УДК 519.2

ДОСЛІДЖЕННЯ ВПЛИВУ ЗБІЛЬШЕННЯ РОЗМІРУ ВИБІРКИ НА РЕЗУЛЬТАТИ ПЕРЕВІРКИ СТАТИСТИЧНИХ ГІПОТЕЗ

Козачук Андрій

CDM a/s

Анотація

Проведення порівняння аналізу нормальності розподілу та ширини довірчих інтервалів даних відхилення прогнозу температури від її реального значення для двох вибірок даних. Показано, що кількість записів у початковій вибірці достатня для висновку про нормальність їх розподілу.

Abstract

Research describes comparison of normality test and size of confidence intervals of temperature forecast accuracy in two data sets. It was shown that number of records in the basic set is big enough to make a conclusion about how well it is modeled by normal distribution, as increased number of data hasn't significally changed its statistical characteristics.

Вступ

Після публікації першого матеріалу про прогноз погоди краулери даних продовжили працювати ще півтора роки. Вибірка даних про прогноз температури у містах Вінниця та Фреденсборг збільшилась майже у 4 рази. Розглянемо, як це вплинуло на результати перевірки статистичних гіпотез та рівень значущості.

Різниця між початковою і розширеною вибіркою

У першу публікацію[1] потрапили дані, зібрані з березня до грудня 2017 року. Після цього система збору прогнозів погоди продовжувала працювати до травня 2019 року, що дало можливість суттєво збільшити вибірку для аналізу. Різниця між кількістю зібраних даних показана у таблиці 1.

Таблиця 1. Порівняння кількості прогнозів по джерелам

Джерело прогнозу	**Загальна кількість прогнозів 2018**	**Загальна кількість прогнозів 2020**
Sinoptik	311	1204
Gismeteo	309	1188
Meteoprog	303	1182
Accuweather	279	1158
DarkSky	303	1182

Аналіз нормальності розподілу

Порівняємо статистичні характеристики даних 2018 і 2020 років. Зауважимо, що пізніша вибірка не є ідеальним продовженням початкової, за два роки сервіси прогнозу погоди могли змінити свої внутрішні моделі та алгоритми, змінивши, таким чином якість свого прогнозу.

Рисунок 1 показує, що візуально дані 2020 року більше нагадують правильний «дзвін» нормального закону розподілу, ніж дані 2018 року, хоча помітно, що правий хвіст графіку густини розподілу вищий за лівий. Те що розподіл не є повністю нормальним підтверджують і дані з таблиці 2, зі збільшенням кількості даних його номінальна «нормальність» навіть зменшилась.

Таблиця 2. Порівняння статистичних характеристик вибірок 2018 і 2020 років.

Характеристика\Джерело	**2018**	**2020**	**Значення для нормально розподілених даних**
Коефіцієнт варіації	8.79	10.84	<0.1
Ексцес	2.44	4.49	[-1;1]
Коефіцієнт асиметрії	0.59	0.30	[-1;1]

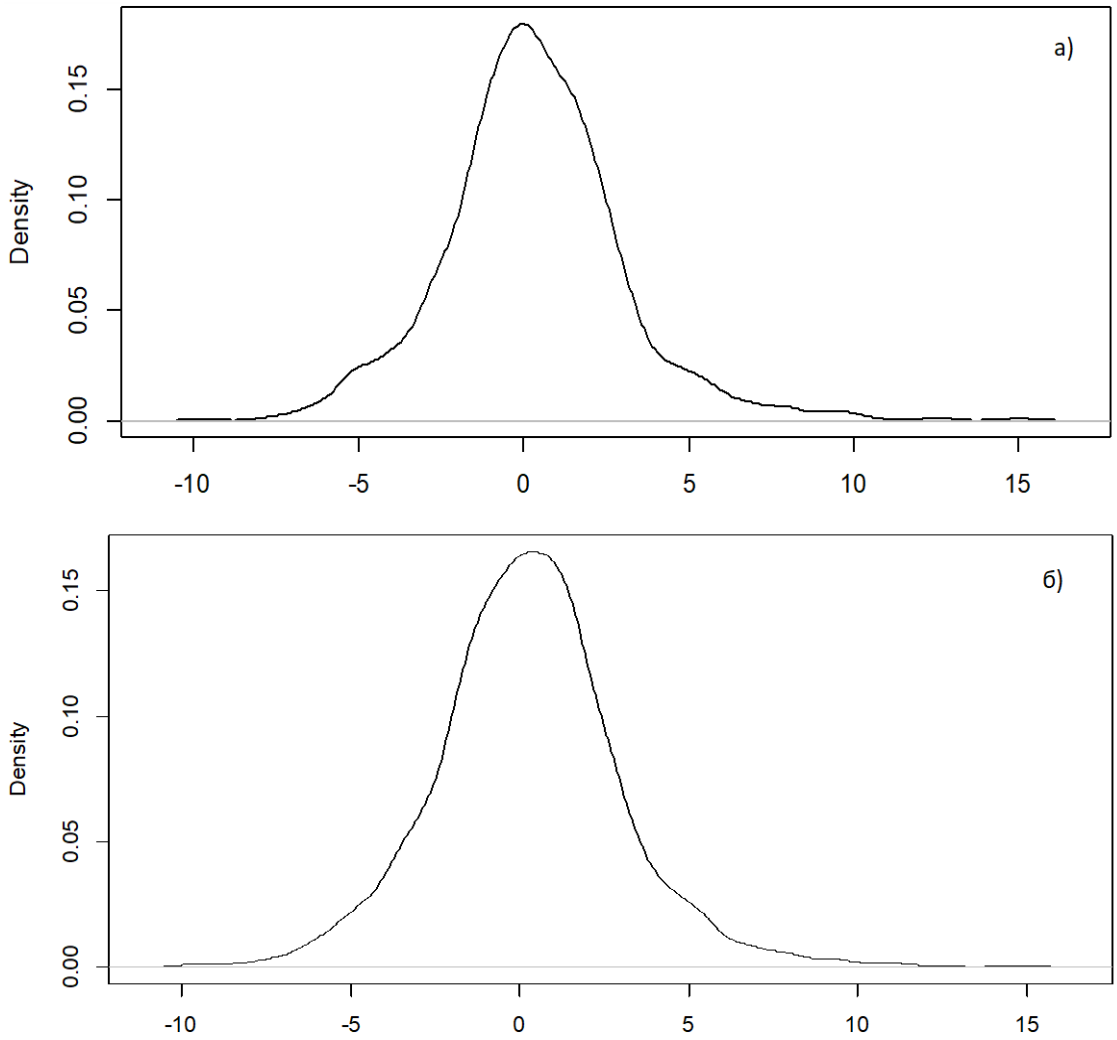

Рисунок 1 – Ядрова оцінка густини розподілу відхилення прогнозу температури від її реального значення: а) для даних 2018 року, б) для даних 2020 року.

Аналіз впливу розміру вибірки на ширину довірчих інтервалів

Збільшення кількості даних майже в 4 рази дозволило вдвічі зменшити ширину довірчих інтервалів розрахованого середнього значення відхилення прогнозу температури від реального значення, як показано у таблиці 3. Даний висновок є справедливим як для даних, що описують прогноз температури повітря у м. Вінниця, так і для данних м. Фреденсборг [2].

Таблиця 3. Порівняння ширини 90% довірчого інтервалу середнього значення відхилення прогнозу температури від реального значення між наборами даних.

Джерело	**Ширина довірчого інтервалу у 2018 р.**	**Ширина довірчого інтервалу у 2020 р.**
Sinoptik	0.78	0.35
Gismeteo	0.75	0.9
Meteoprog	0.76	0.34
Accuweather	1.26	0.52
DarkSky	0.68	0.3

Розглянемо, як збільшення кількості даних вплинуло на результати перевірки наступної статистичної гіпотези. Гіпотеза H_{01} : Імовірність виправданості прогнозу температури на 24 години у всіх джерел однакова.

Використаємо критерій χ^2 [3] для початкового набору даних. Кількість ступенів свободи $df=(M-1)(L-1)$, де M – кількість джерел прогнозу, L – кількість розподілів, що порівнюються. $df = (5-1)(2-1) = 4$.

$$\chi^2 = \sum_{i=1}^{M} \frac{(O_i - E_i)^2}{E_i};$$

де O_i – Кількість виправданих прогнозів для джерела i, E_i – очікувана кількість виправданих прогнозів для джерела i.

$$\chi^2 = \frac{37.21}{233.9} + \frac{77.44}{232.2} + \frac{146.41}{227.9} + \frac{835.21}{209.9} + \frac{4.41}{227.9} = 0.16 + 0.33 + 0.64 + 3.98 + 0.02 = 5.13$$

Критичне значення $\chi^2_{\text{кр}}$ для рівня значущості 0.01 і 4 ступенів свободи: $\chi^2_{\text{кр}} = 13.27$. Нульова гіпотеза H_{01} не може бути відкинута.

Тепер використаємо той самий критерій для розширеного набору даних.

$$\chi^2 = \frac{5535.4}{921.6} + \frac{1413.8}{909.4} + \frac{4.8}{904.8} + \frac{24774.8}{886.4} + \frac{1866.2}{904.8} = 6.01 + 1.55 + 0.01 + 27.95 + 2.06 = 37.58$$

Нульова гіпотеза H_{01} може бути відкинута, на відміну від попередньої ітерації, так як значення критерію вище за критичне.

Висновки

Результати дослідження показали, що розмір початкої вибірки був достатнім для аналізу нормальності розподілу. Збільшення кількості даних в чотири рази дозволило вдвічі зменшити ширину довірчих інтервалів середнього значення відхилення прогнозу температури від її реального значення.

Список використаних джерел

1. Козачук А. В. Аналіз якості прогнозу погоди популярних українських сайтів // Збірник матеріалів одинадцятої міжнародної конференції "Інтернет-Освіта-Наука-2018".- Вінниця.- 2018.

2. Дослідження якості прогнозу погоди 2.0 [Електронний ресурс] – Режим доступа: byandriykozachuk.wordpress.com/2020/04/09/дослідження-якості-прогнозу-погоди-2-0/

3. Гржибовский А. Анализ номинальных данных (независимые наблюдения) [Електронний ресурс] – Режим доступа: https://cyberleninka.ru/article/n/analiz-nominalnyh-dannyh-nezavisimye-nablyudeniya