EPFL

Course: MATH-413
Setter: Olhede

EXAMINATIONS

January 2021

MATH-413

Statistics for data science

Ecole Polytechnique Federale de Lausanne

January 2021

MATH-413

Statistics for data science

Date: Saturday, 23rd January, 2020 Time: 8.15 am–12.15 am

The exam will be distributed on moodle, and solutions uploaded on moodle.

All 6 questions should be answered.

All 6 questions are of equal value.

The exam is formally 4 h long rather than the normal 3 h. You should start uploading after 3 h and 40 mins to ensure that you have the time to upload, and have 4 hours to fully submit your answers to the questions.
Please write out the solutions by hand and scan them using standard EPFL scanning instructions: How to upload handwritten notes.
Convert your solutions to PDF (one PDF per solution so your solutions can be distributed for marking.)
Give each PDF a name including the name of the course and your scipher: e.g. for qn 1: MATH413q1-your-scipher.
You may not collaborate with other students and work together with other students. You may not in any way plagiarise your answer, if you consult online material then it must be referenced appropriately.
You MAY use your lecture notes and any material handed out in the course.
Once you have submitted your answers then you cannot resubmit or change your answers (like leaving a physical exam).
Countdown to submit starts 8.15 and ends 12.15 on January 23rd 2021 UNLESS you have other special circumstances.
For questions related to a technical problem, please call the regular IT support desk at 1234. I will also be available at +41 21 693 78 36.
Please note EPFL regulations for cheating and fraud.

1. Assume we observe random variables X and Y. Assume that X takes values of $\{0,1\}$ and Y takes values uniformly on $(0,1)$. Assume that conditionally on $Y=y$ then the variable X has the distribution

$$\Pr\{X=1\,|\,Y=y\}=y. \tag{1}$$

(a) Please tabulate (write down in a table for the two possible outcomes of X) the joint distribution of X and Y, namely $\Pr\{X=1\cap Y\leq y\}$ and $\Pr\{X=0\cap Y\leq y\}$. Use Bayes' theorem to do so.

(b) Please calculate the conditional expectation of X as well as the marginal expectation of X.

(c) Please calculate the conditional variance of X given $Y=y$ as well as the marginal variance of X. How informative about the distribution of X is knowing the value of Y? Discuss using the results both of the expectation and the variance.

(d) Assume that X could instead be specified given $Z=z$ (where the marginal distribution of X would be different) where $f_Z(z)=5z^4$ if $z\in(0,1)$ and $\Pr\{X=1\,|\,Z=Z\}=z$. Give an intuitive reason for whether you expect the expectation of X to change, and if it would increase or decrease? Note that $\mathrm{E}\{Z\}=5/6$ while $\mathrm{E}\{Y\}=1/2$.

Hint: for a mixed discrete and continuous random vector you may find it useful to write

$$\begin{aligned}\Pr\{X=x\cap y\leq Y\leq y+\delta\}&=\Pr\{X=x\,|\,y\leq Y\leq y+\delta\}\times\Pr\{y\leq Y\leq y+\delta\}\\&=yf_Y(y)\delta+O(\delta^2),\end{aligned} \tag{2}$$

for $0<\delta<<1$ and noting $f_Y(y)$ is continuous.

2. Assume that X and Y are two continuous random variables in $(1,2)^2$. Assume that the joint distribution of X and Y takes the form for some positive constant C of

$$f_{X,Y}(x,y) = C \cdot \begin{cases} 1-(x-1)(y-1) & \text{if} \quad (x,y) \in (1,2)^2 \\ 0 & \text{o/w} \end{cases} \tag{3}$$

(a) Determine the value of C. Show your workings, do not use numerical integration routines, such as maple or mathematica. You may check your answers with a numerical routine.

(b) Determine the marginal distribution of X. Again show your workings, do not use maple or mathematica. Calculate the marginal expectation of X.

(c) For continuous random variable X that exceeds $\alpha > 0$, prove that

$$\mathrm{E}\{X\} = \alpha + \int_{\alpha}^{\infty} \{1 - F_X(x)\} dx. \tag{4}$$

Use it to calculate $\mathrm{E}\{X\}$ when $\alpha = 1$ and compare this to your answer derived from first principles in Qn 2(b).

(d) Assume $Z = -X$ that subceeds (i.e. is less than) $-\alpha < 0$. Show that

$$\mathrm{E}\{Z\} = -\alpha - \int_{-\infty}^{-\alpha} F_Z(z) dz. \tag{5}$$

3. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ denote a random sample of size n from

$$f_X(x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1.$$

(a) Write down the likelihood of $\boldsymbol{X}$.

(b) Find the Maximum Likelihood Estimator of θ, denoting this $\widehat{\theta}$.

(c) Denote the probability that an observation from distribution $f_X(x)$ is equal to 0, θ_0 say, i.e. set $P(X_i = 0) = \theta_0$. Write this probability down as a function of θ.

(d) Find the Maximum Likelihood Estimator of θ_0, writing down any properties about maximum likelihood estimators you use, and checking their applicability to this problem carefully.

4. The mutual information between two continuous random variables X and Y takes the form

$$I(X,Y) = \int\int f_{X,Y}(x,y)\log\left\{\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right\}dxdy.$$

Assume we are considering two possible distributions for X and Y where $(x,y) \in (2,4)^2$: In the first case we have $f_{XY,1}(x,y) = f_{X,1}(x)f_{Y,1}(y)$ where

$$f_{X,1}(y) = f_{Y,1}(y) = \frac{1}{2}, \quad y \in (2,4),$$

and the second possible distribution is

$$f_{XY,2}(x,y) = \frac{1}{36}xy\mathrm{I}((x,y) \in (2,4)^2),$$

where as usual $\mathrm{I}(\cdot)$ denotes the indicator function taking the value unity if the argument is true, zero otherwise.

(a) Determine $f_{X,2}(x)$ from first principles.

(b) Calculate the mutual information $I(X,Y)$ of both the indicated distributions. Interpret this value.

(c) A third distribution we might consider is

$$f_{XY,3}(x,y) = \frac{1}{18}xy\mathrm{I}(2 \leq x < y \leq 4).$$

Determine the mutual information of this distribution and contrast this with your answers to the earlier questions.

(d) In which of these 3 cases is X and Y independent?

5. The variable Y_i is generated from the linear model of

$$Y_i = \beta_0 + \beta_1\sqrt{x_i} + \epsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where ϵ_i is zero-mean and independent over different values i with variance σ^2.

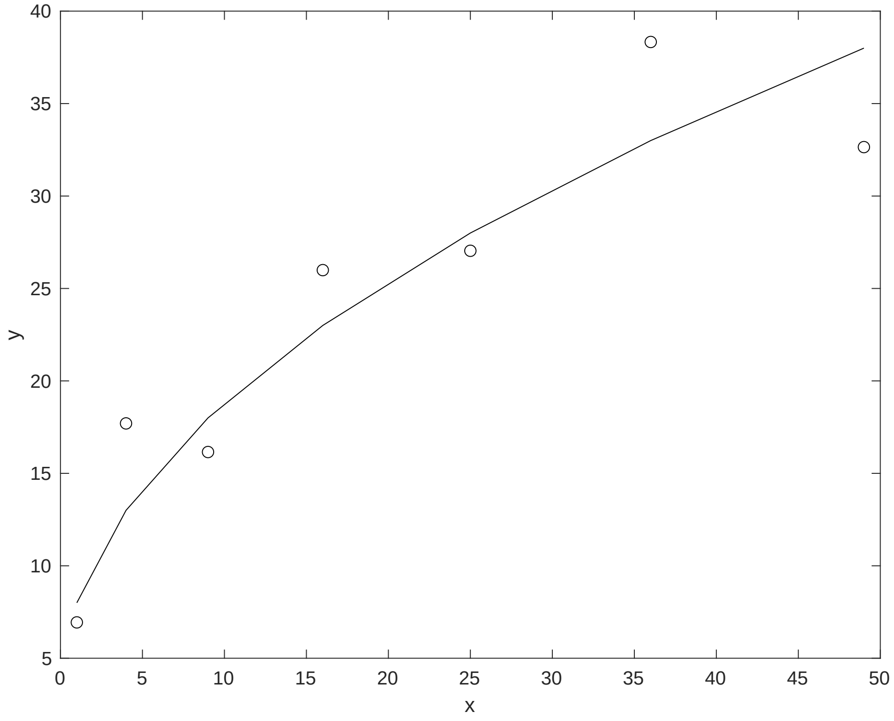
(a) Write down the matrices X, Y, $\beta = (\beta_0, \beta_1)^T$ and ϵ so that the entries are specified the relationship of (6) and

$$Y = X\beta + \epsilon.$$

(b) Determine the unbiased minimum variance estimator of β in terms of X, Y and note any results you need to obtain it; as well as any assumptions on X.

(c) Assume that $x = \begin{pmatrix}1 & 4 & 9 & 16 & 25 & 36 & 49\end{pmatrix}^T$ and $y = \begin{pmatrix}7.8692 & 12.9566 & 18.0343 & 23.3578 & 28.2769 & 32.8650 & 38.3035\end{pmatrix}^T$. Determine the unbiased minimum variance estimator of β for this data set. You may use a calculator or software to calculate the estimates.

(d) Another data set is given by $x = \begin{pmatrix}1 & 4 & 9 & 16 & 25 & 36 & 49\end{pmatrix}^T$ and $y = \begin{pmatrix}6.9384 & 17.7009 & 16.1532 & 25.9923 & 27.0379 & 38.3317 & 32.6461\end{pmatrix}^T$ plotted below. Discuss whether you think second order assumptions hold for this problem.



6. We observe the status of fruit flies in an experiment to determine the effects of a pesticide. n fruit flies are observed, and $Y = y$ corresponds to the number that are observed to be alive at the time point T. Y is the sum of Z_i the outcome for fly i, so that $Y = \sum_i Z_i$.

(a) What distribution is reasonable to assume for Y if each fly is given the same dosage of pesticide? What assumption does that require making on the Z_i?

(b) Each fly i is given a different dosage of pesticide, namely $x_i > 0$. Using the logistic link function write down the linear GLM model for the probability that a fly survives in terms of the constant β_0 and linear coefficient β_1. Interpret the sign of the linear term in the model.

(c) Describe how to fit the parameters of the linear model given data $\{Z_i\}$.

(d) Describe how to test if the linear term is required using the deviance of the GLM. What would be your null and alternative hypothesis, and what is the scientific interpretation of these hypotheses?