https://doi.org/10.63684/dk.2025.4.01

BOGNÁR LÁSZLÓ *

Így szoktuk elrontani az oktatási felméréseinket a statisztikai elemzéseknél és az azokból levont következtetéseknél

Összefoglalás: Az oktatási felmérések statisztikai elemzése során számos lehetőségünk van arra, hogy hibát kövessünk el és téves következtetéseket vonjunk le – és sokszor élünk is ezekkel a lehetőségekkel. Ez a cikk a leggyakoribb hibákat tárgyalja, amelyeket elkövethetünk, miközben azt gondoljuk, hogy pontos elemzéseket végzünk. Bemutatja a tervezett kísérletek és a megfigyeléses adatok közötti különbségeket, valamint a legkedveltebb statisztikai módszerek – mint a t-próba, ANOVA, korreláció, regresszió és khi-négyzet-próba – alkalmazásánál elkövetett tipikus hibákat. Kiemelt figyelmet fordítunk a statisztikai feltételezések ellenőrzésére, a normalitás vizsgálatára, a függetlenég vizsgálatára, és az adatok megbízhatóságára, mert ahogy mondani szokták: a statisztika nem hazudik, de miért ne segítenénk neki egy kicsit.
Kulcsszavak: Oktatási felmérések; statisztikai elemzések; statisztikai hibák; tervezett kísérletek; megfigyeléses adatok; t-próba; ANOVA; korreláció; regresszió; khi-négyzet-próba; normalitásvizsgálat; függetlenségvizsgálat.

* *Dunaújvárosi Egyetem, Informatikai Intézet, Matematika és Számítástudományi Tanszék*
Email: bognarl@uniduna.hu

Bevezetés

Ha már valahogyan túlvergődtünk a kérdések megfogalmazásán, a résztvevők kiválasztásán, és eddig nem követtünk el nagy hibát, akkor már csak a következtető statisztikai elemzések tengernyi lehetősége, feltételei és korlátjai között kell eligazodnunk.

Nem vállalkozom arra, hogy az oktatási felmérésekben alkalmazott módszerek mindegyikénél részletezzem az elkövetett matematikai jellegű hibákat, hiszen az alkalmazott módszerek tárháza rendkívül széles. Csak a leggyakrabban használt módszerekről lesz szó. A bonyolultabb matematikai elemzéseket hagyjuk meg a statisztikusoknak.

Leginkább a koncepcionális hibákról beszélek – arról, amikor rossz módszereket alkalmazunk, vagy annak ellenére vonunk le következtetéseket, hogy nem biztosítottuk az adott módszer alkalmazhatóságának a feltételeit.

Tervezett kísérletek kontra megfigyeléses adatok

A tervezett oktatási kísérletek során szándékosan manipulálunk bizonyos körülményeket, például különböző oktatási módszereket vagy eszközöket alkalmazunk, hogy megfigyelhessük, hogyan hatnak ezek a hallgatói teljesítményre vagy egyéb kimenetelekre. Ilyenkor ellenőrzésünk alatt tartjuk a beavatkozást, és különböző csoportokat hasonlítunk össze, hogy megbízható ok-okozati következtetéseket vonhassunk le.

Ezzel szemben, ha megfigyeléses adatokkal dolgozunk, egyszerűen feljegyezzük, mi történt a természetes oktatási környezetben, anélkül, hogy szándékosan beavatkoznánk. Bár ezek az adatok is hasznosak lehetnek, sokkal óvatosabban kell kezelnünk őket, hiszen a különféle zavaró tényezők miatt nem tudunk biztosan ok-okozati összefüggéseket megállapítani. Ilyenkor az adataink együttállásáról lehet csak beszélni.

Ok-okozati összefüggések feltárása tervezett kísérletekből

Az egyik leggyakoribb hiba az ok-okozati összefüggések feltárásában, amikor csupán megfigyelt adatokból próbálunk ilyen következtetéseket levonni. A korreláció, azaz az adatok együttállása nem bizonyítja, hogy az egyik változó értékének a változása okozza a másik változó értékének a változását. Az ok-okozati kapcsolat csak speciálisan megtervezett kísérletekből, randomizált, kontrollált vizsgálatokból vonható le, ahol kizárhatók a zavaró tényezők.

Egy jó példa a tervezett kísérletre:

Azt szeretnénk vizsgálni, hogy a ChatGPT használatának engedélyezése javítja-e a vizsgaeredményeket egy adott tantárgy adott félévében. Ehhez randomizált, kontrollált kísérletet tervezünk.

A hallgatókat véletlenszerűen két csoportra osztjuk, tankörtől függetlenül. A kísérleti beavatkozás (a ChatGPT használata) csak az újonnan létrehozott csoportokra vonatkozik, ezekkel a csoportokkal dolgozunk a kísérlet során.

Az első csoportnak (kísérleti csoport) engedélyezzük a ChatGPT használatát az órákon és a házi feladatok elkészítésekor, míg a második csoportnak (kontrollcsoport) nem. Fontos, hogy a vizsgákon és a

ZH-kon egyik csoport sem használhatja a ChatGPT-t – ezeknél a megmérettetéseknél minden hallgatónak a saját tudására kell támaszkodnia. Mindkét csoportot ugyanaz a tanár tanítja, ugyanazon tananyagon dolgozik, és ugyanazokat a feladatokat kapja az órákon és a vizsgákon is. Az egyetlen különbség, hogy a kísérleti csoport szabadon használhatja a ChatGPT-t az anyag feldolgozására az órákon és a házi feladatok megoldására. A félév végén összehasonlítjuk a két csoport vizsgaeredményeit, hogy megvizsgáljuk, volt-e hatása a ChatGPT használatának az órákon. A véletlenszerű hozzárendelés biztosítja, hogy a csoportok között ne legyenek szisztematikus különbségek, így, ha az eredmények eltérnek, az a ChatGPT használatának tulajdonítható.

Hogyan lehet ezt a kísérletet elrontani?
Sokféleképpen. Itt van néhány példa:

Önkéntes csoport választás: Ha a diákokat nem véletlenszerűen osztjuk be a kísérleti és a kontrollcsoportokba, hanem hagyjuk, hogy maguk válasszanak, ez súlyos torzítást okozhat, mert azok a diákok, akik önként választják a ChatGPT használatát, valószínűleg motiváltabbak, jobban érdeklődnek a technológia iránt, vagy eleve jobb tanulmányi eredményekkel rendelkeznek. Ez azt eredményezi, hogy az eltérések a vizsgaeredményekben nem feltétlenül a ChatGPT hatását tükrözik, hanem inkább a diákok motivációjának, érdeklődésének vagy előzetes tudásának különbségeiből adódhatnak.

Nem ugyanaz a tanár, tananyag vagy vizsgakörnyezet: Ha a csoportok nem azonos tananyagot kapnak, vagy nem azonos tanár tanítja őket, vagy eltérő vizsgakörnyezetben teljesítenek, az eredményeket ezek a zavaró tényezők torzíthatják.

Évfolyamok vagy egyéb rétegek keverése a csoportokban: Ha a csoportokat úgy osztjuk be, hogy elsőéves és felsőbb évfolyamos hallgatókat keverünk, az eredmények torzulhatnak az eltérő tapasztalati szint miatt. A felsőbb éves hallgatók, akik már több tapasztalattal rendelkeznek, valószínűleg jobban tudják kihasználni a ChatGPT-t, mint az elsőévesek, akik még csak most ismerkednek a felsőoktatási követelményekkel. Ez azt eredményezheti, hogy a teljesítménybeli különbségek nem csak a ChatGPT használatából, hanem az évfolyamok közötti különbségekből is adódnak.

Tantárgyak közötti különbségek figyelmen kívül hagyása: Ha a ChatGPT használatát különböző tantárgyak hallgatói között vizsgáljuk anélkül, hogy ezt megfelelően kontrollálnánk, az eredmények torzulhatnak. Bizonyos tantárgyak, mint például az informatika, nagyobb mértékben profitálhatnak a ChatGPT használatából, míg más tantárgyak, például a művészetek, kevésbé. Ha a csoportok tantárgyi összetétele eltér, akkor az eredmények nem csak a ChatGPT használatát, hanem a tantárgyi különbségeket is tükrözni fogják.

Összefüggések vizsgálata a megfigyelt adatokból

Megfigyelésből származó adatok esetén, ha ok-okozati összefüggéseket nem is, de sokféle egyéb hasznos vizsgálatot végezhetünk. A megfigyelt adataink jelentik a mintát, és természetesen itt sem egyszerűen a mintában látható összefüggéseket akarjuk csak leírni (ez a leíró statisztika dolga), hanem a minta mögötti hallgatói sokaságra vonatkozó összefüggéseket, relációkat szeretnénk találni.

Az alkalmazott statisztikai módszerek mind a tervezett kísérleteknél, mind a megfigyelt adatok elemzésénél ugyanazok lehetnek, de a következtetéseink alapvetően különbeznek, hiszen az egyiknél ok-okozati összefüggéseket tárhatunk fel, míg a másiknál csak az adatok csoportjainak jellegzetességeiről, azok relációjáról, kapcsolataik szorosságáról, az együttállási tendenciákról győződhetünk meg.

A leggyakoribb elemzési feladatok és a hozzájuk tartozó statisztikai módszerek

Egy vagy többmintás összehasonlítások

Az egy- és többmintás összehasonlítások olyan statisztikai módszerek, amelyek segítségével különböző csoportok vagy egy csoport különböző időpontbéli adatait tudjuk összehasonlítani. Ezek a módszerek alapvetően arra szolgálnak, hogy megvizsgáljuk, van-e statisztikailag szignifikáns különbség a csoportok valamilyen jellemzője között. (A szignifikáns különbségről később beszélünk.)

– Egymintás t-próba:
Az egymintás t-próbát akkor használjuk, ha egy csoport átlagát szeretnénk összehasonlítani egy ismert populációs átlaggal. Például, ha tudjuk, hogy az egyetemi hallgatók országos átlagpontszáma egy adott teszten 70 pont, akkor ezt az átlagot összehasonlíthatjuk egy helyi egyetem hallgatóinak tesztátlagával, hogy megtudjuk, szignifikánsan eltér-e az országos átlagtól.

– Kétmintás t-próba:
A kétmintás t-próba két független csoport átlagainak összehasonlítására szolgál. Ez hasznos, ha például azt szeretnénk vizsgálni, hogy a ChatGPT-t használó hallgatók és a nem használók vizsgaeredményei között van-e különbség. A két csoport független egymástól, és a módszer megmutatja, hogy az átlagos teljesítmények eltérnek-e olyan mértékben, amire azt mondhatjuk, hogy ez statisztikailag szignifikáns.

– Páros t-próba:
A páros t-próbát akkor használjuk, ha ugyanannak a csoportnak minden tagját kétszer mérjük meg, például egy beavatkozás előtt és után. Ez hasznos, ha azt szeretnénk vizsgálni, hogy ugyanazon hallgatók eredményei változtak-e, miután egy új oktatási eszközt (pl. ChatGPT) bevezettünk.

– ANOVA (varianciaanalízis):
Az ANOVA akkor jön jól, ha több mint két csoportot szeretnénk összehasonlítani. Például, ha három különböző tanulási módszert (hagyományos oktatás, ChatGPT-használata, egyéb technológiai eszközök használata) szeretnénk összehasonlítani, az ANOVA segít megvizsgálni, hogy van-e szignifikáns különbség a csoportok átlagos teljesítményei között.

Korreláció

A korrelációszámítás az a statisztikai módszer, amely két változó közötti kapcsolat (vagy együttjárás) mértékét méri. Arra használjuk, hogy megállapítsuk, van-e lineáris kapcsolat két változó között, és ha van, milyen irányú és erősségű ez a kapcsolat. A korrelációs együttható az értékét tekintve -1 és +1 között mozog.

Egy vagy többváltozós regresszió

A regresszió egy olyan statisztikai módszer, amelynek a segítségével megvizsgálhatjuk egy vagy több független változó (prediktorok) kapcsolatát egy függő változóval (kimeneti változó, válasz változó). Oktatási kutatásokban a regressziót gyakran használjuk arra, hogy feltárjuk, milyen tényezők befolyásolhatják például a hallgatók teljesítményét, vagy hogy megbecsüljük, mennyire hatékony egy oktatási beavatkozás A lineáris regresszió a regresszió legegyszerűbb formája, ahol feltételezzük, hogy a függő és független változók közötti kapcsolat lineáris. Fontos azonban, hogy megfigyeléses adatok esetén a regresszió csak a változók közötti kapcsolatot mutatja meg, de nem bizonyít ok-okozati összefüggést.

Regresszió kontrollváltozókkal

A regresszió kontrollváltozókkal egy továbbfejlesztett elemzési technika, amely lehetővé teszi, hogy egy adott független változó hatását tisztábban vizsgáljuk a függő változóra, miközben más, zavaró tényezők (kontrollváltozók) hatását kiszűrjük. Ezek a kontrollváltozók olyan tényezők, amelyekről tudjuk, hogy befolyásolhatják a függő változót, de nem közvetlenül érdekelnek minket a jelenlegi kutatásban. Például, ha azt vizsgáljuk, hogy a ChatGPT használata javítja-e a hallgatók vizsgaeredményeit, a kontrollváltozók között lehet a hallgatók előzetes tudása, motivációja vagy akár az évfolyam, amelyre járnak.

A kontrollváltozók beépítése a regressziós modellbe segít elkülöníteni a vizsgált független változó (pl. ChatGPT-használat) hatását a zavaró tényezőktől, így pontosabb eredményeket kapunk.

Két vagy több kategoriális változó közötti függetlenség vizsgálata khi-négyzet-próbával

A khi-négyzet-próba a különböző kategóriákhoz tartozó megfigyelt és várható gyakoriságok különbségét vizsgálja. A megfigyelt gyakoriságok a különböző kategóriákban megfigyelt adatok számát jelenti. Ezeket hasonlítjuk össze azokkal a gyakoriságokkal, amelyeket akkor várnánk, ha a kategoriális változókról tudnánk, hogy függetlenek egymástól. Ha az eltérések jelentősen nagyobbak, mint amit csak a véletlenszerűség miatt várnánk, akkor a khi-négyzet értéke magas lesz, valószínűleg szignifikáns kapcsolat áll fenn a változók között.

Egy alkalmazási példa lehet, amikor azt szeretnénk vizsgálni, hogy van-e kapcsolat a hallgatók órákon való részvétele és a vizsgán elért eredményeik között. A kérdés az, hogy a rendszeresen órákra járó hallgatók eredményei szignifikánsan jobbak-e, mint azokéi, akik ritkábban járnak órákra vagy ez a két dolog független egymástól.

A statisztikai elemzések, következtetések

Lehetőségek az elemzések elrontására, mielőtt még hozzákezdenénk az elemzésekhez

Mielőtt ténylegesen nekikezdenénk az adatok elemzésének, már számos lehetőség kínálkozik arra, hogy elronthassuk az elemzést, ha nem fordítunk kellő figyelmet a kezdeti lépésekre. Ezek közé tartozik a szélsőséges értékek (outlierek) és a gyanús értékek kiszűrésének helyes elvégzése, valamint a kérdőív megbízhatóságának ismételt ellenőrzése, most már a beérkezett válaszok tükrében. Mindkét területen elkövetett hibák könnyen vezethetnek félrevezető eredményekhez és helytelen következtetésekhez.

A szélsőséges értékek kiszűrése

Az outlierek, vagyis szélsőséges értékek, azok az adatok, amelyek lényegesen eltérnek a többi adattól, és jelentős torzítást okozhatnak az elemzés során. Ezeket kiemelt figyelemmel kell kezelni, mert befolyásolhatják az összesített statisztikai mutatókat. Az átlag például jelentősen eltérhet a valós középértéktől, ha egy-két szélsőséges adat hatása érvényesül.

Fontos megjegyezni, hogy nem minden outlier hibás adat. Előfordulhat, hogy egy szélsőséges érték valós választ tükröz, esetleg valami érdekes vagy rendkívüli körülményhez tartozik.

Ezért az outlierek kiszűrésekor nemcsak automatikusan távolítjuk el őket, hanem ha lehetőségünk van rá, elemeznünk kell, hogy valóban hibás adatokról van-e szó, vagy éppen egy ritka körülmény áll fenn. Az outlierek szűréséhez használjunk megbízható statisztikai módszereket (például z-score vagy box-plot módszert), majd gondosan mérlegeljük, hogy mely adatokat töröljük. Az outlierek eltávolítását alaposan dokumentálni kell, hogy később visszakereshető legyen, miért történt meg a kizárás, és hogy az elemzés milyen mértékben változott ennek hatására. A túlzott outlierszűrés probléma lehet. Ha túl sok adatot zárunk ki, az elemzésünk veszít a megbízhatóságából és általánosíthatóságából, hiszen csökkentjük a minta méretét, ami hatással van az eredményekre.

A gyanúsan válaszolók kiszűrése

Fontos, hogy kiszűrjük azokat a válaszokat is, amelyekből arra lehet következtetni, hogy a válaszadók nem vették komolyan a kitöltést, vagy mechanikusan válaszoltak.

Ha egy válaszadó minden kérdésre ugyanazt a választ adja, az arra utalhat, hogy nem figyelmesen töltötte ki a kérdőívet, és a válaszai nem megbízhatók. Ilyen válaszok esetén a válaszok szórása nulla, ami az egyik leggyakoribb jele a nem megbízható válaszadási mintázatnak.

Ha a válaszadó túl gyorsan tölti ki a kérdőívet (pl. jóval kevesebb idő alatt, mint amennyi az ésszerű kitöltéshez szükséges), akkor feltételezhető, hogy nem adott átgondolt válaszokat. Az ilyen válaszok is torzíthatják az elemzés eredményeit.

Ha a válaszadó válaszai között nagy következetlenségek vannak (például ellentmondásos válaszok a különböző kérdéseknél), az arra utalhat, hogy nem figyelmesen válaszolt.

Ezt például úgy lehet észlelni, hogy összehasonlítjuk az egyes kérdésekre adott válaszokat és figyeljük, hogy azok logikailag összhangban vannak-e egymással.

A kérdőív megbízhatóságának vizsgálata a beérkezett válaszok tükrében

A kérdőív megbízhatóságának (reliability) vizsgálata nélkülözhetetlen lépés az adatgyűjtés után. Ez biztosítja, hogy a kérdőív következetesen mérje azokat a fogalmakat, azokat a kutatási kérdéseket, amelyeket mérni szeretnénk. A Cronbach-alfa a leggyakrabban használt mérőszám, amely azt mutatja meg, hogy a kérdőívben szereplő kérdések mennyire vannak összhangban egymással. Az általánosan használt kritérium, hogy ennek az értéke 0.7-nél nagyobb legyen.

Hogyan ronthatjuk el a Cronbach-alfa számítását?

– Például úgy, hogy egész kérdőívre, az összes kérdésre együtt nézzük a Cronbach-alfa értéket, ahelyett, hogy az egyes kutatási kérdésekhez kapcsolódó kisebb kérdéscsoportokat külön vizsgálnánk.

A Cronbach-alfa értékének specialitása, hogy minél több kérdést tartalmaz egy kérdéscsoport, annál nagyobb az alfa értéke, ami miatt könnyen téves következtetéseket vonhatunk le. Ha az egész kérdőívre számolunk alfa értéket, lehet, hogy úgy tűnik, a kérdőív megbízható, miközben az egyes kérdéscsoportokon belül nem kapunk következetes eredményeket.

A helyes megközelítés:

– Minden egyes kutatási kérdéshez tartozó kérdéscsoportnál külön kell ellenőrizni a megbízhatóságot. Ha az adott kérdéscsoport Cronbach-alfa értéke magas (általában 0,7 felett), akkor azt mondhatjuk, hogy az adott kérdéscsoport belső konzisztenciája jó.

A túl alacsony Cronbach-alfa érték kezelése:

– Ha egy adott kérdéscsoport Cronbach-alfa értéke túl alacsony, nem feltétlenül kell az egész kérdéscsoportot kidobni. Első lépésként érdemes megvizsgálni, hogy egyes kérdések hozzájárulnak-e az alacsony értékhez. Egy vagy több olyan kérdést is találhatunk, amelyek nem illeszkednek jól a többihez, és ezek eltávolításával javítható a megbízhatóság.

Hipotézisvizsgálatok.
Mit is jelent a szignifikáns különbség?

A statisztikai elemzések célja, hogy segítsenek nekünk bizonyos döntéseket meghozni. Ennek a leggyakoribb módszere az úgynevezett hipotézisvizsgálat. Igyekszem konyhanyelven összefoglalni, hogyan is működik a döntési mechanizmus.

Hipotézisvizsgálatok

1. A döntések, amiket szeretnénk meghozni, soha nem a kiválasztott mintákra vonatkoznak, hanem azokra a sokaságokra, amikből a mintákat vettük. Úgy is mondhatnánk, hogy azokra a sokaságokra, amelyekre a minták alapján általánosíthatunk. Szóval, ha rosszul határoztuk meg, hogy egy minta alapján melyik sokaságra általánosíthatjuk a döntésünket, akkor azon a legprecízebb hipotézisvizsgálat sem segít.
2. Egy hipotézisvizsgálatnál az eldöntendő kérdések általában ilyenek: *Történt változás az eredetihez képest? Van különbség a kettő között? Különbözik legalább egy a többitől? Van összefüggés ezek között?* (Ezek a kérdések mindig a sokaságnak valamilyen jellemzőjére vonatkoznak, nem a mintáéra.)

3. Általában akkor nyugodnánk meg, akkor lennénk elégedettek, ha azt kapnánk ezekre a kérdésekre válaszul, hogy: *Igen, történt változás. Igen, van különbség. Igen legalább az egyik különbözik. Igen, van összefüggés.* Persze az igazi elégedettségünkhöz még az is kell, hogy az elemzés azt mutassa, hogy ezekben az igen válaszokban nagy bizonyossággal hihetünk, az elemzések nagyon meggyőző bizonyítékot szolgáltatnak erre. Szóval csak ilyenkor hisszük el ezeket az igen válaszokat.
4. A döntési módszer a következő:
   – Tételezzük, fel, hogy: *Nem, nincs itt változás. Nem, nincs különbség a kettő között. Nem, nincs egy sem, amelyik különbözne. Nem, nincs itt összefüggés, ezek függetlenek.* **Egy ilyen feltételezés a nullhipotézis.**
   – Ezután az elemzés során kiszámítjuk, hogy mekkora annak a valószínűsége, hogy ha igaz a nullhipotézisünk (ami, ne felejtsük el, mindig a sokaság valamilyen jellemzőjére vonatkozik), akkor egy ilyen, a nullhipotézisnek megfelelő sokaságból egy véletlen mintavétel során pont olyan mintát kapjunk, mint amilyet éppen kaptunk. **Ez a valószínűség a P-érték.** Abban reménykedünk, hogy ez a valószínűségi érték kicsi (általában 5%-nál kisebb), mert akkor joggal gondolhatjuk, hogy a mi mintánk extrém módon eltér attól, amit várnánk a nullhipotézis fennállása esetén, szóval, valószínűleg ez azért van, mert nem is igaz az, amit a nullhipotézisben feltételeztünk. Ilyenkor azt mondjuk, hogy a feltételezett érték és a tényleges érték között **statisztikailg szignifikáns különbség** van, és inkább elfogadjuk az igenlő válaszokat, amiket **alternatív (vagy kutatási) hipotéziseknek** szoktunk hívni.

A szignifikáns különbség illúziója

A P-érték kiszámítása algoritmusának az a sajátossága, hogy minél nagyobb elemszámú mintából számítjuk ki, annál kisebb P-értéket kapunk. Ez azt jelenti, hogy ha nagy elemszámú (100–1000 fős) mintával dolgozunk, ami a kérdőíves kutatásoknál gyakran előfordul, akkor egészen kicsi eltérés esetén is statisztikailag szignifikáns különbséget kapunk, holott ez a különbség a számunkra a gyakorlatban lehet, hogy érdektelen. Például, ha egy adott tantárgyhoz tartozó vizsgaeredményeket hasonlítunk össze a nők és a férfiak között, akkor egy 500 fős mintából a 4.23-as férfi átlag és a 4.24-es női átlag között is statisztikailag szignifikáns különbséget mutathatunk ki, de ez a különbség olyan kicsi, hogy nem igazán érdekel bennünket. Ezért aztán a statisztikailag szignifikáns különbséges esetén is mindig meg kell nézni, hogy ez a gyakorlatban is szignifikáns különbséget jelent-e a számunkra.

A leggyakrabban elkövetett hibák az egyes statisztikai módszereknél

A teljesség kedvéért ejtsünk szót arról, hogy milyen tipikus hibákat szoktunk elkövetni az egyes statisztikai vizsgálatoknál, a szoftveres elemzéseknél.

– Egy- és kétmintás t-próba, páros t-próba:

Nem megfelelő feltételezések teljesítése: A t-próba egyik legfontosabb feltétele, hogy a sokaság, amelyből a mintát vettük, normális eloszlású. Gyakori hiba, hogy ezt a normalitási feltételt nem ellenőrizzük.

A normalitást vizsgálni kell. Ezt Anderson–Darling-teszttel, Shapiro–Wilk vagy Kolmogorov–Smirnov-tesztekkel lehet ellenőrizni. Ha kiderül, hogy az adatok nem normális eloszlásból származnak, nem-paraméteres módszereket, például a Mann–Whitney U-tesztet kell alkalmazni, amely nem igényli a normális sokasági eloszlást.

Likert-skálás adatok kezelése: Az oktatási felmérésekben gyakran használt Likert-skálák (pl. 1–5 vagy 1–7 közötti értékelések) alapvetően ordinális skálák. Bár ezek a skálák nem folytonos adatokat eredményeznek, sokszor mégis t-próbát alkalmazunk rájuk, ami különösen kis elemszámú minták esetében okozhat nagyobb hibákat.

Ha Likert-skálás adataink vannak, és a mintaméret kicsi, érdemes nem-paraméteres teszteket, például Wilcoxon-rangsoros próbát alkalmazni, mivel ez jobban illeszkedik az ordinális adatok természetéhez. Ha a minta elegendően nagy, a t-próba is alkalmazható, mivel a centrális határeloszlás-tétel ilyenkor érvényesülhet.

Kétmintás t-próbánál a varianciák azonosságának figyelmen kívül hagyása: A kétmintás t-próba egy fontos feltétele, hogy a két minta mögötti sokaságok varianciája megegyezzen. Ha a varianciák különböznek (heteroszkedaszticitás), akkor a t-próba eredményei torzulhatnak.

A varianciák egyenlőségét Levene-teszttel érdemes ellenőrizni. Ha a varianciák eltérnek, akkor a t-próba egy olyan változatát kell használni, amely nem feltételezi a varianciák azonosságát, mint például a Welch-féle t-próba.

A függetlenség feltételezésének megsértése: A kétmintás t-próba feltételezi, hogy a két minta független egymástól. Ez gyakran hibás feltételezés lehet, ha például ugyanazok a résztvevők szerepelnek két különböző mérésben, vagy ha valamilyen kapcsolat van a két minta között.

Ha a minták között valamilyen kapcsolat van (pl. ugyanazok a hallgatók szerepelnek két mérésben),

célszerűbb párosított t-próbát alkalmazni, amely figyelembe veszi a mérések közötti összefüggéseket, és pontosabb eredményt ad.

– ANOVA (varianciaanalízis):

A normális eloszlás feltételének figyelmen kívül hagyása: Az ANOVA egyik alapvető előfeltétele, hogy a sokaságok, amelyből a csoportok mintái származnak, normális eloszlásúak. Gyakran előfordul, hogy ezt a feltételt nem vizsgáljuk meg, különösen kisebb minták esetében okoz nagy hibákat.

Ha az adatok nem követik a normális eloszlást, érdemes nem-paraméteres alternatívákat használni, mint például a Kruskal–Wallis-teszt.

A csoportok varianciáinak egyenlősége (homoszkedaszticitás): Az ANOVA feltételezi, hogy a minták mögötti sokaságok varianciái azonosak (homogének). Ha ez az előfeltétel nem teljesül, a teszt eredményei megbízhatatlanok lesznek, mivel az ANOVA túlérzékeny a heteroszkedaszticitásra.

A varianciák azonosságát Levene-teszttel vagy Bartlett-teszttel ellenőrizhetjük. Ha a varianciák különböznek, érdemes a Welch-féle ANOVA-t használni, amely nem érzékeny a varianciák különbségére.

Túl sok csoport összehasonlítása post hoc teszt nélkül: Az ANOVA csak azt mutatja meg, hogy van-e statisztikailag szignifikáns különbség a csoportok között, de nem mondja meg, hogy pontosan mely csoportok különböznek egymástól.

Gyakori hiba, hogy a kutatók elfelejtik elvégezni a post hoc-teszteket, amelyek lehetővé teszik a páros összehasonlításokat a csoportok között.

Ha az ANOVA szignifikáns eredményt mutat, mindig post hoc-teszteket kell alkalmazni (például Tukey-tesztet), hogy pontosan megállapítsuk, mely csoportok között van különbség.

ANOVA alkalmazása ismételt méréses adatokra: Az ANOVA feltételezi, hogy a csoportok függetlenek egymástól. Ha ismételt mérésekkel dolgozunk, vagyis ugyanazokat a résztvevőket mérjük több időpontban vagy körülmény között, az adatok nem függetlenek, és az egyszerű ANOVA nem ad megbízható eredményt.

Ilyenkor ismételt méréses ANOVA-t (Repeated Measures ANOVA) kell használni, amely figyelembe veszi, hogy ugyanazok az alanyok szerepelnek több mérésben.

Csoportméretek nagy különbsége: Az ANOVA feltételezi, hogy a csoportok méretei hasonlóak. Ha az egyik csoport sokkal nagyobb vagy kisebb mint a többi, az torzíthatja az eredményeket és csökkentheti a teszt statisztikai erejét.

Bár az ANOVA tolerál bizonyos mértékű csoportméret-különbséget, ha a csoportok méretei nagyon eltérnek, érdemes kiegyenlíteni a mintákat, vagy alternatív tesztet alkalmazni, mint a Generalized Linear Model (GLM).

– Korreláció:

Ok-okozati összefüggés feltételezése: Az egyik leggyakoribb hiba, hogy a korrelációból ok-okozati összefüggést vonunk le. A korreláció csak azt mutatja meg, hogy két változó között van valamilyen statisztikai kapcsolat, de nem jelzi, hogy az egyik változó okozza-e a másik változót.
Korreláció kiszámítása nem lineáris összefüggésekre: A Pearson-féle korrelációs együttható csak lineáris kapcsolatot mér a változók között. Ha a kapcsolat nem lineáris (például U-alakú), akkor a Pearson-korreláció nem lesz képes megfelelően jellemezni a kapcsolatot, és hamis következtetéseket vonhatunk le arról, hogy nincs kapcsolat a változók között.

Mielőtt elvégeznénk a Pearson-féle korrelációt, grafikus ábrázolással (pl. szóródásábrával, scattergrammal) meg kell vizsgálni a változók közötti kapcsolat természetét. Ha a kapcsolat nem lineárisnak tűnik, akkor érdemes nem-paraméteres korrelációs módszert használni, mint például a Spearman-féle rangkorreláció.

Túlságosan kis minta használata: A korrelációs elemzés eredményei különösen érzékenyek a mintanagyságra. Kismintás esetben a korrelációs-együttható megbízhatatlan lehet, mivel egy-két kiugró érték erősen befolyásolhatja az eredményt, és a véletlenszerű együttmozgás is könnyen félrevezető eredményt adhat.

Törekedjünk arra, hogy megfelelően nagy mintát használjunk, és ha a minta kicsi, akkor érdemes a megbízhatóságot bootstrap módszerrel fokozni.

Outlierek figyelmen kívül hagyása: Az outlierek, vagyis szélsőséges értékek, komolyan befolyásolhatják a korrelációs együtthatót. Egyetlen kiugró adatpont torzíthatja a kapcsolat erősségét és irányát is, ezért fontos az outlierek azonosítása és kezelése az elemzés előtt.

– Egy és többváltozós regresszió:

Itt a hibák elkövetésének tárháza szinte végtelen. Ezt tényleg jobb, ha egy statisztikusra bízzuk. Csak a legtipikusabb hibákat említem.

Lineáris kapcsolat feltételezése, amikor az nem áll fenn: A regressziós elemzés során gyakran előfeltételezzük, hogy a függő és a független változók közötti kapcsolat lineáris. Azonban ez nem mindig igaz, és ha a kapcsolat nem lineáris, akkor a lineáris regresszió alkalmazása helytelen eredményekhez vezethet. Mielőtt alkalmaznánk a lineáris regressziót, grafikus ábrázolással kell megvizsgálni a változók közötti kapcsolatot. Ha a kapcsolat nem lineáris, akkor más módszerek, nem-lineáris modellek alkalmazása indokolt.

Túl sok független változó bevonása a modellbe (overfitting): A többváltozós regresszió esetében gyakori hiba, hogy túl sok független változót vonunk be a modellbe. Ez a modell túlzott illeszkedéséhez (overfitting) vezethet, ami azt jelenti, hogy a modell túl jól alkalmazkodik a mintához, de nem teljesít jól új adatok esetén.

A jó modell kiválasztásához használhatunk olyan módszereket, mint a lépésenkénti regresszió (Stepwise regression), vagy a legjobb részhalmaz szerinti regresszió (Best subset regression), amelyek segíthetnek minimalizálni a túlzott illeszkedés kockázatát.

Multikollinearitás figyelmen kívül hagyása: A többváltozós regresszió esetében a független változók közötti erős korreláció (multikollinearitás) problémát jelenthet, mivel torzítja a becsült regressziós együtthatókat, és nehezíti a változók hatásának megfelelő értelmezését. A multikollinearitás kimutatására használhatjuk a Variance Inflation Factor-t (VIF). Ha egy változó VIF értéke túl magas (általában 10 felett), az jelzi a multikollinearitást. Ilyenkor célszerű lehet eltávolítani egy vagy több erősen korreláló független változót, vagy módosítani a modellt.

Heteroszkedaszticitás figyelmen kívül hagyása: A regresszió egyik fontos feltétele, hogy a reziduumok (vagy maradékok), azaz a függő változó megfigyelt értékei és a modell által előrejelzett értékek közötti különbségek szórása állandó a megfigyelések során.

Ezt nevezzük homoszkedaszticitásnak. Ha azonban a reziduumok szórása nem állandó – vagyis a szórás az adatok különböző szakaszaiban eltér –, akkor ezt heteroszkedaszticitásnak nevezzük. Ez torzíthatja az eredményeket, mivel ilyenkor a modell egyes adatoknál pontosabb, míg más adatoknál kevésbé pontos lehet, ami csökkenti az elemzés megbízhatóságát és érvényességét.

A rezidumok grafikus ábrázolásával ellenőrizhetjük, hogy fennáll-e heteroszkedaszticitás. Ha heteroszkedaszticitás van jelen, akkor olyan regressziós modelleket kell használnunk, amelyek kezelik ezt a heteroszkedaszticitást.

Hiányzó változók problémája: Ha egy fontos változót kihagyunk a modellből, az úgynevezett hiányzó változó torzítást okozhat (Omitted variable bias), ami azt jelenti, hogy a többi változó becsült együtthatói torzak lesznek, mivel egy fontos magyarázó tényező nincs figyelembe véve.

Extrapoláció túl messzire: A regressziós modell által adott eredményeket gyakran túl messzire extrapoláljuk, olyan tartományokra, ahol nincsenek mintaadatok. Ez különösen veszélyes, mivel a modell nem biztos, hogy jól működik a vizsgált tartományon kívül. Csak a megfigyelt adatok tartományán belül szabad a regressziós modellt alkalmazni. Ha extrapolációra van szükség, először további adatokat kell gyűjteni a kérdéses tartományban.

– Khi-négyzet-próba:

Kis cellaszám figyelmen kívül hagyása: A cella a khi-négyzet-próbában a két vagy több kategoriális változó különböző kategóriáinak kombinációját jelenti. A cellában található érték a megfigyelt gyakoriságot mutatja, amely azt jelzi, hogy az adott kategóriakombinációban hány esetet figyeltünk meg.

Például, ha a „nem" (férfi/nő) és a „szak" (műszaki/természettudományos) kategoriális változókat vizsgáljuk, akkor egy cella a férfiak és a műszaki szakosok kombinációját jelenti, és a cellában lévő szám azt mutatja, hány férfi műszaki szakos hallgató van a mintában. Az egyik leggyakoribb hiba, hogy a táblázat egyes celláiban túl kevés megfigyelés található. A khi-négyzet-próba alapvető feltétele, hogy a várható gyakoriság minden cellában legalább 5 legyen. Ha ennél kisebb, akkor a próba eredménye megbízhatatlan lehet.

Ha a várható gyakoriságok túl alacsonyak, érdemes egyes kategóriákat összevonni, hogy növeljük a cellákban lévő értékeket. Alternatív megoldás lehet a Fisher-féle egzakt-próba használata, amely jobban működik kis minták esetén.

Összefüggés interpretálása függetlenség helyett: Gyakori hiba, hogy amikor a khi-négyzet-próba eredménye szignifikáns, akkor a kutatók összefüggést vagy ok-okozati kapcsolatot feltételeznek a változók között. A khi-négyzet-próba csak a függetlenséget vizsgálja, és nem mutatja meg, hogy a változók között valódi ok-okozati kapcsolat van-e.

Konklúzió

Az oktatási felmérésekben gyakran alkalmazott statisztikai elemzési módszerek helytelen használata számos hibalehetőséget kínál. Bár a statisztikai elemzés a megbízható következtetések levonásának elengedhetetlen eszköze, a módszerek alkalmazhatóságához szükséges feltételek hiánya, könnyen téves eredményekhez vezethet. A leggyakoribb hibák közé tartozik a normalitásvizsgálatok hiánya, a minták függetlenségének megsértése, illetve a különböző statisztikai próbák alkalmazása olyan szituációkban, amikor valami mást módszert kellene használni.

Emellett gyakran követünk el hibákat a megfigyeléses adatokból való ok-okozati következtetések levonásánál, mivel ezek az adatok sokszor nem biztosítják a zavaró tényezők megfelelő kontrollját.

A statisztikai elemzés szabályainak betartása, az alkalmazhatósági feltételek ellenőrzése, a megbízhatóság és érvényesség vizsgálata mind-mind fontosak ahhoz, hogy az eredmények valóban hasznos és pontos következtetésekhez vezessenek. Ha nem figyelünk ezekre, akkor az elemzés téves eredményekhez, és végül félrevezető döntésekhez vezethet.

Irodalom

Bognár L. (2019): *Statisztika Online.* https://www.statisztika-online.hu/

Cohen, J. (1988): *Statistical power analysis for the behavioral sciences* (2nd Ed.). New York: Lawrence Erlbaum Associates.

Field, A. (2018): *Discovering statistics using IBM SPSS statistics* (5th Ed.). London: Sage Publications.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E.–Tatham, R. L. (2010): *Multivariate data analysis* (7th Ed.). New York: Pearson.

Howell, D. C. (2013): *Statistical methods for psychology* (8th Ed.). Wadsworth Cengage Learning.

Kovács E.–Nagy, J. (2015): *Statisztikai elemzések az oktatási kutatásokban: Módszertani áttekintés és gyakorlati példák.* Budapest: Akadémiai Kiadó.

Pallant, J. (2020): *SPSS survival manual* (7th Ed.). McGraw–Hill Education.

Tabachnick, B. G.–Fidell, L. S. (2019): *Using multivariate statistics* (7th Ed.). New York: Pearson.

Zalai E. (2018): *Statisztikai módszerek és alkalmazásaik.* Budapest: Typotex Kiadó.