Enabling People
Education for Future Generations

AI Course

Chapter 4. Quiz

For students

Samsung Innovation Campus

1. What is not correct as an interpretation of the fit model in the following statistical analysis?

   ① In statistical analysis, the one-sided test indicates that the null hypothesis for the parameter is $H_0 : \theta = \theta_0$ and the alternative hypothesis can be defined as $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$.
   ② Type I error is the level of significance and it is an error that occurs when the true null hypothesis is rejected although the null hypothesis is true.
   ③ If the significance probability p is greater than the significance level α ($p > \alpha$), the null hypothesis is rejected and the alternative hypothesis is adopted.
   ④ In order to test the significance of the regression equation, a significance test should be performed on the slope of the regression equation, so the hypothesis is $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$ (where β_1 is the slope of the regression equation).

   Answer:-

   Option 3 is incorrect. When the significance probability p is greater than the significance level α ($p > \alpha$), we FAIL TO REJECT the null hypothesis, not reject it. The correct statement should be: when $p > \alpha$, we fail to reject the null hypothesis.

2. Assuming I get married and have three children, assuming that the random variable X is the number of sons, what is the probability, mean, and variance of having two sons in this case? (In this case, the probability of having a son and daughter is assumed to be the same.)

   ANSWER: Probability = 3/8 = 0.375, Mean = 1.5, Variance = 0.75
   Using binomial distribution with n=3, p=0.5: $P(X=2) = C(3,2) \times (0.5)^2 \times (0.5)^1 = 3/8$, $E(X) = np = 1.5$, $Var(X) = np(1-p) = 0.75$

3. When a random variable X follows $X \sim N(\mu, \sigma^2)$, which of the following is correct as a standardized expression?

   ① $1/x$
   ② $(x - \mu) / \sigma$
   ③ $(x - \sigma) / \mu$
   ④ $(x - \mu)$

   ANSWER: 2
   The standardized expression for a normal random variable $X \sim N(\mu, \sigma^2)$ is $(x - \mu)/\sigma$, which is the Z-score formula.

4. If a sample of 100 people was selected to find out whether satisfaction with two subjects differed according to gender, which statistic would be the most appropriate in this case?

① t

② χ^2

③ Z

④ F

ANSWER: 2 (χ^2)
Chi-square test is most appropriate for testing the relationship between two categorical variables: gender and satisfaction with two subjects.

Among the students taking the class of Information and Statistics, 30 students were randomly selected. When investigating the average of their grades, the mean was 80 and the variance was 9. Find a 95% confidence interval for the average of their grades.

ANSWER: (78.88, 81.12)
Using t-distribution with n=30, mean=80, variance=9 (σ=3): 95% CI = 80 ± $t_{0.025,29}$ × (3/√30) = 80 ± 2.045 × 0.548 ≈ 80 ± 1.12

5. Find the feature that is the most correlated with the target using the code below.

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_breast_cancer


# Load data.
data = load_breast_cancer()
# make DataFrame
df = pd.DataFrame(data.data, columns=[data.feature_names])
df = df.assign(target=pd.Series(1 - data.target))
```

Answer:-

Run the following code to find the most correlated feature:

df.corr()['target'].abs().sort_values(ascending=False)

For the breast cancer dataset, 'worst concave points' or 'worst perimeter' typically shows the strongest correlation with the target