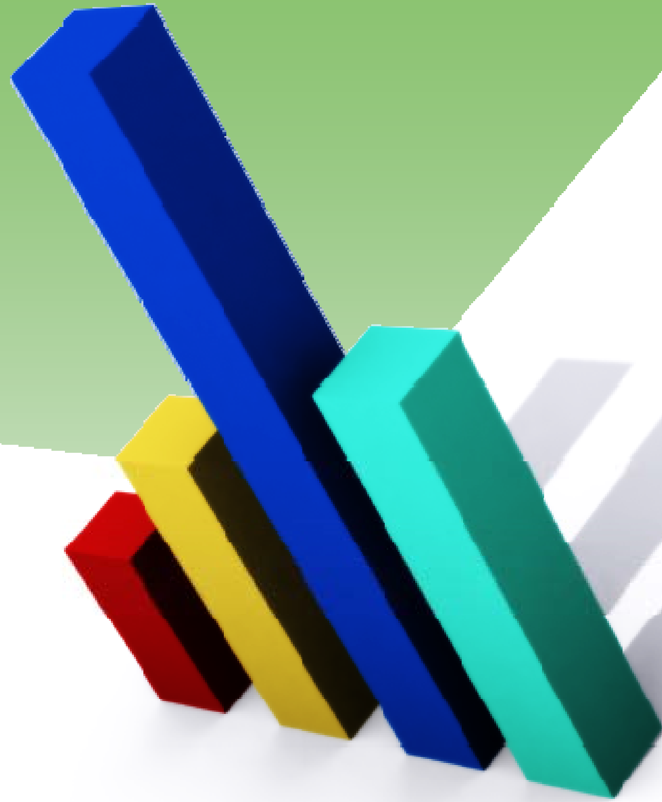
第 9 章



抽樣分配

Recall

- Inferential statistics is a process for transferring the data into information.
- Parameters are used for describing the characters of the target population of interest.
- In general , parameters are unknown (why?)
- The representative sample data are sampled from the target population with the appropriate sampling method.

Recall

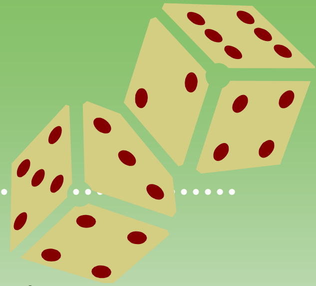
- The corresponding statistic (統計量) for the specified parameter is computed with the sample data.
  - Ex. The sample mean for the population mean
- Q: Why do we need to know the sampling distribution for a statistic before we introduce the inferential statistics?
- A: The sampling distribution for a statistic is a basic component of the inferential statistics !
  - That is, it is difficult to understand the inferential statistics if you are not familiar with the sampling distribution for a statistic.

9.1 樣本平均數的抽樣分配
(The sampling distribution one sample mean)

抽樣分配

- 一個統計量的抽樣分配 (sampling distribution)是由**抽樣**所產生。
- Two ways for generating the sampling distribution (p321)
  1. Repeatedly sampling the sample data with the same sample size from the population, and compute the statistic from each sample data. Based on these statistics from these sample data, the sampling distribution of the statistic could be constructed.
  2. 根據**機率法則**以及**期望值與變異數法則**導出統計量的抽樣分配。
     - 例如，(細想) 投擲一顆以及兩顆骰子…

Ex. 兩顆骰子平均數的抽樣分配



- 母體是藉由投擲一顆公正的骰子無限多次產生，
  隨機變數 X 是指任何一次投擲骰子所出現的點數。

- 隨機變數 X 的機率分配如下：

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

- 母體平均數與變異數的計算為：

$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

平均數的抽樣分配

一個抽樣分配是藉著觀察所有樣本量為2 (例，兩個骰子)的樣本與其平均數所產生的。

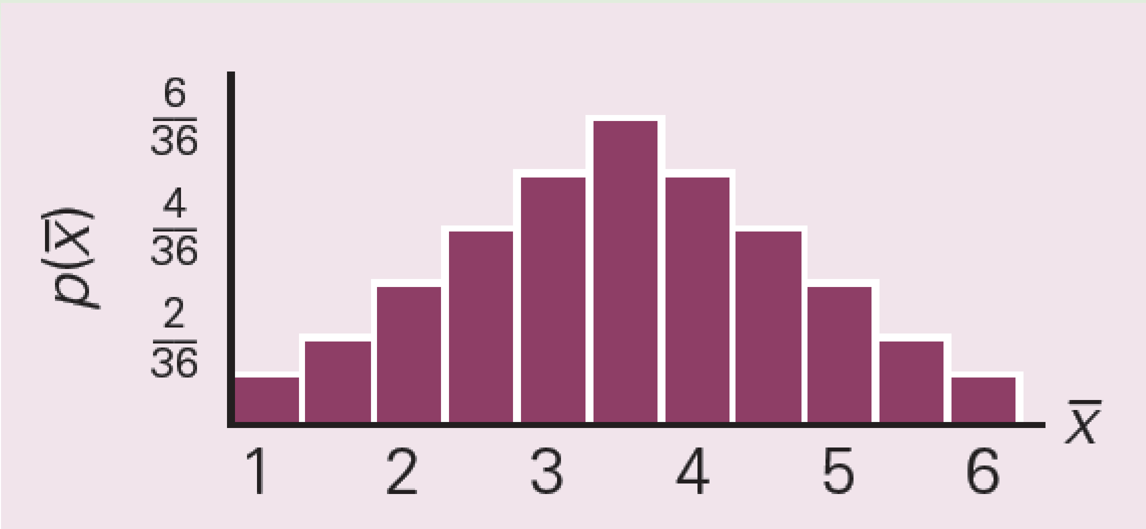
樣本	$\bar{x}$	樣本	$\bar{x}$	樣本	$\bar{x}$
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

一共有36 種不同的大小為2 的可能樣本；$\bar{x}$ 只能有11種不同的可能數值，某些特定的數值(例，$\bar{x}$ =3.5)出現的頻率比其他的數值(例，$\bar{x}$ =1)多。

兩顆骰子平均點數的抽樣分配

抽樣分配結果如下：

$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



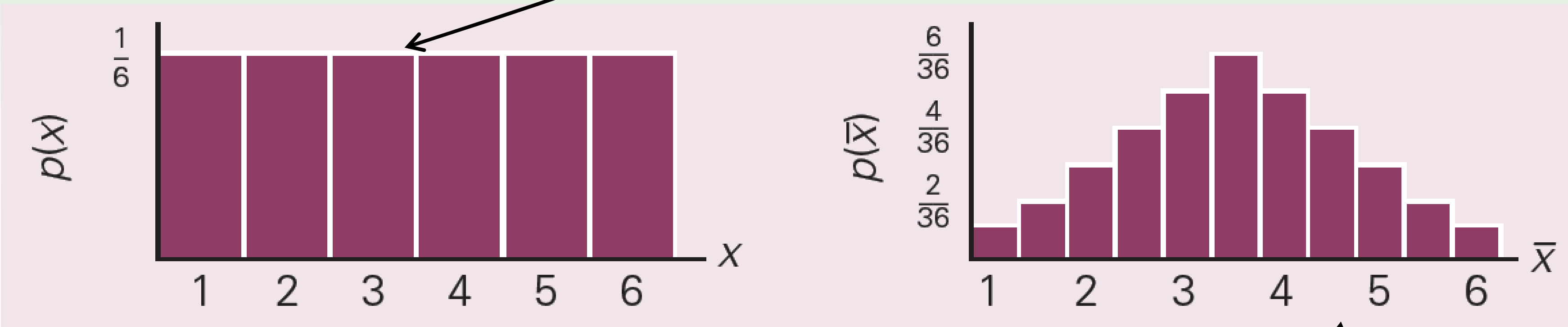
$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

$$\sigma_{\bar{x}}^2 = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{1.46} = 1.21$$

辨別

辨別X的分配



與樣本抽樣的 $\bar{x}$ 。

所以可以得知：
$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}}^2 = \sigma^2/2$$

推論...

- 為了產生 $\overline{X}$ 分配，我們從母體中重複抽出大小為 $n=2$ 的樣本，並且計算每一個樣本的 $\overline{x}$。因此，平均數以符號 $\mu_{\overline{x}}$ 表示，並且變異數以符號 $\sigma^2_{\overline{x}}$ 表示：

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

$$\sigma^2_{\bar{x}} = \sigma^2/2$$

但是使用其他的 n 值，我們會產生一些不同的 X 抽樣分配：

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

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

抽樣分配的標準差被稱為平均數的**標準誤(standard error of the mean)**。

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

Notes

- The sampling distributions of the sample mean for different sample sizes, see Figure 9.3 (p325).
  - Same mean but different variances
  - **Q**: what is the effect of the sample size on the sampling distribution of the sample mean?
- The distribution of the target population for sampling is different from the sampling distribution of the sample mean.

Notes

- The sampling distribution of the sample mean is symmetric.
- **Q**: The above result will be still held if the distribution of the target population for sampling is changed?
- A: No !!!!!

中央極限定理 (Central Limit Theorem, CLT)

- 自任何母體中隨機抽取的樣本，其樣本平均數的抽樣分配，在**樣本大小足夠大**時，會**趨近於**常態分配。
  - 樣本越大，**樣本平均數**的抽樣分配會越**趨近於**常態分配
  - **Q**: How large the sample size is required for CLT?
  - **Importance**: CLT is not mattered for the change of the distribution of the population.

Note: CLT is an important theorem in statistics !!!!!

Notes for CLT

如果母體是常態的，則對所有的n值，$\overline{X}$是常態分配。

如果母體為非常態的，只有在n值比較大的情況下，$\overline{X}$才會近似常態分配。

在許多實務的狀況下，樣本大小30可能足以讓我們使用常態分配近似$\overline{X}$的抽樣分配。

任何母體平均數的抽樣分配

我們可以將我們已有的發現延伸到所有無限大的母體。但是，如果母體是有限的，則標準誤是：

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

其中 *N* 是母體大小並且

$$\sqrt{\frac{N-n}{N-1}}$$

為有限母體校正因子(**finite population correction factor**)。

任何母體平均數的抽樣分配

如果母體大小相對於樣本大小是夠大的，則有限母體校正因子會接近於1，可以忽略它。

我們將視至少比樣本大 20 倍的任何母體為大母體。

在實務中，大部分應用所涉及的母體都符合大母體的條件。

結果 ，有限母體校正因子通常被省略掉 。

樣本平均數的抽樣分配

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

2. $\sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

3. 如果 X 是常態，則 $\overline{X}$ 是常態。如果 X 為非常態，$\overline{X}$ 在樣本是「足夠大」的情況下，近似於常態分配。「足夠大」的定義則依 $\overline{X}$ 非常態性的情況而定。

EMPIRICAL SAMPLING DISTRIBUTION (P328-329)

範例 9.1

- 一間製瓶工廠的領班觀察到
  - 每一個32 盎司瓶子中**汽水的總量**實際上是一個具有平均數 32.2 盎司以及標準差 0.3 盎司的常態分配隨機變數。
- **Q (a)**: 如果有一位顧客買一瓶汽水，這瓶的容量超過32 盎司的機率為何？
- **Q (b)**: 如果有一位顧客買一箱裝**四瓶**的汽水，**這四瓶的平均容量**會大於32盎司的機率為何？

範例 9.1

我們想要找出P(X > 32)，其中 X 是常態分配，μ = 32.2 且 σ =.3，

$$P(X > 32) = P\left(\frac{X - \mu}{\sigma} > \frac{32 - 32.2}{.3}\right)$$

$$= P(Z > -.67)$$

$$= 1 - P(Z < -.67)$$

$$= 1 - .2514 = .7486$$

"一瓶汽水的容量超過32 盎司的機率為75%"

範例 9.1

現在我們想要找出$P(\overline{X} > 32)$，其中 X 是常態分配，$\mu = 32.2$ 且$\sigma = .3$。

我們知道：

1. $\overline{X}$ 服從常態分配

2. $\mu_{\overline{x}} = \mu = 32.2$

3. $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .3/\sqrt{4} = .15$

範例 9.1

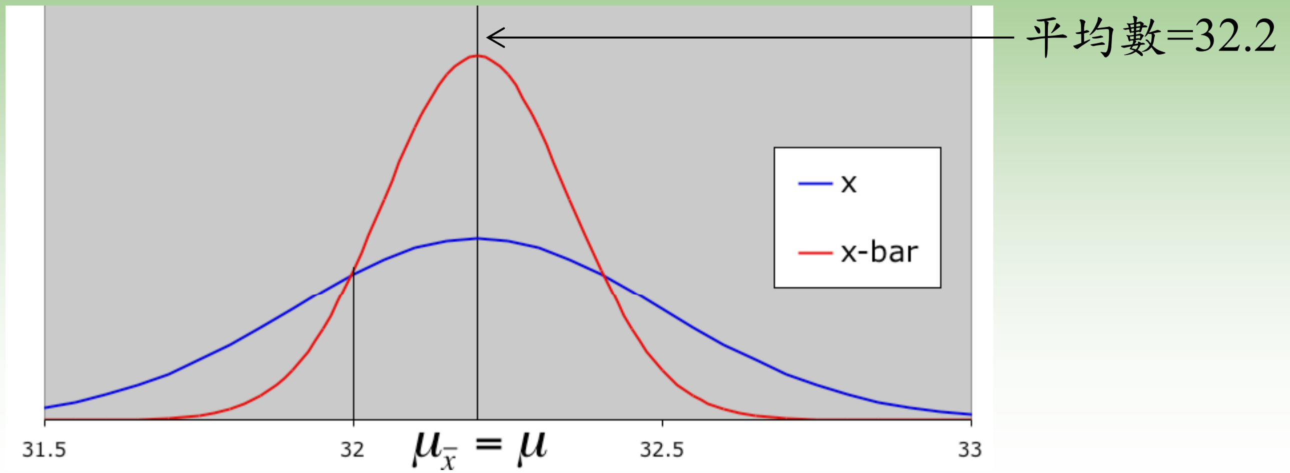
如果有一位顧客買一箱裝四瓶的汽水，**這四瓶的平均容量**會大於32盎司的機率為何？

$$P(\overline{X} > 32) = P\left(\frac{\overline{X} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} > \frac{32 - 32.2}{.15}\right)$$

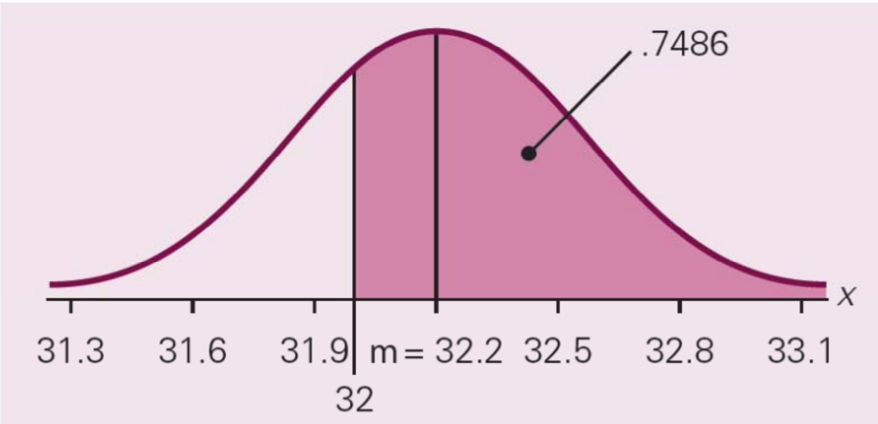
$$= P(Z > -1.33) = 1 - P(Z < -1.33) = 1 - .0918 = .9082$$

"四瓶汽水的平均容量會大於32盎司的機率為91%"

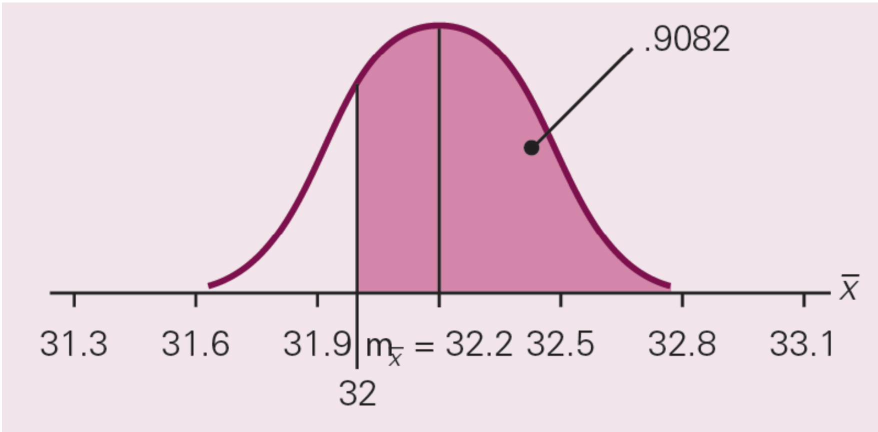
圖9.4　X 分配與$\overline{X}$ 的抽樣分配



容量超過32 盎司的機率為何？



四瓶的平均容量會大於32盎司的機率為何？



開章範例- 商學院畢業生的薪資

- 在一所大型大學的廣告中，商學院的院長宣稱
  - 其學院的畢業生在畢業一年之後的平均週薪是$800 且標準差是$100。
- 一位商學院二年級剛修完一門統計課的學生想要檢視院長所宣稱的平均數是否正確。
- 他調查了25 位一年前畢業的人並記錄他們的週薪
  - 發現樣本平均數是$750。

開章範例- 商學院畢業生的薪資

- 為了詮釋他的發現，當母體平均數是$800 且標準差是$100 時，他必須計算
  - 一個 包含 25 位畢業生的樣本會得到樣本平均數 $750 或更小的機率。
- 計算這個機率之後，他必須導出一些結論。
- X=? Population? Population distribution?

開章範例- 商學院畢業生的薪資

我們想要找出樣本平均數少於$ 750的機率。因此，我們尋找

$$P(\overline{X} < 750)$$

X 的分配，也就是週薪的分配，很可能是正偏的，但又不足以讓 $\overline{X}$ 分配為非常態。因此，我們可以假設 $\overline{X}$ 是常態的，具有平均數 $\mu_{\overline{x}} = \mu = 800$ 且標準差

$\sigma_{\overline{x}} = \sigma / \sqrt{n} = 100 / \sqrt{25} = 20$。

開章範例- 商學院畢業生的薪資

因此，

$$P(\overline{X} < 750) = P\left(\frac{\overline{X} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} < \frac{750 - 800}{20}\right)$$

$$= P(Z < -2.5) = .5 - .4938 = .0062$$

- 當母體平均數是$800 時，觀察到一個樣本平均數如同$750 一樣低的機率是極為低的。
- 因為這個事件不大可能發生，我們將結論
  - 這位院長的宣稱是不被證實的, 即院長的宣稱仍有待商確。

樣本平均數的抽樣分配

樣本平均數的抽樣分配可表達如下所示：

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

使用抽樣分配進行推論

我們可以代數的符號表達為

$$P(-1.96 < Z < 1.96) = .95$$

將 Z 的形式帶入先前的機率陳述之中，我們得到

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < 1.96\right) = .95$$

使用一點簡單的代數處理，我們得到

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

使用抽樣分配進行推論

回到本章開章範例，其中 $\mu = 800, \sigma = 100$，並且 $n = 25$，我們計算

$$P(800 - 1.96\frac{100}{\sqrt{25}} < \overline{X} < 800 + 1.96\frac{100}{\sqrt{25}}) = .95$$

因此我們可以說

$$P(760.8 < \overline{X} < 839.2) = .95$$

這告訴我們有 95% 的機率，一個樣本平均數會落在 760.8 和 839.2 之間。因為計算的樣本平均數是 \$750，我們將結論該院長的宣稱不被樣本統計量所支持的。

使用抽樣分配進行推論

將機率從 .95 改變成 .90 ，則機率的描述改變為

$$P(\mu - 1.645\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.645\frac{\sigma}{\sqrt{n}}) = .90$$

使用抽樣分配進行推論

我們也可以產生這個描述的一般式：

$$P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

在此公式中α (希臘字母 *alpha*) 是$\overline{X}$ 不落在區間內的機率。

為了應用此一公式，我們必須做的只是代入 μ, σ, n 和 α。

使用抽樣分配進行推論

例如，以$\mu = 800, \sigma = 100, n = 25$ 與$\alpha = .01$，我們可以得到

$$P\left(\mu - Z_{.005}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + Z_{.005}\frac{\sigma}{\sqrt{n}}\right) = 1 - .01$$

$$P\left(800 - 2.575\frac{100}{\sqrt{25}} < \overline{X} < 800 + 2.575\frac{100}{\sqrt{25}}\right) = .99$$

$$P(748.5 < \overline{X} < 851.5) = .99$$

Practice the exercises

- Ex 9.7;
- Ex 9.11;
- Ex 9.18

9.2 單一樣本比例的抽樣分配

母體成功比例的估計值是樣本比例。也就是，我們計數在一個**樣本**中成功的次數並計算：

$$\hat{P} = \frac{X}{n}$$

($\hat{P}$ 的讀法是 “p-hat”).

其中 X 是成功的次數，而 n 是樣本大小。

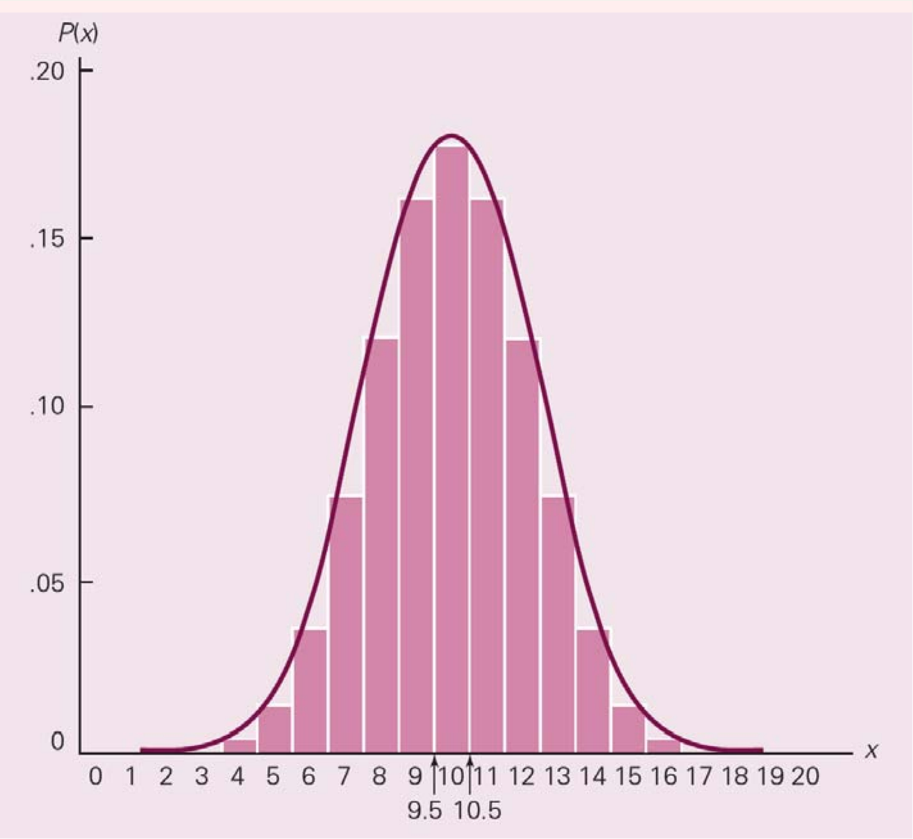
Q: How to compute P($\hat{P} \leqq 0.5$)? See p335.

Q: If X~B(100, 0.05), compute P(X>15)? Difficult…….

常態分配近似二項分配

***n* = 20 與*p* = .5 的二項分配且常態近似疊上**

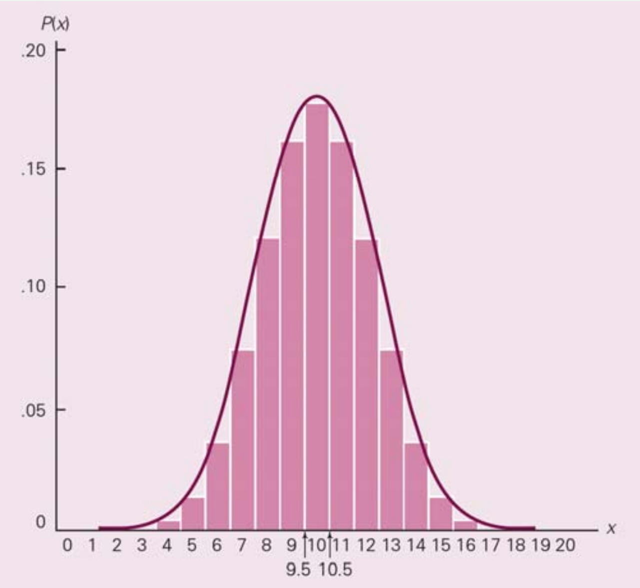
(μ =10 與 σ =2.24)



常態分配近似二項分配

$n = 20$ 與 $p = .5$ 的二項分配且常態近似疊上

(μ =10 and σ =2.24)



如何得知這兩個數值？

在7.4 節中我們指出：

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

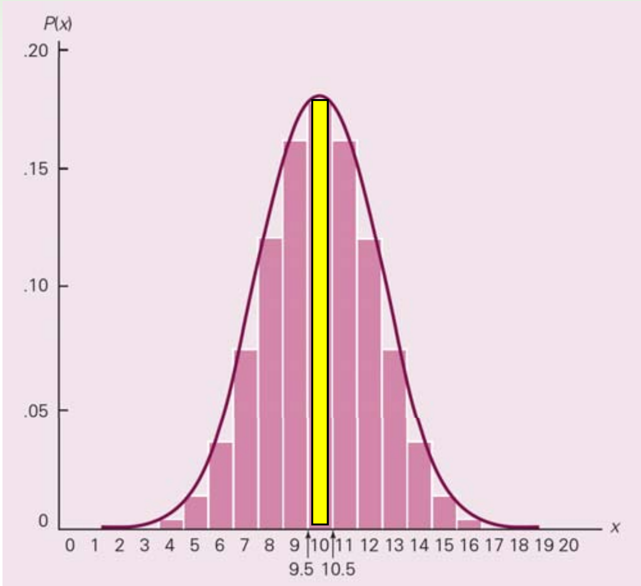
$$\sigma = \sqrt{np(1-p)}$$

因此：

$$\mu = np = 20(.5) = 10 \quad 且 \quad \sigma = \sqrt{np(1-p)} = \sqrt{20(.5)(1-.5)} = 2.24$$

常態分配近似二項分配

為了使用常態分配來計算機率，我們需要找出常態曲線下方介於9.5 與10.5 之間的面積。



$$P(X = 10) \approx P(9.5 < Y < 10.5)$$

其中***Y*** **是**近似**二項隨機變數*X*** 的**常態隨機變數**。

常態分配近似二項分配

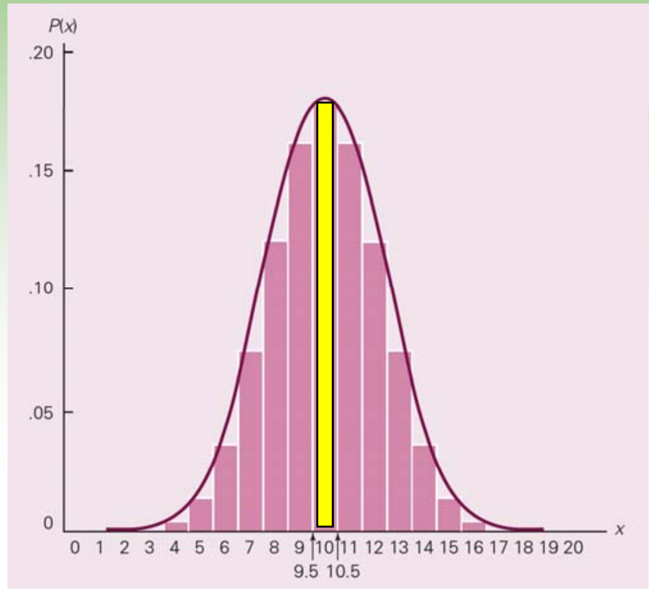
X等於10的實際機率是：

P(X = 10) = .1762

當

P(9.5 < Y < 10.5) = .1742

這近似值是相當好的。



$$P(X = 10) \approx P(9.5 < Y < 10.5)$$

其中***Y*** **是**近似**二項隨機變數** ***X*** 的**常態隨機變數**。

Notes

- Continuity correction factor (p338)
- $P(X \leqq 8)$ is approximated by $P(Y<8.5)$
- $P(X \geqq 14)$ is approximated by $P(Y>13.5)$
- What happen if the continuity correction factor is ignored? See p338-340

NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTION (P341)

單一樣本比例的抽樣分配

使用期望值與變異數法則，我們能夠得到 $\hat{P}$ 的平均數、變異數與標準差。

[$\hat{P}$ 的標準差被稱為**比例的標準誤(standard error of the proportion)**]

$$E(\hat{P}) = p$$

$$V(\hat{P}) = \sigma_{\hat{p}}^2 = \frac{p(1-p)^1}{n}$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

使用此公式，樣本比例可以被標準成標準常態分配：

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

範例 9.2

在上次的選舉，一位州代表得到 52% 的選票。

在選後的一年這位代表發起一個調查，詢問3 00 位民眾的隨機樣本，在下次的選舉中他們是否還會投給他。

如果我們假設他受歡迎的程度沒有改變，則此樣本中會有一半以上的民眾投票給他的機率為何？

範例9.2

將會投給這位代表的受訪者人數是一具有 $n = 300$ 與 $p = .52$ 的二項隨機變數。

我們想要決定樣本比例大於50% 的機率。也就是，我們要找出 $P(\hat{P} > .50)$。

我們現在知道樣本比例 $\hat{P}$ 是服從近似的常態分配，具有平均數 $p = .52$ 與標準差

$$\sqrt{p(1-p)/n} = \sqrt{(.52)(1-.52)/300} = .0288。$$

範例9.2

因此，我們計算

$$P(\hat{P} > .50) = P\left(\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} > \frac{.50 - .52}{.0288}\right)$$

$$= P(Z > -.69) = 1 - P(Z < -.69) = 1 - .2451 = .7549$$

如果我們假設支持度維持在 52%，則在 300位民眾的樣本中，超過半數會投票給此位代表的機率是 .7549。

9.3 兩個獨立樣本平均數差異的抽樣分配

另外一種你很快就會遇到的抽樣分配是**兩個獨立樣本平均數差異的抽樣分配 (sampling distribution of the difference between two independent sample means)**。抽樣計畫需要：

➔ 從**兩個獨立常態**母體中各抽出一個隨機樣本。

如果兩個母體是常態，兩樣本平均數間的**差異**(例，$\bar{X}_1 - \bar{X}_2$)也是常態分配。

注意：如果兩母體皆為**非**常態分配，但是樣本量是”大的“，則 $\bar{X}_1 - \bar{X}_2$ 為近似常態。

兩個獨立樣本平均數差異的抽樣分配

經由使用期望值與變異數法則，我們導出 $\bar{X}_1 - \bar{X}_2$ 抽樣分配的期望值：

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

標準差 [即為兩平均數間差異的標準誤**(standard error of the difference between two means)**]：

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

範例 9.3

假設 Wilfrid Laurier University (WLU) MBA 畢業生與 University of WesternOntario (UWO) MBA 畢業生的起薪是常態分配，且具有平均數與標準差。如果WLU 以及UWO 的 MBA 畢業生的隨機樣本分別被選出，如下所示……

		WLU	UWO
平均數	μ	62,000 $/年	60,000 $/年
標準差	σ	14,500 $/年	18,300 $/年
樣本大小	**n**	50	60

WLU 畢業生的樣本平均起薪超過UWO 畢業生的樣本平均起薪的機率為何？

範例9.3

我們想要決定$P(\overline{X}_1 > \overline{X}_2 > 0)$。我們知道 $\overline{X}_1 > \overline{X}_2$ 是常態分配且具平均數$\mu_1 - \mu_2 = 62{,}000 - 60{,}000 = 2{,}000$ 與標準差

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{14{,}500^2}{50} + \frac{18{,}300^2}{60}} = 3{,}128$$

我們可以將此項變數標準化並且參考附錄B 的表3：

$$P(\overline{X}_1 - \overline{X}_2 > 0) = P\left(\boxed{\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}} > \frac{0 - 2000}{3128} \right) = P(Z > -.64) = .5 + .2389 = .7389$$

Z

"對於50 位從WLU 畢業生與60 位UWO 畢業生的樣本而言，WLU 的樣本平均起薪超過UWO 的樣本平均起薪的機率是.7389"

9.4 由此出發談統計推論

- In order to understand how the inferential statistics is affected by the sampling distribution, we may need to know how we reach this status.
- 在第7 與第8 章中我們介紹過機率分配，它讓我們對隨機變數的數值做機率的描述。
  - 這項計算的先決條件是知道隨機變數的分配與相關的參數。 However these fact are not held in the real life……

由此出發談統計推論

- 範例7.9 中，我們需要知道 Pat Statsdud 猜到正確答案的機率是 20% (p = .2) 以及在10 題 (測驗) 中正確答案的題數 (成功) 是一個二項隨機變數。
- 然後我們可以計算任何成功次數的機率。

由此出發談統計推論

在範例 8.3 中，我們必須知道投資的報酬率服從平均數 10% 與標準差 5% 的常態分配。

這三點資訊讓我們能夠計算隨機變數各種數值的機率。

由此出發談統計推論

下圖 以符號表示機率分配的使用。

簡單地說，對母體與其參數的了解讓我們使用機率分配對母體中的個別成員做機率描述。箭頭的方向表示資料流動的方向。

母體和其參數 $\xrightarrow{\text{機率分配}}$ 個別成員

由此出發談統計推論

在這一章中我們發展了抽樣分配，基於對參數的認知和一些關於分配的資訊，讓我們得以對一個樣本統計量做機率的描述。

母體和其參數 $\xrightarrow{\text{抽樣分配}}$ 統計量

Statistical works 藉著逆轉圖中資料流的方向

由此出發談統計推論

- 在機率分配與抽樣分配兩種應用中，我們必須知道相關參數的數值，這是一個相當難辦到的事情。
- 在實務的世界中，由於參數代表極大母體的敘述性量數，參數幾乎全都是未知的。
- 從第10 章開始，我們將假設大部分的母體參數是未知的。
- 統計實作人員會從母體中抽樣並且計算所需的統計量。
- 統計量的抽樣分配讓我們可以對參數做推論。

由此出發談統計推論

這張圖呈現統計推論的特色