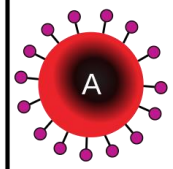
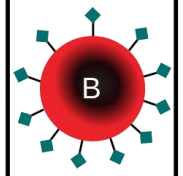
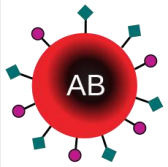
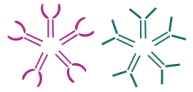
Хи-квадрат, ANOVA и непараметрические тесты

Признаки

Категориальные

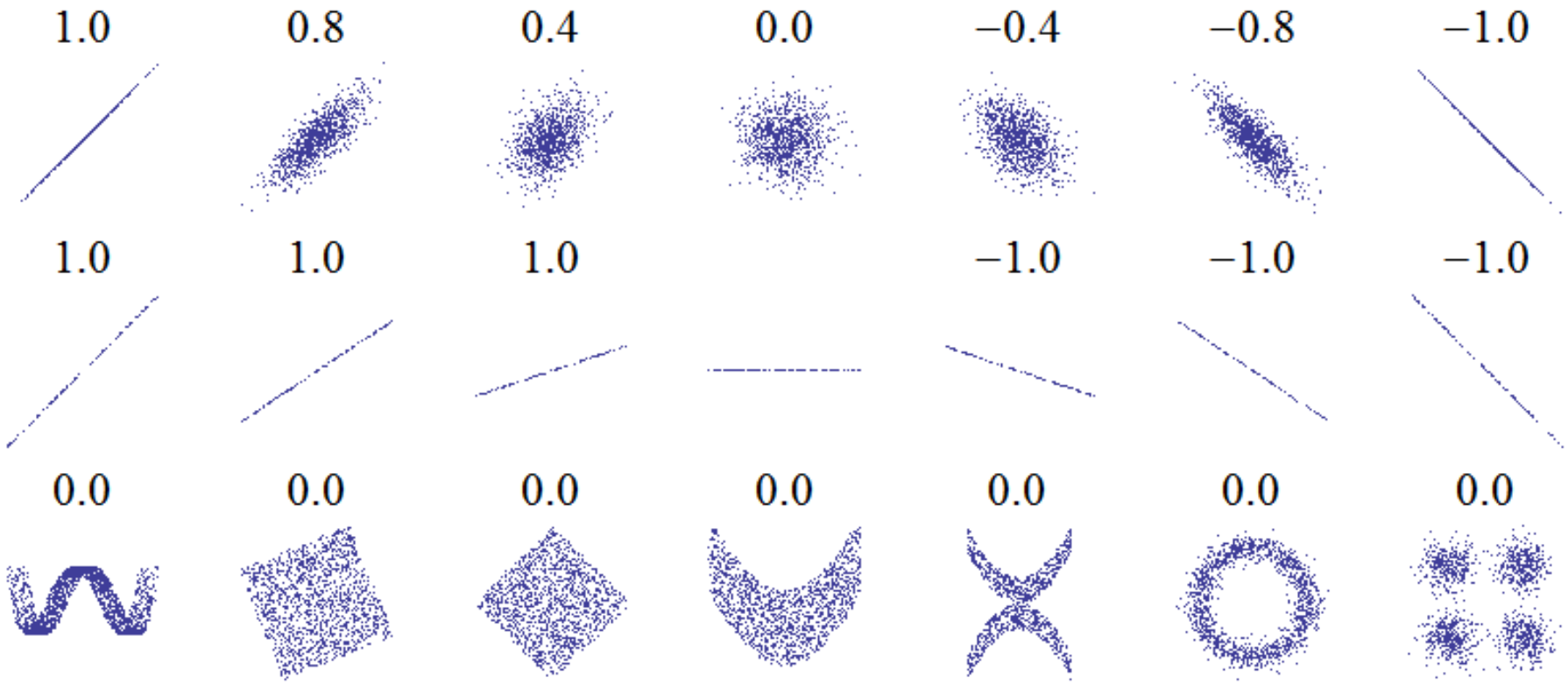
	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in plasma	 Anti-B	 Anti-A	None	 Anti-A and Anti-B
Antigens in red blood cell	A antigen	 B antigen	 A and B antigens	None

Ординальные

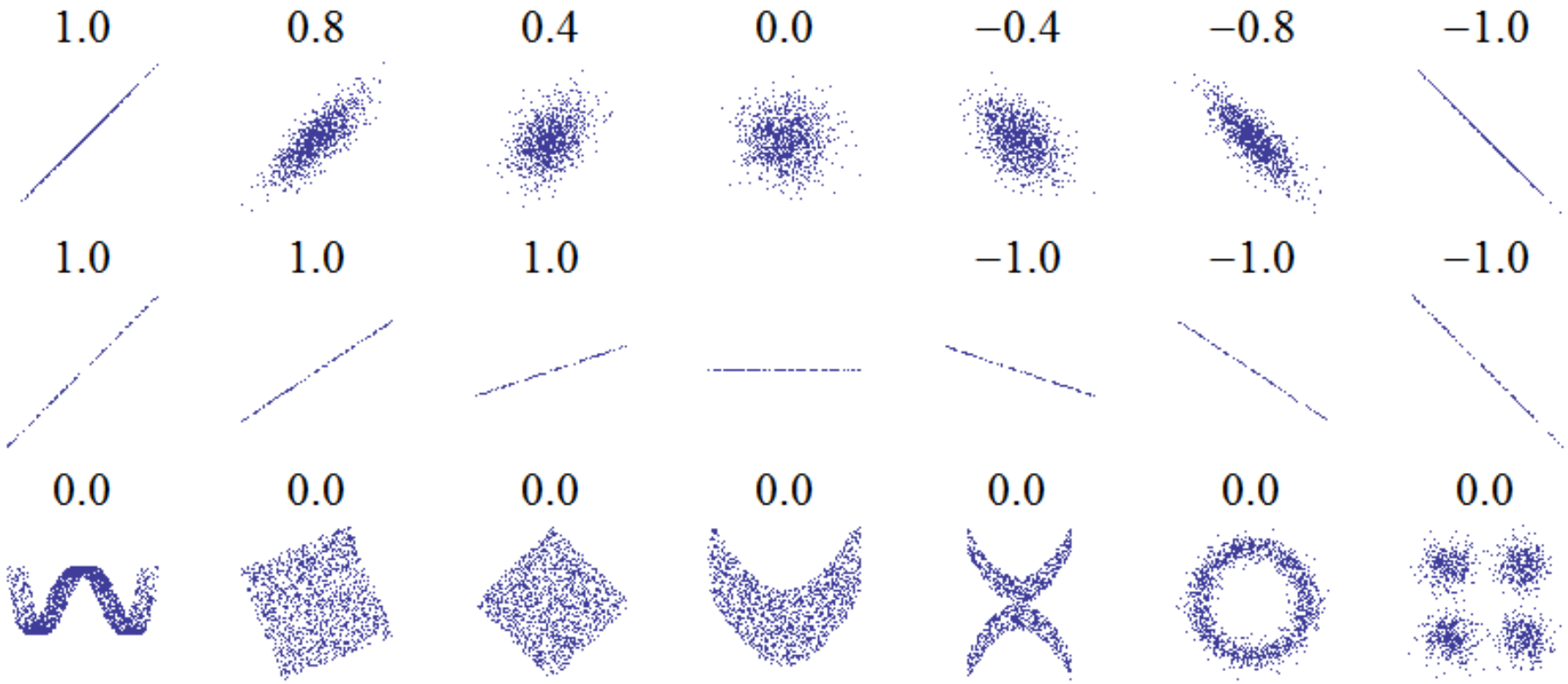
Вещественные

Корреляция

На прошлом занятии мы с вами разобрали корреляцию. Какое требование предъявляет корреляция к переменным, к которым применяется?

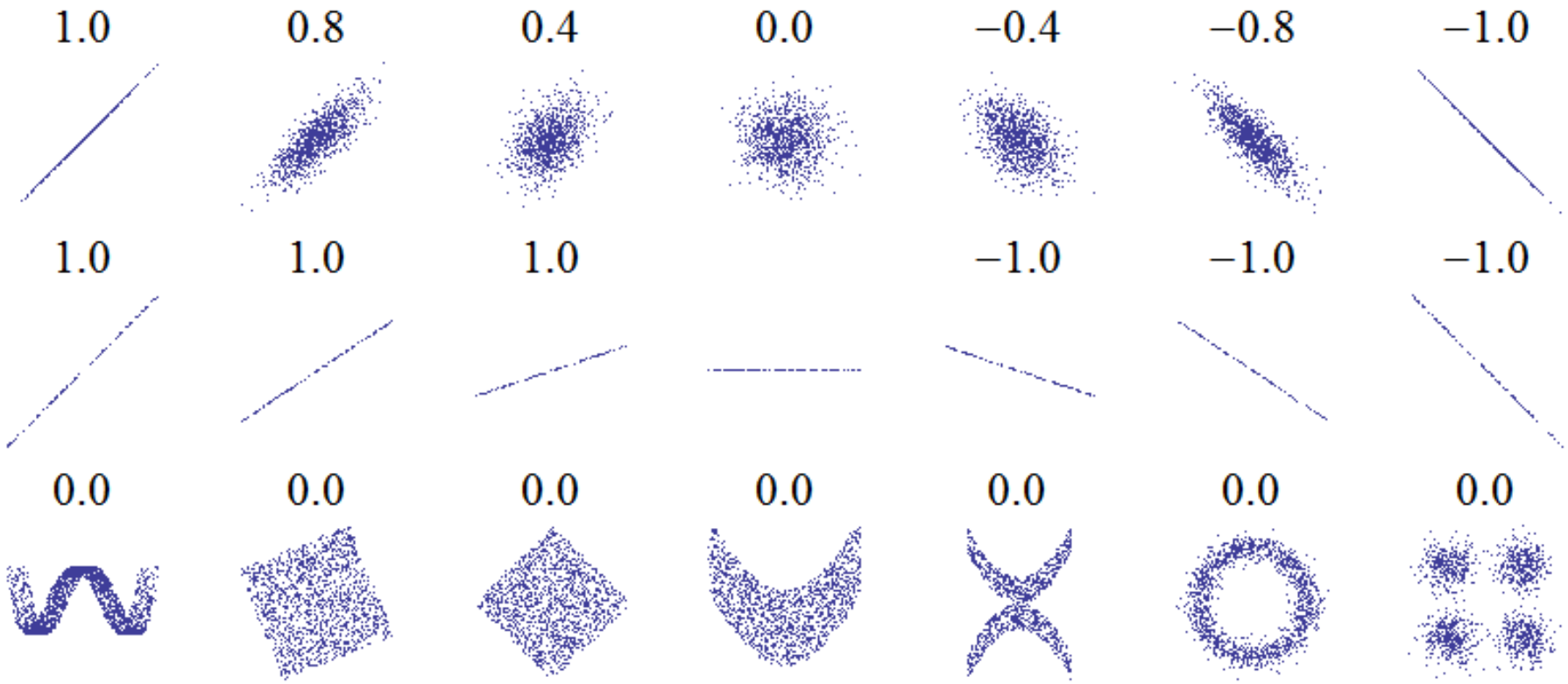


Корреляция



Все разобранные коэффициенты корреляции хорошо применимы к вещественным признакам

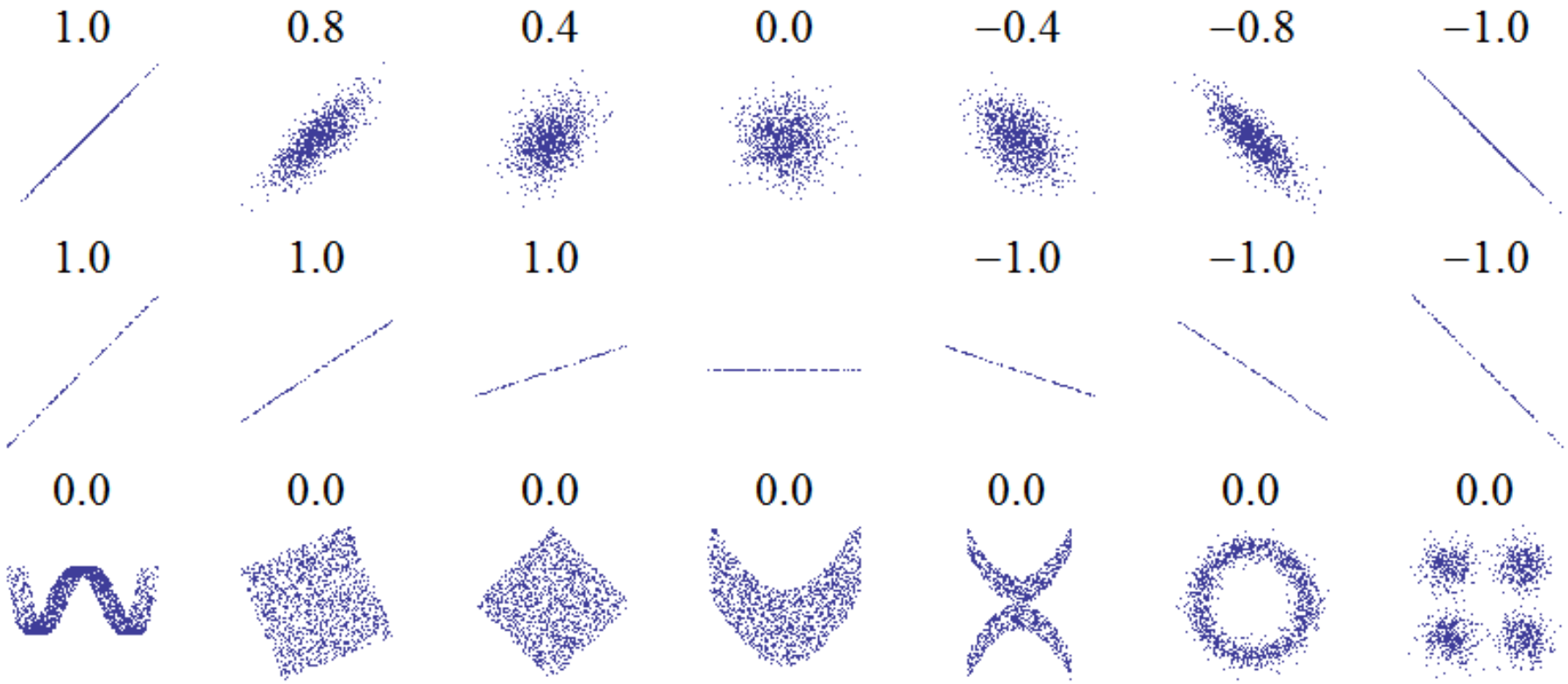
Корреляция



Все разобранные коэффициенты корреляции хорошо применимы к вещественным признакам

Корреляции Спирмена и Тау-Кенделла можно применять к ординальным признакам

Корреляция



Разобранные корреляции нельзя применять к категориальным признакам. И комбинациям категориальных и других.

Связь между двумя категориальными признаками

Допустим у нас есть два категориальных признака и мы хотим проверить, насколько они связаны друг с другом.

	Курит	**Не курит**
Мужчина	17	23
Женщина	25	58

χ^2 тест на сопряженность

H_0: связи между категориями нет

H_1: связь между категориями есть

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

χ^2 тест на сопряженность

H_0: связи между категориями нет

H_1: связь между категориями есть

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

В предположении H_0 вероятность встретить курящего мужчину есть просто произвдение вероятностей встретить курящего человека и мужчину. Аналогично для некурящей женщины и т.д

$$P(AB) = P(A)P(B)$$

χ^2 тест на сопряженность

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

$$P(AB) = P(A)P(B)$$

Откуда взять вероятности P(A) и т.д ?

Оценить из таблицы

χ^2 тест на сопряженность

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

$$P(AB) = P(A)P(B)$$

Откуда взять вероятности P(A) и т.д ?

Оценить из таблицы

χ^2 тест на сопряженность

	Курит	**Не курит**	**Маргинальные вероятности**
Мужчина	17	23	40 / 123 – вероятность встретить мужчину
Женщина	25	58	83/123 - женщину
Маргинальные вероятности	42/123 – курящего человека	81/123	

χ^2 тест на сопряженность

	Курит	Не курит	Маргинальные вероятности
Мужчина	0.111	0.214	40 / 123
Женщина	0.230	0.444	83/123
Маргинальные вероятности	42/123	81/123	

χ^2 тест на сопряженность

Теперь в предположении верности нулевой гипотезы подсчитаем, сколько мы ожидали увидеть людей в каждой ячейке таблицы

χ^2 тест на сопряженность

$$N = 123$$

	Курит	Не курит
Мужчина	0.111 * 123 = 13.653	26.322
Женщина	28.29	54.612

χ^2 тест на сопряженность

У нас теперь есть две таблицы – ожидаемое и наблюдаемое. Кажется, отличаются они не сильно. Но нам нужно получить какое-то одно число, чтобы по нему считать p-value

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

	Курит	Не курит
Мужчина	13.653	26.322
Женщина	28.29	54.612

χ^2 тест на сопряженность

	Курит	Не курит
Мужчина	17	23
Женщина	25	58

	Курит	Не курит
Мужчина	13.653	26.322
Женщина	28.29	54.612

$$V_{ij} = \frac{(O_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}}$$

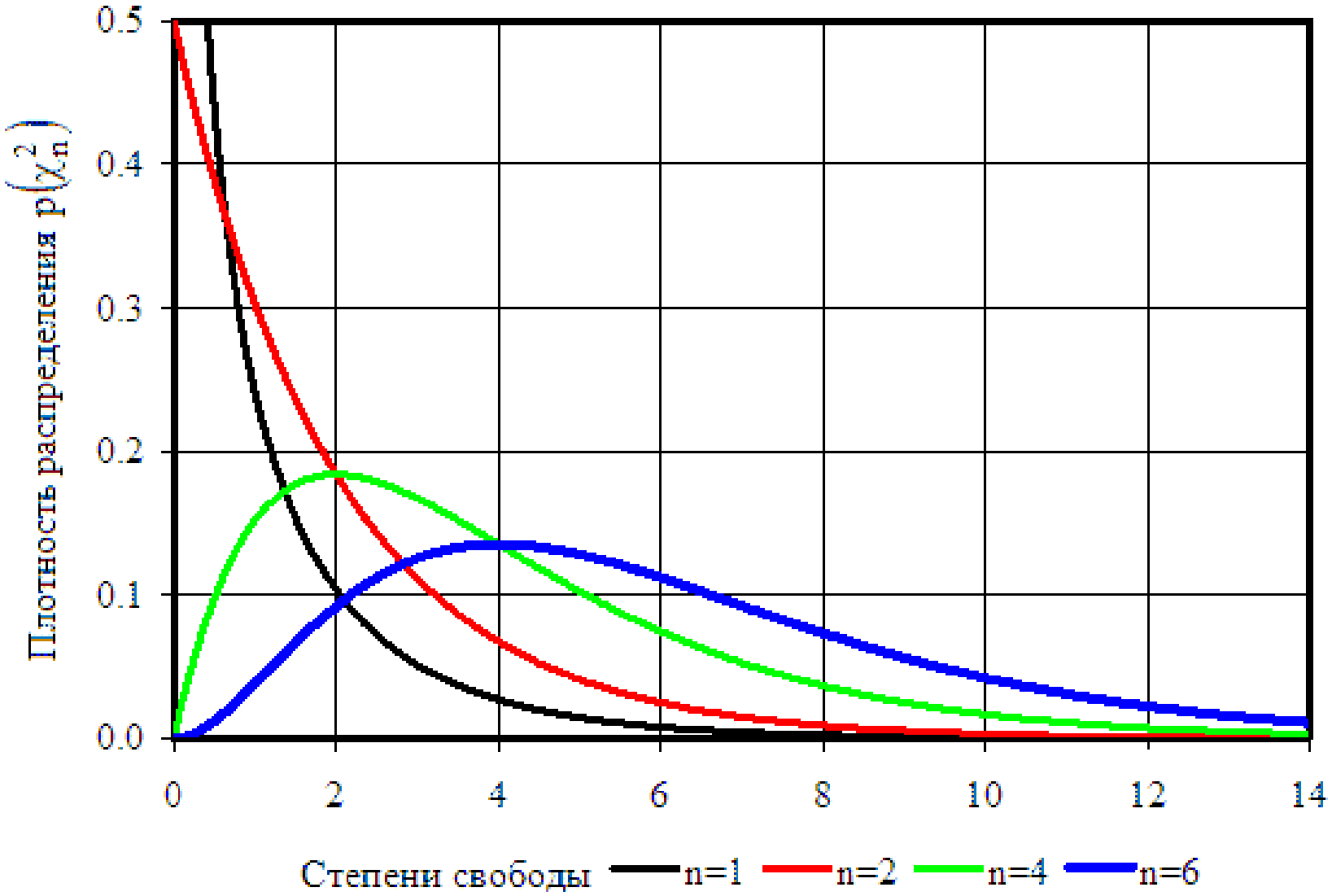
$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}}$$

χ^2 тест на сопряженность

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}}$$

Оказывается, полученное число распределено согласно особому распределению - χ^2 с числом степеней свободы 1

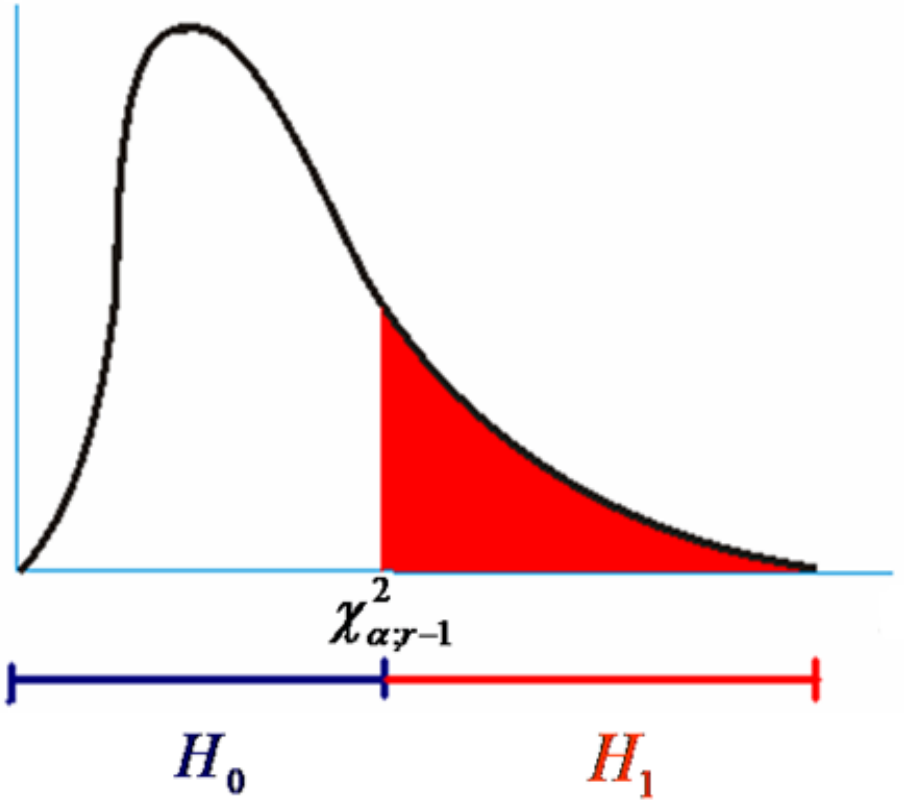
χ^2 – распределение



χ^2 тест на сопряженность

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \mathrm{E}_{ij})^2}{\mathrm{E}_{ij}}$$

Оказывается, полученное число распределено согласно особому распределению - χ^2 с числом степеней свободы 1. Причем, нам нужен только правый хвост (нас интересуют только ситуации более перекошенные чем среднее по больнице, нас не интересуют менее перекошенные)



Отсюда и получаем p-value – 0.1758

χ^2 тест на сопряженность

Можем ли провести вся озвученные выше действия для таблицы не 2х2?

χ^2 тест на сопряженность

	Первая группа крови	Вторая группа крови	Третья группа крови	Четвертая группа крови
Предпочитает ликер	…	…	…	…
Предпочитает пиво	…	…	…	…
Предпочитает сидр	…	…	…	…

χ^2 тест на сопряженность

Можем ли провести вся озвученные выше действия для таблицы не 2х2?

Да, можем для любой таблицы такого вида. Только число степеней свободы будет другое

χ^2 тест на сопряженность

	F_{11}	F_{12}	…	F_{1K}
F_{21}	…	…	…	…
F_{22}	…	…	…	…
…	…	…	…	…
F_{2M}	…	…	…	…

В общем случае применяется для любых двух категориальных переменных – получается таблица К на М, где К – число уровней первой переменной, М – число уровней второй . Полученное по приведенной выше процедуре число всегда следует распределению χ^2, число степеней свободы вычисляется по формуле

$$df = (K-1)(M-1)$$

χ^2 тест на сопряженность

- Используется для определения связи двух категориальных переменных
- Можно использовать для определения связи категориальной и ординальной переменных, но есть более мощные методы
- Не путать с χ^2 тестом на соответствие распределению, который разберем позже. Хотя они и родственны

χ^2 тест на сопряженность - допущения

- Мы должны достаточно точно оценить вероятности + статистика должна сойтись к нужному распределению
- Отсюда вытекает требование, что ожидаемое число в каждой ячейке таблицы должно быть не меньше 5

R chisq.test

Pearson's Chi-squared Test for Count Data

Description

`chisq.test` performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Точный тест Фишера

Если предположения χ^2 не выполняются

	Исход есть	Исхода нет	Всего
Фактор есть	A	B	A+ B
Фактора нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

$$p(table) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

Точный тест Фишера

Левый хвост, сложить вероятности всех таблиц здесь

Все хорошо

Правый хвост, сложить вероятности всех таблиц здесь

Таблица, перекошенная, как наша, но в другую сторону

Наша таблица

Таблицы с еще более перекошенной в другую сторону связью

	Исход есть	Исход а нет	Всего
Фактор есть	A	B	A+B
Фактора нет	C	D	C+D
Всего	A+C	B+D	A+B+C+D

	Исход есть	Исход а нет	Всего
Фактор есть	A	B	A+B
Фактора нет	C	D	C+D
Всего	A+C	B+D	A+B+C+D

Таблицы с еще более перекошенной в нашу сторону связью

Точный тест Фишера

Fisher's Exact Test for Count Data

Description

Performs Fisher's exact test for testing the null of independence of rows and columns in a contingency table with fixed marginals.

Usage

```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
            hybridPars = c(expect = 5, percent = 80, Emin = 1),
            control = list(), or = 1, alternative = "two.sided",
            conf.int = TRUE, conf.level = 0.95,
            simulate.p.value = FALSE, B = 2000)
```

Связь между категориальным и вещественным признаком

Сначала разберем случай, когда категориальный признак – бинарный, то есть принимает только два значения

$$H_0: \mu_a = \mu_b$$

Ничего не напоминает?

Связь между категориальным и вещественным признаком

Сначала разберем случай, когда категориальный признак – бинарный, то есть принимает только два значения

Ничего не напоминает?

t-test $H_0: \mu_a = \mu_b$

Что делать в случае многих групп, чтобы проверить гипотезу:

$$H_0: \mu_a = \mu_b = \mu_c = \cdots = \mu_z$$

Связь между категориальным и вещественным признаком

Что делать в случае многих групп, чтобы проверить гипотезу:

Вариант 1: просто провести попарные тесты между всеми группами. Не забыть сделать поправку на множественное тестирование.

$$H_0: \mu_a = \mu_b \qquad H_0: \mu_a = \mu_c \qquad H_0: \mu_b = \mu_c$$

…

Связь между категориальным и вещественным признаком

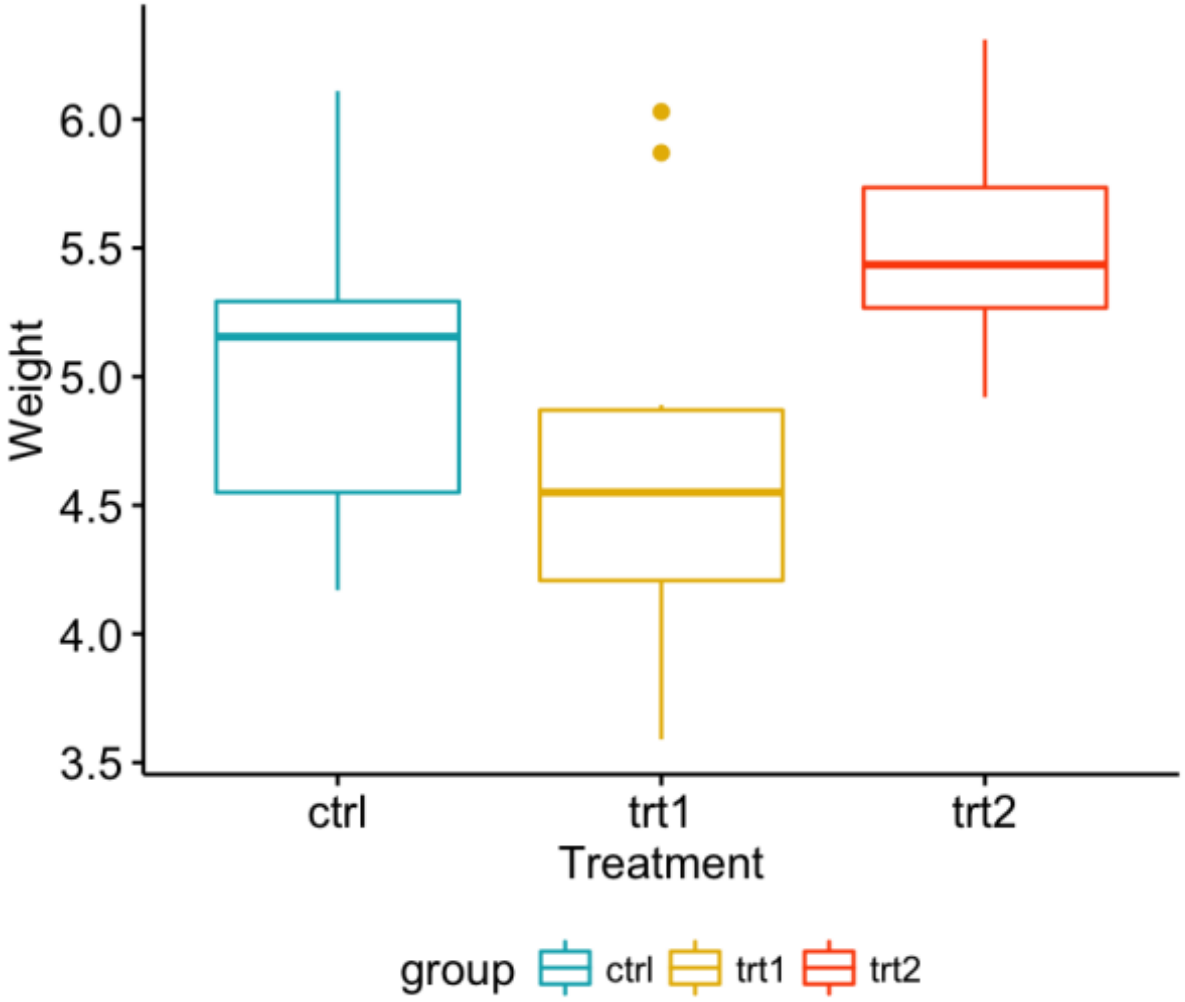
Что делать в случае многих групп, чтобы проверить гипотезу:

Вариант 2: Первый вариант немного излишен, если мы хотим проверить только гипотезу

$$H_0: \mu_a = \mu_b = \mu_c = \cdots = \mu_z$$

Потому будем сначала проверять именно ее, а уже потом, если нам интересно – попарные сравнения. Причем попарные сравнения можно проводить специализированными тестами

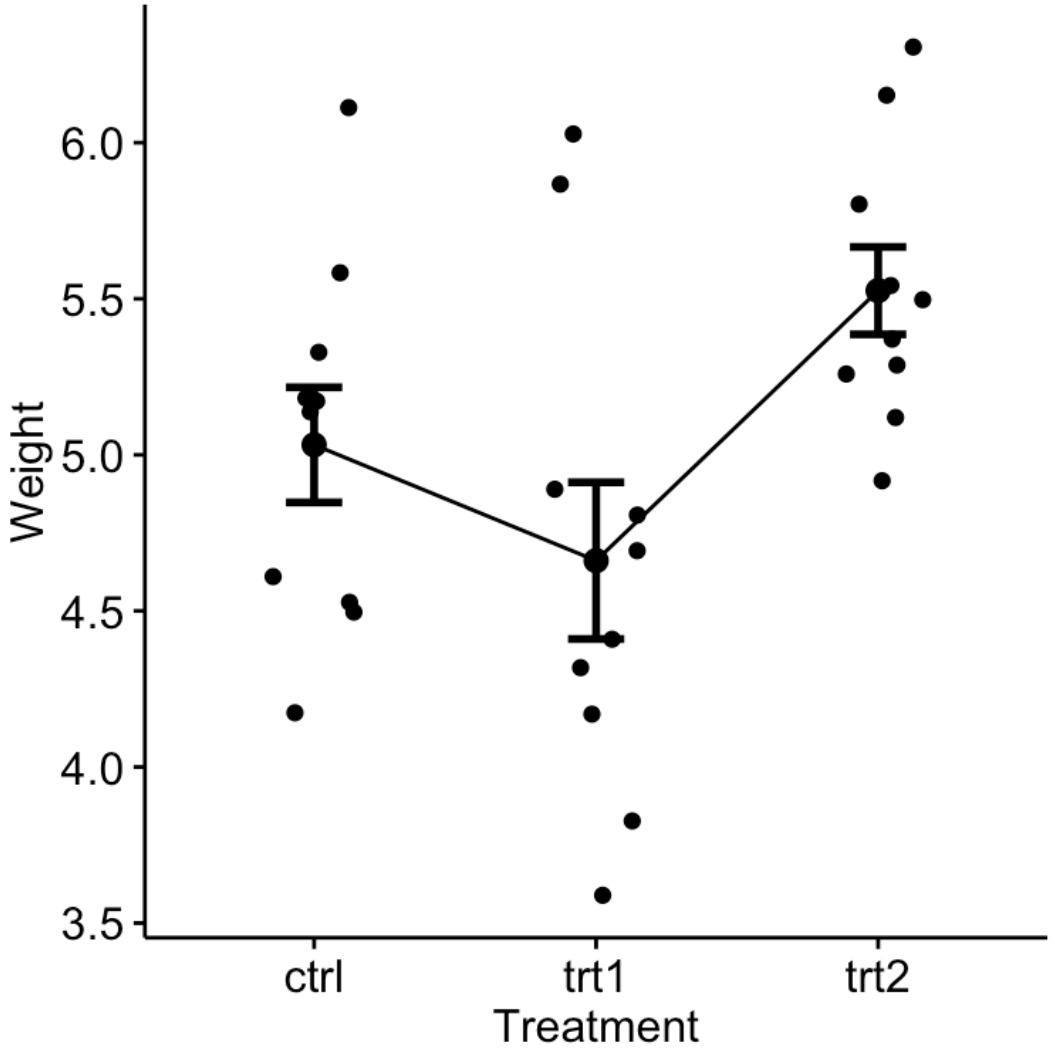
One-way ANOVA



Проверяем гипотезу о том, что средние в группах равны

Часто изображают так, хотя боксплоты показывают медианы, строго говоря)

One-way ANOVA



One-way ANOVA

FactorA	A1	A2	A3
	y_{11}	y_{12}	y_{13}
	y_{21}	...	...
	...	...	...
	y_{n1}		y_{nm}

One-way ANOVA

$$SST = \sum_i \sum_j \left(y_{ij} - \overline{y}\right)^2$$

Сумма квадратов (дисперсия с точностью до числа объектов)

$$SSA = \sum_i \sum_j \left(\overline{y}_{\_j} - \overline{y}\right)^2 = n \sum_j \left(\overline{y}_{\_j} - \overline{y}\right)^2$$

Сумма квадратов, если каждое наблюдение в группе заменить средним в группе – сколько дисперсии объясняется фактором A

$$SSE = \sum_i \sum_j \left(y_{ij} - \overline{y}_{\_j}\right)^2$$

Сколько дисперсии мы не объяснили фактором A

$$SSA = \sum_i \sum_j \left(\bar{y}_{\_j} - \bar{y}\right)^2 = n \sum_j \left(\bar{y}_{\_j} - \bar{y}\right)^2$$

FactorA	A1	A2	A3
	$\bar{y}_{\_1}$	$\bar{y}_{\_2}$	$\bar{y}_{\_3}$
	$\bar{y}_{\_1}$	...	...
	...	...	...
	$\bar{y}_{\_1}$	$\bar{y}_{\_2}$	$\bar{y}_{\_3}$

Сумма квадратов, если каждое наблюдение в группе заменить средним в группе – сколько дисперсии объясняется фактором A

One-way ANOVA

Оказывается, что в предположении верности H_0 верно, что $\frac{SSE}{N-a}$ и $\frac{SSA}{a-1}$ оценивают дисперсию генеральной совокупности, откуда пришли наши группы (мы дополнительно вводим предположение, что дисперсии групп одинаковые). То есть эти оценки должны быть похожи. Разделим их

$$\frac{s_x^2}{s_y^2} \sim F(n, m)$$

Распределение выборочных дисперсий, оцененных по группам размера n и м

$$\frac{\frac{SSA}{a-1}}{\frac{SSE}{N-a}} \sim F(a-1, N-a)$$

One-way ANOVA

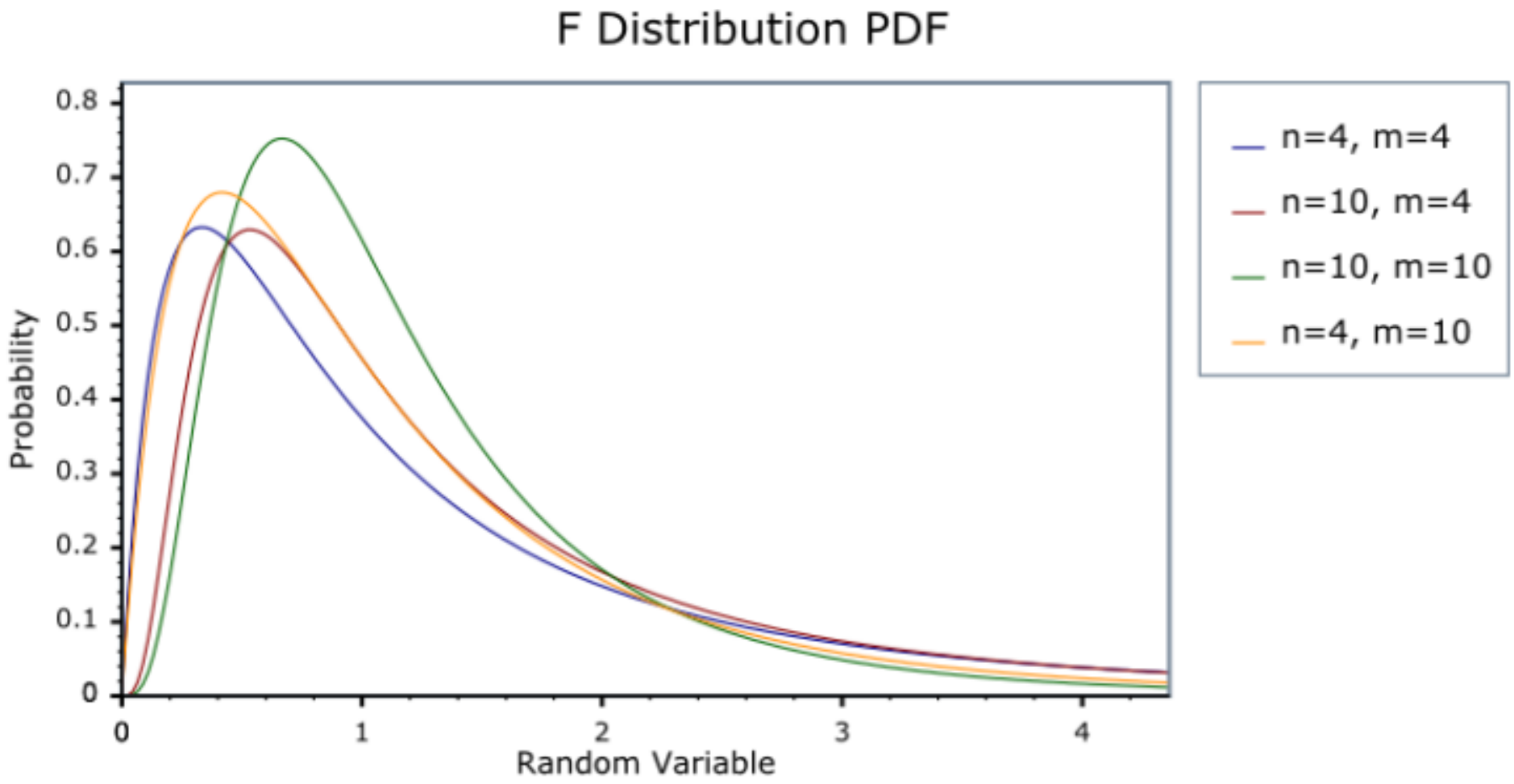
Оказывается, что в предположении верности H_0 верно, что $\frac{SSE}{N-a}$ и $\frac{SSA}{a-1}$ оценивают дисперсию генеральной совокупности, откуда пришли наши группы (мы дополнительно вводим предположение, что дисперсии групп одинаковые). То есть эти оценки должны быть похожи. Разделим одну на другую

$$\frac{s_x^2}{s_y^2} \sim F(n-1, m-1)$$

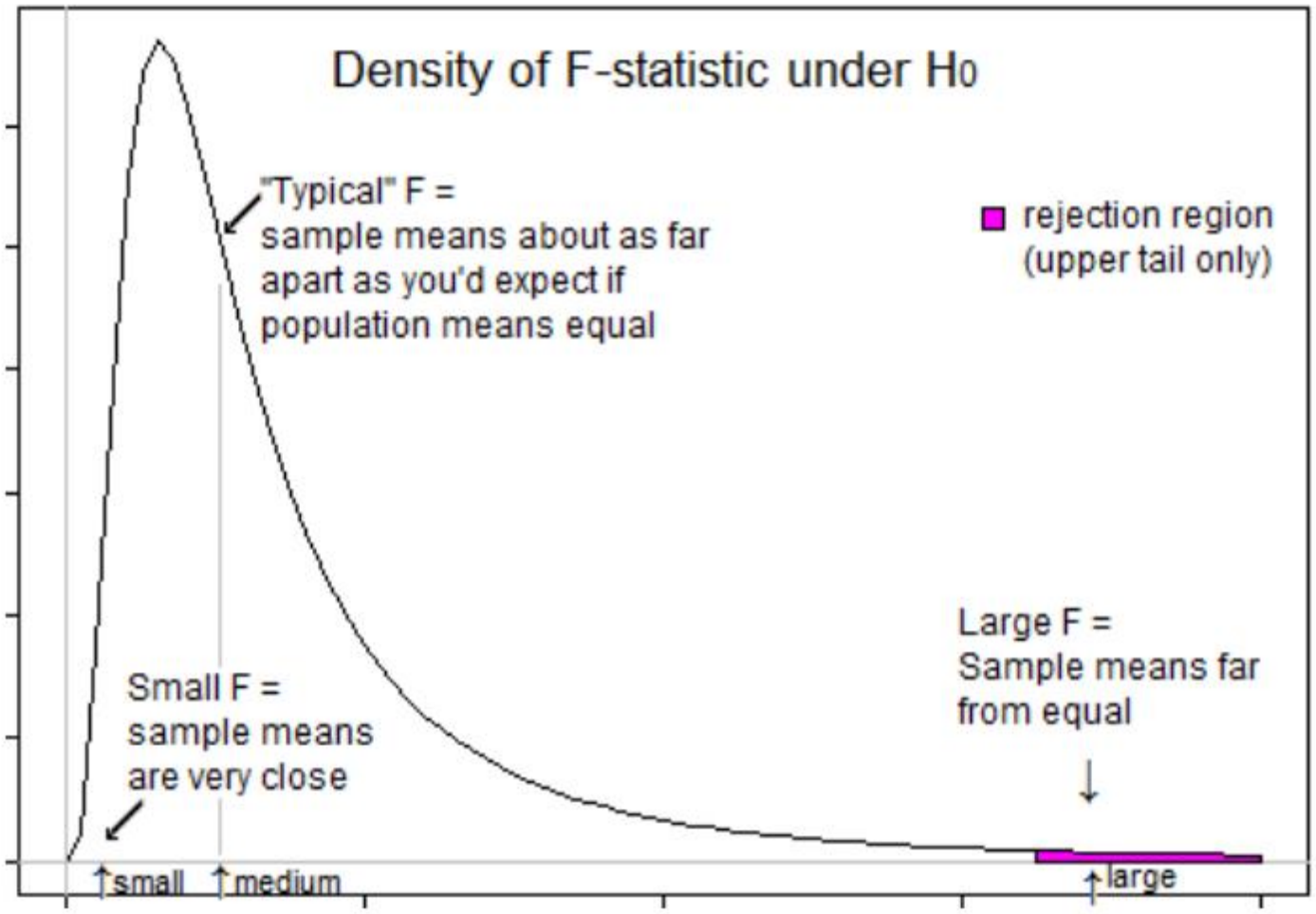
Распределение отношения выборочных дисперсий, оцененных по группам размера n и m

$$\frac{\frac{SSA}{a-1}}{\frac{SSE}{N-a}} \sim F(a-1, N-a)$$

F-распределение



F-распределение



Нам нужен только правый хвост, так как значения F в левом хвосте говорят нам, что средние еще более похожи, чем мы обычно ожидаем. Мы не считаем такие случаи основанием отвергать гипотезу

Предположений 1-way ANOVA

1. Наши наблюдения независимы.
2. Шум в данных распределен нормально
3. Гомоскедастичность – дисперсии в группах одинаковы
4. Для целей нашего исследования фактор одноуровневый – внутри себя он не делится на другие группы. Иначе – n-way ANOVA, repeated measures ANOVA – разберем на следующем занятии

R aov

Fit an Analysis of Variance Model

Description

Fit an analysis of variance model by a call to `lm` for each stratum.

Usage

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,
    contrasts = NULL, ...)
```

R aov

```
res.aov <- aov(weight ~ group, data = PlantGrowth)
# Summary of the analysis
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Бонус

Как проверить гипотезу о равенстве дисперсий между двумя выборками (альтернатива – дисперсии не равны)?

Разделить одно на другое и посчитать p-value отношения по распределению Фишера.

Нюанс – p-value надо домножить на два, так как в этом случае нам важны оба хвоста.

Категориальная переменная и ординальная

По аналогии с вещественным случаем, будем отдельно рассматривать две истории – когда категориальная переменная – бинарная, и когда - нет

Тест Манна-Уитни

Пусть нам даны две выборки X и Y, и мы не можем вводить никаких предположений о нормальности.

В этом случае давайте просто для каждой возможной пары, в которой один объект пришел из X, а второй – из Y, посчитаем следующую величину

$$S(x_i, y_j) = 1\ if\ x_i > y_j\ 0.5\ if\ x_i = y_j\ else\ 0$$

$$U = \sum_i \sum_j S(x_i, y_i)$$

Оказывается, данная величина тоже имеет хорошо подходящее нашим целям распределение и из этого распределения можно посчитать p-value

Тест Манна-Уитни

$$H_0: X \text{ и } Y \text{ пришли из одного распределения}$$

$$H_0: P(X > Y) = 0.5$$

Если мы предполагаем, что формы распределений схожи (а еще лучше - переменные вещественные) то H_0 более приятна

$$H_0: \text{медианы } X \text{ и } Y \text{ совпадают}$$

Тест Манна-Уитни

Wilcoxon Rank Sum and Signed Rank Tests

Description

Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as ‘Mann-Whitney’ test.

Usage

```
wilcox.test(x, ...)

## Default S3 method:
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95,
            tol.root = 1e-4, digits.rank = Inf, ...)

## S3 method for class 'formula'
wilcox.test(formula, data, subset, na.action, ...)
```

Тест Краскала-Уоллиса

Обобщение Манна-Уитни на несколько групп

Тест Краскала-Уоллиса

Kruskal-Wallis Rank Sum Test

Description

Performs a Kruskal-Wallis rank sum test.

Usage

```
kruskal.test(x, ...)

## Default S3 method:
kruskal.test(x, g, ...)

## S3 method for class 'formula'
kruskal.test(formula, data, subset, na.action, ...)
```