Cvičení č. 1, od 18.3.2024:

Při měření rychlosti $N = 50$ aut na Rohanském nábřeží naproti budově MFF jsme získali průměr $\overline{y} = 59.6$ km/h a výběrovou směrodatnou odchylku $s = 11.1$ km/h (měření probíhala v době před instalací dvou nových světelných křižovatek na rozích kancelářské budovy *River Garden*).

Vyšetřete aposteriorní rozdělení střední hodnoty rychlosti aut na Rohanském nábřeží za předpokladu normálního rozdělení rychlostí $\mathcal{N}(\mu, \sigma^2)$. Jako apriorní rozdělení pro μ a $\tau = \sigma^{-2}$ uvažujte (semi-konjugované) rozdělení:

$$p(\mu,\, \tau) = p(\mu)\, p(\tau),$$

kde $p(\mu) \propto 1$ a $\tau \sim \mathsf{Ga}(g,\, h)$, $p(\tau) \propto \tau^{g-1} \exp(-\, h\, \tau)$. Za g, h uvažujte postupně následující volby:

(i) $g = 0.001,\ h = 0.001$;

(ii) $g = 1,\ h = 0.005$;

(iii) $g = 0,\ h = 0$ (nevlastní rozdělení s $p(\tau) \propto \tau^{-1}$).

1. Pro kontrolu spočítejte obyčejný (tj. nebayesovský, frekventistický) 95% konfidenční interval pro neznámou střední hodnotu μ a taktéž pro směrodatnou odchylku σ a inverzní rozptyl τ.

2. Odvoďte marginální aposteriorní rozdělení pro μ a nakreslete jeho hustotu (do jednoho obrázku) pro výše uvedené volby g a h.

   S jakou aposteriorní pravděpodobností je střední hodnota rychlosti vyšší než 55 km/h (opět pro tři různé volby g a h)?

3. Odvoďte marginální aposteriorní rozdělení pro τ a nakreslete jeho hustotu (do jednoho obrázku) pro výše uvedené volby g a h. Do druhého obrázku nakreslete marginální aposteriorní hustoty pro $\sigma = \sqrt{1/\tau}$.

   Uvědomte si, že pro účely kreslení obrázku není potřeba explicitně odvozovat vzorec pro aposteriorní hustotu σ, máte-li k dispozici počítačovou funkci pro výpočet funkčních hodnot aposteriorní hustoty τ.

4. Nakreslete (do tří různých obrázků) *image/contour* graf sdružené aposteriorní hustoty $(\mu,\, \tau)$ pro tři výše uvedené volby hyperparametrů g, h.

   Uvědomte si, že pro výpočet funkčních hodnot sdružené hustoty $p(\mu,\, \tau \,|\, \mathbf{y})$ a její kreslení lze využít vztahu $p(\mu,\, \tau \,|\, \mathbf{y}) = p(\mu \,|\, \tau,\, \mathbf{y})\, p(\tau \,|\, \mathbf{y})$.

5. Nakreslete (do tří různých obrázků) *image/contour* graf sdružené aposteriorní hustoty $(\mu,\, \sigma)$ pro tři výše uvedené volby hyperparametrů g, h.

   Uvědomte si, že pro výpočet funkčních hodnot sdružené hustoty $p(\mu,\, \sigma \,|\, \mathbf{y})$ a její kreslení lze využít vztahu $p(\mu,\, \sigma \,|\, \mathbf{y}) = p(\mu \,|\, \sigma,\, \mathbf{y})\, p(\sigma \,|\, \mathbf{y})$, kde navíc $p(\mu \,|\, \sigma,\, \mathbf{y}) = p\big(\mu \,\big|\, \sigma^{-2},\, \mathbf{y}\big)$.

6. Pro každou výše uvedenou volbu hyperparametrů g a h spočtěte 95% ET (*equal-tail*) věrohodnostní intervaly pro μ, τ i σ.

7. Pro každou výše uvedenou volbu hyperparametrů g a h spočtěte (numericky, pokud si myslíte, že nelze jinak) 95% HPD (*highest posterior density*) věrohodnostní intervaly pro μ, τ i σ. Liší se tyto intervaly od ET věrohodnostních intervalů?

8. Napište krátkou funkci, pomocí níž lze generovat pseudonáhodná čísla ze (sdruženého) aposteriorního rozdělení (μ, τ) (apriorní hyperparametry specifikujte jako argumenty této funkce).

   Opět si uvědomte význam vztahu $p(\mu, \tau \,|\, \mathbf{y}) = p(\mu \,|\, \tau, \mathbf{y})\, p(\tau \,|\, \mathbf{y})$.

   Na základě nasimulovaného výběru z aposteriorního rozdělení (délky alespoň 10 000) spočtěte znovu (nyní Monte Carlo odhady pro) 95% ET věrohodnostní intervaly parametrů μ, τ a σ (opět pro tři volby hyperparametrů g a h). Liší se tyto intervaly od těch spočtených v bodu 6?

9. Výše uvedenou funkci na generování z aposteriorního rozdělení $p(\mu, \tau \,|\, \mathbf{y})$ mírně rozšiřte tak, aby bylo možno generovat též z prediktivního rozdělení rychlosti Y_{n+1} dalšího projíždějícího auta:

$$\begin{aligned} p(y_{n+1} \,|\, \mathbf{y}) &= \int p(y_{n+1}, \mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) \\ &= \int p(y_{n+1} \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) = \int p(y_{n+1} \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau). \end{aligned} \quad (1)$$

   Následně spočtěte (na základě minimálně 10 000 nasimulovaných hodnot z prediktivního rozdělení $p(y_{n+1} \,|\, \mathbf{y})$) 95% ET věrohodnostní interval pro Y_{n+1} (opět při třech volbách hyperparametrů g a h). Jak lze interpretovat věrohodnostní intervaly pro Y_{n+1}?

   Uvědomte si, že ke generování z $p(y_{n+1} \,|\, \mathbf{y})$ můžete využít (pouze na první pohled složitější) generování ze sdruženého rozdělení

$$p(y_{n+1}, \mu, \tau \,|\, \mathbf{y}) = p(y_{n+1} \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y}) = p(y_{n+1} \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y}).$$

10. Pomocí simulace dále aproximujte hodnoty prediktivní hustoty $p(y_{n+1} \,|\, \mathbf{y})$ (opět pro tři různé volby hyperparametrů g a h) a nakreslete je do jednoho obrázku.

    Uvědomte si, že k MC odhadu hodnot prediktivní hustoty, lze využít vztahu (1).

11. Po dalším mírném rozšíření výše uvedené funkce na generování z aposteriorního rozdělení $p(\mu, \tau \,|\, \mathbf{y})$ spočtěte Monte Carlo odhad pravděpodobnosti, že další projíždějící auto překročí nejvyšší povolenou rychlost o více než 30 km/h (tj. řidič spáchá přestupek, po němž následuje odebrání řidičského průkazu)?

    Uvědomte si obdobu vztahu (1):

$$\begin{aligned} \mathsf{P}(Y_{n+1} > 80 \,|\, \mathbf{y}) &= \int \mathsf{P}(Y_{n+1} > 80 \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) \\ &= \int \mathsf{P}(Y_{n+1} > 80 \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau). \end{aligned}$$

Deadline pro odevzdání vypracovaného úkolu (e-mailem na `komarek[AT]karlin.mff...`) je úterý 2.4. ve 14:04 CEST.

Exercise #1, since 18/03/2024:

When measuring the speed of $N = 50$ cars on Rohanské nábřeží, opposite of the MFF building, we obtained the sample mean $\overline{y} = 59.6$ and the sample standard deviation of $s = 11.1$ km/h (the measurements were taken in the period before installation of two traffic lights at the corners of the *River Garden* office building).

Examine the posterior distribution of the mean (expected value) of the speed of cars at Rohanské nábřeží while assuming a normal distribution $\mathcal{N}(\mu, \sigma^2)$ of the speed. As a prior for μ and $\tau = \sigma^{-2}$ consider a (semi-conjugate) distribution:

$$p(\mu,\, \tau) = p(\mu)\, p(\tau),$$

where $p(\mu) \propto 1$ and $\tau \sim \mathsf{Ga}(g,\, h)$, $p(\tau) \propto \tau^{g-1} \exp(-\,h\,\tau)$. For g and h consider the following choices:

(i) $g = 0.001,\ h = 0.001$;

(ii) $g = 1,\ h = 0.005$;

(iii) $g = 0,\ h = 0$ (improper distribution with $p(\tau) \propto \tau^{-1}$).

1. As a check, calculate the ordinary (i.e., non-bayesian, frequentist) 95% confidence interval for the unknown mean μ and also for the standard deviation σ and inverse variance τ.

2. Derive the marginal posterior distribution for *mu* and plot its density (in one figure) for the above choices g and h.

   What is the posterior probability that the mean of the speed is greater than 55 km/h (again for three different choices of g and h)?

3. Derive the marginal posterior distribution for *tau* and plot its density (in one figure) for the above choices g and h. In the second plot, draw the marginal posterior densities for $\sigma = \sqrt{1/\tau}$.

   Note that it is not necessary to explicitly derive the formula for the posterior density of σ for the purpose of plotting once you have a computer function to calculate the functional values of the posterior density of τ.

4. Draw (in three different plots) *image/contour* plot of a joint posterior density of $(\mu,\, \tau)$ for three different choices of g and h.

   Note that you can use the relation $p(\mu,\, \tau \,|\, \mathbf{y}) = p(\mu,\, \tau \,|\, \mathbf{y})\, p(\tau \,|\, \mathbf{y})$ to calculate the functional values of $p(\mu,\ \tau \,|\, \mathbf{y})$.

5. Draw (in three different plots) *image/contour* plot of a joint posterior density of $(\mu,\, \sigma)$ for three different choices of g and h.

   Note that you can use the relation $p(\mu,\, \sigma \,|\, \mathbf{y}) = p(\mu,\, \sigma \,|\, \mathbf{y})\, p(\sigma \,|\, \mathbf{y})$ to calculate the functional values of $p(\mu,\ \sigma \,|\, \mathbf{y})$. Moreover, $p(\mu \,|\, \sigma,\, \mathbf{y}) = p\big(\mu \,\big|\, \sigma^{-2},\, \mathbf{y}\big)$.

6. Calculate the 95% ET (*equal-tail*) credible interval for μ, τ i σ and each choice of g and h.

7. Calculate (numerically, if needed) the 95% HPD (*highest posterior density*) credible interval for μ, τ i σ and each choice of g and h. Do those intervals differ from the ET credible intervals?

8. Write a short function which can be used to generate pseudorandom numbers form the joint posterior distribution (μ, τ) (specify the prior hyperparameters as arguments of the function).

   Note once again that $p(\mu, \tau \,|\, \mathbf{y}) = p(\mu \,|\, \tau, \mathbf{y})\, p(\tau \,|\, \mathbf{y})$.

   Based on the simulated sample from the posterior distribution (of a length at least 10 000), calculate again (now Monte Carlo estimates of) the 95% ET credible intervals of parameters μ, τ a σ (again for the three choices of hyperparameters g and h). Do those intervals differ from the intervals calculated in point 6?

9. Extend the above function to generate also from the posterior predictive distribution of the speed Y_{n+1} of another car:

$$\begin{aligned} p(y_{n+1} \,|\, \mathbf{y}) &= \int p(y_{n+1}, \mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) \\ &= \int p(y_{n+1} \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) = \int p(y_{n+1} \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau). \end{aligned} \quad (2)$$

   Consequently, calculate (based on at least 10 000 simulated values from the predictive distribution $p(y_{n+1} \,|\, \mathbf{y})$) the 95% ET credible intervals for Y_{n+1} (again, while considering the three chioces of hyperparameters g and h). How would you interprete the credible intervals for Y_{n+1}?

   Remind that to generate from the distribution $p(y_{n+1} \,|\, \mathbf{y})$ it might be useful to generate from seemingly a more complicated joint distribution

$$p(y_{n+1}, \mu, \tau \,|\, \mathbf{y}) = p(y_{n+1} \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y}) = p(y_{n+1} \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y}).$$

10. Use the simulation and approximate values of the predictive density $p(y_{n+1} \,|\, \mathbf{y})$ (again for the three choices of hyperparameters g a h) and draw them in one plot.

    Note that the MC estimate of the predictive density can be obtained while using the relationship (2).

11. Extend a bit more the function which generates from the posterior distribution $p(\mu, \tau \,|\, \mathbf{y})$ and calculate the Monte Carlo estimate of a probability that another car exceeds the allowed speed by more than 30 km/h (and the driver loses their driving license for certain period of time).

    Note an analogue of the relationship (2):

$$\begin{aligned} \mathsf{P}(Y_{n+1} > 80 \,|\, \mathbf{y}) &= \int \mathsf{P}(Y_{n+1} > 80 \,|\, \mu, \tau, \mathbf{y})\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau) \\ &= \int \mathsf{P}(Y_{n+1} > 80 \,|\, \mu, \tau)\, p(\mu, \tau \,|\, \mathbf{y})\, \mathrm{d}(\mu, \tau). \end{aligned}$$

Deadline to deliver the report (e-mail to `komarek[AT]karlin.mff...`): Tuesday 2 April at 14:04 CEST.