Rで学ぶデータ解析とシミュレーション

演習問題の解答例

熊谷　悦生

大阪大学大学院基礎工学研究科

平成 20 年 5 月 20 日

データチェックの演習

P.L.Panum(1940) の論文「1846 年のフェロー (Faroe) 諸島での麻疹流行における観測」において, 採用されていたデータを次に示します. このデータを検証して下さい.

年齢	人口	患者数	死亡者数	死亡率
<1	198	154	44	28.6 %
1–9	1440	1117	3	0.3 %
10–19	1525	1183	2	0.2 %
20–29	1470	1140	4	0.4 %
30–39	842	653	10	1.5 %
40–59	1519	1178	46	3.9 %
60–79	752	583	46	7.9 %
80+	118	92	15	16.3 %
合計	7864	6100	170	2.8 %

データチェックの演習

R でのデータ作成:

```
> faroe <- data.frame(matrix(c(198, 154, 44, 1440, 1117, 3, 1525,
+     1183, 2, 1470, 1140, 4, 842, 653, 10, 1519, 1178, 46, 752,
+     583, 46, 118, 92, 15), ncol = 3, byrow = T))
> names(faroe) <- c("population", "patients", "mortality")
> faroe
  population patients mortality
1        198      154        44
2       1440     1117         3
3       1525     1183         2
4       1470     1140         4
5        842      653        10
6       1519     1178        46
7        752      583        46
8        118       92        15
```

データチェックの演習

R での計算結果:

```
> attach(faroe)
> rate01 <- mortality/patients
> summary(rate01)

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001691 0.003303 0.027180 0.073740 0.099940 0.285700

> rate02 <- mortality/population
> summary(rate02)

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.001311 0.002562 0.021080 0.057350 0.077660 0.222200

> rate03 <- patients/population
> summary(rate03)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7753  0.7755  0.7756  0.7763  0.7762  0.7797

> detach(faroe)
```

データチェックの演習

検証結果:

- 各階級における真のデータは全人数と死者数だけ
- Panum は麻疹による一般的な罹患率 77.6 %を知っていたらしい
- 各階級の人数に 0.776 を掛けて, うその患者数を計算

$$\begin{aligned} 198 \times 0.776 &= 153.648 \approx 154, \\ &\vdots \\ 118 \times 0.776 &= 91.568 \approx 92 \end{aligned}$$

- 死者数をうその患者数で割って死亡率を計算

便潜血検査

厚生労働省研究班（主任研究者・津金昌一郎国立がんセンター予防研究部長）による大規模な疫学調査

- 1990 年, 40〜50 歳代の男女約 4 万人を対象に 2003 年までの追跡調査
- 調査時点で 17%が過去 1 年以内に便潜血検査を受診
- 2003 年までに 597 人が大腸がんになり, そのうち 132 人が死亡
- この検査を受けた人は受けなかった人に比べ, 大腸がんによる死亡率は 72%も低かった.

（疑問）検査を受けた人で死亡した人は何人でしょう？調査人数は 4 万人とします.

便潜血検査

死亡率は対象となった人の中で死亡した割合と仮定
検査を受けた人の中で大腸がんで死亡した人数 y

$$\frac{y}{40000 \times 0.17} = (1 - 0.72) \frac{132 - y}{40000 \times (1 - 0.17)}$$

これをまとめると

$$y = \frac{0.17 \times (1 - 0.72) \times 132}{1 - 0.17 \times 0.72} = 7$$

便潜血検査


```
> (kensa.p <- 40000 * 0.17)
[1] 6800
> (kensa.p.d <- round(0.17 * (1 - 0.72) * 132/(1 - 0.17 * 0.72)))
[1] 7
> (kensa.n <- 40000 * (1 - 0.17))
[1] 33200
> (kensa.n.d <- 132 - kensa.p.d)
[1] 125
> kensa.p.d/kensa.p
[1] 0.001029412
> kensa.n.d/kensa.n
[1] 0.00376506
> kensa.p.d/kensa.p/(kensa.n.d/kensa.n)
[1] 0.2734118
```

便潜血検査

得られたデータと死亡率の仮定から作られた分割表

便潜血検査	調査人数	大腸がん	
		死亡	生存
有り	6800	7	?
(割合)	(17%)	(0.1%)	
無し	33200	125	?
(割合)	(83%)	(0.38%)	
合計	40000	132	$465 = 597 - 132$

大腸がんに罹っても生存している人数の内訳は不明

検査で陽性と出た時

ある致命的な感染症にかかる確率は１万分の１です．あなたがこの感染症にかかっているかどうか検査を受けたところ結果は陽性でした．この検査の信頼性は９９%です．実際にこの感染症にかかっている確率はどの程度か計算して下さい．

この検査の９９%の信頼性は, 感染している人が陽性となる割合と感染していない人が陰性となる割合を意味しています．

検査で陽性と出た時

100 万人での分割表

	陽性 (B_1)	陰性 (B_2)	合計
感染 (A_1)	99	1	100
非感染 (A_2)	9,999	989,901	999,900
合計	10,098	989,902	1,000,000

陽性の人が感染している確率:

$$P(A_1 \mid B_1) = \frac{P(A_1 \cap B_1)}{P(B_1)} = \frac{99}{10098} \doteq 0.0098$$

感染している人が陽性となる確率:

$$P(B_1 \mid A_1) = \frac{P(A_1 \cap B_1)}{P(A_1)} = \frac{99}{100} = 0.99$$

検査で陽性と出た時

β の感染率に α の信頼性検査:

	陽性	陰性	合計
感染	$\alpha\,\beta$	$(1-\alpha)\,\beta$	β
非感染	$(1-\alpha)(1-\beta)$	$\alpha(1-\beta)$	$1-\beta$
合計	$\alpha\,\beta+(1-\alpha)(1-\beta)$	$(1-\alpha)\,\beta+\alpha\,(1-\beta)$	1

陽性のうち感染である確率:

$$p \;=\; \frac{\alpha\,\beta}{\alpha\,\beta+(1-\alpha)(1-\beta)}$$

$p = 0.5$ となる条件:

$$p \;=\; 0.5 \quad\Longleftrightarrow\quad \alpha+\beta \;=\; 1$$