

PhD értekezés tézisei

# Téves besorolások figyelembevétele statisztikai elemzésekben

Dr. Hársfalvi Péter

Témavezető: Dr. Reiczigel Jenő



ÁLLATORVOSTUDOMÁNYI EGYETEM

Állatorvostudományi Doktori Iskola

Budapest, 2025

## Témavezető

.....

Dr. Reiczigel Jenő

Állatorvostudományi Egyetem

Biostatisztika Tanszék

Készült .... Példányban. Ez a(z) .... számú példány.

.....

Dr. Hársfalvi Péter

## Tartalom

1.	<i>Bevezetés és célkitűzések</i> .....	4
2.	<i>Az eredmények összegzése</i> .....	6
2.1	Profil-likelihood konfidencia intervallum betegség prevalenciájára, ha a diagnosztikai teszt hibával terhelt .....	6
2.2	Becsült szenzitivitással és specificitással korrigált konfidencia intervallum a relatív kockázatra és a kockázat különbségre .....	6
2.3	Téves besorolások hatása egy- és kétmintás, valószínűségekre vonatkozó statisztikai próbák mintaelemszámára .....	7
2.4	A promising zone mintaelemszám újrameghatározásos elrendezés használata olyan klinikai vizsgálatokban, ahol egy bináris végpont meghatározása diagnosztikai teszten alapul .....	7
2.5	Logisztikus regresszió, ahol a téves besorolások valószínűsége kovariánsfüggő .....	8
3.	<i>Új és újszerű eredmények</i> .....	9
4.	<i>A doktori értekezés témájában született publikációk</i> .....	10
4.1	Impakt faktorral rendelkező tudományos folyóiratban megjelent publikációk	10
4.2	Referált folyóiratban megjelent publikációk .....	10
4.3	Nemzetközi konferenciákon való részvétel .....	10

## 1. Bevezetés és célkitűzések

Hibás mérések előfordulhatnak szinte minden tudományágban. Folytonos változók esetén ezeket “mérési hibáknak” nevezzük, kategóriás változók, például betegségek státuszát jelző bináris változók esetén pedig téves besorolásnak. Míg a folytonos adatok mérési hibáinak kezelésére több módszer áll rendelkezésre, a kategóriás változók téves besorolásai kevesebb figyelemben részesültek, kevesebb tanulmány és módszer foglalkozik velük. Annak ellenére, hogy az utóbbi időben egyre többen hívják fel a figyelmet a téves besorolások figyelembevételére, a megfelelő elemzési módszerek összetettebb modellek esetén továbbra is korlátozottan érhetők el.

Kutatásaink célja kettős volt. Egyrészt, elméleti-analitikus és szimulációs technikák kombinálásával kiterjesztettünk néhány, már létező statisztikai módszert, amelyek korábban nem voltak használhatók téves besorolások előfordulása esetén, hogy ilyen helyzetekben is alkalmazhatók legyenek. Ezt a megközelítést alkalmaztuk a logisztikus regresszió már meglévő módszerének kiterjesztésére, valamint már elérhető konfidencia-intervallum (KI) szerkesztő módszerek kombinálására, hogy a téves besorolások kezelésére való kiterjesztésük révén új alkalmazási területekkel bővíljenek.

Másrészt, új módszereket dolgoztunk ki néhány összetett kutatási probléma kezelésére. Kidolgoztunk egy új profil-likelihood módszeren alapuló KI-t betegségek valódi prevalenciájának becsléséhez olyan esetekre, amikor a betegség diagnosztizáláshoz használt teszt szenzitivitását (Se) és specificitását (Sp) független validációs mintákból becsüljük. Kifejlesztettünk mintaelemszám-számítási módszereket több különböző bináris változóval végzett statisztikai teszthez, valamint egy mintaméret-újraszámolási módszert a népszerű “promising zone” adaptív klinikai vizsgálati elrendezéshez.

Kutatásaink során innovatív módszereken dolgoztunk a statisztikai elemzésekben előforduló téves besorolások által jelentett kihívások

kezelésére, különös tekintettel az orvostudomány és az epidemiológia területén. Az új módszertanok kifejlesztésével kiterjesztettük a már elérhető statisztikai eszközök tárházát azok számára, akik diagnosztikai tesztekkel és bináris változókkal dolgoznak. Minden módszert részletesen megvizsgáltunk szimulációk és való életből származó példák segítségével is, így alátámasztva tényleges gyakorlati hasznukat és robusztusságukat.

Az eredményeink megerősítik, hogy a téves besorolások figyelembevétele kritikus fontosságú nem csupán az adatelemzés során, de a vizsgálatok tervezési szakaszában is. Figyelmen kívül hagyásuk torzított becsléseket, csökkent statisztikai erőt eredményezhet és akár megalapozatlan statisztikai következtetésekhez is vezethet, különösen olyan erőforráshiányos helyzetekben, ahol az elérhető mintanagyság korlátozott. Javasolt módszereink egyszerűen alkalmazható és statisztikailag megalapozott megoldásokat kínálnak amelyek széles körben alkalmazhatók a klinikai vizsgálatoktól a közegészségügyi vizsgálatokig, hivatkozva a meglévő módszerek hiányosságai fölé.

## **2. Az eredmények összegzése**

### ***2.1 Profil-likelihood konfidencia intervallum betegség prevalenciájára, ha a diagnosztikai teszt hibával terhelt***

Kidolgoztunk egy új profil-likelihood konfidencia intervallumot (PLCI) egy betegség prevalencia becslésének olyan eseteire, amikor a használt diagnosztikai teszt szenzitivitását és specificitását a vizsgálati mintától független mintákból becsüljük. Az alapja a likelihood hányados teszt és egy korrekciót is beépítettünk az algoritmusba annak érdekében, hogy az intervallum lefedési valószínűsége extrém megfigyelések esetén is megfelelő legyen. Az intervallum lefedési valószínűségét és várható hosszát szimulációval értékeltük és két másik, ugyanerre a problémára javasolt módszerrel (Lang-Reiczigel és Flor és munkatársai módszereivel) hasonlítottuk össze.

A korrigált PLCI lefedési valószínűség és várható hossz tekintetében is jobban, vagy legalább úgy teljesít, mint az elérhető alternatívák. Valós példák segítségével is szemléltettük a módszer gyakorlati jelentőségét. Nagy mintájú vizsgálatok esetén mindhárom intervallum hasonló becsléseket eredményezett. Az alacsonyabb prevalenciának a becslését célzó vizsgálatok esetén azonban a PLCI és Flor módszere rövidebb intervallumokat eredményezett. Összességében, a PLCI a jó lefedési valószínűséget a rövid várható hosszal kombinálva hatékony alternatívája mindkét versenytársának.

### ***2.2 Becsült szenzitivitással és specificitással korrigált konfidencia intervallum a relatív kockázatra és a kockázat különbségre***

Javasoltunk egy konfidencia intervallumot két asszociációs mérőszám, a relatív kockázat és a kockázat különbség becslésére két korábban kidolgozott módszer, a Zou és Donner javasolta intervallum és Lang és Reiczigel módszerének kombinálásával. Az intervallumunk figyelembe veszi a diagnosztikai tesztek szenzitivitás és specificitás értékeinek esetleges pontatlanságait, ami különösen

fontos olyan vizsgálatoknál, amelyek során diagnosztikai gyors tesztek használunk, mint például a COVID-19 lakossági teszteléséhez használt eszközök, amelyek használata esetén gyakran előfordulhat téves diagnózis.

A módszer egyszerű, számítási igénye alacsony, az intervallum lefedési valószínűségei pedig a nominális értékhez közeli, kivéve olyan esetekben, amikor legalább egy prevalencia (vagy egy klinikai vizsgálatban az egyik válaszadási arány) szélsőséges (0-hoz közeli). Példákon keresztül illusztráltuk az intervallum gyakorlati jelentőségét, használata különösen ajánlott azokban az esetekben amikor egy vizsgálatban a mintanagyság alacsony, illetve a használt diagnosztikai teszt szenzitivitása és specificitása vélhetően pontatlan.

### ***2.3 Téves besorolások hatása egy- és kétmintás, valószínűségekre vonatkozó statisztikai próbák mintaelemszámára***

Kidolgoztunk mintaelemszám meghatározási módszereket a bináris végpontok esetén alkalmazható egy- és kétmintás statisztikai próbákhoz, amelyek figyelembe veszik a potenciális téves besorolásokat. A módszerek használhatók ismert szenzitivitás és specificitás mellett, illetve olyan esetekben is, amikor a diagnosztikai teszt paraméterei becslések, így helyességük bizonytalan.

Eredményeink itt is hangsúlyozzák a téves besorolások figyelembevételének fontosságát a vizsgálattervezés során, ennek figyelmen kívül hagyása ugyanis alacsonyabb statisztikai erejű vizsgálatokhoz vezet.

### ***2.4 A promising zone mintaelemszám újrameghatározásos elrendezés használata olyan klinikai vizsgálatokban, ahol egy bináris végpont meghatározása diagnosztikai teszten alapul***

Javasoltunk egy módosított mintaelemszám számítási formulát a “promising zone” nevezetű adaptív klinikai vizsgálati elrendezés interim elemzéseire olyan esetekre, amikor a vizsgálat binomiális változót használ végpontként és a végpont mérése olyan diagnosztikai

teszt segítségével történik, amelynek ismert szenzitivitása és specificitása.

Szimulációkkal alátámasztottuk, hogy a módszerünk használata egyszerű, segítségével a “promising zone” adaptív elrendezésű klinikai vizsgálatok interim elemzése során a feltételes erő pontosan számolható és a kezelési hatás torzítatlanul becsülhető.

### ***2.5 Logisztikus regresszió, ahol a téves besorolások valószínűsége kovariánsfüggő***

Javasoltunk egy kiterjesztett logisztikus modellt, amely figyelembe veszi a téves besorolásokat és lehetővé teszi a szenzitivitás vagy a specificitás kovariánsfüggését, míg a másik diagnosztikai paraméter konstans. A szimulációs eredmények megerősítették, hogy a modell jól teljesít, a mintanagyság és a statisztikai erő a valódi paraméterektől és a prediktorok eloszlásától függ.

A modell olyan érzékeny kérdőíves adatok elemzésére használható, amikor a válaszok pontossága kovariánsfüggő. A módszer korlátai közé tartoznak a modell specifikációból eredő lehetséges torzítások, az instabilitás, amelyet a prediktorok közötti erős korrelációk okozhatnak, továbbá, hogy feltétele a logisztikus összefüggés.

### 3. Új és újszerű eredmények

1. Javasoltam egy új konfidencia intervallumot betegség prevalencia becslés olyan eseteire, amikor a betegség státuszának megállapításához használt diagnosztikai teszt szenzitivitása és specificitása validációs mintából becsült. Ez az intervallum mind lefedési valószínűség, mind várható hossz tekintetében felülmúlja a többi hasonló célra elérhető alternatívát.

2. Kidolgoztam egy új konfidencia intervallumot kockázat különbség és relatív kockázat becslésére, figyelembe véve a becslések alapját képező betegség státuszának megállapításához használt diagnosztikai teszt szenzitivitását és specificitását. A módszer könnyen alkalmazható, számítási igénye alacsony és jobban teljesít az egyéb, korrekció nélküli módszernél.

3. Egy új mintanagyság meghatározási eljárást javasoltam egy- és kétmintás, valószínűségekre vonatkozó statisztikai próbákra, amely figyelembe veszi a téves besorolásokat és a vizsgálatban várt lemorzsolódási arányt, így pedig segít megakadályozni a vizsgálatok statisztikai erejének jelentős csökkenését.

4. Az általam javított korrekciók az adaptív “promising zone” klinikai vizsgálati elrendezéshez lehetővé teszik a vizsgálatok adatminőség értékelő bizottságai számára, hogy torzítatlan becslést kapjanak a kezelési hatásról és pontos feltételes statisztikai erőt számítsanak az interim elemzések során.

5. Kiterjesztettem a logisztikus modellt azokra az esetekre, amikor a téves besorolások valószínűsége kovariánsfüggő. Ez a módosítás jelentős előnyökkel jár a meglévő módszerekkel szemben bizonyos érzékeny kérdőíves kutatási kérdések elemzése során, amely esetekben a válaszadók vonakodhatnak őszintén válaszolni a kérdésekre és az őszinteség mértéke feltételezhetően bizonyos kovariánsoktól függ.

## **4. A doktori értekezés témájában született publikációk**

### **4.1 Impakt faktorral rendelkező tudományos folyóiratban megjelent publikációk**

Hársfalvi, P., & Reiczigel, J. (2023). Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test. *Preventive Veterinary Medicine*, 214, 105886.

Hársfalvi, P., & Reiczigel, J. (2024). The effect of misclassification on sample size for one and two-sample tests with binary endpoints. *Journal of Biopharmaceutical Statistics*, 1-12.

### **4.2 Referált folyóiratban megjelent publikációk**

Hársfalvi, P., & Singer, J. (2023). Confidence limits for risk differences and risk ratios adjusted for estimated sensitivity and specificity. *Biostatistics & Epidemiology*, 7(1), 2183614.

Hársfalvi, P., & Reiczigel, J. (2023). Profil-likelihood konfidenciaintervallum alkalmazása betegség prevalenciájára, ha a diagnosztikai teszt hibával terhelt= Profile likelihood confidence interval for the prevalence assessed by an imperfect diagnostic test. *STATISZTIKAI SZEMLE*, 101(11), 1033-1043.

### **4.3 Nemzetközi konferenciákon való részvétel**

1. 9th Statistics and Biopharmacy Conference – Párizs - Poszter

A promising zone mintaelemszám újrameghatározásos elrendezés használata olyan klinikai vizsgálatokban, ahol a bináris végpont meghatározása diagnosztikai teszten alapul

2. 44th Annual Conference of the International Society for Clinical Biostatistics – Milánó – Poszter

Profil-likelihood konfidencia intervallum betegség prevalenciájára, ha a diagnosztikai teszt hibával terhelt