

Data Mining - EX5

Deadline: Friday, Azar 02, 1403 - November 22, 2024

Question 1:

The delivery time of goods to customers by an online store is normally distributed, with a mean of 15 days and a standard deviation of 3 days. For the $n = 30$ sample sizes, construct a 95% confidence interval for the population mean delivery time of goods to customers.

Question 2:

A pharmaceutical company is testing the effectiveness of a new drug for reducing blood pressure. The company designs a clinical trial in which the blood pressure of patients is measured before and after taking the drug. After conducting the experiment on 100 patients, the data is analyzed.

The results show that, on average, the patients' blood pressure decreased by 5 mmHg. However, the statistical analysis reveals that the p-value is 0.03, and the margin of error for the average reduction in blood pressure is ± 2 mmHg.

Now, answer the following questions:

1. Given the p-value obtained, can we conclude that the new drug significantly reduces blood pressure? Why?
2. What does the margin of error represent in this experiment, and what information does it provide about the average reduction in blood pressure?
3. If we increase the number of patients in the trial, what do you think will happen to the p-value and margin of error? Why?

Question 3:

In a survey, 1200 users of an online platform were asked if they would consider using its premium version. Out of them, 180 responded positively. For 95% confidence level, construct a confidence interval for the population proportion who might be interested in the premium version.

Question 4:

In a sample of 150 users of an app, 30 deactivated their accounts following recent changes in privacy policies. Test whether the population proportion of users who deactivate their accounts is less than 20%, using a level of significance $\alpha = 0.05$.

Question 5:

Data set	Sample Mean	Sample Standard Deviation	Sample Size
Experimental	$\bar{x}_1 = 35.2$	$s_1 = 6.1$	$n_1 = 1500$
Control	$\bar{x}_2 = 34.8$	$s_2 = 5.9$	$n_2 = 700$

Table 6.3 contains information on the mean duration of user sessions on an app between "experimental" and "control" data sets. Test whether the partition is valid for this variable, using $\alpha = 0.05$.

Question 6:

Table 6.5 contains observed frequencies for education level in "experimental" and "control" data sets. Test whether the partition is valid for this variable, using $\alpha = 0.05$.

Data set	Below High School	High School	Bachelor's and Above	Total
Experimental	500	400	300	1200
Control	150	120	80	350
Total	650	520	380	1550

Question 7:

Table 6.8 shows the amount of time (in minutes) spent by a random sample of customers to receive services through the "website," "mobile app," and "in-person" respectively. Test whether the population mean time spent for receiving services differs among the three methods, using $\alpha = 0.05$.

Website	Mobile App	In-Person
25	35	40
30	40	45
28	38	50
32	36	48

Question 8:

Suppose we are interested in predicting students' scores based on the number of study hours. We ran a regression analysis, resulting in the following equation: "The estimated score equals (20) plus (3 times the number of study hours)."

- Suppose one student studied 5 hours more than another student. What is the estimated difference in their scores?
- Suppose a given student studied 10 hours. What is the estimated score?
- Suppose the above regression equation was based on a sample of students studying between 5 and 15 hours. Now estimate the score of a student who studied for 20 hours, and comment.
- Explain clearly the meaning of the 3 in the equation above.
- Explain clearly the meaning of the 20 in the equation above.