Лекція 05-06.

НЕЗАЛЕЖНІ ЛІНГВІСТИЧНІ ВИПРОБУВАННЯ В ТЕКСТІ

При дослідженні механізмів породження тексту результати окремого лінгвістичного випробування не мають великого значення. Вивчення взаємодії системи, норми та ситуації експлікується за допомогою моделей теорії ймовірностей, які передбачають здійснення масового експерименту, при якому одна і та сама лінгвістична подія повторюється багато разів. Ці випробування, що повторюються, утворюють серії, в кожній з яких подія з'являється або не з'являється певну кількість разів.

Вибір тої чи іншої моделі опису тексту залежить від побудови ймовірнісно-лінгвістичного випробування і, зокрема, від того, як організовано вибір з тексту окремих його одиниць.

1.2 Повторна і безповторна вибірки

Розглянемо такий елементарний приклад. Нехай з тексту взято N фонем, серед яких n голосних та m приголосних, і кожна з фонем записана на окрему картку; картки покладені в урну і перемішані. Випробування, які полягають у витяганні з урни однієї картки, можуть здійснюватись за двома схемами.

За умовами першої схеми кожна вийнята картка повертається до урни, після того як у протоколі фіксується результат кожного випробування. При кожному наступному випробуванні ймовірності появи голосної чи приголосної залишаються незмінними. (Ці ймовірності відповідно дорівнюють n/N та m/N.) Ймовірнісно-лінгвістичний експеримент, який оперує з наслідками взаємно незалежних випробувань, у кожному з яких лінгвістичні події зберігають свої безумовні ймовірності, називається *повторною вибіркою*.

При реалізації другої схеми взяті з урни картки не повертаються. Ймовірність появи голосної чи приголосної у кожному наступному випробуванні залежить від результатів попередніх випробувань. Таким чином, ми маємо справу з залежними випробуваннями, а ймовірність результату кожного з випробувань є умовною. Експеримент, який оперує з послідовністю залежних випробувань, у кожному з яких результати мають умовні ймовірності, називається *безповторною (або без повернень) вибіркою*.

Реальний ймовірнісно-лінгвістичний експеримент може бути здійснений як за допомогою повторної, так і за допомогою безповторної вибірки.

2 ТРИ СХЕМИ НЕЗАЛЕЖНИХ ЛІНГВІСТИЧНИХ ВИПРОБУВАНЬ

Квантитативне мовознавство широко використовує метод серійного спостереження. Суть його полягає в тому, що лінгвістичні одиниці вибираються з тексту групами фіксованої довжини: наприклад, по десять фонем, по сто речень або словоформ тощо. Лінгвістичні одиниці, які утворюють серію, необов'язково повинні бути в тексті одна за одною, вони можуть вибиратись і через певний інтервал.

При розв'язуванні багатьох теоретичних та інженерно-лінгвістичних задач часто потрібно знати ймовірність появи тої чи іншої кількості певних лінгвістичних одиниць у серії.

Якщо лінгвістичні випробування, які утворюють серію, розглядаються як незалежні, то ми можемо здійснювати необхідне прогнозування за допомогою розроблених у теорії ймовірностей трьох систем незалежних випробувань: простої, поліноміальної та пуасонівської.

Проста схема передбачає тільки два результати досліду: появу або не появу ознаки *A*. Прикладом такої схеми є повторна вибірка з тексту приголосних (A) і голосних ($\overline{A}$) фонем.

У *поліноміальній схемі* випробування дає не два, а декілька результатів. За цією схемою здійснюється, наприклад, експеримент, який полягає у виборі з тексту графем трьох видів: букв, знаків пунктуації та пробілів.

У *пуасонівській схемі* незалежні випробування здійснюються відносно декількох сукупностей (підмов, стилів, тематик), у кожній з яких ознака має різну ймовірність. Тому ймовірність лінгвістичного результату змінюється в залежності від того, відносно якої підмови або тематики проводиться дослід.

Математична модель, за якою здійснюється прогнозування результатів простої схеми випробувань, є основою для побудови інших імовірнісних моделей, у тому числі і тих, котрі широко використовуються у квантитативній лінгвістиці.

Квантитативна лінгвістика вивчає та описує лінгвістичні явища за допомогою методів «кількісної» математики (комбінаторика, теорія ймовірності, математична статистика, математичний аналіз, теорія інформації тощо).

2.1 Проста схема незалежних випробувань. Формула Бернуллі

Припустимо, що в деякому тексті з n фонем є m приголосних і n-m голосних. За схемою повторної вибірки проводиться N незалежних випробувань, які полягають у послідовному випадковому витяганні фонеми з тексту. Потрібно визначити ймовірність події, яка полягає в тому, що серед витягнених N фонем рівно x виявляться приголосними, причому порядок слідування голосної і приголосної фонем байдужий.

Вважатимемо появу приголосної подією A, а появу голосної – подією $\overline{A}$. Визначимо ймовірності появи голосної та приголосної. За класичним означенням ймовірності маємо:

$$P(A)=m/n=p, \ P(\overline{A})=(n-m)/n=q.$$

Тепер знайдемо ймовірність того, що при N незалежних випробуваннях подія A з'явиться рівно x разів, якщо ймовірність появи цієї події при кожному окремому випробуванні стала і дорівнює p.

Для цього складемо всі можливі схеми, які утворять послідовність з появи x разів події A та $N-x$ разів не появи цієї події, тобто $AA\ldots A\ \overline{A}\,\overline{A}\ldots\overline{A}$. За теоремою множення ймовірність появи кожної схеми складає $p^x q^{N-x}$, а кількість таких схем дорівнює кількості сполук з N елементів по x, тобто C_N^x. Звідси випливає, що ймовірність появи події A рівно x разів у серії N незалежних випробувань складає

$$P_N(x)=C_N^x p^x q^{N-x}=\frac{N!}{x!(N-x)!}p^x q^{N-x}, \qquad (1)$$

де $p+q=1$. Зауважимо також, що ймовірності (2.1) дорівнюють відповідним членам розкладу за формулою біному виразу $(q+p)^N$.

За допомогою виразу (1), який називається *формулою Бернуллі*, і здійснюється імовірнісне прогнозування результатів у простій схемі незалежних випробувань.

Усі можливі несумісні між собою результати N дослідів полягають у появі 0, 1, 2,..., N разів події A. Тому сума величин (2.1), які є окремими значеннями ймовірностей при x=0, 1, 2, ..., N, дорівнює 1:

$$\sum_{x=0}^{N} P_N(x) = \sum_{x=0}^{N} C_N^x p^x q^{N-x} = (q+p)^N = 1.$$

Розподіл ймовірностей $P_N(x) = C_N^x p^x q^{N-x}$ при x=0, 1, 2, ..., N називається *біноміальним розподілом* (або *біноміальним законом розподілу*) ймовірностей.

При побудові алгоритмів послівного машинного перекладу та інформаційного пошуку постійно виникають задачі, пов’язані з прогнозуванням появи у сегментах заданої довжини певної кількості словоформ, морфем або словосполучень, які належать деяким класам. Формула Бернуллі дозволяє розв’язувати задачі цього типу за умови, що зберігається припущення про взаємну незалежність словоформ, які утворюють даний сегмент.

Часто, щоб одержати достатньо достовірні результати, доводиться проводити велику кількість незалежних випробувань. При цьому величини N та x можуть бути достатньо великими, що робить обчислення за щойно

описаною схемою дуже важкими. У таких випадках обчислення ймовірностей $P_N(x)$ здійснюється за наближеними формулами, які ми розглянемо пізніше.

Часто для розв'язування лінгвістичної або інформаційної задачі необов'язково визначати всі ймовірності появи даної події $0, 1, 2, \ldots, N$ разів. Достатньо вказати найбільш ймовірну кількість появ цієї події. Розглянемо відповідну схему. Зі збільшенням x величина $P_N(x)$ зростає і при деякому x_0 (воно називається *модальним значенням*) досягає свого найбільшого значення $P_N(x_0)$. Після цього зі збільшенням x ймовірність $P_N(x)$ спадає.

Щоб визначити модальне значення x_0, розглянемо поведінку функції $P_N(x)$ послідовним порівнянням двох сусідніх членів розподілу. Нехай $P_N(x_0)$ – найбільше значення ймовірності у розподілі (2.1). Тоді виконуються такі дві нерівності:

$$P_N(x_0-1) \le P_N(x_0),\ P_N(x_0) \ge P_N(x_0+1). \tag{2}$$

Перепишемо першу з нерівностей (2) у вигляді

$$\frac{P_N(x_0)}{P_N(x_0-1)} = \frac{C_N^{x_0} p^{x_0} q^{N-x_0}}{C_N^{x_0-1} p^{x_0-1} q^{N-x_0+1}} = \frac{(N-x_0+1)p}{x_0 q} \ge 1. \tag{3}$$

Замінивши в останній нерівності q на $p-1$, одержимо

$$x_0 \le Np+p. \tag{4}$$

Аналогічно, записавши другу з нерівностей (2.2) у вигляді

$$\frac{P_N(x_0+1)}{P_N(x_0)} = \frac{C_N^{x_0+1} p^{x_0+1} q^{N-x_0-1}}{C_N^{x_0} p^{x_0} q^{N-x_0}} = \frac{(N-x_0)p}{(x_0+1)q} \le 1, \tag{5}$$

одержимо

$$x_0 \ge Np+p-1. \tag{6}$$

Об'єднуючи (4) та (6), одержуємо подвійну нерівність

$$Np+p-1 \le x_0 \le Np+p. \tag{7}$$

Тепер використаємо наведені вище дані про використання іменників і визначимо найімовірнішу кількість появ іменників в англійському реченні з десяти слів. Оскільки $N=10$, $p=1/3$, то за (7) маємо

$$\frac{10}{3}-\frac{2}{3}\le x_0\le\frac{10}{3}+\frac{1}{3}, \text{ або } 2\frac{2}{3}\le x_0\le 3\frac{2}{3}.$$

Таким чином, найбільш ймовірна кількість появ іменників в англійському реченні (сегменті) з 10 слів дорівнює 3. Цей же результат дає розподіл ймовірностей, який ми розглянути раніше.

Знаючи модальне значення x_0, можна визначити потрібні нам імовірності біноміального розподілу. Обчислення їх починається з визначення максимальної ймовірності $P_N(x_0)$:

$$P_N(x_0)=C_N^{x_0}p^{x_0}q^{N-x_0}=\frac{N!}{x_0!(N-x_0)!}p^{x_0}q^{N-x_0}. \qquad (8)$$

Обчислення решти ймовірностей здійснюється за такими рекурентними формулами, що побудовані на використанні виразів (3) та (5):

при $x < x_0$

$$\left.\begin{aligned}
&P_N(x_0-1)=\frac{x_0}{N-(x_0-1)}\cdot\frac{q}{p}\cdot P_N(x_0),\\
&P_N(x_0-2)=\frac{x_0-1}{N-(x_0-2)}\cdot\frac{q}{p}\cdot P_N(x_0-1),\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&P_N(x_{\min}+1)=\frac{x_0+2}{N-x_{\min}-1}\cdot\frac{q}{p}\cdot P_N(x_{\min}+2),\\
&P_N(x_{\min})=\frac{x_{\min}+1}{N-x_{\min}}\cdot\frac{q}{p}\cdot P_N(x_{\min}+1),
\end{aligned}\right\} \tag{9а}$$

при $x > x_0$

$$\left.\begin{aligned}
&P_N(x_0+1)=\frac{N-x_0}{x_0+1}\cdot\frac{p}{q}\cdot P_N(x_0),\\
&P_N(x_0+2)=\frac{N-(x_0+1)}{x_0+2}\cdot\frac{p}{q}\cdot P_N(x_0+1),\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
&P_N(x_{\max}-1)=\frac{N-(x_{\max}-2)}{x_{\max}-1}\cdot\frac{p}{q}\cdot P_N(x_{\max}-2),\\
&P_N(x_{\max})=\frac{N-(x_{\max}-1)}{x_0+1}\cdot\frac{p}{q}\cdot P_N(x_{\max}-1),
\end{aligned}\right\} \tag{9б}$$

де $x_{\min} \ge 0$ та $x_{\max} \le N$.

2.2 Поліноміальна схема

Якщо лінгвістичне випробування має декілька результатів, то їх імовірнісне прогнозування здійснюється за допомогою поліноміальної схеми. Її математична модель будується так.

Припустимо, що деяке лінгвістичне випробування може мати один з k різних попарно несумісних результатів A_1, A_2,..., A_k. Ймовірність кожного з них позначимо відповідно через $P(A_1)=p_1$, $P(A_2)=p_2$,..., $P(A_k)=p_k$. Оскільки подія $A_1+A_2+\cdots+A_k$ є достовірною, то $p_1+p_2+\cdots+p_k=1$. Здійснимо N незалежних випробувань і визначимо ймовірності того, що подія A_1 з'явиться x_1 разів, подія A_2 – x_2 разів,..., подія A_k – x_k разів, де $x_1+x_2+\cdots+x_k=N$.

Вказаний результат одержується різними шляхами, кожний з яких відповідає різним переставленням x_1 разів результату A_1, x_2 разів результату

$A_2,\dots,x_k$ разів результату A_k. Ймовірність появи кожної такої комбінації дорівнює $p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$. Загальна кількість таких комбінацій дорівнює добутку $C_N^{x_1}C_N^{x_2}\cdots C_N^{x_k}$, який приводиться до виразу

$$\frac{N!}{x_1!x_2!\cdots x_k!}.$$

Звідси одержуємо, що при N незалежних випробуваннях ймовірність одержати x_1 разів результат A_1, x_2 разів результат $A_2,\dots,x_k$ разів результат A_k дорівнює

$$P_N(x_1,x_2,\dots,x_k)=\frac{N!}{x_1!x_2!\cdots x_k!}\cdot p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}, \tag{10}$$

де $0\le x_i\le N$, а $\sum_{i=1}^{k}x_i=N$.

У частковому випадку, коли $k=2$, маємо

$$P_N(x_1,x_2)=\frac{N!}{x_1!x_2!}\cdot p_1^{x_1}p_2^{x_2}.$$

Враховуючи, що $x_1+x_2=N$, а $p_1+p_2=1$, і позначаючи x_1 через x, x_2 – через $N-x$, p_1 – через p, а p_2 – через q, приходимо до виразу

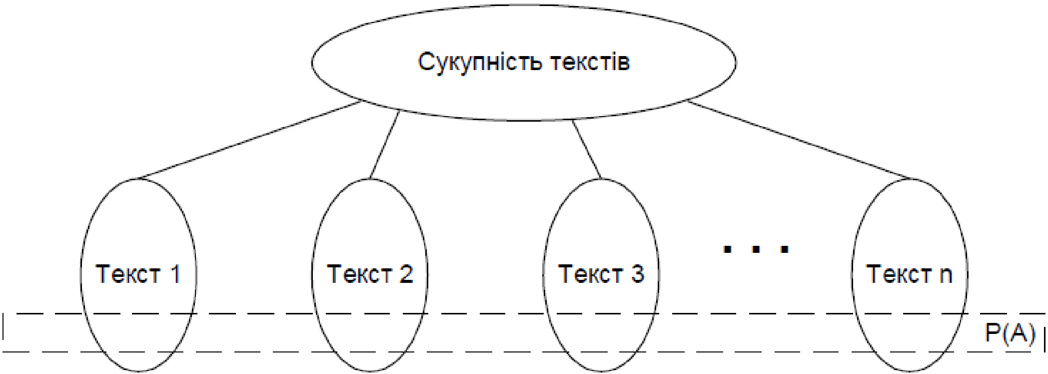
$$P_N(x)=\frac{N!}{x!(N-x)!}p^x q^{N-x}=C_N^x p^x q^{N-x},$$

тобто до формули Бернуллі для простої системи схеми незалежних випробувань. Отже, формула Бернуллі є частинним випадком співвідношення (10).

Як і проста схема, поліномна схема використовується у повторних лінгвістичних вибірках за умови, що величини $N, x_1, x_2, \ldots, x_k$ не дуже великі. За цих умов використання розглянутої схеми дає цінну інформацію не тільки для імовірнісної побудови алгоритмів синтаксичного аналізу іноземного тексту при машинному перекладі. Ці алгоритми дозволяють також визначити оптимальну послідовність викладання синтаксичного матеріалу при навчанні іноземній мові у середній школі та вузі.

2.3 Пуассонівська схема

У лінгвістичній практиці часто доводиться мати справу з такою мовною сукупністю, у якій тексти, що її складають, належать до різних підмов і стилів. Оскільки тексти будуються, виходячи з різних норм, то кожна лінгвістична одиниця має в кожному тексті свою апріорну ймовірність. У підсумку ймовірності появи та не появи певних мовних одиниць міняються від досліду до досліду.



Така ситуація, зображена на малюнку, описується *схемою Пуассона*.
Формальне подання цієї схеми ґрунтується на таких міркуваннях.

Нехай здійснюється N незалежних випробувань, у кожному з яких може з'явитись або не з'явитись подія A. Ймовірності появи події A в 1, 2, ..., N випробуваннях відповідно дорівнюють $p_1, p_2, \ldots, p_N$, а ймовірності її не появи дорівнюють $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, $q_N = 1 - p_N$. Можна показати, що ймовірність появи результату A в серії з N випробувань рівно x разів складає

$$P_N(x) = p_1 p_2 p_3 \cdots p_x q_{x-1} \cdots q_N + p_1 q_2 p_3 \cdots q_{N-1} p_N + \cdots + q_1 q_2 q_3 \cdots q_{N-x} p_{N-x+1} p_{N-x+2} \cdots p_N . \tag{11}$$

Таким чином, потрібна ймовірність є сумою всіх можливих добутків, у кожному з яких p з різними індексами міститься рівно x разів, а q з різними індексами входить $N - x$ разів.

Щоб утворити всі можливі добутки з x ймовірностей p_i та $N - x$ ймовірностей q_i $(i = 1, 2, \ldots, N)$, утворимо добуток біномів

$$(q_1 + p_1 t)(q_2 + p_2 t) \cdots (q_N + p_N t) = \prod_{i=1}^{N} (q_i + p_i t), \tag{12}$$

де t – деякий довільний параметр.

Перемножимо біноми і зведемо подібні члени, тоді одержимо рівність

$$\prod_{i=1}^{N}\left(q_i + p_i t\right) = \sum_{x=0}^{N} P_N(x)\, t^x ,$$

у якій коефіцієнт при t^x є ні що інше, як вираз (2.11).

Розкриємо дужки у лівій частині рівності й зведемо подібні члени, тоді отримаємо всі ймовірності $P_N(0)$, $P_N(1)$, $P_N(2)$, ..., $P_N(N)$, котрі виступають у ролі коефіцієнтів, відповідно, при t^0, t^1, t^2, ..., t^N. Сума всіх ймовірностей $P_N(x)$ дорівнює 1:

$$\sum_{x=0}^{N} P_N(x) = 1 .$$

Зокрема, якщо $p_1 = p_2 = \ldots = p_N = p$, $q_1 = q_2 = \ldots = q_N = q$, маємо

$$(q + pt)^N = \sum_{x=0}^{N} C_N^x p^x q^{N-x} t^x ,$$

звідки випливає формула Бернуллі.

Схему Пуассона, як і дві попередні схеми, доцільно використовувати в лінгвістичному випробуванні тоді, коли ми можемо організувати повторну вибірку, а величини N та x не дуже великі.

У попередніх пунктах ми навчились прогнозувати результати масових лінгвістичних випробувань. Такі прогнози ми можемо поки що здійснювати стосовно повторних вибірок, спираючись на класичне означення ймовірності, тобто за умови, що дослід здійснюється відносно порівняно обмеженої за обсягом сукупності лінгвістичних об'єктів. Така ситуація зустрічається у лінгвістиці порівняно рідко. Найчастіше мовознавцю доводиться мати справу з безповторною вибіркою, яка досліджує лінгвістичні одиниці, що рідко зустрічаються. За таких умов розподіл ймовірностей появи події A підпорядковується гіпергеометричному закону.

2.4 Безповторна лінгвістична вибірка та її опис за допомогою формули Бернуллі

Гіпергеометричний закон може застосовуватись тільки до скінченних генеральних сукупностей, об'єм яких відомий. Оскільки в лінгвістичних задачах об'єм генеральної сукупності текстів, які породжуються відкритою системою мови, звичайно не є скінченною величиною, то застосування вказаного закону для прогнозування результатів лінгвістичних дослідів у безповторних вибірках виявляється нереальним. Разом з цим, за певних умов гіпергеометрична ймовірність добре апроксимується біноміальною ймовірністю. Тому, не боячись порушення математичної строгості, ми будемо здійснювати розрахунок ймовірностей появи події A рівно x разів у нашій безповторній вибірці так, як якщо б мова йшла про повторну вибірку. Іншими словами, ми застосовуємо до безповторних вибірок біноміальний закон.

Будемо розглядати дані S текстів як S серій або вибірок, кожна з яких складається з N незалежних випробувань. Лінгвістична подія A може з'явитись у кожній серії x разів (x=0, 1, 2, …, N). Неважко зауважити, що є групи серій, у яких A появляється 0, 1, 2, ..., N разів. Звідси випливає, що відносна частота появи події A рівно x разів у одній серії визначається

співвідношенням $f_N(x) = S_x / S$, де S_x – кількість серій, у яких подія A з'явиться рівно x разів.

Апріорна ймовірність появи події A в одній навмання взятій серії дорівнює

$$p \approx \frac{\sum xS_x}{NS},$$

і, отже,

$$q \approx 1 - \frac{\sum xS_x}{NS}.$$

У одержаному теоретичному розподілі кожному значенню x співвіднесена не його ймовірність, а деяка теоретично очікувана кількість серій (вибірок) S_x^T, у яких подія A появляється рівно x разів. Оскільки

$$S_x^T = SP_N(x) = SC_N^x p^x q^{N-x}, \qquad (13)$$

то неважко зауважити, що величини S_x^T та $P_N(x)$ зв’язані коефіцієнтом пропорційності S.

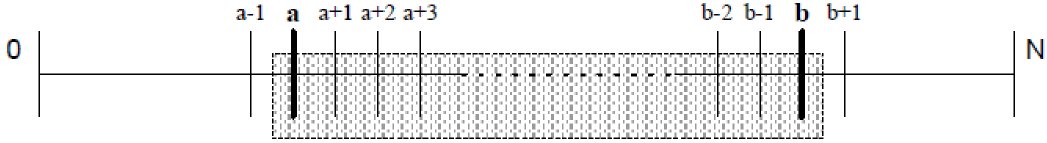
3 ІМОВІРНІСТЬ ПОЯВИ ПОДІЇ В ЗАДАНОМУ ДІАПАЗОНІ КІЛЬКОСТІ ПОЯВ

Розглянуті вище властивості описували числові особливості серії з кількох вибірок. Повернімося тепер до аналізу якісних та кількісних характеристик, якими володіє одна вибірка.

Нехай B'_x – подія, яка полягає в тому, що лінгвістична одиниця A з'явиться на менше a і не більше b разів. Тоді ймовірність $P_N(a \le x \le b)$ цієї події складає

$$P_N(a \le x \le b) = P_N(a) + P_N(a+1) + \cdots + P_N(b-1) + P_N(b) = \sum_{x=a}^{b} P_N(x) = \sum_{x=a}^{b} C_N^x p^x q^{N-x} .$$

Графічно кількість доданків, які необхідно обчислити можна зобразити так:



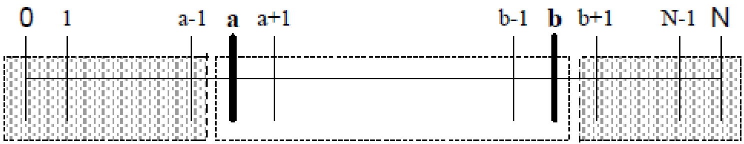
Якщо кількість членів, які відповідають значенням x від a до b, значно більша загальної кількості членів, що відповідають значенням x від 0 до $a-1$ і від $b+1$ до N, то зручніше здійснювати сумування ймовірностей за цими двома послідовностями. У такому разі одержуємо ймовірність протилежної події $\overline{B'_x}$:

$$P\left(\overline{B'_x}\right)=\sum_{x=0}^{a-1} C_N^x p^x q^{N-x}+\sum_{x=b+1}^{N} C_N^x p^x q^{N-x}.$$

Тепер потрібну нам імовірність обчислюємо за формулою

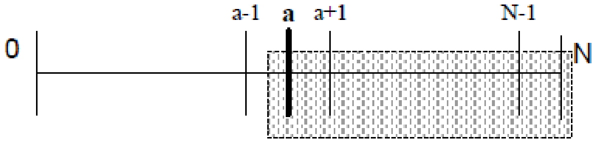
$$P_N(a \le x \le b)=1-P\left(\overline{B'_x}\right)=1-\sum_{x=0}^{a-1} C_N^x p^x q^{N-x}-\sum_{x=b+1}^{N} C_N^x p^x q^{N-x}. \tag{14}$$

Графічно такий підхід можна інтерпретувати наступним чином:



Розглянемо деякі часткові випадки. Припустимо, що необхідно визначити ймовірність того, що деяка лінгвістична одиниця A зустрінеться не менше a разів. Тут

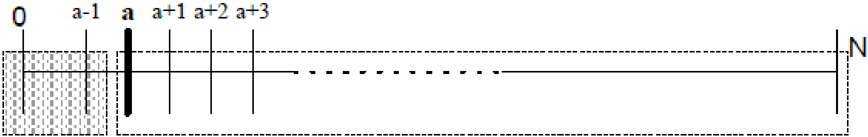
$$P_N(x \geq a) = \sum_{x=a}^{N} C_N^x p^x q^{N-x} \ .$$



Якщо значення a мале, то доцільно скористатись виразом

$$P_N(x \ge a) = 1 - \sum_{x=0}^{a-1} C_N^x p^x q^{N-x},$$

який є частинним випадком формули (2.14).

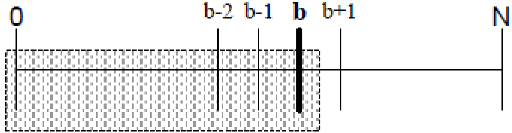


У тому випадку, коли $a = 1$, маємо

$$P_N(1 \le x \le N) = 1 - C_N^0 p^0 q^N = 1 - q^N. \tag{15}$$

Імовірність появи події A не більше b разів також визначається шляхом сумування ймовірностей, у яких подія появляється 0, 1, 2, ..., b разів:

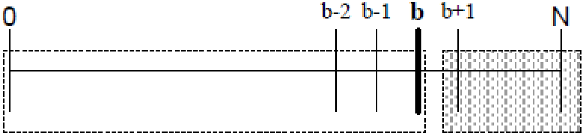
$$P_N(x \le b) = \sum_{x=0}^{b} C_N^x p^x q^{N-x} .$$



Якщо значення b близьке до N, то цю ймовірність доцільно обчислювати за такою формулою:

$$P_N(x \le b) = 1 - \sum_{x=b+1}^{N} C_N^x p^x q^{N-x}, \qquad (16)$$

яка також є частинним випадком формули (2.14).



4 ВИЗНАЧЕННЯ НЕОБХІДНОГО ОБ'ЄМУ ВИБІРКИ

У лінгвістичних дослідженнях і особливо при підготовці лінгвістичних програм машинного перекладу та інформаційного пошуку постійно виникає потреба визначати об'єм вибірки, необхідний для того, щоб забезпечити із заданою ймовірністю появу хоча б один раз потрібної лінгвістичної одиниці.

Для цього перетворимо спочатку формулу $P_N(1 \le x \le N) = 1 - q^N = 1 - (1-p)^N$ до виду $(1-p)^N = 1 - P_N(1 \le x \le N)$. Прологарифмуємо обидві частини рівності і після нескладних перетворень одержимо

$$N = \frac{\lg[1 - P_N(1 \le x \le N)]}{\lg(1-p)}, \qquad (17)$$

де N вказує на необхідний об'єм вибірки.