LATVIJAS UNIVERSITĀTE
FIZIKAS, MATEMĀTIKAS UN OPTOMETRIJAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

Paradoksi un maldīgi spriedumi statistikā

BAKALAURA DARBS

Autore: Svetlana Zariņa
Studentes apliecības nr. sz16022
Darba vadītājs: Mg. Artis Luguzis

RĪGA 2021

ANOTĀCIJA

Šajā darbā tiek aprakstīti vairāki paradoksi un maldīgie spriedumi statistikā, kurus ir būtiski izprast, lai izdarītu korektus secinājumus par datiem dažādās zinātnēs, piemēram, epidemoloģijā, tiesību zinātnē, sociālajās zinātnēs. Kaut arī liela daļa no šiem fenomeniem tika aplūkoti jau 20. gadsimtā, ar paradoksiem saistītie rezultāti vēl joprojām var būt intuitīvi nesaprotami, kaut arī no statistikas viedokļa tie ir korekti. Savukārt, maldīgie spriedumi, kuri var likties loģiski, bet noved pie nepareiza rezultāta, joprojām ir sastopami rakstos. Darbā fenomeni tiek aprakstīti un doti piemēri.

Atslēgvārdi: statistikas paradokss, maldīgie spriedumi, maldīgais spriedums ekoloģisko datu interpretācijā, simpsona paradokss, lorda paradokss, lindleja paradokss, spēlētāja maldīgais spriedums, prokurora maldīgais spriedums

ABSTRACT

This paper is dedicated to multiple statistical paradoxes and fallacies, which are important to be aware of in order to draw valid conclusions about data in various fields such as epidemiology, law and social sciences. Despite the fact that discussions about statistical paradoxes and fallacies started in 20th century already, results related with paradoxes are still not intuitive, however, they are correct from statistical point of view. On the other hand, fallacies seem logical, but to this day they still lead analysts to false inference. In this paper phenomena and examples of them are described.

Keywords: statistical paradox, statistical fallacy, ecological fallacy, simpsons paradox, lord paradox, lindley paradox, gamblers fallacy, prosecutors fallacy

Satura rādītājs

APZĪMĒJUMI

N - populācijas lielums

$P(A)$ - notikuma varbūtība

$P(A|B)$ - nosacītā varbūtība

$Cov(X_i, Y_i)$ - gadījuma lielumu kovariance

σ - standartnovirze

σ^2 - dispersija

$\forall$ - katram

$X \sim \mathcal{N}(\mu 0, \sigma^2)$- mainīgais X ir normāli sadalīts ar vidējo vērtību μ un standartnovirzi σ

IEVADS

Zinātniskajos pētījumos statistika tiek izmantota visa pētījuma laikā un tā ļauj veikt datu apstrādi un analīzi, kas palīdz pētniekam noskaidrot saistības starp fenomeniem un veikt korektu slēdzienizdarīšanu. Sakarā ar to, cik plaši tiek pielietota statistika dažādās jomās, piemēram, epidemoloģijā, tiesību zinātnē, socioloģijā, un ņēmot vērā to, ka biežī vien pētījuma veicējam nav padziļinātas zināšanas statistikā, datu analīze var būt sarežģīta un pētnieks var izdarīt maldīgus spriedumus vai būt pārsteigts par analīzes rezultātiem, ja saskaras ar statistikas paradoksu un pat atteikties no izdarītajiem slēdzieniem, jo nav uzticība datiem.

Kaut arī liela daļa no šiem fenomeniem tika aplūkoti jau 20. gadsimtā[1][8][10], mūsdienās joprojām tiek izdarīti nepareizi secinājumi maldīgo spriedumu ietekmē un daudzi paradoksi pētniekiem nav zināmi.

Bakalaura darba galvenais mērķis ir informēt par statistikas paradoksiem un maldīgajiem spriedumiem un pievērst uzmanību šai aktuālai tēmai.

Darba izstrādei izvirzīti sekojoši uzdevumi:

1. Izpētīt, kādi ir populārākie paradoksi un maldīgie spriedumi statistikā un to aktualitāti;
2. Iepazīties ar izvēlētajiem statistikas paradoksiem un maldīgajiem spriedumiem un aprakstīt tos, pievienojot matemātisko aprakstu;
3. Aprakstīt vai uzģenerēt piemērus fenomeniem.

Darbs sastāv no ievada, 6 nodaļām, secinājumiem, izmantotās literatūras saraksta un pielikumiem. 1.nodaļā tiek apskatīts maldīgais spriedums ekoloģisko datu interpretācijā un aprakstītas divas metodes, kuras var izmantot, strādājot ar ekoloģiskajiem datiem, lai iegūtu korektus secinājumus. 2. nodaļā ir apskatīts Simpsona paradokss un tā variācijas. 3. nodaļā apskatīts Lorda paradokss un paskaidrots, kāpēc rodas atšķirība ANCOVA un divu izlašu t-testa rezultātos.4. nodaļā apskatīts Lindleja paradokss un ievietots īss apraksts par Baiesa un klasisko statistiku labākai paradoksa izprašanai. 5. nodaļā apskatīts spēlētāja maldīgais spriedums un 6. nodaļā prokurora maldīgais spriedums. Visu aprēķinu rezultāti iegūti, izmantojot programmatūru R un veicot aprēķinus.

1. Maldīgais spriedums ekoloģisko datu interpretācijā

1.1 Ekoloģiskā slēdzienizdarīšana

Ekoloģiskā slēdzienizdarīšana ir process, kad no agregētiem datiem (vēsturiski-ekoloģiskajiem datiem) izdara secinājumus par indivīdu uzvedību. Vajadzība izdarīt secinājumus par indivīdu uzvedību no agregētiem datiem ir gan politikā, kad vēlēšanu laikā indivīda līmeņa dati nav pieejami, gan epidemoloģijā, gan ģeogrāfijā, vesture, kad indivīda dati nav pieejami un citās zinātnēs un tā jau ilgu laiku ir bijusi viena no svarīgākajām statistikas problēmām pieminētajās zinātnēs. Šādu datu izmantošanu apgrūtina tas, ka daudz dažādas sakarības indivīda līmenī var veidot vienādu rādītāju, kad dati tiek agregēti jeb agregācijas dēļ zūd svarīga informācija[3].

Indivīda līmeņa dati- visdetalizētākie iegūtie dati par vienu analīzei pakļauto vienību.

1. tabula **Indivīda līmeņa dati**

Vārds	Grupa	Ienākumi, tūkst. EUR
Cilvēks 1	Latvijā dzimušie	1.5
Cilvēks 2	Latvijā dzimušie	2.5
Cilvēks 3	Latvijā dzimušie	2
Cilvēks 4	Latvijā dzimušie	2.7
Cilvēks 5	Latvijā dzimušie	1.9
Cilvēks 6	Ārpus Latvijas dzimušie	2.5
Cilvēks 7	Ārpus Latvijas dzimušie	1.1
Cilvēks 8	Ārpus Latvijas dzimušie	1.3

Agregēti (apkopoti, ekoloģiskie) dati – dati, kas apkopoti atbilstoši noteiktiem vienotiem raksturlielumiem (īpašībām), tie var būt gan gala rezultāts (rādītāju aprēķins), gan starprezultāts. Piemēram, indivīdi ir grupēti pēc dzīvesvietas, tautības vai rakstpratības.

2. tabula **Agregēti dati**

Grupa	Vidējie ienākumi, tūkst. EUR
Latvijā dzimušie	2.12
Ārpus Latvijas dzimušie	1.63

1.2 Maldīgā spieduma apraksts

Ekoloģiskais maldīgais spriedums ir pieņēmums, ka grupā novērotās sakarības noteikti atspoguļo sakarības indivīdu līmenī.

V. Robinsons 1950. gadā pievērsa uzmanību ekoloģiskajam maldīgajam spiedumam, uzsverot atšķirību starp korelāciju grupā un individuālo korelāciju un vērsa analītiķu uzmanību uz to, ka ar metodēm, kuras bija pieejamas tajā laikā korekta ekoloģiskā slēdzienizdarīšana nav iespējama un aicināja analītiķus neizmantot agregētos datus priekš secinājumiem par indivīdiem[1]. Kaut arī šis maldīgais spriedums jau sen tiek apspriests, tas joprojām atkārtojas dažādos rakstos [19].

Piemēram, Eiropā 19. gadsimtā pašnāvību rādītājs bija augstāks valstīs, kur dzīvo vairāk protestanti [2] un bija izdarīts secinājums, ka protenstantu ticība paaugstina pašnāvību risku. Šis ir piemērs ekoloģiskajam maldīgajam spriedumam, kad tiek uzskatīts, ka sakarības, kuras ir novērotas agregētos datos ir attiecināmas arī uz indivīdiem. Netika ņemts vērā kāda reliģija bija katram indivīdam, kurš izdarīja pašnāvību, bet tika salīdzināti divi rādītāji- pašnāvību rādītājs un reliģija valstī. Protestantu valstis atšķirās no Katoļu valstīm pēc daudzām pazīmēm, ne tikai reliģijas- arī šī jaucējfaktoru problēma, kurai vajag pievērst uzmanību katra pētījuma ietvaros- netika ņemta vērā.

1.3 Metodes darbā ar ekoloģiskajiem datiem

Ir dažādi iemesli, kāpēc nākas saskarties ar ekoloģisko maldīgo spriedumu, piemēram, jaucējfaktori, nepareiza izlases izveidošana, informācijas zudums sakarā ar nepareizu agregāciju un citi, tāpēc pētniekiem ir svarīgi būt informētiem par ekoloģisko maldīgo spriedumu un zināt kā strādāt ar ekoloģiskajiem datiem.

Divas vecākās un visplašāk pielietojamās metodes[4] , strādājot ar ekoloģiskajiem datiem ir robežu metode un ekoloģiskā regresija. Šī darba ietvaros apskatīsim tikai šīs 2 metodes.

Visvienkāršāk ir izskaidrot metodes uz reāla piemēra. Uzdevums ir noteikt, cik lielai daļai cilvēku no ārpus ASV dzimušajiem ir augsti ienākumi (virs 50.000 $ gadā). Šeit ir

runa par 2 novērotajiem mainīgajiem- T_i un X_i un diviem mainīgajiem, par kuriem nav datu un kuri interesē pētnieku- B_i^a un B_i^{na}.

3. tabula **Metodes formulas**

	Augsti ienākumi	Zemi ienākumi	Kopā
ASV dzimušie	B_i^a	$1 - B_i^a$	X_i
Ārpus ASV dzimušie	B_i^{na}	$1 - B_i^{na}$	$1 - X_i$
Kopā	T_i	$1 - T_i$	N_i

3. Tabulā attēloti sekojoši dati:

- T_i - iedzīvotāju daļa, kuriem ir augsti ienākumi;
- X_i - iedzīvotāju daļa, kuri ir dzimuši ASV;
- B_i^a - ASV dzimušo iedzīvotāju daļa, kuriem ir augsti ienākumi;
- B_i^{na} - Ārpus ASV dzimušo iedzīvotāju daļa, kuriem ir augsti ienākumi.

1.31. Robežu metode

Robežu metode jeb *Method of bounds* nozīmē, ka pētnieks aizvieto nezināmos B_i^a un B_i^{na} 1. tabulā ar iespējamiem vērtību intervāliem, kurus var izsecināt no marginālajiem tabulas skaitļiem pēc formulām (1.1) un (1.2) [4].

$$B_i^a \in \left[\max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right), \min\left(\frac{T_i}{X_i}, 1\right)\right] \tag{1.1}$$

$$B_i^{na} \in \left[\max\left(0, \frac{T_i - X_i}{1 - X_i}\right), \min\left(\frac{T_i}{1 - X_i}, 1\right)\right] \tag{1.2}$$

Ievieto dotos[5] skaitļus tabulā 3.

4. tabula **1. uzdevums- absolūtie skaitļi**

	Augsti ienākumi	Zemi ienākumi	Kopā
ASV dzimušie	B_i^a	$1 - B_i^a$	3 066 000
Ārpus ASV dzimušie	B_i^{na}	$1 - B_i^{na}$	263 000
Kopā	1 146 000	2 182 000	3 329 000

Pārveido absolūtos skaitļus procentos.

5. tabula **1. uzdevums- relatīvie skaitļi**

	Augsti ienākumi	Zemi ienākumi	Kopā
ASV dzimušie	B_i^a	$1 - B_i^a$	0.92
Ārpus ASV dzimušie	B_i^{na}	$1 - B_i^{na}$	0.079
Kopā	0.344	0.6554	3 329 000

Kā redzams no tabulas 5. marginālie skaitļi ir zināmi, bet iekšējās šūnas nav zināmas. ASV dzimušo iedzīvotāju daļa noteikti pieder intervālam [0,1] jeb tā vērtība ir no 0% līdz 100%. B_i^a aprēķins ir sekojošs:

Atradīsim apakšējo robežu no formulas (1.1)

$$\begin{aligned}
\max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right) &= \max\left((0, \frac{\text{Kopā \% ar augstiem ieņēmumiem-Visi \% dzimušie ārpus ASV}}{\text{Visi \% dzimušie ASV}}\right) \\
&= \max\left(0, \frac{0.344 - (1 - 0.92)}{0.92}\right) \\
&= \max(0, 0.28) \\
&= 0.28 \\
&= 28\%.
\end{aligned}$$

Šis ir vienkārši intuitīvi izskaidrojams- ja no visiem iedzīvotājiem ar augstiem ieņēmumiem atņem visu ārpus ASV dzimušo skaitu (gan ar augstiem ieņēmumiem, gan ar zemiem), tad rezultātā noteikti tiek iekļauti tikai ASV dzimušie ar augstiem ieņēmumiem.

Atradīsim augšējo robežu no formulas (1.1)

$$\begin{aligned}
\min\left(\frac{T_i}{X_i}, 1\right) &= \min\left(\frac{\text{Kopā \% ar augstiem ieņēmumiem}}{\text{Visi \% dzimušie ASV}}, 1\right) \\
&= \min\left(\frac{0.344}{0.92}, 1\right) \\
&= \min(0.37, 1) \\
&= 0.37 \\
&= 37\%.
\end{aligned}$$

ASV dzimušo procents ar augstiem ieņēmumiem nevar būt lielāks par kopējo iedzīvotāju daļu ar augstiem ieņēmumiem.

Ārzemēs dzimušo ar augstiem ienākumiem iedzīvotāju procentam ir jābūt starp 28% un 37%. Taču bieži vien, izmantojot šo metodi, intervāls ir pārāk plats un tādēļ nav informatīvs.

1.32. Ekoloģiskā regresija

Citos avotos [6] piedāvāja risināt šo uzdevumu jeb novērtēt B_i^a un B_i^{na}, izmantojot vienkāršu regresijas modeli:

$$T_i = X_i \cdot B_i^a + (1 - X_i) \cdot B_i^{na} \tag{1.3}$$

Novērtēsim B_i^a un B_i^{na} ar šo metodi. Sākumā 1.3 modelī tiek ievietoti jau zināmi skaitļi:

$$0.344 = 0.92 \cdot B_i^a + (1 - 0.92) \cdot B_i^{na}$$

B_i^{na} ievieto 0, pieņemot, ka nav neviens ārpus ASV dzimis iedzīvotājs ar augstiem ienākumiem. Tad $0.344 = 0.92 \cdot B_i^a + (1 - 0.92) \cdot 0 = \frac{0.344}{0.92} = 0.37$. Atrasta augšējā robeža un rezultāts sakrīt ar rezultātu, kurš iegūts, pielietojot robežu metodi. Tālāk B_i^{na} ievieto 1, pieņemot, ka visi ārpus ASV dzimušie iedzīvotāji ir ar augstiem ienākumiem. Tad $0.344 = 0.92 \cdot B_i^a + (1 - 0.92) \cdot 1 = \frac{0.264}{0.92} = 0.28$. Atrasta apakšējā robeža un tā sakrīt ar robežu metodi.

1.4 Piemērs

Viens ievērojams piemērs ir sakarība starp piedzimšanu valstī, par kuru tiek apkopota statistika un rakstpratību. 1950.gadā Robinsons [1]visiem 48 ASV štatiem ir aprēķinājis, cik procenti no populācijas ir dzimuši ASV un cik procenti no populācijas virs 10 gadiem spēj rakstīt un lasīt. Korelācijas starp 48 pāriem bija 0.53. To sauc par ekoloģisko korelāciju, jo analīzes subjekts ir cilvēku grupa, nevis katrs indivīds atsevišķi. Ekoloģiskā korelācija nosaka, ka ārpus ASV dzimušie ar lielāku varbūtību ir rakstpratīgi nekā vietējie. Bet realitātē indivīdu līmenī aprēķinātā korelācija ir -0.11 un ekoloģiskā korelācija rada nepareizu priekšstatu. Ekoloģiskās korelācijas zīme ir pozitīva, jo ārpus ASV dzimušie visbiežāk dzīvo štatos, kur cilvēkiem ir salīdzinoši augstāka izglītība.

2. Simpsona paradokss

2.1 Apraksts

Simpsona paradokss ir fenomens statistikā, kad saistība starp mainīgajiem X un Y mainās uz pretējo, pazūd vai atšķiras no saistības populācijā, kad tiek pievienots mainīgais Z, neskatoties uz mainīgā Z vērtību. Ja mēs sadalam populāciju apakšgrupās, kur katra apakšgrupa reprezentē mainīgā Z vērtību, var novērot, ka saistība jeb sakarība starp X un Y ir pretēja, salīdzinot ar sakarību kāda ir, kad izskatam visu populāciju, nepievienojot Z mainīgo.

Fenomens tika apskatīt jau kad 1899. gadā iznāca K. Pīrsona un pāris gadus vēlāk Dž. Jula raksts, bet tieši pēc Simpsona 1951.gada raksta *"The interpretation of interaction in contingency tables"* [8], kurā tika izskatīta šādas saistību virziena izmaiņas, šim fenomenam tika piešķirts nosaukums "Simpsona paradokss". No statistikas viedokļa šis paradokss ir izskaidrojams, tomēr tas nav intuitīvs un bieži pārsteidz pētniekus arī mūsdienās, piemēram, 2020. gada COVID-19 Beļģijas pētījuma datos, kur jebkurā vecuma grupā saslimušajiem vīriešiem bija augstāks mirstības koeficients, bet, neiekļaujot analīzē vecumu grupas, tas bija augstāks sievietēm[15].

2.2 Matemātiskais apraksts

Izskatīsim piemēru, kuru pats Simpsons prezentēja savā rakstā[7]. Tabulā 6. attēlots ārstēšanas rezultātu kopsavilkums visai populācijai (N=52), kā arī atsevišķi sievietēm un vīriešiem.

6. tabula **Simpsona piemērs**

	Visi, N=52			Vīrieši, M,N=20			Sievietes, $\neg M$, N=32		
Grupa	Izveseļojās (S)	Neizveseļojās ($\neg S$)	Izveseļošanās %	Izveseļojās (S)	Neizveseļojās ($\neg S$)	Izveseļošanās %	Izveseļojās (S)	Neizveseļojās ($\neg S$)	Izveseļošanās %
Ārstēšana (T)	20	20	50%	8	5	61%	12	15	44%
Kontrole ($\neg T$)	6	6	50 %	4	3	57%	2	3	40%

Ārstēšana izskatās neefektīva kopējā populācijas līmenī (gan ārstējoties, gan neārstējoties rezultāts bija vienāds- 50% no katram grupas izveseļojās), bet tā ir efektīva, kad atsevišķi izskatas sievietes un vīriešus (61% pret 57% vīriešiem, 44% pret 40% sievietēm). Kad uzraksta nosacītās varbūtības, kur T-ārstēšana jeb treatment, M-vīriešu

apakšgrupa, S-atveseļošanās, tad iegūst, ka

$$\begin{aligned} p\left(S \mid T\right) &= p\left(S \mid \neg T\right) \\ \text{, bet tajā pašā laikā } p\left(S \mid T, M\right) &> p\left(S \mid \neg T, M\right) \\ \text{un } p\left(S \mid T, \neg M\right) &> p\left(S \mid \neg T, \neg M\right) \end{aligned} \tag{2.1}$$

Divi apgalvojumi ir būtiski, lai izprastu, kāpēc pozitīva saistība pazūd, kad izskatam visu populāciju. Pirmkārt, atveseļošanās % ir augstāks arī vīriešiem, kuriem nav bijusi ārstēšana, ja salīdzina ar sievietēm, kurām bija ārstēšana (57% pret 44%). No šī var secināt, ka arī dzimums ir svarīgs faktors, no kura ir atkarīga izveseļošanās. Otrkārt, sievietes veido lielāko daļu no grupas, kura bija pakļauta ārstēšanai (27 sievietēs, 13 vīrieši), savukārt vīrieši veido lielāko daļu no grupas, kura netika pakļauta ārstēšanai (7 pret 5). Tātad, populācijā nav korelācija starp ārstēšanu un izveseļošanos, jo vīrieši ar lielāku varbūtību izārstēsies un ar mazāku varbūtību būs grupā, kura tika ārstēta. Tas ir vienkārši izskaidrojams, ja izteikt izveseļošanās varbūtības pie nosacījuma, vai bija ārstēšana un kāda ir apakšgrupa. Kopējā atveseļošanās varbūtība jeb pilna varbūtība [7] var būt izteikta kā svērtais vidējais starp atveseļošanās varbūtībām grupās:

$$p\left(S \mid T\right) = p\left(S \mid T, M\right) p\left(M \mid T\right) + p\left(S \mid T, \neg M\right) p\left(\neg M \mid T\right) \tag{2.2}$$

$$p\left(S \mid \neg T\right) = p\left(S \mid \neg T, M\right) p\left(M \mid \neg T\right) + p\left(S \mid \neg T, \neg M\right) p\left(\neg M \mid \neg T\right) \tag{2.3}$$

Simpsona paradokss var parādītiem dažādos datu tipos, bet parasti tas tiek atspoguļots 2x2 kontingences tabulā[7]. Pieņem, ka $D_i = (a_i, b_i, c_i, d_i)$ ir 4 dimensiju reālo skaitļu vektors, kurš reprezentē 2x2 kontingences tabulu par ārstēšanu un atveseļošanos i-tajā apakšgrupā un

$$D = \sum_{i=1}^{N} D_i = (\sum a_i \sum b_i \sum c_i \sum d_i) \tag{2.4}$$

ir agregētie dati starp N apakšgrupām. Vispārīga datu reprezentācija ir attēlota 7. tabulā.

7. tabula **Abstraktas 2x2 kontingences tabulas reprezentācija ar divām apakšgrupām D_1 un D_2**

	Visi, $D = D_1 + D_2$		Apakšgrupa D_1		Apakšgrupa D_2	
Grupa	Izveseļojās (S)	Neizveseļojās ($\neg S$)	Izveseļojās (S)	Neizveseļojās ($\neg S$)	Izveseļojās (S)	Neizveseļojās ($\neg S$)
Ārstēšana (T)	$a_1 + a_2$	$b_1 + b_2$	a_1	b_1	a_2	b_2
Kontrole ($\neg T$)	$c_1 + c_2$	$d_1 + d_2$	c_1	d_1	c_2	d_2

Pieņem, ka $\alpha(D_i)$ parāda, cik stipra ir saistība starp T un S apakšgrupā Di. $\alpha(D_i)$ =0 nozīmē, ka nav saistības starp mainīgajiem, $\alpha(D_i) > 0$, ka ir pozitīva saistība, $\alpha(D_i) < 0$- negatīva. Aprakstītās saistības var rasties tikai pie šādiem nosacījumiem:

$$\alpha(D_i) = \begin{cases} > 0, \text{ ja } a_i d_i > b_i c_i; \\ = 0, \text{ ja } a_i d_i = b_i c_i; \\ < 0, \text{ ja } a_i d_i < b_i c_i; \end{cases} \tag{2.5}$$

Piemēram, nosacījums $\alpha_i * d_i > b_i * c_i$ apzīmē, ka atveseļošanās varbūtība ārstētai grupai ir lielāka nekā neārstētai grupai:

$$\frac{a_i}{(a_i + b_i)} > \frac{c_i}{(c_i + d_i)} \tag{2.6}$$

2.21. Saistības virziena izmaiņa

Var novērot, ka 6. tabulā ir gadījums, kad $a(D)$ =0, neskatoties uz to, ka $a(D_1) > 0$ un $a(D_2) > 0$. Tas ir gadījums, kad saistība tiek apgriezta *(Association Reversal jeb AR)*. AR rodas tikai, kad ir tāda populācija, kur visās apakšgrupās saistība ir vai nu pozitīva, vai negatīva, vai vienāda ar nulli un sakarība populācijā nesakrīt ar sakarība apakšgrupās. Uzrakstot to matemātiski, datu kopa $D = \sum_{i=1}^{N} D_i$N atbilst vienam no diviem nosacījumiem [7]:

$$\begin{aligned} &\alpha(D) \leq 0 \quad \text{un} \quad \alpha(D_i) \geq 0 \quad \forall 1 \leq i \leq N \\ &\alpha(D) \geq 0 \quad \text{un} \quad \alpha(D_i) \leq 0 \quad \forall 1 \leq i \leq N \end{aligned} \tag{2.8}$$

,kur vismaz viena no nevienādībām ir stingra.

2.22. Jūla saistības paradokss

Vēl viena nozīmīga Simpsona paradoksa variācija rodas tad, kad katrā apakšgrupā nepastāv saistība starp mainīgajiem, bet populācijā pastāv saistība:

$$\alpha(D_i) = 0 \quad \forall 1 \leq i \leq N \quad \text{,bet} \quad \alpha(D) \neq 0 \tag{2.10}$$

Šo speciālo gadījumu mēdz saukt arī par Jūla saistības paradoksu *(Yule's Association Paradox jeb YAP)*. Piemēram, gulēšana drēbēs korelē ar galvassāpēm nākamajā rītā. Toties, ja stratificēt datus atbilstoši izlietotajam alkoholam iepriekšējā vakarā, sakarība pazūd, jo pie vienāda izdzertā alkohola daudzuma cilvēkiem, kuri novilka drēbes galva sāpēc tikpat stipri cik cilvēkiem, kuri nenovilka drēbes. Cits zināms Jūla paradoksa piemērs ir 1973. gada pētījums par iespējamo neobjektivitāti skolu uzņemšanā starp meitenēm un zēniem.

2.23. Mazāka vai lielāka saistība populācijā nekā apakšgrupās

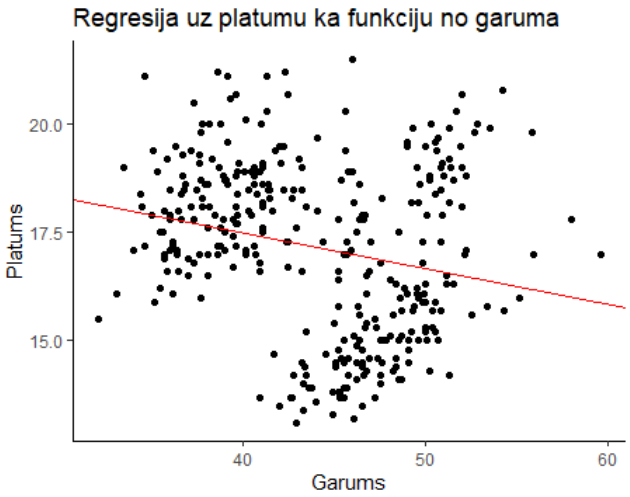
Un pati vispārīgākā Simpsona paradoksa variācija ir, kad saistība starp mainīgajiem populācijā ir mazāka vai lielāka nekā saistība starp mainīgajiem apakšgrupās. Tas gadījums tiek saukts arī par *Amalgamation Paradox jeb AMP*.

$$\alpha(D) > \max_{1 \leq i \leq N} \alpha(D_i) \quad \text{vai} \quad \alpha(D) < \min_{1 \leq i \leq N} \alpha(D_i) \tag{2.12}$$

2.3 Piemērs

Apskatīsim piemēru, kad regresijas zīme mainās uz pretējo, kad populāciju sadala apakšgrupās uz pakotnē *palmerpenguins* pieejamajiem datiem par 3 dažādu šķirņu pingvīniem. Ir apkopoti dati par 344 pingvīniem. Šajā piemērā izskatam pingvīna knābja garumu milimetros un knābja platums milimetros.

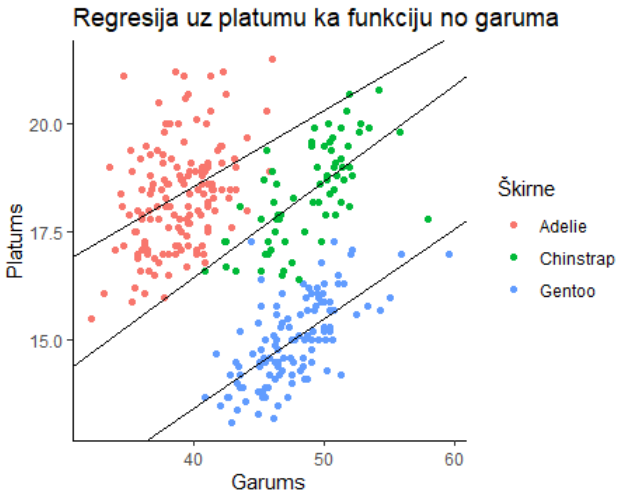
Ja apskatīt izkliedes grafiku knābja garumam un platumam un uztaisīt vienkāršu lineāro regresiju, tad var secināt, ka pastāv negatīva sakarība starp knābja garumu un platumu.



(a)

1. att. **Regresija uz knābja platumu kā funkciju no garuma**

No šī grafika var secināt, ka jo garāks knābis, jo mazāks platums, bet ja apskata katru pingvīna šķirni atsevišķi, tad iegūst pretējus rezultātus. Knābja platums palielinās katrai no trīs šķirnēm, palielinoties knābja garumam.



(a)

2. att. **Regresija uz platumu kā funkciju no garuma pa grupām**

3. Lorda paradokss

3.1 Apraksts

Lorda paradokss apzīmē rezultātu, kad saistība starp nepārtraukto iznākuma mainīgo un diskrēto mainīgo mainās uz pretējo, kad tiek ieviesta vēl viena nepārtraukta kovariate. Viens specifisks piemērs ir, kad divos secīgajos laika periodos izmēra vienu un to pašu nepārtraukto mainīgo. Piemēram, statistikā bieži ir vajadzība salīdzināt vairākas cilvēku grupas vairākos laika periodos un saprast vai parādās atšķirība starp grupām laika gaitā. Statistikā pastāv vairākas metodes kā veikt aprēķinu šādos gadījumos, bet pat priekš 2 grupām un 2 laika periodiem aprēķinu metodes izvēle nav vienkārša.

Zinātnieks F.M. Lords 1967. gadā uz datiem par studentu svaru pēc dzimumiem gada sākumā un gada beigās izpētīja divas analīzes metodes [10]. Pirmā metode bija vidējo salīdzināšana, izmantojot divi izlašu t-testu un otrā metode bija gala svara kovariācijas analīze jeb ANCOVA, izmantojot sākuma svaru kā kovariāti. Pielietojot t-testa metodi rezultātā var secināt, ka nav nekādas izmaiņas ne puišu, ne meiteņu grupā un nav arī atšķirības starp grupām. Kaut arī visu atsevišķu meiteņu un zēnu svars ir mainījies gada laikā, iespējams ievērojami, visas grupas svars nav mainījies. Var secināt, ka nav iemeslu uzskatīt, ka ir jāmaina skolas ēdienkarti vai jāpēta citi faktori, kuri var ietekmēt studentu svaru. Savukārt, ANCOVA rezultāti parāda, ka zēnu grupā svars ir palielinājies ievērojami vairāk nekā meiteņu grupā. Paradokss ir tajā, ka divas metodes dod atšķirīgu rezultātu.

3.2 Matemātiskais apraksts

3.21. Atšķirība parametros

Lorda paradoksa pamatā ir tas, ka interesējošais parametrs divu izlašu t-testā un interesējošais parametrs ANCOVA modelī nav vienādi [9]. Zemāk apskatīsim izvedumus no publikācijas [9]. Atšķirība parametros vai būt paskaidrota ar lineāri kombinēto mainīgo kovarianci.

Pieņemsim, ka $U_1, ..., U_n$ un $W_1, ..., W_m$ ir gadījuma lielumi un $L_1 = \sum_{i=1}^{n} a_i U_i, L_2 = \sum_{j=1}^{m} b_j W_j$ ir to lineārās kombinācijas. Tad kovariācija tiek rēķināta pēc formulas (3.1).

$$Cov(L_1, L_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(U_i, W_j) \tag{3.1}$$

Tā kā $\sigma^2(L_1) = Cov(L_1, L_1)$, tad pārveidojot formulu (3.1). dispersijas formula ir

$$\sigma^2(L_1) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(U_i, U_j) = \sum_{i=1}^{n} a_i^2 \sigma^2(U_i) + 2\sum_{i=1}^{n}\sum_{j>i} Cov(U_i, U_j) \tag{3.2}$$

No šī izveduma ir vienkārši paskaidrot atšķirību t-testa un ANCOVA modeļa testa parametros.

3.22. Divu izlašu t-tests

Pieņemsim, ka Z_i ir pirmais un Y_i ir otrais i-tā subjekta rezultāts. X_i ir grupas indikātors, kur X_i=0 0-tās grupas dalībniekiem un X_i=1 1. grupas dalībniekiem.

Divu izlašu t-tests var būt izteikts kā vienkāršais lineārais modelis.

$$D_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{3.3}$$

kur $D_i = Y_i - Z_i$ ir izmaiņa rezultātā un $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Vienādojumā (3.3) pētnieku interesējošais parametrs ir starpība starp divu grupu vidējām vertībām.

$$\beta_1 = E(D_i \mid X_i = 1) - E(D_i \mid X_i = 0) \tag{3.4}$$

Nulles hipotēze ir $H_0 : \beta_1 = 0$ un $H_1 : \beta_1 > 0.B_1$ var izteikt arī kā

$$\begin{aligned} \beta_1 =& \frac{Cov(X_i, D_i)}{\sigma^2(X_i)}, \text{ jo} \\ Cov(X_i, D_i) =& Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) = \\ & Cov(X_i, \beta_0) + Cov(X_i, \beta_1 X_i) + Cov(X_i, \epsilon_i) = \beta_1 \sigma^2(X_i) \end{aligned} \tag{3.5}$$

3.23. ANCOVA

ANCOVA modeli var aprakstīt ar šo vienādojumu

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \delta_i, \text{ ,kur } \delta_i \sim \mathcal{N}(0, \sigma^2) \tag{3.6}$$

Šajā modelī interesējošais parametrs ir γ_1 jeb starpība 2. rezultātā starp novērojumiem no 0. un 1. grupām ar vienādu 1. rezultātu. Nulles hipotēze ir $H_0 : \gamma_1 = 0$ un $H_1 : \gamma_1 > 0$. Alternatīvs veids kā aprakstīt ANCOVA modeli ir

$$D_i = \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1) Z_i + \delta_i, \tag{3.7}$$

kur no abām pusēm tika atņemts Z_i, ņēmot vērā, ka $D_i = Y_i - Z_i$.

$$\begin{aligned}\gamma_1 =& \frac{Cov(X_i, D_i) + (1-\gamma_1)Cov(X_i, Z_i)}{V(X_i)} \\ =& \beta_1 + (1-\gamma_2)(\frac{Cov(X_i, Z_i)}{\sigma^2(X_i)})\end{aligned} \tag{3.8}$$

Var uzrakstīt, ka $k_1 = \frac{Cov(X_i,Z_i)}{\sigma^2(X_i)} = E(Z_i \mid X_i = 1) - E(Z_i \mid X_i = 0)$,ko var interpretēt kā starpību vidējā 1.rezultātā, kad salīdzina 0. un 1. grupas.

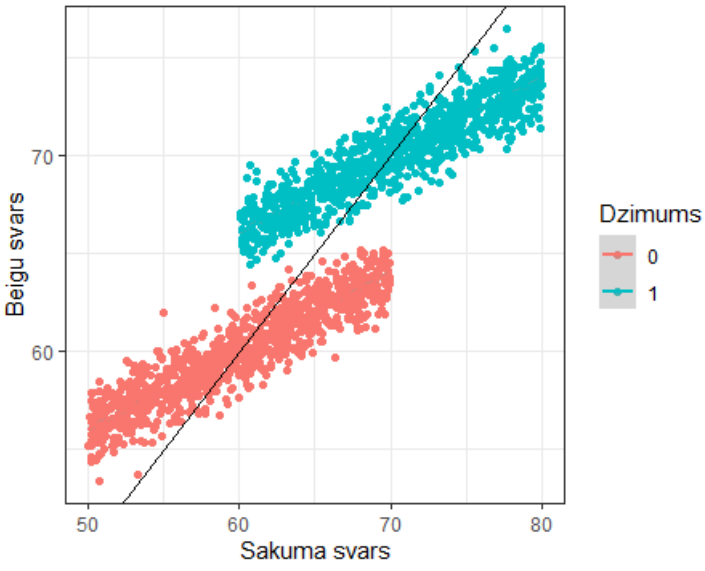
Kā aprakstīts, t-testa un ANCOVA modeļa intereses parametri ir atšķirīgi un starp tiem pastāv šāda saistība:

$$\begin{aligned}\gamma_1 =& \beta_1 + (1-\gamma_2)k_1 \\ \beta_1 =& \gamma_1 + (1-\beta_2)k_1\end{aligned} \tag{3.9}$$

Parametri sakrīt jeb $\gamma_1 = \beta_1$, ja k_1=0 vai γ_2=1. Pirmais nosacījums var būt izpildīts ar randomizāciju(veicot eksperimentu novērojuma vietā), bet otrais nosacījums $\gamma_2 = 1$ nevar būt ietekmēts, visbiežāk pirmais rezulāts ir pozitīvi korelēts ar otro rezultātu.

3.3 Piemērs

Apskata piemēru uz simulētajiem datiem. Studentu, kuri ir uz 45° līnijas svars nav mainījies. Ar Dzimums=0 apzīmē meitenes, Dzimums=1 ir zēni.



(a)

3. att. **Lorda paradokss**

Šajā piemērā t-tests nespēj noraidīt nulles hipotēzi pie nozīmības līmeņa 0.05, savukārt ANCOVA noraida nulles hipotēzi.

Veicot t-testu, pētnieks jautā "Vai ir starpība vidējā svara palielināšanās populācijā" un korekta atbilde ir nē. Savukārt, otrais statistiķis jautā "Vai ir sagaidāms, ka zēni uzņemsies svarā par vairāk kilogramiem nekā meitenes, ja sākotnēji viņi ir vienāda svara?" un korekta atbilde ir jā. Tātad abi varianti ir korekti un paradoksa atisinājums ir pareizi noteikt jautājumu, uz kuru grib atbildēt [11].

4. Lindleja paradokss

4.1 Baiesa un klasiskā (frekventistu) statistika

Kad tiek risināts secinošās statistikas uzdevums, piemēram, parametra novērtēšana, veiktas prognozes un salīdzināti modeļi, tiek pielietotas divas atšķirīgas pieejas aprēķiniem: Baiesa un Klasiskās jeb frekventistu statistikas metodes. Baiesa teorēma par nosacīto varbūtību ir fundamentāla baiesa statistikā, jo tā ļauj skaitliski izteikt, cik pamatoti ir ticēt kādai hipotēzei pie pierādījumiem, kuri šobrīd ir:

$$P(A \mid B) = \frac{(B \mid A) \cdot P(A)}{P(B)} \tag{4.1}$$

Ja piešķirt citu interpretāciju notikumiem formulā 4.1, tad

$$P(H \mid D) = \frac{(D \mid H) \cdot P(H)}{P(D)}, \text{kur} \tag{4.2}$$

H ir jebkura hipotēze, kuras varbūtība var būt datu ietekmēta,
D ir pierādījums jeb jauni dati, kuri netika izmantoti aprioras varbūtības aprēķinā,
$P(H)$ ir apriorā varbūtība jeb H varbūtības novērtējums pirms dati D ir izskatīti,
$P(D)$ ir totāla datu varbūtība, ņemot vērā visas iespējamās hipotēzes,
$P(H|D)$ ir aposteriora hipotēzes varbūtība jeb H varbūtības novērtējums, kad ir ņemts vērā D,
$P(D|H)$ iespējamība jeb varbūtība iegūt datus E, ja H ir patiesa.

Ja apriorā hipotēzes varbūtība un iespējamība ir zināmi visām hipotēzēm, tad Baiesa formula precīzi izrēķina aposterioru varbūtību. Tāds gadījums, piemēram, ir ja metam metamo kauliņu, kurš ir nejauši izvilkts no trauka, kura saturs bija iepriekš zināms. Bet lielākajā daļā eksperiementu apriorās hipotēžu varbūtības nav zināmas. Šajā gadījumā baiesisti izvēlās izdomāt aprioro varbūtību un frekventisti risina uzdevumu, pielietojot ticamības.

Baiesa slēdzienizdarīšanā izmanto gan hipotēzes, gan datu varbūtību un metodes balstās uz apriori varbūtību, kuru izvēlās pētnieks, kā arī iegūto datu varbūtību. Savukārt frekventistu slēdzienizdarīšanā netiek izmantotas vai izvēlētas hipotēzes varbūtības (nav apriorā vai aposteriorā varbūtība) un rezultāts ir atkarīgs no $P(D|H)$ ticamības novērotajiem un nenovērotajiem datiem un tādā veidā tiek nodrošināta ōbjektivitāte".

Frekventisti pieņem, ka kāda hipotēze ir patiesa un ka novērotie dati ir nākuši no kāda sadalījuma [12].

Sarunas par to, kuru metodi labāk izmantot turpinās vēl joprojām. Klasiskā statistika kopš 20. gadsimta ir populārāka un tajā tiek izmantoti tādi plaši zināmi paņēmieni kā p-vērtības, ticamības intervāli, bet baiesa statistiska mūsdienās arvien vairāk tiek izmantota mašīnmācīšanā, ģenētikā, kas iepriekš bija sarežģīti, jo datori nebija tik jaudīgi, cik šobrīd [12].

4.2 Apraksts

Lindleja jeb Džefrija-Lindleja paradokss parāda kā p-vērtības izmantošana hipotēzes pārbaudē klasiskajā statistikā var novest pie pretējiem rezultātiem, salīdzinot ar baiesa hipotēžu pārbaudes testu. D. Lindlejs 1957. gadā publicēja rakstu [13] par paradoksu, kurš jau bija aprakstīts H. Džefrija rakstā 1939. gadā[16]. Lindlejs raksta, ka ir izlase, kuras izmērs ir n un tā ir normāli sadalīta- $\mathcal{N}(\theta, \sigma^2)$ un dispersija- σ^2. Testējot nulles hipotēzi $H_0 : \theta = \theta_0$ (pret alternatīvo hipotēzi $H_1 : \theta \neq \theta_0$), var nonākt pie pretējiem secinājumiem, atkarībā kādu statistisko metodi izvēlēsies pētnieks. Tātad, zinot, ka

$$\overline{x_n} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n}) \text{ un } t_n = \frac{\overline{x_n} - \theta_0}{\sigma/\sqrt{n}},$$

tālāk iegūst p-vērtību, no kuras izdara secinājumus par nulles hipotēzi[12].

Savukārt, ja pielieto Baiesa metodes, tad ir svarīgs sekojošais vienādojums:

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{aposteriorā izredžu attiecība}} = \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{apriotā izredžu attiecība}} \cdot \underbrace{\frac{P(D \mid H_0)}{P(D \mid H_1)}}_{\text{Baiesa faktors}}$$

Vienādojuma kreisajā pusē ir aposteriorā izredžu attiecība jeb secinājumi par hipotēzēm H_0 un H_1 pēc datu D izskatīšanas. Labajā pusē viens no reizinātājiem ir apriorā izredžu attiecība jeb secinājumi par hipotēzēm H_0 un H_1 pirms datu D izskatīšanas. Otrs reizinātājs ir Baiesa faktors un no vienādojuma Baiesa faktoru var interpretēt kā faktoru, kurš izmaina aprioro izredžu attiecību pēc datu izskatīšanas. Baiesa faktors nosaka, cik lielāka varbūtība bija iegūt datus D,ja H_0 ir patiesa, salīdzinot ar H_1 un tāpēc bieži vien tieši Baiesa faktors tiek izmantots, kad tiek veikti Baiesa hipotēžu testi. Ja Baiesa faktors ir tuvu 1, tad dati D praktiski neietekmē mūsu secinājumus par varbūtībām. Parasti ja Baiesa faktors ir lielāks par 10, tad uzskata, ka ir stipri pierādījumi par labu H_0.

Un paradokss, kuru apraksta Lindlejs ir tajā, ka pie t_n fiksētas vērtības un gandrīz jebkuram θ apriorajam sadalījumam, Baiesa faktors tiecās uz bezgalību, kad n tiecās uz

bezgalību, savukārt p-vērtība paliek konsanta jebkuram n. Un var rasties situācija, ka pētnieks var uz 95% būt pārliecināts kā frekventists, ka $\theta \neq \theta_0$ un tajā pašā laikā būt uz 95% pārliecināts, ka $\theta = \theta_0$. Tas var gadīties, piemēram, ja $t_n = 1.96$ and n = 16 818, ja pieņem, ka apriorie svari ir 0.5.

Ir norādīts, ka triviālais paradoksa risinājums ir tāds, ka nav jāsagaida, ka skaitliskie rezultāti baiesa un klasiskās statistikas testos sakritīs, jo tie aprēķina atšķirīgas vērtības, tomēr joprojām nav panākta vienošanās par paradoksa korekto risinājumu un diskusijas par paradoksu turpinās.

4.3 Piemērs

Šajā apakšnodaļā paradokss tiks demonstrēts uz piemēra. Iedomāsimies pilsētu, kur noteiktā laika posmā piedzima 49581 zēni un 48870 meitenes. Zēnu daļa x ir $\frac{49581}{98451} = 0.5036$. Pieņemsim, ka piedzimušo zēnu skaitam ir binomiālais sadalījums ar parametru θ. Grib pārbaudīt, vai θ=0.5. Tātad, $H_0 : \theta = 0.5$ un $H_1 : \theta \neq 0.5$

Kā noskaidroja iepriekš, frekventistu pieeja balstās uz p-vērtības aprēķinu, noteikšanu kāda ir iespēja, ka zēnu daļa ir ne mazāka par x pie nosacījuma, ka H_0 ir patiess. Tā kā piedzimušo bērnu skaits ir liels,zēnu piedzimšanas daļai var izmantot aproksimāciju uz normālo sadalījumu $x \sim \mathcal{N}(\mu, \sigma^2)$, kur

$$\begin{aligned} \mu &= np = n\theta = 98451 \cdot 0.5 = 49225.5 \\ \sigma^2 &= n\theta(1-\theta) = 98451 \cdot 0.5 \cdot 0.5 = 24612.75 \end{aligned} \tag{4.3}$$

Tad iegūst, ka

$$\begin{aligned} & P(X \geq x \mid \mu = 49225,5) \\ &= \int_{x=49581}^{98451} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{u-\mu}{\sigma})^2/2}\, du \\ &= \int_{x=49581}^{98451} \frac{1}{\sqrt{2\pi 24612.75}} e^{-(\frac{u-49225.5}{24612.75})^2/2}\, du \\ &\approx 0,0117 \end{aligned} \tag{4.4}$$

Tātad H_0 hipotēzi var noradīt pie nozīmības līmeņa 0.05.

Izmantojot Baiesa metodi, pieņemam, ka neviena no hipotēzēm nav varbūtiskāka. Baiesa metode balstās uz aprioro varbūtību definēšanu $\pi(H_0) = \pi(H_1) = 0.5$ un vienmērīgo

θ sadalījumu. Tālāk tiek aprēķināta aposteriorā H_0 varbūtība, izmantojot Baiesa teorēmu.

$$
\begin{aligned}
P(H_0 \mid k) &= \frac{P(k \mid H_0)\pi(H_0)}{P(k \mid H_0)\pi(H_0) + P(k \mid H_1)\pi(H_1)} \\
P(k \mid H_0) &= \binom{n}{k}(0,5)^k(1-0.5)^{n-k} = \approx 1.95 \cdot 10^{-4} \approx 1,95 \cdot 0,0001 = 0,000195 \\
P(k \mid H_1) &= \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\,d\theta = \binom{n}{k}B(k)+1, n-k+1) \approx 1.02 \cdot 10^{-5} = 0,0000102
\end{aligned}
\tag{4.5}
$$

Ja ieliek iegūtās vērtības formulā, tad aposteriorā varbūtība $P(H_0 \mid k)$ $\approx$0.95, kas norāda uz to, ka H_0 ir varbūtiskāka par H_1. Tas arī ir paradoksa piemērs- divu metožu rezultāti ir atšķirīgi.

5. Spēlētāja maldīgais spriedums

5.1 Apraksts

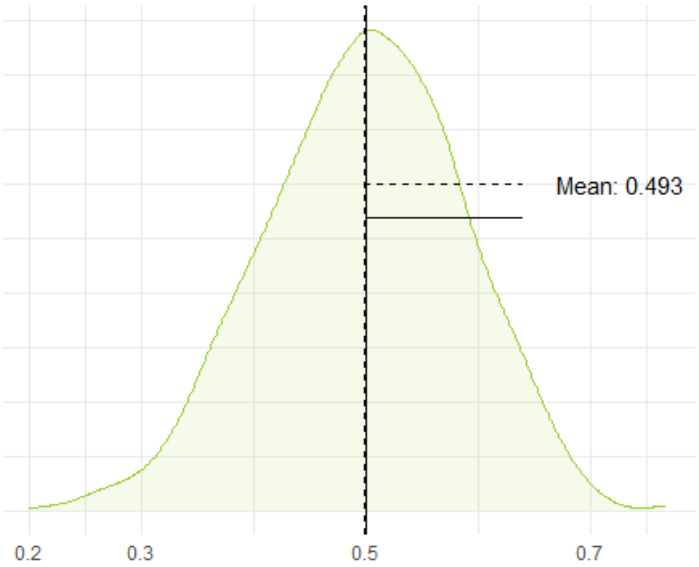
Franču matemātiķis P.S. Laplass bija pirmais, kas sāka runāt par gamblers fallacy savā esejā "Illusions in the Estimation of Probabilities" [17], vēršot uzmanību uz piemēriem no azartspēlēm un loterijām. Kopš tā laika vairāki uzvedības pētījumi ir apstiprinājuši, ka gambler fallacy pastāv un ietekmē cilvēku lēmumus. Cilvēkiem nākas saskarties ar spēlētāja maldīgo spriedumu, kad tie sagaida sistemātisku rezultātu, novērojot nejaušus secīgus notikumus. Piemēram, kad met monētu uzskata, ka, ja uzkrita reverss, tad nākamreiz ar lielāku varbūtību uzkritīs averss.Spēlētāja maldīgais spriedums rodas no ticības "mazo skaitļu likumam".Mazo skaitļu likums ir termins, kuru ieviesa 1971. gadā, lai aprakstītu kā cilvēki pārspīlē, cik stipri maza izlase līdzinās populācijai, no kuras tā ir atlasīta[18]. Kaut arī metot monētu 1000 reizes, avers uzkritīs apmēram 50% gadījumu, nav pamats uzskatīt, ka tāds pats rezultātu sadalījums būs arī pēc 10 metieniem.

5.2 Piemēri

Ar spēlētāja maldīgo spriedumu var saskarties arī pieņemot svarīgus lēmumus. Viens no piemēriem ir dabas katastrofas- cilvēks var uzskatīt, ka tūrisms uz kādu vietu ir dross, jo jau ilgu laiku tur nav bijušas dabas katastrofas. Vai potenciāli vēl bīstamāk ir uzskatīt, ka pēc lielas zemestrīces ir droši ceļot, jo vēl viena zemestrīce visticamāk nebūs.
Vēl viens piemērs ir tiesu process- ja tiesnesis vairākus aizdomās turāmos teroristus ir pasludinājis par nevainīgajiem, tad viņa lēmums par nākamo aizdomās turāmo var būs iespaidots ar uzskatu, ka starp tik daudziem aizdomās turamajiem kaut vienam ir jābūt bīstamam teroristam. Vai arī gluži otrādi- tiesnesis var uzskatīt, ka 5 teroristi pēc kārtas ir pārāk daudz un var uzskats, ka 5. terorists nav vainīgs.
Kad pētīja aizdevumu izsniedzēju darbu, tika atklāts, ka izsniedzēja lēmumi ir negatīvi autokorelēti un izsniedzējs biežāk noraida pieteikumu, ja pirms tam ir apstiprinājis vairākus pieteikumus. Kaut arī katrs aizdevums un aizdevuma pieteicējs ir individuāls, mazāk pieredzējuši lēmumu pieņēmēji uztver pieteikumus kā savstarpēji saistītus un saskaras ar gamblers fallacy.

Var simulēt maldīgo spriedumu, lai pārliecinātos, ka sākotnēji bieži novērojot vienu

iznākumu, nevar secināt, ka varam viennozīmīgi izdarīt secinājumu, ka turpmāk iznākumi būs pretēji. 1000000 reizes tiek simulēta monētas mešana 40 reizes. No tām 935 reizes pirmās 10 reizes tika uzmests averss. Vai tas nozīmē, ka pārējās reizes tiks biežāk uzmests reverss? Tiek novēroti 30 atlikušie metieni 935 simulācijās. Varam novērot, ka vidēji 49% no atlikušajiem metieniem bija reverss.



(a)

4. att. **Piemērs**

6. Prokurora maldīgais spriedums

6.1 Apraksts

Prokurora maldīgais spriedums ir maldīgais uzskats, ka $P(A|B) = P(B|A)$, kuram tika piešķirts šis nosaukums pēc C. Tomsona un E. Šumana raksta 1987. gadā [21]. Tas rodas no varbūtības teorijas neizprašanas un tam ir liela nozīme tiesību zinātnē un epidemoloģijā. Kad prokurors ir ieguvis pierādījumus (piemēram, DNS sakritība) un eksperts izsaka viedokli, ka ir maza varbūtība iegūt šos pierādījumus, ja apsūdzētāis ir nevainīgs, tad var tikt pieļauts prokurora maldīgais spriedums, ka ir maza varbūtība, ka apsūdzētāis ir nevainīgs. Pieņemsim, ka E- pierādījumi, I- apsūdzētais ir nevainīgs, $P(E|I)$- varbūtība, ka tiks iegūti nozīmīgi pierādījumi, pat ja apsūdzētais ir nevainīgs, $P(I|E)$- varbutība, ka apsūdzētāis ir nevainīgs, neskatoties uz pierādījumiem.

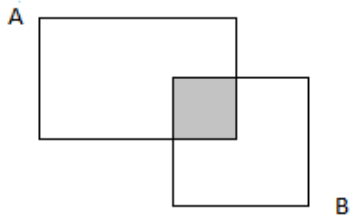
Uzskats, ka $P(E|I) = P(I|E)$ ir aplams, jo pēc Baiesa teorēmas

$$P(I|E) = P(E|I) \cdot \frac{P(I)}{P(E)}, \text{ kur} \tag{6.1}$$

$P(I)$- nevainīguma varbūtība, neņemot vērā pierādījumus E, $P(E)$- varbūtība, ka tiks iegūti pierādījumi. Vienādojums parāda, ka mazs $P(E|I)$ nenosaka viennozīmīgi mazu $P(I|E)$, ja ir liels $P(I)$ un mazs $P(E)$ jeb ja apsūdzētāis bez šī pierādījuma ar lielu iespēju būtu nevainīgs un vispārīgi nav varbūtiski iegūt pierādījumu E.

6.2 Matemātiskais apraksts

Viens no veidiem kā parādīt, ka nosacītā varbūtība nav abpusēja ir izmantojot Venna diagrammas. [20]



(a)

5. att. **Venna diagrammas attēls, kur redz abu figūru laukumus**

Pieņemsim, ka diagrammā 4.attēlā nosacītās varbūtības jēdziens ir prezentēts kā divu laukumu attiecība. Daļas skaitītājs ir laukums, kur divi taisnstūri pārklājas. Daļas saucējs ir

viena vai otra taisnstūra laukums. Ir liela nozīme tam, kura taisnstūra laukums ir saucējā. Varbūtība, ka esi figūrā A pie nosacījuma, ka tu jau atrodies figūrā B ir apmēram 0,25. Savukārt, varbūtība, ka atrodies figūrā B pie nosacījuma, ka jau esi A jeb P(B|A) ir tikai apmēram 0,14.

6.3 Piemērs

Šis maldīgais spriedums ir sastopams arī epidemoloģijā un sabiedrības veselības jomās. Piemēram, Dauna sindroma risks paaugstinās ar mātes vecuma palielināšanos[22], risks ir 17 reizes augstāks starp bērniem, kuru mātēm ir vairāk kā 40 gadi salīdzinot ar bērniem, kuri ir dzimuši mātēm zem 30 gadiem. Var kļūdīties, ja pieņem, ka lielākā daļa sieviešu, kurām piedzimst bērns ar Dauna sindromu ir vecākas, bet dati rāda, ka 51 procents no bērniem ar sindromu ir piedzimuši mātēm zem 30 gadiem, jo sievietēm zem 30 gadiem ievērojami biežāk dzimst bērni.

SECINĀJUMI

Izvirzītais uzdevums izpētīt un aprakstīt populārākos paradoksus un maldīgos spriedumus statistikā tika sasniegts. Tika izpētīti, aprakstīti un izskaidroti uz piemēriem trīs paradoksi- Simpsona, Lorda un Lindleja, kā arī trīs maldīgie spriedumi: maldīgais spriedums ekoloģisko datu interpretācijā, prokurora maldīgais spriedums un spēlētāja maldīgais spriedums.

Darba izstrādes gaitā autors sastapās ar to, ka vienīgie pieejamie avoti par šo tēmu ir raksti un publikācijas, kuros par paradoksiem ir izteikti konfliktējošo viedokļi, piemēram, dažu publikāciju autori uzskata, ka Lorda paradokss ir Simpsona paradoksa variācija.

Izpētes laikā tika secināts, ka paradoksi un maldīgie spriedumi statistikā ir aktuāla problēma un šo tēmu var turpināt pētīt, detalizētāk apskatot jau aprakstītos fenomenus, kā arī apskatot tos, kuri nav aprakstīti šajā darbā.

Izmantotā literatūra un avoti

[1] W. Robinson Ecological Correlations and the Behavior of Individuals. *American Sociological Review* 15(3) (1950): 351-357.

[2] E. Durkheim E *Suicide: A Study in Sociology.* London: Routledge, 1951.

[3] A. A. Schuessler Ecological inference. *Proceedings of the National Academy of Sciences* (1999) 96 (19) 10578-10581.

[4] G. King, O. Rosen, M. A. Tanner. *Ecological inference: new methodological strategies.* Campbridge University Press, 2004.

[5] D.A. Freedman *Ecological Inference and the Ecological Fallacy.* University of California, 1999.

[6] L. Goodman " Ecological Regression and the Behavious of Individuals." *American Sociological Review* (1953) 18: 665-666.

[7] Sprenger, Jan and Weinberger, Naftali,"Simpson's Paradox", *The Stanford Encyclopedia of Philosophy* (2021)

[8] E.H. Simpson "The Interpretation of Interaction in Contingency Tables",*Journal of the Royal Statistical Society*, (1951) 13(2): 238–241

[9] S. Kim "Explaining Lord's Paradox in Introductory Statistical Theory Courses",*International Journal of Statistics and Probability*,(2018) 7(4):1

[10] F.M. Lord "A paradox in the interpretation of group comparisons",*Psychological Bulletin*, (1967) 304-305

[11] J. Pearl Ļord's Paradox Revisited – (Oh Lord! Kumbaya!)",*Journal of Causal Inference*, (2016)

[12] J. Orloff, J. Bloom "Comparison of frequentist and Bayesian inference",*Class 20, Spring 2014*

[13] D. Lindley, A statistical paradox,*Biometrika* (1957), 44 187–192

[14] H. Jeffreys. *Theory of Probability. 1st ed* The Clarendon Press, Oxford (1939)

[15] Wang, Z., Rousseau, R, "COVID-19, the Yule-Simpson paradox and research evaluation" *Scientometrics* (2021) 3501–3511

[16] https://en.wikipedia.org/wiki/Lindley

[17] M.Kovica, S. Kristiansenb "The gambler's fallacy fallacy (fallacy)",*Journal of Risk Research* (2017) 22 2:1-12

[18] Tversky, Amos, and D. Kahneman. "Belief in the law of small numbers", *Psychological Bulletin* (1971) 76 (2): 105–110.

[19] B. Portnov, M. Barchana "On ecological fallacy, assessment errors stemming from misguided variable selection, and the effect of aggregation on the outcome of epidemiological study" *J Expo Sci Environ Epidemio* (2007)17, 106–121

[20] S.W. Huck, *Statistical misconceptions*, Taylor & Francis, 2008.

[21] Thompson, W.C.; Shumann, E.L. (1987). "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy". Law and Human Behavior.

[22] D.Westreich, N.Iliinsky, "Epidemiology Visualized: The Prosecutor's Fallacy". *American Journal of Epidemiology* (2014)1125–1127

A R programmu kodi

```
%SIMPSONA PARADOKSA PIEMĒRS
install.packages("palmerpenguins")
library(tidyverse)
library(palmerpenguins)
set.seed(9450)

penguin_df<-
  palmerpenguins::penguins %>%
  na.omit()



DT::datatable(head(penguin_df))

lin_reg <- lm(bill_depth_mm ~ bill_length_mm, data=penguin_df)




penguin_df %>%
  ggplot(aes(x=bill_length_mm, y=bill_depth_mm)) +
  geom_point() +
  geom_abline(slope = lin_reg$coefficients[[2]],
              intercept = lin_reg$coefficients[[1]],
              color="red") +
  labs(x="Garums", y="Platums",
       title="Regresija uz platumu kā funkciju no garuma") +
  theme_classic()

summary(lin_reg)




chin<-
  penguin_df %>%
  filter(species == "Chinstrap")
adelie<-
  penguin_df %>%
  filter(species == "Adelie")
```

```
gentoo<-
  penguin_df %>%
  filter(species == "Gentoo")

lm_chin<- lm(data=chin, bill_depth_mm ~ bill_length_mm)
lm_adelie<- lm(data=adelie, bill_depth_mm ~ bill_length_mm)
lm_gentoo<- lm(data=gentoo, bill_depth_mm ~ bill_length_mm)

penguin_df %>%
  ggplot(aes(x=bill_length_mm, y=bill_depth_mm,
             color=species)) +
  geom_point() +
  geom_abline(slope = lm_chin$coefficients[[2]],
              intercept = lm_chin$coefficients[[1]],
              color="black") +
  geom_abline(slope = lm_adelie$coefficients[[2]],
              intercept = lm_adelie$coefficients[[1]],
              color="black") +
  geom_abline(slope = lm_gentoo$coefficients[[2]],
              intercept = lm_gentoo$coefficients[[1]],
              color="black") +
  labs(x="Garums", y="Platums",color='Šķirne',
       title="Regresija uz platumu kā funkciju no garuma") +
  theme_classic()

%LORDA
library(ggplot2)
N <- 1000
b <- 10
l <- 50
u <- 70
Siev1 <- runif(N, l, u)
Vir1 <- Siev1 + b
beta1 <- 0.4
SievB0 <- (1 - beta1) * mean(Siev1)
VirB0 <- mean(Vir1) - beta1 * (mean(Siev1) + b)
sds <- 1
Siev2 <- SievB0 + beta1 * Siev1 + rnorm(N, sd=sds)
Vir2 <- VirB0 + beta1 * Vir1 + rnorm(N, sd=sds)
dati <- data.frame(sakuma_svars = c(Siev1, Vir1), gala_svars = c(Siev2, Vir2))
```

```
dati$dif <- dati$gala_svars - dati$sakuma_svars
dati$dzimums = c(rep(0, N), rep(1, N))
legend_title <- "Dzimums"
ggplot(data = dati, aes(sakuma_svars, gala_svars, color = factor(dzimums))) +
  geom_point() + stat_smooth(method = "lm")  +
  geom_abline(intercept = 0, slope = 1) +
  labs(x = "Sākuma svars")+
  labs(y = "Beigu svars")+
 # scale_fill_manual(legend_title,values=c("orange","red"))+
  labs(color = "Dzimums")+
 theme_bw()
summary(lm(dif~dzimums,data= dati)) #t-test
summary(lm(gala_svars ~ dzimums + sakuma_svars, dati)) #ancova
```

A Pielikumi

```
> summary(lm(dif~dzimums,data= dati)) #t-test

Call:
lm(formula = dif ~ dzimums, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5111 -2.9254  0.0298  2.9671  8.8013

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0109472  0.1127243  -0.097    0.923
dzimums     -0.0009778  0.1594162  -0.006    0.995

Residual standard error: 3.565 on 1998 degrees of freedom
Multiple R-squared:  1.883e-08, Adjusted R-squared:  -0.0005005
F-statistic: 3.762e-05 on 1 and 1998 DF,  p-value: 0.9951
```

(a)

6. att. **Lorda paradokss**

```
> summary(lm(gala_svars ~ dzimums + sakuma_svars, dati)) #ancova

Call:
lm(formula = gala_svars ~ dzimums + sakuma_svars, data = dati)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6839 -0.6716  0.0046  0.6793  3.9314

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.344615   0.240019  151.42   <2e-16 ***
dzimums       6.052531   0.059773  101.26   <2e-16 ***
sakuma_svars  0.394649   0.003962   99.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.001 on 1997 degrees of freedom
Multiple R-squared:  0.9677,    Adjusted R-squared:  0.9677
F-statistic: 2.992e+04 on 2 and 1997 DF,  p-value: < 2.2e-16
```

(a)

7. att. **Lorda paradokss**

Bakalaura darbs „Paradoksi un maldīgi spriedumi statistikā” izstrādāts LU Fizikas, matemātikas un optometrijas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā noslēguma darba elektroniskā versija parakstīta ar drošu elektronisko parakstu.

Autors:

_____ _____

(paraksts) (datums)

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītājs:

_____ _____

(paraksts) (datums)

Recenzents:

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā 2021.gada _____ maijā

Dekāna pilnvarotā persona: metodiķe Lāsma Štāle

Darbs aizstāvēts Valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____

Komisijas sekretāre: asociētā profesore Ingrīda Uļjane