Cvičení č. 5.: Provedení faktorové analýzy

Příklad: Výsledky desetiboje z olympiády v Aténách 2004 (Příklad je převzat z knihy Meloun M., Militký J., Hill, M.: Počítačová analýza vícerozměrných dat v příkladech. Academia Praha 2005)
V datovém souboru Desetiboj.sta jsou uloženy výsledky 39 závodníků - mužů, kteří se v roce 2004 zúčastnili desetiboje na olympiádě v Aténách. Zajímají nás výsledky jednotlivých disciplín, tj. proměnné v14 – v23. Budeme se snažit najít menší počet společných faktorů, které vysvětlují variabilitu výsledků závodníků v desetiboji. Přitom budeme uvažovat jenom závodníky, kteří desetiboj dokončili, tj. v proměnné Dokončil je 1.

Řešení v systému STATISTICA:

Nejprve upravíme datový soubor: ponecháme jen ty případy, kdy v proměnné Dokončil je 1 a ponecháme jen proměnné v14 – v23. Upravený datový soubor má tedy 14 proměnných a 28 případů.
Sestavení korelační matice:
Statistiky – Vícerozměrné průzkumné techniky – Faktorová analýza - Proměnné v1 – v10 – OK – OK. Na záložce Popisné statistiky zvolíme Přehled korelací, průměrů, směrodatných odchylek – Korelace

	Korelace (Desetiboj_upraveny.sta) ChD vynechána případově N=28									
Proměnná	Body 100 m	Body skok dálka	Body koule	Body výška	Body 400 m	Body překážky	Body disk	Body tyčka	Body oštěp	Body 1500 m
Body 100 m	1,00	0,71	0,37	0,31	0,63	0,54	0,24	0,26	0,01	0,06
Body skok dálka	0,71	1,00	0,20	0,35	0,67	0,54	0,26	0,28	0,10	0,14
Body koule	0,37	0,20	1,00	0,61	0,21	0,24	0,67	0,03	0,38	-0,13
Body výška	0,31	0,35	0,61	1,00	0,18	0,33	0,52	-0,05	0,21	0,00
Body 400 m	0,63	0,67	0,21	0,18	1,00	0,52	0,16	0,11	0,05	0,54
Body překážky	0,54	0,54	0,24	0,33	0,52	1,00	0,22	0,15	0,08	0,17
Body disk	0,24	0,26	0,67	0,52	0,16	0,22	1,00	-0,18	0,26	-0,22
Body tyčka	0,26	0,28	0,03	-0,05	0,11	0,15	-0,18	1,00	-0,07	-0,19
Body oštěp	0,01	0,10	0,38	0,21	0,05	0,08	0,26	-0,07	1,00	0,25
Body 1500 m	0,06	0,14	-0,13	0,00	0,54	0,17	-0,22	-0,19	0,25	1,00

Některé korelace mezi proměnnými jsou dostatečně vysoké, zřejmě tedy má smysl provádět faktorovou analýzu. Ověříme to pomocí Bartlettova testu sféricity a poté vypočteme Gleasonovu – Staelinovu míru redundance.

Provedení Bartlettova testu:
Logaritmus determinantu výběrové korelační matice získáme v systému STATISTICA takto:
Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné v1 až v10 – OK – OK – Popis. statistiky – Korelační matice Inverzní.

	Inverzní korelační matice (Desetiboj_upraveny.sta) Aktivní proměnné Log(Determinant) korelační matice: -4,9450272									
Proměnná	Body 100 m	Body skok dálka	Body koule	Body výška	Body 400 m	Body překážky	Body disk	Body tyčka	Body oštěp	Body 1500 m
Body 100 m	2,88140	-1,23972	-0,92682	0,10609	-1,00731	-0,407553	0,51392	0,032088	0,23872	0,56349
Body skok dálka	-1,23972	3,55909	1,55223	-1,04009	-1,94474	-0,169294	-0,49625	-0,512693	-0,73880	0,85323
Body koule	-0,92682	1,55223	3,38735	-1,36637	-1,15064	0,156820	-1,21138	-0,382013	-1,03722	0,81110
Body výška	0,10609	-1,04009	-1,36637	2,16403	1,11393	-0,310977	-0,31959	0,156377	0,42814	-0,75409
Body 400 m	-1,00731	-1,94474	-1,15064	1,11393	4,82175	-0,451652	-0,64970	-0,170380	1,07073	-2,82431
Body překážky	-0,40755	-0,16929	0,15682	-0,31098	-0,45165	1,643324	-0,08547	-0,075124	-0,06954	0,02237
Body disk	0,51392	-0,49625	-1,21138	-0,31959	-0,64970	-0,085469	2,51275	0,741222	-0,22313	1,00010
Body tyčka	0,03209	-0,51269	-0,38201	0,15638	-0,17038	-0,075124	0,74122	1,424669	-0,05487	0,57177
Body oštěp	0,23872	-0,73880	-1,03722	0,42814	1,07073	-0,069540	-0,22313	-0,054874	1,65819	-1,09353
Body 1500 m	0,56349	0,85323	0,81110	-0,75409	-2,82431	0,022374	1,00010	0,571771	-1,09353	3,09704

V záhlaví výstupní tabulky je číslo $\ln|\mathbf{R}|$ = -4,9450272.

Otevřeme nový datový soubor o 3 proměnných a 1 případu.

Do Dlouhého jména 1. proměnné napíšeme =-4,9450272 (tj. $\ln|\mathbf{R}|$), do Dlouhého jména druhé proměnné napíšeme =(-137/6)*v1 (tj. $\chi^2 = \frac{11+2p-6n}{6}\ln|\mathbf{R}|$) a Dlouhého jména třetí proměnné napíšeme =VCHi2(0,95;45) (tj. kvantil $\chi^2_{0,95}(45)$).

	1 log(det(R))	2 test. stat.	3 kvantil
1	-4,9450272	112,911454	61,6562334

Testová statistika se realizuje v kritickém oboru, hypotézu o úplné nezávislosti sledovaných 10 proměnných tedy zamítáme na hladině významnosti 0,05.

Výpočet Gleasonovy – Staelinovy míry redundance:
K výstupní tabulce, v níž je uložena korelační matice, přidáme novou proměnnou, která bude obsahovat součty kvadrátů korelačních koeficientů. Do jejího Dlouhého jména napíšeme:

=v1^2+v2^2+v3^2+v4^2+v5^2+v6^2+v7^2+v8^2+v9^2+v10^2

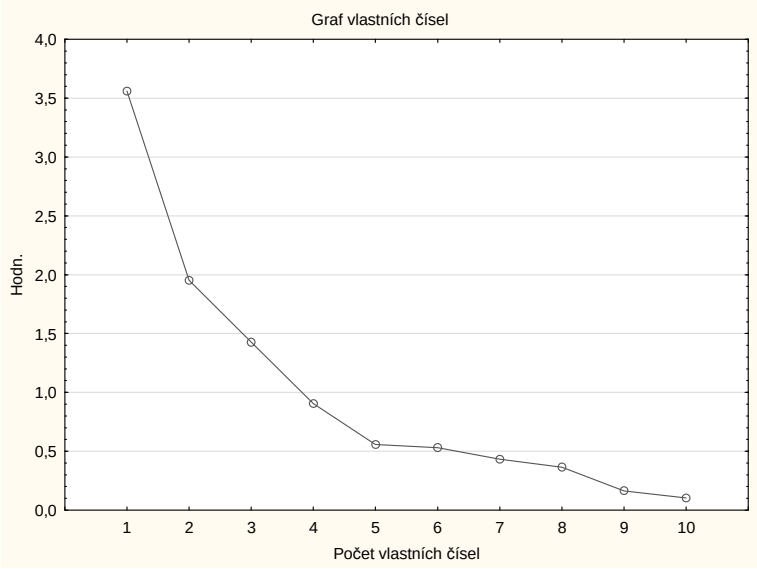
Pomocí Statistiky – Blok sloupců – Součty získáme součet této proměnné. Přidáme další proměnnou a do jejího Dlouhého jména napíšeme: =sqrt((v1-10)/90)

Proměnná	Korelace (Desetiboj_upraveny.sta) ChD vynechána případově N=28 1 NProm	 2 NProm
SOUČET případy 1-10	20,2901091	0,33813392

Vidíme, že koeficient Φ = 0,3381 nabývá dostatečně velké hodnoty pro prokázání korelace v datech.

Vypočteme vlastní čísla výběrové korelační matice, zjistíme procento vysvětleného rozptylu a nakreslíme sutinový graf.
Na záložce Základní nastavení změníme Max. počet faktorů na 10 a Min. vlastní číslo na 0 – OK – na záložce Výklad rozptylu zvolíme Vlastní čísla a poté Sutinový graf.

Hodn.	Vl. čísla (Desetiboj_upraveny.sta) Extrakce: Hlavní komponenty vl. číslo	% celk. rozptylu	Kumulativ. vlast. číslo	Kumulativ. %
1	3,559212	35,59212	3,55921	35,5921
2	1,952914	19,52914	5,51213	55,1213
3	1,426585	14,26585	6,93871	69,3871
4	0,905343	9,05343	7,84405	78,4405
5	0,558752	5,58752	8,40281	84,0281
6	0,531569	5,31569	8,93438	89,3438
7	0,432804	4,32804	9,36718	93,6718
8	0,365741	3,65741	9,73292	97,3292
9	0,164634	1,64634	9,89756	98,9756
10	0,102445	1,02445	10,00000	100,0000



Zkusíme pracovat se čtyřmi faktory., které vysvětlují asi 78 % variability obsažené v datech. Zlom v sutinovém grafu je sice až u 5 faktorů, ale to už je příliš velký počet.

Spočteme komunality pro první čtyři faktory. Na záložce Základní nastavení zadáme Max. počet faktorů 4 – OK. Na záložce Zákl. výsledky zvolíme Rotace faktorů Varimax prostý. Na záložce Výklad rozptylu zvolíme Komunality.

	Komunality (Desetiboj_upraveny.sta) Extrakce: Hlavní komponenty Rotace: Varimax pr.				
Proměnná	Z 1 faktoru	Z 2 faktorů	Z 3 faktorů	Z 4 faktorů	Více R^2
Body 100 m	0,607176	0,688337	0,754538	0,765488	0,652946
Body skok dálka	0,657291	0,701525	0,762419	0,762445	0,719030
Body koule	0,018782	0,721677	0,728789	0,801075	0,704784
Body výška	0,069272	0,610093	0,617912	0,628732	0,537900
Body 400 m	0,820559	0,820841	0,829374	0,848187	0,792606
Body překážky	0,515887	0,565462	0,570450	0,570688	0,391477
Body disk	0,010633	0,754881	0,776694	0,777293	0,602030
Body tyčka	0,030249	0,045444	0,898216	0,898593	0,298083
Body oštěp	0,001017	0,072752	0,074469	0,900852	0,396933
Body 1500 m	0,226628	0,398965	0,607299	0,890702	0,677111

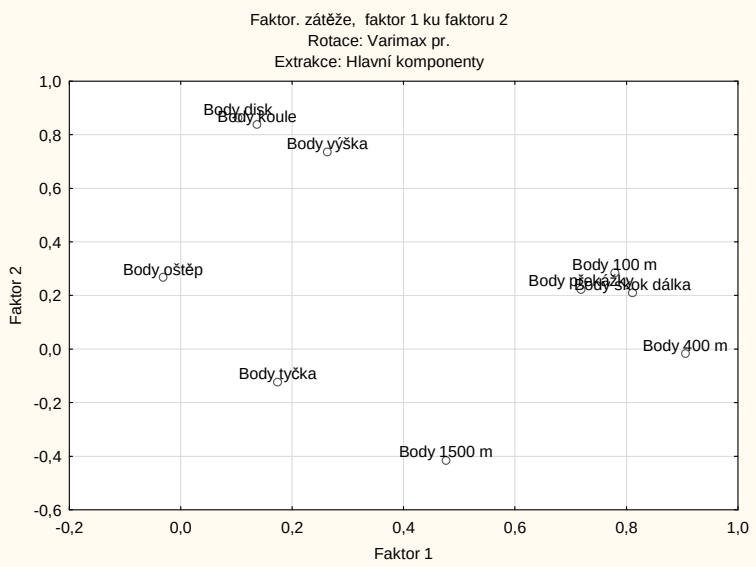
Vidíme, že např. variabilita proměnné Body na 100 m je ze 76,5 % vysvětlena prvními čtyřmi faktory.

Nyní získáme odhad matice rotovaných faktorových zátěží: na záložce Zátěže zvolíme Shrnutí: Faktorové zátěže.

Proměnná	Faktor. zátěže (Varimax pr.) (Desetiboj_upraveny.sta) Extrakce: Hlavní komponenty (Označené zatěže jsou >,700000) Faktor 1	Faktor 2	Faktor 3	Faktor 4
Body 100 m	0,779215	0,284888	0,257295	-0,104642
Body skok dálka	0,810735	0,210319	0,246768	0,005051
Body koule	0,137047	0,838389	0,084332	0,268859
Body výška	0,263195	0,735405	-0,088424	0,104022
Body 400 m	0,905847	-0,016782	-0,092373	0,137160
Body překážky	0,718253	0,222654	0,070625	-0,015438
Body disk	0,103118	0,862698	-0,147691	0,024477
Body tyčka	0,173922	-0,123269	0,923456	0,019420
Body oštěp	-0,031883	0,267834	0,041442	0,909056
Body 1500 m	0,476054	-0,415136	-0,456436	0,532356
Výkl.roz	2,957493	2,422484	1,240182	1,223894
Prp.celk	0,295749	0,242248	0,124018	0,122389

První faktor vysoce koreluje s výsledky krátkých běhů a skoku do dálky. Lze ho označit jako rychlost. Druhý faktor koreluje s výsledky hodu koulí, disku a skoku do výšky. Je možné ho interpretovat jako schopnost zkoncentrovat výbušnou energii do jediného okamžiku. Třetí faktor koreluje s výsledkem skoku o tyči. Vzhledem k vysokému korelačnímu koeficientu ho lze ztotožnit s touto proměnnou. To samé platí o čtvrtém faktoru, který vysoce koreluje s výsledkem hodu oštěpem. Proměnné Body oštěp a Body tyčka jsou tedy unikátní a bez výraznějšího vztahu ke znakům ostatním proměnným.

Faktorovou strukturu můžeme též znázornit graficky v prostoru faktorových zátěží. Vytvoří se shluky jednotlivých proměnných, přičemž každý shluk reprezentuje takovou skupinu disciplín, kterou lze vysvětlit působením stejného faktoru.
Na záložce Zátěže zvolíme Graf zátěží, 2D.



Kvalitu získaného faktorového modelu posoudíme též pomocí odhadnuté korelační a reziduální korelační matice. Na záložce Výklad rozptylu vybereme Reprod./rezid. korelace.

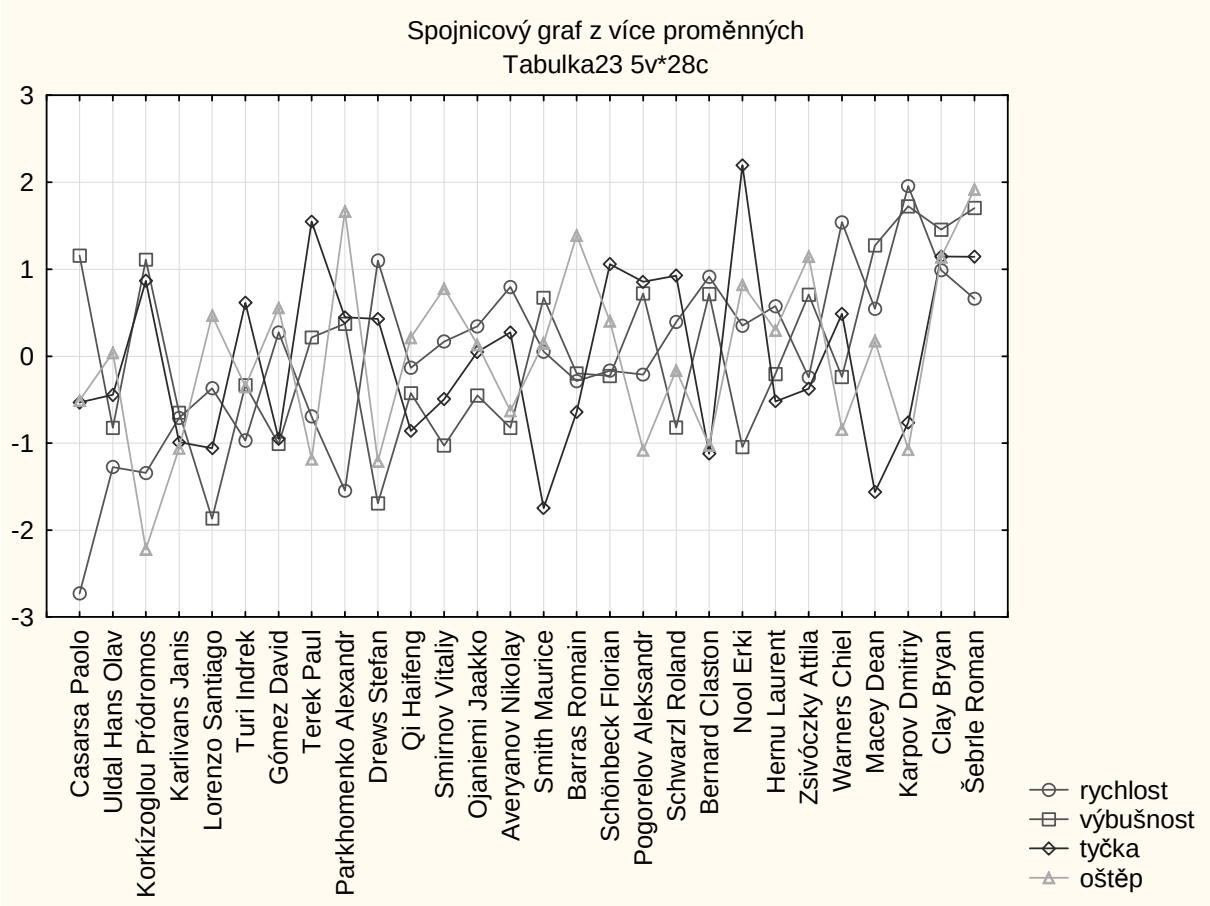
Proměnná	Reprodukované korelace (Desetiboj_upraveny.sta) Extrakce: Hlavní komponenty Body 100 m	Body skok dálka	Body koule	Body výška	Body 400 m	Body překážky	Body disk	Body tyčka	Body oštěp	Body 1500 m
Body 100 m	0,77	0,75	0,34	0,38	0,66	0,64	0,29	0,34	-0,03	0,08
Body skok dálka	0,75	0,76	0,31	0,35	0,71	0,65	0,23	0,34	0,05	0,19
Body koule	0,34	0,31	0,80	0,67	0,14	0,29	0,73	0,00	0,47	-0,18
Body výška	0,38	0,35	0,67	0,63	0,25	0,34	0,68	-0,12	0,28	-0,08
Body 400 m	0,66	0,71	0,14	0,25	0,85	0,64	0,10	0,08	0,09	0,55
Body překážky	0,64	0,65	0,29	0,34	0,64	0,57	0,26	0,16	0,03	0,21
Body disk	0,29	0,23	0,73	0,68	0,10	0,26	0,78	-0,22	0,24	-0,23
Body tyčka	0,34	0,34	0,00	-0,12	0,08	0,16	-0,22	0,90	0,02	-0,28
Body oštěp	-0,03	0,05	0,47	0,28	0,09	0,03	0,24	0,02	0,90	0,34
Body 1500 m	0,08	0,19	-0,18	-0,08	0,55	0,21	-0,23	-0,28	0,34	0,89

Proměnná	Reziduální korelace (Desetiboj_upraveny.sta) Extrakce: Hlavní komponenty (Označená rezidua jsou > ,100000) Body 100 m	Body skok dálka	Body koule	Body výška	Body 400 m	Body překážky	Body disk	Body tyčka	Body oštěp	Body 1500 m
Body 100 m	0,23	-0,05	0,03	-0,07	-0,03	-0,10	-0,05	-0,08	0,05	-0,02
Body skok dálka	-0,05	0,24	-0,11	0,00	-0,04	-0,11	0,03	-0,06	0,06	-0,05
Body koule	0,03	-0,11	0,20	-0,06	0,07	-0,04	-0,06	0,02	-0,08	0,04
Body výška	-0,07	0,00	-0,06	0,37	-0,07	-0,02	-0,16	0,07	-0,07	0,09
Body 400 m	-0,03	-0,04	0,07	-0,07	0,15	-0,12	0,06	0,04	-0,04	-0,01
Body překážky	-0,10	-0,11	-0,04	-0,02	-0,12	0,43	-0,03	-0,01	0,05	-0,04
Body disk	-0,05	0,03	-0,06	-0,16	0,06	-0,03	0,22	0,04	0,01	0,01
Body tyčka	-0,08	-0,06	0,02	0,07	0,04	-0,01	0,04	0,10	-0,08	0,09
Body oštěp	0,05	0,06	-0,08	-0,07	-0,04	0,05	0,01	-0,08	0,10	-0,09
Body 1500 m	-0,02	-0,05	0,04	0,09	-0,01	-0,04	0,01	0,09	-0,09	0,11

K datovému souboru přidáme proměnnou Body celkem. Do jejího Dlouhého jména napíšeme =sum(v1:v10)
Nyní uložíme faktorová skóre. Na záložce Skóre vybereme Uložit faktorová skóre. Uložíme je společně s proměnnou Body celkem. Faktor 1 pak přejmenujeme na rychlost, faktor 2 na výbušnost, faktor 3 na tyčka a faktor 4 na oštěp. Závodníky ještě seřadíme podle dosaženého počtu bodů.

	Faktor. skóre pro 4 1 Body celkem	2 rychlost	3 výbušnost	4 tyčka	5 oštěp
Casarsa Paolo	7404	-2,72916	1,15765	-0,53227	-0,50841
Uldal Hans Olav	7495	-1,27386	-0,82299	-0,44448	0,04313
Korkízoglou Pródromos	7573	-1,34222	1,11152	0,86851	-2,22034
Karlivans Janis	7583	-0,70899	-0,64998	-0,98991	-1,05812
Lorenzo Santiago	7592	-0,36827	-1,86765	-1,05755	0,47114
Turi Indrek	7708	-0,97085	-0,33088	0,61418	-0,34882
Gómez David	7865	0,27301	-1,00958	-0,95076	0,55930
Terek Paul	7893	-0,69327	0,21509	1,54865	-1,18183
Parkhomenko Alexandr	7918	-1,54632	0,37369	0,44675	1,66907
Drews Stefan	7926	1,09819	-1,69251	0,43011	-1,20874
Qi Haifeng	7934	-0,13300	-0,42508	-0,85979	0,21414
Smirnov Vitaliy	7993	0,17004	-1,02645	-0,49304	0,77931
Ojaniemi Jaakko	8006	0,34403	-0,45223	0,05056	0,13424
Averyanov Nikolay	8021	0,79884	-0,82238	0,27567	-0,62499
Smith Maurice	8023	0,05262	0,67125	-1,74580	0,15250
Barras Romain	8067	-0,28532	-0,19668	-0,64335	1,38976
Schönbeck Florian	8077	-0,16632	-0,22737	1,06005	0,40362
Pogorelov Aleksandr	8084	-0,21135	0,72048	0,85437	-1,07971
Schwarzl Roland	8102	0,39334	-0,81860	0,92735	-0,16115
Bernard Claston	8225	0,91525	0,71572	-1,11867	-1,02351
Nool Erki	8235	0,35064	-1,04533	2,19641	0,82584
Hernu Laurent	8237	0,57642	-0,20285	-0,51829	0,29890
Zsivóczky Attila	8287	-0,24175	0,70955	-0,37434	1,14955
Warners Chiel	8343	1,54313	-0,23780	0,48710	-0,84026
Macey Dean	8414	0,54618	1,27155	-1,56035	0,17603
Karpov Dmitriy	8725	1,95784	1,72416	-0,76334	-1,07247
Clay Bryan	8820	0,98805	1,45495	1,14749	1,14111
Šebrle Roman	8893	0,66309	1,70275	1,14475	1,92073

Nyní sestrojíme spojnicový graf faktorových skóre.
Grafy – 2D Grafy – Spojnicové grafy (Proměnné) – Proměnné rychlost – oštěp – OK, zapneme Vícenásobný – OK



Na první pohled zde nedominuje žádný z faktorů. Znamená to, že k vítězství je potřeba souhra všech. Co se týká jednotlivých závodníků, vidíme např., že Roman Šebrle má jedny z nejlepších skóre u všech faktorů, proto také vyhrál na těchto OH.

Podívejme se ještě, jak se změní výsledky, když změníme metodu extrakce faktorů a metodu rotace. Na záložce Detaily zvolíme Centroidovou metodu a na záložce Základní výsledky vybereme Varimax normalizovaný.
Vlastní čísla a procento vysvětleného rozptylu:

	Vl. čísla (Desetiboj_upraveny.sta) Extrakce: Hlavní faktory (Centroid)			
Hodn.	vl. číslo	% celk. rozptylu	Kumulativ. vlast. číslo	Kumulativ. %
1	3,186417	31,86417	3,186417	31,86417
2	1,646201	16,46201	4,832618	48,32618
3	1,261125	12,61125	6,093743	60,93743
4	0,412872	4,12872	6,506616	65,06616

Poněkud pokleslo procento vysvětleného rozptylu, z 78 % na 65 %.

Faktorové zátěže:

Proměnná	Faktor. zátěže (Varimax normaliz.) (Desetiboj_upraveny.sta) Extrakce: Hlavní faktory (Centroid) (Označené zatěže jsou >,700000) Faktor 1	Faktor 2	Faktor 3	Faktor 4
Body 100 m	0,801931	0,169119	0,081081	0,189384
Body skok dálka	0,816232	0,159124	-0,039806	0,129519
Body koule	0,145893	0,944284	0,026339	0,095966
Body výška	0,322112	0,590217	0,107111	-0,103545
Body 400 m	0,807053	0,037249	-0,356080	0,021918
Body překážky	0,636636	0,175576	-0,049498	0,030171
Body disk	0,193462	0,704059	0,192066	-0,262833
Body tyčka	0,218152	-0,096104	0,149190	0,664741
Body oštěp	-0,008349	0,418982	-0,238998	0,011212
Body 1500 m	0,201183	-0,076158	-0,993030	-0,201507
Výkl.roz	2,616572	2,012445	1,251943	0,625656
Prp.celk	0,261657	0,201244	0,125194	0,062566

Na rozdíl od metody hlavních komponent koreluje třetí faktor s proměnnou Body 1500 m, lze ho tedy interpretovat jako vytrvalost.