EPFL

Probability and Statistics for SIC

2016–2017, Spring semester

Probability and Statistics: Exam

23 June 2017

Duration: The exam starts at 8:15 and ends at 11:15.

Exercise 1. **(a)** A fair six-sided die is thrown six times in succession, and the numbers shown on the top faces are recorded. Write down a probability space that represents this random experiment. Is it more likely that all the top faces show the same number, or that these numbers are all different?

(b) If $X \sim \exp(\lambda)$, find the probability density function of $Y = X^2$.

(c) If X and Y are two independent geometric random variables with parameter p, find $\Pr(Y > X)$.

(d) If $Z_1, Z_2 \overset{\text{iid}}{\sim} N(0,1)$ and a, b, c, d, e, f are constants, find the correlation of $X = a+bZ_1+cZ_2$ and $Y = d + eZ_1 + fZ_2$.

(e) If $X_1 \sim N(1,4)$, $X_2 \sim N(-2,9)$ and $\mathrm{corr}(X_1, X_2) = 1$, give the distribution of X_1+X_2+1.

(f) If the joint density of continuous random variables U, V is $f(u,v) = cu^2v$ $(0 < u < 2, 0 < v < 1)$ for some $c > 0$, and $f(u,v) = 0$ elsewhere, find the conditional density of V given that $U = 1$.

(g) If X has a continuous distribution with p quantile x_p, give $\Pr(X \geq x_p)$.

(h) In what sense is the sample median is more robust than the sample average? How can this be quantified?

(i) If, for large n, $\overline{X}$ has an approximate normal distribution with mean $\mu \neq 0$ and variance σ^2/n, give an approximate distribution for $1/\overline{X}$.

(j) If $I_1, \ldots, I_{2n}$ are independent Bernoulli variables with probability p, show that

$$T = n^{-1} \sum_{j=1}^{n} I_{2j-1}(1 - I_{2j})$$

is an unbiased estimator of $p(1-p)$, and give its mean square error.

Français:

(a) On jette six fois de suite un dé équilibré à six faces, et on note les nombres apparaissant sur la face du haut. Donnez l'espace de probabilité de cette expérience aléatoire. Est-il plus probable que les faces du haut des jets montrent le même nombre, ou que toutes ces faces aient un nombre différent?

(b) Si $X \sim \exp(\lambda)$, calculez la fonction de densité de $Y = X^2$.

(c) Si X et Y sont deux variables aléatoires indépendantes géométriques de paramètre p, calculez $\Pr(Y > X)$.

(d) Si $Z_1, Z_2 \overset{\text{iid}}{\sim} N(0,1)$ et a, b, c, d, e, f sont des constantes, calculez la corrélation de $X = a + bZ_1 + cZ_2$ et $Y = d + eZ_1 + fZ_2$.

(e) Si $X_1 \sim N(1,4)$, $X_2 \sim N(-2,9)$ et $\text{corr}(X_1, X_2) = 1$, déterminez la distribution de $X_1 + X_2 + 1$.

(f) Si la densité conjointe de variables aléatoires continues U, V est $f(u,v) = cu^2v$ $(0 < u < 2, 0 < v < 1)$ pour un certain $c > 0$, et $f(u,v) = 0$ sinon, calculez la densité conditionnelle de V sachant que $U = 1$.

(g) Si X a une distribution continue avec p quantile x_p, trouvez $\Pr(X \geq x_p)$.

(h) Que veut dire l'expression 'la médiane empirique est plus robuste que la moyenne empirique'? Comment quantifier ces notions?

(i) Si, pour n grand, $\overline{X}$ a approximativement une distribution normale de moyenne $\mu \neq 0$ et variance σ^2/n, déterminez la distribution approximative de $1/\overline{X}$.

(j) Si $I_1, \ldots, I_{2n}$ sont des variables aléatoires indépendantes de Bernoulli avec probabilité p, montrez que

$$T = n^{-1} \sum_{j=1}^{n} I_{2j-1}(1 - I_{2j})$$

est un estimateur non biasé de $p(1-p)$, et déterminez son erreur moyenne quadratique.

Exercise 2. A credit card transaction is fraudulent with probability $p = 0.01$. When the transaction is legitimate, the amount of money paid, X (thousands of dollars), can be taken to have a standard uniform density on the interval $(0, 1)$, while if the transaction is fraudulent, X has probability density function

$$f(x) = \begin{cases} \frac{9}{(1+9x)^2}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(a) Any transaction attempted with $X > 1$ is automatically classified as fraudulent and refused. What is the probability of this event?

(b) Find the probability that a transaction is fraudulent, given that it is for x thousand dollars.

(c) My bank must reimburse me in case of a fraudulent transaction. What is its expected loss, for transactions that have not been refused?

Français:
Une transaction de carte de crédit est frauduleuse avec probabilité $p = 0.01$. Lorsque la transaction est légitime, son montant, X (milliers de dollars), peut être modélisé par une densité standard uniforme sur l'intervalle $(0, 1)$, tandis que si la transaction est frauduleuse, X a la densité

$$f(x) = \begin{cases} \frac{9}{(1+9x)^2}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(a) Toute transaction tentée avec $X > 1$ est automatiquement classée comme frauduleuse et refusée. Quelle est la probabilité de cet évènement?

(b) Calculez la probabilité qu'une transaction soit frauduleuse, sachant qu'elle est de x milliers de dollars.

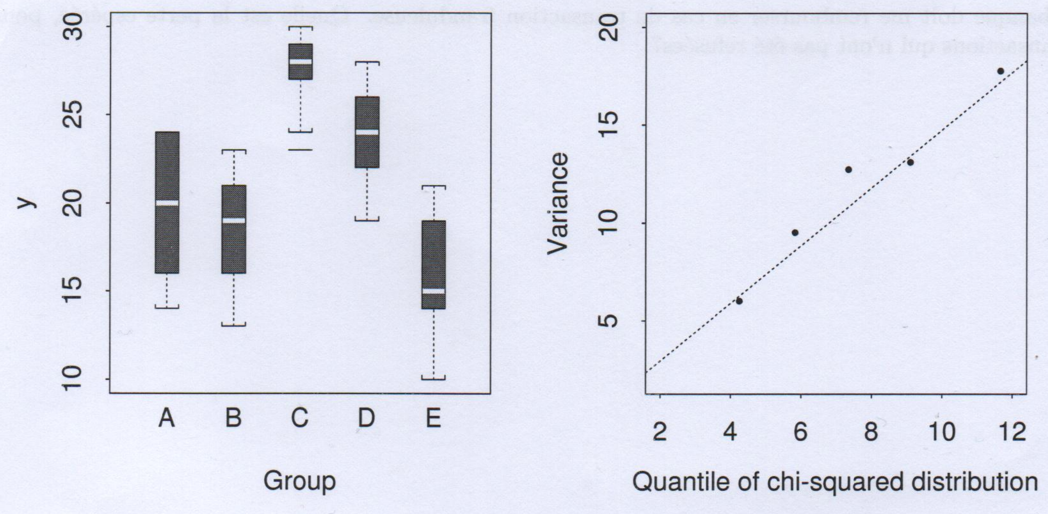
(c) Ma banque doit me rembourser en cas de transaction frauduleuse. Quelle est la perte espérée, pour les transactions qui n'ont pas été refusées?

Exercise 3. In an investigation on the teaching of arithmetic, 45 pupils were divided at random into five groups of nine. Groups A and B were taught in separate classes by the usual method. Groups C, D, and E were taught together for a number of days. On each day C were praised publicly for their work, D were publicly reproved and E were ignored. At the end of the period all pupils took a standard test, with the results given in Table 1 and displayed in the left panel of Figure 1.

Table 1: Data on the teaching of arithmetic.

Group	Test result y									Average	Variance
A (Usual)	17	14	24	20	24	23	16	15	24	19.67	17.75
B (Usual)	21	23	13	19	13	19	20	21	16	18.33	12.75
C (Praised)	28	30	29	24	27	30	28	28	23	27.44	6.03
D (Reproved)	19	28	26	26	19	24	24	23	22	23.44	9.53
E (Ignored)	21	14	13	19	15	15	10	18	20	16.11	13.11

Figure 1: Plots for data on the teaching of arithmetic. The left panel shows the original data, and the right panel plots the ordered variances for the groups against plotting positions for the χ_8^2 distribution.



(a) For the left panel of Figure 1:

(i) explain how the boxplots are constructed.

(ii) What do you learn from the panel?

(b) For the right panel of Figure 1:

(i) explain how a QQ-plot is constructed.

(ii) What do you learn from the panel?

(c) Give a 95% confidence interval for the mean test result for Group C, and discuss its interpretation.

(d) If we assume that the individual test results for Groups A and B have variance 15, is there evidence that the mean test results for these groups differ?

Français:

Lors d'une enquête sur l'enseignement de l'arithmétique, 45 élèves sont divisés au hasard en cinq groupes de taille neuf. Les groupes A et B ont suivi la méthode habituelle dans des classes séparées. Les groupes C, D, et E ont suivi les cours ensemble pendant un certain nombre de jours. A chaque jour, C ont été publiquement félicités pour leur travail, D ont été publiquement blâmés et E ont été ignorés. A la fin de la période d'étude, tous les élèves ont passé un test standard dont les résultats sont donnés dans le tableau 1 et affiché à gauche dans la Figure 1.

(a) Pour le graphique gauche de la Figure 1:

(i) expliquez comment ont été construites les boîtes à moustaches.

(ii) Qu'apprenez-vous de ce graphique?

(b) Pour le graphique droit de la Figure 1:

(i) expliquez comment est construit un QQ-plot.

(ii) Qu'apprenez-vous de ce graphique?

(c) Donnez un intervalle de confiance à 95% pour le résultat du test de la moyenne pour le groupe C, et discutez son interprétation.

(d) Si nous supposons que les résultats des tests individuels pour les groupes A et B ont pour variance 15, y a-t-il des preuves que les résultats des tests pour la moyenne des groupes A et B sont différents?

Exercise 4. Table 2 shows the numbers of goals scored by FC Sion in six rounds of recent Swiss Cup soccer competitions.

Table 2: Numbers of goals scored by FC Sion in the rounds of the Swiss Cup, 2008–2017. A dash (–) indicates that they were defeated in a previous round, and a match with extra time or a penalty shootout is indicated by †; in the case of a penalty shootout, the score given is that at the end of extra time.

Year of final	Number of teams in round					
	64	32	16	8	4	2
2008	7	2	1	–	–	–
2009	5	5	1	2	1†	3
2010	1	1†	–	–	–	–
2011	4	5	3†	2	2	2
2012	5	2	2	3	0	–
2013	1	3	4	2	0	–
2014	3	1	0	–	–	–
2015	3†	1	3	2	1	3
2016	2	2	3†	2†	0	–
2017	6	4	5†	5	0†	0

(a) Briefly discuss the use of the (i) hypergeometric, (ii) binomial, (iii) exponential, and (iv) Poisson distributions to represent football scores.

(b) On the assumption that the numbers of goals may be modelled by a Poisson distribution with parameter λ, derive the maximum likelihood estimate of λ and use it to give a 95% confidence interval for λ based on the data in Table 2. Discuss.

(c) A test of the null hypothesis that λ is constant against the alternative hypothesis that the mean number of goals scored depends upon the round gives p-value $p_{\text{obs}} = 0.003$. Give an interpretation of this number. What do you conclude?

(d) Do you think the analyses in (b) and (c) seem reasonable?

Français:
Le tableau 2 montre les nombres de buts marqués par le FC Sion aux six rounds de la Coupe Suisse de football des récentes années.

(a) Discutez brièvement l'utilisation des distributions (i) hypergéométrique, (ii) binomiale, (iii) exponentielle, et (iv) Poisson pour représenter les nombres de buts.

(b) En supposant que les nombres de buts peuvent être modélisés par une distribution de Poisson de paramètre λ, déterminez l'estimateur du maximum de vraisemblance de λ et ainsi donner un intervalle de confiance à 95% pour λ basé sur les données du tableau 2. Discutez.

(c) Un test d'hypothèse nulle que λ est constant contre l'hypothèse alternative que le nombre moyen de buts marqués dépend du 'round' donne pour p-valeur $p_{\text{obs}} = 0.003$. Donnez une interprétation de ce nombre. Qu'en concluez-vous?

(d) Pensez-vous que les analyses en (b) et (c) soient raisonnables?