

# 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

15 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

В задачах, возникающих в информационном анализе электрокардиосигналов при исследовании воспроизводимости ЭКГ-сигналов и метрологической проверкой пригодности приборов.

## Z-тест (Z)

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

$$\begin{array}{l} S_1 : [n_{11}, n_{12}, \dots, n_{1K}] \\ S_2 : [n_{21}, n_{22}, \dots, n_{2K}] \end{array} \xrightarrow{\forall m=1, \dots, K} \begin{array}{l} H_0 : p_{1m} = p_{2m}; \\ H_1 : p_{1m} \neq p_{2m}. \end{array}$$

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 кодограммы.

Рассматриваются вектора:

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

		J				$n_{\bullet+}$		
		1	2	...	K			
T	Таблица сопряженности	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 **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  $J$ :  $G^2(X_S, X_w) \geq \chi_{1-\frac{\alpha}{2}}^2 \Rightarrow H_0$  отвергается,  $I$  и  $J$  зависимы.

# Тест Фишера (FT)

## Независимость

Для таблицы сопряженности  $\mathbf{T}$ , вычисляется значение  $P$ :

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

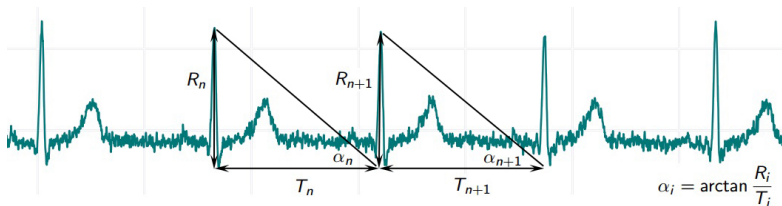
$\{\mathbf{T}^r\}_{r=1}^N$  — таблицы сопряженности с суммой по строкам  $n_{i+}$  и столбцам  $n_{+j}$ .

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

**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

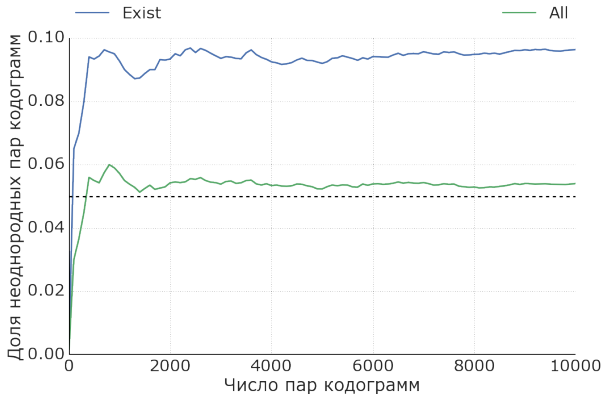
**Данные:** Синтетические данные **M**.

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

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

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

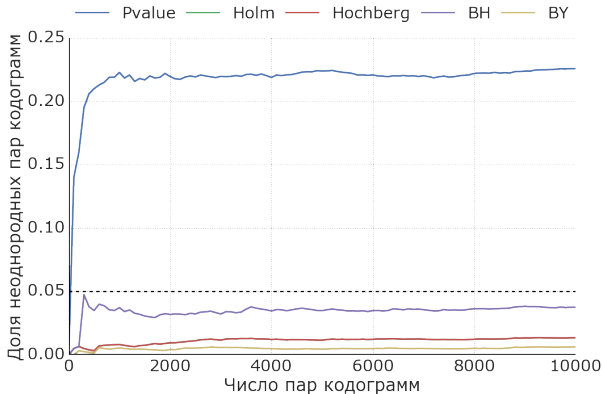
# Корректность Z-тест



**Вывод:** Z-тест корректен при использовании доли отвергнутых триграмм среди всех 216 триграмм.

# Корректность

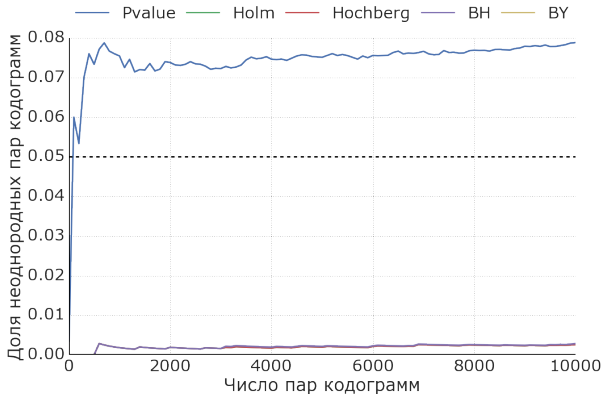
## G-тест



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

# Корректность

## FT-тест



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

# Мощность

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

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

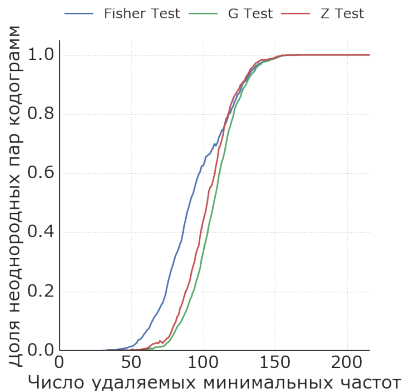
**Эксперимент:** Многократная проверка однородности пары кодограмм, одна из которых выбрана случайно, другая — получена из первой путем зануления максимальных/минимальных частот 3-грамм.

**Тесты:**

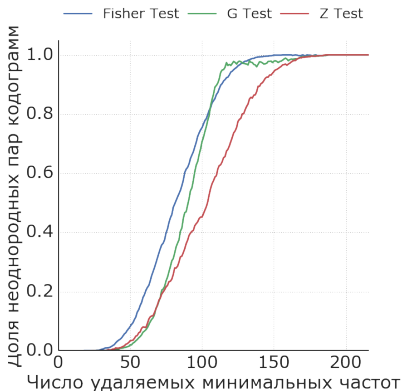
- Тест Фишера;
- G-тест;
- Z-тест.

# Мощность

## Удаление минимальных частот



Данные CardioQvark

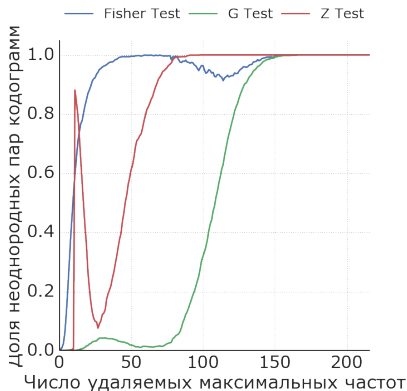


Данные Скринфакс

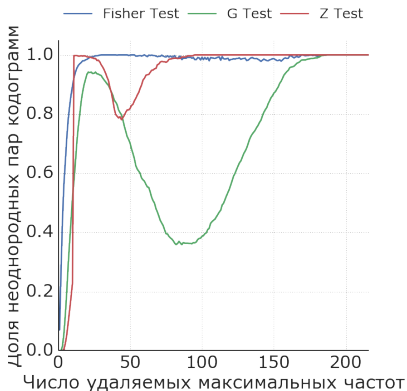
**Вывод:** Тест Фишера является наиболее мощным на данном наборе неоднородных данных.

# Мощность

## Удаление максимальных частот



Данные CardioQvark



Данные Скринфакс

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

# Однородность

## Кодограмм

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

**Цель:** Выяснить, являются ли кодограммы однородными по отдельности.

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

**Тесты:**

- Тест Фишера;
- G-тест;
- Z-тест.



# Однородность

## Кодограмм

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

поправки на множественность

	P-value	Holm	Hochberg	BH	BY	
<b>S</b>	<b>FT</b>	0.736	1.000	1.000	0.844	1.000
	<b>G</b>	0.965	0.993	0.993	0.988	0.993
	<b>Z</b>	0.968	1.000	1.000	1.000	1.000
<b>C</b>	<b>FT</b>	0.856	1.000	1.000	0.976	1.000
	<b>G</b>	0.999	0.999	0.999	0.999	1.000
	<b>Z</b>	0.994	1.000	1.000	1.000	1.000

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

# Однородность

## Кодограмм одного пациента

**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 одного пациента

Доля **однородных** пар кодограмм.

		поправки на множественность					
		P-value	Holm	Hochberg	BH	BY	
Один пациент	S	FT	0.069	1.000	1.000	0.069	0.109
		G	0.168	0.251	0.251	0.171	0.205
		Z	0.163	1.000	1.000	1.000	1.000
	C	FT	0.343	1.000	1.000	0.382	0.497
		G	0.694	0.836	0.836	0.731	0.785
		Z	0.584	1.000	1.000	1.000	1.000
Два пациента	S	FT	0.001	1.000	0.001	0.001	0.001
		G	0.008	0.013	0.013	0.008	0.011
		Z	0.007	1.000	1.000	1.000	1.000
	C	FT	0.149	1.000	1.000	0.157	0.240
		G	0.549	0.693	0.693	0.575	0.633
		Z	0.185	1.000	1.000	1.000	1.000

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

## Сравнение показаний приборов

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

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

**Эксперимент:** Проверка однородности пар кодограмм, полученных из синхронизированных пар ЭКГ.

**Тесты:**

- Тест Фишера;
- G-тест;
- Z-тест.

# Сравнение показаний приборов

## Статистические тесты

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

G-тест: отверг 2 пары кодограмм  $\sim 8\%$

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



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

# Сравнение показаний приборов

## 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  $i = 1, \dots, M$

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

# Сравнение показаний приборов

## Анализ разностей

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

**Тест:** Критерий Колмогорова-Смирнова.

Доля пар кодограмм, для которых вычисленные разности **не однородны**.

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

**Вывод:** можно считать, что разности частот триграмм однородны.

# Сравнение показаний приборов

## Анализ разностей

Генерация шума на вектора частот триграмм:

$$\Delta T_{all} \xrightarrow{\text{bootstrap}} \varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{216}]$$

Для каждого  $n_{\text{original}} = [n_1, n_2, \dots, n_{216}]$  вычислялся вектор частот  $n_{\text{noise}} = n + \varepsilon = [n_1 + \varepsilon_1, n_2 + \varepsilon_2, \dots, n_{216} + \varepsilon_{216}]$

Далее запускались алгоритмы классификации как на оригинальных данных, так и на зашумленных.



# Сравнение показаний приборов

## Анализ разностей

### Средние значения AUC

на обучении (auc\_l) и на контроле (auc\_c).

	auc_l	auc_c
среднее значение показателей качества при наложении шума	0.939	0.914
среднее значение показателей качества на исходных данных	0.940	0.917
средняя разность показателей (с шумом - без шума)	-0.000	-0.004
максимальная разность показателей (с шумом - без шума)	0.043	0.045
минимальная разность показателей (с шумом - без шума)	-0.023	-0.055

**Вывод:** наложение шума не ухудшает качество классификации. 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 с 30-минутным интервалом могут давать неоднородные кодограммы у некоторых обследуемых.
- В экспериментах с параллельной регистрацией ЭКГ двумя приборами показана однородность кодограмм, полученных с помощью систем Скринфакс и CardioQvark.