Research Track II
Fundaments of Statistics

Carmine Tommaso Recchiuto

Testing of Hypotheses

- Hypothesis is usually considered as the principal instrument in research. Its main function is to suggest new experiments and observations.
- Simply put, **a hypothesis is a statement which makes a prediction about something which is not proven.** It is a kind of educated guess.
- In fact, experiments (in robotics and other fields) are carried out with the deliberate object of testing hypotheses.
- Hypothesis testing is the often used strategy for deciding whether a sample data offer such support for a hypothesis that generalisation can be made. Thus hypothesis testing enables us to make probability statements about population parameter(s).
- The hypothesis may not be proved absolutely, but in practice it is accepted if it has withstood a critical testing.

Testing of Hypotheses

- In biology, for example, the purpose of the whole research field is to make predictions about the state of the world and then do experiments to prove or disprove them. Therefore, a strong hypothesis in biology is an easy thing to spot. It might look something like:

"When the salt levels in the water fall, sea trout have a higher rate of mortality."

- This is a simple, clear and testable hypothesis which is provable (or disprovable) by experiments.

- Robotics is a little different from the natural sciences, because it is primarily an engineering field. The engineering process is different from the scientific method, because it doesn't usually *make hypotheses* and *conduct experiments* to prove them. Instead, it *defines design criteria* and then *develops technologies* which achieve those criteria.

Testing of Hypotheses

Most of the research work in robotics is spent on developing new technologies and methodologies. However, even though the primary purpose of your research is not to prove a hypothesis, as it is in natural science research, you still need a hypothesis to conduct experiments which prove that your developments work as well as you claim they do.

Therefore, a typical robotics hypothesis might look something like:

"Our algorithm calculates the solution faster than a naive algorithm can calculate the same problem.“

Testing of Hypotheses

Also in robotics, a good hypothesis will satisfy some characteristics:

- **It is provable or disprovable.** A statement that is unprovable is not a hypothesis, nor is a statement which is always true or false regardless of the results.
- **It is simple and clear**
- **It is testable.** You can't say "our algorithm is more memory efficient than all other algorithms" because you could never test all other algorithms.
- **It's relevant to a problem.** Don't, for example, claim to have developed a low power robot and then design an experiment which tests the hypothesis "Our robot can complete this specific task in under 10 seconds" because this omits the key aspect of "low power robot" in your hypothesis and, thus, your experiment.
- **It is specific.** The following is not a hypothesis: "Our robot is better than our previous robot." What specific aspect of the robot is "better"? By how much is it an improvement over the previous model? How might you be able to demonstrate this?

.

Testing of Hypotheses

This might commonly be phrased in robotics as something like:

"If I apply this technique to this particular robot then the robot will be able to achieve this specific task with a particular level of performance."

Once you've written a draft hypothesis, it's important to take a step back and ask yourself if the hypothesis has a "purpose".

It is not valid to simply hypothesize that a device to perform a certain function can be designed and fabricated. This "hypothesis" naturally begs the question "Why?" and as such is not a legitimate focus for research.“

Do not stick with your first hypothesis and carefully think about it before designing the experiment.
Is there anything you could add which would make the experiment more impressive or more comprehensive?
Any improvements you make to the hypothesis now will save you a lot of time later.

Testing of Hypotheses

Example: Multi-Robot Grasp Planning for Sequential Assembly Operations (Dogar et al 2015)

In this paper, the authors use multiple mobile manipulators with or without re-grasps (i.e. putting down the piece and picking it up in a different way). The hypothesis can be more or less expressed like this:

Our algorithm calculates a good enough manipulation solution with few re-grasps faster than a naive algorithm can calculate the same problem optimally (i.e. with no re-grasps).

For the sake of demonstration, here is an example of what a less specific hypothesis might look like, one which does not define variables. It is completely unspecific as to what "quickly" means:

Our algorithm can calculate a solution quickly.

Another problem might arise if the hypothesis does not include testability. For example, this hypothesis would not be testable, because you could never test all of the other algorithms:

Our algorithm calculates a solution better than all of the other planning algorithms.

Example

Scenario:

- A, B are two path planning techniques
- Score is the planning time
- Data d is a given map, start and goal pose

A(d) = 0.5 s

B(d) = 0.6 s

What does that mean?

2nd Example

Same scenario but four tasks

A(d) = 0.5 s, 0.4 s, 0.6 s, 0.4 s

B(d) = 0.6 s, 0.3s, 0.4 s, 0.5 s

What does that mean?

Mean of the planning time is:

μ_A = 1.9 s/4 = 0.475 s

μ_B = 1.8 s/4 = 0.45 s

Is this enough to say that B is faster than A?

2nd Example

Mean of the planning time is:

$\mu_A = 1.9 \text{ s}/4 = 0.475 \text{ s}$

$\mu_B = 1.8 \text{ s}/4 = 0.45 \text{ s}$

Is this enough to say that B is faster than A?

- Intuitively, you may say that we have just evaluated four tests, thus the two mean values are just some rough estimates. So, in this case, we probably saw too few data to make statements with high confidence.
- But what if the difference between the means was much bigger?

- In other words, how can we make a “generalized”, confident statement that B is better than A?

Some Basic Concepts

Some basic concepts in the context of testing of hypotheses need to be here explained before proceeding.

- ***Null hypothesis and alternative hypothesis***

In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method A with method B about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis.

As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a.

If our sample results do not support this null hypothesis, we should conclude that the alternative hypothesis is true. If we accept H_0, then we are rejecting H_a and if we reject H_0, then we are accepting H_a

Some Basic Concepts

The null hypothesis and the alternative hypothesis are chosen before the sample is drawn (you should avoid the error of deriving hypotheses from the data). Some comments:

- **The alternative hypothesis** is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject, and alternative hypothesis represents all other possibilities.

- ***The level of significance is an important concept***. It is always a percentage (usually 5%) which should be chosen with great care. In case we take the significance level at 5 per cent, then this implies that *H_0 will be rejected when the sampling result (i.e., observed evidence) has less than 5% probability of occurring if H_0 is true. In other words, the 5 per cent level of significance means that researcher is taking a 5 per cent risk of rejecting the null hypothesis when H_0 happens to be true.* Thus the significance level is the maximum value of the probability of rejecting H_0 when it is true and is usually determined in advance before testing the hypothesis.

Some Basic Concepts

But how we can test hypotheses to reject (or accept) the null hypothesis (or the alternative one)?

You might run an experiment and find a certain result. But if you can't repeat that experiment, no one will take your results seriously

For example, in social robotics we may try to assess if a certain algorithm for implementing autonomous conversation is more appreciated by users than another state-of-the-art algorithm; or if some functionalities of a robot may increase or decrease its perceived naturalness or usability. Obviously, when you have tests that also involve human subjects, you can't just perform a single experiment: the results may strictly depend on the considered subject.

However a similar situation may occur also in other contexts, such as mobile robotics. Suppose that you want to evaluate the capability of a certain algorithm to plan a path, or to create a map of the environment, in an environment filled with obstacles. Different algorithms can have different performance with different obstacle configurations.

Hence, how we can properly test our hypothesis? Let's first see some related concepts.

Some Basic Concepts

It’s not sufficient to just repeat the tests and evaluate the overall outcomes: differences may be due to chance.

How many experiments do I need to perform? How may I assess if the differences between the implementations are due to chance, or real differences?

We need to perform a statistical analysis of the result!

Statistical methods are valuable when we need deal with problems that involve complex interactions (e.g. human subjects) or seemingly random phenomena (obstacle position in the environment). Hence, it may be a precious instrument to evaluate the performance of our robotic system in many situations

So, what does it happen when we collect data?

Other Basic Concepts

Mean and Weighted Average. The mean (also know as average), is obtained by dividing the sum of observed values by the number of observations, *n*.

$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

However, this equation can only be used when the error associated with each measurement is the same. Otherwise, the weighted average, which incorporates the standard deviation, should be calculated:

$$X_{wav} = \frac{\sum w_i x_i}{\sum w_i}$$

Other Basic Concepts

Median. The median is the middle value of a set of data containing an odd number of values, or the average of the two middle values of a set of data with an even number of values. The median is especially helpful when separating data into two equal sized bins.

Mode. The mode of a set of data is the value which occurs most frequently.

Standard Deviation. The standard deviation gives an idea of how close the entire set of data is to the average value. Data sets with a small standard deviation have tightly grouped, precise data. Data sets with large standard deviations have data spread out over a wide range of values

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

2nd Example

A(d) = 0.5 s, 0.4 s, 0.6 s, 0.4 s

B(d) = 0.6 s, 0.3s, 0.4 s, 0.5 s

$\mu_A = 1.9\ s/4 = 0.475\ s$

$\mu_B = 1.8\ s/4 = 0.45\ s$

$$\sigma^2_A = \frac{(0.5 - 0.475)^2 + \ldots + (0.4 - 0.475)^2}{4} \qquad \sigma_A = 0.083$$

$$\sigma^2_B = \frac{(0.6 - 0.45)^2 + \ldots + (0.5 - 0.45)^2}{4} \qquad \sigma_B = 0.112$$

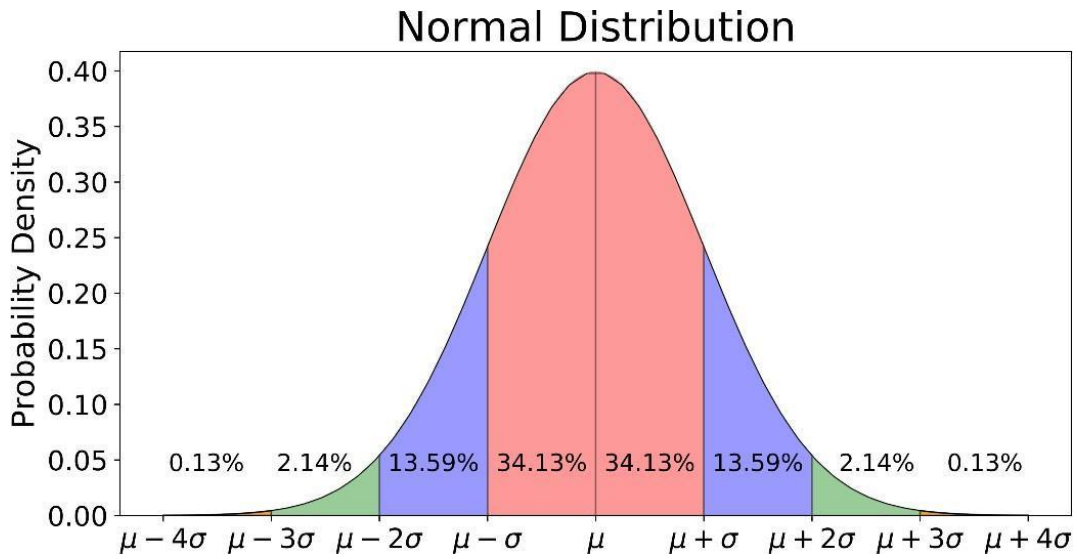
We are still far from a “generalized”, confident statement that B is better than A, but it is evident how the difference between the means, and the standard deviation gives some additional information about the reliability of the measurements.

Normal Distribution

A normal distribution, also known as the Gaussian distribution, or z-distribution or the bell curve is a distribution that occurs naturally in many situations. Usually, the majority of the results are accumulated around the average, while smaller experiments will give results which can be represented in the tails of the curve.

Also, the curve is symmetrical. Half of the data will fall to the left of the mean; half will fall to the right.

Many types of data follow this pattern.



Normal Distribution

Analitically, this outcome can be represented with a probability density function of type:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A PDF is used to specify the probability of the random variable falling within a particular range of values. This probability is given by the integral of this variable's PDF over that range— that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range.

With reference to the figure of the previous slide, 99.7% of the data is within 3 standard deviations (σ) of the mean (μ).

Normal Distribution

However, our data not always represent a normal distribution.

Skewness is a measure of asymmetry and shows the manner in which the items are clustered around the average. In a symmetrical distribution, the items show a perfect balance on either side of the mode, but in a skew distribution the balance is thrown to one side. The amount by which the balance exceeds on one side measures the skewness of the series.

Kurtosis is the measure of flat-toppedness of a curve. The curve may be relatively more peaked or more flat than the normal curve.

Knowing the shape of the distribution curve is crucial to the use of statistical methods in research analysis since most methods make specific assumptions about the nature of the distribution curve.

But before analysing some statistical tests, we still need to acquire some additional concepts about our samples.

Some Other Concepts

Universe/Population: From a statistical point of view, the term 'Universe' refers to the total of the items or units in any field of inquiry, whereas the term 'population' refers to the total of items about which information is desired.

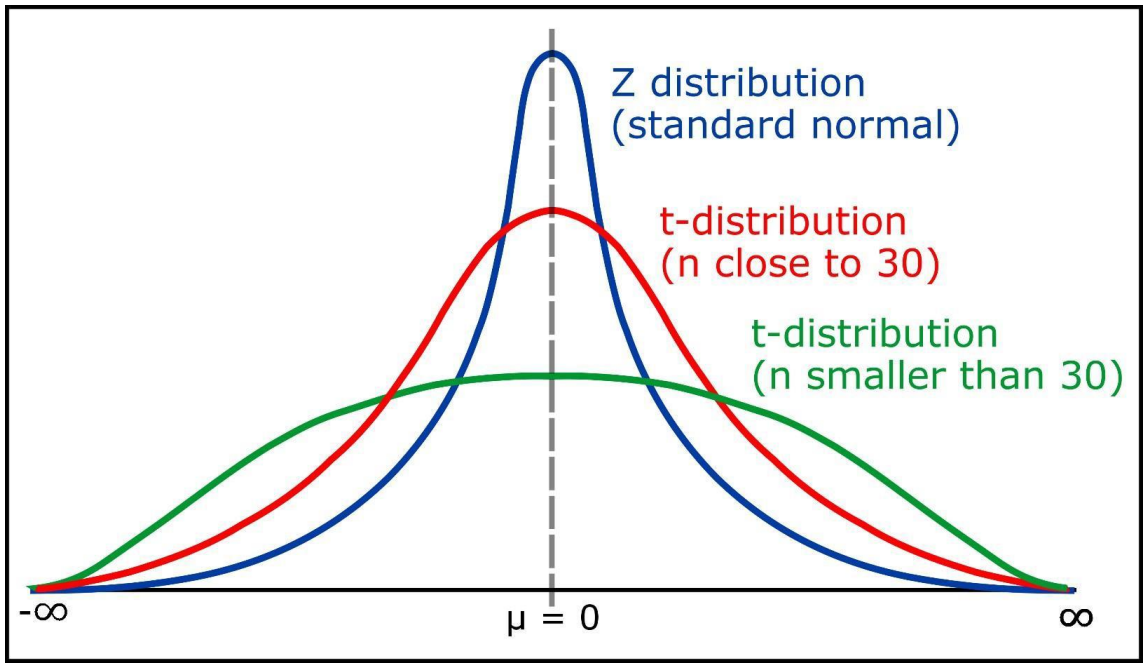
Sampling frame: Thus sampling frame consists of a list of items (belonging to the population) from which the sample is to be drawn. If the population is finite, then it is possible for the frame to be identical with the population. In most cases they are not identical because it is often impossible to draw a sample directly from population

Sampling design: A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. Sampling design is determined before any data are collected.

Statistic(s) and parameter(s): A statistic is a characteristic of a sample, whereas a parameter is a characteristic of a population. Thus, when we work out certain measures from samples, then they are called statistic(s), and they describe the characteristics of a sample. But when such measures describe the characteristics of a population, they are known as parameter(s). For instance, the population mean (m) is a parameter, whereas the sample mean ($\bar{X}$) is a statistic. *To obtain the estimate of a parameter from a statistic constitutes the prime objective of sampling analysis.*

T-Distribution

The *t*-distribution describes the standardized distances of sample means to the population mean when the population standard deviation is not known, and the observations come from a normally distributed population.

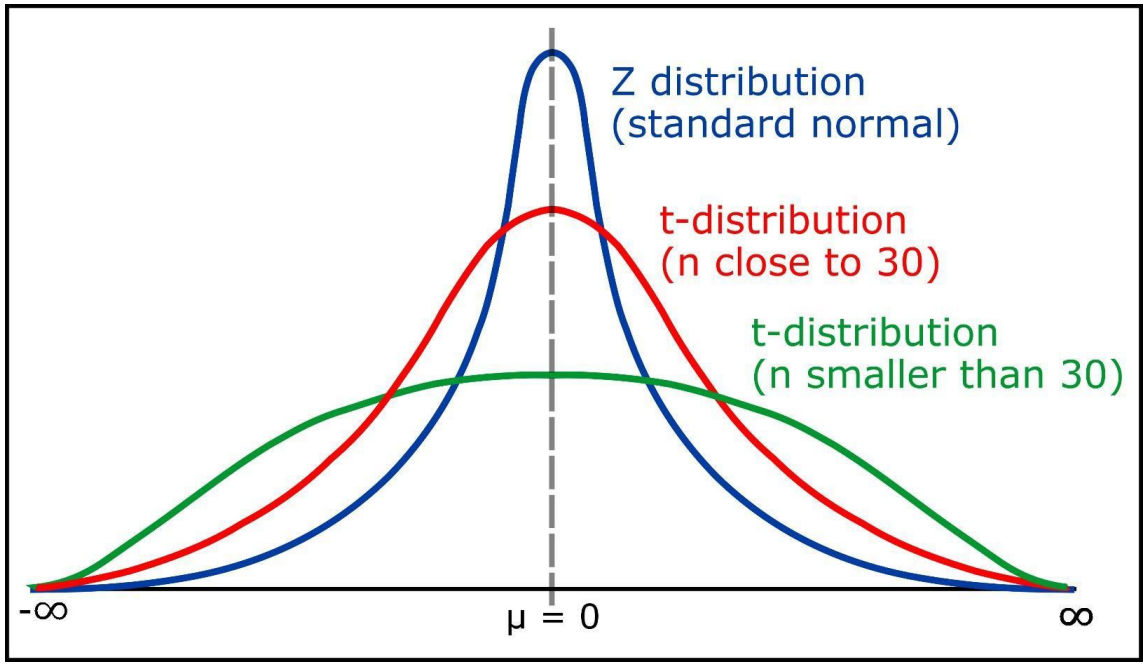


- Like the normal distribution, the *t*-distribution has a smooth shape.
- Like the normal distribution, the *t*-distribution is symmetric.
- Like a standard normal distribution (or z-distribution), the *t*-distribution has a mean of zero.
- The *t*-distribution is most useful for small sample sizes, when the population standard deviation is not known, or both.
- As the sample size increases, the *t*-distribution becomes more similar to a normal distribution.

T-Distribution

A common rule of thumb is that for a sample size of at least 30, one can use the z-distribution in place of a *t*-distribution. The Figure shows a *t*-distribution with 30 degrees of freedom and a z-distribution.

The similarity between the two curves is the reason why a z-distribution is used in statistical methods in place of a *t*-distribution when sample sizes are sufficiently large.



The Sampling Distribution and Standard Deviation of the Mean

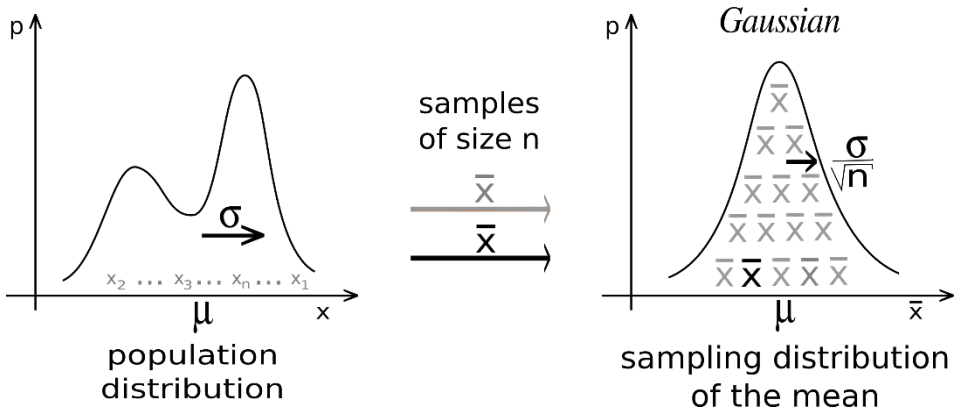
Population parameters follow all types of distributions, some are normal, others are skewed, other very irregular. However, many statistical methodologies, like parametric tests, are based off of the normal distribution. How does this work? Most sample data are not normally distributed.

Imagine an engineering is estimating the mean weight of widgets produced in a large batch. The engineer measures the weight of N widgets and calculates the mean. So far, one sample has been taken. The engineer then takes another sample, and another and another, and continues until a very larger number of samples and thus a larger number of mean sample weights (assume the batch of widgets being sampled from is near infinite for simplicity) have been gathered. The engineer has generated a sample distribution.

As the name suggested, a sample distribution is simply a distribution of a particular statistic (calculated for a sample with a set size) for a particular population. In this example, the statistic is mean widget weight and the sample size is N. If the engineer were to plot a histogram of the mean widget weights, he/she would see a bell-shaped distribution.

Some Other Concepts

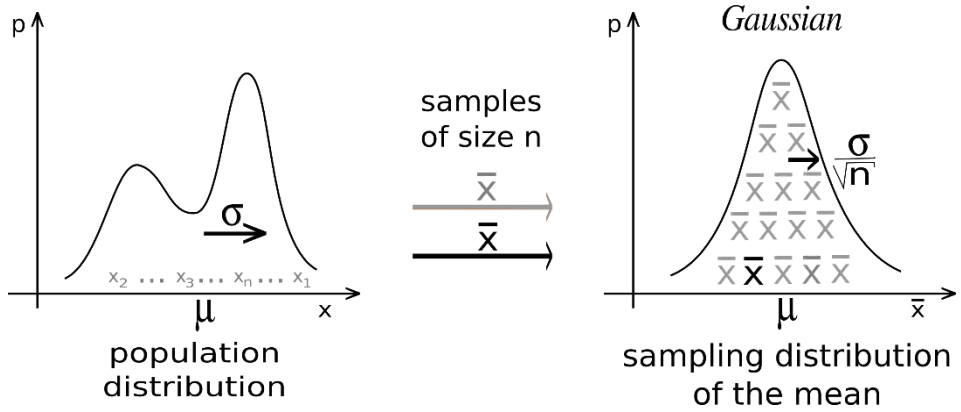
Sampling distribution. If we take certain number of samples and for each sample we compute various statistical measures such as mean, standard deviation, etc., then we can find that each sample may give its own value for the statistic under consideration. All such values of a particular statistic, (e.g., mean), together with their relative frequencies will constitute the sampling distribution of the particular statistic.



Hence, we can have sampling distribution of mean, or the sampling distribution of standard deviation or the sampling distribution of any other statistical measure. It may be noted that each item in a sampling distribution is a particular statistic of a sample. ***However, any sampling distribution tends quite closer to the normal distribution if the number of samples is large (Central Theorem Limit).***

Central Limit Theorem

When sampling is from a normal population, the means of samples drawn from such a population are themselves normally distributed. But when sampling is not from a normal population, the size of the sample plays a critical role. When *n* is small, the shape of the distribution will depend largely on the shape of the parent population, but as *n* gets large ($n > 30$), the shape of the sampling distribution will become more and more like a normal distribution, irrespective of the shape of the parent population.



The theorem which explains this sort of relationship between the shape of the population distribution and the sampling distribution of the mean is known as the central limit theorem. This theorem is by far the most important theorem in statistical inference. It assures that the sampling distribution of the mean approaches normal distribtion as the sample size increases.

The significance of the central limit theorem lies in the fact that it permits us to use sample statistics to make inferences about population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample

Central Limit Theorem

This theorem also gives us a convenient relationship between sample standard deviation (σ, often referred as Standard Error of the Mean, SE) and the standard deviation of the population.

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{N}}$$

An important feature of the standard deviation of the mean is the factor N in the denominator. As sample size increases, the standard deviation of the mean decreases, since the standard deviation of the population does not change.

Example: if we are sampling N=7 from a population in which the mean is 1.36, then the SE will be approximately = 1.36 / sqrt(7) = 0.5

We expect, on average, observed sample means of 1.36, but with a standard deviation of 0.5.

Z-Test

Z score indicates how many standard deviations an observation x is above or below the mean. However, to perform it, I need to have:

- A sample which is randomly chosen from the population
- Mean and variance of the population distribution are known
- Sampling distribution approx. normal (which is, population distributions normal or sample set sufficiently large (N >~ 30)).

In practice, it is quite difficult to verify the conditions for its implementation.

How does it work? Given a μ and σ of a population, we want to test if a sample (from the population) has a significantly different mean than the population. Then:

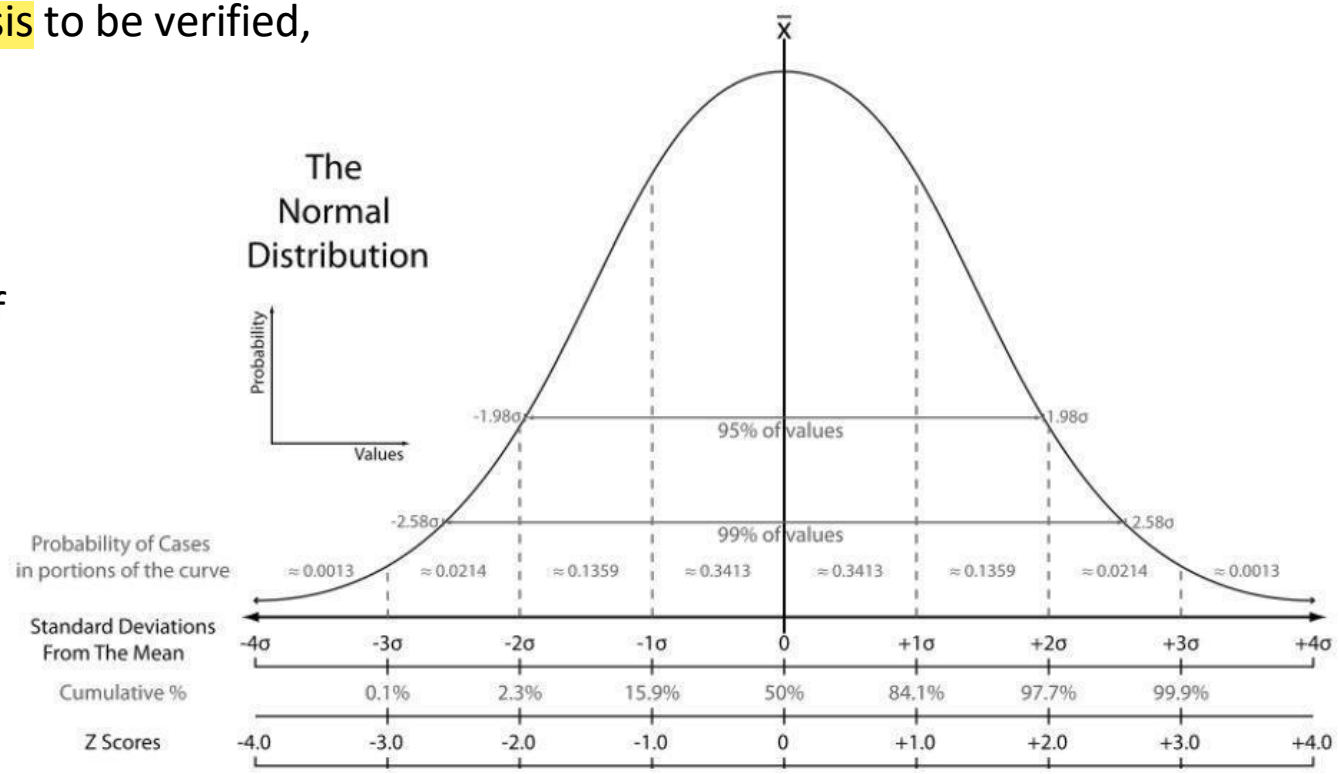
- Take a random sample of size N, and compute the mean (x)
- Compute the Z score
- Look up the Z score in a Z table to obtain the probability that the sample belongs to the population

Z-Test

Z table provides the probability for the null hypothesis to be verified, and it is computed as:

$$Z = \frac{x-\mu}{\sigma}$$

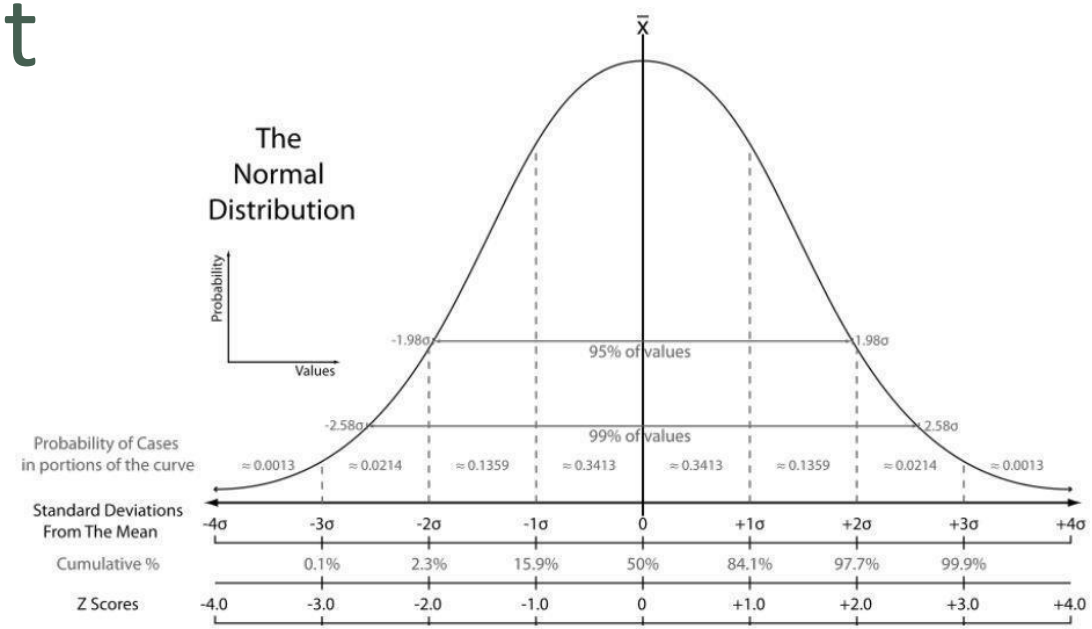
Z=3 : p=99.9% (left tail probability, e.g. probability of getting a z-score less than a given z-score)
Z=0 : p=50%
Z=-1 : p=15.9%
-2<Z<-2 : p=~95%



Z Score Table - Z Table and Z score calculation (z-table.com)

Z-Test

$$Z = \frac{x - \mu}{\sigma}$$



Example:

- Scores of all Italian students in a test
- In Italy: μ=100, σ=12
- A sample of 55 students in Genova obtained an average score
- of 96
- Null hypothesis: Students from Genova are as good as the average Italian students?
- SE = 12 / sqrt(55) = 1.62 -> Z = (96 – 100) / 1.62 = -2.47
- Z-table: the probability of observing a value below -2.47 is approximately 0.68 %
- Reject the null hypothesis

Two-tailed and One-tailed tests

Experimental* and *Control Group: In an experiment, usually we compare data from an experimental group with data from a control group. These two groups should be identical in every respect except one: the difference between a control group and an experimental group is that the independent variable is changed for the experimental group, but is held constant in the control group.

A single experiment may include multiple experimental groups, which may all be compared against the control group.

The purpose of having a control is to rule out other factors which may influence the results of an experiment. Not all experiments include a control group, but those that do are called "controlled experiments."

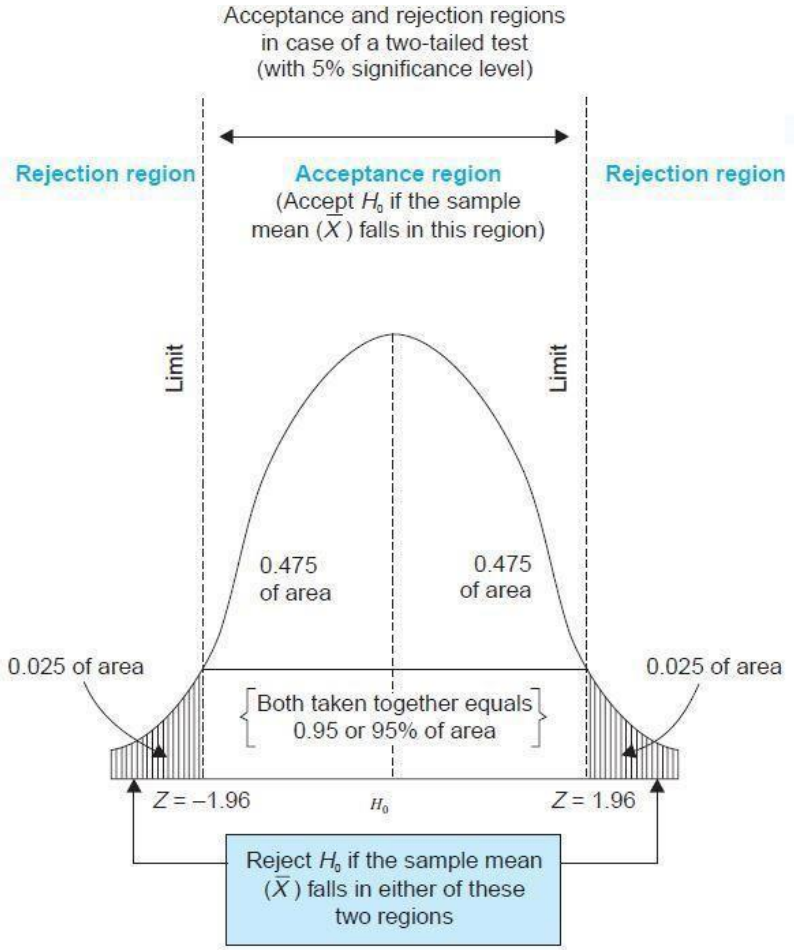
In robotics, this is particularly common, both when we deal with experiments with subjects (e.g., social robotics), or when we want to compare our approach to other state-of-the-art approaches.

Two-tailed and One-tailed tests

When you perform a *z*-test, or a *t*-test (explained later), you check if your test statistic is a more extreme value than expected from the *normal* distribution. There are two main types of tests that can be performed.

Two-tailed and One-tailed tests*:* In the context of hypothesis testing, these two terms are quite important and must be clearly understood. A two-tailed test rejects the null hypothesis if, say, the sample mean is significantly higher or lower than the hypothesised value of the mean of the control group.

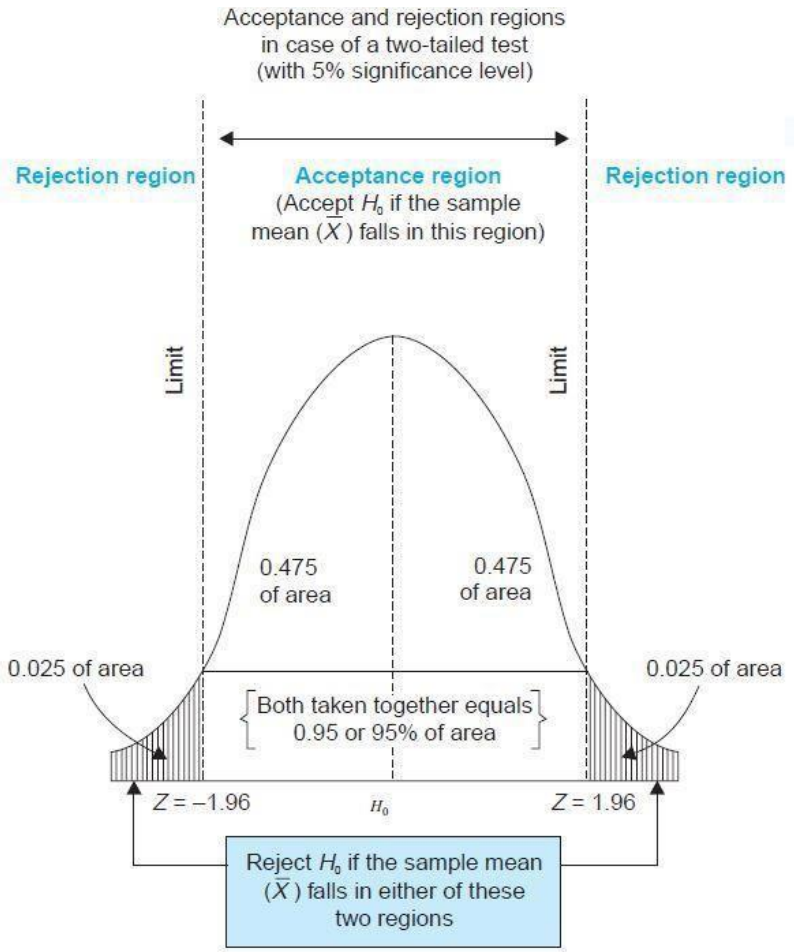
For a two-tailed test, you look at both tails of the distribution: If the test statistic value is either in the lower tail or in the upper tail, you reject the null hypothesis. If the test statistic is within the two reference lines, then you fail to reject the null hypothesis.



Two-tailed and One-tailed tests

If the significance level is 5 per cent and the two-tailed test is to be applied, the probability of the rejection area will be 0.05 (equally splitted on both tails of the curve as 0.025) and that of the acceptance region will be 0.95 as shown in the above curve. If we take m = 100 and if our sample mean deviates significantly from 100 in either direction, then we shall reject the null hypothesis; but if the sample mean does not deviate significantly from m, in that case we shall accept the null hypothesis.

In the figure, Z-scores are standard deviations; this means that we reject the null hypothesis if our sample mean deviates more than 1.96 standard deviations from the mean of our control group.

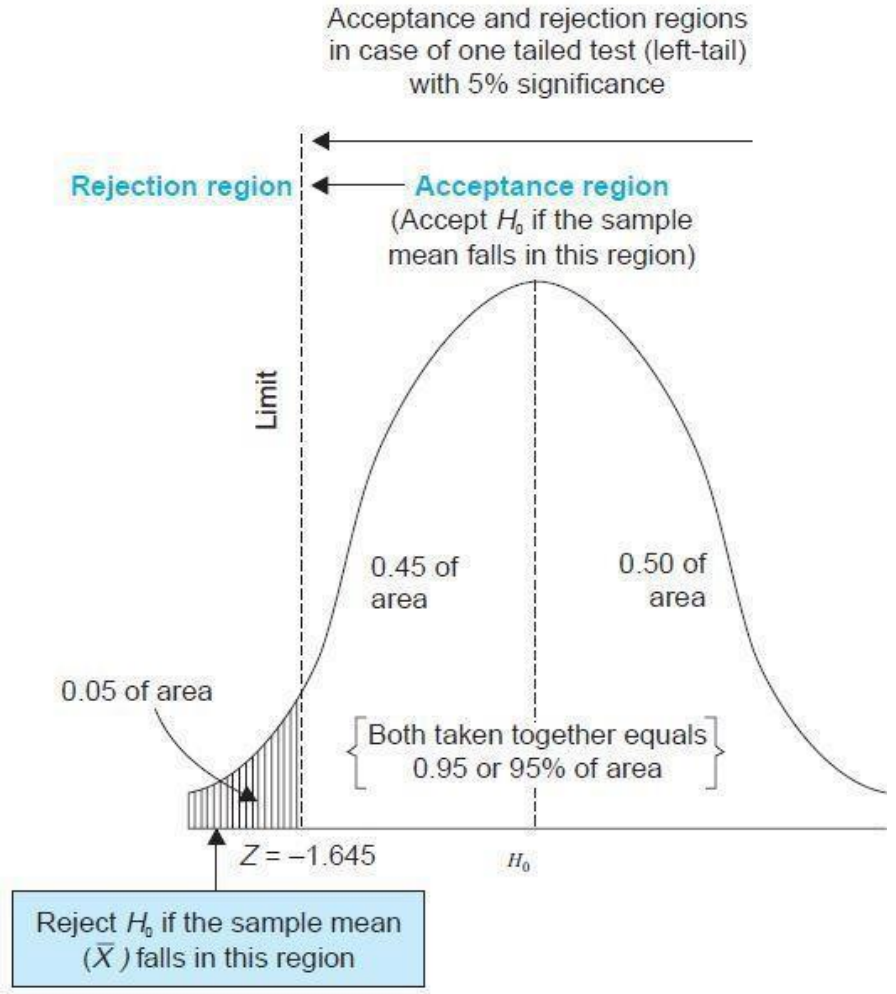


Two-tailed and One-tailed tests

Thus, in a two-tailed test, there are two rejection regions, one on each tail of the curve

For a one-tailed test, you look at only one tail of the distribution. There are situations when only one-tailed test is considered appropriate. A *one-tailed test* would be used when we are to test, say, whether the population mean is either lower than or higher than some hypothesised value (or if the statistics of our experimental group are either lower or higher than the statistics of the control group).

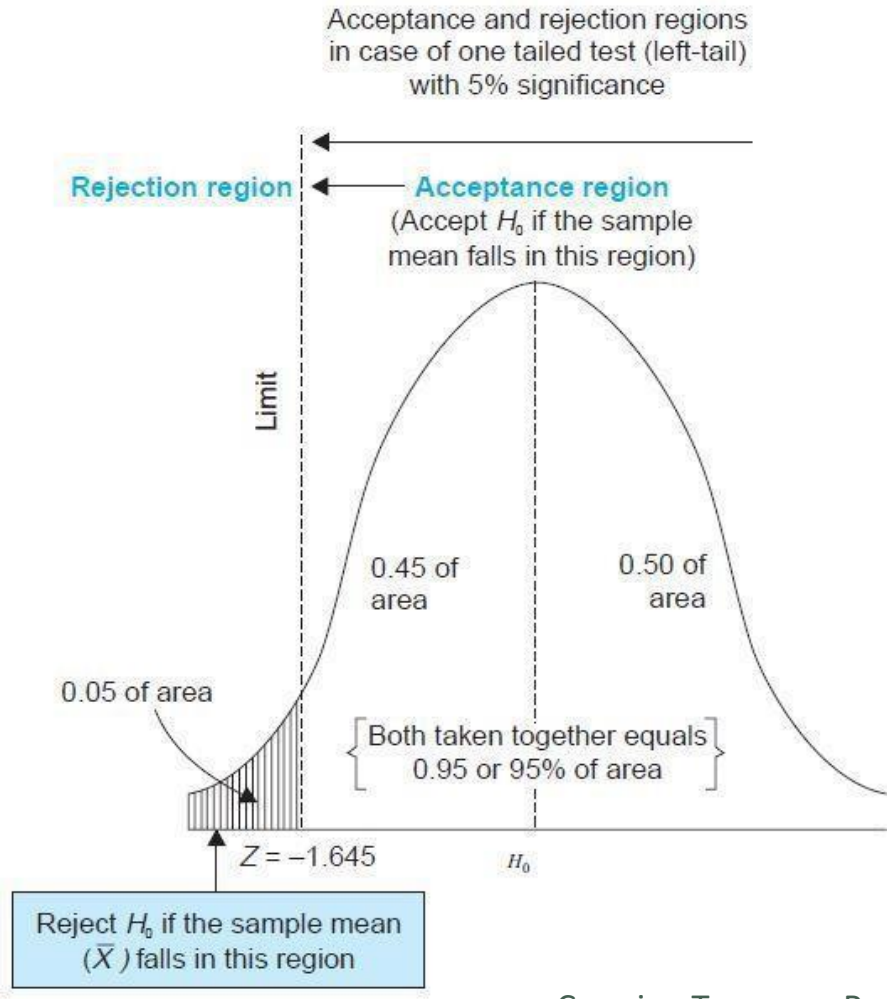
In this case, we can adopt a left-tailed test (or right-tailed test, wherein there is one rejection region only on the left tail or on the right tail)



Two-tailed and One-tailed tests

If the significance level in the given case is kept at 5%, then the rejection region will be equal to 0.05 of area in the left tail as has been shown in the above curve. Given that we are considering only one tail, we could observe how in this case, the deviation of the sample mean from the population mean (or the mean of the control group) should be at least equal to -1.645 (less than before, because just one section of the function is considered).

It should always be remembered that accepting H_0 on the basis of sample information does not constitute the proof that H_0 is true. We only mean that there is no statistical evidence to reject it, but we are certainly not saying that H_0 is true (although we behave as if H_0 is true).



T-Test

T-Test, also known as Student's Test, is based on *t*-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance). The relevant test statistic, *t*, is calculated from the sample data and then compared with its probable value based on *t*-distribution at a specified level of significance for concerning degrees of freedom for accepting or rejecting the null hypothesis.

However, it is still based on the assumption of normality i.e., the source of data is considered to be normally distributed.

How does it work? The t-value is very similar to the z-value

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$

std. dev estimated form the sample

sample size

T-Test

The t-value has to be compared to the values available in a t-table
A t-table shows also a degree of freedom (DoF) which is closely related to the sample size (here: DoF=N-1)

T distribution – Wikipedia

Let's do an example.

confidence level

degree of freedom

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

T-Test

The t-value has to be compared to the values available in a t-table
A t-table shows also a degree of freedom (DoF) which is closely related to the sample size (here: DoF=N-1)

- The average price of a car in city is 12k Euro
- Five cars park in front of a house with an average price of 20,270 Euro and standard deviation of 5,811 Euro
- Null hypothesis (H_0): the cars are not more expensive than in the rest of the city
- DoF=4 (for the one sample t-Test: sample size -1)
- Set confidence level to 95% (5% error probability)
- Since t=3.18 > 2.132 (see t-table) reject H_0
- The cars are significantly more expensive (with 5% error probability)

T-Test

- ✓ Often a 95% or 99% confidence (=5% or 1% significance) level is used
- ✓ t-Test is one of the most frequently used tests in science

However, usually in robotics we need to compare two samples with each other and the population (universe) is unkwnown (e.g. we are comparing two approaches for human-robot interaction, or two path planning algorithms).

In these cases, we want to compare the means of two samples to see if both are drawn from populations with equal means (i.e., from the same population).

Typical null and alternative hypotheses:

$$H_0 \ : \ \mu_1 = \mu_2$$
$$H_1 \ : \ \mu_1 \neq \mu_2 \quad \text{(two-tailored test)}$$
$$H_1 \ : \ \mu_1 < \mu_2 \quad \text{(one-tailored test)}$$
$$H_1 \ : \ \mu_1 > \mu_2 \quad \text{(one-tailored test)}$$

Summary: Procedure for Hypothesis Testing

To test a hypothesis means to tell (on the basis of the data the researcher has collected) whether or not the hypothesis seems to be valid. In hypothesis testing the main question is: whether to accept the null hypothesis or not to accept the null hypothesis? Procedure for hypothesis testing refers to all those steps that we undertake for making a choice between the two actions i.e., rejection and acceptance of a null hypothesis.

- *Making a formal statement:* The step consists in making a formal statement of the null hypothesis (H_0) and also of the alternative hypothesis (H_a). This means that hypotheses should be clearly stated, considering the nature of the research problem. For example, we want to test if our path planning algorithm creates a shorter path to the goal with respect to a well-known state-of-the-art algorithm. We have two hypotheses here:

$$\text{Null hypothesis } H_0\text{: } \mu = \mu_{SOAT}$$
$$\text{Alternative Hypothesis } H_a\text{: } \mu < \mu_{SOAT}$$

The formulation of hypotheses also indicates whether we should use a one-tailed test or a two-tailed test. If H_a is of the type greater than (or of the type lesser than), we use a one-tailed test, but when *Ha* is of the type "different than" then we use a two-tailed test.

Summary: Procedure for Hypothesis Testing

- *Selecting a significance level:* The hypotheses are tested on a pre-determined level of significance and as such the same should be specified. Generally, in practice, either 5% level or 1% level is adopted for the purpose. The factors that affect the level of significance are: (a) the magnitude of the difference between sample means; (b) the size of the samples; (c) the variability of measurements within samples. In brief, the level of significance must be adequate in the context of the purpose and nature of enquiry.

- *Deciding the distribution to use:* After deciding the level of significance, the next step inhypothesis testing is to determine the appropriate sampling distribution, and hence the type of test to be performed. The choice generally remains between normal distribution and the *t*-distribution.

- *Selecting a random sample and computing appropriate values:* Another step is to select a random sample and compute an appropriate value from the sample data concerning the test statistic utilizing the relevant distribution. In other words, draw a sample to furnish empirical data. In practice, this consists in performing the experiments (which may in some cases involves recruiting subjects) and process the acquired data to achieve some statistics (mean, median, standard deviation, ..)

Procedure for Hypothesis Testing

- *Calculation of the probability:* We need to calculate the probability that the sample result would diverge as widely as it has from expectations, if the null hypothesis were in fact true.

- *Comparing the probability:* Yet another step consists in comparing the probability thus calculated with the specified value for the significance level. If the calculated probability is equal to or smaller than the α value in case of one-tailed test (and $\alpha/2$ in case of two-tailed test), then we can reject the null hypothesis (i.e., accept the alternative hypothesis), but if the calculated probability is greater, then we must accept the null hypothesis.

Please remind that in both cases you can commit an error. In case we reject H_0, we run a risk of at most the level of significance

Parametrics and non-Parametric Tests

Several tests of hypotheses (also known as the tests of significance) have been developed, and they can be mainly classified in two categories:

(a) Parametric tests or standard tests of hypotheses;
(b) Non-parametric tests or distribution-free test of hypotheses.

Parametric tests usually assume certain properties of the parent population from which we draw samples. For example, assumptions like: “the observations come from a normal population”, and/or “the sample size is large” must hold before parametric tests can be used. But there are situations when the researcher cannot or does not want to make such assumptions. In such situations we use statistical methods for testing hypotheses which are called non-parametric tests because such tests do not depend on any assumption about the parameters of the parent population.

As a result, non-parametric tests need more observations than parametric tests to achieve the same statistical significance.

Lilliefors Test

In case you are not sure if the data acquired from a sample belongs to a normal distribution, a normality test based on the Kolmogorov–Smirnov test can be applied.

The Lilliefors test is aimed at proving the null hypothesis that the sample belongs to a normal distribution, however, without specifying which normal distribution; i.e., it does not specify the expected value and variance of the distribution

The test proceeds as follows:

- Population mean and population variance are estimated based on the data.

- The maximum discrepancy between the empirical cumulative distribution function and the cumulative distribution function (CDF) of the normal distribution with the estimated mean and estimated variance is computed

- Finally, the test checks if the maximum discrepancy is large enough to be statistically significant, thus requiring rejection of the null hypothesis. To date, tables for this distribution have been computed only by Monte Carlo methods.

Exercises

1) A random variable X follows a normal distribution with $\mu = 6$ and $\sigma = 1.24$. Calculate the following probabilities:

a) $P(X \leq 8)$
b) $P(X \leq 3)$
c) $P(X \geq 7)$
d) $P(7 \leq X \leq 10)$

2) A machine in a factory is responsible for filling boxes of cereal. The weight of cereal in each box is has a mean of 500g and a variance of 20g. A box is picked at random. Calculate the probability that it contains less than 490g of cereal.

Exercises

3) A visual research lab has purchased a digital colour blindness test from a company. Before they can use the test in their research, they must ensure it is properly calibrated. To do this they must check that they get the same results as the company when testing participants with no colour deficiencies. The company states that participants with healthy colour vision will score 15 on the test on average. The research lab tests 13 participants with healthy colour vision. On average they score 12 with a standard deviation of 3.6. Is their machine properly calibrated?

4) The authors of a paper about path planning had tested 20 different environments and found that on average the robot takes 39s to navigate in the environment, with a standard deviation of 4.3 s. They have compared the implemented algorithm with a state-of-the-art approach. The average score in that case is 51s. Is there significant evidence at the 5 % level to support the researcher's claim?