Practical challenge 2

Inference

Foundations of Data Science—DJW—2018/2019

These questions are optional. They are not intended for supervision, and they are not examinable material. They are to give you practical experience of work based on the theoretical material covered in lectures, if you want it, and to reinforce your skills in scientific computing.

If you want guidance, you can (i) ask your fellow students for help in the Moodle forum, (ii) come to the practical classes in the second half of term and ask demonstrators, (iii) wait until model solutions are released, after the practical classes.

Question 1. This question asks you to investigate racial bias in police stop-and-search behaviour. We will restrict attention to records with police_force='cambridgeshire'. We will work with the model

$$\mathbb{P}(Y_i = \mathsf{find}) = \theta_{e_i}$$

where $Y_i \in \{\mathsf{find}, \mathsf{nothing}\}$ is the outcome listed for record i, e_i is the ethnicity, and

$$\theta = \Big(\theta_{\mathsf{Asian}}, \theta_{\mathsf{Black}}, \theta_{\mathsf{Mixed}}, \theta_{\mathsf{Other}}, \theta_{\mathsf{White}}\Big)$$

is an unknown parameter.

```
!wget "https://teachingfiles.blob.core.windows.net/founds/stop-and-search.csv"
police = pandas.read_csv('stop-and-search.csv')
ok = ~pandas.isnull(police['Officer-defined ethnicity']) & \
      (police['police_force'] == 'cambridgeshire')
y = police.loc[ok, 'Outcome'] = 'Nothing found - no further action'
ETHNICITY_LEVELS = ['Asian', 'Black', 'White', 'Mixed', 'Other']
ethnicity_code = {k:i for i,k in enumerate(ETHNICITY_LEVELS)}
e = np.array([ethnicity_code[v] for v in police.loc[ok, 'Officer-defined ethnicity']])
```

(a) As a prior distribution, let θ consist of 5 independent random variables drawn from $\mathrm{Beta}(\delta, \delta)$ where $\delta = 0.5$. Calculate the posterior distribution. Implement a function posterior_sample(size) that generates size independent samples of θ drawn from the posterior distribution. Each sample should be a vector of length 5.

(b) Given a sample of θ, define the overall bias score to be

$$d(\theta) = \max_{e,e'} |\theta_e - \theta_{e'}|.$$

Plot a histogram of the posterior distribution of $d(\theta)$.

(c) Repeat part (b) but for $d_3(\theta)$ instead, which is defined like $d(\theta)$ but restricted to $e, e' \in \{\mathsf{Asian}, \mathsf{Black}, \mathsf{White}\}$. Explain why the two histograms have very different shapes.

(d) Find a Bayesian 95% confidence interval for $d_3(\theta)$ of the form

$$\mathbb{P}\Big(d_3(\theta) \leq c\Big) = 95\%.$$

(e) Find a 95% frequentist confidence interval for $d_3(\theta)$ of the same form as in part (d). Explain your resampling strategy.

(f) Consider testing the hypothesis $\theta_{\mathsf{Black}} = \theta_{\mathsf{Asian}} = \theta_{\mathsf{White}}$. Let the test statistic be $d_3(\hat\theta(x))$, where $\hat\theta(x)$ is the maximum likelihood estimate for θ given dataset x. Plot a histogram showing the distribution of this test statistic assuming that the hypothesis is true. Explain your resampling strategy. Do you accept the hypothesis?

The next question is about computational Bayesian inference. We typically want to report readouts on the posterior distribution $\Pr(\theta\,|\,x) \propto \Pr(\theta) f(\theta)$*, where* $f(\theta) = \Pr(data\,|\,\theta)$*. When we can't solve this analytically, we can compute with it numerically, as follows. First, draw a random sample* $\{\theta_1, \theta_2, \ldots, \theta_n\}$ *from the prior distribution. Second, attach weights to each value in the sample, to get the weighted empirical distribution*

$$\Big\{(\theta_1, w_1), (\theta_2, w_2), \ldots, (\theta_n, w_n)\Big\} \qquad \textit{where } w_i = \frac{f(\theta_i)}{\sum_j f(\theta_j)}.$$

Third, compute whatever probability we're interested in by

$$\mathbb{P}(\theta \in A \mid data) \approx \sum_{i=1}^{n} w_i 1_{\theta_i \in A}, \qquad \mathbb{E}\big(h(\theta) \mid data\big) \approx \sum_{i=1}^{n} w_i h(\theta_i).$$

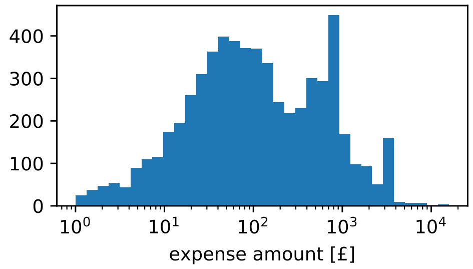
This is like working with a standard empirical distribution, but using weights w_i *rather than giving every observation equal weight* $1/n$*.*

Question 2. This question asks you to look for anomalies in a subset of MP expense claims. Informally, an anomaly is something that lies outside the usual distribution. Formally, we can think of values in the dataset as drawn from a mixture of two distributions, one distribution for the usual state of affairs, another for anomalies. We can express this as

$$\Pr_X(x \mid \xi, \theta, \phi) = \xi \Pr_{\text{usual}}(x \mid \theta) + (1 - \xi) \Pr_{\text{anom}}(x \mid \phi)$$

where $\xi \in [0,1]$ is the probability that the observation is usual, and θ and ϕ respectively parameterize the usual and anomalous distributions.

```
!wget "https://teachingfiles.blob.core.windows.net/founds/expense.csv"
expense = pandas.read_csv('expense.csv',
                          dtype={'Reason If Not Paid': np.str_})  # type hint
x = expense.loc[(expense['Year']=='18_19') & (expense['Category']=='Office Costs'),
                'Amount Claimed'].values
plt.hist(x[x>0], bins=10**np.linspace(.0001, 4.2, 35))
plt.gca().set_xscale("log")
```



For this dataset, consider the probability model

$$\Pr_{\text{usual}}(x \mid \mu, \sigma) = \frac{1_{x>0}}{x\sqrt{2\pi\sigma^2}} e^{-(\log x - \mu)^2 \,/\, 2\sigma^2}$$
$$\Pr_{\text{anom}}(x \mid \alpha) = \frac{\alpha}{2(|x|+1)^{\alpha+1}}$$

and use the following prior distribution on the unknown parameters:

```
def rprior(size=1):
    ξ = np.random.beta(a=9, b=1, size=size)
    μ = np.random.normal(np.log(10**1.4), scale=np.log(10), size=size)
    σ = np.log(6) * np.ones(size)
    α = 1 * np.ones(size)
    return np.column_stack([ξ,μ,σ,α])
```

(a) What are the common names for $\Pr_{\text{usual}}$ and $\Pr_{\text{anom}}$?

(b) Find the posterior distribution of the unknown parameters.

(c) Find the posterior predictive probability $\mathbb{P}_{\text{usual}}(X \le x)$. Differentiate to get the density function, and superimpose it on a histogram of the data.