Statistical Data Analysis Problem sheet 6 Solution - Group 9

Exercise 1

1. Here given,
   significance level

$$\alpha = 0.05 \; for \; observed$$
$$\bar{x} = 1.5 \; with \; n = 15$$

   Where,

$$\sigma^2 = 4 \; and$$
$$H_0 : \theta = 1 \; versus \; H_1 : \mu \neq 1$$

Here, the test statistic is

$$\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}$$

So, the test reject the Null Hypothesis if

$$\left| \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} \right| \geq t_{\alpha/2, n-1}$$

where n-1 is the degrees of freedom and α is the significance level. So, we get

$$\left| \frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} \right| \geq t_{0.025, 14}$$

Now if we use the values(We are using σ value for the *S*), our test statistic

$$Z = \left| \frac{1.5 - 1}{2/\sqrt{15}} \right| = 0.968$$

Now, from the critical value chart, we get

$$t_{0.025, 14} = 1.761$$

Since,

$$|Z| = 0.968 < t_{0.025, 14} = 1.761$$

We fail to reject the Null Hypothesis. We do not have enough evidence at 5% significance level.

---

2. A power function depends on the null hypothesis.

- Here we will define some functions to get the power for different sample size given the standard deviation is σ = 2, Θ = 1.

```
from math import *
def phi(x):
    #'Cumulative distribution function for the standard normal distribution given the sigma = 2, theta =1'
    return (1.0 + erf(x / 2*sqrt(2.0))) / 2.0
```

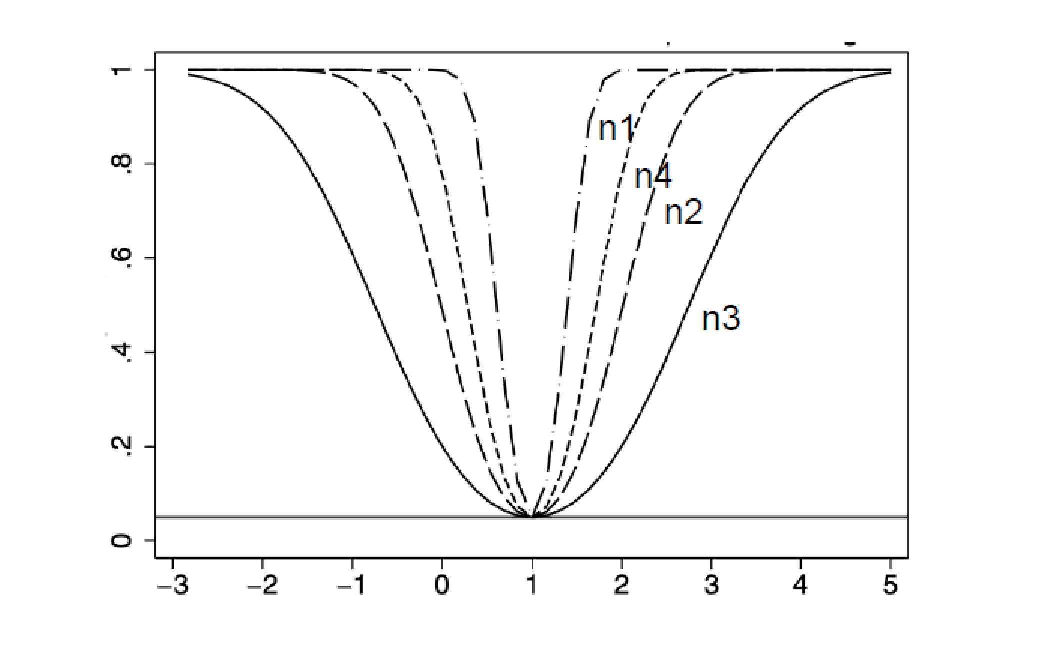
```
import scipy.stats
def critv(n):
  #critical value calculator for signifiance level 0.05
  res = scipy.stats.t.ppf(q=1-0.05/2,df=(n-1))
  return res

```

```
def powerfunc(n):
  #Power function
  return (1-phi(critv(n)-1*sqrt(n))+phi(-critv(n)-1*sqrt(n)))
```

```
#for different sample sizes, the power for the given sigma
samplist = [5,15,30,100]
for i in samplist:
  powerv = powerfunc(i)
  print(f"Power for sample size {i} = {powerv}")
```

```
Power for sample size 5 = 0.29446875765867175
Power for sample size 15 = 0.9580235145565433
Power for sample size 30 = 0.9997004217282532
Power for sample size 100 = 0.9999999999999994
```

Here we can see that when the sample size increases, the power function gets narrower. Thus we can say that the outermost function is associated with the least sample size n_3 = 5, and so on till the innermost function is associated with the largest sample size n_1 =100.



---

3. Here given,

$$\alpha = 0.05,\ emperical\ mean\ \bar{x} = 1.5\ and\ sample\ size\ n = 15$$

- We have to determine the confidence interval for Θ given the values and here we will use z score as σ is known.
- We can derive the confidence interval for Θ when σ is known.
  Here

$$z_{1-\alpha/2}\ denote\ the\ (1-\alpha/2)\ quantile\ of\ the\ distribution\ with\ n-1\ degrees\ of\ freedom.$$

Then using the symmetry of the t distribution, we have

$$1-\alpha = P\bigg(-z_{1-\alpha/2} < \frac{\bar{x}-\theta}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\bigg)$$

$$= P\bigg(-z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \bar{x}-\theta < z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\bigg)$$

$$= P\bigg(\bar{x} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \theta < \bar{x} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\bigg)$$

Then (1 - α) = 95% confidence interval for Θ is:

$$CI_{1-0.05}(\theta) = \bigg(\bar{x} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\bigg)$$

$$CI_{0.95}(\theta) = \bigg(1.5 - z_{0.975} \times \frac{2}{\sqrt(15)}\ ,\ 1.5 + z_{0.975} \times \frac{2}{\sqrt(15)}\bigg)$$

From the z-table we can find the 0.975 quantile value is 1.96.
Now we get

$$CI_{0.95}(\theta) = (0.488, 2.512)$$

Thus, we are 95% confident that Θ is between 0.488 and 2.512.

---

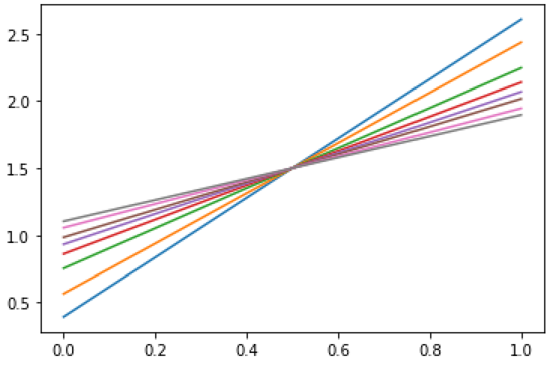
4. If we increase the sample size, the confidence interval will change.

- Let the emprical mean and variance is unchanged.
- Let us find 95% confidence interval for different sample size to see the change.

```
#defining a 95% confidence interval finding function
def confint(x):
  resL = 1.5 - critv(x)*(2/sqrt(x))
  resR = 1.5 + critv(x)*(2/sqrt(x))
  return (resL, resR)
```

```
# creating an increasing sample size list
import matplotlib.pyplot as plt
sampleSize = [15,20,30,40,50,60,80,100]
intervals = []
for i in sampleSize:
  intervs = confint(i)
  intervals.append(intervs)
  print(f"For sample size = {i}, the 95% confidence level is : {intervs}")
  plt.plot(intervs)
plt.show()
```

```
For sample size = 15, the 95% confidence level is : (0.3924369168707169, 2.6075630
For sample size = 20, the 95% confidence level is : (0.5639711871601811, 2.4360288
For sample size = 30, the 95% confidence level is : (0.7531877264838003, 2.2468122
For sample size = 40, the 95% confidence level is : (0.8603689716395658, 2.1396310
For sample size = 50, the 95% confidence level is : (0.9316062897552536, 2.0683937
For sample size = 60, the 95% confidence level is : (0.9833452152569502, 2.0166547
For sample size = 80, the 95% confidence level is : (1.054921802463506, 1.94507819
For sample size = 100, the 95% confidence level is : (1.1031566096982635, 1.896843
```



As we can see, with the increasing sample size, the confidence interval is getting tighter which gives us better precision with a smaller margin of error.

Exercise 2

1. Here,

$$\theta = 25,\ \alpha = 0.05,\ \bar{x} = 26,\ n = 49$$
$$\sigma^2 = 9 \ or\ \sigma = 3$$

Now, the test statistic,

$$\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}$$
$$= \frac{26 - 25}{3/\sqrt(49)}$$
$$= 2.333$$

Using the table we get the critical value for α = 0.05, which is

$$z = 1.645$$

Since, Our test score 2.333 > 1.645, we can reject the null hypothesis.
There is sufficient evidence to conclude that the null hypothesis is incorrect and the alternative hypothesis is Correct. So, we can say with 95% confidence that the population mean Θ > 25.

---

2. *The first type of error occurs when we reject the null hypothesis from the sample result but it's true for the population. SO, we define the Type I Error as the probabilty of rejecting a null hypothesis when it is True in the population. The probability of Type I Error is α.*
Here, given the significant level, α = 0.05. This means that there is a 5% probability that the test will reject the the null hypothesis when it is true which is also called as false positive. So, when we are rejecting the null hypothesis from the sample result we are only 95% confident and, there is still a 5% chance that the population mean Θ ≤ 25.
Although type I can never be avoided entirely, we can reduce the likelihood by increasing the sample size (the larger the sample, the lesser is the likelihood that it will differ substantially from the population). Also, using a lower significance level would lower the probablilty but using a lower α means it will be less likely to find a true difference if exists.

---

3. Here, the test statistic

$$Z_c = \frac{C-\mu}{\sigma/\sqrt{n}}$$
$$\Rightarrow C = \mu + Z_c\frac{\sigma}{\sqrt{n}}$$

Now, given

$$\mu = 25,\ \sigma^2 = 9,\ or\ \sigma = 3,\ n = 49$$
$$and\ for\ \alpha = 0.05,\ Z_c = 1.645$$

So,

$$C = 25 + 1.645 \times \frac{3}{\sqrt{49}} = 25.705$$

Given, true age Θ = 27,

$$Z = \frac{\bar{x}-\theta}{\sigma/\sqrt{n}}$$
$$= \frac{25.705-27}{3/\sqrt{49}}$$
$$= -3.02$$

By using table,

$$for\ Z = -3.02,\ p = 0.0013$$

So, the probabilty of Type II Error is

$$1 - 0.0013 = 0.9987,\ or\ 99.87\%$$

---

4. From Exercise 1.3, we know that, 95% confidence interval for Θ is:

$$CI_{0.95}(\theta) = \left(\bar{x} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

Here,

$$\bar{x} = 26,\ \sigma = 3,\ n = 49,\ from\ table\ we\ get\ z_{0.975} = 1.96$$

So,

$$CI_{0.95}(\theta) = \left(26 - 1.96 \times \frac{3}{\sqrt{49}}\ ,\ 26 + 1.96 \times \frac{3}{\sqrt{49}}\right)$$
$$CI_{0.95}(\theta) = (25.16, 26.84)$$

So, the 95% time the sample mean will be between 25.16 to 26.84.
