Statistics II

Project 1: Cancer screening

....

Screening is a medical examination which can be easily applied to a large group of persons to detect possible diseases at an early stage. Such tests should be highly sensitive and specific, meaning that a positive test result should give high probability that the patient is actually infected and that a negative test results should yield a high probability that the patient is healthy. Screening is particularly useful as a "pre-examination" for diseases which can be cured more easily if detected at an early stage. After a positive screening test other more intensive and possibly more expensive examinations can be performed to check whether the patient is really infected.

In the following we will analyze data from an Australian screening test for intestinal cancer. In a series of six days 38000 patients were tested each day for this carcinoma. For the 3000 patients which had at least one positive test result the further examination showed that $n = 196$ of those had actually cancer. The corresponding distribution of observed patients with cancer Z_k over the number of positive tests k is shown in Table 1.

Table 1: Distribution of observed patients with cancer Z_k.

Number k of positive test	0	1	2	3	4	5	6
Frequency Z_k	?	37	22	25	29	34	49

We would now like to estimate the false-negative ratio $\gamma = P(\text{test negative}|\text{patient infected})$ and with that Z_0.

Let π be the probability for a positive test result for *one* examination of *one* patient and X_i the number of positive test results for patient i in the population of with intestinal cancer infected people. For the first part of this exercise we assume that π does not vary over the population, hence X_i is *iid* binomial distributed, $X_i \sim Bin(N, \pi)$ with $N = 6$, i.e.

$$P(X_i = k) = \binom{N}{k} \pi^k (1-\pi)^{N-k}. \tag{1}$$

(a) Since we are not able to observe $X_i = 0$, we have to use a truncated binomial distribution to describe the observed data. Use

$$P(X_i = k | X_i > 0) = \frac{P(X_i = k)}{P(X_i > 0)}, \quad k = 1, ...6 \tag{2}$$

to show that the log-likelihood function is given by

$$l(\pi) = \sum_{k=1}^{N} Z_k(k\log(\pi) + (N-k)\log(1-\pi)) - n\log(1-(1-\pi)^N). \tag{3}$$

(b) Define a function in `R` which evaluates $l(\pi)$. Use `optim` to maximize this function over π, yielding the maximum likelihood estimator (MLE) $\hat{\pi}$ (use `?optim` for help). From the invariance[1] of the MLE and (2) one can now evaluate the MLE of the false-negative rate $\gamma = P(X_i = 0) = (1-\pi)^N$ as

$$\hat{\gamma} = (1-\hat{\pi})^N. \tag{4}$$

Further, use the relation $\hat{\gamma} = \hat{Z}_0/(\hat{Z}_0 + 196)$ to obtain an estimator for $\hat{Z}_0$. Do you think the result is plausible? What values for the Z_k's would you expect from the estimated value $\hat{\pi}$?

You should have seen that the binomial distribution does not fit the data very well. We therefore drop the assumption that π cannot vary over the population. Instead we assume in the following that π is beta distributed, $\pi \sim Be(\alpha, \beta)$, and that $X_i|\pi \sim Bin(N, \pi)$. Hence, X_i is *iid* beta-binomial distributed $X_i \sim BeB(N, \alpha, \beta)$, i.e.

$$P(X_i = k) = \binom{N}{k} \frac{B(\alpha + k, \beta + N - k)}{B(\alpha, \beta)}, \tag{5}$$

where

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \tag{6}$$

is the so-called beta function (the `R` command is `beta(a,b)`). Again we have to truncate this distribution.

(c) Use (3) to show that the log-likelihood function is given by

$$l(\alpha, \beta) = \sum_{k=1}^{N} Z_k \log\left(\frac{B(\alpha + k, \beta + N - k)}{B(\alpha, \beta)}\right) - n\log\left(1 - \frac{B(\alpha, \beta + N)}{B(\alpha, \beta)}\right). \tag{7}$$

(d) Analogous to part (b) define a `R` function which evaluates $l(\alpha, \beta)$. Find numerically the MLE for $(\hat{\alpha}, \hat{\beta})$. Use (5) to find the relation between $\gamma = P(X_i = 0)$ and α and β. Use this to determine $\hat{\gamma}$ and $\hat{Z}_0$. Compare the results to those of part (b).

[1] Let $\hat{\theta}$ be the MLE of θ and $\phi = h(\theta)$ be a unique transformation, then we have $\hat{\phi} = h(\hat{\theta})$.