Національний університет біоресурсів і природокористування України
Кафедра економічної кібернетики

Методичні вказівки

для дисципліни «Математичні моделі та планування експерименту»

для аспірантів очної, заочної форми навчання за спеціальностями, освітні: неорганічні науки, професійна освіта та біологія, екологія, хімія, компютерні науки, прикладна механіка, ветеринарна медидицина, агрономія, харчові технології

Київ 2018

УДК 519.21.22.25:371.214.114

Висвітлено ряд основних теоретичних епитань курсу та наведено завдання для виконання самостійної роботи з предмета «Математичні моделі та планування експерименту». Рекомендовано для аспірантів неекономічних спеціальностей.

Рекомендовано методичною комісією факультетут ІТ НУБіП України. Протокол № 4 від 22 листопада 2018 року.

Укладачі: д.е.н., проф. А.В. Скрипник,

Рецензенти: к.п.н., доцент Т.Ю. Осіпова,
к.е.н., доцент Л.В. Галаєва

Навчальне видання
«МАТЕМАТИЧНІ МОДЕЛІ ТА ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ»

Методичні вказівки
для аспірантів неекономічних спеціальностей

Укладачі СКРИПНИК Андрій Васильович

Видання здійснено за авторським редагуванням.

ЗМІСТ

Вступ

Представлені методичні рекомендації розраховано на аспірантів різних спеціальностей, які використовують в власних дослідженнях статистичну інформацію. На мій погляд будь яке дисертаційне дослідження містить частку кількісної інформації. Наприклад поняття ймовірності широко використовується в юриспруденції (всі наведені докази мають деяку ступінь достовірності яка доводиться до журі присяжних), що стосується педагогіки то там перевірка інноваційних методів навчання здійснюється на статистичному порівнянні результатів базисної групи і групи що навчається по інноваційним технологіям. Тому навички обробки кількісної інформації потрібне при будь яких дослідженнях, якщо ви їх здійснюєте, або використовуєте результати інших дослідників.

Якщо розглядати статистику з професійної точки зору то вона складається з методології для збору, класифікації, інтеграції, аналізу і інтерпретації кількісної інформації. Слід відмітити, що інформація не завжди буває цифровою, наприклад, стандартна задача хто заробляє більше на фірмі чоловіки або жінки містить як данні о розмірі оплати так і інформацію відносно статі виробника. На цій час існує не менш 3 варіантів вирішення цій проблеми. Слід підкреслити одну надзвичайно важливу особливість статистика: будь яка цифрова інформація містить похибку, що пов’язано з багатьом обставин, що супроводжували збір інформації. Висновки, що робляться на підставі аналізу цій інформації можуть також бути помилковими, тому вводиться рівень значимості (ймовірність того що гіпотеза, яка обґрунтовується на підставі статистичних даних, не вірна). Завжди більш цікаві логічні побудові, які можливо перевіряти за допомогою статистики. Наприклад існує гіпотеза, що дружини зраджують своїм чоловікам не рідше, ніж чоловіки. Одним з можливість варіантів перевірки цієї гіпотези являється оцінка частки дітей які мають батьків, що не відповідають метриці. Якщо ця частка не буде суттєво відрізнятись від нуля, то наша гіпотеза не вірна.

Звичайно, що найбільший попит на статистичні дослідження виникає внаслідок діяльності бізнесу: наприклад, Ви запускаєте нову лінію з виробництва інноваційного м'ясного продукту, який ще не представлений на регіональному рівні. Для визначення ціни інноваційного продукту Вам потрібна різноманітна інформація: щодо розподілу доходів мешканців регіону, щодо рівня закупівельних цін на сировину, щодо рівня цін товарів замінників на регіональному ринку м'ясопродуктів. Однак не тільки бізнесу подрібни статистичні дослідження: значний попит на статистичні дослідження існує з боку політичний партій і політиків, але і будь яки рішення на державному рівні подрібни грунтоватись на статистичних дослідженнях. Звичайно статистики підверджує або відхіляє існуючі теоретичні положення економічної теорії, соціології. Наприклад в Україні існує точка зору, що більшість економічних законів тут не діє. Якщо розглянути наприклад закон існування ринкової рівноваги між попитом та пропозицією то його існування може бути підтверджено на прикладі пропозиції моркви наприклад у 2008 році, коли зменшення пропозиції (неврожай і скорочення посивів моркви) призвело до суттєвого зростання цін тобто морква коштувала як банани, які нам постачаються за тисячі кілометрів, то статистика підтвердила факт, що зменшення пропозиції призводить до зростання цін. Ми всі часто чуємо твердження, що багаті стають багатшими, а бідні біднішими. В цій постанові його не можна перевірити. Потрібно визначити кого вваважати богатимі, а кого біднимі. До речі, це далеко не риторичне питання. Так існує думка, що штаб Хіларі Клінтон перед президентськими виборами 2016 орієнтувався на виборця із середнім доходом, а штаб Трампа на медіанного за доходами виборця. З статистики відомий факт, що у розподілах з правою асиметрією до яких відносяться розподіли доходу завжди середньо значення більш медіанного, а медіанне більш модального. В США тривалий проміжок часу середні доходи росли, а медіанний зменшувалися. Результати останніх виборов в США нам широко відомі -победил Трамп який зробив ставку на медіанного вибірця. Тобто статистиці потрібні чіткі визначення і якісна

інформація. На жаль національна статистика працює надзвичайно неякісно. Наведемо тільки один приклад: відповідно даним Держкомстату рівень диференціації доходів населення у нас надзвичайно малий. Коефіцієнт Джині приблизно дорівнює 0,3 і це один з найкращих світових показників. Тобто наша статистика стверджує що різниця в доходах громадян незначна. Однак значний відсоток авто вищої цінової категорії, наявність двох революцій за останні 15 років, існуюча на цій час політична нестабільність свідчить о значної напруженості в суспільстві, що скоріші базується на суттєвої майнової та дохідної неоднорідності. Тому перевага віддається якісної статики, яку представляють міжнародні організації WB, IMF, FAOSTAT, EUROSTAT, Transparency Agency, яку будьмо використовувати не відмовляюсь, в деяких випадках, від даних Держкомстату.

Інформацію для досліджень можна також отримувати шляхом власних опитувань, проведення власних спостережень. Тепер перейдемо безпосередньо до описової та змістовної статистики. Застосування статистики включає два етапу: перший описання та представлення даних (графічне, табличне), другій використання даних для отримання висновків відносно особливостей середовища, де ці дані о механізмі (моделі) генерації даних, так наприклад, яким шляхом економіка створює розподіл доходів що описується данимі, що отримано в результаті досліджень, або який природний механізм забезпечує результати, що спостерігаються. Перший етап має назву описової статистики (descriptive statistics) другій змістовної статистики (inferential statistics). Описова статистика використовує кількісні та графічні методи для надходження шаблонних рішень поставленої задачі, інтегрування отриманих даних та її уявлення значущім чином. Змістовна статистика використовує дані для оцінок, рішень, передбачення або інших загальних уявлень.

Одне з найбільш розповсюджених застосувань статистичних методів відноситься до опитувань виборців. Воно робиться з метою оцінок шансів на перемогу той або іншої кандидатури на посаду. На цій час в розвинутих країнах опитування проводяться професійними агентствами у яких є

надзвичайно кваліфіковані працівники і відпрацьовані методики досліджень, однак невважаючи на це в останні роки статистичні дослідження, що прогнозували достатнє впевнену перемогу Хіларі Клінтон і відмову населення Великобританії від виходу із ЄС були спростовані реальністю. В обі двох випадках з незначною перевагою перемоклі аутсайдери опитувань: в президентських виборах 2016 року премог Дональд Трамп, а Великобританія підтримала вихід з ЄС (брексіт). Скоріше це свідчить: по перше що методологія опитувань повинна вдосконавлюватися; по друге - будь який прогноз містить статистичну похибку.

Застосування статистичних методів включає два етапу: перший пов'язаний з описом і поданням даних, другій з використанням даних для отримання висновок відносно особливостей середовища отримання інформації та головних механізмах діючих в середовище що створює цю інформацію. Перший етап має назву дискрипутивної статистики, а другій змістовної статистики. Дискрипутивна статистики використовує цифрові і графичні методи для узагальнення та аналізу інформації Змістовна статистика дає можливість зробити оцінки для прийняття рішень, прогнозу або висновок відносно середовища генерації даних. На першому етапі ми розглядаємо інструменти діскрипутивної статистики.

Тема 1. Вибірка та генеральна сукупність
1.1.Кількісні і якісни данні

Генеральна сукупність включає до себе всі об'єкти що потрібно дослідити. Наприклад в розвінутих країнах проводяться дослідження впливу рівня отриманої освіти на рівень доходів коли людина починає працювати. В цьому випадку генеральна сукупність це всі громадяне ,що отримали будь яку освіту. Кількісне уявлення освітного рівня може бути представлено роками навчання, а показник доходу дисконтним доходом протягом життя. Можлив і інший варіант, коли освітний процес представлено вартістю навчання.

Інший приклад, що стосується українських реалій, коли потрібно на законодавчому рівні виявити прожитковий мінімум У випадку України це три різних сукупності для яких прожитковий мінімум може суттєво відрізнятись: працюючи громадяни, пенсіонери та діти, що знаходяться на іждівенні батьків. Однак проводити дослідження всіх представників генеральної сукупності надзвичайно витратне (наприклад середньо статистична вартість одного опитування в Україні приблизно дорівнює 10 USD). Тому робиться вибірково дослідження, що базується на репрезентативної вибірці та містить головні риси генеральної сукупності. Що стосується мінімально потрібного обсягу вибірки то він визначається на підставі максимально припустимої похибки і цьому буде присвячено другій розділ.

Існують дані трех типів: просторові (cross –sectional), часові ряди і панельні данні, що поєлнюють перши дві категорії Класичний приклад просторової інформації це данні відносно продуктивності фермерських господарств (обробляємо площа, урожайність, рентабельність), однак поняття cross –sectional (поперечний переріз) ширше ніж просторовий до нього наприклад відносяться данні о рівні оплати праці працівників деякої фірми на якій то час, якщо це буде вже не визначений час а наприклад декілька років то таки данні вважаються панельними.

Крім того дані діляться на якісні та кількісні.

Табл.1.1 містить як якісни так і кількісни данні про рівень оплати праці 50 працівників (елементів). Кожний елемент містить дві компоненти: тижневу оплату праці і стать працівника. Зарплата і стать є характеристиками кожного елементу (працівника), які змінюються від елементу к елементу. Зарплата – кількісна змінна, а стать якісна. Однак слід підкреслити що в багатьох приложеннях статистики існують методи переходу від якісних до кількісних зміних. Так в економетриці для цифровізації стати використовується бінарна зміна.

Табл.1.1. Данні відносно тижневої оплати праці (USD) та статі працівників

№	Зарп.	Стать	№	Зарп.	Стать	№	Зарп.	Стать
1	236	ж	18	490	м	35	337	ж
2	573	м	19	745	ж	36	1406	м
3	660	ж	20	2033	м	37	530	м
4	1005	м	21	391	ж	38	644	м
5	513	м	22	179	ж	39	776	ж
6	188	ж	23	1629	м	40	440	ж
7	252	ж	24	552	ж	41	548	ж
8	200	ж	25	144	ж	42	751	ж
9	469	ж	26	334	ж	43	618	ж
10	191	ж	27	600	ж	44	822	м
11	675	м	28	592	м	45	437	ж
12	392	ж	29	728	м	46	293	ж
13	346	ж	30	125	ж	47	995	м
14	264	ж	31	401	ж	48	446	ж
15	363	ж	32	759	ж	49	1432	м
16	344	ж	33	1342	м	50	901	ж
17	949	м	34	324	ж			

На підставі отриманих даних побудуємо варіаційний ряд (у порядку зростання) зарплат працівників з у казанням статі. Найменьша зарплата 125 $

стає першою, найбільша 2033 останьою (50). Різниця між найбільшою та найменшою складає діапазон зарплат $2033-$125=$1908.

1.2. Медіана, мода, гістограма

Кількісне значення серединного елементу називається медіаною, якщо кількість непарна то медіаною буде кількісне значення серединного елементу, якщо воно парне, як в цьому випадку, то медіаною буде середньо значення 25 та 26 елементу варіаційного ряду, яке дорівнює $521,5 (медіану також називають 50% процентілем). Крім того в статистиці використаються поняття кванті лей (квантіль містить ¼ елементів вибірці): 25%-перший квантіль (нижній), 50%- другій квантіль (медіана),75%- третій квантіль(верхний). Квантилі щомісячних зарплат подано у табл.1.2.

Табл.1.2. Варіаційний ряд щотижневих зарплат (USD)

№	Зарп.	Стать	Квантілі
1	125	ж	
2	144	ж	
3	179	ж	
…	…	…	
12	334	ж	335,5 -1 квантіль
13	337	ж	25% прцентіль
….			
25	513	м	521,5-медіана
26	530	м	2 квантіль, 50% процентіль
…			
37	745	ж	748-3квантіль,
38	751	ж	75% процентіль
…			
49	1629	м	

50	2033	м	

Відстань між кінцем 1 і початком 3 квантилю називається міжквантильним діапазоном, у нашому випадку він дорівнює: $748-$335,5=$412,5. Наявна у табл.1.2 інформація дозволяє побудувати спрощено графічне уявлення розподілу випадкової змінної що досліджується (box plot) в якому представлені кванти лі розподілу та його найменше та найбільше значення.

Найбільш цікаве питання яке можна вирішити на підставі цих даних отримють ли чоловіки заробітну плату більшу ніж жінки. Для цього потрібно відсортувати окремо масиви чоловіків та жінок.

Потім ми знайдемо діапазони, медіани, кванті лі і межквантільни діапазони для ободвох масівів.

Побудуємо окремі гістограми для чоловіків та жінок (табл.3). Звичайно стандартний метод порівняння здійснюється через оцінку середньох значень, та оцінку загальної дисперсії (t критерій). В подальшому ми розглянемо більш детально порівняння вибірок через порівняння середньох значень і обгунтуємо недоліки і переваги цього методу. Продовжимо порівняння рівней оплати праці чоловіків і жінок на підставі поширених статистичних характеристик. Вихідна Інформація для побудові гістограм для чоловіків, жінок і разом представлено в табл.1.1.

Табл.1.3. Розподіл частот за даними табл.1.1.

Діапазон		частоти		відносні	частоти	
(тис. USD)	Чол.	Жінки	Загалом	%(Чол.)	%(Жін.)	%(Заг.)
0,0-0,5	1	23	24	0,06	0,7	0,48
0,5-1,0	10	10	20	0,58	0,30	0,40
1,0-1,5	4	0	4	0,24	0,00	0,08
1,5-2,0	1	0	1	0,06	0,00	0,02
2,0-2,5	1	0	1	0,06	0,00	0,02

∑	17	33	50	1,00	1,00	1,00

Гістограми розподілу тижневої оплати праці представлено на рис.1.1. Верхня гістограма відповідає розподілу оплати праці чоловіків. Середня жінок і нижня оплаті праці всіх працівників. Що стосується модальних значень то за них приймається середина інтервалу з максимальною кількістю спостережень: для чоловіків це 0,75 тис. USD, для жінок 0,35 тис. USD, для загальної виборці 0,25 тис. USD.

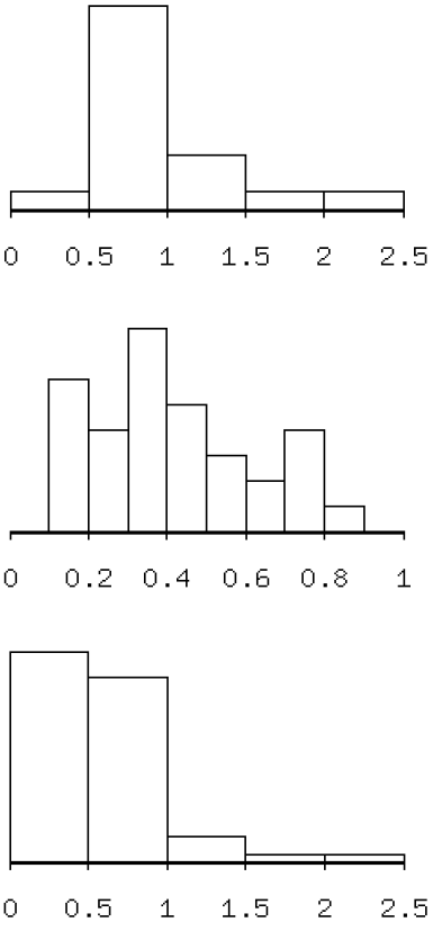


Рис.1.1. Гістограми розподілу тижневої оплати праці (тис. USD) для чоловіків (верхня), жінок (середня),всіх працівників (нижня)

Звичайно гістограма не дозволяє зробити однозначний висновок відносно співвідношення рівня оплати праці чоловіків та жінок, однак слід підкреслити що якщо модальне значення для чоловіків знаходиться в діапазоні 0,5-1,0 тис. USD то для жінок в інтервалі 0,3-0,4 тис.USD. Однак більш

ретельний метод порівняння можна здійснити за допомогою побудови box plot для даних по чоловікам та жінкам (рис.1.2).

1.3. Бокс плот, дісперсія, асіметрія

З представлених box plot слідує, що медіана для чоловіків, а також ніжній і верхній кванті лі для чоловіків суттєво перевищують аналогічні показники для жінок. Крім того максимальне і мінімальне значення для чоловіком суттєво перевищує подібне показники для жінок. Наведени дани відносно рівня оплати праці чоловіков і жінок дозволяють зробити висновок, що базуючись на характеристиках розподілів, чоловіки отримцють більшу щотижневу оплату праці. Однак це не означає, що будь який чоловик отримує щотижневу оплату праці більшу ніж будь-яка жінка.

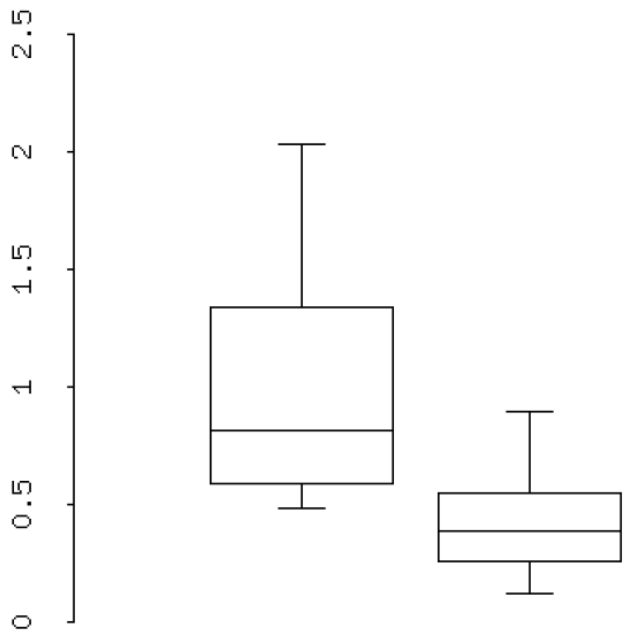


Рис.1.2. Box plot тижневої оплати праці чоловіків і жінок (тис. USD)

Оцінки мінливості

Тощо міра відхилення процесу від очікуваного значення надзвичайно важлива характеристика можна пояснити на наступному прикладі. При вступі на роботу керівник фірми гарантував вам заробітну плату не менш ніж 300 USD. І дійсно по результатам року серене-місячна оплата праці виявилась 325

USD, причому 500 USD було виплачено в кінце року, а за деяки місяци оплата не перевищувала 100 USD. В наслідок того що ви розраховували на стабільний доход в 300 USD вам прийшлось запозичувати кошти під деякий відсоток і ви понесли зайві моральні та матеріальні витрати. Така ситуація склалась тому що при вступі не було оговорено інший важливий показник оплати праці – її стабільність. Найбільш распосюдженою мірою стабільності або мінливості є дисперсія. Дисперсія для будь якої виборці X обсягом N визначається:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}; \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad (1.1)$$

N – 1 використовується в знаменнику при розрахунку дисперсії вибірки, оскільки дисперсія являє собою середнє суми квадратів незалежних відхилень від середнього значення вибірки і тільки N – 1 всіх N відхилень від середнього вибірки може бути незалежно відібрано, останнє відхилення може бути розраховане на базі уже відомих.

Кожна вибірка з буде мати різне середнє, в залежності від тих елементів, які в ній спостерігаються. При цьому, середнє всієї генеральної сукупності являється фіксованим числом, яке не змінюється від вибірки до вибірки. Таким чином відхилення спостережень від генеральної сукупності є незалежними один від одного. В цьому випадку у знаменикі для оцінки дисперсії замінюється N-1 на N.

Якщо вибірка представлено розподілом частот - f_j кількість спостережень на j інтервалі гістограми, де k –кількість інтервалів $\sum_{j=1}^{k} f_j = N$, $\bar{x}_j$ – середньо значення інтервалу то дисперсія визначається:

$$s^2 = \frac{\sum_{j=1}^{k} f_j(\bar{x}_j - \bar{x})^2}{N-1}; \bar{x} = \frac{\sum_{j=1}^{k} f_j \bar{x}_j}{N-1} \qquad (1.2)$$

Дисперсія генеральної сукупності позначається, як σ2 .Для для кінцевої вибірки вона розраховується за формулой (1.2) після заміни N − 1 в знаменнику на N.

В щотижневих даних про заробітку плату, які наведені вище, дисперсія для чоловіків становить 161 894, для жінок 42898, а для всієї вибірки 204792. При цьому слід звернути увагу на те, що одиницею виміру дисперсії в цьому випадку виступають долари США в квадраті – ми беремо суму квадратів доларової різниці в заробітній платі кожного працівника від середнього.

Для отримання оцінку мінливості, яка вимірюється в доларах, а не доларах у квадраті – можна взяти корінь квадратний з дисперсії у рівнянні (1.1;1.2), в результаті чого отримаємо середньо квадратичне відхилення -s. В розглянутій вище вибірці середньо квадратичне відхилення для чоловіків становить 402,3 USD, для жінок 207,1 USD і 452,5 USD для всієї вибірки.

Іншою часто застосованою мірою мінливості виступає коефіцієнт варіації, що визначається як частка середньо квадратичного відхилення від середнього:

$$V = \frac{s}{x} = 100\% \qquad (1.3)$$

Перевагою коефіцієнту варіації є то ще він є без вимірний що дозволяє порівняти показники варіативності різних за масштабом процесів. Наприклад більшими за розмірами компаніям відповідають як більші значення середніх доходів так і прибутків, звичайно, що і показник мінливості дисперсія також залежить від доходів. Використання коефіцієнту варіації дозволяє піти від масштабу процесу і порівняти без вимірні показники варіативності.

Наведемо наступний приклад: середньо добовий доход кафе складає 50 тис. грн., а його дисперсія -400; доход павільйону Кулінічи -20 тис. грн., а дисперсія 25 тис. грн. Порівняти відносні показники варіативності?

Середньо квадратичне відхилення доходу кафе 20 тис. грн., павільйону Кулінічи-5 тис.грн. Кофіцієнт варіації доходу кафе 20/50*100%=40%, для павільйону Кулінічи-5/20*100%=25%/

Звідси можно зробити висновок що Кулінічи більш стабільний бізнес.

Оцінки асиметрії

До асиметричних кількісних даних відносяться дані, для яких розподіл частот, що базується на основі однакових класів не є симетричним, тобто переважають значни відхилення або в бік великих значень або в бік малих. Як правило при дослідженні фінансових показників переважать відхилення в бік великиз значень. Наприклад, дані про заробітну плату представлені на Рис 1.1 не є симетричними ¬– правий хвіст довший ніж лівий. Це має логічне пояснення – тижнева заробітна плата не може бути від'ємною величиною (тобто зліва вона обмежено нулем), чого не можна сказати відносно правого хвосту. Ці дані можуть бути описані через праву асиметрію – асиметрія в сторону довшого хвоста. Така асиметрія спостерігається в бокс плоті на рис 1.2, оскільки верхній вусик довший за нижній. При цьому потрібно звернути увагу на те, що в даних про доходи середнє завжди більше за медіану, а медіана завжди більша за моду. Середні, медіани та моди становлять відповідно 962; 822; та 750 USD для чоловіків, 424; 391 та 250 USD для жінок, 607; 521 та 250 USD для всіх працівників (модальне значення обирається як середнє значення інтервалу з найбільший кількістю спостережень). Середнє завжди буде більше за медіану, а медіана завжди буде більша за моду, коли в даних спостерігається права асиметрія. В тому випадку, коли наявна ліва асиметрія середнє буде менше медіани, а медіана буде менша моди.

Асиметрія уявляє собою усереднений куб відхилення від середнього значення. Асиметрія може дорівнювати нулю, і це означає що розподіл що досліджується є симетричним. Наприклад симетричним є розподіл Гауса та рівномірний розподіл. Асиметрія може розраховуватись, як по вихідному числовому ряду так і по гістограмі:

$$A = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N-1} \qquad (1.4)$$

за вихідним рядом

$$A = \frac{\sum_{j=1}^{k} f_j (\bar{x}_j - \bar{x})^3}{N-1} = \sum_{j=1}^{k} \varphi_j (\bar{x}_j - \bar{x})^3$$, де $\varphi_j = \frac{f_j}{N-1}$

за згрупованими данимі (гістограма)

Позначення в (1.4) аналогічні позначенням у (1.2). По вимірності асиметрія уявляє вихідну одиницю у тре тему ступені. Для прикладу з тижневими зарплатами асиметрія вимірюється в кубічних USD, а оскільки в розподілах як чоловіків так і жінок переважають праві хвости розподілів, асиметрії є додатною. Існує і без вимірний коефіцієнт асиметрії, що дозволяє порівняти ступінь асиметрії процесів різних масштабів. Він розраховується як відношення оцінки асиметрії до кубу оцінки середньо квадратичного відхилення:

$$\gamma = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{s^3 (N-1)} \qquad (1.5)$$

за вихідним рядом

$$\gamma = \frac{\sum_{j=1}^{k} f_j (\bar{x}_j - \bar{x})^3}{s^3 (N-1)} = \frac{\sum_{j=1} \varphi_j (\bar{x}_j - \bar{x})^3}{s^3} = \frac{A}{s^3} \qquad (1.6)$$

за згрупованими данимі (гістограма).

За згрупованими даними спостережень за щомісячнимі зарплатами чоловіків та жинок зробимо оцінку їх середних значень, дисперсій, коефіцієнтів варіації та асиметрії.

$$\bar{x}_ч = 0{,}25 \cdot 0{,}06 + 0{,}75 \cdot 0{,}58 + 1{,}25 \cdot 0{,}24 + 1{,}75 \cdot 0{,}06 + 1{,}25 \cdot 0{,}06 = 0{,}93$$

$$s^2_{ч,94} = 0{,}25^2 \cdot 0{,}06 + 0{,}75^2 \cdot 0{,}58 + 1{,}25^2 \cdot 0{,}24 + 1{,}75^2 \cdot 0{,}06 + 2{,}25^2 \cdot 0{,}06 - 0{,}93^2 = 0{,}14 \Rightarrow s_ч = 0{,}38$$

$$A_ч = 0{,}06(0{,}25 - 0{,}93)^3 + 0{,}58(0{,}75 - 0{,}93)^3 + 0{,}24(1{,}25 - 0{,}93)^3 + 0{,}06(1{,}75 - 0{,}94)^3$$

$$+ 0{,}06(2{,}25 - 0{,}94)^3 = 0{,}02$$

$$V = \frac{0{,}38}{0{,}93}100\% = 40{,}9\%; \gamma_ч = \frac{0{,}02}{0{,}38^3} = 0{,}39$$

Аналогічні розрахунки проведено для жінок:

$$\bar{x}_{ж} = 0{,}4; s_{ж} = 0{,}23; V = 57{,}5\%; A_{ж} = 0{,}01; \gamma_{ж} = 0{,}89$$

Відповідно відносним показникам показники варіативності і асиметрії для жинок перевищують аналогічні показники для чоловіків. Середний рівень оплати праці чоловіків більш ніж вдвічі перевищує аналогічний показник для жінок, тобто попередні висновки, що зроблено на підставі побудови boxplot піддержуються. Однак порівняння середньох буде подальше проведено більш коректно.

Завдання до теми 1.

По даним держкомстату побудовати гистограму доході громадян за 200М рік, де М кількість літерів у вашому імені. Знайти середній дохід, моду медіану і асіметрію доходів.

Тема 2. Квантілі, проценті лі, двохвимірни зміни
2.1. Безвимірні зміни

Відносна позиція деякого спостереження у виборці оцінюється за допомогою апарата що використовувался для побудові boxplota. Наприклад кожному буде цікаво в якому кванті лі за рівнем доходів він знаходиться, на жаль офіційні статистичні спостереження за рівнем доходів скоріші долеки від фактичного розподілу. Найчастиші використовуються поняття квантілей, тобто вісь розподіл ділется на чотирі частині. Надзвтчайно популярно в статистиці порівння першого і останього за рівнем доходів 10% процентілей. Наприклад розглядається споживання мясопрдуктів в цих процентілях і важливою характеристикою вважається відношення споживання представника верхнего процентилю до нижнього.

Однак відносне становіще може вимірюватись за допомогою приведення до безвимірного вигляду параметру, що досліджується Нехай існує вибірка $x_1; x_2; ... x_N$ для якої відоми математичне очікування - μ та середне-квадратичне відхилення - σ. Тоді безвимірна змінна z_i розраховується наступним шляхом:

$$z_i = \frac{x_i - \mu}{\sigma}; i = 1,2,..., N \qquad (2.1)$$

Зміна z_i має нульове математичне очікування та одиничну дисперсію.

Якщо математичне очікування та дисперсія невідомі то воні замінюються на оціки (1.1) що зроблено по вибірце.

$$z_i = \frac{x_i - \bar{x}}{s}; i = 1,2,..., N \qquad (2.2)$$

Безвимірне уявлення використовується при проведенні кластерного аналізу, за допомогою якого будуються кластери або група обєктів що мають близьки характеристики. Наприклад в аграрному секторі на слуху кластер великих великих ферм, яки обробляють значні площи, мають велики прибуткі, значну частку продукції спрямовують на експорт та мають

висококваліфіковани трудові ресурси. Всього нами було перераховано 4 показнику, що мають різні вимірності: га, гривні, відсотки, гривні (вважаємо що рівень кваліфікації визначається рівнем оплати праці). Після здійснення діскрипутивної статистики та приведення всіх даних окремих ферм до без вимірного вигляду між точками у чотиріхвимрному евклідовому просторі, що відповідають окремим фермам, можна розрахувати відстань та сформувати кластери тобто групу ферм що мають близькі значення обраних показників.

Що стосується местоположення окремих об'єктів в виборці, то для розподілу, який наближується до нормального близько 68% спостережень потрапляє в межі одного середньо квадратичного відхилення від середнього, близько 95% значень знаходяться в межах двох середньо квадратичних відхилень від середнього і близько 99,7% спостережень перебувають в межах трьох середньо квадратичних відхилень від середнього. Таким чином, якщо Ви отримуєте оцінку в 52% від максимального тоді як середнє для класу становить 40% з середньо квадратичним відхиленням в 10% і розподіл оцінок являється не біноміальним, то скоріш за все Ви знаходитесь в верхніх 16% класу. Цей розрахунок показує, що 68% класу матиме оцінку в межах одного середньо квадратичного відхилення від 40, тобто між 30 і 50, а 32% матиме оцінку, яка знаходиться за межами цього діапазону. Щоб попасти в 16% кращих вам потрібно зробити припущення про симетричність розподілу. Описані вище відсотки майже повністю зберігаються для нормального розподілу і тільки дещо для колоколообразного розподілу, який не задовольняє критерію нормальності. Також, ці відсотки не зберігаються для бімодального розподілу. Виявляється, що існує правило, розроблене Чебишевим, яка називається нерівністю Чебишева і говорить про те, що ймовірність виходу за межі к средньох квадратичних відхилень від середнього значення не перевищує(1/k)2 для будь-якого розподілу. Таким чином, в межах двох середньо квадратичних відхилень від середнього значення знаходиться найменше ніж 75% розподілу, а за ці межи потрапляє не більш ніж 25% розподілу.

2.2. Двохвимірні дані: Коваріація та кореляція

Набір даних, що містить тільки одну змінну, яка уявляє інтерес для дослідника, як у випадку з даними про заробітну плату називається одновимірним набором даних. Набори даних, які містять дві змінні, такі як заробітна плата та стать працівника називаються двохвимірними. Індекс споживчих цін та рівень інфляції, представлений в таблицях 1.4; 1.5 являються багатомірними, при цьому кожний набор даних містить чотири змінні – індекс споживчих цін або рівень інфляції для чотирьох країн.

У випадку двомірних або багатомірних наборів даних ми часто зацікавлені в тому, чи є елементи, що мають більші значення одної чи декількох змінних також більші значення для інших змінних. Наприклад, ми можемо бути зацікавлені в тому, чи люди з більшою кількістю років навчання мають вищий дохід, тобто дійсно лі освіта позитивно впливає на рівень життя. Так, ми можемо отримати для всіх домогосподарств Канади дві кількісні змінні, їх дохід (в доларах США) та кількість років освіти для кожного представника домогосподарства. Нехай Xi буде значення середньо річного доходу і-го домогосподарства та Yi кількість років навчання представника і-го домогосподарства.

Тепер розглянемо випадкову вибірку з N домогосподарств з парою спостережень (Xi, Yi) для i = 1, 2, 3, . . . , N .

Розрахуємо для кожної з визначених зміних середне значення та дисперсію.

Зверніть увагу на те, що вибірка складається з пар спостережень тобто кожне домогосподарство характеризується кількістю років навчання і доходом. Ми зацікавлени в питанні впливаю тривалісті років навчання на рівень доходів і який характер впливу прямий або обернений.

Оскільки вважається, що освіта являється формою інвестицій, які приносять свої доходи у формі більшого рівня оплати праці протягом всього життя, можна очікувати, наприклад, що дохід домогосподарства буде більший чим більше буде кількість років навчання.

Тобто, можна очікувати, що більші значення X будуть йти в парі з більшими значеннями Y, коли Xi високий, тоді асоційований з ним Yi також повинен бути більший, і навпаки.

Іншим прикладом являється індекс споживчих цін та рівень інфляції для пари країн. Нам може стати цікаво дізнатися, чи високі ціни та висока інфляція в США пов'язана з високим рівнем цін та інфляції в Канаді. В табл.2.1,2.2 представлено інтегральний індекс продовольчих цін і щорічний індекс інфляції для Канади, США, Великобританії, Японії за 1975-1996 роки. Індекс річної інфляції може бути представлено через інтегральний індекс споживчих цін наступним шляхом:

$$\pi = \frac{100(P_t - P_{t-1})}{P_{t-1}} \qquad (2.3)$$

де $P_t -$ індекс споживчих цін на час t, $\pi -$ індекс інфляції.

Табл.2.1. Інтегральний індекс споживчих цін для Канади, США, Великобританії, Японії за 1975-1996 роки (100%-1980 рік)

	Canada	U.S.	U.K.	Japan
1975	65.8	65.3	51.1	72.5
1976	70.7	69.0	59.6	79.4
1977	76.3	73.5	69.0	85.9
1978	83.1	79.1	74.7	89.4
1979	90.8	88.1	84.8	92.8
1980	100.0	100.0	100.0	100.0
1981	112.4	110.3	111.9	104.9
1982	124.6	117.1	121.5	107.8
1983	131.8	120.9	127.1	109.8
1984	137.6	126.0	133.4	112.3
1985	143.0	130.5	141.5	114.6
1986	149.0	133.0	146.3	115.3
1987	155.5	137.9	152.4	115.4
1988	161.8	143.5	159.9	116.2

1989	169.8	150.4	172.4	118.9
1990	177.9	158.5	188.7	122.5
1991	187.9	165.2	199.7	126.5
1992	190.7	170.2	207.2	128.7
1993	194.2	175.3	210.4	130.3
1994	194.6	179.9	215.7	131.2
1995	198.8	184.9	223.0	131.1
1996	201.9	190.3	228.4	131.3

Source: International Monetary Fund, International Financial Statistics.

Табл.2.2. Річной індекс інфляції для Канади, США, Великобританії, Японії за 1975-1996 роки

	Canada	U.S.	U.K.	Japan
1975	10.9	9.1	24.1	11.8
1976	7.5	5.7	16.6	9.4
1977	8.0	6.5	15.9	8.2
1978	8.9	7.6	8.2	4.1
1979	9.2	11.3	13.5	3.8
1980	10.2	13.6	17.9	7.8
1981	12.4	10.3	11.9	4.9
1982	10.8	6.2	8.6	2.7
1983	5.8	3.2	4.6	1.9
1984	4.3	4.3	5.0	2.2
1985	3.9	3.6	6.1	2.0
1986	4.2	1.9	3.4	0.6
1987	4.4	3.6	4.2	0.1
1988	4.0	4.1	4.9	0.7
1989	5.0	4.2	7.8	2.3
1990	4.8	5.4	9.5	3.1
1991	5.6	4.2	5.8	3.3

1992	1.5	3.0	3.7	1.7
1993	1.8	3.0	1.6	1.3
1994	0.2	2.6	2.4	0.7
1995	2.2	2.8	3.4	-0.1
1996	1.6	2.9	2.4	0.1

Source: International Monetary Fund, International Financial Statistics.

Одним з шляхів дізнатися це – буде побудова діаграм розсіяння з канадським індексом споживчих цін та рівнем інфляції на горизонтальних осях та індексом споживчих цін США та рівнем інфляції на відповідних вертикальних осях. Відповідно діаграму розсіяння індексу споживчих цін можна побачити на рис.2.1 та рівень інфляції на рис.2.2.

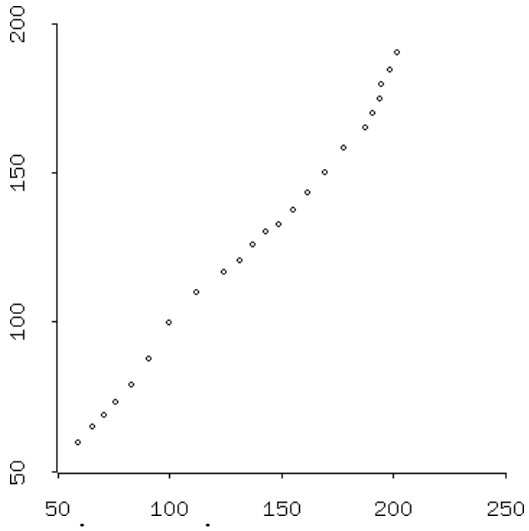


Рис.2.1. Діаграма розсіяння інтегральних канадського індексів споживчих цін – канадського горизонтальна ось та США вертикальна ось

З побудованих діаграм можна побачити, що і рівень споживчих цін і рівень інфляції в двох країнах мають позитивний зв’язок. Такий висновок робиться на підставі розрахунку коваріації між двома зміни мі:

$$s_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1} \qquad (2.4)$$

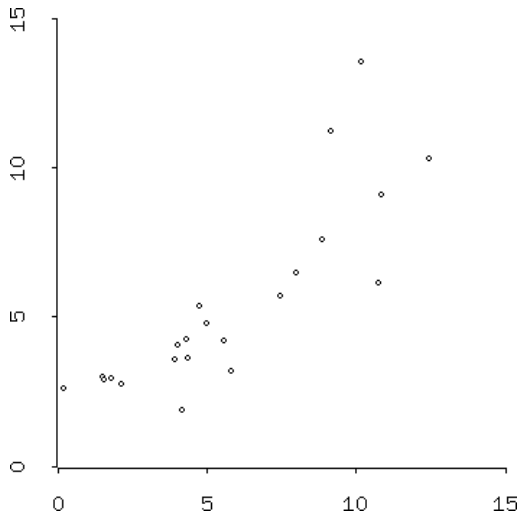


Рис.2.2. Діаграма розсіяння щорічного канадського індексу інфляції - горизонтальна ось та індексу інфляції США-вертикална ось

Знаменик у останьому виразу змінюється на N у випадку повної сукупності обєктів. Слід підкреслити, що коваріація є вимірна величина, так в прикладі впливу рівня освти на рівень доходів вона вимірюється в доларах* роки.

Для будь-якої вибірки пари змінних X та Y , sx,y має єдине числове значення, яке може бути позитивним, негативним або дорівнювати нулю. Позитивне значення являється індикатором того, що наявні значення для X та Y позитивно пов’язані, тобто їх підйоми та падіння відбуваються одночасно. Негативне значення коваріації означає що зростанню однієї змінної відповідає зменшення іншої. З наведених на рис.1.5,1.6 графіків слідує що між інфляційними процесами в Канаді та США спостерегался прямий взаємозв’язок тому оцінка коваріації буде позитивною.

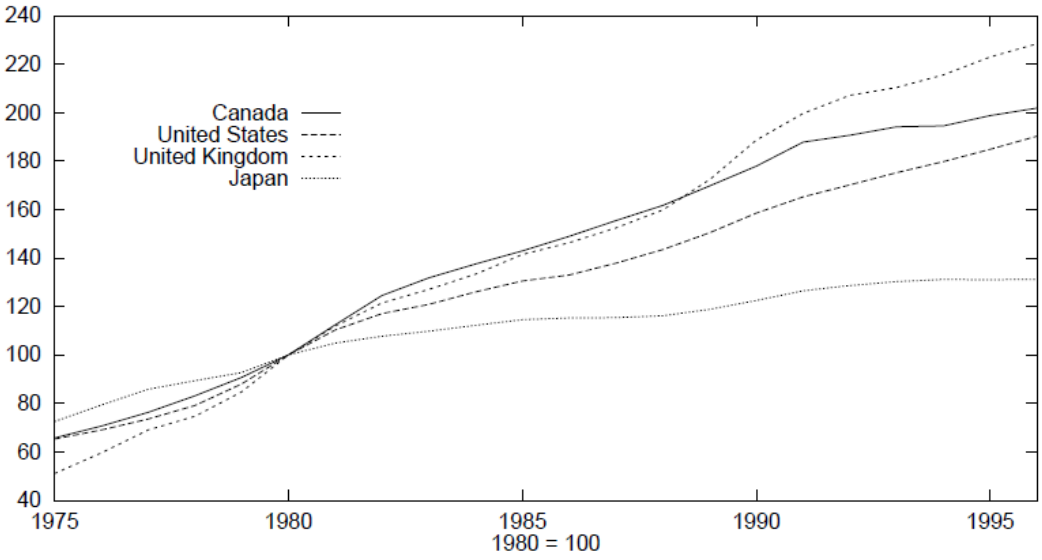


Рис.2.3. Динамика інтегральних індексів споживчіх цін декількох розвинутих країн

Крім того цікавим є то ще в 70-80 роки минулого століття в розвинутих країнах спостерігались інтенсівни інфляційні процеси варіативність яких суттєво зменшилась к концу 90 років минулого століття. На цій час річна інфляція в цих країна в межах 5%, а в Японії вона близька до нуля, а в деякі роки від’ємна (дефляція).

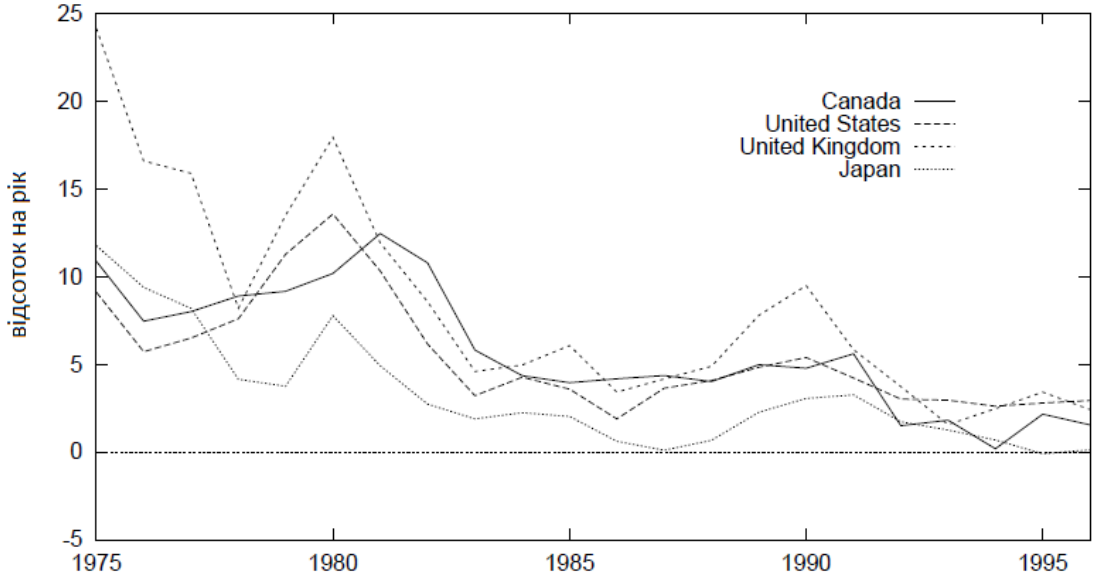


Рис.2.4. Дінамика щорічних індексів інфляції для декількох розвинутих країн

Позитивне значення sx,y (прямий зв'язок) сигналізує, що Xi має тенденцію бути більше його середнього значення, кожний раз коли Yi більше свого середнього значення. Змінні X та Y мають обернений зв’язок коли sx,y відємне. Це означає, що Xi має тенденцію бути менш власного середнього значення, кожний раз коли Yi більш власного середнього значення.

У випадку, коли ніякого зв’язку між змінними X та Y немає, sx,y наближується до нуля. В нашому прикладі про доход та кількість років навчання для домогосподарства, можна очікувати, що випадкова вибірка надасть sx,y позитивного значення, і це справді те, що було виявлено в фактичних вибірках для канадських домогосподарств. Зверніть увагу, що рівняння (9) може бути використане для розрахунку sx,x— коваріації змінної X з нею самою. Легко побачити з рівнянь (1) та (9), що це дасть дисперсію вибірки X, яка позначається через s2 . Таким чином можна сказати, що концепція дисперсії являється лише окремим випадком більш загальної концепції коваріації.

Нажаль величину коваріації надзвичайно важко економічно трактувати. Наприклад якщо ковариація між рівнем доходів та освітним рівнем дорівнює 1000 доларів*рік це тільки означає що зростання років навчання сприяє зростанню рівня доходів. Більш зрозумілою є щільно пов’язаною з коваріацією поняття кореляції (або коефіцієнт кореляції), між двома змінними. Коефіцієнт кореляції між двома змінними X та Y, позначається як rx,y або ry,x і визначається як:

$$r_{xy} = \frac{s_{x,y}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2.5)$$

де sx та sy – середньо квадратичні відхилення вибірки X та Y, розраховані використовуючи формулу (1.1) і знаходження кореню квадратного з отриманої величини. Перевагою коефіцієнту кореляції є то що він на відміну від коваріації є без вимірна та масштабована величина. Це

означає що він вимірює щільність лінійного взаємозв'язку між двома випадковими змінними.

Що не очевидно (і не буде доказано на цей час) це те, що для двох будь-яких змінних X та Y :

$-1 \le r_{x,y} \le +1.$

Оскильки масіви вихідної інформації містять похибку, оцінки коефіцієнту кореляції також містять похибки. Інтуітивно зрозумило шо похибка зменшується з зростанням кількості спостережень. В подальшому нами будуть це питання формалізовано за допомогою рівня значимості.

2.3. Рівень значимості лінійного взаємозвязку

Звичайно вважається що ступінь взаємозв'язку зростає по мірі наближення модуля коефіцієнту до одиниці. Однак міра лінійного взаємозв'язку залежить не тільки від значення коефіцієнту кореляції, але і також від кількості спостережень по яким здійснюється оцінка. При малій кількості спостережень зростає похибка, тому, наприклад, значення коефіцієнту кореляції 0,9 зовсім не означає наявність щільного лінійного взаємозв'язку при незначній кількості спостережень. Рівень взаємозв'язку визначався на підставі критерію Стьюдента [11]:

$$t = \frac{r}{\Delta r}, \qquad (2.6)$$

де r – оцінка коефіцієнта кореляції, Δr – похибка, яка визначається:

$$\Delta r = \sqrt{\frac{1 - r^2}{n - 2}}, \qquad (2.7)$$

де n – кількість спостережень. Критерій значімості α визначається на підставі співвідношення:

$$|t| \ge t_{\alpha; n-2}, \qquad (2.8)$$

де $t_{\alpha; n-2}$ – табличне значення розподілу Стьюдента на рівні значімості α

.

За допогою рівня значимості ми перевіряємо гіпотезу наявності лінійного взаємозв’язку.

Нульова гіпотеза означає, що лінійний взаємозвязоу відсутен. За допогою рівня значимості ми відхіляємо нульову гіпотезу на рівні значимості α (це ймовірність похибки при відхиленні нульової гіпотези). Це означає, що впевненість з якої ми відхилі лі нульову гіпотезу дорівнює $p = 1 - \alpha, \alpha = 0{,}1; 0{,}05; 0{,}01; 0{,}001.$

Найбільший рівень значимості, який дозволяє вважати зміни взаємозалежні не перевищує 0,2. У табл.2.5. подано приклад розрахунку коефіцієнта кореляції між річними спостереженнями за експортом (col 1), імпортом (col 2) та ВВП (col 3) України протягом 1996 – 2006 років. Крім значення коефіцієнта кореляції в таблиці наведені кількість спостережень та рівень значимості (ймовірність відсутності лінійного взаємозв’язку). Матриця симетрична, тому може використовуватись у трикутному вигляді.

Рівень значимості менш 0,1 позначається однією зіркою, менш 0,01 двома, менш 0,001 трьома.

Оскільки рівень значущості при всіх коефіцієнтах кореляції менше за 0,01, то можна вважати що всі зміни, що досліджуються, мають щільний лінійний взаємозв’язок, що обумовлено загальним економічним трендом.

Табл.2.5. Розрахунок кореляції в програмі «Statgraphics»

Col_2	0,8036 *	
	(11)	
	0,0029	
Col_3	0,8752**	0,9894***
	(11)	(11)
	0,0004	0,0000

Розглянемо, як впливає кількість спостережень на оцінку рівня значущості.

Нехай рівень лінійного взаємозв'язку (коефіцієнт кореляції) у першому випадку дорівнюю 0,9, кількість спостережень 4; у другому випадку коефіцієнт кореляції дорівнює 0,3 при кількості що дорівнює 100. Оцінити та порівняти рівень лінійного взаємозв'язку у першому та другому випадках.

Відповідно виразу (1.13) знайдемо похибки розрахунку коефіцієнту кореляції. У першому випадку похибка дорівнює 0,3, у другому 0,096. Значення t параметра (1.12) у першому випадку дорівнює 2,91, у другому 3,13. Порівняємо розраховані значення з критичними значеннями розподілу Стьюдента. Кількість ступенів свободи (n-2) у першому випадку дорівнює 2, у другому 98. Відповідні критичні значення, що задовольняють умову (1.14), у першому випадку дорівнює 1,89 на рівні значущості 0,1, у другому випадку дорівнює 2,63 на рівні значущості 0,005. Тобто у другому випадку рівень лінійного взаємозв'язку більш суттєвий, ніж у першому завдяки значно більшій кількості спостережень. Слід підкреслити що при великої кількості спостережень (більш 1000) щільний лінійний взаємозв'язок спостерігається при значенні коефіцієнту варіації в діапазоні 0,1-0,2.

Очевидно з виразу (1.11), що знак коефіцієнта кореляції такий самий, як і знак коваріації між двома змінними, оскільки середньо квадратичні відхилення не можуть бути негативними. Позитивна коваріація означає позитивну кореляцію, негативна коваріація означає негативну кореляцію і нульова коваріація означає, що X та Y не корелюють. Також з формули (1.11) очевидно, що rx,y не залежить від одиниць в яких X та Y вимірюються – це число без вимірвним.

Тобто, коефіцієнт кореляції між будь-якими двома змінними повинен знаходитись в інтервалі [-1,+1]. Значення в плюс, мінус одиницю означає, що дві змінні мають функціональну залежність (лінійна функція):

$$y = \beta_0 + \beta_1 x \qquad (2.8)$$

де β_1 додатний, коли коефіцієнт кореляції дорівнює 1 та відємний, коли він дорівнює -1. Якщо r_{xy} дорівнює нулю, то X та Y взагалі не корелюють.

Якшо коефіцієнт кореляції по модулю більш нуля але менш одиниці то в залежності (1.15) з'являється похибка лінійної моделі:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (2.9)$$

Дослідження останього рівняння представляє предмет регресійного аналізу.

Розглянемо взаємозв'язок між рівнем цін і рівнем інфляції в Канаді та США. Коефіцієнт кореляції між рівнями індексами споживчих цін в Канаді та США, які відображені на рис. 1.3 становить 0,996, що дуже близько до +1, це означає що точки спостережень знаходяться майже на одній лінії. Менша кореляція присутня між рівнем інфляції двох країн, що видно з більшого «розсіяння» точок на рис.1.4 навколо уявної прямої лінії, яку можна провести через них. В даному випадку коефіцієнт кореляції становить 0,839, що значно нижче коефіцієнта кореляції для інтегральних споживчих цін. Оскільки оцінки кореляції як і будь які інші статистичні оцінки містять похибку, цю похибку потрібно враховувати при оцінки щільності лінійного взаємозв'язку.

Якщо в регресійному рівнянні (1.16) зміни x,y є функції часу то відповідно принципу причинності зміна y не повинна випереджати зміну x

$$y(t+\tau) = \beta_0 + \beta_1 x(t) + \varepsilon \qquad (2.10.)$$

де $\tau \geq 0$ величина запізнення (лагу) процесу y(t) відносно x(t). Для визначення запізнення використається взаємно-кореляційна функція процесів.

В якості прикладу існування часового лагу між причиною та її наслідками можно розглянути процес інвестування в економіку, якщо наприклад вважати, що в деякий час t в економіку країни інвестовано I(t), то потрібен деякий час поки що виробництво буде налагоджено, а продукція буде реалізовано. Тобто якщо x(t) обсяг інвестувань на час t ці інвестиції сприяють зростанню доходів фірми y(t) тільки з деяким лагом тобто на час $t+\tau$

Величина лагу визначається за допомогою максимуму взаємно кореляційної функції [5]:

$$R_{xy}(\tau)=\frac{\sum_{t=1}^{T-\tau}\left(x(t)-\bar{x}\right)\left(y(t+\tau)-\bar{y}\right)}{T\sigma_x\cdot\sigma_y} \qquad (2.11)$$

Взаємно-кореляційна не є парною. Якщо ми шукаємо взаємно кореляційну функцію у вигляді (1.18), то вважається, що зміна x(t) впливає на зміну y(t) з деяким лагом, що визначається в процесі досліджень по максимальному значенню $\max R_{xy}(\tau)$. Однак цілком можливо що це припущення є хибним – максимальне значення досягається при $\tau<0$. Тоді слід поміняти змінні x(t) та y(t) у рівнянні (1.16), внаслідок існування принципу причинності. Якщо бог створил чоловіка, то спочатку був бог, а потім чоловік.

Програма «Statgraphics» реалізує процедуру розрахунку взаємно-кореляційної функцій відповідно виразу (1.18) як для додаткових так і для від'ємних τ.

Процеси x(t) та y(t) задано в табл.2.3Потрібно знайти зсув процесу x(t) відносно процесу y(t).

Табл.2.3 вихідна інформація для взаємно-кореляційних функцій.

T	1	2	3	4	5	6	7	8	9	10
x(t)	0	3	4	6	3	1	0	-2	-4	-5
y(t)	1	2	5	8	11	13	9	7	5	1

Знайдемо значення взаємно-кореляційної функції при $\tau=-3,-2,-1,0,1,2,3$ або використовуємо програму «Statgraphics» для побудови взаємно-кореляційну функцію процесів. Рис.1. 7. Максимум взаємно-кореляційної функції спостерігається при запізненні y(t) відносно x(t) на 2 часових лага.

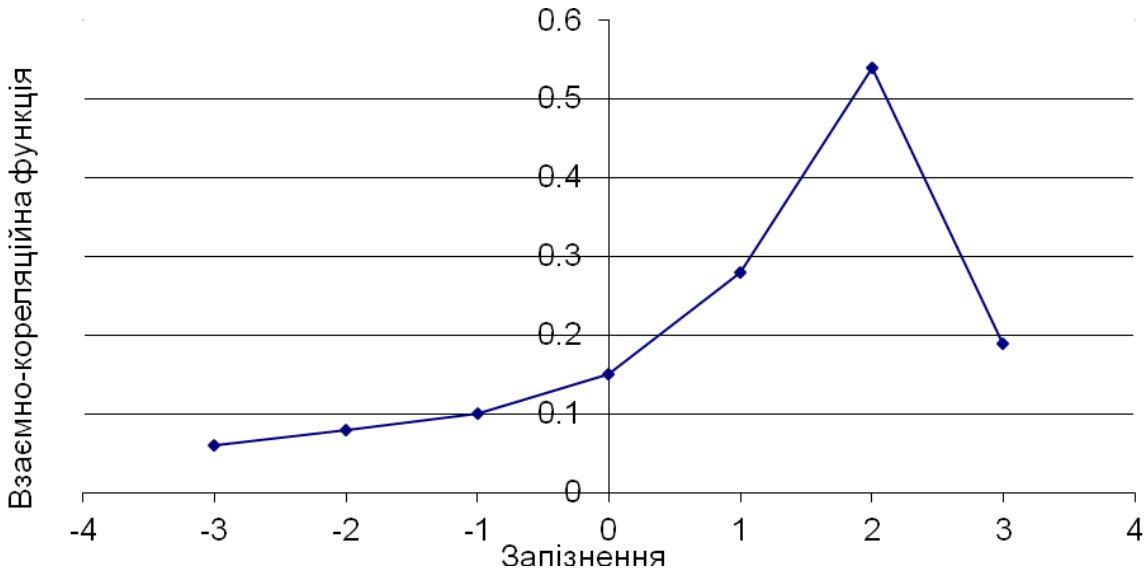


Рис. 2.6. Визначення лагу по графіку взаємно кореляційної функції

З урахуванням лагу буде розглядатись натупне регресійне рівняння.

$$y(t) = \beta_0 + \beta_1 x(t-2) + \varepsilon$$

Завдання до теми 2

По даним світово статистики спостережень за ВВП та індексом спріяття корупції оцінити рівень лінійного взаємизмязку між цімі показниками (кількість країн не менш 100. Якщо кількість літерів у прозвіще парна використати дані 2018 року, непарна 2019 року.

Тема 3. Ймовірність в статистиці

3.1. Простір подій

Межа між статистикою та ймовірністю надзвичайно важко помітна. Кваліфіковано трактувати деякі висновки що зроблено на підставі статистики можна тільки на підставі теорії ймовірностей. В цьому розділі будуть наведено приклади як теорія ймовірностей допомогає трактувати факти відомі з статистики. Наведемо приклад наближений до реального життя. Припустимо, що друг попросив Вас інвестувати $10 000 в спільний бізнес. Головним аргументом вашого другу є статистика згідно з якою більше 60% інвестицій в цей вид бізнесу призводять до успіху. Хоча представлення вашого друга про потенційний прибуток являлись переконливими, ви провели власний аналіз і дізнаєтесь, що він ініціював три попередні бізнес-проекти, всі з яких закінчили невдачею. Розглядаю це питання ви повинні задатися питанням об ймовірністю реалізації трьох невдач в виборці з трьох спроб з генеральної сукупності в якої більш ніж 60% елементів є вдачами. Неважко оцінити що ймовірність трех півторних невдалих інвестувань 0,4*0,4*0,4=0,064, тобто тільки приблизно 6,5% і це ні зовсім відповідає припущенню про 60% ймовірність успіху проекту. Цей шлях роздумів являється невід'ємною частиною процесу статистичних висновків, тому що ми постійно питаємо себе, яка ймовірність реалізації для конкретної вибірки, якщо вона зберігає характеристики загальної вибірки.

Більшість статистичних висновків включають постановку гіпотез про характеристики генеральної сукупності (яку ми пізніше назвемо нульовою гіпотезою) з подальшим аналізом того, чи вибірка має менший або більший шанс реалізації, якщо ця гіпотеза вірна.

Давайте почнемо вивчення поняття ймовірності з генеральної сукупності, чиї характеристики нам відомі і дослідимо, яка ймовірність або шанс отримати різни вибірки з відомої генеральної сукупності.

Пространства подій

Припустимо, що ми підкидаємо звичайну монету і спостерігаємо чи випаде вона орлом чи решкою. Релевантною вибіркою тут являється безкінечна послідовність підкидань монети. Для кожного підкидання існує невизначеність - чи буде результатом орел чи решка. Кожне підкидання монети являється прикладом випадкового експерименту, який може бути визначений як діяльність, що має два або більше можливих результати, з попередньою невизначеністю відносно того, який результат буде мати перевагу. Різні можливі результати випадкового підкидання називаються базиснимі результатами. Набір базових результатів для випадкового експеріменту називається простіром елементрних подій вибірки подкідання монетки. Вибірковий простір для підкидання звичайної монети позначимо через S, містить два базових результати - H (орел) і T (решка). Він репрезентує одну з нескінченої кількості вибірок підкидання монети. Набір базісних результатів можна записати:

$$S = \{H, T\} \quad (3.1)$$

Ці базісні результати також називаються точками вибірци або простими подіями. Вони являються взаємно виключними (несумісними) відносно один одного, тобто тільки один з ним може трапитись, та взаємного вичерпними, тобто щонайменше один з них повинен трапитись.

Тепер припустимо, що ми підкидуємо дві монети одночасно і записуємо якою стороною вони впали. В такому випадку існує чотири базисних результатів (якщо має зміст нумерувати монети): два орла, орел і решка, решка і орел, дві решки. Таким чином, простір елементарних подій або повна група подій для двох випадкових підкидань буде;

$$S = \{HH, HT, T H, T T\} \quad (3.2)$$

Підмножина цього простору називається подією. Наприклад, розглянемо подію «щонайменше один орел». Вона буде складатися з елементарних подій:

$$E1 = \{HH, HT, T H\} \quad (3.3)$$

що містить три з чотирьох елементарних. Іншою подією буде «обидві сторони однакові». Ця подія, яку ми назвемо Е2, буде складатись з наступних елементарних подій:

$$E2 = \{HH, T\,T\}. \quad (3.4)$$

Подія, що не містить подію Ej називається комплементарною (доповнюючу) подією до події Ej, яку позначимо Ejc. Таким чином комплементарними подіями до Е1 та Е2 являються, відповідно:

$$Ec = \{T\,T\}; \; E2c = \{HT;TH\} \quad (3.5)$$

Відповідно Набір елементарних подій які належать як події Ei так і події Ej називається перетин подій Ei та Ej . Перетин подій Е1 та Е2с є

$$E_1 \cap E_2^c = \langle HT, TH \rangle = E_2^c$$

Перетин Е1с та Е2с не містить ні яких елементів і це позначується:

$$E_1^c \cap E_2^c = O$$

Де О означає відсутні від елементів. Коли перетин двох подій є нульовою подією то вважаємо що ці події не сумісни, тобто реалізація однієї виключає реалізацію іншої. Перетин події та її компліментарної події є нульова подія. Компліментарну подію у відчізняної літературі називають протилежною.

Набір елементарних подій до якого належать як кожне з елементарних події Ei так і кожне з елементарних події Ej (одна одна враховуються один раз) називається поєднянням подій Ei та Ej. Наприклад поєдняння подій Е2 та Е1 є повна група подій:

В подальшому для спрощення ми використаємо поняття суми випадкових подій, хоч це не зовсім коректно.

Одномірний, двомірні та багатомірні простори вибірки

Простір вибірки отриманий з простого підкидання монети представляє собою одновимірний простір вибірки – існує тільки один вимір випадкового випробування. Коли ми підкидуємо дві монети одночасно, простір вибірки має два виміри – результат першого підкидання першої монети та результат

підкидання другої монети. Часто корисно відображувати двомірні простори вибірки у табличній формі:

	Перша	
Друга	Н	Т
Н	НН	ТН
Т	НТ	ТТ

Кожна з чотирьох клітинок таблиці представляє результат підкидання першої монети за яким слідує результат підкидання другої монети. Даний простір вибірки також може мати форму дерева:

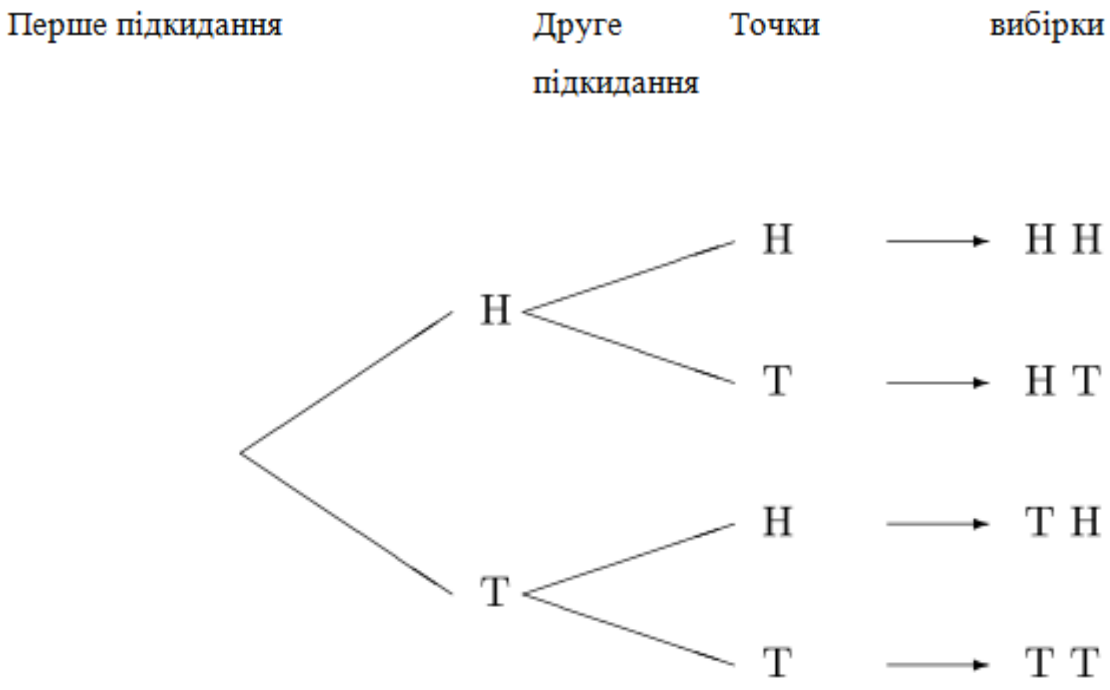


Рис.3.1. Дерево рішень для двократного кідання монетки

Більш цікавим прикладом може бути операція доставки запчастин для установок буріння нафтових свердловин транснаціональною фірмою. Відповідне випадкове випробування – доставка запчастини (випадковість має два виміру: 1) може бути доставлена не та запчастина, 2) тривалість поставки

також випадкова величина. Це, також, двомірне випадкове випробування, сутність якого можна відобразити наступною таблицею (табл.3.1):

Табл.3.1.Таблиця розподілу подій при постачанні запчпастин

		Time of Delivery		
		S	N	M
Order	C	C S	C N	C M
Status	I	I S	I N	I M

Статут замовлення має дві категорії: «правильна запчастина» (C) та «неправильна запчастина» (I). Час доставки має три категорії: «той самий день» (S), «наступний день» (N) та «більше ніж один день» (M). Існує шість точок вибірки базових результатів. Верхній ряд в таблиці представляє подію «правильна запчастина» і нижній ряд представляє події «неправильна запчастина». Кожне з цих подій містить по три точок вибірки. Перша колонка зліва таблиці представляє подію «той самий день доставки», середня колонка представляє подію «наступний день доставки» і третя колонка представляє подію «більше одного дня доставки». Кожна з цих трьох подій вміщає дві точки вибірки або базових результати. Подія «правильна запчастина доставлена менше ніж за два дня» буде розташовуватись зліва – дві точки вибірки в першому ряді, (C S) та (C N). Доповнення даної події, «неправильна запчастина або більше одного дня доставки» буде представлене іншими результатами (C M), (I S), (I N) та (I M).

Потрібно звернути увагу, що базові результати кожної клітинки вище наведеної таблиці є перетинанням двох подій. Наприклад (CS) є перетинання події C або «правильна запчастина» та події S «той самий день доставки», (IN) є перетинанням події I, «неправильна запчастина» та події N «наступний день доставки». Подія «правильна запчастина» є поєднянням трьох простих подій:

(C S) ∪ (C N) ∪ (C M).Простір вибірки доставки запчастин також можна представити у формі дерева:

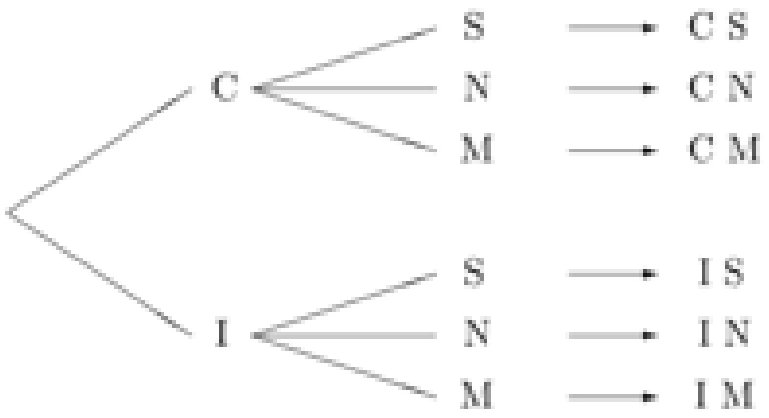


Рис.3.2. Дерево постачання запчастин на сверлину

Введемо формальне визначення ймовірності хоч це поняття ми вже використовували раніші. Якщо ми призводимо деякий експеримент для якого можливо n різних ісходів то в результаті експерименту виявиться один з ісходів, кожний з яких має власну ступінь невизначеності. Найпрстійший випадок кральний кубік на грані якого нанесено цифри від 1 до 6. Результативною вважається цифра на верх неї грані. В цьому випадку кожна з цифр може з однаковими шансами реалізуватися і тому всі ймовірності вважаються рівними 1/6. Тобто у випадку коли всі елементарні події кількістю n мають однакові шанси на реалізацію то ймовірність кожної з них дорівнює 1/n. Однак це найпростіший випадок, який також використовуються коли тільки відома загальна кількість можливих елементарних подій, але немає ніякої інформації відносно співвідношення шансів їх реалізації. В цьому випадку всі ймовірності вважаються рівнимі. Якщо подія E складається з m(E) елементарних подій, які мають рівни шанси реалізації, а загальна кількість елементарних подій дорівнює n то ймовірність E:

$$p(E)=m(E)/n \qquad (3.6)$$

Введемо деяку аксіоматику для ймовірності. Для будь якої події Е ймовірність знаходиться в межах від 0 до 1:

$$0 \le p(E) \le 1 \qquad (3.7)$$

Подія яка не коли не реалізується називається неможливою: р(НП)=0.

Подія яка реалізується в будь якому випадку є достовірною: р(ДП)=1

Якщо поді Е є поєднення несумісних елементарних подій Е1, Е2,…, Ек то:

$$p(E) = \sum_{i=1}^{k} p(E_i) \qquad (3.8)$$

Якщо події не є несуміснимі (тобто є сумісними) то вираз (3.8) ускладнюється. Наприклад для суми двох подій Е1 та Е2:

$$p(E)=p(E1)+p(E2)-p(E1)p(E2) \qquad (3.9)$$

Наведемо приклад. Ймовірність вступу до НАТО України в найближчі 10 років дорівнює 0,8, а вступу до ЄС -0,4. Яка ймовірність що відбудеться хоч дна з цих подій? Нехай шукана подія Е сума подій вступи до НАТО та ЄС Е1 та Е2. Ці події не є несумісними тому потрібно використати вираз (9):

$$P(E)=0,8+0,4-0,8*0,4=0,88$$

Поєднення події та її компланарної події створює достовірну подію (тому що завжди реалізується або деяка подія або компланарна (протилежна) подія:

$$p(E)+p(Ec)=1$$

$$p(E)=1-p(Ec) \qquad (3.10)$$

Останній вираз надзвичайно зручно використовувати для практичних розрахунків. Наприклад у вас було 20 незахищених сексуальних контактів, імовірність отримати ВІЛ в кожному з них 0,01(більш детальну інформацію можна отримати на сайті «digecon.com.ua»). Яка ймовірність, що у вас ВІЛ?

Прямий метод розрахунку надзвичайно тривалий (можна отримати ВІЛ після першого, другого та всіх інших контактів до 20 включно). Однак розрахувати компліментарну подію -не отримати ВІЛ за 20 контактів нескладно. Ймовірність неотримати ВІЛ за один контакт дорівнює

1-0,01=0,99. А за 20 контактів - 0,9920=0,82. Звідси відповідно (3.10) ймовірність отримати ВІЛ дорівнює 0,18.

Ймовірність та шанси (odds)

Крім поняття ймовірності, коли мова йде про прогноз деяких можливих подій, використається поняття шансів їх реалізації. Наприклад актуальною як для США так і всього світу є можливість імпічменту (дострокової відставки) президента Трампа. На 17.08.18 шансі на імпічмент і збереження посади до наступних виборів по даним ринку політичних прогнозів (Predictlt.Org.) мали відношення 2/3, однак після недавних свідчень колишнього адвокату Трампу на тему про плати мовчання його колишніх коханок воно зросло до 9/11. Перерахунок в ймовірності здійснюється наступним шляхом. Позначимо шанси за імпічмент a, збереження посади b, тоді ймовірність імпічменту до свідчень:

p(I1)=a/(a+b)=2/(2+3)=0,4 після p(I2)=9/(9+11)=0,45

Тобто ймовірність імпічменту зросла на 5%.

Як встанавлівати ймовірності в окремих випадках, коли це важливо для прийняття рішень. Призначення ймовірності відповідає розподілу одиничної маси в полі елементарних подій коли окремі частки пропорційні частоті появи цих подій. Наприклад для «справедливої» монети в разі дворазового підкидання це варіанти НН, ТТ, НТ, ТН яки мають однакові шанси на реалізацію і значить ймовірність кожного 0,25. Якщо визначити процес призначення ймовірностей то можна виділити дві категорії об’єктивна та суб’єктивна. К об’єктивним відноситься випадок коли ми маємо чітку модель експерименту: наприклад підкидання монетки, гра в рулетку, або держану лотерею. Наприклад при європейському варіанті гри в рулетку (один зеро), відповідно (3.6) ймовірність виграшу при ставці на парне число дорівнює 18/37. Наприклад ймовірність ½ для орла - решки для монетки може бути перевірено експериментальною. Для цього потрібно достатньо тривалий проміжок часу кидати монетку та розглядати відносну частоту випадіння

наприклад орла. Як слідує з рис.3.3 відносна частота випадіння орла по мірі зростання кількості експериментів наближується до ½.

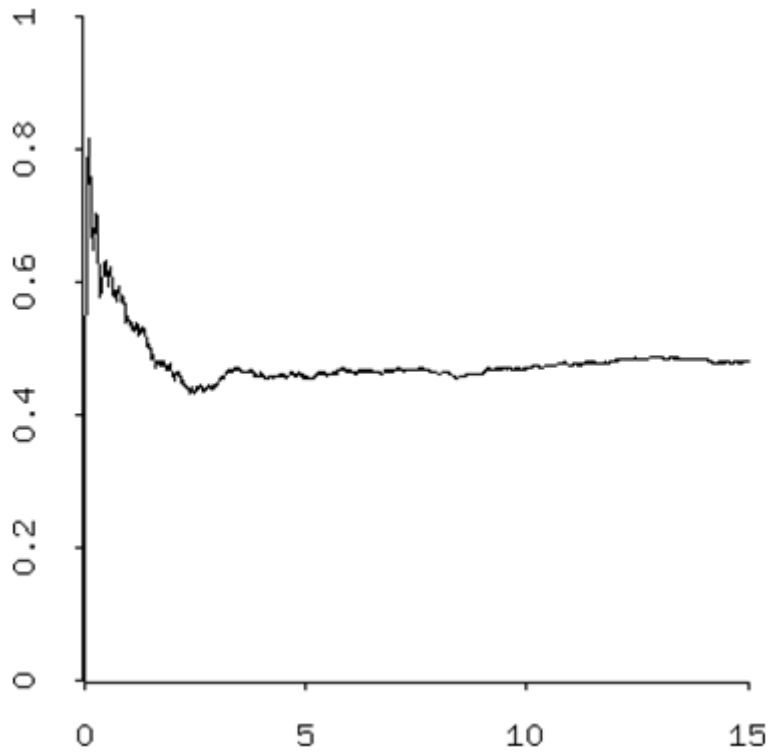


Рис.3.3. Графік відношення кількості орлов до загальной кількості пікірувань.

Тобто піводячи підсумки можна зробити висновок що існує два обєктивних підходу до визначення ймовірності: перший коли відома модель експеріменту (монетка, рулетка, лотерея) другій, коли прізводиться декілька вибірок для яких досліджується відносна частка реалізації події що досліджується, яка в результаті зростання обсягу вибірки прямує до деякої величини. Всі інши варіанти призначення ймовірності є субєкутивнимі однак це не означає що вони не можуть бути кориснимі. В якості прикладу розглянемо світову фінансовр –економічну кризу 2008-2009 року, в наслідок якої в Україні відбулоссь падіння ВВП більш ніж на 20%. Всі національні інституції що призвано оцінівати можливи варіанти розвитку подій вважали вплоть до другої половіни 2008 року що світова криза зовсім не вплине на економіку України (тобто ймовірність кризи національної економіки наближується до нуля). Однак непрямі ознаки кризи достатньо явно

проглядались і деякі наближені і суб'єктивні оцінки ймовірності падіння української економіки повіни були бути зроблено. В цьому випадку наслідки для України кризи 2008 -2009 були б суттєво меншими.

Призначення ймовірностей в двовимірних просторах вибірки можуть бути легко візуалізовано використовуючи приклад постачання запчастей для нафтових свердлін (табл.3.1). Наступна таблиця , яка ілюструє наш попередній приклад міжнародної доставки запчастин до місць буріння нафти (табл.3.2).

Табл.3.2. Таблиця ймовірностей двовимірного розподілу подій при постачанні запчастин

		Час доставки			
		S	N	M	Сума
Статут Замовлення	C	.600	.24	.120	.96
	I	.025	.01	.005	.04
Сума		.625	.25	.125	1.00

Ймовірності були призначено шістьом елементарним подія як чисто суб'єктивно так і використовуючи дані про частоти. Ці ймовірності, представлені цифрами в центральному прямокутнику, повинні в сумі дорівнювати одиниці, оскільки вони охоплюють весь простір вибірки – щонайменше одна з усіх точок вибірки повинна реалізуватись (повна група подій). Вони називаються сумісними ймовірностями (joint probabilities), оскільки кожна з них є перетином двох подій – подія «статус замовлення» (C чи I) та подія «час доставки» (S, N, чи M). Ймовірності в правій колонці та вздовж нижнього ряду називаються маргінальними ймовірностями (marginal probabilities). Ті, що знаходяться на правому краї, дають ймовірності подій «правильно» та «неправильно». Вони є сумою сумісних ймовірностей вздовж відповідних рядків і повинні в сумі становити одиницю, оскільки доставлене замовлення може бути або правильним або неправильним. Маргінальні

ймовірності вздовж нижнього ряду є ймовірностями подій «той самий день доставки» (S), «наступний день доставки» (N) та «більше одного дня доставки» (M). Вони є перетинами сумісних подій в відповідних колонках та повинні також в сумі дорівнювати одиниці тому що всі замовлення в кінцевому результаті доставляються. З таблиці можна дізнатись, що ймовірність правильного замовлення, яке доставлено в менше ніж два дня становить 0,60+0,24 = 0,84 і ймовірність незадовільного виконання (неправильна доставка або доставка с терміном два дня і більше) становить 0,12+0,025+0,01+0,005=0,16=(1-0,84).

3.2 Умова ймовірність

Можна спитати, яка ймовірність відправки правильного замовлення, коли доставка відбувається в той самий день. Потрібно звернути увагу, що це відрізняється від ймовірності як відправки правильного замовлення так і доставки в той самий день. Ймовірність отримання правильного замовлення з умовою доставки в той самий день тому називається умовною ймовірністю. Дві варіанту можуть трапитись, коли доставка відбувається в той самий день: відправлене замовлення може бути правильним, або неправильним. Як видно з таблиці вага ймовірності отримання замовлення в той же день 0,600+0,0250=0,625. Відношення ймовірності отримати вірне замовлення в той же день до сумарної ймовірності отримати замовлення в той же день 0,600/0,625=0,96 називається умовною ймовірністю отримати вірне замовлення. А відношення отримати невірне замовлення до загальної ймовірності є умовна тримати невірне замовлення 0,025/0,625=0,04 присвоєне події «неправильне замовлення». Умовна ймовірність (conditional probability) визначається як:

$$p(C/S) = \frac{p(C \cap S)}{p(S)} \qquad (3.11)$$

Якщо задано ймовірність події S та відома умовна ймовірність C але сумісна ймовірність C I S невідома то можна прорахувати ймовірність спільної родії наступним шляхом:

$$P(C \cap S) = P(C|S)P(S). \qquad (3.12)$$

Таким чином, якщо ми знаємо умовну ймовірність C, за умовою S (вона дорівнює 0,96) і ймовірність події C дорівнює 0,625, але при цьому не задана сумісна ймовірність C та S, можна розрахувати сумісну ймовірність, як добуток: 0,625 * 0,960 = 0,600.

3.3 Задача про розподілений борг

Розглянемо задачу о розподіленому боргі, тобто існує два варіанту боргових забовязань: перший дати в борг 100 USD одному чоловіку, другій дати в боргу ту же суму чотирем. Ймовірність неповернення 10%. Потрібно прояснити в якому випадку ризики кредитора більше. Розглядається одноразове кредитування на короткий термін під 20%. Вважається, що існує тільки два варіанту повна віддача боргу, або повне не повернення. В якости показнику ступеня ризику проекту використаємо математичне очікування та дисперсію.

Розглянемо перший випадок-один боржник. Сума яка повинна бути повернено дорівнює 100*1,2=120 USD з ймовірністю 0,9 у випадку неповернення кредитор отримує 0 з ймовірністю 0,1.

Для оцінок ступеня ризику кредитору використовуються стандартні оцінки для математичного очікування, дисперсії та коефіцієнту варіації:

$$\bar{x} = \sum_{i=1}^{i=n} x_i p_i; \sigma^2 = \sum_{i=1}^{n} x_i^2 p_i - \bar{x}^2; V_1 = \frac{\sigma}{x} 100\%$$

Звідси очікувана сума повернення у першому випадку дорівнює дорівнює $\bar{x}_1 = 120 \cdot 0{,}9 = 108 USD$;

$$\sigma_1^2 = 120^2 0{,}9 - 108^2 = 1296 \Rightarrow \sigma_1 = 36 USD \Rightarrow V_1 = \frac{36}{108} 100\% = 33{,}3\%$$

Зробимо відповідні оцінки для 4 позичальників. В цьому випадку потрібно розрахувати ймовірність повернення всіх чотирьох позичальників (m=4); трьох (m=3); двох (m=2); одного(m=1); ні одного (m=0). За допомогою біноміального розподілу отримаємо:

$$p(m) = C_4^m p^m (1-p)^{4-m}; m = 4;3;..;0$$

Сума, що повертається кожним з позичальників дорівнює 25*1,2=30USD, звідси сума що повертається: S(m)=30*m. Суми, що повертаються, та ймовірності повернення представлено в наступної таблиці:

Табл.3.3. Величини повернення та їх ймовірності у випадку 4 позичальників

m	4	3	2	1	0
S(m)	120	90	60	30	0
p(m)	0,6561	0,2916	0,0486	0,0036	0,0001

Знайдемо очікувану суму повернення та його дисперсію у випадку 4 позичальників:

$$\bar{x}_2 = \sum_{m=0}^{4} S(m) \cdot p(m) = 120 \cdot 0{,}6561 + 90 \cdot 0{,}2916 + 60 \cdot 0{,}0486 + 30 \cdot 0{,}0036 + 0 \cdot 0{,}0001 = 108 USD$$

$$\sigma_2^2 = \sum_{m=0}^{4} S^2(m) \cdot p(m) - \bar{x}^2 = 120^2 \cdot 0{,}6561 + 90^2 \cdot 0{,}2916 + 60^2 \cdot 0{,}0486 + 30^2 \cdot 0{,}0036 - 108^2 = 324 USD$$

$$\sigma_2 = 18 USD; V_2 = 16{,}7\%$$

На підставі розрахунків можна зробити висновок, що очікувана сума повернення в обі двох випадках однакова, однак дисперсія у випадку одного боржника в 4 рози більш тому ризик позичання у другому випадку суттєво менший. По суті має місто звичайний ефект диверсифікації портфелю позичальників.

Як вже згадувалось раніше характеристики процесу необов'язково мають кількісний приклад (наприклад стать співпрацівника деякої фірми, позитивний або негативний відгук на дисертаційне дослідження). Зараз ми розглядаємо випадкови результати експерименту які мають цифрові значення. Наприклад виграшні номера розіграшу державної лотереї 6 з 53 можуть бути яки чисел від 1 до 53, залишки на рахунку банку може бути деяке число, з великою кількістю знаків після коми. Всі ці значення які нам невідоми до результатів експерименту (розіграш лотереї, кількість коштів на рахунку, ринкові індексі) називаються випадковими змінними. Випадкові зміни

можуть бути дискретнимі (наприклад кількість сонячних днів у 2020 році) або неперервнимі як наприклад індекс цін споживчого ринку на наступний місяц, однак завдяки округленню неперервна величина може бути подано у вигляді декретної. Суттевою відмінністю дискретної та неперервної випадкової величини є то ще перший дискретні можуть бути перераховані (кожному значенню надано порядковий номер), а неперервні ні. Слід підкреслити що дискретні випадкові величини це не обов'язково цілі числа, наприклад кількість сонячних днів на наступний рік можна подати як відношення кількості днів до тривалості року. У випадку неперервних випадкових величин надати порядковий номер кожному стану неможливо. Крім того між двома і значеннямі неперервної випадковим величини можна завжди розтушувати трете значення.

3.3. Ймовірнісні розподіли випадкових величин

Розподіл дискретної випадкової величини X, яка приймає значення xi (i=1,2,3,…,n) задається ймовірностями кожного з дискретних значень $P(X = x_i)$. Це функція щільності розподілу дискретної величини, приклад такої функції подано в табл. 3.4 для кількості тижнів на лікарняному протягом наступного року. Функція розподілу дискретної випадкової величини X задається як ймовірність того ще випадкова величина не перевищує (менш або дорівнює) деяке з можливих значень $P(X \le x_i)$ для всіх x_i. Функція розподілу для дискретної випадкової величини подано в табл.3.4. Звичайно, що в цьому випадку останній елемент в останьому стовпчику повинен дорівнювати одиниці.

Табл. 3.4. Розподіл ймовірностей кількості тижней на лікарняному

x_i	1	2	3	4	5
$P(x_i)$	0,2	0,3	0,3	0,1	0,1

$P(X \le x_i)$	0,2	0,5	0,8	0,9	1,0

Розподіл неперервних випадкових величини неможливо подати через значення ймовірностей в окремих точках контініума (вона дорівнює нулю), тому має зміст розподіл за допомогою ймовірності знаходження випадкової величини на деякому інтервалі. Функція щільності $\varphi(x)$ розподілу є позитивно визначеная функція площа під якою що обмежено двома значеннями випадкової зміної відповідає ймовірності знаходження випадкової змінної на заданому інтервалі. Функція щільності розподілу обладає наступнимі властивостями:

$$\varphi(x) \ge 0;$$
$$\int_{-\infty}^{+\infty} \varphi(x)dx = 1$$
$$\int_{a}^{b} \varphi(x)dx = P(a \le x \le b) \qquad (3.13)$$

Функція розподілу або кумулятивна функція розподілу визначається через функцію щільності розподілу:

$$F(x) = \int_{-\infty}^{x} \varphi(u)du = P(X \le x)$$

де $-\infty \le x \le +\infty$

Функція розподілу дорівнює ймовірності що випадкова величина X менш або дорівнює будь якого визначеного x. Відповідно F(x) площі, що розміщено під функцією щільності розподілу зліва від значення x.

Завдання до 3 розділу

Функція щільності розподілу задано у виглґяді:

f(x)=Cx при (0<x<N)

f(x)=0 при x<0; x>N (N-кількість літерів у прізвіще)
Знайти C, математичне очікування та дисперсію.

Тема 4. Математичне очікування та варіація

4.1 Коваріація і кореляція

Розглянемо алгоритм розрахунку математичного очікуваної випадкової величини та її варіації. В подальшому ми розглянемо на прикладі казино економічний зміст математичного очікування. Для дискретної випадкової величини математичне очікування та дисперсія визначається:

(4.1)

Розрахуємо математично очікування та дисперсію для випадку кількості лікарняних (табл.3.4):

$$E(X) = 1 \cdot 0{,}2 + 2 \cdot 0{,}3 + 3 \cdot 0{,}3 + 4 \cdot 0{,}1 + 5 \cdot 0{,}1 = 2{,}6$$

$$\sigma^2(X) = 1 \cdot 0{,}2 + 4 \cdot 0{,}3 + 9 \cdot 0{,}3 + 16 \cdot 0{,}1 + 25 \cdot 0{,}1 - 2{,}6^2 = 1{,}44$$

Розглянемо зміст математичного очікування та дисперсії на прикладі казино. Розглянемо європейську рулетку (кількість чисел 37: 0,1,2,…,36), і простіший тип гри червоне – чорне або парне –непарне. Виграшний коефіцієнт 2. Розрахуємо ці характеристики у випадку одного двох та чотири гравців.

Нехай один гравец робит ставку в 100 USD на червоне. Виграш в цьому випадку складає 200 USD у випадку випадіння червоного з ймовірністю

$$p = \frac{18}{37} \approx 0{,}4865$$

Ймовірність прогришу

$$q = \frac{19}{37} \approx 0{,}5135$$

Відповідно (4.1) математичне очікування та дисперсія у випадку одного гравця:

$$E(X_1) = 200 \cdot 0{,}4865 \approx 97{,}3USD$$

$$\sigma(X_1) = 200^2 \cdot 0{,}4865 - 97{,}3^2 \approx 9992{,}7 \Rightarrow \sigma(X_1) \approx 99{,}96USD$$

У випадку двох та більш гравців використаємо біноміальний розподіл:

$$P_n^m = C_n^m p^m q^{n-m}$$

де n - кількість гравців, m –кількість виграшів, $C_n^m = \frac{n!}{m!(n-m)!}$ - кількість сполучень з n елементів по m. Ставка на одного гравця 100/n, виграш 100m/n.

Випадок двох гравців подано в табл.4.1

Табл.4.1. Виграші та їх ймовірності у випадку двох гравців (ставка 50 USD, виграш 100 USD)

Кількість виграшів (m)	2	1	0
Виграш (USD)	200	100	0
Ймовірн.	$p^2 = 0{,}4865^2 \approx 0{,}2367$	$2pq = 2 \cdot 0{,}486 \cdot 0{,}514 \approx 0{,}5000$	$q^2 = 0{,}514^2 \approx 0{,}2633$

Звідси очікуваний виграш та його дисперсія дорівнюють:

$$E(X_2) = 200 \cdot 0{,}2367 + 100 \cdot 0{,}5 \approx 96{,}3 USD$$
$$\sigma^2(X_2) = 200^2 0{,}2367 + 100^2 \cdot 0{,}5 - 96{,}3^2 \approx 5194{,}3 \Rightarrow \sigma \approx 72{,}1 USD$$

Розглянемо випадок 4 гравців (ставка 25 USD, виграш 50 USD)

Табл. 4.2. Виграші та їх ймовірності у випадку чотирьох гравців

Кільк. вигр.(m)	4	3	2	1	0
Виграш (USD)	200	150	100	50	0
Ймовір.	0,0560	0,2365	0,3745	0,2635	0,0695

$$E(X_4) = 200 \cdot 0{,}056 + 150 \cdot 0{,}2365 + 100 \cdot 0{,}3745 + 50 \cdot 0{,}2635 \approx 96{,}3 USD$$
$$\sigma^2(X_4) = 200^2 0{,}056 + 150^2 \cdot 0{,}2365 + 100^2 \cdot 0{,}3745 + 50^2 \cdot 0{,}2635 - 96{,}3^2 \approx 2691{,}3 \Rightarrow \sigma \approx 51{,}88 USD$$

Цій приклад наявно показує, що по міре зростання кількісті гравців очікуваний виграш залишається сталою величиною (прибуток казіно складє

2,7% від величини ставкі для європейських казино), а дисперсія виграшу зменшується.

Для дискретної випадкової поняття стандартизованої величини впроваджується наступним щляхом:

$$z_i = \frac{x_i - E(X)}{\sigma(X)} \tag{4.2}$$

Стандартизовано випадкова дискретна величина має математичне очікування, що дорівнює 0, і дисперсію, яка дорівнює 1:

$$E(Z) = 0; \sigma^2(Z) = 1$$

Для неперервної випадкові величини математичне очікування та дисперсія розраховуються за допомогою відомої функції щільності розподілу:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot \varphi(x) dx$$

$$\sigma^2(X) = \int_{-\infty}^{+\infty} (x - E(x))^2 \varphi(x) dx = \int_{-\infty}^{+\infty} x^2 \varphi(x) dx - E^2(x) \tag{4.3}$$

Функції щільності розподілу повинні задовольняти умовам (3.13)

Наведемо приклад розрахунку математичного очікування та дисперсії. Нехай функція розподілу задається у вигляді:

$$\varphi(x) = \begin{cases} 0; x < 0 \\ 2x; 0 \le x \le 1 \\ 0; x > 1 \end{cases}$$

Потрібно розрахувати математичне очікування, дисперсію, коефіцієнт варіації, ймовірність що випадкова величина менш 0,9; більш 0,1.

Відповідно (4.3) знайдемо математичне очікування та дисперсію:

$$E(x) = \int_0^1 x \cdot 2x \cdot dx = \frac{2}{3}$$

$$\sigma^2(X) = \int_0^1 x^2 \cdot 2x \cdot dx = \frac{1}{2}$$

$$F(x) = \int_0^x 2x \cdot dx = x^2$$

$$P(x \le 0{,}9) = F(0{,}9) = 0{,}81; P(x > 0{,}1) = 1 - F(0{,}1) = 0{,}99$$

У першому підрозділі було показано, що ко варіація і кореляція міра лінійного взаємозвязку між двома випадковими змінними, що задано у вигляді числових рядів. В цьому розділі ми розглянемо взаємозмязок між двома випадковимі змінимі що задано їх спільним розподілом. Для двох випадкових зміних коваріація визначається наступним шляхом:

$$\sigma(X;Y) = E\{(x_i - E(X)(y_j - E(Y)\} = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - E(X))(y_j - E(Y)P(x_i; y_j)$$

$$P(x_i; y_j) = P(X = x_i \cap Y = y_j)$$

(4.4)

$$\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_i; y_j) = 1$$

Розглянемо розрахунок коваріації і кореляції для випадкових зміних заданих у табличному вигляді.

X/Y	5	10	$\sum$
2	0,1	0,4	0,5
3	0,3	0,2	0,5
$\sum$	0,4	0,6	1,0

Дискретна величина X приймає значення 2;3 з ймовірностями 0,5 і 0,5, а У приймає значення 5;10 з ймовірностями 0,4 і 0,6. Оскілько для розрахунку коваріації потрібни математичні очікування X і У знайдемо математичні очікування та дісперсії X і У.

$$E(x) = 2 \cdot 0{,}5 + 3 \cdot 0{,}5 = 2{,}5$$
$$E(y) = 5 \cdot 0{,}4 + 10 \cdot 0{,}6 = 8$$

$$\sigma^2(X) = 2^2 \cdot 0{,}5 + 3^2 \cdot 0{,}5 - 2{,}5^2 = 0{,}25 \Rightarrow \sigma(X) = 0{,}5$$
$$\sigma^2(Y) = 5^2 \cdot 0{,}4 + 10^2 \cdot 0{,}6 - 8^2 = 6 \Rightarrow \sigma(Y) \approx 2{,}45$$

Випішемо у наявному вигляді спільні ймовірності розподілу:

P(X=2;Y=5)=0,1; P(X=2;Y=10)=0,4

P(X=3;Y=5)=0,3; P(X=3;Y=10)=0,2

Розрахуємо відповідно (4.4) величину коваріації:

$$\sigma(X;Y) = (2-2{,}5)\cdot(5-8)\cdot 0{,}1 + (2-2{,}5)(10-8)\cdot 0{,}4 + (3-2{,}5)(5-8)\cdot 0{,}3 +$$
$$+ (3-2{,}5)(10-8)\cdot 0{,}2 = -0{,}5$$

Коефіцієнт кореляції двох випадкових змінних визначається $\rho(X;Y)$:

$$\rho(X;Y) = \frac{\sigma(X;Y)}{\sigma(X)\cdot\sigma(Y)} \qquad (4.5)$$

Для наведених вище даних:

$$\rho(X;Y) = \frac{-0{,}5}{0{,}5\cdot 2{,}45} \approx -0{,}41$$

Знак коефіцієнту показує наявність оберненого співвідношення між випадковимі змінимі. Коефіцієнт кореляції між двома випадковими змінимі дорівнює коваріації між стандартізованимі формами цих зміних. Це відбувається тому що дисперсія стандартизованої змінної дорівнює одиниці.

Коваріація неперервних змінних дорівнює:

$$\sigma(X;Y) = E\{(x-E(x))(y-E(y)\} = \int\int (x-E(x))(y-E(y))\, f(x;y)dxdy \qquad (4.6)$$

Спільна функція щільності розподілу у випадку двох змінних задається:

$$f(x;y) = P(X \le x \cap Y \le y)$$

де $\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} f(x;y)dy = 1$

Якщо дві випадкові зміни є статистично незалежними коефіцієнт кореляції і звичайно коваріація обов'язково дорівнюють є нулю, однак протилежне не завжди виконується тому що можливо між змінним і існує нелінійний статистичний взаємозв'язок.

4.2 Лінійна функція випадкових змінних

Розглянемо лінійну функцію випадкової змінної X.

W=a+bX

Здійснемо операцію математичного очікування та дисперсії від ободвох частин цього рівняння. При цьому використуються наступна аксіоматика:

Математичне очікування від сталої величини дорівнює цій сталої

Сталий множник виноситься за знак математичного очікування

Дисперсія сталої величини дорівнює нулю

Сталий множник виноситься за знак дисперсії в квадраті

Отримаємо наступний вираз:

$$E(w) = E(a + bX) = a + bE(X)$$
$$\sigma^2(W) = \sigma^2(a + bX) = b^2\sigma^2(X) \qquad (4.7)$$

Сума і різниця випадкових зміних

Розглянемо, як розраховується математичне очікування та дисперсія суми, різниці та добутку двох випадкових змінних. Розглянемо суму двох випадкових змінних:

$$Z_1 = X + Y \qquad (4.8)$$

Математичне очікування та дисперсія визначаються:

$$E(Z_1) = E(X) + E(Y)$$
$$\sigma^2(Z_1) = \sigma^2(X) + \sigma^2(Y) + 2\sigma(X,Y) = \sigma^2(X) + \sigma^2(Y) + 2\sigma(X)\sigma(Y)\rho(X,Y) \qquad (4.9)$$

Для різниці двох випадкових змінних:

$$Z_1 = X - Y$$

$$E(Z_1) = E(X) - E(Y)$$
$$\sigma^2(Z_1) = \sigma^2(X) + \sigma^2(Y) - 2\sigma(X,Y) = \sigma^2(X) + \sigma^2(Y) - 2\sigma(X)\sigma(Y)\rho(X,Y) \qquad (4.10)$$

Розглянемо наступний приклад, що базується на статистикі податкових надходжень. Відоми математичні очікування від двох податків: податок на додану вартість (ПДВ) та податок на доходи підприємств (ППП). Важаємо що бюджет формується тільки від ціх двох податків. Потрібно проаналізувати статистичні характеристики бюджету у випадку наявності лінійного взаємозв’язку між цим і податками (коефіцієнт кореляції дорівнює 0,8) , у випадку її відсутності і у випадку оберненого взаємозв’язку.

Математичне очікування надходжень від ПДВ 650 млрд. грн., надходжень від податку на прибуток ППП 400 млрд. грн., відповідні стандартні відхилення дорівнюють 160 і 80 млрд.грн.(табл.). Якщо розглядати коефіцієнт варіації сумарних надходжень , як показник ступеня ризику, та порівняти з коефіцієнтами варіації ПДВ та ППП то виявиться, що для

сумарних надходжень він в будь якому випадку менш ніж для окремих податків. Наведемо варіант розрахунку диспесії з оберненим взаємозвязков (математичне очікування у всіх випадках однаково.

Табл.4.3. Статистичні характеристики суми двох випадкових змінних при різних рівнях лінійного взаємозв'язку (млрд. грн.)

	$E(X)$	$\sigma^2(X)$	$\sigma(X)$	V(%)
X_1	640	25 600	160	25
X_2	400	6 400	80	20
$Z_1 = X_1 + X_2 (\rho = 0{,}8)$	1040	42 240	205	19,7
$Z_2 = X_1 + X_2 (\rho = 0{,}0)$	1040	32 000	178,9	17,2
$Z_3 = X_1 + X_2 (\rho = -0{,}8)$	1040	21 760	147,5	14,2

$$E(Z_3) = E(X_1) + E(X_2) = 640 + 400 = 1040 млрд.грн$$
$$\sigma^2(Z_3) = \sigma^2(X_1) + \sigma^2(X_2) + 2\sigma(X_1)\sigma(X_2)\rho(X_1; X_2) = 25600 + 6400 - 160 \cdot 80 \cdot 0{,}8 = 21760$$

Дисперсія суми у випадку оберненого взаємозв'язку суттєво менш ніж у випадку прямого або повної відсутності.

Розглянемо випадок різниці двох випадкових змінних. Наявним економічним прикладом цього є випадок співвідношення (різниці) виручки від реалізації - X_1 та собівартості - X_2 в результаті якої фірма отримує прибуток - Z. Як правило виручка та витрати мають прямий взаємозв'язок внаслідок того що на них впливають однакові макроекономічни процеси (наприклад інфляція) .

Результати розрахунку статистичних характеристик двох випадкових процесів подано у табл.4.4.

Табл. 4.4. Статистичні характеристики різниці двох випадкових змінних при різних рівнях лінійного взаємозв'язку (млн. грн.)

	$E(X)$	$\sigma^2(X)$	$\sigma(X)$	V(%)
X_1	28	49	7	25

X_2	24	25	6	25
$Z_1 = X_1 - X_2(\rho = 0{,}8)$	4	6,8	2,6	64
$Z_2 = X_1 - X_2(\rho = 0{,}4)$	4	40,4	6,4	160
$Z_3 = X_1 - X_2(\rho = 0{,}0)$	4	74	8,6	215

У випадку різниці двох випадкових змінних дисперсія результату зменшується по міре зростання лінійного взаємозв'язку між випадковими змінними.

У загальному випадку коли розраховуються характеристики суми N незалежних змінних, то математичне очікування та диспесія визначаються наступним шляхом:

$$Z = X_1 + X_2 + ... + X_n$$
$$E(Z) = E(X_1) + E(X_2) + ... + E(X_n)$$
$$\sigma^2(Z) = \sigma^2(X_1) + \sigma^2(X_2) + ...\sigma^2(X_n) \qquad (4.11)$$

4.3. Випадок корельованиз змінних

У випадку незалежних змінних не мають значення знаки складових.

У випадку існування залежності між складовими рівняння для математичного очікування залишається у том же вигляді, але змінюється вигляд для розрахунку загальної дисперсії:

$$\sigma^2(Z) = \sum_{i=1}^{n}\sum_{j=1}^{n} sign\{E(X_i)\}sign\{E(X_j)\}\sigma(X_i)\sigma(X_j)\rho_{ij}$$

$$\rho_{ij} = 1,(i = j)$$

Добуток двох випадкових змінних

У практичних приложеннях зустрічаються випадки коли потрібно проаналізувати статистичні характеристики добутку двох випадкових змінних. Наявним прикладом цього може слугувати виручка фірми яка є добутком двох випадкових змінних ціни реалізації та фізичному обсягу реалізованої продукції. Аналогічна проблема актуальна і для аграрного сектору де обсяг деякої товарної продукції (наприклад пшениці) є добутком урожайності та площ, що відведених під цю продукцію.

Представимо процес, що досліджується - Z у вигляді добутку двох випадкових процесів з відомими статистичнимі характеристиками $X_1; X_2$:

$$Z = X_1 \cdot X_2 \tag{4.12}$$

Тоді математичне очікування та дисперсія добутку визначаються наступним шляхом:

$$E(Z) = E(X_1)E(X_2) + \sigma(X_1; X_2)$$
$$\sigma^2(Z) = E^2(X_1) \cdot \sigma^2(X_2) + E^2(X_2) \cdot \sigma^2(X_1) + 2E(X_1)E(X_2)\sigma(X_1; X_2) \tag{4.13}$$

Розглянемо задачу оцінки варіативності виробництва пшениці в Україні, країнах колишнього СРСР. Дані відносно показників варіативності площ та урожайностей подано у табл. 4.5.

Табл.4.5. Діскрипутивна статистика площ (млн. га) під пшеницю (Україна, країни колишнього СРСР , Світ) по даним 2000-2012 років

Characteristics	Kazakhstan	Russia	Ukraine	FSR	World	Share of FSR in the world (%)
Average	12.3	23.3	5.9	41.5	216.7	19.1
Variance	1.4	3.6	1.4	12.9	18.6	2.0
Standard deviation	1.2	1.9	1.2	3.6	4.3	1.3

Coeff. of variation (%)	9.7	8.1	19.6	8.6	2.0	7.1

Source: Own calculations based on data from (FAO, 2014)

Як слідує з наведених даних середня величина площ під пшеницею в Україні приблизно дорівнює 6 млн. га з достатньо високою варіативність (коефіцієнт варіації 18,6%). Цій показник суттєво більш ніж в цілому в світе, або в будь якої з країн колишнього СРСР.

Перейдемо до розгляду статистичних характеристик урожайності, що представлено в наступної таблиці. Для України показник урожайності декілька гірше ніж світовий однак варіативність суттєво перевищує світовий рівень(табл..4.6).

Табл. 4.6. Діскрипутивна статистика урожайності (т/ га) пшениці (Україна, країни колишнього СРСР , Світ) по даним 2000-2012 років.

Characteristics	Kazakhstan	Russia	Ukraine	FSR	World
Average	1,1	2,0	2,8	1,9	2,9
Variance	0,05	0,05	0,3	0,1	0,03
Standard deviation	0,22	0,23	0,6	0,24	0,16
Coeff. of variation	20,9%	11,5%	20,5%	13,1%	5,6%

В подальшому для аналізу процесу виробництва пшениці в Україні та країнах колишнього СРСР.Використовуються наступні позначення : Var (X)- дисперсія процесу X, cov(X;Y)- коваріація процесів X і У, S- площа посевів пшениці (млн. га), Y-урожайність (т/га), Pv- обсяг річного виробництва пшениці. Існує співвідношення:

$$Pv=Y*S$$

Табл. 4.7. Аналіз складових варіативності виробництва пшениці вУкраїні, країнах колишнього СРСР , Світі по даним 2000-2012 років.

Indicators		Kazakhstan	Russia	Ukraine	FSR	EU	World
	$\overline{Y}_i$	1.07	2.01	2.77	1.85	5.14	2.9

Yields (tonnes/ha)	$Var\left(\overline{Y}_i\right)$	0.05	0.05	0.32	0.06	0.09	0.03
	$\overline{S}_i^{\,2} Var\left(Y_i\right)$	7.6	28.8	11.35	101.29	64.04	1229.3
	$d_1(\%)$	66.3%	36.3%	28.1%	38.6%	60.2%	61.0%
$\rho\left(Y_i, S_i\right)$		0.32	0.89	0.85	0.87	0.38	0.72
$\mathrm{cov}\left(Y_i, S_i\right)$		0.09	0.39	0.56	0.76	0.09	0.50
Areas (million ha)	$\overline{S}_i$	12.26	23.33	5.94	41.53	26.12	216.69
	$Var\left(\overline{S}_i\right)$	1.40	3.57	1.36	12.90	0.65	18.63
	$\overline{Y}_i^{\,2} Var\left(S_i\right)$	1.62	14.42	10.46	44.14	17.03	156.95
	$d_2(\%)$	14.1%	18.1%	25.9%	16.8%	16.0%	7.8%
Mutual variation	$2\overline{S}_i \overline{Y}_i \mathrm{cov}\left(Y_i, S_i\right)$	2.25	36.21	18.57	116.77	25.33	628.64
	$d_3(\%)$	19.6%	45.6%	46.0%	44.5%	23.8%	31.2%
Model estimates	$E\left(Pv_i\right)$	13.26	47.27	17.04	77.57	134.29	629.42
	$Var\left(Pv_i\right)$	11.46	79.46	40.38	262.20	106.40	2014.90
Actual data	$E\left(Pv_i\right)$	13.26	47.27	17.04	77.57	134.29	629.42
	$Var\left(Pv_i\right)$	12.65	80.66	31.01	257.13	104.24	2016.52
Error of estimation	$\xi\left\{Var\left(Pv_i\right)\right\}$	9.4%	1.5%	30.2%	2.0%	2.1%	0.1%

Notes: d_1, d_2, d_3 1 частки дисперсії виробництва пшениці що пояснюються варіативністю урожайності, площ та їх спільним ефектом.

В таблиці наведено частки дисперсії процесу виробництва пшениці що пояснюються варіативністю урожайності, площ та спільним ефектом.

Завдання до теми 4.

Для американської рулеткі (два зіроу, 38 ціфр) оцінити доходность казино, при грі парноє/непарное і при ставці на одну ціфру.

Тема 5. Оцінка обсягу вибірки та закони розподілу випадкових величин

5.1 Закони розподілу випадкових величин

При проведенні експерименту надзвичайно важливим питанням є визначення обсягу вибірки, що потрібен для отримання надійних оцінок. Спочатку зробимо спробу оцінити обсяг вибірки коли потрібно з із заданою точністю визначити наприклад частку населення що підтримує деякого кандидатуру. Однак проведення кожного дослідження коштує деяку суму тому краще зазлегідь визначиться з мінімальним обсягом спостережень, що забеспечіть необхідну точність. Відповідно центральної предельної теоремі що вибіркова оцінка частки, що підтримує одного з кандидатів $\hat{p}$ при достатньому обсязі спостережень розподілено нормально відносно частки p в генеральної сукупності. Щоб побудувати довірчи інтервали для цій частки нам потрібно оцінити середньо квадратичне відхилення цій частки. Оскілько кількість вибірців X, що підтримує кандидатуру при загальної кількості вибірців n розподілено відповідно біноміального розподілу то дисперсія X дорівнює:

$$\sigma^2(X) = np(1-p)$$

Прорахуємо дисперсію частки:

$$\sigma^2(\hat{p}) = \sigma^2(\frac{X}{n}) = \frac{1}{n^2}\sigma^2(X) = \frac{p(1-p)}{n}$$

Змінюваємо p на його оцінку $\hat{p}$ отримаємо для середне-квадратичного відхилення частки оцінку:

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Незміщено оцінка:

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

Тоді $100(1-\alpha)$ довірчи інтервали для частки вибірцев в генерального сукупності:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (5.1)$$

$z_{\alpha/2}$ – табличне значення хвостів стандартного нормального розподілу, яке відповідає ймовірності $\alpha/2$.

Наведемо приклад.

У випадкової вибірці з 1000 потенційних виборців 210 підтримали діючого президенту. Це дає точкову оцінку для частки $\hat{p} = 0{,}21$. Тоді незміщена оцінка середньо квадратичного відхилення:

$$\hat{\sigma}_{\hat{p}} = \sqrt{0{,}21 \cdot 0{,}79 / 999} \approx 0{,}013$$

Розрахуємо 95% $\alpha/2 = 0{,}025$ довірчи інтервали для загальної сукупності (всіх вибірців). З таблиці хвостів нормального розподілу $z_{0{,}025} = 1{,}96$.

Звідси 95% довірчи інтервали для частки:

$$0{,}21 - 1{,}96 \cdot 0{,}013 \le p \le 0{,}21 + 1{,}96 \cdot 0{,}013 \Rightarrow 0{,}185 \le \hat{p} \le 0{,}235$$

Це означає, що підтримка колишнього президенту на час опитування з 95% ймовірності складає від 18,5% до 23,5% населення.

Відхилення від очікуваного значення до верх неї або нижньої межи має назву похибки виборці. З виразу (11) слідує, що похибка частки виборці дорівнює:

$$\Delta = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad (5.2)$$

Розрахуємо якім потрібен бути обсяг вибірці, що похибка частки не перевищувала деякий фіксований відсоток d. Отримаємо оцінку $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

.

Ця оцінка здійснюється до визначення обсягу вибірці n, тому обсяг вибірці не є відомим. Визначмо p при якому його серенеквадратичне відхилення має максимальне значення. При фіксованому n задача зводиться к максимізації виразу:

$$y(p) = p - p^2$$

Знайдемо критичні точкі:

$$y' = 1 - 2p = 0 \Rightarrow p = 1/2$$

Перевірімо, що це є максимум: $y'' = -2 < 0$

Тобто для знадходження обсягу вибірці ми використаємо припущенн повної невизначеності (максимуму ентропії). Звідси:

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{\sqrt{0{,}5 \cdot 0{,}5}}{\sqrt{n}} = \frac{0{,}5}{\sqrt{n}}$$

Оскилько похибка оцінки частки обмежено величиною d, то цю умову можна подати:

$$z_{\alpha/2} \cdot \frac{0{,}5}{\sqrt{n}} \le d$$

Звідсі достатній обсяг вибірці для забезпечення точності d на рівні значимості α:

$$n \ge \frac{0{,}25 \cdot z^2_{\alpha/2}}{d^2} \qquad (5.3)$$

Наприклад для забезпечення 2% максимальної похибки з 5% рівнем значимости $(d = 0{,}02; z_{0{,}025} = 1{,}96)$ потрібна вибірця не менш ніж:

$$n \ge \frac{0{,}25 \cdot 1{,}96^2}{0{,}02^2} = 2401$$

Якщо задати максимальну похибку в 1% то обсяг вибірці зрасте в 4 разі і настильні зрасте вартість проведення спостережень.

Розглянемо питання планування обсягу спостережень який відповідає заданим вимогам точності . Мова й де про оцінку математичного очікування. Існують дві оцінки математичного очікуванн для достатньо великих виборок (більш100 спостережень) і для малих виборок.

Для великих виборок 100 (1-α) довірчи інтервали для математичного очікування:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le E(x) \le \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \qquad (5.4)$$

Похибка оцінки h дорівнює;

$$h = \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \qquad (5.5)$$

Звідси обсяг виборкі необхідний для отримання оцінки математичного очікування з точністю h і з довірчою ймовірністю 100(1-α):

$$n = \frac{z_{\alpha/2}^2}{h^2}\sigma^2 \qquad (5.6)$$

Розглянемо це на прикладі оцінки очікуваної урожайності пшениці в деякому регіоні. Нехай нам потрібна ця оцінка з 99% ймовірністю та з похибкою що не перевищує 3 ц (h= 3 ц/га). Значенья середньо квадратичного відхилення для урожайності в даному регіоні дорівнює 15 ц/га ($\sigma = 15 ц/га$). З таблиці хвостів нормального розподілу знайдемо значення нормованої зміної $z_{0,005} = 2{,}57$

Звідси обсяг вибірці можна оцінити:

$$n = \frac{2{,}57^2 \cdot 15^2}{3^2} = 165$$

Тобто для того щоб оцінити очікувану регіональну урожайність з похибкою, що не перевищує 3 ц/га на рівні значімосты 1% потрібни використати спостереження не менш ніж на 165 фермах.

У випадку незначної кількості спостережень (n<100) заміст нормального розподілу використовується розподіл Стьюденту:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} t_{n-2;\alpha/2} \le E(x) \le \bar{x} + \frac{\sigma}{\sqrt{n}} t_{n-2;\alpha/2} \qquad (5.7)$$

де $t_{n-2;\alpha/2}$ – критичне значення розподілу Стьюденту з n-2 ступенями свободи та рівнем значимості $\alpha/2$.

Кожній випадковій величині відповідає деяка множина чисел – значень, яких вона може набувати. У результаті дослідів ці значення можуть набуватись з різною ймовірністю. Правило, що встановлює зв'язок між

можливими значеннями випадкової величини і їхніми ймовірностями, названо законом розподілу випадкової величини.

Закон розподілу дискретної випадкової величини – це відповідність між значеннями випадкової величини та їх ймовірностями. Може бути заданий таблично, графічно або аналітично.

Нехай випадкова дискретна величина X може набувати одне із n значень

$$x_1, x_2, ..., x_n.$$

Водночас кожне з цих значень величина X набуває з певною ймовірністю – відповідно

$$p_1, p_2, ..., p_n,$$

тобто p_1 – це ймовірність події "випадкова величина X набула значення x_1" ($X = x_1$),

p_2 – ймовірність події $X = x_2$,

…

p_n – ймовірність події $X = x_n$.

Звівши ці дані в таблицю, в першому рядку якої вказані значення, що їх набуває випадкова величина X, в другій – їхню ймовірність, отримаємо

X	x_1	x_2	…	x_i	…	x_n
P	p_1	p_2	…	p_i	…	p_n

Таблицю називають таблицею розподілу випадкової величини X. Переважно числа в першому рядку таблиці розподілу розміщують у порядку зростання.

Оскільки в результаті досліду випадкова величина X набуває одне з цих значень, сума несумісних подій

$$\{X = x_1\} + \{X = x_2\} + ... + \{X = x_n\}$$

є достовірною подією, ймовірність якої дорівнює 1. Тому для таблиці розподілу будь-якої випадкової величини X справедлива рівність

$$p_1 + p_2 + ... + p_n = 1 \qquad (5.8)$$

Наприклад, аеропорт може одночасно прийняти не більше п’яти літаків. Ймовірність того, що на певний момент часу на злітних смугах є один літак дорівнює 0,3, два літаки – 0,25, три літаки – 0,15, чотири літаки – 0,1, п’ять літаків – 0,05. Ймовірність того, що на певний момент часу на злітних смугах не буде жодного літака, дорівнює 0,15. В цьому прикладі випадковою величиною (позначимо її Y) є кількість літаків в аеропорту. Таблиця розподілу випадкової величини Y така:

Y	0	1	2	3	4	5
P	0,15	0,3	0,25	0,15	0,1	0,05

Отже, для заповнення таблиці розподілу випадкової величини X, потрібно виписати всі значення, які може приймати випадкова величина $(x_1, x_2, ..., x_n)$ і розрахувати відповідні ймовірності $(p_1, p_2, ..., p_n)$.

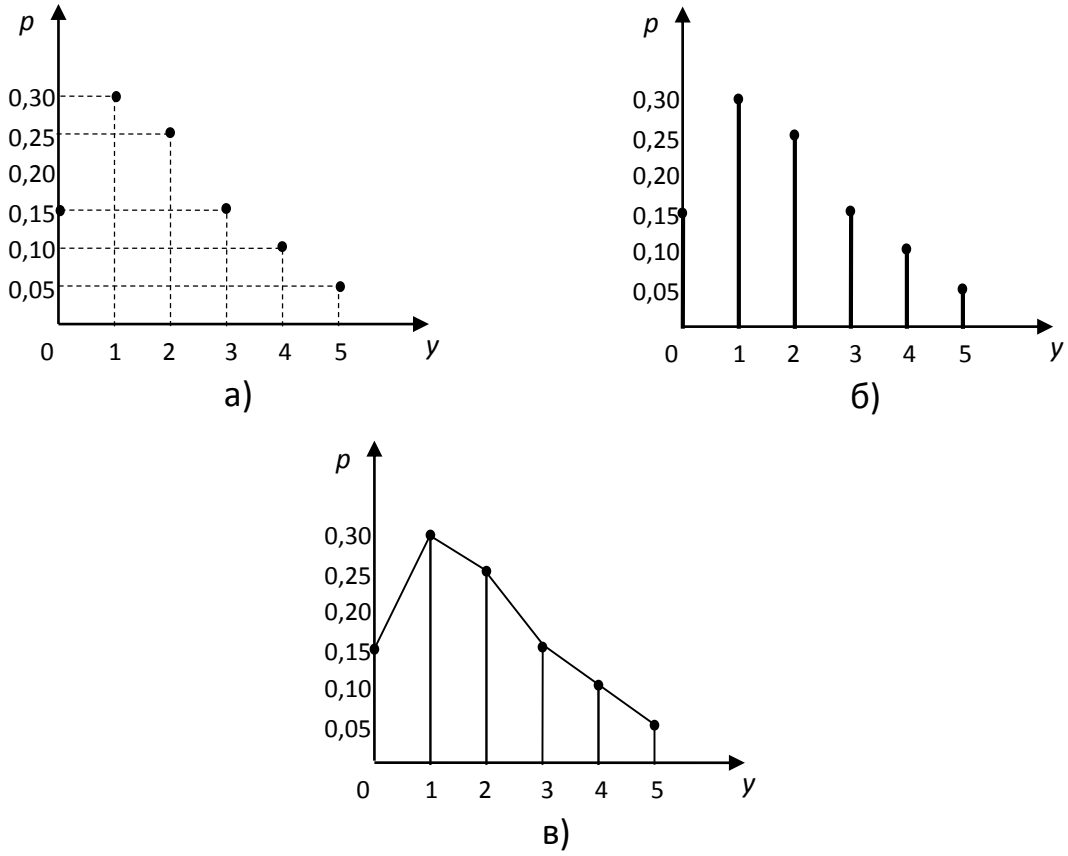


Рис. 5.1. Закон розподілу випадкової величини Y

Для більш наочного зображення закону розподілу часто використовують графічний спосіб. Для цього в системі координат по осі абсцис відкладають значення, яких набуває випадкова величина, по осі ординат – їхні ймовірності. На площині (x, p) позначають точки (x_1, p_1), (x_2, p_2),…, (x_n, p_n).

Аналітично закон розподілу випадкової величини може бути заданий за допомогою деякої функції, за якою можна знайти ймовірність p відповідного значення x_k, тобто

$$p_k = f(x_k),\ k = 1,2,\ldots,n.$$

Всі значення неперервної випадкової величини перебрати неможливо, тому її задають функцією розподілу.

Функцією розподілу (інтегральним законом розподілу) випадкової величини X називається функція

$$F(x) = P(X < x),$$

де $P(X < x)$ – ймовірність того, що випадкова величина X набуде значення, менше x.

Якщо неперервна випадкова величина може набувати будь-якого значення з проміжку (a, b), то

$$P(a < X < b) = F(b) - F(a),$$

тобто ймовірність прийняття випадковою величиною X значень з проміжку (a, b), дорівнює приросту функції розподілу.

Властивості функції розподілу:

$$0 \le F(x) \le 1.$$

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \quad F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

Ймовірність того, що випадкова величина X набуде значення з проміжку (α, β), дорівнює різниці значень функції розподілу на кінцях цього проміжку

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha).$$

$F(x)$ – неспадна функція, тобто якщо $x_1 \le x_2$, то $F(x_1) \le F(x_2)$.

Функцією щільності ймовірностей (диференційною функцією розподілу) $f(x)$ неперервної випадкової величини називають похідну першого порядку від її інтегральної функції розподілу

$$f(x) = F'(x).$$

Графік функції $f(x)$ – це крива розподілу. Часто для спрощення функцію щільності ймовірностей $f(x)$ називають густиною розподілу.

Ймовірність того, що випадкова величина X набуде значення з інтервалу (a, b), можна знайти за формулою

$$P(a < X < b) = \int_a^b f(x)dx$$.

Якщо функція щільності ймовірностей $f(x)$ відома, то інтегральну функцію $F(x)$ можна обчислити як

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Властивості густини розподілу

Густина розподілу як похідна від неспадної функції є невід'ємною $f(x) \geq 0$.

Інтеграл від густини розподілу, обчислений в області її існування, дорівнює одиниці, тобто

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$.

Якщо випадкова величина X розподілена на проміжку [a;b], то

$$\int_a^b f(x)dx = 1$$.

Ймовірність того, що випадкова величина X набуде значення з проміжку (α;β) дорівнює інтегралу від густини розподілу випадкової величини X, обчисленому в межах цього інтервалу

$$P(\alpha < X < \beta) = \int_\alpha^\beta f(x)dx$$.

Геометрично це означає, що ймовірність потрапляння випадкової величини X в проміжок (α, β) дорівнює площі криволінійної трапеції, обмеженої функцією $f(x)$, віссю Ox та прямими $x = \alpha$ та $x = \beta$.

Біноміальний розподіл

Біноміальний закон розподілу. Випадкова змінна розподілена за біноміальним законом з параметрами n та p ($n \in N,\ 0 < p < 1$), якщо вона набуває значень 0, 1,2,…, n з ймовірностями:

$$p(X = k) = C_n^k p^k q^{n-k} . \qquad (5.9)$$

Функція розподілу такої випадкової змінної має вигляд

$$F(x) = \begin{cases} 0, & x < 0, \\ \sum_{k=0}^{[x]} b(k, n, p), & 0 \le x < n, \\ 1, & n \le x. \end{cases} \qquad (5.10)$$

Математичне очікування і дисперсія для біноміального розподілу:

$$E(x) = np; \sigma^2 = npq \Rightarrow \sigma = \sqrt{npq}$$

Графік біномнорозподіленої випадкової величини зображено на рис. 5.2.

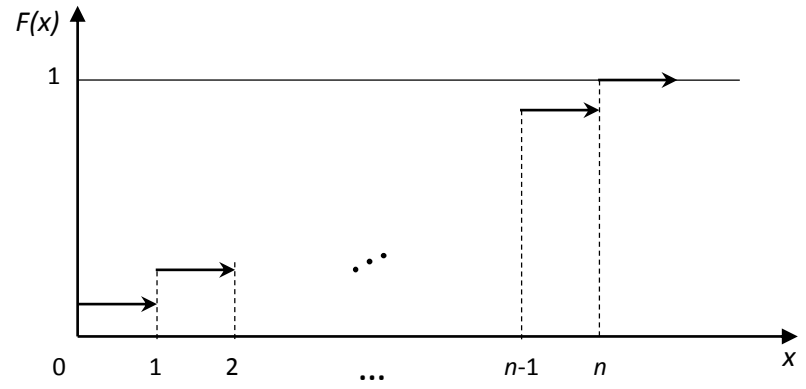


Рис. 5.2. Графік функції розподілу випадкової величини, що

Розподілено за біноміальним законом

Прикладом біномнорозподіленої випадкової величини може бути кількість страхових випадків, що настали упродовж певного часу, кількість покупців у супермаркеті, кількість пасажирів, перевезених залізничним транспортом упродовж певного періоду.

Приклад 1.

Нехай страхова компанія має 10 договорів страхування життя з 10 клієнтами. Ймовірність настання страхового випадку в цій групі клієнтів однакова і дорівнює 0,01. Величина виплат кожному клієнту однакова і дорівнює 2 тис. грошових одиниць. Побудувати закон розподілу для випадкової величини X – суми виплат за цим пакетом договорів.

Таблиця 5.1.Закон розподілу випадкової величини виплат

X	P	Розрахунок
0	0,904382	0,9910=0,904382
2000	0,091352	10·0,999·0,011=0,091352
4000	0,004152	(10·9)/(1·2)0,998·0,012=0,004152
…	…	…
k·2000	…	$C_{10}^{k}\,0{,}01^{k}\,0{,}99^{10-k}$
…	…	…
20 000	10-20	(0,01)10=10-20

Розподіл Пуассона

Якщо дискретна випадкова величина X може набувати тільки цілі невід'ємні значення з ймовірностями

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, \qquad (5.11)$$

то кажуть, що вона розподілена за законом Пуассона з параметром λ.

Аналітично закон розподілу такої величини задається за допомогою наступної функції:

$$F(x)=\begin{cases}0, & x<0,\\ \sum\limits_{k=0}^{[x]} p(k,\lambda), & x\geq 0.\end{cases} \qquad (5.12)$$

Для такої випадкової величини математичне сподівання і дисперсія рівні між собою і дорівнюють параметру λ.

$$E(x)=\lambda;\sigma^2=\lambda$$

Графік випадкової величини розподіленої за законом Пуассона зображено на рис. 5.3.

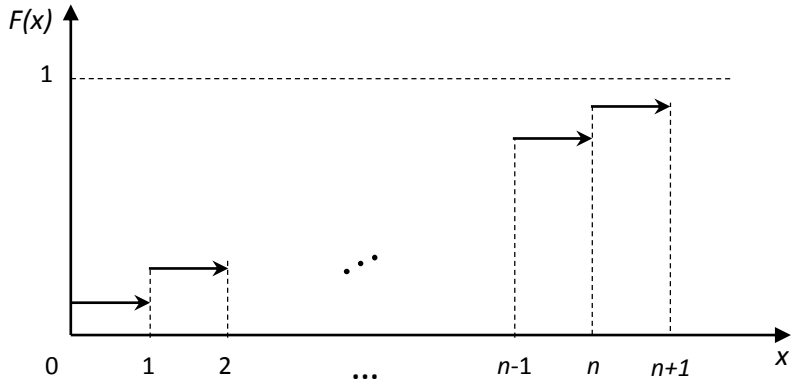


Рис. 5.3. Графік функції розподілу випадкової величини, розподіленої за законом Пуассона

Приклади випадкових величин, які розподілені за законом Пуассона: кількість клієнтів, яким будуть надані послуги у банку за певний час, кількість договорів, що будуть підписані підприємством за рік, кількість бракованої продукції у готовій продукції.

Приклад 2. Побудувати закон розподілу випадкової величини розподіленої за законом Пуассона при значенні параметра $\lambda = 3$.

Таблиця 5.2. Закон розподілу випадкової величини, розподіленої за законом Пуассона ($\lambda = 3$)

X	P	Розрахунок
0	0,0497871	$P(X=0)=\frac{3^0 e^{-3}}{0!}$
1	0,1493612	$P(X=1)=\frac{3^1 e^{-3}}{1!}$
2	0,2240418	$P(X=2)=\frac{3^2 e^{-3}}{2!}$
3	0,2240418	$P(X=3)=\frac{3^3 e^{-3}}{3!}$
4	0,1680314	$P(X=4)=\frac{3^4 e^{-3}}{4!}$
5	0,1008188	$P(X=5)=\frac{3^5 e^{-3}}{5!}$
6	0,0504094	$P(X=6)=\frac{3^6 e^{-3}}{0!}$
7	0,021604	$P(X=7)=\frac{3^7 e^{-3}}{0!}$
8	0,0081015	$P(X=8)=\frac{3^8 e^{-3}}{8!}$
9	0,00000017	$P(X=9)=\frac{3^9 e^{-3}}{9!}$
…	…	…

Рівномірний розподіл.

Випадкова величина X розподілена рівномірно на проміжку $[a,b]$, якщо усі її можливі значення належать цьому проміжку і функція розподілу має вигляд (2.13), а щільність її ймовірностей у цьому проміжку постійна, тобто

$$F(x)=\begin{cases}0, & x<a,\\ \dfrac{x-a}{b-a}, & a\le x\le b,\\ 1, & x>b.\end{cases} \tag{5.13}$$

$$f(x)=\begin{cases}0, & x<a,\\ \dfrac{1}{b-a}, & a\le x\le b,\\ 0, & x>b.\end{cases} \tag{5.14}$$

Графік функції рівномірно розподіленої випадкової величини при $a>0$ зображено на рис. 5.4.

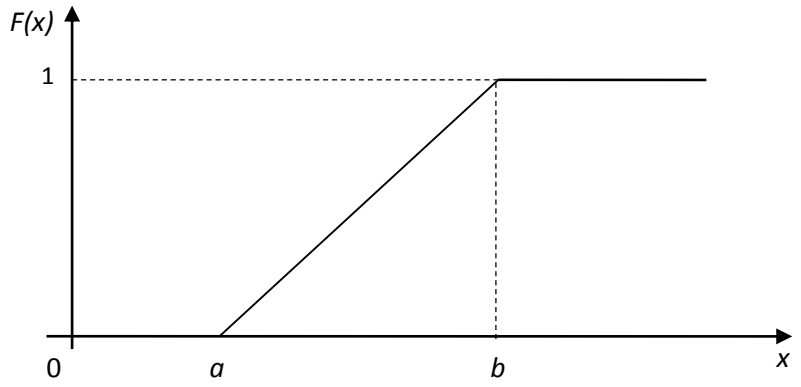


Рис. 5.4. Графік функції розподілу випадкової величини, рівномірно розподіленої на проміжку $[a,b]$

Графік функції щільності розподілу такої випадкової величини при $a>0$ зображено на рис. 5.4.

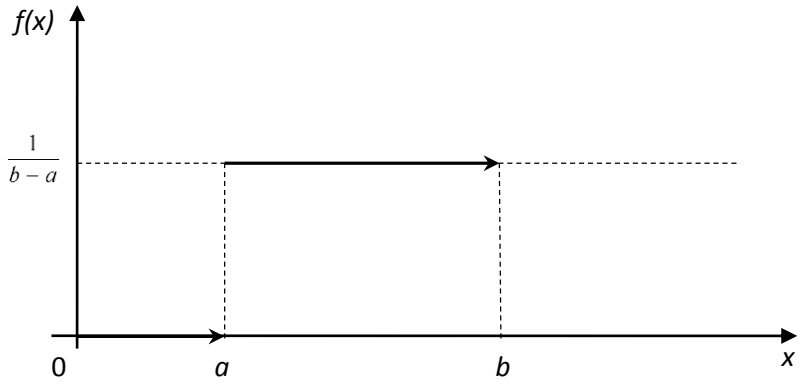


Рис. 5.5. Графік функції щільності розподілу випадкової величини, рівномірно розподіленої на проміжку $[a,b]$

За рівномірним законом розподілу можуть бути розподілені такі велични, як вартість акцій, обсяг продаж за певний період (рік, місяць, день), величина збитків у випадку аварії автомобіля, частка ринку певного товару.

Математичне очікування і дисперсія для рівномірного розподілу:

$$E(x)=\frac{a+b}{2};\sigma^2=\frac{(b-a)^2}{12}$$

Показниковий (експоненціальний) розподіл. Випадкову величину X називають розподіленою за експоненціальним законом, якщо її функція розподілу має вигляд

$$F(x)=\begin{cases}0, & x<0,\\ 1-e^{-\lambda x}, & x\geq 0.\end{cases} \qquad (5.15)$$

де λ – параметр розподілу, $\lambda>0$.

Графік функції (5.13) зображено на рис. 5.5.

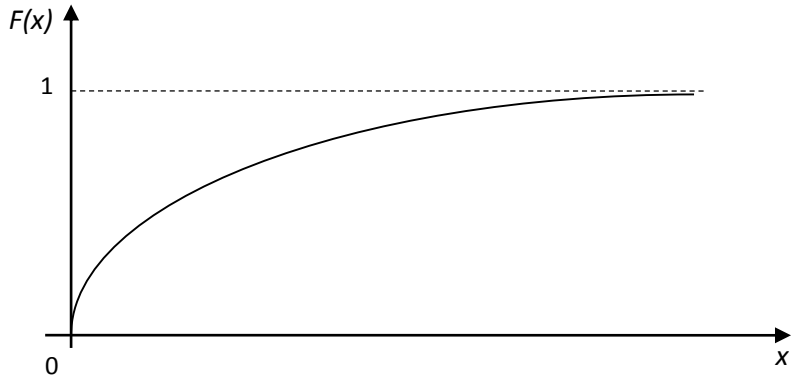


Рис. 5.6. Графік функції розподілу випадкової величини, розподіленої за експонентним законом

Густина розподілу випадкової величини розподіленої за експонентним законом:

$$f(x)=\begin{cases} 0, & x<0, \\ \lambda e^{-\lambda x}, & x\geq 0. \end{cases} \qquad (5.16)$$

Графік функції щільності розподілу такої випадкової величини зображено на рис. 5.6.

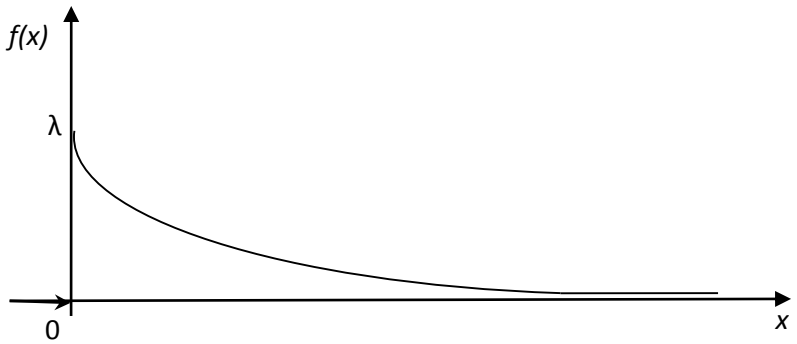


Рис. 5.7. Графік функції щільності розподілу випадкової величини, розподіленої за експоненціальним законом

Математичне очікування і дісперсія для експоненціального розподілу:

$E(x)=1/\lambda;\sigma^2=1/\lambda^2$

Прикладом випадкових величин, розподілених за показниковим законом можуть бути: час обслуговування одного клієнта банкоматом, тривалість телефонної розмови, проміжок часу між появою клієнтів у медичному центрі тощо.

5.2.Нормальний розподіл (Гаусса).

Випадкова величина X розподілена нормально, якщо функція її розподілу наступна:

$$F(x) = N(x; a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-a)^2}{2\sigma^2}} dt, \tag{5.17}$$

$$-\infty < x < +\infty; \ -\infty < a < +\infty; \ \sigma > 0.$$

Графік функції стандартной нормально розподіленої випадкової величини ($a = 0; \sigma = 1$) зображено на рис. 5.7. В подальшому позначується N(0;1).

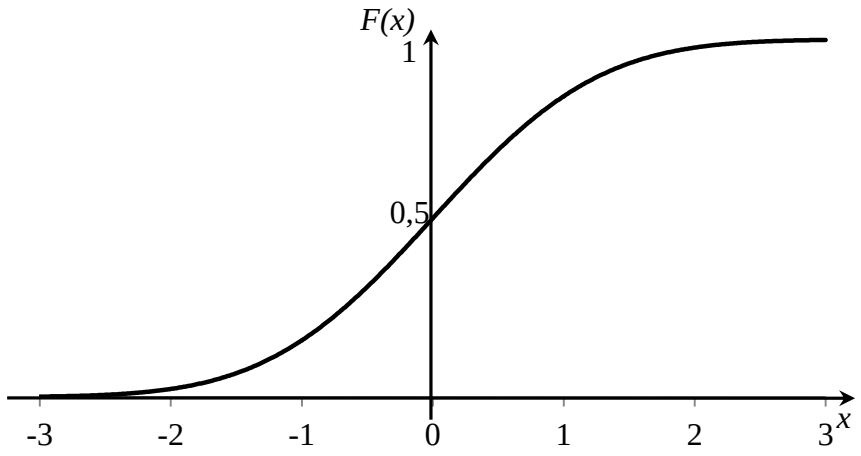


Рис. 5.8. Графік функції розподілу випадкової величини,розподіленої за нормальним законом розподілу при $a = 0; \ \sigma = 1$

Щільність розподілу рівномірно розподіленої випадкової величини

$$f(x) = n(x; a, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-a)^2}{2\sigma^2}}, \tag{5.18}$$

$$-\infty < x < +\infty; \ -\infty < a < +\infty; \ \sigma > 0.$$

Графік функції щільності розподілу такої випадкової величини зображено на рис. 5.9.

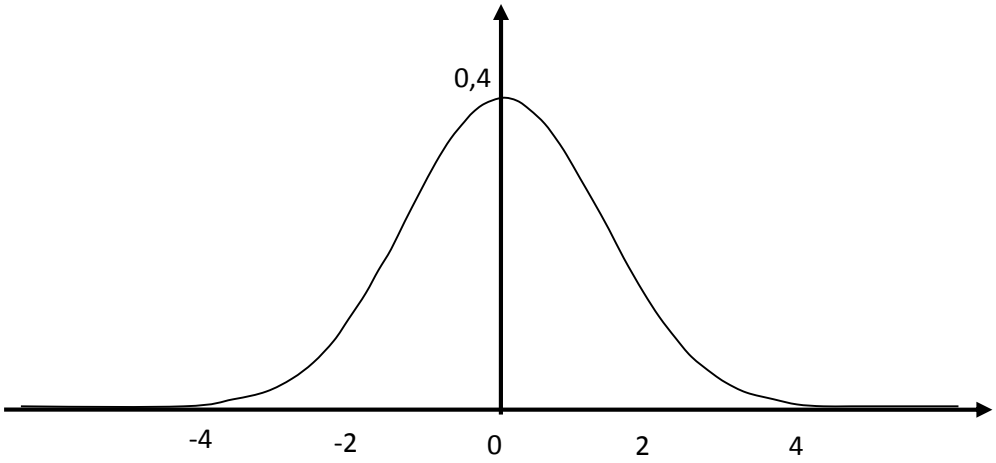


Рис. 5.9. Графік функції щільності розподілу випадкової величини, розподіленої за нормальним законом розподілу при $a = 0;\ \sigma = 1$

Випадкова величина розподілена за нормальним законом називається нормованою, якщо параметри розподілу $a = 0,\quad \sigma = 1$.

За нормальним законом розподілу можуть бути розподілені всі величини, що отримано обсяги виробництва продукції за день (місяць, квартал, рік), ціна товару, послуги, цінних паперів, величина інфляції тощо.

Закон розподілу для мішаної випадкової величини

Закон розподілу для мішаної випадкової величини задається комбіновано, тобто для її дискретних значень задаються ймовірності, а для значень із неперервної підмножини функція розподілу, або функція густини розподілу.

На практиці страхові компанії досить часто використовують схеми страхування, в яких передбачене повне відшкодування втрат клієнта при умові, що ці збитки не перевищують певного рівня (наприклад рівня а). Якщо ж збитки клієнта при страховому випадку є більшою від а, то йому виплачують відшкодування рівне а.

В таких схемах випадкова величина страхової виплати може розглятися як така, що має мішаний розподіл.

Якщо випадкова величина X може приймати дискретні значення x1, x2, x3,... з ймовірностями p1, p2, p3,... і значення з інтервалу (a,b), для якого задається функція густини розподілу f(x), то повинна виконуватися умова

$$p_1 + p_2 + p_3 + ... + \int_a^b f(x)dx = 1,$$

яка означає, що подія, яка полягає в тому, що випадкова величина набуває якогось значення з дискретної чи неперервної підмножин, є достовірною.

Приклад 3. Побудувати закон розподілу для випадкової величини виплат за одним договором при таких допущеннях, одержаних із аналізу статистичних даних та умов договору:

1. Ймовірність настання одного страхового випадку протягом періоду дії договору дорівнює 0,2;

2. Ймовірність настання більш ніж одного страхового випадку дорівнює 0;

3. Величина відшкодування при настанні страхового випадку не перевищує 2000 грн;

4. Ймовірність настання страхового випадку із втратами більшими рівними за 2000 грн дорівнює 0,05;

5. Функція густини виплат для значень з інтервалу (0;2000) є обернено пропорційною до x+100, тобто $f(x) = \frac{k}{x+100}$.

Випадкова величина виплат x приймає два дискретні значення x1=0 та x2=2000 з ймовірностями p1=0,8 та p2=0,05 відповідно і значення з неперервного інтервалу (0,2000), якому відповідає функція густини виду

$f(x) = \frac{k}{x+100}$. Для знаходження параметра k використовуємо умову $p_1 + p_2 + \int_0^{2000} f(x)dx = 1.$

$$0{,}8 + 0{,}05 + \int_0^{2000} \frac{k}{x+100}dx = 1;$$

$$\int_0^{2000} \frac{k}{x+100}dx = 0{,}15;$$

$$k \cdot \ln(x+100)\Big|_0^{2000} = 0{,}15;$$

$$k \cdot (\ln(2100) - \ln(100)) = 0{,}15;$$

$$k \cdot \ln(\frac{2100}{100}) = 0{,}15;$$

$$k = \frac{0{,}15}{\ln(21)} \approx 0{,}049269;$$

Отже, $p_1 = P(X = 0) = 0{,}8$; $p_2 = P(X = 2000) = 0{,}05$.

Для $x \in (0;2000)$ $f(x) = \frac{0{,}049269}{x+100}$.

Побудуємо інтегральну функцію розподілу $F(x)$.

Для $x \le 0$ $F(x) = P(X < x) = 0$ (ймовірність того, що величина виплат буде від'ємною рівна 0);

для $0 < x \le 2000$

$$F(x) = P(X \le x) = P(X = 0) + P(0 < X < x) = 0{,}8 + \int_0^x \frac{0{,}049269}{x+100} dx =$$

$$= 0{,}8 + 0{,}049269 \ln(x+100)\Big|_0^x = 0{,}8 + 0{,}049269\ (\ln(x+100 - \ln(100)) =$$

$$= 0{,}8 + 0{,}49269 \ln(1 + 0{,}01x);$$

для $x > 2000$

$$F(x) = P(X < x) = P(X < 2000) + P(X = 2000) + P(X > 2000) =$$

$$= 0{,}95 + 0{,}05 + 0 = 1.$$

Отже,
$$F(x) = \begin{cases} 0, & \text{для} \quad x < 0, \\ 0{,}8 + 0{,}049269 \cdot \ln(1 + 0{,}01x), & \text{для} \quad 0 \le x \le 2000, \\ 1, & \text{для} \quad x > 2000. \end{cases}$$

5.3. Граничні теореми теорії ймовірності

Нехай X – послідовність незалежних однаково розподілених випадкових величин з математичним сподіванням M(X)=a та дисперсією $D(X) = \sigma^2$.

Позначимо через Fn(x) функцію розподілу нормованої суми $\dfrac{\sum_{k=1}^{n} x_k - an}{\sigma\sqrt{n}}$,

тобто
$$F_n(x) = P\left(\frac{\sum_{k=1}^{n} x_k - an}{\sigma\sqrt{n}} < x \right). \qquad (5.19)$$

Позначимо через Φ(x) функцію розподілу стандартного нормального закону.

Тобто
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt .$$

Центральна гранична теорема: Нехай X – послідовність незалежних однаково розподілених випадкових величин з математичним сподіванням M(X)=a та дисперсією $D(X) = \sigma^2$. Тоді

$$P\left(\frac{\sum_{k=1}^{n} x_k - an}{\sigma\sqrt{n}} < x \right) \xrightarrow[n\to\infty]{} \Phi(x).$$

Нормальний закон розподілу має важливе значення у практиці економетричних розрахунків, оскільки часто зустрічається в ситуаціях, коли випадкова величина визначається великою кількістю незалежних випадкових факторів, жодний з яких не справляє переважаючого впливу.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Функція називається функцією Лапласа.

Властивості функції Лапласа:

$\Phi_0(-x) = -\Phi_0(x)$.

$\Phi_0(+\infty) = 1$.

$\Phi_0(-\infty) = 0$

Табличні значення функції Лапласа подано в доповненнях до методичних вказивік, крім того надано таблиця хвостів нормального розподілу:

$P(x' > x) = \Phi'(x)$

Інтегральна теорема Муавра-Лапласа

Якщо ймовірність настання певної події в кожній з n спроб однакова і рівна p і не дорівнює нулю та одиниці $(0 < p < 1)$, то $P_n(k_1, k_2)$ – ймовірність того, що подія появиться від k_1 до k_2 разів наближено рівна (чим більше n тим точніше)

$$P_n(k_1 k_2) = \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{t^2}{2}} dt = \Phi(x_2) - \Phi(x_1), \quad (5.20)$$

де $x_1 = \frac{k_1 - np}{\sqrt{npq}}, \quad x_2 = \frac{k_2 - np}{\sqrt{npq}}$, $\Phi(x)$ – функція Лапласа.

Локальна теорема Муавра-Лапласа

При достатньо великому n ймовірність того, що подія А відбудеться рівно k разів наближено дорівнює

$$P(\mu = k) \approx \frac{1}{\sqrt{npq}} \varphi(x), \quad (5.21)$$

де $x = \frac{k - np}{\sqrt{npq}}; \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ – щільність розподілу стандартного нормального закону.

Локальна теорема Муавра-Лапласа дозволяє наближено обчислити ймовірність появи події k разів в спробах, яка є незручною для обчислення за формулою Бернуллі при великих n.

Нехай ймовірність появи події А є функцією від n, тобто pn=p(n).

Теорема Пуассона. Нехай $p_n \xrightarrow[n\to\infty]{} 0$, так що $np_n \xrightarrow[n\to\infty]{} \lambda > 0$, тоді при достатньо великому n ймовірність того, що подія А відбудеться в схемі Бернуллі k разів, наближено дорівнює

$$P(\mu = k) \approx \frac{\lambda^k e^{-\lambda}}{k!}. \quad (5.22)$$

Приклад Портфель договорів страхової компаній містить 1000 договорів страхування життя. Ймовірність появи страхового випадку за період дії кожного договору в кожного клієнта однакова і дорівнює 0,02. Знайти ймовірність того, що протягом цього періоду буде 24 звернення до страхової компанії.

$$np = 20;$$

$$\sqrt{npq} = \sqrt{1000 \cdot 0{,}02 \cdot 0{,}98} = \sqrt{19{,}6} \approx 4{,}4272;$$

$$X = \frac{24 - 20}{4{,}4272} = \approx 0{,}9035;$$

$$P_{1000}(24) \approx \frac{\varphi(0{,}9035)}{4{,}4272} \cdot = \frac{0{,}2652}{4{,}4272} = 0{,}05991.$$

Приклад. За тих же умов обчислити $P_{1000}(16).$

$$X = \frac{16 - 20}{4{,}4272} \approx -0{,}9035;$$

$$P_{1000}(16) = \frac{(0{,}2652)}{4{,}4272} = 0{,}05991;$$

$$\varphi(-0{,}9035) = \varphi(0{,}9035) = 0{,}2652.$$

Значення функції $\varphi(x)$ беремо з таблиці Додатку 1.

Приклад За умов прикладів 6, 7 (n=1000, p=0,02) обчислити ймовірність того, що кількість страхових випадків буде знаходитися в межах від 16 до 24.

$$\sqrt{npq} = \sqrt{1000 \cdot 0{,}02 \cdot 0{,}98} = 4{,}4272;$$

$$x_1 = \frac{16 - 20}{4{,}4272} = -0{,}9035;$$

$$x_2 = \frac{24 - 20}{4{,}4272} = -0{,}9035;$$

$$P_{1000}(16;24) = 2\Phi(0{,}9035) - 1 = 0{,}6336$$

(Значення функції Лапласа беремо з таблиці Додатку 1).

Приклад За умов попереднього прикладу обчислимо ймовірність того, що кількість страхових випадків не більша 24.

Нехай X – випадкова величина кількості страхових випадків. Тоді

$$P(X \le 24) = P(0 \le X \le 24) = P_{1000}(0;24) = \Phi(x_2) - \Phi(x_1);$$

$$x_1 = \frac{0-20}{4{,}4272} = -4{,}5172;$$

$$x_2 = \frac{24-20}{4{,}4272} = 0{,}9035;$$

$$P(X \le 24) = \Phi(0{,}9035) - \Phi(-4{,}5175) = 0{,}3168 - (-0{,}5000) = 0{,}8168$$

.

Слід зауважити, що при малих значеннях добутку $p \cdot n$ (менших від 20), значення $P_n(k)$, обчислені за допомогою локальної теореми Лапласа, містять велику похибку. Так, наприклад, при $p = 0{,}02 \quad P_{500}(12) = 0{,}1038$ за локальною теоремою Лапласа, а за формулою Бернуллі $P_{500}(12) = 0{,}09555$, тобто відносна похибка становить майже 10%. В таких випадках розрахунок значень ймовірностей для випадкових величин, розподілених за біномним законом, краще виконувати використовуючи теорему Пуассона.

Використовуючи теорему Пуассона $p = 0{,}02$, отримаємо:

$$P_{500}(12) \approx \frac{(500 \cdot 0{,}02)^{12}}{12! e^{500 \cdot 0{,}02}} = \frac{10^{12}}{12! \cdot e^{10}} = \frac{\left(\frac{10}{e}\right)^{10} \cdot 100}{12!} = 0{,}09478.$$

За формулою Бернуллі $P_{500}(12) = 0{,}09555$.

Табл. 5.3. Основні характеристики розподілів

ДИСКРЕТНІ РОЗПОДІЛИ				
Тип розподілу	Пара-метри	Можливі значення k	Ймовірність	Математичне сподівання, дисперсія
Бернуллі	$0 < p < 1$	$k = 0{,}1$	$P(X = 0) = 1 - p$ $P(X = 1) = p$	$M(X) = p$ $D(X) = p(1-p)$

Біномний	$n \in N$ $0 < p < 1$	$k = 0,1,...,n$	$P(X = k) = C_n^k p^k q^{n-k}$	$M(X) = np$ $D(X) = np(1-p)$
Пуассона	$\lambda > 0$	$k = 0,1,...$	$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$	$M(X) = \lambda$ $D(X) = \lambda$
Геоме-тричний	$0 < p < 1$	$k = 0,1,...$	$P(X = k) = (1-p)^{k-1} p$	$M(X) = 1/p$ $D(X) = (1-p)/p^2$

НЕПЕРЕРВНІ РОЗПОДІЛИ			
Тип розподілу	Параметри	Щільність розподілу	Математичне сподівання, дисперсія
Рівномірний	$a \in R$ $b \in R$	$f(x) = \begin{cases} 0, & x < a, \\ \frac{1}{b-a}, & a \le x \le b, \\ 0, & x > b. \end{cases}$	$M(X) = \frac{a+b}{2}$ $D(X) = \frac{(b-a)^2}{12}$
Нормальний	$a \in R$ $\sigma > 0$	$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-a)^2}{2\sigma^2}}$	$M(X) = a$ $D(X) = \sigma^2$
Показниковий	$\lambda > 0$	$f(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \ge 0. \end{cases}$	$M(X) = \frac{1}{\lambda}$ $D(X) = \frac{1}{\lambda^2}$

Завдання до розділу 51

1.Монета підкидається 50N раз. знайти 99% довірчі інтервали для кількості випадінь решки (N-кількість літерів у прозвіщи).

2. Знайти довірчі 95% інтервали для середнього показнику урожайності по даним будь якої області.

Розділ 6. Однофакторні економетричні моделі

У процесі дослідження економічних явищ виникає потреба функціонально пов’язати змінні, за допомогою яких у відносно спрощеній формі описується це явище. Статистичну залежність, яка подається у вигляді певної математичної функції із випадковою змінною, називають економетричною моделлю. Економетричні моделі широко застосовуються у всіх сегментах соціально-економічного простору: від мікро-, мезо- та макроекономічних процесів до суспільно-політичних явищ.

Кожне економічне явище або процес описується низкою факторів, які, у свою чергу, взаємопов’язані між собою. Так, споживання на ринку благ деякого товару розглядається як функція доходу споживача, що залежить від ціни на товар у певний момент часу. У свою чергу, ціна на товар залежить від затрат на сировину, електроенергію, транспорт, а споживчі витрати можуть бути функцією доходу. На підставі наявних економічних знань можна встановити причинно-наслідкові взаємозв'язки між різними сторонами економічного явища. Ці пропозиції потім будуть розглядатись в якості початкових гіпотез, які будуть підтверджені або відхилені в результаті проведення економетричних досліджень.

Прикладом простих економетричних моделей можуть бути: дослідження функціональної залежності заощаджень громадян від їх доходу; дослідження, яким чином попит на товар залежить від його ціни; як впливає на надходження до бюджетів різних рівнів роздрібна торгівля; як отримати інформацію про перемогу того чи іншого кандидата на президентських виборах і, відповідно до цього, чи буде і як буде трансформована економіка країни і т. д.

Прикладом більш складних моделей можуть бути: моделі попиту на товар, який є функцією ціни, середнього доходу і ціни на конкуруючі товари; надходження до бюджету як функція роботи промисловості, сільського

господарства, транспорту, торгівлі, енергетики, банківської системи та інших чинників.

При побудові макро- та мікроекономічних моделей використовують чотири типи функціональних рівнянь:

- функцію поведінки, яка відображає переваги, що склалися у суспільстві. Наприклад, розподіл доходу між заощадженням (S) та споживанням (C) можна подати у вигляді залежності заощаджень від доходу (S=S(y)) або залежності споживання від доходу (C=C(y));
- технологічні функції, які характеризують технологічні умови виробництва (виробничі функції);
- інституціональні функції, що визначають інституціонально встановлені залежності між параметрами моделі. Так, сума податкових надходжень (T) є функцією доходу (y) та встановленої державою податкової ставки (Ty): T=Ty·y;
- дефініційні функції виражають залежності, що витікають із вербального визначення економічних явищ. Прибуток від виробництва у кейнсіанській концепції спрямовується домашніми господарствами на споживання, виплату податків і заощадження: y ≡ C+T+ S.

Таким чином, в економетрії кінцевим продуктом дослідження економічних процесів та явищ виступають моделі у вигляді рівнянь або систем одночасних рівнянь (так званих симулятивних моделей).

Першим кроком побудови будь-якої економетричної моделі деякого процесу є теоретичне (якісне) обґрунтування казуальне наслідкових взаємозв'язків досліджуваного явища. Такий підхід дозволяє висунути деякі гіпотези, які в процесі дослідження можна підтвердити або відхилити.

В основі економетричного моделювання лежить спосіб спрощення економічної дійсності до певної кількості найбільш суттєвих взаємозв'язків. Модель складається з двох груп елементів: відомих або так званих екзогенних (незалежних) або предетермінованих змінних та невідомих або ендогенних (залежних) змінних.

Економетричні моделі повинні відповідати певним вимогам і мати ряд відповідних властивостей:

- вони повинні бути по можливості відносно простими, оскільки для більшості випадків економічні закони описуються у відносно простій математичній формі (для правильної інтерпретації);
- вони повинні бути адекватно побудованими (із заданим рівнем точності);
- вони повинні не тільки правильно пояснювати економічний процес чи явище для минулих періодів, тобто проводити їх аналіз, але і давати достовірний прогноз для майбутніх періодів.

Не допускається побудова моделі, де ендогенна змінна є сумою екзогенних змінних. Наприклад, ендогенна змінна «валовий дохід підприємства», а екзогенні змінні «доходи від рослинництва та тваринництва даного підприємства».

Після з'ясування ролі та місця складових економічного процесу в рамках певної економічної теорії приступають до аналізу та дослідження зв'язків між змінними. Кількість зв'язків залежить від ряду обмежень: умов, при яких ця модель конструюється, а також від деталізації або класу моделі.

У загальному вигляді всі економетричні моделі можна розділити:

- за пояснювальною здібністю – на каузальні та некаузальні;
- за часовою ознакою – на статичні та динамічні;
- за формою побудови – на моделі, які складаються з одного рівняння, та моделі, які складаються з системи одночасних рівнянь (як їх ще називають – симулятивні моделі).

Каузальні моделі – це моделі, які досліджують причинно-наслідкові залежності між змінними. Тип даних – перехресні (ознаки, впорядковані у просторі).

Некаузальні моделі – це моделі, які досліджують залежність показника від зміни часу. Тип даних – часові ряди (ознаки, впорядковані за часом).

Некаузальні моделі або динамічні моделі часових рядів

Часовим рядом називається хронологічна послідовність спостережень певної ознаки у відповідні моменти часу. Отже, моделі часових рядів – це моделі, у яких у ролі незалежної змінної виступає вектор часу. Серед моделей часових рядів виділяють моделі тренду, сезонності, циклічності, автокореляцію та більш складні, які включають усі складові одночасно.

Тренд – це тривала тенденція зміни показника, який є детермінованою компонентою, відображається аналітичною функцією.

Сезонна компонента характеризує стійкі внутрішньорічні коливання рівнів ознаки, які є квартальними або місячними даними. В основі сезонних моделей лежать їх несезонні аналоги, доповнені засобами відображення сезонних коливань. Сезонні моделі можуть відображати як відносно постійні коливання, так і такі, які динамічно змінюються залежно від тренду. Першу форму відносять до класу адитивних, а другу – до класу мультиплікативних моделей. Стандартний період сезонних коливань – один рік.

Циклічна компонента відображає коливання економічних процесів упродовж тривалих періодів. Серед циклів, які найбільше впливають на економіку, виділяють: короткострокові (3–5 років) цикли Кітчіна (пов’язані з напругою на фінансовому ринку, зумовленою міжгалузевим переливом капіталу в межах ділового циклу)(пов’язані із запізненнями у часі інформації, що впливає на прийняття рішень підприємствами); середньострокові – ділові цикли або цикли Жугляра (близько 10 років, пов’язані зі строком життя машин і міжгалузевим переливом капіталу) (до часових запізнень, що характерні для циклів Кітчіна, включають і часові запізнення між прийняттям інвестиційних рішень і введенням в дію відповідних виробничих потужностей) та цикли Кузнеця (15-20 років, технологічні, інфраструктурні цикли, в рамках яких відбувається масове оновлення основних технологій); довгострокові (майже 50 років) – довгі хвилі Кондратьєва (пов’язані з інноваційними процесами в національній та глобальній економічних системах, а також зміною застарілих основних будівель, споруд та інфраструктури).

Модель тренду описується рівнянням виду: $Y_t = T_t + \varepsilon_t$,

де T_t – рівняння тренду певного виду (наприклад, лінійного, що характеризується сталим зростанням з часом – $T_t = \alpha + \beta \cdot t$; експоненціального, що характеризується сталим відносним зростанням – $T_t = A \cdot e^{\beta \cdot t}$); ε_t – стохастична компонента.

Сучасне програмне забезпечення пропонує більш ніж 20 елементарних монотонних функцій для апроксимації трендів. Використання поліномів другої та вищих ступенів може бути обґрунтовано тільки на інтервалі монотонності.

Модель сезонності: $Y_t = S_t + \varepsilon_t$,

де S_t – сезонна (періодична) компонента; ε_t – стохастична компонента.

Модель циклічності: $Y_t = C_t + \varepsilon_t$,

де C_t – циклічна (періодична) компонента; ε_t – стохастична компонента.

Більш складні моделі:

адитивна модель: $Y_t = T_t + S_t + C_t + \varepsilon_t$;

мультиплікативна модель: $Y_t = T_t \times S_t \times C_t + \varepsilon_t$.

До моделей часових рядів відносять також авторегресійні моделі – AR(p), моделі рухомого середнього – MA(q) та більш складні моделі – ARMA(p,q), ARIMA(p,d,q), ARCH(p,q), GARCH(p,q) та інші.

Каузальні моделі

Каузальними моделями з одним регресійним рівнянням описуються більшість економічних процесів та явищ. Наприклад, споживання деякого продукту на ринку можна описати рівнянням виду:

$$\ln C_t = \alpha_0 + \alpha_1 \cdot P_t + \alpha_2 Y_t + \varepsilon_t,$$

де P_t – ціна товару; Y_t – середній дохід на душу населення; a_0, a_1, a_2 – коефіцієнти рівняння.

Модель буде визначено, якщо будуть розраховані (оцінені) параметри a_0, a_1, a_2.

Симулятивні моделі або моделі одночасних рівнянь складаються з тотожностей та регресійних рівнянь. Прикладом симулятивної моделі може бути така модель:

$$C_t = a_1 + a_2 \cdot Y_t + \varepsilon_t,$$
$$I_t = b_1 + b_2 \cdot (Y_{t-1} - Y_{t-2}) + \varepsilon_t,$$
$$Y_t = C_t + I_t + G_t,$$

де Y_t – величина ВВП на час t; Y_{t-1} – величина ВВП на час $(t-1)$; Y_{t-2} – величина ВВП на час $(t-2)$; G_t – урядові витрати; I_t – інвестиції.

Це макроекономічна модель динамічного переходу економіки від одного сталого стану до іншого під дією зовнішніх або внутрішніх факторів (зростання інвестицій або державних витрат). Особливість цієї моделі полягає в тому, що рівень інвестицій залежить від рівня ВВП минулих періодів.

Модель буде визначено, якщо будуть розраховані параметри моделі: a_1, a_2, b_1, b_2.

Ці коефіцієнти мають такий зміст:

a_1 – автономне споживання (яке не залежить від ВВП);

a_2 – гранична схильність до споживання;

b_1 – автономне інвестування при сталому рівні ВВП;

b_2 – коефіцієнт, що визначає приріст інвестування залежно від приросту ВВП.

6.2. Статистичне оцінювання, довірчі інтервали

У практичних економетричних дослідженнях, які мають справу з рядом спостережень, параметри генеральної сукупності, як правило, не відомі, тому обчислюються оцінки цих параметрів на основі вибіркових даних. У математичний статистиці розглядаються два види оцінок:

- точкова оцінка;
- інтервальна.

При точковому оцінюванні робиться приблизний розрахунок, в результаті якого приходимо до єдиного результату оцінювання значення цього параметра генеральної сукупності. Часто нам доводиться приймати приблизне припущення, що той чи інший параметр генеральної сукупності дорівнює відповідній вибірковій статистиці (наприклад, середнє $E(x)$ ідентичне його оцінці $\overline{X}$).

При інтервальному оцінюванні йдеться про приблизний розрахунок інтервалу, в межах якого повинно знаходитись невідоме значення параметра генеральної сукупності. Існують різні методи оцінюван-ня. Серед них можемо виділити: метод найменших квадратів (МНК); метод максимальної правдоподібно-сті (МП); метод моментів (ММ) та метод Бейєса. Останні два не є предметом детального розгляду в даному посібнику, тому зупинимось на двох перших.

Розглянемо окремо питання побудови довірчих інтервалів для параметрів, що оцінюються у процесі використання регресійного апарату. При цьому звичайно використовуються два розподілу: нормальний розподіл та розподіл Стьюденту. Нормальний розподіл задається двома параметрами математичним очікуванням та середньоквадратичним відхиленням $N(\overline{x};\sigma^2)$, звичайно табулюється нормальний розподіл з нульовим математичним очікуванням та одиничною дисперсією N(0,1). Побудуємо 95% довірчі інтервали звичайно для цій змінної. Для цього використаємо таблицю хвостів нормального розподілу. Таблиця хвостів нормального розподілу дозволяє по

заданому значенню змінної знайти ймовірність Pr, того що випадкова зміна перевищує задане значення $\Pr(z > z_\alpha) = \alpha$. Можливе і обернений шлях, коли по значенню ймовірності α (рівень значущості) знаходять значення змінної z_α. Оскільки розподіл симетричний то рівень значущості в цьому випадку дорівнює 0,025. Відповідне табличне значення $z_{0,025} = 1{,}96$. Звідси випадкова змінна, що підпорядкується розподілу N(0,1) у загальному випадку має наступні довірчи інтервали:

$$\Pr(-z_{\alpha/2} \le z \le z_{\alpha/2}) = 1 - \alpha$$

для $\alpha = 5\%,$ (ймовірність 95%), $z_{\alpha/2} = 1{,}96$ і відповідно:

$$\Pr(-1{,}96 \le z \le 1{,}96) = 0{,}95$$

Якщо ми маємо 100 випадкових чисел що підпорядкуються розподілу N(0,1) ми очікуємо, що приблизно 95% з них належить проміжку (-1,96;1,96). Нехай ми маємо n спостережень, $x_1, x_2, \ldots x_n$, що належать розподілу $N(\mu; \sigma^2)$. Задача полягає у знаходженні довірчих інтервалів для математичного очікування по заданої вибірки. Оцінка середнього має розподіл $N(\mu; \sigma^2 / n)$. Тоді $z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$ має розподіл N(0,1). Довірчі інтервали для нормованої змінної z:

$$\Pr(-z_{\alpha/2} \le \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}) = 1 - \alpha$$

Звідси при відомої дисперсії довірчі інтервали для математичного очікування:

$$\Pr(\bar{x} - z_{\alpha/2}\sigma / \sqrt{n} \le \mu \le \bar{x} + z_{\alpha/2}\sigma / \sqrt{n}) = 1 - \alpha$$

У випадку 100 спостережень з N(0,1) маємо наступні 95% довірчі інтервали для математичного очікування:

$$\Pr(-0{,}196 \le \mu \le 0{,}196) = 0{,}95$$

Якщо дисперсія не є відомою то здійснюється перехід від дисперсії до її незміщеної оцінки, а заміст нормального розподілу використовується розподіл Стьюденту. Незміщена оцінка дисперсії розраховується наступним шляхом:

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 /(n-1)$$

Довірчі інтервали для математичного очікування:

$$\Pr(\bar{x} - t_{\alpha/2;n-1} s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2;n-1} s/\sqrt{n}) = 1-\alpha$$

Слід підкреслити, що якщо у попередньому випадку (нормальний розподіл) значення нормованої змінної не залежать від кількості спостережень, то у випадку невідомої дисперсії використовується розподіл Стьюденту що залежить від кількості спостережень (кількість степенем свободи n-1). Для незначної кількості спостережень $z_{\alpha/2} ... ma ... t_{\alpha/2}$ суттєво відрізняються.

Статистичним або емпіричним розподілом вибірки називають таблицю значень ознаки, що розташовані у порядку зростання та відповідних їм частот (відносних частот).

x_i	x_1	…	x_k
Частоти, n_i	n_1	…	n_k

Розрізняють:

- дискретні розподіли;
- інтервальні розподіли або гістограма (послідовність інтервалів та відповідних їм частот).

Для побудови гістограми потрібно визначить кількість інтервалів розподілу та крок розподілу, вважаємо що досліджувана вибірка розташована

у порядку зростання тобто відомі: обсяг вибірки – n, максимальне - $x_{\max}$ та мінімальне значення $x_{\min}$. Тоді крок визначається (формула Стерджеса):

$$h = (x_{\max} - x_{\min}) / (1 + 3{,}322 \cdot \log_{10}(n))$$

Якщо це число не ціле то береться ціле значення та додається одиниця.

Початок першого інтервалу визначається:

$$a_0 = x_{\min} - h / 2$$

Кількість інтервалів m визначається з умови, що $x_{\max}$ попадає в останній інтервал. Якщо n_j кількість влучень в j інтервал (частота) то виконується умова: $\sum_{j=1}^{m} n_j = n.$ Відносна частота визначається: $w_j = n_j / n$ ($\sum_{j=1}^{m} w_j = 1$). Емпірична функція розподілу (кумулята) визначається:

$$F(x) = \sum_{j-1}^{x=aj} w_j$$

Табл.6.1. Етапи побудови гістограми

Номер інтервалу	1	2	……..	m
Межі інтервалу	$(a_0; a_1]$	$(a_1; a_2]$	……..	$(a_{m-1}; a_m]$
Середина інтервалу	$y_1 = \frac{a_0 + a_1}{2}$	$y_2 = \frac{a_1 + a_2}{2}$	……	$y_m = \frac{a_{m-1} + a_m}{2}$
Частота	n_1	n_2	……	n_m
Відносна частота	$w_1 = n_1 / n$	$w_2 = n_2 / n$	……	$w_m = n_m / n$
Кумулята	w_1	$w_1 + w_2$	$w_1 + w_2 +$ $+ \ldots$	$w_1 + w_2 +$ $+ \ldots + w_m$

Статистичні характеристики по згрупованій вибірці можна визначити наступним шляхом:

$$\overline{X}_в = \sum_{j=1}^{m} w_j \cdot y_j$$

$$\sigma_в^2 = \sum_{j=1}^{m} w_j \cdot y_j^2 - \overline{X}_в^2$$

$$\gamma = \sum_{j=1}^{m} w_j \cdot (y_j - \overline{X}_в)^3 / \sigma_в^3$$

$$\mu = \sum w_j \cdot (y_j - X_в)^4 / \sigma_в^4 - 3$$

Приклад (дискретний розподіл).

За тиждень продано 50 пар взуття, що мають наступні розміри (табл.). Побудувати ряд розподілу та емпіричну функцію розподілу.

Табл.6.2. Приклад побудови функції розподілу

Розмір взуття	37	38	40	42	43
Частота	10	12	15	10	3
Відносна частота	0,2	0,24	0,3	0,2	0,06
Кумулята	0,2	0,44	0,74	0,94	1,0

При необмеженому зростанні кількості спостережень та одночасному зростанні кількості інтервалів відносна частота наближується до функції щільності розподілу:

$$f(y_j) \approx \frac{n_j}{h \cdot n}$$

В сучасних програмних засобах використовується зручне та спрощене уявлення розподілу досліджує мої змінної за допомогою графічного уявлення (Box plot). На рисунку вказано масштаб зміної (лева вісь), максимально та мінімальне значення, перший квантиль, медіана та третий квантіль. Розглянемо уявлення розподілу за допомогою Box plot на підставі даних за щомісячнимі доходами 20 домогосподарств Київської області за 2014 рік табл.

Таблиця 6.3. Варіаційний ряд щомісячних доходів 20 домогосподарств Київської області

N	X	N	X	N	X	N	X
1	4	6	7,5	11	13	16	27
2	4	7	9	12	15	17	28
3	5	8	10	13	16	18	31
4	6	9	10,5	14	18	19	31
5	7	10	11	15	24	20	34

Перший квантиль відповідає 5 значенню варіаційного ряду, медіана 10 значенню, третий квантиль 15 значенню. Звідсі Boxplot будується по наступної інформації:

$$x_{\min} = 4; x_{0,25} = 7; x_{0,5} = x_m = 11; x_{0,75} = 24; x_{\max} = 34$$

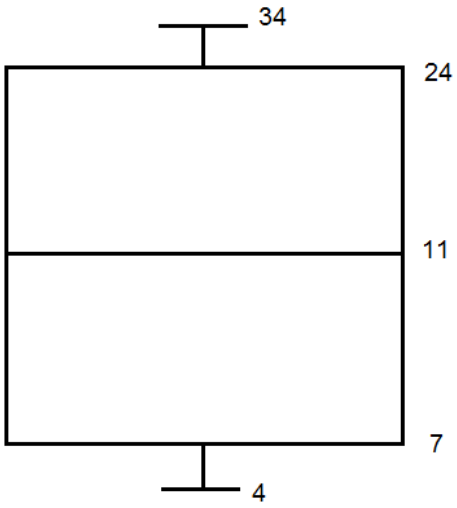


Рис.6.1. Box plot

6.3 Лінійна Залежність між двома змінними

Прикладом найпростішої (однофакторної) моделі є модель з двома змінними, яка описується функцією виду:

$$y_i = f(x) + \varepsilon_i . \qquad (6.1)$$

Це так звана однофакторна модель лінійної регресії, яка є найбільш розповсюдженим видом залежності між економічними змінними. Функція (6.1) може бути або лінійною, або нелінійною. Якщо модель лінійна, то це означає, що саме пряма лінія визначає відображення цієї залежності. Отримати вичерпну інформацію про зв'язок між двома змінними можна за допомогою кореляційного та регресійного аналізу.

Перед тим, як будувати рівняння регресії, з'ясовують для себе, яку змінну потрібно вибрати як аргумент, а яку у вигляді функції, спираючись на економічну теорію. Якщо нас цікавить щільність зв'язку між змінними, тобто як тісно пов'язані між собою два ряди спостережень, що характеризують змінні x і y, розраховують коефіцієнт кореляції. Коефіцієнт кореляції служить мірою лінійного взаємозв'язку між двома величинами, що вимірюються.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (6.2)$$

де $r_{x,y} \epsilon (-1,+1);$

n- кількість спостережень.

Таким чином, коефіцієнт кореляції є безрозмірною величиною і являє собою нормовану коваріацію.

На відміну від коваріації (яка також є мірою взаємозв'язку двох змінних), на величину коефіцієнта кореляції не впливають лінійні перетворення вимірної шкали, тобто якщо вихідні дані помножити чи збільшити на певну величину. При значних додатних значеннях коефіцієнту кореляції взаємозв'язок вважається прямим (одночасно зростають, або зменшуються обидві зміни), при від'ємних оберненим (при зростанні однієї інша зменшується).

Значення коефіцієнта кореляції характеризує ступінь лінійного зв'язку між змінними x і y. Незначна величина коефіцієнту кореляції не означає, що відсутній взаємозв'язку, а тільки підтверджує відсутність лінійного зв'язку. Цілком можливо, що існує нелінійне перетворення однієї з змінних яке перетворює незалежни зміни до залежних (збільшує коефіцієнт кореляції).

Таким чином, коефіцієнт кореляції вимірює також і якість узгодженості статистичних даних з прийнятою гіпотезою про лінійність зв'язку. Слід підкреслити, що оскилькі будь-яки емпіричні дані містять похибку, оцінка коефіцієнту кореляції (3.2) також мітить статистичну похибку, яка оцінюється:

$$\Delta r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} \qquad (6.3)$$

Коефіцієнт кореляції буде значимим, якщо величина t-критерію Стьюдента:

$$\hat{t} = \frac{|r_{xy}|}{\Delta r} = \frac{|r_{xy}| \cdot \sqrt{n-2}}{\sqrt{1-r_{x,y}^2}} \qquad (6.4)$$

буде більшою деякого критичного значення. Ця величина розподілена за законом Стьюдента з ($n-2$) ступенями свободи. При

$$\hat{t} \geq t_{n-2,\alpha} \qquad (6.5)$$

гіпотеза $H_0 : \rho = 0$ (відсутність лінійного взаємозв'язку) відхиляється.

При цьому визначається рівень значимості лінійного взаємозв'язку α (ймовірність, що нульова гіпотеза відхилено помилково), або ймовірність впевненості прийняття нульової гіпотези - $p = 1-\alpha$.

Рівень значимості α приймає значення: 0,2;0,1;0,05;…0,001.

Найбільш часто використаємо значення (банківський сектор, страхування, соціальні дослідження) $\alpha = 0{,}05$. Величини рівня значимості більш ніж $t_{n-2;0,05}$ свідчать про недостатню високу щільність лінійного взаємозв'язку.

Наведемо приклад оцінки рівня лінійного взаємозв'язку.

Приклад

Розглянемо рівень лінійного взаємозв'язку між щомісячною вартістю продуктової корзини та аналогічним показником девальвації гривні. По даним останніх 6 місяців відповідно (3.2) коефіцієнт кореляції дорівнює 0,7. Відповідно (6.2) похибки оцінки дорівнює 0,357, звідси оцінка t критерію відповідно (6.4) дорівнює 1,96. За допомогою таблиці розподілу Стьюдента при 4 ступенях свободи знайдемо $t_{4;0,2} = 1{,}53$. Тобто за умовою (6.5) лінійний взаємозв'язок між вартісю корзини та показником девальвації існує на рівні значимості 20% (можливо існують інші причини зростання ціни продуктової корзини).

Вираз (6.4) дозволяє оцінити мінімальне значення коефіцієнту кореляції (у випадку прямого взаємозвязку), що забезпечує відхилення нульової гіпотези:

$$|r_{xy}| = \frac{t_{n-2;\alpha}}{\sqrt{(t^2_{n-2;\alpha} + n - 2}} \qquad (6.6)$$

Мінімальні значення коефіцієнту кореляції для рівня значимості 0,05 (5%) наведено у табл.6.4.

Табл.6.4. Мінімальне значення коефіцієнту кореляції що забезпечує рівень значимості лінійного взаємозв'язку 5%.

n	5	10	20	30	40	60	80	100
r_{xy}	0,88	0,63	0,44	0,36	0,31	0,25	0,21	0,17

Як слідує з наведених даних, навидь значення коефіцієнту, що перевищує 0,8 не забезпечує достатньо щільного рівня взаємозв'язку при малої кількості спостережень. Графічно дані табл.6.4 представлено на рис. 6.2.

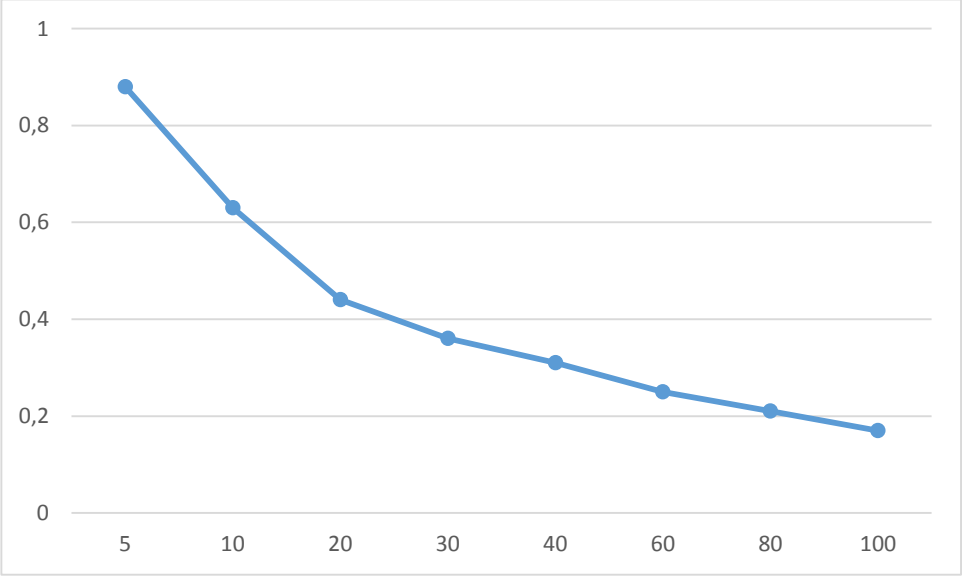


Рис.6.2. Значення коефіцієнта кореляції для відхилення нульової гіпотези на рівні значимості 0,05

6.4 Метод найменших квадратів

В основі класичного регресійного аналізу лежить метод найменших квадратів (МНК).

У методі найменших квадратів застосовується критерій, який приводить до єдиного розв'язку.

Розглянемо однофакторну (одновимірну) лінійну регресію виду:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ (i = 1,2,...,n), \qquad (6.7)$$

де β_0, β_1 – параметри регресії; ε_i – похибка.

В якості підібраної прямої застосуємо рівняння:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i, \qquad (6.8)$$

де $\widehat{\beta_0}; \widehat{\beta_1}$ визначені як оцінки параметрів регресії.

Як міру відхилення від прямої $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ у МНК застосовують суму квадратів відхилень:

$$\varepsilon_i = Y_i - \widehat{Y_i}$$

$$F = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2 \qquad (6.9)$$

Необхідною умовою існування мінімуму цього функціонала, або точкою екстремуму, буде рівність частинних похідних нулю:

$$\frac{\partial F}{\partial \beta_0} = 0$$

$$\frac{\partial F}{\partial \beta_1} = 0 \qquad (6.10)$$

Підставляя в (6.10) (6.9) отримаємо систему нормальних рівнянь:

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases} \qquad (6.11)$$

Звідки

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\cdot\left(\sum_{i=1}^{n} Y_i\right)}{\sum_{i=1}^{n} X_i{}^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\cdot\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\left(\sum_{i=1}^{n} X_i Y_i\right) - n\cdot\bar{X}\cdot\bar{Y}}{\left(\sum_{i=1}^{n} X_i{}^2\right) - n\cdot\bar{X}^2}, \qquad (6.12)$$

а $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\hat{X}$.

Приведемо ще декілька корисних формул:

$$\bar{X} = \sum_{i=1}^{n} X_i / n; \bar{Y} = \sum_{i=1}^{n} Y_i / n; x_i = X_i - \bar{X}; y_i = Y_i - \bar{Y}; \sum_{i=1}^{n} x_i = 0; \sum_{i=1}^{n} y_i = 0$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}; \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} x_i^2 + n\bar{X}^2; \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} y_i^2 + n\bar{Y}^2 \qquad (6.13)$$

Знак який з'явився над регресійними коефіцієнтами означає, що ми отримуємо тільки оцінки регресійних коефіцієнтів (істіни значення нам не відоми. У подальшому для істинних значень нами будуть оцінюватись довірчи інтервали.

Крім оцінок регресійних коефіцієнтів, що зроблено методім найменших квадратів у сучасному програмному забеспеченні існують робастні оцінки, які дають декілька інши оцінки регресійних коефіцієнтів, та оцінки максимальної правдоподібності, які у випадку нормального розподілу співпадають з оцінками, що отримано за допомогою МНК. Використовується наступна умова для отримання робастних оцінок регресійної залежності:

$$F^1 = \sum_{i=1}^{n} |\varepsilon_i|_i = \sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_i| \Rightarrow \min \qquad (6.14)$$

Оцінка, що отримано за допомогою (6.14) є більш стійкими до пливу точок, що значно випадають з головної тенденції, тобто ці точки менш впливають на оцінки регресійних коефіцієнтів, ніж у випадку МНК.

В якості прикладу ефективності МНК оцінювання розглянемо оцінку математичного очікування (середньо значення) та доведемо, що це є оцінка МНК. Цільова функція в цьому випадку:

$$F(\beta_0) = \sum_{i=1}^{n} (x_i - \beta_0)^2$$

Умова мінімуму цільової функції:

$$\frac{dF}{d\beta_0} = -2\sum(x_i - \beta_0) = 0 \Rightarrow \beta_0 = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6.15)$$

Звідси слідує, що середньо значення є також оцінкою МНК.

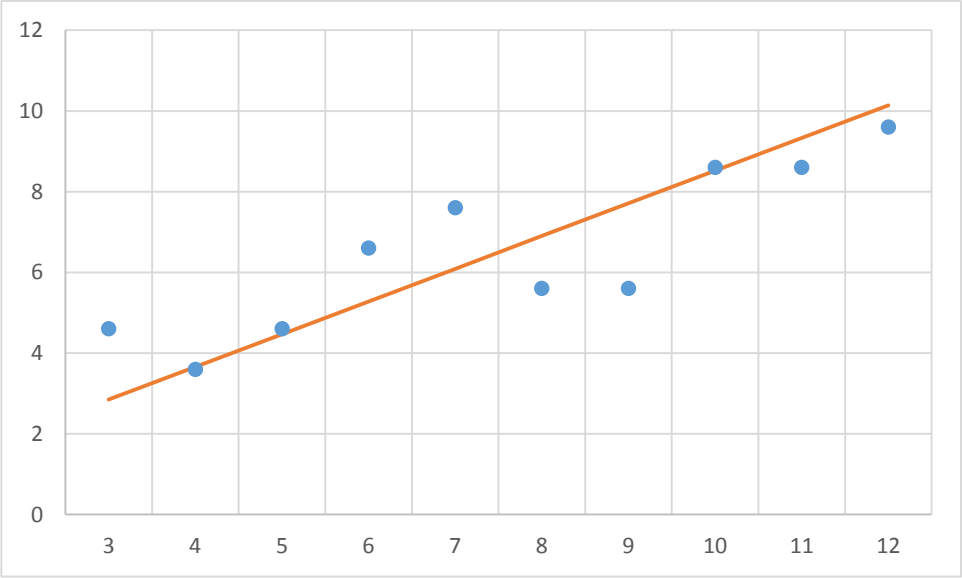


Рис. 6.3. Діаграма розсіювання та графік лінійної залежності рівня споживання від рівня доходів

6.2 Стандартні помилки та довірчі інтервали оцінок параметрів регресії

Стандартна похибка та довірчий інтервал кутового коефіцієнта β_1.

Введемо поняття стандартної похибки одно факторної моделі:

$$s = \sqrt{\frac{\sum_{1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{1}^{n}\varepsilon_i^2}{n-2}} \qquad (6.16)$$

Стандартна похибка використовується для вирішення багатьох питань: розрахунку довірчих інтервалів для істинних значень регресійних коефіцієнтів, оцінки довірчих інтервалів для прогнозних значень на рівні математичного очікування.

Довірчим інтервалом випадкової величини β_1 називають інтервал з межами $(\beta_1 - \Delta\beta_1)$ і $(\beta_1 + \Delta\beta_1)$, в якому з певною, наперед заданою, ймовірністю

$P=1-\alpha$ знаходиться істинне значення параметра β_1 . При цьому ймовірність $P=1-\alpha$ називається довірчою ймовірністю, а α – рівнем значімості.

Тоді стандартна похибка оцінки регресійного коефіцієнту визначається:

$$\sigma_{\beta 1}=\frac{s}{(\sum(X_i-\overline{X})^2)^{1/2}} \qquad (6.17)$$

Для оцінки довірчого інтервалу розраховується похибка, яка визначається:

$$\Delta\beta_1=t_{n-2;\alpha/2}\cdot\sigma_{\beta 1} \qquad (6.18)$$

Звідси довірчій інтервал на рівні значущості α (з ймовірністю $p=1-\alpha$) для регресійного коефіцієнту β_1:

$$\widehat{\beta}_1+t_{n-2;\alpha/2}\sigma_{\beta 1}\le\beta_1\le\widehat{\beta}_1+t_{n-2;\alpha/2}\sigma_{\beta 1} \qquad (6.19)$$

де $t_{n-2;\alpha/2}$ – квантиль розподілу Стьюденту.

Важливим є також тест перевірки нульової гіпотези відносно величини регресійного коефіцієнту β_1. Нульова гіпотеза-H_0 означає що істинне значення регресійного коефіцієнту ні відрізняється від нуля. Однак за допомогою отримання оцінок як регресійного коефіцієнту так і його похибки ми можемо підтвердити або відхилити цю гіпотезу. Спочатку розраховується величина t критерію:

$$t_1=\frac{\widehat{\beta}_1}{\sigma_{\beta 1}} \qquad (6.20)$$

Якщо модуль цього значення перевищує відповідний квантіль розподілу Стьюденту $t_{n-2;\alpha}$ то нульову гіпотезу потрібно відхилити на рівні значущості α .

Стандартна помилка для регресійного коефіцієнту β_0 визначається:

$$\sigma_{\beta 0} = s\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}} \qquad (6.21)$$

Довірчій інтервал на рівні значущості α (з ймовірністю $p = 1-\alpha$) для регресійного коефіцієнту β_0:

$$\hat{\beta}_0 + t_{n-2;\alpha/2}\sigma_{\beta 0} \le \beta_0 \le \hat{\beta}_0 + t_{n-2;\alpha/2}\sigma_{\beta 0} \qquad (6.22)$$

Величина t критерію відносно коефіцієнту β_0 розраховується аналогічно (3.22), що дозволяє підтримати або відхилити гіпотезу H_0:

$$t_0 = \frac{\hat{\beta}_0}{\sigma_{\beta 0}} \qquad (6.23)$$

Приклад 3

На підставі даних табл.6.3 зробити наступні оцінки:

- оцінку стандартної похибки;
- знайти помилки оцінок регресійних коефіцієнтів;
- прийняти або відхилити нульові гіпотези відносно їх значень (t критерій);
- знайти 90% довірчі інтервали для істинних оцінок регресійних коефіцієнтів.

Рішення

Відповідно (3.21) та даним табл.6.2 оцінка стандартної похибки при n=10:

$$s = \sqrt{\frac{\sum_{1}^{n} \varepsilon_i^2}{n-2}} = \sqrt{\frac{2{,}5}{8}} \approx 0{,}56; s^2 \approx 0{,}31$$

2)Знайдемо помилку $\sigma_{\beta 1} = \frac{s}{(\sum (X_i - \overline{X})^2)^{1/2}} = \frac{0,56}{\sqrt{52,5}} \approx 0,14$

Знайдемо помилку $\sigma_{\beta 0} = s\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}} = 0,56\sqrt{\frac{52,5+10\cdot 7,5^2}{10\cdot 52,5}} \approx 0,25$

Зробимо оцінки t критерію відповідно (3.22) та (3.25):

$t_0 = \frac{0,42}{0,25} \approx 1,68; t_1 = \frac{0,81}{0,14} \approx 5,79$

Відповідні критичні значення розподілу Стьюденту: $t_{8;0,2} = 1,4; t_{8;0,001} = 5,04 \Rightarrow t_0 > t_{8;0,2}; t_1 > t_{8;0,001}$.

Звідси рівень значимості на якому відхіляється нульова гіпотеза для автономного споживання 20%, а рівень значимості для відхилення нульової гіпотези для автономного споживання 0,1%.

Оскільки потрібно розрахувати 90% довірчі інтервали з розподілу Стьюденту знайдемо критичне значення $t_{8;0,05} = 2,31$. Звідси довірчі інтервали для автономного та маргінального споживання:

$0,42 - 2,31\cdot 0,25 \le C_a \le 0,42 + 2,31\cdot 0,25 \Rightarrow -0,16 \le C_a \le 1,0$

$0,81 - 2,31\cdot 0,14 \le C_m \le 0,81 + 2,31\cdot 0,14 \Rightarrow 0,49 \le C_m \le 1,13$

На перший погляд довір чи інтервали за надто великі, щоб робити будь які висновки. Одна з можливих причин мала кількість спостережень, що збільшує стандартну похибку та критичне значення розподілу Стьюденту. Для порівняння $t_{8;0,05} = 2,31$ але $t_{60;0,05} = 2,0$, що суттєво зменшує довірчи інтервали.

Довірчий інтервал для дисперсії генеральної похибки моделі σ^2

В деяких випадках потрібно мати уявлення відносно довірчих інтервалів генеральної похибки моделі оцінкою якої є стандартна похибка S. При побудові довірчого інтервалу для параметра σ^2 виходять з того, що

відповідна статистика $n \cdot s^2 / \sigma^2$ має χ^2 – розподіл з $\nu = n - 2$ ступенями свободи. При цьому довірчий інтервал вибирають таким чином, щоб

$$P\left(\chi^2 < \chi_{\left(1-\frac{\alpha}{2};n-2\right)}\right) = P\left(\chi^2 > \chi_{\left(\frac{\alpha}{2};n-2\right)}\right) = \alpha \qquad (6.24)$$

Тоді довірчий інтервал $(p = 1-\alpha)$ для генеральної дисперсії σ^2 для рівня значимості α визначається з формули:

$$\frac{n \cdot s^2}{\chi^2_{\alpha/2;n-2}} \le \sigma^2 \le \frac{n \cdot s^2}{\chi^2_{1-\alpha/2;n-2}} \qquad (6.25)$$

Приклад 4

Знайти 90% довірчи інтервали для дисперсії генеральної похибки моделі залежності споживання від доходу (табл.6.3).

Рішення

За умовою відомої ймовірності знайдемо рівень значимості:

$p = 0{,}9 \Rightarrow \alpha = 1 - 0{,}9 = 0{,}1 \Rightarrow \alpha/2 = 0{,}05$

З таблиці розподілу знайдемо критичні значення розподілу та інші параметри нерівності (6.26):

$\chi^2_{0{,}05;8} = 15{,}5; \chi^2_{0{,}95;8} = 2{,}73; n = 10; s^2 = 0{,}31$

Отримаємо наступну нерівність для дисперсії похибки моделі:

$$\frac{10 \cdot 0{,}31}{15{,}5} \le \sigma^2 \le \frac{10 \cdot 0{,}31}{2{,}73} \Rightarrow 0{,}2 \le \sigma^2 \le 1{,}1 \; ; \; p(0{,}2 \le \sigma^2 \le 1{,}1) = 0{,}9$$

Використання однофакторної моделі з метою аналіза та прогнозу

Одними з основних цілей економічних досліджень є аналіз економічного явища чи процесу на основі вибіркових спостережень.

Розглянемо можливості аналізу детальніше. Нехай існує одно факторне регресійне рівняння, в якому У ендогенна змінна, Х екзогенна детермінована змінна:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (6.26)$$

Здійснімо операцію математичного очікування від об двох частин (6.26), врахуємо що очікувана величина похибки дорівнює нулю, а невідомі значення регресійних коефіцієнтів змінюємо на їх оцінки:

$$E(Y(X_i)) = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (6.27)$$

Введемо вимірність змінних: [Y]; [X]. Якщо подставите до (6.27) значенья X=0 , то отримаємо:

$$E(Y(0)) = \hat{\beta}_0 \qquad (6.28)$$

Звідси слідує, що вимірність регресійного коефіцієнту β_0 дорівнює вимірності Y. Існують випадки, коли цій регресійний коефіцієнт не має ніякого економічного змісту. Наприклад якщо розглядати стандартну задачу ріелтора: залежність вартості житла від площі, то коефіцієнт $\hat{\beta}_0$ може приймати як від'ємне так і додатне значення та забезпечує найкраще наближення модельних даних (що розраховуються відповідно (6.26)) до даних спостережень, тому що вартість житла при нульовому значенні площі не має змісту. Існують і протилежні випадки. Наприклад, якшо досліджувати середний тариф для місцевого такси (вартість проїзду 1 км), то β_0 уявляє середнюю плату за посадку.

Розглянемо економічний зміст коефіцієнту β_1 . Якщо знайти маргінальну величину залежної (її похідну) змінної то отримаємо:

$$\frac{d(E(Y)}{dX} = \beta_1 \Rightarrow [\beta_1] = \frac{[Y]}{[X]}$$

Це означає, що коефіцієнт при незалежної змінної дорівнює маргінальної величині залежної змінної з вимірністю що дорівнює відношенню вимірності залежної змінної до вимірності незалежної змінної. Розгляне це на прикладі погодинної оплати праці. Нехай залежна зміна (Y) це

оплата праці, а незалежна (Х) кількість відпрацьованих годин, тоді коефіцієнт при Х має вимірність грн/год і уявляю маржинальну оплату праці (плату за додатково відпрацьовану годину). У прикладі про маржинальну схильність до споживання (табл.6.3) вільний член вимірюється у тис. грн.(автономна схильність до споживання) а коефіцієнт при прибутку не має вимірності (грн./грн.), тобто уявляє собою частку від доходу (81%) що втрачається на споживання.

Прогноз по одно факторної моделі

Виконання прогнозу надзвичайно відповідальна та складна задача. Справа в тому що на довготривалому макроекономічному прогнозі в деяких випадках базується стратегія розвитку країни, тому будь які похибки та некоректна трактовка прогнозу можуть привести до непередбачених, та катастрофічних наслідив. Наприклад стратегія розвитку Російської Федерації базувалась на прогнозі, що зроблено у 2008 році, що попит на природний газ як внутрішній, так і з боку сусідніх країн буде ста більно зростати і к 2014 року досягне 600 млрд. куб. метрів. Однак сланцева революція у США, енергозберігаючі технології та розвиток оновлює мої енергетики в ЄС, економічний спад в Російської Федерації призвели до фактичного споживання у 2014 році у обсязі 400 млрд. куб. метрів. А це означає що третя частина від виробничого потенціалу виявилася незатребуваний, рівень надходжень до бюджету від експорту енергоносій суттєво зменшився (на 40 млрд. USD на рік) відносно запланованого рівня. Приблизно така ж сума було втрачено на планування та розробку нових шляхів газопостачання в ЄС, яки з великою ймовірністю вже не реалізуються. Аналогічні помилки характерну і для нашої країни-найбільш наявний приклад це побудова за кошти ЄС (500 млн. євро) у 90 роках минулого століття непрацюючого нафта гону Одеса – Броді. На тій проміжок часу цих коштів було б достатньо для побудови транс українського шосе світового рівня. Наведени приклади свідчать, що виконання прогнозів надзвичайно складна та відповідальна задача, де неможна гарантувати

відсутність похибки. Тому похибка прогнозу невід'ємна частина його здійснення.

Введемо наступні визначення, що необхідні для прогнозування:

Базисний інтервал - інтервал незалежної змінної X на якому розтушевано спостереження. Наприклад для задачі на оцінку маржинальної схильності до споживання (табл.6.3) базисний інтервал це є інтервал доходів від 5 до 12 тис. грн.

Горизонт прогнозування це інтервал між останньою точкою базисного інтервалу та останьою точкою прогнозного інтервалу.

Прогноз може здійснюватись як для значень що знаходяться в межах базисного інтервалу так і для точок що знаходяться по за його межами.

При будь-якому прогнозуванні потрібно враховувати наступне:

- при використанні будь якої моделі неможна врахувати всі фактори, що можливо вплинуть на кінцевий результат;
- для будь якого прогнозу характерна похибка, яка збільшується з збільшенням горизонту прогнозування.

Однак у більшості випадків прогнози, що виконуються провідними світовими науковими інституціями для широкого суспільного кола для спрощення друкуються без довірчих інтервалів. Це не означає що ці інституції мають алгоритми, що дозволяють запобігти похибок прогнозу. Як правіло в цьому випадку в якості прогнозного показнику виступає математичне очікування. В умовах надзвичайно нестабільної української економіки важко робити будь які прогнози з горизонтом прогнозування що перевищую один рік. На користь цього припущення свідчить постійна коректування прогнозів що здійснюється на часовому проміжку горизонту прогнозування. У подальшому ми будимо розглядати тільки прогоноз з довірчимі інтервалами. Ймовірність попадання фактичного значення до довірчого інтервалу (реалізація прогнозу задається у процесі дослідження). Наічастиші використовуються 95% та 90% довірчи інтервали.

Перейдемо до оцінок довірчих інтервалів. Нехай у деякої точці X^* нам потрібно знайти прогнозне значення та 90% довірчи інтервали. Для оцінки прогнозного значення використовуємо рівняння (6.27):

$$\hat{Y}(X^*) = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad (6.29)$$

Оскільки оцінки регресійних коефіцієнтів мають статистичні похибки (то прогнозне значення на рівні математичного очікування має також похибку:

$$Var(\hat{Y}(X^*)) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \right] \qquad (6.30)$$

Невідоме значення дисперсії змінюється на його оцінку квадрат стандартної похибки (6.16). Звідси похибка для прогнозного значення в точці X^* розраховується:

$$\sigma(X^*) = s \cdot \left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \right]^{1/2} \qquad (6.31)$$

На підставі останього виразу можна оцінити параметри, вибірки що визначають похибку прогнозного значення, це у першу чергу стандартна похибка - s, та обсяг вибірки - n. Крім того похибка залежить від відстані точки, для якої розраховується прогнозне значення від середнього значення пояснюючої змінної. Найменьше значення похибки спостерегається при середньому значенні незалежної змінної:

$X^* = \overline{X}$

Наступний крок розрахунки довірчих інтервалів для прогнозних значень:

$$\Delta_\alpha(X^*) = t_{n-2;\alpha/2}\sigma_Y(X^*) \Rightarrow \hat{Y}(X^*) - \Delta_\alpha(X^*) \le Y(X^*) \le \hat{Y}(X^*) + \Delta_\alpha(X^*) \quad (6.32)$$

Приклад 5

В якості прикладу розрахунку оцінімо довірчі інтервали для задачі залежності споживання від доходу домогосподарства (табл.6.1).

Рішення

Оцінімо 90 % довірчі інтервали для споживання. В цьому випадку:

$$p = 0{,}9 \Rightarrow \alpha/2 = (1-p)/2 = 0{,}05 \Rightarrow t_{8;0{,}05} = 2{,}31$$

На першому етапі ми розраховуємо модельні (прогнозні) значення споживання:

$$\hat{Y}(X^*) = 0{,}42 + 0{,}81X^*$$

Для побудови довірчих інтервалів краще подати вихідну інформацію у вигляді варіаційного ряду по X з сталим кроком. Нехай крок по доходу дорівнює однієї тисячі грн.

Крім існуючих даних 10 спостережень довірчи інтервали розраховуються для трех спостережень по за межами базисного інтервалу: 14; 15; 16 тис. грн. (табл.6.5).

Табл.6.5. Послідовність розрахунку довірчих інтервалів

N	X^*	$\hat{Y}(X*)$	$\sigma(X^*)$	$\Delta_{0,9}(X^*)$	$\hat{Y}(X^*) - \Delta_{0,9}(X^*)$	$\hat{Y}(X^*) + \Delta_{0,9}(X^*)$
1	4	3,7	0,32	0,6	3,0	4,4
2	5	4,5	0,26	0,49	3,9	5,1
2	6	5,3	0,21	0,42	4,8	5,8
4	7	6,1	0,18	0,42	5,7	6,5
5	8	6,9	0,21	0,49	6,5	7,3
6	9	7,7	0,18	0,6	7,2	8,2
7	10	8,5	0,26	0,74	7,9	9,1
8	11	9,3	0,32	0,49	8,6	10,0
9	12	10,1	0,39	0,9	9,2	11,0
10	13	11,0	0,46	1,1	9,9	12,1

11	14	11,8	0,53	1,2	10,6	13,0
12	15	12,6	0,61	1,41	11,2	14,0
13	16	13,4	0,68	1,57	11,8	15,0

В сучасних економетричних програмах довірчи інтервали входять в програмний сервіс при графичному уявленні, якщо будувати їх самостійно за допомогою Екселя то краще обірати сталий крок відносно незалежної змінної X. Наприклад у випадку що розглядається базисний інтервал задається відрізком[4;12]. Дискретність достатня для побудови графіка дорівнює 1 тис. грн.

Горізонт прогнозування дорівнює 4. Тобто для якісної побудови довірчих інтервалів достатньо в цьому випадку розрахувати довірчи інтервали для (16-4)+1=13 точок. Графичне уявлення довірчих інтервалів для прогнозних значень подано на рис.6.4.

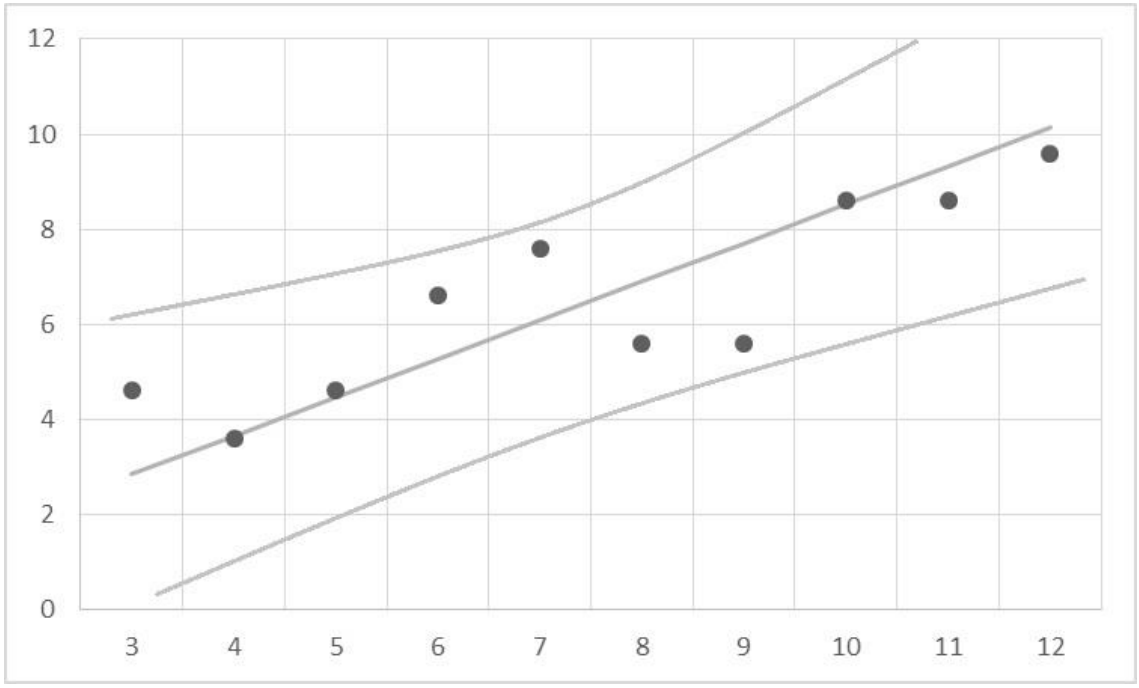


Рис.6.4. Побудова довірчих інтервалів для задачі про маргінальну схильність до споживання.

6.3. Однофакторний дисперсійний аналіз ANOVA

Неважко довести на підставі властивостей оцінок регресії, що загальна сума квадратів відхилень (TSS) залежної змінної Y від її вибіркового середнього значення складається із суми квадратів відхилень, що обумовлена регресією (RSS), та суми квадратів залишків (ESS):

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i+\widehat{Y}_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2$$

Або $$\boldsymbol{TSS=RSS+ESS} \qquad (6.33)$$

Або поділив всі члени рівняння на n та зробив перехід до дисперсій:

$$\sigma^2(Y)=\sigma^2(\widehat{Y})+\sigma^2(\varepsilon) \qquad (6.34)$$

Дисперсия залежной змінної дорівнює суме поясненої дисперсії та дисперсії похибки.

Приклад 6

По даним залежності споживання від доходу, оцінити частку поясненої та непоясненої дисперсії у загальної дисперсії процесу.

Рішення

По даним таблиць (6.2) та (6.3) знайдемо складові виразу (6.34):

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2=458{,}9-10\cdot 6{,}5^2=36{,}4 \qquad (6.35)$$

$$\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2=\sum_{i=1}^{n}\widehat{Y}_i^2-n\overline{\widehat{Y}}^2=456{,}4-10\cdot 6{,}5^2=33{,}9$$

$$\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}\varepsilon_i^2=2{,}5 \qquad (6.36)$$

При розрахунках враховувалось що: $\overline{Y}=\overline{\widehat{Y}}$.

Звідсі загальна дисперсія дорівнює: 3,64; поясненная дисперсія 3,39; непоясненная дисперсія дорівнює 0,25. Частка поясненої дисперсії в загальної

складає 93,1%, непоясненої - 6,9%. Тобто модель достатньо адекватно описує процес, що досліджується.

В табл. 6.5 та 6.6 представлено таблиці дисперсного аналізу.

Таблиця 6.5. Таблиця дисперсійного аналізу

Джерело розсіювання	Сума квадратів відхилень	Число ступенів свободи	Середня сума квадратів відхилень
Регресія	$RSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{RSS}{1}$
Помилка	$ESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	n-2	$MSE = \frac{ESS}{n-2} = s^2$
Загальна варіація	$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2$	n-1	–

У пакетах прикладного програмного забезпечення зокрема, ППП «Statgraphics», ППП «Statistica» та інших, таблиця ANOVA має такий вигляд:

Табл. 6.6. Analysis of Variance (ANOVA)

Source	Sum of Squares (SS)	Df	Mean Square (MS)
Model або Regression	RSS	1	MSR
Error або Residual	ESS	n-2	MSE
Total (corr.)	TSS	n	–

MSE – середній квадрат відносно регресії (s^2) – дає оцінку залишкової дисперсії відносно регресії, яка базується для однофакторної моделі на n-2 ступенях свободи, тобто s^2 є оцінкою σ_ε^2. У моделі з кращим підбиранням s^2 має менші значення. Отже, s^2 може виступати мірою адекватності підібраної моделі.

MSR – середня сума квадратів відхилення, обумовлена регресією.

6.4. Критерії адекватності однофакторної економетричної моделі

Для першого наближення підібрана модель вважається адекватною, якщо RSS>ESS. Як критерії адекватності регресійної моделі застосовують статистику R^2 – коефіцієнт детермінації та F-критерій Фішера.

Коефіцієнт детермінації

Коефіцієнт детермінації є відношення поясненої дисперсії до загальної дисперсії ендогенної(залежної) зміної та розраховується за формулою:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} = \frac{b^2{}_1 \sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}, \quad (6.37)$$

де R – множинний коефіцієнт кореляції.

У лінійній моделі множинний коефіцієнт кореляції R виступає як міра ступеня лінійності зв'язку між Y та X, оскільки:

$$ESS = \sum_{i=1}^{n}(Y_i - \widehat{Y})^2 = (1 - R^2)\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2 \quad (6.38)$$

Звідки:

$$ESS = (1 - R^2) \cdot TSS \quad (6.39)$$

Або

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \quad (6.40)$$

Визначення. Коефіцієнтом детермінації називають вираз:

$$R^2 = \frac{RSS}{TSS} \quad (6.41)$$

який показує частку поясненої дисперсії в загальній дисперсії

$R^2 \in (0,\ 1)$, або $R^2 \in (0,\ 100\%)$.

Із визначення безпосередньо випливає нерівність $\mathbf{0 \le R^2 \le 1}$.

Чим ближче $\boldsymbol{R^2}$ до 1, тим підібрана модель вважається кращою. Якщо $\boldsymbol{R^2} = 0$, то це означає, що залежність між Y та X відсутня, тобто X не впливає на Y. Однак на практиці використовується критичне значення коефіцієнту кореляції (квадрат коефіцієнту кореляції дорівнює коефіцієнту детермінації) - вираз (3.6), яке дозволяє прийняти або відхилити нульову гіпотезу відносно існування лінійного взаємозв'язку між двома змінними:

$$\left|r_{xy}\right| = \frac{t_{n-2;\alpha}}{\sqrt{(t^2_{n-2;\alpha} + n - 2}}$$

В іншому випадку, коли R^2=1, в системі існує функціональний зв'язок між Y та X, тобто всі точки будуть лежати на підібраній прямій, що є можливим лише, коли виконується умова ESS=0.

Приклад 7

Зробити оцінку коефіцієнту детермінації для задачи залежности споживання від доходу

Рішення

Коефіцієнт детермінації визначається як відношення поясненої дисперсії до загальної. По даним прикладу 6 це відношення дорівнює 93,1%.

Критерій Фішера

Для прийняття рішення про адекватність підібраної лінії регресії, поряд з коефіцієнтом детермінації, розраховують F-статистику, яка використовуються для перевірки якості оцінюваної регресії в цілому.

$$F = \frac{RSS}{ESS/(n-2)} \qquad (6.42)$$

Для багатофакторної моделі:

$$F = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 \Big/ m}{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \Big/ (n-m-1)}, \qquad (6.43)$$

де m – кількість незалежних змінних моделі.

Для однофакторної моделі

$$F = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \Big/ (n-2)} \qquad (6.44)$$

або

$$F = \frac{MSR}{MSE}. \qquad (6.45)$$

Приклад 8

Розрахувати критерій Фішера для задачі про залежність споживання від доходу.

Рішення

Підставляє до виразу (6.45) дані задачі (6), що представлено виразом (6.36) отримаємо:

$$F = \frac{33,9}{2,5/8} \approx 109,4$$

Це відношення відповідає F-розподілу з $\nu_1 = 1$ і $\nu_2 = 10 - 2 = 8$ ступенів свободи(табл.1.4). За умовою 5% рівня значимості:

$$F_{0,05}(1;8) = 5,3$$

Якщо виконується умова, при якій $F > F_\alpha(\nu_1;\nu_2)$, регресія вважається значимою.

6.4. Використання одно факторних нелінійних моделей

У попередніх розділах нами розглядалися випадок одно факторної лінійної залежності. Однак в реальної економіки спостерігаються не тільки лінійні залежності. Наприклад, якщо розглядати залежність урожайності від кількості внесених добрів, то лінійною буде тільки початкова частка залежності (незначна кількість добрів на одиницю площі), а маргінальна корисність буде спадною функцією від кількості внесених добрів.

Побудова нелінійних моделей виконується у декілька етапів на першому етапі виконується лінеаризація, на другому за допомогою стандартної процедури МНК здійснюються оцінки лінеарізованої регресії, на тре тему етапі відбувається повернення до початкового представлення.

Найбільш широкій спектр одно факторних залежностей представлено в програмі STATGRAPHICS (рис.6.6). Вони мають наступний аналітичний вигляд:

Square Y model: Y = (a + b*X)^2

Exponential model: Y = exp(a + b*X)

Reciprocal-Y model: Y = 1/(a + b*X)

Squared root-Y model: Y = sqrt(a + b*X)

Square root-X model: Y = a + b*sqrt(X)

Double square root model: Y = (a + b*sqrt(X))^2

Logarithmic-Y square root-X model: Y = exp(a + b*sqrt(X))

Reciprocal-Y square root-X: Y = 1/(a + b*sqrt(X))

Squared-Y square root-X: Y = sqrt(a + b*sqrt(X))

Logarithmic-X model: Y = a + b*ln(X) (6.46)

Square root-Y logarithmic-X model: Y = (a + b*ln(X))^2

Multiplicative model: Y = a*X^b

Reciprocal-Y logarithmic-X model: Y = 1/(a + b*ln(X))

Squared-Y logarithmic-X model: Y = sqrt(a + b*ln(X))

Reciprocal-X model: Y = a + b/X

Square root-Y reciprocal-X model: Y = (a + b/X)^2

S-curve model: Y = exp(a + b/X)

Squared-Y reciprocal-X model: Y = sqrt(a + b/X)

Squared-X model: Y = a + b*X^2

Square root-Y squared-X model: Y = (a + b*X^2)^2

Reciprocal-Y squared-X: Y = 1/(a + b*X^2)

Double-squared: Y = sqrt(a + b*X^2)

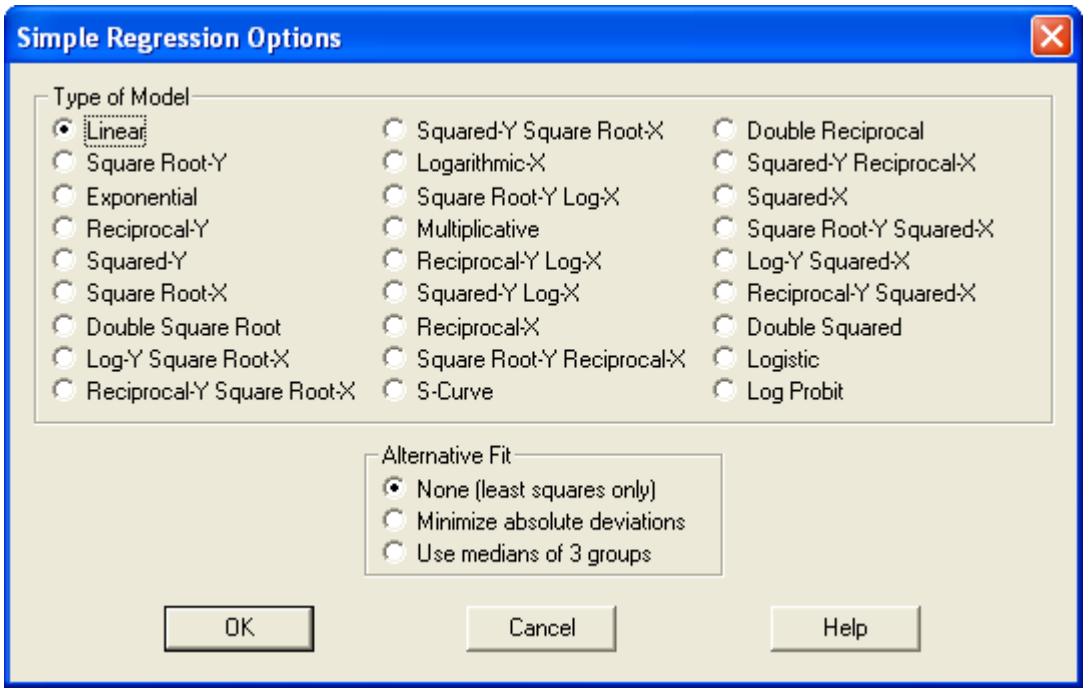


Рис. 6.7. Панель однофакторних залежностей програми STATGRAPHICS

Звичайно до цих монотонних функцій додається і лінійна залежність (рис.6.7). Якщо вважати всі фактори додатними (x>0) то всі запропоновані залежності мають монотонний характер (зростаючі або спадні). Взагалі же першим кроком побудови моделі, як це згадувалося раніше, є теоретичний аналіз досліджуваного явища. В економіці та в природокористуванні зустрічаються випадки немонотонних залежностей. Так, наприклад, в оподаткуванні широко розповсюджена немонотонна залежність обсягу податкових надходжень від ставки оподаткування (крива Лафера). На першому етапі зростання податкової ставки супроводжується зростанням

надходжень, однак подальше зростання ставки має негативний вплив на економіку, база оподаткування скорочується, що при значних податкових ставках призводить до зменшення надходжень. Аналогичну залежніть можна використати при дослідженні впливу внесення добрів на зростання крожайності. При цьому на початковому етапі ми повінни висунути гіпотезу відносно механізму впливу факторной зміної на результуючу. Звичайно таки залежності досліджуються за допомогою квадратичного поліному:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2; \beta_1 > 0; \beta_2 < 0.. \qquad (6.47)$$

Однак використання поліномів більш високої степені з метою прогнозу не дає позитивного результату, тому що в цьому випадку за рахунок варіативності поліному значної ступені відбувається поєднання детермінованої (трендової) складової та випадкової складової. Тому використання полиномові степені більш двох призводить до помилкових висновків, у першу чергу тому, що не відбувається розділення на випадкову та трендову складову. В значної мірі економетричний аналіз використовується звичайно для виявлення тенденцій економічних процесів (детермінована складова) та відхилення від цих тенденцій (випадкова складова). Підрозділ економетрики, що займається прогнозом та аналізом економічних процесів має назву аналізу часових рядів. Розглянемо детальніше побудову двох трендових залежностей: лінійного тренду (стабільне абсолютне економічне зростання) та експоненційного тренду (стабільне відносне економічне зростання). В цьому випадку незалежна змінна – час t в роках (існують випадки і іншої дискретності), а залежна зміна x(t). Зручне уявлення лінійного та експоненціального тренду:

$$x_l(t) = \beta_0 + \beta_1(t - t_0) + \varepsilon(t) \qquad (6.48)$$

$$x_e(t) = Ae^{\alpha(t-t_0)} \qquad (6.49)$$

t_0 – початок базисного інтервалу, β_0 – очікуване значення змінної що досліджується на початок базисного інтервалу ($\beta_0 = x_l(0)$),

β_1 – очікуваний щорічний приріст.

А-очікуване початкове значення ($A = x_e(0), \alpha -$ річні темпи зростання.

Оцінка коефіцієнтів лінійного тренду (6.48) здійснюється звичайним МНК . Для оцінки експоненціального тренду потрібно провести попередню лінеаризацію що можливо здійснити за допомогою логарифмування обох частин рівняння (6.49):

$$\ln x = \ln A + \alpha(t - t_0) + \varepsilon(t) \qquad (6.50)$$

Позначимо:

$$\ln x = y; \ln A = \beta_0; \alpha = \beta_1; t - t_0 = x. \qquad (6.51)$$

Тоді отримаємо рівняння аналогічне рівнянню лінійної регресії:

$$y = \beta_0 + \beta_1 x + \varepsilon(t) \qquad (6.52)$$

Після отримання оцінок $\hat{\beta}_0, \hat{\beta}_1$ відповідно (3.21) потрібно перейти до початкової моделі:

$$x_e(t) = e^{\beta 0} e^{\beta 1(t-t0)} \qquad (6.53)$$

Розглянемо побудову лінійного та експоненціального тренду на прикладі розвитку КНР з 2000 по 2014 роки (табл.6.8).

Табл.6.6. Валовий внутрішній продукт КНР на часовому інтервалі 2000-2014 роки (тріліони USD)

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
1,2	1,3	1,5	1,6	1,9	2,3	2,7	3,5	4,6	5,1	6	7,5	8,5	9,5	10,4

Послідовність розрахунків представлена в табл.6.9 та відповідає алгоритму, що представлено на табл.6.9. При розрахунках експоненціального тренду потрібно виконати логарифмування ВВП, у представленому прикладі використовуються натуральні логарифми.

Послідовність розрахунків розрахунки лінійного та експоненціального трендів подано у табл. 6.7.

Табл.6.7. Таблиці розрахунків параметрів лінійного та експоненціального трендів для ВВП КНР.

Роки, T	ВВП (X), трн.USD	t=T-2000	t-E(t) (I)	X(t)-E(x)(J)	I*J	$(t-E(t))^2$	lnX	lnX-E(lnX)(K)	K*I
2000	1,2	0	-7,00	-3,31	23,15	49,00	0,18	-1,07	7,47
2001	1,3	1	-6,00	-3,21	19,24	36,00	0,26	-0,99	5,92
2002	1,5	2	-5,00	-3,01	15,03	25,00	0,41	-0,84	4,22
2003	1,6	3	-4,00	-2,91	11,63	16,00	0,47	-0,78	3,12
2004	1,9	4	-3,00	-2,61	7,82	9,00	0,64	-0,61	1,82
2005	2,3	5	-2,00	-2,21	4,41	4,00	0,83	-0,42	0,83
2006	2,7	6	-1,00	-1,81	1,81	1,00	0,99	-0,26	0,26
2007	3,5	7	0,00	-1,01	0,00	0,00	1,25	0,00	0,00
2008	4,6	8	1,00	0,09	0,09	1,00	1,53	0,28	0,28
2009	5,1	9	2,00	0,59	1,19	4,00	1,63	0,38	0,76
2010	6	10	3,00	1,49	4,48	9,00	1,79	0,54	1,63
2011	7,5	11	4,00	2,99	11,97	16,00	2,01	0,77	3,06
2012	8,5	12	5,00	3,99	19,97	25,00	2,14	0,89	4,45
2013	9,5	13	6,00	4,99	29,96	36,00	2,25	1,00	6,01
2014	10,4	14	7,00	5,89	41,25	49,00	2,34	1,09	7,65
	4,51	7			192,00	280,00	1,25		47,48

В результаті розрахунків ми отримали наступні трендові залежності та параметри адекватності (табл.6.10).

Табл.6.8. Параметри адекватності трендових залежностей

	Модель	$R^2(\%)$	F	Стандартна похибка
Лінійна	$\widehat{X}_L(T) = -0{,}33 + 0{,}69(T-2000)$	92,8	85,3	0,88
Експоненціальна	$\widehat{X}_E(T) = 1{,}06e^{0{,}17(T-2000)}$	98,5	454,4	0,4

Всі параметри адекватності що наведено у табл. 6.10 кращі для експоненціального тренда. Крім того з розгляду похибок моделі лінійного тренду слідує, що між німі існує взаємозв'язок. З початку базисного інтервалу лінійна модель занижує ВВП КНР, в середині завишає, у кінці снову занижує

(рис.1.5). Особливо слід підкреслити суттєву різницю у стандартних похибках. Це означає, що довірчі інтервали при одному рівні значимості та для одного горизонту прогнозування для експоненціального тренду будуть вдвічі менш ніж для лінійного. Значний інтерес уявляє також аналіз похибок.

Зробимо аналіз показників зростання обох моделей. Спочатку зробимо оцінку похибки регресійного коефіцієнту β_1:

$$\sigma_{\beta 1} = \frac{s}{(\sum (X_i - \overline{X})^2)^{1/2}}$$

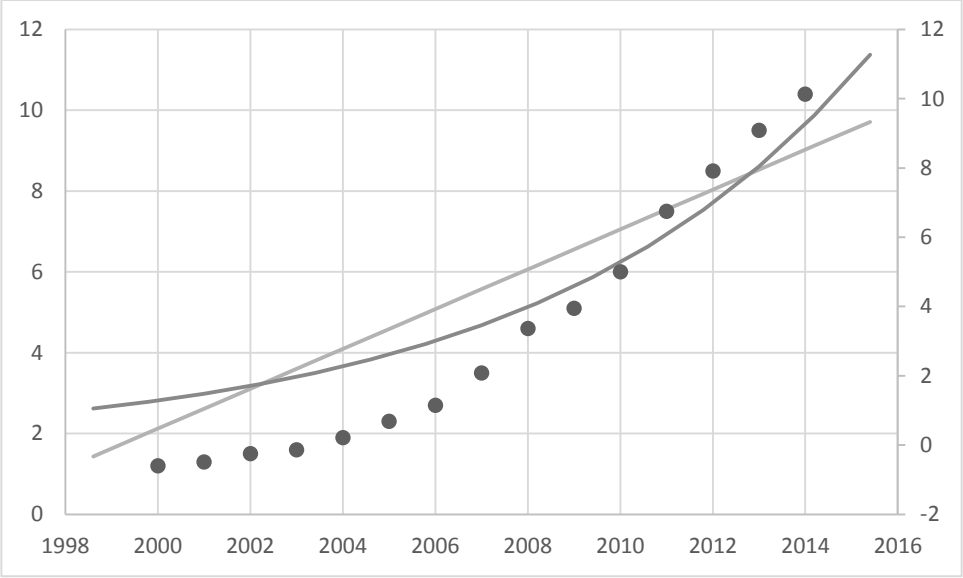


Рис. 6.8. Модель лінійного та експоненціального тренду розвитку економіки КНР.

Для моделі лінійного тренду з використанням даних табл.6.9, 6.10:

$$\sigma_{\beta 1} = \frac{s}{(\sum (X_i - \overline{X})^2)^{1/2}} = \frac{0{,}88}{\sqrt{280}} \approx 0{,}05 - \textit{лінійний},$$

$$\sigma_{\beta 1} = \frac{0{,}4}{\sqrt{280}} \approx 0{,}02 - \textit{експоненціальний}$$

t критерій для моделі лінійного тренду дорівнює 13,5 для експоненціального тренду 8,5. Це означає що нульова гіпотеза в обох випадках може буте відхилена на рівні значимості менш ніж 0,001.

Після відхилення нульової гіпотези можна трактувати зміст регресійних коефіцієнтів: для лінійної моделі щорічне зростання ВВП складало 0,69 трил.

USD, тоді, як відповідно експоненціальної моделі середні темпи зростання ВВП на заданому інтервалі складали 17%.

Виникає питання чому темпи зростання, що отримано по моделі суттєво перевищують темпи зростання економіки Китаю (7-9%). На наш погляд відповідь на ці питання полягає в тому що темпи зростання реального ВВП оцінюються з урахуванням інфляційних процесів, чого ні відбувається при фіксованому курсі та ВВП номінованому у доларах США.

Розглянемо ще декілька варіантів використання залежностей (6.46).

Приклад 9

Підібрати адекватну аналітичну залежність обсягу еякуляту від віку кнура. Вхідна інформація представлено в табл.6.9.

Розв’язок

Серед монотонних залежностей, що представлено в моделі STATGRAPHICS обираємо ту що має найкращі показники адекватності (коефіцієнт детермінації (6.41), критерій Фішера (6.44).

Табл. 6.9. Обсяг еякуляту та вік кнура.

Кличка	Напо-леон	Ес-мер	Овесі-йон	Роял Турк	Денні	Енорм	Ла 1	Ла 2	Вб 4	Вб 8
Вік, тижні X	61	65	68	62	60	72	74	66	62	61
Обсяг еякуляту, мл. Y	420	325	280	320	401	200	182	305	274	406
Обсяг еякуляту модельний, мл. $\hat{Y}$	379	291	248	353	410	207	191	275	353	379

Модельна похибка	41	34	32	-33	-9	-7	-9	30	-79	27

Із переліку залежностей наведеного вище набору 6.46 обираємо №3:

3. Reciprocal-Y model: Y = 1/(a + b*X)

Оцінки регресійних коефіцієнтів та їх стандартні похибки представлено в наступної таблиці:

Табл. 6.10. Оцінка регресійних коефіцієнтів та їх похибки.

	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept (a)	-0,00951855	0,00186893	-5,09304	0,0009
Slope (b)	0,000199262	0,0000286361	6,95841	0,0001

Підставляє значення регресійних коефіцієнтів у обрану модель отримаємо:

$$\hat{Y} = \frac{1}{-0.009518 + 0.000199 \cdot X} \qquad (6.54)$$

На підставі даних табл.6.11 (P-Value<0,0001) робимо висновок про то ще регресійні коефіцієнти не дорівнюють нулю, тобто нульова гіпотеза повинна бути відхилена з достовірністю, що перевищує 0,9999. На підставі даних таблиці .10 зробимо оцінку стандартної похибки (формула 1.22). Для цього використаємо гіперболічну модель в аналітичному вигляді (6.48) та розрахуємо модельні значення обсягу еякуляту в залежності від віку кнура (табл. 6.9).

Наступним кроком, розрахуємо стандартну похибку (6.11), вона дорівнює 40,1. Однак стандартна похибка, що розраховано за допомогою програми STATGRAPHICS має інше значення (Standard Error of Est. = 0,00041979). Це відбувається за рахунок лінеаризації, яка полягає в зміні залежної змінної на обернену величину.

Аналогічним шляхом (в межах лінеаризованої моделі) розраховано інші параметри (коефіцієнт детермінації, критерій Фішера), які потрібно перерахувати для вихідної моделі. Наприклад, якщо перерахувати коефіцієнт детермінації по даним табл.6.10, то він дорівнює 70,8%, а не 85,8% (R-squared = 85,8205 percent) у лінеаризованої моделі.

На діаграмі розсіювання (рис.6.8) представлено графік функції (6.48) та 95% довірчі інтервали для математичного очікування (помаранчевий колір), та окремих значень (чорний колір). Довірчий інтервал для математичного очікування розраховується відповідно (6.30), для окремих значень під. радикалом виявляється додаткова одиниця. Аналіз варіації та критерій Фішеру для лініарізованої моделі представлено в таблице 6.11.

Таблиця 6.11. Аналіз варіації та критерій Фішера.

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	0,00000853266	1	0,00000853266	48,42	0,0001
Residual	0,00000140979	8	1,76224E-7		
Total (Corr.)	0,00000994245	9			

Correlation Coefficient = 0,926394
R-squared = 85,8205 percent
R-squared (adjusted for d.f.) = 84,0481 percent
Standard Error of Est. = 0,00041979
Mean absolute error = 0,00031985
Durbin-Watson statistic = 2,40054 (P=0,6907)

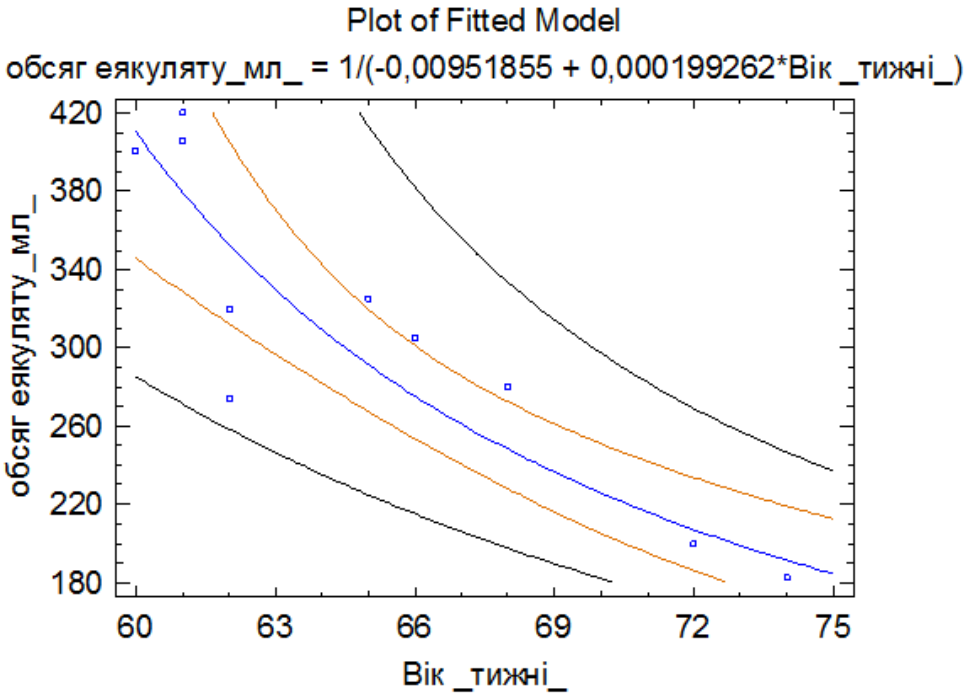


Рис. 6.9. Графічне уявлення даних спостережень та модельної залежності обсягу еякуляту від віку кнура.

Висновок:

Як слідує з наведеної моделі, залежність обсягу еякуляту від часу носить спадний характер, причому швидкість зменшення обсягу еякуляту зменшується з віком. Наведена функціональна залежність дозволяє прогнозувати обсяги еякуляту в залежності від віку кнура.

Приклад 10

Підібрати адекватну трендову залежність обсягу виробництва страусиних яєць від часу, що прийшов із початку несення птиці (табл.6.13).

Табл. 6.13. Дані спостережень

Час, тижні X	10	20	50	90	100	150	200	250	300
Обсяг, штукі Y	58	60	65	66	66	68	69	70	71

Розв’язок

Серед монотонних залежностей, що представлено у STATGRAPHICS обираємо ту, що має найкращі показники адекватності (коефіцієнт

детермінації (6.41), критерій Фішера (6.44). По цім критеріям нами обрано 10 модель (логарифмічна) з набору (6.47):

Logarithmic-X model: Y = a + b*ln(X)

Оцінки регресійних коефіцієнтів та їх стандартної похибки представлено в наступної таблиці (табл.6.14).

На підставі даних табл.6.14(P-Value<0,0001) робимо висновок про те, що регресійні коефіцієнти не дорівнюють нулю, тобто нульова гіпотеза повинна бути відхилено з достовірністю, що перевищує 0,9999.

Табл. 6.14. Оцінка регресійних коефіцієнтів та їх похибки

	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept (a)	49,2141	0,710506	69,2663	0,0000
Slope (b)	3,76599	0,155774	24,1759	0,0000

Маємо наступну залежність кількості страусиних яєць від часу:

$$\hat{Y} = 49{,}21 + 3{,}77 \cdot \ln X \qquad (6.55)$$

Представлена залежність показує, що швидкість зростання обсягів виробництва зменшується з часом залишаючись додатною величиною:

$$\frac{d\hat{Y}}{dX} = \frac{3{,}77}{X} \qquad (6.56)$$

Графічно дані спостережень та залежність (6.49) представлено на (рис.6.9). Аналіз варіації та критерій Фішера представлено в табл. 6.15.

Високе значення коефіцієнту детермінації та критерію Фішера свідчать, що модель адекватно описує спостерігаємо явище

Табл. 6.15. Аналіз варіації та критерій Фішера

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	153,056	1	153,056	584,47	0,0000
Residual	1,83308	7	0,261869		
Total (Corr.)	154,889	8			

Correlation Coefficient =	0,994065
R-squared =	98,8165 percent
R-squared (adjusted for d.f.) =	98,6474 percent
Standard Error of Est. =	0,511732
Mean absolute error =	0,327352
Durbin-Watson statistic =	2,59532 (P=0,7062)
Lag 1 residual autocorrelation =	-0,326681

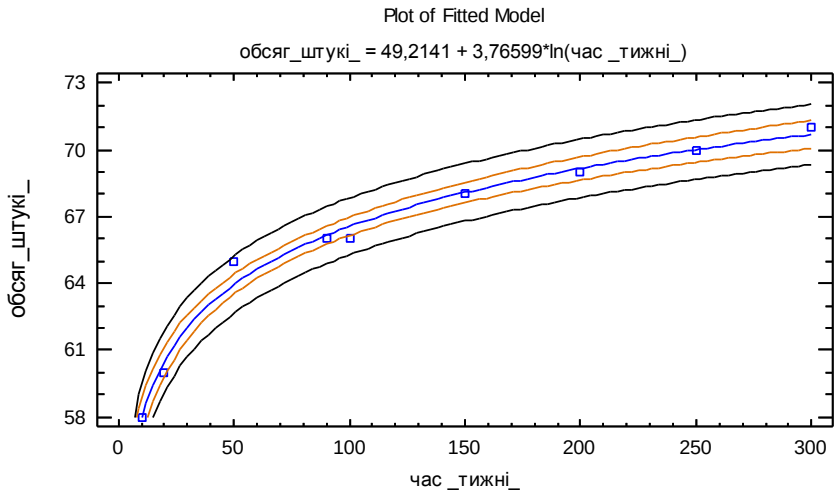


Рис. 6.10. Графічне уявлення даних спостережень та модельної залежності обсягу страусиних яєць від часу.

Висновок

Побудована динамічна модель залежності кількості яєць від часу, адекватно описує спостигнуте явище. Швидкість зростання обсягу виробництва є спадна функція часу.

Похибка апроксимації

Сучасне програмне забезпечення крім наведених вище критеріїв адекватності представляє і ряд інших критеріїв які надають можливість більш повно оцінити процес апроксимації даних спостережень.

MSE – середня квадратична похибка,

$$MSE = \sum_{i=1}^{T} (x(i) - \hat{x}(i))^2 / T$$;

де $x(i)$ – фактичне, $\hat{x}(i)$ – прогнозне значення;

MAE – середня абсолютна похибка,

$$MAE = \sum_{i=1}^{T} | x(i) - \hat{x}(i) | / T$$;

MPE – середня відсоткова похибка,

$$MPE = \sum_{i=1}^{T} \frac{(x(i) - \hat{x}(i))}{x(i)} * 100\% / T$$;

MAPE – середня абсолютна відсоткова похибка,

$$MAPE = \sum_{i=1}^{T} \left| \frac{x(i) - \hat{x}(i)}{x(i)} \right| * 100\% / T$$;

Завдання до теми 6

1.По даним світової статистики проаналізувати розвиток країни за останні 30 років що має ВВП на душу населення 20 +N тис. USD (N - кількість літерів у прізвіще). Викоритата моделі лінейного та екпоненційного тренду.

2.За допомогою економетричних методів праналізувати розвиток України (ВВП на душу населення) за останні 20 років

\

Тема 7. Множинна регресія

7.1 МНК оцінювання множинної регресії

Розглянемо випадок коли на ендогенну зміну впливає декілька факторів. Наприклад, на прибуток аграрного підприємства впливають як фінансові результати попереднього року, так і обсяг отриманих кредитів, крім того важливо врахувати обсяги землі, що орендує підприємство.

Слід підкреслити, що на першому кроці побудови економетричної моделі потрібно, на підставі економічної теорії виявити вісь діапазон факторів, що впливає на досліджуваний показник (ендогенну змінну).

Потім, слід виявити загальну аналітичну форму моделі (наприклад адитивна, мультиплікативна) а вже потім перейти безпосередньо до економетричних досліджень.

В подальшому нами буде розглянуто приклади з початкового етапу досліджень. Тому розглянемо безпосередньо алгоритм побудови множинної регресії. Нехай крім досліджуваної змінної y(j), $j = \overline{(1,n)}$ існує к причинних факторів $x_i, i = \overline{(1,k)}$ кожний з яких послідовно приймає n значень $x_i(j), (j = \overline{1,n})$

.

Лінійне регресійне рівняння аналогічно виразу (3.7) має наступний вигляд:

$$y(j) = \beta_0 + \beta_1 \cdot x_1(j) + \beta_2 \cdot x_2(j) + ... + \beta_k \cdot x_k(j) + \varepsilon(j), \; j = \overline{(1,n)} \qquad (7.1)$$

Рішення регресійного рівняння (знаходження невідомих регресійних коефіцієнтів) здійснюється за допомогою стандартного МНК з використанням матричної алгебри. Сума квадратів похибок:

$$RSS = \sum_{j=1}^{n} \varepsilon^2(j) = \sum (y(j) - \beta_0 - \sum_{j=1}^{k} \beta_i x_i(j))^2$$

За допомогою умови першого порядку отримаємо:

$$\partial(\sum_{j=1}^{n} \varepsilon^2(j)) / \partial\beta_0 = -2\sum_{j=1}^{n} \varepsilon(j) = 0$$

$$\partial(\sum_{j=1}^{n} \varepsilon^2(j)) / \partial\beta_i = -2\sum \varepsilon(j) x_i(j) = 0, i = 1,2,...,k \qquad (7.2)$$

Що еквівалентно системі рівнянь:

$$\sum_{j=1}^{n} y(j) = \beta_0 n + \beta_1 \sum_{j=1}^{n} x_1(j) + ... + \beta_k \sum_{j=1}^{n} x_k(j)$$

$$\sum_{j=1}^{n} y(j)x_1(j) = \beta_0 \sum_{j=1}^{n} x_1(j) + \beta_1 \sum x_1^2(j) + ... + \beta_k \sum_{j=1}^{n} x_1(j)x_k(j)$$

.....

.....

$$\sum_{j=1}^{n} y(j)x_k(j) = \beta_0 \sum_{j=1}^{n} x_k(j) + \beta_1 \sum_{j=1}^{n} x_1(j)x_k(j) + ... + \beta_k \sum_{j=1}^{n} x_k^2(j) \qquad (7.3)$$

На першому етапі будується вхідна матриця X розмірністю (к+1)*n, перший стовпчик якої складають одиниці, другій зміна x_1, останній зміна x_k:

$$X = \begin{pmatrix} 1...x_1(1)...x_2(1).....x_k(1) \\ 1...x_1(2)...x_2(2).....x_k(2) \\ ..... \\ 1...x_1(n)...x_2(n).....x_k(n) \end{pmatrix} \qquad (7.4)$$

Наступним кроком розраховується матриця $X'X$, де X' матриця транспоновано відносно X:

$$X'X = \begin{pmatrix} n...\sum_{j=1}^{n} x_1(j)...\sum_{j=1}^{n} x_2(j).....\sum_{j=1}^{n} x_k(j) \\ .....\sum x_1^2(j)...\sum_{j=1}^{n} x_1(j)x_2(j).....\sum_{j=1}^{n} x_1(j)x_k(j) \\ ..... \\ .....\sum_{j=1}^{n} x_k^2(j) \end{pmatrix} \qquad (7.5)$$

Матриця $X'X$ симетрична матриця розмірністю (к+1)*(к+1), тому у виразі (6.59) відображено головна діагональ та елементи по над головною діагоналлю.

Вектор регресійних коефіцієнтів $\overline{\beta}(\beta_0, \beta_1, ..., \beta_k)$ розраховується наступним шляхом:

$$\overline{\beta} = (X' \cdot X)^{-1}(X' \cdot Y) \qquad (7.6)$$

Матриця $(X' \cdot Y)$ розраховується наступним шляхом:

$$X'Y = \begin{pmatrix} \sum_{j=1}^{n} y(j) \\ \sum_{j=1}^{n} y(j)x_1(j) \\ \sum_{j=1}^{n} y(j)x_2(j) \end{pmatrix} \qquad (7.7)$$

Оскільки у виразі (6.60) присутня обернена матриця, визначник матриці (6.59) повинен відрізнятись від нуля. А це означає що стовпчики вхідної матриці не повинні бути взаємозалежні. У протилежному випадку визначник матриці (6.59) наближується до нуля і регресійні коефіцієнти не будуть коректно визначени. Най простіший шлях, якщо два вхідних факторів корелюють, то потрібно залишити тільки один. Існують і інші варіанти, які будуть розглянуто пізніше.

У випадку двох пояснюючих змінних матриця $X'X$ має вигляд:

$$X'X = \begin{pmatrix} n\ldots\ldots\ldots\sum_{j=1}^{n} x_1(j)\ldots\ldots\ldots\sum_{j=1}^{n} x_2(j) \\ \sum_{j=1}^{n} x_1(j)\ldots\sum_{j=1}^{n} x_1^2(j)\ldots\ldots\ldots\sum_{j=1}^{n} x_1(j)x_2(j) \\ \sum_{j=1}^{n} x_2(j)\ldots\sum_{j=1}^{n} x_1(j)x_2(j)\ldots\sum_{j=1}^{n} x_2^2(j)\ldots\ldots \end{pmatrix} \qquad (7.8)$$

Приклад 11

Розглянемо алгоритм знаходження багатофакторної регресійної залежності на прикладі ефективності діяльності аграрних підприємств. В якості показника ефективності використовується прибуток підприємства (залежна змінна-у (тис. грн.)), в якості пояснюючих факторів площа земельних ділянок - x_1 (га) та витрати x_2 (тис. грн.). П'ятнадцять фермерських підприємств, що працюють у близьких умовах мають наступні показники доходу, площ посивів, та витрат на проведення посівної компанії табл.1.15. Побудувати регресійну залежність доходу від площі посивів та витрат.

Розв’язок

Послідовність кроків розв’язку задачі представлено у табл. 6.16. На першому кроку ми знайдемо елементи матриць (6.59) і (6.61). на наступному кроці потрібно знайти матрицю обернену до матриці $X'X$. На останньому кроці робляться оцінки регресійних коефіцієнтів відповідно виразу (6.60), та оцінки похибки моделі.

Табл.7.1. Вихідні дані та побудова ко варіаційної матриці у випадку двох незалежних змінних.

Ферма	Прибуток (у тис. 141рн..)	S (x1,га)	Витрати (x2, тис.грн)	$\sum_{j=1}^{n} x_1 x_2$	$\sum_{j=1}^{n} x_1^2$	$\sum_{j=1}^{n} x_2^2$	$\sum_{j=1}^{n} x_1 y$	$\sum_{j=1}^{n} x_2 y$	Прибуток (у тис. 141рн..)	e	e2
1	2300	100	200	20000	10000	40000	230000	460000	2216	83	6895
2	2900	110	300	33000	12100	90000	319000	870000	2825	74	5545
3	3500	200	250	50000	40000	62500	700000	875000	3468	31	987
4	1600	50	200	10000	2500	40000	80000	320000	1718	-118	14010
5	2000	70	210	14700	4900	44100	140000	420000	1968	31	980
6	2400	60	300	18000	3600	90000	144000	720000	2326	73	5338
7	3000	120	300	36000	14400	90000	360000	900000	2925	74	5587
8	4100	200	400	80000	40000	160000	820000	1640000	4231	-131	17384
9	5000	300	350	105000	90000	122500	1500000	1750000	4974	25	644
10	2500	140	200	28000	19600	40000	350000	500000	2615	-115	13418
11	2200	120	180	21600	14400	32400	264000	396000	2314	-114	13140
12	2600	150	160	24000	22500	25600	390000	416000	2512	87	7740
13	3000	180	180	32400	32400	32400	540000	540000	2912	87	7578
14	2400	130	200	26000	16900	40000	312000	480000	2516	-116	13483
15	2400	90	250	22500	8100	62500	216000	600000	2371	28	802
$\sum$	41900	2020	3680	521200	331400	972000	6365000	10887000	41900	0	113539

	15	2020	3680
X’X	2020	331400	521200
	3680	521200	972000

	0,97520	-0,00088	-0,00322
(X’X)-1	-0,00088	0,00002	-0,00001
(3.64)	-0,00322	-0,00001	0,00002

	41900
x’Y	6365000
	10887000

	202,0773
beta	9,971921
	5,088467

Табл.7.2. Дані для розрахунку критеріїв адекватності та стандартизованих оцінок регресійних коефіцієнтів

Ферма	$(y-\bar{y})^2$	$(x_1-\bar{x}_1)^2$	$(x_2-\bar{x}_2)^2$	$(\hat{y}-\bar{y})^2$
1	8628906,3	23256,25	67600	9123666
2	5463906,3	20306,25	25600	5817602
3	3018906,3	2756,25	44100	3129080
4	13231406	41006,25	67600	12384280
5	10481406	33306,25	62500	10685123
6	8051406,3	37056,25	25600	8471421
7	5006406,3	17556,25	25600	5346503
8	1293906,3	2756,25	3600	1011333
9	56406,25	2256,25	12100	69106,22
10	7493906,3	12656,25	67600	6873104

11	9226406,3	17556,25	78400	8543173
12	6956406,3	10506,25	90000	7428245
13	5006406,3	5256,25	78400	5403536
14	8051406,3	15006,25	67600	7405913
15	8051406,3	26406,25	44100	8212985
Σ	100018594	267643,75	760400	99905070

В результаті розрахунків ми отримали наступне регресійне рівняння залежності доходу (у) від площ посивів x_1 та витрат x_2:

$$\hat{y} = 202{,}1 + 10x_1 + 5{,}1x_2 \qquad (7.9)$$

Отримані регресійні коефіцієнти можна трактувати наступним шляхом: зростання площа посевов на 1 га призводить до зростання доходу ферми на 10 тис. грн., зростання витрат на посівну компанію на 1 тис. грн. призводить до зростання доходу на 5,1 тис. грн.

7.2. Параметри адекватності моделі множинної регресії

Параметри адекватності розраховуються аналогічно лінійної однофакторної моделі (6.38),(6.41), (6.44):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} = \frac{b^2{}_1 \sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2},$$

$$R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\varepsilon_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \qquad (7.10)$$

$$F = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 \Big/ m}{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \Big/ (n-m-1)} \qquad (7.11)$$

Стандартна похибка розраховується з урахуванням кількості входів:

$$s = \sqrt{\frac{\sum_{1}^{n}\varepsilon_i^2}{n-(k+1)}} \qquad (7.12)$$

Розрахунки можна здійснити як за допомогою MS Excel, так і спеціалізованих програм: STATGRAPHICS, STATISTICS та інших.

Приклад 12

Наведемо приклади розрахунку параметрів адекватності моделі. Для цього використаємо попередний приклад регресійної моделі доходу фермерських господарств (табл.6.17).

Розв'язок

Спочатку розрахуємо стандартну похибку. У цьому випадку кількість незалежних зміних дорівнює 2 (к=2), тоді відповідно виразу (6.55) та табл.6.17:

$$s = \sqrt{\frac{113539}{15-(2+1)}} \approx 97{,}3 тис.грн \qquad (7.13)$$

Коефіцієнт детермінації та критерій Фішера дорівнюють відповідно (6.62) та (6.63):

$$R^2 = 0{,}989; F = 544{,}1$$

Значення параметрів адекватності моделі свідчать, що модель достатньо адекватно представляє залежність доходу ферм від площ що обробляються та витрат на посівну.

7.3.Оцінка довірчих інтервалів для коефіцієнтів множинної регресії

Довірчі інтервали для оцінок регресійних коефіцієнтів, що отримано за допомогою системи рівнянь (6.59) залежать від рівня лінійного взаємозв'язку між екзогенним змінними.

Визначимо вхід x_1 за допомогою інших входів:

$$x_1 = \hat{x}_1 + \nu_1$$

Позначимо коефіцієнт детермінації останього рівняння R_1^2.

Дісперсія оцінки регресійного коефіцієнту β_1 визначається:

$$\text{var}(\beta_1) = s^2 / \sum (x_1(j) - \bar{x}_1)^2 (1 - R_1^2) \qquad (7.14)$$

Останній вираз показував, що при зростанні коефіцієнту детермінації одного з входів по іншим похибка оцінки регресійного коефіцієнту зростає. У подальшому це буде використане при визначенні шляхів зменшення колініарності.

Для визначення похибок оцінок регресійних коефіцієнтів існує інший шлях, що базується на розрахунках здійснених при оцінки регресійних коефіцієнтів. Введемо поняття коваріаційної матриці векторної змінної $\bar{x}(x_1; x_2, ..., x_k)$:

$$\sigma_{ij} = E((x_i - \bar{x}_i)(x_j - \bar{x}_j)) \qquad (7.15)$$

Це симетрична квадратна матриця вимірності к*к на головної діагоналі якої розтушоване дисперсії складових процесу. При множенні коваріаційної матриці на $1/(\sigma_i \sigma_j)$ отримаємо симетричну кореляційну матрицю на головної діагоналі розтушоване 1, а інші елементи коефіцієнти кореляції складових процесу, що досліджується.

Коваріаційна матриця похибок регресійних коефіцієнтів:

$$\sum_{\beta} = E((\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)) = s\sqrt{(X'X)^{-1}_{jj}} \qquad (7.16)$$

Якщо позначити діагональні елементи матриці $(X'X)^{-1}$ як $a_{00}; a_{11}, ..., a_{kk}$ то похибка оцінки регресійних коефіцієнтів визначається:

$$\sigma_{\beta 0} = s\sqrt{a_{00}}; \sigma_{\beta 1} = s\sqrt{a_{11}}; ..., \sigma_{\beta k} = s\sqrt{a_{kk}} \qquad (7.17)$$

Значущість регресійних коефіцієнтів можна перевірити якщо врахувати, що $(\beta_j - \hat{\beta}_j) / \sigma_{\beta j}$ відповідає розподілу Стьюдента з $\nu = n - (k+1)$ ступенями

свободи. Як правило першою перевіряється гіпотеза ріності нулю кожного з регресійних коефіцієнтів. Якщо виконується умова:

$$t = \frac{\left|\widehat{\beta}_j\right|}{\sigma_{\beta j}} \geq t_{v;\alpha} \qquad (7.18)$$

То нульова гіпотеза відхиляється на рівні значущості α. Аналогичним шляхом перевіряється гіпотеза що регресійний коефіцієнт дорівнює деякому відомому з інших джерел значенню:

$$t = \frac{\left|\beta_j - \widehat{\beta}_j\right|}{\sigma_{\beta j}} > t_{v;\alpha} \qquad (7.19)$$

Довірчі інтервали для значень регресійних коефіцієнтів з довірчою ймовірністю $1-\alpha$:

$$\widehat{\beta}_j - t_{v;\alpha/2}\sigma_{\beta j} \leq \beta_j \leq \widehat{\beta}_j + t_{v;\alpha/2}\sigma_{\beta j} \qquad (7.20)$$

Приклад 13

Зробити перевірку нульової гіпотези для кожного з регресійних коефіцієнтів прикладу 11.

Розв'язок

Значення стандартної похибки відповідно (6.69) дорівнює 97,3 тис. грн. За допомогою виразів (6.73) та діагональних елементів оберненої матриці (6.64) отримаємо похибки оцінок регресійних коефіцієнтів:

$$\sigma_{\beta 0} = 97{,}3\sqrt{0{,}9752} \approx 96{,}1; \sigma_{\beta 1} = 97{,}3\sqrt{0{,}00002} \approx 0{,}44; \sigma_{\beta 2} = 97{,}3\sqrt{0{,}00002} \approx 0{,}44$$

Величини регресійних коефіцієнтів подано у рівнянні (3.65), за допомогою (3.74) знайдемо значення статистики Стьюденту:

$$t_0 = \frac{202{,}1}{96{,}1} \approx 2{,}1; t_1 = \frac{10{,}0}{0{,}44} \approx 22{,}7; t_2 = \frac{5{,}1}{0{,}44} \approx 11{,}6$$

Критичне значення розподілу Стьюдентуз 12 ступенями свободи на рівні значущості 0,1 дорівнює 1,8; на рівні значущости 0,001 -4,3. Це означає, що нульова гіпотеза для коефіцієнту β_0 може бути відхилено на рівні значимості 0,1; тоді як для коефіцієнтів $\beta_1; \beta_2$ на рівні значимості 0,001. Відхилення нульової гіпотези для коефіцієнту β_0 можна трактувати, як

існування інших джерел доходів фермерських господарств не пов'язаних з вирощуванням зернових. Однак доходи, що пов'язані з головним напрямком господарювання мають більш стабільний характер (малий рівень значущості). У сучасних програмних засобах розраховується також рівень значущості (P-Value) для оцінок коефіцієнтів, який показує ймовірність відхилення нульової гіпотези. Для вільного члену регресійного рівняння P-Value дорівнює 0,06; для двох інших коефіцієнтів менш ніж 0,001.

Довірчі інтервали для прогнозних значень

Для оцінки довірчих інтервалів для прогнозних значень множинної регресії використовується поняття квадратичної форми, що побудовано на матриці $(X' \cdot X)^{-1}$.

Нехай у випадку к входів потрібно оцінити похибку прогнозу в точці з координатами.

$\bar{x}^* = \begin{pmatrix} 1 \\ x_1^* \\ \dots \\ x_k^* \end{pmatrix}$. Похибка на рівні значимості α дорівнює:

$$\Delta = s \cdot t_{\nu;\alpha./2} \cdot \left((\bar{x}^*)' \cdot (X'X)^{-1} \bar{x}^* \right)^{0,5} \qquad (7.21)$$

Для випадку двох входів цей вираз можна представити у вигляді:

$$\bar{x}^* = \begin{pmatrix} 1 \\ x_1^* \\ x_2^* \end{pmatrix} \qquad (7.22)$$

$$\Delta(\bar{x}^*) = s \cdot t_{n-3;\alpha/2} \cdot \sqrt{a_{11} + 2a_{12}x_1^* + 2a_{13}x_2^* + a_{22}x_1^*x_1^* + a_{33}x_2^*x_2^* + 2a_{23}x_1^*x_2^*}$$

Де $a_{ij} -$ елементи матриці $(X'X)^{-1}$. Тоді довірчі інтервали для прогнозних значень на рівні значущості α в точці $\bar{x}^*$:

$$\hat{y}(\bar{x}^*) - \Delta(\bar{x}^*) \le y(\bar{x}^*) \le \hat{y}(\bar{x}^*) + \Delta(\bar{x}^*) \qquad (7.23)$$

Приклад 14

На підставі прикладу 11 розрахувати довірчи інтервали для доходу фермерського господарства з площиною посевів 200 га та витратами 500 тис. грн.

Розв'язок

На підставі оцінок регресійних коефіцієнтов отримано наступне регресійне рівняння:

$$\hat{y} = 202{,}1 + 10x_1 + 5{,}1x_2$$

Знайдемо очікуване значення доходу для ферми з площею посевів 200 га та витратами 500 тис. грн.:

$$\hat{y}(200;500) = 202{,}1 + 10 \cdot 200 + 5{,}1 \cdot 500 = 4752{,}1 тис.грн$$

Виберемо 90% довірчий інтервал для прогнозного значення. Звідси $\alpha = 1 - 0{,}9 = 0{,}1 \Rightarrow \alpha / 2 = 0{,}05$.

З таблиць розподілу Стьюденту $t_{12;0,05} = 1{,}78$

За допомогою (6.78) розрахуємо похибку прогнозу: $x_1^* = 200; x_2^* = 500; s = 97{,}3.$ Значення елементів оберненої матриці подано у (6.64):

$$\Delta(200;500) = 97{,}3 \cdot 1{,}78(0{,}9752 - 2 \cdot 0{,}00088 \cdot 200 - 2 \cdot 0{,}0032 \cdot 500 + 0{,}00002 \cdot 200^2 +$$
$$+ 0{,}00002 \cdot 500^2 - 2 \cdot 0{,}00001 \cdot 200 \cdot 500)^{0{,}5} \approx 191{,}5 тис.грн$$

Звідсі 90% довірчій інтервал для прогнозних значень доходу для ферми з площиною угідь 200 га та витратами 500 тис. грн:

$$4560{,}6 тис.грн \le y(200;500) \le 4943{,}6 тис.грн$$

В програмі STATGRAPHICS передбачено побудова довірчих інтервалів.

7.4. Коефіцієнти еластичності та стандартизовані коефіцієнти регресії

Крім стандартного підходу до трактовки регресійних коефіцієнтів як приріст результуючої змінної на одиничний приріст пояснюючих змінних (маржинальний підхід) існують інші варіанти трактування результатів регресійного аналізу. В цьому випадку застосовуються стандартизовані

регресійні коефіцієнти (тобто коефіцієнти що враховують варіативність як результуючої так і пояснюючих змінних), та показник еластичності результуючої по кожної з пояснюючих змінних. Почнемо з показнику еластичності.

Показник еластичності відповідає відносному приросту результуючої змінної на 1% приріст пояснюючої змінної. кількох задачах потрібно прорахувати реакцію залежної змінної на приріст однієї з незалежних змінних. Еластичність по j входу визначається:

$$E_j = \frac{\Delta_j y}{y} / \frac{\Delta x_j}{x_j} \qquad (7.24)$$

де $\Delta_j y -$ прирост результуючий змінної, що обумовлено приростом j пояснюючої змінної. Прирост по j зміної визначається:

$$\Delta_j y = \beta_j \cdot \Delta x_j$$

У випадку лінійної залежності показник еластичності залежить від точки в якої він розраховується, тому вважається що найбільш адекватно процес відображає еластичність точці з координатами $(\bar{x}; \bar{y})$ що належить регресійному рівнянню:

$$E_j = \hat{\beta}_j \frac{\overline{x_j}}{\bar{y}} \qquad (7.25)$$

Перейдемо до стандартизованого регресійного коефіцієнту, який розраховується наступним шляхом:

$$\beta'_j = \hat{\beta}_j \cdot \frac{\sigma_{xj}}{\sigma_y} \qquad (7.26)$$

Приклад 15

По даним задачі про фермерські господарства (табл.6.16; табл.17) оцінити показники еластичності та стандартизованих регресійних коефіцієнтів доходу по площі та витратам на посівну.

Розв’язок

Для оцінок еластичності отримаємо оцінки середніх показників доходів, площ, та витрат (табл.) . Оцінки еластичності доходу по площі та витратам:

$$E_1 = 10\frac{252,5}{2793,3} = 0,9; E_2 = 5,1\frac{460}{2793,3} = 0,84$$

Оцінки стандартизованих регресійних коефіцієнтів:

$$\beta_1' = 10\frac{133,6}{833} = 1,6; \beta_2' = 5,1\frac{225,2}{833} = 1,4$$

	Дохід (у тис. грн)	Площа (х1 га)	Витрати (х2 тис. грн)
Середньо значення	2793,3	252,5	460
Стандартне відхилення	833	133,6	225,2

Слід підкреслити, що у попередньому прикладу коефіцієнт еластичності не є сталою величиною, а залежить від точки в якої про зводяться розрахунки, а показник еластичності розраховується в точці яку умовно можна вважати центром ваги $O(\bar{x}_1;\bar{x}_2;...;\bar{x}_к;\bar{y})$. Однак існують функції які мають стали еластичності. До таких функцій відноситься виробнича функція Коба-Дугласа. Розглянемо приклад оцінки параметрів виробничої функції за допомогою економе річного аналізу.

Приклад 16

Розрахувати виробничу функцію Коба-Дугласа по відомим показникам обсягів виробництва - У (млн. грн.), основним фондам - К (млн. грн.), обсягам використаної робочої сіли - L (тис. люд. год.), що представлено в таблиці 6.18. Перевірити гіпотезу сталої віддачи від обсягів виробництва.

Таблиця 7.3. Вхідна інформація для оцінки виробничої функції

T (роки)	Y(млн. грн.) (z)	K (млн .грн.)(x_2)	L (люд. год.)(x_3)	$\hat{Y}$	$Y-\hat{Y}$
2000	13(2,565)	10(2,303)	3(1,099)	16,8	-3,8
2001	16(2,773)	15(2,700)	2(0,693)	17,5	-1,5
2002	22(3,091)	17(2,033)	4(1,386)	24,1	-2,1
2003	26(3,250)	20(2,996)	5(1,609)	28,2	-2,2
2004	23(3,135)	21(3,045)	3(1,099)	23,9	-0,9
2005	30(3,401)	24(3,170)	5(1,609)	30,8	-0,8
2006	31(3,434)	28(3,332)	4(1,386)	30,5	0,5
2007	27(3,296)	26(3,250)	3(1,099)	26,5	0,5
2008	26(3,250)	29(3,367)	2(0,693)	24,0	2,0
2009	28(3,332)	33(3,497)	2(0,693)	25,5	2,5

Розв'язок

Якщо виробнича функція залежить від часу T, то вона має наступний вигляд:

$$Y(T,K,L) = B \cdot e^{\gamma(T-2000)} \cdot K^{\alpha} \cdot L^{\beta}, \qquad (7.27)$$

де B сталій множник, $\gamma-$ невідомий відсоток річного зростання, що обумовлено техніческім прогресом, $\alpha-$ показник еластичності по капіталу, $\beta-$ показник еластичності по обсягам праці. Для лінеаризації рівняння (6.83) потрібно взяти логарифм від обох частин цього рівняння.

Зробимо позначення:

$$LnY = z; LnB = \beta_0; \gamma = \beta_1; T - 2000 = x_1; \alpha = \beta_2; LnK = x_2; \beta = \beta_3; LnL = x_3$$

Отримаємо наступне лінеаризоване рівняння :

$$z = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \hat{\beta}_3 \cdot x_3 \qquad (7.28)$$

На першому етапи розрахуємо стаціонарну функцію ($\beta_1 = 0$) Кобба-Дугласа (немає залежності від часу) та перевіримо гіпотезу сталої віддачи обсягу виробництва ($\beta_2 + \beta_3 = 1$).

$$z = \hat{\beta}_0 + \hat{\beta}_2 \cdot x_2 + \hat{\beta}_3 \cdot x_3 \qquad (7.29)$$

В табл. 1.3 наведено значення змінних $\hat{\beta}_0; \hat{\beta}_2; \hat{\beta}_3$. Для розрахунків використано програма STATGRAPHICS.

Оцінки регресійних коефіцієнтів та їх похибок наведено в наступній таблиці (табл. 6.18).

Табл. 7.4. Оцінка регресійних коефіцієнтів та їх похибок

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	1,31644	0,414871	3,17312	0,0156
X2	0,476734	0,114384	4,16784	0,0042
X3	0,370137	0,15307	2,41809	0,0462

Рівень значущості розрахованих коефіцієнтів (P-Value) відповідно таблиці 6.18 не перевищує 4,6 %. Тобто всі розраховані коефіцієнти відрізняються від нуля на рівні значимості, що не перевищує 5%. В таблиці 1.19 подано показники адекватності моделі. Коефіцієнт детермінації (R-squared) приведено нижче.

Однак, слід підкреслити наступну особливість використання програми STATGRAPHICS, коефіцієнт детермінації та критерій Фішера розраховуються не для початкової моделі (6.83), а для лінеаризованої (3.84). Це також стосується і надзвичайно важливої характеристики моделі – стандартної похибки.

Таблиця 7.5. Оцінка регресійних коефіцієнтів та їх похибок

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	0,522759	2	0,261379	9,98	0,0089

Residual	0,183325	7	0,0261893
Total (Corr.)	0,706084	9	

R-squared =	74,0363 percent
R-squared (adjusted for d.f.) =	66,6181 percent
Standard Error of Est. =	0,161831
Mean absolute error =	0,0991287
Durbin-Watson statistic =	1,77883 (P=0,1462)
Lag 1 residual autocorrelation =	-0,0913896

Відповідно до даних таблиці 6.20 лінеаризована регресійна залежність має вигляд:

$$\ln \widehat{Y} = 1{,}316 + 0{,}478 \cdot \mathrm{x}_2 + 0{,}37 \cdot \mathrm{x}_3 \qquad (7.30)$$

Це відповідає наступному вигляду виробничої функції:

$$\widehat{Y} = 3{,}73 \cdot K^{0.478} \cdot L^{0.37} \qquad (7.31)$$

В таблиці 6.20 представлені модельні значення обсягу виробництва в залежності від обсягів праці та капіталу, що дозволяє оцінимо стандартну похибку, вона дорівнює 2,33 млн. грн. Слід зазначити, що вона не відповідає значенню отриманого за допомогою програми STATGRAPHICS, де вона становить (Standard Error of Est. = 0,161831). Це відбувається через те, що остання оцінка отримана із лінеарізованого рівняння (логарифмованих вихідних даних).

Розглянемо загальну постанову задачі з урахуванням фактору часу. Суттєво покращилися всі параметри адекватності моделі (табл.7.6).

Таблиця 7.6. Оцінка регресійних коефіцієнтів та їх похибок

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value

Model	0,6512	2	0,217	23,8	0,001
Residual	0,0548	7	0,009		
Total (Corr.)	0,7060	9			

R-squared = 92,244 percent

R-squared (adjusted for d.f.) = 88,366 percent

Standard Error of Est. = 0,0955371

Mean absolute error = 0,0527715

Durbin-Watson statistic = 1,82189 (P=0,1699)

Lag 1 residual autocorrelation = -0,197993

Оцінки регресійних коефіцієнтів та їх похибок з урахуванням часу наведено в наступній таблиці (табл. 6.22).

Таблиця 7.7. Оцінка регресійних коефіцієнтів та їх похибок

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	2,13	0,33	3,173	0,0156
X1	0,0794	0,02116	3,75	0,0095
X2	0,0559	0,131	0,427	0,683
X3	0,435	0,092	4,73	0,0032

Отримаємо наступне лінеаризоване рівняння :

$$z = 2{,}13 + 0{,}0794 \cdot x_1 + 0{,}0559 \cdot x_2 + 0{,}435 \cdot x_3 \qquad (7.32)$$

Яке відповідає наступному вигляду виробничої функції:

$$\widehat{Y} = 8{,}41 \cdot e^{0{,}0794(T-2000)} \cdot K^{0{,}0559} \cdot L^{0{,}435} \qquad (7.33)$$

Останній вираз означає практично 8% зростання обсягів виробництва щорично, однак внесок капіталу существенно зменьшився і нульову гипотезу відносно еластичності від капіталу неможна відхилити.

Висновок

За даними спостережень побудовано виробничу функцію шляхом визначення еластичності по праці, капіталу та коефіцієнту масштабу. Виявилось, що у заданої виробничої функції не виконується умова сталої віддачі масштабу виробництва, вмвлено також суттевий вплив научно техничного прогресу, що забеспечує 8% зростання обсягів виробництва.

Приклад 17

Агент з продажу нерухомості накопичив за 2007 рік деяку статистику для більш реалістичного визначення цін будинків в передмісті м. Харків (ціни приведено в доларах США внаслідок значного інфляційного тренду гривні).

Для цього він зробив припущення про існування прямої лінійної залежності між ціною нерухомості - у ($ тис.), площею будинку - $x_1(м^2)$, та площею прибудинкової ділянки - $x_3(com.(100м^2))$ та лінійної оберненої залежності між ціною і віком будинку $x_2(роки)$.

Знайти лінійну залежність ціни від визначених показників, оцінити правильність та достовірність зроблених припущень (рівень значимості), оцінити параметри адекватності моделі, оцінити еластичність ціни по вхідним параметрам моделі.

Вхідна інформація представлено в наступної таблиці 7.8.

Таблиця 7.8. Вхідна інформація для оцінки вартості нерухомості

№	Ціна продажу ($1000)	Площа будинку ($м^2$)	Вік будинку (роки)	Площа ділянки (com.)
1	89,5	200	5	4,1
2	79,9	148	10	6,8
3	85,1	205	8	6,3
4	56,9	125	7	5,1
5	66,6	180	8	4,2
6	82,5	143	12	8,6
7	126,3	275	1	4,9

8	79,3	165	10	6,2
9	119,9	243	2	7,5
10	87,6	202	8	5,1
11	112,6	220	7	6,3
12	120,8	190	11	12,9
13	78,5	123	16	9,6
14	74,3	140	12	5,7
15	74,8	167	13	4,8
$\sum_{1}^{15}$	1334,6	2726	132	98.1

Розв’язок

Розв’яжемо задачу за допомогою програми STATGRAPHICS.

В якості залежної величини візьмемо ціну продажу, а в якості факторних ознак: площа будинку, вік будинку, площа ділянки.

Оцінки регресійних коефіцієнтів та їх похибок наведено в наступній таблиці (табл. 7.9).

Таблиця 7.9. Оцінка регресійних коефіцієнтів та їх похибок

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	-19,4819	19,0154	-1,02453	0,3297
Площа будинку	0,423469	0,073508	5,76086	0,0002
Вік будинку	-0,107541	0,870496	-0,12354	0,9041
Площа ділянки	4,91069	0,89066	5,51354	0,0003

Як бачимо із наведених у табл.7.9. розрахунків, найменш значимою ознакою є вік будинку (коефіцієнт P-Value = 0.904).

Найбільш впливовими ознаками є площа будинку та площа ділянки. Коефіцієнт детермінації (R-squared) приведено нижче. Через те, що за для побудови регресійної моделі лінералізація даних не відбувалася, результат розрахунку стандартної помилки у STATGRAPHICS є вірним.

Таблиця 7.10. Параметри адекватності моделі

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	5760,1	3	1920,03	42,61	0,0000
Residual	450,61	10	45,061		
Total (Corr.)	6210,71	13			

R-squared = 92,7446 percent

R-squared (adjusted for d.f.) = 90,568 percent

Standard Error of Est. = 6,71275

Mean absolute error = 4,09148

Durbin-Watson statistic = 2,26222 (P=0,6981)

Lag 1 residual autocorrelation = -0,154647

Відповідно даним таблиці 6.25 регресійна залежність має вигляд:

$$\hat{Y} = -19{,}4819 + 0{,}423x_1 - 0{,}108x_2 + 4{,}91x_3 \qquad (7.34)$$

Де коефіцієнти при змінних означають оцінки регресійних коефіцієнтів рівняння (6.90). Розглянемо економічний зміст отриманих регресійних коефіцієнтів.

Вільний член рівняння (6.90), що відповідає нульовим значенням незалежних змінних не можливо трактувати, тому, що нерухомість з нульовою площею ділянки та нульовою площею будинку немає змісту. Тому вільний член має виключно розрахунковий характер та використовується для зменшення стандартної похибки.

Що стосується інших коефіцієнтів то вони мають очікувані знаки: з зростанням площі будинки та ділянки вартість нерухомості зростає, а з зростанням віку експлуатації зменшується тобто припущення, що зроблено при побудові моделі підтверджуються.

Величина коефіцієнту при першому вході свідчить, що зростання площі будинку на 1 квадратний метр у середньому збільшує ціну будинку на $423, при другому вході, що збільшення віку експлуатації на 1 рік веде до зменшення ціни на $102, зростання площі ділянки на 1 сотку веде до зростання ціни на $4911. Достовірність припущень перевіряється на підставі перевірки нульової гіпотези (істинне значення регресійних коефіцієнтів дорівнює нулю). Розраховується значення t відношення :

$$t = \left\| \frac{\hat{\beta}}{\sigma_\beta} \right\|$$

та порівнюється з критичним значенням розподілу Стьюденту $t_{(n-(k+1));\alpha}$. У випадку задачі з оцінкою нерухомості (n=15;k=3; $\alpha = 0{,}005$) $t_{11;0{,}005} = 3{,}1$.

Значення оцінок регресійних коефіцієнтів та їх похибок наведено в табл. 6.25:

На підставі наведених даних можна зробити припущення, що нульова гіпотеза відкидається для першої (площа будинку) та третьої (площа ділянки) змінної, однак її не можна відхилити для другої змінної, тобто, вік будинку можна в моделі не враховувати, якщо це не призведе до суттєвого погіршення параметрів адекватності моделі до яких відносяться коефіцієнт детермінації R^2=92,7%, критерій Фішеру F=42,6; стандартна похибка s=$6,7 тис.

Оцінка коефіцієнту детермінації означає що 92,7% дисперсії ціни визначається моделлю, адекватність моделі визначається по значенню параметру Фішера, який порівнюється з табличним значенням розподілу Фішера з ступенями свободи $\nu_1 = k - 1 = 2; \nu_2 = n - k = 12$.

Крім того потрібно задати рівень значимості (ймовірність похибки), який обирається рівним 0,01. Тоді відповідне значення розподілу Фішера дорівнює: $F_{2;11;0{,}01} = 7{,}2$. Оскільки модельне значення перевищує табличне, модель можна вважати адекватною.

Для розрахунку показників еластичності по всім трьом змінним використовуються середні показники вхідних змінних, які, на підставі

табл.6.25: $\bar{x}_1 = 181{,}7 м^2$; $\bar{x}_2 = 8{,}8 років$; $\bar{x}_3 = 6{,}5 com$. Значення функції в цій точці відповідно дорівнює:

$$y = -16{,}06 + 0{,}415 \cdot 181{,}7 - 0{,}236 * 8{,}8 + 4{,}831 * 6{,}5 = \$88{,}7 \text{ тис.}$$

Звідси можна відповідно виразу (6.81) зробити оцінки еластичностей:

$$E_1 = 0{,}415 \cdot 181{,}7 / 88{,}7 \approx 0{,}85;$$

$$E_2 = -0{,}236 \cdot 8{,}8 / 88{,}7 \approx -0{,}02;$$

$$E_3 = 4{,}831 \cdot 6{,}5 / 88{,}7 \approx 0{,}35$$

Висновок

Оцінки еластичностей свідчать про то ще найбільш впливовим показником є площа будинку, впливом віку будинку можна нехтувати. Модель (6.88) може бути використана для проведення оцінок вартості продажу по відомим параметрам нерухомості.

Штучна зміна використовується коли потрібно перевести якісний показник в кількісний. Наприклад потрібно проаналізувати як вплинуло зміна законодавства що проведено наприклад у 2000 році на економічні показники. Для цього ми формуємо додаткову пояснюючу зміну яка до 2000 року дорівнює нулю, а з 2000 року дорівнює одиниці. Або в межах попередньої задачі потрібно оцінити як впливає на вартість будинку з ділянкою наявність басейну. Для цього формується додатковий вхід (четвертий вхід) який дорівнює нулю у випадку відсутності басейну та 1 у протилежному випадку. Така пояснююча зміна має назву уявної (dummy).

Приклад 18

У юридичної фірмі працює 15 юристів (6 жінок, 9 чоловіків), кожний з яких відпрацював за остатній місяць деяку кількість годин та отримав за це деяку суму коштів (табл. 6.26). Існує лі ризниця в оплаті праці по гендерної ознаці?

Табл.7.11. Вихідні характеристики задачі оплати праці по гендерному признаку

№ (j)	Кількість відпрацьованих годин	Стать (уявна зміна)	Оплата праці тис. грн.	$D(j)x(j)$
1	200	ж (0)	20	0
2	205	ж (0)	21	0
3	190	ж (0)	18	0
4	190	ж(0)	20	0
5	250	ж (0)	24	0
6	270	ж (0)	28	0
7	190	ч (1)	23	190
8	150	ч (1)	18	150
9	300	ч (1)	35	300
10	250	ч (1)	31	250
11	170	ч (1)	20	170
12	180	ч (1)	22	180
13	180	ч (1)	23	180
14	210	ч (1)	24	210
15	220	ч (1)	25	220

Розв’язок

Нехай y(j) -оплата праці, x(j) –кількість відпрацьованих годин, D(j)-уявна змінна. Шукаємо залежність оплати праці від кількості відпрацьованих годин у наступному вигляді:

$$y(j) = \beta_0 + (\beta_1 + \beta_2 \cdot D(j))x(j) + \varepsilon(j) \qquad (7.35)$$

де β_1 – маргінальна погодинна оплата праці жінок, β_2 – різниця в оплаті праці чоловіків та жінок. Розкрив дужки правої частини (6.89) отримаємо:

$$y(j) = \beta_0 + \beta_1 x_1(j) + \beta_2 \cdot x_2(j) + \varepsilon(j)$$
$$x_1(j) = x(j); x_2(j) = D(j)x(j) \qquad (7.36)$$

В результаті розрахунків отримано наступни параметри адекватності моделі (табл.7.12).

Таблиця 7.12. Параметри адекватності моделі Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	306,415	2	153,207	162,43	0,0000
Residual	11,3188	12	0,943231		
Total (Corr.)	317,733	14			

R-squared = 96,4377 percent

R-squared (adjusted for d.f.) = 95,8439 percent

Standard Error of Est. = 0,971201

Mean absolute error = 0,749859

Durbin-Watson statistic = 2,57414 (P=0,8025)

Lag 1 residual autocorrelation = -0,352412

Оцінки регресійних коефіцієнтів та їх похибок представлено у табл.7.13.

Таблиця 7.13. Оцінка регресійних коефіцієнтів та їх похибок

		Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
CONSTANT	0,82981	1,36599	0,607477	0,5549
Col_2	0,096684	0,00641404	15,0738	0,0000
Col_3	0,0186567	0,00238184	7,8329	0,0000

В результаті розрахунків отримано наступне рівняння залежності оплати праці від відпрцьованих годин та стати:

$$\hat{y}(i) = 0{,}83 + 0{,}097x_1 + 0{,}019x_2$$

Кожний з регресійних коефіцієнтів значимий на рівні значущості меньш ніж 0,0001. Це означає що середня оплата праці для жинок складає 97 грн. за годину, а для чоловіків на 19 грн більше, тобто чоловики дийсно заробляють приблизно на 20% більш ніж жінки.

7.5.Приведено значення коефіцієнту детермінації

При здійсненні розрахунків за допомогою програми Statgraphics крім коефіцієнту детермінації розраховується приведений на кількість входів коефіцієнт детермінації. Справа в тому що зростання коефіцієнту детермінації може відбувається за рахунок входів які не мають ніякого відношення до досліджує мого процесу. Наприклад для будь якого досліджує мого явища додатковим входом, що сприяє зростанню коефіцієнту детермінації може слугувати білий шум. Для обмеження кількості входів, які мають незначний вплив на залежну зміну використовується приведений на кількість входів коефіцієнт детермінації (adjusted for d.f. – табл. 6.27), який для множинної регресії завжди менш ніж звичайний коефіцієнт детермінації. Приведений коефіцієнт детермінації позначається як $\overline{R}^2$ та визначається:

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2 /(n-(k+1))}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 /(n-1)} \qquad (7.37)$$

Приведений коефіцієнт детермінації визначається через звичайний наступним шляхом:

$$\overline{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-(k+1)} \qquad (7.38)$$

Приклад 19

Розрахувати приведений коефіцієнт детермінації для прикладу 18.

Розв’язок

Вважаємо відомим коефіцієнт детермінації з попереднього прикладу - 0,971. Підставляємо у (6.94) n=15, k=2:

$$\overline{R}^2 = 1 - \frac{(1-0{,}964)14}{12} \approx 0{,}971$$

7.6. Кореляційна матриця та часткова кореляція

Для оцінки щільності лінійного взаємозв'язку між змінними нами у попередніх розділах було впроваджено оцінку вибіркового коефіцієнту кореляції. Якщо зміни впливають одна на то за допомогою звичайної кореляції важко оцінити вплив наприклад кожної з них на залежну зміну. Тому використовується коефіцієнт часткової кореляції. Для введення поняття коефіцієнту часткової кореляції визначимо поняття кореляційної матриці елементи якої є звичайні корреляційни коефіцієнти змінних $x_1, x_2, ..., x_k$:

$$r_{ij} = \begin{pmatrix} 1...r_{12}.....r_{1k} \\ r_{21}..1.....r_{2k} \\ ..... \\ r_{k1}...r_{k2}.....1 \end{pmatrix} \qquad (7.39)$$

Кореляційна матриця симетрична, а в якості однієї змінної може використовуватись залежна зміна

Вибірковий коефіцієнт часткової кореляції між змінними $x_i; x_j$ при фіксованих значеннях інших к-2 змінних оцінюється наступним щляхом:

$$r_{ij/1,2,...k} = \frac{-q_{ij}}{q_{ii}q_{jj}} \qquad (7.40)$$

де q_{ij}, q_{ii}, q_{jj} алгебраїчні доповнення елементів r_{ij}, r_{ii}, r_{jj} кореляційної матриці. Слід підкреслити, що оцінка рівня значущості для часткової кореляції аналогічна оцінки для звичайної кореляції.

У випадку трьох змінних (к=3) частковий коефіцієнт кореляції між першою та третьою змінними дорівнює:

$$r_{13/2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} \qquad (7.41)$$

Приклад 20

Кореляційна матриця для трех випадкових зміних має наступний вигляд (табл.). Побудувати та проаналізувати матрицю часткових кореляцій.

$$r_{ij} = \begin{pmatrix} 1...0,6...0,5 \\ 0,6..1....0,7 \\ 0,5..0,7...1 \end{pmatrix}$$

Розв'язок

$$r'_{12} = \frac{0,6 - 0,5 \cdot 0,7}{\sqrt{(1-0,25)(1-0,49)}} \approx 0,4; r'_{13} = \frac{0,5 - 0,6 \cdot 0,7}{\sqrt{(1-0,36)(1-0,49)}} \approx 0,14; r'_{23} = \frac{0,7 - 0,6 \cdot 0,5}{\sqrt{(1-0,25)(1-0,36)}} = 0,58$$

$$r'_{ij} = \begin{pmatrix} 1...0,4...0,14 \\ 0,4..1....0,58 \\ 0,14..0,58...1 \end{pmatrix}$$

Всі значення кореляційних коефіцієнтів суттєво зменшились. Нехай рівень значимості 5% тоді критичне значення коефіцієнта кореляції дорівнює 0,2. Це означає, що перший та третій процеси не можна вважати взаємозалежнимі, тому що ця залежність обумовлено взаємозками з другим процесом.

7.7 Використання лагових змінних у регресійному аналізі

Для України як країни з відкритою економікою надзвичайно суттєвим є вплив зовнішніх факторів на економічні показники в середині країни. Наприклад, світова фінансова криза 2008-2009 року суттево вплинула на українську економіку, крім інших причин, завдяки значному зростанню вартості кредитних ресурсів яки зарубіжни фінансові інституції представляли

для української економіки. Тобто, якщо розглядати вартість фінансових ресурсів як незалежну зміну –x(t), а темпи економічного розвитку як залежну-y(t), то з початку суттєво зросла вартість ресурсів (від 7% до 30%), а потім з лагом τ приблизно півроку почалося падіння темпів економічного развитку. Регресійну залежність між темпами економічного розвитку та вартістю фінансових ресурсів можна подати у вигляді:

$$y(t) = \beta_0 + \beta_1 x(t-\tau) + \varepsilon(t) \qquad (7.42)$$

Величина зсуву визначається за допомогою максимуму взаємно кореляційної функції [5]:

$$R_{xy}(\tau) = \frac{\sum_{t=1}^{T-\tau}\left(x(t)-\bar{x}\right)\left(y(t+\tau)-\bar{y}\right)}{T\sigma_x \cdot \sigma_y} \qquad (7.43)$$

На відміну від автокореляційної функції взаємно-кореляційна не є парною. Якщо ми шукаємо взаємно кореляційну функцію у вигляді (3.98), то вважається, що зміна x(t) впливає на зміну y(t) з деяким лагом, що визначається в процесі досліджень по максимальному значенню $\max R_{xy}(\tau)$. Однак цілком можливо що це припущення є хибним – максимальне значення досягається при $\tau < 0$. Тоді слід поміняти змінні x(t) та y(t) і повторити процедуру. Програма «Statgraphics» реалізує процедуру розрахунку взаємно-кореляційної функцій як для додатних так і для від'ємних τ.

Приклад 21

Процеси x(t) та y(t) задано в таблиці. Знайти зсув процесу y(t) відносно процесу x(t) та побудувати регресійну модель залжності y(t) від x(t)

T	1	2	3	4	5	6	7	8	9	10
x(t)	0	3	4	6	3	1	0	-2	-4	-5
y(t)	1	2	5	8	11	13	9	7	5	1

Рішення

Відповідно виразу (6.99) знайдемо середні квадратичні відхилення. Знайдемо значення взаємно-кореляційної функції при $\tau = -3,-2,-1,0,1,2,3$ або

використовуємо програму «Statgraphics» для побудови взаємно-кореляційну функцію процесів. рис 6.10. Максимум взаємно-кореляційної функції спостерігається при запізненні y(t) відносно x(t) на 2 часових лага.

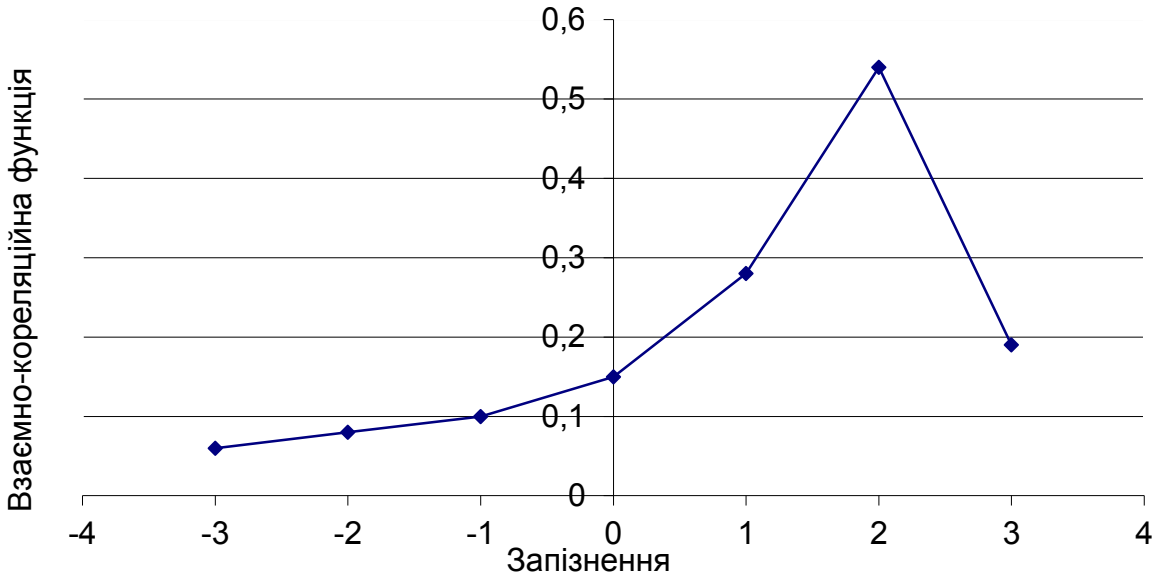


Рис. 7.1. Графік взаємно кореляційної функції

В цьому випадку регресійне рівняння має наступний вигляд:

$$y(t) = \beta_0 + \beta_1 x(t-2) + \varepsilon(t) \qquad (7.44)$$

Звідсі слідує що для того щоб зробити оцінку МНК регресійних коефіцієнтів наведеного рівняння потрібно відкінути два перших члена ряду y(t) і два останних члену ряду x(t).

В результаті розрахунків отримано наступню таблицю регресійних коефіцієнтів та їх похибок.

Табл. 7.14. Таблиця регресійних коефіцієнтів та їх похибок

	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept	4,68	0,318055	14,7147	0,0000
Slope	1,43733	0,103875	13,8372	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	96,8403	1	96,8403	191,47	0,0000
Residual	3,03467	6	0,505778		
Total (Corr.)	99,875	7			

Correlation Coefficient = 0,98469

R-squared = 96,9615 percent

R-squared (adjusted for d.f.) = 96,4551 percent

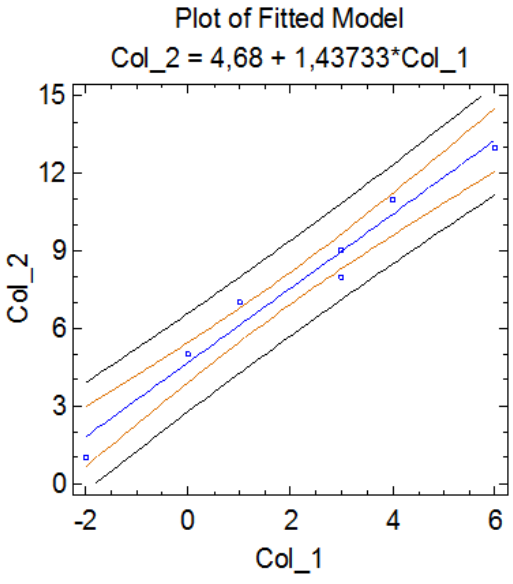
Standard Error of Est. = 0,711181

Mean absolute error = 0,525333

Durbin-Watson statistic = 2,42981 (P=0,6174)

Lag 1 residual autocorrelation = -0,338636

Col_2 = 4,68 + 1,43733*Col_1



Існує інше більш узагальнене визначення нестаціонарності процесів. Для цього необхідно ввести поняття автокореляціїйної функції яка визначається наступним шляхом [5].

$$R(t_0,\tau)=\frac{1}{\sigma^2 T}\times\sum_{t=t_0+1}^{t=t_0+T-\tau}(x(t+\tau)-\bar{x})(x(t)-\bar{x}) \quad (7.45)$$

Вважається що умова стаціонарності виконується, коли $R(t_0,\tau)$ є тільки функція τ. Автокореляційна функція є парною тому розраховується тільки для додатних τ.

На практиці більшість процесів, які досліджуються в економіці є нестаціонарними – зростають практично всі номінальні показники: доходи та витрати населення, рівень заощаджень, компоненти грошових агрегатів та інші. Крім того, ознаками нестаціонарності є збільшення волатильності курсів основних світових валют на час політичних криз. Перевірка на стаціонарність здійснюється за допомогою програмного засобу Econometrics Views (тест Дики-Фуллера [6]). Приведення до стаціонарності досліджуваних часових рядів виконується шляхом використання оператору першої різниці $\Delta x(t)=x(t)-x(t-1)$. Якщо цієї процедури достатньо, щоб зробити початковий ряд стаціонарним, то він має назву інтегрованого ряду першого порядку [10]. Якщо потрібно повторити процедуру 2 рази, то маємо інтегрований ряд другого порядку. При побудові регресійних моделей яки залежать як від пояснюючої зміної так і від попередних значень пояснюємої зміної регресійна залежність шукається у наступному вигляді:

$$y(t)=\beta_0+\sum_{j=1}^{m}\alpha_j y(t-j)+\sum_{i=0}^{k}\beta_i x(t-i)+\varepsilon(t) \quad (7.46)$$

де m-порядок авторегресії, к- порядок запізнення пояснюючої зміної.

7.8. Порушення класичних положень, узагальнений метод найменших квадратів

Наведемо положення на яких базується оцінка регресійних коефіцієнтів:

$$\begin{aligned} &1) E(\varepsilon_i) = 0 \\ &2) \operatorname{var}(\varepsilon_i) = \sigma^2 \\ &3) E(\varepsilon_i \varepsilon_j) = 0, i \neq j \\ &4) E(x_i \varepsilon_i) = 0 \end{aligned} \qquad (7.47)$$

Випадок коли дисперсія похибки не є сталою величиною називається хетероскедатичностю. Хетероскедатичність не впливає на оцінку регресійних коефіцієнтів, однак призводить до зростання дисперсії цих оцінок.

У випадку порушень наведених положень використовується узагальнений метод найменьших квадратів. Якщо у випадку виконання (6.103) коваріаційна матриця похибок $\sigma^2 I_m$ то у випадку невиконання умовов вона перетворюється у $\sigma^2\Omega$. Вважається що для будя якої позитивно визначено ї матриці Ω існує несингулярна матриця P така що $P \cdot P' = \Omega$. Перетворюємо початкову модель

$$y = X\beta + \varepsilon$$

Шляхом множення обох частин на P^{-1}ми отримаємо:

$$P^{-1}y = P^{-1}X\beta + P^{-1}\varepsilon$$

Позначимо зміни, що множаться на P^{-1}зірками (*):

$$y^* = X^*\beta + \varepsilon^*$$

Звідси несмещена оцінка для цього випадку:

$$\hat{\beta}_{BLUE} = (X*'X)^{-1}X*'y* = (X'\Omega X)^{-1}X\Omega^{-1}y$$

Дісперсія оцінки:

$$\text{var}(\widehat{\beta}_{BLUE}) = \sigma^2 (X *' X)^{-1} = \sigma^2 (X' \Omega X)^{-1}$$

Завдання для теми 7.

По даним світової статистики знайти залежність основного показнику економічного розвитку від від індексу спріяття корупції, індексу розвитку демократії та індексу ефективності уряду. Проаналізовати отриману залежність (задача може бути замінено на іншу множінну регресію, що відповідає теме дісертаційного дослідження).

ТЕМА 8. Оптимізаційні рішення

8.1. Неоклассична теорія фірми

Неокласична теорія фірми побудована на припущенні, що ціль фірми полягає в максимізації прибутку. Вважається, що випускається тільки один від продукції ціною р. Шляхом вибору обсягів факторів виробництва $\bar{x} = (x_1, x_2, ..., x_n)$, при заданій виробничій функції, заданої ціні випуску р і цінах факторів виробництва $\bar{w} = (w_1, w_2, ..., w_n)$ максимізується прибуток П, що дорівнює річному доходу R за вирахуванням витрат виробництва С:

П=R-C,

де річний дохід обчислюється як річної обсяг продукція, помножений на ціну випуску:

$$R = pq = pf(\bar{x})$$

Витрати виробництва дорівнюють загальним виплатам за всі види витрат:

$$C = \sum_{j=1}^{n} w_j x_j = \bar{w} \cdot \bar{x}$$

Вирішуючи довгострокову задач, фірма вільна вибирати любий вектор витрат з простору витрат, тому завдання формулюється таким чином:

$\max_x$ П(x)= $pf(\bar{x}) - \bar{w} \cdot \bar{x}$ при умові $x_j \geq 0, j = 1,2,...,n$,

або в розгорнутій формі:

$$\max_{x1,x2,...xn} П(x_1, x_2, ...x_n) = pf(x_1, x_2, ...x_n) - \sum_{j=1}^{n} w_j x_j \qquad (8.1)$$

при умові

$x_1 \geq 0, x_2 \geq 0, ..., x_n \geq 0$.

Ця задача є завданням нелінійного програмування, у якій як інструментальні змінні виступає вектор обсягів факторів виробництва $\bar{x}$, цільова функція виражається функцією прибутку П(x); єдиним обмеженням є умова не відємності компонент $\bar{x}$ і цінових показників ((n+1) параметрів $\bar{w}$ і р.

В протилежність довгостроковому завданню, для якого характерно, що всі витрати можна довільно варіювати, при короткостроковій з'являються обмеження на вибір витрат, як, наприклад, знижені ліміти на певні витрати із-за договірних зобов'язань. У короткостроковому завданні фірма повинна вибрати вектор витрат із заданої підмножини простору витрат, отже додається ряд обмежень:

$$g_i(x_1, x_2, ..., x_n) \leq b_i \quad i = 1,2,...,m, \qquad (8.2)$$

де ці m нерівностей виражають обмеження на витрати для певного короткострокового періоду.

В умовах довгостроковості необхідними умовами для максимізації прибутку є умови Куна-Таккера при одному ресурсі:

$$\frac{\partial П}{\partial x} = p\frac{\partial f}{\partial x}(x) - w \leq 0, \frac{\partial П}{\partial x} \cdot x = (p\frac{\partial f}{\partial x}(x) - w)x = 0 \quad x \geq 0.$$

Таким чином, для n ресурсів:

$$p\frac{\partial f}{\partial x_j}(x) \leq w_j, j = 1,2,...,n$$

і

$$\begin{cases} p\dfrac{\partial f}{\partial x_j}(\bar{x}) = w_j, x_j > 0 \\ p\dfrac{\partial f}{\partial x_j}(\bar{x}) < w_j, x_j = 0 \end{cases} \quad j = 1,2,...,n, \tag{8.3}$$

де $p\dfrac{\partial f(\bar{x})}{\partial(x_j)}$ є вартістю граничного продукту в точці $\bar{x}$, тобто вартість додаткового випуску, отриманого при використанні одиниці додаткового ресурсу j-го виду.

Передбачимо, що всі ресурси були дійсно використані ($x_j \rhd 0$), тоді умови першого порядку будуть мати вигляд:

$$p\frac{\partial f}{\partial x_j}(\bar{x}) = w_j$$

Тобто вартість граничних продуктів дорівнює платі за витрати чинників виробництва. Умови першого порядку

$$\psi_j(x) = p\frac{\partial f}{\partial x_j}(x) - w_j = 0,\ j = 1,2,...,n$$

Остання умова означає що маржінальна ефективність використання j ресурсу у грошовому виразі дорівнює ціні цього ресурсу.

8.2. Оптімізаційна задача розподілу часу

Розглянемо оптімізаційну задача розподілу щодобової праці, що вирішує кожна людина за допомогою функції власного добробуту. В даному випадку функція добробуту будується по надзвичайно простому принципу: кожна людина бажає більше заробляти та менш працювати. Оптимізація здійснюється по відношенню к добовому розподілу часу якій поділюється на час необхідний до обслуговання организму -12 год., робочій час –t, відпочинок -τ :

$$24 = t + \tau + 12 \Rightarrow t + \tau = 12$$

Погодинна оплата праці –w, тоді доход за добу дорівнює y=wt. Якщо прийняти для функції особистого добробуту форму виробничої функції, а еластичність по доходу вважати рівної α, а по вільному часу β то фунцію особистого добробуту -z можна подати у вигляді:

$$z(t,\tau) = y^{\alpha}\tau^{\beta} = (tw)^{\alpha}\tau^{\beta} \Rightarrow \max$$

Функція Лагранжу що включає цільову функцію та обмеження:

$$L(t,\tau,\lambda) = z(t,\tau) + \lambda(12 - t - \tau)$$

Потрібно розвязати наступну систему рівнянь:

$$\frac{\partial L}{\partial t} = 0 \Rightarrow \alpha(tw)^{\alpha-1} w\tau^{\beta} = 1$$
$$\frac{\partial L}{\partial \tau} = 0 \Rightarrow \beta(tw)^{\alpha}\tau^{\beta-1} = 1$$
$$\frac{\partial L}{\partial \lambda} = 0 \Rightarrow t + \tau = 12$$

Поділімо перше рівняння на друге: $\tau = \frac{t\beta}{\alpha}$ та підставі в останнє рівняння:

$$t(1 + \beta/\alpha) = 12$$

Враховуючі низький рівень життя більшості українцев вважаємо показник еластичності по доходу вдвічи більш ніж показник еластичності по вільному часу, крім того вважаємо сталий відгук масштабу $\alpha + \beta = 1 \Rightarrow \alpha = 2/3; \beta = 1/3$.

В цьому випадку тривалість робочого часу дорівнює 8 годинам, відапочинку 4 годинам.

Аграрний сектор економіки, як і будь-яка інша галузь економіки, має свої особливості. Однією з головних особливостей аграрного виробництва є значний ступінь ризику, на який наражаються інвестиції в аграрні інновації. Однак перспективи розвитку аграрного ринку та стабільне підвищення попиту на аграрну продукцію, що спостерігається в останні роки, робить інвестування в цю галузь досить привабливим. При оцінці ризику інвестування в аграрне виробництво розглядають питання змін природно-кліматичних умов, політичної та макроекономічної стабільності, наявності інфраструктури,

якості трудового потенціалу, ефективності державного регулювання та інших факторів [5, 6]. Актуальними є шляхи зменшення погодних ризиків, що впливають як на показники урожайності, так і на прибутковість через варіативність цінового та виробничого факторів [5]. Серед можливих варіантів вирішення проблеми найбільшу увагу доцільно приділити шляхам диверсифікації аграрного виробництва [6].

Однак специфічні риси аграрного бізнесу у сукупності з незавершеністю ринкових перетворень у нашої країні створюють особливо несприятливі умови, звичайно, для аграрного сектору. Основною складовою аграрних ризиків залишається ризик урожайності, який значною мірою обумовлено низьким рівнем капіталізації аграрного виробництва. Тому актуальними є шляхи зменшення ризиків зміни погодних умов, що впливають як на показники урожайності, так і прибутковості, шляхом варіативності цінового фактору.

Метою представленої роботи є дослідження можливостей диверсифікації стратегії підприємства при врахуванні як очікуваного показнику прибутковості для кожної з культур, так і показнику ступеню ризику.

8.3 Стандартна оптимізаційна задачу лінійного програмування

Спочатку розглянемо стандартну оптимізаційну задачу лінійного програмування в сукупності з оптимальним планом виробництва зернових.

Нехай цільова функція - це прибуток, та існує два обмеження: на площу та бюджетне обмеження на витрати. Нехай існує N культур, для яких потрібно визначити площі $x_1, x_2, \dots, x_N$, що максимізують прибуток, при цьому відомі наступні характеристики культур: прибуток з одного гектару - $c_1, c_2, \dots, c_N$, витрати на один гектар $b_1, b_2, \dots, b_N$. Загальна величина витрат не повинна перевищувати B (бюджетне обмеження), а загальна площа - S. Останнє обмеження у ринкових умовах не є обов'язковим, землю можна орендувати в необмежних обсягах, якщо дозволяє бюджетне обмеження. Позначимо

цільову функцію $w(x_1, x_2, \dots, x_N)$. Тоді стандартна оптимізаційна задача лінійного програмування при наявності двох обмежень [4]:

$$w = \sum_{i=1}^{N} c_i x_i \Rightarrow max$$

$$\sum_{i=1}^{N} b_i x_i \leq B \qquad (8.4)$$

$$\sum_{i=1}^{N} x_i \leq S$$

Неважко довести, що ця задача має монокультурне рішення. Тобто, щоб отримати максимальний прибуток, потрібно вирощувати тільки одну культуру, яка визначається умовою максимальної рентабельності: $\max(c_i/b_i)$. Площа в цьому випадку визначається на підставі бюджетного обмеження:

$$x_i^* = \min(B \,/\, b_i; S) \qquad (8.5)$$

В цьому випадку значення цільової функції $\max(w) = c_i x_i^*$.

Так для обох господарств такої монокультурою є соняшник, для якого показник рентабельності належить проміжку від 80 до 90 %.

Щоб ввести інші культури використовують штучні обмеження, які не мають нічого спільного з максимізацією прибутку.

Однак в такій постановці не враховуються ризики аграрного виробництва, які виникають як внаслідок непередбачених погодних умов, так і внаслідок нестабільності цін як на фактори виробництва, так і на аграрну продукцію.

8.4. Нелінійна оптімізаційна задача

Середі багатьох показників ризику, які використовуються в економіці, найбільш поширеним є фактор дисперсії прибутку, наявність якого дозволяє оцінити ймовірність збитків, які є небезпечними для аграрного виробництва.

Нехай додатково до умов задачі **Ошибка! Источник ссылки не найден.** відомими є оцінки дисперсії прибутків з 1 га для кожної з культур:

$\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2$. Тоді дисперсія прибутку підприємства за умовою незалежності прибутків від окремих культур [3]:

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2 x_i^2 \tag{8.6}$$

Для того щоб додати нелінійне обмеження на ступінь ризику (дисперсію), потрібно визначити кількісно її максимальне значення. Для цього використовується коефіцієнт варіації загального прибутку: $V = \sigma / \overline{w}$. Якщо базуватись на функції нормального розподілу прибутку, то при $V = 0{,}1$ ймовірність збитків практично дорівнює нулю, тоді як при $V = 0{,}6$ ймовірність збитків досягає 5%, тобто при відомій величині прибутку підприємства ми задаємо діапазон можливих дисперсій - $\sigma^2(V)$, що відповідають зростанню коефіцієнту варіації від 0,1 до 0,6 з кроком 0,1. Тобто ми маємо наступний алгоритм завдання дисперсії прибутку на підставі очікуваної величини прибутку та заданої величини коефіцієнту варіації:

$$\sigma^2 = V^2 \overline{w}^2, V = 0{,}1; 0{,}2; \dots, 0{,}6$$

Постановка задачі нелінійної оптимізації (максимізації прибутку) з урахуванням ступеню ризику має наступний вигляд:

$$\begin{gathered} w = \sum_{i=1}^{N} c_i x_i \Rightarrow max \\ \sum_{i=1}^{N} b_i x_i \leq B \\ \sum_{i=1}^{N} x_i \leq S \\ \sum_{i=1}^{N} \sigma^2{}_i x^2{}_i \leq \sigma^2(V) \end{gathered} \tag{8.7}$$

В нашому розпорядженні були дані відносно урожайності, номінальних цін на фактори виробництва та готову продукцію для двох господарств за 2004-2013 роки (2014 рік невключений до розгляду внаслідок

макроекономічної нестабільності). На першому етапі ми привели всі ціни до цін 2013 рік, щоб зменшити варіативність показників, що обумовлено інфляційними процесами.

Обидва господарства культивують однакову продукцію рослинництва: пшеницю (1), кукурудзу (2), соняшник (3), ячмінь (не розглядається), сою (4). Внаслідок низької рентабельності ячменя оптимальним рішенням для будь-якої дисперсії є виключення цієї культури (нульова площа), і до розгляду у подальшому не приймається. В результаті проведених розрахунків ми отримали близькі значення для доходів та витрат для обох господарств. Тому в цільовій функції та в лівих частинах обмежень задаються однакові коефіцієнти. Оскільки підприємства відрізняються масштабами виробництва, то праві частини обмежень (бюджетне обмеження, площа, дисперсія задаються окремо). Наведемо наступні коефіцієнти для цільової функції та обмежень (тис. грн.) у векторній формі [4]:

$$\bar{c}(3{,}6;\ 4{,}9;\ 8{,}0;\ 4{,}1)$$

$$\bar{b}(7{,}9;\ 9{,}4;\ 8{,}9;\ 6{,}4)$$

$$\sigma^2(5{,}0;\ 10;\ 440;\ 70)$$

$$S_1 \leq 3000 \text{ га.};\ S_2 \leq 1400 \text{ га.};$$

$$B_1 \leq 25 \text{ млн. грн.}; B_2 \leq 10 \text{ млн. грн.}$$

Значення дисперсії сумарного прибутку задаються у 6 варіантах для кожного господарства відповідно заданого раніше алгоритму. Для першого більшого по витратам та масштабам виробництва ми отримали наступні рішення, що максимізують прибуток при заданих обмеженнях (табл.8.1). Крім звичайних показників, які використовуються в аграрному виробництві: очікуваний прибуток, витрати, рентабельність, варіація прибутку, нами використовується прибуток на рівні значимості 5 %. Такий показник з метою зменшення ступеня ризику прийняття управлінських рішень використовується у банківській та страховій справі. Він дозволяє отримати нижню межу прибутку, вихід за яку має ймовірність 5%. У випадку аграрного виробництва

це відповідає 1 випадку на 20 років. Прибуток на рівні значимості 5% розраховується наступним шляхом:

$$w_{0,05} = w - 1{,}64\sigma$$

Вважаємо, що цей показник ризику цілком припустимий для аграрного виробництва. Всі дані для розрахунку представлено у (табл.8.1). Як і слідує з теоретичних положень максимальні показники рентабельності та прибутку досягаються при максимальному показнику ризику (дисперсія).

Однак така стратегія не є оптимальною, тому що суттєво підвищує ймовірність збитків. На наш погляд, оптимальним є план, що максимізує прибуток на рівні значимості 5%. (рис.8.1) В цьому випадку практично гарантований прибуток повинен бути не менший, ніж 3,2 млн. грн.. Найбільші площі в цьому випадку виділяються під кукурудзу та пшеницю. Слід підкреслити, що навіть на рівні значимості 5% рентабельність плану перевищує 25%.

Розглянемо можливі варіанти структури посівів для другого господарства. В цьому випадку показник ризику залишається на попередньому рівні, а бюджетне обмеження зменшується до 10млн. грн. Тобто розглядається варіант можливості господарювання при жорсткому бюджетному обмеженні та при значному ризику. Жорстке бюджетне обмеження може спонукати керівництво підприємства до шляху швидкого виходу з важкого фінансового стану за рахунок культивації культур, що мають очікувану високу прибутковість.

Табл.8.1. Ризики структура площ та показники прибутковості для першого господарства

Дисперсія $(\boldsymbol{\sigma}, \text{млн. грн.})$	$\mathbf{0,5 \cdot 10^7}$ $\mathbf{(2,24)}$	$\mathbf{10^7}$ $\mathbf{(3,16)}$	$\mathbf{2 \cdot 10^7}$ $\mathbf{(4,47)}$	$\mathbf{3 \cdot 10^7}$ $\mathbf{(5,48)}$	$\mathbf{4 \cdot 10^7}$ $\mathbf{(6,32)}$	$\mathbf{5 \cdot 10^7}$ $\mathbf{(7,07)}$
Структура Площа (га)	(694;472; 18;56)	(982;668; 25;80)	(1388;945; 35;113)	(1167;1334; 77;199)	(790;1610; 115;271)	(524;1805; 141;322)

	(1240)	(1795)	(2481)	(2777)	2786	(2792)
Прибуток (млн. грн.)	5,2	7,3	10,4	12,2	12,8	13,2
Витрати (млн. грн.)	10,4	14,8	20,9	23,7	24,1	24,4
Рентабельність %	49,7	49,4	49,8	51,3	53,0	54,0
$w_{0,05}$*(млн.грн.)*	1,5	2,1	3,1	3,2	2,4	1,6
Варіація прибутку (%)	0,43	0,43	0,43	0,45	0,49	0,54

Джерело: [4]

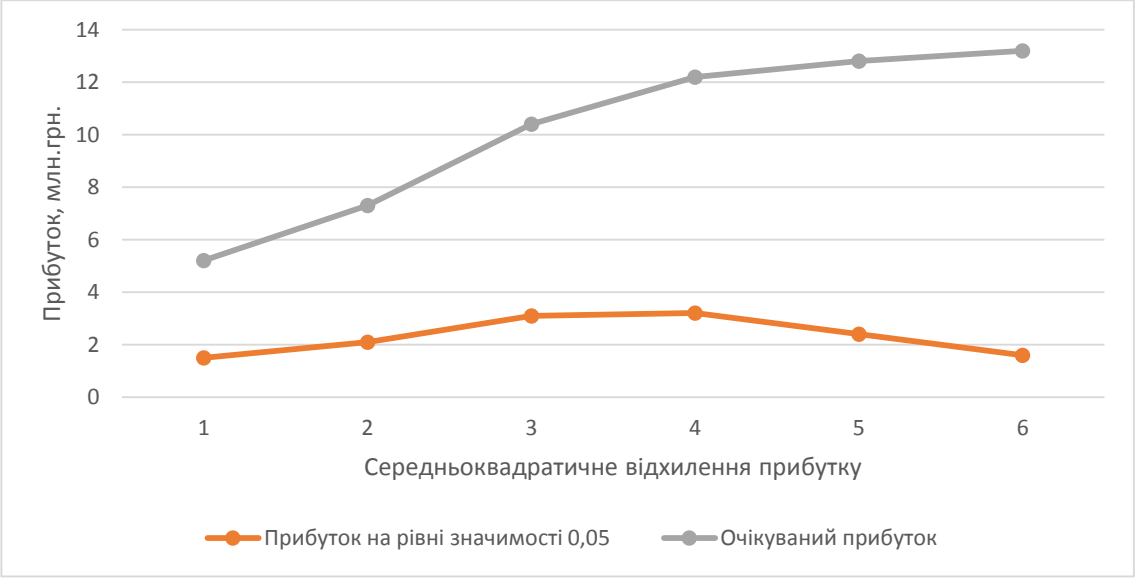


Рис. 8.1. Залежність прибутку від ступеня ризику

Однак цей шлях має надзвичайно високу ступінь ризику. Із зростанням ризику внаслідок переходу до більш рентабельних культур, прибуток зростає не так швидко, як його дисперсія (Табл. 8.), а це суттєво збільшує ймовірність збитків для підприємства, що обрало цю стратегію. При значному ступені ризику (коефіцієнт варіації прибутку перевищує 60%), прибуток на рівні значимості 5% стає від'ємним, що свідчить, що підприємство вступило в зону великого ризику.

Табл. 8.2. Показники прибутковості для другого господарства

Дисперсія ($\boldsymbol{\sigma}$, млн. грн.)	$\mathbf{0,5 \cdot 10^7}$ $(\mathbf{2,24})$	$\mathbf{10^7}$ $(\mathbf{3,16})$	$\mathbf{2 \cdot 10^7}$ $(\mathbf{4,47})$	$\mathbf{3 \cdot 10^7}$ $(\mathbf{5,48})$	$\mathbf{4 \cdot 10^7}$ $(\mathbf{6,32})$	$\mathbf{5 \cdot 10^7}$ $(\mathbf{7,07})$
Структура Площа (га)	(443;551; 33;84) (1111)	(108;800; 64;149) (1121)	(0;746; 148;261) (1155)	(0;645; 203;333) (1181)	(0;569; 244;387) (1200)	(0;504; 279;433) (1216)
Прибуток (млн.грн.)	4,9	5,4	5,9	6,1	6,3	6,5
Витрати (млн.грн.)	9,5	9,8	10	10	10	10
Рентабельність %	51,5	55,1	59	61	63	65
$w_{0,05}$*(млн.грн.)*	1,2	0,2	-1,4	-2,9	-4,1	-5,1
Варіація прибутку (%)	0,45	0,57	0,76	0,9	1,0	1,1

Джерело: [4]

Наведені розрахунки показують, чому основна товарна позиція для визначення продовольчої безпеки (пшениця) в ринкових умовах залишається в структурі посівів значної частки аграрних підприємств. Це пов'язано з співвідношенням прибутковості та ризику, що гарантує підприємству стабільні прибутки.

Не тривіальне рішення для оптимізаційної задачі максимізації прибутку аграрного в сукупності із структурою посівів зернових можна отримати шляхом обмеження дисперсії прибутку аграрного підприємства. Якщо не використовувати обмеження на дисперсію та штучні обмеження на площі окремих культур, рішенням, що максимізує прибуток, буде монокультура з максимальним показником рентабельності.

Нами пропонується впровадження замість величини очікуваного прибутку показник прибутку на заданому рівні значимості, що дозволяє враховувати як показники прибутковості, так і показники ризику отримання збитків.

Для аграрного виробництва виконується загальноекономічне співвідношення відносно існування прямої залежності між прибутковістю та ступенем ризику. Однак, якщо використовувати критерій прибутковості на

заданому рівні значимості, залежність перестає бути монотонною та існує величина ризику, що забезпечує максимальну прибутковість на заданому рівні значимості.

Оптимізаційна задача з урахуванням погодних ризиків

Стандартна задача знаходження оптимального розподілу площ базується на середніх погодних умовах, характерних для даного регіону. На наш погляд, кінцевий план повинен будуватись з урахуванням погодних ризиків, тобто не обов'язково максимізувати прибутки підприємства, а, можливо, представляти план, ризики реалізації якого у будь-якому випадку не призведуть до катастрофічних наслідків. Тому до стандартного підходу знаходження оптимального розв'язку при обмежених ресурсах додається матриця впливу погодних умов на рентабельність окремих видів культур, де ρ_{ij} – рентабельність і-тої культури при j-тих погодних умовах $(1 \leq \text{і} \leq n)$, $(1 \leq j \leq m)$, n – кількість культур, а m – кількість погодних умов. Погодні умови можна задати ймовірнісним розподілом $\text{P}(\theta_1), \text{P}(\theta_2), \dots, \text{P}(\theta_m)$. Звичайно, повинна виконуватись умова повноти погодних умов $\sum_j^m P(\theta_j) = 1$. Крім того, відмінністю пропонованого підходу від традиційного використання оптимізаційних задач в аграрному секторі є те, що прибуток від окремих культур буде розраховуватись на підставі витрат на виробництво цих культур та матриці рентабельності, яка залежить від погодних умов. Витрати на виробництво містять фінансове обмеження сільськогосподарського підприємства.

Цільова функція при заданих j-их погодних умовах задається наступним шляхом:

$$W_j = \sum_{i=1}^{n} c_{ij} x_i \qquad (8.8)$$

де індекс j – відповідає погодним умовам θ_j, і – індекс культури, c_{ij} – показник прибутковості в грн./га i культури при j погодних умовах, x_i – площа під i культурою в га, n-кількість культур.

Стандартна матриця обмежень записується в наступному вигляді:

$$\sum_{i=1}^{n} a_{ki} x_i \leq b_k, k = 1,2, \dots, l \qquad (8.9)$$

деякі величина витрат на один га к ресурсу для і культури, b_k – загальна величина к-го ресурсу ($1 \leq k \leq l$), l – кількість використаних ресурсів.

Нехай у подальшому перше обмеження відповідає фінансам, тоді коефіцієнти першої нерівності системи (8.9) - вартості зрощування культур. які виробляються, на 1 га. Крім того, вважається відомою матриця рентабельності розмірністю (nxm), тоді коефіцієнти цільової функції розраховуються на підставі матриці рентабельності та першого рядка обмежень (фінансового обмеження) [2, 3]:

$$c_{ij} = a_{1i} \cdot \rho_{ij} \qquad (8.10)$$

Це означає, що при постановці задачі лінійного програмування при зміні погодних умов змінюється тільки цільова функція, хоча можливі і варіанти зміни обмежень, коли керівник приймає рішення, що базується на довгостроковому погодному прогнозі відносно нестандартного методу вирощування культур або взагалі відносно змін їх номенклатури.

Розглянемо рішення задачі лінійного програмування для підприємства з урахуванням погодних ризиків як вже згадувалось раніше.

Оптимальне рішення при погодних умовах θ_j позначимо $\overline{x_j}(x_{1j}, x_{1j} \dots . x_{nj})$, відповідно прибуток від плану $\overline{x_j}$ позначимо:

$$W_{ij} = W_j(\overline{x_j}) \qquad (8.11)$$

Знайдемо математичне очікування прибутку плану $\overline{x_j}$. Якщо при погодних умовах θ_lбуде реалізовано план $\overline{x_j}$, то прибуток дорівнює:

$$W_{ij} = W_l(\overline{x_j}) \qquad (8.12)$$

Тоді математичне очікування прибутку при реалізації плану $\overline{x_j}$ буде дорівнювати:

$$E_j = \sum_{l=1}^{n} P(\theta_l) W_{lj} \tag{8.13}$$

Крім того, розрахуємо показник ступеня ризику – дисперсію прибутку, що відповідає плану $\overline{x_j}$:

$$\sigma_j^2 = \sum_{l=1}^{n} P\,(\theta_l) W_{lj}^2 - E_j^2 \tag{8.14}$$

Оскільки існують m можливих погодних умов, то існує і m оптимальних планів.

Вибір кінцевого рішення повинен базуватися не тільки на математичних очікуваннях планів при різних погодних умовах, але і на показниках їх похибок внаслідок нестабільних погодних умов. До цих показників у першу чергу відноситься дисперсія, що відповідає рішенням $\overline{x_1}$, $\overline{x_2}$, ..., $\overline{x_n}$ m. Крім того, визначені характеристики дають можливість оцінити імовірність збитків при кожному з прийнятих варіантів рішень.

Якщо вважати, що величина прибутку підпорядковується нормальному розподілу, то імовірність того, що підприємство отримає збитки розраховується відповідно:

$$P_n\big(W_j < 0\big) = 1 - F\left(\frac{E_j}{\sigma_j}\right) \tag{8.15}$$

де F(x) – функція Лапласа $F(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-t^2} dt$

Якщо ж розподіл не є відомим, використовується нерівність Чебишова:

$$P_{\text{ч}}\big(W_j \le 0\big) \approx \frac{1}{2} \frac{\sigma_j^2}{E_j^2} \tag{8.16}$$

Кінцеве рішення приймається з урахуванням фінансового та майнового стану підприємства.

Розглянемо цей алгоритм на прикладі. Вихідна інформація в (табл.8.3):

Фінансові резерви – 1млн. грн.

Людські ресурси – 50000 люд/год.

Площа – 200га.

Ймовірність настання певних погодних умов:

$P(\theta_1)$ = 0,3 – добрі;

$P(\theta_2)$= 0,5 – середні;

$P(\theta_3)$= 0,2 – погані.

Табл.8.3 Вихідна інформація для побудови оптимальних рішень

Культура	**Затрати на 1 га**		**рентабельність**		
	Люд/год	Тис грн	θ_1	θ_2	θ_3
Озима пшениця	20	5	0,4	0,2	0,1
Ячмінь	15	3	0,4	0,4	0,05
Картопля	40	6	0,5	0,4	0,2

Джерело: [3]

Нехай $x_1; x_2; x_3 -$ площі під озиму пшеницю, ячмінь та картоплю.

Сформулюємо обмеження за трьома критеріями: обмеження за фінансовими резервами; обмеження за площею; обмеження за людськими ресурсами (Табл.3):

$$\begin{gathered} 5x1 + 3x2 + 6x3 \leq 1000 \\ x1 + x2 + x3 \leq 200 \\ 20x1 + 15x2 + 40x3 \leq 5000 \end{gathered} \quad (8.17)$$

Відповідно до трьох погодних умов (перший рядок обмежень та вираз (8.17) побудуємо цільові функції для кожного з трьох типів погодних умов:

$$\begin{gathered} W_1 = 2\,x_1 + 1{,}2x_2 + 3x_3 \\ W_2 = 1\,x_1 + 1{,}2x_2 + 2{,}4x_3 \\ W_3 = 0{,}5\,x_1 + 0{,}15x_2 + 1{,}2x_3 \end{gathered} \quad (8.18)$$

Кожна з цільових функцій (8.18) та обмежень (8.19) дозволяє знайти оптимальний розв’язок для кожного типу погодних умов. За допомогою звичайного симплекс методу отримаємо наступні рішення:

$$1)\ \bar{x}_1 = \begin{pmatrix} 125 \\ 0 \\ 62{,}5 \end{pmatrix}; \quad (8.19)$$

$$2)\ \bar{x}_2 = \begin{pmatrix} 0 \\ 120 \\ 80 \end{pmatrix};$$

$$3)\ \bar{x}_3 = \begin{pmatrix} 125 \\ 0 \\ 0 \end{pmatrix};$$

Оскільки оптимальний план не обов'язково відповідає погодним умовам, на які він був розрахований, знайдемо очікуване значення цільових функцій (тис. грн.) при кожній з погодних умов (перший індекс відповідає плану, другий погодним умовам):

$$w_{11}(\bar{x}_1) = 437{,}5;\ w_{12}(\bar{x}_1) = 275;\ w_{13}(\bar{x}_1) = 137{,}5$$
$$w_{21}(\bar{x}_2) = 384;\ w_{22}(\bar{x}_2) = 336;\ w_{23}(\bar{x}_2) = 114 \qquad (8.20)$$
$$w_{31}(\bar{x}_3) = 375;\ w_{32}(\bar{x}_3) = 300;\ w_{33}(\bar{x}_3) = 150$$

Зробимо оцінку математичного очікування, дисперсії та коефіцієнту варіації для кожного із оптимальних для даних погодних умов планів:

$E_{ij} = \sum_{j=1}^{n} p_i W_{ji}$;

$E_1 = \sum_{j=1}^{3} p_i W_{1j} = 0{,}3 * 437{,}5 + 0{,}5 * 275 + 0{,}2 *$ $137{,}5 = 296{,}25$;

$E_2 = \sum_{j=1}^{3} p_i W_{2j} = 0{,}3 * 384 + 0{,}5 * 336 + 0{,}2 * 114 =$ 306 ;

$E_3 = \sum_{j=1}^{3} p_i W_{3j} = 0{,}3 * 375 + 0{,}5 * 300 + 0{,}2 * 150 =$ $292{,}5$.

$\sigma_1^2 = \sum_{j=1}^{3} p_i W_{1j} - E_1^2 = 0{,}3 * 437{,}5^2 + 0{,}5 * 275^2 +$ $0{,}2 * 137{,}5^2 - 296{,}25^2 = 106{,}1^2$,

$\sigma_2^2 = \sum_{j=1}^{3} p_i W_{2j} - E_2^2 = 0{,}3 * 384^2 + 0{,}5 * 336^2 +$ $0{,}2 * 114^2 - 306^2 = 98{,}2^2$;

$\sigma_3^2 = \sum_{j=1}^{3} p_i W_{3j} - E_3^2 = 0{,}3 * 375^2 + 0{,}5 * 300^2 +$ $0{,}2 * 150^2 - 292{,}5^2 = 78{,}3^2$;

$E_{ij} = \sum_{j=1}^{n} p_i W_{ji}$;

$$E_1 = \sum_{j=1}^{3} p_i W_{1j} = 0{,}3 * 437{,}5 + 0{,}5 * 275 + 0{,}2 *$$

$137{,}5 = 296{,}25$;

$$E_2 = \sum_{j=1}^{3} p_i W_{2j} = 0{,}3 * 384 + 0{,}5 * 336 + 0{,}2 * 114 =$$

306 ;

$$E_3 = \sum_{j=1}^{3} p_i W_{3j} = 0{,}3 * 375 + 0{,}5 * 300 + 0{,}2 * 150 =$$

$292{,}5$.

Зробимо оцінку ймовірності отримати збитки при послідовному використанні оптимальних планів для різних типів погодних умов. На першому етапі зробимо припущення відносно нормальності розподілу прибутків та оцінимо можливості отримати збитки для кожного з планів:

$$P_n(W_1 < 0) = 1 - f\left(\frac{E_1}{\sigma_1}\right) = 1 - F\left(\frac{296{,}25}{106{,}1}\right) = 1 -$$

$0{,}997 = 0{,}003$,

$$P_n(W_2 < 0) = 1 - f\left(\frac{E_2}{\sigma_2}\right) = 1 - F\left(\frac{306}{98{,}2}\right) = 1 - 0{,}999 =$$

$0{,}001$,

$$P_n(W_3 < 0) = 1 - f\left(\frac{E_3}{\sigma_3}\right) = 1 - F\left(\frac{292{,}5}{78{,}3}\right) = 1 -$$

$0{,}9999 = 0{,}0001$,

Зробимо подібні оцінки без припущення відносно нормальності розподілу прибутків на підставі нерівності Чебишова:

$$P_ч(W_1 < 0) = 0{,}064;\ P_ч(W_2 < 0) = 0{,}052;\ P_ч(W_3 <$$

$0) = 0{,}036$,

Крім того, розрахуємо для кожного з оптимальних планів прибуток на рівні значущості 10% при умові нормальності розподілу:

$$P_{0,1}(j) = E_j - 1{,}28\sigma_j \Rightarrow P_{0,1}(1) = 296{,}25 - 1{,}28 \cdot$$

$106{,}1 = 160{,}4;\ P_{0,1}(2) = 180{,}3;\ P_{0,1}(3) = 192{,}3.$

Зведемо розраховані кількісні показники в табл.8.4.

По наведеним даним можна зробити наступні висновки: на перший погляд оптимальний план для других (середніх) погодних умов має деякі переваги (максимальний очікуваний прибуток), однак відповідно прибутку на

рівні значущості 10% та ймовірності збитків ступінь ризику для третього варіанту є суттєво нижчим.

Табл.8.4. Показники ступеня ризику оптимальних планів

Оптимальний план	**Очікуваний прибуток – E_j (тис. грн.)**	**Прибуток на рівні значущості 10%(тис. грн.)**	**Ймовірність Збитків $P_{\text{п}}(P_{\text{ч}})$ (%)**
$\bar{x}_1$	296,25	160,4	0,3 (6,4)
$\bar{x}_2$	306	180,3	0,1 (5,2)
$\bar{x}_3$	292,5	192,3	0,01 (3,6)

Джерело: [2, 3]

Якщо представити очікувані результати в залежності від прийнятого плану та погодних умов у вигляді матриці (табл.8.5.), то можна використати критерій Вальда (maxmin).

Табл.8.5. Матриця очікуваних прибутків в залежності від прийнятого плану та погодних умов

Оптимальний план	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\theta_3}$
$\bar{x}_1$	437,5	275	137,5
$\bar{x}_2$	384	336	114
$\bar{x}_3$	375	300	150

Джерело: [3]

При використанні критерію Вальда спочатку знаходиться мінімальне значення по кожному з рядків (найгірший варіант розвитку подій), а потім обирається найкращій варіант з найгірших:

$$\min(W_{ij}) = \begin{pmatrix} 137{,}5 \\ 114 \\ 150 \end{pmatrix}; \max(\min(W_{ij})) = 150$$

Тобто при використанні критерію Вальда варіантом з найменшим ризиком є оптимальний план для найгірших погодних умов. Цей план пропонує використання тільки однієї культури (картоплі), для якої вважається, що рентабельність за будь-яких умов не зменшується нижче за 20%.

Отриманий оптимальний план пропонує не повне використання наявних площ завдяки обмеженню на обсяг праці.

Розглянемо питання чи варто орієнтуватися на найгірші погодні умови та що вважати повною групою несумісних погодних умов. Звичайно, що ділення на три групи погодних умов є надзвичайно спрощеним, тому що найгірші погодні умови мають ймовірність 0,2 , що відповідає повторюваності 5 років. Однак надзвичайно жарке літо 2010 року мало оцінюватися як більш рідкісна подія (зі значно більшим терміном повторюваності), тому виникає питання – на скільки градацій потрібно розбивати можливі погодні умови (а точніше, до якої ймовірності несприятливих погодних умов).

Завдання до теми 8

Побудуйте власну функцію корисності в залежності від доходу та часу . Як співвідносяться ваші показники еластичності по доходу та вільному часу.

Тема 9 Приклад порівняння двох виборок

9.1. Порівняння середніх .

На практиці зустрічаються богатьо задач коли потрібно порівняти дани по двом вибіркам, наприклад : якій сорт пшениці дає більшу врожайність, хто отримує більш чоловіки, або жінки та богать інших прикладів.

Перший найбільш розповсюджений метод це порівняння серелних. Він здійснюється наступним шляхом . Стандартним шляхом знаходяться оцінки середнього та математичного очікування по обом виборкам:

$\bar{x}_1; s_1^2; \bar{x}_2; s_2^2$ Якщо кількість спостережень в первої вибірці дорівнює n_1 а в другої n_2 то похибки оцінки середнього для першої

виборці $s_1/\sqrt{n_1}$ для другої $s_2/\sqrt{n_2}$. Нехай у нас мали вибірці ($n_1;n_2<30$)

Наступним кроком розраховуэться параметр:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Розраховується t статистика:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Діапазон відхилення нульової гіпотези (однакові математичні очікування для обох вибірок) на рівні значимості α

$t > t_{\alpha;\nu}, \nu = n_1 + n_2 - 2$

Розглянемо порівняння середних на натупному прикладі:

X_1	35	38	33	45	40	34	36	35		
X_2	44	46	30	38	43	44	38	32	37	38
Z	9	8	-3	-7	3	10	2	-3		

Отримаємо наступні оцінки для середнього та дисперсії для кожної з виборок:

$\bar{x}_1 = 37; \sigma_1^2 = 15{,}4; n_1 = 8$

$\bar{x}_2 = 39; \sigma_2^2 = 28; n_2 = 10$

$s_p^2 = \frac{15{,}4 \cdot 7 + 28 \cdot 9}{16} \approx 22{,}5$

$t = \frac{39 - 37}{\sqrt{\frac{22{,}5}{10} + \frac{22{,}5}{8}}} \approx 0{,}88$

$t_{16;0,2} = 1{,}34$

Це означає неможливість відхилити нульову гіпотезу навіть на рівні значимості 0,2. Тобто у нас відсутні аргументи вважати математичні очікування виборок різнимі.

Якщо обсяг вибірці перевищує 30 спостережень, то алгоритм порівняння середних декілька зпрощується.

9.2. Порівняння середніх привеликої кількості спостережень

Нехай $n_1 = n_2 = 100$, а середньо та дисперсія залишаються, як у попередньому прикладі. В цьому випадку порівняння відбувається з критичнимі значеннями нормального розподілу z_α:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Якщо $z > z_\alpha$, то нульова гіпотеза відхиляється на рівні значимості $\alpha = 0{,}2; 0{,}1; 0{,}05; 0{,}01; 0{,}0001$

9.2

Розглянемо попередній приклад за умовою, що кількість спостережень дорівнює 100 для кожної вибірці:

$$z = \frac{39 - 37}{\sqrt{\frac{28}{100} + \frac{15}{100}}} \approx 3{,}0$$

$$z_{0,01} = 2{,}33$$

Оскилько $z > z_{0,01}$ то нульову гіпотезу можна відхилити на рівне значимості 0,01. Тобто з впевненістю 99% можна вважати що урожайність у другому випадку більш ніж у першому. При цьому слід помятати, що ми суттєво збільшілі кількість спостережень.

Розглянемо, що один варіант порівняння виборок. Вважаємо що в кожної вибірці тільки по 8 спостережен. Побудуємо ряд різниц $Z=X_2-X_1$. Нульва гіпотеза в цьому випадку відповідає рівності нулю математичного очікування ряду z:

$$t = \bar{z} \Big/ s_z / \sqrt{n}$$

Вона відхіляється на рівні значимості α, якщо $t > t_{n-1;\alpha}$

Знайдемо середне значення та дисперсію ряду різниць. Для оцінки дисперсії зручніші заміст звичайного виразу використати еквівалентний:

$$s_z^2 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})}{n-1} = \frac{\sum_{i=1}^{n} z_i^2 - (\sum_{i=1}^{n} z_i)^2 / n}{n-1} = \frac{325 - 19^2/8}{7} = 40$$

$$\sum_1^n z_i^2 = 81 + 64 + 9 + ... + 9 = 325$$

$$\sum_1^n z_i = 9 + 8 - 3 - 7 + ... - 3 = 19 \Rightarrow \bar{z} = 19/8 \approx 2{,}4$$

$$t = \frac{2{,}4}{6{,}33/\sqrt{8}} = 1{,}1$$

$$t_{7;0,2} = 1{,}42$$

$$t < t_{7;0,2}$$

Це означає що ми не в змозі відхилити нульову гіпотезу.

9.3. Порівнняння розподілів

Розглянемо останній можливий варіант порівняння виборок

Табл.1.1. Данні відносно тижневої оплати праці (USD) та статі працівників

№	Зарп.	Стать	№	Зарп.	Стать	№	Зарп.	Стать
1	236	ж	18	490	м	35	337	ж
2	573	м	19	745	ж	36	1406	м
3	660	ж	20	2033	м	37	530	м
4	1005	м	21	391	ж	38	644	м
5	513	м	22	179	ж	39	776	ж
6	188	ж	23	1629	м	40	440	ж
7	252	ж	24	552	ж	41	548	ж
8	200	ж	25	144	ж	42	751	ж
9	469	ж	26	334	ж	43	618	ж
10	191	ж	27	600	ж	44	822	м
11	675	м	28	592	м	45	437	ж
12	392	ж	29	728	м	46	293	ж
13	346	ж	30	125	ж	47	995	м
14	264	ж	31	401	ж	48	446	ж

15	363	ж	32	759	ж	49	1432	м
16	344	ж	33	1342	м	50	901	ж
17	949	м	34	324	ж			

На підставі отриманих даних побудуємо варіаційний ряд (у порядку зростання) зарплат працівників з у казанням статі. Найменьша зарплата 125 $ стає першою, найбільша 2033 останьою (50). Різниця між найбільшою та найменшою складає діапазон зарплат $2033-$125=$1908. Кількісне значення серединного елементу називається медіаною, якщо кількість непарна то медіаною буде кількісне значення серединного елементу, якщо воно парне, як в цьому випадку, то медіаною буде середньо значення 25 та 26 елементу варіаційного ряду, яке дорівнює $521,5 (медіану також назвають 50% процентілем). Крім того в статистиці використаються поняття кванті лей (квантіль містить ¼ елементів вибірці): 25%-перший квантіль (нижній), 50%-другій квантіль (медіана),75%- третій квантіль(верхний). Квантилі щомісячних зарплат подано у табл.1.2.

Табл.1.2. Варіаційний ряд щотижневих зарплат (USD)

№	Зарп.	Стать	Квантілі
1	125	ж	
2	144	ж	
3	179	ж	
…	…	…	
12	334	ж	335,5 -1 квантіль
13	337	ж	25% прцентіль
….			
25	513	м	521,5-медіана
26	530	м	2 квантіль, 50% процентіль
…			

37	745	ж	748-3квантіль,
38	751	ж	75% процентіль
…			
49	1629	м	
50	2033	м	

Відстань між кінцем 1 і початком 3 квантилю називається міжквантильним діапазоном, у нашому випадку він дорівнює: $748-$335,5=$412,5. Наявна у табл.1.2 інформація дозволяє побудувати спрощено графічне уявлення розподілу випадкової змінної що досліджується (box plot) в якому представлені кванти лі розподілу та його найменше та найбільше значення.

Найбільш цікаве питання яке можна вирішити на підставі цих даних отримують ли чоловіки заробітну плату більшу ніж жінки. Для цього потрібно відсортувати окремо масиви чоловіків та жінок.

Потім ми знайдемо діапазони, медіани, кванті лі і межквантільни діапазони для ободвох масівів.

Побудуємо окремі гістограми для чоловіків та жінок (табл.3). Звичайно стандартний метод порівняння здійснюється через оцінку середньох значень, та оцінку загальної дисперсії (t критерій). В подальшому ми розглянемо більш детально порівняння вибірок через порівняння середньох значень і обгунтуємо недоліки і переваги цього методу. Продовжимо порівняння рівней оплати праці чоловіків і жінок на підставі поширених статистичних характеристик. Вихідна Інформація для побудові гістограм для чоловіків, жінок і разом представлено в табл.1.1.

Табл.1.3. Розподіл частот за даними табл.1.1.

Діапазон		частоти		відносні	частоти	
(тис. USD)	Чол.	Жінки	Загалом	%(Чол.)	%(Жін.)	%(Заг.)
0,0-0,5	1	23	24	0,06	0,7	0,48

0,5-1,0	10	10	20	0,58	0,30	0,40
1,0-1,5	4	0	4	0,24	0,00	0,08
1,5-2,0	1	0	1	0,06	0,00	0,02
2,0-2,5	1	0	1	0,06	0,00	0,02
∑	17	33	50	1,00	1,00	1,00

Гістограми розподілу тижневої оплати праці представлено на рис.1.1. Верхня гістограма відповідає розподілу оплати праці чоловіків. Середня жінок і нижня оплаті праці всіх працівників. Що стосується модальних значень то за них приймається середина інтервалу з максимальною кількістю спостережень: для чоловіків це 0,75 тис. USD, для жінок 0,35 тис. USD, для загальної виборці 0,25 тис. USD.

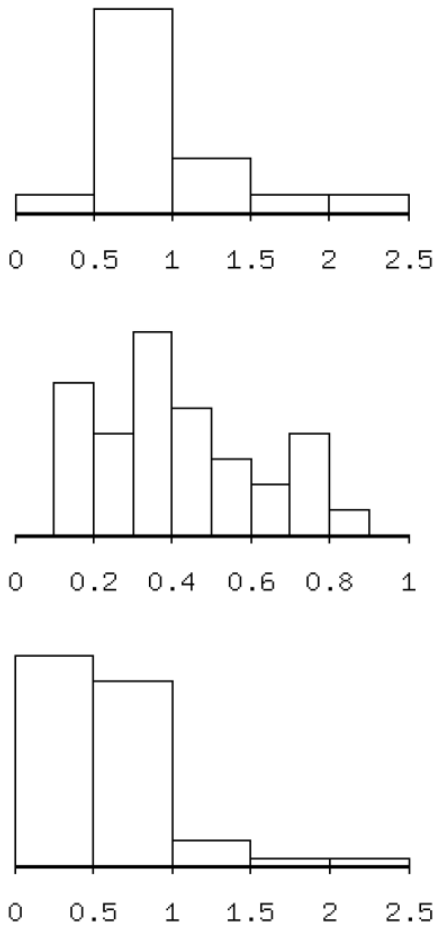


Рис.1.1. Гістограми розподілу тижневої оплати праці (тис. USD) для чоловіків (верхня), жінок (середня),всіх працівників (нижня)

Звичайно гістограма не дозволяє зробити однозначний висновок відносно співвідношення рівня оплати праці чоловіків та жінок, однак слід підкреслити що якщо модальне значення для чоловіків знаходиться в діапазоні 0,5-1,0 тис. USD то для жінок в інтервалі 0,3-0,4 тис.USD. Однак більш ретельний метод порівняння можна здійснити за допомогою побудови box plot для даних по чоловікам та жінкам (рис.1.2).

9.4 Порівняння за допомогою бокс плота.

З представлених box plot слідує, що медіана для чоловіків, а також ніжній і верхній кванті лі для чоловіків суттєво перевищують аналогічні показники для жінок. Крім того максимальне і мінімальне значення для чоловіком суттєво перевищує подібне показники для жінок. Наведени дани відносно рівня оплати праці чоловіков і жінок дозволяють зробити висновок, що базуючись на характеристиках розподілів, чоловіки отримцють більшу щотижневу оплату праці. Однак це не означає, що будь який чоловик отримує щотижневу оплату праці більшу ніж будь-яка жінка.

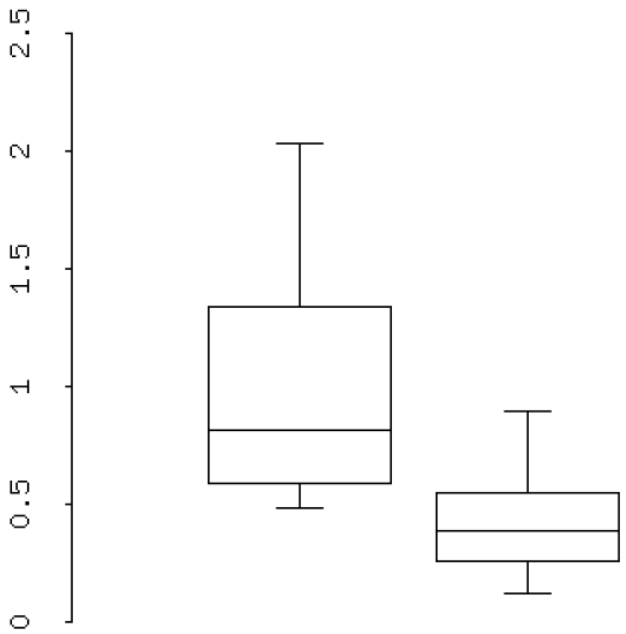


Рис.1.2. Box plot тижневої оплати праці чоловіків і жінок (тис. USD)

9.5. Розподіл хи квадрат - χ^2

Вважаємо що всім є відомий нормальний розподіл який відноситься до двох параметричних (задається двома параметрами: математичним очікуванням - $\bar{x}$ та дисперсією $\sigma^2 \Rightarrow N(\bar{x};\sigma^2)$. В таблицях надаються значення стандартного нормального розподілу з нульовим математичним очікуванням та одиничною дисперсією $N(0;1)$. Характеристики нормального розподілу ізлагаться не будут. Перейдемо до розподілу хи квадрат - χ^2 (chi-squaer). Цій одно параметричний розподіл створюється як сума квадратів випадкових величин, що підпорядковуються стандартному нормальному розподілу. Він повністю визначається кількістью випадкових величин $N(0;1)$, що додаються (кількістю ступенев свободи ν). Математичне очікування для цього розподілу дорівнює

Математичне очікування кількістю ступенев свободи ν, дисперсія 2ν, модальне значення ν-2 при ν>2. Випадкова зміна,що підпорядковується цьому розподілу не може бути відємною величиною (це квадрат або сума квадратів(від дійсної величини). Розподіл χ^2 найчастіші використовується для перевірок відповідності розподілу, що спостерігається в результаті проведення експерименту, деякому теоретичному розподілу, який слідує з проведених дослідником теоретичного аналізу явища, що досліджується. Такі попередні пропозиції звичайно вважаються нульовою гіпотезою. Алгоритм перевірки:

Задається рівень значимості α.

Розраховується статистика:

$$\chi^2 = \sum_{i=1}^{k} \frac{(y_i - e_i)^2}{e_i}; k - кільк.град.спост.$$

$$\sum_{i=1}^{k} y_i = n - кільк.спост.$$

$$y_i - кільк.спост.кат.A_i; e_i = n \cdot p(A_i) - очік.кільк.спост.кат.A_i$$

$$\sum_{i=1}^{k} p(A_i) = 1 - повн.груп.подій$$

Умова відхилення нульової гіпотези відносно розподілу А (p(A)):

$\chi^2 > \chi^2_{\alpha;\nu}; \nu = k - 1 - r$

де r кількість параметрів розподілу А, що оцінюється на підставі виборці.

Розподіл Стьюденту.

Розподіл Стьюденту або t розподіл використовується у богаьох практичних приложеннях статистики. Розподіл студенту як і розподіл χ^2 впроваджується на підставі нормального розподілу. Нехай змінна Z підпорядковується нормальному розподілу, а зщміна U розподілу χ^2 з ν ступенями свободи. Тоді зміна $t = \frac{Z}{\sqrt{U/\nu}}$ підпорядкується t розподілу з ν ступенями свободи. Якщо маємо n незалежних спостережень з нормально розподіленої випадкової величини $N(\mu;\sigma^2)$, то випадкова величина

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ має розподіл Стьюденту з n-1 ступенями свободи (де $\bar{x}; s^2$ – оцінка середнього та дисперсії з виборці у n спостережень.)

Розподіл Стьюденту використуэться в статистичних дослідженнях для побудови довірчих інтервалов для математичного очікування.

Завдання до теми 9

Порівняйте доходи чоловіків та жінок в колективі де ви процюєте.

Література

1. Кальна-Дубенюк Т.П., Литовченко А.М. Оцінка ефективності інформаційно-консультаційного забезпечення поширення інноваційних біотехнологій в умовах ризику та невизначеності // Економіка АПК. 2014. № 1. с. 70–75.

2. Скрипник А.В., Герасимчук Н.. Економічні і фінансові ризики / А.В. Скрипник, Н.. Герасимчук, Житомир: Видавництво ЖДУ, 2013. 371 с.

3. Скрипник А.В., Кравченко К.Я. Врахування погодного ризику при рішенні стандартних оптимізаційних задач аграрного виробництва // Науковий вісник НУБіП України. 2012. № 1 ((177)). с. 344–354.

4. Ус С. Оптиммізація галузі рослинництва сільськогосподарських підприємств // Вісник ХНАУ (серія економічні науки). 2016. № 1. с. 210–129.

5. Gradium A., Pannell D. Risk attitudes and risk perceptions of crops producers in Western Australia під ред. Babcock, B.A., R.W. Fraser, J.N. Lekakis, Amsterdam:, 2003. 113–134 с.

6. Hoag D.L. Applied Risk Management in Agricultuer / D.L. Hoag, Washington DC: CRS Press, 2010. 380 с.

ДОДАТКИ

ДОДАТОК А

Таблиця значень функції Лапласа $\Phi(x) = -\frac{1}{\sqrt{2\pi}}\int_{\infty}^{x} e^{\frac{z^2}{2}} dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,5000	0,36	0,6406	0,72	0,7642	1,08	0,8599
0,01	0,5040	0,37	0,6443	0,73	0,7673	1,09	0,8621
0,02	0,5080	0,38	0,6480	0,74	0,7703	1,10	0,8643
0,03	0,5120	0,39	0,6617	0,75	0,7734	1,11	0,8665
0,04	0,5160	0,40	0,6564	0,76	0,7764	1,12	0,8686
0,05	0,5199	0,41	0,6691	0,77	0,7794	1,13	0,8708
0,06	0,5239	0,42	0,6628	0,78	0,7823	1,14	0,8729
0,07	0,5279	0,43	0,6664	0,79	0,7852	1,15	0,8749
0,08	0,5319	0,44	0,6700	0,80	0,7881	1,16	0,8770
0,09	0,5359	0,45	0,6736	0,81	0,7910	1,17	0,8790
0,10	0,5398	0,46	0,6772	0,82	0,7939	1,18	0,8810
0,11	0,5438	0,47	0,6808	0,83	0,7967	1,19	0,8830
0,12	0,5478	0,48	0,6844	0,84	0,7995	1,20	0,8849
0,13	0,5517	0,49	0,6879	0,85	0,8023	1,21	0,8869
0,14	0,5557	0,50	0,6915	0,86	0,8051	1,22	0,8883
0,15	0,5596	0,51	0,6950	0,87	0,8078	1,23	0,8907
0,16	0,5636	0,52	0,6985	0,88	0,8106	1,24	0,8925
0,17	0,5675	0,53	0,7019	0,89	0,8133	1,25	0,8944
0,18	0,5714	0,54	0,7054	0,90	0,8159	1,26	0,8962
0,19	0,5753	0,55	0,7088	0,91	0,8186	1,27	0,8980
0,20	0,5793	0,56	0,7123	0,92	0,8212	1,28	0,8997
0,21	0,5832	0,57	0,7157	0,93	0,8238	1,29	0,9015
0,22	0,5871	0,58	0,7190	0,94	0,8264	1,30	0,9032

0,23	0,5910	0,59	0,7224	0,95	0,8289	1,31	0,9049
0,24	0,5948	0,60	0,7257	0,96	0,8315	1,32	0,9066
0,25	0,5987	0,61	0,7291	0,97	0,8340	1,33	0,9082
0,26	0,6026	0,62	0,7324	0,98	0,8365	1,34	0,9099
0,27	0,6064	0,63	0,7357	0,99	0,8389	1,35	0,9115
0,28	0,6103	0,64	0,7389	1,00	0,8413	1,36	0,9131
0,29	0,6141	0,65	0,7422	1,01	0,8438	1,37	0,9147
0,30	0,6179	0,66	0,72454	1,02	0,8461	1,38	0,9162
0,31	0,6217	0,67	0,7486	1,03	0,8485	1,39	0,9177
0,32	0,6255	0,68	0,7517	1,04	0,8508	1,40	0,9192
0,33	0,6293	0,69	0,7549	1,05	0,8531	1,41	0,9207
0,34	0,6331	0,70	0,7580	1,06	0,8554	1,42	0,9222
0,35	0,6368	0,71	0,7611	1,07	0,8577	1,43	0,9236
x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
1,44	0,9251	1,73	0,9582	2,04	0,9793	2,62	0,9956
1,45	0,9265	1,74	0,9591	2,06	0,9803	2,64	0,9959
1,46	0,9279	1,75	0,9599	2,08	0,9812	2,66	0,9961
1,47	0,9292	1,76	0,9608	2,10	0,9821	2,68	0,9963
1,48	0,9306	1,77	0,9616	2,12	0,9830	2,70	0,9965
1,49	0,9319	1,78	0,9625	2,14	0,9838	2,72	0,9967
1,50	0,9332	1,79	0,9633	2,16	0,9846	2,74	0,9969
1,51	0,9345	1,80	0,9641	2,18	0,9854	2,76	0,9973
1,52	0,9357	1,81	0,9649	2,20	0,9861	2,78	0,9973
1,53	0,9370	1,82	0,9656	2,22	0,9868	2,80	0,9974
1,54	0,9382	1,83	0,9664	2,24	0,9875	2,82	0,9976
1,55	0,9394	1,84	0,9671	2,26	0,9881	2,84	0,9977
1,56	0,9406	1,85	0,9678	2,28	0,9887	2,86	0,9979
1,57	0,9418	1,86	0,9686	2,30	0,9893	2,88	0,9980
1,58	0,9429	1,87	0,9693	2,32	0,9898	2,90	0,9981

1,59	0,9441	1,88	0,9699	2,34	0,9904	2,92	0,9982
1,60	0,9452	1,89	0,9706	2,36	0,9909	2,94	0,9984
1,61	0,9463	1,90	0,9713	2,38	0,9913	2,96	0,9985
1,62	0,9474	1,91	0,9719	2,40	0,9918	2,98	0,9986
1,63	0,9484	1,92	0,9726	2,42	0,9922	3,00	0,99865
1,64	0,9495	1,93	0,9732	2,44	0,9927	3,20	0,99931
1,65	0,9505	1,94	0,9738	2,46	0,9931	3,40	0,99966
1,66	0,9515	1,95	0,9744	2,48	0,9934	3,60	0,999841
1,67	0,9525	1,96	0,9750	2,50	0,9938	3,80	0,999928
1,68	0,9535	1,97	0,9756	2,52	0,9941	4,00	0,999968
1,69	0,9545	1,98	0,9761	2,54	0,9945	4,50	0,999997
1,70	0,9554	1,99	0,9767	2,56	0,9948	5,00	0,999997
1,71	0,9564	2,00	0,9772	2,58	0,9951		
1,72	0,9573	2,02	0,9783	2,60	0,9953		

ДОДАТОК Б

Таблиця «хвостів» нормального розподілу

$\frac{x-\mu}{\sigma}$	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,496	0,492	0,488	0,484	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,409	0,4052	0,4103	0,3974	0,3936	0,3897	0,3589
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,352	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,33	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,305	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,281	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,242	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,209	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,166	0,1635	0,1611
1	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,123	0,121	0,119	0,117
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,102	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,063	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,025	0,0244	0,0239	0,0233
2	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,017	0,0166	0,0162	0,0158	0,0154	0,015	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,011
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,008	0,0078	0,0076	0,0073	0,0071	0,007	0,0068	0,0066	0,0064
2,5	0,0062	0,006	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,,440	0,0043	0,0042	0,004	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,003	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,002	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3	0,0014									
3,1	0,001									
3,2	0,0007									
3,3	0,0005									
3,4	0,0003									
3,5	0,0002									
3,6	0,0002									
3,7	0,0001									
3,8	7E-05									
3,9	5E-05									
4	3E-05									

ДОДАТОК В

Критичні точки розподілу Стьюдента (t – розподілу)

Число ступенів свободи, k	Рівень значущості, a						
	0,20	0,10	0,05	0,02	0,01	0,002	0,001
1	3,08	6,31	12,7	31,82	63,,66	127,32	636,62
2	1,89	2,92	4,30	6,97	9,93	14,09	31,60
3	1,64	2,35	3,18	4,54	5,84	7,45	12,94
4	1,53	2,13	2,78	3,75	4,60	5,60	8,61
5	1,48	2,02	2,57	3,37	4,03	4,77	6,86
6	1,44	1,94	2,45	3,14	3,71	4,32	5,96
7	1,42	1,90	2,36	3,00	3,50	4,03	5,41
8	1,4	1,86	2,31	2,90	3,36	3,83	5,04
9	1,38	1,83	2,26	2,82	3,25	3,69	4,78
10	1,37	1,81	2,23	2,76	3,17	3,58	4,59
11	1,36	1,80	2,20	2,72	3,11	3,50	4,44
12	1,36	1,78	2,18	2,68	3,05	3,43	4,32
13	1,35	1,77	2,16	2,65	3,01	3,37	4,22
14	1,34	1,76	2,14	2,62	2,98	3,33	4,14
15	1,34	1,75	2,13	2,60	2,95	3,29	4,07
16	1,34	1,75	2,12	2,58	2,92	3,25	4,02
17	1,33	1,74	2,11	2,57	2,90	3,22	3,97
18	1,33	1,73	2,10	2,55	2,88	3,20	3,92
19	1,33	1,73	2,09	2,54	2,86	3,17	3,88
20	1,33	1,73	2,09	2,53	2,85	3,15	3,85
21	1,32	1,72	2,08	2,52	2,83	3,14	3,82
22	1,32	1,72	2,07	2,51	2,82	3,12	3,79
23	1,32	1,71	2,07	2,50	2,81	3,10	3,77
24	1,32	1,71	2,06	2,49	2,80	3,09	3,75
25	1,32	1,71	2,06	2,48	2,79	3,08	3,73

26	1,32	1,71	2,06	2,48	2,78	3,07	3,71
27	1,31	1,70	2,05	2,47	2,77	3,06	3,69
28	1,31	1,70	2,05	2,47	2,76	3,05	3,67
29	1,31	1,70	2,04	2,46	2,76	3,04	3,66
30	1,31	1,70	2,04	2,46	2,75	3,03	3,65
40	1,30	1,68	2,02	2,42	2,70	2,97	3,55
60	1,30	1,67	2,00	2,39	2,66	2,91	3,46
120	1,29	1,66	1,98	2,36	2,62	2,86	3,37
∞	1,28	1,64	1,96	2,33	2,58	2,81	3,29