About Sampling for Statistics

關於統計「抽樣」

抽樣(sampling)基本概念

抽樣是指自母體取得樣本的程序或方法

抽樣的「隨機」條件

- 母體中任一的元素都有機會被抽中
- 樣本被抽出的機率為已知，或是可被計算
- 不同樣本之間，被抽出的過程彼此是「獨立事件」

抽樣方法：

非隨機抽樣方法

隨機抽樣方法

非隨機抽樣方法

●便利抽樣

樣本的選取，主要考量以方便性為主。

●判斷抽樣

根據研究者自己判斷如何選擇樣本，又稱為「立意式抽樣法」，在人文社會科學的領域中，問卷的調查對象常採用這種方式。

●滾雪球抽樣

主要針對調查對象數量稀少，甚至不知道在哪，此時先根據已知的少數樣本做調查，再從這些樣本所提供的管道取得其他樣本資訊。

隨機抽樣方法

●簡單隨機抽樣

樣本中任何一個元素，被選到的機率都相同的收樣方式：

抽出放回 → 每次抽樣都是獨立事件；每次抽出機率都相同。

抽出不放回～抽出不放回，所以前後次抽樣會受到影響，機率值會隨著前次抽樣〝不放回〞而有不同。

●系統抽樣

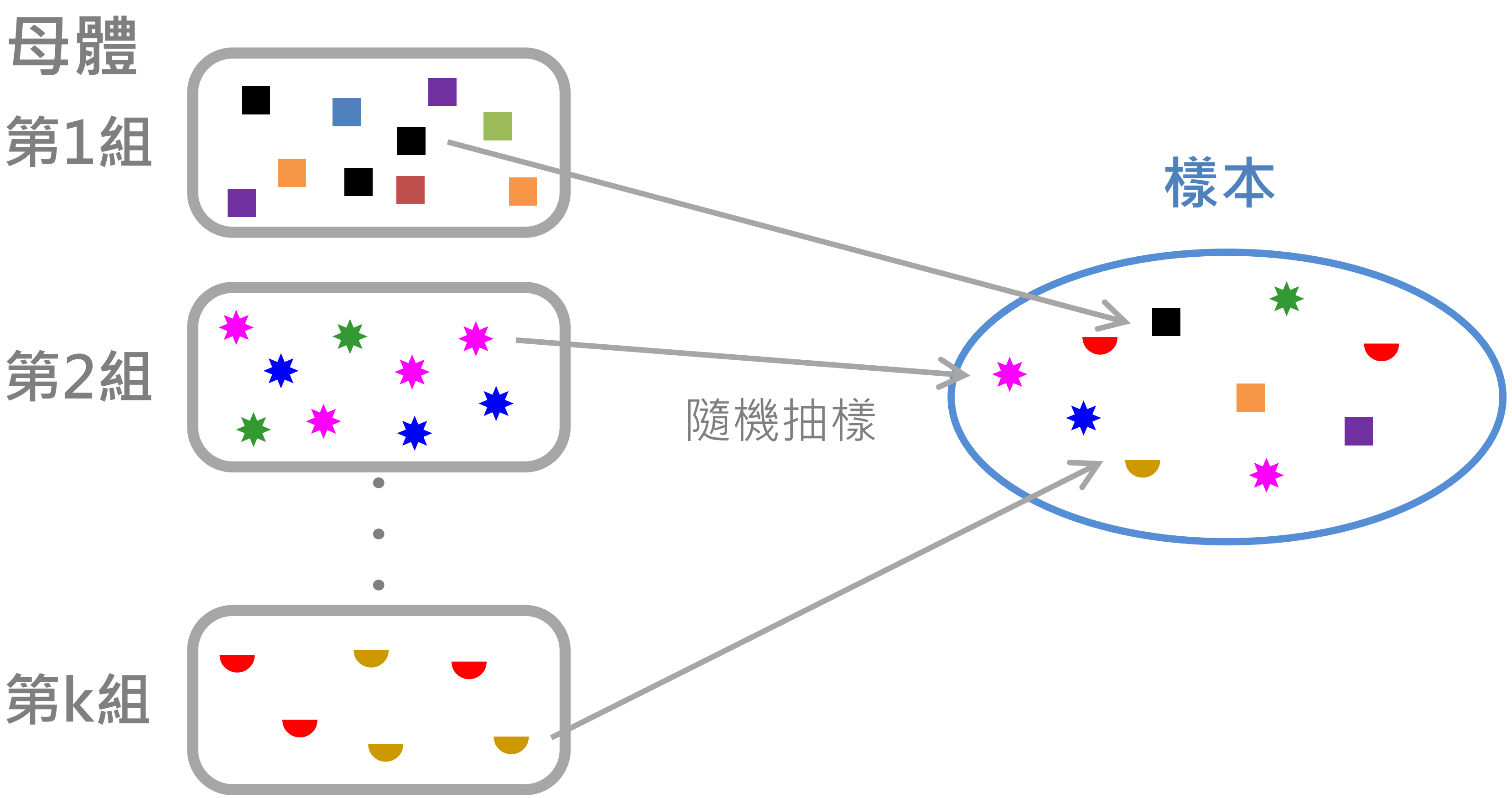
將母體各元素排列後，每隔一間隔選取一樣本，直到選滿為止。例如利用電話簿、名冊、通訊錄等「排列」，每隔10位選取一個「樣本」。

但此種方式不適用在具有週期性、季節性的資料，因為抽樣可能會集中在高點或是低點，影響樣本代表性。

隨機抽樣方法

●分組隨機抽樣

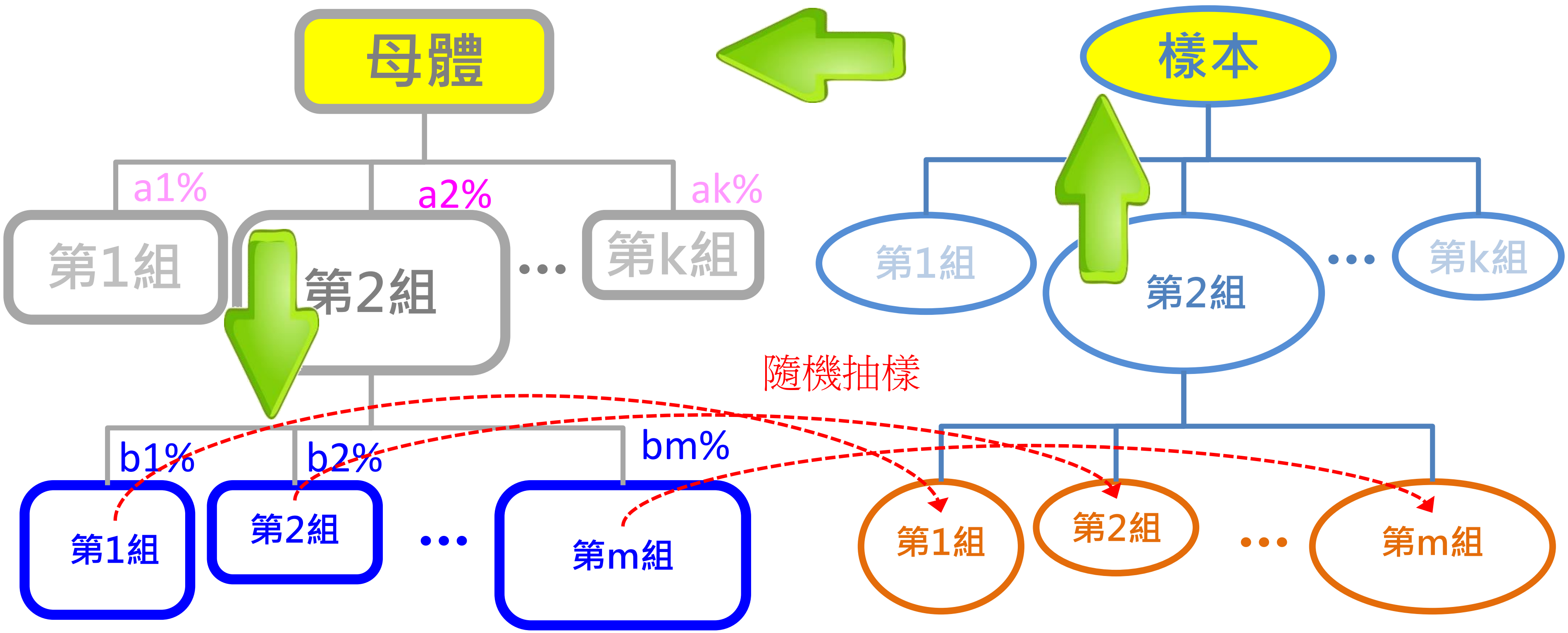
母體特質分散於各組（例如某地區調查對象的性別、居住地、收入等），在抽樣前，可將母體先分為數組，在依據各組於母體所佔的比例多少來隨機抽取樣本。



隨機抽樣方法

●分層隨機抽樣

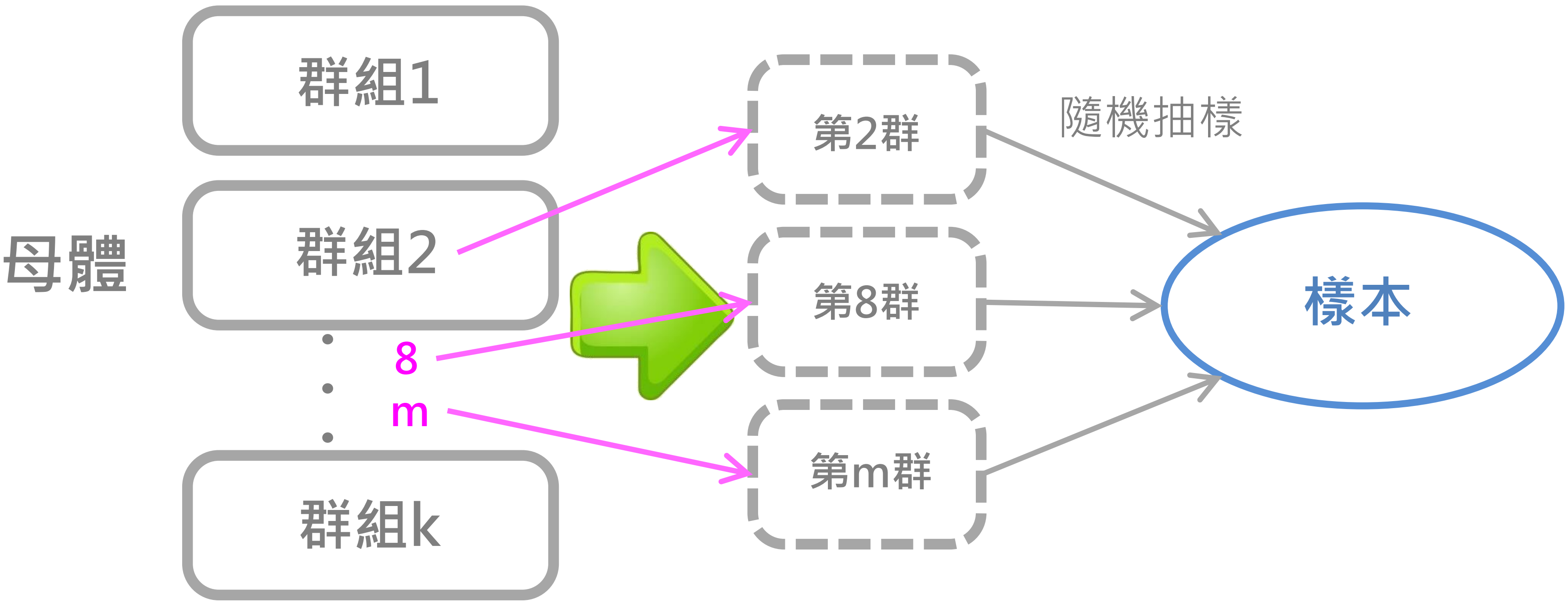
先將母體分為數層，再以每層特性進行分組抽樣。在抽樣時，每個層級單位都有階層關係(hieratical relationship)，利用樣本整體結果來推論母體。



隨機抽樣方法

●群落抽樣

將母體分為k組群落（群體），先從這k組中抽出m群，再從這m群中進行抽樣調查。因為調查範圍可以縮小，更容易控制時間、費用、但能提高調查品質。特別是在群落內差異性高，但群落間差異性小的調查中，可以抽樣幾個群落，就能獲得良好的母體推論。



抽樣分配的基本概念

抽樣分配與樣本分配有何不同？

母體：4人的年齡組合

- **Assume there is a population …**
- Population size N=4
- Random variable, X, is age of individuals
- Values of X: 18, 20, 22, 24 (years)

資料來源：Statistics for Managers Using Microsoft Excel 7th Edition, Global Edition, CH7

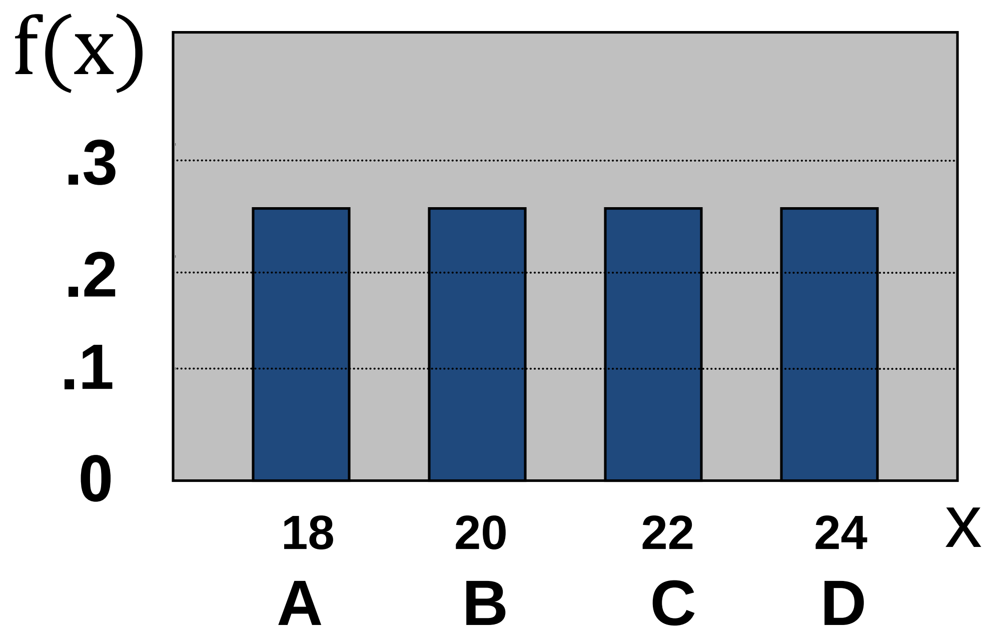
母體分配與參數：平均數(μ)與標準差(σ)

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N} = \frac{18+20+22+24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$

Uniform Distribution



樣本抽樣：抽2個人(n=2)

Now consider all possible samples of size n=2

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

$$\bar{x} = \frac{x_1 + x_2}{2}$$

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

16 Sample Means

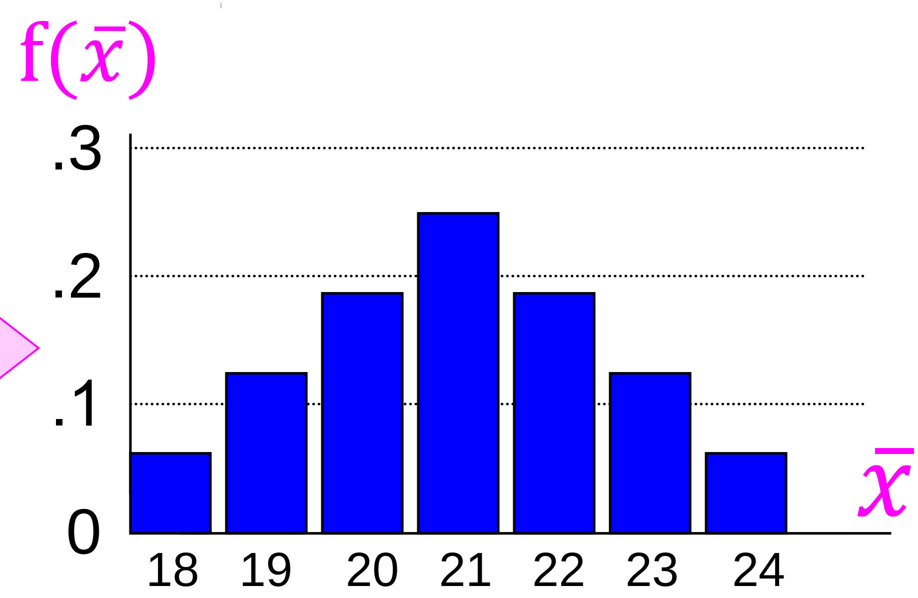
抽樣分配

Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

「抽樣分配」$\bar{x}$ 的平均數與標準差

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{x}} = \frac{18+19+19+\cdots+24}{16} = 21$$

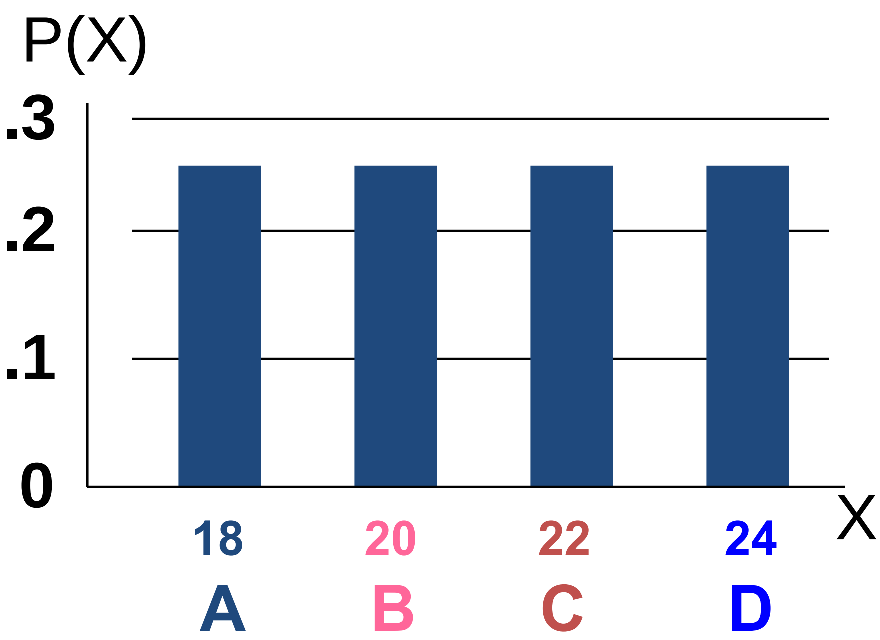
$$\sigma_{\bar{x}} = \sqrt{\frac{(18-21)^2+(19-21)^2+\cdots+(24-21)^2}{16}} = 1.58$$

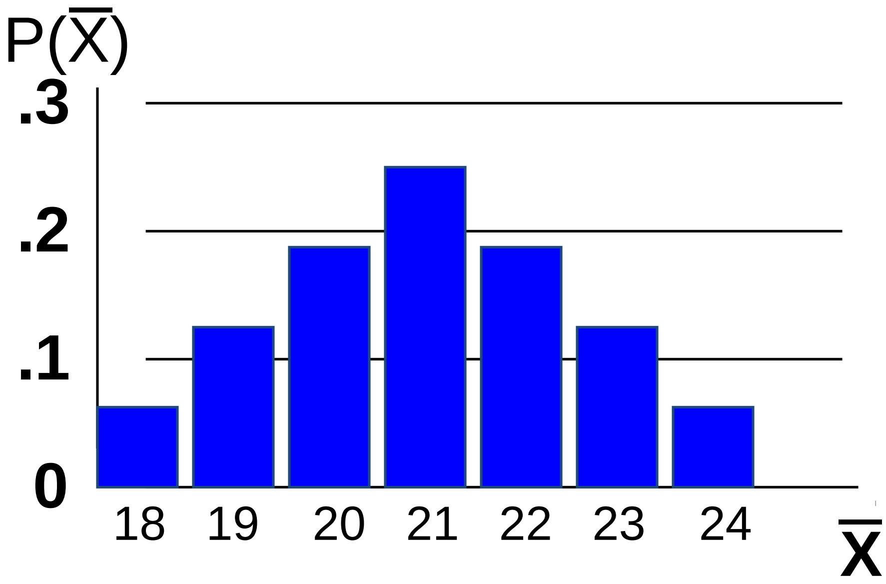
Note: Here we divide by 16 because there are 16 different samples of size 2.

母體分配與抽樣分配

Population
N = 4
$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution
n = 2
$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$

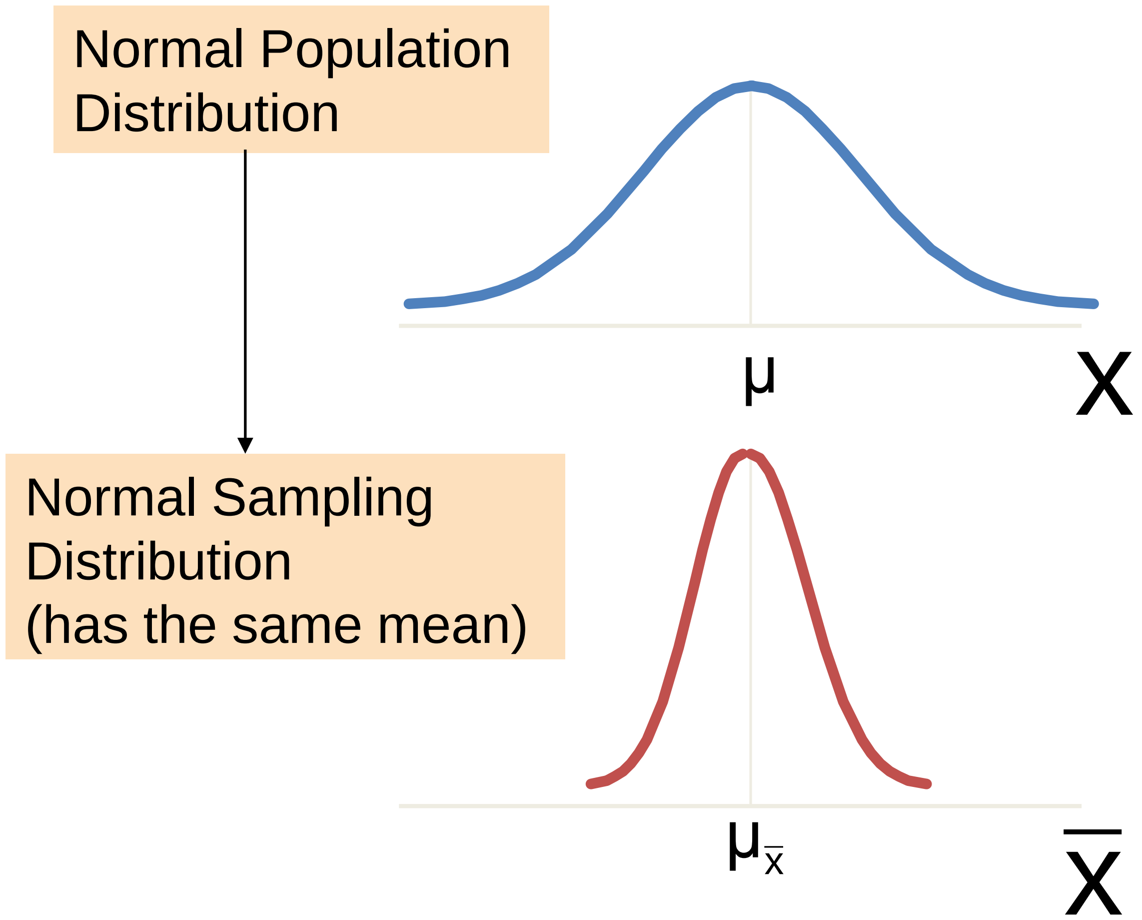




Sampling Distribution Properties

- $$\mu_{\bar{x}} = \mu$$

(i.e. $\bar{X}$ is unbiased)



Sample Mean Sampling Distribution: Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

  (This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

Sample Mean Sampling Distribution: If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

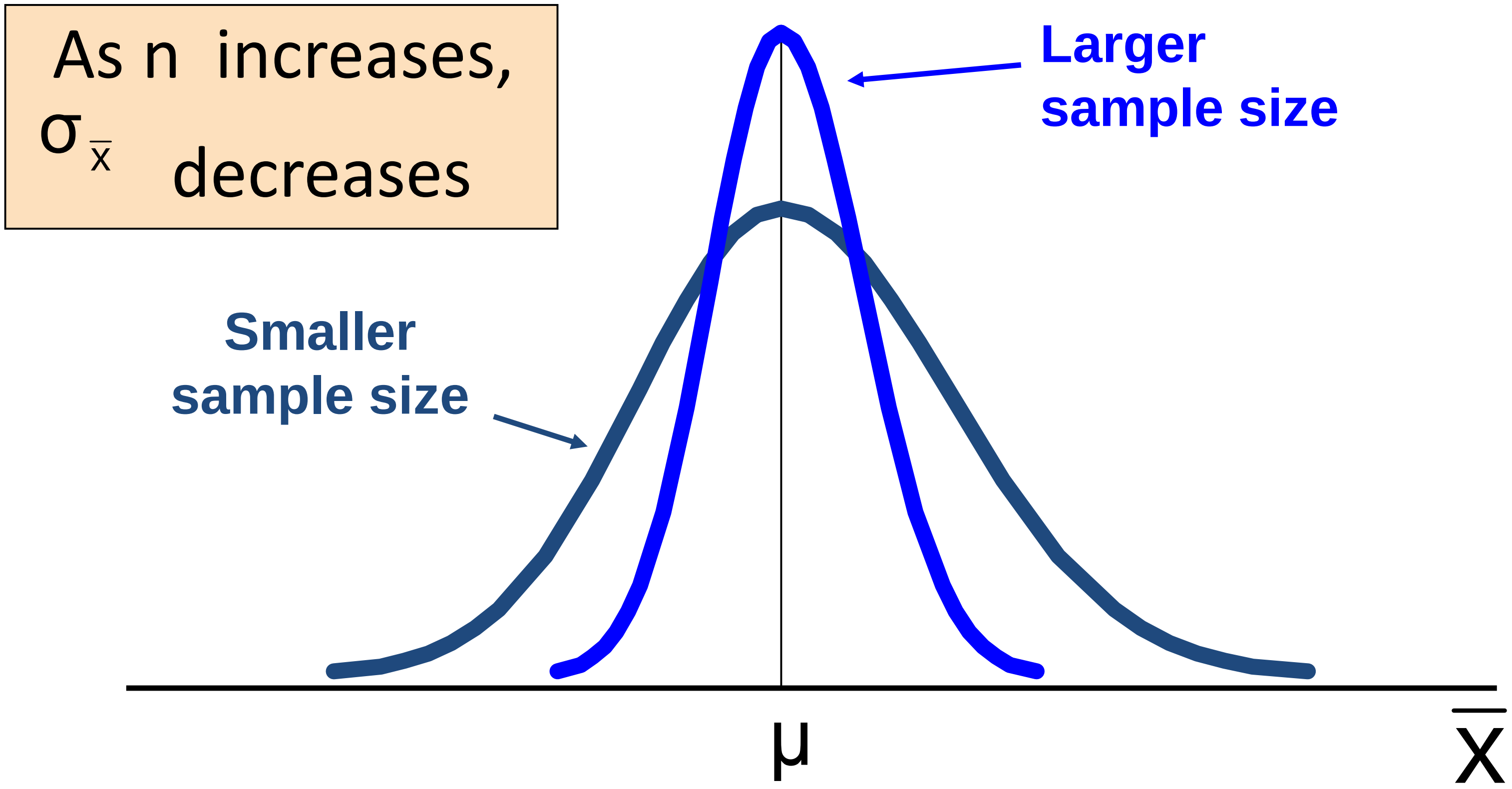
$\bar{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

Sampling Distribution Properties



Determining An Interval Including A Fixed Proportion of the Sample Means

Find a symmetrically distributed interval around μ that will include 95% of the sample means when $\mu = 368$, $\sigma = 15$, and $n = 25$.

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.

Determining An Interval Including A Fixed Proportion of the Sample Means

- Calculating the lower limit of the interval

$$\bar{X}_L = \mu + Z\frac{\sigma}{\sqrt{n}} = 368 + (-1.96)\frac{15}{\sqrt{25}} = 362.12$$

- Calculating the upper limit of the interval

$$\bar{X}_U = \mu + Z\frac{\sigma}{\sqrt{n}} = 368 + (1.96)\frac{15}{\sqrt{25}} = 373.88$$

- 95% of all sample means of sample size 25 are between 362.12 and 373.88

The Central Limit Theorem(C.L.T.)

中央極限定理

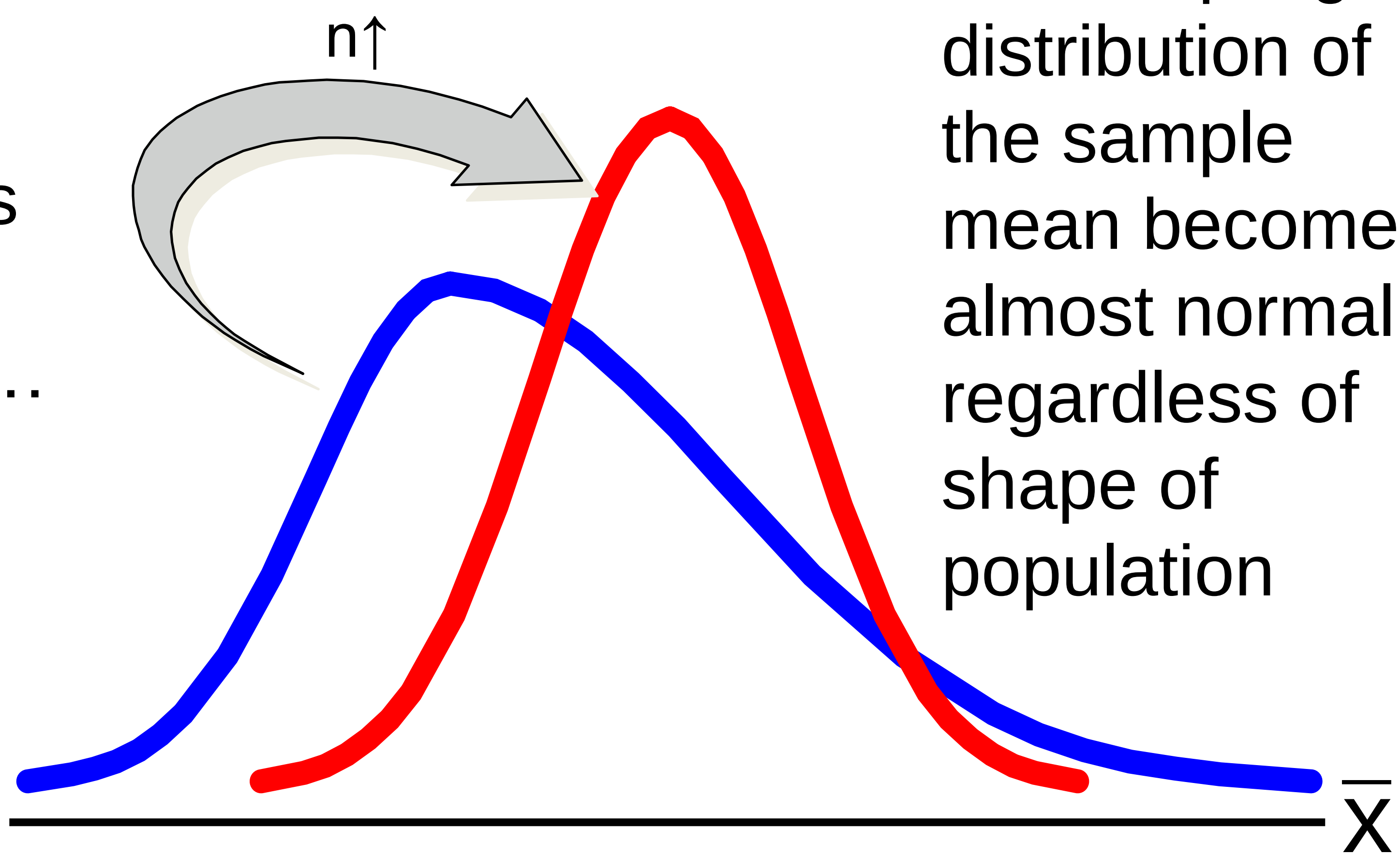
大數法則(law of large number)

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots . + x_n}{n} \approx \mu$$

大數法則主要含意代表著，當抽樣的樣本數越多，所獲得的結論越可靠

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution of the sample mean becomes almost normal regardless of shape of population

Sample Mean Sampling Distribution: If the Population is not Normal

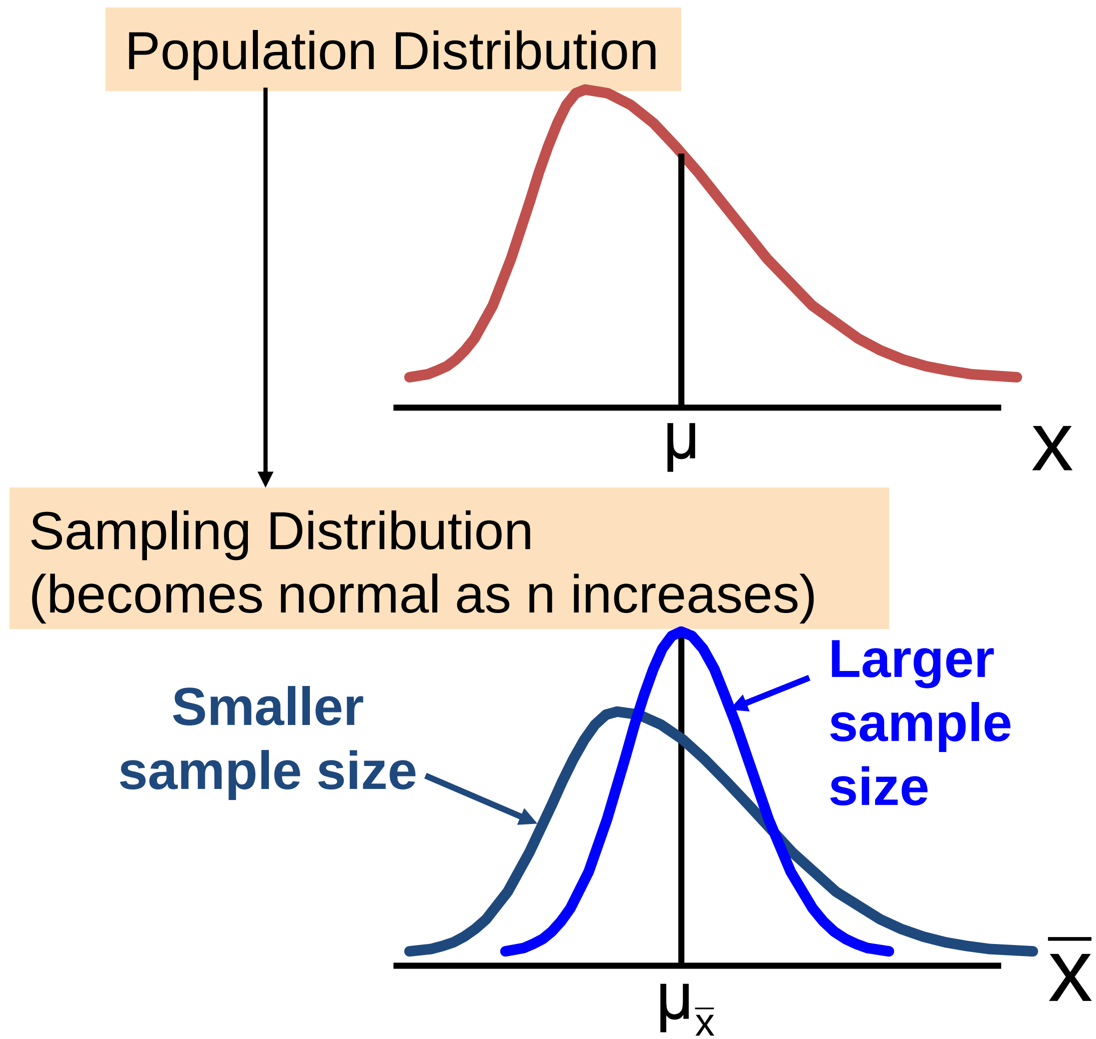
Sampling distribution properties:

Central Tendency

$$\mu_{\bar{x}} = \mu$$

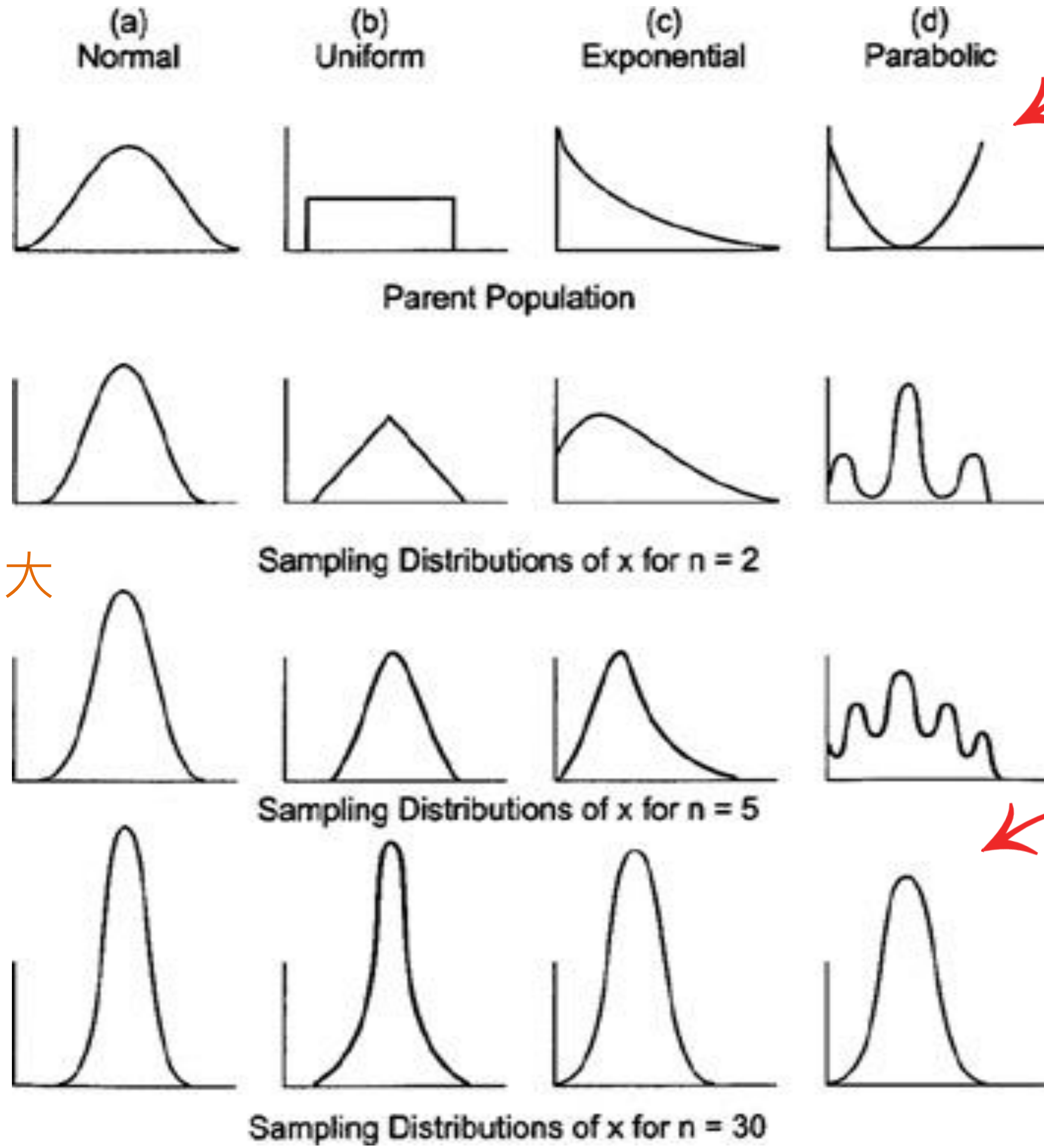
Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



C.L.T

n 越來越大



(a) Normal
(b) Uniform
(c) Exponential
(d) Parabolic
Parent Population
Sampling Distributions of x for n = 2
Sampling Distributions of x for n = 5
Sampling Distributions of x for n = 30

中央極限定理(C.L.T.)

當樣本數夠大時，樣本平均數的抽樣分配會近似常態分配：

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow Z = \frac{\overline{\mathrm{x}} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

此定理適用於母體為任何分配的形狀，但若母體分配為常態時，則不論n的大小，$\bar{\mathrm{x}}$ 的抽樣分配皆為常態分配。

實務上，當樣本數 n ≧ 30 時，中央極限定理變成立。

案例

假設某廠牌罐裝奶粉每罐平均重量為500克，變異數為120，現品管人員抽取30罐檢驗其重量，試問：

(1)抽取之30罐的樣本平均重量與母體平均數之差在3公克之內的機率為何？

(2)以母體平均數為中心，涵蓋95%的樣本平均數的區間為何？

「應用統計學 二版」p193，李德治、童惠玲 著，博碩文化

案例解說

平均重量500公克 → $\mu = 500$ ；變異數120 → $\sigma^2 = 120$

抽取30罐 → n=30 ；→ 〝平均數的抽樣分配〞呈常態分配

→ 列表、標準化、查表

(1) $P(500-3 \leqq \bar{x} \leqq 500+3) = P(497 \leqq \bar{x} \leqq 503)$

$$= P\left(\frac{497-500}{\sqrt{\frac{120}{30}}} \leqq Z \leqq \frac{503-500}{\sqrt{\frac{120}{30}}}\right) = P(-1.5 \leqq Z \leqq 1.5) = 0.8664$$

(2) $P(\mu - a \leqq \bar{x} \leqq \mu + a) = 0.95$

$$\Rightarrow P\left(\frac{(500-a)-500}{\sqrt{\frac{120}{30}}} \leqq Z \leqq \frac{(500+a)-500}{\sqrt{\frac{120}{30}}}\right) = 0.95$$

查表得 $P(-1.96 \leqq Z \leqq 1.96) = 0.95 \Rightarrow \frac{a}{\sqrt{\frac{120}{30}}} = 1.96$

所以 $a = 3.92 \rightarrow \bar{x} = [500 - 3.92,\ 500 + 3.92] = [496.08,\ 503.92]$

The End