Hypothesis Testing

By

Dr Shaik A Qadeer

Professor MJCET

Statistical inference terminology

- Inferential statistics: It is a process of estimating population parameters from samples
- Parameters: A characteristics of population(mean, SD etc)
- Sampling Error: The amount of error in the estimation of a population parameters that is based on sample statistics

Central limit theorem

- Let $S_{1,}$ $S_{2,\ldots}$ S_k be samples of size n drawn from an independent and identically distributed population with mean μ and standard deviation σ.
- Let $X_{1,}$ $X_{2,\ldots}$ X_k be the sample means.
- “According to the CLT, the distribution of $X_{1,}$ $X_{2,\ldots}$ X_k follows a normal distribution with mean μ and standard deviation of $\sigma/\sqrt{n}$. That is, the sampling distribution of mean will follow a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.”

Hypothesis Testing

- Its an example of inferential statistics
- Hypothesis is a claim
- Hypothesis testing is to either reject or retain a null hypothesis using data
- Hypothesis testing consists of two complementary statements
    1. Null Hypothesis (H_0)– an existing belief
    2. Alternate Hypothesis (H_A) – what we intend to establish with new evidences

Hypothesis Testing..

- Types of Hypothesis testing: Parametric and non-parametric

  1. Parametric tests: They make use population parameters such as mean, standard deviations etc. Example: Z-test, T-test, ANOVA etc.

  2.Non- Parametric tests: They make use data distribution to comment on the claim. Example Chi-Square etc.

- Few examples of the null hypothesis are as follows:

  1. Children who drink the health drink Complan are likely to grow taller

  2. Women use camera phone more than men (Freier, 2016)

  3. Vegetarians miss few flights (Siegel, 2016)

  4. Smokers are better sales people

Hypothesis Testing..

- The steps for hypothesis tests are as follows:

  1. Define null and alternate hypothesis. Hypothesis is described using a population parameter.

  2. Identify the test statistic to be used for testing the validity of the null hypothesis

  3. Decide the criteria for rejection and retention of null hypothesis. This is called significance value.

  4. Calculate the p-value (probability value), which is the conditional probability of observing the test statistic value when the null hypothesis is true.

  5. Take the decision to reject or not

Hypothesis Testing..

- For all the following examples, we will use the following notations:
  1. μ - population mean
  2. σ - population standard deviation
  3. X - sample mean
  4. S - sample standard deviation
  5. n - sample size

Z-test(Parametric HT)

- Z-test is used when:

  1. We need to test the value of **population mean**, given that population variance is known. And mean and standard deviations are given

  2. The population is a **normal distribution** and the population variance is known

  3. If the sample size is large and the population variance is known then normal distribution can be relaxed. That is, the assumption of normal distribution can be relaxed for large samples (n>30)

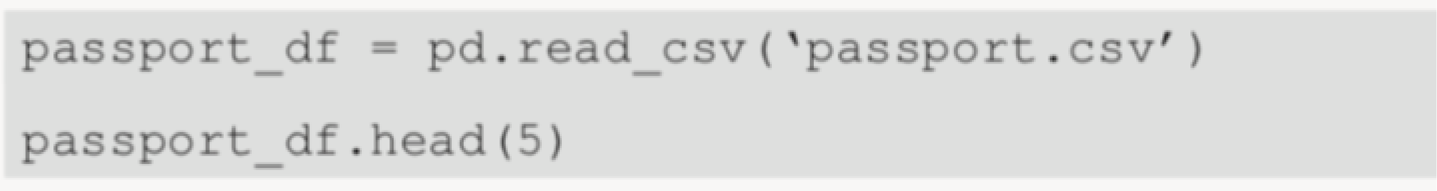
- Z- statistic is calculated as $Z = \frac{X - \mu}{\sigma / \sqrt{n}}$

Z-test(Parametric HT): Example

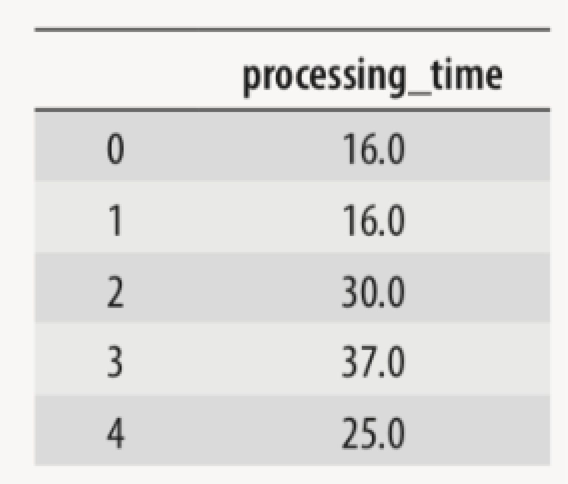
- Question: A passport office claims that the passport applications are processed within 30days of submitting the application form and all necessary documents. The file *passport.csv* contains processing time of 40 passport applicants. The population standard deviation of the processing time is 12.5 days. Conduct a hypothesis test at significance level ∝ =0.05 to verify the claim made by the passport office.
- Solution: In this case, the population mean (claim made by passport office=30days) and standard deviation(=12.5days) are known. The dataset in *passport.csv* contains observations of actual processing time of 40 passports. We can calculate the mean of these observations and calculate Zstatistic. If calculated mean with Z test is greater than the given ∝, then it can be concluded that the processing time is less than 30 as claimed by passport office.

Z-test(Parametric HT): Solution

- Load the data and display first 5 records from *passport.csv*.



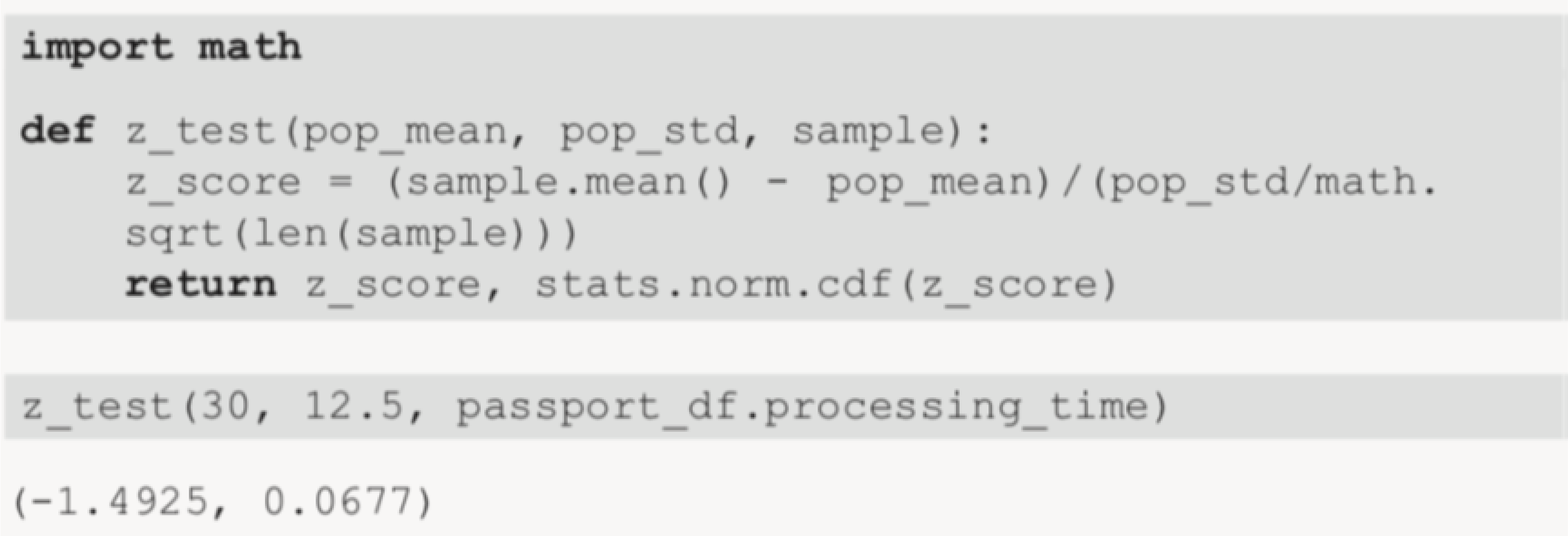
```
passport_df = pd.read_csv(‘passport.csv’)

passport_df.head(5)
```



	processing_time
0	16.0
1	16.0
2	30.0
3	37.0
4	25.0

Z-test(Parametric HT): Solution

- Conducting Z-test for the above hypothesis test



```
import math

def z_test(pop_mean, pop_std, sample):
    z_score = (sample.mean() - pop_mean)/(pop_std/math.
    sqrt(len(sample)))
    return z_score, stats.norm.cdf(z_score)
```

```
z_test(30, 12.5, passport_df.processing_time)
```

```
(-1.4925, 0.0677)
```

- The first value of the result is Z-statistic value or Z-score and second value is the corresponding p-value.

Z-test(Parametric HT): Solution Result:

- As the p-value is more than 0.05,
- Since 6.77% is greater than the significance value 5%, there is not enough evidence to reject null hypothesis.
- Hence, the null hypothesis is retained.

.

T-test(Parametric HT):

- There are 3 types of T-test: One-Sample, Two-Sample and Paired Sample
- In T-Test, poplation mean is given, but standard deviation is not given.

One Sample T-test(Parametric HT):

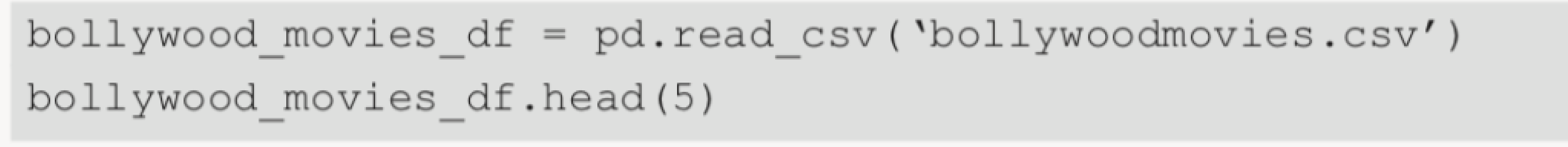
- It is used if one population mean is given, but SD is missing

One Sample T-test(Parametric HT):Example

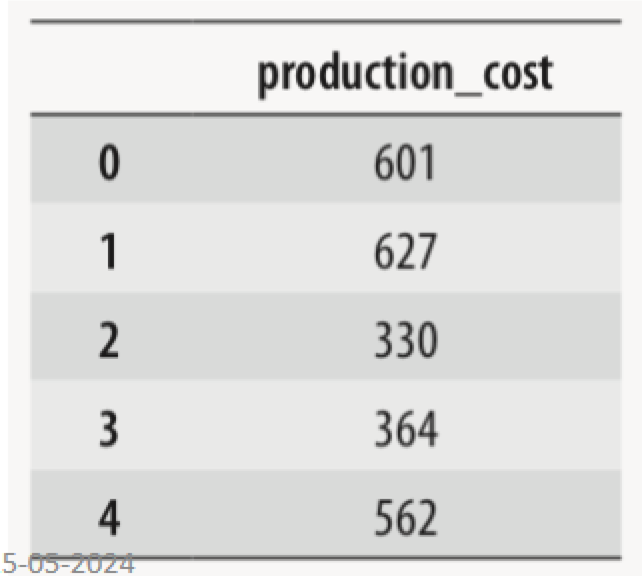
- Questions: Aravind Productions (AP) is a newly formed movie production house based out of Mumbai, India. AP was interested in understanding the production cost required for producing Bollywood movie. The industry believes that the production house will require INR 500 million on average. It is assumed that the Bollywood movie production cost follows a normal distribution. The production costs of 40 Bollywood movies in millions of rupees are given in *bollywoodmovies.csv* file. Conduct and appropriate hypothesis test at ∝ =0.05 to check whether the belief about average production cost is correct.
- Solution: The population mean is 500 and the sample set for actual production cost is available in the file *bollywoodmovies.csv.* The population standard deviation is not known.

One Sample T-test(Parametric HT):Example Solution..

- Reading data:



```
bollywood_movies_df = pd.read_csv(‘bollywoodmovies.csv’)
bollywood_movies_df.head(5)
```



	production_cost
0	601
1	627
2	330
3	364
4	562

One Sample T-test(Parametric HT):Example Solution..

- Defining the hypothesis :

$$H_0 : \mu = 500$$
$$H_A : \mu \neq 500$$

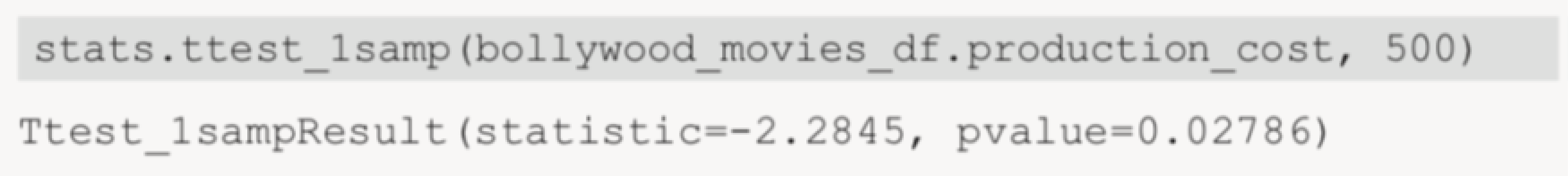
- The built in function for this takes two parameters

  1. an array_like – sample observation

  2. Given popmean– expected value in null hypothesis

One Sample T-test(Parametric HT):Example Solution..

- Conducting the test:



```
stats.ttest_1samp(bollywood_movies_df.production_cost, 500)
```

```
Ttest_1sampResult(statistic=-2.2845, pvalue=0.02786)
```

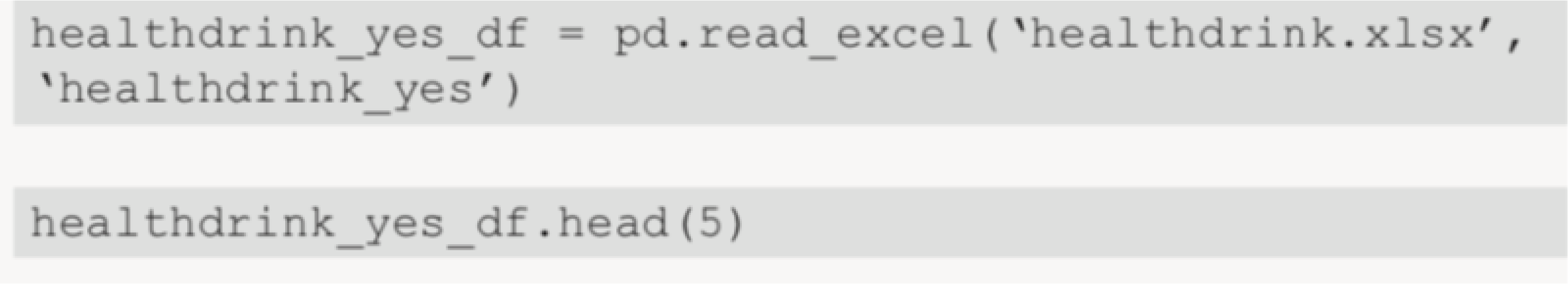
- t-statistic value = -2.2845, and p-value = 0.02786
- p-value is less than 0.05
- Reject the null hypothesis

Two-Sample-T-test

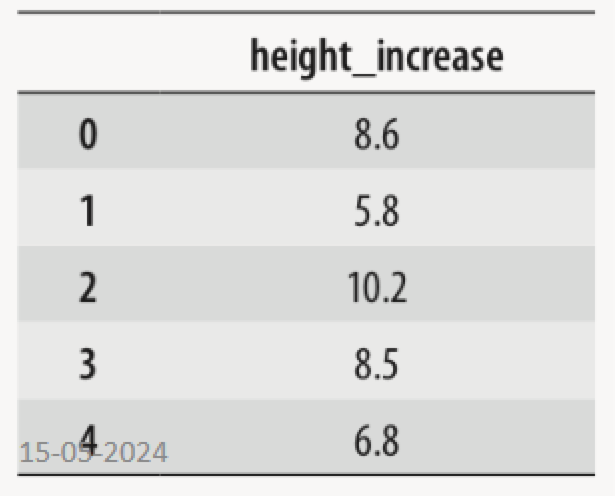
- It is used to test the difference between two population means, if SD is not given.
- Example: A company claims that children who drink their health drink will grow taller than the children who do not drink that health drink. Data in the file *healthdrink.xlsx* shows average increase in height over one-year period from two groups: one drinking the health drink and the other not drinking the health drink. At ∝ =0.05, test whether the increase in height for the children who drink the health drink is different than those who do not drink health drink.

Two-Sample-T-test: Solution

- Reading the data with the tab *healthdrink_yes* as parameter and then display first five records.



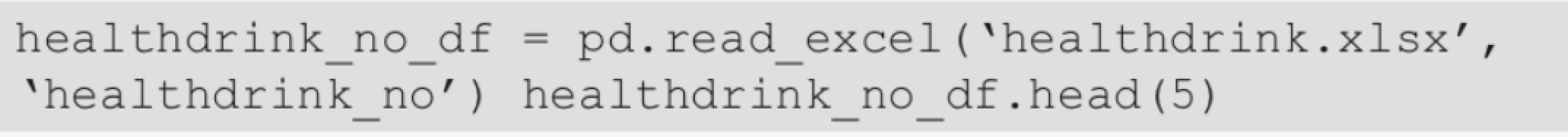
```
healthdrink_yes_df = pd.read_excel('healthdrink.xlsx',
'healthdrink_yes')

healthdrink_yes_df.head(5)
```



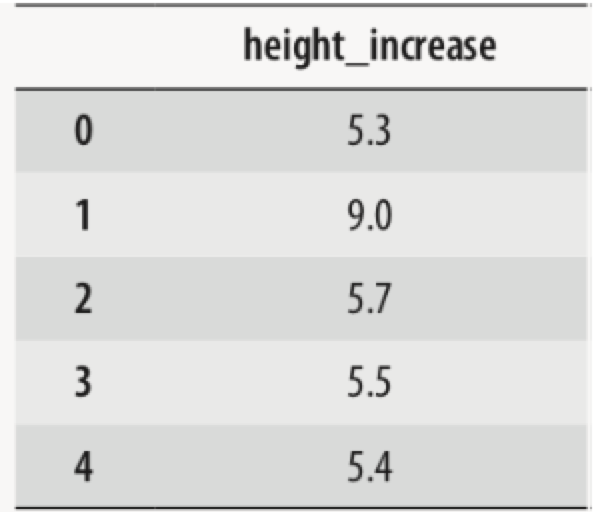
	height_increase
0	8.6
1	5.8
2	10.2
3	8.5
4	6.8

Two-Sample-T-test: Solution..

- Reading the data from the tab *healthdrink_no* in data and then display first records



```
healthdrink_no_df = pd.read_excel(‘healthdrink.xlsx’,
‘healthdrink_no’) healthdrink_no_df.head(5)
```



	height_increase
0	5.3
1	9.0
2	5.7
3	5.5
4	5.4