Problem Set 1- Solutions
Precision Care Medicine(BME 580.480/680)

September 18, 2019

1. Suppose that the number of patients arriving to the ER obeys a Poisson arrival process with arrival rate of 5 patients per hour. What is the probability that at least 1 patient (1 or 2 or 3, etc.....) arrive in 3 hours? Please show your work.

   Solution: $\lambda = 5, T = 3$. Recall that $Pr(k \text{ arrivals in 3 hours}) = \frac{(3\lambda)^k e^{-3\lambda}}{k!} \quad k = 0, 1, 2, \ldots.$
   Then $Pr(k \geq 1) = 1 - Pr(k = 0) = 1 - e^{-3\lambda} = 1 - e^{-15}$.

2. Suppose that the number of patients arriving to the ER obeys a Poisson arrival process with arrival rate of 3 patients per 30 minutes. What is the probability that two patients will arrive between 9AM-9:30AM and 3 patients will arrive between 10AM-10:15AM? Please show your work.

   Solution: As the number of arrivals in two different time segments are independent:

$$P[\text{2 arrivals 9-9:30 AM} \cap \text{3 arrivals 10-10:15 AM}] = P[\text{2 arrivals 9-9:30 AM}]P[\text{3 arrivals 10-10:15 AM}] \quad (1)$$

   Let X be the number of arrivals in a time segment of length Δt, when the rate of arrivals is λ. Its probability mass function is given by:

$$P(X = k) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!} \quad (2)$$

   In this case, $\lambda = \frac{3}{30}$. In the first segment, $\Delta t = 30$, and in the second segment $\Delta t = 15$. Substituting these known values, and using Equations 1 and 2:

$$P[\text{2 arrivals 9-9:30 AM}]P[\text{3 arrivals 10-10:15 AM}] = \frac{3^2 e^{-3}}{2!} \frac{1.5^3 e^{-1.5}}{3!} \quad (3)$$

3. Suppose y is a Gaussian random variable with mean $\mu = 3$ and variance $\sigma^2 = 1$. Let $z = ay + b$. What is the probability density function for z conditioned on y?

   Solution: We know that any variable that is a linear combination of Gaussian variables is itself Gaussian. So we just need to compute the mean and variance of z.
   $E(z) = aE(y) + b = 3a + b, \quad var(z) = a^2 var(y) = a^2$. Therefore,
   z is a Gaussian random variable with mean $\mu = 3a + b$ and variance $\sigma^2 = a^2$.
   This yields: $f(z) = \frac{1}{\sqrt{2\pi a^2}} e^{\frac{-(z-3a-b)^2}{2a^2}}$

4. Suppose y is a Gaussian random variable with mean $\mu = 0$ and variance $\sigma^2 = 1$. Let $z = 3y$. Write an expression in the form of an integral to compute $Pr(z > 1)$. No need to solve the integral.

   Solution: Let Φ denote the cumulative density function of the standard Gaussian.

$$P(z > 1) = P(3y > 1) = P(y > \frac{1}{3}) = 1 - \Phi(\frac{1}{3}) \quad (4)$$

$$P(z > 1) = 1 - \int_{-\infty}^{\frac{1}{3}} \frac{e^{\frac{-z^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dz = \int_{\frac{1}{3}}^{\infty} \frac{e^{\frac{-z^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dz \quad (5)$$

5. Suppose two independent fair coins are tossed in the air. If either coin lands on tails, then that coin is tossed again until it lands on heads. Let y_i be the number of tosses of coin i for $i = 1, 2$. What is the probability that the number of total tosses equals 10, i.e., $Pr(y_1 + y_2 = 10)$? Please show your work.

   Solution: $p_1 = p_2 = 0.5$. We know that $Pr(y_i = k) = (1-p)^{k-1}p$ for $y_i = 1, 2$..
   $Pr(y_1 + y_2 = 10) = \sum_{k=1}^{9} Pr(y_1 = k)Pr(y_2 = 10 - k|y_1 = k)$
   $= \sum_{k=1}^{9} Pr(y_1 = k)Pr(y_2 = 10 - k))$, because y_1 and y_2 are independent.
   $Pr(y_1 + y_2 = 10) = \sum_{k=1}^{9}(1-p)^{k-1}p(1-p)^{10-k-1}p$
   $= p^2(1-p)^8 \sum_{k=1}^{9} 1 \quad = 9p^2(1-p)^8$

6. Suppose you toss a coin 50 times. The coin has probability of heads equal to p (unknown) and you observe heads for the first 25 tosses and tails for the last 25 tosses.

   (a) Derive the data likelihood function as a function of p

   (b) Compute the maximum likelihood estimate for p. Show all your work.

   Solution:

   (a) The likelihood function is:
   $L(y_1, ..., y_{50}; p) = \prod_{i=1} 25p \cdot \prod_{j=26}^{50}(1-p) = p^{25}(1-p)^{25}$.

   (b) To compute the ML estimate, we take the derivative of the likelihood or the log likelihood and set to 0. Then solve for p_{ML}.
   $log(L) = 25log(p) + 25log(1-p)$.
   $\frac{dlog(L)}{dp} = \frac{25}{p} - \frac{25}{1-p} = 0$ Then we get that
   $\frac{1}{p} = \frac{1}{1-p} \rightarrow p = 1 - p \rightarrow p = 0.5$.
   Therefore $p_{ML} = 0.5$.
   To check that it is a maximum and not a minimum, we take the second derivative of the log likelihood function at p_{ML} and see if that is less than 0:
   $\frac{d^2log(L)}{dp^2} = -25\frac{1}{p^2} - 25\frac{1}{(1-p)^2} =< 0$ for all $p > 0$ and $p < 1$
   Therefore $p_{ML} = 0.5$.

7. Suppose you toss a coin 100 times. The coin has probability of heads equal to p (unknown) and you observe heads for 20 tosses and tails for 80 tosses.

   (a) Derive the data likelihood function as a function of p

   (b) Compute the maximum likelihood estimate for p. Show all your work.

   Solution:

   (a) The likelihood function is:
   $L(y_1, ..., y_{100}; p) = p^{20}(1-p)^{80}$.

   (b) To compute the ML estimate, we take the derivative of the likelihood or the log likelihood and set to 0. Then solve for p_{ML}.
   $log(L) = 20log(p) + 80log(1-p)$.
   $\frac{dlog(L)}{dp} = \frac{20}{p} - \frac{80}{1-p} = 0$ Then we get that
   $\frac{1}{p} = \frac{4}{1-p} \rightarrow 4p = 1 - p \rightarrow p = 0.2$.
   Therefore $p_{ML} = 0.2$.
   To check that it is a maximum and not a minimum, we take the second derivative of the log likelihood function at p_{ML} and see if that is less than 0:
   $\frac{d^2log(L)}{dp^2} = -20\frac{1}{p^2} - 80\frac{1}{(1-p)^2} =< 0$ for all $p > 0$ and $p < 1$
   Therefore $p_{ML} = 0.2$.

8. Suppose two independent fair coins are tossed twice. For each coin, if it lands on tails, then you receive 5, else you pay 10. Let r equal the amount of money you have left (can be positive or negative) after your two tosses of the two coins. Calculate your expected earnings. Show your work.

   Solution:
   There are 3 possible outcomes: $r \in \{-20, -5, 10\}$, corresponding to 2 tails, 1 heads and 1 tails, and 2 heads. From combinatorics, we see that there is 1 way to make 2 tails, 2 ways to make 1 heads and 1 tails, and 1 way to make 2 heads. Therefore we obtain the complete probability mass function of r:

r	P(r)
-20	0.25
-5	0.5
10	0.25

(6)

   Expectation is given by $\sum_r rP(r)$, and thus:

$$E[r] = -20 \times 0.25 - 5 \times 0.5 + 10 \times 0.25 = -5 \tag{7}$$

   Thus you are expected to lose 5 dollars.

9. Suppose that the time in days until ICU discharge for a patient population recovering from cardiac arrest follows a density $f(x) = \lambda e^{-x/3}$ for $x > 0$

   (a) What value of λ would makes this density a valid pdf?
   (b) What is the probability that a patient stays in the ICU for longer than a week ?

   Solution:

   (a) In order to be a valid pdf, the density must integrate to 1:

$$1 = \int_0^\infty \lambda e^{-x/3} dx = -3\lambda e^{-x/3}\Big|_0^\infty = 3\lambda \rightarrow \lambda = 1/3$$

   (b) Probability that a patient stays in the ICU for longer than a week:

$$Pr(X > 7) = 1 - Pr(X \leq 7) = 1 - \int_0^{x=7} 1/3 e^{-x/3} dx = e^{-7/3} = 0.097$$

10. Derive the maximum likelihood estimate for λ in the Poisson process.

   Solution: Recall the Poisson probability mass distribution is:

$$P(X = n) = \frac{(\lambda \Delta t)^n e^{-\lambda \Delta t}}{n!}$$

   The likelihood function:

$$L(\lambda; n, \Delta t) = \frac{(\lambda \Delta t)^n e^{-\lambda \Delta t}}{n!}$$
$$log(L) = nlog(\lambda \Delta t) - \lambda \Delta t - log(n!)$$
$$\frac{dlog(L)}{d\lambda} = \frac{n}{\lambda} - \Delta t$$

   At local maximum λ_{ML}:

$$\frac{dlog(L)}{d\lambda} = 0 \rightarrow \lambda_{ML} = \frac{n}{\Delta t}$$

   Check for see if this is a maximum or minimum:

$$\frac{d^2 log(L)}{d\lambda^2} = -n/\lambda^2 < 0$$

   Thus, $\lambda_{ML} = n/\Delta t$.

11. Suppose you play the coin-tossing game with your friend again like in problem 8. However, this time, there is only one coin and the coin is biased with the probability of head is 0.3. To compensate for the biased coin, your friend says that she will give you 6 dollars if the coin lands on head and you give her 4 dollars if it comes up tails. There will be 100 rounds in total. What is your expected total earnings and variance of your total earnings? Show your work.

    Solution:
    Let X denote the random variable that represents your earning for one round. Then we have X = 6 with probability P(X=6) = 0.3 and X = -4 with P(X=-4) = 0.7. Your expected earning and variance for round i ($i = 1, 2, .., 100$) is:

    $$E(X_i) = \sum x(P(X_i = x) = 6(0.3) + (-4)(0.7) = -1$$

    $$Var(X_i) = E(X_i^2) - E(X_i)^2 = 6^2(0.3) + (-4)^2(0.7) - (-1)^2 = 21$$

    For 100 rounds:

    $$E(\sum_{i=1}^{100} X_i) = \sum_{i=1}^{100} E(X_i) = 100.(-1) = -100$$

    $$Var(\sum_{i=1}^{100} X_i) = \sum_{i=1}^{100} Var(X_i) = 100.21 = 2100$$

    Thus you are expected to lose 100 dollars but the variance of your total earnings is 2100 dollars.

12. Suppose that the systolic blood pressures (SBP) of Hopkins students are normally distributed with mean 110 (mmHg) and variance 16. Suppose that 20 students are randomly selected for a health exam. Suppose that the only criterion for hypertension in a young adult is defined as SBP $>$130 mmHg. Calculate the probability that 3 students or more are hypertensive?

    Solution: ** Since the integral of the probability density function (pdf) for the normal distribution is not straightforward, an answer that shows the derived formula (without actual calculation) is acceptable.

    Let X be the random variable that represents the SBP of a Hopkins student. X belongs to a normal distribution with $\mu = 110$ and $\sigma^2 = 16$ ($X \sim Normal(\mu = 110, \sigma^2 = 16)$). Recall that the pdf of normal random variables is:

    $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

    The probability that one student is hypertensive:

    $$p_{hbp} = P(X > 130) = 1 - P(X \leq 130) = 1 - \int_0^{130} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx$$

    Let Y be the random variable that represents the number of students having hypertension. We can model 20 students in the health exam as a binomial process, with probability of each student being hypertensive is p_{hbp}:

    $$Y \sim Binomial(n = 20, p = p_{hbp})$$

    The probability that 3 students or more are hypertensive out of 20 students is:

    $$P(Y \geq 3) = 1 - P(Y < 3) = 1 - (P(Y = 0) + P(Y = 1) + P(Y = 2))$$

    $$= 1 - \sum_{k=0}^{2} \binom{20}{k} p_{hbp}^k (1 - p_{hbp})^{20-k}$$

    ** For the actual calculation of p_{hbp}, one could use Taylor's series to approximate the integral of $e^{-x^2/2}$. Another (and much easier) way is to use R's function $pnorm$:

    $$p_{hbp} = P(X > 130) = 1 - P(X \leq 130) = 1 - pnorm(130, mean = 110, sd = 4) = 2.87e^{-7}$$

13. Suppose that heart rates obtained from ICU patient data is normally distributed with a mean of 86 beats/second and standard deviation of 49 beats/second. Knowing that this includes heart rates that are physiologically impossible, you decide that only the heart rates between 20 beats/second and 220 beats/second are valid.

a) What proportion of the data will be within the valid range?

Suppose that you select a random patient and find that during their ICU stay, there were 1000 heart rate measurements.

b) What is the probability that 90 percent of the data points are within the valid range.

Solution:

a) Let X denote the random variable that represents the value of a heart rates data point. We need to find $P(20 \leq X < 220)$. One could use $pnorm()$ like in the previous problem and get:

$$P(20 \leq X < 220) = pnorm(220, 86, 49) - pnorm(20, 86, 49) = 0.908$$

Or 91 percent of all ICU patient data are within the valid range. Another method is to Z-score and use a Z-score table $pnorm()$ to get the area under the normal distribution curve. Of course, one could feel free to use Taylor series.

$$z = \frac{x - \mu}{\sigma}$$
$$z_1 = \frac{20 - 86}{49} = -1.347$$
$$z_2 = \frac{220 - 86}{49} = 2.734$$

z_1 corresponds to the left most area under the standard normal distribution of $pnorm(-1.347, 0, 1) = 0.0890$, and z_2 corresponds to the left most area under the standard normal distribution of $pnorm(2.734, 0, 1) = 0.9968$. Therefore, the proportion of data within the valid range is 0.9968 - 0.0890 = 0.9078 - about 91% of the data is therefore valid given the standard distribution.

b) Let Y denote the random variable that represents the number of data points of all 1000 data points that are within the valid range. Assume that each data point is iid. According to part a, the probability of a single data point is valid is 0.91. We can model the process as a binomial process, with probability of each data point being physiologically possible is $p_{valid} = 0.91$. We need to find the find the joint probability that 900 data points are valid and 100 data points are invalid:

$$P(Y = 900) = \binom{1000}{900}(0.91)^{900}(0.09)^{100}$$

(**One could use R $dbinom()$ function or $choose()$ function to calculate the exact number:

$$choose(1000, 900)0.91^{900}0.09^{100} = 0.023$$

or

$$(dbinom(1000, 900, 0.91)) = 0.023.)$$

14. Suppose that on average 1 in-hospital cardiac arrests occur per day. What is the probability that the hospital experiences:

a) no in-hospital cardiac arrest for an entire week.

b) 5 in-hospital cardiac arrest in 3 days.

Solution:

a) The rate of cardiac arrest cases per week is 1*7 = 7 cardiac arrests per week. The Poisson distribution equation is:

$$P(\boldsymbol{X} = k) = f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

λ = 7 arrests / 1 week = 7 We are interested in the probability that no arrests occurred P(X = 0).

$$P(X=0) = \frac{e^{-7}7^0}{0!} = 9.119e^{-4}$$

b) The rate of cardiac arrest cases per 3 days is λ = 3 arrests / every 3 day period.

$$P(X=5) = \frac{e^{-3}3^5}{5!} = 0.101$$

15. Derive the maximum likelihood estimates for mean and standard deviation for a normal distribution N(μ, σ^2).
Solution:
https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood

16. A probability density function for a random variable is given by:

$$f(x) = \begin{cases} k(x^2+x) & \text{if } 0 \leq \text{x} \leq 1, \\ 0 & \text{else} \end{cases}$$

a) Find the value k that makes f(x) a probability density function.
b) Find the cumulative density function of f(x)?
c) Use the CDF to find Pr(0.25 $\leq$ x $\leq$ 0.75).
d) Find the mean, variance, and standard deviation of x.
Solution:
a) Find the value k that makes f(x) a probability density function. a valid PDF integrates to 1.

$$1 = \int_{-\infty}^{\infty} f(x)dx = k\int_0^1 (x^2+x)dx = k(\frac{x^3}{3} + \frac{x^2}{2})|_0^1$$

$$1 = k(\frac{5}{6})$$

$$k = \frac{6}{5}$$

b) Find the cumulative density function of f(x)?
The CDF for interval [0,1]

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_0^x f(t)dt$$

$$= \frac{6}{5}(\frac{x^3}{3} + \frac{x^2}{2})$$

c) Use the CDF to find Pr(0.25 $\leq$ x $\leq$ 0.75).
Pr(0.25 $\leq$ x $\leq$ 0.75) = Pr(X $\leq$ 0.75) - Pr(X $\leq$ 0.25) = F(0.75) - F(0.25)
d) Find the mean, variance, and standard deviation of x.
mean =

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$= \frac{6}{5}\int_0^1 x(x^2+x)dx$$

$$= \frac{6}{5}\int_0^1 x^3 + x^2 dx$$

$$= \frac{6}{5}(\frac{x^4}{4} + \frac{x^3}{3})|_0^1 = \frac{7}{10}$$

variance =

$$Var(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$$

$$= \int_{-\infty}^{\infty} x^2 f(x)dx - (\frac{7}{10})^2$$

$$= \frac{6}{5}\int_0^1 x^2(x^2 + x)dx - (\frac{7}{10})^2$$

solving it out yields:

$$= \frac{1}{20}$$

standard deviation is simply the squre root of variance

$$= \frac{1}{2\sqrt{5}}$$