greatlearning
Power Ahead

A Short Monograph on Statistical Inference: Estimation and Hypothesis Testing

TO SERVE AS A REFRESHER FOR PGP-DSBA

Index

Contents

List of Figures

List of Tables

1. Problem of Inference

1.1 Definitions and Interpretation

Before the subject of statistical inference is properly introduced, several important concepts need to be revisited.

Population: A population is the largest possible collection of items (animate or inanimate) with a set of common characteristics. Size of a population is denoted by N. N may be finite or infinite.

The set of common characteristics define a population. The population of Indian citizens is different from the population of residents of India. Not all Indian citizens may be residing in India; similarly, not all residents in India are Indian citizens. These two populations have a significant overlap, but they are not identical.

In both cases, population size is very large, but they are not infinite.

Now consider the population of all electric bulbs ever produced and will ever be produced. This population is truly infinite.

Similarly, we can define population of (i) all current clients of Airtel, (ii) population of all current clients of Jio and (iii) population of all cell phone users. The population in (iii) contains both (i) and (ii), while (i) and (ii) are two distinct populations, possibly with some overlap.

Sample: A sample is a subset of the population. Size of a sample is denoted by *n*, *n* is always finite.

A sample may or may not be a random sample.

Random Sample: A random sample is a subset of the population where each unit of the population has equal chance of getting selected.

Consider the population of all engineering college graduates in India who graduated between 2005 and 2015, both years included. While choosing a sample from that population, one may prefer only male graduates passing before 2010. Another may prefer only female graduates from the southern states. None of these samples will be random samples.

Only a random sample is a true representative of the population. All statistical inference problem must be based on random samples from the population. Otherwise there is no validity of the inference procedure.

From now on, unless otherwise mentioned, sample will always indicate random sample.

Population Parameter: A population is determined by its parameters. Population parameters are constants. A normal population is determined by its mean μ and variance σ^2 (or standard deviation σ). If these two parameters are known, then all information about this population is completely known. A binomial population is determined by its success proportion π. Population parameters are typically denoted by Greek alphabets.

Sample Statistic: A statistic is a quantity derived from the sample. Sample statistics are denoted by Roman alphabets. Sample mean is a statistic denoted by $\bar{x}$, sample median is another statistic denoted by $\tilde{x}$, sample standard deviation is denoted by *s*. The value of the sample statistic depends on the sample. Since from the same population, more than one random sample of size *n* may be taken, the numerical value of a sample statistic will vary.

Let us consider a small illustrative example.

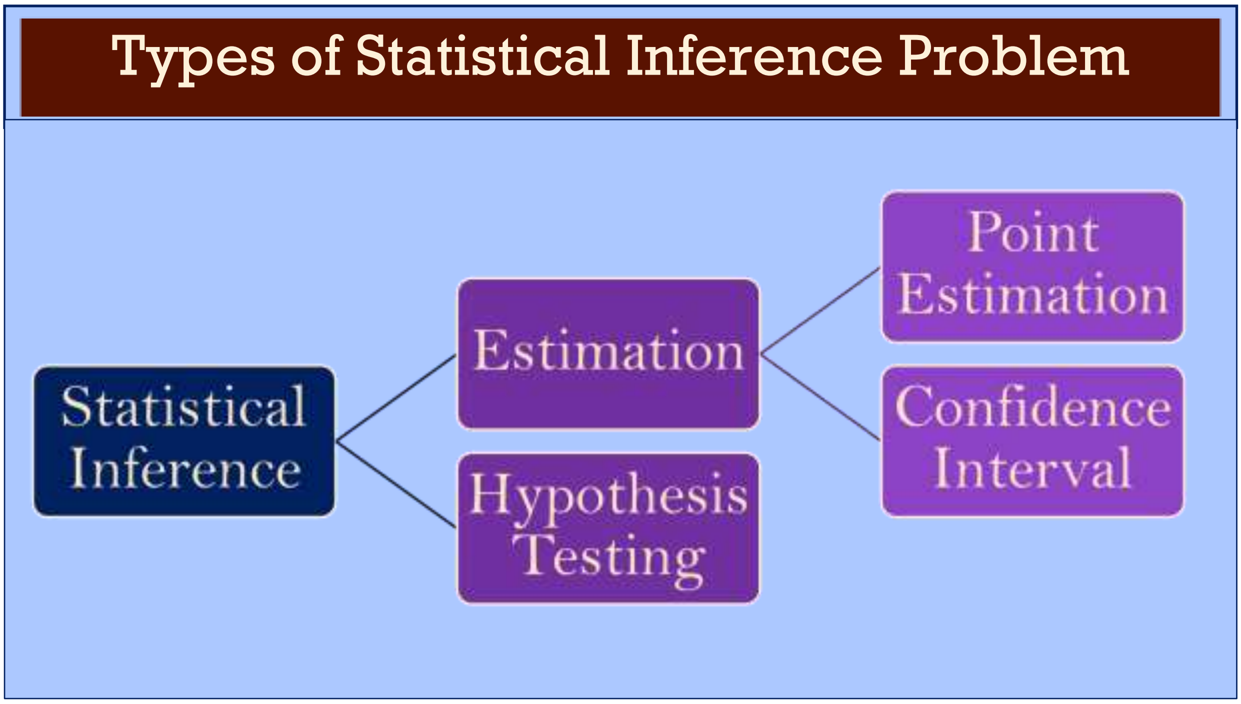
Population of size N = 5: [8, 3, 9, 0, 4]. Population mean = 4.8

Sample 1 of size *n* = 3: [8, 0, 4]. Sample average = 4

Sample 2 of size *n* = 3: [3, 4, 9]. Sample average = 5.33

1.2 The Problem of Inference

The problem of statistical inference comprises of gathering knowledge about a population through a random sample. Since a population is completely known through its parameters, the problem of statistical inference reduces to assessing the population parameters through the sample statistics.



Types of Statistical Inference Problem

Before the problem of inference is taken up, the following important points need to be mentioned.

I. The population is NEVER observed. The true parameter values of the population will always remain unknown.

II. A random sample is used to infer about the population parameters. Sample statistics are used to estimate the population parameters or to test hypotheses about the population parameters

III. The population parameters are (unknown) constants but the sample statistics are random variables. If a different sample is chosen from the same population, values of the statistics will change (refer to the illustrative example in Sec 1.1). A random variable has an associated probability distribution. A sample statistic also has a probability distribution.

IV. Accuracy is important for statistical inference. The procedures discussed subsequently ensures that the inference process has the highest possible accuracy. Accuracy of a procedure is dependent on the variance of the sample statistic used. A statistic with lower variance is more preferable.

Following theorem is pivotal to the statistical inference problem.

1.3 Central Limit Theorem (CLT)

Consider any population with mean μ and standard deviation σ. A random sample of size *n* is chosen from this population. Let $\bar{x}$ and s denote the sample average and sample standard deviation respectively.

If the sample size is large enough ($n \geq 30$) then, $\bar{x}$ follows a **normal distribution** with mean μ and standard deviation $\sigma/\sqrt{n}$

Let us examine the statement before its applications and importance in statistical inference is considered.

The three most important aspects of Central Limit Theorem are

I. The population from which the random sample is drawn need not be a normal population. It may be a binomial population; it may be a positively or negatively skewed population. The CLT will still hold

II. Provided the sample size is large enough, the sample mean will follow a normal distribution.

III. As mentioned in Sec 1.1, a normal distribution is characterized by its mean and standard deviation (sd.). CLT ensures that the distribution of sample average has mean identical to the (not-necessarily normal) population mean μ. But the most important observation is that, sd. of sample average is much smaller $(\sigma/\sqrt{n})$ compared to the population sd. σ.

Distribution of a sample average (mean)

As mentioned before, from a population of size N, a sample of size *n* may be chosen in many ways. For each chosen sample, of size *n*, a sample average may be computed. Each sample average may be numerically different. The histogram constructed with the sample averages indicate the form of the distribution of the sample average.

Experiment I: Follow the steps below.

i) Generate data from a normal distribution. Assume that is the whole population. N = 10000

ii) Take a sample of size *n* = 1000. Compute sample average

iii) Repeat this 500 times

iv) Construct a histogram with 500 sample means

```
# Generate data from a N(5.5, 2.4) distribution
import random
import numpy as np
import statistics
random.seed(573) # Fixes a random seed
Population1 = np.random.normal(5.5, 2.4, 10000)
print(statistics.mean(Population1))
                >>5.511437018449265
print(statistics.stdev(Population1))
                >>2.414308095166452

# Generate 500 samples each of size 1000 from Population1

from numpy.random import randint
sample_list = [sum(random.sample(list(Population1), 1000))/1000 for _ in range(500)]
```

```
print(statistics.mean(sample_list))
                    >>5.511080605286483
print(statistics.stdev(sample_list))
                    >>0.07380174031776532


fig, (ax1, ax2) = plt.subplots(1,2, figsize = (12,5))

plt.rcParams.update({'font.size': 14})

ax1.hist(Population1, 15, color="blue",rwidth=0.85)
ax1.set_xlabel('Distribution of N(5.5, 2.4) Population')
ax1.set_ylabel('Frequency')

ax2.hist(sample_list, 15, color='blue',rwidth=.85)
ax2.set_xlabel('Distribution of Sample Average from N(5.5, 2.4) Population')

plt.tight_layout()
ax1.grid(axis='y', alpha=0.75)
ax2.grid(axis='y', alpha=0.75)
plt.show()
```

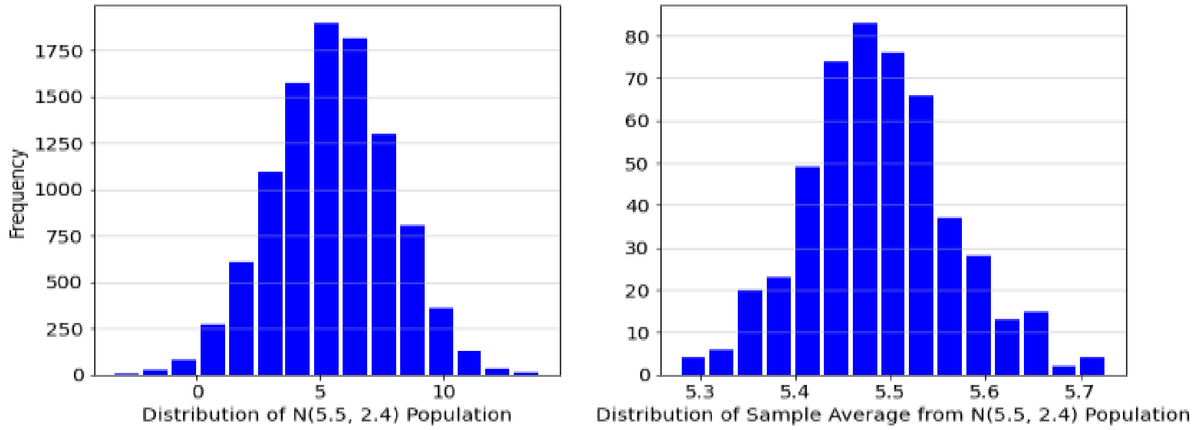


Figure 1: Comparison of normal distribution and sampling distribution of sample average

Compare the population mean (5.51) and the mean from the distribution of sample average (5.51). For all essential matter, the two means are identical.

Compare now the two standard deviations. Std dev for the population is 2.4; std dev of distribution of sample mean is 0.07 = 2.4/√1000. Look at the x-axes in the two histograms.

A smaller standard deviation ensures that the distribution is more concentrated around the men value. The range of values visible on the x-axes of the two histograms is a direct proof of that.

To understand this concept better, repeat the above exercise with different value of random seed, population size N, sample size *n* and the number of times the sample of size *n* is generated.

Experiment II: Follow the steps below.

i) Generate data from a positively skewed distribution. Chi-square distribution with degrees of freedom γ = 3 is used. Assume that is the whole population. N = 10000
ii) Take a sample of size *n* = 1000. Compute sample average
iii) Repeat this 500 times
iv) Construct a histogram with 500 sample means

Note that the chi-square distribution (a common and useful distribution whose many applications will be discussed later in this monograph as well as in other monographs also) is defined by its degrees of freedom γ. The mean of $\chi^2(\gamma)$ is γ and variance is 2 γ.

```
# Generate data from a chi-square(3) distribution

random.seed(573)
Population2 = np.random.chisquare(3,10000)
print(statistics.mean(Population2))
                    >>2.9824936484682776
print(statistics.stdev(Population2))
                    >>2.405164616119428

# Generate 500 samples each of size 1000 from Population2

from numpy.random import randint
sampleavevector = [sum(random.sample(list(Population2), 1000))/1000 for _ in range(500)]
print(statistics.mean(sampleavevector))
                    >>2.9860118717055184
print(statistics.stdev(sampleavevector))
                    >>0.07210415155627425


fig, (ax1, ax2) = plt.subplots(1,2, figsize = (12,5))
plt.rcParams.update({'font.size': 14})
ax1.hist(Population2, 15, facecolor='green', rwidth=0.85)
ax1.set_xlabel('Distribution of N(5.5, 2.4) Population')
ax1.set_ylabel('Frequency')

ax2.hist(sampleavevector, 15, facecolor='green', rwidth=0.85)
ax2.set_xlabel('Distribution of Sample Average from N(5.5, 2.4) Population')
```

```
plt.tight_layout()
ax1.grid(axis='y', alpha=0.75)
ax2.grid(axis='y', alpha=0.75)
plt.show()
```

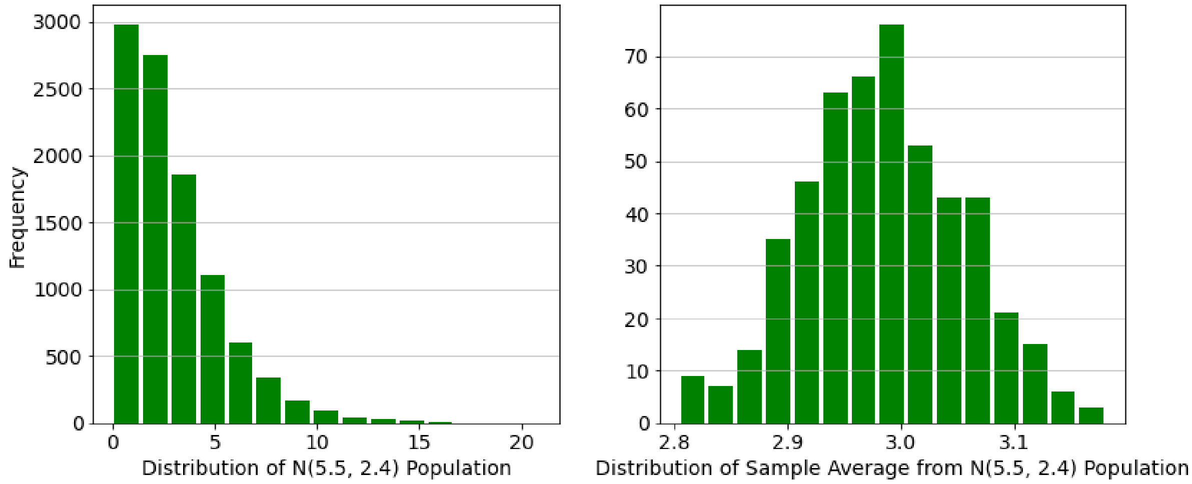


Figure 2: Comparison of chi-square distribution and sampling distribution of sample average

Compare the population mean and the sample mean. They are identical for all practical purposes, as expected. Population std dev is √6 = 2.45. Std dev of the distribution of sample average is 0.07 = 2.45/√1000

To understand this concept better, repeat the above exercise with different value of random seed, population size N, sample size *n* and the number of times the sample of size *n* is generated.

The most important observation here is that the distribution of sample mean is symmetric. In fact, it can be shown mathematically that the distribution follows normal with the above mentioned mean and std dev.

This is the focal point of Central Limit Theorem. It underscores the fact that, *whatever be the distribution in the population*, if the sample size is large enough, the distribution of sample mean **always follows a normal distribution**.

Distribution of a sample statistic is known as a **sampling distribution**.

Another very important contribution of CLT is to emphasise that the std dev of sample mean is much smaller than the std dev of the parent population. The larger the sample is, the smaller will be the std dev of the sampling distribution.

Standard deviation of a sample statistic (standard deviation from a sampling distribution) is known as its **Standard Error (s.e.)**.

S.E. of sample mean is $({}^{\sigma}/_{\sqrt{n}})$, if the standard deviation of the population is denoted by σ.

Caveat: CLT is applicable for sample means only. If a sample statistic cannot be shown as an average, (i.e. if it is not a sum divided by the count), CLT will not be applicable.

1.4 Application of Central Limit Theorem (CLT)

Consider the following two situations.

1. The baggage limit for an airline is set at 50 kg per person. The weight of the baggage of an individual passenger follows a normal distribution with a mean of 45 kg and a standard deviation of 17 kgs. What is the probability that a randomly chosen passenger's baggage will be over the limit?

2. When in a hurry, the airline does not weigh each passenger's baggage, but check whether average baggage weight of 100 passengers is within the limit. What is the probability that average baggage weight of 100 passengers will be over the limit?

To highlight the difference between the situations, the keywords are highlighted. When the behaviour of a sample average is concerned, CLT needs to be applied.

Solution

(1) X: Baggage weight ~ N(45, 17)
Prob[X > 50] = 1 - NORM.DIST(50,45,17,1) = 0.38

```
import scipy.stats
p=scipy.stats.norm(45, 17).cdf(50)
1-p
                              >>0.384
```

There is a 38% chance that a randomly selected passenger's baggage will exceed the limit

(2) $\bar{X}$~ N(45, 17/10 = 1.7)
Prob[$\bar{X}$ > 50] = 1 - NORM.DIST(50,45,1.7,1) = 0.0016

```
p=scipy.stats.norm(45, 1.7).cdf(50)
1-p
                    >> 0.0016
```

There is a 0.16% chance that average of passengers' baggage will exceed the limit

2. Problem of Estimation

Estimation problem involves making a statement about an unknown population parameter based on a sample statistic. If the numerical value of a sample average is used instead of the unknown mean of the population (from where the sample is taken), then it is said that the population mean is estimated by the sample mean.

When one single numerical value is proposed for the unknown parameter, the estimation problem is known as point estimation.

If a range of values is proposed for the unknown population parameter, then the estimation problem is known as interval estimation.

2.1 Point Estimation

When a single numerical value is used to estimate the unknown population parameter, it is called point estimation.

Usually point estimate of an unknown population parameter is the corresponding sample statistic. For example:

a) Population mean μ is estimated by sample mean $\bar{x}$.
b) Population median is estimated by sample median $\tilde{x}$.
c) Population proportion of success π is estimated by sample proportion of success p

One important property for an estimate is **Unbiasedness**.

If the mean of the sampling distribution is equal to the population parameter, the sample statistic is said to be unbiased for the population parameter.

CLT proves that $\bar{x}$ is unbiased for μ. It can be shown (not by CLT) that, for a symmetric distribution, $\tilde{x}$ is an unbiased estimate for population median.

Application of CLT ensures that p is unbiased for π.

Consider the case of population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$. The divisor is the population size N. The sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$ is defined with $(n - 1)$ in the divisor. In this form s^2 is unbiased for σ^2. However, if n is used in the definition of s^2, it does not remain an unbiased estimate of σ^2 any more.

Sample standard deviation s is NOT an unbiased estimate of σ, though s is used as a point estimate of σ.

2.2 Interval Estimation or Confidence Interval

When a range of values is used to estimate an unknown population parameter, it is called interval estimation.

To some extent point estimation is intuitive. Sample average to be used in place of unknown population mean is only sensible. However, when it comes to a range of values to be recommended, more than a simple intuition is needed.

The confidence interval is developed by exploiting the properties of sampling distribution of the sample statistic.

2.2.1 Construction of Confidence Interval

Consider the problem of interval estimation of a normal population mean μ

The sample statistic considered is the sample mean $\bar{x}$ since it provides an unbiased estimate of μ. The sampling distribution of $\bar{X}$ is normal. An interval of the form $\bar{x} \pm E$ will be defined so that

$$\text{Prob}[\bar{X} - E < \mu < \bar{X} + E] = 100(1 - \alpha)\%$$

This is known as $(100 - \alpha)$ % Confidence Interval for the population mean.

The most common value for α is 0.05 and typically 95% confidence intervals are constructed. Another common value for α is 0.01, leading to a 99% confidence interval.

The confidence interval constructed above is symmetric about $\bar{x}$. For overwhelming majority of cases, the confidence intervals are constructed so that they are symmetric about the sample statistic. This ensures several optimal properties of the confidence intervals, any discussion of which is beyond the scope of this monograph.

Let us revisit the sampling distribution of sample mean. Let us also assume that the parent population is not a normal population. A random sample of size *n* is taken from the population, where *n* is large enough so that CLT is applicable. It may be concluded that

$$\bar{X} \sim N(\mu, \text{s.e.}(\bar{X})$$

It is also known that if $X \sim N(\mu, \sigma)$, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$. Hence $Z = \frac{\bar{X} - \mu}{s.e.(\bar{X})} \sim N(0, 1)$. Substituting the value of $\text{s.e.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, the following result is obtained:

$$\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} \sim N\,(0, 1)$$

This distribution is pivotal in construction of the confidence intervals, as well as performing hypothesis tests, that we will see later.

For a standard normal distribution Prob[-1.96 < Z < 1.96] = 0.95. This can be easily verified.

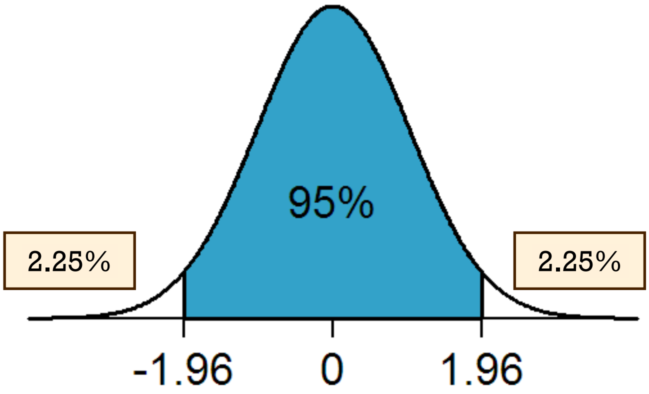


Figure 3: Standard normal distribution with 95% critical points (source: Google image)

One important assumption is to be made now. The problem here is to construct a confidence interval for the unknown normal mean μ, but it is assumed that the sd. σ is known. This may not be as unlikely an assumption as it seems. For a tight manufacturing process which is in place for a long time, the variability in the process may be controlled, but with small changes in temperature or humidity, the mean may change.

Assuming σ is known,

$$\text{Prob}\ [\ -1.96\ <\ \frac{\sqrt{n}\,(\bar{X}-\mu)}{\sigma}\ <\ 1.96] = 0.95$$

Note that in the above formulation, the only unknown quantity is μ. After an algebraic manipulation, the above reduces to

$$\text{Prob}\ [\ \bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu\ < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\ ] = 0.95$$

Therefore, the 95% confidence interval for normal mean μ is given by $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. The associated probability (here 95%) is known as the confidence coefficient.

Example 1:

The caffeine content (in mg) was examined for a random sample of 50 cups of black coffee dispensed by a new machine. The mean of the sample is found to be 110 mg. It is known that the standard deviation from all the machines of that manufacturer is 7 mg. Construct a 95% confidence interval for μ, the mean caffeine content for cups dispensed by the machine.

Solution:

Sample mean $\bar{x}$ = 110 mg

Sample size *n* = 50

Population std dev σ is known to be 7

Hence the 95% C.I of population mean μ is $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 110 \pm \frac{7}{\sqrt{50}} = [109.01, 110.99]$

2.2.2 Interpretation of Confidence Interval

Before the confidence intervals are interpreted, two important points needs to be recalled from Sec 1.

I. The unknown population parameter is a constant. This parameter value will never be known

II. From a population of size N, a very large number of random samples may be selected. For each of them, the value of sample statistic will be different

Confidence interval provides an interval, or a range of values, which is expected to cover the true unknown parameter. In the current scenario, in the form of the confidence interval, the sample mean x̄ will change from one random sample to the next. Recall the two illustrative examples in Sec 1. Hence, for each different sample, a different confidence interval is obtained.

If a large number of confidence interval is constructed with 95% confidence coefficient, then 95% of those intervals will cover the true unknown population mean; in the rest 5% of cases the interval will NOT cover the population parameter.

In practice, only one sample is chosen from the population and only one confidence interval is constructed. It is never known whether that interval actually covers the population parameter, because μ will always remain unknown.

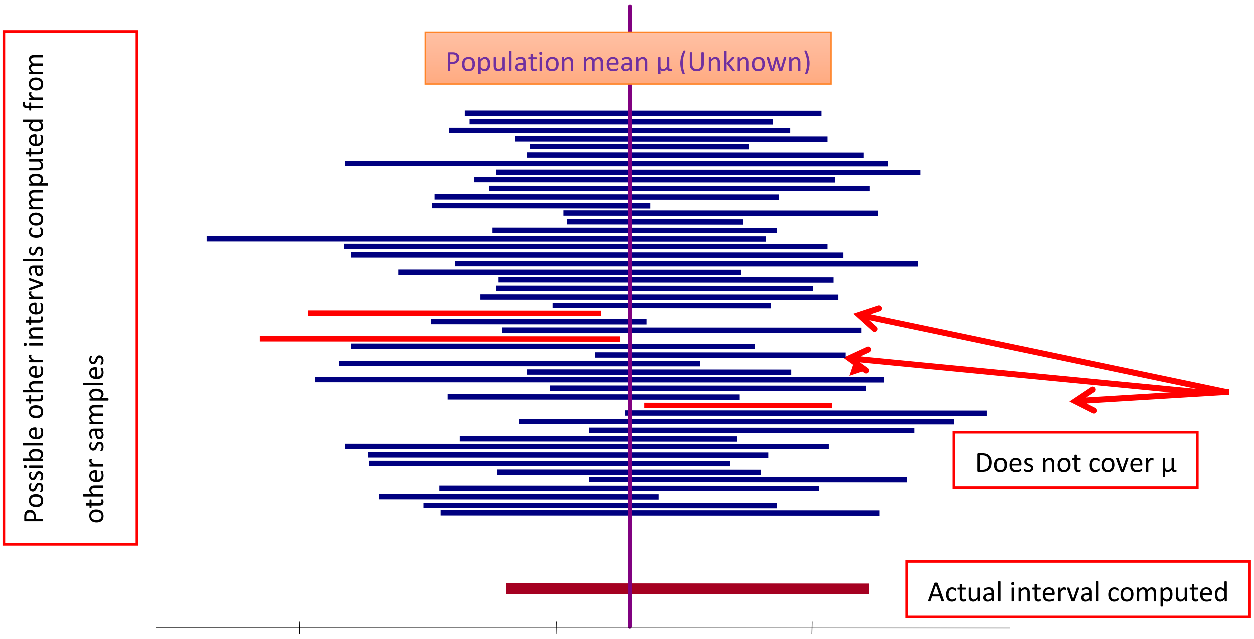

Possible other intervals computed from other samples
Population mean μ (Unknown)
Does not cover μ
Actual interval computed

Figure 4: Illustrative interpretation of confidence interval

Experiment III: Follow the steps below.

i) Refer to Experiment I. Take the set of 500 independent random samples, which has already been created. Recall that each sample is of size n = 1000 and is generated from a normal population with given parameters ($\mu = 5.5$, $\sigma = 2.4$).

ii) Convert this into a problem of construction of confidence interval for a normal mean. Assume μ is unknown, but $\sigma = 2.4$ is known.

iii) Compute sample average from each sample.

iv) Apply the formula $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$, where $\bar{x}$ is computed from each of the 500 samples of size 1000.

v) Count how many of these intervals actually contain $\mu = 5.5$, and divide it by 500 to get the proportion of intervals that contains the population mean.

```
from numpy.random import randint
sampleavevector = [sum(random.sample(list(Population1), 1000))/1000 for _ in range(500)]
print(statistics.mean(sampleavevector))
                >>5.503842705172149
print(statistics.stdev(sampleavevector))
                >>0.0740453580482261
import numpy as np
import math
Conflimt = np.zeros( (500, 2) )
counter  = 0
for i in range(0,Conflimt.shape[0]):
   Conflimt[:,0] =sample_list[i] - 1.96*2.4/math.sqrt(1000)
   Conflimt[:,1] =sample_list[i] + 1.96*2.4/math.sqrt(1000)
   if (((Conflimt[:,0] <5.5).any() & (Conflimt[:,1] >5.5).any())):
      counter = counter+1
counter
Confcoeff = counter/500
print(Concoeff)
                >>0.954
```

Note that in this case 479 out of 500 confidence intervals, i.e. 95.4% intervals actually cover the true population mean.

In practice only one sample will be taken from the population and only one confidence interval will be constructed.

The interpretation of a 95% confidence interval is that, if the process is repeated a large number of times, then the intervals so constructed, will contain the true population parameter 95% of times.

The confidence interval that is constructed from the single sample, **MAY or MAY NOT** contain the true population mean.

To understand these concepts better, repeat Experiment III with different value of random seed, population size N, sample size *n* and the number of times the sample of size *n* is generated. Also change the confidence coefficient.

Interpretation of confidence interval in Example 1: 95% of the time, the mean caffeine content for cups dispensed by the machine is between 109.01 mg and 110.99 mg.

2.2.3 The Confidence Coefficient

The most common choice of confidence coefficient is 95%. But that is only convention. There is no reason that any other confidence coefficient cannot be used. Often 99% confidence interval is also constructed.

Definition: The 100α-*th* percentile point of a probability distribution is that point, below which lies $100\alpha\%$ probability. In other words, if x^* is the 100α-*th* percentile point of a distribution, then $P[X < x^*] = 100\alpha\%$.

Consider a standard normal distribution Z. The 95-th percentile point of a standard normal distribution is determined by $z_{0.95}$ so that $P[Z < z_{0.95}] = 0.95$.

```
from scipy.stats import norm
norm.ppf(.95)
                >>1.64485
```

$z_{0.95} = 1.645$

Similarly, it can be easily seen that $z_{0.975} = 1.96$, i.e. $P[Z < z_{0.975}] = 0.975$.

Because normal distribution is symmetric around 0, the following observations hold

- $P\,[Z > 1.645] = 0.05$
- $P\,[Z > 1.96] = 0.025$
- $P\,[Z < -1.645] = 0.05$
- $P\,[Z < -1.96] = 0.025$

The confidence coefficients depend on the percentile points of a distribution. While constructing 95% confidence intervals, the relevant percentile points are taken so that above the larger lies 100 α/2% of probability, and below the lower one lies 100 α/2% of probability.

Example (i): $\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$ gives 90% confidence interval of the normal mean when the std dev is known. (Why?)

Example (ii): $\bar{x} \pm 2.576 \frac{\sigma}{\sqrt{n}}$ gives 99% confidence interval of the normal mean when the std dev is known. (Why?)

By varying the percentile points, confidence intervals of any given confidence coefficients may be constructed.

2.2.4 Properties of The Confidence Interval

For all analytical problems involving estimation, confidence intervals are always presented along with the point estimate. The most important properties of the confidence interval are given below:

I. Length of a confidence interval is the difference of the upper and the lower limit of the interval.
II. Half length of the confidence interval is known as the Margin of Error. Often the margin of error is stated, instead of the confidence interval. This is often seen for approval rating of President (in US) or prediction of election result (in India or other democratic countries).
III. If std dev increases, the length of the confidence interval increases, provided the confidence coefficient and the sample size remain the same
IV. If the sample size increases, the length of the confidence interval decreases, provided the confidence coefficient and the std dev remain the same
V. If the confidence coefficient increases, the length of the confidence interval increases, provided the sample size and the std dev remain the same

Why 100% confidence interval is never used?

The importance of a confidence interval is in identifying a region that covers the population parameter with a certain known probability. A 100% confidence interval will include **All** possible values. Hence there will be no insight into the problem.

2.2.5 Construction of Confidence Interval for Population Mean when σ is unknown

The discussion so far is based on the assumption that the population std dev is known. However, in the majority of cases that assumption will not be satisfied. If the population σ is not known, it will be estimated from the sample. In the form of the confidence interval, the quantity $\frac{\sigma}{\sqrt{n}}$ will be replaced by $\frac{s}{\sqrt{n}}$, which is the estimated std. error of $\bar{x}$.

It has been stated in Sec 2. that $\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$

However, when σ is replaced by its sample estimate s, the same quantity does not follow a standard normal distribution any more.

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t\,(n - 1)$$

This is known as Student's t distribution, after an English statistician Gosset, who used the pseudonym Student.

2.2.5.1 t distribution

Just like the normal distribution, t-distribution is also very useful in statistical inference. It is a symmetric distribution around 0. The parameter of t-distribution is known as degrees of freedom (d.f.). Degrees of freedom can take any integer value between 1 and infinity. For a very large d.f., t distribution is almost identical to standard normal distribution. The d.f. of t-distribution is a function of the sample size *n*.

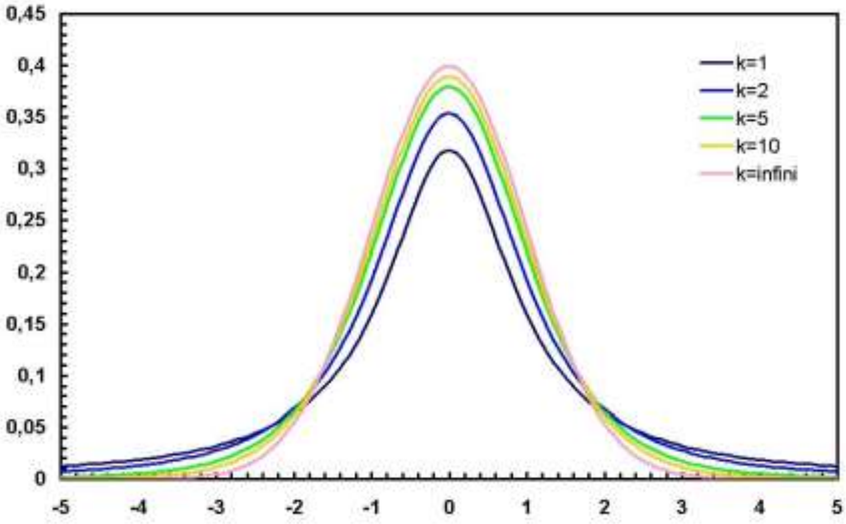


Figure 5: Shape of the t-distribution for various d.f.(k). (Source: Google image)

The construction of confidence interval for μ follows the same rule with one modification: The appropriate percentile point of the t (*n – 1*) distribution is used instead of the normal percentile point.

Example 2:

The caffeine content (in mg) was examined for a random sample of 50 cups of black coffee dispensed by a new machine. The mean of the sample is found to be 110 mg and the sample standard deviation is estimated to be 7 mg. Construct a 95% confidence interval for μ, the mean caffeine content for cups dispensed by the machine.

Solution:

Sample mean $\bar{x}$ = 110 mg

Sample size n = 50

Sample std dev s = 7

The d.f. = 50 – 1 = 49

The 97.5th percentile point of a t(49) distribution is 2.01

```
from scipy.stats import t
t.ppf([0.975], 49)
                    >>2.009
```

Hence the 95% C.I of population mean μ is $\bar{x} \pm 2.01 \frac{s}{\sqrt{n}} = 110 \pm \frac{7}{\sqrt{50}}$ = [108.01, 111.99]

Interpretation of confidence interval in Example 2: 95% of the time, the mean caffeine content for cups dispensed by the machine is between 109.01 mg and 111.99 mg.

Comparison of this confidence interval in Example 1 shows that the interval constructed for Example 2 is wider.

2.2.6 Confidence Intervals for Parameters (other than Normal Means)

So far only the confidence interval of normal population mean is explicitly derived, for both cases when the population standard deviation is known and when it is unknown. The latter case is more common which demands application of *t*-distribution with appropriate degrees of freedom.

However, confidence interval is associated with every estimation problem, such as estimation of binomial success probability, estimation of normal variance, as well as estimation of all regression parameters and parameters of any other model building exercise. The general approach to construction of confidence interval is to use the appropriate sample statistic to estimate the population parameter and use the proper percentile point of the sampling distribution.

No further confidence interval will be explicitly derived at this point. Construction of confidence interval and hypothesis testing enjoy a close relationship. In the next section, when hypothesis testing problems are taken up, associated confidence intervals will be indicated in a few cases.

3. Hypothesis Testing

In the previous section the problem of estimation is considered, where there is no previous knowledge of the population parameter. The problem is simpler in that case. A random sample is taken, a sample statistic is computed and an appropriate point and interval estimate is suggested.

Often the interest is not in the numerical value of the point estimate of the parameter, but in knowing whether the population parameter is less than a certain value or greater than a certain value.

Why the numerical value of the sample statistic is not enough to address that question? Think about the sampling distribution of the statistic. The numerical value of the statistic will change from one sample to the next. The standard error will depend on the sample size. One single numerical value is not reliable enough for inference.

Consider another example where you are interested in knowing whether two or more populations have equal means. It is not the numerical values of the population parameters that are of interest, but it is their relative values. Similarly, one may be interested in knowing whether two or more populations have same variance.

A hypothesis is a conjecture about a parameter of a population.

Based on belief or previous knowledge (process / domain / observation) a hypothesis is formulated.

Let us consider the following examples.

I. A marketing manager in a grocery superstore is interested in knowing whether the newly introduced free delivery within 3-mile radius of the store has increased sales

II. In the current WFH situation, an HR manager of a company with global presence wants to see whether the effect of lockdown is the same in India, US, Germany and UK.

III. A food processing giant claims that their instant noodles contains only trace amount of lead which can never be carcinogenic while the NABL asserts that the amount is enough for causing harmful effects with prolonged use

IV. Ministry of Women and Child Development is afraid that lockdown has an adverse effect on women and children, it is the cause of higher proportion of domestic violence

V. Indian government claims that over last five decades there is a significant drop in the total fertility rate but the rate is not uniform across all religions

It is clear that estimation is not enough to arrive at a conclusion for these examples. Each one of these statements involve conjectures about one or more parameters of one or more populations. Statistical testing procedure is developed to tackle such problems.

Objective of the hypothesis testing procedure is to SET a value for the parameter(s) and perform a statistical TEST to see whether that value is tenable **in the light of the evidence gathered from the sample**.

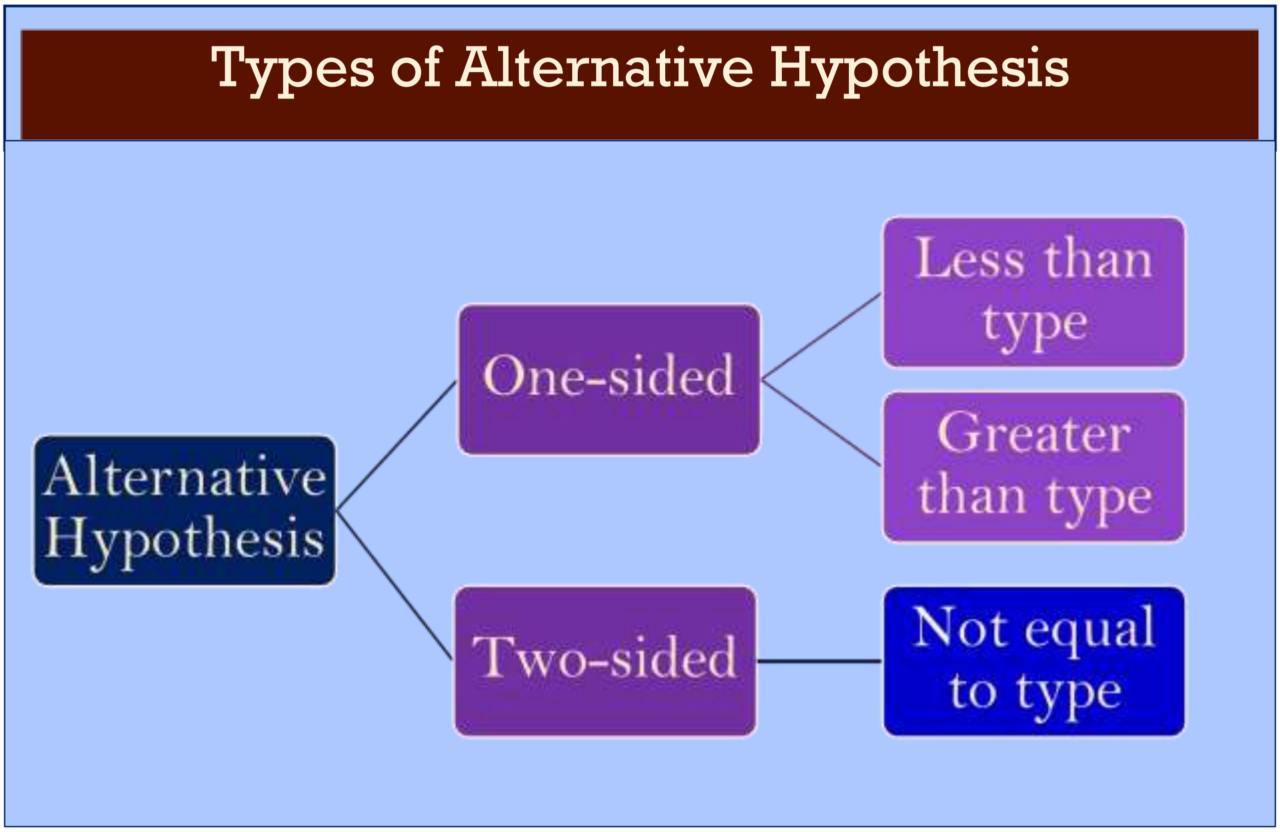
3.1 Basic Concepts and Definitions of Hypothesis Testing

Before the hypothesis testing procedure is developed, several important concepts need to be clarified.

Null hypothesis (H_0) is the presumed current state of the matter or status quo.

It may be a prevalent opinion, or it may be based on previous knowledge. It can also be an assumption or a prevailing theory, based on which a change will be measured.

Alternative hypothesis (H_A) is the rival opinion or research hypothesis or an improvement target.



Null hypothesis is assumed to be true till reasonably strong evidence to the contrary is found.

Based on a random sample a decision is made whether there exists reasonably strong evidence against H_0. An appropriate sample statistic is used to define a test statistic. The evidence is tested against a pre-determined Decision Rule. If H_0 is not supported (automatically) H_A comes into force.

A Test Statistic is a random variable based on H_0 and sample statistic.

It (usually) follows a standard distribution, eg. normal, t, F, chi-square, etc. and is used to make a choice between H_0 and H_A. Since hypothesis testing is done on the basis of sampling distribution, the decisions made are probabilistic. Hence it is very important to understand the **errors associated with hypothesis testing**.

Errors are NOT mistakes.

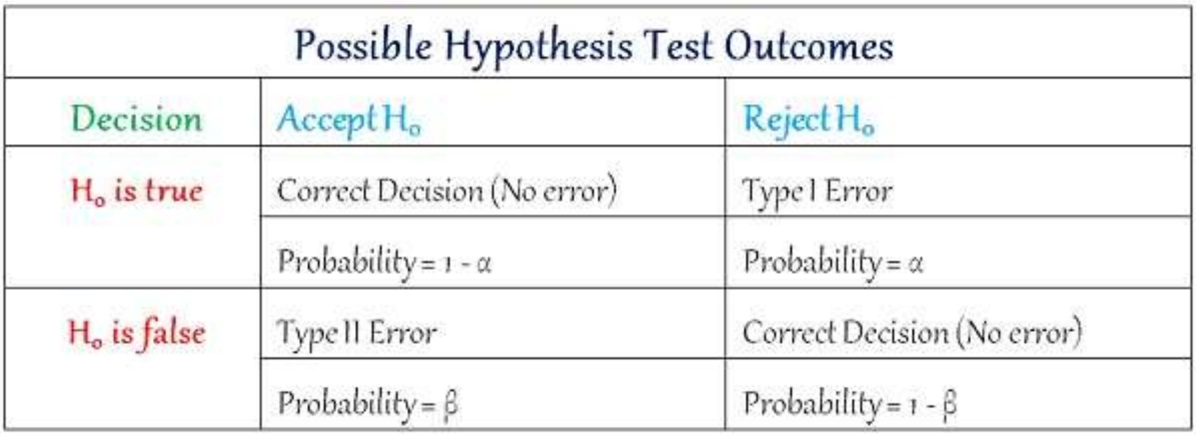
Type I error: If H_0 is TRUE but based on observed sample and decision rule H_0 is REJECTED

Type II error: If H_0 is FALSE but based on observed sample and decision rule, H_0 is NOT REJECTED

Type I error is assumed to be more critical and is kept at a suitably low level. Prob[Rejecting H_0 when H_0 is true] is denoted by α. This is also known as the level of significance of the test. This has a close relationship with the confidence coefficient, discussed in Section 2.2.3, which will be explained later.

Type II error depends on the alternative hypothesis and is denoted by β. Prob[Not rejecting H_0 when H_A is true] is defined as **Power** of the test.

When a hypothesis test protocol is set up, the objective is to set it up in such a way that the power of the test is maximized.



Possible Hypothesis Test Outcomes

Decision	Accept H_o	Reject H_o
H_o is true	Correct Decision (No error)	Type I Error
	Probability = 1 - α	Probability = α
H_o is false	Type II Error	Correct Decision (No error)
	Probability = β	Probability = 1 - β

Figure 6: Errors of hypothesis testing (Source: Google image)

The steps of hypothesis testing are given below:

I. Identify the population parameter regarding which the hypothesis is to be constructed

II. Set up Null and Alternative hypotheses: H_0 and H_A. Check the type of alternative (less than type, greater than type, not equal to type)

III. Choose the correct test statistic. Test statistic does not depend on the type of alternative

IV. Distribution of test statistic is determined **ASSUMING** H_0 is true

V. REJECTION RULE is set a priori. Rejection rule depends on the type of alternative and the probability of type I error

VI. Compute the numerical value of the test statistic, check whether the numerical value falls in the rejection region

VII. If the numerical value falls in the rejection region, reject H_0, otherwise H_0 is not rejected

Before examples of hypothesis testing problems are discussed, note that the null and the alternative hypotheses partition the whole parameter space. That means, each and every possible value of the population parameter is either part of the null hypothesis or part of the alternative hypothesis.

3.2 Setting up of null and alternative hypotheses: Illustrations

In this section several illustrative examples are shown to familiarize the reader to set up the null and alternative hypotheses. Once these are clearly identified, and the appropriate test statistic is chosen, the other steps are almost mechanical.

In this section no actual hypothesis test will be performed.

Example 1. The Environmental Protection Agency releases figures on SPM in urban localities. For New Delhi the average figure for the past year was 528.8 g/m^3. After the odd-even rotation policy with private automobiles was implemented the Government claims that the amount of SPM has reduced significantly in New Delhi.

Solution: The population parameter regarding which the hypothesis is to be tested is the population mean μ. The given condition is that mean SPM is 528.8 g/m^3. Hence

H_0: $\mu = 528.8$ or H_0: $\mu \geq 528.8$

The new claim or the expected change is that mean SPM is less than what is was before. Hence

H_A: $\mu < 528.8$

Important points to note:

- The = sign is always included in the null hypothesis.
- H_0: $\mu = \mu_0$ against H_A: $\mu < \mu_0$ is identical to H_0: $\mu \geq \mu_0$ against H_A: $\mu < \mu_0$
- H_0: $\mu = \mu_0$ against H_A: $\mu > \mu_0$ is identical to H_0: $\mu \leq \mu_0$ against H_A: $\mu > \mu_0$

There is deep mathematical reason for the above which is beyond the scope of this monograph.

Example 2: Suppose that in the big metros in India average price of real estate is Rs 3000 per square foot. When I moved to Delhi I was told that in Delhi real estate price is much higher than national average. To test that claim I took a sample of 47 apartments and looked at their prices.

Solution: The population parameter regarding which the hypothesis is to be tested is the population mean μ. The given condition is that mean price per square foot is Rs 3000. Hence

H_0: $\mu = 3000$ or H_0: $\mu \leq 3000$

The new claim is that mean price in Delhi is more than the national average. Hence

H_A: $\mu > 3000$

Example 3: Mean length of lumber is specified to be 8.5m for a certain building project. A construction engineer wants to see whether the shipments she received adhere to that specification.

Solution: The population parameter regarding which the hypothesis is to be tested is the population mean μ. The given condition is that the mean length of lumber is 8.5m. Hence

H_0: $\mu = 8.5$

The shipment will not adhere to specification if the average length is either less or more than 8.5 m. Hence

H_A: $\mu \neq 8.5$

This is an example of two-sided alternative. The null hypothesis is a point and the alternative is the entire parameter space, except that single point.

Example 4: There is a belief that 20% of men on business travel abroad brings a significant other with them. A chain hotel thinks that number to be too low.

Solution: The population parameter regarding which the hypothesis is tested is the population proportion π. The belief is that 20% men are accompanied on business travel abroad. Hence

H_0: $\pi = 0.2$ or H_0: $\pi \leq 0.2$

The hotel chain claims that the null value is too low. Hence

H_A: $\pi > 0.2$

Example 5: A survey on Indian college students shows that 70% or more male students smoke. The social scientists believe that the proportion of smokers among female students is significantly low.

Solution: The population parameter regarding which the hypothesis is tested is the proportion π of female smokers. The survey finds at least 70% men are smokers. Hence

H_0: $\pi \geq 0.7$

The research hypothesis is that among females the proportion of smokers is less. Hence

H_A: $\pi < 0.7$

Example 6: A survey is done on Indian college students to determine, among other things, whether female students smoke less than the male students.

Solution: This problem is different from all the problems discussed so far. Here there are two different populations, male and female college students. There is no value given for any of the population proportions, both are unknown. In Example 5, the proportion of male smokers was known. This is called a Two-sample Problem.

The null hypothesis is that the proportion of smokers among both male and female students is equal. The alternative is clearly stated to be that females smoke less. The alternative is the research hypothesis, which the survey targets. Hence

H_0: $\pi F \geq \pi M$ against H_A: $\pi F < \pi M$

Example 7: A company produces industrial discs. Each batch of disc is specified to have 0.03cm thickness. The company inspects the discs in 8 locations and determines that the thickness adheres to specifications on the average but the measurements vary across locations. It is not acceptable if the variance of a batch is more than 0.001cm. The company wants to see if the manufacturing process is under control.

Solution: Here is an example of a problem where the population parameter of interest is the population variance σ^2. The manufacturing process comprises both mean and variance to remain under control limits. It is clearly stated that the mean adheres to process specification but the variance may or may not be.

Let us examine the nature of the null hypothesis. The company wants to check whether the variance is less than or equal to 0.001 cm. This is their status quo and this is what the company claims too. Unless there is very strong evidence against this, the company does not want to make any adjustment in the manufacturing process. Hence the null and alternative hypotheses will be

H_0: $\sigma^2 \leq 0.001$ against H_A: $\sigma^2 > 0.001$

Note that the test of variability is always set up in terms of variance, not in terms of standard deviation.

Example 8: A tea chain is interested in estimating the difference in the average daily consumption of regular black tea and all other exotic teas (which includes all varieties of green tea and herbal tea). The manager of one shop randomly selects 20 regular black tea drinkers and asks them how many cups of regular black tea they consume. She also selects 15 drinkers of exotic teas and asks how many cups of such tea do the drink.

Solution: This is also a two-sample problem, where the difference in the two *independent* population means is of interest. Since the store manager is interested in knowing whether there is any difference in the population means, the alternative hypothesis is two-sided.

H_0: $\mu_B = \mu_E$ against H_A: $\mu_B \neq \mu_E$

Important points to note: Hypothesis test of difference of two independent population means is a difficult problem. While performing the test, several assumptions need to be checked. These will be discussed in detail at the appropriate place.

Example 9: A marketing manager wants to see whether television viewing time for husbands is more than the wives in the same household. In order to do that, she takes a random sample of 80 households and find out TV viewing time for both spouses in those households.

Solution: This problem is known as a paired-sample problem, where each household is assumed to be one unit and two measurements are taken on each unit. Here the two populations are not independent. Though the form of the null and alternative hypothesis is identical to the problem of two independent samples, the testing procedure differs. This will be discussed later at the appropriate place.

The null and the alternative hypotheses are

H_0: $\mu_H \leq \mu_W$ against H_A: $\mu_H > \mu_W$

Example 10: A survey was done after the draft Education Policy 2020 was published with 578 college teachers. Each of them was asked whether they voted for the ruling BJP in 2019 or not and whether they are in favour of, against or indifferent to the NEP. The following table shows the result. Does it show evidence that favouring NEP is independent of voting for BJP?

	New Education Policy			
Voted for BJP	Favors	Indifferent	Against	Total
Yes	205	27	3	235
No	64	69	210	343
Total	269	96	213	578

Table-1 (Survey Results on Education Policy)

Solution: Here the form of the null and alternative hypothesis is completely different from what has been discussed so far.

H_0: Favouring NEP is independent of voting for BJP against

H_A: Favouring NEP is NOT independent of voting for BJP against

Note that for all problems involving cross-classified data, the null and the alternative hypotheses are of this form. Again, these will be discussed in detail in appropriate places.

In addition to the problems mentioned above, there are many other different situations where hypothesis testing plays important roles. It is not possible for one short monograph to even list all of those. Many other types of cases will be dealt with in subsequent monographs, but there will still be more and different cases left out. To my knowledge no book exists, that deals with all types of hypothesis testing problems comprehensively.

3.3 Critical Point, Level of Significance, Rejection Region, P-value

Once the null and alternative hypotheses are set up in a proper manner, next step is to come to the conclusion whether the null hypothesis can be rejected. The only tool to take a decision is to depend on the random sample and the appropriate sample statistic. Recall that each and every sample statistic has a sampling distribution, that depends on the population parameter.

The process of hypothesis testing may be thought of as a reverse process of estimation. In estimation, the unknown parameter is estimated directly. A random sample is collected from the population and an appropriate quantity is computed from the sample. This is used to estimate the unknown population parameter. Because of the sampling distribution, i.e. because the value of the sample statistic may change from one sample to another, an associated confidence interval is also proposed.

If the problem is to estimate a population mean, a sufficiently large random sample is taken, sample mean computed and that numerical value is proposed as an estimate of population mean. Apply CLT and the associated confidence interval is also available.

Hypothesis testing is the reverse approach. First a null hypothesis is set, i.e. a value of the unknown population mean is put forth. Goal is to keep this value of the population parameter, till there is reasonably strong evidence that the hypothesised value may not be correct. A rejection rule for the null hypothesis needs to be put in place so that the evidence can be tested.

The rejection rule (rejection region) depends on the alternative hypothesis. The following is shown in terms of population mean. But the same rule applies to all hypothesis testing problems. With the population parameter, the form of the test statistic will change.

$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$

(less than type alternative)

- Reject H_0 if the value of test statistic is **too small**

$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$

(greater than type alternative)

- Reject H_0 if the value of test statistic is **too large**

$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$

(two-sided alternative)

- Reject H_0 if the value of test statistic is either **too small or too large**

A threshold value is determined from the sampling distribution of the sample statistic. The threshold depends on the pre-determined significance level α = Prob[Type I Error]. The test statistic for testing hypotheses for population mean is $\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t\,(n-1)$.

The test statistic is always evaluated assuming $\mu = \mu_0$.

If H_A: $\mu < \mu_{0,}$ the value of the test statistic is computed and checked whether it falls in the left tail (the red zone) in Fig 7. The zone is determined so that the probability of being there is α (or less), when H_0 is true. The critical value on the left-hand side is the $100\alpha^{th}$ percentile point.

If H_A: $\mu > \mu_{0,}$ the value of the test statistic is computed and checked whether it falls in the right tail (the red zone) in Fig 7. The zone is determined so that the probability of being there is α (or less), when H_0 is true. The critical value on the left hand side is the $100(1-\alpha)^{th}$ percentile point.

If H_A: $\mu \neq \mu_0$, the value of the test statistic is computed and checked whether it falls in the right tail or left tail (there are two red zones) in Fig 7(last). If the test statistic has a symmetric distribution (such as Z or t distributions), then the rejection regions are symmetric, i.e. the size of each zone is exactly half of the total. The zone is determined so that the probability of being in the left tail is $\alpha/2$ (or less) and the probability of being in the right tail is also $\alpha/2$ (or less), when H_0 is true. The critical points are $100\alpha/2$th (left) and $100(1 - \alpha/2)$th (right) critical points.

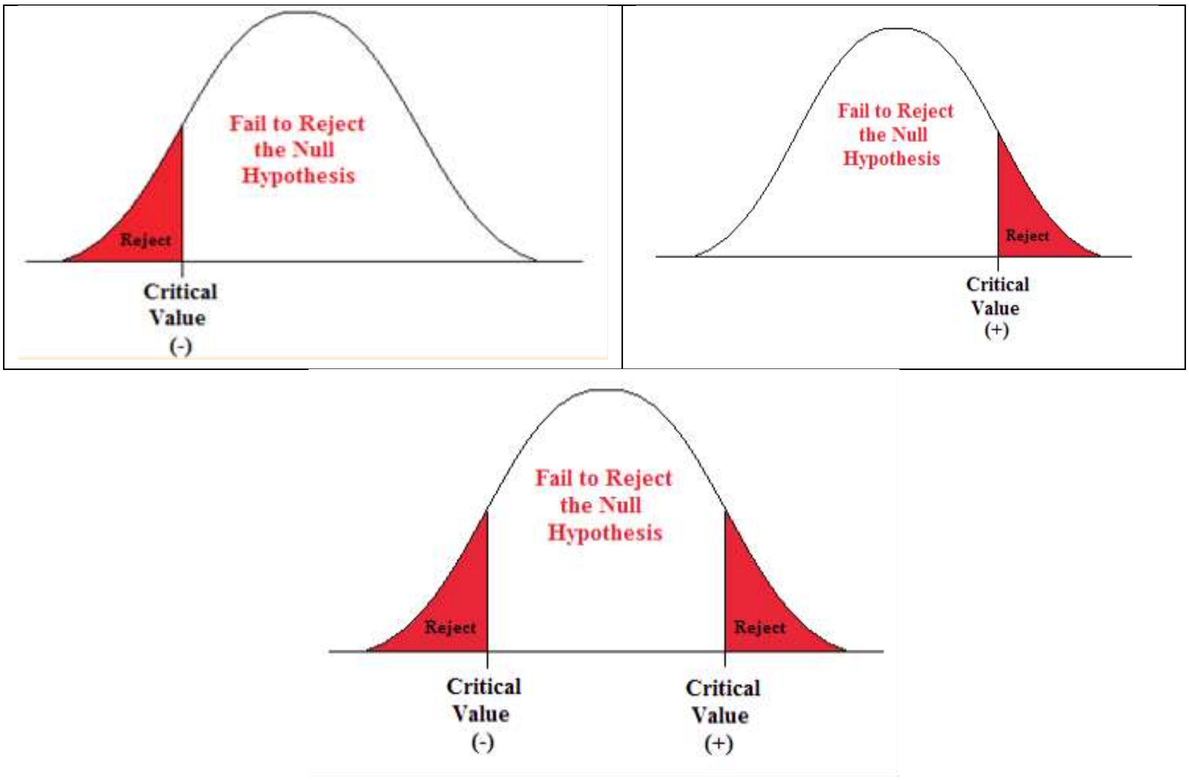

Fail to Reject the Null Hypothesis
Reject
Critical Value (-)
Fail to Reject the Null Hypothesis
Reject
Critical Value (+)
Fail to Reject the Null Hypothesis
Reject
Reject
Critical Value (-)
Critical Value (+)

Figure-7: Acceptance and Rejection Region (Source: Google)

What is the intuitive justification of such decision rule?

Consider H_A: $\mu < \mu_0$. If the population mean is truly μ_0 (or greater than μ_0), the probability of being in the left hand (red) region is very small. But the sample that has been collected, causes the sample statistic to fall in that (unlikely) zone. Hence the decision is that, the null hypothesis looks untenable, in the light of the sample.

However, it is not impossible (i.e. probability is not 0, though small), that even when the population mean is truly μ_0, this particular sample may still be realized. Hence α is the probability of rejecting the null hypothesis, when it is true or probability of type I error.

Why type I error probability cannot be taken to be 0?

Theoretically, all values of the test statistic may be realized when H_0 is true, albeit values with negligible probability. If probability of type I error is set at 0, never can H_0 be rejected.

The above approach of decision making is known as rejection region approach.

An alternative approach to decision making is through p-value. P-value of a test is defined as the probability of observing a more extreme value than the observed test statistic. Extreme value depends on the type of alternative.

If H_A: $\mu < \mu_{0,}$ the p-value of the test is Prob[t(df) < observed value of test statistic]

If H_A: $\mu > \mu_{0,}$ the p-value of the test is Prob[t(df) > observed value of test statistic]

If H_A: $\mu \neq \mu_0$, the p-value of the test is Prob[t(df) > observed value of test statistic or

t(df) < (-ve) observed value of test statistic]. Since t-distribution is symmetric, p-value may also be defined as 2[Prob[t(df) > observed value of test statistic]]

Once p-value is determined, the final decision may be taken by comparing with a given significance level α. Typically the significance level is fixed at 5% or at 1% level.

3.4 Form of Test Statistics used Frequently

The following table specifies test statistics for the different hypothesis testing problems that will be discussed in subsequent sections. In addition to these, there are innumerable statistics for many different types of testing problems. Some of them will be discussed in other monographs. For the rest, reference books need to be consulted.

Hypothesis Testing Problem	Test Statistic	Notations Explained
Test for population mean H_0: $\mu = \mu_0$	When population std dev σ known: $\frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma} \sim N(0, 1)$	n: sample size $\bar{x}$: sample mean s : sample standard deviation
	When population std dev σ unknown: $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s} \sim t(n-1)$	
Test for population proportion H_0: $\pi = \pi_0$	$\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0, 1)$	p: sample proportion n: sample size
Test for equality of two population means H_0: $\mu_1 = \mu_2$	When the two populations are independent and the two std devs may be assumed to be equal: $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}} \sim t(n_1 + n_2 - 2)$	$\bar{x}_1$: sample mean of first sample n_1: sample size of first sample s_1: sample std dev of first sample $\bar{x}_2$: sample mean of second sample n_2: sample size of second sample s_2: sample std dev of second sample s_p^2: pooled variance s_p: pooled std dev $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ d = difference of two values in one unit s_d = sample std dev of the differences n: sample size of paired sample
	When the two populations are independent but the population std devs may not be assumed to be equal $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}} \sim t$ Degrees of freedom expression is complex, hence not given explicitly	
	When the two populations are not independent $\frac{\sqrt{n}(\bar{d})}{s_d} \sim t(n-1)$	

Test for equality of two population proportions H$_0$: $\pi_1 = \pi_2$	$\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p^*(1-p^*)\,(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0, 1)$	n_1: sample size of first sample n_2: sample size of second sample p_1: proportion in first sample $= \frac{x_1}{n_1}$ p_2: proportion in second sample $= \frac{x_2}{n_2}$ $p^* = \frac{x_1 + x_2}{n_1 + n_2}$
Test for equality of two population variances H$_0$: $\sigma_1^2 = \sigma_2^2$	$\frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 2)$ F distribution will be introduced later	s_1^2: sample variance of first sample s_2^2: sample variance of second sample
In a contingency (cross-classified) table H$_0$: The row and column variables are independent	$\sum_{\text{all cells}} \frac{(\text{obs frq} - \text{exp frq})^2}{\text{exp frq}}$ $\sim \chi^2_{(r-1)(c-1)}$ X^2 distribution will be introduced later	r: number of rows in the contingency table c: number of columns in the contingency table

Table-2: Test Statistics of various hypothesis

3.5 Connection between Hypothesis Testing and Confidence Interval

Consider Figure 7. Does this indicate any connection between hypothesis testing for two-sided alternative and confidence interval? The total space is partitioned into two: the rejection region (where the null hypothesis is rejected) and the space at the centre, where the null hypothesis cannot be rejected. The probability of being in this region, under H$_0$ is 95% and the probability of being in the rejection region is 5%. This observation suggests that the technique of hypothesis testing against a two-sided alternative and constructing a confidence interval is actually approaching the same problem from two opposite sides.

If a 100(1 – α)% (e.g. 95%) confidence interval of a population parameter includes the null value of the parameter, then the null hypothesis cannot be rejected at 100α% (e.g. 5%) level of significance. If the 100(1 – α)% (e.g. 95%) confidence interval of a population parameter does not include the null value of the parameter, then the null hypothesis is rejected at 100α% (e.g. 5%) level of significance.

If the null hypothesis is rejected in favour of a two-sided alternative, a natural extension of the problem is to provide an estimate of the unknown parameter. We have discussed in Section 2.2 why an interval estimate is better than a point estimate.

Confidence intervals are not associated with one-sided hypothesis testing problems.

In the next section many and varied examples are provided with extensive explanations.

Important Note: The value of the test statistic does not change for a two-tailed test or a one-tailed test. Only the p-value or the probability value associated with the test statistic changes. As a default, while conducting hypothesis test, **Python will calculate the results considering two tailed test only**. When p-value for one tail test is to be computed, the python p-value needs to be divided by 2.

i.e **P-Value (One tailed test) = P-Value(Two tailed test)/2**

4.Miscellaneous Illustrations and Examples

PYTHON HAS SOME LIMITATIONS FOR DOING HYPOTHESIS TESTING. SINCE IT DOES NOT ALLOW TO SPECIFY THE ALTERNATIVE AND ALWAYS ASSUMES THAT THE ALTERNATIVE IS BOTH SIDED, **P-VALUE FOR ONE-SIDED ALTERNATIVE NEEDS TO BE COMPUTED MANUALLY.**

Example 1. A certain food aggregator ZYX is facing stiff competition from its main rival SWG during Corona period. To retain business, ZYX is advertising that, within a radius of 5 km from the restaurant where the order is placed, it can deliver in 40 minutes or less on the average. The delivery times in minutes of 30 deliveries are given in the file fastfood.csv. Assuming the delivery distribution is approximately normal, is there enough evidence that ZYX's claim is true?

It is clearly a one-sided hypothesis testing problem, concerning population mean μ, the average delivery time.

The null hypothesis may be a prevalent opinion, or an assumption. In this case the opinion expressed in advertisement is that average delivery time is 40 minutes or less. Hence

$$H_0\text{: } \mu \leq 40 \text{ against } H_A\text{: } \mu > 40$$

Let us fix the significance level at $\alpha = 0.05$.

Note: Adjustment required to convert 2-sided p-value to the appropriate one-sided p-value.

```
import pandas as pd
FastFood=pd.read_csv('FastFood.csv')
from scipy import stats
from scipy.stats import ttest_1samp
mean=FastFood.Time.mean()
print("The mean Delivery Time is {}".format(round(mean,3)))
            >> The mean Delivery Time is 42.3

t_statistics, p_value1 = stats.ttest_1samp(FastFood['Time'], popmean=40)
p_value =p_value1/2           # converting two tailed test value to one tailed test (right tail)

print("t_statistic= {} and pvalue= {}".format(round(t_statistics,3),round(p_value,7)))

          >>t_statistic= 5.581 and pvalue= 2.5e-06

alpha= 0.05
if (p_value/2 < alpha):
    print("We Reject Null Hypothesis. True mean is greater than {}".format(40))
else:
    print("We Fail to Reject Null Hypothesis")
        >>We Reject Null Hypothesis. True mean is greater than 40
```

For the sample average delivery time is computed as is $\bar{x}$ = 42.3. The population stdev is not known. Hence the sample standard error is estimated from data and $\frac{\sqrt{n}(\,\bar{x}-\mu_0)}{s}$ is used as the test statistic. Note that sample size *n* = 30 and the df associated with the t-distribution is *n* – 1 = 29. Numerical value of the test statistic is 5.581 and the p-value is $2.5\text{x}10^{-6}$.. The rejection region is on the right-hand side. The rejection rule in this case is: Reject H_0, if the numerical value of the test statistic is too large.

The small p-value indicates that, had H_0 been true, the probability of observing this value of the t-statistic is extremely small.

Conclusion: The null hypothesis is rejected at 5% level of significance. Hence, at 5% level of significance, there is not enough evidence that the mean delivery time within 5 km radius is indeed 40 min or less, as claimed by ZYX in their advertisement.

Example 2. Do you feel frustrated when your favourite TV program is interrupted by commercials? Faced by criticisms in consumers' forums, major channels published a report that as per their policy, maximum duration of commercials is 7 minutes per 30-minute slot. That means in that slot, actual programming minutes is 23 or more. Consumers' Forum took an independent sample of 20 30-minute programs and measured the number of programming minutes in each of them. The data is given in program.csv. Assuming that the population distribution is normal, is there enough evidence that the channels' official policy is being followed?

It is clearly a one-sided hypothesis testing problem, concerning population mean μ, the average programming minute in a 30-minute slot on TV.

The null hypothesis in this case is a stated policy that the actual programming minutes is 23 or more. Hence

$$H_0: \mu \geq 23 \text{ against } H_A: \mu < 23$$

```
import pandas as pd
Program=pd.read_csv('Program.csv')
Mean=Program.Minutes.mean()
print("The mean Program Minutes is {}".format(round(Mean,3)))
        >>The mean Program Minutes is 22.5
from scipy import stats
t_statistics, p_value1  = stats.ttest_1samp(Program['Minutes'], popmean=23)
```

```
p_value =p_value1/2  # converting two tailed test value to one tailed test (left tail )
print("t_statistic= {} and pvalue= {}".format(round(t_statistics,3),round(p_value,3)))
                    >>t_statistic= -1.998 and pvalue= 0.03
alpha= 0.05
if (p_value < alpha):
   print("Reject Null Hypothesis. True Mean is less than {}".format(23))
else:
   print("We fail to Reject Null Hypothesis")
   >> Reject Null Hypothesis. True Mean is less than 23
```

For the sample average programming minute is computed as is $\bar{x} = 22.5$. The population stdev is not known. Hence the sample standard error is estimated from data and $\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$ is used as the test statistic. Note that sample size $n = 20$ and the df associated with the t-distribution is $n - 1 = 19$. Numerical value of the test statistic is -1.998 and the p-value is 0.03. The rejection region is on the left-hand side. The rejection rule in this case is: Reject H_0, if the numerical value of the test statistic is too small.

If the level of significance is fixed at $\alpha = 0.05$, p-value is less than α. Hence the null hypothesis is rejected.

Conclusion: At 5% level of significance there does not seem enough evidence that the stated policy is being followed.

Suppose now the level of significance is fixed at $\alpha = 0.01$, p-value is NOT less than α. Hence the null hypothesis is NOT rejected.

Conclusion: At 1% level of significance there is not enough evidence to claim that the stated policy is not followed.

The final conclusion depends on the assumed level of significance.

Example 3(a). The cost of a certain high-quality diamond is $5600 per carat in mid-west USA. A jeweller contacted New York City jewellers to understand whether mean price there differs significantly from mid-west. He contacted 25 jewellers and the price per carat for the same type of diamond is given in diamonds.csv. Is there any reason to believe that price of diamond is different in two places?

This is a two-sided alternative since the jeweller wants to check whether prices *differ* in two places.

The null value is given to be $5600. Hence

$$H_0: \mu = 5600 \text{ against } H_A: \mu \neq 5600$$

Let us fix the level of significance at $\alpha = 0.05$

```
import pandas as pd
from scipy import stats
Diamonds=pd.read_csv("Diamond.csv")
Mean=Diamonds.Price.mean()
print("The mean of Price is {}".format(Mean))
        >> The mean of Price is 5835.0
t_statistics, p_value = stats.ttest_1samp(Diamonds['Price'], popmean=5600)
print("t_statistic= {} and pvalue= {}".format(round(t_statistic,3),round(p_value,3)))
        >> t_statistic= 2.259 and pvalue= 0.033

alpha= 0.05
if (p_value < alpha):
   print("We Reject Null Hypothesis. True Mean is not equal to {}".format(5600))
else:
   print("We fail to Reject Null Hypothesis")
        >> We Reject Null Hypothesis. True Mean is not equal to 5600.
```

The rejection region is both-sided. The null hypothesis is rejected if either the value of the t-statistic is too large or too small. In this case the sample size is *n* = 25 and the df associated with the t-distribution is 24. The p-value is 0.033. At 5% level of significance, the null hypothesis is rejected.

Conclusion: At 5% level average price of diamonds differs significantly between mid-west and NYC.

If α is fixed at 0.01, what would be the conclusion? Why?

Example 3(b). Construct a 95% confidence interval for average price of diamonds in NYC. Based on this confidence interval, can you conclude whether H_0: μ = 5600 may be rejected in favour of H_A: μ ≠ 5600.

```
import numpy as np
import scipy.stats


def mean_confidence_interval(data, confidence):
   a = 1.0 * np.array(data)
   n = len(a)
   m, se = np.mean(a), scipy.stats.sem(a)
   h = se * scipy.stats.t.ppf((1 + confidence) / 2., n-1)
   return print("The lower confidence interval value is {} and the upper confidence interval
value is {}.".format(round(m-h,2),round(m+h,2)))

mean_confidence_interval(Diamonds.Price,confidence=0.95)
```

>>The lower confidence interval value is 5620.31 and the upper confidence interval value is 6049.69.

Note that the 95% confidence interval for μ, the mean price of diamond, is [5620.31, 6049.69]. This interval does not contain the null value 5600. Hence, the null hypothesis will be rejected at 5% level of significance.

But this does not provide any insight into the numerical value of the test statistic, nor the p-value.

Example 3(c). Construct a 99% confidence interval for average price of diamonds in NYC. Based on this confidence interval, can you conclude whether H_0: $\mu = 5600$ may be rejected in favour of H_A: $\mu \neq 5600$.

```
mean_confidence_interval(Diamonds.Price,confidence=0.99)
```

>> The lower confidence interval value is 5544.05 and the upper confidence interval value is 6125.95.

The 99% confidence interval for μ, the mean price of diamond, is [5544.05, 6125.95]. This interval contains the null value 5600. Hence the null hypothesis H_0: $\mu = 5600$ cannot be rejected.

Note also that the p-value of the test is 0.03. If the level of significance α is set at 0.01, then the p-value is not less than the significance level.

Points to note from Examples 1 – 3 (one-sample problems)

(1) Given a set of data, the value of the test statistic and the p-value of the test is fixed
(2) Final conclusion regarding rejection of the null hypothesis depends on the chosen level of significance
(3) For a two-sided alternative and a given α, confidence interval and testing of hypothesis has a one-to-one correspondence

Example 4. SAT verbal scores of two groups of students are given in SATVerbal.csv. The first group, College, contains scores of students whose parents have at least a bachelor's degree and the second group, High School, contains scores of students whose parents do not have any college degree. The Education Department is interested to know whether the sample data support the hypothesis that students show a higher population mean verbal score on SAT if their parents attain a higher level of education.

This is a two-sample problem where the College and High School populations are independent. The two sample sizes are also different. The testing problem may be set up as

H_0: $\mu_C = \mu_{HS}$ against H_A: $\mu_C > \mu_{HS}$ or, equivalently, H_0: $\mu_C \leq \mu_{HS}$ against H_A: $\mu_C > \mu_{HS}$

Before proceeding any further, the standard deviations need to be compared.

$s_C = 59.4$ and $s_{HS} = 51.7$. Their ratio is 1.15

For now, an empirical rule is applied: If the ratio of two sample standard deviations is between 0.7 and 1.4, the population standard deviations may be assumed to be equal.

Hence the appropriate test statistic is $\frac{(\bar{x}_C - \bar{x}_{HS}) - (\mu_C - \mu_{HS})}{s_p\sqrt{(\frac{1}{n_C} + \frac{1}{n_{HS}})}} \sim t(n_C + n_{HS} - 2)$

Note that, under H_0, $(\mu_C - \mu_{HS}) = 0$

The alternative is right-sided. The null hypothesis will be rejected if the value of the test statistic is too large.

```
import pandas as pd
from scipy import stats
SATVerbal=pd.read_csv('SATVerbal.csv')
X1_mean=SATVerbal["College"].mean()
X2_mean=SATVerbal["High School"].mean()

print("The mean Verbal score of College students is {} wheras the mean score of High Schoo
l students is {}".format(X1_mean,X2_mean))

        >>The mean Verbal score of College students is 525.0 wheras the mean score of High
School students is 487.0

p_value=p_value1/2                    # Right Tailed test
print("t_statistic= {} and pvalue= {}".format(round(t_statistic,3),round(p_value,3)))
                >> t_statistic= 1.767 and pvalue= 0.044
alpha =0.05
if (p_value < alpha):
   print("We Reject Null Hypothesis. True difference in means is greater than {}".format(0))
else:
   print("We fail to Reject Null Hypothesis")
                >> We Reject Null Hypothesis. True difference in means is greater than 0
```

The p-value of the test is 0.044. If level of significance is fixed at 5%, then the null hypothesis is rejected.

Conclusion: At 5% level of significance it may be concluded that for the students whose parents have college education mean of SAT verbal score is higher than the other group.

Now fix α = 1%. Does that change the conclusion?

Alternative formulation:

H_0: $\mu_{HS} = \mu_C$ against H_A: $\mu_{HS} < \mu_C$ or, equivalently, H_0: $\mu_{HS} \geq \mu_C$ against H_A: $\mu_{HS} < \mu_C$

Null hypothesis will be rejected if the test statistic is too small.

Complete the test.

Note that the p-value does not change on changing the alpha. Here the p-value of the test is 0.044. If level of significance is fixed at 1%, then we fail to reject the Null Hypothesis..

Example 5. Typical prices of single-family homes in Florida are given for a sample of 15 metropolitan areas (in 1000 USD) for 2002 and 2003. Data is provided in Florida.csv. Is there any significant evidence that the increase in price is more than 10,000 USD in one year?

This is a problem of paired sample, since two observations are taken on each sampled unit.

Define $\mu_d = \mu_{2003} - \mu_{2002}$. Noting that the prices are given in 1000 USD, the null and the alternative hypotheses are formed as H_0: $\mu_d \leq 10$ against H_A: $\mu_d > 10$

```
import pandas as pd
Florida=pd.read_csv('Florida.csv')
diff=Florida.Jan_2003-Florida.Jan_2002
Florida["diff"]=diff
Mean=Florida["diff"].mean()
print("The Mean of the difference in house price from 2003 to 2002 is {}.".format(Mean)
        > The Mean of the difference in house price from 2003 to 2002 is 15.0.
from scipy import stats
t_statistic, p_value1 = stats.ttest_1samp(Florida["diff"],popmean=10)
p_value = p_value1/2  # Right Tail test
print("t_statistic= {} and pvalue= {}".format(round(t_statistic,3),round(p_value,3)))

        >>t_statistic= 1.696 and pvalue= 0.056

Alpha=0.05
if (p_value/< alpha):
   print("We Reject Null Hypothesis, True difference in means is greater than {}".format(10))
else:
   print("We fail to Reject Null Hypothesis")
        >> We Reject Null Hypothesis, True difference in means is greater than 10
```

P-value of the test is 0.056. If the significance level is fixed at 5%, the null hypothesis is not rejected (even though it is on the borderline)

Conclusion: At 5% level of significance it may be concluded that there is no evidence that the price increase is more than 10000 USD from Jan 2002 to Jan 2003.

If α is fixed at 1%, would there be any change in the conclusion?

Alternative formulation:

Define $\mu_d^* = \mu_{2002} - \mu_{2003}$ and formulate the hypothesis testing problem accordingly. Solve the same.

Example 6. In the lockdown period, because of working from home and increased screen time, many opted for listening to FM Radio for entertainment rather than watching Cable TV. An advertisement agency collected data (TVRadio.csv) on both types of users and would like to know whether there is any difference between TV and Radio usage.

This is a two-sample problem with sample sizes being equal. (*For a paired sample, both sets of observations are taken on the same sampled unit. There is no evidence here that the TV time and radio time are measured on the same person*)

H_0: $\mu_{TV} = \mu_{Radio}$ against H_A: $\mu_{TV} \neq \mu_{Radio}$ or, equivalently, H_0: $\mu_{TV} - \mu_{Radio} = 0$ against H_A: $\mu_{TV} - \mu_{Radio} \neq 0$

The population stdevs are assumed to be equal. (Verify this assumption yourselves)

```
import pandas as pd
from scipy import stats
TVRadio=pd.read_csv('TVRadio.csv')
X1_mean=TVRadio["Cable_TV"].mean()
X2_mean=TVRadio["FM_Radio"].mean()

print("The mean usage of Cable_TV is {} wheras the mean usage of FM_Radio is {}".format
(X1_mean,X2_mean))
         >> The mean usage of Cable_TV is 18.8 wheras the mean usage of FM_Radio is 20.0

t_statistic, p_value  = stats.ttest_ind(TVRadio['Cable_TV'], TVRadio['FM_Radio'], equal_var
=True)
print("t_statistic= {} and pvalue= {}".format(round(t_statistic,3),round(p_value,3)))
          >> t_statistic= -0.606 and pvalue= 0.549
```

P-value of this test is 0.55. Hence the null hypothesis cannot be rejected.

Conclusion: At 5% level of significance it may be concluded that there is no evidence that during the lockdown period TV time and radio time differ significantly.

Points to note from Examples 4 – 6 (two-sample problems)

(1) Decide first whether the problem is for paired sample or independent samples. If the sample sizes are different, the problem *cannot* be a paired sample
(2) For paired sample, two sets of observations must come from each unit sampled
(3) For two independent sample problem, check whether the population std deviations may be assumed to be equal. Apply empirical rule or do a formal test of hypothesis, which is discussed subsequently.
(4) Instruct the software to do an appropriate test

Example 7. Among professional drivers sleep deprivation is a serious work hazard. A research paper claims that more than half of the night-time drivers drive even when they feel drowsy. A sample of 500 night-time drivers were asked whether they ever driven when drowsy, and the data is given in Drowsy.csv. Is there evidence that the research paper's claim is justified?

Clearly this is a test for population proportions. The null and alternative hypotheses may be defined as H_0: $\pi \leq 0.50$ against H_A: $\pi > 0.50$

```
Drowsy=pd.read_csv("Drowsy.csv")
Drowsy["Drive While Drowsy?"].replace(["Yes","No"],[1,0],inplace=True)
a=Drowsy.sum()
Proportion = a/Drowsy.size
Proportion
                    >> Drive While Drowsy?    0.464   # Proportion of Drowsy/ Total Sample Size

from statsmodels.stats.proportion import proportions_ztest
z_stat,p_value=proportions_ztest(232, 500, value=0.5, alternative='larger', prop_var=False)
print("t_statistic= {} and pvalue= {}".format(round(z_stat,3),round(p_value,3)))
          >> t_statistic= -1.614 and pvalue= 0.947
```

Sample proportion is 46.4% and the p-value of the test is 0.947. As P_value is greater than 0.05, Hence the null hypothesis cannot be rejected.

Conclusion: At 5% level of significance it may not be concluded that more than 50% of the night-time drivers feel drowsy while driving.

Note that the value of the sample proportion is less than the null value of the parameter. Since the alternative is greater than type, the null would be rejected if the value of the test statistic is too large, which is not possible given the sample values. The alternative does not depend on the sample value. Both alternative and significance level need to be fixed before the random sample is taken.

4.1 Chi-square Test for Independence or Goodness of Fit Statistics

So far the hypothesis testing problems considered are all parametric problems. The null and the alternative hypotheses are statements about the population parameters, such as population mean, population variance or population proportions. Now consider the example below.

	Male	Female	Total
Smoker	120	100	220
Non-smoker	60	125	185
Total	180	225	405

Table-3 (2x2 Contingency table)

This is a 2x2 cross-classification or contingency table. This describes two attributes at two levels each and the number of observations at each cell. An *r* x *c* contingency will be defined with two attributes, one at *r* levels and the other at *c* levels.

Here the hypothesis of interest is to see whether the two attributes are independent. In Table 2, one may want to know whether smoking is independent of gender.

Formally

H_0: Smoking and gender is independent against H_A: Smoking and gender is not independent

Note that, the alternative cannot be of less-than or greater-than type.

For a 2x2 table, this problem reduces to a comparison of two independent proportions. However, for an *r* x *c* table, the test statistic has a definitive form, called a Goodness-of-Fit statistic. The form of the statistic is shown below.

$$\sum\nolimits_{\text{all cells}} \frac{(\text{obs frq} - \text{exp frq})^2}{\text{exp frq}} \quad \sim \chi^2_{(r-1)(c-1)}$$

For each cell, the observed frequencies are as derived from the sample. The expected frequencies are computed under the null hypothesis of independence.

	C1	C2	Total
R1	n_{11}	n_{12}	n_{10}
R2	n_{21}	n_{22}	n_{20}
Total	n_{01}	n_{02}	$n_{00} = n$

Table-4 (2x2 Contingency table- cell reference)

The formula for expected frequency in the (i,j)th cell is $\frac{n_{i0}n_{0j}}{n}$. The squared differences between observed and expected frequencies are small if H_0 holds, it will be large if H_0 is not true. Hence the rejection region in this case is **always on the right hand side**.

The distribution of the test statistic under the null hypothesis follows a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom. For a 2x2 table, the df = 1

This test is applicable and works well under certain conditions. It is clear from the definition of the statistic that if expected frequency is 0, the value of the statistic is infinity. If the expected frequency is very small, even then the numerical value of the test statistic will be unstable. The rule of thumb is, therefore, that for expected frequency less than 5, caution must be used, especially if the p-value of the test is small.

4.1.1 χ^2 distribution

The χ^2 distribution is a non-symmetric, positively skewed distribution which can take only positive values and is controlled by one degrees of freedom.

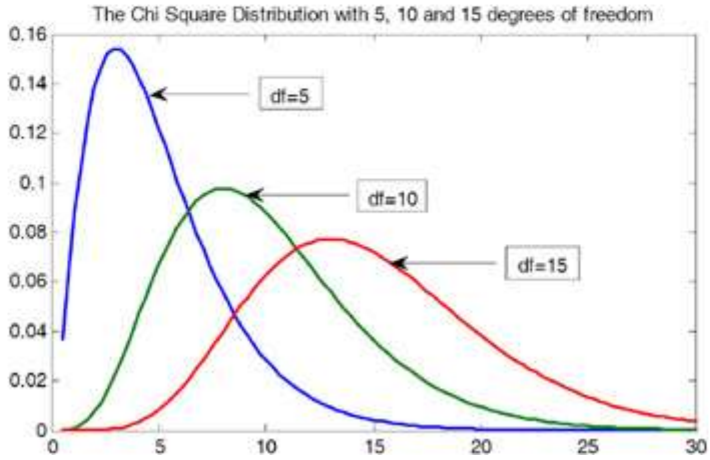


Figure. 8: χ^2 distribution for various degrees of freedom (Source: Google)

Since the χ^2 distribution comprises of only squared quantities, the rejection region is only at the right-hand tail area.

Example 8. The following table summarizes beverage preference (Beverage.csv) across different age-groups. Does beverage preference depend on age?

	Beverage Preference		
Age	Tea/Coffee	Soft Drink	Others
21 - 34	25	90	20
35 - 55	40	35	25
> 55	24	15	30

Table-5 (Beverage Preference)

The null and the alternative hypotheses are

H_0: Beverage preference is independent of age against H_A: Beverage preference depends on age

Since this is a 3x3 contingency table, the df of the chi-square distribution is (3 – 1)(3 – 1) = 4

```
import pandas as pd
```

```
Beverage=pd.read_csv('Beverage.csv')
Beverage=Beverage.drop("Age",axis=1)
from scipy.stats import chi2
from scipy.stats import chi2_contingency
chi2, p, dof, exp_freq = chi2_contingency(Beverage)
print("chi2= {} , p_value= {}, dof={}".format(round(chi2,3),round(p,3), dof))
        >> chi2= 49.158 , p_value= 0.0, dof=4
```

The p-value of the test is very small. Hence the null hypotheses is rejected.
Conclusion: At 5% level of significance it may be concluded that beverage preference is not independent of age.

Example 9. The following table summarizes the number of hypertensive persons among males and females (Hypertension.csv) sampled from a certain population. Is there any evidence that the distribution is different among males and females?

	Male	Female
Normal	35	40
High normal	84	65
Mild hypertensive	73	75
Moderate hypertensive	91	87
Severe hypertensive	60	50
Total	343	317

Table-6 Distribution of Hypertension among males and females

This is also a problem of goodness of fit test. In fact, goodness of fit test statistic has an extensive number of applications and testing equivalency of two (or more) distributions is also included among them.

H_0: Distribution of hypertension is identical among males and females against H_A: Distribution of hypertension is not equivalent among them

```
import pandas as pd

Hypertension=pd.read_csv('Hypertension.csv')
Hypertension= Hypertension.drop("Type",axis=1)
chi2, p, dof, expected = chi2_contingency(Hypertension)
print("chi2= {} , p_value= {}, dof={}".format(chi2, p, dof))
        >> chi2 = 2.7622, df = 4, p_value = 0.5984
```

The p-value of the test is quite large.

Conclusion: At 5% level of significance it may be concluded that null hypothesis of identical distribution of hypertension in both groups, cannot be rejected.

4.2 Testing whether two population variances are equal

(***This part may be skipped for now. But it is strongly recommended that you familiarize yourself with this procedure before you tackle ANOVA***)

It has been already noted that before a test of two population means is facilitated, it needs to be checked whether two population variances are equal. (*In fact, when equality of more than two population means is to be considered, it needs to be checked whether the population variances are all equal. This will be taken up in a subsequence monograph on ANOVA*)

The test of equality of two population variances depends on their ratio.

H_0: $\sigma_1^2 = \sigma_2^2$ against H_A: $\sigma_1^2 \neq \sigma_2^2$ may equivalently written as H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ against H_0: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

The unbiased estimates of the population variances are the sample variances. These statistics are used in constructing the appropriate test statistic.

Define $$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Under H_0, $$F = s_1^2/s_2^2$$

H_0 is rejected if the numerical value of F is too large or too small.

4.2.1 F distribution

F distribution is a non-symmetric, positively skewed distribution which can take only positive values and is controlled by two degrees of freedom.

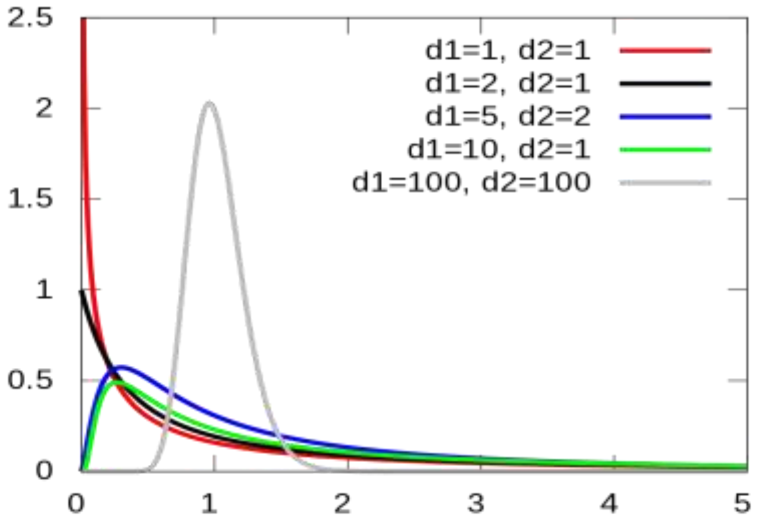


Figure 9: F distribution for various degrees of freedom (Source: Google)

F distribution originates from the ratio of two squared quantities, adjusted by associated denominator. Recall that the sample variance may be written as $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$, which is an unbiased estimate of the population variance. The associated denominator here is (n – 1).

In the expression of F in Section 4.1, the first degrees of freedom is the sample size (less 1) from the first sample and the second degrees of freedom is the sample size (less 1) from the second sample. Since variances are squared quantities, and hence always positive, the numerical value of the F distribution cannot be negative.

The rejection region may fall in either tail of the distribution. For a right-sided alternative, at α level of significance, the rejection region will be in the right tail. For a left-sided alternative, the rejection region will be in the left tail. Note that, F distribution being non-symmetric, the rejection regions will not be symmetric (unlike the normal and t-distribution). For the inequality alternative, the rejection region will be on both sides.

Example 10. The variance of a process is an important quality of the process. A large variance implies that the process needs better control and there is opportunity to improve. The data (Bags.csv) includes weights for two different sets of bags manufactured from two different machines. Consider a statistical test to determine whether there is a significant difference between the bag weights for the two machines. Which machine, if any, provides a greater opportunity for quality improvement?

The null hypothesis specifies that the ratio of the two population variances are equal. If the null hypothesis is rejected, then one machine's variance is certainly greater than the other, and that particular process provides a greater opportunity for improvement.

H$_0$: $\sigma_1^2 = \sigma_2^2$ against H$_A$: $\sigma_1^2 \neq \sigma_2^2$ or, equivalently H$_0$: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ against H$_0$: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

```
import numpy as np
import scipy
def f_test(x, y):
    x = np.array(x)
    y = np.array(y)
    f = np.var(x, ddof=1)/np.var(y, ddof=1) #calculate F test statistic
    dfn = x.size-1 #define degrees of freedom numerator
    dfd = y.size-1 #define degrees of freedom denominator
    p = (1-scipy.stats.f.cdf(f, dfn, dfd)) #find p-value of F test statistic
    p1 = p*2      # Converting one-tail to two-tail test
    return(print("f_stat is {} and p_value is {}" .format(round(f,3),round(p,8))))

#perform F-test
f_test(Bags.dropna()['Machine 1'], Bags.dropna()['Machine 2'])
```

```
➢   f_stat is 8.895 and p_value is 5.1e-06
```

The p-value of the test is 5.1x10^{-6}, much smaller compared to and standard significance level, α = 0.05 or even α = 0.01. The null hypothesis of equality is rejected.

Please note that we have created a user defined function for one tailed test. To convert the o-value of one tailed to two-tailed test, we will multiply p_value by 2.

Conclusion: At 5% level of significance it may be concluded that the process variance of the two machines are not equal.

The ratio of variances is 8.89, which indicates that the variance of Machine 1 is more than 8 times than that from Machine 2. Therefore, it may be concluded that Machine 1 provides more opportunity for improvement.

5.Further Applications of Statistical Inference

In the previous sections the simplest problems on statistical inference have been dealt with. The two main divisions of inference, namely construction of confidence interval and hypothesis testing will be applicable in all subsequent topics covered.

Among the topics covered in subsequent modules, are Analysis of Variance and various model building procedures. In all of these, many different types of hypotheses will be tested. Take for example regression model building. The assumed model is fitted to the data and the associated hypothesis tested is that, the response does not depend on the predictors.

All model building procedures involve estimation of model parameters. Associated with their point estimations, standard errors are also estimated so that the confidence intervals may also be constructed. In time series models, confidence intervals for forecasted values are computed by default in many applications.

References:

Stat 500 Applied Statistics. Penn State University sites:

https://online.stat.psu.edu/stat500/lesson/4

https://online.stat.psu.edu/stat500/lesson/5

https://online.stat.psu.edu/stat500/lesson/6a

https://online.stat.psu.edu/stat500/lesson/6b

https://online.stat.psu.edu/stat500/lesson/7

https://online.stat.psu.edu/stat500/lesson/8

Statistical Learning Center videos on YouTube:
https://www.youtube.com/watch?v=D6oTsaSC5NM&list=PLJUchALB4eUOxyE8LB13M9Nv3rEYoqWLD&index=3

https://scipy-lectures.org/packages/statistics/index.html