MTA SJTB STAB ajánlás a reprezentatív minta kifejezés használatáról és a mintavételből származó adatok jellemzéséről:

Véges sokaságok vizsgálatakor cél és elvárás a minta reprezentativitása[1]. Azonban a gyakorlatban nincs egyetértés a reprezentativitás pontos definíciójáról, mérhetőségéről, mérési módszereiről.

A STAB ajánlása:

Tudományos munkákban a reprezentativitás fogalmának használatakor adjuk meg annak pontos értelmezését.

Ne használjuk a reprezentatív minta kifejezést a mintavételből származó adatok jellemzésére, anélkül, hogy pontosan leírnánk, miként biztosítottuk[2], hogyan ellenőriztük azt.

Általánosan ajánlott (tudományos célú kutatásoknál követelmény) a mintavételből származó adatok jellemzése, ennek általánosan ajánlható részei:

- *az adatfelvételi folyamat lényeges elemeinek (célsokaság, mintavételi keret, mintaválasztás módja, minta mérete, adatgyűjtés módja, eszközei), feldolgozási módszereinek (pl. adatellenőrzés, outlierkezelés, imputálás, újrasúlyozás) ismertetése;*
- *az adatfelvétel során bekövetkezett hibák (pl. lefedettségi hiba, nemválaszolás, mintavételi hiba) azonosítása, mértékük jellemzése;*
- *ezek értékelése a felvétel célja szempontjából, továbbá figyelmeztetés a felhasználóknak, ha indokolt (pl. valamilyen aggregálási szint alatti becslések bizonytalansága, fogalmak eltérő értelmezése).*

A minősítésnek objektív mutatókon kell alapulnia. Ha az adatállomány minősége nem teszi lehetővé, hogy a sokaságra következtessünk, úgy az eredményekből adódó állításainkat csak a mintába került elemekre fogalmazzuk meg.

Olyan tartalmú, részletezettségű leírást adjunk, amiből a felhasználó meg tudja ítélni az eredmények adott célra való használhatóságát

[1] A reprezentatívitás mint cél arra utal, hogy a sokaság bizonyos egységei (a minta) azzal a céllal kerülnek kiválasztásra, hogy a sokaság egységeit (sokaság egészét) megfelelő módon képviseljék. Így a minta elemzéséből a sokaság jellemzőire következtethetünk. Véletlen mintavételi módszerek választása esetén van tudományos alapja a mintajellemzőkből a sokaságra való következtetésnek, a mintavételi hiba becslésének, ilyenkor a kiválasztott véletlen minta reprezentálja a sokaságot (mintavételi keretet). Ez az alapja pl. annak a nézetnek, mely csak a véletlen módszerrel kijelölt mintát tekinti reprezentatívnak.

[2] Ugyanis a reprezentativitásra (bármely értelmezésben) nem csak a minta választásakor, hanem az adatfelvételi folyamat minden lépésében oda kell figyelnünk, pl. a keretválasztáskor is.

INDOKLÁS

Statisztikai adatfelvétel (survey) egésze

Egy véges sokaság vizsgálatának hagyományos eszköze egy mintavételes statisztikai adatfelvétel, azaz a sokaság egészének jellemzése az adott sokaság egy részére – mintára - kiterjedő adatfelvétel alapján. A mintából a sokaságra való következtetés megalapozottságának megítéléséhez a kijelölt mintaelemek megfigyeléséből származó adatállomány és a célsokaság jellemzőinek összevetésére lenne szükség a vizsgált jellemzők szempontjából, ami – értelemszerűen - közvetlenül nem lehetséges. Az eredmények megbízhatóságáról a statisztikai adatfelvételi folyamat alapján szerezhetünk információkat.

A statisztikai adatfelvételi folyamat – bár statisztikai célú - igen összetett, tervezése és végrehajtása során a statisztikai mellett etikai, jogi, pszichológiai, társadalom- és gazdaságtudományi szempontokra is tekintettel kell lennünk. A statisztikai adatfelvételek különböző alkalmazási területeire átfogó, standard elvárásnak tekintett irányelvek, ajánlások jöttek létre, ilyenek pl. ISO 20252, American Assiciation for Public Opinion Research (AAPOR) vagy az European Society for Opinion and Marketing Research (ESOMAR) ajánlásai és az európai hivatalos statisztika és a KSH módszertani anyagai, melyek listája az anyag végén található.

Az adatfelvételi folyamat részei nem függetlenek egymástól, egymáshoz illeszkedve, csak együtt biztosíthatják a megbízható eredményt. A mintavétel, a begyűjtött adatok kezelése is része az adatfelvételi folyamatnak, az ajánlás indoklásaként ezzel foglalkozunk bővebben.

A mintavétel, feldolgozás és kapcsolódó folyamatszakaszok

A következőkben a véges sokaságból történő mintavételes adatfelvételek megfelelő végrehajtásához igyekszünk támpontokat adni. A terület meglehetősen széles, tehát a jelen dokumentum csak a vonatkozó irodalommal együtt adhat elegendően pontos eligazítást. Egy mintás adatfelvétel végső soron akkor bizonyul hasznosnak, ha segítségével megalapozott következtetéseket vonhatunk le a vizsgált sokaságra vonatkozóan. A minta használhatóságát tehát érdemes a becslés minősége szempontjából megítélni, hiszen egyes köztes lépések (pl. mintavétel) az adatok felhasználói számára önmagukban aligha bírnak relevanciával. Egy rossz mintavételi keret vagy a válaszhiány szakszerűtlen kezelése egy hatékony minta mellett is használhatatlanná teheti a becsléseket, ezért a célunk eléréséhez a mintavételi eljáráson túl a teljes adat-előállítási folyamat minőségét kell biztosítanunk. Akkor biztosan sikerrel jártunk, ha a folyamat végén előálló adatbázis változóinak pontosságát megalapozottan tudjuk a matematikai-statisztikai eszköztár segítségével jellemezni. Elfogadott feltevések mellett a becslések torzítása nem jelentős és számíthatóak a mintavételi hibák, akkor az adat-előállítási folyamatot megfelelően hajtottuk végre. Nem feltétlenül követeljük meg tehát a végső adatbázisunk változóira a becslések pontosságát (csupán a pontosság becsülhetőségét), hiszen a mintavételes technikához immanens módon hozzátartozik a véletlen hiba lehetősége, különösen kis minták esetén.

Természetesen például egy szakdolgozat-írónak vagy egy önálló kutatónak egészen mások az anyagi, időbeli, felkészültségbeli lehetőségei, mint a statisztikák publikálására szakosodott

hivataloknak, közvélemény-kutatóknak. Az utóbbiaktól joggal várható el, hogy az adat-előállítási folyamat minden lépésénél tudományos igényességű megoldásokra törekedjenek. Az előbbiek viszont gondolkodhatnak úgy, hogy nem kívánnak állításokat megfogalmazni és publikálni a teljes célsokaságra vonatkozóan, megelégszenek csak a minta jellemzésével. Például a szakdolgozókkal szemben sokszor csak az az elvárás, hogy valamilyen elemzési eszköztár korrekt használatát képesek legyenek bemutatni. Ez nem valószínűségi mintán is megvalósítható. A minta, a mintából származó adatok – ajánlás szerinti - jellemzésére ezekben az esetekben is szükség van, illetve a magyarázatra, hogy miért döntöttek a nem valószínűségi minta mellett. Azonban, ha a teljes sokaságra levont következtetéseiket publikálásra szánják, akkor velük szemben is jogos elvárás, hogy minden adat-előállítási lépést elfogadható minőségben hajtsanak végre.

Követendő lépések[3]

Célsokaság

Az adatfelvételek első lépése a *célsokaság* definiálása: meg kell határoznunk, hogy mi az a véges sokaság, amelyre vonatkozóan következtetéseket akarunk levonni.

Az adatfelvétel tervezése előtt gyűjtsünk információt a célsokaságról. Egyrészt lehet, hogy már létezik olyan állomány (pl. más adatfelvételből származó mikroadat-állomány, adminisztratív adat, big data), amiből a szükséges információ kinyerhető és így nem lesz szükség adatfelvételre; ha mégis, az összegyűjtött információ segíthet a felvételi folyamat tervezésében (pl. mintavételi keret, módszer választásában, rétegezési lehetőség), végrehajtásában.

Mintavételi keret

Ezután állítjuk elő a *mintavételi keretet*. Ennek során többnyire keresünk egy olyan nyilvántartást[4], mely a lehető legnagyobb mértékben lefedi a célsokaságot. Bizonyos esetekben – például összeírási nehézségek fennállásakor – tervezhetünk hiányos lefedettségű keretet is, azonban ilyenkor vizsgálni kell a lefedettségi hiány becslésekre gyakorolt várható hatását. Amennyiben egy részsokaságot nem fed le a keret, és arra vonatkozóan elérhető másik keret, úgy megfontolandó a keretek együttes használata. A mintavételi keretben kell, hogy legyen használható információ az egyedek elérhetőségéről. A mintavételi keret fontos minőségi jellemzője még a pontosság: a duplikátumok, a célsokaságba nem tartozó (out-of-scope) elemek, hibás információk számát minimálisra kell szorítani. A keretben lévő információk legyenek továbbá naprakészek a felvétel vonatkozási idejéhez viszonyítva. A keret alkalmazhatóságával szorosan összefügg, hogy az egységek azonosításán túl rendelkezésre állnak-e olyan információk, melyek hatékonyabbá tehetik a mintavételi tervet vagy a súlyozást. Nem adható számszerű útmutatás arra nézve, hogy a fenti jellemzőknek milyen mértékben kell teljesülnie. Azt javasoljuk, hogy minden esetben készítsünk leírást, amelyben – ha lehetséges, akkor számszerűen – **jellemezzük a mintavételi keretet a**z összes fenti tulajdonság - **lefedettség, pontosság, időszerűség** - **szempontjából**. Ez a felvétel

[3] A lépések nem feltétlenül ebben a sorrendben következnek a gyakorlatban.

[4] A teljesség kedvéért megjegyezzük, hogy cenzus, minta és térkép is szolgálhat mintavételi keretként.

végrehajtójának is segít eldönteni, hogy érdemes-e továbblépni a minta tervezése felé, és az adatok felhasználóit is orientálja, hogy a saját elemzési céljaikhoz kielégítő-e a felvétel.

Mintavételi terv, kiválasztás, mintaelemszám

A mintavételi tervezés[5] során meghatározzuk a minta nagyságát és kiválasztásának módját. **Tervezzünk valószínűségi mintát, amennyiben a feltételek biztosíthatóak**, mivel csak ebben az esetben tudjuk a becslések pontosságát a matematikai statisztikai eszköztár segítségével jellemezni. A mintaelemek cseréje (pótcímezés) nem megengedett.

Ha csak nem valószínűségi mintát (pl. kvótás minta, véletlen séta, toborzás, hógolyó módszer, kényelmi minta stb.) **tudunk kivitelezni, akkor feltétlenül ki kell fejtenünk, hogy miért számítunk arra, hogy a vizsgálat eredményeit ez érdemben nem torzítja.** Az ilyen sejtéseket esetleg alátámaszthatja valamilyen szimuláció vagy megalapozott feltevés a sokaság jellemzőiről. Ha ilyesmit nem tudunk prezentálni, akkor a későbbiekben csak a minta elemeire, nem pedig a teljes célsokaságra vonatkozóan vonhatunk le következtetéseket.

A minta jellemzőinek tervezése során figyelemmel kell lennünk az esetleges pontossági elvárásokra, az adatgyűjtés szempontjaira, a felvételi keretben elérhető információkra, a várható nemválaszolásra, a kerethibákra és a költségvetési korlátokra. E szempontok sokszor elkerülhetetlenül rontják a hatékonyságot (többlépcsős minta, tervezett lefedettségi hiány, kis mintanagyság stb.), azonban a mintavételi rétegzés, allokáció és kiválasztás mindegyikénél lehet optimalitásra törekedni. Ennek mikéntjéről lásd például Cochran (1977), Hunyadi (2001) vagy Särndal et al. (1992) műveit. A minta jó minősége azt jelenti, hogy a fenti szempontok figyelembevétele után a felvétel legfontosabbnak ítélt változóinak várható mintavételi hibája (és torzítása) minimális. Sokcélú felvételnél ugyanakkor meg kell fontolni, hogy a legfontosabb változókra optimális minta a többi változóra nem hat-e hátrányosan.

Folyamatos/longitudinális felvételeknél érdemes időben stabil hatékonyságú mintát tervezni. Az ilyen felvételeknél többnyire meg kell határoznunk a rotációs sémát is, mely akkor megfelelő, ha eleget tesz a változások becslésének pontosságára vonatkozó elvárásoknak, figyelembe véve az adatszolgáltatók terhelhetőségét és a várható lemorzsolódást.

Minden esetben **tegyük megismerhetővé a mintavételi terv leírását**. A leírás tartalmazza a mintavétel minden fontos jellemzőjét (elsődleges és másodlagos mintavételi egységek, rétegek kialakítása, kiválasztás módja, allokáció, minta mérete stb.). A terv alapján megvalósítható kell, hogy legyen a mintakiválasztás folyamata. A dokumentumban célszerű megindokolni, hogy az adott tervezési megoldást milyen cél vagy korlát teljesülése érdekében alkalmazzuk. A minta kiválasztásának természetesen pontosan követnie kell a mintavételi tervet.

Adatgyűjtés

Az adatgyűjtés célja, hogy a mintába került elemektől teljes körűen (vagy a lehető legnagyobb arányban) begyűjtse a kívánt adatokat. Az adatgyűjtés eszközei, módja, folyamata lényeges

[5] A különböző – ebben a részben is említett – mintavételi eljárásokat itt nem ismertetjük, azok a hivatkozott szakkönyvekben megtalálhatóak.

része az adatfelvételi folyamatnak, befolyásolja az eredmények érvényességét, a mintába kerültek válaszadási hajlandóságát. Az adatgyűjtés jellemzése során ki kell térni az adatgyűjtés mérőeszközére, módjára, idejére, az adatgyűjtésben résztvevők számára, az adatgyűjtés előrehaladására, monitoringjára és az esetlegesen tapasztalt váratlan eseményekre, amelyek hatással lehetnek az adatok minőségére, és az információ segíthet az esetleges korrekcióban (pl. nemválaszolás nagysága, tapasztalt jellemzői, kerethiba, számítógépes rendszerek hiányossága az adatgyűjtés közben, egyes területeken tapasztalható összeírási nehézségek stb.).

Az adatfelvételek elkerülhetetlen velejárója a nemválaszolás[6], ami a becslésekben súlyos torzítást okozhat, ezért a nemválaszolás korrekt jellemzése és a felmerülő hibák kezelése elengedhetetlen. **A mintás adatfelvételek dokumentációjához szervesen hozzátartozik a válaszolási arány (vagy ennek komplementere, a nemválaszolási arány) számítása.** A mutató számlálójában a becsléshez használható kérdőívek száma, a nevezőjében pedig a kiválasztott minta célsokasághoz tartozó elemeinek száma található (Statistics Canada, 2001, 11.o.). A számos rossz tapasztalatra tekintettel külön felhívjuk a figyelmet arra, hogy a nevezőben mindig az eredetileg kiválasztott minta nagysága szerepel.

Editálás

Az adatfelvételek során gyakran előfordul, hogy egy-egy érték hibásan kerül rögzítésre. Az editálás folyamán meg kell kísérelnünk ezek helyreállítását. Célszerű rekordonként megkeresni az inkonzisztens vagy kiugró értékeket, és azokat konzekvens módon javítani. A nagy súlyú rekordoknál és a szisztematikusan előforduló hibák kijavításának nagyobb jelentősége van. Automatizált hibakereséssel és javítási szabályok alkalmazásával elkerülhetjük az önkényes javításokat, noha az egyedi mérlegelést sokszor nem tudjuk teljesen kiküszöbölni. **Törekedjünk a kézenfekvőbb hibák** (elütések, számok nagyságrendjének/előjelének/mértékegységének tévesztése, rossz kódolás) **korrekciójára**, elkerülve ezzel a túlzott, az adatok valódiságának önkényes megítélésén alapuló editálást. Az editált adatokat mindig meg kell jelölni, az eredeti értékeket el kell tárolni.

Súlyozás, kalibrálás

A mintából történő becslések előkészítéséhez a legtöbb esetben a mintát súlyozni kell. A mintavételi terv alapján előállítható az egyedek mintába kerülési valószínűsége, ezek reciproka a design súly.

Az egység szintű meghiúsulást (unit nonresponse – a kérdőív egésze hiányzik[7]), általában súlyozással kezeljük. Az átsúlyozás a mintavétel során előálló design súlyok jellemzően többszöri módosítását jelenti, melyek eredményeképpen becslésre alkalmas súlyok jönnek létre. A végső becslő súlyok figyelembe veszik az eltérő kiválasztási és megvalósulási arányokat, a lefedettségi hiányt, és ismert sokasági paraméterekhez (sarokszámokhoz) vannak illesztve (kalibrálás).

[6] Az egybeírás nem véletlen: az angol „nonresponse” kifejezés tükörfordítását tekintsük szakkifejezésnek.
[7] vagy nem elfogadható minősége miatt kerül törlésre

Asúlyozás első lépése a megvalósult minta design súlyainak módosítása a lehetséges torzítás mértékének csökkentése érdekében. A súlyozáshoz használjunk olyan megbízható segédváltozókat, amik összefüggenek a megvalósulással és a vizsgálat célváltozóival egyaránt. A megvalósulás valószínűségét becsülhetjük homogén válaszadói csoportok módszerével, logisztikus regresszióval. Ezen súlyozás eredményét jó minőségű kivitelezés esetén akár végső becslősúlyként is használhatjuk. Azonban ezt a lépést a gyakorlatban sokszor nem, vagy nem kielégítően tudjuk végrehajtani, mivel a kereteinkben általában nem áll rendelkezésre kellő számú és minőségű meghiúsulás-kompenzáláshoz használható változó.[8] Ilyenkor fokozott jelentősége van az ezt követő kalibrálásnak, melynek során úgy változtatjuk a súlyokat, hogy a mintából a végső súlyokkal számított paraméterek[9] közel essenek az ún. sarokszámokhoz. A kalibrálást a lehető legkisebb torzítást és szórást eredményező, azaz a szakirodalom (pl. Deville et al., 1992) által elismert módszerek valamelyikével kell végrehajtani (pl. raking ratio, GREG). Emellett figyelnünk kell a sarokszámok helyes megválasztására is. A sarokszámok rendszerint olyan sokasági értékek vagy becslések, melyek forrását sokkal megbízhatóbbnak tekintjük a kérdéses mintánál (pl. adott nemű, korú, iskolai végzettségű személyek száma a népszámlálás vagy valamely regiszter szerint). Olyan sarokszámokat válasszunk, melyek tartalmilag pontosan egyeznek a kalibráló változókkal, a lehető legközelebbi a vonatkozási idejük, a forrásuk pontos, valamint minél erősebben összefüggenek a nemválaszolás tényével és a felvétel által vizsgált változókkal. A gyakorlatban az alkalmas sarokszám-források szűkössége eléggé behatárolja a lehetőségeket. A súlyok terjedelmét célszerű abszolút vagy relatív korlátok közé szorítani.

A súlyok megfelelőségét a súlyozás fentiek szerinti végrehajtása biztosítja. A súlyok ne szóródjanak extrém módon, ne legyen közöttük negatív érték. Legyen kielégítő a sarokszámokhoz való illeszkedés. A becslősúlyok értéke jó esetben nem kerül nagyon távol az elsődleges súlytól. A súlyok eloszlása lehetőleg ne a szélsőértékek környékén sűrűsödjön.

A fenti feltételek figyelembevétele mellett is előfordulhat, hogy a kalibráló változók között nem szereplő változó eloszlása távol esik a valóságtól, különösen kisebb elemszámú (rész)minták és olyan változók esetén, amik gyengén korrelálnak a kalibráló változókkal. A kalibrálásba nem beépített, de a felvétel szempontjából fontosnak ítélt változók eloszlását a fentiek miatt szükséges alaposabban vizsgálni[10], az eredményeket nem kevésbé megbízható felvételek, nyilvántartások adataival összehasonlítva validálni.

Készítsünk a felhasználók számára elérhető dokumentációt arról, hogy milyen módszert alkalmaztunk a nemválaszolás kezelésére és a kalibrálás során, mik voltak és honnan származtak a felhasznált segédváltozók, és mi lett a súlyok ellenőrzésének az eredménye.

[8] Az is előfordul, hogy valamilyen okból dizájn súlyokat sem számítunk és meghiúsulás-kompenzálást sem végzünk, hanem bizonyos tulajdonságokkal rendelkező csoportok sokasági és mintabeli létszámának aránya adja az elsődleges súlyokat a kalibráláshoz. Ez a megoldás elfogadható, de nem ajánlott.

[9] A paraméter az esetek többségében valamilyen csoport létszámát (pl. 15-30 éves nők száma), ritkább esetben valamilyen értékösszeget (pl. a munkajövedelmek összértéke) jelent.

[10] Vizsgáljuk meg, hogy egy-egy változó kalibrált becslése hihető-e, illetve nem vált-e „hihetetlenné" az elsődleges súlyokkal készített becsléshez képest. Próbaként érdemes olyan súlyozásokat is készíteni, amelyekben nem használjuk fel az összes sarokszámot, mivel ebben az esetben tudjuk ellenőrizni, hogy a kalibrálásból kihagyott változóra készített becslés és a fel nem használt sarokszám nem kerültek-e túlságosan messze egymástól.

Imputálás

A tétel szintű meghiúsulást (item nonresponse – csak 1-1 kérdésre hiányzik a válasz) többnyire imputálással korrigáljuk. Itt az a különös helyzet áll fenn, hogy a legelterjedtebb módszerek (átlaggal imputálás, törlés, az előző időszaki megfigyelés továbbvitele, determinisztikus regresszió stb.) valójában csak erős feltételek fennállása esetén használhatók. A gyakorlatban előforduló adatbázisok imputálásakor inkább csak néhány szofisztikáltabb módszertől (többszörös imputálás, egyes hot deck módszerek, modell alapú módszerek, sztochasztikus regresszió) várhatjuk, hogy nem okoznak jelentős torzítást a becslésekben. Mindezekről jó leírást találhatunk például Enders (2010) vagy Molenberghs et al. (2015) műveiben. Ha mégis a kevésbé cizellált módszerek mellett döntünk, akkor indokoljuk meg, hogy miért állhatnak fent a használatuk feltételei (pl. teljesen véletlenszerűen hiányzó adatok stb.). **Emellett írjuk le azt is, hogy az egyes változók hiányzó értékeit milyen más változók és modellek használatával imputáltuk.** Az imputálással bevitt értékeket szükséges megjelölni az adatbázisban.

Outlier-azonosítás és kezelés

Erősen ferde eloszlású célváltozók esetén az átlagostól jelentősen eltérő kiugró értékek (outlier) jelenléte a mintában alapvetően befolyásolhatja a mintából származó becslések pontosságát, különösen kisebb elemszámú (rész)minták mellett. Akár egyetlen kiugró érték is óriási növekedést okozhat a varianciában, ami szükségessé teszi az outlier azonosítását és kezelését. Érdemes több outlier-azonosítási technikát szimultán alkalmazni, mivel egyetlen módszer nem feltétlenül azonosítja az összes kiugró értéket. Az outlierek megfelelő beazonosításához elengedhetetlen a célváltozó által megfigyelt jelenség, folyamat kellő szakmai ismerete. A kiugró értékek súlyának módosításával vagy robusztus becslőfüggvények alkalmazásával csökkenthető a becslések átlagos négyzetes hibája. A kiugró érték azonosításában és annak kezelésében szubjektív módon meghatározott paraméterek jelenhetnek meg, emiatt különös óvatosság indokolt, elkerülendő, hogy az outlier kezelés egy előzetes szakértői hipotézis alátámasztásának eszköze legyen. Az eredményekre gyakorolt lényeges hatása miatt az alkalmazott módszerek átláthatósága különös jelentőségű.

A torzítás kockázata

Az adatfelvétel egyes folyamatszakaszaiban kifejtett minden erőfeszítés ellenére a becsléseket torzítás terhelheti. Emiatt különösen fontos a torzítás kockázatát jellemezni. A válaszadási arány a leggyakrabban használt és értékes mértéke ennek: a magas válaszadási arány növeli annak valószínűségét, hogy a megvalósult minta megfelelően reprezentálja a célsokaságot, míg az alacsony válaszadási arány a jelentősebb torzítás valós kockázatát jelezheti. A válaszadási arány önmagában azonban nem jellemzi kellőképpen a torzítás kockázatát. Külső információk megfelelő bevonásával megbízhatóbb kép alakítható ki. Ez a többi között magába foglalhatja a mintából származó becslések külső forrásból származó adatokkal történő összehasonlítását, a könnyen és nehezen elért válaszadók összehasonlítását, a súlyozáshoz használt változók és a célváltozó közötti korreláció mérését, alternatív súlyozással kapott eredmények vizsgálatát, a lehetséges torzítás terjedelmének vizsgálatát. Élő kutatások témája az ún. reprezentativitás-indikátorok számításának módja, ami egy mutatóval jellemzi a minta megbízhatóságát.

Becslés, mintavételi hiba

A mintából valamely sokasági paraméter becsléséhez becslőfüggvényt kell választanunk. Ennek lehetőségei és a választás szempontjai megismerhetők az irodalomból. A becslőfüggvény választása is hatással van a becslés pontosságára.

A mintából számított becslés értéke mintavételi ingadozásnak van kitéve, mintáról mintára változhat. Ennek mértéke a becsült jellemző (mint valószínűségi változó várható értéke körüli) szórásától függ, véletlen minta esetén becsülhetjük. A mintavételi hiba leggyakrabban használt mutatói a standard hiba és a relatív hiba (CV). Véletlen minta esetén mindenképpen számítsuk ki legalább a főbb mutatókra. **A mintavételi hiba fontos információ a felhasználónak, a dokumentáció fontos része.**

Kivonat a kulcsfontosságú döntésekről

(1) Mintavételi keret

- A mintavételi keret legyen friss, naprakész, tartalmazzon jó minőségű információkat a mintavételi tervezéshez, adatgyűjtéshez és súlyozáshoz.
- A mintavételi keret lefedettségi hiánya és annak hatása a becslésekre legyen minimális. A hiány legyen vizsgálat tárgya.
- Leírásban: célsokaság és keret, illetve információ a kettő eltéréséről, lefedettségi többlet és hiány

(2) Mintavételi terv, kiválasztás

- A célsokaság minden elemének (a lefedettségi hiánytól eltekintve, lásd 1. pont) legyen ismert és pozitív mintába kerülési valószínűsége. Ennek megfelelően a nem valószínűségi technikák (pl. kvótás, véletlen sétás minta) alkalmazása nem megengedett. A mintaelemek cseréje nem megengedett.
- A mintavételi technikák megfelelő alkalmazásával (pl. rétegzés, kiválasztás módja, minta allokáció) javítsuk a minta hatékonyságát.
- Leírásban: mintavételi terv főbb jellemzői.

(3) Mintaelemszám

- A megvalósult minta elemszáma legyen kellően nagy ahhoz, hogy kielégítse az elemzési célok igényeit. A mintaelemszám tervezésénél vegyük figyelembe a mintavételi terv hatékonyságát és a várható meghiúsulást.
- Leírásban: kiválasztott és megvalósult mintaelemszámok, kiválasztási arányok, egység szintű nemválaszolási arányok, fontosabb változókhoz tételszintű nemválaszolási arányok.

(4) Adatgyűjtés

- Az adatgyűjtés folyamata megtervezett, monitorozott, dokumentált.
- Mérőeszközök teszteltek, a kívánt információt gyűjtik.

- Minél teljesebb körű adatgyűjtés a mintaelemektől.
- Leírásban: adatgyűjtés módja, eszközei.

(5) Editálás, outlier kezelés

- Az editálást jellemezze a kézenfekvőbb hibák korrekciója, elkerülve a túlzott editálást, az adatok valódiságának önkényes megítélését.
- Erősen ferde eloszlású célváltozó esetén a kiugró értékek súlyának módosításával vagy robusztus becslőfüggvények alkalmazásával csökkenthető a becslések átlagos négyzetes hibája. Az eredményekre gyakorolt lényeges hatása miatt az alkalmazott módszerek átláthatósága különös jelentőségű.
- Leírásban: kritériumok, módszerek, editált értékek aránya (változónként), outlier azonosítás és kezelés módszere, azonosított és kezelt egységek/változók száma.

(6) Súlyozás, kalibrálás

- A végső becslő súlyok figyelembe veszik az eltérő kiválasztási és megvalósulási arányokat, a lefedettségi hiányt, és ismert sokasági elemszámokhoz vannak illesztve (kalibrálás).
- A súlyozás a megvalósult minta design súlyainak módosítása a lehetséges torzítás mértékének csökkentése érdekében. A súlyozáshoz használjunk olyan megbízható segédváltozókat, amik összefüggenek a megvalósulással és a vizsgálat célváltozóival egyaránt.
- A kalibráláshoz használt sarokszámok minősége legyen jobb az adott felvételénél. A sarokszámok feleljenek meg a felvétel célsokaságának.
- Leírásban: súlyozás és újrasúlyozás módszerei, kalibrálás esetén sarokszámok.

(7) Imputálás

- A legelterjedtebb módszerek a pótlásra jellemzően erős feltételek fennállása esetén használhatók. Javasolt kifinomultabb eljárások alkalmazása (sztochasztikus regresszió, többszörös imputálás).
- Leírásban: Imputált egységek, változóként imputált értékek száma, aránya, imputálás módszere.

Hivatkozások

Cochran, William G. (1977): Sampling Techniques. John Wiley & Sons, Inc.

Deville, Jean-Claude – Särndal, Carl-Erik (1992): Calibration Estimators in Survey Sampling. http://www.jstor.org/stable/2290268, 2019.10.14-én.

Enders, Craig K. (2010): Applied Missing Data Analysis. The Guilford Press, New York.

Hunyadi László (2001): A mintavétel alapjai. BKÁE egyetemi jegyzet. Számalk Kiadó, Budapest.

Molenberghs, Geert – Fritzmaurice, Garrett – Kenward, Michael G. – Tsiatis, Anastasios – Verbeke, Geert (2015): Handbook of Missing Data Methodology. CRC Press.

Särndal, Carl-Erik – Swensson, Bengt – Wretman, Jan (1992): Model Assisted Survey Sampling. Springer-Verlag.

Statistics Canada (2001): Standards and Guidelines for Reporting Nonresponse Rates. Statistics Canada Technical Report.

Átfogó szakmai ajánlások különböző célú statisztikai adatfelvételekhez

ISO 20252:2019 Market, opinion and social research, including insights and data analytics – Vocabulary and service requirements

https://www.iso.org/obp/ui/#iso:std:73671:en

(Fentinek csak a bevezető és fogalmi része olvasható, többihez csak fizetés ellenében férhetnénk hozzá. Jó proxy a 2018-as szavazásra bocsájtott változat: https://www.amsrs.com.au/documents/item/2481)

AAPOR (American_Association_for_Public_Opinion_Research) 2015: Code of Ethics

https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx

ICC/ESOMAR (2016): International Code on Market, Opinion and Social Research and Data Analytics

(ICC: International Chamber of Commerce; ESOMAR: European Society for Opinion and Marketing Research)

https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf

Az európai hivatalos statisztikára:

Eurostat (2008): Survey sampling reference guidelines. Introduction to sample design and estimation techniques. Risto Lehtonen and Karl Djerf. Cat. No. KS-RA-08-003-en

https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-RA-08-003

Eurostat 2014: ESS Handbook for Quality Reports

https://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf

Eurostat (2013): Handbook on precision requirements and variance estimation for ESS households surveys (2013 Edition) http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-RA-13-029

Gazdaságstatisztikai felvételeknél a kiemelt témák mindegyikével foglalkozik az Memobust e-kézikönyv: Handbook on Methodology of Modern Business Statistics (Eurostat 2017)

https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en

Eurostat: European statistics Code of Practice - revised edition 2017

https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142

Központi Statisztikai Hivatal, Budapest:

KSH 2014: Minőségi irányelvek a Központi Statisztikai Hivatal statisztikai folyamataira

http://www.ksh.hu/docs/bemutatkozas/hun/minosegi_iranyelvek_2014.pdf