機率 & 統計

A Comprehensive Tutorial to Probability & Statistics

盧政良 (Arthur)

預備知識

微積分

線性代數

關於微積分

- 至少掌握微分、積分的概念。

  變化量 累積量

- 還債的機會
  - 台大數學系朱樺老師：

    http://www.math.ntu.edu.tw/~hchu/Calculus/

自行閱讀Chapter 1,2,3,5。

關於線性代數

- 至少掌握向量、內積、聯立方程組、反矩陣的概念。
- 還債的機會
    - 3Blue1Brown:

https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

不確定性

用機率的語言去描述不確定性

沒有完整的資訊

原因？成本太高，甚至根本不可行。

無法預知未來

尤其是金融市場！！！

擲一次硬幣

公正的硬幣：正面與反面出現的機率相同

均匀分配

Uniform Distribution

https://en.wikipedia.org/wiki/Discrete_uniform_distribution

https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)

擲很多次硬幣

二項式分配

Binomial Distribution

https://en.wikipedia.org/wiki/Binomial_distribution

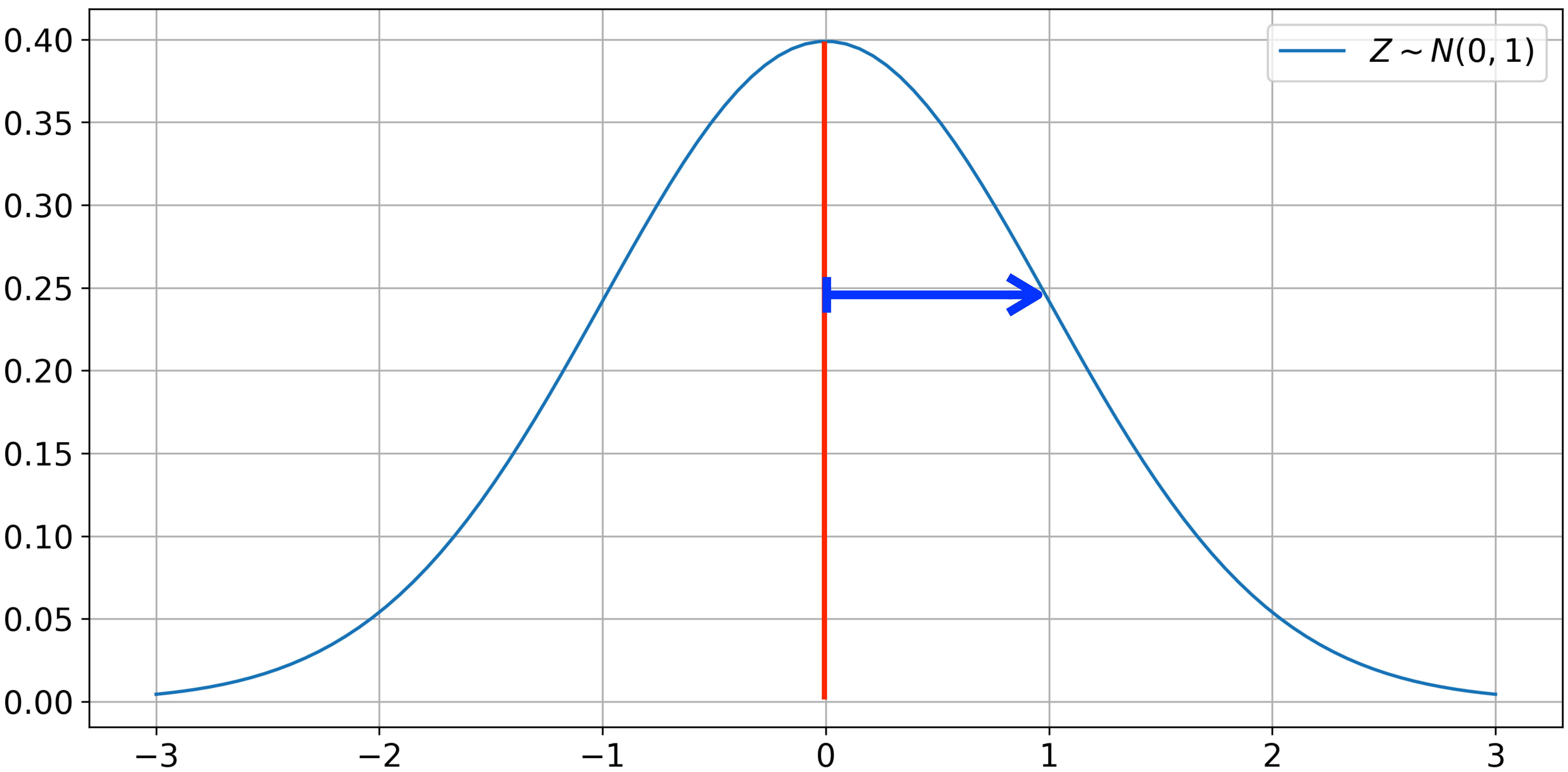
擲了無窮多次硬幣

常態分配

Normal Distribution

https://en.wikipedia.org/wiki/Normal_distribution

又稱高斯分配 (Gauss Distribution)

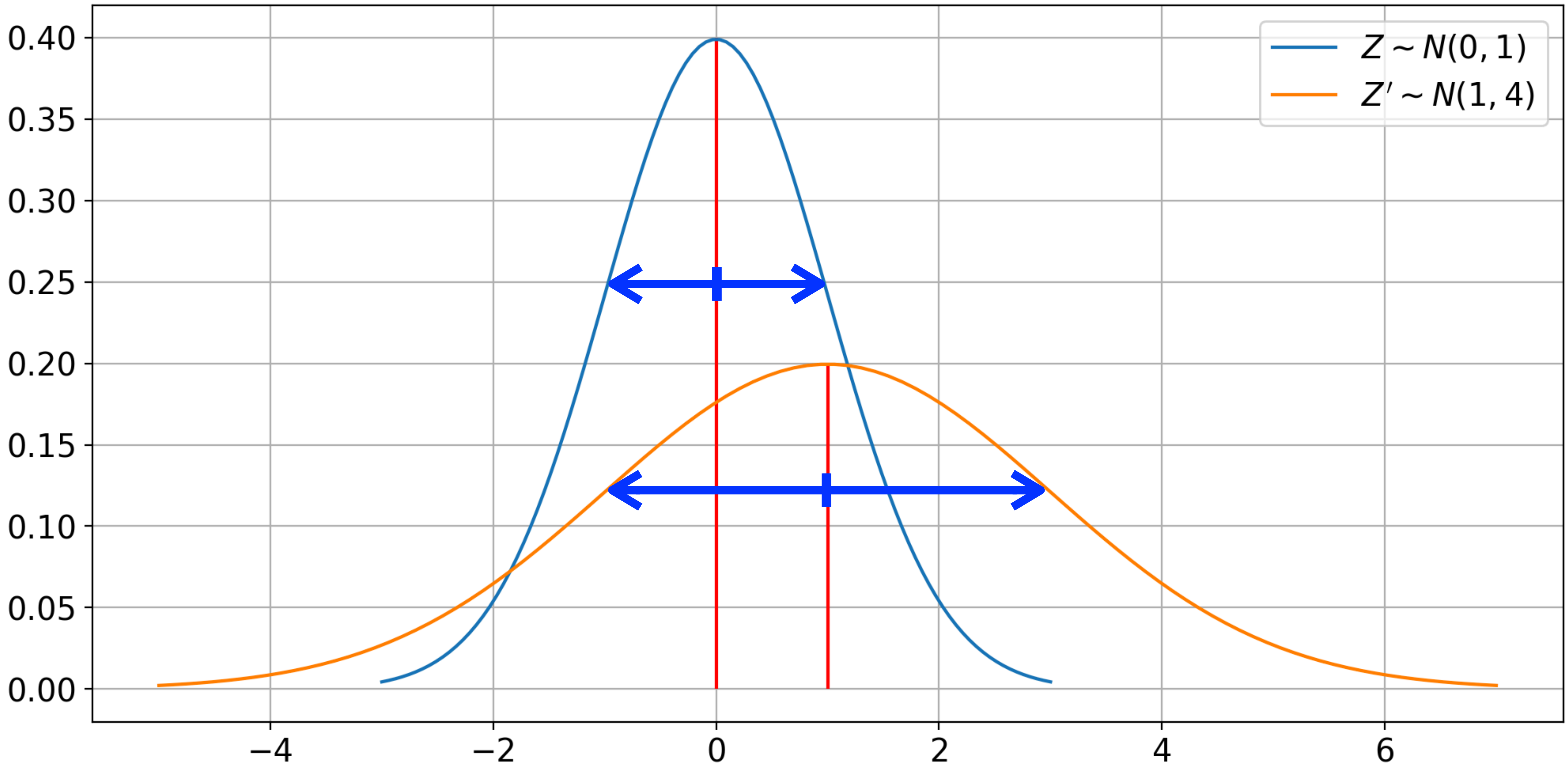


平均數

$$Z \sim N(0, 1)$$

變異數

Z是一個隨機變數，遵守一個標準常態分配。



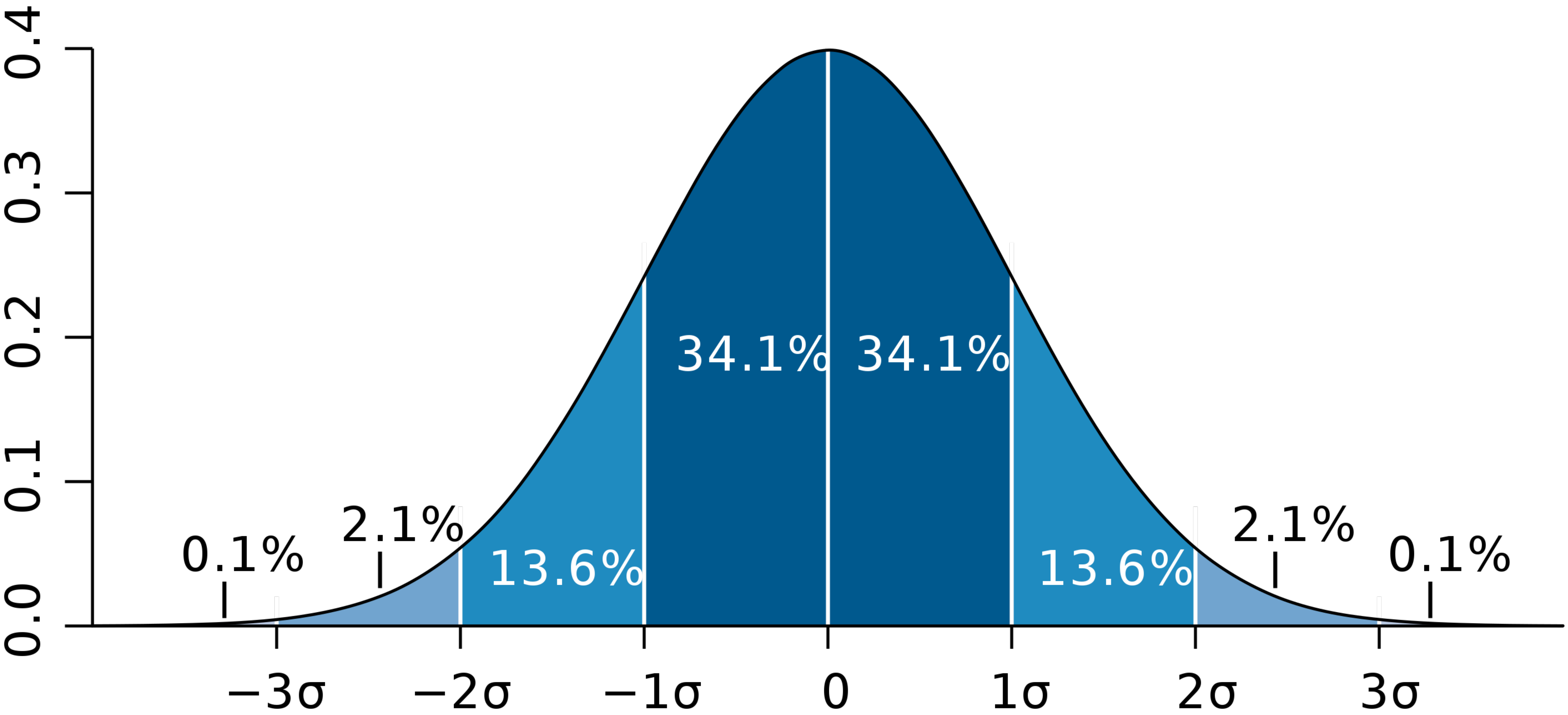
scaling

$$Z' = 1 + 2 \times Z \sim N(1,4)$$

shift

敘述統計

- 中心
  - 算術平均值、幾何平均數、調和平均數、中位數、眾數
- 分散程度
  - 變異數、標準差、四分位差、百分位數、最大值、最小值



https://upload.wikimedia.org/wikipedia/commons/thumb/8/8c/Standard_deviation_diagram.svg/2880px-Standard_deviation_diagram.svg.png

機率論

Probability Theory

https://en.wikipedia.org/wiki/Probability_axioms

符號定義

- 樣本：ω
- 樣本空間：Ω

  $\omega \in \Omega$

- 事件：E
- 事件空間：$\mathscr{F} = 2^{\Omega}$

  事件空間為樣本空間中所有元素的冪集

- 機率測度：$P : \mathscr{F} \rightarrow \mathbb{R}$
- 機率空間：$(\Omega, \mathscr{F}, P)$

擲公正硬幣一次

- 考慮一個公正硬幣，則 $\Omega = \{\text{H, T}\}$。
- 其事件空間 $\mathscr{F} = \{\text{非H且非T, 出現H, 出現T, 出現H或T}\}$
- 所對應的機率測度為：
  - $P(A = \phi) = 0$
    空集合
  - $P(A = H) = P(A = T) = 0.5$
  - $P(A = H \text{ or } T) = 1$

擲公正硬幣兩次

- 考慮一個公正硬幣，則 $\Omega = \{\mathrm{H}, \mathrm{T}\} \times \{\mathrm{H}, \mathrm{T}\}$。
- 其事件空間 $\mathscr{F} = ?$

樹狀圖？

- 所對應的機率測度為：
  - $P(A = (H, H)) = P(A = (T, T)) = 0.25$
  - $P(A = (H, \cdot\,)) = P(A = (T, \cdot\,)) = P(A = (\,\cdot\,, H)) = P(A = (\,\cdot\,, T)) = 0.5$

來當個賭徒吧

假設出現正面獲得一元，出現反面則損失一元。

隨機變數

Random Variable

$$X : \omega \in \Omega \to \mathbb{R}$$

$X \sim$ 某分配

每次擲硬幣的報酬是多少？

每次擲硬幣的期望報酬是多少？

$$\mathbf{E}[\, g(X)\,] = \begin{cases} \sum_{\Omega} g(x) P(x) \\ \int_{\Omega} g(x)\, dF(x) \end{cases}$$

隨機變數的函數

離散版本

連續版本

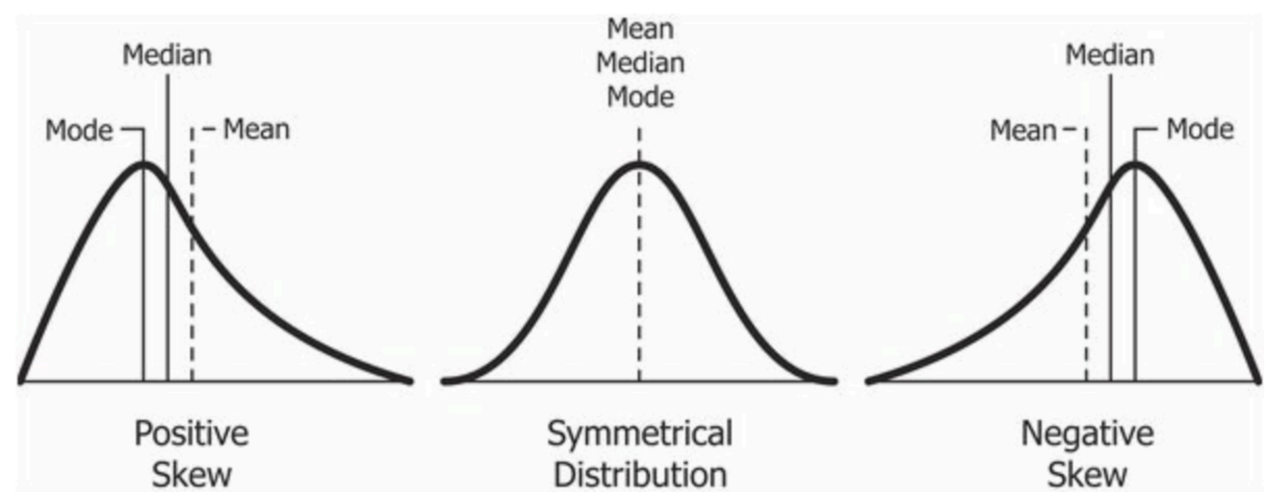
$$\begin{aligned}\mathbf{E}[\,X\,] &= X(\mathsf{H}) \times P(X(\mathsf{H})) + X(\mathsf{T}) \times P(X(\mathsf{T}))\\ &= 1 \times 0.5 + (-1) \times 0.5\\ &= 0\end{aligned}$$

動差

Moment

https://en.wikipedia.org/wiki/Central_moment

- 第一動差：$\mu_X = \mathbf{E}[\, X \,]$
- 第二動差：$\mathbf{Var}(X) = \mathbf{E}\left[(X - \mu_X)^2 \right]$
  - 標準差：$\sigma_X = \sqrt{\mathbf{Var}(X)}$
  - 標準化：$Z = \dfrac{X - \mu_X}{\sigma_X}$
    z score



https://en.wikipedia.org/wiki/Skewness#/media/File:Relationship_between_mean_and_median_under_different_skewness.png

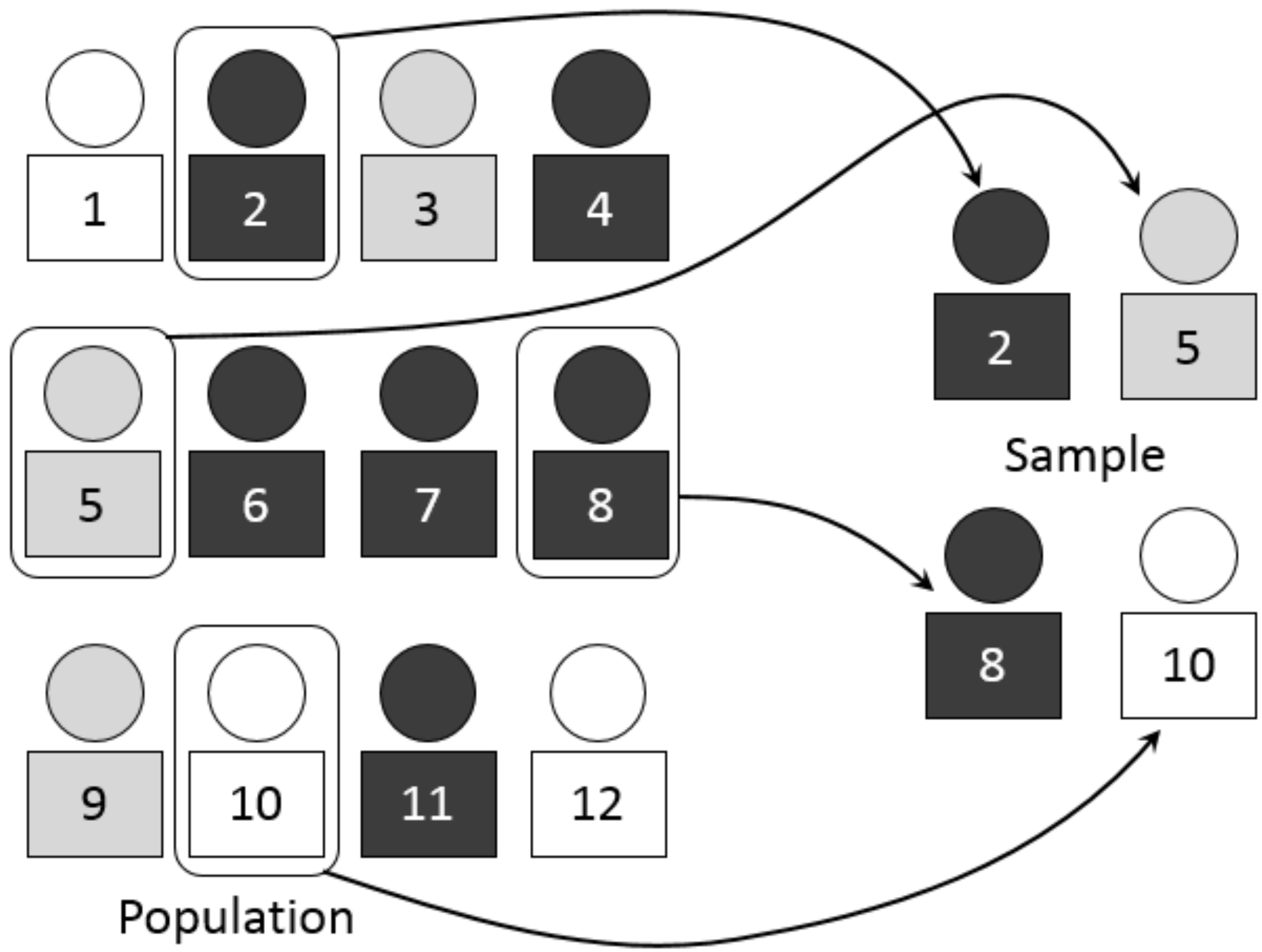
- 第三動差：skewness $= \mathbf{E}[\, Z^3 \,]$
  偏態
  https://en.wikipedia.org/wiki/Skewness
- 第四動差：kurtosis $= \mathbf{E}[\, Z^4 \,]$
  峰態
  https://en.wikipedia.org/wiki/Kurtosis

抽樣

Sampling

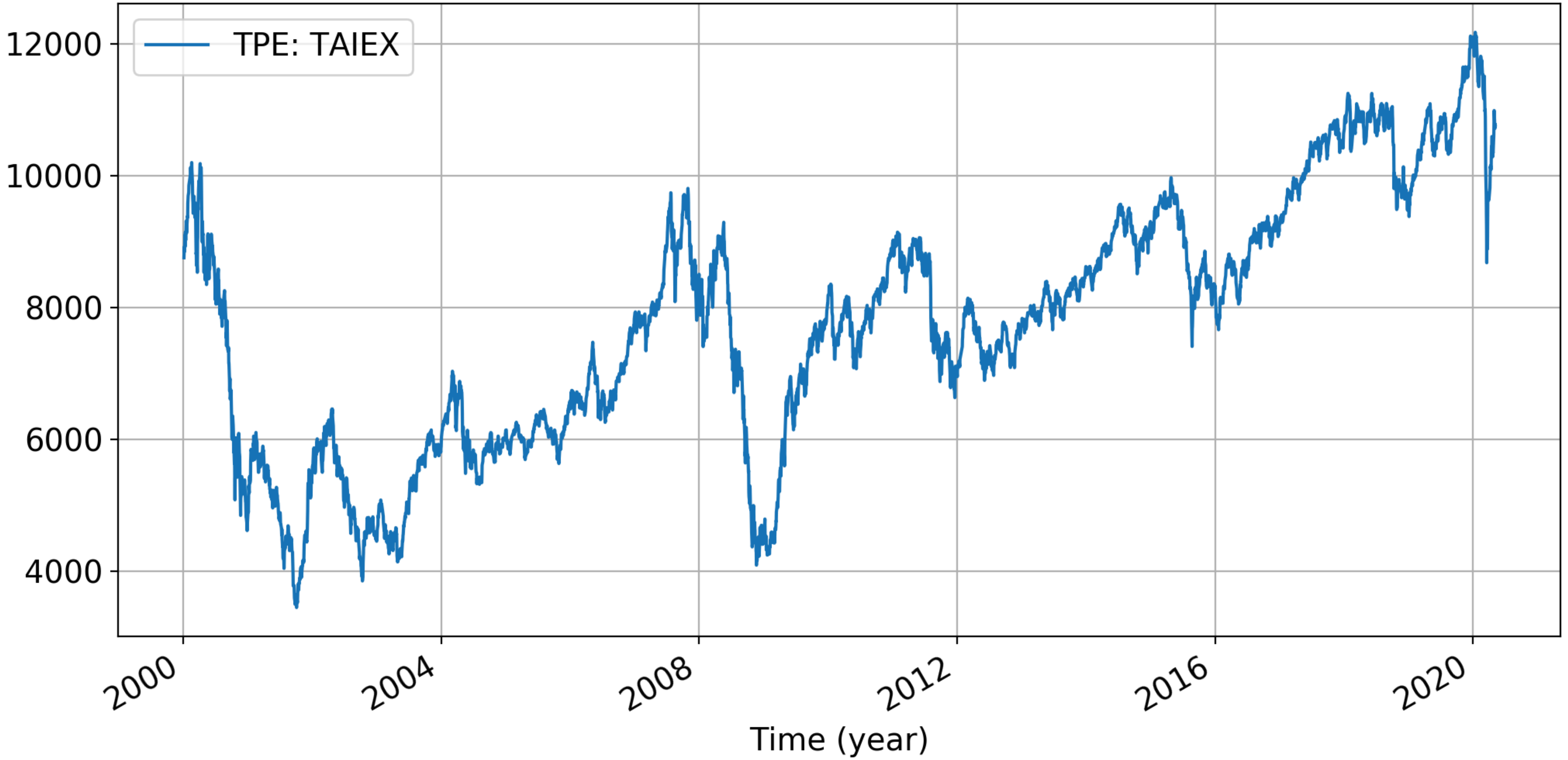
從母體隨機抽取樣本

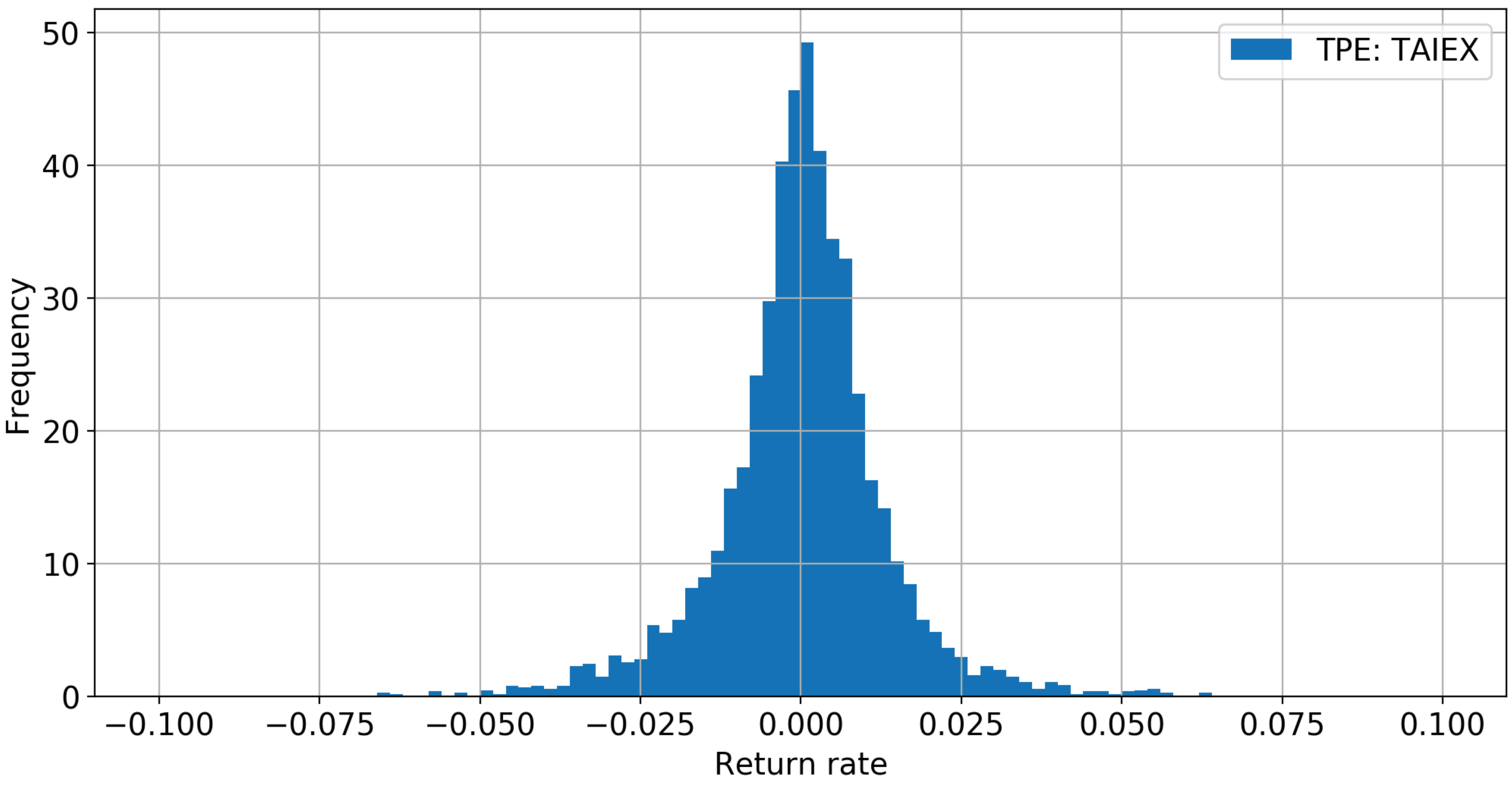
Population Sample



https://zh.wikipedia.org/wiki/%E6%8A%BD%E6%A8%A3#/media/File:Simple_random_sampling.PNG

台灣加權指數歷史收盤價 (未還原)





TPE: TAIEX
Frequency
Return rate
0
10
20
30
40
50
−0.100
−0.075
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100

```
descriptive_stats = twii.calc_stats()
descriptive_stats.display()
```

```
Stat                 twii
-------------------  ----------
Start                2000-01-04
End                  2020-05-06
Risk-free rate       0.00%

Total Return         23.05%
Daily Sharpe         0.16
Daily Sortino        0.25
CAGR                 1.03%
Max Drawdown         -66.22%
Calmar Ratio         0.02

MTD                  -1.98%
3m                   -8.30%
6m                   -7.54%
YTD                  -10.19%
1Y                   -1.12%
3Y (ann.)            2.74%
5Y (ann.)            1.88%
10Y (ann.)           3.58%
Since Incep. (ann.)  1.03%

Daily Sharpe         0.16
Daily Sortino        0.25
Daily Mean (ann.)    3.32%
Daily Vol (ann.)     21.31%
Daily Skew           -0.18
Daily Kurt           3.74
Best Day             6.74%
Worst Day            -9.46%
```

https://en.wikipedia.org/wiki/Sharpe_ratio

https://en.wikipedia.org/wiki/Sortino_ratio

https://en.wikipedia.org/wiki/Compound_annual_growth_rate

https://breakingdownfinance.com/finance-topics/performance-measurement/calmar-ratio/

是否為常態分佈？

檢定

Hypothesis Testing

Normality Test

- Shapiro-Wilk test
  https://en.wikipedia.org/wiki/Shapiro-Wilk_test
- Anderson-Darling test
  https://en.wikipedia.org/wiki/Anderson-Darling_test
- Kolmogorov-Smirnov test
  https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test
- Jarque-Bera test
  https://en.wikipedia.org/wiki/Jarque-Bera_test
- Pearson's χ^2 test
  https://en.wikipedia.org/wiki/Pearson-_chi-squared_test

(´ ༎ຶД༎ຶ `)

虛無假設

Null Hypothesis

H_0: 是常態分佈

對立假設

Alternative Hypothesis

H_1: 不是常態分佈

統計量

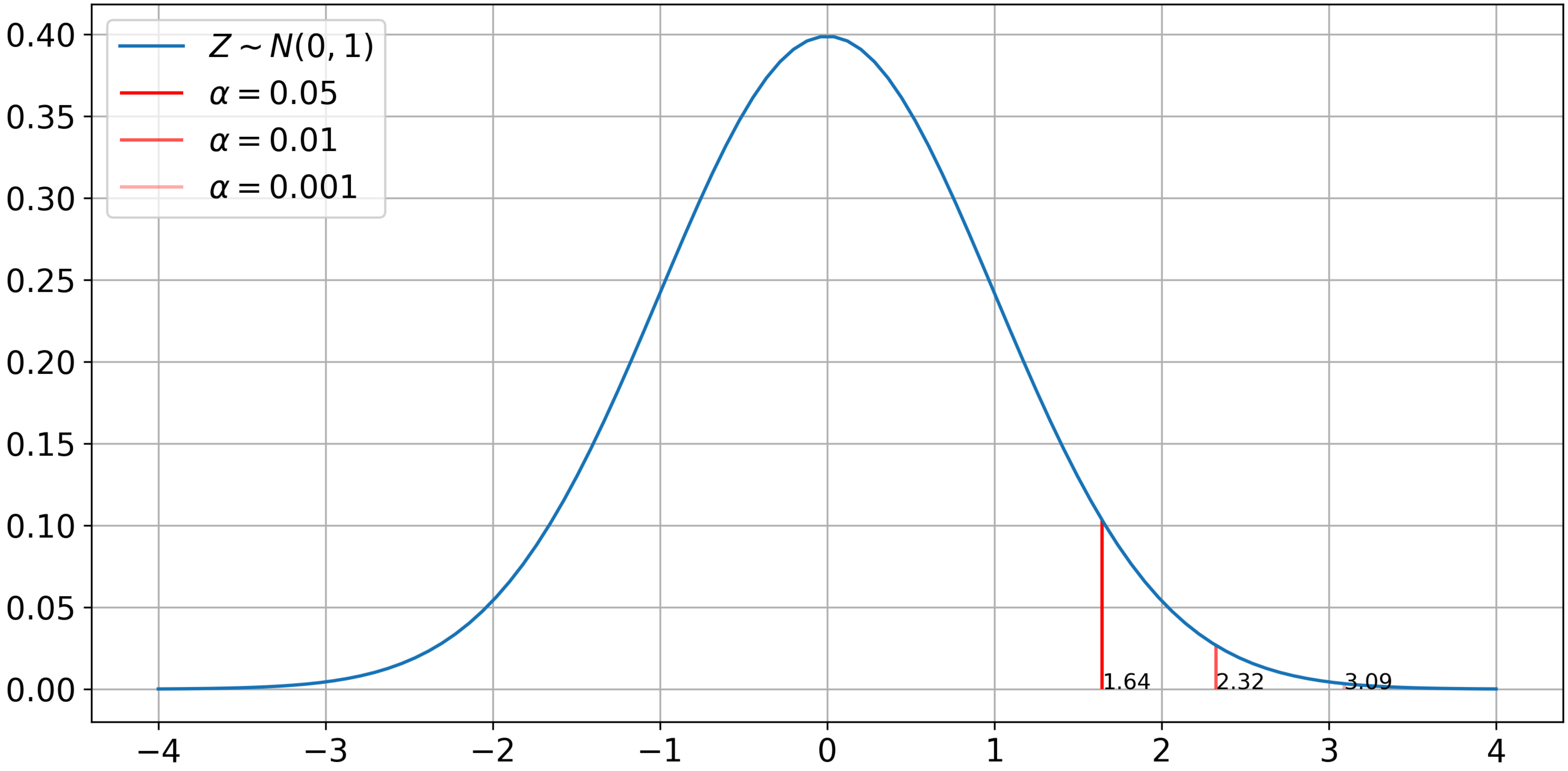
Statistic

顯著性

Statistical Significance

$$\alpha$$

常見的設定是 $\alpha = 0.05$ 或者是 $\alpha = 0.01$



$Z \sim N(0, 1)$
$\alpha = 0.05$
$\alpha = 0.01$
$\alpha = 0.001$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1.64
2.32
3.09
−4
−3
−2
−1
0
1
2
3
4

型一誤差

Type I Error

p-value

當p-value < α 時，

此統計量有顯著差異，拒絕 H_0！

當p-value >= α 時，

此統計量無顯著差異，**無法拒絕** H_0！

也就是說，我找不到足夠的證據來證明 H_0 不成立！

一開始可能會誤以為，既然沒有拒絕，代表 H_0 是對的。
但是這是錯誤的解讀。

"... is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

– Ronald Fisher (1890–1962)

型二誤差

Type II Error

https://en.wikipedia.org/wiki/Type_III_error

β

Table of error types		**Null hypothesis (H_0) is**	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)

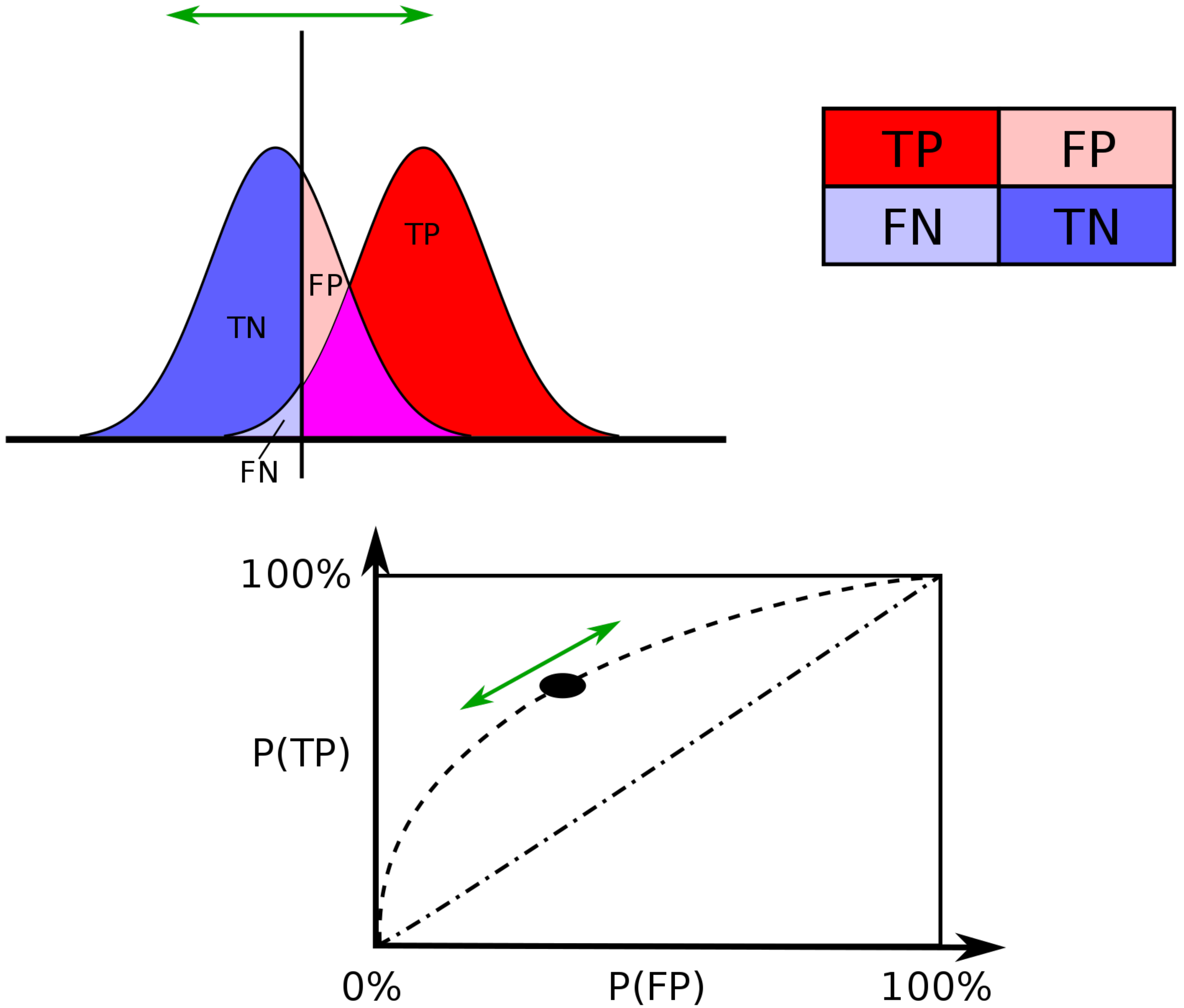
Confusion Matrix

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

ROC曲線

Receiver Operating Characteristic

https://en.wikipedia.org/wiki/Receiver_operating_characteristic



https://upload.wikimedia.org/wikipedia/commons/thumb/4/4f/ROC_curves.svg/1920px-ROC_curves.svg.png

型三誤差

Type III Error

https://en.wikipedia.org/wiki/Type_III_error

"correctly rejecting the null hypothesis for the wrong reason"

Bayes

https://en.wikipedia.org/wiki/Bayes'_theorem

先驗機率

Prior Probability, Belief

條件機率

Conditional Probability

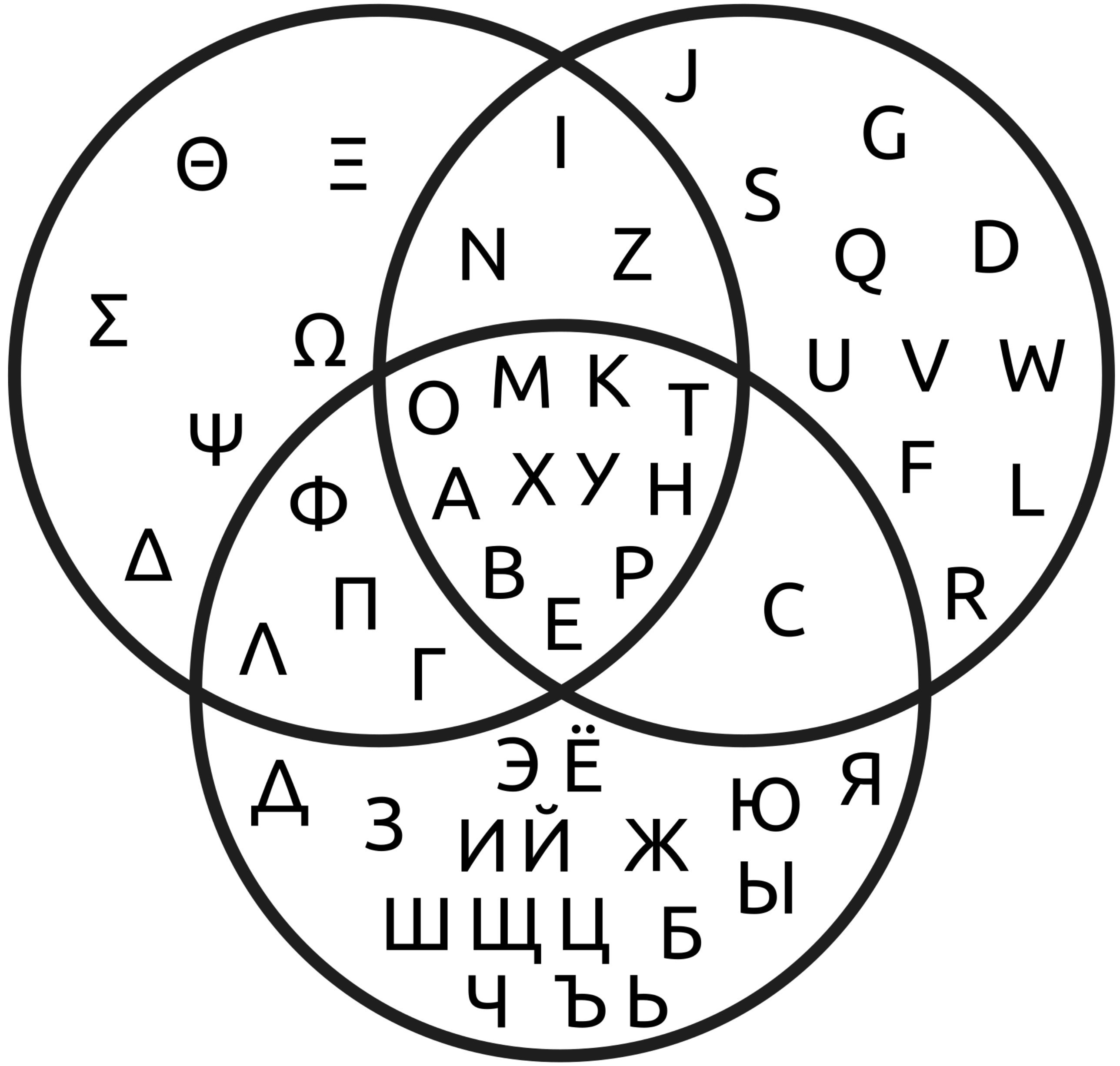
交集

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

在事件B的條件下，發生事件A的機率

文氏圖

https://en.wikipedia.org/wiki/Venn_diagram



https://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Venn_diagram_gr_la_ru.svg/1920px-Venn_diagram_gr_la_ru.svg.png

$$
\begin{aligned}
P(A \cap B) &= P(A \mid B)\, P(B) \\
&= P(B \mid A)\, P(A)
\end{aligned}
$$

Symptom	Cancer		Total
	No	Yes	
No	99989	0	99989
Yes	10	1	11
Total	99999	1	100000

Which can then be used to calculate the probability of having cancer when you have the symptoms:

$$
\begin{aligned}
P(\text{Cancer}|\text{Symptoms}) &= \frac{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer})}{P(Symptoms)} \\
&= \frac{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer})}{P(\text{Symptoms}|\text{Cancer})P(\text{Cancer}) + P(\text{Symptoms}|\text{Non-Cancer})P(\text{Non-Cancer})} \\
&= \frac{1 \times 0.00001}{1 \times 0.00001 + (10/99999) \times 0.99999} \\
&= \frac{1}{11} \\
&\approx 9.1\%
\end{aligned}
$$

https://en.wikipedia.org/wiki/Bayes%27_theorem#A_more_complicated_example

獨立事件

Independence

$$P(A \cap B) = P(A)\,P(B)$$

很重要，但是很難辦到！

案例：投籃命中率

案例：上漲下跌

題外話們

凱利賭徒

https://en.wikipedia.org/wiki/Kelly_criterion

平賭

Martingale

https://en.wikipedia.org/wiki/Martingale_(probability_theory)

https://rich01.com/blog-pos-22/

反平賭