Department of Mathematics, IST — Probability and Statistics Unit

Probability Theory

2nd. Test ("Recurso") 1st. Semester — **2014/15**

Duration: **1h30m** **2015/01/24** — **9:45 AM**, Room P1

- Please justify your answers.
- This test has TWO PAGES and THREE GROUPS. The total of points is 20.0.

Group I — Independence **3.5 points**

The water of a certain reservoir is depleted at a constant rate of 1000 units daily. The reservoir is refilled by randomly occurring rainfalls; they occur according to a Poisson process with rate 0.2 per day; and the amount of water added to the reservoir by a rainfall is 5000 units with probability 0.8 or 8000 units with probability 0.2. The present water level is at 5000 units.

(a) Define in some detail the r.v. $W(t)$ representing the water level of the reservoir at time t. **(1.0)**

- **Auxiliary r.v. and stochastic process**
  X_i = water refill by the i^{th} rainfall
  $X_i \stackrel{i.i.d.}{\sim} X$, $i \in \mathbb{N}$
  $$P(X=x) = \begin{cases} 0.8, & x = 5000 \\ 0.2, & x = 8000 \\ 0, & x \neq 5000, 8000 \end{cases}$$
  $N(t)$ = number of rainfalls up to time (in days) t
  $\{N(t),\, t \geq 0\} \sim PP(\lambda = 0.2)$ and independent of the X_i
  $N(t) \sim \text{Poisson}(\lambda\, t = 0.2t)$
- **Defining $W(t)$**
  Since the water of the reservoir is depleted at a constant rate of 1000 units daily, *etc.*, the water level of the reservoir at time t is given by
  $$W(t) = \max\left\{0, 5000 - 1000\,t + \sum_{i=1}^{N(t)} X_i\right\}$$
  (current level − water depleted + water from the rainfalls).[1]

(b) Find the probability that the reservoir will be empty within the next five days? **(1.0)**

- **Event**
  Capitalizing on the fact that the r.v. X_i are positive, the reservoir will be empty within the next five days iff

[1]Note: $\sum_{i=1}^{0} X_i = 0$; $\sum_{i=1}^{N(t)} X_i$ is a compound PP.

$$\begin{aligned} W(5) &= 0 \\ 5000 - 1000 \times 5 + \sum_{i=1}^{N(5)} X_i &\leq 0 \\ \sum_{i=1}^{N(5)} X_i &\leq 0 \\ N(5) &= 0, \end{aligned}$$

that is, iff it does not rain in those 5 days.

- **Requested probability**

$$\begin{aligned} P[W(5)=0] &= P[N(5)=0] \\ &\stackrel{N(5)\sim Poisson(0.2\times 5)}{=} F_{Poisson(1)}(0) \\ &\stackrel{tables}{=} 0.3679. \end{aligned}$$

(c) What is the probability that the reservoir will be empty within the next ten days? **(1.5)**

- **Event**
  Considering that the r.v. X_i take values 5000 and 8000, the reservoir will be empty within the next ten days iff

$$\begin{aligned} W(10) &= 0 \\ 5000 - 1000 \times 10 + \sum_{i=1}^{N(10)} X_i &\leq 0 \\ \sum_{i=1}^{N(10)} X_i &\leq 5000 \\ N(10) &= 0 \quad \text{or} \quad (N(10) = 1 \text{ and } X_1 = 5000). \end{aligned}$$

- **Requested probability**

$$\begin{aligned} P[W(10)=0] &\stackrel{N(10)\perp\!\!\!\perp X_i}{=} P[N(10)=0] + P[N(10)=1] \times P(X_1 = 5000) \\ &\stackrel{N(10)\sim Poisson(0.2\times 10)}{=} F_{Poisson(2)}(0) + [F_{Poisson(2)}(1) - F_{Poisson(2)}(0)] \times 0.8 \\ &\stackrel{tables}{=} 0.1353 + (0.4060 - 0.1353) \times 0.8 \\ &= 0.35186. \end{aligned}$$

Group II — Expectation **10.0 points**

1. Let: $X \sim \text{Uniform}(0,1)$; $0 < a < b < 1$; $Y = X^{-1} \times I_{(0,b)}(X)$; $Z = (1-X)^{-1} \times I_{(a,1)}(X)$.

(a) Show that $Y, Z \notin L^1$ and yet $Y \times Z \in L^1$. **(2.0)**

- **R.v.**
  $X \sim \text{Uniform}(0,1)$
  $$Y = X^{-1} \times I_{(0,b)}(X) = \begin{cases} X^{-1}, & 0 < X < b \quad (0 < b < 1) \\ 0, & \text{otherwise} \end{cases}$$

$$Z = (1-X)^{-1} \times I_{(a,1)}(X) = \begin{cases} (1-X)^{-1}, & a < X < 1 \quad (0<a<1) \\ 0, & \text{otherwise} \end{cases}$$

Obs.: $0 < a < b < 1$.

- **To prove**
  $Y, Z \notin L^1$
- **Proof**
  Since $0 < a, b < 1$,

$$\begin{aligned} E(|Y|) &= E[|X^{-1} \times I_{(0,b)}(X)|] \\ &\overset{Cor. 4.81}{=} \int_{-\infty}^{+\infty} |x^{-1} \times I_{(0,b)}(x)| \times f_X(x)\, dx \\ &\overset{X\sim Uniform(0,1)}{=} \int_0^b \frac{1}{x}\, dx \\ &= \ln(x)|_0^b \\ &= +\infty \\ E(|Z|) &= E[|(1-X)^{-1} \times I_{(a,1)}(X)|] \\ &\overset{Cor. 4.81}{=} \int_{-\infty}^{+\infty} |(1-x)^{-1} \times I_{(a,1)}(x)| \times f_X(x)\, dx \\ &\overset{X\sim Uniform(0,1)}{=} \int_a^1 \frac{1}{1-x}\, dx \\ &= -\ln(1-x)|_a^1 \\ &= +\infty, \end{aligned}$$

i.e., Y and Z are not integrable r.v. ($Y, Z \notin L^1$). QED

- **Checking whether $Y \times Z \in L^1$**
  Recalling that $0 < a < b < 1$, we obtain

$$\begin{aligned} E(|Y \times Z|) &= E[|X^{-1} \times I_{(0,b)}(X) \times (1-X)^{-1} \times I_{(a,1)}(X)|] \\ &\overset{Cor. 4.81}{=} \int_{-\infty}^{+\infty} |x^{-1} \times I_{(0,b)}(x) \times (1-x)^{-1} \times I_{(a,1)}(x)| \times f_X(x)\, dx \\ &\overset{X\sim Uniform(0,1)}{=} \int_a^b \frac{1}{x(1-x)}\, dx \\ &= \int_a^b \left(\frac{1}{x} + \frac{1}{1-x}\right) dx \\ &= \ln(x)|_a^b - \ln(1-x)|_a^b \\ &= \ln(b/a) + \ln[(1-a)/(1-b)] \\ &< +\infty, \end{aligned}$$

that is, $Y \times Z$ is an integrable r.v. ($Y \times Z \notin L^1$). QED

(b) Does the previous example contradict Cauchy-Schwarz's inequality? Justify your answer. (**0.5**)

- **Comment**
  Cauchy-Schwarz's inequality provides sufficient conditions on r.v. Y and Z to be dealing with an integrable product $Y \times Z$: we must be dealing with $Y, Z \in L^2$.

  However, in the previous example we have $Y, Z \notin L^1$ and therefore $Y, Z \notin L^2$, by Lyapunov's inequality ($L^2 \subseteq L^1$), therefore we cannot even aply such inequality let alone contradict it.

2. (a) Prove (from SCRATCH!) that if X is an absolutely continuous and nonnegative r.v. in L^1 then $P(X \geq a) \leq \frac{E(X)}{a}$, for $a > 0$. (**1.5**)

- **To prove**
  X absolutely continuous, $X \geq 0$, $X \in L^1 \quad \Rightarrow \quad P(X \geq a) \leq \frac{E(X)}{a}$, for $a > 0$
- **Proof**

$$\begin{aligned} P(X \geq a) &= \int_a^{+\infty} f_X(x)\, dx \\ &\overset{x\geq a,\, a>0,\, x/a\geq 1}{\leq} \int_a^{+\infty} \frac{x}{a} f_X(x)\, dx \\ &\overset{x\, f_X(x)\geq 0}{\leq} \frac{1}{a}\int_0^{+\infty} x\, f_X(x)\, dx \\ &= \frac{E(X)}{a}. \end{aligned}$$

QED

(b) Now, admit that $X \in L^2$. (**1.0**)

For which values of a can the previous upper bound be surely improved?

- **Another upper bound for $P(X \geq a)$**
  By invoking Markov's inequality with $p = 2$, we get

$$\begin{aligned} P(X \geq a) &\overset{X\geq 0, a>0}{=} P(|X| \geq a) \\ &\leq \frac{E(X^2)}{a^2}. \end{aligned}$$

- **Comparing the upper bounds**
  By taking advantage of the characteristics of the r.v. X, we can add that the latter bound is stricter than the one we obtained in (a) for

$$\begin{aligned} a \;:\; & \frac{E(X^2)}{a^2} < \frac{E(X)}{a} \\ & a > \frac{E(X^2)}{E(X)}. \end{aligned}$$

3. Let $X_1, X_2, \ldots$ be i.i.d. r.v. to X, whose p.d.f. is $f_X(x) = \frac{1}{2} \times e^{-|x|}$, $x \in \mathbb{R}$. Suppose we continue sampling from this population until a negative observation appears. Let S be the sum of the observations thus obtained (including the negative one). (**2.5**)

Obtain the expected value and the variance of S.

Hint: Recall that $E(S) = E[E(S|N)]$ and $V(S) = V[E(S|N)] + E[V(S|N)]$, where N is the random number of observations collected (including the negative one).

- **R.v. and common p.d.f., expected value and variance**
  $X_i \overset{i.i.d.}{\sim} X$, $i \in \mathbb{N}$

$f_X(x) = \frac{1}{2} \times e^{-|x|}$, $x \in \mathbb{R}$
Moreover, since the p.d.f. is symmentric around the origin,

$$\begin{aligned} E(X) &= 0 \\ V(X) &= E(X^2) \\ &\overset{Cor. 4.81}{=} \int_{-\infty}^{+\infty} x^2 \times f_X(x)\,dx \\ &= \int_{-\infty}^{+\infty} x^2 \times \frac{1}{2} \times e^{-|x|}\,dx \\ &= \int_{-\infty}^{0} x^2 \times \frac{1}{2} \times e^{x}\,dx + \int_{0}^{+\infty} x^2 \times \frac{1}{2} \times e^{-x}\,dx \\ &= 2\int_{0}^{+\infty} x^2 \times \frac{1}{2} \times e^{-x}\,dx \\ &= \int_{0}^{+\infty} x^{3-1} \times e^{-x}\,dx \\ &\overset{(2.45)}{=} \Gamma(3) \\ &= (3-1)! \\ &= 2. \end{aligned}$$

- **Other r.v.**
  N = number of observations collected until a negative observation appears (including the negative one)
  $N \sim \text{Geometric}(p)$, where $p = P(X < 0) = \frac{1}{2}$ because the p.d.f. is symmentric around the origin.
  $S = \sum_{i=1}^{N} X_i$ = sum of the observations thus obtained (including the negative one)
- **Requested expected value and variance**
  [Taking into account that N is independent of the X_i,] $E(S|N)$ is a r.v. that takes value

$$\begin{aligned} E(S|N=n) &= E\left(\sum_{i=1}^{n} X_i\right) \\ &= n \times E(X) \\ &= 0 \end{aligned}$$

with probability 1, for any $n \in \mathbb{N}$. Furthermore, $V(S|N)$ is a r.v. that takes value

$$\begin{aligned} V(S|N=n) &= V\left(\sum_{i=1}^{n} X_i\right) \\ &= n \times V(X) \\ &= 2n, \end{aligned}$$

with probability $P(N=n) = (1-p)^{n-1}\,p$, $n \in \mathbb{N}$. As a consequence,

$$\begin{aligned} E(S) &= E[E(S|N)] \\ &= 0 \end{aligned}$$

$$\begin{aligned} V(S) &= V[E(S|N)] + E[V(S|N)] \\ &= V(0) + E(2N) \\ &\overset{N\sim Geo(p)}{=} 2 \times \frac{1}{p} \\ &\overset{(p=1/2}{=} 4. \end{aligned}$$

4. Visitors of a natural park choose at random one out of 6 paths and do it independent of one another.

(a) Find the probability that paths 1 and 2 are chosen twice each, paths 3, 4 and 5 are chosen once each, and path 6 is not chosen at all by the 7 visitors. **(1.5)**

- **Random vector**
  $\underline{N} = (N_1, \ldots, N_6)$
  N_i = number of times path i is chosen by n visitors, $i = 1, \ldots, 6$
- **Distribution**
  $\underline{N} \sim \text{Multinomial}_{d-1}(n, \underline{p})$
- **Parameters**
  $d = 6$
  $n = 7$
  $\underline{p} = (p_1, \ldots, p_6) = (1/6, \ldots, 1/6)$ (because the paths are chosen at random)
- **Joint p.f.**

$$\begin{aligned} P(\underline{N} = \underline{n}) &= P(N_1 = n_1, \ldots, N_6 = n_6) \\ &\overset{Def. 4.208}{=} \frac{7!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} \left(\frac{1}{6}\right)^{n_i}, \end{aligned}$$

  for $n_i \in \{0, 1, \ldots, 7\}$, $i = 1, \ldots, 6$, such that $\sum_{i=1}^{6} n_i = 7$.
- **Requested probability**

$$\begin{aligned} P(N_1 = 2, N_2 = 1, N_3 = 1, N_4 = 1, N_5 = 1, N_6 = 0) &= \frac{7!}{2!\,2!\,1!\,1!\,1!\,0!} \left(\frac{1}{6}\right)^{2+2+1+1+1+0} \\ &\simeq 0.004501. \end{aligned}$$

(b) Obtain the expected value and variance of the number of visitors who choose path 1 given that 3 of the 7 visitors chose path 6. **(1.0)**

- **R.v.**
  $(N_1 \mid N_6 = n_6) \overset{Rem. 4.223}{\sim} \text{Binomial}\left(n - n_6, \frac{p_1}{1-p_6}\right)$
- **Requested expected value and variance**

$$\begin{aligned} E(N_1 \mid N_6 = 5) &\overset{Rem. 4.223 \text{ or } Prop. 4.222}{=} (7-3)\,\frac{\frac{1}{6}}{1-\frac{1}{6}} \\ &\simeq 0.8 \\ V(N_1 \mid N_6 = 5) &= (7-3)\,\frac{\frac{1}{6}}{1-\frac{1}{6}}\left(1 - \frac{\frac{1}{6}}{1-\frac{1}{6}}\right) \\ &\simeq 0.64 \end{aligned}$$

Group III — Convergence of sequences of r.v. **6.5 points**

1. The Pareto distribution, named after the Italian economist Vilfredo Pareto, was originally used to model the wealth of individuals, X. We say that $X \sim \text{Pareto}(b, \alpha)$ if **(3.0)**

$$f_X(x) = \frac{\alpha\, b^\alpha}{x^{\alpha+1}},\ x \geq b,$$

where $b > 0$ is the minimum possible value of X and $\alpha > 0$ is called the Pareto index. Let:

- $\{X_1, X_2, \dots\}$ be a sequence of i.i.d. individual wealths with a common Pareto$(b = 1, \alpha)$ distribution;
- $Y_n = \frac{1}{n^{\frac{1}{\alpha}}} \max_{i=1,\dots,n} X_i$ be the scaled maximum wealth of the n first individuals.

Show that $\{Y_1, Y_2, \dots\}$ converges in distribution to Y, whose c.d.f. is given by $F_Y(y) = e^{-y^{-\alpha}},\ y > 0$.[2]

- **R.v.**
  $X_i \overset{i.i.d.}{\sim} X,\ i \in \mathbb{N}$
  $X \sim \text{Pareto}(1, \alpha),\ \alpha > 0$
- **Common p.d.f. and c.d.f.**

$$f_X(x) = \begin{cases} 0, & x < 1 \\ \frac{\alpha}{x^{\alpha+1}}, & x \geq 1 \end{cases}$$

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \int_1^x \alpha\, t^{-\alpha-1}\, dt = -\, t^{-\alpha}\big|_1^x = 1 - x^{-\alpha}, & x \geq 1 \end{cases}$$

- **Range**
  $\mathbb{R}_X = [1, +\infty)$
- **Another r.v.**
  $Y_n = \frac{1}{n^{\frac{1}{\alpha}}} \max_{i=1,\dots,n} X_i = n^{-\frac{1}{\alpha}} X_{(n)}$
- **Range**
  $\mathbb{R}_{Y_n} = \left[n^{-\frac{1}{\alpha}}, +\infty\right)$
  $\mathbb{R}_{Y_n} \to \mathbb{R}^+$, as $n \to +\infty$
- **C.d.f.**
  For $y \geq n^{-\frac{1}{\alpha}}$,

$$\begin{aligned} F_{Y_n}(y) &= P\left[n^{-\frac{1}{\alpha}} X_{(n)} \leq y\right] \\ &= P\left[X_{(n)} \leq n^{\frac{1}{\alpha}}\, y\right] \\ &\overset{Example\ 3.67}{=} \left[F_X\left(n^{\frac{1}{\alpha}}\, y\right)\right]^n \\ &= \left[1 - \left(n^{\frac{1}{\alpha}}\, y\right)^{-\alpha}\right]^n \\ &= \left(1 - \frac{y^{-\alpha}}{n}\right)^n. \end{aligned}$$

[2] I.e., the Pareto distribution belongs to the domain of attraction of the Fréchet distribution.

- **Checking the convergence in distribution**
  Since
  (i) $\lim_{n\to+\infty} F_{Y_n}(y) = \lim_{n\to+\infty} \left(1 + \frac{-y^{-\alpha}}{n}\right)^n = e^{-y^{-\alpha}},\ y > 0,$
  (ii) $F_Y(y) = e^{-y^{-\alpha}},\ y > 0$, is the c.d.f. of the absolutely continuous r.v. Y,[3]
  (iii) $\lim_{n\to+\infty} F_{Y_n}(y) = F_Y(y)$, for all the continuity points of the c.d.f. of Y,
  we can conclude that

$$Y_n \overset{d}{\to} Y.$$

2. Based on information from a previous study, a group of ornithologists believe that approximately 55% of the red tail hawk population consists of female hawks.

(a) What is the relative error if we apply the DeMoivre-Laplace local limit theorem to approximate the probability that exactly 20 red tail hawks are females if the ornithologists sample 40 red tail hawks? Comment. **(1.5)**

- **R.v.**

$$X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ red tail hawk is a female} \\ 0, & \text{otherwise} \end{cases}$$

  $X_i \overset{i.i.d.}{\sim} X,\ i = 1, 2, \dots, n\ (n = 40)$
  $X \sim \text{Bernoulli}(p = 0.55)$
- **Another r.v.**
  $S_n = \sum_{i=1}^n X_i =$ number of females in a sample of size n red tail hawks
  $S_n \sim \text{Binomial}(n, p)$
  $P(S_n = s) = \binom{n}{s} p^s\, (1-p)^{n-s},\ s = 0, 1, \dots, n$
- **Requested probability**

$$\begin{aligned} P(S_{40} = 20) &= \binom{40}{20} \times 0.55^s \times (1 - 0.55)^{40-20} \\ &= 0.102542. \end{aligned}$$

- **Approximate value**
  Assuming the DeMoivre-Laplace local limit theorem (Theorem 5.177) can be applied then

$$\begin{aligned} P(S_{40} = 20) &\overset{(5.68)}{\simeq} \frac{1}{\sqrt{40 \times 0.55 \times (1 - 0.55)}} \times \phi\left[\frac{20 - 40 \times 0.55}{\sqrt{40 \times 0.55 \times (1 - 0.55)}}\right] \\ &= \frac{1}{\sqrt{40 \times 0.55 \times (1 - 0.55)}} \\ &\quad \times \frac{1}{\sqrt{2\,\pi}} \exp\left\{-\frac{1}{2}\left[\frac{20 - 40 \times 0.55}{\sqrt{40 \times 0.55 \times (1 - 0.55)}}\right]^2\right\} \\ &\simeq 0.103599. \end{aligned}$$

[3] The footnote leads us to believe that $Y \sim$ Fréchet(α) (this is indeed the case!).

- **Relative error and comment**
  It is given by

$$\frac{0.103599 - 0.102542}{0.102542} \times 100\% = 1.030797\%$$

  and rather small, suggesting that the approximation following from the DeMoivre-Laplace local limit theorem is very reasonable in this case.

(b) Find the minimal sample size n so that the sample will contain more than 50% females, with a probability of at least 90%. (**2.0**)

Hint: It may be useful to invoke the Lindeberg-Lévy Central Limit Theorem.

- **R.v.**
  $X_i = \begin{cases} 1, & \text{if the } i^{th} \text{ red tail hawk is a female} \\ 0, & \text{otherwise} \end{cases}$
  $X_i \overset{i.i.d.}{\sim} X,\ i = 1, 2, \ldots$
  $E(X) = \mu = p = 0.55$
  $V(X) = \sigma^2 = p(1-p) < \infty$

- **Another r.v.**
  $S_n = \sum_{i=1}^{n} X_i =$ number of females in a sample of size n red tail hawks

- **Approximate distribution of S_n**
  Let $\{Z_1, Z_2, \ldots\}$ be the sequence of the standardized partial sums, where $Z_n = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$. Then, according to the Lindeberg-Lévy CLT, $Z_n \overset{d}{\to}$ Normal$(0,1)$. Moreover, for a sufficiently large value of n,

$$P(S_n \leq s) \simeq \Phi\left(\frac{s - n\mu}{\sqrt{n\sigma^2}}\right).$$

- **Requested number**
  This number can be obtained approximately:

$$\begin{aligned}
n \in \mathbb{N} \ : \ & P(S_n > 0.5n) \geq 0.9 \\
& 1 - \Phi\left[\frac{0.5n - n \times p}{\sqrt{n \times p \times (1-p)}}\right] \geq 0.9 \\
& \Phi^{-1}(1 - 0.9) \geq \sqrt{n}\,\frac{0.5 - p}{\sqrt{p \times (1-p)}} \\
& -\Phi^{-1}(0.9) \geq \sqrt{n}\,\frac{0.5 - 0.55}{\sqrt{0.55 \times (1 - 0.55)}} \\
& n \geq \left[\frac{1.2816 \times \sqrt{0.55 \times (1 - 0.55)}}{|0.5 - 0.55|}\right]^2 \\
& n \geq 162.607357,
\end{aligned}$$

  hence, the minimal sample size is $n = 163$.
  [Note that $1 - \Phi\left[\frac{0.5 \times 163 - 163 \times 0.55}{\sqrt{163 \times 0.55 \times (1 - 0.55)}}\right] \simeq 0.900279 \geq 0.9$.]