Exercises II

1. Use the sample of opinions about the height of Eiffel collected from this course. Take the minimum and maximum value of each opinion, and make a uniform density, $\mathrm{U}(\min_i, \max_i)$, from each, $i = 1, \ldots, n$. Combine them into a single probability density by using the product method and the method of sum. Write the density function and *sketch* its features graphically in each case. The height of Eiffel is 324m (with antenna). The listed heights of other 'comparable' monuments may have affected the opinions? The **mi**nimum and **ma**ximum values are listed here pairwise:

```
mi=c(100,240, 60,140,150,250,250,200,100,180, 70, 70,100,300,290,150,100,300,190)
ma=c(300,320,200,450,250,450,350,600,300,300,150,130,500,400,310,500,400,400,250)
```

If each of the n uniform distributions $\mathrm{U}(a_i, b_i)$ is weighted by $1/n$, the sum is

$$\pi(x) = \sum_{i=1}^{n} \frac{1}{n} \frac{1_{\{a_i<x<b_i\}}(x)}{b_i - a_i}.$$

The plot can be produced in R (it was enough to sketch by pencil) by typing:

```
mi=c(100,240,60,140,150,250,250,200,100,180,70,70,100,300,290,150,100,300,190)
ma=c(300,320,200,450,250,450,350,600,300,300,150,130,500,400,310,500,400,400,250)
y <- numeric()
x <- seq(50,700,1)
for(i in 1:length(x)){y[i]<- sum((1/19)*(mi<x[i])*(ma>x[i])*(1/(ma-mi)))}
plot(x,y,type="l",xlab="height of Eiffel (m)",ylab="probability density")
points(c(324,324),c(0,0.0065),col="red",type="l",lwd=3)
```

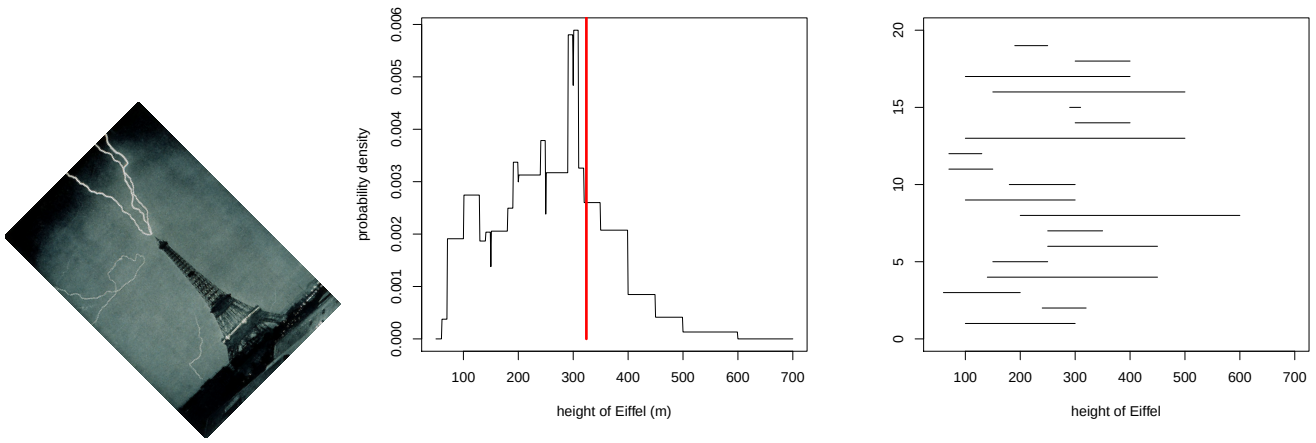

Figure 1: What's the height of Eiffel? Näsinneula, Tampere 1971: 168m. Tallinn TV Tower 1980: 312m. Petronas Tower, Kuala Lumpur, 1998: 452m. Sears Tower, Chicago, 1974: 442m. Empire State Building, New York, 1931: 381m. WTC towers, New York, 1972: 417m.

If we take a product, we are hopefully left with an overlapping interval that is common to all distributions. Elsewhere the resulting density is zero. For the overlap, the density is uniform and only needs

to be normalized: $\pi(x) = \frac{1_{\{\max(a)<x<\min(b)\}}(x)}{\min(b)-\max(a)}$. In this case, the maximum of the lower bounds was 300, and the minimum of upper bounds 130, which means the overlapping set is empty!

2. Let $\pi(X \mid N, p) = \text{Bin}(N, p)$ and $\pi(p) = \text{Beta}(\alpha, \beta)$. Calculate analytically from the Bayes formula that the posterior distribution $\pi(p \mid X)$ is $\text{Beta}(X + \alpha, N - X + \beta)$.

$$\pi(p \mid x) = \frac{\binom{N}{x} p^x (1-p)^{N-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 \binom{N}{x} p^x (1-p)^{N-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \mathbf{d}p}$$

$$= \frac{p^{x+\alpha-1}(1-p)^{N-x+\beta-1}}{\int_0^1 p^{x+\alpha-1}(1-p)^{N-x+\beta-1}\mathbf{d}p}$$

$$= \frac{\Gamma(N+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(N-x+\beta)} p^{x+\alpha-1}(1-p)^{N-x+\beta-1} = \text{Beta}(x+\alpha, N-x+\beta)$$

Here we used the general result from Beta(a,b)-distributions that, of course, the density must integrate to one, so that

$$1 = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}\mathbf{d}p$$

which gives a useful result:

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 p^{a-1}(1-p)^{b-1}\mathbf{d}p$$

and this is applied in the calculation of the integral in the denominator.

3. Use the sample data from this course: number of left-handed $X = 2$ / $N = 15$ (males) and $X = 0$ / $N = 3$ (females). Find the posterior density of the percentage of left-handed in the population of female and male students. Assume 'uninformative' prior. Is it really uninformative, considering what you really knew before the sample already? Summarize the results. According to literature, (Tiede-lehti, Google...), about 5-20% of the population is thought to be left-handed. Formulate a prior density reflecting this evidence, and recalculate the posteriors. Is there a difference? Judge the weight of the prior information against the weight of the sample.

With uniform prior, the two posterior distributions are $\text{Beta}(3, 14)$ and $\text{Beta}(1, 4)$. The uniform prior gives equal weight to all possible percentages, even though we know well by experience that the percentage is low, at least almost surely below 50%. Because the sample size is so small, the inferences remain very uncertain and the prior has a large influence. Using the literature information, we could formulate a prior based on that. For example, by using normal approximation, we could take $m = 0.125$, and $1.96s = 0.075$, which gives $s = 0.038$. Then, by solving the parameters of a corresponding Beta-distribution:

$$\alpha = -m(mm - m + ss)/(ss) \quad , \quad \beta = (mm - m + ss)(m-1)/(ss)$$

we get $\text{Beta}(\alpha, \beta) = \text{Beta}(9.34, 65.40)$ which is very informative prior, corresponding to a prior sample of size ≈ 75. The posteriors are then $\text{Beta}(2 + 9.34, 13 + 65.40)$ and $\text{Beta}(9.34, 3 + 65.40)$ which are

only slightly different from the prior, which means we have not learned much from these data!

4. According to birth statistics, there were 319,157 boys and 306,376 girls born during 1990-1999 in Finland. As Laplace, try to analyze the percentage θ of boys born in a large number of births, based on this evidence, by using paper and pencil. You need to approximate the beta-density posterior by e.g. normal density by matching the relevant parameters. Try to calculate $P(\theta > 0.5 \mid \text{data})$. You may use tabulated values (e.g. from some software) for cumulative normal distribution.

In this case, it is enough to assume uniform prior, because the data set has very large sample size which means it will dominate our inferences (unless we use a ridiculously informative prior). The posterior of boy percentage is then $\text{Beta}(319157 + 1, 306376 + 1)$ which has the following mean and variance (check from a suitable reference of distributions):

$$\mu = \frac{\alpha}{\alpha + \beta} = \frac{319158}{319158 + 306377} = 0.5102161$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{97782670566}{2.447685e + 17} = 3.994904e - 07\ , \quad \sigma \approx 0.0006320525$$

Thus, a normal approximation to this would be $\text{N}(0.5102161, 0.0006320525^2)$. By using some software, e.g. R (`pnorm`), we get a numerical value for the cumulative probability at 0.5, approximately: $P(\theta < 0.5 \mid \mu, \sigma) = 4.571023e - 59$, so that $P(\theta > 0.5 \mid \mu, \sigma) \approx 1$. The conclusion of $\theta > 0.5$ seems to be very sure!

5. Consider the binomial model $P(X \mid N, p) = \text{Bin}(N, p)$ and the posterior density for p, given X. This could result from an experiment where we decide to collect N samples and then observe X of them to be positive. Show that the posterior density is the same for the following experiment: a population has a fraction p of positives, and we continue testing until we obtain X positives. Why is this so?

Assume we use the same prior for p. Difference of posteriors is then possible only if the likelihood functions (discarding constant terms) are different. Let N denote the number of samples needed to get X positives. The likelihood function in such experiment is:

$$P(N \mid X, p) = \binom{N-1}{X-1}(1-p)^{N-1-(X-1)}p^{X-1}p \quad \propto (1-p)^{N-X}p^X$$

which is of the same form (for p) as with the binomial model $P(X \mid N, p)$. (This is also known as Negative Binomial model). Therefore, posterior densities will be identical if the prior is identical in both cases.