Bayesian Data Analysis (bda)
Cédric Huwyler
Last Update: November 29, 2023

Fachhochschule
Nordwestschweiz

Exercise Solutions to *Bayes Rules!*

☞ The solutions provided here are inofficial and have not been checked carefully for typos and overall correcteness at all. Please send feedback and corrections to `cedric.huwyler@fhnw.ch`.

2 Bayes' Rule

2.1 a) $P(B|A) > P(B)$, b) $P(B|A) < P(B)$, c) $P(B|A) > P(B)$, d) $P(B|A) > P(B)$

2.2

a) $P(B|A) = 0.73$, b) $P(A) = 0.2$, c) $P(D) = 0.15$

d) $P(D|C) = 0.91$, e) $P(E \cap F) = 0.38$, f) $P(E|F) = 0.95$

2.3

a) Y is not binomial, but rather follows a Poisson distribution.

b) Y is binomial, given that the events of blooming are independent: $f(k) = \binom{27}{k} 0.9^k \, 0.1^{27-k}$.

c) Y is not binomial, but rather follows a geometric distribution.

d) Y is not binomial. Possibly (given enough data about Henry), it might follow an exponential or a Gamma distribution.

e) Y is not binomial, it cannot be run as a sequence of independent Bernoulli experiments.

f) Y is binomial: $f(k) = \binom{60}{k} 0.8^k \, 0.2^{60-k}$. Each show or no-show is an independent Bernoulli experiment because the guests do not know each other.

2.4 $P(A) = 0.05$, $P(\neg A) = 0.95$, $P(B|A) = 0.7$, $P(B|\neg A) = 0.03$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)} = \frac{0.7 \cdot 0.05}{0.7 \cdot 0.05 + 0.03 \cdot 0.95} \approx 0.55$$

Chances that Vampires exist are 55%. If Bella believed to 90% that Vampire's exist, the posterior probability would be 99.5%!

2.5 I = 'tree is infected', E = 'tree is elm', M = 'tree is maple', O = 'tree is of other species'

a) $P(I) = 0.18$

b) $P(M) = P(M|I)\,P(I) + P(M|\neg I)\,P(\neg I) = 0.8 \cdot 0.18 + 0.1 \cdot 0.82 \approx 0.226$

c) $P(I|M) = \frac{P(M|I)\,P(I)}{P(M)} = \frac{0.8 \cdot 0.18}{0.226} \approx 0.637$

d) Given the data (tree is a maple) the probability has risen from 18% prior to 63.7% (posterior). This makes sense since maples are over-represented in the infected trees and under-represented in the healthy tree group.

2.6 We need to know the proportion of restaurants with fewer than four stars that Sandra does not like, otherwise we cannot compute the evidence.

2.7 NB = 'person is non-binary', R = 'Matt swipes right'

a) $P(NB) = P(NB|R)\,P(R) + P(NB|\neg R)\,P(\neg R) = 0.2 \cdot 0.08 + 0.1 \cdot 0.92 = 0.108$

b) $P(R|NB) = \frac{P(NB|R)\,P(R)}{P(NB)} = \frac{0.2 \cdot 0.08}{0.108} = 0.148$

2.8 M = 'morning flight', D = 'flight is delayed'

a) $P(D|M) = \frac{P(M|D)\,P(D)}{P(M)} = \frac{0.4 \cdot 0.15}{0.3} = 0.2$

b) $P(M|\neg D) = \frac{P(\neg D|M)\,P(M)}{P(\neg D)} = \frac{(1-P(D|M))\,P(M)}{P(\neg D)} = \frac{0.8 \cdot 0.3}{0.85} \approx 0.282,$

wobei das Zwischenresultat aus a) benutzt wurde.

2.9

a) In the exercise, the prior probabilities $P(G) = 0.4$ and $P(\neg G) = 0.6$ are given, where G stands for 'good mood' and $\neg G$ for bad mood. In addition, the conditional probabilites $P(x \text{ texts}|G)$ and $P(x \text{ texts}|\neg G)$ are given. One could either list the conditional probabilities in a table, as in

	good mood	bad mood	total
0 texts	0.05	0.13	0.18
1-45 texts	0.84	0.86	1.7
>45 texts	0.11	0.01	0.12
Total	1	1	

or list the joint probabilities $P(x \text{ texts } \cap\ G) = P(x \text{ texts}|G)\ P(G)$ and $P(x \text{ texts } \cap\ \neg G) = P(x \text{ texts}|\neg G)\ P(\neg G)$, as in

	good mood	bad mood	total
0 texts	0.02	0.078	0.098
1-45 texts	0.336	0.516	0.852
>45 texts	0.044	0.006	0.05
Total	0.4	0.6	1

The latter table is the one asked for in the exercise.

b) $P(G) = 0.4$ (prior)

c) $P(> 45|G) = 0.11$ (likelihood)

d)

$$P(G| > 45) = \frac{P(> 45|G)\, P(G)}{P(> 45)} = \frac{P(> 45|G)\, P(G)}{P(> 45|G)\, P(G) + P(> 45|\neg G)\, P(\neg G)} = \frac{0.11 \cdot 0.4}{0.11 \cdot 0.4 + 0.01 \cdot 0.6} = 0.88$$

2.10 R = 'student is from rural area', U = 'student is from urban area'

a) $P(\text{LGBTQ}) = P(\text{LGBTQ}|R)\, P(R) + P(\text{LGBTQ}|U)\, P(U) = 0.1 \cdot 0.085 + 0.105 \cdot 0.915 \approx 0.1046$

b) $P(R|\text{LGBTQ}) = \dfrac{P(\text{LGBTQ}|R)\, P(R)}{P(\text{LGBTQ})} = \dfrac{0.1 \cdot 0.085}{0.1046} \approx 0.081$

c) Hier kann entweder noch einmal marginalisiert oder das Zwischenresultat aus a) verwendet werden. Mit dem Zwischenresultat aus a):

$$P(R|\neg\text{LGBTQ}) = \frac{P(\neg\text{LGBTQ}|R)\, P(R)}{P(\neg\text{LGBTQ})} = \frac{(1 - P(\text{LGBTQ}|R))\, P(R)}{1 - P(\text{LGBTQ})} \approx \frac{0.9 \cdot 0.085}{1 - 0.1046} \approx 0.0854$$

Mit erneuter Marginalisierung:

$$\begin{aligned} P(R|\neg\text{LGBTQ}) &= \frac{P(\neg\text{LGBTQ}|R)\, P(R)}{P(\neg\text{LGBTQ})} = \frac{P(\neg\text{LGBTQ}|R)\, P(R)}{P(\neg\text{LGBTQ}|R)\, P(R) + P(\neg\text{LGBTQ}|U)\, P(U)} \\ &= \frac{0.9 \cdot 0.085}{0.9 \cdot 0.085 + 0.895 \cdot 0.915} \approx 0.0854 \end{aligned}$$

2.11

a) $f(y|\pi) = \dbinom{6}{y}\, \pi^y\, (1-\pi)^{6-y}$

b) $f(y = 4|\pi = 0.3) = \dbinom{6}{4}\, 0.3^4\, 0.7^2 \approx 0.06$

c) $f(\pi|y = 4) \propto f(y = 4|\pi)\, f(\pi)$

Posteriors computed in R (calculation of evidence can be replaced by normalisation):

$$f(\pi|y = 4) \approx \begin{cases} 0.112 & \pi = 0.3 \\ 0.624 & \pi = 0.4 \\ 0.264 & \pi = 0.5 \end{cases}$$

The posterior distribution has shifted considerably towards a higher acceptance probability.

2.12

a) Binomial pmf: $f(y|\pi) = \binom{7}{y} \, \pi^y \, (1-\pi)^{7-y}$

b) $f(\pi|y=1) \propto f(y=1|\pi)\, f(\pi) = \binom{7}{1} \, \pi \, (1-\pi)^6 \, f(\pi)$

Posteriors computed in R (calculation of evidence can be replaced by normalisation):

$$f(\pi|y=1) \approx \begin{cases} 0.286 & \pi = 0.1 \\ 0.538 & \pi = 0.25 \\ 0.176 & \pi = 0.4 \end{cases}$$

c) The posterior model contains a significantly reduced probability for $\pi = 0.4$ that is distributed more or less equally onto the lower values for π.

d) $f(\pi|y=1) \approx \begin{cases} 0.288 & \pi = 0.1 \\ 0.241 & \pi = 0.25 \\ 0.471 & \pi = 0.4 \end{cases}$

Even though Kris had a similar performance as Miles, the data gathered was not enough to shift the mode from $\pi = 0.4$ to the lower values for π.

2.13

a) Naively, because of $47/80 \approx 0.6$ one would think that values around $\pi = 0.6$ get more support and support is removed from the lower and higher values of π.

b) $f(\pi|y=47) \propto f(y=47|\pi)\, f(\pi) = \binom{80}{47} \, \pi^{47} \, (1-\pi)^{33} \, f(\pi)$

Posteriors computed in R (calculation of evidence can be replaced by normalisation):

$$f(\pi|y=47) \approx \begin{cases} 0.00065 & \pi = 0.4 \\ 0.114 & \pi = 0.5 \\ 0.834 & \pi = 0.6 \\ 0.052 & \pi = 0.7 \end{cases}$$

This compares well to the guess in a). Because of the large amount of data (compared to previous exercises), the effect is quite large.

c) $f(\pi|y=470) \approx \begin{cases} 5 \times 10^{-26} & \pi = 0.4 \\ 2.7 \times 10^{-6} & \pi = 0.5 \\ 0.99997 & \pi = 0.6 \\ 1.0 \times 10^{-10} & \pi = 0.7 \end{cases}$

The likelihood has now an even stronger impact and the posterior is almost fully concentrated on $\pi = 0.6$.

2.14

a) Prior model:

π	0.15	0.25	0.5	0.75	0.85
$n(\pi)$	3	3	8	3	3
$f(\pi)$	0.15	0.15	0.4	0.15	0.15

b) $f(\pi|y=3) \propto f(y=3|\pi)\, f(\pi) = \binom{13}{3}\, \pi^3\, (1-\pi)^{10}\, f(\pi)$

Computed with R:

$$f(\pi|y=3) \approx \begin{cases} 0.355 & \pi = 0.15 \\ 0.470 & \pi = 0.25 \\ 0.174 & \pi = 0.5 \\ 0.00022 & \pi = 0.75 \\ 1.9 \times 10^{-6} & \pi = 0.85 \end{cases}$$

c) Since $3/13 \approx 0.25$, a lot of weight is shifted to $\pi = 0.25$ and values close to it from the high values for π that are now very unlikely to be true. It seems now most likely that the bus is late in 25% of the time.

2.15

a) The weights in $f(\pi)$ would be shifted towards higher values of π.

b) Even more weight would be shifted to low values of π, in particular to $\pi = 0.6$.

c) $f(\pi|y=10) \propto f(y=10|\pi)\, f(\pi) = \binom{15}{10}\, \pi^{10}\, (1-\pi)^5\, f(\pi)$

Computed with R:

$$f(\pi|y=10) \approx \begin{cases} 0.281 & \pi = 0.6 \\ 0.428 & \pi = 0.65 \\ 0.208 & \pi = 0.7 \\ 0.083 & \pi = 0.75 \end{cases}$$

d) The experimental data more or less confirm the prior expectations, only slightly increasing the belief that $\pi = 0.65$ is the real hatchling survival rate.

2.16

a) In the article several proportions of fake articles are mentioned, ranging from below 2% up to 99%. One could elicit a more or less symmetrical prior (assuming that the true answer lies in the mean - if there is no bias in the selection of statements through the author of the article).

Proposition:

π	0.02	0.2	0.4	0.7	0.99
$f(\pi)$	0.05	0.4	0.3	0.2	0.05

b) The prior introduced in a) includes also the extremes mentioned in the article. The main difference to the provided prior is the slight asymmetry of the proposed prior towards lower values, otherwise the two priors are very similar.

c) $f(\pi|y) \propto f(y|\pi)\, f(\pi) = \binom{15}{y}\, \pi^y\, (1-\pi)^{15-y}\, f(\pi)$

Computed with R, since solving analytically for y is probably impossible:

π	0.2	0.4	0.6
$f(\pi\|y=1)$	0.766	0.230	0.0045
$f(\pi\|y=2)$	0.545	0.436	0.019
$f(\pi\|y=3)$	0.299	0.638	0.063
$f(\pi\|y=4)$	0.126	0.715	0.159
$f(\pi\|y=5)$	0.042	0.639	0.319
$f(\pi\|y=6)$	0.011	0.465	0.523

At least 6 artworks would need to be forged.

3 The Beta-Binomial Bayesian Model

3.1 Since $E(\pi) = \frac{\alpha}{\alpha+\beta}$, the knowledge of $E(\pi)$ allows the specification of the ratio

$$\frac{\alpha}{\beta} = \frac{E(\pi)}{1-E(\pi)}$$

a) $\frac{\alpha}{\beta} = \frac{0.4}{0.6} = \frac{2}{3}$, a reasonable prior could be between Beta$(6,9)$ (broader) and Beta$(12,18)$ (a bit more narrow).

b) $E(\pi) = 0.8 \Rightarrow \frac{\alpha}{\beta} = 4$. Using `summarize_beta()` from the `bayesrules` package, Var$(\pi) = 0.05$ is approximately reached with Beta$(2, 0.5)$.

c) $E(\pi) = 0.9 \Rightarrow \frac{\alpha}{\beta} = 9$. The desired range is approximately reached with Beta$(63,7)$.

d) One could use something like Beta$(0.8, 0.8)$ here.

3.2

a) $\frac{\alpha}{\beta} = \frac{0.8}{0.2} = 4$, a reasonably narrow prior could be Beta$(64,16)$.

b) $E(\pi) = 0.9 \Rightarrow \frac{\alpha}{\beta} = 9$. Trial and error with the given ratio yields approximately Beta$(\frac{1}{10}, \frac{1}{90})$.

c) $E(\pi) = 0.85 \Rightarrow \frac{\alpha}{\beta} = \frac{17}{3}$. A reasonable candidate seems to be Beta$(8 \cdot \frac{17}{3}, 8)$.

d) $E(\pi) = 0.3 \Rightarrow \frac{\alpha}{\beta} = \frac{3}{7}$. Beta$(3,7)$ could be a good guess, or an even broader Beta$(1.5, 3.5)$.

3.3

a) Uniform distribution: Beta$(1,1)$

b) $E(\pi) = \frac{1}{1+1} = 0.5$. This does align with all values for π being equally likely, however not with not knowing π at all.

c) SD$(\pi) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}} = \sqrt{\frac{1}{12}} \approx 0.289$

d) e.g. Beta$(1.5, 1.5)$ with SD $= 0.25$.

e) e.g. Beta$(0.5, 0.5)$ with SD $= 0.35$.

3.4 Clues: symmetric if $\alpha = \beta$, right-skewed if $\alpha < \beta$ and left-skewed if $\alpha > \beta$.

a) Beta$(0.5, 0.5)$, b) Beta$(2, 2)$, c) Beta$(6, 2)$, d) Beta$(1, 1)$, e) Beta$(0.5, 6)$, f) Beta$(6, 6)$

3.5 Clues: symmetric if $\alpha = \beta$, right-skewed if $\alpha < \beta$ and left-skewed if $\alpha > \beta$.

- left-skewed ($\alpha > \beta$), with decreasing variance: a) Beta$(1, 0.3)$, c) Beta$(4, 2)$, f) Beta$(6, 3)$
- (slightly) right-skewed ($\alpha < \beta$): e) Beta$(5, 6)$
- linear (α=1 and $\beta = 2$ or vice versa): d) Beta$(2, 1)$
- symmetric ($\alpha = \beta$): b) Beta$(3, 3)$

3.6

a) Visually, (e) has the most of the mass concentrated around small values of π and thus has the smallest value for $E(\pi)$. Similarly, (c) has the largest value for $E(\pi)$ since most of its mass is at high values for π. The means are

(e) $E[\pi] = \frac{0.5}{6.5} \approx 0.077$

(c) $E[\pi] = \frac{6}{8} = 0.75$

b) There is some ambiguity in the answer here because example (a) is bimodal with modes 0 and 1 and therefore has both the smallest and the biggest mode. Of the unimodal examples, (e) has the smallest mode at $\pi = 0$ and (c) has the biggest mode (value with highest probability is most right).

Analytically, the mode of a Beta(α, β) distribution is given as

$$\text{Mode}[\pi] = \frac{\alpha - 1}{\alpha + \beta - 2}, \quad \text{for } \alpha, \beta > 1,$$

however this formula cannot be used for (a) and (e) since at least one parameter of the involved beta distributions is not larger than one. For (c) we find Mode$[\pi] = \frac{5}{6} \approx 0.833$.

c) It looks like (e) has the smallest standard deviation. Since it has the most weight at the extremes, (a) has probably the highest standard deviation.

3.8

a) Since $f(\pi)$ is a probability density function,

$$\int_0^1 f(\pi)\, \text{d}\pi = \frac{\Gamma[\alpha + \beta]}{\Gamma[\alpha]\, \Gamma[\beta]} \int_0^1 \pi^{\alpha - 1} (1 - \pi)^{\beta - 1}\, \text{d}\pi = 1$$

holds, and conversely the integral

$$\int_0^1 \pi^{\alpha - 1} (1 - \pi)^{\beta - 1}\, \text{d}\pi = \frac{\Gamma[\alpha]\, \Gamma[\beta]}{\Gamma[\alpha + \beta]}$$

can be computed for any α and β. The expectation $E[f(\pi)]$ can then be computed by making use of this property and also that the Gamma function satisfies $\Gamma[\alpha + 1] = \alpha\, \Gamma[\alpha]$:

$$\begin{aligned} E[\pi] &= \int_0^1 \pi \; f(\pi) \, \mathsf{d}\pi = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \int_0^1 \pi^{\alpha}(1-\pi)^{\beta-1} \, \mathsf{d}\pi = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \frac{\Gamma[\alpha+1]\,\Gamma[\beta]}{\Gamma[\alpha+\beta+1]} \\ &= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \frac{\alpha\,\Gamma[\alpha]\,\Gamma[\beta]}{(\alpha+\beta)\,\Gamma[\alpha+\beta]} = \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

b) The mode of a probability density $f(\pi)$ is $\pi_* = \arg\max_\pi f(\pi)$, the value of π with the highest value for $f(\pi)$, or with calculus:

$$\frac{\mathsf{d}}{\mathsf{d}\pi} f(\pi) \,\big|_{\pi=\pi_*} = 0.$$

The derivative can be computed as

$$\begin{aligned} \frac{\mathsf{d}}{\mathsf{d}\pi} f(\pi) &= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \frac{\mathsf{d}}{\mathsf{d}\pi} \left[\pi^{\alpha-1}(1-\pi)^{\beta-1}\right] \\ &= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \left[(\alpha-1)\,\pi^{\alpha-2}\,(1-\pi)^{\beta-1} - \pi^{\alpha-1}\,(\beta-1)\,(1-\pi)^{\beta-2}\right], \end{aligned}$$

and setting it to zero yields

$$(\alpha-1)\,\pi^{\alpha-2}\,(1-\pi)^{\beta-1} = \pi^{\alpha-1}\,(\beta-1)\,(1-\pi)^{\beta-2}.$$

Dividing by $\pi^{\alpha-2}(1-\pi)^{\beta-2}$ and solving for π leads to

$$\pi_* = \frac{\alpha-1}{\beta+\alpha-2}$$

Note that to prove that π_* is actually a maximum, we would need to compute the second derivative and assert that it is negative, however we save us the trouble and observe visually that the Beta(α, β) distribution is concave for $\alpha, \beta > 1$.

c) To compute the variance, we use that

$$\mathsf{Var}[\pi] = E[(\pi - E[\pi])^2] = E[\pi^2] - E[\pi]^2.$$

We have already shown that $E[\pi] = \frac{\alpha}{\alpha+\beta}$ and only need to compute $E[\pi^2]$ with the same method as for the expectation. Specifically,

$$\begin{aligned} E[\pi^2] &= \int_0^1 \pi^2 \; f(\pi) \, \mathsf{d}\pi = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \int_0^1 \pi^{\alpha+1}(1-\pi)^{\beta-1} \, \mathsf{d}\pi = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \frac{\Gamma[\alpha+2]\,\Gamma[\beta]}{\Gamma[\alpha+\beta+2]} \\ &= \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\,\Gamma[\beta]} \, \frac{\Gamma[\alpha+2]\,\Gamma[\beta]}{\Gamma[\alpha+\beta+2]} = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]} \, \frac{(\alpha+1)\,\alpha\,\Gamma[\alpha]}{(\alpha+\beta+1)\,(\alpha+\beta)\,\Gamma[\alpha+\beta]} \\ &= \frac{\alpha\,(\alpha+1)}{(\alpha+\beta+1)\,(\alpha+\beta)}, \end{aligned}$$

where the property $\Gamma[\alpha+1] = \alpha\,\Gamma[\alpha]$ has been applied twice. Then,

$$
\begin{aligned}
\text{Var}[\pi] &= E[\pi^2] - E[\pi]^2 = \frac{\alpha\,(\alpha+1)}{(\alpha+\beta+1)\,(\alpha+\beta)} + \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\,(\alpha+1)\,(\alpha+\beta) - \alpha^2\,(\alpha+\beta+1)}{(\alpha+\beta+1)\,(\alpha+\beta)^2} \\
&= \frac{\alpha\beta}{(\alpha+\beta+1)\,(\alpha+\beta)^2}
\end{aligned}
$$

3.9

a) Beta$(8,2)$: $E[\pi] = 0.8$, Mode$[\pi] = \frac{7}{8} = 0.875$, Var$[\pi] = \frac{16}{11 \cdot 10^2} \approx 0.015$

   Beta$(1,20)$: $E[\pi] = \frac{1}{21} \approx 0.0476$, Mode$[\pi] = 0$, Var$[\pi] = \frac{20}{22 \cdot 21^2} \approx 0.002$

b) See R notebook

c) In the eyes of the first person, almost everybody uses the word pop, and in the eyes of the second person, practically nobody uses the word pop.

3.10

a) Likelihood: $L(\pi|y=12) = \binom{50}{12}\,\pi^{12}\,(1-\pi)^{38}$

   Posteriors:

$$
P(\pi|y=12) \propto \pi^{\alpha-1}\,(1-\pi)^{\beta-1}\,\pi^{12}\,(1-\pi)^{38} = \pi^{\alpha+12-1}\,(1-\pi)^{\beta+38-1},
$$

   where all scalar factors have been omitted. When properly normalized, this is a Beta$(\alpha+12, \beta+38)$ distribution, thus a Beta$(20,40)$ distribution for the first person and a Beta$(13,58)$ distribution for the second person.

b) See R notebook

c) The first person expects now a proportion around 30%, the second person a proportion around 20%. The posteriors are now much closer than the priors, given the data.

3.11

a) $E[\pi] = 0.25 \Rightarrow \frac{\alpha}{\beta} = \frac{0.25}{0.75} = \frac{1}{3} \Rightarrow \beta = 3\alpha.$

   Plugging this into the formula for the mode yields Mode$[\pi] = \frac{\alpha-1}{4\alpha-2} = \frac{5}{22}$, solving for α results in $\alpha = 6$ and $\beta = 18$ (plot in R notebook).

b) Since $n = 50$ and $y = 15$, the posterior is a Beta$(6+15, 18+50-15) =$ Beta$(21,53)$ distribution.

c) $E[\pi] = \frac{21}{74} \approx 0.284$, Mode$[\pi] = \frac{20}{72} \approx 0.278$, SD$[\pi] = \sqrt{\frac{21 \cdot 53}{(75 \cdot 74^2)}} \approx 0.052.$

d) The mean and mode of the posterior are a bit closer to the mean and mode of the likelihood than to the prior.

3.12

a) $E[\pi] = 0.15 \Rightarrow \frac{\alpha}{\beta} = \frac{0.15}{0.85} \Rightarrow \alpha = k \cdot 0.15, \beta = k \cdot 0.85$ with k in $\mathbb{R}$.

The prior could be represented by a Beta$(15, 85)$ distribution (see R notebook).

b) $n = 90, y = 30 \Rightarrow$ The posterior follows a Beta$(15 + 30, 85 + 90 - 30) =$ Beta$(45, 145)$ distribution.

c) $E[\pi] = \frac{45}{190} \approx 0.237$, Mode$[\pi] = \frac{44}{188} \approx 0.234$, SD$[\pi] = \sqrt{\frac{45 \cdot 145}{190^2 \cdot 191}} \approx 0.031.$

d) The mean and mode of the posterior are more or less in the middle between the mean and mode of the likelihood and prior.

3.13

a) With no further assumptions given, we assume that the prior is symmetric around the expectation $E[\pi] = 0.475$, thus $\frac{\alpha}{\beta} = \frac{0.475}{0.525} \approx 0.9$. A good approximation could be a Beta$(36, 40)$ distribution.

b) $n = 200, y = 80 \Rightarrow$ The posterior follows a Beta$(36 + 80, 40 + 200 - 80) =$ Beta$(116, 160)$ distribution.

c) $E[\pi] = \frac{116}{276} \approx 0.420$, Mode$[\pi] = \frac{115}{274} \approx 0.420$, SD$[\pi] = \sqrt{\frac{116 \cdot 160}{276^2 \cdot 275}} \approx 0.030.$

d) The posterior is closer to the likelihood. This makes sense because 200 people are a lot of data.

3.14

Since a binomial likelihood with y out of n observations being a success transforms a Beta(α, β) prior to a Beta$(\alpha' = \alpha + y, \beta' = \beta + n - y)$ posterior, we can compute that

$$\begin{aligned} y &= \alpha' - \alpha = 11 - 2 = 9, \\ n &= \beta' - \beta + y = 24 - 3 + 9 = 30, \end{aligned}$$

i.e. 9 observations out of 30 where a success. This can be recovered with (see R notebook)

```
summarize_beta_binomial(alpha=2, beta=3, n=30, y=9).
```

3.15

Similar as in exercise 3.14,

$$\begin{aligned} y &= \alpha' - \alpha = 100 - 1 = 99, \\ n &= \beta' - \beta + y = 3 - 2 + 99 = 100, \end{aligned}$$

i.e. 99 observations out of 100 where a success. This can be recovered with (see R notebook)

```
summarize_beta_binomial(alpha=1, beta=2, n=100, y=99).
```

3.16

a) The prior model expects only extremly high values for π very close to 1. On the other hand, the likelihood is centered around a low $\pi = 0.25$. For the person who elicited the prior, the measured data must have come with a surprise.

b) The posterior has its mode a bit closer to the mode of the prior than the mode of the likelihood.

c) See R notebook. After some trial and error: `plot_beta_binomial(alpha=95, beta=2, n=43, y=10)`

3.17

a) The prior model expects values for π around 0.5, however with a very high standard deviation. The likelihood is much more narrow (indicating a large amount of data), centered around $\pi \sim 0.1$.

b) The posterior is very close to the likelihood that obviously presented strong evidence.

c) See R notebook. After some trial and error: `plot_beta_binomial(alpha=3, beta=3, n=88, y=11)`

3.18

a) Patrick's prior is an underestimate - the posterior is strongly pulled to the right by the likelihood (see R notebook).

b) Even though similar to Patrick's town, also in Harold's town 75% attend the protest, he has less data and consequently the posterior is less influenced by the prior than in Patrick's case.

c) Harold's posterior's mean is slightly closer to the mean of the posterior and also a bit broader. This is because the likelihood is wider and less data is available to narrow the posterior.

4 Balance and Sequentiality in Bayesian Analyses

4.1

a) $\alpha = \beta \Rightarrow$ symmetric (centering π on 0.5)

b) $\alpha > \beta \Rightarrow$ slightly left-skewed, $E[\pi] = \frac{3}{5} = 0.6$ (somewhat favoring $\pi > 0.5$)

c) $\alpha \ll \beta \Rightarrow$ strongly right-skewed, $E[\pi] = \frac{1}{11} \approx 0.09$ (strongly favoring $\pi < 0.5$)

d) $\alpha < \beta \Rightarrow$ right-skewed, $E[\pi] = \frac{1}{4} = 0.25$ (somewhat to strongly favoring $\pi < 0.5$)

e) $\alpha \gg \beta \Rightarrow$ left-skewed, $E[\pi] = \frac{17}{19} \approx 0.89$ (strongly favoring $\pi > 0.5$)

4.2

- Prior is right-skewed $\Rightarrow \alpha < \beta \Rightarrow$ arguments in c), d) and e) agree with this.
- Likelihood is perfectly symmetric $\Rightarrow y = n/2 \Rightarrow$ only e) agrees with this.
- e) presents the correct arguments.

4.3 (see R notebook for plots)

a) e.g. Beta$(1, 40)$

b) A common choice here is the uniform prior Beta$(1, 1)$.

c) e.g. Beta$(10, 2)$

d) e.g. Beta$(6, 3)$

e) e.g. Beta$(1, 5)$

4.4 (see R notebook for plots)

a) Kimya is not very sure, but thinks that chances are more towards the shop being closed.

b) Fernando is convinced that the shop is closed.

c) Ciara thinks that the shop might rather not be open anymore, is however not so sure.

d) Taylor is convinced that the shop is open.

4.5 See R notebook

4.6

- Kimya: Beta$(1+3, 2+7-3)$ = Beta$(4,6)$, $E[\pi] = 0.4$
- Fernando: Beta$(0.5+3, 1+7-3)$ = Beta$(3.5, 5)$, $E[\pi] \approx 0.41$
- Ciara: Beta$(3+3, 10+7-3)$ = Beta$(6, 14)$, $E[\pi] = 0.3$
- Taylor: Beta$(2+3, 0.1+7-3)$ = Beta$(5, 4.1)$, $E[\pi] \approx 0.55$

All posterior means are extremely close to the simulated values and only differ in the second digit behind the decimal point.

4.7 In the following it is used that (see book)

$$E[\pi|Y=y] = \frac{\alpha+\beta}{\alpha+\beta+n} E[\pi] + \frac{n}{\alpha+\beta+n} \frac{y}{n}.$$

In particular, we compute $\kappa := \frac{\alpha+\beta}{\alpha+\beta+n}$ to argue whether the posterior is closer to the prior (κ closer to 1) or closer to the likelihood (κ closer to 0).

a) $\kappa \approx 0.33$ (the data have more influence on the posterior)

b) $\kappa \approx 0.96$ (the prior has considerably more influence on the posterior)

c) $\kappa \approx 0.67$ (the prior has more influence on the posterior)

d) $\kappa = 0.50$ (the posterior is an equal compromise between the data and the prior)

e) $\kappa \approx 0.10$ (the data have considerably more influence on the posterior)

4.8 See R notebook. The plots support the results in exercise 4.7.

4.9

a) $E[\pi] = \frac{7}{9} \approx 0.78$, SD$[\pi] = \sqrt{\frac{14}{9^2 \cdot 10}} \approx 0.13$ $\Rightarrow$ Reasonable values for π are values approximately between 70-90%.

b) $Y = 19$, $n = 20$ leads to a Beta$(26, 3)$ posterior with mean $E[\pi] = \frac{26}{29} \approx 0.90$ and standard deviation SD$[\pi] = \sqrt{\frac{78}{29^2 \cdot 30}} \approx 0.05$ $\Rightarrow$ Reasonable values for π are approximately between 85-95%. The expectation was raised a bit and the uncertainty was reduced considerably by this data (prior and data agree).

c) $Y = 1$, $n = 20$ leads to a Beta$(8, 21)$ posterior with mean $E[\pi] = \frac{8}{29} \approx 0.28$ and standard deviation SD$[\pi] = \sqrt{\frac{168}{29^2 \cdot 30}} \approx 0.08$ $\Rightarrow$ Reasonable values for π are approximately between 20-36%. The expectation was lowered considerably and again the uncertainty was reduced by this data, however a bit less than in a) (prior and data disagree a bit).

d) $Y = 10$, $n = 20$ leads to a Beta$(17, 12)$ posterior with mean $E[\pi] = \frac{17}{29} \approx 0.59$ and standard deviation SD$[\pi] = \sqrt{\frac{204}{29^2 \cdot 30}} \approx 0.09$ $\Rightarrow$ Reasonable values for π are approximately between 50-68%. The expectation was lowered and the uncertainty was reduced (prior and data disagree strongly).

4.10 (see R notebook for plots)

a) $y = \alpha' - \alpha = 8$, $n = \beta' - \beta + y = 10$

b) $y = \alpha' - \alpha = 3$, $n = \beta' - \beta + y = 13$

c) $y = \alpha' - \alpha = 2$, $n = \beta' - \beta + y = 16$

d) $y = \alpha' - \alpha = 7$, $n = \beta' - \beta + y = 10$

e) $y = \alpha' - \alpha = 3$, $n = \beta' - \beta + y = 6$

f) $y = \alpha' - \alpha = 29$, $n = \beta' - \beta + y = 31$

4.11 (see R notebook for plots)

a) Beta$(11, 4)$, b) Beta$(1, 2)$, c) Beta$(101, 31)$, d) Beta$(21, 101)$, e) Beta$(235, 235)$

4.12 (see R notebook for plots)

a) Beta$(20, 5)$, b) Beta$(10, 3)$, c) Beta$(110, 32)$, d) Beta$(30, 102)$, e) Beta$(244, 236)$

4.13

a) See the R notebook for a plot of the prior.

b) The politician is 100% certain that not less than 50% has voted for them and has equal expectations for any value between 50-100%.

c) Posterior:

$$p(\pi|\text{data}) \propto p(\text{data}|\pi)\, f(\pi) = \binom{100}{0} \pi^0 (1-\pi)^{100}\, f(\pi) = \begin{cases} 0, & 0 < \pi < 0.5 \\ 2\,(1-\pi)^{100} & 0.5 \leq \pi < 1 \end{cases}.$$

See the R notebook for a plot of the posterior.

d) The posterior by definition has the same support as the prior. Since the likelihood is highly right-skewed, the prior will also be highly right-skewed, however with a support starting from $\pi = 0.5$. Hence there will be a tiny peak around $\pi = 0.5$ that will be strongly exaggerated when normalizing the posterior.

4.14

a) We know that for a Beta(α, β) distribution, Mode$(\pi) = \frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$. Measuring y positive outcomes out of n Bernoulli experiments leads us to a Beta$(\alpha + y, \beta + n - y)$ posterior with the mode Mode$(\pi|Y = y) = \frac{\alpha+y-1}{\alpha+\beta+n-2}$.

The goal of the exercise is to write the posterior mode as a weighted sum of the prior mode and the observed sample success rate:

$$\text{Mode}(\pi|Y = y) = a \cdot \text{Mode}(\pi) + (1 - a) \cdot \frac{y}{n}.$$

Notice that we have used here that the weights in front of the prior mode and the success rate should sum to 1. This leads to

$$\left(\text{Mode}(\pi) - \frac{y}{n}\right) a = \text{Mode}(\pi|Y = y) - \frac{y}{n},$$

and thus

$$a = \frac{\text{Mode}(\pi|Y = y) - \frac{y}{n}}{\text{Mode}(\pi) - \frac{y}{n}} = \frac{\frac{\alpha+y-1}{\alpha+\beta+n-2} - \frac{y}{n}}{\frac{\alpha-1}{\alpha+\beta-2} - \frac{y}{n}} = \frac{\frac{\alpha n-\alpha y-\beta y+2y-n}{n\,(\alpha+\beta+n-2)}}{\frac{\alpha n-\alpha y-\beta y+2y-n}{n\,(\alpha+\beta-2)}} = \frac{\alpha + \beta - 2}{\alpha + \beta + n - 2}.$$

Consequently,

$$1 - a = \frac{\alpha + \beta + n - 2 - (\alpha + \beta - 2)}{\alpha + \beta + n - 2} = \frac{n}{\alpha + \beta + n - 2},$$

leaving us with the weights in the equation given in the exercise.

b) Since $\lim\limits_{n\to\infty} a = 0$ (n dominating only in denominator) and $\lim\limits_{n\to\infty} (1 - a) = 1$ (n dominating both in nominator and denominator),

$$\lim_{n\to\infty} \text{Mode}(\pi|Y = y) = \frac{y}{n}.$$

With a large number of data, the posterior mode converges towards the observed sample success rate.

4.15

a) $n = 1, y = 1 \Rightarrow$ Beta$(3, 3)$

b) $n = 1, y = 1 \Rightarrow$ Beta$(4, 3)$

c) $n = 1, y = 0 \Rightarrow$ Beta$(4, 4)$

d) $n = 1, y = 1 \Rightarrow$ Beta$(5, 4)$

All at once: $n = 4, y = 3 \Rightarrow$ Beta$(2, 3) \mapsto$ Beta$(5, 4)$.

4.16

a) $n = 5, y = 3 \Rightarrow$ Beta$(5, 5)$

b) $n = 5, y = 1 \Rightarrow$ Beta$(6, 9)$

c) $n = 5, y = 1 \Rightarrow$ Beta$(7, 13)$

d) $n = 5, y = 2 \Rightarrow$ Beta$(9, 16)$

All at once: $n = 20, y = 7 \Rightarrow$ Beta$(2, 3) \mapsto$ Beta$(9, 16)$.

4.17 Prior: Beta$(4, 3)$, Data: 1) $n = 1, y = 0$, 2) $n = 10, y = 3$, 3) $n = 100, y = 20$

a) See R notebook for plots. The employees are quite uncertain with regards to what proportion of customers will click the ad, they vaguely expect that around 60% will click it.

b) First employee: Beta$(4, 4)$, second employee: Beta$(7, 10)$, third employee: Beta$(24, 83)$.

c) See R notebook.

d) The first employee's posterior is still very vague given the limited data, slightly shifting to a lower expected proportion. The second employee's posterior is more strongly influenced by the data (sample success ratio of 30%), moves to a mode of around 40% and is a bit less vague than the posterior. The third employee's posterior becomes very informative because of the strong evidence, the vague prior has only a minor influence, he settles on a mode of around 20%.

4.18

a)
- Day 0: Beta$(4, 3)$
- Day 1: $n = 1, y = 0 \Rightarrow$ Beta$(4, 4)$
- Day 2: $n = 10, y = 3 \Rightarrow$ Beta$(7, 11)$,
- Day 3: $n = 100, y = 20 \Rightarrow$ Beta$(27, 91)$,

b) The following are the modes and variances after each day:

- Day 0: Mode$[\pi] = 0.6$, SD$[\pi] \approx 0.17$
- Day 1: Mode$[\pi] = 0.5$, SD$[\pi] \approx 0.17$
- Day 2: Mode$[\pi] = 0.375$, SD$[\pi] \approx 0.11$
- Day 3: Mode$[\pi] \approx 0.22$, SD$[\pi] \approx 0.04$

After the first day the employee's understanding of π evolved only slightly and was still very vague, after the second day it became clear that the actual value of π will be lower than expected and after the third day it was very clear that a reasonable value for π lies somewhere around 22%.

c) The employee has collected $n = 111$ measurements with $y = 23$ successes. This promotes the Beta$(4, 3)$ prior to a Beta$(27, 91)$ posterior, similar to the one achieved on the third day of sequential updating.

4.19

a) Uniform Beta$(1,1)$ prior, $y=4, n=14 \Rightarrow$ Beta$(5,11)$ posterior, $E[\pi]\approx 0.31$, Mode$[\pi]\approx 0.29$

b) $y=6, n=15 \Rightarrow$ Beta$(11,20)$ posterior, $E[\pi]\approx 0.35$, Mode$[\pi]\approx 0.34$

c) $y=29, n=63 \Rightarrow$ Beta$(40,54)$ posterior, $E[\pi]\approx 0.43$, Mode$[\pi]\approx 0.42$

d) Uniform Beta$(1,1)$ prior, $y=39, n=92 \Rightarrow$ Beta$(40,54)$ posterior, $E[\pi]\approx 0.43$, Mode$[\pi]\approx 0.42$

4.20 In a frequentist approach for the previous exercises, we would estimate an underlying rate $\hat{\pi}$ with $\hat{\pi}_i = \frac{y_i}{n_i}$ for each sample i with size n_i and y_i positive outcomes. The different sample rate estimators could then be combined in a parallel manner (similar to exercise 4.17) with the formula for the combined mean:

$$\hat{\pi} = \frac{\sum\limits_i n_i\hat{\pi}_i}{\sum\limits_i n_i}. \tag{1}$$

It can easily be shown, that this is equivalent to the sample rate estimator we would get with union of all the samples:

$$\hat{\pi} = \frac{\sum\limits_i n_i \dfrac{y_i}{n_i}}{\sum\limits_i n_i} = \frac{\sum\limits_i y_i}{\sum\limits_i n_i}.$$

In a similar fashion, one could perform sequential updates. Given that we are currently 'believing' in a population rate $\hat{\pi}_N$ based on n_N measurements and collect n_{N+1} new measurements with y_{N+1} positive outcomes, we can update the rate according to (**??**) to

$$\hat{\pi}_{N+1} = \frac{n_N\hat{\pi}_N + y_{N+1}}{n_N + n_{N+1}}.$$

One can therefore also perform a sequential analysis with a frequentist approach and more data gives a tighter rate estimator $\hat{\pi}$ for the true underlying population rate π. While in the Bayesian approach a probability distribution for the rate π is estimated, in the frequentist approach only a point estimator $\hat{\pi}$ is recovered. In contrast to the Bayesian approach, prior information cannot be used in the frequentist approach and the method relies entirely on data. Given enough data, the Bayesian posterior mean and the frequentist estimator will converge to a common value.

5 Conjugate Families

5.1

a) Mode$[\lambda] = \frac{s-1}{r} = 4$, $E[\lambda] = \frac{s}{r} = 7 \;\Rightarrow \left|\begin{array}{ccc} s-1 & = & 4r \\ s & = & 7r \end{array}\right| \Rightarrow\; 3r = 1 \;\Rightarrow\; r = \frac{1}{3},\; s = \frac{7}{3}$

b) Mode$[\lambda] = \frac{s-1}{r} = 10$, $E[\lambda] = \frac{s}{r} = 12 \;\Rightarrow \left|\begin{array}{ccc} s-1 & = & 10r \\ s & = & 12r \end{array}\right| \Rightarrow\; 2r = 1 \;\Rightarrow\; r = \frac{1}{2},\; s = 6$

c) Mode$[\lambda] = \frac{s-1}{r} = 5$, Var$[\lambda] = \frac{s}{r^2} = 3 \;\Rightarrow\; \left|\begin{array}{ccc} s-1 & = & 5r \\ s & = & 3r^2 \end{array}\right| \;\Rightarrow\; 3r^2 - 5r - 1 = 0$

$\Rightarrow r = \frac{5+\sqrt{37}}{6} \approx 1.85,\; s = 5r + 1 \approx 10.24$

(only one positive solution for r)

d) Mode$[\lambda] = \frac{s-1}{r} = 14$, Var$[\lambda] = \frac{s}{r^2} = 6 \;\Rightarrow\; \left|\begin{array}{ccc} s-1 & = & 14r \\ s & = & 6r^2 \end{array}\right| \;\Rightarrow\; 6r^2 - 14r - 1 = 0$

$\Rightarrow r = \frac{14+\sqrt{220}}{12} \approx 2.40,\; s = 14r + 1 \approx 34.64$

(only one positive solution for r)

e) $E[\lambda] = \frac{s}{r} = 4$, Var$[\lambda] = \frac{s}{r^2} = 12 \;\Rightarrow\; \left|\begin{array}{ccc} s & = & 4r \\ s & = & 12r^2 \end{array}\right| \;\Rightarrow\; 4r = 12r^2 \overset{r>0}{\Rightarrow} r = \frac{1}{3},\; s = \frac{4}{3}$

f) $E[\lambda] = \frac{s}{r} = 22$, Var$[\lambda] = \frac{s}{r^2} = 3 \;\Rightarrow\; \left|\begin{array}{ccc} s & = & 22r \\ s & = & 3r^2 \end{array}\right| \;\Rightarrow\; 22r = 3r^2 \overset{r>0}{\Rightarrow} r = \frac{22}{3} \approx 7.33,\; s = \frac{22^2}{3} \approx$ 161.33

5.2 (for plots see R notebook)

a) $\sum_i y_i = 29, n = 3 \;\Rightarrow\; L(\lambda|\vec{y}) = \dfrac{\lambda^{29} e^{-3\lambda}}{3!\; 7!\; 19!}$

b) $\sum_i y_i = 36, n = 4 \;\Rightarrow\; L(\lambda|\vec{y}) = \dfrac{\lambda^{36} e^{-4\lambda}}{12!\; 12!\; 12!\; 0!}$

c) $\sum_i y_i = 12, n = 1 \;\Rightarrow\; L(\lambda|\vec{y}) = \dfrac{\lambda^{12} e^{-\lambda}}{12!}$

d) $\sum_i y_i = 65, n = 5 \;\Rightarrow\; L(\lambda|\vec{y}) = \dfrac{\lambda^{65} e^{-5\lambda}}{16!\; 10!\; 17!\; 11!\; 11!}$

5.3

a) $f(\lambda|\vec{y}) =$ Gamma$(24 + 29, 2 + 3) =$ Gamma$(53, 5)$

b) $f(\lambda|\vec{y}) =$ Gamma$(24 + 36, 2 + 4) =$ Gamma$(60, 6)$

c) $f(\lambda|\vec{y}) =$ Gamma$(24 + 12, 2 + 1) =$ Gamma$(36, 3)$

d) $f(\lambda|\vec{y}) =$ Gamma$(24 + 65, 2 + 5) =$ Gamma$(89, 7)$

5.4

a) $f(\lambda|\vec{y}) =$ Gamma$(2 + 29, 2 + 3) =$ Gamma$(31, 5)$

b) $f(\lambda|\vec{y}) =$ Gamma$(2 + 36, 2 + 4) =$ Gamma$(38, 6)$

c) $f(\lambda|\vec{y}) =$ Gamma$(2 + 12, 2 + 1) =$ Gamma$(14, 3)$

d) $f(\lambda|\vec{y}) =$ Gamma$(2 + 65, 2 + 5) =$ Gamma$(67, 7)$

5.5

a) $E[\lambda] = \frac{s}{r} = 5, \text{Var}[\lambda] = \frac{s}{r^2} = 0.25^2 \ \Rightarrow r = \frac{5}{0.25^2} = 80, \ s = 5r = 400$

b) `pgamma(q, s, r)` computes the cumulative distribution function $P(\lambda \leq q)$ of a Gamma(s, r) pdf at position q, i.e. it integrates the density function from zero to q. The prior probability that the rate of text messages λ per hour is larger than 10 is $P(\lambda > 10) = 1 - P(\lambda \leq 10)$. The probability $P(\lambda \leq 10)$ is so small that `1 - pgamma(10, 400, 80)` evaluates to exactly zero (however this is the effect of numeric precision).

5.6 (see R notebook for a)-c))

d) Because the prior was chosen so narrow, the data was not enough to draw away the posterior by a large amount, more evidence is needed.

5.7

a) The prior places high probability on very low numbers of goals. A mean of 4 goals is expected, however most often a lower number of goals is expected to be scored, most often no goals at all. More than 10 goals are already very unlikely (see R notebook for plots).

b) The number of events (goals) in a fixed timespan is modeled.

c) It seems that the data suggests a different distribution than the prior. Only very rarely are no goals shot and more than 4 goals are also rarely scored.

d) Posterior model: Gamma$(s + \sum_i y_i, r + n) =$ Gamma$(1 + 146, 0.25 + 52) =$ Gamma$(147, 52.25)$

e) The posterior is more localized than the prior and weight is moved from the left side (zero goals) towards the center (1,2 or 3 goals).

5.8 (See R notebook for plots)

Given the known variance σ, the number of measured data points n and the average value $\bar{y}$ of the data points, the likelihood can according to the book be expressed as

$$L(\mu|\vec{y}) \propto \text{exp}\left[-\frac{(\bar{y} - \mu)^2}{2\sigma^2/n}\right].$$

a) $n = 3, \sigma = 10, \bar{y} \approx 7.67 \ \Rightarrow \ L(\mu|\vec{y}) \propto\approx \text{exp}\left[-\frac{(7.67-\mu)^2}{66.67}\right]$

b) $n = 4, \sigma = 6, \bar{y} = -3.675 \ \Rightarrow \ L(\mu|\vec{y}) \propto\approx \text{exp}\left[-\frac{(-3.675-\mu)^2}{18}\right]$

c) $n = 2, \sigma = 5, \bar{y} = 9.25 \ \Rightarrow \ L(\mu|\vec{y}) \propto\approx \text{exp}\left[-\frac{(9.25-\mu)^2}{25}\right]$

d) $n = 5, \sigma = 0.6, \bar{y} = 1.118 \ \Rightarrow \ L(\mu|\vec{y}) \propto\approx \text{exp}\left[-\frac{(1.118-\mu)^2}{0.144}\right]$

5.9 (See R notebook for plots)

a) $\mu \sim N(\theta = 7.2, \tau = 2.6)$

b) Yes.

c) Yes, however slightly less likely than 7.6 dollars.

d) That the stock price goes down on average ($\mu < 0$) is highly unlikely with $P(\mu = 0) \approx 0.0028 = 0.28\%$.

e) $P(\mu > 8) = 1 - P(\mu \leq 8) \approx 0.379 = 37.9\%$.

5.10 (See R notebook for plots)

d) Since the prior is pretty vague, the posterior is drawn more to the likelihood. Before the data came in, we expected the stock to go up by a mean of 7.2 dollars, with nonzero probabilities between $\sim 0 - 15$ dollars. The posterior mean is however with 1.15 dollars much lower, and the standard deviation is approximately one third of before. A negative average change (stock goes down on average) is now also likely. Average changes larger than ~ 5 are not expected anymore.

e) $P(\mu = 0) \approx 0.100 = 10.0\%$.

f) $P(\mu > 8) = 1 - P(\mu \leq 8) \approx 10^{-13}$ (practically zero).

5.11 (See R notebook for plots)

Prior on grade point average: $\mu \sim N(\theta = 80, \tau = 4)$.

Known standard deviation of student scores: $\sigma = 3$.

a) $n_1 = 32, \bar{y}_1 = 86$

$\mu|\vec{y}_1 \sim N\left(\theta \, \frac{\sigma^2}{n_1\tau^2+\sigma^2} + \bar{y}_1 \, \frac{n_1\tau^2}{n_1\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n_1\tau^2+\sigma^2}\right) \approx N\left(85.90, 0.53\right)$

b) $n_2 = 32, \bar{y}_1 = 82$

$\mu|\vec{y}_2 \sim N\left(\theta \, \frac{\sigma^2}{n_2\tau^2+\sigma^2} + \bar{y}_2 \, \frac{n_2\tau^2}{n_2\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n_2\tau^2+\sigma^2}\right) \approx N\left(81.97, 0.53\right)$

c) Variant 1: Sequentially

- Posterior values after first results: $\theta_1 = 85.90$, $\tau_1 = 0.53$
- Update after second results:

  $\mu|\vec{y} \sim N\left(\theta_1 \, \frac{\sigma^2}{n_2\tau_1^2+\sigma^2} + \bar{y}_2 \, \frac{n_2\tau_1^2}{n_2\tau_1^2+\sigma^2}, \frac{\tau_1^2\sigma^2}{n_2\tau_1^2+\sigma^2}\right) \approx N\left(83.97, 0.37\right)$

Variant 2: Combined mean

- $\bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2}{n1+n2}$, $n = n_1 + n_2$
- $\mu|\vec{y} \sim N\left(\theta \, \frac{\sigma^2}{n\tau^2+\sigma^2} + \bar{y} \, \frac{n\tau^2}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right) \approx N\left(83.97, 0.37\right)$

5.12 (See R notebook for plots)

Prior on hippocampal volume: $\mu \sim N(\theta = 0.5, \tau = 0.4)$.

Known standard deviation of individual volumes: $\sigma = 0.5$.

a) $n = 25, \bar{y} \approx 7.60$

b) Posterior model: $\mu | \vec{y} \sim N\left(\theta \frac{\sigma^2}{n\tau^2+\sigma^2} + \bar{y} \frac{n\tau^2}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right) \approx N(7.54, 0.097)$

c) The posterior is almost equivalent to the likelihood and far away from the prior (prior mean 6.5, posterior mean 7.5). In particular, there is practically no overlap between prior and posterior understanding of reasonable values of average hippocampal volume in the control group μ.

5.13 (See R notebook for plots)

a) Assuming a normal distribution of μ around $\theta = 30$ with standard deviation $\tau = 15$. Choosing $\tau = 20$ seems a bit too wide, allowing for temperatures beyond 60° Celsius.

b) The distribution of temperatures is not perfectly normal and a bit right-skewed.

c) Posterior model: $\mu | \vec{y} \sim N\left(\theta \frac{\sigma^2}{n\tau^2+\sigma^2} + \bar{y} \frac{n\tau^2}{n\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right) \approx N(24.00, 0.02)$

d) Because there is so much data available (around 2.5 years), the likelihood is extremely convincing and the posterior has been drawn to practically be the likelihood, concentrating around extremely narrow values for the average temperature around $24.00 \pm 0.02°$ C.

5.14

a) If we would proceed in a similar way as for binomially distributed quantities, we would first sample e.g. 10'000 candidates for μ from the prior distribution $N(\mu;\ \theta, \tau)$ and then for each particular value of μ sample a set of values (here our strategy becomes already unclear, before we only sampled one value) for Y from $N(Y;\ \mu, \sigma)$ and compute the mean $\bar{y}$. We could then assess the posterior probability distribution for μ given a particular value of $\bar{y} = \bar{y}_0$ by selecting only sampled $\bar{y}$ equal to $\bar{y}_0$. However here we fail, because the numbers are continuous and we will most certainly not find a sampled value in our continuous data that exactly matches $\bar{y}_0$.

b) See R notebook.

5.15

a) Binomial($\theta;\ \pi = 0.3,\ n = 16$)

b) Poisson($\theta;\ \lambda = 1$)

c) Beta($\theta;\ \alpha = 5,\ \beta = 8$)

d) Normal($\theta;\ \mu = 0,\ \sigma^2 = 1$)

5.16

a) Gamma($\theta;\ s = 16,\ r = 2$)

b) Normal($\theta;\ \mu = 12,\ \sigma^2 = 18$)

c) Poisson($\theta;\ \lambda = 0.3$)

5.17 For the Gamma-Poisson conjugate family we have $E[\lambda] = \frac{s}{r}$ (prior mean) and $E[\lambda|y] = \frac{s+\sum_i y_i}{r+n} = \frac{s+n\bar{y}}{r+n}$ (posterior mean) for a Gamma(s, r) prior distribution and n Poisson-distributed measurements y_i.

Our goal is to provide a relation of the form

$$E[\lambda|y] = \alpha\, E[\lambda] + (1-\alpha)\,\bar{y},$$

where α is a relative weight parameter. Plugging in the above expressions leads to

$$\frac{s+n\bar{y}}{r+n} = \alpha\frac{s}{r} + \bar{y} - \alpha\bar{y} \;\Rightarrow\; \alpha = \frac{\frac{s+n\bar{y}}{r+n} - \bar{y}}{\frac{s}{r} - \bar{y}} = \frac{r}{r+n},\; 1-\alpha = \frac{n}{r+n},$$

and thus

$$E[\lambda|y] = \frac{r}{r+n}\,E[\lambda] + \frac{n}{r+n}\,\bar{y}.$$

5.18

a) Since we deal with the count of the number of events in a fixed timespan, a Poisson distribution should be used.

b) Prior distribution parameters:

$E[\lambda] = \frac{s}{r} = \frac{1}{2}$, Var$[\lambda] = 0.25^2 = \frac{s}{r^2}$ $\Rightarrow s = \frac{r}{2}$, $\frac{1}{2r} = 0.25^2$ $\Rightarrow r = 8, s = 4$

$\Rightarrow \lambda \sim$ Gamma$(4, 8)$

Posterior model (see R notebook): Gamma$(17, 28)$.

The posterior is more in sync with the likelihood than with the prior. This might make sense because the sample mean of $\bar{y} = 0.65$ is already towards the tail of the prior and $n = 20$ is a large number compared with the vague prior.

This can also be seen when considering the weights given in Exercise 5.17:

$$\alpha = \frac{r}{r+n} = \frac{8}{8+20} \approx 0.29, \quad 1-\alpha = \frac{n}{r+n} = \frac{20}{8+20} \approx 0.71.$$

$n = 20$ datapoints are a considerable amount compared with $r = 8$.

c) $E[\lambda] \approx 0.61$, SD$[\lambda] \approx 0.15$

d) The initial assumption of 0.5 ± 0.25 insects per square meter can now be refined to 0.61 ± 0.15 insects per square meter given the measured data. The mean has shifted towards a larger value than initially expected and the standard deviation as a measure of uncertainty has been almost cut in half.

5.19

a) Prior: $\theta \sim$ Beta(α, β) with $f(\theta) \propto \theta^{\alpha-1}\,(1-\theta)^{\beta-1}$

Likelihood: $Y|\theta \sim$ Geometric(θ) with $f(y|\theta) \propto \theta\,(1-\theta)^{y-1}$

The geometric distribution is related to the binomial distribution and computes the probability of the event that the k-th trial of a Bernoulli experiment with individual probability θ is the first success.

Posterior: $f(\theta|y) \propto f(y|\theta)\,f(\theta) \propto \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,\theta\,(1-\theta)^{y-1} = \theta^{\alpha}\,(1-\theta)^{\beta+y-2}$.

The posterior follows thus a Beta$(\alpha+1, \beta+y-1)$ distribution.

b) By definition of conjugate distributions (likelihood distribution is such tahat prior and posterior follow the same distribution), the Beta and Geometric distribution are conjugated.

6 Approximating the Posterior

6.1

a)
1. Create a grid over the quantities in question.
2. Evaluate prior and likelihood over all points of the grid.
3. Compute unnormalized posterior by multiplying prior and posterior on each grid point
4. Divide the unnormalized posterior by the sum of all values at all grid points to get the normalized posterior.

The approximation becomes more accurate if the density of the grid is increased. The density can be increased by increasing the number of grid points along each dimension.

6.2 (not sketched, but described and pointing to figures in the book)

a) Chain is mixing too slowly: There is strong correlation among the chain values and the chain does not manage to explore the full parameter space. This results in a bad approximation of the posterior outside the explored range (as in top of Fig. 6.12 in the book).

b) Chain has high correlation: The chain does not look like white noise, there is a considerable correlation between ~10-20 or more consecutive lags (top of Fig. 6.12 and Fig. 6.16 in the book).

c) Chain has tendency to get stuck: Trace plot contains parts that are constant or almost constant for a while (bottom of Fig. 6.12 in the book).

d) Chain has no problems: Trace plot looks like uncorrelated white noise (e.g. Fig. 6.15 in the book).

6.3

a) Chain is mixing too slowly: Not the full posterior might be explored, the approximation might yield zero or close to zero probabilities at locations where the posterior distribution is non-zero.

b) Chain has high correlation: The chain could spend significantly more time in certain regions and consequently it could take much more steps for it to reach a good smooth approximation of the posterior.

c) Chain has tendency to get stuck: The approximation of the posterior distribution could contain artificial peaks or modes at parameters where the chain got stuck - these peaks do not reflect the true posterior.

6.4

a) In a realistic setting (no conjugated prior), we have no way of checking directly, whether the chain has converged to a state where it can reasonably approximate the posterior. MCMC diagnostics show us indicators for whether the chain is mixing quickly (good approximation) or slowly (approximation might be bad.

b) If prior and likelihood are not conjugated, the posterior distribution cannot be computed analytically. Since the grid approximation suffers from the curse of dimensionality, a method that explores the parameter space more intelligently is required: MCMC.

c) RStan needs to know only prior and likelihood and then takes care about everything else! (see next chapter)

d) (up to you)

6.5 - **6.7** see R notebook

6.8

a) Examples: healthcare data (many indicators for one person), genomics (e.g. degree of association of particular gene expressions with observed body properties), further partition of estimated parameters (e.g. infection rate per country or hourly number of cars passing by), facial recognition (e.g. probability for each person), . . .

b) Say that we want to estimate a proportion in $[0, 1]$. With 10 grid points, the grid points are 0.1 apart from each other. When estimating a proportion in $[0, 1] \times [0, 1]$ (a square), we need 100 grid points to get the same density, i.e. grid points that are horizontally and vertically 0.1 apart from each other. To achieve the same density, 1000 grid points are then needed for a cube and in general 10^n for n-dimensional data. For e.g. a 200-dimensional space (which is not unsual), this would be 10^{200}, an infeasible number of measurements to collect. Thus the number of datapoints needed to get a sufficient density of grid points (and 10 grid points is not that much) grows exponentially with the number of dimensions.

6.9

a) Compared to the closed form posterior, both are inaccurate and inefficient, i.e. they will only approximate the true posterior to a certain extent (which is hard to judge since when these methods are needed, the true posterior cannot be calculated in closed form) and they require much more effort to compute than the closed form solution (estimation of a large number of grid points / computing of a large number of chain values).

b) They are both available even when a closed form solution for the posterior is not available.

c) It is simple and well-understood. If we have used enough grid points, then we know that we have a good approximation of the posterior that we can sample from.

d) The grid approximation becomes too inefficient when trying to explore a high-dimensional configuration space. MCMC suffers less from this limitation.

6.10 (very original exercise - I am not completely sure whether I solved it correctly - up for discussion!)

A Markov chain has the important property, that the probability of an event E_n following a 'chain' of events E_i only depends on the event just before:

$$p(E_n \mid E_1, E_2, \ldots, E_{n-1}) = p(E_n \mid E_{n-1}),$$

i.e. all the knowledge needed to compute $p(E_n)$ is given by knowing that E_{n-1} occurred. Thus, the question to ask for each exercise is: 'Does it suffice to know θ_{i-1} to compute $p(\theta_i)$'?

a) How likely it is that I go to a Thai restaurant is probably influenced how much I consider it before. I most likely will not go to a Thai two days in a row and my mood for Thai might grow higher the longer I have not eaten Thai. However there is probably much more occuring through my day (and probably with the mood of my peers if it is a joint dinner) than I could have known the day before, so even though it might be a Markov chain to some extent, I think it's in general not.

b) This is clearly not a Markov chain. Today's chance of winning the lottery is not influenced by yesterday's chance, thus the chain does not have the Markov property.

c) This seems to be a Markov chain. I will probably update my win probability after each game, such that all the games are contained in the previous win probability θ_{n-1}.

6.11 - 6.18 see R notebook

7 MCMC under the hood

7.1

a) Step 1: Sample μ from posterior $f(\mu|y)$

Step 2: Go there (add this value to the chain or - in the jargon of the book - tour)

b) pro: very easy to implement

con: suffers from curse of high dimensionality - very inefficient in high dimensional setting

7.2

a)

Step 1: Given current position μ, sample new position μ' from proposal distribution $q(\mu'|\mu)$.

Step 2: Decide whether to go to this position - go there with probability

$$\alpha = \min\left(1, \frac{f(\mu')\,L(\mu'|y)}{f(\mu)\,L(\mu|y)}\;\frac{q(\mu|\mu')}{q(\mu'|\mu)}\right).$$

b) In contrast to Monte Carlo, in the Metropolis-Hastings algorithm, a proposed step can be rejected.

c) The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm when the probabilities $P(\mu \to \mu')$ and $P(\mu' \to \mu)$ are equal, i.e. q is symmetric with $q(\mu'|\mu) = q(\mu|\mu')$. Then the acceptance rule simplifies to

$$\alpha = \min\left(1, \frac{f(\mu')\,L(\mu'|y)}{f(\mu)\,L(\mu|y)}\right),$$

i.e jump there at once if μ' corresponds to higher posterior plausibility and otherwise jump there with a probability α equal to the ratio of posterior plausibilities.

7.3 b, c

7.4

a) See Fig. 7.8 in the book for an example. In general, it produces a curve that is slowly increasing or decreasing.

b) If w is too small, the chain is very slowly climbing up the nearest peak and will then probably stay very close to this peak once it reached it. Consequently, the chain produces highly correlated samples. Besides only exploring a very small part of the distribution, the chain will never be able to jump to other modes of the probability distribution.

c) See Fig. 7.8 in the book for an example. In general, the trace plot will contain many sections that are flat, i.e. where the chain stays at the same location for some time.

d) The proposed locations will often lie far away from the modes of the probability distribution and will be rejected with high probability. Thus the chain will stay at the same location for a long time - and therefore produces highly correlated (the same) samples.

e) See Fig. 7.8 in the book for an example. In general, the trace plot will look like white noise.

f) In the previous book chapter, diagnostics were explored to distinguish chains that are mixing slowly (produce highly correlated samples) from chains that are mixing quickly (produce samples with low correlations). One could look at such diagnostics (effective sample size ratio, autocorrelation, R-hat) or also inspect the trace plot and the resulting histogram visually to decide whether the chosen value of w is reasonable.

7.5

a) False. Per definition, independence sampling proposes a new candidate position independent from the current value.

b) True, since the proposal model is not conditioned on any current value.

c) True, it is a special case of the Metropolis-Hastings algorithm with a proposal distribution $q(x'|x) = q(x')$.

d) In general false, because this requires $q(x') = q(x)$ for any pair (x, x'). This implies that $q(x)$ is constant. This is only the case for a uniform proposal distribution.

7.6 (see R notebook)

7.7 $\lambda = 2$, $\lambda' = 2.1$

a)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{\lambda^2}{\lambda'^2}\,\frac{e^{\frac{(\lambda-\lambda')^2}{2\cdot 1^2}}}{e^{\frac{(\lambda'-\lambda)^2}{2\cdot 1^2}}} = \frac{\lambda^2}{\lambda'^2} \approx 0.907$$

Jump to $\lambda' = 2.1$ with a probability of 90.7% (symmetric proposal distribution).

b)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{\lambda'}}{e^{\lambda}}\,\frac{e^{\frac{(\lambda-\lambda')^2}{2\cdot 0.2^2}}}{e^{\frac{(\lambda'-\lambda)^2}{2\cdot 0.2^2}}} = \frac{e^{\lambda'}}{e^{\lambda}} \approx 1.105$$

Jump to $\lambda' = 2.1$ with a probability of 100% (symmetric proposal distribution).

c)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{-10\lambda'}}{e^{-10\lambda}}\,\frac{\frac{1}{2\cdot 0.5}}{\frac{1}{2\cdot 0.5}} = \frac{e^{-10\lambda'}}{e^{-10\lambda}} \approx 0.368$$

Jump to $\lambda' = 2.1$ with a probability of 36.8% (symmetric proposal distribution).

d)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{-\lambda'^4}}{e^{-\lambda^4}}\,\frac{\lambda'\,e^{-\lambda'\cdot\lambda}}{\lambda\,e^{-\lambda\cdot\lambda'}} = \frac{e^{-\lambda'^4}}{e^{-\lambda^4}}\,\frac{\lambda'}{\lambda} \approx 0.033$$

Jump to $\lambda' = 2.1$ with a probability of 3.3%.

e) Only in scenario b). Since the posterior distribution is proportional to an exponential function, we want to go to higher values of λ (and consequently to higher plausibilities) as quickly as possible (see corresponding plots in R notebook).

7.8 $\lambda = 1.8$, $\lambda' = 1.6$

a)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{\lambda}{\lambda'}\,\frac{e^{\frac{(\lambda-\lambda')^2}{2\cdot 3^2}}}{e^{\frac{(\lambda'-\lambda)^2}{2\cdot 3^2}}} = \frac{\lambda}{\lambda'} \approx 1.125$$

Jump to $\lambda' = 1.6$ with a probability of 100% (symmetric proposal distribution).

b)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{3\lambda'}}{e^{3\lambda}}\,\frac{e^{\frac{(\lambda-\lambda')^2}{2\cdot 0.5^2}}}{e^{\frac{(\lambda'-\lambda)^2}{2\cdot 0.5^2}}} = \frac{e^{3\lambda'}}{e^{3\lambda}} \approx 0.549$$

Jump to $\lambda' = 1.6$ with a probability of 54.9% (symmetric proposal distribution).

c)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{-1.9\lambda'}}{e^{-1.9\lambda}}\,\frac{\frac{1}{2\cdot 0.3}}{\frac{1}{2\cdot 0.3}} = \frac{e^{-1.9\lambda'}}{e^{-1.9\lambda}} \approx 1.462$$

Jump to $\lambda' = 1.6$ with a probability of 100% (symmetric proposal distribution).

d)

$$\frac{f(\lambda')\,L(\lambda'|y)}{f(\lambda)\,L(\lambda|y)}\,\frac{q(\lambda|\lambda')}{q(\lambda'|\lambda)} = \frac{e^{-\lambda'^4}}{e^{-\lambda^4}}\,\frac{\lambda'\,e^{-\lambda'\cdot\lambda}}{\lambda\,e^{-\lambda\cdot\lambda'}} = \frac{e^{-\lambda'^4}}{e^{-\lambda^4}}\,\frac{\lambda'}{\lambda} \approx 45.889$$

Jump to $\lambda' = 1.6$ with a probability of 100%.

e) In scenarios a), c) and d). In a) the posterior distribution is proportional to an inverse power function, diverging to infinity at $\lambda = 0$. Consequently, we want to move to smaller values as quickly as possible. Similarly, in the inverse exponential function in c), one will want to move left as fast as possible. In d), the bulk of probability mass lies between -1 and 1, thus we want to move from 1.8 to lower values as fast as possible.

7.9-7.14 See R notebook.

7.15-7.17 Provide and explore your own examples.

8 Posterior Inference & Prediction

8.1

- Estimation: Given a prior and measured data, what are likely values for model parameters? (e.g. using credible intervals)
- Hypothesis testing: How does the belief in different hypotheses change in the light of measured data?
- Prediction: Given models updated by measured data, what are likely outcomes for new measurements?

8.2

a) Unknown uncertainty (variance and skewness)

b) λ lies between 1 and 3.4 with a probability of 95%. Remember that probability represents a degree of belief and should not be interpreted in a frequentist way.

8.3

a) One-sided hypothesis test: p: proportion of dogs in the park without a licence, H_0: $p \leq 0.4$ (null hypothesis), H_1: $p > 0.4$ (alternative hypothesis)

b) This is an estimation problem and not related to hypothesis testing.

c) One-sided hypothesis test: p: proportion of voters who support the new regulation, H_0: $p \leq 0.6$ (null hypothesis), H_1: $p > 0.6$ (alternative hypothesis).

d) This could be formulated as a two-sided hypothesis test:

   k: number of times this particular mode of argument is used on the page, H_0: $k = 3$ (null hypothesis), H_1: $k \neq 3$ (alternative hypothesis).

   However, this is probably not what Sarah is interested in (k being exactly 3 or different) and probably more an estimation problem. Given her prior (3 times per page) and her data from 90 pages, Sarah is interested in what number of times the mode of argument actually turns up (i.e. a posterior credible interval for values of k).

While it is extremely important in frequentist statistics which is the null hypothesis (H_0) and which the alternative one, this does not matter in the Bayesian approach, they are just two different hypotheses.

8.4

a) Posterior odds: ratio for posterior probabilities of two hypotheses given the measured data

b) Prior odds: ratio of prior probabilities of two hypotheses

c) Bayes factor: ratio or posterior odds to prior odds. By what factor did the measured data increase our understanding of how much more likely one hypothesis is than another? The Bayes factor thus quantifies the impact of the given data on our belief.

8.5

a)
- Sampling variability: The actual value of the outcome of the sample (e.g. number of positive events in the Binomial case, number of occuring events in the Poisson case, mean and standard deviation of the sample in the Normal case) depends on the particular values sampled (sampling noise).
- Posterior variability: The sampling variability applies for one single model (i.e. for a fixed set of model parameters). However, we face uncertainty in the estimation of these parameters and are not exactly sure which is the best model. This introduces another source of variance.

These concepts are similar to prediction intervals in frequentist prediction where there is a confidence interval for model parameters (e.g. regression coefficients) and sampling variability has to be added on top (e.g. since linear regression only predicts the expectation).

b) Let us assume that I want to plan the amount of spaghetti needed for a snow camp that takes place every year. For the initial snow camp I had to come up with a reasonable and safe prior, overestimating the amount of spaghetti needed per person. However, already after the first snow camp I had some data about average spaghetti consumption and could update my prior to a more sensible posterior. This continued over the years and I could improve my posterior model more and more, estimating the mean spaghetti consumption per head μ and its standard deviation σ with some uncertainty due to the limited amount of available data (posterior variability). This does not mean that I know the perfect amount of spaghetti needed per person now, since there is an inherent sampling variability left: What sample of people comes to the snowcamp? Are many of them people who generally eat a lot or are many of them generally happy with less? How was the day? Was it strenuous such that people are hungrier? Is the sauce particularly delicious such that people eat more?

c) A posterior predictive model is only conditional on the data, since its total uncertainty is computed by integration over all posterior models (using the full domain of model parameters) weighted by their plausibility. This can be understood when looking at equation (8.4) of the book.

8.6 (see R notebook)

a) $\pi \in [0.16, 0.75]$ b) $\pi \in [0.31, 0.58]$ c) $\lambda \in [0.0032, 0.5836]$ (rounded to nearest values within interval)

8.7 (see R notebook)

a) $\lambda \in [0.0011, 1.0596]$ b) $\mu \in [6.1, 13.9]$ c) $\mu \in [-4.3, -1.7]$ (rounded to nearest values within interval)

8.8 (see R notebook)

a) $\lambda \in [0, 0.599]$ (rounded to nearest values within interval)

b) $\lambda \in [0.005, 0.737]$ (rounded to nearest values within interval)

c) Especially at the upper end the intervals are quite different. Since there is more probability mass located around $\lambda = 0$, the highest posterior density credible interval (HPDCI) extends less into the low-probability tail. Because there is so much more probability mass at $\lambda = 0$, the HPDCI represents the uncertainty in λ better.

d) $\mu \in [-16.92, 9.08]$

e) $\mu \in [-16.92, 9.08]$

f) The intervals are the same, since the normal distribution is symmetric.

In general it is not straightforward to compute the HPDCI. However in the two example cases of this exercise the computation is simple: The Gamma(1,5) distribution diverges at $\lambda = 0$ and it is clear that the HPDCI needs to start at $\lambda = 0$. Since the normal distribution is symmetric, it can easily be argued that the HPDCI and the middle posterior density credible interval are the same.

8.9 (see R notebook)

prior: $\pi \sim$ Beta$(1, 0.8)$, posterior: $\pi|y \sim$ Beta$(4, 3)$

H_0: $\pi \leq 0.4$ (null hypothesis), H_a: $\pi > 0.4$ (alternative hypothesis)

a) $P(H_a|y) = \int_{0.4}^{1} \text{Beta}(\pi; 4, 3)\ \text{d}\pi = 1 - \int_{0}^{0.4} \text{Beta}(\pi; 4, 3)\ \text{d}\pi = 0.8208$

b) posterior odds $= \dfrac{P(H_a|y)}{P(H_0|y)} = \dfrac{0.8208}{0.1792} \approx 4.58$

It is 4.58 times more plausible that the alternative hypothesis is true after the data y were measured.

c) prior odds $= \dfrac{P(H_a)}{P(H_0)} \approx \dfrac{0.665}{0.335} \approx 1.98$

Before the data y were measured, it was 1.98 times more plausible that the alternative hypothesis is true.

d) BF $= \dfrac{\text{posterior odds}}{\text{prior odds}} \approx \dfrac{4.58}{1.98} \approx 2.31$

The plausibility of H_a has increased by a factor of 2.31 in the light of the measured data.

e) Before the data we believed that the alternative hyptothesis was twice as likely as the null hypothesis. Now, after we measured some data, we believe that the alternative hypothesis is around 4.6 times as likely as the null hypothesis. The presence of the data has increased the odds by a factor of 2.31.

8.10 (see R notebook)

prior: $\mu \sim N(10, 10^2)$, posterior: $\mu|\vec{y} \sim N(5, 3^2)$

H_0: $\mu \geq 5.2$ (null hypothesis), H_a: $\mu < 5.2$ (alternative hypothesis)

a) $P(H_a|\vec{y}) = \int_0^{5.2} N(\mu; 5, 3^2)\,\mathrm{d}\mu \approx 0.527$

b) posterior odds $= \dfrac{P(H_a|\vec{y})}{P(H_0|\vec{y})} = \dfrac{0.527}{0.473} \approx 1.11$

It is 1.11 times more plausible that the alternative hypothesis is true after the data $\vec{y}$ were measured.

c) prior odds $= \dfrac{P(H_a)}{P(H_0)} \approx \dfrac{0.316}{0.684} \approx 0.46$

Before the data $\vec{y}$ were measured, it was 0.46 times more plausible that the alternative hypothesis is true.

d) BF $= \dfrac{\text{posterior odds}}{\text{prior odds}} \approx \dfrac{1.11}{0.46} \approx 2.42$

The plausibility of H_a has increased by a factor of 2.42 in the light of the measured data.

e) Before the data we believed that the alternative hyptothesis was around half as likely as the null hypothesis. Now, after we measured some data, we believe that the alternative hypothesis is equally likely as the null hypothesis (or even a bit preferred). The presence of the data has increased the odds by a factor of 2.42.

8.11

a) $\pi|y \sim \text{Beta}(\alpha + y, \beta + n - y) \quad \Rightarrow \quad f(\pi|y) = \dfrac{\Gamma[\alpha+\beta+n]}{\Gamma[\alpha+y]\,\Gamma[\beta+n-y]}\,\pi^{\alpha+y-1}\,(1-\pi)^{\beta+n-y-1}$

b) $y'|\pi \sim \text{Bin}(n', \pi) \quad \Rightarrow \quad f(y'|\pi) = \dbinom{n'}{y'}\,\pi^{y'}\,(1-\pi)^{n'-y'}$

c)

$$\begin{aligned}
f(y'|y) &= \int_0^1 f(y'|\pi)\,f(\pi|y)\,\mathrm{d}\pi \\
&= \int_0^1 \binom{n'}{y'}\,\pi^{y'}\,(1-\pi)^{n'-y'}\,\frac{\Gamma[\alpha+\beta+n]}{\Gamma[\alpha+y]\,\Gamma[\beta+n-y]}\,\pi^{\alpha+y-1}\,(1-\pi)^{\beta+n-y-1}\mathrm{d}\pi \\
&= \binom{n'}{y'}\,\frac{\Gamma[\alpha+\beta+n]}{\Gamma[\alpha+y]\,\Gamma[\beta+n-y]}\int_0^1 \pi^{\alpha+y+y'-1}\,(1-\pi)^{\beta+n-y+n'-y'-1}\,\mathrm{d}\pi \\
&= \binom{n'}{y'}\,\frac{\Gamma[\alpha+\beta+n]}{\Gamma[\alpha+y]\,\Gamma[\beta+n-y]}\,\frac{\Gamma[\alpha+y+y']\,\Gamma[\beta+n-y+n'-y']}{\Gamma[\alpha+\beta+n+n']},
\end{aligned}$$

where the property $\displaystyle\int_0^1 \pi^{\alpha-1}\,(1-\pi)^{\beta-1}\,\mathrm{d}\pi = \frac{\Gamma[\alpha]\,\Gamma[\beta]}{\Gamma[\alpha+\beta]}$ has been used.

d) $f(y'|y=14) = \dbinom{20}{y'}\,\dfrac{\Gamma[110]}{\Gamma[18]\,\Gamma[92]}\,\dfrac{\Gamma[18+y']\,\Gamma[112-y']}{\Gamma[130]}$

e) $f(y'|y=14) = \dbinom{4}{y'}\,\dfrac{\Gamma[110]}{\Gamma[18]\,\Gamma[92]}\,\dfrac{\Gamma[18+y']\,\Gamma[96-y']}{\Gamma[114]}$

For a sketch see R notebook.

8.12

- $y|\lambda \sim$ Pois(λ) (y events in a fixed time interval) with $f(y|\lambda) = \dfrac{\lambda^y}{y!}\, e^{-\lambda}$
- $\lambda \sim$ Gamma(s, r) with $f(\lambda) = \dfrac{r^s}{\Gamma[s]}\, \lambda^{s-1}\, e^{-r\lambda}$ for $\lambda > 0$

a) $\lambda|y \sim$ Gamma$(s+y, r+1) \quad \Rightarrow \quad f(\lambda|y) = \dfrac{(r+1)^{s+y}}{\Gamma[s+y]}\, \lambda^{s+y-1}\, e^{-(r+1)\,\lambda}$

b) $y'|\lambda \sim$ Pois$(\lambda) \quad \Rightarrow \quad f(y'|\lambda) = \dfrac{\lambda^{y'}}{y'!}\, e^{-\lambda}$

c)

$$\begin{aligned} f(y'|y) &= \int_0^\infty f(y'|\lambda)\, f(\lambda|y)\, \mathsf{d}\lambda \\ &= \int_0^\infty \frac{\lambda^{y'}}{y'!}\, e^{-\lambda}\, \frac{(r+1)^{s+y}}{\Gamma[s+y]}\, \lambda^{s+y-1}\, e^{-(r+1)\,\lambda}\, \mathsf{d}\lambda \\ &= \frac{1}{y'!}\, \frac{(r+1)^{s+y}}{\Gamma[s+y]} \int_0^\infty \lambda^{s+y+y'-1}\, e^{-(r+2)\,\lambda} \mathsf{d}\lambda \\ &= \frac{1}{y'!}\, \frac{\Gamma[s+y+y']}{\Gamma[s+y]}\, \frac{(r+1)^{s+y}}{(r+2)^{s+y+y'}}, \end{aligned}$$

where the property $\int_0^\infty \lambda^{s'-1}\, e^{-r'\lambda} = \dfrac{\Gamma[s']}{r'^{s'}}$ has been used.

d) $f(y'|y=7) = \dfrac{1}{y'!}\, \dfrac{\Gamma[57+y']}{\Gamma[57]}\, \dfrac{51^{57}}{52^{57+y'}}$

e) See R notebook.

8.13

- $y|\mu \sim N(\mu, \sigma^2)$ with $f(y|\mu) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right)$
- $\mu \sim N(\theta, \tau^2)$ with $f(\mu) = \dfrac{1}{\sqrt{2\pi\tau^2}}\, \exp\left(-\dfrac{(\mu-\theta)^2}{2\tau^2}\right)$

a) $\mu|\vec{y} \sim N\left(\theta\dfrac{\sigma^2}{n\tau^2+\sigma^2} + \bar{y}\dfrac{n\tau^2}{n\tau^2+\sigma^2}, \dfrac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)$

For a single measurement y:

$\mu|y \sim N\left(\theta\dfrac{\sigma^2}{\tau^2+\sigma^2} + y\dfrac{\tau^2}{\tau^2+\sigma^2}, \dfrac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right)$

$\Rightarrow f(\mu|y) = \dfrac{1}{\sqrt{2\pi\tau'^2}}\, \exp\left(-\dfrac{(\mu-\theta')^2}{2\tau'^2}\right)$, with $\theta' = \theta\frac{\sigma^2}{\tau^2+\sigma^2} + y\frac{\tau^2}{\tau^2+\sigma^2}$ and $\tau' = \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}$

b) $y'|\mu \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad f(y'|\mu) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left(-\dfrac{(y'-\mu)^2}{2\sigma^2}\right)$

c)

$$\begin{aligned}
f(y'|y) &= \int_{-\infty}^{\infty} f(y'|\mu)\ f(\mu|y)\ \mathsf{d}\mu \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\ \exp\left(-\frac{(y'-\mu)^2}{2\sigma^2}\right)\ \frac{1}{\sqrt{2\pi\tau'^2}}\ \exp\left(-\frac{(\mu-\theta')^2}{2\tau'^2}\right)\ \mathsf{d}\mu \\
&= \frac{1}{2\pi\sigma\tau'} \int_{-\infty}^{\infty} \exp\left(-\frac{(y'-\mu)^2}{2\sigma^2} - \frac{(\mu-\theta')^2}{2\tau'^2}\right)\ \mathsf{d}\mu \\
&= \frac{1}{2\pi\sigma\tau'} \int_{-\infty}^{\infty} \exp\left(-\frac{y'^2-2y'\mu+\mu^2}{2\sigma^2} - \frac{\mu^2-2\mu\theta'+\theta'^2}{2\tau'^2}\right)\ \mathsf{d}\mu \\
&= \frac{1}{2\pi\sigma\tau'}\ \exp\left(-\frac{y'^2}{2\sigma^2}\right)\ \exp\left(-\frac{\theta'^2}{2\tau^2}\right)\ \int_{-\infty}^{\infty} \exp\left(\frac{2y'\mu-\mu^2}{2\sigma^2} - \frac{\mu^2-2\mu\theta'}{2\tau'^2}\right)\ \mathsf{d}\mu.
\end{aligned}$$

The integral can then be computed as

$$\begin{aligned}
&\int_{-\infty}^{\infty} \exp\left(\frac{2y'\mu-\mu^2}{2\sigma^2} - \frac{\mu^2-2\mu\theta'}{2\tau'^2}\right)\ \mathsf{d}\mu \\
=\ &\int_{-\infty}^{\infty} \exp\left(\frac{2y'\mu\tau'^2-\mu^2\tau'^2-\mu^2\sigma^2+2\mu\theta'\sigma^2}{2\sigma^2\tau'^2}\right)\ \mathsf{d}\mu \\
=\ &\int_{-\infty}^{\infty} \exp\left(\frac{-\mu^2\left(\sigma^2+\tau'^2\right)+2\mu\left(\theta'\sigma^2+y'\tau'^2\right)}{2\sigma^2\tau'^2}\right)\ \mathsf{d}\mu \\
=\ &\int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\mu^2-2\mu\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)\right)\ \mathsf{d}\mu \\
\overset{(*)}{=}\ &\int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\mu^2-2\mu\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}+\left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\right)\ \exp\left(\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right) \\
=\ &\exp\left(\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\ \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\mu-\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\ \mathsf{d}\mu \\
=\ &\exp\left(\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\ \int_{-\infty}^{\infty} \exp\left(-\frac{\left(\mu-\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2}{2\frac{\sigma^2\tau'^2}{\sigma^2+\tau'^2}}\right)\ \mathsf{d}\mu \\
\overset{(**)}{=}\ &\exp\left(\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\ \left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\ \sqrt{2\pi\frac{\sigma^2\tau'^2}{\sigma^2+\tau'^2}},
\end{aligned}$$

where in (*) the quadratic complement has been used to arrive at a quadratic equation (similar to the derivation in the book in section 5.3.4) and in (**) the property

$$\int_{-\infty}^{\infty} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\ \mathsf{d}x = \sqrt{2\pi\sigma^2}$$

has been applied. Thus,

$$\begin{aligned}
f(y'|y) &= \frac{1}{2\pi\sigma\tau'}\sqrt{2\pi\frac{\sigma^2\tau'^2}{\sigma^2+\tau'^2}}\ \exp\left(-\frac{y'^2}{2\sigma^2}\right)\ \exp\left(-\frac{\theta'^2}{2\tau^2}\right)\ \exp\left(\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau'^2\right)}}\ \exp\left(-\frac{y'^2}{2\sigma^2}-\frac{\theta'^2}{2\tau^2}+\frac{\sigma^2+\tau'^2}{2\sigma^2\tau'^2}\left(\frac{\theta'\sigma^2+y'\tau'^2}{\sigma^2+\tau'^2}\right)^2\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau'^2\right)}}\ \exp\left(-\frac{1}{2\sigma^2\tau'^2}\left[y'^2\tau'^2+\theta'^2\sigma^2-\frac{1}{\sigma^2+\tau'^2}\left(\theta'\sigma^2+y'\tau'^2\right)^2\right]\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau'^2\right)}}\ \exp\left(-\frac{1}{2\sigma^2\tau'^2(\sigma^2+\tau'^2)}\right.\\
&\quad \left.\left[y'^2\tau'^2\sigma^2+y'^2\tau'^4+\theta'^2\sigma^4+\theta'^2\sigma^2\tau'^2-\theta'^2\sigma^4-2\theta'\sigma^2y'\tau'^2-y'^2\tau'^4\right]\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau'^2\right)}}\ \exp\left(-\frac{\sigma^2\tau'^2}{2\sigma^2\tau'^2(\sigma^2+\tau'^2)}\left[y'^2+\theta'^2-2\theta'y'\right]\right)\\
&= \frac{1}{\sqrt{2\pi\left(\sigma^2+\tau'^2\right)}}\ \exp\left(-\frac{(y'-\theta')^2}{2\left(\sigma^2+\tau'^2\right)}\right).
\end{aligned}$$

Even though the calculations are tedious, the result is surprisingly simple. Given the posterior mean $\theta' = \theta\frac{\sigma^2}{\tau^2+\sigma^2} + y\frac{\tau^2}{\tau^2+\sigma^2}$ and variance $\tau' = \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}$, the predictive distribution is centered around the same posterior mean θ', however with a broader variance $\sigma^2 + \tau'^2$, adding the additional sampling variability σ^2 to the posterior variance τ'^2.

d) $\theta' = \theta\frac{\sigma^2}{\tau^2+\sigma^2} + y\frac{\tau^2}{\tau^2+\sigma^2} = -10 \cdot \frac{1^2}{1^2+3^2} = -1$, $\tau' = \frac{\tau^2\sigma^2}{\tau^2+\sigma^2} = \frac{1^2 \cdot 3^2}{1^2+3^2} = 0.9$

$\Rightarrow \sigma^2 + \tau^2 = 3^2 + 0.9^2 = 9.81$

$\Rightarrow y'|y \sim N\left(-1, 9.81\right)$

e) See R notebook.

8.14

a) We are modelling a proportion for a binary outcome $\Rightarrow$ data follow a binomial distribution and it is reasonable to assume a beta prior.

b) My knowledge here is quite limited and I would assume a vague prior with a mean of 25%, thus $E[\pi] = \frac{\alpha}{\alpha+\beta} = 0.25$ or $\beta = \frac{1-0.25}{0.25}\alpha = 3\alpha$. Trying out different values of α, a Beta$(2, 6)$ seems to represent my expectations best.

c) My chosen Beta$(2, 6)$ shares the right-skewedness with the proposed linear Beta$(1, 2)$ prior, however the Beta$(1, 2)$ prior places much more weight at $\pi = 0$, where I assume that at least a percentage of the US population does not believe in climate change.

d) 15% (see notebook)

e) $y = 150$, $n = 1000$ $\quad\Rightarrow\quad$ Posterior model: $\pi \sim$ Beta$(\alpha + y, \beta + n - y) =$ Beta$(151, 852)$

New posterior mean: 15% with a standard deviation of 1% (plot see R notebook)

Posterior credible interval: $\pi \in [0.129, 0.173]$

8.15

a) π is in $[0.129, 0.173]$ with a plausibility of 95%, thus the plausibility that $\pi > 0.1$ is even larger than 95% and the alternative hypothesis is clearly favored by large.

b) $P(H_a) = \int_{0.1}^{1} f(\pi|y)\,\mathrm{d}\pi = 1 - \int_{0}^{0.1} f(\pi|y)\,\mathrm{d}\pi \approx 0.9999997$

c) BF $\sim 7.5 \times 10^4$. This is a huge Bayes factor, much larger than 1, there is an overwhelming amount of evidence for the alternative hypothesis.

d) The proportion π of US-americans who do not believe that climate change is real is larger than 10%, most likely somewhere between $\sim 13 - 17\%$.

8.16 - **8.18** See R notebook

8.19

a) The flipper length is a continuous quantity that is approximately normally distributed around a mean μ (see R notebook for a quick test of normality). Thus, a normal-normal model seems most appropriate.

b) It is a bit unclear, what 'as low as' or 'as high as' means in this context. Therefore I prefer to choose a standard deviation of 30 that results in a slightly more vague prior than one with standard deviation of 25 that would be approximately spot-on. Thus my choice is a $N(200, 30^2)$ prior.

c) $n = 151$ points with a sample mean of $\bar{y} = 189.95$.

d) $\mu \in [188.9, 190.1]$ (see R notebook). Note: The fact that σ is not given might be a mistake in the book (authors contacted).

8.20

a) H_0: $\mu \in [200, 220]$ mm, H_a: $\mu \notin [200, 220]$ mm

b) μ is below 200 with at least a plausibility of 95%. Thus it makes sense to clearly prefer H_a.

c) $P(H_0) = \int_{200}^{220} f(\mu|\vec{y})\,\mathrm{d}\mu = \int_{0}^{220} f(\mu|\vec{y})\,\mathrm{d}\mu - \int_{0}^{200} f(\mu|\vec{y})\,\mathrm{d}\mu \approx 0$ (very small number).

Consequently, $p(H_a) \approx 1$.

e) The population average flipper length u lies somewhere between 188.9 and 190.1 mm with 95% plausibility. The plausibility that it lies instead between 200 and 220 mm is negligibly small.

8.21

a) The number of loons observed per 100-hour period is a discrete, positive quantity. Here, the Gamma-Poisson model seems most appropriate and we will continue the exercise by using it, however it is not clear, whether the measured data follows a Poisson distribution (see R notebook).

b) $E(\lambda) = \frac{s}{r} = 2 \Rightarrow s = 2r$, $\mathrm{Var}(\lambda) = \frac{s}{r^2} = \frac{2r}{r^2} = \frac{2}{r} = 1 \quad \Rightarrow \quad r = 2, s = 4 \quad \Rightarrow \quad$ Gamma$(4, 2)$ prior

c) $\bar{y} = 1.5, n = 18$

d) Posterior model: Gamma $(s + \sum_i y_i, r + n)$ = Gamma $(s + n\bar{y}, r + n)$ = Gamma $(31, 20)$

Middle 95% posterior credible interval: $\lambda \in [1.05, 2.14]$ (see R notebook).

The elicitation of $\lambda = 2$ was not too bad, it is still in the middle 95% posterior credible interval.

8.22

a) H_0: $\lambda < 1$, H_a: $\lambda \geq 1$

b) λ is in $[1.05, 2.14]$ with a 95% posterior plausibility, thus our belief in H_a should be much stronger than the one in H_0.

c) $P(H_0) = \int_0^1 f(\lambda|\vec{y})\, \mathrm{d}\lambda \approx 0.013$. The plausibility of the null hypothesis is extremely small, the alternative hypothesis should be clearly preferred (odds ratio of 73).

d) λ is expected to lie within the interval $[1.05, 2.14]$ with 95% plausibility. Our hypothesis that only 1 loon per 100 hour period is expected has a plausibility of only 1.3%, with an odds ratio of only 1:73 and should thus rather be rejected.

e) See R notebook.

8.23 See R notebook.

8.24

a) See R notebook.

b) $E[y'] \approx 1.55$, Var$[y'] \approx 1.62$, Mode$[y'] = 1$

c) $P(0 \text{ loons}) \approx 22.1\%$

9 Simple normal regression

9.1

a) β_0 and β_1 take values over the entire real range. If we make a symmetric assumption around a center (mean), then the choice of normal distribution as a prior is appropriate.

b) σ should only take positive values, what cannot be represented by a normal distribution.

c) I'm not sure whether the used notions here align with the general literature. I assume that with 'vague prior', they mean things such as a uniform distribution over the entire space. In the temperature case, this means that also extremely negative temperatures could occur (such as -1000 degrees), what is at least in the context of the model impossible. A weakly informative prior is very vague (in terms of making rather broad assumptions), but still makes more assumptions than an uninformative prior. For temperatures, one can get the typical temperature range from the given data and then elicit a normal distribution that includes this range, but is still a bit broader (as apparently done by rstanarm).

9.2

a) X: arm length, Y: height

b) X: distance between home and work, Y: carbon footprint

c) X: age, Y: vocabulary level

d) X: sleep habits, Y: reaction

9.3 a) positive, b) positive, c) negative, d) probably negative

9.4 The smaller σ, the stronger the relationship and the more accurate a prediction.

9.5

a) X: age in years, Y: orange juice consumption in gallons

b) $Y \sim N(\mu, \sigma^2)$, where μ is the (global) average orange juice consumption and σ the (global) standard deviation.

c) $Y_i \sim N(\mu_i, \sigma^2)$, with $\mu_i = \beta_0 + \beta_1 X$ as the local mean

d) β_0: any real value - one might say that orange juice consumption at age zero will also be zero, however we are fitting a linear model to probably quite non-linear data, so the linear model could also start with a negative value at age zero.

β_1: any real value - consumption might increase or decrease with age, this remains to be found out, I personally have no idea whether kids, adults or seniors drink more orange juice.

σ: any positive real value.

e) Normal distributions seem a good choice for β_0 and β_1 and for σ the usual exponential distributions seems best.

Prior assumptions:

- β_0: As in the book, it probably makes more sense to define a centered intercept. At a middle age of 40, an average person probably drinks 1 glass of orange juice per day (3 dl = 0.08 gallons) and the consumptions probably varies between zero and 0.3 gallons ($\sim$1.1 l) of orange juice per day. The upper limit is quite high (at least for Europe..), but staying more vague is probably a good recipe here. Thus I elict the prior $\beta_{0,c} \sim N(\mu = 0.08, \sigma = 0.1)$ (see notebook for a plot), even though it goes a bit into negative values. Choosing a right-skewed prior distribution might also make sense here.
- β_1: Difficult! Extreme cases: 1) the consumption goes up from zero gallons to 0.3 gallons in 80 years ($\beta_1 = 0.3/80 \sim 0.04$), 2) the consumption jumps to 0.3 gallons in the childhood years and drops to zero in 80 years ($\beta_1 = -0.3/80 \sim -0.04$). On average it may just stay constant (baseline assumption). Thus I elicit the prior $\beta_1 \sim N(\mu = 0, \sigma = 0.02)$ (see notebook for a plot).
- σ: It will probably be very hard to predict orange juice consumption from age only. I elicit an average standard deviation of $\sigma = 0.16$ (2 glasses) and hence the rate of the exponential distribution is $\lambda = 1/0.16 = 6.25$, thus $\sigma \sim \text{Exp}(\lambda = 6.25)$.

9.6

a) X: temperature today, Y: temperature tomorrow

b) $Y \sim N(\mu, \sigma^2)$, where μ is the global average temperature and σ is the global standard deviation.

c) $Y_i \sim N(\mu_i, \sigma^2)$, with $\mu_i = \beta_0 + \beta_1 X$

d) β_0: any real value, β_1: any real value (although probably a positive one), σ: any positive real value

e) Also here, normal distributions seem a good choice for β_0 and β_1 and an exponential distribution seems a good choice for σ.

Prior assumptions:

- β_0: The intercept is likely centered around zero. A systematical bias for a lower to go to a higher or a higher to go to a lower temperature is not expected (maybe for extremely high temperatures). This bias is most certainly only small, I assume here a broad standard deviation of 5° centigrade and thus $\beta_0 \sim N(\mu = 0, \sigma = 5)$ in centigrades. If β_0 is to be centered, then μ is the center itself. This prior assumption is probably too vague, but better safe than sorry.
- β_1: I assume that the temperature will most likely stay the same, except if it is extraordinary high. As an unlikely extreme, 40° might mostly jump to 30° (slope of 0.75) or in the other extreme, 30° might mostly jump to 40° (slope of 1.3) - also these assumptions are probably too vague. This aligns more or less with, $\beta_1 \sim N(\mu = 1, \sigma = 0.1)$.
- σ: Probably the temperatures can only predicted within a standard deviation of 5° centigrades when taking only today's temperature as a feature. With $E[\sigma] = 5$, I use $\sigma \sim \text{Exp}(\lambda = 0.2)$.

9.7 a) False, b) True

9.8 Take the code given in the book and essentially replace a) height $\sim$ age, data = bunnies, b) Clicks $\sim$ Snaps, data = songs, c) Happiness $\sim$ Age, data = dogs.

9.9

a) Elicited priors: $\beta_{0,c} \sim N(\mu = 5000, \sigma = 2000)$, $\beta_1 \sim N(\mu = 10, \sigma = 5)$, $\sigma \sim \text{Exp}(0.0005)$.

   Model: $Y_i \sim N(\mu_i, \sigma)$, $\mu_i = \beta_0 + \beta_1 X$

b) See notebook.

c) See notebook.

d) The slope of the regression line varies to a much lower extent than the intercept, hence the different models for the local mean μ_i look almost like parallel lines. This is also visible in the generated data - the slope looks similar for all point clouds, however the intercept and especially the standard deviation vary a lot, for this reason the generated four point clouds look very different.

9.10-9.20 See notebook.

10 Evaluating Regression Models

10.1 How fair is the model? How wrong is the model? (in terms of model assumptions) How accurate are the posterior predictive models?

10.2-10.4 (find examples for yourself)

10.5 Every model is a theoretical concept and an approximation of reality. It will never be able to cover all aspects of reality. Nevertheless, a non-perfect model is better than no model at all and if we can use it to get at least approximate results for a problem, we can say it is useful.

10.6 Independence of the Y_i, linearity of $\mu_i = \beta_0 + \beta_1 X_i$, constant variance σ independent of X (homoscedasticity)

10.7

a) Sample a value from $N\left(\mu_1^{(1)}, \sigma^{(1)}\right)$, where $\mu_1^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_1$ and $X_1 = -1.8$

b) See notebook.

10.8

a) Goal: To check whether the model simulates predictions with a similar distribution to the actual data (i.e. is the model close to the unknown underlying 'model' generating the data?)

b) The distribution of the simulated data should lie 'close enough' to the simulated data. What 'close enough' means is use-case specific.

10.9

a) The median absolute error tells us the median deviation of the predicted $\hat{Y}_i$ from the actual Y_i.

b) While the magnitude of the median absolute error is use-case specific and depends on the typical size of Y_i, the scaled median absolute error divides the median absolute error by this typical scale and reports relative median absolute error in percent of this typical magnitude.

c) The within-50 statistics tells us, how many of the actual values for Y_i were between the upper and lower quartile of the predictive distribution $N(\mu_i, \sigma)$.

10.10

a) The darker density represents the actual distribution of data (normalized histogram). The lighter-colored densities represent our simulated predictions.

b) The densities of original data and simulation overlap more or less in area and show the same structure (e.g. in terms of modes, skewness, etc.). A good fitting model will produce samples that are close to the original data distribution.

c) The densities do not overlap in area and are e.g. shifted or show modes at different positions.

10.11

a) Data: whether Reem likes our recipes

b) Model: our preference to produce / generate only recipes with anchioves

c) Ask multiple people to evaluate recipes and divide them into folds.

d) The model would learn the preferences of only a subset of people and then test the recipees on the other people that were not in the training dataset.

10.12

a) I find division into four a bit artificial: Here's three: Create folds, train/test models on all splits, calculate cross-validation estimates.

b) The model might overfit and just learn the training data without generalizing to a broader setting.

c) (for yourself)

10.13-10.21 See notebooks.

11 Extending the Normal Regression Model

11.1 More than one variable might be needed to explain an outcome.

11.2

a) $\beta_1 = 0$, $\beta_2 = 0$ and $\beta_3 = 0$ automatically implies that the car is a Ford, representing our baseline model.

b) Average difference in miles per gallon between a Subaru and a Ford.

c) Average miles per gallon for a car of the Ford brand.

11.3

a) β_0: Average weight of a Mr. Stripey tomato, β_1: Relationship between the weight of a Mr. Stripey tomato and the number of days it has been growing, β_2: Average difference between a Roma and a Mr. Stripey tomato.

b) There is no difference in the average weight of both types of tomato.

11.4

a) The relationship between tomato weight and number of days it has been growing depends on the type of tomato, the slope might be steeper for one type than the other.

b) β_3 is only 'on' if the tomato is of Roma type and represents the correction to the slope β_1 if the tomato is of Roma type.

11.5

a) (look for your own example here)

b) (look for your own example here)

c) 1. Does it make sense from the context? (would you find it reasonable from your knowledge of the problem?) 2. (Bayesian) two-sided hypothesis test whether the interaction coefficient is larger than zero.

11.6

a) More predictors could explain more of the variance of the outcome.

b) A predictor might not be associated with the outcome and might not help to explain its variance, but rather contribute to overfitting to noise in the data. In addition, a model with less predictors is better understandable and explainable.

c) The childs body size, because I expect it to be correlated with shoe size.

d) The categorical variable whether the child knows how to swim. I do not expect any correlation of the ability to swim with the shoe size.

11.7

a) Easily understandable and explainable (simple structure, as few predictors as necessary), low prediction error.

b) Complicated model structure with many predictors (hard to understand how the model makes its predictions), high prediction error.

11.8

- Evaluating the predictive accuracy using visualisations: Posterior predictive intervals
- Evaluating the predictive accuracy using cross-validation: mae, mae scaled, within 50, within 95, ELPD

11.9 Low-capacity models have high bias (do not capture the underlying structure of the data well) and high-capacity models have a high variance (overfit to noise in the data). In practice, there is a trade-off between bias and variance and we want to select a model that has both a reasonably low (but not the lowest) bias and variance.

11.10-11.15 See notebook

12 Poisson & Negative Binomial Regression

12.1

a) Y: Count of the number of cars passing through on a road, X_1: time of day, X_2: working day / weekend

b) A link function g determines the connection between the expectation of the model distribution and the linear model through $g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$

c) Counts can only be positive. $\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$ is equivalent to $\mu = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...}$ and hence makes sure that the expectation of counts is always positive.

d) 1) Structure: Conditioned on X, the observed data points Y_i are independent. 2) Y_i represents the number of counts in a fixed interval and follows a Poisson distribution, 3) The expectation is linked to the linear model through a log-link, 4) $E[Y] = \text{Var}[Y] = \lambda$.

12.2 a) both, b) both, c) beither, d) Negative Binomial regression, e) both

12.3

a) Poisson regression makes the assumption that conditional expectation and variance of the outcome are equal.

b) Negative Binomial regression regression uses an additional parameter r that controls the amount of overdispersion (larger conditional variance than expectation). In the limit of large r, Negative Binomial regression converges to Poisson regression, for small r it is overdispersed.

c) If $E[Y|X] \approx \text{Var}[Y|X]$, we might better use the simpler (Poisson) model, since the parameter r in the Negative Binomial regression model adds additional complexity that might lead to more variance of the model.

12.4

a) e^{β_0}: The expected number of likes per hour if the Person who wrote the tweet has zero followers and did not include any emojis.

b) e^{β_1}: The multiplicative change in the number of likes per hour per follower of the Person who wrote the tweet.

c) e^{β_2}: The multiplicative correction of the expected number of likes per hour if an emoji is included in the tweet.

d) $\log(\lambda) = \beta_0 + 300\,\beta_1$

12.5-12.10 See notebook.

13 Logistic Regression

13.1 a) logistic regression, b) normal regression, c) normal regression

13.2 a) odds for rain tomorrow 4:1, b) odds for 2 heads in a row: 1:3, c) odds for bus on time: 1:1, d) odds for person being left-handed: 1:4

13.3 a) $\pi = 95\%$, b) $\pi = 33\%$, c) $p = 73\%$, d) $p = 12\%$

13.4

a) odds of belief in climate change $= e^{1.43-0.02\ \text{age}} = 4.18\ e^{-\ \text{age}/50}$

b) For an increase in one year of age, the odds of belief in climate change decrease by $e^{-0.02} \approx 0.98$, i.e. decrease by approximately 2%.

c) odds of belief in climate change for 60 years $= 4.18\ e^{-60/50} \approx 1.26, \quad \pi = \frac{\text{odds}}{1+\text{odds}} \approx 55.7\%$ (This is low!)

d) odds of belief in climate change for 20 years $= 4.18\ e^{-20/50} \approx 2.80, \quad \pi = \frac{\text{odds}}{1+\text{odds}} \approx 73.7\%$

13.5

a) Overall accuracy $= \frac{50+620}{50+30+620+300} = 67\%$

b) Sensitivity $= \frac{TP}{TP+FN} = \frac{620}{620+30} = 95.4\%$

c) Specificity $= \frac{TN}{TN+FP} = \frac{50}{50+300} = 14.3\%$

d) Increase the probability cut-off. With a higher cut-off, the model will put more people in the negative ('non-believer') group.

13.6-13.14 See notebook.

14 Naïve Bayes

14.1 The assumption of conditional independence is naïve.

14.2 a) Naïve Bayes, b) both, c) both

14.3 a) does not require MCMC, b) is wrong if there is substantial correlation among predictors

14.4-14.11 See notebook.

15 Hierarchical Models are Exciting

15.1

a) All data are shoveled into the same model and the group structure is ignored. This violates the independence assumption in regression models and discards group-level information.

b) Data from each group go into the fit of a different model and the fact that the data are about the same topic is ignored. This makes it impossible to predict for new groups and ignores any overall trend common to all groups in the nature of the overall topic.

c) The group structure is preserved with a hierarchical model that accounts both for within-group variability (modeling individual trends) and between-group variability (modeling overall trends).

15.2

a) Whenever data are grouped: We might for example have blood pressure data of different subjects for different types of activity. Both the individual health situation and the stressfulness of the activity have an impact on the resulting blood pressure.

b) Since group information is ignored, trends and levels in individual groups are ignored. In the above example, we would ignore that each individual subject has their own health situation.

c) Yes, we cannot predict for previously unseen groups (not in the training set) and overall trends are ignored (in terms of the example in a): every models learns again that certain types of activities are more stressful than others).

15.3-15.8 See notebook.

16 (Normal) Hierarchical Models without Predictors

16.1

- Complete pooling: posterior mean ratings are all equal, somewhere around 80.
- No pooling: posterior mean ratings are all centered in the centers of the boxplots.
- Hierarchical: posterior means are close to the centers of the boxplots, however drawn a bit towards the global mean of around 80. This also depends on the number of data points per group (see shrinkage) which cannot be inferred from the plot.

16.2

a) Grouping variable: Probably not all expeditions are covered and there are multiple climbers on the same expedition, influencing each others success.

b) Not a grouping variable: All seasons of interest are covered.

c) Probably not a grouping variable: It seems that all processing methods are covered by including 'Other'. If not all processing methods are covered in the data, then one could consider a hierarchical model.

d) Grouping variable: Not all farms can be covered in a sample and the farm on which the coffee is produced is definitely correlated with total cup points.

16.3

a) Each person has their own particular speed. I expect a relatively low within-group variability and a higher between-group variability. Discarding the information who typed leads to much more unexplained variability.

b) μ_j: Typical average typing speed of person j, μ global average typing speed of all involved people, σ_y: Variance in different measured typing speeds for the same person, σ_μ: variance in average typing speeds for all the involved people.

16.4

a) Four rather narrow distributions with significant overlap

b) Four well-separated narrow distributions with low within-group variability

c) Four broad distributions with high overlap

16.5 a) $\sigma_y \approx \sigma_\mu$, b) $\sigma_y < \sigma_\mu$, c) $\sigma_y > \sigma_\mu$

16.6-16.13 See notebook.

17 (Normal) Hierarchical Models with Predictors

17.1

a) We should use a hierarchical random intercepts and slopes model:

$$\begin{aligned}
Y_{ij}|\beta_{0j}, \beta_{1j}, \sigma_y &\sim N(\mu_{ij}, \sigma_y^2) \quad \text{with} \quad \mu_{ij} = \beta_{0j} + \beta_{1j} X_{ij} \\
\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \bigg| \beta_0, \beta_1, \sigma_0, \sigma_1 &\sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{\Sigma}\right) \\
\beta_{0c} &= N(m_0, s_0^2) \\
\beta_1 &= N(0, s_1^2) \\
\sigma_y &= \text{Exp}(l_y) \\
\mathbf{\Sigma} &= \text{(decomposition of covariance)}
\end{aligned}$$

b) Hierarchical random intercepts model:

$$\begin{aligned}
Y_{ij}|\beta_{0j}, \beta_1, \sigma_y &\sim N(\mu_{ij}, \sigma_y^2) \quad \text{with} \quad \mu_{ij} = \beta_{0j} + \beta_1 X_{ij} \\
\beta_{0j}|\beta_0, \sigma_0 &\sim N\left(\beta_0, \sigma_0^2\right) \\
\beta_{0c} &= N(m_0, s_0^2) \\
\beta_1 &= N(0, s_1^2) \\
\sigma_y &= \text{Exp}(l_y) \\
\sigma_0 &= \text{Exp}(l_0)
\end{aligned}$$

c) Hierarchical random slopes model:

$$\begin{aligned}
Y_{ij}|\beta_0, \beta_{1j}, \sigma_y &\sim N(\mu_{ij}, \sigma_y^2) \quad \text{with} \quad \mu_{ij} = \beta_0 + \beta_{1j} X_{ij} \\
\beta_{1j}|\beta_1, \sigma_1 &\sim N\left(\beta_1, \sigma_1^2\right) \\
\beta_{0c} &= N(m_0, s_0^2) \\
\beta_1 &= N(0, s_1^2) \\
\sigma_y &= \text{Exp}(l_y) \\
\sigma_1 &= \text{Exp}(l_1)
\end{aligned}$$

17.2

a) A plot similar to Fig. 17.8 (b) in the book.

b) $\sigma_y > \sigma_0$: More within-group variation than between-group variation, i.e. the reaction times among subjects vary more than the intercepts of the linear models: there is little distinction between different subjects, knowledge of the individual subject does not help much when predicting.

c) A plot similar to Fig. 17.8 (a) in the book.

d) $\sigma_y < \sigma_0$: Less within-group variation than between-group variation, i.e. the reaction times among subjects vary less than the intercepts of the linear models: there is more distinction between different subjects, knowledge of the individual subject helps a lot when predicting.

17.3

a) Plot similar to Fig. 17.11 (c) in the book.

b) There is a positive correlation between intercept and slope. People who have longer reaction times in general will respond more strongly to sleep deprivation.

c) Plot similar to Fig. 17.11 (a) in the book.

d) There is a negative correlation between intercept and slope. People who have longer reaction times in general will respond less strongly to sleep deprivation.

17.4 X_{ij}: height of puppy i of mother j, Y_{ij}: weight of puppy i of mother j

a) Similar as in 17.1 b)

b) Similar as in 17.1 a)

c) In the random intercepts model, it is assumed that the weight of all puppies increases by the same amount when their height increases by one unit, the only difference being in the weight baseline (intercept) given by the average weight of puppies of mother j. In contrast, in the random intercepts and slopes model the weight increase following a height increase is also specific to mother j and not global anymore.

17.5

- Model 1: `weight ~ height + (1 | mother)`
- Model 2: `weight ~ height + (height | mother)`

17.6-17.15 See R notebook.

18 Non-normal Hierarchical Regression and Classification

See notebook for all exercises.

19 Adding More Layers

19.1 Political value index, high school graduation rate are state-level predictors, language and antifamily are book-level predictors. Evidence can be found in the R notebook.

19.2 Height in meters and first ascent year are peak-level predictors, team count is expedition-level predictor and age and expedition role are climber-level predictors. Evidence can be found in the R notebook.

19.3

a) factory and worker

b) Factories → factory 1, factory 2, factory 3, factory 4 → worker 1, worker 2, worker 3 (for each factor) → widget 1, widget 2, widget 3, widget 4, widget 5 (for each worker)

c) Yes. The workers are within factories and the widgets within workers.

19.4

a) β_0 is the baseline number of defects when no information about factory or worker is available.

b) f_{0k} is the change in defect number baseline when it is known at what factory the widget is being produced, p_{0j} is the change in defect number baseline when the worker is known who produced the widget.

c) σ_y denotes the variation of widget defect for a given worker within a given factory, i.e. the inherent variability of workers that cannot be explained by variability within factories or workers. σ_b denotes the variation in the number of defects base rates among different workers and σ_f the variation in the number of defects base rates among different companies. In the given example, $\sigma_b \gg \sigma_y > \sigma_f$, i.e. most of the variability is explained by a variability between the skill of different workers and not by the technical environment in the factory, nor by the variability within the produced widgets of one worker. In short: There are better workers and worse workers and they produce consistent results, hire the best ones instead of improving conditions for workers in your factory (maybe not very realistic).

19.5-19.12 See R notebook.