大數據統計與預測 第六章作業

Created by 黃彥鈞(Weber Huang)

1. The frequentist definition of probability :

```
If an experiment is repeated n times under essentially identical conditions, and
if the event A occurs m times, then as n grows large, the ratio m/n approaches a
fixed limit that is the probability of A.
```

3. Mutually exclusive vs independent events :

```
If variables are  mutually exclusive, we can say the intersection is a null-
event, that is, they absolutely cannot both happen in the same time. For
example, when tossing a coin, the result can either be heads or tails but cannot
be both.

Otherwise, if variables are independent, it means that the occurrence of
variable A won't influence variable B. For example, when tossing two coins, the
result of one flip does not affect the result of the other.
```

8. Mexican-American birth research :

Probability of infant's gestational age > 37 weeks is, $P(A) = 0.142$
Probability of infant's birth weight > 2500 grams is, $P(B) = 0.051$
Both conditions occurrence probability is, $P(A \bigcap B) = 0.031$

(a) Plot a Venn's graph to illustrate the conditions :

Venss

(b) Are A and B independent?

```
If A and B are indepentent, mathematically,
```

$A \bigcap B = A \times B$
$A \times B = 0.142 \times 0.051 = 0.007242$

```
But we have known that from the data :
```

$A \bigcap B = 0.031$

```
The outcome is, A and B aren't insependent.
```

(c) For a randomly selected M-American newborn, what is the P that A or B or both occur ?

```
The answer is the union of probability of both A and B, that is,
```

$$A \bigcup B = (A + B) - A \bigcap B = (0.142 + 0.051) - 0.031 = 0.162$$

(d) What is the probability that event A occurs given that event B occurs?

```
Mathematically, we have to calculate P( A | B ), and P(B) should not be 0
```

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{0.031}{0.051} = 0.6078431$$

9. Natality Statistics of American, 1992

Age	Probability
<15	0.003
15-19	0.124
20-24	0.236
25-29	0.290
30-34	0.220
35-39	0.085
40-44	0.014
45-49	0.001
Total	1.000

Since the total probability is the sum of all probability and each row won't influence each other, so we consider those data are **independent** and **mutually exclusive** .

(a) What is the probablility of women giving birth >= 24 yrs old?

```
We have to add up first 3 values that are under 24 of the table :
```

$$Answer : 0.003 + 0.124 + 0.236 = 0.363$$

(b) What is the probability that she was 40 or older?

```
We have to add up last 2 values that are higher than 40 of the table :
```

$$Answer : 0.014 + 0.001 = 0.015$$

(c) We want to calculate P(A|B), that A is mother who was not 20 yet, B was mother under 30 yrs old.

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{(0.003+0.124)\times(0.03+0.124+0.236+0.29)}{0.03+0.124+0.236+0.29} = 0.127$$

(d) We want to calculate P(A|B), that A is mother who was under 40 , B was mother older than 35 yrs old.

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{0.085\times(0.085+0.014+0.001)}{0.085+0.014+0.001} = 0.085$$

13. A sensitive of a screening test of detecting the breast cancer is 0.85, while its specificity is 0.80.

$$Sensitive = P(T^+|D^+)$$

$$Specificity = P(T^-|D^-)$$

(a) Probability of FN?

$$FN = P(T^-|D^+) = 1 - Sensitive = 1 - P(T^+|D^+) = 1 - 0.85 = 0.15$$

(b) Probability of FP?

$$FP = P(T^+|D^-) = 1 - Specificity = 1 - P(T^-|D^-) = 1 - 0.80 = 0.20$$

(c) $P(D^+) = 0.0025$, what is the probability that the woman has cancer given that her mammogram is positive?

$$P(D^+|T^+) = \frac{P(D^+ \bigcap T^+)}{P(T^+)} = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+)+P(T^+|D^-)P(D^-)} = \frac{0.085\times0.0025}{0.085\times0.0025+0.20\times(1-0.0025)} = 0.00106403$$

15. Radionuclide ventriculograhy - diagnostic test for detecting coronary artery disease

	Disease	**Disease**	
Test	Present	Absent	Total
+	302	80	382
-	179	372	551
Total	481	452	933

(a) Sensitive? Specificity?

$$Sensitive = P(T^+|D^+) = \frac{P(T^+ \bigcap D^+)}{P(D^+)} = \frac{302}{481} = 0.6278586$$

$$Specificity = P(T^-|D^-) = \frac{P(T^- \bigcap D^-)}{P(D^-)} = \frac{372}{452} = 0.8230088$$

(b) $P(D^+) = 0.10$, calculate the one has disease given that he or she T^+

$$P(D^+|T^+) = \frac{P(D^+ \bigcap T^+)}{P(T^+)} = \frac{0.6278586\times0.10}{0.6278586\times0.10+(1-0.8230088)\times(1-0.10)} = 0.2827199$$

(c) What is the $P(T^-)$?

$$P(T^-) = \frac{551}{933} = 0.5905681$$

19. A community-base study of respiratory illness during the first year of life, north American

Socioeconomic status	Number of children	Number with symptoms
Low	79	31
Middle	122	29
High	192	27

(a) Probability suffering from the persistent respiratory illness symptoms in each socioeconomic status.

$$P(L) = \frac{31}{79} = 0.3924051$$

$$P(M) = \frac{29}{122} = 0.2377049$$

$$P(H) = \frac{27}{192} = 0.140625$$

(b) odd ratio

$$OR(L) = \frac{0.3924051}{1-0.3924051} = 0.6458334$$

$$OR(M) = \frac{0.2377049}{1-0.2377049} = 0.3118279$$

$$OR(H) = \frac{0.140625}{1-0.140625} = 0.1636364$$

(c) Did the status really matter?

```
Yeah, there is negative association between socioeconomic status and respiratory
symptoms since the higher the status is, the lower the odd ratio of the symptoms
is.
```