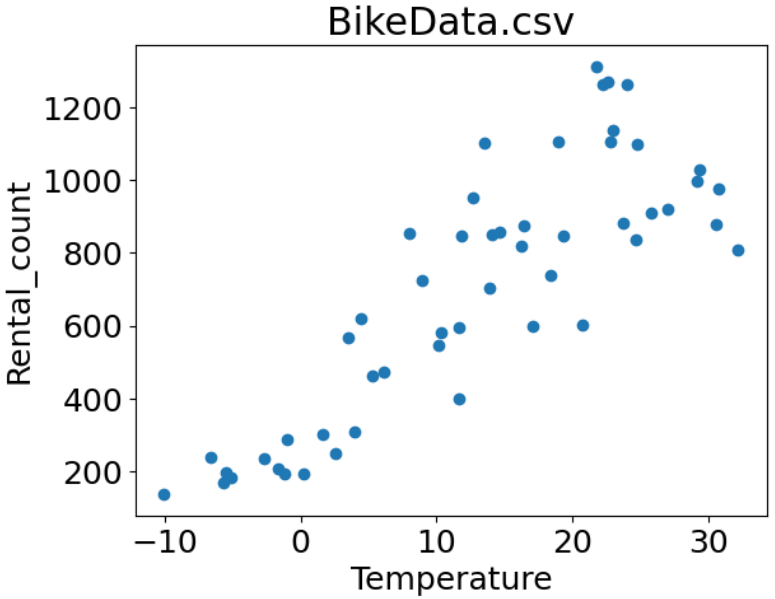
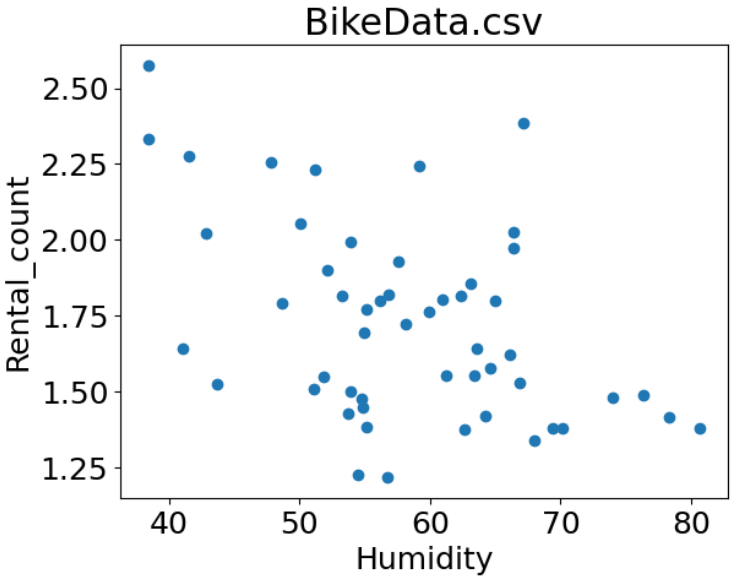
Корреляционная зависимость

Корреляционная зависимость (связь) – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

Например: Чем выше температура на улице, тем больше велосипедов сдается в аренду.
С увеличением влажности количество арендованных велосипедов уменьшается.





Наглядное представление о связи двух переменных дает **график рассеяния** (scatter plot), на котором каждый объект представляет собой точку, координаты которой заданы значениями двух переменных. Таким образом, множество объектов представляет собой на графике множество точек. По конфигурации этого множества точек можно судить о характере связи между двумя переменными.

Корреляционный анализ

Корреляционный анализ – статистический метод, используемый для измерения корреляционной связи. Цель -- отвечает на два вопроса:

1) Как сильна корреляционная связь?

2) Какое направление имеет корреляционная связь?

! Нет цели установить функциональный вид связи.
! Нет цели установить причинно следственную связь (что на что влияет).

Для ответа на вопросы 1 и 2 использую **коэффициент корреляции**, который является статистической мерой линейной связи двух переменных.

Этапы корреляционного анализа:

1. **Классификация типа данных,** т.к. для различных типов данных используются различные коэффициенты корреляции.

2. **Оценка силы и направления связи** с помощью соответствующего эмпирического значения коэффициента корреляции.

3. **Проверка гипотезы об отсутствии связи признаков**. Поскольку эмпирический коэффициента корреляции это случайная величина (определяется по выборке), то ненулевое значение коэффициента корреляции может быть вызвано случайными факторами. Поэтому проверять нулевую гипотезу, что коэффициент корреляции =0.

Отступление. Шкалы измерения

Для дальнейшего изложения материала нам понадобится различать исследуемые переменные по их типу:

Номинальные (категориальные) — для классификации измеряемых признаков. Могут характеризоваться текстовым значением, но всегда можно сопоставить с числом (числовые метки).
Примеры: пол, группа крови, образование, любые признаки типа «да - нет».
Доступные операции: равны или нет.

Порядковые (ординальные) — для сравнения признаков. Доступны операции сравнения.
Примеры: данные типа «лучше - хуже», оценки, разряды, уровень образования и т. д.

Количественные — все остальные. Они могут отличаться доступными арифметическими операциями, имеющими осмысленное значение. Например, дата (операции сравнения и вычитания), размер (все операции).

Коэффициент корреляции

Коэффициент корреляции *r* . -- статистическая мера линейной связи двух переменных.

Коэффициент корреляции *r*

Изменяется от -1 до 1

Значение |r|	Сила связи
0.0 < 0.2	Очень слабая корреляция
0.2 < 0.5	Слабая корреляция
0.5 < 0.7	Средняя корреляция
0.7< 0.9	Высокая корреляция
0.9 < 1	Очень высока корреляция

Если коэффициент корреляции равен 0, обе переменные линейно независимы друг от друга.

Направление связи
Прямая или *положительная корреляция* ($r>0$) более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого.
Обратная или *отрицательная корреляция* ($r<0$) более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – высокие значения другого.

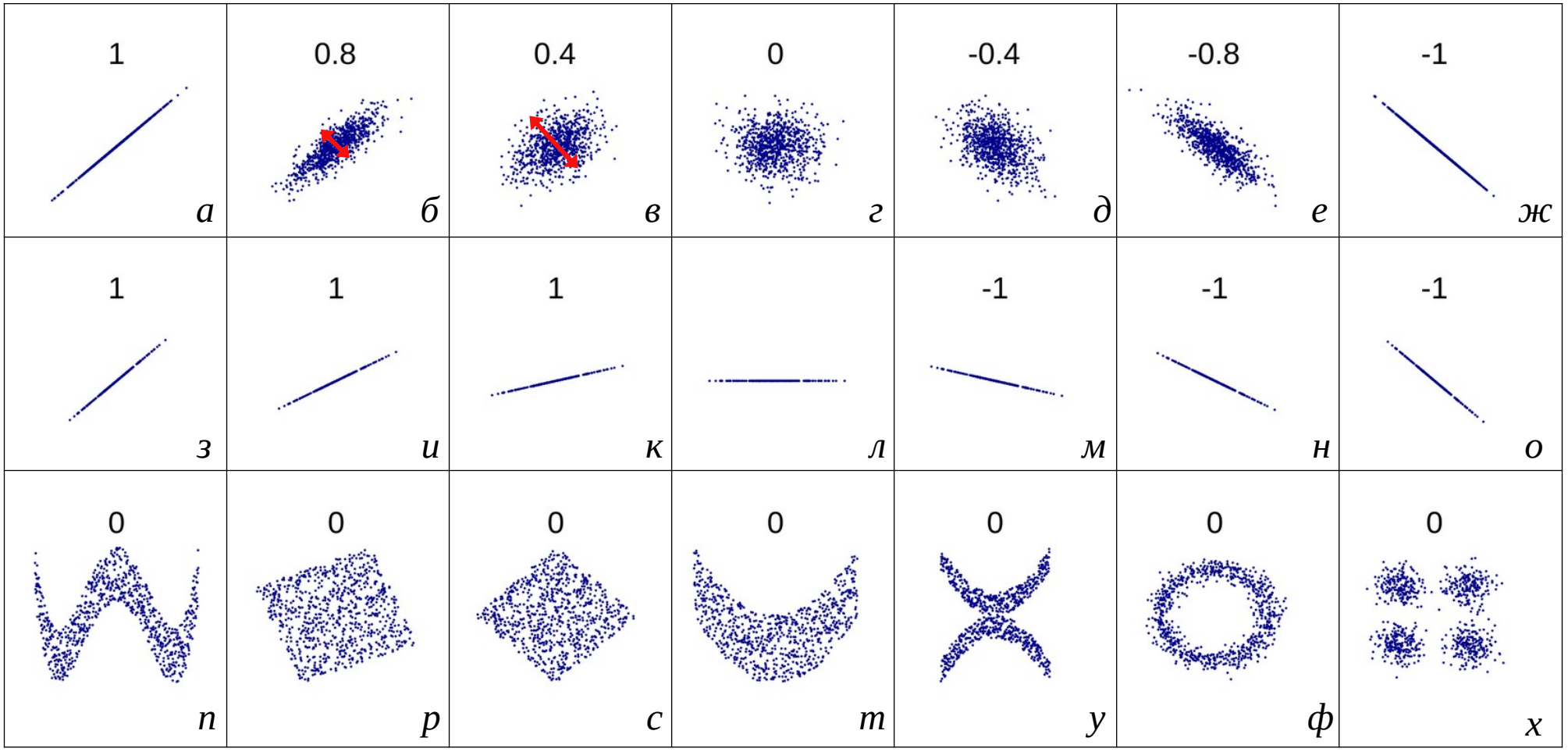
В зависимости от типа данных используют различные коэффициент корреляции:

- Для признаков, измеренных в номинальной шкале, применяются ***таблицы сопряженности, статистика Фишера-Пирсона*** X^2, различные *меры связи* (коэффициенты Юла, Крамера);
- Для признаков, измеренных в порядковой шкале, применяются *ранжирование* и ***коэффициенты корреляции Спирмана и Кендэлла***;
- Для данных, измеренных в количественных шкалах, применяют ***коэффициент корреляции Пирсона***.

Более точную таблицу с видами коэффициентов корреляций можно найти здесь Шпаргалка

Примеры корреляционной связи

Чем сильнее связь, тем меньше точки отклоняются от прямой (см. рис. ⟷)



а, *з*-*к*) строгая положительная корреляция;

б) высокая положительная корреляция;

в) слабая положительная корреляция;

г, *л*) нулевая корреляция;

д) слабая отрицательная корреляция;

е) высокая отрицательная корреляция;

ж, *м*-*о*) строгая отрицательная корреляция;

п-*х*) нелинейная корреляция

Коэффициент корреляции Пирсона

Коэффициент парной корреляции Пирсона -- статистическая мера линейной связи между двумя переменными.

$$r_{xy} = \frac{M[(X - M[X])(Y - M[Y])]}{\sigma[X]\sigma[Y]}$$

Числитель -- сумма площадей для всех *i*.
Если точек в I и III квадрантах больше, то сумма площадей будет положительной (положительная корреляция). В противном случае — сумма будет отрицательной (отрицательная корреляция).

Знаменатель нормирует значение суммарной площади.

Свойства:

1) $r_{xy} \in [-1, 1]$;

2) $r_{xy} = r_{yx}$;

3) *Если X,Y независимые, то* $r_{xy} = 0$, *обратное неверно*!!!

4) $r_{xy} = \pm 1$ *функциональная линейная связь*

5) *Определяет факт наличия и тесноту (силу) только линейной связи*

В I и III квадрантах площади S_i будут положительными (+*+ и -*-), в II и IV — отрицательные (+*- и -*+).

$$S_i = (x_i - M[X])(y_i - M[Y])$$

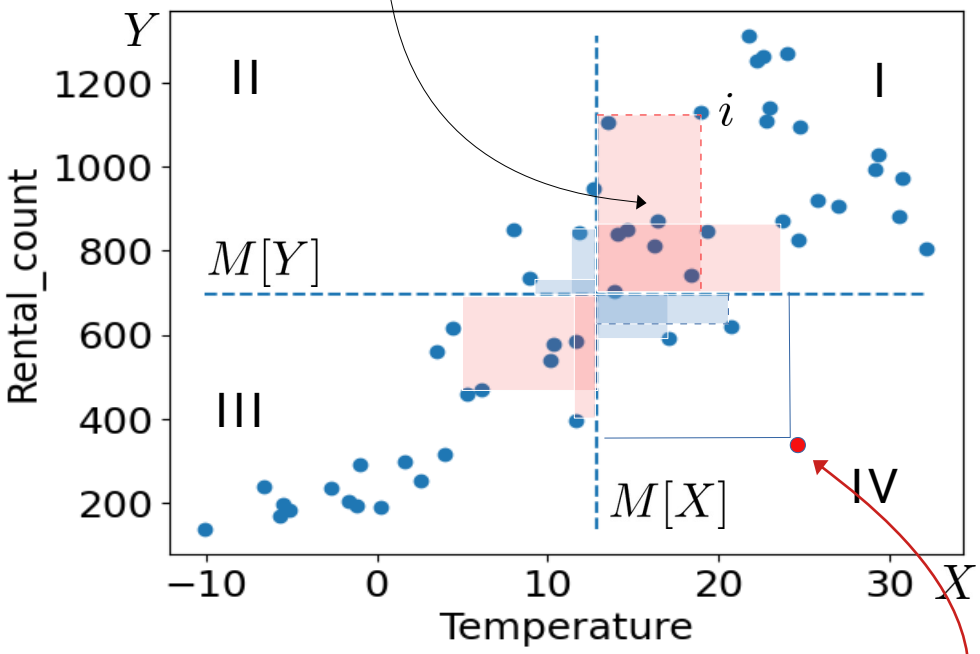


Рис. 1 Геометрическая интерпретация коэффициента корреляции Пирсона

Очевидно, что если существует выбросы (сильно отличающиеся значения), то значение коэффициента корреляции Пирсона будет сильно искажаться.

Анализ корреляций количественных признаков

Несмещенная оценка для коэффициента парной корреляции
(***эмпирический коэффициент корреляции***)

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Если ***n*<15**

$$\hat{r}^*_{xy} = \hat{r}_{xy}\left[1 + \frac{1 - \hat{r}^2_{xy}}{2(n-3)}\right]$$

Проверяют *значимость* эмпирического коэффициента корреляции, т. е. проверяют гипотезу:

$H_0 : r_{xy} = 0, \quad \longrightarrow$ r_{xy} **не значимо отличается от нуля**, поэтому линейной связи нет

$H_1 : r_{xy} \neq 0 \quad \longrightarrow$ r_{xy} **значимо отличается от нуля**, поэтому линейная связь существует

Если выборки взяты из нормально распределенных генеральных совокупностей, то в качестве критерия используют:

$$Z = \frac{\hat{r}_{xy}\sqrt{n-2}}{\sqrt{1 - \hat{r}^2_{xy}}} \in F_T(n-2)$$

Функция распределения Стьюдента с (n-2) степенями свободы

Если p-value для вычисленного критерия меньше уровня значимости, то нулевая гипотеза отвергается в пользу альтернативной, т. е. коэффициент корреляции значимо отличается от нуля и существует линейная связь.
В противном случае, линейно связи нет.

Важно запомнить

- Коэффициент корреляции Пирсона (r) вычисляется для **количественных** переменных.
- Коэффициент корреляции Пирсона (r) оценивает только **линейную связь** переменных. Нелинейную связь данный коэффициент выявить не может.
- Коэффициент корреляции Пирсона очень чувствителен к выбросам.
- Значение корреляции Пирсона можно вычислить для произвольных выборок. **Для проверки значимости коэффициента корреляции** требуется, чтобы выборки были взяты из **нормально распределенных генеральных совокупностей**.
- Корреляция не подразумевает наличия причинно-следственной связи между переменными.
- Нельзя путать коэффициент корреляции Пирсона с критерием Пирсона ХИ-квадрат.

Коэффициенты ранговой корреляции

Коэффициенты ранговой корреляции — непараметрический (*не требующая нормальность выборок*) копия коэффициента корреляции Пирсона. Наиболее часто используются **коэффициенты ранговой корреляции Спирмана и Кенделла.**

При вычислении коэффициентов ранговой корреляции не использует исходные данные, а берет только их ранги.

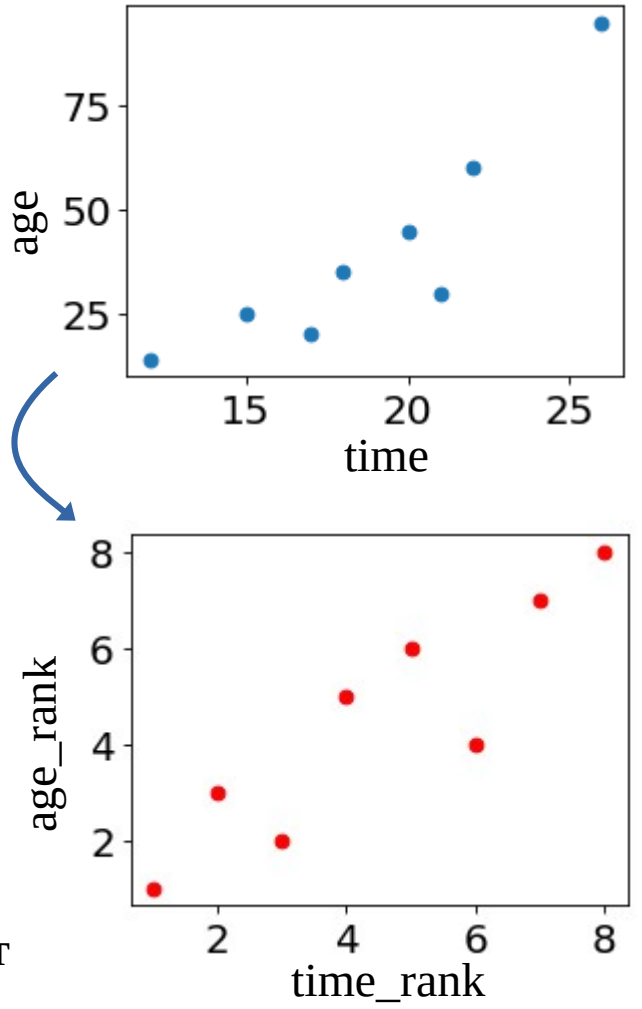
Например, имеются модельные данные о времени реакции людей различного возраста:

time	age
12	14
15	25
17	20
18	35
20	45
21	30
22	60
26	95

Трансформация значений в их ранги

time	t-rank	age	age-rank
12	1	14	1
15	2	25	3
17	3	20	2
18	4	35	5
20	5	45	6
21	6	30	4
22	7	60	7
26	8	95	8

Рис. Изменение графиков рассеяния после трансформации данных



Если переменные независимы,то и их порядок (ранги) будут случайными. Все возможные сочетания равновероятны.

Коэффициент ранговой корреляции Спирмана

Коэффициент корреляции Спирмана использует разницу между рангами двух переменных для установления зависимости между признаками (рангами). Если признаки зависимы, то разница между рангами будет минимальной. Если переменные независимы,то и их порядок (ранги) будут случайным. Все возможные сочетания рангов равновероятны.

Как вычисляется? 1) по той же формуле, что и коэффициент Пирсона, только вместо исходных значений берутся их ранги.

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$x_i = \text{t-rank}_i$ $\quad \bar{x} = \overline{\text{t-rank}} = 4.5$

$y_i = \text{age-rank}_i$ $\quad \bar{y} = \overline{\text{age-rank}} = 4.5$

$$\hat{r}_{xy} = 0.9$$

time	t_rank	age	age_rank	*d*	d^2
12	1	14	1	0	0
15	2	25	3	-1	1
17	3	20	2	1	1
18	4	35	5	-1	1
20	5	45	6	-1	1
21	6	30	4	2	4
22	7	60	7	0	0
26	8	95	8	0	0

$$\sum d_i^2 = 8$$

2) Существует **альтернативная формула** для определения коэффициента корреляции Спирмана:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

n — число точек

d_i — разность между рангами двух переменных

Для полностью совпадающих рангов сумма d_i будет = 0

$$r_s = 1 - \frac{6 \cdot 8}{8(64-1)} = 0.9$$

Коэффициент ранговой корреляции Спирмана

Коэффициент корреляции Спирмана использует разницу между рангами двух переменных для установления зависимости между признаками (рангами). Если признаки зависимы, то разница между рангами будет минимальной. Если переменные независимы,то и их порядок (ранги) будут случайным. Все возможные сочетания рангов равновероятны.

Как вычисляется? 1) по той же формуле, что и коэффициент Пирсона, только вместо исходных значений берутся их ранги.

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$x_i = \text{t-rank}_i$ $\quad \bar{x} = \overline{\text{t-rank}} = 4.5$

$y_i = \text{age-rank}_i$ $\quad \bar{y} = \overline{\text{age-rank}} = 4.5$

$$\hat{r}_{xy} = 0.9$$

time	t_rank	age	age_rank	*d*	d^2
12	1	14	1	0	0
15	2	25	3	-1	1
17	3	20	2	1	1
18	4	35	5	-1	1
20	5	45	6	-1	1
21	6	30	4	2	4
22	7	60	7	0	0
26	8	95	8	0	0

$$\sum d_i^2 = 8$$

2) Существует **альтернативная формула** для определения коэффициента корреляции Спирмана:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

n — число точек

d_i — разность между рангами двух переменных

Для полностью совпадающих рангов сумма d_i будет = 0

$$r_s = 1 - \frac{6 \cdot 8}{8(64-1)} = 0.9$$

Коэффициент ранговой коорреляцийи Кеддэлла

Коэффициент (ранговой) корреляции Кендэлла — непараметрический копия коэффициента корреляции Пирсона, как и коэффициент корреляции Спирмана (они очень похожи).

Коэффициент корреляции Кендэлла лучше работает при **малом количестве элементов выборки с большим количеством рангов**.

В качестве **меры сходства рангов** используется не разность рангов, как у Спирмана, а **минимальное число перестановок**, которые надо сделать, чтобы одно упорядочение объектов превратить в другое.

Как вычисляется?

$$\tau = \frac{C - D}{C + D}$$

Например, два доктора распределили пациентов по состоянию здоровья:

	Ранг состояния здоровья	
Пациент	Врач №1	Врач №1
1	1	**3**
2	2	**1**
3	3	**4**
4	4	**2**
5	5	**6**
6	6	**5**

	Больше или меньше **3**	Больше или меньше **1**	Больше или меньше **4**	Больше или меньше **2**	Больше или меньше **6**
2	-				
3	+	+			
4	-	+	-		
5	+	+	+	+	
6	+	+	+	+	-

C (concordant pairs) , т. е. «+»
D (discorconrdant pairs) , т. е. «-»

Альтернаитвная формула:

$$\tau = \frac{2(C - D)}{n(n-1)}$$

C (т.е. «+») = 11 и *D* (т. е. «-») = 4

$$\tau = \frac{11 - 4}{11 + 4} = 0.47$$

Проверка значимости коэффициентов ранговой корреляции

Проверяют *значимость* эмпирического коэффициента корреляции, т. е. проверяют гипотезу:

$H_0 : r_s = 0 \text{ or } \tau = 0, \quad \longrightarrow$ **r_s и τ не значимо отличается от нуля**, поэтому линейной связи нет

$H_1 : r_s \neq 0 \text{ or } \tau \neq 0 \quad \longrightarrow$ **r_s и τ значимо отличается от нуля**, поэтому линейная связь существует

1) Для проверки гипотезы используют тот факт, что **при верной H_0 и больших n** (>40) случайные величины:

$$Z = r_s\sqrt{n-1} \in N(0,1) \qquad \text{и} \qquad Z = \tau\sqrt{\frac{9n(n+1)}{2(2n+5)}} \in N(0,1)$$

По этим формулам вычисляется выборочное значение Z и соответствующее ему значение p-value, которое сравнивается с уровнем значимости α.
Если p-value < α, то нулевая гипотеза отвергается, т. е. коэффициенты корреляции значимо отличаются от нуля, значит корреляционная связь существует.

2) **Для малых n** хороших критериев не существует, поэтому пользуются негласным правилом: «Появление больших(по модулю) наблюдаемых значений ранговой корреляции свидетельствует против гипотезы независимости в пользу связи между признаками».

Важно запомнить

- Коэффициенты ранговой корреляции Спирмана и Кендэлла используются для измерения взаимозависимости **между признаками, значения которых могут быть упорядочены или проранжированы по степени убывания** (или возрастания) данного качества у исследуемых объектов.
- Для определения значимости коэффициентов ранговой корреляции **не требуется нормальность выборочных данных**.
- Коэффициенты ранговой корреляции Спирмена и Кендэлла позволяет определить тесноту (силу) и направление корреляционной связи между двумя признаками или двумя профилями (иерархиями) признаков.

$$r_s \in [-1, 1] \qquad \text{и} \qquad \tau \in [-1, 1]$$

Поэтому сила и направление связи определяются аналогично коэффициенту Пирсона.

Анализ корреляций категориальных признаков

Корреляция между категориальными переменными не может быть измерена с помощью коэффициентов Пирсона, Спирмена и Кендалла.

Используется **таблица сопряженности** (кросс-таблицы, contingency tables)

Gender	Metric Gender	Works	Metric Works
Male	0	Yes	1
Female	1	No	0
Female	1	Yes	1
Male	0	Yes	1
Female	1	Yes	1
Male	0	No	0
Male	0	Yes	1

		Metric Works		
		0	1	
Metric Gender	0	1 = A	3 = C	4
	1	1 = B	2 = D	3
		2	5	7

crosstab в модуле scipy.stats есть функция chi2_contingency

Для данной таблицы выдвигается гипотеза о независимости признаков. В этом случае, если гипотеза верна: $p(A_i B_j) = p(A_i) \cdot p(B_j)$

$$\frac{n_{ij}}{n} = \frac{n_i}{n}\frac{n_j}{n} \quad \Longrightarrow \quad n_{ij} = \frac{n_i n_j}{n}$$

Наблюдаемые частоты

Ожидаемые частоты

Т.е. задача сводится в сравнении частот некоторого закона распределения == проверке гипотезы о законе распределения по частотам, т. е. надо использовать критерий H^2-Пирсона

Анализ корреляций категориальных признаков

Наблюдаемые частоты

Ожидаемые частоты

		Metric Works		
		0	1	
Metric Gender	0	1/(8/7)	3/(20/7)	4
	1	1/(6/7)	2/(15/7)	3
		2	5	7

	A(0,0)	B(1,0)	C(0,1)	D(1,1)
наблюдаемые	1	1	3	2
теоретические	8/7	6/7	20/7	15/7

Из лекции по проверке непараметрических гипотез:

$$\chi^2 = \sum_{j=1}^{m} \frac{(n_j - np_j(\theta))^2}{np_j(\theta)} = \sum_{j=1}^{m} c_j \left(\frac{n_j}{n} - p_j(\theta)\right)^2, \qquad c_j = \frac{n}{p_j(\theta)}$$

$$\chi^2 = \sum \frac{(T-H)^2}{T} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\frac{n_i n_j}{n} - n_{ij}\right)^2}{\frac{n_i n_j}{n}} \in F_H^2((n_1-1)(n_2-1))$$

T — теоретические (ожидаемые) частоты
H — наблюдаемые частоты
n_1, n_2 — число возможных значений первой и второй категориальных переменных

Для вычисленного значения критерия вычисляется p-value, которое сравнивается с уровнем значимости.

Взаимосвязь категориального и числового признаков

В случае бинарного категориального признака (т. е. с двумя возможными исходами) и числового признака можно применить **бисериальный коэффициент корреляции (biserial correlation)**. Этот коэффициент корреляции является специфическим случаем коэффициента корреляции Пирсона.

Например, рассмотрим связь между временем подготовки (time) к тесту и результату (results) (сдал=Pass/провалил=Fail)

time	result	metric
2	Fail	0
3	Pass	1
16	Fail	0
17	Pass	1
5	Pass	1
6	Pass	1
14	Fail	0

$$r_{pb} \in [-1, 1]$$

Поэтому сила и напрявление связи определяются аналогично коэффициенту Пирсона.

Как вычисляется?

1) либо как коэффициент Пирсона для времени и меток теста:

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$x_i = \text{time}_i$$

$$y_i = \text{metric}_i$$

2) либо по соотношению:

$$r_{pb} = \frac{\bar{x}_2 - \bar{x}_1}{s_x}\sqrt{\frac{n_1 n_2}{(n_1 + n_2)^2}}$$

$\bar{x}_1, n_1$ -- среднее значение и количество тех, кто сдал

$\bar{x}_2, n_2$ -- среднее значение и количество тех, кто не сдал

s_x -- оценка для стандартного отклонения всех переменных

pointbiserialr из модуля scipy.stats

Проверка значимости бисериального коэффициентов корреляции

Проверяют *значимость* эмпирического коэффициента корреляции, т. е. проверяют гипотезу:

$H_0 : r_{pb} = 0 \longrightarrow$ r_{pb} **не значимо отличается от нуля**, поэтому линейной связи нет

$H_1 : r_{pb} \neq 0 \longrightarrow$ r_{pb} **значимо отличается от нуля**, поэтому линейная связь существует

Для проверки значимости бисериального коэффицента корреляции rpb требуется, чтобы метрическая переменная имела **нормальный закон распределения**. В это случае используют статистический критерий:

$$Z = r_{pd}\sqrt{\frac{n_1 + n_2 - 2}{1 - r_{pb}^2}} \in F_T(n_1 + n_2 - 2)$$

Распределение Стьюдента

По этим формулам вычисляется выборочное значение Z и соответствующее ему значение p-value, которое сравнивается с уровнем значимости α.

Если p-value < α, то нулевая гипотеза отвергается, т. е. коэффициенты корреляции значимо отличаются от нуля, значит корреляционная связь существует.

Источники информации, используемые в презентации:

1. https://www.kaggle.com/code/marfedorovna/lesson-2-correlation-analysis -- приложение к python.

2. https://www.youtube.com/watch?v=G5FkaxWBtkM -- отличное видео в целом о корреляционом анализе.

3. Еще раз ссылка на шпаргалку

4. На сайте https://studfile.net/preview/5855743/page:31/ приведена следующая таблица:

Типы сравниваемых шкал		Коэффициент корреляции
1	2	
Дихотомическая	Дихотомическая	Дихотомический (φ)
Дихотомическая	Порядковая (ранговая)	Рангово-бисериальный (r_{rb})
Дихотомическая	Интервальная	Точечно-бисериальный (r_{pb})
Ранговая	Ранговая	Коэффициент Пирсона (r_{xy}) Коэффициент Спирмена (r_s) Коэффициент Кендалла (τ)
Ранговая	Интервальная	Коэффициент Пирсона (r_{xy})
Интервальная	Интервальная	Коэффициент Пирсона (r_{xy})