Lecture 3: Probability based on Set Theory

9/29/2021

- **Last week**
  - Review: conditional probability

    $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$ (given that $\Pr(B) \neq 0$).

    Equivalent relationship: $\Pr(A \cap B) = \Pr(A \mid B)\Pr(B)$

    $= \Pr(B \mid A)\Pr(A)$

    解讀: The probability of event A given that B must occur.
  - When two events are mutually exclusive, they can not be independent.
    - Mutually exclusive (互斥) $\mathrm{P}(A \cap B) = \mathrm{P}(\phi) = 0$
    - Independence (獨立)

      $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \Leftrightarrow \mathrm{P}(A \mid B) = \mathrm{P}(A) \Leftrightarrow \mathrm{P}(B \mid A) = \mathrm{P}(B)$
    - $\mathrm{P}(A \cap B) = \mathrm{P}(\phi) = 0 \neq P(A)P(B)$

重要題型: Check whether two events are independent

Criteria: (choose either one rule)

1. $\Pr(A \cap B) = \Pr(A)\Pr(B)$
2. $\Pr(A \mid B) = \Pr(A)$ or $\Pr(B \mid A) = \Pr(B)$

Ex: A_1 = pass the exam A_2 = go to a cram school or get a tutor

Facts: $P(A_1) = 0.2$, $P(A_2) = 0.7$, $P(A_1 \cap A_2^C) = 0.02$ (pass but self-study)

Are the two events independent?

方法一: 比較

$P(A_1 \cap A_2^C) = 0.02$

$P(A_1)P(A_2^C) = 0.2 \times 0.3 = 0.06$

Since $P(A_1 \cap A_2^C) \neq P(A_1)P(A_2^C)$, A_1 and A_2 are NOT independent.

方法二: 由條件機率判斷

- $\Pr(A_2 \mid A_1) = \dfrac{\Pr(A_1 \cap A_2)}{\Pr(A_1)} = 0.9 \neq \Pr(A_2) = 0.7$ → A_1 and A_2 are NOT independent.

  關聯性的方向

  $\Pr(A_2 \mid A_1) = 0.9 > \Pr(A_2) = 0.7$

  在 A_1 中，A_2 的比例更高 (通過考試的人之中，有補習的比例高過整體有補習的比例)，代表正相關

- $\Pr(A_1 \mid A_2) = \dfrac{\Pr(A_1 \cap A_2)}{\Pr(A_2)} = \dfrac{0.18}{0.7} = 0.257 > \Pr(A_1) = 0.2$

  補習的人之中，通過考試的比例大於整體錄取率，代表正相關

Remark: You can use any one of the above methods to justify your answer.

- **Useful techniques of counting (計數的技巧)**
  - $P(A) = \dfrac{\#\text{ in A}}{\#\text{ in S}}$
  - We apply techniques of permutations (排列) or combinations (組合) to count the number of events

Ex 1. A couple have 3 children and we would like to know their genders

Event A: two girls;

Event B: the first one is a boy

Method 1: List all possible outcomes

$\omega_1 : MMM$,

$\omega_2 : MMF$, $\omega_3 : MFM$, $\omega_4 : MFF$,

$\omega_5 : FMM$, $\omega_6 : FMF$, $\omega_7 : FFM$,

$\omega_8 : FFF$

Method 2: Use some "tricks" to simplify the calculation

Total # in S = $2\times 2\times 2 = 8$

# of A → $\binom{3}{2} = 3$ → the positions of the two girls: (1,2), (1,3), (2,3)

# of B =1*2*2 = 4

$P(A) = \frac{3}{8}$,

$P(B) = \frac{4}{8}$

Ex 2. A couple have 5 children and we would like to know their genders

Event A = exactly 3 males and 2 females

Method 1: list all possible outcomes

Total number in the sample space = $2^5 = 32$

Possible positions of the two females

(1,2), (1,3), (1,4), (1,5),

(2,3), (2,4), (2,5),

(3,4), (3,5),

(4,4)

Use the combination rule

in A = $\binom{5}{3} = \frac{5!}{3!2!} = \frac{5*4}{2} = 10$ → $\Pr(A) = \frac{10}{32}$

→ $\Pr(A) = \frac{10}{32}$

Note:

In Ex1, Ex2, each position has only two possible outcomes.

What if one position can have more than 2 possible outcomes?

Ex3: Record the glucose level (血糖) 5 times, each of which has 3 levels: n, h, l

Total number in the sample space = 3^5

Event A = 2 normal, 2 high and one low

in A = $\frac{5!}{2!2!1!} = \frac{5*4*3*2*1}{2*2} = 30$ → combination rule

$\Pr(A) = \frac{30}{3^5}$

If you try to do direct calculations:

first fix the position of "normal", then locate "high" and "low"

Normal: (1,2) → 3 positions left for 2 "high" and one "low

NNHHL, NNHLH, NNLHH → 3 outcomes

Normal: (1,3)

NHNHL, NHNLH, NLNHH → 3 outcomes

Normal: (1,4)

Normal: (1,5)

Normal: (2,3)

Normal: (2,4)

Normal: (2,5)

Normal: (3,4)

Normal: (3,5)

Normal: (4,5)

in A = $\binom{5}{2} \times \binom{3}{2} = 10 \times 3$ 種

Conclusion:

Applying the rule of combination or permutation will simplify the work of counting.

Example:

Using the independence property to approximate the answer of sensitive questions

Design the problem

Question A: non-sensitive question (The last digit of your ID number is odd)

Question B: sensitive question

→ the major interest (e.g. any experience of cheating in exams)

Procedure:

- Flip a coin (or other random experiment)
  - If head occurs, answer question A
  - If tail occurs, answer question B.

Suppose 100 students participated the game and 60 answered YES.

Useful fact: Since A and B are independent, $\Pr(B \mid A) = \Pr(B)$

Pr(Yes) = Pr(solve question A and answer "yes")

+ Pr(solve question B and answer "yes")

= Pr(head occurs and ID is odd) + Pr(tail occurs and cheat)

≈ 1/2 * 1/2 + Pr(cheat | tail occurs) Pr(tail occurs)

= 1/2 * 1/2 + Pr(cheat) Pr(tail occurs) (based on independence)

≈ 1/2 * 1/2 + Pr(cheat) * 1/2

≈ 0.6

Answer: Pr(cheat) ≈ (0.6-0.25)/0.5 = 0.7

Remark: Such an approximation will be more accurate if there are more people participating in the experiment.

Applications of Probability Theory in Biomedical studies

a. 孟得爾遺傳率 (Mendelian Inheritance)

b. 流行病學的應用 (Epidemiology) → related to "conditional probability"

機率在基礎遺傳學的應用

Elementary Genetics (Optional)

The hereditary characteristics of an organism are determined by units called *genes*. Genes occur in pairs in an individual and come in contrasting forms. These forms are called *alleles*. For example, consider the gene that determines the height of a pea plant. This gene has two alleles, *T* for tallness and *t* for dwarfism. Thus there are three possible genetic compositions, or *genotypes*, with respect to this trait. They are *TT*, *Tt*, and *tt*. When the two genes are of the same form, we say that the organism is *homozygous* for the given trait; otherwise, it is *heterozygous*. A trait that will appear when the allele for the trait is present is called a *dominant* trait, and the allele is the dominant allele. Its contrasting trait or allele is said to be *recessive*. In the case of pea plants, the allele for tallness is dominant. Thus the genotypes *TT* and *Tt* will result in a tall plant, while the genotype *tt* will result in a dwarfed plant. Notationally, dominant alleles are denoted by capital letters, and recessive alleles are written as lowercase letters. For each trait the offspring inherits one gene randomly from each parent.

Ex1: genotype for height (suppose a single gene determines the phenotype)

"homozygous" (同型合子) vs. "heterozygous" (異型合子)

孟德爾的豌豆實驗就是一種完全顯性。例如，表現出高莖豌豆或矮莖豌豆都受同一基因座上的基因控制(T或t)，這對位於同一基因座的基因稱為「對偶基因(alleles)」或「等位基因」。TT(顯性同型合子，dominant homozygous)、Tt (異型合子，heterozygous)兩種基因型(genotype)表現高莖性狀(顯性性狀)；而tt (隱性同型合子，recessive homozygous)則表現矮莖(隱性性狀)。當只要有1個顯性基因出現時，此性狀將表現出顯性性狀。動物一些致死、半致死或缺陷基因也呈現完全顯性遺傳，且有害基因通常為隱性的。

T: tall (capital letter → dominant 顯性)

t: short (small letter → recessive 隱性)

EX: Both parents are heterozygous

Father: Tt

Mother: Tt

Offspring: 2*2 cases

TT (pure dominant)

tT Tt

tt (pure recessive)

A= the offspring is tall (TT, Tt, tT) → probability = 3/4

Note: you can plot a tree or a two-by-two table to describe all possible situations

EXAMPLE 2.2.3. Each member of a couple has alleles for both brown and blue eyes. In genetic terms they are heterozygous for eye color. In the case of eye color, the allele for brown eyes, which we denote by B, is dominant over that for blue eyes, b. That is, anyone with the B allele will have brown eyes. At conception, each parent contributes one allele for eye color. Hence we can view the experiment of determining the eye color of a child as a two-stage process. Stage 1 represents the inheritance of an allele from the mother; stage 2

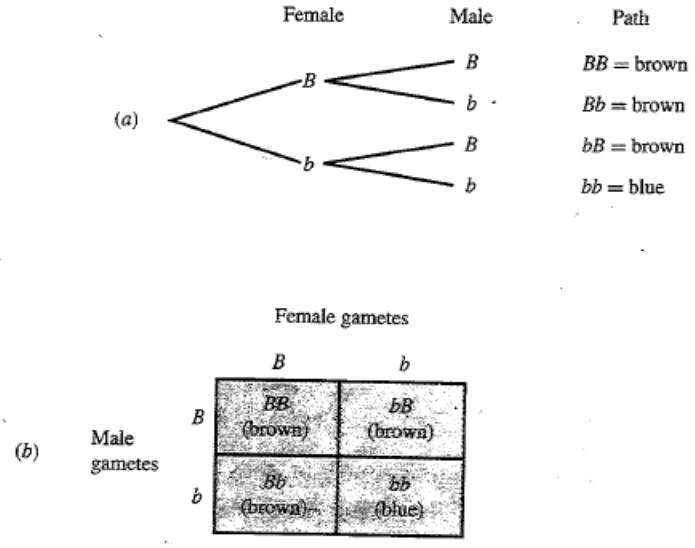


FIGURE 2.2
(a) Tree diagram for the inheritance of eye color from a couple, each of whom is heterozygous for eye color. (b) A biologist's representation of the problem as a Punnett square.

Example 2.2.3 Eye color (One trait, dominant gene)

Brown eye → dominant (B)

Blue eye → recessive (b)

Father: Bb, Mother: Bb

Offspring: Pr(Brown color) = 3/4; Pr(Blue color) = 1/4;

Example 2.2.4 Color of plant (One trait, no dominant gene)

EXAMPLE 2.2.4. The plant known as the four-o'clock can have red, white, or pink flowers. The allele for redness is denoted by R and that for white by r. A red flower has two R alleles and is said to be homozygous for color; a white flower is homozygous with genotype rr. When pure white plants are bred to pure red ones, the resulting flower has genotype Rr. Since there is no dominant allele, the resulting flower is pink. When two of these heterozygous plants are bred, the outcomes given in the tree of Figure 2.3 result. Each of the four paths through the tree is equally likely. By using classical probability we can conclude that the probability of obtaining a white flower from the cross-match is $\frac{1}{4}$.

Genotype: Red → R

White → r

Phenotype: RR → Red;

rr → White;

Rr, rR → Pink

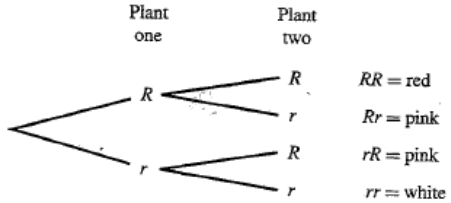


FIGURE 2.3
Outcomes that result when two heterozygous four-o'clock flowers are cross-matched.

Example 2.2.5 Two traits → 圖更複雜

skin (Albinis-白化症)

Earlobes (耳垂)

Trees can be used in a genetic setting to study more than one trait simultaneously. To do so, we simply extend the idea developed in the previous two examples.

EXAMPLE 2.2.5. In humans, the allele for normal skin pigmentation *S* is dominant over that for albinism *s*. The allele for free earlobes *F* is dominant over that for attached lobes *f*. A woman has genotype *SsFF*, and her husband has genotype *ssFf*. Hence the woman has normal skin pigmentation and free earlobes; her husband is albino with free earlobes. What are the possible outcomes for their offspring? We visualize this as a four-stage experiment with the following stages:

1. Inherit an allele for skin pigmentation from the mother
2. Inherit an allele for skin pigmentation from the father
3. Inherit an allele for ear formation from the mother
4. Inherit an allele for ear formation from the father

- Skin color
  - Normal skin → S
  - albinism → s
- Earlobe type
  - Free earlobes → F
  - attached earlobes → f

Parents' characteristics:

Mother: SsFF

Father: ssFf

Offspring's characteristics:

- skin color: 2*1 = 2 possible outcomes (膚色:母 2 種;父: 1 種，共 2 種可能)
- earlobe type: 1*2 = 2 possible outcomes (耳垂:母 1 種;父: 2 種，共 2 種可能)

Two traits combined: 2*2 = 4 possible outcomes (可任選樹狀圖或是 2-by-2 table)

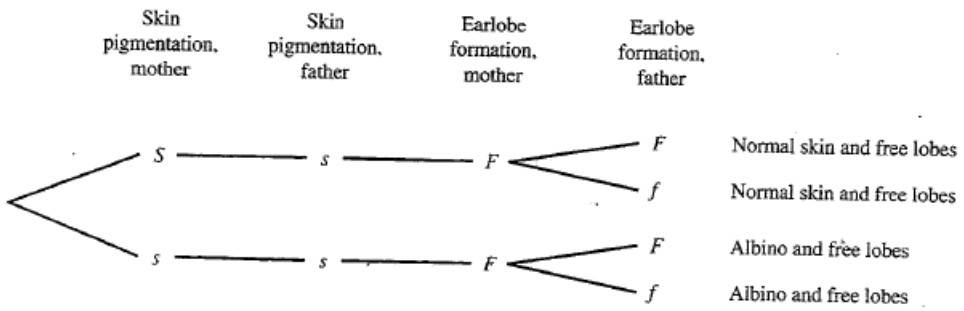


FIGURE 2.4
A four-stage tree used to study two traits simultaneously.

Offsprings have

- 4 possible genotypes: SsFF, SsFf, ssFF, ssFf
- 2 possible phenotypes:
  - "normal skin and free lobes"
  - "albino skin and free lobes"

		Earlobe	
		FF	Ff
Skin	Ss	SsFF	SsFf
	ss	ssFF	ssFf

Useful rules

* Rule of Addition (加法原則)

– 互斥事件的聯集機率 = 個別機率相加總

If $A \cap B = \phi$, $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

* Rule of multiplication (乘法原則)

- 交集的機率可寫成邊際機率與條件機率的乘積

for calculating the probability of an intersection

$$\Pr(A_1 \cap A_2) = \Pr(A_1)\Pr(A_2 \mid A_1)$$

$$\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1)\Pr(A_2 \cap A_3 \mid A_1)$$

$$= \Pr(A_1)\Pr(A_2 \mid A_1)\Pr(A_3 \mid A_1 \cap A_2)$$

…

$$\Pr(A_1 \cap A_2 \ldots \cap A_k) = \Pr(A_1)\Pr(A_2 \mid A_1)\Pr(A_3 \mid A_1 \cap A_2)\ldots\Pr(A_k \mid A_1 \cap \ldots \cap A_{k-1})$$

Note:

$$\Pr(A_2 \cap A_3) = \Pr(A_2)\Pr(A_3 \mid A_2)$$

$$\Pr(A_2 \cap A_3 \mid A_1) = \Pr(A_2 \mid A_1)\Pr(A_3 \mid A_2 \cap A_2)$$

Example: RH Blood type

Facts: Pr(RH + gene) = 0.61, Pr(RH - gene) = 0.39

Q: Please compute Pr(Ph – blood)

Pr(Ph – blood) = Pr(RH – gene & RH – gene) = 0.39*0.39 = 0.1521

Q: Please compute Pr(Ph + blood)

Method 1:

Pr(Ph + blood) = Pr(RH + gene & RH + gene)+ Pr(RH + gene & RH - gene)

Pr(RH + gene & RH + gene) = 0.61*0.61 = 0.3721

Pr(RH + gene & RH - gene) = 2*0.61*0.39 = 0.4758

Pr(Ph + blood) = 0.3721+0.4758 = 0.8479

Method 2:

Pr(Ph + blood) = 1 - Pr(Ph – blood) = 1 – 0.1521 = 0.8479

Example: calculation the probability of "blood incompatibility"

母嬰血型不合的溶血病

Rh 因子引起的問題，稱作 新生兒溶血性疾病(haemolytic disease of the newborn)，又叫胎性母紅血球增多病(erythroblastosis fetalis)，這是因為顯性遺傳的 Rh-D 抗原，在 RhD- 的母親，假如胎兒的紅血球是 RhD+ (因為父親是 RhD+ 的原故) ， 在第一次生產之後，胎兒的紅血球會刺激婦女產生 anti-D antibody

The use of the multiplication rule in a genetics setting is illustrated in Example 3.6.2.

EXAMPLE 3.6.2. When a mother is Rh negative and her child is Rh positive, a blood incompatibility exists that may lead to erythroblastosis fetalis, a condition in which the mother forms an antibody against fetal Rh which leads to the destruction of fetal red blood cells. What is the probability that a randomly selected child will have this condition?

One way for the child to have this problem is for the father to be Rh-positive heterozygous (+ − or − +) and pass a positive gene to the child while the mother is Rh negative. To find the probability of this combination of events we must find $P[(A_1 \text{ and } A_2) \text{ and } A_3]$ where A_1 denotes the event that the father is Rh-positive heterozygous, A_2 that the father passes a positive gene to the child, and A_3 that the mother is Rh negative. Notice that events A_1 and A_2 are not independent; the fact that the father is positive heterozygous does have a bearing on the child's ability to obtain a positive gene from this source. Via the multiplication rule,

$$P[A_1 \text{ and } A_2] = P[A_2 \mid A_1]P[A_1]$$

From Exercise 10 of Section 3.5. we know that $P[A_1] \doteq .48$. Since one gene is inherited at random from the father, $P[A_2 \mid A_1] = .5$. Hence

$$P[A_1 \text{ and } A_2] \doteq .5(.48) = .24$$

Since the mother's gene type has no effect on the father or on his ability to convey a positive gene to the child, A_3 is independent of A_1 and A_2. From Example 3.5.2, we know that $P[A_3] \doteq .15$. Hence by definition of independence,

$$P[(A_1 \text{ and } A_2) \text{ and } A_3] \doteq .24(.15) = .0360$$

Analysis 1: 分析問題

Pr(blood incompatibility)

= Pr(mother has RH – blood & child has RH + blood)

Note: We can NOT separate the intersection since the two events (mother and child) are NOT independent.

Analysis 2: 媽媽的狀況

Mother → Pr(RH – blood) = Pr(Rh - gene, Rh – gene) = 0.1521

Analysis 3: 爸爸的狀況 → 兩個可能性 (加法原則)

Father must have Rh + gene

(Rh + gene, Rh + gene) or (Rh + gene, Rh - gene)

Analysis 4: 父母對孩子患病的影響

The child must be Rh+ heterozygous (from mother: Rh – gene;from father:Rh + gene)

Q: Compute Pr(blood incompatibility)

Situation 1:

Pr(mother is Rh negative homozygous & father is Rh positive homozygous)

= Pr(mother is Rh negative homozygous) * Pr(father is Rh positive homozygous)

(Assume mother and father are independent)

Pr(father is Rh positive homozygous) = 0.61*0.61 = 0.37

Pr(mother is Rh negative homozygous) = 0.39*0.39 = 0.15

Probability of "situation 1" = 0.37*0.15 = 0.056

Situation 2: 乘法原則的應用

Pr(mother is Rh negative homozygous & father is Rh positive heterozygous & passes PH+ gene to the child) (the intersection of 3 sets)

A_1 = father is RH+, RH- → $\Pr(A_1) = 2*0.39*0.61 = 0.4758$

A_2 = mother is RH-,RH- → $\Pr(A_2) = 0.39*0.39 = 0.15$

A_3 = father passes RH+ gene to child

$\Pr(A_2 \cap A_1 \cap A_3)$

= $\Pr(A_2 \cap A_1) * \Pr(A_3 \mid A_2 \cap A_1)$ (first given the condition of parents)

= $\Pr(A_2 \cap A_1) * \Pr(A_3 \mid A_1)$ (remove the mother's information since it is irrelavant)

$= \Pr(A_2)\Pr(A_1) * \Pr(A_3 \mid A_1)$ (the blood types of mother and father are independent)

= 0.0724*0.5 = 0.036

Probability of "situation 2" = 0.036

A: Pr(blood incompatibility) = 0.056+0.036 = 0.092

(We add up the two probabilities since they are disjoint)

Conclusion:

By randomly selecting an infant, the probability of having "erythroblastosis fetalis" is 0.092. 隨機抽取一個嬰兒 (父母資訊未給定)有此溶血症問題的機率為 0.092

(或說在人群裡, 胎兒發生溶血症的例 = 0.092)

試劑的評估

False-positive and false-negative – 檢驗方法可能會犯的兩種錯誤

Terminology associated with diagnostic tests

		True state	
		Condition absent (−)	**Condition present (+)**
Test results	Condition found (+)	True − but tests + False-positive result P [false positive] = α	True + and tests + No error
	Condition not found (−)	True − and tests − No error	True + but tests − False − negative result P[false negative] = β

Event A: true status is positive

Event B: test status is positive

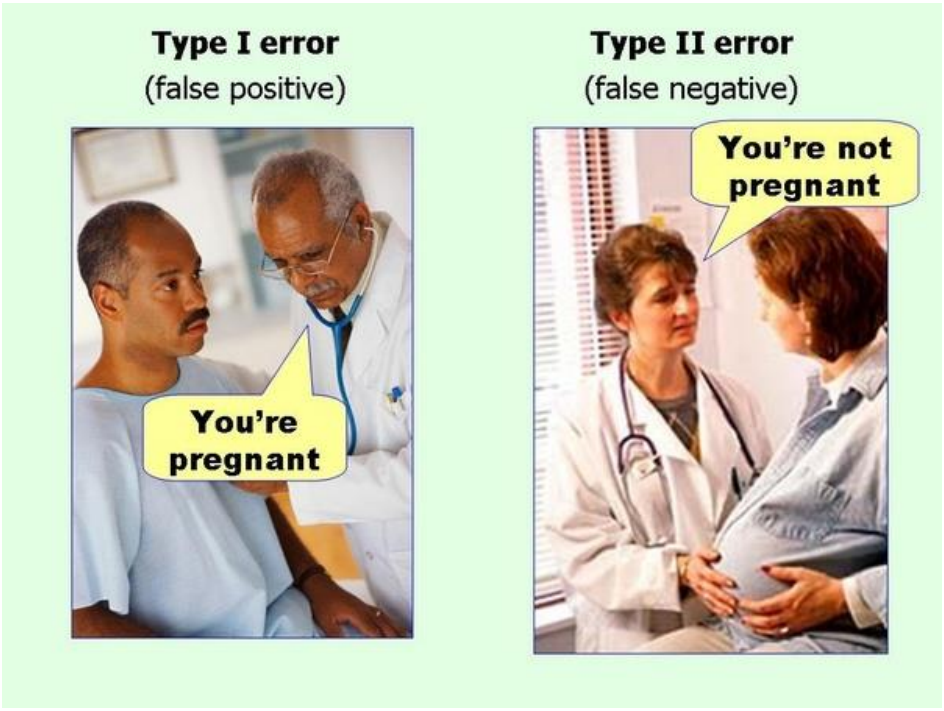
- false positive rate

  $\alpha = \Pr(\text{test positive} \mid \text{true negative}) = \Pr(B|A^c)$

  (沒有愛滋卻檢查出有; 沒有唐氏症卻檢查出有) → 虛驚一場

- false negative rate → $\beta = \Pr(\text{test negative} \mid \text{true positive}) = \Pr(B^C \mid A)$

  (有乳癌卻檢查不到) → 錯失良機



2. "specificity" and "sensitivity" – 評估檢驗方法的兩種標準

Event A: true status is positive; Event B: test status is positive

- Specificity (特異性):
  - Specificity is the ability to exclude persons who do not have the disease.
  - $\Pr(\text{test negative} \mid \text{true negative}) = \Pr(B^C \mid A^c)$
- Sensitivity (敏感度)
  - Sensitivity is the ability to detect a disease if it is really present.
  - $\Pr(\text{test positive} \mid \text{true positive}) = \Pr(B \mid A)$

Breast cancer diagnostic tests (乳房攝影, 超音波, 3-D MRI)

Sensitivity & Specificity
Mammogram Vs Ultrasound Vs MRI

	Sensitivity	Specificity
Mammogram	82%	99%
Ultrasound	86%	98%
MRI 3T	100%	94%

Haitham Elsamaloty et al . AJR 2009; 192:1142-1148, Increasing the accuracy of detection of Breast Cancer with 3-T MRI.

常用的資料呈現方法: two-by-two table

Made-up example:

	True + (A)	True -	Total
Test + (B)	80	10	90
Test -	20	90	110
Total	100	100	200

$\alpha = \Pr(\text{test positive} \mid \text{true negative}) = \Pr(B \mid A^c) = 0.1$

Specificity = $\Pr(\text{test negative} \mid \text{true negative}) = \Pr(B^C \mid A^c) = 0.9$

$\beta = \Pr(\text{test negative} \mid \text{true positive}) = \Pr(B^C \mid A) = 0.2$

sensitivity = $\Pr(\text{test positive} \mid \text{true positive}) = \Pr(B \mid A) = 0.8$

Example: Testing the gender of a fetus (胎兒)

	Sex		
Pregnancy Zone	**Male (true −)**	**Female (true +)**	
Present (test +)	51	78	129 (Random)
Absent (test −)	96	75	171 (Random)
	147 (Random)	153 (Random)	300 (Fixed)

definition the false-positive rate is

$$\alpha = P[\text{test} + \mid \text{true} -]$$

To estimate this conditional probability we must estimate P[true −] and P[test + and are true −]. Using the relative frequency approach to probability, $P[\text{true} -] \doteq 147/300$ and $P[\text{test} + \text{ and true} -] \doteq 51/300$. The definition of conditional probability yields

$$\alpha \doteq \frac{51/300}{147/300} = \frac{51}{147} = .3469$$

estimated false-positive rate. To estimate β, note that of the 153 true-positive subjects, 75 tested negative. Hence

$$\beta \doteq \frac{75}{153} = .4902$$

參考補充檔案 about COVID Test

ROC curve (will not be on the exam) 補充，不考

- ROC curve: Receiver operating characteristic curve
- A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

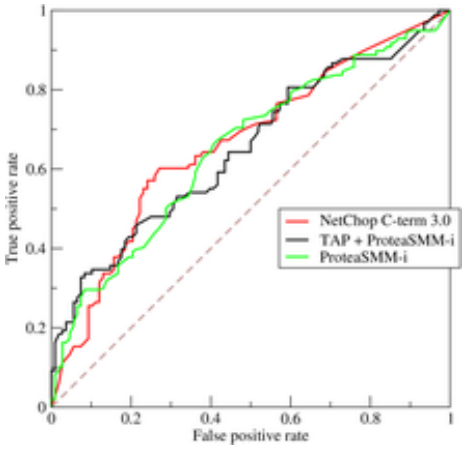
分類模型（又稱分類器，或診斷）是將一個實例映射到一個特定類的過程。ROC分析的是二元分類模型，也就是輸出結果只有兩種類別的模型，例如：（陽性 / 陰性）（有病 / 沒病）（垃圾郵件 / 非垃圾郵件）（敵軍 / 非敵軍）。

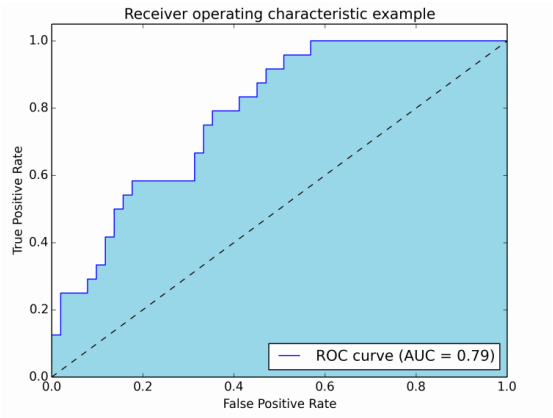
當**訊號偵測**（或**變數測量**）的結果是一個連續值時，類與類的邊界必須用一個閾值（英語：threshold）來界定。舉例來說，用血壓值來檢測一個人是否有高血壓，測出的血壓值是連續的實數（從0~200都有可能），以收縮壓140 / 舒張壓90為閾值，閾值以上便診斷為有高血壓，閾值未滿者診斷為無高血壓。二元分類模型的個案預測有四種結局：

1. **真陽性**（TP）：**診斷為有**，**實際上也有**高血壓。
2. **偽陽性**（FP）：**診斷為有**，**實際卻沒有**高血壓。
3. **真陰性**（TN）：**診斷為沒有**，**實際上也沒有**高血壓。
4. **偽陰性**（FN）：**診斷為沒有**，**實際卻有**高血壓。

Given a threshold: (閥值，決定有病沒病的切點)

- X-axis: the false positive rate (FPR) → smaller, the better
- Y-axis: the true positive rate (TPR) = sensitivity → larger, the better
- For a single test, we will select a threshold if it produces a point near to the upper-left corner (X small, Y large).
- We can use AUC (area under the ROC curve) to compare several tests.





Relative Risk – 相對風險 (判斷某個風險因子是否真和疾病有關聯)

Event D : have a disease (i.e. lung cancer)

Event E : exposure to a risk factor (i.e. smoking)

$$\text{RR} = \text{Relative risk} = \frac{\Pr(D|E)}{\Pr(D|E^C)} = \frac{a/(a+c)}{b/(b+d)}.$$

		Disease Status		
		Diseased	Non-diseased	Total
Exposure Status	Exposued	a	c	E =a+c
	Non-Exposed	b	d	E^C =b+d
	Total	D	D^C	

Remark:

If RR > 1 → "exposure" increases the risk of getting the disease

Living in a radiation house results in higher chance of getting leukemia.

表 91、以 Poisson regression 分析輻射曝露的癌症風險分析

癌症	輻射組人數	對照組人數	RR	95% CI	ERR	95% CI of ERR	p
口腔癌	25	297	0.84	0.56-1.27	-0.16	-0.44-0.27	0.4100
子宮頸癌	16	159	1.01	0.6-1.69	0.01	-0.40-0.69	0.9689
甲狀腺癌	22	149	1.48	0.95-2.32	0.48	-0.05-1.32	0.0840
白血病	**32**	**194**	**1.65**	**1.14-2.4**	0.65	0.14-1.40	**0.0083****
何杰金病	3	31	0.97	0.3-3.18	-0.03	-0.70-2.18	0.9617
卵巢癌	8	66	1.22	0.58-2.54	0.22	-0.42-1.54	0.5987
肝癌	52	515	1.01	0.76-1.34	0.01	-0.24-0.34	0.9436
乳癌	48	558	0.86	0.64-1.16	-0.14	-0.36-0.16	0.3331
肺癌	43	488	0.88	0.65-1.21	-0.12	-0.35-0.21	0.4331

- **Odds ratio (勝算比的比率)**

"odds" (勝算比): $\frac{\Pr(D)}{\Pr(D^C)}$

Odds for the exposed group : $\frac{a}{c}$;

Odds for the non-exposed (E^C) group: $\frac{b}{d}$

Odds ratio = $\frac{a/c}{b/d} = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}$

Remarks:

- If the odds ratio is close to 1, it means that the exposure has no effect.
- If the odds ratio is far from 1, it means that the exposure influences the disease status.

EXAMPLE 3.4.3. A study of age of the mother at the birth of the child as a risk factor in the development of sudden infant death syndrome (SIDS) is conducted. A total of 7330 women who were under the age of 25 when the child was born were selected for study. Of these, 29 had children afflicted with SIDS. Of the 11,256 women selected for study who were 25 or older when their children were born, 15 had children with SIDS. These data are shown in Table 3.4. From this table we see that

$$P[D|E] = \frac{29}{7330} \quad \text{and} \quad P[D|E'] = \frac{15}{11{,}256}$$

Age as a risk factor in the development of SIDS

		SIDS		
		Yes	No	
Age	Under 25 years	29	7,301	7,330 (Fixed)
	25 years or over	15	11,241	11,256 (Fixed)

Example 3.4.3: 母親年齡與嬰兒猝死症 (SIDS)

E : mother's age is under 25

E^C : mother's age is 25 or older

Relative risk $= \dfrac{\Pr(D|E)}{\Pr(D|E^C)} = \dfrac{29/7330}{15/11256} = \dfrac{29*11256}{15*7330} = 2.97$

Odds ratio: $\dfrac{29*11241}{15*7301} = 2.96$

Supplement knowledge about "data collection"

- The indices discussed above are "probabilities" which cannot be known in reality.
- By sampling, we can collect sample data and then estimate these quantities.

Sampling scheme 1: fix total

Example:

	Sleep before 12:00 am	Sleep after 12:00 am	total
Age: 20-35	a	c	
Age: 35-50	b	d	
Total			1000 = fixed

Remarks:

- Multinomial sampling
- RR and odds ratio are both estimable.

Sampling scheme 2: fix row totals

	diseased	Not diseased	total
Aspirin (E)	a	c	a+c (fixed)
Placebo (E^C)	b	d	b+d (fixed)
Total			

Remarks:

- Binomial sampling
- RR and odds ratio are both estimable.
  - RR = Relative risk $= \dfrac{a/(a+c)}{b/(b+d)} = \dfrac{\text{Risk for E}}{\text{Risk for E}^C}$
  - Odds ratio = $\dfrac{a/c}{b/d} = \dfrac{\text{odds for E}}{\text{odds for E}^C} = \dfrac{a \times d}{b \times c}$

Sampling scheme 3: fix column totals ~ **Case-control data**

	Diseased	Healthy
Smoke	a	c
Non-smoke	b	d
Total	a+b = fixed	c+d=fixed

Remarks:

- Binomial sampling
  - The diseased group is sampled from patients in the hospital
  - The healthy group is sampled from outside of the hospital
- It is not possible to estimate RR = $\dfrac{a/(a+c)}{b/(b+d)} = \dfrac{\text{Risk for E}}{\text{Risk for E}^C}$
  - Not estimable since it is not appropriate to compute a+c or b+d
- It is OK to estimate the odds ratio
  - $\dfrac{a/b}{c/d} = \dfrac{\text{odds for D}}{\text{odds for D}^C} = \dfrac{a \times d}{b \times c}$
- Applications of case-control studies
  - Rare disease
  - Save time & money

Bayes' theorem (貝氏定理)

- Partition the sample space into disjoint sets
  - $S = A_1 \cup A_2 ... \cup A_K$ & $A_j \cap A_k = \phi$ $(j \neq k)$
- Given $\Pr(A_j)$ & $\Pr(B \mid A_j)$,
  - We can derive $\Pr(A_j \mid B)$

Theorem: Given that $A_1, ... A_n$ "partition" (切割) the sample space,

$$\Pr(A_j \mid B) = \frac{\Pr(B \cap A_j)}{\text{P(B)}} = \frac{\Pr(B \cap A_j)}{\sum_{k=1}^{K} \Pr(B \cap A_k)} = \frac{\Pr(B \mid A_j)\Pr(A_j)}{\sum_{k=1}^{K} \Pr(B \mid A_k)\Pr(A_k)}$$

Note: "partition" means that $A_i \cap A_j = \phi$ for $i \neq j$, and $A_1 \cup ... A_n = S$.

Example: arthritis (關節炎) diagnostic test

A_1 = arthritis; A_1^C = no arthritis; B = test positive

- Pr(arthritis) = Pr(A_1) = 0.1;
- Pr(test+| arthritis) = $\Pr(B \mid A_1)$ = 0.85;
- Pr(test+|no arthritis) = $\Pr(B \mid A_1^C)$ = 0.04

Question: Pr(arthritis |test +) = $\Pr(A_1 \mid B)$ = sensitivity of the test

Solution:

Pr(test+) = Pr(test+ & A_1^C) + Pr(test+ & A_1)

= Pr(test+| A_1^C)Pr(A_1^C)+Pr(test+| A_1)Pr(A_1)

= 0.04*0.9+0.85*0.1

= 0.121 = Pr(B)

Pr(有關節炎 & test +) = Pr(A_1 &B) = 0.85*0.1 = 0.085

Pr(有關節炎|test +) = $\Pr(A_1 \mid B)$ = 0.085/0.121 = 0.7

Example: Bayes Theorem

A chip manufacturing plant has 3 machines producing chips.

Machine 1 produces 30% of the output and of these, 2% are defective;

Machine 2 produces 45% of the output and of these, 1% are defective;

Machine 3 produces the remaining 25% chips and of these 3% are defective.

Q:

Find the probability that a randomly selected chip produced by this plant is defective.

If a randomly selected is defective, what is the probability that it is from Machine 3?

Solution. Define events.

A= randomly selected chip is defective;

B1=chip was produced by machine 1, P(B1)=0.3;

B2=chip was produced by machine 2, P(B2)=0.45;

B3=chip was produced by machine 3, P(B3)=0.25.

P(A|B1)=0.02, P(A|B2)=0.01, P(A|B3)=0.03.

By the Total Probability Formula:

$$P(A)= P(A|B1)xP(B1) + P(A|B2)xP(B2) + P(A|B3)xP(B3)=$$

$$= 0.02 \times 0.3 + 0.01 \times 0.45 + 0.03 \times 0.25 = 0.018.$$

By Bayes formula,

$$P(B3|A) = \frac{P(A|B3)xP(B3)}{P(A|B1)P(B1) + P(A|B2)P(B2) + P(A|B3)P(B3)} =$$

$$= \frac{0.03 \times 0.25}{0.02 \times 0.3 + 0.01 \times 0.45 + 0.03 \times 0.25} = 0.42$$

John Tukey (1915~2000)

- American mathematician
- developed the FFT algorithm
- created box plot and stem-leaf plot.
- He is also credited with coining the term 'bit'.
- He also contributed to statistical practice and articulated the important distinction between exploratory data analysis and confirmatory data analysis, believing that much statistical methodology placed too great an emphasis on the latter. Though he believed in the utility of separating the two types of analysis, he pointed out that sometimes, especially in natural science, this was problematic and termed such situations uncomfortable science.
- He emphasized the importance of **having methods of statistical analysis that are robust to violations of the assumptions underlying their use**.

John Tukey - Wikipedia

John Wilder **Tukey** was an American mathematician and statistician, best known for the development of the Fast Fourier Transform (FFT) algorithm and box plot.

John Tukey and the Beginning of Interactive Graphics

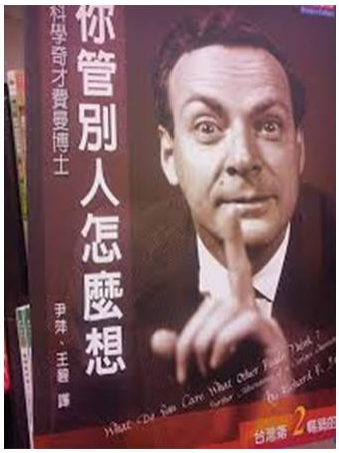
Exploratory Data Analysis / John Tukey

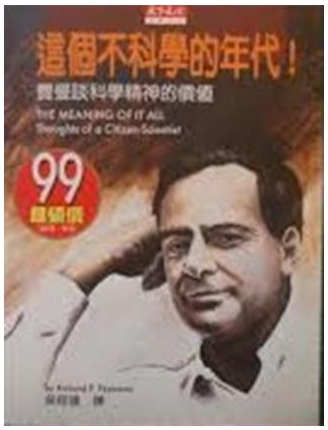
More than 30 years ago, visualization cracked its way into stat.

-

An interesting story between Richard Feynman and John Tukey

https://youtu.be/Cj4y0EUlU-Y?list=PLE80C041156D48764 → 2:30





Richard Feynman on thinking processe...
generallythinking.com

就像數一、二、三那麼簡單……

我小的時候有個朋友叫伯尼・沃克。我們倆在家裡都有「實驗室」，常常做「實驗」。有一次，那時我們大約十一二歲吧，我倆在討論什麼。我說，「思考不過就是在內心對自己講話罷了。」

「真的？」伯尼說，「你知道汽車齒輪箱裡的奇怪形狀的齒輪吧？」
「知道啊，那又怎麼樣呢？」
「好，現在告訴我，你是怎麼對自己『說』它的形狀的？」
所以我從伯尼那兒學到。思維除了語言之外，還可以是視覺的。
在上大學的時候，我開始對夢發生了興趣。在做夢的時候，眼睛是閉著的，可是一切都這麼清晰逼真，完全像是通過視網膜而成的影像。這是由於視網膜被另外的東西激發了（比方說被腦子）呢，還是在腦裡有「控制中心」，在夢中失控了呢？儘管我對腦的功能非常感興趣，可從未從心理學那兒得到任何令人滿意的答案。心理學都在做那些圓夢之類的生意。

在普林斯頓上研究所的時候，有一篇其實愚蠢無比的心理學論文引起了廣泛的討論。作者推斷大腦中控制時間感的是一個含鐵的化學反應。我對自己說，「他見鬼的怎麼得到這個結論？」

原來，他的妻子有長期的體溫不正常，有時高有時低。不知怎的，他想出來試驗她的時間感。他讓她不看鐘錶而數秒鐘，然後記下她數六十秒所花的時間。他讓她（可憐的女人！）成天從早到晚地數，發現她發燒的時候數得快，不發燒的時候數得慢。於是他推論，腦中控制時間感的機制一定是在發燒時跑得更快。

作為一個很「科學」的人，那傢伙知道化學反應的速度是隨反應能量和環境溫度而變化的。他測量了他太太讀秒的速度變化和體溫，推測出溫度和速度的相對變化，然後從化學書裡找出那些反應速度與溫度變化有近似的化學反應。他發現最接近的是含鐵的反應。於是，他就推出時間感是由一個含鐵的化學反應來決定的。

我覺得那完全是胡說八道——那長長的一連串推論中，任何一步都有無數出錯的可能。不過，他提出的問題是非常有趣的；究竟什麼來決定時間感呢？當你試圖以某一種速度來讀秒，是什麼來決定這個速度呢？你又怎麼能讓自己改變它呢？我決定來研究這個問題。我先不看鐘錶，以勻速來數一、二、三，直到六十。數完後一看鐘，花了四十八秒。不過這並不是問題，只要能以一定的勻速計數，絕對的時間是無關緊要的。我又重複了一次，這回花了四十九秒，接下來是四十八、四十七、四十八、四十九、四十八、四十八……所以看來我可以用相當準確的速度來默數。

如果我坐在那兒不默數，只是估計一分鐘的長短，結果就差得很多。因此，憑空估計一分鐘是很不準確的，有默數的幫助則會好很多。

好，現在我知道自己可以用一定的速度默數，下一個問題是哪些因素會影響它呢？

我猜想心率可能是一個因素。於是我便上上下下跑樓梯，跑得心跳極快，然後衝回房間，趴在床上默數到六十。

我還試驗了在跑樓梯的同時默數六十。

同學看見我上竄下跳，都樂了，「嘿，幹嘛呢？」我不能回答他們（這使我明白自己不能一邊說話一邊默數）。我只是埋頭起勁地跑，活像個瘋子。

（那些夥伴已經對我的瘋癲行為習以為常了。另一次，一個傢伙來我的宿舍，我正在做一個實驗忘了鎖門。他看見我穿著一件厚羊皮襖，探身到窗外的冰天雪地之中，一手托著一隻碗，另一手在不停地攪拌，還大聲嚷著，「別打擾我！別打擾我！」那次我是在做一個瓊脂實驗：我好奇如果瓊脂在不斷被攪拌時，在低溫下是否還會凝成膠凍。）

話說回來，在我試了跑上跑下和躺在床上默數之後，想不到的結果是：心率沒有影響。而且運動使我很熱，這樣看來體溫也沒什麼影響。我沒找到任何影響默數速度的因素。跑樓梯不一會就變得枯燥了，我就在做其他事的同時默數。比如，在洗衣服的時候，我會填寫有幾件襯衣，幾條褲子。我可以在「襯衣」一欄寫三，在褲子一欄寫四……可碰上襪子就糟了——襪子數目太多了。我在數三十六、三十七、三十八時，還有一大堆的三十九、四十、四十一……，這怎麼辦？

後來我發現我可以把它們分到不同的空間位置，比如一個四方形：左下角一雙，右下角一雙，這邊一雙，那邊一雙——行了，一共八雙。

同樣，我發現我可以數報紙的條數，只要把它們分成三、三、三再加一就能得十；然後三個那樣的組再加一組就可得一百。這樣，我默數到六十時可以說，「到點了，有一百一十三條。」更奇妙的是，我竟可以一邊默數，一邊閱讀文章，而默數的速度並不變化！事實上，除了說話之外，我可以一邊做任何事一邊默數。

我又試了邊打字邊默數。這回，我發現數六十需要的時間變了。我大為興奮，終於發現了一個可以改變默數速度的因素了！我繼續做實驗。

我一邊打字一邊默數，十九、二十、二十一……沒問題……二十七、二十八、二十九……沒問題——碰上一個不懂的詞，心裡會一動，「這是什麼詞」，然後明白過來，「噢，是它呀」——然後接著數三十、三十一、三十二，等到六十時，我已經遲了。

經過仔細自我觀察和琢磨，我找出真相了：我分心了。其實默數的速度並沒有變，而是在碰到難詞的時候由於注意力的轉移，默數停了一小會，而我自己一開始並沒有注意到而已。

第二天早上，我在早飯時向同桌的夥伴講了這一系列實驗。**我說，除了說話，我可以一邊默數一邊做任何事情。**

一個叫約翰·吐其的說，「我不信你可以邊閱讀邊數，也不相信你不能邊說邊數。我敢打賭，你並不能邊閱讀邊數，但你能邊說邊數！」

於是我演示了一遍。他們拿來一本書，我一邊看一邊數。到了六十我叫停——果然是四十八秒，我的老時間，然後我正確地複述出書裡講什麼。

吐其驚訝不已。我們拿他做實驗，測了他數六十的平均時間。他開始說話，「瑪麗有隻小羊羔，我愛講啥就講啥，一點問題也沒有，不知為什麼你們就不行……」他「哇啦哇啦」說個不停，最後叫道，「到點了！」我們一看，他默數的時間和平時一模一樣！我簡直不能相信！

我們討論了一會，發現了新東西。**原來吐其默數的方法和我不同，他在默數時是想像一個寫著數字的紙條在跳動，這樣他可以在嘴上唸，「瑪麗有個小羊羔。」這下可弄清楚了：因為他是在用視覺默數，所以他可以說話但不能閱讀，我正好相反，我是用聲音來默數，所以我不能同時說話。**

這個發現之後，我又嘗試能否在默數時大聲地讀書——這是我們兩人都不能做的。我想這會用著腦中既不管視覺也不管語言的區域，所以我想用手指，因為它由觸覺來控制。不一會，我成功地用手指來數，同時大聲地讀書。不過我進一步想讓一切都是意識過程，而不包含動作，所以我試著一邊唸書一邊想像手指在動著數。

我一直無法成功。或許是我的努力不夠，或許是它確實不可能。自此以後，我從來沒碰上誰能做到它。

通過那個試驗，吐其和我發現，原來像默數這麼簡單的事情，看上去似乎應該大家都一樣，其實每個人也有自己獨特的方法。而且我們發現腦功能可以用客觀、外部的方法來

檢測，比方說，不必依賴他對自己的分析和陳述，你可以觀察在默數時一個人能做什麼或不能做什麼，這樣的測試是客觀和公正的，沒法做假。

用自己已知的東西來解釋新的概念是人之常情。概念是一層一層的：這個是由那個組成，而那個又是由其他組成。因此，像默數這個概念，各人也可以不同。

我常常想起這個實驗。特別是在我教很艱深的諸如巴塞爾積分方程時，我會看見方程式的數字、符號是五彩的——我也不知為什麼。我會在腦海中看見方程就像傑克和艾曼德教科書裡的一樣，但是 J 是棕色的，N 是紫色的，X 是黑色的，到處飄浮著。我不知學生們是怎麼看它的。