Definition and purpose of the statistical tests

Introduction

In statistics the *null-hypothesis significance testing* is an *inference* method for estimation of how likely the difference of an observed random value from the expected is caused by random variation vs a systematic effect.

The first step is selection of the *null-hypothesis* and the *alternative hypothesis*. Under the null-hypothesis the observed random variable or a random variable derived from the observed one follows a specific model distribution, and the probability of observation of the extreme values from the tail(s) of the model distribution are much less likely than from the core of the distribution.

The second step is selection of the *confidence level*, which can be expressed in percents or fraction of unity. For instance, the confidence level of 95% can be written as $\alpha = 0.95$. Then, under the null hypothesis this confidence level defines one or two *critical values* limiting the range of the values of the test random variable, which are likely to be observed. The values outside this range are unlikely to be observed under the null-hypothesis, thus the test value outside that range forces to reject the null-hyporhesis and to accept the alternative hypothesis. Failure to reject the null-hypothesis does not mean that the null-hypothesis is valid and the alternative hypothesis should be rejected - instead, the test result is inconclusive.

There is a convention to express the test result in terms of *p-value*. In short, it is the probability of observation of values, at least, as exteme as the calculated test value. Basically, lower the p-value less likely it is that the null-hypothesis is valid.

1-sided right-tailed test

Let the test model distribution have the cummulative probability density function be $\Phi(x)$, then for the confidence level α there is the *upper critical value* t_{up} such that

$$\mathtt{Pr}[X \leq t_{up}] = \Phi(t_{up}) = \alpha \Rightarrow t_{up} = \Phi^{-1}(\alpha)$$

For the observed test value *t* the associated p-value is

$$p = \mathtt{Pr}[X \geq t] = 1 - \Phi(t)$$

Thus, the null-hypothesis should be rejected if

$$(t > t_{up} = \Phi^{-1}(\alpha)) \equiv (p = 1 - \Phi(t) < 1 - \alpha)$$

1-sided left-tailed test

In this case, there is the *lower critical value* t_{low} such

$$\mathtt{Pr}[X \geq t_{low}] = 1 - \Phi(t_{low}) = \alpha \Rightarrow t_{low} = \Phi^{-1}(1 - \alpha)$$

For the observed test value *t* the associated p-value is

$$p = \mathtt{Pr}[X \leq t] = \Phi(t)$$

Thus, the null-hypothesis should be rejected if

$$(t < t_{low} = \Phi^{-1}(1-\alpha)) \equiv (p = \Phi(t) < 1-\alpha)$$

2-sided test

In this case both the *upper* and the *lower critical values* are taken into account:

$$\mathtt{Pr}[t_{low} \leq X \leq t_{up}] = \Phi(t_{up}) - \Phi(t_{low}) = \alpha$$

For simplicity, the critical values are chosen such that the cummulative probabilities of the left and right tail are equal

$$\Phi(t_{low}) = 1 - \Phi(t_{up}) = \frac{\alpha}{2} \Rightarrow t_{low} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right); \mathtt{AND}; t_{up} = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$$

The definition of the p-value in this case depends on the relation of the test value to the *median* of the distribution. For the test value greater than the median

$$\Phi(t > \mathtt{Median}) > \frac{1}{2} \Rightarrow p = 2 * (1 - \Phi(t))$$

and for the test value less than the median

$$\Phi(t < \mathtt{Median}) < \frac{1}{2} \Rightarrow p = 2 * \Phi(t)$$

Thus, the null-hypothesis should be rejected if

$$\left(t < t_{low} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right); \mathtt{OR}; t > t_{up} = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right) \equiv (p < 1-\alpha)$$

Z-test

For a random variable following the Gaussian distribution $\mathbb{N}(\mu, \sigma)$ the mean of a sample $\bar{x}$ of length *N* also follows a Gaussian distribution $\mathbb{N}(\mu, \sigma/\sqrt{N})$. Then, the test value

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim \mathbb{N}(0, 1)$$

follows the standard normal distribution, a.k.a. Z-distribution.

Thus, the null- and alternative hypothesis are:

- 2-sided test
  - Null: sample mean equals the population mean (no signicant difference)
  - Alternative: sample mean is not equal to the population mean (significantly different)
- 1-sided right-tailed:
  - Null: sample mean is less than or equal to the population mean
  - Alternative: sample mean is greater than the population mean
- 1-sided left-tailed:
  - Null: sample mean is greater than or equal to the population mean
  - Alternative: sample mean is less than the population mean

Note, that the both population mean and standard deviation are supposed to be known.

Student's t-test

If the population standard deviation is not known, a different test should be used concerning the comparison of the sample mean with the population mean:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}} = \sqrt{\frac{N(N-1)}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} * (\bar{x} - \mu) \sim t_{N-1}$$

which follows the Student's t-distribution with N-1 degrees of freedom, where *s* is the Bessel corrected sample standard deviation used as an estimator of the population standard deviation.

The test hypothesis are the same as in the case of Z-test.

Chi-squared test

For a random variable following the Gaussian distribution $\mathbb{N}(\mu, \sigma)$ the Bessel corrected sample variance s^2 (as an estimator of the population variance) follows the chi-squared distribution with N-1 degrees of freedom, where N is the sample length:

$$t = \frac{(N-1) * s^2}{\sigma^2} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{N-1}$$

Thus, the null- and alternative hypothesis are:

- 2-sided test
  - Null: sample standard deviation equals expected value from the population distribution (no signicant difference)
  - Alternative: sample standard deviation is not equal to the expectation (significantly different)
- 1-sided right-tailed:
  - Null: sample standard deviation is less than or equal to the expected value from the population distribution
  - Alternative: sample standard deviation is greater than the expectation
- 1-sided left-tailed:
  - Null: sample mean is greater than or equal to the expected value from the population distribution
  - Alternative: sample mean is less than the expectation

Note, that the population standard deviation is supposed to be known.

Unpaired Student's t-test

This test is useful for comparison of the two samples' means under the assumption, that:

- the both samples are either pullled from the same population, or from two different populations with equal, or comparable variance
- the respective population(s) follow(s) the Gaussian distribution
- two samples are independent from each other

Let one sample (X_1) be of the length N_1 and the second (X_2) - of the length N_2. Then, the test variable

$$ t = \frac{\bar{x}_1 - \bar{x}*2}{\sqrt{\frac{N_1 + N_2}{N_1 * N_2} \times \frac{(N_1 - 1) * s_1^2 + (N_2 - 1) * s_2^2}{N_1 + N_2 - 2}}} \sim t*{N_1+N_2-2} $$

follows the Student's t-distribution with $N_1 + N_2 - 2$ degrees of freedom.

The null- and alternative hypothesis are:

- 2-sided test
  - Null: the samples' means are equal (no significant difference)
  - Alternative: the samples' means are significantly different
- 1-sided right-tailed
  - Null: the first sample mean is less than or equal to the second sample mean
  - Alternative: the first sample mean is greater than the second sample mean
- 2-sided left-tailed
  - Null: the first sample mean is greater than or equal to the second sample mean
  - Alternative: the first sample mean is less than the second sample mean

The criterion of the comparability of the sample's variance is $\frac{1}{2} \leq \frac{s_1}{s_2} \leq 2$.

Welch t-test

If the relation $\frac{1}{2} \leq \frac{s_1}{s_2} \leq 2$ is not observed, the samples' variances are considered to be too different, thus, instead of the unpaired Student's t-test the Welch t-test should be used. The test variable is

$$ t = \frac{\bar{x}_1 - \bar{x}*2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N^2}}} \sim t*{df} $$

where the number of the degrees of freedom is

$$\mathtt{df} = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N^2}\right)^2}{\frac{s_1^4}{N_1^2*(N_1-1)} + \frac{s_2^4}{N_2^2*(N_2-1)}}$$

The null- and alternative hypothesis are the same as in the previous case.

Paired Student's t-test

The paired Student's t-test is applied when the two samples represent two independent measurements on the same subjects / objects. For instance, measurements with two different instruments or methods, or measurements with the same instrument / method on the same subjects before and after specific treatment. Thus, it shows if the systematic difference is significant in comparison with the random variation. By definition, the both samples are of the same length N.

In this test, the *difference* random variable is constructed first
$Y = X_1 - X_2;:;y_i = x_{1,i} - x_{2,i};\forall;1 \leq i \leq N$, and the test variable is

$$t = \frac{\bar{y} - \mu_0}{s_y/\sqrt{N}} \sim t_{N-1}$$

where μ_0 is the *expected* measurement bias value, or expected threshold value, for which the difference between the methods is supposed to be signicant.

The null- and alternative hypothesis are:

- 2-sided
  - Null: the difference in the methods is about μ_0, or the both methods yeild equal results for μ_0 = 0
  - Alternative: the difference between the methods results is significantly different from μ_0, or the methods produce significantly different results for μ_0 = 0
- 1-sided righ-tailed
  - Null: the difference between the methods results does not exceed μ_0
  - Alternative: the difference between the methods results significantly exceeds μ_0
- 1-sided left-tailed
  - Null: the difference between the methods results is greater than or equal to μ_0
  - Alternative: the difference between the methods results is significantly less than μ_0

In practice, μ_0=0 is often used.

F-test on variance equality

If the two samples are pulled from the population(s) following Gaussian distribution, or, at least, the samples' variances follow (scaled) chi-squared distribution, the test variable

$$t = \frac{s_1^2}{s_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} = \frac{s_1^2}{s_2^2} \times \delta \sim F_{N_1-1,N_2-1}$$

follows the F-distribution with $N_1 - 1$ and $N_2 - 1$ degrees of freedom, where N_1 and N_2 are the lengths of the samples, and σ_1 and σ_2 are the standard deviations of the underlying population distributions.

In practice, the underlying populations are often supposed to have the same variance / standard deviation, i.e $\delta = 1$, in which case the test hypothesis are:

- 2-sided
  - Null: the samples variances are equal (no significantly different)
  - Alternative: the samples variances are significantly different
- 1-sided right-tailed
  - Null: the first sample variance is less than or equal to the second sample variance
  - Alternative: the first sample variance is significantly greater than the second sample variance
- 1-sided left-tailed
  - Null: the first sample variance is greater than or equal to the second sample variance
  - Alternative: the first sample variance is significantly less than the second sample variance

ANOVA F-test

In general, ANOVA F-test concerns the statistical significance of the difference between the groups (samples) pooled, expectedly, from the same population in comparison with the variation within the samples. It's advantage is that it can be applied to an arbitrary number of groups / samples. In the case of only two groups, the test value is

$$ t = (N_1 + N_2 - 2) \frac{N_1 * (\bar{x} *1 - \bar{x})^2 + N_2 * (\bar{x}2 - \bar{x})^2}{(N_1 - 1) * s_1^2 + (N_2 - 1) * s_2^2} \sim F{1, N_1 + N_2 - 2} \sim t*{N_1 + N_2 -2}^2 $$

where the total mean is

$$\bar{x} = \frac{N_1 * \bar{x}_1 + N_2 * \bar{x}_2}{N_1 + N_2}$$

In this interpretation, this test does not provide any benefits in comparison to the unpaired Student's t-test or Welch t-test. However, it has a different interpretation: if two samples are pooled from the same population, but, for some reason, they have very different sample variances. For instance, one sample has unusually lower variance than statistically expected, and / or the other sample has unusually higher variance than statistically expected. In this case, the sample are likely to have significantly different samples means. I.e. this test can be used for the assesment of the *homoscedasticity* of the samples - if they have comparable variances.

Thus, the test is 1-sided right-tailed with the hypothesis as:

- Null: the both samples have equal variances (no significant difference)
- Alternative: the samples variances are significantly different

Levene's test

In this test the observed random variables are transformed as

$$ X_1 \rightarrow Y_1 ; : ; y_{1,i} = |x_{1, i} - \bar{x} *1/ \newline X_2 \rightarrow Y_2 ; : ; y*{2,i} = |x_{2, i} - \bar{x}_2| $$

where */x/* is the absolute value of the variable *x*. Then the ANOVA F-test is applied to the tranformed random variable.

The benefit of this test is that it can be applied for the comparison of the variances of the samples pooled from *a priori* different populaions, with the populations' means expected to be different. The tranformation above minimizes the effect of the populations' means difference, whereas it emphasises the influence of the samples variances.

The statistical test hypothesis are the same.

Brown-Forsythe test

In the Brwon-Forsythe test a different transformation is used:

$$ X_1 \rightarrow Y_1 ; : ; y_{1,i} = |x_{1, i} - \tilde{x} *1/ \newline X_2 \rightarrow Y_2 ; : ; y*{2,i} = |x_{2, i} - \tilde{x}_2| $$

where $\tilde{x}$ is the *median* value of the sample. Use of the median instead of the sample's mean results in more robust test, especially in the case of heavy-tailed (high kurtosis) or very assymetric (large positive or negative value skewness) underlying distribution(s).

The statistical test hypothesis are the same.