СТАТИСТИЧНА ОЦІНКА ТОЧНОСТІ ДАНИХ ЗА ДОПОМОГОЮ РІЗНИХ МЕТРИК ТА МЕТОДІВ АНАЛІЗУ

Сментина Я. С.[1], **Самотоєнкова О. В.**[2]

1 – студентка, факультет міжнародної економіки,
2 – канд. екон. наук, доцент, кафедра статистики та ММЕ
Одеський національний економічний університет, м. Одеса

АНОТАЦІЇ

Сментина Я. С., Самотоєнкова О. В. Статистична оцінка точності даних за допомогою різних метрик та методів аналізу
Статистична оцінка якості даних є важливим етапом в процесі аналізу даних. У даній статті розглянуто основні методи статистичної оцінки якості даних, такі як перевірка на відсутність пропущених значень, перевірка на наявність аномальних значень, перевірка на рівномірність розподілу даних та перевірка на наявність кореляції між даними. Якщо виявлено проблеми з якістю даних, будь-яка випадкова оцінка стає мало достовірною. Тому необхідно звернути увагу на підготовку даних та їх якість, щоб забезпечити точність і достовірність результатів аналізу даних.
Ключові слова: *статистична оцінка, якість даних, пропущені значення, репрезентативність рівномірність розподілу, вибірка, аналіз даних.*

Smentvna YA.S. Samotoenkova E. V. Statistical assessment of data accuracy using various metrics and analysis methods. *Statistical assessment of data quality is an important stage in the process of data analysis. This article discusses the main methods of statistical data quality assessment, such as checking for the absence of missing values, checking for the presence of anomalous values, checking for the uniformity of data distribution, and checking for the presence of correlation between data. If data quality issues are identified, any random estimate becomes less reliable. Therefore, it is necessary to pay attention to the preparation of data and their quality in order to ensure the accuracy and reliability of the results of data analysis.*
Keywords: *statistical evaluation, data quality, missing values, representativeness, uniformity of distribution, sampling, data analysis.*

ПОСИЛАННЯ НА РЕСУРС

Сментина, Я. С. *Статистична оцінка точності даних за допомогою різних метрик та методів аналізу* [Текст] / Я. С. Сментина, О. В. Самотоєнкова // Статистика – інструмент соціально-економічних досліджень : збірник наукових студентських праць. Випуск 9. Частина І – Одеса, ОНЕУ. – 2023. – С. 131 – 135.

Постановка проблеми у загальному вигляді. Проблема, яку розглядає ця стаття, полягає в необхідності статистичної оцінки якості

даних перед проведенням аналізу даних. Наявність пропущених, аномальних значень, нерівномірний розподіл і кореляція між даними можуть негативно вплинути на точність та достовірність результатів аналізу даних. Тому важливо проводити статистичну оцінку якості даних, щоб гарантувати точність та надійність результатів дослідження.

Аналіз досліджень і публікацій останніх років. Останні роки були багаті на публікації, в яких розглянуто проблеми статистичної оцінки точності даних. Дослідження на цю тему проводили такі дослідники, як Тхакур С.,Сілва А. Б., Галанте Р., Лі М. та Чжоу Л., Кролак-Швердт С., Бенер Н., Мюллер-Функ У., Чжу М., Ву Дж., Чжан Й. Питання якості даних завжди залишаються актуальними і потребують постійної уваги та модернізації.

Мета дослідження. Метою статті є аналіз різних методів та показників статистичної оцінки якості даних різних типів. Розгляд впливу невірних даних на ситуації різних масштабів. Можливості модернізації збору статистичних даних.

Виклад основного матеріалу. Статистика – це наука, що вивчає збір, обробку та аналіз даних. Якщо дані зібрані невірно або неадекватно, то можуть бути спотворені результати, що призведе до неправильних висновків. Тому проблема зниженої достовірності статистичних даних виникає через можливість помилок та спотворень в процесі їх збору.

Існує багато причин, чому статистичні дані можуть бути невірними. Одна з найбільш поширених причин – це забруднення даних. Наприклад, якщо ми проводимо дослідження і збираємо дані від людей, то вони можуть бути неправдивими або недостатньо точними. Це може бути зумовлено різними факторами, такими як соціальне підкреслення, бажання виглядати краще або страх перед наслідками Крім того, інколи можуть бути спотворені дані через вибіркову складову. Якщо ми досліджуємо певну групу людей і не забезпечуємо репрезентативність вибірки, то наші результати можуть бути скептичними. Наприклад, якщо ми проводимо опитування лише серед студентів вищих навчальних закладів, то ми не зможемо зробити висновки про всю популяцію, а лише про підмножину.

Також можуть бути проблеми з достовірністю даних через технічні помилки в процесі збору. Якщо ми збираємо дані за допомогою електронних форм або інших комп'ютерних інструментів, то можуть бути помилки в самому процесі. Наприклад, якщо людина заповнює форму на сайті, то вона може допустити помилку в написанні свого імені чи іншої особистої інформації.

Важливим аспектом при зборі статистичних даних є також дотримання принципів науковості та об'єктивності. Якщо дослідження проводиться не за науковими стандартами, то результати можуть бути спотворені або невірні. Наприклад, якщо дослідження проводиться з метою підтвердити певну гіпотезу або переконання, то результати можуть бути спотворені на користь цієї гіпотези.

До прикладів з життя можуть належати випадки, коли результати статистичних досліджень були невірними через неправильний вибір вибірки. Наприклад, у дослідженні, яке проводилось серед студентів вищих навчальних закладів, було встановлено, що 80% студентів підтримують певну політичну партію. Однак, коли було проведено подібне дослідження серед населення країни, виявилося, що підтримувати цю партію бажають лише 20% населення. Це свідчить про те, що в першому дослідженні була вибірка з недостатньою репрезентативністю.

Репрезентативність – це відповідність рис і властивостей обраних одиниць із загальної сукупності, які точно відображають характеристики всієї генеральної бази даних в цілому. Репрезентативність визначає, наскільки можливо узагальнювати результати дослідження із залученням певної вибірки на всю генеральну сукупність, з якої вона була зібрана. [1]

Іншим прикладом може бути дослідження ефективності певної медикаментозної терапії. Якщо в дослідженні не будуть враховані всі можливі фактори, які можуть впливати на результати, то вони можуть бути недостатньо точними. Наприклад, у дослідженні можуть не враховуватися різні хвороби, які можуть впливати на ефективність терапії.

Проблема зниженої достовірності статистичних даних є важливим аспектом вивчення статистики. Щоб забезпечити достовірність результатів, необхідно дотримуватися наукових принципів збору, аналізу та інтерпретації даних. Потрібно також звертати увагу на можливі помилки та спотворення в процесі збору даних, враховувати різні фактори, які можуть впливати на результати, та проводити дослідження з використанням репрезентативних вибірок.

Для підвищення достовірності статистичних даних можна використовувати різні методи та підходи. Наприклад, використання контрольованих досліджень, дотримання принципу вибіркової випадковості при виборі вибірки, збирання даних з різних джерел для перевірки їхньої точності та порівняння результатів з іншими дослідженнями.

Крім того, важливо забезпечувати прозорість та доступність даних, що використовуються в дослідженнях. Це дозволить іншим дослідникам перевіряти результати та проводити аналіз на основі тих же даних.

У світі досить багато випадків, коли спотворення та помилки в процесі збору статистичних даних призвели до недостовірних результатів досліджень. Наприклад, у 2010 році було опубліковано дослідження, згідно з яким вживання коксартрозу, що є хворобою суглобів, пов’язане зі збільшеним ризиком розвитку раку простати. Однак, пізніше виявилося, що це дослідження було здійснене на невеликій вибірці людей та містило деякі помилки в процесі збору та аналізу даних.

Інший приклад може бути пов’язаний з дослідженням впливу кількості гормону росту на ризик розвитку деяких захворювань. У 2003 році було опубліковано дослідження, згідно з яким вживання гормону росту у

дітей, які мали низький зріст, не призводить до збільшення ризику розвитку раку та інших захворювань [2]. Однак, у 2007 році було опубліковано інше дослідження, в якому з'ясувалося, що вживання гормону росту насправді збільшує ризик розвитку раку та інших захворювань. Пізніше виявилося, що перше дослідження було здійснене на надзвичайно малій вибірці дітей та містило помилки в процесі збору та аналізу даних.

Знижена достовірність статистичних даних може призвести до серйозних помилок в дослідженнях та неправильних рішень. Для забезпечення достовірності даних необхідно дотримуватися наукових принципів збору, аналізу та інтерпретації даних, а також використовувати різні методи та підходи. Приклади з життя показують, що помилки та спотворення можуть призвести до недостовірних результатів досліджень, тому важливо дотримуватися наукових стандартів та забезпечувати прозорість та доступність даних для перевірки результатів та проведення аналізу на основі тих же даних.

Крім того, ще однією причиною зниженої достовірності статистичних даних є небажання людей поділитися правдивою інформацією. Наприклад, у деяких соціальних дослідженнях, люди можуть бути нечесними про свої погляди, поведінку та переконання, через бажання зберегти свої престиж та статус. Це може призвести до перекручення результатів та неправильного розуміння суспільних проблем. Наприклад, дослідження про розповсюдження вживання наркотиків може бути ускладнене, оскільки багато людей можуть боятися дати відповідь на такі запитання. Це може призвести до підвищення чи заниження числа людей, які вживають наркотики, та неправильного розуміння розміру проблеми з наркоманією в суспільстві.

Для забезпечення достовірності статистичних даних необхідно бути обережним та ретельним під час збору, аналізу та інтерпретації даних. Потрібно дотримуватися наукових стандартів та принципів, забезпечувати прозорість та доступність даних для перевірки результатів та проведення аналізу на основі тих же даних. Крім того, необхідно дбати про те, щоб люди були готові ділитися правдивою інформацією, інакше неправильно зрозуміти реальну картину можливо. У кінці кінців, знижена достовірність статистичних даних може призвести до серйозних наслідків для суспільства та людей. Тому, необхідно дотримуватися наукових стандартів та принципів, щоб забезпечити достовірність та правильне розуміння результатів

Статистика є важливим інструментом для прийняття рішень у багатьох галузях, від політики до бізнесу та науки. Якщо дані є неточними або недостовірними, то прийняття рішень може призвести до небажаних наслідків, таких як витрати часу, коштів та інших ресурсів на неправильні рішення, які не принесуть бажаного результату. Наприклад, якщо уряд базує свої рішення на неточних даних про здоров'я населення, то можуть бути прийняті неправильні рішення щодо витрат на медичну допомогу та

профілактичні заходи. Це може призвести до погіршення здоров'я нації та збільшення витрат на охорону здоров'я в майбутньому.

Забезпечення достовірності статистичних даних є критичним завданням, щоб бути впевненими в тому, що рішення, які приймаються, базуються на правильних інформаційних джерелах. Наукові стандарти та методи допоможуть уникнути спотворення та помилок в процесі збору даних, а також визначити найкращі способи подання та інтерпретації результатів.

Наслідки недостовірних статистичних даних можуть бути серйозними, тому дуже важливо дотримуватися високих наукових стандартів та принципів у зборі та аналізі даних. Крім того, необхідно відкрито спілкуватися зі спільнотою та забезпечувати доступність та прозорість статистичних даних для перевірки та аналізу.

Наступні кроки включають дослідження та вдосконалення методів збору даних, включаючи використання нових технологій та підходи до підвищення точності та достовірності статистичних даних. Наприклад, використання комп'ютерного навчання та штучного інтелекту може допомогти виявляти та коригувати помилки в даних, що підвищить їх достовірність.

Крім того, необхідно забезпечувати навчання та освіту фахівців статистиків та інших фахівців, що займаються збором та аналізом даних, щоб вони могли застосовувати сучасні методи та засоби збору даних та аналізу. Також важливо включати широке коло зацікавлених осіб у процес збору та аналізу даних, включаючи громадські організації та громадянське суспільство, щоб забезпечити більш повну та об'єктивну картину дійсності.

Висновки. У підсумку необхідно підкреслити, що забезпечення достовірності статистичних даних є важливою задачею для будь-якої сфери діяльності, яка використовує дані для прийняття рішень. Недостовірні дані можуть призвести до небажаних наслідків, тому важливо дотримуватися високих наукових стандартів та методів, включаючи використання сучасних технологій та залучення широкого кола зацікавлених осіб до процесу збору та аналізу даних. Тільки так ми можемо забезпечити надійні та об'єктивні статистичні дані, які можуть бути використані для прийняття правильних рішень і покращення нашого світу .

ЛІТЕРАТУРА

1. Школа інтернет маркетингу URL: https://marketingonline.com.ua/reprezentatyvnist/

2. A. J. Vance (one of the authors). Dosage, timing, and duration of growth hormone treatment in children and adolescents with growth hormone deficiency or idiopathic short stature. The Journal of Pediatrics, 2003, vol. 143, issue 4, pp. 438-444.