STSB6816 Test 1 of 2022

Mathematical Statistics and Actuarial Science; University of the Free State

2022/04/07

Time: 180 minutes; Marks: 50

MEMORANDUM

Instructions

- Answer all questions in a single R Markdown document. Please knit to PDF or Word at the end and submit both the PDF/Word document and the .Rmd file for assessment, in that order.
- Label questions clearly, as it is done on this question paper.
- All results accurate to about 3 decimal places.
- Show all derivations, formulas, code, sources and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, no devices, and no communication capable websites may be accessed prior to submission. You may not (nor even appear to) attempt to communicate or pass information to another student.

Introduction

The data is provided on https://ufs.blackboard.com. **It consists of the queueing times of customers at two cashiers in a small store. Your task today is to establish whether the two cashiers are similar in serving speed, or whether there is a meaningful difference between them on average.**

These times (in minutes) are assumed to follow a gamma distribution (on the basis that they are the sum of exponential waiting times of independent customers arriving uniformly over time). You are encouraged to fit gamma distributions to these times as part of this process, although other approaches will get partial credit.

Question 1

1.1) Before touching the data, derive the Jeffreys prior for the gamma distribution parameters: shape α and rate λ (or scale β if you prefer). **[5]**

```
cat("$$\\begin{aligned}
g&=-\\log(f(x))\\\\
&=-\\alpha \\log \\lambda + \\log \\Gamma(\\alpha) - (\\alpha - 1)\\log x + \\lambda
x \\\\
\\frac{\\partial g}{\\partial\\alpha}&=-\\log\\lambda + \\psi(\\alpha) - \\log x\\\\
```

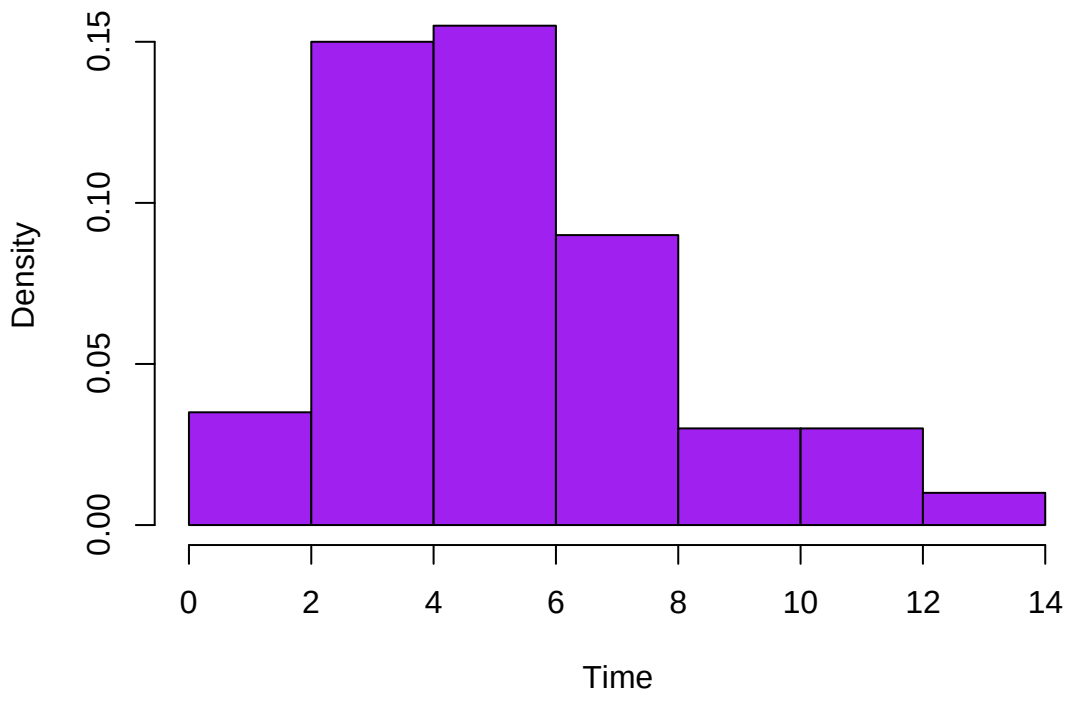
```
\\frac{\\partial g}{\\partial\\lambda}&=-\\alpha\\lambda^{-1} + x\\\\
\\pi&=\\left|\\begin{matrix}
\\psi'(\\alpha) & -\\lambda^{-1} \\\\
-\\lambda^{-1} & \\alpha\\lambda^{-2}
\\end{matrix}\\right|^{0.5}\\\\
&=\\lambda^{-1}\\sqrt{\\alpha\\psi'(\\alpha)-1}
\\end{aligned}$$")
```

$$
\begin{aligned}
g &= -\log\big(f(x)\big) \\
&= -\alpha\log\lambda + \log\Gamma(\alpha) - (\alpha-1)\log x + \lambda x \\
\frac{\partial g}{\partial \alpha} &= -\log\lambda + \psi(\alpha) - \log x \\
\frac{\partial g}{\partial \lambda} &= -\alpha\lambda^{-1} + x \\
\pi &= \begin{vmatrix} \psi'(\alpha) & -\lambda^{-1} \\ -\lambda^{-1} & \alpha\lambda^{-2} \end{vmatrix}^{0.5} \\
&= \lambda^{-1}\sqrt{\alpha\psi'(\alpha)-1}
\end{aligned}
$$

Negative log density [1]. Partial derivatives [2]. Information matrix [1]. Determinant and square root [1].

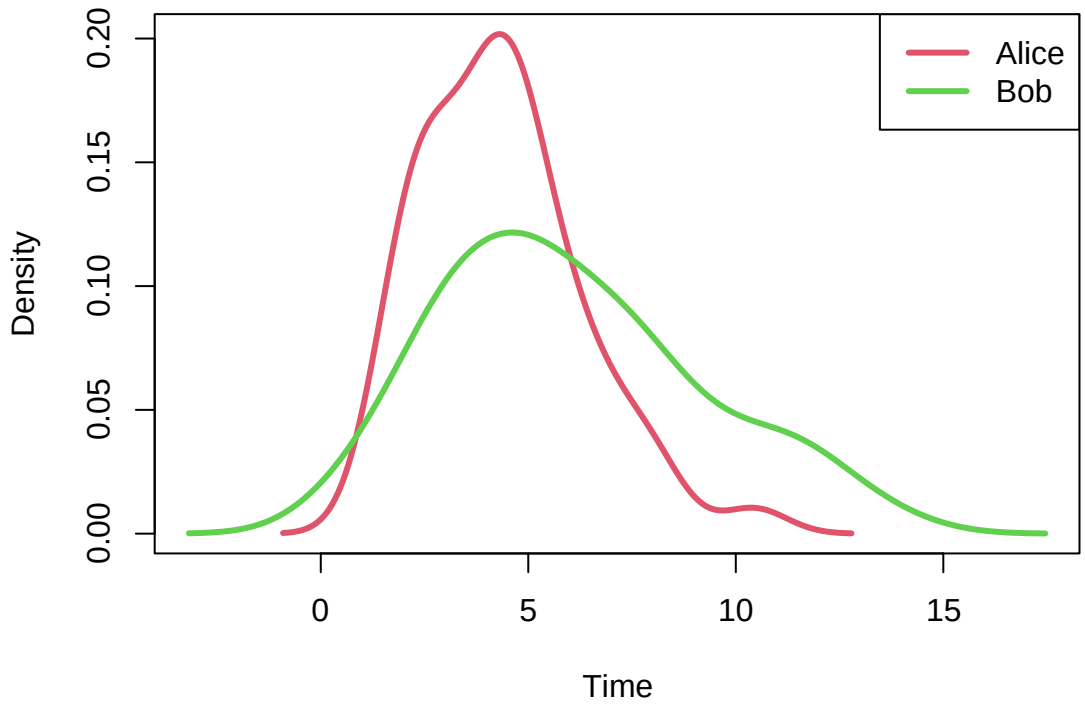
1.2) Read in the data set and explore it visually. You could start with a histogram or density plots and by drawing a box plot of Times against the Cashier name. Give a very short summary of what you see. **[5]**

```
"STSB6816Test1Data2022.xlsx" |> openxlsx::read.xlsx("TestData") -> d
Cashiers <- unique(d$Cashier)

par(mar=c(5,5,1,1))
d$Time |> hist(col = 'purple', main = '', xlab = 'Time', freq = FALSE)
```

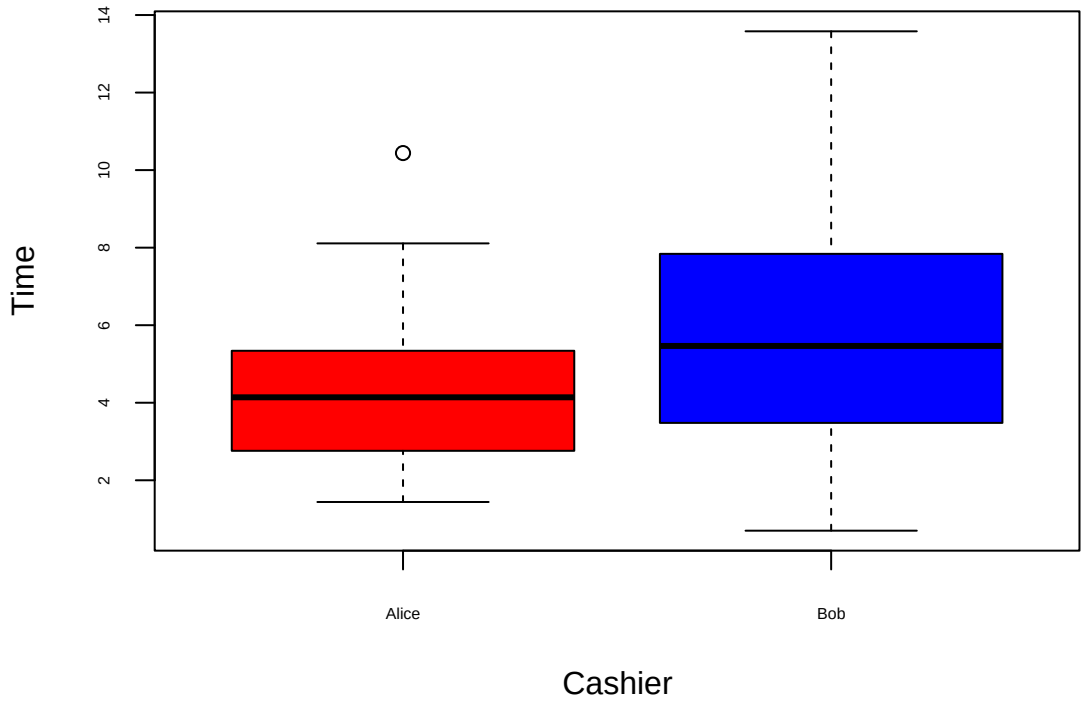



```
d$Time |> tapply(d$Cashier, density) -> densities
densities |> sapply(with, x) -> x
densities |> sapply(with, y) -> y
```

```r
plot(range(x), range(y), type = 'n', main='', xlab = 'Time', ylab = 'Density')
for (Cas in 1:2) {
  lines(x[,Cas], y[,Cas], col = Cas+1, lwd = 3)
}
legend('topright', legend = Cashiers, col = (1:2)+1, lwd = 3)
```




```r
boxplot(Time ~ Cashier, data=d, cex.axis=0.5, col=c('red','blue'))
```

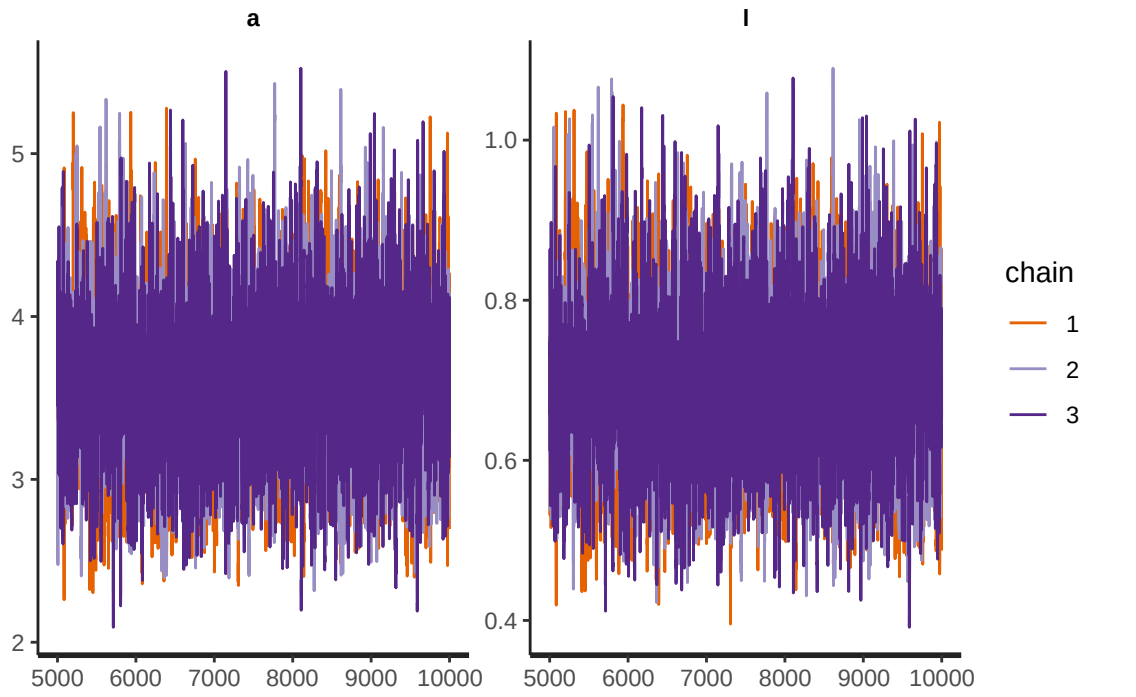


Load data [1], Histogram/density plot [1], Box plot [2], Discussion saying something about skewness or Bob seeming (not is) slower [1]. Any statement suggesting that Bob actually is slower based on this plot alone gets -3.

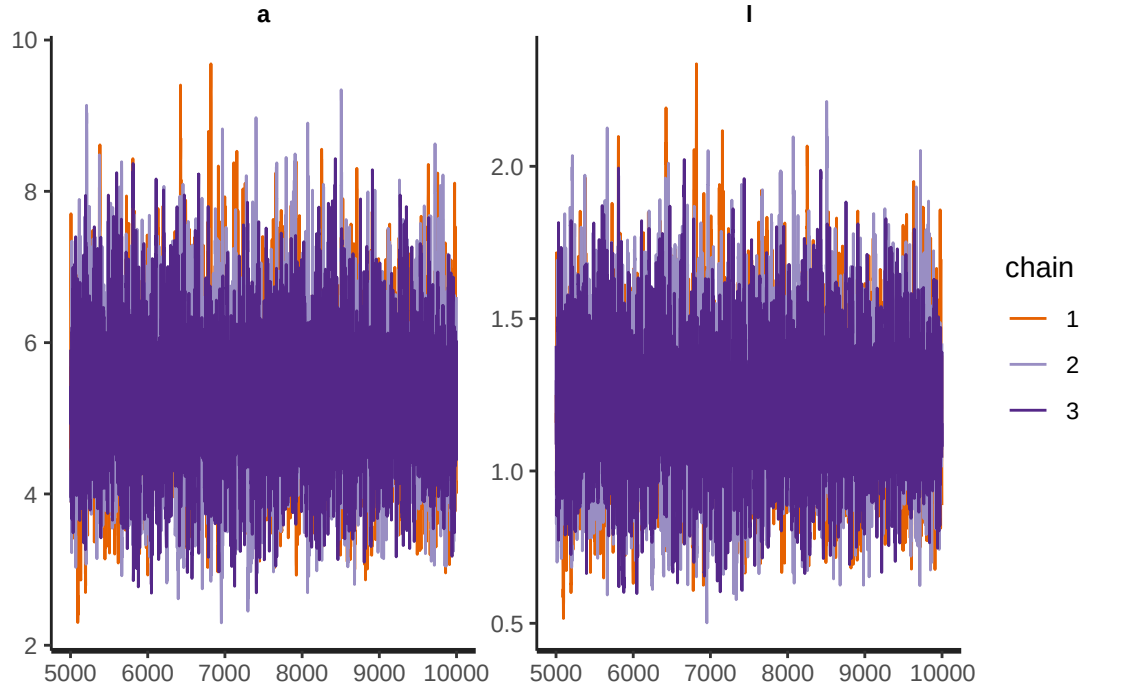
1.3) Fit distributions (ideally gamma distributions) to the times as a whole, as well as to the times of each individual cashier. Give parameter estimates, with uncertainty, for your fits. For full credit you must use the Jeffreys prior: $\lambda^{-1}\sqrt{\alpha\psi'(\alpha)-1}$ (trace plots showing good convergence are highly recommended for simulation fits). **[16]**

Hint: to add non-standard terms to the log posterior in Stan you can use *target +=* and then Stan math functions. Stan math functions are similar to R math functions (identical in this case) including the *trigamma()* function $(\psi'(\quad))$.

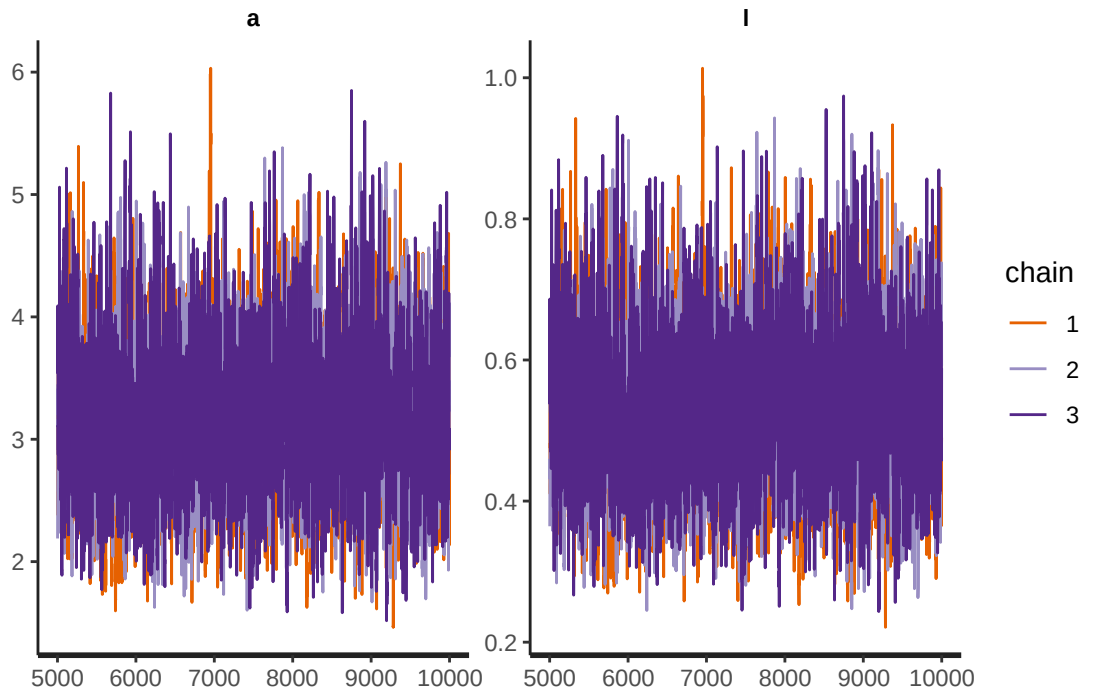
```
library(rstan)
mycores <- 3
options(mc.cores = mycores)

// This Stan block defines a Gamma model, by Sean van der Merwe, UFS
data {
 int<lower=1> n;              // number of observations
 real<lower=0> y[n];     // observations
}
// The parameters of the model
parameters {
 real<lower=0> a;
 real<lower=0> l;
}
model {
 y ~ gamma(a, l);
 target += 0.5*log(a*trigamma(a)-1) - log(l);
}

saveRDS(Gamma, file = 'Gamma.Rds')

ModelFit <- vector('list', 3)
ModelFit[[1]] <- sampling(Gamma, list(n=nrow(d), y=d$Time), iter = 10000, chains =
mycores)
CasRows <- which(d$Cashier == Cashiers[1])
ModelFit[[2]] <- sampling(Gamma, list(n=length(CasRows), y=d$Time[CasRows]), iter =
10000, chains = mycores)
CasRows <- which(d$Cashier == Cashiers[2])
ModelFit[[3]] <- sampling(Gamma, list(n=length(CasRows), y=d$Time[CasRows]), iter =
10000, chains = mycores)
names(ModelFit) <- c("Both", Cashiers)

ModelFit |> lapply(\(mf) {
  traceplot(mf) |> print()
  summary(mf)$summary[1:2,] |> kable(digits = 3) |> print()
})
```



	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	3.586	0.009	0.487	2.70	3.248	3.566	3.906	4.598	3104.231	1.000
l	0.694	0.002	0.102	0.51	0.623	0.690	0.761	0.907	3071.947	1.001



	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	5.262	0.018	1.019	3.435	4.549	5.189	5.911	7.429	3074.653	1
l	1.210	0.004	0.245	0.772	1.038	1.193	1.364	1.731	3045.467	1



	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a	3.157	0.012	0.605	2.101	2.734	3.106	3.531	4.478	2637.250	1.001
l	0.528	0.002	0.109	0.335	0.451	0.519	0.597	0.764	2676.487	1.001

$Both NULL

$Alice NULL

$Bob NULL

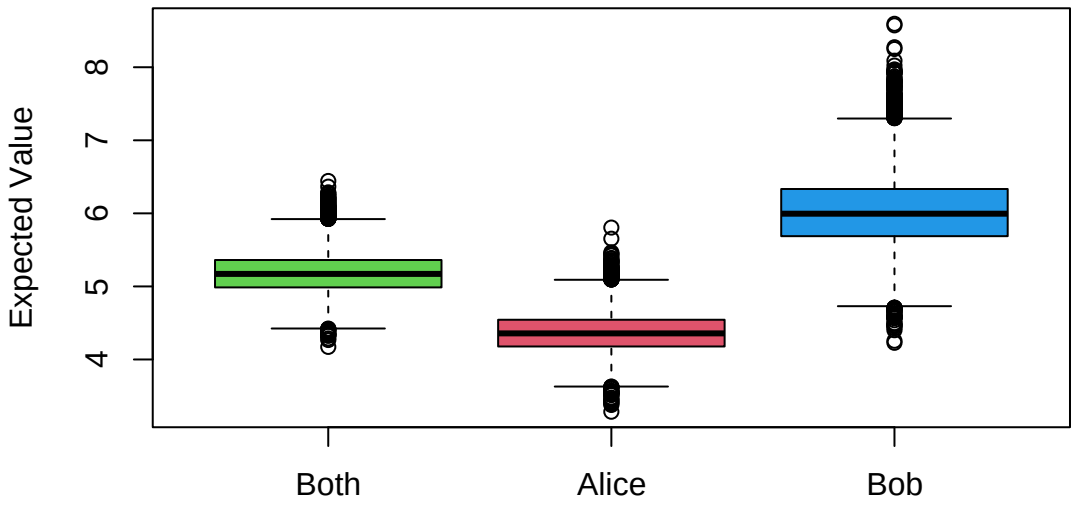
Fitting gamma model to all the data [2] (other model gets 1), giving parameter estimates with uncertainty [2], using the Jeffreys prior as given [2], giving a good trace plot [2]. Repeating for Alice and Bob [(1+1+1+1)x2].

1.4) For each fit, regardless of distribution, describe the modelled **average** wait time of customers in general. Either use a visual tool (e.g. density/box plot) or an estimate with interval or both. **[4]** [1 bonus mark for putting the three summaries in 1 table or plot]

```
ModelFit |> sapply(\(mf) {
  mf |> extract() -> sims
  sims$a/sims$l
}) -> mus
mus |> summary() |> kable(digits = 3)
```

Both	Alice	Bob
Min. :4.172	Min. :3.285	Min. :4.228
1st Qu.:4.986	1st Qu.:4.178	1st Qu.:5.688
Median :5.171	Median :4.357	Median :5.995
Mean :5.181	Mean :4.366	Mean :6.023
3rd Qu.:5.362	3rd Qu.:4.544	3rd Qu.:6.334
Max. :6.444	Max. :5.805	Max. :8.595

```
mus |> boxplot(ylab = 'Expected Value', col = c(3,2,4))
```



Using the formula for expected value of the relevant distribution to transform the parameters [2]. Summarising the results neatly [2]. Combined summary gets 1 bonus.

1.5) What is the probability that one of the cashiers is more than a minute slower than the other on average? What does your answer imply practically? **[4]**

```
(abs(mus[,2] - mus[,3]) > 1) |> mean()
```

```
[1] 0.8859333
```

```
max(mean((mus[,2] - mus[,3]) > 1), mean((mus[,3] - mus[,2]) > 1))
```

```
[1] 0.8859333
```

Comparing the difference to 1 before calculating the probability [2]. Combining the two sides (makes no difference in this case but still important for appearing objective) [1]. Saying that the probability is high and that there is probably a difference [1].

1.6) Consider a random future customer joining Alice's line. What is the probability that they will wait longer than 10 minutes according to your model? **[3]**

```
AliceSims <- extract(ModelFit[[2]])
nsims <- nrow(mus)
AlicePreds <- rgamma(nsims, AliceSims$a, AliceSims$l)
mean(AlicePreds > 10)
```

```
[1] 0.01126667
```

Generating random future Alice customers based on the posterior simulations [2], calculating the probability [1].

1.7) Consider 10 random future customers, 6 in Alice's line and 4 in Bob's. What is the probability that the one that has to wait the longest will wait more than 10 minutes? **[5]**

Hint: You must construct sets of 10 customers by taking 6 Alice predictions and 4 Bob predictions and concatenating them, then check whether the longest wait time of the 10 is longer than 10 minutes. You must do this at least 1000 times and average the results to get a probability estimate.

```
BobSims <- extract(ModelFit[[3]])
less_sims <- round(nsims*4/6)
BobPreds <- rgamma(less_sims, BobSims$a[1:less_sims], BobSims$l[1:less_sims])
WeightedPreds <- rbind(matrix(AlicePreds, 6), matrix(BobPreds, 4))
WeightedPreds |> apply(2, max) -> longest_wait
mean(longest_wait > 10)
```

```
| [1] 0.4372
```

Combining predictions in a weighted fashion [2], generating sets of 10 customers and finding the longest wait of each set [2], getting final probability [1].

1.8) Briefly explain more approaches that might allow you compare the models and possibly answer the core question of whether there is a meaningful difference in speed between the cashiers. Do not give any code or implementation, just words and references. **[8]**

The approach already implemented involved separately fitting models on the two sets of data and calculating a probability of a difference, so no marks for mentioning that approach.

The alternatives involve comparing the models of all the data together: one model with one set of parameters and one model with two sets of parameters [2].

In all cases we compare the predictive errors of the models and see which fits better [2], although in rare cases we could show that one model fits and the other does not, which is really nice when it happens.

Methods of comparison include: information criteria, cross validation, Bayes factors, visual comparisons of predictive fit [4].

Points total

The points on the test add up to **50**
