



Use ChatGPT to Assist Statistical Teaching and Learning: A Guide to Prompt Engineering

Tse-min Lin

University of Texas at Austin

January 6, 2025

National Sun-Yat Sen University

ChatGPT & Wittgenstein's Philosophy of Language

“The philosopher Ludwig Wittgenstein wrote two influential books on language, ”Tractatus Logico-Philosophicus“ and ”Philosophical Investigations.“ As a language model, which of Wittgenstein's books do you think is closer to your model?”

ChatGPT's Response

“As a language model, I am designed to process and generate natural language, so my functioning is more similar to the approach taken in "Philosophical Investigations." While "Tractatus Logico-Philosophicus" is a highly structured and logical work that seeks to establish a system of language and meaning, its focus is on formal logic and mathematical notation, rather than natural language. In contrast, "Philosophical Investigations" is more concerned with the way that natural language is used in everyday contexts, and the complex relationships between language, thought, and the world. Like "Philosophical Investigations," my primary focus is on understanding and generating natural language, and I am designed to learn from and adapt to the diverse ways that language is used in different contexts. While my functioning is of course much simpler than that of the human mind, Wittgenstein's approach to language as a complex and ever-evolving social practice that is used in various contexts for different purposes is a guiding principle of my design.”

Jensen Huang: Domain Expertise versus Programming Skills



Scope of This Presentation

- Focus on teaching and learning of statistics
- Cover mostly basic statistics, with some advanced materials
- Use ChatGPT 3.5 / 4o
- Write code in R, but can easily extended to Python and Stata

Tips for Prompt Engineering

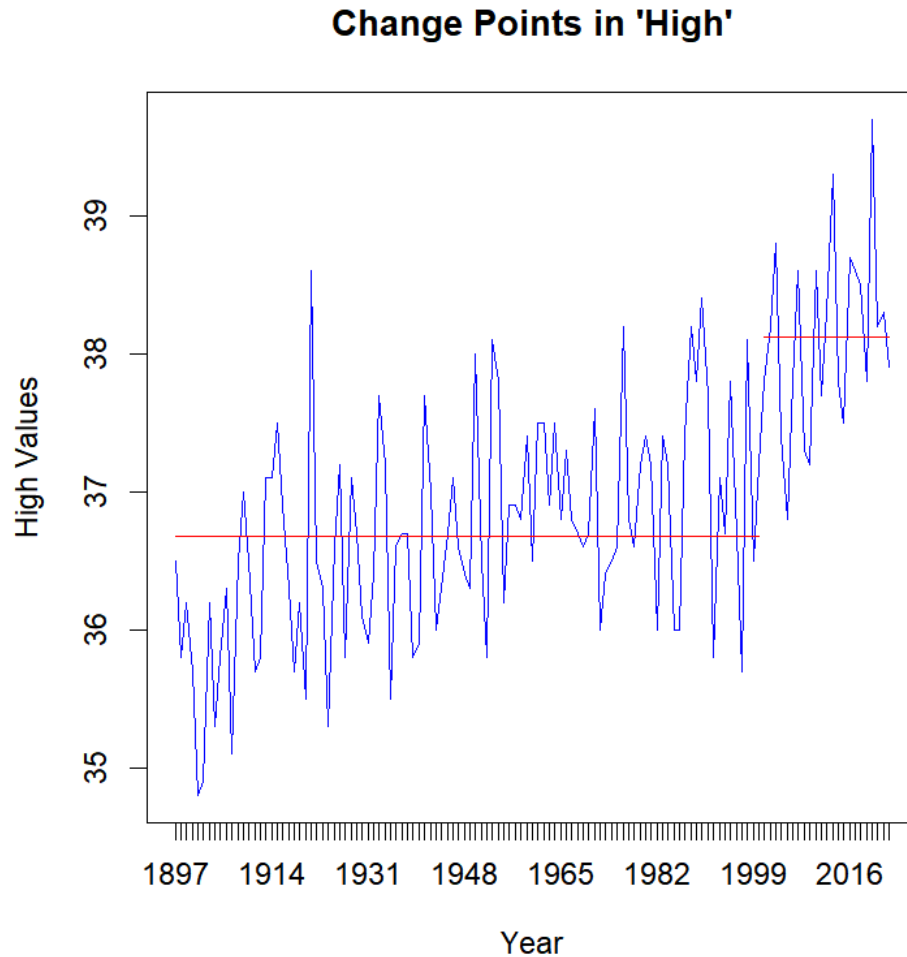
- GPT can help with programming, but you must have domain knowledge.
- Interactions with ChatGPT are not exactly repeatable.
- When doing empirical analysis, provide the path to your data file.
- If desirable, provide the exact names of `package::function`.
- You may have to separately install the necessary packages.
- Read GPT's explanation carefully to make sure it does what you want.
- Click the "Copy code" button to copy the generated code then paste it to R-Studio to run.
- If GPT provides code in separate parts, ask it to combine them.
- If the code is installing a package that is already loaded, you may have to comment out the `install.packages` line to avoid restarting R.
- When errors occur, copy the error messages and paste them to GPT for corrected code.
- Feel free to enter follow-up instructions to improve the results.
- In case of a catch-22 situation, use your domain expertise.

Example 1: Change Points of Time Series

Write R code to read the data in C:\Users\lint\Desktop\GPT\Taipei.csv. Use `changepoint::cpt.mean` to identify the change point in “high”. Plot the result. Print the change points in years.

Example 1: Change Point in Taipei's Annual Highest Temperature (1897-2023)

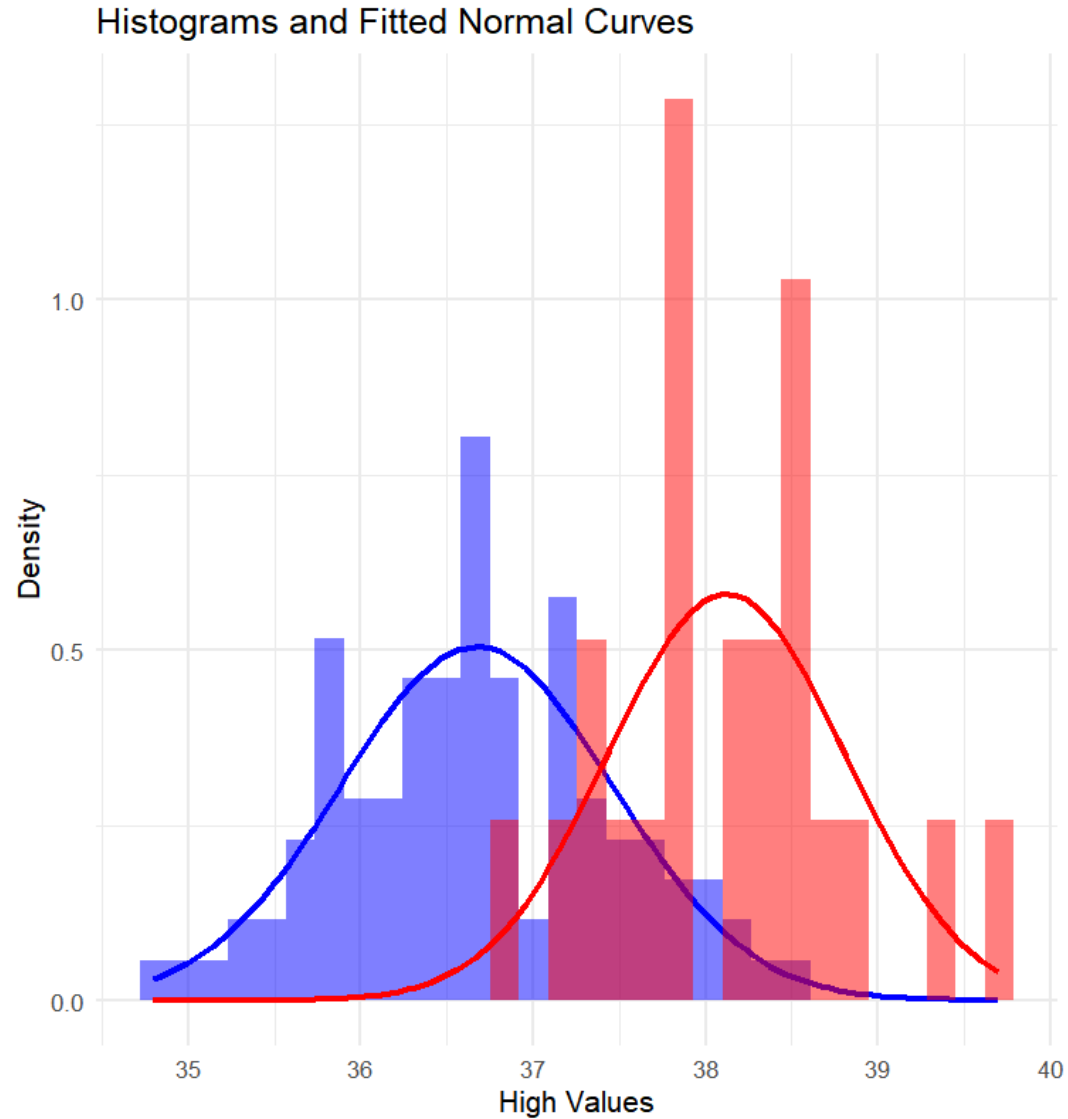
- Change point: 2000



Example 2: Descriptive Statistics before & after the Change Point

Write R code to read the data in C:\Users\lint\Desktop\GPT\Taipei.csv. Plot the histogram for the variable "high" for year<2000 and year>2000, respectively, together with a fitted normal curve. Calculate the mean and standard deviation of "high" for the two periods, respectively, and conduct a difference of means test assuming both equal and unequal variance.

Example 2: Histograms & Fitted Normal Curves



Example 2: Difference of Means Test (Equal Variance)

Difference of Means Test (Equal Variance):

```
> print(t_test_equal_var)
```

Two Sample t-test

data: data_pre_2000\$high and data_post_2000\$high

t = -8.1271, df = 124, p-value = 3.802e-13

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.801796 -1.096052

sample estimates:

mean of x mean of y

36.67282 38.12174

Example 2: Difference of Means Test (Unequal Variance)

Difference of Means Test (Unequal Variance):

```
> print(t_test_unequal_var)
```

Welch Two Sample t-test

data: data_pre_2000\$high and data_post_2000\$high

t = -8.8814, df = 36.217, p-value = 1.269e-10

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.779722 -1.118126

sample estimates:

mean of x mean of y

36.67282 38.12174

Example 3: Regression Analysis

Write R code to read the data in C:\Users\lint\Desktop\GPT\Taipei.csv. Regress "high" on "year". Use `stargazer::stargazer` to show the result in a table. Plot "high" versus "year" with a fitted regression line.

Example 3: Regression Result

=====

Dependent variable:

high

year 0.017***

(0.002)

Constant 4.383

(3.462)

Observations 127

R2 0.414

Adjusted R2 0.410

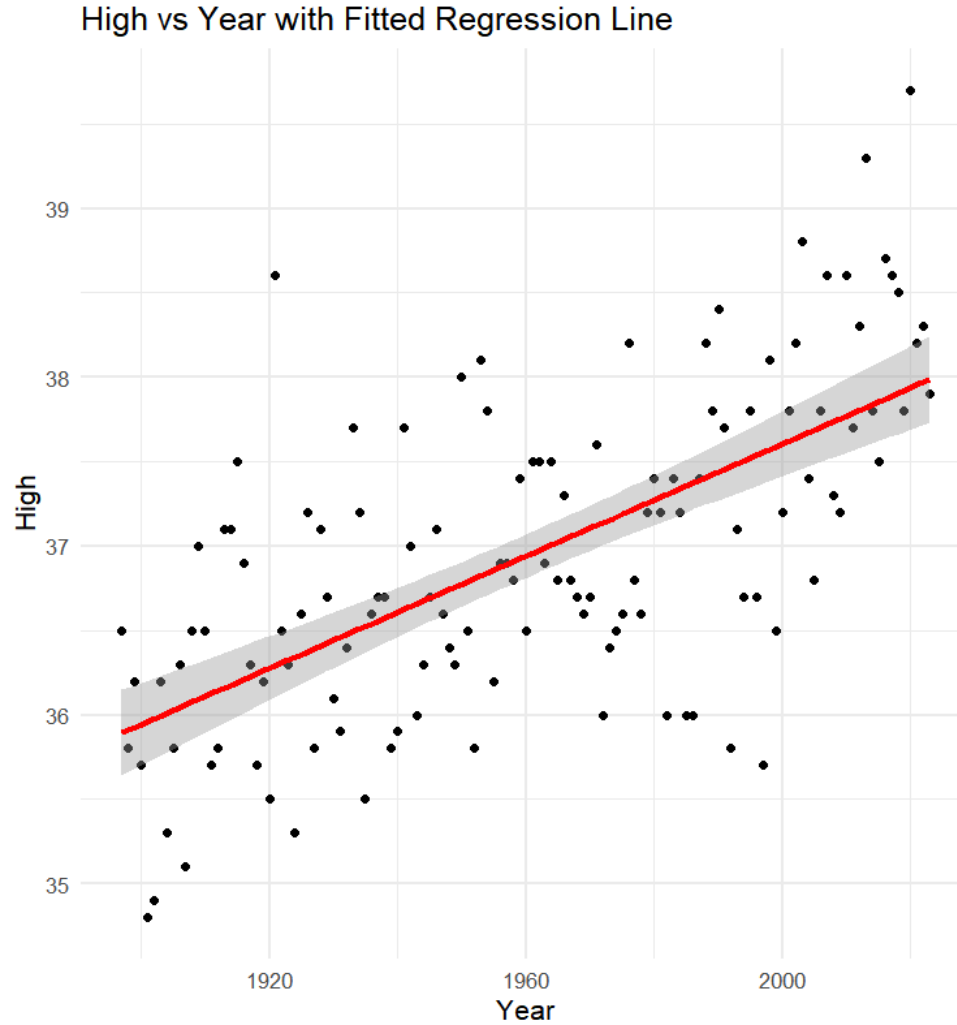
Residual Std. Error 0.730 (df = 125)

F Statistic 88.457*** (df = 1; 125)

=====

Note: *p<0.1; **p<0.05; ***p<0.01

Example 3: Regression Plot



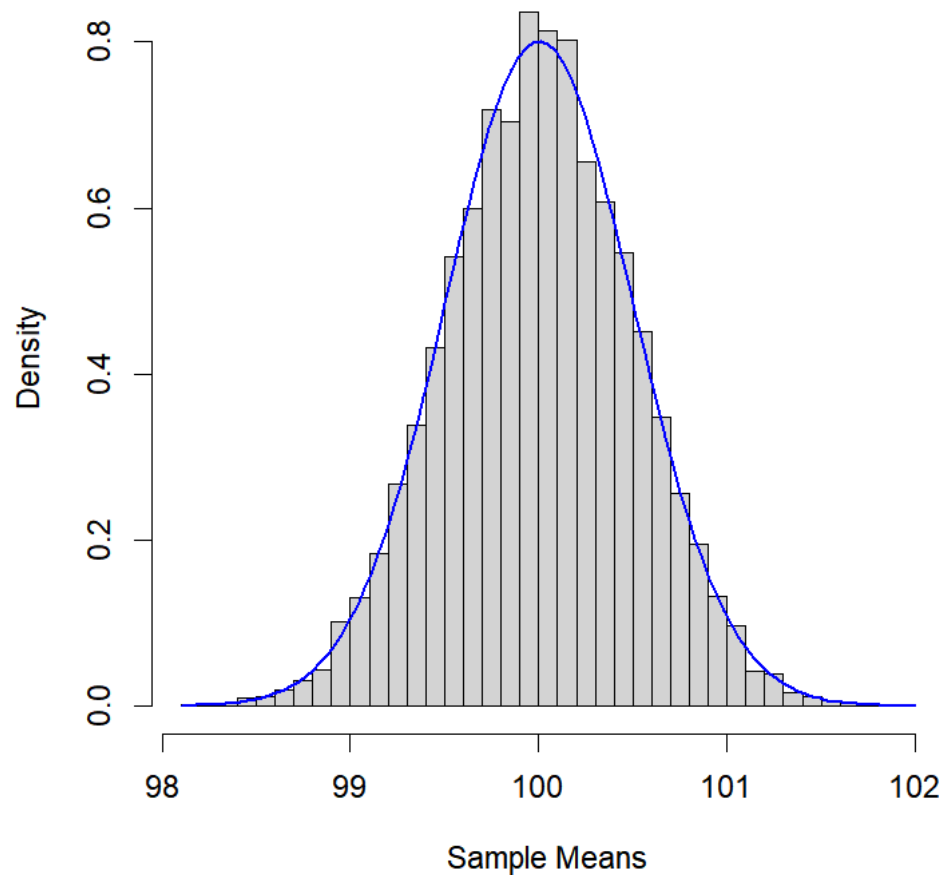
Example 4: Sampling Distribution of the Sample Mean

Please write R code to simulate 1600 observations from a normal distribution with an expected value of 100 and a standard deviation of 20. Calculate the mean of the sample. Repeat this 10000 times. Plot a histogram of the sample means with a fitted normal curve. Calculate the mean and standard deviation of the sample means.

Sampling Distribution of the Sample Mean

- Mean=100.003 Standard deviation=0.4979269

Histogram of Sample Means with Fitted Normal Curve



Theoretical vs. Simulated Characteristics of the Sampling Distribution

- Theoretically, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- $E(\bar{X}) = \mu = 100$
- $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{1600}} = 0.5$
- Simulated results:
- Mean of $\bar{X} = 100.003$
- Standard deviation of $\bar{X} = 0.4979269$

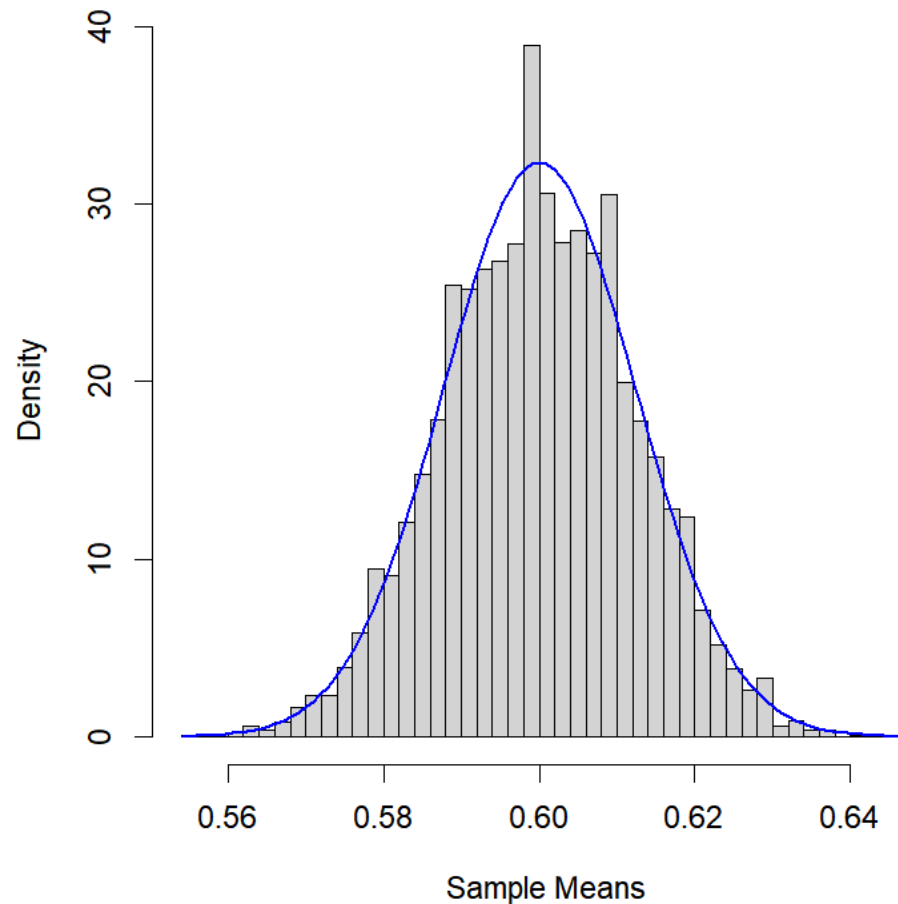
Example 5: Sampling Distribution of the Sample Proportion

Please write R code to simulate 1600 observations from a Bernoulli distribution with parameter 0.6. Calculate the mean of the sample. Repeat this 10000 times. Plot a histogram of the sample means with a fitted normal curve. Calculate the mean and standard deviation of the sample means.

Sampling Distribution of the Sample Proportion

- Mean=0.600 Standard deviation=0.01234397

Histogram of Sample Means with Fitted Normal Curve



Theoretical vs. Simulated Characteristics of the Sampling Distribution

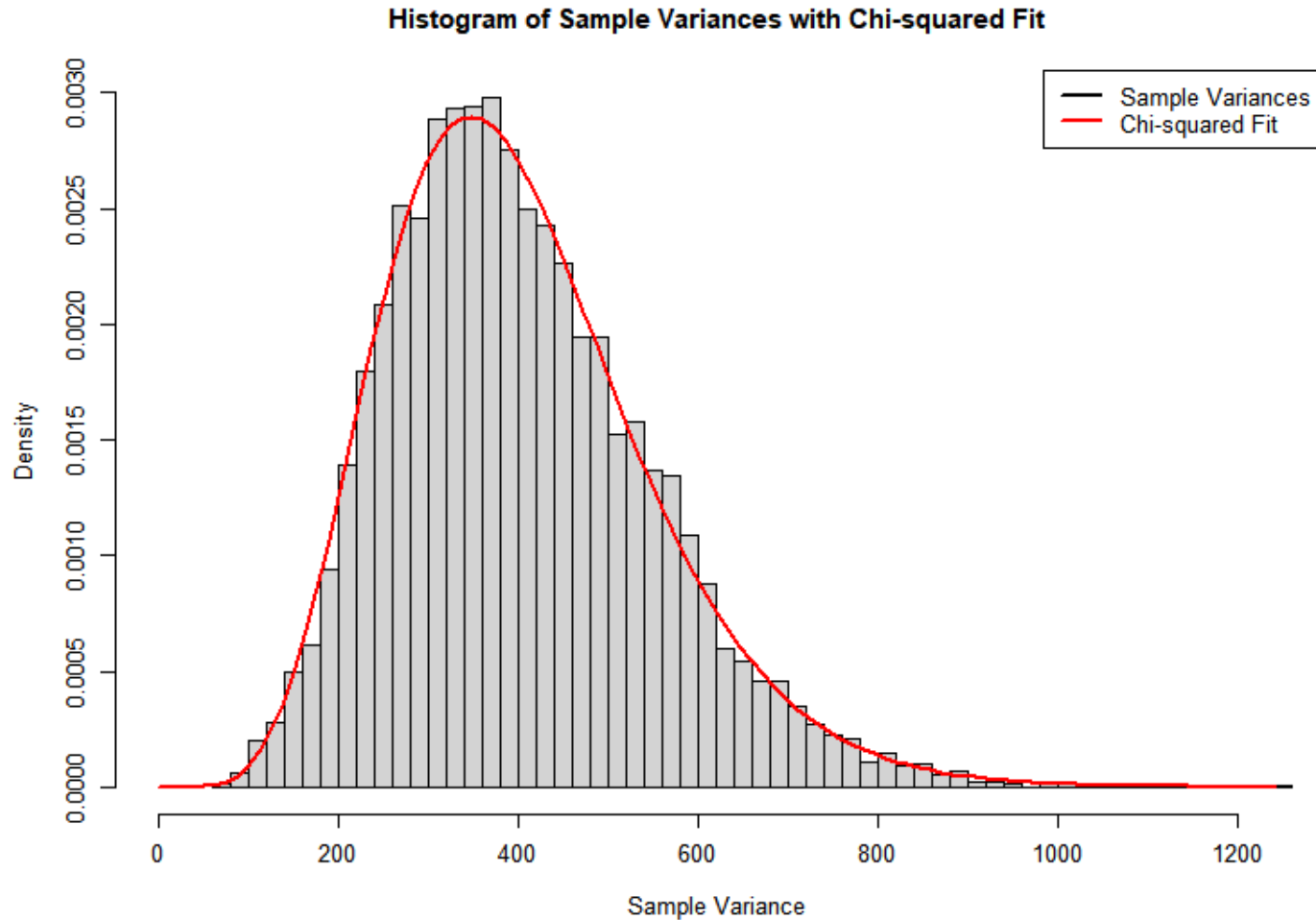
- According to the CLT, $P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$.
- $E(P) = \pi = 0.6$
- $SD(P) = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.6(1-0.6)}{1600}} = 0.0122$
- Simulated Results:
- Mean of P = 0.600
- Standard deviation of P = 0.01234397

Example 6: Sampling Distribution of the Sample Variance

Please write R code to simulate $n=16$ observations from a normal distribution with an expected value of 100 and a standard deviation of 20. Calculate the sample variance. Repeat this 10000 times. Calculate the mean of the sample variances. Plot a histogram of the sample variances and fit it with a Chi-squared distribution with $df=15$.

Sampling Distribution of the Sample Variance

- Mean of the sample variances: 401.8384



Theoretical vs. Simulated Characteristics of the Sampling Distribution

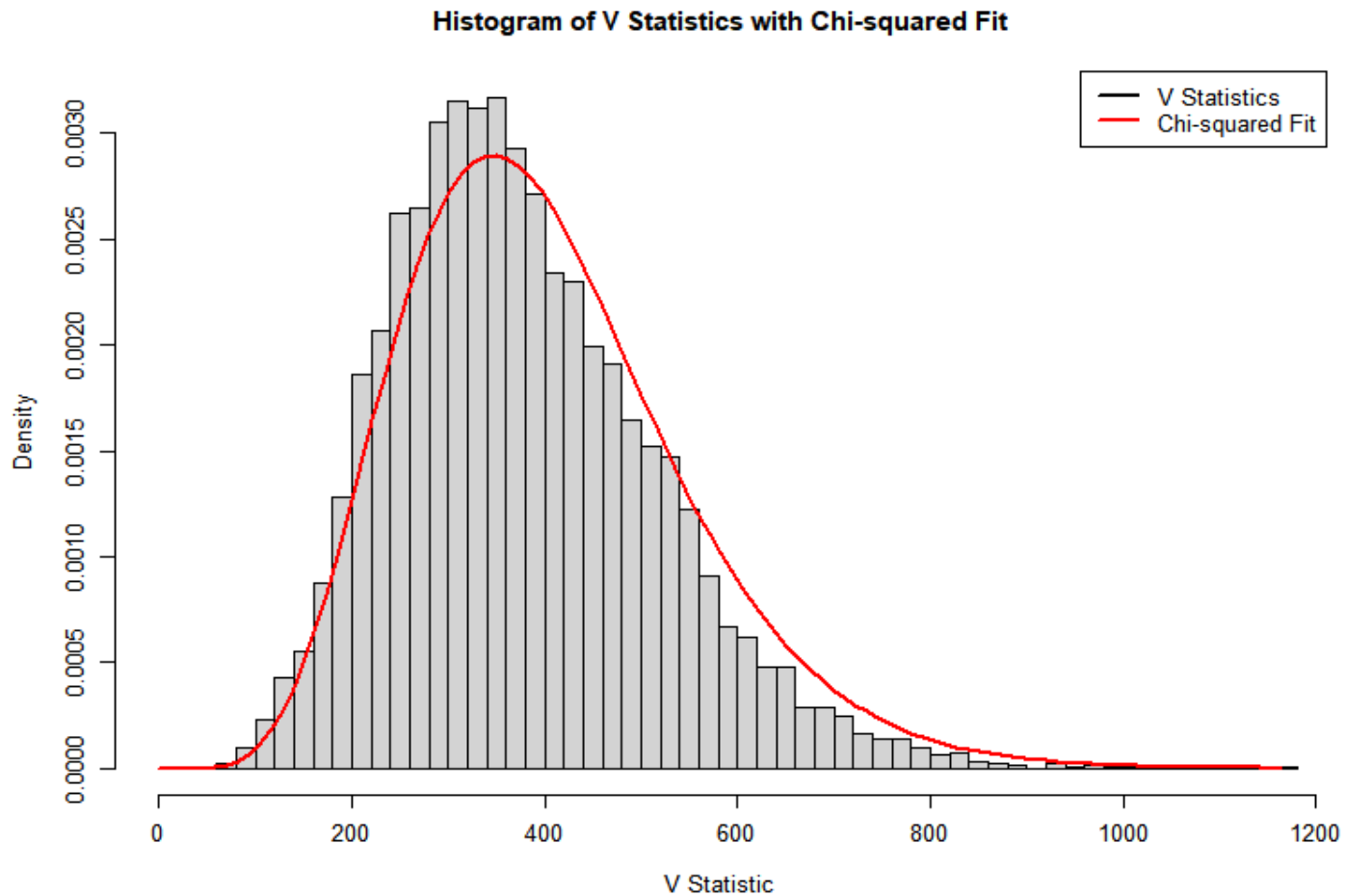
- $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
- Theoretically,
- $\frac{n-1}{\sigma^2} s^2 \sim \chi^2(n-1)$
- $E(s^2) = \sigma^2 = 400$
- s^2 is an *unbiased* estimator of σ^2 .
- Simulated Result:
- Mean of $s^2 = 401.8384$

Example 7: Why is the Denominator of Variance $n-1$ rather than n ?

Please write R code to simulate $n=16$ observations from a normal distribution with an expected value of 100 and a standard deviation of 20. Calculate the sum of squared deviations from the sample mean and divide it by n . Call this statistics V . Repeat this 10000 times. Calculate the mean of the V statistics. Plot a histogram of the V statistics and fit it with a Chi-squared distribution with $df=15$.

Sampling Distribution of V

- Mean of the sample variances: 376.7235



Theoretical vs. Simulated Characteristics of the Sampling Distribution

- $V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$
- Theoretically,
- $E(V) = \frac{n-1}{n} E(s^2) = \frac{n-1}{n} \sigma^2 = 375$
- V is a *biased* estimator of $\sigma^2 = 400$.
- Simulated Result:
- Mean of $V = 376.7235$
- 請參考：林澤民，〈[統計學中算變異量為什麼要除以n-1？什麼是「自由度」？](#)〉

Example 8: Simulating Multilevel Data

Write R code to simulate multilevel data with $N_2=20$ level-2 groups and $N_1=30$ level-1 units per group. There are two level-1 variables, x_1 and x_2 , and one level-2 variable, z . The coefficient for x_1 is fixed at 1, the coefficient for x_2 is fixed at 2, and the coefficient of z is fixed at 3. The intercept contains a fixed effect of 0.5 and a level-2 random effect. Generate Y and estimate the model.

Simulating Multilevel Data

Random effects:

Groups	Name	Variance	Std.Dev.
group	(Intercept)	0.8418	0.9175
Residual		0.9587	0.9791

Number of obs: 600, groups: group, 20

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	0.62967	0.21299	2.956
x1	0.92261	0.04154	22.210
x2	2.03558	0.03983	51.112
z	2.72240	0.19675	13.837

Correlation of Fixed Effects:

(Intr)	x1	x2	
x1	0.001		
x2	-0.003	-0.041	
z	0.192	0.011	0.003

Example 9: Simulating Panel Data with AR(1)

Write R code to simulate panel data with $N=20$ cross-sectional units and $T=30$ temporal units. There are two time-varying variables, x_1 and x_2 , and one time-invariant variable, z . The coefficient for x_1 is fixed at 1, the coefficient for x_2 is fixed at -2, and the coefficient of z is fixed at 3. The intercept contains a fixed effect of 0 and a time-invariance random effect. The idiosyncratic error term is an AR1 process with a coefficient of 0.5. Estimate the model with `nlme::lme` to account for the autocorrelation.

Simulating Panel Data with Autocorrelation

Random effects:

Formula: $\sim 1 \mid \text{id}$

(Intercept) Residual

StdDev: 0.9041356 1.130561

Correlation Structure: AR(1)

Formula: $\sim \text{time} \mid \text{id}$

Parameter estimate(s):

Phi

0.5074867

Fixed effects: $Y \sim x1 + x2 + z$

Value Std.Error DF t-value p-value

(Intercept) 0.1866778 0.22084959 578 0.84527 0.3983

x1 0.9272788 0.03613220 578 25.66350 0.0000

x2 -1.9614738 0.03517054 578 -55.77036 0.0000

z 2.7684981 0.20401623 18 13.56999 0.0000

Correlation:

(Intr) x1 x2

x1 0.001

x2 -0.004 -0.010

z 0.192 0.013 0.005

Example 10 : Text Analysis of *Dream of the Red Chamber*

請寫R的程式碼來分析兩個文本。第一個文本稱為「前80回」，位於C:\Users\lint\Desktop\GPT\Dream of the Red Chamber 80.txt。第二個文本稱為「後40回」，位於C:\Users\lint\Desktop\GPT\Dream of the Red Chamber 40.txt。首先，請找出前80回最常出現的20個字（標點符號不算），算出每個字出現次數佔前80回總字數的百分比。其次，就同樣這20個字，算出每個字在後40回中出現次數佔後40回總字數的百分比。請用表格列出這20個字及其在前80回和後40回中的比例。

《紅樓夢》前八十回和後四十回高頻字列表

Character Percentage_80 Percentage_40

1	了	2.8659661	2.9563964
2	的	2.0486983	2.3493326
3	不	2.0179377	2.1334687
4	一	1.7495361	1.4756385
5	來	1.5054615	1.6799840
6	道	1.4670610	1.5950889
7	人	1.3818158	1.5528547
8	我	1.2931528	1.1650676
9	是	1.2853119	1.5938091
10	說	1.2644027	1.4423631
11	他	1.1053723	0.9551763
12	這	1.0937114	1.0131950
13	你	1.0331953	0.8557765
14	去	0.8480284	0.8378589
15	個	0.8331507	0.6531375
16	也	0.8202835	0.8672949
17	子	0.8098289	0.6151693
18	玉	0.8019880	0.8775335
19	有	0.7943481	0.8664417
20	兒	0.7897239	0.8758271

Example 11：《紅樓夢》文本分析之二

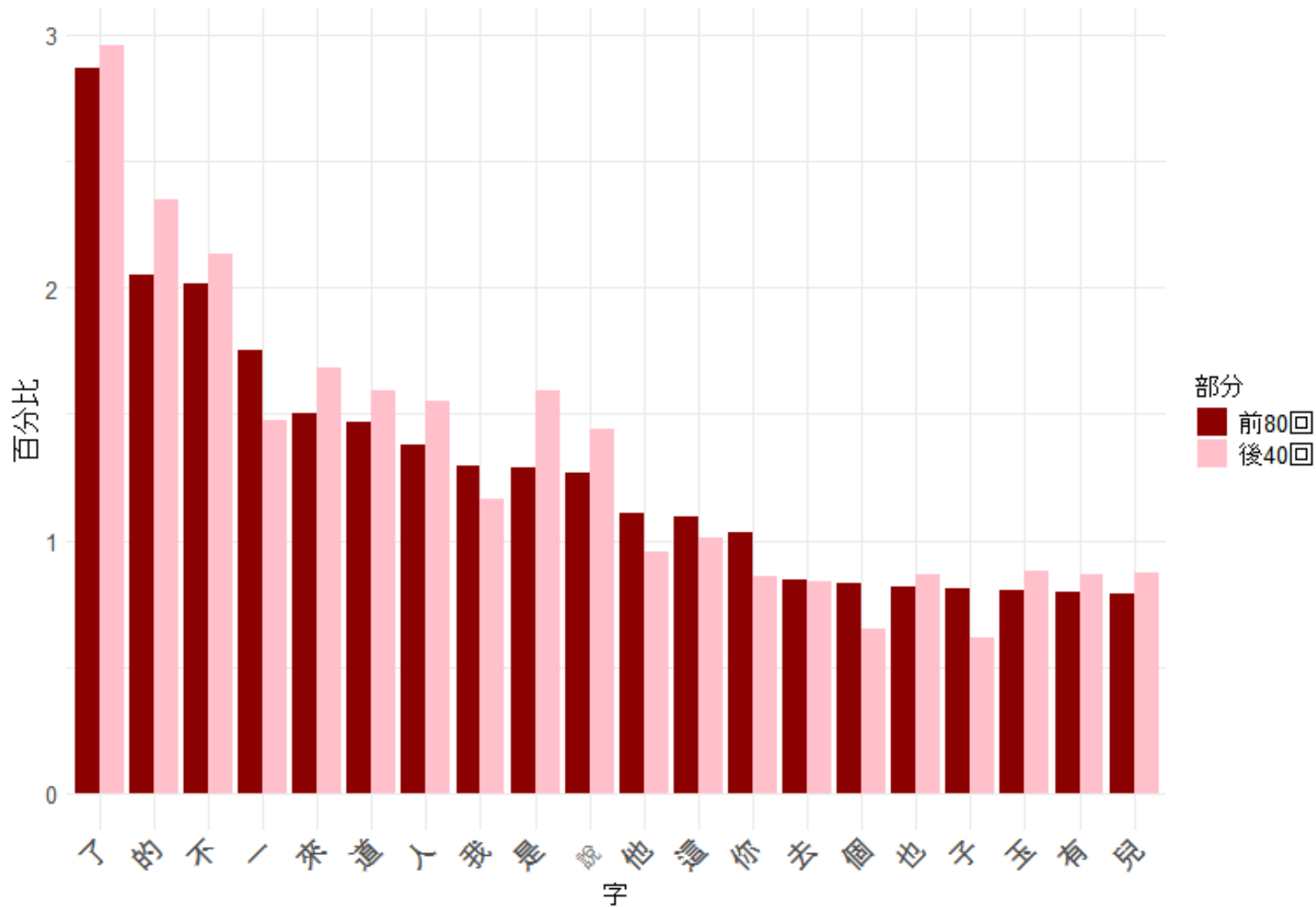
再來，畫長條圖，將每個字的兩個百分比，用不同顏色的長條並列。特別注意：每個字要有兩個不同顏色的長條並列以便比較，不要相疊。前**80**回的長條用深紅色，後**40**回的長條用粉紅色。**20**個字的順序請依照前**80**回百分比的大小順序由大而小排列。最後，請為長條圖加上「紅樓夢前八十回最常用**20**字使用頻率與後四十回比較圖」的標題，並務必加上中文的圖例標籤。

修正錯誤並調整、美化長條圖

- 執行前頁指令後，可能需要再逐一輸入下列指令以修正錯誤並調整、美化長條圖：
- 錯誤發生在 `gather(result, key = "Part", value = "Percentage", -Character)`：沒有這個函式 "gather"
- 20個字的順序請依照前80回百分比的大小順序由大而小排列。同時也請把圖中的字體加大。
- 你還是沒把長條的順序自左自右依照前80回百分比的大小順序由大而小排列呀。
- 結果你把後40回的長條放左邊，前80回放右邊了，請調整，把前80回的長條放左邊，後40回放右邊。

《紅樓夢》前八十回和後四十回高頻字比較

紅樓夢前八十回最常用20字使用頻率與後四十回比較圖

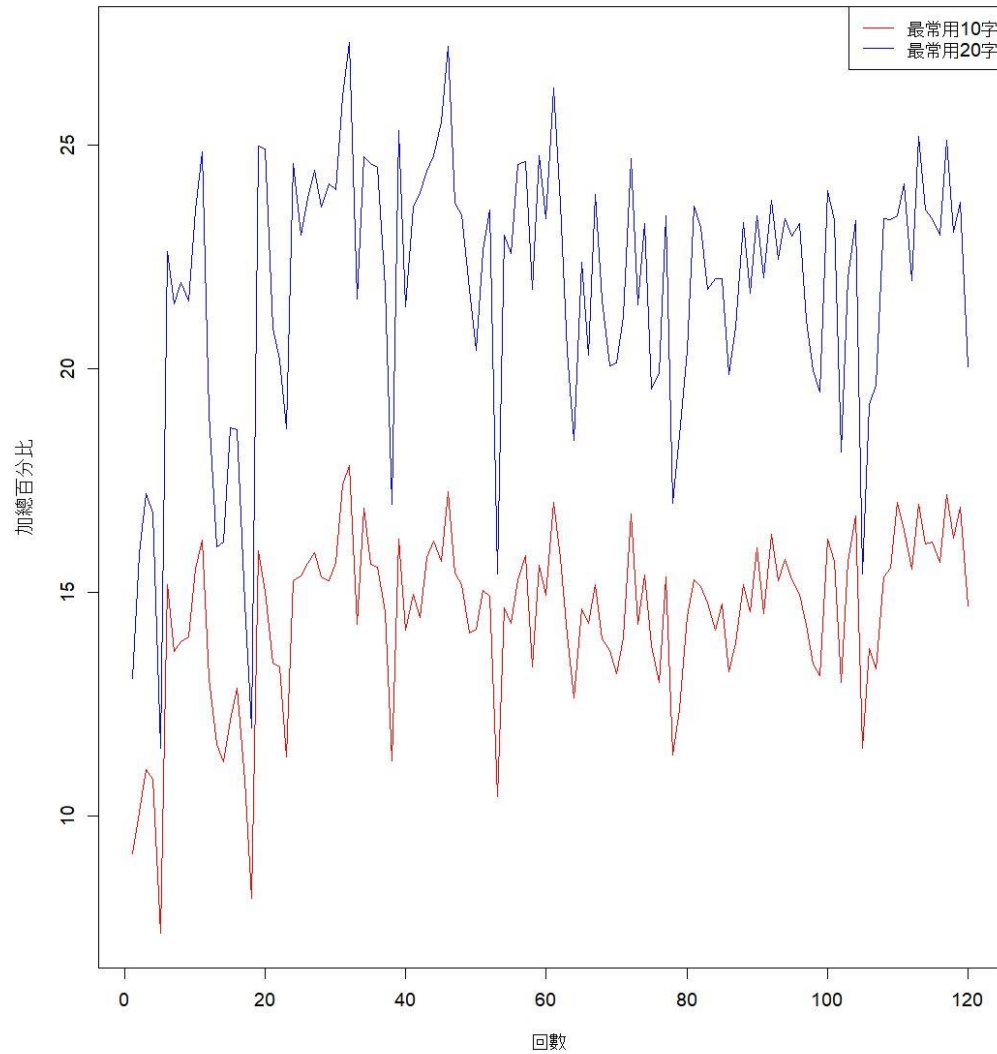


Example 12：更多《紅樓夢》文本分析

- （一）請寫R的程式碼來分析兩個文本。第一個文本稱為「前80回」，位於C:\Users\lint\Desktop\GPT\Dream of the Red Chamber 80.txt。首先，請找出前80回最常出現的20個中文字（標點符號不算），算出這20個字每一個字在前80回中出現的次數和其佔前八十回總字數的百分比。請用表格列出這20個字、其出現次數、和百分比。
- （二）請繼續寫R程式碼來分析另一個文本。這個文本位於C:\Users\lint\Desktop\GPT\Dream of the Red Chamber 120.txt。此文本包含120回，也就是120章，每章的開頭是「第？回」，其中「？」為該回的數目，由「一」排到「一二零」。請將此文本分割成獨立的120個資料框，每個資料框以其回數命名。
- （三）繼續。請就前八十回中出現頻率最高的20個字，算出這20個字的每一個字在全部120回中每回出現次數佔該回總字數的百分比。對這20個字的每一個字建立一個 $T=120$ 的時間序列，畫出每一個序列的時間路徑圖。
- （四）繼續。請將前80回中出現頻率最高的10個字的时间序列加總起來，稱之為top10。再請將前80回中出現頻率最高的20個字的时间序列加總起來，稱之為top20。
- （五）繼續。請將top10和top20這兩個時間序列的時間路徑畫在同一張圖內，並加上標題和圖例。
- （六）繼續。請對top10和top20這兩個時間序列做改變點分析。
- （七）繼續以上的程式碼。我們有120回的資料，每回有20個代表頻率的百分比，你能用這個資料矩陣算出每兩回之間的「餘弦距離」，並畫出「餘弦相似度」矩陣嗎？

《紅樓夢》高頻字回數序列路徑圖

紅樓夢常用字字頻時間序列圖



《紅樓夢》 各回高頻字餘弦相似度矩陣

