42.510 Statistical Learning for Data Science FALL 2025

Assignment 1 solutions

Probability review

1. A multiple choice quiz has 15 questions with 4 choices for each. If you guess every answer, what is the probability that you score at least 50% on the quiz?

   Solution: This can be modeled using a binomial distribution, where the probability of success is $\frac{1}{4}$, and we require 8 or more successes in 15 trials. If X denotes the number of correct answers among 15 questions then the probability is given by

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{15} \binom{15}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{15-k} \approx 0.0173.$$

2. An instructor is going to give the grades A, B, C, D, F according to the following scale

   (a) A: grade $> \mu + 1.5\sigma$;

   (b) B: $\mu + 0.5\sigma <$ grade $\leq \mu + 1.5\sigma$;

   (c) C: $\mu - 0.5\sigma <$ grade $\leq \mu + 0.5\sigma$;

   (d) D: $\mu - 2\sigma <$ grade $\leq \mu - 0.5\sigma$;

   (e) F: grade $\leq \mu - 2\sigma$.

   What percentage of students get each letter grade assuming that the grades follow a normal distribution $\mathcal{N}(\mu, \sigma^2)$?

   Solution: Let Φ denote the cumulative distribution function of $\mathcal{N}(0,1)$ distribution and F denote the cumulative distribution function of a $\mathcal{N}(\mu, \sigma^2)$ distribution. For any value of α, it holds that $F(\mu + \alpha\sigma) = \Phi\left(\frac{\mu + \alpha\sigma - \mu}{\sigma}\right) = \Phi(\alpha)$. Also $\lim_{x \to \infty} F(x) = 1$ and $\lim_{x \to -\infty} F(x) = 0$.

   (a) A: $\lim_{x \to \infty} F(x) - F(\mu + 1.5\sigma) = 1 - \Phi(1.5) = 0.06681$.

   (b) B: $F(\mu + 1.5\sigma) - F(\mu + 0.5\sigma) = \Phi(1.5) - \Phi(0.5) = 0.24173$.

   (c) C: $F(\mu + 0.5\sigma) - F(\mu - 0.5\sigma) = \Phi(0.5) - \Phi(-0.5) = 0.38292$.

   (d) D: $F(\mu - 0.5\sigma) - F(\mu - 2\sigma) = \Phi(-0.5) - \Phi(-2) = 0.28579$.

   (e) E: $F(\mu - 2\sigma) - \lim_{x \to -\infty} F(x) = \Phi(-2) - 0 = 0.02275$.

3. Let X be a Poisson random variable with mean 25. Define $p = \mathbb{P}(X \geq 32)$.

   (a) Use Markov's inequality to obtain an upper bound on p.

   (b) Use Chebyshev's inequality to obtain an upper bound on p.

   (c) Approximate p using the central limit theorem (also give a brief justification why Poisson may be approximated by a normal distribution).

   Solution:

   (a) Using Markov's inequality we have $\mathbb{P}(X \geq 32) \leq \frac{25}{32} = 0.781$.

   (b) Using Chebyshev's inequality we have $\mathbb{P}(X \geq 32) \leq \mathbb{P}(|X-25| \geq 7) \leq \frac{25}{7^2} = 0.510$.

   (c) Note that a Poisson random variable with mean λ can be thought of as a limiting sum of n independent Bernoulli trials of probabilities p such that $np \approx \lambda$. Hence we may use the Central limit Theorem here to obtain

   $$\begin{aligned}\mathbb{P}(X \geq 32) &= \mathbb{P}\left(\frac{X-25}{\sqrt{25}} \geq \frac{32-25}{\sqrt{25}}\right)\\ &\approx \mathbb{P}(Z \geq 1.4) = 0.08076.\end{aligned}$$

   where $Z \sim N(0,1)$ and we use CLT to approximate the quantity.

   Note: A answer with continuity correction using 31.5 instead of 32 is accepted as well.

4. Two types of coins are produced at a factory: a fair coin and a biased one that comes up heads 55 percent of the time. We have one of these coins, but do not know whether it is a fair coin or a biased one. In order to ascertain which type of coin we have, we shall perform the following statistical test: we shall toss the coin 1000 times. If the coin lands on heads 525 or more times, then we shall conclude that it is a biased coin, whereas if it lands on heads less than 525 times, then we shall conclude that it is a fair coin. If the coin is actually fair, what is the probability that we shall reach a false conclusion? What would it be if the coin were biased?

   Solution: Define X as the number of heads in 1000 tosses. If the coin were actually fair, then a Normal approximation to the Binomial distribution is

   $$X \overset{a.}{\sim} \mathcal{N}(500, 250),$$

   and

   $$\mathbb{P}(X > 524.5) = 1 - \mathbb{P}(X < 524.5) \approx 0.0606.^{1}$$

   In other words, the probability that a fair coin will yield 525 or more heads in 1000 tosses is 0.0606, which is the probability that we shall reach a false conclusion.

   If the coin were biased, then

   $$X \overset{a.}{\sim} \mathcal{N}(550, 247.5),$$

   and the probability of reaching a false conclusion is

   $$\mathbb{P}(X < 524.5) \approx 0.053.$$

[1] Again we can either use statistical software, or rescale it to standard Normal to get $1 - \Phi(1.55) \approx 0.0606$.

5. Let $x_1, \ldots, x_n$ be realizations of iid random variables with pdf f_θ for some parameter $\theta > 0$. For each of the following cases find the MLE of θ. You may assume $\theta > 0$.

   (a) $f_\theta(x) = \begin{cases} \theta x^{-(\theta+1)} & x \geq 1, \\ 0 & x < 1. \end{cases}$

   Solution: The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{-(\theta-1)}$$

   for $x_1, \ldots, x_n \geq 1$. Taking logarithm we have

$$\ell(\theta) = \ln L(\theta) = n \ln \theta - (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

   Taking derivative and equating to 0 we have from $\ell(\theta)' = 0$ that

$$\begin{aligned} \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i = 0, \\ \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} \ln x_i}. \end{aligned}$$

   It is easy to check that $\ell(\hat{\theta}_{\text{MLE}}) = -\frac{n}{(\hat{\theta}_{\text{MLE}})^2} < 0$ and hence $\hat{\theta}_{\text{MLE}}$ is a maxima.

   (b) $f_\theta(x) = \begin{cases} \sqrt{\theta} x^{\sqrt{\theta}-1} & 0 \leq x \leq 1, \\ 0 & \text{otherwise}. \end{cases}$

   Solution: The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \theta^{\frac{n}{2}} \left(\prod_{i=1}^{n} x_i \right)^{\sqrt{\theta}-1}$$

   for $0 \leq x_1, \ldots, x_n \leq 1$. Taking logarithm we have

$$\ell(\theta) = \ln L(\theta) = \frac{n}{2} \ln \theta + (\sqrt{\theta} - 1) \sum_{i=1}^{n} \ln x_i.$$

   Taking derivative and equating to 0 we have from $\ell(\theta)' = 0$ that

$$\begin{aligned} \frac{n}{2\theta} + \frac{1}{2\sqrt{\theta}} \sum_{i=1}^{n} \ln x_i = 0, \\ \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \left(-\frac{n}{\sum_{i=1}^{n} \ln x_i} \right)^2. \end{aligned}$$

   Again it is easy to check that $\ell(\hat{\theta}_{\text{MLE}}) < 0$ and hence $\hat{\theta}_{\text{MLE}}$ is a maxima.

(c) $f_\theta(x) = \begin{cases} \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right) & x \geq 0, \\ 0 & x < 0. \end{cases}$

Solution: The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\theta^{2n}} \left(\prod_{i=1}^{n} x_i\right) \exp\left(-\frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2\right).$$

for $x_1, \ldots, x_n \geq 0$. Taking logarithm we have

$$\ell(\theta) = \ln L(\theta) = -2n \ln\theta + \sum_{i=1}^{n} \ln x_i - \frac{1}{2\theta^2} \sum_{i=1}^{n} x_i^2.$$

Taking derivative and equating to 0 we have from $\ell(\theta)' = 0$ that

$$-\frac{2n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} x_i^2 = 0,$$

$$\Rightarrow \quad \hat{\theta}_{\text{MLE}} = \left(\frac{1}{2n} \sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

Again we can check that $\ell(\hat{\theta}_{\text{MLE}}) < 0$ and hence $\hat{\theta}_{\text{MLE}}$ is a maxima.

6. Find bias, standard error and MSE of estimators of λ and θ below.

(a) Let $X_1, \ldots, X_n$ be i.i.d. random variables $\sim$ Poisson(λ) and let $\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Solution: We can compute $\mathbb{E}X_1 = \lambda, \mathbb{E}X_1^2 = \lambda(\lambda+1)$. Hence $\mathbb{V}ar(X_1) = \lambda$. Therefore

$$\begin{aligned} \mathbb{E}(\hat{\lambda}_n) &= \lambda, \\ \text{Bias}(\hat{\lambda}_n) &= \mathbb{E}(\hat{\lambda}_n) - \lambda = \lambda - \lambda = 0, \\ \text{se}(\hat{\lambda}_n) &= \sqrt{\mathbb{V}ar(\hat{\lambda}_n)} = \sqrt{\mathbb{V}ar(X_1)/n} = \sqrt{\lambda/n}. \end{aligned}$$

Now,

$$\text{MSE}(\hat{\lambda}_n) = \text{Bias}((\hat{\lambda}_n))^2 + \text{se}((\hat{\lambda}_n))^2 = \lambda/n.$$

(b) Let $U_1, \ldots, U_n$ be i.i.d. random variables $\sim$ Unif$(0, \theta)$ and let $\hat{\theta}_n = \max\{U_1, \ldots, U_n\}$.

Solution: First we find the distribution of $\hat{\theta}_n$. Note that the support of $\hat{\theta}_n$ is $(0, 1)$. For $0 < u < 1$,

$$F(u) = \mathbb{P}(\hat{\theta}_n \leq u) = (\mathbb{P}(U_1 \leq u))^n = \frac{u^n}{\theta^n}.$$

Hence the p.d.f. of $\hat{\theta}_n$ is

$$f(u) = \frac{n}{\theta^n} u^{n-1}, \quad 0 < u < 1.$$

We can compute

$$\mathbb{E}(\hat{\theta}_n) = \int_0^\theta u \cdot \frac{n}{\theta^n} u^{n-1} \mathrm{d}u = \frac{n}{n+1}\theta, \qquad \mathbb{E}(\hat{\theta}_n^2) = \int_0^\theta u^2 \cdot \frac{n}{\theta^n} u^{n-1} \mathrm{d}u = \frac{n}{n+2}\theta^2,$$

$$\mathbb{V}ar(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n^2) - (\mathbb{E}(\hat{\theta}_n))^2 = \frac{n}{(n+2)(n+1)^2}\theta^{2\cdot}$$

Hence

$$\begin{aligned}
\mathrm{Bias}(\hat{\theta}_n) &= \mathbb{E}(\hat{\theta}_n) - \theta = -\frac{1}{n+1}\theta, \\
\mathrm{se}(\hat{\theta}_n) &= \sqrt{\mathbb{V}ar(\hat{\theta}_n)} = \sqrt{\frac{n}{n+2}\frac{\theta}{n+1}}. \\
\mathrm{MSE}(\hat{\theta}_n) &= (\mathrm{Bias}(\hat{\theta}_n))^2 + \mathbb{V}ar(\hat{\theta}_n) = \frac{2}{(n+1)(n+2)}\theta^{2.}
\end{aligned}$$

7. A person claims to be able to taste whether tea or milk was added first to a cup of English tea. To test her claim, 12 cups of visually indistinguishable tea are prepared, of which 6 of the cups are prepared tea-first, the other 6 milk-first. Being aware of this experimental setup, she would always try to pick 6 of the cups as tea-first, and the other 6 as milk-first. After tasting each cup of tea, she correctly identifies 5 of the tea-first cups (making 1 mistake), and 5 of the milk-first cups (also making 1 mistake).

   Compute the p-value, that is, the probability that one can do at least as well as her by guessing, and hence perform a hypothesis test at the $\alpha = 0.05$ level.

   Solution: Since the *tea-tasting lady* always picks exactly 6 of the cups as tea-first, and the rest as milk-first, if she identifies k of the tea-first cups correctly, then she will automatically get k of the milk-first cups right.

   So, to do at least as well as her means to either (i) correctly identify 5 of the tea-first cups (and hence also 5 of the milk-first cups); or (ii) correctly identify 6 of the tea-first cups (and hence also 6 of the milk-first cups) – that is, identify everything correctly.

   There are $\binom{12}{6} = 924$ ways to pick 6 cups from the 12 cups (ignoring the order in which they are chosen). The probability of getting (ii) by random guessing is just $1/924$. The probability of (i) is $\binom{6}{5}\binom{6}{1}/924$, as the $\binom{6}{5}$ accounts for the number of ways to get 5 of the 6 tea-first cups right, and the $\binom{6}{1}$ accounts for the number of ways to get 1 of the tea-first cups *wrong*.

   Adding up the two probabilities, we find that the p-value is $37/924 \approx 0.040 < \alpha$, therefore we conclude that there is significant evidence in support of her claim.

Statistical Decision theory

8. (KL-divergence) Let f, g be pdf's which are both supported on $A \subset \mathbb{R}$. A measure of "distance" (which is not actually a distance, mathematically) between the two distributions is given by

$$KL(f||g) = -\int_A f(x) \ln\left(\frac{g(x)}{f(x)}\right) \mathrm{d}x.$$

   This is called *Kullback-Leibler divergence*, or, *KL divergence* (or, also *relative entropy*) between f and g or X and Y where $X \sim f, Y \sim g$. Find the KL-divergence between the following distributions. Show computations. HINT: $\mathbb{E}(X) = \int_A xf(x)$, i.e. the expected value of distribution X is equal to the integral of $xf(x)$ over its support.

(a) What is the KL-divergence between $X \sim \mathcal{N}(a, \sigma^2)$ and $Y \sim \mathcal{N}(b, \sigma^2)$?

Solution: We know the support for the normal/gaussian distribution to be $x \in \mathbb{R}$, so the limits of our integral will be $-\infty$ and ∞.

Substitute the pdf for normal distribution into the KL equation, except the first term as we know the area under a pdf is 1, and its expected value is a, anyway for now we have:

$$KL(f||g) = -\int_{-\infty}^{\infty} f(x) \ln \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-b)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2\sigma^2}}} \right) \mathrm{d}x,$$

cancel some terms to get:

$$KL(f||g) = -\int_{-\infty}^{\infty} f(x) \ln \left(\frac{e^{-\frac{(x-b)^2}{2\sigma^2}}}{e^{-\frac{(x-a)^2}{2\sigma^2}}} \right) \mathrm{d}x,$$

continuing simplifying:

$$KL(f||g) = -\int_{-\infty}^{\infty} f(x) \left(\frac{((x-a)^2 - (x-b)^2)}{2\sigma^2} \right) \mathrm{d}x,$$

expand:

$$KL(f||g) = \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} f(x)(2ax - 2bx + b^2 - a^2)\mathrm{d}x,$$

not easy to get step... we use the property $\mathbb{E}(X) = \int_A x f(x)$, i.e. the expected value of distribution X is equal to the integral of $xf(x)$ over its support. The expected value here we need is a, since we are integrating over the probability space for $f(x)$, $X \sim \mathcal{N}(a, \sigma^2)$.

$$KL(f||g) = \frac{2a^2 - 2ba + b^2 - a^2}{2\sigma^2},$$

simplify to get:

$$KL(f||g) = \frac{(a-b)^2}{2\sigma^2}.$$

(b) What is the KL-divergence between $X \sim \text{Exp}(a)$ and $Y \sim \text{Exp}(b)$?

Solution: Limit here is 0 to ∞

Substitute in the pdf, except the first one again, so we can do expectation trick:

$$KL(f||g) = -\int_{0}^{\infty} f(x) \ln \left(\frac{be^{-bx}}{ae^{-ax}} \right) \mathrm{d}x,$$

simplify:

$$KL(f||g) = -\int_{0}^{\infty} f(x) \left(\ln(b) - \ln(a) + (a-b)x\right) \mathrm{d}x,$$

using $\mathbb{E}X = 1/a$, we get the answer:

$$KL(f||g) = \ln(a) - \ln(b) - 1 + \frac{b}{a}.$$

9. For a given loss function L, the risk R is given by the expected loss

$$R(f) = \mathbb{E}[L(Y; f(X))],$$

where $f = f(X)$ is a function of the random predictor variable X.

(a) Consider a regression problem and the squared error loss

$$L(Y, f(X)) = (Y - f(X))^2.$$

Derive the expression of $f = f(X)$ which minimizes the associated risk.

Solution: 2 ways, either use integral and marginal probabilities, see bishop page 46, or conditional expectations, we provide the (easier) solution using conditional expectations:

$$R(f) = \mathbb{E}\left((Y - f(X))^2 | f(X) = f(x)\right) = \mathbb{E}\left(Y^2 | f(X) = f(x)\right) - 2f(X)\mathbb{E}(Y|f(X) = f(x)) + f(X)^2$$

get partial derivative wrt Y:

$$\frac{\partial R(f)}{\partial Y} = -2\mathbb{E}\left(Y|f(X) = f(x)\right) + 2f(X) = 0$$

answer will be:

$$f(x) = \mathbb{E}(Y|f(X) = f(x))$$

(b) What if we use the absolute (L1) loss instead as below

$$L(Y, f(X)) = |Y - f(X)|,$$

what will be the minimizer?

Solution: The conditional median $f(x) = \mathit{median}\left(Y|f(X) = f(x)\right)$.

10. Consider a binary classification problem where $Y \in \{-1, 1\}$. Consider the following imbalanced loss function: $L(-1, -1) = L(1, 1) = 0$, $L(-1, 1) = C_+ > 0$ (cost of false positive), $L(1, -1) = C_- > 0$ (cost of false negative). Compute the Bayes classifier/predictor at $\mathbf{X} = \mathbf{x}$.

Solution: Given a loss function L as above the Bayes' classifier at $\mathbf{X} = \mathbf{x}$ is:

$$\begin{aligned}\widehat{Y}(\mathbf{x}) &= \arg \min_{z \in \{-1,1\}} \sum_{y \in \{-1,1\}} L(y, z)\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}) \\ &= \operatorname{sgn}(C_-\mathbb{P}(Y = +1|\mathbf{X} = \mathbf{x}) - C_+\mathbb{P}(Y = -1|\mathbf{X} = \mathbf{x})).\end{aligned}$$

Hence we have the Bayes classifier:

$$\widehat{Y}(\mathbf{x}) = \begin{cases} +1 & \text{if } C_+\mathbb{P}(Y = -1|\mathbf{X} = \mathbf{x}) \geq C_-\mathbb{P}(Y = +1|\mathbf{X} = \mathbf{x}) \\ -1 & \text{otherwise.} \end{cases}$$

Additional practice questions: NO NEED TO SUBMIT

11. Choose a number X at random from the set of numbers $\{1, 2, 3, 4, 5\}$. Now choose a number at random from the subset no larger than X, that is, from $\{1, ..., X\}$. Call this second number Y.

    (a) Find the joint mass function of X and Y.

    Solution: Note that, for any $i, j \in \{1, 2, 3, 4, 5\}$,

    $$\mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j \mid X = i).$$

    Then, since for all $i \in \{1, \cdots, 5\}$, $\mathbb{P}(X = i) = 1/5$ and j is necessarily lower than i for being drawn with positive probability for Y; we have for any $i, j \in \{1, 2, 3, 4, 5\}$,

    $$\mathbb{P}(X = i,\, Y = j) = \begin{cases} \frac{1}{5i} & \text{if } j \leq i, \\ 0 & \text{if } j > i. \end{cases}$$

    (b) Find the conditional mass function of X given that $Y = i$. Do it for $i = 1, 2, 3, 4, 5$.

    Solution: By the Bayes formula and the law of total probability, we have for any $i \in \{1, 2, 3, 4, 5\}$, and jli

    $$\begin{aligned} \mathbb{P}(X = j \mid Y = i) &= \frac{\mathbb{P}(X = j, Y = i)}{\mathbb{P}(Y = i)} = \frac{\mathbb{P}(Y = i \mid X = j)\mathbb{P}(X = j)}{\sum_{j=1}^{5} \mathbb{P}(Y = i \mid X = j)\mathbb{P}(X = j)} \\ &= \frac{\frac{1}{5j}}{\frac{1}{5}\sum_{k=i}^{5}\frac{1}{k}} = \frac{1/j}{\sum_{k=i}^{5}(1/k)}. \end{aligned}$$

    For $j > i$, $\mathbb{P}(X = j \mid Y = i) = 0$.
    Let $Y = 1$. For $j \in \{1, 2, 3, 4, 5\}$ we have

    $$\mathbb{P}(X = j \mid Y = 1) = \frac{\frac{1}{5j}}{1\frac{1}{5} + \frac{1}{2}\frac{1}{5} + \frac{1}{3}\frac{1}{5} + \frac{1}{4}\frac{1}{5} + \frac{1}{5}\frac{1}{5}} = \frac{1/j}{137/60} = \frac{60}{137j}.$$

    Let $Y = 2$. For $j \in \{2, 3, 4, 5\}$ we have

    $$\mathbb{P}(X = j \mid Y = 2) = \frac{\frac{1}{5j}}{0 + \frac{1}{2}\frac{1}{5} + \frac{1}{3}\frac{1}{5} + \frac{1}{4}\frac{1}{5} + \frac{1}{5}\frac{1}{5}} = \frac{1/j}{77/60} = \frac{60}{77j}.$$

    Let $Y = 3$. For $j \in \{3, 4, 5\}$ we have

    $$\mathbb{P}(X = j \mid Y = 3) = \frac{\frac{1}{5j}}{0 + 0 + \frac{1}{3}\frac{1}{5} + \frac{1}{4}\frac{1}{5} + \frac{1}{5}\frac{1}{5}} = \frac{1/j}{47/60} = \frac{60}{47j}.$$

    Let $Y = 4$. For $j \in \{4, 5\}$ we have

    $$\mathbb{P}(X = j \mid Y = 4) = \frac{\frac{1}{5j}}{0 + 0 + 0 + \frac{1}{4}\frac{1}{5} + \frac{1}{5}\frac{1}{5}} = \frac{1/j}{9/25} = \frac{25}{9j}.$$

Let $Y = 5$. For $j = 5$, we have

$$\mathbb{P}(X = j \mid Y = 5) = \frac{\frac{1}{25}}{0 + 0 + 0 + 0 + \frac{1}{5}\frac{1}{5}} = 1.$$

(c) Are X and Y independent? Why?

Solution: Check that $\mathbb{P}(X = 1) = 1/5$ and $\mathbb{P}(Y = 5) = 1/25$. But $\mathbb{P}(X = 1, Y = 5) = 0 \neq \mathbb{P}(X = 1)\mathbb{P}(Y = 5)$. Hence X and Y are NOT independent.

Alternatively: *Reductio ad absurdum.* Both X and Y take values in $\{1, \cdots, 5\}$. Assume that X and Y are independent. Then, the domain of definition of (X, Y) is $\{1, \cdots 5\} \times \{1, \cdots 5\}$. But some couples, e.g. $(1, 4)$, are not in the domain since Y is constrained to be lower than X. So, X and Y are not independent.

12. A complex machine is able to operate effectively as long as at least 3 of its 5 motors are functioning. If each motor independently functions for a random amount of time with density function

$$f(x) = xe^{-x}, \quad x > 0,$$

compute the density function of the length of time that the machine functions.

Solution: For $i = 1, \ldots, 5$, let X_i denote the random variable corresponding amount of time the i^{th} motor functions. Hence X_i's are iid with pdf f, as given above. For any $t > 0$,

$$\begin{aligned} p_t = \mathbb{P}(\text{motor } i \text{ functions at time } t) &= \mathbb{P}(X_i > t) \\ &= \int_t^\infty xe^{-x}\mathrm{d}x \\ &= (-xe^{-x})\Big|_{x=t}^{x=\infty} + \int_t^\infty e^{-x}\mathrm{d}x = (t+1)e^{-t}. \end{aligned}$$

The distribution of the length of time the machine works is non-negative, hence for $t > 0$,

$$\begin{aligned} F_{\text{CM}}(t) :&= \mathbb{P}(\text{the machine stops working before or at time } t) \\ &= 1 - \mathbb{P}(\text{the machine functions after time } t) \\ &= 1 - \mathbb{P}(\text{at least three of the motors work at time } t) \\ &= 1 - \sum_{k=3}^{5} \binom{5}{k} p_t^k (1 - p_t)^{5-k} \\ &= 1 - 10p_t^3 + 15p_t^4 - 6p_t^5 \\ &= 1 - 10(t+1)^3e^{-3t} + 15(t+1)^4e^{-4t} - 6(t+1)^5e^{-5t}. \end{aligned}$$

Hence the pdf of length of time the machine works for any $t > 0$ is given by

$$\begin{aligned} f_{\text{CM}}(t) &= \frac{\mathrm{d}F_{\text{CM}}(t)}{\mathrm{d}t} \\ &= -30(t+1)^2e^{-3t} + 30(t+1)^3e^{-3t} + 60(t+1)^3e^{-4t} - 60(t+1)^4e^{-4t} \\ &\qquad - 30(t+1)^4e^{-5t} + 30(t+1)^5e^{-5t} \\ &= 30t(t+1)^2e^{-3t} - 60t(t+1)^3e^{-4t} + 30t(t+1)^4e^{-5t}. \end{aligned}$$

13. Ten hunters are waiting for ducks to fly by. When a flock of ducks flies overhead, the hunters fire at the same time, but each chooses his target at random, independently of the others. If each hunter independently hit their target with probability 0.6, compute the expected number of ducks that are hit. Assume that the number of ducks in a flock is a Poisson random variable with mean 6. You may provide the final expression or an approximation to the solution.

    Solution: We approach this question in parts. Let N be a random variable that describes the number of ducks observed, where $N \sim \text{Pois}(6)$. Let D be a random variable describing the number of ducks hit by the hunters. We can see that D depends (is conditional) on N. By the law of total expectation (or the tower property), $\mathbb{E}[D] = \mathbb{E}[\mathbb{E}[D|N]]$. Thus we can write that

    $$\mathbb{E}[D] = \sum_{n=1}^{\infty} \mathbb{E}[D|N = n]\mathbb{P}(N = n).$$

    We know from the pmf of a Poisson distribution that $\mathbb{P}(N = n) = \frac{e^{-6}6^n}{n!}$. Now we find $\mathbb{E}[D|N = n]$. Since each duck is hit with independent probability p, $D \sim \text{Bin}(n, p)$.

    Denote q as the probability that a given duck is *not* hit by a given hunter. Then a duck is hit unless all 10 hunters do not hit it; that is, a duck is hit with probability $p = 1 - q^{10}$. A duck is not hit if the hunter does not target it, or if the hunter targets it and misses. Since a hunter targets a duck at random, he targets a given duck with probability $\frac{1}{n}$. Then

    $$\begin{aligned} q &= \mathbb{P}(\text{Hunter does not target duck}) + \mathbb{P}(\text{Hunter misses duck}|\text{Hunter targets duck}) \\ &= (1 - \frac{1}{n}) + (0.4 \times \frac{1}{n}) \\ &= 1 - \frac{0.6}{n} \end{aligned}$$

    and $p = 1 - \left(1 - \frac{0.6}{n}\right)^{10}$. Since $D \sim \text{Bin}(n, p)$, $\mathbb{E}[D|N = n] = np$. Putting it all together we have

    $$\mathbb{E}[D] = \sum_{n=1}^{\infty} n\left(1 - \left(1 - \frac{0.6}{n}\right)^{10}\right) \cdot \frac{e^{-6}6^n}{n!}.$$

    We can evaluate this expression numerically by setting an appropriate upper limit for the summation. If we plot $\mathbb{E}[D|N = n]\mathbb{P}(N = n)$ over a range of values for n, we can see that the values become very small for $n > 25$. If we evaluate the summation for $n \in \{1, 2, \ldots 25\}$, we have $\mathbb{E}[D] \approx 3.6989$.[2]

14. Let $X_n, n = 1, 2, \ldots$ be random variables such that $\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2}$ and $\mathbb{P}\left(X_n = n\right) = \frac{1}{n^2}$. Does X_n converge in probability? Does it converge in L^2?

    Solution: Let $\epsilon > 0$. Then for all $n \in \mathbb{N}$, such that $\epsilon > \frac{1}{n}$, we have

    $$\mathbb{P}(|X_n| > \epsilon) \le \mathbb{P}\left(|X_n| > \frac{1}{n}\right) = \mathbb{P}(X_n = n) = \frac{1}{n^2} \to 0 \quad \text{as } n \to \infty.$$

    Hence $X_n \to 0$ in probability. On the other hand,

[2] See example code at `http://www.r-fiddle.org/#/fiddle?id=1PhPYebV&version=1`.

$$\mathbb{E}(X_n - 0)^2 = \mathbb{E}X_n^2 = \frac{1}{n^2} \times \left(1 - \frac{1}{n^2}\right) + n^2 \times \frac{1}{n^2} \to 1 \quad \text{as } n \to \infty.$$

Thus X_n does not converge in L^2.

15. If $Z \sim \mathcal{N}(0,1)$, find $\text{Cov}(Z, Z^2)$.

Solution: First, recall that since $Z \sim N(0,1)$, $\mathbb{E}[Z] = 0$ and $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathbb{E}[Z^2] = 1$. We also know that for any random variables X and Y, $\mathbf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

So, $\mathbf{Cov}(Z, Z^2) = \mathbb{E}[Z^3] - \mathbb{E}[Z]\mathbb{E}[Z^2] = \mathbb{E}[Z^3]$, that is the third moment of the standard normal random variable. To evaluate it, we can use the moment-generating function Φ of Z:

$$\Phi_Z(t) = \mathbb{E}[e^{tZ}] = e^{\frac{t^2}{2}}$$

Then,

$$\begin{aligned}\mathbb{E}[Z^3] &= \left(\frac{d^3\Phi_Z}{dt^3}\right)\Big|_{t=0} \\ &= \left((t + t^2 + t^3)e^{\frac{t^2}{2}}\right)\Big|_{t=0} \\ &= 0.\end{aligned}$$

16. Consider testing $\mathbf{H}_0 : \mu = 0$ vs $\mathbf{H}_\text{A} : \mu \neq 0$ based on a random sample of size n from a $\mathcal{N}(\mu, 1)$ distribution.

(a) Calculate the p-values for the following three cases:
(i) $\overline{x} = 0.1$, $n = 100$; (ii) $\overline{x} = 0.1$, $n = 400$; (iii) $\overline{x} = 0.1$, $n = 900$.

(b) Given the significance level $\alpha = 0.01$, conduct hypothesis tests for the three cases in (a).

Solution: (a) To compute the p-value, we first assume that $\mathbf{H}_0$ is true, that is, $\mu = 0$.

Then for (i), the p-value is

$$\Pr\left(|Z| \geq \left|\frac{0.1 - 0}{1/\sqrt{100}}\right|\right) = \Pr(|Z| \geq 1) = 0.3173;$$

for (ii), $\Pr(|Z| \geq 2) = 0.0455$; for (iii), $\Pr(|Z| \geq 3) = 0.0027$.

Note that the p-values are two-sided, since the alternative hypothesis is two-sided.

(b) For (i) and (ii), the p-value is greater than α, so we do not reject $\mathbf{H}_0$; for (iii), the p-value is smaller than α, so we reject $\mathbf{H}_0$.

17. A certain component is critical to the operation of an electrical system and must be replaced immediately upon failure. If the mean lifetime of this type of component is 100 hours and its standard deviation is 30 hours, how many of these components must be in stock so that the probability that the system is in continual operation for the next 2,000 hours is at least 0.95?

    Solution: If there are n components in stock, and $X_1, \dots, X_n$ denote the lifetimes of these components, then we want

    $$\Pr(\sum_{i=1}^{n} X_i > 2500) \approx 0.95.$$

    Using CLT where we have $Z \sim N(0,1)$, we want n such that,

    $$\begin{aligned} &\Pr\left(\frac{\sum_{i=1}^{n} X_i - 100n}{30\sqrt{n}} > \frac{2500-100n}{30\sqrt{n}}\right) \approx 0.95 \\ \text{or,} \;\; &\Pr\left(Z > \frac{2500-100n}{30\sqrt{n}}\right) \approx 0.95 = \Pr(Z > -1.645) \\ \text{or,} \;\; &\frac{2500-100n}{30\sqrt{n}} \approx -1.645. \end{aligned}$$

    Solving for n we get $n \approx 23.$

18. Consider non-negative integers a, b. Show that if $a \leq b$, then $a \leq \sqrt{ab}$. Now consider a two class classification problem where the elements of the i-th class are generated from a pdf $f_i, i = 1, 2$. For simplicity consider that the elements lie in $\mathbb{R}$. Suppose the decision regions (say, $\mathcal{R}_1$ and $\mathcal{R}_2$) for classification are chosen by minimizing the probability of misclassification. Show that this misclassification probability satisfies:

    $$\mathbb{P}(\text{misclassification}) \leq \int_{-\infty}^{\infty} \sqrt{f_1(x) f_2(x)}\, \mathrm{d}x.$$

    Solution: It is easy to check that if $a \leq b$ and $a, b \geq 0$ then $a^2 \leq ab$. Hence taking the positive square root on both sides we get $a \leq \sqrt{ab}$.

    Let the two classes be $\mathcal{G}_1, \mathcal{G}_2$ with respective decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$ chosen by minimizing probability of misclassification. Then

    $$\begin{aligned} \mathbb{P}(\text{misclassification}) &= \int_{-\infty}^{\infty} \mathbb{P}(\text{misclassification}, \mathbf{X} \in \mathrm{d}\mathbf{x}) \\ &= \int_{-\infty}^{\infty} \left[\mathbf{1}_{\{f_1(\mathbf{x}) \geq f_2(\mathbf{x})\}} f_2(\mathbf{x}) + \mathbf{1}_{\{f_1(\mathbf{x}) \leq f_2(\mathbf{x})\}} f_1(\mathbf{x})\right] \mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{R}_1} f_2(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \end{aligned}$$

    since $\mathcal{R}_1 \cup \mathcal{R}_2 = \mathbb{R}$ and they are disjoint. Therefore using the inequality we proved above,

    $$\begin{aligned} \mathbb{P}(\text{misclassification}) &= \int_{\mathcal{R}_1} f_2(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} f_1(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &\leq \int_{\mathcal{R}_1} \sqrt{f_1(\mathbf{x}) f_2(\mathbf{x})}\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} \sqrt{f_1(\mathbf{x}) f_2(\mathbf{x})}\, \mathrm{d}\mathbf{x} \\ &= \int_{-\infty}^{\infty} \sqrt{f_1(\mathbf{x}) f_2(\mathbf{x})}\, \mathrm{d}\mathbf{x}. \end{aligned}$$