Sampling Distribution & Estimator

Central Limiting Theorem
t-distribution
estimators
sample size

參數 vs 統計量

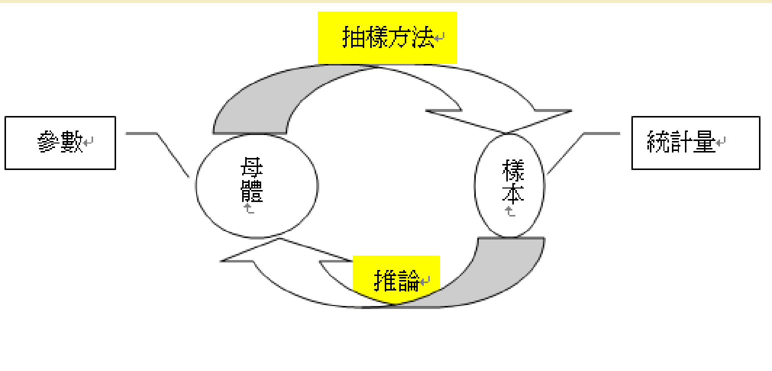


圖 7-1　母體與樣本間的關係

Terms

估計（Estimation）：根據樣本資料所取得的統計量抽樣分配，來推估母群體參數的統計方法。

估計量（Estimator）：在估計過程中，用來推估母群體參數的統計量。

估計值（Estimate）：估計量所決定的值。

統計量

統計量： 為樣本資料特徵的測量值。

樣本統計量為隨機樣本 (random samples) 的函數，而隨機樣本是由 n 個隨機變數所組成的，

故樣本統計量亦為一隨機變數為樣本之函數，亦為一隨機變數，

抽樣分配

樣本統計量為一隨機變數，其機率分配稱為抽樣分配。

樣本平均數

$$\bar{X}=\frac{X_1+X_2+X_3+...+X_n}{n}=\frac{1}{n}\sum X_i$$

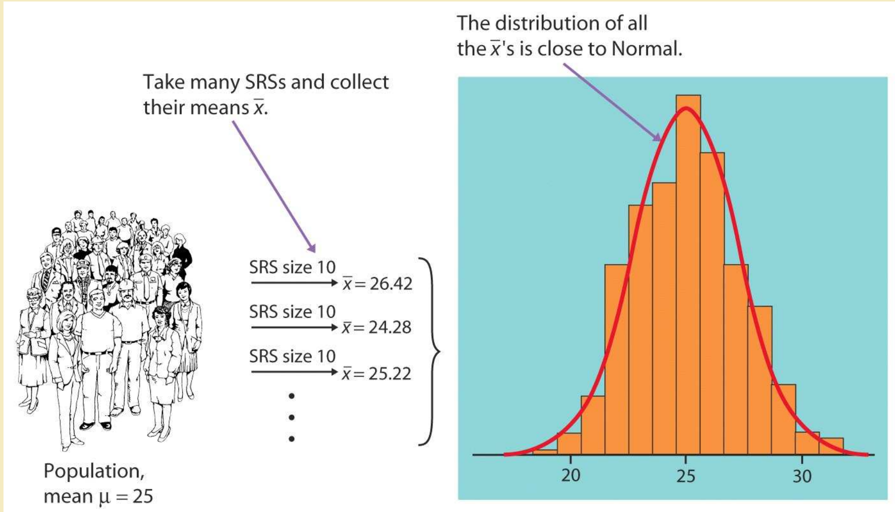
一般統計量的抽樣分配，多根據重覆抽樣（實驗）結果來瞭解其機率模式。

- 一大數法則，中央極限定理

統計量的抽樣分配行為

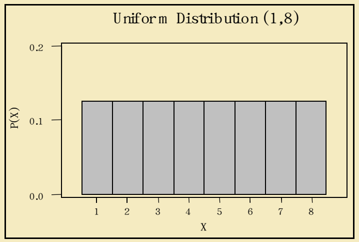
樣本數不大時，我們可重覆抽樣多次，每次計算統計量的樣本值，再以直方圖呈現統計量的抽樣分配行為。

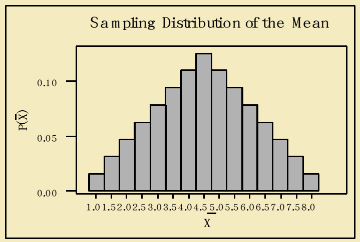
- 一從同一母體中，重覆抽樣多次，每次樣本數為10。
- 一計算每組樣本的平均數 。
- 一以直方圖圖示所有的平均數 $\bar{X}_{10}$。
- 一檢視圖形的形狀、中心、分散度以及離群點和其他變異處。



Properties of the Sampling Distribution of the Sample Mean

- Comparing the population distribution and the sampling distribution of the mean:
  - ✓The sampling distribution is more bell-shaped and symmetric.
  - ✓Both have the same center.
  - ✓The sampling





Relationships between Population Parameters and the Sampling Distribution of the Sample Mean

The **expected value of the sample mean** is equal to the population mean:

$$E(\bar{X}) = m_{\bar{X}} = m_X$$

The **variance of the sample mean** is equal to the population variance divided by the sample size:

$$V(\bar{X}) = s^2_{\bar{X}} = \frac{s^2_X}{n}$$

The **standard deviation of the sample mean, known as the standard error of the mean**, is equal to the population standard deviation divided by the square root of the sample size:

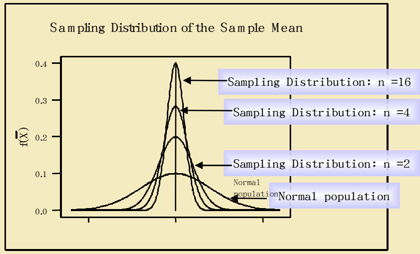
$$SD(\bar{X}) = s_{\bar{X}} = \frac{s_X}{\sqrt{n}}$$

Sampling from a Normal Population

When sampling from a **normal population** with mean μ and standard deviation σ, the sample mean, X, has a **normal sampling distribution:**

$$\bar{X} \sim N(m, \frac{s^2}{n})$$

This means that, as the sample size increases, the sampling distribution of the sample mean remains centered on the population mean, but becomes more compactly distributed around that



law of large number

大數法則

$$P\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1$$ ¿

亦即當樣本數夠大時，樣本平均數會趨近於母體平均的機率爲 1 。

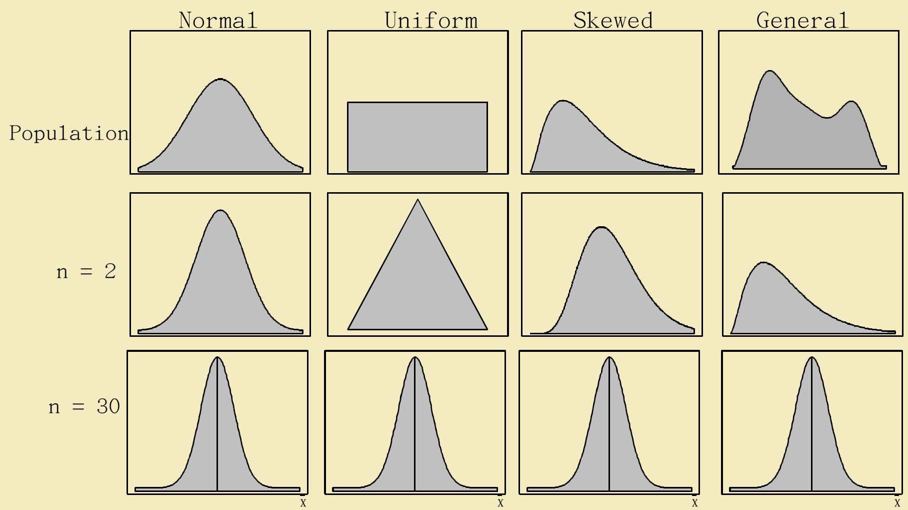
The Central Limit Theorem

Theorem 5.4-1 If $\bar{X}$ is the mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from a distribution with a finite mean μ and a finite positive variance σ^2, then the distribution of

$$W = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}$$

is $N(0,1)$ in the limit as $n \to \infty$.

The Central Limit Theorem Applies to Sampling Distributions from Any Population



Remark

If $X_i \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, for $i=1,\cdots,n$, then

$$U=\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n),$$

and

$$W=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

ex:

ex ：某班同學統計成績呈現常態分配 N(72,81) ，問
(1) 隨機抽 1 人，分數超過 80 之機率？
(2) 隨機抽 10 人，平均分數超過 80 機率？
(3) 同 (2) ，總分超過 800 分之機率？
(4) 若母體非常態，如何做？

ex ：由下列分配中，抽 30 個 R.S. ，求各組樣本均數大於之機率。

(1) $X \sim Poisson(4.5)$

(2) $X \sim Bin(9, 0.5)$

Student’s T Distribution

Definition:

If r is a positive integer and $T=\frac{Z}{\sqrt{U/r}}$,

where $Z \sim N(0,1), U \sim \chi^2(r)$, and they are independent,

then we say that T has a t distribution with r degree of freedom.

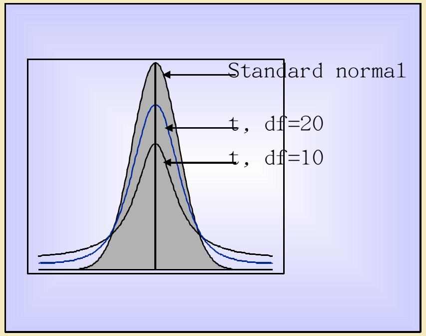
Denote it $T \sim t(r)$

Gossett, William Sealy published “t-test” in Biometrika 1908 to measure the confidence interval, the deviation of “small samples” from the “real”.

Suppose the underlying distribution is normal with unknown σ^2.

Student's t Distribution

- The t is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of t is 0.
- The variance of t is greater than 1, but approaches 1 as the number of degrees of freedom increases. The t is flatter and has fatter tails than does the standard normal.
- The t distribution approaches a standard



Remark

If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, s. If the population is normal, the resulting statistic:

$$t=\frac{\bar{X}-\mu}{s/\sqrt{n}}$$

has a ***t distribution with (n − 1) degrees of freedom***.

$$T=\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}/(n-1)}}.$$

We know that

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \text{ is } N(0,1),\quad U=\frac{(n-1)S^2}{\sigma^2} \text{ is } \chi^2(n-1),$$

and these two are independent. So, $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

Estimation

在統計的應用上我們常可依主客觀條件假設母群體是甚麼分佈，而分佈裡常含有未知參數 (unknown parameter)。 因此應用前我們必須先估計這些參數。

我們可從這母群體中取得一組個隨機樣本，根據這組隨機樣本來估計參數。一個參數的點估計量 (point estimator) 是一個不含未知參數的隨機樣本之函數，所以點估計量本身是一個統計量，當樣本取得後就能計算出該參數的估計值。

一個參數可以有一個或以上之估計量，例如估計母群體期望值時，樣本平均值與中位數都是“合理”的估計量。 此時我們該如何取捨呢？

Methods

點估計 (Point Estimation)

- 根據樣本資料所求得之單一個估計值，以推估未知的母群體參數。
- 單一數值無法評估其估計的準確度。

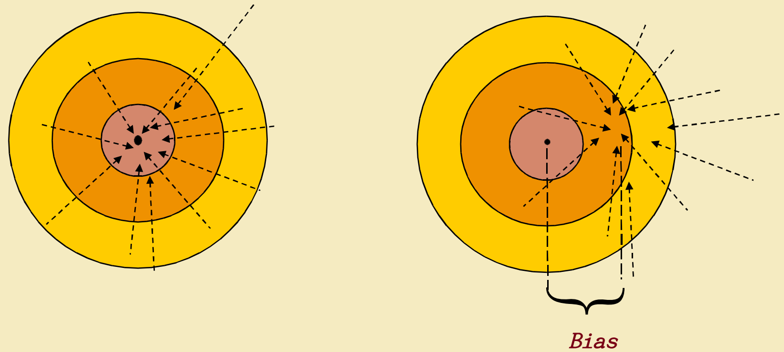
區間估計 (Interval Estimation)

- 根據所求得之點估計量的抽樣分配特質，求出兩個數值以構成一區間，並利用此一區間推估未知的母群體參數範圍。
- 區間估計是以區間估計值來推估母群體參數的真實值落於該區間的機率大小。

Estimation

- **Consider the following statements:**
  - ✓ $\overline{x}$ = 550
    - A single-valued estimate that conveys little information about the actual value of the population mean.
  - ✓ **We are 99% confident that μ is in the interval [449,551]**
    - An interval estimate which locates the population mean within a narrow interval, with a high level of confidence.
  - ✓ **We are 90% confident that μ is in the interval [400,700]**
    - An interval estimate which locates the population

Unbiased and Biased Estimators



An **unbiased** estimator is on target on average.

A **biased** estimator is off target on average.

some often used point Estimators

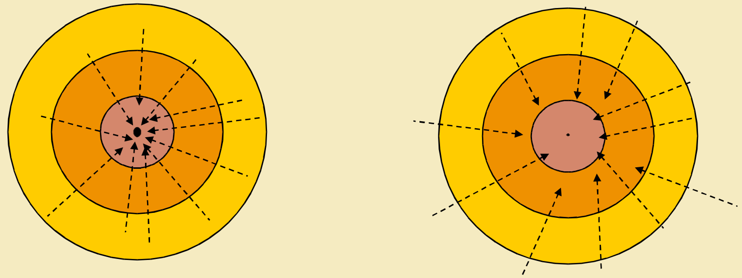
An **estimator** of a population parameter is a sample statistic used to estimate the parameter. The most commonly-used estimator of the:

Population Parameter		Sample Statistic
Mean (μ)	is the	Mean ($\overline{X}$)
Variance (σ^2)	is the	Variance (s^2)
Standard Deviation (σ)	is the	Standard Deviation (s)
Pro		$\hat{p}$

- Desirable properties of estimators include:
  - ✓ Unbiasedness
  - ✓ Efficiency
  - ✓ Consistency
  - ✓ Sufficiency

Efficiency

An estimator is **efficient** if it has a relatively small variance (and standard deviation).



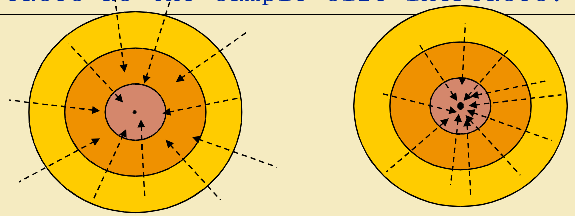
An **efficient** estimator is, on average, closer to the parameter being estimated..

An **inefficient** estimator is, on average, farther from the parameter being estimated.

Consistency and Sufficiency

An estimator is said to be **consistent** if its probability of being close to the parameter it estimates increases as the sample size increases.

Consistency



n =10 n = 100

An estimator is said to be **sufficient** if it contains all the information in the data about the parameter it estimates.

Properties of the Sample Mean

For a normal population, both the sample mean and sample median are unbiased estimators of the population mean, but the sample mean is both more efficient (because it has a smaller variance), and sufficient. Every observation in the sample is used in the calculation of the sample mean, but only the middle value is used to find the sample median.

In general, the sample mean is the best estimator of the population mean. The sample mean is the most efficient unbiased estimator of the population mean. It is also a consistent estimator.

Properties of the Sample Variance

The ***sample variance*** (the sum of the squared deviations from the sample mean divided by *(n–1)* is an ***unbiased estimator*** of the population variance. In contrast, the ***average squared deviation*** from the sample mean is a ***biased*** (though *consistent*) estimator of the population variance.

$$E(s^2)\quad E\quad \frac{(x\quad \bar{x})^2}{(n\quad 1)}\qquad {}^2$$

$$E\quad \frac{(x\quad \bar{x})^2}{n}\qquad {}^2$$

Confidence Interval or Interval Estimate
母群體平均值之區間估計

A ***confidence interval*** or ***interval estimate*** is a range or interval of numbers believed to include an unknown population parameter. Associated with the interval is a measure of the ***confidence*** we have that the interval does indeed contain the parameter of interest.

- A confidence interval or interval estimate has two components:
  - ✓ A range or interval of values
  - ✓ An associated *level of confidence*

Confidence Interval for μ When σ Is Known

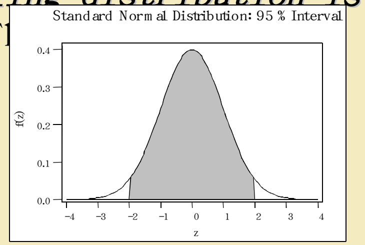
- **If the population distribution is normal, *the sampling distribution of the mean is normal.***
  - **If the sample is sufficiently large, regardless of the shape of the population distribution, *the sampling distribution is normal*** (Central Limit Th

In either case:

$$P\left(-1.96<\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}<1.96\right)=0.95$$

or

$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)=0.95$$



Confidence Interval for μ when σ is Known (Continued)

Before sampling, there is a 0.95 probability that the interval

$$1.96\frac{\;}{\sqrt{n}}$$

will include the sample mean (and 5% that it will not).

Conversely, after sampling, approximately 95% of such intervals

$$\bar{x}\quad 1.96\frac{\;}{\sqrt{n}}$$

will include the population mean (and 5% of them will not).

That is, **$\bar{x}$ $1.96\frac{\;}{\sqrt{n}}$ is a 95% confidence interval for .**

R–sampling

The **sample** function is used to generate a random sample from a given population. It can be used to sample with or without replacement by using the **replace** argument (the default is F). The only obligatory argument is a vector of data which will constitute the population from which the sample will be drawn.

The default is to create a sample equal in size to the population but by using the **size** argument any sample size can be specified. A vector of probabilities can also be supplied in the **prob** argument. This vector has to be equal in length to the size of the population and it will automatically be normalized if its elements do not sum up to one. The default is for every element in the population to have equal chance of being chosen

R: sampling from a population

- #Generating a random sample from a Poisson distribution with lambda=3
- set.seed(124)
- **pois <- rpois(100, lambda=3)**
- **pois[1:10]**
- **mean(pois)**
- **var(pois)**

- #Generating a random sample from a Binomial distribution with size=20 and prob=.2 set.seed(124)
- **binom <- rbinom(100, 20, .2)**
- **binom[1:10]**
- **mean(binom)**
- **sd(binom)**

- CLT: sampling n random sample for the population “a” , and repeating the processing nsample times. See the distribution of xbar.

```
a=c(1:6)
n=3
nsample=20
xbar=rep(0, nsample)
for (i in 1:nsample)
{
rsample=sample(a,n,replace=T)
xbar[i]=mean(rsample)
print(c(rsample, xbar[i]))
}
hist(xbar)
```

R: CLT

```
#function for central limit
   theorem
#sampling wwith replacement
#a : original population
#n : sampling size
#N: repeated sampling
```

```
central=function(a, N, n ) {
   mu.a=mean(a)
   sd.a=sd(a)
   x.bar=rep(NA, N)
   for (i in 1:N) {
    x.bar[i]=mean(sample(a, n,
    replace=TRUE))
   }
   x.barnorm=(x.bar-
   mu.a)/sd.a*sqrt(n)
   hist(x.barnorm, freq=FALSE,
   main=paste("R: N(0,1),
   n=",n))
   points(density(x.barnorm),
   cex=.2, col="blue")
}

central(c(1:10),N=30, n=4)
```

區間估計（2）

區間的大小可分為 90% ， 95% ， 99% 信賴區間或信賴界線。所謂 95% 信賴區間，以拿平均數來說，每抽一次樣本，就利用某種公式，算得其 95% 的信賴區間，如此重複很多次，在這些信賴區間裡，將會有 95% 包括母體的平均數。

在 100 次中有 95 次會包含母體平均數，也就有 5 次沒有包括母體平均數。這稱為顯著水準，就是 1 減去信賴水準。若以 95% 信賴水準而言，顯著水準就是 5% 。通常用 {alpha} 表示顯著水準。

區間估計

研究者想知道電池的使用壽命，抽樣若干個電池，發現壽命的平均數為 50 小時。母體的平均數會是多少？會是介於哪段區間？

研究者關心初生男嬰體重的變異數，隨機抽樣若干位初生男嬰，得體重的變異數為 40000 公克，母體變異數會介於哪段區間內？

以上這些研究問題，都在探討母體某個參數倒底介於哪段區間內，這就是所謂的區間估計。

θ 的 $(1-\alpha)$100% 信賴區間

P[l < θ < u] = $(1-\alpha)$100%

- l ：信賴下限（Lower Limit of Confidence）
- u ：信賴上限（Upper Limit of Confidence）
- θ ：母群體參數（Parameter）
- $1-\alpha$ ：信賴水準（Confidence Level）
- α ：可能誤差（Possible Error）
- 若重覆進行 100 次的抽樣與估算，其中約有 $(1-\alpha)\times 100$ 個估計值會涵蓋母群體的參數。

μ 的信賴區間

適用時機： $\sigma 2$ 已知

雙側信賴區間

$$P[\bar{x}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}<\mu<\bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]=(1-\alpha)100$$

右側信賴區間

$$P[\mu<\bar{X}+Z_{\alpha}\frac{\sigma}{\sqrt{n}}]=(1-\alpha)100$$

左側信賴區間

$$P[\bar{x}-Z_{\alpha}\frac{\sigma}{\sqrt{n}}<\mu]=(1-\alpha)100$$

（ 1−a ） 100% 的信心下之樣本數

在（ 1−a ） 100% 信心，用樣本平均數估計母體平均數的誤差 e 在 $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ 之內。 即

$$n=\left(z_{\alpha/2}\,\sigma/e\right)^2$$

ex: 如果希望有 95% 的信心，利用估計 m ，誤差不會超過 3 小時，則需要多大的樣本數？

平均數的區間估計 (ex)

某廠牌手機電池的待機時數近似常態分佈，變異數為 100 。現隨機抽取 25 個電池，檢查其待機時數，得到平均數為 50 。試求母體平均數的 95% 信賴區間。

母體平均數的 95% 信賴區間為：

$$50-1.96\times\frac{10}{\sqrt{25}}<\mu<50+1.96\times\frac{10}{\sqrt{25}}\Rightarrow 46.08<\mu<53.92$$

例子（有限母體）

某公司想瞭解員工每天上網的時間，該公司員工共 50 位，抽樣記錄了 10 位員工，結果發現平均數為 60 分鐘。已知上網時間近似常態分佈，標準差為 20 。求母體平均數的 90% 信賴區間。

樣本數 10 佔母體數 50 的比例高達 1/5

不放回法抽樣

平均數的區間估計

不放回法抽樣且母體變異數已知

有限母體（ finite population ）

$$\bar{X} \sim N\left(\mu\ ,\ \frac{\sigma^2}{n}\times\frac{N-n}{N-1}\right)$$

母體平均數 m 的（ 1-a ） 100% 信賴區間是

$$\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\times\sqrt{\frac{N-n}{N-1}}<\mu<\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\times\sqrt{\frac{N-n}{N-1}}$$

$$60-1.645\times\frac{20}{\sqrt{10}}\times\sqrt{\frac{50-10}{50-1}}<\mu<60+1.645\times\frac{20}{\sqrt{10}}\times\sqrt{\frac{50-10}{50-1}}$$

$$\Rightarrow 50.6<\mu<69.4$$

例子

承上題，如果希望在 90% 的信心下，誤差不超過 5 分鐘，則必須抽樣多少人？

作法

$$n=\frac{50\times1.645^2\times20^2}{49\times5^2+1.645^2\times20^2}=23.5$$

如果樣本數為 24 ，在 90% 的信心下，誤差不超過 5 分鐘。

樣本數

在（ 1-a ） 100% 的信心下，用樣本平均數估計母體平均數的誤差 $e=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\times\sqrt{\frac{N-n}{N-1}}$ 之內。

整理後得

$$n=\frac{Nz_{\alpha/2}^2\sigma^2}{(N-1)e^2+z_{\alpha/2}^2\sigma^2}$$

區間估計（ σ^2 未知 大樣本）

當母體不是常態分佈，只要樣本數夠大，那麼樣本平均數的抽樣分佈也會非常接近 t 分佈。

令 X1 ， ， Xn 來自常態分佈，但其變異數未知，則

$$T=\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$$

是自由度為 n － 1 的 t 分佈。介於 -ta/2 和 ta/2 的機率為 1-a 。

如果母體既不是常態分佈，且樣本數又少，那麼對母體平均數的估計就變得十分不穩定。

適用時機： σ^2 未知

雙側信賴區間

$$P[\bar{x}-t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}<\mu<\bar{X}+t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}]=(1-\alpha)100$$

右側信賴區間

$$P[\mu<\bar{X}+t_{n-1,\alpha}\frac{s}{\sqrt{n}}]=(1-\alpha)100$$

左側信賴區間

$$P[\bar{x}-t_{n-1,\alpha}\frac{s}{\sqrt{n}}<\mu]=(1-\alpha)100$$

兩平均數差異區間估計

母體變異數已知

若有兩個獨立的常態分佈母體，其平均數分別為 m1 和 m2 ，變異數為 σ_1^2 和 σ_2^2 ，則

$$\bar{X}_1-\bar{X}_2 \sim N\left(\mu_1-\mu_2,\ \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

如果不是常態母體時，基於中央極限定理，只要兩個樣本數均很大，公式亦可成立。

$$\bar{X}_1- \quad \{\bar{X}$$

$$Z=\frac{(¿¿_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

兩平均數差異區間估計

男生與女生的智商平均數差異有多大？經過一段時間的實驗教學之後，實驗班和正常組的學業成績的平均數差異多大。在減肥課程訓練之前，量一下體重，經過一段時間的訓練之後，再量體重，兩個體重平均數是否有差異。

以上所面臨的問題就是兩個母體平均數的差異。可用兩個樣本的平均數的差異當作兩母體平均數差異的點估計。

同樣的，必須理解該點估計（統計量）的抽樣分佈，才能進行區間估計。

兩平均數差異區間估計

某研究者想瞭解喝啤酒對注意力的影響，他隨機分派各 50 人至實驗組和控制組中。實驗組要喝一瓶啤酒，控制組則喝一瓶開水。然後測試他們的注意力，總分 0 至 100 分，分數越高表示注意力越好如果依照過去的經驗，喝啤酒或喝白開水的人的注意力的變異數都是 25 。現得到實驗組的平均數為 55 ，控制組為 58 。求實驗組與控制組的平均數差異的 95% 信賴區間。

$$(55-58)-1.96\times\sqrt{\frac{25}{50}+\frac{25}{50}}<\mu_1-\mu_2<(55-58)+1.96\times\sqrt{\frac{25}{50}+\frac{25}{50}}$$

兩平均數差異區間估計

母體變異數未知：大樣本

雖然母體變異數未知，但如果兩個母體是常態分佈，且樣本數 n1 和 n2 夠大（如均大於 25），仍可用 Z 分佈。

$$(\bar{x}_1-\bar{x}_2)-z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}<\mu_1-\mu_2<(\bar{x}_1-\bar{x}_2)+z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

兩平均數差異區間估計（例子）

在一項關於速讀訓練的實驗中，研究者隨機分派各 5 位受試者到實驗組（接受速讀訓練課程）和控制組（只接受和速讀無關的一些活動），為期 10 小時後，測其速讀成績，得實驗組和控制組的樣本平均數分別為 70 和 60，樣本變異數分別為 100 和 50，求實驗組與控制組的平均數差異的 95% 信賴區間。已知實驗組和控制組的速讀成績均呈常態分佈，且變異數相等。

兩平均數差異區間估計

母體變異數未知但相等：小樣本

當兩母體是常態分佈，樣本數很小，若可以假設兩母體的變異數和雖未知但卻相等，那麼

$$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}}$$

是自由度 n1+n2−2 的 t 分佈

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

稱為合併的變異數

第二節 兩平均數差異區間估計 (14)

合併的變異數為

$$S_p^2=\frac{4\times100+4\times50}{5+5-2}=75$$

$$(70-60)-2.306\times\sqrt{\frac{75}{5}+\frac{75}{5}}<\mu_1-\mu_2<(70-60)+2.306\times\sqrt{\frac{75}{5}+\frac{75}{5}}$$

由於這段區間（−2.63, 22.63）包含了 0，因此實驗組的母體平均數有可能等於控制組的平均數。

兩平均數差異區間估計

母體變異數未知且不等：小樣本

如果常態分佈母體的變異數未知，而且也不相等，當小樣本時，

$$T=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

並不是 t 分佈，只是近似 t 分佈，且自由度為：

$$v=\frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/\left(n_1-1\right)+\left(s_2^2/n_2\right)^2/\left(n_2-1\right)}$$

如果是大樣本，可以比較不關心兩母體是否為常態分佈，因為可以仰賴中央極限定理。

如果是小樣本，兩母體就必須是常態分佈。萬一兩母體不是常態分佈，樣本數又很小，並不適合用此處的區間估計方式，應該改用無母數（ non-parametric ）統計方式。

兩平均數差異區間估計（例子）

承上例，如果我們懷疑兩常態母體的變異數並不相等，求實驗組與控制組的平均數差異的 95% 信賴區間。

作法

$$v=\frac{(100/5+50/5)^2}{(100/5)^2/(5-1)+(50/5)^2/(5-1)}=7.2$$

$$(70-60)-2.365\times\sqrt{\frac{100}{5}+\frac{50}{5}}<\mu_1-\mu_2<(70-60)+2.365\times\sqrt{\frac{100}{5}+\frac{50}{5}}$$

paired samples 平均數區間估計

成對觀測值的平均數差異

如果兩個樣本是成對地發生，那麼這兩個樣本必定有關連，而非兩個獨立樣本。這種成對觀測值（又稱相依樣本， paired samples or dependent samples ）平均數差異的區間估計和上述兩獨立樣本有所不同。

paired samples

將每一對的數值相減，稱為 d1 ， ， dn ，這些差異均可視為來自隨機樣本 D1 ， ， Dn 的值。而這些隨機樣本是從平均數 mD = m1 − m2 ，和變異數 σ_D^2 的常態分佈母體抽樣而來。

用 S_D^2 取代 σ_D^2 ，

$$T=\frac{\bar{D}-\mu_D}{\sqrt{S_D^2/n}}$$

是自由度為 n-1 的 t 分佈。

paired sample 兩平均數差異區間估計

研究者關心夫妻的智力會有多大的差異，他隨機抽取了 10 對夫妻。估計夫妻間智力平均數差異的 95% 信賴區間。

編號	1	2	3	4	5	6	7	8	9	10	平均數	標準差
夫	120	110	95	100	105	100	125	90	85	95	102.50	12.75
妻	125	115	90	120	115	95	115	100	80	80	103.50	16.67
夫－妻	-5	-5	5	-20	-10	5	10	-10	5	15	-1.00	9.72

$$-1-2.26\times\frac{9.72}{\sqrt{10}}<\mu_D<-1+2.26\times\frac{9.72}{\sqrt{10}}$$

paired samples

$$1-\alpha=P\left[-t_{\alpha/2}<\frac{\bar{D}-\mu_D}{\sqrt{S_D^2/n}}<t_{\alpha/2}\right]$$

$$¿\,P\left[\bar{D}-t_{\alpha/2}\frac{S_D}{\sqrt{n}}<\mu_D<\bar{D}+t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right]$$

mD 的（ 1−a ） 100% 的信賴區間

$$\bar{d}-t_{\alpha/2}\sqrt{\frac{s_D^2}{n}}<\mu_D<\bar{d}+t_{\alpha/2}\sqrt{\frac{s_D^2}{n}}$$

變異數的區間估計

一母體變異數

$\frac{(n-1)S^2}{\sigma^2}$ 是自由度為 n − 1 的卡方分佈，

$$1-\alpha=P\left[\chi^2_{1-\alpha/2}<\frac{(n-1)S^2}{\sigma^2}<\chi^2_{\alpha/2}\right]$$

$$¿\,P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right]$$

σ 2 的 (1—a) 100% 的信賴區間

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}}<\sigma^2<\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

σ^2的信賴區間

適用時機： 資料為常態分配時

雙側信賴區間

$$P\left[\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] = (1-\alpha)100$$

右側信賴區間

$$P\left[\sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}}\right] = (1-\alpha)100$$

左側信賴區間

$$P\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha}} < \sigma^2\right] = (1-\alpha)100$$

P的信賴區間

適用時機：當 np 至少大於 5 或 $n(1-p)$ 至少大於 5 時

雙側信賴區間

$$P\left[\overline{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} < p < \overline{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] = (1-\alpha)100$$

右側信賴區間

$$P\left[p < \overline{p} + Z_{\alpha}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right] = (1-\alpha)100$$

左側信賴區間

$$P\left[\overline{p} - Z_{\alpha}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} < p\right] = (1-\alpha)100$$