Assignment 2 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. We use discrete distributions when

   (a) the outcomes fall in a fixed range

   (b) the number of possible outcomes is countable

   (c) the probability values corresponding to the outcomes are discrete

   (d) the number of iterations of the experiment are fixed

   Sol. (b)
   Option (a) is incorrect since even if all the outcomes fall in a fixed range they may be continuous in nature, i.e., they may take any value in the specified interval.

2. Based on a survey, it was found that the probability that a student likes to play football was 0.25 and the probability that a student likes to play cricket is 0.43. It was also found that the probability that a student likes to play both football and cricket is 0.12. What is the probability that a student does not like to play either?

   (a) 0.32

   (b) 0.2

   (c) 0.44

   (d) 0.56

   Sol. (c)
   Given P(football) = 0.25, P(cricket) = 0.43, and P(football $\cap$ cricket) = 0.12.

   We are interested in the probability of students who do not like to play either football or cricket, i.e., P((football $\cup$ cricket)$'$).

   From basic set theory, we have P(football $\cup$ cricket) = P(football) + P(cricket) - P(football $\cap$ cricket) = 0.25 + 0.43 - 0.12 = 0.56.

   Also, the two events, a student likes to play football or cricket (football $\cup$ cricket) and a student does not like to play either football or cricket ((football $\cup$ cricket)$'$) are mutually exclusive. Therefore, we have P((football $\cup$ cricket)$'$) = 1 - P(football $\cup$ cricket) = 1 - 0.56 = 0.44.

3. Can the value of a probability density function be greater than one? What about the cumulative distribution function?

   (a) PDF: yes, CDF: yes

(b) PDF: yes, CDF: no

(c) PDF: no, CDF: yes

(d) PDF: no, CDF: no

Sol. (b)
The maximum value of any CDF is one. Thus, the CDF cannot take on values greater than one. For PDFs, consider the uniform distribution in the range 0 to 0.5. It should be clear that the PDF for this uniform distribution is $f(x) = 2$ for $0 \leq x \leq 0.5$ and $f(x) = 0$ elsewhere. Clearly, the PDF can take on values greater than one.

4. We are told that for a particular coin, the probability of observing k heads in n tosses is x ($0 \leq k \leq n$, $0 < x < 1$) and the probability of observing k heads in $n+1$ tosses is y ($0 < y < 1$). What is the probability, p, of the coin showing up heads in a single toss?

(a) $\frac{y\binom{n}{k}}{x\binom{n+1}{k}}$

(b) $\frac{x\binom{n+1}{k}}{y\binom{n}{k}}$

(c) $\frac{y\binom{n}{k}}{x\binom{n+1}{k}} - 1$

(d) $\frac{x\binom{n+1}{k} - y\binom{n}{k}}{x\binom{n+1}{k}}$

Sol. (d)
Given

$$\binom{n+1}{k} p^k (1-p)^{n-k+1} = y \text{ and } \binom{n}{k} p^k (1-p)^{n-k} = x$$

We have

$$p^k (1-p)^{n-k+1} = \frac{y}{\binom{n+1}{k}} \text{ and } p^k (1-p)^{n-k} = \frac{x}{\binom{n}{k}}$$

Dividing, we have

$$(1-p) = \frac{y\binom{n}{k}}{x\binom{n+1}{k}}$$

$$p = 1 - \frac{y\binom{n}{k}}{x\binom{n+1}{k}}$$

$$p = \frac{x\binom{n+1}{k} - y\binom{n}{k}}{x\binom{n+1}{k}}$$

5. The probability that you can hit the bullseye of a dart board in a single throw is 0.6. Given 10 throws, what is the probability of you hitting the bullseye at most 4 times? Also, what is the probability of you hitting the bullseye for the first time on your fourth attempt?

(a) 0.166, 0.038

(b) 0.054, 0.038

(c) 0.166, 0.064

(d) 0.054, 0.064

Sol. (a)
Let $P_{\leq 4}$ be the probability of hitting the bullseye at most 4 times and P_{f4} be the probability of hitting the bullseye for the first time on the fourth attempt.

To calculate $P_{\leq 4}$, we are given $n = 10$, $k = 4$, and $p = 0.6$. We have

$$P_{\leq 4} = \sum_{i=0}^{i=4} \binom{10}{i} 0.6^i 0.4^{10-i}$$

Expanding, we have

$$P_{\leq 4} = \binom{10}{0}0.6^0 0.4^{10-0} + \binom{10}{1}0.6^1 0.4^{10-1} + \binom{10}{2}0.6^2 0.4^{10-2} + \binom{10}{3}0.6^3 0.4^{10-3} + \binom{10}{4}0.6^4 0.4^{10-4}$$

$$P_{\leq 4} = 0.4^{10} + 10 * 0.6 * 0.4^9 + 45 * 0.6^2 * 0.4^8 + 120 * 0.6^3 * 0.4^7 + 210 * 0.6^4 * 0.4^6$$

$$P_{\leq 4} = 0.0001048576 + 0.001572864 + 0.010616832 + 0.042467328 + 0.111476736$$

$$P_{\leq 4} = 0.166$$

To calculate the probability of hitting the bullseye the first time on the fourth attempt, we have

$$P_{f4} = (1-p)^{4-1} p = 0.4^3 * 0.6 = 0.0384$$

6. It is known that 45% of the population support a particular candidate for election. What is the probability that in a poll, more than half of the 600 randomly sampled people express support for that candidate? (Hint: Use the normal approximation to the binomial distribution. In calculating the probability value, first calculate the z-score (https://en.wikipedia.org/wiki/Standard_score) and then use the following z-table: http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf).

   (a) 0.0082
   (b) 0.0069
   (c) 0.9931
   (d) 0.9918

   Sol. (b)
   The number of people supporting the candidate (k) is distributed according to the binomial distribution with $n = 600$ and $p = 0.45$. Since n is large, we apply the normal approximation to the binomial distribution.

   Mean, $\mu = np = 270$.

   Standard deviation, $\sigma = \sqrt{np(1-p)} = 12.186$.

   We are interested in calculating the probability that more than half of the sampled people support the candidate, i.e., P($k > 300$) or 1 - P($k \leq 300$), where k is the number of 'successes'.

   Standardising to get the z-value, we have

$$z = \frac{k - \mu}{\sigma} = \frac{300 - 270}{12.186} = 2.46$$

The probability that more than half of the sampled people support the candidate

$$P(k > 300) = 1 - P(k \leq 300) = 1 - p(z \leq 2.46)$$

From the provided z-table, we find the corresponding probability value as 0.9931. Thus, the required probability is 1 - 0.9931 - 0.0069.

7. It is known that the weight of 1000 g rice packets produced by a certain company follows the normal distribution $\mathcal{N}(1000, 25)$. Given a sample of size 100, what is the distribution of the sample mean?

   (a) $\mathcal{N}(1000, 2.5)$
   (b) $\mathcal{N}(1000, 0.5)$
   (c) $\mathcal{N}(1000, 0.25)$
   (d) $\mathcal{N}(1000, 6.25)$

   Sol. (c)
   Original distribution: $\mathcal{N}(1000, 25)$
   We know that the sampling distribution of the sample mean with sample size of n follows the normal distribution

$$\mathcal{N}\left(\mu, (\frac{\sigma}{\sqrt{n}})^2\right) \text{ or } \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

   Thus, the distribution of the sample mean for the given distribution is

$$\mathcal{N}\left(1000, \frac{25}{100}\right) = \mathcal{N}(1000, 0.25)$$

8. Past experience shows that in a particular class, average scores of students in a particular subject is 69%. After adopting new methods of teaching relying more on multimedia content, we want to test whether the new methods help improve the average scores. Which among the following pair of null and alternate hypotheses do you consider suitable in this scenario?

   (a) null: $\mu = 69$; alternate: $\mu > 69$
   (b) null: $\mu \leq 69$; alternate: $\mu > 69$
   (c) null: $\mu > 69$; alternate: $\mu \leq 69$
   (d) null: $\mu = 69$; alternate: $\mu \neq 69$

   Sol. (b)
   The first option does not cover the entire range of values as required. The third option is incorrect because the default case is the one where the class average is 69%. The last option is incorrect because we are interested to know whether the new teaching methods increase the average score. In this test, on the basis of rejecting the null hypothesis, it would not be clear whether the class average improved or not.

9. A soft drinks company wants to introduce a new product. In setting up the manufacturing process it needs to verify that the average amount of sugar in each 500 ml bottle is less than or equal to 55 g. It randomly samples 10 bottles and observes the following values of sugar in the samples: 53.8, 56.1, 54.5, 54.8, 55.2, 55.1, 54.7, 55.8, 55.9, 54.5. Perform the z-test on this data and report the z_{stat} value and p value. Assume known standard deviation of 0.3 g. (Hint: Use the following z-table: http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf).

(a) z_{stat} : 0.42; $p - value$: 0.6628

(b) z_{stat} : 0.23; $p - value$: 0.5910

(c) z_{stat} : 0.23; $p - value$: 0.4090

(d) z_{stat} : 0.42; $p - value$: 0.3372

Sol. (d)
Null hypothesis: $\mu_0 \leq 55$

Alternate hypothesis: $\mu_0 > 55$

$$\bar{x} = \frac{53.8 + 56.1 + 54.5 + 54.8 + 55.2 + 55.1 + 54.7 + 55.8 + 55.9 + 54.5}{10} = 55.04$$

$$z_{stat} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{55.04 - 55}{\frac{0.3}{\sqrt{10}}} = 0.42$$

Using the z-table, we observe the probability value corresponding to $z_{stat} = 0.42$ as 0.6628. However, we need the value corresponding to $z_{stat} > 0.42$. Thus, the p-value = 1 - 0.6628 = 0.3372.

10. The above soft drinks company wants to ensure that the variance in the amount of sugar in each 500 ml bottle of its new product is within an acceptable limit of 0.3 g. It takes a random sample of 65 bottles from its manufacturing unit and measures the amount of sugar in each bottle. Which hypothesis test should the company use to verify that the variance in the amount of sugar is less than 0.3 g?

    (a) z-test

    (b) t-test

    (c) chi-square test

    (d) proportion z-test

    Sol. (c)
    Among the above four mentioned tests, the chi-square test is used for testing the variance of a sample.