AccelerateAI

Data Science Global Bootcamp **Assignment 04**

**Hypothesis Testing**

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?

**Ans 1 :**

**Approach :**

1. We will proceed with the t-test as the sample size is less than 30 and variance is unknown
2. Also, we will apply a two-tailed test, as the question doesn’t suggest a greater than or less than estimation

**Hypothesis values :**

* Null Hypothesis : µ = 30
* Alternate Hypothesis : µ ≠30

**Inputs :**

* s = 16
* = 32.7
* µ = 30
* n = 10
* DoF = 2n-1 = 2\*10 -1 = 19
* α = 5% which gets divided into 2.5% on both sides or .025

**Test Statistic and T-alpha-DoF :**

t = (µ - ) / (s/sqrt(n))

= (30 – 32.7) / (16/sqrt(10))

= -2.7/ 5.059

= -0.5337

t(α /DoF) =  [**+/- 2.093024**](https://goodcalculators.com/student-t-value-calculator/)

**Conclusion:**

*Since t is not less than the negative t(α /DoF), we cannot reject the null hypothesis (that true average time to completion is 30 nanoseconds 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.

1. State the null and alternative hypothesis,
2. find the p-value, state your conclusion, and
3. identify the Type I and Type II errors.

**Ans 2 :**

**Approach :**

1. We will proceed with the z-test for proportions
2. Also, we will apply a one-tailed test, as the we are trying to prove that the actual proportion is lower

**Hypothesis values :**

* Null Hypothesis : p = .92
* Alternate Hypothesis : p < .92

**Inputs :**

* p = .92
* = 174/200 = .87

**Test Statistic:**

* Z = ( – p)/sqrt(p(1-p)/n)
* = -.05 / sqrt(0.000368)
* = -.05 / 0.019183
* = -2.6064

**Evaluation and Conclusion:**

* Z = 2.6064
* P(Z) = 0.00457 or .457% [1-st.norm.cdf(2.6064)]
* Since p value is less than p\_alpha (5%), we reject the null hypothesis that 92% of adults in the US have a cellphone. We are inclined to accept the alternate hypothesis that the proportion is lower than 92%.

**Identification of Type 1 and Type 2 Errors:**

* Z = 2.6064
* P(Z) = 0.00457 or .457% [1-st.norm.cdf(2.6064)]
* **Type 1 Error = P(Z) = .457%**
* **For Type 2 Error, since True population mean or the Power of the test is not given, we won’t be able to calculate the same.**

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.

**Ans 3 :**

**Approach :**

1. We will proceed with the z-test for proportions
2. Also, we will apply a one-tailed test, as the we are trying to prove that the actual proportion is higher than the hypothesis value
3. We will consider p as the proportion that we get a heads.

**Hypothesis values :**

* Null Hypothesis : p = 0.5
* Alternate Hypothesis : p >0.5

**Inputs :**

* p = 0.5
* = 0.6

**Test Statistic:**

* Z = ( – p)/sqrt(p(1-p)/n)
* = 0.1 / sqrt(0.5 \* 0.5 /100)
* = 0.1 / 0.05
* = 2

**Evaluation and Conclusion:**

* Z = 2
* P(Z) = 0.0227 or 2.27% [1-st.norm.cdf(2)]
* Since p value is less than p\_alpha (5%), **we reject the null hypothesis that it is a fair coin**. We may be inclined to accept that it is in favor of heads.

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?

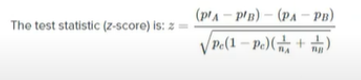
Use the sample data to test for a difference in home ownership rates in these two sectors. Use a 5% significance level. Interpret and summarize your results.

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

**Ans 4 :**

**Approach :**

1. We will proceed with the z-test for comparing proportions between two samples (two-tailed)
2. We will consider p as the proportion of home ownership.



1. Converting the table to proportions :

|  |  |  |
| --- | --- | --- |
|  | Proportions | Sample size |
| NW Sector | 0.6899225 | 129 |
| SW Sector | 0.8617886 | 123 |

**Hypothesis values :**

* pA = pB

**Inputs :**

* pA - pB = 0 (As we assume that both are equal
* = 0.69
* = 0.86
* pC = (x1+x2)/(n1+n2) = (89+106)/(129+123) = 195 / 252 = 0.7738

**Test Statistic:**

* Z =( - – (pA - pB) )/sqrt( [pC (1-pC){1/nA + 1/nB}])
* = (.69 - .86 – 0 ) / sqrt((.77 \* .23) \* {1/129 + 1/123})
* = -0.17/ 0.0028125251
* = -3.2055

**Evaluation and Conclusion:**

* Z = -3.2055
* P(Z) = 0.000674 or .06% [1-st.norm.cdf(-3.2055)]
* Since p value is less than p\_alpha (for two tailed) (2.5%), **we reject the null hypothesis that the sample proportions are equal**



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 pvalue. Assume ratings are at least approximately normally distributed.

*Submitted via Python Jupyter notebook.*

[*https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics*](https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics)

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***.

1. 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?
2. 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.
3. Comment on the sample size used in this study. Specifically, is the sample unnecessarily large? Is it too small? Explain your reasoning.

*Submitted via Python Jupyter notebook.*

[*https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics*](https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics)

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.

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

*Submitted via Python Jupyter notebook.*

[*https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics*](https://github.com/nageshn16/AccelerateAIGlobalBootcamp/tree/main/Assignments/04-Statistics)

All data files are posted at the link below:

https://github.com/Accelerate-AI/Data-Science-GlobalBootcamp/tree/main/ClassAssignment/Assignment04

