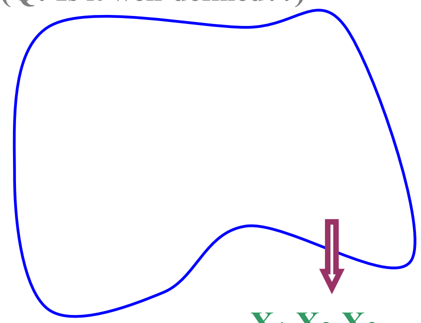
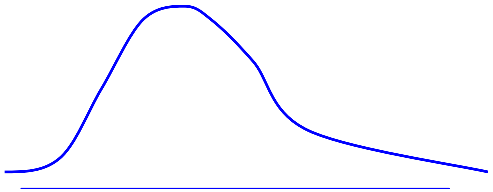
Lecture 6. Sampling Distributions

抽樣分布

Population (Q: Is it well-defined??)



Distribution $X_1, X_2, X_3, \ldots, X_n$



Distributional function 分布函數

$$F(\theta)$$

Random sampling 隨機取樣

Characteristics of the population (distribution), e.g., mean? median? mode? standard deviation? skewness? kurtosis?

Parameter(s) 參數，母數

Statistic(s) 統計量

- Constructed to **estimate** a parameter (characteristic).
- E.g., the **sample mean** is constructed to estimate the **population mean** if the sample is **randomly** collected. Similarly, the **sample variance** is constructed to estimate the **population variance**, … etc.

- The ‘**variation**’ **property** of a statistic under repeated samplings.（若能允許在同樣的狀況下重複作抽樣，則每一次抽樣的結果必定會有所不同。此變動性使得每一次的統計量實現值(**realization**)都會不一樣）

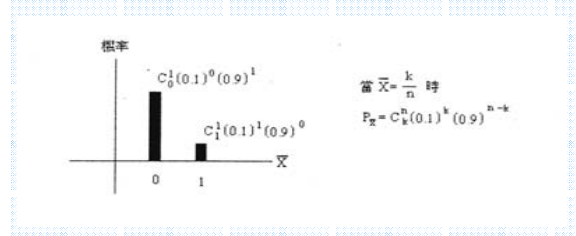
- **Sampling distribution** of a **statistic**

-- sampling distribution of the **sample mean**

-- sampling distribution of the **sample variance**

Example 1

假設 X 表投擲一個銅板的結果，若是正面(head), 則令 X=1；反面(tail), 則令 X=0；設 P(X=1)=p. 此隨機變數 X 服從於 Bernoulli 分布, X~Ber(p)。如果 p=0.1，則 X 的分布圖為：



Q:獨立地丟擲銅板兩次，得 X_1 和 X_2，則$(X_1+X_2)/2$=????

(sample mean (樣本平均數) $\equiv \bar{x}$)

Ans: 事實上$(X_1+X_2)/2$ 不會是一個數值，而是一個分布

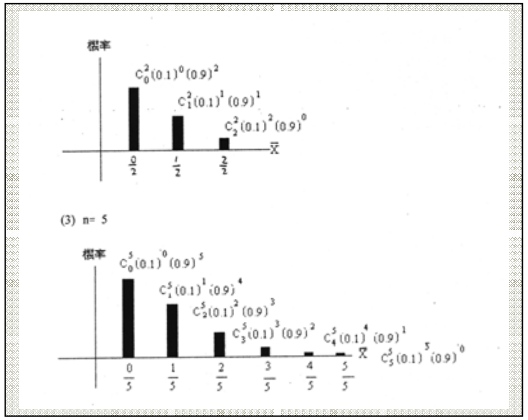
亦即$(X_1+X_2)/2=0$ 之機率為 $0.9\times0.9 =$ 0.81

$(X_1+X_2)/2=0.5$ 之機率為 $2\times0.9\times0.1 = 0.18$

$(X_1+X_2)/2=1$ 之機率為 $0.1\times0.1 =$ 0.01

所有的可能加起來：1.00

上述分布稱為：樣本數為 2 時樣本平均數($\bar{x}$)之抽樣分布。



Example 1. (Cont.)

繼續此問題，若問：獨立地丟擲銅板五次，得 $X_1,X_2,\ldots,X_5$，

則 $(X_1+X_2+X_3+X_4+X_5)/5$=????

↓

sample mean (樣本平均數) $\equiv \bar{x}$

同樣地，$(X_1+X_2+X_3+X_4+X_5)/5$=不是一個數值，而是一個分布，亦即

$(X_1+X_2+X_3+X_4+X_5)/5=$

0	之機率為	${}_5C_0(0.1)^0\ (0.9)^5 =0.59049$
0.2	之機率為	${}_5C_1(0.1)^1\ (0.9)^4 =0.32805$
0.4	之機率為	${}_5C_2(0.1)^2\ (0.9)^3 =0.07290$
0.6	之機率為	${}_5C_3(0.1)^3\ (0.9)^2 =0.00810$
0.8	之機率為	${}_5C_4(0.1)^4\ (0.9)^1 =0.00045$
1.0	之機率為	${}_5C_5(0.1)^5\ (0.9)^0 =0.00001$

所有的可能加起來：1.00000

上述分布稱為：「樣本數為 5」時，樣本平均數($\bar{x}$)之抽樣分布。 分布圖為：(See the last page)

Example 1. (Cont.)

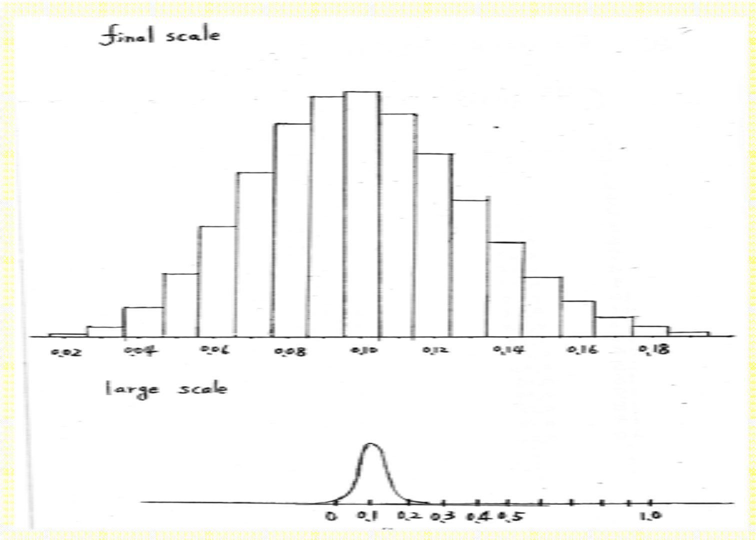
再問：獨立地丟擲銅板 100 次，得 $X_1,X_2,\ldots,X_{100}$，則 sample mean (樣本平均數) $\bar{x} \equiv (X_1+X_2+\ldots+X_{100})/100$=?

同樣地，$\bar{x}$ 有其分布；依公式

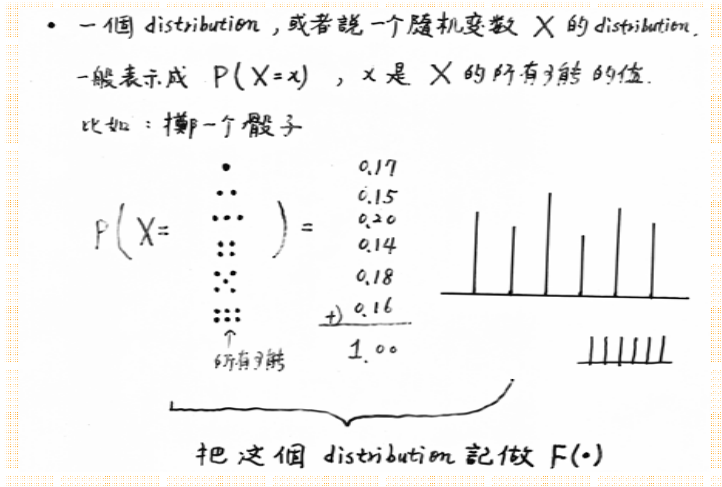
$P(\Sigma X_i=x)={}_{100}C_x(0.1)^x(0.9)^{100-x}$ ，得下表：

x	$\bar{X}$	Probability	x	$\bar{X}$	Probability	x	$\bar{X}$	Probability
0	0.00	0.00002656	8	0.08	0.11482303	16	0.16	0.01929172
1	0.01	0.00029513	9	0.09	0.13041628	17	0.17	0.01059153
2	0.02	0.00162320	10	0.10	0.13186535	18	0.18	0.00542653
3	0.03	0.00589160	11	0.11	0.11987759	19	0.19	0.00260219
4	0.04	0.01587460	12	0.12	0.09878801	20	0.20	0.00117099
5	0.05	0.03386580	13	0.13	0.07430210	21	0.21	0.00049566
6	0.06	0.05957873	14	0.14	0.05130383	22	0.22	0.00019776
7	0.07	0.08889525	15	0.15	0.03268244	23	0.23	0.00007452

所有可能機率加起來等於 1.00。分布圖為：

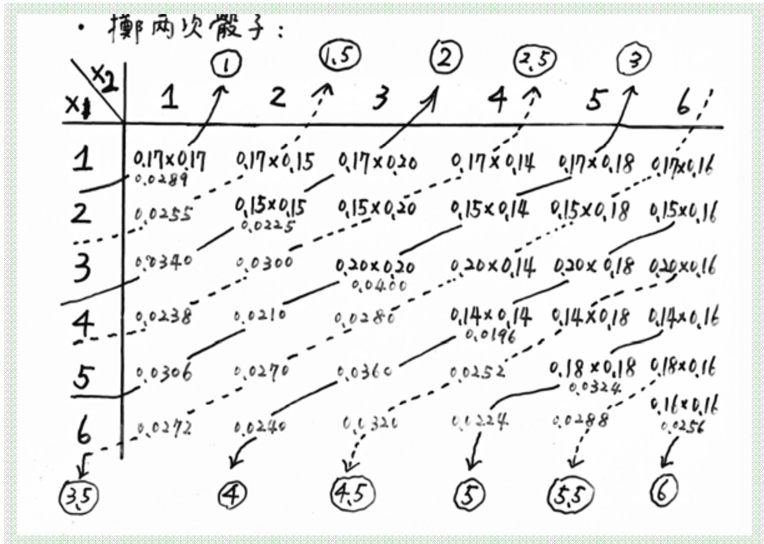


Example 2



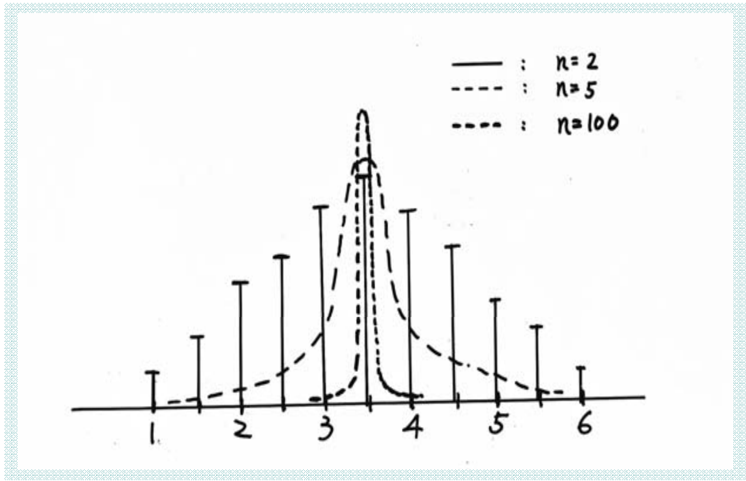
Question:從 F(•)中抽取兩個樣本（比如，擲骰子兩次），得到 X_1 和 X_2，則 $\bar{X} = (X_1 + X_2)/2$ 的抽樣分布(sampling distribution)為何??

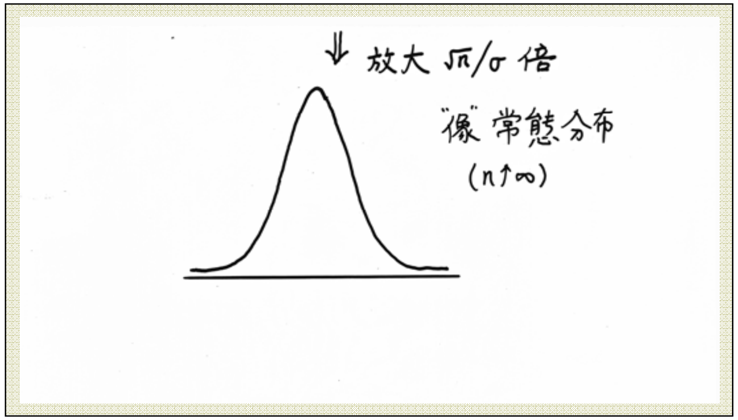
Answer: 考慮以下之所有可能組成(configurations)：



$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5
組成	X_1=1 X_2=1	X_1=1 X_2=2; X_1=2 X_2=1	X_1=1 X_2=3; X_1=3 X_2=1; X_1=2 X_2=2	X_1=4 X_2=1; X_1=3 X_2=2; X_1=2 X_2=3; X_1=1 X_2=4;	X_1=5 X_2=1; X_1=4 X_2=2; X_1=3 X_2=3; X_1=2 X_2=4; X_1=1 X_2=5	X_1=6 X_2=1; X_1=5 X_2=2; X_1=4 X_2=3; X_1=3 X_2=4; X_1=2 X_2=5 ; X_1=1 X_2=6
機率	0.0289	0.0255×2 =0.0510	0.0340×2 +0.0225 =0.0905	0.1076	0.1432	0.1644
$\bar{X}$	4.0	4.5	5.0	5.5	6.0	
組成	X_1=6 X_2=2; X_1=5 X_2=3; X_1=4 X_2=4; X_1=3 X_2=5; X_1=2 X_2=6	X_1=6 X_2=3; X_1=5 X_2=4; X_1=4 X_2=5; X_1=3 X_2=6;	X_1=6 X_2=4; X_1=5 X_2=5; X_1=4 X_2=6	X_1=6 X_2=5; X_1=5 X_2=6	X_1=6 X_2=6	
機率	0.1396	0.1144	0.0772	0.0576	0.0256	Σ=1.0000

$\bar{X}$ 之抽樣分布(sampling distribution)圖形：





樣本平均數之抽樣分布

(Sampling Distribution of the Sample Mean)

• 在前面的兩個 examples 裡，原始分布(寫成 F(•))均為離散型 (discrete-type)，當樣本數(n)很小時，欲計算 $\bar{X}$ 之抽樣分布，理論上並不困難。但是當樣本數(n)很大時，$\bar{X}$ 的抽樣分布之計算，必須借助中央極限定裡(Central Limit Theorem, CLT)以求得 $\bar{X}$ 的抽樣分布的一個近似(approximation)!

• 當原始分布 F(•)為連續型 (continuous-type)時，情況類似。

• 當原始分布 F(•)為常態分布 $N(\mu,\sigma^2)$時? ⇩

CONCLUSIONS: If there is a random sample drawn from a population with distribution F(•), i.e., $X_1,X_2,\ldots,X_n \sim F(\bullet)$ &

$E(X)=\mu$，$Var(X)=\sigma^2$

- F(•)= $N(\mu,\sigma^2)$, then $\bar{X} \sim N(\mu,\sigma^2/n)$, **exactly** (for all **n**)!!
- F(•)≠ Normal, then $\bar{X} \sim N(\mu,\sigma^2/n)$, **approximately** (for large **n**)!! **(CLT)**

樣本變異數之抽樣分布★

(Sampling Distribution of the Sample Variance)

• Specifically, if $X_1,X_2,\ldots,X_n \sim N(\mu,\sigma^2)$

We have $(X_1-\mu)/\sigma \sim N(0,1)$ ⇐ **(standardization!),**

$(X_2-\mu)/\sigma \sim N(0,1)$, ……………………,

$(X_n-\mu)/\sigma \sim N(0,1)$;

or, $(X_1-\mu)^2/\sigma^2 \sim \chi^2_1$

$(X_2-\mu)^2/\sigma^2 \sim \chi^2_1$, ……………………,

$(X_n-\mu)^2/\sigma^2 \sim \chi^2_1$;

Note: $E(\chi^2_v)=v$, $Var(\chi^2_v)=2v$, v=degree of freedom, d.f.

$\Sigma_i (X_i-\mu)^2/\sigma^2 \sim \chi^2_n$ or $(1/\sigma^2)\, \Sigma_i (X_i-\mu)^2 \sim$

$(1/\sigma^2)\, \Sigma_i (X_i-\bar{X})^2 \sim \chi^2_{n-1}$

or, $[\Sigma_i (X_i-\bar{X})^2]/(n-1) \sim \sigma^2 \chi^2_{n-1}/(n-1)$ [see Q1]

$E[\Sigma_i (X_i-\bar{X})^2/(n-1)]=\sigma^2\, E[\chi^2_{n-1}]/(n-1)=\sigma^2$

We say the statistic $s^2 \equiv \Sigma_i (X_i-\bar{X})^2/(n-1)$ **(sample variance) is an unbiased estimator of** σ^2 **(**樣本變異數 s^2 為母體變異數 (population variance) σ^2 的一個**不偏估計式(值)**。)

Q1: $\Sigma_i (X_i-\mu)^2/\sigma^2=(1/\sigma^2)\, \Sigma_i (X_i-\bar{X})^2+(n/\sigma^2)(\bar{X}-\mu)^2$, or conveniently, $T_n=T_{n-1}+T_1$. It can be shown that: (1) T_n is distributed as χ^2_n and T_1 is a χ^2_1-variate. (2) T_1 and T_{n-1} are mutually independent. (3) Then, from (1) and (2), T_{n-1} is χ^2_{n-1}.

(REF: Seber and Lee (2003) Linear Regression Analysis, 2nd ed. Chapter 2. (Wiley, New York).