$ = 11.1433 is $\sigma^2 = \frac{0.815 * (4)}{11.143}$ = 0.293

The corresponding values of σ are $\sigma = 0.541$ and $\sigma = 2.59$

Hence, for the 95% confidence interval standard deviation is from 0.541 years to 2.59 years

As $\sigma = 1$ is in the interval [0.541,2.59]

We can conclude that manufacture claim $\sigma = 1$ year is valid

Problem 8

For a chi-squared distribution, find X 2 such that

1. P(X2>$X\alpha$ 2) =0.99 when ν =4
2. P(X2>$X\alpha$ 2) =0.025 when ν =19
3. P(37.652< X2 < $X\alpha$ 2) = 0.045 when ν =25

Solution:

1. P($\chi^2 > \chi^2_\alpha)$ =0.99 when v=4

P($\chi^2_\alpha < \chi^2)$ =1- $P(\chi^2_\alpha > \chi^2)$

$P(\chi^2_\alpha > \chi^2$)=1-0.99=0.01

From chi square distribution table for $\alpha = 0.01$ and v=4

χ^2 =13.2767

2. $P(\chi^2_\alpha > \chi^2$)=1-0.025=0.975

From chi square distribution table for $\alpha = 0.975$ and v=19

χ^2 =32.852

3. P(37.652< X2 < $X\alpha$ 2) = 0.045 when ν =25

P(37.652< X2 < $X\alpha$ 2) = P(χ^2 >37.652)-P($\chi^2 > \chi^2_\alpha$)

P($\chi^2 > \chi^2_\alpha$)=P(χ^2 >37.652)-0.045 = P($\chi^2 > \chi^2_{0.05})$ -0.045 = 0.05-0.045 = 0.005

From chi square distribution table for $\alpha = 0.975$ and v=25

Problem 9

A normal population with unknown variance has a mean of 20. Is one likely to obtain a random sample of size 9 from this population with a mean of 24 and a standard deviation of 4.1? If not, what conclusion would you draw?

Solution:

Standard deviation = 4.1

mean $\bar{x}$ = 24

$\mu = 20$

sample size = n =9

$H_0 : \mu = 20$ null hypothesis

$H_1 : \mu \neq 20$ alternate hypothesis

The test static value under null hypothesis t = $\dfrac{\left(\bar{x} - \mu_0\right)}{\dfrac{\sigma}{\sqrt{n}}}$ = $\dfrac{(24 - 20)}{\dfrac{4.1}{\sqrt{9}}}$ = 2.927

For $\alpha = 0.01$ from t-distribution table for n-1=8 degrees of freedom

Matlab code

```
tinv(0.01,8)
```

```
ans = -2.8965
```

P(Z>2.927)<0.01

As 2.927>|-2.8965| the null hypothesis is rejected at significance level $\alpha = 0.01$

Therefore there is sufficient evidence to conclude that $\mu = 20$ is unlikely and the reasonable conclusion from the sample is $\mu > 20$

Problem 10

In Chennai suburbs, power cuts during summer months (and even otherwise) were quite common. In one such suburb, there was a complete blackout and complaints on the duration of the power cut were quite variable. The electricity board conducted a survey of 25 randomly chosen families and found that the mean duration of the power cut was 12 hours and the sample variance was 4 (hours)2. Construct a 98% CI on the variance assuming the population of families suffering power cuts to be normally distributed (for statistical purposes only).

Solution:

n=25, mean=12, variance = 4

degrees of freedom = 25-1 = 24

For chi-square distribution

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

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

Since 98% confidence interval, α = 0.02

The critical values of χ^2 at $\frac{\alpha}{2}$ = 0.01 and $1-\frac{\alpha}{2}$ = 0.99 and for 24 degrees of freedom are

The matlab code

```
chi2inv(0.01,24)
chi2inv(0.99,24)
```

The corresponding limits on σ^2 for $\chi^2_{0.01}$ =10.8564 is $\sigma^2 = \frac{(4) * 24}{10.8564}$ = 8.843 and for $\chi^2_{0.99}$ = 42.9798 is $\sigma^2 = \frac{4 * 24}{42.9798}$ = 2.234

The corresponding values of σ are $\sigma = 2.97$ and $\sigma = 1.149$

Hence, for the 98% confidence interval variance is from 2.234 hours to 8.843 hours

Question 11:

The goodness of fit is used to test whether results may have come from a specified population. So we compare the actual results with those predicted from the hypothesized population. An experiment is conducted by tossing a dye 96 times and recording the values shown on the face in the table below. It is hypothesized that the results belong to the population of face values formed from tossing of a fair die.

The goodness of fit test between observed (o_i) and hypothesized (e_i) frequencies is determined from the chi-square distribution with k degrees of freedom as follows

$${}_{k}^{2} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Face Value	1	2	3	4	5	6
Experiment (O_i)	19	15	17	18	14	
Hypothesized (e_i)						

a. Fill in the missing information for face value 6.

b. Fill the hypothesized values (e_i) for the different face values in the above table

c. Identify the degrees of freedom k

d. For the proposed population to fit the experimental data well, should the ${}_{k}^{2}$ value be high or low? If you conclude that there is indeed goodness of fit, what can you conclude about the fairness of the die?

Solution:

```
clear
```

a) Total number of experiments = 96. To find O_6

```
O_6 = 96 - sum([19,15,17,18,14])
```

```
O_6 = 13
```

Therefore, frequency of face value 6 = 13

Experimental values O = [19,15,17,18,14,13]

```
o = [19,15,17,18,14,13]
```

```
o = 1×6

   19    15    17    18    14    13
```

b) Hypothesized values:

When a fair die is tossed 96 times, there is equal probability for each of the 6 face values to show up.

So frequency of each face value will be 96/6 = 16

Hypothesised values H = [16,16,16,16,16,16]

```
e = [16,16,16,16,16,16]
```

```
e = 1×6

   16    16    16    16    16    16
```

c) Degrees of freedom k = 6

d) Given $\chi_k^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$

Critical value: 12.592, for given degrees of freedom

Formula for $\chi_k^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$

= 1.1875

We accept H0 as 12.592 ≥ 1.1875.

We have statistically significant evidence at α=0.05 to show that H0 is true.

If the experimental data is well fit, then $o_i - e_i$ tends to 0.

Therefore χ_k^2 value tends to 0.

So to fit well, χ_k^2 value should be low

If there is indeed goodness of fit, then we can assume the dye is very fair and has equal probability of getting any of the face values.

Problem 12

Two F-distributions are plotted below (shaded region denotes probability). Fig. 1a has 5 numerator degrees of freedom and unknown ν denominator degrees of freedom.

Note: Only Parts a-c refer to Fig.1

1. Use the information given in Fig. 1(**a)** and find ν**.**
2. After interchanging the numerator and denominator degrees of freedom, we get Fig. 1(b). Find the numerical value of **A**.
3. Find the mean and variances for both the distributions.
4. If the numerator and denominator degrees of freedom are *identical*, find f such that

P(F>f,,)=P(F>1/f1-, ,)

Then what is ?

Solution:

a) For Fig. 1(a), numerator degrees of freedom, m = 5

We take $P(F > f_{\alpha,m,n}) = \alpha$

From above eqn and fig 1, we get $\alpha = 0.15$ and $f_{0.15,5,n} = 1.730$

From this, we get value of n = 40

```
fcdf(1.730,5,40,"upper")
```

```
ans = 0.1500
```

b) For Fig. 1(b), given $\alpha = 0.85$

Interchanging degrees of freedom, we get m = 40, n = 5

```
A = finv(0.15,40,5)
```

```
A = 0.5781
```

Therefore, numerical value of A = 0.5781

c) When degree of freedom in the numerator is m and that in the denominator is n, the mean and variance is given by

$$\mu = \frac{n}{n - 2} \text{ and } \sigma^2 = \frac{2 * m^2 * (m + n - 2)}{m * (n - 2)^2 * (n - 4)}$$

For fig 1, m = 5, n = 40

```
m = 5
```

```
m = 5
```

```
n = 40
```

```
n = 40
```

```
mu = n/(n-2)
```

```
mu = 1.0526
```

```
V = (2*(m^2)*(m+n-2))/(m*((n-2)^2)*(n-4))
```

```
V = 0.0083
```

For fig 2, m = 40, n =5

```
m = 40
```

```
m = 40
```

```
n = 5
```

```
n = 5
```

```
mu = n/(n-2)
```

```
mu = 1.6667
```

```
V = (2*(m^2)*(m+n-2))/(m*((n-2)^2)*(n-4))
```

```
V = 382.2222
```

Therefore, for fig 1, mean = 1.0526 and variance = 0.0083

And for fig 2, mean = 1.667 and variance = 382.222

d) Given numerator and denominator degrees of freedom are identical

Also, $P(F > f_{\alpha,m,m}) = P\left(F > \frac{1}{f_{1-\alpha,m,m}}\right)$

We know that $f_{\alpha,m,n} = \frac{1}{f_{1-\alpha,n,m}}$

In this case, since m = n, $f_{\alpha,m,m} = \frac{1}{f_{1-\alpha,m,m}}$

Therefore α can take any value in (0, 1)

Problem 13.

Answer the following briefly

1. How will demonstrate that your experimental data are affected by only random errors?
2. If a hypothesis is rejected at a particular significance level, can it be accepted at a higher level of significance? When increasing the level of significance, is the confidence interval becoming broader or narrower?
3. Find the degrees of freedom (dof) in T-distribution such that P(Z>2.086404)=P(Tdof>2.5). Why is the value corresponding to the T-distribution higher than the Z value?

Solution:

Part (A):

Random errors are statistical fluctuations (in either direction) in **the** measured **data** due to **the** precision limitations of **the** measurement device. **Random errors can** be evaluated through statistical analysis and **can** be reduced by averaging over a large number of observations (see standard **error**).

Part (B):

As you draw larger and larger random samples from the same population, the **confidence intervals** tend to **become narrower**.

- As you **increase** the **confidence level** for a given same sample, say from 95% to 99%, the range **becomes wider**.
- To have greater confidence that an interval contains the parameter, it makes sense that the range must become wider.
- Conversely, a narrower range is less likely to include the parameter, which lowers your confidence.
- A confidence interval for the mean says nothing about the dispersion of values around the mean.

Part (C):

P(Z>2.086404)=P(Tdof>2.5)

As we increase the level of significance, the confidence interval becomes narrower.

t = (x-mu)/(s/sqrt(n))

z = (x-mu)/sigma

P(Z > 2.086404)=P(Tdof > 2.5)

```
t_min = 2.5;
z_min = 2.086404;

n = (t_min/z_min)*(t_min/z_min); %% degrees of freedom (n)
```

The reason the value corresponding to the T-distribution higher than the Z value:

In the expression of t-score, we have an additional factor of square root of (degrees of freedom); and as n is greater than equal to 1, the multiplying factor is greater than 1.

Hence, t- value is higher than z - value for corresponding distributions.

Part (D):

Using $f_{0.90,4,6}$ how will you find $f_{0.10,4,6}$? Demonstrate your procedure.

Solution:

t-static $t_{\alpha} = -(1 - t_{\alpha})$ for same degrees of freedom because t-distribution is symmetric about a mean of zero

```
fcdf(0.90,4,6)
ans = 0.4812
fcdf(0.10,4,6)
ans = 0.0215
```

Part (E):

If there are n independent standard normal variables Z1, Z2,...Zn, what will be the distribution formed by $\sum_{i=1}^{n} Z_i^2$

Solution:

Let Y = $\sum_{i=1}^{n} Z_i^2$

Since Z1,Z2,....Zn are independent the moment generating functions of Y are

$$M_Y(t) = M_{z_1^2}(t) * M_{z_2^2}(t) *.......*M_{z_n^2}(t)$$

Each Z_i^2 follows χ_i^2 that is chi-square distribution and therefore it has mgf equal to $(1-2t)^{-\frac{1}{2}}$

M_Y (t) = $(1-2t)^{\frac{-n}{2}}$. This is the mgf of Γ(n/2 , 2), and it is called the chi-square distribution with n degrees of freedom.

Problem 14.

Two monitoring stations in a river test concentrations of a pollutant. In the first station, 15 samples had a standard deviation of 3.07 ppm while in the second station, 12 samples had a standard deviation of 0.80. Construct a 98% confidence interval for $\frac{\sigma_1}{\sigma_2}$.

Solution:

First station : m=15, $s_1 = 3.07$

Second station: n=12, $s_2 = 0.80$

$\alpha = 0.02$

100(1-α) percentage confidence interval on the ratio of variances is given by

$$f_{1-\frac{\alpha}{2},n-1,m-1}\frac{s_1^2}{s_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le f_{\frac{\alpha}{2},n-1,m-1}\frac{s_1^2}{s_2^2}$$

Substituting in above equation

$$f_{0.99,11,14}\frac{3.07^2}{0.80^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le f_{0.01,11,14}\frac{3.07^2}{0.80^2}$$

Matlab code

```
finv(0.99,11,14)
ans = 3.8640
finv(0.01,11,14)
ans = 0.2329
min=0.2329*(3.07/0.80)^2
min = 3.4298
max=3.8640*(3.07/0.80)^2
max = 56.9028
```

Therefore $3.430 \le \frac{\sigma_2^2}{\sigma_1^2} \le 56.903$

$1.852 \le \frac{\sigma_2}{\sigma_1} \le 7.543$

Problem 15.

A normal distribution of marks has mean 50 and standard deviation 15. Find the following using Matlab

- Show that P(X>90)=P(X<10)
- For what value of marks **M** is the probability 0.5?
- For what value of marks **M** is the probability 0.75?
- For what value of marks **M** is the probability 0.25?
- Solve for **v** such that P(Z>**v**) = 0.025
- Solve for **w** such that Pα(X<**w**) = 0.5*P(Z<w-mu/20)

Solution:

P(X>90)=P(X<10)

The matlab code for P(X>90) and P(X<10)

```
1-normcdf(90,50,15)
ans = 0.0038
```

ans=0.0038

```
normcdf(10,50,15)
ans = 0.0038
```

ans=0.038

P(X>90)=P(X<10)=0.0038

P(X)=0.5

```
norminv(0.5,50,15)
ans = 50
```

ans=50

for M=50 the probability is 0.5

P(X)=0.75

```
norminv(0.75,50,15)
ans = 60.1173
```

ans=60.1173

for M=60 the probability is 0.75

P(X)=0.25

```
norminv(0.25,50,15)
```

```
ans = 39.8827
```

ans=39.8827

for M=39.88 ≈ 40 the probability is 0.25

P(Z>v)=0.025

P(Z<v)=1-0.025=0.975

```
norminv(0.975,50,15)
```

```
ans = 79.3995
```

ans=79.3995

v=79.39 ≈ 79.4

P(X<w)=0.5P(z<w- $\mu/20)$

```
1-chi2cdf(10,5)
```

```
ans = 0.0752
```

Problem 16

Using the T-distribution, what is the degrees of freedom such that

- **a.** P(T>1.25)=0.15 and P(T<1.25)=0.85
- **b.** Find P(T_{35}>0.8)

Solution:

a) Given $P(T > 1.25) = 0.15$ and $P(T < 1.25) = 0.85$

From f-table for $\alpha = 0.15$, degrees of freedom = 3

b) Find $P(T_{35} > 0.8)$

```
ans_b = tcdf(0.8,34,'upper')
```

```
ans_b = 0.2146
```

Therefore, $P(T_{35} > 0.8) = 0.2146$

Problem 17

For the following probability values (based on the upper tail) taken from a t-distribution with 20 degrees of freedom, find the corresponding t-values.

a. p=0.5 **b.** p=0.3 **c.** p=0.7 **d.** p= 0.80 **e.** p=0.2

Solution:

a) Given $P(T_{20} > t) = 0.5$

```
syms t
nu = 20
```

```
nu = 20
```

```
t = tinv(0.5,nu-1)
```

```
t = 0
```

Therefore, corresponding t-value is 0

b) Given $P(T_{20} > t) = 0.3$

```
t = tinv(0.7,nu)
```

```
t = 0.5329
```

Therefore, corresponding t-value is 0.5329

c) Given $P(T_{20} > t) = 0.7$

```
t = tinv(0.3,nu)
```

```
t = -0.5329
```

Therefore, corresponding t-value is -0.5329

d) Given $P(T_{20} > t) = 0.8$

```
t = tinv(0.2,nu)
```

```
t = -0.8600
```

Therefore, corresponding t-value is -0.86

e) Given $P(T_{20} > t) = 0.8$

```
t = tinv(0.8,nu)
```

```
t = 0.8600
```

Therefore, corresponding t-value is 0.86

Problem 18

A random sample of class marks are taken in an inspection and the values of the random variables are recorded as follows

$$\mathbf{X} = [0\ 15\ -5\ 20\ 10\ 9\ 14]$$

- **a.** Sample size
- **b.** Sample mean
- **c.** Sample standard deviation
- **d.** Sample variance
- **e.** Construct the two-sided 99% confidence interval for mean μ
- **f.** Construct the one-sided 95% confidence interval for the lower bound on μ
- **g.** Construct the one-sided 98% confidence interval for the upper bound on μ

Solution:

Given $X = [0\,, 15, -5, 20, 19, 9, 14]$

```
X = [0,15,-5,20,10,9,14]
```

```
X = 1×7

   0    15    -5    20    10     9    14
```

```
n = numel(X)
```

```
n = 7
```

```
xbar = mean(X)
```

```
xbar = 9
```

```
sigma = std(X)
```

```
sigma = 8.7560
```

```
V = var(X)
```

```
V = 76.6667
```

a) Sample size = 7

b) Sample mean = 9

c) Sample standard deviation = 8.756

d) Sample variance = 76.667

e) Construct the two-sided 99% confidence interval for mean

```
alpha = 0.01;
z = tinv(1-(alpha/2),6)
```

```
z = 3.7074
```

```
SE = sigma/sqrt(n)
```

```
SE = 3.3094
```

```
L = xbar - (z*SE)
```

```
L = -3.2695
```

```
U = xbar + (z*SE)
```

```
U = 21.2695
```

Therefore, lower bound = -3.2695 and upper bound = 21.2695

f) Construct the one-sided 95% confidence interval for the lower bound

```
alpha = 0.05;
z = tinv(1-alpha,6)
```

```
z = 1.9432
```

```
SE = sigma/sqrt(n)
```

```
SE = 3.3094
```

```
L = xbar - (z*SE)
```

```
L = 2.5692
```

Therefore, lower bound = 2.5692

g) Construct the one-sided 98% confidence interval for the upper bound

```
alpha = 0.02
```

```
alpha = 0.0200
```

```
z = tinv(1-alpha,6)
```

```
z = 2.6122
```

```
SE = sigma/sqrt(n)
```

```
SE = 3.3094
```

```
U = xbar + (z*SE)
```

```
U = 17.6451
```

Therefore, upper bound = 17.6451

Problem 20

Random samples of size 100 are drawn, with replacement, from two populations P1 and P2 and their means $\overline{X}_1$ and $\overline{X}_2$ are computed. If $\mu_1 = 10$ $\sigma_1 = 2$, $\mu_2 = 8$ and $\sigma_2 = 1$, find

- **a.** the probability that the difference between a given pair of sample means is less than 1.5
- **b.** The probability that the difference between a given pair of samples means is greater than 1.75 but less than 2.5.

Solution:

Given population P_1 with $\mu_1 = 10$, $\sigma_1 = 2$ and population P_2 with $\mu_2 = 8$, $\sigma_2 = 1$

Random samples are taken from P_1 and P_2 such that $n_1 = n_2 = 100$

a) Find P($\bar{X}_1$ - $\bar{X}_2$ < 1.5)

We know, $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 10 - 8 = 2$

Also, $\sigma_{\bar{X}_1 - \bar{X}_2}{}^2 = \sigma_{\bar{X}_1}{}^2 + \sigma_{\bar{X}_2}{}^2 = \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2} = \frac{4}{100} + \frac{1}{100} = 0.05$

Since n = 100 > 30, we assume the sample distributions to be normal distributions

```
mu = 2
```

```
mu = 2
```

```
sigma = sqrt(0.05)
```

```
sigma = 0.2236
```

```
ans_a = normcdf(1.5,mu,sigma)
```

```
ans_a = 0.0127
```

Therefore P($\bar{X}_1$ - $\bar{X}_2$ < 1.5) = 0.0127

b) Find P(1.75 < $\bar{X}_1$ - $\bar{X}_2$ < 2.5)

$P(1.75 < \bar{X}_1 - \bar{X}_2 < 2.5) = P(\bar{X}_1 - \bar{X}_2 < 2.5) - P(\bar{X}_1 - \bar{X}_2 < 1.75)$

```
ans_b = normcdf(2.5,mu,sigma) - normcdf(1.75,mu,sigma)
```

```
ans_b = 0.8556
```

$P(1.75 < \bar{X}_1 - \bar{X}_2 < 2.5) = 0.8556$

Problem 21

A machine is considered to be unsatisfactory if it produces more than8% defectives. It is suspected that the machine is unsatisfactory. A random sample of 120 items produced by the machine contains 14 defectives. Does the sample evidence support the claim that the machine is unsatisfactory? Use α = 0.01.

Solution:

Let Y be the number of observed defectives. This follows a binomial distribution. However, because np0 and nq0 are greater than 5, we can use a normal approximation to the binomial to test the hypothesis.

So we need to test

versus

Let the point of estimate of p be

, the same sample proportion.

Then the value of the TS is

For,

. Hence the rejection is

Decision:

As 0.137 < 2.33, ***we do not reject the null hypothesis***.

We conclude that evidence **does not support the claim** that machine performance is unsatisfactory.

Problem 22

A physician claims that the variance in cholesterol levels of adult men in a certain laboratory is at least 100. A random sample of 25 adult males from this laboratory produced a sample standard deviation of cholesterol levels as 12. Test the physician's claim at 5% level of significance.

[4]

$$H_o : \sigma^2 = 100$$

$$H_1 : \sigma^2 > 100$$

$T = \left(N - 1\right)\left(s/\sigma_o\right)^2$ = 34.56

```
chi2inv(1-0.05, 24)
```

chi2inv(1-0.05, 24) = 36.4150

$$T < \chi^2_{\alpha, N-1}$$

t-value = 0.0752

t-value > 0.05 (alpha)

Therefore, we fail to reject the null hypothesis

Problem 23

Find out how to conduct a t-test on difference of population means involving small samples and a) unknown but equal population variances and b) unknown and unequal population variances. In such cases how do you find the degrees of freedom?

Solution:

1. The two independent samples are simple random samples from two distinct populations.
2. For the two distinct populations:

● if the sample sizes are small, the distributions are important (should be normal)

● if the sample sizes are large, the distributions are not important (need not be normal)

The test comparing two independent populations means with unknown and possibly unequal population standard deviations is called the Aspin-Welch t-test.

Case (A) unknown but equal population variances

$$t = \frac{\bar{x}_2 - \bar{x}_1 + \mu_1 - \mu_2}{\sqrt{\left(\frac{s^2}{n_1} + \frac{s^2}{n_2}\right)}}$$

degrees of freedom : $\mathrm{df} = n_1 + n_2 - 2$

Case (B) unknown and unequal population variances

$$t = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\left(\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}\right)}}$$

The **degrees of freedom (** $\nu)$ associated with this variance estimate is approximated using the ***Welch-Satterthwaite equation***:

$$\nu \approx \frac{\left(\frac{s^2{}_1}{n_1} + \frac{s^2{}_2}{n_2}\right)^2}{\left(\frac{s^4{}_1}{n_1{}^2\nu_1} + \frac{s^4{}_2}{n_2{}^2\nu_2}\right)}$$

Here,

$\nu_1 = N_1 - 1$, the degrees of freedom associated with the first variance estimate.

$\nu_2 = N_2 - 1$, the degrees of freedom associated with the 2nd variance estimate.

1. Welch's *t*-test is more robust than Student's *t*-test and maintains type - I error rates close to nominal for unequal variances and for unequal sample sizes under normality.
2. Furthermore, the power of Welch's *t*-test comes close to that of Student's *t*-test, even when the population variances are equal and sample sizes are balanced. Welch's *t*-test can be generalized to more than 2-samples, which is more robust than one-way analysis of variance (ANOVA).

Problem 24

The intelligence quotients (IQs) of 17 students from one area of a city showed a sample mean of 106 with a sample standard deviation of 10, whereas the IQs of 14 students from another area chosen independently showed a sample mean of 109 with a sample standard deviation of 7. Is there a significant difference between the IQs of the two groups at α = 0.02? Assume that the population variances are equal.

Solution:

$$H_o : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

$S_p^2 = \left(n_1 - 1\right) * s_1^2 + \left(n_2 - 1\right) * s_2^2 / \left(n_1 + n_2 - 2\right)$ = 77.137

$T = \mu_1 - \mu_2 / \left(S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$ = -0.946

$$T < t_{0.02, n_1 + n_2 - 2}$$

```
tcdf(-0.946,29)
```

t value = 0.176

t value > 0.02 (alpha)

Therefore, we fail to reject the null hypothesis.

Problem 25

Show how to find the one degree of freedom chi-square distribution value when the upper 100 α percentage point is specified. For example, if the 15% point in the chi-square distribution is required, how will you find it from the normal distribution?

Solution:

- A random variable has a Chi-square distribution if it can be written as a sum of squares of x1, x2, x3, ..., xn, where these are mutually independent standard normal random variables.
- The importance of the Chi-square distribution stems from the fact that sums of this kind are encountered very often in statistics, especially in the estimation of variance and in hypothesis testing.
- The Chi Squared distribution ChiSq(n) can be approximated by a Normal distribution for large n. The ChiSq(n) distribution is the sum of n independent (Normal(0,1))2 distributions, so ChiSq(a) + ChiSq(b) = ChiSq(a+b). A (Normal(0, 1))2 = ChiSq(1) distribution is highly skewed (skewness = 2.83).
- Central Limit Theorem says that ChiSq(n) will look approximately Normal when n is rather large. A good rule of thumb is that n > 50 or so to get a pretty good fit. In such cases, we can make the following approximation by matching moments (i.e. using the mean and standard deviation of a ChiSq(n) distribution in a Normal distribution):

  ChiSq(n) » Normal (v, sqrt(2v))

- The ChiSq(n) distribution peaks at x = n-2, whereas the Normal approximation peaks at n, so acceptance of this approximation depends on being able to allow such a shift in the mode. Of course as n gets large, the difference becomes relatively small.
- 15% point for chi-square distribution is square of 15% point of normal distribution

Problem 26

An environmental engineer carries out a t-test for 9 samples from a polluted lake and obtains a |t|-value of 2.306. However he has forgotten to take the t-tables and has only the F-tables. How will he estimate the required probability value using the F-tables?

Given $n = 9$, $|t|$_value $= 2.306$, degrees of freedom v = 8

$$t_v = \frac{Z}{\sqrt{\frac{\chi_v^2}{v}}} = \frac{\sqrt{\frac{\chi_1^2}{1}}}{\sqrt{\frac{\chi_v^2}{v}}} = \sqrt{F_{1,v}}$$

$$F_{1,v} = {t_v}^2$$

$$F_{1,8} = {t_8}^2 = 2.306^2 = 5.318$$

```
alpha = fcdf(5.318,1,8,"upper")
```

```
alpha = 0.0500
```

Therefore, we get $\alpha = 0.05$

Problem 27

Eagle Eye is used in cricket to track the trajectory of the ball. The equipment has been tested rigorously on many overseas cricket pitches over 5 years. After a large number of tests it uses the standard deviation (σ) in bounce of the ball pitched at good length as 50 cm in its tracking calculations. The Eagle Eye tracker is then tested through 5 independent trials in India and the measured standard deviation in the cricket bounce for the ball pitched on good length based on the measurements carried out is 25.74 cm. Can the Eagle Eye be used reliably to track the ball (to give lbw decisions) on Indian pitches?

Solution:

Mean = 50

Standard deviation = 25.74

n = 5

```
CI = tinv(0.975, 4)*25.74/sqrt(4);
T = 50/CI*tinv(0.975, 4);
x = tinv(0.975, 5);

%%%%%%%%%

mu = 50;
sigma = 25.74;
n = 5;

rng default   % For reproducibility
x = normrnd(mu,sigma,n,1);
xbar = mean(x);
se = std(x)/sqrt(n);
nu = n - 1;
conf = 0.95;
alpha = 1 - conf;
pLo = alpha/2;
pUp = 1 - alpha/2;
crit = tinv([pLo pUp], nu);
ci = xbar + crit*se;
```

T-stat = 3.8850 ....(Case:1 ... Alpha=0.25)

ci = [8.0187 105.3006] ....(Case:2 ... Alpha=0.50)

Hence, we can not use the Eagle Eye reliably to track the ball (to give lbw decisions) on Indian pitches.

Problem 28

How will you estimate the probability in distributions involving chi-square using the F distribution tables?

Solution:

When m_1 is in the degrees of freedom in the numerator and m_2 in the denominator,

$$F_{m_1,m_2} = \left(\frac{\frac{\chi^2_{m_1}}{m_1}}{\frac{\chi^2_{m_2}}{m_2}} \right)$$

Here we are able to express F-distribution in terms of χ^2 distribution

If we take limit m_2 as infinity

$$F_{m_1,m_2} = \lim_{m_2 \longrightarrow \infty} \left(\frac{\frac{\chi^2_{m_1}}{m_1}}{\frac{\chi^2_{m_2}}{m_2}} \right) = \frac{\chi^2_{m_1}}{m_1}$$

$$\chi^2_{m_1} = F_{m_1,\infty} * m_1$$

Therefore, using F-distribution table, and we know the degrees of freedom, we can find the χ^2 value and thus the probability of he distribution can be found

If there are a large number of observations (i.e. v2 is large), then the shape of the F distribution is very similar to the chi squared distribution with v1 degrees of freedom although there is a shift in position (chi squared equals v1 F, and for 1 degree of freedom, they are both the same as v1 = 1). If both v1 and v2 are large, the F distribution also resembles the normal distribution, with a mean of 1.

So to find out value of chi-square with n degrees of freedom using the F table we can look for values of F (n, m) where m is a very large number.

Problem 29

a) Show that $f_{1-\alpha,m_1,m_2} = \dfrac{1}{f_{\alpha,m_2,m_1}}$

Proof of Part (a):

$$\frac{1}{F_{(m,n)}} = F_{(n,m)}$$

by the reciprocal of the chi square distributions, then let $X \sim F_{(n,m)}$

That means $F_{(n,m)} \sim \dfrac{1}{X}$

$$\alpha = P(X \leq F_{\alpha,(m,n)})$$

$$\alpha = P\left(\frac{1}{X} \geq \frac{1}{F_{\alpha,\,(m,n)}}\right)$$

$$\alpha = 1 - P\left(\frac{1}{X} \leq \frac{1}{F_{\alpha,\,(m,n)}}\right)$$

$$\alpha = 1 - P\left(F_{(n,m)} \leq \frac{1}{F_{\alpha,\,(m,n)}}\right)$$

$$\alpha = 1 - F\left(\frac{1}{F_{\alpha,\,(m,n)}}\right)$$

Thus,

$$1 - \alpha = F_{(n,m)}\left(\frac{1}{F_{\alpha,\,(m,n)}}\right)$$

$$F_{1-\alpha,\,(n,m)} = \left(\frac{1}{F_{\alpha,\,(m,n)}}\right)$$

Hence proved.

b) From above relation, $f_{\alpha,m_1,m_2} = \dfrac{1}{f_{1-\alpha,m_2,m_1}}$

Substituting $\alpha = 0.90$, $m_1 = 4$, $m_2 = 6$

$$f_{0.90,4,6} = \frac{1}{f_{0.10,6,4}}$$

Cross-multiplying, $f_{0.10,6,4} = \dfrac{1}{f_{0.90,4,6}} = 4.01$

```
ans_d = finv(0.9,6,4)
```

```
ans_d = 4.0097
```

Problem 30

A sample of 40 alumni from IIT Madras (batch of 2010) working in Indian Organizations is selected to find whether the average annual income may be taken to belong to a population with mean of Rs. 30 lakhs. Population variance is not known. The population may not be assumed to be normally distributed.

(a) Explain briefly how will you go about testing the hypothesis that μ = 30 lakhs.

(b) If global slowdown of the economy are being considered, what would be the alternate hypothesis?

To compare between IITs, a random sample of 50 alumni from IITX (also batch of 2010) is also taken.

(c) How will you go about testing the null hypothesis that the mean salaries of students graduating from both Institutions are the same and the alternate hypothesis that the average salary of IITM is higher than IITX?

State the assumptions that have been made.

Solution:

a)

Given $n = 40$, Also to be noted that population is not normally distributed

Since n=40, which is greater than 30, we can assume that the sample means of the population form a normal curve.

Null Hypothesis: $H_0 : \mu = 30 \text{ lakhs}$

We use t-distribution. If $t_{0.39}$ is in the acceptance region, then the null hypothesis is accepted. Else null hypothesis is rejected.

b)

If a global slowdown of economy is considered, then average annual income in population means will be less than 30 lakhs

Alternate Hypothesis: $H_1 : \mu < 30 \text{ lakhs}$

c)

Assumption: Populations variances are equal for both samples and they are unknown

Null Hypothesis: $H_0 : \mu_1 = \mu_2$

Alternate Hypothesis: $H_1 : \mu_1 > \mu_2$

$$t = \frac{\mu_{\bar{X}_1} - \mu_{\bar{X}_2}}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

We can use t-value and degrees of freedom to find probability

if probability <= α , we reject null hypothesis. Else it is accepted