Example sheet 2

Inference

Foundations of Data Science—DJW—2017/2018

Practical work is *not* assessed as coursework, and there is no submission deadline (apart from whatever your supervisor sets), and the practical session is optional. However, the topics and methods in the practical questions are examinable material.

- You are not required to answer every question—but you are expected to spend around 6 solid hours per example sheet, including the time you spend reviewing notes and attending the practical classes. Attempt Question 5 last.
- These questions are verbose, but you should be able to answer them with no more than 12 lines of code each. In inference, asking a question the right way is often harder than answering it. You answers should include clear statements of what you have found, not just code and numbers.
- There is an accompanying Python notebook `ex2.ipynb`, which contains code snippets that you may find useful (though if you wish to use another programming language you are welcome to do so). You can find the link on the course webpage.
- In a Jupyter Notebooks, you can run your code cells in any order. For the sanity of your supervisor, before saving your work, please use Kernel | Restart & Clear Output and re-run your code in the order it appears in the notebook. This will uncover bugs that arise from assigning a variable in one place and using it earlier in the notebook.
- Do not waste time wrestling with Python; instead, post questions at the `allanswered.com` forum, linked to on the course webpage.

For supervisors: This example sheet should be supervised some time after 30 October. Model answers can be found on the course webpage.

In the first four questions you will investigate racial bias in police stop-and-search behaviour. You will make inferences, and quantify your uncertainty about those inferences. The dataset is `https://teachingfiles.blob.core.windows.net/founds/stop-and-search.csv`, and 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 row i, e_i is the ethnicity, and

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

is an unknown parameter.

Question 1 (Bayesian confidence interval).

(a) Let θ consist of 5 independent random variables drawn from the Beta(δ, δ) distribution, where $\delta = 0.5$. Calculate the posterior distribution of θ. 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 maximum discrepancy to be

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

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

(c) You should have found that the posterior distribution of $d(\theta)$ is highly variable, because the dataset has few cases of $e_i = \mathsf{Mixed}$ and none of $e_i = \mathsf{Other}$. Plot the histogram again, but showing only the maximum discrepancy of $(\theta_{\mathsf{Asian}}, \theta_{\mathsf{Black}}, \theta_{\mathsf{White}})$. We'll call this $d_3(\theta)$.

(d) Find a 95% confidence interval of the form

$$\mathbb{P}\big(d_3(\theta) < c\big) = 95\%$$

Question 2 (Frequentist confidence interval).

(a) Give a formula for the maximum likelihood estimator $\hat{\theta}$.

(b) Use the bootstrap resampling method to estimate, given $\varepsilon > 0$, the probability

$$\mathbb{P}\big(d_3(\theta) \in [d_3(\hat{\theta}) - \varepsilon,\, d_3(\hat{\theta}) + \varepsilon]\big)$$

(c) Find a 95% confidence interval for $d_3(\theta)$.

Question 3 (Frequentist hypothesis testing). In science and in policy making, it is often useful to frame questions in the following way. "My default hypothesis is H_0. I'm planning a data-gathering exercise, and based on the data X I gather I might stick with H_0 or I might reject it." In programming terms, the data scientist has to define a hypothesis-testing function $\mathsf{reject\_H0}(x)$ which returns either True or False.

We don't want $\mathsf{reject\_H0}(x)$ to return True when H_0 actually is the case. For example, H_0 might be "There is no racial bias" and the police commissioner does not want to spend money to correct bias if H_0 is indeed the case. It's typically impossible to know for certain from the observed data whether H_0 is actually the case or not, but what we can do instead is set a probability threshold e.g.

$$\mathbb{P}\big(\mathsf{reject\_H0}(X) = \mathsf{True}\big) \leq 5\% \text{ if } H_0 \text{ is true.}$$

A common way to implement $\mathsf{reject\_H0}$ is to invent some real-valued test statistic $T(x)$, and define

```
def reject_H0(x):
    if T(x) > thresh:
        return True
    else:
        return False
```

Whatever function T is chosen, thresh should be set so that $\mathbb{P}(T(X) > \mathsf{thresh}) \leq 5\%$ if H_0 is true. We should try to design T such that $\mathbb{P}(T(X) > \mathsf{thresh})$ is large if there truly is bias.

(a) Let H_0 be the hypothesis $\theta_{\mathsf{Asian}} = \theta_{\mathsf{Black}} = \theta_{\mathsf{White}}$, and let $X = (F_{\mathsf{Asian}}, F_{\mathsf{Black}}, F_{\mathsf{White}})$ where $F_e \in \{0, 1, \ldots, n_e\}$ is the number of outcomes where $Y_i = \mathsf{find}$ among members of group e, and n_e is the number of individuals in that group. If H_0 is true, then we can generate a resampled version X^* by sampling three groups of individuals from the pooled population of $n_{\mathsf{Asian}} + n_{\mathsf{Black}} + n_{\mathsf{white}}$ individuals from any of those three groups. Implement a function $\mathsf{X\_star}(n)$ that generates n samples of X^*.

(b) Let the test statistic be $T(x) = d_3(\hat{\theta}(x))$, where $\hat{\theta}(x)$ is the maximum likelihood estimator of θ when the observed data is x. Find the threshold value thresh such that

$$\mathbb{P}\big(T(X^*) > \mathsf{thresh}\big) = 5\%.$$

Given the actual data we observed, do we reject H_0 i.e. does $\mathsf{reject\_H0}()$ return True?

(c) Another way to express the output of a hypothesis test is as a *significance level* p. This is defined as

$$p = \mathbb{P}\big(T(X^*) > T(x)\big)$$

where x is the actual observed data. (This has the property that $p < 5\%$ if and only if $T(x) > \mathsf{thresh}$.) Compute the significance level for this dataset.

Question 4 (Natural parameters). We might have a prior belief that the coefficients of θ are very close to each other, but we might have no idea what that common value is. In Question 3 we expressed this belief through our resampling method, drawing outcomes from the pooled population.

Invent a Bayesian prior distribution for θ that expresses this same belief. Your distribution should have the property that any θ in $[0,1]^5$ is possible, but that those θ with small $d(\theta)$ are more likely. Generate samples from this distribution and plot the joint distribution of $(\theta_{\text{Asian}}, \theta_{\text{Black}})$. *Hint. Generate suitable random variables in $\mathbb{R}^5$, and then transform them to values in the range $[0,1]$.*

In Part II Machine Learning and Bayesian Inference, you will learn how to generate samples from an arbitrary Bayesian posterior distribution.

Question 5 (Model selection). A data scientist might want to know which of two models is better. In this question we will compare two different ways of answering this. We will do so in a controlled situation: we'll simulate the data ourselves, so we know exactly which model is true.

The true model (unknown to the data scientist analysing the data) is

$$Y \sim \text{Normal}(5 + 3x_1 + 0.1x_2, 1).$$

The dataset at `https://teachingfiles.blob.core.windows.net/founds/model_selection_sample.csv` was generated from this distribution.

(a) As a data scientist, you have been given the dataset, and you believe the underlying model is either

$$\text{Model A:} \qquad Y_i \sim \text{Normal}(\alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i}, \sigma^2)$$

or

$$\text{Model B:} \qquad Y_i \sim \text{Normal}(\alpha + \beta x_{1,i}, \sigma^2).$$

For each model, find the maximum likelihood estimators for all the parameters. *Hint. Let $\sigma^2 = e^\gamma$, and run the optimization over γ rather than σ or σ^2. This ensures $\sigma^2 > 0$ and it makes the log likelihood differentiable so that scipy.optimize.fmin runs nicely.*

(b) One way to select a model is to pick the model that gives better predictions. Normally the data scientist should set aside some of her data for validation, but in our case we have a black box that can synthesize as much validation data as we want. For every (x_1, x_2) on a grid of 10×10 points in the range $x_1 \in [-1.5, 1.5]$, $x_2 \in [1.5, 1.5]$, compute the expected value according to the fitted Model A, and also generate 1000 simulated values of Y at every grid point. Compute the mean square error

$$\text{MSE} = \text{average}\Big((\text{simulated} - \text{expected})^2\Big)$$

across all these 100×1000 values. Repeat the exercise for Model B. Which model gives the smaller mean square error? **[NOTE. The original example sheet erroneously said 10000 simulated values at every grid point, when it should have said 1000.]**

(c) Another way to select a model is by quantifying confidence in the fit. Models A and B differ only in that Model B sets $\beta_2 = 0$. A frequentist would ask "When I fit model A, what is the range of values of $\hat{\beta}_2$ that I might expect to see?" By resampling from the original dataset, generate samples of $\hat{\beta}_2$, and plot a histogram. Give a confidence interval for β_2. *Hint. See section 2.5 of the notes for two different ways you might resample.*

(d) How would a Bayesian say we should decide between the two models?

(e) Can you devise a hypothesis test for deciding between them?

This example is taken from *To explain or to predict?* by Galit Shmueli, 2010, `https://projecteuclid.org/download/pdfview_1/euclid.ss/1294167961`.