8.2. Hypothesis Testing (Hypoteesien testaus)

8.2.1. Main Steps in Testing (Test. pääpiirtet)

In hypothesis testing one attempts to decide based upon data, whether the parameters of a population agree with a prespecified assumption, called hypothesis (hypoteesi).

The null hypothesis (nollahyoteesi) H_0 is an assertion about the value of a population parameter based upon prior knowledge, which we hold true unless we have sufficient statistical evidence to conclude otherwise. The alternative hypothesis (vastahypotheesi) H_1 is the *negation* of the null hypothesis. This is usually the research hypothesis, which the researcher wants to show holds true.

Example.

A machine fills paint buckets with 1000g each. The standard deviation of the fill is known to be 10g. In order to assess whether the machine works correctly one would set up the hypothesis:

$$H_0 : \mu = 1000g \text{ and } H_1 : \mu \neq 1000g.$$

H_1 above is called two-tailed or two-sided (kaksisuuntainen), because it may hold true for μ both smaller and larger than 1000g.

One-tailed or one-sided (yksisuuntainen) alternative hypotheses would have been either

$$H_1 : \mu > 1000g$$

or

$$H_1 : \mu < 1000g.$$

Because the choice between H_0 and H_1 depends upon the information given in the sample, the decision is based upon so called test statistics (testisuure), which are usually based upon the estimators of the parameters to be tested. This requires knowing the sampling distribution of the test statistics (testisuuren otantajakauma).

Example. (paint buckets continued)

A natural candidate for the test statistic is the sample mean. If the weight of the paint buckets is normally distributed, then

$$\bar{X} \sim N\left(1000g, \frac{(10g)^2}{n}\right) \text{ if } H_0 \text{ holds true.}$$

Next one determines the values of the test statistics, which shall then be regarded as deviating too much from the presumed value, for the sample to be taken from a population with parameters according to H_0. This set of values is called the critical region (hylkäysalue, kriittinen alue) C. The critical region of a test depends upon the form of the alternative hypothesis, the sampling distribution of the test statistics, and the so called significance level (merkitsevyys-/ riskitaso) α, which denotes the probability of rejecting H_0 even though in fact it holds true. Commonly used significance levels are $\alpha = 0.05$ (denoted by *), $\alpha = 0.01$ (denoted by **), and $\alpha = 0.001$ (denoted by ***).

Example. (paint buckets continued)
In the case of the two-sided alternative $H_1 : \mu \neq 1000g$, a critical region at significance level α is

$$C_{\bar{X}} = \{\bar{X} || \bar{X} - 1000g| > k\}$$

where

$$P_{H_0}(|\bar{X} - 1000g| > k) = \alpha.$$

Example. (paint buckets continued)

Consider taking a sample of 25 buckets and a significance level of $\alpha = 0.05$. In searching for the critical region $C_{\bar{X}}$ recall from our discussion of confidence intervals that the margin of error for $k = |\bar{X} - \mu|$ in $100(1-\alpha)\%$ of all samples is

$$k = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{10}{\sqrt{25}} = 3.92.$$

The critical region for the two-sided test is therefore:

$$C_{\bar{X}} = \{\bar{X} || \bar{X} - 1000| > 3.92\}.$$

An easier way to perform the very same test is considering the standardized test statistics

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ if } H_0 \text{ holds true.}$$

In terms of this new test statistic the critical region becomes:

$$C_Z = \{|Z| > z_{0.05/2}\} = \{|Z| > 1.96\},$$

such that H_0 will be rejected if $|Z| > 1.96$.

Example. (paint buckets continued)

If the sample mean is $\bar{X} = 1006g$, then

$$Z = \frac{1006 - 1000}{10/\sqrt{25}} = 3 > 1.96$$

(also: $|\bar{X} - 1000| = 6 > 3.92$)

$\Rightarrow$ H_0 is rejected at significance level 0.05.

$\Rightarrow$ The machine appears to malfunction.

Hypothesis Testing in a nutshell:

1. Set up the null hypothesis H_0.
2. Set up the alternative hypothesis H_1.
3. Determine the test statistic and its distribution under H_0.
4. Decide upon a significance level α and determine the corresponding critical region.
5. Take a sample, compute the test statistic and compare its value with the critical region.
6. Make the test decision (acceptance/ rejection) and interpret the result.

Type I and Type II Errors
(Testauksen liityvät virhemahdollisuudet)

After the test decision is made, we cannot be sure it was correct, because it was only based on observations of the sample rather than the full population.

Type I error (α): Rejection error (hylkäämisvirhe), rejecting the correct H_0 hypotesis.

Type II error (β): Acceptance error (hyväksymisvirhe), accepting the false H_0 hypotesis.

		Reality	
		H_0	H_1
Decision	H_0	Correct decision	Type II error (β)
	H_1	Type I Error (α)	Correct decision

Acceptance and rejection of H_0 form a partition of the sample space of the decisions we may choose to take, such that the probability of correctly rejecting H_0 when H_0 is false is $\pi := 1 - \beta$. This is the so called power of the test (testin voimakkuus).

p-values

Statistical programs usually report so called p-values (p-arvot). They express the probability to obtain the observed or an even more extreme value of the observed statistic if H_0 holds true. These probabilities are called observed significance levels (havaitu merkitsevyystaso) and denote the highest α, which do not yet lead to a rejection of H_0. That is, they may be compared with the chosen significance level as follows:

Accept H_0 at significance level α if $p \geq \alpha$.
Reject H_0 at significance level α if $p < \alpha$.

Example. (paint buckets continued)
The value of the test-statistic was 3, the p-value of which is obtained by using the standard normal cumulative distribution function $\Phi(z)$=NORMSDIST(z) as follows:

$$p = P(|Z| > 3) = P(Z > 3 \text{ or } Z < -3) = 2P(Z > 3)$$
$$= 2[1 - P(Z \leq 3)] = 2[1 - \underbrace{0.9987}_{\Phi(3)}] = 0.0026,$$

which implies that applying any significance level lower than 0.26%, that is $\alpha > 0.0026$, will lead to a rejection of H_0.

8.2.2. Tests for the Mean (Keskiarvotestejä)

1. One sample (Yksi ostos)

In the case of one sample one investigates whether the sample mean supports the value of the population mean as stated in the null hypothesis. The test assumes, that the sample observations are normally distributed ($X_1, \ldots, X_n \sim NID(\mu, \sigma^2)$) or n is large. The null hypothesis is then of the form

$$H_0 : \ \mu = \mu_0.$$

a) The test statistic for known population variance σ^2 is:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ if } H_0 \text{ holds true.}$$

The critical regions C are of the form:

$$\begin{aligned} H_1 : \mu > \mu_0 \ &\Rightarrow \ C = \{Z | Z > z_\alpha\} \\ H_1 : \mu < \mu_0 \ &\Rightarrow \ C = \{Z | Z < -z_\alpha\} \\ H_1 : \mu \neq \mu_0 \ &\Rightarrow \ C = \{Z | |Z| > z_{\alpha/2}\} \end{aligned}$$

Example: See paint buckets above.

b) The test statistic for unknown population variance σ^2 is:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ if } H_0 \text{ holds true, where}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}\right).$$

The fractiles of the t-distribution are used to set up critical regions C as follwows:

$$\begin{aligned} H_1 : \mu > \mu_0 &\Rightarrow C = \{T | T > t_\alpha(n-1)\} \\ H_1 : \mu < \mu_0 &\Rightarrow C = \{T | T < -t_\alpha(n-1)\} \\ H_1 : \mu \neq \mu_0 &\Rightarrow C = \{T | |T| > t_{\alpha/2}(n-1)\} \end{aligned}$$

Example. (paint buckets continued)

Let us assume that we do not know the variance of the paint buckets weight and that our sample of 25 buckets yields $S = 11.2$. A 5% critical region for the two-sided test is then

$$C = \{T | |T| > t_{0.05/2}(24)\} = \{T | |T| > \text{TINV}(0.05; 24) = 2.064\}.$$

$$\text{Now} \quad t = \frac{1006 - 1000}{11.2/\sqrt{25}} \approx 2.68 \in C.$$

$\Rightarrow$ H_0 is rejected, that is we decide $\mu \neq 1000g$.

2. Two independent samples
(Kaksi riippumatonta otosta)

In the case of two independent samples we compare two sample means from different populations and attempt to clarify whether both population means are identical. The test assumes $X_{11}, \ldots, X_{1n} \sim N(\mu_1, \sigma_1^2)$ and $X_{21}, \ldots, X_{2n} \sim N(\mu_2, \sigma_2^2)$, and that both samples are taken independently. The null hypothesis is now of the form

$$H_0 : \ \mu_1 = \mu_2.$$

a) Both σ_1^2 and σ_2^2 are known, then:

$$Z = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ if } H_0 \text{ holds true.}$$

Critical regions are based upon the fractiles of the standard normal distribution:

$$\begin{aligned} H_1 : \mu_1 > \mu_2 &\ \Rightarrow\ C = \{Z | Z > z_\alpha\} \\ H_1 : \mu_1 < \mu_2 &\ \Rightarrow\ C = \{Z | Z < -z_\alpha\} \\ H_1 : \mu_1 \neq \mu_2 &\ \Rightarrow\ C = \{Z | |Z| > z_{\alpha/2}\} \end{aligned}$$

b) σ_1^2 and σ_2^2 unknown, but of equal size, that is we may assume $\sigma_1^2 = \sigma_2^2 =: \sigma^2$. We estimate σ_1^2 and σ_2^2 using

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2}{n_1 - 1} \text{ and } S_2^2 = \frac{\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2}{n_2 - 1},$$

and the common population variance σ^2

$$\text{with} \quad S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

The test statistic is now

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \text{ under } H_0.$$

Critical regions are based upon the fractiles of the t-distribution:

$$H_1 : \mu_1 > \mu_2 \;\Rightarrow\; C = \{T | T > t_\alpha(n_1 + n_2 - 2)\}$$
$$H_1 : \mu_1 < \mu_2 \;\Rightarrow\; C = \{T | T < -t_\alpha(n_1 + n_2 - 2)\}$$
$$H_1 : \mu_1 \neq \mu_2 \;\Rightarrow\; C = \{T | |T| > t_{\alpha/2}(n_1 + n_2 - 2)\}$$

c) If σ_1^2 and σ_2^2 are both unknown and possibly of unequal size, the test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ approx. } \sim t(df) \text{ under } H_0,$$

where the degrees of freedom (vapausasteet) df may be calculated from

$$df = \frac{(S_1^2/n1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)}$$

rounded down to the nearest integer.

Note. The statistical formula collection writes for df:

$$\frac{1}{df} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}, \text{ where } c = \frac{S_1^2/n_1}{(S_1^2/n_1) + (S_2^2/n_2)}.$$

This gives the same result, because:

$$\begin{aligned} 1 - c &= \frac{(S_1^2/n_1) + (S_2^2/n_2)}{(S_1^2/n_1) + (S_2^2/n_2)} - \frac{S_1^2/n_1}{(S_1^2/n_1) + (S_2^2/n_2)} \\ &= \frac{S_2^2/n_2}{(S_1^2/n_1) + (S_2^2/n_2)}, \text{ such that} \\ \frac{1}{df} &= \frac{(S_1^2/n_1)^2/(n_1 - 1)}{(S_1^2/n1 + S_2^2/n_2)^2} + \frac{(S_2^2/n_2)^2/(n_2 - 1)}{(S_1^2/n1 + S_2^2/n_2)^2} \\ &= \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}. \end{aligned}$$

Sampling distribution of the sample variance: The Chi-square (χ^2) distribution

Let $Z_1, \ldots, Z_n$ be independent $N(0,1)$ random variables. Then

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n),$$

where $\chi^2(n)$ denotes the χ^2-distribution with degrees of freedom $df = n$.

Note.
We will shortly use that for $X_i \sim \text{NID}(\mu, \sigma^2)$:

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1),$$

where $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ denotes the sample variance of $X_1, \ldots, X_n$.

The tail fractiles χ^2_α of the χ^2-distribution for $P(\chi^2 > \chi^2_\alpha(df)) = \alpha$ are tabulated (see next page) and implemented in Excel as:

$$\chi^2_\alpha(df) = \text{CHIINV}(\alpha; df).$$

Table. Tail fractiles χ^2_α of the χ^2-distribution: $P(\chi^2 > \chi^2_\alpha(df)) = \alpha$.

					χ_α^2(df)					
df / α	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.001
1	0.000039	0.000157	0.000982	0.003932	0.0158	2.706	3.841	5.024	6.635	10.827
2	0.0100	0.0201	0.0506	0.103	0.211	4.605	5.991	7.378	9.210	13.815
3	0.0717	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	16.266
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	18.466
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.086	20.515
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	22.457
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	24.321
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	26.124
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	27.877
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	29.588
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	31.264
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	32.909
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	34.527
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	36.124
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	37.698
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	39.252
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	40.791
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	42.312
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	43.819
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	45.314
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	46.796
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	48.268
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	49.728
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	51.179
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	52.619
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	54.051
27	11.808	12.878	14.573	16.151	18.114	36.741	40.113	43.195	46.963	55.475
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	56.892
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	58.301
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	59.702
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	73.403
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	86.660
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	99.608
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	112.317
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	124.839
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	137.208
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	149.449

Example:

$\chi^2 \sim \chi^2(16) \Rightarrow \ 0.05 = P(\chi^2 \geq 26.296).$

The sampling distribution for the ratio of sample variances: Fisher's F-distribution

Let $U \sim \chi^2_m$ and $W \sim \chi^2_n$ be independent, then the ratio

$$F = \frac{U/m}{W/n} \sim F(m, n)$$

where $F(m, n)$ denotes the F-distribution with degrees of freedom *df* $= m$ and n (Fisherin F-jakauma vapausastein m ja n).

Note.
This implies by the remark on χ^2-distributions, that the ratio of the sample variances S^2_U and S^2_W is in independent samples distributed as

$$\frac{S^2_U/\sigma^2_U}{S^2_W/\sigma^2_W} \sim F(n_U - 1, n_W - 1),$$

where n_U and n_W denote the sample size of U and W, and σ^2_U, σ^2_W the population variances.

Tail fractiles $F_\alpha(m, n)$ for the probabilities $P[F > F_\alpha(m, n)] = \alpha$ of the F-distribution are tabulated for various combinations of m, n and α (see next page) and implemented in excel as FINV($\alpha; m; n$).

Example.

$F \sim F(10, 20) \Rightarrow \ 0.10 = P(F \geq 1.94).$

Table. Tail fractiles $F_{0.10}(m, n)$ of the F-distribution for $P[F > F_{0.10}(m, n)] = 0.10$.

$F_{0.1}(m,n)$

n \ m	1	2	3	4	5	10	20	40	120	∞
1	39.86	49.50	53.59	55.83	57.24	60.19	61.74	62.53	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.39	9.44	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.23	5.18	5.16	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	3.92	3.84	3.80	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.30	3.21	3.16	3.12	3.10
10	3.29	2.92	2.73	2.61	2.52	2.32	2.20	2.13	2.08	2.06
20	2.97	2.59	2.38	2.25	2.16	1.94	1.79	1.71	1.64	1.61
40	2.84	2.44	2.23	2.09	2.00	1.76	1.61	1.51	1.42	1.38
120	2.75	2.35	2.13	1.99	1.90	1.65	1.48	1.37	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.60	1.42	1.30	1.17	1.00

Table. Tail fractiles $F_{0.05}(m, n)$ of the F-distribution for $P[F > F_{0.05}(m, n)] = 0.05$.

$F_{0.05}(m,n)$

n \ m	1	2	3	4	5	10	20	40	120	∞
1	161.4	199.5	215.7	224.6	230.2	241.9	248.0	251.1	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.40	19.45	19.47	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.79	8.66	8.59	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	5.96	5.80	5.72	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.74	4.56	4.46	4.40	4.36
10	4.96	4.10	3.71	3.48	3.33	2.98	2.77	2.66	2.58	2.54
20	4.35	3.49	3.10	2.87	2.71	2.35	2.12	1.99	1.90	1.84
40	4.08	3.23	2.84	2.61	2.45	2.08	1.84	1.69	1.58	1.51
120	3.92	3.07	2.68	2.45	2.29	1.91	1.66	1.50	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	1.83	1.57	1.39	1.22	1.00

The F test for equality of spread
(Varianssien σ_1^2 ja σ_2^2 yhtä suuruuden testaus)

In order to decide whether test procedure 2b) or 2c) is appropriate for comparing the means of two independent samples one applies the so called F test with the hypothesis

$$H_0: \ \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1: \ \sigma_1^2 \neq \sigma_2^2.$$

The test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \text{ if } H_0 \text{ holds true,}$$

so the critical region of the two sided test is:
$C = \{F|F < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)\} \cup \{F|F > F_{\alpha/2}(n_1 - 1, n_2 - 1)\}.$
Since tables usually provide only a few critical values for the upper tail, it is common practice to label the samples such that $S_1^2 \geq S_2^2$. That way we are sure that $F \geq 1$ and the critical region of the two sided test becomes:

$$C = \{F|F > F_{\alpha/2}(n_1 - 1, n_2 - 1)\} \text{ for } S_1^2 \geq S_2^2.$$

Note. Unlike the t procedures for means, the F test and other procedures for standard deviations are extremely sensitive to nonnormal distributions. This lack of robustness does not improve in large samples.

Example.

Does calcium reduce blood pressure?

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

Let's first test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$ at α=20%. The test statistic is:

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2} = \frac{8.743^2}{5.901^2} \approx 2.195$$

with critical region:

$$C = \{F | F > F_{\alpha/2}(n_1 - 1, n_2 - 1)\}.$$

Now by looking up from a table, or calling FINV(0.1;9;10) in excel, we find that

$$F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.1}(9, 10) = 2.347,$$

which leads us to accept equal variances even at $\alpha = 20\%$, since $2.195 < 2.347$.

Alternatively, we might have calculated the p-value of the F-test as $p = 0.237$ by calling 2*FDIST(2.195;9;10) in excel, which leads to the same conclusion, since $0.237 > 0.2$.

Example: (continued).

So we assume equal variances, which means that in testing for equality of means we should apply procedure 2b) with test statistic

$$t = \frac{\bar{x_1} - \bar{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{where}$$

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$= \frac{9 \cdot 8.743^2 + 10 \cdot 5.901^2}{10 + 11 - 2} \approx 54.536$$

$$\text{such that} \quad s = \sqrt{54.536} = 7.385.$$

Therefore:

$$t = \frac{5.000 - (-0.273)}{7.385\sqrt{\frac{1}{10} + \frac{1}{11}}} = \frac{5.273}{3.227} = 1.634.$$

Example: (continued).
Now we are interested in demonstrating that the treatment with calcium does actually have a positive impact upon the reduction of blood pressure as measured by X. That is we choose to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$ with critical region

$$C = \{t | t > t_\alpha(n_1 + n_2 - 2) = t_\alpha(19)\}.$$

Now by looking up from a table or calling TINV(2*α; 19) in excel we find that

$$(t_{0.1}(19) = 1.328) < (t = 1.634) < (t_{0.05}(19) = 1.729),$$

which means that we may reject H_0 at 10%, but not at a significance level of $\alpha = 5\%$.

Alternatively, we might have calculated the p-value of the t-test as $p = 0.059$ by calling TDIST(1.634;19;1) in excel, (1 is for one-sided test, TDIST(1.634;19;2) would have given the p-value for the two-sided test). So the observed signficance level (the highest α, at which we do not declare t as significant) is 5.9%, confirming our conclusion above.

Example: (continued).
If the F-test had rejected the null hypothesis of equal variances, we would have calculated the t-statistics according to c) as

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.000 - (-0.273)}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.604$$

with degrees of freedom df calculated from

$$df = \left\lfloor \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)} \right\rfloor .$$

Now

$$s_1^2/n_1 = 8.743^2/10 = 7.609,$$
$$s_2^2/n_2 = 5.901^2/11 = 3.166,$$

such that

$$df = \left\lfloor \frac{(7.609 + 3.166)^2}{7.609^2/9 + 3.166^2/10} \right\rfloor = \lfloor 15.6 \rfloor = 15$$

with critical region for the one-sided test

$$C = \{t | t > t_\alpha(15)\}.$$

Now again by looking up from a table or calling TINV(2*α; 15) in excel we find that

$$(t_{0.1}(15) = 1.341) < (t = 1.604) < (t_{0.05}(15) = 1.753),$$

which is again significant at $\alpha = 10\%$, but not enough to reject H_0 at $\alpha = 5\%$.

3. Paired observation t test
(Kaksi toisistaan riippuvaa otosta)

We call measuring a statistical variable twice on the same subject under different circumstances paired observations or matched samples (kaltaistetut otokset). For example, suppose two different flavours are rated by the same people. Then the rating of each flavour is a sample, but they are no longer independent. Assuming that the difference between the flavours is normally distributed, that is $D := X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$, we may apply the one sample t-test upon D. The null hypothesis is

$$H_0 : \mu_D = 0 \quad \Leftrightarrow \quad \mu_1 = \mu_2$$

with test statistic:

$$T = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t(n-1) \text{ if } H_0 \text{ holds true,}$$

where $\bar{D}$ is the sample mean of the differences, and s_D is the standard deviation of the differences. The critical regions are the same as for the one sample t-test with μ replaced by μ_D, and μ_0 replaced by 0.

8.2.3. Testing Population Proportions (Prosenttilukutestejä)

1. One sample (Yksi ostos)

In the case of one sample one investigates whether the sample proportion $\hat{P}$ of elements of type A supports the value of the population proportion Π as stated in the null hypothesis. The null hypothesis is of the form:

$$H_0 : \ \Pi = \Pi_0.$$

The test statistic is:

$$Z = \frac{\hat{P} - \Pi_0}{\sqrt{\frac{\Pi_0(100-\Pi_0)}{n}}} \sim N(0,1) \quad \text{if } H_0 \text{ holds true.}$$

The critical regions C are of the form:

$$\begin{aligned} H_1 : \Pi > \Pi_0 \ &\Rightarrow \ C = \{Z | Z > z_\alpha\} \\ H_1 : \Pi < \Pi_0 \ &\Rightarrow \ C = \{Z | Z < -z_\alpha\} \\ H_1 : \Pi \neq \Pi_0 \ &\Rightarrow \ C = \{Z | |Z| > z_{\alpha/2}\} \end{aligned}$$

Example.

The support for a political party in the previous elections was 18.4%. A random sample of $n = 1493$ eligible voters yields a support of $\hat{P} = 22.7\%$. Did the support for the party increase?

We test $H_0 : \Pi = 18.4$ against $H_1 : \Pi > 18.4$.

Choosing a significance level of $\alpha = 1\%$ yields:

$$z_{0.01} = \Phi^{-1}(0.99) = \text{NORMSINV}(0.99) \approx 2.33$$

$$\text{such that} \quad C = \{Z | Z > 2.33\}.$$

Now:

$$Z = \frac{22.7 - 18.4}{\sqrt{\frac{18.4(100-18.4)}{1493}}} \approx 4.3 \in C.$$

H_0 is therefore rejected at a significance level of 1%, and we regard the support for the party as having increased.

2. Two samples (Kaksi otosta)

In the case of two samples we compare the sample proportions of type A elements $\hat{P}_1$ and $\hat{P}_2$ and attempt to clarify, whether the population proportions Π_1 and Π_2 are of equal size. The null hypothesis is of the form:

$$H_0 : \ \Pi_1 = \Pi_2.$$

The test statistic is now

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(100 - \hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \text{ under } H_0,$$

where $\hat{P} = \dfrac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1 + n_2}$ denotes the combined type A proportion of samples 1 and 2.

The critical regions are:

$$\begin{aligned} H_1 : \Pi_1 > \Pi_2 \ &\Rightarrow \ C = \{Z | Z > z_\alpha\} \\ H_1 : \Pi_1 < \Pi_2 \ &\Rightarrow \ C = \{Z | Z < -z_\alpha\} \\ H_1 : \Pi_1 \neq \Pi_2 \ &\Rightarrow \ C = \{Z | |Z| > z_{\alpha/2}\} \end{aligned}$$

Example.

80 out of 200 randomly selected female students and 47 out of 100 randomly selected male students smoke. Is there a difference in the fraction of smokers? We test

$$H_0: \ \Pi_F = \Pi_m \quad \text{against} \quad H_1: \ \Pi_F \neq \Pi_m.$$

Choosing a significance level of $\alpha = 5\%$ yields:

$$z_{0.025} = \Phi^{-1}(0.975) \approx 1.96$$

$$\text{such that} \quad C = \{Z \| Z| > 1.96\}.$$

Now:

$$\hat{P} = \frac{200 \cdot 40 + 100 \cdot 47}{200 + 100} = \frac{80 + 47}{300} \cdot 100 \approx 42.3$$

such that:

$$Z = \frac{40 - 47}{\sqrt{42.3 \cdot 57.7 \left(\frac{1}{200} + \frac{1}{100}\right)}} \approx -1.16 \notin C.$$

We accept therefore H_0 and regard the difference in smoking behaviour between female and male students as nonsignificant at 5% level.

8.2.4. Tests related to statistical dependence

1. χ^2 independence test

Recall from our discussion of contingency tables that we could use Pearson's χ^2 statistics

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}},$$

where f_{ij} and eij denoted the observed and expected frequencies of cell (G_i, E_j), in order to assess dependence between two statistical variables x and y. It turns out that this may be used as a test statistic for statistical independence as long as the following conditions are satisfied:

1. No more than 20% of the expected frequencies e_{ij} are smaller than 5.
2. All expected frequencies satisfy $e_{ij} > 1$.

The hypotheses of the χ^2 independence test are:

H_0: x and y are statistically independent
H_1: x and y are statistically dependent

The test statistic is

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2[(I-1)(J-1)] \text{ under } H_0,$$

where I denotes the number of rows and J the number of columns in the contingency table for x and y.

The critical region is:

$$C = \{\chi^2 | \chi^2 > \chi^2_\alpha[(I-1)(J-1)]\}.$$

Note. In the special case of two way tables (nelikentäjä), that is $I = J = 2$, the test statistic simplifies to

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{1\bullet}f_{2\bullet}f_{\bullet 1}f_{\bullet 2}} \sim \chi^2(1) \text{ under } H_0.$$

Example: (study progress continued.)
In the section about contingency tables we discussed dependence between working besides studying and study progress. The contingency table had $I = 3$ rows and $J = 2$ columns. The hypotheses to be tested are:

H_0 : Working and Study Progress are statistically independent,
H_1 : Working and Study Progress are statistically dependent.

Choosing a significance level of $\alpha = 5\%$, we obtain for the critical region:

$$C = \{\chi^2 | \chi^2 > \chi^2_{0.05}(2) = \text{CHIINV}(0.05;2) = 5.99\}.$$

Now we obtained earlier $\chi^2 = 14.2 \in C$, such that we reject H_0 and consider the variables as statistically dependent. Alternatively, we may calculate the p-value of the χ^2-test as $p = 0.0008$ by calling CHIDIST(14.2;2) in excel, which shows that we may reject H_0 even at much higher significance levels (e.g. $\alpha = 0.1\%$ would be feasable).

2. Correlation test

Pearson's correlation coefficient r may be used to estimate the population correlation ρ between two statistical variables x and y, if both are normally distributed. The null hypothesis to be tested is

$H_0: \ \rho = 0$ (x, y are linearly independent)

The test statistic is:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \text{ under } H_0$$

with critical regions:

$$H_1 : \rho > 0 \ \Rightarrow \ C = \{T | T > t_\alpha(n-2)\}$$
$$H_1 : \rho < 0 \ \Rightarrow \ C = \{T | T < -t_\alpha(n-2)\}$$
$$H_1 : \rho \neq 0 \ \Rightarrow \ C = \{T | |T| > t_{\alpha/2}(n-2)\}$$

Note:
When both r is small and n is large, the test statistic above may be approximated by

$Z = r\sqrt{n} \sim N(0,1)$ under H_0, with critical regions:

$$H_1 : \rho > 0 \ \Rightarrow \ C = \{Z | Z > z_\alpha\}$$
$$H_1 : \rho < 0 \ \Rightarrow \ C = \{Z | Z < -z_\alpha\}$$
$$H_1 : \rho \neq 0 \ \Rightarrow \ C = \{Z | |Z| > z_{\alpha/2}\}$$

Example: (softdrink campaign continued.)
In our earlier investigation of the correlation between the number of adverts seen and bottles of softdrinks bought we found a correlation coefficient of $r = 0.68$ based upon $n = 8$ subjects. In order to clarify, whether the advert has a positive impact upon sales we test

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho > 0.$$

Choosing $\alpha = 5\%$ yields the critical region:

$$C = \{T | T > t_\alpha(n-2) = t_{0.05}(6) = 1.943\}.$$

$$\text{Now:} \qquad t = \frac{0.68\sqrt{8-2}}{\sqrt{1-0.68^2}} = 2.27 \in C,$$

so we reject H_0 and assume positive linear dependence between adverts and sales.

Note. The t-statistic above may also applied to Spearmans rank correlation coefficient, provided that n is large. In that case we don't need to assume normally distributed variables. The null hypothesis is then:

$H_0 : \ \rho_S = 0$ (x, y are monotonically independent)

Testing Regression Coefficients

When fitting a regression line in a scatterplot, one often wants to know whether x does indeed have an impact upon y. This concerns testing the slope coefficient of the regression, because it tells the average impact of a change in x upon y. Now the sample regression parameter b_1 may be regarded as an estimator of the population regression parameter β_1. The common null hypothesis is:

$$H_0: \ \beta_1 = 0 \ (x, y \text{ are linearly independent})$$

When assuming normally distributed regression errors $e_i = y_i - \widehat{y}_i$, the test statistic is:

$$T = \frac{b_1}{s(b_1)} \sim t(n-2) \text{ under } H_0,$$

where $s(b_1)$ denotes the standard error of the slope coefficient b_1:

$$s(b_1) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2/(n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}},$$

which is generally provided with the computer output of any regression analysis software. Note that

$$T = \frac{b_1}{s(b_1)} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y}\sqrt{\frac{n-2}{1-r^2}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

the same as the test statistic for testing for a significant correlation coefficient r. That is, testing for a significant correlation and testing for a significant slope of the regresion line are equivalent.

Example:(Flat size and price continued.) We calculated earlier the variance of flat size as $s_x^2 = 660.933$ and the variance of flat price as $s_y^2 = 10\,504.267$. The correlation between flat size and price was found to be $r_{xy} = 0.9446$. Below is some regression output of the statistical software package SPSS:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	145,092	31,865		4,553	,002
	Size	3,766	,463	,945	8,138	,000

a. Dependent Variable: Price

The program reports the same regression coeffients as we found earlier, $b_0 \approx 145.1$ and $b_1 \approx 3.77$. We could have obtained the standard error of the slope coeffient $s(b_1) \approx 0.463$ ourselves by calculating:

$$s(b_1) = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{10\,504.267}{660.933}}\cdot\sqrt{\frac{1-0.9446^2}{10-2}} \approx 0.463.$$

The t-values are obtained by dividing the regression coefficients by their corresponding standard errors, and Sig. stands for observed significance levels, that is the p-values of the t-statistics, e.g. for the slope coefficient b_1: p=TDIST(8.138;8;2)≈0.00004. (in excel)

8.2.5. χ^2 Test for Goodness of Fit

Consider randomly sampled variables, which are classified into k categories, the counts of which we denote by $f_1, f_2, \ldots, f_k$. Now the goal is to clarify, whether the observed counts f_i agree with some expected counts $e_i (= np_i)$. Usually the idea is to clarify, whether the random variable X follows some known probability distribution. The hypotheses to be tested are:

$H_0 : X$ follows a given probability distribution,
$H_1 : X$ does not follow the given distribution.

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2(k - s - 1) \text{ under } H_0,$$

where k denotes the number of categories and s the number of parameters to be estimated from the data.
The critical region is:

$$C = \{\chi^2 | \chi^2 > \chi^2_\alpha(k - s - 1)\}.$$

Example. Consider the number of accidents per day randomly sampled over 100 days:

# accidents	0	1	2 (or more)
days	50	30	20

Is the number of accidents Poisson-distributed?

Recall that $X \sim$Poi$(\lambda) \Rightarrow E(X) = \lambda$, so:

$$\hat{\lambda} = \bar{x} = \frac{50 \cdot 0 + 30 \cdot 1 + 20 \cdot 2}{100} = 0.7.$$

H_0 : # accidents $\sim$ Poi(0.7),
H_1 : # accidents not $\sim$ Poi(0.7).

Now $k = 3$ and $s = 1$, because we had to calculate one distribution parameter (λ). Choosing a significance level of $\alpha = 5\%$ yields then for the critical region:

$$C = \{\chi^2 | \chi^2 > \chi^2_\alpha(k-s-1) = \chi^2_{0.05}(1) = 3.84\}.$$

Now under the Poisson distribution:

$$p_i = P(\#\text{ accidents}{=}i) = \frac{\lambda^i}{i!}e^{-\lambda} = \frac{0.7^i}{i!}e^{-0.7}$$

such that

$$p_0 = 0.4966, \quad p_1 = 0.3476,$$
$$p_{\geq 2} = 1 - p_0 - p_1 = 0.1558,$$

and the expected frequencies $e_i = np_i$ become (n=100):

$$e_0 = 49.66, \quad e_1 = 34.76, \quad e_{\geq 2} = 15.58.$$

Therefore:

$$\chi^2 = \frac{(50-49.66)^2}{49.66} + \frac{(30-34.76)^2}{34.76} + \frac{(20-15.58)^2}{15.58}$$
$$= 1.91 \notin C \text{ (since } 1.96 < 3.84\text{)}.$$

So we accept H_0 and regard the number of accidents as Poisson-distributed.