Conditional probability
Bayes theorem

27.10.2005

GE02 day 3 part 2

Yurii Auchenko

Erasmus MC Rotterdam

Colour blindness: experiment

- Experiment: drawing a random subject from a total population of N people
- In this subject, we observe the following features
  - Sex = {M, F}
  - Colour-blindness = {D, U}

- We finally aim to predict the risk (the probability) that this random subject is colour-blind

Relations between events

- Note:
  - M and F are mutually exclusive

    P(M&F) = 0
  - D and U are mutually exclusive

    P(D&U) = 0
  - Sex and colour blindness are not:

    P(M&U) > 0

    P(M&D) > 0

    P(F&U) > 0

    P(F&D) > 0

Numbers

- Let
  - number of affected is N_D
  - number of unaffected is $N_U = N - N_D$
  - number of males is N_M
  - number of females is $N_F = N - N_M$
- We also know
  - number of affected males, $N_{D\&M}$
  - number of affected females, $N_{D\&F}$

Probabilities

- Then the probability that a random subject is colour-blind is
  - N_D/N
- But we know well that frequency of colour-blindness in males is higher then in female!
  - Or, to say it more formal, probability that a person is colour-blind, depends on sex

Using more information in risk prediction

- Our risk prediction may gain accuracy if we utilize the information on sex

- What is the probability that a random male is affected? Or, better to say, what is probability of being affected GIVEN the person is male?

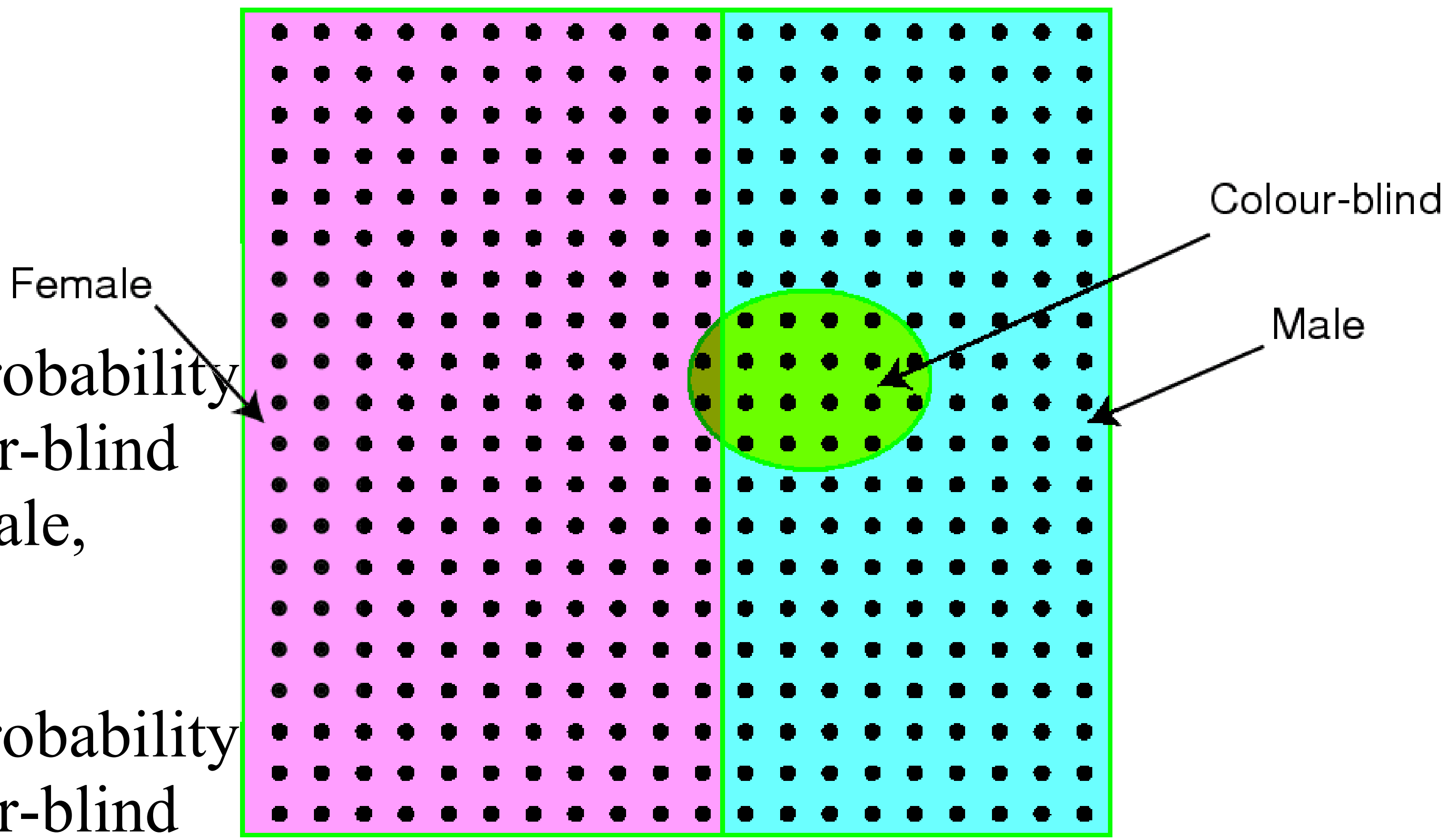
  - $P(D|M) = N_{M\&D}/N_M = P(M\&D)/P(M)$

Conditional probability

- Probability of being colour-blind given sex
  - $P(D|M)$
  - is an example of **conditional probability**
- **There are many genetic probabilities that are conditional**
  - **transmission probabilities**
  - **penetrances**
  - **...**
- **Generally, $P(A|B) = P(A\&B)/P(B)$**

Problem

- Compute
  - P(D)
  - P(D|M)
  - P(D|F)
- Compute probability that a colour-blind person is male,
  - P(M|D)
- Compute probability that a colour-blind person is female,
  - P(F|D)



Solution

- N = 400
- P(M) = 180/400 = 9/20
- P(F) = 220/400 = 11/20

- **P(D) = 20/400 = 1/20 = 5%**
- **P(D|M) = 18/180= 1/10 = 10%**
- **P(D|F) = 2/220 = 1/110 = 0.9%**
- **P(M|D) = 18/20 = P(M&D)/P(D)**
- **P(F|D) = 2/20 = P(F&D)/P(D)**

Task

- There are three bowls full of cookies. Bowl #1 has 10 chocolate chip cookies and 30 plain cookies, while bowl #2 has 20 of each.
  - What is probability to pick up a plain cookie from bowl #1?
  - … #2?
  - What is probability to pick up a a bowl at random and then cookie at random and then to discover that it is a plain one?
  - If you pick up a bowl at random and then a cookie at random and discover that it was a plain one, what is probability that you picked it up from the bowl #1?
  - … from bowl #2?

Answer

- Denote bowl as B and cookie as C
  - $P(C{=}plain|B{=}1) = N_{plain\ in\ \#1}/N_{\#1} = 30/40 = 3/4$
  - $P(C{=}plain|B{=}2) = N_{plain\ in\ \#2}/N_{\#2} = 20/40 = 1/2$
  - $P(C{=}plain) = N_{plain}/N = 50/80 = 5/8$
  - $P(B{=}1|C{=}plain) = N_{plain\ in\ \#1}/N_{plain} = 30/50 = 3/5$
  - $P(B{=}2|C{=}plain) = N_{plain\ in\ \#2}/N_{plain} = 20/50 = 2/5$

Problem

- Let in population there are 2 alleles, M and N
- Frequency of M, $P(M)=0.05$
- Penetrances (conditional probability of having disease given genotype) are
  - $P(D|MM)=1.0$
  - $P(D|MN)=0.7$
  - $P(D|NN)=0.03$
- Assuming HWE, what is the frequency of disease in the population?

Solution

- Frequency of M, P(M)=0.05. Thus, assuming HWE,
  - P(MM) = 0.0025, P(MN) = 0.095, P(NN) = 0.9025
  - Of MM, who make 0.0025 of the population, all are ill, thus, they contribute 0.0025 to the frequency of the diseas
  - Of MN, who make 9.5% of the population, 70% are ill, thus, they contribute 0.095*0.7 = 0.0665 to the frequency of the disease
  - Of NN, 3% are ill, they contribute 0.9025*0.03 = 0.0271 to the disease

Solution

- Thus, the frequency of disease is

0.0025 (these ill among MM) +
0.0665 (among MN) +
0.0271 (among NN) = 0.0961 =

9.61% of the population are ill

Formula of total probability

- We were following schema

P(M) 0,05

g	P(g)	P(D|g)	P(g)*P(D|g)
MM	0,0025	1,0000	0,0025
MN	0,0950	0,7000	0,0665
NN	0,9025	0,0300	0,0271
		P(D)=	0,0961

And the computations were done using the formula

$$P(D) = \sum_{g=MM,MN,NN} P(D \mid g)P(g) =$$

$$P(D \mid MM)P(MM) + P(D \mid DM)P(DM) + P(D \mid DD)P(DD)$$

Task

- Use the total probability formula to find out the chance to pick up a a bowl at random and then cookie at random and then to discover that it is a CHOCOLATE one

Answer

$P(C{=}choc|bowl{=}1)P(bowl{=}1)$ +

$P(C{=}choc|bowl{=}2)P(bowl{=}2)$ =

$\frac{1}{4}\,\frac{1}{2} + \frac{1}{2}\,\frac{1}{2} = 3/8$

Problem

- For the same disease and gene:
  - if we observe an ill person, what is the probability it would have genotype MM, MN or NN?
  - ...to put it formally, what are the genotypic probabilities given a person is ill, P(MM|D), P(MN|D) and P(NN|D)?
  - These are the probabilites of the genotypes in a "population" of ill people!

Solution

- Probability of disease, P(D) = 0.0961
- This probability was made of three components:
  - 0.0025 (these ill from MM) + 0.0665 (from MN) + 0.0271 (from NN) = 0.0961
- Thus, the proportion of
  - MM is 0.0025/0.0961 = 0.026 = 2.6%
  - MN is 0.0665/0.0961 = 0.6922 = 69.22%
  - NN is 0.0271/0.0961 = 0.2818 = 28.18%

Bayes' formula

- We were following the schema

- And the computations were done using the formula

$$P(g \mid D) = \frac{P(D \mid g)P(g)}{P(D)} = \frac{P(D \mid g)P(g)}{\sum_{g=MM,MD,DD} P(D \mid g)P(g)}$$

Total probability and Bayes' formulas

- Two sets of events are considered:
  - "Hypothesis" H_i for which a prioi probabilities, $P(H_i)$ are known. E.g. genotypes were "hypotheses" in our example. These hypotheses must be mutually exclusive.
  - Event(s) of interest, A, e.g. disease. For this event, conditional probabilites given hypotheses, $P(A|H_i)$

Total probability & Bayes' formulae

- Total probability (of event A)

$$P(H_i \mid A) = \sum_i P(A \mid H_i) P(H_i)$$

- Probability of hypothesis H_i, given A

$$P(H_i \mid A) = \frac{P(A \mid H_i) P(H_i)}{P(A)} = \frac{P(A \mid H_i) P(H_i)}{\sum_i P(A \mid H_i) P(H_i)}$$

Task

- You pick up a bowl at random, and then pick up a cookie at random. The cookie turns out to be a plain one.
- Use Bayes' formula to find out what is the probability that you picked the cookie out of bowl #1

Answer

- H_1 – bowl number 1
- H_2 – bowl number 1
- A – plain cookie
- $P(H_1) = P(H_2) = ½$
- $P(A| H_1) = ¾$
- $P(A| H_2) = ½$

$$P(H_1 \mid A) = \frac{P(A \mid H_1)P(H_1)}{P(A)} = \frac{P(A \mid H_1)P(H_1)}{\sum_{i=1,2} P(A \mid H_i)P(H_i)}$$

$= (¾ \; ½) / (¾ \; ½ + ½ \; ½) = (3/8) / (5/8) = 3/5$

Task

- In a population, the frequency of obese people is 25%, overweight is observed in 40% and normalweight people have frequency of 25%. The frequency of hypertension in these groups is 45, 30 and 20%, respectively
  - What is the total frequency of hypertension in the population?
  - If a random person is hypertensive, what is the best quess about his (her) weight?
  - If a random person is not hypertensive, what is the best quess about his (her) weight?

Solution

- Denote
  - H1=obese, H2=overweight and H3=normal
  - A = hypertensive, B=not hypertensive
- Probabilities
  - $P(H1)=0.25$, $P(H2)=0.4$ and $P(H3)=0.35$
  - $P(A|H1)=0.45$, $P(A|H2)=0.3$ and $P(A|H3)=0.2$
  - $P(B|H1)=1 – P(A|H1) = 0.55$, $P(B|H2)=0.7$ and $P(B|H3)=0.8$

Solution: frequency of hypertension

- Probabilities
  - P(H1)=0.25, P(H2)=0.4 and P(H3)=0.35
  - P(A|H1)=0.45, P(A|H2)=0.3 and P(A|H3)=0.2

$$P(A)=\sum_{i=1,2,3} P(A/H_i)P(H_i)$$

$$P(A/H_1)P(H_1)+P(A/H_2)P(H_2)+P(A/H_3)P(H_3)$$

$$0.25\cdot 0.45+0.4\cdot 0.3+0.35\cdot 0.2=0.3$$

Solution: weight group frequencies in hypertensive subjects

- Probabilities
  - P(H1)=0.25, P(H2)=0.4 and P(H3)=0.35
  - P(A|H1)=0.45, P(A|H2)=0.3 and P(A|H3)=0.2

$$P(H_i/A)=\frac{P(A/H_i)P(H_i)}{P(A)}$$

$$P(H_1/A)=\frac{P(A/H_1)P(H_1)}{P(A)}=\frac{0.25\cdot 0.45}{0.3}=0.37$$

$$P(H_2/A)=\frac{P(A/H_2)P(H_2)}{P(A)}=\frac{0.4\cdot 0.3}{0.3}=0.4$$

$$P(H_3/A)=\frac{P(A/H_3)P(H_3)}{P(A)}=\frac{0.35\cdot 0.2}{0.3}=0.23$$