

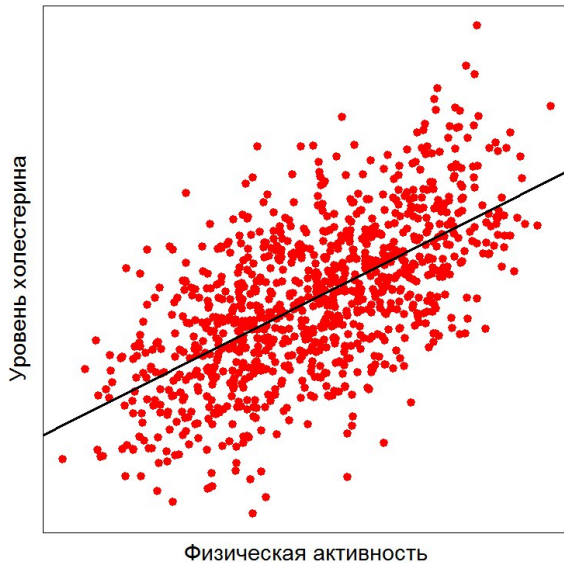
# Прикладной статистический анализ данных

## 11. Причинно-следственные связи

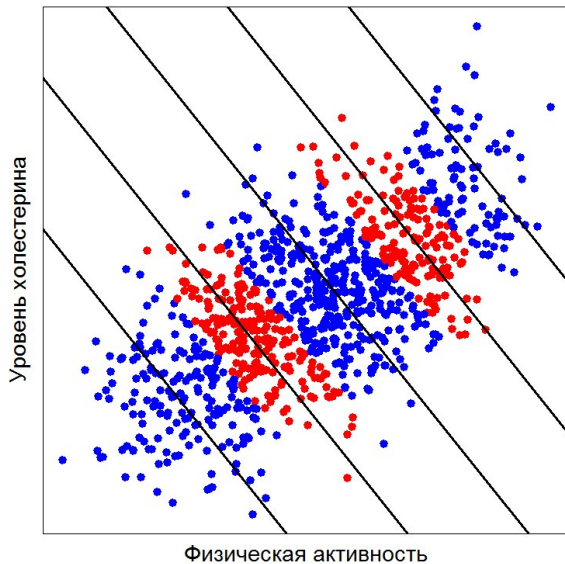
Бахтеев Олег  
psad@phystech.edu

2021

## Исследование уровня холестерина



## Исследование уровня холестерина



# Парадокс Симпсона

Пример 1:

$\Sigma$	лекарство	плацебо
выздоровели	273	289
не выздоровели	77	61
	78%	83%

плацебо на 5%  
эффективнее

мужчины	лекарство	плацебо
выздоровели	81	234
не выздоровели	6	36
	93%	87%

лекарство на 5%  
эффективнее

женщины	лекарство	плацебо
выздоровели	192	55
не выздоровели	71	25
	73%	69%

лекарство на 4%  
эффективнее

# Парадокс Симпсона

Какой из двух выводов верен?

Предположение: верны выводы по отдельным подгруппам, потому что они основаны на более детальной информации.

Это предположение неверно — всё зависит от того, как признак, по которому происходит разбиение на подгруппы, связан с остальными анализируемыми признаками.

# Парадокс Симпсона

Пример 2:

Лекарство снижает давление, но имеет множество побочных эффектов.

$\Sigma$	лекарство	плацебо
выздоровели	273	289
не выздоровели	77	61
	78%	83%

плацебо на 5%  
эффективнее

низкое давление в конце лечения	лекарство	плацебо
выздоровели	81	234
не выздоровели	6	36
	93%	87%

лекарство на 5%  
эффективнее

высокое давление в конце лечения	лекарство	плацебо
выздоровели	192	55
не выздоровели	71	25
	73%	69%

лекарство на 4%  
эффективнее

# Причинные графы

Отношения причинности могут быть представлены в виде направленного графа, вершины которого соответствуют признакам, а наличие пути говорит о существовании причинно-следственной связи.

**Путь** — последовательность вершин, где каждая вершина соединена со следующей ребром.

**Направленный путь** — путь, в котором все ребра имеют одинаковое направление.

# Элементы причинного графа

$X \rightarrow Y \rightarrow Z$  — цепочка

Пример:

- $X$  — бюджет школы
- $Y$  — средний балл учеников
- $Z$  — доля поступающих в ВУЗы

Свойства:

- 1  $X$  и  $Y$ ,  $Y$  и  $Z$  — зависимы:  
 $\exists x, y : \mathbf{P}(Y = y | X = x) \neq \mathbf{P}(Y = y)$   
 $\exists y, z : \mathbf{P}(Z = z | Y = y) \neq \mathbf{P}(Z = z)$
- 2  $Z$  и  $X$  — скорее всего, зависимы
- 3  $Z \perp X | Y$  — условно независимы:  $\forall x, y, z$

$$\mathbf{P}(Z = z | X = x, Y = y) = \mathbf{P}(Z = z | Y = y)$$

(если  $Y$  фиксировано, то  $X$  и  $Z$  независимы)



# Элементы причинного графа

$$X \leftarrow Y \rightarrow Z \text{ — вилка}$$

Пример:

- $X$  — продажи мороженого
- $Y$  — средняя дневная температура воздуха
- $Z$  — число преступлений

Свойства:

- ①  $X$  и  $Y$ ,  $Y$  и  $Z$  — зависимы
- ②  $X$  и  $Z$  — скорее всего, зависимы
- ③  $X \perp Z | Y$  — условно независимы

# Элементы причинного графа

$$Y \rightarrow X \leftarrow Z \text{ — коллайдер}$$

Пример (заболевание вирусом):

- $X$  — осложнения
- $Y$  — возраст
- $Z$  — хронические болезни

Свойства:

- 1  $Y$  и  $X$ ,  $Z$  и  $X$  — зависимы
- 2  $Y$  и  $Z$  — независимы
- 3  $Y \not\perp Z | X$  — условно зависимы

## d-разделимость

Путь  $P$  блокируется переменной  $Z$ , если:

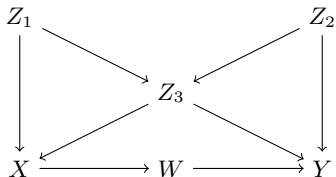
- ①  $P$  содержит  $A \rightarrow B \rightarrow C$ ,  $A \leftarrow B \rightarrow C$ ,  $B \in Z$
- ②  $P$  содержит  $A \rightarrow B \leftarrow C$ ,  $B \notin Z$  и все потомки  $B \notin Z$

Если  $Z$  блокирует все пути из  $X$  в  $Y$ , то  $X$  и  $Y$  **d-разделимы**:

$$X \perp Y | Z.$$

## d-разделимость

Пример:

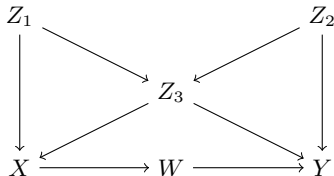


Упорядоченная пара вершин	d-разделяющее множество
$(Z_1, W)$	$X$

(условие 1: цепочка)

## d-разделимость

Пример:



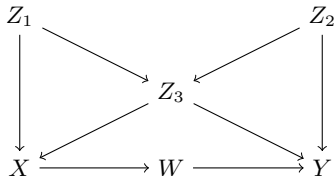
Упорядоченная пара вершин	d-разделяющее множество
$(Z_1, W)$	$X$
$(Z_1, Y)$	$\{Z_3, X, Z_2\}, \{Z_3, W, Z_2\}$

$(X, W, Z_3$ : цепочка)

$(Z_2$ : вилка)

## d-разделимость

Пример:



Упорядоченная пара вершин	d-разделяющее множество
$(Z_1, W)$	$X$
$(Z_1, Y)$	$\{Z_3, X, Z_2\}, \{Z_3, W, Z_2\}$
$(X, Y)$	$\{W, Z_3, Z_1\}$

( $W$ : цепочка)

( $Z_1, Z_3$ : вилка)

# Алгоритм индуктивной причинности

Вход: множество вершин  $V$

- ①  $\forall A, B \in V$  ищем множество  $S_{AB}: A \perp B | S_{AB}$ ,  $A, B \notin S_{AB}$ . Если такого  $S_{AB}$  не существует, соединяем  $A$  и  $B$  ребром.
- ②  $\forall A, B$ , не связанных ребром и имеющих общего соседа  $C$ , проверяем:  $C \in S_{AB}$ ? Если нет, то заменяем пару рёбер  $A - C, C - B$  на пару ориентированных рёбер  $A \rightarrow C, C \leftarrow B$
- ③ Рекурсивно применяем следующие два правила:
  - ▶ если из  $A$  в  $B$  есть ориентированный путь  $A \rightarrow \dots \rightarrow B$ , то  $A - B$  заменяем на  $A \rightarrow B$ ;
  - ▶ если  $A$  и  $B$  не соединены,  $A \rightarrow C, C - B$ , то  $C - B$  заменяем на  $C \rightarrow B$ .

Выход: ориентированный (возможно, частично) граф  $G$ .

## Алгоритм индуктивной причинности

Правила (1) и (2) применять в чистом виде невозможно — число перебираемых множеств экспоненциально растёт с числом вершин графа. Поэтому используются сокращающие перебор эвристики.

Признаки	дискретные	непрерывные
Распределение	мультиномиальное	нормальное
Критерий условной независимости	хи-квадрат для трёхмерных таблиц сопряжённости	Стьюдента для частной корреляции
Критерий качества графа	$BIC$	



## Причинность по Грейнджеру

Между рядами  $x_1, \dots, x_T$  и  $y_1, \dots, y_T$  существует **причинная связь Грейнджера**  $x_t \rightarrow y_t$ , если дисперсия ошибки оптимального прогноза  $\hat{y}_{t+1}$  по  $y_1, \dots, y_t, x_1, \dots, x_t$  меньше, чем только по  $y_1, \dots, y_t$ .

### Причинность по Грейнджеру

- может следовать из причинно-следственной связи;
- не является достаточным условием причинно-следственной связи.

$x_1, \dots, x_T$  и  $y_1, \dots, y_T$  **взаимосвязаны**, если  $x_t \rightarrow y_t$  и  $y_t \rightarrow x_t$ .

## Критерий Грейнджера

$$y_t = \alpha + \sum_{i=1}^{k_1} \phi_{1i} y_{t-i} + \sum_{i=1}^{k_2} \phi_{2i} x_{t-i} + \varepsilon_t.$$

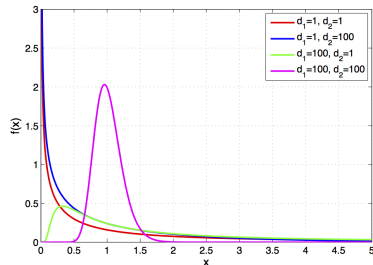
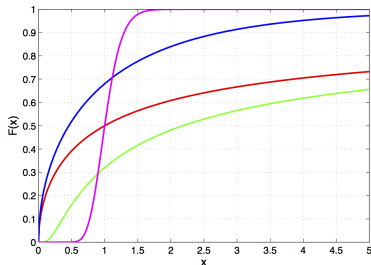
$k_1$  и  $k_2$  выбирается по информационному критерию.

$$x_t \rightarrow y_t \Rightarrow \exists \phi_{2i} \neq 0.$$

нулевая гипотеза:  $H_0: \phi_{21} = \dots = \phi_{2k_2} = 0;$

альтернатива:  $H_1: H_0$  неверна;

статистика:  $F = \frac{(RSS_r - RSS_{ur})/k_2}{RSS_{ur}/(T - k_1 - k_2 - 1)};$   
 $F \sim F(k_2, T - k_1 - k_2 - 1)$  при  $H_0$ .



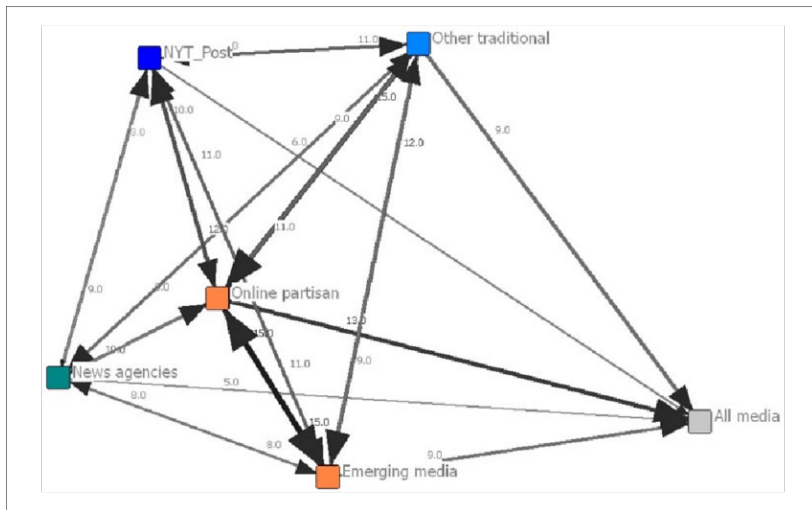
# Многомерный критерий Грейнджера

Зависимость между признаками  $x$  и  $y$  может оцениваться с учётом возможной зависимости от всех остальных признаков:

$$y_t = \alpha + \sum_{i=1}^{k_1} \phi_{1i} y_{t-i} + \sum_{i=1}^{k_2} \phi_{2i} x_{t-i} + \sum_{j=1}^m \sum_{i=1}^{k_{j+2}} \phi_{(j+2)i} z_{t-i}^j + \varepsilon_t.$$

Для задач с большим количеством признаков могут использоваться регуляризаторы (лассо, ридж).

## Граф причинности по Грейнджеру



К критерию Грейнджера применима поправка на множественную проверку гипотез

## Причинно-следственная связь и обусловленность

$$X \leftarrow Y \rightarrow Z.$$

- $X$  — продажи мороженого
- $Y$  — средняя дневная температура воздуха
- $Z$  — число преступлений

$X$  и  $Z$  коррелируют. Как понять, зависит ли число преступлений от продажи мороженого?

# Интервенция

$X$  коррелировано с  $Y \nRightarrow X$  влияет на  $Y$ .

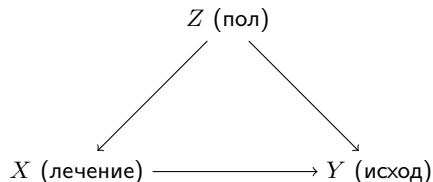
Влияние обычно оценивают в эксперименте, когда объектам искусственно назначают разные уровни  $X$ , но эксперимент можно провести не всегда:

- погода  $\rightarrow$  лесные пожары — не можем управлять  $X$
- теленасилие  $\rightarrow$  жестокость — тяжело фиксировать уровень  $X$  и создать условия для измерения  $Y$
- потребление алкоголя  $\rightarrow$  успеваемость школьников — неэтично

В таких случаях мы вынуждены использовать обсервационные данные, по которым мы хотим оценить эффект **интервенции**: что будет с  $Y$ , если мы установим значение  $X$  равным  $x$ ?

Обозначение:  $do(X = x)$ .

# Интервенция



Оценку эффективности лекарства можно сформулировать в терминах интервенций:

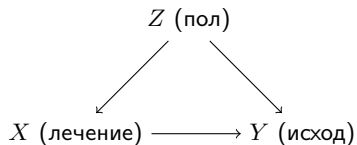
$$ACE = \mathbf{P}(Y = \text{выздоровление} | do(X = \text{лекарство})) - \\ - \mathbf{P}(Y = \text{выздоровление} | do(X = \text{плацебо})) .$$

(average conditional effect).

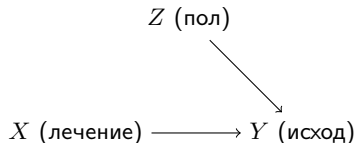
# Хирургия графа

**Хирургия графа** — удаление всех ребер, входящих в  $X$ .

Пример 1, исходный граф  $G$ :



Оперированный граф  $G_m$ :



$$\mathbf{P}(Y = y | do(X = x)) = \mathbf{P}_m(Y = y | X = x)$$



## Хирургия графа

В оперированном графе:

$$\mathbf{P}_m(Z = z) = \mathbf{P}(Z = z),$$

$$\mathbf{P}_m(Y = y | X = x, Z = z) = \mathbf{P}(Y = y | X = x, Z = z),$$

так как рёбра, входящие в  $Z$  и  $Y$ , не изменились  $\Rightarrow$

$$\begin{aligned}\mathbf{P}(Y = y | do(X = x)) &= \mathbf{P}_m(Y = y | X = x) = \\ &= \sum_z \mathbf{P}_m(Y = y | X = x, Z = z) \mathbf{P}_m(Z = z) = \\ &= \sum_z \mathbf{P}(Y = y | X = x, Z = z) \mathbf{P}(Z = z).\end{aligned}$$

## Хирургия графа

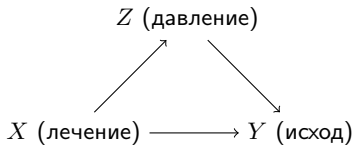
В примере 1 по полученной формуле:

$$\mathbf{P}(Y = \text{выздоровление} | do(X = \text{лекарство})) = 0.832,$$

$$\mathbf{P}(Y = \text{выздоровление} | do(X = \text{плацебо})) = 0.7818$$

$$\Rightarrow ACE = 0.05.$$

В примере 2  $G = G_m$ :



Значит,

$$\mathbf{P}(Y = y | do(X = x)) = \mathbf{P}_m(Y = y | X = x) = \mathbf{P}(Y = y | X = x)$$

$$\mathbf{P}(Y = \text{выздоровление} | do(X = \text{лекарство})) = 0.78,$$

$$\mathbf{P}(Y = \text{выздоровление} | do(X = \text{плацебо})) = 0.83$$

$$\Rightarrow ACE = -0.05.$$

## Поправочная формула

Поправочная формула позволяет вычислить эффект интервенции обуславливанием по вершинам  $Z$ :

$$\mathbf{P}(Y = y | do(X = x)) = \sum_z \mathbf{P}(Y = y | X = x, Z = z) \mathbf{P}(Z = z).$$

Что это за вершины?

**Формула причинного эффекта:**

$$\mathbf{P}(Y = y | do(X = x)) = \sum_z \mathbf{P}(Y = y | X = x, PA = z) \mathbf{P}(PA = z),$$

где  $PA$  — родители вершины  $X$ .

beginframed-разделимость Путь  $P$  блокируется переменной  $Z$ , если:

- ①  $P$  содержит  $A \rightarrow B \rightarrow C$ ,  $A \leftarrow B \rightarrow C$ ,  $B \in Z$
- ②  $P$  содержит  $A \rightarrow B \leftarrow C$ ,  $B \notin Z$  и все потомки  $B \notin Z$

Если  $Z$  блокирует все пути из  $X$  в  $Y$ , то  $X$  и  $Y$  **d-разделимы**:

$$X \perp Y | Z.$$

## d-разделимость

Путь  $P$  **блокируется** переменной  $Z$ , если:

- ①  $P$  содержит  $A \rightarrow B \rightarrow C$ ,  $A \leftarrow B \rightarrow C$ ,  $B \in Z$
- ②  $P$  содержит  $A \rightarrow B \leftarrow C$ ,  $B \notin Z$  и все потомки  $B \notin Z$

Если  $Z$  блокирует все пути из  $X$  в  $Y$ , то  $X$  и  $Y$  **d-разделимы**:

$$X \perp Y | Z.$$

## Неизвестные родители



Социоэкономический статус — ненаблюдаемая величина; как оценить эффект интервенции по  $X$ ?

## Критерий задней двери (КЗД)

Для упорядоченной пары вершин  $(X, Y)$  в ациклическом графе  $G$  множество вершин  $Z$  удовлетворяет **критерию задней двери**, если:

- $Z$  не содержит потомков  $X$
- $Z$  блокирует все пути между  $X$  и  $Y$ , содержащие  $X \leftarrow$ .

Если  $Z$  удовлетворяет КЗД для  $(X, Y)$ , то

$$\mathbf{P}(Y = y | do(X = x)) = \sum_z \mathbf{P}(Y = y | X = x, Z = z) \mathbf{P}(Z = z)$$

(формула задней двери).

## Критерий задней двери (КЗД)

Чтобы вычислять меньше условных вероятностей, ФЗД можно упростить:

$$\begin{aligned}\mathbf{P}(Y = y | do(X = x)) &= \sum_z \mathbf{P}(Y = y | X = x, Z = z) \mathbf{P}(Z = z) = \\ &= \sum_z \frac{\mathbf{P}(X = x, Y = y, Z = z)}{\mathbf{P}(X = x | Z = z)}\end{aligned}$$

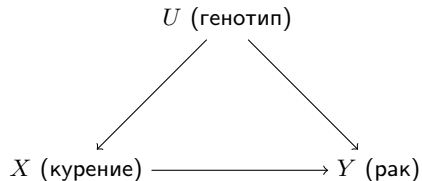
В таком виде

- метод называется **обратное вероятностное взвешивание**
- знаменатель  $\mathbf{P}(X = x | Z = z)$  — propensity score.



# Неизвестные родители

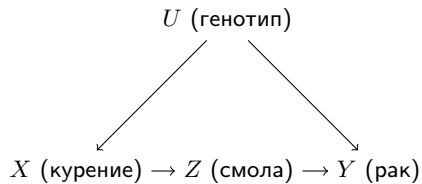
Вызывает ли курение рак?



$\Sigma$	курильщики	некурящие
нет рака	341	59
есть рак	39	361
	15%	90.25%

курильщики болеют  
на 75.25% реже

# Курение



смола	курильщики	некурящие
нет рака	323	1
есть рак	57	19
	15%	95%

курильщики болеют  
на 80% реже

нет смолы	курильщики	некурящие
нет рака	18	38
есть рак	2	342
	10%	90%

курильщики болеют  
на 80% реже

Курить полезно?

# Курение

У курильщиков смола в 95% случаев вместо 5%; у курильщиков смола увеличивает риск рака с 10% до 15%; у некурящих — с 90% до 95%.

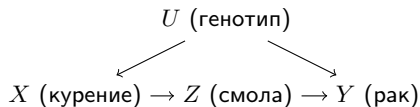
Курить вредно?

Поможет граф!

## Курение

Поправочная формула (КЗД для пустого множества и для  $X$ ):

$$\begin{aligned}\mathbf{P}(Z = z | do(X = x)) &= \mathbf{P}(Z = z | X = x), \\ \mathbf{P}(Y = y | do(Z = z)) &= \sum_{x'} \mathbf{P}(Y = y | Z = z, X = x') \mathbf{P}(X = x')\end{aligned}$$



$$\begin{aligned}\mathbf{P}(Y = y | do(X = x)) &= \\ &= \sum_z \mathbf{P}(Y = y | do(Z = z)) \mathbf{P}(Z = z | do(X = x)) = \\ &= \sum_z \sum_{x'} \mathbf{P}(Y = y | Z = z, X = x') \mathbf{P}(Z = z | X = x) \mathbf{P}(X = x').\end{aligned}$$

## Критерий передней двери (КПД)

Для упорядоченной пары вершин  $(X, Y)$  в ациклическом графе  $G$  множество вершин  $Z$  удовлетворяет **критерию передней двери**, если:

- $Z$  перекрывает все направленные пути из  $X$  в  $Y$
- нет незакрытых путей через заднюю дверь из  $X$  в  $Z$
- все пути через заднюю дверь из  $Z$  в  $Y$  блокируются  $X$

Если  $Z$  удовлетворяет КПД для  $(X, Y)$ , то

$$\begin{aligned} \mathbf{P}(Y = y | do(X = x)) &= \\ &= \sum_z \mathbf{P}(Z = z | X = x) \sum_{x'} \mathbf{P}(Y = y | X = x', Z = z) \mathbf{P}(X = x') \end{aligned}$$

(формула передней двери).

# Литература

- причинные графы и выводы по ним — Pearl
- восстановление графов по статическим данным — Nagarajan, глава 2
- причинность по Грейнджеру — Kirchgassner, глава 3

Kirchgassner G., Wolters J., Hassler U. *Introduction to modern time series analysis*, 2013.

Nagarajan R., Scutari M., Lebre S. *Bayesian Networks in R with Applications in Systems Biology*, 2013.

Pearl J., Glymour M., Jewell N.P. *Causal Inference in Statistics: A Primer*, 2016.

## Контрольная 2

- Время: после семинара (27 апреля)
- Темы: занятия 5-9