Statistics 2: Computer Practical 2

1 Independent but not identically distributed random variables

Statistics 2 mainly focuses on the scenario where the data is modelled as a vector of observations of i.i.d. random variables, as this scenario is sufficiently rich mathematically to capture many fundamental statistical ideas.

In practice, Statistics is often used to understand statistical relationships within data. For example the relationship between some recorded variables and a response

We consider in this practical a specific variant of this where each response is modelled as a Bernoulli random variable with a mean that depends on the other recorded variables in a simple way. The Bernoulli random variables in the model are independent but not identically distributed. This is an extension of what was considered in CP1, where the only recorded variable was whether the clinic had instituted the policy of chlorine hand-washing.

2 Statistical analysis of EET one year after FEP

We consider the data, but not the analysis, of

Leighton SP, Krishnadas R, Chung K, Blair A, Brown S, Clark S, et al. (2019) Predicting one-year outcome in first episode psychosis using machine learning. PLoS ONE 14(3): e0212846.

Specifically, we are interested in understanding how certain factors measured at initial presentation statistically affect whether a person is in employment, education or training (EET) one year after first episode pyschosis (FEP). Motivation given by Leighton et al. (2019) includes

- Evidence suggests that finding employment is more important than any specific mental health intervention.
- If we can correctly identify those with poor EET outcomes at their initial presentation, we could apply . . . vocational interventions at an earlier stage.

If you are interested in further context, you can read the paper.

The data provided by the authors of the paper is richer than what we consider here. Specifically, we will consider the records of 130 people with the following measurements:

1. `M0_Emp`: Whether the person is in EET at initial presentation.
2. `Female`: Whether the person is female or male (the only recorded values).
3. `Parent`: Whether the person is a parent.
4. `Age`: Age in years.
5. `PANSS_G`: Total score on the PANSS General Psychopathology scale.
6. `PANSS_P`: Total score on the PANSS Positive scale.
7. `PANSS_N`: Total score on the PANSS Negative scale.
8. `Y1_Emp`: Whether the person is in EET after one year.

All binary variables have 0 corresponding to “No” and 1 corresponding to “Yes”.

We will be interested in trying to understand (statistically) how the first 7 variables affects the 8th variable.

You can read about the Positive and Negative Syndrome Scale if you are interested.

3 Logistic regression maximum likelihood estimators

Consider the model

$$Y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\sigma(\theta^T x_i)), \qquad i \in \{1, \dots, n\},$$

where $x_1, \dots, x_n$ are d-dimensional real vectors of explanatory variables, and σ is the standard logistic function

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

The data consists of observed realizations of $(Y_1, \dots, Y_n)$, $\mathbf{y} = (y_1, \dots, y_n)$, as well as $(x_1, \dots, x_n)$. We define the $n \times d$ matrix $X = (x_{ij})$.

You should be clear that the x_i are *not* modelled as observations of random variables. They are deterministic quantities that are observed. Only $Y_1, \dots, Y_n$ are random variables.

Question 1. [1 mark] Show that the log-likelihood function is

$$\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)).$$

Solution The likelihood function is

$$L(\theta; \mathbf{y}) = \prod_{i=1}^{n} \sigma(\theta^T x_i)^{y_i} [1 - \sigma(\theta^T x_i)]^{(1-y_i)}.$$

Hence the log-likelihood function is

$$\ell(\theta; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\sigma(\theta^T x_i)) + (1 - y_i) \log(1 - \sigma(\theta^T x_i)).$$

Question 2. [1 mark] Show that each component of the score is

$$\frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n} [y_i - \sigma(\theta^T x_i)] x_{ij}, \qquad j \in \{1, \dots, d\}. \tag{1}$$

I suggest that you use the fact that $\sigma'(z) = \sigma(z)[1 - \sigma(z)]$.

Solution First we consider the partial derivative w.r.t. θ_j of the ith term in the log-likelihood. That is,

$$\begin{aligned}
\frac{\partial \ell(\theta; y_i)}{\partial \theta_j} &= y_i \frac{\partial}{\partial \theta_j} \log(\sigma(\theta^T x_i)) + (1 - y_i) \frac{\partial}{\partial \theta_j} \log(1 - \sigma(\theta^T x_i)) \\
&= \left\{ y_i \frac{\sigma'(\theta^T x_i)}{\sigma(\theta^T x_i)} - (1 - y_i) \frac{\sigma'(\theta^T x_i)}{1 - \sigma(\theta^T x_i)} \right\} \frac{\partial}{\partial \theta_j} \theta^T x_i \\
&= \left\{ y_i \frac{\sigma'(\theta^T x_i)}{\sigma(\theta^T x_i)} - (1 - y_i) \frac{\sigma'(\theta^T x_i)}{1 - \sigma(\theta^T x_i)} \right\} x_{ij} \\
&= y_i [1 - \sigma(\theta^T x_i)] x_{ij} - (1 - y_i) \sigma(\theta^T x_i) x_{ij} \\
&= y_i x_{ij} - \sigma(\theta^T x_i) x_{ij} \\
&= [y_i - \sigma(\theta^T x_i)] x_{ij}
\end{aligned}$$

It follows then that

$$\frac{\partial \ell(\theta; \mathbf{y})}{\partial \theta_j} = \sum_{i=1}^{n} [y_i - \sigma(\theta^T x_i)] x_{ij}.$$

In fact, Equation 1 implies that the score can be written as

$$\nabla \ell(\theta; \mathbf{y}) = X^T[\mathbf{y} - \mathbf{p}(\theta)],$$

where $\mathbf{p}(\theta)$ is the vector $(p_1(\theta), \dots, p_n(\theta)$ where $p_i(\theta) = \sigma(\theta^T x_i)$.

Question 3. [1 mark] Show that the Hessian matrix entries are

$$\frac{\partial^2 \ell(\theta; \mathbf{y})}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{n} \sigma(\theta^T x_i)[1 - \sigma(\theta^T x_i)] x_{ij} x_{ik}, \qquad j, k \in \{1, \dots, d\}. \tag{2}$$

Solution

For arbitrary $j, k \in \{1, \dots, d\}$, we have

$$\begin{aligned}
\frac{\partial^2 \ell(\theta; \mathbf{y})}{\partial \theta_j \partial \theta_k} &= \frac{\partial}{\partial \theta_k} \sum_{i=1}^{n} [y_i - \sigma(\theta^T x_i)] x_{ij} \\
&= \sum_{i=1}^{n} -\sigma'(\theta^T x_i) x_{ij} x_{ik} \\
&= -\sum_{i=1}^{n} \sigma(\theta^T x_i)[1 - \sigma(\theta^T x_i)] x_{ij} x_{ik}.
\end{aligned}$$

In fact, Equation 2 implies that the Hessian can be written as

$$\nabla^2 \ell(\theta; \mathbf{y}) = -X^T D(\theta) X,$$

where $D(\theta)$ is the $n \times n$ diagonal matrix where the ith diagonal entry is $p_i(\theta)[1 - p_i(\theta)]$ for $i \in \{1, \dots, d\}$.

Notice that for this particular model, the Hessian $\nabla^2 \ell(\theta; \mathbf{y})$ does not depend on $\mathbf{y}$.

It is not difficult to see that the Hessian is negative definite for every $\theta \in \mathbb{R}^d$. Indeed, let $z \in \mathbb{R}^d$ such that $z \neq 0$. Then,

$$z^T \nabla^2 \ell(\theta; \mathbf{y}) z = -(Xz)^T D(\theta)(Xz) < 0,$$

since $D(\theta)$ is a diagonal matrix with positive diagonal entries and is therefore positive definite.

The following function return the log-likelihood, score vector and Hessian matrix associated with a given θ, X and $\mathbf{y}$. You do not need to understand *how* exactly they compute these quantities in order to complete the practical, but you should understand *what* they compute.

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

ell <- function(theta, X, y) {
  p <- as.vector(sigma(X%*%theta))
  sum(y*log(p) + (1-y)*log(1-p))
}

score <- function(theta, X, y) {
  p <- as.vector(sigma(X%*%theta))
  as.vector(t(X)%*%(y-p))
}
```

```
hessian <- function(theta, X) {
  p <- as.vector(sigma(X%*%theta))
  -t(X)%*%((p*(1-p))*X)
}
```

4 FEP-EET data maximum likelihood estimate

First we load the FEP-EET data, and inspect the first few rows of data.

```
fep.eet <- read.csv("FEP_EET.csv")
head(fep.eet)
```

```
##   M0_Emp Female Parent Age PANSS_G PANSS_P PANSS_N Y1_Emp
## 1      0      1      0  18      39      10      19      0
## 2      1      1      0  18      29      12      12      1
## 3      0      0      0  18      39      15      16      0
## 4      0      1      0  17      51      29      26      0
## 5      1      1      0  16      45      10      25      1
## 6      1      1      0  18      35      24      10      1
```

In order to calculate the ML estimate, we should extract the matrix X of explanatory variables and the vector $\mathbf{y}$ of responses.

```
X.raw <- as.matrix(fep.eet[,1:7])
y <- fep.eet$Y1_Emp
```

It is useful to add a column of 1s to X, so that there is an "intercept" term in the model. This is what we did in CP1 with the mortality data, so that we could distinguish between a "baseline" probability and the impact of chlorine hand-washing. Mathematically, the value of θ_1 determines the probability when the explanatory variables are all 0.[1]

```
X <- cbind(1, X.raw)
head(X)
```

```
##        M0_Emp Female Parent Age PANSS_G PANSS_P PANSS_N
## [1,] 1      0      1      0  18      39      10      19
## [2,] 1      1      1      0  18      29      12      12
## [3,] 1      0      0      0  18      39      15      16
## [4,] 1      0      1      0  17      51      29      26
## [5,] 1      1      1      0  16      45      10      25
## [6,] 1      1      1      0  18      35      24      10
```

```
d <- 8
```

We are now in a position to compute the maximum likelihood estimate. The following code computes the ML estimate using R's general-purpose `optim` function. Unlike some simpler numerical optimization problems, this is already challenging enough that one must choose the options somewhat carefully.

```
maximize.ell <- function(ell, score, X, y, theta0) {
  optim.out <- optim(theta0, fn=ell, gr=score, X=X, y=y, method="BFGS",
                     control=list(fnscale=-1, maxit=1000, reltol=1e-16))
  optim.out$par
}
```

[1] Even if this is impossible: in this dataset the youngest Age is 15 and the minimum possible score for the PANSS scale overall is 30.

```r
mle <- maximize.ell(ell, score, X, y, rep(0,d))
mle
```

```
## [1]  0.34557722  3.53898725 -0.56967106  0.60441618 -0.03680464  0.02056926
## [7] -0.08936738 -0.04371038
```

It appears that being in EET at month 0 greatly improves the probability of being in EET after 1 year. Being female appears to lower the probability, while being a parent increases it. Age seems to have a small negative effect. With respect to the PANSS scores, higher general scores appear to improve the probability, while higher positive and negative scores seem to decrease it. Naturally, we don't necessarily have much confidence in any of these statements: the true parameter values could be 0 or even have different signs to the estimated values.

Although the parameter values for age and the PANSS scores are quite small, it's worth bearing in mind that these parameters multiply larger integers, e.g. ages are between 15 and 45 in the data we are analyzing.

5 Confidence intervals for each parameter

So far all we have done is find the maximizer of the log-likelihood function, i.e. the ML estimate of θ. What you will have to do now, is produce observed "Wald"-type confidence intervals for each of the components of θ.

In lectures we have seen that for regular statistical models with a one-dimensional parameter θ, the ML estimator $\hat{\theta}_n$ is *asymptotically normal* with

$$I_n(\hat{\theta}_n)^{1/2}(\hat{\theta}_n - \theta) = \sqrt{nI(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \to_{\mathcal{D}(\cdot;\theta)} Z \sim N(0,1).$$

This convergence in distribution justifies the construction of Wald confidence intervals for θ.

In this computer practical, the statistical model has d-dimensional parameters and the observed random variables are independent but not identically distributed. Nevertheless, for this model under some appropriate regularity assumptions on $x_1, x_2, \ldots$, the ML estimator $\hat{\theta}_n$ is *asymptotically (multivariate) normal* in the sense that $I_n(\theta)^{1/2}(\hat{\theta}_n - \theta)$ converges in distribution to a vector of d independent standard normal random variables, where $I_n(\theta)$ is the Fisher information *matrix*

$$I_n(\theta) = -\mathbb{E}[\nabla^2 \ell(\theta; Y_1, \ldots, Y_n); \theta].$$

In our case, the Fisher information matrix is precisely the negative Hessian of the log-likelihood, because the Hessian of the log-likelihood does not depend on $\mathbf{y}$.

One can deduce from this multivariate asymptotic normality that for $j \in \{1, \ldots, d\}$,

$$\frac{\hat{\theta}_{n,j} - \theta_j}{\sqrt{(I_n(\theta)^{-1})_{jj}}} \to_{\mathcal{D}(\cdot;\theta)} Z \sim N(0,1),$$

where $\hat{\theta}_{n,j}$ denotes the jth component of $\hat{\theta}_n$ and θ_j denotes the jth component of θ.

Notice that $(I_n(\theta)^{-1})_{jj}$ is the jth diagonal entry of the inverse of the Fisher information matrix, and is not in general equal to $(I_n(\theta)_{jj})^{-1}$, the inverse of the jth diagonal entry of the Fisher information matrix[2]. In `R` you can compute numerically the inverse of a matrix using the `solve` command.

As in the one-dimensional parameter case, when the function $\theta \mapsto (I(\theta)^{-1})_{jj}$ is continuous, one can replace $(I_n(\theta)^{-1})_{jj}$ with $(I_n(\hat{\theta}_n)^{-1})_{jj}$ to obtain

$$\frac{\hat{\theta}_{n,j} - \theta_j}{\sqrt{(I_n(\hat{\theta}_n)^{-1})_{jj}}} \to_{\mathcal{D}(\cdot;\theta)} Z \sim N(0,1).$$

[2] This is statistically interesting, as it captures the fact that our estimators are less precise in the presence of more parameters.

Question 4. [3 marks] Compute the lower and upper endpoints of observed asymptotically exact $1-\alpha$ "Wald" confidence intervals for each component of θ, for $\alpha = 0.05$. I recommend that you write a function, so you can use it in later questions.

```
compute.CI.endpoints <- function(X, y, alpha) {
  mle <- maximize.ell(ell, score, X, y, rep(0,d))

  # some code here

  # compute the lower and upper endpoints
  # lower and upper should be vectors whose length is the same length as mle
  lower <- mle - 1 # obviously wrong
  upper <- mle + 1 # obviously wrong
  return(list(lower=lower,upper=upper))
}

ci <- compute.CI.endpoints(X, y, 0.05)
```

Solution

```
compute.CI.endpoints <- function(X, y, alpha) {
  mle <- maximize.ell(ell, score, X, y, rep(0,d))
  inv.FIM <- solve(-hessian(mle, X))
  standard.deviations <- sqrt(diag(inv.FIM))
  z.alpha <- qnorm(1-alpha/2)
  lower <- mle - z.alpha*standard.deviations
  upper <- mle + z.alpha*standard.deviations
  return(list(lower=lower,upper=upper))
}

compute.CI.endpoints(X, y, 0.05)
```

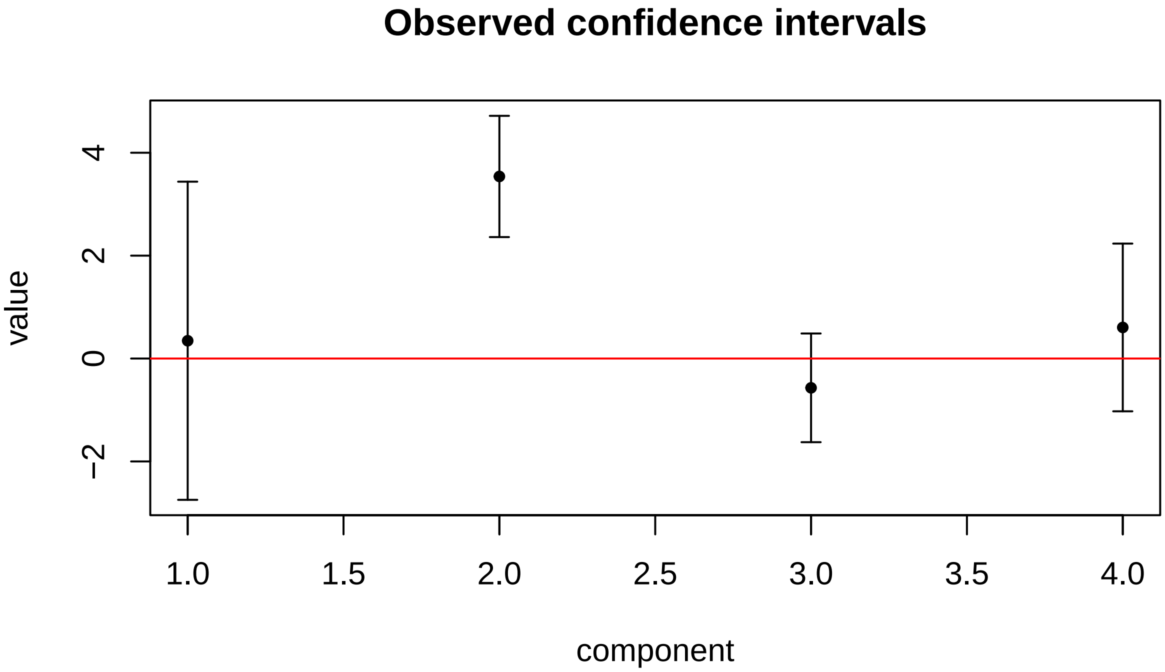
```
## $lower
##                  MO_Emp      Female      Parent         Age     PANSS_G
## -2.74657539  2.36058528 -1.62605347 -1.02547757 -0.12795070 -0.05071875
##     PANSS_P     PANSS_N
## -0.17565573 -0.14518954
##
## $upper
##                    MO_Emp       Female       Parent          Age      PANSS_G
##  3.437729826  4.717389228  0.486711343  2.234309938  0.054341425  0.091857258
##      PANSS_P      PANSS_N
## -0.003079036  0.057768778
```

If you have written the `compute.CI.endpoints` function above, the following code will visualize the observed confidence intervals in two separate plots.

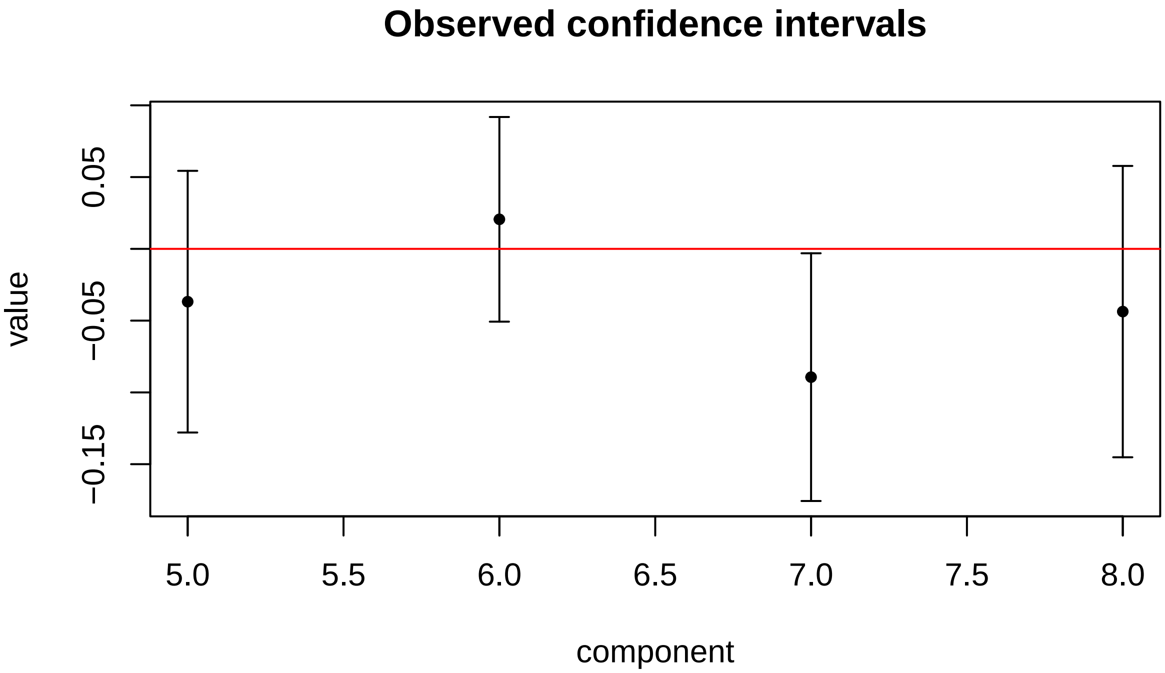
```
ci <- compute.CI.endpoints(X, y, 0.05)

plot.ci <- function(mle, CI.L, CI.U, components) {
  plot(components, mle[components], pch=20, main="Observed confidence intervals",
       xlab="component", ylab="value", ylim=c(min(CI.L[components]), max(CI.U[components])))
  arrows(components, CI.L[components], components, CI.U[components], length=0.05, angle=90, code=3)
  abline(h=0.0, col="red")
  axis(side=1, at=components, labels = FALSE)
}

plot.ci(mle, ci$lower, ci$upper, 1:4)
```



```
plot.ci(mle, ci$lower, ci$upper, 5:8)
```



6 Repeated experiments

Now we will perform repeated experiments under the assumption that $\theta^* = (-0.7, 3.5, 0, 0, 0, 0, -0.08, 0)$. Notice that if the true value of a component of θ is exactly 0, this means that the corresponding explanatory variable has no effect on the probability of the response. For this reason, it is often interesting to look at whether an observed confidence interval includes or excludes the value 0.

In order to perform repeated experiments, one needs to be able to simulate data according to the model. The following function should be helpful.

```
# generate data associated with the matrix X when theta is the true value of the parameter
generate.ys <- function(X, theta) {
  n <- dim(X)[1]
  rbinom(n, size = 1, prob=sigma(X%*%theta))
}
```

Question 5. [2 marks] Under the assumption that $\theta^* = (-0.7, 3.5, 0, 0, 0, 0, -0.08, 0)$, letting $\alpha = 0.05$ and using the same X matrix, approximate:

1. the coverage of the asymptotically exact $1 - \alpha$ confidence interval for θ_7,
2. the probability that the asymptotically exact $1 - \alpha$ confidence interval for θ_7 excludes 0,
3. the coverage of the asymptotically exact $1 - \alpha$ confidence interval for θ_8,
4. the probability that the asymptotically exact $1 - \alpha$ confidence interval for θ_8 excludes 0.

I suggest you write functions using the following template.

```
# returns the proportion of trials in which the condition holds
prop.condition <- function(trials, X, theta, component) {
  count <- 0
  for (i in 1:trials) {
    # simulate synthetic data using X and theta
    # compute observed confidence interval(s)
    # if condition is satisfied, increment count
    #   e.g. if theta[component] is in the relevant observed confidence interval,
    #        or if the relevant observed confidence interval excludes 0
  }
  count/trials
}
```

Solution

```
prop.inside <- function(trials, X, theta, component) {
  count <- 0
  for (i in 1:trials) {
    synthetic.y <- generate.ys(X, theta)
    ci <- compute.CI.endpoints(X, synthetic.y, 0.05)
    # check whether theta.star[component] is in the relevant CI
    if (ci$lower[component] < theta[component] && theta[component] < ci$upper[component]) {
      count <- count + 1
    }
  }
  count/trials
}

prop.exclude.zero <- function(trials, X, theta, component) {
  count <- 0
  for (i in 1:trials) {
    synthetic.y <- generate.ys(X, theta)
    ci <- compute.CI.endpoints(X, synthetic.y, 0.05)
    # check whether the relevant CI excludes 0
    if (ci$lower[component]*ci$upper[component] > 0) {
      count <- count + 1
    }
  }
  count/trials
}
```

```
theta.star <- c(-0.7,3.5,0,0,0,0,-0.08,0)
trials <- 1000
prop.inside(trials, X, theta.star, 7)
```

```
## [1] 0.946
```

```
prop.exclude.zero(trials, X, theta.star, 7)
```

```
## [1] 0.492
```

```
prop.inside.8.estimate <- prop.inside(trials, X, theta.star, 8)
# the event that the CI for theta_8 excludes 0 is the complement
# of the event that theta_8 is in the CI, since theta_8 is 0.
prop.excluding.zero.8.estimate <- 1 - prop.inside.8.estimate
prop.inside.8.estimate
```

```
## [1] 0.959
```

```
prop.excluding.zero.8.estimate
```

```
## [1] 0.041
```

Now for the purpose of performing repeated experiments, we will assume that we have twice as much data, in the sense that there are two Bernoulli observations for each row of the X matrix. One can equivalently construct a larger X matrix with two copies of each row in the original X matrix, as follows:

```
big.X <- rbind(X, X)
```

Question 6. [2 marks] Under the assumption that $\theta^* = (-0.7, 3.5, 0, 0, 0, 0, -0.08, 0)$, letting $\alpha = 0.05$ and using now the `big.X` matrix, approximate:

1. the coverage of the asymptotically exact $1 - \alpha$ confidence interval for θ_7,
2. the probability that the asymptotically exact $1 - \alpha$ confidence interval for θ_7 excludes 0,
3. the coverage of the asymptotically exact $1 - \alpha$ confidence interval for θ_8,
4. the probability that the asymptotically exact $1 - \alpha$ confidence interval for θ_8 excludes 0.

Explain any differences between the results for the previous question.

Solution

```
prop.inside(trials, big.X, theta.star, 7)
```

```
## [1] 0.94
```

```
prop.exclude.zero(trials, big.X, theta.star, 7)
```

```
## [1] 0.778
```

```
prop.inside.8.estimate <- prop.inside(trials, big.X, theta.star, 8)
prop.excluding.zero.8.estimate <- 1 - prop.inside.8.estimate
prop.inside.8.estimate
```

```
## [1] 0.944
```

```
prop.excluding.zero.8.estimate
```

```
## [1] 0.056
```

The coverage of the CIs is still very close to $1 - \alpha$, as it should be. This also implies that the probability that the CI for θ_8 excludes zero is approximately 0.05. The probability that the CI for θ_7 excludes zero increases, since θ_7 is indeed non-zero and the width of the observed CIs decreases with larger n.

---

7 Epilogue

Analyzing this data using this model does suggest a possible statistical relationship between a higher positive scale PANSS score and reduced probability of being in EET after one year. You may notice that the statistical evidence is much weaker than in CP1, where we looked at the relationship between chlorine hand-washing and mortality after childbirth.

It is not possible to give a causal interpretation to this statistical relationship. For example, it is not clear from the data and the analysis how or why higher positive scale scores have any effect on EET: this could be because such scores are associated with characteristics that hinder ability to engage in EET, or that those characteristics are more highly stigmatized, or various other possible explanations.

The study motivates understanding the statistical relationship between PANSS scores and EET as a route to improve vocational interventions. You may be interested to know that the effect of such interventions, e.g. Individual Placement and Support is often also analyzed from a statistical perspective.

Finally, you may be happy to learn that some of what we have done in this computer practical can be done with less effort using standard R functionality. For example, for this dataset one could produce the ML estimate and particular observed confidence intervals via three simple commands. However, using statistical software without understanding what it is doing it can easily lead to serious errors in decision-making.

```
model <- glm(Y1_Emp ~ ., family=binomial, data=fep.eet)
summary(model)
confint(model, level=0.95)
```

Note that the observed confidence intervals produced are slightly different to observed Wald confidence intervals, so you will not get any marks if you report these as your answers for Q4. The point of this practical is for you to learn how such quantities can be calculated.

If you are interested in linear and generalized linear models, like this one, there is a third year unit you can take. These models are very widely used.