大數據統計與預測 第八章 與 第九章 作業

Created by Weber, YC Huang (黃彥鈞) m946108006

2019-11-08

Chapter 8

1. What is statistical inference?

通常母體的特徵值參數是未知的，而跟跟樣本部分訊息推測母體特徵值的過程，我們稱之為統計推論。(如推測母體平均數、變異數、標準差...等)。

4. What is the standard error of a sample mean? How does the standard error compare to the standard deviation of the population?

標準誤，即樣本**平均數**抽樣分配的標準差，是描述對應的樣本平均數抽樣分布的離散程度及衡量對應樣本平均數抽樣誤差大小的尺度。

SD 是指原始母體資料之標準差；而SE 則是樣本統計量之標準差。

5. Explain the central limit theorem.

從同一母群體取出樣本數為n之無限多組樣本，當「樣本平均數抽樣分佈」抽樣之樣本數n趨近於無限大時，依據「中央極限定理」，其分佈具有以下特性：

- 樣本平均數抽樣分佈會趨近常態分佈。
- 樣本平均數抽樣分佈之平均數會等於母群體平均數。
- 本平均數抽樣分佈的標準差，又稱「平均數之標準誤」，會等於母群體標準差除以樣本數 n 的平方根。 (隨著n增加，平均數之標準誤會隨之變小。)

6. What happens to the amount of sampling variability among a set of sample means $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots$ as the size of the samples increases?

抽樣變異會隨樣本量上升而降低。

8. Among adults in the United States, the distribution of albumin levels (albumin is a type of protein) in cerebrospinal fluid is roughly symmetric with mean $\mu = 29.5$ mg/100 ml and standard deviation $\sigma = 9.25$ mg/100 ml [5]. Suppose that you select repeated samples of size 20 from this population and calculate the mean for each sample.

(a) If you were to select a large number of random samples of size 20, what would be the mean of the sample means?

(b) What would be their standard deviation? What is another name for this standard deviation of the sample means?

(c) How does the standard deviation of the sample means compare with the standard deviation of the albumin levels themselves?

(d) If you were to take all the different sample means and use them to construct a histogram, what would be the shape of their distribution?

(e) What proportion of the means of samples of size 20 are larger than 33 mg/100 ml?

(f) What proportion of the means are less than 28 mg/100 ml?

(g) What proportion of the means are between 29 and 31 mg/100 ml?

(a) 根據中央極限定理，樣本夠大的情況下，抽樣平均值分布的平均趨近於母體平均值，29.5

(b) 抽樣平均值的標準差又稱為(標準誤) 值為: $\sigma/\sqrt{n}$

(c) SD 是指原始母體資料之標準差；而SE 則是樣本統計量之標準差。SD 用來衡量母體資料中，資料與母體平均之離散程度；而 SE 用以估計，樣本平均與母體平均之差距

(d) 得到常態分佈之鐘型曲線

(e) $P(\overline{\mathrm{x}} > 33)$? `1-pnorm(33,29.5,9.25)=0.3525748`

(f) $P(\overline{\mathrm{x}} < 28)$? `pnorm(28,29.5,9.25)=0.4355891`

(g) $P(29 \leqslant \overline{\mathrm{x}} < 31)$? `pnorm(31,29.5,9.25)-pnorm(29,29.5,9.25)=0.08596487`

11. In Norway, the distribution of birth weights for infants whose gestational age is 40 weeks is approximately normal with mean $\mu = 3500$ grams and standard deviation $\sigma = 430$ grams [7].

(a) Given a newborn whose gestational age is 40 weeks, what is the probability that his or her birth weight is less than 2500 grams?

(b) What value cuts off the lower 5% of the distribution of birth weights?

(c) Describe the distribution of means of samples of size 5 drawn from this population. List three properties.

(d) What value cuts off the lower 5% of the distribution of samples of size 5?

(e) Given a sample of five newborns all with gestational age 40 weeks, what is the probability that their mean birth weight is less than 2500 grams?

(f) What is the probability that only one of the five newborns has a birth weight less than 2500 grams?

(a) $P(X < 2500) =$ `pnorm(2500,3500,430)` =0.01002045

(b) `qnorm(0.05,3500,430)=2792.713`

(c)

(d)

(e) `pnorm(2500,3500,430/sqrt(5))=0.000`

(f) `5*pnorm(2500,3500,430)*(1-pnorm(2500,3500,430))^4=0.04812403`

12. For the population of females between the ages of 3 and 74 who participated in the National Health Interview Survey, the distribution of hemoglobin levels has mean $\mu = 13.3$ g/100 ml and standard deviation $\sigma = 1.12$ g/100 ml [8].

(a) If repeated samples of size 15 are selected from this population, what proportion of the samples will have a mean hemoglobin level between 13.0 and 13.6 g/100 ml?

(b) If the repeated samples are of size 30, what proportion will have a mean between 13.0 and 13.6 g/100 ml?

(c) How large must the samples be for 95% of their means to lie within ± 0.2 g/100 ml of the population mean μ?

(d) How large must the samples be for 95% of their means to lie within ± 0.1 g/100 ml of the population mean?

(a) $\sigma/\sqrt{n} = 0.2891828$, $P(13 \leqslant \bar{x} < 13.6) =$ `pnorm(13.6,13.3,0.29)-pnorm(13,13.3,0.29)=0.6990895`

(b) $\sigma/\sqrt{n} = 0.2044831$, $P(13 \leqslant \bar{x} < 13.6) =$ `pnorm(13.6,13.3,0.2)-pnorm(13,13.3,0.2)=0.8663856`

(c) $95\%\ n\ size = 120.4726 \approx 120.4$

```
# Way to calculate 99% 95% 90% sample size
sample_determine <- function(x, sigma, ci_rate, ci){
  S <- 0
  if (ci_rate == 0.99){
    S <- (2.58*sigma/(ci/2))^2
  }
  else if (ci_rate == 0.95) {
    S <- (1.96*sigma/(ci/2))^2
  }
  else {
    S <- (1.645*sigma/(ci/2))^2
  }
  cat(ci_rate*100,'% n_size :',S)
}

sample_determine(13.3, 1.12, 0.95, 0.4) # ci = 0.2+0.2
# 95 % n_size : 120.4726
```

(d) $95\%\ n\ size = 481.8903 \approx 481.9$

```
# Way to calculate 99% 95% 90% sample size
sample_determine <- function(x, sigma, ci_rate, ci){
  S <- 0
  if (ci_rate == 0.99){
    S <- (2.58*sigma/(ci/2))^2
  }
  else if (ci_rate == 0.95) {
    S <- (1.96*sigma/(ci/2))^2
  }
  else {
    S <- (1.645*sigma/(ci/2))^2
```

```
  }
  cat(ci_rate*100,'% n_size :',S)
}

sample_determine(13.3, 1.12, 0.95, 0.2) # ci = 0.1+0.1
# 95 % n_size : 481.8903
```

Chapter 9

1. Explain the difference between point and interval estimation.

點估計，用樣本數據來估計母體參數， 估計結果使用一個點的數值表示「最佳估計值」 ；區間估計，使用區間來估計未知的母群體參數，以導出對於母群體的推論

2. Describe the 95% confidence interval for a population mean μ. How is the interval interpreted?

有95%信心估計母群體平均數，在樣本平均數 ± 1.96 * (標準誤) 的範圍內。 我們有 95%信心區間會包含母體平均

4. Describe the similarities and differences between the t distribution and the standard normal distribution. If you were trying to construct a confidence interval, when would you use one rather than the other?

t 分布與常態分佈一樣，分布狀態都是單峰對稱；不同的地方在於，t 分布雙尾分布資料較多，極端值可能比常態分佈多，根據自由度的不同，t 分布形狀也會有差異。

t分布用於小樣本，總體變異數未知的情況，若變異數已知則應使用常態分佈。

5. The distributions of systolic and diastolic blood pressures for female diabetics between the ages of 30 and 34 have unknown means. However, their standard deviations are $\sigma_s = 11.8$ mm Hg and $\sigma_d = 9.1$ mm Hg, respectively [8].

(a) A random sample of ten women is selected from this population. The mean systolic blood pressure for the sample is $\bar{x}_s = 130$ mm Hg. Calculate a two-sided 95% confidence interval for μ_s, the true mean systolic blood pressure.
(b) Interpret this confidence interval.
(c) The mean diastolic blood pressure for the sample of size 10 is $\bar{x}_d = 84$ mm Hg. Find a two-sided 90% confidence interval for μ_d, the true mean diastolic blood pressure of the population.
(d) Calculate a two-sided 99% confidence interval for μ_d.
(e) How does the 99% confidence interval compare to the 90% interval?

(a) $130 \pm 1.96 \times (11.8/\sqrt{10}) = (122.6863, 137.3137)$

(b) 我們有95%的信心，母體平均將落在以上區間(122.6863,137.3137)

(c) $84 \pm 1.65 \times (9.1/\sqrt{10}) = (79.25184, 88.74816)$

(d) $84 \pm 2.58 \times (9.1/\sqrt{10}) = (76.5756, 91.4244)$

(e) 99% 信賴區間較 90% 來的大，區間越大我們就有足夠之信心說明母體平均落於

6. Consider the t distribution with 5 degrees of freedom.

   (a) What proportion of the area under the curve lies to the right of $t = 2.015$?

   (b) What proportion of the area lies to the left of $t = -3.365$?

   (c) What proportion of the area lies between $t = -4.032$ and $t = 4.032$?

   (d) What value of t cuts off the upper 2.5% of the distribution?

(a) `pt(2.015, 5, lower.tail = FALSE)=0.05000309`

(b) `pt(-3.365, 5, lower.tail = TRUE)=0.009999236`

(c) `pt(4.032, 5, lower.tail = TRUE)-pt(-4.032, 5, lower.tail = TRUE)=0.9899986`

(d) `qt(0.025,5, lower.tail = FALSE)=2.570582`

8. Before beginning a study investigating the ability of the drug heparin to prevent bronchoconstriction, baseline values of pulmonary function were measured for a sample of 12 individuals with a history of exercise-induced asthma [9]. The mean value of forced vital capacity (FVC) for the sample is $\bar{x}_1 = 4.49$ liters and the standard deviation is $s_1 = 0.83$ liters; the mean forced expiratory volume in 1 second (FEV_1) is $\bar{x}_2 = 3.71$ liters and the standard deviation is $s_2 = 0.62$ liters.

   (a) Compute a two-sided 95% confidence interval for μ_1, the true population mean FVC.

   (b) Rather than a 95% interval, construct a 90% confidence interval for the true mean FVC. How does the length of the interval change?

   (c) Compute a 95% confidence interval for μ_2, the true population mean FEV_1.

   (d) In order to construct these confidence intervals, what assumption is made about the underlying distributions of FVC and FEV_1?

(a)

```
# T distribution calculate confidence interval
t_dis_ci <- function(x, sd, n, ci_rate){
  ci <- 1 - ((1 - ci_rate)/2)
  error <- qt(ci,df = n-1)*s/sqrt(n)
  left <- x - error
  right <- x + error

  cat('(',left,',',right,')')
}
```

`t_dis_ci(4.49,0.83,12,0.95) = ( 3.962643 , 5.017357 )`

(b) `t_dis_ci(4.49,0.83,12,0.9)=( 4.059705 , 4.920295 )`

區間變小了，信心下降QQ

(c) `t_dis_ci(3.71,0.62,12,0.95)=( 3.316071 , 4.103929 )`

(d) 假設母體分布為常態，

9. For the population of infants subjected to fetal surgery for congenital anomalies, the distribution of gestational ages at birth is approximately normal with unknown mean μ and standard deviation σ. A random sample of 14 such infants has mean gestational age $\bar{x} = 29.6$ weeks and standard deviation $s = 3.6$ weeks [10].

(a) Construct a 95% confidence interval for the true population mean μ.

(b) What is the length of this interval?

(c) How large a sample would be required for the 95% confidence interval to have length 3 weeks? Assume that the population standard deviation σ is known and that $\sigma = 3.6$ weeks.

(d) How large a sample would be needed for the 95% confidence interval to have length 2 weeks?

(a)

```
# T distribution calculate confidence interval
t_dis_ci <- function(x, sd, n, ci_rate){
  ci <- 1 - ((1 - ci_rate)/2)
  error <- qt(ci,df = n-1)*s/sqrt(n)
  left <- x - error
  right <- x + error

  cat('(',left,',',right,')')
}
```

`t_dis_ci(29.6,3.6,14,0.95)=( 27.52142 , 31.67858 )`

(b) 4.15716

(c) $95\% \ n \ size = 22.12762 \approx 22.1$

```
# Way to calculate 99% 95% 90% sample size
sample_determine <- function(x, sigma, ci_rate, ci){
  S <- 0
  if (ci_rate == 0.99){
    S <- (2.58*sigma/(ci/2))^2
  }
  else if (ci_rate == 0.95) {
    S <- (1.96*sigma/(ci/2))^2
  }
  else {
    S <- (1.645*sigma/(ci/2))^2
  }
  cat(ci_rate*100,'% n_size :',S)
}

sample_determine(29.6, 3.6, 0.95, 3)
# 95 % n_size : 22.12762
```

(d) $95\% \ n \ size = 49.78714 \approx 49.8$

```
# Way to calculate 99% 95% 90% sample size
sample_determine <- function(x, sigma, ci_rate, ci){
  S <- 0
  if (ci_rate == 0.99){
    S <- (2.58*sigma/(ci/2))^2
```

```
  }
  else if (ci_rate == 0.95) {
    S <- (1.96*sigma/(ci/2))^2
  }
  else {
    S <- (1.645*sigma/(ci/2))^2
  }
  cat(ci_rate*100,'% n_size :',S)
}

sample_determine(29.6, 3.6, 0.95, 2)
# 95 % n_size : 49.78714
```

11. When eight persons in Massachusetts experienced an unexplained episode of vitamin D intoxication that required hospitalization, it was suggested that these unusual occurrences might be the result of excessive supplementation of dairy milk [12]. Blood levels of calcium and albumin for each individual at the time of hospital admission are shown below.

Calcium (mmol/l)	Albumin (g/l)
2.92	43
3.84	42
2.37	42
2.99	40
2.67	42
3.17	38
3.74	34
3.44	42

(a) Construct a one-sided 95% confidence interval—a lower bound—for the true mean calcium level of individuals who experience vitamin D intoxication.

(b) Compute a 95% lower confidence bound for the true mean albumin level of this group.

(c) For healthy individuals, the normal range of calcium values is 2.12 to 2.74 mmol/l and the range of albumin levels is 32 to 55 g/l. Do you believe that patients suffering from vitamin D intoxication have normal blood levels of calcium and albumin?

(a) $Calcium\ lower\ ci = 2.845492$

```
cal = c(2.92,3.84,2.37,2.99,2.67,3.17,3.74,3.44)
x = mean(cal)
s = sd(cal)
print(paste('sample mean :',x,'sample std :',std))
# "sample mean : 3.1425 sample std : 0.510678819947388"
```

$\overline{\mathrm{x_c}} = 3.1425;\ s_c = 0.5106788$

$one\ sized\ 95\%\ lower\ bound\ : \overline{\mathrm{x_c}} - 1.645 \times (\frac{\sigma}{\sqrt{n}})$

```r
one_side_95 <- function(x, s, n){
  upper <- x+1.645*(s/sqrt(n))
  lower <- x-1.645*(s/sqrt(n))
  cat('upper ci :',upper,';lower ci :',lower)
}
one_side_95(3.1425,0.5106788,8)
# upper ci : 3.439508 ;lower ci : 2.845492
```

(b) $Albumin\ lower\ ci = 38.61814$

```r
cal = c(43,42,42,40,42,38,34,42)

# one sided ci
one_side_95 <- function(obs){
  x <- mean(obs)
  s <- sd(obs)
  n <- length(obs)
  cat('sample mean :',x,';sample std :',s,';n size :',n,'\n')

  upper <- x+1.645*(s/sqrt(n))
  lower <- x-1.645*(s/sqrt(n))
  cat('upper ci :',upper,';lower ci :',lower)
}

one_side_95(cal)
# sample mean : 40.375 ;sample std : 3.020761 ;n size : 8
# upper ci : 42.13186 ;lower ci : 38.61814
```

(c) Calcium 平均值落在信賴區間 2.845 以上，並沒有落在正常值 (2.12-2.74)，需要醫院進一步評估；另一方面，我們無法從信賴區間得出 Albumin 之相關結論