Computational Statistics and Data Analysis (MVComp2)

Problem sets

1 Session 1

1.1 Round 1

1. Given two events A and B such that $P(A) = 0.5$, $P(B) = 0.3$, and $P(A \cap B) = 0.2$, find:
   - $P(A \cup B)$
   - $P(\overline{A} \cap \overline{B})$ where $\overline{A}$ and $\overline{B}$ are the complements of A and B, respectively.
2. In a class of 40 students, 25 like chocolate ice cream, 15 like vanilla ice cream, and 8 like both. How many students don't like either flavor?
3. A deck of cards contains 52 cards. How many ways can you draw 5 cards?
4. A pizza place offers 10 toppings. How many different three-topping pizzas can you order?
5. How many ways can 7 books be arranged on a shelf?

1.1.1 Answers:

1.
   - $P(A \cup B) = 0.5 + 0.3 - 0.2 = 0.6$
   - $P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B) = 0.4$
2. $25 + 15 - 8 = 32$ students like either flavor. $40 - 32 = 8$ students don't like either flavor.
3. $C(52, 5) = \frac{52!}{5! \times 47!} = 2,598,960$
4. $C(10, 3) = \frac{10!}{3! \times 7!} = 120$
5. $7! = 5,040$

1.2 Round 2

1. In a certain town, 60% of the families own a cat, 40% own a dog, and 25% own both a cat and a dog. If a randomly chosen family owns a dog, what's the probability that they also own a cat?
2. A bag contains 4 red balls, 5 blue balls, and 6 green balls. If a ball is drawn at random, it is found to be neither red nor green. What is the probability that it's blue?
3. There are two boxes. Box 1 contains 3 red and 7 blue marbles, while Box 2 contains 6 red and 4 blue marbles. If a box is chosen at random (with equal probability) and a marble is drawn, what is the probability that the marble is red?
4. In a factory, Machine A produces 60% of the items, and Machine B produces the remaining 40%. Past records show that 1% of the items produced by Machine A are defective, and 1.5% of the items produced by Machine B are defective. If an item is chosen at random and found to be defective, what's the probability that it was produced by Machine A?

1.2.1 Answers:

1. $P(\text{Cat} \mid \text{Dog}) = \frac{0.25}{0.4} = 0.625$
2. $P(\text{Blue}) = 1$
3. $P(\text{Red}) = 0.5 \times \frac{3}{10} + 0.5 \times \frac{6}{10} = 0.45$
4. Bayes theorem: $P(A \mid D) = \frac{P(D|A) \times P(A)}{P(D)}$. $P(D)$ can be found from the law of total probability: $P(D) = P(D \mid A) \times P(A) + P(D \mid B) \times P(B) = 0.012$. And so $P(A|D) = \frac{0.01 \times 0.60}{0.012} = 0.5$.

1.3 Round 3

1. A school library has 500 books, of which 120 are fiction and 75 are science fiction. If 50 books are both fiction and science fiction, how many books are neither fiction nor science fiction?
2. In a bouquet, there are 5 roses, 4 lilies, and 6 daisies. How many ways can you choose 3 flowers if at least one of them has to be a lily?
3. A password requires exactly 8 characters and can consist of 26 lowercase letters and 10 digits. How many possible passwords are there if it must start with a letter?
4. A class has 15 students. How many ways can you form a 4-student committee if two specific students can't be in the committee together?
5. In a drawer of socks, there are 8 blue socks, 5 red socks, and 7 green socks. If you pull out one sock and it's not green, what's the probability it's blue?

1.3.1 Answers:

1. 355 books are neither fiction nor science fiction.
2. $C(15, 3) - C(11, 3)$
3. 26×36^7
4. Both students excluded: $C(13, 4)$; One student included: $C(2, 1) \times C(13, 3)$. Total: $C(13, 4) + C(2, 1) \times C(13, 3)$.
5. $\frac{8}{13}$

2 Session 2

2.1 Round 1

Find the mean, variance, and standard deviation of the RV Y given in the following table

y	$p(y)$
0	1/8
1	1/4
2	3/8
3	1/4

2.1.1 Answers:

2.1.1.1 Mean

$$\mu = E[Y] = \sum_{y=0}^{3} yp(y) = 1.75$$

2.1.1.2 Variance

$$\sigma^2 = E[(Y-\mu)^2] = \sum_{y=0}^{3} (y-\mu)^2 p(y) = 0.9375$$

2.1.1.3 Standard deviation

$$\sigma = \sqrt{\sigma^2} = 0.97$$

2.2 Round 2

The manager of an industrial plant is planning to buy a new machine of either type A or type B. If t denotes the number of hours of daily operation, the number of daily repairs Y_1 required to maintain a machine of type A is a random variable with mean and variance both equal to $.10t$. The number of daily repairs Y_2 for a machine of type B is a random variable with mean and variance both equal to $.12t$. The daily cost of operating A is $C_A(t) = 10t + 30Y_1^2$; for B it is $C_B(t) = 8t + 30Y_2^2$. Assume that the repairs take negligible time and that each night the machines are tuned so that they operate essentially like new machines at the start of the next day. Which machine minimizes the expected daily cost if a workday consists of (a) 10 hours and (b) 20 hours?

2.2.1 Answers:

Expected daily cost for A is

$$\begin{aligned} E[C_A(t)] &= E[10t + 30Y_1^2] = 10t + 30E[Y_1^2] \\ &= 10t + 30\left(\mathrm{Var}[Y_1] + E[Y_1]^2\right) \\ &= 10t + 30\left(0.10t + (0.10t)^2\right) \\ &= 13t + 0.3t^2 \end{aligned}$$

Similarly

$$\begin{aligned} E[C_B(t)] &= E[8t + 30Y_2^2] = 8t + 30E[Y_2^2] \\ &= 8t + 30\left(\mathrm{Var}[Y_2] + E[Y_2]^2\right) \\ &= 8t + 30\left(0.12t + (0.12t)^2\right) \\ &= 11.6t + 0.432t^2 \end{aligned}$$

(a) $t = 10$, s.t. $E[C_A(10)] = 160$ and $E[C_B(10)] = 159.2$. Choose B.

(b) $t = 20$, s.t. $E[C_A(20)] = 380$ and $E[C_B(20)] = 404.8$. Choose A.

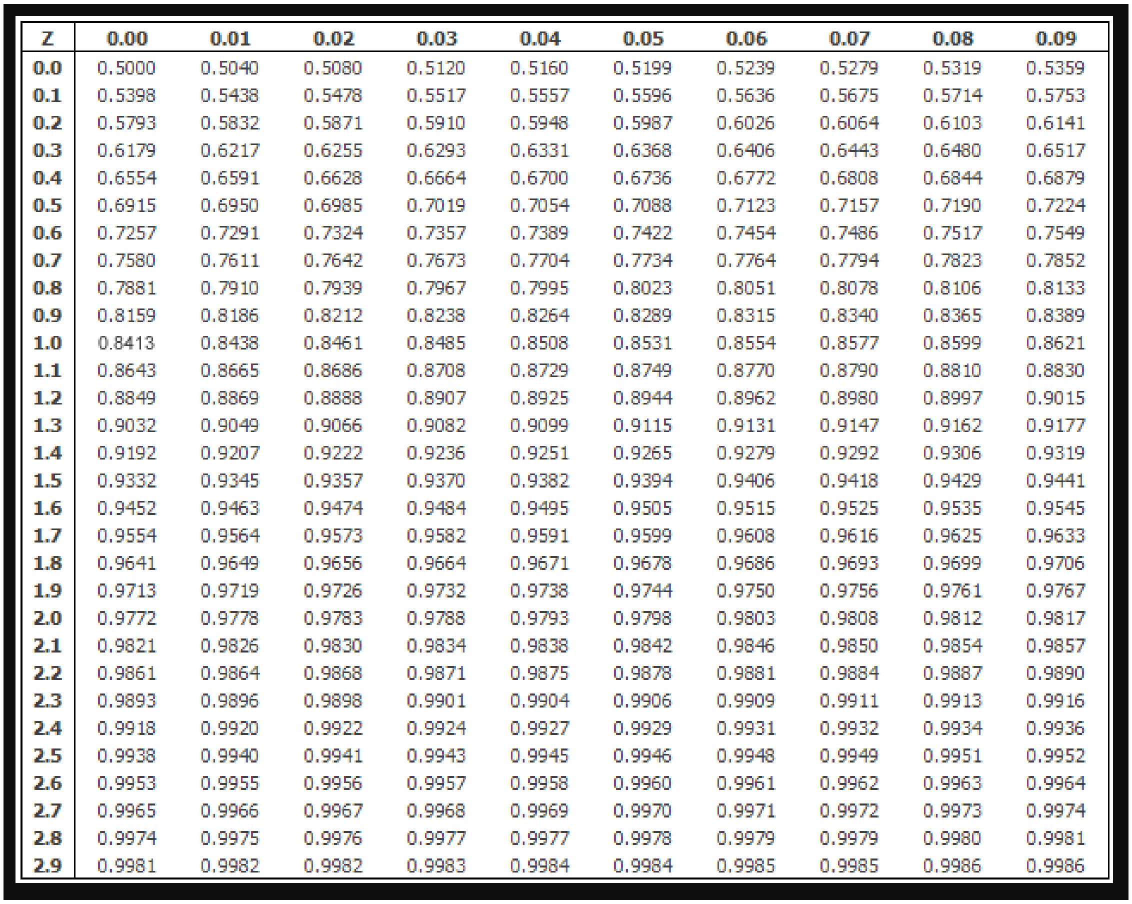
3 Session 3

3.1 Round 1

1. An experiment consists of tossing a fair die until a 6 occurs four times. What is the probability that the process ends after exactly ten tosses with a 6 occurring on the ninth and tenth tosses?

2. The number of people entering the intensive care unit at a hospital on any single day possesses a Poisson distribution with a mean equal to five persons per day. What is the probability that the number of people entering the intensive care unit on a particular day is equal to 2?

3. The length of time required to complete a college achievement test is found to be normally distributed with mean 70 minutes and standard deviation 12 minutes. When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test? Hint: You can use the following table for the cumulative distribution of the standard RV, $Z = \frac{X-\mu}{\sigma}$. The label for rows contains the integer part and the first decimal place of Z; The label for columns contains the second decimal place of Z; The values within the table are the cumulative probabilities.



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Figure 1: Standard normal table

3.1.1 Answers:

1. Event E: 10 tosses, including a 6 on the 9th, a 6 on the 10th, and two other 6 between 1st and 8th. Probability to land on a 6: 1/6. For the first 8 tosses, use the Binomial with $N = 8$ and $\pi = 1/6$.

$$P(X = E) = \binom{8}{2} \pi^2 (1 - \pi)^6 \cdot \pi \cdot \pi \approx 0.0072.$$

2. Poisson distribution with parameter $\lambda = 5$ (in units of persons per day). RV X: number of people entering the IC unit on any single day.

$$P(X = 2) = \frac{\lambda^2}{2!} \mathrm{e}^{-\lambda} \approx 0.084.$$

3. Find the upper bound of the probability interval, x_{max}, such that the cumulative distribution function yields $F(x_{\text{max}}) = 0.9$. First, convert X to a standard normal distribution, $Z = \frac{X-\mu}{\sigma}$. From the table we find $F(z = 1.28) = 0.8997$, which is close to the desired value 0.9. To convert from a standard normal to the original RV we use $x = \sigma z + \mu$. We obtain $x = 85.36$ minutes, which we round up to 86 minutes.

4 Session 4

5 Round 1

Suppose Y is a random variable representing a coin toss, where the event $Y = 1$ corresponds to heads and $Y = 0$ corresponds to tails. Let $\theta = p(Y = 1)$ be the probability of heads. You toss the coin N times, and observe N_1 heads and N_0 tails.

1. Use maximum likelihood estimation to infer $\hat{\theta}$.
2. What is the MLE if you've only observed a single toss, which landed on head?

5.0.1 Answers:

1. The probability distribution of the RV is the Binomial distribution. Let $N = N_1 + N_0$. Its log-likelihood is given by

$$\begin{aligned} l(\theta) &= \ln \mathcal{L}(\theta) \\ &= \ln \left[\binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_0} \right] \\ &= \ln \binom{N}{N_1} + N_1 \ln \theta + N_0 \ln(1-\theta). \end{aligned}$$

   Take the derivative of l wrt θ and set it to zero

$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0.$$

   Solve for θ to get

$$\begin{aligned} N_1(1-\hat{\theta}) &= N_0\hat{\theta} \\ (N_0 + N_1)\hat{\theta} &= N_1 \\ \hat{\theta} &= \frac{N_1}{N_0 + N_1}. \end{aligned}$$

2. $\hat{\theta} = 1$.

6 Round 2

Same setting (Suppose Y is a random variable representing a coin toss, where the event $Y = 1$ corresponds to heads and $Y = 0$ corresponds to tails. Let $\theta = p(Y = 1)$ be the probability of heads. You toss the coin N times, and observe N_1 heads and N_0 tails.)

3. To mitigate the issue of question 2, revisit question 1 by incorporating a prior. The prior will follow the beta distribution, $p(\theta) = \mathrm{Beta}(\theta|a, b)$, where $a, b > 1$ encourages values of θ near $\frac{a}{a+b}$. Use maximum-a-posteriori estimation to infer $\widehat{\theta}$.
4. Set $a = b = 2$, interpret the estimator that you obtain.

6.0.1 Answers:

3. The Beta distribution is given by $p(\theta) = \text{Beta}(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$. For the likelihood we use a Bernoulli distribution with N_1 heads and N_0 tails. This leads to the posterior

$$\begin{aligned}\ln p(\theta|\mathcal{D}) &\propto \ln p(\mathcal{D}|\theta) + \ln p(\theta)\\ &= N_1 \ln\theta + N_0 \ln(1-\theta) + (a-1)\ln\theta + (b-1)\ln(1-\theta)\end{aligned}$$

Set the derivative wrt θ to zero to yield

$$\hat{\theta} = \frac{N_1 + a - 1}{N_1 + N_0 + a + b - 2}.$$

4. With $a = b = 2$ we get

$$\hat{\theta} = \frac{N_1 + 1}{N_1 + N_0 + 2}$$

which is called *add-one smoothing*, and is used to avoid the zero-count problem. The zero-count problem is linked to overfitting and the black swan paradox.