**Sample:** To find things about a population of interest it is common practice to take a sample.

A sample is a selection of objects or observations taken from the population of interest. For example a population might be all apples in an orchard at a given time. We wish to know how big the apples are. We can't measure all of them so we take a sample of some of them and measure them. Different sampling methods: simple, random, convenience etc;

**Sampling Error:** Inference is when we draw conclusions about the population from the sample because the sample is only a selection of objects from the population it will never be a perfect representation of the population. Different samples of the same population will give different results this is called sampling error or variation due to sampling. There will always be sampling error.

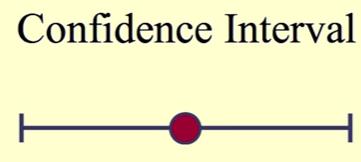
**Confidence Intervals:** When we express an estimate of a population parameter it is good practice to give it as a confidence interval. A confidence interval communicates how accurate our estimate is likely to be. Say we wish to find out how big the apples are in our orchard. We put this as an investigative question what is the mean weight of all the apples in the orchard. We take a sample and calculate the sample mean this is the best estimate of the population mean. We use a confidence interval to express the range in which we are pretty sure the population parameter lies. In this case the population parameter is the mean weight for all the apples in the orchard.

The width of a confidence interval depends on two things the variation within the population of interest and the size of the sample.

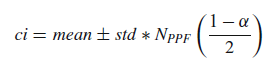
Greater variation in the population leads to a wider confidence interval. Lesser the variation in the population leads to a smaller confidence interval.

Sample size also effects the confidence interval. The effect of sampling error is reduced with larger samples. When we take a large sample the confidence interval can be smaller.

**Example:** Suppose you want to know what percentage of people in the India love Cricket. The only one thing you can do in order to get a 100% correct answer to that question is by asking each one of the citizens in the India whether or not they love cricket. Instead what we do is by getting a random sample of people and get the percentage of people who love cricket in that sample, but then we won’t be 100% confident that this number is right or how far is this number from the real answer, so, what we’ll try to achieve is get an interval, for example, a possible answer to that question may be: “I am 95% confident that the percentage of people that love soccer in the U.S. is between 58% and 62%”. That’s where the name Confidence Interval come from, we have an interval, and we have some confidence about it. If the sample percentage will fall 95% of the time between real percentage-3 and real percentage +3, then the real percentage will be 95% of the times between sample percentage -3 and sample percentage +3. If we took a sample and got 63%, we can say that we 95% confident that the real percentage is between 60% (63 -3) and 66% (63+3). This is the Confidence Interval, the interval is 63+-3 and the confidence is 95%.



The α% *confidence interval* (*CI*) reports the range that contains the true value for the parameter with a likelihood of α%. If the sampling distribution is symmetrical and unimodal (i.e., decaying smoothly on both sides of the maximum), it will often be possible to approximate the confidence interval by:



where *std* is the standard deviation, and *NPPF* the *percentile point function* (*PPF*) for the standard normal distribution.

**UNIT-4**

**Continuous Distributions Derived from the Normal Distribution**

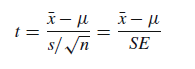
Some frequently encountered continuous distributions are closely related to the normal distribution:

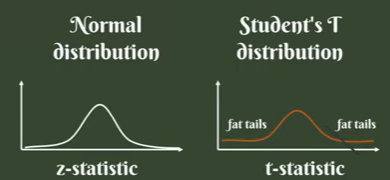
* ***t*-Distribution**: The sample distribution of mean values for samples from a normally distributed population. Typically used for small sample numbers, when the true mean/SD are not known.
* **Chi-Square distribution**: For describing variability of normally distributed data.
* ***F*-distribution**: For comparing variabilities of two sets of normally distributed data.

**t-Distribution (or) Student’s distribution:** William Gosset was an English statistician, developed different methods for the selection of the best yielding varieties of barley – an important ingredient when making beer. Gosset found big samples tedious, so he was trying to develop a way to extract small samples but still come up with meaningful predictions.

Gosset introduced the t-statistic also named as Student’s t. This distribution allowed inference through small samples with an unknown population variance. This distribution has plenty of real time applications.

Visually T-distribution looks similar to normal distribution but has generally has fatter tails. Fatter tails allows for a higher dispersion of variables as there is more uncertainty. As z-statistic is related to standard normal distribution, t-statistic is related to T-distribution:



****

**Figure: T-Distribution vs Normal Distribution**

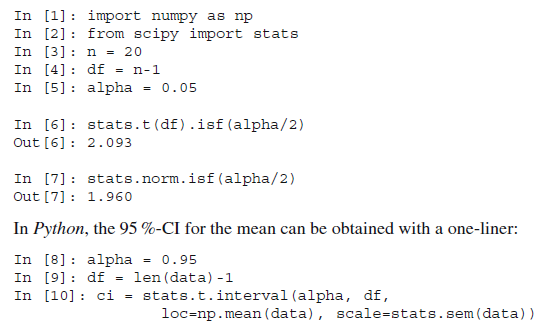
****

***t*-Distribution with DoF=1, DoF=5 and normal**

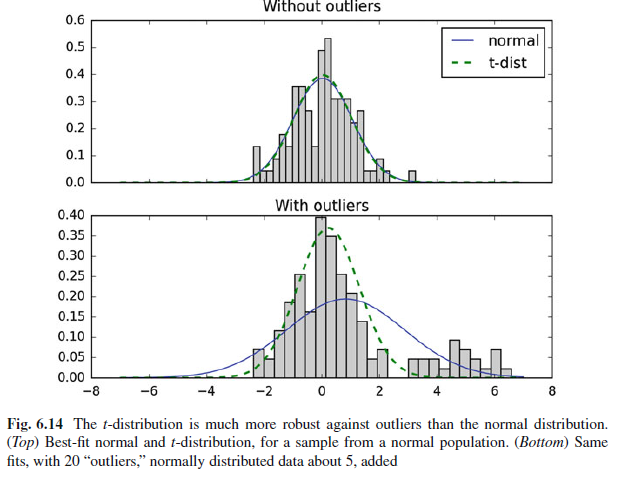
Very frequent application of the *t*-distribution is in the calculation of confidence intervals for the mean.



The following example shows how to calculate the *t*-values for the 95%-CI, for n = 20. The lower end of the 95% CI is the value that is larger than 2.5% of the distribution; and the upper end of the 95%-CI is the value that is larger than 97.5% of the distribution. These values can be obtained either with the *percentile point function* (*PPF*), or with the *inverse survival function* (*ISF*). For comparison, calculate the corresponding value from the normal distribution:

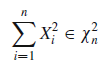


*t*-distribution has longer tails than the normal distribution, it is much less affected by extreme cases.



***Chi-Square Distribution:***

The chi-square distribution is related to the normal distribution in a simple way: if a random variable *X* has a normal distribution (*X ϵ* *N*(0; 1)), then *X*2 has a chisquare distribution, with one degree of freedom(*X2 ϵ* ). The sum of squares of *n* independent and standard normal random variables have a chi-square distribution with *n* degrees of freedom:

**



**Chi-Square Distribution**

**Application:** A pill producer is ordered to deliver pills with a standard deviation of σ =0.05. From the next batch of pills *n* = 13 random samples have a weight of 3.04, 2.94, 3.01, 3.00, 2.94, 2.91, 3.02, 3.04, 3.09, 2.95, 2.99, 3.10, 3.02 g.

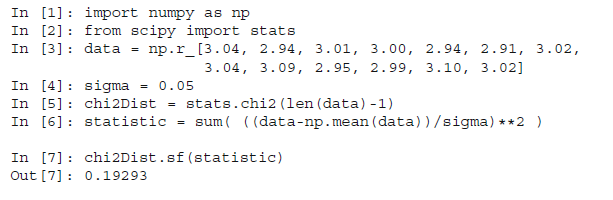
*Question* Is the standard deviation larger than allowed?

*Answer* Since the chi-square distribution describes the distribution of the summed squares of random variates from a *standard normal distribution*, we have to normalize our data before we calculate the corresponding CDF-value:



**Interpretation:**If the batch of pills is from a distribution with a standard deviation of σ =0.05, the likelihood of obtaining a chi-square value as large or larger than the one observed is about 19%, so it is not atypical. In other words, the batch matches the expected standard deviation.

**Note** The number of the DOF is *n*-1, because we are only interested in the shape of the distribution, and the mean value of the *n* data is subtracted from all data points.



***F-Distribution:*** This distribution is named after Sir Ronald Fisher, who developed the *F* distribution for use in determining critical values in ANOVAs (“ANalysis Of VAriance”). If we want to investigate whether two groups have the same variance, we have to calculate the ratio of the sample standard deviations squared:

****

where *Sx* is the sample standard deviation of the first sample, and *Sy* the sample standard deviation of the second sample.

The distribution of this statistic is the *F distribution*. For applications in ANOVAs, the cutoff values for an *F* distribution are generally found using three variables:

• ANOVA numerator degrees of freedom

• ANOVA denominator degrees of freedom

• significance level

An ANOVA compares the size of the variance between two different samples. This is done by dividing the larger variance by the smaller variance. The formula for the resulting *F* statistic is:



where and are the chi-square statistics of sample one and two respectively, and *r*1 and *r*2 are their degrees of freedom.

****

***Figure: F*-distribution**

Take for example the case where we want to compare the precision of two methods to measure eye movements. The two methods can have different accuracy and different precision. As shown in the below Figure, the *accuracy* gives the deviation between the real and the measured value, while the *precision* is determined by the variance of the measurements. With the test we want to determine if the precision of the two methods is equivalent, or if one method is more precise than the other.

****

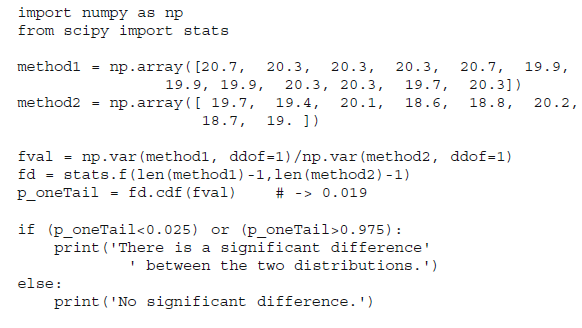
**Accuracy and precision of a measurement are two different characteristics**

**Example:** When you look 20o to the right, you get the following results:

Method 1: [20.7, 20.3, 20.3, 20.3, 20.7, 19.9, 19.9, 19.9, 20.3, 20.3, 19.7, 20.3]

Method 2: [ 19.7, 19.4, 20.1, 18.6, 18.8, 20.2, 18.7, 19. ]

The *F* statistic is *F* = 0:244, and has *n-*1 and *m-*1 degrees of freedom, where *n* and *m* are the number of recordings with each method. The code sample below shows that the *F* statistic is in the tail of the distribution (p\_oneTail=0.019), so we reject the hypothesis that the two methods have the same precision.

****