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

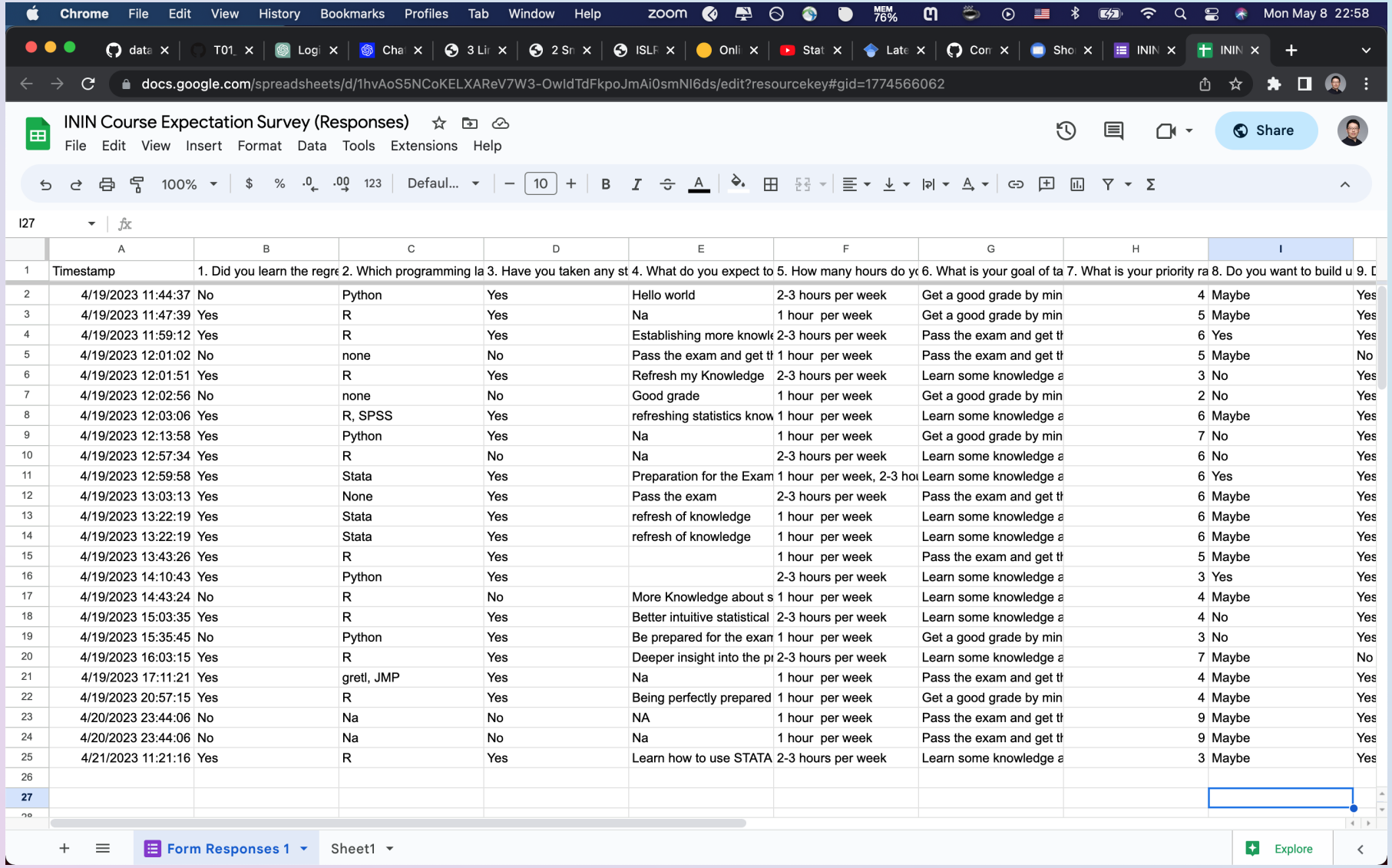
1. Introduction to `data.table`
2. Univariate Statistics
3. Bivariate Statistics
4. Multivariate Statistics
5. Regression Analysis
6. Summary

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`

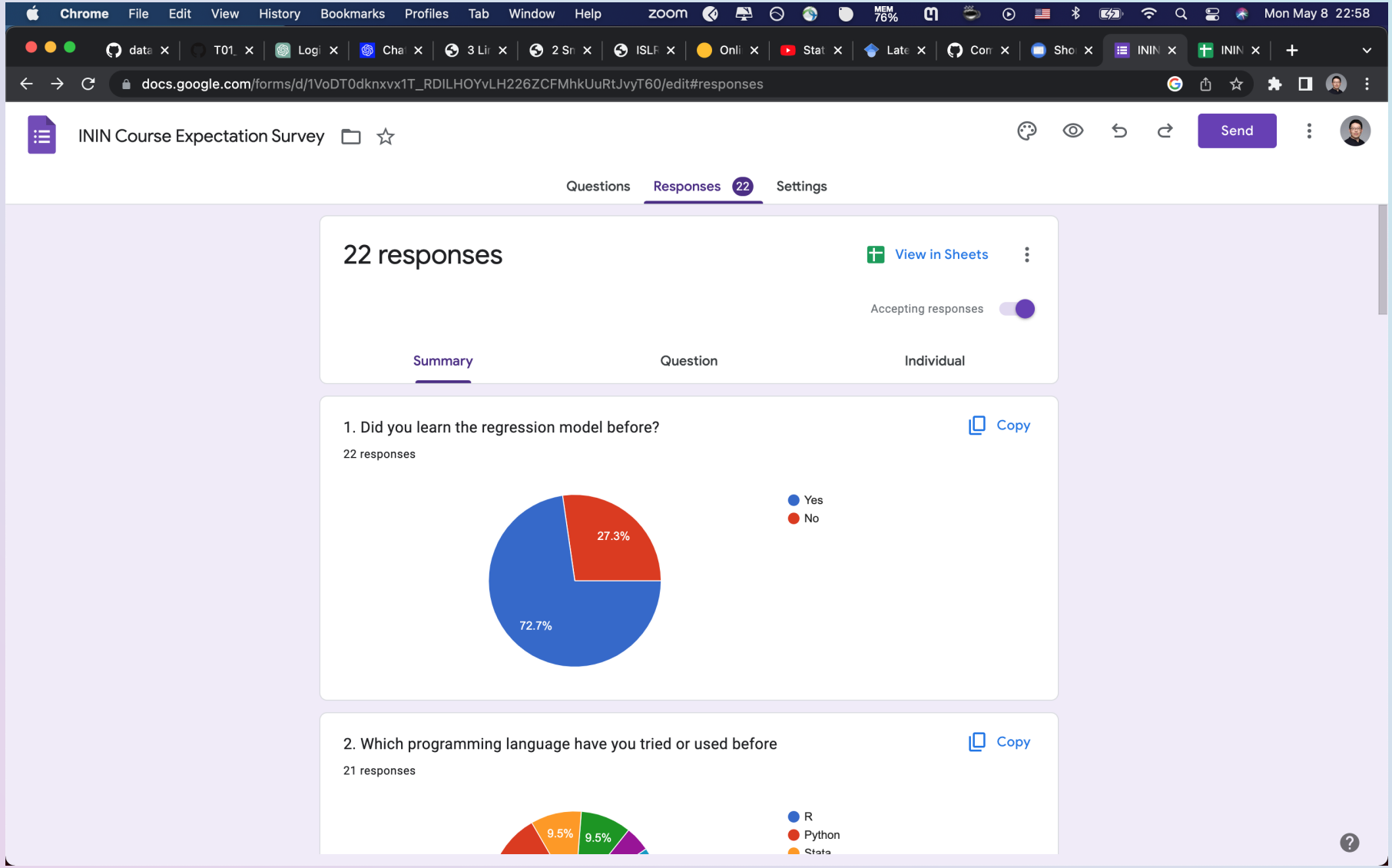
In-class Lab 1.1



ININ Course Expectation Survey (Responses)

	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 exam	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										

In-class Lab 1.1



In-class Lab 1.1 🦋

```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.

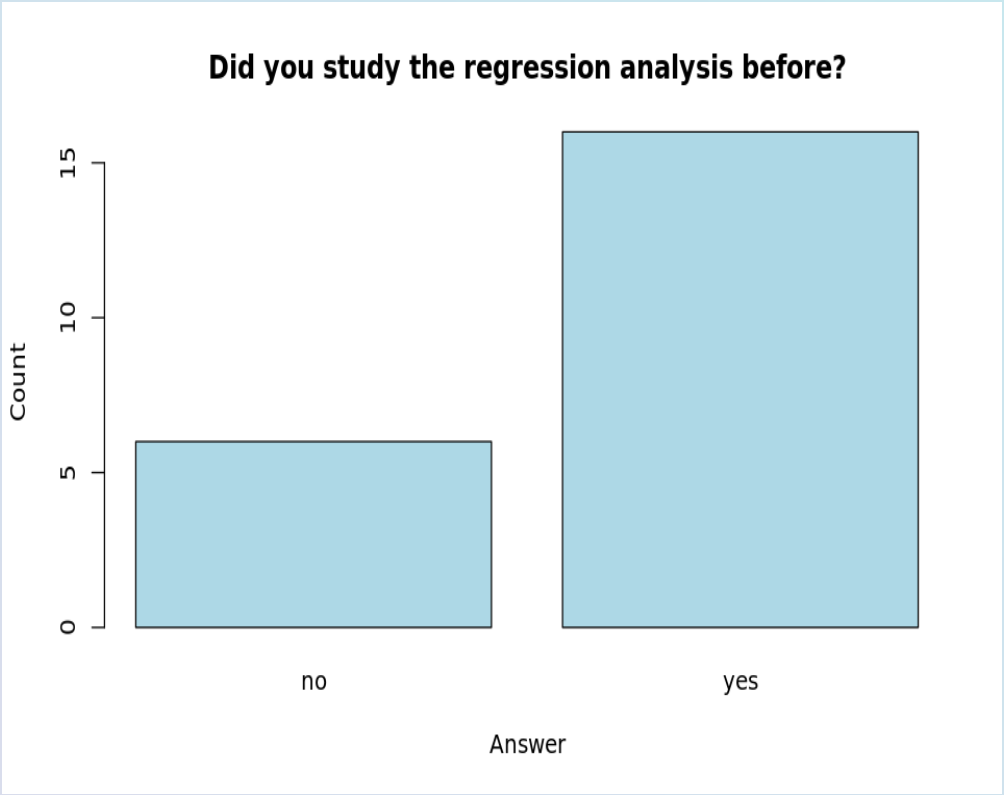
q1	N
no	6
yes	16

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

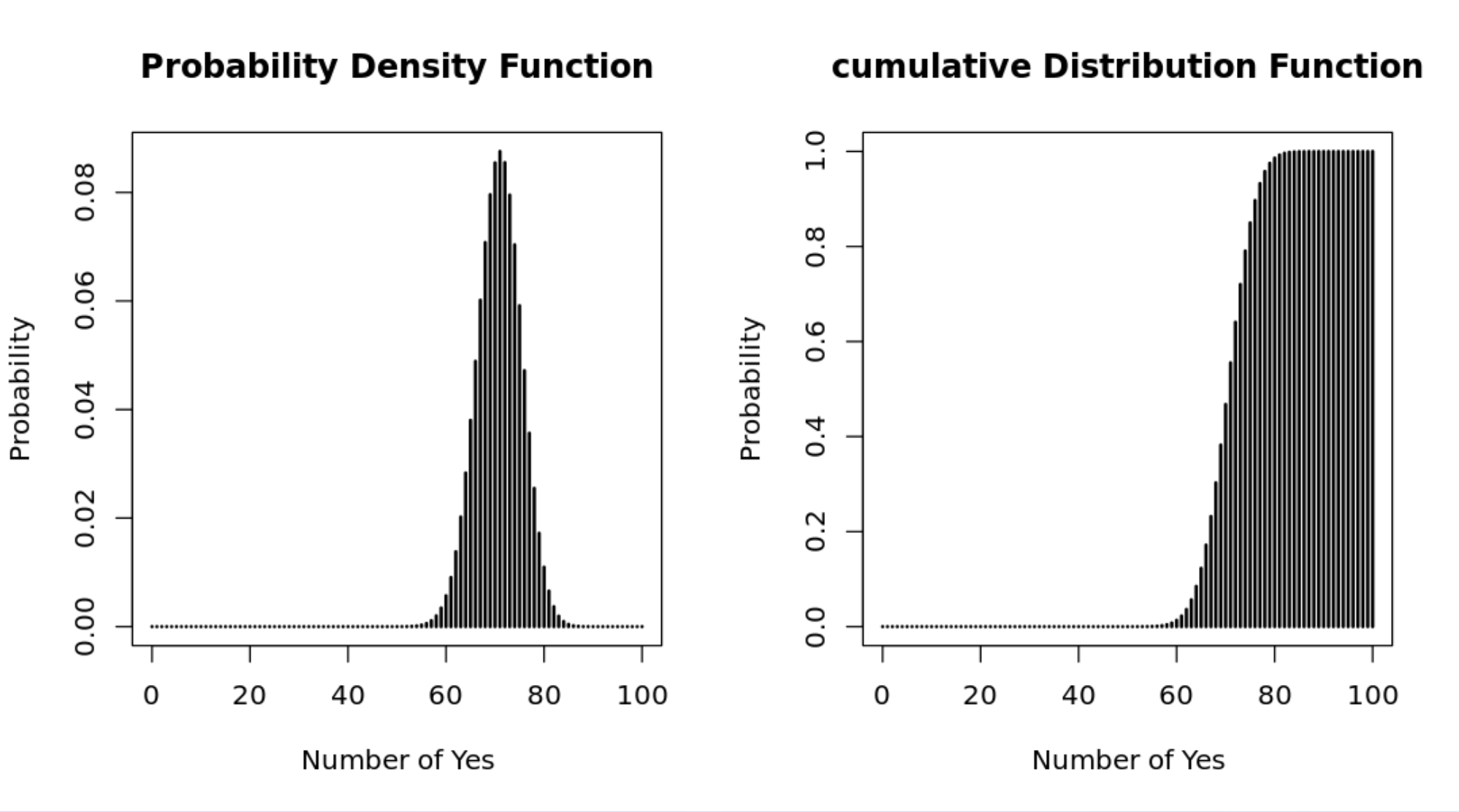
- `dbinom(x, size, prob)` is the function to calculate the probability of x successes in size trials with the probability of success prob.
- `pbinom(x, size, prob)` is the function to calculate the cumulative probability of x successes in size trials with the probability of success prob.

```
# 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 $100 \times 0.7 = 70$.

3. Bivariate Statistics

4. Multivariate Statistics

5. Regression Analysis

6. Summary