AccelerateAI

Data Science Global Bootcamp

Assignment 04

Q1: A company named Outel Semiconductors has developed a new microprocessor. It wants to test how fast one of these new chips can conduct a certain benchmark calculation. Suppose that the time it takes to complete the calculation is normally distributed. After 10 runs, the sample average time to completion is 32.7 nanoseconds, and the sample variance is 16 nanoseconds. Can Outel claim that true average time to completion is 30 nanoseconds at 95% confidence level?

Solution: This is a one sample test on mean.

$\mu = 30$

Null Hypothesis: $\mu = 30$
Alternate Hypothesis: $\mu < 30$ (this is what Outel wants to prove)

N = 10, hence we will conduct a one sample t-test (with dof = 9)

Sample mean $\bar{x} = 32.7$
Sample s.d. = 16

$$t = \frac{(\bar{x} - 30)}{16/\sqrt{10}}$$

Use python to get the p-value for the corresponding t-stat from the t-distribution.

p-val = scipy.stats.t.cdf(x=32.7, loc=30, scale=16, df=9)
= 0.565

Since the p-value is >> 0.05, we cannot reject the Null hypothesis at 95% confidence level.

Hence, Outel Semiconductors can claim that the average time for completion is 30 nanoseconds or less , at 95% confidence level.

Q2: Marketers believe that 92% of adults in the United States own a cell phone. A cell phone manufacturer believes that the number is actually lower. 200 American adults are surveyed, of which, 174 report having cell phones. Use a 5% level of significance.

a) State the null and alternative hypothesis,
b) find the p-value, state your conclusion, and
c) identify the Type I and Type II errors.

Solution:

a) This is a one sample test of proportion. Hypothesized value of proportion (of cell phone owners is 0.92)

Null Hypothesis, H0: $p = 0.92$
Alternate Hypothesis: Ha: $p < 0.92$

b) Given, $p = 174/200$

$Z = (p - p0)/\sqrt{p(1-p)/n}$, where $p0 = 0.92$
p-value = 0.0046

Because $p < 0.05$, we reject the null hypothesis in favour of the alternate hypothesis. There is sufficient evidence to conclude that fewer than 92% of American adults own cell phones.

c) Type I Error: To conclude that fewer than 92% of American adults own cell phones when, in fact, 92% of American adults do own cell phones (reject the null hypothesis when the null hypothesis is true).

Type II Error: To conclude that 92% of American adults own cell phones when, in fact, fewer than 92% of American adults own cell phones (do not reject the null hypothesis when the null hypothesis is false).

Q3: When a coin is tossed 100 times, suppose we get 60 heads and 40 tails. Test if the coin is fair versus the alternative it is loaded in favour of heads, using significance level alpha = 0.05.

Solution: We want to test

$$\mathrm{H}_0 : p = \frac{1}{2} \text{ (fair coin) versus } \mathrm{H}_a : p > \frac{1}{2} \text{ (here } p_0 = \frac{1}{2} = 0.5).$$

Since

$$\hat{p} = \text{ observed proportion } = \frac{60}{100} = 0.6,$$

Step 1

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{(0.6) - (0.5)}{\sqrt{\frac{(0.5)(0.5)}{100}}} = \frac{(0.1)(10)}{0.5} = 2.$$

Steps 2 and 3 For this one sided alternative, we reject H_0 if $z > z_\alpha$. From Table A, $z_\alpha = 1.645$. Since $z = 2 > 1.645$, we reject the null hypothesis H_0 that the coin is fair.

Q4. The table below contains data from a survey of 500 randomly selected households. Researchers would like to use the available sample information to test whether home ownership rates vary by household location. For example, is there a nonzero difference between the proportions of individuals who own their homes (as opposed to those who rent their homes) in households located in the SW and NW sectors of this community?
Use the sample data to test for a difference in home ownership rates in these two sectors. Use a 5% significance level. Interpret and summarize your results.

	Home Ownership		
	No	Yes	Grand Total
NW Sector	40	89	129
SW Sector	17	106	123

Solution:

Here we are to compare proportion for 2 independent samples.

Given:

Home ownership proportion in NW sector, pa = 89/129 = 0.6899

Home ownership proportion in NW sector, p_b = 106/123 = 0.8618

Null Hypothesis, H_0: pa = pb

Alternate Hypothesis, H_a: pa ≠pb

The pooled proportion: $pc = \frac{P_A + P_B}{n_A + n_B}$ P_A = 129, P_B = 106

Pc = $\frac{89+106}{129+123} = 0.7736$

Z = $\frac{pa-pb}{\sqrt{pc(1-pc)(\frac{1}{na}+\frac{1}{nb})}}$ = -3.2597

p-val = st.norm.cdf(x=-3.2579) = 0.0056

Since p-val is much smaller than 0.05, we reject the Null hypothesis, and conclude that the home ownership rates are not equal in the two regions.

Q5. Twenty people have rated a new beer on a taste scale of 0 to 100. Their ratings are in the file ***Q5_Beer_Taste.xlsx***. Marketing has determined that the beer will be a success if the average taste rating exceeds 76. Using a 5% significance level, is there sufficient evidence to conclude that the beer will be a success? Discuss your result in terms of a p-value. Assume ratings are at least approximately normally distributed.

Please see solution here: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment04/AAI-DS-Assignment04%20-%20Solution.ipynb

Q6. A market research consultant hired by a leading soft drink company wants to determine the proportion of consumers who favor its low-calorie drink over the

leading competitor's low-calorie drink in a particular urban location. A random sample of 250 consumers from the market under investigation is provided in the file ***Q6_Lowcalorie_Drink.xlsx***.

a. Find a 95% confidence interval for the proportion of all consumers in this market who prefer this company's drink over the competitors. What does this confidence interval tell us?
b. Does the confidence interval in part a support the claim made by one of the company's marketing managers that more than half of the consumers in this urban location favor its drink over the competitor's? Explain your answer.
c. Comment on the sample size used in this study. Specifically, is the sample unnecessarily large? Is it too small? Explain your reasoning.

Please see solution here: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment04/AAI-DS-Assignment04%20-%20Solution.ipynb

Q7. A large buyer of household batteries wants to decide which of two equally priced brands to purchase. To do this, he takes a random sample of 100 batteries of each brand. The lifetimes, measured in hours, of the batteries are recorded in the file ***Q7_Battery_life.csv***. Before testing for the difference between the mean lifetimes of these two batteries, he must first determine whether the underlying population variances are equal.

a. Perform a test for equal population variances. Report a p-value and interpret its meaning.
b. Based on your conclusion in part a, which test statistic should be used in performing a test for the difference between population means?

Please see solution here: https://github.com/Accelerate-AI/Data-Science-Global-Bootcamp/blob/main/ClassAssignment/Assignment04/AAI-DS-Assignment04%20-%20Solution.ipynb