Computational Statistics & Probability

Problem Set 1 - Bayesian Inference
Due: 23:59:59 16.nov.2022

Fall 2022

Instructions

Assignments must be submitted through Canvas. See the course Canvas page for policies covering collaboration, acceptable file formats (`.Rmd` & `.pdf`), and late submissions. Completed assignments must include executable code (`.Rmd`) and a corresponding knitted markdown file (`pdf`). An R Markdown cheat sheet is available.

1. COVID Home Test

A home COVID-19 antigen test developed by BinaxNOW is accepted in the US for travel. BinaxNOW reports that in a clinical trial their test correctly gave a positive result 84.6% of the time and correctly gave a negative result 98.5% of the time. We are assuming a PCR result is ground truth: if a PCR test is positive, then the patient has COVID-19; if a PCR test is negative, then the patient does not have COVID-19.

Suppose that 10% of the people in your community are currently infected. (This is your base rate for exposure.) Now suppose you take a BimaxNOW COVID-19 antigent test.

a) Given that your BinaxNOW test is positive, what is the probability that you have COVID-19?

```
# ANSWER
# The following three questions are answered by applying Bayes' theorem.  Since the
# variable of interest is binary (i.e., you either have COVID-19 or you don't),
# for the first question you can apply:
#
```

$$P(C \mid T) = \frac{P(T \mid C) \times P(C)}{(P(T \mid C) \times P(C)) + (P(T \mid \overline{C}) \times P(\overline{C}))}$$

```
#
# where P(C) is the prior probability of contracting COVID-19, P(T|C) is the likelihood
# of a positive home test given that you have COVID-19, and  C-overbar is the event of
# not having COVID-19, read "not-C", such that P(not-C) = 1- P(C). The denominator
# is the marginal likelihood.  To be explicit, if we instead treat T and C as binary
# random variables, then we have:
#
#  C = 1 encodes the event "has COVID-19"
#  C = 0 encodes the event "does not have COVID-19"
#  T = 1 encodes the event "tested positive for COVID-19 antigens"
#  T = 0 encodes the event "tested negative for COVID-19 antigens"
#
# ANSWER 1a:
# The question asks you to compute the conditional probability Pr(C=1 | T=1),
# which requires three values to compute an answer:
#
prior <- 0.10                  # Pr(C = 1)
```

```
likelihood <- 0.846            # Pr(T = 1 | C = 1)
neglikelihood <- 1- 0.985      # Pr(T = 1 | C = 0) = 1 - Pr(T = 0 | C = 0)

marginal_likelihood <- (prior * likelihood) + ((1-prior) * (neglikelihood))
(prior * likelihood) / marginal_likelihood
```

```
## [1] 0.8623853
```

b) Given that your BinaxNOW test is negative, what is the probability that you have COVID-19?

```
# ANSWER 1b:
# The question asks you to compute P(C=1 | T=0). Note
# here that the likelihood is 1 - P(T=1 | C =1) = 1- 0.846:

prior <- 0.10
likelihood <- 1-  0.846
neglikehood <- 0.985

marginal_likelihood <- (prior * likelihood) + ((1-prior) * (1-neglikelihood))
((prior * likelihood) / marginal_likelihood )
```

```
## [1] 0.01707506
```

```
# The probability that you have Covid-19 given a negative result is low, approximately
#  1.7%.

# Nevertheless, this low probability depends on the prior.  Is the decision
# to let you fly too dependent on the prior?  We explore that question, next.
```

c) Suppose instead the base rate for infection is 30%. This is extreme: The peak daily number of cases in Germany (so far) was 250,000 in April 2022. People are considered infectious for 10 days. So, no more than 2.5M people were infected with COVID in Germany during this peak, which is 3% of the population. Do you think a negative test result of a BinaxNOW home test is sufficient to conclude that you do not have COVID-19? Why or why not?

```
# ANSWER 1c:
# The question asks you to compute Pr(C=1 | T=0).  We simply change the prior
# and recompute the posterior probability with the given likelihoods:
prior <- 0.30

marginal_likelihood <- (prior * likelihood) + ((1-prior) * (1-neglikelihood))
((prior * likelihood) / marginal_likelihood )
```

```
## [1] 0.06279734
```

```
# Even with a prior 10x higher than an exceedingly conservative estimate of the
# peak exposure (so far) to infection, the probability of having COVID-19 given
# a negative BinaxNOW home test is about 6.3%.
#
# So, at the time of testing (t0), you might cautiously accept that you do not have
# COVID-19. It is another question whether the airline or health safety officials
# ought to be satisfied with this error rate under such extreme conditions, however.
#
# Further, if you have specific knowledge that you were exposed to the virus,
# that knowledge should dominate the analysis given here, as your specific
# knowledge of exposure dominates the less specific base-rate (prior) of exposure.
#
```

```
# Finally, if you remain in a population with a 30% exposure for t0+t length of time,
# your confidence that you do not have COVID-19 should decrease as a function of t.
# How much should your confidence decrease over time? That depends on the
# transmissibility of the variant of Corona-19 virus you were exposed to.  A simple
# set of differential equations, informed by the epidemiology of that variant,
# can be given to estimate your risk over time, and this estimate can be fed
# into your probabilistic model.  While modeling dynamical systems is beyond
# the scope of this course, we will nevertheless see an example of how to do
# this within our Bayesian framework later in the course.
```

2. Swing Voters

Imagine a country where there are only two political parties, Red and Blue, which divide the electorate equally. One difference between registered Blue voters and registered Red voters is their willingness to vote for the opposing party's candidate. Blue voters vote Red 20% of the time, otherwise they vote Blue. Red voters vote Blue 10% of the time, otherwise they vote Red. Voters who switch are called *swing voters*.

Smith was a swing voter in the last election but you do not know whether he is Red or Blue. (Nobody changes parties.) What is the probability that Smith will be a swing voter in the next election?

```
# ANSWER 2:
# The form of the question is a conditional probability. Given that Smith was a
# swing voter in the previous election (swing1), what is the probability he is
# a swing voter in the next (swing2), that is:
#
```

$$P(swing\_2 \mid swing\_1) = \frac{P(swing\_2 , swing\_1)}{P(sing\_1)}$$

```
# which is then answered by calculating the joint probability P(swing1, swing2)
# and the marginal probability P(swing1).

# The marginal probability P(swing1) is simply the probability of a voter in
# this equally-divided country being a swing voter, which is
#
p_twins <- 0.5*0.1 + 0.5*0.2
p_twins
```

```
## [1] 0.15
```

```
# The probability that a Blue voter is a swing voter in two successive elections is
# 0.2 * 0.2 = 0.04. The probability that a Red voter is a swing voter in two
# successive elections is 0.1 * 0.1 = 0.01. There is an equal chance that a voter
# is Red or Blue, so P(swing1, swing2) =
p_joint <- 0.5 * 0.04 + 0.5*0.01
p_joint
```

```
## [1] 0.025
```

```
# Finally, the conditional probability P(swing2 | swing1) is
p_joint/p_twins
```

```
## [1] 0.1666667
```

```
# Observe that P(swing2|swing1) > P(swing1).  Although we do not know which party
# Smith belongs to, learning that he was a swing voter in the last election
# provides some information about which party Smith belongs to, which is then
```

```
# factored into the estimate of the probability that he will be a swing voter
# in this election.
```

3. More Precision

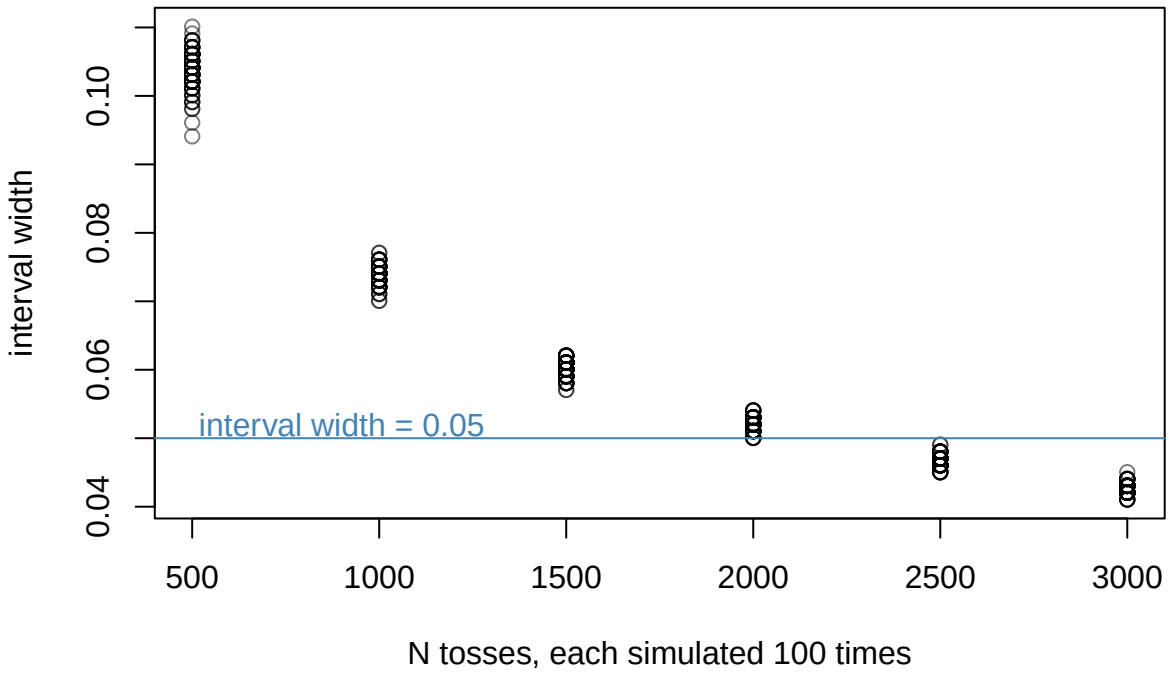
Suppose you want a very precise estimate of the proportion of the Earth's surface that is covered in water. Specifically, suppose you would like the 99% percentile interval of the posterior distribution of p (the estimated proportion of water) to have a width of no greater than 0.05 – that is, the distance between the lower and upper bound on p should be no greater than 0.05. How many times must you toss the globe to achieve this precision? An exact count is unnecessary. I am primarily interested in your approach.

```
# ANSWER 3:
# Here is one approach, which wraps the code for globe tossing into
# a function (with a flat prior), calculates the PI interval at 0.99, then
# returns the numeric difference between the upper (PI_99[2]) and lower
# (PI_99[1]) bounds.

interval_width <- function(N){
  p_true <- 0.71
  W <- rbinom( 1, size = N, prob=p_true)
  p_grid <- seq( from=0 , to=1 , length.out=1000 )
  prior <- rep(1, 1000)
  prob_data <- dbinom( W, size=N, prob=p_grid)
  posterior <- prob_data * prior
  posterior <- posterior / sum(posterior)
  samples <- sample( p_grid, prob=posterior, size=1e4, replace=TRUE)
  PI_99 <- PI( samples, 0.99)
  return(as.numeric( PI_99[2] - PI_99[1]))
}

N_list <- c(500, 1000, 1500, 2000, 2500, 3000)
N_list <- rep( N_list, each=100 )
width <- sapply( N_list, interval_width)

plot(N_list , width, col=grau(),
     xlab="N tosses, each simulated 100 times", ylab="interval width")
abline( h=0.05, col="steelblue")
text(900, 0.052, "interval width = 0.05", col="steelblue")
```




```
# A simulation of 2500 tosses of the globe repeated 100 times yields a maximum
# interval length less than 0.05, calculated by `max(width[401:500])`:
```

```
## The max interval width of 2500 tosses over 100 simulations is:  0.04904905
```