Task 11.1

Load the data from 'Heart.csv', a dataset collecting a sample of medical conditions about persons from US (each row describes a person, each column describes whether the persons have or not some condition. E.g., 'AHD = true' means that that person has a heart disease).

In [1]:
```
# import the necessary stuff
import pandas as pd
import numpy as np
import statsmodels.api as sm
import scipy.stats.distributions as dist
from scipy import stats
```

In [2]:
```
# load the dataset, and check whether the loading was successful
database = pd.read_csv('Heart.csv')
database.head()
```

Out[2]:

	Unnamed: 0	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAng	Oldpeak	Slope
0	1	63	1	typical	145	233	1	2	150	0	2.3	3
1	2	67	1	asymptomatic	160	286	0	2	108	1	1.5	2
2	3	67	1	asymptomatic	120	229	0	2	129	1	2.6	2
3	4	37	1	nonanginal	130	250	0	0	187	0	3.5	3
4	5	41	0	nontypical	130	204	0	2	172	0	1.4	1

Task 11.2

Assume that the proportion of the population in Ireland that havs heart disease is 42%. Create a statistical test that decides, using the just loaded dataset, whether there are more people suffering from heart disease in the US than in Ireland.

In [3]:
```
# setting up the notation:
#
# null hypothesis  = H0 = "the proportion of US population that has AHD is <= 0.42"
# alternative hyp. = H1 = "the proportion of US population that has AHD is >  0.42"
```

In [4]:
```
# compute the empirical proportion in the dataset
empirical_p = len(database[database['AHD'] == 'Yes']) / len(database)

# debug
print('proportion of people in the dataset that has AHD = {}'.format(empirical_p))
```

```
proportion of people in the dataset that has AHD = 0.45874587458745875
```

Important consideration, from theoretical perspectives: even if the empirical proportion is ~0.46, we should be careful and do not say immediately "H1 is true", since we need to check whether the deviation from 0.42 is *statistically significant*. E.g., if we were having a database of only 1 person, and

that person has AHD, just by looking at the empirical mean one would say that everybody in US has heart problems.

This means there is the need for taking into account the amount of information that there is in the dataset, and thus there is the need for doing statistical hypothesis testing.

Continue thus with setting up a p-test as in the tasks below.

Task 11.2.1

As a first step (this should always be the first step, by the way), decide which significance level the test should be (a typical choice is 5%, i.e., 0.05). Remember that selecting a significance level of X% means that there is a probability of X% of rejecting H0 under the assumption that H0 is true (i.e., were H0 true, we would have X% of chances of saying 'H0 false', a type I error). Remember also that decreasing the probability of type I errors increases though the probability of type II errors.

See also https://en.wikipedia.org/wiki/Statistical_significance and https://en.wikipedia.org/wiki/Type_I_and_type_II_errors for more information.

In [5]:
```
significance_level = 0.05
```

Task 11.2.2

Setup a single population proportion test, consisting of:

1. computing the standard score (i.e., the estimated number of standard deviations by which the estimated proportion differs from the hypothesized one) as $$

z

\frac {\text{estimated proportion} - \text{least favorable proportion in H0}} {\text{standard error of the estimate}}

$$where the estimated proportion is the empirical mean estim$$
$$the least favorable proportion in H0 is that proportion among all the ones in H0 that makes the a$$

\text{standard error of the estimate}

\sqrt {

```
\frac
{
    \text{least favorable proportion in H0}
    \cdot
    ( \text{1 - least favorable proportion in H0} )
}
{n}
```

} $$ with n the number of samples in the dataset. (See also https://en.wikipedia.org/wiki/Standard_score)

2. computing a p value, i.e., the probability of obtaining a z score at least as extreme given that the null hypothesis is true, implicitly assuming that, because of the central limit theorem, this statistics is approximately normally distributed. Hint: think at what p is from a graphical perspective (i.e., a probability, and thus an area), at the fact that we are assuming z to be normal, and considering that we have been measuring a specific z. See also https://en.wikipedia.org/wiki/Z-test.

Quite more info and mathematical details in https://courses.lumenlearning.com/suny-wmopen-concepts-statistics/chapter/hypothesis-test-for-a-population-proportion-1-of-3/ and following pages, even if this may be an overkill.

In [6]:
```
# setting up the least favorable proportion in H0
lfp = 0.42

# estimating the standard deviation
standard_deviation = np.sqrt((lfp*(1-lfp))/len(database))

# computing the z-score
z = (empirical_p - lfp)/standard_deviation

# computing the p value associated to the z-score
# hint: this is a double-tailed thing

p_value = 2 * dist.norm.cdf(-z)

# debug
print('value of the s-dev:   {}'.format(standard_deviation))
print('value of the z-score: {}'.format(z))
print('p value:              {}'.format(p_value))
```

```
value of the s-dev:   0.028354195386919447
value of the z-score: 1.3664952949196816
p value:              0.1717835566635938
```

Task 11.2.3

Draw some conclusions from the computed p-value and selected significance level.

The p-value is bigger than the significance level 0.05 selected before. So, we cannot reject H0 -- meaning that we cannot conclude that there is a significant difference in the proportions of populations having heart diseases in Ireland and the US.

At the same time the p value is not very big, so somehow the conclusion is not very strong.

Task 11.3

Create a statistical test that decides, using the loaded dataset, whether there is any statistical difference between the population proportion of males and females having heart diseases in UK.

Do virtually the same statistical steps as before, with the only change that the standard deviation for the estimates is $$

```
\text{standard error of the estimate}
=
\sqrt
{
    \text{estimated total proportion assuming H0}
    \cdot
    ( \text{1 - estimated total proportion assuming H0} )
    \cdot
    \left(
        \frac{1}{n_m} + \frac{1}{n_f}
    \right)
}
```

$$

In [7]:
```
# setting up the notation:
#
# null hypothesis  = H0 = "the proportions of US males and females that have AHD are th
# alternative hyp. = H1 = "the proportions of US males and females that have AHD are di
```

In [8]:
```
# fix the database so that the values are more readable
database['Gender'] = database.Sex.replace({1: "Male", 0: "Female"})

# extract another database that has only two columns and rows
# showing only population proportions and population totals
database2 = database.groupby("Gender")['AHD'].agg([lambda z: np.mean(z=='Yes'), "size"]
database2.columns = ["HeartDisease", 'Total']

# for readability
empirical_p_f = database2.HeartDisease.Female
empirical_p_m = database2.HeartDisease.Male
n_f           = database2.Total.Female
n_m           = database2.Total.Male
#
# note that this stays the same as before
empirical_p   = len(database[database['AHD'] == 'Yes']) / len(database)

# debug
print('pf = {}'.format(empirical_p_f))
print('pm = {}'.format(empirical_p_m))
print('nf = {}'.format(n_f))
print('nm = {}'.format(n_m))
print('p  = {}'.format(empirical_p))
database2
```

```
pf = 0.25773195876288657
pm = 0.5533980582524272
nf = 97
nm = 206
p  = 0.45874587458745875
```

Out[8]:

Gender	HeartDisease	Total

Gender	HeartDisease	Total
Female	0.257732	97
Male	0.553398	206

In [9]:
```
# estimating the standard deviation
standard_deviation = np.sqrt(empirical_p*(1-empirical_p)*(1/n_m + 1/n_f))

# computing the z-score
z = (abs(empirical_p_m - empirical_p_f))/standard_deviation

# computing the p value associated to the z-score
p_value =  2 * dist.norm.cdf(-z)

# debug
print('value of the s-dev:   {}'.format(standard_deviation))
print('value of the z-score: {}'.format(z))
print('p value:              {}'.format(p_value))
```

```
value of the s-dev:   0.0613604495249707
value of the z-score: 4.818512605081534
p value:              1.4463238972316502e-06
```

Task 11.3.3

Draw some conclusions from the computed p-value and selected significance level.

The p-value is much smaller than the significance level 0.05 selected before. So, we can safely reject H0 -- meaning that we can conclude that there is a significant difference in the proportions of male and female populations having heart diseases in US.