用R進行資料分析

Data Analysis with R

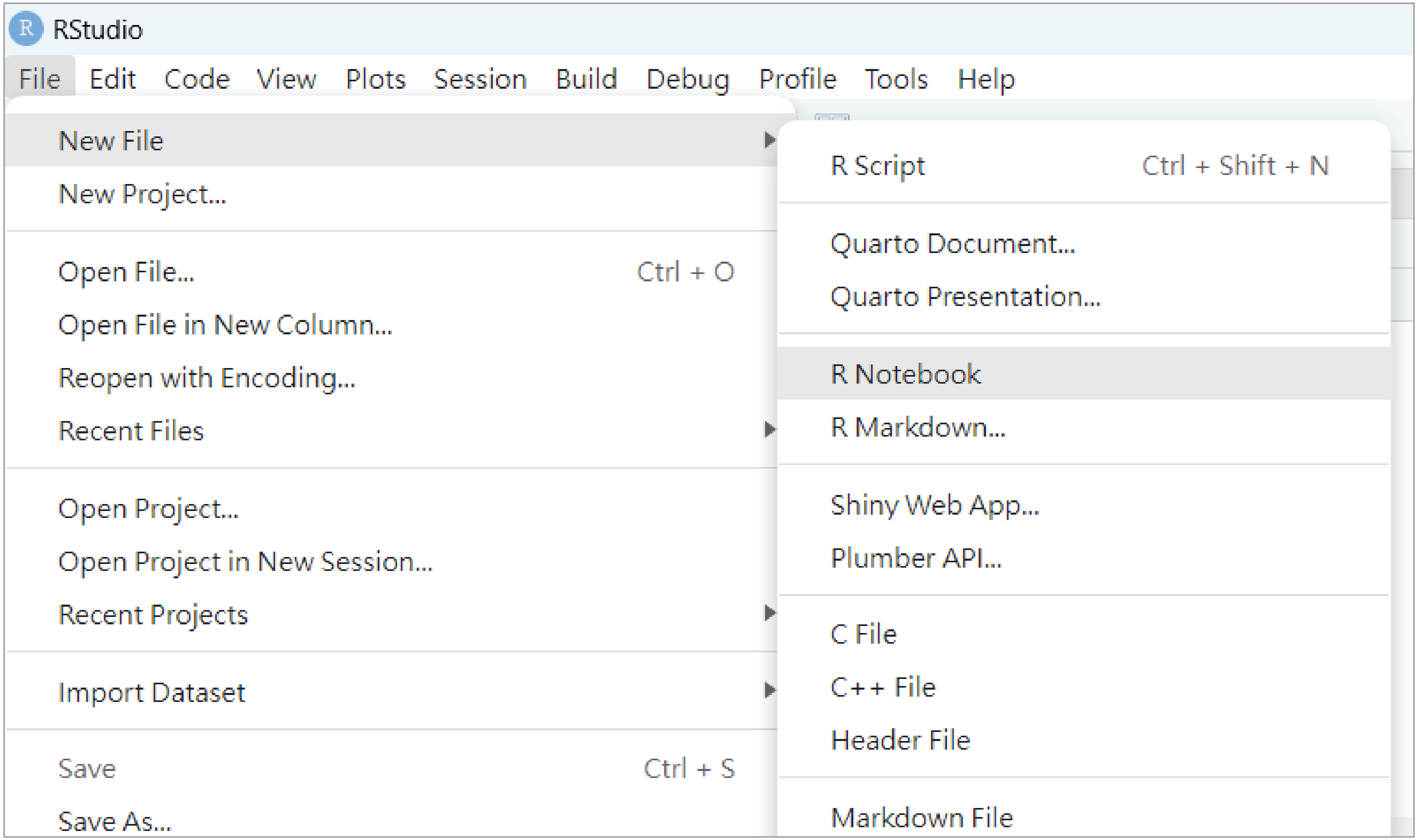
授課教師：溫在弘

E-mail: wenthung@ntu.edu.tw

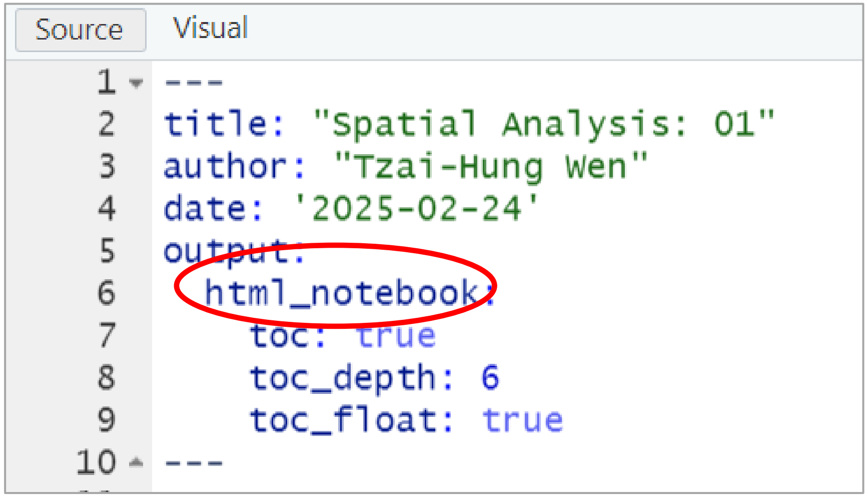
本週大綱

- 假設檢定
- 機率分布
- 統計圖表

Creating R Notebook files



R Notebook files



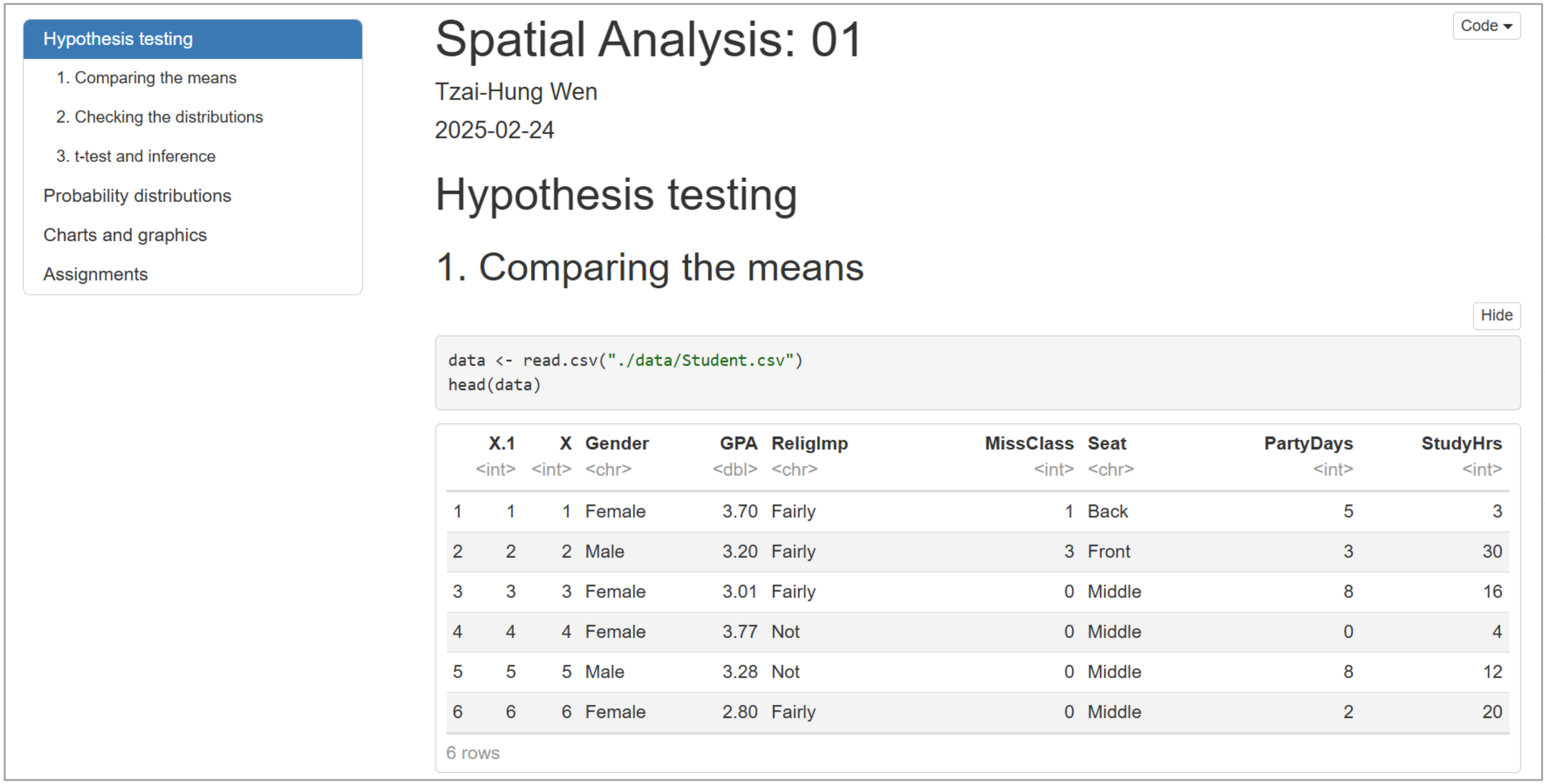
```
---
title: "Spatial Analysis: 01"
author: "Tzai-Hung Wen"
date: '2025-02-24'
output:
  html_notebook:
    toc: true
    toc_depth: 6
    toc_float: true
---
```

Lab1.nb.html

Lab1.Rmd

An R Notebook is an R Markdown document with chunks that can be executed *independently* and *interactively*, with output visible *immediately* beneath the input.

1. 假設檢定



Spatial Analysis: 01

Tzai-Hung Wen

2025-02-24

Hypothesis testing

1. Comparing the means

```
data <- read.csv("./data/Student.csv")
head(data)
```

	X.1	X	Gender	GPA	ReligImp	MissClass	Seat	PartyDays	StudyHrs
	<int>	<int>	<chr>	<dbl>	<chr>	<int>	<chr>	<int>	<int>
1	1	1	Female	3.70	Fairly	1	Back	5	3
2	2	2	Male	3.20	Fairly	3	Front	3	30
3	3	3	Female	3.01	Fairly	0	Middle	8	16
4	4	4	Female	3.77	Not	0	Middle	0	4
5	5	5	Male	3.28	Not	0	Middle	8	12
6	6	6	Female	2.80	Fairly	0	Middle	2	20

6 rows

text

code

output

Hypothesis Tests for Means

1. Determine the null and alternative hypotheses.
2. Verify necessary data conditions, and if met, summarize the data into an appropriate test statistic.
3. Assuming the null hypothesis is true, find the *p*-value.
4. Decide whether or not the result is statistically significant based on the *p*-value.
5. Report the conclusion in the context of the situation.

Key Concepts of Hypothesis Testing

- t-distribution
- Sampling distribution
- Standard error
- Null / alternative hypothesis
- One-tailed / two-tailed test
- Significance level (e.g. alpha = 5%)
- Confidence interval (e.g. 95% C.I.)
- p-value
- Type I and Type II errors

Basic Statistics Review

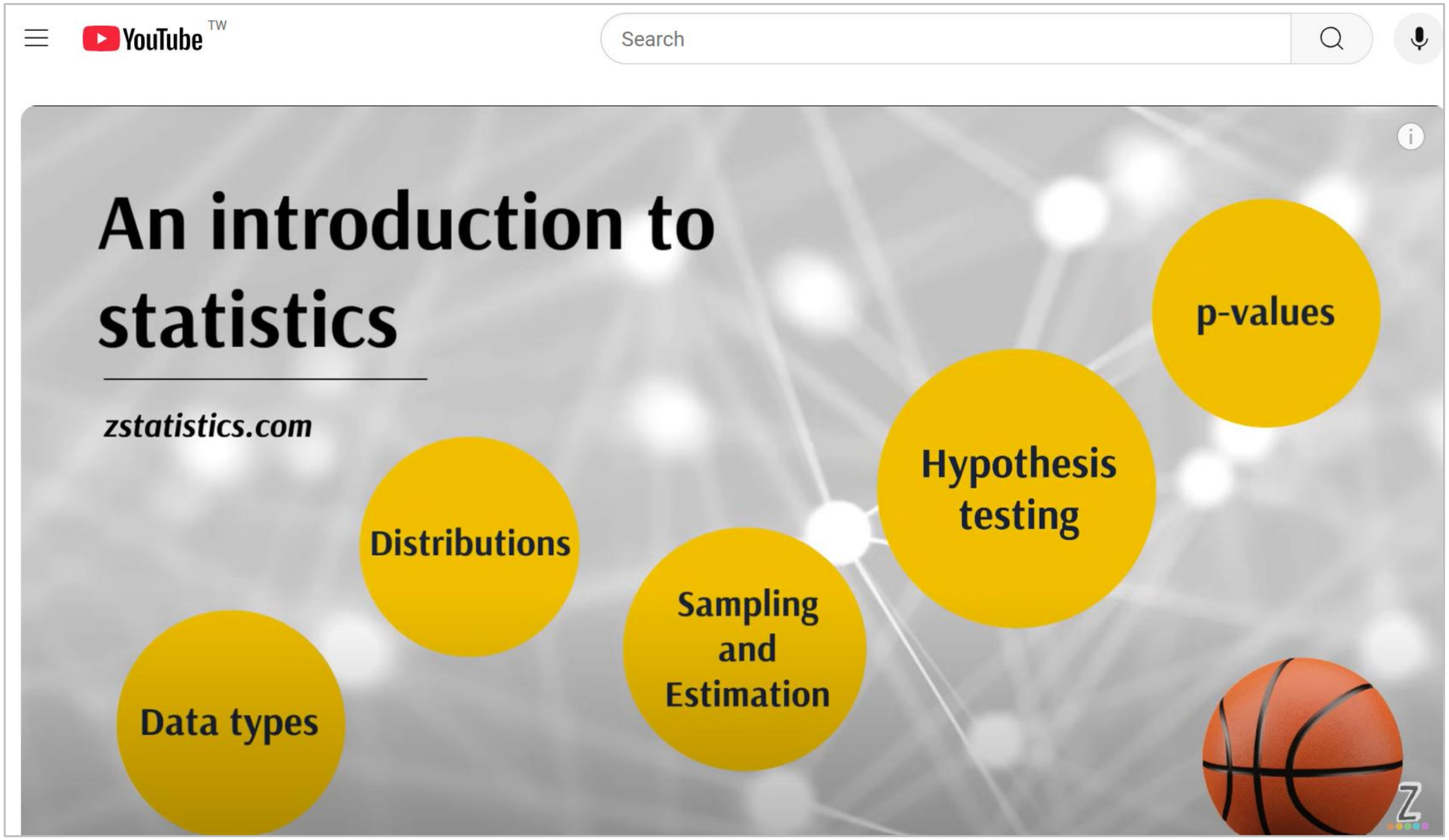
https://online.stat.psu.edu/statprogram/reviews/statistical-concepts

Statistics Online

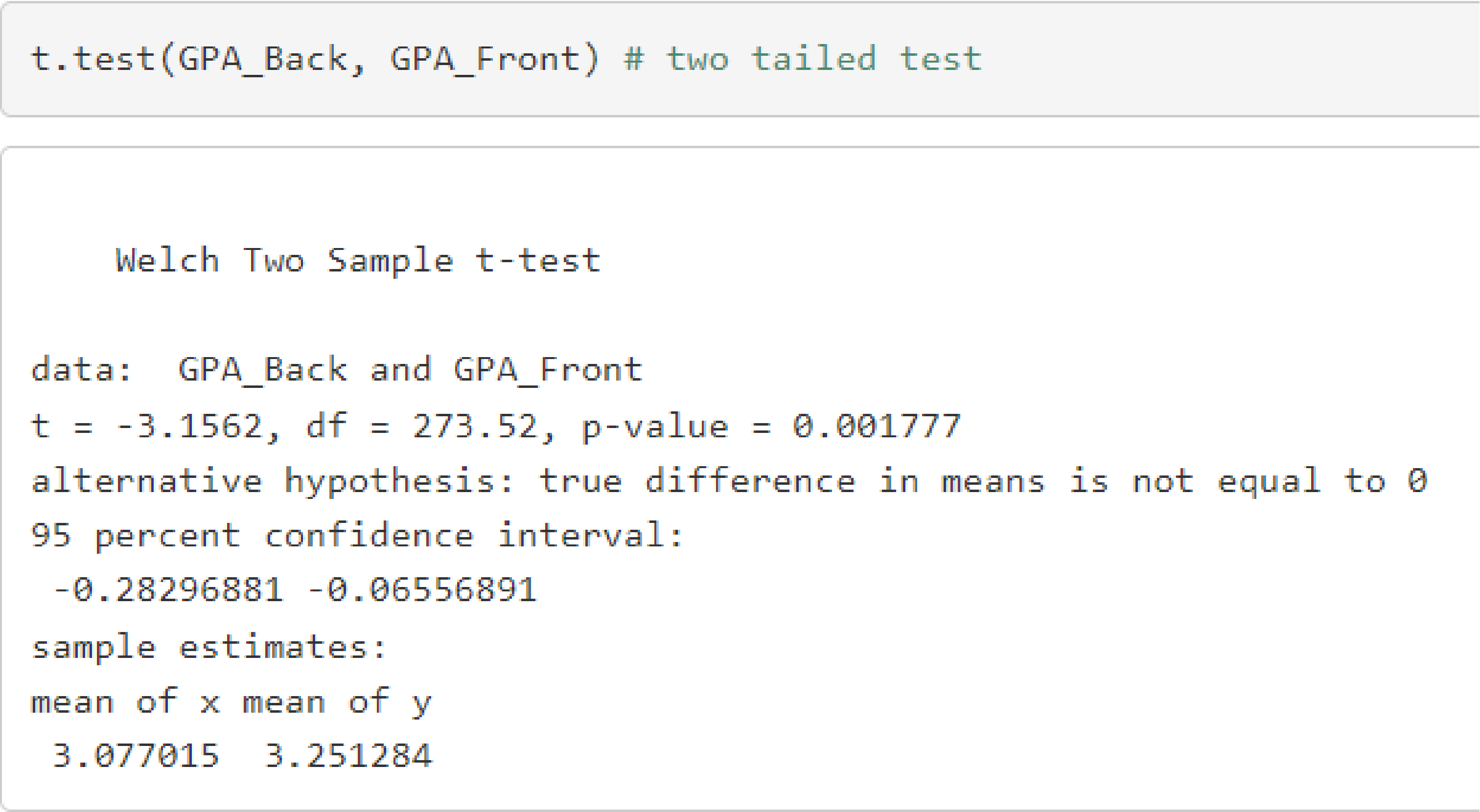
Basic Statistical Concepts

Basic Statistics Review

https://www.youtube.com/watch?v=kyjlxsLW1Is
(42:08)



Two-sample t-test



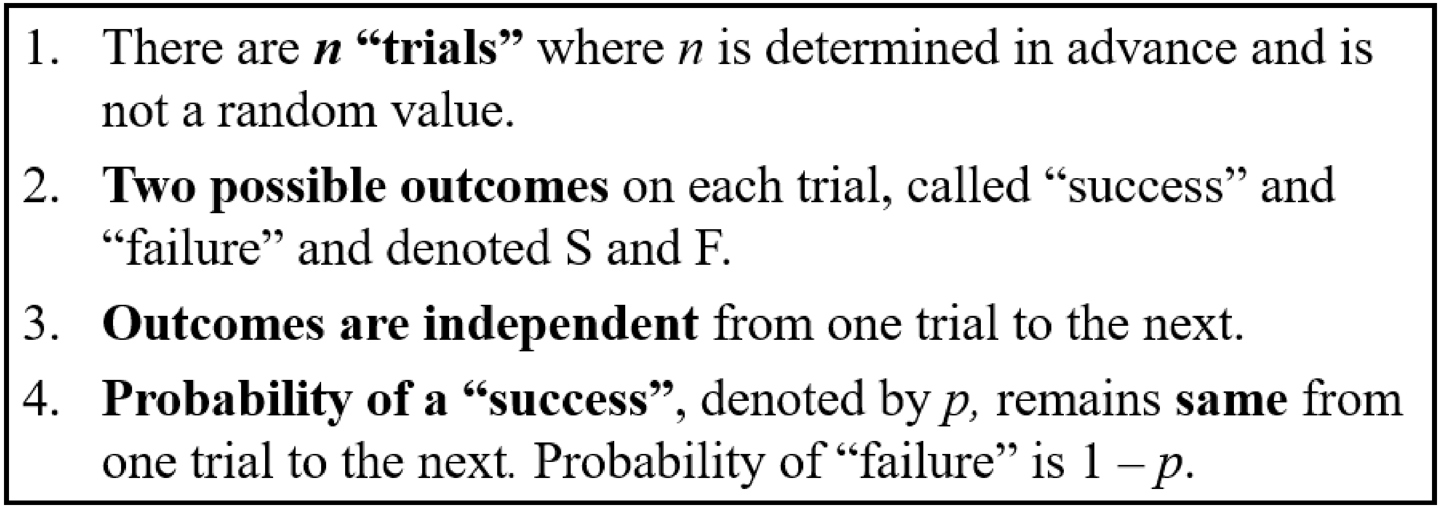
```r
t.test(GPA_Back, GPA_Front) # two tailed test
```

```

    Welch Two Sample t-test

data:  GPA_Back and GPA_Front
t = -3.1562, df = 273.52, p-value = 0.001777
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.28296881 -0.06556891
sample estimates:
mean of x mean of y 
 3.077015  3.251284 
```

2. 機率分布

Binomial Random Variables

Conditions for a binomial experiment:



1. There are ***n*** **“trials”** where *n* is determined in advance and is not a random value.
2. **Two possible outcomes** on each trial, called “success” and “failure” and denoted S and F.
3. **Outcomes are independent** from one trial to the next.
4. **Probability of a “success”**, denoted by *p,* remains **same** from one trial to the next. Probability of “failure” is 1 – *p*.

A **binomial random variable** is defined as *X*=number of successes in the *n* trials of a binomial experiment.

Binomial Probability Distribution

Finding Binomial Probabilities

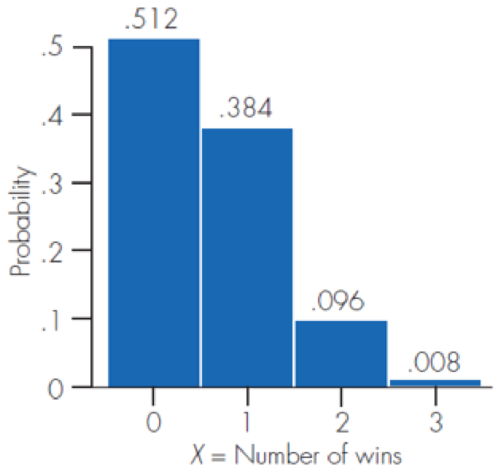
$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

Probability of Two Wins in Three Plays

p = probability win = 0.2; plays of game are *independent*.

***X* = number of wins in three plays.**

What is ***P*(*X* = 2)**?



$$P(X = 2) = \frac{3!}{2!(3-2)!} .2^2 (1-.2)^{3-2}$$

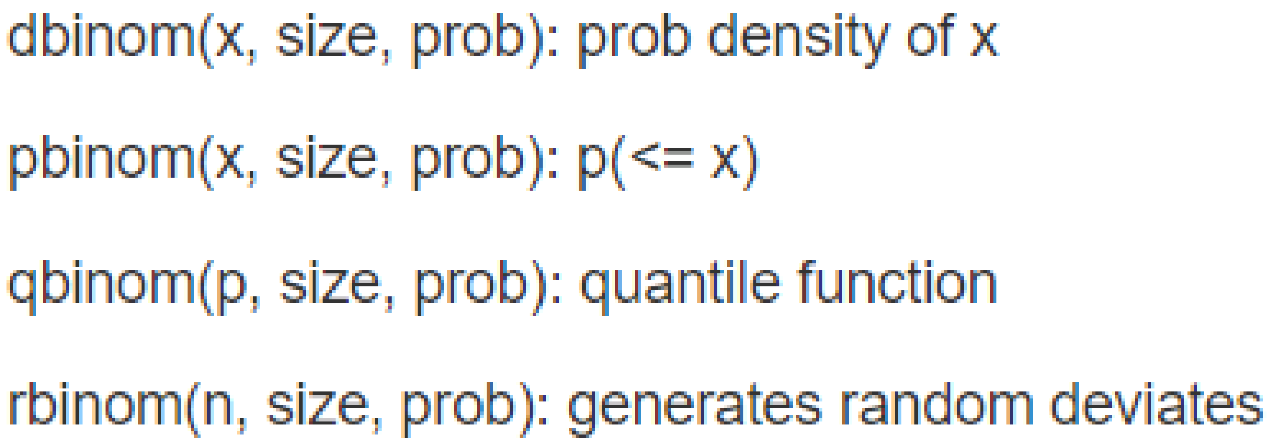
$$= 3(.2)^2(.8)^1 = 0.096$$

Creating Binomial Prob. Distributions

- 例題：某網購公司規定消費者在一週內(7 days)購買商品可全額退款。根據過去記錄，平均每週會有2件退貨商品。
- 請繪製該公司在1個月內(30 days)退貨商品數量的機率分布圖 (probability distribution function, PDF)。
- 估計一個月的退貨商品數超過10件的機率。

Built-in functions

Binomial



dbinom(x, size, prob): prob density of x

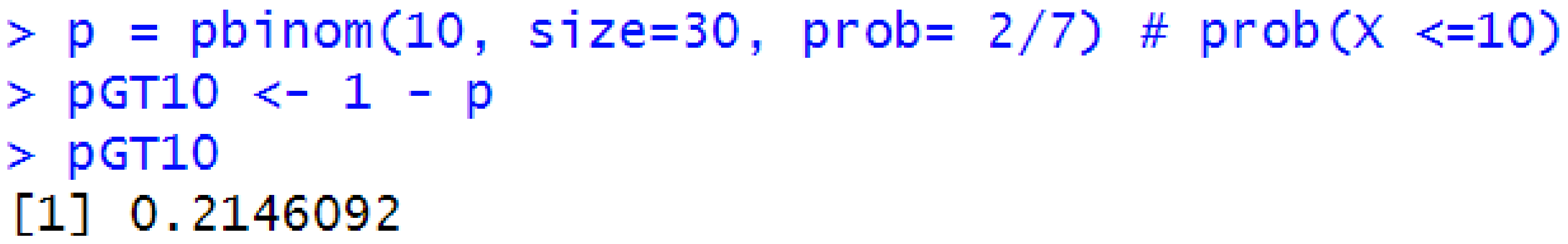
pbinom(x, size, prob): p(<= x)

qbinom(p, size, prob): quantile function

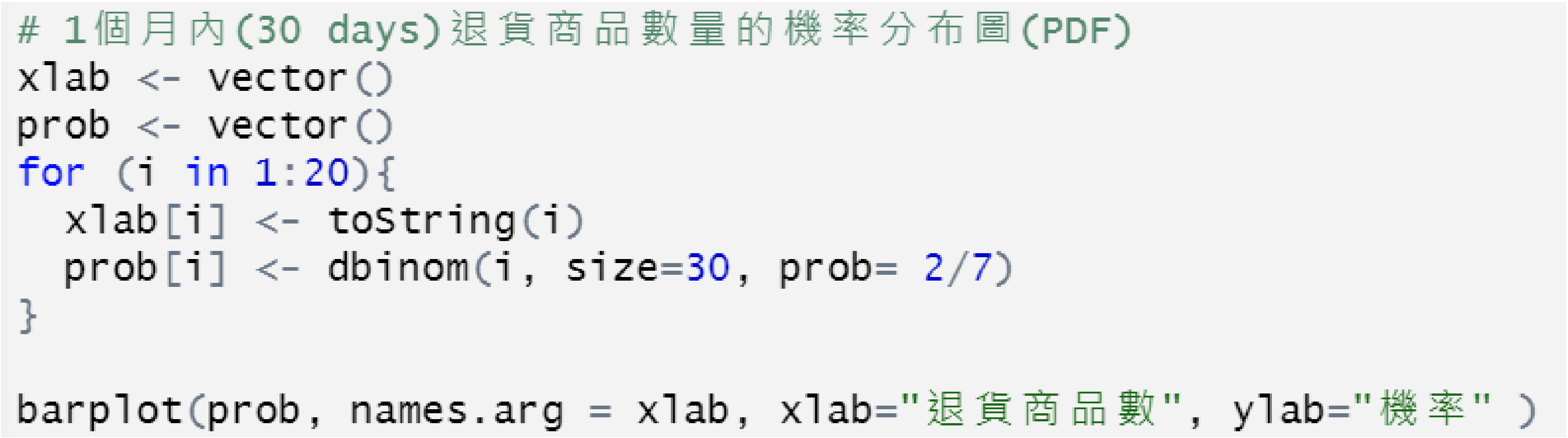
rbinom(n, size, prob): generates random deviates

Normal

```
pnorm(35, mean=30, sd=5)     # CDF, P(x<=35)
dnorm(35, mean=30, sd=5)     # Likelihood, P(x=35)
qnorm(0.7, mean=100, sd=15) # given exceedance prob. --> x
rnorm(20, mean=100, sd=15)  # generating samples
```

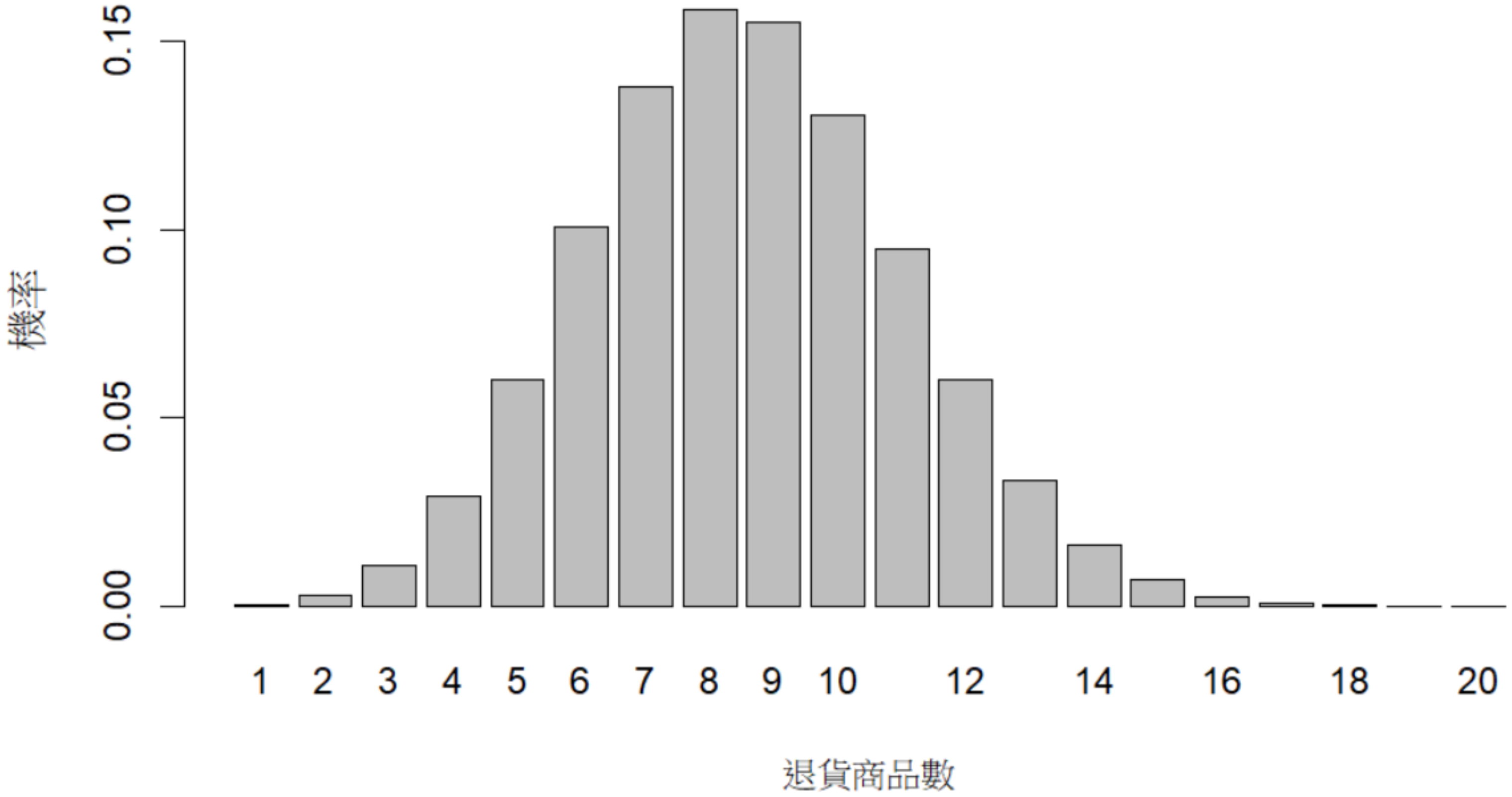


```
> p = pbinom(10, size=30, prob= 2/7) # prob(X <=10)
> pGT10 <- 1 - p
> pGT10
[1] 0.2146092
```

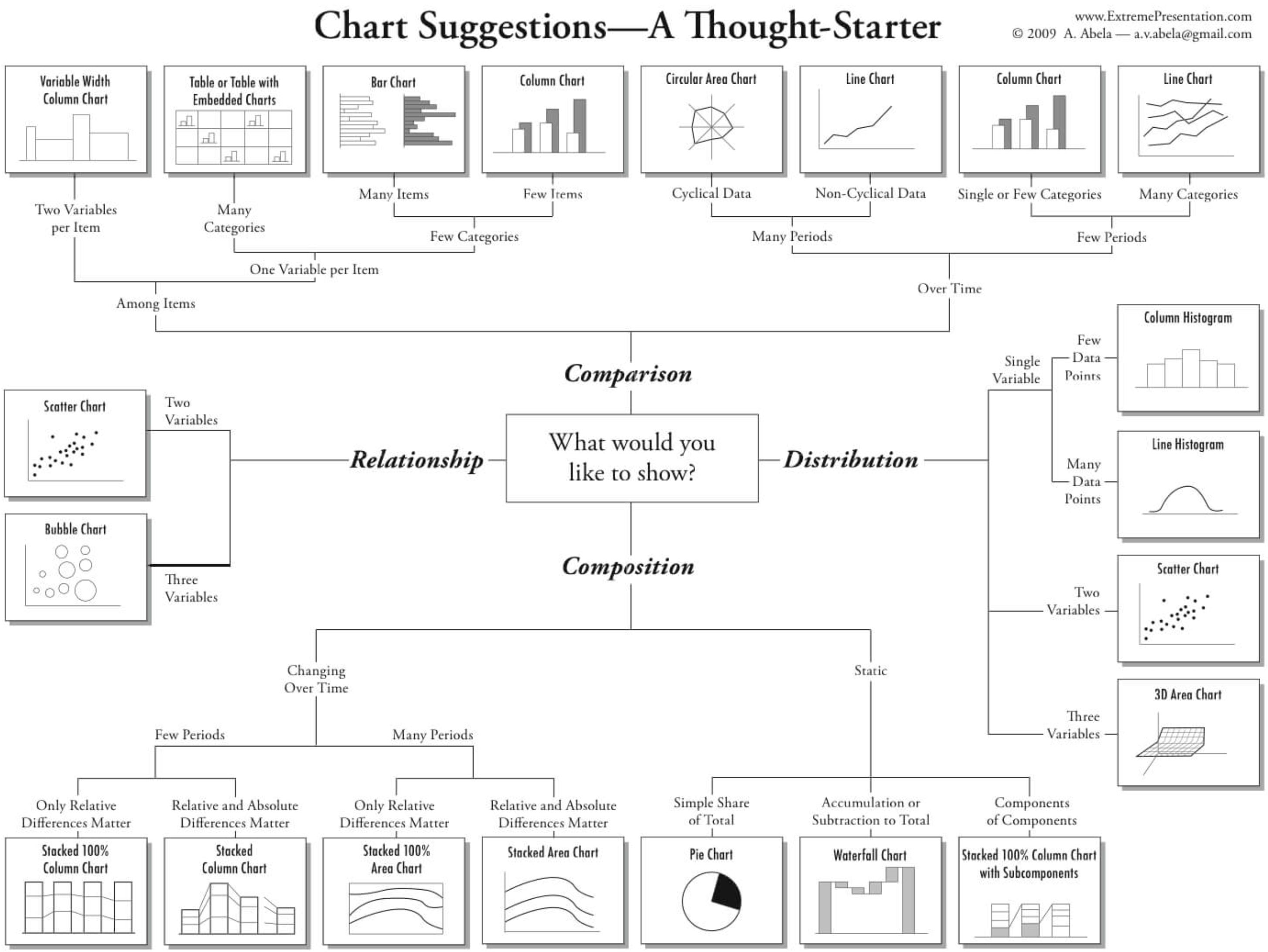


```
# 1個月內(30 days)退貨商品數量的機率分布圖(PDF)
xlab <- vector()
prob <- vector()
for (i in 1:20){
  xlab[i] <- toString(i)
  prob[i] <- dbinom(i, size=30, prob= 2/7)
}

barplot(prob, names.arg = xlab, xlab="退貨商品數", ylab="機率" )
```

1個月內退貨商品數量的機率分布圖



3. 統計圖表



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Commonly-used Plots

- Distribution: Histogram; Box plot
- Composition: Pie chart; Stacked bar chart
- Comparison: Bar chart; Box plot
- Relationship: Scatter plot (bubble chart); Heat map

- Combined:
  - Scatter plot + marginal Histogram / Boxplot

Using ggplot2 package

- Based on grammar of graphics (Wilkinson, 2005)
- Plot specification at a high level of abstraction
- It is very flexible
- theme system for polishing plot appearance
- mature and complete graphics system
- many users, active mailing list

Plots: Using ggplot2 package

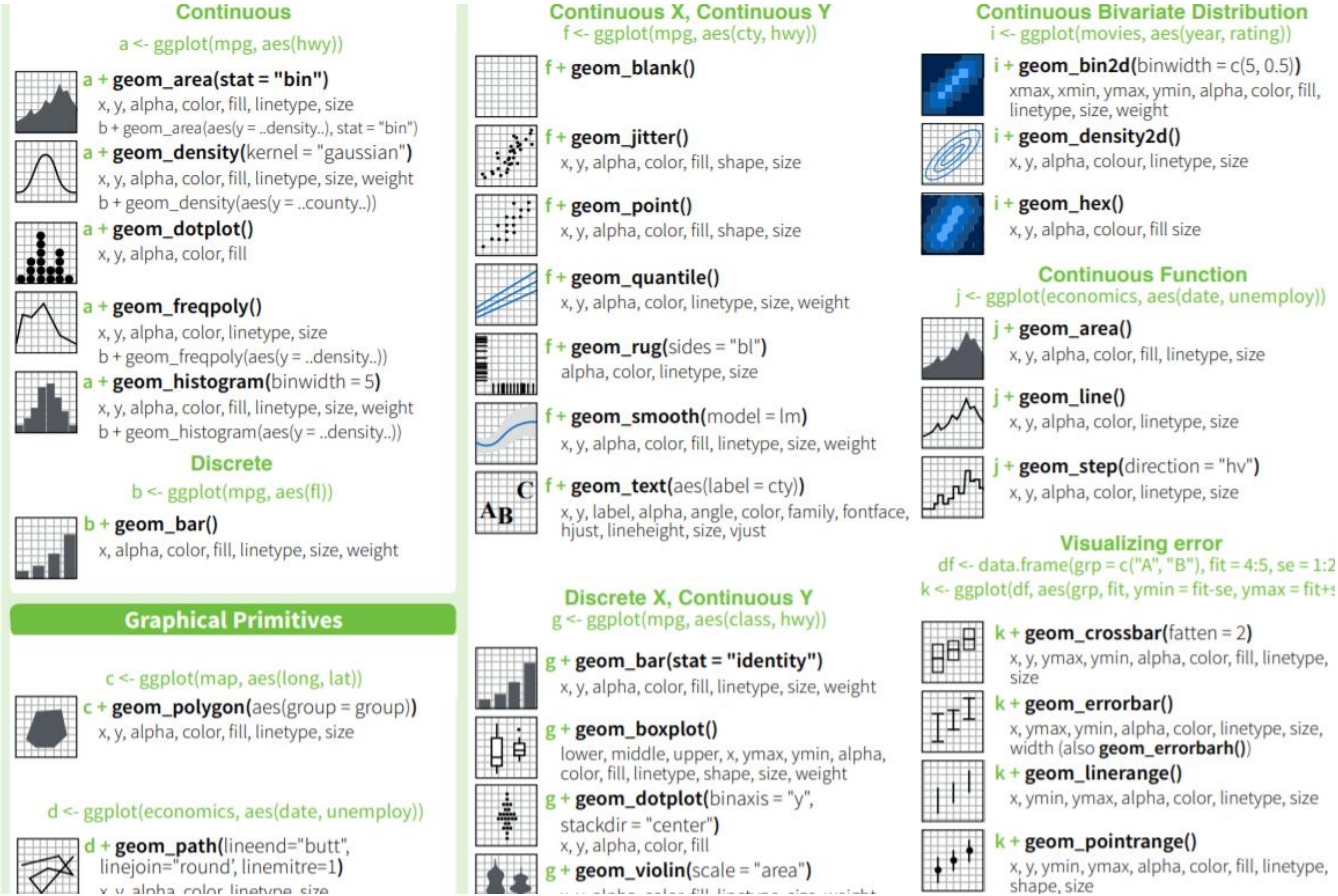
ggplot(檔名 ）＋aes(欄位設定)

＋ geometric objects (geom_) 設定圖表格式

（例如：geom_histogram(),

geom_boxplot()...）

geometric objects (geom_)



Continuous

a <- ggplot(mpg, aes(hwy))

a + **geom_area(stat = "bin")**
x, y, alpha, color, fill, linetype, size
b + geom_area(aes(y = ..density..), stat = "bin")

a + **geom_density(**kernel = "gaussian"**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_density(aes(y = ..county..))

a + **geom_dotplot()**
x, y, alpha, color, fill

a + **geom_freqpoly()**
x, y, alpha, color, linetype, size
b + geom_freqpoly(aes(y = ..density..))

a + **geom_histogram(**binwidth = 5**)**
x, y, alpha, color, fill, linetype, size, weight
b + geom_histogram(aes(y = ..density..))

Discrete

b <- ggplot(mpg, aes(fl))

b + **geom_bar()**
x, alpha, color, fill, linetype, size, weight

Graphical Primitives

c <- ggplot(map, aes(long, lat))

c + **geom_polygon(**aes(group = group)**)**
x, y, alpha, color, fill, linetype, size

d <- ggplot(economics, aes(date, unemploy))

d + **geom_path(**lineend="butt", linejoin="round', linemitre=1**)**
x, y, alpha, color, linetype, size

Continuous X, Continuous Y

f <- ggplot(mpg, aes(cty, hwy))

f + **geom_blank()**

f + **geom_jitter()**
x, y, alpha, color, fill, shape, size

f + **geom_point()**
x, y, alpha, color, fill, shape, size

f + **geom_quantile()**
x, y, alpha, color, linetype, size, weight

f + **geom_rug(**sides = "bl"**)**
alpha, color, linetype, size

f + **geom_smooth(**model = lm**)**
x, y, alpha, color, fill, linetype, size, weight

f + **geom_text(**aes(label = cty)**)**
x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

Discrete X, Continuous Y

g <- ggplot(mpg, aes(class, hwy))

g + **geom_bar(stat = "identity")**
x, y, alpha, color, fill, linetype, size, weight

g + **geom_boxplot()**
lower, middle, upper, x, ymax, ymin, alpha, color, fill, linetype, shape, size, weight

g + **geom_dotplot(**binaxis = "y", stackdir = "center"**)**
x, y, alpha, color, fill

g + **geom_violin(**scale = "area"**)**
x, y, alpha, color, fill, linetype, size, weight

Continuous Bivariate Distribution

i <- ggplot(movies, aes(year, rating))

i + **geom_bin2d(**binwidth = c(5, 0.5)**)**
xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight

i + **geom_density2d()**
x, y, alpha, colour, linetype, size

i + **geom_hex()**
x, y, alpha, colour, fill size

Continuous Function

j <- ggplot(economics, aes(date, unemploy))

j + **geom_area()**
x, y, alpha, color, fill, linetype, size

j + **geom_line()**
x, y, alpha, color, linetype, size

j + **geom_step(**direction = "hv"**)**
x, y, alpha, color, linetype, size

Visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2
k <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+s

k + **geom_crossbar(**fatten = 2**)**
x, y, ymax, ymin, alpha, color, fill, linetype, size

k + **geom_errorbar()**
x, ymax, ymin, alpha, color, linetype, size, width (also **geom_errorbarh()**)

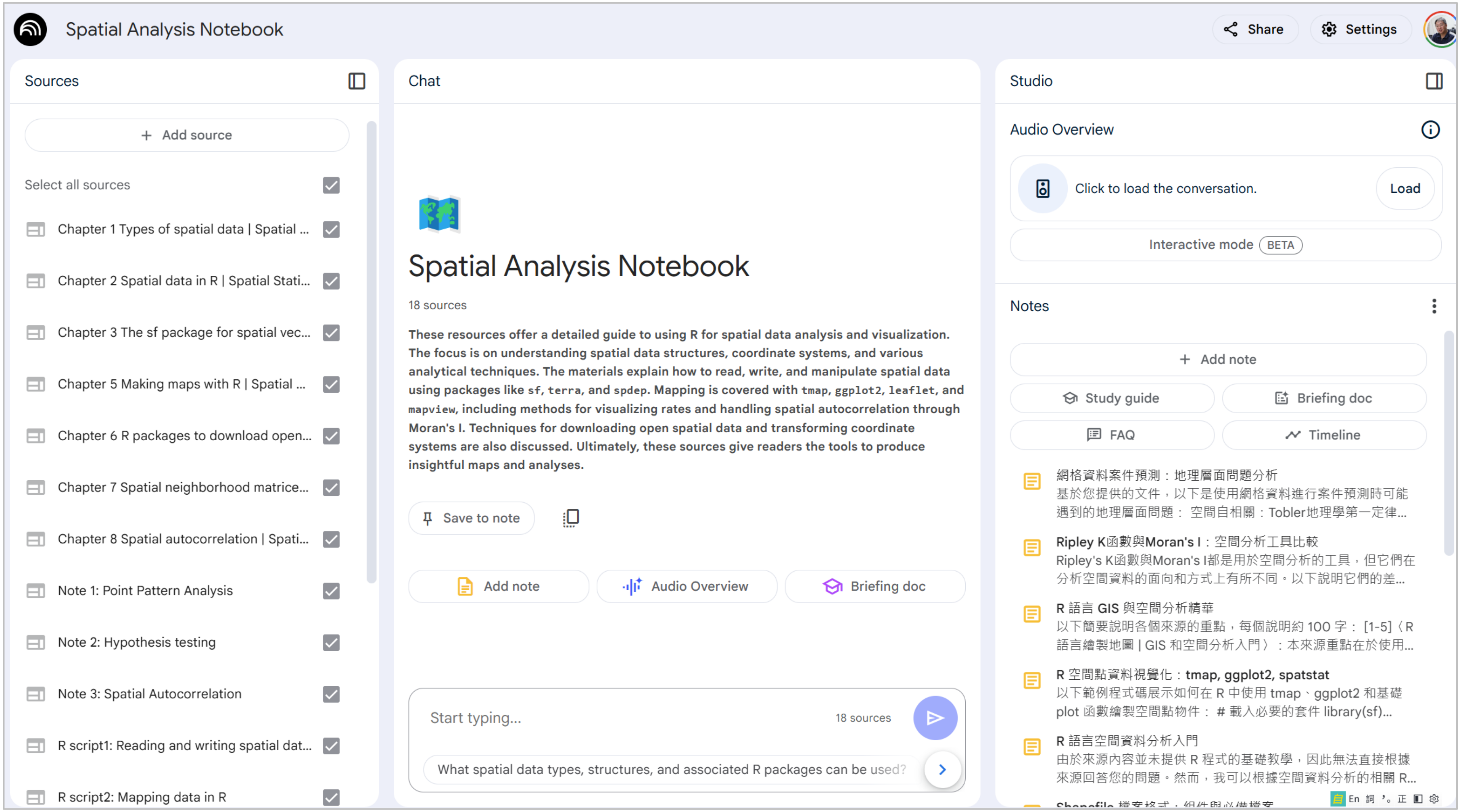
k + **geom_linerange()**
x, ymin, ymax, alpha, color, linetype, size

k + **geom_pointrange()**
x, y, ymin, ymax, alpha, color, fill, linetype, shape, size

課後釋疑問答：Using NotebookLM

填表，提供 gmail address

https://docs.google.com/spreadsheets/d/1SgWy3qtFE0USzHqQhQ4cQIfgd9Fec7eItXV6bzvQ_CM/edit?gid=0#gid=0



Select all sources ☑

- Chapter 1 Types of spatial data | Spatial ... ☑
- Chapter 2 Spatial data in R | Spatial Stati... ☑
- Chapter 3 The sf package for spatial vec... ☑
- Chapter 5 Making maps with R | Spatial ... ☑
- Chapter 6 R packages to download open... ☑
- Chapter 7 Spatial neighborhood matrice... ☑
- Chapter 8 Spatial autocorrelation | Spati... ☑
- Note 1: Point Pattern Analysis ☑
- Note 2: Hypothesis testing ☑
- Note 3: Spatial Autocorrelation ☑
- R script1: Reading and writing spatial dat... ☑
- R script2: Mapping data in R ☑

Spatial Analysis Notebook

18 sources

These resources offer a detailed guide to using R for spatial data analysis and visualization. The focus is on understanding spatial data structures, coordinate systems, and various analytical techniques. The materials explain how to read, write, and manipulate spatial data using packages like `sf`, `terra`, and `spdep`. Mapping is covered with `tmap`, `ggplot2`, `leaflet`, and `mapview`, including methods for visualizing rates and handling spatial autocorrelation through Moran's I. Techniques for downloading open spatial data and transforming coordinate systems are also discussed. Ultimately, these sources give readers the tools to produce insightful maps and analyses.

What spatial data types, structures, and associated R packages can be used?

Audio Overview

Click to load the conversation.

網格資料案件預測：地理層面問題分析
基於您提供的文件，以下是使用網格資料進行案件預測時可能遇到的地理層面問題： 空間自相關：Tobler地理學第一定律...

Ripley K函數與Moran's I：空間分析工具比較
Ripley's K函數與Moran's I都是用於空間分析的工具，但它們在分析空間資料的面向和方式上有所不同。以下說明它們的差...

R 語言 GIS 與空間分析精華
以下簡要說明各個來源的重點，每個說明約 100 字： [1-5]〈R 語言繪製地圖 | GIS 和空間分析入門〉：本來源重點在於使用...

R 空間點資料視覺化：tmap, ggplot2, spatstat
以下範例程式碼展示如何在 R 中使用 tmap、ggplot2 和基礎 plot 函數繪製空間點物件： # 載入必要的套件 library(sf)...

R 語言空間資料分析入門
由於來源內容並未提供 R 程式的基礎教學，因此無法直接根據來源回答您的問題。然而，我可以根據空間資料分析的相關 R...

Shapefile 檔案格式：組件與必備檔案

本週作業

- **1. 機率分布：**

  (a).某一都市有10萬人口，假設流行一種新興疾病，每人每年被感染機率 p = 0.01，沒有免疫與任何預防措施。請繪製該市的每年感染人數頻率分布圖。

  (b).該市衛生當局定義若某年的感染人數超過960人，該年則視為疫情爆發。市長的任期是4年，若任期內爆發疫情事件，就必須辭職下台。請評估市長在任期四年內，因疫情爆發而辭職的機率?

- **2. 繪製統計圖表與統計檢定 (Student.csv)**

  (a). 比較不同性別(Gender)，讀書時間(StudyHrs)是否有差異？

  (b). 比較不同性別(Gender)，對於虔誠信仰宗教的比例(ReligImp)是否有差異？