Exam 2023

2024-01-23

Exercise 1

Assume $X \sim Gamma(\alpha, \beta)$ in shape-rate parametrization and $Y|X = x \sim exp(x)$ in rate parametrization.

1.

By the law of total expectation and the Gamma functional equation

$$EY = E(E(Y|X)) = E(1/X) = \int_0^\infty \frac{1}{x}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}dx = \frac{\beta}{\alpha-1}\int_0^\infty \frac{\beta^{\alpha-1}}{\Gamma(\alpha-1)}x^{(\alpha-1)-1}e^{-\beta x}dx = \frac{\beta}{\alpha-1}$$

We check by simulation that we are correct

```
alpha <- 3
beta <- 2
X <- rgamma(10000, shape = alpha, rate = beta)
Y <- rexp(10000, rate = X)
mean(Y)
```

```
## [1] 1.014341
```

Which is fairly close to the theoretical value of 1.

2

Since the Y's are independent the joint density is

$$p((y_1, y_2, \cdots, y_n|\theta) = \prod_{i=1}^n \theta \exp(-y_i\theta) = \theta^n \exp(-\theta\sum_{i=1}^n y_i)$$

We have that

$$p(\theta) \propto \theta^{\alpha-1}\exp(-\beta\theta)$$

And so

$$p(\theta|y_1, y_2, \cdots my_n) \propto \theta^{\alpha-1}\exp(-\beta\theta)\cdot\theta^n\exp(\theta\sum_{i=1}^n y_i) = \theta^{n+\alpha-1}\exp(-\theta(\sum_{i=1}^n y_i + \beta))$$

Which we recognise as an unormalized $Gamma(n+\alpha, \sum_{i=1}^n y_i + \beta)$ distribution. Hence the posterior of θ is a $Gamma(n+\alpha, \sum_{i=1}^n y_i + \beta)$ distribution. Since the prior and posterior distributions of γ belong to the same family, this tells us that the Gamma distribution is a conjugate prior for θ with exponential likelihood.

3.

We simulate from the posterior to find a credible interval for θ.

```r
set.seed(1)
alpha <- 1.2
beta <- 6
sum_y <- 831
n <- 85
shape <- alpha + n
rate <- beta + sum_y

theta <- rgamma(1e5, shape = shape, rate = rate)
quantile(theta, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## 0.08238669 0.12577965
```

Hence a 95% credible interval for θ is [0.082386690.12577965]. Recall that the expected waiting times is exactly $1/\theta$ as θ is the rate parameter. We find a posterior mean and credible interval by transforming θ accordingly:

```r
mean(1/theta)
```

```
## [1] 9.82225
```

```r
quantile(1/theta, c(0.025, 0.975))
```

```
##      2.5%     97.5%
##  7.950412 12.137883
```

Hence the posterio mean for the expected waiting times is 9.82225 months and a 95 % credible interval is $[7.950412, 12.137883]$.

4.

We have that

$$E(\tilde{Y}|Y) = E(E(\tilde{Y}|\theta)|Y)$$

And since

$$E(\tilde{Y}|\theta) = \frac{1}{\theta}$$

We obtain by the law of total probability

$$E(\tilde{Y}|Y) = E(1/\theta|Y)$$

But this is just a special case of what we had in exercise 1. Hence we see that

$$E(\tilde{Y}|Y) = E(1/\theta|Y) = \frac{\sum_{i=1}^{n} y_i + \beta}{\alpha + n - 1}$$

We can also see this more directly. The posterior predictive distribution is the marginal distribution of Y given that θ has the posterior distribution. We know then that θ follows a $Gamma(n + \alpha, \sum_{i=1}^{n} y_i + \beta)$, and we know from question 1, that the mean of Y in this case is

$$E(\tilde{Y}|Y) = \frac{\sum_{i=1}^{n} y_i + \beta}{n + \alpha - 1}$$

Exercise 2

1.

```
load("januar23.Rdata")
mod <- lmer(log(humidity) ~ width * depth + (1|plank), data=plankData)
```

The model is a linear mixed model with the logarithm of humidity as the response and the width, depth and their interaction (width and depth are used as factors) as fixed effects plank as a random effect. This means that each plank has a random intercept, and each width/depth combination has an intercept for the logarithm of humidity. The depths and widths are (probably) chosen deliberately and we want to say something about the drying of the wood dependent for each width/depth pair. It is therefore reasonable to model them as fixed effects. We are not specifically interested in the humidity of individual planks, but rather from planks as a whole. To be able to interpret the planks as a random sample from the population of planks it is reasonable to model the plank as a random effect. Furthermore, measurements from the same plank are probably correlated, and therefore to make valid inference it makes sense to model plank as a random effect.

2.

We test the hypothesis $H_0 : EY \in L_w + L_d$ by doing a likelihood ratio test. We compute the p-value both by comparing to an asymptotic $\chi^2(8)$ and to simulated LRT statistics. This is done via the `PBmodcomp` function from `pbkrtest`. We run 2000 simulations

```
null_mod <- lmer(log(humidity) ~ width + depth + (1|plank), data = plankData)
PBmodcomp(mod, null_mod, nsim = 2000)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00303531 (tol = 0.002, component 1)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.00423296 (tol = 0.002, component 1)
```

```
## Bootstrap test; time: 54.33 sec; samples: 2000; extremes: 877;
## large : log(humidity) ~ width * depth + (1 | plank)
## log(humidity) ~ width + depth + (1 | plank)
##          stat df p.value
## LRT    8.1507  8  0.4189
## PBtest 8.1507     0.4388
```

We find that the p-value not significant using both methods. The p-values are almost identical using both methods. All in all we can not reject H_0.

3.

Looking at the summary of the model we see that

```
summary(null_mod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(humidity) ~ width + depth + (1 | plank)
##    Data: plankData
##
## REML criterion at convergence: -382.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6284 -0.5623  0.0480  0.5591  2.7631
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  plank    (Intercept) 0.03126  0.1768
##  Residual             0.01129  0.1063
## Number of obs: 300, groups:  plank, 20
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  1.537957   0.042737  35.986
## width2       0.048957   0.015029   3.258
## width3      -0.076369   0.015029  -5.082
## depth3       0.221928   0.019402  11.438
## depth5       0.271392   0.019402  13.988
## depth7       0.217295   0.019402  11.200
## depth9      -0.007873   0.019402  -0.406
##
## Correlation of Fixed Effects:
##        (Intr) width2 width3 depth3 depth5 depth7
## width2 -0.176
## width3 -0.176  0.500
## depth3 -0.227  0.000  0.000
## depth5 -0.227  0.000  0.000  0.500
## depth7 -0.227  0.000  0.000  0.500  0.500
## depth9 -0.227  0.000  0.000  0.500  0.500  0.500
```

We obtain the estimated standard deviation for the i'th observation by taking the square root of the sum of the estimated residual variance and the within-plank variance. That is

$$sd(Y_i) = \sqrt{0.03126 + 0.01129} = 0.2062765$$

We obtain the within-plank correlation by dividing the within-plank variance by the total variance, i.e.

$$\rho = \frac{0.03126}{0.03126 + 0.01129} = 0.7346651$$

4.

Let the variable $z = |5 - \text{depth}|$, where we treat depth as numeric for the construction of z. The hypotesis that plank humidity is symmetric in depth is equivalent to the hypothesis that $H_1 : EY \in H_z + H_w$ where z is used as a factor. We carry out the test below via pbmodcomp.

```
plankData <- plankData %>% mutate(num_depth = as.numeric(as.character(depth)), z = factor(abs(num_depth-
null_mod2 <- lmer(log(humidity) ~ z + width + (1|plank) , data = plankData)
PBmodcomp(null_mod, null_mod2)
```

```
## Bootstrap test; time: 30.94 sec; samples: 1000; extremes: 905;
## large : log(humidity) ~ width + depth + (1 | plank)
## log(humidity) ~ z + width + (1 | plank)
##          stat df p.value
## LRT    0.2264  2  0.8930
## PBtest 0.2264     0.9051
```

We get none-significant p-values, and can not reject H_1.

5.

a.

We compute a percentile bootstrap interval for ρ.

```
B <- 1000
rho_boot <- numeric(B)
sims <- simulate(null_mod, nsim = B)
plankData_copy <- plankData

get_rho <- function(model){
  frame <- as.data.frame(VarCorr(model)) %>% tibble()
  plank_var <- frame %>% filter(grp == "plank") %>% select(vcov) %>% as.numeric()
  tot_var <- sum(frame$vcov)
  rho <- plank_var/tot_var
  return(rho)
}

for (i in 1:B){
  plankData_copy$log_hum <- sims[[i]]
  mod_sim <- lmer(log_hum ~ width + depth + (1|plank), data = plankData_copy)
  rho_boot[i] <- mod_sim %>% get_rho()
}
```

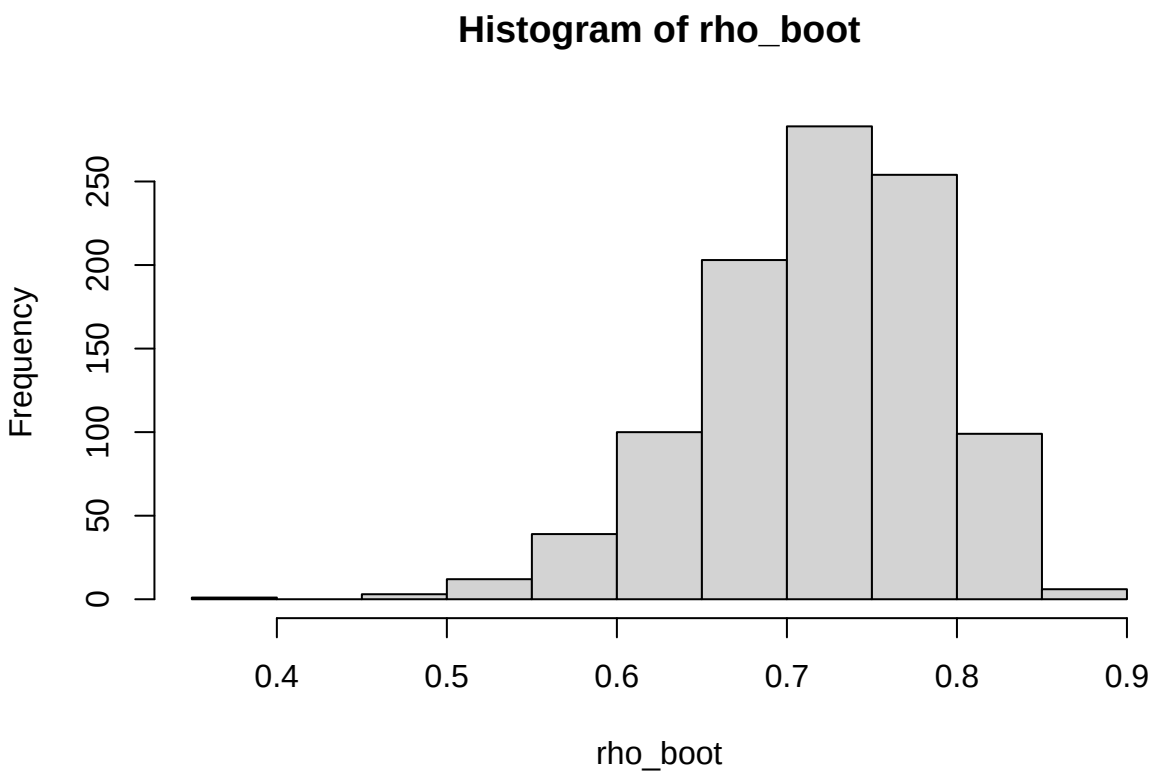
Below we compute the percentile interval

```
quantile(rho_boot, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.5681829 0.8286848
```

So a 95% bootstrap percintle confidence interval for ρ is $[0.5690622, 0.8293972]$. If we visualise the distribution of the bootstrapped rhos we see

```
hist(rho_boot)
```



This is quite asymetric, so as an alternative to the percentile interval, we can compute the basic bootstrap interval as well, which does not require the statistic to be symmetric, but just pivotal. We find that

```
rho_mod <- get_rho(null_mod)
c(2*rho_mod - quantile(rho_boot,0.975), 2*rho_mod - quantile(rho_boot,0.025))
```

```
##     97.5%      2.5%
## 0.6405287 0.9010306
```

So a 95% percent confidence interval for ρ is $[0.6338992, 0.9126809]$.

Alternatively we could compute a credible interval by obtaining simulations from the posterior distribution where we put stans defualt priors on parameters.

```
bayes <- brm(log(humidity) ~ depth + width + (1|plank) , data = plankData, refresh = 0)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
sims <- as.data.frame(bayes)
```

We compute credible intervals by taking the simulated values of σ^2 and σ^2_{plank} and computing

$$\rho = \frac{\sigma^2_{plank}}{\sigma^2_{plank} + \sigma^2}$$

for each simulation. We form a credible by taking quantiles:

```
rhos <- sims %>% tibble() %>%
  select(sd_plank__Intercept, sigma) %>% mutate(rho = sd_plank__Intercept^2/(sd_plank__Intercept^2 + sig
```

```r
quantile(rhos$rho, c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.6080231 0.8620821
```