Chapter 2

Statistical inference

Statistical inference (*tilastollinen päättely*) is the mathematical theory behind estimates and their distributions. Estimates can be constructed in a way that statistical hypothesis can be tested against their distributions. Estimate and its distribution is the link between model (i.e. distribution and its parameters) and data.

2.1 Likelihood

Likelihood (*uskottavuus*) is the key concept in statistical inference. The theory is developed by R.A. Fisher at the beginning of the 20th century. Likelihood deals with data, model, and parameters. First of all, we need to have a model. Model is the statistical distribution that we believe the random variable Y should obey, so the model is probability density function $\mathrm{f}_Y(\cdot)$. Model has parameters and their values are unknown. In likelihood problems the parameter vector is often noted with $\boldsymbol{\theta}$, although individual distributions usually have traditional conventions with the parameter symbols. For example, normal distribution has $\boldsymbol{\theta} = (\mu, \sigma^2)$.

The final component in likelihood is data. Very seldom we are doing inference based on single observation y, almost always the data consists of observations $y_1, \ldots, y_n$. In that case the data is a vector of observations, $\boldsymbol{y}$. In more general case the data is vector of multidimensional observations, i.e. matrix $\mathbf{Y}$.

We are not dealing with random processes here, so the observations y_i are identically distributed and the model or its parameters are not assumed to change with time. If there is (auto)correlation between consecutive observations (y_i, y_{i+k}) we are dealing with time series (*aikasarja*), but here we do not consider such cases. We limit ourselves to independent observations, so together with the assumption of non-varying model we deal with i.i.d. observations $\boldsymbol{y} = (y_1, \ldots, y_n)$.

The idea of likelihood is quite simple and straightforward. Let us say that we have reasons to believe that our data is from process that can be described with normal

distribution with fixed and known variance of 1. The unknown parameter is the expectancy μ. What if we have one observation y_1? We cannot say much, but our best guess would be that $\mu = y_1$, as in Fig. 2.1 a).

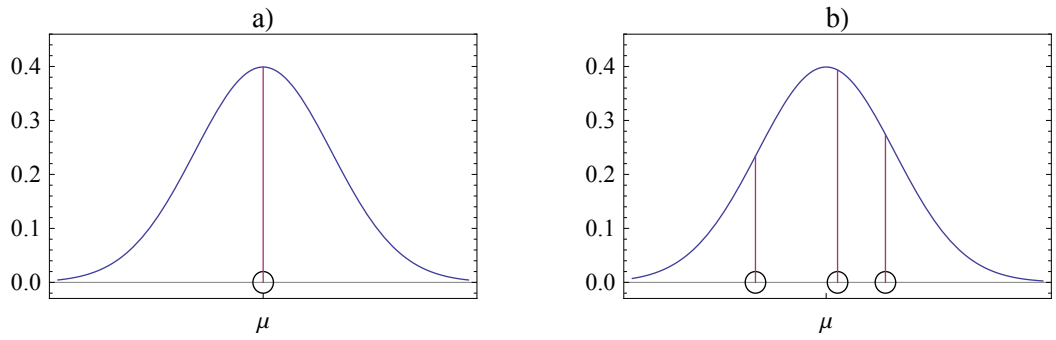

Figure 2.1: Example of normal model with one observation (a) and with three observations (b).

Next, we consider case with three observations $\boldsymbol{y} = (y_1, y_2, y_3)$, as in Fig. 2.1 b). Intuitively, we should place our normal distribution so that it would somehow fit to all three observations in the best possible way. What is the best possible way? If our model $Y \sim \mathcal{N}(\mu, 1)$ is correct, the probability (density) of observing $Y = y_1$ can be computed from $\mathrm{f}_Y(y_1; \mu, 1)$. As the observations are i.i.d., the joint probability of observing all three can be computed as a product of individual probabilities (densities), $\mathrm{f}_Y(\boldsymbol{y}; \mu, 1) = \mathrm{f}_Y(y_1; \mu, 1) \times \mathrm{f}_Y(y_2; \mu, 1) \times \mathrm{f}_Y(y_3; \mu, 1)$. Please note that with likelihood and related fields both the data and the parameters are usually written out with the pdf as $\mathrm{f}_Y(\boldsymbol{y}; \boldsymbol{\theta})$. The abovementioned procedure is, in a nutshell, the likelihood principle.

2.1.1 Likelihood function

Following the previous procedure we can formulate the likelihood function $\mathrm{L}(\cdot)$ in a more formal way. Likelihood function is

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \mathrm{c}(\boldsymbol{y})\, \mathrm{f}_Y(\boldsymbol{y}; \boldsymbol{\theta}), \tag{2.1}$$

where the pdf is the joint density function for $\boldsymbol{y}$. Note the small change of paradigm — likelihood function is used to estimate the unknown parameter vector $\boldsymbol{\theta}$, so that is the main parameter of the function, the observed data $\boldsymbol{y}$ is a 'secondary parameter'.

The function $\mathrm{c}(\boldsymbol{y})$ in Eq. (2.1) can be any function involving only the data and not the parameter vector, and in that sense the likelihood function is not uniquely defined. Any function $\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}) \propto \mathrm{f}_Y(\boldsymbol{y}; \boldsymbol{\theta})$ is likelihood function. This fact can be used to clean out unnecessary constants (i.e. terms independent of $\boldsymbol{\theta}$) from the likelihood, making it a bit simpler.

If we have i.i.d. observations, as we do in almost all the examples here, the likelihood function is the product of the one-dimensional distributions:

$$\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}) \propto \prod_{i=1}^{n} \mathrm{f}_Y(y_i; \boldsymbol{\theta}) \text{ , if } \boldsymbol{y} \text{ is i.i.d.} \tag{2.2}$$

The likelihood function is used together with maximum likelihood principle (*suurimman uskottavuuden periaate*). The principle simply states, that we should find values (i.e. estimates) for our unknown parameters $\boldsymbol{\theta}$ so that it will maximize the likelihood function for observed data $\boldsymbol{y}$. As L is defined through the joint probability density, we are essentially maximizing the probability of parameter values, given the data.

In the example in Fig. 2.1 b) we have three observed values: $-1.2, 0, 0.7$. The likelihood function is $\mathrm{L}(\mu) \propto \exp\left(-\left((-1.2-\mu)^2 + (0-\mu)^2 + (0.7-\mu)^2\right)/2\right)$. It is not too hard to see that setting $\mu = -1/6$ will maximize the likelihood, see Fig. 2.2.

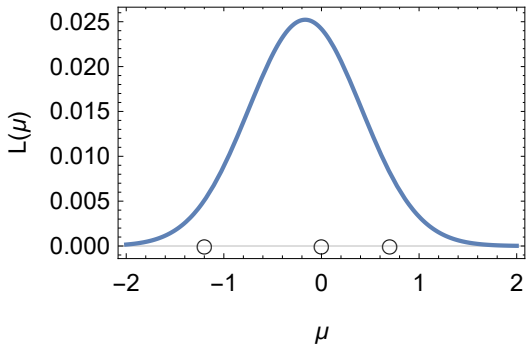

Figure 2.2: Likelihood function of normal model with three observations as in Fig. 2.1 b).

Log-likelihood function

The likelihood function is a product of pdf's, and the aim is to maximize that. Taking any monotonic and increasing function of L will not alter the values where the function reaches its extrema points. The logarithm function can be used to reduce the likelihood into simpler form, because logarithm of product is sum of logarithms. Therefore, maximum likelihood problems are often solved through log-likelihood function (*log-uskottavuusfunktio*). Log-likelihood function $\mathrm{l}(\cdot)$ is simply

$$\mathrm{l}(\boldsymbol{\theta}; \boldsymbol{y}) = \log\left(\mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}\right), \tag{2.3}$$

where $\log$ stands for natural logarithm. Another convenient property of logarithm is that $\log(\exp(x)) = x$. Many statistical distributions belong to the so-called exponential family, normal distribution being one of them, so the exponential form

in likelihood function is quite common. With log-likelihood one can change from product of exponentials to sum without exponent functions.

With log-likelihood function our example in Fig. 2.1 b) would reduce to task of maximizing $\mathrm{l}(\mu) \propto -((-1.2-\mu)^2 + (0-\mu)^2 + (0.7-\mu)^2)$, see Fig. 2.3.

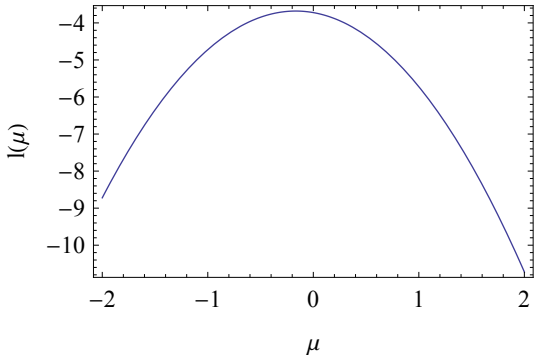

Figure 2.3: Log-likelihood function of normal model with three observations as in Fig. 2.1 b).

2.1.2 Maximum likelihood estimate

The concept of likelihood defines the maximum likelihood (ML) principle (*suurimman uskottavuuden periaate*) in statistics. The maximum likelihood estimate (MLE) of the unknown parameter in our probability model, given the data, is the value $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood (or log-likelihood) function:

$$\mathrm{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq \mathrm{L}(\boldsymbol{\theta}; \boldsymbol{y}) \; \forall \boldsymbol{\theta}. \tag{2.4}$$

This $\hat{\boldsymbol{\theta}}$ is the point-estimate (*piste-estimaatti*) to $\boldsymbol{\theta}$.

In most of the cases the likelihood and log-likelihood functions are at least twice differentiable over the whole parameter space. If this is the case, the MLE can be found by studying the first and second derivatives of the (log-)likelihood function. Extrema points of continuous and differentiable functions have zero value of the first derivative. Furthermore, if the extremum point is maximum, the value of the second derivative is negative.

The conditions described before form the so-called likelihood equation. In the general case the parameter is a vector (of length d here), and the vector of first partial derivatives is called the score function $\mathrm{u}(\cdot)$:

$$\mathrm{u}(\boldsymbol{\theta}; \boldsymbol{y}) = \nabla \mathrm{l}(\boldsymbol{\theta}; \boldsymbol{y}) = \left(\frac{\partial \mathrm{l}}{\partial \theta_1}, \ldots, \frac{\partial \mathrm{l}}{\partial \theta_d} \right), \tag{2.5}$$

and the Hessian matrix $\mathbf{H}$ is the matrix of second order partial derivatives:

$$\mathbf{H} = \nabla \nabla^T \mathrm{l}(\boldsymbol{\theta}; \boldsymbol{y}) = \left[\frac{\partial^2 \mathrm{l}}{\partial \theta_i \, \partial \theta_j} \right]_{ij}. \tag{2.6}$$

With these notations, the MLE satisfies the likelihood equation, i.e. $\mathrm{u}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) = \mathbf{0}$ and $\mathbf{H}$ at $\hat{\boldsymbol{\theta}}$ is negative definite.

Properties of maximum likelihood estimate

MLE has some nice properties which make it even more important in statistics. We list the most important here, invariance and asymptotic properties. First, MLE is invariant in re-parametrization. If we would change our parameter of interest so that we would use parameter $\phi := \mathrm{g}(\theta)$, the MLE of the re-parametrized model would still be $\hat{\phi} = \mathrm{g}(\hat{\theta})$.

What is even more important with MLE is that we know its asymptotic distribution, and it is the normal distribution. The proof of that relies on the central limit theorem, but is far too cumbersome for us. So, without proof, we state that

$$\hat{\boldsymbol{\theta}} \xrightarrow{\sim} \mathcal{N}_d \left(\boldsymbol{\theta}, -\mathbf{H}^{-1} \right). \tag{2.7}$$

That means, at least, four things. First of all, it states that if we have 'enough' data, the MLE will approximately obey normal distribution. Note that as the parameter here is a vector, the distribution is multidimensional.

Second, the MLE is unbiased. This means that the expectation of MLE is the 'true' $\boldsymbol{\theta}$. Third, the MLE is efficient. This concept has not been mentioned here, but it means that the variance of MLE is the smallest possible over all estimators.

Fourth consequence is very important in practice — we have a asymptotic variance for the MLE, so we know how much it typically varies around true $\boldsymbol{\theta}$. This is the basis for confidence intervals and statistical tests. The asymptotic variance for vector parameter is expressed through the expectation of the Hessian matrix, i.e. the second partial derivatives of the log-likelihood function. While this may seem a bit cumbersome, the good thing is that we usually do not need to derive estimators and their variances ourselves. Somebody else has gone through the trouble and done that for us using the abovementioned equations. For many practical cases the formulas can be reduced to quite simple forms, for example that the variance of mean $\overline{x}$ for normal model is σ^2/n.

2.2 Statistical tests

From estimators and their distributions we can continue to statistical tests and confidence intervals. Let us first deal with confidence intervals.

2.2.1 Confidence intervals

The MLE is a point-estimate, it gives us the most probable value for the unknown parameter of our model. In the same manner, any statistics, whether MLE or any

other $t := t(\boldsymbol{y})$, are point-estimates. On the other hand, the data that we have observed, $\boldsymbol{y}$, is just one possible outcome of the random process. If we would repeat the experiment or redo the observations, we would get different data vector $\boldsymbol{y}^*$. Following the thought, we would also get another value for the statistics, t^*, that would probably differ from the original t. As the observations $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are both realizations of a random variable $\boldsymbol{Y}$, also the *estimates* t and t^* are realizations of a random *estimator* $T := t(\boldsymbol{Y})$.

For that reason, often the point-estimate alone is not enough for us for data-analysis purposes. A more interesting would be to know an interval where the statistics would most probably be, even if we would repeat the experiment over and over again. This interval is called confidence interval (CI; *luottamusväli*), or credible interval in Bayesian inference.

The p 100 % confidence interval (e.g. 95 %) for parameter $\boldsymbol{\theta}$ is the region where the true value of parameter lies, with p 100 % confidence. More formally

$$\mathrm{P}(\boldsymbol{\theta} \in \Omega_p) = p, \tag{2.8}$$

although there are some philosophical issues in frequentist probability concept that require slightly different formulation*. The Eq. (2.8) does not define how the area Ω_p is chosen. There are some options for that, but with symmetric distributions (of T) all the options lead to the same conclusion — the area Ω_p should be chosen so that it is a symmetric interval around the $\boldsymbol{\theta}$, and only $(1-p)$ 100 % of the density is left out from the tails of the pdf. Thus, CI for one-dimensional parameter and symmetric distribution is such that

$$\mathrm{P}(\hat{\theta} - c \leq \theta \leq \hat{\theta} + c) = p. \tag{2.9}$$

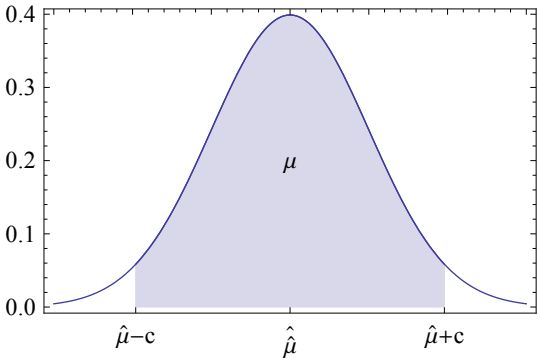

Figure 2.4: Confidence interval $(\hat{\mu} - c, \hat{\mu} + c)$ for μ, when data is from normal distribution.

*Actually, in frequentist sense the parameter value is an unknown but constant value, and probability is not meaningful for it. The interval should be formulated using statistics as random variable, $T := t(\boldsymbol{Y})$. Still, in practice the interpretation is more or less the same, and in Bayesian concept it is allowed to speak about the probability of the parameter.

Confidence interval for mean

Mean $\overline{y}$ is the most common statistics. With normal distribution as model, it is the MLE for expected value, but the same is true for many other (symmetric) distributions and their location parameters. And, due to the asymptotic behavior of mean, normal distribution is at least its asymptotic distribution.

The CI for mean and (asymptotic) expectancy μ is

$$\mathrm{P}(\overline{y} - \xi\frac{s}{\sqrt{n}} \leq \mu \leq \overline{y} + \xi\frac{s}{\sqrt{n}}) = p, \tag{2.10}$$

where the term $s/\sqrt{n}$ is the standard error of the sample, divided by the square root of the number of observations, i.e. the 'standard error of the mean'. The coefficient ξ depends on the selected confidence level p. The ξ is selected so, that the probability in standard normal pdf $\phi(\cdot)$ from $-\xi$ to ξ is p, i.e.

$$\int_{-\xi}^{\xi} \phi(x)dx = p. \tag{2.11}$$

For 95 % CI (i.e. $p = 0.95$) this value is 1.96, and similarly 2.58 for 99 % CI. To be exact, the Eq. (2.10) with ξ from normal distribution is only the asymptotic result. If the probability model actually is normal distribution, the ξ-values should be taken from the Student's t-distribution with $n - 1$ degrees of freedom. The difference is not large, in practice it is something to be taken into account if sample size is, say, less than 10. Example of normal and t-distributions are shown in Fig. 2.5.

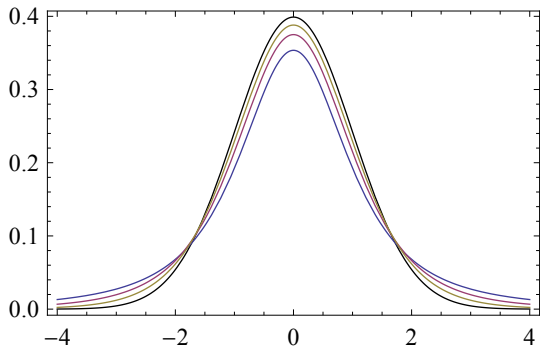

Figure 2.5: Standard normal distribution (black) and t-distribution with 2 (blue), 4 (red), and 9 (yellow) degrees of freedom.

2.2.2 Tests

With statistical tests we can check the likelihood of our hypothesis against the observed data, and make conclusions that are based on quantitative results. For tests we need suitably constructed test statistics $t(\boldsymbol{y})$ and a hypothesis, the so-called null

hypothesis H_0 (*nollahypoteesi*). The null hypothesis needs to define the probability model for test statistics, i.e. we must be know how $T|H_0$ is distributed.

If the data shows that our null hypothesis is very unlikely to be true, then we conclude that the alternative hypothesis H_1 (*vastahypoteesi*) seems more plausible. While the null hypothesis defines either one point in the parameter space, or at least some (small) set of parameters, the alternative hypothesis is its complement and does not define single value for the parameter, rather a single value that the parameter is not. For example, one could test with the mean from normally distributed data if (H_0) the $\mu = c$ or, (H_1) the $\mu \neq c$.

p-value of a test

The principle of statistical tests lies in the distribution of $T|H_0$ and in the likelihood of observed t. As said, we must know the pdf of $T|H_0$, i.e. $\mathrm{f}_{T|H_0}(t)$. With that knowledge we can calculate the probability of observing *as extreme value* of T as we have, or *even more extreme*, on the condition that H_0 is true. We return to the question of 'even more extreme' in the next section, but for now we just formulate that

$$\begin{aligned}\mathrm{P}(T \text{ more extreme as } t|H_0) &= \int_{t \text{ more extreme}} \mathrm{f}_{T|H_0}(x)dx \\ &= 1 - \int_{t \text{ less extreme}} \mathrm{f}_{T|H_0}(x)dx = p. \quad (2.12)\end{aligned}$$

Now, the philosophy is that if it is not that unlikely to observe such values of the statistic t if H_0 is true, we should not reject it. We do not say that H_0 is proven, but that there is no evidence that it should be rejected. If the p-value is very small it is quite unlikely to observe such value of t if H_0 is true. In that case we have two possibilities — either H_0 is not true, or a very unlikely event has happened. When the p-value is small enough, we tend to rule out the very unlikely event and say that H_0 is rejected and H_1 is accepted with certain p-value. See Fig. 2.6 for an example of test statistics where $T|H_0$ obeys χ^2-distribution and the corresponding p-value.

A certain conservative attitude is adopted with testing, and typical p-values where the H_0 is rejected are 0.10, 0.05 and 0.01. In times before computers it was common that just these three p-values were used, because tabulated values were looked up from tables containing these three cases. Nowadays one can as easily compute the exact p-value for the test and report that.

Limitations of statistical tests

With statistical tests one needs to understand their capabilities and limitations. Tests are quite good to quantify observed facts when there is moderate amount of data in hand. With just a few observations the uncertainty is usually so large,

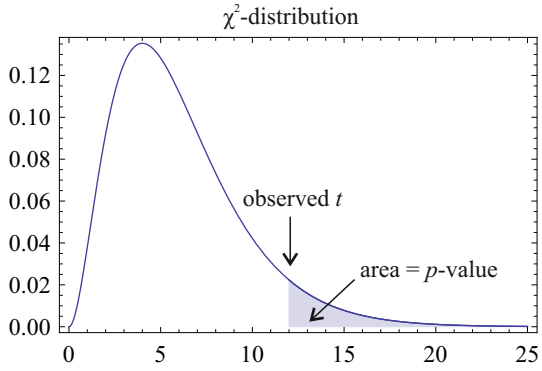

Figure 2.6: χ^2-distribution, observed test statistics t and the area corresponding to p-value of the one-tailed test.

that it is very hard to reject H_0. With large amount of data the problem is the opposite — it is quite easy to reject H_0. This is because the test usually states that there is evidence of deviation from H_0. What the test does not quantify that well is how large the deviation from H_0 is, and especially, does it have any practical consequences. For example, if one tests the correlation between two variables, H_0 is that there is no correlation, i.e. $\rho = 0$. With almost any kind of data, the parameter ρ probably deviates slightly from zero. When the number of observations increase, the test becomes stronger and picks up smaller and smaller differences from zero. Therefore, with large data it is easy to conclude that the correlation is not zero, and thus there is correlation, but the amount of correlation can be very small and not significant within the physical/real-world context behind the data. That said, statistical tests are very useful with moderate number of observations and with moderate deviations from H_0 when it is difficult to see without statistics if the deviation is 'unusual' or not.

Rejection areas

We need to define what we mean in Eq. (2.12) by areas where t is 'even more extreme'. That depends on the distribution of the test statistics, and on the alternative hypothesis. First, if the test statistics can have both negative and positive values, the distribution must be symmetric around zero. This is the case, for example, if the test statistics has normal or t-distribution under H_0. If we cannot say beforehand if it is impossible to have smaller (larger) values of t than assumed in H_0, our alternative hypothesis must be two-tailed (*kaksisuuntainen*), i.e. $H_0 : \theta = c$, $H_1 : \theta \neq c$. In this case (symmetric distribution, two-tailed H_1), the rejection area for test is such

that

$$\mathrm{P}(T \geq \mathrm{abs}(t)|H_0) = 2\int_{\mathrm{abs}(t)}^{\infty} \mathrm{f}_{T|H_0}(x)dx = 2\int_{-\infty}^{-\mathrm{abs}(t)} \mathrm{f}_{T|H_0}(x)dx$$
$$= 1 - \int_{-\mathrm{abs}(t)}^{\mathrm{abs}(t)} \mathrm{f}_{T|H_0}(x)dx = p. \quad (2.13)$$

If we have some a priori knowledge so that we can rule out, for example, positive values of t, we have one-tailed (*yksisuuntainen*) alternative hypothesis $H_1 : \theta < c$ and the rejection area is

$$\mathrm{P}(T \leq t|H_0) = \int_{-\infty}^{t} \mathrm{f}_{T|H_0}(x)dx = 1 - \int_{t}^{\infty} \mathrm{f}_{T|H_0}(x)dx = p, \quad (2.14)$$

and in similar manner for alternative hypothesis $H_1 : \theta > c$ but with integration limits changed accordingly.

The test statistics might have distribution that is only valid for positive values, for example χ^2- or F-distribution. These distributions are not symmetric, and we have to choose carefully the rejection area. If our statistics is close to zero and we have one-tailed H_1, the test is defined as

$$\mathrm{P}(T \leq t|H_0) = \int_{0}^{t} \mathrm{f}_{T|H_0}(x)dx = 1 - \int_{t}^{\infty} \mathrm{f}_{T|H_0}(x)dx = p. \quad (2.15)$$

With observed test statistics 'large' and with one-tailed H_1, the test is

$$\mathrm{P}(T \geq t|H_0) = \int_{t}^{\infty} \mathrm{f}_{T|H_0}(x)dx = 1 - \int_{0}^{t} \mathrm{f}_{T|H_0}(x)dx = p. \quad (2.16)$$

If we cannot rule out beforehand the small or large values of t, we must choose two-tailed test. Then, as we observe t to be either (i) close to zero or (ii) large, we choose (i) Eq. (2.15) or (ii) Eq. (2.16) and multiply the p-value in the correct equation by two to get the two-tailed p-value.

Mean tests

To list some tests, let us first consider the mean test, i.e. test for the expected value. The data is $\boldsymbol{y}$, and the statistics of interest is the mean value $\overline{y}$. The null hypothesis if of form $\mu = \mu_0$. For practical reasons we rather use the test statistics

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}, \quad (2.17)$$

where s is the sample standard deviation. From Eq. (2.7) we know that the asymptotic distribution of $T|H_0$ is standard normal distribution. We can formally say that

$$H_0 : \mu = \mu_0 \implies T \overset{\text{approx.}}{\sim} \mathcal{N}(0, 1). \quad (2.18)$$

Actually, if we know that the distribution of data is normal, we can replace the asymptotic distribution with the exact one: $T \sim t_{n-1}$, i.e. the Student's t-distribution with $n-1$ degrees of freedom.

In Fig. 2.7 there are 10 random numbers that are sampled from $\mathcal{N}(0.1, 1)$ distribution. Our H_0 is that $\mu = \mu_0 = 0$, and that distribution is drawn in subfigure a) together with the data. The test statistics t is calculated ($t \approx 1.34$) and the areas $]-\infty, -t]$ and $[t, \infty]$ drawn in subfigure b) together with the distribution of $T|H_0$, the t-distribution with 9 degrees of freedom. The p-value, i.e. the colored area in subfig b), is 0.212. Therefore, we do not have enough evidence against $H_0 : \mu = 0$ and we cannot reject that possibility, although we actually know that the data comes from distribution with $\mu = 0.1$.

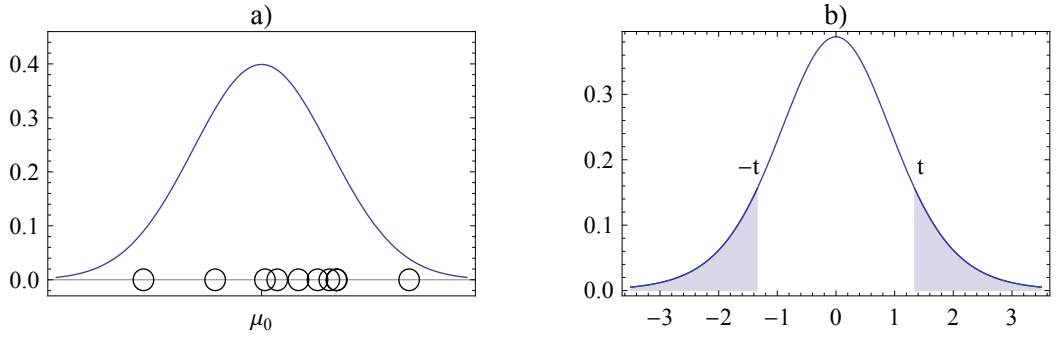

Figure 2.7: Data and H_0-distribution in left (a), observed value of t and the distribution according to H_0 in right (b).

Similar mean test can be also constructed for two samples and the difference of their mean values. One has to assume that the samples have the same distributions (expect for the location parameter) and that their variances σ_1^2 and σ_2^2, while unknown, are equal. In that case,

$$H_0 : \mu_1 - \mu_2 = d_0 \implies T = \frac{(\overline{y}_1 - \overline{y}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}, \tag{2.19}$$

where pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{2.20}$$

In what follows we will shortly describe some tests, but the list is not by far complete. You will notice that almost all the distributions for test statistics are either Student's t-distribution, χ^2-distribution or F-distribution. This is simply because all these distributions are derived from normal distribution — t-distribution from the ratio of normal variable and its standard deviation, χ^2-distribution from sum of squared normal variables, and F-distribution from ratio of normal variables.

Variance tests

For variance of one normal distributed sample the test is

$$H_0 : \sigma^2 = \sigma_0^2 \implies T = (n-1)\frac{s^2}{\sigma_0^2} \sim \chi^2_{n-1}, \tag{2.21}$$

and rejection areas for two-tailed test can be computed using Eq. (2.15) or (2.16) and adjusting p-value to $2p$.

For two normal distributed samples the test for equal variance is

$$H_0 : \sigma_1^2 = \sigma_2^2 \implies T = \frac{s_1^2}{s_2^2} \sim \mathcal{F}_{n_1-1,n_2-1}, \tag{2.22}$$

and the alternative hypothesis will define the rejection area to either Eq. (2.15) or (2.16).

Correlation test

The linear correlation, i.e. the value of correlation coefficient ρ and its sample statistics $r = \mathrm{cor}(\boldsymbol{x}, \boldsymbol{y})$, can be tested against being zero. The test is

$$H_0 : \rho = 0 \implies T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}, \tag{2.23}$$

and rejection area is defined by Eq. (2.13) for two-tailed, and by Eq. (2.14) for one-tailed test.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov (K-S) test is our first non-parametric test. It can be used to test if the observed distribution differs from theoretical distribution, and the test is valid for all (continuous) distributions. The test is based on the empirical CDF and the theoretical CDF. The test statistics t is defined as $t = \sqrt{n}D$, where D is the maximum difference between the two CDF's, see Fig. 2.8.

The K-S test is always one-tailed, and the test statistics have Kolmogorov distribution if H_0 that the sample comes from the theoretical distribution is true, rejection area is defined as in Eq. (2.16).

There is a similar version for K-S test between two empirical distributions, check e.g. Wikipedia for the details.

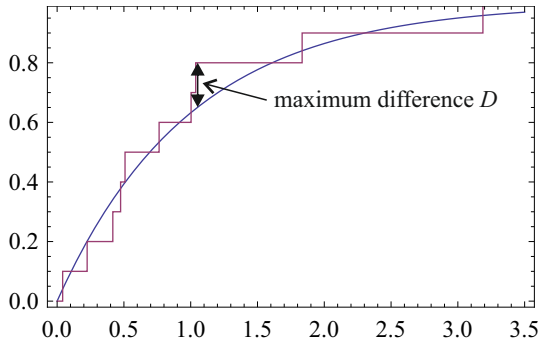

Figure 2.8: Empirical and theoretical cumulative distribution functions and the Kolmogorov-Smirnov difference D.

Goodness-of-fit test

Goodness-of-fit test can be used for discrete variables. It is formulated as

$$H_0 : \text{Empirical distribution obeys the theoretical one} \implies$$
$$T = n \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \sim \chi^2_{n-1-m}, \quad (2.24)$$

and large values speak against H_0 as in Eq. (2.16). The terms o_i are the observed probabilities (proportions) of class/value/category i in the sample, and terms e_i are the expected probabilities if H_0 is true. The variable m in the degrees of freedom for the χ^2-distribution is the number of unknown parameter values estimated from the data for the theoretical distribution. For example, if we want to test if the observed proportions come from uniform (discrete) distribution, we do not need to estimate any parameter values from the data, and $m = 0$.

Independence test

The same test statistics as above can be used to test the independence between two-dimensional categorical variable, i.e. proportions in two-way contingency tables (cross tabulations, *ristiintaulukko*). Every observation has two properties, A and B, and it can be associated to one cell in the contingency table. The proportions of the associations are counted, resulting the following table

$A \setminus B$	1	...	k	Σ
1	o_{11}	...	o_{1k}	A_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
m	o_{m1}	...	o_{mk}	A_m
Σ	B_1	...	B_k	1

The expected proportions, if the two properties A and B are independent, can be estimated from the product of the marginal proportions: $e_{ij} = A_i B_j$. The test statistics is computed over all the rows and columns, and

$$H_0 : A \perp\!\!\!\perp B \implies T = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(m-1)(k-1)}, \tag{2.25}$$

and large values speak against H_0 as in Eq. (2.16).