MINERVA UNIVERSITY

Inferential Statistics Report

1. Introduction

As we see in daily life, there are a lot of differences between men and women. Some are identifiable, while others are more based on personalities. I have always seen tall men back in my home country, but coming to a diverse international space. I saw a lot of tall women as well. And mostly, I encountered women being taller than men. And I got confused if there is actually a difference between the height of men and women or is it just a perception, Hence the research question for my assignment - "Is there a difference between men's and women's height?".To answer this question, we will use a difference of means test and a confidence interval will be created.

2. Dataset

The dataset used for this assignment was featured in Openintro Statistics Textbook (CDC,2019). It is a sample of a youth risk behavior surveillance system. The data set has 100 respondents and various questions were asked related to their physical strength, sleep, height, weight, gender, and driving habits.
However, for this paper, we will focus only on two variables namely Gender, which is considered a qualitative nominal variable. It does not have any order and is just based on categories - male or a female. The other variable is heights which is considered quantitative continuous as the heights are numbers with decimals and have an indefinite number of values between two numbers.[1]

3. Analysis

Hypothesis

To answer our main question about whether there is a difference between the height of men and women, we will perform a difference of means significance test. We will consider 2 groups, men and women, which can be seen in appendix A. And by default, define our significance level = 0.05, which shows the probability of rejecting the null hypothesis when it is true and committing a type 1 error. So here, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.
Before we proceed, let us properly define our hypothesis.

[1] #variables: I distinguished between the two different type of variable which were important for my research question.

Null hypothesis (H0) : There is no difference of means between the height of female and males. ($\bar{x}_2$-$\bar{x}_1$= 0)

Alternate hypothesis (HA) : There is a difference of means between the heights of females and males. ($\bar{x}_2 \neq \bar{x}_1$)

Note: It is a two-tailed test because the difference of means can be either way. It can be that men are taller than women or vice versa.

Summary Statistics

The summary statistics have been found using pandas in python. Before we go on further. We need to have descriptive statistics. Appendix A shows the relevant coding on the calculation of various statistics, and table 1 below indicates only relevant statistics for our significance test.
The histograms in Figures 1 and 2, which can be referred to in Appendix B, give us precise details about our statistics. We can see a clear difference between the mean values of heights of men and women. Further considering the median and mode are also quite different. However, the standard deviation and range are quite close, showing that the data does not vary much. Figure 3 shows a clear comparison graphically. And looking at the graphs, they do not look much skewed, except one outlier in male heights; other than that, they are more oriented towards a normal distribution.[2]

Table 1: Summary statistics for the heights of the two sample groups: female and male.

	Female	**Male**
Count	n_1 =51	n_2= 48
Mean	$\bar{x}_1$= 1.64	$\bar{x}_2$ =1.74
Median	1.65	1.75
Mode	1.60	1.78
Standard Deviation	s_1=0.07	s_2 =0.07

[2] #descriptivestats: I calculated mean,median,mode, range to better interpret my data and used them later to conduct significance tests.

Range	0.30	0.37

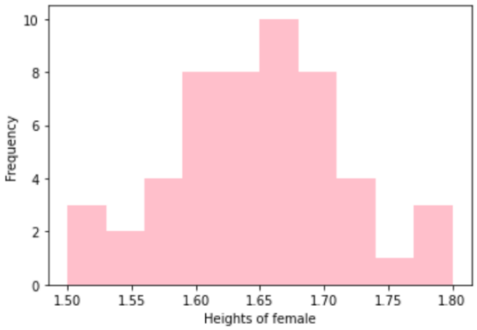


Fig 1. Histogram showing the heights of female

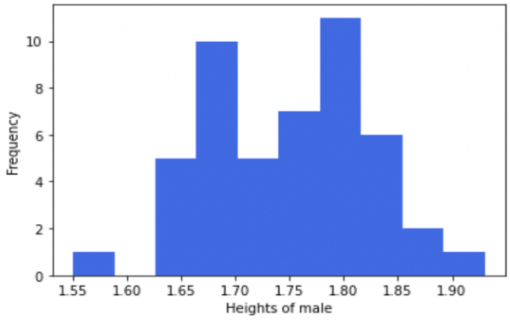


Fig 2. Histogram showing the heights of male

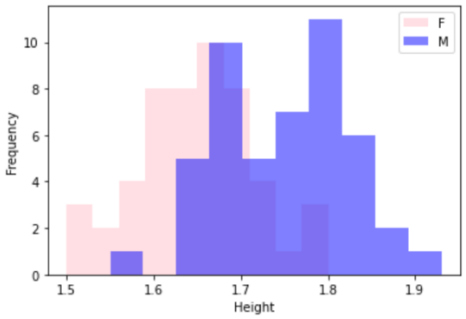


Fig 3. Histogram showing comparison between heights of female and males[3]

Conditions for inference:

Even though we have a sample size greater than 30 (n1,n2>30) but we still don't know the population standard deviation and we are considering the standard error using sample standard deviation hence we will be using a ***t-distribution*** as our sampling distribution.
For this, we will consider 3 conditions:

- **Independent:**
  Following the 10% rule and considering the whole population of more than a billion, we can clearly see that the sample data is less than 10% which makes our data to be independent.

[3] #dataviz: I generated a histogram to examine the distributions, and look at the skewness.

- **Normal Distribution:**
  As the distribution has a sample size of greater than 30 (n>30) and the values for mean, median and mode are quite close, we can conclude that the distribution is similar to a normal distribution.

- **Random Sample:**
  As it is nowhere mentioned, if the dataset is random or not, we will consider this as a limitation and just assume the sample to be random.

  As the conditions for inference are met, we will use the t distribution and apply central limit theorem.[4]

Difference of Means Test

Now, we will conduct a difference of mean significance test and for this, first, we need to calculate the p-value. And before calculating the p-value, we need to find the T-score which can be found using the formula $T = (\bar{x}_2 - \bar{x}_1) \div \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where the denominator is the standard error and the degrees of freedom is considered the minimum of one of the sample size subtracting 1. df =47. See Appendix C for calculations in python.
We get a T score = 6.71 and the p-value as 2.25×10^{-8}<0.05, and the value here is too small, which proves that there is convincing evidence to reject the null hypothesis and favor the alternate hypothesis.[5]

Let us also consider the practical significance, for which we will measure the hedges g because it helps us get more accurate results and eliminate the upward bias that we get using cohen's D. To find the value for hedges g, we first need to calculate pooled standard deviation $sp = \sqrt{\frac{(n-1)s_1^2 + (n-2)\, s_2^2}{n_1 + n_2 - 2}}$. Check Appendix C for calculations. We get hedges g = 1.34 which means that there is a large effect size on the data means there is a big difference.[6]

Confidence Intervals

---

[4] #distributions: I identified and justified a kind of distribution of my sampling distributions by checking all conditions carefully.
[5] #probability: I calculated the p value which states the conditional probability- that is it is the probability of getting the data observed or more extreme data if the null hypothesis is true.
[6] #significance: I calculated whether my dataset was statistically significant or not and then interpreted the results. Additionally, I also calculated the effect size for looking at the practical significance.

Now we will calculate the confidence interval for the mean of heights of men and women. We choose a 95% confidence interval as it will give us plausible range values. As we have discussed all the conditions for inference met. So we can use the t distribution and apply the Central Limit Theorem. Now let us calculate the interval using the formula $CI = \bar{x} \pm t * (s/\sqrt{n})$. Note that 95% confidence interval is a range of values that you can be 95% certain the interval contains the true mean of the population.
For calculations consider appendix D. Below are the confidence intervals for men and women.

- Female : [1.62, 1.65]
- Male: [1.72, 1.75]

We can clearly see that both the confidence intervals for the group do not overlap each other which states that there is a difference in their mean values and proves that we reject the null hypothesis and supports our difference of means test.[7]

4. Results and Conclusions

From all the previous calculations done above, We get a T score = 6.71, p-value = 2.25×10^{-8} and hedges g= 1.34 . All these values show that there is enough evidence to infer that there is a difference between the mean of heights of females and males. Hence, we can reject the null hypothesis. Also, the p-value shows that the difference is statistically significant, and as well as the hedges g shows the effect size which is large, conveying that the difference is practically significant as well. Also, the t score is large which makes the difference more certain.
The confidence interval for males [1.72, 1.75] and females [1.62, 1.65] are not overlapping, which further supports our inference that we can reject the null hypothesis which states that there is no difference between the means of females and males. And we can now answer the research question that yes, there is a difference between the heights of males and females. Also, looking at the means we can also conclude that males are taller than females.

Our inference was inductive because we took sample sizes and then discussed the population. So we went from specific to more general and the conclusion was not contained within the scope of the premises. And as we used confidence intervals, significance levels this implies that we are not 100% sure but strongly believe in our conclusions. Our inductive argument is strong and reliable because we justified all the

[7] #confidenceinterval:I calculated and constructed an appropriate confidence interval with clearly stated steps. And I interpreted exactly the meaning of the confidence interval and applied this into following tests to support my significance test.

three conditions for the t- distribution and applied the central limit theorem. The effect size was also large which also convinces that the inference was reliable.[8]

Word Count: 1498 words

References

CDC.(2019). YRBSS Data and documentation.
https://www.cdc.gov/healthyyouth/data/yrbs/data.html

Reflection

I once got a 1 in distributions where I did not answer the question at all, we had to calculate the p value using the t distribution and I misinterpreted the question because of my unclear concepts related to the type of distribution but the feedback on my poll helped me out a lot. The feedback described a clear difference between a z distribution and t distribution and and when to use which one. For example, a t distribution is used when we do not know the population standard deviation and the sample size is small and the feedback referred back to the book for revising the concept. I did that, which was super helpful for this assignment, now I clearly know when to use a t-distribution.

Appendix

The full Jupyter notebook file and the data can be accessed in the zipped folder submitted as a secondary file.

Appendix A: Dividing the dataset into subgroups and analyzing

[8] #induction: I identified why statistical inference are inductive in nature and discussed about the strength and reliability of the inference.

In [120]:
```
#dividing the dataset into subgroups - male and female
height_of_female = df['height'][df['gender']=='female']
height_of_male = df['height'][df['gender']=='male']
print(height_of_female.head(10))
print(height_of_male.head(10))
```

```
0     1.50
5     1.57
11    1.68
12    1.67
14    1.60
15    1.60
16    1.68
18    1.73
19    1.63
20    1.63
Name: height, dtype: float64
1     1.78
2     1.75
3     1.68
4     1.70
6     1.78
7     1.63
8     1.63
9     1.83
10    1.69
13    1.78
Name: height, dtype: float64
```

In [21]:
```
#print the summary statistics: mean,count,SD for height of female
height_of_female.describe()
```

Out[21]:
```
count    51.000000
mean      1.648431
std       0.070011
min       1.500000
25%       1.600000
50%       1.650000
75%       1.700000
max       1.800000
Name: height, dtype: float64
```

In [25]:
```
#print the summary statistics: mean,count,SD for height of male
height_of_male.describe()
```

Out[25]:
```
count    48.000000
mean      1.747917
std       0.076990
min       1.550000
25%       1.700000
50%       1.750000
75%       1.785000
max       1.930000
Name: height, dtype: float64
```

In [133]:
```
#import library
import statistics
print("statistics for female heights")
print("median=", statistics.median(height_of_female))
print("mode=",statistics.mode(height_of_female))
print("range=",(max(height_of_female)-min(height_of_female)))
```

```
statistics for female heights
median= 1.65
mode= 1.6
range= 0.30000000000000004
```

In [131]:

```
#using library to get relevant statistics
print("statistics for male heights")
print("median=",statistics.median(height_of_male))
print("mode=",statistics.mode(height_of_male))
print("range=",(max(height_of_male)-min(height_of_male)))
```

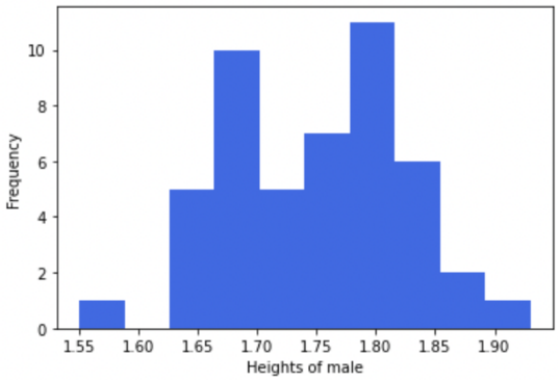
```
statistics for male heights
median= 1.75
mode= 1.78
range= 0.3799999999999999
```

Appendix B: Visualizing Data

In [47]:

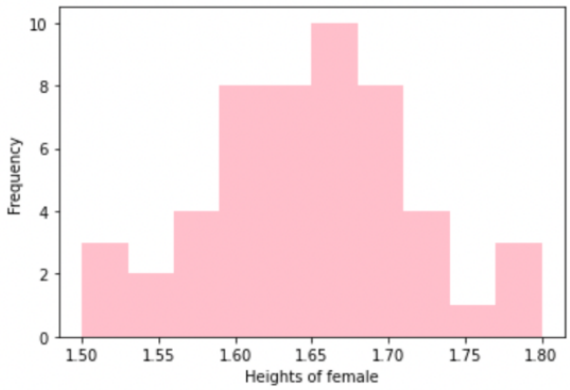
```
#plotting a histogram using the library metplotlib for males height
import matplotlib.pyplot as plt
plt.hist(height_of_male, color =  'royalblue')
plt.xlabel("Heights of male")
plt.ylabel("Frequency")
```

Out[47]: Text(0, 0.5, 'Frequency')



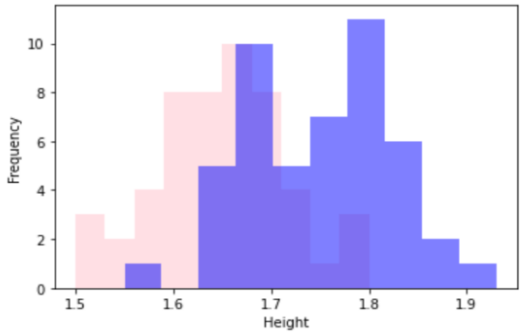
In [48]:
```
#plotting a histogram for females height
plt.hist(height_of_female, color = 'pink')
plt.xlabel("Heights of female")
plt.ylabel("Frequency")
```

Out[48]: Text(0, 0.5, 'Frequency')



In [111]:
```
#plotting a histogram with both variables to show comparison
plt.hist(height_of_female, color = 'pink',alpha= 0.5)
plt.hist(height_of_male, color = 'blue', alpha = 0.5)
plt.xlabel("Height")
plt.ylabel("Frequency")
```

Out[111]: Text(0, 0.5, 'Frequency')



Appendix C: Difference of means test

```
#this code has been used from class Session 13.2 https://sle-collaboration.minervaproject.com/
#?url=https%3A//sle-authoring.minervaproject.com/api/v1/worksheets/c942ce85-aaca-4b51-b9de-d14
#b96d31616/&userId=11969&name=Sana+Mehta&avatar=https%3A//s3.amazonaws.com/picasso.fixtures/
#Sana_Mehta_11969_2021-08-30T23%3A44%3A59.323Z&noPresence=1&readOnly=1&isInstructor=0&signatu
#re=25b3277b489db7d96e36710ca965f652f7c937e4e4d4b95327d6db62d9c5df7b

import numpy as np
from scipy import stats

#defining a function
def difference_of_means_test(data1,data2,tails):

#finding the count of data
    n1 = len(data1)
    n2 = len(data2)

#finding the mean of the two groups
    x1 = np.mean(data1)
    x2 = np.mean(data2)

#finding the standard deviation of the dataset
    s1 = np.std(data1,ddof=1) #bessels correction using n-1 as denominator
    s2 = np.std(data2,ddof=1)

#finding the standard error
    SE = np.sqrt((s1**2/n1 + s2**2/n2))


#finding the t score using standard eror
    Tscore = ((x2-x1)/SE)

#calculating degrees of freedom given the conservative estimate from openintro
    df = min(n1,n2)-1

#calculating p value using stats library
    pvalue = tails*stats.t.cdf(-Tscore,df)

#calculating pooled standard deviation using the formula
    SDpooled = np.sqrt((s1**2*(n1-1) + s2**2*(n2-1))/(n1+n2-2))
```

```
#calculating cohens d using the formula
    Cohensd = (x2-x1)/SDpooled

#calculating hedges g from cohens d
    hedgesg = Cohensd*((4*(n1+n2)-9)-(3))/(4*(n1+n2)-9)

#printing all the values calculated using the function
    print('t=',Tscore)
    print('p =', pvalue)
    print('d =',Cohensd)
    print('g=',hedgesg)

#calling the function
difference_of_means_test(height_of_female,height_of_male,2)
```

```
t= 6.713450954353862
p = 2.22514300865667e-08
d = 1.3539963495503458
g= 1.3435002538173975
```

Appendix D: Confidence Intervals

In [121]:

```
#importing relevant packages
#defining a function for calculating confidence interval
from scipy import stats
import numpy as np
def confidence_interval(mean,SD,n,level):
    df = n-1
    t = stats.t.ppf(1-(1-level)/2,df) #finding t score value where 1- (1-level)/2 shows
    #sum of two tails and subtrating it from 1 because to give the correct side of the area on the graph
    interval=((mean-t*SE,mean+t*SE))
    print("confidence_interval=",interval)

confidence_interval(1.64,0.7,51,0.95) #for heights of female
confidence_interval(1.74,0.7,48,0.95) #for heights of male
```

```
confidence_interval= (1.6203121827675093, 1.6596878172324905)
confidence_interval= (1.7202809988026393, 1.7597190011973607)
```