



Факультет компьютерных наук

НИС Python

Москва 2025

Лекция 2

Исследование взаимосвязей

Преподаватель: Меликян Алиса Валерьевна, amelikyan@hse.ru
кандидат наук, доцент Департамента программной инженерии

Взаимосвязи между переменными

При изучении взаимосвязи между парой переменных одна из них считается независимой (или объясняющей), а вторая – зависимой (или объясняемой). Качество модели взаимосвязи переменных может быть оценено посредством коэффициентов связи.

В зависимости от задачи исследования, типов шкал переменных и полноты имеющихся данных необходимо выбрать подходящий метод исследования взаимосвязи.

Таблица сопряжённости

Связь между категориальными переменными, то есть переменными, относящимися к номинальной шкале или порядковой шкале с не очень большим количеством категорий, лучше всего представить в форме таблиц сопряжённости. Таблицы сопряжённости позволяют выявить наличие статистических, а не причинно-следственных зависимостей.

Таблица сопряжённости с частотами

Content Rating	prime_genre_upd				Total
	Games	Entertain	Education	Other	
12+	741	108	8	298	1,155
17+	177	98	7	340	622
4+	2,079	285	432	1,637	4,433
9+	865	44	6	72	987
Total	3,862	535	453	2,347	7,197



Таблица сопряжённости с частотами и процентами по столбцам

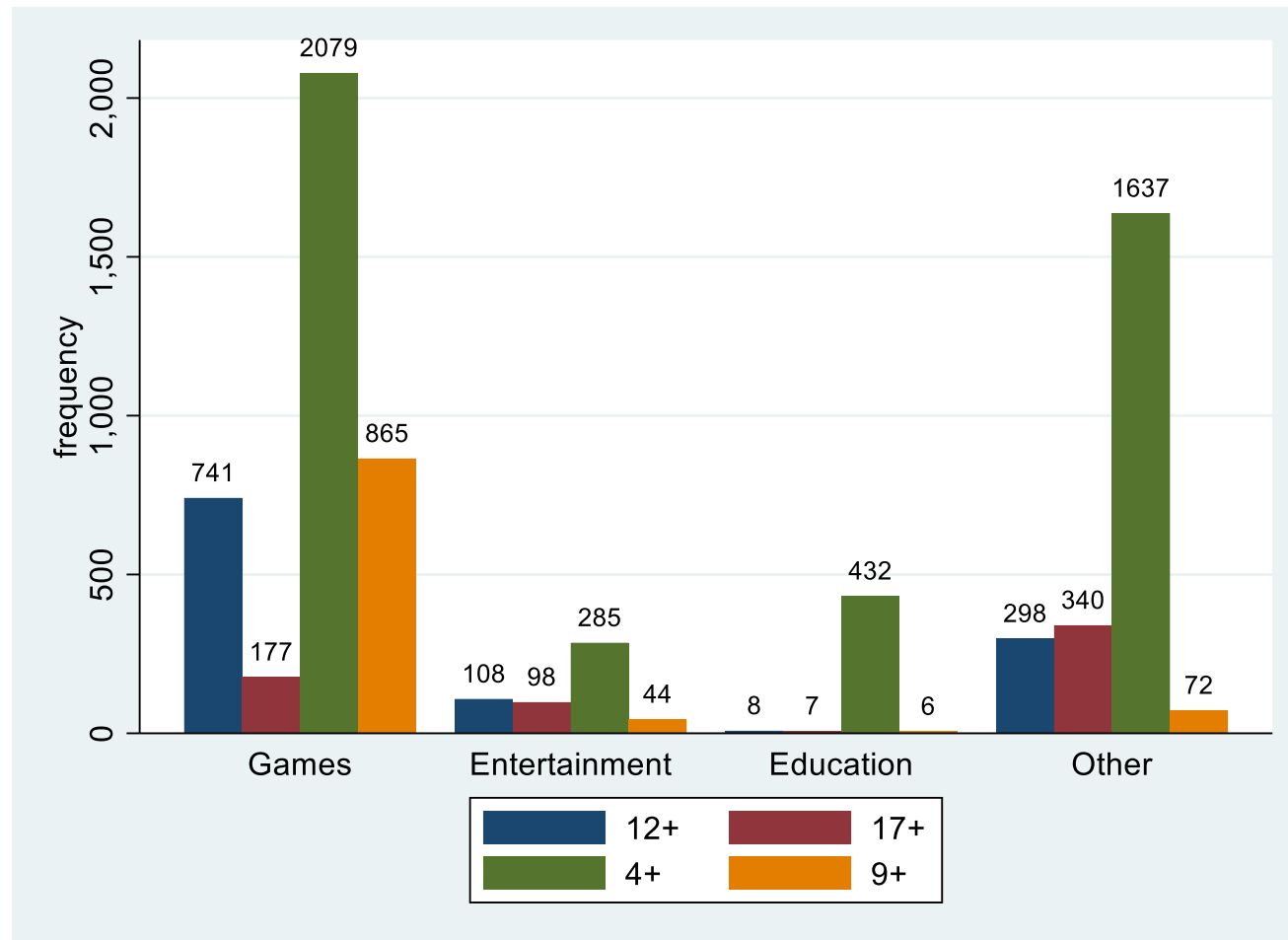
Content Rating	prime_genre_upd				Total
	Games	Entertain	Education	Other	
12+	741	108	8	298	1,155
	19.19	20.19	1.77	12.70	16.05
17+	177	98	7	340	622
	4.58	18.32	1.55	14.49	8.64
4+	2,079	285	432	1,637	4,433
	53.83	53.27	95.36	69.75	61.60
9+	865	44	6	72	987
	22.40	8.22	1.32	3.07	13.71
Total	3,862	535	453	2,347	7,197
	100.00	100.00	100.00	100.00	100.00



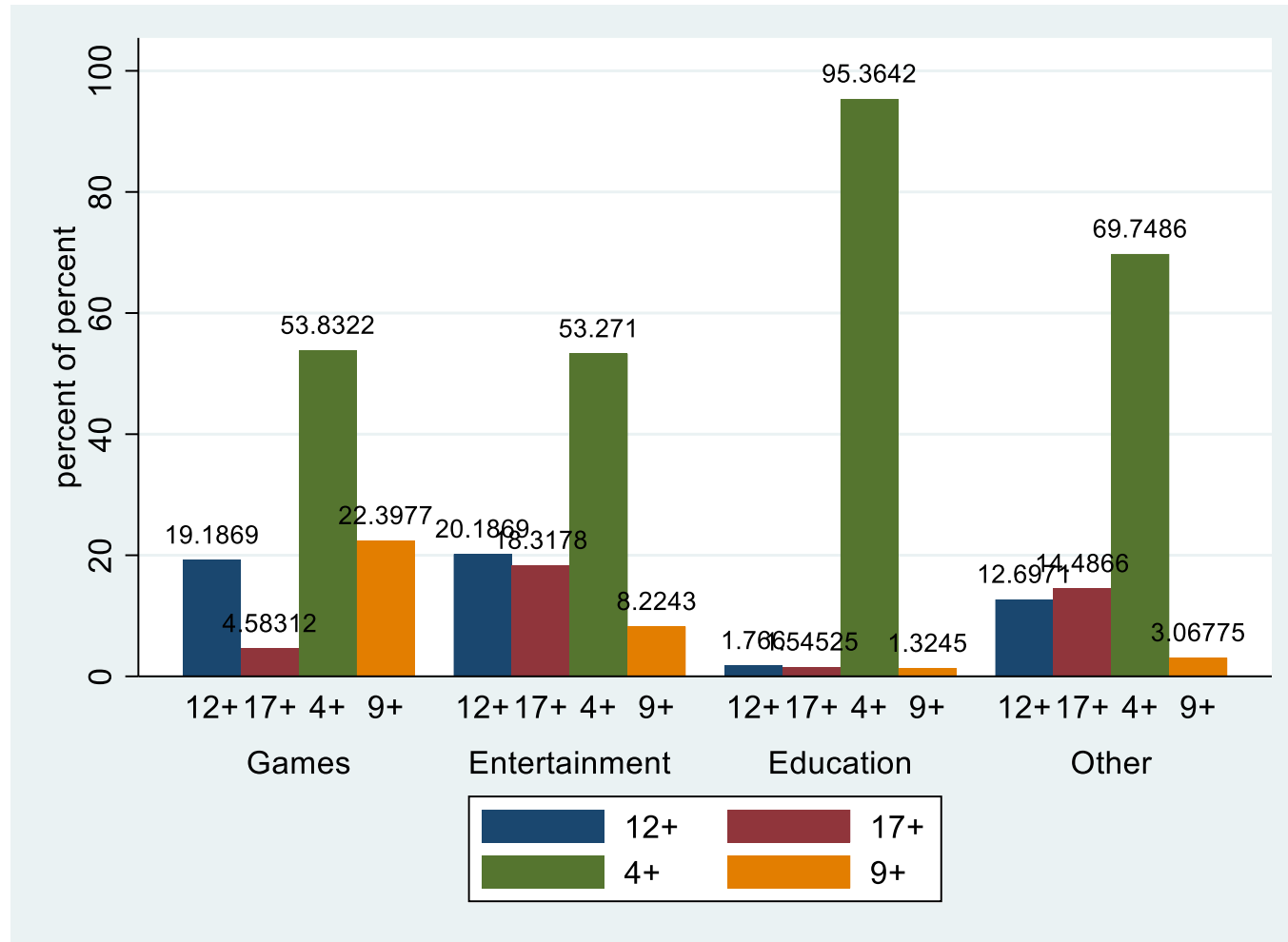
Таблица сопряжённости с частотами и процентами по строкам

Content Rating	prime_genre_upd				Total
	Games	Entertain	Education	Other	
12+	741 64.16	108 9.35	8 0.69	298 25.80	1,155 100.00
17+	177 28.46	98 15.76	7 1.13	340 54.66	622 100.00
4+	2,079 46.90	285 6.43	432 9.75	1,637 36.93	4,433 100.00
9+	865 87.64	44 4.46	6 0.61	72 7.29	987 100.00
Total	3,862 53.66	535 7.43	453 6.29	2,347 32.61	7,197 100.00

Столбчатая диаграмма с частотами и группировкой



Столбчатая диаграмма с процентами и группировкой



Статистика

Описательная

Описание данных с использованием графиков, частотных таблиц, мер средней тенденции и разброса.

Инференциальная (дедуктивная)

Позволяет установить, в какой мере взаимосвязи, выявленные в выборке, описывают характеристики генеральной совокупности (насколько они статистически значимы). Инструменты дедуктивной статистики используют теорию вероятности для оценки вероятности того, что каждая закономерность или тренд, который вы наблюдаете в данных, не являются случайными.

Формулировка гипотез

Гипотеза – это предположение, выдвигаемое для объяснения каких-либо явлений. По итогам проведённого исследования гипотеза может быть принята или отвергнута.

Нулевая гипотеза (H_0) – принимаемое по умолчанию предположение о том, что не существует связи между двумя наблюдаемыми событиями, феноменами. Она считается верной пока нельзя доказать обратное.

Альтернативная гипотеза (H_1) – это утверждение, являющееся логическим отрицанием нулевой гипотезы. Альтернативная гипотеза предполагает существование взаимосвязи или отличия между изучаемыми переменными.

Нулевая гипотеза

H_0 : средние значения в двух группах не различаются/
не наблюдается взаимосвязи между переменными.

Любые наблюдаемые отличия или взаимосвязи случайны.

Альтернативная гипотеза

H1: средние значения в двух группах различаются / наблюдается взаимосвязи между переменными.

Наблюдаемые отличия или взаимосвязи не случайны.

Этапы проверки гипотез

1. Генерация гипотезы (или гипотез) - предположение о том, что некоторые тенденции имеют место в генеральной совокупности.
2. Сбор необходимых данных.
3. Сопоставление статистической модели с реальными данными – модель позволит проверить изначальные предположения.
4. Оценка того, поддерживает ли модель первоначальные предположения.

Результат проверки гипотез:

H_0 отвергается, H_1 принимается;

H_1 отвергается, H_0 не отвергается, но это не значит, что она принимается.

Подробнее про этапы проверки гипотез

1. Формулировка нулевой гипотезы (H_0) и альтернативной гипотезы (H_1).
2. Выбор подходящего статистического теста для проверки гипотезы.
3. Выбор уровня значимости (альфа). Как правило – 5%.
4. Вычисление значения выборочной статистики.
5. На основе определённых теоретических распределений (t-распределение, F-распределение, распределение Хи-квадрат, распределения Стьюдента) с учетом числа наблюдений и заданного уровня значимости определение критического значения статистики, которое делит интервал на область принятия и отвержения нулевой гипотезы.
6. Определение того, попадает ли выборочное значение статистики, полученное на этапе 4, в область принятия или отклонения нулевой гипотезы.
7. Принятие статистического решения о том принимается или отвергается нулевая гипотеза.
8. Интерпретация полученного статистического решения с точки зрения проблемы исследования (статистически значимый результат не всегда значим с содержательной точки зрения).

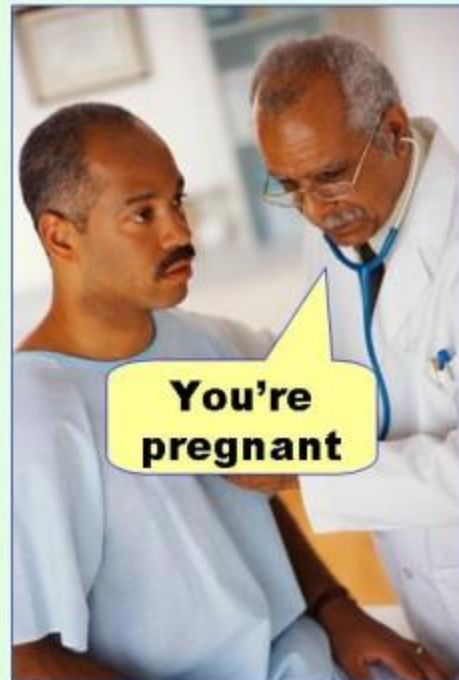
Уровень значимости и ошибка I-го рода

	H_0 верна	H_0 не верна
H_0 принимается	H_0 верно принята	H_0 неверно принята (ошибка II-го рода)
H_0 не принимается	H_0 неверно отвергнута (ошибка I-го рода)	H_0 верно отвергнута

Уровень значимости – вероятность ошибки I-го рода. Выбирается исследователем заранее, обычно 5% (0,05). Обозначается Sig. (significance) value, p-value, Pr.

Ошибки I-го и II-го рода

Type I error
(false positive)



Type II error
(false negative)



P-значение (p-value)

Это число от 0 до 1. Показывает насколько мы можем быть уверены в том, что наша экспериментальная гипотеза (H_1) верна. Чем ближе значение p-value к 0, тем больше у нас уверенности в том, что H_1 истинна. Вопрос в том, насколько маленьким должно быть значение p , чтобы мы были достаточно уверены в истинности H_1 , какой порог мы можем использовать, чтобы принять правильное решение?

Обычно используется порог 0,05 (уровень значимости). Это означает, что если мы проведем один и тот же анализ/эксперимент на других выборках, то только 5% этих экспериментов дадут противоположный результат.

Уровень значимости (level of significance)

Если очень важно получить максимально надёжный результат, то можно использовать меньший порог, например 0,00001. Это значит, что мы будем не правы только один раз в 100 000 экспериментов.

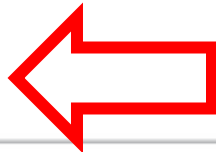
Если мы допускаем большую вероятность ошибки, например, мы прогнозируем, прибудет ли автобус вовремя, тогда мы можем использовать больший порог, например 0,2. Это значит, что мы не будем правы только 2 раза из 10. Самый распространенный порог – 0,05.

	Rate	Log Frequency	Synonyms	Mutual-Information	Imageability	Arousal	Log Senses	Log AgeOfAcq
Rate	-	-.273**	.242*	-.281*	-.254*	.029	-.046	.255*
LogFrequency	-.273**	-	.381***	.221	-.208	-.046	.357***	-.482***
Synonyms	.242*	.381***	-	-.003	-.486***	.195	.592***	-.043
Mutual-Information	-.281*	.221	-.003	-	.506***	-.111	.031	-.367***
Imageability	-.254*	-.208	-.486***	.506***	-	-.071	-.369***	-.238
Arousal	.029	-.046	.195	-.111	-.071	-	.046	.097
LogSenses	-.046	.357***	.592***	.031	-.369***	.046	-	-.178
LogAgeOfAcq	.255*	-.482***	-.043	-.367***	-.238	.097	-.178	-

***: $p < .0001$

** : $p < .01$

* $p < .05$.



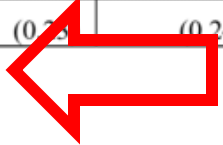
doi:10.1371/journal.pone.0147924.t002





	CONTRI- BUTION1	BELIEF1	MEETING DUM	BIOGAS DUM	ANYTREE CF	NUMTREES FARM	MONITOR ING	NUMTREES CF
CONTRI- BUTION1	1.00							
BELIEF1	0.54 (0.00)***	1.00						
MEETING DUM	0.01 (0.92)	-0.03 (0.64)	1.00					
BIOGAS DUM	0.01 (0.79)	0.02 (0.76)	0.00 (0.99)	1.00				
ANYTREE CF	0.09 (0.10)*	0.07 (0.23)	0.07 (0.20)	-0.01 (0.86)	1.00			
NUMTREES FARM	0.09 (0.09)*	0.21 (0.00)***	0.18 (0.00)***	0.20 (0.00)***	0.26 (0.00)***	1.00		
MONITOR- ING	0.01 (0.90)	0.01 (0.91)	0.14 (0.01)***	0.08 (0.13)	0.42 (0.00)***	0.28 (0.00)***	1.00	
NUMTREES CF	0.09 (0.10)*	0.07 (0.23)	0.07 (0.24)	0.00 (0.93)	1.00 (0.00)***	0.26 (0.00)***	0.42 (0.00)***	1.00

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$



Проверка гипотез

H1: существует взаимосвязь между посещаемостью и успеваемостью студентов

- $p\text{-value} = 0.0000001 \rightarrow H1$ принимается на уровне значимости 1%
- $p\text{-value} = 0.049 \rightarrow H1$ принимается на уровне значимости 5%
- $p\text{-value} = 0.051 \rightarrow H1$ отвергается на уровне значимости 5%, но принимается на уровне значимости 10%
- $p\text{-value} = 0.73 \rightarrow H1$ отвергается на уровне значимости 10%

Тест Хи-квадрат

При проведении теста хи-квадрат проверяется взаимная независимость двух переменных таблиц сопряжённости и благодаря этому косвенно выясняется зависимость обоих переменных. Две переменные считаются взаимно независимыми, если наблюдаемые в ячейках частоты совпадают с ожидаемыми.

Критерий Хи-квадрат основан на предположении о том, что если признаки независимы, то частоты в клетках таблицы сопряжённости распределены пропорционально.

Тест Хи-квадрат: сценарий 1

Полное отсутствие взаимосвязи между переменными.

Group * Counsellor Crosstabulation

			Counsellor		Total
			John	Jane	
Group	male	Count	50	50	100
		Expected Count	50.0	50.0	100.0
	female	Count	50	50	100
		Expected Count	50.0	50.0	100.0
Total	Count		100	100	200
	Expected Count		100.0	100.0	200.0

Тест Хи-квадрат: сценарий 2

Сильная взаимосвязь между переменными.

Group * Counsellor Crosstabulation

			Counsellor		Total
			John	Jane	
Group	male	Count	100	0	100
		Expected Count	50.0	50.0	100.0
	female	Count	0	100	100
		Expected Count	50.0	50.0	100.0
Total	Count		100	100	200
	Expected Count		100.0	100.0	200.0

Тест Хи-квадрат: сценарий 3

sex * counsellor Crosstabulation

			counsellor		Total
			John	Jane	
sex	male	Count	10	4	14
		Expected Count	6.5	7.5	14.0
		% within sex	71.4%	28.6%	100.0%
	female	Count	4	12	16
		Expected Count	7.5	8.5	16.0
		% within sex	25.0%	75.0%	100.0%
Total	Count		14	16	30
	Expected Count		14.0	16.0	30.0
	% within sex		46.7%	53.3%	100.0%

$$7.5 = \frac{14 * 16}{30}$$

$$E_i = \frac{(\text{row total} \times \text{column total})}{\text{Table total}}$$

Расчёт показателя Хи-квадрат

Коэффициент Хи-квадрат фиксирует степень расхождения реальных и ожидаемых частот. Вычисляется по формуле:

$$X^2 = \sum_{ij} \frac{(n_{ij}^e - n_{ij}^t)^2}{n_{ij}^t}$$

n_{ij}^e - значение в ячейке таблицы сопряжённости, построенной по имеющимся данным

n_{ij}^t - значение, которое находилось бы в ячейке, если бы признаки были независимыми

$$n_{ij}^t = \frac{n_{i*} n_{*j}}{n}$$

Пример

$$\begin{aligned}(2310 \cdot 1958) / 4413 &= 1024,9 \\ (2310 \cdot 1354) / 4413 &= 708,8 \\ (2103 \cdot 1958) / 4413 &= 933,1\end{aligned}$$

<u>SEX OF STUDENT * INDEX OF STUDENTS</u> <u>VALUING CHEM Crosstabulation.</u> <u>TIMSS 2007. Russia</u>			INDEX OF STUDENTS VALUING CHEM (C-SVS)			Total
			HIGH	MEDIUM	LOW	
SEX OF STUDENT	GIRL	Count	1014	771	525	2310
		Expected Count	1024,9	708,8	576,3	2310,0
	BOY	Count	944	583	576	2103
		Expected Count	933,1	645,2	524,7	2103,0
Total		Count	1958	1354	1101	4413
		Expected Count	1958,0	1354,0	1101,0	4413,0

$$\chi^2 = \frac{(1014 - 1024,9)^2}{1024,9} + \frac{(771 - 708,8)^2}{708,8} + \frac{(525 - 576,3)^2}{576,3} + \frac{(944 - 933,1)^2}{933,1} + \frac{(583 - 645,2)^2}{645,2} + \frac{(576 - 524,7)^2}{524,7} = 21,306$$

$$df = (\text{rows} - 1) * (\text{columns} - 1) = (2 - 1) * (3 - 1) = 2$$

Chi-square Distribution Table

Пример

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69

<https://www.di-mgt.com.au/chisquare-calculator.html>

Выводы по рассмотренному примеру

H0: Нет взаимосвязи между полом учеников и оценками по химии.

H1: Есть взаимосвязь между полом учеников и оценками по химии.

В таблице статистического распределения Хи-квадрат мы можем найти, что критическое значение для степеней свободы $df=(2-1)*(3-1)=2$ и уровня значимости $p = 0,05$ составляет 5,99. Поэтому, если между переменными нет взаимосвязи, статистика хи-квадрат не должна быть больше 5,99. В нашем случае $21,3 > 5,99$, поэтому мы должны отклонить H0 и принять H1 – существует взаимосвязь между переменными.

Свойства показателя Хи-квадрат

- Значение показателя может находиться в интервале от 0 до $+\infty$. Чем больше выборка, тем больше значение (при прочих равных).
- По значению показателя нельзя определить силу и направление взаимосвязи, а также какая из переменных зависимая и какая независимая.
- Для применимости теста хи-квадрат надо убедиться в наличии достаточного количества частот в ячейках таблицы сопряжённости (не менее 95% ячеек должны содержать ожидаемую частоту больше 5).

Тесты на нормальность

Тесты Колмогорова-Смирнова (для выборок больше 50 наблюдений) и Шапиро-Уилка (для выборок от 3 до 50 наблюдений) позволяют проверить, соответствует ли реальное распределение переменной нормальному распределению, т.е. является ли различие между двумя частотными распределениями значимым или случайным.

H_0 : распределение значений переменной X не значимо отличается от нормального распределения;

H_1 : распределение значений переменной X значимо отличается от нормального распределения.

Тест Шапиро-Уилка

Рассчитывается статистика критерия Шапиро-Уилка W .

Порядок расчёта статистики:

http://www.statistics4u.info/fundstat_eng/ee_shapiro_wilk_test.html (на английском)

http://www.machinelearning.ru/wiki/index.php?title=Критерий_Шапиро-Уилка (на русском)

Тест Колмогорова-Смирнова

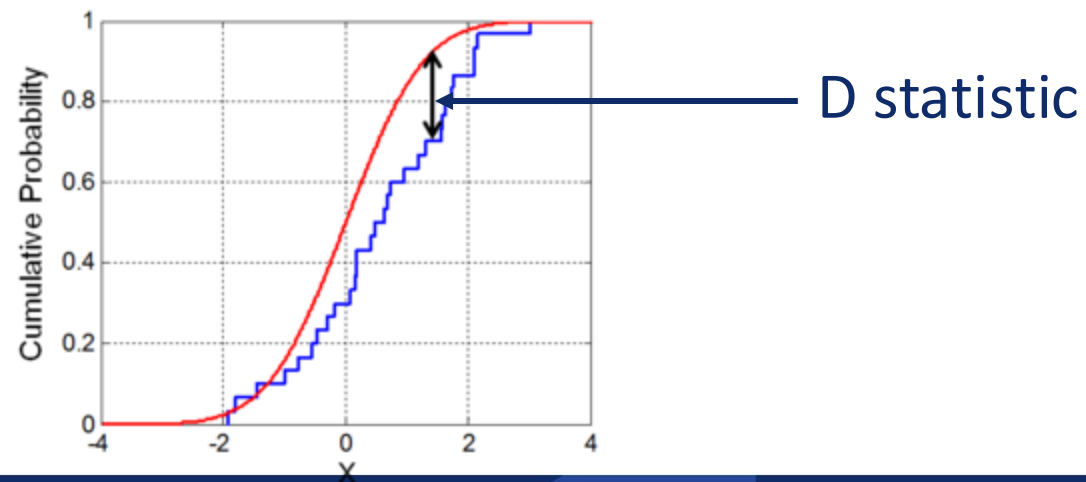
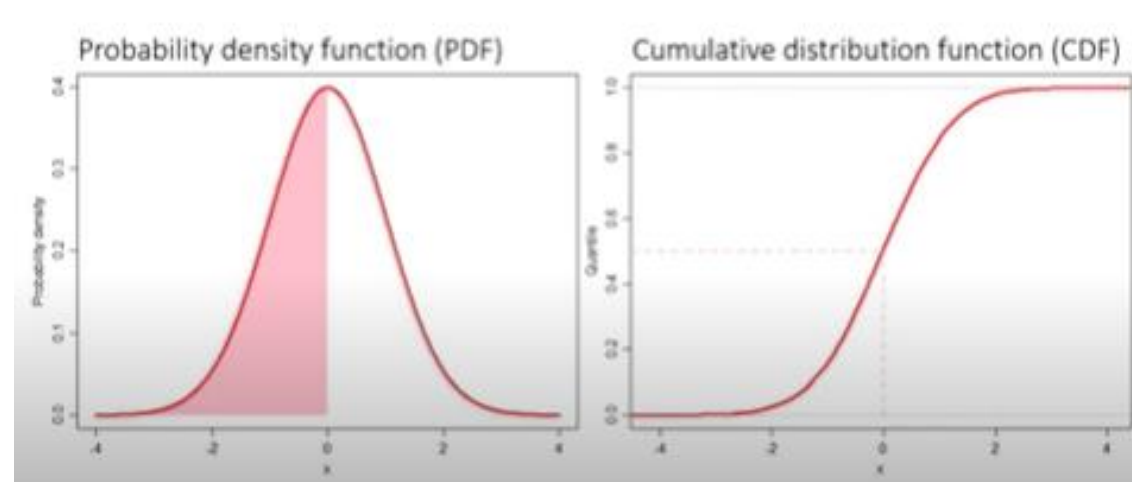


График квантиль-квантиль (Q-Q plot)

График квантиль-квантиль (Q-Q plot) – это графический метод сравнения двух распределений путем сопоставления их квантилей. Обычно используется для сравнения распределения наблюдаемых данных с теоретическим распределением, таким как нормальное распределение.

На графике Q-Q точки располагаются таким образом, что по горизонтальной оси откладываются теоретические квантили (или квантили одного из сравниваемых распределений), а по вертикальной оси — эмпирические квантили другого наблюдаемого распределения. Если данные идеально соответствуют выбранному распределению, точки будут выстроены вдоль прямой линии, часто с наклоном, зависящим от масштаба распределений.

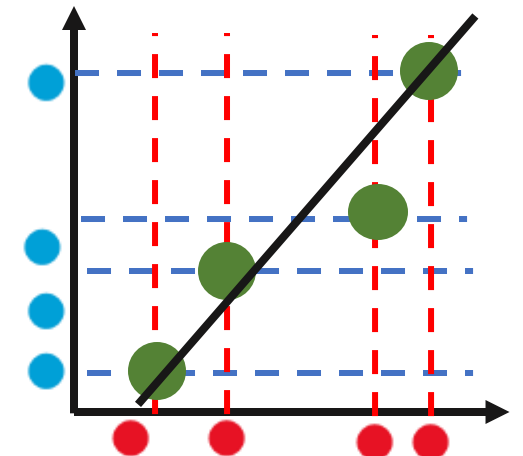
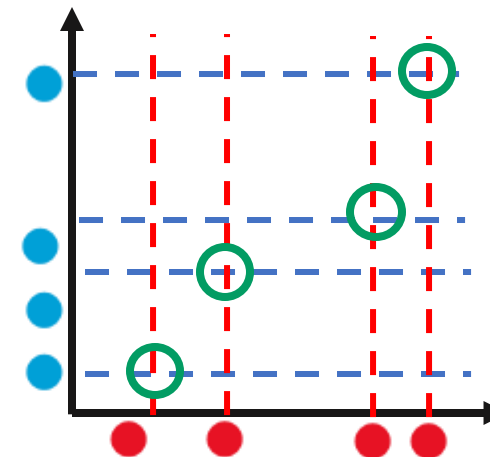
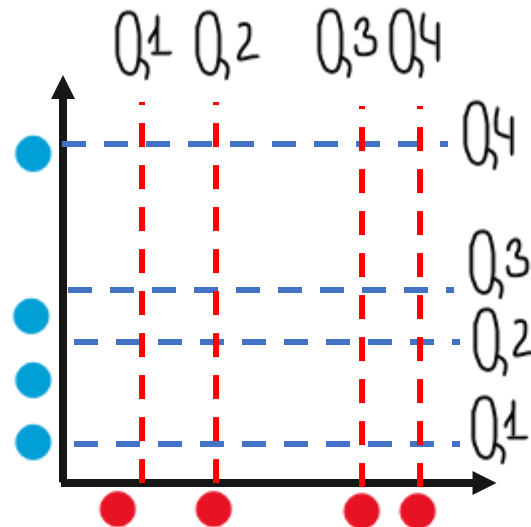
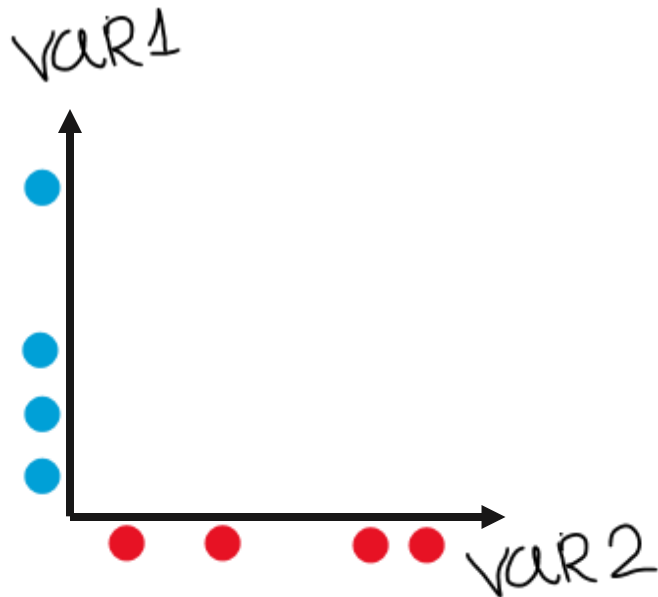
График квантиль-квантиль (Q-Q plot)

1) Расположить точки вдоль осей

2) Нарисовать квантильные линии

3) Отметить пересечение линий квантилей

4) Добавить референтную линию (например, регрессионную прямую)

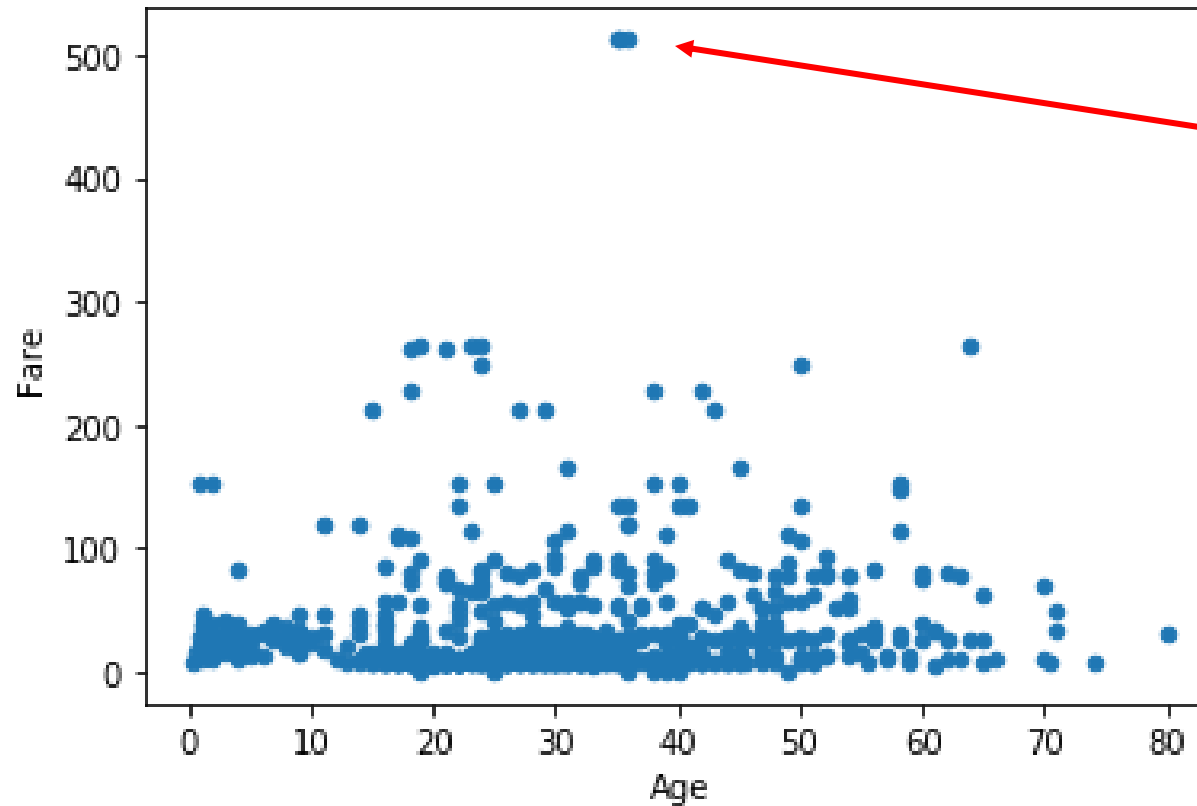


Этапы корреляционного анализа

1. Построить диаграмму рассеяния, чтобы убедиться в существовании линейной взаимосвязи между переменными;
2. Выбрать подходящий коэффициент корреляции в зависимости от особенностей переменных;
3. Рассчитать коэффициент корреляции и соответствующее ему p-value;
4. Интерпретировать статистическую значимость, силу и направление взаимосвязи.

Диаграмма рассеяния

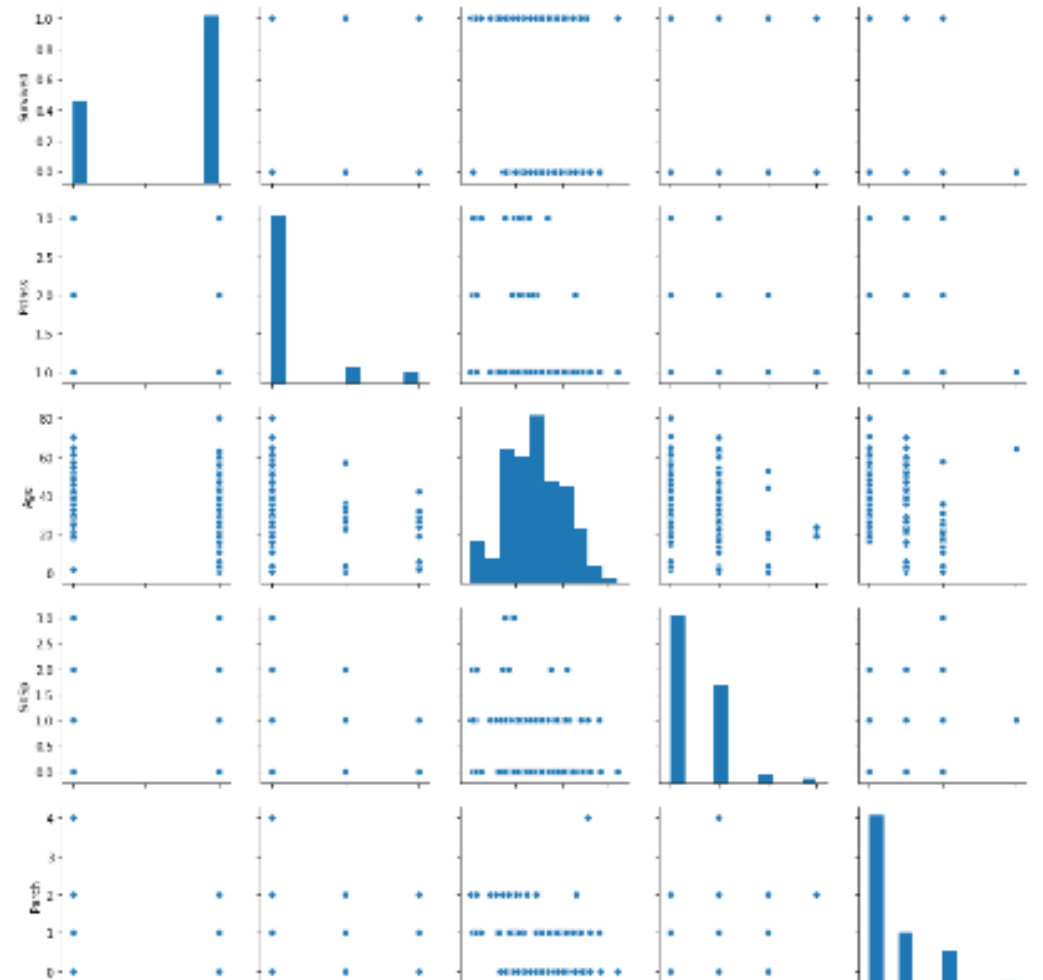
Зависимая
переменная



Потенциальные
выбросы

Независимая переменная

Матрица диаграмм рассеяния



Коэффициенты корреляции

1. Коэффициент корреляции Пирсона,
2. Коэффициент корреляции Спирмена,
3. Коэффициент корреляции Кендалла.

Коэффициент корреляции Пирсона

Измеряет силу линейной взаимосвязи между двумя переменными.

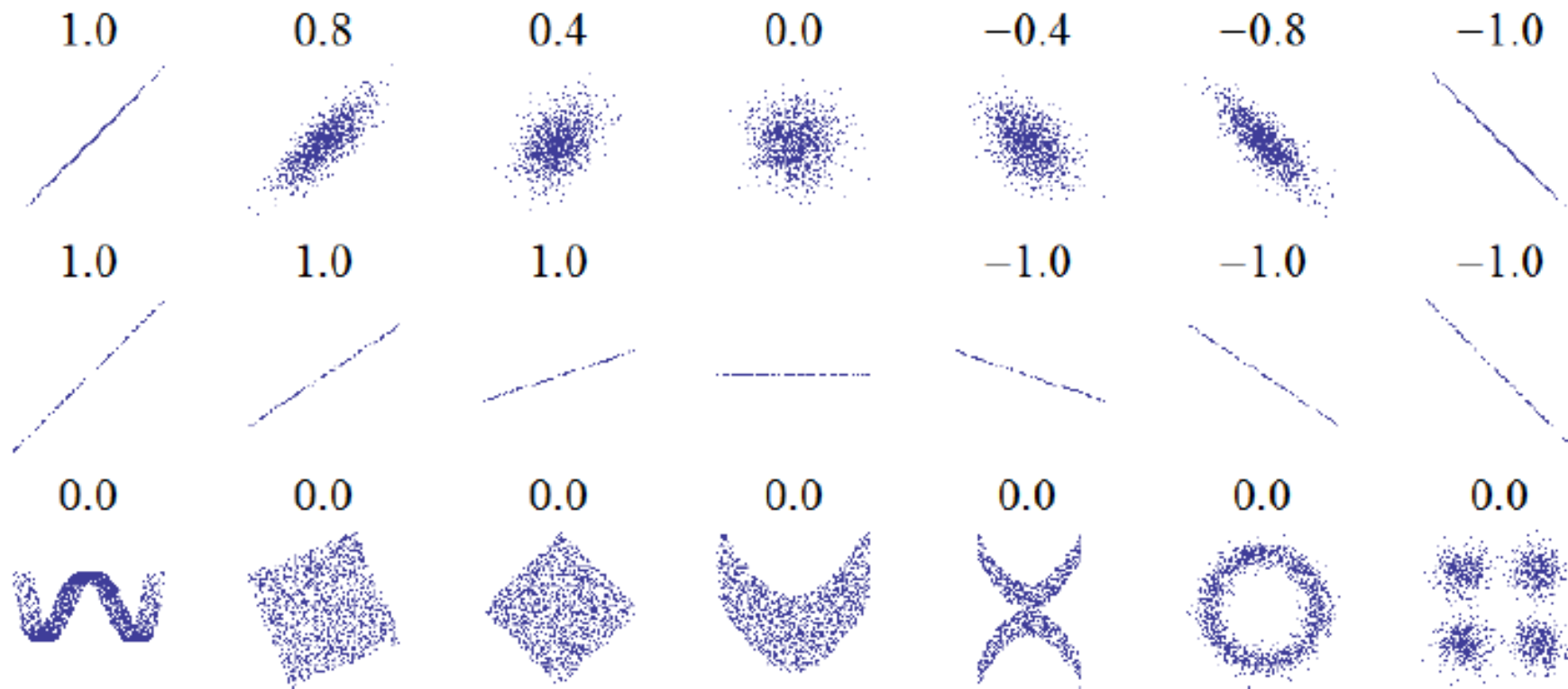
$$r_{xy} = \frac{\sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{(n-1) \cdot \sigma_x \cdot \sigma_y}$$

- Принимает значения от -1 до +1.
- Может быть подсчитан для метрических переменных, значения которых подчиняются закону нормального распределения.
- Измеряет направление (знак) и силу (величина) связи.

Интерпретация значений коэффициента корреляции

Значение коэффициента корреляции	Интерпретация
$0 < r \leq 0,2$	Очень слабая корреляция
$0,2 < r \leq 0,5$	Слабая корреляция
$0,5 < r \leq 0,7$	Средняя корреляция
$0,7 < r \leq 0,9$	Сильная корреляция
$0,9 < r \leq 1$	Очень сильная корреляция

Вид диаграммы рассеяния для разных значений r



Проверка значимости коэффициента корреляции Пирсона

1. $H_0: r = 0$, $H_1: r \neq 0$
2. Подсчитывается t-критерий

$$t = r \sqrt{\frac{n - 2}{1 - r^2}}$$

3. По таблице распределения Стьюдента, определяется критическое значение t при заданном уровне значимости и степенях свободы.
4. Если фактическое значение превышает критическое, то гипотеза H_0 отклоняется и принимается H_1 .
5. В Python вычисляется вероятность верности H_0 , которую необходимо сравнить с заданным пороговым уровнем значимости (0,05 или 0,01).

Ранговые коэффициенты корреляции

- Применяются для измерения связи между порядковыми и метрическими переменными, распределения которых отличаются от нормального.
- При подсчёте вместо исходных значений используются ранги.
- Наиболее популярные ранговые коэффициенты: Спирмена и Кендалла.

К-т Кендалла подсчитывается при большом числе связанных рангов, при небольшом числе связанных рангов подсчитывается к-т Спирмена.

Коэффициент корреляции Спирмена

Можно использовать для расчёта корреляции между интервальными переменными, распределение которых отличается от нормального и порядковыми переменными. Сперва исходные значения переменных ранжируются. Затем расчёты проводятся с рангами. Подходит когда при ранжировании не появляется много связанных рангов (т.е. почти все значения переменных уникальны).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Коэффициент корреляции Спирмена

Sales (x_i)	Advertisement (y_i)	Rank for Sales (x_i)	Rank for Advertisement (y_i)	$d_i = x_i - y_i$	d_i^2
90	7	2	2	0	0
85	6	3	3	0	0
68	2	8	7	1	1
75	3	6	6	0	0
82	4	4	5	-1	1
80	5	5	4	1	1
95	8	1	1	0	0
70	1	7	8	-1	1
Total	—	—	—	0	4

Связанные ранги

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})
15	2	40	6
20	3.5	30	4
28	5	50	7
12	1	30	4
40	6	20	2
60	7	10	1
20	3.5	30	4
80	8	60	8

Коэффициент корреляции Кендалла

Рекомендуется рассчитывать для небольших выборок с большим числом связанных рангов при ранжировании значений переменных.

$$Kendall's \tau = \frac{C - D}{C + D}$$

C – число совпадений

D – число инверсий

Коэффициент корреляции Кендалла

Var 1	Var 2	C	D
1	2	10	1
2	1	10	0
3	4	8	1
4	3	8	0
5	6	6	1
6	5	6	0
7	8	4	1
8	7	4	0
9	10	2	1
10	9	2	0
11	12	0	1
12	11		

$$Kendall's \tau = \frac{60 - 6}{60 + 6}$$

$$Kendall's \tau = \frac{54}{66}$$

$$Kendall's \tau = .818$$

Коэффициент корреляции Кендалла

Var 1	Var 2	C	D
1	12	0	11
2	2	9	1
3	3	8	1
4	4	7	1
5	5	6	1
6	6	5	1
7	7	4	1
8	8	3	1
9	9	2	1
10	10	1	1
11	11	0	1
12	1		

$$Kendall's \cdot tau = \frac{45 - 21}{45 + 21}$$

$$Kendall's \cdot tau = \frac{24}{66}$$

$$Kendall's \cdot tau = .364$$

Сравнение коэффициентов корреляции

Коэффициент корреляции	Переменные
Пирсона	Обе переменные метрические, распределение значений которых не отличается от нормального
Спирмена	Переменные метрические не нормально распределённые или порядковые, большинство их значений уникальны
Кендалла	Переменные метрические не нормально распределённые или порядковые, их значения не обязательно уникальны

Коэффициент частной корреляции

При изучении корреляции двух переменных необходимо учитывать возможное воздействие на них со стороны других переменных. Частная корреляция позволяет измерить связь между парой переменных после удаления линейных воздействий третьей переменной.

Частный коэффициент корреляции (Partial correlation coefficient) – это мера зависимости между двумя переменными при фиксации (исключении, корректировке) эффектов одной или нескольких переменных.

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2} \sqrt{1 - r_{yz}^2}}$$

Пример частной корреляции 1

В 1873 году Поль Брока обнаружил сильную взаимосвязь между полом и размером мозга: мозг женщин в целом меньше, чем у мужчин. В то время это использовалось как доказательство того, что женщины интеллектуально уступают мужчинам. Очевидная проблема заключалась в том, что эта взаимосвязь не учитывала размеры тела: у людей с более крупным телосложением мозг больше, независимо от интеллектуальных способностей.

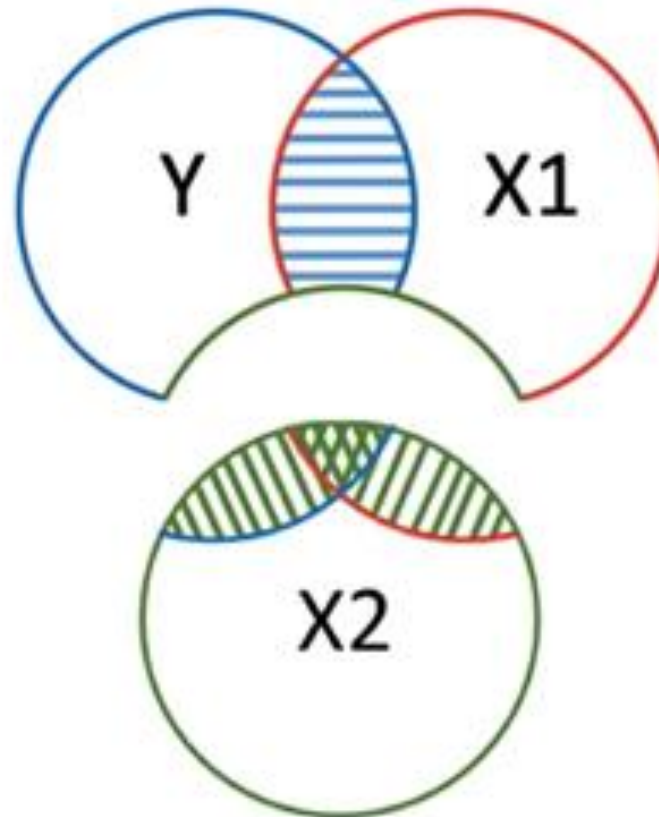
В 1981 году Стивен Джей Гулд заново проанализировал данные Брока и показал, что сильная зависимость между полом и размером мозга исчезает, если учитывать размер тела.

Пример частной корреляции 2

Наблюдается корреляция между размером обуви и спортивными успехами. Но на самом деле скрытая переменная – рост: высокие люди обычно имеют и больший размер обуви, и лучшие спортивные показатели в определенных видах спорта. Если учесть рост, связь между размером обуви и успехами в спорте почти исчезает.

Частная корреляция

Partial Correlation between Y and X1
controlling for X2



Корреляции с бинарными переменными

Коэффициент корреляции	Переменная 1	Переменная 2
Точечно-бисериальная корреляция	Бинарная	Метрическая
Фи (φ) коэффициент корреляции	Бинарная	Бинарная

Точечно-бисериальная корреляция

Частный случай коэффициента корреляции Пирсона, позволяющий оценить связь между дихотомической и метрической переменными.



$$r = \frac{Mean_{group\ 1} - Mean_{group\ 2}}{SD_{sample}} \sqrt{\frac{n_{group\ 1} n_{group\ 2}}{n^2}}$$









Mean – среднее значение метрической переменной в группе

SD – стандартное отклонение

n – число наблюдений

Точечно-бисериальная корреляция

	2	failed	0
	3	passed	1
	16	failed	0
	17	passed	1
	5	passed	1
	6	passed	1
	14	failed	0
	7	passed	1

Mean value of the persons who failed

Mean value of the persons who passed

$$r_{pb} = \frac{\bar{x}_2 - \bar{x}_1}{s_x} \cdot \sqrt{\frac{n_1 \cdot n_2}{n^2}}$$

Number of people who have passed

Number of those who have failed

Total number

$$r = \frac{10.6 - 7.6}{5.6} \sqrt{\frac{5 * 3}{8^2}} = 0.25$$

Фи коэффициент корреляции

Используют для измерения тесноты связи переменных в таблицах 2x2.

	Y = 0	Y = 1	Total
X = 0	n_{00}	n_{01}	n_{0*}
X = 1	n_{10}	n_{11}	n_{1*}
Total	n_{*0}	n_{*1}	n

$$\varphi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1*}n_{*1}n_{0*}n_{*0}}}$$



Название теста	Цель	Тип данных	Сравниваемые группы	Ключевые допущения	Когда использовать
Хи-квадрат (Chi-Square Test)	Исследовать взаимосвязь между категориальными переменными	Категориальные (номинальные/порядковые)	2 категориальные переменные	Ожидаемая частота ≥ 5 в каждой ячейке	Сравнение наблюдаемых и ожидаемых частот (например, пол и предпочтение в голосовании)
Корреляция Пирсона (Pearson Correlation)	Измерение линейной связи между переменными	Непрерывные	2 переменные	Нормальность, линейность	Анализ корреляции между ростом и весом
Корреляция Спирмена (Spearman Correlation)	Измерение монотонной зависимости	Порядковые / Непрерывные	2 переменные	Множество совпадающих рангов может снизить точность коэффициента корреляции	Непараметрическая альтернатива корреляции Пирсона
Корреляция Кендалла (Kendall Correlation)	Измерение порядковой ассоциации	Порядковые	2 переменные	Большое количество связанных рангов снижает мощность и точность расчёта. τ -b Кендалла больше подходит для порядковых переменных с малым числом значений, чем τ -a	Небольшие выборки или наличие множества совпадающих рангов
Точечно-бисериальная корреляция (Point-Biserial Correlation)	Измерение связи между непрерывной и бинарной переменной	Непрерывная + Бинарная	1 непрерывная, 1 бинарная	Бинарная переменная должна быть строго дихотомической	Корреляция между результатами теста и исходом «сдал/не сдал»
Коэффициент Фи (Phi Correlation Coefficient)	Измерение силы связи между двумя бинарными переменными	Бинарные (дихотомические)	2 бинарные переменные	Переменные должны быть строго бинарными	Корреляция между двумя «да/нет» переменными



Факультет компьютерных наук

НИС Python

Москва 2025

Спасибо за внимание!