UNIT - II

SUMMARIZING DATA & REVISITING PROBABILITY (NOS 2101)

PROBABILITY:

Probability is the chance of occurrence anything

$$P(A) = S/p$$

Where S is Sample size or no of positive outcomes and p is the population size or total no of outcomes.

Probability distribution:

Which describes how the values of a random variable are distributed.

Binominal distribution: The collection of all possible outcomes of a sequence of coin tossing.

Normal distribution:

The means of sufficiently large samples of a data population.

Example :

Probability of ace of Diamond in a pack of 52 Cards when 1 card is Pulled out at random.

Other mathematical variables.

A random variable is a real-valued function defined on the points of a Sample Space.

RANDOM VARIABLES ARE TWO BROAD CATEGORIES:

→ Random variable with discrete values

→ Bivariate Random variable

DISCRETE:

Specified finite or countable list of values, endowed with a probability mass function, characteristic of a probability distribution;

CONTINUOUS:

Any Numerical Value in an interval or Collection of intervals, via a probability density function that is characteristic of a probability distribution;

MIXTURE OF BOTH TYPES:

The realizations of a random variable, that is, the results of randomly choosing values according to the variable's Probability distribution function, are called random variates.

2 DC ANOVA test of 3 different datasets which are subset of sepal Length

```
iris.1 ← iris[1:10, 1:1]
iris.2 ← iris[11:20, 1:1]
iris.3 ← iris[21:30, 1:1]
names(iris.1)
```

iris.1	iris.2	iris.3
5.1	5.4	5.4
4.9	4.8	5.1
4.7	4.8	4.6
4.6	4.3	5.1
5.0	5.8	4.8
5.4	5.7	5.0
4.6	5.4	5.0
5.0	5.1	5.2
4.4	5.7	5.2
4.9	5.1	4.7
4.86	5.21	5.01

Mean = $\bar{x}$

sd =

0.291357 0.4817791 ↓

iris.2

0.2469818

Total sum of squares

SST

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\left(x_{ij}-\bar{\bar{x}}\right)^2$$

$\bar{\bar{x}}$ = 5.1 + 4.9 + 4.7 + 4.6 + 5.0 + 5.4 + 4
+ 5.0 + 4.4 + 4.9 + 5.4

Stacking:

```
iris.new ← data.frame(cbind(iris.1,
                            iris.2, iris.3)

iris.stacked ← stack(iris.new)

names(iris.stacked)
[1] "values" "ind"

iris.anova ← aov(values ~ ind, data =
                 Y ~ X      iris.stac

sum((iris.stacked$values - sst)^2
```

SST = 4.018667

```
sst ← mean(iris.stacked$values)
sst = 5.026667
```

$\bar{\bar{x}}$ SST = 5.026667

mean (all is stacked values)

$= (x_{ij} - \bar{\bar{x}})^2$

SST = 4.018667

Find (SSTR):- Treatment sum of squares

$10*(4.86 - 5.026667)^2 + [10*(5.21 - 5.026667)^2$

$+ [10*(5.01 - 5.026667)]^2$

$10*(-0.166667)^2 + 10*(0.183333)^2 +$

$10*(-0.016667)^2$

$10*0.02777789 + 10*0.03361099 +$

0.0002777889

$= 0.0027777889 + 0.3361099 +$

$0.2777789 = 0.616666689$

SSTR = 0.617

Total sum of squares
(SST)

$$SS\left(x_{ij} - \bar{x}_j\right)^2$$

iris.1

5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0
4.4 4.9

m1 ← mean(iris.1)

m1

4.86

iris.1 - m1

0.24 0.04 -0.16 -0.26 0.14 0.54 -0
0.14 -0.46 0.04

0.0576 0.0016 0.0256 0.0676 0.019
0.2916 0.0676 0.0196 0.2116 0.00

(iris.1-m1)^2 =

sum((iris.1-m1)^2), = 0.764 | m3 = 5.0
sum((iris.2-m2)^2) = 2.089 | m2 = 5.2
sum((iris.3-m3)^2) = 0.549

0.764 + 2.089 + 0.549

SSE = 3.402

(iris.2-m2) =

0.19 -0.41 -0.41 -0.91 0.59 0.49

0.19 -0.11 0.49 -0.11

(iris.2-m2)^2

0.0361 0.1681 0.1681 0.8281

0.3481 0.2401 0.0361 0.0121

0.2401 0.0121

sum((iris.2-m2)^2)

= 2.089

(iris.3-m3)

0.39 0.09 -0.41 0.09 -0.21 -0.01

-0.01 0.19 0.19 -0.31

(iris.3-m3)^2

0.1521 0.0081 0.1681 0.0081

0.0441 0.0001 0.0001 0.0361

0.0361 0.0961

SST = SSTR + SSE

$4.018667 = 0.617 + 3.402$

4.019

round(4.018667, 3)

$= 4.019$

Total mean square = $\frac{4.019}{N-1} = \frac{4.019}{30-1} = \frac{4.019}{29}$

$= 0.1385862$

Mean square treatment (MSTR)

$\frac{SSTR}{c-1}$ = 'c' = no colum

$= \frac{0.617}{(3-1)} = \boxed{0.3085}$

$= \underline{0.3083}$

pval ← pf(3.354131, 2, 27)

pval

= 0.95

Mean square error (MSE)

$$\frac{SSE}{N-C} \quad \frac{3.402}{30-3} =$$

$$= 0.126$$

F distribution:-

$$\frac{MSTR}{MSE} = \frac{0.3085}{0.126}$$

$$= 2.448413 \approx 2.447.$$

critical f: value:-

$df1 = C-1$ 2

$df2 = N-C$ 27

cY

$F_{2,27} =$

For p-value:-

$= pt(t, df = n-1)$

$qf(.95, df = 2, df2 = 27)$

p: 3.354131

The 95% percentile of the F-distribution with (2, 27) degrees of freedom is 3.354131

2.447 - 3.354731 *

P-value look up:

Given an observed value of a test statistic tobs that is supposed (under the null hypothesis) be 'F' distributed with df1 numerator degrees of freedom and df2 denominator degrees of freedom the p-value for a lower-tailed test

a) a lower-tailed test:

is the lower tail area given by

Pf(tobs, df1, df2)

Pf(2.447, df1 = 2, df2 = 27)

= 0.8944789

1 - Pf = 0.1055211

= 0.105

c) an upper-tailed test is the upper tail area given by

1 - pf(tobs, df1, df2)

c) a two-tailed test is twice whichever one-tailed p-value is smaller

plow ← pf(tobs, df1, df2)

2 * min(plow, 1 - plow)

Example:- in DeGroot & Schervish, the (obviously made up) "observed" value of the test statistic they call it V) is 3.0. the numerator degrees of freedom is 5 and denominator degrees of freedom is 20.