Statistical Foundations of Data Science (COMS4055A, COMS7058A)
Class Test 2

24 April 2024, 14h15–16h15, JD du Plessis

Name: ____________________ Row: ______ Seat: ______ Signature: __________

Student Number: ______________ ID Number: ______________

Instructions

- Answer all questions in pen. **Do not write in pencil.**
- This test consists of 3 pages. Ensure that you are not missing any pages.
- A formula sheet is provided
- You are allocated 2 hour to complete this test.
- Ensure your cellphone is switched off.
- You may use a calculator during the test.
- Round off to an appropriate number of decimal places and simplify your answers fully.
- Please manage your time appropriately. Some questions are difficult but worth few marks, do not linger on a difficult question if you have not answered others.

Question 1 Probability [35 Marks]

1. A certain disease affects 1 in every n people. There is a test for the disease with the following properties:

   - If a person has the disease, the test is positive with probability p_t of the time (true positive rate).
   - If a person does *not* have the disease, the test is falsely positive p_f of the time (false positive rate).

   A randomly selected person takes the test and receives a **positive** result. What is the probability that this person actually has the disease? If

   (a) $n = 1000$, $p_t = 0.98$ and $p_f = 0.05$. That is one in 1000 people has the disease 98% of infected people test positive and only 5% of uninfected people test positive. [3]

   (b) $n = 2000$, $p_t = 0.99$ and $p_f = 0.02$. That is one in 2000 people has the disease 99% of infected people test positive and only 2% of uninfected people test positive. [3]

   (c) Construct a general formula in terms of n, p_t and p_f.

   [4]

   (a) $n = 1000$, $p_t = 0.98$ and $p_f = 0.05$. That is one in 1000 people has the disease 98% of infected people test positive and only 5% of uninfected people test positive. [3]

   (b) $n = 2000$, $p_t = 0.99$ and $p_f = 0.02$. That is one in 2000 people has the disease 99% of infected people test positive and only 2% of uninfected people test positive. [3]

   (c) Construct a general formula in terms of n, p_t and p_f.[4]

$$\frac{0.001 \times 0.98}{0.001 \times 0.98 + 0.999 \times 0.05} = 0.01924209699$$

$$\frac{0.0005 \times 0.99}{0.0005 \times 0.99 + 0.9995 \times 0.02} = 0.02416402245$$

$$\frac{\frac{1}{n}p_t}{\frac{1}{n}p_t + \frac{n-1}{n}p_f} = \frac{p_t}{p_t + (n-1)p_f}$$

2. In a group of 4 randomly selected people, assume that each person is equally likely to have any of the 12 zodiac star signs, and that star signs are independent between people.

   (a) What is the probability that all 4 people have different star signs? [4]

   (b) What is the probability that at least two people share the same star sign? [2]

   (a)

$$\frac{12}{12} \cdot \frac{11}{12} \cdot \frac{10}{12} \cdot \frac{9}{12} = 0.57291666666$$

(b)

$$1 - 0.57291666666 = 0.427083333333$$

3. Let $f(x)$ be a probability density function (PDF) defined as follows:

$$f(x) = \begin{cases} c \cdot x^2 & \text{for} \quad 0 \leq x \leq 2, \\ c \cdot (4 - x^2) & \text{for} \quad 2 < x \leq 4, \\ 0 & \text{otherwise.} \end{cases}$$

where c is a constant.

(a) Find the value of c that makes $f(x)$ a valid probability density function. [3]

(b) Once you have the value of c, calculate the mean μ of the distribution. [3]

(c) Calculate the variance σ^2 of the distribution. [4]

Error in question, this "distribution" goes negative. However if you apply the standard formulas:

(a)

$$\int_0^2 cx^2 dx + \int_2^4 c4 - cx^2 dx = 1$$
$$c = \frac{-1}{8}$$

(b)

$$\mathbb{E}[X] = \int_0^2 \frac{-1}{8} x^3 dx + \int_2^4 \frac{-1}{8} 4x - \frac{-1}{8} x^3 dx$$
$$\mathbb{E}[X] = 4$$

(c)

$$V(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{-4}{3}$$

4. (a) The weights of apples in a large orchard are approximately normally distributed with a mean of 150 grams and a standard deviation of 20 grams. What proportion of apples weigh more than 150 grams? [2]

(b) A student scores 72 on a test where the class mean is 65 and the standard deviation is 5.
Find the student's z-score and intepret it. [2]

(c) The heights of adult men in a population are normally distributed with a mean of 175 cm and a standard deviation of 8 cm. What is the probability that a randomly selected man is taller than 183 cm? [2]

(d) The marks on a certain test are normally distributed. It is known that:

- 20% of students scored below 40 marks
- 10% of students scored above 70 marks

Find the mean score on the test. [3]

(a) The weights of apples in a large orchard are approximately normally distributed with a mean of 150 grams and a standard deviation of 20 grams. What proportion of apples weigh more than 150 grams? [2]
0.5, normals are symmetric about their means.

(b) A student scores 72 on a test where the class mean is 65 and the standard deviation is 5.
Find the student's z-score and intepret it. [2]
$Z = 1.4$, 1.4 standard deviations above the mean, better than 91.924 percet of students.

(c) The heights of adult men in a population are normally distributed with a mean of 175 cm and a standard deviation of 8 cm. What is the probability that a randomly selected man is taller than 183 cm? [2]
Z score of 1. Equates to 0.15866 probability.

(d) The marks on a certain test are normally distributed. It is known that:

- 20% of students scored below 40 marks
- 10% of students scored above 70 marks

Find the mean score on the test. [3]

$$40 = \mu - 0.84\sigma$$
$$70 = \mu + 1.28\sigma$$

$\mu \approx 52$

Question 2 Combinatorics [6 Marks]

5. How many integers between 1 and 1000 inclusive are divisible by

   (a) At least one of 2,3 and 5. [3]

   (b) Exactly two of 2,3 and 5. [3]

   (a) At least one of 2,3 and 5.[3] $500 + 333 + 200 - 166 - 100 - 66 + 33 = 734$

   (b) Exactly two of 2,3 and 5.[3] $166 + 100 + 66 - 2 * 33 = 266$

Question 3 Inference [29 Marks]

6. A manufacturer claims that 90% of their light bulbs last at least 1,000 hours. A quality control engineer decides to test this claim using two different sample sizes.

   (a) In a small-scale test, the engineer tests a random sample of 10 light bulbs and finds that 7 of them last at least 1,000 hours.

      i. State the null and alternative hypotheses. [2]

      ii. Assuming the manufacturer's claim is correct, let X be the number of bulbs in the sample that last at least 1,000 hours. What is the distribution of X under the null hypothesis? [3]

      iii. Compute $P(X \leq 7)$ under the null hypothesis. [2]

      iv. At the 5% significance level, should the engineer reject the manufacturer's claim? Justify your answer. [2]

   (b) In a larger test, the engineer tests 1,000 light bulbs and finds that 850 of them last at least 1,000 hours.

      i. Again, state the null and alternative hypotheses. Does it change? [2]

      ii. Let $\hat{p}$ be the observed proportion of bulbs that last at least 1,000 hours. Use the normal approximation to the binomial distribution to test the hypothesis. [4]

      iii. At the 5% significance level, what conclusion should the engineer draw? [2]

(a) i.

   ii. $H_0 : P = 0.9$, $H_1 : p < 0.9$ $\alpha = 0.05$ (technically you use a different α

   iii. $Binomial(10, 0.9)$ [3]

   iv.

   $$P(X \leq 7) = 1 - P(X = 8) - P(X = 9) - P(X - 10) = 0.0701908264$$

   v. Fail to reject at 5 percent.

(b) i. $H_0 : P = 0.9$, $H_1 : p < 0.9$ $\alpha = 0.05$ (technically you use a different α. Doesn't change.

   ii. X is normal with mean 900 and variance $100 * 0.9 * 0.1 = 90$, standard deviation is $\sqrt{90} = 9.48683298051$. This gives a Z-score of $\frac{850-900}{\sqrt{90}} = -5.27046276695$. P-value is basically 0

   iii. We reject the null hypothesis.

7. A local call center receives an average of 10 calls per hour. Assume that the number of calls follows a Poisson distribution.

   (a) What is the probability that the call center receives exactly 7 calls in an hour? [2]

   (b) What is the probability that the call center receives fewer than 5 calls in an hour? [2]

(c) What is the probability that the call center receives more than 3 calls in a half hour period? [2]

(a) What is the probability that the call center receives exactly 7 calls in an hour? [2]

$$\frac{10^7 e^{-10}}{7!} = 0.09007922571$$

(b) What is the probability that the call center receives fewer than 5 calls in an hour? [2]

$$\frac{10^0 e^{-10}}{0!} + \frac{10^1 e^{-10}}{1!} + \frac{10^2 e^{-10}}{2!} + \frac{10^3 e^{-10}}{3!} + \frac{10^4 e^{-10}}{4!} = 0.02925268807$$

(c) What is the probability that the call center receives more than 3 calls in a half hour period? [2]

$$1 - \frac{5^0 e^{-5}}{0!} - \frac{5^1 e^{-5}}{1!} - \frac{5^2 e^{-5}}{2!} - \frac{5^3 e^{-5}}{3!} = 0.55950671493$$

8. Prove that

(a) $\mathbb{E}[(X-\mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[X]^2$ [3]

(b) $\mathbb{E}[(X-\mu)^3] = \mathbb{E}[x^3] - 3\mu\mathbb{E}[X]^2 + 2\mathbb{E}[X]^3$ [3]

(a) $\mathbb{E}[(X-\mu)^2] = \mathbb{E}[x^2] - \mathbb{E}[X]^2$[3]

$$\begin{aligned}
\mathbb{E}[(X-\mu)^2] &= \mathbb{E}[(X^2 - 2\mu X + \mu^2)] \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}[x^2] - \mathbb{E}[X]^2
\end{aligned}$$

(b) $\mathbb{E}[(X-\mu)^3] = \mathbb{E}[x^3] - 3\mu\mathbb{E}[X]^2 + 2\mathbb{E}[X]^3$ [3]

$$\begin{aligned}
\mathbb{E}[(X-\mu)^3] &= \mathbb{E}[(X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3)] \\
&= \mathbb{E}[X^3] - 3\mu\mathbb{E}[X^2] + 3\mathbb{E}[X]^2\mathbb{E}[X] - \mu^3)] \\
&= \mathbb{E}[X^3] - 3\mu\mathbb{E}[X^2] + 3\mathbb{E}[X]^2\mu - \mathbb{E}[X]^3)] \\
&= \mathbb{E}[x^3] - 3\mu\mathbb{E}[X]^2 + 2\mathbb{E}[X]^3
\end{aligned}$$