EPFL
Probability and Statistics for SIC
2015–2016, Spring semester

# Probability and Statistics: Exam

20 June 2016

**Duration:** The exam starts at 12:15 and ends at 15:15.

First name:

Last name:

SCIPER number:

| Exercise | Points | Indicative marks |
|---|---|---|
| 1 | | /17 points |
| 2 | | /9 points |
| 3 | | /6 points |
| 4 | | /12 points |
| Total: | | /44 points |

**PROTOCOL**:

- Personal documents may not be brought into the exam.
- A simple calculator may be used but use of other electronic devices, including mobile phones, is forbidden. You may not share a calculator.
- Do NOT unstaple pages.
- French versions of the questions are provided for your convenience. The English version is the reference in case of discrepancies between the two versions.
- Answers may be given in English or in French.
- Explain your reasoning! An unjustified answer will be treated as incorrect.
- Write your answers in the exam scripts. If you need more space or scratch paper, use the blank pages at the end of the script, or ask for another blank page and staple it into the script.
- The assistants will reply to questions only if there is a typo. If you find a question unclear, explain how you understand it when giving your solution.

**Exercise 1. (a)** A tetrahedral die has four sides numbered 1, 2, 3, 4. Let $X$ denote the sum of the bottom faces when two fair tetrahedral dice are thrown independently. Write down a probability space that represents this random experiment, and use it to obtain $\Pr(X = 3 \mid X \leq 3)$.

**(b)** In the novel *War and Peace*, Bezukhov and Dolokhov fight a duel with pistols. Bezukhov is a poor shot, and hits his target with probability 0.2, while Dolokhov hits his target with probability 0.8. If Bezukhov shoots first, and then they take turns at firing until one of them is hit, what is the probability that Bezukhov wins the duel?

**(c)** Let $X$ denote the number of heads occurring when I toss a fair coin, and suppose that, conditional on $X = x$, $Y \sim \text{Pois}(x + 1)$. Find $\text{E}(Y)$, and compute $\Pr(X = 1 \mid Y = y)$.

**(d)** If $X_1, X_2$ are independent Bernoulli variables with probability $p$, find the distribution of $Z = \max(X_1, X_2) - \min(X_1, X_2)$.

**(e)** Recall that the moment-generating function of a random variable $X$ is $M_X(t) = \text{E}(e^{tX})$, defined for all $t \in \mathbb{R}$ such that $M_X(t)$ is finite. If $a, b$ are constants, find the moment-generating function of $Z = a + bX$ in terms of $M_X(t)$.

**(f)** Is it possible for data $x_1, \ldots, x_n$ to have average $\bar{x} = 0.3$, sample variance $s^2 = 0$ and inter-quartile range 0.1? If so, give an example. If not, why not?

**(g)** If $X \sim \exp(\lambda)$, find the $p$ quantile of $1/X$.

**(h)** Suppose that $Y_1, \ldots, Y_n$ are independent Bernoulli variables with success probability $\theta$. If $\theta$ has prior density $\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a, b)$ $(0 < \theta < 1)$, where $a, b > 0$, find the posterior density of $\theta$ given that $Y_1 = y_1, \ldots, Y_n = y_n$.

**(i)** If the joint density of $X, Y$ is $f(x, y) = c(1 + 4xy)$ $(0 < x, y < 1)$ and $f(x, y) = 0$ elsewhere, find $c$ and the marginal density of $X$.

**Français**:

**(a)** Un dé en forme de thétraèdre a ses quatre faces numérotées 1, 2, 3, 4. Soit $X$ la somme des faces inférieures de deux dés à quatre faces lancés de manière indépendante. Donnez un espace de probabilité qui correspond à cette expérience aléatoire et utilisez-le pour obtenir $\Pr(X = 3 \mid X \leq 3)$.

**(b)** Dans le roman *Guerre et paix*, Bezukhov and Dolokhov s'affrontent en duel avec des pistolets. Bezukhov est mauvais tireur et atteint sa cible avec une probabilité de 0.2, alors que Dolokhov atteint sa cible avec une probabilité de 0.8. Si Bezukhov tire en premier, et qu'ensuite ils tirent à tour de rôle jusqu'à ce que l'un d'eux soir touché, quelle est la probabilité que Bezukhov gagne le duel ?

**(c)** Soit $X$ le nombre de fois où face est obtenu lors du lancé d'une pièce de monnaie équilibrée, et supposons que, étant donné $X = x$, $Y \sim \text{Pois}(x + 1)$. Trouvez $\text{E}(Y)$, et calculez $\Pr(X = 1 \mid Y = y)$.

**(d)** Si $X_1, X_2$ sont des variables de Bernoulli indépendantes de probabilité $p$, trouvez la distribution de $Z = \max(X_1, X_2) - \min(X_1, X_2)$.

**(e)** Pour rappel, la fonction génératrice des moments d'une variable aléatoire $X$ est $M_X(t) = \text{E}(e^{tX})$, pour tous les $t \in \mathbb{R}$ tels que $M_X(t)$ est fini. Si $a, b$ sont des constantes, trouvez la fonction génératrice des moments de $Z = a + bX$ en fonction de $M_X(t)$.

**(f)** Est-il possible pour des données $x_1, \ldots, x_n$ d'avoir une moyenne $\bar{x} = 0.3$, une variance $s^2 = 0$ et une étendue interquartiles de 0.1 ? Si oui, donnez un example. Si non, pourquoi pas ?

**(g)** Si $X \sim \exp(\lambda)$, donnez le $p$-quantile de $1/X$.

**(h)** Soient $Y_1, \ldots, Y_n$ des variables aléatoire de Bernoulli indépendantes avec probabilité de succès $\theta$. Si $\theta$ a la densité à priori $\pi(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a, b)$ $(0 < \theta < 1)$, où $a, b > 0$, donnez la densité a posteriori de $\theta$ étant donné $Y_1 = y_1, \ldots, Y_n = y_n$.

**(i)** Si la fonction de densité conjointe de $X, Y$ est $f(x, y) = c(1 + 4xy)$ si $0 < x, y < 1$ et $f(x, y) = 0$ sinon, donnez $c$ et la densité marginale de $X$.

**Exercise 2.** Two successive software downloads take times (minutes) $X_1 \sim \mathcal{N}(8, 3^2)$ and $X_2 \sim \mathcal{N}(16, 4^2)$. The download times are independent.

**(a)** What is the distribution of the total download time $T = X_1 + X_2$?

**(b)** What is the probability that the total download time exceeds 30 minutes?

**(c)** Given that $X_1 = 10$, find the probability that the total download time exceeds 30 minutes.

**(d)** Given that the total download time was 30 minutes, what is the probability that $X_1$ was less than 7 minutes?

Hint: recall that if

$$\begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix} \right\},$$

then $W_1 \mid W_2 = w \sim \mathcal{N}\{\mu_1 + \omega_{12}\omega_{22}^{-1}(w - \mu_2), \omega_{11} - \omega_{12}^2/\omega_{22}\}$.

**Français**: Deux téléchargements successifs par une application durent (en minutes) $X_1 \sim \mathcal{N}(8, 3^2)$ and $X_2 \sim \mathcal{N}(16, 4^2)$. Les téléchargements sont independents.

**(a)** Quelle est la distribution de la durée totale de téléchargement $T = X_1 + X_2$ ?

**(b)** Quelle est la probabilité que la durée totale de téléchargement excède 30 minutes ?

**(c)** Etant donné $X_1 = 10$, donnez la probabilité que la durée totale de téléchargement excède 30 minutes.

**(d)** Etant donné que la durée totale de téléchargement était de 30 minutes, quelle est la probabilité que $X_1$ était inférieur à 7 minutes ?

**Exercise 3.** The number of failures of an operating system observed over a period of time $t$ has the Poisson distribution with mean $\lambda t$.

**(a)** Find the maximum likelihood estimate of $\lambda$ and the observed information, based on data $(t_1, x_1), \ldots, (t_n, x_n)$ for independent periods of observation.

**(b)** Use your results from (a) to give a 90% confidence interval for $\lambda$ based on the data below:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t$ (days) | 2.5 | 2.0 | 7.0 | 0.5 | 3.0 | 4.0 | 3.0 | 3.0 | 4.0 | 7.0 |
| $x$ | 0 | 1 | 3 | 0 | 5 | 0 | 0 | 2 | 3 | 10 |

**(c)** Suppose that in addition to the data in (a), data $y_1, \ldots, y_m$ are available representing the times to first failure for the operating system. Assuming that these times are independent, and independent of the data in (a), find the maximum likelihood estimate of $\lambda$ when the sets of data are combined.

**Français**: Le nombre d'erreurs d'un système d'exploitation observées au cours d'une période de temps $t$ suit une loi de Poisson de moyenne $\lambda t$.

**(a)** Donnez l'estimateur du maximum de vraisemblance de $\lambda$ ainsi que l'information observée sur la base des données $(t_1, x_1), \ldots, (t_n, x_n)$ pour des périodes d'observation indépendantes.

**(b)** Utilisez vos résultats de (a) pour donner un intervalle de confiance à 90% pour $\lambda$ sur la base des données suivantes :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t$ (jours) | 2.5 | 2.0 | 7.0 | 0.5 | 3.0 | 4.0 | 3.0 | 3.0 | 4.0 | 7.0 |
| $x$ | 0 | 1 | 3 | 0 | 5 | 0 | 0 | 2 | 3 | 10 |

**(c)** Supposons qu'en complément des données de (a), on dispose également des données $y_1, \ldots, y_m$ qui correspondent aux durées jusqu'aux premiers échecs du système d'exploitation. En supposant que ces durées sont indépendantes, et indépendantes des données de (a), donnez l'estimateur du maximum de vraisemblance de $\lambda$ quand les jeux de données sont combinés.

**Exercise 4.** According to US federal law, action should be taken if levels of lead in 10% or more samples of drinking water in a town exceed 15 parts per billion (ppb). The data shown in Figure 1 below are samples of tap water in Flint, Michigan, where residents complained about discoloured drinking water. Three hundred sampling kits were sent out to households chosen at random, and 271 households each returned three samples of tap water: sample 1, taken from the first water to come out of the tap; sample 2, taken after letting the water run for 15 seconds; and sample 3, taken after letting the water run for two minutes.

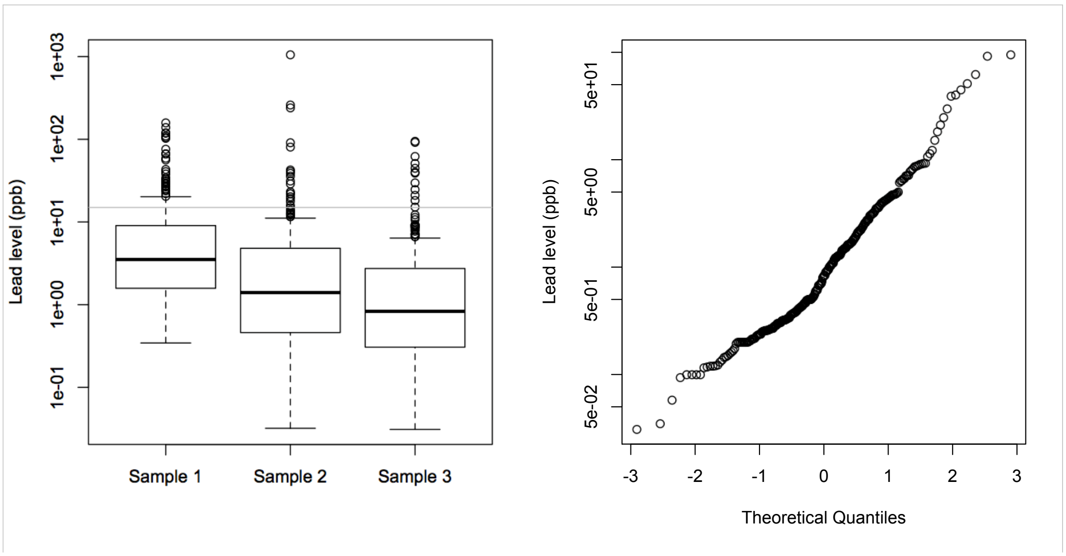


Figure 1: Flint lead data. Left: boxplots of the sampled lead levels (ppb), with horizontal line showing 15 ppb. Right: lead levels (ppb) for sample 1. Note the log scales in both panels.

**(a)** Consider the left panel of Figure 1.

(i) Explain how a boxplot is constructed.

(ii) What do you learn from the panel?

**(b)** The level 15 ppb is exceeded by 45 of the elements of sample 1, and

$$\sum_{r=45}^{271} \binom{271}{r} 0.1^r 0.9^{271-r} \approx 0.0003.$$

What does this calculation represent? Do you think that action should be taken?

**(c)** What do you learn from the right panel of Figure 1?

**(d)** Let $x_1, \ldots, x_{271}$ represent the log lead levels in sample 1. Given that their average is 1.40 and their sample variance is 1.68, give a 95% confidence interval for the mean log level of lead.

**(e)** On what assumptions is the interval in (d) based? Do you find them reasonable?

**Français**: D'après une loi fédérale américaine, des démarches doivent être entreprises quand les taux de plomb de 10% ou plus des échantillons d'eau potable d'une la ville excèdent 15 parties par milliard (ppb). Les données présentées dans la Figure 1 ci-dessus sont des échantillons d'eau du robinet de Flint, Michigan, dont les habitants se sont plaints à cause de la décoloration de l'eau potable. Trois cents kits d'échantionnage ont été envoyés à des foyers choisis aléatoirement et 271 d'entre-eux ont renvoyé trois échantillons d'eau du robinet : échantillon 1, prélevé de l'eau provenant directement du robinet ; échantillon 2, prélevé après avoir laissé coulé l'eau pendant 15 secondes ; et l'échantillon 3, prélevé après avoir laissé coulé l'eau pendant deux minutes.

**(a)** Considérez le graphique de gauche dans la Figure 1.

(i) Expliquez comment une boîte à moustaches est construite.

(ii) Que déduisez-vous de ce graphique ?

**(b)** Le niveau de 15 ppb est dépassé par 45 élements de l'échantillon 1, et

$$\sum_{r=45}^{271} \binom{271}{r} 0.1^r 0.9^{271-r} \approx 0.0003.$$

Que représente ce calcul ? Pensez-vous que des démarches doivent être entreprises ?

**(c)** Que déduisez-vous du graphique de droite dans la Figure 1 ?

**(d)** Soient $x_1, \ldots, x_{271}$ les logs des niveaux de plomb de échantillon 1. Etant donné que leur moyenne est 1.40 et que leur variance est 1.68, donnez un intervalle de confiance à 95% pour la moyenne des logs des niveaux de plomb.

**(e)** Sur quelles hypothèses se basent l'intervalle obtenu en (d) ? Les trouvez-vous raisonnables ?