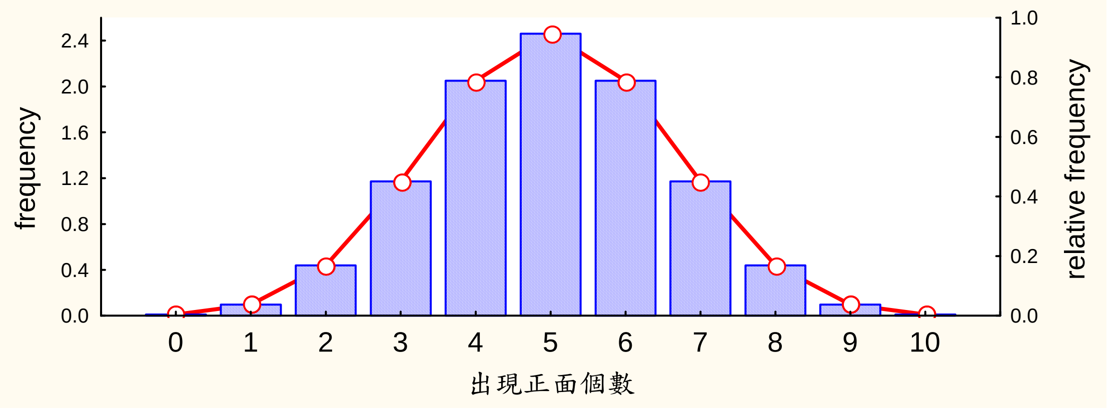
第六章　常態分布

一、什麼是常態分布

前面介紹過間斷機率的分布，以二項分布為例，如果投擲一均勻的的硬幣 10 次，我們預期每次正面反面的機率相同，都是 0.5。下表列出投擲 10 次硬幣出現正面次數所有可能的樣本空間、其預期頻率以及其相對頻度(機率)：

出現正面次數 (sample space 樣本空間)	出現正面次數頻率	相對頻度(機率)
0	0.009765625	0.000977
1	0.09765625	0.009766
3	1.171875	0.117188
4	2.05078125	0.205078
5	2.4609375	0.246094
6	2.05078125	0.205078
7	1.171875	0.117188
8	0.439453125	0.043945
9	0.09765625	0.009766
10	0.009765625	0.000977
Total	10	1

根據上表，我們把所有可能出現的結果(即樣本空間)，依理論上預期頻度與相對頻度畫成如下的直方圖



上圖就是間斷型變數(二項分布)頻度的直方圖，如果我們把它想像成型的變數，那麼介於兩個整數之間的無限個數字就變的有意義了。這樣我們就可以把每一個出現次數的機率連接起來(如上圖紅色線條)，這樣大致上就呈現一個鐘型的機率分布平滑曲線 (bell-shaped distribution)。

如果重複這樣的實驗無限多次，理論上就會成為上圖所示的分布：以出現 5 次正面(平均值)的機率最高；以平均值 5 次為中心呈兩側對稱，離開平均值越遠，其出現的頻率或機率就越低。這個曲線所顯示的機率分布，若變數是連續性的變數，這個機率的分布就是常態分布。

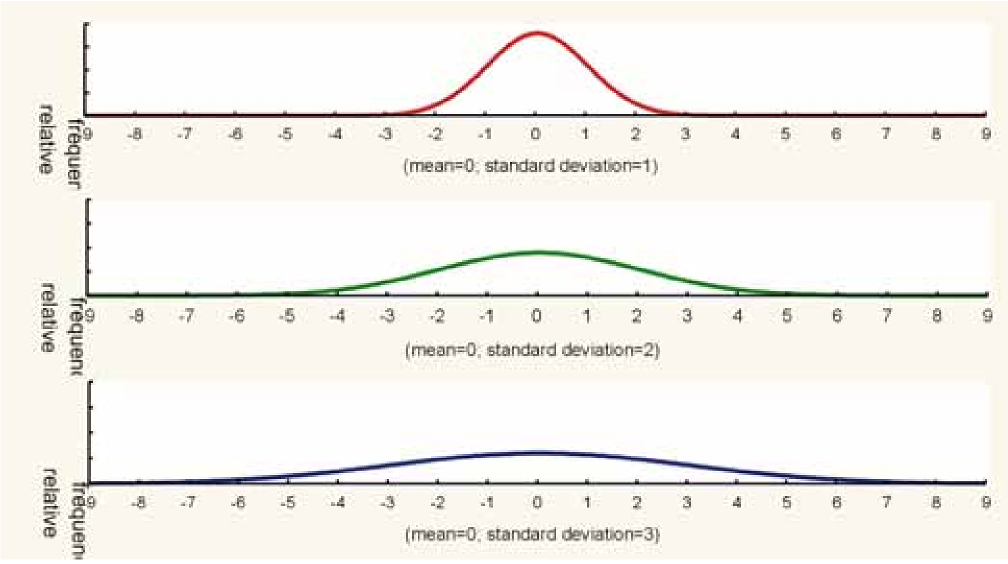
常態分布 (normal distribution)

常態分布為連續型的機率分布，又稱高斯（Gaussian）分布，它的機率密度函數方程式為： $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$

以 X~$N(\mu,\sigma^2)$表示之。常態分布是一曲線家族，大致上鐘型形狀不會在的中心點

位置與高矮胖瘦會不一。其所在位置及高矮胖瘦取決於兩個參數；即平均值μ(決定中心點所在位置)與標準差σ(決定鐘型曲線的形狀高矮胖瘦)。

標準差σ決定這個分布的高矮胖瘦



鐘形受μ及σ之影響

μ值影響鐘形中心位置

σ值影響鐘形形狀

» σ越大，則資料愈分散，鐘形越低寬

» σ越小，則資料愈集中，鐘形越高窄

常態分布是最常被使用的分布是因為常態分布是往後運用統計來做推論時重要的假設前提：假設某個參數的母族群是常態分布。當然許多自然界的現象或特徵值參數得分布多為常態分布，例如許多物理的、生物的及社會學的特徵值通常呈現常態分布。

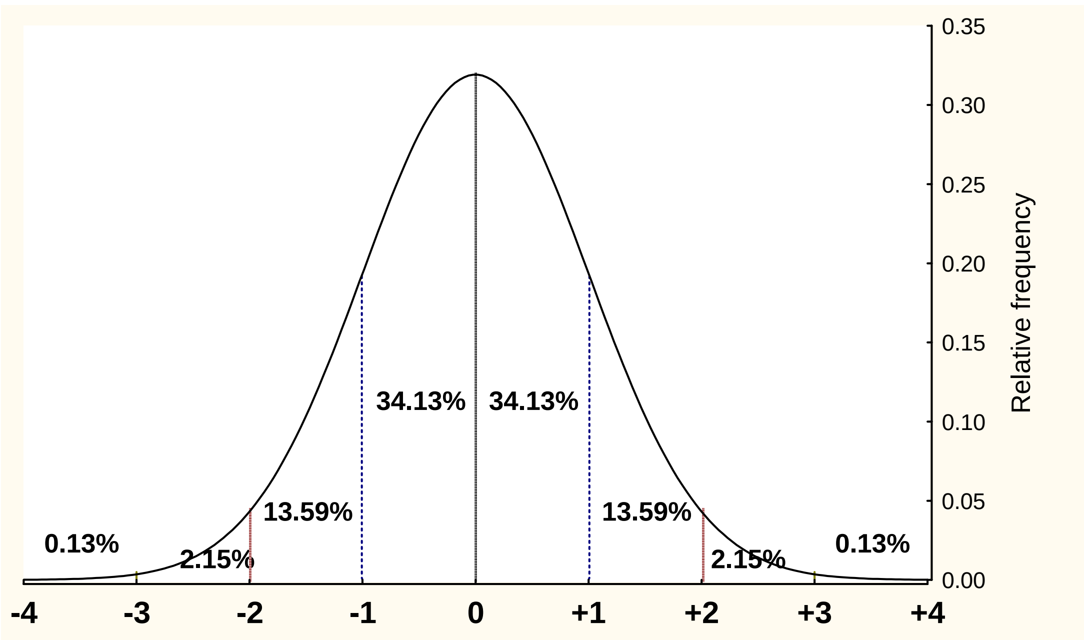
機率分布函數

- 常態分布曲線函數：$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2};-\infty<x<\infty$
- 常態曲線下總面積=1 ($\int_{-\infty}^{\infty} f(x)\,dx = 1$)
- 因此介於任兩數(a, b)之間的面積(即機率)可以求得 $p(a \le x \le b)=\int_a^b f(x)dx$

二、標準常態分布

標準常態分布 (standard normal distribution)

雖然常態分布是一個家族，但標準的常態分布只有一個，是常態分布家族中的一個特例，就是當常態分布的平均值為 0 ($\mu=0$)且標準差為 1 ($\sigma=1$)時，就稱為標準的常態分布($X \sim N(0,1)$)。



$\mu-4\sigma$ $\mu-3\sigma$ $\mu-2\sigma$ $\mu-1\sigma$ μ $\mu+1\sigma$ $\mu+2\sigma$ $\mu+3\sigma$ $\mu+4\sigma$

如果一個母族群是呈標準常態分布，平均值$\mu = 0$標準差$\sigma = 1$，則理論上族群中有各有 50 %的組成元素大於或小於平均值 0(以 0 為中心對稱)；有 68.26% $(34.13\% \times 2)$的組成元素x會分布在距離平均值加減一個標準差$\mu \pm 1\sigma$ $(0 \pm 1$即$+1$與-1之間的範圍內；有 95.44%的x會分布在距離平均值加減兩個標準差$\mu \pm 2\sigma$ (0 ± 2)即$+2$與-2的範圍內；有 99.74%的x會分布在距離平均值加減三個標準差$\mu \pm 3\sigma$ (0 ± 1)即$+3$與-3的範圍內。相反的，我們如果從母族群逢基指定一個x，也可以同樣的方式由x值得知，大於或小於x值的機率是多少被抽樣取到的機率有多少。想要知道這些值，只需要查表、透過各式各樣的統計軟體或是上網都可以輕易的查到，不必經由複雜的計算。

例如從標準常態分布母族群逢基取樣本大小為 n 的樣本，其均值大於或小於 0 的機率各為 0.5；樣本均值大於 1 的機率為 0.16，樣本均值小於 1 的機率為 0.84；樣本均值小於 2 而大於-1 $(-1 < x < 2)$的機率為 0.82；樣本均值小於-1 而大於-2 $(-2 < x < -1)$的機率為 0.14。

常態分布在自然界有很多，但標準常態分布必須滿足$\mu = 0$, $\sigma = 1$兩個條件，在自然的狀況少見，通常為數學物理學上的理論值或人為制定的標準，如某些測量儀器的誤差標準，有些輸出入物品某添加物的含量或檢定標準等。所以常態分布經常必須標準化才能利用標準常態分布的性質。

三、常態分布標準化

我們把($X \sim N(0,1)$)的常態分布當成標準是因為任何常態家族經過適當的轉換都可以變成標準的常態分布。那麼我們就可以便利的利用標準常態分布的特性得到我們想知道的訊息。這個轉換的程序我們稱為標準化。

標準化的程序分為兩個步驟：其一是要如何把平均值變成 0；其二是要如何把

標準差變成 1。設有一個常態分布母族群其平均值為 μ ，我們只要把所有 x 都減去平均值($x_i-\mu$)，新的平均值就變成 0。 回想一下敘述統計，這時新的母族群(由元素 $x_i-\mu$ 構成的母族群)的標準差並沒有改變還是 σ 。然後新的母族群的每個($x_i-\mu$)再除以母族群的標準差($(x_i-\mu)/\sigma$)，則新的母族群的標準差變成 1。我們回想一下敘述統計當每個數加減或乘除一個常數時對平均值及標準差的影響。

標準化的完整過程為:

$$\mathrm{z}=\frac{\mathrm{x}-\mu}{\sigma}$$

$$\mu_{new}=\frac{\sum(\frac{x-\mu}{\sigma})}{N}=\frac{1}{\sigma}\frac{(\sum x-\mu)}{N}=\frac{1}{\sigma}\frac{(\sum x-N\cdot\mu)}{N}=\frac{1}{\sigma}(\frac{\sum x}{N}-\mu)=\frac{1}{\sigma}(\mu-\mu)=0$$

$$\sigma_{new}=\frac{\sqrt{\sum(\frac{x-\mu}{\sigma}-0)^2}}{\sqrt{N}}=\frac{\sqrt{\frac{1}{\sigma^2}\sum(x-\mu)^2}}{\sqrt{N}}=\frac{1}{\sigma}\cdot\sqrt{\frac{\sum(x-\mu)^2}{N}}=\frac{1}{\sigma}\cdot\sigma=1$$

轉換後的值一班以 z 來表示，所以標準常態分布 (Standard Normal Distribution)又稱 Z 分布 $Z \sim N(0,1)$

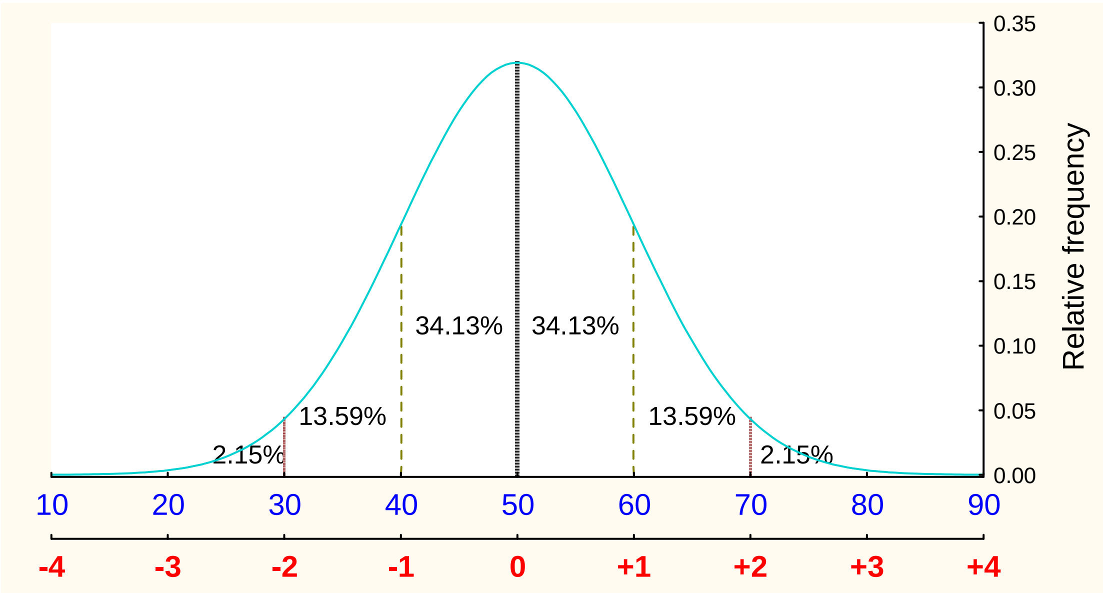
例如，假設有一 $\mu=50$, $\sigma=10$ 的常態分布

經 $z=\frac{x_i-\mu}{\sigma}=\frac{x_i-50}{10}$ 標準化後，成為 $\mu=0$, $\sigma=1$ 的 Z 分布 $Z \sim N(0,1)$

以實際的例子來看，假設有一母群是由 2, 5, 6, 9 組成經 $z=\frac{x-\mu}{\sigma}$ 轉換後如下表，在重新計算其平均值與標準差分別為 0 與 1

x	2	5	6	9	$\mu_x = 5.5$	$\sigma_x = 2.5$
$z = \dfrac{x-\mu}{\sigma}$	-1.4	-0.2	0.2	1.4	$\mu_z = 0$	$\sigma_z = 1$

又有某一人供造林樹種之純林最大徑圍為常態分布其平均值是 50 cm 標準差為 10 cm，下圖以兩種不同的尺度來顯示此一常態分布：



$\mu-4\sigma \quad \mu-3\sigma \quad \mu-2\sigma \quad \mu-1\sigma \quad \mu \quad \mu+1\sigma \quad \mu+2\sigma \quad \mu+3\sigma \quad \mu+4\sigma$

上面是原始尺度，下面為標準化後之尺度。標準化後之尺度±1，±2，±3我們可以簡單的理解成”離開平均值 50 cm ±1，±2，±3個標準差σ(分別為$\mu\pm1\sigma$，$\mu\pm2\sigma$，$\mu\pm3\sigma$)的距離”；也就是1×10，2×10，3×10 cm。

例題 1

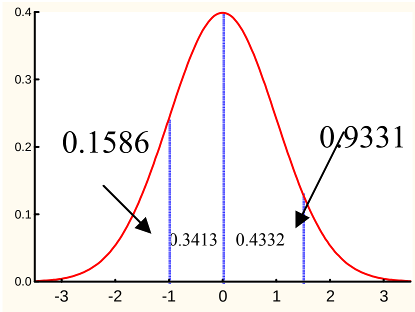
試求標準常態分布下 $x = -1 \sim 1.5$ 的機率？

[解]：

$P(-1 < x < 1.5) = P(-1 < x < 0) + P(0 < x < 1.5)$
$= 0.3413 + 0.4332 = 0.7745$

or

$P(x < 1.5) - P(x < -1) = 0.9331 - 0.1586$
$= 0.7745$



例題 2

一標準常態分布，

(1) 如 $P(x < x_a) = 0.025)$ ， $P(x < x_b) = 0.975$ ；試求 x_a 與 x_b 值各為多少？

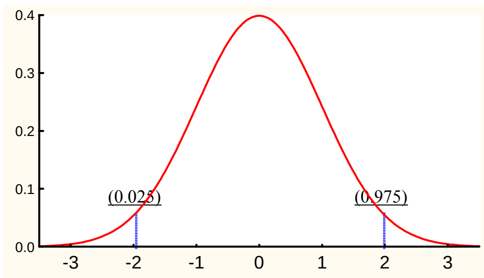
(2) 如 $0.025 < P(x) < 0.975$ ，試求值的範圍？

[解]：

從單尾 Z 值表中找出機率值為 0.025 與 0.975 在找出其對應之 Z 值

(1) $P(x < -1.96) = 0.025$
$P(x < 1.96) = 0.975$

所以 $x_a = -1.96$
$x_b = +1.96$



(2) $P(x < -1.96) = 0.025$
$P(x < 1.96) = 0.975$

所以 $-1.96 < x <= 1.96$

例題 3：某校在一次的統計學測驗中，參加的學生有 50 名。假設此次測驗的成績合於常態分配，且其平均成績為 80 分，標準差 5 分。試求：

(1) 成績在 65 分與 75 分之間的人數

(2) 成績在 90 分以上的人數

[解]：

(1) 令 x 代表統計學成績，依題意知 $\mu = 80, \sigma = 5$，於是：

$$P(65 < x < 75) = P(\frac{65-80}{5} < Z < \frac{75-80}{5}) = P(-3 < Z < -1)$$
$$= 0.1587 - 0.0013 = 15.74\%$$

(2) 成績在 90 分以上的人數

$$P(x > 90) = P(Z > \frac{90-80}{5}) = P(Z > 2) = 1 - P(Z < 2)$$
$$= 1 - 0.9772 = 2.28\%$$

故人數約為 50(人) × 2.28%=1.14(人)　1 人

例題 4：某品牌家電用品的使用壽命為平均數 3 年，標準差 1 年的常態分配。若其保證期間為二年，試問退貨比例為多少？

[解]：

設 X 為該品牌家電用品使用的壽命長度，依題意知 μ =3, $\sigma = 1$ 於是，

$$\mathrm{P}(x < 2) = \mathrm{P}(Z < \frac{2-3}{1}) = \mathrm{P}(Z < -1) = 0.1587$$

亦即退貨的比例約為 0.1587

例題 5：若性向測驗之成績呈常態分配，某測驗結果其 μ =506，σ =81，

試求 (1) 分數低於 574 者佔全體之比例；

(2) 第 30 百分位數

[解]：令 X 表測驗成績，

(1) 依題意，成績低於 574 之比例為：

$$P(x<574)=P(Z<\frac{574-506}{81})=P(Z<0.8395)=0.799$$

亦即測驗成績低於 574 者大約佔 79.9 %

(2) 第 30 百分位數為 x 即表示常態機率

$$P(\frac{x-506}{81})=0.3$$

查表可得 $P(Z<-0.52)=0.3015$ ；$P(Z<-0.53)=0.2981$

所以我們取約 $P(Z<-0.524)\cong 0.3$

即 $\frac{x-506}{81}=-0.524$

故第 30 百分位數 $x=506+81\times(-0.524)=463.56$

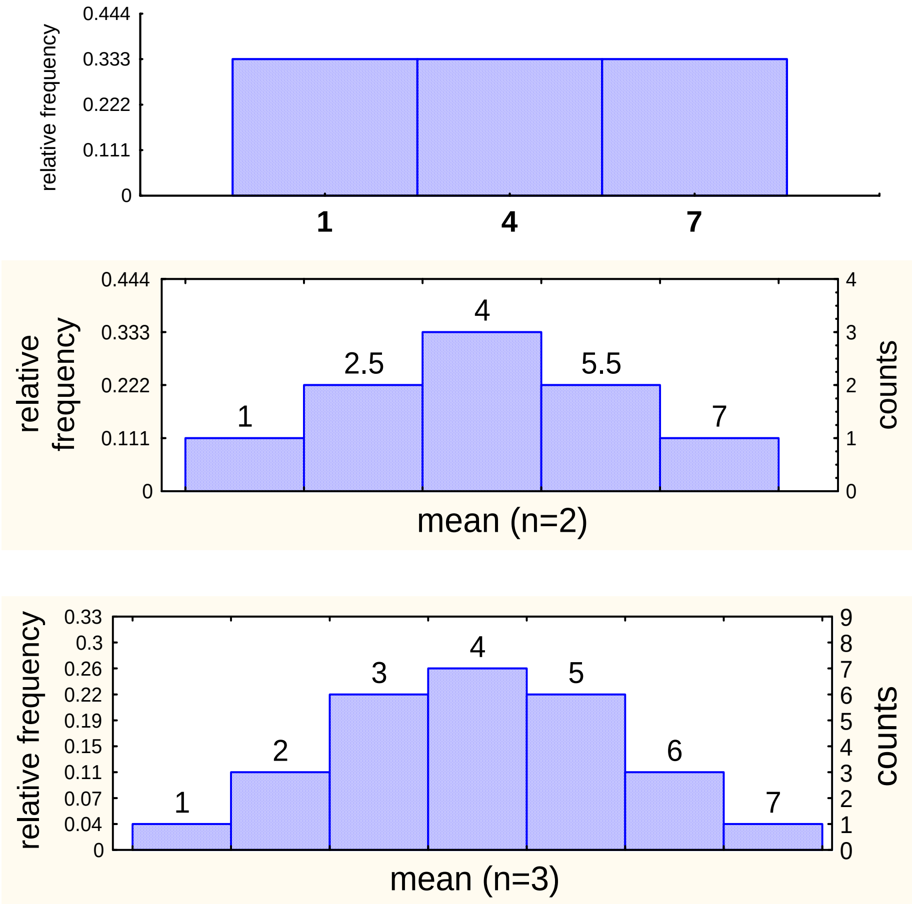
四、樣本平均值的分布

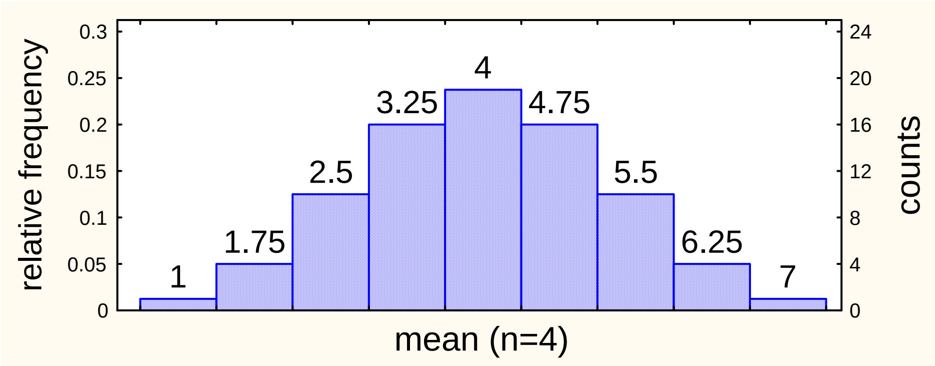
統計實際操作的方法是利用抽樣的方法取得樣本，利用樣本所得到的平均數、標準差、變異數等統計值(sample statistics) 來了解母族群的平均數、標準差、變異數等參數(parameters)。因此，我們更關心抽樣樣本統計值分布的情形。例如，從一個常態分布的母族群抽樣的樣本當然會符合常態分布。但，即便某些特徵值的母族群並非常態分布，只要取樣的樣本數夠大時(通常樣本數 n 大於 30 時)，其樣本平均值的分布也會是常態分布，這便是中央極限定理(central limit theorem)。

例如我們可以假設一個只含有 1,4,7 的母族群，以實際的取樣來觀察樣本均值的分布情形。我們分別自母族群取出樣本數等於 2、3 和 4(n=2,3,4)的樣本，將所

有可能的樣本組合逐一列出；例如當 n=2 時，所有可能的樣本有(1,1)、(1,4)、(1,7)、(4,1)、(4,4)、(4,7)、(7,1)、(7,4)和(7,7)一共有 9 種可能的樣本。n=3 時，所有可能的樣本有(1,1,1)、(1,1,4)、(1,1,7)、(1,4,1)、(1,4,4)、(1,4,7)、(1,7,1)、(1,7,4)、(1,7,7)、(4.1,1)、(4,1,4)、(4,1,7)、(4,4,1)……(7,7,7)。再分別計算這些樣本的平均值，例如樣本(1,1)的平均值為 1($\bar{x}=1$)、樣本(1,4)的平均值為 2.5($\bar{x}=2.5$)、樣本(1,7)的平均值為 4($\bar{x}=4$)……樣本(1,7)的平均值為 7($\bar{x}=7$)，依此類推。然後計算每個樣本平均值出現的頻率。

這樣我們可以依不同的樣本數畫出如下圖的直方圖，看看樣本平均值分布的情形與樣本數大小有何關係。





接著我們分別再算出樣本均值的平均值、樣本均值的變異數及標準誤差 (standard error) (樣本均值的標準差稱為標準誤差)，如下表：

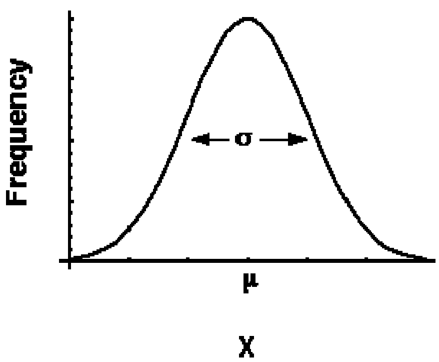
母族群	3	$\mu = 4$	$\sigma^2 = 6$	$\sigma = 2.4495$
樣本數	樣本空間	均值	變異數	標準誤差(樣本均值的標準差)
	N	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}^2$	$\sigma_{\bar{x}}$
n= 2	3^2=9	4	3	1.7321
n= 3	3^3=27	4	2	1.4142
n= 4	3^4=81	4	1.5	1.2247

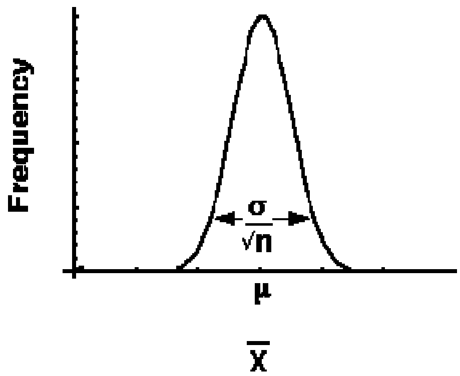
我們發現即使不是常態分布的母族群，從這個母族群抽樣出來的樣本均值的分布，會隨著抽樣樣本數(n)的增大而越接進常態分布。且不管樣本數大小，樣本平均值的平均值等於母族群的均值 $\mu_{\bar{x}} = \mu$ ；而樣本平均值的變異數隨著樣本數增大而減小且 $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}$ ；因此樣本平均值的標準差(標準誤差) 隨著樣本數增大而減小 $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ 。

如果原來的母族群為常態分布，其樣本平均值所形成的族群當然為常態分

布，且其均值不變($\mu_{\bar{x}} = \mu$)；但其樣本平均值的標準差隨樣本數而減小($\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)；

如下圖所示：左圖為常態分布知母族群中 x 分布情形，下圖為其樣本大小為 n 時樣本平均值 $\bar{x}$ 的分布情形。





上述的情形我們可以歸納出來所謂的中央極限定理 (Central Limit Therrem) ：當樣本數夠大(通常 n 大於 30)時，不論母族群機率分布如何，從母族群抽樣的樣本平均值($\bar{x}$)的分布近似常態分布即 $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$ 其中 $\mu_{\bar{x}} = \mu$，$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$，所以 $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$。

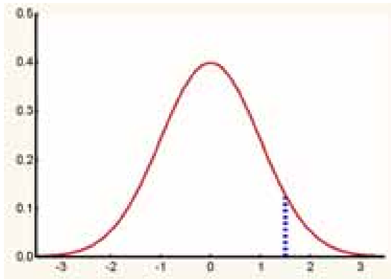
練習例題

例題 1：有一品牌溫度計宣稱其溫度計的標準差為 1，如果我們用此溫度計測量純水的冰點理論的平均值應為 0^{o}C，重複測量結果有些值會小於 0^{o}C，有些值會大於 0^{o}C，試問若溫度計準確度如廠商宣稱，測得水結冰溫度大於 -0.8 ^{o}C 小於 0.8 ^{o}C 的機率有多少？

[解]：

測值的分布為標準常態分布 $X \sim N(0,1)$

$P(-0.5 > Z > 0.5) = 0.5763$



例題 2：設某學校學生之身高為平均數 μ =170 cm，標準差 σ =12 cm 的常態分配，若由此學校隨機抽出 16 個學生為一個樣本，則所抽出樣本之平均數在母群體平均數 6 cm 以內的機率為何？

[解]：

平均身高 X 為常態分配，且平均值與標準誤差為 μ =170，$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{12}{\sqrt{16}} = 3$

$\overline{\mathrm{X}}$ 位於 $\mu \pm 6 = 170 \pm 6\,\mathrm{cm}$ 以內範圍（亦即介於 164 至 176 cm 之間）的機率為：

$$P(164 < \bar{x} < 176) = P(\frac{164-170}{3} < Z < \frac{176-170}{3}) = P(-2 < Z < 2) = 0.9545$$

例題 3：設某班級學生之統計學成績呈常態分配，其平均數為 72，標準差為 9。

(a) 自該班中隨機抽出 1 人，其分數超過 80 之機率。

(b) 自該班中隨機抽出 9 位學生，則此 9 位學生之平均成績超過 80 的機率。

(c) 自該班中隨機抽出 25 位學生，則此 25 位學生之平均成績介於 70 到 75 的機率。

[解]：

令 X 代表該班學生之統計學成績 $X \sim N(72,9)$

(a) 已知 $X \sim N(72,9)$

$$P(x > 80) = P(Z > \frac{80-72}{9}) = P(Z > 0.89) = 0.1867$$

(b) 隨機抽出 10 位學生其平均分數超過 80 的機率為：

$$P(\bar{x} > 80) = P(Z > \frac{80-72}{(9/\sqrt{9})}) = P(Z > 2.67) = 0.0038$$

(c) 自該班中隨機抽出 20 位學生，平均成績介於 70 到 75 的機率為：

$$P(70 < \bar{x} < 75) = P(\frac{70-72}{(9/\sqrt{25})} < Z < \frac{75-72}{(9/\sqrt{25})}) = P(-1.11 < Z < 1.67) = 0.8190$$

例題 4：據調查，大學畢業生初進公司的起薪為一平均 28000 元、標準差為 1500 元的常態分布，若你畢業後將就業，問：

(1) 你希望你的起薪能高於 30000 元，機率有多大？

(2) 你的起薪有 90%的機率會高於多少元呢？

(3) 若從你同屆的大學畢業生中，隨機抽取 20 人的起薪為樣本，則此樣本平均起薪介於 27000 至 30000 的機率為何？

(4) 有 50%的機率此 20 人的平均起薪會低於多少呢？

(5) 有 90%的機率此 20 人的平均起薪會高於多少呢？

[解] ：

(1) 你希望你的起薪能高於 30000 元

$$P(x > 30000) = P(Z > \frac{30000-28000}{1500}) \cong P(Z > 1.33) = 1 - P(Z < 1.33)$$
$$= 1 - 0.9082 = 0.0918$$

(2) 你的起薪有 90%的機率會高於多少元

$$P(Z > \frac{x-28000}{1500}) = 1 - P(Z < \frac{x-28000}{1500}) = 0.9$$

即 $P(Z < \frac{x-28000}{1500}) = 0.1$ 查表得 $P(Z < -1.28) = 0.1003$

$$\frac{x-28000}{1500} = -1.28$$
$$x = 26080$$

(3) 若從你同屆的大學畢業生中，隨機抽取 20 人的起薪為樣本，則此樣本平均起薪介於 27000 至 30000 的機率為

$$P(27000 < \bar{x} < 30000) = P(\frac{27000-28000}{1500/\sqrt{20}} < Z < \frac{30000-28000}{1500/\sqrt{20}})$$
$$= P(-2.98 < Z < 5.96) = P(Z < 5.96) - P(Z < -2.98) = 0.9986$$

(4) 有 50%的機率此 20 人的平均起薪會低於多少

$$P(Z < \frac{\bar{x}-28000}{1500/\sqrt{20}}) = 0.5$$

查表得 $P(Z < 0) = 0.5$，所以 $\frac{\bar{x}-28000}{1500/\sqrt{20}} = 0$，$\bar{x} = 28000$

(5) 有 90%的機率此 20 人的平均起薪會高於多少

$$P(Z > \frac{\bar{x}-28000}{1500/\sqrt{20}}) = 1 - P(Z < \frac{\bar{x}-28000}{1500/\sqrt{20}}) = 0.9$$

$P(Z < \frac{\bar{x}-28000}{1500/\sqrt{20}}) = 0.1$ 查表得 $P(Z < -1.28) = 0.1003$

$\frac{\bar{x}-28000}{1500/\sqrt{20}} = -1.28$ 所以 $\bar{x} = 28000 - 1.28 \times \frac{1500}{\sqrt{20}} \cong 27571$

五、主要參考文獻

1. 方世榮。2005。統計學導論，華泰文化事業股份有限公司，台北。
2. 林惠玲、陳正倉。1999。應用統計學 ，雙葉書廊有限公司，臺北。
3. 沈明來。2001。生物統計學入門(第四版)，九州圖書文物有限公司，台北。