Ljubljana, 2012

Priporočila za urejanje podatkovne datoteke

Verzija 1.1

Arhiv družboslovnih podatkov,
UL, Fakulteta za družbene vede,
Kardeljeva ploščad 5,
1000 Ljubljana
tel: +386 01 5805 292,
e-mail: arhiv.podatkov@fdv.uni-lj.si,
http://www.adp.fdv.uni-lj.si/

Pripravile: Sanja Lužar, Maja Ojsteršek in Irena Vipavc Brvar

KAZALO

Podana priporočila in primeri so nekoliko bolj usmerjeni na uporabnike statističnega programa SPSS, vendar se lahko ustrezno uporabljajo tudi za druge statistične programe.

1. Poimenovanje spremenljivk

Dolžina imen spremenljivk v SPSS-u (Name) naj **ne presega 8 znakov** (pri tem so všteti vsi znaki). V imenih **ne uporabljajte presledkov ali pik**. Za ločnico se lahko uporabi podčrtaj. **Prvi znak naj bo črka**.

Kljub temu, da novejše verzije omogočajo daljše imenovenje, uporabnik, ki bo do datoteke dostopal s starejšo verzijo SPSS ali drugim programom, datoteke ne bo mogel smiselno uporabljati.

Spremenljivke naj bodo smiselno poimenovane. Zaradi lažje prepoznave spremenljivk in vprašanj v vprašalniku je najpogostejša raba zapisa spremenljivk kar oznaka za vprašanje (npr. V - vprašanje ali Q – question) in zaporedna številka vprašanja ter po potrebi dodatek za večkratni odgovor.

Npr. V1, V2, V3, V4.1, V4.2, V4.3, V5, V6… Pri čemer je V1 vprašanje, V4.1 pa podvprašanje vprašanja V4, ali eden izmed več možnih odgovorov vprašanja V4.

Primer 1: Označevanje po vprašanjih

4.00 MEDNARODNE OPERACIJE IN MISIJE

4.01 Koliko po vašem mnenju Slovenija pripomore k preprečevanju konfliktov po svetu?

1 - sploh ne pripomore
2 - ne pripomore
3 - niti pripomore, niti ne pripomore
4 - pripomore
5 - zelo pripomore
9 - ne vem, b.o.

Vir: SJM092

Takšno označevanje se lahko uporablja le pri vprašanjih, kjer je možen samo en odgovor.

Vprašanje 4.01 v vprašalniku predstavlja spremenljivka V4_01 v podatkovni datoteki. Vsi možni odgovori v vprašalniku (1, 2, 3, 4, 5, 9) predstavljajo vrednosti spremenljivke V4_01.

Primer 2: Označevanje po vprašanjih in odgovorih

15. ALI MENIŠ, DA JE V NAŠI DRUŽBI BOLJE POSKRBLJENO ZA MLADE ALI ZA STAREJŠE?

A: Obkroži ENO številko pred izbranim odgovorom.

1. Mnogo bolje je poskrbljeno za starejše
2. Nekoliko bolje je poskrbljeno za starejše
3. Približno enako dobro je poskrbljeno za mlade kot za starejše
4. Nekoliko bolje je poskrbljeno za mlade
5. Mnogo bolje je poskrbljeno za mlade

A: Ne beri! *99 – ne vem, b.o.*

Vir: MLA10

Ta način označevanja spremenljivk se običajno uporablja pri vprašanjih, kjer je možnih več odgovorov.

Spremenljivka V15_1 je predstavlja odgovor 1. v vprašalniku, spremenljivka V15_2 odgovor 2, itd. Vsaka od teh spremenljivka ima lahklo le dve vrednosti: *je izbral* (1) ali *ni izbral (0)*.

2. Opis spremenljivk

Vsaka spremenljivka naj vsebuje tudi opis spremenljivke (Label). Zaradi zahtev dolgotrajne hrambe so dolžine opisov spremenljivk (Label) **omejene na 60 znakov**.

3. Vrednosti spremenljivk

Za vsako spremenljivko naj bodo **določene in poimenovane** tudi njene vrednosti (Values), razen tam kjer je to razvidno že iz samega podatka (npr. pri letnici rojstva, dohodku). Opis je lahko tako numeričen kot alfanumeričen (npr. pri kodah držav).

3.1. Vrednosti spremenljivk – odprti odgovori

Odprte odgovore smiselno pretvorimo v nove vrednosti, glede na obstoječe šifrante. Kadar gre za odprti tekst, ki ga ne želimo pretvoriti, predlagamo, da se besedilo nekoliko očisti nepotrebnih znakov in tipkarskih napak.

V primeru, da so besedila odprtih odgovorov daljša, se jih v datoteki podatkov nadomesti z zaporednimi številkami, besedilo odgovorov pa se ločeno hrani in distribuira v **šifrantu** (v katerem so odgovori pod ustreznimi zaporednimi številkami). V opis spremenljivke v podatkovni datoteki (Label) se doda opomba o tem, kje so odgovori dostopni (npr. za odgovore glej dokument xx).

4. Manjkajoče vrednosti

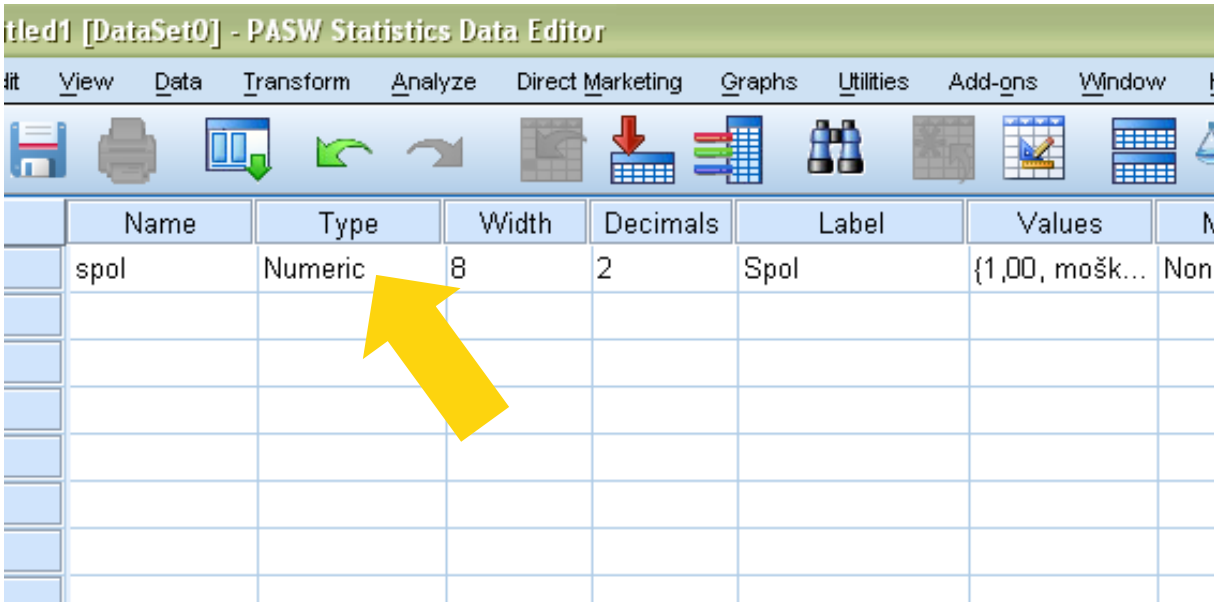
Pazimo na dosledno dokumentiranje manjkajočih vrednosti spremenljivk. Kadar kode niso podane v vprašalniku, priporočamo **uporabo standardnih oznak** za manjkajoče vrednosti, npr.:

Vrednost	Ime vrednosti
99	brez odgovora
98	ne vem
97	se ne nanaša, preskok
95	napaka
-1	zakrita vrednost

5. Tip spremenljivk in merska lestvica

Razmislimo o tem katere spremenljivke je smiselno ohraniti kot opisne in katere raje spremenimo v številske z ustreznimi opisi.

Npr. spremenljivko spol (vrednosti M in Ž) je lažje uporabiti v analizah kadar je označena kot številska »numeric« in ne opisna »string«. Torej bi imela vrednosti odgovorov 1 in 2 z ustreznimi opisi (1=Moški, 2=Ženski).
Pazimo na ustrezno oznako v podatkovni datoteki.



V SPSS-u lahko določimo tudi mersko lestvico spremenljivke. Lahko izbiramo med imensko (nominalno) spremenljivko, pri kateri v vrednostih ni nobene urejenosti, urejenostno (ordinalno), pri kateri lahko vrednosti uredimo po nekem redu »od manj do več« in številsko, pri čemer SPSS ne loči med razmično in razmernostno obliko.

Korektna oznaka naših spremenljivk je pomembna pri prenosu v druge programe, ki omogočajo pregled rezultatov. Če bo spremenljivka npr. imena »imensko« oznako programi ne bodo ponudili računanja povprečja.

6. Preoblikovane spremenljivke

Poleg spremenljivk vezanih neposredno na vprašalnik bodo v datoteki verjetno tudi dodatne spremenljivke. Po končanem postopku čiščenja in ustreznega imenovanja razmislimo o tem katere spremenljivke hraniti. V ADP priporočamo, da se hranita dve različni datoteki – ena daljša, neanonimizirana, z identifikatorji, ki bi jih mogoče nekoč uporabili pri ponovitvi raziskave, in ena krajša – torej tista, ki bo vključevala le relevantne spremenljivke.

Kadar so v datoteki podatkov preoblikovane ali izvedene spremenljivke (npr. iz spremenljivke *letnica rojstva* kreirana spremenljivka *starostni razredi*), moramo le-te ustrezno dokumentirat (zapisati kako smo spremenljivko izračunali). To je zlasti pomembno pri pripravi indeksov. Tako bo uporabnik poznal spremenljivko, ki jo je v svojih analizah uporabljal raziskovalec.

Zaradi lažjega razumevanja in uporabe spremenljivk, je priporočljivo, da je v primeru osnovnih pretvorb ime nove spremenljivke sestavljeno iz imena primarne spremenljivka, ki mu dodamo »_r« (recode). Tako bi iz npr. spremenljivke starost (letnica rojstva) kreirali spremenljivko starost_r (starostni razredi). Ker v tem primeru z dodajanjem »_r« pridobimo v imenu še dva dodatna znaka in tako skupno ime presega 8 znakov, ga smiselno skrajšamo v npr. star_r.

7. Spremenljivke uteži

Spremenljivke, ki predstavljajo uteži, je potrebno ustrezno dokumentirat v opisu spremenljivke (npr. utež po spolu). Kadar je potrebna daljša obrazložitev, jo vpišemo v dokumentacijo za podatkovno datoteko oziroma v obrazec za opis raziskave, ki ga izpolnimo pri predaji podatkov v ADP. Lahko gre tudi za samostojen dokument, ki ga pripnemo opisu raziskave.

8. Manjkajoči odgovori pri vseh enotah spremenljivke

Pogosto se pri pripravi spremenljivk za večkratne možnosti odgovora zgodi, da zadnjih nekaj spremenljivk nima vrednosti. V tem primeru lahko tako spremenljivko izločimo iz datoteke.

Sicer velja pravilo, da so vse spremenljivke iz vprašalnika vključene v podatkovno datoteko. Kakršno koli spremembo je potrebno ustrezno dokumentirati.

9. Katerih spremenljivk ne hranimo?

Iz datoteke izločimo spremenljivke, za katere menimo, da jih uporabnik ne bo uporabil. To so npr. spremenljivke generirane zaradi definicije programa anketiranja (npr. pri CATI ali CAPI anketah).

Dodatno izločimo vse spremenljivke, ki niso ustrezno dokumentirane.

10. Anonimiziranje podatkov

Iz datoteke podatkov, ki jo boste posredovali v hrambo ADP primarno izločite neposredne identifikatorje kot so imena, naslovi, telefonske številke.

Dodatno je potrebno v datoteki, ki bo dana v javno distribucijo, zakriti posredne »pokazatelje« oseb (npr. če imamo spremenljivko ime podjetja in drugo spremenljivko z vlogo v podjetju, lahko hitro ugotovimo kateri odgovori pripadajo direktorju). Lahko se celo odločimo za pripravo več distribucijskih datotek na tak način, da bomo vključili kombinacijo spremenljivk, in tako ne bomo neposredno razkrili respondenta, še vedno pa bo takšna datoteka relevantna za analizo.

Glede na obliko spremenljivke jo lahko anonimiziramo v celoti ali delno. Npr. pri spremenljivkah kot je občina respondeta uporabimo popolno anonimizacijo, kar pomeni, da vse vrednosti spremenljivke zamenjamo z vrednostjo -1 in opisom spremenljivke »zakrita vrednost«. Kadar obstaja možnost za prepoznavanje dela skupine pa lahko uporabimo delno anonimizacijo. Zakrijemo samo vrednosti z nizko stopnjo odgovora. Kadar ima npr. določena kategorija odgovora vrednosti manjše od 30, se odločimo za zakritje te vrednosti in druge najmanjše. Če bomo zakrili samo eno vrednost se bo iz skupne vsote to vrednost dalo izračunati. Vedno torej zakrijemo dve ali več vrednosti odgovora.

V ADP hranimo tri vrste datoteke: originalno datoteko, ki smo jo prejeli od dajalca, datoteko za distribucijo v znanstveno-raziskovalne namene (SUF - Scientific use files) in javno dostopno datoteko (PUF - Public use files). Zadnja je seveda najbolj anonimizirana. Izločene so tudi kategorije odgovorov pri krajevnih identifikatorjih (kot sta Občina, Kraj, Poštna številka).

Ker ADP lahko ustrezno omeji dostop do datotek za različne uporabnike svetujemo, da v primerih, ko niste prepričani ali je spremenljivko potrebno anonimizirati ali ne, le-to raje pustite v datoteki in ADP nanjo opozorite.

Pomembno je, da ne zakrijemo preveč vrednosti saj lahko tako izločimo morebitne pomembne informacije za raziskovalca, po drugi strani pa tudi, da ne zakrijemo premalo vrednosti in tako ne zagotovimo primerne stopnje anonimnosti. Na tem mestu svetujemo, da se držite pravila manj je več in se posvetujte z ADP. Z našimi

izkušnjami in znanjem bomo hitro videli katere vrednosti je potrebno zakriti in katere ne.

11. Končna priporočila

Ko je datoteka podatkov pripravljena je priporočljivo, da na njej izračunamo osnovne statistike in še enkrat preverimo, da:

- so imena spremenljivk omejena na 8 znakov
- imajo vse spremenljivke opise (do 60 znakov)
- so dodane manjkajoče vrednosti
- so spremenljivke smiselnega tipa
- nimamo »čudnih« vrednosti, kategorij, ki jih ni v vprašalniku
- odprti odgovori ne vsebujejo nepotrebnih znakov (////, **…)
- so vrednosti spremenljivk primerno anonimizirane

12. Vprašanja? Kontakti…

- **Spletna stran:** www.adp.fdv.uni-lj.si
- **E-mail:** arhiv.podatkov@fdv.uni-lj.si
- **Facebook:** Arhiv družboslovnih podatkov
- **Twitter:** @ArhivPodatkov

13. Nadaljnje reference

Seznam drugih priporočil za pripravo podatkov in ostalih gradiv raziskave:

- Managing data (CESSDA – Council of European Social Science Data Archives)
- Create and Manage data (UK Data archive)
- Guide for Social Science Data Preparation and Archiving (ICPSR – Inter-university Consortium for Political and Social Research)
- Tools and guidelines (IHNS – International Household Survey Network)

Priporočila so namenjena urejanju podatkovnih datotek, tako za raziskovalce kot tudi za vse, ki bi jim nasveti kakorkoli koristili.
Pridržujemo si pravico do morebitnih napak.