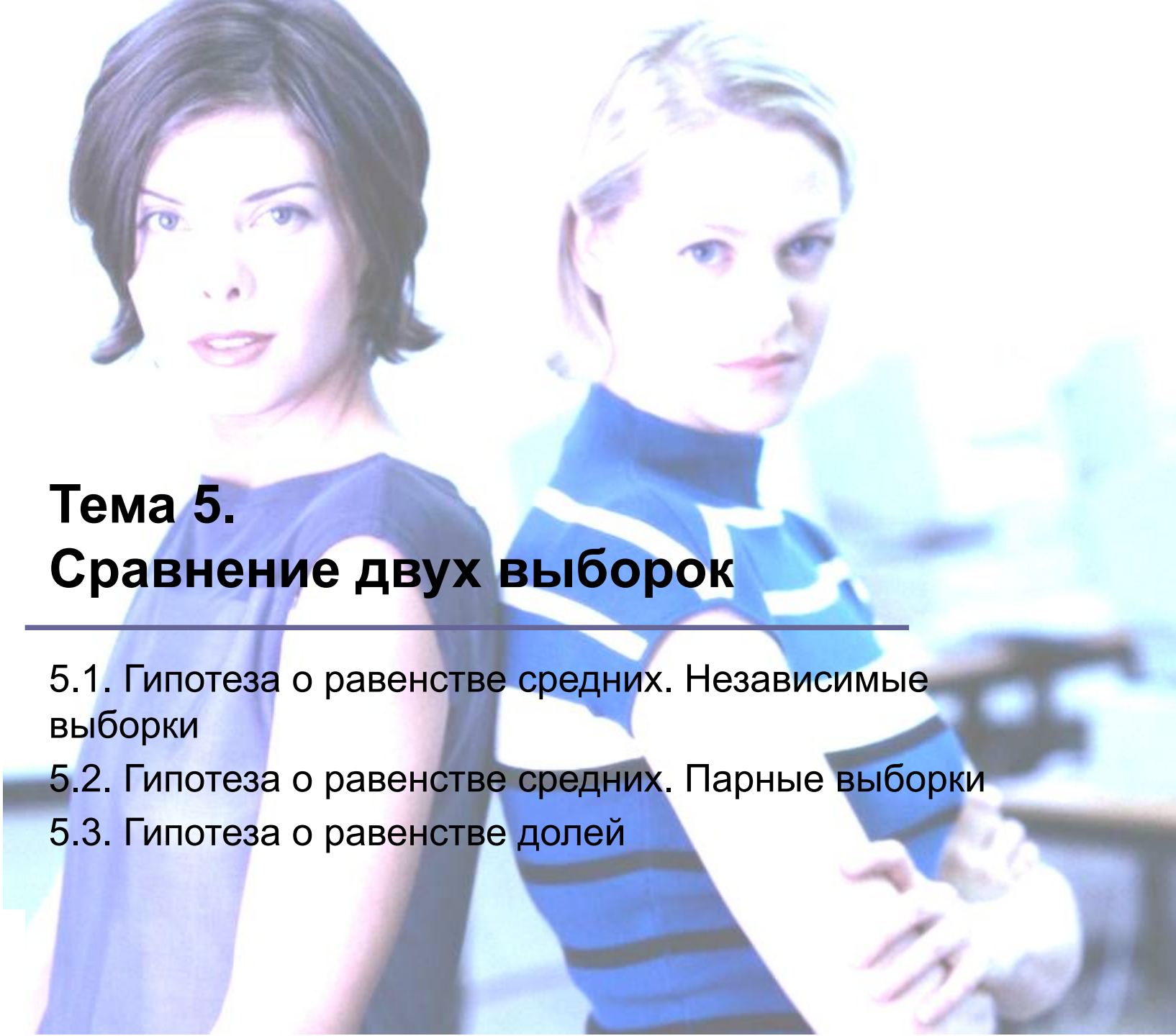
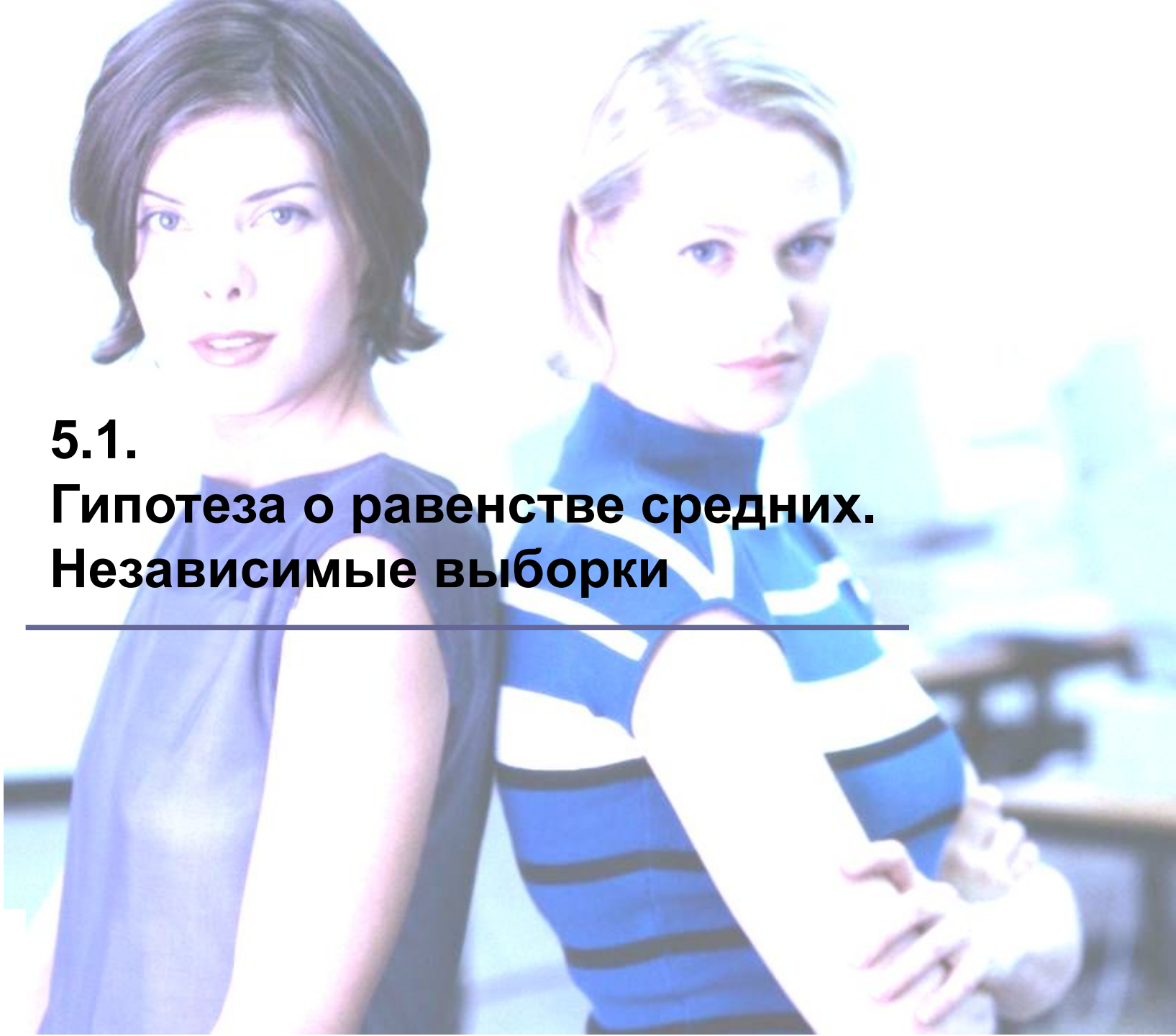


Тема 5.
Сравнение двух выборок

5.1. Гипотеза о равенстве средних. Независимые выборки

5.2. Гипотеза о равенстве средних. Парные выборки

5.3. Гипотеза о равенстве долей



5.1.
Гипотеза о равенстве средних. Независимые выборки

Пример

Представьте себе, что вы — региональный менеджер по продажам компании BLK Foods и хотите сравнить объемы продаж BLK-колы, выставленной на обычных полках и на специализированных стеллажах. Для этого вы создаете выборку, состоящую из 30 магазинов компании BLK Foods, в которых объявлена полная распродажа товаров. Затем вы случайным образом делите эту выборку пополам: 15 магазинов относите к первой группе, а остальные 15 — ко второй. Менеджеры магазинов из первой группы размещают бутылки с BLK-колой на обычных полках среди других прохладительных напитков. В то же время менеджеры магазинов из второй группы должны расположить бутылки с BLK-колой на специализированных стеллажах и разместить на них рекламу. Как определить, одинаковы ли объемы продаж BLK-колы в магазинах из этих двух групп?

Независимые выборки. Описание проблемы

Что мы имеем

1. Две случайные выборки, полученные из двух генеральных совокупностей
2. Выборки являются независимыми. Это значит, что между субъектами в каждой из выборок нет связи.
3. Выборки извлечены из нормальной генеральной совокупности. Если объем каждой выборки больше 30, то это требование не обязательно.

Что мы хотим

Проверить гипотезу о равенстве средних двух генеральных совокупностей:

$$H_0 : a_1 = a_2$$

Гипотеза

Нулевая гипотеза:

$$H_0 : a_1 = a_2$$

Альтернативная гипотеза:

$$H_1 : a_1 \neq a_2$$

Односторонние гипотезы

Нулевая гипотеза:

$$H_0 : a_1 \leq a_2 \qquad H_0 : a_1 \geq a_2$$

Альтернативная гипотеза:

$$H_1 : a_1 > a_2 \qquad H_1 : a_1 < a_2$$

Критерий Стьюдента для проверки равенства средних. Статистика

Для проверки гипотезы используется статистика:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

где $\overline{x}_1 \quad \overline{x}_2$ - выборочные средние

$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$ - объединенная дисперсия двух выборок

$n_1 \quad n_2$ - объемы выборок

$\mathrm{df} = n_1 + n_2 - 1$ - степени свободы

Последовательность действий

Шаг 1. Сформулировать основную и альтернативную гипотезы.

Шаг 2. По выборке сосчитать значение статистики.

Шаг 3. Задать уровень значимости α.

Шаг 4. По таблице найти критические значения и построить критическую область.

Шаг 5. Сравнить полученное значение с критической областью. Если значение попало в критическую область – отклонить основную гипотезу, не попало – принять.

Шаг 6. Написать ответ.

Пример По данным выборочного обследования домохозяйств необходимо определить существенно ли различается среднедушевой доход домохозяйств в Волгоградской и Саратовской областях

Регион	Среднедушевой доход домохозяйства
Волгоградская область	6000
Волгоградская область	9900
Волгоградская область	17800
Волгоградская область	10000
Волгоградская область	5000
Волгоградская область	5500
Волгоградская область	3645
Волгоградская область	6900
Волгоградская область	6200
Волгоградская область	12167
Волгоградская область	8100
Волгоградская область	5880
Волгоградская область	6900
Волгоградская область	5000

151 домохозяйство

Саратовская область	9667
Саратовская область	8648
Саратовская область	7400
Саратовская область	6197
Саратовская область	7028
Саратовская область	22500
Саратовская область	13000
Саратовская область	5300
Саратовская область	6800
Саратовская область	5650
Саратовская область	8000
Саратовская область	5451
Саратовская область	7768
Саратовская область	8713

142 домохозяйства

Шаг 1. Сформулировать основную и альтернативную гипотезы.

$H_0 : a_1 = a_2$

Среднедушевой доход в Саратовской и Волгоградской областях одинаков

$H_1 : a_1 \neq a_2$

Среднедушевой доход в Саратовской и Волгоградской областях отличается

Шаг 2. По выборке сосчитать значение статистики.

1. Вычисляем выборочные средние (СРЗНАЧ)

$\overline{x}_1 = 8044$ Средний среднедушевой доход в Волгоградской области

$\overline{x}_2 = 8891$ Средний среднедушевой доход в Саратовской области

2. Вычисляем выборочные дисперсии (ДИСП)

$s_1^2 = 17563297$ Выборочная дисперсия в Волгоградской области

$s_2^2 = 62988196$ Выборочная дисперсия в Саратовской области

3. Вычисляем общую выборочную дисперсию по формуле

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)} = \frac{(151-1)s_1^2 + (142-1)s_2^2}{(151-1)+(142-1)} = 39573300$$

4. Вычисляем t-статистику по формуле

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{8044 - 8891}{\sqrt{39573300\left(\frac{1}{151} + \frac{1}{142}\right)}} = -1{,}15$$

Шаг 3. Задать уровень значимости α. (вероятность того, что мы ошибемся, отвергая $H_0 : a_1 = a_2$)

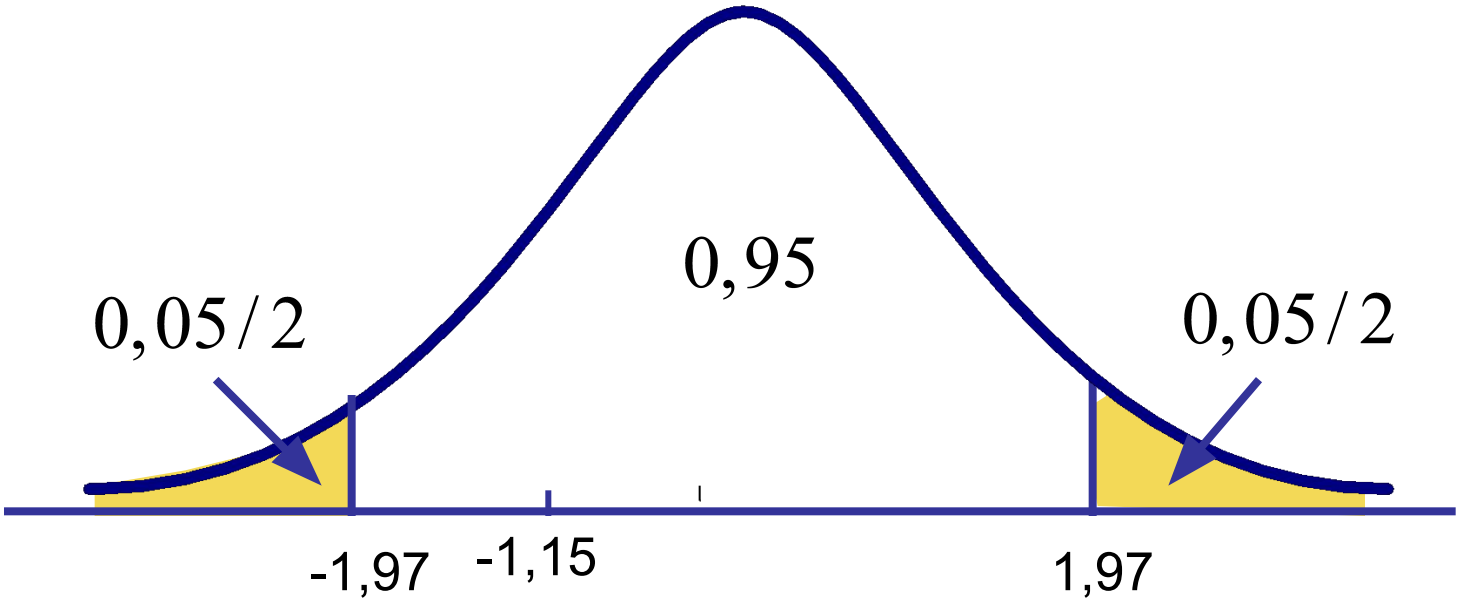
.

Пусть $\alpha = 0,05$

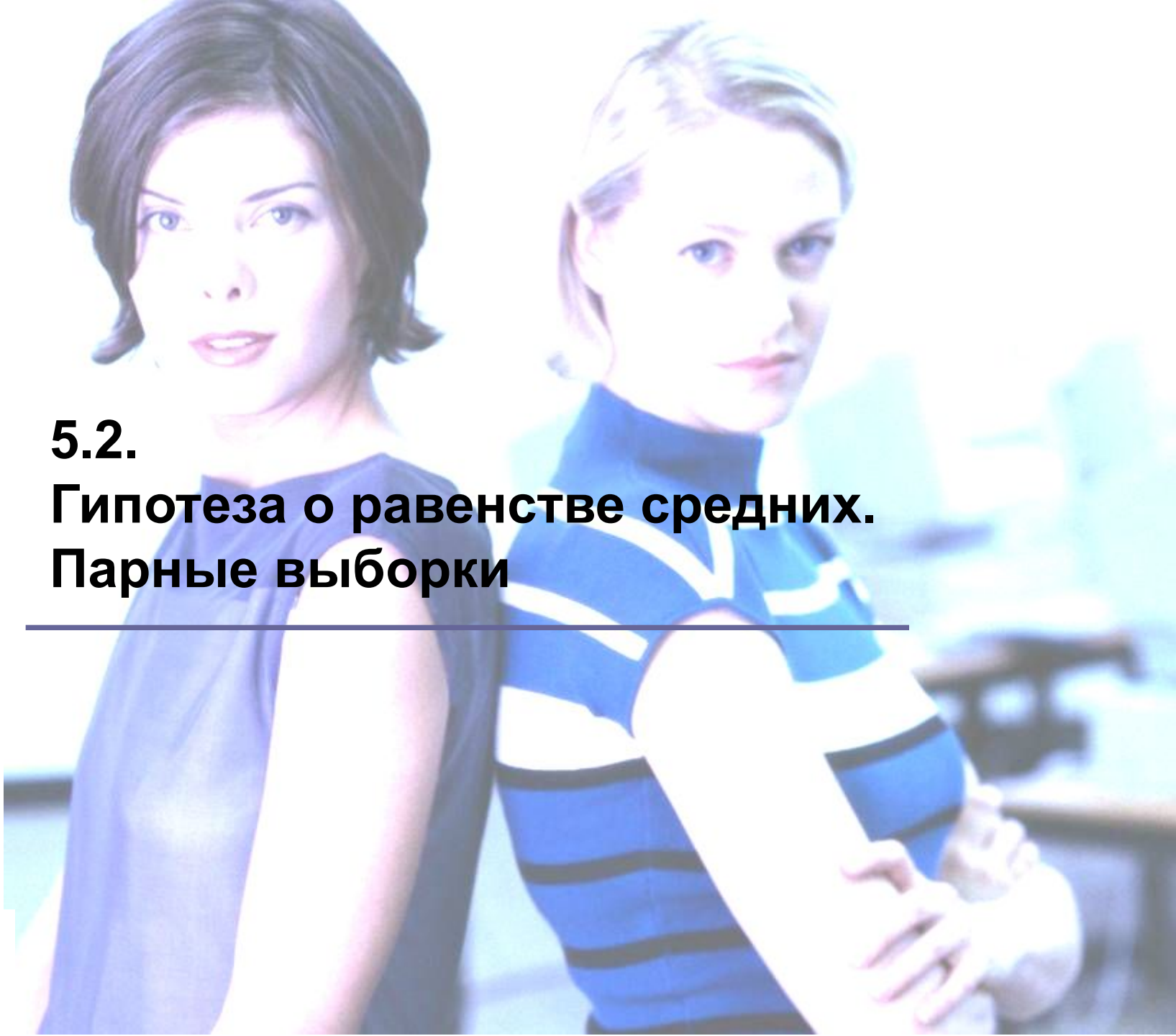
Шаг 4. По таблице найти критические значения и построить критическую область.

Критерий двусторонний.
= СТЬЮДРАСПОБР(0,05;151+142-2)

Критическое значение 1,968151



Вывод: Нет оснований отвергать основную гипотезу. Среднедушевой доход в Саратовской и Волгоградской областях одинаков



5.2.
Гипотеза о равенстве средних. Парные выборки

Пример

Предположим, что некая компания разрабатывает новое программное обеспечение для финансовых расчетов. Поскольку одним из основных критериев качества программного обеспечения является скорость вычислений, разработчики стремятся к тому, чтобы их пакет не уступал по своим возможностям лидерам рынка программ, но превосходил их по скорости расчетов. Если новый пакет окажется эффективным, он будет приводить к тем же результатам, что и другие программы, но за более короткое время.
Для оценки программного обеспечения разработчики провели эксперимент, в ходе которого один и тот же набор задач решали как с помощью стандартных программ, так и с помощью нового пакета. Поскольку измерения для каждой конкретной задачи проводились согласованно, для оценки эффективности пакета необходимо сравнить не средние значения двух независимых выборок, а среднюю разность между соответствующими элементами.

Парные выборки. Описание проблемы

Что мы имеем

1. Две случайные выборки, полученные из двух генеральных совокупностей
2. Выборки являются парными (зависимыми)
3. Обе выборки взяты из нормально распределенных генеральных совокупностей. Если объем каждой выборки больше 30, то это требование не обязательно.

Что мы хотим

Проверить гипотезу о разности средних двух генеральных совокупностей:

$$H_0 : a_1 = a_2$$

Статистика для парных выборок

Для проверки гипотезы используется статистика:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \qquad \mathrm{df} = n - 1$$

где

d - разность между двумя значениями x – y в одной паре

$\bar{d}$ - среднее для парных разностей для выборки

s_d - стандартное отклонение разностей для выборки

n - количество пар

Пример. Тренинг студентов

Студент	До	После	d	d^2
1	90	93	-3	9
2	91	90	1	1
3	93	89	4	16
4	89	88	1	1
5	85	88	-3	9
6	89	86	3	9
7	83	84	-1	1
8	88	83	5	25
9	84	83	1	1
10	82	80	2	4
11	83	77	6	36
12	81	76	5	25
13	72	74	-2	4
14	70	70	0	0
15	71	69	2	4
			Σ=21	Σ=145

Группа из 15 студентов прошла тест до тренинга и после. Результаты теста в таблице. Проверим гипотезу для парных выборок на отсутствие влияния тренинга на подготовку студентов на уровне значимости 0,05.

Решение. Подсчитаем разности и их квадраты.

Решение

Шаг 1. Основная и альтернативная гипотезы:

$H_0 : a_1 \geq a_2$ результаты теста не лучше, чем были до тренинга

$H_1 : a_1 < a_2$ результаты теста выше

Решение

Шаг 2. По выборке сосчитаем значение статистики.

$$s_d = \sqrt{\frac{\sum d^2 - \frac{\left(\sum d\right)^2}{n}}{n-1}} \quad \Rightarrow \quad s_d = \sqrt{\frac{145 - \frac{21^2}{15}}{15-1}} = 2{,}87$$

Можно использовать функцию ДИСП

Решение

Статистика принимает значение:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} = \frac{1,4}{\frac{2,87}{\sqrt{15}}} = 1,889$$

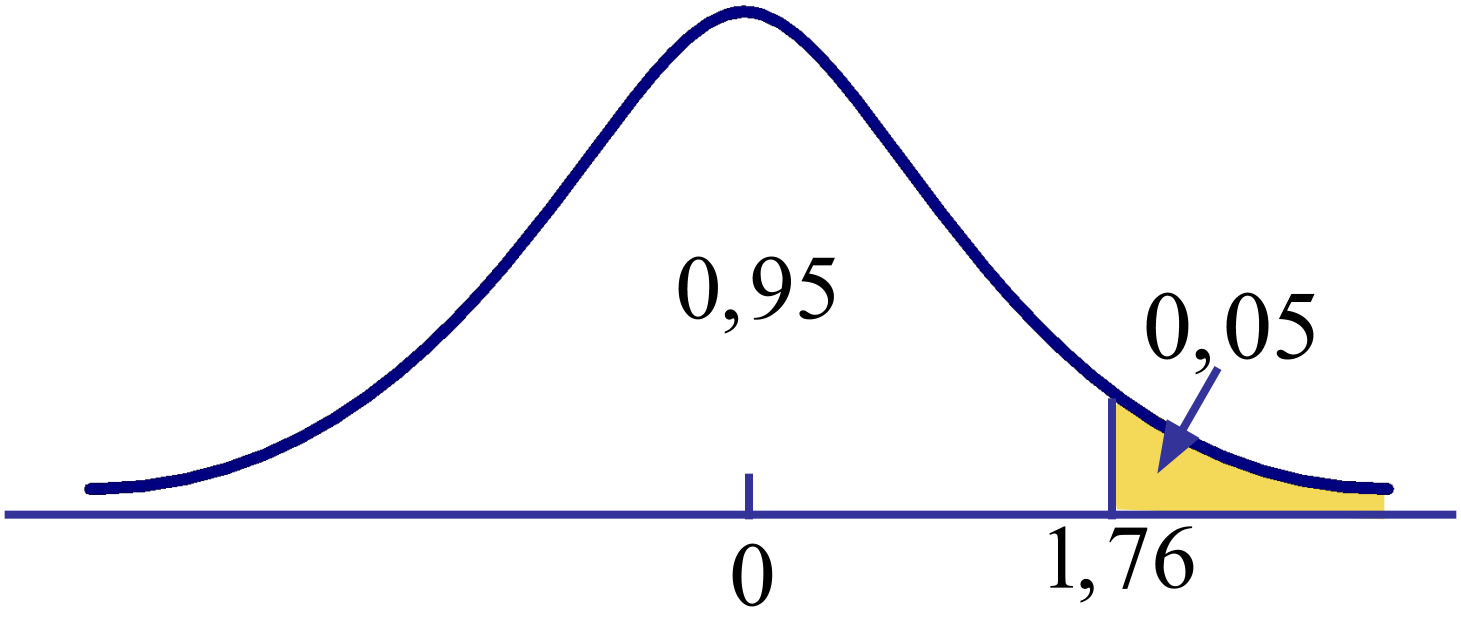
Среднее значение разностей получено делением 21 на 15 и равно 1,4.

Решение

Шаг 3. Задан уровень значимости α=0,05.

Шаг 4. По таблице или в Excel для степеней свободы df = 15 – 1=14 находим критическое значение t = 1,76 и строим критическую область:

=СТЬЮДРАСПОБР(0,1;14)



Решение

Шаг 5. Сравним полученное значение с критической областью.

$$1,889 > 1,76$$

Полученное значение статистики попало в критическую область.

Шаг 6. Формулируем вывод.

Нулевая гипотеза отвергается. Это означает, что влияние тренинга значимо на уровне значимости 0,05.

5.3. Гипотеза о равенстве долей

Пример

На одном из островов компании T. C, Resort Properties принадлежат два отеля: Beachcomer и Windsurfer. На вопрос “Планируете ли вы вернуться в наш отель снова?” 163 из 227 постояльцев отеля Beachcomer ответили: “Да”, в то же время 154 из 262 постояльцев отеля Windsurfer на этот вопрос ответили: “Нет”. Можно ли утверждать, что при уровне значимости, равном 0,05, между степенью удовлетворенности постояльцев обоих отелей (вероятностью, что в следующем сезоне они вернутся в отель) значимой разницы нет?

Гипотезы

Требуется проверить предположение о равенстве долей в двух генеральных совокупностях.

Нулевая гипотеза:

$$H_0 : p_1 = p_2$$

Альтернативная гипотеза:

$$H_1 : p_1 \neq p_2$$

Гипотезы

Требуется проверить превышает ли доля успехов в одной группе долю успехов в другой

Нулевая гипотеза:

$$H_0 : p_1 \leq p_2$$

Альтернативная гипотеза:

$$H_1 : p_1 > p_2$$

II

Нулевая гипотеза:

$$H_0 : p_1 \geq p_2$$

Альтернативная гипотеза:

$$H_1 : p_1 < p_2$$

III

Обозначения

$n_1 \quad n_2$ - объемы выборок

$m_1 \quad m_2$ - количество «успехов» в каждой выборке

$\hat{p}_1 = \frac{m_1}{n_1}$ - доля «успехов» в первой выборке

$\hat{p}_2 = \frac{m_2}{n_2}$ - доля «успехов» во второй выборке

$\bar{p} = \frac{m_1 + m_2}{n_1 + n_2}$ - общая доля «успехов» в обеих выборках

Статистика

В качестве статистики выбираем следующую случайную функцию:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

Статистика z имеет нормальное распределение, поэтому для проверки гипотезы пользуемся таблицей нормального распределения или функцией Excel НОРМСТОБР.

Пример

На одном из островов компании T. C, Resort Properties принадлежат два отеля: Beachcomer и Windsurfer. На вопрос “Планируете ли вы вернуться в наш отель снова?” 163 из 227 постояльцев отеля Beachcomer ответили: “Да”, в то же время 154 из 262 постояльцев отеля Windsurfer на этот вопрос ответили: “Да”. Можно ли утверждать, что при уровне значимости, равном 0,05, между степенью удовлетворенности постояльцев обоих отелей (вероятностью, что в следующем сезоне они вернутся в отель) значимой разницы нет?

Решение

Вычислим необходимые значения:

$$\hat{p}_1 = \frac{m_1}{n_1} = \frac{163}{227} = 0,718$$

$$\hat{p}_2 = \frac{m_2}{n_2} = \frac{154}{262} = 0,588$$

$$\overline{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{163 + 154}{227 + 262} = \frac{317}{489} = 0,648$$

Решение

Шаг 1. Основная и альтернативная гипотезы:

$$H_0 : p_1 = p_2 \qquad H_1 : p_1 \neq p_2$$

Шаг 2. По выборке сосчитаем значение статистики.

$$z = \frac{(0,718 - 0,588)}{\sqrt{0,648(1-0,648)\left(\frac{1}{227} + \frac{1}{262}\right)}} = 3,01$$

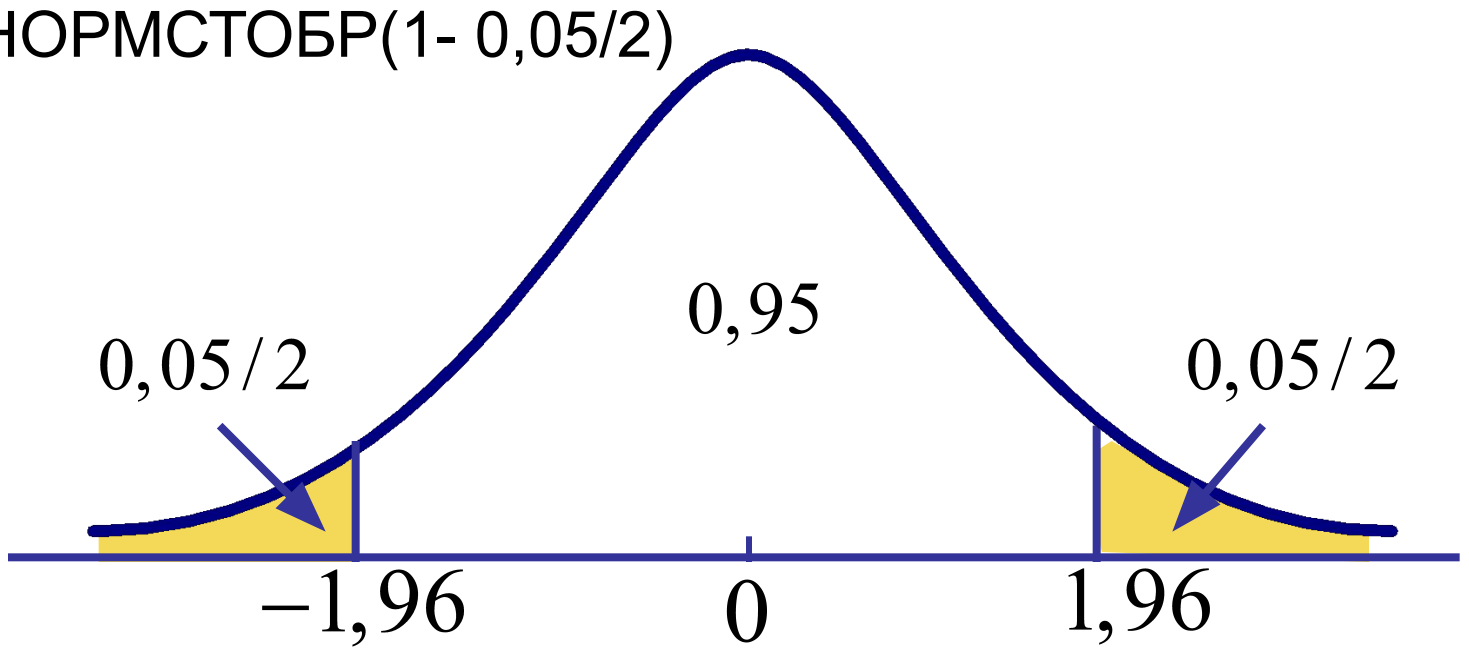
Решение

Шаг 3. Задан уровень значимости α=0,05.

Шаг 4. По таблице нормального распределения находим критические значения z = - 1,96 и z = 1,96 строим критическую область:

$$z < -1{,}96 \qquad z > 1{,}96$$

=НОРМСТОБР(1- 0,05/2)



Решение

Шаг 5. Сравним полученное значение с критической областью.

$$3,01 > 1,96$$

Полученное значение статистики попало в критическую область.

Шаг 6. Формулируем вывод. О**твергаем основную гипотезу. Два отеля значительно различаются по качеству обслуживания. В отеле** Beachcomer качество выше.