“An Introduction to Statistical Software R”

by

Geethu Gopinath
Department of Statistics
Christ College(Autonomous), Irinjalakuda

Random Number Generation From Normal Distribution

- **rnorm()-** Generate a vector of random numbers which are normally distributed
- **Syntax:-**

  rnorm(n, mean, sd)

  n- Number of Observations

  mean- mean of the data set

  sd- Standard deviation of the data set

❑ Create a vector of 1000 random numbers with mean =90 and sd=5

R commands

```
x=rnorm(1000,90,5)    # Generate random numbers
hist(x, breaks=50)    # Create histogram with 50 bars
```

Resident Data Set

- Resident data sets are data sets that come with R.
- We use **data function** to access resident data sets.

  data(name of the data set)

- data(rivers) # Read data with data()

  rivers

- data() # To get list of all data sets in the base package
- data(package=packages(all.available=TRUE))

PLOTS TO CHECK NORMALITY IN R

BOX PLOT & Q-Q PLOT

- Measures how well the data is distributed in the data set.
- It divides the data set into three quartiles.
- Boxplot can also be used to check for symmetry or lack of it.

R commands

```
x=rnorm(100,1,1)
boxplot(x)
qqnorm(x)
qqline(x)
```

```
y=rexp(100,1)
boxplot(y)
qqnorm(y)
qqline(y)
```

Shapiro-Wilk Test

- Test to check Normality in R
- **Null Hypothesis-** Population is normally distributed.
- **Interpretation-** If the p value is less than the chosen alpha level, then the null hypothesis is rejected
- **Conclusion-** There is evidence that the data tested are not normally distributed.
- **R syntax:- shapiro.test(data)**

```
 data(rivers)
 rivers
shapiro.test(rivers)
```

Parametric tests

- It’s a statistical test which makes certain assumptions about the distribution of the population or about the parameters of the population

Example

- One sample t test
- Two sample t test
- Chi-Square test
- ANOVA

One sample t test

- One sample t tests compares the mean of a sample to a pre-specified value.
- **Null Hypothesis**: There is no difference in the sample mean and the hypothesized mean.
- ❑ We have the potato yield from 12 different farms. We know that the standard potato yield for the given variety is 20. Test if the potato yield from these farms is significantly better than the standard yield.

Syntax :-

```
t.test(y, mu = mean)
 y   - name of the variable
mu - mean specified by the null hypothesis
```

R commands

```
x=c(21.5,24.5,18.5,17.2,14.5,23.2,22.1,20.5,19.4,18.1,24.1,18.5)
 t.test(x,mu=20,alternative="greater")
```

Two sample t test

- Syntax:-

  t.test (A, B, mu= mean)

  A-name of 1st variable

  B- name of 2nd variable

  mu- specified mean

Paired t test

- Used to compare the means between two related groups of samples.
- **Null Hypothesis-** There is no significant difference between the observations in each pairs.
- 10 mice received a treatment X during 3 months. We want to know whether the treatment X has an impact on the weight of the mice.
- Syntax:-

```
 # Weight of the mice before treatment
    before=c(200.1, 190.9,192.7,213,241.4,196.9,172.2,185.5,205.2,193.7)
# Weight of the mice after treatment
    after=c(392.9,393.2,345.1,393,434,427.9,422,383.9,392.3,352.2)
  t.test(before,after,paired=T)
```

Chi-Square test

- Used to determine if two categorical variables have a significant correlation between them.
- Both the variables should be from same population.
- Categorical- Yes/No, Male/Female etc.
- **Null Hypothesis-** The two variables are independent

Alternative- The two variables relate to each other.

- Test the hypothesis whether the students smoking habit is independent of their exercise level at 5% significant level.
- **Syntax:-**

```
library(MASS)
tbl=table(survey$Smoke,survey$Exer)
tbl
chisq.test(tbl)
```

Interpretation:-As the p value 0.4828 is greater than the 0.05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

ANOVA

- Extension of independent two-sample t-test.
- **Null hypothesis-** The means of the different groups are the same.
- Test the effects of 3 types of fertilizer on crop yield.
- Syntax

```
one.way=aov(yield~fertilizer, data=crop.data)
summary(one.way)
```

Non-Parametric tests

- If there is no knowledge about the distribution of the population or parameters of the population.

  Example
- Sign test
- Wilcoxon's test
- Kruskal Wallis test

Different situations for parametric and non- parametric tests

Sl. No	Situation	Parametric tests	Non-parametric tests
1	Comparison between two independent populations	t-test	Wilcoxon’s rank sum tests
2	Comparison between two correlated populations	Paired t-test	Wilcoxon’s signed rank tests
3	Comparisons among several independent populations	One-way ANOVA	Kruskal Wallis tests
4	Comparisons among several correlated populations	Two-Way ANOVA	Freidman’s Two-way ANOVA

What we learn?

- Random number Generation
- Plots to check Normality
- Resident Data Set
- Parametric Tests
- Non Parametric Tests

THANK YOU