Mohammad Javad Ranjbar

810101173

Homework 3

Statistical Inference, Fall 2022

Question #1:

$SE = \frac{\sigma}{\sqrt{n}} = \frac{90000}{\sqrt{81}} = 10000, \bar{x} = 800000$

n is smaller than 10 percent of all the housed in NY and also $n \geq 30$. Therefore, we can use the normal distribution and confidence interval.

$$\bar{x} - z^{*}SE < \mu < \bar{x} + z^{*}SE$$

a) For 98% confidence, we have $z^{*} = 2.33$ (> qnorm(0.01) [1] -2.326348)

$$800000 - 2.33 * 10000 < \mu < 800000 + 2.33 * 10000 \Rightarrow 776700 < \mu < 823300$$

b) For 95% confidence, we have $z^{*} = 1.96$ (> qnorm(0.025) [1] -1.959964)

$$800000 - 1.96 * 10000 < \mu < 800000 + 1.96 * 10000 \Rightarrow 780400 < \mu < 819600$$

c) For 90% confidence, we have $z^{*} = 1.64$ (> qnorm(0.05) [1] -1.644854)

$$800000 - 1.64 * 10000 < \mu < 800000 + 1.64 * 10000 \Rightarrow 783600 < \mu < 816400$$

d) For 50% confidence, we have $z^{*} = 0.67$(> qnorm(0.25) [1] -0.6744898)

$$800000 - 0.67 * 10000 < \mu < 800000 + 0.67 * 10000 \Rightarrow 793300 < \mu < 806700$$

e) P% of random samples of 81 houses will yield CIs that capture the true average price of houses in NY.

f) In each CIs we will be P% sure that those intervals would contain the true population mean(μ)

g) For 98% confidence we have $z^{*} = 2.58$(> qnorm(0.005) [1] -2.575829):

$$\text{Margine of error} = z^{*}SE = z^{*}\frac{\sigma}{\sqrt{n}} = 2.58\frac{90000}{\sqrt{n}} = 5000 \Rightarrow \sqrt{n} > 46.44 \Rightarrow n > 2156.6736$$
$$\Rightarrow n = 2157$$

h) $\frac{\text{Margine of error}_1}{\text{Margine of error}_2} = \frac{z^{*}\frac{\sigma}{\sqrt{n_1}}}{z^{*}\frac{\sigma}{\sqrt{n_2}}} = \frac{\sqrt{n_2}}{\sqrt{n_1}} \Rightarrow \frac{5000}{2500} = \frac{\sqrt{n_2}}{46.44} \Rightarrow n_2 = 8649$

Question #2:

1- the original research was based on high school students and comparing them with college students is wrong.
2- The null hypothesis should have an equal sign.
3- The alternative hypothesis should have a not equals or > sign.

The correct way to set up these hypotheses is:

$$H_0: x = 7 \text{ hours}$$
$$H_A: x > 7 \text{ hours}$$

Question #3:

n might be smaller than 10% of the number of children in that city but $n < 30$. Therefore, we cannot use a normal distribution for this question. We use Student's t-distribution with the below hypothesis:

$$H_0: \mu = 5\ years$$
$$H_A: \mu \neq 5\ years$$

a) $SE = \frac{s}{\sqrt{n}} = \frac{2.2}{\sqrt{20}} = 0.491935$

$$T = \frac{observation - null}{SE} = \frac{4.6 - 5}{0.491935} = 0.8131156$$

$df = n - 1 = 20 - 1 = 19$

Now we calculate the result:
```
> 2*pt(0.8131156, df = 19, lower.tail = FALSE)
[1] 0.4262241
```

Because $p - value = 0.43 > \alpha$ we fail to reject H_0.

b) Another solution is using CI. For 95% confidence, we have $t^* = 2.1$(
```
> qt(0.025,df=19)
[1] -2.093024
```
)

$$4.6 - 2.1 * 0.49 < \mu < 4.6 + 2.1 * 0.49 \Rightarrow 3.5 < \mu < 5.6$$

We are 95% confident that the average number of years to learn piano for children is 3.5 to 5.6 years.

c) In both solutions, we fail to reject H_0. In the first solution, the p-value is more than the determined α, meaning there is no considerable difference to reject H_0. In the second solution, Because 5 or the first hypothesis is in this interval we fail to reject H_0.

Question #4:

a) The number of adults(n) is less than 10% of the number of all the adults and also $n \geq 30$. Therefore, we can use normal distribution and the confidence interval.

b)

$$H_0: T = 98.6 \text{ Fahrenheit}$$
$$H_A: T > 98.6 \text{ Fahrenheit}$$

c) For 98% confidence, we have $z^* = 2.33$ (
```
> qnorm(0.01)
[1] -2.326348
```
)

$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.6824}{\sqrt{52}} = 0.09463185$

$98.2846 - 2.33 * 0.09463185 < \mu < 98.2846 + 2.33 * 0.09463185 \Rightarrow 98.064 < \mu < 98.505$

Based on the above confidential interval the study's suggested body temperature is rejected.

d) $Z = \frac{observation - null}{SE} = \frac{98.2 - 98.6}{0.09463185} = -4.226907 \Rightarrow p - value = 0$

Because $\alpha = 0.02 > p - value$ we can reject the H_0.

Question #5:

Because $np \geq 10 \Rightarrow 50 \geq 10$ and $n(1 - p) \geq 10 \Rightarrow 50 \geq 10$ we can approximate with the normal model. $\sigma = npq = 5$

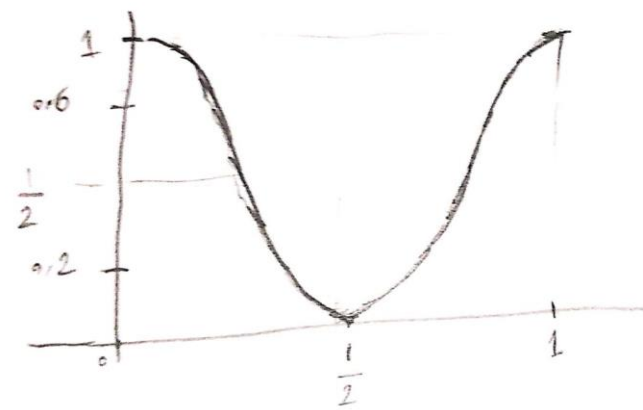
a) α is the probability of rejecting the null hypothesis when it's true.

$$|X-50|>10 \Rightarrow \{X>60 \; or \; X<40 \Rightarrow Z>\frac{60-50}{5}=2 \; or \; Z<\frac{40-50}{5}=-2 \Rightarrow |Z|>2$$

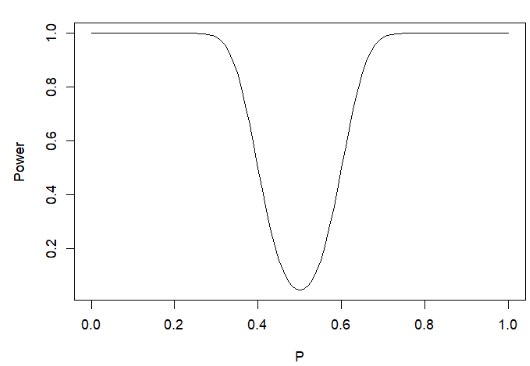
```
> 2*pnorm(-2)
[1] 0.04550026
```

In order to H_0 get rejected the $p-value<\alpha$ therefore $\alpha>0.045$

b) Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$. Now, if the actual value p is really far from the current value the probability of rejecting H_0 would be high. For instance, when $\mathrm{P}=1$ or $\mathrm{P}=0$ we have a high chance of rejecting H_0. However, when the actual value of P is close to the value suggested by $H_0(0.5)$ the chance of rejecting it is very low. For example, when P is around 0.5 the probability is close to zero. Based on the above explanation, The plot of power versus p, should have a parabola shape(U-shaped).



The accurate figure, plotted using R is as followed:



Question #6:

a) The number of samples is less than 10% of number population and also $n \geq 30$. Therefore, we can use normal distribution and the confidence interval.

$$H_0: \mu \geq 28$$
$$H_A: \mu < 28$$

$$\mathrm{SE} = \frac{\mathrm{s}}{\sqrt{\mathrm{n}}} = \frac{5.6}{\sqrt{50}} = 0.7919596$$

a) For 90% confidence we have $z^* = 1.64$ (`> qnorm(0.05)` `[1] -1.644854`)

$$25.9 - 1.64 * 0.7919596 < \mu < 25.9 + 1.64 * 0.7919596 \Rightarrow 24.60 < \mu < 27.19$$

Based on the confidence interval, $\mu \geq 28$ is not present in this interval. Therefore, the H_0 is rejected.

b) $\mu = 27$

$$\text{Power} = 1 - \text{ Type II error} = 1 - \beta = 1 - p(z \leq z_a - \frac{\mu_a - \mu}{\mathrm{SE}}) \Rightarrow \text{power} = 1 - p(z \leq 1.645 - \frac{28-27}{0.8}) = 1 - p(z \leq 0.395) = 1 - 0.6535786 = 0.3464214$$

c) No, we reject H_0. Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β.

Question #7:

$$\mu = 27 \Rightarrow Power = 1 - TypeIIerror = 1 - \beta = 1 - p\left(z \leq z_a - \frac{\mu_a - \mu}{\mathrm{SE}}\right) \Rightarrow$$

$$\text{power} = 1 - p(z \leq 1.645 - \frac{28-27}{0.8}) = 1 - p(z \leq 0.395) = 1 - 0.6535786 = 0.3464214$$

$$\mu = 26 \Rightarrow power = 1 - p\left(z \leq 1.645 - \frac{28-26}{0.8}\right) = 1 - p(z \leq -0.855) = 1 - 0.1962756 = 0.8037244$$

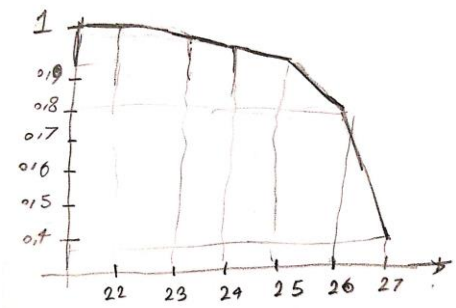
$$\mu = 25 \Rightarrow \text{power} = 1 - p(z \leq 1.645 - \frac{28-25}{0.8}) = 1 - p(z \leq -2.105) = 1 - 0.01764565 = 0.9823544$$

$$\mu = 24 \Rightarrow \text{power} = 1 - p(z \leq 1.645 - \frac{28-24}{0.8}) = 1 - p(z \leq -3.355) = 1 - 0.0003968249 = 0.9996032$$

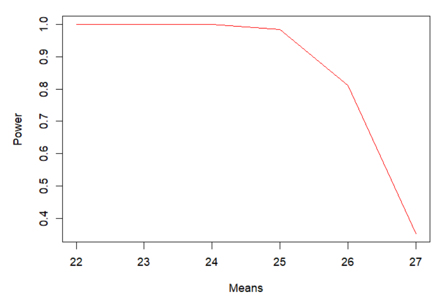
$$\mu = 23 \Rightarrow \text{power} = 1 - p(z \leq 1.645 - \frac{28-23}{0.8}) = 1 - p(z \leq -4.605) = 1 - 0.000002062329 = 0.9999979$$

$$\mu = 22 \Rightarrow \text{power} = 1 - p(z \leq 1.645 - \frac{28-22}{0.8}) = 1 - p(z \leq -5.855) = 1 - 0 = 1$$

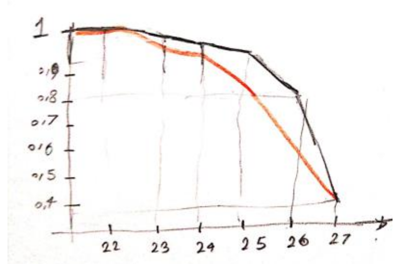
a) power of a test is the probability of correctly rejecting H0, and the probability of doing so is 1-β. As H0 gets further away from the actual mean the chance of correctly rejecting it will get higher. for example, H_0 suggests the μ=28 and when the actual mean value is 22(very far) we have a high effect size, and the probability of rejecting H_0 is very high and close to 1. However, when the actual mean value is 27(very close) we have a low effect size and the probability of rejecting H0 is very low and close to 0. The approximated figure is as followed:



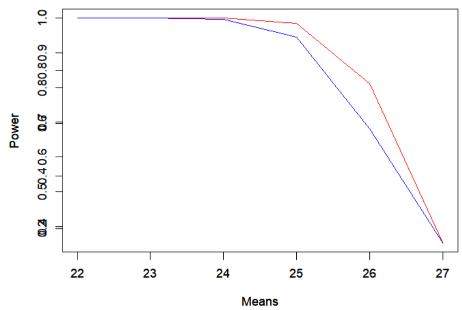
The accurate figure, plotted using R is as followed:



b) alpha=0.01. the power approaches one with a smaller slope.

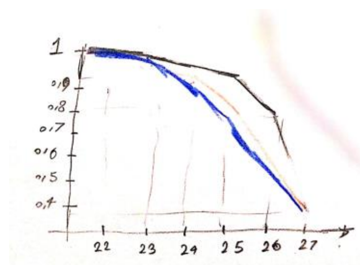


The accurate figure, plotted using R is as followed:

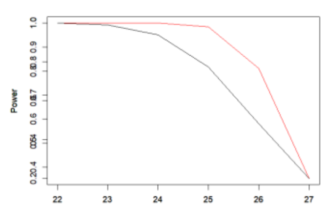



```
a<-.05
n<-50
means_list<-seq(22,27,1)
plot(means_list, (pnorm(qnorm(a,0,1,lower.tail =TRUE) -
                        (means_list-28)/(o/(sqrt(n))))),col="red",type="l",xlab= "Means",ylab = "Power")
par(new=TRUE)
a<-.01
plot(means_list, (pnorm(qnorm(a,0,1,lower.tail =TRUE) -
                        (means_list-28)/(o/(sqrt(n))))),col="blue",type="l",xlab= "Means",ylab = "Power")
```

c) With lower n=20, the power approaches one with a smaller slope.



The accurate figure, plotted using R is as followed:




```
a<-.05
n<-50
means_list<-seq(22,27,1)
plot(means_list, (pnorm(qnorm(a,0,1,lower.tail =TRUE) -
                        (means_list-28)/(o/(sqrt(n))))),col="red",type="l",xlab= "Means",ylab = "Power")
par(new=TRUE)
n<-20
plot(means_list, (pnorm(qnorm(a,0,1,lower.tail =TRUE) -
                        (means_list-28)/(o/(sqrt(n))))),col="black",type="l",xlab= "Means",ylab = "Power")
```

Question #8:

n is smaller than 10% of the number of parents but $n > 30$. Therefore, we can use a normal distribution for this question.

a) Around 97% of intervals include the real mean value.

```
> student_mean<-mean(df$child)
> intervals_with_mean<-0
> n_samples<-60
> CI<-0.97
> n_intervals<-20000
> for (x in 1:n_intervals)
+ {
+     temp_samples<-sample(df$child,n_samples)
+     temp_mean<-mean(temp_samples)
+     temp_sd<-sd(temp_samples)/sqrt(n_samples)
+     high_tresh<-temp_mean+abs(qnorm((1-CI)/2))*temp_sd
+     low_tresh<-temp_mean-abs(qnorm((1-CI)/2))*temp_sd
+     if(student_mean<high_tresh & student_mean>low_tresh)
+     {
+         intervals_with_mean<-intervals_with_mean+1
+     }
+ }
> print(intervals_with_mean/n_intervals)
[1] 0.96995
```

b) n is smaller than 10% of the number of children but $\mathrm{n} < 30$. Therefore, we cannot use a normal distribution for this question. We use Student's t-distribution with the below hypothesis:
Around 90% of intervals include the real mean value.

```
> student_mean<-mean(df$child)
> intervals_with_mean<-0
> n_samples<-10
> CI<-0.9
> n_intervals<-10000
> degree_f<-n_samples-1
> for (x in 1:n_intervals)
+ {
+     temp_samples<-sample(df$child,n_samples)
+     temp_mean<-mean(temp_samples)
+     temp_sd<-sd(temp_samples)/sqrt(n_samples)
+     high_tresh<-temp_mean+abs(qt((1-CI)/2,degree_f))*temp_sd
+     low_tresh<-temp_mean-abs(qt((1-CI)/2,degree_f))*temp_sd
+     if(student_mean<high_tresh & student_mean>low_tresh)
+     {
+         intervals_with_mean<-intervals_with_mean+1
+     }
+ }
> print(intervals_with_mean/n_intervals)
[1] 0.8942
```

c) We used CIs to test these hypotheses. H_0 is rejected and power Is equal to 1.

```
> parent_mean<-mean(df$parent)
> n_samples<-70
> alpha<-0.05
> H0<-60
> CI<-1-alpha
> temp_samples<-sample(df$parent,n_samples)
> temp_mean<-mean(temp_samples)
> temp_sd<-sd(temp_samples)/sqrt(n_samples)
> high_tresh<-temp_mean+abs(qnorm((1-CI)/2))*temp_sd
> low_tresh<-temp_mean-abs(qnorm((1-CI)/2))*temp_sd
> if(H0<high_tresh & H0>low_tresh)
+ {
+     cat("the real mean with the value of: ", parent_mean," is between", low_tresh,
+         " and ",high_tresh, ". Therefore, The H0 is not rejcted")
+ } else
+ {
+     cat("the real mean with the value of: ", parent_mean,
+         " is between", low_tresh, " and ",high_tresh, ". Therefore, The H0 is rejcted")
+ }
the real mean with the value of:  68.30819  is between 67.93216  and  68.55355 . Therefore, The H0 is rejcted> cat("th
e power is equal to:", 1-pnorm(qnorm(alpha,0,1,lower.tail =TRUE) - (parent_mean-H0)/temp_sd))
the power is equal to: 1
```

d) n is smaller than 10% of the number of parents but $\mathrm{n} < 30$. Therefore, we cannot use a normal distribution for this question. We use Student's t-distribution with the below hypothesis:
We used CIs to test these hypotheses. The power is equal to 1 and the H_0 is rejected.

```
> parent_mean<-mean(df$parent)
> n_samples<-10
> alpha<-0.05
> CI<-1-alpha
> degree_f<-n_samples-1
> H0<-60
> temp_samples<-sample(df$parent,n_samples)
> temp_mean<-mean(temp_samples)
> temp_sd<-sd(temp_samples)/sqrt(n_samples)
> high_tresh<-temp_mean+abs(qt((1-CI)/2,degree_f))*temp_sd
> low_tresh<-temp_mean-abs(qnorm((1-CI)/2,degree_f))*temp_sd
> if(H0<high_tresh & H0>low_tresh)
+ {
+     cat("the real mean with the value of: ", parent_mean," is between", low_tresh, " and ",high_tresh, ". Therefore,
 The H0 is not rejcted")
+ } else
+ {
+     cat("the real mean with the value of: ", parent_mean,
+         " is between", low_tresh, " and ",high_tresh, ". Therefore, The H0 is rejcted")
+ }
the real mean with the value of:  68.30819  is between 65.00882  and  70.01821 . Therefore, The H0 is rejcted> cat("th
e power is equal to:", 1-pt(qt(alpha,df=degree_f,lower.tail =TRUE) - (parent_mean-H0)/temp_sd,df=degree_f))
the power is equal to: 1>
```

Also, if we use p-value we get the same result.

```
> parent_mean<-mean(df$parent)
> n_samples<-10
> alpha<-0.05
> CI<-1-alpha
> degree_f<-n_samples-1
> H0<-60
> temp_samples<-sample(df$parent,n_samples)
> temp_mean<-mean(temp_samples)
> temp_sd<-sd(temp_samples)/sqrt(n_samples)
> p_value<-2*pt((temp_mean-H0)/temp_sd,degree_f)
> if(p_value>alpha)
+ {
+     cat("the real mean with the value of: ", parent_mean," is between", low_tresh, " and ",high_tresh, ". Therefore,
 The H0 is not rejcted")
+ } else
+ {
+     cat("the real mean with the value of: ", parent_mean,
+         " is between", low_tresh, " and ",high_tresh, ". Therefore, The H0 is rejcted")
+ }
the real mean with the value of:  68.30819  is between 65.00882  and  70.01821 . Therefore, The H0 is not rejcted> cat
("the power is equal to:", 1-pt(qt(alpha,df=degree_f,lower.tail =TRUE) - (parent_mean-H0)/temp_sd,df=degree_f))
the power is equal to: 1>
```

e) the H_0 is wrong and should be rejected. And this is algin with the actual value of mean which is 68. And, also both normal and T distribution work similar in this case.