資料科學方法論- HW1

魏上傑

2023-03-09

1. Independently toss a coin n times. Let p = P(head). Model the number of heads by a Binomial distribution.

(a) What is the probability of obtaining x heads, $0 \le x \le n$?

$$\binom{n}{x} p^x (1-p)^{n-x}$$

(b) Suppose n is even. What is the probability of obtaining n/2 heads?

$$\binom{n}{n/2} p^{n/2} (1-p)^{n/2}$$

(c) Using The Stirling formula to approximate.

$$\begin{aligned}
&\binom{n}{n/2} p^{n/2} (1-p)^{n/2} \\
&\approx \frac{n!}{(\frac{n}{2})!(\frac{n}{2})!} p^{n/2} (1-p)^{n/2} \\
&\approx \frac{\sqrt{2n\pi}(n/e)^n}{\sqrt{n\pi}(n/2e)^{n/2}\sqrt{n\pi}(n/2e)^{n/2}} p^{n/2} (1-p)^{n/2} \\
&\approx \frac{\sqrt{2}(n/e)^n}{\sqrt{n\pi}(n/2e)^n} p^{n/2} (1-p)^{n/2} \\
&\approx \frac{2^{n+\frac{1}{2}}}{\sqrt{n\pi}} p^{n/2} (1-p)^{n/2}
\end{aligned}$$

(d)

```r
approxi <- function(n, p){
  return  ((2^(n+0.5))*p^(n/2)*(1-p)^(n/2)/sqrt(n*pi))
}

exact <- function(n,p){
  return (choose(n, n/2)*p^(n/2)*(1-p)^(n/2))
}

n <- c(20,100,500,1000)
p <- 0.5
for (i in 1:length(n)){
  approx_prob <- approxi(n[i], p)
  exact_prob <- exact(n[i],p)
  cat("n=",n[i], "approximate probability=",approx_prob,
      "exact probability= ", exact_prob,"\n")
}
```

```
## n= 20 approximate probability= 0.1784124 exact probability=  0.1761971
## n= 100 approximate probability= 0.07978846 exact probability=  0.07958924
## n= 500 approximate probability= 0.03568248 exact probability=  0.03566465
## n= 1000 approximate probability= 0.02523133 exact probability=  0.02522502
```

n	20	100	500	1000
Approx.	0.1784124	0.07978846	0.03568248	0.02523133
Exact	0.1761971	0.07958924	0.03566465	0.02522502

從表格可以看出利用 Stirling formula 逼近的結果會些微高估實際的值。

(e)

```r
approxi <- function(n, p){
  return  ((2^(n+0.5))*p^(n/2)*(1-p)^(n/2)/sqrt(n*pi))
}

exact <- function(n,p){
  return (choose(n, n/2)*p^(n/2)*(1-p)^(n/2))
}
```

```r
n <- c(20,100,500,1000)
p <- 1/3
for (i in 1:length(n)){
  approx_prob <- approxi(n[i], p)
  exact_prob <- exact(n[i],p)
  cat("n=",n[i], "approximate probability=",approx_prob,
      "exact probability= ", exact_prob,"\n")
}
```

```
## n= 20 approximate probability= 0.05494141 exact probability=  0.0542592
## n= 100 approximate probability= 0.0002209601 exact probability=  0.0002204084
## n= 500 approximate probability= 5.811985e-15 exact probability=  5.809079e-15
## n= 1000 approximate probability= 6.693894e-28 exact probability=  6.692221e-28
```

n	20	100	500	1000
Approx.	0.05494141	0.0002209601	5.811985e-15	6.693894e-28
Exact	0.0542592	0.0002204084	5.809079e-15	6.692221e-28

從表格可以看出利用 Stirling formula 逼近的結果會些微高估實際的值。

(f) Conclude from (d) and (e)

可以看出利用 Stirling formula 逼近的結果會些微高估實際的值。另外，p=1/2 的計算結果明顯高於 p=1/3 的計算結果。

2. Independently toss a coin 100 times.

(a) If 50 heads appear, is this coin fair?

```r
n <- 100 #large sample
p <- 0.5
p_hat <- 50/n
zstat <- (p_hat - p) / sqrt(p*(1-p)/n)
p_value <- 2*pnorm(-abs(zstat)) # pnorm calculate the cdf of normal
p_value
```

```
## [1] 1
```

$H_0 : p = \frac{1}{2}$

The null hypothesis is that the coin is fair, since the p-value is large, we do not reject H0, and thus this coin is fair.

(b) If 5 heads appear, is this coin fair? Hint: use CLT to conduct hypothesis testings

```
n <- 100
p <- 0.5
p_hat <- 5/n
zstat <- (p_hat - p) / sqrt(p*(1-p)/n)
p_value <- 2*pnorm(-abs(zstat))
p_value
```

```
## [1] 2.257177e-19
```

The null hypothesis is that the coin is fair, since the p-value is small, we reject H0, and thus this coin is not fair.

3.

Find a sequence of p_i such that $\sum p_i(1-p_i) \to \infty$ Hint: check $p_i = i^{-1}, p_i = i^{-2}$

```
# 定義兩個序列
p1 <- (1:(10^5)) ^ -1
p2 <- (1:(10^5)) ^ -2

# 計算 p_i(1-p_i)
s1 <- sum(p1 * (1 - p1))
s2 <- sum(p2 * (1 - p2))

# 印出結果
cat("For p_i = i^-1, the sum p_i(1-p_i) is:", s1, "\n")
```

```
## For p_i = i^-1, the sum p_i(1-p_i) is: 10.44522
```

```
cat("For p_i = i^-2, the sum p_i(1-p_i) is:", s2, "\n")
```

```
## For p_i = i^-2, the sum p_i(1-p_i) is: 0.5626008
```

According to the results above, it seems that $p_i = i^{-1}$ will have $\sum p_i(1-p_i) \to \infty$, but not for $p_i = i^{-2}$

4.

樣本平均值的分佈趨近於常態分佈的條件是:

- 母體分佈是有限期望值和有限變異數的分佈。
- 樣本數目足夠大，通常是樣本數目大於 30。

然而，standard Cauchy 的期望值和變異數不存在，inverse chi-square 的期望值 (ν>2) 和變異數 (ν>4) 則有其定義範圍，另外，inverse chi-square 變異數只與自由度有關，不會隨著樣本數目的增加而變小，因此它們違反了 CLT 的假設。

5.

```
data(cars)


str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```
tail(cars)
```

```
##    speed dist
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
lm(dist~speed, data=cars) %>% summary()
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

由上面結果可以看出，speed of cars 對 distances taken to stop the cars 有正相關與統計顯著性，速度每增加 1 單位會造成距離增加 3.93 單位，兩者之間的正向關係符合我們的預期。

6.

```
data("chickwts")
str(chickwts)
```

```
## 'data.frame':    71 obs. of  2 variables:
##  $ weight: num  179 160 136 227 217 168 108 124 143 140 ...
##  $ feed  : Factor w/ 6 levels "casein","horsebean",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
head(chickwts)
```

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

```
tail(chickwts)
```

```
##    weight   feed
## 66    352 casein
## 67    359 casein
## 68    216 casein
## 69    222 casein
## 70    283 casein
## 71    332 casein
```

```
lm(weight~feed, data=chickwts) %>% summary()
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
```

```
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

由上面結果可以看出，以 casein 為基準組，horsebean, linseed, meatmeal soybean 對小雞體重的效果都不如 casein，而且結果具統計顯著性，例如用 horsebean 餵食的小雞，平均而言體重比用 casein 餵食的小雞少 163.383 單位。

使用 sunflower 餵食的小雞雖然似乎比使用 casein 餵食的小雞好點，但不具統計顯著性，換言之用 sunflower 或用 casein 餵食小雞在統計上並無顯著差異。