Lecture 3: Probability based on Set Theory

9/30/2020

- **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)

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

Check whether two events are independent → very important

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?

$P(A_1 \cap A_2^C) = 0.02$ vs. $P(A_1)P(A_2^C) = 0.2 \times 0.3 = 0.06$ → not equal (a)

$\Pr(A_2 \mid A_1) = \dfrac{\Pr(A_1 \cap A_2)}{\Pr(A_1)} = 0.9 > \Pr(A_2) = 0.7$ (b)

(可以看出正相關，因為在 A_1 中有 A_2 的比例更高)

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

Remark: You can use any one of (a), (b) or (c) to justify your answer.

- **Useful techniques of counting**
  - $P(A) = \frac{\#\text{ in A}}{\#\text{ in S}}$
  - 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 = $2^5 = 32$

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) → $\Pr(A) = \frac{10}{32}$

Use the combination rule

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

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

Event A = 2 normal, 2 high and one low

Total number in the sample space = 3^5

$$\#\text{ in A} = \frac{5!}{2!2!1!} = \frac{5*4*3*2*1}{2*2} = 30$$ → 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” → 3 outcomes
Normal: (1,3)
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 = 10 * 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 (the 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 ← Approximation will be more accurate if there are more people participating in the experiment.

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)

T: tall (capital letter → dominant 顯性)

t: short (small letter → recessive 隱性)

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

EX: Both parents are heterozygous

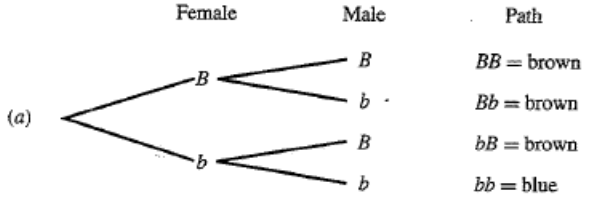
Father: Tt Mother: Tt → Offspring: 2*2 cases

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



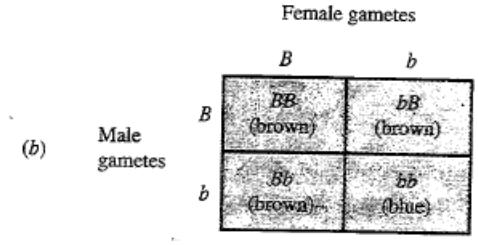


FIGURE 2.2
(*a*) Tree diagram for the inheritance of eye color from a couple, each of whom is heterozygous for eye color. (*b*) A biologist's representation of the problem as a Punnett square.

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

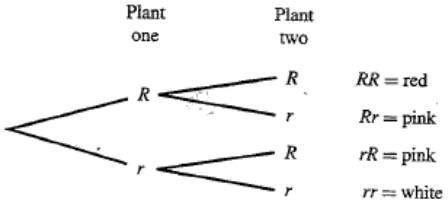


FIGURE 2.3
Outcomes that result when two heterozygous four-o'clock flowers are cross-matched.

Example 2.2.5 Two traits: skin (Albinis-白化症) & Earlobes (耳垂) → 圖更複雜

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)

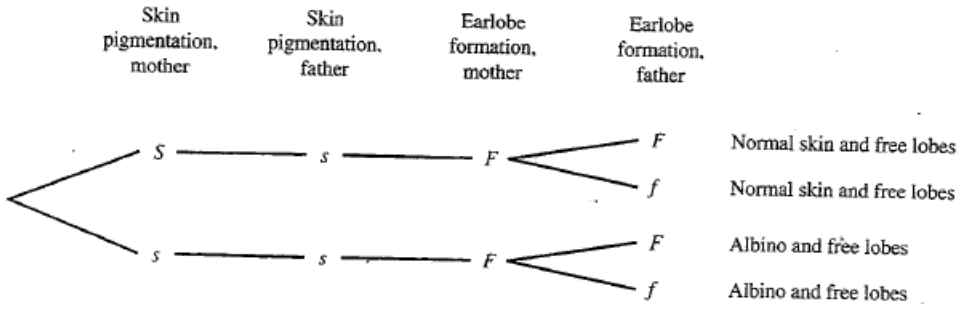


FIGURE 2.4
A four-stage tree used to study two traits simultaneously.

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

Useul rules

a. Rule of Addition – for calculating the probability of a union

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

b. Rule of multiplication – 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:

$$\Pr(\text{Ph + blood}) = 1 - \Pr(\text{Ph – blood}) = 1 - 0.1521 = 0.8479$$

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:

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:

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

Analysis 3: Father must have Rh + gene

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

Analysis 4:

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

Q: Comopute Pr(blood incompatibility)

Situation 1: **(Modified)**

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:

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)

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"

1. 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

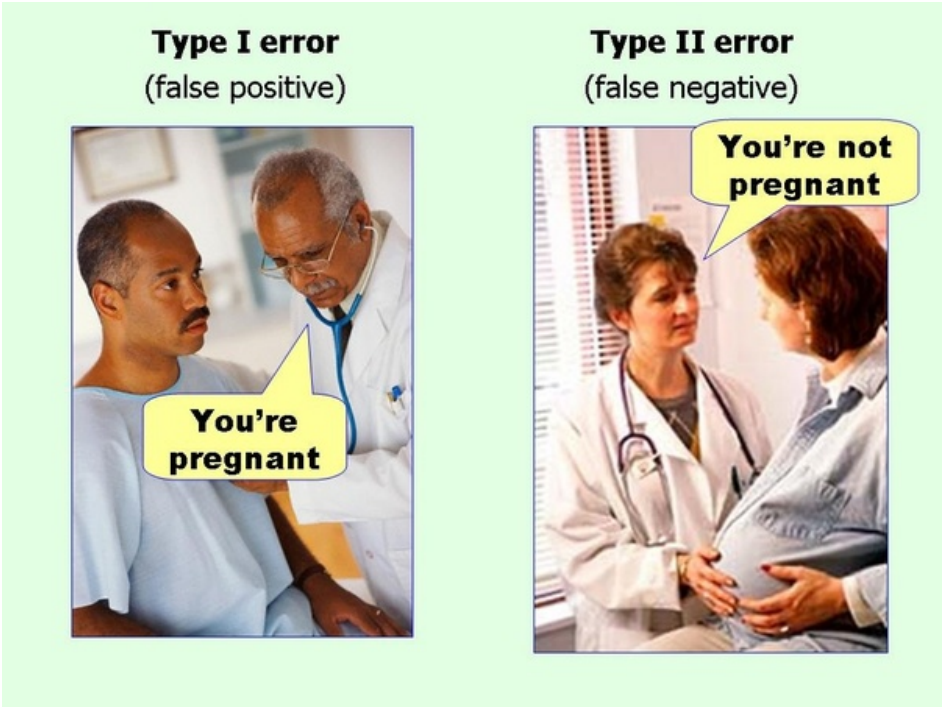
- false positive rate

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

  (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[true −] $\doteq$ 147/300 and P[test + and true −] $\doteq$ 51/300. The definition of conditional probability yields

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

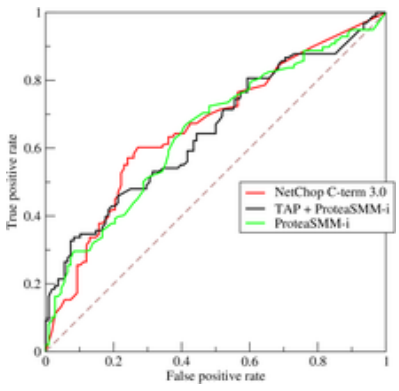
estimated false-positive rate. To estimate β, note that of the 153 true-positive subjects, 75 tested negative. Hence

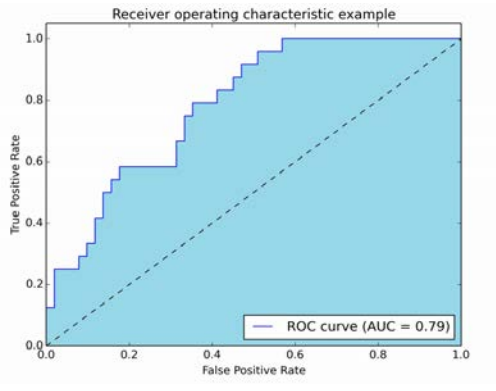
$$\beta \doteq \frac{75}{153} = .4902$$

- **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.

Given a threshhold: (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 threshhold if it produces a point near to the upper-left corner (X small,Ylarge).
- We can use AUC 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 +) = $\Pr(A_1 \mid B)$

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:

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

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

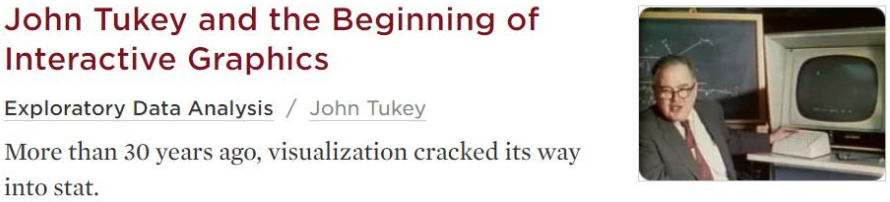
By Bayes formula,

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

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

John Tukey (1915~2000)

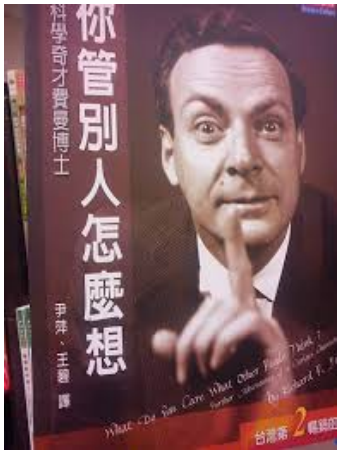
- American mathematician
- developed the FFT algorithm
- created box plot and stem-leaf plot.
- He is also credited with coining the term 'bit'.
- He also contributed to statistical practice and articulated the important distinction between exploratory data analysis and confirmatory data analysis, believing 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**.

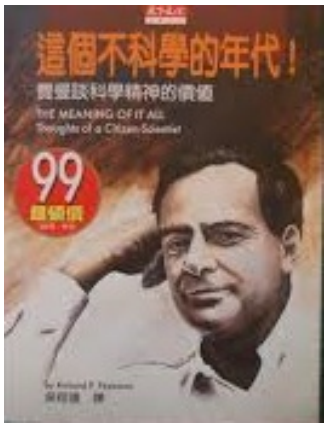


-

An interesting story between Richard Feynman and John Tukey

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





Feynman said that there are no miracle people, and anyone can do what he did if they put their mind to it (my thoughts here). Yet there's one domain in which Feynman clearly had a natural gift in — curiosity! This is exemplified by the little experiments he describes in the video below, where he learned how accurate his sense of time was and what things affected

this sense. He'd count to a minute in his head and learn that when he got to 48, a minute had passed. Then he tested what else he could do while doing this, and he could read but not talk. At the end of the video he says "Now I'm starting to talk like a psychologist, and I know nothing about that!" Let's test that theory. Here's the video.
For the lazy, when Feynman told mathematician John Tukey about this, Tukey could do the reverse — talk but not read. The reason was that Feynman would talk to himself in his head, while Tukey would see an image of a clock ticking over. Feynmann suggests this could be because people think differently, and if you're having trouble getting a point across, it might be because what your saying is more difficult to translate into the other person's favoured modality than it is your own.

I don't know if he's right about that latter point, but he's certainly right about the rest. We have multiple cognitive "modules" in the brain which are specialised to different functions, and it's possible to bring different modules to bear on a task. For example, our working memory, which is the cognitive process in use whenever you're consciously "doing" something (like Feynman's counting task) has a number of different components. I discuss these here. Each of these components has limitations, but your brain can use all the components at the same time.

When Feynman started counting in his head he was employing the phonological loop, and when counting lines in a book he's using the visuo-spatial sketch pad. These are different "modules," that's why he could do both tasks at the same time. Talking uses the phonological loop, so when he tried that, he's asking too much of the module (which in most people would be fully occupied by the counting) causing him to mess up on the task.

For Tukey, the reverse is true. He visualised a clock, occupying the visuo-spatial sketch pad but leaving the phonological loop free. So he could talk freely but as soon as he tried to read, he messed up.

Some experiments even take advantage of this fact, by having participants count out-loud as they perform some other task, so they occupy the phonological loop as they test some other cognitive module.

It's also true that different people have different preferences in terms of how the process information, and cultural differences play a big role in this. So at the end of the video, Feynman was being a little unfair on himself when he said he knew nothing about psychology!