



# *Causality*

Выполняли: Бехруз Аъзам, Анастасия Безрукова, Анна Фролова

## Когда статистика - не панацея?

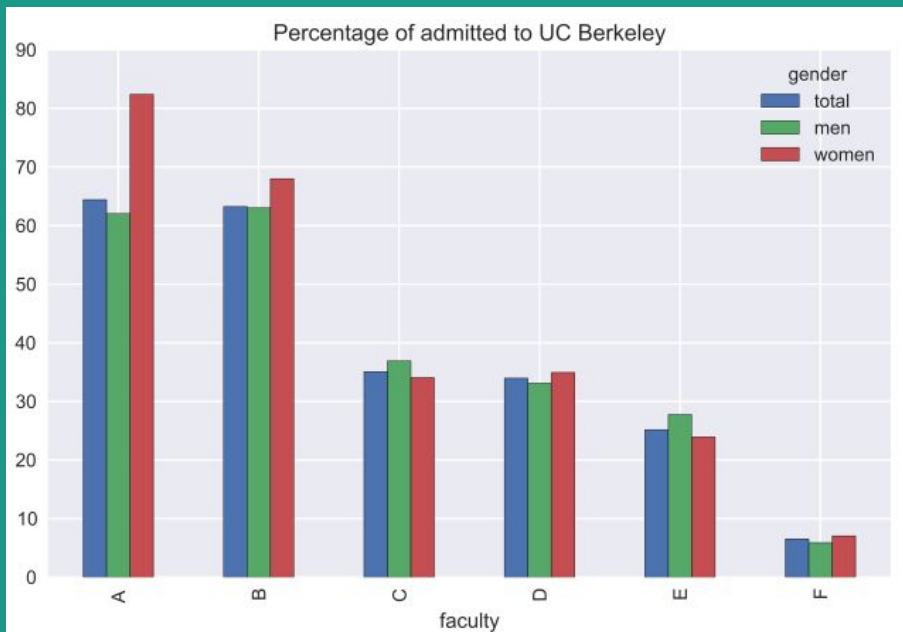
- Статистика показывает то, как люди ведут себя сейчас. С её помощью можно ответить на вопросы о вероятности какого-то события. Например, оценить вероятность попадания в аварию водителей 16 и 18 лет
- Однако когда происходит переход от наблюдения к действию, статистика не помогает сделать правильный вывод о целесообразности и эффективности принятой меры
- Это связано с тем, что данные статистики - это конечный результат огромной совокупности факторов, чей вклад также нужно понимать и оценивать.

## Парадокс Симпсона: Университет Беркли, 1973

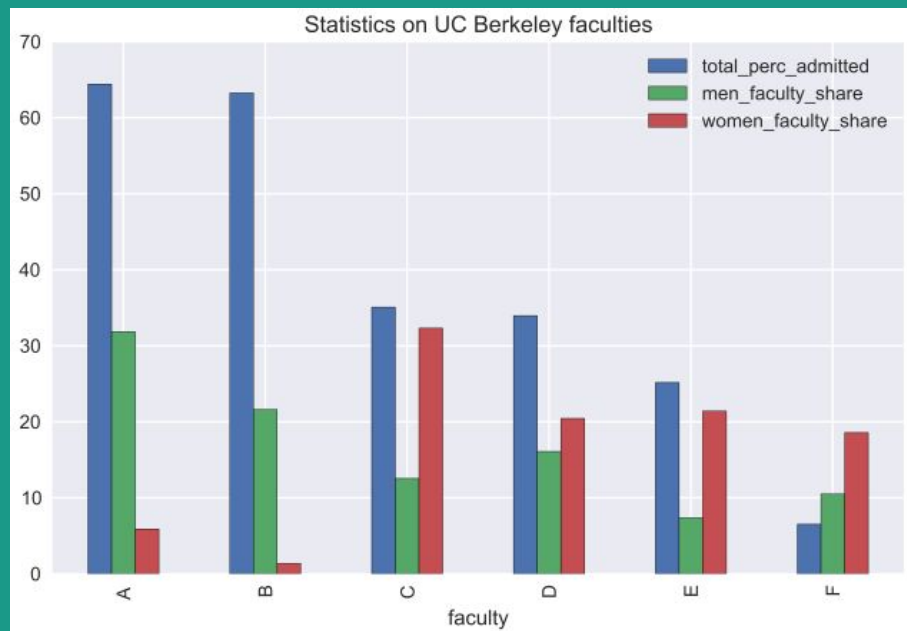
param	accepted	applied	perc_admitted
gender			
men	1192	2590	46.02
women	557	1835	30.35

gender	men		women		total	
param	accepted	applied	accepted	applied	accepted	applied
faculty						
A	512	825	89	108	601	933
B	353	560	17	25	370	585
C	120	325	202	593	322	918
D	138	417	131	375	269	792
E	53	191	94	393	147	584
F	16	272	24	341	40	613

# Парадокс Симпсона: Университет Беркли, 1973



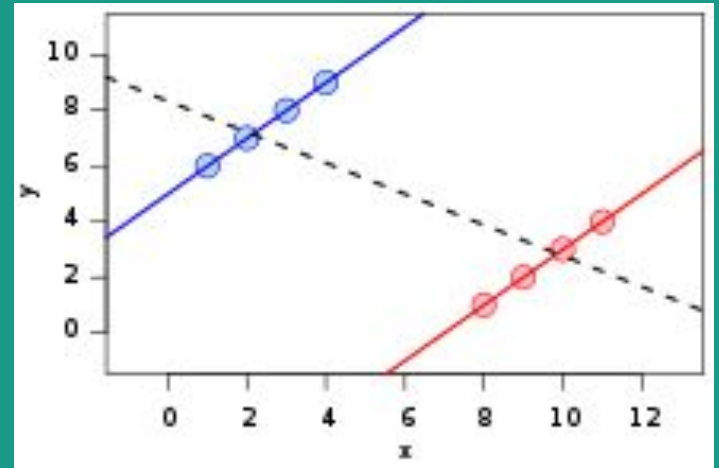
# Парадокс Симпсона: Университет Беркли, 1973



# Парадокс Симпсона:

**Парадокс Симпсона** — явление в статистике, когда при наличии двух групп данных, в каждой из которых наблюдается одинаково направленная зависимость, при объединении этих групп направление зависимости меняется на противоположное.

Причина парадокса заключается в некорректном усреднении двух групп данных с различной долей контрольных наблюдений.



## Парадокс Симпсона

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax	Fully Vax	
All ages			214	301	Vax don't work!

## Парадокс Симпсона

Age	Population (%)		Severe cases		Efficacy vs. severe disease
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	67.5%



# Парадокс Симпсона

Age	Population (%)		Severe cases		Efficacy vs. severe disease
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	<b>67.5%</b>
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	<b>91.8%</b>
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	<b>85.2%</b>

# Парадокс Симпсона

Age	Population (%)		Severe cases/100k		Severe Case Risk	Efficacy
	% Not Vax	% Fully Vax	Not Vax	Fully Vax	Ratio w/ 30-39 <u>UnVax</u>	vs. severe disease
12-15	62.1%	29.9%	0.30	0.00	1/20x	100%
16-19	21.9%	73.5%	1.60	0.00	1/4x	100%
20-29	20.5%	76.2%	1.50	0.00	1/4x	100%
30-39	16.2%	80.9%	6.20	0.20	1	96.8%
40-49	13.2%	84.4%	16.50	1.00	2.7x	93.9%
50-59	10.0%	88.0%	40.20	2.90	6.5x	92.8%
60-69	8.8%	89.8%	76.60	8.70	12.4x	88.7%
70-79	4.2%	94.6%	190.10	19.80	30.7x	89.6%
80-89	5.6%	92.6%	252.30	47.90	40.7x	81.1%
90+	6.1%	90.5%	510.9	38.60	82.4x	92.4%

# Каузальная модель

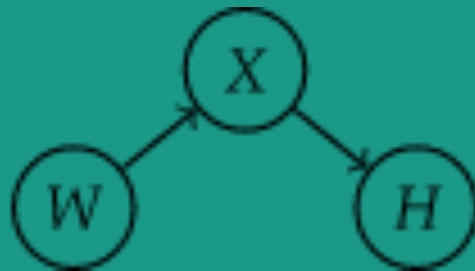
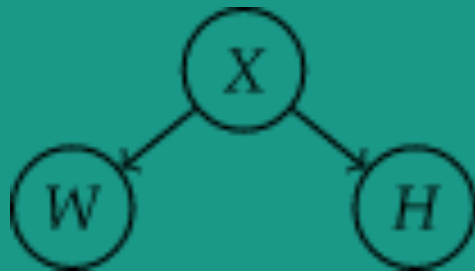
- Каузальная модель — это математическая модель, представляющая причинно-следственные связи внутри отдельной системы или группы.
- Формально мы можем сказать, что каузальная модель - это набор переменных  $X_1, \dots, X_d$ , где каждая переменная определяется как  $X_i = f_i(P_i, U_i), i=1, \dots, d$ .

# Каузальные графы

Часто для демонстрации причинно-следственных связей используют каузальные графы.

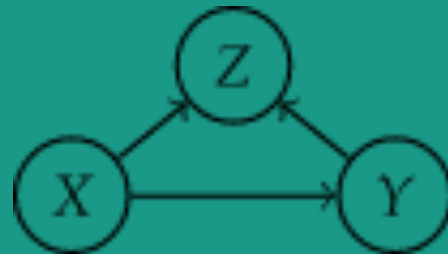
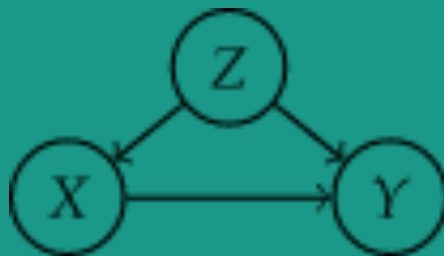
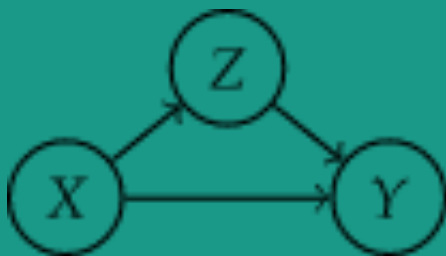
Вершина - событие, фактор

Две вершины A и B соединены ребром (стрелкой,  $A \rightarrow B$ ), если изменение A вызывает изменение B



## Каузальные графы: виды

- 1) Цепь
- 2) Вилка
- 3) Коллайдер



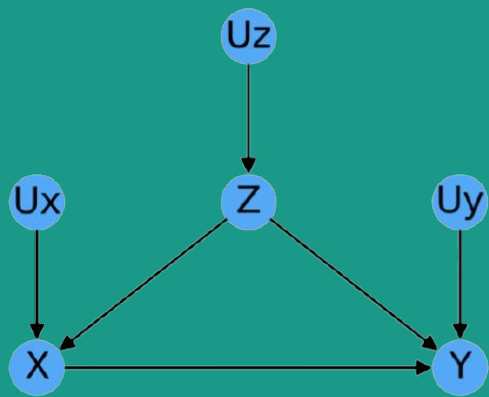
## Вмешательства и причинно- следственные связи



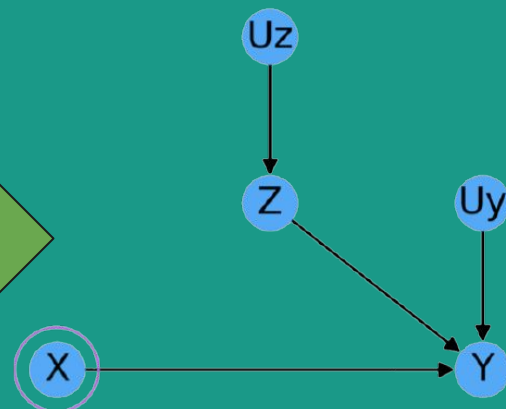
## Пример парадокса Симпсона

Recovery	Drug	No Drug
Men	81/87 (93%)	234/270 (87%)
Women	192/263 (73%)	55/80 (69%)
Total	273/350 (78%)	289/350 (83%)

# Do-операция



$P(X, Y, Z)$



$P_m(X, Y, Z)$



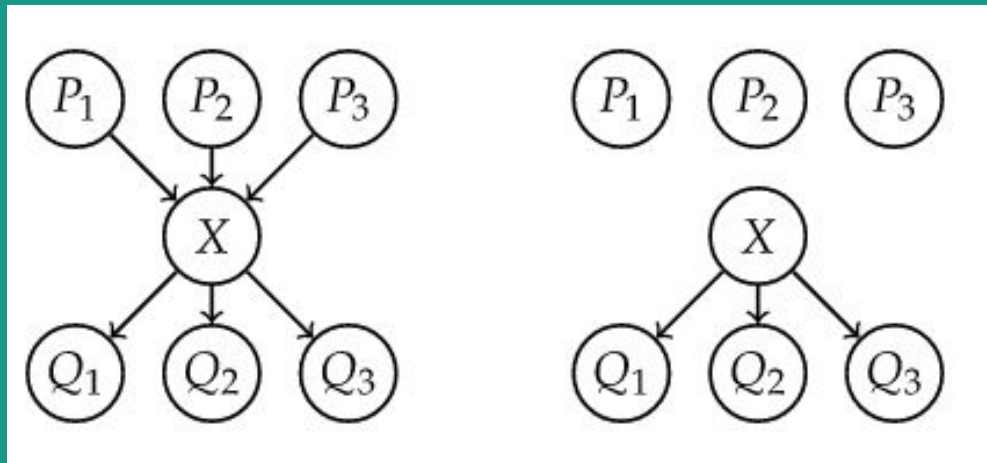
# Do-операция

Имеем:

- причинно-следственную модель
- несбалансированные данные

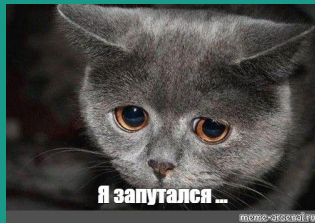
Хотим:

- Уменьшить количество влияющих факторов
- Проверить гипотезу о влиянии X на Y



$$P\{Y=y \mid \text{do}(X:=x)\} \neq P\{Y=y \mid X=x\}$$

**Запутывание**



## Корректирующая формула

$$\mathbb{P}\{Y = y \mid \text{do}(X := x)\} = \sum \mathbb{P}\{Y = y \mid X = x, PA = z\} \mathbb{P}\{PA = z\}$$

*Почему не формула полной вероятности?*

Recovery	Drug	No Drug
Men	81/87 (93%)	234/270 (87%)
Women	192/263 (73%)	55/80 (69%)
Total	273/350 (78%)	289/350 (83%)

$$P(\text{Recovery}|\text{do}(\text{Drug})) = 0.93 \cdot \frac{87 + 270}{700} + 0.73 \cdot \frac{263 + 80}{700} = 0.832$$

$$P(\text{Recovery}|\text{do}(\text{No Drug})) = 0.87 \cdot \frac{87 + 270}{700} + 0.69 \cdot \frac{263 + 80}{700} = 0.7818$$

Комбинаторный взрыв!



—

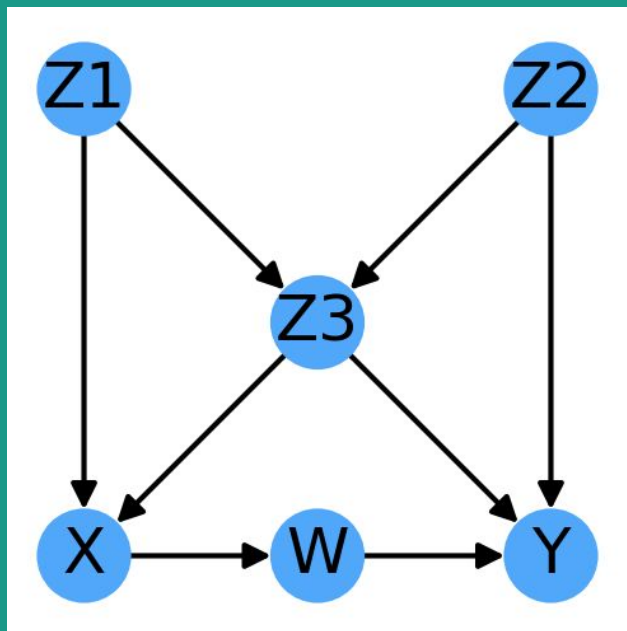
## “Backdoor” критерий

Задача: найти контролирующее множество  $Z$  для корректного вычисления  $P\{Y | do(X = x)\}$ .

Критерий:

1.  $X$  не должен быть предком ни одной вершины из  $Z$
2.  $Z$  блокирует все пути (неориентированные) из  $X$  в  $Y$  такие, что они содержат входящее ребро в  $X$

## Пример



**Можно ли избежать  
нежелательных причинно-  
следственных связей на  
этапе сбора данных?**

—

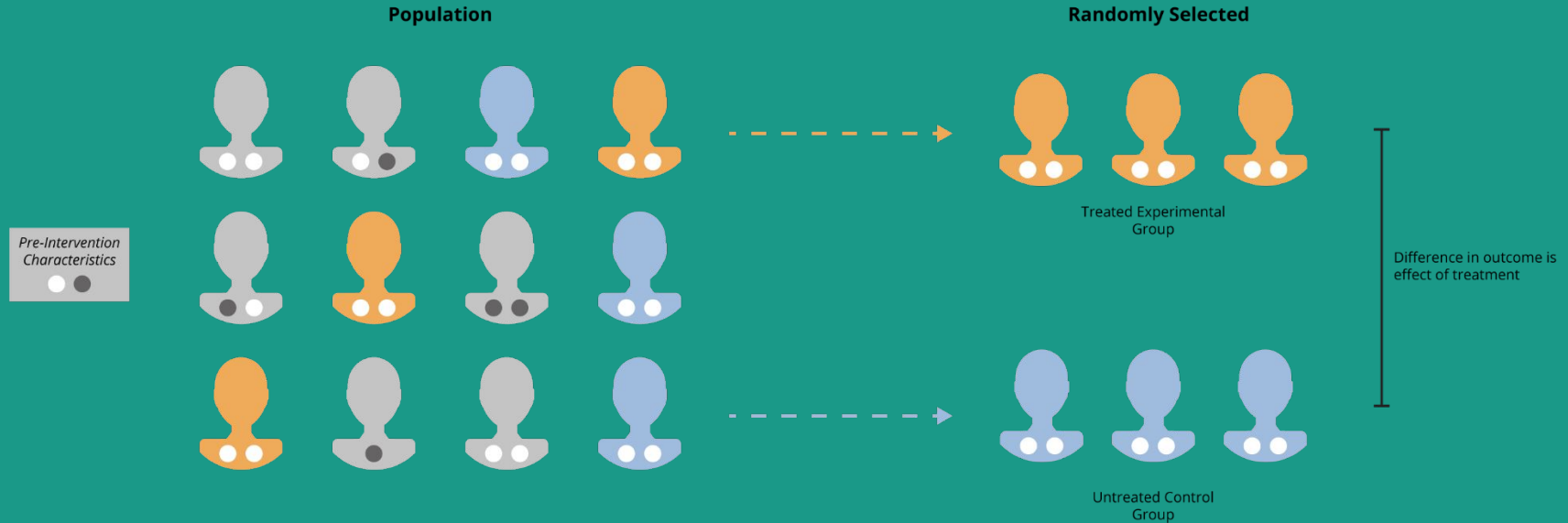


# Causal Questions


---

- Вызывает ли курение рак?
- Улучшают ли дополнительные занятия результаты тестов учащихся?
- Делают ли физические упражнения людей счастливее?
- Повышает ли грудное вскармливание IQ ребенка?

# Randomized experiments



# Treatment effect


$$\hat{y} = b_0 + b_1 T$$

$$b_1 = \bar{y}_T - \bar{y}_C$$

 $\bar{y}_T$ 

- среднее значение результатов в группе лечения

 $\bar{y}_C$ 

- среднее значение в контрольной группе

# Magic of randomization



# Магия в действии, пример

---

- Есть многообещающая новая учебная программа для обучения математике воспитанников детских садов
- Хотим знать, эффективна ли эта учебная программа.



# Данные о воспитанниках

Observations: 335

Variables: 10

```
$ ID          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
$ FEMALE      <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, ...
$ MINORITY    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, ...
$ MOM_ED      <chr> "Some college", "Vocational/technical program", "Some co...
$ DAD_ED      <chr> "Vocational/technical program", "Some college", "Bachelo...
$ SES_CONT     <dbl> -0.27, -0.03, 0.48, -0.03, -0.66, 1.53, 0.20, 0.07, -0.3...
$ READ_pre    <dbl> 27.4, 32.5, 48.2, 43.9, 36.1, 95.8, 33.8, 33.1, 32.2, 44...
$ MATH_pre     <dbl> 18.7, 30.6, 31.6, 31.4, 24.2, 49.8, 27.1, 27.4, 25.1, 41...
$ Trt_rand     <dbl> 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, ...
$ Trt_non_rand <dbl> 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, ...
```

**MINORITY** - принадлежность к меньшинству

**SES** - socio-economic status (оценка социально экономического статуса)

**READ\_pre/MATH\_pre** - reading and math scores

**Trt\_rand** - переменная-показатель (в группе лечения ( $trt\_rand == 1$ ), в контрольной группе ( $trt\_rand == 0$ ))

```
ed_data %>%  
  count(Trt_rand)
```

```
# A tibble: 2 x 2  
  Trt_rand     n  
    <dbl> <int>  
1         0  168  
2         1  167
```

*Разделим детей на группы случайно (по Trt\_rand). Сравним характеристики в каждой из групп.*

```
ed_data %>%  
  group_by(Trt_rand) %>%  
  summarise_if(is.numeric, mean) %>%
```

```
# A tibble: 2 x 6  
  Trt_rand FEMALE MINORITY SES_CONT READ_pre MATH_pre  
    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  
1         0  0.542   0.327   0.296   46.8    39.0  
2         1  0.569   0.293   0.320   48.2    39.9
```

# Результат

---

В нашем гипотетическом рандомизированном эксперименте:

- Внедрили бы в лечебную группу новую учебную программу
- Измерили бы баллы по предметам
- Сравнили бы результаты
- Сделали бы вывод об успешности новой программы



*А что, если мы разделим детей на группы не рандомно (по Trt\_non\_rand)*

```
ed_data %>%  
  group_by(Trt_non_rand) %>%  
  summarise_if(is.numeric, mean) %>%  
  select(-c(ID, Trt_rand))
```

```
# A tibble: 2 x 6
```

	Trt_non_rand	FEMALE	MINORITY	SES_CONT	READ_pre	MATH_pre
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	0.565	0.226	0.912	51.7	43.6
2	1	0.545	0.395	-0.300	43.3	35.3

# Результат

---

- Дисбаланс по многим переменным до лечения
- Различия в результатах тестов в конце года, было бы трудно определить, были ли эти различия вызваны вмешательством или некоторыми из этих других различий до лечения.

Recovery	Drug	No Drug
Men	81/87 (93%)	234/270 (87%)
Women	192/263 (73%)	55/80 (69%)
Total	273/350 (78%)	289/350 (83%)

# Causal Questions

---

- Вызывает ли курение рак?
- Улучшают ли дополнительные занятия результаты тестов учащихся?
- Делают ли физические упражнения людей счастливее?
- Повышает ли грудное вскармливание IQ ребенка?

I USED TO THINK  
CORRELATION IMPLIED  
CAUSATION.



THEN I TOOK A  
STATISTICS CLASS.  
NOW I DON'T.



SOUNDS LIKE THE  
CLASS HELPED.

WELL, MAYBE.

