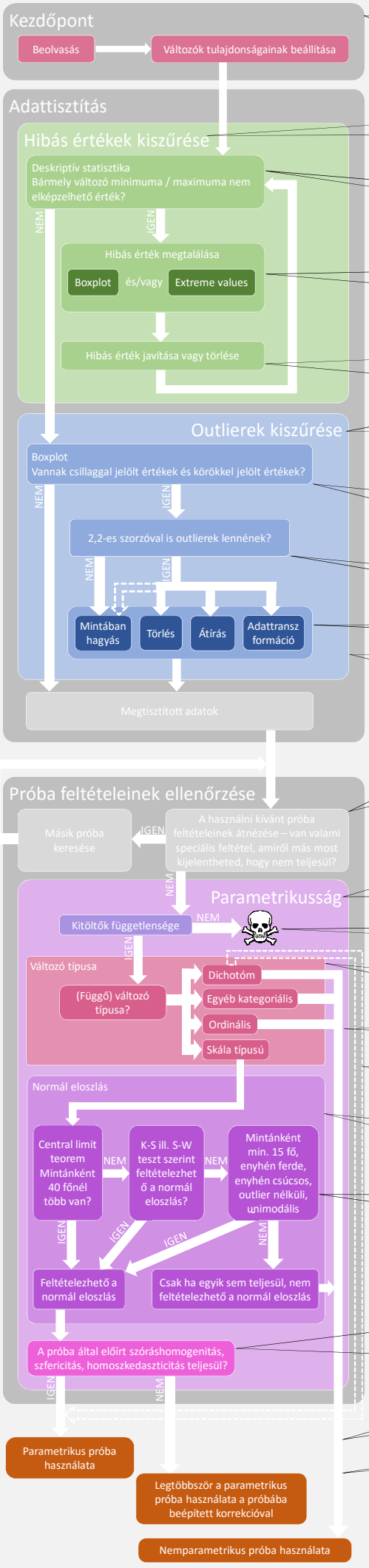

Ismertek a kérdések, az adatok, és van elképzelés arról, milyen próbát akarsz végezni.
Adatok: Érdemes lekérni leíró statisztikákat, hogy átfogó képet kapjunk az adatainkról. Hány fő van a mintáinkban, vannak-e hiányzó értékek, átlagok, szórások, stb. Ezeket egy részét úgyis közölni kell a minta leírása során.
Próbák: Egy jó kutatásban, már a kísérletet is úgy tervezzük meg, hogy figyelembe vesszük, a bejövő adatokon milyen elemzéseket akarunk majd végezni, de legkésőbb itt, mielőtt még az adatokon bármit is dolgoztunk volna, döntenünk kell a statisztikákról, hiszen a próbáknak különböző feltételei vannak, különböző kérdésekhez különböző minták tartoznak, ezért elemzéstől függően különböző ellenőrzéseket kell végeznünk.

A hibás értékek olyan adatok, melyek az adott skálán nem fordulhatnak elő, valamilyen hiba (például elgépelés) miatt kerültek csak az adatbázisunkba. Példa: nemi adatoknál 3-as érték, vagy 0-tól 50-ig terjedő skálán 61-es érték, vagy kor változóban 132 év. Ha hibás érték van az adataink között, azzal mindenképpen kezdeni kell valamit.

Hogy van-e hibás érték, legkönnyebben egy leíró statisztikával ellenőrizhetjük. Kérjük ki a minimumot és a maximumot, és ha ezek minden változó esetén valamilyen elképzelhető értékek (például nem minimuma 1 és maximuma 2), akkor mehetünk tovább.

Ha felfedeztünk egy hibás értéket, az javítanunk kell. Egy nagyobb adatbázisban viszont nem könnyű megtalálni ezeket. Két módszert használhatunk, a Boxplotot és a Extreme values, ami az Explore statisztika Outliers opciójával kérhető ki (ez utóbbi elnevezés félrevezető, hiszen nem feltétlenül outliereket sorol fel, hanem az öt legnagyobb és öt legkisebb értéket). Mind a két megjelenítési formánál látjuk majd a hibás érték sorszámát. Nem kell mind a kettőt kikérni, mikor egyik, mikor másik hasznosabb.

Ha megtaláltunk egy hibás értéket, azzal mindenképpen kezdeni kell valamit, nem szabad egyszerűen a mintában hagyni. Ha szerencsénk van, utána tudunk nézni, mi lehetett a jó érték (például papíron kitöltött kérdőívnél kikereshetjük), és ki tudjuk javítani azt a helyes adatra. Ha nincs szerencsénk, és nem tudunk ennek utánanézni, akkor törölni kell. Itt nem opció, hogy az értéket átírjuk valami másra. Sőt tippelni sem érdemes, mert több kárt teszünk, ha tévedünk, mint amennyi hasznot hajtunk azzal, hogy „megmentünk" egy adatot. Ha végeztünk, kérjünk ki még egy leíró statisztikát, hogy ellenőrizzük, minden hibás értéket eltávolítottunk.

Az outlier olyan érték, mely az adott skálán elképzelhető, de annyira szélsőséges, hogy torzítaná a statisztikáinkat (eltolná az átlagot, megnövelné a szórást stb.). Ilyen lehet például egy egyetemista mintában egy 35 éves kísérleti személy – létezik ilyen, de a többségtől jóval idősebb. Itt az ő szélsőséges értékének van információtartalma, ezért jó lenne megtartani, mint valaki, akinek magas a kora a mintában, de valahogy meg kellene szüntetni a szélsőségességének torzító hatását.

Mennyire kell szélsőségesnek lenni egy értéknek ahhoz, hogy outliernek tekintsük. Több definíció is létezik, mi az outlier labelling rule szerint dolgozunk (másik ismert szabály az átlag +/- 2 vagy 3 szórás). Az outlier labelling rule szerint outlier az, aki a középső ötven százalék másfélszeresénél távolabbra van az alsó és felső negyedelő pontoktól – ezeket a Boxplot körökkel jelöli. A Boxplot jelöl értékeket csillaggal is, ezek az értékek a középső ötven százalék háromszorosánál vannak távolabb a quartilisektől.

Bár az SPSS másfélszeres szorzóval dolgozik, helyesebb lenne 2,2-eddel számolni, és azokat outliernek tekinteni, akik a középső ötven százalék 2,2-szeresénél vannak távolabb a quartilisektől.

Ha megtaláltuk az outliereket, el kell döntenünk mit csinálunk velük, amire több megoldás is létezik.

Bár az outlierek szűrését általában egységesen, az összes adaton közösen, az adatfeldolgozás elején szoktuk elvégezni, egyes hipotézisek, melyekhez a teljes adatbázisnak csak egy részét használjuk mintának , szükségessé tehetik, hogy később az adott almintán is ellenőrizzük az outliereket. Például kiszűrtük a 35 éves kísérleti személyt az egyetemista mintából, de ha az egyik hipotézishez csak az elsősöket használjuk, akkor lehet, hogy már a 28 éves is outlier lenne, és ki kellene szűrni.

Általában törekszünk arra, hogy parametrikus próbát használjunk, hiszen legtöbbször ezek a legérzékenyebb elemzések, ezért a parametrikusság feltételei külön kiemelve szerepelnek a tananyagban. De a parametrikusságon kívül minden próbának vannak egyéni feltételei, melyeknek teljesülni kell ahhoz, hogy használhassuk őket. Ilyen lehet például a minimális elemszámra, a minták függetlenségére vagy összefüggésére, vagy a változók közötti kapcsolat meglétére, hiányára, erősségére, linearitásra vonatkozó feltételek. Mivel a feltételek próbánként eltérőek, ezért ezen a diagramon nem szerepel részletesen az ellenőrzésük, érdemes azonban még a parametrikusság feltételeinek ellenőrzése előtt áttekinteni őket.

A parametrikusságnak négy feltétele van: Függetlenség, legalább intervallum skála típusú adatok, normál eloszlás, szóráshomogenitás. Ellenőrzésük során érdemes ebben a sorrendben haladni, hiszen ha nem teljesül a függetlenség, akkor az egy egészen speciális eset, és speciális próbák használatát kívánja meg. Ha az adatok nem skála típusúak, akkor biztosan nem használhatunk parametrikus próbát, és egyébként bizosan nem teljesül/nem értelmezhető a normalitás és szóráshomogenitás sem (például nem(!) ellenőrizzük egy dichotóm változó normalitását. Végül a szóráshomogenitás tesztek is érzékenyek a normalitásra, ezért más tesztet kell alkalmazni a normál eloszlás feltételének teljesülése és sérülése esetén.

Ha nem teljesül a függetlenség, azt nem igazán lehet korrigálni, Azért nem feltétlenül biztos halál: vannak olyan kísérleti elrendezések, melyek tudatosan a kísérleti személyek egymásra hatását figyelembe véve dolgoznak, és vannak statisztikák, melyekkel az ilyen adatokat elemezni lehet.

A feltételek ellenőrzésekor általában a függő változó parametrikusságát kell ellenőriznünk (például nem ellenőrizzük egy t-próbánál, hogy a csoportosító változó normál eloszlást követ-e), de lehet olyan elemzés is, melynél a független változókra is van feltételük (például regresszió elemzés)

Ha a függő változó nem skála típusú, nem teljesül/nem értelmezhető a normál eloszlás, ezért ilyenkor nem is ellenőrizzük.

Bizonyos parametrikus próbák megengedik, hogy a függő változó dichotóm legyen. Normalitást attól még továbbra sem ellenőrzünk dichotóm adaton.

Itt válik először kiemelkedően fontossá, hogy a feltételeknek mindig az adott hipotézishez tartozó mintán kell teljesülnie, nem az egész adatbázison. Tehát ha a különböző hipotéziseinkhez és az alkalmazandó próbákhoz különböző minták tartoznak, akkor minden próbához külön ellenőrizni kell, hogy az ahhoz tartozó mintán teljesülnek-e a feltételek.

A normál eloszlás ellenőrzésének három elégséges módszere van. A három elemzés közül BÁRMELYIK teljesül, feltételezhető a normál eloszlás. Éppen ezért nem kell mind a hármat elvégezni, ha bármelyik módszer alapján teljesül a normál eloszlás, a másik két ellenőrzést nem kell elvégezni.

A szóráshomogenitás a a parametrikusság feltételei közül a legkevésbé egységes. Egyrészről van, ahol a minták szórásának hasonlóságát ellenőrizzük (szóráshomogenitás), van ahol a minták közötti különbségek szórásának hasonlóságát (szfericitás), van, ahogy az egyik változó szintjein a másik változó szórásának hasonlóságát (homoszkedaszticitás), van ahol nem kell szóráshomogenitást ellenőrizni. Másrészről próbánként eltérő módon ellenőrizzük. Harmadrészről az SPSS-ben a legtöbb próbába be van építve a szóráshomogenitás ellenőrzése, és általában nem is kell külön próbát használnunk a feltétel sérülése esetén, mert a próbákban korrekció is be van építve ilyen esetekre. Éppen ezért sokszor nem is szoktuk külön előre ellenőrizni ezt a feltételt, hanem kikérjük a próbát, ott megnézzük a szóráshomogenitás tesztet, és ha teljesül a feltétel, akkor az eredeti elemzéshez tartozó eredményeket, ha nem teljesül, akkor a korrekcióhoz tartozó táblázatokat értelmezzük.

Ha nem teljesül a normalitás és szóráshomogenitás sem, olyan nemparametrikus tesztet érdemes használni, melyet a normalitás sérülése esetén választanánk. (A legtöbb ilyen teszt rangsorolással dolgozik, mely megoldja a szóráshomogenitás sérülésének kérdését is egyben)

Nemparametrikus próbák között is van többféle, melyek különböző körülményekre a legalkalmasabbak, és különböző módszerrel dolgoznak.

Ne feledjétek, ez a flowchart csak útmutatóként használható, a statisztikai elemzés mindig értelmező döntések sorozatából áll, így az itt felvázolt lépések mechanikus követése nem feltétlenül vezet jó megoldáshoz.

Várhelyi Klára - stathelp.hu