15th February 2023 — Practical Day 1

χ^2 distribution

Let $X_1, X_2, \ldots, X_n$ be n mutually independent standard normal ~~variables~~ random variables. Then the sum of squares $\boxed{U = \sum_{i=1}^{n} (X_i^2)}$ is said to follow χ^2 distribution with n degrees of freedom.

It has the following pdf

$$f(u) = \begin{cases} \frac{1}{2^{n/2}\Gamma_{n/2}} \left[e^{-u/2} u^{n/2 - 1} \right] & 0 < u < \infty \\ 0 & \text{o.w.} \end{cases}$$

The distribution is characterized by its degrees of freedom (n).

We will write it as $U \sim \chi^2_n$

t Distribution — 22nd February 2023 — Practical Day 2

Let X be a standard normal random variable & Y be another random variable following chisq dist^n with degrees of freedom n. Furthermore, X & Y are independent. Then the new variable $\boxed{t = \frac{X}{\sqrt{Y/n}}}$ is said to follow t-distribution, with n degrees of freedom.

It has the following pdf

$$f(t) = \begin{cases} \frac{1}{\sqrt{n}\, B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{\left(\frac{-n-1}{2}\right)} & t \in \mathbb{R} \\ 0 & \text{o.w.} \end{cases}$$

We write it as $t \sim t_n$

Q2) For $n = 10, 20, 30, \ldots, 60$ plot the pdf of a t_n distribution in seperate panels of the same graph & comment.

c1 := t [-5 to 5 with step value 0.01 make patterned data]

c2 := f(t) (n=10)
c3 := f(t) (n=20)
⋮
(n=60)

Calc → Prob Dist → t
Prob density
Input → t
Optional storage → c2

We notice that as $n \to \infty$, t-dist^n tends to be normal dist^n.

Graph : same as chisq.
[Apply overlaid graphs to understand skewness and kurtosis]

Skewness → Symmetric
Area under tail decreasing with increasing n.
▸ Mesokurtic
Kurtosis : Area under tail / heavy or thick tail
As k ↓ , ⟼ mesokurtic
Area under tail ↑ Heavy / thick tail ↑

Q3) Plot the pdf of the t-distribution having 10 df and the std normal dist^n on the same graph and comment.
[-4 to 4 gap 0.01]
Kurtosis more for t-dist^n.

→ F-distribution
let X and Y be two independently distribution χ^2 random variables with degrees of freedom m & n respectively. Then the new random variable

$F = \frac{(x/m)}{(y/n)}$ is said to follow F-distribution with (m, n) degrees of freedom. It has the following pdf.

$$F(f) = \begin{cases} \frac{(m/n)^{m/2}}{B(m/2, n/2)} f^{\left(\frac{m-2}{2}\right)} \left(1 + \frac{m}{n} f\right)^{\left(\frac{m+n}{2}\right)} & 0 < f < \infty \\ 0 & 0.w \end{cases}$$

We write it as $F \sim F_{m,n}$

Q4) Plot the pdf of the F distribution having df $(n1, n2)$ for the following choices of n_1 & n_2

i) $n1 = 10$ & $n2 = 10, 20, 30, 40$
ii) $n1 = 20$ & $n2 = 10, 20, 30, 40$
iii) $n1 = 30$ & $n2 = 10, 20, 30, 40$
iv) $n1 = 40$ & $n2 = 10, 20, 30, 40$

+ve skewed
Kurtosis decrease

```
c1 := f
c2 := n2 = 10
```

1st March 2023

Tests & Confidence intervals related to univariate normal popⁿ. (A1)

Practical Day 3

Q5) Nine items of a sample had the following values 45, 47, 50, 52, 48, 47, 49, 53 & 51. Does the mean of the normal population from which these 9 items are drawn differ significantly from an assumed popⁿ mean of 47.5 (i) when popⁿ s.d is ~~unknown~~ is given to be 2.5 and (ii) when popⁿ s.d is unknown For case (ii), obtain 95% confidence interval for the popⁿ mean.

Let X be a random variable denoting the value of a randomly selected item.

Given that $X \sim N(\mu, \sigma^2)$; Sample size $n = 9$.

let $X_1, X_2, \ldots, X_n$ be a random sample from the distⁿ of X. We want to test $H_0 : \mu = 47.5$ against $H_1 : \mu \neq 47.5$

ii) Since σ is unknown, here, the apt test is one sample t-test. Here the test statistic is given by $t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

then $t \sim t_{n-1}$

$|t_{obs}| > t_{\alpha/2;\, n-1}$. t denotes upper $(\alpha/2)^{th}$ quantile of t-dist^n with $(n-1)$ df.

Calc $\longrightarrow$ st^d deviation (Input C1)

$s = 2.61937$ and $t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = -1.84523$

$\alpha = 0.05$ and $\alpha/2 = 0.025$

$P(t > t_{\alpha/2;\, n-1}) = \alpha/2 \Rightarrow P(t < t_{\alpha/2;\, n-1}) = 0.975$

$\Rightarrow t_{\alpha/2;\, n-1} = P_t^{-1}(0.975)$

$= 2.306$

and $|t_{obs}| = 1.84523$

$\therefore |t_{obs}| < t_{\alpha/2;\, n-1}$ is what we get.

$\therefore$ On the basis of the given sample we **do not** reject H_0 at 5% level of significance

$\therefore$ We can conclude that the pop^n mean is not significantly different from 47.5

① Test of μ when σ^2 is known.

i) Given that $\sigma = 2.5$. Since σ is known, the apt test should be 1-sample Z-test.

Here the test statistic is given by

$$Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} \sim N(0,1)$$

Both-tailed test

We will reject H_0 at level α if $|Z_{obs}| > Z_{\alpha/2}$ where $Z_{\alpha/2}$ denotes the upper $\alpha/2^{th}$ quantile of $N(0,1)$

1-sample ⇒ 1 set of sample
2-sample ⇒ 2 set of sample

Calculation: $n = 9$; $\mu_0 = 47.5$; $\sigma = 2.5$; $\bar{X} = \frac{1}{n}\sum X_i$ = 49.11 (using minitab)

$$Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} = 1.932$$

$\alpha = 1 - 0.95 = 0.05$

$\therefore Z_{\alpha/2} = Z_{0.025}$

[Calc → Prob dist^n → Normal → Inverse cumulative]

$P(Z > Z_{\alpha/2}) = \alpha/2 \Rightarrow P(Z < Z_{\alpha/2}) = 1 - \alpha/2 = 0.975$

$\Rightarrow Z_{\alpha/2} = \Phi_Z^{-1}(0.975)$

$= 1.959$

[Input constant → 0.975 ... Optional storage : blank]

$\therefore Z_{obs} = 1.933$. $\therefore |Z_{obs}| = 1.933 < 1.959 = Z_{0.025}$

Hence, on the basis of the given sample, we do not reject H_0 at 5% level of significance.

∴ We can conclude that the population mean is not significantly different from 47.8.

15th March 2023 (A) (A2) Practical Day 4

Q6) A firm manufacturing rivets to limit variation in their lengths as much as possible. The length (in cms) of 10 rivets manufactured by a new process is 2.10, 1.94, 2.0, 2.07, 2.12, 1.96, 1.93, 1.98, 2.20, 1.88. In the past the std deviation of length of rivets by the firm has been 0.145 cms. Examine whether the new process seems to be superior to the old in case (i) the mean length is known to be 2.00 cm and (ii) the mean length is unknown. Obtain 95% confidence interval for the std deviation when the mean length is unknown. Assume the length of rivets to be normally distributed.

Sol^n:

Let X be a random variable denoting the length (in cm), of a randomly selected rivet manufactured by the new process of the firm.

In the past, the standard deviation of the rivets manufactured by the firm has been 0.145 cm $= \sigma_0$ (say)

Let $X_1, X_2, \ldots, X_{10} \overset{iid}{\sim} X$ be a random sample of size 10.

We are given a random sample of size 10. The sample values are 2.10, 1.91, 2.00, 2.07, 2.12, 1.96, 1.93, 1.98, 2.20, 1.88.

Given mean length is 2.00 cm. That is the population mean μ_0 is given.

We will be testing σ^2, the following hypothesis $H_0 : \sigma = 0.145$ against $H_1 : \sigma < 0.145$

Reason behind the formulation of such an hypothesis : We need to check if the new process (whose sample values are given) is superior than the old or pre-existing process of manufacturing by the firm. which means superiority implies less variation and hence the hypothesis.

(a) Under H_0,

Consider $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$

let $T = \sum_{i=1}^{n}\left(\frac{x_i - \mu_0}{\sigma_0}\right)^2 = \left(nS_0^2/\sigma_0^2\right)$

Clearly, $T \sim \chi^2_n$ if

we know if $S_0^2 < (0.145)^2$, i.e a lower value of $\frac{S_0^2}{0.145^2}$ less than 1 will indicate departure from H_0.

Given level of significance is $\alpha = 0.05$.

Thus $P(\text{Type 1 error}) = \alpha \Rightarrow P(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha$

$$\Rightarrow P\left(T < k_1 \mid \sigma_0 = 0.145\right) = \alpha$$

$$\Rightarrow k_1 = F^{-1}_{\chi^2_n}(\alpha) = \chi^2_{n;1-\alpha}$$ the lower α-point

Hence the test rule is reject H_0 at α level of significance iff $T_{obs} < \chi^2_{n;1-\alpha}$

We see $T_{obs} = \frac{n s_0^2}{\sigma_0^2}$; $s^2 = 0.00942$

$\therefore T_{obs} = 4.48038$

and using minitab we find that $\chi^2_{n;1-\alpha} = 3.94030$

Hence we see $T_{obs} = 4.48 \not< 3.9403 = \chi^2_{n;1-\alpha}$

So we do not reject H_0 at α level of significance

$\therefore H_0$ is accepted when μ is 2cm.

Hence, the new process of manufacture do not differ significantly with respect to the old one in terms of superiority.

(b) Here the mean length is unknown.

Consider $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ where $\bar{x}$ = sample mean.

Now consider the test statistic $T = \frac{(n-1)s^2}{\sigma_0^2}$ under H_0

Then $T = \sum\left(\frac{x_i - \bar{x}}{\sigma_0}\right)^2 \Rightarrow T \sim \chi^2_{n-1}$ where $n = 10$.

Under H_0, $T \sim \chi^2_{n-1}$

Now $P_{H_0}[T_o < c] = \alpha \Rightarrow$ c is the lower alpha point

Hence we will reject H_0 iff at α level of significance

iff $T_{obs} < \chi^2_{n-1;1-\alpha}$

We see $T_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$; Using minitab, we find

$(n-1)s^2 = 0.0709$ and so $T_{obs} = 4.3262$.

Now $\chi^2_{n-1;1-\alpha} = 3.32511$.

We find that $T_{obs} = 4.32 \not< 3.32 = \chi^2_{n-1;\alpha}$.

Thus, we do not reject H_0 at α level of significance

$\therefore H_0$ is accepted when mean length is unknown.

Thus, the new process of manufacture do not differ significantly with respect to the old one in terms of superiority.

[let D be a dist. Hence D_α means it is a point such that $Pr(D_{statistic} > D_\alpha) = \alpha$.

এজন্যই মোহাম্মু inv. cumulative-এতে এর ক্ষেত্রে '1' ব্যবহৃত হয়ে থাকে, সেই অনুযায়ী representation টা বদলাতে থাকবে।

Example. $K = \chi^2_{n;1-\alpha}$ means $Pr(\chi^2_{statistic} > K) = 1-\alpha$

$\Rightarrow 1 - Pr(\chi^2_{statistic} < K) = 1-\alpha \Rightarrow K = Pr^{-1}_{\chi^2_n}(\alpha)$

$\Rightarrow \chi^2_{n;1-\alpha} = Pr^{-1}_{\chi^2_n}(\alpha)$

Under null hypothesis H_0, consider the test statistic T defined as $\frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{13}+\frac{1}{12}}}$ where

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}$$ where $s_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x})^2$

and $s_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_j-\bar{y})^2$. where $\bar{x}, \bar{y}$ are sample means.

Now we know $\frac{(n_1-1)s_1^2}{\sigma^2} \sim \chi^2_{n_1}$ and $\frac{(n_2-1)s_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$

So $(n_1+n_2-2)s^2/\sigma^2 \sim \chi^2_{n_1-1} + \chi^2_{n_2-1} = \chi^2_{n_1+n_2-2}$

Now $\bar{x} \sim N(\mu_1, \frac{\sigma^2}{n_1})$ and $\bar{y} \sim N(\mu_2, \frac{\sigma^2}{n_2})$ under H_0

$$\Rightarrow (\bar{x}-\bar{y}) \sim N\left(\mu_1-\mu_2, \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right) \text{ under } H_0$$
$$= N\left(0, \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

$$\Rightarrow \frac{\bar{x}-\bar{y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1).$$

and $(n_1+n_2-2)s^2/\sigma^2 \sim \chi^2_{n_1+n_2-2}$

$$\therefore T = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{(\bar{x}-\bar{y})\Big/\left[\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]}{\sqrt{\frac{s^2(n_1+n_2-2)}{\sigma^2}\Big/(n_1+n_2-2)}}$$

$$\Rightarrow T = \frac{N(0,1)}{\sqrt{\chi^2_{n_1+n_2-2}/n_1+n_2-2}} \sim t_{n_1+n_2-2}$$

$$\Rightarrow T \sim t_{n_1+n_2-2} \text{ under } H_0$$

So accordingly, we will be rejecting H_0 iff we get $T_{obs} < t_{-\alpha;\, n_1+n_2-2}$ at $\alpha = 0.05$ level of significance.

Using 2-sample t-test in minitab, we find $T_{obs} = 0.89$ and value of $s = 8.3805$

and $t_{\alpha;\, n_1+n_2-2} = -1.71$ [Here $\alpha = 0.05$, $n_1 = 13$, $n_2 = 12$]

Thus, we see $T_{obs} = 0.89 \not< -1.71 = t_{\alpha;\, n_1+n_2-2}$

Conclusion: Thus we do not reject H_0 at α level of significance. Hence, the average daily milk production (in pounds) do not differ significantly between cows fed with field watered alfalfa and dewatered alfalfa.

CI types

CI for μ (σ unknown) : 1-sample T (using minitab)

CI for μ (σ known) : 1-sample Z (using minitab)

CI for σ (μ known) : χ^2_n (manually)

CI for σ (μ unknown) : χ^2_{n-1} (manually)

CI for $(\mu_1 - \mu_2)$ ($\sigma_1 = \sigma_2$ unknown) : 2-sample T (using minitab)

CI for σ_1/σ_2 (μ_1, μ_2 unknown (Sample std will be provided)) : 2-variances (using minitab)

CI for σ_1/σ_2 (μ_1, μ_2 known) : $F_{n_1;n_2}$ (manually)

$$\left(\frac{s_1^2}{s_2^2} \left(F_{\alpha; n_1, n_2} \right)^{-1}, \frac{s_1^2}{s_2^2} \left(F_{1-\alpha_2; n_1, n_2} \right)^{-1} \right).$$

Practical Day6.
5th April 2023.

9) It is known that the mean diameter of rivets produced by two factories is practically the same but std deviation may differ. For 22 rivets produced by factory I, the std deviation is 2.9 mm, while for 16 rivets produced by factory II the s.d is 3.8 mm. Do you think that the products of factory I are of better quality than that of factory II? Also obtain 95% confidence interval

B2b

Soln:

Let X denote the diameter of a randomly selected rivet produced by factory I.
Y denote the diameter of a randomly selected rivet produced by factory II.
X, Y are independent as the factories are different.

Let $X \sim N(\mu, \sigma_1^2)$ and $Y \sim N(\mu, \sigma_2^2)$ where μ, σ_1, σ_2 are unknown.

We are to test for the population variance ratio. For better quality, the factory having lesser variance will be preferred.

So let us test $H_0: \frac{\sigma_X}{\sigma_Y} = 1 \ (= \xi_0)$ vs $H_1: \frac{\sigma_X}{\sigma_Y} < 1 \ (= \xi_0)$

Now let $X_1, \ldots, X_{22}$ be a random sample of size 22 taken from the distribution of X

and let $Y_1, \ldots, Y_{16}$ be a random sample of size 16 taken from the distribution of Y.

From given problem we know sample standard deviation of following X - distribution is

$$s_1 = 2.9$$

and that following Y - distribution is $s_2 = 3.8$

consider the test statistic $F = \frac{S_{X_0}^2}{S_{Y_0}^2}$ where

$$S_{X_0}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad \text{and}$$

$$S_{Y_0}^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

where $$F = \frac{(n_1-1) S_{X_0}^2 / (n_1-1)\sigma_{X_0}^2}{(n_2-1) S_{Y_0}^2 / (n_2-1)\sigma_{Y_0}^2} = \frac{S_{X_0}^2}{S_{Y_0}^2} \frac{\sigma_Y}{\sigma_X}$$

$$= \frac{S_{X_0}^2}{S_{Y_0}^2} \cdot \frac{1}{\xi_0} = \frac{S_{X_0}^2}{S_{Y_0}^2}$$

Hence

$$F \sim F_{n_1-1, n_2-1}$$

Small value of T goes in favour of H_1. Hence, a left - tailed test based on T will be appropriate

Rejection Rule: We reject H_0 if $F_{obs} < F_{1-\alpha; n_1-1, n_2-1}$ where

$F_{n_1-1, n_2-1; 1-\alpha}$ denotes lower α^{th} point of F_{n_1-1, n_2-1} distribution.

Using minitab, we found $F_{obs} = 0.58$ and $F_{1-\alpha; n_1-1, n_2-1}$

Hence we fail to reject H_0 since we do not have significant evidence to conclude that products of factory I is of better quality than factory II

Practical Day 7

Methods of estimation

1. In a sample of size 8 from the exponential pop^n with specification $f(x) = \frac{1}{2}\exp\left(-\frac{1}{2}(x-\theta)\right) \quad \theta \leq x < \infty$, the observations were 3.71, 9.36, 16.24, 2.67, 6.78, 8.92, 10.22, 4.31. Give the moment estimate of θ from sample.

MME

In this problem, we are given a random sample of size 8 drawn from an exponential pop^n with specification

$$f(x) = \frac{1}{2}\left(\exp\left(-\frac{1}{2}(x-\theta)\right)\right) \quad \theta \leq x < \infty$$

We are to obtain the moment estimate of θ from the sample.

In method of moments, we equate the sample moments with the corresponding pop^n moment. If there are k unknown parameters to be estimated then generally the 1st raw moment about zero and the 'second' to the k^{th} central moments are equated.

$X_1, X_2, \ldots, X_n$ drawn from dist^n given by $f(x, \theta)$

Since here, we have to estimate θ, hence by equating the sample mean with the pop^n mean we get the eq^n.

(1st order raw moment) pop^n mean $\leftarrow \mu_1' = \bar{x}$ ← sample mean

where $\mu_1' = E(X)$, X following exp dist^n characterized by given pdf $f(x)$ and

$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$ for a given sample

$x_1, \ldots, x_n$ drawn from the dist^n of X.

Now $E(X) = \frac{1}{2}\int_{\theta}^{\infty} x e^{-\frac{1}{2}(x-\theta)}\,dx = \frac{1}{2}\,x\int_{\theta}^{\infty} e^{-\frac{1}{2}(x-\theta)}dx$

$$- \frac{1}{2}\int_{\theta}^{\infty} \frac{e^{-\frac{1}{2}(x-\theta)}}{\left(-\frac{1}{2}\right)}\,dx$$

$$= \left[\frac{1}{2}\,x\cdot\frac{e^{-\frac{1}{2}(x-\theta)}}{\left(-\frac{1}{2}\right)}\right]_{\theta}^{\infty} + \int_{\theta}^{\infty}\frac{e^{-\frac{1}{2}(x-\theta)}}{\left(-\frac{1}{2}\right)}\Big]_{\theta}^{\infty}$$

$$= [0 + \theta] + \left(+\frac{1}{2}\right) = \left(\theta + \frac{1}{2}\right)$$

Now $E(X) = \bar{x}$ so $\theta + 2 = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\hat{\theta}_{MME} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) - 2$$

Now from the given sample data, the sample mean is $\bar{x} = 7.77625$

Hence $\hat{\theta}_{MME} = \bar{x} - 2 = 5.77$ which is the required moment estimate of the θ from the sample.

Q2) In drawing 25 balls with replacement from an urn containing white and black balls. 11 are found to be white. Find the maximum likelihood estimate of P, the proportion of black balls in the urn. It is known that P can assume only one of the following values 0.3, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70. What would be MLE of P in this case?

MLE

Let X be a random variable denoting the no of black balls obtained out of 25 balls drawn randomly from an urn with replacement, containing white and black balls.

So $X \sim Bin(25, p)$, where p is the proportion of black balls in an urn.

Here it is given that out of 25 balls drawn, 11 are found to be white, so 14 are black.

Now we have only one sample where there are 14 black balls.

Hence $P(X=14) = \binom{25}{14} p^{14}(1-p)^{11}$

Now $L(p) = \prod_{i=1}^{1} f(p; \underset{\sim}{x}) = P(X=14)$

is the likelihood function which we is to be checked for all different values of p given in the question.

[Construct a table p vs $L(p)$]

p	$L(p)$
⋮	⋮

Using minitab, the table of values are noted.

We find that $L(p)$ takes maximum value when $p = 0.55$.

So the required maximum likelihood estimate is $\hat{p} = 0.55$

■

93. The length of life recorded in hours for 10 electron tubes, were 980, 1020, 995, 1015, 990, 1030, 975, 950, 1050, 870. Assume the lifetime is distributed in the form $f(x) = \frac{1}{\theta} \exp\left(\frac{-x}{\theta}\right)$; $\theta > 0$; $0 < x < \infty$. Obtain the MLE of θ. Also estimate the probability that the electron tube will survive at least 100 hours.

MLE

Let X be a random variable denoting the length of life (in hours) of a randomly selected electron tube.

Here $X \sim \text{Exp}(\text{mean} = \theta)$ given by the

pdf $\quad f(x) = \frac{1}{\theta} \exp(-x/\theta) \quad \theta > 0, \quad 0 < x < \infty$

Here we given the length of life for ten electron tubes. We have to obtain the MLE of θ.

We know that for an observed sample $(\underset{\sim}{x}) = (x_1, x_2, \ldots, x_n)$ drawn from the given dist^n, the likelihood function is given by ~~$L(\theta|x)$~~

given by $\quad L(\theta|\underset{\sim}{x}) = \prod_{i=1}^{n} f(\theta; x_i)$

$$= \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum x_i}$$

Hence the log-likelihood function is given by

$$\rightarrow L_1 = \log L = a \quad -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0$$

likelihood equation

Diff w.r.t θ

$$\Rightarrow \quad -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \quad n = \frac{1}{\theta}\sum_{i=1}^{n}(x_i)$$

$$\rightarrow \theta = \left(\frac{\sum x_i}{n}\right)$$. Now check the second derivative criteria

$L_1''(\theta) = \left(-\frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i\right) < 0$ so we get $\theta = \left(\frac{1}{n}\sum x_i\right)$ is the value for which $L(\theta)$ is maximum

Thus, $\hat{\theta}_{MLE} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$ as the maximum likelihood estimator of θ.

capitals e thakbe X_i's তখনই $\hat{\theta}$ is a test statistic in general

Hence the ML estimate is (calculate)

$$\hat{\theta}_{MLE}\Big|_{\underset{\sim}{x}} = 987.5 \text{ hrs}$$

Q1) The following table gives the frequency dist^n of the number of red blood corpuscles (RBC) per cell of a hemocytometer. Obtain the MLE for the true average no of RBC per cell using these data. Also obtain the MLE of the probability that no of RBC per cell is at most one

RBC count	0	1	2	3	4	5	+
Cells count	143	156	68	27	5	1	400

Let X be a random variable denoting the no of red blood corpuscles in a randomly chosen cell of a hemocytometer.

$\therefore X \sim \text{Poisson}(\lambda)$ where λ is unknown, follows a pdf $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$ and $\lambda > 0$.

We are to obtain MLE of λ.

So for an observed sample as given, there are total 400 observations, and the sample data is $\underset{\sim}{x} = (x_1, x_2, \ldots, x_{400})$. So the likelihood function is given by $L(\lambda|\underset{\sim}{x}) = \prod_{i=1}^{400} f(\lambda|x_i) = \left(e^{-400\lambda}\lambda^{\sum x_i}\right) \Big/ \left(\prod_{i=1}^{400}(x_i!)\right)$ (! = factorial)

Hence the log likelihood function is given by $L_1 = \log L$

$= -400\lambda + \sum_{i=1}^{400} x_i \log\lambda - \log(A)$ where A is the denominator

$\therefore$ Differentiating L_1 w.r.t λ we get the critical pt when of $L(\lambda|\underset{\sim}{x})$ equated with zero

$$-400 + \frac{1}{\lambda}\sum_{i=1}^{400} x_i \Rightarrow \lambda = \left(\frac{\sum_{i=1}^{400} x_i}{400}\right)$$

Now double differentiating L_1 w.r.t λ we get $L_1'' = -\frac{1}{\lambda^2}\sum x_i$

$\Rightarrow L_1''\Big|_{\lambda = \frac{1}{400}\sum x_i} < 0$ so clearly at $\lambda = \frac{1}{400}\sum_{i=1}^{400} x_i$,

L_1 is maximum and so is L. Thus we get the maximum likelihood estimator $\hat{\lambda} = \frac{1}{400}\sum_{i=1}^{400} X_i$. So, now using the available sample data, we get the maximum likelihood estimate

$$\lambda = \frac{1}{400}\sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda}_{MLE}\Big|_{\underset{\sim}{x}} = \frac{1}{400}\cdot\left[0(143) + 1(156) + 2(68) + 3(27) + 4(5) + 5\right]$$

$$= 398/400 = \underline{0.995} \text{ (Ans)}$$

Now Probability that no of RBC per cell is at most one is $P(X=0) + P(X=1)$. Now for this we will use the MLE so that we get MLE of the required probability, i.e using $\hat{\lambda}_{MLE}\Big|_{\underset{\sim}{x}}$ we get $P(x=0) + P(x=1)$

$= e^{-\hat{\lambda}_{MLE}} + \hat{\lambda}_{MLE}\cdot e^{-\hat{\lambda}_{MLE}} = e^{-\hat{\lambda}}(1+\hat{\lambda}) = \underline{0.737}$ (Ans)

Practical Day 8

1) Below are given the yields in gm per plot (plot size = 1/2000 acre) for three varieties of seed cotton:

Variety 1	V2	V3
77	109	46
70	106	70
63	137	73
81	79	65
95	134	61
81	79	46
101	126	98

Test if the varieties differ significantly among themselves. Take $\alpha = 0.05$

Anova 1-way layout. We have data on yields classified according to 3 varieties of seed cotton. Hence we will use 1-way layout of ANOVA to model the data.

Total 21 plots have been observed where each variety has been yielded over 7 plots.

Let us consider the model,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \begin{matrix} j = 1, 2, \dots, n_i \\ i = 1, 2, \dots, K \end{matrix}$$

where y_{ij} : yields in j^{th} plot receiving i^{th} variety

μ : general effect.

α_i : additional effect due to i^{th} variety $i = 1, 2, \dots, K$

ε_{ij} : random error associated with y_{ij}

Assumptions $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

K populations are homoscedastic.

We are interested to see whether yields vary over varieties of seed cotton. Hence, set of hypotheses will be

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_K = 0 \quad \text{vs} \quad H_1 : \text{not } H_0$$

Test statistic : $F = \frac{MSB}{MSE} \overset{H_0}{\sim} F_{k-1, n-k}$

where, $MSB = \frac{\sum n_i (\bar{y}_{io} - \bar{y}_{oo})^2}{k-1}$ and

$$MSE = \frac{\sum_{ij} (y_{ij} - \bar{y}_{io})^2}{n-k}$$

A right tailed test based on F would be appropriate. We reject H_0 at level α if (observed F) $> F_{\alpha; k-1; n-k}$

where,

$F_{\alpha; k-1; n-k}$ denotes the upper α-point of $F_{k-1; n-k}$ distribution.

[Steps : Stat → Anova → 1-way Anova → Fill in the required fields]

Calculations

k : no of varieties of seeds

n_i : total no of plots of i^{th} varieties seed $i = 1(1)3$

$\therefore n = \sum n_i = 21$ = total no of plots.

$\therefore F_{\alpha; k-1; n-k} = F_{0.05; 2; 19}$

Using minitab we found the values of the following table

Anova table

Source	df	Sum of squares	MS	obsd F	$F_{\alpha; k-1; n-k}$	Decision
Variety	2	SSB = 7089	MSB = 3544	9.85	3.55	Rej H_0
Error	18	SSE = 6477	MSE = 360			
Total	20	TSS = 13566				

We have rejected H_0 at 5% level of significance. Therefore, on the basis of the data, we can conclude at 5% level that yield varies significantly over varieties of seed cotton.

[$F_{\alpha; m; n} \Rightarrow P(F > F_{\alpha; m; n}) = \alpha \Rightarrow F_{\alpha; m; n} = P^{-1}_{F_{m;n}}(1-\alpha)$
similarly for t as well and for any dist^n]