STATISTICS

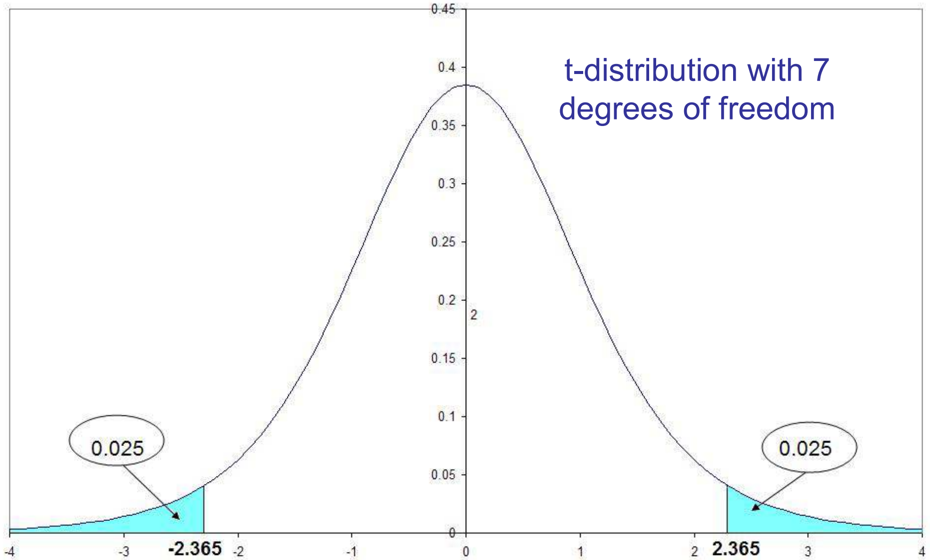
F-test
two-sample t-test
Cochran-test

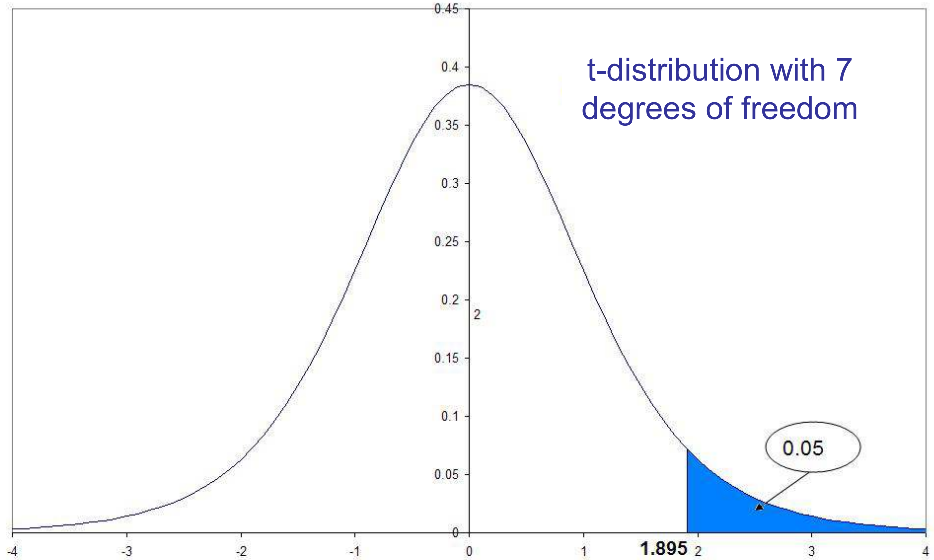
Variance analysis (ANOVA)

One- and two-sided tests

- Two-sided tests
  - H_0: no change
  - H_a: there is change (any direction)

- One-side test
  - H_0: the average did not increase
  - H_a: the average increased





In case of p-values: *p*(one-sided)=*p*(two-sided)/2

Interpretation of the significance

- Significant difference: $p<\alpha$, $p<0.05$. It is stated that the compared populations are different. The probability error of the decision is small (maximum α – this is the error of the first kind or Type I error).
- Non-significant difference: $p>\alpha$, $p>0.05$. In this case, all you can say is that there is not sufficient information to detect the difference. Maybe
  - actually there is no difference;
  - there is a difference, however only a few number of elements were available;
  - there was a standard deviation;
  - it was wrong of the test method;
  - …
- Statistical significance should always be thought through whether for example it is significant in agricultural point of view;
- When providing statistical significance, indication of the p-value is also advisable;

F-test for detecting identity of variances of two normally distributed random variables

- Our hypothesis for the identity of the variances of two independent random variables of normal distribution with unknown expectation and variance is checked by the so-called F-test.

- H_0: $\sigma_1^2 = \sigma_2^2$
- H_1: $\sigma_1^2 > \sigma_2^2$

To be peformed before t-test!

- The test is always carried out as a one-sided test (it could be carried out otherwise as well)

- Test statistics: $F_{sz} = \frac{s_1^{*2}}{s_2^{*2}}$ where $s_1^{*2} > s_2^{*2}$

numerator: $DF_1 = n_1 - 1$ denominator: $DF_2 = n_2 - 1$

- If H_0 fulfils, then F_{sz} is of F-distribution with degrees of freedom n_1-1, n_2-1

- Decision principle: for $F_{sz} \leq F_\alpha$ 0-hypothesis is accepted, otherwise not.

Two-sample *t*-test

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If the samples are independent, normally distributed and their standard deviations do not differ significantly, they be seen as two parts of a single sample.

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} \cdot \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

The test statistics received is of t-distribution with degrees of freedom $n_1 + n_2 - 2$

Conditions of the *t*-test:

- For one-sample t-test:
- the random variable is normally distributed;
- the sample elements are independent;

- For two-sample t-test, in addition:
- standard deviations of the same two random variables are identical;

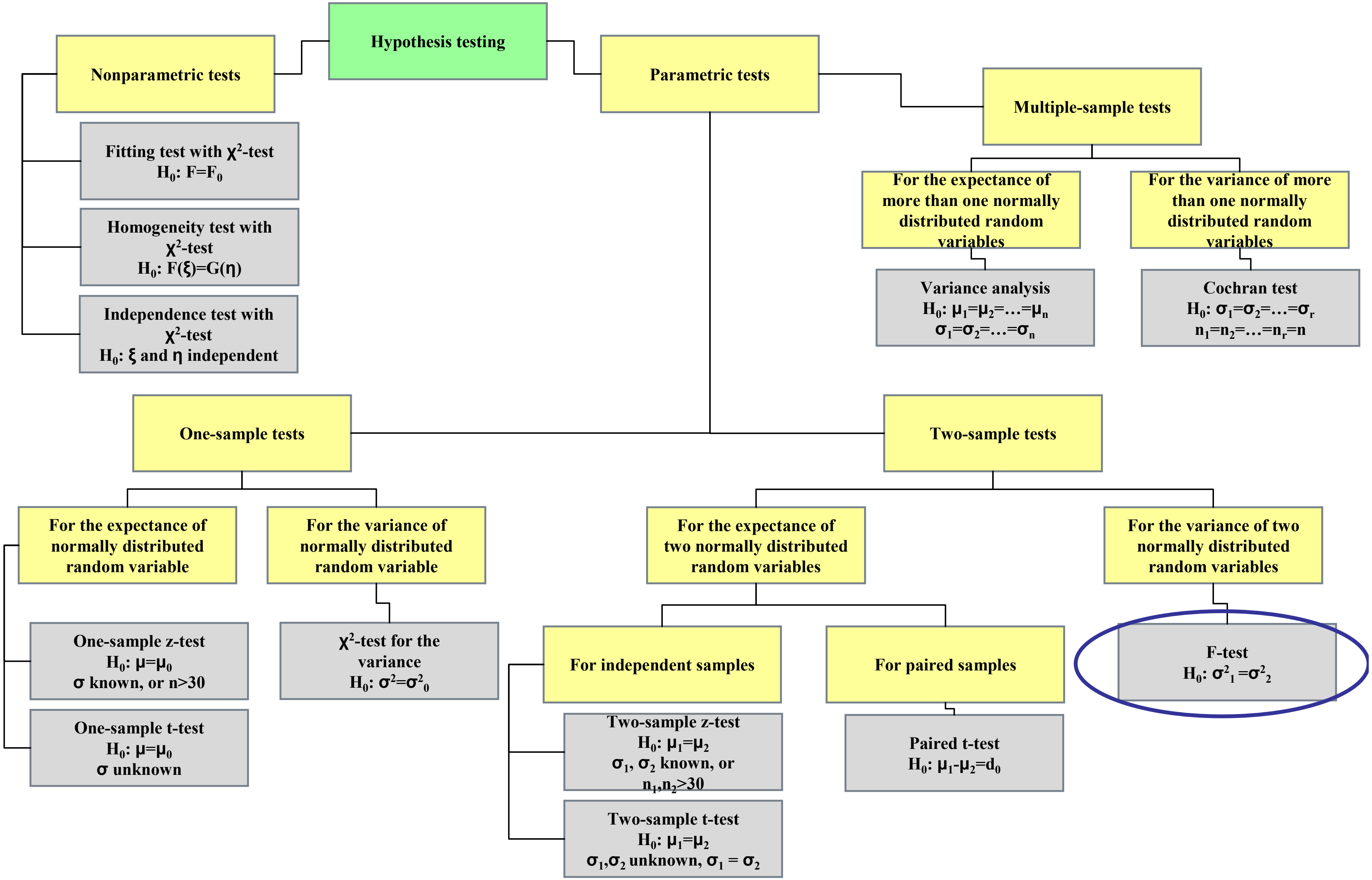
Table of the *t*-distribution

and the test statistics of the one-sample t-test

$$t = \frac{\bar{x}(- m)}{s/\sqrt{n}}$$

degree of freedom	0.05	0.025	0.01	0.005
	critical values belongong to p			
1	6.314	12.706	31.821	63.656
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
30	1.697	2.042	2.457	2.750
40	1.684	2.021	2.423	2.704
50	1.676	2.009	2.403	2.678
60	1.671	2.000	2.390	2.660
80	1.664	1.990	2.374	2.639
100	1.660	1.984	2.364	2.626
150	1.655	1.976	2.351	2.609

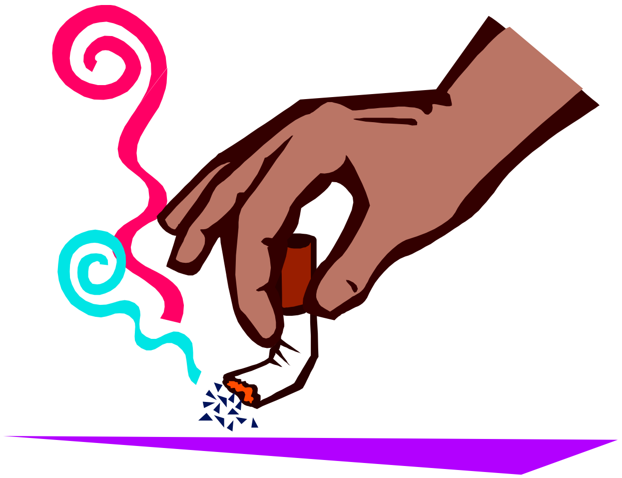
F-test for detecting identity of variances of two normally distributed random variables



Task (*F*-test)

CO emissions of cigarettes from two different brands were tested. The data were as follows. May we assume that the standard deviation of the CO emission of the two brands are the same?

	„A"	„B"
n	11	10
Mean	16,4 mg	15,6 mg
s*	1,2 mg	1,1 mg



Solution of the task (*F*-test)

H_0: $\sigma_1 = \sigma_2$

H_1: $\sigma_1 > \sigma_2$

$\alpha = 0{,}05$, $DF_1 = 10$, $DF_2 = 9$

$F_{0,05} = 3{,}13$

$$F_{sz} = \frac{1{,}2^2}{1{,}1^2} = 1{,}19$$

- Since F_{sz} occurs in the acceptance range, there is no reason to reject H_0 in the 5% significance level.

Table of *F*-test

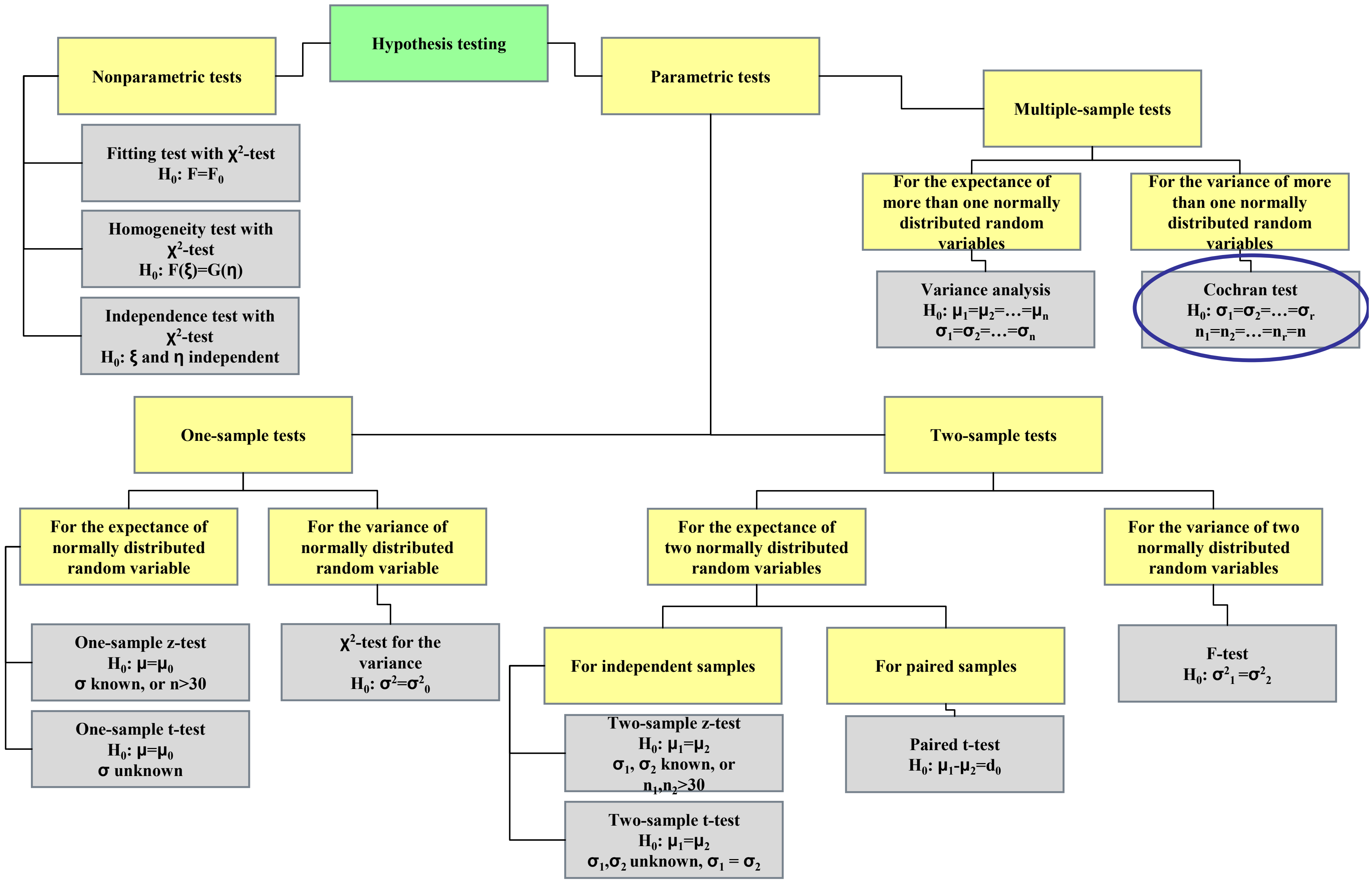
Critical values of *F*-test at the 95% probability level

Nevező	Számláló szabadságfoka																						
szab. foka	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
2	18.5	19	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.6	8.58	8.57	8.56	8.54	8.54	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6	5.96	5.93	5.91	5.87	5.84	5.8	5.77	5.74	5.71	5.7	5.68	5.66	5.65	5.64	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.7	4.68	4.64	4.6	4.56	4.53	4.5	4.46	4.44	4.42	4.4	4.38	4.37	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06	4.03	4	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.6	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.34	3.31	3.28	3.23	3.2	3.15	3.12	3.08	3.05	3.03	3	2.98	2.96	2.94	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.1	3.07	3.02	2.98	2.93	2.9	2.86	2.82	2.8	2.77	2.76	2.73	2.72	2.71
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.7	2.67	2.64	2.61	2.59	2.56	2.55	2.54
12	4.75	3.88	3.49	3.26	3.11	3	2.92	2.85	2.8	2.76	2.72	2.69	2.64	2.6	2.54	2.5	2.46	2.42	2.4	2.36	2.35	2.32	2.31	2.3
14	4.6	3.74	3.34	3.11	2.96	2.85	2.77	2.7	2.65	2.6	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.2	2.16	2.13	2.09	2.07	2.04	2.02	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2	1.98	1.95	1.93	1.92
20	4.35	3.49	3.1	2.87	2.71	2.6	2.52	2.45	2.4	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.9	1.87	1.85	1.84
22	4.3	3.44	3.05	2.82	2.66	2.66	2.47	2.4	2.35	2.3	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.8	1.78
23	4.28	3.42	3.03	2.8	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.2	2.14	2.1	2.04	2	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
25	4.24	3.38	2.99	2.76	2.6	2.49	2.41	2.34	2.28	2.24	2.2	2.16	2.11	2.06	2	1.96	1.92	1.87	1.84	1.8	1.77	1.74	1.72	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.1	2.05	1.99	1.95	1.9	1.85	1.82	1.78	1.76	1.72	1.7	1.69
28	4.2	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.23	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
32	4.15	3.3	2.9	2.67	2.51	2.4	2.32	2.25	2.19	2.14	2.1	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
34	4.13	3.28	2.88	2.65	2.49	2.38	2.3	2.23	2.17	2.12	2.08	2.05	2	1.95	1.89	1.84	1.8	1.74	1.71	1.67	1.64	1.61	1.59	1.57
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.1	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
38	4.1	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.8	1.76	1.71	1.67	1.63	1.6	1.57	1.54	1.53
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2	1.95	1.9	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.6	1.57	1.54	1.51	1.49
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.1	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.5	1.48
46	4.05	3.2	2.81	2.57	2.42	2.3	2.22	2.14	2.09	2.04	2	1.97	1.91	1.87	1.8	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
48	4.04	3.19	2.8	2.56	2.41	2.3	2.21	2.14	2.08	2.03	1.99	1.96	1.9	1.86	1.79	1.74	1.7	1.64	1.61	1.56	1.53	1.5	1.47	1.45
50	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07	2.02	1.98	1.95	1.9	1.85	1.78	1.74	1.69	1.63	1.6	1.55	1.52	1.48	1.46	1.44
100	3.94	3.09	2.7	2.46	2.3	2.19	2.1	2.03	1.97	1.92	1.88	1.85	1.97	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.3	1.28
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.9	1.86	1.83	1.77	1.72	1.65	1.6	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.8	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.9	1.85	1.81	1.78	1.72	1.67	1.6	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
1000	3.85	3	2.61	2.38	2.22	2.1	2.02	1.95	1.89	1.84	1.8	1.76	1.7	1.65	1.58	1.53	1.47	1.41	1.36	1.3	1.26	1.19	1.13	1.08
∞	3.84	2.99	2.6	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.4	1.35	1.28	1.24	1.17	1.11	1

Cochran test for detecting identity of variances of more than two normally distributed random variables

- Let *r* normally distributed random variables are given
- H_0: $\sigma_1 = \sigma_2 = \ldots = \sigma_r$
- H_1: **standard deviation of the variable having the maximum standard deviation, significantly differs from those of the others**
- **Cochran test can be applied if the element numbers of the random variables (*n*) are the same**.
- Corrected empirical variance of the *j*-th sample: s_j^{*2}
- s_{max}^{*2} is the maximum corrected empirical variance among the s_j^{*2} values.
- Test statistics: $g_{sz} = \dfrac{s_{max}^{*2}}{\sum_{j=1}^{r} s_j^{*2}}$
- Degree of freedom: $DF = n - 1$
- Knowing α, *DF* and *r* , $\Rightarrow$ g_{krit} can be determined from the table of Cochran test;
- Decision: if $g_{sz} \leq g_{krit}$, then H_0 is accepted, otherwise not.

Cochran test for detecting identity of variances of more than two normally distributed random variables



Task (Cochran test)*

- Műselyem For silk tensile strength testing of 20 pieces (r = 20), 10-item data for each r (n = 10), the following corrected empirical standard deviations were calculated for the tensile strength (see the table below). Is it presumable that there is no significant difference between standard deviations of the random variables studied, if the level of significance is 5% ?

i	1.	2.	3.	4.	5.	6.	7.	8.	9	10.
s_i^{*2}	*24,9*	*8,4*	*21,2*	*8,0*	*8,4*	*6,0*	*26,3*	*26,7*	*6,8*	*12,5*
i	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.
s_i^{*2}	*12,5*	*11,4*	*4,8*	*22,2*	*22,6*	*16,1*	*10,9*	*9,6*	*60,5*	*10,9*

* Source: Kövesi J.: Kvantitatív módszerek, Oktatási segédanyag, BME MBA Mérnököknek program. (Quantitative methods, Educational aid, BME MBA for Engineers programme.) Budapest, 1998

Solution of the task (Cochran-test)

- H_0: standard deviations are identical
- H_1: the highest standard deviations significantly differs from the others

n = 10
DF = n-1= 10-1=9
r = 20, α = 5%

$g_{krit} = 0,135$

$$g_{sz} = \frac{s^{*2}_{max}}{s^{*2}_{1} + s^{*2}_{2} + \ldots + s^{*2}_{r}}$$

$$g_{sz} = \frac{60,5}{330,7} = 0,183$$

i	s_i^2	i	s_i^2
1	24,9	11	12,5
2	8,4	12	11,4
3	21,2	13	4,8
4	8,0	14	22,2
5	8,4	15	22,6
6	6,0	16	16,1
7	26,3	17	10,9
8	26,7	18	9,6
9	6,8	19	60,5
10	12,5	20	10,9

$g_{sz} > g_{krit} \Rightarrow$ H_0 is rejected, namely the highest standard deviation significantly differs from the others.

Cochran-test, critical values of G_{krit} at the 5% probability level

	r															
DF	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	-	0,96	0,91	0,85	0,79	0,72	0,68	0,64	0,59	0,55	0,48	0,39	0,34	0,28	0,23	0,17	0,10
2	0,98	0,89	0,76	0,69	0,63	0,57	0,52	0,48	0,45	0,395	0,34	0,28	0,24	0,20	0,16	0,115	0,067
3	0,96	0,82	0,68	0,61	0,55	0,50	0,44	0,42	0,38	0,335	0,285	0,23	0,20	0,16	0,13	0,092	0,053
4	0,92	0,77	0,63	0,56	0,495	0,44	0,39	0,37	0,34	0,295	0,245	0,20	0,17	0,14	0,11	0,078	0,043
5	0,90	0,73	0,58	0,52	0,455	0,40	0,36	0,335	0,31	0,27	0,22	0,18	0,15	0,125	0,098	0,068	0,038
6	0,86	0,70	0,56	0,48	0,42	0,37	0,335	0,31	0,29	0,245	0,20	0,165	0,135	0,11	0,087	0,06	0,033
7	0,85	0,68	0,53	0,46	0,40	0,355	0,32	0,295	0,27	0,23	0,19	0,155	0,13	0,105	0,083	0,057	0,031
8	0,83	0,65	0,52	0,44	0,38	0,34	0,305	0,28	0,255	0,22	0,18	0,145	0,12	0,098	0,076	0,052	0,039
9	0,81	0,63	0,50	0,42	0,37	0,325	0,29	0,27	0,245	0,21	0,17	0,135	0,115	0,092	0,07	0,049	0,038
10	0,79	0,61	0,48	0,41	0,355	0,31	0,28	0,26	0,235	0,20	0,16	0,13	0,11	0,088	0,067	0,047	0,036
12	0,77	0,58	0,46	0,39	0,34	0,29	0,27	0,245	0,225	0,19	0,15	0,122	0,105	0,082	0,063	0,043	0,034
14	0,75	0,56	0,445	0,375	0,33	0,28	0,255	0,235	0,215	0,18	0,145	0,115	0,098	0,078	0,058	0,041	0,033
16	0,73	0,54	0,43	0,365	0,315	0,27	0,25	0,225	0,205	0,175	0,138	0,11	0,092	0,075	0,055	0,040	0,032
18	0,72	0,53	0,42	0,355	0,305	0,26	0,245	0,22	0,20	0,17	0,133	0,105	0,09	0,072	0,053	0,038	0,031
20	0,70	0,52	0,41	0,35	0,295	0,255	0,235	0,215	0,195	0,165	0,13	0,10	0,087	0,07	0,051	0,037	0,0305
25	0,68	0,51	0,395	0,34	0,29	0,245	0,23	0,205	0,185	0,155	0,122	0,098	0,083	0,067	0,050	0,036	0,03
30	0,67	0,49	0,38	0,33	0,28	0,235	0,215	0,20	0,175	0,145	0,118	0,092	0,08	0,063	0,048	0,034	0,03
40	0,65	0,47	0,365	0,31	0,265	0,22	0,205	0,185	0,16	0,135	0,11	0,085	0,074	0,058	0,045	0,032	0,029
50	0,63	0,44	0,35	0,295	0,25	0,21	0,195	0,175	0,15	0,125	0,105	0,08	0,07	0,055	0,043	0,03	0,028
60	0,61	0,43	0,34	0,285	0,24	0,205	0,185	0,165	0,145	0,12	0,098	0,075	0,065	0,052	0,04	0,028	0,027
90	0,58	0,41	0,32	0,265	0,225	0,19	0,17	0,15	0,13	0,11	0,09	0,07	0,058	0,047	0,037	0,026	0,025
120	0,56	0,39	0,30	0,25	0,215	0,175	0,16	0,14	0,125	0,10	0,085	0,065	0,055	0,044	0,035	0,024	0,024
240	0,53	0,37	0,28	0,225	0,19	0,16	0,14	0,13	0,11	0,09	0,075	0,055	0,048	0,038	0,03	0,02	0,022
∞	0,49	0,33	0,25	0,19	0,17	0,135	0,12	0,11	0,095	0,075	0,065	0,049	0,042	0,032	0,025	0,018	0,01

Column 1: number of degrees of freedom;
Heading: r = number of groups (samples);

Mixed relationships – repetition

Similarity with ANOVA

Indications

$x_{ij} - \bar{x}$ = full difference (d_{ij})

$(x_{ij} - \bar{x}_j)$ = internal difference (B_{ij})

$(\bar{x}_j - \bar{x})$ = external difference (K_{ij})

σ^2 = full variance

σ_B^2 = internal variance

σ_K^2 = external variance

Calculation of variance

$$\sigma^2 = \frac{\sum_{i=1}^{n_j}\sum_{j=1}^{m}(x_{ij}-\bar{x})^2}{n} = \frac{S}{n}$$

S: Total sum of squares

$$\sigma_B^2 = \frac{\sum\sum(x_{ij}-\bar{x}_j)^2}{n} = \frac{\sum n_j\sigma_j^2}{n} = \frac{S_B}{n}$$

S_B: Internal sum of squares

$$\sigma_K^2 = \frac{\sum n_j(\bar{x}_j-\bar{x})^2}{n} = \frac{S_K}{n}$$

S_K: External sum of squares

Relationships

$$d_{ij} = B_{ij} + K_{ij}$$

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

full difference internal difference external difference

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

full variance internal variance external variance

$$S = S_B + S_K$$

Total sum of squares Internal sum of squares External sum of squares

Example:

In a college, bachelor training occurs in 4 professions. The time spent on the students' daily learning is as follows.

Profession	Time spent on daily learning (hr)		Students (%)
	Mean x_j	*St. deviation σ_j	
Human resources	1,5	1,2	24
Management	2,25	0,8	26
International management	1,75	1,5	20
Finance & accounting	2,75	1,3	30

*Standard deviation

Calculate $\sigma_B, \sigma_K, \sigma$ and interpret them!

Solution

$$\bar{x} = 0{,}24 \cdot 1{,}5 + 0{,}26 \cdot 2{,}25 + 0{,}2 \cdot 1{,}75 + 0{,}3 \cdot 2{,}75 = 2{,}12$$

$$\sigma_K^2 = 0{,}24 \cdot (1{,}5 - 2{,}12)^2 + \ldots + 0{,}3 \cdot (2{,}75 - 2{,}12)^2 = 0{,}2431$$

$$\sigma_k = 0{,}49$$

$$\sigma_B^2 = 0{,}24 \cdot 1{,}2^2 + 0{,}26 \cdot 0{,}8^2 + 0{,}2 \cdot 1{,}5^2 + 0{,}3 \cdot 1{,}3^2 = 1{,}469$$

$$\sigma_B = 1{,}212$$

$$\sigma^2 = \sigma_B^2 + \sigma_K^2$$

$$\sigma^2 = 1{,}469 + 0{,}2431 = 1{,}7121 \rightarrow \sigma = 1{,}308$$

Indicators of mixed relationships

Variance-ratio: show that to what extent (in percentage) the classification of a quality or areal criterion affects dispersion of a quantitative criterion.

$$H^2 = \frac{\sigma_K^2}{\sigma^2} = 1 - \frac{\sigma_B^2}{\sigma^2} = \frac{S_K}{S} = 1 - \frac{S_B}{S}$$

Quotient of standard deviations (square root of variance-ratio): shows that how strong is the relationship between the non-quantitative (grouping) and quantitative criteria.

$$H = \sqrt{H^2} = \sqrt{\frac{\sigma_K^2}{\sigma^2}} = \frac{\sigma_K}{\sigma} = \sqrt{1 - \frac{\sigma_B^2}{\sigma^2}} = \sqrt{\frac{S_K}{S}} = \sqrt{1 - \frac{S_B}{S}}$$

Interpretation of the indicators of the mixed relationships

$$0 < H < 1$$

$$0 < H^2 < 1$$

Stochastic relationships

$$H = H^2 = 0$$

Total independence, total lack of relationship

$$H = H^2 = 1$$

Function-like, deterministic relationship

Procedures based on ranking (a type of non-parametric tests)

- What if the conditions of the t-test (normality, identity of variances) do not fulfil???
  - applying transformations (log, square root, arcsin, …);
  - non-parametric tests – procedures based on ranking;

- Non-parametric tests can be used if
  - the conditions of the parametric tests do not fulfil;
  - we can not control (small sample size);
  - we do not want to control;
  - ordinal variables (how glad I am for spring??? – little, medium, very much);

- Only the magnitude of the data matters; it is unimportant how much is one data bigger than the other;
- Calculation: based on ranking;
- **BUT:** not the same null hypothesis is tested as the parametric tests. So they cannot be regarded as non-parametric "equivalents " of parametric tests;

Table of Khi^2 distribution

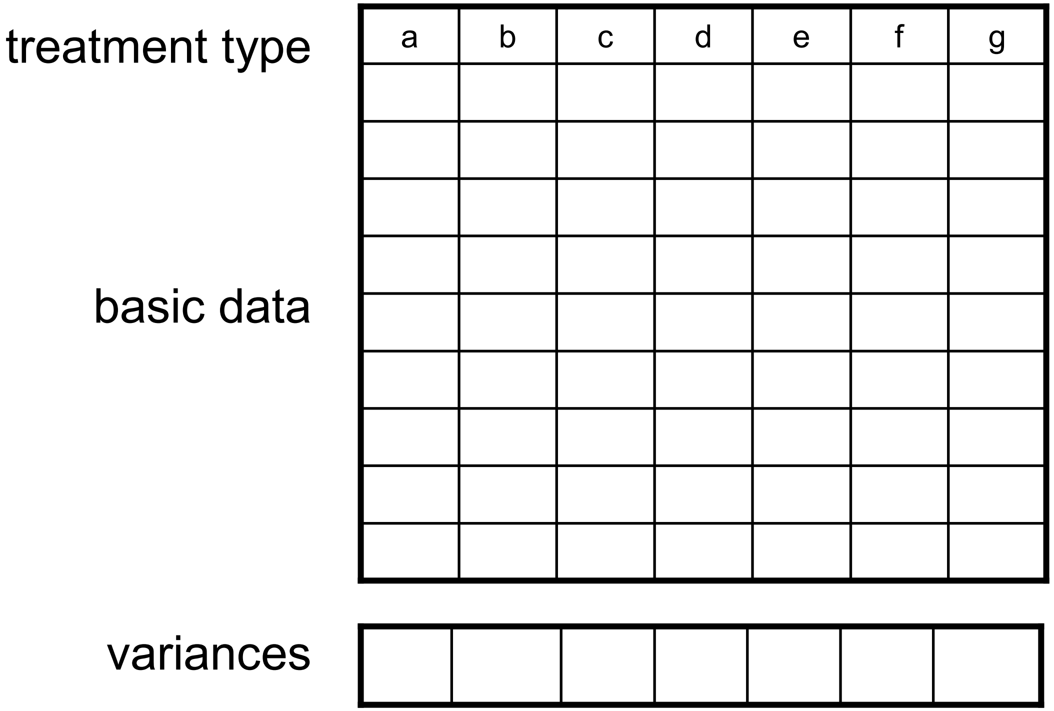
One-side and two-sided test

degree of freedom	0.995	0.975	0.95	0.05	0.025	0.01	0.005
			critical values belongong to p				
1	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	15.308	16.928	41.337	44.461	48.278	50.994
29	13.121	16.047	17.708	42.557	45.722	49.588	52.335
30	13.787	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	74.222	77.929	124.342	129.561	135.807	140.170

Studying more than two groups

One-way ANOVA

It's base is a single *F*-test, which compares ”between groups” variance (characterizing the differences of the means) to the ”within groups” variance (characterizing random differences)

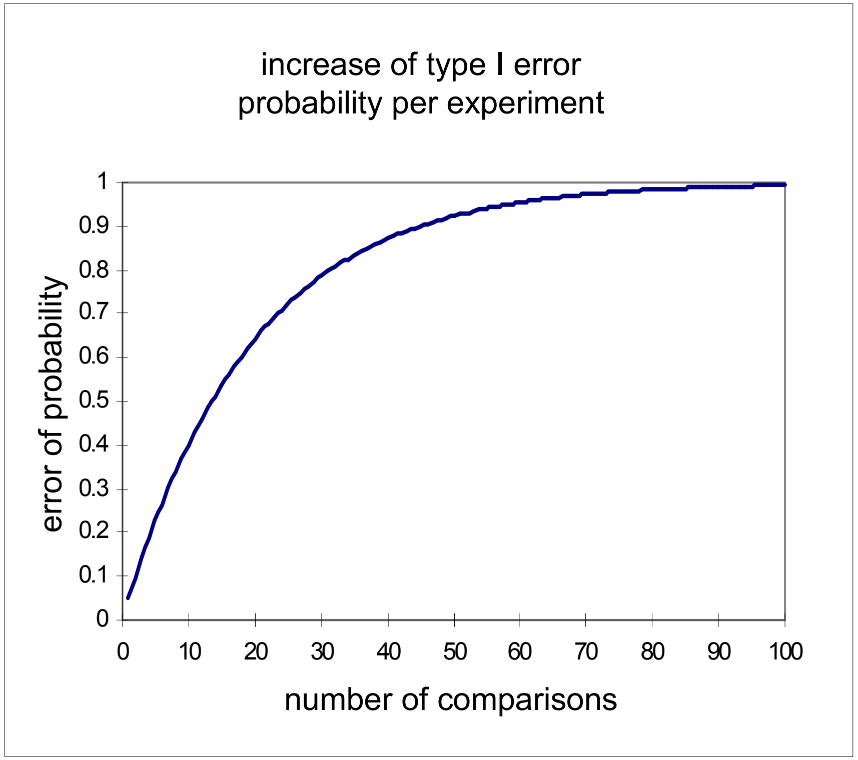


Why not perform pairwise comparisons?

- inefficient;
- it may distort our decisions, because when making pairwise comparisons, random may also cause "significant" results;

  if e.g. α=0,05, then on average, in every 20-th case, we're making a type I error, that is we reject a true 0-hypothesis;

  In other words: **we do not know which of the significant results are attributable to the random and which reflect a real difference.**



- A lot of false comparison "inflates" the significance levels;

Repeated pairwise comparisons, joint probabilities

Number of indepedent decisions	*Nominal significance level*	*Probability of correct decision*	*Probability of wrong decision*
1	**0.05**	**0.950**	**0.050**
2	**0.05**	**0.903**	**0.098**
3	**0.05**	**0.857**	**0.143**
4	**0.05**	**0.815**	**0.185**
5	**0.05**	**0.774**	**0.226**
6	**0.05**	**0.735**	**0,265**
7	**0.05**	**0.698**	**0.302**
8	**0.05**	**0.663**	**0.337**
9	**0.05**	**0.630**	**0.370**
10	**0.05**	**0.599**	**0.401**
20	**0.05**	**0.358**	**0.642**
40	**0.05**	**0.129**	**0.871**

Variance analysis – ANOVA (1)

- Let r be pieces of normally distributed random variables
- It is assumed that the random variables have the same variance, i.e. $\sigma_1 = \sigma_2 = \ldots = \sigma_r$. This is an important condition for implementating ANOVA. Existence of this condition can be tested by Cochran test.
- H_0: $\mu_1 = \mu_2 = \ldots = \mu_r$
- H_1: at least one expected value significantly differs from the others
- $n_1, n_2, \ldots, n_r$ element numbers of the independent items of the random variables, while n is the sum of the element number of the items.
- x_{ij} is the j-th element of the i-th item (i=1, 2,…,r), (j=1, 2,…,n_i)
- $\overline{x}$ is the average of the elements of all items, $\overline{x_i}$ is the average of the az i-edik item

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{n}\sum_{i=1}^{r} n_i \overline{x_i}$$

Variance analysis – ANOVA (1)

- Let form the following statistics:

$$SSK = \sum_{i=1}^{r} n_i \left(\overline{x_i} - \overline{x}\right)^2$$

External sum of squares

$$SSB = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i}\right)^2$$

Internal sum of squares

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x}\right)^2$$

Total sum of squares

- *SST=SSK+SSB;*
- If H_0 is true (and the identity of standard deviation fulfils), then
  - ✓ *SSB* is of χ^2-distribution with degree of freedom *r*-1 , while *SSK* is of χ^2-distribution with degree of freedom *n-r* ;
  - ✓ *SSK* is independent from *SSB*; $s_k^2 = \frac{SSK}{r-1}$ external variance and $s_b^2 = \frac{SSB}{n-r}$ internal variance are independent form each other, their expectance value equal to each toher and to the unknown variance of the population;
- To decide on variance equality, *F*-test is applied. If H_0 fulfils, the test staticstics is of *F*-distribution with degree of freedom *r*-1, *n-r* ; $F_{sz} = s_k^2 / s_b^2$

Variance analysis – ANOVA table

- Calculations can be arranged into a so called ANOVA table

Name of sum of squares	Sum of squares	Degree of freedom	Assessment of st. dev.*	F-value	p-value
Between groups	$\sum_{i=1}^{r} n_i (\bar{x}_i - \bar{x})^2$	r-1	s_k^2	s_k^2/s_b^2	p
Within groups	$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	n-r	s_b^2	-	-
Total	$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	n-1	-	-	-

*standard deviation

- Two ways of decisions are possible:
  - H_0 is accepted, if $F_{sz} \leq F_{krit}$, otherwise H_0 is rejected;
  - H_0 is accepted, if $p > \alpha$, otherwise H_0 is rejected;
- *p*-value is the maximum type I error probability (significance level) at which the null hypothesis might be accepted;

Analysis of more than one groups consists of *two* steps

- To determine whether there is a significant difference between the results of the set of groups;
- If it is so, then look for significant differences between the groups:
  - Difference may be not only in the form of difference between groups;

Basic idea of the analysis: variance is estimated in two ways in all samples

- The idea of ANOVA comes from **R.A. Fisher,** who worked at an agricultural experimental station in England, between 1918-25.
- **His ingenious Recognition:** in experiments performed with several groups, null hypothesis H_0 can also be examined so that the population variance is estimated in two ways and the look whether these these estimates are in good agreement or not.

1. on the dispersions of within groups/samples we can conclude the *variance of the population*
2. on the dispersions of the groups/sample averages we can conclude the same *variance* .

The sum of squares can be divided into additive elements

- Distance of the sub-sample elements from the high average of the whole sample is estimated by the sum of squares:

  $\Sigma (x_{high\ average} - x_i)^2$;
- The sum of squares can be particioned by the methods of algebra
  (can be divided into parts additively)
- **Each portion is decomposed so that they comply with the *specific proportion of standard deviation***
- **<u>The "internal" variance corresponds to the random, while of the variance "between the averages" meets the difference between the groups</u>**

One-way ANOVA

- Several independent samples are given
- Objective: comparison of averages
- **Conditions:**
  - The individuals get randomly to one or the other group, **the sample are independent** (one individual cen get into one and only one group).
  - The variable comprising the values to be compared is continuous.
  - **The samples come from normally distributed population**.
  - **The populations from which the samples come, have equal variance.**

- Null hypothesis:
  - The independent samples come from a population with the same distribution, i.e.

  the population averages are the same

Methodology

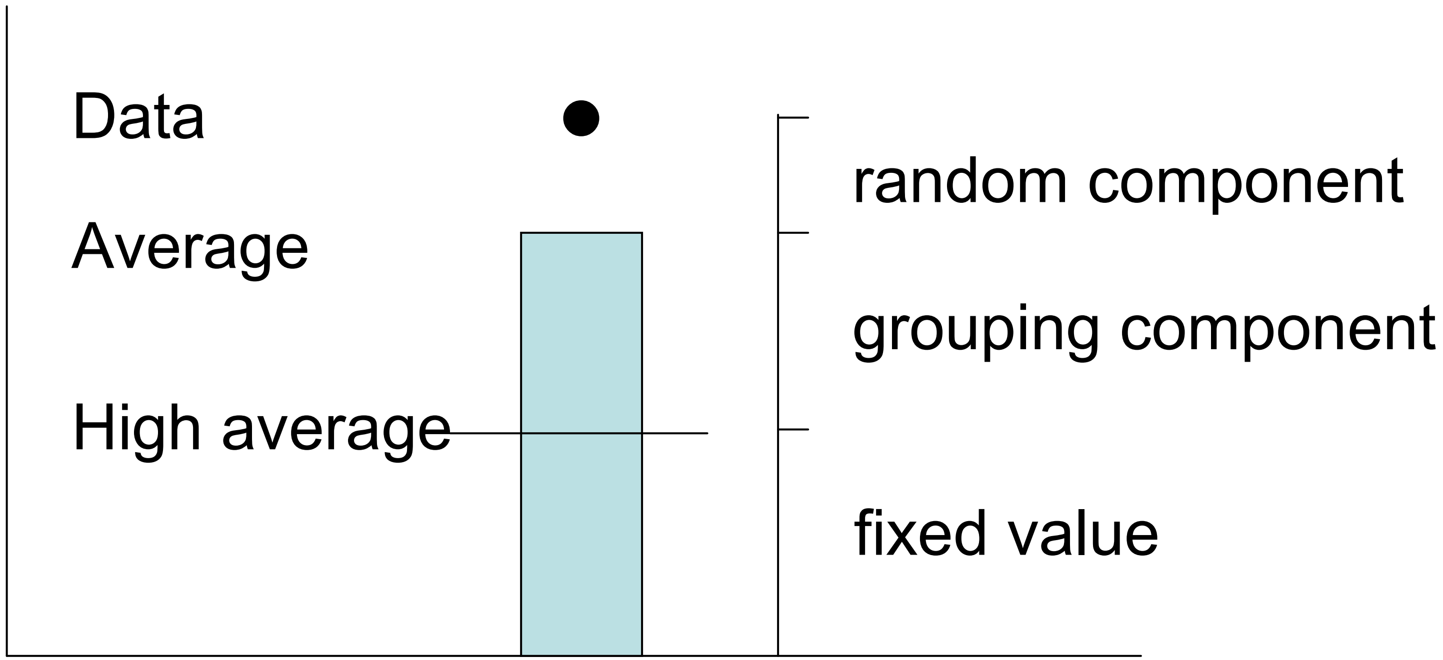
- ANOVA gets the total variance of the whole data set from two sources:
  - Between groups variance
  - Within groups variance
- **Starting null hypothesis:** the population averages are the same;

  The above assumption is equivalent to the following: between groups and within groups variances are the same in the population. By comparing these two variances, one can conclude the identity of the averages.
- **'New' null hypothesis:** between groups and within groups variances are the same in the population.
- **Testing:** estimation of the two variances are shown in the table below. The test statistic is the quotient of the two variances. Testing: *F*-test (one-sided).
- It gives a *p*-value:
  - if p>0.05, the we accept the identity of the averages (H_0)
  - if p<0.05, the there is at least one different in the averages

Calculations of analysis of variance used to be summarized in a table

Reason of dispersion	Sum of squares	Degress of freedom	Variance	F-value
Between groups	$Q_k = \sum_{i=1}^{t} n_i(\overline{x_i} - \overline{x})^2$	t-1	$s_k^2 = \frac{Q_k}{t-1}$	$F = \frac{s_k^2}{s_b^2}$
Within groups	$Q_b = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$	N-t	$s_b^2 = \frac{Q_b}{N-t}$	
Total	$Q = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$	N-1		

Illustration to the resolution of the sum of squares



Reasoning of ANOVA

(Analysis of Variance=ANOVA)

- Samples were taken from normal distribution (each sample has *n* element);
- Indipendent samples;
- Random samples (randomization);
- Null hypothesis: the samples come from one population; $(v_1=v_2=v_3=\ldots=v_n)$
- Consequence of the null hypothesis : $(s_1^2=s_2^2=s_3^2=\ldots=s_n^2)$
- Two independent estimates are made from the samples to the standard deviation, more exactly variance of the population (σ^2);
- The quotient of the two estimates of variance follows $F_{1,2}$ distribution ($F_{1,2} = s_1^2/s_2^2$);

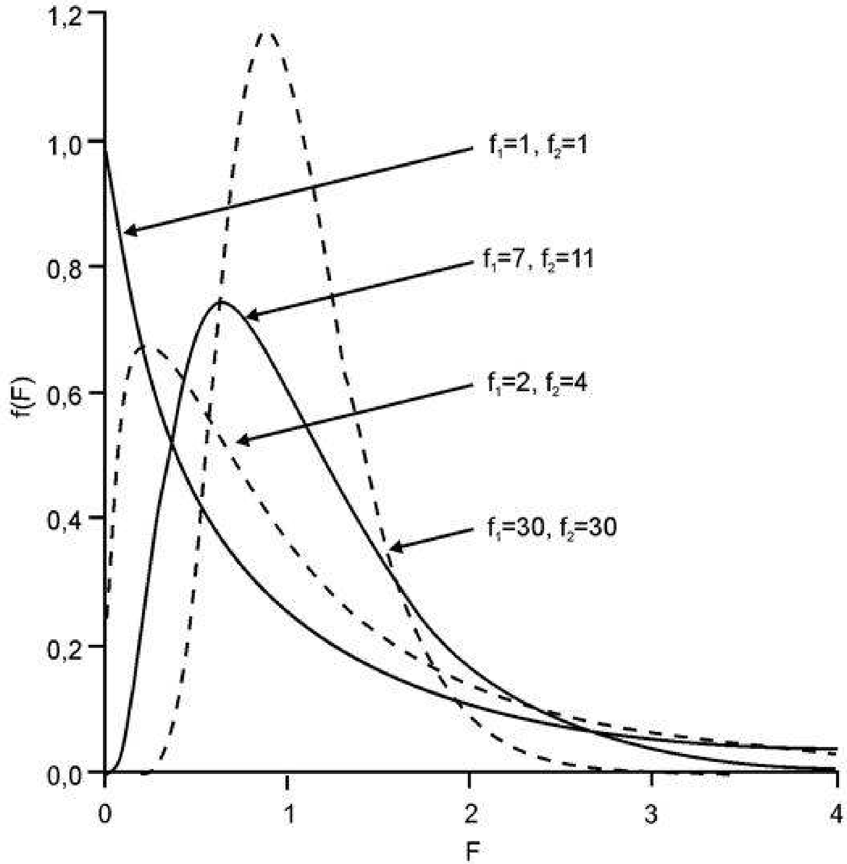
Reasoning of ANOVA (continued)

- If the samples come from one population (null-hypothesis is true), then expectance value of the distribuiton of $F_{1,2}$ is: $v(F_{1,2}) = 1$;
- If $p<0,05$ for $F_{1,2} = 1$, then the null-hypothesis is rejected;
- If the null-hypothesis is rejected, then we look for the groups that do not come from one distribution.
- Pre-planned (a priori), or post (a posteriori) comparisons are performed;

Distribution of the quotient of two variances; Fisher–Snedecor distribution

Quotient of sum of squares made from normally distributed samples

$F_{(m,n)}=s_{1(m)}^2/s_{2(n)}^2$



Relationship between ANOVA and *t*-test

- In the denominator of the *t*-test formula the standard deviation of the mean is found;
- The numerator comprises a value corresponding to the standard deviation: difference of means of two samples;
- This is none other than the difference of the two figures separately from their joint mean, divided by n-1, which for n = 2 equals one;
- In the numerator and denominator two estimates are found for the same value, and the quotient of their square is of F distribution;

Formula of the *t*-test, and its conversion

$$t = \frac{m_1 - m_2}{\dfrac{s_{1,2}}{\sqrt{n_1 + n_2 - 2}}}$$

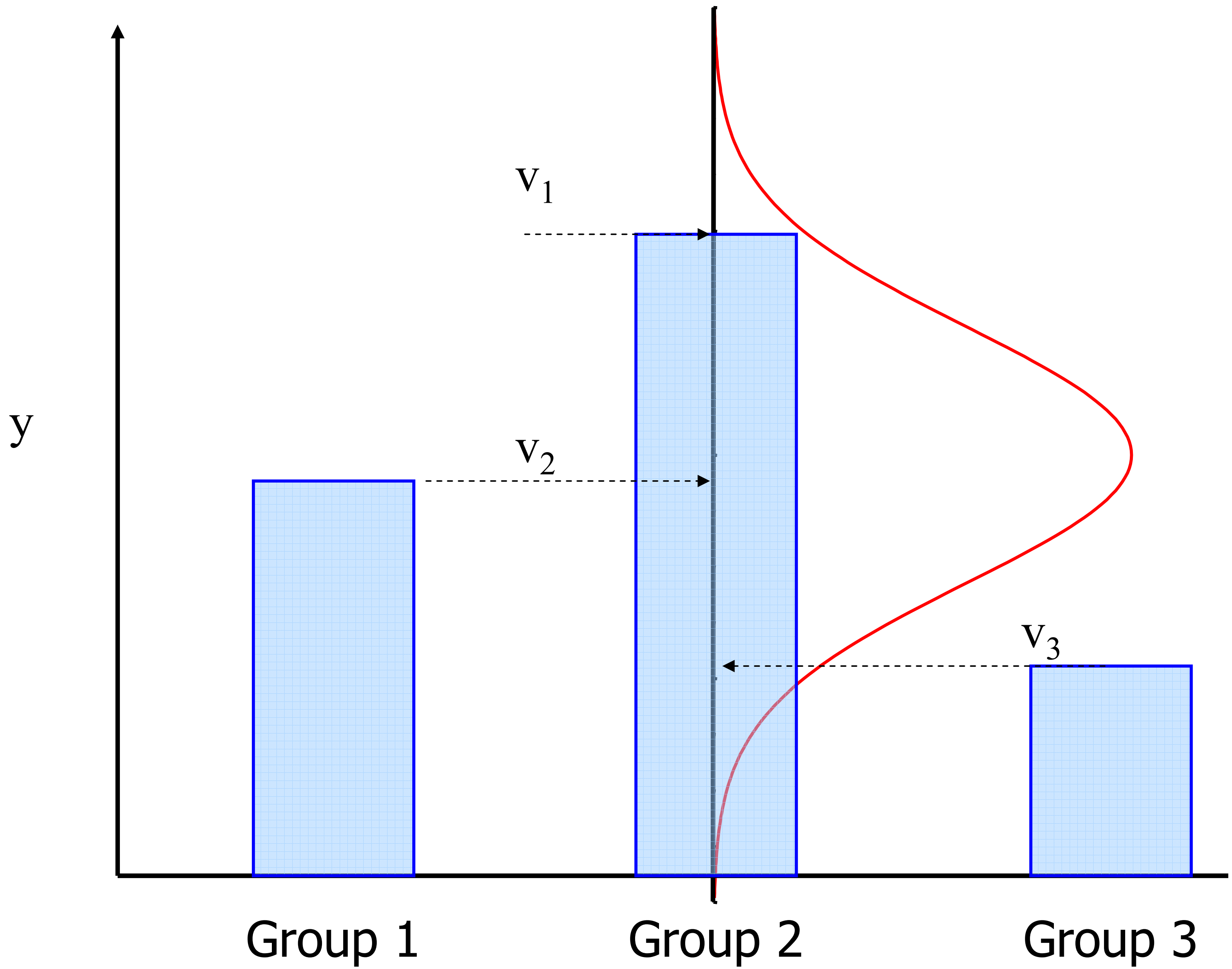
If both sides of the formula is squared:

$$t^2_{n_1+n_2-2} = \frac{\dfrac{(m_1 - m_2)^2}{1^2}}{\dfrac{s^2_{1,2}}{n_1 + n_2 - 2}}$$

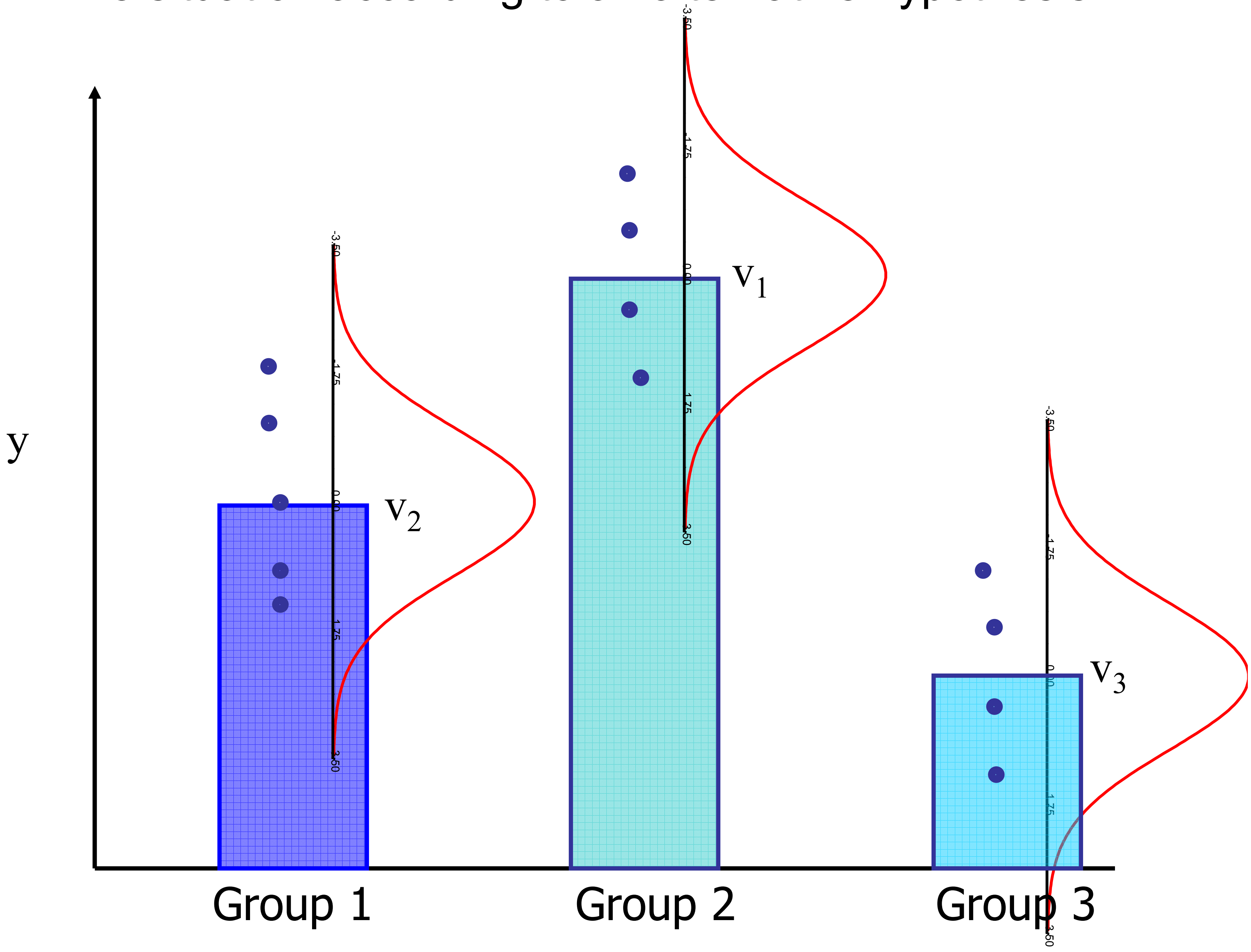
Then at the right hand side of the formula we receive the quotient of two variances, namely:

$$t^2_{n_1+n_2-2} = F_{1,n_1+n_2-2}$$

The situation according to the null hypothesis



The situation according to an alternative hypothesis



Reasoning following the significant ANOVA

Two or more statistical decision in one analysis?

- What happens in type I error, if two completely independent experiments are performed végzünk, when two independent samples are compared.
- In this case, two independent hypothesis tests and two significance tests are performed, all at **α=0,05** level. Since two independent investigations are concerned, the two significance testing can also be considered independent.
- If both null-hypothesis are valid, then the probability that at least, one of the null hypotheses is (incorrectly) rejected:
  - Let $P(s_1)=0.05$ is the above probability for the first test and $P(s_2)=0.05$ the second above probability.
    The probability of the joint occurrenc eof the two events $P(s_1)*P(s_2)$, i.e., 0.05*0.05=**0.0025**
- The three possible events: s_1 occurs alone, s_2 occurs alone, while s_1 and s_2 occur together.
- In the case of two independent experiments, the probability that at least in one of them the null hypothesis is (incorrectly) rejected:

  p= 0.05+0.05-0.0025= **0.0975**, which is **significantly higher** than 0.05 accepted for a single significance test.
- And if the experiments and comparisons are not independent?

Repeted pairwise comparisons, joint probabilities

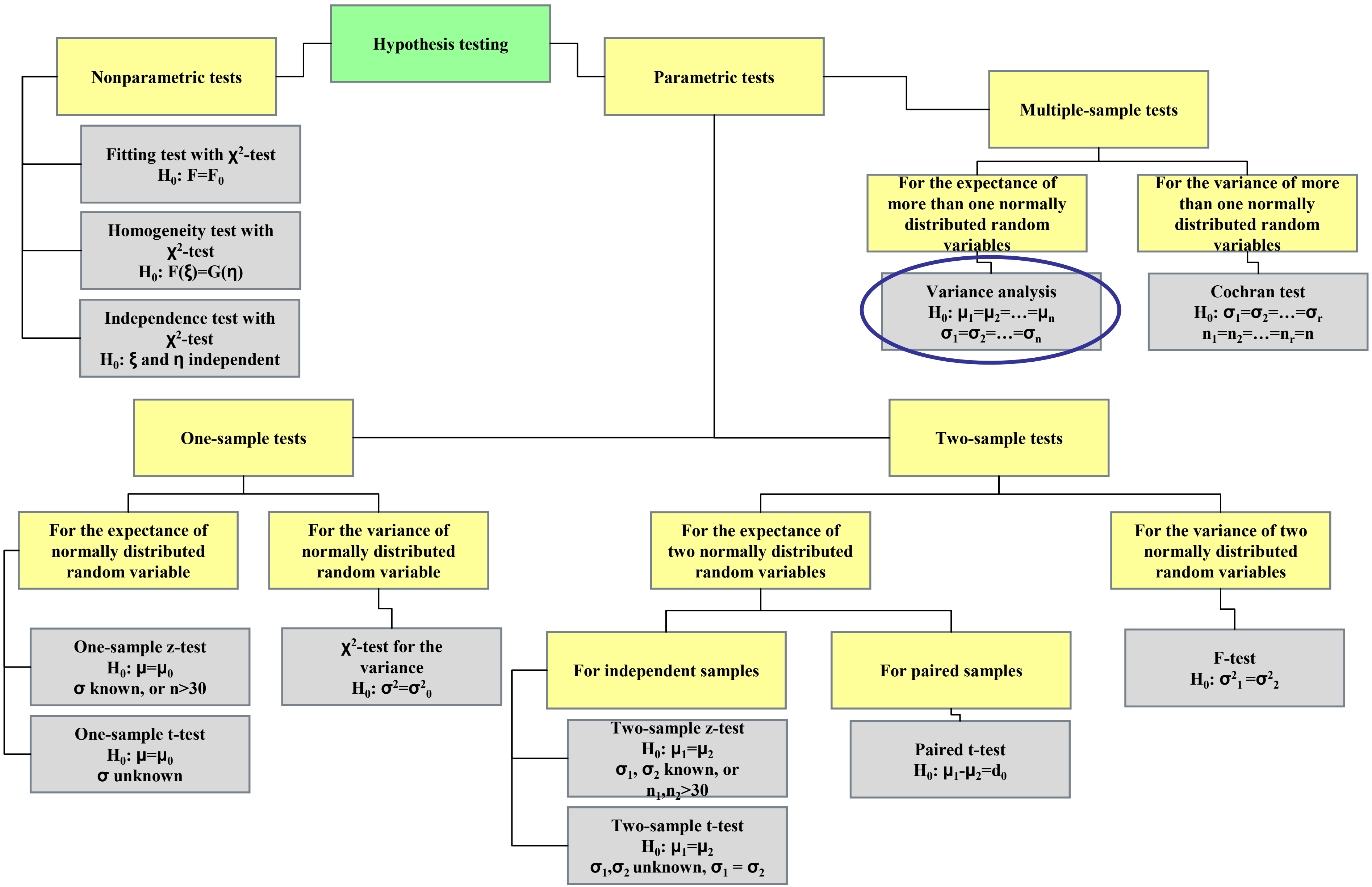
Number of independent decisions	*Nominal significance level*	*Probability of correct decision*	*Probability of wrong decision*
1	**0.05**	**0.950**	**0.050**
2	**0.05**	**0.903**	**0.098**
3	**0.05**	**0.857**	**0.143**
4	**0.05**	**0.815**	**0.185**
5	**0.05**	**0.774**	**0.226**
6	**0.05**	**0.735**	**0.265**
7	**0.05**	**0.698**	**0.302**
8	**0.05**	**0.663**	**0.337**
9	**0.05**	**0.630**	**0.370**
10	**0.05**	**0.599**	**0.401**
20	**0.05**	**0.358**	**0.642**
40	**0.05**	**0.129**	**0.871**

If there are many groups?

- If the above discussion is performed for k=10 independent tests then $p=1-(1-0.05)^{10}=0.4$
- By increasing the number of independent analyses we significantly increase the probability that such effects exist, however in reality they do not exist
- **Regarding every is possible significance tests, the tests are not independent, although the samples were independent.**

.

ANOVA for detecting identity of expectance values of more than two normally distributed random variables



Task (ANOVA)*

- Egy At 3 shops of a retail chain it was examined whether the same amount was paid for a purchase. Each store selected six random amount paid [in dollars] (see the table below). Assuming that the payments are normally distributed and the standard deviations equal: is there a difference between the three shops?

1. bolt	2. bolt	3. bolt
12,05	15,17	9,48
23,94	18,52	6,92
14,63	19,57	10,47
25,78	21,4	7,63
17,52	13,59	11,90
18,45	20,57	5,92

- H_0: expectance value of the purchase is the same in the three shops
- H_1: expectance value of the purchase is not the same in the three shops

* Source: Curwin, J., Slater, R.: Quantitative Methods for Business Decisions, Third Edition, Chapman & Hall, London, 1991

Solution of the task (ANOVA) (1)

High average: **15.195**

SSK= 6*(18.73-15.195)^2 + + … = **378.4**

$$SSB = 5.288^2 \cdot 5 + 3.106^2 \cdot 5 + 2.281^2 \cdot 5 =$$

= **214.1**

Shop 1	Shop 2	Shop 3
12.05	15.17	9.48
23.94	18.52	6.92
14.63	19.57	10.47
25.78	21.4	7.63
17.52	13.59	11.90
18.45	20.57	5.92

mean: **18.73** **18.14** **8.72**

Corrected empirical standard deviation: 5.288 3.106 2.281

Solution of the task (ANOVA) (2)

Nomination	Sum of squares	*Deg. of freedom	**Est. of st. dev.	*F*-value	*p*-value
Between groups	378,4	2	189,2	13,26	0,0005
Within groups	214,1	15	14,3	-	-
Total	592,5	17	-	-	-

*Degree of freedom;
**Estimation of standard deviation

- $\alpha = 0{,}05$, r-1 = 2, n-r = 15
- $F_{krit} = 3{,}68$
- $F_{sz} > F_{krit}$, i.e. H_0 is rejected;

Tasks: the name of the statistical method to be used, conditions of its carrying out, and if further processes can also be used, then what is their ranking.

1. Some statistics of previous lottery drawings can be downloaded from the website of the Szerencsejáték Rt. E.g. how many times the numbers have been pulled out until now. How could be examined whether it has not been fraud; in other words, whether certain numbers were pulled out in a significantly higher or smaller frequency?

2. A company offers a new reagent claiming to more effectively increasing the conductivity of a solution (it does not matter, why and how). Decide whether this statement is true or not (method or methods)?

3. An entrepreneur sells an excipient, which (according to his statetment) increases wheat yields. What method (or methods) can you decide that the statement is true?

4. We would like to compare dog breeds. Assume that there is a system of criteria, according to which the animals examined can be graded from 0 to 4:
0 - mini, 1 - low, 2 - medium - large 3, 4 - huge.
Eight selected types of 366 dogs were analysed. What kind of statistical test can be applied for detecting size difference among the types?

Always look on the bright side of things!

We finished for today, goodbye!

ямарваа нэг зүйлийн гэгээлэг
талыг нь үргэлж олж харцгаая

өнөөдөртөө ингээд дуусгацгаая, баяртай

让我们总是从光明的一面来看待事物吧！

今天的课程到此结束，谢谢！

دعونا ننظر دائما إلى الجانب المشرق من
الأشياء!

انتهينا لهذا اليوم، وداعا!