Quiz - 4

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Course: Biostatistics | Instructor: Dr. Gaurav

Duration: 60 minutes | Each question weights 5 Marks

Answer in one or two sentences only. Negative marks for lengthy explanations (>2 sentences)

1. A machine learning model has an AUC ROC of 0.8. Can you definitively say the model is good at classifying positive and negative cases?
   Answer: An AUC ROC of 0.8 suggests the model has good discriminatory power, but the assessment of whether it's "good" depends on the context and the specific problem being addressed. Generally, an AUC ROC above 0.7 is considered acceptable for many applications.
2. A new blood test claims to be 95% accurate at detecting a rare disease (meaning there's a 5% chance of a false positive). The disease itself only affects 1% of the population. If someone tests positive on this blood test, what is the chance they actually have the disease?

   1. **Given values:**
      - Probability of having the disease (A): $P(A) = 0.01$
      - Probability of a false positive (B given not A): $P(B|\neg A) = 0.05$
      - Complement of the disease probability (not having the disease): $P(\neg A) = 1 - P(A) = 0.99$
   2. **Calculate total probability of testing positive (B):**
      - $P(B) = P(B|A) \times P(A) + P(B|\neg A) \times P(\neg A)$
      - $P(B) = 1 \times 0.01 + 0.05 \times 0.99 = 0.01 + 0.0495 = 0.0595$
   3. **Use Bayes' theorem to find the probability of having the disease given a positive test result (A given B):**
      - $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$
      - $P(A|B) = \frac{1 \times 0.01}{0.0595}$
      - $P(A|B) \approx \frac{0.01}{0.0595} \approx 0.168 \approx 16.8\%$

   So, there's roughly a 16.8% chance that someone who tests positive actually has the disease.

3. In a study following cancer patients, you observe a patient who is still alive after 5 years. They are then lost to follow-up due to moving away. Can you definitively say this patient survived for more than 5 years?

Answer : No, we can't definitively say the patient survived for more than 5 years. The patient's status beyond the last observed time point (5 years) is uncertain due to being lost to follow-up.

4. Two groups (A and B) are followed in a survival study. Group A has a higher median survival time on the Kaplan-Meier curve compared to Group B. However, the hazard ratio for group B is less than 1 (p-value < 0.05) in a Cox proportional hazards model. How can you explain this seemingly contradictory finding?
   Answer : The higher median survival time in Group A on the Kaplan-Meier curve suggests longer survival up to a certain point. However, the hazard ratio less than 1 for Group B in the Cox model indicates a lower risk of the event (e.g., death) compared to Group A over the entire follow-up period, considering other factors. These findings highlight different aspects of survival analysis: median survival time focuses on specific time points, while the hazard ratio accounts for changes in risk over time.
5. You are analyzing a dataset on blood pressure measurements. The data appears highly skewed towards lower values. You decide to log-transform the data to achieve normality. However, after transformation, you observe a few negative values. How can you explain this, and does it invalidate the log transformation?
   Answer : The appearance of negative values after log-transformation often indicates that the original data contained zero or negative values. Logarithm is undefined for zero and negative numbers, resulting in negative values after transformation. This occurrence highlights the need to handle zero or negative values appropriately before log-transformation, such as adding a constant or using alternative transformations like the Box-Cox method. While the presence of negative values post-transformation complicates interpretation, it doesn't necessarily invalidate the log transformation, but it does warrant careful consideration and potentially alternative approaches.
6. A researcher conducts a hypothesis test to compare the mean weight loss of two different diet programs. They obtain a statistically significant p-value ($p < 0.05$), rejecting the null hypothesis of equal mean weight loss. However, the observed

Quiz - 4

INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Course: Biostatistics | Instructor: Dr. Gaurav

Duration: 60 minutes | Each question weights 5 Marks

difference in mean weight loss between the two groups is very small. Can you definitively conclude that one diet program is superior to the other?

Answer : No, obtaining a statistically significant p-value does not necessarily mean the observed difference is practically significant. A small observed difference in mean weight loss suggests limited practical significance, and therefore, definitive conclusions about the superiority of one diet program over the other cannot be made based solely on statistical significance. Other factors such as clinical relevance, effect size, and practical significance should also be considered.

7. You are comparing the effectiveness of two pain medications (A and B) for post-surgical pain relief. You plan to use the Mann-Whitney U test because the pain scores are not normally distributed. However, the sample sizes for the two groups are quite different (Group A: n = 20, Group B: n = 8). Is the Mann-Whitney U test still a suitable choice in this scenario?

   Answer : Yes, the Mann-Whitney U test can still be a suitable choice even with unequal sample sizes, as it is robust to differences in sample size. However, it's important to note that larger sample sizes generally provide more precise estimates, so the interpretation of the results should consider the imbalance in sample sizes between the two groups.

8. A study finds a prevalence of 10% for a chronic disease in a population of 10,000 people. One year later, a follow-up study identified 1,000 new cases of the disease. Can you calculate the incidence rate of the disease for that year?

Given:

- Total population: 10,000 people
- Prevalence of the disease: 10% (0.10)
- Number of new cases identified during the year: 1,000

First, let's find out how many people had the disease at the beginning of the year:

Number of existing cases = Prevalence * Total population
Number of existing cases = 0.10 * 10,000
Number of existing cases = 1,000

Now, we have 1,000 existing cases at the beginning of the year.

The incidence rate formula is:
$$\text{Incidence Rate} = \frac{\text{Number of new cases}}{\text{Population at risk}}$$

Population at risk = Total population - Number of existing cases
Population at risk = 10,000 - 1,000
Population at risk = 9,000

Now, plug in the values:
$$\text{Incidence Rate} = \frac{1{,}000}{9{,}000}$$
$$\text{Incidence Rate} = \frac{1}{9}$$

Let's calculate:
$$\text{Incidence Rate} = 0.1111$$

9. Two researchers are analyzing the body weight of adults in a large population. Researcher A finds the data is normally distributed, with a mean of 70 kg and a standard deviation of 10 kg. Researcher B, analyzing the same data, reports a median of 70 kg and a quartile range (IQR) of 20 kg. Are their findings necessarily contradictory?

   Answer : No, their findings are not necessarily contradictory. Researcher A's report of a normal distribution with a mean and standard deviation and Researcher B's report of a median and interquartile range (IQR) describe different aspects of the data's distribution. It's possible for the data to have a normal distribution (as described by Researcher A) while also having a median and IQR as reported by Researcher B.

Quiz - 4

Course: Biostatistics | Instructor: Dr. Gaurav

Duration: 60 minutes | Each question weights 5 Marks

10. You are sampling heart rates from a population known to have a slightly skewed distribution towards higher values. Because of the central limit theorem, you can be confident that the sample mean, regardless of sample size, will be normally distributed. Do you agree with this statement?

    Answer : No, the central limit theorem does not guarantee that the sample mean will always be normally distributed regardless of the population's distribution. It ensures that the distribution of the sample means approaches a normal distribution as the sample size increases, provided certain conditions are met, such as independent and identically distributed samples. However, the distribution of the original population remains a factor in determining whether the sample mean will be normally distributed.