Institutions and Innovation

Tutorial 01 - Statistics Review

Prof. Dr. Cornelia Storz
Fei Wang (Michael) ❤️ AI
Goehte University Frankfurt
Summer Semester 2023

Roadmap of this tutorial

1. Introduction to `data.table`
2. Univariate Statistics
3. Bivariate Statistics
4. Multivariate Statistics
5. Regression Analysis
6. Summary

1. Introduction to `data.table`

1.1. What is `data.table`?

- `data.table` is a package in R that provides an enhanced version of data.frame. It is widely used for fast aggregation of large datasets, low latency add/update/remove of columns, quicker ordered joins, and a fast file reader. `data.table` is an extension of `data.frame` package in R.
- check benchmark: https://h2oai.github.io/db-benchmark/
    - 100 GB data
    - 155 seconds
    - out of memory for `Pandas`

In-class Lab 1.1 🦋

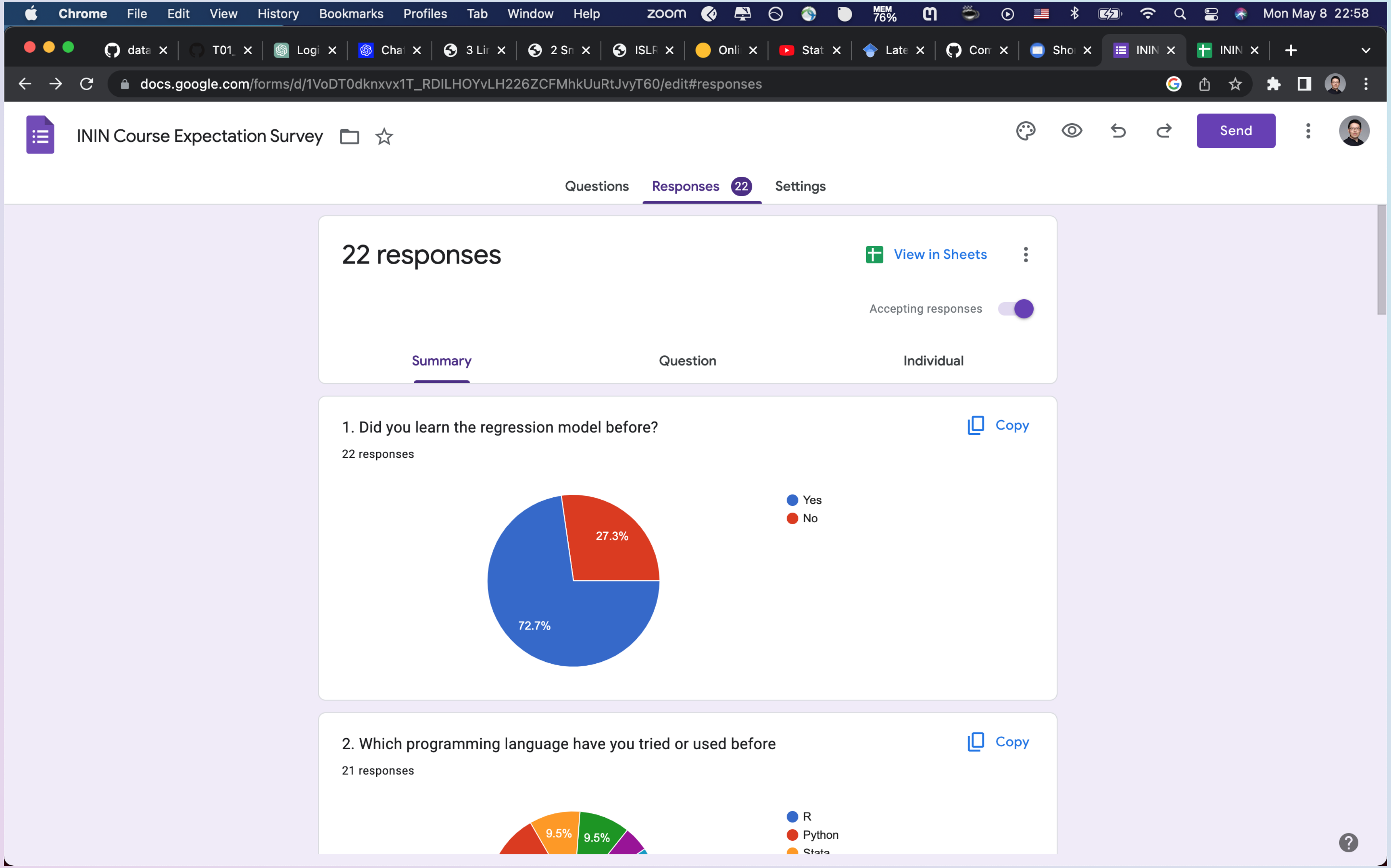
docs.google.com/spreadsheets/d/1hvAoS5NCoKELXAReV7W3-OwldTdFkpoJmAi0smNl6ds/edit?resourcekey#gid=1774566062

ININ Course Expectation Survey (Responses)

	A	B	C	D	E	F	G	H	I	
1	Timestamp	1. Did you learn the regre	2. Which programming la	3. Have you taken any st	4. What do you expect to	5. How many hours do yo	6. What is your goal of ta	7. What is your priority ra	8. Do you want to build u	9. D
2	4/19/2023 11:44:37	No	Python	Yes	Hello world	2-3 hours per week	Get a good grade by min	4	Maybe	Yes
3	4/19/2023 11:47:39	Yes	R	Yes	Na	1 hour per week	Get a good grade by min	5	Maybe	Yes
4	4/19/2023 11:59:12	Yes	R	Yes	Establishing more knowle	2-3 hours per week	Pass the exam and get th	6	Yes	Yes
5	4/19/2023 12:01:02	No	none	No	Pass the exam and get th	1 hour per week	Pass the exam and get th	5	Maybe	No
6	4/19/2023 12:01:51	Yes	R	Yes	Refresh my Knowledge	2-3 hours per week	Learn some knowledge a	3	No	Yes
7	4/19/2023 12:02:56	No	none	No	Good grade	1 hour per week	Get a good grade by min	2	No	Yes
8	4/19/2023 12:03:06	Yes	R, SPSS	Yes	refreshing statistics know	1 hour per week	Learn some knowledge a	6	Maybe	Yes
9	4/19/2023 12:13:58	Yes	Python	Yes	Na	1 hour per week	Get a good grade by min	7	No	Yes
10	4/19/2023 12:57:34	Yes	R	No	Na	2-3 hours per week	Learn some knowledge a	6	No	Yes
11	4/19/2023 12:59:58	Yes	Stata	Yes	Preparation for the Exam	1 hour per week, 2-3 hou	Learn some knowledge a	6	Yes	Yes
12	4/19/2023 13:03:13	Yes	None	Yes	Pass the exam	2-3 hours per week	Pass the exam and get th	6	Maybe	Yes
13	4/19/2023 13:22:19	Yes	Stata	Yes	refresh of knowledge	1 hour per week	Learn some knowledge a	6	Maybe	Yes
14	4/19/2023 13:22:19	Yes	Stata	Yes	refresh of knowledge	1 hour per week	Learn some knowledge a	6	Maybe	Yes
15	4/19/2023 13:43:26	Yes	R	Yes		1 hour per week	Pass the exam and get th	5	Maybe	Yes
16	4/19/2023 14:10:43	Yes	Python	Yes		2-3 hours per week	Learn some knowledge a	3	Yes	Yes
17	4/19/2023 14:43:24	No	R	No	More Knowledge about s	1 hour per week	Learn some knowledge a	4	Maybe	Yes
18	4/19/2023 15:03:35	Yes	R	Yes	Better intuitive statistical	2-3 hours per week	Learn some knowledge a	4	No	Yes
19	4/19/2023 15:35:45	No	Python	Yes	Be prepared for the exan	1 hour per week	Get a good grade by min	3	No	Yes
20	4/19/2023 16:03:15	Yes	R	Yes	Deeper insight into the pr	2-3 hours per week	Learn some knowledge a	7	Maybe	No
21	4/19/2023 17:11:21	Yes	gretl, JMP	Yes	Na	1 hour per week	Pass the exam and get th	4	Maybe	Yes
22	4/19/2023 20:57:15	Yes	R	Yes	Being perfectly prepared	1 hour per week	Get a good grade by min	4	Maybe	Yes
23	4/20/2023 23:44:06	No	Na	No	NA	1 hour per week	Pass the exam and get th	9	Maybe	Yes
24	4/20/2023 23:44:06	No	Na	No	Na	1 hour per week	Pass the exam and get th	9	Maybe	Yes
25	4/21/2023 11:21:16	Yes	R	Yes	Learn how to use STATA	2-3 hours per week	Learn some knowledge a	3	Maybe	Yes
26										
27										

Form Responses 1 | Sheet1

In-class Lab 1.1 🦋



In-class Lab 1 🦋

```r
# library
library(data.table)

# read the dataset from url
# url: https://shorturl.at/eixVX
csv_url <- "https://shorturl.at/eixVX"
survey <- fread(csv_url)

# check the data
str(survey)
head(survey)
summary(survey)
```

2. Univariate Statistics

2.1. What is Univariate Statistics?

- Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.
- Methods:
  - Discrete data: frequency table, bar chart, pie chart
  - Continuous data: histogram, box plot, summary statistics

2.2. Discrete Data

For discrete data, we can use

- frequency table
- bar chart
- pie chart to visualize the data.

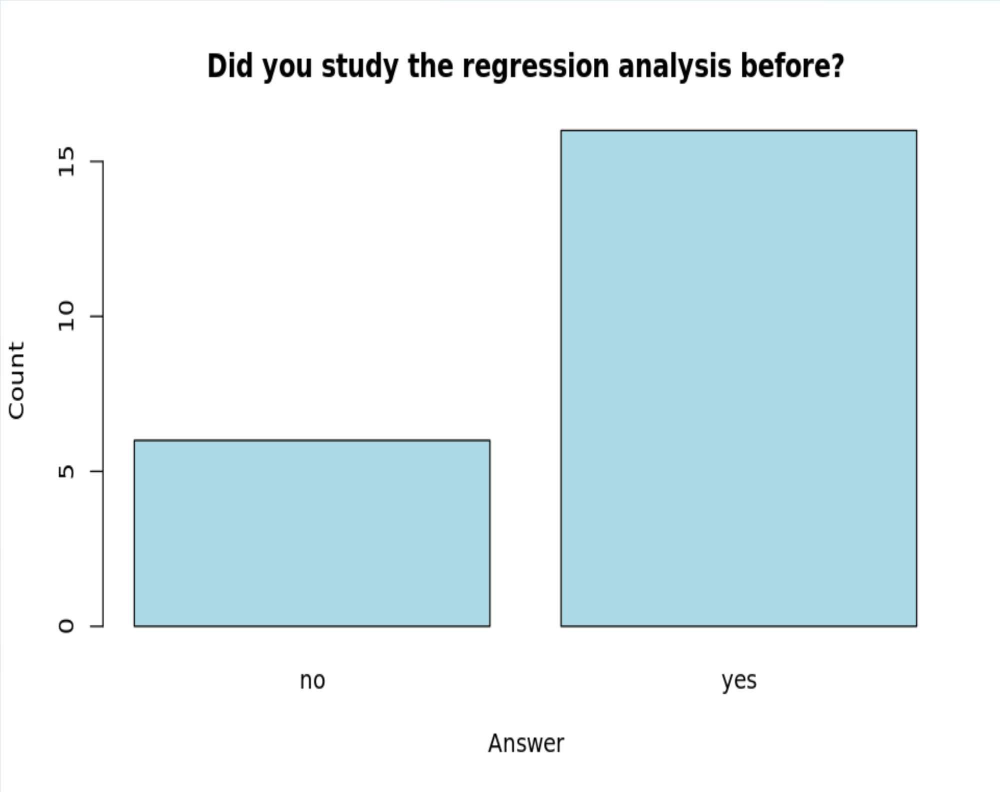
q1	N
no	6
yes	16

2.2.1. Bar plot

```
# use basic R function to get the frequency table
survey %>%
    with(table(q1)) %>%
    kable()

# using prop.table function to get the percentage
survey %>%
    with(table(q1)) %>%
    prop.table() %>%
    kable()

options(repr.plot.width = 8, repr.plot.height = 5)
survey %>%
    with(table(q1)) %>%
    barplot(main = "Did you study the regression analysis before?",
            xlab = "Answer",
            ylab = "Count",
            col = "lightblue")
```



2.2.2. Binomial Distribution

- Binomial distribution is a discrete probability distribution that expresses the probability of one set of two outcomes, as a function of the number of trials.
- In our survey, 70% of the students have studied the regression analysis before. We can use binomial distribution to calculate the probability of the number of students who have studied the regression analysis before.
- One class has 100 students. What is the probability that 30 of them have studied the regression analysis before?

2.2.2. Binomial Distribution

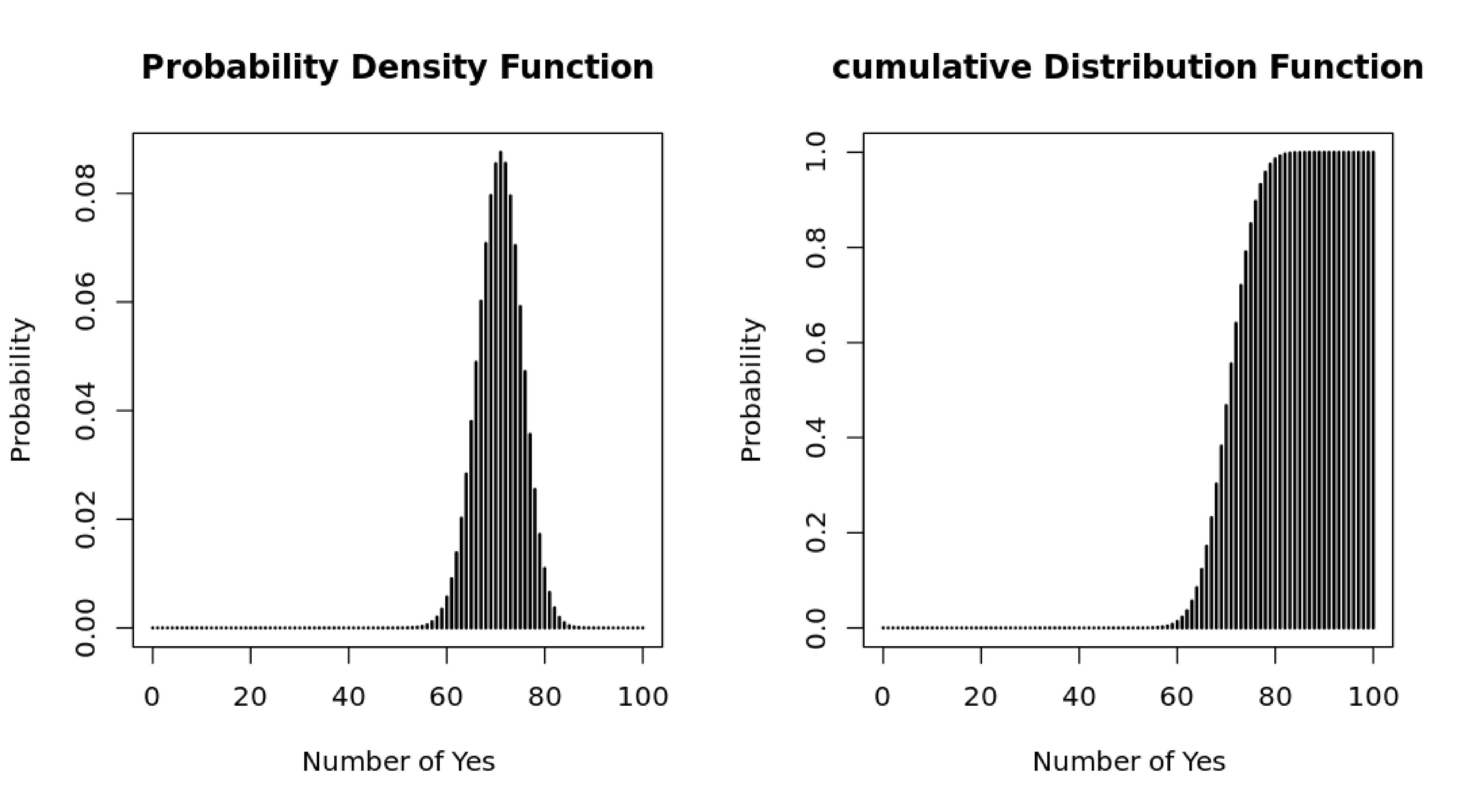
- `dbinom(x, size, prob)` is the function to calculate the probability of x successes in size trials with the probability of success prob.
- `pbinom(x, size, prob)` is the function to calculate the cumulative probability of x successes in size trials with the probability of success prob.

```
# probability of 30 students have studied the regression analysis before
dbinom(30, 100, 0.7)  # discrete probability
pbinom(30, 100, 0.7)  # cumulative probability
```

- The formula of binomial distribution is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

2.2.2. Binomial Distribution (discrete probability)



2.2.2. Binomial Distribution (discrete probability)

- Properties of binomial distribution:
  - The mean of binomial distribution is np.
  - The variance of binomial distribution is $np(1-p)$.
  - The standard deviation of binomial distribution is $\sqrt{np(1-p)}$.
- For instance, the mean of the number of students who have studied the regression analysis before is $100 \times 0.7 = 70$.

2.2.2. From Binomial Distribution to Poisson Distribution

> Now, suppose you are a restaurant owner. You want to know how many customers will come to your restaurant from 10am to 11am. How could you estimate the number of customers?

2.2.2. From Binomial Distribution to Poisson Distribution

> Now, suppose you are a restaurant owner. You want to know how many customers will come to your restaurant from 10am to 11am. How could you estimate the number of customers?

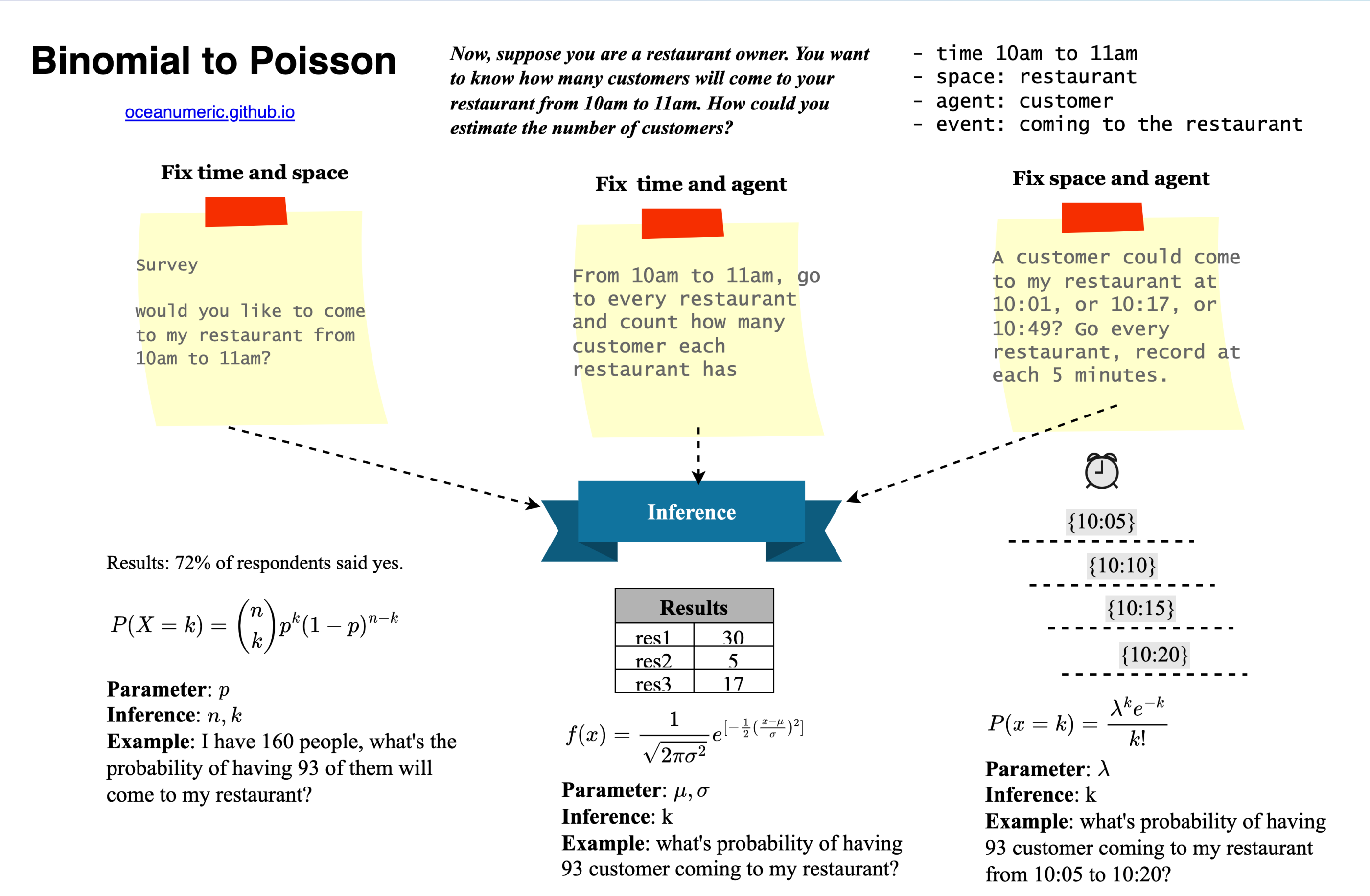
1. Do a survey?
    - fixed time and space
2. Use the competitor's data?
    - different restaurant
    - fixed time and people

2.2.2. From Binomial Distribution to Poisson Distribution

> Now, suppose you are a restaurant owner. You want to know how many customers will come to your restaurant from 10am to 11am. How could you estimate the number of customers?

- Our univariate dimension:
  - time: 10am to 11am
  - space: the restaurant
  - people: customers
  - event: customers come to the restaurant

2.2.2. From Binomial Distribution to Poisson Distribution



2.2.2. From Binomial Distribution to Poisson Distribution

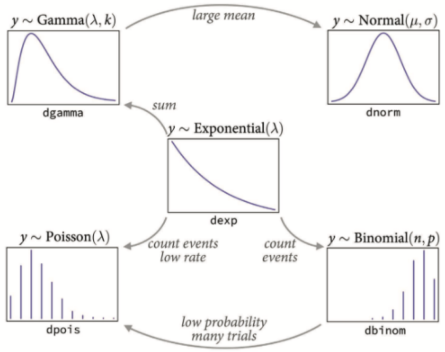
Binomial to Poisson

oceanumeric.github.io

Parameter: μ, σ
Inference: k
Example: what's probability of having 93 customer coming to my restaurant?

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2]}$$

Binomial	discrete	`dbinom(30, 100, 0.72)`
Poisson	discrete	`dpois(x,lambda)`
Gaussian	continuous	`dnorm(u,sigma)`



$$\begin{aligned} P(X=x) &= \lim_{n\to\infty} \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty} \frac{n(n-1)(n-x+1)}{x!} \frac{\lambda^x}{n^x} \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty} \frac{n(n-1)(n-x+1)}{n^x} \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty} \frac{n(n-1)(n-x+1)}{n^x} \frac{\lambda^x}{x!} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-x} \\ &= \frac{e^{-\lambda}\lambda^x}{x!} \end{aligned}$$

$$P(x=k) = \frac{\lambda^k e^{-k}}{k!}$$

Parameter: λ
Inference: k
Example: what's probability of having 93 customer coming to my restaurant from 10:05 to 10:20?

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Parameter: p
Inference: n, k
Example: I have 160 people, what's the probability of having 93 of them will come to my restaurant?

Notes or R-code

2.2.3. Poisson Distribution

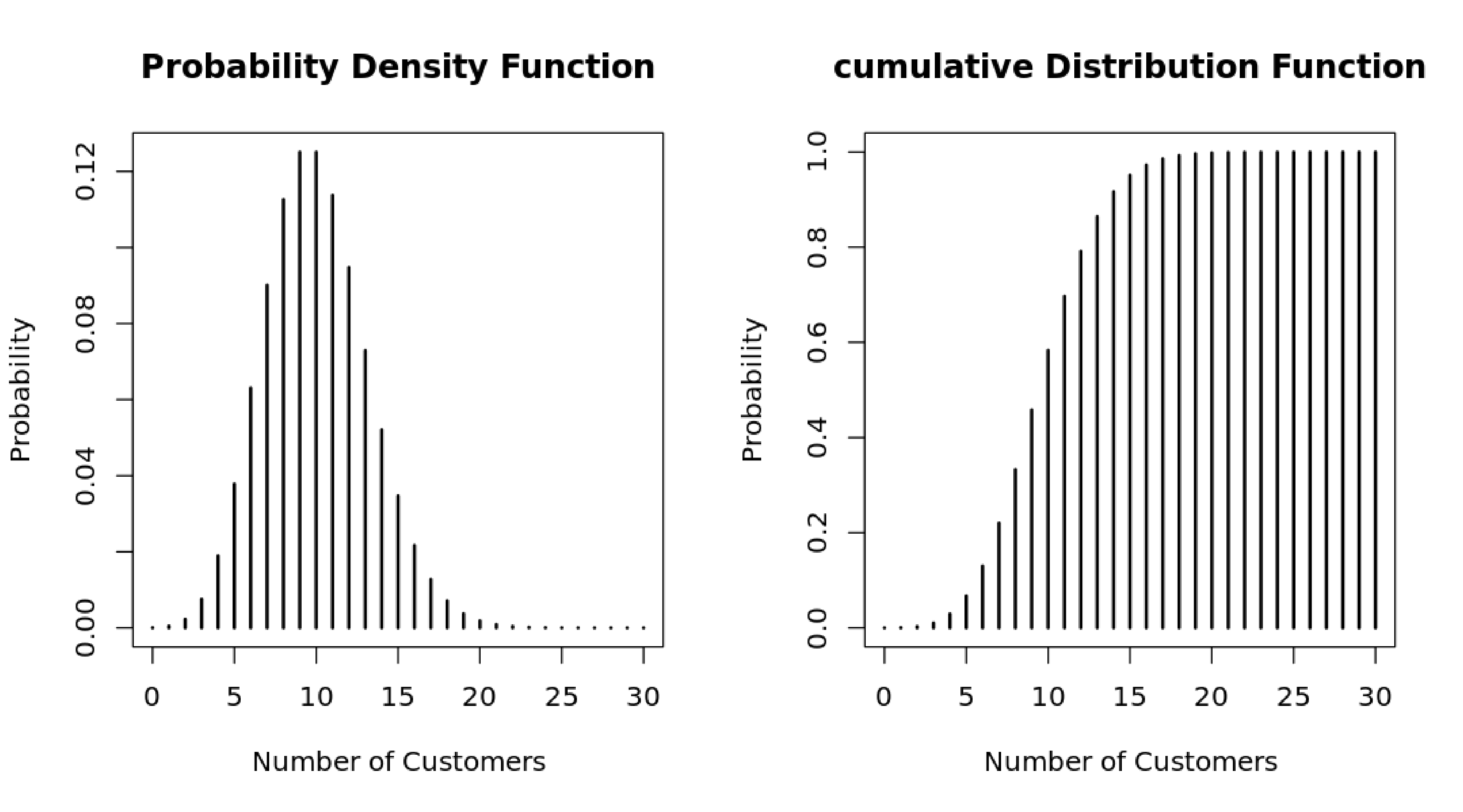
- Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.
- For example,
  - the number of customers who come to the restaurant from 10am to 11am
  - the number of students who come to the class from 10am to 11am
  - the number of cars that pass through a crossroads between 10am and 11am

2.2.3. Poisson Distribution

- Parameters of Poisson distribution:
  - λ is the mean number of events in an interval
- Inference:
  - k is the number of events in an interval
  - e is the Euler's number ($e = 2.71828\ldots$)

$$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

2.2.3. Poisson Distribution

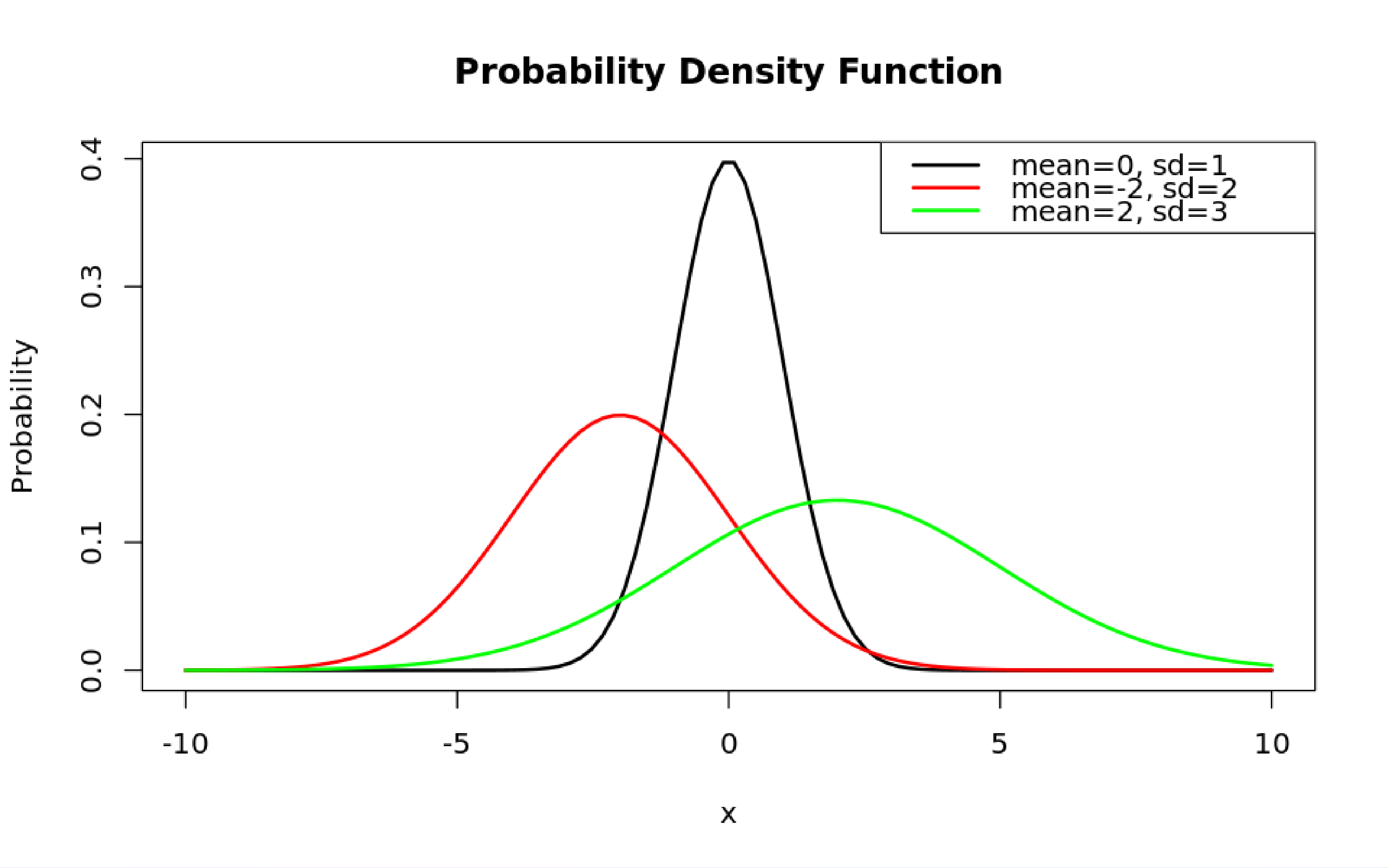


2.2.4. Normal Distribution

- Normal distribution is also called Gaussian distribution. It is a continuous probability distribution that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

2.2.4. Normal Distribution



3. Bivariate Statistics

3.1. What is Bivariate Statistics?

- Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. It is basically the analysis of two variables simultaneously, in order to determine the empirical relationship between them.
    - cross tabulation (contingency table) for discrete data
    - correlation for continuous data
    - covariance for continuous data

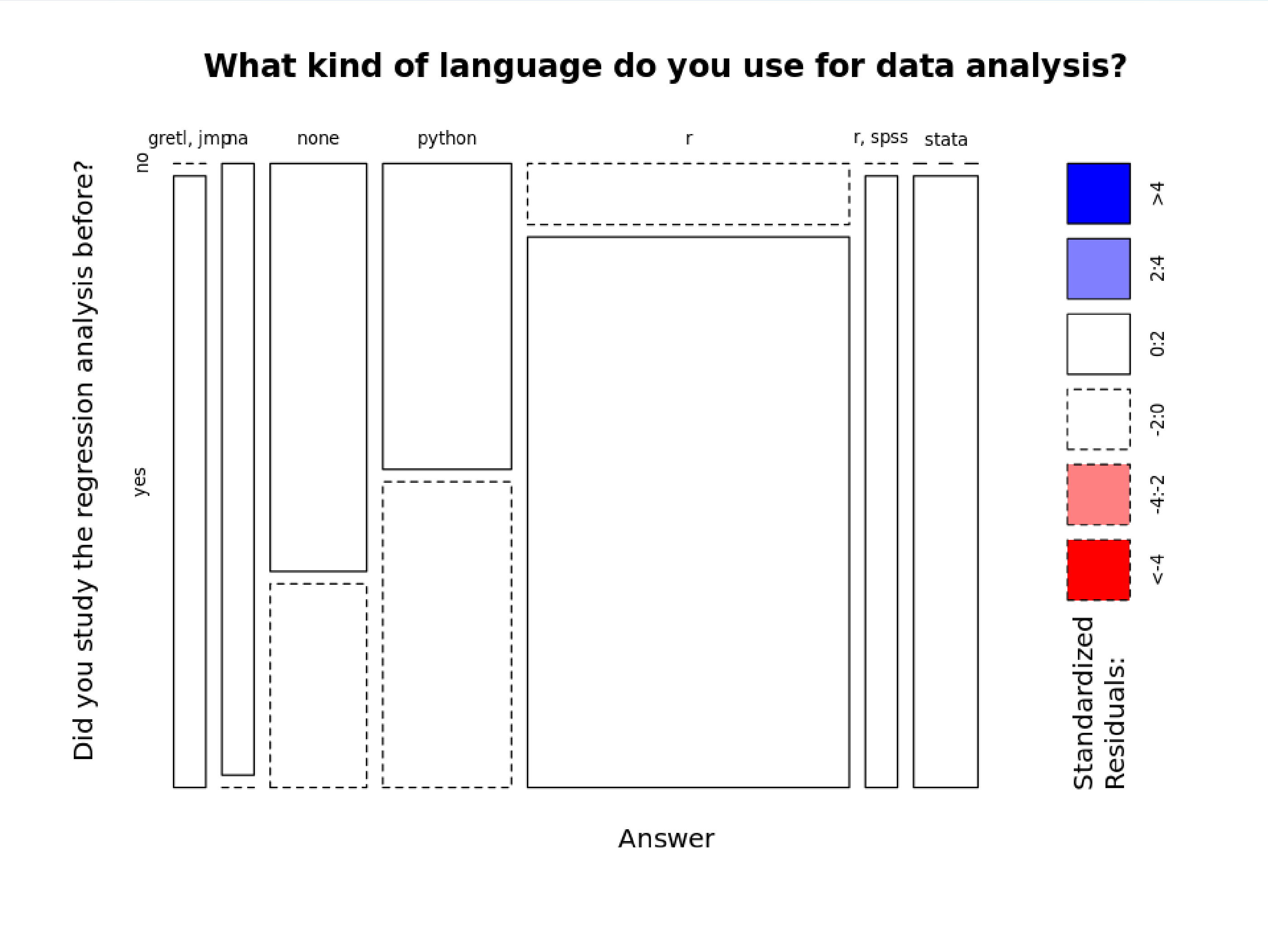
3.2. Cross Tabulation

- Cross tabulation is a tool that allows you compare the relationship between two variables. It is also known as contingency table. It is a table showing the distribution of one variable in rows and another in columns, used to study the correlation between the two variables.

	gretl, jmp	na	none	python	r	r, spss	stata
no	0	1	2	2	1	0	0
yes	1	0	1	2	9	1	2

3.2. Cross Tabulation

- We could use mosaic plot to visualize the cross tabulation.



4. Multivariate Statistics

4.1. What is Multivariate Statistics?

- Multivariate analysis is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. Typically, it involves a combination of two or more variables.
- Methods:
  - multiple regression
  - factor analysis
  - cluster analysis
  - Neural Networks

5. Regression Analysis

5.1. What is Regression Analysis?

- Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for **forecasting**, time series modelling and finding **the causal effect relationship** between the variables.
    - we will focus on forecasting in this tutorial
    - and causal effect relationship in the next tutorial

5.2 Linear Regression

- There many resources on the internet about linear regression. Here, we will focus on the implementation of linear regression in R.
  - Recommendation:
    - An Introduction to Statistical Learning
    - Regression in R
- Basic idea: dependent variable is a linear function of independent variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon; \quad \epsilon \sim N(0, \sigma^2)$$

- The simplest neural network is linear regression.

In-class Lab 2 🦋

In-class Lab 2 🦋 Boston Housing Dataset

```
===============================================
                      Dependent variable:
                  ---------------------------
                             medv
-----------------------------------------------
lstat                      -0.950***
                           (0.039)

Constant                   34.554***
                           (0.563)

-----------------------------------------------
Observations                  506
R2                           0.544
Adjusted R2                  0.543
Residual Std. Error     6.216 (df = 504)
F Statistic         601.618*** (df = 1; 504)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

6. Summary