1 Example Sheet 2

As before, I assume the following imports before all code fragments:

```
import random
import numpy as np
import pandas as pd
import scipy.stats as stats
import scipy.optimize as opt
from sklearn.linear_model import LinearRegression
```

1. Define a function `rxy()` that produces a random pair of values (X, Y) which, when shown in a scatter plot, produces a smiley face like this. Also plot the marginal distributions of X and Y.

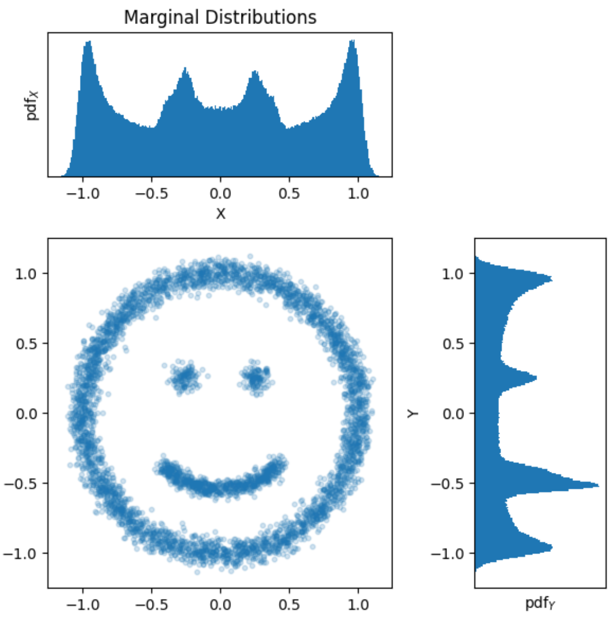
```
def circle(n):
        theta = 2 * np.random.random(n) * np.pi
        mu = np.column_stack([np.cos(theta), np.sin(theta)])
        return stats.norm.rvs(mu, 0.05)

def smile(n):
        theta = (np.random.random(n) - 1.5) * np.pi / 2
        mu = np.column_stack([np.cos(theta), np.sin(theta)])
        return 0.5 * stats.norm.rvs(mu, 0.05)

def left_eye(n):
        return stats.norm.rvs([-0.25, 0.25], 0.05, (n, 2))

def right_eye(n):
        return stats.norm.rvs(0.25, 0.05, (n, 2))

def face(n):
        p = [0.8, 0.15, 0.025, 0.025]
        return np.array([circle(n), smile(n), lefteye(n), righteye(n)])[
                np.random.choice([0,1,2,3], p=p, size=n), np.arange(n)]
```



2. Consider this code for generating random variables X and Y:

```
x = np.random.uniform()
y = np.random.geometric(p=x)
```

Derive the marginal likelihood $\Pr_Y(y)$ and the conditional likelihood $\Pr_X(x \mid Y = y)$.

The marginal likelihood $\Pr_Y(y)$ is gained by integrating $\Pr_Y(y \mid X = x)\Pr_X(x)$ over the valid range for X – in this case $[0, 1]$. We can do this using integration by parts.

$$\begin{aligned}
\Pr_Y(y) &= \int_0^1 \Pr_Y(y \mid X = x) \cdot \Pr_X(x)\mathrm{d}x \\
&= \int_0^1 (1-x)^{y-1}x \cdot 1\mathrm{d}x \\
&= \int_0^1 (1-x)^{y-1}x\mathrm{d}x \\
&= \left[-\frac{1}{y}x(1-x)^y\right]_0^1 + \frac{1}{y}\int_0^1 (1-x)^2\mathrm{d}x \\
&= -\frac{1}{y(y+1)}\left[(1-x)^{y+1}\right]_0^1 \\
&= \frac{1}{y(y+1)}
\end{aligned}$$

The conditional likelihood $\Pr_X(x \mid Y = y)$ can be derived using Bayes rule:

$$\begin{aligned}
\Pr_X(x \mid Y = y) &= \frac{\Pr_X(x) \cdot \Pr_Y(y \mid X = x)}{\Pr_Y(y)} \\
&= \frac{1 \cdot (1-x)^{y-1}x}{\frac{1}{y(y+1)}} \\
&= xy(y+1)(1-x)^{y-1}
\end{aligned}$$

3. I sample $x_1, \ldots, x_n$ from Uniform$[0, \theta]$. The parameter θ is unknown and I shall use $\Theta \sim \text{Pareto}(b_0, \alpha_0)$ as my prior, where $b_0 > 0$ and $\alpha_0 > 1$ are known. This has the cumulative distribution function:

$$\mathbb{P}(\Theta \leq \theta) = \begin{cases} 1 - \left(\frac{b_0}{\theta}\right)^{\alpha_0} & \text{if } \theta \geq b_0 \\ 0 & \text{if } \theta < b_0 \end{cases}$$

   (a) Calculate the prior likelihood for Θ.

   The prior likelihood for θ is given by $\frac{\partial \mathbb{P}}{\partial \theta}$. This is equal to:

$$\begin{aligned}
\text{pdf}_\Theta(\theta) &= \begin{cases} \frac{\alpha_0 b_0^{\alpha_0}}{\theta^{\alpha_0 - 1}} & \text{if } \theta \geq b_0 \\ 0 & \text{otherwise} \end{cases} \\
&= \frac{\alpha_0 b_0^{\alpha_0}}{\theta^{\alpha_0 - 1}} 1_{\theta \geq b_0}
\end{aligned}$$

   (b) Show that the posterior distribution of $(\Theta | x_1, \ldots, x_n)$ is Pareto and find its parameters.

   We can find the posterior distribution using Bayes rule.

$$
\begin{aligned}
\Pr(\theta|x_1,\ldots,x_n) &= \frac{\Pr(\theta)\cdot\Pr(x_1,\ldots,x_n|\theta)}{\Pr(x_1,\ldots x_n)} \\
&= \frac{\frac{\alpha_0 b_0^{\alpha_0}}{\theta^{\alpha_0-1}}1_{\theta\geq b_0}\frac{1}{\theta^n}1_{\theta\geq\max_i x_i}}{\Pr(x_1,\ldots x_n)} \\
&= \frac{\alpha_0 b_0^{\alpha_0}}{\Pr(x_1,\ldots x_n)}\frac{1}{\theta^{\alpha_0+n-1}}1_{\theta\geq b_0}1_{\theta\geq\max_i x_i} \\
&= \kappa\frac{1}{\theta^{\alpha_0+n-1}}1_{\theta\geq\max(b_0,\max_i x_i)} \text{ for some constant } \kappa \\
&= \frac{\kappa}{(\alpha_0+n)\left(\max\left(b_0,\max_i x_i\right)\right)^{\alpha_0+n}}\text{Pareto}(\max\left(b_0,\max_i x_i\right),\alpha_0+n)
\end{aligned}
$$

Since the whole function integrates to 1 over $(0,\infty)$ and $\text{Pareto}(\max\left(b_0,\max_i x_i\right),\alpha_0+n)$ also integrates to 1 over $(0,\infty)$, we can conclude that $\kappa=(\alpha_0+n)\left(\max\left(b_0,\max_i x_i\right)\right)^{\alpha_0+n}$ and therefore the distribution itself is $\text{Pareto}(\max\left(b_0,\max_i x_i\right),\alpha_0+n)$.

Therefore the posterior distribution for θ is $\text{Pareto}(\max\left(b_0,\max_i x_i\right),\alpha_0+n)$..

(c) Find a 95% posterior confidence interval for Θ.

We can find the inverse cumulative distribution function for pareto and use this to find θ such that $\mathbb{P}(\Theta<\theta)=0.95$.

$$
\begin{aligned}
p &= 1-\left(\frac{\max\left(b_0,\max_i x_i\right)}{\theta}\right)^{\alpha_0+n} \\
1-p &= \left(\frac{\max\left(b_0,\max_i x_i\right)}{\theta}\right)^{\alpha_0+n} \\
\frac{1}{\sqrt[\alpha_0+n]{\frac{1}{1-p}}} &= \frac{\max\left(b_0,\max_i x_i\right)}{\theta} \\
\theta &= \sqrt[\alpha_0+n]{\frac{1}{1-p}}\max\left(b_0,\max_i x_i\right)
\end{aligned}
$$

Setting $p=0.95$ gives $\text{cdf}_\theta\left(\sqrt[\alpha_0+n]{20}\max\left(b_0,\max_i x_i\right)\right)=0.95$

Therefore a 95% confidence interval for Θ is $\left[\max\left(b_0,\max_i x_i\right),\ \sqrt[\alpha_0+n]{20}\max\left(b_0,\max_i x_i\right)\right]$

(d) Find a different 95% posterior confidence interval. Which is better? Why?

Alternatively we could find values such that $\mathbb{P}(\Theta<\theta)=0.025$ and $\mathbb{P}(\Theta<\theta)=0.975$. This would give a 95% confidence interval

Plugging these values into the inverse cdf calculated earlier gives $\text{cdf}_\theta\left(\sqrt[\alpha_0+n]{\frac{40}{39}}\max\left(b_0,\max_i x_i\right)\right)=0.025$ and $\text{cdf}_\theta\left(\sqrt[\alpha_0+n]{40}\max\left(b_0,\max_i x_i\right)\right)=0.975$

Therefore another 95% confidence interval for Θ is $\left[\sqrt[\alpha_0+n]{\frac{40}{39}}\max\left(b_0,\max_i x_i\right),\ \sqrt[\alpha_0+n]{40}\max\left(b_0,\max_i x_i\right)\right]$.

Neither of these 95% confidence intervals is "better" than the other. However, I would argue that the second is more useful than the first in many cases.

4. I have a collection of numbers $x_1,\ldots,x_n$ which I take to be independent samples from the $\mathcal{N}(\mu,\sigma_0^2)$ distribution. Here σ_0 is known, and μ is unknown. Using the prior distribution $M\sim\mathcal{N}(\mu_0,\rho_0^2)$ for μ, show that the posterior distribution is:

$$
\Pr_M(\mu|x_1,\ldots,x_n)=\kappa e^{-\frac{(\mu-c)^2}{2\tau^2}}
$$

where κ is a normalising constant and where you should find formulae for c and τ in terms of σ_0, μ_0, ρ_0 and the x_i. Hence deduce that the posterior distribution is $\mathcal{N}(c, \tau^2)$.

Using Bayes rule:

$$\begin{aligned}
\Pr_M(\mu|x_1,\ldots,x_n) &= \frac{\Pr_M(\mu)\cdot\Pr_X(x_1,\ldots,x_n|\mu)}{\Pr_X(x_1,\ldots,x_n)} \\
&= \frac{\frac{1}{\sqrt{2\pi}\rho_0}e^{-\frac{(\mu-\mu_0)^2}{2\rho_0^2}}\prod_{i=1}^n\frac{1}{\sqrt{2\pi}\sigma_0}e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}}}{\Pr(x_1,\ldots,x_n)} \\
&= \kappa_0 e^{-\left(\frac{(\mu-\mu_0)^2}{2\rho_0^2}+\frac{\sum_{i=1}^n(\mu-x_i)^2}{2\sigma_0^2}\right)} \text{ for some constant } \kappa_0 \\
&= \kappa_0 e^{-\frac{\sigma_0^2\mu^2-2\sigma_0^2\mu_0\mu+n\rho_0^2\mu^2-2\rho_0^2\sum_{i=1}^n x_i}{2\sigma_0^2\rho_0^2}+\kappa_1} \text{ for some constants } \kappa_0,\kappa_1 \\
&= \kappa_0\ln\kappa_1 e^{-\frac{\left(\sigma_0^2+n\rho_0^2\right)\mu^2-2\left(\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i\right)\mu}{2\sigma_0^2\rho_0^2}} \text{ for some constants } \kappa_0,\kappa_1 \\
&= \kappa_2 e^{-\frac{\mu^2-2\frac{\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i}{\sigma_0^2+n\rho_0^2}\mu}{2\frac{\sigma_0^2\rho_0^2}{\sigma_0^2+n\rho_0^2}}} \text{ for some constant } \kappa_2 \\
&= \kappa_2 e^{-\frac{\left(\mu-\frac{\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i}{\sigma_0^2+n\rho_0^2}\right)^2}{2\frac{\sigma_0^2\rho_0^2}{\sigma_0^2+n\rho_0^2}}+\frac{\left(\frac{\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i}{\sigma_0^2+n\rho_0^2}\right)^2}{2\frac{\sigma_0^2\rho_0^2}{\sigma_0^2+n\rho_0^2}}} \text{ for some constant } \kappa_2 \\
&= \kappa_3 e^{-\frac{\left(\mu-\frac{\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i}{\sigma_0^2+n\rho_0^2}\right)^2}{2\frac{\sigma_0^2\rho_0^2}{\sigma_0^2+n\rho_0^2}}} \text{ for some constant } \kappa_3
\end{aligned}$$

This is of the form:

$$\Pr_M(\mu|x_1,\ldots,x_n) = \kappa e^{-\frac{(\mu-c)^2}{2\tau^2}}$$

With

$$c = \frac{\sigma_0^2\mu_0+\rho_0^2\sum_{i=1}^n x_i}{\sigma_0^2+n\rho_0^2} \qquad \tau = \frac{\sigma_0\rho_0}{\sqrt{\sigma_0^2+n\rho_0^2}}$$

These values pass sanity checks: for $n = 0$ the posterior distribution is the prior distribution $\mathcal{N}(\mu_0, \rho_0^2)$ and as $n \to \infty$, the distribution approaches $\mathcal{N}\left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sigma_0^2}{n}\right)$ – the distribution of the mean of the x_i.

5. I repeatedly attempt a task and each time I attempt it I succeed with probability θ and fail with probability $1-\theta$. The parameter θ is unknown, so I model it as a random variable Θ. Ever the optimist, my prior for Θ is heavily biased in favour of large values for θ:

$$\Pr_\Theta(\theta) = \varepsilon 1_{\theta\le\frac{1}{2}} + (2-\varepsilon)1_{\theta\ge\frac{1}{2}}$$

for some known small value $\varepsilon > 0$; this implies $\mathbb{P}\left(\Theta \le \frac{1}{2}\right) = \frac{\varepsilon}{2}$.

But I experience an unbroken run of n failures. How big does n need to be for me to concede there's a 50% posterior probability that $\Theta \le \frac{1}{2}$? How big would it need to be if $\varepsilon = 0$?

For the purposes of this question I will assume the *only observations we have* are this unbroken run of n failures. Without this assumption, the question is unanswerable.

The probability of $\Theta < \frac{1}{2}$ given a sequence of n failures is given by:

$$\varepsilon \int_0^{\frac{1}{2}} (1-\theta)^n \, \mathrm{d}\theta$$

The probability of $\theta \geq \frac{1}{2}$ given a sequence of n failures is given by:

$$(2-\varepsilon) \int_{\frac{1}{2}}^1 (1-\theta)^n \, \mathrm{d}\theta$$

If the probability of $\theta < \frac{1}{2}$ is 50% then these two probabilities must be equal to each other: they are the only two cases and the probability of all cases sums to 1 so the other probability must be 100% - 50% = 50%. We can therefore equate them and solve.

$$\begin{aligned}
\varepsilon \int_0^{\frac{1}{2}} (1-\theta)^n \, \mathrm{d}\theta &= (2-\varepsilon) \int_{\frac{1}{2}}^1 (1-\theta)^n \, \mathrm{d}\theta \\
\frac{\varepsilon}{n+1} \left[-(1-\theta)^{n+1} \right]_0^{\frac{1}{2}} &= \frac{2-\varepsilon}{n+1} \left[-(1-\theta)^{n+1} \right]_{\frac{1}{2}}^1 \\
\varepsilon \left(1 - \frac{1}{2}^{n+1} \right) &= (2-\varepsilon) \frac{1}{2}^{n+1} \\
\varepsilon &= \frac{1}{2}^n \\
n &= -\lg \varepsilon
\end{aligned}$$

So to concede that there is a 50% posterior probability that $\Theta \leq \frac{1}{2}$, we need to observe an unbroken run of $-\lg \varepsilon$ failures.

If $\varepsilon = 0$, then we could never be convinced our probability of failure was $\leq \frac{1}{2}$. If the probability of an event in the prior distribution is zero, then the probability of that event in the posterior distribution is always zero.

6. I have a collection of numbers

$$[4.3, 2.8, 3.9, 4.1, 9, 4.5, 3.3]$$

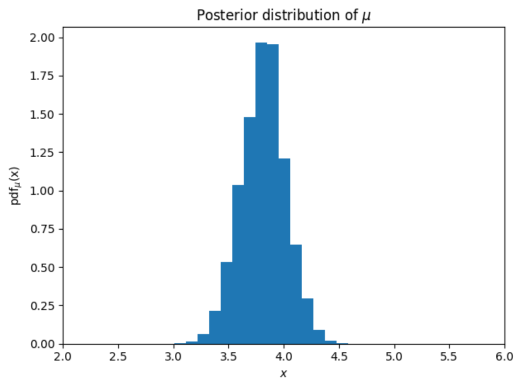
which look like they mostly come from a Gaussian distribution, but with the occasional outlier. Model the data as

$$X \sim \begin{cases} \mathcal{N}\left(\mu, 0.5^2\right) & \text{with probability } 99\% \\ \text{Cauchy} & \text{with probability } 1\% \end{cases}$$

Use a $\mathcal{N}(0, 5^2)$ prior distribution for μ. Give pseudocode to plot the posterior distribution.

```
n = 5000000
observed = [4.3, 2.8, 3.9, 4.1, 9, 4.5, 3.3]
data = np.full((n, len(observed)), observed)
p = np.random.random(size=data.shape)
logprob = np.empty_like(data)
logprob[p >= 0.99] = stats.cauchy.logpdf(data[p >= 0.99])
mus = stats.norm.rvs(0, 5, n)
logprob[p < 0.99] = stats.norm.logpdf(np.column_stack([mus]), data, 0.5)[p < 0.99]
```

```
logprob = np.sum(logprob, axis=1)
prob = np.exp(logprob - np.max(logprob))
"""
bins is so large because we're plotting a lot of data outside the
graph. We can resolve this by either cropping data or zooming into
the part we care about by generating points with scipy.stats.truncnorm
"""
plt.hist(mus, weights=prob, bins=500, density=True)
plt.show()
```



7. In lecture notes section 2.6 we investigated a dataset of police stop-and-search actions. Let the outcome for record i be $y_i \in \{0,1\}$, where 1 denotes that the police found something and 0 denotes that they found nothing. Consider the probability model $Y_i \sim B(1, \beta_{\text{eth}})$ where eth_i is the recorded ethnicity for the individual involved in record i, ans where the parameters β_{As}, β_{Blk}, β_{Mix}, β_{Oth}, β_{Wh} are unknown. As a prior distribution, suppose that the five β parameters are all independent $\beta\left(\frac{1}{2}, \frac{1}{2}\right)$ random variables.

   (a) Write down the joint prior density for $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$.

   Since the β are independent, the joint probability density function is the product of the individual probability density functions. Therefore, using the probability density function for the $\beta\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution; the joint probability density function is given by:

$$
\begin{aligned}
\Pr(\beta_{\text{As}}, \ldots, \beta_{\text{Wh}}) &= \prod_{\text{eth}} \frac{\beta_{\text{eth}}^{\frac{1}{2}-1}\left(1-\beta_{\text{eth}}\right)^{\frac{1}{2}-1}}{\left(\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}+\frac{1}{2}\right)}\right)} \\
&= \prod_{\text{eth}} \frac{\beta_{\text{eth}}^{-\frac{1}{2}}\left(1-\beta_{\text{eth}}\right)^{-\frac{1}{2}}}{\left(\frac{\Gamma\left(\frac{1}{2}\right)^2}{\Gamma(1)}\right)} \\
&= \prod_{\text{eth}} \frac{\beta_{\text{eth}}^{-\frac{1}{2}}\left(1-\beta_{\text{eth}}\right)^{-\frac{1}{2}}}{\left(\frac{\Gamma\left(\frac{1}{2}\right)^2}{1}\right)} \\
&= \prod_{\text{eth}} \frac{1}{\Gamma\left(\frac{1}{2}\right)^2\sqrt{\beta_{\text{eth}}(1-\beta_{\text{eth}})}}
\end{aligned}
$$

   (b) Find the joint posterior distribution of $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$ given the y data.

Using Bayes rule:

$$\begin{aligned}
\Pr(\beta_{\text{As}},\dots,\beta_{\text{Wh}}|Y) &= \frac{\Pr(\beta_{\text{As}},\dots,\beta_{\text{Wh}})\cdot\Pr(Y|\beta_{\text{As}},\dots,\beta_{\text{Wh}})}{\Pr(Y)} \\
&= \kappa_1\prod_{\text{eth}}\frac{1}{\Gamma\left(\frac{1}{2}\right)^2\sqrt{\beta_{\text{eth}}(1-\beta_{\text{eth}})}}\beta_{\text{eth}}^{\sum_{i=1}^n y_i 1_{\text{eth}_i=\text{eth}}}(1-\beta_{\text{eth}})^{\sum_{i=1}^n(1-y_i)1_{\text{eth}_i=\text{eth}}} \\
&= \kappa_2\prod_{\text{eth}}\beta_{\text{eth}}^{\left(\sum_{i=1}^n y_i 1_{\text{eth}_i=\text{eth}}\right)-\frac{1}{2}}(1-\beta_{\text{eth}})^{\left(\sum_{i=1}^n(1-y_i)1_{\text{eth}_i=\text{eth}}\right)-\frac{1}{2}} \\
&= \prod_{\text{eth}}\beta\left(\frac{1}{2}+\sum_{i=1}^n y_i 1_{\text{eth}_i=\text{eth}}, \frac{1}{2}+\sum_{i=1}^n(1-y_i)1_{\text{eth}_i=\text{eth}}\right)
\end{aligned}$$

Where the β on the line above denotes the probability density function of a β distribution with the given parameters.

8. I am prototyping a diagnostic test for a disease. In healthy patients, the test result is $\mathcal{N}(0, 2.1^2)$. In sick patients it is $\mathcal{N}(\mu, 3.2^2)$, but I have not yet established a firm value for μ. In order to estimate μ, I trialled the test on 30 patients whom I know to be sick and the mean test result was 10.3. I subsequently apply the test to a new patient and get the answer 8.8. I wish to know whether this new patient is healthy or sick.

   (a) In this question there are two unknown quantities: μ and $h \in \{\text{healthy}, \text{sick}\}$ the status of the new patient. Model the former as a random variable M with prior distribution $\mathcal{N}(5, 3^2)$ and the latter as a random variable H with prior distribution

   $$\Pr_H(h) = 0.99 \times 1_{h=\text{healthy}} + 0.01 \times 1_{h=\text{sick}}$$

   Write down the joint prior likelihood for (M, H).

   Since M and H are independent, the joint prior likelihood is the product of their individual likelihoods.

   $$\Pr(M, H) = (0.99\cdot 1_{h=\text{healthy}} + 0.01\cdot 1_{h=\text{sick}})\cdot\frac{1}{\sqrt{2\pi}\cdot 3}e^{-\frac{(\mu-5)^2}{2\cdot 3^2}}$$

   (b) In this question the data consists of 31 values, test results $x_1, \dots, x_{30}$ from the known sick patients and test result y from the new patient. Write down the data likelihood $\Pr(x_1, \dots, x_{30}, y|\mu, h)$.

   $$\begin{aligned}
   \Pr(x_1,\dots,x_{30},y|\mu,h) &= \Pr(y \mid \mu, h)\cdot\Pr(x_1,\dots,x_30 \mid \mu, h) \\
   &= \left(\frac{1_{h=\text{healthy}}}{\sqrt{2\pi}2.1}e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{1_{h=\text{sick}}}{\sqrt{2\pi}3.2}e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right)\prod_{i=1}^{30}\frac{1}{\sqrt{2\pi}\cdot 3.2}e^{-\frac{(\mu-x_i)^2}{2\cdot 3.2^2}} \\
   &= \frac{1}{\sqrt{2\pi}^{31}\cdot 3.2^{30}}\left(\frac{1_{h=\text{healthy}}}{2.1}e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{1_{h=\text{sick}}}{3.2}e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right)e^{-\frac{\sum_{i=1}^{30}(\mu-x_i)^2}{2\cdot 3.2^2}}
   \end{aligned}$$

   (c) Find the posterior density of (M, H). Leave your answer as an unnormalised density function. It should simplify to be a function of $\bar{x}$ and y where $\bar{x}$ is the mean test result for the known sick patients.

$$
\begin{aligned}
\Pr(\mu, h|x_1, \ldots, x_{30}, y) &= \frac{\Pr_{M,H}(\mu, h)\Pr\left(x_1, \ldots, x_{30}, y|\mu, h\right)}{\Pr\left(x_1, \ldots, x_{30}, y\right)} \\
&= \kappa_0 \Pr_{M,H}(\mu, h)\Pr\left(x_1, \ldots, x_{30}, y|\mu, h\right) \\
&= \kappa_1 e^{-\frac{(\mu-5)^2}{2\cdot 3^2}} \left(\frac{0.99 \cdot 1_{h=\text{healthy}}}{2.1} e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{0.01 \cdot 1_{h=\text{sick}}}{3.2} e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right) e^{-\frac{\sum_{i=1}^{30}(\mu-x_i)^2}{2\cdot 3.2^2}} \\
&= \kappa_2 e^{-\frac{(\mu-5)^2}{2\cdot 3^2}} \left(\frac{0.99 \cdot 1_{h=\text{healthy}}}{2.1} e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{0.01 \cdot 1_{h=\text{sick}}}{3.2} e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right) e^{-\frac{30\mu^2 - 2\mu\sum_{i=1}^{30} x_i}{2\cdot 3.2^2}} \\
&= \kappa_3 e^{-\frac{(\mu-5)^2}{2\cdot 3^2}} \left(\frac{0.99 \cdot 1_{h=\text{healthy}}}{2.1} e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{0.01 \cdot 1_{h=\text{sick}}}{3.2} e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right) e^{-\frac{\left(\mu - \frac{\mu\sum_{i=1}^{30} x_i}{30}\right)^2}{2\cdot\left(\frac{3.2}{\sqrt{30}}\right)^2}} \\
&= \kappa_3 e^{-\frac{(\mu-5)^2}{2\cdot 3^2}} \left(\frac{0.99 \cdot 1_{h=\text{healthy}}}{2.1} e^{-\frac{y^2}{2\cdot 2.1^2}} + \frac{0.01 \cdot 1_{h=\text{sick}}}{3.2} e^{-\frac{(\mu-y)^2}{2\cdot 3.2^2}}\right) e^{-\frac{(\mu-\bar{x})^2}{2\cdot\left(\frac{3.2}{\sqrt{30}}\right)^2}}
\end{aligned}
$$

(d) Give pseudocode to compute the posterior distribution of H, i.e. compute $\mathbb{P}\left(H = h|\text{data}\right)$ for both $h =$ healthy and $h =$ sick.

```
xs = [...]
y = ...
n = 100000

def logprob(xs, y, mus, h):
        broadcast_xs = np.full((n, len(xs)), xs)
        broadcast_mu = np.full((len(xs), n), mus).transpose()
        logpdfs = stats.norm.logpdf(broadcast_xs, broadcast_mu, 3.2)
        logpr_x = np.sum(logpdfs, axis=1)
        if h == 'healthy':
                logpr_y = stats.norm.logpdf(y, 0, 2.1)
        else:
                logpr_y = stats.norm.logpdf(y, mu, 3.2)
        return logpr_x + logpr_y

mus = stats.norm.rvs(3, 5, n)

healthy_logpr = logprob(xs, y, mus, 'healthy')
healthy_logpr = np.exp(healthy_logpr - np.max(healthy_logpr))

sick_logpr = logprob(xs, y, mus, 'sick')
sick_logpr = np.exp(sick_logpr - np.max(sick_logpr))

fig, (ax1, ax2) = plt.subplots(2)

ax1.hist(mu, weights=healthy_logpr, density=True, bins=k)
ax2.hist(mu, weights=healthy_logpr, density=True, bins=k)
```

9. In the lecture notes on linear modelling, we proposed a linear model for a temperature increase:

$$\mathbf{temp} \approx \alpha + \beta_1 \sin(2\pi\mathbf{t}) + \beta_2 \cos(2\pi\mathbf{t}) + \gamma(\mathbf{t} = 2000)$$

Suggest a probability model for **temp**. Suggest Bayesian prior distributions for the unknown parameters α, β_1, β_2 and γ. Give pseudocode to find a 95% confidence interval for γ.

I propose that the residuals are normally distributed around the prediction made by the linear model mean 0 and standard deviation σ. Therefore the model can be rewritten as

$$\mathbf{temp} \sim \alpha + \beta_1 \sin(2\pi\mathbf{t}) + \beta_2 \cos(2\pi\mathbf{t}) + \gamma(\mathbf{t} = 2000) + \mathcal{N}(0, \sigma^2)$$

I suggest the following prior distributions for α, β_1, β_2 and γ:

$$\begin{aligned} \alpha &\sim \mathcal{N}(10, 1) \\ \beta_1 &\sim \mathcal{N}(-1, 0.5) \\ \beta_2 &\sim \mathcal{N}(-6, 0.5) \\ \gamma &\sim \mathcal{N}(0, 0.1) \end{aligned}$$

```
n = 1000000
features = np.array([
                np.ones_like(climate.t),
                np.sin(2 * np.pi * climate.t),
                np.cos(2 * np.pi * climate.t),
                climate.t - 2000
                ])

coefs = np.array([
                stats.norm.rvs(10, 1, n),
                stats.norm.rvs(-1, 0.5, n),
                stats.norm.rvs(-6, 0.5, n),
                stats.norm.rvs(0, 0.1, n)
                ])

br_features = np.full((n, 4, climate.t.size), features)

br_coefs = np.full((climate.t.size, 4, n), coefs)

predictions = np.sum(br_features.transpose() * br_coefs, axis=1)
residuals = predictions - np.full((n, df.t.size), df.temp).transpose()
logprob = stats.norm.logpdf(residuals, 0, np.std(residuals, axis=0))
logprob = np.sum(logprob, axis=0)
prob = np.exp(logprob - np.max(logprob))
gammas = coefs[:, 3]
order = np.argsort(gammas)
gammas = gammas[order]
prob = prob[order] / np.sum(prob)
print(prob[np.cumsum(prob) > 0.025][0], prob[np.cumsum(prob) < 0.975][-1])
```

The output is 0.012644294207450358, 0.03886258682519564.

2 Supplementary question sheet 2

10. Consider this code for generating random variables $X \to Y \to Z$.

```
x = np.random.uniform()
y = np.random.binomual(n=1, p=x)
z = np.random.normal(loc=y, scale=epsilon)
```

Show that:

$$\Pr_Y(1|X=x, Z=z) = \frac{x}{x + (1-x)e^{\frac{1-2z}{2\varepsilon^2}}}$$

$$\begin{aligned}
\Pr_Y(1|X=x, Z=z) &= \frac{\Pr_X(x)\Pr_Y(1|x), \Pr_Z(Z|X,Y)}{\Pr_X(x)\Pr_Z(z|x)} \\
&= \frac{1 \cdot x \cdot \frac{1}{\sqrt{2\pi}\varepsilon} e^{-\frac{(z-1)^2}{2\varepsilon^2}}}{1 \cdot \left(\frac{x}{\sqrt{2\pi}\varepsilon} e^{-\frac{(z-1)^2}{2\varepsilon^2}} + \frac{(1-x)}{\sqrt{2\pi}\varepsilon} e^{-\frac{z^2}{2\varepsilon^2}}\right)} \\
&= \frac{x \cdot e^{-\frac{(z-1)^2}{2\varepsilon^2}}}{x \cdot e^{-\frac{(z-1)^2}{2\varepsilon^2}} + (1-x) \cdot e^{-\frac{z^2}{2\varepsilon^2}}} \\
&= \frac{x}{x + (1-x) \cdot e^{\frac{(z-1)^2 - z^2}{2\varepsilon^2}}} \\
&= \frac{x}{x + (1-x) \cdot e^{\frac{1-2z}{2\varepsilon^2}}}
\end{aligned}$$

How does $\Pr_Y(1|X=x, Z=z)$ depend on x and z when $\varepsilon \approx 0$? What if ε is very large?

If $\varepsilon \approx 0$ then $\Pr_Y(1|X=x, Z=z) \approx e^{\frac{2z-1}{2\varepsilon^2}}$. This assigns a very high probability if $z > \frac{1}{2}$ and a very low probability if $z < \frac{1}{2}$.

If ε is very large then $\Pr_Y(1|X=x, Z=z) \approx x$ – as the standard deviation of z is so high that z tells us very little.

11. Suppose we're given a function $f(x) \geq 0$ and we want to evaluate

$$\int_{x=a}^{b} f(x)\,\mathrm{d}x$$

Here's an approximation method: (i) draw a box that contains $f(x)$ over the range $x \in [a,b]$, (ii) scatter points uniformly at random in this box, (iii) return $A \times p$ where A is the area of the box and p is the fraction of points that are under the curve. Explain why this is a special case of Monte Carlo integration.

Consider taking a sample (x,y) from inside a box A. The probability that $y < f(x)$ for some function f is equal to the proportion of the box which is less than f. This is the definition of the integral of f (potentially clipped) over $[a,b]$. Therefore $P(y < f(x)) = \frac{\int_a^b f(x)\mathrm{d}x}{A}$. Consider now a binomial random variable Z which takes n such samples. By definition, Z therefore has the distribution $B(n, P(y < f(x))) = B\left(n, \frac{\int_a^b f(x)\mathrm{d}x}{A}\right)$.

By the expectation of binomial random variables:

$$\mathbb{E}(Z) = nP(y < f(x))$$

By splitting Z up into n individual bernoulli trials (x_i, y_i), we can rewrite Z as:

$$Z = \sum_{i=1}^{n} 1_{y_i < f(x_i)}$$

We can now combine these equations:

$$
\begin{aligned}
nP(y < f(x)) &= \mathbb{E}\left(\sum_{i=1}^{n} 1_{y_i < f(x_i)}\right) \\
nP(y < f(x)) &\approx \sum_{i=1}^{n} 1_{y_i < f(x_i)} \\
P(y < f(x)) &\approx \frac{1}{n}\sum_{i=1}^{n} 1_{y_i < f(x_i)} \\
\mathbb{E}\left(P(y < f(x))\right) &\approx \frac{1}{n}\sum_{i=1}^{n} 1_{y_i < f(x_i)}
\end{aligned}
$$

This is in the form of a Monte Carlo approximation; where we estimate the probability $P(y < f(x))$ by drawing samples from a distribution which is 1 with probability $P(y < f(x))$.

Notice that $\frac{1}{n}\sum_{i=1}^{n} 1_{y_i < f(x_i)}$ is the fraction of points under the curve – p and that by definition, $P(y < f(x)) = \frac{\int_a^b f(x)\mathrm{d}x}{A}$. Next, substitute these into the above formula:

$$
\begin{aligned}
\mathbb{E}\left(\frac{\int_a^b f(x)\mathrm{d}x}{A}\right) &\approx p \\
\frac{\int_a^b f(x)\mathrm{d}x}{A} &\approx p \\
\int_a^b f(x)\mathrm{d}x &\approx A \times p
\end{aligned}
$$

This proves that the proposed approximation technique is valid.

12. I have a biased coin, with unknown probability of heads θ. I toss it n times with outcomes $x_1, x_2, \ldots, x_n$ where $x_i = 1$ indicates heads and $x_i = 0$ indicates tails. My prior belief is $\Theta \sim \mathcal{U}[0,1]$. Here are two approaches to applying Bayes's rule:

- *One-shot Bayes.* Use Bayes's rule to compute the posterior of Θ, given data $(x_1, \ldots, x_n)$, using prior $\Theta \sim \mathcal{U}[0,1]$ and assuming that coin tosses are independent.

- *Sequential Bayes.* Use Bayes's rule to compute the posterior of Θ given data x_1 using the uniform prior; let the posterior density be $p_1(\theta)$. Apply Bayes's rule again to compute the posterior of Θ given data x_2, but this time using $p_1(\theta)$ as the prior; let the posterior density be $p_2(\theta)$. Continue applying Bayes's rule in this way until we have found $p_n(\theta)$.

State the posterior distribution found by one-shot Bayes. prove by induction on n that sequential Bayes gives the same answer.

The posterior distribution for θ found by one-shot Bayes is:

$$
\Theta \sim \beta\left(1 + \sum_{i=1}^{n} x_i, 1 + n - \sum_{i=1}^{n} x_i\right)
$$

I will prove by induction that the posterior distribution found by Sequential Bayes is the same:

For $n = 0$:

$$\begin{aligned}
\Pr_\Theta(\theta \mid \,) &= \frac{\Pr_\Theta(\theta)\Pr(1 \mid \theta)}{\Pr(1)} \\
&= \Pr_\Theta \\
&= 1 \\
&= \theta^0(1-\theta)^0 \\
&\sim \beta\left(1 + \sum_{i=1}^{0} x_i, 1 + 0 - \sum_{i=1}^{0} x_i\right)
\end{aligned}$$

Therefore the result holds for $n = 0$.

Assume next that it holds for $n = k$ and calculate the posterior probability for $n = k = 1$:

$$\begin{aligned}
\Pr_\Theta(\theta \mid x_1, \dots, x_{k+1}) &= \frac{\Pr_\Theta(\theta \mid x_1, \dots, x_k)\Pr(x_{k+1} \mid \theta, x_1, \dots, x_k)}{\Pr(x_{k+1})} \\
&= \kappa \cdot \theta^{\sum_{i=1}^{k} x_i}(1-\theta)^{n-\sum_{i=1}^{k} x_i} \cdot \theta^{x_{k+1}}(1-\theta)^{1-x_{k+1}} \\
&= \theta^{\sum_{i=1}^{k+1} x_i}(1-\theta)^{n-\sum_{i=1}^{k+1} x_i} \\
&\sim \beta\left(1 + \sum_{i=1}^{k+1} x_i, 1 + n - \sum_{i=1}^{k+1} x_i\right)
\end{aligned}$$

Therefore if the result holds for $n = k$, it must also hold for $n = k + 1$. Since it holds for $n = 0$, we can conclude that for all $n \in \mathbb{N}$:

$$\Pr(\Theta \mid x_1, \dots, x_n) \sim \beta\left(1 + \sum_{i=1}^{n} x_i, 1 + n - \sum_{i=1}^{n} x_i\right)$$

For the general case I will prove that the posterior probability for $\Pr(\theta_0, \dots, \theta_m \mid x_1, \dots x_n)$ obtained via sequential bayes is the same as the result obtained using one-shot Bayes.

For $n = 0$, the result from Sequential Bayes bayes is equal to:

$$\begin{aligned}
\Pr(\theta_1, \dots, \theta_m \mid 1) &= \frac{\Pr(\theta_1, \dots, \theta_m)\Pr(1 \mid \theta_1, \dots, \theta_m)}{\Pr(1)} \\
&= \Pr(\theta_1, \dots, \theta_m)
\end{aligned}$$

This is clearly the same as the result from one-shot Bayes.

Assume that for $n = k$, the result from Sequential Bayes is the same as the result from One-Shot Bayes:

$$\begin{aligned}
\Pr(\theta_1, \dots, \theta_m \mid x_1, \dots x_{k+1}) &= \frac{\Pr(\theta_1, \dots, \theta_m \mid x_1, \dots, x_k)\Pr(x_{k+1} \mid \theta_1, \dots, \theta_m, x_1, \dots, x_k)}{\Pr(x_{k+1})} \\
&= \frac{\Pr(\theta_1, \dots, \theta_m \mid x_1, \dots x_{k+1})\Pr(x_{k+1} \mid \theta_1, \dots, \theta_m)}{\Pr(x_{k+1})} \\
&= \frac{\frac{\Pr(\theta_1, \dots, \theta_m)\Pr(x_1, \dots, x_k \mid \theta_1, \dots, \theta_m)}{\Pr(x_1, \dots, x_k)}\Pr(x_{k+1} \mid \theta_1, \dots, \theta_m)}{\Pr(x_{k+1})} \\
&= \frac{\Pr(\theta_1, \dots, \theta_m)\Pr(x_1, \dots, x_{k+1} \mid \theta_1, \dots, \theta_m)}{\Pr(x_1, \dots, x_{k+1})}
\end{aligned}$$

This is the posterior distribution obtained by One-Shot Bayes. Therefore if the posterior distribution obtained from Sequential Bayes with k independent observations is equal to the posterior distribution obtained from One-Shot Bayes with k independent observations; then the posterior distribution obtained from Sequential Bayes with $k+1$ independent observations is the same as the posterior distribution obtained from One-Shot Bayes with $k+1$ independent observations. Since the posterior distributions obtained with zero observations were the same, we can conclude by mathematical induction that for all $n \in \mathbb{N}$; the posterior distributions obtained from One-Shot Bayes and Sequential Bayes are the same.

13. In the setting of question 7, I wish to measure the amount of police bias. Given a 5-tuple of parameters $\beta = (\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$, I define the overall bias score to be

$$d(\beta) = \max_{e,e'} |\beta_e - \beta_{e'}|$$

If $d(\beta)$ is large, then there is some pair of ethnicities with very unequal treatment. As a Bayesian, I view β as a random variable taking values in $[0,1]^5$, therefore $d(\beta)$ is a random variable also. To investigate its distribution, I sample β from the posterior distribution that I found in question 7, I compute $d(\beta)$, and plot a histogram. The output from the left is bizarre. To help me understand what's going on, I plot histograms of each of the individual β_e coefficients, shown here on the right.

Explain the results.

The beta distribution is U-shaped only when both parameters α and β are strictly less than 1. If there are any measurements for a given ethnicity then one of the the parameters become strictly greater than 1: the parameters are $\frac{1}{2} + \sum_{i=1}^{n} y_i 1_{\text{eth}_i=\text{eth}}$ and $\frac{1}{2} + \sum_{i=1}^{n}(1-y_i)1_{\text{eth}_i=\text{eth}}$.

Since the distribution for mixed is U-shaped, we can conclude that there are no records of any people of mixed ethnicity being stop-and-searched in the police dataset. This has skewed the posterior distribution of $d(\beta)$.

All ethnicities except for Mixed have data. Other has the least data, while Black, White and Asian having a sufficient data for the distribution to be moderately low.

14. Consider the outlier model from question 6. How likely is it that the datapoint with value 9 is an outlier.

I will consider an "outlier" to be a datapoint generated according to the Cauchy distribution.

Let Z be $1_{X \sim \text{Cauchy}}$. Therefore the probability that the datapoint with value 9 is an outlier is $\Pr(Z = 1 \mid X = 9)$. We can write this as a probability model using Bayes rule for continuous random variables.

$$\begin{aligned} \Pr(Z = 1 \mid X = 9) &= \frac{\Pr(Z = 1, X = 9)}{\Pr(X = 9)} \\ &= \frac{\Pr(Z = 1, X = 9)}{\Pr(Z = 1, X = 9) + \Pr(Z =, X = 9)} \end{aligned}$$

The functions on both side are well behaved and therefore we can calculate this explicitly. $\Pr(Z = 1, X = 9)$ is the probability density function of a Cauchy distribution evaluated at value 9. Using the probability density function for a Cauchy distribution:

$$\begin{aligned} \Pr(Z = 1, X = x) &= \Pr(Z = 1) \Pr(X = x \mid Z = 1) \\ \Pr(Z = 1, X = x) &= 0.01 \cdot \frac{1}{\pi(1 + x^2)} \end{aligned}$$

The probability density function for $\Pr(Z = 0, X = x)$ can be found by integrating the product of the prior distribution $\Pr(\mu)$ and the distribution $\Pr(X \mid Z = 0)$ between $-\infty$ and ∞.

$$\begin{aligned}\Pr(Z = 0, M = \mu, X = x) &= \Pr(Z = 0)\Pr(M = \mu \mid Z = 0)\Pr(X = x \mid Z = 0, M = \mu)\\ &= 0.99 \cdot \frac{1}{\sqrt{2\pi \cdot 5^2}} e^{-\frac{\mu^2}{2\cdot 5}} \cdot \frac{1}{\sqrt{2\pi \cdot 0.5}} e^{-\frac{(x-\mu)^2}{2\cdot 0.5^2}}\end{aligned}$$

We can therefore integrate this with respect to μ to work out $\Pr(Z = 0, X = x)$.

$$\begin{aligned}\Pr(Z = 0, X = x) &= \int_{-\infty}^{\infty} 0.99 \cdot \frac{1}{\sqrt{2\pi} \cdot 5} e^{-\frac{\mu^2}{2\cdot 5^2}} \cdot \frac{1}{\sqrt{2\pi} \cdot 0.5} e^{-\frac{(x-\mu)^2}{2\cdot 0.5^2}} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{\mu^2}{5^2} + 4(\mu - x)^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot 25}\left(\mu^2 + 100\mu^2 - 200\mu x + 100x^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot 25}\left(101\mu^2 - 200\mu x + 100x^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot \frac{25}{101}}\left(\mu^2 - \frac{200}{101}\mu x + \frac{100}{101}x^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot \frac{25}{101}}\left(\mu^2 - 2\frac{100}{101}\mu x + \frac{100^2}{101^2}x^2 + \frac{100}{101^2}x^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot \frac{25}{101}}\left(\left(\mu - \frac{100}{101}x\right)^2 + \frac{100}{101^2}x^2\right)} \mathrm{d}\mu\\ &= \frac{1.98}{10\pi} \int_{-\infty}^{\infty} e^{-\frac{1}{2\cdot \frac{25}{101}}\left(\mu - \frac{100}{101}x\right)^2 - \frac{2}{101}x^2} \mathrm{d}\mu\\ &= \frac{1.98}{\sqrt{2\pi}\sqrt{101}} e^{-\frac{2}{101}x^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\frac{5}{\sqrt{101}}} e^{-\frac{1}{2\cdot\left(\frac{5}{\sqrt{101}}\right)^2}\left(\mu - \frac{100}{101}x\right)^2} \mathrm{d}\mu\end{aligned}$$

Since the integral is the integral of a $\mathcal{N}\left(x, \frac{5}{\sqrt{101}}\right)$ distribution, we know that it must integrate to 1. Using this knowledge we can determine the probability density function for $\Pr(Z = 0, X = x)$.

$$\Pr(Z = 0, X = x) = \frac{1.98}{\sqrt{202\pi}} e^{-\frac{2}{101}x^2}$$

Substituting these likelihoods into the original equation gives us the likelihood for $\Pr(Z = 1 \mid X = x)$. We can then evaluate this for $x = 9$ to obtain the required result.

$$\begin{aligned}\Pr(Z = 1 \mid X = x) &= \frac{\frac{0.01}{\pi(1+x^2)}}{\frac{0.01}{\pi(1+x^2)} + \frac{1.98}{\sqrt{202\pi}} e^{-\frac{2}{101}x^2}}\\ \Pr(Z = 1 \mid X = 9) &= \frac{\frac{0.01}{82\pi}}{\frac{0.01}{82\pi} + \frac{1.98}{\sqrt{202\pi}} e^{-\frac{162}{101}}}\\ &\approx 0.00245\end{aligned}$$

This result seemed surprising – I was expecting a higher probability. However, when I verified it experimentally (using the code below) the results obtained were very similar and after 500 million generated variables suggested that $\Pr(Z = 1 \mid X = 9) \approx 0.00245016$.

```python
n = 5000000
k = 100
# I split work up so data fits in RAM and we can get a progress bar
x = 9
total = 0
for i in tqdm.tqdm(range(k)):
    total += np.mean(stats.norm.pdf(9, stats.norm.rvs(0, 5, n), 0.5))
pbar = total / k
print(0.01*stats.cauchy.pdf(9) / (0.01*stats.cauchy.pdf(9) + 0.99*pbar))
```

15. I have a coin, which might be biased. I toss it n times and get x heads. I am uncertain whether or not the coin is biased. Let $m \in \{\text{fair}, \text{biased}\}$ indicate which of the two cases is correct; and if it is biased let θ be the probabiilty of heads. The probability of x heads is thus

$$\Pr(x \mid m, \theta) = \begin{cases} \binom{n}{x} \theta^x (1-\theta)^{n-x} & \text{if } m = \text{biased} \\ \binom{n}{x} \frac{1}{2}^x \left(1 - \frac{1}{2}\right)^{n-x} & \text{if } m = \text{unbiased} \end{cases}$$

As a Bayesian I shall represent my uncertainty about m with a prior distribution, $\Pr_M(\text{fair}) = p$, $\Pr_M(\text{biased}) = 1 - p$. If it is biased, my prior belief is that the probability of heads is $\Theta \sim \mathcal{U}[0, 1]$.

(a) Write down the prior distribution for the pair (M, Θ), assuming independence as usual.

$$\Pr_{M,\theta}(m, \theta) = p \cdot 1_{m=\text{fair}} + (1-p) \cdot 1_{m=\text{biased}}$$

(b) Find the posterior distribution of (M, Θ) given x.

We can work out the probability $\Pr_X(x)$ using integration by parts:

$$\begin{aligned}
\Pr(x) &= \binom{n}{x} \int_0^1 \theta^x (1-\theta)^{n-x} \mathrm{d}\theta \\
&= \binom{n}{x} \frac{1}{n+1-x} \left[\theta^x (1-\theta)^{n+1-x}\right]_0^1 + \binom{n}{x} \frac{x}{n+1-x} \int_0^1 \theta^{x-1} (1-\theta)^{n+1-x} \mathrm{d}\theta \\
&= \binom{n}{x} \frac{x}{n+1-x} \int_0^1 \theta^{x-1} (1-\theta)^{n+1-x} \mathrm{d}\theta \\
&= \dots \\
&= \binom{n}{x} \frac{x!(n-x)!}{n!} \int_0^1 (1-\theta)^n \mathrm{d}\theta \\
&= -\frac{n!x!(n-x)!}{n!x!(n-x)!} \frac{1}{n+1} \left[(1-\theta)\right]_0^1 \\
&= \frac{1}{n+1}
\end{aligned}$$

Therefore we have the prior probability for $\Pr(x)$.

We can use this alongside the prior distribution for the pair (M, Θ) to work out the posterior probability $\Pr(M, \Theta \mid x)$ using Bayes rule.

$$\begin{aligned}
\Pr(M, \Theta \mid x) &= \frac{\Pr(M, \Theta)\Pr(x \mid M, \Theta)}{\Pr(x)} \\
&= (n+1)\binom{n}{x}\left(p \cdot 1_{m=\text{fair}}\frac{1}{2}^x\left(1-\frac{1}{2}\right)^{n-x} + (1-p)1_{m=\text{biased}}\theta^x\left(1-\theta\right)^{n-x}\right) \\
&= (n+1)\binom{n}{x}\left(\frac{p}{2^n}\cdot 1_{m=\text{fair}} + (1-p)\theta^x\left(1-\theta\right)^{n-x}\cdot 1_{m=\text{biased}}\right)
\end{aligned}$$

16. (a) Suppose we have a single observation x, drawn from a $\mathcal{N}(\mu+\nu, \sigma^2)$, where μ and ν are unknown parameters and σ^2 is known. Explain why the maximum likelihood estimates for μ and ν are non-identifiable.

Since μ and ν only occur in the distribution as $\mu+\nu$, when we find maximum likelihood estimates; we are actually only finding a maximum-likelihood estimator $\widehat{\mu+\nu}$. Individually the features have no meaning – if we increase $\hat{\mu}$ by 1 and decrease $\hat{\nu}$ by 1; the resulting estimator for $\widehat{\mu+\nu}$ is unchanged. Since the parameters have no individual meaning, we call them non-identifiable.

(b) For μ use $\mathcal{N}(\mu_0, \rho_0^2)$ as prior and for ν use $\mathcal{N}(\nu_0, \rho_0^2)$ where μ_0, ν_0 and ρ_0 are known. For the posterior density of (μ, ν). Calculate the parameter values $(\hat{\mu}, \hat{\nu})$ where the posterior density is maximum.

We can use Bayes Rule to work out the probability of (μ, ν) given x. We then find the maxima of the (logarithmic) probability by partially differentiating with respect to each parameter and solving the simultaneous equations.

$$\begin{aligned}
\Pr(\mu, \nu) &= \frac{\Pr(\mu)\Pr(\nu \mid \mu)\Pr(x \mid \mu, \nu)}{\Pr(x)} \\
&= \kappa \Pr(\mu)\Pr(\nu)\Pr(x \mid \mu, \nu) \\
&= \kappa \frac{1}{\sqrt{2\pi}\rho_0} e^{-\frac{(\mu-\mu_0)^2}{2\rho_0^2}} \frac{1}{\sqrt{2\pi}\rho_0} e^{-\frac{(\nu-\nu_0)^2}{2\rho_0^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu-\nu)^2}{2\sigma^2}} \\
&= \kappa e^{-\frac{(\mu-\mu_0)^2}{2\rho_0^2}} e^{-\frac{(\nu-\nu_0)^2}{2\rho_0^2}} e^{-\frac{(x-\mu-\nu)^2}{2\sigma^2}} \\
\ln \Pr(\mu, \nu) &= \ln \kappa - \frac{(\mu-\mu_0)^2}{2\rho_0^2} - \frac{(\nu-\nu_0)^2}{2\rho_0^2} - \frac{(x-\mu-\nu)^2}{2\sigma^2}
\end{aligned}$$

We must now partially differentiate with respect to each parameter and solve the simultaneous equations:

$$\begin{aligned}
\frac{\partial \ln \Pr(\mu, \nu)}{\partial \mu} &= -\frac{\mu-\mu_0}{\rho_0^2} - \frac{\nu+\mu-x}{\sigma^2} \\
0 &= -\frac{\hat{\mu}-\mu_0}{\rho_0^2} - \frac{\hat{\nu}+\hat{\mu}-x}{\sigma^2} \\
\frac{\partial \ln \Pr(\mu, \nu)}{\partial \nu} &= -\frac{\nu-\nu_0}{\rho_0^2} - \frac{\nu+\mu-x}{\sigma^2} \\
0 &= -\frac{\hat{\nu}-\nu_0}{\rho_0^2} - \frac{\hat{\nu}+\hat{\mu}-x}{\sigma^2}
\end{aligned}$$

Equating these expressions gives:

$$\begin{aligned}
-\frac{\hat{\mu}-\mu_0}{\rho_0^2} - \frac{\hat{\nu}+\hat{\mu}-x}{\sigma^2} &= -\frac{\hat{\nu}-\nu_0}{\rho_0^2} - \frac{\hat{\nu}+\hat{\mu}-x}{\sigma^2} \\
\hat{\mu}-\mu_0 &= \hat{\nu}-\nu_0 \\
\hat{\mu} &= \hat{\nu}-\nu_0+\mu_0
\end{aligned}$$

Substituting this value of $\hat{\mu}$ into the equation for $\hat{\nu}$ gives:

$$\begin{aligned}
0 &= -\frac{\hat{\nu}-\nu_0}{\rho_0^2} - \frac{\hat{\nu}+\hat{\nu}-\nu_0+\mu_0-x}{\sigma^2} \\
0 &= -\hat{\nu}\left(\sigma^2+2\rho_0^2\right)+\nu_0\left(\sigma^2+\rho_0^2\right)-\rho_0^2\mu_0+\rho_0^2 x \\
\hat{\nu} &= \frac{\nu_0\left(\sigma^2+\rho_0^2\right)+\rho_0^2 x-\rho_0^2\mu_0}{\sigma^2+2\rho_0^2}
\end{aligned}$$

Substituting this into the equation for $\hat{\mu}$ gives:

$$\hat{\mu} = \frac{\mu_0\left(\sigma^2+\rho_0^2\right)+\rho_0^2 x-\rho_0^2\nu_0}{\sigma^2+2\rho_0^2}$$

(c) An Engineer friend tells you “Bayesianism is the Apple of inference. You just work out the posterior and everything Just Works$^{\mathrm{TM}}$ and you don’t need to worry about irritating things like non-identifiability”. What do you think?

Non-identifiability must still be resolved in Bayesianism to get worthwhile results for three reasons:

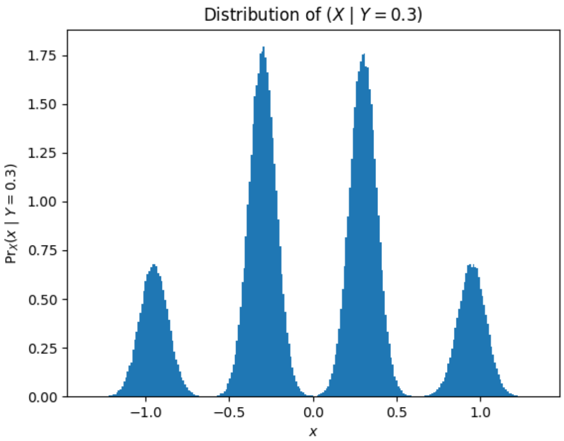
- If the features we use are non-identifiable then the posterior distributions we create will be uninterpretable – which means the model is useless.
- When we use Bayesianism we have to create prior distributions. If the features are non-identifiable then creating meaningful priors is much more difficult – we cannot use intuition and must instead reason about their values when combined with other features. This will decrease the quality of the priors chosen.
- Bayesianism is computationally expensive. If we increase the number of features and decrease the quality of our prior distributions then computing the posterior distributions will become computationally intractable very quickly. If we do not invest sufficient compute then the posteriors will not converge or will have such low resolution that we cannot reason about the emergent features we do care about.

17. Here’s my answer to question 1:

```
k = np.random.choice(4, p=[.6, .3, .05, .05], size=n)
t = np.random.uniform(size=n)
x = np.column_stack([...])
y = np.column_stack([...])
xy = np.column_stack([x[np.arange(n), k], y[np.arange(n), k]])
xy = np.random.normal(loc=xy, scale=.08)
```

Compute the distribution of $(X \mid Y = 0.3)$. Give your answer as a histogram.

```
# remove the last line of Damons code then add the following:
probs = stats.norm.pdf(0.3, xy[:, 1], 0.08)
xs = stats.norm.rvs(xy[:, 0], 0.08)
plt.hist(xs, weights=probs, density=True, bins=...)
```



3 2019 Paper 6 Question 7

https://www.cl.cam.ac.uk/teaching/exams/pastpapers/y2019p6q7.pdf

(a) Let $X_1, \ldots, x_n$ be independent binary random variables, $\mathbb{P}(X_i = 1) = \theta$, $\mathbb{P}(X_i = 0) = 1 - \theta$ for some unknown parameter θ. Using $\mathcal{U}[0,1]$ as the prior distribution for θ, find the posterior distribution.

This is a standard result: the posterior distribution for $(\theta|X)$ is:

$$\beta\left(1 + \sum_{i=1}^{n} X_i, 1 + n - \sum_{i=1}^{n} X_i\right)$$

We can find the posterior distribution using Bayes rule:

$$\begin{aligned}\Pr(\theta \mid X) &= \frac{\Pr(\theta)\Pr(X_1, \ldots, X_n \mid \theta)}{\Pr(X_1, \ldots, X_n)} \\ &= \kappa \cdot 1 \cdot \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n - \sum_{i=1}^{n} X_i} \\ &= \kappa' \frac{n+1}{n - \sum_{i=1}^{n} X_i} \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n - \sum_{i=1}^{n} X_i}\end{aligned}$$

The distribution excluding κ' is the probability density function for a β distribution with parameters $(1 + \sum_{i=1}^{n} X_i, 1 + n - \sum_{i=1}^{n} X_i)$.

Since both the whole distribution integrates to 1 over the range $\theta \in [0,1]$ and $\frac{n+1}{n - \sum_{i=1}^{n} X_i}\theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i}$ integrates to 1 over the range $[0,1]$, we can conclude that $\kappa' = 1$ and therefore the posterior distribution is $\beta\left(1 + \sum_{i=1}^{n} X_i, 1 + n - \sum_{i=1}^{n} X_i\right)$.

(b) I have collected a dataset of images, and employed an Amazon Mechanical Turk worker to label them. the labels are binary, nice or nasty. To assess how accurate the worker is, I first picked 30 validation images at random, found the true label myself and compared to the worker's label. The worker was correct on 25 and incorrect on 5.

Let θ be the probability that the worker labels an image incorrectly. Using $\beta(0.1, 0.5)$ as the prior distribution for θ, find the posterior.

Let 1_i be the indicator variable for whether the worker labelled the i^{th} image correctly.

We can find the posterior distribution using Bayes rule:

$$\begin{aligned}
\Pr(\theta \mid 1_1, \dots, 1_{30}) &= \frac{\Pr(\theta)\Pr(1_1, \dots, 1_{30} \mid \theta)}{\Pr(1_1, \dots, 1_{30})} \\
&= \kappa\theta^{0.1-1}(1-\theta)^{0.5-1} \cdot \theta^5(1-\theta)^{25} \\
&= \kappa\theta^{5.1-1}(1-\theta)^{25.5-1} \\
&= \kappa' \binom{29.6}{4.1} \theta^{5.1-1}(1-\theta)^{25.5-1}
\end{aligned}$$

The equation excluding κ' is the probability density function for a beta distribution $\beta\,(5.1, 25.5)$.

Since both the posterior probability and the β component integrate to 1 over the range $\theta \in [0, 1]$; we can conclude that $\kappa' = 1$. Therefore the posterior probability distribution is equal to the probability density function for $\beta\,(5.1, 25.5)$ and therefore the posterior probability has a beta distribution parameters $(5.1, 25.5)$.

(c) I next ask the worker to label a new test image, and they tell me the image is nice. Let $z \in \{\text{nice}, \text{nasty}\}$ be the true label and let the prior distribution for z be $\Pr(\text{nice}) = 0.1$, $\Pr(\text{nasty}) = 0.9$.

For both $z = \text{nice}$ and $z = \text{nasty}$, find:

$$\mathbb{P}(\text{worker says nice} \mid z, \theta)$$

Hence find the posterior distribution of (z, θ). Your answer may be left as an unnormalised density function.

$$\begin{gathered}
\mathbb{P}(\text{worker says nice} \mid \text{nice}, \theta) = 1 - \theta \\
\mathbb{P}(\text{worker says nice} \mid \text{nasty}, \theta) = \theta \\
\mathbb{P}(\text{worker says nice} \mid z, \theta) = 1_{z=\text{nice}}\,(1-\theta) + 1_{z=\text{nasty}}\theta
\end{gathered}$$

We can find the posterior distribution of (z, θ) using Bayes rule:

$$\begin{aligned}
\Pr(z, \theta \mid \text{worker says nice}, 1_i, \dots, 1_{30}) &= \frac{\Pr(z, \theta)\Pr(\text{worker says nice}, 1_i, \dots, 1_{30} \mid z, \theta)}{\Pr(\text{worker says nice}, 1_i, \dots, 1_{30})} \\
&= \kappa\,(0.1 \cdot 1_{z=\text{nice}}(1-\theta) + 0.9 \cdot 1_{z=\text{nasty}}\theta)\,\theta^{5.1-1}(1-\theta)^{25.5-1}
\end{aligned}$$

(d) Find the posterior distribution of z:

We can find the posterior distribution for z by integrating the posterior distribution for (z, θ) with respect to θ between 0 and 1.

Let $\Pr(\theta, \alpha, \beta)$ be the pdf for the beta distribution with parameters (α, β) evaluated at θ.

$$\begin{aligned}
&\Pr(z \mid \text{worker says nice}, 1_1, \ldots, 1_{30}) \\
&= \int_0^1 \Pr(z, \theta \mid \text{worker says nice}, 1_i, \ldots, 1_{30}) \mathrm{d}\theta \\
&= \int_0^1 \kappa \left(0.1 \cdot 1_{z=\text{nice}}(1-\theta) + 0.9 \cdot 1_{z=\text{nasty}}\theta\right) \theta^{5.1-1}(1-\theta)^{25.5-1} \mathrm{d}\theta \\
&= \int_0^1 \kappa \left(0.1 \cdot 1_{z=\text{nice}}\theta^{5.1-1}(1-\theta)^{26.5-1} + 0.9 \cdot 1_{z=\text{nasty}}\theta^{6.1-1}(1-\theta)^{25.5-1}\right) \mathrm{d}\theta \\
&= \int_0^1 0.1\kappa \binom{30.6}{4.1}^{-1} 1_{z=\text{nice}} \Pr(\theta, 5.1, 26.5) + 0.9\kappa \binom{30.6}{5.1}^{-1} 1_{z=\text{nasty}} \Pr(\theta, 6.1, 25.5) \mathrm{d}\theta \\
&= 0.1\kappa \binom{30.6}{4.1}^{-1} 1_{z=\text{nice}} \int_0^1 \Pr(\theta, 5.1, 26.5) \mathrm{d}\theta + 0.9\kappa \binom{30.6}{5.1}^{-1} 1_{z=\text{nasty}} \int_0^1 \Pr(\theta, 6.1, 25.5) \mathrm{d}\theta
\end{aligned}$$

Since the integral of any beta distribution between 0 and 1 is 1, both the integrals evaluate to 1. We substitute this in and then absorb some more terms into κ. This gives the posterior probability:

$$0.1\kappa \binom{30.6}{4.1}^{-1} 1_{z=\text{nice}} + 0.9\kappa \binom{30.6}{5.1}^{-1} 1_{z=\text{nasty}}$$

Since these probabilities must sum to 1, we can conclude that:

$$\kappa = \frac{\binom{30.6}{5.1} \cdot \binom{30.6}{4.1}}{0.1 \binom{30.6}{5.1} + 0.9 \binom{30.6}{4.1}}$$

Therefore the posterior probability for z is given by:

$$\Pr(z \mid \text{data}) = 0.1 \frac{\binom{30.6}{5.1}}{0.1 \binom{30.6}{5.1} + 0.9 \binom{30.6}{4.1}} 1_{z=\text{nice}} + 0.9 \frac{\binom{30.6}{4.1}}{0.1 \binom{30.6}{5.1} + 0.9 \binom{30.6}{4.1}} 1_{z=\text{nasty}}$$

This equates to a probability of 35.7% that the image is nice and 64.3% that the image is nasty. This passes the sanity check – the worker has shown that it is often correct and therefore the worker saying an image is nice would mean the posterior probability of that image being nice should be larger than the prior probability. Additionally, if the robot had been more correct, the posterior probability for nice worked out using this method would have been higher – and if it had been less correct it would be lower.

(e) My colleague has more grant money and she can employ 3 workers to rate each image. On a test set of 30 images, she found that they all agreed on 15 images, worker 1 was the odd one out on 8 of the images, worker 2 was the odd one out on 4, and worker 3 was the odd one out on 3.

Let θ_i be the probability that worker i labels an image incorrectly. Find the posterior distribution of $(\theta_1, \theta_2, \theta_3)$. Your answer may be left as an un-normalised density function.

Let data be the described observations.

We can work out the posterior probability using Bayes' rule:

$$\Pr(\theta_1, \theta_2, \theta_3 \mid \text{data}) = \frac{\Pr(\theta_1, \theta_2, \theta_3)\Pr(\text{data} \mid \theta_1, \theta_2, \theta_3)}{\Pr(\text{data})}$$

The probability of observing the data we have seen is equal to the product of the probabilities of the observations. So if worker 1 and 2 guessed the same but 3 disagreed the probability of that observation would be $\theta_1\theta_2(1-\theta_3)+(1-\theta_1)(1-\theta_2)\theta_3$. Therefore the total probability of observing all the data given $\theta_1, \theta_2, \theta_3$ is:

$$\begin{aligned}\Pr(\text{data} \mid \theta_1, \theta_2, \theta_3) =& (\theta_1\theta_2\theta_3 + (1-\theta_1)(1-\theta_2)(1-\theta_3))^{15}\times \\ & ((1-\theta_1)\theta_2\theta_3 + \theta_1(1-\theta_2)(1-\theta_3))^{8}\times \\ & (\theta_1(1-\theta_2)\theta_3 + (1-\theta_1)\theta_2(1-\theta_3))^{4}\times \\ & (\theta_1\theta_2(1-\theta_3) + (1-\theta_1)(1-\theta_2)\theta_3)^{3}\end{aligned}$$

Denote the above probability $p(\theta_1, \theta_2, \theta_3)$

Therefore, using Bayes rule, the posterior probability is equal to:

$$\begin{aligned}\Pr(\theta_1, \theta_2, \theta_3 \mid \text{data}) &= \frac{\Pr(\theta_1, \theta_2, \theta_3)\Pr(\text{data} \mid \theta_1, \theta_2, \theta_3)}{\Pr(\text{data})} \\ &= \kappa\theta_1^{-0.9}(1-\theta_1)^{0.5}\theta_2^{-0.9}(1-\theta_2)^{0.5}\theta_3^{-0.9}(1-\theta_3)^{0.5}p(\theta_1, \theta_2, \theta_3)\end{aligned}$$