СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ТАБЛИЧНОМ ПРОЦЕССОРЕ MS EXCEL: АНАЛИЗ И СРАВНЕНИЕ ВЫБОРОК

1.1 Средства для решения задач статистического анализа данных в MS Excel. Надстройка «Анализ данных»

Для решения многих распространенных задач статистического анализа данных (вычисление статистических показателей, оценка различия между выборками, корреляционный анализ и т.д.) в MS Excel имеется специальный инструмент – надстройка **Анализ данных**. Кроме того, имеется большой набор функций для вычислений (математические, статистические и т.д.).

Чтобы иметь возможность пользоваться надстройкой **Анализ данных**, ее необходимо *активировать*. Для этого требуется выполнить следующее.

1 Выбрать **Файл – Параметры**.

2 В появившемся окне **Параметры Excel** выбрать **Надстройки**.

3 Убедиться, что в поле **Управление** выбрано **Надстройки Excel**. Нажать кнопку **Перейти**, расположенную рядом с этим полем.

4 В появившемся окне **Надстройки** установить флажок **Пакет анализа** (для интерфейса на английском языке – **Analysis Toolpak**). Нажать **OK**.

5 Убедиться, что на вкладке **Данные** имеется кнопка **Анализ данных** (обычно – в правой части ленты).

1.2 Построение таблиц частот и гистограмм

Для построения таблиц частот и гистограмм на основе выборки в Excel применяется инструмент **Гистограмма** из пакета **Анализ данных**.

Пример 1 – В ходе социологического исследования получены данные (выборка) о возрасте 25 работников государственных предприятий:

32	48	62	54	59	31	29	47	42	37	51	52	43
47	61	54	53	41	42	36	48	41	49	57	46	

Построить таблицу частот и гистограмму, т.е. подсчитать количество работников в возрасте до 20 лет, от 20 до 30, от 30 до 40, от 40 до 50, от 50 до 60, от 60 до 70, старше 70 лет.

1 В ячейку A1 ввести заголовок, например, «Работники государственных предприятий». Чтобы заголовок состоял из нескольких строк и имел удобный вид, настроить формат ячейки следующим образом:

- выбрать ячейку A1;
- из меню **Формат** выбрать элемент **Ячейки**;
- перейти на вкладку **Выравнивание**. Установить: **По горизонтали – По центру, По вертикали – По центру**. Установить флажок **Переносить по словам**. Нажать **OK**;
- отрегулировать ширину столбца.

2 В ячейки A2–A26 ввести данные (значения возраста).

3 В ячейку B1 ввести заголовок, например, «Возраст». В ячейки B2–B7 ввести границы интервалов гистограммы: 20, 30, 40, 50, 60, 70.

4 Получить таблицу частот и гистограмму. Для этого выбрать элемент меню **Данные – Анализ данных**. Из появившегося списка выбрать инструмент **Гистограмма**. В появившемся окне выполнить следующую настройку:

– указать диапазоны ячеек, где находятся данные для построения гистограммы: **Входной интервал** – **A1:A26**, **Интервал карманов** – **B1:B7**;

Примечание – Чтобы указать диапазон ячеек, можно провести по соответствующим ячейкам указателем мыши или просто набрать обозначения ячеек на клавиатуре. Следует обратить внимание, что если обозначения ячеек набираются на клавиатуре, то можно использовать только латинские буквы (но не русские).

– установить флажок **Метки**, так как в первых ячейках диапазонов данных (т.е. в ячейках A1 и B1) указаны не данные, а заголовки;
– установить флажок **Вывод графика** (для построения гистограммы);
– в области **Параметры вывода** указать, куда требуется вывести результаты. Если установить переключатель **Новый рабочий лист**, то будет создан новый рабочий лист, и на него будут выведены результаты. Если установить переключатель **Выходной интервал**, а в поле рядом с ним указать ячейку, то результаты будут выведены на текущий рабочий лист, начиная с заданной ячейки;
– для получения результатов нажать кнопку **OK**.

1.3 Вычисление статистических показателей

Для вычисления отдельных статистических показателей применяются имеющиеся в Excel статистические функции, а для вычисления всех статистических показателей сразу – инструмент **Описательная статистика** из пакета **Анализ данных**.

Пример 2 – Используя функции MS Excel, по данным примера 1 найти основные статистические показатели для возраста работников: среднее, стандартное отклонение, дисперсию.

Примечание– Формулы для расчета этих величин следующие:

Среднее: $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Стандартное отклонение: $$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Дисперсия: $$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

1 Перейти на рабочий лист с исходными данными, введенными для примера 1.

2 В ячейке C1 (или в любой другой свободной ячейке) ввести надпись «Среднее».

3 В ячейке D1 вычислить среднее. Для этого выбрать **Формулы – Вставить функцию.** Выбрать категорию **Статистические** (или **Полный алфавитный перечень**). Выбрать функцию **СРЗНАЧ**. В появившемся окне функции **СРЗНАЧ** в поле **Число1** ввести A2:A26, т.е. диапазон ячеек, для которых требуется вычислить среднее. Нажать **ОК**. Вычисляется среднее. Для данного примера $\overline{X}$ =46.48 года.

4 Аналогично в любых свободных ячейках вычислить стандартное отклонение и дисперсию, используя функции **СТАНДОТКЛОН.В** и **ДИСП.В**.

Примечание – Буква «В» в именах функций **СТАНДОТКЛОН.В** и **ДИСП.В** означает, что вычисления выполняются по выборке.

Примечание – Стандартное отклонение и дисперсия – показатели *изменчивости (разброса)* исследуемой величины. Другими словами, чем больше стандартное отклонение и дисперсия, тем более различны отдельные значения исследуемой величины (в данном примере – значения возраста работников).

Примечание – Применяются также некоторые другие статистические показатели. Например, *мода* вычисляется как середина интервала таблицы частот, где частота максимальна. В Данном примере мода составляет M_o = 45 лет. *Медиана* – «центральное» значение в выборке, упорядоченной по возрастанию:

29	31	32	36	37	41	41	42	42	43	46	47	47
48	48	49	51	52	53	54	54	57	59	61	62	

В данном примере медиана M_e = 47 лет.

Пример 3 – Для выборки из примера 1 вычислить стандартную ошибку для среднего. На ее основе вычислить показатель точности.

Формула стандартной ошибки для среднего:

$$S_{\overline{X}} = \frac{\sigma}{\sqrt{n}},$$

где σ – стандартное отклонение, n – объем выборки (в данном примере n=25).

Формула показателя точности:

$$A = \frac{S_{\overline{X}}}{\overline{X}} \cdot 100.$$

Для данного примера $S_{\overline{X}}$ = 1.83 года, A = 3.93%. Значение A < 3% соответствует очень высокой точности, A = 3…5% – приемлемая точность. Если A > 5%, то точность недостаточна, и следует увеличить объем выборки. В данном примере точность приемлемая.

Пример 4 – Для выборки из примера 1 вычислить доверительный интервал для среднего с доверительной вероятностью 0,95 (или 95%).

Формула для вычисления доверительного интервала:

$$\overline{X} \pm t_{\alpha;n-1} \cdot S_{\overline{X}},$$

где $\overline{X}$ – среднее, $S_{\overline{X}}$ – стандартная ошибка для среднего, $t_{\alpha;n-1}$ – квантиль распределения Стьюдента. Здесь α – уровень значимости: если доверительная вероятность равна 0,95, то α = 0,05. Величина n-1 – число степеней свободы (для данного примера n-1=24).

Чтобы найти $t_{\alpha;n-1}$, используют специальные статистические таблицы или функцию MS Excel **СТЬЮДЕНТ.ОБР.2Х** со следующими параметрами: **Вероятность**: 0,05; **Степени свободы**: 24. В данном примере $t_{\alpha;n-1} = t_{0,05;24}$=2,0639.

Доверительный интервал: 46,48 ± 2,0639·1,83 = (42,71; 50,25). Это означает, что с вероятностью 0,95 средний возраст в *генеральной совокупности* (т.е. во всей исследуемой категории работников) находится в диапазоне от 42,71 до 50,25 лет. При этом 0,05 – вероятность того, что средний возраст в генеральной совокупности составляет менее 42,71 или превышает 50,25 года.

Пример 5 – Для выборки из примера 1 найти статистические показатели, используя инструмент **Описательная статистика**.

1 Выбрать элемент меню **Данные – Анализ данных**. Из списка инструментов выбрать инструмент **Описательная статистика**.

2 В появившемся окне ввести необходимые параметры:

- указать ячейки с исходными данными: **Входной интервал** – A1:A26, **Группирование – По столбцам**;
- установить флажок **Метки в первой строке**, так как в первой ячейке диапазона данных (т.е. в ячейке A1) указаны не данные, а заголовок;
- установить флажки **Итоговая статистика** (для получения всех статистических показателей) и **Уровень надежности** (для получения данных, необходимых для вычисления доверительного интервала). В поле **Уровень надежности** указать доверительную вероятность в процентах: **95**;
- в области **Параметры вывода** указать, куда требуется вывести результаты, как показано в примере 1;
- для получения результатов нажать кнопку **ОК**. Вычисляются статистические показатели.

3 По выходным данным инструмента **Описательная статистика** вычислить границы доверительного интервала. Величина, обозначенная как **Уровень надежности** – это величина $t_{\alpha;n-1} \cdot S_{\bar{x}}$ (см. вычисление доверительного интервала в примере 4). Таким образом, границы доверительного интервала вычисляют как **Среднее** ± **Уровень надежности**.

Задание

Имеется выборка значений прочности образцов некоторого материала:

68	75	34	92	57	29	38	49	84	68	71	42
67	73	37	51	97	42	83	43	54	68	67	

Для этой выборки:

- построить гистограмму и таблицу частот (указание: используя функции **МИН** и **МАКС**, найти диапазон значений выборки, затем выбрать интервалы для построения гистограммы);

– вычислить статистические показатели: среднее, стандартное отклонение, дисперсию, стандартную ошибку для среднего, 95-процентный доверительный интервал для среднего. Решить эту задачу, используя функции MS Excel и инструмент **Описательная статистика**. Убедиться, что результаты одинаковы.

1.4 Оценка значимости различий между независимыми выборками

Для решения задач, связанных с оценкой статистической значимости различий между независимыми выборками, применяются инструменты, входящие в состав пакета **Анализ данных**: **Двухвыборочный F-тест для дисперсии, Двухвыборочный t-тест с одинаковыми дисперсиями, Двухвыборочный t-тест с различными дисперсиями**.

Пример 6 – Анализируются данные о работниках государственных и негосударственных предприятий. Имеются значения возраста 25 работников государственных предприятий (см. пример 1) и 19 работников негосударственных предприятий:

48	29	37	32	47	52	38	34	41	32
47	53	42	35	37	29	51	42	53	

Определить, является ли статистически значимым различие в среднем возрасте этих двух категорий работников. Другими словами, требуется определить, можно ли утверждать, что работники одной из исследуемых категорий значительно старше, чем другой.

Задача решается в два этапа. Сначала требуется определить, является ли статистически значимым различие между *дисперсиями*. Затем, с учетом результата первого этапа, требуется определить, значимо ли различие между *средними*.

Для решения задачи в MS Excel введем в ячейки A1 и B1 какие-либо заголовки (например, «Государственные предприятия» и «Негосударственные предприятия»), а в ячейки A2:A26 и B2:B20 – сами выборки.

Алгоритм решения задачи

а) Сравнение дисперсий на основе F-критерия (критерия Фишера)

Вычисляют *F-статистику*:

$$F = \frac{\sigma_1^2}{\sigma_2^2},$$

где σ_1^2, σ_2^2 – дисперсии, вычисленные по выборкам. Для вычисления F делят большую дисперсию на меньшую, т.е. $\sigma_1^2 > \sigma_2^2$.

Для данного примера σ_1^2 = 83,51, σ_2^2 = 67,11 (вычислены с использованием функции **ДИСП.В**), F = 1.24.

Затем по значению F и числам степеней свободы n_1-1 и n_2-1 определяют расчетный *уровень значимости P*. Для этого используют функцию **F.РАСП.ПХ** со следующими параметрами: **X**: F; **Степени свободы 1**: n_1-1; and **Степени свободы 2:** n_2-1 (для данного примера n_1-1 = 24, n_2-1 = 18). Получим P=0,32. Упрощенно говоря, эта величина – вероятность того, что различие между дисперсиями *незначимо*. Если $P < \alpha$ (где α – заданный уровень значимости, обычно α = 0,05), то различие между дисперсиями признают статистически *значимым*, так как вероятность его незначимости (P) очень мала. В данном примере $P > \alpha$, поэтому различие между дисперсиями следует признать *незначимым* (недостаточно оснований для отклонения гипотезы о незначимости). Это означает, что разброс возрастов в рассматриваемых категориях работников различается незначительно.

Примечание – Данное объяснение – упрощенное. Более строгое математическое описание см. в литературе по математической статистике.

б) Сравнение средних на основе t-критерия (критерий Стьюдента)

Вычисляют *t-статистику*. Для этого используют разные формулы в зависимости от результата первого этапа, т.е. от того, значимо ли различие между дисперсиями.

Если различие между дисперсиями незначимо:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1 \cdot n_2}}}.$$

Если различие между дисперсиями значимо:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

где n_1, n_2 – объемы выборок;

$\overline{X}_1, \overline{X}_2$ – средние (определяются с помощью функции **СРЗНАЧ**);

σ_1^2, σ_2^2 – дисперсии (определяются с помощью функции **ДИСП.В**).

Для данной задачи различие между дисперсиями незначимо, поэтому используется первая формула: t = 2,06.

Кроме того, для дальнейших расчетов требуется найти число степеней свободы (k).

Если различие между дисперсиями незначимо:

$k=n_1+n_2-2$.

Если различие между дисперсиями значимо:

$$k = \frac{\left(\sigma_1^2 / n_1 + \sigma_2^2 / n_2\right)^2}{\frac{\left(\sigma_1^2 / n_1\right)^2}{n_1 - 1} + \frac{\left(\sigma_2^2 / n_2\right)^2}{n_2 - 1}}$$ (округляется до целого).

Для данного примера $k=n_1+n_2-2 = 25 + 19 - 2 = 42$.

Затем по *абсолютному* значению t и числу степеней свободы k вычисляют *уровень значимости* P. Для этого используют функцию **СТЬЮДЕНТ.РАСП.2Х** со следующими параметрами: **X**: **ABS**(*t*); **Степени свободы**: *k*. Получим P=0,046. Это вероятность того, что различие между средними статистически *незначимо*. В данном примере $P < 0,05$, т.е. различие между средними статистически *значимо*.

В данном примере средние равны $\overline{X}_1$ =46,48 года, $\overline{X}_2$ = 41,00 года. Это означает, что средний возраст работников первой группы (государственные предприятия) *значимо* больше, чем средний возраст работников второй группы (негосударственные предприятия).

Решение задачи с использованием инструментов пакета «Анализ данных»

а) Сравнение дисперсий на основе F-критерия (критерия Фишера)

Из меню **Данные – Анализ данных** выбрать инструмент **Двухвыборочный F-тест для дисперсии**.

В появившемся окне установить следующие параметры: **Интервал переменной 1** – A1:A26, **Интервал переменной 2** – B1:B20, **Альфа** – 0,05. Установить флажок **Метки**. В области **Параметры вывода** указать, куда требуется вывести результаты. Нажать **ОК**. Результаты приведены на рисунке 1.1

Расчетный уровень значимости (P) обозначен как **P(F<=f) одностороннее**. В данном случае $P > 0,05$. Это значит, что различие между дисперсиями двух рассматриваемых генеральных совокупностей (работники государственных и негосударственных предприятий) статистически незначимо. Этот вывод необходимо учесть на следующем шаге.

б) Сравнение средних на основе t-критерия (критерий Стьюдента)

Так как различие между дисперсиями оказалось *незначимым*, требуется использовать инструмент **Двухвыборочный t-тест с одинаковыми дисперсиями**.

Примечание – Если различие между дисперсиями *значимо*, то для сравнения средних необходимо использовать инструмент **Двухвыборочный t-тест с различными дисперсиями**.

Параметры для инструмента задаются так же, как и при сравнении дисперсий. Результаты приведены на рисунке 1.2.

	A	B	C
1	Двухвыборочный F-тест для дисперсии		
2			
3		*Работники государственных предприятий*	*Работники негосударственных предприятий*
4	Среднее	46,48	41
5	Дисперсия	83,51	67,11111111
6	Наблюдения	25	19
7	df	24	18
8	F	1,244354305	
9	P(F<=f) одностороннее	0,320703044	
10	F критическое одностороннее	2,149664535	
11			

Рисунок 1.1 – Решение задачи оценки значимости различий между выборками (сравнение дисперсий)

	A	B	C
1	Двухвыборочный t-тест с одинаковыми дисперсиями		
2			
3		*Работники государственных предприятий*	*Работники негосударственных предприятий*
4	Среднее	46,48	41
5	Дисперсия	83,51	67,11111111
6	Наблюдения	25	19
7	Объединенная дисперсия	76,48190476	
8	Гипотетическая разность средних	0	
9	df	42	
10	t-статистика	2,058835639	
11	P(T<=t) одностороннее	0,022873461	
12	t критическое одностороннее	1,681952358	
13	P(T<=t) двухстороннее	0,045746921	
14	t критическое двухстороннее	2,018081679	
15			

Рисунок 1.2 – Решение задачи оценки значимости различий между выборками (сравнение средних)

Расчетный уровень значимости (P) обозначен как **P(T<=t) двухстороннее**. В данном случае $P < 0{,}05$. Это означает, что различие между средними статистически *значимо*. Из значений самих средних (строка **Среднее**) видно, что среднее значение для первой из исследуемых генеральных совокупностей (работники государственных предприятий) больше, чем для второй.

Задание – Имеются выборки значений прочности образцов двух материалов.
Первый материал:

68	75	34	92	57	29	38	49	84	68	71	42
67	73	37	51	97	42	83	43	54	68	67	

Второй материал:

63	58	69	74	65	74	63	80	75	69	71	41
77	69	49	58	77	79	52	49	63	69	52	77

Определить, значимо ли различаются материалы по средней прочности. Если различие значимо, указать, какой материал прочнее.

1.5 Оценка значимости различий между зависимыми выборками

Под «зависимыми выборками» обычно понимают одну и ту же выборку до и после некоторого воздействия. Цель сравнения зависимых выборок обычно состоит в оценке результата такого воздействия.

Пример 7 – Известна концентрация некоторого загрязнителя в воздухе вблизи 17 заводов:

88	86	92	75	69	80	97	74	83
69	93	80	68	101	93	79	67	

После того, как на заводах были проведены мероприятия по снижению загрязнения, концентрация загрязнителя в тех же местах оказалась следующей:

85	77	91	74	70	77	92	75	83
65	90	72	65	94	90	79	64	

Определить, является ли статистически значимым различие концентраций до и после мероприятий по снижению загрязнения.

Для решения задачи в MS Excel введем в ячейки A1 и B1 какие-либо заголовки (например, «До» и «После»), а в ячейки A2:A18 и B2:B18 – величины концентраций.

Алгоритм решения задачи

Сначала вычисляют разности величин в выборках, т.е. разности концентраций до и после мероприятий:

3	9	1	1	-1	3	5	-1	0
4	3	8	3	7	3	0	3	

Здесь, например, для первого завода 88 – 85 = 3, для второго 86 – 77 = 9, и т.д.

Вычисляют среднюю разность ($\overline{X}$) и стандартное отклонение разностей (σ), используя функции **СРЗНАЧ** и **СТАНДОТКЛОН.В**. Для данной задачи $\overline{X} = 3$, σ=2,96.

Затем вычисляют *t-критерий* (критерий Стьюдента):

$$t = \frac{\overline{X}\sqrt{n}}{\sigma}.$$

Для данной задачи n=17, t = 4,18.

Затем по *абсолютному* значению t и числу степеней свободы n-1 вычисляют *уровень значимости P*. Для этого используют функцию **СТЬЮДЕНТ.РАСП.2Х** со следующими параметрами: **X**: **ABS**(*t*); **Степени свободы**: *n*-1. Получим P=0,0007. Это вероятность того, что различие между средними в совокупностях, из которых взяты выборки, статистически *незначимо*. В данном примере $P < 0,05$, т.е. различие между средними статистически *значимо*. Кроме того, видно, что величины во второй выборке (после мероприятий), как правило, меньше, чем в первой. Это значит, что мероприятия по снижению загрязнения были успешными: концентрация загрязнителя после мероприятий значительно ниже, чем до них.

Решение задачи с использованием инструментов пакета «Анализ данных»

Выбрать инструмент **Парный двухвыборочный тест для средних**. Его использование аналогично рассмотренным выше.

Расчетный уровень значимости (P) обозначен как **P(T<=t) двухстороннее**. В данном случае $P < 0,05$. Это означает, что различие между средними *стати-*

стически значимо. Кроме того, видно (из строки **Среднее**), что среднее во второй выборке (после мероприятий) меньше, чем в первой.

Задание – Измерены значения прочности образцов некоторого материала:

68	75	34	92	57	29	38	49	84	68	71	42
67	73	37	51	97	42	83	43	54	68	67	

Затем эти образцы были подвергнуты некоторой обработке для повышения их прочности. После обработки прочность оказалась следующей:

72	70	38	97	61	28	38	52	87	72	77	44
72	70	35	58	96	42	87	46	58	65	69	

Определить, обеспечила ли обработка значимое повышение прочности.

1.6 Оценка значимости различий между частотами

Пример 8 – Некоторые изделия могут изготавливаться с использованием двух технологий (технология А и технология В). Проверено 284 изделия, изготовленных по технологии А, и 217 изделий, изготовленных по технологии В. Дефекты обнаружены, соответственно, в 83 и 46 изделиях. Требуется определить, является ли статистически значимым различие между частотами, т.е. можно ли утверждать, что технологии значимо различаются по уровню надежности.

Алгоритм решения задачи

Вычислить частоты дефектов: Q_1=83/284=0,2923, Q_2=46/217=0,2120. Вычислить также общую частоту дефектов: Q = (83+46)/(284+217) = 0,2575.

Затем вычисляют *t-критерий* (критерий Стьюдента):

$$t = \frac{Q_1 - Q_2}{\sqrt{Q \cdot (1 - Q) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}}.$$

Для данной задачи n_1=284, n_2=217, t = 2.0361.

Затем по *абсолютному* значению t и числу степеней свободы n_1+n_2-2 вычисляют *уровень значимости P*. Для этого используют функцию **СТЬЮДЕНТ.РАСП.2Х** со следующими параметрами: **X**: **ABS**(*t*); **Степени свободы**: n_1+n_2-2 (для данной задачи число степеней свободы равно 501). Получим P=0,0423. Это вероятность того, что различие между частотами в совокупностях, из которых взяты выборки, статистически *незначимо*. В данном примере P < 0,05, т.е. различие между частотами статистически *значимо*. Кроме того, видно, что частота дефектов для технологии В меньше, чем для технологии А. Это значит, что технология В значимо более эффективна, т.е. обеспечивает значимо более высокую надежность, чем технология А.

Задание – Из 100 изделий, изготовленных из материала М1, отказали 11 изделий. Из 140 изделий, изготовленных из материала М2, отказали 16 изделий. Найти, значимо ли различаются материалы по надежности.