Workshop: Data Analytic using R

Radhendushka Srivastava

rsrivastava@iitb.ac.in

Department of Mathematics
Indian Institute of Technology Bombay

March 5, 2022

Statistical Data Analysis Framework

- Objective of an experimental/behavioral/physical phenomenon study.
- Planning of experiment and methods of collection of data (sampling survey, designing the experiment, observing the phenomenon)
- Given a data, what kind of statistical analysis is possible?
- Randomness in data?
- How to make meaningful inferences from a data?
- Probabilistic understanding of randomness in data.
- Statistical modelling of the random data under suitable set of conditions.
- How to begin?
- Exploratory data analysis!

Data analysis

- Data: Variable types
  - Discrete variable (Nominal, Ordinal, Numeric)
  - Continuous variable.
- Univariate analysis of each variable.
- Bivariate analysis of variables.
- Statistically modelling response variable as a function of dependent variables.
- Parameter estimation and validation.
- Model diagnostic: Validating assumptions of the statistical model, such as Normality, homogeneity, outlier detection.
- Model refinement using statistical significant parameters.
- Data refresh mechanism and maintenance.

An example: Heart Disease Data Set

The heart disease depends on several conditions of a patient.

- Consider the popular heart data set: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
- Choose "cleve.mod" data set.
- The data set contains 14 variables.
- An objective of the study is to predict whether a patient with certain conditions have heart disease or not.
- Several probabilistic/statistical models have been built for this purpose and more than 60 articles have cited this data. For more detail, visit the above mentioned website.

Heart disease data set: Attributes

Attribute	Description
Age	Age (Continuous)
Sex	Male, Female (Dis. Nominal)
CP	Chest pain (angina, abnang, notang, asympt) (Dis. Nominal
Trestbps	Resting blood pressure (Continuous)
chol	Cholesteral (Continuous)
fbs	Fasting blood sugar < 120 (true or false) (Dis. Nominal)
restecg	Resting ecg (norm, abn, hyper) (Dis. Ordinal)
thalach	Max heart rate achieved (Continuous)
exang	Exercise induced angina (true or false) (Dis. Nominal)
oldpeak	ST depression induced by exercise (Continuous)
slope	Slope of peak exercise ST (up, flat, down) (Dis. Ordinal)
ca	Number of vessels colored (Dis. Numeric)
thal	Thalassemia (norm, fixed, rever) (Dis. Nominal)
target	Healthy (buff) or with heart-disease (sick) (Dis. Nominal)

Discrete data

- A *random* variable is said to be discrete if it takes finitely (*countably*) many values.
- A discrete random variable may be one of the following type.
  - Nominal: Typically have finitely many categories (coding to numeric values does not express the magnitude).
  - Ordinal: Typically have finitely many options but the options have sense of order (coding to numeric values does not express the magnitude but certainly express the order).
  - Numeric: Have finitely (*countably*) many values and value shows the true magnitude.

Let X be a discrete random variable taking values $\{x_1, x_2, \ldots\}$. The Probability Mass Function (pmf) of X is

$$P(X = x_i) = p_i, \tag{1}$$

where $p_i \geq 0$ and $\sum_i p_i = 1$.

Estimation of probability mass function

Given a discrete random variable (X) in a data set, an estimator of probability mass function of it is

$$\text{Estimate of } P(X = x_i) = \hat{p}_i = \frac{\text{Number of times } x_i \text{ appears in the data}}{\text{Total number of the data points}},$$

- Given a data, the discrete random variable will have only finitely many values.
- Sum of the estimates will be 1.
- This estimator is also referred as frequency based pmf.

PMF tables, Visualization

Given a discrete variable in a detaset, we can estimate the probability mass function and present it in table format. We can also visualize the probability mass function of a discrete random variable as Bar plot, Pie chart.

We have several discrete variables in the heart disease data set.

We will now move to R studio to compute the frequency, pmf tables, bar and pie chart of some of the discrete variables.

Joint pmf of two discrete random variable

Joint pmf: Let X and Y be two discrete random variables taking values from $\{x_1, x_2, \ldots, \}$ and $y_1, y_2, \ldots$. The joint probability mass function of random variables X and Y is

$$P(X = x_i, Y = y_j) = p_{ij},$$

where $p_{ij} \geq 0$ and $\sum_i \sum_j p_{ij} = 1$.

Marginal pmf: Given the joint pmf, the marginal pmf of random variables are obtained as

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} = p_i^{(x)}$$

$$P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} = p_j^{(y)}$$

Independence: The random variables X and Y are said to be independent when

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \quad \text{for all } i, j.$$

Joint pmf of two discrete random variable..

A typical way of expressing the joint pmf of two random variables are

	y_1	y_2	$\cdots$	Total
x_1	p_{11}	p_{12}	$\cdots$	$p_1^{(x)}$
x_2	p_{21}	p_{22}	$\cdots$	$p_2^{(x)}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$
Total	$p_1^{(y)}$	$p_2^{(y)}$	$\cdots$	1

- When the response variable is discrete, one should explore the joint pmf structure on response and other variables.
- Given a dataset, the number of choices for the discrete variables are always finite.
- Marginal pmf of variable are obtained from joint pmf table by the total of row and columns.

Estimation of Joint pmf

Given a discrete variables X and Y in a data set, an estimator of joint pmf is

$$\hat{p}_{ij} = \frac{\text{Number of times } (x_i, y_j) \text{ appears together the data}}{\text{Total number of the data points}},$$

- Estimator of $\hat{p}_{ij}$ gives information about probability of happening the event (x_i, y_j) simultaneously.
- This estimator is also referred as frequency based probability.
- Several interesting questions can be answered during the data analysis based joint pmf estimates.

Heart data: Joint pmf estimates and visualization

- Target variable (healthy or sick) is influenced by several other variables, like gender, chest pain, etc.
- We estimate the bivariate joint probabilities.
- One can visualize the joint probabilities using the grouped bar plot (*Stack bar plot*).

We now move to R studio and explain it.

Testing of hypothesis

(Null hypothesis)H_0	:	An assertion about the population that experimenter wish to establish
(Alternative hypothesis)H_1	:	An assertion that is compliment of the null hypothesis

- How to make a statistical test?
- A test statistic that discriminates the null and alternative hypothesis.
- Construct rejection and acceptance region based on test statistic.
- Decision: Reject null hypothesis if test statistics falls in critical (rejection) region.
- Possibilities of error and its remedy.

Testing of hypothesis..

- Two types of error in decision making
  1. Type I error: Reject H_0 when H_0 is true.
  2. Type II error: Accept H_0 when H_1 is true.
- Simultaneous minimizing the probability of both the errors is not possible.
- Type I error is more serious.
- Fix the probability of type I error. Known as level of significance. (*Need to know the distribution of test statistic under* H_0)
- Minimize the probability of type II error. (*Need to know the distribution of test statistic under* H_1)
- **p value:** Observed level of significance. Probability of rejecting H_0 under null distribution for given test statistic.

χ^2 test of association

Let X and Y be discrete random variables with two choices and the joint pmf is

	y_1	y_2	Total
x_1	p_{11}	p_{12}	$p_1^{(x)}$
x_2	p_{21}	p_{22}	$p_2^{(x)}$
Total	$p_1^{(y)}$	$p_2^{(y)}$	1

(Null hypothesis) H_0 : Variables X and Y are independent

(Alternative hypothesis) H_1 : Variables X and Y are associated

From the given data set, we have the following frequency table

	y_1	y_2	Total
x_1	O_{11}	O_{12}	$O_1^{(x)}$
x_2	O_{21}	O_{22}	$O_2^{(x)}$
Total	$O_1^{(y)}$	$O_2^{(y)}$	N

χ^2 test of association...

Compute the χ^2 test statistics as

$$\chi^2_{test} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{2}$$

where

$$E_{ij} = N \hat{p}_i^{(x)} \hat{p}_j^{(y)}.$$

E_{ij} is the expected frequency under the null hypothesis.

- The probability distribution of χ^2_{test} statistics under H_0 is same as distribution of χ^2 random variable with degree of freedom (number of rows -1) (number of columns -1).
- Decision: Reject H_0 if the p-value$=$ $P[\chi^2_{(2-1)(2-1)} > \chi^2_{test}] < 0.05$.
- Level of significance (probability of type I error is set as 0.05).

Heart Data: χ^2 test of association

- The response variable (target) which is discrete variable with categories sick and healthy.
- An interesting task is to verify whether target variable and other discrete variables like chest pain, gender, ca, etc are independent or associated.
- We now move to r studio to apply the chi square association test for these discrete variables.

Testing equality of proportions

- Let p_1 and p_2 be the proportions of an attribute in two populations.
- Let two samples of size n_1 and n_2 with that specific attribute are given from the two populations, respectively.

We are interested in testing the following hypothesis:

$$\begin{aligned} (\text{Null hypothesis}) H_0 &: p_1 = p_2 \\ (\text{Alternative hypothesis}) H_1 &: p1 \neq p_2 \end{aligned}$$

A test statistic is given by

$$Z_{stat} = \frac{\hat{p}_1 - \hat{p}_2}{s.e.(\hat{p}_1 - \hat{p}_2)},$$

Testing equality of proportions..

Here

$$\hat{p}_1 = \frac{\text{Number of times the attribute appears in the first sample}}{n_1}$$

$$\hat{p}_2 = \frac{\text{Number of times the attribute appears in the second sample}}{n_2},$$

$$\hat{p} = \frac{\text{Number of times the attribute appears in the both the samples}}{n_1 + n_2}$$

$$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

- **Decision:** Reject the null hypothesis if the p-value$=2P(Z > |Z_{stat}|) < 0.05$, where Z is a standard normal random variable.
- Chosen level of significance: Probability of type I error is 0.05.
- We now move to R studio to apply this test on heart data.

Continuous Random Variable

- A random variable X is said to be a continuous random variable if $P(X = x) = 0$ for all real x.
- The probability density function f of continuous random variable is used to compute the probability as follows.

$$P(a < X < b) = \int_a^b f(x)dx.$$

- Examples:
  - A standard normal random variable Z has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

We compute $P(Z > a) = \int_a^\infty f(x)dx$.

Continuous Random Variable..

- Examples contd:
  - A χ^2 random variable with k degree of freedom has the probability density function

$$f(x) = \begin{cases} \frac{1}{2^{k/2}} \frac{1}{\Gamma(\frac{k}{2})} x^{k/2-1} e^{-x/2}, & 0 < x < \infty. \\ 0 & \text{otherwise} \end{cases}$$

    We compute $P(\chi^2_k > a) = \int_a^\infty f(x)dx$.
- Note that, even for testing related to discrete variables, the null distribution of test statistic happens to be continuous.
- Given a data, how to decide whether the variable is continuous or numeric discrete?

Basic Statistics for continuous variable

1. Measure of central tendency
   - Mean, Median, Mode
2. Measure of dispersion
   - Variance, Standard deviation, Range, Inter quartile range
3. Skewness.
4. Kurtosis.

We now move to R studio to compute the basic statistics for continuous variable on Heart data.

Testing equality of means

- Let μ_1 and μ_2 be the mean of a continuous variable in two populations.
- Let two samples of size n_1 and n_2 with that continuous variables are given from the two populations, respectively.

We are interested in testing the following hypothesis:

$$\begin{aligned} \text{(Null hypothesis)} H_0 &: \mu_1 = \mu_2 \\ \text{(Alternative hypothesis)} H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

A test statistic is given by

$$t_{stat} = \frac{\bar{x}_1 - \bar{x}_2}{s.e.(\bar{x}_1 - \bar{x}_2)},$$

Testing equality of means..

Here

$$\bar{x}_1 = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}, \qquad \bar{x}_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_1},$$

$$sd_1 = \sqrt{\frac{\sum_{j=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{n_1 - 1}}, \qquad sd_2 = \sqrt{\frac{\sum_{j=1}^{n_1}(x_{2i} - \bar{x}_2)^2}{n_2 - 1}},$$

$$s.e.(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

- **Decision:** Reject the null hypothesis if the p-value=$2P(t > |t_{stat}|) < 0.05$, where t is Student-t random variable with $(n_1 + n_2 - 2)$ degree of freedom.
- Chosen level of significance: Probability of type I error is 0.05.
- We now move to R studio to apply this test on heart data.

Another example: Bike sharing data

- Bike sharing systems are new generation of traditional bike rentals.
- Riding bike plays important role in traffic, environmental and health issues.
- Data source: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
- Bike Sharing data has 16 variables.
- We choose a subset of 5 variables for simpler analysis of the data.
- The total count of rented bikes depends upon several weather condition.

Bike sharing data variables

Attribute	Description
weathersit	1 (Clear), 2 (Mist and cloudy). (Dis. Nominal)
temp	Normalized temperature in Celsius. (Continuous)
hum	Normalized humidity. (Continuous)
windspeed	Normalized wind speed. (Continuous)
cnt	count of total rental bikes. (Continuous)

- Move to R studio for computation of basic statistics of the continuous variables.

Histogram

Estimation of pdf using Histogram:

- Construct bins: group the range of variable into mutually exclusive classes.
- Guidelines for number of bins. Number of bins $= K$
- Count the number of observations lie in the bin, say $m_1, m_2, \ldots, m_K$.
- The bar plot of m vs bin is called frequency histogram.
- The bar plot of $\frac{m_i \times length(bin_i)}{n}$ vs bin_i is called density histogram.
- It can be shown that density histogram is a consistent estimator of pdf of the continuous random variable.
- We now construct the density histogram of "Count" variable in R.

Correlation and Scatter plot

- Let $(x_i, y_i)_{1 \leq i \leq n}$ be samples of two continuous random variable.
- **Correlation** between the two variables is computed as

$$Cor(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Correlation measures the linear relationship.
- $-1 \leq Cor(x, y) \leq 1$.
- Negative correlation: Increase in one variable amounts to decrease in other.
- Positive correlation: Both variables increases (decreases) together.
- Scatter plot visualizes the pattern between two variables.

Linear regression model between two variables

Let $(x_i, y_i)_{1 \le i \le n}$ be samples of two continuous random variable.

Regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

here β_0 and β_1 are the parameters and ϵ_i's are i.i.d. Normal random variables with mean 0 and variance σ^2.

- How to estimate the parameters β_0, β_1 and σ^2?
- Minimize squared error loss function (least square function) to get the estimate.

$$\text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Linear regression model..

- The estimator turns out to be

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\end{aligned}$$

- The estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The testing of significance of the variable is tested using $t_s tats$.
- Measure of goodness of fit: R^2 is a one such measure of goodness of fit.
- Regression model for two variables: $R^2 = (Cor(x, y))^2$.

Diagnostic of the model

- Compute the residual vector

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n.$$

- The noise sequence ϵ_i's are not observed.
- If the model is correct, the residuals e_i's should be a good proxy for ϵ_i.
- Verify that e_i's have constant variance. Plot(e_i^2) and it should be constant without any pattern.
- Verify that e_i's are normally distributed. Plot the histogram of the residual to check this.
- Residuals analysis can be performed using other forms of residuals like standardized residual, studentized residual, etc.

Multiple Linear Regression Model

Multiple Linear Regression Model: Given data points $(y_i, x_{1i}, x_{2i}, \ldots, x_{pi})$, for $i = 1, 2, \ldots, n$, multiple linear regression model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \tag{3}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are real unknown parameters and ϵ_i's are i.i.d. Normal random variables.
Multiple linear regression model can be in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = X\beta + \epsilon, \tag{4}$$

Assumptions

- Design matrix $\mathbf{X}$ is a full column matrix (i.e. $n > p$)
- Noise sequence $\{\epsilon_i\}_{i \geq 1}$ are independent and identically distributed (i.i.d.) Normal random variables with mean 0 and variance σ^2
- How to estimate parameters β_i's and σ^2.
- Minimize Quadratic loss function/ Least Square function!

Least Square Estimator

Consider the quadratic loss function/ least square function

$$Q(\beta_0, , \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

We aim to minimize the least square function to estimate the parameters β_i's.

In matrix form, the least square function can be viewed as

$$Q(\beta) = (y - X\beta)'(y - X\beta) \tag{5}$$

Least square estimator turns out to be

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{6}$$

and an estimator of σ^2 is given by

$$\hat{\sigma^2} = \frac{1}{n - p - 1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2. \tag{7}$$

Measure of Accuracy

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2 = (n-p-1)\hat{\sigma}^2$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- Lower the RSS, better the goodness of fit.
- R^2 closer to 1 shows that data exhibits linear pattern.
- **Caution**: R^2 small only indicates non-existence of linear pattern. However, there can be a non-linear pattern.

Model Diagnostics

- Compute the residual vector

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}, \quad i = 1, 2, \ldots, n.$$

- The noise sequence ϵ_i's are not observed.
- If the model is correct, the residuals e_i's should be a good proxy for ϵ_i.
- Verify that e_i's have constant variance. Plot(e_i^2) and it should be constant without any pattern.
- Verify that e_i's are normally distributed. Plot the histogram of the residual to check this.
- Residuals analysis can be performed using other forms of residuals like standardized residual, studentized residual, etc.