

**АЛЬТЕРНАТИВНИЙ КОМПЛЕКС ДИСПЕРСІЙНОГО АНАЛІЗУ ЯК
МЕТОД ВИЗНАЧЕННЯ ПРИЧИНИ ВІДСУТНОСТІ ДОСТОВІРНОСТІ
ЗВ'ЯЗКУ МІЖ КАТЕГОРІАЛЬНИМИ ДАНИМИ**

СІТКАР А.Д.

асистент кафедри біохімії, фармакології та

фізичних методів лікування з курсом аналітичної медицини

Медичний факультет ДВНЗ "Ужгородський національний університет"

м. Ужгород, Україна

РОСТОКА Л.М.

кандидат медичних наук, доцент,

в.о. зав. кафедри біохімії, фармакології та

фізичних методів лікування з курсом аналітичної медицини

Медичний факультет ДВНЗ "Ужгородський національний університет"

м. Ужгород, Україна

Актуальність. Під час проведення статистичного аналізу науковець може виявити відсутність взаємозв'язку між досліджуваними явищами. Якщо апріорі потрібно вказати, що його немає, то проблем не виникає і, як наслідок, формулюються відповідні висновки. У випадку, коли гіпотеза дослідження вимагає виявити достовірний зв'язок, то за його відсутності, частіше, подальший аналіз даних припиняється, а статистично невірогідні результати просто не вказують в матеріалах роботи. Але слід зазначити, що необхідно не тільки оцінювати достовірність отриманих результатів, але й шукати причини їх виникнення. Це особливо важливо, коли на збір та реєстрацію даних витрачено багато часу, зусиль, а в деяких випадках і коштів, а очікуваних результатів роботи не отримано.

Під час перевірки гіпотези про наявність зв'язку між явищами загальноприйнятим є розподіл методів статистичного аналізу на наступні

групи[4]: 1) кореляційний аналіз – використовують, коли змінні представлені кількісними або якісними ординальними даними (даний метод використовують найбільш часто); 2) дисперсійний аналіз - використовують, коли залежна змінна (результативна ознака) представлена кількісними або якісними ординальними даними, а незалежна змінна (факторіальна ознака, предиктор) представлена якісними номінальними даними; 3) аналіз таблиць спряженості – використовують, коли предиктори представлені якісними номінальними даними, а залежна змінна відображає кількість спостережень, для яких ознака відсутня або присутня. Слід зазначити, що такий поділ є умовним, оскільки прямою функцією деякий із методів є не аналіз зв'язку, а перевірка гіпотези про достовірність різниці між групами (дисперсійний аналіз). Інші методи (критерій χ^2 -Пірсона) оцінюють не стільки взаємозв'язок, а скільки асоціацію тому, що їх використовують при роботі із якісними даними та частотами.

Дисперсійний аналіз дає змогу визначити достовірність різниці між вибірками, коли кількість груп більше двох. Також, за допомогою нього, можна визначити частку факторного впливу (%) предиктору на параметри розсіювання у досліджуваних групах [3]. Наприклад, вплив (у %) прийому певної дози Йоду на рівень тиреоїдних гормонів. Тобто дисперсійний аналіз може оцінити наявність зв'язку лише опосередковано. Аналіз таблиць спряженості, залежно від кількості градацій, дає змогу перевірити як гіпотезу про рівність пропорцій у групах (таблиці 2×2 , залежно від того чи вони взаємопов'язані, чи незалежні), так і достовірність взаємозв'язку між якісними даними (таблиці $m \times n$, де хоча б одна ознака має рівно або більше трьох градацій). Але для характеристики сили взаємозв'язку, у даному випадку, необхідно використовувати спеціальні методи (ф-коефіцієнт спряженості, коефіцієнт асоціації Юла, V-коефіцієнт Крамера, K-коефіцієнт Чупрова та ін.) [5]. Слід окремо відмітити критерій χ^2 -Пірсона: для впорядкованих градацій (для тренду). Його доцільно застосовувати у випадку, коли одна ознака дихотомічна, а інша має більше трьох градацій, які є впорядкованими [9]. Наприклад, взаємозв'язок між

наявністю кровотеч із варикозно-розширених вен стравоходу (є, немає) та ступенем фіброзу печінки (I, II, III, цироз).

Мета роботи. Вказати умови використання статистичних методів аналізу зв'язку між явищами та критерії вибору певного методу залежно від типу даних та їх розподілу. Визначити альтернативний комплекс дисперсійного аналізу як метод, який дозволяє оцінити причини отриманих результатів при аналізі зв'язку, перенести алгоритм його обчислення в комп'ютерну програму.

Матеріали та методи. Використано бібліосемантичний та контент-аналіз літератури, в якій представлено методики аналізу зв'язку між явищами. Методику використання альтернативного комплексу дисперсійного аналізу показано на прикладі статистичної обробки даних досліджень кафедри біохімії, фармакології та фізичних методів лікування медичного факультету ДВНЗ "УжНУ". Для створення алгоритму обчислення альтернативного комплексу дисперсійного аналізу використано комп'ютерну програму Microsoft Excel 2007.

Результати дослідження. Методом, який дозволяє сукупно оцінити різні параметри взаємозв'язку є регресійний аналіз, зокрема множинний [9]. При його обчисленні ми також визначаємо ступінь кореляції параметрів, який є необхідним для побудови регресійної моделі. Також ми отримуємо особливий показник, коефіцієнт детермінації, який визначає сумарну дію предикторів на дисперсію (%) залежної змінної. Його використовують для оцінки рівня апроксимації, необхідної для побудови регресійної моделі. Коефіцієнт детермінації також може опосередковано вказати на причину відсутності взаємозв'язку між явищами, визначаючи факторний вплив предикторів [6, 8]. Коефіцієнт детермінації можна використовувати, коли дані є кількісними, а взаємозв'язок між змінними майже лінійним. Також можна провести багатфакторний дисперсійний аналіз. Але вище вказані методи неможна застосувати щодо якісних, (категоріальних) даних (табл. 1), тим більше між ними не може бути лінійного зв'язку як такого, доречніше говорити про асоціацію ознак.

**Методи перевірки гіпотези про наявність зв'язку (асоціації) між
категоріальними (якісними) даними**

Два ряди змінних	Достовірність зв'язку	Кр χ^2 -Пірсона: таблиця 2x2 (з поправкою Йейтса); Точний кр Фішера (кр Фішера-Ірвіна); Кр χ^2 -Пірсона: великі таблиці спряження; Коеф кореляції $r(r)$ -Спірмена; Коеф кореляції τ -Кендалла; Коеф Фехнера; Бісерійний (точковий або ранговий) кр.
	Сила зв'язку (асоціації)	Коеф кореляції $r(r)$ -Спірмена; Коеф кореляції τ -Кендалла; Коеф Фехнера; Тетрахоричний показник зв'язку; Поліхоричний показник зв'язку; Q-коеф асоціації Юла; Со-коеф коваріації; ϕ -коеф спряженості, контингенції; V-коеф Крамера; Бісерійний (точковий або ранговий) кр; Кореляційне співвідношення Пірсона
Більше двох рядів змінних		Коеф конкордації Кендалла; Багатомірний непараметричний дисперсійний аналіз Фрідмана; Лог-лінійний аналіз таблиць спряження
Примітки: кр - критерій, коеф - коефіцієнт.		

Приклад проведення статистичного аналізу та частина матеріалів досліджень [1, 2, 10].

Мета роботи. Оцінити нутритивний статус та соматометричні дані досліджуваної вибірки людей у взаємозв'язку з глікемічним індексом (ГІ) та калорійністю вживаних ними продуктів. Провести аналіз взаємозв'язку раціону харчування та вегетативного тону. **Матеріали і методи дослідження.** Проведено опитування добровольців ($n=105$, середній вік - $27,7 \pm 2,6$, частка чоловіків – 32,4%, жінок – 67,6%) за допомогою спеціальних анкет. Похідними показниками даних анкети стали середній ГІ (сГІ) та середня калорійність (сК) вживаних продуктів у перерахунку на 100 г продукту, а також усереднені ГІ та калорійність за одним прийомом (табл. 2). Опитувані були розподілені на 2 групи: I – ті, що вживали переважно продукти з високим та середнім ГІ (60%); II – в раціоні яких переважали продукти з низьким і нульовим ГІ (40%). **Критерії оцінки показників:** Глікемічний індекс: <10 – нульовий, 10-40 – низький, 40-70 – середній, >70 - високий; ІМТ (індекс маси тіла): <16 – виражений дефіцит маси, 16-18,5 – недостатня маса, 18,5-24,99 – норма, 25-

30 – надмірна маса, 30-35 – ожиріння 1 ст., 35-40 – ожиріння 2 ст., >40 – ожиріння 3 ст.

Таблиця 2

Структура вибірки (%) за віком, ІМТ та ГІ вживаних продуктів

Показник		Загальні (n=105)	Чоловіки n=34 (32,4%)	Жінки n=71 (67,6%)
Вік	Юнацький	57	53	59
	Зрілий 1 період	16	21	14
	Зрілий 2 період	27	26	27
ІМТ	Виражений дефіцит	0,95	0	1
	Недостатня маса	11,43	0	17
	Норма	60,95	62	61
	Надмірна маса	15,25	26	10
	Ожиріння 1 ст.	9,52	9	10
	Ожиріння 2 ст.	0	0	0
	Ожиріння 3 ст.	1,9	3	1
ГІ середній	Низький	34	38	32
	Середній	65	59	68
	Високий	1	3	0
ГІ за прийом	Низький	2	3	1
	Середній	46	50	44
	Високий	52	47	55

Статистичний аналіз даних. Для аналізу взаємозв'язку між ІМТ (у вигляді порядкових даних) та переважним вживанням продуктів з певним ГІ складено таблиці спряженості в різних комбінаціях та використано наступні критерії (за критичний рівень вважали $\alpha=0,05$): 1) χ^2 -Пірсона: таблиці 2x2, відповідає тетрагоричному показнику зв'язку; 2) χ^2 -Пірсона: mxn (довільні таблиці, 2x3 та 3x3), відповідає полігоричному показнику зв'язку; 3) Точний критерій Фішера (двосторонній); 4) Показник відносного ризику; 5) Показник відношення шансів; 6) Q-коефіцієнт асоціації Юла; 7) ϕ -коефіцієнт спряженості (контингенції); 8) C-коефіцієнт спряженості Пірсона; 9) K-коефіцієнт Чупрова; 10) V-коефіцієнт Крамера. Перші три критерії характеризують наявність та достовірність зв'язку, 4 і 5 використовуються для когортних досліджень та досліджень "випадок-контроль" відповідно, 6-10 характеризують силу взаємозв'язку. Результати розрахунків показано в таблиці 3.

Розраховані критеріальні значення вищевказаних методів

Пункт	Критерій	Статистика критерію	
1	χ^2 -Пірсона: 2x2	0,327 (3,841)	$p > 0,05$
2	χ^2 -Пірсона: 2x2	0,135 (3,841)	$p > 0,05$
3	χ^2 -Пірсона: 2x3	1,76 (5,991)	$p > 0,05$
4	χ^2 -Пірсона: 3x3	3,763 (9,488)	$p > 0,05$
5	Точний критерій Фішера	0,68368	$p > 0,05$
6	Коефіцієнт ϕ , Чупрова, Крамера	0,056	Зв'язок несуттєвий
7	Коефіцієнт спряженості Пірсона	0,056	
8	Коефіцієнт асоціації Юла	0,117	
9	Показник відношення шансів	1,265 (0,565-2,833)	Недостовірний
10	Показник відносного ризику	1,153 (0,703-1,891)	Недостовірний

В дужках (пункт 1-4) вказані критичні (табличні) значення критерію при $p=0,05$, тобто, для достовірності зв'язку розраховане значення повинне бути більше за критичне. Для точного критерію Фішера (пункт 5) обчислюється безпосередньо значення P , тобто, для достовірності зв'язку, воно повинне бути $< 0,05$. Значення критеріїв оцінки сили взаємозв'язку (пункт 6-8) повинні бути не менше 0,1 (за шкалою Чеддока). Щодо показників відношення шансів та відносного ризику (пункт 9, 10), то для них в дужках вказаний 95% довірчий інтервал (якщо він включає 1, то показники не достовірні). Як видно з таблиці, наявність взаємозв'язку між досліджуваними явищами, характер їх сили, показник шансів та ризику не є достовірними ($p > 0,05$).

Для поглибленого статистичного аналізу та оцінки факторного впливу (вживання продуктів з певним глікемічним індексом, частота вживання фаст-фуду) на масу тіла досліджуваних осіб було використано дисперсійний аналіз (альтернативний комплекс). Аналіз проводили за допомогою програми Microsoft Excel 2007 із введенням відповідних алгоритмів розрахунку статистики критерію (рис. 1).

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
2	BA	N	+	+		KA	3			MA1	15	NA1	35	CA	0,0917749		NJA	2		
3	+	8	4	3		KB	2			MA2	20	NA2	55				NJB	1		
4	-	7	16	3		KA*KB	6			MA3	6	NA3	15				NJAB	2		
5	-	17	38	8																
6																				
7	M	41			CZ	24,473		N X	5	MIA	NIA	MIA2	MIA2/NIA#FIA=MIA/NIA				S2A	0,045367		
8	N	105			CX	0,5171		N Z	99	A1	15	35	225	6,4286	0,4285714		S2B	0,077778		
9	M2/N	16,01			NX2	0,0207		S2X	0,1034	A2	20	55	400	7,2727	0,3686364		S2AB	0,157332		
10	CY	24,99			NZ2	0,9793		S2Z	0,2472	A3	6	15	36	2,4	0,4					
11											41	105		16,101						
12			MI	NI	MI2	MI2/NI	FI=MI/NI			MB1	15	NB1	42	CB	0,0777778					
13	A1	B1	8	18	64	3,5556	0,4444			MB2	26	NB2	63							
14		B2	7	17	49	2,8824	0,4118													
15	A2	B1	4	17	16	0,9412	0,2353													
16		B2	16	38	256	6,7368	0,4211			MIB	NIB	MIB2	MIB2/NIE#IB=MIB/NIB							
17	A3	B1	3	7	9	1,2857	0,4286			B1	15	42	225	5,3571	0,3571429					
18		B2	3	8	9	1,125	0,375			B2	26	63	676	10,73	0,4126984					
19		41	105			15,527					41	105		16,087						
20																				
21			F0,05	F0,01	P					A1	B1	18	0,4444	0,4286	0,3571429	0,8349	0,0492	0,002421	0,043583	
22	FX	0,4184	2,21	3,02	0,8353					B2	17	0,4118	0,4286	0,4126984	0,8022	-0,039	0,001523	0,025895		
23	FA	0,1856	2,99	4,6	0,8308					A2	B1	17	0,2353	0,3636	0,3571429	0,6258	-0,095	0,009027	0,153454	
24	FB	0,3146	3,84	5,64	0,5758					B2	38	0,4211	0,3636	0,4126984	0,8115	0,0352	0,001239	0,047068		
25	FAB	0,6364	2,99	4,6	0,5308					A3	B1	7	0,4286	0,4	0,3571429	0,819	0,0619	0,003332	0,026825	
26										B2	8	0,375	0,4	0,4126984	0,7655	-0,0472	0,00223	0,01784		
27																				
28										Г	0,3905						CAB	0,314654		
29																				

Рис. 1. Приклад алгоритму обчислень альтернативного комплексу дисперсійного аналізу в Microsoft Excel 2007.

Було побудовано спеціальну таблицю (табл. 3): по вертикалі – частота переважного вживання продуктів з певним ГІ (фактор В), по горизонталі – частота вживання фаст-фуду (фактор А), у клітинках – частота випадків (у чисельнику – кількість осіб з неадекватною масою тіла, в знаменнику – загальна кількість випадків).

Таблиця 4

Макет таблиці для початкових обчислень альтернативного комплексу дисперсійного аналізу якісних (категоріальних) даних

Фактори В↓ і А→	Не вживають	Рідко	Часто
Низький ГІ	8/18	4/17	3/7
Високий ГІ	7/17	16/38	3/8

За результатами аналізу розраховане значення $F=0,418$ (F критичне = 2,21, $p=0,835$). Сумарний вплив факторів А і В становить 2%, вплив неврахованих випадкових факторів – 98%. Значення F у даному випадку характеризує узагальнену факторіальну девіацію. Хоча за результатами

обчислень вплив досліджуваних факторів є недостовірним, але це не означає відсутність їх впливу окремо, або в комбінації [7].

Розраховане значення F для фактору А становить 0,186 ($F_{кр}=2,99$, $p=0,831$), для фактору В - 0,315 ($F_{кр}=3,84$, $p=0,576$), для комбінації факторів А і В - 0,636 ($F_{кр}=2,99$, $p=0,531$). З цього випливає, що вплив фактору А, фактору В чи їх комбінації масу тіла досліджуваних осіб не є достовірним ($p>0,05$).

Статистичний висновок. Таким чином, взаємозв'язок між характером харчування (частотою вживання фаст-фуду та продуктів з певним глікемічним індексом) та ІМТ відсутній. Причиною таких результатів може бути значний вплив випадкових факторів (регулярні фізичні навантаження або навпаки гіподинамія, наявність захворювання, віковий фактор та ін.), значна варіабельність показників нутритивного статусу (ГІ та калорійність продуктів, кратність прийому їжі), різна їх частота в групах (між- та внутрігрупова різниця), значний (98%) сумарний вплив цих факторів. Проблеми з розміром вибірки немає як такої, оскільки за формулою Лера (швидка формула з 80% потужністю та 0,5 стандартизованою різницею, тобто найгірший випадок) кількість досліджуваних у кожній групі повинна становити 64 (в даному дослідженні мінімальна кількість становила приблизно 40 осіб при складанні таблиць спряженості 2x2, з врахуванням того факту, що розподіл ознак за ГІ був нормальним (за критерієм Колмогорова-Смірнова)). Незважаючи на загальновідомий факт, що вживання високоглікемічних продуктів (швидкі вуглеводи) та транс-жирів (фаст-фуд) здатні призводити до ожиріння, вплив цих факторів у даній вибірці є недостовірним.

Альтернативний комплекс дисперсійного аналізу на вищевказаному прикладі дав змогу визначити достовірність факторіального впливу предикторів та частку неврахованих факторів саме якісних (категоріальних) даних, що створює передумови для перегляду даних досліджень, доопрацювання та залучення нових даних, виявлення помилок.

Висновок. Отже, під час проведення статистичного аналізу необхідно не тільки оцінювати достовірність отриманих результатів, але й шукати причини

їх виникнення. Навіть при отриманні недостовірних результатів їх все рівно необхідно обґрунтовувати. Слід також пам'ятати, що статистична достовірність (значущість) та клінічна важливість результатів це зовсім різні поняття. Наприклад, якщо деякий лікарський препарат недостовірно проявляв побічну дію на печінку (а під час проведення статистичного аналізу ми самі задаємо критичний рівень значущості), то це не означає, що негативного впливу даного препарату немає зовсім, або він незначний. Напроти, в деяких випадках ця побічна дія має грати ключову роль в наданні рекомендацій щодо тактики лікування та застосування конкретних засобів, які ґрунтуються на принципах доказової медицини. Таким чином, врахування більшості факторів впливу на здоров'я за допомогою статистичних засобів дозволить прогнозувати перебіг захворювання в кожного окремого пацієнта, що і визначає персоналізований підхід до лікування та профілактики.

Використана література

1. Вегетативний статус людини в асоціації з раціоном харчування як фактора виникнення цукрового діабету II типу / [Кузьмак М.В., Мимренко А.А., Росток Л.М., Сіткар А.Д.]. // Сучасна наука: проблеми і перспективи (частина I): матеріали III-ї Міжнародної науково-практичної конференції м. Київ, 6-7 жовтня 2017 року. — Київ.: МЦНД, 2017. — С. 43-44.

2. Використання глікемічного індексу для оцінки стану харчування студентів у порівнянні із соматометричними показниками / [Сіткар А.Д., Бедевельський М.С., Примак С.Л., Лях О.І., Йонаш К.І.] // Матеріали XIII науково-практичної конференції з міжнародною участю студентів та молодих вчених «Науковий потенціал молоді – прогрес медицини майбутнього». 22–25 квітня 2015 р. / Укладач Смоланка В.В. — Ужгород: Видавництво УжНУ «Говерла», 2015. — С. 85-86.

3. Гланц С. Медико-биологическая статистика / С. Гланц. — М.: Практика, 1998. — 459 с.

4. Лапач С.Н. Статистика в науке и бизнесе / С.Н. Лапач, А.В. Чубенко, П.Н. Бабич. — К.: МОРИОН, 2002. — 640 с.
5. Медик В.А. Статистика в медицине и биологии / В. А. Медик, М.С. Токмачев, Б. Б. Фишман ; под. ред. Ю.М. Комарова : руководство: в 2 т. — М.: Медицина, 2000. — Т. 1. Теоретическая статистика. — 412 с.
6. Медична інформатика. Методичні вказівки до лабораторного практикума / [Шуста В.С., Кедюлич В.М., Куриця І.Ю., Сливка О.Г.]. — Ужгород: "Говерла". — 2011. — 41 с.
7. Мінцер О.П. Оброблення клінічних і експериментальних даних у медицині / О.П. Мінцер, Ю.В. Вороненко, В.В. Власов. — К.: Вища школа, 2003. — 350 с.
8. Наследов А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А.Д. Наследов. — СПб.: Питер, 2005 — 416 с.
9. Петри А. Наглядная медицинская статистика / А. Петри, К. Сэбин ; пер. В.П. Леонова. — М.: ГЭОТАР-Медиа, 2009. — 168 с.
10. Порівняльний аналіз глікемічного індексу та калорійності продуктів харчування для складання дієт / [Мимренко А.А., Кузьмак М.В., Росток Л.М., Сіткар А.Д.]. // Сучасна наука: проблеми і перспективи (частина I): матеріали III-ї Міжнародної науково-практичної конференції м. Київ, 6-7 жовтня 2017 року. — Київ.: МЦНД, 2017. — С. 45-46.