Лабораторная работа 6
Компьютерный анализ медицинских данных

Вопросы

Основные понятия математической статистики. Статистические таблицы. Графическое изображение данных. Основные статистические характеристики. Инструменты статистического анализа.

Цель работы: *научиться обрабатывать статистические данные с помощью встроенных функций Excel; изучить возможности Пакета анализа.*

Указания к выполнению работы:

Математическая статистика имеет дело с ***совокупностью*** объектов , которые обладают некоторым набором ***признаков (показателей, характеристик)***. Это т.н. ***статистическая совокупность***.

Статистическая совокупность может включать ***все изучаемые объекты,*** в этом случае она называется ***генеральной совокупностью*** или только ***часть объектов***, тогда она называется ***выборкой***

Случайная выборка из генеральной совокупности

Чтобы по выборке можно было делать выводы о свойствах всей генеральной совокупности, она должна быть представительной (***репрезентативной***). Это обеспечивается в тех ситуациях, когда выборка является случайной. Модель случайной выборки должна отвечать следующим требованиям:

1) каждый из объектов, составляющих генеральную совокупность, должен иметь одинаковую возможность быть представленным в выборке;

2) все n измерений, изучаемого показателя, образующих выборку, должны быть независимыми, т. е. результаты каждого измерения не должны зависеть от предыдущих измерений.

Измерения изучаемого показателя, составляющие выборку называются вариационным рядом и обычно, помещаются в таблицу.

Чаще всего используются два вида статистических таблиц: простые и групповые

Простые таблицы содержат перечень отдельных измерений, входящих в состав совокупности. В групповых таблицах измерения объединяются в определенные группы в соответствии с каким-либо признаком.

Например: простая таблица содержит ряд измерений веса

61	57	65	57	59	63	61	59	64	56	67	62	57	63	66	58

Простая ранжированная таблица (данные размещены по возрастанию)

56	57	57	57	58	59	59	61	61	62	63	63	64	65	66	67

Групповая таблица с размером диапазонов разброса - 5

Диапазоны группировки	55-60	61-65	66-70
Кол-во значений в группе	7	7	2

Дискретным вариационным рядом распределения называется ранжированная совокупность вариант признака (x_i) с соответствующими им частотами (m_i) или частностями($p_i=m_i/\sum m_i$).

Для рассмотренного примера ряд имеет вид:

x_i	56	57	58	59	61	62	63	64	66	67	Σ
m_i	1	3	1	2	2	1	2	1	1	1	15
p_i	1/15	1/5	1/15	2/15	2/15	1/15	2/15	1/15	1/15	1/15	1

1.Расчет основных статистических характеристик.

Для анализа статистических данных можно использовать различные программные пакеты. Для несложных расчетов описательной статистики применим **MS *Excel,*** который имеет широкий набор средств работы с данными. Наиболее часто используемые статистические функции встроены в основное ядро ***Excel*** и доступны с момента запуска программы. Другие более специализированные функции входят в дополнительную подпрограмму, называемую «**пакетом анализа**». Команды и функции пакета анализа называют ***Инструментами анализа***. Рассмотрим нескольких основных встроенных статистических функций и наиболее полезные инструменты анализа данных из пакета.

Вычисление среднего значения.

Формула среднего значения

Функция **СРЗНАЧ** (или **AVERAGE**) вычисляет выборочное (или генеральное) среднее, то есть среднее арифметическое значение признака выборочной (или генеральной) совокупности. Аргументом функции **СРЗНАЧ** является набор чисел, как правило, задаваемый в виде интервала ячеек, например, **=СРЗНАЧ (A3:A201).**

Вычисление дисперсии и среднего квадратического отклонения.

Для оценки разброса данных используются такие статистические характеристики, как дисперсия ***D*** и среднее квадратическое (или стандартное) отклонение **σ**. Стандартное отклонение есть квадратный корень из дисперсии:. Большое стандартное отклонение указывает на то, что значения измерения сильно разбросаны относительно среднего, а малое – на то, что значения сосредоточены около среднего.

В *Excel* имеются функции, отдельно вычисляющие выборочную дисперсию $D_в$ и стандартное отклонение $\sigma_в$ и генеральные дисперсию $D_г$ и стандартное отклонение $\sigma_г$. Поэтому, прежде чем вычислять дисперсию и стандартное отклонение, следует четко определиться, являются ли ваши данные генеральной совокупностью или выборочной. В зависимости от этого нужно использовать для расчета $D_г$ и $\sigma_г$,$D_в$ и $\sigma_в$.

Для вычисления выборочной дисперсии $D_в$ и выборочного стандартного отклонения $\sigma_в$ имеются функции **ДИСП** (или **VAR**) и **СТАНДОТКЛОН** (или **STDEV**). Аргументом этих функций является набор чисел, как правило, заданный диапазоном ячеек, например, **=ДИСП (B1:B48).**

Для вычисления генеральной дисперсии $D_г$ и генерального стандартного отклонения $\sigma_г$ имеются функции **ДИСПР** (или **VARP**) и **СТАНДОТКЛОНП** (или **STDEVP**), соответственно.

Аргументы этих функций такие же как и для выборочной дисперсии.

Вычисление объема совокупности.

Объем совокупности выборочной илигенеральной – это число элементов совокупности. Функция **СЧЕТ** (или **COUNT**) определяет количество ячеек в заданном диапазоне, которые содержат числовые данные. Пустые ячейки или ячейки, содержащие текст, функция **СЧЕТ** пропускает. Аргументом функции **СЧЕТ** является интервал ячеек, например: **=СЧЕТ** (C2:C16).

Для определения количества непустых ячеек, независимо от их содержимого, используется функция **СЧЕТ3**. Ее аргументом является интервал ячеек.

Расчет моды и медианы.

Мода – это значение признака, которое чаще других встречается в совокупности данных. Она вычисляется функцией МОДА (или MODE). Ее аргументом является интервал ячеек с данными.

Медиана – это значение признака, которое разделяет совокупность на две равные по числу элементов части. Она вычисляется функцией МЕДИАНА (или MEDIAN). Ее аргументом является интервал ячеек.

Вычисление размаха варьирования. Наибольшее и наименьшее значения.

Размах варьирования R – это разность между наибольшим x_{max} инаименьшим x_{min} значениями признака совокупности (генеральной или выборочной): $R=x_{max}-x_{min}$. Для нахождения наибольшего значения x_{max} имеется функция **МАКС** (или **MAX**), а для наименьшего x_{min} – функция **МИН** (или **MIN**). Их аргументом является интервал ячеек. Для того, чтобы вычислить размах варьирования данных в интервале ячеек, например, от A1 до A100, следует ввести формулу: **=МАКС (A1:A100)-МИН (A1:A100).**

Выявление отклонения случайного распределения от нормального.

Нормально распределенные случайные величины широко распространены на практике, например, результаты измерения любой физической величины подчиняются нормальному закону распределения. Нормальным называется распределение вероятностей непрерывной случайной величины, которое описывается плотностью

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

где σ^2 дисперсия, **a**- среднее значение случайной величины X .

Для оценки отклонения распределения данных эксперимента от нормального распределения используются такие характеристики как асимметрия **A** и эксцесс **E**. Для нормального распределения **A**=0 и **E**=0.

Асимметрия показывает, на сколько распределение данных несимметрично относительно нормального распределения: если **A**>0, то большая часть данных имеет значения, превышающие среднее ; если **A**<0, то большая часть данных имеет значения, меньшие среднего . Асимметрия вычисляется функцией **СКОС**. Ее аргументом является интервал ячеек с данными, например, **=СКОС (A1:A100).**

Эксцесс оценивает ***«крутость»***, т.е. величину большего или меньшего подъема максимума распределения экспериментальных данных по сравнению с максимумом нормального распределения. Если **E**>0, то максимум экспериментального распределения выше нормального; если **E**<0, то максимум экспериментального распределения ниже нормального. Эксцесс вычисляется функцией **ЭКСЦЕСС**, аргументом которой являются числовые данные, заданные, как правило, в виде интервала ячеек, например: **=ЭКСЦЕСС (A1:A100).**

Задание 1.

Одним и тем же вольтметром было измерено 25 раз напряжение на участке цепи. В результате опытов получены следующие значения напряжения в вольтах: 32, 32, 35, 37, 35, 38, 32, 33, 34, 37, 32, 32, 35, 34, 32, 34, 35, 39, 34, 38, 36, 30, 37, 28, 30.

1. Наберите результаты эксперимента в строку 1 таблицы листа Excel,
2. Создайте ранжированный (по возрастанию), ряд во 2й строке листа с помощью сортировки.
3. В строках ниже постройте дискретный вариационный ряд распределения.
4. По вариантам признака (x_i) и частотам (m_i) дискретного ряда распределения постройте гистограмму распределения.

Задание 2.

По данным из первого задания найдите выборочные среднюю, дисперсию, стандартное отклонение, размах варьирования, моду, медиану. Проверьте отклонение от нормального распределения, вычислив асимметрию и эксцесс.

1.Активизируйте новый лист книги Excel, введите результаты эксперимента в столбец А.
2.В ячейку В1 наберите «Среднее», в В2 – «выборочная дисперсия», в В3 – «стандартное отклонение», в В4 – «Максимум», в В5 – «Минимум», в В6 – « Размах варьирования», в В7 – «Мода», в В8 – «Медиана», в В9 – «Асимметрия», в В10 – «Эксцесс». Выровняйте ширину этого столбца с помощью ***Автоподбора*** ширины.
3.Выделите ячейку С1 и нажмите на знак «=» в строке формул. С помощью ***Мастера функций*** в категории ***Статистические*** найдите функцию **СРЗНАЧ**, затем выделите интервал ячеек с данными и нажмите *Enter*.
4.Выделите ячейку С2 и нажмите на знак «=» в строке формул. С помощью помощью ***Мастера функций*** в категории ***Статистические*** найдите функцию **ДИСП**, затем выделите интервал ячеек с данными и нажмите ***Enter***.
5.Проделайте самостоятельно аналогичные действия для вычисления стандартного отклонения, максимума, минимума, моды, медианы, асимметрии и эксцесса.
6.Для вычисления размаха варьирования в ячейку С6 следует ввести формулу: **=МАКС (А1:А25)-МИН(А1:А25)**.

2. Инструменты статистического анализа: Генерация случайных чисел, Гистограмма, Описательная статистика.

Загрузка *Пакета анализа*.

Пакет анализа без дополнительных установок автоматически не загружается при запуске ***Excel***. Он входит в так называемую ***Надстройку*** – набор дополнительных подпрограмм, к которым относятся, например, уже известные вам ***Мастер диаграмм*** и ***Мастер функций*.** Для загрузки ***Пакет анализа*** необходимо:

1)в ***Основном меню*** выбрать пункт ***Сервис***;

2)выбрать пункт ***Надстройки***;

3)в появившемся списке ***Надстроек*** активизировать переключатель ***AnalysisToolPak-VBA*** и нажать **ОК**.

После этого в меню ***Сервис*** добавится пункт ***Анализ данных*.** К этому пункту следует обращаться для вызова ***Пакета анализа***.

Инструмент: *Генерация случайных чисел.*

В ***Excel*** имеется встроенная функция **СЛЧИСЛ** (или **RAND**) для генерации равномерно распределенных случайных чисел в интервале [0,1].

Пакет анализа позволяет генерировать случайные числа с различными типами распределений: равномерное, нормальное, Бернулли, биномиальное, Пуассона и дискретное (определенное пользователем). Для генерации случайных чисел следует:

1)в меню ***Сервис*** выбрать команду ***Анализ данных***;

2)в появившемся диалоговом окне ***Анализ данных*** в группе ***Инструменты анализа*** выбрать пункт ***Генерация случайных величин***и нажать ***ОК***;

3)в появившемся диалоговом окне ***Генерация случайных чисел*** следует заполнить поля ввода:

- в полях ***Число переменных*** и ***Число случайных чисел*** указать нужное количество столбцов и сколько чисел вы хотите получить в каждом столбце;

- в поле ***Распределение*** следует выбрать один из имеющихся типов распределения случайных чисел;

- в группе ***Параметры*** следует указать диапазон чисел, т.е. min и max числа распределения для ***Равномерного распределения***; или среднее значение и стандартное отклонение для ***Нормального распределения*** и т.д.

- поле ***Случайное*** рассеивание заполняется только в том случае, если вам необходимо несколько раз воспроизводить одну и туже последовательность случайных чисел;

- в поле ***Выходной интервал*** указывается место, куда следует поместить последовательность чисел, как правило, это интервал ячеек (или столбец целиком).

Инструмент: *Гистограмма.*

Графическое представление результатов обработки статистических данных обычно оформляется в виде гистограммы. Совокупность данных разбивается на частичные интервалы, называемые нормальными. Интервалы разбиения могут быть любой ширины, но обязательно они должны следовать в порядке возрастания. Интервалы разбиения откладываются по оси абсцисс гистограммы. На оси ординат гистограммы откладывается число значений, попавших в интервал разбиения. Это число значений признака совокупности называется частотой.

Для построения гистограммы:

1) в начале следует задать частичные интервалы разбиения;

2) затем в меню ***Сервис*** выбрать команду ***Анализ данных*** и указать инструмент анализа –***Гистограмма*** и нажать ***ОК***;

3) в диалоговом окне ***Гистограмма*** следует указать:

- в группе ***Входные данные*** в поле ***Входной интервал*** – интервал ячеек с данными, а в поле ***Интервал карманов*** – интервал ячеек с частичными интервалами разбиения;
- в группе ***Параметры вывода*** указывается интервал ячеек для вывода частот и отмечается галочкой переключатель ***Вывод графика***.

После нажатия ***ОК*** инструмент***Гистограмма*** выводит два столбца: карман и частота. Сама гистограмма выводится правее столбца частот. Форматирование гистограммы производится так же, как и любой диаграммы в ***Excel ..***

Инструмент: *Описательная статистика.*

В пакете анализа ***Excel*** содержится инструмент ***Описательная статистика***, который создает таблицу основных статистических характеристик для совокупности данных. В этой таблице будут содержаться следующие характеристики: среднее, стандартная ошибка, дисперсия, стандартное отклонение, мода, медиана, размах варьирования интервала, максимальное и минимальное значения, асимметрия, эксцесс, объем совокупности, сумму всех элементов совокупности, доверительный интервал (уровень надежности). Инструмент ***Описательная статистика*** существенно упрощает статистический анализ тем, что нет необходимости вызывать каждую функцию для расчета статистических характеристик отдельно.

Для того, чтобы вызвать ***Описательную статистику***, следует:

1) в меню ***Сервис*** выбрать команду ***Анализ данных***;

2) в списке ***Инструменты анализа*** диалогового окна ***Анализ данных*** выбрать инструмент ***Описательная статистика*** и нажать ***ОК***;

3) в появившемся диалоговом окне ***Описательная статистика*** необходимо:

- в группе ***Входные данные*** в поле ***Входной интервал*** указать интервал ячеек, содержащих данные;
- если первая строка во входном диапазоне содержит заголовок столбца, то в *поле* ***Метки в первой строке*** следует поставить галочку;
- активизировать переключатель (поставить галочку) ***Итоговая статистика***, если нужен полный список характеристик;
- активизировать переключатель ***Уровень надежности*** и указать надежность в %, если необходимо вычислить доверительный интервал.

Задание 3.

Сгенерировать 500 случайных чисел, распределенных нормально. Построить гистограмму и полный список статистических характеристик с помощью инструмента ***Описательная статистика***.

1. Выполните команду ***Сервис→Анализ данных→Генерация случайных чисел***;

2..В диалоговом окне ***Генерация случайных чисел*** введите в поле число переменных: 1; в поле Число случайных чисел 500; выберите ***Распределение Нормальное***; задайте любое среднее значение (желательно около 100) и небольшое стандартное отклонение (не

больше 10); в поле Выходной интервал укажите абсолютный адрес столбца A2.
Нажмите **ОК.**

3. Теперь постройте гистограмму по совокупности случайных чисел. Сначала нужно задать интервалы решения. Пусть длины интервалов будут одинаковыми и равны 3. Для автоматического составления интервалов разбиения наберите в ячейку В2 начальное число, например, 75 для наших случайных чисел. Затем выполните команду ***Правка→Заполнить→Прогрессия.*** В появившемся диалоговом окне заполните данные:
   - в группе переключателей поле ***Расположение*** установите *по столбцам*;
   - в поле ***Шаг*** наберите 3;
   - в поле ***Предельное значение*** наберите 125;
   - в группе переключателей ***Тип*** установите ***арифметическая*** и нажмите **ОК**.

   В результате столбец В будет содержать интервалы разбиения (карманы).
4. Выполните команду ***Сервис→Анализ данных→Гистограмма.*** В появившемся диалоговом окне ***Гистограмма*** заполните:
   - входной интервал появится, если щелкнуть мышью по столбцу А;
   - интервал карманов появится, если щелкнуть мышью по столбцу В;
   - поставьте галочку в поле метки;
   - укажите столбец С в поле ***Выходной интервал***;
   - активизируйте переключатель ***Вывод графика***; если это поле не содержит галочки, нажмите **ОК.**
5. Построение гистограммы займет от 5 до 10 минут. За это время письменно ответьте на контрольные вопросы. В результате вычисления получатся столбец под названием *Карман*, который дублирует ваш столбец интервалов разбиения, и столбец под название *Частота* с рассчитанными частотами. После того, как появилась гистограмма, измените ее размеры с помощью мыши так, чтобы хорошо были видны все столбцы и подписи.
6. Теперь осталось получить таблицу статистических характеристик с *помощью* ***Описательной статистики***. Выполните команду ***Сервис□Анализ данных□Описательная статистика***. В появившемся диалоговом окне ***Описательная статистика*** укажите:
   - в поле ***Входной интервал*** появится адрес, если выделить мышью интервал сданными или с клавиатуры набрать адрес A2: A501;
   - в поле ***Группирование*** активизировать переключатель ***по столбцам***;
   - активизировать переключатель ***Метки в первой строке***;
   - в группе ***Параметры вывода*** укажите ***Выходной интервал***, щелкнув мышью по какой-либо пустой ячейке ниже столбца частот, например, по С 25;
   - активизируйте переключатель ***Итоговая статистика*** (если в этом поле нет галочки);
   - активизируйте переключатель ***Уровня надежности*** и установите 95%;
   - снимите галочки с полей ***наименьший*** и ***наибольший*** и нажмите **ОК**.

Результаты покажите преподавателю.

Задание 4

В результате выборочного обследования студентов ВУЗа получено 100 измерений роста, данные в приведенной ниже таблице

138	173	166	172	167	172	155	157	162	156
151	177	166	153	156	150	169	148	157	174
168	163	174	150	164	173	177	153	162	163
178	179	166	159	161	175	185	171	156	166
162	166	172	154	164	178	155	169	175	160
155	174	174	155	152	152	156	173	164	160

170	169	180	165	168	165	163	160	165	164
167	176	186	163	146	154	167	153	155	171
160	177	175	166	160	173	163	171	172	150
163	162	171	157	160	149	164	156	141	169

Необходимо:

1. ранжировать ряд данных,
2. сгруппировать по диапазонам размером 5см.
3. построить гистограмму по сгруппированным данным
4. с помощью инструментов анализа провести анализ данных (как в Задании 3)
5. сформулировать выводы

Контрольные вопросы.

1. Для чего предназначена функция СРЗНАЧ?
2. С помощью каких характеристик оценивают разброс статистических данных? Какие функции в *Excel* их вычисляют? В чем отличие функции оценки разброса данных для генеральной и выборочной совокупности?
3. В чем отличие функций СЧЕТ и СЧЕТЗ?
4. Что такое мода и какая функция ее вычисляет?
5. Что такое медиана и какая функция ее вычисляет?
6. Как вычислить размах варьирования?
7. С помощью каких характеристик оценивают отклонение случайного распределения от нормального? Какой смысл этих характеристик и какие функции в *Excel* их вычисляют?
8. Что такое *Инструменты Анализа*? Как загрузить *Пакет Анализа*?
9. Опишите последовательность действий, которые необходимо совершить для генерации случайных чисел распределенных нормально.
10. Как построить гистограмму?
11. Для чего предназначен инструмент *Описательная статистика*?