UNIVERSITY OF CAMBRIDGE

MPhil in Data Intensive Science

Wednesday 29th December 2023

**THIS IS NOT A REAL EXAM PAPER**

S1: Principles of Data Science - Example Questions

M. Kenzie

*These are set of practise questions put together by one of the MPhil students, Mate Balogh. They are not officially endorsed and the solutions have not been explicitly checked. The weighting of marks is a guide and should be taken with a pinch of salt. The length of the real exam will be shorter than this.*

*Attempt **all** questions.*

*The appropriate number of marks allocated to each question, or each part of a question, is indicated in the right margin inside square,* [], *brackets.*

*Answers to **Section A** questions should be kept short and concise. Relevant formula can be assumed. **Section B** questions will take longer and consequently carry a higher fraction of the marks.*

*The paper contains **8** sides including this one.*

*You may refer to the Mathematical Formulae Handbook supplied, which gives values of constants and contains mathematical formulae, which you may quote without proof. You may also use an approved calculator.*

*Answers should be written in the booklets provided. If several booklets are used join them together using a treasury tag. Write your **candidate number**, not your name, on the cover of **each** booklet. A separate master (yellow) cover sheet should also be completed.*

| STATIONERY REQUIREMENTS | SPECIAL REQUIREMENTS |
|---|---|
| Booklets and treasury tags | Mathematical Formulae Handbook |
| Rough workpad | Approved calculator |
| Yellow master coversheet | |

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator.

SECTION A

*Answers should be concise and relevant formulae may be assumed without proof.*

A1 Show that for any distribution the first moment about the mean (also called the first central moment) is zero. [4]

A2 Show that the expectation of the $\chi^2$ distribution is equal to the number of degrees of freedom. [4]

A3 A bag contains four balls that either white or black. Two balls are drawn from the bag at random without replacement. If both of the balls drawn are black what is the probability that all of the ball in the bag are black? [*Use Bayes' theorem.*] [4]



SECTION B

*Significant marks are given for showing your working.*

B1 The random variables $X_1, \ldots, X_k$ are drawn from Poisson distributions with rates $\lambda_1, \ldots, \lambda_k$.

(a) Derive the distribution of $Z = \sum_{i=1}^{k} X_i$ [5]

(b) Derive the conditional distribution of $X_1$ given that $Z = n$ [5]

B2 $Y_1, Y_2, \ldots$ are i.i.d random variables with zero mean and $\sigma^2$ variance.

(a) $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$ is the sample mean of the first $n$ values. Show that

$$E[\bar{Y}_n^2] = \frac{\sigma^2}{n} \tag{1}$$

[2]

(b) Let $q^4 = E[Y_i^4]$. Show that

$$E[\bar{Y}_n^4] = \frac{1}{n^4}[nq^4 + 3n(n-1)\sigma^4] \tag{2}$$

and thus deduce that there exists some constant $C$ (which you don't have to find) such that

$$E\left[\sum_{n=k}^{\infty} \bar{Y}_n{}^4\right] \leq \frac{C}{k} \tag{3}$$

[*Only evaluate the highest order term and approximate the sum with an integral.*] [4]

(c) Compare what the results 1 and 3 say about the sample mean's deviation from the true mean 0 when the sample size is large. Explain your reasoning in 200 words or less. [10]

B3 $U_1, \ldots, U_k$ are independent random variables drawn from a uniform distribution with $U_i \in [0, 1]$

(a) If $Q_n = U_1^n$, show that the p.d.f. of $Q_n$ is $q_n(x) = \frac{1}{n} x^{\frac{1}{n}-1}$
$\left[\textit{It might be easier to compute the c.d.f. of } Q_n \textit{ first.}\right]$ [5]

(b) Show that $S_1 = -\ln U_1$ is exponentially distributed with parameter 1.

(c) Show that $S_2 = -\ln U_1 \cdot U_2$ has p.d.f $s_n(x) = xe^{-x}$. [5]
$\left[\textit{Be careful with the integration limits!}\right]$

(d) Show that

$$S_n = -\ln \prod_{i=1}^{n} U_i$$

has a p.d.f

$$s_n(x) = \frac{x^{n-1}}{(n-1)!} e^{-x}$$

[5]

(e) What is the distribution of $T_n$, the product of $n$ standard uniform random variables?

$$T_n = \prod_{i=1}^{n} U_i$$

[10]

(f) The characteristic function of the standard Cauchy distribution is given by

$$\phi(t) = e^{-|t|} \tag{4}$$

(a) using the inverse Fourier transformation

$$p(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itX} \phi(t) dt$$

show that the PDF of a standard Cauchy distribution is given by:

$$p(X) = \frac{1}{\pi \left(1 + X^2\right)} \tag{5}$$

[10]

(b) Using (4) show that the mean of the standard Cauchy distribution is undefined. [5]

(TURN OVER

(c) Using (5) directly, what do you get for the mean? [5]

(d) Reminder that if $X_1, \ldots, X_n$ are independent random variables with characteristic functions $\phi_1(t), \ldots, \phi_n(t)$ and $a_1, \ldots, a_n$ are constants, then the characteristic function of the linear combinations of the $X_i$s with weights $a_i$ is given by:

$$\phi_{comb}(t) = \prod_{i=1}^{n} \phi_i(a_i t) \tag{6}$$

Given $k$ independent random samples drawn from the standard Cauchy distribution, what is the distribution of their sample mean? [5]

(e) What does the Central Limit Theorem say about the result of (4)? [5]

B4 We have an unfair coin with some unknown probability $\theta$ of coming up heads. The value $\theta$ is drawn from a uniform distribution on $[0, 1]$.

(a) $X_1, X_2$ are the results of the first two flips (1 if heads, 0 if tails). Find the joint and marginal pmfs of $X_1, X_2$. [6]

(b) Are $X_1$ and $X_2$ independent? What is their correlation? [4]

(c) What is the posterior probability distribution of $\theta$ right after we observe $X_1 = 1$? [5]

(d) Show that if we flip the coin $n$ times and get $k$ heads, then the posterior distribution of $\theta$ is given by:

$$p(\theta|n, k) = (n + 1)\binom{n}{k}\theta^k(1 - \theta)^{n-k} \tag{7}$$

[*Use induction.*] [10]

B5 The pdf of a Pareto distributed random variable $X$ with shape $\alpha > 0$ is given by $p(x) = \frac{C}{x^{\alpha+1}}$ where $x \geq 1$

(a) Compute C. For what values of $\alpha$ are the mean and variance of $X$ finite? [4]

(b) Given $X_1, X_2, \ldots, X_N$ independent random values drawn from a Pareto distribution, what is the maximum likelihood estimator for the shape parameter $\alpha$? [6]

B6 You know that $X$ and $Y$ have a linear relationship of the form $Y = m \cdot X$. You have measured some values $y$ at points $x$.

(a) Show that ordinary least squares estimation gives $\hat{m} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2}$. [3]

(b) How does the expected value of $\hat{m}$ change if you add some noise to the measurements $y$? How about adding some noise to $x$? [6]

B7 A minimum-variance unbiased estimator (MVUE) is an unbiased estimator for a parameter that has lower variance than any other unbiased estimator. Why is this an

interesting object? What happens if you relax the restriction on bias? How does the MVUE relate to the concept of an efficient estimator? Provide your answer in 200 words or less. [15]

B8 The geometric distribution describes the number of independent trials up to and including the first success given that the probability of an individual trial succeeding is $p$.

2Show that the PMF of the geometric distribution is given by

$$g(i) = P[\text{first success on trial i}] = (1-p)^{i-1}p$$

[2]

4Given that the Characteristic function for a random variable $X$ is $\phi_X(t) = E[e^{itX}]$, show that the Characteristic function of the geometric distribution is:

$$\phi_g(t) = \frac{pe^{it}}{1-(1-p)e^{it}} \tag{8}$$

[4]

(a) Show that the mean and variance of the geometric distribution are given by:

$$E[i] = \frac{1}{p}$$

and

$$v(i) = \frac{1-p}{p^2}$$

[5]

(b) Given some samples from a geometric distribution $x_1, x_2, \ldots, x_N$, what is the method of moments estimate for $p$? [2]

(c) Given the samples above, what is the maximum likelihood estimate for $p$? [3]

(d) You repeatedly roll a $d$ sided die and keep track of how many distinct faces of the die you have seen. $T_i$ is the number of additional rolls necessary to see the $i$th face after having seen $i-1$ already. Show that $T_i$ follows a geometric distribution. What is its parameter $p$? [4]

(e) What is the mean and variance of the number of rolls needed to get all possible outcomes on a 4 sided die? [4]

(f) Using the approximation of the central limit theorem and the values in Table 1 (see the end of the paper), how many times do you have to roll the die to be 95% sure you'll see all possible outcomes? [3]

(g) A fast-food chain is running a promotion where every meal comes with a ticket randomly chosen from four varieties. You really want the grand prize for collecting all four types, but after 30 meals you are still missing one. What is your conclusion about the promotional tickets? [3]

(TURN OVER

B9 $X_1, X_2, \ldots, X_N$ are independent random variables drawn from an exponential distribution with probability density $p(x) = Ce^{-\lambda x}, 0 \leq x$

(a) Compute $C$ and the mean and variance of $X_i$ [5]

(b) Show that the maximum likelihood estimator for the rate parameter is

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} X_i} \tag{9}$$

[3]

(c) Given the pdf of the sum

$$\sum_{i=1}^{N} X_i \sim s_N(t) = \frac{\lambda^N t^{N-1}}{(N-1)!} e^{-\lambda t}$$

and the following property of the gamma function for integer $n$

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t} dt = (n-1)!$$

Show that the expected value of the estimate from 9 is:

$$E[\hat{\lambda}] = \frac{N}{N-1}\lambda$$

[5]

(d) Construct an unbiased estimator for $\lambda$ and compute its variance [6]

(e) Compute the minimum-variance bound for an efficient estimator of $\lambda$. [4]

(f) Is your unbiased estimator from part 4 efficient? [2]

B10 We take $N = 2n + 1$ iid samples from a probability distribution with a smooth pdf $f(x)$ and cdf $F(x)$.

(a) Show that the probability that the sample median $\hat{m}$ falls between $x$ and $x + dx$ is

$$P[x \leq \hat{m} \leq x + dx] = \frac{(2n+1)!}{(n!)^2} F(x)^n (1 - F(x))^n f(x) dx \tag{10}$$

[5]

(b) Using Stirling's approximation $n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$, show that

$$\frac{(2n+1)!}{(n!)^2} \approx \frac{(2n+1)4^n}{\sqrt{\pi n}}$$

[5]

(c) using a Taylor series expansion around the distribution median $m = F^{-1}(\frac{1}{2})$ show that

$$4^n(F(m+\epsilon)(1-F(m+\epsilon)))^n = (1 - \frac{4\epsilon^2 n f(m)^2}{n} + o(\epsilon^3))^n \tag{11}$$

[5]

(d) The expression in 11 has a sharp peak around $\epsilon = 0$. To estimate the width of the peak show that for large $n$

$$\frac{1}{e} = (1 - 4c\epsilon^2)^n \tag{12}$$

holds when $\epsilon = \frac{1}{2c\sqrt{n}}$ [4]

(e) Are additional terms in the Taylor expansion in part 3 necessary to determine the asymptotic behaviour as $n \to \infty$? [2]

(f) For large $N$, the following term can be approximated by a (unnormalized) Gaussian.

$$\left(1 - \frac{4\epsilon^2 n f(m)^2}{n}\right)^n$$

What are its mean and variance? [4]

(g) Show that for large $N$, the mean and variance of the sample median are given by:

$$E[\hat{m}] = m$$

and

$$V(\hat{m}) = \frac{1}{4Nf(m)^2}$$

[6]

(h) What is the relative efficiency of using the sample median versus the sample mean as an estimator for the mean of a normal distribution? [4]

END OF PAPER

| Degrees of freedom | $\chi^2$ value | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.004 | 0.02 | 0.06 | 0.15 | 0.46 | 1.07 | 1.64 | 2.71 | 3.84 | 6.63 | 10.83 |
| 2 | 0.10 | 0.21 | 0.45 | 0.71 | 1.39 | 2.41 | 3.22 | 4.61 | 5.99 | 9.21 | 13.82 |
| 3 | 0.35 | 0.58 | 1.01 | 1.42 | 2.37 | 3.66 | 4.64 | 6.25 | 7.81 | 11.34 | 16.27 |
| 4 | 0.71 | 1.06 | 1.65 | 2.20 | 3.36 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 | 18.47 |
| 5 | 1.14 | 1.61 | 2.34 | 3.00 | 4.35 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 | 20.52 |
| 6 | 1.63 | 2.20 | 3.07 | 3.83 | 5.35 | 7.23 | 8.56 | 10.64 | 12.59 | 16.81 | 22.46 |
| 7 | 2.17 | 2.83 | 3.82 | 4.67 | 6.35 | 8.38 | 9.80 | 12.02 | 14.07 | 18.48 | 24.32 |
| 8 | 2.73 | 3.49 | 4.59 | 5.53 | 7.34 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 | 26.12 |
| 9 | 3.32 | 4.17 | 5.38 | 6.39 | 8.34 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 | 27.88 |
| 10 | 3.94 | 4.87 | 6.18 | 7.27 | 9.34 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 | 29.59 |
| p-value (probability) | 0.95 | 0.90 | 0.80 | 0.70 | 0.50 | 0.30 | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |

Table 1: $\chi^2$ distribution values