Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«Московский государственный университет леса»

О.М.Полещук
Е.Г.Комаров

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Практикум
для всех специальностей МГУЛ

Москва

Издательство Московского государственного университета леса

2013

УДК 519.21
ББК 22.171

П49 Полещук О.М., Комаров Е.Г. Лабораторный работы по математической статистике: Учебно-методическое пособие для студентов всех специальностей МГУЛ. – М.: ФГБОУ ВПО МГУЛ, 2013. – 76 с.: ил.

Рецензенты: доктор технических наук, профессор Домрачев В.Г.,
доктор технических наук, профессор Ретинская И.В.

Работа подготовлена на кафедре высшей математики факультета электроники и системотехники МГУЛ.

Одобрено и рекомендовано к изданию в качестве учебно-методического пособия редакционно-издательским советом университета.

Автор: Ольга Митрофановна Полещук, профессор
Евгений Геннадиевич Комаров, профессор

Учебно-методическое пособие

Полещук Ольга Митрофановна
Комаров Евгений Геннадиевич

Лабораторные работы по математической статистике

В авторской редакции

Компьютерный набор и верстка авторов

По тематическому плану внутривузовских изданий учебной литературы на 2013 год.

Лицензия ЛР № 020718 от 02.02.1998 г.
Лицензия ЛР № 00326 от 14.02.2000 г.

Подписано к печати 01.06.13 Формат 60 × 90/16
Бумага 80 г/ $м^2$ Ризография
Гарнитура «Таймс» Заказ № 93 Тираж 100экз.
Объем 4,75 п. л.

Издательство Московского государственного университета леса, e-mail: izdat@mgul.ac.ru, 141005, Мытищи-5, Московская обл., 1-ая Институтская, 1, МГУЛ, телефон: (495) 588-57-62.
По вопросам приобретения литературы издательства ФГБОУ ВПО МГУЛ обращаться в отдел реализации, телефон: (498) 687-41-33, e-mail: kurilkina@mgul.ac.ru.

Оглавление

ВВЕДЕНИЕ

Математическая статистика - это раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений с целью выявления существующих закономерностей. Выводы о закономерностях, которым подчиняются явления, изучаемые методами математической статистики, всегда основываются на ограниченном, выборочном числе наблюдений.

Оценив неизвестные величины или зависимости между ними по данным наблюдений, исследователь выдвигает ряд гипотез (предположений) о том, что рассматриваемое явление можно описать той иди иной вероятностной
теоретической моделью. Далее, используя методы математической статистики, можно дать ответ на вопрос, какую из гипотез или моделей следует принять. Именно эта модель и есть закономерность изучаемого явления. Таков типичный путь исследования на основе аппарата математической статистики.

Предлагаемый курс лабораторных работ направлен на обучение студентов методам обработки статистических данных, их анализа и управления с помощью компьютера. Лабораторный практикум включает в себя девять лабораторных работ по основным темам, предусмотренным учебной программой по дисциплине. Для его выполнения необходимо программное обеспечение MS EXCEL.

В ходе выполнения каждой лабораторной работы студент оформляет и сдает отчет, который должен содержать: название лабораторной работы; цель работы; постановку задания; результаты решения на компьютере; анализ полученного решения, интерпретация результатов; выводы и заключения по заданию.

ЛАБОРАТОРНАЯ РАБОТА № 1
СТАТИСТИЧЕСКАЯ ОБРАБОТКА ДАННЫХ

Цель: Научиться основным методам обработки данных, представленных выборкой, путем построения гистограммы, определения выборочного среднего, выборочной дисперсии, выборочной медианы и моды.

Вероятностная модель ставит в соответствие результатам наблюдений

$$x_1, x_2, ..., x_n \tag{1}$$

последовательность случайных величин

$$X_1, X_2, ..., X_n \ . \tag{2}$$

Предполагается, что случайные величины $X_1, X_2, ..., X_n$ независимы и имеют одно и то же распределение с функцией распределения $F(x)$. Полагают, что наблюдения (1) являются значениями величин (2) при осуществлении вероятностного эксперимента. Несмотря на различие объектов (1) и (2), в математической статистике принято называть и то и другое ***выборкой из генеральной совокупности.***

Количество наблюдений n называется объемом выборки.

Произвольная случайная величина X характеризуется своей функцией распределения вероятностей $F(x)$. Если эта функция неизвестна, но известна выборка (1), числовые данные которой являются значениями случайной величины X, то возможно построить ***эмпирическую функцию распределения вероятностей*** $F_n(x)$, которая служит оценкой теоретической функции распределения вероятностей

$F(x)$. Если обозначить через $\mu_n(x)$ число тех значений $x_1, x_2, ..., x_n$, которые меньше или равны x, то

$$F_n(x) = \frac{\mu_n(x)}{n}\ . \qquad (3)$$

Если объем выборки n большой, то для представления о виде ее распределения строится ***гистограмма.***

Вводим в первый столбец (ячейки A1…) исходные данные. Для элементов выборки находим минимальный и максимальный элементы, которые ограничивают интервал, содержащий все элементы выборки. Для этого запишем в первую строку второго столбца (B1) слово Максимум, а во вторую строку второго столбца (B2) слово Минимум. В соседних ячейках C1 и C2 определим функции **MAX** и **MIN**. Для этого ставим курсор в C1 и вызываем мастер функций, нажав на кнопку *fx* , в открывшемся окне в поле «Категория» выбираем **СТАТИСТИЧЕСКИ**Е, и ниже ищем функцию МАКС и вызываем ее двойным щелчком по названию. В качестве аргумента функции (в графе «Число 1») обведем область данных (ячейки A1…). Поле «Число 2» оставляем пустым. Нажимаем «ОК». Ставим курсор в ячейку C2 и аналогично вводим функцию **МИН**. В некоторых случаях для удобства обработки интервал расширяется, но не существенно.

Следующим шагом является разбиение построенного интервала на 5-10 более мелких интервалов. Если разбиение построено удачно, то гистограмма будет напоминать график плотности (если она существует) распределения вероятностей случайной величины, значениями которой являются элементы выборки. Если разбиение мелкое, то гистограмма не дает представления о плотности распределения вероятностей из-за случайных флуктуаций. Если разбиение крупное, то гистограмма также не дает представления о плотности распределения вероятностей из-за того, что теряется много информации.

Чтобы построить интервалы разбиения (группировки), нужно от максимального значения выборки вычесть минимальное значение и полученный результат разделить на число интервалов. Полученное значение называется шагом разбиения. Чтобы получить верхние границы интервалов группировки, нужно последовательно прибавлять шаг разбиения, начиная от минимального значения выборки.

В ячейки D1… вводим верхние границы интервалов группировки. Для вычисления частот n_i используется функция **ЧАСТОТА**, находящаяся в категории **СТАТИСТИЧЕСКИЕ**. Введем ее в ячейку E1. В строке «Массив данных» введем диапазон выборки (ячейки A1…). В строке «Массив интервалов» введем диапазон верхних границ интервалов группировки (ячейки D1…). Результат функции является массивом и выводится в ячейках E1... Для полного вывода (не только первого числа в E1) нужно выделить ячейки E1…, обведя их мышью, и нажать F2, а далее одновременно CTRL+SHIFT+ENTER. Результат – частоты n_i, которые показывают, сколько элементов выборки попало в каждый из интервалов разбиения.

Для построения гистограммы в EXCEL 2003 нужно из меню **ВСТАВКА** выбрать **ДИАГРАММА** (или нажать на соответствующий значок **МАСТЕР ДИАГРАММ** на основной панели), при этом курсор должен стоять в свободной ячейке. Далее выбрать тип диаграммы: **ГИСТОГРАММА**, вид по выбору, нажать ДАЛЕЕ, в строке **ДИАПАЗОН** обвести частоты E1.., перейти на вкладку **РЯД**, в строке **ПОДПИСИ ОСИ** X ввести интервалы в ячейках D1…, нажать ДАЛЕЕ ввести название **ГИСТОГРАММА**, подписи осей: ось X - **ИНТЕРВАЛЫ** и ось Y - **ЧАСТОТА**, нажать ГОТОВО. Для создания полигона перейти на пустую ячейку и сделать то же самое, только вместо типа диаграммы **ГИСТОГРАММА**, выбрать **ГРАФИК**.

При использовании EXCEL 2007 для создания диаграммы необходимо выделить блок данных, на основании которых строится диаграмма. В выделяемый блок данных включить не только числовые данные, но и заголовки строк (столбцов), в которых они расположены. Заголовки будут использованы в качестве подписей по осям (меток) и для формирования условных обозначений (легенды). При выделении блоков с данными для построения диаграмм необходимо соблюдать два правила:

1. Выделенный фрагмент должен состоять из равновеликих столбцов.
2. В выделенном фрагменте не должно быть объединенных ячеек.

Для построения гистограммы необходимо перейти на вкладку **ВСТАВКА**, открыть список **ГИСТОГРАММА** выбрать нужную гистограмму. Гистограмма строится сразу. Иногда необходимо выделить построенную диаграмму и провести изменение размера шрифта или растянуть диаграмму для лучшего чтения данных в поле диаграммы. Если вызвать контекстное меню в поле всей диаграммы, то меню предлагает три отдельных шага в построении диаграммы (в предыдущих версиях было четыре шага): Изменить тип диаграммы; выбрать данные; переместить диаграмму.

В мастере функций *fx* существуют специальные функции, позволяющие вычислять выборочные характеристики.

Функция **СРЗНАЧ** вычисляет выборочное среднее (оценку теоретического математического ожидания) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Функция **ДИСП** вычисляет выборочную дисперсию (оценку теоретической дисперсии) $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$.

Функция **СТАНДОТКЛОН** вычисляет квадратный корень из выборочной дисперсии.

Функция **МЕДИАНА** вычисляет выборочную медиану (оценку медианы) заданной выборки. Медианой случайной величины называется то ее значение, которое делит распределение на две равновероятные половины. В качестве выборочной медианы $\tilde{m}$ в выборке объема $2n+1$ берут значение $x_{(n+1)}$ в вариационном ряде. Если объем выборки равен $2n$, то в качестве выборочной медианы $\tilde{m}$ берут $\frac{1}{2}(x_{(n)} + x_{(n+1)})$.

Функция **МОДА** вычисляет выборочную моду (оценку моды). Модой случайной величины называется ее наиболее вероятное значение.

В Excel можно генерировать случайные числа, имеющие разные законы распределения. Для этого можно использовать надстройку **АНАЛИЗ ДАННЫХ** и пункт **ГЕНЕРАЦИЯ СЛУЧАЙНЫХ ЧИСЕЛ**. Если вы хотите сгенерировать, например, 100 случайных чисел из нормального распределения, то в поле **ЧИСЛО ПЕРЕМЕННЫХ** введите 1; в поле **ЧИСЛО СЛУЧАЙНЫХ ЧИСЕЛ** введите 100; в списке **РАСПРЕДЕЛЕНИЕ** выберите **НОРМАЛЬНОЕ**; введите параметры нормального распределения – **СРЕДНЕЕ** и **СТАНДАРТНОЕ ОТКЛОНЕНИЕ.** В качестве типа распределения можно выбрать, например, **РАВНОМЕРНОЕ, БИНОМИАЛЬНОЕ** или **РАСПРЕДЕЛЕНИЕ ПУАССОНА**. Введя для каждого распределения соответствующие параметры, получим сгенерированные случайные числа.

Задания

Выборка состоит из 50 значений некоторой случайной величины. Построить гистограмму, вычислить выборочное среднее, выборочную дисперсию (исправленную), выборочные медиану и моду.

1.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	0.865	0.932	2.303	1.51	0.605	3.181	2.547	0.773	2.982	1.64	3.248

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	1.711	4.801	4.085	1.802	1.592	3.519	2.284	2.514	3.276	3.999	2.859	2.182	0.056

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	2.717	0.029	1.132	1.816	3.004	1.464	1.656	4.096	1.81	2.349	3.015	0.878	2.741

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	4.867	4.43	1.916	2.415	2.407	4.284	0.706	3.098	0.283	0.616	3.594	2.088	0.641

2.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	0.122	-0.359	0.053	-0.903	-2.371	1.087	0.759	2.113	5.384	2.617	2.97

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	2.724	2.831	2.346	-1.089	1.138	-0.511	2.393	0.636	-0.289	-0.446	-0.033	2.116	0.51

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	1.179	3.524	-0.411	1.004	3.215	2.785	-4.802	-3.314	1.412	-0.232	-1.395	1.198	2.542

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	1.612	1.023	-0.529	0.182	-0.348	0.736	3.036	1.361	2.027	2.48	0.967	1.558	1.324

3.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-1.338	0.122	-2.625	-1.733	0.682	-2.196	0.223	-0.831	-4.894	-2.761	0.11

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	-0.527	-0.979	0.068	0.391	-1.359	-1.921	0.794	1.937	-1.249	1.354	0.054	-1.774	-1.149

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	1.343	1.939	1.278	-0.372	-0.721	-0.731	-0.679	-1.327	3.126	-0.854	-0.937	-2.778	0.337

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	-2.081	0.282	3.309	0.862	0.342	1.774	3.885	1.307	0.905	0.85	0.046	-0.691	1.918

4.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-0.263	-1.213	-2.124	-0.824	-1.833	1.022	-0.107	-0.018	-0.664	3.053	1.798

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	-0.672	2.504	1.565	2.29	1.544	0.726	-1.365	1.687	-0.571	0.656	-0.852	0.448	-4.465

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-0.33	-2.262	-1.523	-0.758	-2.043	0.66	-1.859	-2.738	0.02	-2.133	-0.143	-0.384	1.611

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	2.055	-2.704	-1.776	-0.471	-0.488	0.679	4.905	-3.749	0.98	-2.49	-0.751	1.8	-2.027

5.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	0.253	3.468	-2.413	3.27	0.683	-1.509	-0.294	-0.682	-0.648	-2.21	2.707

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	-2.01	2.848	0.168	-2.001	-1.058	-0.927	-1.063	0.527	-0.563	-2.016	-0.886	-0.658	1.427

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-0.911	-0.806	1.243	-1.039	3.053	-0.205	-1.037	-0.107	-2.193	-1.681	-2.199	2.263	2.131

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	-0.239	-0.241	-1.711	0.065	-0.102	0.576	2.813	-0.128	3.4	1.69	-2.676	3.568	0.129

6.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-2.255	4.727	2.332	0.578	-0.019	1.192	-1.949	1.893	0.683	-0.637	-0.084

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	-1.15	-0.103	2.594	0.46	2.169	2.135	-0.149	-1.53	-1.073	1.396	-1.041	0.212	-0.636

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-0.674	0.862	1.208	-0.071	2.759	-0.546	2.72	0.048	-0.701	-0.034	1.564	0.644	1.301

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	1.842	1.507	1.402	2.382	-0.618	0.751	-0.496	-0.537	-0.352	-0.24	-3.298	1.066	0.195

7.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-5.65	2.492	-1.634	4.298	-2.39	3.629	2.128	3.072	-0.626	3.484	-0.011

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	3.968	-1.269	-1.432	-3.21	-4.115	2.53	0.829	2.146	0.891	0.368	-9.371	-4.943	-0.417

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-4.854	-1.304	-0.948	-2.528	-5.092	1.429	2.047	0.366	-4.127	-0.101	1.016	4.364	0.802

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	-1.595	0.583	2.488	1.578	-4.117	1.013	-1.65	-0.89	0.21	-3.219	-0.576	-0.91	1.773

8.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-5.155	-2.374	2.578	2.265	-0.774	-1.057	1.451	-6.647	1.678	-1.685	-4.161

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	3.196	-4.252	-3.777	7.613	1.521	4.667	1.827	3.559	-0.741	5.575	-7.618	1.792	-0.059

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	2.773	4.256	0.229	4.407	-4.64	7.729	5.589	1.217	1.316	-3.656	-2.756	0.582	-0.121

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	4.212	-4.789	3.045	0.194	0.964	2.82	-2.747	-0.575	1.153	-0.059	2.267	-3.96	4.343

9.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	4.848	2.896	2.26	-0.985	1.096	5.162	2.655	2.437	5.242	3.461	3.066

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	1.883	3.825	1.658	1.06	4.475	-0.217	2.094	1.834	2.429	4.285	4.333	1.378	0.577

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	1.759	0.816	4.409	3.157	4.032	5.137	2.924	4.108	6.047	-0.296	2.047	3.438	2.516

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	3.412	4.349	3.462	2.136	2.181	5.722	1.031	2.254	0.554	2.202	-1.089	5.441	0.93

10.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-0.028	-0.698	-0.033	-1.084	-0.157	-0.402	-0.077	2.043	-0.187	-0.663	1.143

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	1.57	-1.86	1.445	-1.499	-1.78	-0.719	-0.314	0.67	-0.349	-1.99	-0.128	0.341	0.837

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	1.84	-0.832	0.066	-1.528	-3.215	1.715	-0.646	-2.264	-2.064	0.173	0.519	-1.198	1.868

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	0.614	1.456	0.48	-1.622	1.092	-1.505	-0.134	-0.489	-1.258	-2.129	0.318	0.672	-0.359

11.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	-4.267	-2.319	-3.054	-1.187	-2.417	-1.588	1.494	1.209	-1.066	-3.996	0.949

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	-5.813	-0.886	-3.126	-1.994	-1.615	-0.888	0.259	1.868	-1.616	-1.624	-2.581	-1.453	-1.014

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-0.696	-2.353	-0.969	-1.113	-2.688	-0.591	-0.934	-3.471	0.142	0.428	-5.595	-2.793	-2.209

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	1.241	1.726	-2.003	-1.004	-0.658	-0.042	-0.458	-1.338	-0.895	-2.11	0.215	-2.095	-1.4

12.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	1.236	1.302	1.489	1.382	2.026	1.464	1.554	2.961	1.975	3.678	0.394

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	1.871	2.178	3.265	2.358	0.255	0.565	3.956	2.46	1.798	1.039	0.884	0.332	0.852

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	2.804	2.651	2.484	0.845	1.428	3.246	1.821	1.502	1.389	2.157	2.247	1.729	1.875

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	2.18	1.105	1.679	0.483	1.982	3.313	0.983	2.508	1.687	0.794	2.451	2.062	1.006

13.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	2.316	2.249	1.829	1.547	2.294	2.151	1.066	2.793	2.716	3.044	2.811

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	2.617	1.755	1.914	2.685	2.791	2.358	1.789	1.844	2.136	2.864	3.958	2.069	2.565

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	2.557	3.891	3.479	-0.294	2.662	1.559	1.526	2.089	1.301	2.347	1.125	3.9	-0.726

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	2.87	4.179	3.085	3.176	1.029	0.424	2.464	3.474	3.165	1.237	2.509	2.496	1.198

14.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	0.938	2.26	-0.147	1.051	2.046	-1.322	1.808	-0.549	0.959	-0.552	-0.742

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	0.924	-1.417	0.376	0.913	1.966	-0.729	0.115	0.976	0.686	0.696	1.021	1.323	1.431

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	1.397	0.76	-0.193	0.966	0.349	-0.363	2.614	-0.736	-1.194	0.886	-0.929	0.089	1.374

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	-0.069	-0.41	-0.368	1.597	-0.523	-0.752	1.125	-0.238	1.041	1.164	1.02	0.646	0.449

15.

№ наблюдения	1	2	3	4	5	6	7	8	9	10	11
Значение X	0.643	-0.713	-1.043	4.572	-0.43	-1.92	1.139	-0.589	-1.805	-1.555	-0.029

№	12	13	14	15	16	17	18	19	20	21	22	23	24
X	0.048	0.397	0.737	0.357	-1.027	-2.422	2.117	1.457	-1.734	0.216	2.546	-0.48	-3.047

№	25	26	27	28	29	30	31	32	33	34	35	36	37
X	-2.117	-1.988	-0.776	0.093	-0.894	-0.496	0.809	0.328	0.608	-1.734	-1.658	-1.655	-0.611

№	38	39	40	41	42	43	44	45	46	47	48	49	50
X	-1.97	-4.261	1.914	1.187	0.214	-1.509	-1.39	-1.317	0.417	-0.603	1.953	-2.067	0.57

ЛАБОРАТОРНАЯ РАБОТА № 2
ОСНОВНЫЕ РАСПРЕДЕЛЕНИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Цель: Исследовать основные распределения, используемые в математической статистике: нормальное распределение, распределение хи-квадрат, распределения Стьюдента и Фишера.

При статических исследованиях широко используются случайные величины, имеющие нормальное распределение, распределение χ^2 (хи-квадрат), распределения Стьюдента и Фишера.

Случайная величина X имеет нормальное распределение с параметрами m (математическое ожидание) и σ^2 (дисперсия), если плотность распределения имеет вид:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2} \right].$$

В качестве примера на рис. 1 изображены два графика плотности нормального распределения с одинаковым математическим ожиданием и разными дисперсиями.

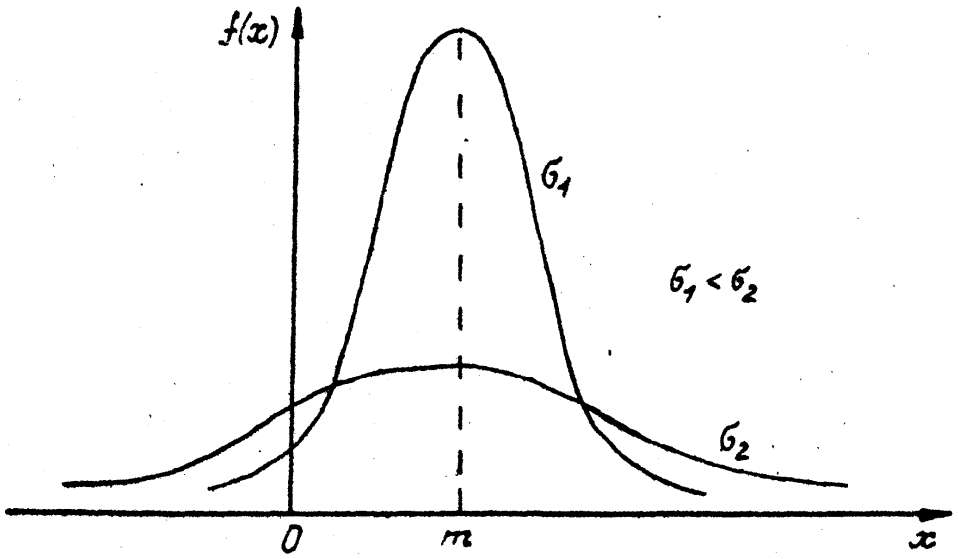

Рис. 1. Графики плотностей нормального распределения с одинаковым математическим ожиданием m и разными дисперсиями σ_1^2, σ_2^2.

Нормальное распределение с параметрами 0 и 1 называют стандартным нормальным распределением и обозначают $N(0,1)$.

Нормально распределенная случайная величина с большой вероятностью принимает значения, близкие к своему математическому ожиданию, что выражается правилом сигм:

$$P(|X-m|<k\sigma)=\begin{cases}0{,}6827, & k=1\\ 0{,}9545, & k=2\\ 0{,}9973, & k=3\end{cases}$$

Чаще всего используется правило трех сигм, т.е. $k=3$.

Построим график плотности нормального распределения и исследуем влияние параметров m (математическое распределение) и σ (среднеквадратическое отклонение) на него.

Запускаем программу Excel и задаем значения параметров m и σ .

Для этого в ячейки первой строки первого столбца (A1) и второй строки первого столбца (A2) вводим подписи **m=** и **sig=**, а в первую строку второго столбца (B1) и во вторую строку второго столбца (B2) вводим их числовые значения. Для построения графика протабулируем в третьем и четвертом столбцах (соответственно C и D) функцию плотности нормального распределения на интервале (a,b) с некоторым выбранным шагом h (интервал (a,b) лучше выбрать так, чтобы его серединой было значение m). Для этого вводим в C1 надпись **X=**, а в D1 надпись **f=**. Вводим в C2 значение a, а в C3 значение $a+h$. После этого обводим, выделяя, ячейки C2 и C3 и захватив за нижний правый угол рамки вокруг ячеек C2 и C3, перетягиваем его вниз до ячейки C($2+\frac{b-a}{h}$), что позволит автоматически занести в столбец значения от a до b с шагом h. Ставим курсор в ячейку D2 и вызываем функцию плотности нормального распределения. Для этого нажимаем кнопку мастера функций *fx* и выбираем категорию **СТАТИСТИЧЕСКИЕ** и функцию **НОРМРАСП**. Вводим ссылкой на переменную X: «C2» (для ввода ссылки достаточно щелкнуть мышью по ячейке с данной адресацией), ссылкой на m и σ - «B1» и «B2». Эти ссылки абсолютные, так как ячейки со значениями m и σ всегда B1 и B2, поэтому пишется знак $ (чтобы быстро относительную ссылку сделать абсолютной нужно после ввода ссылки нажать F4). В поле «Интегральное» ставим 0 или «**ложь»**, нажимаем «ОК». В ячейке D2 появляется результат $f(a)$– значение плотности нормального распределения, а в строке формул –
запись **=НОРМРАСП(C2;B1;B2;ложь)**. За нижний правый угол ячейки D2 автозаполняем результат на ячейки D2-D($2+\frac{b-a}{h}$). Если требуется построить функцию распределения вероятностей нормального

распределения, то в поле «Интегральное» нужно поставить *a* или «**истина**».

Строим график плотности нормального распределения по данным. Ставим курсор в любой свободной ячейке. При использовании EXCEL 2003 вызываем **МАСТЕР ДИАГРАММ**, выбрав пункты меню **ВСТАВКА/ДИАГРАММА**. Выбираем тип диаграммы **ГРАФИК** и его вид– левый график в верхнем ряду. Нажимаем ДАЛЕЕ. Ставим курсор в поле ДИАПОЗОН и обводим курсором ячейки D2-D($2+\frac{b-a}{h}$). Далее переходим на закладку РЯД, ставим курсор в поле ПОДПИСИ ОСИ Х и обводим диапазон данных C2-C($2+\frac{b-a}{h}$). Нажимаем ГОТОВО. Получаем график плотности нормального распределения.

При использовании EXCEL 2007 установите курсор на ячейку, где хотите расположить график и вверху в меню переключитесь на вкладку **ВСТАВКА**. Затем нажмите на кнопку **ГРАФИК**, выпадет несколько их видов. Выбрать можно любой, например, первый – классический график. На листе появится новый объект - чистый график. Когда он выделен, то верхняя панель с иконками действий имеет другой вид, специально для работы с графиками. Чтобы заполнить график, нажмите на кнопку **ВЫБРАТЬ ДАННЫЕ**. Отобразится окно выбора данных для графика. В нем имеется поле **ВЫБОР ДАННЫХ ДЛЯ ДИАГРАММЫ**. В конце поля необходимо нажать на кнопку выбора диапазона. Далее следует выделить мышкой таблицу с данными и подписями строк и столбцов и снова кликнуть на кнопку выбора диапазона данных. Нажимаем **«Ок»**. График построен.

Следующим шагом является исследование влияния параметров m и σ на вид графика. Для этого увеличиваем значение математического ожидания m в ячейке B1 и нажимаем «Enter». График плотности

нормального распределения должен сместиться вправо. Теперь уменьшаем значение математического ожидания m в ячейке B1 и нажимаем «Enter». График плотности нормального распределения должен сместиться влево. Возвращаем в B1 первоначальное значение m и начинаем изучать влияние среднеквадратического отклонения σ (или дисперсии σ^2) на график плотности. Увеличивая значение σ в ячейке B2, можно наблюдать растяжение графика. Уменьшая значение σ, можно наблюдать сжатие графика.

После такого исследования можно сделать соответствующие выводы о влиянии параметров m и σ на вид графика плотности нормального распределения.

Пусть $X_1, X_2, ..., X_n$ - независимые случайные величины с общим распределением $N(0,1)$ (нормальное распределение с нулевым математическим ожиданием и единичной дисперсией). Тогда случайная величина

$$\chi^2 = \sum_{i=1}^{n} X_i^2$$

имеет распределение χ^2 с n степенями свободы или распределение χ_n^2.

Плотность распределения вероятностей χ_n^2 имеет вид:

$$f_{\chi_n^2}(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}} \cdot \Gamma\left(\dfrac{n}{2}\right)} \cdot x^{\frac{n}{2}-1} \cdot e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \le 0 \end{cases},$$

где

$$\Gamma(n) = \int_0^{\infty} x^{n-1} e^{-x} dx .$$

Графики плотности распределения χ^2 с n степенями свободы асимметричны и, начиная с $n=2$, имеют по одному максимуму в точке $x=n-2$ (рис. 2). Причем с ростом n кривая плотности приближается к симметричной функции.

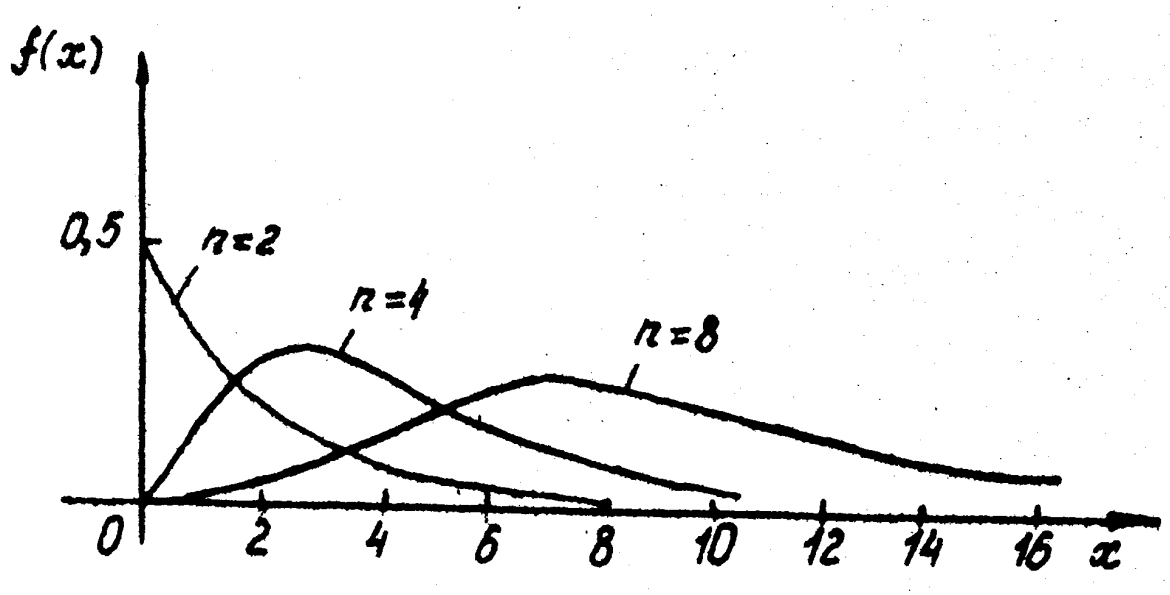

Рис.2. Графики плотности распределения χ^2 с n степенями свободы.

Интервал для построения необходимо выбрать с нулевой левой границей и произвести его разбиение. Пусть точки разбиения занимают, например, ячейки B1-B15. Для того, чтобы найти значение функции распределения вероятностей χ^2 с n степенями свободы, нужно нажать кнопку мастера функций *fx* , выбрать категорию **СТАТИСТИЧЕСКИЕ** и функцию **ХИ2РАСП**. Вызванная функция находит значения вероятностей $P(X>x)$, которые заполняют, например, ячейки C1-C15. Поскольку нам нужна функция распределения вероятностей $F(x)=P(X\le x)=1-P(X>x)$, то ее значения мы получим в ячейках D1-D15 следующим образом: D1=1-C1,…,D15=1-C15. Для определения значений плотности распределения χ^2 с n степенями свободы необходимо воспользоваться формулой для

приближенного вычисления первой производной от функции распределения. Если шаг разбиения равен h, а функция распределения вероятностей в соседних точках разбиения имеет значения соответственно $g((i-1)h)$ и $g(ih)$, то значение плотности распределения вероятностей в точке разбиения ih равна $\frac{g(ih)-g((i-1)h)}{h}$. Значения плотности распределения вероятностей получаем в ячейках, например, начиная с E2 по формуле E2= (D2-D1)/h. Остальные ячейки автозаполняем. Таким образом, можно найти значения плотности χ^2 с n степенями свободы на выбранном отрезке, кроме нуля (левой границы отрезка).

Действуя аналогично построению графика плотности нормального распределения, строим график плотности распределения χ^2 с n степенями свободы и изучаем влияние степени свободы на вид графика, уменьшая или увеличивая n.

После исследования делаем соответствующие выводы.

Пусть случайная величина Y имеет распределение $N(0,1)$, а независимая от Y случайная величина Z принадлежит χ^2_n (имеет распределение χ^2 с n степенями свободы). Тогда случайная величина

$$X=\frac{Y}{\sqrt{\frac{Z}{n}}}$$

имеет распределение Стьюдента с n степенями свободы (t_n-распределение) и плотностью распределения вероятностей:

$$f_{t_n}(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ x\in(-\infty,+\infty).$$

Графики плотности случайной величины, имеющей распределение Стьюдента, при любом $n = 1,2,....$ симметричны относительно оси ординат (рис.3), поэтому при любом $n = 1,2,.....$ математическое ожидание равно нулю.

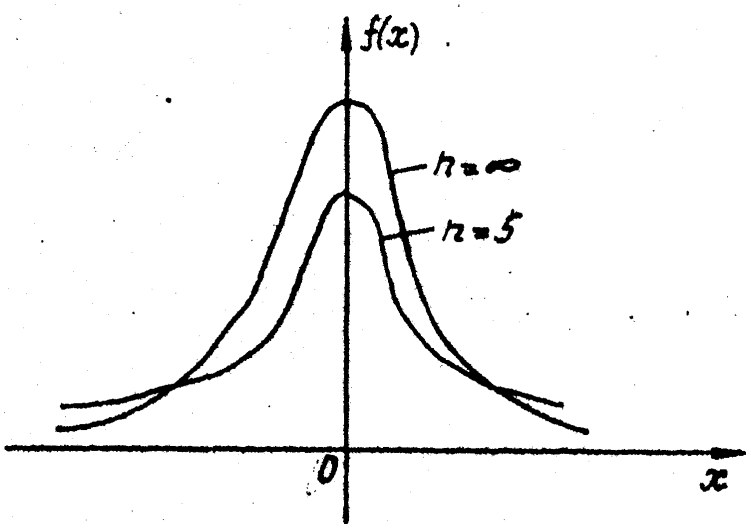

Рис.3. Графики плотности распределения Стьюдента.

С ростом n распределение Стьюдента приближается к $N(0,1)$.

Интервал для построения функции распределения Стьюдента t_n необходимо выбрать симметричным относительно нуля. Разбиение построить так, чтобы среди точек разбиения был нуль. Пусть точки разбиения занимают, например, ячейки B1-B15, среди них ячейки B1-B7 содержат отрицательные значения, ячейка B8 содержит нуль и ячейки B9-B15 содержат положительные значения.

Для того, чтобы найти значения функции распределения Стьюдента t_n с n степенями свободы, нужно нажать кнопку мастера функций fx, выбрать категорию **СТАТИСТИЧЕСКИЕ** и функцию **СТЬЮДРАСП**. Функция имеет дополнительный чисто вычислительный параметр «Хвосты», который не связан с распределением Стьюдента, а связан с

выводом полученных результатов программой EXCEL. Его всегда задаем равным 1. В этом случае вызванная функция находит значения вероятностей $P(X > x)$ при неотрицательном x, которые заполняют, например, соответственно ячейки C9-C15. В ячейке C9 обязательно должно быть значение 0.5, поскольку это вероятность $P(X > 0) = 0.5$. Поскольку нам нужна функция распределения вероятностей $F(x) = P(X \le x) = 1 - P(X > x)$, то ее значения мы получим в ячейках D9-D15 следующим образом: D9=1-C9,…,D15=1-C15. Значения функции распределения в ячейках D1-D7 получим по формуле: D1=C15, D2=C14,….D7=C9, поскольку функция распределения Стьюдента из-за симметричности распределения обладает свойством $P(X \le -x) = P(X > x)$.

Для нахождения значений плотности распределения Стьюдента t_n необходимо воспользоваться формулой для приближенного вычисления значений первой производной (по аналогии с распределением χ^2).

Далее строим плотность распределения вероятностей t_n с n степенями свободы (значение берется произвольно) и изучаем влияние степени свободы на вид графика, уменьшая или увеличивая значение n.

После исследования делаем соответствующие выводы.

Пусть U и V - независимые случайные величины, распределенные по закону χ^2 с n_1 и n_2 степенями свободы соответственно. Тогда случайная величина

$$X = \frac{\dfrac{U}{n_1}}{\dfrac{V}{n_2}}$$

имеет распределение Фишера с n_1 и n_2 степенями свободы (F_{n_1,n_2}-распределение) и плотностью распределения вероятностей

$$f(x)=\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\cdot n_1^{\frac{n_1}{2}}\cdot n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right)\cdot\Gamma\left(\dfrac{n_2}{2}\right)}\cdot x^{\frac{n_1}{2}-1}\cdot\left(n_2+n_1x\right)^{-\frac{n_1+n_2}{2}}, x>0\\ 0, x\le 0\end{cases}.$$

Графики плотности распределения случайной величины асимметричны, имеют длинные "хвосты" и достигают максимума вблизи точки $x=1$ (рис.4).

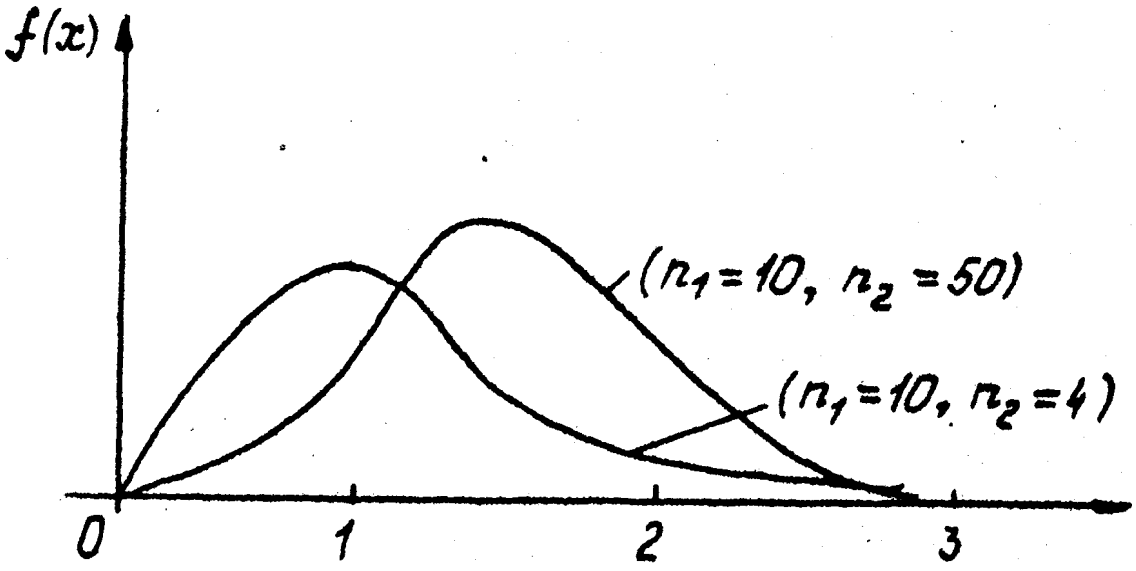

Рис.4. Графики плотности распределения Фишера.

Интервал для построения необходимо выбрать с нулевой левой границей. Пусть точки разбиения занимают, например, ячейки B1-B15. Для того, чтобы найти значения функции распределения Фишера F_{n_1,n_2} с n_1 и n_2 степенями свободы, нужно нажать кнопку мастера функций *fx* , выбрать категорию **СТАТИСТИЧЕСКИЕ** и функцию **FРАСП**. Вызванная функция находит значения вероятностей $P(X>x)$, которые заполняют, например, ячейки C1-C15. Поскольку нам нужна функция распределения

вероятностей $F(x)=P(X\le x)=1-P(X>x)$, то ее значения мы получим в ячейках D1-D15 следующим образом: D1=1-C1,…,D15=1-C15.

Для определения значений плотности распределения вероятностей необходимо воспользоваться формулой для приближенного вычисления значений первой производной.

Далее строим график плотности распределения F_{n_1,n_2} и изучаем влияние степеней свободы n_1 и n_2 на вид графика. Для этого фиксируем значение n_1 и уменьшаем или увеличиваем значение n_2. После этого фиксируем значение n_2 и уменьшаем или увеличиваем значение n_1.

После исследования делаем соответствующие выводы.

ЛАБОРАТОРНАЯ РАБОТА № 3
ОПРЕДЕЛЕНИЕ ТОЧЕЧНЫХ ОЦЕНОК НЕИЗВЕСТНЫХ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

Цель: Научиться определять точечные (числовые) оценки для неизвестных параметров распределений.

Как известно, статистикой называется функция от результатов наблюдений (функция от выборки). Статистики используются для построения точечных оценок неизвестных параметров различных законов распределений.

Рассмотрим задачу определения точечных оценок для параметра равномерного распределения.

Пусть выборка $x_1,x_2,...,x_n$ принадлежит равномерному на $[0,\theta]$ распределению. Найдем оценку для неизвестного параметра θ двумя способами $\theta=x_{(n)}$ и $\theta=2\bar{x}$.

Вводим в первый столбец (ячейки A1…) исходные данные. Для элементов выборки находим максимальный элемент, которые ограничивает интервал, содержащий все элементы выборки. Для этого запишем в первую строку второго столбца (B1) слово Максимум, а в соседней ячейке C1 определим функцию **MAX**. Для этого ставим курсор в C1 и вызываем мастер функций, нажав на кнопку *fx* , в открывшемся окне в поле **КАТЕГОРИЯ** выбираем **СТАТИСТИЧЕСКИЕ**, и ниже ищем функцию **МАКС** и вызываем ее двойным щелчком по названию. В качестве аргумента функции (в графе «Число 1») обведем область данных (ячейки A1…). Поле «Число 2» оставляем пустым. Нажимаем «ОК».

Получаем оценку $\theta = x_{(n)}$.

Функция **СРЗНАЧ** вычисляет выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Умножая найденное значение на 2, получаем оценку $\theta = 2\overline{x}$.

Рассматривается выборка из сдвинутого экспоненциального распределения с плотностью $f(x-\theta)$, $\theta > 0$. В качестве оценок неизвестного параметра θ предлагается $x_{(1)}$ и $\overline{x} - 1$.

Вводим в первый столбец (ячейки A1…) исходные данные. Для элементов выборки находим минимальный элемент, который ограничивает интервал, содержащий все элементы выборки. Для этого запишем в первую строку второго столбца (B1) слово Минимум. В соседней ячейке C1 определим функцию **MIN**. Для этого ставим курсор в C1 и вызываем мастер функций, нажав на кнопку *fx* , в открывшемся окне в поле «Категория» выбираем «Статистические», и ниже ищем функцию **МИН** и вызываем ее двойным щелчком по названию. В качестве аргумента функции (в графе «Число 1») обведем область данных (ячейки A1…). Поле «Число 2» оставляем пустым. Нажимаем «ОК».

Получаем оценку $\theta = x_{(1)}$.

Функция **СРЗНАЧ** вычисляет выборочное среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Вычитая из найденного значения единицу, получаем оценку $\theta = \overline{x} - 1$.

Задания для нахождения оценок равномерного распределения

1	**0.15, 0.39, 0.18, 0.06, 0.98, 0.32, 0.88, 1.14, 1.34, 0.10, 0.18, 0.56, 1.24, 1.76, 0.92, 1.56, 0.22, 1.32, 1.19, 0.29, 0.32, 0.34, 0.10, 0.56, 0.56, 0.22, 0.27, 1.32, 1.34, 1.10, 0.56, 0.15, 0.39, 1.18, 0.06, 0.34, 1.10, 0.18, 0.32, 0.38, 1.10, 0.56, 0.29, 0.76, 0.99**
2	**0.38, 0.08, 0.19, 0.87, 0.99, 0.65, 0.04, 0.24, 0.98, 0.29, 0.32, 0.34, 0.17, 0.36, 0.47, 0.76, 0.99, 0.56, 0.08, 0.19, 0.29, 0.32, 0.38, 0.10, 0.56, 0.29, 0.76, 0.99, 0.56, 0.22, 0.27, 0.32, 0.34, 0.17, 0.46, 0.24, 0.31, 0.84, 0.02, 0.24, 0.98, 0.29, 0.32, 0.34, 1.17, 0.56, 1.19, 0.29, 1.32, 1.34, 1.10, 0.56**
3	**0.92, 0.41, 0.43, 0.22, 0.29, 0.32, 0.34, 0.10, 0.56, 0.32, 0.38, 1.10, 0.56, 1.29, 0.76, 0.99, 0.56, 0.22, 0.27, 0.32, 0.34, 1.10, 0.56, 0.24, 0.98, 0.29, 0.32, 0.34, 1.17, 0.56, 1.19, 0.29, 0.32, 0.34, 1.10, 0.56, 0.24, 0.76, 0.42, 1.02, 0.24, 0.98, 1.29, 0.32, 0.34, 1.17, 0.56 1.24, 0.76**
4	**0.18, 0.06, 0.98, 0.32, 0.88, 0.12, 0.47, 0.92, 1.18, 1.15, 1.39, 1.18, 1.06, 1.19, 0.29, 1.32, 1.34, 1.10, 1.56, 1.32, 1.34, 1.10, 0.56, 1.24, 0.79, 0.92, 0.56, 1.24, 0.76, 0.92, 0.56, 0.22, 1.32, 0.51, 0.99, 0.87, 1.04, 0.17, 1.12, 0.38, 1.04, 1.17, 1.12, 1.38, 1.47, 0.92, 1.18, 1.15, 1.39**

5	**0.32, 0.34, 1.10, 0.56, 0.24, 0.79, 0.92, 0.47, 0.92, 0.18, 0.15, 0.39, 1.18, 0.06, 0.98, 0.32, 0.88, 0.12, 0.51, 0.99, 0.87, 1.04, 0.17, 1.12, 1.38, 1.19, 0.29, 1.32, 0.34, 1.10, 0.56, 1.08, 1.19, 0.29, 0.32, 1.34, 1.10, 0.56, 1.24, 0.76, 0.99, 0.56, 0.08, 1.19, 0.29, 0.32, 0.34, 1.10, 0.56**
6	**0.10, 0.56, 0.29, 0.76, 0.99, 0.56, 0.22, 0.27, 0.32, 0.34, 0.10, 0.56, 0.24, 0.76, 0.11, 0.44, 0.32, 0.38, 0.10, 0.56, 0.29, 0.76, 0.99, 0.56, 0.22, 0.67, 0.32, 0.19, 0.29, 0.32, 0.34, 0.10, 0.56, 0.34, 0.10, 0.56, 0.32, 0.34, 0.10, 0.56, 0.24, 0.76, 0.99, 0.56, 0.08, 0.19, 0.29, 0.32, 0.38, 0.56, 0.22, 0.67**
7	**0.15, 0.39, 0.18, 0.06, 0.68, 0.37, 0.81, 0.92, 0.17, 0.22, 0.32, 0.51, 0.99, 0.87, 0.94, 0.17, 0.12, 0.38, 0.04, 0.17, 0.12, 0.34, 0.37, 0.10, 0.18, 0.15, 0.39, 0.18, 0.26, 0.56, 0.22, 0.27, 0.32, 0.34, 0.10, 0.88, 0.21, 0.34, 0.15, 0.56, 0.24, 0.79, 0.92, 0.47, 0.92, 0.18, 0.19, 0.29, 0.32, 0.34, 0.10, 0.49**
8	**0.77, 0.29, 0.35, 0.44, 0.10, 0.56, 0.24, 0.76, 0.92, 0.68, 0.24, 0.76, 0.99, 0.56, 0.68, 0.19, 0.29, 0.38, 0.08, 0.19, 0.87, 0.99, 0.87, 0.04, 0.24, 0.98, 0.29, 0.32, 0.34, 0.17, 0.56, 0.24, 0.76, 0.46, 0.19, 0.29, 0.32, 0.34, 0.10, 0.56, 0.02, 0.24, 0.98, 0.29, 0.32, 0.38, 0.10, 0.56, 0.29, 0.76, 0.99, 0.59**
9	**0.22, 1.32, 1.19, 0.29, 0.32, 0.34, 0.10, 0.56, 0.56, 0.22, 0.27, 1.32, 1.34, 1.10, 0.56, 0.15, 0.39, 1.18, 0.06, 0.34, 1.10, 0.18, 0.15, 0.39, 0.18, 0.06, 0.98, 0.32, 0.88, 1.14, 1.34, 0.10, 0.18, 0.56, 1.24, 1.76, 0.92, 1.56, 0.32, 0.38, 1.10, 0.56, 0.29, 0.76, 0.99**
10	**0.17, 0.46, 0.24, 0.31, 0.84, 0.02, 0.24, 0.98, 0.29, 0.32, 0.34, 1.17, 0.56, 1.19, 0.29, 1.32, 1.34, 1.10, 0.560.38, 0.08, 0.19, 0.87, 0.99, 0.65, 0.04, 0.24, 0.98, 0.29, 0.32, 0.34, 0.17, 0.36, 0.47, 0.76,**

	0.99, 0.56, 0.08, 0.19, 0.29, 0.32, 0.38, 0.10, 0.56, 0.29, 0.76, 0.99, 0.56, 0.22, 0.27, 0.32, 0.34
11	**1.17, 0.56, 1.19, 0.29, 0.32, 0.34, 1.10, 0.56, 0.24, 0.76, 0.42, 1.02, 0.24, 0.98, 1.29, 0.32, 0.34, 1.17, 0.56 1.24, 0.76, 0.92, 0.41, 0.43, 0.22, 0.29, 0.32, 0.34, 0.10, 0.56, 0.32, 0.38, 1.10, 0.56, 1.29, 0.76, 0.99, 0.56, 0.22, 0.27, 0.32, 0.34, 1.10, 0.56, 0.24, 0.98, 0.29, 0.32, 0.34**
12	**0.79, 0.92, 0.56, 1.24, 0.76, 0.92, 0.56, 0.22, 1.32, 0.51, 0.99, 0.87, 1.04, 0.17, 1.12, 0.38, 1.04, 1.17, 1.12, 1.38, 1.47, 0.92, 1.18, 1.15, 1.390.18, 0.06, 0.98, 0.32, 0.88, 0.12, 0.47, 0.92, 1.18, 1.15, 1.39, 1.18, 1.06, 1.19, 0.29, 1.32, 1.34, 1.10, 1.56, 1.32, 1.34, 1.10, 0.56, 1.24**
13	**0.87, 1.04, 0.17, 1.12, 1.38, 1.19, 0.29, 1.32, 0.34, 1.10, 0.56, 1.08, 1.19, 0.29, 0.32, 1.34, 1.10, 0.56, 1.24, 0.76, 0.99, 0.56, 0.08, 1.19, 0.29, 0.32, 0.34, 1.10, 0.56, 0.32, 0.34, 1.10, 0.56, 0.24, 0.79, 0.92, 0.47, 0.92, 0.18, 0.15, 0.39, 1.18, 0.06, 0.98, 0.32, 0.88, 0.12, 0.51, 0.99**
14	**0.37, 0.10, 0.18, 0.15, 0.39, 0.18, 0.26, 0.56, 0.22, 0.27, 0.32, 0.34, 0.10, 0.88, 0.21, 0.34, 0.15, 0.56, 0.24, 0.79, 0.92, 0.47, 0.92, 0.18, 0.19, 0.29, 0.32, 0.34, 0.10, 0.49, 0.15, 0.39, 0.18, 0.06, 0.68, 0.37, 0.81, 0.92, 0.17, 0.22, 0.32, 0.51, 0.99, 0.87, 0.94, 0.17, 0.12, 0.38, 0.04, 0.17, 0.12, 0.34**
15	**0.68, 0.19, 0.29, 0.38, 0.08, 0.19, 0.87, 0.99, 0.87, 0.04, 0.77, 0.29, 0.35, 0.44, 0.10, 0.56, 0.24, 0.76, 0.92, 0.68, 0.24, 0.76, 0.99, 0.56, 0.24, 0.98, 0.29, 0.32, 0.34, 0.17, 0.56, 0.24, 0.76, 0.46, 0.19, 0.29, 0.32, 0.34, 0.10, 0.56, 0.02, 0.24, 0.98, 0.29, 0.32, 0.38, 0.10, 0.56, 0.29, 0.76, 0.99, 0.59**

Задания для нахождения оценок сдвинутого экспоненциального распределения

1	**2.56, 6.24, 4.76, 5.92, 3.56, 0.22, 1.32, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 2.34, 1.10, 0.18, 1.15, 1.39, 1.18, 2.06, 6.98, 4.32, 5.88, 3.14, 2.34, 1.10, 0.18, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.561.15, 1.39, 1.18, 2.06**
2	**5.32, 2.34, 1.10, 2.56, 6.24, 4.79, 5.92, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 4.32, 5.88, 3.12, 2.51, 6.99, 4.87, 5.04, 3.17, 5.12, 3.38, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 2.08, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 6.24, 8.76, 5.99, 7.56, 2.08, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56**
3	**1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 6.24, 8.76, 5.76, 7.02, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56 6.24, 4.76, 5.92,12.56, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56**
4	**1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 5.32, 2.34, 1.10, 2.56, 6.24, 4.79, 5.92, 2.56, 6.24, 4.76, 5.92, 3.56, 0.22, 1.32, 2.51, 6.99, 4.87, 5.04, 3.17, 5.12, 3.38, 5.04, 3.17, 5.12, 3.38, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 4.32, 5.88, 3.12, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06**
5	**3.38, 2.08, 1.19, 2.87, 6.99, 4.87, 5.04, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 6.24, 8.76, 5.99, 7.56, 2.08, 1.19, 0.29, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56,6.24, 8.76, 5.76, 7.02, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.561.19, 0.29, 9.32, 2.34, 1.10, 8.56**

6	**4.32, 2.38, 1.10, 2.56, 6.29, 4.76, 7.99, 4.56, 1.22, 3.67, 8.32, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 2.34, 1.10, 9.56, 6.24, 8.76, 5.99, 7.56, 2.08, 1.19, 0.29, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.11, 2.44, 4.56, 1.22, 3.67, 8.32, 2.34, 1.10, 9.56**
7	**2.34, 1.10, 0.18, 1.15, 1.39, 1.18, 2.06, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 5.32, 2.34, 1.10, 2.56, 6.24, 4.79, 5.92, 3.47, 0.92, 7.18, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 3.15, 1.39, 1.18, 2.06, 2.56, 6.24, 4.76, 5.92, 3.56, 0.22, 1.32, 2.51, 6.99, 4.87, 5.04, 3.17, 5.12, 3.38, 5.04, 3.17, 5.12, 3.38**
8	**6.24, 8.76, 5.99, 7.56, 2.08, 1.19, 0.29, 3.38, 2.08, 1.19, 2.87, 6.99, 4.87, 5.04, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 6.24, 8.76, 5.76, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 7.02, 3.24, 5.98, 3.29, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.92, 12.56**
9	**7.56, 0.22, 9.27, 8.32, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 2.34, 1.10, 9.56, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 5.32, 2.34, 1.10, 2.56, 6.24, 4.79, 5.92, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 4.32, 5.88, 3.12**
10	**6.24, 8.76, 5.76, 7.02, 3.24, 5.98, 3.29, 6.24, 8.76, 5.99, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 7.56, 2.08, 1.19, 0.29, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 3.38, 2.08, 1.19, 2.87, 6.99, 4.87, 5.04, 3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 6.24, 8.76, 5.76, 7.02, 3.24, 5.98, 3.29**
11	**3.24, 5.98, 3.29, 8.32, 2.34, 1.17, 8.56, 3.29, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 1.19, 0.29, 9.32, 2.34,**

	1.10, 8.56, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.92, 12.56, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.92, 12.56
12	**6.24, 8.76, 5.99, 7.56, 2.08, 1.19, 0.29, 2.34, 1.10, 0.18, 1.15, 1.39, 1.18, 2.06, 2.56, 6.24, 4.76, 5.92, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 3.56, 0.22, 1.32, 2.51, 6.99, 4.87, 5.04, 3.17, 5.12, 3.38, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.92, 12.56**
13	**6.24, 8.76, 5.76, 7.02, 3.24, 5.98, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 4.32, 5.88, 3.12, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 3.291.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 9.56, 6.24, 4.76, 5.92, 12.56**
14	**2.34, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 1.10, 0.18, 1.15, 1.39, 1.18, 2.06, 6.98, 4.32, 5.88, 3.14, 2.34, 1.10, 0.18, 1.32, 2.38, 1.10, 2.56, 6.29, 4.76, 5.99, 7.56, 0.22, 9.27, 8.32, 2.34, 1.10, 2.56, 6.24, 4.76, 5.92, 3.56, 0.22, 1.32, 2.51, 6.99, 4.87, 5.04, 3.17, 5.12, 3.38, 5.04, 3.17, 5.12, 3.38**
15	**4.56, 1.22, 3.67, 1.19, 0.29, 9.32, 2.34, 1.10, 8.56, 8.32, 2.34, 1.10, 9.56, 5.32, 2.34, 1.10, 2.56, 6.24, 4.79, 5.92, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 4.32, 5.88, 3.12, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06, 11.98, 4.32, 5.88, 3.12, 3.47, 0.92, 7.18, 3.15, 1.39, 1.18, 2.06**

ЛАБОРАТОРНАЯ РАБОТА № 4
ОПРЕДЕЛЕНИЕ ИНТЕРВАЛЬНЫХ ОЦЕНОК НЕИЗВЕСТНЫХ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

Цель: Научиться определять интервальные оценки с заданным уровнем доверия (доверительные интервалы) для неизвестных параметров распределений.

В ряде задач требуется не только найти для параметра θ подходящую оценку, но и указать, к каким ошибкам может привести замена параметра его оценкой. Другими словами, требуется оценить точность и надежность оценки. Такого рода задачи особенно актуальны при малом числе наблюдений, когда точечная оценка в значительной мере случайна и приближенная замена этой оценкой истинного значения параметра θ может привести к серьезным ошибкам.

Для определения ***точности*** оценки θ в математической статистике пользуются ***доверительными интервалами***, а для определения ***надежности - доверительными вероятностями.***

Интервал (θ_1, θ_2) называется α***-доверительным интервалом*** или ***доверительным интервалом с доверительной вероятностью*** $1-\alpha$, если

$$P(\theta_1 \le x \le \theta_2) = 1-\alpha.$$

Рассмотрим задачу построения доверительных интервалов для неизвестных математического ожидания и дисперсии. Как известно, α-доверительный интервал для математического ожидания a выглядит следующим образом:

$$\bar{x} - \frac{\Delta_{n-1,\alpha} s_1}{\sqrt{n}} \le a \le \bar{x} + \frac{\Delta_{n-1,\alpha} s_1}{\sqrt{n}},$$

где s_1 -среднеквадратическое отклонение (корень квадратный из исправленной выборочной дисперсии), $\Delta_{n-1,\alpha}$ находится из таблицы для вероятностей $P\left(\left|t_{n-1}\right| > \Delta_{n-1,\alpha}\right) = \alpha$ распределения t_{n-1} (распределение Стьюдента с $n-1$ степенью свободы).

Вводим в первый столбец, например, ячейки A1…A25 исходные данные. Задаем уровень значимости $\alpha = 0.05$. Далее для получения результатов подписываем ячейки, как на рис. 5.

	F	G	H	I
1		Уровень значимости		0,05
2		Интервал	Левая граница	Правая граница
3		Матожидание		
4		Дисперсия		

Рис. 5. Пример подписи ячеек для получения результатов.

Для вычисления величины $\frac{\Delta_{n-1,\alpha} s_1}{\sqrt{n}}$ служит функция **ДОВЕРИТ** категории **СТАТИСТИЧЕСКИЕ** с тремя параметрами «Альфа» - уровень значимости α, «Станд. откл» - s_1 -среднеквадратическое отклонение (корень квадратный из исправленной выборочной дисперсии), «Размер» - объем выборки n.

Таким образом, вводим в ячейку H3 функцию:

=СРЗНАЧ(A1:A25)-ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)

а в ячейку I3 функцию:

=СРЗНАЧ(A1:A25)+ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25).

Как известно, α-доверительный интервал для дисперсии σ^2 выглядит следующим образом:

$$\partial^{-1}_{n-1,\frac{\alpha}{2}}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \le \sigma^2 \le \partial^{-1}_{n-1,1-\frac{\alpha}{2}}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 .$$

Нам потребуется функция **ХИ2ОБР** (категория **СТАТИСТИЧЕСКИЕ**), которая вычисляет обратное значение $x_{n-1,p}$ односторонней вероятности распределения хи-квадрат $P\left(\chi^2_{n-1} > x_{n-1,p}\right) = p$.

В данном конкретном случае $P\left(\chi^2_{n-1} > \partial_{n-1,\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$, $P\left(\chi^2_{n-1} > \partial_{n-1,1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$.

ХИ2ОБР имеет два параметра: первый «Вероятность» содержит доверительную вероятность соответственно $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$, второй – степень свободы *n*-1.

Вводим в соответствии с формулой для доверительного интервала σ^2 в ячейку H4 запись:

=ДИСП(A1:A25)*24/ХИ2ОБР(0,025;24),

а в ячейку I4 запись:

=ДИСП(A1:A25)*24/ХИ2ОБР(0,975;24).

Получаем значения границ доверительных интервалов для σ^2.

Задания

Станок производит детали, измерения которых приведено ниже. С доверительной вероятностью 0.95 построить доверительные интервалы для математического ожидания и дисперсии размера деталей.

1.

43.8	**43.9**	**46.3**	**44.6**	**47.5**	**42.0**	**44.5**	**45.0**	**46.8**	**45.3**	**41.8**	**42.3**	**47.9**	**45.5**	**44.4**	**43.1**	**42.8**
41.9	**42.8**	**46.0**	**45.3**	**41.8**	**42.3**	**47.9**	**45.5**	**46.3**	**44.6**	**47.5**	**42.0**	**44.5**	**43.8**	**43.9**	**46.3**	**44.6**

2.

49.0	**48.8**	**49.2**	**50.2**	**49.5**	**49.8**	**49.9**	**49.3**	**49.6**	**49.5**	**49.7**	**49.0**	**48.8**	**51.8**	**49.1**	**48.3**	**50.0**
49.0	**48.4**	**48.5**	**49.6**	**49.5**	**49.7**	**49.0**	**48.8**	**51.8**	**49.5**	**49.8**	**49.9**	**49.3**	**49.6**	**48.8**	**49.2**	**50.2**

3.

42.1	**41.9**	**42.3**	**43.1**	**42.5**	**42.7**	**42.9**	**42.3**	**42.6**	**42.5**	**42.7**	**42.0**	**41.9**	**44.5**	**42.2**	**41.5**	**42.9**
42.1	**41.5**	**41.6**	**41.9**	**42.3**	**43.1**	**42.5**	**42.7**	**42.9**	**42.3**	**42.6**	**42.5**	**42.7**	**42.0**	**42.3**	**43.1**	**42.5**

4.

23.0	**23.9**	**22.2**	**23.0**	**23.7**	**21.4**	**23.6**	**21.9**	**23.0**	**21.9**	**21.8**	**23**	**21.3**	**22.6**	**22.9**	**23.7**	**21.8**
22.4	**23.0**	**22.8**	**21.4**	**23.6**	**21.9**	**23.0**	**21.9**	**21.8**	**23**	**23.0**	**23.9**	**22.2**	**23.0**	**23.7**	**21.4**	**23.6**

5.

25.6	**25.5**	**25.5**	**24.7**	**25.0**	**25.8**	**25.2**	**25.0**	**25.0**	**25.3**	**25.4**	**25.1**	**25.2**	**25.3**	**24.8**	**25.1**	**24.6**
25.2	**25.8**	**24.8**	**25.5**	**24.7**	**25.0**	**25.8**	**25.2**	**25.0**	**25.0**	**25.3**	**25.4**	**25.1**	**25.2**	**25.0**	**25.0**	**25.3**

6.

15.3	**15.9**	**14.9**	**15.4**	**15.8**	**14.4**	**15.7**	**14.7**	**15.3**	**14.7**	**14.7**	**15.3**	**14.4**	**15.1**	**15.3**	**15.7**	**14.7**
15.0	**15.3**	**15.2**	**15.9**	**14.9**	**15.4**	**15.8**	**14.4**	**15.7**	**14.7**	**15.3**	**14.7**	**14.7**	**15.3**	**14.4**	**15.3**	**15.2**

7.

51.2	**49.9**	**50.7**	**52.3**	**51.3**	**51.2**	**52.4**	**51.6**	**51.5**	**51.0**	**51.8**	**50.9**	**50.7**	**52.0**	**50.2**	**51.1**	**51.0**
51.2	**52.0**	**49.9**	**50.7**	**52.3**	**51.3**	**51.2**	**52.4**	**51.6**	**51.5**	**51.0**	**51.8**	**50.9**	**50.7**	**52.4**	**51.6**	**51.5**

8.

21.8	**21.6**	**21.7**	**20.8**	**20.8**	**20.8**	**21.2**	**20.7**	**19.8**	**22.2**	**21.9**	**21.5**	**20.9**	**20.9**	**20.9**	**21.6**	**21.2**
22.2	**20.6**	**21.7**	**20.8**	**20.8**	**20.8**	**21.2**	**20.7**	**19.8**	**22.2**	**21.9**	**21.5**	**20.9**	**20.9**	**20.9**	**21.8**	**21.6**

9.

22.7	**25.7**	**25.5**	**25.2**	**22.7**	**23.6**	**26.4**	**24.2**	**23.7**	**23.5**	**24.7**	**24.8**	**24.0**	**24.3**	**24.7**	**23.1**	**24.0**
22.1	**24.4**	**26.5**	**25.7**	**25.5**	**25.2**	**22.7**	**23.6**	**26.4**	**24.2**	**23.7**	**23.5**	**22.7**	**25.7**	**25.5**	**25.2**	**22.7**

10.

44.1	**44.2**	**44.6**	**45.0**	**44.6**	**43.0**	**41.4**	**46.6**	**45.8**	**42.2**	**44.4**	**47.0**	**43.6**	**40.7**	**41.8**	**41.9**	**43.3**
44.3	**43.1**	**43.6**	**44.6**	**45.0**	**44.6**	**43.0**	**41.4**	**46.6**	**45.8**	**42.2**	**44.1**	**44.2**	**44.6**	**45.0**	**44.6**	**43.0**

11.

49.5	**49.5**	**46.7**	**47.2**	**49.1**	**48.7**	**51.2**	**46.1**	**50.5**	**48.9**	**49.3**	**50.4**	**47.2**	**48.5**	**49.4**	**48.8**	**50.1**
46.5	**51.2**	**46.2**	**46.1**	**50.5**	**49.5**	**48.7**	**51.2**	**47.2**	**49.1**	**49.5**	**46.7**	**47.2**	**49.1**	**48.7**	**47.2**	**49.1**

12.

27.3	**27.4**	**27.9**	**27.6**	**29.1**	**27.8**	**28.0**	**31.1**	**28.9**	**32.7**	**25.4**	**28.7**	**29.4**	**31.8**	**29.8**	**25.1**	**25.8**
33.3	**30.0**	**28.5**	**27.3**	**27.4**	**27.9**	**27.6**	**29.1**	**27.8**	**28.0**	**31.1**	**28.9**	**32.7**	**25.4**	**28.7**	**31.1**	**28.9**

13.

31.6	**31.4**	**31**	**31.4**	**33.1**	**33.0**	**32.9**	**31.4**	**31.9**	**33.5**	**32.3**	**32.0**	**31.9**	**32.6**	**32.7**	**32.2**	**32.3**
32.6	**31.6**	**32.1**	**32.9**	**31.4**	**31.9**	**33.5**	**32.3**	**32.0**	**31.6**	**31.4**	**31**	**31.4**	**33.1**	**33.0**	**32.9**	**31.4**

14.

40.9	**37.7**	**38.9**	**39.6**	**39.1**	**39.6**	**40.0**	**39.6**	**38.9**	**38.9**	**40.7**	**38.4**	**39.0**	**38.1**	**39**	**37.3**	**40.6**
38.3	**37.3**	**38.2**	**40.9**	**37.7**	**38.9**	**39.6**	**39.1**	**39.6**	**40.0**	**39.6**	**38.9**	**38.9**	**40.7**	**39.6**	**40.0**	**39.6**

15.

36.2	**35.9**	**36.1**	**34.5**	**34.5**	**34.6**	**35.2**	**34.3**	**32.8**	**36.9**	**36.4**	**35.8**	**34.6**	**34.7**	**34.8**	**35.9**	**35.3**
37	**34.3**	**36.0**	**36.2**	**35.9**	**36.1**	**34.5**	**34.5**	**34.6**	**35.2**	**34.3**	**32.8**	**36.9**	**36.4**	**34.5**	**34.6**	**35.2**

ЛАБОРАТОРНАЯ РАБОТА № 5
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ
(КРИТЕРИЙ ХИ КВАДРАТ ПИРСОНА)

Цель: Познакомиться с критериями согласия на примере критерия хи квадрат Пирсона, используемого для проверки согласованности полученных данных с гипотезой об их принадлежности к известному закону распределения.

Пусть $F(x,\theta)$ - функция распределения вероятностей случайной величины с неизвестным параметром θ. Значения случайной величины являются элементами выборки

$$x_1, x_2, ..., x_n .$$

Любое предположение, выделяющее некоторое подмножество, которому может принадлежать неизвестный параметр θ ($F(x,\theta)$), в математической статистике называют ***гипотезой***.

Так, например, статистическая гипотеза утверждает: неизвестный параметр θ распределения вероятностей $F(x,\theta)$ принадлежит заданному подмножеству $H \subset \Theta$ множества возможных значений параметра θ. Подмножество, которое является дополнением к подмножеству H, называется ***альтернативой к гипотезе*** H.

Гипотеза называется простой, если подмножество H состоит из одного единственного значения θ, в противном случае гипотеза называется сложной.

Практическое применение математической статистики состоит в проверке фактического соответствия реальных результатов экспериментов предполагаемой гипотезе. С этой целью строится процедура проверки гипотезы (критерий согласия), который позволяет по результатам наблюдений принимать или отвергать гипотезу.

Построение начинается с определения двух непересекающихся подмножеств A_0 и A_1. Правило проверки гипотезы формулируется следующим образом: если значения соответствующей статистики (функции от полученных наблюдений $f(x_1, x_2, ..., x_n)$) принадлежат A_0, то гипотеза H принимается; если значения соответствующей статистики $f(x_1, x_2, ..., x_n)$ принадлежат A_1, то гипотеза H отвергается.

Подмножество A_1 (область непринятия гипотезы H) называется ***критической областью***.

Подмножество A_0 (область принятия гипотезы H) называется ***областью дополнительных значений или областью принятия гипотезы***.

Поскольку события, состоящие в том, что значения функции от наблюдений принадлежат к A_0 или A_1, являются случайными, то применение процедуры проверки гипотезы сопряжено с ошибками двух родов: отвергнуть гипотезу, когда она верна (***ошибка первого рода***); принять гипотезу, когда она неверна (***ошибка второго рода***).

Если

$$P\left(f\left(x_1, x_2, ..., x_n\right) \in A_0\right) = 1 - \alpha,$$

то вероятность ошибки первого рода равна α. Значение вероятности ошибки первого рода называют уровнем значимости критерия.

Если вероятность ошибки второго рода равна β, то величину $1 - \beta$ называют ***мощностью критерия***.

При заданном уровне значимости критическую область строят так, чтобы мощность критерия была максимальной, тем самым уменьшают ошибку второго рода.

Одной из важнейших групп критериев проверки статистических гипотез являются критерии согласия, которые по выборочным данным проверяют предположение о принадлежности генеральной совокупности к некоторому распределению. Одним из наиболее мощных критериев согласия является критерий Пирсона или критерий хи-квадрат.

Будем проверять простую гипотезу H_0, состоящую в том, что исследуемая случайная величина X имеет функцию распределения $F(x)$ (или в частном случае дискретное распределение $\left(p_1, p_2, ..., p_m\right)$, $\sum_{i=1}^{m} p_i = 1$).

Разобьем числовую ось на непересекающиеся полуинтервалы

$$R = (y_0, y_1] \cup (y_1, y_2] \cup \ldots \cup (y_k, y_{k+1}),\ y_0 = -\infty, y_{k+1} = +\infty .$$

Определим $n_i, i = \overline{1,k+1}$ - числа попаданий элементов выборки $x_1, x_2, \ldots, x_n$ в полуинтервалы $(y_{i-1}, y_i],\ i = \overline{1,k+1}$.

Очевидно, что $n_i, i = \overline{1,k+1}$ можно представить в виде следующей суммы (аналогичное представление мы уже использовали при определении эмпирической функции распределения):

$$n_i = \sum_{j=1}^{n} I_{\{x_j \in (y_{i-1}, y_i]\}},$$

где

$$I_{\{x_j \in (y_{i-1}, y_i]\}} = \begin{cases} 1, x_j \in (y_{i-1}, y_i] \\ 0, x_j \notin (y_{i-1}, y_i] \end{cases},\ i = \overline{1,k+1},\ j = \overline{1,n}.$$

Обозначим через $p_i, i = \overline{1,k+1}$ вероятности $P(X \in (y_{i-1}, y_i]) = F(y_i) - F(y_{i-1}), i = \overline{1,k+1}$.

Вычислим $M(n_i)$:

$$M(n_i) = nM\left(I_{\{x_1 \in (y_{i-1}, y_i]\}}\right) = np_i .$$

Таким образом, если наша гипотеза H_0 верна, то $n_i, i = \overline{1,k+1}$ должны быть близки к $M(n_i) = np_i$. Мер этой близости можно предложить много, но К.Пирсоном была предложена следующая статистика:

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}.$$

При $n \to \infty$ распределение статистики χ^2 стремится к распределению χ^2_k.

Пусть $\chi^2_{k,\alpha}$ - решение уравнения

$$P\left(\chi^2_k > \chi^2_{k,\alpha}\right) = \alpha .$$

Поскольку $\chi^2 \to \chi^2_k$, то $\mathrm{P}\left(\chi^2 > \chi^2_{\alpha,k}\right) \to \mathrm{P}\left(\chi^2_k > \chi^2_{\alpha,k}\right) = \alpha$, то есть можно считать, что $\mathrm{P}\left(\chi^2 > \chi^2_{k,\alpha}\right) \approx \alpha$.

Если α мало, то событие $\chi^2 > \chi^2_{k,\alpha}$ является практически невозможным. Если $\chi^2 \le \chi^2_{k,\alpha}$, то считается, что полученные данные $x_1, x_2, ..., x_n$ не противоречат гипотезе H_0. Ошибка первого рода равна α.

Таким образом, область принятия гипотезы H_0:

$$\chi^2 \le \chi^2_{k,\alpha} .$$

Критическая область:

$$\chi^2 > \chi^2_{k,\alpha} .$$

Если известен тип распределения, но неизвестны его l параметров, то предельным распределением для

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}$$

является χ^2_{k-l}.

Замечания.

1) Необходимым условием применения критерия χ^2 является наличие в каждом из интервалов по крайней мере 5 - 10 наблюдений. Поэтому все частоты $n_i < 5$ следует объединить путем объединения соседних интервалов, в этом случае соответствующие им теоретические частоты (np_i) также надо сложить.

2) При проверке гипотезы о законе распределения контролируется лишь ошибка первого рода.

Предположим, что мы имеем выборку из непрерывного распределения объемом 40.

Будем проверять гипотезу H_0, состоящую в том, выборка принадлежит нормальному распределению.

Открываем электронную таблицу, вводим данные выборки в нее в ячейки A2-A41 и делаем надписи для расчетных параметров в соответствии с рис. 6.

	A	B	C	D	E	F	G
1	Выборка	Параметры:	Интервалы	Частота	Вер-ть	Теор.част.	Критерий
2	64	Объем					
3	56						
4	69	Максимум					
5	78						
6	78	Минимум					
7	83						
8	47	Среднее					
9	65						
10	77	СКО					
11	57						

Рис. 6. Пример ввода данных и подписей для расчетных параметров.

Вычисляем параметры по выборке. Для этого вводим в ячейку B2 **=СЧЁТ(A2:A41)** . Функции можно вводить с помощью мастера функций *fx* из категории «Статистические». Ссылки на ячейки можно ввести, щелкнув по ним мышью. В B4 вводим **=МАКС(A2:A41)**, в B6 **=МИН(A2:A41)**, в B8**=СРЗНАЧ(A2:A41)**, в B10**=СТАНДОТКЛОН(A2:A41)**.

Разбиваем полученный интервал [**МИН, МАКС**] на интервалы группировки и введем в ячейки C2-C11 границы интервалов (для примера предполагаем, что у нас 9 интервалов разбиения, поэтому их границы размещаются в ячейках C2-C11). Для вычисления частот n_i используется функцию **ЧАСТОТА**. Для этого в D3 вводим формулу **=ЧАСТОТА(A2:A41;C3:C11)**, затем обводим курсором ячейки D3-D11,

выделяя их и нажимаем F2, а затем одновременно Ctrl+Shift+Enter. В результате в ячейках D3-D11 окажутся значения частот.

Для расчета теоретической вероятности

$$p_i = \mathrm{P}\left(X \in (y_{i-1}, y_i]\right) = F(y_i) - F(y_{i-1})$$

вводим в ячейку E3

=НОРМРАСП(C3;B8;B10;1)-НОРМРАСП(C2;B8;B10;1).

С помощью этой формулы вычисляем разность $F(y_i) - F(y_{i-1})$ между значениями функции нормального распределения (функция **НОРМРАСП** категории «Статистические»). В скобках стоят следующие параметры: первый (например, C3) – значение границы интервала, второй «Среднее» - (ссылка на ячейку B8), третий - «Стандартное откл» (ссылка на ячейку B10), четвертый - «Интегральная» - 1. Автозаполняем эту формулу на E3-E10 перемещая нижний правый угол E3 до ячейки E10. В последней ячейке столбца E11 для соблюдения условия нормировки вводим дополнение предыдущих вероятностей до единицы. Для этого вводим в ячейку E11: **=1-СУММ(E3:E10)**

Для расчета теоретической частоты np_i вводим в ячейку F3 формулу: **=E3*B3** и автозаполняем ее на F3-F11.

Для вычисления элементов суммы $\chi^2 = \sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}$ критерия Пирсона вводим в G3 значение **=(D3-F3)*(D3-F3)/F3** и автозаполняем его на

диапазон G3-G11.

Находим значение критерия

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}$$

и критическое значение $\chi^2_{k,\alpha}$.

Для этого вводим в F12 **Сумма**, а в F13 подпись **Критич**. Вводим в ячейку G12 **=СУММ(G3:G11)**, а в ячейку G13 **=ХИ2ОБР(0,05;6)**. Первый параметр в скобках- заданное значение α. Для примера мы взяли α=0,05. Второй параметр – число степеней свободы. У нас 6 степеней свободы (*k*-*l*-1)=(9-2-1)=6, так как *k*=9 – число интервалов группировки, а *l*=2, так как у нормального распределения два неизвестных параметра распределения – математическое ожидание и дисперсия.

Таким образом, область принятия гипотезы H_0:

$$\chi^2 \le \chi^2_{k,\alpha}.$$

Критическая область:

$$\chi^2 > \chi^2_{k,\alpha}.$$

Если мы принимаем основную гипотезу, то это решение можно еще раз подтвердить, построив графики плотностей эмпирического и теоретического распределений. Ставим курсор в любую свободную ячейку и вызываем мастер диаграмм (Вставка/Диаграмма). Выбираем тип диаграммы «График» и вид «График с маркерами» самый левый во второй строке, нажимаем «Далее». Ставим курсор в поле «Диапазон» и удерживая кнопку CTRL обводим мышью область ячеек D3-D11 а затем F3-F11. Переходим на закладку «Ряд» и в поле «Подписи оси X» обводим область C3-C11. Нажимаем «Готово». Анализируем, достаточно ли хорошо совпадают графики, что говорит о соответствии данных нормальному закону.

Задания

Взяв данные из лабораторной работы № 4, проверить гипотезу об их принадлежности к нормальному распределению.

ЛАБОРАТОРНАЯ РАБОТА № 6
КРИТЕРИЙ ФИШЕРА СРАВНЕНИЯ ДИСПЕРСИЙ В ДВУХ НОРМАЛЬНЫХ ВЫБОРКАХ

Цель: Научиться проверять гипотезу об однородности (равенстве) дисперсий двух нормальных выборок.

Критерий Фишера используется в случае, если нужно проверить различается ли разброс данных (дисперсии) у двух выборок. Основной характеристикой критерия является уровень значимости α или доверительная вероятность $1-\alpha$.

Имеются две независимые выборки, принадлежащие нормальным законам распределения:

$$X_{11}, X_{12}, ..., X_{1n_1} \in N\left(a_1, \sigma_1^2\right),$$

$$X_{21}, X_{22}, ..., X_{2n_2} \in N\left(a_2, \sigma_2^2\right).$$

Через a_1, σ_1^2 обозначены соответственно математическое ожидание и дисперсия случайных величин из первой выборки, а через a_2, σ_2^2 - математическое ожидание и дисперсия случайных величин из второй выборки.

В экспериментальных исследованиях часто возникает вопрос о сравнении a_1 и a_2, σ_1^2 и σ_2^2. Очевидно, что это сравнение будет происходить на основе соответствующих выборочных характеристик

$$\overline{X_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki},\ S_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n}\left(X_{ki} - \overline{X_k}\right)^2, k = \overline{1,2}.$$

Начнем с задачи сравнения σ_1^2 и σ_2^2. Хорошо известно, что

$\frac{(n_1-1)S_1^2}{\sigma_1^2} \in \chi^2_{n_1-1}$ (χ^2-распределению с n_1-1 степенями свободы),

$\frac{(n_2-1)S_2^2}{\sigma_2^2} \in \chi^2_{n_2-1}$ (χ^2-распределению с n_2-1 степенями свободы.

Как известно, отношение $\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2\lambda}$ имеет распределение Фишера с

n_1-1 и n_2-1 степенями свободы, $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$. В числителе стоит выборочная дисперсия, которая больше. Тогда

$$\mathrm{P}\left(\Delta_{1-\frac{\alpha}{2}} \le \frac{S_1^2}{S_2^2\lambda} \le \Delta_{\frac{\alpha}{2}}\right) = 1-\alpha,$$

$\Delta_{\frac{\alpha}{2}}, \Delta_{1-\frac{\alpha}{2}}$ находятся из таблицы критических точек распределения Фишера с n_1-1 и n_2-1 степенями свободы.

Таким образом, для λ получаем доверительный интервал

$$\frac{S_1^2}{S_2^2\Delta_{\frac{\alpha}{2}}} \le \lambda \le \frac{S_1^2}{S_2^2\Delta_{1-\frac{\alpha}{2}}}.$$

В качестве основной гипотезы будем рассматривать $H_0 : \sigma_1^2 = \sigma_2^2$, в качестве конкурирующей гипотезы будем рассматривать $H_1 : \sigma_1^2 \ne \sigma_2^2$.

Критическая область строится на основе доверительного интервала для параметра λ. Подставляя вместо параметра λ единицу, получим:

$$\frac{S_1^2}{S_2^2\Delta_{\frac{\alpha}{2}}} \le 1,\ 1 \le \frac{S_1^2}{S_2^2\Delta_{1-\frac{\alpha}{2}}},$$

$$\Delta_{\frac{\alpha}{2}} \ge \frac{S_1^2}{S_2^2},\ \Delta_{1-\frac{\alpha}{2}} \le \frac{S_1^2}{S_2^2}$$

или

$$\Delta_{1-\frac{\alpha}{2}} \le \frac{S_1^2}{S_2^2} \le \Delta_{\frac{\alpha}{2}}.$$

Если это неравенство выполняется, то считается, что полученные наблюдения не противоречат гипотезе H_0, и она принимается. В противном случае гипотеза H_0 отвергается и принимается конкурирующая гипотеза H_1. То есть, если значение статистики $\frac{S_1^2}{S_2^2}$, вычисленное в соответствии с полученными данными, принадлежит отрезку $\left[\Delta_{\frac{\alpha}{2}}, \Delta_{1-\frac{\alpha}{2}}\right]$ (области принятия гипотезы), то основная гипотеза H_0 принимается.

Критической областью в этом случае является двусторонняя область

$$\frac{S_1^2}{S_2^2} \in \left(-\infty, \Delta_{\frac{\alpha}{2}}\right) \cup \left(\Delta_{1-\frac{\alpha}{2}}, +\infty\right).$$

В качестве основной гипотезы рассмотрим $H_0 : \sigma_1^2 = \sigma_2^2$, в качестве конкурирующей гипотезы рассмотрим $H_1 : \sigma_1^2 > \sigma_2^2$.

Тогда областью принятия гипотезы H_0 является отрезок $\frac{S_1^2}{S_2^2} \in \left[\Delta_{\frac{\alpha}{2}}, \Delta_{1-\frac{\alpha}{2}}\right]$, а критической областью является односторонняя область $\frac{S_1^2}{S_2^2} \in \left(\Delta_{1-\frac{\alpha}{2}}, +\infty\right)$.

При проверке основной гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ при конкурирующей гипотезе $H_1 : \sigma_1^2 > \sigma_2^2$ область принятия гипотезы строят еще, исходя из условия

$$\mathrm{P}\left(\frac{S_1^2}{S_2^2 \lambda} \le \Delta_\alpha\right) = 1 - \alpha \text{ или } \mathrm{P}\left(\frac{S_1^2}{S_2^2 \lambda} > \Delta_\alpha\right) = \alpha,$$

Тогда, подставляя $\lambda = 1$, получаем, что областью принятия гипотезы H_0 является интервал $\frac{S_1^2}{S_2^2} \in (-\infty, \Delta_\alpha]$, а критической областью интервал $\frac{S_1^2}{S_2^2} \in (\Delta_\alpha, +\infty)$.

Рассмотрим основную гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ и конкурирующую гипотезу $H_1 : \sigma_1^2 < \sigma_2^2$.

Тогда областью принятия гипотезы H_0 по-прежнему является отрезок $\frac{S_1^2}{S_2^2} \in \left[\Delta_{\frac{\alpha}{2}}, \Delta_{1-\frac{\alpha}{2}}\right]$, а критической областью является односторонняя область $\frac{S_1^2}{S_2^2} \in \left(-\infty, \Delta_{\frac{\alpha}{2}}\right)$.

Или областью принятия гипотезы H_0 является интервал $\frac{S_1^2}{S_2^2} \in [\Delta_\alpha, +\infty)$, а критической областью интервал $\frac{S_1^2}{S_2^2} \in (-\infty, \Delta_\alpha)$.

Построенные критерии называются соответственно двусторонним и односторонними ***критериями Фишера***.

Для проверки двустороннего критерия Фишера в Excel можно использовать функцию **ФТЕСТ**(массив1;массив2), где массив1 – диапазон со значениями первой выборки, массив2 – диапазон со значениями второй выборки. Результатом выполнения этой функции оказывается уровень значимости, соответствующий степени различия дисперсий или вероятность того, что различия дисперсий недостоверны. Поскольку обычно уровень значимости α принимается 0.05 все значения функции **ФТЕСТ**, меньшие 0.05, будут свидетельствовать о достоверных отличиях между дисперсиями. Расчетные уровни значимости можно перевести в привычную форму критерия Фишера с помощью функции **FРАСПОБР** (вероятность;

степени свободы 1; степени свободы 2), где вероятность – уровень значимости, рассчитанный функцией **ФТЕСТ** или ссылка на ячейку, содержащую формулу этой функции, степени свободы 1 – число степеней свободы для выборки с большей дисперсией, степени свободы 2 – число степеней свободы для выборки с меньшей дисперсией.

Для решения задачи можно использовать надстройку **«Анализ данных»**. Далее нужно выбрать **«Двухвыборочный F-тест для дисперсий».** В открывшемся окне в полях **«Интервал переменной 1»** и **«Интервал переменной 2»** вводят ссылки на данные двух выборок. Далее вводят уровень значимости в поле **«Альфа»**. По умолчанию α =0.05. В разделе **«Параметры вывода»** ставят метку около **«Выходной интервал»** и, поместив курсор в появившееся поле, щелкают левой кнопкой в любой свободной ячейке. Вывод результата будет осуществляться, начиная с этой ячейки. Нажав на **«Ок»**, получаем таблицу результатов, в которой указаны средние и дисперсии для каждой выборки, число степеней свободы, значение F – критерия $(\frac{S_1^2}{S_2^2})$, односторонний критический уровень значимости в строке **«P(F<=f) одностороннее»** и критическое значение F-критерия (Δ_α). Если значение F - критерия меньше, чем F -критическое, то с заданной вероятностью можно считать, что дисперсии равны. Если значение F - критерия больше, чем F -критическое, то дисперсии различны.

Задания

Два однотипных станка изготавливают изделия по одному и тому же образцу. На основании измерений изделий, произведенных станками, при доверительной вероятности 0.95 проверить гипотезу об одинаковой точности станков.

1.

33.6	35.3	36.0	37.6	35.9	33.2	36.3	37.1	37.9	39.6	35.1	36.6	34.0	31.2	39.4
35.4	32.7	33.1	36.0	37.6	35.9	33.2	36.3	37.1	37.9	39.6	33.6	35.3	36.0	37.6

36.6	36.9	35.4	38.0	37	37.7	36.8	35.1	37.3	35.2	36.3	36.0	35.4	34.7	36.7
37.0	36.1	37.3	36.6	36.9	35.4	38.0	37	37.7	36.8	35.1	37.3	35.2	36.8	35.1

2.

23.1	22.8	22.6	20.4	21.4	22.5	21.4	22.3	22.3	22.0	21.4	22.9	23.0	22.6	21.6
22.8	22.5	23.1	22.8	22.6	20.4	21.4	22.5	21.4	22.3	22.3	22.0	21.4	23.1	22.8

22.3	23.3	22.2	23.2	22.9	23.1	22.5	23.1	22.6	23.2	22.3	22.3	22.0	21.9	23.0
23.1	22.5	23.1	22.6	23.2	22.3	22.3	22.0	21.9	23.0	22.7	22.3	23.3	22.2	23.2

3.

43.0	43.3	42.6	43.5	42.7	42.5	41.2	42.0	43.6	42.6	42.5	43.7	42.9	42.8	42.3
42.2	42.0	42.0	43.6	42.6	42.5	43.7	42.9	42.8	42.3	43.1	43.0	43.3	42.6	43.5

43.5	41.1	42.3	42.2	42.5	43.4	43.4	41.9	41.5	42.1	41.7	43.5	42.8	43.3	43.8	42.7
43.3	44.3	41.1	43.5	41.1	42.3	42.2	42.5	43.4	43.4	41.9	41.5	42.1	41.7	43.5	42.8

4.

32.7	32.3	32.5	31.9	33.1	31.9	32.3	32.8	32.3	32.8	31.9	32.7	32.5	32.8	34.5	32.7
32.3	33.8	32.7	32.3	32.5	31.9	33.1	31.9	32.3	32.8	32.3	32.8	31.9	32.7	32.7	32.3

32.0	32.5	32.5	31.8	32.7	32.5	31.5	33.0	33.1	30.7	31.8	32.0	33.4	33.6	32.1	32.5
32.7	32.9	32.0	32.5	32.5	31.8	32.7	32.5	31.5	33.0	33.1	30.7	31.8	32.0	33.4	32.7

5.

23.4	23.2	21.7	23.0	22.3	22.5	22.8	21.9	22.6	22.5	23.2	23.2	22.7	22.2	23.5	22.3
24.2	21.4	24.0	23.4	23.2	21.7	23.0	22.3	22.5	22.8	21.9	22.6	22.5	23.2	23.2	22.7

25.1	22.5	22.1	23.3	23.0	22.8	19.9	23.2	22.6	23.5	20.8	22.0	22.6	22.5	23.9	21.4
22.8	21.1	23.7	25.1	22.5	22.1	23.3	23.0	22.8	19.9	23.2	22.6	23.5	20.8	22.0	25.1

6.

21.3	28.5	31.1	28.9	29.1	24.8	23.6	27.7	29.7	29.1	25.2	27.8	27.7	25.2	25.5
25.6	26.1	21.3	28.5	31.1	28.9	29.1	24.8	23.6	27.7	29.7	29.1	25.2	27.8	27.7

26.9	**27.6**	**28.7**	**26.8**	**27.3**	**27.3**	**28.6**	**28.2**	**24.5**	**27.4**	**26.3**	**26.3**	**26.8**	**26.1**	**27.1**
25.9	**28.7**	**26.9**	**27.6**	**28.7**	**26.8**	**27.3**	**27.3**	**28.6**	**28.2**	**24.5**	**27.4**	**26.3**	**26.3**	**26.9**

7.

12.6	**12.2**	**11.8**	**13.0**	**12.4**	**12.0**	**12.1**	**12.6**	**12.7**	**12.3**	**13.0**	**12.7**	**12.9**	**12.7**	**12.2**	**12.9**
12.1	**12.6**	**12.7**	**12.3**	**13.0**	**12.7**	**12.9**	**12.7**	**12.2**	**12.9**	**11.8**	**13.0**	**12.4**	**12.0**	**12.1**	**12.6**

12.5	**12.3**	**13.1**	**13.3**	**11.8**	**13.2**	**12.0**	**11.9**	**12.3**	**12.5**	**12.9**	**12.5**	**11.8**	**12.5**	**12.7**	**12.9**
13.4	**12.3**	**12.5**	**12.3**	**13.1**	**13.3**	**11.8**	**13.2**	**12.0**	**11.9**	**12.3**	**12.5**	**12.9**	**12.5**	**11.8**	**12.3**

8.

43.9	**41.6**	**42.3**	**43.4**	**42.3**	**43.4**	**41.6**	**43.2**	**42.8**	**43.3**	**46.8**	**43.1**	**42.3**	**45.3**	**46.0**
40.3	**45.8**	**40.9**	**43.9**	**41.6**	**42.3**	**43.4**	**42.3**	**43.4**	**41.6**	**43.2**	**42.8**	**43.3**	**46.8**	**43.1**

41.9	**42.7**	**43.1**	**43.9**	**43.0**	**41.7**	**43.2**	**43.6**	**44.0**	**44.9**	**42.6**	**43.4**	**42.1**	**40.7**	**44.8**
42.8	**41.5**	**41.6**	**41.9**	**42.7**	**43.1**	**43.9**	**43.0**	**41.7**	**43.2**	**43.6**	**44.0**	**44.9**	**42.6**	**43.4**

9.

23.1	**22.4**	**25.4**	**22.0**	**21.6**	**23.6**	**21.8**	**22.8**	**24.1**	**22.7**	**22.3**	**22.4**	**24.4**	**22.2**	**22.0**
23.9	**22.5**	**23.1**	**22.4**	**25.4**	**22.0**	**21.6**	**23.6**	**21.8**	**22.8**	**24.1**	**22.7**	**22.3**	**22.4**	**24.4**

24.1	**19.4**	**23.0**	**24.1**	**26.2**	**20.8**	**22.5**	**24.2**	**23.6**	**23.6**	**24.3**	**24.9**	**25.1**	**25.0**	**23.8**
21.9	**24.2**	**24.1**	**19.4**	**23.0**	**24.1**	**26.2**	**20.8**	**22.5**	**24.2**	**23.6**	**23.6**	**24.3**	**24.9**	**25.1**

10.

20.6	**22.6**	**20.8**	**21.8**	**23.1**	**21.7**	**21.3**	**21.4**	**23.4**	**21.2**	**21.0**	**22.9**	**21.5**	**22.8**	**22.9**	**22.8**
20.8	**21.8**	**23.1**	**21.7**	**21.3**	**21.4**	**23.4**	**21.2**	**21.0**	**22.9**	**21.5**	**22.8**	**22.9**	**22.8**	**20.6**	**22.6**

22.5	**22.2**	**22.9**	**22.8**	**21.9**	**21.3**	**22.8**	**22.6**	**20.4**	**23.8**	**23.7**	**24.3**	**23.9**	**23.5**	**21.8**	**22.1**
23.6	**23.8**	**22.5**	**22.2**	**22.9**	**22.8**	**21.9**	**21.3**	**22.8**	**22.6**	**20.4**	**23.8**	**23.7**	**24.3**	**23.9**	**23.5**

11.

23.7	**24.7**	**24.4**	**22.5**	**23.8**	**23.8**	**22.5**	**22.6**	**22.7**	**22.9**	**22.8**	**23.5**	**22.9**	**23.0**	**24.5**	**23.4**	**25.3**
21.7	**23.3**	**23.7**	**24.7**	**24.4**	**22.5**	**23.8**	**23.8**	**22.5**	**22.6**	**22.7**	**22.9**	**22.8**	**23.5**	**22.9**	**23.0**	**24.5**

24.1	**26.5**	**24.5**	**19.8**	**20.5**	**28.0**	**24.7**	**23.2**	**21.6**	**21.2**	**20.0**	**21.1**	**25.5**	**25.1**	**24.8**	**21.1**	**22.4**

26.5	**23.3**	**24.1**	**26.5**	**24.5**	**19.8**	**20.5**	**28.0**	**24.7**	**23.2**	**21.6**	**21.2**	**20.0**	**21.1**	**25.5**	**25.1**	**24.8**

12.

23.6	**25.5**	**25.8**	**24.3**	**26.9**	**25.9**	**26.6**	**25.7**	**24.0**	**26.2**	**24.1**	**25.2**	**24.9**	**24.3**	**23.6**	**25.6**
25.9	**25.0**	**23.6**	**25.5**	**25.8**	**24.3**	**26.9**	**25.9**	**26.6**	**25.7**	**24.0**	**26.2**	**24.1**	**25.2**	**24.9**	**24.3**

27.1	**25.2**	**24.5**	**21.3**	**23.4**	**27.4**	**24.9**	**24.7**	**27.5**	**25.7**	**25.3**	**24.1**	**26.1**	**23.9**	**23.3**	**26.7**
22.0	**27.1**	**25.2**	**24.5**	**21.3**	**23.4**	**27.4**	**24.9**	**24.7**	**27.5**	**25.7**	**25.3**	**24.1**	**26.1**	**23.9**	**23.3**

13.

18.7	**14.4**	**13.2**	**17.3**	**19.3**	**18.7**	**14.8**	**17.4**	**17.3**	**14.8**	**15.1**	**15.2**	**15.7**	**15.4**	**16.9**	**15.6**
13.2	**17.3**	**19.3**	**18.7**	**14.8**	**17.4**	**17.3**	**14.8**	**15.1**	**15.2**	**15.7**	**15.4**	**16.9**	**15.6**	**18.7**	**14.4**

17.6	**16.5**	**18.4**	**14.8**	**16.4**	**16.8**	**18.0**	**17.0**	**14.6**	**15.0**	**18.8**	**17.1**	**16.4**	**15.5**	**15.3**	**14.7**
18.4	**14.8**	**16.4**	**16.8**	**18.0**	**17.0**	**14.6**	**15.0**	**18.8**	**17.1**	**16.4**	**15.5**	**15.3**	**14.7**	**17.6**	**16.5**

14.

44.9	**45.2**	**44.2**	**45.3**	**44.4**	**45.4**	**44.0**	**44.0**	**43.5**	**43.3**	**45.0**	**44.6**	**44.9**	**44.6**	**44.5**	**41.9**
44.2	**45.3**	**44.4**	**45.4**	**44.0**	**44.0**	**43.5**	**43.3**	**45.0**	**44.6**	**44.9**	**44.6**	**44.5**	**41.9**	**44.9**	**45.2**

41.6	**44.1**	**41.6**	**43.6**	**43.8**	**42.9**	**41.5**	**45.2**	**45.5**	**44.6**	**42.0**	**44.3**	**44.9**	**46.8**	**44.8**	**43.5**
44.7	**45.8**	**41.6**	**44.1**	**41.6**	**43.6**	**43.8**	**42.9**	**41.5**	**45.2**	**45.5**	**44.6**	**42.0**	**44.3**	**44.9**	**46.8**

15.

41.7	**39.5**	**42.9**	**42.8**	**43.4**	**43.0**	**42.6**	**40.9**	**41.2**	**42.7**	**42.9**	**42.1**	**40.9**	**41.0**	**41.6**	**43.1**	**45.3**
41.5	**42.5**	**41.7**	**39.5**	**42.9**	**42.8**	**43.4**	**43.0**	**42.6**	**40.9**	**41.2**	**42.7**	**42.9**	**42.1**	**40.9**	**41.0**	**41.6**

41.9	**43.2**	**42.8**	**39.1**	**42.0**	**40.9**	**40.9**	**41.4**	**40.7**	**41.7**	**40.5**	**43.3**	**38.6**	**42.2**	**43.5**	**42.4**
42.5	**40.4**	**41.9**	**43.2**	**42.8**	**39.1**	**42.0**	**40.9**	**40.9**	**41.4**	**40.7**	**41.7**	**40.5**	**43.3**	**38.6**	**42.2**

ЛАБОРАТОРНАЯ РАБОТА № 7
КРИТЕРИЙ СТЬЮДЕНТА СРАВНЕНИЯ МАТЕМАТИЧЕСКИХ ОЖИДАНИЙ В ДВУХ НОРМАЛЬНЫХ ВЫБОРКАХ

Цель: Научиться проверять гипотезу об однородности (равенстве) математических ожиданий двух нормальных выборок.

Критерий Стьюдента используется для проверки предположения о том, что математические ожидания (средние значения) двух выборок значимо различаются. Существует три разновидности критерия: один – для связанных выборок, и два для несвязанных выборок (с одинаковыми и разными дисперсиями). Если выборки не связаны, то предварительно нужно проверить гипотезу о равенстве дисперсий, чтобы определить, какой из критериев использовать.

Решим задачу сравнения средних a_1 и a_2 в двух выборках

$$X_{11}, X_{12}, ..., X_{1n_1} \in N\left(a_1, \sigma_1^2\right),$$

$$X_{21}, X_{22}, ..., X_{2n_2} \in N\left(a_2, \sigma_2^2\right).$$

при условии равенства дисперсий $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Рассмотрим статистику

$$\overline{X_1} - \overline{X_2} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}\,.$$

Поскольку обе выборки принадлежат нормальным распределениям, то статистика $\overline{X_1} - \overline{X_2}$ тоже принадлежит нормальному закону распределения с параметрами $\left(a_1 - a_2, \dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}\right)$.

Соответственно статистика

$$\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

имеет стандартное (с параметрами $(0,1)$) нормальное распределение.

Известно, что

$$\frac{(n_1-1)S_1{}^2+(n_2-1)S_2{}^2}{\sigma^2}\in\chi^2_{n_1+n_2-2},$$

$$S_k^2=\frac{1}{n_k-1}\sum_{i=1}^{n}\left(X_{ki}-\overline{X_k}\right)^2, k=\overline{1,2}.$$

Поскольку

$$\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\in N(0,1),$$

то отношение

$$\frac{\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}}=\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

имеет распределение Стьюдента ($t_{n_1+n_2-2}$-распределение) с n_1+n_2-2 степенями свободы. Построим доверительный интервал для разности a_1-a_2. С вероятностью $1-\alpha$ выполняется

$$-\Delta_{n_1+n_2-2,\alpha}\le\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\le\Delta_{n_1+n_2-2,\alpha},$$

где $\Delta_{n,\alpha}$ находится из таблицы для вероятностей $P(|t_n|>\Delta_{n,\alpha})=\alpha$ распределения Стьюдента.

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}.$$

После соответствующих преобразований получаем:

$$\overline{x_1} - \overline{x_2} - \Delta_{n_1+n_2-2,\alpha} s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le a_1 - a_2 \le \overline{x_1} - \overline{x_2} + \Delta_{n_1+n_2-2,\alpha} s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

В качестве основной гипотезы будем рассматривать $H_0 : a_1 = a_2$, в качестве конкурирующей гипотезы будем рассматривать $H_1 : a_1 \neq a_2$.

Подставляя в доверительный интервал ноль вместо $a_1 - a_2$, получаем:

$$-\Delta_{n_1+n_2-2,\alpha} \le \frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le \Delta_{n_1+n_2-2,\alpha}.$$

Таким образом, если значение статистики

$$\frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

принадлежит отрезку $\left[-\Delta_{n_1+n_2-2,\alpha}, \Delta_{n_1+n_2-2,\alpha}\right]$, то гипотеза $H_0 : a_1 = a_2$ принимается.

Если значение статистики

$$\frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

попадают в двустороннюю критическую область $\left(-\infty, -\Delta_{n_1+n_2-2,\alpha}\right) \cup \left(\Delta_{n_1+n_2-2,\alpha}, +\infty\right)$, то принимается конкурирующая гипотеза $H_1 : a_1 \neq a_2$.

Так же, как и в случае сравнения дисперсий, имеются два способа решения задачи. Можно использовать стандартную функцию **ТТЕСТ**(массив1; массив2; хвосты; тип), где массив1 – диапазон со значениями первой выборки, массив2 – диапазон со значениями второй

выборки, хвосты- вид критерия: 1- односторонний критерий, 2- двусторонний критерий, тип – тип критерия: 1- выборки связаны, 2- несвязанные выборки с равными дисперсиями, 3-несвязанные выборки с неравными дисперсиями. Результатом выполнения этой функции оказывается уровень значимости, соответствующий степени различия математических ожиданий, или вероятность того, что различия недостоверны. Поскольку обычно уровень значимости α принимается 0.05 все значения функции **ТТЕСТ**, меньшие 0.05, будут свидетельствовать о достоверных отличиях между математическими ожиданиями.

Для решения задачи можно использовать надстройку **АНАЛИЗ ДАННЫХ**. Далее нужно выбрать один из трех тестов – **ПАРНЫЙ ДВУХВЫБОРОЧНЫЙ t-ТЕСТ ДЛЯ СРЕДНИХ** (для связанных выборок), **ДВУХВЫБОРОЧНЫЙ t-ТЕСТ С ОДИНАКОВЫМИ ДИСПЕРСИЯМИ** или **ДВУХВЫБОРОЧНЫЙ t-ТЕСТ С РАЗНЫМИ ДИСПЕРСИЯМИ.** Последние два теста для несвязанных выборок. В открывшемся окне в полях **«Интервал переменной 1»** и **«Интервал переменной 2»** вводят ссылки на данные двух выборок. Если имеются подписи данных, то ставят флажок у надписи **«Метки».** Далее вводят уровень значимости в поле **«Альфа»**. По умолчанию α =0.05. Поле **«Гипотетическая средняя разность»** оставляют пустым. В разделе **«Параметры вывода»** ставят метку около **«Выходной интервал»** и, поместив курсор в появившееся поле, щелкают левой кнопкой в любой свободной ячейке. Вывод результата будет осуществляться, начиная с этой ячейки. Нажав на **Ок**, получаем таблицу результатов, в которой указаны средние и дисперсии для каждой выборки, число степеней свободы, значение t-статистики $T=(\dfrac{\overline{x_2}-\overline{x_1}}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}})$, критические значения этой статистики (одностороннее критическое значение t, которое находится из

уравнения $P(t_n > t) = \alpha$ и двустороннее критическое значение t, которое находится из уравнения $P(|t_n| > t) = \alpha$) и критические уровни значимости **«P(T ≤ t) одностороннее» «P(|T| ≤ t) двухстороннее».** Если по модулю t-статистика T меньше критического, то математические ожидания с заданной вероятностью равны.

Задания (данные взять из заданий лабораторной работы № 6)

Два однотипных станка изготавливают изделия. На основании измерений изделий, произведенных станками, при доверительной вероятности 0.95 проверить гипотезу о равенстве номинальных размеров изделий (математических ожиданий), произведенных разными станками.

ЛАБОРАТОРНАЯ РАБОТА № 8
КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Цель: Научиться аппроксимировать набор наблюдений линейной функцией.

Цель регрессионного анализа состоит в определении общего вида уравнения регрессии, построении оценок неизвестных параметров, входящих в уравнение регрессии, и проверке статистических гипотез о регрессии. В зависимости от формы связи между переменными различают ***линейную*** и ***нелинейную*** регрессию. Наиболее простым является случай, когда регрессия линейна.

Рассмотрим задачу наилучшей аппроксимации набора наблюдений

$x_i, y_i, i = \overline{1,n}$ линейной функцией $f(X) = a + bX$ в смысле минимизации функционала

$$F = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2 .$$

Запишем необходимые условия экстремума:

$$\frac{\partial F}{\partial a} = -2\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\frac{\partial F}{\partial b} = -2\sum_{i=1}^{n} x_i \left(y_i - a - bx_i\right) = 0$$

или

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\sum_{i=1}^{n} x_i \left(y_i - a - bx_i\right) = 0 .$$

Раскрыв скобки, получим:

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i .$$

Решая систему уравнений, находим неизвестные a и b:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i .$$

Добавим к постановке задачи некоторые статистические данные и запишем линейное регрессионное уравнение в виде:

$$Y_i = a + bX_i + \varepsilon_i, i = \overline{1,n},$$

где X_i - неслучайная (детерминированная) величина, Y_i, ε_i - случайные величины, ε_i - ошибки регрессии.

Основные гипотезы:

1. $Y_i = a + bX_i + \varepsilon_i$, $i = \overline{1,n}$ - спецификация модели.

2. X_i - детерминированная величина; вектор $(X_1, X_2, ..., X_n)$ не коллинеарен вектору $(1,1,...,1)$.

3. $M(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$, $i = \overline{1,n}$, $M(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.

Часто добавляется условие

4. ε_i - нормально распределенная случайная величина, $M(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$.

Как утверждает ***теорема Гаусса – Маркова***, в этих предположениях оценки неизвестных параметров модели

$$b'' = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^{\,2} - \left(\sum_{i=1}^{n} X_i\right)^2}$$

и

$$a'' = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} X_i,$$

полученные по МНК, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Непосредственно из 1) - 4) следует, что Y_i - нормально распределенная случайная величина, $M(Y_i) = a + bX_i, D(Y_i) = \sigma^2$.

Нетрудно проверить, что

$$M(b'')=b,\ M(a'')=a, D(b'')=S_b^2=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}, D(a'')=S_a^2=\sigma^2\frac{\sum_{i=1}^{n}x_i^2}{n\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)},$$

$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$,

поэтому

$$a''\in N\left(a,\sigma^2\frac{\sum_{i=1}^{n}x_i^2}{n\left(\sum_{i=1}^{n}x_i^2-n\overline{x}^2\right)}\right),\ b''\in N\left(b,\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\right).$$

Обозначим через $e_i, i=\overline{1,n}$ разницу между действительным значением переменной Y и модельным значением этой переменной, то есть

$$e_i=Y_i-a-bX_i, i=\overline{1,n}.$$

Несмещенной оценкой дисперсии ошибок σ^2 является

$$S^2=\frac{1}{n-2}\sum_{i=1}^{n}e_i^2.$$

Нетрудно показать, что S^2 независима с a'' и b'', а $\frac{(n-2)S^2}{\sigma^2}\in\chi^2_{n-2}$.

Построим статистику для проверки гипотезы $H_0: b=b_0$ против альтернативной гипотезы $H_1: b\neq b_0$.

Поскольку $b''-b\in N\left(0,\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\right)\in N\left(0,\sigma_b^2\right)$, то $\frac{b''-b}{\sigma_b}\in N(0,1)$. Из условия $\frac{(n-2)S^2}{\sigma^2}\in\chi^2_{n-2}$ следует, что

$$\frac{\dfrac{b''-b}{\sigma_b}}{\sqrt{\dfrac{(n-2)S^2}{\sigma^2(n-2)}}}=\frac{b''-b}{\dfrac{\sigma_b S}{\sigma}}=\frac{b''-b}{S_b}\in t_{n-2}$$

(распределению Стьюдента с $n-2$ степенями свободы).

Таким образом, для проверки гипотезы $H_0: b=b_0$ против альтернативной гипотезы $H_1: b\neq b_0$ будет использоваться статистика $\dfrac{b''-b}{S_b}$.

Построим доверительный интервал для b, используя распределение t_{n-2} и его двусторонние квантили $t_{n-2,\alpha}$, которые находятся из таблицы для вероятностей $\mathrm{P}\left(|t_{n-2}|\le t_{n-2,\alpha}\right)=1-\alpha$ или $\mathrm{P}\left(|t_{n-2}|> t_{n-2,\alpha}\right)=\alpha$:

$$\mathrm{P}\left(-t_{n-2,\alpha}\le\frac{b''-b}{S_b}\le t_{n-2,\alpha}\right)=1-\alpha,$$

откуда следует $b''-t_{n-2,\alpha}S_b\le b<b''+t_{n-2,\alpha}S_b$.

Если b_0 принадлежит отрезку $\left\lfloor b''-t_{n-2,\alpha}S_b, b''+t_{n-2,\alpha}S_b\right\rfloor$, то принимается гипотеза H_0, в противном случае принимается гипотеза H_1. Если требуется проверить наличие связи между переменными X и Y, то используется статистика $\dfrac{b''}{S_b}$, тем самым проверяется равенство нулю коэффициента b. Если в границы построенного при этом доверительного интервала попадает ноль, (то есть нижняя граница доверительного интервала отрицательна, а верхняя положительна), то коэффициент b принимается равным нулю и делается вывод об отсутствии связи между

переменными X и Y. Другими словами, при $\left|\frac{b''}{S_b}\right| > \hat{t}_{n-2,\alpha}$ делается вывод о достоверной связи между переменными X и Y, при $-t_{n-2,\alpha} \le \frac{b''}{S_b} \le t_{n-2,\alpha}$ делается вывод об ее отсутствии.

Можно показать, что

$$\frac{a'' - a}{S_a} \in t_{n-2}$$

и использовать эту статистику для проверки аналогичных гипотез относительно коэффициента a.

Рассмотрим статистику $\frac{Y-(a+bX)}{\sigma}$, которая принадлежит стандартному нормальному распределению - $N(0,1)$. При известной σ^2 (дисперсии ошибок) можно было бы использовать $N(0,1)$ для прогнозирования значений Y в виде доверительных интервалов.

Поскольку σ^2 неизвестно, то будем использовать ее оценку S^2, для которой известно, что $\frac{(n-2)S^2}{\sigma^2} \in \chi^2_{n-2}$.

Таким образом,

$$\frac{\frac{Y-(a+bX)}{\sigma}}{\frac{S}{\sigma}} = \frac{Y-(a+bX)}{S} \in t_{n-2}$$

и используется для построения доверительных интервалов с целью прогнозирования значений Y:

$$a + bX - t_{n-2,\alpha} S \le Y \le a + bX + t_{n-2,\alpha} S .$$

Количественным показателем качества построенной линейной модели является коэффициент детерминации

$$R^2 = 1 - \frac{\sum_{i-1}^{n} (\tilde{y}_i - y_i)^2}{\sum_{i-1}^{n} (y_i - \bar{y})^2}.$$

Коэффициент детерминации показывает, какая доля общей дисперсии Y объясняется уравнением регрессии.

$$0 \le R^2 \le 1.$$

Чем ближе R^2 к 1, тем лучше построенная регрессионная модель согласуется с исходными данными.

Для построения регрессии в Excel, создаем файл исходных входных и выходных данных и начинаем с построения корреляционного поля, позволяющего визуализировать наличие связи между этими данными. Выбираем меню **ВСТАВКА/ ДИАГРАММА**, тип диаграммы: **ТОЧЕЧНАЯ** вид: **ТОЧЕЧНАЯ ДИАГРАММА.** Нажимаем кнопку **ДАЛЕЕ**. В появившемся диалоговом окне указываем диапазон значений и расположение данных: **В СТОЛБЦАХ**. Нажимаем кнопку **ДАЛЕЕ**. В следующем диалоговом окне указываем название диаграммы, наименование осей. Нажимаем **ДАЛЕЕ** и **ГОТОВО**. Построенная таким образом диаграмма рассеяния представляет собой совокупность пар точек, абсциссами которых являются значения переменной X, а ординатами значения переменной Y.

В меню **СЕРВИС** выбираем **АНАЛИЗ ДАННЫХ** и **РЕГРЕССИЯ.** Указываем входной интервал Y (для примера A2: A26) и входной интервал X (для примера B2: B26), а также параметры вывода, остатки, нормальную вероятность как показано на рис. 7.

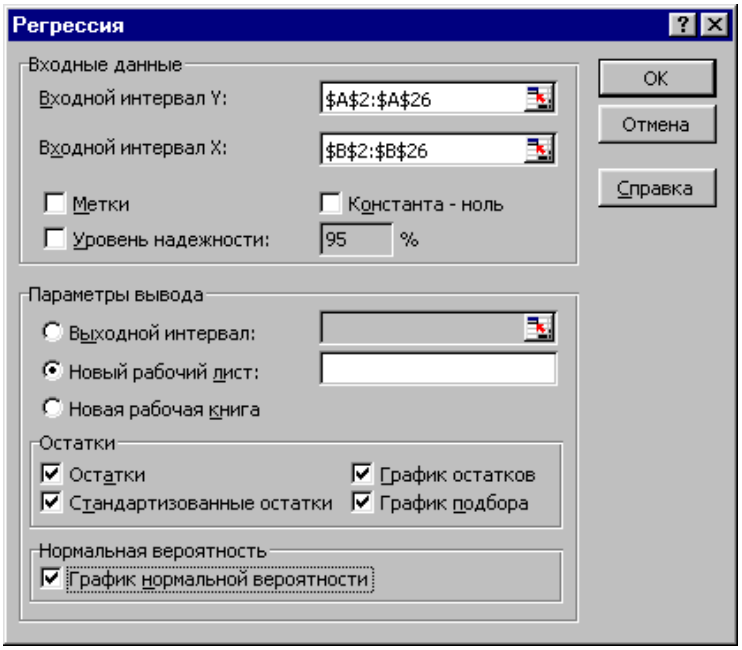

Рис. 7. Диалоговое окно **Регрессия**.

В диалоговом окне задаются следующие параметры:

Входной интервал Y – диапазон ячеек, содержащий данные результативного признака;

Входной интервал X – диапазон ячеек, содержащий данные факторного признака;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет;

Константа-ноль – данный флажок необходимо установить, чтобы линия регрессии прошла через начало координат;

Уровень надежности – этот флажок необходимо использовать, если требуется уровень надежности отличный от 95%, принятый по умолчанию;

Выходной интервал – верхняя левая ячейка интервала, в который будут помещаться результаты вычислений.

Excel автоматически сгенерирует результаты по регрессионной статистике. Ниже в качестве примеров приведены возможные результаты и их расшифровки. !

Регрессионная статистика

Множественный R 0,969525973

R-квадрат 0,939980612

Нормированный R-квадрат 0,935363736

Стандартная ошибка 14,22893673

Наблюдения 15.

Полученное значение коэффициента детерминации говорит об очень хорошей согласованности построенной регрессионной модели и исходных данных (соответственно об очень хорошей связи исследуемых факторов X и Y).

Результаты дисперсионного анализа будут представлены в виде:

Дисперсионный анализ					
	df	***SS***	***MS***	***F***	***Значимость F***
Регрессия	1	41220,72106	41220,72106	203,5966782	2,55346E-09
Остаток	13	2632,014326	202,4626405		
Итого	14	43852,73538			

	Коэффициенты	***Стандартная ошибка***	***t-статистика***	***P-Значение***	***Нижние 95%***	***Верхние 95%***
Y-пересечение	4,746	7,003	0,678	0,510	-10,384	19,876
Переменная X 1	9,595	0,672	14,269	0,000	8,142	11,048

df – степени свободы (degree of freedom);

SS – сумма квадратов отклонений (Sum of squares);

MS – средний квадрат отклонения (Mean square);

F – отношение дисперсий (факторной к остаточной)

$$F = \frac{S^2_{факт}}{S^2_{ост}} = \frac{\sum_{i-1}^{n}\left(\tilde{y}_i - \overline{y}\right)^2}{\sum_{i-1}^{n}\left(y_i - \tilde{y}_i\right)^2}.$$

Значимость *F* – критическое значение квантиля распределения Фишера, которое используется для проверки нулевой гипотезы, состоящей в том, что факторная и остаточная дисперсии равны. По сути дела нулевая гипотеза означает, что на результативный признак Y в равной степени влияют и независимая (факторная) переменная X и необъясненные факторы. В таком случае ***уравнение регрессии не значимо.*** Чтобы уравнение регрессии было значимым необходимо, чтобы факторная дисперсия превышала остаточную дисперсию в несколько раз.

В примере, приведенном выше, ***F*** больше, чем **Значимость *F*** (критическое значение), значит регрессионная модель адекватна. Регрессионная сумма ***SS=***41220,72106 (объясненная регрессией) намного больше остаточной ***SS=***2632,014326 (не объясненной регрессией, вызванной случайными факторами), что тоже говорит о хорошей регрессии.

Коэффициенты – значения коэффициентов;

Стандартная ошибка – стандартная ошибка коэффициентов;

t-статистика – значение статистики критерия;

P–значение – уровень значимости отклонения гипотезы равенства коэффициента нулю (вероятность принять равенство коэффициента нулю);

Нижние 95% - нижняя граница доверительного интервала, в котором находится значение коэффициента;

Верхние 95% - верхняя граница доверительного интервала, в котором находится значение коэффициента.

Приведенные в качестве примера результаты позволяют проверить значимость коэффициентов регрессии: свободного члена и коэффициента при переменной X. Значение коэффициента при X 9,595 больше, чем его стандартная ошибка. К тому же этот коэффициент является значимым, о

чем можно судить по значениям показателя ***P-значение*** в таблице, которые меньше заданного уровня значимости α=0,05. Для свободного члена ситуация диаметрально противоположная. В построенный для него доверительный интервал попадает ноль, что говорит о том, что он незначим и может быть принят равным нулю.

Есть возможность вывести таблицу стандартных и простых остатков, где для каждого значения ряда выводится предсказанное значение, с которым сопоставляется остаток, представляющий разность между прогнозным и реальным значением.

Простым и наглядным способом проверки удовлетворительности регрессионной модели является графическое представление отклонений, которое Excel представляет в виде графика остатков. Если регрессионная модель близка к реальной зависимости, то отклонения будут носить случайный характер и их сумма будет близка! к нулю. Если необходимо получить дополнительную информацию и графики остатков, установите соответствующие флажки в диалоговом окне.

Задания

Построить уравнение регрессии $Y = a + bx$.

1. **2.** **3.**

№	X	Y	№	X	Y	№	X	Y
1	-1.132	1.554	1	-0.132	1.791	1	-0.332	-1.65
2	-0.204	4.601	2	0.796	1.51	2	0.596	2.31
3	0.858	2.943	3	1.858	4.17	3	1.658	4.953
4	1.715	-1.157	4	2.715	3.007	4	2.515	8.62
5	2.494	-6.048	5	3.494	5.875	5	3.294	7.49
6	4.013	-1.194	6	5.013	7.187	6	4.813	6.09
7	4.964	11.465	7	5.964	9.005	7	5.764	11.958
8	6.167	-6.257	8	7.167	14.865	8	6.967	14.975
9	7.658	-11.07	9	8.658	12.008	9	8.458	18.518
10	8.243	10.243	10	9.243	11.718	10	9.043	21.794
11	9.296	10.995	11	10.296	16.744	11	10.096	17.245
12	10.259	-11.17	12	11.259	18.789	12	11.059	20.881
13	11.275	-10.84	13	12.275	12.863	13	12.075	18.787
14	12.202	-9.78	14	13.202	20.862	14	13.002	16.334

15	12.687	15.066	15	13.687	15.309	15	13.487	28.613

4. **5.** **6.**

№	X	Y	№	X	Y	№	X	Y
1	-1.118	-2.972	1	-2.132	1.115	1	-3.132	1.902
2	-1.062	0.729	2	-1.204	4.44	2	-2.204	5.319
3	0.054	-8.757	3	-0.142	3.101	3	-1.142	4.086
4	2.359	7.554	4	0.715	-0.742	4	-0.285	0.329
5	3.561	5.017	5	1.494	-5.4	5	0.494	-4.25
6	3.56	1.466	6	3.013	-0.09	6	2.013	1.211
7	6.348	17.308	7	3.964	10.076	7	2.964	-8.679
8	6.617	7.144	8	5.167	-4.507	8	4.167	-2.99
9	8.108	10.657	9	6.658	-8.872	9	5.658	-7.207
10	8.538	27.801	10	7.243	-7.87	10	6.243	-6.146
11	8.312	18.64	11	8.296	-8.307	11	7.296	-6.477
12	10.895	28.057	12	9.259	-8.192	12	8.259	-6.266
13	10.381	23.008	13	10.275	-7.557	13	9.275	-5.53
14	12.18	28.91	14	11.202	-6.22	14	10.202	-4.099
15	12.973	24.253	15	11.687	-11.36	15	10.687	-9.191

7. **8.** **9.**

№	X	Y	№	X	Y	№	X	Y
1	-3.132	0.412	1	1.868	12.669	1	-4.132	-4.539
2	-2.204	2.204	2	2.796	-9.23	2	-3.204	-3.306
3	-1.142	-1.282	3	3.858	0.753	3	-2.142	-2.018
4	-0.285	-2.529	4	4.715	-2.106	4	-1.285	-1.223
5	0.494	-2.995	5	5.494	-4.044	5	-0.506	-0.739
6	2.013	-0.325	6	7.013	1.903	6	1.013	-0.823
7	2.964	-5.018	7	7.964	5.065	7	1.964	-1.873
8	4.167	-6.295	8	9.167	3.807	8	3.167	1.055
9	5.658	-3.773	9	10.658	3.374	9	4.658	0.989
10	6.243	-6.772	10	11.243	0.563	10	5.243	0.116
11	7.296	-5.079	11	12.296	2.511	11	6.296	6.213
12	8.259	-5.519	12	13.259	1.934	12	7.259	0.856
13	9.275	-6.505	13	14.275	4.609	13	8.275	0.743
14	10.202	-7.864	14	15.202	9.673	14	9.202	5.137
15	10.687	-8.578	15	15.687	6.565	15	9.687	2.209

10. **11.** **12.**

№	X	Y	№	X	Y	№	X	Y
1	1.868	4.322	1	-1.118	7.252	1	1.132	2.573
2	2.796	11.126	2	-1.062	10.683	2	3.204	14.963
3	3.858	4.536	3	0.054	4.075	3	5.142	4.317
4	4.715	11.52	4	2.359	5.463	4	7.285	17.39
5	5.494	15.855	5	3.561	11.377	5	9.506	26.456

6	7.013	19.365	6	3.56	13.401	6	10.987	30.205
7	7.964	20.027	7	6.348	12.638	7	13.036	32.238
8	9.167	15.986	8	6.617	18.804	8	14.833	25.086
9	10.658	23.784	9	8.108	8.975	9	16.342	36.133
10	11.243	22.948	10	8.538	26.612	10	18.757	37.62
11	12.296	18.02	11	8.312	23.302	11	20.704	29.791
12	13.259	31.75	12	10.895	21.718	12	22.741	53.538
13	14.275	29.538	13	10.381	20.979	13	24.725	50.432
14	15.202	24.758	14	12.18	17.243	14	26.798	43.52
15	15.687	27.018	15	12.973	19.821	15	29.313	50.701

13. **14.** **15.**

№	X	Y	№	X	Y	№	X	Y
1	-1.121	-4.575	1	-0.121	18.34	1	-1.118	-3.108
2	0.391	3.839	2	2.391	4.469	2	-1.062	8.195
3	0.587	1.864	3	3.587	11.81	3	0.054	0.521
4	2.114	2.832	4	6.114	16.8	4	2.359	4.907
5	3.131	17.286	5	8.131	34.17	5	3.561	-6.499
6	4.528	7.376	6	10.52	34.84	6	3.56	1.929
7	4.806	6.239	7	11.80	-0.329	7	6.348	-4.912
8	6.165	17.959	8	14.16	33.95	8	6.617	-4.048
9	7.464	13.944	9	16.46	27.52	9	8.108	-11.76
10	7.454	17.99	10	17.45	29.17	10	8.538	-6.738
11	9.392	27.978	11	20.39	40.67	11	8.312	-11.246
12	9.685	22.938	12	21.68	35.38	12	10.89	-10.066
13	11.138	25.206	13	24.13	50.74	13	10.38	-16.524
14	11.684	26.74	14	25.68	41.61	14	12.18	-19.811
15	12.627	36.957	15	27.62	73.25	15	12.97	-23.647

ЛАБОРАТОРНАЯ РАБОТА № 9
ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Цель: С помощью однофакторного дисперсионного анализа научиться изучать влияние некоторого фактора на исследуемый признак.

Однофакторный дисперсионный анализ используется в тех случаях, когда рассматриваются три и более независимые выборки (группы данных), которые формируются на основе группировочного фактора.

Например, изучается работа некоторого оборудования в разных климатических условиях. В этом случае группировочным фактором является климат, влияние которого изучается на работу оборудования (исследуемый признак). Более простым аналогом однофакторного дисперсионного анализа является критерий Стьюдента, который рассматривался в лабораторной работе № 7. Суть однофакторного дисперсионного анализа состоит в изучении компонент общей дисперсии, которая складывается из межгрупповой дисперсии (факторной), обусловленной различием средних значений в группах, и внутригрупповой дисперсии, обусловленной случайными факторами. Если межгрупповая дисперсия вносит значительный вклад в общую дисперсию и соответственно в разы превосходит внутригрупповую дисперсию (что говорит о значительном разбросе групповых средних значений), то логично сделать вывод о существенном влиянии группировочного фактора.

Проводя анализ, мы фактически проверяем гипотезу о равенстве внутригрупповых выборочных средних. Если эта гипотеза верна, то считается, что выборки принадлежат одной и той же генеральной совокупности, и, соответственно, отсутствует влияние группировочного фактора на исследуемый признак. Если гипотеза неверна, то делается обратный вывод. Для того, чтобы принять или опровергнуть основную гипотезу, строится критерий Фишера.

Предположим, что рассматривается k выборок ($n = \sum_{i=1}^{k} n_i$ - суммарный объем выборок)

$$X_{11}, X_{12}, ..., X_{1n_1}$$

$$X_{21}, X_{22}, ..., X_{2n_2}$$

……..

$$X_{k1}, X_{k2}, \ldots, X_{kn_k},$$

Выборки сформированы, исходя из некоторого фактора, влияние которого и будет изучаться на признак, отраженный в генеральной совокупности.

Например, рассматривается совокупность программистов, работающих в крупных банках и имеющих одинаковое образование. Эта совокупность разбивается на выборки, число которых соответствует числу банков. Основная гипотеза будет состоять в том, что на зарплату программиста не влияет место работы. Таким образом, фактор – место работы, исследуемый признак – зарплата. Задачей однофакторного дисперсионного анализа будет являться изучение этого фактора на исследуемый признак.

Из самого названия анализа уже понятно, что будут изучаться дисперсии, а именно – общая дисперсия, межгрупповая и средняя внутригрупповая.

Выборочное среднее по всей совокупности - $\overline{X}$ и общая выборочная дисперсия - σ^2 (оценка общей дисперсии σ^2) определяются соответственно по формулам

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij},\ \sigma^2 = \frac{\sum_{i=1}^{k}\sum_{j-1}^{n_i}\left(X_{ij} - \overline{X}\right)^2}{n}.$$

Межгрупповая выборочная дисперсия $\sigma^2_{beweent}$ (оценка межгрупповой дисперсии $\sigma^2_{beweent}$) определяется по формуле

$$\sigma^2_{beweent} = \frac{\sum_{i=1}^{k}\left(\overline{X_i} - \overline{X}\right)^2 n_i}{n},$$

где $\overline{X_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, , $i = \overline{1,k}$ - внутригрупповые выборочные средние.

Средняя внутригрупповая выборочная дисперсия σ^2_{within} (оценка средней внутригрупповой дисперсии σ^2_{within}) определяется по формуле

$$\sigma^2_{within} = \frac{\sum_{i=1}^{k} \sigma_i^2 n_i}{n},$$

где $\sigma_i^2 = \frac{\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X_i}\right)^2}{n_i}, i = \overline{1,k}$ - внутригрупповые выборочные дисперсии.

Основная формула для дисперсий

$$\sigma^2 = \sigma^2_{between} + \sigma^2_{within}$$

Эмпирическим коэффициентом детерминации называется

$$\eta^2 = \frac{\sigma^2_{between}}{\sigma^2}.$$

Чем ближе коэффициент к 1, тем существеннее влияние группировочного фактора на исследуемый признак. Чем ближе коэффициент к 0, тем меньше это влияние.

Из общего курса математической статистики известно, что $\sigma^2, \sigma^2_{between}, \sigma^2_{within}$ являются смещенными (асимптотически несмещенными) оценками теоретических дисперсий. Несмещенные оценки теоретических дисперсий $\sigma^2, \sigma^2_{between}, \sigma^2_{within}$ получаются делением соответствующих сумм на число степеней свободы, то есть

$$S^2 = \frac{\sum_{i=1}^{k}\sum_{j-1}^{n_i} \left(X_{ij} - \overline{X}\right)^2}{n-1}, \; S^2_{beweent} = \frac{\sum_{i=1}^{k} \left(\overline{X_i} - \overline{X}\right)^2 n_i}{k-1}, \; S^2_{within} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X_i}\right)^2}{n-k}.$$

Также известно, что $\dfrac{\dfrac{\sum_{i=1}^{k}\left(\overline{X_i}-\overline{X}\right)^2 n_i}{\sigma^2_{between}(k-1)}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X_i}\right)^2}{\sigma^2_{within}(n-k)}} = \dfrac{\dfrac{S^2_{beweent}}{\sigma^2_{between}}}{\dfrac{S^2_{within}}{\sigma^2_{within}}} = \dfrac{S^2_{beweent}}{\lambda S^2_{within}} = F_{k-1,n-k}$

(распределение Фишера), $\lambda = \dfrac{\sigma^2_{between}}{\sigma^2_{within}}$.

Как уже было сказано, проводя однофакторный дисперсионный анализ, мы проверяем гипотезу о равенстве внутригрупповых средних. Для этого мы изучаем вклад межгрупповой дисперсии в общую дисперсию. Межгрупповая дисперсия показывает разброс внутригрупповых средних вокруг общего среднего. Если внутригрупповые средние отличаются, то разброс будет большим, соответственно увеличивается вклад $\sigma^2_{between}$ в общую дисперсию. Другими словами, если $\sigma^2_{between} \le \sigma^2_{within}$, то мы принимаем гипотезу об отсутствии существенного влияния группировочного фактора. Если $\sigma^2_{between} > \sigma^2_{within}$, то мы признаем существенное влияние группировочного фактора. Итак,

H_0: Отсутствие существенного влияния группировочного фактора на исследуемый признак ($\sigma^2_{between} \le \sigma^2_{within},\ \lambda \le 1$).

H_1: Наличие существенного влияния группировочного фактора на исследуемый признак ($\sigma^2_{between} > \sigma^2_{within}, \lambda > 1$).

$$P\left(\frac{S^2_{beweent}}{\lambda S^2_{within}} \le \Delta\right) = 1-\alpha.$$

$$\frac{S^2_{beweent}}{S^2_{within}} \le \lambda\Delta,\ P\left(F_{k-1,n-k} > \Delta\right) = \alpha,$$

Если $\frac{S^2_{beweent}}{S^2_{within}} \le \Delta$ **(область принятия основной гипотезы), то принимаем гипотезу** H_0**: Отсутствие существенного влияния группировочного фактора на исследуемый признак.**

Если $\frac{S^2_{beweent}}{S^2_{within}} > \Delta$ **(критическая область), то принимаем гипотезу** H_1**: Наличие существенного влияния группировочного фактора на исследуемый признак.**

Для выполнения лабораторной работы создаем файл с исходными данными, выбираем **АНАЛИЗ ДАННЫХ** и **ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ** (Рис. 8).

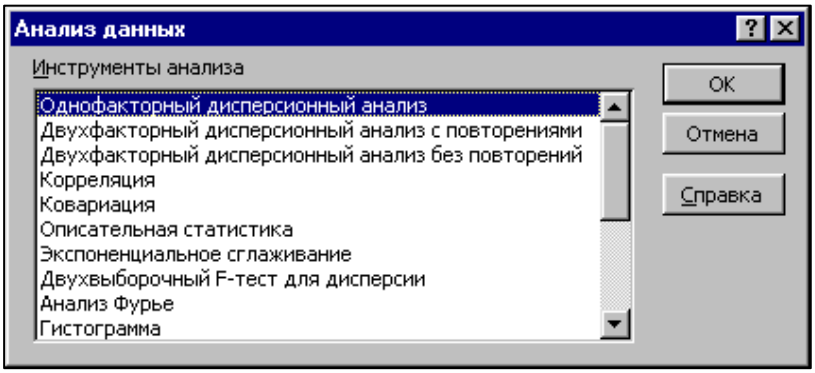

Рис. 8. Выбор инструмента анализа.

В диалоговом окне режима (Рис. 9) указываем входной интервал, способ группирования, выходной интервал, метки в первой строке/ Метки в первом столбце, альфа (уровень значимости).

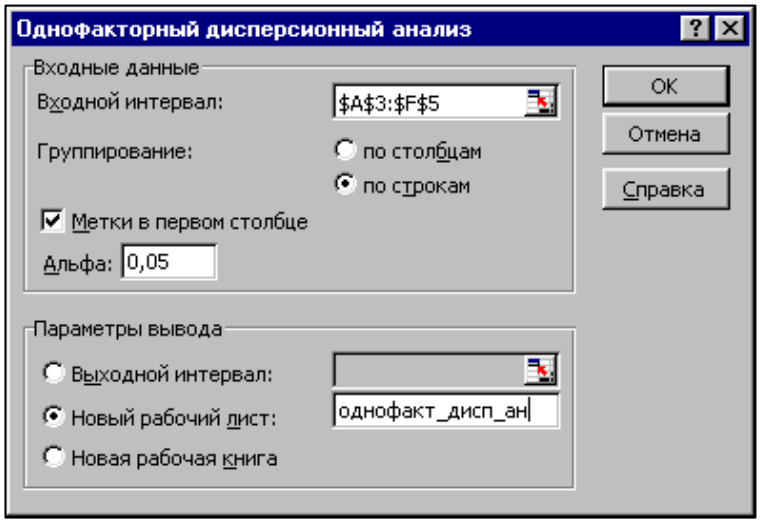

Рис. 9. Диалоговое окно однофакторного дисперсионного анализа.

Входной диапазон – это ссылка на ячейки, содержащие анализируемые данные. Ссылка должна состоять как минимум из трех смежных диапазонов данных, организованных в виде столбцов или строк.

Группирование. Установите переключатель в положение «по столбцам» или «по строкам» в зависимости от расположения данных во входном диапазоне.

Метки в первой строке/ Метки в первом столбце. Установите переключатель в положение «Метки в первой строке», если первая строка во входном диапазоне содержит названия столбцов. Установите переключатель в положение «Метки в первом столбце», если названия строк находятся в первом столбце входного диапазона. Если входной диапазон не содержит меток, то необходимые заголовки в выходном диапазоне будут созданы автоматически.

Новый лист. Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга. Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Выходной диапазон. Введите ссылку на ячейку, с которой будут показаны результаты анализа. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Нажать кнопку ОК.

Пример полученных результатов представлен на рис. 10.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	I группа (контр.)	5	1673	334,6	56,8		
6	II группа	5	1812	362,4	220,8		
7	III группа	5	1885	377	276,5		
8							
9	ANOVA						
10	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
11	Между группами	4640	2	2319,8	12,55983	0,0011415	3,885290312
12	Внутри групп	2216	12	184,7			
13							
14	Итого	6856	14				
15							
16							
17							

однофакт_дисп_ан | Лист1 | Лист2 | Лист3 | Лист4 | Лист5

Рис. 10. Результаты однофакторного дисперсионного анализа.

Результаты содержат выборочные средние и дисперсии по группам. Таблица ANOVA содержит суммы квадратов отклонений (SS), число степеней свободы (df), средние суммы квадратов отклонений (MS), значение критерия F ($\frac{S^2_{beweent}}{S^2_{within}}$), $P-Значение$ (определяет вероятность того, что полученная взаимосвязь между фактором и результатом может считаться случайной). Если $P-Значение < \alpha = 0{,}05$, то исследуемый (группировочный) фактор статистически значим. F критическое (Δ)

позволяет сделать вывод о наличии или отсутствии существенного влияния группировочного фактора на исследуемый признак..

В приведенном выше примере F=12,55893 > F критическое=3,885290312, что говорит о наличии существенного влияния группировочного фактора на исследуемый признак. $P-Значение$ =0,0011415< 0,05, что говорит о статистической значимости группировочного фактора.

Рекомендуемая литература

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2005. – 479 с. : ил.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 2005. – 404 с. : ил.

3. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. – М.: Высшая школа, 2000. – 480 с.

4. О.М.Полещук Основы теории вероятностей и математической статистики: учеб. пособие.-М: ГОУ ВПО МГУЛ, 2007.-140 с. : ил.

5. О.М.Полещук Основы теории вероятностей, математической статистики и случайных процессов: Учебное пособие для студентов всех специальностей МГУЛ. – М.: ФГБОУ ВПО МГУЛ, 2012. – 256 с.: ил.