Statistics 2: Computer Practical 1

1 Ignaz Semmelweis

Ignaz Semmelweis was a Hungarian physician who collected and published data on deaths after childbirth in Vienna in the 1840s. There were two clinics at the general hospital, and these clinics had very different mortality rates. Semmelweis sought to understand the cause of this difference and propose an intervention, chlorine hand-washing, to reduce mortality rates.

In this practical, we will analyze some of the data he collected. It was very unclear scientifically at that time why hand-washing would have any health benefit; one of the strengths of Statistics is to demonstrate relationships between variables that are difficult to understand from a purely theoretical perspective. Unfortunately, Statistics and the mathematics of random variation were also not understood at that time either, and the reaction of his contemporaries was not positive. See also: Semmelweis reflex.

2 Binomial maximum likelihood estimators

Consider the model $Y \sim \text{Binomial}(n, p)$. The maximum likelihood estimator of p is $\hat{p}(Y) = Y/n$, and this estimator is unbiased.

Question 1. [1 mark] Derive the variance of $\hat{p}(Y)$.

Solution We use the fact that $Y = \sum_{i=1}^{n} B_i$ where $B_1, \ldots, B_n$ are independent Bernoulli(p) random variables. We find

$$\text{var}(Y; p) = \text{var}\left(\sum_{i=1}^{n} B_i; p\right) = n\text{var}(B_1; p) = np(1-p).$$

It follows that

$$\text{var}(\hat{p}(Y); p) = \text{var}\left(\frac{Y}{n}; p\right) = \frac{1}{n^2}\text{var}(Y; p) = \frac{p(1-p)}{n}.$$

Question 2. [1 mark] Perform repeated experiments to verify empirically that the variance you have derived in Question 1 is correct for $(n, p) = (13, 0.31)$.

Solution

```r
n <- 13
p <- 0.31
estimates <- rbinom(1000000, size = n, prob = p)/n
var(estimates)
```

```
## [1] 0.01648933
```

```r
p*(1-p)/n
```

```
## [1] 0.01645385
```

The theoretical value and the sample variance of 1000000 realizations are very similar, so it seems we have not made an obvious mistake.

3 Clinic data

We now consider the yearly mortality data for the two Viennese clinics.

```r
year.data <- read.csv("year_data.csv")
knitr::kable(year.data) # or just year.data, for a less pretty display
```

year	births	deaths	clinic
1841	3036	237	1
1842	3287	518	1
1843	3060	274	1
1844	3157	260	1
1845	3492	241	1
1846	4010	459	1
1841	2442	86	2
1842	2659	202	2
1843	2739	164	2
1844	2956	68	2
1845	3241	66	2
1846	3754	105	2

We will model the total number of deaths in clinic i as a Binomial(n_i, p_i) random variable Y_i where n_i is the total number of births in clinic i and p_i is the mortality rate for clinic i. We shall assume that Y_1 and Y_2 are independent random variables.

To ease calculations, it is useful to calculate n_i and the realization y_i for $i \in \{1, 2\}$.

```r
n1 <- sum(year.data[year.data$clinic==1,]$births) # number of births in clinic 1
y1 <- sum(year.data[year.data$clinic==1,]$deaths) # number of deaths in clinic 1
c(n1, y1)
```

```
## [1] 20042  1989
```

```r
n2 <- sum(year.data[year.data$clinic==2,]$births) # number of births in clinic 2
y2 <- sum(year.data[year.data$clinic==2,]$deaths) # number of deaths in clinic 2
c(n2, y2)
```

```
## [1] 17791   691
```

We see that there were 20042 births and 1989 deaths in Clinic 1, while there were 17791 births and 691 deaths in Clinic 2. These numbers are indeed quite different.

Question 3. [1 mark] Compute and report the maximum likelihood (ML) estimates $\hat{p}_1(y_1)$ and $\hat{p}_2(y_2)$ of the parameters p_1 and p_2, using the data.

Solution

```r
p1.hat <- y1/n1
p1.hat
```

```
## [1] 0.09924159
```

```r
p2.hat <- y2/n2
p2.hat
```

```
## [1] 0.03883986
```

One of the key questions when facing discrepancies such as this, is whether or not the different observed rates can be attributed to random variation rather than systematic differences.

Question 4. [1 mark] Under the assumption that $p_1 = p_2 = p$, show that

$$W := \hat{p}_1(Y_1) - \hat{p}_2(Y_2),$$

has mean 0 and find its variance.

Solution We observe that if $p_1 = p_2 = p$ then

$$\mathbb{E}[\hat{p}_1(Y_1); p] = \mathbb{E}[\hat{p}_2(Y_2); p] = p,$$

and so

$$\mathbb{E}[W; p] = \mathbb{E}[\hat{p}_1(Y_1) - \hat{p}_2(Y_2); p] = \mathbb{E}[\hat{p}_1(Y_1); p] - \mathbb{E}[\hat{p}_2(Y_2); p] = p - p = 0.$$

To find the variance, we notice that if U and V are independent then

$$\mathrm{var}(U - V) = \mathrm{var}(U) + \mathrm{var}(V),$$

so in particular

$$\begin{aligned}\mathrm{var}(W; p) &= \mathrm{var}(\hat{p}_1(Y_1); p) + \mathrm{var}(\hat{p}_2(Y_2); p) \\ &= \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\end{aligned}$$

Question 5. [2 marks] Use Chebyshev's inequality to show that it is unlikely to have seen such a large difference between the observed mortality rates for the two clinics under the assumption that the true underlying mortality rates are actually the same.

Solution

We can apply Chebyshev's inequality here to give

$$\mathbb{P}(|W| > k; p) \leq \frac{\mathrm{var}(W; p)}{k^2} = \frac{p(1-p)}{k^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

We can bound $p(1-p)$ uniformly by $1/4$ since the maximizer of $p \mapsto p(1-p)$ is $p = 1/2$. Then we can plug in the observed value $w = \hat{p}_1 - \hat{p}_2 > 0$ for k, to give

$$\mathbb{P}(|W| > w; p) \leq \frac{1}{4w^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

```r
w <- p1.hat - p2.hat
bound <- 1/(4*w^2)*(1/n1 + 1/n2)
bound
```

```
## [1] 0.007270606
```

The probability of seeing such a large difference is therefore less than 0.007271.

4 Intervention: chlorine hand washing

The main difference Semmelweis identified between the two clinics was that people at the first clinic also performed autopsies. He decided to introduce systematic chlorine hand-washing between autopsy work and the examination of patients, suspecting some unknown "cadaverous material" was responsible for the increased mortality rate in the first clinic.

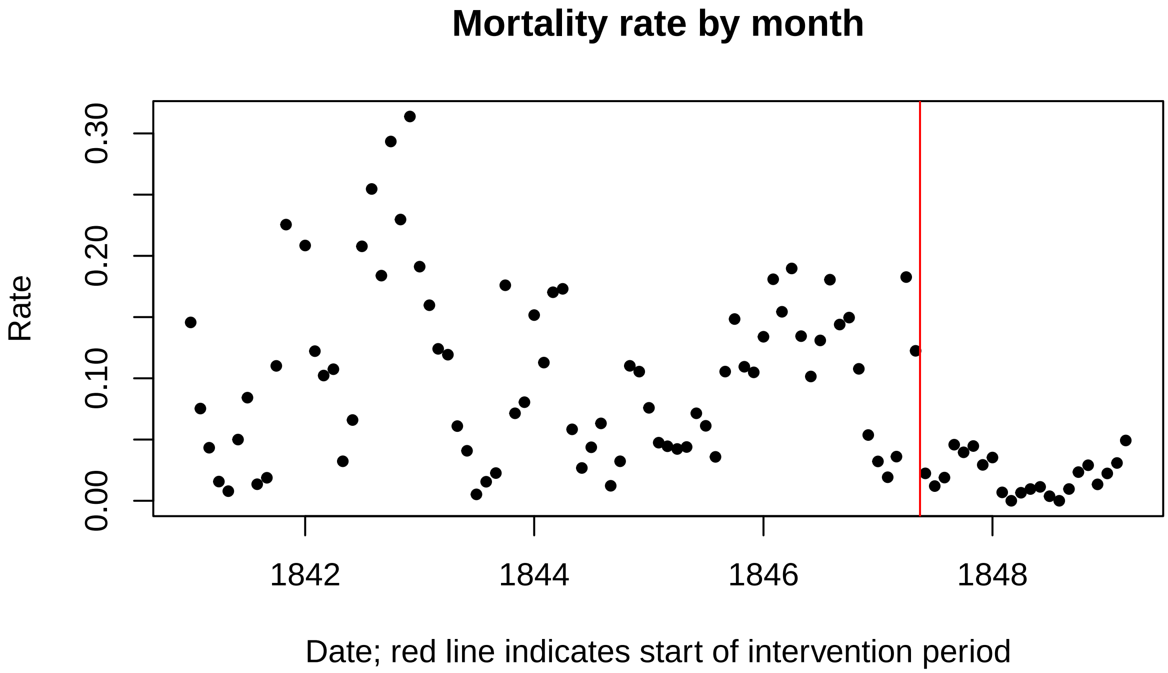
We consider now aggregated monthly data before and after the intervention. We model the total number of deaths before the intervention as a $\mathrm{Binomial}(n_1, p_1)$ random variable Y_1 where n_1 is the total number of births before the intervention and p_1 is the mortality rate before the intervention. Similarly, we model the total number of deaths after the intervention as a $\mathrm{Binomial}(n_2, p_2)$ random variable Y_2 where n_2 is the total number of births after the intervention and p_2 is the mortality rate after the intervention. As before, we shall assume that Y_1 and Y_2 are independent random variables.

```
month.data <- read.csv("month_data.csv")
month.data <- month.data[!is.na(month.data$births),]

month.data$rate <- month.data$deaths/month.data$births
month.data$date <- as.Date(month.data$date)

intervention.date <- as.Date("1847-05-15")

plot(month.data$date, month.data$rate, pch=20, main="Mortality rate by month",
     xlab="Date; red line indicates start of intervention period", ylab="Rate")
abline(v=intervention.date, col="red")
```



```
before.intervention <- month.data[month.data$date < intervention.date,]
after.intervention <- month.data[month.data$date > intervention.date,]

n1 <- sum(before.intervention$births)
y1 <- sum(before.intervention$deaths)
n2 <- sum(after.intervention$births)
y2 <- sum(after.intervention$deaths)
```

Question 6. [2 marks] Compute and report the ML estimates for the mortality rates before and after the intervention. As in Question 5, show that it is unlikely to see such a large difference between the observed mortality rates before and after the intervention if the true underlying mortality rates are the same.

Solution The same reasoning as in Q5 applies.

```
p1.hat <- y1 / n1
p2.hat <- y2 / n2
c(p1.hat, p2.hat)
```

```
## [1] 0.10525778 0.02153146
```

To obtain the bound, we can use the same calculation as before

```
w <- p1.hat - p2.hat
bound <- 1/(4*w^2)*(1/n2+1/n1)
bound
```

```
## [1] 0.007229788
```

5 A first logistic regression

We consider again the aggregated monthly data before and after the chlorine hand-washing intervention.

Modelling binary outcomes or responses is a fairly common task in data analysis. Typically one is interested in a number of different explanatory variables. Here we have considered only one: whether or not chlorine hand-washing was standard practice at the time.

A straightforward way to define the probability associated with binary outcomes, in the presence of explanatory variables, is to model

$$Y_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i, g(x_i, \theta)),$$

where $g : \mathbb{R}^d \times \Theta \to [0, 1]$ is a function of the explanatory variables ($x_i \in \mathbb{R}^d$) and the statistical parameter (θ). That is, the Y_i are independent Binomial random variables and the probability associated with mortality depends on x_i and θ. Notice that the data for such a model will consist of triples (y_i, x_i, n_i).

This is a very general class of models. In order to define a concrete model, one must specify the specific function g. In logistic regression, one defines

$$g(x_i, \theta) = \sigma(\theta^T x_i) = \sigma\left(\sum_{j=1}^{d} \theta_j x_{ij}\right),$$

where $\sigma : \mathbb{R} \to (0, 1)$ is the *standard logistic function*

$$\sigma(z) = \frac{1}{1 + \exp(-z)},$$

which is strictly increasing. In the logistic regression model $\Theta = \mathbb{R}^d$, i.e. θ has the same dimension as the x_i.

For Semmelweis' hand-washing data, we will take $d = 2$ and assign values to x_1 and x_2 that make interpreting θ easier. We define $x_1 = (1, 0)^T$ so that $g(x_1, \theta) = \sigma(\theta_1)$, and then $x_2 = (1, 1)^T$ so that $g(x_2, \theta) = \sigma(\theta_1 + \theta_2)$. This makes it easy to interpret the sign of the true values of θ_1 and θ_2. In particular, if $\theta_1 < 0$ (resp. $\theta_1 > 0$) then the probability of mortality before the intervention is less than half (resp. greater than half). Similarly, if $\theta_2 < 0$ (resp. $\theta_2 > 0$) then the probability of mortality decreases (resp. increases) after the intervention.

The following code maps x_1 and x_2 to the R variables `x1` and `x2`, and σ to the R function `sigma`.

```
x1 <- c(1,0)
x2 <- c(1,1)

sigma <- function(z) {
  1/(1+exp(-z))
}
```

Question 7. [2 marks] Write a function `ell` that corresponds to the log-likelihood function for θ, and optimize the function using the `optim` function. Explain the relationship between the maximum likelihood estimates $(\hat{p}_1, \hat{p}_2)$ and $\hat{\theta}$.

Solution

```r
ell <- function(theta) {
  dbinom(y1, n1, sigma(sum(theta*x1)), log=TRUE) +
    dbinom(y2, n2, sigma(sum(theta*x2)), log=TRUE)
}
optim.out <- optim(c(0,0), ell, control=list(fnscale=-1, reltol=1e-10))
optim.out$par
```

```
## [1] -2.140145 -1.676325
```

An alternative is

```r
ell <- function(theta) {
  y1*log(sigma(sum(theta*x1))) + (n1-y1)*log(1-sigma(sum(theta*x1))) +
    y2*log(sigma(sum(theta*x2))) + (n2-y2)*log(1-sigma(sum(theta*x2)))
}
optim.out <- optim(c(0,0), ell, control=list(fnscale=-1, reltol=1e-10))
optim.out$par
```

```
## [1] -2.140145 -1.676325
```

We see that the corresponding estimates of the before and after intervention mortality rates are almost the same. In fact, mathematically they are the same, since (p_1, p_2) is a bijective reparameterization of θ.

```r
c(sigma(optim.out$par[1]), sigma(sum(optim.out$par)))
```

```
## [1] 0.10525572 0.02153152
```

```r
c(p1.hat, p2.hat)
```

```
## [1] 0.10525778 0.02153146
```

6 Epilogue

Analyzing this data using this model does suggest a statistical relationship between chlorine hand-washing and reduced mortality rates. In general, it is difficult to give causal interpretations to statistical relationships, but in this case one may have more reason to do so as the explanatory variable of interest is controlled directly (as opposed to being passively observed). Of course, one may still be concerned that the intervention is not responsible for the drop in mortality and some other change occurred at roughly the same time in the hospital.

The bounds obtained in Questions 5 and 6 are very conservative, due to the use of Chebyshev's inequality. In fact, it is extremely unlikely to see such a large difference in observed mortality rates if the the model is correct and the true mortality rates are the same.

In many challenging scenarios with high-dimensional explanatory variables x it is increasingly common to use similar models but with more complicated functions g and very high-dimensional statistical parameters θ. For example, this is a starting point for the development of artificial neural networks and deep learning algorithms, which you may have heard about.