Statistics for Data Science: Midterm Exam

19 November 2018

Duration: The exam starts at 12:00 and ends at 14:30.

Name :

Sciper :

Question	1	2	3	4 (Bonus)	**Total**
Points					

Instructions

1. You have to hand in the exam as well as *all* your rough work before exiting the exam room.
2. Calculators, mobile phones or any other electronic devices are not allowed. You are entitled to have 2 handwritten A4 sheets. No additional material/format is allowed further to that.
3. Always justify your calculations. If these miss important details, or if your notation is inconsistent, you risk being penalised.
4. Your solutions may be written in English or in French.
5. You can use theorems/propositions/lemmas/corollaries from the course, provided these are clearly stated and their conditions are verified, unless otherwise required by the question.
6. Page 2 contains a list of definitions, theorems or results from the course that *may or may not* be useful for solving some of the exam questions. Please take some time to consult it.
7. Answer all questions 1-3 to get full marks. Question 4 is a bonus and can be answered for extra points but is not mandatory. All 4 questions are worth *equal* amounts of points.

Good luck!

Selected results and definitions from the course

The definitions, theorems and results listed on that page *may or may not* be useful for solving the exam questions. If you need to use any of them, you can simply do so by citing it. For theorems/propositions/lemmas/corollaries make sure that their conditions are verified, unless otherwise required by the question.

Definition 1 (Moment Generating Function)**.** The MGF $M_X : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ of a real-valued random variable X is defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \ \forall t \in \mathbb{R}.$$

Theorem 1 (Ratio of χ^2 and Fisher Distribution)**.** Let $X \sim \chi^2_k$ and $Y \sim \chi^2_n$ be independent random variables. Then,

$$\frac{X/k}{Y/n} \sim F_{k,n}.$$

Theorem 2 (Delta Method)**.** Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n, \theta \in \mathbb{R}$ for all n and a_n diverges to infinity. Let g be continuously differentiable at θ. Then,

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z.$$

Definition 2 (Density of Gaussian Distribution)**.** The density function of the Gaussian distribution $N(0,1)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \ x \in \mathbb{R}.$$

Theorem 3 (Sum of Gaussian Squares and Chi-Square Distribution)**.** Let $X_1, \cdots, X_k \overset{i.i.d.}{\sim} N(0,1)$. Then, $\sum_{i=1}^k X_i^2 \sim \chi^2_k$.

Definition 3 (Mean Squared Error)**.** Let $\hat{\theta}$ be an estimator of a parameter θ. The mean squared error of $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right].$$

Definition 4 (Poisson Distribution)**.** The probability density function f of a Poisson distribution with rate $\lambda > 0$ is given by

$$f : \begin{cases} \mathbb{N} \cup \{0\} \to \mathbb{R}, \\ k \mapsto e^{-\lambda} \frac{\lambda^k}{k!}. \end{cases}$$

Proposition 1 (Densities are Normalised)**.** Let X be a continuous random variable with probability density function $f : \mathbb{R} \to [0, +\infty[$. Then, we have

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Proposition 2 (MGF of Poisson Distribution)**.** The moment generating function of the Poisson distribution with rate λ is given by

$$M(t) = e^{\lambda(e^t - 1)}, \quad t \in \mathbb{R}.$$

Question 1

The goal of this exercise is to get an expression for the *mean* and *variance* of the *chi-squared distribution* χ^2_k with $k \in \mathbb{N}$ degrees of freedom. To this end, we will compute and make use of its *moment generating function* (denoted by MGF hereafter).

(a) Let $X \sim \mathcal{N}(0,1)$. Show that the MGF of X^2 is given by :

$$M_{X^2}(t) = \begin{cases} \frac{1}{\sqrt{1-2t}}, & \text{for } t < 1/2, \\ \infty, & \text{otherwise.} \end{cases}$$

What is the distribution of X^2 ?

Hint : A change of variable could help.

(b) Prove that, for two *independent* random variables Y and Z, we have

$$M_{Y+Z}(t) = M_Y(t) M_Z(t), \quad \forall t \in \mathbb{R}.$$

(c) Use (a) and (b) to compute the MGF of a random variable $U \sim \chi^2_k$, $k \in \mathbb{N}$.

(d) Using the MGF computed in (c), compute the mean and variance of the χ^2_k distribution.

Solution.

(a) From the definition of the MGF, we have :

$$\begin{aligned} M_{X^2}(t) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx^2} e^{-\frac{x^2}{2}} \, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-(1-2t)\frac{x^2}{2}} \, dx, \quad t \in \mathbb{R}. \end{aligned}$$

For $t \geq 1/2$, the integral is undefined and the MGF infinite. The domain of definition of M_{X^2} is hence $]-\infty, 1/2[$. On this domain, we have $1 - 2t > 0$. We can hence perform the following change of variables :

$$z = \sqrt{1-2t}x, \quad dz = \sqrt{1-2t}dx.$$

This yields :

$$\begin{aligned} M_{X^2}(t) &= \frac{1}{\sqrt{1-2t}} \underbrace{\left[\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{z^2}{2}} \, dz \right]}_{=1} \\ &= \frac{1}{\sqrt{1-2t}}, \quad t < \frac{1}{2}, \end{aligned}$$

which is indeed well-defined. We have hence

$$M_{X^2}(t) = \begin{cases} \frac{1}{\sqrt{1-2t}}, & \text{for } t < 1/2, \\ \infty, & \text{otherwise.} \end{cases}$$

Finally, X^2 being the square of a Gaussian distribution, we have $X^2 \sim \chi^2_1$. We have hence characterised the MGF of a $\sim \chi^2_1$ distribution.

(b) We have

$$M_{Y+Z}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right], \quad t \in \mathbb{R}.$$

Since Y and Z are assumed independent then so are e^{tX} and e^{tY} as transformations of independent random variables. Because of this independence, we have

$$\mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right], \quad t \in \mathbb{R},$$

which yields the desired result :

$$M_{Y+Z}(t) = M_Y(t)M_Z(t), \quad t \in \mathbb{R}.$$

(c) If $U \sim \chi^2_k$ with $k \in \mathbb{N}$, we know that $U = \sum_{i=1}^k X_i^2$ where $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Generalising the result proven in (b) to k random variables (trivial) and using (a) yields immediately

$$M_{\sum_{i=1}^k X_i^2}(t) = \begin{cases} \left(\frac{1}{\sqrt{1-2t}}\right)^k, & \text{for } t < 1/2, \\ \infty, & \text{otherwise.} \end{cases}$$

(d) Let $U \sim \chi^2_k$. Differentiating the MGF derived in (c) yields :

$$M'_U(t) = k\left(\frac{1}{\sqrt{1-2t}}\right)^{k+2}, \quad t < \frac{1}{2},$$

and

$$M''_U(t) = k(k+2)\left(\frac{1}{\sqrt{1-2t}}\right)^{k+4}, \quad t < \frac{1}{2}.$$

We have hence

$$\mathbb{E}[U] = M'_U(0) = k,$$

and

$$\begin{aligned} \mathrm{Var}(U) &= \mathbb{E}[U^2] - \mathbb{E}[U]^2 \\ &= M''_U(0) - k^2 \\ &= k^2 + 2k - k^2 \\ &= 2k. \end{aligned}$$

Question 2

Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution Poisson(λ) and rate $\lambda > 0$.

(a) Find the maximum likelihood of λ, denoted by $\widehat{\lambda}$, and its asymptotic distribution.

(b) Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Find a function g such that

$$\sqrt{n}\left(g(\overline{X}) - g(\lambda)\right) \xrightarrow{d} N(0,1).$$

(c) Use the asymptotic distributions in (a) and (b) to derive two confidence intervals. Are they comparable or is one of them better than the other ?

Solution.

(a) MLE is the sample mean and we have by the CLT that $\sqrt{n}(\overline{X}_n - \lambda) \to N(0, \mathrm{var}X_1)$ with $\mathrm{var}X_1 = \lambda$. Alternatively, students can use the asymptotic distribution of the MLE. This yields the same result since $I_1(\lambda) = 1/\lambda$.

(b) The delta method tells us that $\sqrt{n}(g(\overline{X}_n) - g(\lambda)) \to N(0, [g'(\lambda)]^2\lambda)$, hence we want g such that $g'(\lambda) = \frac{1}{\sqrt{\lambda}}$. By integration, $g(\lambda) = 2\sqrt{\lambda}$.

(c) Using LLN and Cramer-Slutzsky on (a) we have $\sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\widehat{\lambda}}} \to N(0,1)$. CI follows easily :

$$(\overline{X}_n \pm \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \sqrt{\overline{X}_n}).$$

Part (b) can be used readily. We gradually update the CI's for

$$\begin{aligned} 2\sqrt{\lambda} \quad &\ldots \quad (2\sqrt{\overline{X}_n} \pm \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}}) \\ \sqrt{\lambda} \quad &\ldots \quad (\sqrt{\overline{X}_n} \pm \frac{1}{2\sqrt{n}} q_{1-\frac{\alpha}{2}}) \\ \lambda \quad &\ldots \quad \left(\max(0, \sqrt{\overline{X}_n} - \frac{1}{2\sqrt{n}} q_{1-\frac{\alpha}{2}})^2, (\sqrt{\overline{X}_n} + \frac{1}{2\sqrt{n}} q_{1-\frac{\alpha}{2}})^2\right) \end{aligned}$$

The CI's are not really comparable, the second one is shifted to the right compared to the first one (easy to see), but we can't really say which one is better. They both contain one element slowing down the asymptotics of CLT.

Question 3

Let $X_1, \ldots, X_n$ be i.i.d. random variables with Gaussian distribution $N(0, \sigma^2)$. We are interested in testing the hypothesis pair

$$H_0 : \sigma^2 = 1 \quad \text{vs.} \quad H_1 : \sigma^2 \neq 1.$$

(a) Find the maximum likelihood estimate of σ^2, denoted $\hat{\sigma}^2$. What is the *exact* distribution of $n\hat{\sigma}^2$ under H_0?
In the following, denote by F the cumulative distribution function of this distribution and treat it as a known function.

(b) Use the distribution derived in (a) to construct a statistical test. Express the p-value of the test in terms of F.

(c) Assuming that the true value of the variance is $\sigma^2 = 2$, express the power of the test in terms of F.

Solution.

(a) We find by the usual calculation that $\widehat{\sigma}^2 = \frac{1}{n}\sum X_i^2$ (full points awarded for this, the exact distribution is rather tied with (c)). The distribution under H_0 is $n\widehat{\sigma}^2 \sim \chi^2_n$.

(b) We reject on level α if $n\widehat{\sigma}^2 \geq q_{1-\alpha}$, where $q = F^{-1}$. p-value is the smallest level on which we reject, i.e.

$$\text{p-value} = \inf\left\{\alpha, n\widehat{\sigma}^2 = q_{1-\alpha}\right\} = \inf\{\alpha, F(n\widehat{\sigma}^2) = 1-\alpha\} = 1 - F(n\widehat{\sigma}^2)\,.$$

(c)

$$\begin{aligned}\text{power} &= P_{H_1}(n\widehat{\sigma}^2 \geq q_{1-\alpha}) = P_{H_1}\left(\sum_{i=1}^n X_i^2 \geq q_{1-\alpha}\right) \\ &= P_{H_1}\left(\frac{\sum X_i^2}{2} \geq \frac{q_{1-\alpha}}{2}\right) = P\left(\chi^2_n \geq \frac{n}{2}\right) = 1 - F\left(\frac{q_{1-\alpha}}{2}\right).\end{aligned}$$

Question 4 (Bonus) :

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ with $\theta \in \Theta =]-\infty, b]$, for a known $b < +\infty$. Given this restriction on the parameter space, can you construct an estimator of θ that performs better than the empirical mean $\overline{X}$ in terms of mean squared error ?
Be sure to prove your claims !
Solution. Let $\hat{\theta} = \bar{X}\mathbf{1}\{\bar{X} \leq b\} + b\mathbf{1}\{\bar{X} > b\}$ and note that $\bar{X} = \hat{\theta} + (\bar{X} - b)^+$. It follows that

$$\mathrm{MSE}(\bar{X}, \theta) = \mathbb{E}(\bar{X} - \theta)^2 = \underbrace{\mathbb{E}(\hat{\theta} - \theta)^2}_{=\mathrm{MSE}(\hat{\theta}, \theta)} + \underbrace{\mathbb{E}((\bar{X} - b)^+)^2}_{>0} + 2\underbrace{\mathbb{E}[(\hat{\theta} - \theta)(\bar{X} - b)^+]}_{\geq 0}$$

The second term is strictly positive because $\mathbb{P}_\theta[\bar{X} > b] > 0$ for any θ. To see that the last term on the right is non-negative, we note that when $(\bar{X} - b)^+ > 0$ then $(\hat{\theta} - \theta) = (b - \theta) \geq 0$ yielding a non-negative product. Otherwise, the product is zero. Hence the random variable inside the expectation is non-negative.

———————— END OF THE EXAM PAPER ————————