Statistical tests:

estimation of the sample statistics:

I. t-test:

The sample $x=\{x_1,x_2, \ldots, x_n\}$ *is extracted from a normally distributed population, mean* μ *– known.*

t-test is used to estimate the average value of an individual sample and compare the empirical average with the target value (theoretical).

$H_0: \ \bar{x} = \mu$

statistical criterion:

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

t_{n-1} *– criteria statistics has Student* (T_{n-1}) *distribution with n-1 degrees of freedom, symmetrical; the number of degrees of freedom is the difference in the number of observations and number of estimated parameters*

II. z-test: *(the sample is extracted from a normally distributed population,* μ,σ*-known), statistical criteria:*

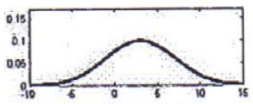
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has a normal distribution N(0,1)

III. *Two samples cases for* ***mean*** *with the different and the same volume:*

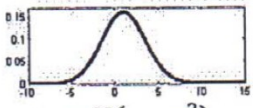
$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$



Two sample cases

$x_1 \sim N(\mu_1, \sigma_1^2)$



$x_2 \sim N(\mu_2, \sigma_2^2)$

$$x_1 - x_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$D^2(X+Y) = D^2X + D^2Y$$

$$D^2(aX) = a^2 \cdot D^2X$$

Confidence interval for $\mu_1 - \mu_2$ (σ_1, σ_2 known)

$$\bar{x}_1 - \bar{x}_2 \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Confidence interval for $\mu_1 - \mu_2$ ($\sigma_1 = \sigma_2$ unknown)

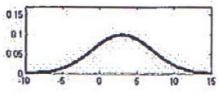
$$\bar{x}_1 - \bar{x}_2 \pm t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

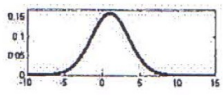
Large sample estimate for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 \pm z_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

IV. *Two samples cases for* ***variance*** *with the different and the same volume:*

Two samples comparing variances

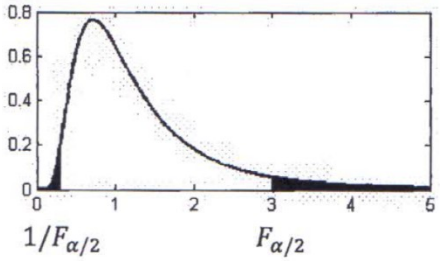
$x_1 \sim N(\mu_1, \sigma_1^2)$ 

$x_2 \sim N(\mu_2, \sigma_2^2)$ 

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

Confidence interval for σ_2^2/σ_1^2

$$\frac{1}{F_{\alpha/2}} \cdot \frac{s_2^2}{s_1^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq F_{\alpha/2} \cdot \frac{s_2^2}{s_1^2}$$



V. ***Pooled variance*** *(совокупная дисперсия):*

Samples $x^{(j)}=\{x_1{}^{(j)}, x_2{}^{(j)}, \ldots, x_{n_j}{}^{(j)}\}$, $j=1,\ldots,k$ *are extracted from a normally distributed population* $N(\mu, \sigma^2)$; μ, σ^2 *– known. Sample variances are* $s_1^2, s_{2,\ldots,}^2 s_k^2$.

Pooled variance estimator *is*

$$S_p^2 = \frac{(n_1-1)s_1^2+\cdots+(n_k-1)s_k^2}{(n_1-1)+\cdots+(n_k-1)},$$

$y=(n_1 - 1) + \cdots + (n_k - 1)$ *– number of degrees of freedom*

MatLab

Hypothesis Tests

t-test, F-test, chi-square goodness-of-fit test, and more

Functions

Distribution Tests

adtest	Anderson-Darling test
chi2gof	Chi-square goodness-of-fit test
crosstab	Cross-tabulation
dwtest	Durbin-Watson test
fishertest	Fisher's exact test
jbtest	Jarque-Bera test
kstest	One-sample Kolmogorov-Smirnov test
kstest2	Two-sample Kolmogorov-Smirnov test
lillietest	Lilliefors test
runstest	Run test for randomness

Distribution tests

One-sample Kolmogorov-Smirnov test

`examgrades.mat` - contains five samples volume 120, normal or not?

kstest

`h = kstest(x)` returns a test decision for the null hypothesis that the data in vector `x` comes from a standard normal distribution, against the alternative that it does not come from such a distribution, using the one-sample Kolmogorov-Smirnov test. The resulth is 1if the test rejects the null hypothesis at the 5% significance level, or `0` otherwise.

`h = kstest(x,Name,Value)` returns a test decision for the one-sample Kolmogorov-Smirnov test with additional options specified by one or more name-value pair arguments. For example, you can test for a distribution other than standard normal, change the significance level, or conduct a one-sided test.

`[h,p] = kstest(__)` also returns the *p*-value p of the hypothesis test, using any of the input arguments from the previous syntaxes.

`[h,p,ksstat,cv] = kstest(__)` also returns the value of the test statistic `ksstat` and the approximate critical value `cv` of the test.

Two-samples Kolmogorov-Smirnov test

kstest2

`h = kstest2(x1,x2)` returns a test decision for the null hypothesis that the data in vectors `x1` and `x2` are from the same continuous distribution, using the two-sample Kolmogorov-Smirnov test. The alternative hypothesis

is that x1 and x2 are from different continuous distributions. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

`h = kstest2(x1,x2,Name,Value`

`[h,p] = kstest2(__)` from the previous syntaxes.

`[h,p,ks2stat] = kstest2(__)` also returns the test statistic `ks2stat`.

Chi-square goodness-of-fit test
chi2gof

The chi-square goodness-of-fit test determines if a data sample comes from a specified probability distribution, with parameters estimated from the data.

The test groups the data into bins, calculating the observed and expected counts for those bins, and computing the chi-square test statistic

$$\chi^2 = \sum_{i=1}^{n} (O_i - E_i)/E_i$$

for theoretical (population) distribution need to add name and parameters `pd = makedist(distname,Name,Value)`

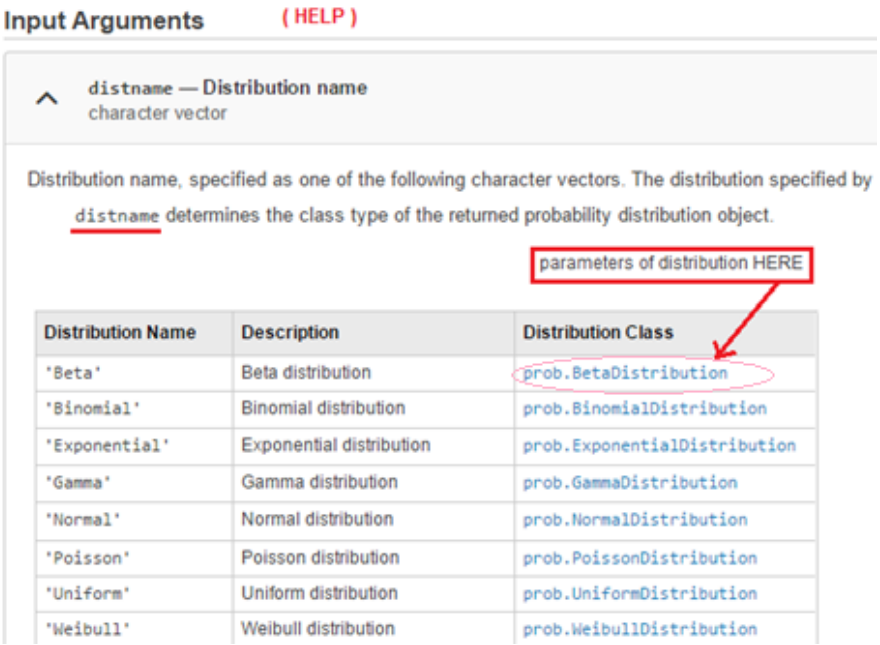
`h = chi2gof(x)` returns a test decision for the null hypothesis that the data in vector x comes from a normal distribution with a mean and variance estimated from x, using the chi-square goodness-of-fit test. The alternative hypothesis is that the data does not come from such a distribution. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

`h = chi2gof(x,Name,Value)` returns a test decision for the chi-square goodness-of-fit test with additional options specified by one or more name-value pair arguments. For example, **you can test for a distribution other than normal**, or change the significance level of the test.

`[h,p] = chi2gof(__)` also returns the *p*-value p of the hypothesis test, using any of the input arguments from the previous syntaxes.

`[h,p,stats] = chi2gof(__)` also returns the structure `stats`, containing information about the test statistic.

In Addition



Location Tests

Location Tests

friedman	Friedman's test
kruskalwallis	Kruskal-Wallis test
multcompare	Multiple comparison test
ranksum	Wilcoxon rank sum test
sampsizepwr	Sample size and power of test
signrank	Wilcoxon signed rank test
signtest	Sign test
ttest	One-sample and paired-sample t-test
ttest2	Two-sample t-test
ztest	z-test

`ranksum` - *Wilcoxon rank sum test*

`p = ranksum(x,y)` returns the *p*-value of a two-sided Wilcoxon rank sum test. `ranksum` tests the null hypothesis that data in x and y are samples from continuous distributions with **equal medians**, against the alternative that they are not. The test assumes that the two samples are independent. x and y can have different lengths.

This test is equivalent to a Mann-Whitney U-test.

`[p,h] = ranksum(x,y)` also returns a logical value indicating the test decision. The result h = 1 indicates a rejection of the null hypothesis, and h = 0 indicates a failure to reject the null hypothesis at the 5% significance level.

`[p,h,stats] = ranksum(x,y)` also returns the structure `stats` with information about the test statistic.

`[__] = ranksum(x,y,Name,Value)` returns any of the output arguments in the previous syntaxes, for a rank sum test with additional options specified by one or more `Name,Value` pair arguments.

`ttest`

$t = (\bar{x} - \mu)/(s/\sqrt{n})$ – *statistics for one sample t-test*

`h = ttest(x)` returns a test decision for the null hypothesis that the data in x comes from a **normal** distribution **with mean equal to zero** and **unknown variance**, using the one-sample *t*-test. The alternative hypothesis is that the population distribution does not have a mean equal to zero. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

`h = ttest(x,y)` returns a test decision for the null hypothesis that the data in `x - y` comes from a normal distribution with mean equal to zero and unknown variance, using the paired-sample *t*-test.

`h = ttest(x,y,Name,Value)` returns a test decision for the paired-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`h = ttest(x,m)` returns a test decision for the null hypothesis that the data in x comes from a normal distribution with mean `m` and unknown variance. The alternative hypothesis is that the mean is not `m`.

`[h,p,ci,stats] = ttest(x,m,Name,Value)` returns the confidence interval `ci`, and the structure `stats` containing information about the test statistic, also returns the *p*-value, p, returns a test decision for the one-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

ttest2

$t = (\bar{x} - \bar{y})/(\sqrt{s_x^2/n_x + s_y^2/n_y})$ *-statistics for two sample t-test* $(df=n_x + n_y - 2)$

and pooled standard deviation $s = \sqrt{s_x^2(n_x - 1) + s_y^2(n_y - 1)}/(n_x + n_y - 2)$

`h = ttest2(x,y)` returns a test decision for the null hypothesis that the data in vectors x and y comes from independent random samples from normal distributions with equal means and equal but unknown variances, using the two-sample *t*-test. The alternative hypothesis is that the data in x and y comes from populations with unequal means. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and `0` otherwise.

`h = ttest2(x,y,Name,Value)` returns a test decision for the two-sample *t*-test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct the test without assuming equal variances.

`[h,p] = ttest2(__)` also returns the *p*-value, p, of the test, using any of the input arguments in the previous syntaxes.

`[h,p,ci,stats] = ttest2(__)` also returns the confidence interval on the difference of the population means, `ci`, and the structure `stats` containing information about the test statistic.

ztest

The *z*-test is a parametric hypothesis test used to determine whether a sample data set comes from a population with a particular mean. The test assumes that the sample data comes from a population with a normal distribution and a known standard deviation.

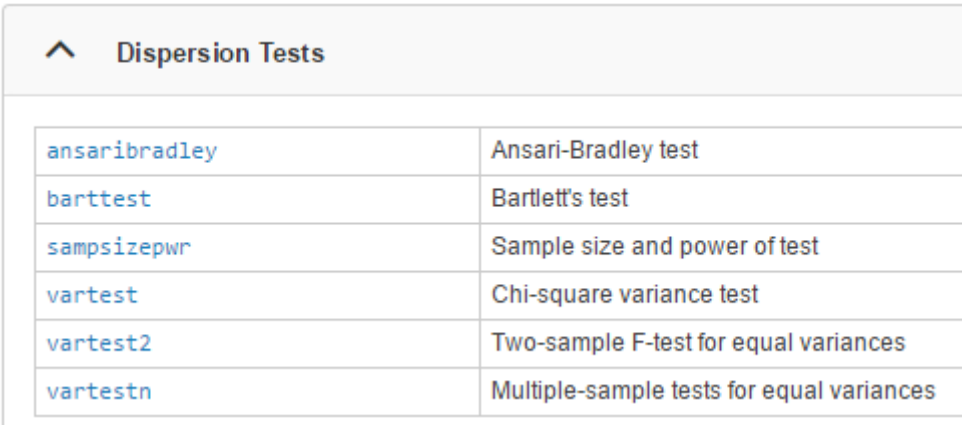
The test statistic is

$t = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ *– statistics for sample z-test, normal distributed*

where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation

multcompare (by yourself)

Dispersion tests



Dispersion Tests

ansaribradley	Ansari-Bradley test
barttest	Bartlett's test
sampsizepwr	Sample size and power of test
vartest	Chi-square variance test
vartest2	Two-sample F-test for equal variances
vartestn	Multiple-sample tests for equal variances

vartest

`h = vartest(x,v)` returns a test decision for the null hypothesis that the data in vector **x comes from a normal distribution with variance v**, using the chi-square variance test. The alternative hypothesis is that x comes from a normal distribution with a different variance. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise.

`h = vartest(x,v,Name,Value)` performs the chi-square variance test with additional options specified by one or more name-value pair arguments. For example, you can change the significance level or conduct a one-sided test.

`[h,p] = vartest(__)` also returns the *p*-value of the test, p, using any of the input arguments in the previous syntaxes.

`[h,p,ci,stats] = vartest(__)` also returns the **confidence interval** for the true variance, **`ci`**, and the structure `stats` containing information about the test statistic.

vartest2 *(by yourself)*

Examples in MatLab statistical functions

t-test in ML:

Two way to check H_0
The probability of observing a t-statistic greater than or equal to the value computed from a sample can be found with the tcdf function:

```
mu = 1; % Population mean
sigma = 2; % Population standard deviation
n = 100; % Sample size
x = normrnd(mu,sigma,n,1); % Random sample from population
xbar = mean(x); % Sample mean
s = std(x); % Sample standard deviation
t = (xbar-mu)/(s/sqrt(n)) % t-statistic
t =
   0.2489
p = 1-tcdf(t,n-1) % Probability of larger t-statistic
```

This probability is the same as the p-value returned by a t-test of the null hypothesis that the sample comes from a normal population with mean μ:

```
[h,ptest] = ttest(x,mu,0.05,'right')
```

Example. To compare theoretical and empirical Probability function

```
x = normrnd(10,2,20,1);
[f,xf] = ecdf(x); % prepare vectors for empirical cdf

stairs(xf,f)    %  piecewise function for [f,xf]  - empirical  cdf
hold on % option for adding new plot
xx=linspace(5,15,100);  % 100 points at equal intervals in[1,15]
yy = normcdf(xx,10,2); % theoretical normal cumulative distributed function - cdf
plot(xx,yy,'r:')
hold off
legend('Empirical cdf','Normal cdf',2)
```

Example. To determine original outcomes. Exponential distribution.

```
x = 0.5:0.2:1.5 % Outcomes
x =
   0.5000  0.7000  0.9000  1.1000  1.3000  1.5000
p = expcdf(x,1) % Cumulative probabilities
p =
   0.3935  0.5034  0.5934  0.6671  0.7275  0.7769
expinv(p,1) % Return original outcomes
ans =
   0.5000  0.7000  0.9000  1.1000  1.3000  1.5000
```

Example. Binomial distribution. The inverse cdf returns the first outcome whose cdf value equals or exceeds the input cumulative probability.

the binoinv function can be used to compute inverses of binomial cumulative probabilities:

```
x = 0.5:0.2:1.5; % Outcomes
p = binocdf(x,10,0.2) % Cumulative probabilities
binoinv(p,10,0.2) % Return binomial outcomes
```

ans =
0 0 0 1 1 1

Example.The inverse cdf is useful in hypothesis testing, where critical outcomes of a test statistic are computed from cumulative significance probabilities. Norminv can be used to compute a 95% confidence interval under the assumption of normal variability:
p = [0.025 0.975]; % Interval containing 95% of [0,1]
x = norminv(p,0,1) % Assume standard normal variability
x =
-1.9600 1.9600 % 95% confidence interval
n = 20; % Sample size
y = normrnd(8,1,n,1); % Random sample (assume mean is unknown)
ybar = mean(y);
*ci = ybar + (1/sqrt(n))*x % Confidence interval for mean*
ci =7.6779

Example. Normplot
x = exprnd(10,100,1);
normplot(x)

Example. are the three groups significantly different from one another?
load carsmall
whos
x = [MPG Horsepower Displacement Weight];
gplotmatrix(x,[],Model_Year,[],'+xo') %(as subplot)
%as a group, are the three years significantly different from one another? The manova1 function can answer that question.
[d,p,stats] = manova1(x,Model_Year)

d - an estimate of the dimension of the group means. If the means were all the same, the dimension would be 0, indicating that the means are at the same point. If the means differed but fell along a line, the dimension would be 1. In the example the dimension is 2, indicating that the group means fall in a plane but not along a line. This is the largest possible dimension for the means of three groups.

p - a vector of p-values for a sequence of tests. The first p-value tests whether the dimension is 0, the next whether the dimension is 1, and so on. In this case both p-values are small. That's why the estimated dimension is 2.

stats - a structure containing several fields, described in the following section

Tests in ML
kstest – Kolmogorov-Smirnov test on one sample normality N(0,1)
kstest2 – two-sample Kolmogorov-Smirnov test to compare the distributions of the values;
chi2gof - performs a chi-square goodness-of-fit test of the default null hypothesis that the data in vector x are a random sample from a normal distribution with mean and variance estimated from x
vartest – a chi-square test of the hypothesis that the data in the vector X comes from a normal distribution with given variance
vartest2– two given independent samples come from normal distributions with the same variance
and so on…