Worksheet 1

Foundations of Bayesian Methodology

Group: Goliath

Spring Semester 2022

Exercise 4 (Bayes theorem)

$$
\begin{aligned}
P[A \mid B, I] &= \frac{P[A, B, I]}{P[B, I]} && \text{conditional probability} \\
&= \frac{P[B \mid A, I] \cdot P[A, I]}{P[B, I]} && \text{apply chain rule to } P[A, B, I] \\
&= \frac{P[B \mid A, I] \cdot P[A \mid I] \cdot P[I]}{P[B, I]} && \text{apply chain rule to } P[A, I] \\
&= \frac{P[B \mid A, I] \cdot P[A \mid I] \cdot P[I]}{P[B \mid I] \cdot P[I]} && \text{apply chain rule to } P[B, I] \\
&= \frac{P[B \mid A, I] \cdot P[A \mid I]}{P[B \mid I]} && P[I] \text{ cancels out}
\end{aligned}
$$

Exercise 5 (Application of the Bayes theorem)

Given information:

- Sensitivity: $P[T^+ \mid D^+] = 0.96$
- Specificity: $P[T^- \mid D^-] = 0.97$
- Prior: $P[D^+] = 0.002$

From given information, we can derive:

- $P[T^- \mid D^+] = 1 - P[T^+ \mid D^+] = 0.04$
- $P[T^+ \mid D^-] = 1 - P[T^- \mid D^-] = 0.03$
- $P[D^-] = 1 - P[D^+] = 0.998$

$$
\begin{aligned}
P[D^- \mid T^+] &= \frac{P[D^-, T^+]}{P[T^+]} \\
&= \frac{P[T^+ \mid D^-] \cdot P[D^-]}{P[T^+ \mid D^-] \cdot P[D^-] + P[T^+ \mid D^+] \cdot P[D^+]} \\
&= \frac{0.03 \times 0.998}{0.03 \times 0.998 + 0.96 \times 0.002} \\
&\approx 0.94
\end{aligned}
$$

This diagnostic test seems reliable at the first glance since the sensitivity is 0.96 and the specificity is 0.97. From a Bayesian perspective, however, we see that the probability that someone is healthy given he or she is tested positive is 0.94. In other words, if someone is tested positive, he or she is actually

healthy with a probability of 0.94. This diagnostic test in this sense is not reliable at all. The prevalence of this disease in the population is 0.002, which indicates a rare disease. The PPV (positive predictive value) is an important parameter for the evaluation of diagnostic tests. It is defined as being diseased when having a positive test result. In this case it is calculated by:

$$PPV = P[D^+ \mid T^+] = 1 - P[D^- \mid T^+] = 1 - 0.94 = 0.06$$

The low PPV indicates that the ratio of patients truly diagnosed as positive to all those who had positive test result is low, seen in the alternative expression Altman and Bland (1994):

$$PPV = \frac{TP}{TP + FP} = 0.06$$

With TP = True positives and FP = False Positives. On the other hand, the NPV designates the ratio of truly diagnosed negatives to all who had a negative test result. It is worth mentioning that depending on the category of the test outcome, one or the other value might be of higher importance. For instance, for SARS-CoV-2 antigen rapid tests, it is more important that negative predictions are true, hence the NPV should be considered, which is related to the specificity. On the other hand, for e.g. SARS-CoV-2 rapid antibody tests, the PPV is more important and thus the sensitivity should be considered first.

Furthermore, all these values are impacted by the prevalence of the disease. As the prevalence increases, the PPV also increases but the NPV decreases Tenny and Hoffman (2017), even though the test's specificity and sensitivity remain identical. This seemingly unintuitive result highlights the importance of the Bayesian methodology.

Exercise 6 (Monte Carlo: random sample vs the true distribution)

```
library(coda) # trace plot
library(pCalibrate) # calibration of p-values
```

```
set.seed(44566) # set seed for reproducibility

# X ~ N(160, 20^2)
mu <- 160
sigma <- 20

# Generate a Monte Carlo sample of size 1000
mc.samples <- rnorm(1000, mean = mu, sd = sigma)
```

6.1 True Values Given that the random variable X follows a normal distribution with $\mu = 160$ and $\sigma = 20$, we can know that:

- Expectation of X: 160
- Standard deviation of X: 20
- Variance of X: 400
- Median of X: 160

```
# Report the (0.025, 0.5, 0.975) quantiles of X
qnorm(c(0.025, 0.5, 0.975), mean = mu, sd = sigma)
```

```
## [1] 120.8007 160.0000 199.1993
```

```r
# Plot the traceplot of the MC sample for X
traceplot(mcmc(mc.samples), ylab="X")
```

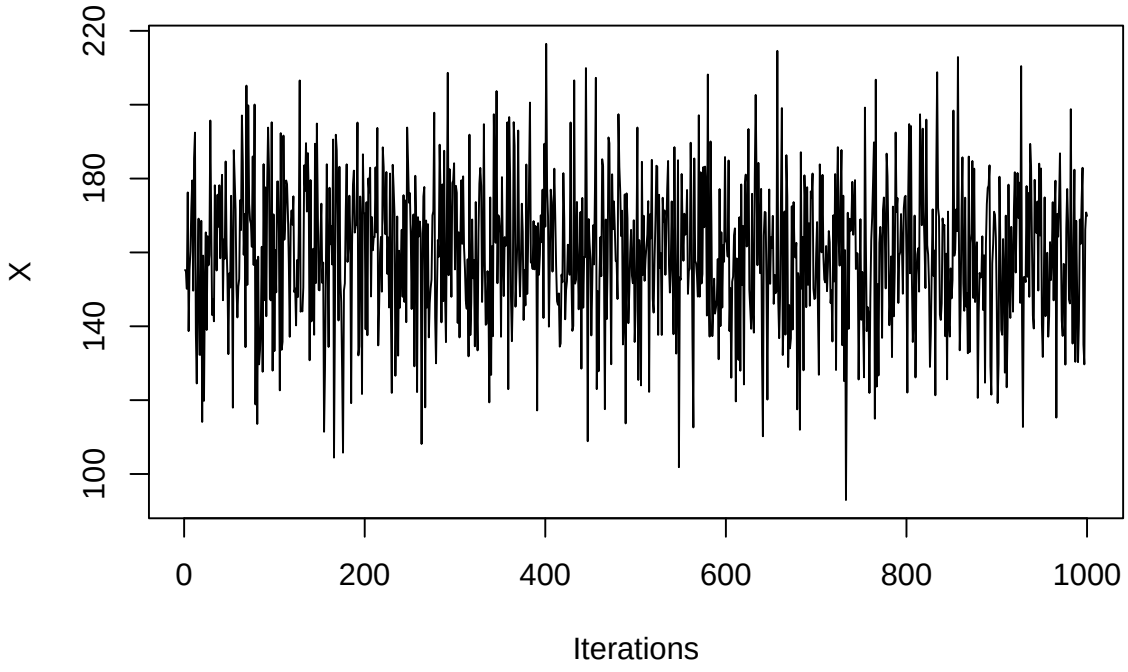

Figure 1: Traceplot of the MC sample for X

6.2 Traceplot

```r
hist(mc.samples, freq = FALSE, breaks = 20, xlab="X") # generate a histogram
lines(density(mc.samples), col = "red", lwd = 2) # add empirical density
lines(seq(100, 220), dnorm(seq(100, 220), mean = mu, sd = sigma),
      col = "blue", lwd = 2) # add true density
legend(x = "topright", legend = c("empirical density", "true density"),
       col = c("red", "blue"), lwd = 2) # add legend
```

6.3 Histogram and True Density Curve

```r
# Summary statistics of MC sample
sample.mean <- mean(mc.samples)
sample.sd <- sd(mc.samples)
sample.var <- var(mc.samples)
sample.med <- median(mc.samples)
```

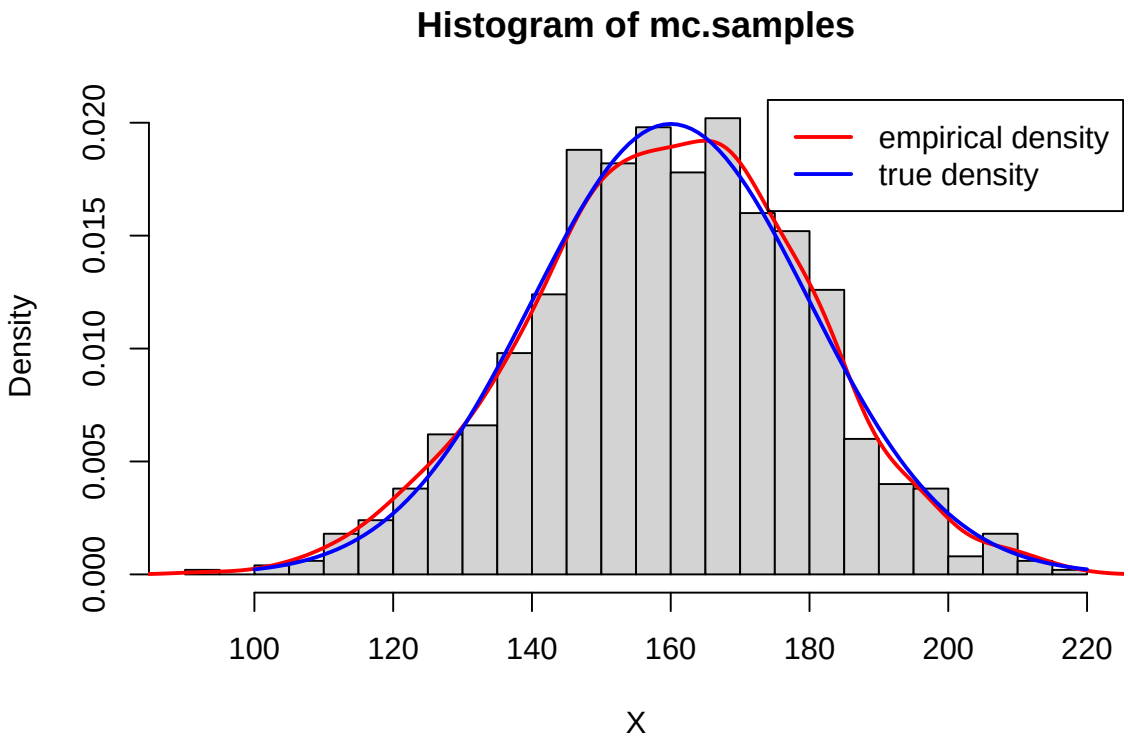

Figure 2: Histogram of the MC sample for X

```
cat(sprintf("The sample mean is %.4f
            \nThe sample standard deviation is %.4f
            \nThe sample variance is %.4f\n
            \nThe sample median is %.4f\n\n",
            sample.mean, sample.sd, sample.var, sample.med))
```

6.4 Sample Values

```
## The sample mean is 159.6127
##
## The sample standard deviation is 19.8155
##
## The sample variance is 392.6537
##
##
## The sample median is 159.8035
```

```
# Sample quantiles
quantile(mc.samples, probs = c(0.025, 0.5, 0.975))
```

```
##     2.5%      50%    97.5%
## 119.6477 159.8035 197.4104
```

6.5 Probabilities

$$P[X > 175] = 1 - P[X \leq 175]$$

Table 1: Descriptive statistics

	Mean	Standard deviation	Variance	Median	2.5% quantile	2.5% quantile	2.5% quantile
True	160.0000	20.0000	400.0000	160.0000	120.8007	160.0000	199.1993
Sample	159.6127	19.8155	392.6537	159.8035	119.6477	159.8035	197.4104

$$P[150 < X < 180] = P[X \leq 180] - P[X \leq 150]$$

```
# Compute the empirical cumulative distribution function for MC sample
Fn <- ecdf(mc.samples)

# Calculate P[X>175] from empirical CDF
1 - Fn(175)
```

```
## [1] 0.225
```

```
# Calculate P[150<X<180] from empirical CDF
Fn(180) - Fn(150)
```

```
## [1] 0.536
```

```
# Calculate P[X>175] from true CDF
pnorm(175, mean = mu, sd = sigma, lower.tail = FALSE)
```

```
## [1] 0.2266274
```

```
# Calculate P[150<X<180] from true CDF
pnorm(180, mean = mu, sd = sigma) - pnorm(150, mean = mu, sd = sigma)
```

```
## [1] 0.5328072
```

Exericse 7 (Bayes Factor)

7(a) Since the normal prior is conjugate to the normal likelihood with **known variance**, we can avoid integration and use the following formula to compute the marginal distribution:

From **Exercise 4** we derived:

$$P[A \mid B, I] = \frac{P[B \mid A, I] \cdot P[A \mid I]}{P[B \mid I]}$$

Apply the Bayes theorem to our context:

$$f(y \mid \mu, H_1) = \frac{f(\mu \mid y, H_1) f(y \mid H_1)}{f(\mu \mid H_1)}$$

Rearrange the above equation:

$$f(y \mid H_1) = \frac{f(y \mid \mu, H_1) f(\mu \mid H_1)}{f(\mu \mid y, H_1)}$$

We leave out the additional conditioning on H_1 for readability:

$$f(y \mid H_1) = \frac{f(y \mid \mu) f(\mu)}{f(\mu \mid y)}$$

We know that:

$$\begin{aligned} y \mid \mu &\sim N(\mu, \kappa^{-1}) \\ \mu &\sim N(v, \lambda^{-1}) \end{aligned}$$

The likelihood function:

$$f(y \mid \mu) \propto \exp\left(-\frac{\kappa}{2}(y-\mu)^2\right)$$

The prior:

$$f(\mu) \propto \exp\left(-\frac{\lambda}{2}(\mu - v)^2\right)$$

The posterior density of μ:

$$\begin{aligned} f(\mu \mid y) &\propto f(y \mid \mu) \cdot f(\mu) \\ &\propto \exp\left(-\frac{\kappa}{2}(y-\mu)^2 - \frac{\lambda}{2}(\mu - v)^2\right) \\ &\propto \exp\left(-\frac{\kappa+\lambda}{2}\left(\mu - \frac{y\kappa + v\lambda}{\kappa+\lambda}\right)^2\right) \end{aligned}$$

Therefore:

$$\mu \mid y \sim N\left(\frac{y\kappa + v\lambda}{\kappa+\lambda}, \frac{1}{\kappa+\lambda}\right)$$

$$\begin{aligned} f(y \mid H_1) &= \frac{f(y \mid \mu) f(\mu)}{f(\mu \mid y)} \\ &= \frac{\frac{1}{\sqrt{2\pi}}\sqrt{\kappa}\exp(-\frac{\kappa}{2}(y-\mu)^2) \cdot \frac{1}{\sqrt{2\pi}}\sqrt{\lambda}\exp(-\frac{\lambda}{2}(\mu - v)^2)}{\frac{1}{\sqrt{2\pi}}\sqrt{\kappa+\lambda}\exp\left(-\frac{\kappa+\lambda}{2}(\mu - \frac{y\kappa+v\lambda}{\kappa+\lambda})^2\right)} \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\kappa\lambda}{\kappa+\lambda}}\exp\left(-\frac{\kappa}{2}(y-\mu)^2 - \frac{\lambda}{2}(\mu - v)^2 + \frac{\kappa+\lambda}{2}\left(\mu - \frac{y\kappa+v\lambda}{\kappa+\lambda}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\kappa\lambda}{\kappa+\lambda}}\exp\left(-\frac{\kappa\lambda}{2(\kappa+\lambda)}(y-v)^2\right) \end{aligned}$$

7(b)

$$f(y \mid H_0) = \frac{1}{\sqrt{2\pi}}\sqrt{\kappa}\exp\left(-\frac{\kappa}{2}(y-\mu_0)^2\right)$$

The Bayes factor is given:

$$\begin{aligned} BF_{01}(y) &= \frac{f(y \mid H_0)}{f(y \mid H_1)} \\ &= \frac{\frac{1}{\sqrt{2\pi}}\sqrt{\kappa}\exp\left(-\frac{\kappa}{2}(y-\mu_0)^2\right)}{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\kappa\lambda}{\kappa+\lambda}}\exp\left(-\frac{\kappa\lambda}{2(\kappa+\lambda)}(y-v)^2\right)} \\ &= \sqrt{\frac{\kappa+\lambda}{\lambda}}\exp\left(-\frac{\kappa}{2}\left((y-\mu_0)^2 - \frac{\lambda}{\kappa+\lambda}(y-v)^2\right)\right) \end{aligned}$$

7(c)

$$\begin{aligned} \lim_{\lambda\to 0} BF_{01}(y) &= \lim_{\lambda\to 0}\sqrt{\frac{\kappa+\lambda}{\lambda}}\exp\left(-\frac{\kappa}{2}\left((y-\mu_0)^2 - \frac{\lambda}{\kappa+\lambda}(y-v)^2\right)\right) \\ &= \lim_{\lambda\to 0}\sqrt{\frac{\kappa+\lambda}{\lambda}}\exp\left(-\frac{\kappa}{2}(y-\mu_0)^2\right) \\ &= \infty \end{aligned}$$

7(d) Plug the parameters into BF_{01}:

$$\begin{aligned}
BF_{01} &= \sqrt{\frac{\kappa+\lambda}{\lambda}} \exp\left(-\frac{\kappa}{2}\left((y-\mu_0)^2 - \frac{\lambda}{\kappa+\lambda}(y-v)^2\right)\right) \\
&= \sqrt{\frac{1+\frac{1}{2}}{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left((1-0)^2 - \frac{\frac{1}{2}}{1+\frac{1}{2}}(1-2)^2\right)\right) \\
&= \sqrt{3}\exp\left(-\frac{1}{3}\right) \\
&\approx 1.241
\end{aligned}$$

$$\begin{aligned}
\frac{P[H_0 \mid y]}{P[H_1 \mid y]} &= BF_{01}(y)\frac{P[H_0]}{P[H_1]} \\
\frac{P[H_0 \mid y]}{1 - P[H_0 \mid y]} &= BF_{01}(y)\frac{P[H_0]}{P[H_1]} \\
P[H_0 \mid y] &= \frac{BF_{01}(y)\frac{P[H_0]}{P[H_1]}}{1 + BF_{01}(y)\frac{P[H_0]}{P[H_1]}} \\
&= \frac{1.241}{1+1.241} \\
&\approx 0.554
\end{aligned}$$

The prior probability of H_0 ($\mu = \mu_0$) is 0.5, and after observing the data y such a probability is updated to 0.554 (posterior probability). The higher the BF_{01} the higher is the favor for H_0. It is only ≈ 1.241, which means that there is barely no evidence and gets also reflected in hardly updated posterior probability.

Exercise 8 (Calibration of *p*-values: pCalibrate)

```
df <- data.frame(matrix(c(14, 9, 23, 1, 5, 6, 15, 14, 29), nrow = 3, byrow = TRUE))
colnames(df) <- c("Responders", "Non-responders", "Total")
rownames(df) <- c("Secukinumab", "Placebo", "Total")
knitr::kable(df, "pandoc", align = "c", caption = "Contingency table")
```

Table 2: Contingency table

	Responders	Non-responders	Total
Secukinumab	14	9	23
Placebo	1	5	6
Total	15	14	29

```
tab <- matrix(c(14, 9, 1, 5), nrow = 2, byrow = TRUE)
(minBF <- twoby2Calibrate(x=tab, type="two.sided", alternative="simple")$minBF)
```

```
## [1] 0.2937708
```

```
(p.value <- twoby2Calibrate(x=tab, type="two.sided", alternative="simple")$p.value)
```

```
##       p.pb       p.ce       p.bl      p.mid      p.lie
## 0.08007663 0.13908046 0.08007663 0.07586207 0.06580151
```

```
formatBF(BF=minBF)
```

```
## [1] "1/3.4"
```

We see that $BF = \frac{1}{3.4} < 1$ so the BF is decreasing the prior odds of H_0 (i.e. the drug has no effect).

```
BF2pp(1/3.4, prior.prob=0.5)
```

```
## [1] 0.2272727
```

$$\underbrace{\frac{P[H_0 \mid y]}{P[H_1 \mid y]}}_{\text{Posterior odds}} = BF_{01}(y) \underbrace{\frac{P[H_0]}{P[H_1]}}_{\text{Prior odds}}$$

The Bayes factor BF_{01} transforms the prior odds $\frac{P[H_0]}{P[H_1]}$ to a posterior odds $\frac{P[H_0|y]}{P[H_1|y]}$

$$\frac{P[H_0 \mid y]}{1 - P[H_0 \mid y]} = BF_{01}(y) \frac{P[H_0]}{1 - P[H_0]}$$

$$P[H_0 \mid y] = \frac{BF_{01}(y) \frac{P[H_0]}{1-P[H_0]}}{1 + BF_{01}(y) \frac{P[H_0]}{1-P[H_0]}}$$

The prior probability of no effect $P[H_0]$ is equal to 50%. After observing the data, the probability is updated to 22.7% (posterior probability $P[H_0 \mid y]$).

Bibliography

Altman, D. G, and J M. Bland. 1994. "Statistics Notes: Diagnostic Tests 2: Predictive Values." *BMJ* 309 (6947): 102–2. https://doi.org/10.1136/bmj.309.6947.102.

Held, Leonhard, and Bove Sabanes Daniel. 2014. *Applied Statistical Inference: Likelihood and Bayes*. Springer Berlin Heidelberg.

Tenny, Steven, and Mary R Hoffman. 2017. "Prevalence."