MATH50013 Probability and Statistics for JMC, Autumn 2021

Coursework

Due on **Thursday, December** 16^{th} **at 17:00 GMT**

Instructions (please read these fully):

- This coursework will be collected via Turnitin at Blackboard > Course Content > Assessments > Coursework.
- You may hand-write solutions but only if your handwriting is really good. A typed report is nicer.
- For questions that ask you to make a figure please embed the figure in the page where you are answering the question (preferred) or else attach your figures as an appendix at the end (if attaching at the end please refer to which page and figure in the main part of your writeup, e.g. "See Fig. 2 on p. 6.")
- On all figures you need axis labels and be sure to have labeled tick marks on the x and y axes to give a sense of scale. If there are multiple things in one plot put labels on them, or add a legend, or somehow describe the different elements in the figure caption.
- To receive marks you must show your work and/or explain your reasoning. Show how you worked out the solution rather than just give the final expression or number. The general advice is: do not make the marker try to guess what you were thinking. For numerical computations, don't paste your code but describe concisely what your algorithm is doing.
- You cannot discuss the problems with your classmates or anyone else.
- If you have questions of clarification do not hesitate to ask me, but please ask privately (either on the discussion forum or by email) rather than as a public post. If needed, I will post clarifications to Blackboard and send an announcement.

Other details:

- This coursework is worth 10% of the total module mark.

1. This question is about estimating integrals by sampling. You are interested in calculating $I = \int_{-\infty}^{\infty} g(x)dx$ for some function $g(x)$. Suppose you have a way of generating independent samples of a random variable X with pdf $f(x)$. The algorithm is:

   Step 1. Generate a sample of size n, $(x_1, x_2, \ldots, x_n)$, from the X distribution.
   Step 2. Give each sample a "weight" $w_i = g(x_i)/f(x_i)$. In other words, each w_i is a sample of the random variable $W = g(X)/f(X)$.
   Step 3. Compute the sample mean of the weights.

   (Note: to keep things well-behaved f must be chosen so that $f(x) > 0$ whenever $g(x) \neq 0$.)

   (a) Show that the sample mean of the weights, $\overline{W} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}$, is an unbiased estimator of the integral I. *[6 points]*

   (b) Show that the variance of this estimator is proportional to $\left(\mathrm{E}\left[\dfrac{g(X)^2}{f(X)^2}\right] - I^2\right)$. How does the standard deviation of the estimator scale with the sample size n? *[6 points]*

   (c) For a non-negative $g(x)$, show that the ideal choice for the RV X is to take $f(x) \propto g(x)$ (ideal in the sense that the variance of the estimator is 0, so the estimate always gives you the true value of I, even using a single sample). Why is it impractical to use this "ideal" choice? *[6 points]*

   Now you will numerically implement the above algorithm to estimate the integral $I = \displaystyle\int_{-\infty}^{\infty} \exp(-x^4)dx$, where the samples x_i are drawn from a normal distribution, i.e. $X_i \sim \mathrm{N}(\mu, \sigma^2)$.

   (d) Start by sampling $n = 20$ values from a standard normal distribution[1]. Using your sample, compute an estimate of the integral I and a 95% confidence interval for I. When finding the confidence interval, you may assume that the weights are normally distributed (though this is not actually true). Do not paste your code into your report but make sure you describe step by step how you calculated the estimate of I and

the confidence interval. *[15 points]*

(e) Repeat part (d) for cases where X is drawn from a normal distribution with a different variance, $\mathrm{N}(0, \sigma^2)$, for $\sigma = 0.5, 1.0, 1.5, 2.0, \ldots, 5$. Don't report numbers but display your results in a plot with σ on the x-axis and I on the y-axis. For each of the 10 values of σ the plot should have a dot at your point estimate of I and also an error bar that shows the 95% confidence interval. Add a horizontal line to the plot at the true value of $I = 1.81280$. *[10 points]*

(f) Explore how the uncertainty of the estimate varies with the sample size n by repeating part (d) for $n = 2^k$, where $k = 3, 4, 5, \ldots, 12$. For each n, rather than compute the confidence interval, repeat the experiment 100 times and find the standard deviation of your 100 estimates of I.

Display your results in a log-log plot with n on the x-axis and the standard deviation of the estimates on the y-axis. There should be a point showing the standard deviation of the I estimator for each of the 10 values of n.

Add a line, passing through the $k = 3$ point, that shows your predicted scaling with n from part (b) (i.e. if you predicted that the standard deviation scales as n^λ plot the line $y = \sigma_{n=8} \left(\frac{n}{8}\right)^\lambda$, where $\sigma_{n=8}$ is the standard deviation of your 100 estimates of I for the $n = 8$ case). *[10 points]*

(g) Discuss why this method becomes unreliable if the pdf $f(x)$ is much "narrower" or "broader" than $g(x)$. It might help to look at the estimator's variance from part (b) — you can write the expectation as an integral and look at plots of the integrand as a function of x for various choices of σ^2. You can also plot the weight as a function of x for various σ. *[7 points]*

[1] To generate the x_i values you have two options. One is to use a pseudorandom number generator for the standard normal distribution, e.g. in python you can do (other languages will be very similar):

```
import numpy as np
np.random.randn() #returns a single sample drawn from a standard normal distribution
np.random.randn(100) #returns an array of 100 independent samples
```

The other option is to download a text file containing standard normal random values from Blackboard > Course Content > Assessments > Coursework > randomnormals.txt. The file contains one million values, one per line. If you want me to post these in another format (e.g. as raw binary bytes, or compressed) please contact me.

2. Let $F(x)$ be some continuous function that increases monotonically from 0 to 1. Define the RV X by the transformation $X = F^{-1}(U)$, where U is uniformly distributed between 0 and 1 (i.e. $U \sim \mathrm{U}(0,1)$), and F^{-1} is the inverse of F. Show that X has $F(x)$ as its cdf, i.e. $F_X(x) = F(x)$, or, equivalently, that X has pdf $f_X(x) = F'(x)$. *[10 points]*

3. Galaxies in the universe have masses that are distributed approximately as a power law above some minimum threshold. That is, if X is the mass of a random galaxy the pdf of X is

$$f_X(x) = \begin{cases} c\left(\frac{x}{x_{\min}}\right)^{-\lambda} & x \geq x_{\min} \\ 0 & x < x_{\min}, \end{cases} \tag{1}$$

where $x_{\min}$ is a fixed constant, and λ is an unknown parameter with $\lambda > 1$.

(a) Show that the constant c is equal to $\dfrac{\lambda - 1}{x_{\min}}$. *[5 points]*

For the rest of the problem set $x_{\min} = 1$. This is equivalent to measuring the mass of a galaxy in units of the minimum mass.

(b) Given the measured masses of n galaxies, derive the maximum likelihood estimate for λ (assume independence of the galaxies). *[8 points]*

In the next two parts you will empirically determine the sampling distribution for the maximum likelihood estimator you found in part (b).

Question 2 points to a method for numerically generating random samples from arbitrary distributions: For the target distribution, first work out the inverse of its cdf, F_X^{-1}. Then generate[2] a sample u from the uniform distribution between 0 and 1 and plug it into this inverse to obtain the sample $x = F_X^{-1}(u)$.

(c) Work out the cdf of the power law distribution and its inverse (also known as the quantile function). *[6 points]*

(d) Create three density histograms showing the sampling distribution of your maximum likelihood estimator $\hat{\lambda}$ from part (b) for the cases $(n, \lambda) = (10, 2)$, $(10, 4)$, and $(50, 4)$.

[2]Like in the Question 1 footnote, you can generate standard uniform samples using a pseudorandom number generator (e.g. in python, `u=np.random.rand()`). Alternatively you can download the file randomuniforms.txt, which contains one million samples from $\mathrm{U}(0,1)$.

That is, generate n samples of X when the true value of the parameter is λ. Then plug this sample into your estimator to get $\hat{\lambda}$. Repeat this 10^4 times to get a sample $(\hat{\lambda}_1, \ldots, \hat{\lambda}_{10^4})$, then create a density histogram from these 10^4 values. (Don't throw away your samples of $\hat{\lambda}$, you will need them again in the next part.)

In your report you should put a plot that shows all three density histograms on the same axes. Only draw the outlines of the histograms — don't shade in the bars, otherwise they will overlap each other and it will be hard to read[3].

The binning of the histograms should be as follows. For the $(n, \lambda) = (10, 2)$ and $(50, 4)$ cases you should use 90 equally-sized bins that cover the interval between 1 and 10 (so each bin has width 0.1). For the $(10, 4)$ case, use 45 bins between 1 and 10 (each bin has width 0.2).

Also, add two vertical lines to the plot showing the true values of λ at $\lambda = 2$ and 4.

[15 points]

(e) Two astronomers are arguing over the true value of λ. One thinks it's 2, the other thinks it's 4. They ask you to design a hypothesis test to test the null hypothesis H_0: $\lambda = 2$ against the alternative H_1: $\lambda = 4$. The test will be based on the observation of a sample of $n = 10$ galaxies and you decide to use the maximum likelihood estimator as the test statistic, i.e. $T = \hat{\lambda}(X_1, X_2, \ldots, X_{10})$.

i. Using your samples from part (d), determine an appropriate rejection region so that the Type I error is 5% and explain your reasoning and procedure. This will inevitably be approximate since you don't have an infinite number of samples, but 10^4 should be enough to make the error negligible.
Hint: it might help to think about the empirical cdf. *[8 points]*

ii. Given your rejection region, use your samples from part (d) to find the power of the test? (Again, it will be slightly approximate since $10^4 < \infty$.) Explain your reasoning. *[6 points]*

iii. The scientists perform their observations and obtain the following masses. Carry out the hypothesis test and state a conclusion about λ. Determine the p value for this observation with the help of your samples from part (d).

```
[1.00, 1.06, 15.69, 1.09, 4.04, 2.20, 2.28, 1.10, 1.46, 1.47]
```

[12 points]

[3] If you are using `matplotlib` in python to make the plots you can do this with the `histtype='step'` argument to `hist`.