**Stepik Course Part 2**

**Module 1 Аналіз номінативних даних**

**Lecture 1 Номінативні дані**

**Номінативні змінні** – використовуються для розподілу на групи, при цьому цифри не несуть математичного змісту, наприклад у вибірці чоловіки отримають цифру 1, а жінки 2.

**Рангові** (частковий випадок якісної змінної) – розподіл вибірки по рангу, наприклад результат забігу.

**Логістична регресія** – вид аналізу, при якому **залежною змінною** являється **номінативна змінна**.

Для номінативних даних можна знайти тільки **моду –** значення, яке зустрічається найчастіше (може бути декілька мод, якщо однакова кількість).

Основні типи гіпотез при роботі з номінативними даними:

* **перевірка гіпотези про розподіл номінативної змінної** – наскільки спостережуваний розподіл відрізняється від очікуваного;
* **перевірка гіпотези про взаємодію двох номінативних змінних** – чи є взаємозв’язок між двома номінативними змінними;

**Lecture 2 Відстань Хі-квадрат Пірсона**

Для перевірки гіпотези про відмінність певного розподілу частот від теоретичного використовуються наступні характеристики:

* **відстань Хі-квадрат Пірсона**;
* **розподіл Хі-квадрат Пірсона**;
* **критерій Хі-квадрат Пірсона**;

**O (observe)** – позначається частота номінативного значення, яка спостерігаються;

**E (expected)** – частота значення, яка очікується.

**Відстань Хі-квадрат Пірсона** –характеризує, наскільки сильно певний розподіл частот відрізняється від теоретичного,рахується по формулі:

**Lecture 3 Розподіл Хі-квадрат Пірсона**

Якщо вірна **нульова гіпотеза** (розподіл спостережуваної номінативної змінної дорівнює очікуваному розподілу), то розподіл різниці між очікуваним і спостережуваним значенням (відхилення) буде **нормальним розподілом** з середнім µ = 0 і D = 1.

**Розподіл Хі-квадрат з k-ступенями свободи** – розподіл суми квадратів k незалежних стандартних нормальних випадкових величин (µ = 0 і D = 1).

При малих значення ступенів свободи (k) найбільша кількість значень близька до нуля і далі різко спадає.

При великому k розподіл Хі-квадрат буде прямувати до нормального розподілу.

**k (df) ступені свободи** – число ймовірних значень номінативної змінної мінус 1, рахується по формулі: k = n – 1

**Lecture 4 Розрахунок p-рівня значимості**

Для розрахунку p-рівня значимості потрібно порахувати значення хі-квадрат, далі скористатись онлайн калькулятором:

<https://gallery.shinyapps.io/dist_calc/> - поставити «Distribution» в положення «Chi-Squared**»**, зазначити число степенів свободи (Degrees of freedom) і значення пораховане хі-квадрат в «а».

Якщо значення p-рівня значимості менше 0.05, то можна відхилити нульову гіпотезу і прийняти альтернативну.

**Lecture 5 Аналіз таблиць спряженості**

**Критерій хі-квадрат** дозволяє перевірити гіпотезу про взаємодію двох номінативних змінних (наприклад хворі, які приймали препарат А виздоровіли, а хворі, які приймали препарат В - ні).

**Нульова гіпотеза** – дві номінативні змінні не зв’язані між собою.

**Альтернативна гіпотеза** – дві номінативні змінні зв’язані між собою.

Для перевірки зв’язку між номінативними змінними потрібно:

* приймаємо нульову гіпотезу і заповнюємо таблицю спряженості при очікуваних значеннях;
* заповняємо таблицю спряженості спостережуваними значеннями;
* рахуємо хі-квадрат для даної таблиці;

Очікувані значення беруться наступним чином:

* зафіксовуємо одну з номінативних змінних за основу (яку конкретно не має значення – формула симетрична);
* з урахуванням відношення, яке вийшло, рахуємо які значення мали б бути у другої номінативної змінної (наприклад дві змінні - стать і професія; 26 чоловіків (63.4%) і 15 жінок (36.6%), тоді таке ж відношення чоловіків до жінок мало б бути у професіях; рахуємо 63.4% чоловіків для обох професій і відповідно 36.6% для жінок);

Для таблиць спряженості df = (n - 1)·(m - 1), де n – кількість стовбців в таблиці, m – кількість строк.

**Поправка Йейтса** – корегує значення p-рівня значимості. Зазвичай, застосовується, коли частота (значення) певної номінативної змінної менше 10.

**Вимоги до таблиць спряженості**:

* всі спостережувані змінні мають бути незалежні;
* мінімальна кількість спостережень (частота) в кожній комірці має бути більше 5;

**Аналіз залишків** дозволяє оцінити, які конкретно частоти значимо відхиляються від очікуваних. Для цього потрібно побудувати **mosaic** графік. Даний графік візуально відображає різницю між спостережуваними частотами і очікуваними. Ширина стовбців на графіку показує кількість спостережень для кожної змінної.

**Lecture 6 Точний критерій Фішера**

**Точний критерій Фішера** – застосовується на малих вибірках, де частоти певних величин менше 5 і відповідно не можна застосовувати критерій хі-квадрта.

Точний критерій Фішера розраховує ймовірність отримати спостережувані або ще більш виражені відхилення. Даний метод можна використовувати і на великих вибірках.

**Module 2 Логістична регресія і непараметричні методи**

**Lecture 7 Логістична регресія**

**Логістична регресія** – метод пошуку взаємозв’язку між номінативною залежною змінною, яка має тільки дві градації (два можливих значення) і незалежною змінною (предиктором), яка може бути як кількісна так і номінативна. Також можливі варіанти, при яких незалежні змінні (предиктори) різних типів.

Логістична трансформація – перетворення ймовірності діапазону [0, 1] в діапазон [-∞, +∞].

**Шанс** (odds) – відношення ймовірності успіху (Y=1) до ймовірності невдачі (Y=0). Може приймати значення [0, +∞].

Якщо значення odds:

* ≈ 1 то шанс на усміх і невдачу рівний;
* > 1 то шанс на успіх більший;
* < 1 то шанс на невдачу більший;

**Логарифм шансів** трансформує значення ймовірностей з діапазону [0, +∞] в діапазон [-∞, +∞].

Якщо значення (натурального) логарифма:

* < 0 то значення шансів менше 1
* > 0 то значення шансів більше 1
* = 0 то значення шансів рівне 1

Великі значення (натурального) логарифма, означають, що ймовірність позитивного результату приближається до 1.

**Lecture 8 Модель без предикторів. Intercept only model**

**Intercept only model** - передбачаємо шанс, використовуючи тільки одне число (intercept). В моделі без предиктора, intercept – це натуральний логарифм шансу позитивного результату по всій вибірці. Взявши експоненту від intercept, отримаємо шанс (odds). Далі можна визначити ймовірність.

Така регресійна модель (intercept only model) вирішує задачу про розподіл частот певної ознаки – наскільки спостережуваний розподіл відрізняється від очікуваного (рівномірного) і показує який шанс більший (позитивний чи негативний). У цьому основна відмінність від хі-квадрата, який покаже тільки, чи відрізняється розподіл від очікуваного.

**Lecture 9 Модель з одним номінативним предиктором**

Спочатку потрібно побудувати таблицю спряженості для залежного і незалежного параметрів. Це дає змогу визначити ймовірності, шанси і їх логарифми для залежного і незалежного параметрів.

При проведенні логістичної регресії в пакетах статистики (R, python і т.д.) предиктор автоматично розділяється на два фактора (наприклад предиктор стать «Sex» ділиться на два «SexMale» і «SexFemale», кожен зі значеннями 0 і 1, відповідно чоловіки отримають 1 в «SexMale» і 0 в «SexFemale», а жінки навпаки). Результатом логістичної регресії будуть значення Intercept і певного коефіцієнту (фактору «SexMale» або «SexFemale», зазвичай вибирається по алфавіту).

Фактор якого немає серед коефіцієнтів і буде Intercept (натуральний логарифм шансів позитивного результату для даного фактору).

Інший коефіцієнт – це натуральний логарифм відношення шансів позитивного результату для фактора, який є серед коефіцієнтів до шансів позитивного результату фактора, якого немає серед коефіцієнтів (intercept). Щоб отримати логарифм шансів позитивного результату для фактора, який є серед коефіцієнтів, потрібно цей коефіцієнт добавити до значення Intercept.

В загальному нас цікавить чи являється певний предиктор статистично значимим, тобто чи дозволяє модель з предиктором краще передбачати значення залежної змінної ніж нульова модель (без предиктора).

Для цього нам потрібно визначити якість наших моделей, тобто порівняти модель без предиктора з моделями з предиктором:

* будуємо модель без предикторів;
* будуємо моделі з предиктором;
* порівнюємо моделі з предиктором із нульовою моделлю (порівняння проводиться по параметру **Deviance** в результатах логічної регресії - чим нижче значення, тим краще);

**Lecture 10 Модель з двома номінативними предикторами**

Як і у випадку моделі з одним номінативним предиктором, після проведення логістичної регресії, шукаємо яких факторів немає серед коефіцієнтів (наприклад серед коефіцієнтів є Sex[T.male], Pclass[T.Second], Pclass[T.Third], значить відсутні Sex[T.female] і Pclass[T.First], тобто intercept в даному випадку – це «жінки в першому класі»). Intercept – це натуральний логарифм шансів позитивного результату для даного фактору (в даному випадку – це логарифм шансів вижити для жінок в першому класі).

Всі інші коефіцієнти, це значення переходів, які являються собою порівняння шансів базового рівня (шансів вижити для жінок в першому класі) з усіма іншими (в даному випадку:

* Sex[T.male] – це логарифм відношення шансів вижити для чоловіків в першому класі, до шансів вижити жінок в першому класі;
* Pclass[T.Second] – логарифм відношення шансів вижити для жінок в другому класі до шансів вижити для жінок в першому;
* Pclass[T.Third] – логарифм відношення шансів вижити для жінок в третьому класі до шансів вижити для жінок в першому;).

**Lecture 11 Взаємодія номінативних предикторов**

Взаємодія двох номінативних змінних показує зв'язок між одним предиктором і залежною змінною в залежності від другого предиктора (в даному випадку показує як стать впливає на те, чи пасажир вижив в залежності від класу білету:

* Коефіцієнт Sex[T.male]:Pclass[T.Second] – різниця логарифмів відношень шансів вижити чоловіків до жінок в другому і першому класах;
* Коефіцієнт Sex[T.male]:Pclass[T.Third] - різниця логарифмів відношень шансів вижити чоловіків до жінок в третьому і першому класах).

**Lecture 12 Коли потрібно використовувати непараметричні методи?**

**t-критерій Стьюдента** – порівняння двох вибірок між собою (по середнім значенням). Вимоги до t-тесту:

* **незалежність** спостережень;
* **гомогенність** дисперсії (бажано);
* **нормальний розподіл** досліджуваної ознаки в генеральній сукупності;
* **мінімальний об’єм** вибірки (n < 30);

**Перевірка на нормальність** – найбільш часто використовуваний тести Shapiro-Wilk та Колмогорова-Смирнова:

* **Нульова гіпотеза**: вибірка взята з нормально розподіленої генеральної сукупності;
* **Альтернативна гіпотеза**: вибірка взята з НЕ нормально розподіленої генеральної сукупності;

Часто перед порівнянням двох груп або при дисперсійному аналізі відкидаються з розподілу значення, які більші за ±2sd (стандартні відхилення).

При дуже великий об’ємах вибірок t-розподіл буде нормальним навіть, якщо генеральна сукупність має НЕ нормальний розподіл (відхилення в певну сторону або навіть однорідний розподіл).

Навіть при об’ємі вибірки n > 30 потрібно перевірити нормальність розподілу досліджуваної ознаки.

Найпопулярнішим непараметричним критерієм для порівняння двох груп є – **критерій Манна - Уїтні**. Замість порівняння середніх значень у двох вибірках критерій порівнює суму рангів (НЕ медіани).

Критерій Манна - Уїтні варто застосовувати:

* **розподілу** хоча б в одній з вибірок значно **відрізняється від нормального**;
* присутні помітні **викиди** в даних.
* у деяких завданнях потужність тесту навіть вище, ніж t критерію (наприклад, коли обох вибірках спостерігається **помітна асиметрія** в однаковому напрямку).

Критерій Манна - Уїтні НЕ варто застосовувати на вибірках різного розміру, з різним напрямком асиметрії.

**Критерій Краскела-Уоллеса** – непараметричний аналог дисперсійного аналізу. Порівнює декілька груп по дисперсіям середніх значень рангів в цих групах. Якщо вірна нульова гіпотеза, то розподіл даної дисперсії можна описати за допомогою розподілу Хі-квадрат. Якщо дисперсія велика, то спостереження відрізняються.

Критерій Краскела-Уоллеса варто застосовувати:

* **розподілу** вибірок суттєво **відрізняється від нормального**;
* **НЕ гомогенна дисперсія**;

Зі збільшенням обсягу вибірки ймовірність того, що p-значення при перевірці на нормальність розподілу виявиться менше 0,05 збільшується.

**Module 3 Кластерний аналіз і метод головних компонент**

**Lecture 13 Кластерний аналіз методом k-середніх**

**Unsupervised learning** (навчання без вчителя) – метод навчання в якому відсутня цільова ознака.

**Кластеризація** – розбиття початкового набору об’єктів на декілька підмножин об’єктів, які складаються з схожих елементів.

**k-means** – метод кластерного аналізу, який розділяє n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Робота алгоритму:

1. Вибрати кількість кластерів, яка здається оптимальною для наших даних;
2. Випадково вибрати початкові позиції центроїді кластерів;
3. Для кожної точки нашого набору даних порахувати, до якого центроїду вона ближче;
4. Перемістити кожен центр ваги в центр вибірки, яку ми віднесли до цього центроїду
5. Обновляємо позиції центроїдів і належність точок до кластерів;

Позиція центроїда кластера – середнє для всіх значень в кластері (а НЕ геометричний центр).

<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/> - візуалізація роботи алгоритму k-means.

**Lecture 14 Помилки кластерного аналізу**

В деяких випадках алгоритм k-means зійдеться не оптимально. Алгоритм знайде локальний мінімум (який відповідає заданим умовам – при черговій зміні позицій ні одна з точок не змінить своєї відповідності) і сприйме його за глобальний мінімум.

Для вирішення проблеми локальних мінімумів існують різні рішення, але найтиповішими варіантами є:

1. розміщення центроїдів не випадково, а максимально далеко один від одного;
2. проведення кластерного аналізу велику кількість разів, і якщо центроїди приходять в одні і ті ж точки, то в даних є певна структура (кластери), а якщо центроїди кожен раз займають інші позиції, то кластерів немає;

**Lecture 15 Вибір оптимально числа кластерів**

**Внутрішньо групова сума квадратів** (**within cluster sum of squares**) – сума квадратів відхилень кожного спостереження від центра кластера.

**Загальна внутрішньо групова сума квадратів** (**total** **within cluster sum of squares**) – сума всіх внутрішньо групових сум квадратів.

Для знаходження оптимального числа кластерів потрібно по черзі починаючи з двох кластерів (k = 2) проводити кластеризацію і розраховувати загальну внутрішньо групову суму квадратів. Якщо додавання ще одного кластера (k) сильно понижає значення загальної внутрішньо групової суми квадратів, а додавання наступного кластера (k + 1) не дає відчутних змін, то варто зупинитись на цьому значення (k).

Якщо побудувати графік зміни загальної внутрішньо групової суми квадратів від числа кластерів, то оптимальне число кластерів можна визначити по зламі кривої. Якщо графік плавно опускається, то в даних немає явно вираженої структури.

Після знаходження оптимального числа кластерів, варто добавити в дані нову змінну – номер кластера і порівняти кластери між собою по наявним даним.

**Lecture 16 Ієрархічна кластеризація**

**Ієрархічна кластеризація** – сукупність методів кластеризації. Алгоритм роботи даних методів:

* вибираємо кожну точку, як центр свого кластера;
* рахуємо відстань між точками (зазвичай – Евклідову відстань);
* об’єднуємо найближчі попарно;
* повторюємо попередній крок, доки всі точки не об’єднаються в один кластер;

**Дендрограмма** – це візуалізація ієрархічної кластеризації. На такому графіку добре видно, чи є в даних певна структура і якщо є, то на скільки кластерів варто ділити дані.

Ієрархічна кластеризація діляться на два типа:

1. **Агломеративні методи** – нові кластери створюються шляхом об'єднання дрібніших кластерів, нагадує дерево від листків до стовбура;
2. **Дивізіонні методи** (**divisive –** розбіжність) – нові кластери створюються шляхом ділення більших кластерів на більш дрібні, нагадує дерево від стовбура до листків;

Методи кластеризації (ієрархічна кластеризація і k-means) можна комбінувати між собою. Також можна поділити вибірку на дві частини і в кожній застосувати різний метод кластеризації.

**Lecture 17 Введення в метод головних компонент**

**Метод головних компонент** (**principal components analysis**) – метод зниження розмірності даних, шляхом проекції даних, які мають лінійну кореляцію на пряму регресії. При зменшенні розмірності ми втрачаємо частину даних. Чим більший коефіцієнт кореляції, тим менше інформації втрачається при такому переході. Уся страчена інформація зберігається в другій головній компоненті (**principal component 2**), яка має вигляд прямої, яка перпендикулярна першій головній компоненті (прямій регресії).

Розмірність до якої потрібно зменшуватись, як правило вибирають, щоб зберегти 90% дисперсії в даних.

Графік **biplot** – відображає спостереження в новому просторі. По x – principal component 1, а по y – principal component 2.

Якщо в даних присутня сильна **мультиколінеарність**, то можна такі предиктори **об’єднати методом головних компонент**.

**Факторний аналіз** – метод зниження розмірності даних. Ідея методу – деякі змінні можуть бути згруповані в **фактори** (наприклад характеристики вага, зріст, сила, знання англійської, німецької та російської мови можна виділити в два фактори: фізична підготовка і знання мов).