Homework Stats 2

Week 2

6733172621 Patthadon Phengpinij

Collaborators. ChatGPT (for LaTeX styling and grammar checking)

1 Null Hypothesis Significance Test

TO SUBMIT

Problem 4. Hamtaro and his casino:

After opening HamHub for a short while, the website was also banned by the government since it contains some 'immoral' videos. Hamtaro then moves on and follows his other passionate dream of creating a gambling empire. Therefore, he hones his skills on public gambling websites which can be easily found even if they are illegal.

After playing for a while, he notices that the online gambling business has great business potential since the risk of gambling websites being banned is much lower than his previous entertainment business. Thus, he decides to open his own online casino.

At the opening date, he offers only a dice game. The rule is simple, the player selects a number and rolls a die. The player will receive a reward if the rolled number is the same as the one he chooses. Hamtaro wants to maximize his profit by cheating using a biased die. Since it is an online casino, he could easily change the biasness of the die after the player selects a number. However, the player is not a fool and would notice if it is too biased.

As a player,

4.1 Formulate the null hypothesis H_0 and alternative hypothesis H_A to investigate the biasness of the dice.

Solution. From the problem statement, we want to investigate whether the die is biased or not. Let p be the probability of rolling the selected number. We can formulate the hypotheses as follows:

$$\boxed{H_0 : p = \frac{1}{6} \quad \text{(The die is fair) and } H_A : p < \frac{1}{6} \quad \text{(The die is biased)}}$$

The null hypothesis H_0 states that the die is fair, meaning the probability of rolling the selected number is $\frac{1}{6}$. The alternative hypothesis H_A states that the die is biased, meaning the probability of rolling the selected number is less than $\frac{1}{6}$.

4.2 Should the H_A be one-sided or two-sided? What are the differences and benefits over another in this problem?

Solution. In this problem, $\boxed{\text{the alternative hypothesis } H_A \text{ should be one-sided}}$. A one-sided hypothesis test is appropriate here because we are only interested in detecting if the die is biased towards rolling the selected number more frequently than expected.

4.3 The player found the selected number is rolled out 3 out of 30 attempts. If he wants no more than 10% of type-I error, can he reject the H_0? Justify your answer.

Solution. This is a hypothesis testing problem for a binomial distribution.

$$H_0 : p = \frac{1}{6} \; ; X \sim \text{Binomial}(30, \frac{1}{6})$$

We need to determine whether the observed value $X = 3$ falls within the rejection region for a significance level of $\alpha = 0.10$. Because we are conducting a one-sided test, we need to consider only the lower tail of the distribution.

$$\text{The rejection region is } \alpha = 0.10$$

Calculating the cumulative probabilities, we find:

- The lower tail critical value k such that $P(X \leq k) \leq 0.1$

Consider the cumulative probabilities:

$$P(X \leq k) = \sum_{i=0}^{k} P(X = i) = \sum_{i=0}^{k} \binom{30}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{30-i}$$

Calculating $P(X \leq k = 3)$:

$$P(X \leq 3) \approx 0.184$$

Since $P(X \leq 3) \approx 0.184 > 0.10$, we do not reject H_0.

4.4 If the player plays 200 games, what is the rejection region if he wants no more than 10% type-I error?

Solution. Like before, we need to determine the rejection region for a significance level of $\alpha = 0.10$. Consider the cumulative probabilities:

$$P(X \leq k) = \sum_{i=0}^{k} P(X = i) = \sum_{i=0}^{k} \binom{200}{i} \left(\frac{1}{6}\right)^{i} \left(\frac{5}{6}\right)^{200-i}$$

Using python to calculate the cumulative probabilities:

```python
number_of_trials = 200
p_null = 1/6
alpha = 0.10
cumulative_prob = 0.0

for k in range(number_of_trials + 1):
    cumulative_prob += math.comb(number_of_trials, k) * (
    p_null ** k) * ((1 - p_null) ** (number_of_trials - k)
    )
    if cumulative_prob >= alpha:
        lower_critical_value = k
        break

print('Lower critical value:', lower_critical_value - 1)
```

from the result, we find that the rejection region is:

$$\boxed{\text{Rejection Region: } X \leq 26}$$

4.5 What would be the result in 4. if the true distribution is approximated by the Normal distribution?

Solution. Using the Normal approximation to the Binomial distribution, we have:

$$\mu = np = 200 \times \frac{1}{6} = \frac{100}{3}$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{200 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{\frac{1000}{18}} = \frac{10\sqrt{5}}{3}$$

Using z-score table to find the critical value for $\alpha = 0.10$:

$$P(Z \leq z_\alpha) = 0.10 \implies z_\alpha \approx -1.28$$

Calculating the lower critical value using the z-score:

$$X = \mu + z_\alpha \sigma = \frac{100}{3} + (-1.28) \times \frac{10\sqrt{5}}{3} \approx 26.05$$

Thus, the rejection region using the Normal approximation is:

$$\boxed{\text{Rejection Region: } X \leq 26}$$

As a Hamtaro,
(**Hint:** Problem 6 and 7 are related to test power)

4.6 The mastermind Hamtaro observes that players will play no more than 200 games a day. He knows that some players studied **COMP ENG MATH 2** and might perform hypothesis testing to check whether Hamtaro cheats. Hamtaro assumes that the players will use a significant level of 0.01. He thinks that it is safe enough if the probability of being caught by a player is less than 0.05. What should be the lowest probability of rolling the selected number? (How much bias can he put in the dice) Answer in floating number with a precision of 3.

Solution. First, we need to determine the rejection region for a significance level of $\alpha = 0.01$ with $n = 200$. Using python to calculate the cumulative probabilities:

```python
number_of_trials = 200
p_null = 1/6
alpha = 0.10
cumulative_prob = 0.0

for k in range(number_of_trials + 1):
    cumulative_prob += math.comb(number_of_trials, k) * (p_null ** k) * ((1 - p_null) ** (number_of_trials - k))
    if cumulative_prob >= alpha:
        lower_critical_value = k
        break

print('Lower critical value:', lower_critical_value - 1)
```

From the result, we find that the rejection region is:

$$\text{Rejection Region: } X \leq 21$$

Next, we need to find the lowest probability p such that the power of the test is at least 0.95 (i.e., the probability of being caught is less than 0.05).

```
target_power = 0.95

for p in np.arange(0, 1, 0.001):
    cumulative_prob = 0.0
    for k in range(22):
        cumulative_prob += math.comb(number_of_trials, k)
    * (p ** k) * ((1 - p) ** (number_of_trials - k))
    power = 1 - cumulative_prob
    if power >= target_power:
        lowest_p = p
        break

print('Lowest probability p:', round(lowest_p, 3))
```

From the result, the lowest probability p that Hamtaro can use is:

$$\boxed{p = 0.148}$$

4.7 What if Hamtaro accepts the probability of being caught equal to 0.01 instead? Answer in floating number with the precision of 5.

Solution. From the previous calculation, we need to find the lowest probability p such that the power of the test is at least 0.99 (i.e., the probability of being caught is less than 0.01).

```
target_power = 0.99

for p in np.arange(0, 1, 0.001):
    cumulative_prob = 0.0
    for k in range(22):
        cumulative_prob += math.comb(number_of_trials, k)
    * (p ** k) * ((1 - p) ** (number_of_trials - k))
    power = 1 - cumulative_prob
    if power >= target_power:
        lowest_p = p
        break

print('Lowest probability p:', round(lowest_p, 3))
```

From the result, the lowest probability p that Hamtaro can use is:

$$\boxed{p = 0.167}$$

TO SUBMIT

Problem 6. Hamtaro and his casino:

Hamtaro also have a factory. He tried to boost the factory productivity by replacing the old machines with a new type-II variant. However, there is a concern from the local factory managers that Hamtaro might get bamboozled, since they do not observe an increase in productivity compared to the previous one. Therefore, to ease their concern, he decided to conduct a z-testing.

Given that the number of goods produced each day by the old machines was $x \sim \mathcal{N}(5000, 20^2)$:

6.1 Formulate the null and alternative hypothesis for determining whether the new machine is better than the previous one at a significant level = 0.05.

Solution. From the problem statement, we want to investigate whether the new machine increases productivity compared to the old machine. Let μ be the mean productivity of the new machine. We can formulate the hypotheses as follows:

$$\boxed{H_0 : \mu = 5000 \quad \text{(The new machine does not increase productivity)}}$$

$$\boxed{H_A : \mu > 5000 \quad \text{(The new machine increases productivity)}}$$

The null hypothesis H_0 states that the new machine does not increase productivity, meaning the mean productivity is equal to 5000. The alternative hypothesis H_A states that the new machine increases productivity, meaning the mean productivity is greater than 5000.

For the following problems, we will use the data from 4 factories provided below. The following information is given for the testing:

```python
from scipy.stats import norm
import numpy as np

# 30 days of product quantity in 4 factories

fac_0 = np.array([
    4993.89323126, 5021.67118211, 5023.54710937,
    4999.11746331, 5001.53450095, 4986.27990953,
    4987.12362188, 5004.91535087, 4999.97591193,
    5038.09176163, 4993.94184053, 5026.52644680,
    5040.62862593, 4979.91124088, 5008.59143715,
    5016.45331659, 5013.63203948, 5010.84253735,
    5014.99772195, 5002.39462129, 5047.80507624,
    5007.23005532, 5019.87205007, 5005.76363012,
    4997.09106036, 4982.80291132, 5037.18158407,
    4996.54197735, 5007.57964251, 4971.18880247
])
```

```python
fac_1 = np.array([
    5036.80041897, 4989.33779117, 4971.68709581,
    5041.92493487, 5041.64823146, 5026.33602398,
    5009.58334612, 4989.05382998, 5031.17423169,
    4992.20198911, 4970.63425555, 5007.17615704,
    4993.84416738, 5028.59671588, 5009.91388156,
    5049.64187466, 5015.12711371, 5032.29005130,
    5013.66869347, 4988.21257317, 5020.44276181,
    4985.62886808, 5022.46800468, 5042.35501669,
    5001.61539080, 5012.14209858, 5006.14666402,
    4999.93219541, 5002.77927647, 5002.20750425
])

fac_2 = np.array([
    5029.95293241, 5019.47959949, 4976.84278360,
    4985.22792264, 4994.97618684, 5026.75059569,
    5015.71350753, 5008.46632673, 5037.96915682,
    4990.38948551, 4988.70822060, 5032.42440206,
    5036.41040953, 5003.75236158, 5002.62361815,
    4998.89320570, 5000.51153033, 5002.19196574,
    5023.74534474, 5032.03601587, 5000.10614764,
    4989.74566985, 4985.97436664, 4973.63380449,
    5028.58100504, 4997.84267810, 5011.42021980,
    5018.71432385, 4969.03296199, 5009.23456565
])

fac_3 = np.array([
    4962.36508403, 5015.91734917, 5030.86885403,
    5012.74787091, 5036.94455211, 4995.21037570,
    5029.84241184, 5015.68062582, 4996.43546786,
    4999.57614716, 5006.88735305, 5035.10432486,
    5017.33437936, 5006.70625696, 5007.97827037,
    4981.80482708, 5020.78603239, 4993.12742287,
    4996.10718141, 4988.00315629, 5003.00004152,
    4949.54117305, 5008.62500480, 5004.09075453,
    5026.56246304, 5011.02296759, 5010.67413795,
    4990.58062539, 5009.64435703, 5001.94134280
])
```

For the problem 6.2., we will create a `NDArray` that contains all the data from the 4 factories, `fac_whole`. And, we also create `factories`, which is a dictonary that maps the factory name to its corresponding data.

```python
fac_whole = np.concatenate((fac_0, fac_1, fac_2, fac_3))

factories = {
    "Factory 0": fac_0,
    "Factory 1": fac_1,
    "Factory 2": fac_2,
    "Factory 3": fac_3,
    "All Factories": fac_whole
}
```

6.2 From the testing, can Hamtaro conclude that factory productivity increased as a whole?

Solution. To determine whether Hamtaro can conclude that factory productivity increased as a whole, we need to perform a z-test using the provided data. Assuming we have the sample mean $\bar{x}$ and sample size n, we can calculate the z-score as follows:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where $\mu_0 = 5000$ is the mean under the null hypothesis, and $\sigma = 20$ is the standard deviation of the old machine.

Next, we compare the calculated z-score to the critical value for a one-tailed test at a significance level of $\alpha = 0.05$. Considering the z-table, the critical value is approximately:

$$P(Z \leq z) = 0.05 \implies z_\alpha \approx 1.645$$

.
Calculating the z-score with the provided data will give us the final conclusion.

$$z = \frac{\bar{x} - 5000}{20/\sqrt{n}}$$

Next, using python to calculate the z-score and compare it with the critical value:

```python
mu_0 = 5000
sigma_0 = 20
z_alpha = norm.ppf(0.95)

for name, data in factories.items():
    n = len(data)
    sample_mean = np.mean(data)

    z_score = (sample_mean - mu_0) / (sigma_0 / np.sqrt(n))

    decision = "Reject" if z_score > z_alpha else "Accept"

    print(f"{name}: z-score = {z_score:.4f}, Decision: {decision}")
```

The result from the code above are as follows:

Factory	z-score	Decision
Factory 0	2.1647	Reject
Factory 1	3.0542	Reject
Factory 2	1.7468	Reject
Factory 3	1.5072	Accept
All Factories	4.2365	Reject

The results show that for "All Factories", the z-score is greater than the critical value, leading to the rejection of the null hypothesis.

Thus, Hamtaro can conclude that factory productivity increased as a whole.

6.3 Can Hamtaro say the same for each individual factory?

Solution. Using the results from problem 6.2., we can see that for each individual factory:

Factory	z-score	Decision
Factory 0	2.1647	Reject
Factory 1	3.0542	Reject
Factory 2	1.7468	Reject
Factory 3	1.5072	Accept
All Factories	4.2365	Reject

Thus, Hamtaro cannot say the same for each individual factory. (Factory 3 does not show a significant increase in productivity.)

6.4 Repeat 6.1 - 6.3 again but with a t-test. Is there any difference from the z-test? What, and why does it happen?

Solution. To perform a t-test, we will use the sample standard deviation instead of the population standard deviation. The t-score is calculated as follows:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where s is the sample standard deviation.
The critical value for a one-tailed t-test at a significance level of $\alpha = 0.05$ with $n-1$ degrees of freedom can be found using the t-distribution table.

```python
from scipy.stats import t

for name, data in factories.items():
    n = len(data)
    sample_mean = np.mean(data)
    sample_std = np.std(data, ddof=1)

    t_score = (sample_mean - mu_0) / (sample_std / np.sqrt(n))
    t_alpha = t.ppf(0.95, df=n-1)

    decision = "Reject" if t_score > t_alpha else "Accept"

    print(f"{name}: t-score = {t_score:.4f}, Decision: {decision}")
```

The result from the code above are as follows:

Factory	t-score	Decision
Factory 0	2.3262	Reject
Factory 1	2.8827	Reject
Factory 2	1.8045	Reject
Factory 3	1.5589	Accept
All Factories	4.3410	Reject

The results of using the t-test are similar to those of the z-test.

The decisions are the same for all factories.

The reason for this similarity is that the sample size is **relatively large** (n=30 for each factory), which makes the t-distribution approach the normal distribution. Thus, the results of the t-test and z-test converge, leading to the same conclusions.