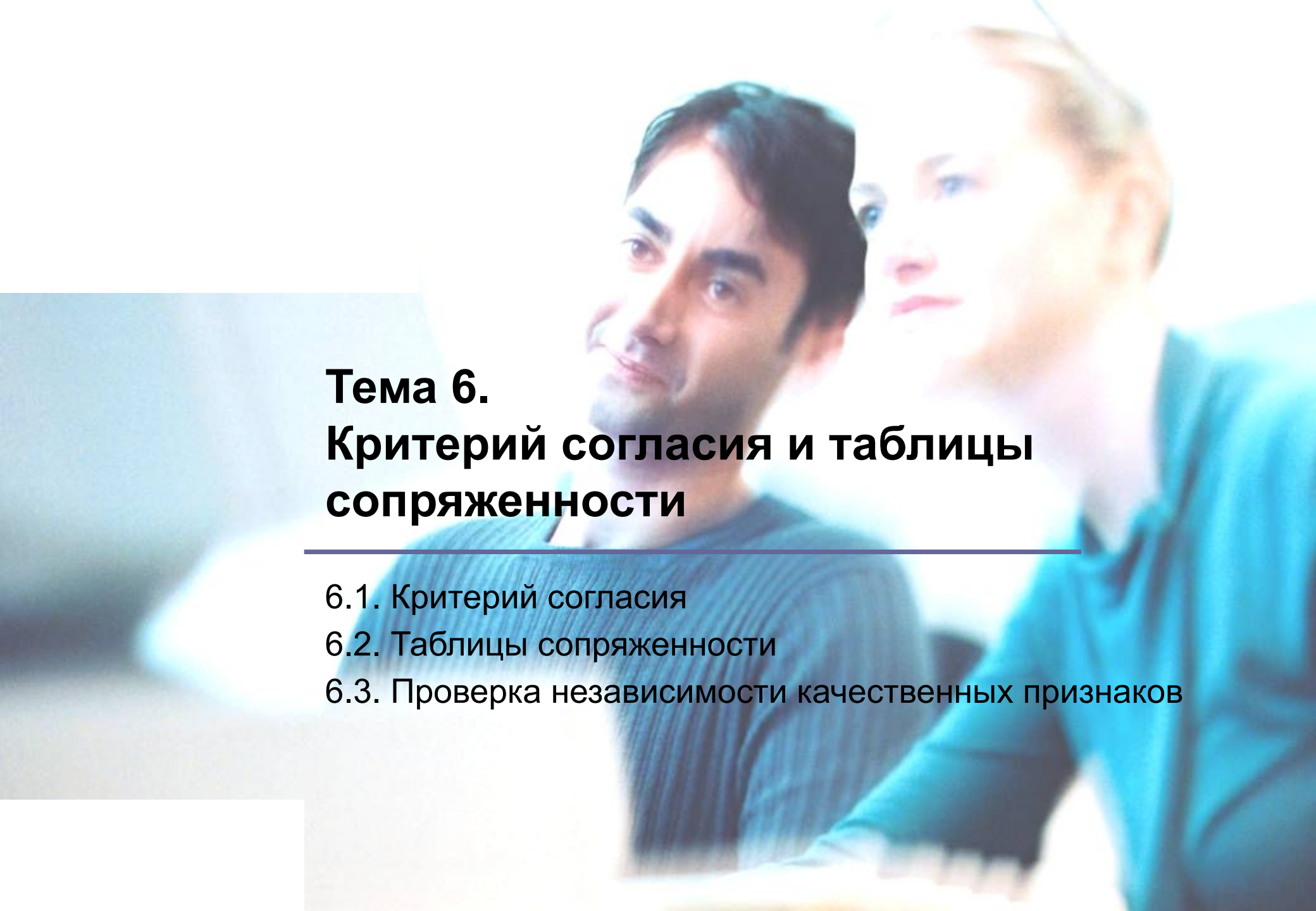
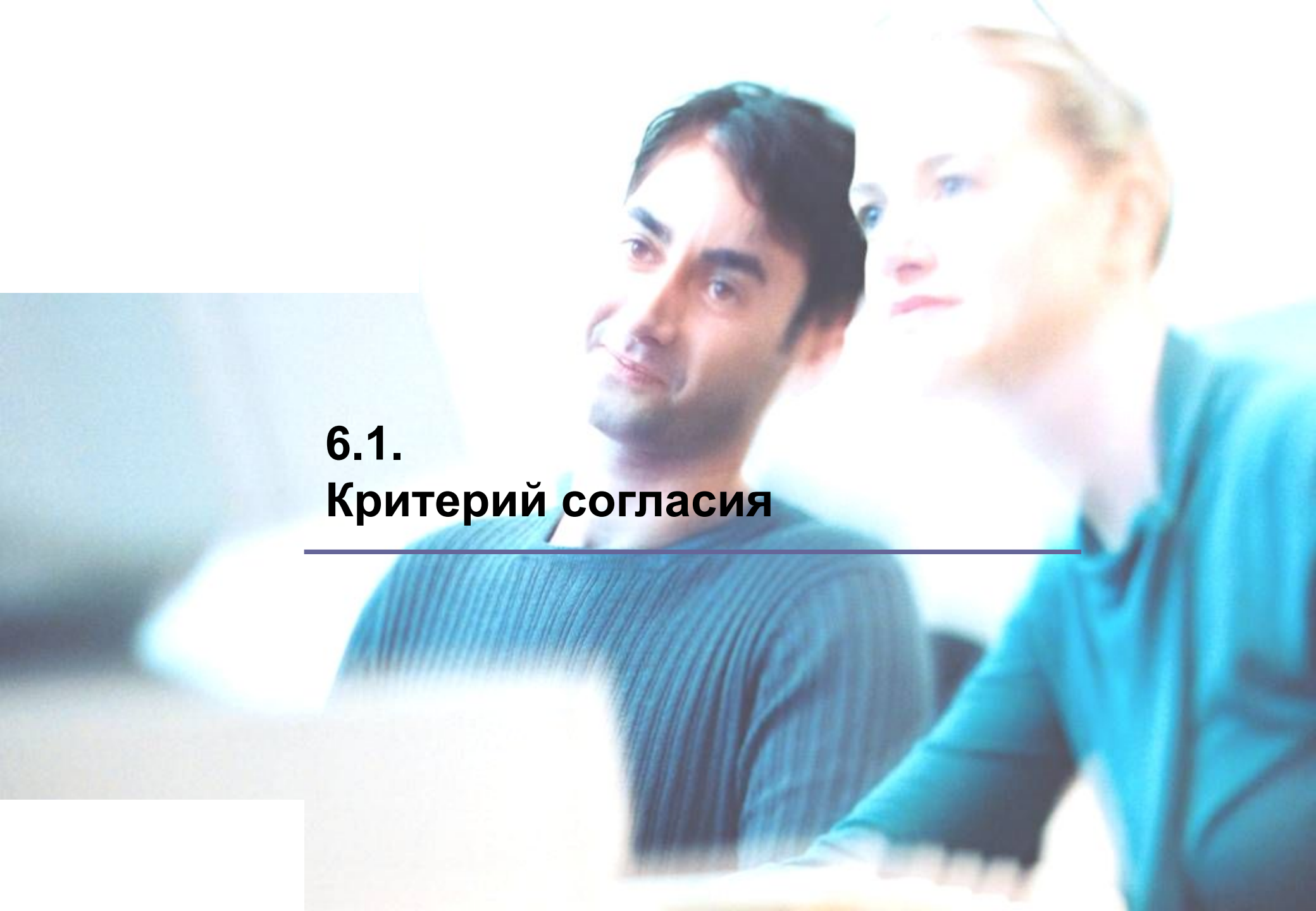


Тема 6.
Критерий согласия и таблицы сопряженности

6.1. Критерий согласия

6.2. Таблицы сопряженности

6.3. Проверка независимости качественных признаков



6.1. Критерий согласия

Пример. Вкусовые предпочтения

Маркетолог хочет узнать, какому из пяти вкусов нового напитка отдают предпочтение покупатели. Ниже приведены данные, полученные из опроса 100 человек:

Вишня	Клубника	Апельсин	Лайм	Виноград
32	28	16	14	10

Если нет каких-либо особых вкусовых предпочтений, то каждый вид напитка покупают с одинаковой частотой. В таком случае каждая частота должна быть равна 100/5 = 20, то есть *приблизительно* по 20 человек выберут каждый вид сока.

Вишня	Клубника	Апельсин	Лайм	Виноград	
32	28	16	14	10	Наблюдаем
20	20	20	20	20	Ожидаем

Наблюдаемые и ожидаемые частоты

Наблюдаемые частоты - частоты полученные по выборке.

Ожидаемые частоты - частоты, полученные путем вычисления на основе теоретических представлений о предполагаемом распределении.

Вишня	Клубника	Апельсин	Лайм	Виноград
32	28	16	14	10
20	20	20	20	20

Наблюдаемые частоты

Ожидаемые частоты

Что проверяет критерий согласия

Критерий согласия позволяет выяснить, насколько согласуются между собой наблюдаемые частоты и ожидаемые, иными словами, существенны или нет различия между ними.

Гипотезы для примера с предпочтениями запишутся так:

H_0: У покупателей нет предпочтений по поводу вкусов сока.

H_1: У покупателей есть предпочтения.

Необходимые условия

1. Выборка случайна.
2. Наблюдаемая частота должна быть не меньше 5.

Статистика

Для проверки гипотезы используется статистика :

$$X = \sum \frac{(H - O)^2}{O}$$

H – наблюдаемая частота

O – ожидаемая частота

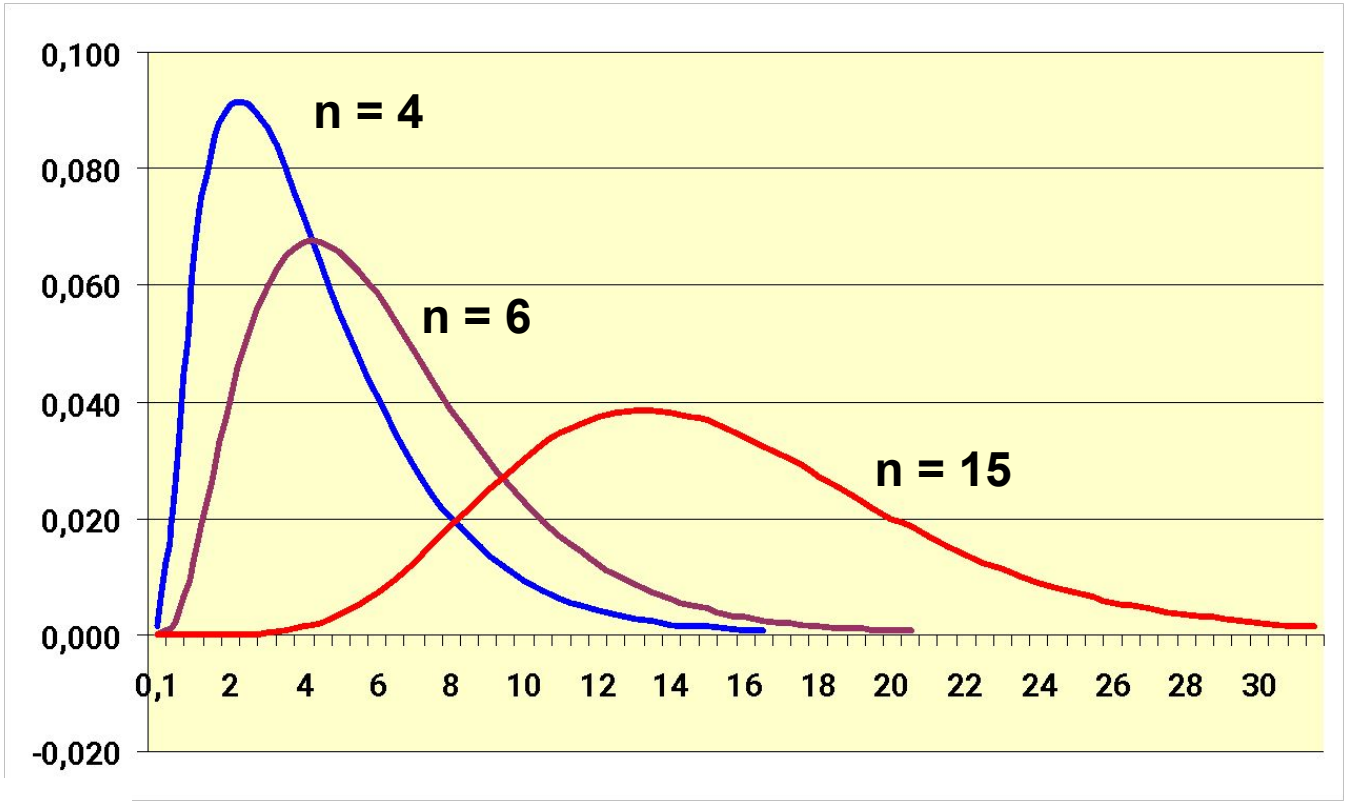
Если значение X велико, гипотезу H_0 следует отвергнуть (расхождения между наблюдаемыми и ожидаемыми частотами значительны)

Для уточнения понятия «велико надо» знать распределение X.

В условиях нулевой гипотезы статистика имеет χ^2-распределение с числом степеней свободы df = *n* – 1 (где n – число слагаемых в сумме)

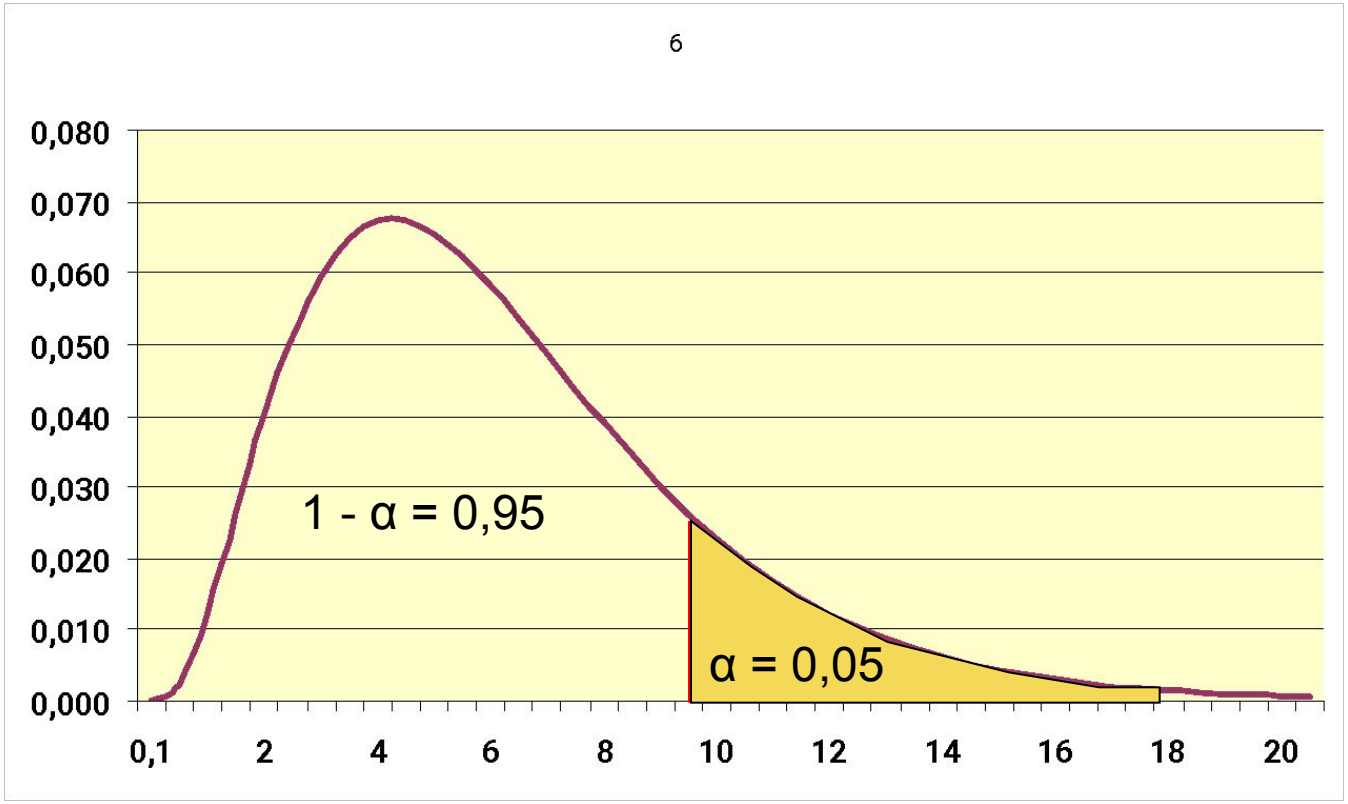
Вид χ^2 распределения

В зависимости от числа степеней свободы n вид распределения изменяется. При увеличении n распределение приближается к нормальному.



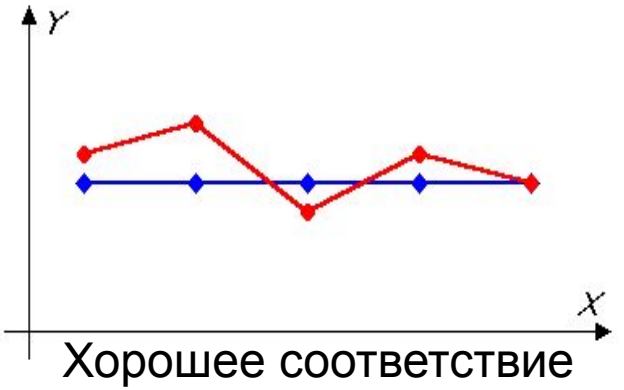
Критическая область

Этот критерий имеет только правостороннюю критическую область. Критическая область соответствует значениям статистики, для которых значение велико. Это означает, что данные плохо согласуются.

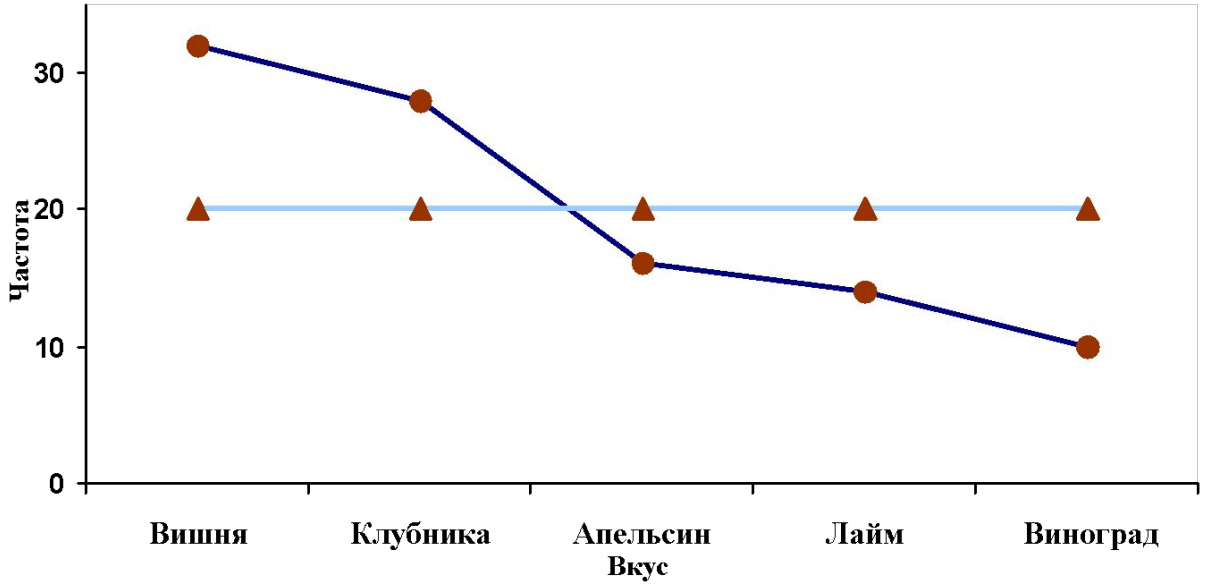


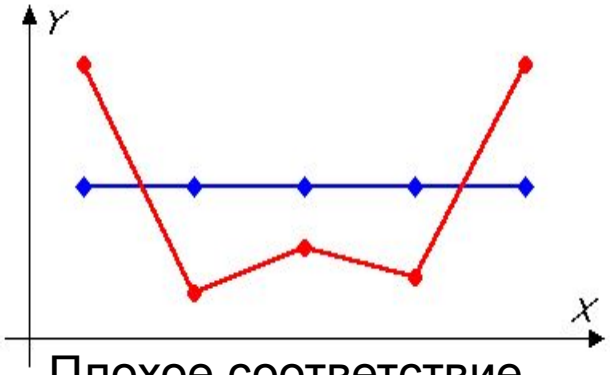
Что значит «частоты согласуются»

Если наблюдаемые и ожидаемые значения близки друг к другу, значение X будет небольшим. Гипотеза H_0 не будет отвергнута. Имеется хорошее соответствие наблюдаемых данных и исследовательской модели.



Хорошее соответствие





Плохое соответствие

Решение задачи

Шаг 1. Нулевая и альтернативная гипотезы:

H_0: У покупателей нет предпочтений по поводу вкусов сока.

H_1: У покупателей есть предпочтения.

Шаг 2. Уровень значимости α=0,05.

Шаг 3. Критическое значение равно 9,488 (по таблице χ^2-распределения или с помощью функции Excel,df = 5 – 1 = 4 и α = 0,05).

=ХИ2ОБР(0,05;4)

Шаг 4. По выборке находим значение статистики:

$$\chi^2 = \sum \frac{(H-O)^2}{O} = \frac{(32-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(10-20)^2}{20} = 18$$

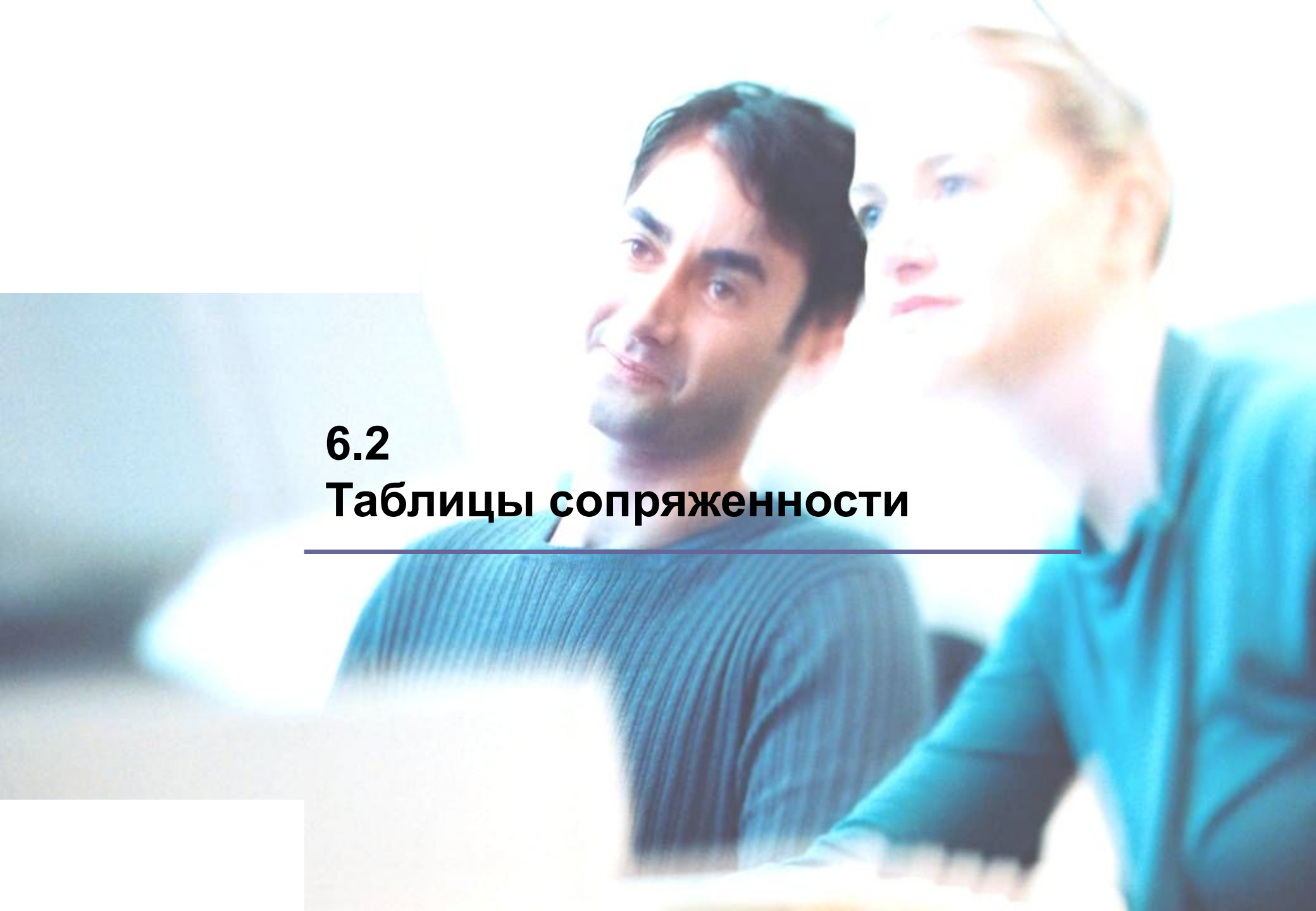
Шаг 5. Сравним полученное значение с критической областью: 18 > 9,488. Значение попало в критическую область.

Шаг 6. Формулируем ответ. **Существуют значимые предпочтения покупателей по поводу вида напитка.**

Применение критерия согласия

1. Для проверки гипотезы о согласовании наблюдаемого распределения и теоретического. Это было в примере с напитками. Наиболее часто проверяют согласование наблюдаемого распределения с нормальным, т.к. многие критерии предполагают нормальность распределения.

2. Для проверки гипотезы о совпадении законов распределения двух генеральных совокупностей. Предположение о виде теоретического распределения (теоретическая модель данных) в этом случае не требуется. Критерий дает нам представление о «расстоянии между двумя наборами данных» и на основе значения этого расстояния позволяет делать вывод о «согласии» между двумя распределениями.



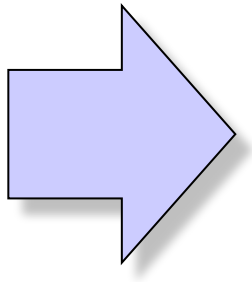
6.2
Таблицы сопряженности

Обработка данных

Данные эксперимента

Номер респондента	Признак 1 Пол?	Признак 2 Курит?
1	Мужчина	Курит
2	Женщина	Не курит
3	Женщина	Курит
4	Мужчина	Курит
5	Мужчина	Не курит
6	Женщина	Не курит
7	Мужчина	Не курит
8	Мужчина	Курит
9	Женщина	Не курит
10	Женщина	Не курит

Таблица сопряженности



	Курит	Не курит
Мужчина	3	2
Женщина	1	4

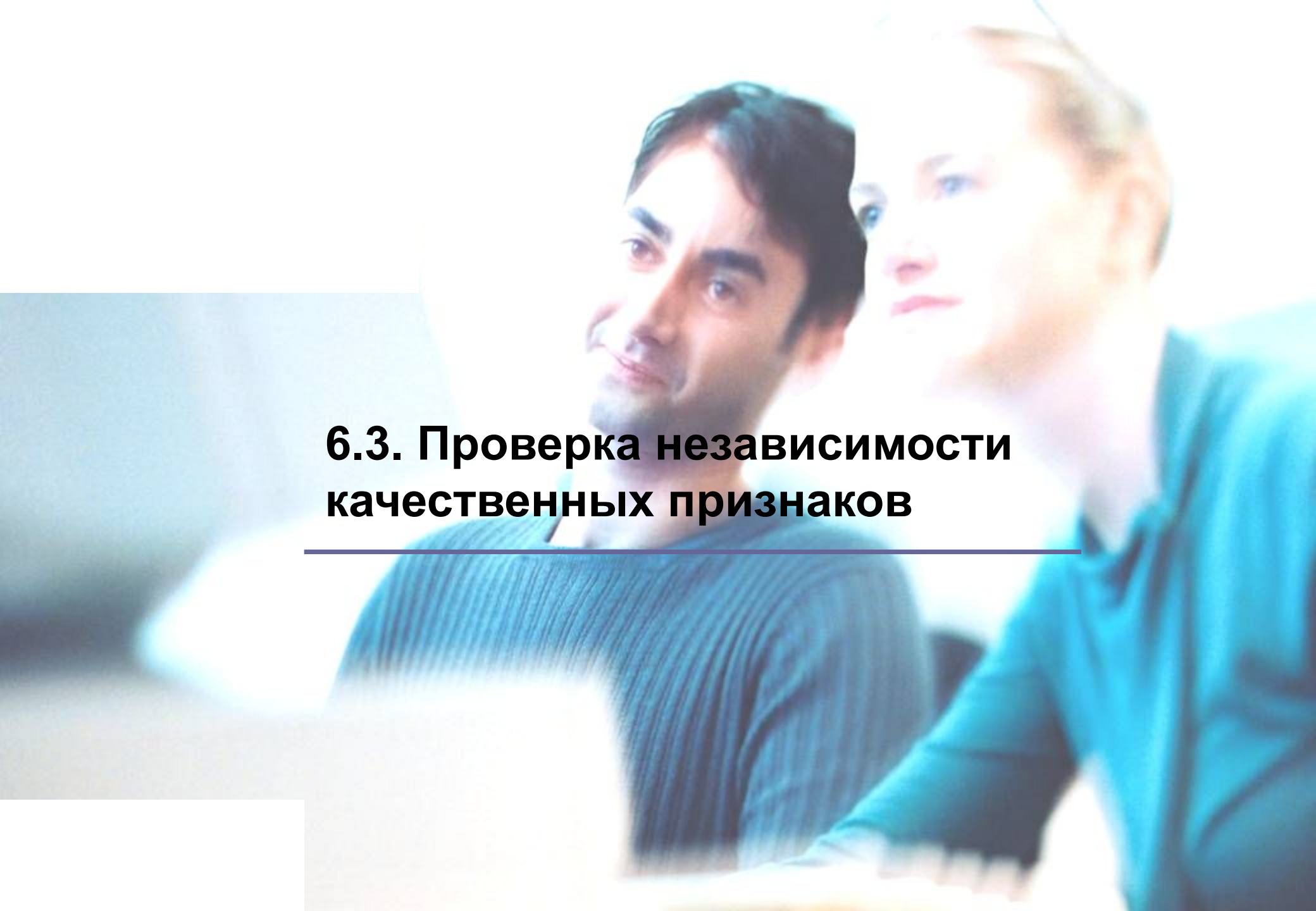
Таблица сопряженности составляется для двух признаков и содержит частоты для каждого набора значений.

В общем виде таблица сопряженности состоит из *r* рядов и *c* столбцов.

Каждая клетка таблицы определяется номером ее ряда (Row) и столбца (Column).

Признак 1.	**Признак 2. Отношение к новому препарату**		
Категория персонала	Согласны	Не согласны	Воздержались
Медсестры	F_{11}	F_{12}	F_{13}
Врачи	F_{21}	F_{22}	F_{23}

Данная таблица имеет два ряда и три столбца: $r = 2$, $c = 3$.



6.3. Проверка независимости качественных признаков

Наблюдаемые частоты (Observed frequencies)

В результате эксперимента мы получаем наблюдаемые частоты. Подсчитаем суммы по срокам и столбцам.

	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры	100	80	20	200
Врачи	50	120	30	200
ВСЕГО	150	200	50	400

Шаг 1. Гипотезы

Критерий согласия используется для проверки гипотезы о независимости качественных признаков.

Гипотезы выглядят так:

H_0 : признаки независимы.

H_1 : признаки зависимы.

Ожидаемые частоты (Expected frequencies)

Вычислим теоретические ожидаемые частоты (в предположении независимости признаков).

A – случайно выбранный медработник – медсестра

B – случайно выбранный медработник согласен с эффективностью препарата

$A \cap B$ -случайно выбранный медработник – медсестра, согласная с эффективностью препарата

Если события A и B независимы, то

$$P(A \cap B) = P(A) \cdot P(B)$$

A – случайно выбранный медработник – медсестра
B – случайно выбранный медработник согласен с эффективностью препарата

$$P(A) = \frac{200}{400} \qquad P(B) = \frac{150}{400} \qquad P(A \cap B) = \frac{200}{400} \cdot \frac{150}{400} = \frac{3}{16}$$

	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры				**200**
Врачи				200
ВСЕГО	**150**	200	50	**400**

A – случайно выбранный медработник – медсестра
B – случайно выбранный медработник согласен с эффективностью препарата

$$P(A)=\frac{200}{400} \qquad P(B)=\frac{150}{400} \qquad P(A\cap B)=\frac{200}{400}\cdot\frac{150}{400}=\frac{3}{16}$$

На 400 человек ожидаемая частота медсестер согласных с эффективностью препарата

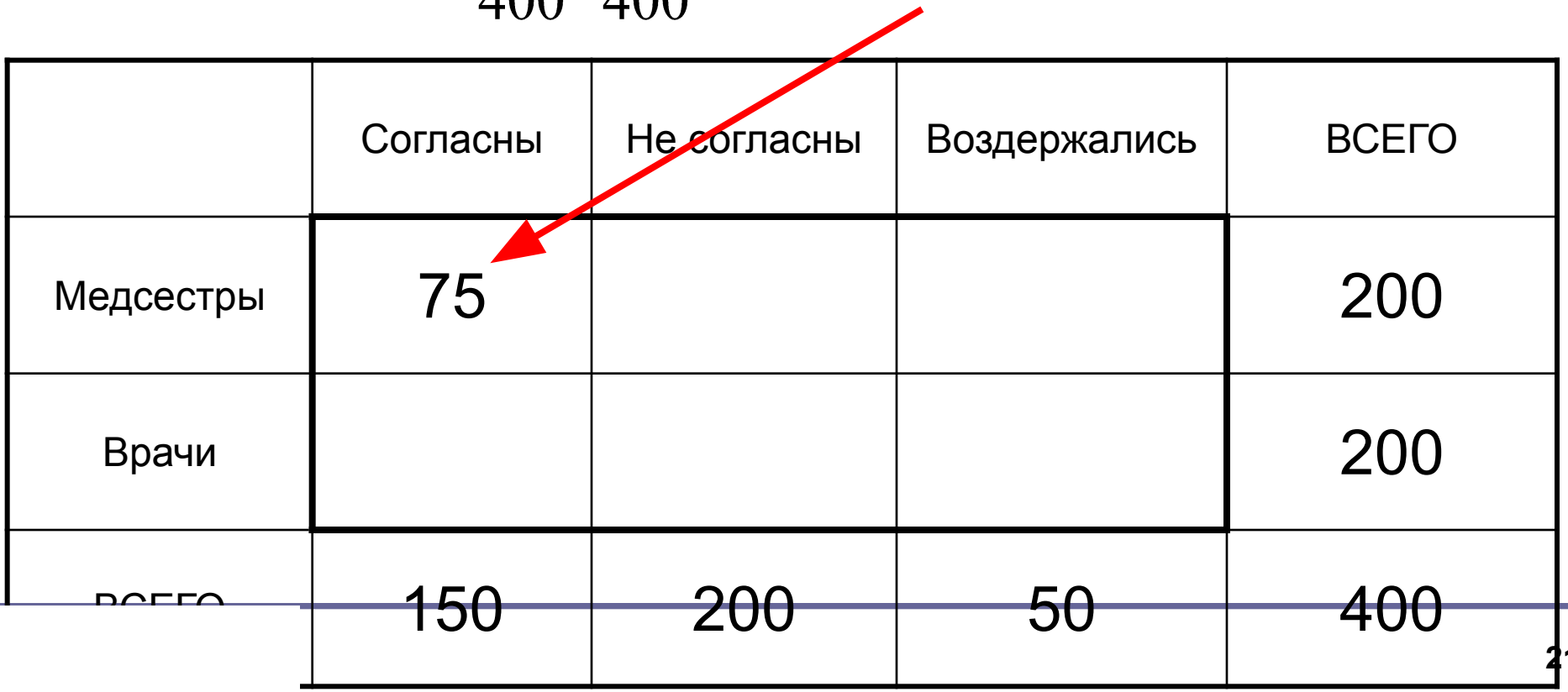
$$\frac{3}{16}\cdot 400=75$$

	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры				**200**
Врачи				200
ВСЕГО	**150**	200	50	**400**

Ожидаемые частоты (Expected frequencies)

Вычислим теоретические частоты (в предположении независимости признаков). В первую клетку надо поставить частоту:

$$\frac{200}{400} \cdot \frac{150}{400} \cdot 400 = 75$$



	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры	75			200
Врачи				200
ВСЕГО	150	200	50	400

Ожидаемые частоты (Expected frequencies)

Вычислим теоретические частоты.

$$\frac{200}{400} \cdot \frac{50}{400} \cdot 400 = 25$$

	Согласны	Не согласны	Воздержались	ВСЕГО
Медсестры	75	100	25	200
Врачи	75	100	25	200
ВСЕГО	150	200	50	400

Критерий проверки гипотезы

Наблюдаемые частоты				Ожидаемые частоты		
100	80	20		75	100	25
50	120	30		75	100	25

Если бы признаки были независимыми, то частоты должны быть распределены так, как показано в таблице ожидаемых частот. **Критерий согласия** позволяет оценить, насколько сильно различаются наблюдаемые частоты от ожидаемых. Если сильно, тогда мы признаем наличие зависимости признаков.

$$X = \sum \frac{(H - O)^2}{O}$$

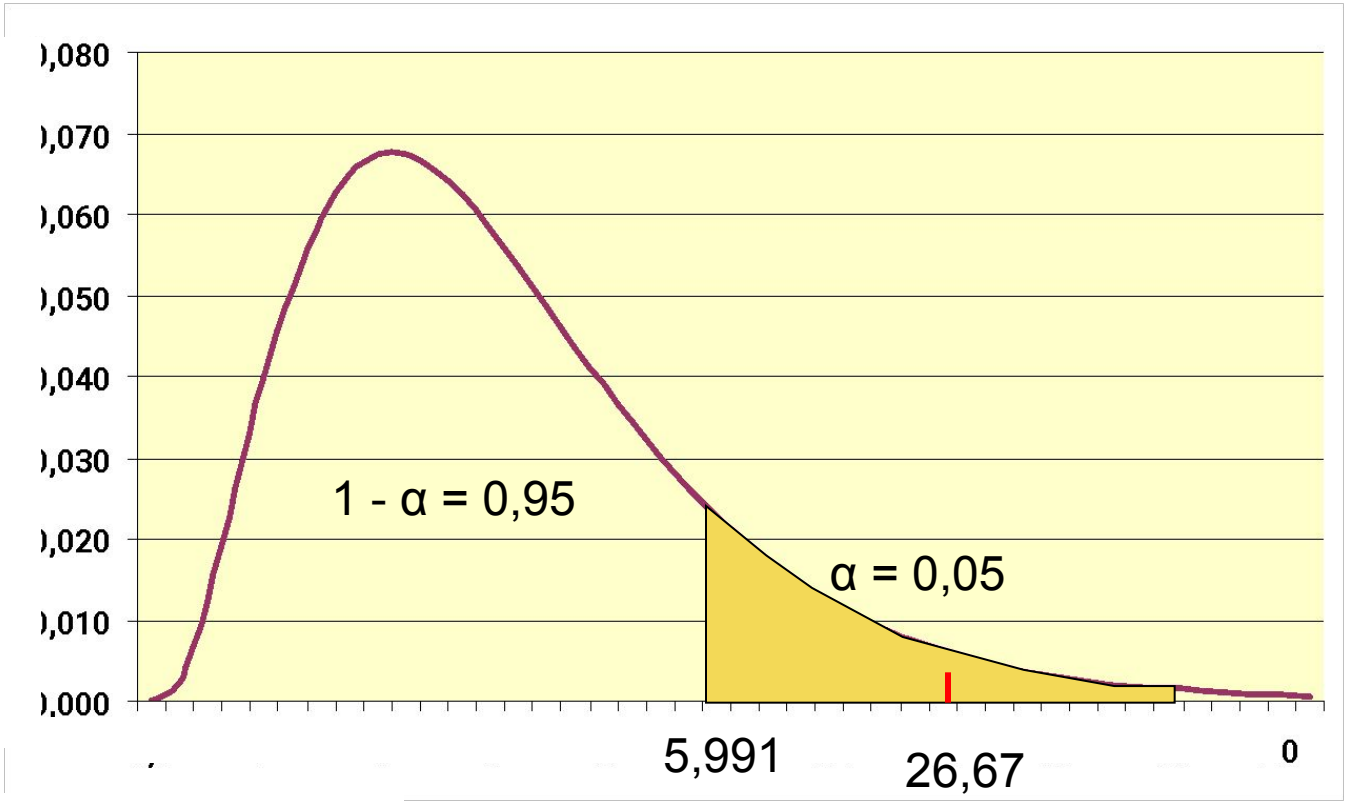
Вычисление статистики

Наблюдаемые частоты				Ожидаемые частоты		
100	80	20		75	100	25
50	120	30		75	100	25

$$\chi^2 = \frac{(100-75)^2}{75} + \frac{(80-100)^2}{100} + \frac{(20-25)^2}{25} +$$
$$+ \frac{(50-75)^2}{75} + \frac{(120-100)^2}{100} + \frac{(30-25)^2}{25} = 26{,}67$$

Уровень значимости и критическая область

В условиях нулевой гипотезы статистика имеет χ^2-распределение с числом степеней свободы df = (r – 1)(c – 1) = (2 – 1)(3 – 1) = 2. Зададим α = 0,05, критическое значение равно 5,991.



=ХИ2ОБР(0,05;2)

Получение выводов

Поскольку значение статистики попало в критическую область, 26,67 > 5,991, мы отклоняем гипотезу о независимости признаков.

Вывод. Признаки зависимы. Отношение к новому лекарству существенно зависит от категории персонала.