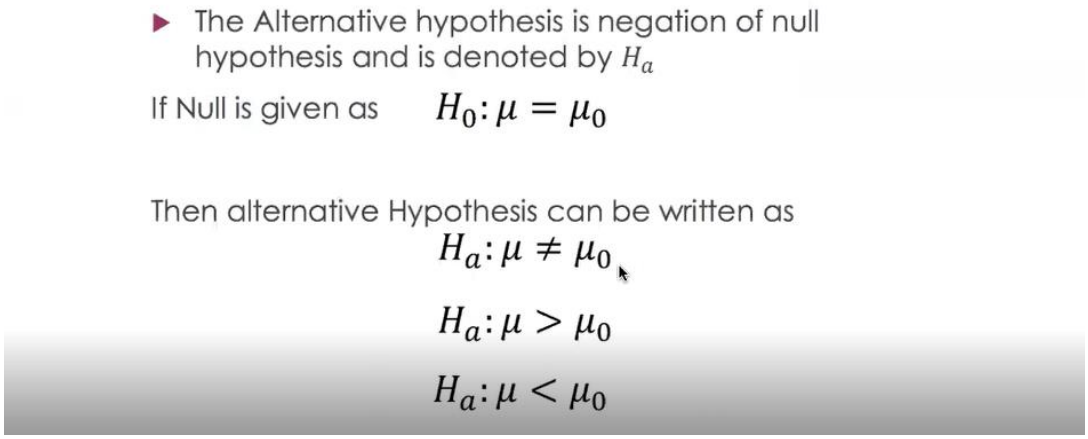
Hypothesis

What is a Hypothesis?

An assumption is called a hypothesis, and the statistical tests used for this are called statistical hypothesis tests.

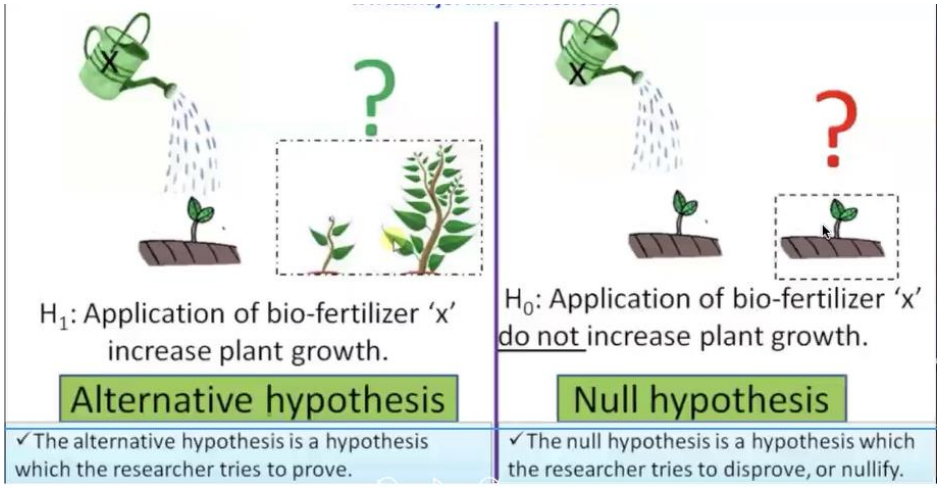
Two hypotheses:

- H_0 is the Null Hypothesis.
- H_1 or H_A is the alternative Hypothesis.



Note:

- The Null Hypothesis assumes that there is a "NO" difference between two sets of values.
- The alternative hypothesis used in hypothesis testing is contrary to the null hypothesis.



Hypothesis Testing's Mechanism

Analyze the performance of the students of the university on an overall basis.

Here, the population mean is H_0 and the percentage is 75.

And H_1/H_A is the population mean

H_0→ when μ_0 is equal to 75 percent

H_1→ when μ_0 is not equal to 75 percent

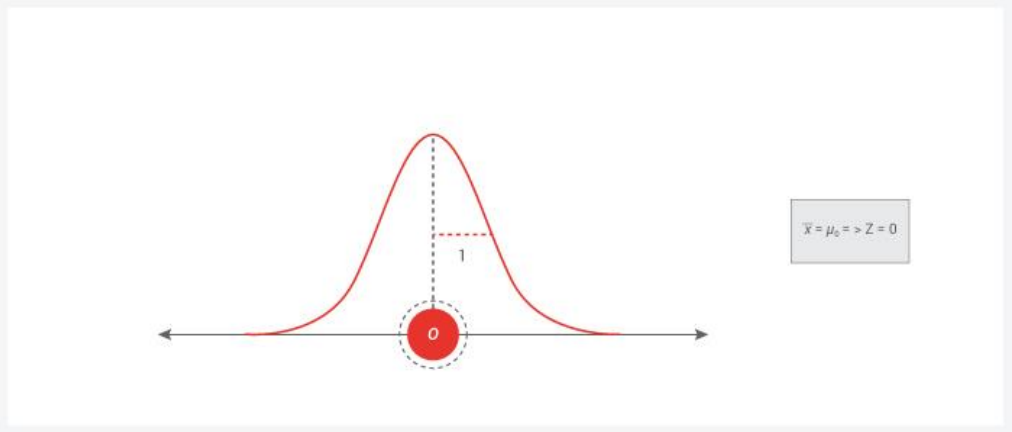
The formula of Z-test :

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

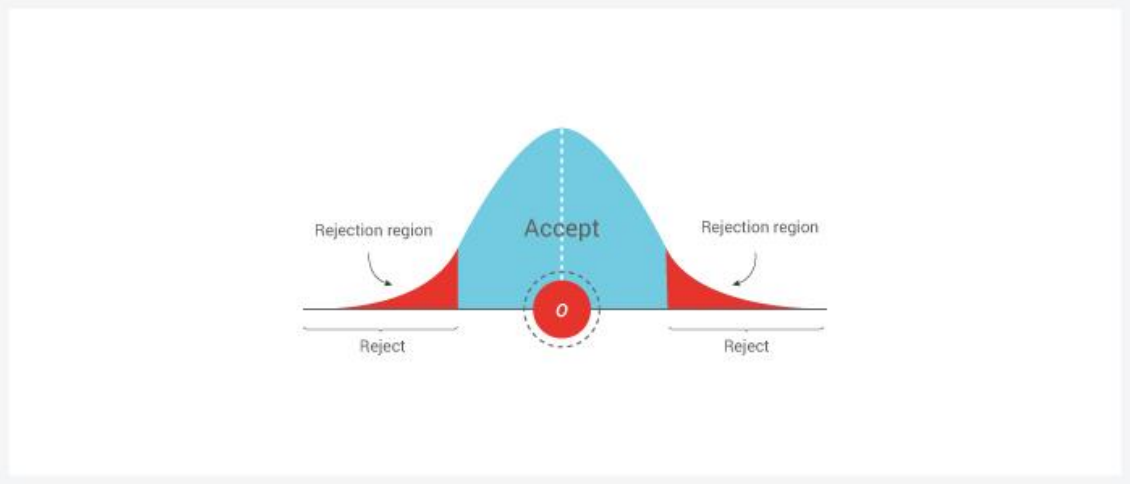
Here $\bar{x}$ is the sample mean, 'μ' is hypothesized mean, 's' is the standard error, and 'n' is the sample size.

So if the sample mean is close enough to hypothesized mean, then Z will be close to 0.

In this case, we will accept the Null Hypothesis (As demonstrated in the image below). Otherwise, we will reject it.



If the z value lies in middle area, then we assume the null hypothesis; otherwise, we reject the null hypothesis.



Hypothesis Testing Example

Example

- A Criminal trial is an example of hypothesis testing without the statistics.
- In a trial, a judicial system must decide between 2 hypotheses. The null hypothesis is

    - H_0: The defendant is not guilty.
- Another hypothesis is
    - H_1: The defendant is not innocent(guilty).

The jury doesn't know which is the correct hypothesis. They must decide based on the evidence presented.

There are two scenarios:

1. There is enough evidence to support the other hypothesis.
2. There is not enough evidence to support the other hypothesis.

P-Value

- It is is the level of marginal significance representing a given event's probability of occurrence.
- P-Value tables or spreadsheet/statistical software can be used to calculate the p-value.
- The smaller p-value indicates stronger evidence favoring the alternative hypothesis.

A p-value is a number between 0 and1:

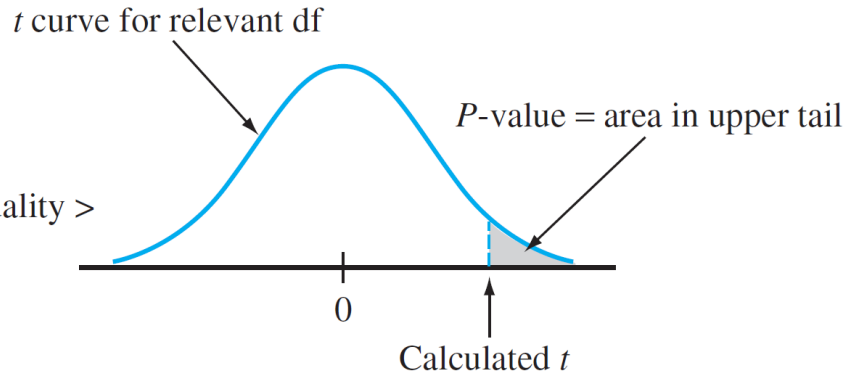
- A p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis; the null hypothesis is rejected.
- A p-value (> 0.05) indicates weak evidence against the null hypothesis; the null hypothesis is not rejected.
- p-values very close to the cut-off (0.05).

P-Values for t-Tests

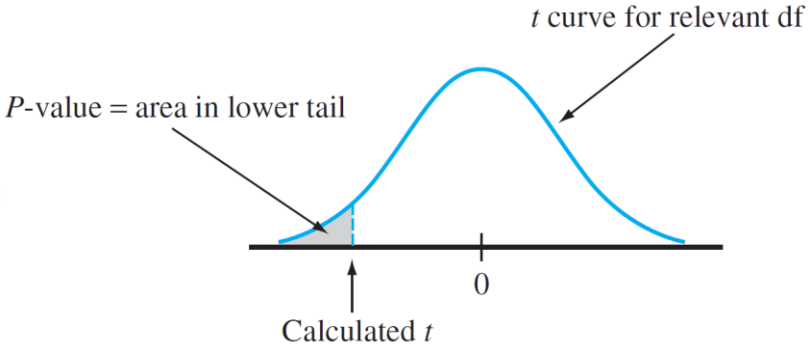
Just as the P-value for a z test is a z curve area, the P-value for a t-test will be a t-curve area.

The following figure illustrates the three different cases.
The number of df for the one-sample t-test is n – 1.

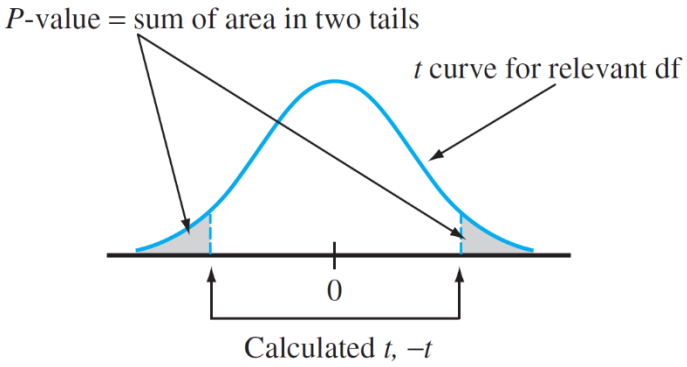
1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <

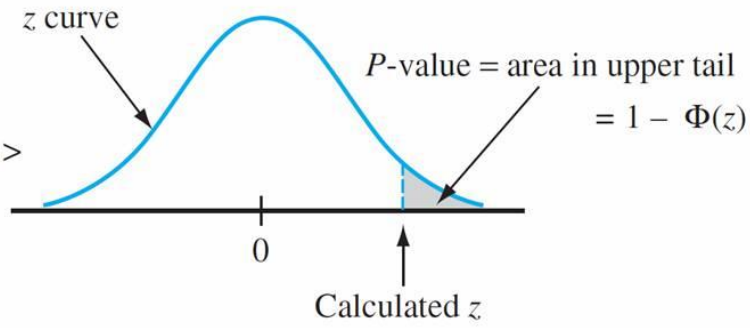


3. **Two-tailed test**
   H_a contains the inequality ≠

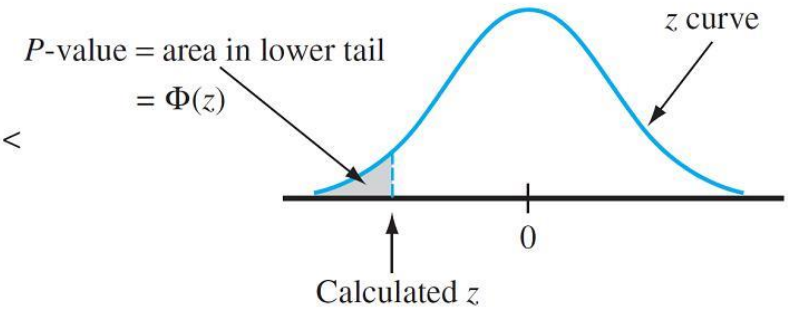


P-Values for z Tests

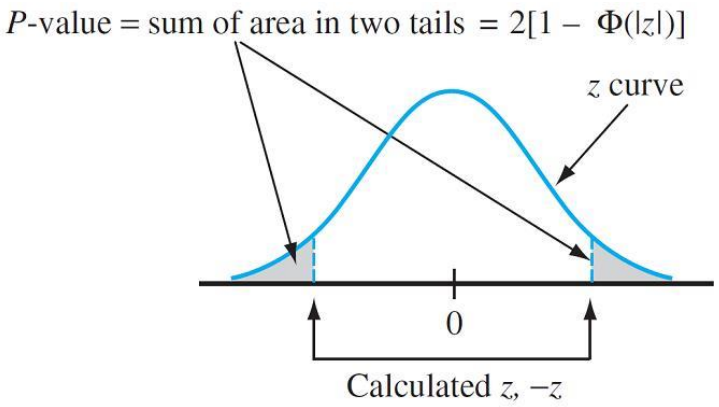
1. **Upper-tailed test**
   H_a contains the inequality >



2. **Lower-tailed test**
   H_a contains the inequality <



3. **Two-tailed test**
   H_a contains the inequality ≠



Example:

The silicon wafers' target thickness of the integrated circuit is 245 μm.

A sample of 50 wafers is chosen, and the thickness of each one is determined, resulting in a sample mean thickness of 246.18 μm and a sample standard deviation of 3.60 μm.

Does this data suggest that actual average wafer thickness is something other than the target value?

1. Parameter of interest: μ = true average wafer thickness
2. Null hypothesis: H_0: μ = 245
3. Alternative hypothesis: H_a: μ ≠ 245
4. Formula for test statistic value:

$$z = \frac{\bar{x} - 245}{s/\sqrt{n}}$$

5. Calculation of test statistic value:

$$z = \frac{246.18 - 245}{3.60/\sqrt{50}} = 2.32$$

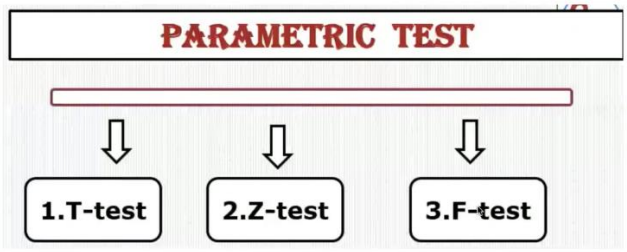
6. Determination of P-value: Because the test is two-tailed,
   P-value = 2(1 – F(2.32)) = .0204

7. Conclusion: Using a significance level of .01, H_0 would not be rejected since .0204 > .01.

   At this significance level, there is insufficient evidence to conclude that the actual average thickness differs from the target value.

T-Stats

A t-test is a statistical method used to see if two sets of data are significantly different. T-tests are used to involve data analysis that has applications in business, science, and many other disciplines.



Student T distribution

The T distribution is bell-shaped & symmetric, like the normal distribution, but has more massive tails. The T distribution is the family of distributions that looks identical to the normal distribution curve.

- Only a bit shorter and fatter.
- It is used in place of the normal distribution when we have small samples. (n<30)
- The T distribution similar is to the normal distribution if the sample size increases.

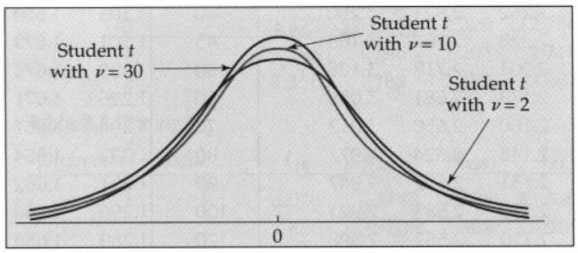
Here the letter t is used to represent the random variable, hence the name. The density function for the Student t distribution is as follows,

$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2}$$

ν (nu) is called the degrees of freedom, and

Γ (Gamma function) is Γ (k)=(k-1)(k-2)...(2)(1)

In much the same way that μ and σ define the normal distribution, ν, the degrees of freedom, defines the student t distribution:

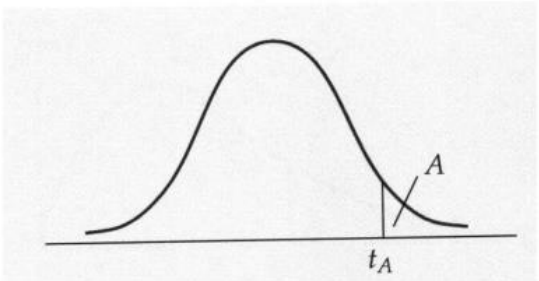


As the number of degrees of freedom increases, the t distribution approaches the standard normal distribution.

Determining Student values

- The Student 't' distribution is used extensively in statistical inferences.
- The value of a Student 't' random variable with V degrees of freedom is:

$$P(t > t_{A,\nu}) = A$$

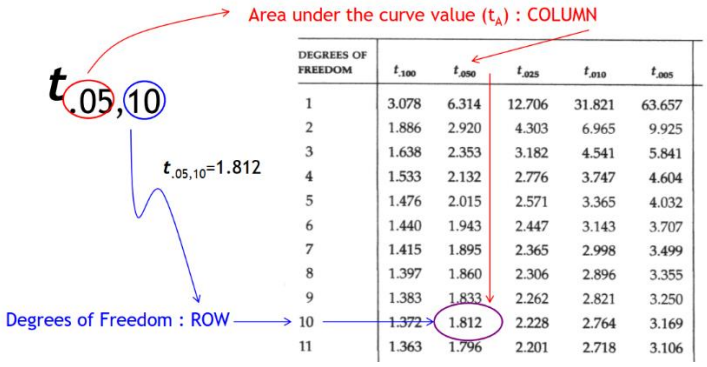


- The values for 'A' are pre-determined "critical" values, typically in the 10%, 5%, 2.5%, 1% and 1/2% range

Using the t table for all values

Example

- The value of t with degrees of freedom 10, so that the area under the Student 't' curve is 0.05:



DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106

Student T Distribution using Python

```
import scipy.stats
import numpy as np
import matplotlib.pyplot as plt
x= np.linspace(-10, 10, 100)
df = 3.34
mean, var = scipy.stats.t.stats(df, moments='mv')
print('mean: {:.2f}, variance: {:.2f}'.format(mean, var))
plt.plot(x, scipy.stats.t.pdf(x,df))
plt.show()
```

mean: 0.00, variance: 2.49

The t distribution has the following properties:

- The distribution's mean = 0 .
- v / (v - 2) is equal to variance, where v is the degrees of freedom.
- The variance is although close to 1, but it is always greater than 1 when degrees of freedom is greater.
- The t distribution is similar to the standard normal distribution, with infinite degrees of freedom.

Student t distribution and Probability:

When ‘n’ - the sample size, is drawn from a population having a normal distribution, using the equation of t-distribution, the sample mean is transformed into a t statistic. Following is the equation:

$$t = [X - \mu] / [S / \sqrt{n}]$$

where, μ is the population mean, X is the sample mean, and the sample size is n, and n – 1 is degrees of freedom here and s is the standard deviation of the sample.

Examples

Question: The Company by the name Acme Corporation manufactures light bulbs. The CEO of the company claims that an average Acme light bulb lasts for 300 days. So for testing, 15 bulbs are selected by a researcher randomly. An average sample bulb lasts for 290 days, with 50 days of standard deviation. What is the probability that 15 randomly selected bulbs do not have an average life of more than 290 days?

Solution

First we need to calculate the t statistic :

$$\text{t-distribution ‘t’} = [X - \mu] / [S / \sqrt{n}]$$
$$t = (290 - 300) / [50 / \sqrt{14}]$$
$$t = -10 / 12.909945 = - 0.7745966$$

where, μ is the population mean, X is the sample mean, the standard deviation of the sample is s and the sample size is ‘n.’

As we are aware of the t statistic, we can select the "T score" from the Random Variable dropdown box and enter the following data:

- 14 - 1 = 13. (The degrees of freedom)

- - 0.7745966. is t statistics

Question: IQ test's scores are normally distributed, with 100 as the population mean. Suppose 20 people are selected randomly and then tested.15 is the standard deviation in the sample group. What is the probability that in the sample group, an average test score will be at-most 110?

Answer:

- 20 - 1 = 19 is equal to degrees of freedom.
- 100 is the population mean.
- 110 is the sample mean
- 15 is the standard deviation

0.996 is the Cumulative Probability

So the chance that the sample average will be no higher than 110 is 99.6%

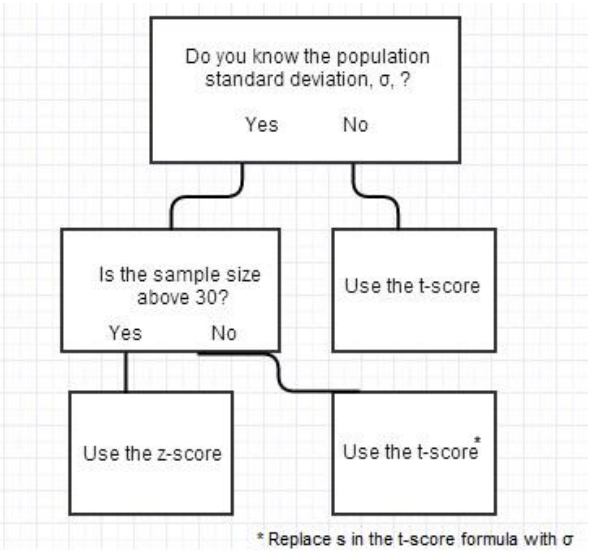
T-Stats vs. Z-Stats: Overview

- Z-tests
  - With the help of z-score, we can come to know how far the data point is from the mean in the standard deviation.
  - A z-test compares a sample to a defined population and is typically used for dealing with problems relating to larger samples (n > 30).
  - Z-score can also be used to test the hypothesis, and if the standard deviation is known, then it is beneficial.
- T-tests
  - Similar to the z-tests, t-tests are also used to test a hypothesis, but it is most useful when there is a need to know the significant difference between 2 independent sample groups.
  - In other words, the t-test asks whether the difference between the means of two groups is unlikely to have occurred because of the random chance. Usually, t-tests are the most appropriate when dealing with problems with a limited sample size (n < 30).

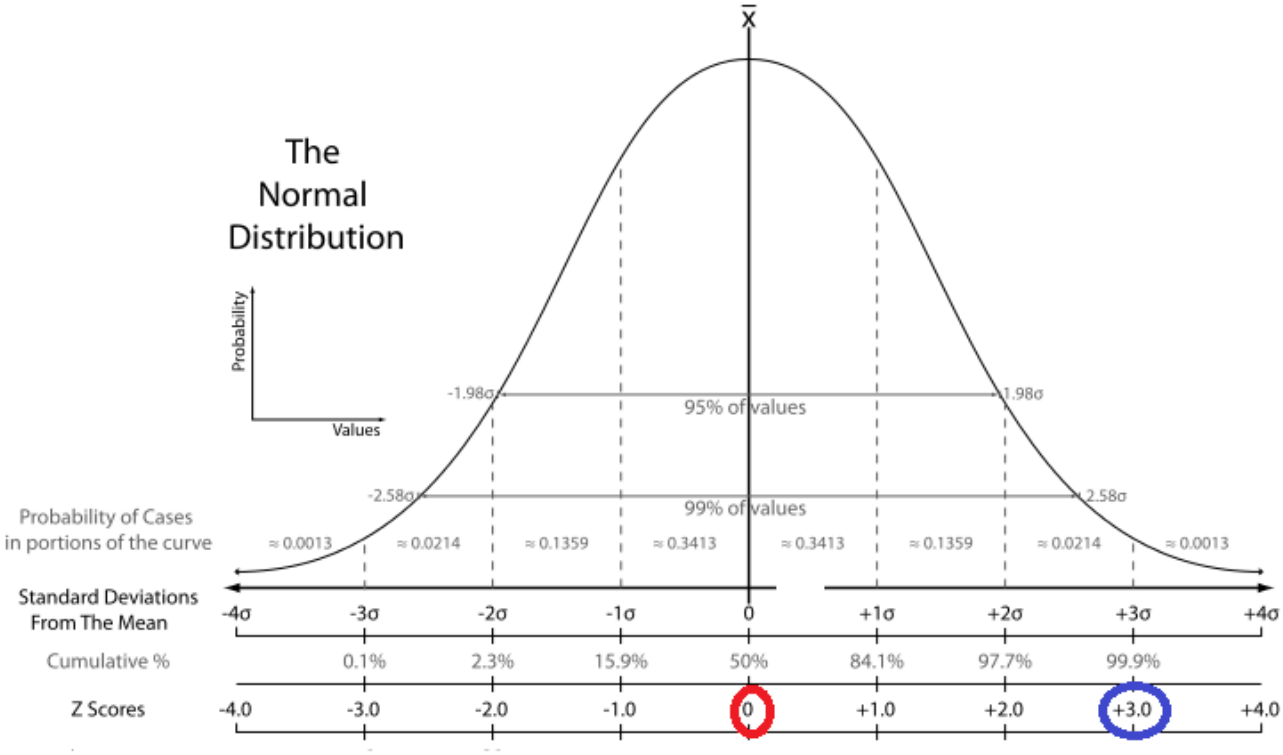
When to use a t-tests vs. z-tests

the t-test can be used when your sample:

* Has a sample size below 30,
* Has an unknown population standard deviation.



Note: A z-test tells you how many standard deviations away from the mean your result is. You can use the z-table to determine what percentage of the population will fall below or above your result.



The z-score is calculated using the formula:

$$z = (X-\mu)/\sigma$$

Where:

- σ is the population standard deviation and
- μ is the population mean.

Note: The t-score is calculated using the formula:

$$T = (X - \mu) / [S/\sqrt{(n)}]$$

Where the standard deviation of the sample is S.

Note: Like z-scores, t-scores are also a conversion of individual scores into a standard form. However, t-scores are used when you don't know the population standard deviation. You estimate by using your sample.

Type 1 & Type 2 Error
There are two possible errors:

1) A Type 1 Error:

* A Type 1 error occurs when we reject a true null hypothesis. Tha t is, a Type 1 error occurs when the jury convicts an innocent pers on.

- **In False Positive Error,** the null hypothesis is true. A type 1 error occurs but is rejected.

2) A Type 2 Error:

* When the null hypothesis is not rejected, then a Type 2 error oc curs.

- A type II error, or false negative, is where a test result indicates that a condition failed, while it was successful.

Example 1

Q. Regulations from the Environmental Protection Agency say that soil used in play areas should not have Lead levels that exceed 400 parts per million (ppm). Here an agent will run a test of significance on the mean Lead level in the soil. If the Lead level is higher than 400 ppm, then the soil is unsafe, and construction should not be continued.

Hypotheses for this test:

H_0: $\mu \leq 400$ ppm (that is soil is safe)

H_a: $\mu > 400$ ppm (that is soil is not safe)

(Here μ is the mean lead level in the soil at the new site).
Which of the following would be a Type I error in this setting?

Choose one answer:

A. The soil is safe, and the sample result is below 400\text{ ppm}400 ppm400, start a text, space, p, p, m, end text, so construction continues.

B. The soil is safe, and the sample result is significantly higher than 400\text{ ppm}400 ppm400, start a text, space, p, p, m, end text, so construction stops.

C. The soil is unsafe, and the sample result is below 400\text{ ppm}400 ppm400, start a text, space, p, p, m, end text, so construction continues.

D. The soil is unsafe, and the sample result is significantly higher than 400\text{ ppm}400 ppm400, start a text, space, p, p, m, end text, so construction stops.

Solution:

1 / 3

Type I and Type II error

	H_0 true	H_a true
Fail to reject H_0	correct conclusion	Type II error
Reject H_0	Type I error	correct conclusion

2 / 3

In this setting

Type I error: Rejecting a true null hypothesis

If H_0 is true, then the soil is actually safe. So a Type I error would occur if the sample result is significantly higher than 400 ppm, and construction stops when the soil is actually safe.

Type II error: Failing to reject a false null hypothesis

If H_0 is false, then H_a is true, and the soil is actually unsafe. So a Type II error would occur if the sample result isn't significantly higher than 400 ppm and construction continues when the soil is actually unsafe.

3 / 3

Answer

In this setting, a Type I error would be:

The soil is actually safe, and the sample result is significantly higher than 400 ppm, so construction stops.

Example 2

Q. According to a report from the United States Environmental Protection Agency, burning one gallon of gasoline typically emits about 8.9 kg of CO_2. A fuel company wants to test a new type of gasoline designed to have lower CO_2 emissions. Here are their hypotheses:

$H_o : \mu = 8.9$ kg

$H_a : \mu < 8.9$ kg

(where μ is the mean amount of CO_2 emitted by burning one gallon of this new gasoline).

Choose any of the following conditions through which the company commits a Type I error?

Choose one answer:

A. The mean amount of CO_2 emitted by the new fuel is 8.9 kg, and they fail to conclude it is lower than 8.9 kg

B. The mean amount of CO_2 emitted by the new fuel is lower than 8.9 kg, and they find it is smaller than 8.9 kg

C. The mean amount of CO_2emitted by the new fuel is 8.9 kg, and they conclude it is lower than 8.9 kg

D. The mean amount of CO_2emitted by the new fuel is lower than 8.9 kg, and they fail to find it is smaller than 8.9 kg

Solution:

1 / 3

Type I and Type II error

	H_0 true	H_a true
Fail to reject H_0	correct conclusion	Type II error
Reject H_0	Type I error	correct conclusion

2 / 3

In this setting

Type I error: Rejecting a true null hypothesis

If H_0 is true, then the mean amount of CO_2 emitted is actually $8.9\ \mathrm{kg}$. So a Type I error would occur if the sample result is significantly lower than $8.9\ \mathrm{kg}$, and they conclude that emissions are lower than $8.9\ \mathrm{kg}$

Type II error: Failing to reject a false null hypothesis

If H_0 is false, then H_a is true, and the mean amount of CO_2 emitted is actually lower than $8.9\ \mathrm{kg}$. So a Type II error would occur if the sample result is not significantly lower than $8.9\ \mathrm{kg}$, and they don't conclude that emissions are lower than $8.9\ \mathrm{kg}$

3 / 3

Answer

In this setting, a Type I error would be:

The mean amount of CO_2 emitted by the new fuel is actually $8.9\ \mathrm{kg}$, and they conclude it is lower than $8.9\ \mathrm{kg}$.

Bayes Statistics (Bayes Theorem)

Bayes Statistics is used for calculating conditional probabilities.

$$P(Ak \mid B) = \frac{P(Ak \cap B)}{[P(A1 \cap B) + P(A2 \cap B) + \ldots + P(An \cap B)]}$$

$$P(Ak \cap B) = P(Ak) P(B \mid Ak)$$

It can also be written as follows:

$$P(Ak \mid B) = \frac{P(Ak) P(B \mid Ak)}{[P(A1) P(B \mid A1) + P(A2) P(B \mid A2) + \ldots + P(An) P(B \mid An)]}$$

When to Apply Bayes' Theorem:

- When the sample space is divided(partitioned) into a set of events { A1, A2,…. An }.
- An event B is present, for which P(B) > 0 exists within the sample space.
- $P(A_k \mid B)$ is the form to compute a conditional probability.

One of the two sets of probabilities is mentioned:

- For each A_k, Probability, $P(A_k \cap B)$.
- For each A_k, Probability, $P(A_k)$ and $P(B \mid A_k)$

Example:

Problem

There is a marriage ceremony in the desert and Marie is getting married tomorrow. In past years, it has rained only five days a year. The weatherman has a weather report of raining tomorrow. The weatherman forecasts rain 90% of the time when it rains. When it failed to rain, the weatherman incorrectly predicts 10% of the times. What's the probability that it would rain on the marriage day of Marie?

Solution: The sample space is defined as - it rains or it does not rain. Furthermore, a 3rd event occurs when the weatherman predicts rain.

Event A. It rains on Marie's wedding.

Event B. It does not rain on Marie's wedding.

Event C. HERE THE RAIN IS PREDICTED.

- P (A) = 5/365 =0.0136985 [In a year only 5 days it might rain.]
- P (b) = 360/365 = 0.9863014 [It does not rain 360 days out of the year.]
- P (C | A) = 0.9 [When it rains, the weatherman predicts rain 90% of the time.]
- P (C | B) = 0.1 [When it FAILS TO rain, then the prediction is 10% of rain]

We want to know P (A | C), that is the probability it will rain on the wedding's day of Marie,

$$P(A \mid C) = \frac{P(A)\,P(B \mid A)}{P(A)\,P(C \mid A) + P(B)\,P(C \mid A2)}$$

$$P(A \mid C) = \frac{(0.0014)\,(0.09)}{[\,(0.014)(0.09) + (0.986)(0.1)\,]}$$

$$P(A \mid C) = 0.1111$$

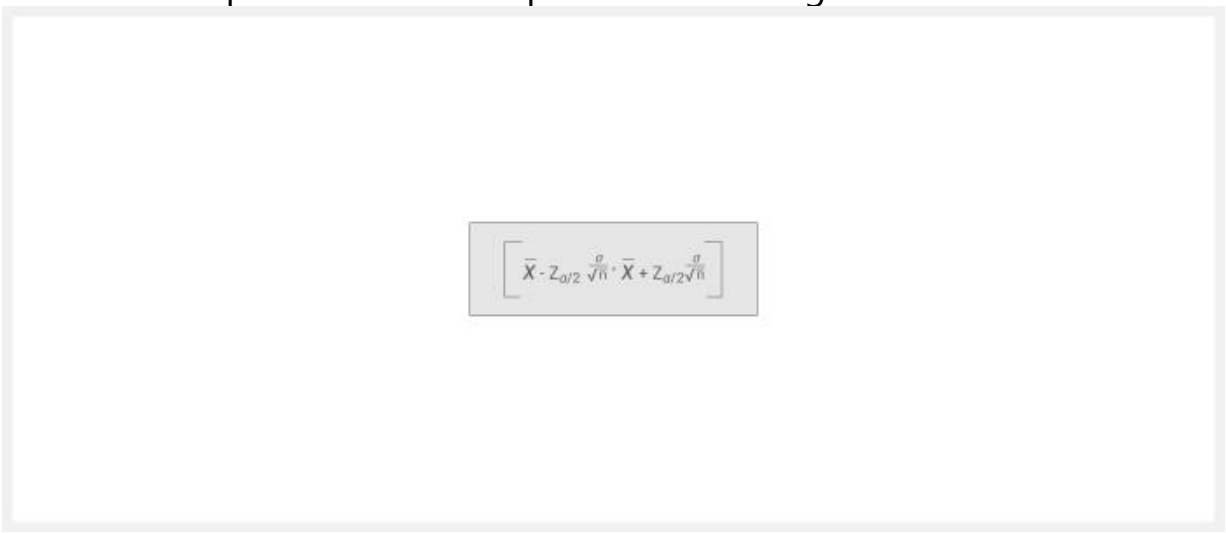
Even when the weatherman predicts rain, it might rain only about 11% of the time. So, there is a good chance that Marie might not get rained on at her wedding.

Confidence Interval(CI)

- It is the range within which we expect the population parameter to be.
- A confidence interval around the sample statistic is computed in such a way that it has a specific chance of containing the value of the corresponding population parameter.
- It also indicates that the range that's likely to contain the true population parameter, so the CI focuses on the population.
- They are often used with a margin of error.

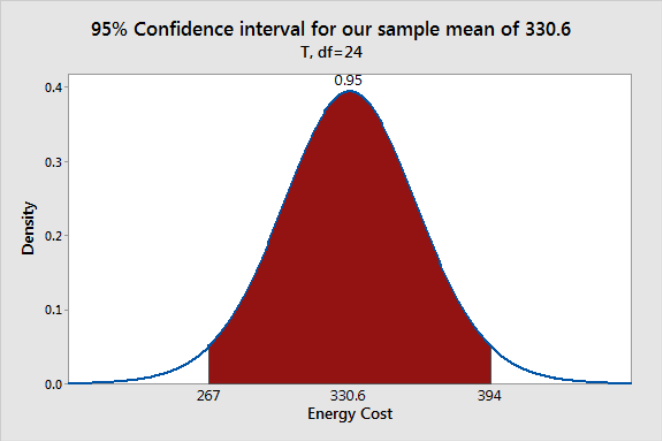
- It is denoted by 1 – α, here α is a value between 0 and 1 and is called the Confidence Level of Interval.

For example:
if we say that α is 10% and 90% is the parameter is inside the interval, and if we are 95% sure, then α is 5%.
It can be computed with the help of the formula given below:



The common CI are 90%, 95%, and 99%. With respective to α of 10%, 5%, and 1%.
Or we say α = 0.1%, 0.05% & 0.01%

Confidence Intervals and the Margin of Error



Here the area of the shaded region is the range of sample means that is 95% of the time using sample mean.
This range from 267 to 394 is the 95% confidence interval.

The CI is equal to 1, which is the alpha level. And if 0.05 is the level of significance, then 95% is the corresponding CI.

- If the significance (alpha) level is higher than the P-value, then we can say that the hypothesis test can be significantly correct.
- If the null hypothesis value is not found in the confidence interval, then the results are statistically significant.
- If alpha is higher than the P-value, then the hypothesis value will not be found in the confidence interval.

For our example, the P-value (0.031) is less than the significance level (0.05), which indicates that our results are statistically significant. Similarly, our 95% confidence interval [267 394] does not include the null hypothesis mean of 260, and we draw the same conclusion.

Interpreting confidence levels and confidence intervals

- When we create a confidence interval, it's essential to be able to understand the meaning of the confidence level we used and the interval obtained.
- The confidence level refers to the long-term success rate of the method, that is, how often this type of interval will capture the parameter of interest.
- A specific confidence interval gives a range of plausible values for the parameter of interest.

- Let's look at a few examples that demonstrate how to interpret confidence levels and confidence intervals.

Example 1: Interpreting a confidence level
A random sample of 500 voters was asked by a political pollster whether or not they support the incumbent candidate. The pollster will take the results of the sample and construct a 90% confidence interval for the exact proportion of all voters who support the candidate.
Which of the following is a correct interpretation of the 90% confidence level?

Choose all answers that apply:
A. If the pollster repeats this process and constructs 20 intervals from separate independent samples, we can expect about 18 of those intervals to contain the true proportion of voters who support the candidate.

B. About 90% of people who support the candidate will respond to the poll.

C. About 90% of the intervals produced will capture the true proportion of voters who support the candidate if the pollster repeats this process many times.

The answers:
Option A and Option C

Explanation: The confidence level tells us the long-term capture rate of these intervals over repeated samples. They do not tell us what percent of people will or won't respond to the poll.

Example 2: Interpreting a confidence interval
A baseball coach was curious about the true mean speed of fastball pitches in his league. The coach recorded the speed in kilometers per hour of each fastball in a random sample of 100 pitches and constructed

a 95% confidence interval for the mean speed. The resulting interval was (110, 120).
Which of the following is a correct interpretation of the interval (110, 120)?

Choose all answers that apply:
A. If the coach took another sample of 100 pitches, there's a 95% chance the sample mean would be between 110 and 120 km/hr.

B: About 95% of pitches in the sample were between 110 and 120 km/hr.

C: We're 95% confident that the interval (110, 120) captured the true mean pitch speed.

The answer:
Option C.

Explanation:

- A confidence interval doesn't estimate the sample result from an upcoming sample. So, we can't use this interval to make predictions about the sample mean from a new sample of 100 pitches.
- A confidence interval doesn't describe the distribution of the sample data used to build the interval. So, we can't say that 95% of pitches in the sample were between 110 and 120 km/hr.
- The true value of the estimated parameter is captured by a confidence interval, which in this case, is the true mean pitch speed in the league. The confidence level tells us the long-term capture rate of these intervals over repeated samples.

Calculating the Confidence Interval

Question:

Step1: Given the number of observations **n**, their mean **X**, and standard deviation **s.**

Number of observations: **n = 40,** Mean: **x = 175,** Standard Deviation: **s = 20**

Step 2: decide what Confidence Interval we want: 95% or 99% are common choices.

Then to find is the "Z" value for that Confidence Interval:

Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

1.960 is the Z value for 95%

S**tep 3**: use that Z in this formula for the Confidence Interval

$$\bar{x} \pm Zs \textbf{ sqrt(n)}$$

Here, **Z** is the Z-value taken from the table above, x is the mean, and **s** is the standard deviation, n is the number of observations

$$175 \pm 1.960 \times 20\sqrt{\textbf{40}}$$

Which is:

175cm ± 6.20cm

In other words: from 168.8cm to 181.2cm

Chi-Square test

In probability and statistics, the chi-squared distribution (also chi-square or X^2-distribution) with the degrees of freedom 'k' is the distribution of a sum of the squares of k independent standard normal random variables.

- The x^2(chi-square) distribution is a continuous probability distribution that is widely used in statistical inference.
- The x^2(chi-square) distribution is related to the standard normal distribution:
    - If a random variable Z has the standard normal distribution, then Z^2 has the x^2(chi-square) distribution with one degree of freedom.

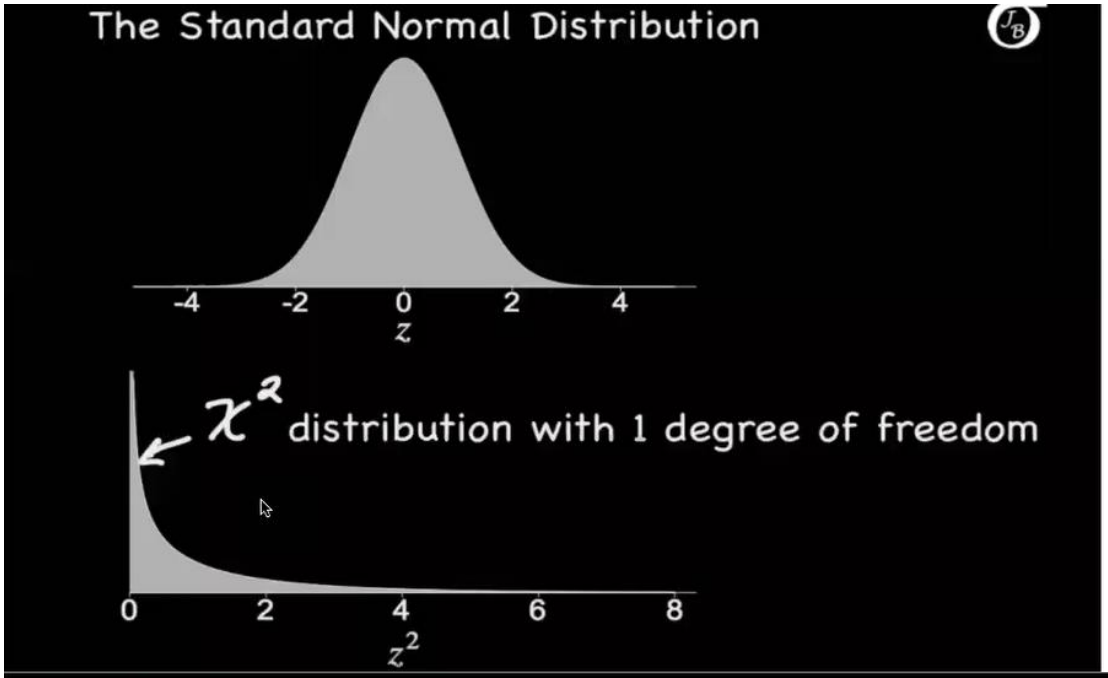
Chi-Square Distribution using Python

```
%matplotlib inline
import scipy.stats
import numpy as np
import matplotlib.pyplot as plt

x=np.linspace(5,10,100)
df=99
mean,var,skew, kurt=scipy.stats.chi2.stats(df, moments='mvsk')
print('mean: {:.2f}. var: {:.2f}, Skewness: {:.2f}. Kurtosis: {:.2f}'.format(mea
n, var, skew, kurt))

plt.plot(x, scipy.stats.chi2.pdf(x,df))
plt.show()
```

```
mean: 99.00. var: 198.00, Skewness: 0.28. Kurtosis: 0.12
```



If $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables, then

$$Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

has a χ^2 distribution with k degrees of freedom.

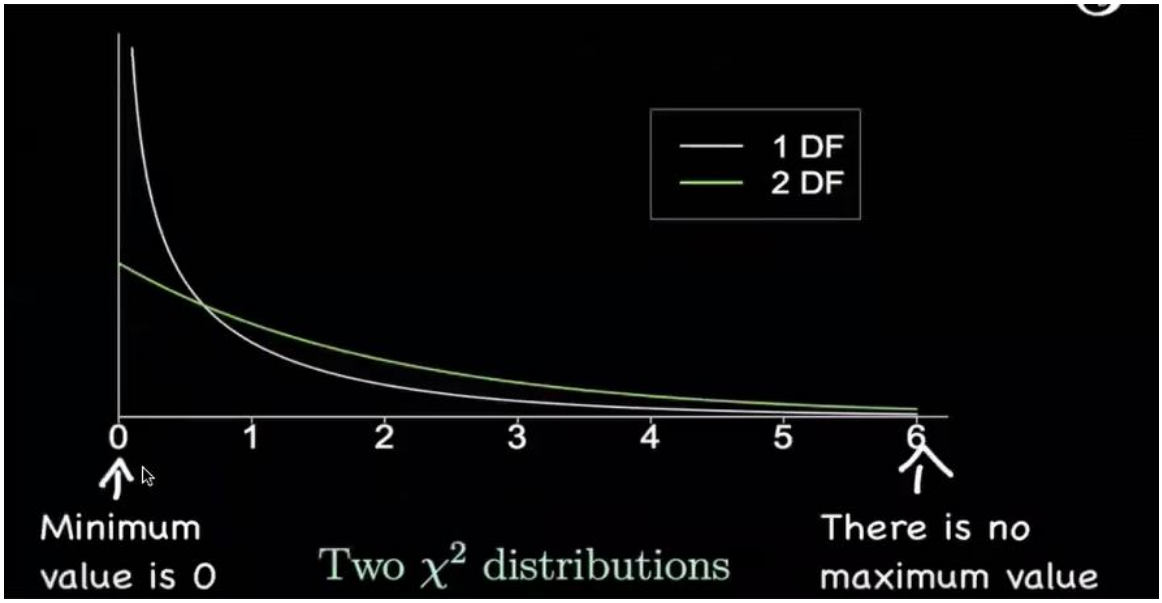
Usually a positive whole number

The pdf of the χ^2 distribution with k degrees of freedom:

$$f(x) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} \quad \text{for } x \geq 0$$

$\mu = k$

$\sigma^2 = 2k$



Chi-Square for Goodness of Fit Test

Chi-Square test for testing goodness of fit is used to decide whether there is any difference between the experimental value & the expected (theoretical) value.

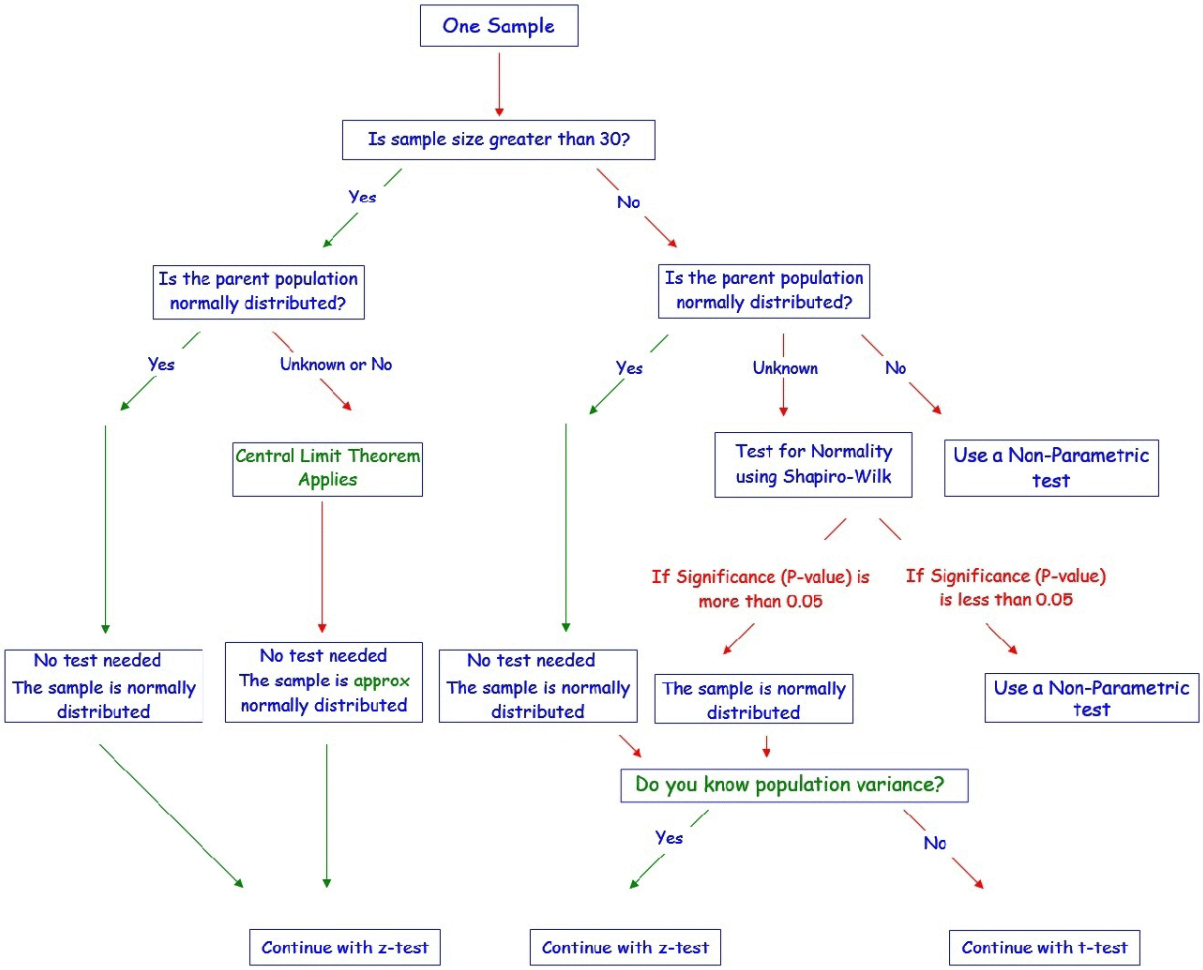
A chi-squared test, also written as x2 test, is any statistical hypothesis test where the sampling distribution of the test statistics is chi-squared distribution when the null hypothesis becomes true.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

χ^2 = the test statistic $\sum$ = the sum of

O = Observed frequencies E = Expected frequencies

When to use which statistical distribution?



Analysis Of Variance (ANOVA)

- It is a parametric statistical technique that is used to compare datasets.
- R.A. Fisher is the person who invented this technique. It is also referred to as Fisher's NOVA.

Assumptions to use ANOVA

1. The independence of a case assumption means that the sample should be selected randomly. There should be no pattern in the selection of the sample.
2. Normality: Distribution of each group should be normal. Kolmogorov-Smirnov or the Shapiro-Wilk test may be used to confirm the normality of the group.

3. Homogeneity means between the group; the variance must be the same. To test the correlation between groups, Levene's test is used.

ANOVA has 3 types:

- If we start to compare more than three groups based on a single factor, then it is known as a one-way analysis of variance.
  For example, if the mean output of 3 workers is compared whether they are the same based on the working hours of the three workers.

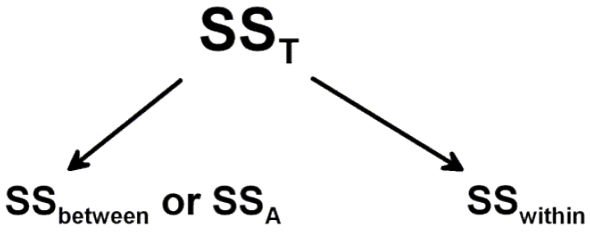
- Two-way analysis: When factor variables are more than two, then it is said to be the two-way analysis of variance (ANOVA). For example, based on working conditions and working hours, we can compare whether or not the mean output of the three workers is the same.

- K-way analysis: When factor variables are k, then it is said to be the k-way analysis of variance (ANOVA).

Degrees of Freedom

```
### Calculate the degrees of freedom
DFbetween = k - 1
DFwithin = N - k
DFtotal = N - 1
```

Partitioning of Variance in the ANOVA



Note: SS stands for Sum of Squares.

Calculating using Python

1) The sum of squares between (SSbetween)
2) The sum of squares within (SSwithin)
3) The sum of squares total (SSTotal)

1) Sum of Squares Between is the variability because of the interaction between the groups.

$$SSbetween = \frac{\sum(\sum k_i)^2}{n} - \frac{T^2}{N}$$

2) The variability in the data because of differences within people. The calculation of Sum of Squares Within can be carried out according to this formula:

$$SSwithin = \sum Y^2 - \frac{\sum(\sum a_i)^2}{n}$$

3) This is the total data variability.

$$SStotal = \sum Y^2 - \frac{T^2}{N}$$

Sum of Squares Between, Within, and Total

```
# Between
SSbetween = (sum(df.groupby('group').sum()['weight']**2)/n) \
    - (df['weight'].sum()**2)/N

# Within
#sum_y_squared = sum([value**2 for value in df['weight'].values])
SSwithin = sum_y_squared - sum(df.groupby('group').sum()['weight']**2)/
n

# Total
SStotal = sum_y_squared - (df['weight'].sum()**2)/N
```

Means Square Errors and F-value

```
MSbetween = SSbetween/DFbetween
MSwithin = SSwithin/DFwithin

F = MSbetween/MSwithin
### Obtaining the p-value
```

```
p = stats.f.sf(F, DFbetween, DFwithin)
```

Example: ANOVA in Python using Statsmodels

```
#import scipy, numpy, and matplotlib

data = pd.read_csv('PlantGrowth.csv')
data.boxplot('weight', by='group')

Output:
<matplotlib.axes._subplots.AxesSubplot at 0x2a11624a2e8>

#First, we import statsmodels API and ols:

# ols is used to set up a model using a formula
mod = ols('weight ~ group', data=data).fit()

# anova_lm to carry out the ANOVA
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)

Output:
            sum_sq    df         F    PR(>F)
Group     3.7563534   3.0  4.848588  0.151
Residual  11.49210  23.0       NaN       NaN
```

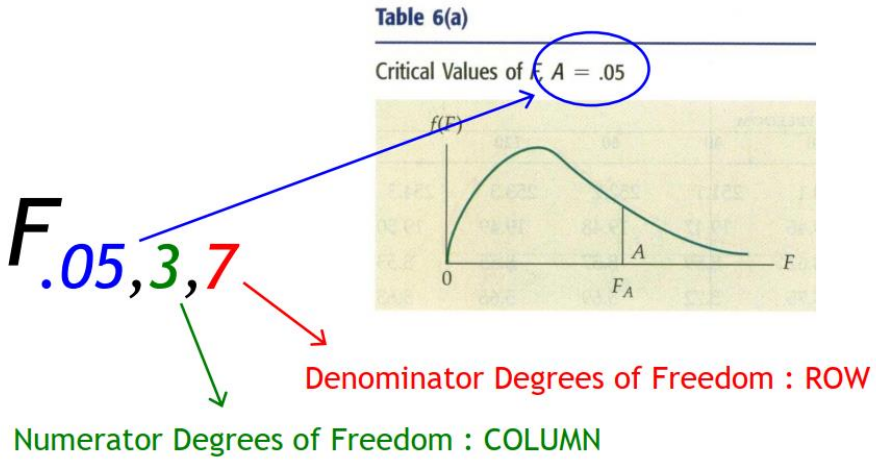
F Distribution

F-Test (variance ratio test)
When we run a regression analysis, we get f value to find out the means between two populations. It's similar to a T statistic from a T-Test. A T-test will tell you if a single variable is related statistically, and an F test will tell you if a group of variables is jointly significant.

- F-test is used to test the two independent estimations of population variances(S_1^2 & S_2^2).

- F-test is used by comparing the ratio of the two variances S_1^2 & S_2^2.
- The samples must be independent.
- F-test is a small sample test.
    - F = (Larger estimate of population variance) / (Smaller estimate Of population variance)
- The variance ratio = S_1^2 & S_2^2
- F-test never is -ve because the upper value is greater than lower.
- Degree of freedom for larger population variance is V1 and smaller V2
- The null hypothesis of two population variance are equal, i.e.,
    - H_O: S_1^2 = S_2^2

Determining the Values of F



$$F_{1-A,\nu_1,\nu_2} = \frac{1}{F_{A,\nu_2,\nu_1}}$$

F Distribution using Python

```
#impolrt scipy, numpy and matplotlib
```

```
x=np.linspace(-10, 10, 100)
dfn = 29
dfd = 18
mean, var, skew, kurt = scipy.stats.f.stats(dfn, dfd, moments='mvsk')
print('mean: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, ske
w, kurt))
plt.plot(x, scipy.stats.f.pdf(x,dfn, dfd))
plt.show()


mean: 1.12, skewness: 0.28, kurtosis: 1.81
```

Note:

- The Student 't' distribution is robust, which means that if the population is non-normal, the results of the t-test and confidence interval estimate are still valid provided that the population is not extremely non-normal.
- To check this requirement, draw a histogram of the data and see how bell-shaped the resulting figure is. If a histogram is extremely skewed (say in that case of an exponential distribution), that could be considered "extremely non-normal," and hence, t-statistics would not be valid in this case.

Example

Question: From a population of women**,** suppose you randomly select 7 women, and from the population of men, 12 men are selected.

Population	Population standard deviation	Sample standard deviation
Women	30	35
Men	50	45

To calculate f statistics.

Answer: The f statistic can be calculated from the sample standard deviations and population, using the following equation: $f = [s_1^2/\sigma_1^2] / [s_2^2/\sigma_2^2]$

where Standard deviation of the sample drawn from population 1 is s_1 and s_2 in the denominator is the standard deviation of the sample drawn from population 2, σ_1 is the standard deviation of population 1, Population 2's standard deviation is σ_2.

As we can see from the equation, there are two ways to compute an f statistic from these data. If the data of women appears in the numerator, we can compute f statistic as follows:

$$f = (55^2 / 20^2) / (45^2 / 50^2)$$

$$f = (3025 / 400) / (2025 / 2500).$$

$$f = 1.361 / 0.81 = 1.68$$

For calculations, the numerator degrees of freedom v_1 are 7 - 1 or 6; and the degrees of freedom for denominator v_2 are 12 - 1 or 11.

On the other hand, if the men's data appears in the numerator, we can calculate the f statistic as follows:

$$f = (45^2 / 50^2) / (55^2 / 20^2)$$

$$f = (2025 / 2500) / (3025 / 400)$$

$$f = 0.812 / 1.3610 = 0.5955$$

For this calculation, the denominator degrees of freedom v_2 is 7 - 1 or 6 and the numerator degrees of freedom v_1 is 12 - 1 or 11

When we are trying to find the cumulative probability associated with an f statistic, you need to know v_1 and v_2.

Question: Find the cumulative probability related to each of the f statistics from the above example:

Answer: First, we need to find the degrees of freedom for each sample. Then, probabilities can be found.

- The sample of women's degrees of freedom is equal to n - 1 = 7 - 1 = 6.
- The sample of men's degrees of freedom is equal to n - 1 = 12 - 1 = 11.

Therefore, when data of women appear in the numerator, then v_1 is equal to 6; and then v_2 is equal to 11. And, the f statistic is equal to 1.68. So, 0.78 is the cumulative probability.

When data of men appear in the numerator, then v_1 is equal to 11; and then v_2 is equal to 6. And, the f statistic is equal to 0.595. Thus the cumulative probability is **0.22.**