Исследование зависимости/независимости двух номинальных признаков

1. Имеются сведения (Macmillan Publishing Company) о влиянии прививки на холерную инфекцию. Среди 1630 человек, привитых от инфекции, заболело 5 человек; среди 1033 непривитых заболело 11 человек. Имеется ли зависимость между наличием прививки (признак A) и заболеваемостью холерой (признак B)? Если зависимость установлена, опишите её, используя коэффициенты контингенции и ассоциации.

2. Имеются следующие данные о специализации и поле 900 английских студентов

Специализация	Муж.	Жен.
Искусствоведение	165	185
Естественные науки	168	92
Социально-экономические науки	115	105
Музыка	32	38

Выясните, имеется ли зависимость между специализацией и полом студента. Оцените силу связи этих признаков.

3. В 2009г. центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Среди ста вопросов анкеты были, в частности, такие: 1)какое из шести перечисленных описаний точнее всего соответствует материальному положению вашей семьи; 2)удовлетворены ли вы своим здоровьем.
На первый вопрос предлагались ответы:

-- денег не хватает даже на питание (категория A_1);

-- на питание денег хватает, но не хватает на покупку одежды и обуви (категория A_2);

-- на покупку одежды и обуви денег хватает, но не хватает на покупку бытовой техники (категория A_3);

-- денег вполне хватает на покупку крупной бытовой техники, но не можем купить новый автомобиль (категория A_4);

-- денег хватает на все, кроме таких дорогих приобретений, как квартира, дом (категория A_5);

-- материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом (категория A_6).

Ответы на второй вопрос: удовлетворен (категория B_1) и не удовлетворен (категория B_2).

Результаты опроса представлены в таблице сопряженности признаков A (материальное положение семьи) и B (удовлетворенность состоянием своего здоровья).

	B_1	B_2
A_1	83	154
A_2	278	354
A_3	470	299
A_4	204	76
A_5	46	20
A_6	13	3

Оцените меры прогноза Гутмана, дайте трактовку полученных результатов.

Домашнее задание

1. Задача №11 стр. 203 (учебник Кибзун А.И., Горяинова Е.Р., Наумов А.В.)
2. Задача №9 стр. 202 (учебник Кибзун А.И., Горяинова Е.Р., Наумов А.В.) + вычислить для рассматриваемых показателей коэффициент Пирсона и коэффициент Крамера
3. Изучается взаимосвязь между зоркостью правого и левого глаза. Зоркость каждого глаза соответствует одной из четырёх категорий – высшая, вторая, третья и низшая. По результатам обследования 3242 мужчин в возрасте 30-39 лет получены следующие данные о зоркости.

Правый/Левый	высшая	вторая	третья	низшая
высшая	821	112	85	35
вторая	116	494	145	27
третья	72	151	583	87
низшая	43	34	106	333

Оцените меры прогноза Гутмана. Прокомментируйте полученный результат.

18.04

$H_0: F_W(x,y) = F_X(x) \cdot F_Y(y)$ – независимые

1. Имеются сведения (Macmillan Publishing Company) о влиянии прививки на холерную инфекцию. Среди 1630 человек, привитых от инфекции, заболело 5 человек; среди 1033 непривитых заболело 11 человек. Имеется ли зависимость между наличием прививки (признак A) и заболеваемостью холерой (признак B)? Если зависимость установлена, опишите её, используя коэффициенты контингенции и ассоциации.

A \ B	Заболел	Не заболел	
Привит	5	1625	$1630 = n_{1\cdot}$
Не привит	11	1022	$1033 = n_{2\cdot}$
	16 $= n_{\cdot 1}$	2647 $= n_{\cdot 2}$	2663

сумма либо по вертикали, либо по горизонтали

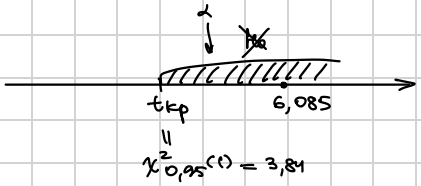
$H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ → независим

$H_A: \exists (i,j): p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}$

$$\chi^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij})^2}{n_{i\cdot}\cdot n_{\cdot j}} - 1\right) \sim \chi^2((k-1)\cdot(m-1))$$

Для случая $m=2, k=2$: $\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} \sim \chi^2(1)$

$\chi^2_{0,95}(1) = 3,84$



$$\chi^2 = \frac{2663(5\cdot 1022 - 1625\cdot 11)}{1630\cdot 1033\cdot 16\cdot 2677} = 6,085$$

Арсений Семкин решал, там на 10 было нрзм

Коэф. контингенции

A \ B	B	$\bar{B}$	
A	a	b	a+b
$\bar{A}$	c	d	c+d
	a+c	b+d	n

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \in [-1;1]$$ двусторонняя

$|\Phi| \in [0;0,3]$ – слабая связь

Таблица Чеддока → $|\Phi| \in [0,3;0,5]$ – умеренная

Шкала Чеддока

Таблица 3. Шкала Чеддока

Коэффициент корреляции $\lvert r\rvert$	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99	1,0
Характеристика связи	Слабая	Умеренная	Заметная	Тесная	Очень тесная	Функциональная

потому что число отрицат. Коэф. ассоциации Юля (односторон. связь)

$-0,555 = Q = \frac{ad - bc}{ad + bc}$

$A \Rightarrow \bar{B}$, $\bar{B} \not\Rightarrow A$

A отсутствие ⇒ B отсутств.

но B отсутствие ⇏ A отсутствие

$\chi^2 = 20,38$

2. Имеются следующие данные о специализации и поле 900 английских студентов

Специализация	Муж.	Жен.	
Искусствоведение	165	185	350
Естественные науки	168	92	260
Социально-экономические науки	115	105	220
Музыка	32	38	70
	480	480	900

Критерий Пирсона

$H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$ – независимы

$H_A: \exists (i,j): p_{ij} \neq p_{i\cdot} \cdot p_{\cdot j}$

$$\chi^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij})^2}{n_{i\cdot}\cdot n_{\cdot j}} - 1\right) \sim \chi^2((k-1)(m-1))$$

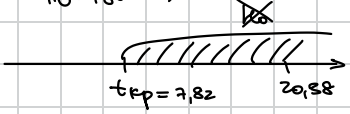
семя очень хорошо рассказал

10 минимум

Выясните, имеется ли зависимость между специализацией и полом студента. Оцените силу связи этих признаков.

$$= \left(\frac{165^2}{350\cdot 480} + \frac{185^2}{350\cdot 420} + \frac{168^2}{260\cdot 480} + \frac{92^2}{260\cdot 420} \cdots + \frac{38^2}{70\cdot 480} - 1\right)\cdot 900 = 20,38 = t_н$$

$t_{кр} = \chi^2_{0,95}(3\cdot 1) = \chi^2_{0,95}(3) = 7,82$



$$P = \sqrt{\frac{\hat{\chi}^2}{\hat{\chi}^2 + n}} = \sqrt{\frac{20,38}{20,38 + 900}} = 0,149 \in [0;1]$$ зависимы

Коэф. Крамера

$$C = \sqrt{\frac{\hat{\chi}^2}{n\cdot \min\{(m-1),(k-1)\}}} = \sqrt{\frac{20,38}{900\cdot 1}} \approx 0,15$$

если $P, C \in [0;0,3)$ – слабая

$P, C \in [0,3;0,7)$ – умерен.

$p_1 c \in [0{,}7\ ;\ 1]$ — значительная

3. В 2009г. центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Среди ста вопросов анкеты были, в частности, такие: 1)какое из шести перечисленных описаний точнее всего соответствует материальному положению вашей семьи; 2)удовлетворены ли вы своим здоровьем.
На первый вопрос предлагались ответы:

-- денег не хватает даже на питание (категория A_1);

-- на питание денег хватает, но не хватает на покупку одежды и обуви (категория A_2);

-- на покупку одежды и обуви денег хватает, но не хватает на покупку бытовой техники (категория A_3);

-- денег вполне хватает на покупку крупной бытовой техники, но не можем купить новый автомобиль (категория A_4);

-- денег хватает на все, кроме таких дорогих приобретений, как квартира, дом (категория A_5);

-- материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом (категория A_6).

Ответы на второй вопрос: удовлетворен (категория B_1) и не удовлетворен (категория B_2).

Результаты опроса представлены в таблице сопряженности признаков A (материальное положение семьи) и B (удовлетворенность состоянием своего здоровья).

	B_1	B_2	
A_1	83	154	237
A_2	278	354	632
A_3	470	299	769 ←
A_4	204	76	280
A_5	46	20	66
A_6	13	3	16
	1094	906	

Оцените меры прогноза Гутмана, дайте трактовку полученных результатов.

Первый прогноз — наивный прогноз по модальному значению

Вероятность ошибки: $\hat{p}_1 = 1 - \frac{769}{2000} = 0{,}6155$

Воспользуемся знанием о распределении людей в категории B:

$\hat{p}_2 = 1 - \frac{470+354}{2000} = 0{,}588$

$\lambda_A = \frac{\text{Вер-ть ошибки 1 прогн} - \text{Вер-ть ошибки 2 прогноза}}{\text{Вер ошибки 1 прогноза}} = \frac{0{,}6155 - 0{,}588}{0{,}6155} = 0{,}045$

↓ мера прогноза

Если учитывать состояние здоровье, то прогноз улучшается на 4%.

Гутман

$\hat{p}_1 = 1 - \frac{1094}{2000} = 0{,}453$

$\hat{p}_2 = 1 - \frac{154+354+470+204+46+13}{2000} = 0{,}3795$

$\lambda_B = \frac{0{,}453 - 0{,}3795}{0{,}453} = 0{,}1622$ улучшился на 16,22%

$\lambda = \frac{\lambda_A + \lambda_B}{2} = 0{,}1$