hypothesis test
A method for gaining confidence in a hypothesis (H_A) by succeeding to reject a contradictory hypothesis, the *null hypothesis* (H_0). The amount of confidence gained by performing the test depends on the *power of the test* which, in turn, depends on the chosen *level of significance*. For instance, when succeeding to reject a null hypothesis at an $\alpha = 0.01$ level of significance, the *probability* of (erroneously) rejecting a true null hypothesis is 1%.

The following procedure is used to conduct a hypothesis test:

(1) A null hypothesis H_0 is stated.

(2) An *alternative hypothesis* H_A is stated.

(3) A level of significance α is stated.

(4) An *experiment* is conducted, resulting in *observations*.

(5) A *test statistic* T is computed from the observations.

(6) If $P(T) > 1 - \alpha$, H_0 is rejected and H_A may be accepted.

Hypothesis tests use specific *probability distributions*, like the χ^2-distribution ($\rightarrow$ *chi-square distribution*) or the *t-distribution* to calculate the *improbability* of an observation given a null hypothesis. I.e., the value $P(T) = F_X(T)$ of the *cumulative distribution function* (CDF) of the distribution X is exactly the improbability of the observation from which the test statistic T was computed. The *complement* $1 - T$ of T is commonly called the *p-value*.

The level of significance α is the complement of the threshold where an observation is deemed so improbable that the null hypothesis has to be rejected. For instance, if $\alpha = 0.05$, then a value of $F_X(T) \geq 0.95$ would lead to the rejection of H_0. (There are also left-tailed tests, where the condition for rejecting H_0 would be $F_X(T) \leq \alpha$).

The intervals between $1 - \alpha$ and 1 and/or 0 and α are called the *critical regions* of a test. A test statistic T whose probablity falls within a critical region means that H_0 has to be rejected. The interval outside of the critical region(s) may be considered to be a

confidence interval at a *confidence level* of c for the observed data to be explained by the null hypothesis.

Some tests use *two-tailed* probability distributions. In this case, there are two critical regions and, subsequently, two conditions for rejecting H_0: $F_X(T) \leq \frac{\alpha}{2}$ and $F_X(T) \geq c + \frac{\alpha}{2}$. See figure HTD for illustrations. For an illustration of a left-tailed test, see *z-test for location*.

When the null hypothesis H_0 can be rejected during a hypothesis test with a significance level of α, the result of the test is stated as "the null hypothesis could be rejected at a α level of significance" or, alternatively, "the alternative hypothesis could be accepted at a α level of significance". Both wordings indicate that there is a probability of $p = \alpha$ of the null hypothesis still being valid while the data gathered in the experiment arose by chance. In other words, there is a probability of $p = \alpha$ of committing a *type I error*.

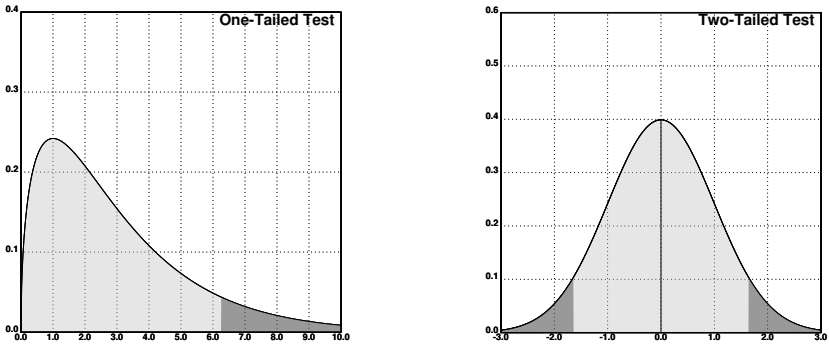

Figure **HTD:** hypothesis test distributions; left: one-tailed $\chi^2(3)$-distribution with one critical region with probability $p = \alpha$; right: two-tailed Z-distribution with two critical regions with probability $p = \frac{\alpha}{2}$ each; both panels: light gray area: confidence interval, dark gray area(s): critical region(s); $\alpha = 0.1$, $c = 0.9$

When the null hypothesis cannot be rejected during a test, the test has no result. In particular, failure to reject H_0 does not mean that the null hypothesis is valid or the alternative hypothesis is invalid. A probability for the test statistic T that is close to $1 - \alpha$ often indicates that further research is advisable.

Example: It is to be shown that a six-sided die is not balanced, i.e. some faces will show up more often than others. Then the null

hypothesis H_0 would be that the die is balanced, so that in a series of *trials* that is a multiple of 6 (total of faces) each face shows up the same number of times. The alternative hypothesis H_A would be that the die is not balanced, so some faces show up more often than others. The experiment would be to cast the die $n = 6m$ times, expecting the following *frequency distribution*, which describes an ideal model of a balanced die. Setting $m = 20$:

Eyes	1	2	3	4	5	6	Total
Expected Frequency	20	20	20	20	20	20	120

The level of significance is set at $\alpha = 0.1$, giving a threshold of $p = 0.9$. The actual experiment may then yield the following outcomes:

Eyes	1	2	3	4	5	6	Total
Observed Frequency	27	15	23	16	26	13	120

From the observation and the expectation a χ^2 statistic (→ *chi-square statistic*) can be created:

$$\mathrm{X}^2 = \sum_{i=1}^{6} \frac{(O_i - 20)^2}{20} = 9.2$$

where each O_i is an observed frequency. The CDF of the χ^2-distribution with 5 *degrees of freedom* can then be used to compute the improbability of the outcome of the experiment while H_0 holds:

$$F_{\chi^2(5)}(9.2) \approx 0.899$$

Because $0.899 < 0.9$, the null hypothesis cannot be rejected at a $\alpha = 0.1$ level of significance. The data from the experiment is insufficient to conclude that the die is not balanced. However, the test result is so close to the threshold that further experiments seem advisable.

I,J