Example sheet 1

Models, distributions

Foundations of Data Science—DJW—2017/2018

This example sheet is based on material that will be lectured up to and including 16 October.

- Questions 1 and 2 mix pen-and-paper and practical work, and there are online notebooks with relevant Python code snippets (though Python is not compulsory). Please use the `allanswered.com` forum for sharing advice and answering each other's code-related questions, e.g. "how do I install Python?" or "why does this code fail?" or "what are we meant to implement here?" Links to notebooks and the forum can be found on the class webpage.
- At the practical session on 18 October, I will pick some interesting questions from the forum and discuss them. The 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.
- 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. It's up to you and your supervisor to select appropriate questions to focus on.

For supervisors: This example sheet should be supervised some time after 18 October. Model answers, including full code for the practical work, are online on the course webpage.

Question 1. Given a dataset $x_1, \dots, x_n$, the empirical distribution function is

$$\hat{F}(x) = \frac{1}{n}\bigl(\text{number of the } x_i \text{ that are } \geq x\bigr).$$

In this question you will investigate noise and uncertainty in the empirical distribution function, using a dataset of web server response times available at `https://teachingfiles.blob.core.windows.net/founds/weblog_sizes.txt`.

(a) Plot the empirical distribution of this dataset.

(b) When plotted on log-log axes, it seems that the empirical distribution might consist of two straight lines, i.e.

$$\log \hat{F}(x) \approx -\beta \log x - \gamma \max(\log x - \theta, 0).$$

Compute the maximum likelihood estimators $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\theta}$, for the unknown parameters in this function.

(c) A good way to assess uncertainty is to split the data. Split the dataset into 20 equal-sized pieces, and plot the empirical distribution of each piece, superimposed on the same chart. Show also the fitted distribution function

$$G(x) = \exp\bigl(-\hat{\beta} \log x - \hat{\gamma} \max(\log x - \hat{\theta}, 0)\bigr).$$

(d) Let X be a random variable with distribution G, and let $Y = 1_{X \geq x}$ for some fixed value of x. Calculate the mean and variance of Y.

(e) Let $\tilde{F}_n(x)$ be the empirical distribution function for a random sample of size n drawn from G. Show that $\mathbb{E}\tilde{F}_n(x) = G(x)$ for all x. Calculate $\operatorname{Var} \tilde{F}_n(x)$.

(f) Repeat the plot from (c), and superimpose a ribbon that shows, at every x, a 95% confidence interval for $\tilde{F}_m(x)$. You should choose m appropriately, so that the ribbon is informative about the 20 empirical distributions you plotted. Explain your choice.

(g) Would you expect roughly 19 out of the 20 empirical distribution plots (i.e. 95% of them) to lie entirely within the ribbon? Explain your reasoning.

Question 2. I took some measurements of wasp activity this summer, to learn if it was getting worse and whether I should take action. On three successive days, at the same time each day, I recorded the timestamp in seconds of every departure from the nest, over a 10 minute interval. The data is at `https://teachingfiles.blob.core.windows.net/founds/wasp.csv`.

(a) Plot the data in some appropriate way. For example, you could plot the number of wasps in every 10 second interval.

For the rest of this question, restrict attention to day 1. Let $x_1, \dots, x_n$ be wasp inter-departure times, i.e. the first wasp departs at time `t[0]`, the second at `t[1]=t[0]`$+x_1$, the third at `t[2]=t[1]`$+x_2$ and so on. In many counting processes found in nature, the inter-event times are independent. We will investigate whether this is the case for the wasp dataset. This is so that I can properly quantify the variability in wasp activity, which will be important when I come to analyse whether wasp activity is becoming significantly worse, but that is a matter for another time.

(b) Compute the mean μ and variance σ^2 of wasp inter-departure times. *Note: 'mean' and 'variance' are words for describing random variables, not datasets. You should interpret them in a way that makes sense.*

(c) Compute the quartiles of wasp inter-departure times; write q_{25} for the 25%ile etc.

(d) If inter-departure times were independent, then knowing one inter-departure time X_i would be uninformative about the subsequent inter-departure time X_{i+1}. Compute $\mu_{0:25} = \mathbb{E}(X_{i+1} \mid X_i < q_{25})$, and $\mu_{25:50} = \mathbb{E}(X_{i+1} \mid q_{25} \leq X_i < q_{50})$, and so on. Also compute $\operatorname{Var}(X_{i+1} \mid X_i < q_{25})$ and so on. Do your answers support the belief that wasp inter-departure times are independent?

(e) Let $y_i^m = x_i + \cdots + x_{i+m-1}$, the sum of m consecutive inter-departure times, and let Y^m be a typical value of this. If the inter-departure times were independent, then $\operatorname{Var} Y^m$ would be equal to $m\sigma^2$ where σ^2 is the variance you found in (b). Plot $\operatorname{Var} Y^m$, and superimpose $m\sigma^2$, both as a function of m. Do your answers support the belief that wasp inter-departure times are independent?

In (d) you used the data to compute $\mu_{0:25}$ and $\mu_{25:50}$, and found that $\mu_{0:25} - \mu_{25:50}$ was not precisely equal to zero. This is not suprising, because of inherent noise in the data. What matters is *how much* different from zero.

(f) Let $y_1, \dots, y_p$ be the inter-departure times that you averaged to obtain $\mu_{0:25}$, and let $z_1, \dots, z_q$ be the inter-departure times that you averaged to obtain $\mu_{25:50}$. Repeat the following resampling procedure 10,000 times:

- Pick p values sampled with replacement from the concatenated list $y_1, \dots, y_p, z_1, \dots, z_q$ and let $\mu^*_{0:25}$ be the average.
- Pick q values sampled with replacement and let $\mu^*_{25:50}$ be the average.
- Record $\mu^*_{0:25} - \mu^*_{25:50}$.

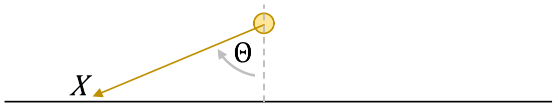
Plot a histogram of the 10,000 values you obtain. On your plot, mark the value $\mu_{0:25} - \mu_{25:50}$. Does this plot suggest that $\mu_{0:25}$ and $\mu_{25:50}$ are significantly different? *The ideas behind this will be studied in Section 3, but it's a good idea to explore them numerically before we study them formally.*

(g) How might you assess your confidence in your answer to (e)?

(h) Are wasp inter-departure times independent?

Question 3. Let U be a uniform random variable on $[0,1]$. Let $Y = U(1-U)$. Calculate $\mathbb{P}(Y \geq y)$, and hence find the density of Y.

Question 4. A point lightsource at coordinates $(0,1)$ sends out a ray of light at an angle Θ chosen uniformly in $[-\pi/2, \pi/2]$. Let X be the point where the ray intersects the horizontal line through the origin. What is the density of X?



X is known as the Cauchy distribution. It is unusual in that it has no mean.

Question 5. The variance of a random variable X is defined to be $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}X)^2$. Show that $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2$. *Note: by convention, $\mathbb{E}$ is taken to have lower precedence than multiplication and power, and higher precedence than addition and subtraction. So the expression of interest is* $(\mathbb{E}(X^2)) - (\mathbb{E}(X))^2$.

Question 6. (a) Let X and Y be random variables. Show that $\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}X$. (You can take X and Y to be discrete random variables, though the result is still true when they are continuous.)

(b) Let $N \sim \text{Poisson}(\lambda)$. Let X_i be a collection of independent $\text{Exp}(\mu)$ random variables. Let $S = \sum_{i=1}^{N} X_i$. Find the mean and variance of S.

Question 7. If $X_1, \ldots, X_n$ are independent samples from Uniform$[0, \theta]$, find the maximum likelihood estimator for θ. *Hint. When you write out the density function, make sure your function is correct for all values of x, including $x > \theta$.*

Question 8. Rewrite the random variable

```
def rxy($p$,$\sigma$):
    X = 0 with probability $1-p$, and 1 with probability $p$
    Y is Gaussian with mean X and variance $\sigma^2$
    return (X,Y)
```

so that your code first generates Y and then generates X conditional on Y. *(This type of model is used for finding clusters in data. There might be a small number of clusters (here we have two, at 0 and 1), and each observation is located near a cluster. A typical task is to to assign each observation to its cluster.)*

Question 9. A recent paper, *Fairness testing: testing software for discrimination* by Galhotra, Brun and Meliou, in ESEC/FSE 2017, defines discrimination as follows. Consider a piece of software that takes several features as inputs (e.g. name, age, race, employment history, level of education, gender), and produces yes/no output (e.g. whether to grant bail). They say it is 'causally unfair' with respect to a particular feature like race, if there are two feature vectors differing only by race which produce different outputs. They go on to define a 'causal discrimination score' to be the fraction of feature vectors in the dataset which are causally unfair with respect to the feature of interest. Is this a reasonable definition?

This is an open-ended question, but you should answer it concretely, with illuminating examples rather than with long-winded text. Think of the point you want to make, then see if you can translate your point into statements about causal diagrams and latent variables.