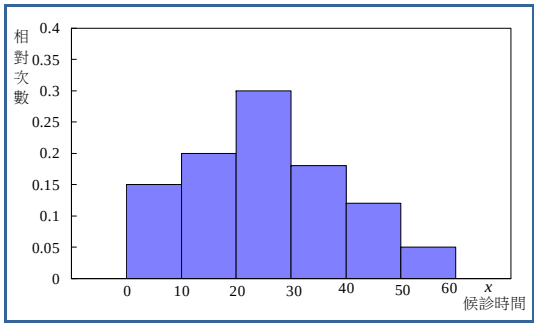
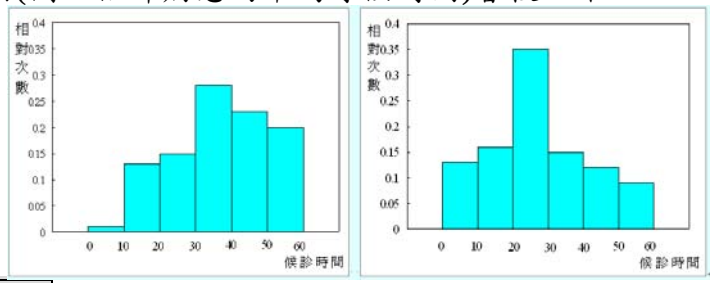
第 10 章　簡單隨機抽樣與抽樣分配

- **統計推論的目的**：從已知的、部分的樣本特性，去推論未知的、整體的母體特性。統計推論的結果是否精確，受到三個因素的影響：(1)樣本數的大小；(2)抽樣方法；(3)推論方法。
- **抽樣的重要性**：人們在研究某些問題或現象時，有時並不直接探討母體，而係經由對樣本的研究分析，以獲致某些樣本統計量，然後再利用這些樣本統計量去推測母體的參數，主要是因為：(1)有限的資源，為了節省時間與經費；(2)毀壞性的實驗，實驗完後物品已無法再使用；(3)概念性的母體，無法全部觀察，(4)樣本較母體小，在資料搜集與整理時較容易且較精確。
  - 統計推論係利用樣本統計量去推論母體的特質，而樣本是否具有代表性[可以無偏地代表母體]會受抽樣方法的影響，因此抽樣方法非常重要。

- 在抽樣時，除了樣本具有母體代表性這個條件外，我們通常喜好所取得的樣本為**隨機樣本**(random sample)，因為隨機樣本可以簡化統計推論的方法。
- 例子：假設某牙醫診所去年診察病患 4,500 人，這些病患等候看診時間的相對次數直方圖如下[這就是所謂的母體分配：母體分配一般而言是看不到的，若看得到就不會有統計推論的問題了；這裡假設可以看得到母體分配僅是為了說明代表性樣本的概念]



- 『樣本具有母體代表性』指的是『樣本的分配特性與母體幾乎相同』。
- 假設我們從所有病患中，抽出兩組樣本數各 50 人的樣本，兩組樣本的直方圖如下所示。由直方圖可以看出，右圖的樣本分配與母體分配較接近，若用這組樣本來進行統計推論(例如估計病患的平均等候時間)會較正確。



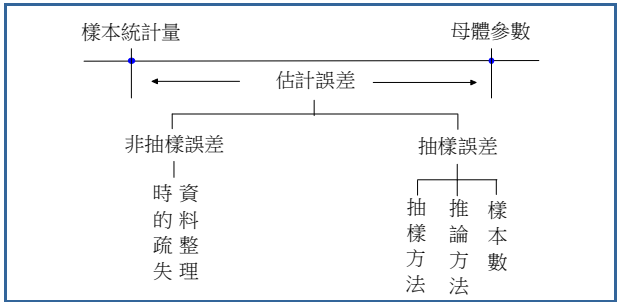
- 註：抽樣的困難之處在於我們並無法知道母體的分配，因此無法在抽樣之後判斷樣本是否具有母體代表性。所以在抽樣前就必須設計出一個可以抽出代表性樣本的方法。

抽樣誤差(sampling error)與非抽樣誤差(nonsampling error)

- 當我們利用樣本統計量去推論母體參數時[例如想知道 4,500 個病患的平均候診時間，但利用抽出的 50 人樣本所計算的樣本統計量來估計]，由於母體中的各個元素本身即有所不同，因此不管如何抽樣，樣本統計量與母體參數之間總會有一些差異，這個差異就稱為**估計誤差**(estimation error)。
  - 估計誤差可再細分為抽樣誤差與非抽樣誤差。
- **抽樣誤差**：因為抽樣所造成的誤差。此種誤差來自於抽樣過程的機遇(chance)、抽樣方法、及推論方法的不同。
  - 由於母體的元素各不相同，根據樣本計算的樣本統計量本來就有機會不等於母體參數，這種純粹是因為抽樣的問題所產生的誤差稱為**隨機抽樣誤差**(random sampling error)。**隨機抽樣誤差**可以透過樣本個數的增加來降低。
  - 抽樣方法本身如果就不易抽出代表性樣本，產生抽樣誤差當然也就不意外了。

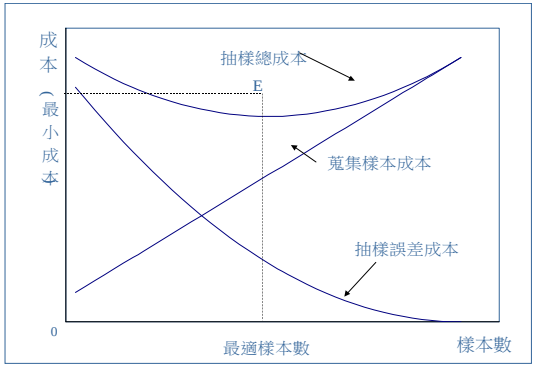
- 統計推論的方法是否恰當也會影響到估計誤差，例如：若母體的分配為常態，則用樣本平均數或樣本中位數來估計母體平均數應是相同的[但我們還是偏好樣本平均數]；但若母體的分配是非對稱分配，則利用樣本中位數來估計母體平均數所產生的估計誤差會比樣本平均數大。

- **非抽樣誤差**：非抽樣誤差主要來自調查時的執行與事後在記錄、整理資料時所發生的錯誤[例如：數據紀錄錯誤、受訪者亂答]。
  - 我們通常假設這種誤差不存在。



抽樣成本與抽樣誤差

- 要降低抽樣誤差，最好的方法就是進行母體普查，但是普查的成本相對於抽樣而言是很高的。但樣本數太小又可能造成抽樣誤差太大，亦可能帶來極大的成本(推論錯誤造成決策錯誤)。
- 若把兩個成本同時考量進來，理論上應該會有一個最適的抽樣樣本數。但實際上，最適樣本數應該有多少則是見仁見智了。



簡單隨機抽樣

- 簡單隨機抽樣：簡單隨機抽樣是指抽樣母體中所有可能被抽出的樣本組被抽出的機率均相等的抽樣方法。[或是說：每個母體元素被抽到的機率均相同，而且所有樣本均是被**獨立**地抽出]
  - 簡單隨機抽樣的實施方式
    ①抽籤式
    ②以亂數表抽取樣本
    ③用電腦做隨機抽樣
  - 簡單隨機抽樣所抽出的樣本稱為**隨機樣本**(random sample)

抽樣分配(sampling distribution)

- 若我們對母體特性(分配型態、母體平均數、母體變異數)做一些假設，則在還沒進行抽樣之前，我們即可預知根據抽樣的樣本所計算出的樣本統計量會呈現何種分配，而這種分配即稱為**抽樣分配**。[假設我們要對某個母體進行隨機抽樣，在真正開始進行抽樣，得到一組樣本前，我們即可知道樣本統計量具備何種分配]

- **母體參數**：母體參數是描述母體資料特性的統計測量數，一般簡稱為參數或母數。參數是我們想要得知的，是統計的核心。
- **樣本統計量**：樣本統計量為樣本的實數值函數。通常用來描述樣本資料特性，或用來推論母體參數。
  - 若假設 $X_1,...,X_n$ 為自母體 X 中抽出的隨機樣本，定義實數值函數 $T(X_1,...,X_n)$，則 T 為一樣本統計量(觀察到樣本即可算出一個確定數值的公式)。若還沒有開始真正進行抽樣，則 $X_1,...,X_n$ 到底會出現何種數值還是未知，因此將隨機樣本 $X_1,...,X_n$ 視為 n 個(獨立的)隨機變數，故樣本統計量 $T(X_1,...,X_n)$ 亦為隨機變數[故樣本統計量會具有分配]。
- 例子：想要瞭解台灣全體小學生的平均身高[母體參數]，自全體小學生中抽出 100 個學生測量其身高，令樣本為 $X_1,X_2,...,X_{100}$，則 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 、 $\hat{X}=\max_i(X_i)$ 、 $\hat{\mu}=\frac{1}{2}(X_1+X_{100})$ 均為**樣本統計量**

- **抽樣分配**：樣本統計量為隨機樣本的函數，而隨機樣本是由 n 個獨立隨機變數 $X_1, X_2, \cdots X_n$ 所組成的，故樣本統計量亦為一隨機變數，其機率分配稱為抽樣分配。
  - 例子：樣本平均數 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 為隨機樣本 $X_1, X_2, \cdots X_n$ 的函數，故 $\overline{X}$ 為隨機變數，$\overline{X}$ 的分配稱為抽樣分配。抽樣分配的特性會取決於母體分配的特性或母體參數。
- 例子：虛構一母體，自該母體中抽樣，看樣本平均數的抽樣分配
  - 假設某汽車公司共 5 位接待小姐(母體)的薪資(千元)分別為

    22　25　25　28　30

    底下整理出母體資料的次數分配(左)與相對次數分配(右)

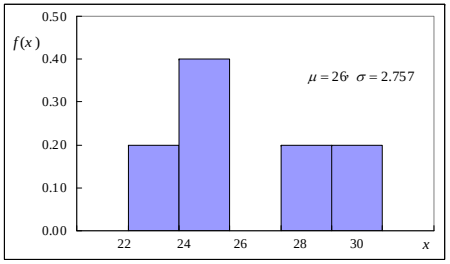
x	f
22	1
25	2
28	1
30	1
	$N=5$

x	$f(x)$
22	$1/5=0.2$
25	$2/5=0.4$
28	$1/5=0.2$
30	$1/5=0.2$
	$\sum f(x)=1$

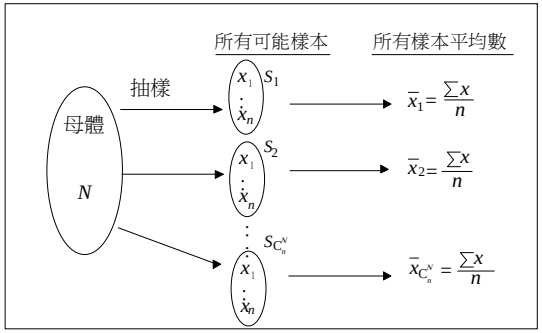
  - 根據以上次數分配，可計算母體的平均數與變異數分別為

$$\mu = \frac{1}{5}\times 22 + \frac{2}{5}\times 25 + \frac{1}{5}\times 28 + \frac{1}{5}\times 30 = 26$$

$$\sigma^2 = \frac{1}{5}\times(22-26)^2 + \frac{2}{5}\times(25-26)^2 + \frac{1}{5}\times(28-26)^2 + \frac{1}{5}\times(30-26)^2 = 7.6$$



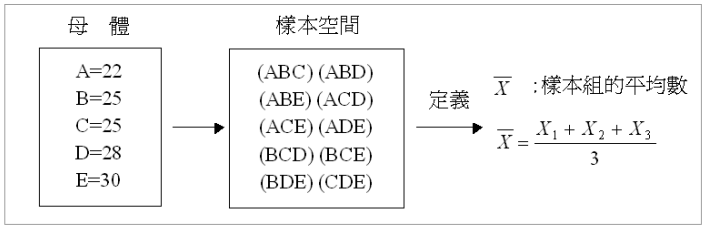
  - 若母體有 N 個元素，抽出 n 個為一組樣本，則所有可能的樣本組有 C_n^N 個。根據這 C_n^N 組樣本，我們就可計算出 C_n^N 個樣本平均數 $\overline{X}$，進而觀察 $\overline{X}$ 的分配狀況。



  - 若抽樣方法為簡單隨機抽樣，則每組樣本出現的機率都為 $1/C_n^N$，因此這 C_n^N 組樣本的樣本平均數 $\overline{X}$ 之分配為

$\overline{x}$	$f(\overline{x})$
$\overline{x}_1$	$1/C_n^N$
$\overline{x}_2$	$1/C_n^N$
$\vdots$	$\vdots$
$\overline{x}_{C_n^N}$	$1/C_n^N$
$\overline{X}$ 的平均數與變異數	$E(\overline{X})$ 、$V(\overline{X})$

  - 在展示小姐薪資的例子中，$N=5$、$n=3$，因此可能樣本組共有 $C_3^5 = 10$ 個。若令 $A=22$、$B=25$、$C=25$、$D=28$、$E=30$，所有可能樣本組應如下圖所示

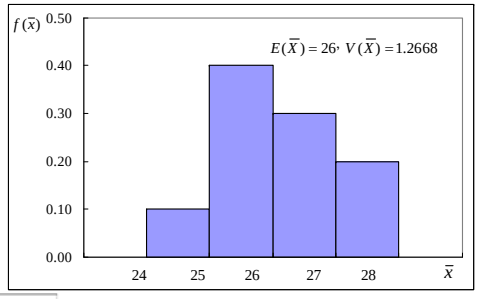


  - 這 10 組樣本的樣本平均數可計算如下頁上表，並可進一步整理成下頁下表的 $\overline{X}$ 之抽樣分配表[現在你已經看到了樣本平均數的抽樣分配(直方圖)]。$\overline{X}$ 之平均數與變異數為

$$E(\overline{X}) = 24\times\frac{1}{10} + 25\times\frac{2}{10} + 25.6667\times\frac{2}{10} + 26\times\frac{1}{10} + 26.6667\times\frac{2}{10} + 27.66767\times\frac{1}{10} = 26$$

$$V(\bar{X}) = (24-26)^2 \times \tfrac{1}{10} + (25-26)^2 \times \tfrac{2}{10} + (25.6667-26)^2 \times \tfrac{2}{10} + (26-26)^2 \times \tfrac{1}{10}$$
$$+ (26.6667-26)^2 \times \tfrac{2}{10} + (27.66767-26)^2 \times \tfrac{1}{10} = 1.2668$$

樣本	樣本平均數 $\bar{x}$
$(ABC) = (22,25,25)$	24
$(ABD) = (22,25,28)$	25
$(ABE) = (22,25,30)$	25.6667
$(ACD) = (22,25,28)$	25
$(ACE) = (22,25,30)$	25.66667
$(ADE) = (22,28,30)$	26.66667
$(BCD) = (25,25,28)$	26
$(BCE) = (25,25,30)$	26.66667
$(BDE) = (25,28,30)$	27.66667
$(CDE) = (25,28,30)$	27.66667



$\bar{x}$	$\bar{x}^2$	$f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
24	576	$1/10 = 0.10$	57.6
25	625	$2/10 = 0.20$	125
25.6667	659.7779	$2/10 = 0.20$	131.7556
26	676	$1/10 = 0.10$	67.6
26.6667	711.1113	$2/10 = 0.20$	142.2223
27.6667	765.4446	$2/10 = 0.20$	159.0999
		$\sum f(\bar{x}) = 1.00$	677.2668

樣本平均數的抽樣分配(sampling distribution)

- **樣本平均數的抽樣分配**：設母體為隨機變數 X，其機率分配為 $f(x)$，若自母體中簡單隨機抽取 n 個元素為一組樣本，表為 $(X_1, X_2, ..., X_n)$，若令 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，則 $\bar{X}$ 為樣本平均數。其機率分配表為 $f(\bar{x})$，稱為樣本平均數的抽樣分配。
  - 若我們已經取得一組樣本[例如前述例子中之樣本 $(A,B,C) = (22,25,25)$]，則可以算出一個確定的平均數 ($\bar{X} = 24$)，此時當然沒有所謂的樣本平均數的分配問題。
  - 樣本平均數(或任何樣本統計量)的抽樣分配，是站在還沒有進行抽樣前的角度來看樣本平均數的分配；因為還沒進行(隨機)抽樣，我們不知道會抽出怎麼樣的樣本，因此將可能會抽出的樣本視為一組隨機變數 $(X_1, X_2, ..., X_n)$，既然樣本平均數是根據樣本所計算出來的 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，所以 $\bar{X}$ 亦為隨機變數，故我們可以探討 $\bar{X}$ 的(抽樣)分配。

- $\bar{X}$ 的平均數與變異數：$\bar{X}$ 之抽樣分配的平均數與變異數稱為 $\bar{X}$ 的平均數與變異數。以符號 $\mu_{\bar{X}}$ 或 $E(\bar{X})$ 及 $\sigma^2_{\bar{X}}$ 或 $V(\bar{X})$ 分別表示。
  - 假設 $(X_1, X_2, ..., X_n)$ 為自母體 $X \sim (\mu, \sigma^2)$ 中隨機抽取的樣本[將 $(X_1, X_2, ..., X_n)$ 視為 n 個具有相同平均數 μ 與變異數 σ^2 的獨立隨機變數]
  - $\bar{X}$ 之**抽樣分配的平均數**：$\bar{X}$ 抽樣分配的平均數等於母體平均數，即 $E(\bar{X}) = \mu_{\bar{X}} = \mu$ ；因為

$$E(\bar{X}) = E(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) = \tfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \tfrac{1}{n}\sum_{i=1}^{n} \mu = \tfrac{1}{n}(n\mu) = \mu$$

  - $\bar{X}$ 之**抽樣分配的**的變異數($\sigma^2_{\bar{X}}$)與標準差($\sigma_{\bar{X}}$)

(1) 若母體的元素個數有無窮多個(**無限母體**)，則 $\bar{X}$ 之變異數與標準差分別為

$$\boxed{\sigma^2_{\bar{X}} = V(\bar{X}) = \tfrac{\sigma^2}{n}} \qquad \boxed{\sigma_{\bar{X}} = \sqrt{\tfrac{\sigma^2}{n}}}$$

證明：由於 $X_1, X_2, ..., X_n$ 為獨立隨機變數，因此

$$V(\bar{X}) = V(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) = \tfrac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \tfrac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$
$$= \tfrac{1}{n^2}(n\sigma^2) = \tfrac{\sigma^2}{n}$$

(2) 若母體的元素個數有限(**有限母體**)，則 $\bar{X}$ 之變異數與標準差分別為(N 為母體元素個數)

$$\sigma^2_{\bar{X}} = V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \qquad \sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

註 1: $\frac{N-n}{N-1}$ (<1)稱為**有限母體修正因子**(finite population correction factor: FPF)；有限母體下之樣本平均數抽樣分配的變異數會比無限母體小。

註 2：若 N 相對於 n 而言很大的話，則 $\frac{N-n}{N-1} \approx 1$，可將 FPF 忽略不計[實際上都是這種情況，故 FPF 不重要]

- 既然我已經知道了 $\bar{X}$ 的平均數與標準差，根據柴比氏定理，我們已經可以約略估算樣本平均數 $\bar{X}$ 位於某個區間的機率，但這個機率太過粗略，我們有沒有辦法得到比較精確的結果？當然有，我們區分成底下兩種情況討論
  - 假設母體為常態分配：實際分配(exact distribution)
  - 不假設母體為常態分配：漸近分配(asymptotic distribution)
- **常態母體**之樣本平均數 $\bar{X}$ 的抽樣分配：
  - 若母體 X 為常態分配，平均數為 μ，標準差為 σ，亦即 $X \sim N(\mu, \sigma^2)$，則從母體 X 中所抽出的樣本 $(X_1, X_2, ..., X_n)$ 可視為 n 個獨立的常態隨機變數，每個隨機變數都具有 $X_i \sim N(\mu, \sigma^2)$ 的分配。
  - 由於樣本平均數 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，為常態隨機變數的線性函數，因此 $\bar{X}$ 亦具有常態分配。

  - 由於 $\bar{X}$ 的平均數為 $\mu_{\bar{X}} = \mu$、變異數為 $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$，搭配上 $\bar{X}$ 為常態分配的性質，我們可知道

    $$\bar{X} \sim N(\mu, \tfrac{\sigma^2}{n})$$

  - 這個抽樣分配稱為**實際分配**，只要我們假設母體具有常態分配，不管樣本數 n 是多少[這與不假設母體為常態時之狀況不同]，樣本平均數 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$。
  - 欲得到精確分配，一定得假設母體具常態分配。
- 例子：假設鹿谷鄉所產的 1 斤裝茶葉之重量呈常態分配，其平均數為 600g、標準差為 9.5g。若現在從鹿谷鄉所產的茶葉中隨機抽出 n 包，並計算其樣本平均數 $\bar{X}$，則 $\bar{X}$ 之分配為何？
  - 若 $n = 25$：因母體為常態分配，故 $\bar{X}$ 亦為常態分配；又因 $\mu_{\bar{X}} = \mu = 600$、$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{9.5^2}{25} = 3.61$，因此 $\bar{X} \sim N(600, 3.61)$

  - 若 $n = 49$：因 $\mu_{\bar{X}} = \mu = 600$、$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{9.5^2}{49} = 1.842$，因此 $\bar{X} \sim N(600, 1.842)$
  - 若 $n = 100$：因 $\mu_{\bar{X}} = \mu = 600$、$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{9.5^2}{100} = 0.9025$，因此 $\bar{X} \sim N(600, 0.9025)$
  - 有沒有發現：樣本個數 n 越多，$\bar{X}$ 抽樣分配的變異數越小；這就表示，當樣本個數 n 越多，樣本平均數與母體平均數較接近的機率會比較高(抽樣誤差較小)。
- **非常態母體**之樣本平均數 $\bar{X}$ 的抽樣分配：如果母體的分配並非常態，則我們無法得知樣本平均數 $\bar{X}$ 的精確分配，但我們可以知道 $\bar{X}$ 的漸近分配(樣本個數 $n \to \infty$ 時之抽樣分配)
  - 由於 $\bar{X}$ 的平均數為 $\mu_{\bar{X}} = \mu$、變異數為 $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$，若將 $\bar{X}$ 標準化後可得 $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$，而且 $E(Z) = 0$、$V(Z) = 1$。

  - **中央極限定理**(Central Limit Theorem; CLT)：假設有一母體之平均數為 μ、變異數為 $\sigma^2 < \infty$，若自母體隨機抽取一組樣本 $(X_1, X_2, ..., X_n)$，則當樣本個數 $n \to \infty$ 時，

    $$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

  - 中央極限定理的涵義：無論母體為何種分配(但要求變異數 $\sigma^2 < \infty$)，只要樣本個數 $n \to \infty$ 時，標準化的樣本平均數就會具有標準常態分配(這個分配稱為漸近分配)。
  - 中央極限定理的實際應用：實務上我們當然不可能取得樣本個數為 ∞ 的樣本，因此在實際應用時，只要樣本個數 n 夠大，就將標準化的樣本平均數視為具有標準常態分配

    $$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

而這個分配隱含，只要樣本個數 n 夠大

$$(\bar{X}-\mu)\sim N(0,\tfrac{\sigma^2}{n}) \Leftrightarrow \bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})$$

亦即，不管母體分配為何，只要樣本個數 n 夠大，我們可用 $\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})$ 來逼近 $\bar{X}$ 的抽樣分配[近似的分配]。

- 什麼叫做『樣本個數 n 夠大』:
  (1) CLT 所描述的是在樣本個數 $n\to\infty$ 時才會成立的分配，因此樣本個數 n 越多，樣本平均數標準化後的分配愈接近標準常態分配。
  (2) 但 n 到底需要多大，樣本平均數標準化後的分配才會接近標準常態分配，則取決於母體原先的分配；不同的母體分配會有不同的樣本個數要求[有些分配所要求的 n 很小；但有些分配儘管 n 已經很大，卻還是無法近似]。
  (3) 課本上(還有很多其他教科書也一樣)定義 $n\geq 30$ 就是『樣本個數 n 夠大』(大樣本)是沒有意義的。

- **總結**：根據以上說明可知，樣本平均數的抽樣分配到底該用實際分配還是漸近分配完全取決於我們對母體分配的假設
  - **實際分配**：假設隨機樣本 $(X_1,X_2,...,X_n)$ 來自於 $X\sim N(\mu,\sigma^2)$ 的常態母體，則不管樣本個數 n 為何，樣本平均數 $\bar{X}$ 的抽樣分配為

    $$\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})$$

  - **漸近分配**：若僅知母體平均數為 μ、變異數為 $\sigma^2<\infty$，但不知母體分配形態，只要樣本個數 n 夠大，則樣本平均數 $\bar{X}$ 的抽樣分配可以底下分配來近似($\bar{X}$ 之分配近似常態分配)

    $$\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})$$

  - 雖然實際分配與漸近分配看來完全相同，但其本質上確有很大的差異。實際分配是『精確的抽樣分配』，但漸近分配僅是『近似的抽樣分配』(樣本個數夠多時近似才會較佳)。

- 例子：假設高雄市計程車每天的燃料費用為一常態分配，平均數為 313 元，標準差為 101 元。
  - 令 X 代表高雄市計程車每天的燃料費用，則由題意可知

    $$X\sim N(313,101^2)$$

  - 隨機抽取一台計程車，其燃料費超過 400 元的機率為

    $$P(X>400)=P(\tfrac{X-\mu}{\sigma}>\tfrac{400-313}{101})=P(Z>0.86)=0.1949$$

  - 若某家車行有 8 輛計程車，則 8 輛計程車平均燃料費超過 400 元的機率為何？

    8 輛計程車平均燃料費 $\bar{X}$ 的抽樣分配為

    $$\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})\Rightarrow \bar{X}\sim N(313,\tfrac{101^2}{8})$$

    故平均燃料費超過 400 元的機率為

    $$P(\bar{X}>400)=P(\tfrac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}>\tfrac{400-313}{\sqrt{101^2/8}})=P(Z>2.44)=0.0073$$

    [這題還問了『8 輛計程車平均燃料費為多少？』怪怪的]

該車行每天燃料費少於 2000 元的機率為何？

令 $S=\sum_{i=1}^{8}X_i$ 代表該車行每天燃料費總和，則 $S=8\bar{X}$。因 $\bar{X}\sim N(313,\tfrac{101^2}{8})$，故 S 亦為常態分配，其平均數 $\mu_S=8\mu_{\bar{X}}$、變異數 $\sigma_S^2=64\sigma_{\bar{X}}^2$，亦即

$$S\sim N(8\times 313,64\times\tfrac{101^2}{8})=N(2504,81608)$$

車行每天燃料費少於 2000 元的機率為

$$P(S<2000)=P(\tfrac{S-\mu_S}{\sigma_S}<\tfrac{2000-2504}{\sqrt{81608}})=P(Z<-1.76)=0.039$$

- 若計程車每天燃料費用的分配未知，隨機抽查 60 輛計乘車每天的燃料費，則總和介於 1.7 萬和 2 萬間的機率為何？

  解答：令 $\bar{X}$ 代表這 60 輛計程車燃料費的樣本平均，因母體分配未知，只好以底下分配來近似 $\bar{X}$ 的抽樣分配

  $$\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})\Rightarrow \bar{X}\sim N(313,\tfrac{101^2}{60})$$

令 $T=\sum_{i=1}^{60} X_i$ 代表該這 60 輛計乘車每天燃料費總和，則 $S=60\bar{X}$。因 $\bar{X}$ 近似 $N(313,\frac{101^2}{60})$，故 T 亦近似常態分配，其平均數 $\mu_T=60\mu_{\bar{X}}$、變異數 $\sigma_T^2=3600\sigma_{\bar{X}}^2$，亦即

$$T\sim N(60\times 313, 3600\times\tfrac{101^2}{60})=N(18780, 782.3^2)$$

60 輛計乘車每天燃料費總和介於 1.7 萬和 2 萬間的機率為

$$P(17000<T<20000)=P(\tfrac{17000-18780}{782.3}<\tfrac{T-\mu_T}{\sigma_T}<\tfrac{20000-18780}{782.3})$$

$$=P(-2.27<Z<1.56)=0.929$$

- 例子[10.16]：(抽樣分配與統計推論的聯結)劉大、許二、張三、吳四、劉七 5 個人合夥開餐廳，由許二與張三負責經營。餐廳滿座時每日營業收入可達 30000 元，但一年下來，公司的損益表卻顯示每日營業收入的平均值為 18000、標準差為 3000 元，與其他股東的認知有極大差異。

  - 吳四認為許二與張三可能中飽私囊，因此採分層抽樣，抽取 36 個營業日，算出樣本的平均每日營業額為 26000。請問許二與張三是否中飽私囊？
  - 若許二與張三並未中飽私囊，則 36 個營業日之營業額的樣本平均數 $\bar{X}$ 應可以底下分配近似(因未假設母體為常態)

$$\bar{X}\sim N(\mu,\tfrac{\sigma^2}{n})\Rightarrow \bar{X}\sim N(18000,\tfrac{3000^2}{36})=N(18000,500^2)$$

  - 若許二與張三並未中飽私囊，則 $\bar{X}$ 應該有很高的機率會超過 26000；反之，若 $\bar{X}$ 比 26000 大的機率很低，就代表許二與張三可能中飽私囊。由於

$$P(\bar{X}>26000)=P(\tfrac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}>\tfrac{26000-18000}{500})=P(Z>16)=0.0000$$

$\bar{X}$ 比 26000 大的機率確實很低，因此我們可以斷定『許二與張三可能中飽私囊』

樣本比例的抽樣分配

- 問題描述：假設有一母體有 N 個元素，其中 K 個為 A 類別，亦即母體中 A 類別的比例為 $p=K/N$。若從母體中抽出一組個數為 n 的樣本，則樣本中 A 類別所佔比例的抽樣分配為何？
  - 例子：已知一批產品中瑕疵品的比例(母體比例)為 p，從該批產品中抽出個數為 n 的一組樣本，則樣本中瑕疵品的比例(樣本比例)之抽樣分配為何？
- 點二項分配(柏努利分配)：若從母體中抽出 1 個元素，可將這個抽樣的結果用一個柏努利隨機變數 X 來描述。若抽出的元素為 A 類別，則 $X=1$；若抽出的元素非 A 類別，則 $X=0$。由於母體中 A 類別的比例為 p，因此柏努利隨機變數 X 的平均數與變異數分別為

$$E(X)=1\times p+0\times(1-p)=p$$

$$V(X)=(1-p)^2\times p+(0-p)^2\times(1-p)=p(1-p)$$

- 樣本比例：從母體中抽出一組個數為 n 的樣本，若我們將每一個樣本的結果以柏努利隨機變數 X_i [出象只有 1 與 0 兩種可能]來表示，則該組樣本可表達為 $(X_1,X_2,...,X_n)$；若從還未真正進行抽樣的觀點來看，應將 $(X_1,X_2,...,X_n)$ 視為 n 個隨機變數。
  - 若已知該組樣本中 A 類別元素有 k 個($k=0,1,...,n$)，則樣本比例 $\hat{p}=k/n$ 可用另一種方式來表達。[在此須假設 $n\le K$]
  - 由於 A 類別元素有 k 個，所以 $(X_1,X_2,...,X_n)$ 中應有 k 個 1、$n-k$ 個 0，因此 $X_1+X_2+\cdots+X_n=k$，故樣本比例可表達為 n 個柏努利隨機變數的平均

$$\hat{p}=\frac{k}{n}=\frac{\sum_{i=1}^{n}X_i}{n}$$

  - 由於 $(X_1,X_2,...,X_n)$ 為 n 個隨機變數，而 $\hat{p}$ 為 $(X_1,X_2,...,X_n)$ 的函數，故亦為隨機變數；因此，可探討 $\hat{p}$ 的抽樣分配

- $\hat{p}$ 應具什麼抽樣分配：在間斷隨機變數中，我們講過兩個有關類別資料之隨機變數的分配：二項分配(抽出後放回)與超幾何分配(抽出後不放回)。[注意：$\hat{p}$ 的數值取決於 A 類別個數 k]
  - 所以你應該可以猜到，$\hat{p}$ 的抽樣分配應是兩者其中之一。
  - 若抽樣時採取抽出後放回(或母體元素個數 N 為無窮大)，則 $(X_1, X_2, ..., X_n)$ 可視為 n 個獨立的柏努利隨機變數(成功機率均為 p)，此時 $\hat{p}$ 具有二項分配。
  - 若抽樣時採取抽出後不放回，則 $(X_1, X_2, ..., X_n)$ 並非 n 個獨立的柏努利隨機變數，此時 $\hat{p}$ 具有超幾何分配。但若母體元素個數 N 相對於樣本個數 n 而言很大，則抽出後不放回近似抽出後放回，此時 $\hat{p}$ 的分配近似二項分配。
  - 若試行次數 n 很大時，二項分配會趨近常態分配，因此，當樣本個數 n 很大時，$\hat{p}$ 的抽樣分配會近似常態分配(這個結果還是根據中央極限定理而來的，底下會說明)

- **母體元素個數 N 為無窮大(或抽出後放回)時 $\hat{p}$ 的抽樣分配**
  - 可將 $\hat{p}$ 視為 $(X_1, X_2, ..., X_n)$ 等 n 個獨立的柏努利隨機變數的平均數。$\hat{p}$ 應具二項分配，其分配函數為

    $$P(\hat{p} = \tfrac{k}{n}) = C_k^n p^k (1-p)^{n-k}, k = 0,1,...,n$$

    說明：當 $\hat{p} = \frac{k}{n}$ 時，代表 n 個樣本中 A 類別元素有 k 個，而抽出了 n 個元素中有 k 個為 A 類別的機率為 $C_k^n p^k (1-p)^{n-k}$，因此 $P(\hat{p} = \frac{k}{n}) = C_k^n p^k (1-p)^{n-k}$ 。
  - $\hat{p}$ 之抽樣分配的平均數為 $E(\hat{p}) = \mu_{\hat{p}} = p$

    $$E(\hat{p}) = E(\tfrac{1}{n}\sum_{i=1}^n X_i) = \tfrac{1}{n}\sum_{i=1}^n E(X_i) = \tfrac{1}{n}\sum_{i=1}^n p = \tfrac{1}{n}(np) = p$$
  - $\hat{p}$ 之抽樣分配的變異數為 $V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{pq}{n}$ ，其中 $q = 1-p$

    $$V(\hat{p}) = V(\tfrac{1}{n}\sum_{i=1}^n X_i) = \tfrac{1}{n^2} V(\sum_{i=1}^n X_i) = \tfrac{1}{n^2}\sum_{i=1}^n V(X_i)$$
    $$= \tfrac{1}{n^2}\sum_{i=1}^n pq = \tfrac{1}{n^2}(npq) = \tfrac{pq}{n}$$

- 例子：若有 1 枚銅板出現正面的機率為 p，丟這枚銅板 5 次，以 $\hat{p}$ 來代表出現正面的比例，則 $\hat{p}$ 的抽樣分配為何？

  解答：由於每次投擲銅板出現正面的機率均為 p，而且不互相影響(獨立)，因此 $\hat{p}$ 的抽樣分配為二項分配

  $$P(\hat{p} = \tfrac{k}{5}) = C_k^5 p^k (1-p)^{5-k}, k = 0,1,...,5$$

  亦即，$P(\hat{p}) = \begin{cases} C_0^5 p^0 (1-p)^{5-0} & ,\hat{p} = \frac{0}{5} \\ C_1^5 p^1 (1-p)^{5-1} & ,\hat{p} = \frac{1}{5} \\ C_2^5 p^2 (1-p)^{5-2} & ,\hat{p} = \frac{2}{5} \\ C_3^5 p^3 (1-p)^{5-3} & ,\hat{p} = \frac{3}{5} \\ C_4^5 p^4 (1-p)^{5-4} & ,\hat{p} = \frac{4}{5} \\ C_5^5 p^5 (1-p)^{5-5} & ,\hat{p} = \frac{5}{5} \end{cases}$

  而 $\hat{p}$ 之平均數與變異數分別為 $\mu_{\hat{p}} = p$ 與 $\sigma_{\hat{p}}^2 = \frac{pq}{n} = \frac{p(1-p)}{5}$

- **母體元素個數 N 有限(且抽出後不放回)時 $\hat{p}$ 的抽樣分配**
  - 回顧超幾何分配：設有一元素個數為 N 的有限母體，其中 A 類別有 K 個。現在自該母體中抽取 n 個(抽出後不放回)，令 X 為 n 個中 A 類別的個數，則 X 具有超幾何分配

    $$P(X = k) = \frac{C_k^K C_{n-k}^{N-K}}{C_n^N}, \quad k = 0,1,...,n$$

    X 的平均數與變異數分別為[參閱課本 p.204]

    $$E(X) = n \cdot \tfrac{K}{N}$$
    $$V(X) = n \cdot \tfrac{K}{N} \cdot \tfrac{N-K}{N} \cdot \tfrac{N-n}{N-1}$$

    若以 $p = K/N$ 代表母體中 A 類別的比例，則 X 的平均數與變異數可表達為[$\frac{N-n}{N-1}$ 為有限母體修正因子]

    $$E(X) = np$$
    $$V(X) = np(1-p)\tfrac{N-n}{N-1} = npq\tfrac{N-n}{N-1}$$

- 由於母體元素個數 N 有限且抽出後不放回，此時樣本 $(X_1, X_2, ..., X_n)$ 不再是 n 個獨立的柏努利隨機變數，因此 $\hat{p}$ 的分配不再是二項分配，而是超幾何分配。
- 在超幾何分配中，X 代表抽取的 n 個樣本中 A 類別的個數，因此樣本比例可表示為 $\hat{p} = \frac{X}{n}$，當 $X = k$ 時，$\hat{p} = \frac{k}{n}$，故 $P(\hat{p} = \frac{k}{n}) = P(X = k)$，因此 $\hat{p}$ 的抽樣分配之機率函數為

$$P(\hat{p} = \tfrac{k}{n}) = \frac{C_k^K C_{n-k}^{N-K}}{C_n^N}, \quad k = 0,1,...,n$$

- $\hat{p}$ 之抽樣分配的平均數為 $E(\hat{p}) = \mu_{\hat{p}} = p$

$$E(\hat{p}) = E(\tfrac{X}{n}) = \tfrac{1}{n}E(X) = \tfrac{1}{n}(np) = p$$

- $\hat{p}$ 之抽樣分配的變異數為 $V(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{pq}{n}\frac{N-n}{N-1}$

$$E(\hat{p}) = V(\tfrac{X}{n}) = \tfrac{1}{n^2}V(X) = \tfrac{1}{n^2}npq\tfrac{N-n}{N-1} = \tfrac{pq}{n}\tfrac{N-n}{N-1}$$

- 在超幾何分配中我們知道：若母體元素個數 $N \to \infty$ [或 N 相對於 n 而言很大]，則超幾何分配會趨近二項分配[因為抽出後不放回與抽出後放回的抽樣方法差不多]
- 因此，若 $N \to \infty$ [或 N 相對於 n 而言很大]時，樣本比例的抽樣分配都可用二項分配來描述。

- **當樣本個數 $n \to \infty$ 時樣本比例 $\hat{p}$ 的抽樣分配**
  - 此時當然是討論母體元素個數 N 無窮大的狀況，否則怎麼會有無窮大的樣本個數 n。
  - 樣本 $(X_1, X_2, ..., X_n)$ 為 n 個獨立隨機變數，每個 X_i 的平均數為 p、變異數為 pq。或者說『$(X_1, X_2, ..., X_n)$ 是從平均數為 p、變異數為 pq 的母體中抽出的隨機樣本』。
  - 由於 $\hat{p}$ 可表達為 $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$，因此可將 $\hat{p}$ 視為隨機樣本 $(X_1, X_2, ..., X_n)$ 的平均數。

- 中央極限定理說的是：若我們從一個變異數有限的母體中抽出一組隨機樣本，則當樣本個數 $n \to \infty$ 時，標準化的樣本平均數[樣本比例是一種平均數]就會具有標準常態分配。
- 由於 $\mu_{\hat{p}} = p$、$\sigma_{\hat{p}}^2 = \frac{pq}{n}$，因此標準化的樣本比例為 $\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$，根據中央極限定理，則當樣本個數 $n \to \infty$ 時，標準化的樣本比例會具有標準常態分配，亦即

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

- 因此只要樣本個數 n 足夠大($np > 5$ 及 $nq > 5$)，標準化的樣本比例都可用標準常態分配來近似

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \Rightarrow \hat{p} - p \sim N(0, \tfrac{pq}{n}) \Rightarrow \boxed{\hat{p} \sim N(p, \tfrac{pq}{n})}$$

- 連續性調整：欲以常態分配來近似樣本比例 $\hat{p}$ 介於 $\frac{k_1}{n} \le \hat{p} \le \frac{k_2}{n}$ 間的機率，必須進行連續性調整，亦即計算常態分配介於 $\frac{k_1 - 0.5}{n} \le \hat{p} \le \frac{k_2 + 0.5}{n}$ 間的機率。[但實際上的意義已經不大，因為當 n 很大時，$\frac{0.5}{n} \to 0$]
- **註：**在此運用中央極限定理的動機不是因為不知道母體分配，而是因為用二項分配來計算機率太過複雜。

- 例子：96 年全國捐血人次共 1,800,550，捐血的血液為 O 型者佔 44%，A 型者佔 26.6%。若某捐血站每天有 250 人來捐血
  - O 型血所佔比例 $\hat{p}$ 的抽樣分配為：
    因樣本個數 $n = 250$ 已經夠大，所以 $\hat{p}$ 可用常態分配近似

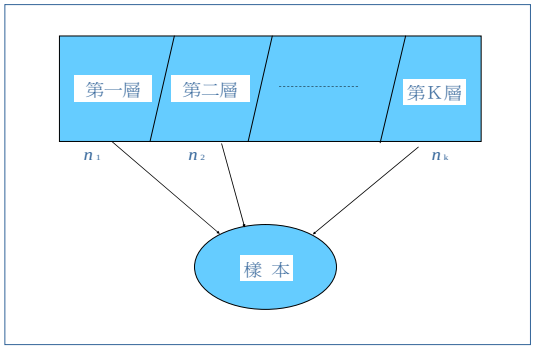
$$\hat{p} \sim N(0.44, \tfrac{0.44 \times 0.56}{250}) = N(0.44, 0.03^2)$$

  - O 型血所佔比例小於 35%的機率為何？

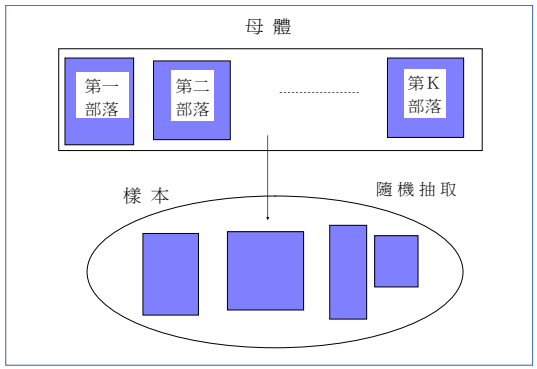
$$P(\hat{p} \le 0.35) = P(\tfrac{\hat{p} - 0.44}{0.03} \le \tfrac{0.35 - 0.44}{0.03}) = P(Z \le -3) = 0.0013$$

其他抽樣方法

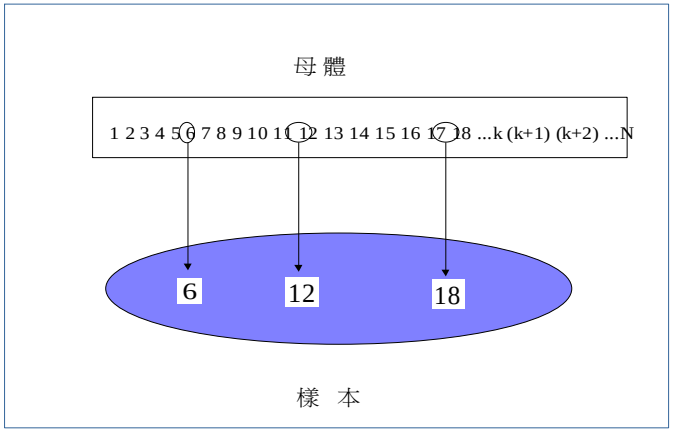
- 分層抽樣的意義：分層抽樣是將母體依其特性或依與調查目的有關的性質分成幾個類或組，母體中的每一個個體或元素都屬於其中的一層，而且是唯一的一層。分層之後再從各層中簡單隨機抽取樣本。[確保每層中都有元素被抽出，以使樣本具有母體代表性]



- 部落抽樣的意義：部落抽樣是先將母體中相鄰的某些群體劃分為 n 個不同的部落(cluster)，母體中的每一個元素均屬於其中的一個部落，且是唯一的一個部落。然後再從這些部落中隨機抽取部落，並對抽出的部落進行普查的抽樣方法，又稱集團抽樣。
  - 例如：要瞭解農家所得，可先抽取出鄉鎮，接下來對整個鄉鎮進行普查。



- 系統抽樣的意義：系統抽樣法是自母體自然隨機排列的資料中，每隔一定間隔選取一個樣本，直至抽滿 n 個樣本為止。
  - 例如：欲調查產品的瑕疵品比率，可在生產線上每隔 50 個抽取一個產品，直到抽滿 n 個樣本為止。



- 分段抽樣的意義：分段抽樣法是將母體按照某些特性或某種分類標準分為數個部落或層別，先由這些部落或層別中抽出幾個部落或層別，此為第一段。然後再由已經抽出的部落或層別，依特性或分類標準再抽出部落或層別，此為第二階段，如此依序為之，最後再依隨機或系統或其它方法抽出樣本。