```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
from scipy import stats
```

Analyzing the average heights of NBA *Players*

```
df2 = pd.read_csv('players.csv')
df2.head()
```

	Name	Games Played	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB
0	AJ Price	26	324	133	51	137	37.2	15	57	26.3	16	24	66.7	6	26
1	Aaron Brooks	82	1885	954	344	817	42.1	121	313	38.7	145	174	83.3	32	134
2	Aaron Gordon	47	797	243	93	208	44.7	13	48	27.1	44	61	72.1	46	123
3	Adreian Payne	32	740	213	91	220	41.4	1	9	11.1	30	46	65.2	48	114
4	Al Horford	76	2318	1156	519	965	53.8	11	36	30.6	107	141	75.9	131	413

```
df2.shape
```

```
(490, 34)
```

Hypothesis Testing

One Sample Significance Test for Mean is extremely similar to that for Proportion. We will go through almost an identical process.

The hypotheses are defined as follows:

- **Null Hypothesis**: The average height of an NBA player is 200.66 cm.
- **Alternate Hypothesis**: The average height of an NBA player is not 200.66 cm.

Significance Level, α is at 0.05. Assuming Null Hypothesis to be true.

```
h0_mean = 200.66   #google search
```

```
  h1_mean = df2['Height'].mean()                      #z=(x-mu)/sig/sqrt(n)
  h1_mean                                              #z=(x-mu)/sigma
```

```
    197.44075829383885
```

```
sigma = df2['Height'].std()/np.sqrt(len(df2))
sigma
```

```
    0.3948442447237618
```

```
z = (h1_mean - h0_mean)/sigma
z
```

```
    -8.15319394718129
```

```
#p_val = (1 - stats.norm.cdf(abs(z))) #ONE TAIL  I.E LOWER TAIL PART
p_val = (1 - stats.norm.cdf(abs(z)))*2   #TWO TAIL            #pval or prob value
p_val
```

```
    4.440892098500626e-16
```

The p value obtained is much lesser than the significance level α. We therefore reject the null hypothesis and accept the alternate hypothesis (the negation). We can therefore arrive at the following conclusion from this analysis:

The average height of NBA Players is NOT 6'7".

Analyzing DEPRESSION in India by Gender

Are men as likely to commit suicide as women?

This is the question we will attempt at answering in this section. To answer this question, we will use suicide statistics shared by the National Crime Records Bureau (NCRB), Govt of India. To perform this analysis, we need to know the sex ratio in India. The Census 2011 report states that there are 940 females for every 1000 males in India.

Let p denote the fraction of women in India.

H0:MEN AND WOMEN ARE EQUALLY LIKELY TO DEPRESS (NULL)

H1:MEN AND WOMEN ARE NOT EQUALLY LIKELY TO DEPRESS (ALTERNATE)

```
p = 940/(940+1000)   # Female population proportion
p
```

```
0.4845360824742268
```

```
df = pd.read_excel('Suicides.xlsx')
df.head()
```

	State	Year	Type_code	Type	Gender	Age_group	Total
0	A & N Islands	2001	Causes	Illness (Aids/STD)	Female	0-14	0
1	A & N Islands	2001	Causes	Bankruptcy or Sudden change in Economic	Female	0-14	0
2	A & N Islands	2001	Causes	Cancellation/Non-Settlement of Marriage	Female	0-14	0
	A & N			Physical Abuse (Rape/Incest			

```
df.shape
```

```
(237519, 7)
```

```
df['Gender'].value_counts()
```

```
Male      118879
Female    118640
Name: Gender, dtype: int64
```

Step 2: Decide on the Statsitical Test

We will be using the One Sample Z-Test here.

Step 3: Compute the p-value

```
h0_prop = p
h0_prop
```

```
0.4845360824742268
```

```
h1_prop = df['Gender'].value_counts()['Female']/len(df)
h1_prop
```

```
0.49949688235467476
```

```
sigma_prop = np.sqrt((h0_prop * (1 - h0_prop))/len(df))
sigma_prop
```

```
0.0010254465276083747
```

```
z = (h1_prop - h0_prop)/sigma_prop
z
```

```
14.589546580591277
```

```
p_val = (1-stats.norm.cdf(z))*2       # pval<aplha
p_val
```

```
0.0
```

Analyzing Literacy Rates

Two Sample test

```
from google.colab import files
uploaded=files.upload()
```

```
Choose Files No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving cities.csv to cities (1).csv
```

```
df3 = pd.read_csv('cities.csv')
df3.head()
```

	name_of_city	state_code	state_name	dist_code	population_total	populati
0	Abohar	3	PUNJAB	9	145238	
1	Achalpur	27	MAHARASHTRA	7	112293	
2	Adilabad	28	ANDHRA PRADESH	1	117388	
3	Adityapur	20	JHARKHAND	24	173988	

```
df3['state_name'].value_counts()
```

```
UTTAR PRADESH                63
WEST BENGAL                  61
MAHARASHTRA                  43
ANDHRA PRADESH               42
MADHYA PRADESH               32
TAMIL NADU                   32
GUJARAT                      29
RAJASTHAN                    29
BIHAR                        26
KARNATAKA                    26
HARYANA                      20
PUNJAB                       16
NCT OF DELHI                 15
ORISSA                       10
JHARKHAND                    10
CHHATTISGARH                  9
KERALA                        7
UTTARAKHAND                   6
ASSAM                         4
JAMMU & KASHMIR               3
PUDUCHERRY                    2
MANIPUR                       1
MEGHALAYA                     1
ANDAMAN & NICOBAR ISLANDS     1
CHANDIGARH                    1
NAGALAND                      1
TRIPURA                       1
MIZORAM                       1
HIMACHAL PRADESH              1
Name: state_name, dtype: int64
```

```
punjab = df3[df3['state_name'] == 'PUNJAB']['effective_literacy_rate_total']
delhi = df3[df3['state_name'] == 'NCT OF DELHI']['effective_literacy_rate_total']
```

```
punjab_mean = punjab.mean()
punjab_std = punjab.std()
```

```
punjab_mean, punjab_std
```

```
(83.44062499999998, 5.381935796408821)
```

```
delhi_mean = delhi.mean()
delhi_std = delhi.std()

delhi_mean, delhi_std
```

```
(83.658, 4.6569551671206195)
```

From the above calculations, it can be seen that the mean and the standard deviations of Punjab and Delhi literacy rates differ slightly. The next step is to determine if this difference is a statistically significant one.

For hypothesis testing, the following are defined:

- **Null Hypothesis:** The true mean literacy rate for Punjab and Delhi are the same.
- **Alternate Hypothesis:** The true mean literacy rate for Punjab and Delhi are not the same.

The threshold value of α is assumed to be 0.05. Assuming Null Hypothesis is true.

```
h0_mean = 0
mean_diff = delhi_mean - punjab_mean
sigma_diff = np.sqrt((delhi_std**2)/len(delhi)  + (punjab_std**2)/len(punjab))
mean_diff, sigma_diff
```

```
(0.2173750000000183, 1.8044784525904138)
```

Since we are dealing with sample sizes less than 30, using the t-statistic will be more appropriate. To use student's t though, we need to calculate the degree of freedom. This is done as follows:

```
deg = (((delhi_std**2)/len(delhi)  + (punjab_std**2)/len(punjab)) ** 2) / ((((delhi_std**2)/
deg
```

```
28.82681788840003
```

```
z = (mean_diff - h0_mean) / sigma_diff
z
```

```
0.12046417051307332
```

```
p = (1-stats.t.cdf(z, deg))*2
p
```

```
0.904951180450877
```

The value of p obtained here is much higher than the significance level α. Therefore, we cannot reject the null hypothesis. It stands.

The true mean literacy rate for Punjab and Delhi are the same.