

# Статистические тесты однородности символьных последовательностей для информационного анализа электрокардиосигналов

Жариков Илья Николаевич

Московский физико-технический институт  
Факультет управления и прикладной математики  
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. К. В. Воронцов

29 июня 2016 г.

# Цель работы

## Цель работы

Построить статистические тесты для проверки однородности символьных последовательностей, представленных векторами частот  $k$ -грамм.

## Требования

Толерантность к разреженности векторов частот слов.

## Применение

Исследование воспроизводимости дискретизированных ЭКГ-сигналов и метрологическая проверка технологии информационного анализа электрокардиосигналов.

## Z-тест (Z)

Предполагается, что  $n \sim \text{Bin}(l, p)$ .

$$S_1 : [n_{11}, n_{12}, \dots, n_{1K}] \quad \forall m = 1, \dots, K \quad H_0 : p_{1m} = p_{2m};$$

$$S_2 : [n_{21}, n_{22}, \dots, n_{2K}] \quad \xrightarrow{\hspace{1.5cm}} \quad H_1 : p_{1m} \neq p_{2m}.$$

Z-статистика (для  $k$ -граммы под номером  $m$ ):

$$Z_m = \frac{\frac{n_{1m}}{l_1} + \frac{1}{2l_1} - \frac{n_{2m}}{l_2} - \frac{1}{2l_2}}{\sqrt{\frac{n_{1m} + n_{2m}}{l_1 + l_2} \cdot \frac{l_1 + l_2 - n_{1m} - n_{2m}}{l_1 + l_2} \cdot \left(\frac{1}{l_1} + \frac{1}{l_2}\right)}}.$$

Пусть  $U_\beta$  —  $\beta$ -квантиль распределения  $N(0, 1)$ .

**Критерий:**

**$k$ -граммы:**  $|Z_m| \geq U_{1-\frac{\alpha}{2}} \Rightarrow H_0$  отвергается.

**кодограммы:**  $\sum_{m=1}^K \left[ |Z_m| \geq U_{1-\frac{\alpha}{2}} \right] > \alpha \cdot K \Rightarrow S_1$  и  $S_2$  различны.

# Постановка задачи

## Независимость

Однородность  $\Leftrightarrow$  Номер  $k$ -граммы не зависит от номера символьной последовательности.

Рассматриваются вектора:

$$\mathbf{I} = [i_1, \dots, i_q, \dots, i_L] \quad \text{и} \quad \mathbf{J} = [j_1, \dots, j_q, \dots, j_L].$$

Т Таблица сопряженности

		J				$n_{\bullet+}$
		1	2	...	K	
I	1	$n_{11}$	$n_{12}$	...	$n_{1K}$	$n_{1+}$
	2	$n_{21}$	$n_{22}$	...	$n_{2K}$	$n_{2+}$
$n_{+\bullet}$		$n_{+1}$	$n_{+2}$	...	$n_{+K}$	$n$

Проверяемая гипотеза:

$H_0$  : I и J независимы;

$H_1$  : I и J зависимы.

# G-тест (G)

## Независимость

По таблице сопряженности  $\mathbf{T}$ , вычисляется значение статистики:

$$G^2(I, J) = 2 \cdot \sum_{j=1}^K \sum_{i=1}^2 n_{ij} \ln \left( \frac{n_{ij} n}{n_{i+} n_{+j}} \right).$$

В условиях истинности  $H_0$ :  $G^2 \sim \chi_{(2-1)(K-1)}^2$ .

Пусть  $\chi_{\beta}^2$  —  $\beta$ -квантиль распределения  $\chi_{(K-1)}^2$ .

### Критерий:

для  $I$  и  $J$ :  $G^2(I, J) \geq \chi_{1-\frac{\alpha}{2}}^2 \Rightarrow H_0$  отвергается,  $I$  и  $J$  зависимы.

# Тест Фишера (FT)

## Независимость

Для таблицы сопряженности  $T$ , вычисляется значение  $P$ :

$$P = \frac{\prod_{i=1}^2 n_{i+}! \cdot \prod_{j=1}^K n_{+j}!}{n! \cdot \prod_{i=1}^2 \prod_{j=1}^K n_{ij}!}$$

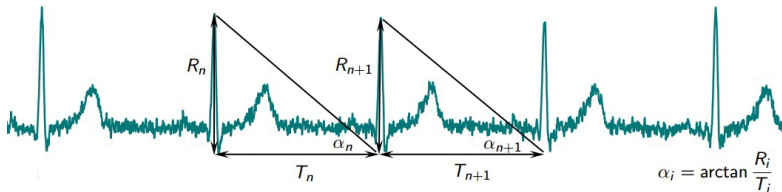
$\{T^r\}_{r=1}^N$  — таблицы сопряженности с суммой по строкам  $n_{i+}$  и столбцам  $n_{+j}$ .

$$\text{p-value} = \sum_{h \in \mathcal{B}} P_h, \quad h \in \mathcal{B} \Leftrightarrow P_h \leq P, \quad \mathcal{B} \subseteq \{1, 2, \dots, N\}.$$

**Критерий:**

для  $I$  и  $J$ :  $\text{p-value} \leq \alpha \Rightarrow H_0$  отвергается,  $I$  и  $J$  зависимы.

## Электрокардиограмма (ЭКГ)



↓ сочетания знаков приращений  $(R, T, \alpha) \Leftrightarrow \{A, B, C, D, E, F\}$

ABCDEFADCEFD BCFDEAF CBCFADECFDEFACBDFECAFDECF  
 ADCEBCFADECF ABCFEDAFC EBCDFDBCEFAFCDEBF EFDCA  
 AFDCBCDFAEDFC DBDFEFACDBFE ABCDFAAEBDCFE CBFDFEF

## Данные

**S**  $|S| = 7626$

Набор ЭКГ, полученных с помощью прибора Скринфакс;

**C**  $|C| = 4918$

Набор ЭКГ, полученных с помощью прибора CardioQvark;

**E**  $|E| = 2 \cdot 23$

Набор ЭКГ, полученных при одновременной записи сигнала с помощью приборов CardioQvark и Скринфакс;

**M**  $|M| = 1000$

Синтетические данные.



## Корректность тестов

**Данные:** Синтетические данные **M**.

**Цель:** Проверить **корректность** тестов.

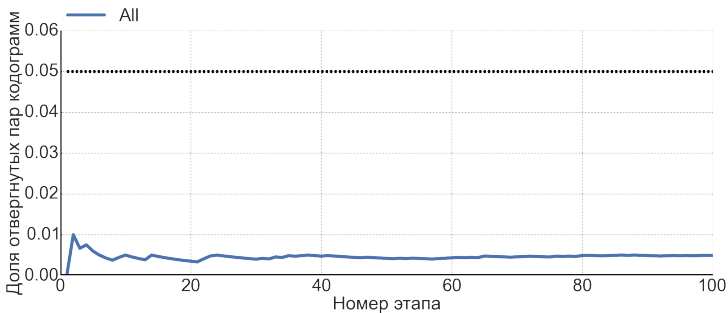
**Эксперимент:** Многократная проверка однородности пары кодограмм, выбранных случайным образом.

- Тесты:**
- Тест Фишера;
  - G-тест;
  - Z-тест.

**Корректным** тестом является тест, у которого ошибка первого рода не превосходит заявленный уровень значимости.

# Корректность тестов

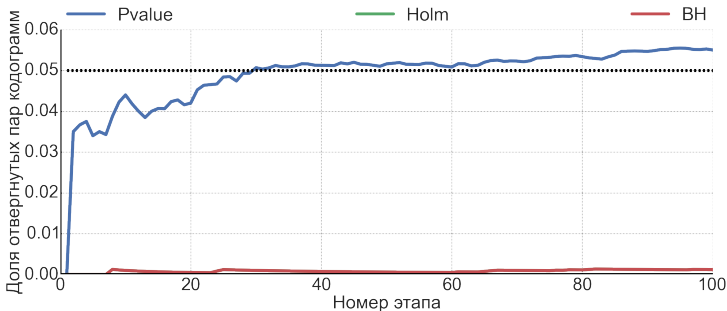
## Z-тест



**Вывод:** Z-тест корректен.

# Корректность тестов

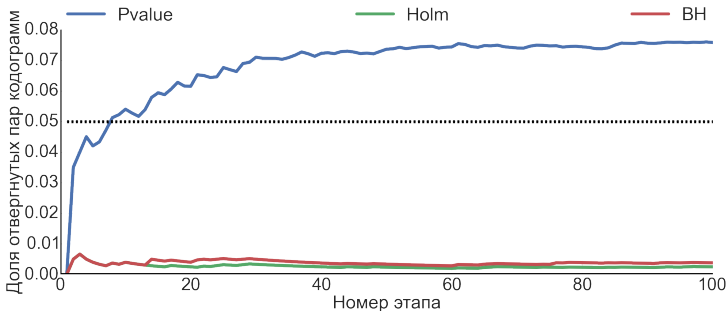
## G-тест



**Вывод:** G-тест корректен с учетом поправок на множественность тестирования.

# Корректность тестов

## FT-тест



**Вывод:** Тест Фишера корректен с учетом поправок на множественность тестирования.

# Мощность тестов

**Данные:** Данные Скринфакс **S** и CardioQvark **C**.

**Цель:** Выяснить, какой критерий является наиболее мощным.

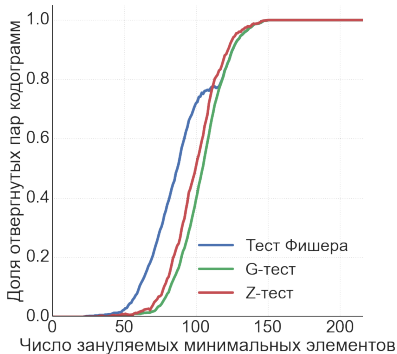
**Эксперимент:** Многократная проверка однородности пары векторов частот триграмм, один из которых посчитан по случайно выбранной кодограмме, а другой — получен из первого путем зануления максимальных/минимальных частот 3-грамм.

**Тесты:**

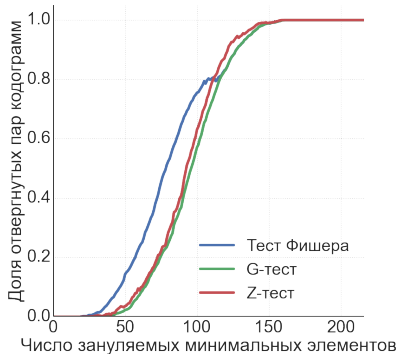
- Тест Фишера;
- G-тест;
- Z-тест.

# Мощность тестов

## Удаление минимальных частот



Данные CardioQvark

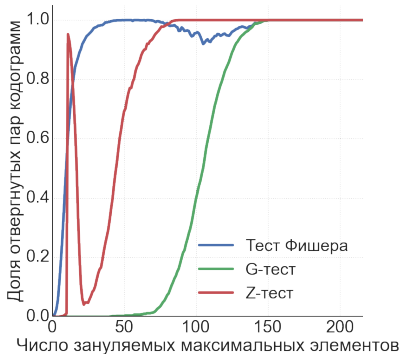


Данные Скринфакс

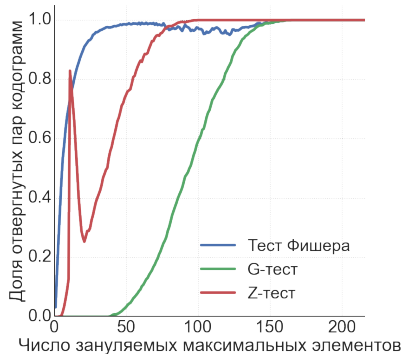
**Вывод:** Тест Фишера является наиболее мощным на данном наборе неоднородных данных.

# Мощность тестов

## Удаление максимальных частот



Данные CardioQvark



Данные Скринфакс

**Вывод:** Тест Фишера является наиболее мощным на данном наборе неоднородных данных.

## Однородность кодограммы в одном обследовании

**Данные:** Данные Скринфакс **S** и CardioQvark **C**.

**Цель:** Проверить однородность кодограмм в пределах одного обследования.

**Эксперимент:** Многократная проверка однородности пары кодограмм, которые являются частями выбранной случайно кодограммы.

**Тесты:**

- Тест Фишера;
- G-тест;
- Z-тест.



## Однородность кодограммы в одном обследовании

Доля **однородных** в пределах одного обследования кодограмм.

поправки

		Pvalue	Holm	BH
S	Тест Фишера	0,744	1,000	0,863
	G-тест	0,973	0,998	0,994
	Z-тест	0,974	—	—
C	Тест Фишера	0,887	1,000	0,977
	G-тест	0,999	0,999	0,999
	Z-тест	0,997	—	—

**Вывод:** кодограммы в пределах одного обследования однородны.

# Однородность кодограмм при повторных обследованиях

**Данные:** Данные Скринфакс **S** и CardioQvark **C**.

**Цель:** Выяснить, есть ли различие в результатах при сравнении кодограмм одного пациента и кодограмм разных пациентов.

**Эксперимент:** Многократная проверка однородности пары кодограмм одного пациента и разных пациентов.

**Тесты:**

- Тест Фишера;
- G-тест;
- Z-тест.

# Однородность кодограмм при повторных обследованиях

Доля **однородных** пар кодограмм.

			поправки		
		Pvalue	Holm	BH	
Кодограммы одного пациента	<b>S</b>	Тест Фишера	0,105	1,000	0,109
		G-тест	0,290	0,400	0,292
		Z-тест	0,255	—	—
	<b>C</b>	Тест Фишера	0,356	1,000	0,388
		G-тест	0,760	0,903	0,830
		Z-тест	0,728	—	—
Кодограммы разных пациентов	<b>S</b>	Тест Фишера	0,004	1,000	0,004
		G-тест	0,047	0,097	0,049
		Z-тест	0,038	—	—
	<b>C</b>	Тест Фишера	0,077	1,000	0,080
		G-тест	0,357	0,630	0,391
		Z-тест	0,343	—	—

**Вывод:** данные разных пациентов более неоднородны, чем данные одного и того же пациента.

# Однородность кодограмм, снятых разными приборами

**Данные:** Синхронизированные данные Скринфакс и CardioQvark **E**.

**Цель:** Выяснить, являются ли данные рассматриваемых приборов однородными.

**Эксперимент:** Проверка однородности синхронизированных пар кодограмм.

- Тесты:**
- Тест Фишера;
  - G-тест;
  - Z-тест.

# Однородность кодограмм, снятых разными приборами

## Статистические тесты

Тест Фишера: отверг 7 пар кодограмм  $\sim 30\%$

G-тест: отверг 1 пару кодограмм  $\sim 4\%$

Z-тест: отверг 0 пар кодограмм  $\sim 0\%$



Показания приборов **однородны**, то есть данные, полученные с двух приборов, можно смешивать при формировании обучающих выборок.

# Однородность обучающей выборки

## Анализ разностей

Для каждой пары синхронизированных ЭКГ **E** (всего пар 23, обозначим через  $M$ ) вычислялась разница между векторами RR-интервалов, амплитуд R-зубцов и частот триграмм.

**Итого:**  $\Delta T^i$ ,  $\Delta R^i$ ,  $\Delta n^i$ , где  $i = 1, \dots, M$

$$\Delta T_{all} = \bigcup_{i=1}^M \Delta T^i, \quad \Delta R_{all} = \bigcup_{i=1}^M \Delta R^i, \quad \Delta n_{all} = \bigcup_{i=1}^M \Delta n^i$$

# Однородность обучающей выборки

## Анализ разностей

**Цель:** Выяснить, являются ли разности частот триграмм, разности RR-интервалов и разности амплитуд R-зубцов однородными.

**Тест:** Критерий Смирнова.

Доля пар кодограмм, для которых вычисленные разности  
не однородны.

	p-value	Holm	BH
$\Delta n$	0.156	0.004	0.036
$\Delta R$	0.996	0.996	0.996
$\Delta T$	0.993	0.989	0.993

**Вывод:** разности частот триграмм однородны.

# Однородность обучающей выборки

## Генерация шума

Генерация шума на вектора частот триграмм:

$$\Delta T_{all} \xrightarrow{\text{bootstrap}} \epsilon = [\epsilon_1, \epsilon_2, \dots, \epsilon_{216}]$$
$$\mathbf{n}_{\text{original}} = [n_1, n_2, \dots, n_{216}]$$
$$\downarrow$$
$$\mathbf{n}_{\text{noise}} = \mathbf{n} + \epsilon = [n_1 + \epsilon_1, n_2 + \epsilon_2, \dots, n_{216} + \epsilon_{216}]$$

Далее запускались алгоритмы классификации (задача классификации здоровый-больной для 45 различных болезней):

SA синдромный алгоритм;

LR\_PCA логистическая регрессия на главных компонентах;

RF случайный лес.



# Однородность обучающей выборки

## Качество классификации

Средние значения **AUC** на контроле при обучении на данных разного типа.

	Обучение на исходных данных			Обучение на зашумленных данных			
	SA	LR_PCA	RF	SA	LR_PCA	RF	
Без шума	0,86234	0,95531	1,00000	0,86195	0,95459	0,98233	
С шумом	0,86189	0,95232	0,99370	0,86200	0,95239	0,97435	
Наибольшая разница	0,00394	0,01192	0,01513	0,00602	0,00942	0,04692	
Средняя разница	0,00045	0,00299	0,00631	0,00005	0,00220	0,00798	

**Вывод:** для данных различных приборов можно применять одни и те же алгоритмы классификации.

## Результаты, выносимые на защиту

- Предложены статистические тесты для проверки однородности символьных последовательностей.
- Показано, что кодограммы в пределах одного обследования (600 кардиоциклов), как правило, однородны. Однако обследования одного и того же пациента могут давать неоднородные кодограммы.
- В экспериментах с синхронной регистрацией ЭКГ двумя приборами показана однородность кодограмм, полученных с помощью систем Скринфакс и CardioQvark.

# Однородность обучающей выборки

## Качество классификации

Средние значения **AUC** на обучении и на контроле на данных одного типа.

	Обучение и контроль на данных одного типа					
	AUC на обучении			AUC на контроле		
	SA	LR_PCA	RF	SA	LR_PCA	RF
Без шума	0,86169	0,95741	1,00000	0,86065	0,93676	0,95423
С шумом	0,86220	0,95553	1,00000	0,86137	0,93508	0,94415
Наибольшая разница	0,04267	0,02311	0,00000	0,04518	0,00878	0,05531
Средняя разница	-0,00051	0,00188	0,00000	-0,00072	0,00168	0,01008

## Синтетические данные

Соединение всех кодограмм множества  $S$ .



Подсчет частот встречаемости всевозможных сочетаний из трех символов алфавита  $\mathcal{A}$ .



Нормировка:  $\sum_{b \in \mathcal{A}} p_{vb} = 1, \quad \forall v \in \mathcal{A} \times \mathcal{A}$ .



Матрица переходных вероятностей (размера  $36 \times 6$ ).

	A	B	C	D	E	F
AA	$p_{AAA}$	$p_{AAB}$	$p_{AAC}$	$p_{AAD}$	$p_{AAE}$	$p_{AAF}$
AB	$p_{ABA}$	$p_{ABB}$	$p_{ABC}$	$p_{ABD}$	$p_{ABE}$	$p_{ABF}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
FF	$p_{FFA}$	$p_{FFB}$	$p_{FFC}$	$p_{FFD}$	$p_{FFE}$	$p_{FFF}$

## Синтетические данные

Данные генерируются согласно следующей процедуре:

- Шаг 1.** Первый и второй символ каждой кодограммы выбирается с одинаковой вероятностью из возможных шести.
- Шаг 2.** В зависимости от двух последних символов кодограммы, согласно вычисленной матрице переходных вероятностей выбирается следующий символ из распределения, заданного соответствующей строкой данной матрицы.
- Шаг 3.** Повторяются действия **Шага 2** до тех пор, пока кодограмма не достигнет нужной длины.

# Поправки на множественность тестирования

## FWER и FDR

Пусть  $H_1, H_2, \dots, H_m$  — семейство проверяемых гипотез.

$M_0$  — множество индексов верных гипотез.

	Верных $H_i$	Неверных $H_i$	Всего
Принятых $H_i$	$U$	$T$	$m - R$
Отвергнутых $H_i$	$V$	$S$	$R$
Всего	$m_0$	$m - m_0$	$m$

# Поправки на множественность тестирования

## FWER и FDR

### FWER

Групповой вероятностью ошибки первого рода (familywise error rate) называется величина:

$$\text{FWER} = P(V > 0).$$

### FDR

Ожидаемой долей ложных отклонений гипотез (false discovery rate) называется величина:

$$\text{FDR} = \mathbb{E} \left( \frac{V}{\max(R, 1)} \right).$$

# Поправки на множественность тестирования

## Метод Холма (Holm)

Пусть  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$  — достигаемые уровни значимости, упорядоченные по возрастанию.

**Метод Холма** — нисходящая процедура со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{\alpha}{m-1}, \dots, \alpha_i = \frac{\alpha}{m-i+1}, \dots, \alpha_m = \alpha,$$

Метод контролирует FWER на уровне значимости  $\alpha$ .



# Поправки на множественность тестирования

## Метод Бенджамини-Хохберга (BH)

**Метод Бенджамини-Хохберга** — восходящая процедура со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m}, \alpha_2 = \frac{2\alpha}{m}, \dots, \alpha_i = \frac{i\alpha}{m}, \dots, \alpha_m = \alpha,$$

Метод контролирует FDR на уровне значимости  $\alpha$ .