Команда №5

Проект по статистике

MathsHub, 12 марта 2024г

Над проектом работали

Варламов Никита	**Брагин Олег**
Гусев Сергей	**Раева Марина**

Google Collab с подробным описанием хода решения можно найти здесь

В проекте использованы библиотеки:

Задание 1

Менеджер сайта, предоставляющего независимым продавцам площадку для продаж, решил провести АБ-тест, выбрав в качестве метрики время обработки заказов продавцами. Для контрольной группы продавцов оставили предыдущий интерфейс работы с заказами, а для тестовой группы внедрили интерактивный дэшборд заказов.
Полученные результаты времени обработки заказов в часах для тестовой и контрольной групп представлены в csv-файлах *time_order_processing_test.csv* и *time_order_processing_control.csv.*
Проверьте гипотезу менеджера о том, что использование интерактивного дэшборда уменьшило время обработки заказов.

Подсказка: это независимые выборки.

Формулируем гипотезы

Нулевая гипотеза

$H_0: \mu_1 = \mu_2$

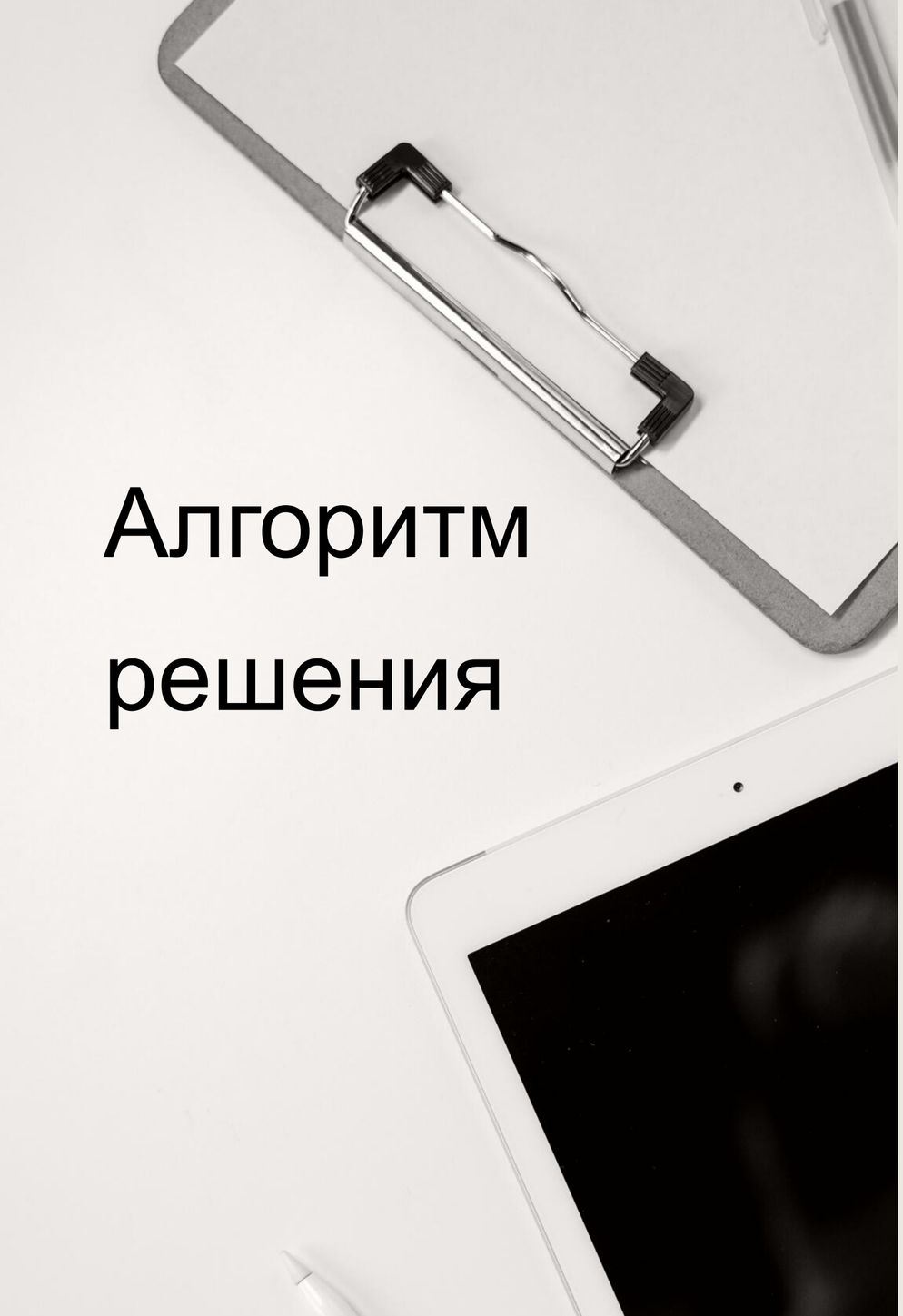
использование альтернативного дашборда не оказало никакого влияния на время обработки заказов продавцом.

Альтернативная гипотеза

$H_1: \mu_1 > \mu_2$

использование изменений в интерфейсе площадки агрегатора уменьшило время обработки заказов.

Уровень значимости альфа определим равным $\alpha = 0.05$



Алгоритм решения

01

Обзор данных

Загружаем CSV-данные

Проводим первичный обзор данных

Вывод некоторых статистических сведений

Визуально оцениваем плотность распределения данных (гистограмма + KDE/ЯОП)

02

Статистическая оценка данных

Сравним стандартные отклонений выборок

Применим двухвыборочный T-тест для независимых выборок

Интерпретируем результаты

```
[ ] # Загружаем CSV-даныне
    control = pd.read_csv(r'C:\Downloads\time_order_processing_control.csv')
    test = pd.read_csv(r'C:\Downloads\time_order_processing_test.csv')
```

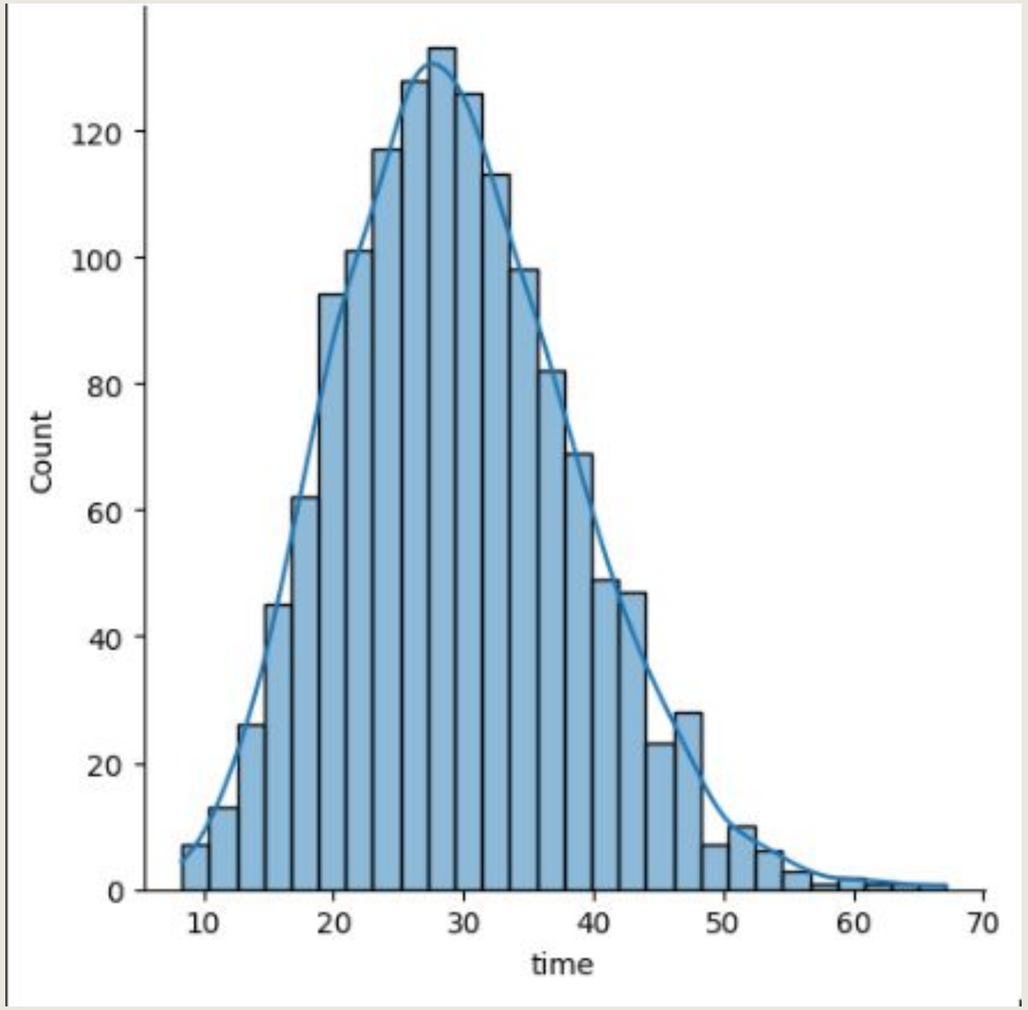
```
# Проводим первичный обзор данных
# 1. Просмотр датафрейма, вывод нескольких первых строк.
#control.head(5)
test.head(5)
```

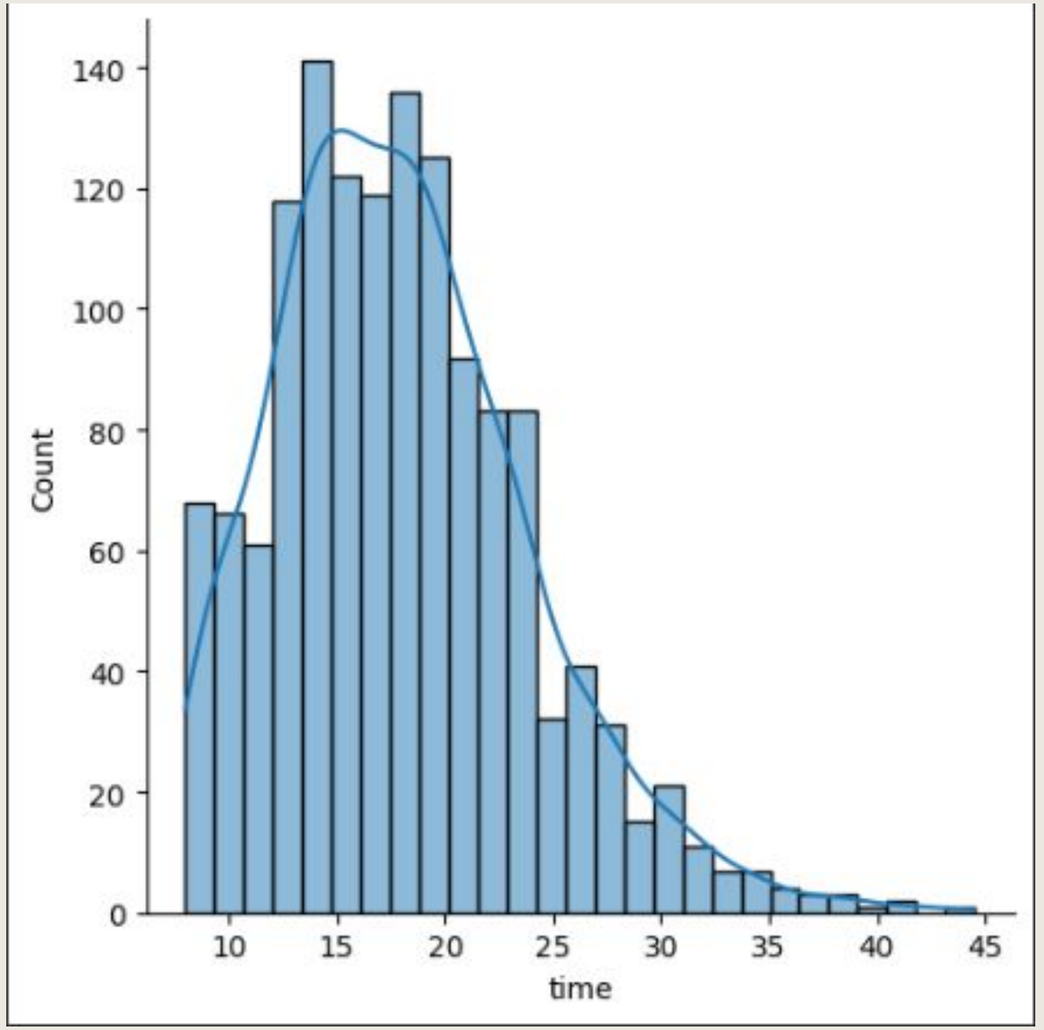
	Unnamed: 0	time
0	0	16.56
1	1	23.76
2	2	12.12
3	3	13.80
4	4	15.08

```
# 2. Вывод некоторых статистических сведений
#control.describe()
test.describe()
```

	Unnamed: 0	time
count	1393.000000	1393.000000
mean	696.000000	18.009160
std	402.268774	5.857845
min	0.000000	8.000000
25%	348.000000	13.840000
50%	696.000000	17.480000
75%	1044.000000	21.520000
max	1392.000000	44.520000

```python
# 3. Визуально оцениваем плотность распределения данных (гистограмма + KDE/ЯОП)
# Kernel Density Estimation, KDE или ядерная оценка плотности - один из непараметрических способов оценки плотности
# случайной величины, особенностью которого является сглаживание данных.
g_c = sns.displot(control, x="time", kde=True)
g_t = sns.displot(test, x="time", kde=True)
```





```
# Сравним стандартные отклонений выборок, для определения необходимости применением теста Уэлча.
control_std = control['time'].std( )
print(f'Стандартное отклонение контрольной выборки: {control_std:.4f}')
test_std = test['time'].std( )
print(f'Стандартное отклонение тестовой выборки: {test_std:.4f}')
```

```
Стандартное отклонение контрольной выборки: 8.9799
Стандартное отклонение тестовой выборки: 5.8578
```

Разница между стандартными отклонениями двух выборок, очевидно и ожидаемо, не существенна (equal_var=True).

Применим двухвыборочный Т-тест для независимых выборок.
index = 1 for control, index = 2 for test
Нулевая гипотеза **H_0: mu_1 = mu_2,** т.е.использование альтренативного дэшборда не оказало никакого влияния на время обработки заказов продавцом.
Альтернативная гипотеза **H_1: mu_1 > mu_2,** т.е. использование изменений в интерфейсе площадки агрегатора уменьшило время обработки заказов. Уровень значимости альфа определим равным α = 0.05

```python
# T-тест для сравнения среднего значения (независимые выборки)
t_statistic, p_value = st.ttest_ind(control['time'], test['time'], equal_var=True, alternative = "greater")

print("Результаты t-теста:")
print(f"t-статистика: {t_statistic:.5f}")
print(f"p-значение: {p_value}")
```

Интерпретация результатов:

alpha = 0.05

if p_value < alpha:

    print("\nОтвергаем нулевую гипотезу. Использование интерактивного дэшборда значимо уменьшило время обработки заказов.")

else:

    print("\nНе отвергаем нулевую гипотезу. Разница в среднем времени обработки заказов имеет случайный характер и статистически не значима. Положительное влияние изменений в интерфейсе не доказано"

Результаты теста

```
Результаты t-теста:
t-статистика: 40.01056
p-значение: 2.9949647128013707e-277

Отвергаем нулевую гипотезу. Использование интерактивного дэшборда значимо уменьшило время обработки заказов.
```

Внедрение интерактивного дэшборда значимо уменьшило время обработки заказов.

Задание 4 (дополнительное)

Постройте 90%-е доверительные интервалы по выборкам из задания 1 для среднего времени обработки заказа продавцами, использующими и не использующими интерактивный дашборд.

Формула доверительного интервала

$$\overline{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Расчет доверительного интервала производится по формуле: среднее значение +/- критическое значение * стандартное отклонение / квадратный корень из размера выборки

Зададим доверительный уровень, в соответствии с условиями задачи:
confidence_level = 0.9
alpha = 0.1

```
[ ]  # Средние и стандартные отклонения для двух выборок
     mean_control_group = control['time'].mean()
     std_control_group = control['time'].std()

     mean_test_group = test['time'].mean()
     std_test_group = test['time'].std()

     # Размер выборок
     n_control_group = len(control)
     n_test_group = len(test)

     # Определение критического значения Z_a/2 для 90% уровня доверия
     z_control_group = 1.65
     z_test_group = 1.65
```

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

Z-таблица критических значений

Полная формула доверительного интервала с указанием стандартных критических значений Z

FORMULA FOR THE CONFIDENCE INTERVAL

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{X}$ = sample mean
N = sample size
σ = standard deviation

For 90% confidence interval: $z_{\alpha/2} = 1.65$

95% confidence interval: $z_{\alpha/2} = 1.96$

99% confidence interval: $z_{\alpha/2} = 2.58$

```python
# Расчёт доверительных интервалов выборок
confidence_interval_low_control = mean_control_group - z_control_group * (std_control_group / np.sqrt(n_control_group))
confidence_interval_high_control = mean_control_group + z_control_group * (std_control_group / np.sqrt(n_control_group))

confidence_interval_low_test_group = mean_test_group - z_test_group * (std_test_group / np.sqrt(n_test_group))
confidence_interva_high_test_group = mean_test_group + z_test_group * (std_test_group / np.sqrt(n_test_group))

print("Доверительный интервал для среднего времени обработки заказа без использования интерактивного дашборда:")
print(f"({confidence_interval_low_control:.3f}, {confidence_interval_high_control:.3f})")
print("Доверительный интервал для среднего времени обработки заказа с использованием интерактивного дашборда:")
print(f"({confidence_interval_low_test_group:.3f}, {confidence_interva_high_test_group:.3f})")
```

```
Доверительный интервал для среднего времени обработки заказа без использования интерактивного дашборда:
(29.106, 29.900)
Доверительный интервал для среднего времени обработки заказа с использованием интерактивного дашборда:
(17.750, 18.268)
```

Инструменты статистики scipy позволяют произвести подобный расчет в одну формулу, которая принимаем в себя три значения:

- доверительный уровень (90% в нашем случае)
- среднее значение выборки (Mean)
- стандартную ошибку среднего (Standard Error of Mean - SEM)

Размер выборок контрольной и тестовой групп достаточно велик, поэтому, согласно ЦПТ, построим доверительный интервал с использованием нормального распределения.

```
interval_con = st.norm.interval(confidence=0.9, loc=np.mean(control['time']), scale=st.sem(control['time']))
interval_test = st.norm.interval(confidence=0.9, loc=np.mean(test['time']), scale=st.sem(test['time']))
print(f'Для контрольной выборки 90% доверительный интервал составляет: ({(interval_con[0]):.4f} , {(interval_con[1]):.4f})')
print(f'Для тестовой выборки 90% доверительный интервал составляет: ({(interval_test[0]):.4f} , {(interval_test[1]):.4f})')
print('То есть, в 90 процентах случаев наш доверительный интервал будет включать истинное среднее значение времени обработки заказа.')
```

```
Для контрольной выборки 90% доверительный интервал составляет: (29.1071 , 29.8986)
Для тестовой выборки 90% доверительный интервал составляет: (17.7510 , 18.2673)
То есть, в 90 процентах случаев наш доверительный интервал будет включать истинное среднее значение времени обработки заказа.
```

Задание 2

Этот же маркетплейс предоставил с 1 мая часто заказывающим клиентам бесплатную доставку, действующую для всех заказов до конца календарного месяца, если в этом месяце клиент уже сделал пять заказов (то есть начиная с шестого заказа). В файле *clients_orders.csv* представлена информация о количестве заказов, которое тысяча случайно отобранных часто заказывающих клиентов сделали в апреле и мае. Проверьте гипотезу о том, что клиенты стали делать больше заказов после введения бесплатной доставки с шестого заказа.

Подсказка: это зависимая (парная) выборка.

Формулируем гипотезы

Нулевая гипотеза

$H_0: \mu_1 = \mu_2$

или $\mu_d = 0$

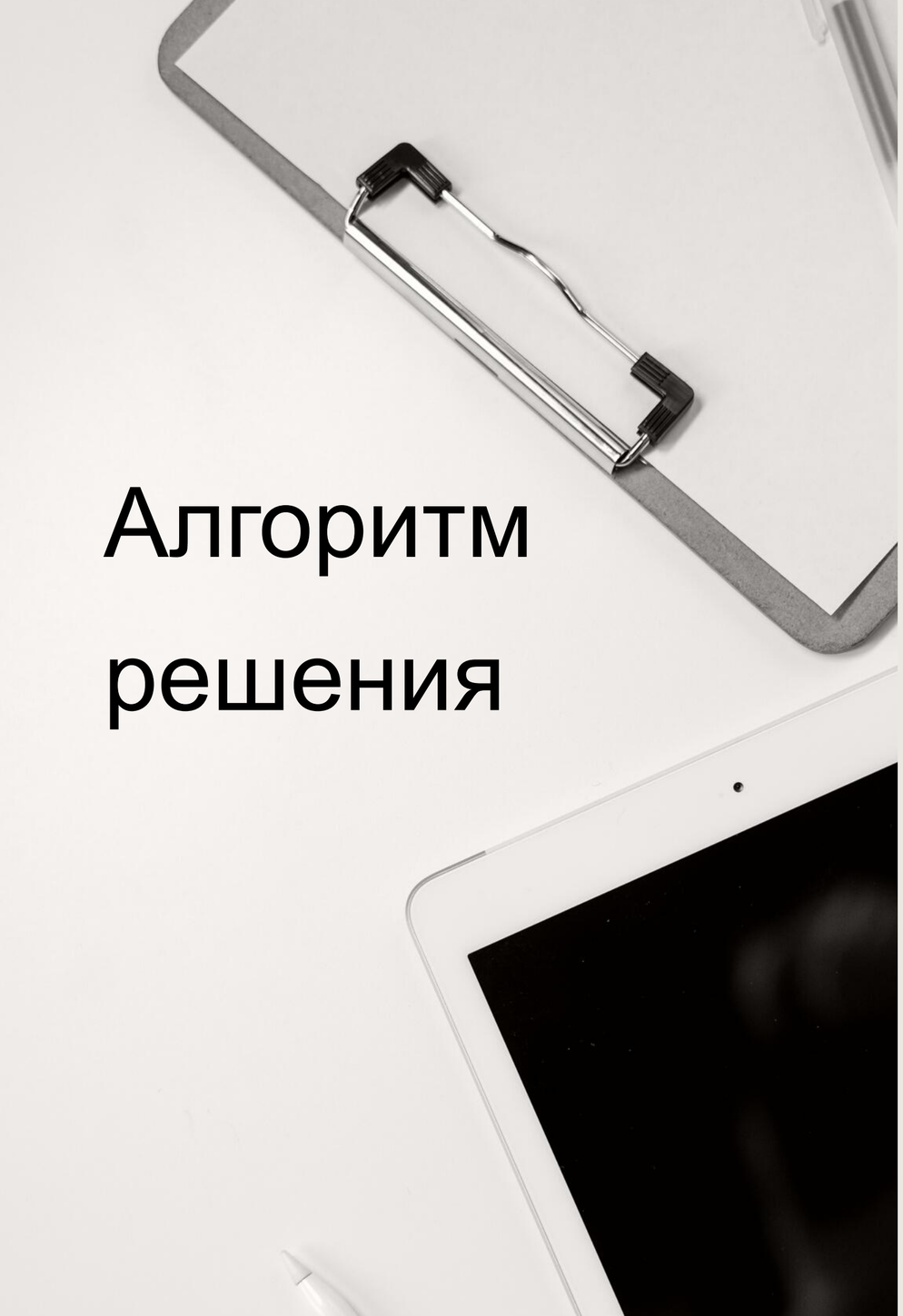
введение бесплатной доставки с 6-ого заказа не повлияло на число заказов за текущий месяц.

Альтернативная гипотеза

$H_1: \mu_1 < \mu_2$

введение нового бонуса увеличило число заказов за месяц.

Уровень значимости альфа определим равным $\alpha = 0.05$



Алгоритм решения

01

Обзор данных

Загружаем CSV-данные

Проводим первичный обзор данных

Вывод некоторых статистических сведений

Визуально оцениваем плотность распределения данных (гистограмма + KDE/ЯОП)

02

Статистическая оценка данных

Применим двухвыборочный T-тест для зависимых выборок.

Интерпретируем результаты

```
[ ] # Импорт данных
    orders = pd.read_csv(r'C:\Downloads\clients_orders.csv')
```

```
# Проводим первичный обзор данных
# 1. Просмотр датафрейма, вывод нескольких первых строк.
orders.head(10)
```

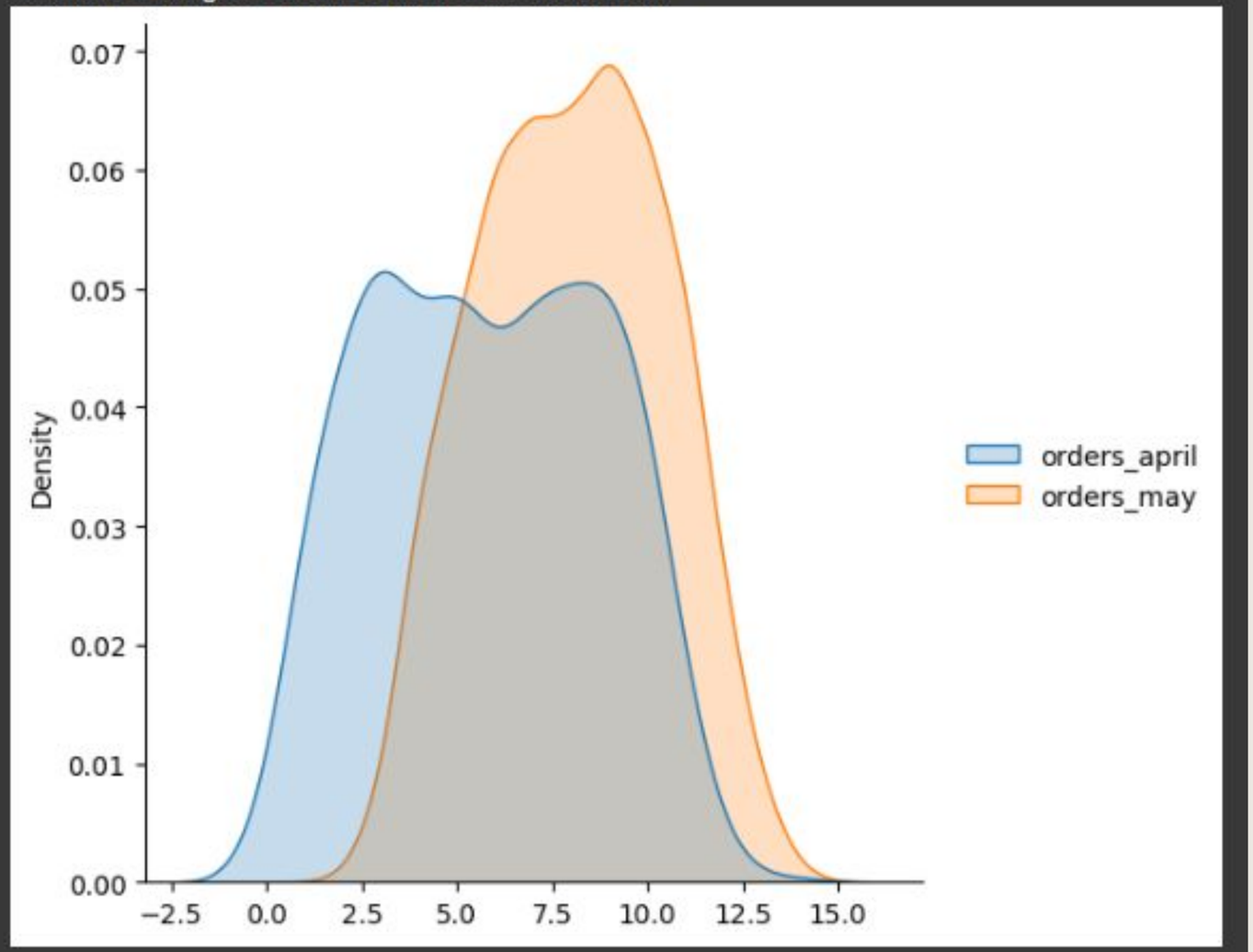
	Unnamed: 0	orders_april	orders_may
0	0	1	13
1	1	5	7
2	2	10	8
3	3	11	6
4	4	1	10
5	5	5	8
6	6	2	11
7	7	9	6
8	8	3	8
9	9	3	7

```
# 2. Вывод некоторых статистических сведений
orders.describe()
```

	Unnamed: 0	orders_april	orders_may
count	1000.000000	1000.000000	1000.000000
mean	499.500000	5.762000	7.982000
std	288.819436	3.006392	2.414437
min	0.000000	0.000000	2.000000
25%	249.750000	3.000000	6.000000
50%	499.500000	6.000000	8.000000
75%	749.250000	8.000000	10.000000
max	999.000000	14.000000	14.000000

```python
# 3. Визуально оцениваем плотность распределения данных (гистограмма + KDE/ЯОП)
orders_test = orders[['orders_april','orders_may']]
sns.displot(data=orders_test, kind="kde", fill=True)
```

```
<seaborn.axisgrid.FacetGrid at 0x2c42589c890>
```



```
[ ] # Применим двухвыборочный T-тест для зависимых выборок.
    # index = 1 for april, index = 2 for may
    # Нулевая гипотеза H_0: mu_1 = mu_2, или mu_d = 0, т.е. введение бесплатной доставки с 6-ого заказа не повлияло на число заказов за текущий месяц.
    # Альтернативная гипотеза H_1: mu_1 < mu_2, т.е. введение нового бонуса увеличило число заказов за месяц.
    # уровень значимости альфа = 0.05
```

```
# T-тест для сравнения среднего значения (зависимые выборки)
t_statistic_rel, p_value_rel = st.ttest_rel(orders['orders_may'], orders['orders_april'], alternative = "greater")

print("Результаты t-теста:")
print(f"t-статистика: {t_statistic_rel:.5f}")
print(f"p-значение: {p_value_rel}")

# Интерпретация результатов
alpha = 0.05
if p_value_rel < alpha:
    print("\nОтвергаем нулевую гипотезу. Клиенты стали делать больше заказов полсе введения акции на бесплатную доставку.")
else:
    print("\nНе отвергаем нулевую гипотезу. Введение бонуса на бесплатную доставку с 6-ого заказа не повлияло на число оформленных заказов пользователями за месяц.")
```

Результаты теста

```
Результаты t-теста:
t-статистика: 18.17200
p-значение: 2.80660251589552 54e-64

Отвергаем нулевую гипотезу. Клиенты стали делать больше заказов полсе введения акции на бесплатную доставку.
```

Акция действительно помогла увеличить количество заказов в мае

Задание 3

В файле vendors.csv представлена анонимизированная информация о продавцах маркетплейса: тип продукции (goods) и время осуществления продаж через этот маркетплейс (experience).

Переменная goods принимает значения:

clothes — одежда

electronics — электроника и техника

outdoor — товары для активного отдыха

beauty — косметика и товары для ухода

pets — товары для животных

Переменная experience принимает значения:

0-1 — до одного года продаж

1-3 — от одного до трёх лет продаж

3-5 — от трёх до пяти лет продаж

5- — от пяти лет продаж

Проверьте гипотезу о том, что стаж продаж на маркетплейсе не зависит от типа реализуемой продукции.

Формулируем гипотезы

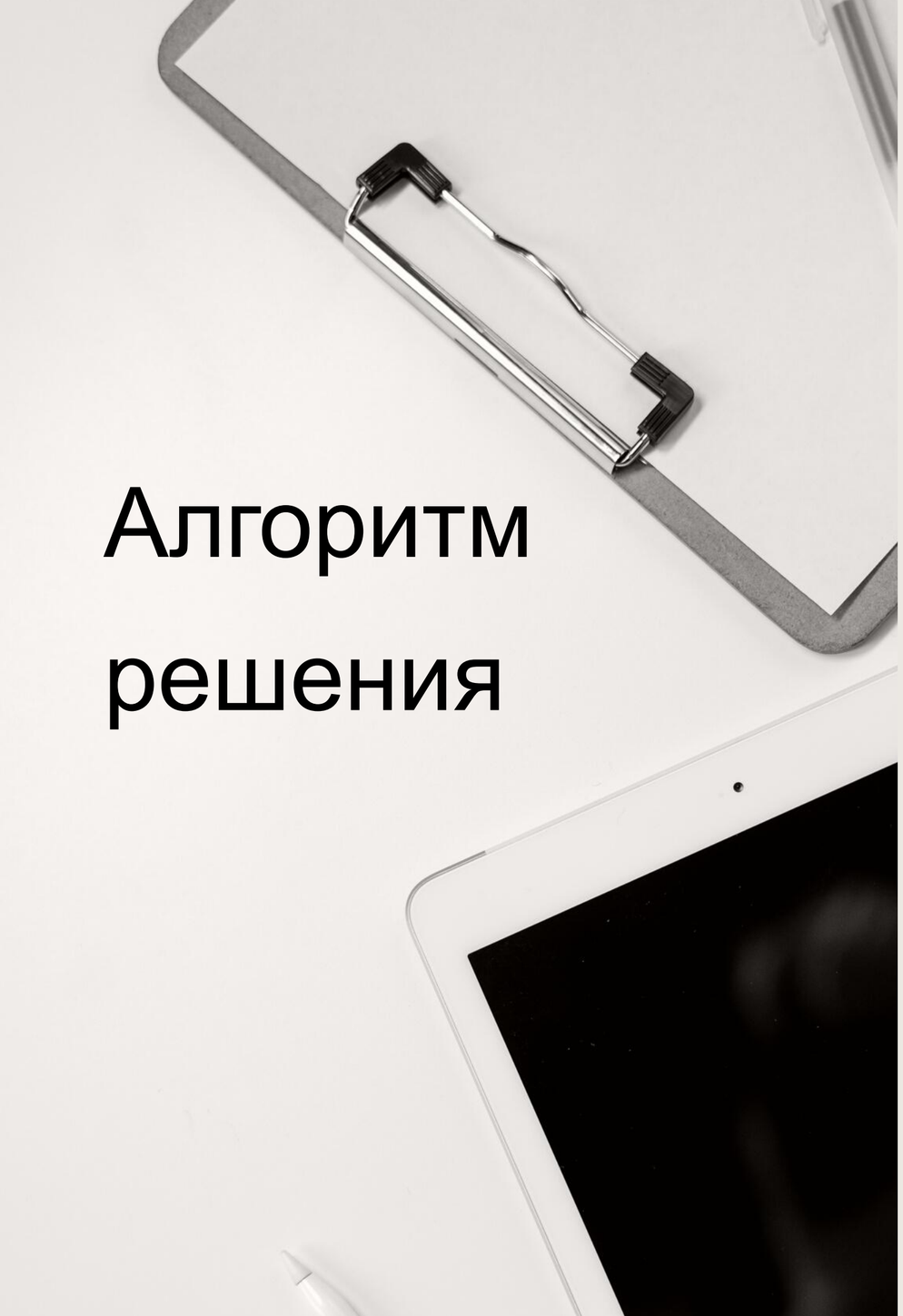
Нулевая гипотеза

H_0: Стаж продаж на маркетплейсе не зависит от типа реализуемой продукции.

Альтернативная гипотеза

H_1: Стаж продаж на маркетплейсе зависит от типа реализуемой продукции.

Уровень значимости альфа определим равным $\alpha = 0.05$



Алгоритм решения

01

Обзор данных

Загружаем CSV-данные

Проводим первичный обзор данных

Строим таблицу сопряженности

02

Статистическая оценка данных

Вычислим Хи-квадрат для данного случая

Интерпретируем результаты

```
[ ] # Импорт данных
    vendors = pd.read_csv(r'C:\Downloads\vendors.csv')
```

```
# Обзор датафрейма
vendors.head(10)
```

	Unnamed: 0	goods	experience
0	0	beauty	5-
1	1	beauty	3-5
2	2	outdoor	0-1
3	3	outdoor	0-1
4	4	clothes	3-5
5	5	clothes	1-3
6	6	clothes	1-3
7	7	clothes	5-
8	8	clothes	1-3
9	9	clothes	5-

```
# Построение таблицы сопряженности
table = pd.crosstab(vendors['goods'], vendors['experience'])
table
```

experience	0-1	1-3	3-5	5-
goods				
beauty	40	65	22	26
clothes	104	129	42	46
electronics	68	67	29	31
outdoor	36	46	23	26
pets	13	17	4	5

```python
res = chi2_contingency(table)

print(f'хи-квадрат стастистика: {res.statistic:.5f}')
print(f'p-value: {res.pvalue:.5f}')
print('степеней свободы:', res.dof)
print('рассчитанные ожидаемые значения:\n', res.expected_freq)
```

```
хи-квадрат стастистика: 9.17797
p-value: 0.68766
степеней свободы: 12
рассчитанные ожидаемые значения:
 [[ 47.59594756  59.08462455  21.88319428  24.43623361]
 [ 99.85816448 123.96185936  45.91179976  51.2681764 ]
 [ 60.66150179  75.30393325  27.89034565  31.14421931]
 [ 40.75208582  50.58879619  18.73659118  20.92252682]
 [ 12.13230036  15.06078665   5.57806913   6.22884386]]
```

```python
# Интерпретация результатов
# Проверка гипотезы путем сравнения Хи-квадрат статисткиик с критическим значением χ2(5%)_12.
chi2_05 = 21.026
if res.statistic > chi2_05:
    print("\nОтвергаем нулевую гипотезу. Существует статистически значимая зависимость стажа продаж от типа реализуемой ЧЯКАВС продукции.")
else:
    print("\nНе отвергаем нулевую гипотезу. Нет достаточных (значимых) оснований полагать, что стаж продаж на маркетплейсе зависит от типа реализуемой продукции.")
```

```
хи-квадрат стастистика: 9.17797
p-value: 0.68766
степеней свободы: 12
рассчитанные ожидаемые значения:
 [[ 47.59594756  59.08462455  21.88319428  24.43623361]
 [ 99.85816448 123.96185936  45.91179976  51.2681764 ]
 [ 60.66150179  75.30393325  27.89034565  31.14421931]
 [ 40.75208582  50.58879619  18.73659118  20.92252682]
 [ 12.13230036  15.06078665   5.57806913   6.22884386]]

Не отвергаем нулевую гипотезу. Нет достаточных (значимых) оснований полагать, что стаж продаж на маркетплейсе зависит от типа реализуемой продукции.
```

```
# Проверка гипотез путем сравнения значения p-value с уровнем значимости α
if res.pvalue < alpha:
    print("\nОтвергаем нулевую гипотезу. Существует статистически значимая зависимость стажа продаж от типа реализуемой продукции.")
else:
    print("\nНе отвергаем нулевую гипотезу. Нет достаточных (значимых) оснований полагать, что стаж продаж на маркетплейсе зависит от типа реализуемой продукции.")
```

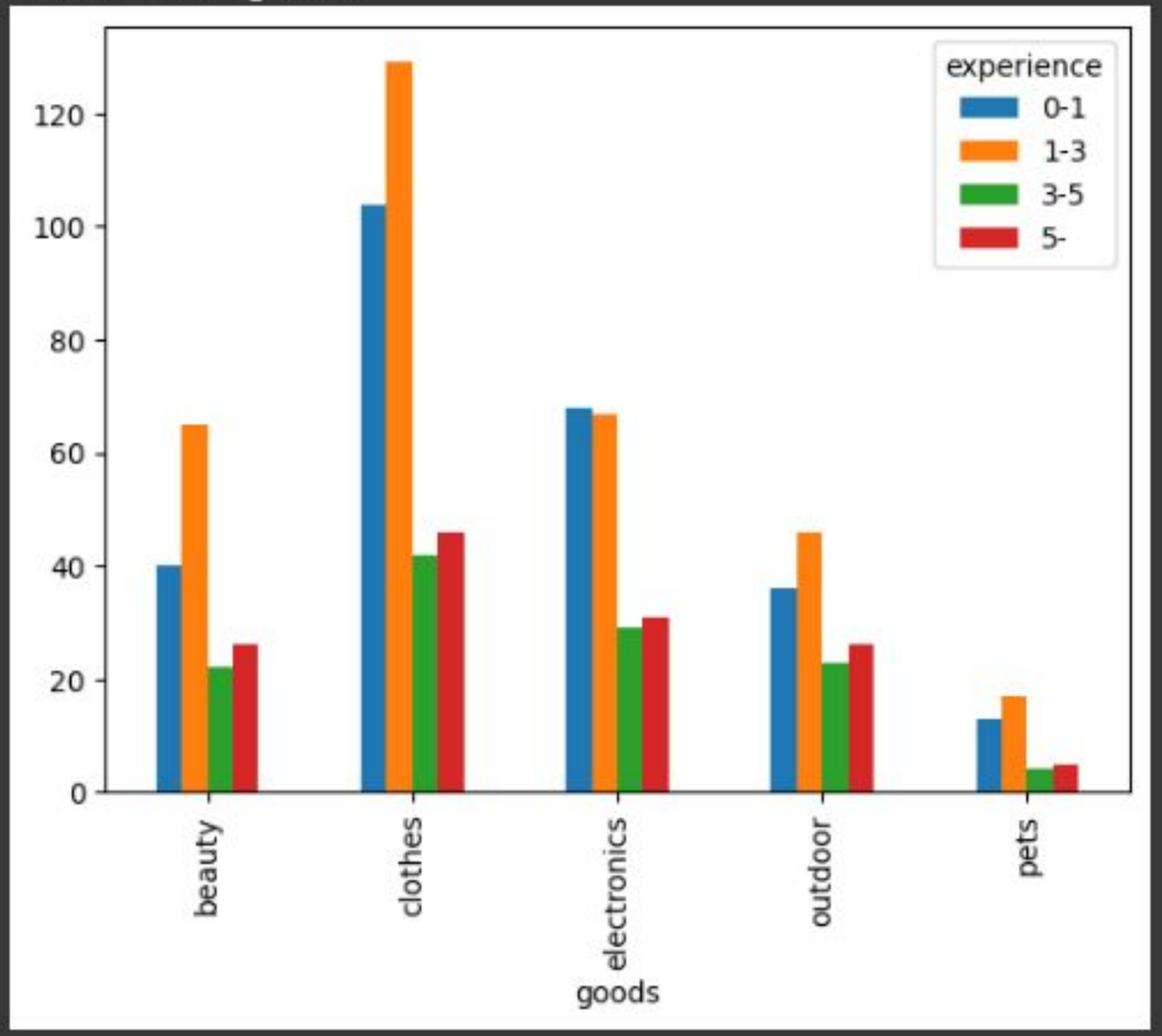
```
Не отвергаем нулевую гипотезу. Нет достаточных (значимых) оснований полагать, что стаж продаж на маркетплейсе зависит от типа реализуемой продукции.
```

	Significance level (α)							
Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Критические значения распределения Хи-квадрат

```python
# Построение столбчатой диаграммы
table.plot(kind='bar')
```

```
<Axes: xlabel='goods'>
```



Спасибо за внимание