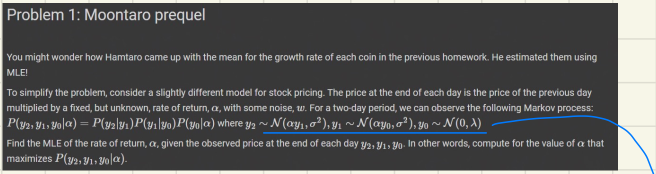


Problem 1: Moontaro prequel

You might wonder how Hamtaro came up with the mean for the growth rate of each coin in the previous homework. He estimated them using MLE!

To simplify the problem, consider a slightly different model for stock pricing. The price at the end of each day is the price of the previous day multiplied by a fixed, but unknown, rate of return, α, with some noise, w. For a two-day period, we can observe the following Markov process: $P(y_2, y_1, y_0|\alpha) = P(y_2|y_1)P(y_1|y_0)P(y_0|\alpha)$ where $y_2 \sim \mathcal{N}(\alpha y_1, \sigma^2), y_1 \sim \mathcal{N}(\alpha y_0, \sigma^2), y_0 \sim \mathcal{N}(0, \lambda)$

Find the MLE of the rate of return, α, given the observed price at the end of each day y_2, y_1, y_0. In other words, compute for the value of α that maximizes $P(y_2, y_1, y_0|\alpha)$.

1) $$P(y_2, y_1, y_0 | \alpha) = P(y_2|y_1)\, P(y_1|y_0)\, P(y_0|\alpha)$$

$$= \left[\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_2-\alpha y_1}{\sigma}\right)^2}\right]\left[\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_1-\alpha y_0}{\sigma}\right)^2}\right]\left[\frac{1}{\sqrt{\lambda}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{y_0-0}{\sqrt{\lambda}}\right)^2}\right]$$

$$= \frac{1}{\sigma^2(2\pi)} \cdot \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}\left(\left(\frac{y_2-\alpha y_1}{\sigma}\right)^2+\left(\frac{y_1-\alpha y_0}{\sigma}\right)^2+\left(\frac{y_0}{\sqrt{\lambda}}\right)^2\right.}$$

consider $$\ln\left(\frac{1}{\sigma^2(2\pi)} \cdot \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}\left(\left(\frac{y_2-\alpha y_1}{\sigma}\right)^2+\left(\frac{y_1-\alpha y_0}{\sigma}\right)^2+\left(\frac{y_0}{\sqrt{\lambda}}\right)^2\right)}\right) = \ln\left(\frac{1}{\sigma^2(2\pi)}\right) + \ln\left(\frac{\lambda^{-\frac{1}{2}}}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\left(\frac{y_2-\alpha y_1}{\sigma}\right)^2+\left(\frac{y_1-\alpha y_0}{\sigma}\right)^2+\left(\frac{y_0}{\sqrt{\lambda}}\right)^2\right.$$

find maximize α → $$\frac{d}{d\alpha}\left(\ln\left(\frac{1}{\sigma^2(2\pi)}\right) + \ln\left(\frac{\lambda^{-\frac{1}{2}}}{\sqrt{2\pi}}\right) - \frac{1}{2}\left(\left(\frac{y_2-\alpha y_1}{\sigma}\right)^2+\left(\frac{y_1-\alpha y_0}{\sigma}\right)^2+\left(\frac{y_0}{\sqrt{\lambda}}\right)^2\right)\right) = 0$$

$$0 + 0 - \frac{1}{\cancel{2}}\left(-\cancel{2}y_1\left(\frac{y_2-\alpha y_1}{\sigma}\right) - \cancel{2}y_0\left(\frac{y_1-\alpha y_0}{\sigma}\right) + 0\right) = 0$$

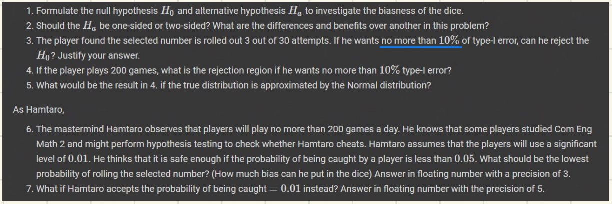
$$y_1\left(\frac{y_2-\alpha y_1}{\cancel{\sigma}}\right) + y_0\left(\frac{y_1-\alpha y_0}{\cancel{\sigma}}\right) = 0$$

$$y_1y_2 - \alpha y_1^2 + y_0y_1 - \alpha y_0^2 = 0$$

$$\frac{y_1y_2 + y_0y_1}{y_0^2 + y_1^2} = \alpha$$

#

5)



1. Formulate the null hypothesis H_0 and alternative hypothesis H_a to investigate the biasness of the dice.
2. Should the H_a be one-sided or two-sided? What are the differences and benefits over another in this problem?
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.
4. If the player plays 200 games, what is the rejection region if he wants no more than 10% type-I error?
5. What would be the result in 4. if the true distribution is approximated by the Normal distribution?

As Hamtaro,

6. The mastermind Hamtaro observes that players will play no more than 200 games a day. He knows that some players studied Com 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.
7. What if Hamtaro accepts the probability of being caught $= 0.01$ instead? Answer in floating number with the precision of 5.

1) H_0 : The dice is unbiased

H_a : The dice is biased, making player lose easier

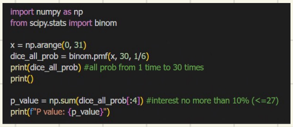
2) H_a should be One-sided because as a player, we just interest that the dice is biased or not and make it harder for they to win $(p < \frac{1}{6})$

3) significance : 0.1

from scipy.stat.binom

we will get p value = 0.24 > 0.1

∴ not reject H_0

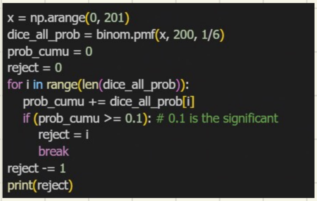


```
import numpy as np
from scipy.stats import binom

x = np.arange(0, 31)
dice_all_prob = binom.pmf(x, 30, 1/6)
print(dice_all_prob) #all prob from 1 time to 30 times
print()

p_value = np.sum(dice_all_prob[:4]) #interest no more than 10% (<=27)
print(f"P value: {p_value}")
```

4) From this code, The reject region is number rolled out ≤ 26



```
x = np.arange(0, 201)
dice_all_prob = binom.pmf(x, 200, 1/6)
prob_cumu = 0
reject = 0
for i in range(len(dice_all_prob)):
    prob_cumu += dice_all_prob[i]
    if (prob_cumu >= 0.1): # 0.1 is the significant
        reject = i
        break
reject -= 1
print(reject)
```

5) $P(Z < z) = 0.1$

$z = -1.282$

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

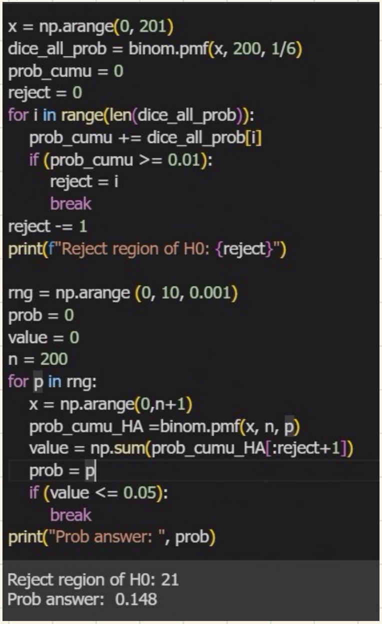
$\sigma^2 = np(1-p) = 200 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{250}{9}$

$z = \frac{x_i - \mu}{\sigma}$

$-1.282 = \frac{x_i - \frac{100}{3}}{\sqrt{\frac{250}{9}}}$

$x_i \approx 26.58$ $\therefore$ The reject region is number rolled out ≤ 26 (Same as 4))

6)

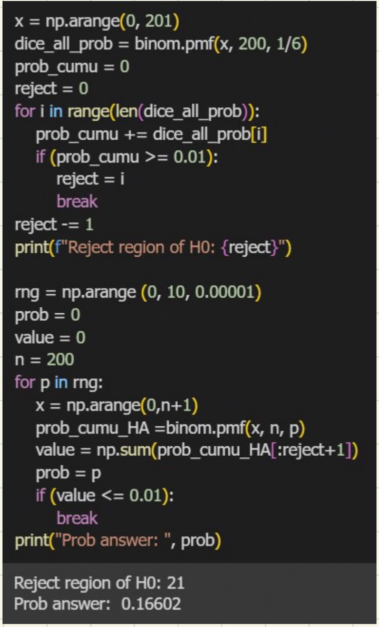


```
x = np.arange(0, 201)
dice_all_prob = binom.pmf(x, 200, 1/6)
prob_cumu = 0
reject = 0
for i in range(len(dice_all_prob)):
    prob_cumu += dice_all_prob[i]
    if (prob_cumu >= 0.01):
        reject = i
        break
reject -= 1
print(f"Reject region of H0: {reject}")

rng = np.arange (0, 10, 0.001)
prob = 0
value = 0
n = 200
for p in rng:
    x = np.arange(0,n+1)
    prob_cumu_HA =binom.pmf(x, n, p)
    value = np.sum(prob_cumu_HA[:reject+1])
    prob = p
    if (value <= 0.05):
        break
print("Prob answer: ", prob)
```

```
Reject region of H0: 21
Prob answer:  0.148
```

7)

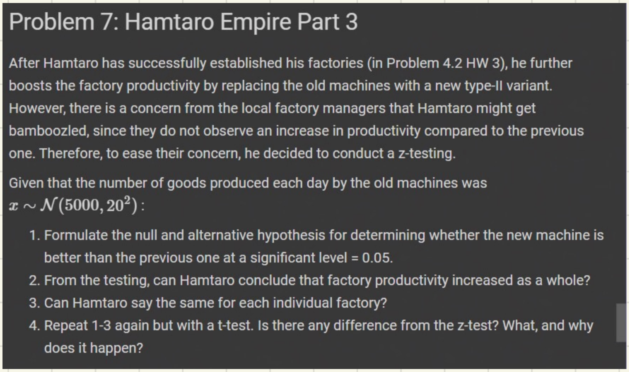


```
x = np.arange(0, 201)
dice_all_prob = binom.pmf(x, 200, 1/6)
prob_cumu = 0
reject = 0
for i in range(len(dice_all_prob)):
    prob_cumu += dice_all_prob[i]
    if (prob_cumu >= 0.01):
        reject = i
        break
reject -= 1
print(f"Reject region of H0: {reject}")

rng = np.arange (0, 10, 0.00001)
prob = 0
value = 0
n = 200
for p in rng:
    x = np.arange(0,n+1)
    prob_cumu_HA =binom.pmf(x, n, p)
    value = np.sum(prob_cumu_HA[:reject+1])
    prob = p
    if (value <= 0.01):
        break
print("Prob answer: ", prob)
```

```
Reject region of H0: 21
Prob answer:  0.16602
```

7)



Problem 7: Hamtaro Empire Part 3

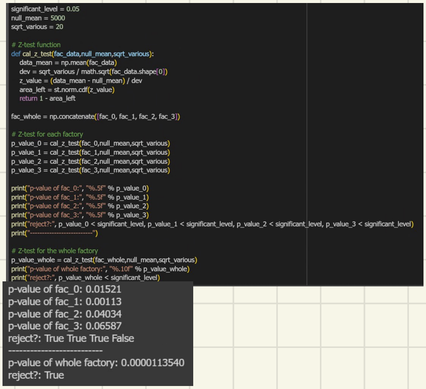
After Hamtaro has successfully established his factories (in Problem 4.2 HW 3), he further boosts the factory productivity by replacing the old machines with a new type-II variant. However, there is a concern from the local factory managers that Hamtaro might get bamboozled, since they do not observe 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 $x \sim \mathcal{N}(5000, 20^2)$:

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.
2. From the testing, can Hamtaro conclude that factory productivity increased as a whole?
3. Can Hamtaro say the same for each individual factory?
4. Repeat 1-3 again but with a t-test. Is there any difference from the z-test? What, and why does it happen?

1) H_0 : New machine is not better than the previous one $N(5000, 20^2)$

H_a : New machine is better than the previous one



```
significant_level = 0.05
null_mean = 5000
sqrt_various = 20

# Z-test function
def cal_z_test(fac_data,null_mean,sqrt_various):
    data_mean = np.mean(fac_data)
    dev = sqrt_various / math.sqrt(fac_data.shape[0])
    z_value = (data_mean - null_mean) / dev
    area_left = st.norm.cdf(z_value)
    return 1 - area_left

fac_whole = np.concatenate([fac_0, fac_1, fac_2, fac_3])

# Z-test for each factory
p_value_0 = cal_z_test(fac_0,null_mean,sqrt_various)
p_value_1 = cal_z_test(fac_1,null_mean,sqrt_various)
p_value_2 = cal_z_test(fac_2,null_mean,sqrt_various)
p_value_3 = cal_z_test(fac_3,null_mean,sqrt_various)

print("p-value of fac_0:", "%.5f" % p_value_0)
print("p-value of fac_1:", "%.5f" % p_value_1)
print("p-value of fac_2:", "%.5f" % p_value_2)
print("p-value of fac_3:", "%.5f" % p_value_3)
print("reject?:", p_value_0 < significant_level, p_value_1 < significant_level, p_value_2 < significant_level, p_value_3 < significant_level)
print("------------------------")

# Z-test for the whole factory
p_value_whole = cal_z_test(fac_whole,null_mean,sqrt_various)
print("p-value of whole factory:", "%.10f" % p_value_whole)
print("reject?:", p_value_whole < significant_level)
```

```
p-value of fac_0: 0.01521
p-value of fac_1: 0.00113
p-value of fac_2: 0.04034
p-value of fac_3: 0.06587
reject?: True True True False
-------------------------
p-value of whole factory: 0.0000113540
reject?: True
```

2) For whole factory

reject H_0 because p value < significant level

"New machine is better than the previous one"

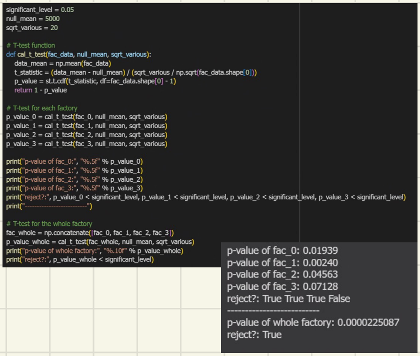
3) For each factory

reject H_0 of factory number 0, 1, 2

but do not reject H_0 of factory number 3

4) H_0 : New machine is not better than the previous one

H_a : New machine is better than the previous one



```
significant_level = 0.05
null_mean = 5000
sqrt_various = 20

# T-test function
def cal_t_test(fac_data, null_mean, sqrt_various):
    data_mean = np.mean(fac_data)
    t_statistic = (data_mean - null_mean) / (sqrt_various / np.sqrt(fac_data.shape[0]))
    p_value = st.t.cdf(t_statistic, df=fac_data.shape[0] - 1)
    return 1 - p_value

# T-test for each factory
p_value_0 = cal_t_test(fac_0, null_mean, sqrt_various)
p_value_1 = cal_t_test(fac_1, null_mean, sqrt_various)
p_value_2 = cal_t_test(fac_2, null_mean, sqrt_various)
p_value_3 = cal_t_test(fac_3, null_mean, sqrt_various)

print("p-value of fac_0:", "%.5f" % p_value_0)
print("p-value of fac_1:", "%.5f" % p_value_1)
print("p-value of fac_2:", "%.5f" % p_value_2)
print("p-value of fac_3:", "%.5f" % p_value_3)
print("reject?:", p_value_0 < significant_level, p_value_1 < significant_level, p_value_2 < significant_level, p_value_3 < significant_level)
print("------------------------")

# T-test for the whole factory
fac_whole = np.concatenate([fac_0, fac_1, fac_2, fac_3])
p_value_whole = cal_t_test(fac_whole, null_mean, sqrt_various)
print("p-value of whole factory:", "%.10f" % p_value_whole)
print("reject?:", p_value_whole < significant_level)
```

```
p-value of fac_0: 0.01939
p-value of fac_1: 0.00240
p-value of fac_2: 0.04563
p-value of fac_3: 0.07128
reject?: True True True False
-------------------------
p-value of whole factory: 0.0000225087
reject?: True
```

For whole factory

reject H_0 because p value < significant level

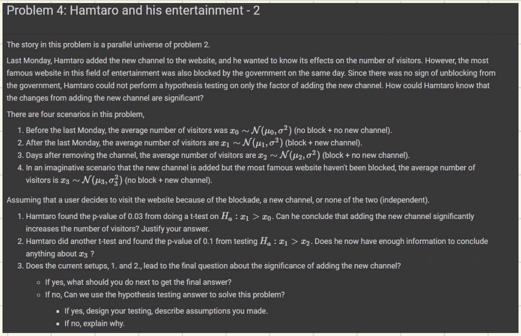
"New machine is better than the previous one"

For each factory

reject H_0 of factory number 0, 1, 2

but do not reject H_0 of factory number 3

The result is the same as z-test but p-value has a little different becuase Student't distribution and Normal distribution are similar, but not identical



Problem 4: Hamtaro and his entertainment - 2

The story in this problem is a parallel universe of problem 2.

Last Monday, Hamtaro added the new channel to the website, and he wanted to know its effects on the number of visitors. However, the most famous website in this field of entertainment was also blocked by the government on the same day. Since there was no sign of unblocking from the government, Hamtaro could not perform a hypothesis testing on only the factor of adding the new channel. How could Hamtaro know that the changes from adding the new channel are significant?

There are four scenarios in this problem,

1. Before the last Monday, the average number of visitors was $x_0 \sim \mathcal{N}(\mu_0, \sigma^2)$ (no block + no new channel).
2. After the last Monday, the average number of visitors are $x_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ (block + new channel).
3. Days after removing the channel, the average number of visitors are $x_2 \sim \mathcal{N}(\mu_2, \sigma^2)$ (block + no new channel).
4. In an imaginative scenario that the new channel is added but the most famous website haven't been blocked, the average number of visitors is $x_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$ (no block + new channel).

Assuming that a user decides to visit the website because of the blockade, a new channel, or none of the two (independent).

1. Hamtaro found the p-value of 0.03 from doing a t-test on $H_a : x_1 > x_0$. Can he conclude that adding the new channel significantly increases the number of visitors? Justify your answer.
2. Hamtaro did another t-test and found the p-value of 0.1 from testing $H_a : x_1 > x_2$. Does he now have enough information to conclude anything about x_3 ?
3. Does the current setups, 1. and 2., lead to the final question about the significance of adding the new channel?
    - If yes, what should you do next to get the final answer?
    - If no, Can we use the hypothesis testing answer to solve this problem?
        - If yes, design your testing, describe assumptions you made.
        - If no, explain why.

1. No, because it has other factor relevent (block and non-block)

2. No, because p value he get is 0.1 that is not less than the common significant level so he can't reject Hypothesis and make a conclusion

3. Yes, from 2). you will get that adding channal do not make visitors increase significantly and from 1). Adding channal and block has impact to the visitors number significantly

<u>Ans</u> Blocking effects to the visitors number significantly but Adding channal does not effect.