廣義線性模型 (HDAS7009-02)

平均值的抽樣分布 (Sampling distributions of the Mean)

Tsai, Dai-Rong

Table of contents

1 中央極限定理 (Central Limit Theorem)

從平均值為 μ，標準差為 σ 的母體中，隨機地抽取大小為 n 的獨立樣本。當樣本數 n 夠大時，其樣本平均值 $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ 減掉平均值 μ 再除以標準差 $\frac{\sigma}{\sqrt{n}}$，將會趨近平均值為 0，標準差為 1 的常態分佈 (normal distribution)。

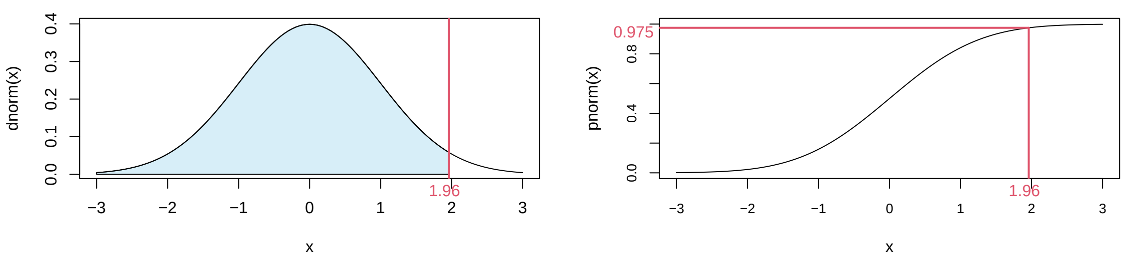
$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{n\to\infty} N(0,1)$$

2 Probability distributions in R

Distributions in the `{stats}` package

$$
\begin{pmatrix}
\texttt{d} & \text{(probability density)} \\
\texttt{p} & \text{(cumulative distribution)} \\
\texttt{q} & \text{(quantile function)} \\
\texttt{r} & \text{(random sample)}
\end{pmatrix}
\otimes
\begin{pmatrix}
\texttt{unif} & \text{(Uniform)} \\
\texttt{binom} & \text{(Binomial)} \\
\texttt{pois} & \text{(Poisson)} \\
\texttt{exp} & \text{(Exponential)} \\
\texttt{norm} & \text{(Normal)} \\
\texttt{t} & \text{(Student's t)} \\
\texttt{chisq} & \text{(Chi-squared)} \\
\vdots &
\end{pmatrix}
$$

See `?distribution` for more distributions.




```r
pnorm(1.96)
```

```
[1] 0.9750021
```

```r
qnorm(0.975)
```

```
[1] 1.959964
```

3 模擬隨機抽樣

模擬從不同統計分布進行 1000 次抽樣，樣本數 (sample size) 分別設定 5、10、30，計算各組抽樣結果的平均值繪製其分布圖，並與中央極限定理的理論分布做比較。

3.1 離散均勻分布 (Discrete uniform distribution)

模擬 1000 組投擲一顆公正骰子 5、10、30 次所出現的點數。先以 5 次為例：

1. 生成資料模擬投擲骰子的結果

```r
set.seed(2025)

n <- 5    # 大小為 n 的獨立樣本
m <- 1000 # 重複抽取 m 次

sample_unif <- matrix(sample(1:6, size = m * n, replace = TRUE),
                      nrow = m, ncol = n)

rownames(sample_unif) <- paste("sample", 1:m, sep = "")
colnames(sample_unif) <- paste("obs", 1:n, sep = "")

dim(sample_unif)
```

```
[1] 1000    5
```

```r
head(sample_unif)
```

```
        obs1 obs2 obs3 obs4 obs5
sample1    5    2    3    3    5
sample2    4    1    2    3    5
sample3    4    3    1    3    5
sample4    2    6    4    4    5
sample5    1    2    4    4    1
sample6    6    1    5    5    6
```

2. 計算平均值

```r
sample_unif_mean <- rowMeans(sample_unif)
head(sample_unif_mean)
```

```
sample1 sample2 sample3 sample4 sample5 sample6
    3.6     3.0     3.2     4.2     2.4     4.6
```

```r
# Equivalent:
sample_unif_mean <- apply(sample_unif, 1, mean) # 1: by row; 2: by column
head(sample_unif_mean)
```

```
sample1 sample2 sample3 sample4 sample5 sample6
    3.6     3.0     3.2     4.2     2.4     4.6
```

Note

若要計算每一列的 mean，rowMeans() 會比 apply(x, 1, mean) 更有效率。然而 apply 功能性更廣泛，它可以對每一列或每一行計算更複雜的統計量。例如：

```r
# Standard deviation
head(apply(sample_unif, 1, sd))
```

```
 sample1  sample2  sample3  sample4  sample5  sample6
1.341641 1.581139 1.483240 1.483240 1.516575 2.073644
```

```r
# 25th percentile
head(apply(sample_unif, 1, \(x) quantile(x, 0.25)))
```

```
sample1 sample2 sample3 sample4 sample5 sample6
      3       2       3       4       1       5
```

```r
# Mean absolute deviation
head(apply(sample_unif, 1, \(x) mean(abs(x - mean(x)))))
```

```
sample1 sample2 sample3 sample4 sample5 sample6
   1.12    1.20    1.04    1.04    1.28    1.44
```

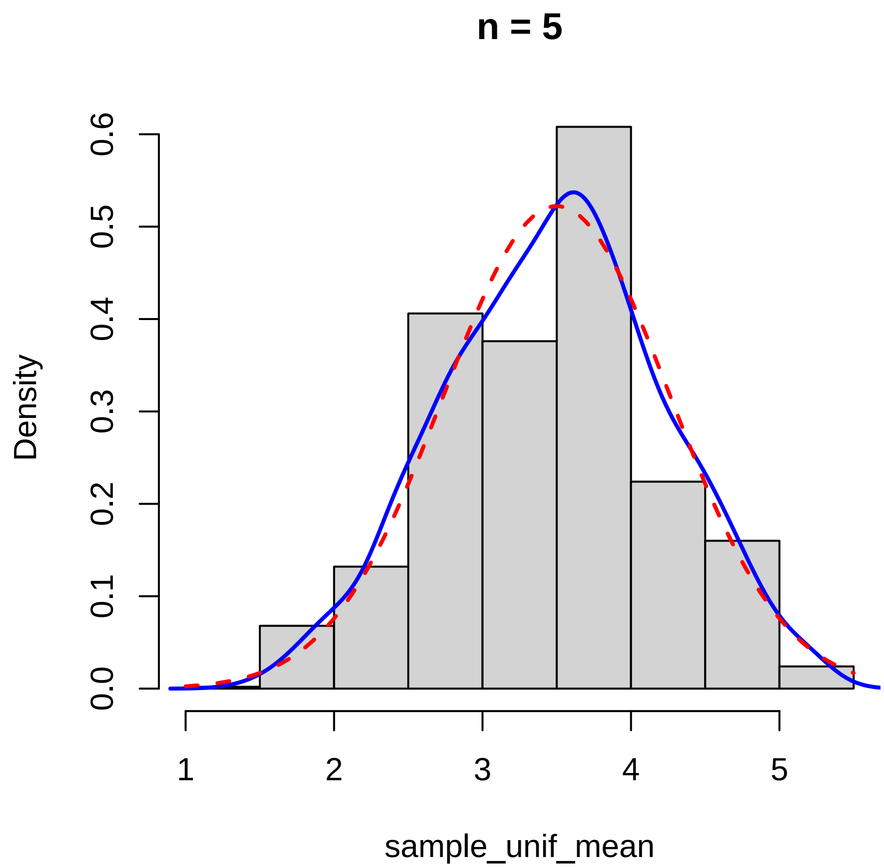
3. 繪製分布圖並與中央極限定理的理論分布做比較

   我們先來計算中央極限定理的理論分布所需要的參數。一顆公正骰子出現的點數服從最小值 1 最大值 6 的離散均勻分布，其平均值與變異數分別是：

$$\text{mean} = \frac{1+6}{2} = 3.5, \quad \text{variance} = \frac{6^2-1}{12} \approx 2.917$$

   因此中央極限定理的理論分布為 $Normal(\mu = 3.5, \sigma = \sqrt{\frac{2.917}{n}})$

```r
hist(sample_unif_mean, prob = TRUE, main = paste("n =", n))
lines(density(sample_unif_mean), col = "blue", lwd = 2)
curve(dnorm(x, mean = 3.5, sd = sqrt(2.917 / n)), add = TRUE,
      col = "red", lwd = 2, lty = 2)
```

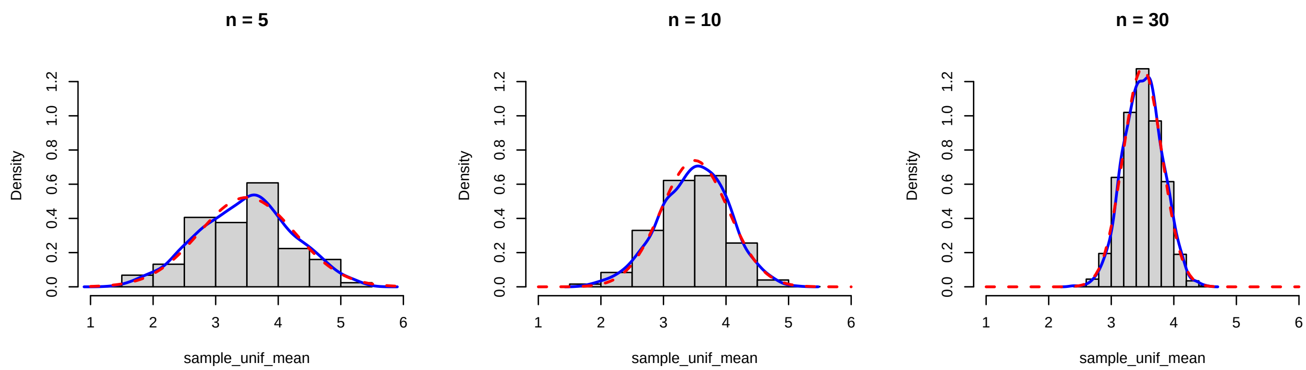


4. 比較不同樣本數 (5、30、100) 下的抽樣分布

```
par(mfrow = c(1, 3))

for(n in c(5, 10, 30)) {
  sample_unif <- matrix(sample(1:6, size = m * n, replace = TRUE),
                        nrow = m, ncol = n)
  sample_unif_mean <- rowMeans(sample_unif)
  hist(sample_unif_mean, prob = TRUE, main = paste("n =", n),
       xlim = c(1, 6), ylim = c(0, 1.3))
  lines(density(sample_unif_mean), col = "blue", lwd = 2)
  curve(dnorm(x, mean = 3.5, sd = sqrt(2.917 / n)), add = TRUE,
```

```
        col = "red", lwd = 2, lty = 2)
}
```



3.2 二項式分布 (Binomial distribution)

模擬 1000 組同時投擲 10 枚正面機率為 0.1 的不公正銅板 5、10、30 次，紀錄每次出現正面的次數。

```
N <- 10
p <- 0.1
```

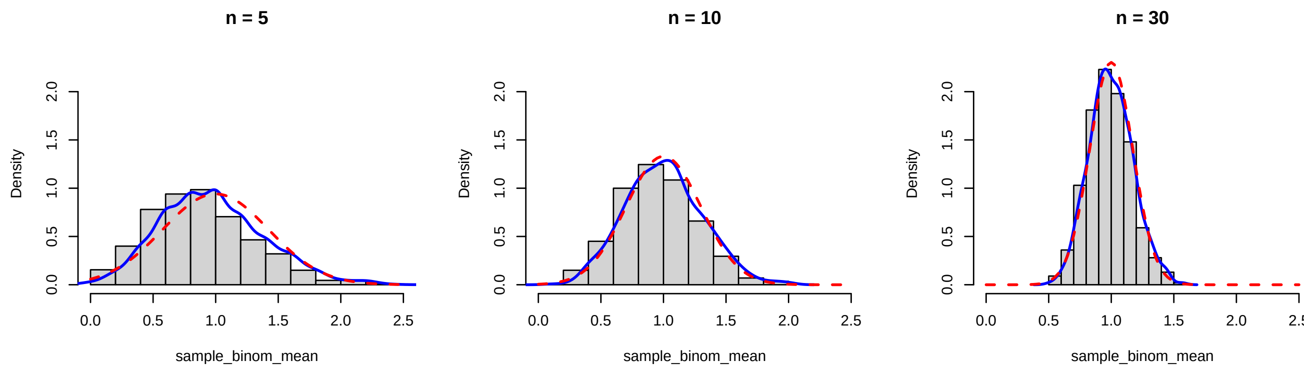
同時投擲 10 枚正面機率為 0.1 的不公正銅板，出現正面的次數服從 Binomial($N = 10, p = 0.1$)，其平均值與變異數分別是：

$$\text{mean} = Np, \quad \text{variance} = Np(1-p)$$

```
par(mfrow = c(1, 3))

for(n in c(5, 10, 30)) {
  sample_binom <- matrix(rbinom(m * n, size = N, prob = p), # rbinom(n, size, prob)
                         nrow = m, ncol = n)
  sample_binom_mean <- rowMeans(sample_binom)
  hist(sample_binom_mean, prob = TRUE, main = paste("n =", n),
       xlim = c(0, 2.5), ylim = c(0, 2.3))
  lines(density(sample_binom_mean), col = "blue", lwd = 2)
```

```
  curve(dnorm(x, mean = N*p, sd = sqrt(N*p*(1-p) / n)), add = TRUE,
        col = "red", lwd = 2, lty = 2)
}
```



3.3 卜瓦松分布 (Poisson distribution)

某路段一個月平均發生 2 次車禍，模擬 1000 組 5、10、30 個月內每個月的車禍發生次數。

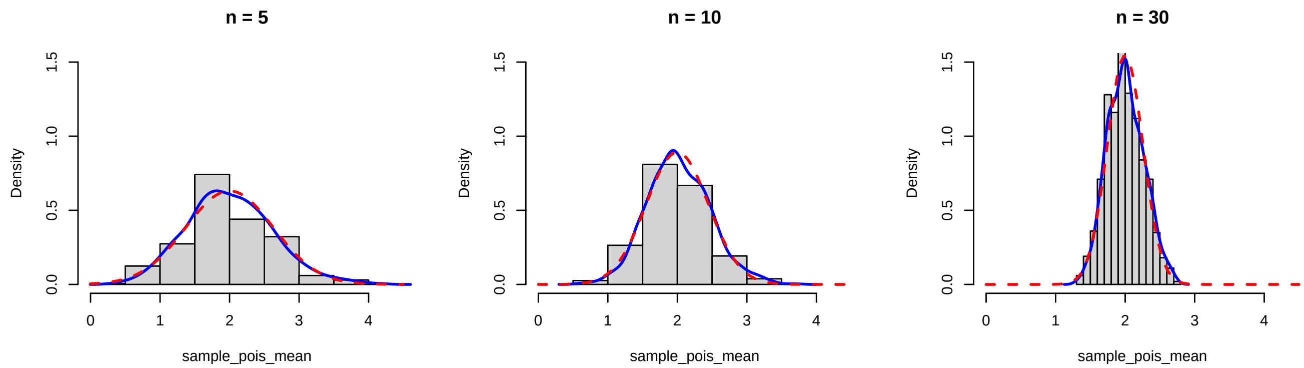
```
lambda <- 2
```

某路段一個月平均發生 2 次車禍，每個月的車禍發生次數服從 $\text{Poisson}(\lambda = 2)$，其平均值與變異數分別是：

$$\text{mean} = \text{variance} = \lambda$$

```
par(mfrow = c(1, 3))

for(n in c(5, 10, 30)) {
  sample_pois <- matrix(rpois(m * n, lambda = lambda), # rpois(n, lambda)
                        nrow = m, ncol = n)
  sample_pois_mean <- rowMeans(sample_pois)
  hist(sample_pois_mean, prob = TRUE, main = paste("n =", n),
       xlim = c(0, 4.5), ylim = c(0, 1.5))
  lines(density(sample_pois_mean), col = "blue", lwd = 2)
  curve(dnorm(x, mean = lambda, sd = sqrt(lambda / n)), add = TRUE,
        col = "red", lwd = 2, lty = 2)
}
```



Exercise (1)

若大學畢業生的平均月薪資是 \$25,000，標準差是 \$5,000。

1. 若隨機選取 100 名畢業生，這些人的平均薪資超過 \$26,000 的機會有多大?

$$\mu = 25000,\ \sigma = 5000$$

所以 $\bar{X}$ 的期望值是 \$25,000，而 $\bar{X}$ 的抽樣分布的標準差，也就是標準誤 (standard error) 為

$$se = \frac{\sigma}{\sqrt{n}} = \frac{5000}{\sqrt{100}} = 500$$

根據中央極限定理，$\bar{X} \sim \text{Normal}(25000, 500)$，所以

$$P(\bar{X} > 26000) = P(Z > \frac{26000 - 25000}{500}) = P(Z > 2)$$

表示 $\bar{X} = 26,000$ 比期望值 25,000 超過 2 個標準差。以下指令都可以求出 $Z > 2$ 的機率：

```
1 - pnorm(26000, mean = 25000, sd = 500)
```

```
[1] 0.02275013
```

```
pnorm(26000, mean = 25000, sd = 500, lower.tail = FALSE)
```

```
[1] 0.02275013
```

```r
1 - pnorm(2) # default: mean = 0, sd = 1
```

```
[1] 0.02275013
```

```r
pnorm(2, lower.tail = FALSE)
```

```
[1] 0.02275013
```

2. 若隨機選取 100 名畢業生，這些人的平均薪資在 $24,000 到 $26,000 之間的機會有多大?

$$
\begin{aligned}
P(24000 < \bar{X} < 26000) &= P(\frac{24000-25000}{500} < Z < \frac{26000-25000}{500}) \\
&= P(-2 < Z < 2) = P(Z < 2) - P(Z < -2)
\end{aligned}
$$

```r
pnorm(2) - pnorm(-2)
```

```
[1] 0.9544997
```

3. 若大學畢業生的月薪資服從常態分布，平均值是 $25,000，標準差未知。若隨機選取 49 名畢業生，計算樣本標準差為 $s = 3500$，這些人的平均薪資超過 $26,000 的機會有多大? 平均薪資在 $24,000 到 $26,000 之間的機會有多大?

$$
\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\bar{X}-25000}{\frac{3500}{\sqrt{49}}} = \frac{\bar{X}-25000}{500} \sim t_{(df=49-1)}
$$

$$
P(\bar{X} > 26000) = P(T > \frac{26000-25000}{500}) = P(T > 2)
$$

```r
1 - pt(2, df = 49-1)
```

```
[1] 0.02558797
```

$$
\begin{aligned}
P(24000 < \bar{X} < 26000) &= P(\frac{24000-25000}{500} < T < \frac{26000-25000}{500}) \\
&= P(-2 < T < 2)
\end{aligned}
$$

```r
pt(2, df = 49-1) - pt(-2, df = 49-1)
```

```
[1] 0.9488241
```

Exercise (2)

1991 至 1995 年英國 Bristol 的兩位心臟外科醫師，共對 181 位 5 歲以下患有先天性心臟病的小孩作矯正手術，結果有 43 個小孩死亡。英國同時期的全國平均手術死亡率是 12%。

假設 Birstol 的兩位心臟外科醫師的技術和英國其他心臟外科醫師沒有不同，那麼在 181 次手術發生 43 次或更多次死亡的機率是多少?

```r
1 - pbinom(42, 181, 0.12)
```

```
[1] 8.296321e-06
```

```r
pbinom(42, 181, 0.12, lower.tail = FALSE)
```

```
[1] 8.296321e-06
```

```r
sum(dbinom(43:181, 181, 0.12))
```

```
[1] 8.296321e-06
```

4 信賴區間 (Confidence Interval)

- 利用樣本觀察值計算一個母體參數區間的上界與下界，稱為信賴界限 (confidence limits)，使得在重複抽取樣本時，未知參數落在計算的信賴界限的比例達到需要的準確度，稱為信賴水準 (level of confidence)。
- 一個 95% 信賴區間 (confidence interval) 意味著，如果我們利用 100 個來自相同的母群體的不同樣本，計算 100 個信賴區間，我們可以期待在這 100 個區間裡，95 個會包含母群體平均值。
- Constructing a $(1-\alpha)100\%$ confidence interval about μ
  - when σ is known

$$\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

  - when σ is unknown

$$\bar{X} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Exercise (3)

為了想知道台灣高中男生的平均身高，我們蒐集了 12 位某市立高中男生的身高體重。請計算台灣高中男生的平均身高的 95% 信賴區間。

```
dat <- readxl::read_excel("data/basketball_team.xls", sheet = "sheet 1")
```

Preview

```
head(dat)
```

```
# A tibble: 6 x 4
     No Grade Height Weight
  <dbl> <dbl>  <dbl>  <dbl>
1     4     1    187     75
2     5     3    182     70
3     6     2    180     70
4     7     1    175     60
5     8     2    181     71
6     9     2    193     87
```

Data Structure

```
str(dat)
```

```
tibble [12 x 4] (S3: tbl_df/tbl/data.frame)
 $ No    : num [1:12] 4 5 6 7 8 9 10 11 12 13 ...
 $ Grade : num [1:12] 1 3 2 1 2 2 3 3 2 1 ...
 $ Height: num [1:12] 187 182 180 175 181 193 187 186 197 193 ...
 $ Weight: num [1:12] 75 70 70 60 71 87 81 73 77 71 ...
```

1. 計算身高平均值 $(\bar{X})$ 與樣本標準差 (s)

   ```
   (xbar <- mean(dat$Height))
   ```

   ```
   [1] 187.9167
   ```

```
(s <- sd(dat$Height))
```

```
[1] 8.061788
```

2. 由於母體標準差 σ 未知，信賴區間需要利用 t 分布推估，自由度為

$$\text{degree of freedom} = n - 1 = 12 - 1 = 11$$

3. 95% 信賴區間，$\alpha = 0.05$。若無特別註明，信賴區間一般都是指”雙尾”信賴區間，故 t 分布累計機率左右各佔 $0.05/2 = 0.025$。此時 $t_{\alpha/2} =$

```
qt(1 - 0.05/2, df = 12-1)
```

```
[1] 2.200985
```

4. 95% 信賴區間

   - lower limit

```
xbar - qt(1-0.05/2, 12-1) * s/sqrt(12)
```

```
[1] 182.7945
```

   - upper limit

```
xbar + qt(1-0.05/2, 12-1) * s/sqrt(12)
```

```
[1] 193.0389
```

Note

我們也可以用 {stats} 套件的 t.test() 函數一次計算 平均值 ($estimate)、標準誤 ($stderr)、信賴區間 ($conf.int) 等統計量。

```
test_ht <- t.test(dat$Height, conf.level = 0.95)
test_ht
```

```

	One Sample t-test

data:  dat$Height
t = 80.747, df = 11, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 182.7945 193.0389
sample estimates:
mean of x
 187.9167
```

```r
test_ht$estimate
```

```
mean of x
 187.9167
```

```r
test_ht$stderr
```

```
[1] 2.327238
```

```r
test_ht$conf.int
```

```
[1] 182.7945 193.0389
attr(,"conf.level")
[1] 0.95
```