

Введение

При выполнении курсового проекта по математической статистике возникает много вопросов как по поводу теоретического обоснования применяемых процедур, так и по поводу их практической реализации. В данном пособии приводятся подробные схемы вычислений в рамках популярного компьютерного приложения MS Excel. Теоретические основы применяемых процедур даются в пособии [4].

Работу над курсовым проектом следует начать с изучения главы I “Предварительные понятия и определения” теоретического пособия [4]. Эта глава будет весьма полезна при подготовке ответов на контрольные вопросы. Выполнение каждого задания лучше всего начинать с изучения теоретического обоснования тех процедур, которые рассматриваются в этом задании. Причем желательно изучить весь материал заранее, до проведения соответствующего занятия в компьютерном классе.

В конце каждого задания приведены варианты контрольных вопросов, которые могут быть заданы студенту при сдаче всего курсового проекта.

Задания

- Задание 1.** Вычислить выборочные характеристики – среднее, дисперсию, стандартное отклонение, асимметрию, эксцесс.
- Задание 2.** Построить гистограмму выборки с подогнанной нормальной (равномерной, экспоненциальной) плотностью.
- Задание 3.** Построить эмпирическую функцию распределения выборки с подогнанной нормальной (равномерной, экспоненциальной) функцией распределения.
- Задание 4.** Проверить гипотезу нормальности (равномерности, экспоненциальности) выборочных данных.
- Задание 5.** Проверить гипотезу однородности по одновыборочному критерию Стьюдента.
- Задание 6.** Проверить гипотезу однородности по критерию знаков.
- Задание 7.** Проверить гипотезу однородности по двухвыборочному критерию Стьюдента.
- Задание 8.** Проверить гипотезу однородности по критерию Вилкоксона.
- Задание 9.** Проверить гипотезу равенства дисперсий.
- Задание 10.** Проверить гипотезу однородности по критерию хи-квадрат.
- Задание 11.** Построить доверительные границы для среднего значения нормального распределения.
- Задание 12.** Построить доверительные границы для дисперсии нормального распределения.
- Задание 13.** Построить доверительные границы для вероятности успеха.
- Задание 14.** Проверить гипотезу независимости признаков по критерию сопряженности хи-квадрат.
- Задание 15.** Проверить гипотезу независимости по критерию Стьюдента.
- Задание 16.** Построить линии регрессии.

Данные

Все примеры выполнения заданий основаны на следующих данных.

К заданиям 1 – 4, 13.

| | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|
| 119,3 | 122,1 | 120 | 120,2 | 122,8 | 121,8 | 120,8 | 124,6 |
| 121,9 | 120,3 | 120,2 | 120,7 | 120,9 | 120,2 | 121,7 | 119 |
| 122,2 | 119,2 | 120 | 121,2 | 120 | 121 | 123 | 124,2 |
| 121,5 | 120,7 | 121 | 120,1 | 120,1 | 122 | 119,6 | 120,3 |
| 121,5 | 119,6 | 120,6 | 121,4 | 124,1 | 119,4 | 122 | 123,7 |
| 122,9 | 121,9 | 121,9 | 119,9 | 121,4 | 122,7 | 118,8 | 123 |
| 119,8 | 121,4 | 120,1 | 121 | 120,9 | 120,9 | 121,9 | 121,4 |
| 120,8 | 119,4 | 120,1 | 119,9 | 120,3 | 119,9 | 120,8 | 122,6 |
| 121,9 | 121,4 | 120 | 120,6 | 122,5 | 119,4 | 119,2 | 120,1 |
| 119,3 | 123,4 | 120,1 | 121,1 | 120,2 | 121,3 | 120,6 | 122,1 |
| 120,3 | 120,7 | 119,1 | 119,8 | 121,7 | 118 | 122,1 | |
| 122,7 | 120,3 | 120,9 | 120,4 | 121,5 | 120,7 | 124,2 | |
| 120,1 | 122,8 | 123,1 | 121,2 | 119,8 | 118,6 | 121,4 | |

Число интервалов для гистограммы – 10, первая граница – 117,05, длина интервала – 1.

К заданиям 5, 6.

| | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| До | 186 | 170,4 | 179,4 | 154,3 | 152,9 | 179,7 |
| После | 157,1 | 152,6 | 149,3 | 156,2 | 159,7 | 154,1 |

| | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| До | 173,3 | 162,9 | 168,8 | 161,8 | 162,3 | 163,2 |
| После | 148,7 | 148,3 | 157,4 | 161 | 151,7 | 168,5 |

| | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| До | 173,3 | 173,1 | 178,8 | 169,8 | 166 | 167,2 |
| После | 164,9 | 168 | 157,7 | 129,9 | 146,7 | 143 |

К заданиям 7, 8, 9.

| | | | | | | |
|-------------|-------|-------|-------|-------|-------|-------|
| 1-я выборка | 159,3 | 158,9 | 163,1 | 169,8 | 161,2 | 158,1 |
| 2-я выборка | 143,4 | 158,8 | 167,3 | 163,4 | 163,8 | 174,9 |

| | | | | | | |
|-------------|-----|-------|-------|-------|-------|-------|
| 1-я выборка | 191 | 151,4 | 143,5 | 166,5 | 173,1 | 184,1 |
| 2-я выборка | 166 | 148,7 | 163,6 | 199,9 | | |

К заданию 10.

| Группы | 1-я | 2-я | 3-я | 4-я | 5-я | 6-я |
|-------------|-----|-----|-----|-----|-----|-----|
| 1-я выборка | 6 | 15 | 28 | 39 | 10 | 2 |
| 2-я выборка | 5 | 11 | 12 | 17 | 13 | 2 |

К заданиям 11, 12.

| | | | | | | | |
|------|------|------|------|------|------|------|------|
| 3,9 | 4,1 | 4,19 | 4,09 | 3,9 | 4,87 | 4,51 | 4,52 |
| 4,94 | 4,83 | 4,65 | 3,55 | 4,29 | 4,62 | 3,8 | |
| 5,26 | 4,58 | 4,7 | 4,32 | 3,49 | 3,81 | 5,13 | |

К заданиям 14 – 16.

| X | Y | X | Y | X | Y | X | Y |
|-------|------|-------|------|-------|------|-------|------|
| 119,3 | 56,1 | 120 | 56,9 | 122,8 | 52 | 120,8 | 49,8 |
| 121,9 | 63,1 | 120,2 | 56,3 | 120,9 | 58,4 | 121,7 | 54,1 |
| 122,2 | 54,2 | 120 | 54,2 | 120 | 54,1 | 123 | 55,3 |
| 121,5 | 55,8 | 121 | 53,7 | 120,1 | 55,8 | 119,6 | 55,7 |
| 121,5 | 56,7 | 120,6 | 55,5 | 124,1 | 53,6 | 122 | 51,5 |
| 122,9 | 54,5 | 121,9 | 56 | 121,4 | 60,8 | 118,8 | 49,5 |
| 119,8 | 57,6 | 120,1 | 55,5 | 120,9 | 54,3 | 121,9 | 52,5 |
| 120,8 | 52,2 | 120,1 | 53,6 | 120,3 | 50,6 | 120,8 | 52,3 |
| 121,9 | 53,7 | 120 | 55,8 | 122,5 | 55,1 | 119,2 | 55,6 |
| 119,3 | 57,6 | 120,1 | 52,1 | 120,2 | 55 | 120,6 | 50,4 |
| 120,3 | 57,1 | 119,1 | 53,7 | 121,7 | 54,5 | 122,1 | 52,3 |
| 122,7 | 56,8 | 120,9 | 56 | 121,5 | 54,9 | 124,2 | 53,2 |
| 120,1 | 58,4 | 123,1 | 57,4 | 119,8 | 53,3 | 121,4 | 57,8 |
| 122,1 | 60,7 | 120,2 | 53,6 | 121,8 | 57,2 | 124,6 | 54,5 |
| 120,3 | 56,9 | 120,7 | 53,5 | 120,2 | 54,4 | 119 | 52,5 |
| 119,2 | 52,7 | 121,2 | 52,7 | 121 | 55,4 | 124,2 | 49,9 |
| 120,7 | 56,8 | 120,1 | 53,9 | 122 | 56,1 | 120,3 | 52,3 |
| 119,6 | 51,8 | 121,4 | 53,7 | 119,4 | 49,8 | 123,7 | 53,5 |
| 121,9 | 58,1 | 119,9 | 53,8 | 122,7 | 54,9 | 123 | 52,8 |
| 121,4 | 55,3 | 121 | 57,9 | 120,9 | 53,4 | 121,4 | 53,3 |
| 119,4 | 60,2 | 119,9 | 56,3 | 119,9 | 51,8 | 122,6 | 48,9 |
| 121,4 | 54,2 | 120,6 | 53,1 | 119,4 | 54,3 | 120,1 | 51,6 |
| 123,4 | 55 | 121,1 | 54,2 | 121,3 | 54,7 | 122,1 | 51,8 |
| 120,7 | 53,3 | 119,8 | 50,4 | 118 | 51,5 | | |
| 120,3 | 56,4 | 120,4 | 55,4 | 120,7 | 56,2 | | |
| 122,8 | 55,3 | 121,2 | 53,6 | 118,6 | 50,4 | | |

Первая граница для признака X – 119,65, шаг – 1,6;
для признака Y – 52,55, шаг – 1,5.

Задание 0.

Основы математической статистики.

Постановка задачи.

Изучите основные понятия и методы, необходимые для выполнения курсового проекта по математической статистике.

Вероятностные характеристики: функция распределения, функция плотности, квантиль распределения, верхняя квантиль распределения, нормальная модель, экспоненциальная модель, модель равномерного распределения, биномиальная модель.

Основные понятия математической статистики: выборка, статистика, оценка, решающая функция, состоятельность, несмещенность, задача проверки гипотезы, вероятности ошибок 1-го и 2-го рода, размер критерия, уровень значимости, критический уровень значимости.

Теоретические основы.

См. стр. 5-22 пособия [4].

Контрольные вопросы.

1. Что такое функция распределения (функция плотности)?
Ответ: см. [4] стр. 14.
2. Что такое функция надежности?
Ответ: см. [4] стр. 14.
3. Какая случайная величина имеет нормальное распределение (показательное, равномерное, биномиальное, хи-квадрат, Стьюдента)?
Ответ: см. [4] стр. 16-22.
4. Запишите формулу нормального закона распределения (экспоненциального, равномерного, биномиального).
Ответ: см. [4] стр. 16-22.
5. Какой смысл несут параметры нормального распределения (экспоненциального, биномиального, хи-квадрат, Стьюдента)?
Ответ: см. [4] стр. 16-22.
6. Чему равны среднее значение и дисперсия экспоненциального распределения (нормального, биномиального)?
Ответ: см. [4] стр. 16-22.
7. Что такое квантиль распределения (верхняя квантиль)?
Ответ: см. [4] стр. 15.
8. Как связаны функция распределения и её верхняя квантиль?
Ответ: см. [4] стр. 15.
9. Найдите по таблице значение верхней 7%-й квантили для распределения хи-квадрат при 15 степенях свободы (для нормального распределения, для распределения Стьюдента).
Ответ: см. [4] стр. 17-20.
10. Что такое выборка?
Ответ: см. [4] стр. 5.
11. Что такое оценка?
Ответ: см. [4] стр. 10.
12. Дайте определение состоятельности оценки и проинтерпретируйте смысл этого определения.
Ответ: см. [4] стр. 12.
13. Можно ли сказать, что состоятельная оценка лучше не состоятельной оценки?

14. Дайте определение несмещенности оценки и проинтерпретируйте смысл этого определения.

Ответ: см. [4] стр. 10.

15. Можно ли сказать, что несмещенная оценка лучше смещенной оценки?

16. Как следует выбирать нулевую гипотезу?

Ответ: см. [4] стр. 7.

17. Как определяется вероятность ошибки 1-го рода? Что такое размер критерия?

Ответ: см. [4] стр. 7.

18. Что такое уровень значимости?

Ответ: см. [4] стр. 7.

19. Какой уровень значимости лучше выбрать – 5% , 10% или 1%?

Ответ: см. [4] стр. 7-8.

20. Как часто мы будем ошибаться, если будем применять критерий уровня $\alpha = 0,03$.

Ответ: см. [4] стр. 7.

21. Как построить критерий заданного уровня, основываясь на значениях некоторой статистики T ?

Ответ: см. [4] стр. 8.

22. Можно ли признать новый метод лечения лучше старого, если при клинических испытаниях результативность нового метода составила 85%, а старого – 70%? Что ещё нужно знать, что бы правильно ответить на этот вопрос?

Ответ: см. [4] стр. 8.

23. Что такое критический уровень значимости? Чем он отличается от уровня значимости?

Ответ: см. [4] стр. 7, 9.

24. Следует ли принять гипотезу, если критический уровень значимости равен $\alpha_{\text{крит}} = 0,18$?

Ответ: см. [4] стр. 9.

Задание 1.

Выборочные характеристики.

Постановка задачи.

Вычислить основные статистические характеристики выборочных данных:

1. Среднее арифметическое.
2. Дисперсию.
3. Стандартное отклонение.
4. Коэффициент асимметрии.
5. Коэффициент эксцесса.

Вычисления.

В пакете Excel имеется возможность вычисления всех четырех выборочных моментов. Для этого можно использовать встроенные функции СРЗНАЧ, ДИСПР, СКОС и ЭКСЦЕСС, вызываемые либо посредством мастера функций (кнопка f_x – категория «Статистические»), либо введенные непосредственно в ячейку листа Excel. Формат вызова всех функций одинаков:

=ФУНКЦИЯ (x1; x2 ; ...) или =ФУНКЦИЯ (Массив данных).

Второй способ вызова, конечно, удобнее. Здесь Массив данных – это область листа Excel вида A1 : B43, содержащая обрабатываемые данные.

Функция ДИСПР дает смещенную оценку дисперсии, для подсчета несмещенной оценки предназначена функция ДИСП. Стандартное отклонение вычисляется посредством функций СТАНДОТКЛОНП и СТАНДОТКЛОН, квадраты которых суть функции ДИСПР и ДИСП, соответственно.

Функции СКОС и ЭКСЦЕСС вычисляют почти несмещенные оценки соответствующих характеристик. Для наших целей эти функции не совсем подходят. Мы будем вычислять эти коэффициенты исходя из определений.

Пример.

Лист “Моменты” с выборочными характеристиками может выглядеть следующим образом.

| | А | В | С | Д | Е | Ф | Г | Н |
|---|---------------------------|--------|------------|---|---|---|---------|-----------|
| 1 | Выборочные моменты | | | | | | 1,1950 | 0,2687 |
| 2 | | | | | | | Кубы | 4-е степ. |
| 3 | Объем выборки n | 101 | | | | | -4,9302 | 8,3911 |
| 4 | Среднее \bar{x} | 121,00 | | | | | 0,7242 | 0,6503 |
| 5 | Дисперсия S^2 | 1,773 | | | | | 1,7195 | 2,0599 |
| 6 | Ст.Отклонение S | 1,33 | | | | | 0,1235 | 0,0615 |
| 7 | Асимметрия g_1 | 0,506 | Значимо | | | | 6,8376 | 12,9778 |
| 8 | Экссесс g_2 | 0,051 | Не значимо | | | | -1,7366 | 2,0873 |
| 9 | | | | | | | ... | |

Если выборочные данные хранятся на листе “Данные” в области “В3:В103”, то для заполнения листа можно использовать следующий

Порядок вычислений.

- 1) В ячейке В3 найти количество данных
 - =СЧЁТ (Данные!В3:В103)
- 2) в ячейке В4 вычислить среднее арифметическое
 - =СРЗНАЧ (Данные!В3:В103)
- 3) в ячейке В5 вычислить дисперсию
 - =ДИСПР (Данные!В3:В103)
- 4) в ячейке В6 вычислить стандартное отклонение
 - =СТАНДОТКЛОНП (Данные!В3:В103)

Для отыскания асимметрии и эксцесса провести вспомогательные вычисления в столбцах Г и Н:

- 5) в ячейках Г3 и Н3 ввести формулы вычисления 3-ей $((x_i - \bar{x})^3)$ и 4-ой $((x_i - \bar{x})^4)$ степеней отклонения выборочных данных от среднего
 - =(Данные!В3-\$B\$3)^3
 - =(Данные!В3-\$B\$3)^4
 (обратите внимание на знаки \$);

- 6) скопировать ячейки G3 и H3 до 103-й строки (параллельно выборочным данным);
- 7) в ячейках G1 и H1 вычислить средние арифметические столбцов G и H
 - =СРЗНАЧ (G3:G104)
 - =СРЗНАЧ (H3:H104)
- 8) в ячейке B7 вычислить асимметрию
 - =G1/B6^3
- 9) в ячейке B8 вычислить эксцесс
 - =H1/B6^4-3
- 10) в ячейки C7 и C8 занести результаты сравнений асимметрии и эксцесса с критическими значениями (см. пособие [4], стр. 26). В нашем случае асимметрия не попадает (– значимо отличается от нуля), а эксцесс попадает в допустимый интервал (– не значимо отличается от нуля). Таким образом, имеются некоторые основания сомневаться в нормальности распределения данных.

Замечание. Вычисленные посредством функций СКОС и ЭКСЦЕСС, выборочные асимметрия и эксцесс равны 0,5147 и 0,0144.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Является ли выборочное среднее (дисперсия, стандартное отклонение, коэффициент асимметрии, эксцесс) несмещенной оценкой?
Ответ: см. [4] стр. 11.
3. Является ли выборочное среднее (дисперсия, стандартное отклонение, коэффициент асимметрии, эксцесс) состоятельной оценкой?
Ответ: см. [4] стр. 12.
4. Что такое состоятельность и несмещенность?
Ответ: см. [4] стр. 10, 12.
5. Как можно исправить смещение выборочной дисперсии? Будет ли после такого исправления несмещенным стандартное отклонение? Будет ли состоятельной несмещенная оценка дисперсии?
Ответ: см. [4] стр. 11.
6. Какую информацию несет коэффициент эксцесса (среднее значение, дисперсия, асимметрия, коэффициент корреляции)?
Ответ: см. [4] стр. 25-27.
7. По какой формуле вычисляется дисперсия (среднее, асимметрия, эксцесс)?
Ответ: см. [4] стр. 25-26.
8. Что означают слова “значимо” и “не значимо” в представленном отчете?
Ответ: см. [4] стр. 26-27.
9. На каком принципе основан выбор границ, по которым проверялась гипотеза нормальности?
Ответ: см. [4] стр. 27.

Задание 2.

Гистограмма выборки.

Постановка задачи.

Построить график гистограммы выборки с подогнанной ожидаемой функцией плотности.

Теоретические основы.

См. стр. 28-30 пособия [4].

Вычисления.

Пакет Excel располагает функцией ЧАСТОТА, предназначенной для подсчета количеств попаданий в заданные интервалы разбиения числовой прямой. Это не совсем обычная функция. Она относится к классу так называемых функций массива. Для её вызова необходимо

- 1) в столбце, например, A2 : A10 указать интервалы группировки;
- 2) напротив первой границы, например, в ячейку B2 ввести формулу вычисления частоты
 - =ЧАСТОТА (Данные; Область границ)
 - область границ должна содержать ссылку на ячейки с границами + дополнительная ячейка для крайнего правого интервала, например, A2 : A11;
- 3) выделить вертикальный диапазон ячеек, начиная с ячейки, содержащей формулу вычисления частоты, и заканчивая дополнительной ячейкой, соответствующей правой крайней границе (например, B2 : B11);
- 4) ввести формулу как формулу массива
 - последовательно нажать сочетания клавиш

| |
|----|
| F2 |
|----|

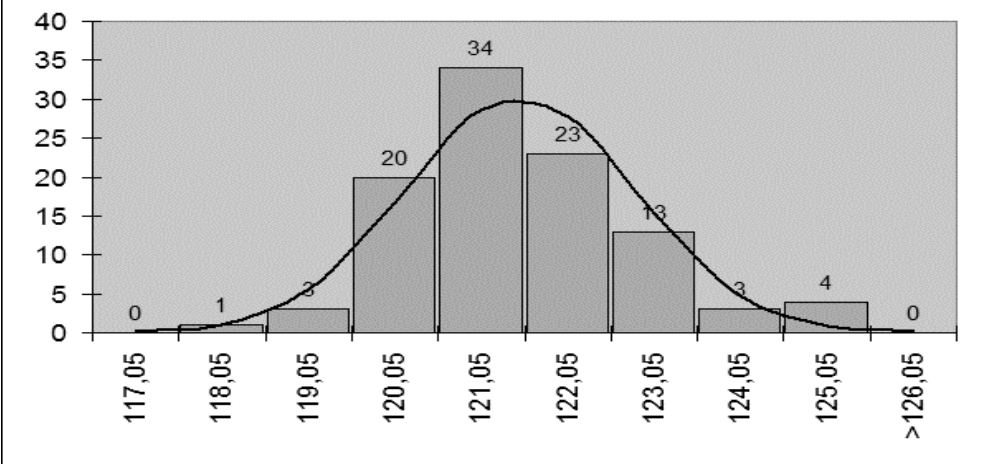
 –

| |
|----------------------|
| Ctrl + Shift + Enter |
|----------------------|

 - в результате формула {заклученная в фигурные скобки} будет введена во все ячейки диапазона.

Правильность применения функции можно проконтролировать, сравнив сумму значений выделенного диапазона (указывается в нижней строке состояния окна Excel) с объемом выборки.

Пример.

| | А | В | С | Д | Е |
|----|---|---------|-----------|-----------|--------|
| 1 | Границы | Частоты | Плотность | | |
| 2 | 117,05 | 0 | 0,11 | | |
| 3 | 118,05 | 1 | 1,05 | Среднее | 121,00 |
| 4 | 119,05 | 3 | 5,55 | Дисперсия | 1,773 |
| 5 | 120,05 | 20 | 16,70 | Объем | 101 |
| 6 | 121,05 | 34 | 28,57 | | |
| 7 | 122,05 | 23 | 27,80 | | |
| 8 | 123,05 | 13 | 15,39 | | |
| 9 | 124,05 | 3 | 4,85 | | |
| 10 | 125,05 | 4 | 0,87 | | |
| 11 | >125,05 | 0 | 0,09 | | |
| 12 | Всего | 101 | 100,99 | | |
| 13 | | | | | |
| 14 |  | | | | |
| 15 | | | | | |
| 16 | | | | | |
| 17 | | | | | |
| 18 | | | | | |
| 19 | | | | | |
| 20 | | | | | |
| 21 | | | | | |
| 22 | | | | | |
| 23 | | | | | |
| 24 | | | | | |

📖 Порядок вычислений.

- 1) Заполнить ячейки А2, А3 значениями первых двух границ;
- 2) выделить ячейки А2, А3 и, захватив мышкой точку в правом нижнем углу выделения, протащить мышку до ячейки А10 (число введенных границ будет на 1 меньше необходимого числа групп);
- 3) в столбце Е (ячейки Е3, Е4, Е5) ввести значения среднего, дисперсии и объема выборки, вычисленные в задании 1;
- 4) в ячейку В2 (напротив 1-ой границы) ввести формулу

- =ЧАСТОТА (Данные!В3:В103;А2:А11)
- 5) скопировать введенную формулу как формулу массива:
 - выделить диапазон ячеек со 2-ей по 11-ую в столбце В (на одну ячейку больше, чем в столбце А);
 - последовательно нажать сочетания клавиш **F2** – **Ctrl + Shift + Enter**
 - в результате формула {заклученная в фигурные скобки} будет введена во все ячейки выделенного диапазона;
- 6) для контроля просуммировать все значения во втором столбце
 - результат должен равняться объему выборки (в нашем примере см. ячейку В12);
- 7) в столбце А в ячейке под последней границей (А11) ввести выражение
 - >125,05
 - здесь, конечно, нужно указывать свою последнюю границу или символ ∞ ;
- 8) в ячейку С2 ввести формулу вычисления плотности нормального закона (в средней точке интервала)
 - =E\$5*exp(-(A2-1/2-E\$3)^2/(2*E\$4))/КОРЕНЬ(2*ПИ()*E\$4)*1
 - число “1” это длина интервала (в случае необходимости заменить другим числом);
 - не забывайте про знак \$, обеспечивающий неизменность ссылки при параллельном копировании;
- 9) скопировать эту ячейку вниз в столбце С до ячейки С11, соответствующей последнему интервалу (>116,05);
- 10) исправить формулу для плотности в последнем интервале
 - заменить выражение A11-1/2 на A10+1/2
 - попробуйте самостоятельно объяснить такую замену;
- 11) просуммировать значения плотности в столбце С
 - результат должен быть приблизительно равен объему выборки (ячейка С12).

Теперь уже все готово для построения графика:

- 12) выделить ячейки A2 : C11;
- 13) вызвать “Мастера Диаграмм”;
- 14) выбрать тип диаграммы «График|гистограмма» из категории “Нестандартные”;
- 15) Готово;
- 16) привести вида графика к стандартному виду:
 - щелкнуть правой кнопкой мыши по одному из столбиков и, выбрав раздел меню «Формат рядов данных»
 - в закладке \\Параметры//, уменьшить зазор до 5;
 - в закладке \\Подписи данных//, включить в подписи “значения”;
 - щелкнуть правой кнопкой мыши по графику плотности и, выбрав раздел меню «Формат рядов данных»
 - в закладке \\Вид//, добавить возможность “Сглаживания линии” и убрать “Маркеры” на линиях;
 - изменить, если очень захочется, остальные параметры диаграммы (например, убрать “Легенду”).

Замечание 1. При сравнении с показательным распределением в пункте 8 данной схемы следует изменить вычисление функции плотности

$$\text{➤ } = \exp(- (A2-1/2) / *E\$3)) / E\$3 \text{ .}$$

Замечание 2. При сравнении с равномерным на отрезке [0; 1] распределением в пункте 8 данной схемы следует изменить вычисление функции плотности

$$\text{➤ } = 1 \text{ .}$$

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Как строится гистограмма?
 Ответ: см. [4] стр. 28.
3. Каким образом следует выбирать интервалы группировки при построении гистограммы?
 Ответ: см. [4] стр. 29.
4. Сколько интервалов нужно выбирать?
 Ответ: см. [4] стр. 29.
5. Как связаны значения гистограммы и функции плотности?
 Ответ: см. [4] стр. 28.
6. Оцените вероятность попадания в интервал $[119,05; 122,05)$.
7. Почему следует сравнивать гистограмму с нормальной плотностью?
 Ответ: см. [4] стр. 30.
8. Выпишите формулу плотности нормального закона (равномерного, экспоненциального)?
 Ответ: см. [4] стр. 16-21.
9. Чему полагаются равными параметры нормального закона (равномерного, экспоненциального)?
 Ответ: см. [4] стр. 30.

Задание 3.

Эмпирическая функция распределения.

Постановка задачи.

Построить график эмпирической функции распределения с подогнанной ожидаемой функцией распределения.

Теоретические основы.

См. стр. 31-32 пособия [4].

Вычисления.

Если попытаться построить ЭФР средствами Excel, упорядочив сначала данные и сопоставив затем каждому упорядоченному значению $x_{(k)}$ значение $(k-1)/n$, то вместо горизонтальных получим наклонные ступеньки. Чтобы избежать этого недостатка, можно каждое значение вариационного ряда повторить дважды, при этом первому из этих значений сопоставить ЭФР $F_n(x_{(k)}) = (k-1)/n$, а второму $F_n(x_{(k)}) = k/n$.

Вычисление нормальной функции распределения описано ниже в главе “Встроенные функции Excel”. Здесь кратко только скажем, что для этого можно использовать функции НОРМРАСП и НОРМСТРАСП из категории “Статистические”.

Функция распределения экспоненциального закона вычисляется с помощью простой функции EXP.

Кроме того, предполагается, что уже вычислены среднее значение и дисперсия выборки (задание 1).

Пример.

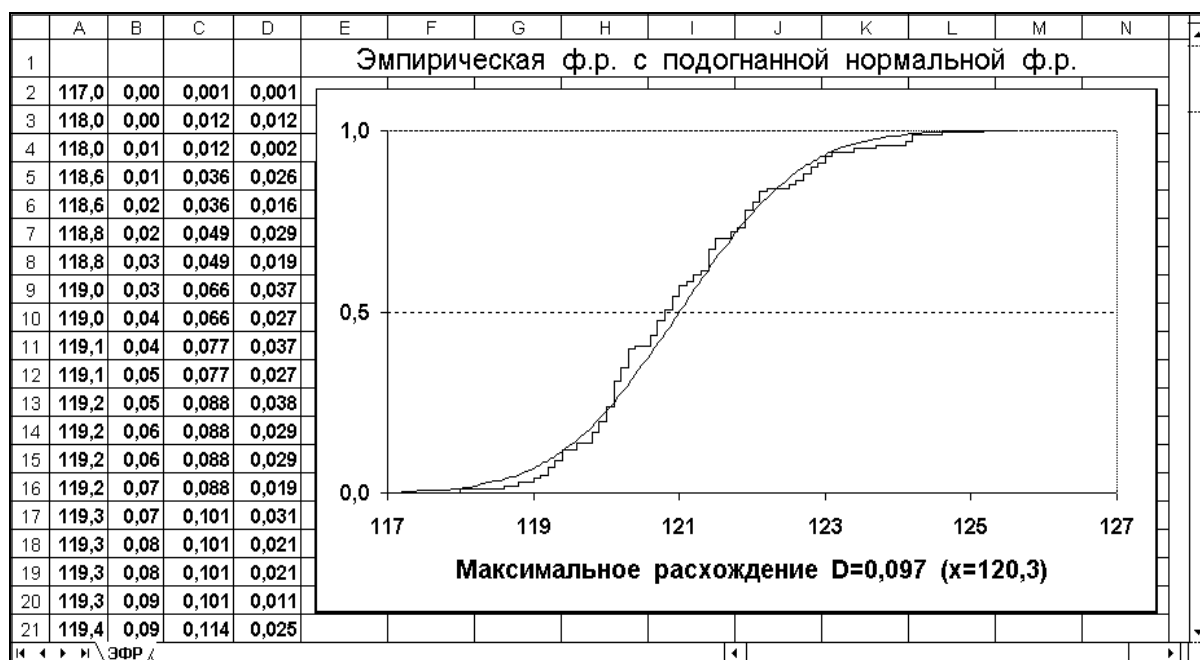


Рис. 2

☞ Порядок вычислений.

- 1) Скопировать исходные данные в буфер обмена;
- 2) перейти на лист “ЭФР” и, установив курсор в ячейку А3, вставить данные из буфера обмена;
- 3) повторить процесс восстановления данных, начиная с ячейки А104
 - установить курсор в ячейку А104;
 - вставить данные из буфера обмена
 - всего получится 202 значения с 3-й по 204-ю ячейки;
- 4) упорядочить значения в столбце А
 - кликнуть мышкой по кнопке

| | |
|---|---|
| А | ↓ |
| Я | |

;
- 5) ввести в ячейку В3 формулу
 - $= (\text{СТРОКА}(\text{В}3) - 1) / 202 - 1 / 101$

- функция «СТРОКА» возвращает номер строки указанного аргумента, то есть в данном случае в ячейке В3 получится значение $(3-1)/202-1/101 = 0$;
- 6) ввести в ячейку В4 формулу
 - $= (\text{СТРОКА}(\text{В3}) - 1) / 202$
 - получится значение $(3-1)/202 = 1/101$;
- 7) выделить обе ячейки В3 и В4 и скопировать их параллельно всем данным до ячейки В204
 - в последней ячейке должно получиться значение 1;
- 8) добавить в ячейку А2 значение, на единицу меньшее значения ячейки А3 и сопоставить ему значение 0 в ячейке В2;
- 9) добавить в ячейку А205 значение, на единицу большее значения ячейки А204 и сопоставить ему значение 1 в ячейке В205.

Ввести формулы вычисления нормального распределения:

- 10) в ячейки F4, F5 (те, которые скрыты графиком) скопировать среднее и стандартное отклонение, соответственно
 - $=\text{МОМЕНТЫ!В4}$
 - $=\text{МОМЕНТЫ!В6}$
- 11) в ячейку С2 ввести формулу нормального распределения
 - $=\text{НОРМРАСП}(\text{А2}; \$\text{F}\$4; \$\text{F}\$5; 1)$
- 12) в ячейку D2 ввести формулу вычисления расхождения между ЭФР и ожидаемой функцией распределения
 - $=\text{ABS}(\text{С2}-\text{В2})$
- 13) скопировать обе ячейки С2 и D2 вплоть до 205-й строки;
- 14) вычислить максимальное расхождение, например, в ячейке F6
 - $=\text{МАКС}(\text{D2}:\text{D205})$

Теперь уже можно рисовать графики:

- 15) выделить все значения в ячейках А2 : С205;
- 16) вызвать “Мастера Диаграмм”;

- 17) выбрать «Точечную» диаграмму – без маркеров со сглаживающей линией (третья по порядку среди точечных диаграмм);
- 18) при выборе представления диаграммы, после двух нажатий кнопки Далее, удалить “Легенду” и добавить “Заголовок по оси X”:
 - МАКСИМАЛЬНОЕ РАСХОЖДЕНИЕ $D=...$
(указав здесь полученное значение Δ из ячейки F6);
- 19) Готово;
- 20) установить параметры диаграммы, как в примере.

Замечание. Если бы параметры нормальной модели не оценивались по выборочным данным, а были бы в точности равны этим оценкам, то при полученном здесь расхождении $\Delta=0,097$ гипотезу нормальности следовало бы принять с критическим уровнем значимости $> 0,20$ (см. таблицу 6.2 сборника таблиц [1]). Это надо воспринимать как хороший знак и не более того. Если неизвестные значения параметров оцениваются по выборке, то критический уровень значимости становится зависящим от неизвестных параметров и трудно ожидать, что даже в предположениях гипотезы критерий будет иметь приемлемый размер.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Что такое вариационный ряд?
Ответ: см. [4] стр. 31.
3. Дайте определение эмпирической функции распределения?
Ответ: см. [4] стр. 31.
4. Почему некоторые ступеньки ЭФР высокие, а некоторые низкие?
Ответ: см. [4] стр. 31.
5. Почему одни ступеньки ЭФР длинные, а другие короткие?
Ответ: см. [4] стр. 31.
6. Постройте ЭФР по следующим данным: 1; 2; 1; 3; 1; 5; 1; 3.
7. Выпишите формулу для функции распределения нормального закона (равномерного, экспоненциального).
Ответ: см. [4] стр. 16-21.
8. Можно ли утверждать, что ЭФР является состоятельной оценкой истинной функции распределения? Что сие означает?
Ответ: см. [4] стр. 31.
9. Можно ли утверждать, что ЭФР является несмещенной оценкой истинной функции распределения? Что сие означает?
Ответ: см. [4] стр. 31.
10. Докажите несмещенность ЭФР.
11. Можно ли по значению максимального расхождения между ЭФР и ожидаемой функцией распределения принять или отвергнуть гипотезу о виде истинной функции распределения?
Ответ: см. [4] стр. 32.

Задание 4.

Критерий согласия хи-квадрат.

Постановка задачи.

Требуется проверить гипотезу $H_0 : F \in \Psi$ о том, что функция распределения выборочных данных F принадлежит нормальному семейству распределений Ψ (экспоненциальному, равномерному семейству).

Теоретические основы.

См. стр. 33-36 пособия [4].

Вычисления.

Пакет Excel предоставляет возможность вычисления как значений функции надежности $\bar{K}_m(x) = 1 - K_m(x)$, так и значений p -квантилей хи-квадрат распределения. Эти функции называются ХИ2РАСП и ХИ2ОБР.

Для вычисления нормального распределения можно использовать функцию НОРМРАСП. Подробнее см. ниже в главе “Встроенные функции Excel”.

Интервалы группировки и частоты попадания в эти интервалы могут быть взяты из задания 2.

Ниже приведен фрагмент листа Excel с примером вычислений, проводимых при построении критерия хи-квадрат для проверки гипотезы нормальности.

Пример.

| | A | B | C | D | E | F | G | H |
|----|--------------|-------------------|------------------|---------------------|----------------|---|---|---|
| 1 | Грани- цы | Выбор. частоты | Веро- ятность | Ожидаем. частоты | Хи- квадрат | | | |
| 2 | | | 0 | | | | | |
| 3 | 117,05 | 0 | 0,001 | 0,15 | 0,151 | Критический уровень значимости | | |
| 4 | 118,05 | 1 | 0,013 | 1,19 | 0,031 | | | |
| 5 | 119,05 | 3 | 0,071 | 5,86 | 1,395 | <div>9 ст.свободы</div> <div>$\alpha_9 =$ 0,101</div> <div>7 ст.свободы</div> <div>$\alpha_7 =$ 0,041</div> | | |
| 6 | 120,05 | 20 | 0,237 | 16,76 | 0,624 | | | |
| 7 | 121,05 | 34 | 0,514 | 27,98 | 1,293 | | | |
| 8 | 122,05 | 23 | 0,784 | 27,27 | 0,669 | | | |
| 9 | 123,05 | 13 | 0,938 | 15,51 | 0,407 | | | |
| 10 | 124,05 | 3 | 0,989 | 5,15 | 0,896 | | | |
| 11 | 125,05 | 4 | 0,999 | 0,99 | 9,076 | <div>$\alpha_{\text{крит}} =$ 0,05</div> | | |
| 12 | >125,05 | 0 | 1 | 0,12 | 0,119 | | | |
| 13 | Всего | 101 | | 101,00 | | Вывод | | |
| 14 | | | | | | Гипотеза нормальности не может быть принята или отвергнута | | |
| 15 | | | | $\chi^2=$ | 14,66 | | | |

🔗 Порядок вычислений.

- 1) Скопировать ячейки A2:B11 с листа “Гисто”, содержащие выборочные частоты, на рабочий лист в ячейки A3:B12.
- 2) В ячейке подсчитать B13 общее число данных
➤ =СУММ(A3:B12)

В столбце C (Вероятность) вычислить значение гипотетической функции распределения $F_0(x_i)$:

- 3) в ячейку C3 (напротив первой границы) ввести формулу
➤ =НОРМРАСП(A3; Моменты!\$B\$4; Моменты!\$B\$6; 1)
– напомним, что в ячейках B4 и B6 на листе “Моменты” хранятся среднее и стандартное отклонение;
- 4) скопировать ячейку C3 во все ячейки столбца C вплоть до ячейки, соответствующей последней границе (C11);

- 5) в ячейке C2 указать значение 0 (соответствует $F_0(-\infty)$), а в ячейке C12 – значение 1 (соответствует $F_0(+\infty)$).

В столбце D (Ожидаемые частоты) вычислить теоретические частоты $p_i = n \cdot (F_0(x_i) - F_0(x_{i-1}))$:

- 6) в ячейку D3 ввести формулу
➤ = \$B\$13 * (C3 - C2)
- 7) скопировать ячейку D3 в столбце D до ячейки D12 (напротив границы “>125,05”);
- 8) для контроля в ячейке D13 (Всего) вычисляется сумма значений в столбце D (должно получиться число 1).

В столбце E (Хи-квадрат) вычислить слагаемые статистики X^2 :

- 9) в ячейку E3 ввести формулу
➤ = (B3 - D3) ^ 2 / (\$B\$13 * D3)
– в ячейке \$B\$13 хранится объем выборки;
- 10) скопировать ячейку E3 в столбце E до ячейки E12;
- 11) в ячейке E15 вычислить сумму значений столбца E – искомое значение статистики X^2 .

Вычислить уровни значимости $\alpha_{r-3}, \alpha_{r-1}$:

- 12) в ячейки G6, G8 ввести формулы
➤ =ХИ2РАСП (E15; 10 - 1)
➤ =ХИ2РАСП (E15; 10 - 3)
“10” – это число групп, для каждого студента оно может быть разным.

Вывод в ячейке F11 сделан в соответствии с правилом, описанным в пункте VII на стр. 36 пособия [4].

Замечание 1. При проверке гипотезы экспоненциальности необходимо заменить ячейки, в которых вычисляются значения вероятностей попадания в интервалы (пункт 2) в соответствии с формулой экспоненциального распределения

$$\triangleright = 1 - \text{EXP}(-A3 / \text{Моменты!} \$B\$4)$$

Критический уровень значимости вычисляется при $r-1$ и при $r-2$ степенях свободы.

Замечание 2. При проверке гипотезы равномерности необходимо, во-первых, выбрать равномерное разбиение отрезка $[0;1]$. Во-вторых, нужно заменить ячейки, в которых вычисляются значения вероятностей попадания в интервалы (пункт 2) в соответствии с формулой равномерного распределения

$$\triangleright = A3$$

Критический уровень значимости вычисляется при $r-1$ степени свободы.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Как строится критерий согласия хи-квадрат?
Ответ: см. [4] стр. 33.
3. Почему критерий называется критерием согласия?
Ответ: см. [4] стр. 36.
4. Выпишите формулу тестовой статистики критерия согласия хи-квадрат. Почему эту статистику можно считать мерой близости выборочных данных к выдвинутой гипотезе?
Ответ: см. [4] стр. 33.
5. Какое распределение имеет статистика критерия хи-квадрат?
Ответ: см. [4] стр. 35.
6. Почему иногда приходится вычислять два критических уровня значимости?
Ответ: см. [4] стр. 36.
7. Чему равен критический уровень значимости при проверке гипотезы о равномерном (нормальном, экспоненциальном) распределении?
Ответ: см. [4] стр. 36.
8. Почему при построении критерия хи-квадрат нельзя выбирать интервалы группировки в зависимости от выборочных данных?
Ответ: см. [4] стр. 36.

Задание 5.

Одновыборочный критерий Стьюдента.

Постановка задачи.

Двухвыборочный вариант. Имеются две выборки $(x_1, \dots, x_n), (y_1, \dots, y_n)$ одинакового объема. Известно, что распределения в этих выборках подчинены нормальному закону и, кроме того, каждое i -ое наблюдение x_i в 1-ой выборке зависит (в вероятностном смысле) от соответствующего i -ого наблюдения y_i во второй выборке. Требуется проверить гипотезу однородности выборок. Точнее, требуется проверить гипотезу о том, что среднее значение разности выборок равно нулю (меньше нуля, больше нуля).

Одновыборочный вариант. Имеется одна выборка из нормального распределения. Требуется проверить гипотезу о том, что среднее значение этого распределения не превосходит заданной величины $C_{\text{норм}}$.

Теоретические основы.

См. стр. 37-40 пособия [4].

Вычисления.

В приложении Excel для работы с распределением Стьюдента имеются встроенные функции

| | |
|--------------|-----------------------------------|
| СТЮДРАСП, | вычисляющая функцию надежности, и |
| СТЮДРАСПОБР, | вычисляющая верхние квантили. |

Подробнее см. ниже в главе “Встроенные функции Excel”.

Пример.

Измерялось верхнее артериальное давление у 10 пациентов до и после лечения. Требуется проверить эффективность лечения.

Если эти пациенты страдают гипертонической болезнью, то ожидается, что давление после лечения будет понижаться. Поэтому в качестве альтернативы здесь нужно выдвинуть утверждение о том, что среднее значение разности давлений до и после лечения будет больше нуля.

| | A | B | C | D | E | F | G | H |
|----|-------|-------|---------------|---|---|--------------|------------------------|--------------|
| 1 | До | После | Раз- ность | | | | | |
| 2 | 162,8 | 139 | 23,8 | | Выборочные характеристики | | | |
| 3 | 186,9 | 189 | -2,1 | | | | | |
| 4 | 167,2 | 162 | 5,2 | | | До | После | Разность |
| 5 | 166,5 | 168,6 | -2,1 | | Среднее \bar{x} | 168,94 | 152,88 | 16,06 |
| 6 | 173 | 164,9 | 8,1 | | Ст.Отклон. s | 7,48 | 19,39 | 13,96 |
| 7 | 164,1 | 137,9 | 26,2 | | Ош.средн. (+-) m | 2,49 | 6,46 | 4,65 |
| 8 | 158,3 | 121,7 | 36,6 | | Объем выборки n | 10 | 10 | 10 |
| 9 | 168,4 | 129,9 | 38,5 | | Статистика Стьюдента | | | |
| 10 | 174,8 | 160,5 | 14,3 | | | | 3,45 | |
| 11 | 167,4 | 155,3 | 12,1 | | | | | |
| 12 | | | | | Гипотеза | Альтернатива | $\alpha_{\text{крит}}$ | Принимается |
| 13 | | | | | Не изменилось | Стало меньше | 0,004 | Альтернатива |
| 14 | | | | | Вывод: выборочные данные высоко значимо подтверждают эффективность лечения | | | |

☞ Порядок вычислений.

- 1) Ввести данные в столбцы A и B.
- 2) В столбце C вычислить разности значений до и после лечения
 - в ячейку C2 ввести формулу
 - =A2-B2
 - скопировать ячейку C2 параллельно данным столбца A.
- 3) Вычислить основные характеристики столбцов A, B и C
 - средние значения (ячейки F5, G5, H5)
 - функция СРЗНАЧ;
 - стандартные отклонения (ячейки F6, G6, H6)

- функция СТАНДОТКЛОНП;
 - объемы выборок (ячейки F8, G8, H8)
 - функция СЧЕТ;
 - стандартные ошибки среднего – формула
 $=F6/КОРЕНЬ(F8)$
 введенная в ячейку F7 и скопированная в ячейки G7, H7.
- 4) В ячейке G11 вычислить статистику Стьюдента
- $=H5/H7$
 $\{ = \bar{u} / \frac{s}{\sqrt{n-1}} \}.$
- 5) В ячейке G11 вычислить критический уровень значимости
- $=СТЮДРАСП(G11; H8-1; 1)$
- $\{ = 1 - S_{n-1}(t), \text{ “1” – число хвостов для односторонней альтернативы } H_1 : \mu_1 - \mu_2 > 0 \}.$
- 6) В ячейке H15 сделать вывод о предпочтении гипотезы или выбранной альтернативы.

Замечание 1. Если по каким-либо причинам вычислялось стандартное отклонение, основанное на несмещенной оценке дисперсии (функция «СТАНДОТКЛОН»), то в формуле для ошибки среднего необходимо делитель $\sqrt{n-1}$ заменить на \sqrt{n} . Аналогичную замену нужно производить и при вычислениях T .

Замечание 2. Описанную выше схему с очевидными изменениями можно применить и в случае одновыборочного варианта задания. Например, если перед исследователем стояла задача полного излечения гипертонических больных, то необходимо было бы проверить гипотезу о том, что среднее значение верхнего артериального давления у пациентов, прошедших курс лечения, будет больше 125 при альтернативе меньше 125. Всё отличие вычислений в этом случае будет состоять в рассмотрении разности давлений после лечения у каждого пациента минус нормативная граница 125.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Как вычисляется статистика одновыборочного критерия Стьюдента?
Ответ: см. [4] стр. 39.
3. Почему к рассматриваемым данным нельзя применить двухвыборочный критерий Стьюдента?
Ответ: см. [4] стр. 38.
4. Когда следует применять критерий Стьюдента, а когда критерий знаков?
Ответ: см. [4] стр. 38, 40.
5. Чему равен критический уровень значимости для критерия Стьюдента при двухсторонней альтернативе?
Ответ: см. [4] стр. 39.
6. Можно ли к рассматриваемым данным применить критерий однородности хи-квадрат?
Ответ: см. [4] стр. 48.
7. Можно ли однозначно утверждать, что пребывание в спортивном летнем лагере повышает спортивную форму, если средний вес случайно отобранной части студентов после пребывания в лагере уменьшился на 7 кг?
Ответ: см. [4] стр. 8.
8. Что еще нужно знать, чтобы правильно ответить на предыдущий вопрос?
9. Найдите по таблице критический уровень значимости при двухсторонней альтернативе, если значение $t = 2,1$ и объем выборки $n = 29$.
Ответ: см. [4] стр. 20, 39 или ниже стр. 76-77.

Задание 6.

Критерий знаков.

Постановка задачи.

Двухвыборочный вариант. Имеются две выборки $(x_1, \dots, x_n), (y_1, \dots, y_n)$ одинакового объема. Известно, что каждое i -ое наблюдение x_i в 1-ой выборке зависит (в вероятностном смысле) от соответствующего i -ого наблюдения y_i во второй выборке. Распределение выборок неизвестно. Требуется проверить гипотезу однородности выборок.

Одновыборочный вариант. Имеется одна выборка. Требуется проверить гипотезу, что некоторое фиксированное событие происходит чаще, чем противоположное к этому событию утверждение (например, лечение чаще приводит к выздоровлению).

Теоретические основы.

См. стр. 40-42 пособия [4].

Вычисления.

При малых значениях n критический уровень значимости может быть вычислен с использованием простого калькулятора. Например, если в эксперименте наблюдалось $m = 8$ в $n = 10$ испытаниях, то критический уровень значимости можно вычислить так:

| m | 8 | 9 | 10 | Сумма | 2^{10} | $\alpha_{\text{крит}}$ |
|----------------------------|----|----|----|-------|----------|-------------------------|
| $C_{10}^m = C_{10}^{10-m}$ | 45 | 10 | 1 | 56 | 1024 | $56/1024 \approx 0,055$ |

Пакет Excel имеет встроенную функцию БИНОМРАСП (см. стр. 74), которая позволяет вычислять вероятность $P\{M \leq m | n, p\}$ для биномиального распределения. Нам необходима вероятность противоположного события, причем значение m должно быть учтено при вычислении этой вероятности. Таким образом, для вычисления $\alpha_{\text{крит}}$ следует использовать функцию

$$=1 - \text{БИНОМРАСП}(m - 1; n; 0,5; 1)$$

(единица отнимается с целью учета значения m).

Пример.

Рассмотрим данные, которые использовались для иллюстрации одновыборочного критерия Стьюдента.

| | A | B | C | D | E | F |
|----|-------|-------|--------|---|--|-------|
| 1 | До | После | Эффект | | | |
| 2 | 162,8 | 139 | 1 | | Число наблюдений n | 10 |
| 3 | 186,9 | 189 | 0 | | Число успехов m | 8 |
| 4 | 167,2 | 162 | 1 | | | |
| 5 | 166,5 | 168,6 | 0 | | Уровень значимости | |
| 6 | 173 | 164,9 | 1 | | $\alpha_{\text{крит}} =$ | 0,055 |
| 7 | 164,1 | 137,9 | 1 | | | |
| 8 | 158,3 | 121,7 | 1 | | Наличие эффекта | |
| 9 | 168,4 | 129,9 | 1 | | слабо значимо | |
| 10 | 174,8 | 160,5 | 1 | | Вывод: по-видимому, препарат способствует уменьшению давления. Для уточнения необходимо провести дополнительное исследование. | |
| 11 | 167,4 | 155,3 | 1 | | | |

☞ Порядок вычислений.

- 1) В столбце C указать наличие эффекта для каждой пары данных;
- 2) в ячейке F2 вычислить количество пар данных;
- 3) в ячейке F3 подсчитать число пар с наличием эффекта;
- 4) в ячейке F6 вычислить критический уровень значимости
 $\rightarrow =1-\text{БИНОМРАСП}(F3-1; F2; 0,5; 1)$
- 5) сделать вывод о степени влиянии лечения на артериальное давление.

Замечание 1. Проведенные вычисления можно разместить на том же листе, где строился одновыборочный критерий Стьюдента.

Замечание 2. Если для этих данных построить нижнюю 95%-доверительную границу для вероятности эффекта (см. Задание 13), то получим $\underline{p} = 0,493$, что говорит в пользу гипотезы, поскольку интервал $(0,493; 1]$ не попадает полностью в область альтернативы $(0,5; 1]$ (см. способ проверки гипотезы, основанный на доверительной границе, описанный в [4]). Если же построить 90%-ю границу, то получим $\underline{p} = 0,505$, свиде-

тельствующее в пользу альтернативы. Это объясняет тот странный вывод, что сделан нами по результатам статистической обработки.

Замечание 3. Описанную схему можно применять также для проверки гипотезы о вероятности “успеха” при биномиальных испытаниях – одно-выборочный вариант критерия. Это связано с тем, что при применении критерия знаков вывод основывается исключительно на количестве положительных эффектов и не зависит от того, как это количество было подсчитано.

Например, если бы перед исследователем стояла задача окончательного излечения больных гипертонией, то для представленных данных мы имели бы лишь один положительный эффект: $m=1$. В этом случае

$$\alpha_{\text{крит}} = 1 - \text{БИНОМРАСП}(0; 10; 0,5; 1) = 0,999.$$

То есть, несколько преждевременно говорить, что лечение приводит к выздоровлению.

В качестве другого **примера** рассмотрим ситуацию, когда при составлении договора купли-продажи заказчиком была оговорена верхняя граница в 8% для доли бракованной продукции. При поступлении товара заказчик проводит контрольные измерения n единиц продукции. По результатам испытаний требуется проверить гипотезу $H_0: p \geq 0,08$ (опять же гипотеза противоположна ожиданиям). Все вычисления в данном случае будут аналогичны вышеприведенным. Единственное отличие возникнет при нахождении критического уровня значимости – здесь нужно, во-первых, вычислять не функцию надежности, а функцию распределения биномиального закона (объясните самостоятельно), и, во-вторых, заменить граничное значение гипотезы 0,5 на значение 0,08. Например, если среди 37 проконтролированных изделий было обнаружено 1 некондиционное (т.е. 2,7% от объема контроля), то критический уровень значимости равен

$$\alpha_{\text{крит}} = \text{БИНОМРАСП}(1; 37; 0,08; 1) = 0,193.$$

Другими словами, нет достаточных оснований утверждать, что продукция удовлетворяет требованиям заказчика (гипотеза $H_0: p \geq 0,08$ не отвергается). Может быть, надо провести еще ряд контрольных замеров.

Еще один **пример** см. пособие [4] стр. 42.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Чему равна статистика критерия знаков?
Ответ: см. [4] стр. 41.
3. Чему равен критический уровень значимости критерия знаков?
Ответ: см. [4] стр. 41.
4. Когда следует применять критерий Стьюдента, а когда критерий знаков?
Ответ: см. [4] стр. 38, 40.
5. Чем, как Вы думаете, обусловлен неоднозначный вывод в нашем первом примере?
6. Вычислите с помощью калькулятора значение критического уровня значимости, если число успехов равно 6 при 9 испытаниях.
Ответ: см. выше стр. 33.
7. Проверьте гипотезу о том, что вероятность рождения мальчика равна 0,515, если среди 1000 новорожденных детей 509 оказались мальчики.
Ответ: см. [4] стр. 41-42.

Задание 7.

Двухвыборочный критерий Стьюдента.

Постановка задачи.

Имеются две выборки (x_1, \dots, x_{n_1}) , (y_1, \dots, y_{n_2}) , относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону с одинаковыми для обеих выборок дисперсиями. Требуется проверить гипотезу однородности выборок, то есть гипотезу совпадения средних значений.

Теоретические основы.

См. стр. 42-43 пособия [4].

Вычисления.

При построении этого критерия применяются те же вспомогательные функции Excel, что для одновыборочного критерия Стьюдента (задание 5).

Пример.

В качестве иллюстрации рассмотрим пример изучения верхнего артериального давления у пациентов, прошедших два различных курса лечения. В первую группу мы отнесли данные о пациентах, лечившихся по новой методике, во вторую – по стандартной методике. Естественные ожидания авторов новой методики состоят в том, что среднее первой выборки будет меньше среднего второй выборки. Таким образом, необходимо проверить гипотезу $H_0 : \mu_1 \geq \mu_2$ при альтернативе $H_1 : \mu_1 < \mu_2$ (напомним: альтернатива суть ожидания исследователя). Критерий Стьюдента для такой гипотезы полностью совпадает с критерием Стьюдента для простой гипотезы $H_0 : \mu_1 = \mu_2$ при той же альтернативе.

| | A | B | C | D | E | F | G | H |
|----|----------------|----------------|---|---|--|-------------------|------------------------|------------------|
| 1 | 1-я выборка | 2-я выборка | | | Выборочные характеристики | | | |
| 2 | 139 | 119 | | | | | | |
| 3 | 136 | 143 | | | | | | |
| 4 | 126 | 139 | | | | 1-я выборка | 2-я выборка | |
| 5 | 143 | 130 | | | Среднее \bar{x} | 144,6 | 131,375 | |
| 6 | 154 | 144 | | | Ст.Отклон. s | 10,03 | 10,26 | |
| 7 | 156 | 114 | | | Ош.средн. (+-) m | 2,49 | 6,46 | |
| 8 | 155 | 135 | | | Объем выборки n | 10 | 8 | |
| 9 | 143 | 127 | | | Уровень значимости | | | 0,05 |
| 10 | 157 | | | | Статистика Стьюдента | | | |
| 11 | 139 | | | | | | | 2,59 |
| 12 | | | | | | | | |
| 13 | | | | | Гипотеза | Альтернати- ва | $\alpha_{\text{крит}}$ | Прини- мается |
| 14 | | | | | Равны | 1-я меньше | 0,99 | Гипотеза |
| | | | | | Вывод: выборочные данные не свидетельствуют в пользу новой методики. | | | |

☞ Порядок вычислений здесь вполне идентичен вычислениям, проводимым при реализации одновыборочного критерия Стьюдента. Отличия заключаются только в том, что не надо вычислять разности, а статистика Стьюдента (ячейка G11) вычисляется как

$$\begin{aligned} & \rightarrow = (F5 - G5) * \text{КОРЕНЬ} (F8 * G8 * (F8 + G8 - 2) / ((F8 + G8) * (F8 * F6 + G8 * G6))) \\ & \{ = (\bar{x} - \bar{y}) * \sqrt{n_1 * n_2 * (n_1 + n_2 - 2) / ((n_1 + n_2) * (n_1 s_1^2 + n_2 s_2^2))} \}. \end{aligned}$$

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Как вычисляется статистика двухвыборочного критерия Стьюдента?
 Ответ: см. [4] стр. 43
3. Почему к рассматриваемым данным нельзя применить одновыборочный критерий Стьюдента?
 Ответ: см. [4] стр. 37-38, 42.
4. Когда следует применять критерий Стьюдента, а когда критерий Вилкоксона?
 Ответ: см. [4] стр. 42-43.
5. Чему равен критический уровень значимости для критерия Стьюдента при двухсторонней альтернативе; при односторонней альтернативе типа – «в первой группе больше»?
 Ответ: см. [4] стр. 43.
6. Можно ли к рассматриваемым данным применить критерий однородности хи-квадрат?
 Ответ: см. [4] стр. 48.
7. Что такое ошибка среднего? Какую смысловую нагрузку она несет применительно к рассматриваемому критерию?
 Ответ: см. [4] стр. 55.
8. Какой критерий следует применять для проверки гипотезы равенства средних значений, если известно, что обе выборки получены из экспоненциального распределения?
 Ответ: см. [4] стр. 43, 48.

Задание 8.

Критерий Вилкоксона.

Постановка задачи.

Имеются две выборки (x_1, \dots, x_{n_1}) , (y_1, \dots, y_{n_2}) , относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда ожидается, что значения в 1-й выборке будут меньше значений во второй выборке.

Теоретические основы.

См. стр. 43-46 пособия [4].

Вычисления.

Для нахождения ранга значения x в ряду данных Q в среде Excel можно использовать функцию РАНГ(x ; Q ; 1) (третий аргумент функции, равный 1, указывает на порядок расположения по возрастанию). Всем совпадающим значениям в ряду Q функция РАНГ присваивает одинаковое значение, равное меньшему рангу. Чтобы исправить её в соответствии с вышеописанной схемой, можно поступить следующим образом.

1. Для каждого значения x 1-й выборки подсчитать количество данных (в обеих выборках), совпадающих с x , воспользовавшись функцией

$$\text{СЧЁТЕСЛИ}(Q; x).$$

2. Если таких значений наберется K штук, то их средний ранг будет равен $R + (K-1)/2$, где R – результат применения функции РАНГ к значению x .

Для вычисления приближенного значения уровня значимости с помощью нормальной аппроксимации можно воспользоваться функцией НОРМРАСП.

Пример.

Обратимся снова к данным, иллюстрировавшим применение двухвыборочного критерия Стьюдента. Для этих данных как раз ожидаемо, что первая выборка (результат лечения новым препаратом) будет “левее” второй выборки. Поэтому при отсутствии информации о нормальности распределения выборок вполне уместно будет применить критерий Вилкоксона.

| | A | B | C | D | E | F | G |
|----|----------------|----------------|---|----------------------|-------------------------------|---------------------------------|-------|
| 1 | 1-я выборка | 2-я выборка | Число равных | Ранги 1-й выборки | | | |
| 2 | 139 | 119 | 3 | 9 | | Среднее статистики Вилкоксона | |
| 3 | 134 | 143 | 1 | 6 | | 95,5 | |
| 4 | 126 | 139 | 1 | 3 | | Дисперсия статистики Вилкоксона | |
| 5 | 143 | 130 | 3 | 12 | | 126,67 | |
| 6 | 154 | 144 | 1 | 15 | | | |
| 7 | 156 | 114 | 1 | 17 | | Приближенное значение | |
| 8 | 155 | 135 | 1 | 16 | | уровня значимости | |
| 9 | 143 | 127 | 3 | 12 | | $\alpha_{\text{крит}} =$ | 0,972 |
| 10 | 157 | | 1 | 18 | | | |
| 11 | 139 | | 3 | 9 | | | |
| 12 | | | С _{крит} =75 | | | | |
| 13 | Количество | | Статистика Вилкоксона | | Гипотезу однородности следует | | |
| 14 | 10 | 8 | 117 | | принять на уровне > 95% | | |
| 15 | | | Вывод: выборочные данные не свидетельствуют в пользу новой методики. | | | | |
| 16 | | | | | | | |

☞ Порядок вычислений.

- 1) В ячейке C2 вычислить количество совпадений с 1-ым элементом 1-ой выборки
➤ =СЧЕТЕСЛИ(\$A\$2:\$B\$11; A2)
- 2) в ячейку D2 ввести функцию вычисления ранга 1-ого элемента 1-ой выборки
➤ =РАНГ(A2; \$A\$2:\$B\$11; 1) + (C2-1) / 2
(не забудьте про знаки \$);
- 3) скопировать ячейки C2 и D2 параллельно данным столбца A;
- 4) в ячейках A14 и B14 подсчитать количество данных n_1 и n_2 в каждой выборке;

- 5) найти в сборнике таблиц [1] значение критической константы $C_{\text{крит}}$ для полученных значений n_1 и n_2 (ячейка D12);
- 6) в ячейке B14 вычислить сумму рангов столбца D (статистику Вилкоксона)
 - =СУММ (D2 : D11)
- 7) сравнив полученное значение с $C_{\text{крит}}$, сделать вывод об однородности или неоднородности выборок
 - в нашем случае нет оснований утверждать, что в первой выборке значения меньше, чем во второй:
 $w = 117 \nless C_{\text{крит}} = 75$.

С целью подтверждения вывода вычислить приближенное значение критического уровня значимости:

- 8) в ячейке G3 вычислить среднее значение статистики W
 - =A14 * (A14+B14+1) / 2+0,5
- 9) в ячейке G5 вычислить дисперсию статистики W
 - =A14*B14*(A14+B14+1) / 12
- 10) в ячейке G9 вычислить критический уровень значимости
 - =НОРМРАСП (C14; G3; КОРЕНЬ (G5) ; 1)
- 11) сделать вывод об однородности выборок (ячейка G14).

Замечание. Если бы мы в качестве альтернативы рассмотрели гипотезу H_1 : «2-ая выборка сдвинута влево», то нам пришлось бы отвергнуть гипотезу однородности в пользу альтернативы – $\alpha_{\text{крит}} \approx 0,028$. Однако, как говорится, “после выборки критическим значением не размахивают”. Как гипотеза, так и альтернатива должны выбираться до получения выборки. На худой конец, можно было бы выдвинуть двухстороннюю альтернативу – $\alpha_{\text{крит}} \approx 0,056$ и, в случае её принятия, скажем, на 10%-ом уровне, выбрать направленность соотношения между выборками визуально.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Как вычисляется статистика критерия Вилкоксона?
 Ответ: см. [4] стр. 44.
3. При каких альтернативах следует прибегать к критерию Вилкоксона?
 Ответ: см. [4] стр. 43.
4. Как присваивать ранги совпадающим значениям?
 Ответ: см. [4] стр. 45.
5. Чему равен критический уровень значимости критерия Вилкоксона?
 Ответ: см. [4] стр. 45.
6. Найдите по таблице критическое значение для объемов выборок $n_1 = 12$ и $n_2 = 14$.
7. Какой критерий следует применять, если в качестве альтернативы к гипотезе однородности выдвинуто утверждение о том, что первая выборка получена из нормального распределения, а вторая из экспоненциального?
 Ответ: см. [4] стр. 48.

Задание 9.

Проверить гипотезу равенства дисперсий по критерию Фишера.

Постановка задачи.

Имеются две выборки (x_1, \dots, x_{n_1}) , (y_1, \dots, y_{n_2}) , относящиеся к двум независимым группам наблюдений одной и той же характеристики, подчиняющейся нормальному закону. Требуется сравнить дисперсии наблюдений в этих групп.

Теоретические основы.

См. стр. 46-47 пособия [4].

Вычисления.

В среде Excel для отыскания критического уровня значимости можно воспользоваться функцией FРАСП из категории “Статистические”, которая вычисляет значения функции надежности $1 - F_{k,m}(f)$ распределения Фишера. Способ вызова этой функции вполне тривиален.

Пример.

Воспользуемся снова данными из задания 7. В качестве альтернативы возьмем утверждение о несовпадении дисперсий. При выполнении задания можно оба эти критерия объединить, расположив приводимые ниже вычисления на том же самом листе, где строился двухвыборочный критерий Стьюдента.

| | A | B | C | D | E | F | G | H |
|----|----------------|----------------|---|---|--|---------------------------|------------------------|------------------|
| 1 | 1-я выборка | 2-я выборка | | | | | | |
| 2 | 139 | 119 | | | | Выборочные характеристики | | |
| 3 | 134 | 143 | | | | | | |
| 4 | 126 | 139 | | | | 1-я выборка | 2-я выборка | |
| 5 | 143 | 130 | | | | | | |
| 6 | 154 | 144 | | | | | | |
| 7 | 156 | 114 | | | Дисперсия s^2 | 100,64 | 105,23 | |
| 8 | 155 | 135 | | | Объем n | 10 | 8 | |
| 9 | 143 | 127 | | | | Уровень значимости | | 0,10 |
| 10 | 157 | | | | | Статистика Фишера | | |
| 11 | 139 | | | | | | 0,744 | |
| 12 | | | | | | | | |
| 13 | | | | | Гипотеза | Альтерна- тива | $\alpha_{\text{крит}}$ | Прини- мается |
| 14 | | | | | Дисперсии равны | Не равны | 0,66 | Гипотеза |
| | | | | | Вывод: дисперсии обеих выборок одинаковы. | | | |

☞ Порядок вычислений.

- В ячейках F7 и G7 вычислить дисперсии обеих выборок
 - использовать функцию ДИСПР;
- в ячейках F8 и G8 вычислить объемы выборок
 - использовать функцию СЧЕТ;
- в ячейке G11 вычислить значение статистики Фишера
 - $=F7 * (G8 - 1) / (G7 * (F8 - 1))$;
- в ячейке G14 вычислить критический уровень значимости при двухсторонней альтернативе
 - $=2 * (1 - \text{ФРАСП}(G11; F8 - 1; G8 - 1))$
(здесь приведена формула для случая, когда “верхняя” дисперсия меньше “нижней”, как и получено в эксперименте);
- в ячейке H14 отдать предпочтение гипотезе или альтернативе в зависимости от значения ячейки G14 и выбранного уровня значимости (ячейка H9).

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Чему равна статистика Фишера?
Ответ: см. [4] стр. 47.
3. Какое распределение имеет статистика Фишера?
Ответ: см. [4] стр. 47.
4. Чему равен критический уровень значимости при односторонней альтернативе H_1 : «дисперсия 1-ой выборки меньше дисперсии 2-ой выборки»?
Ответ: см. [4] стр. 47.
5. Можно ли применить критерий Фишера к проверке гипотезы о равенстве дисперсий для данных из задания 5?
6. Можно ли применять критерий Фишера, как предварительный тест для проверки условий применимости критерия Стьюдента?
Ответ: см. [4] стр. 47.
7. Найдите по таблице критическое значение для статистики Фишера при уровне значимости $\alpha = 0,01$ и объёмах выборок $n_1 = 21$, $n_2 = 25$.

Задание 10.

Критерий однородности хи-квадрат.

Постановка задачи.

Имеются две выборки (x_1, \dots, x_{n_1}) , (y_1, \dots, y_{n_2}) , относящиеся к двум независимым группам наблюдений одной и той же характеристики. Требуется проверить гипотезу однородности выборок в ситуации, когда неизвестна модель распределения выборок и нет никакой информации о соотношении между этими выборками.

Теоретические основы.

См. стр. 48-50 пособия [4].

Пример.

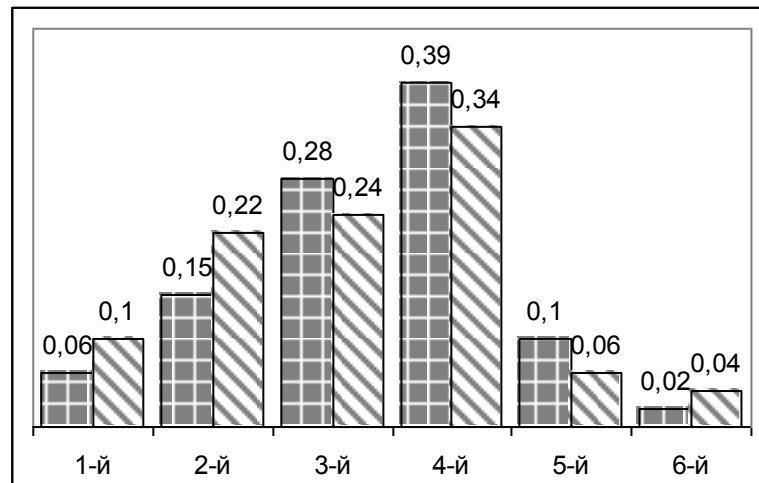
Изучалось влияние жевательных резинок “Обрит-без” на степень усвоения школьниками усложненного материала по географии. Знания оценивались по 6-тибальной шкале (шесть интервалов группировки). Выборка жующих состояла из 100 школьников, выборка крепко сжавших рот – из 50 школьников. В таблице ниже в столбцах В и D приведены количества школьников, попавших в ту или иную группу (получивших ту или иную оценку).

| | A | B | C | D | E | F | G |
|----|---------------------|---|--------|-----------------|--------|--------------------------|----------------|
| 1 | Инт-вал (группа) | Частоты | | | | Сумма частот | Хи- квадрат |
| | | выборка 1 | | Выборка 2 | | | |
| 2 | | Абсол. | Относ. | Абсол. | Относ. | | |
| 3 | 1-й | 6 | 0,06 | 5 | 0,10 | 11 | 0,727 |
| 4 | 2-й | 15 | 0,15 | 11 | 0,22 | 26 | 0,942 |
| 5 | 3-й | 28 | 0,28 | 12 | 0,24 | 40 | 0,200 |
| 6 | 4-й | 39 | 0,39 | 17 | 0,34 | 56 | 0,223 |
| 7 | 5-й | 10 | 0,10 | 3 | 0,06 | 13 | 0,615 |
| 8 | 6-й | 2 | 0,02 | 2 | 0,04 | 4 | 0,500 |
| 9 | Всего | 100 | 1 | 50 | 1 | 150 | 3,208 |
| 10 | | | | | | | |
| 11 | | Статистика X 2 | | Степени свободы | | Уровень значимости | |
| 12 | | X 2 = | 3,208 | r -1 = | 5 | $\alpha_{\text{крит}}$ = | 0,67 |
| 13 | | Вывод: Степень обучаемости не за- висит от методики. | | | | | |

☞ Порядок вычислений.

- 1) В ячейках B9, D9 и F9 найти общее количество индивидов в первой и второй выборках и общее число индивидов;
- 2) в ячейке C3 вычислить относительную частоту попадания в 1-ую группу 1-ой выборки
➤ =B3/B9
- 3) скопировать ячейку C3 в ячейки C4:C8 столбца C и ячейки E3:E8 столбца E;
- 4) с целью проверки правильности вычислений в ячейках C9 и E9 найти сумму относительных частот – должна получиться 1;
- 5) в ячейке F3 вычислить сумму ячеек столбцов B и D
➤ =B3+D3;
- 6) в ячейке G3 вычислить хи-квадрат расхождение между частотами 1-ой группы
➤ =\$B\$9*\$D\$9*(C3-E3)^2/F3
- 7) скопировать ячейку G3 в ячейки G4:G8;

- 8) в ячейке G9 вычислить статистику хи-квадрат (для удобства восприятия значение в этой ячейке повторено в C12)
 - =СУММ(G3:G8)
- 9) в ячейке G12 вычислить критический уровень значимости и сделать вывод о значимости отклонения данных от гипотезы однородности
 - =ХИ2РАСП(G9; 6-1)
 - (”6” – число интервалов группировки);
- 10) построить график относительных частот
 - выделить ячейки A3:A8, C3:C8, E3:E8 и вызвать функцию построения обычных гистограмм, соответствующим образом изменив ее вид



Поскольку $\alpha_{\text{крит}}$ велико, сделан вывод о справедливости гипотезы однородности выборок.

Замечание 1. Кроме вывода об однородности или неоднородности групп, здесь полезно визуально сравнить распределения в группах. Для этого можно совместить гистограммы обеих выборок (пункт 10). Следует только помнить, что, поскольку объемы выборок могут быть различны, то гистограммы должны быть построены по относительным (деленным на объемы выборок) частотам.

Замечание 2. Построенный критерий не зависит от способа, каким были получены частоты. Этот критерий можно использовать и для проверки однородности двух выборок, когда частоты представляют собой количества выборочных данных, удовлетворяющих произвольным взаимоисключающим условиям. Например, в пособии “Курсовой проект ...” [2] рассматривается задача сравнения двух общин по группам крови ($r = 4$). Другой пример. В медицинской практике очень часто требуется сравнить новую методику лечения со старой методикой по результатам клинических наблюдений. При этом пациентов, прошедших курс лечения подразделяют, например, на 3 группы – а) не выздоровели, б) выздоровели, но через год болезнь повторилась, и в) выздоровели без последующего рецидива.

Больше двух выборок.

Критерий однородности может быть применен и в случае, когда число выборок s больше двух. Пусть v_{ij} – число исходов в j -ой выборке ($j = \overline{1, s}$), попавших в i -ый интервал группировки ($i = \overline{1, r}$), см. таблицу на стр. 50 [4].

☞ Порядок вычислений.

- 1) Составить вспомогательную таблицу, в ячейках которой вычисляются соответствующие слагаемые статистики X^2

$$\triangleright \kappa_{ij} = \frac{v_{ij}^2}{v_{i\bullet} \cdot v_{\bullet j}}$$

| Выборка \ Интервал | 1 | ... | s | Всего |
|--------------------|---------------|-----|---------------|-------------------------------|
| 1 | κ_{11} | ... | κ_{1s} | |
| \vdots | ... | | | |
| r | κ_{r1} | ... | κ_{rs} | |
| Всего | | | | $Q = \kappa_{\bullet\bullet}$ |

- 2) просуммировать все ячейки этой таблицы: $Q = \kappa_{\bullet\bullet}$;
- 3) вычислить значение статистики хи-квадрат: $\chi^2 = n^* (Q - 1)$;
- 4) найти критический уровень значимости: $\alpha_{\text{крит}} \approx 1 - K_g(\chi^2)$,
при числе степеней свободы $\mathcal{G} = (r - 1)(s - 1)$;
- 5) принять или отвергнуть гипотезу однородности.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Что означает (в вероятностном смысле) однородность выборок?
Ответ: см. [4] стр. 37.
3. Чему равна статистика критерия однородности хи-квадрат? Почему эта статистика может трактоваться как мера близости к гипотезе?
Ответ: см. [4] стр. 48.
4. Чему равен критический уровень значимости критерия однородности?
Ответ: см. [4] стр. 49
5. Можно ли с помощью этого критерия проверить гипотезу о том, что 1-ая половина данных из задания 1 имеет такое же распределение, как и 2-ая половина данных? Как в этом случае следует строить таблицу частот?
Ответ: см. [4] стр. 48.
6. При клинических испытаниях из 80 пациентов, лечившихся по новой методике, 85% полностью выздоровели. Можно ли сказать, что новая методика лучше старой, если из 50 пациентов, лечившихся по старой методике, 35 пациентов полностью выздоровели?

Задание 11.

Построить интервальную оценку для среднего значения нормального распределения.

Постановка задачи.

Имеется выборка (x_1, \dots, x_n) из нормального распределения. Требуется построить 95%-доверительный интервал (верхнюю границу, нижнюю границу) для неизвестного среднего μ этого распределения.

Теоретические основы.

См. стр. 51-54 и стр. 54-55 пособия [4].

Вычисления.

При работе с пакетом Excel для нахождения квантилей t^α можно воспользоваться встроенной функцией “СТЮДРАСПОВР” (категория “Статистические”), которая вычисляет квантиль одностороннего распределения Стьюдента. Процесс построения доверительных границ состоит из следующих этапов.

1. По выборочным данным находятся среднее \bar{x} и дисперсия s^2 .
2. Вычисляется стандартная ошибка среднего $m = s / \sqrt{n-1}$.
3. Находится квантиль распределения Стьюдента t^α или $t^{\alpha/2}$.
4. Строятся доверительные границы $\bar{x} \pm m \cdot t^{\alpha/2}$.

Пример.

Ниже в таблице приведено построение доверительных границ по 22 значениям нормальной сл.в.. Таблица справа приведена для пояснений.

| | | | Пояснения | |
|------------------------|----------------|-----------------|------------------------------------|--|
| Характеристика | | | Формула | Функция Excel |
| Среднее | | 4,366 | \bar{x} | СРЗНАЧ (...) |
| Дисперсия | | 0,237 | s^2 | ДИСПР (...) |
| Станд.Откл. | | 0,487 | s | СТАНДОТКЛОНП (...) |
| Ошибка среднего | | 0,106 | $m = s / \sqrt{n-1}$ | |
| Число данных | | 22 | n | СЧЕТ (...) |
| Уровень α | | 0,05 | | |
| Надежность Q | | 95% | $(1-\alpha) \cdot 100\%$ | |
| Квантили | $t^{\alpha/2}$ | 2,093 | | СТЮДРАСПОБР (α ; $n-1$) |
| | t^{α} | 1,729 | | СТЮДРАСПОБР ($2 \cdot \alpha$; $n-1$) |
| Доверительный интервал | | (3,912 ; 4,254) | $\bar{x} \pm m \cdot t^{\alpha/2}$ | |
| Верхняя граница | | 4,225 | $\bar{x} + m \cdot t^{\alpha}$ | |
| Нижняя граница | | 3,942 | $\bar{x} - m \cdot t^{\alpha}$ | |

☞ Порядок вычислений не требует дополнительных пояснений (см. правую таблицу и описание процесса построения выше). Заметим только, что при вызове функций все ссылки, обозначенные многоточием, следует заменить ссылкой

Данные!B139:D146.

Замечание. Не надо строить все границы – ограничьтесь только той, которая требуется в задании.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Что такое доверительное множество?
Ответ: см. [4] стр. 51.
3. Дайте интерпретацию определения нижней (верхней) доверительной границы.
Ответ: см. [4] стр. 51.
4. Какую надежность будет иметь двухсторонний доверительный интервал, если он построен на основе 90%-ой нижней границы и 95%-ой верхней границы?
Ответ: см. [4] стр. 52.
5. Можно ли с помощью двухстороннего доверительного интервала проверить гипотезу о том, что истинное значение оцениваемого параметра будет больше 18?
Ответ: см. [4] стр. 52.
6. Можно ли утверждать, что чем выше надежность, тем выше качество доверительного множества?
Ответ: см. [4] стр. 52.
7. Приведите формулы доверительных границ (доверительного интервала) для среднего значения нормального распределения.
Ответ: см. [4] стр. 54-55.
8. Найдите по таблице распределения Стьюдента квантиль $t^{0,03}$ для $n = 25$.
Ответ: см. ниже стр. 77.
9. Что такое стандартная ошибка среднего?
Ответ: см. [4] стр. 55.
10. Можно ли, основываясь на записи вида $a \pm m$, построить доверительный интервал для среднего значения?

Задание 12.

Построить интервальную оценку для дисперсии нормального распределения.

Постановка задачи.

Имеется выборка (x_1, \dots, x_n) из нормального распределения. Требуется построить 95%-доверительный интервал (верхнюю границу, нижнюю границу) для неизвестной дисперсии σ^2 этого распределения.

Теоретические основы.

См. стр. 51-54 и стр. 55-56 пособия [4].

Вычисления.

В пакете Excel реализована функция надежности распределения хи-квадрат $\text{ХИ2РАСП}(x; m) = 1 - K_m(x)$ и обратная к ней функция $\text{ХИ2ОБР}(p; m)$. Квантиль хи-квадрат распределения $t^{1-p}(m)$ можно вычислить как

$$t^{1-p}(m) = \text{ХИ2ОБР}(1-p; m).$$

Пример.

Воспользуемся данными из предыдущего задания. Схема построения доверительных границ для дисперсии состоит из следующих шагов.

1. По выборочным данным находится дисперсия s^2 .
2. По таблицам или воспользовавшись функцией ХИ2ОБР , находят квантили хи-квадрат распределения $t^{1-p}(m)$ и $t^p(m)$ для $p = 0,05$ и $p = 0,025$.
3. Строятся доверительные границы $\frac{n s^2}{t^p(m)}$ и $\frac{n s^2}{t^{1-p}(m)}$.

| | | | |
|------------------------|-----------------------|-----------------------------|---|
| Характеристика | | Формула | Функция Excel |
| Дисперсия | 0,237 | s^2 | ДИСПР (...) |
| Число данных | 22 | n | СЧЕТ (...) |
| Уровень α | 0,05 | | |
| Надежность Q | 95% | | |
| Квантили | | $p = 0,05$ или $p = 0,025$ | |
| $t^{0,975}(21)=10,283$ | $t^{0,95}(21)=11,591$ | | ХИ2ОБР (p ; $n-1$) |
| $t^{0,025}(21)=35,479$ | $t^{0,05}(21)=32,671$ | | ХИ2ОБР ($1-p$; $n-1$) |
| Верхняя граница | 0,450 | $\frac{ns^2}{t^{0,95}(21)}$ | $\frac{22 \cdot 0,237}{11,591}$ |
| Нижняя граница | 0,160 | $\frac{ns^2}{t^{0,05}(21)}$ | $\frac{22 \cdot 0,237}{32,671}$ |
| Доверительный интервал | (0,147; 0,507) | | $\left(\frac{22 \cdot 0,237}{35,479}, \frac{22 \cdot 0,237}{10,283} \right)$ |

☞ Порядок построения вполне очевиден из приведенной пояснительной таблицы.

Замечание 1. Не надо строить все границы – ограничьтесь только той, которая требуется в задании.

Замечание 2. Мы воспользовались теми же данными, по которым было построено доверительное утверждение для среднего значения. Однако так поступать нельзя, поскольку мы тем самым утверждаем, что с надежностью $(1 - \alpha) \cdot 100\%$ выполняется составное утверждение

$$\underline{\mu} \leq \mu \leq \bar{\mu} \quad \text{и} \quad \underline{\sigma}^2 \leq \sigma^2 \leq \bar{\sigma}^2,$$

но это не верно.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Приведите формулы доверительных границ (доверительного интервала) для дисперсии нормального распределения.
Ответ: см. [4] стр. 56.
3. Найдите по таблице 5%-ю и 90%-ю верхние квантили хи-квадрат распределения для объема выборки $n = 35$.
Ответ: см. ниже стр. 75.
4. Проверьте гипотезу $H_0 : \sigma^2 = 0,55$ о значении истинной дисперсии на уровне 5% при альтернативе $H_1 : \sigma^2 < 0,55$, воспользовавшись результатами примера.

Ответ: см. [4] стр. 52.

Задание 13.

Построить интервальную оценку для вероятности успеха

Постановка задачи.

В эксперименте подсчитывалось число успешных реализаций некоторого события (например, число доброкачественных изделий). Требуется построить доверительную границу для вероятности p этого события.

Теоретические основы.

См. стр. 51-54 и стр. 56-58 пособия [4].

Вычисления.

В пакете Excel имеется возможность вычисления функции

$$\text{БИНОМРАСП}(t; n; p; b),$$

которая при четвертом параметре $b=1$ вычисляет сумму всех биномиальных вероятностей до t включительно. Для построения доверительных пределов можно воспользоваться методом поиска решений (подраздел «Подбор параметра» раздела «Сервис» главного меню Excel), решая уравнения

$$\text{БИНОМРАСП}(t-1; n; p; 1) = 1 - \alpha \quad (\text{или} \quad = 1 - \alpha/2) -$$

для нижней границы и

$$1 - \text{БИНОМРАСП}(t; n; p; 1) = 1 - \alpha \quad (\text{или} \quad = 1 - \alpha/2) -$$

для верхней границы, относительно параметра p . Схема применения метода приведена ниже в примере.

Для подсчета числа выборочных данных, удовлетворяющих заданному условию, можно воспользоваться функцией

$$\text{СЧЁТЕСЛИ}(\text{Данные}, \text{Условие}),$$

которая выдает количество чисел в массиве Данные, удовлетворяющих заданному Условию.

Пример.

Сначала рассмотрим пример построения приближенных доверительных границ для вероятности выпуска доброкачественной продукции. Данные представляют собой измерения прочности дисков турбин авиадвигателей (те же данные, что были использованы нами при первичной статистической обработке). Норма прочности должна задаваться конструктором. Так как нам эта норма не известна, то мы (в иллюстративных целях) будем задавать её произвольно. Изделие считается кондиционным, если его прочность больше $C_{\text{норм}}$.

| | А | | В |
|----|---|-----------------------|-----------------------|
| 1 | Норма $C_{\text{норм}}$ | | 119,67 |
| 2 | Общее число данных n | | 101 |
| 3 | Число кондиционных t | | 87 |
| 4 | Оценка вероятности выпуска хорошего изделия \tilde{p} | | 0,861 |
| 5 | Стандартная ошибка m | | 0,0344 |
| 6 | $\alpha =$ | | 0,05 |
| 7 | Квантили | $t^{\alpha} = 1,6448$ | $t^{\alpha/2} = 1,96$ |
| 8 | Доверительный интервал | | (0,794 ; 0,929) |
| 9 | Верхняя граница | | 0,918 |
| 10 | Нижняя граница | | 0,805 |

| Формула | Функция Excel |
|---|-------------------------|
| | |
| n | СЧЁТ (...) |
| t | СЧЁТЕСЛИ (... , >B1) |
| $\frac{t}{n}$ | В3/В2 |
| $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ | КОРЕНЬ (В4* (1-В4) /В2) |
| | |
| $\Phi^{-1}(1-\alpha)$ | НОРМСТОБР |
| $\tilde{p} \pm m \cdot t^{\alpha/2}$ | В4±В5*А8 |
| $\tilde{p} + m \cdot t^{\alpha}$ | В4+В5*А7 |
| $\tilde{p} - m \cdot t^{\alpha}$ | В4-В5*А7 |

☞ Порядок вычислений вполне очевиден из приведенной справа пояснительной таблицы. Многоточия “...” при обращении к функциям следует заменить ссылками на область данных.

Теперь перейдем к построению точных границ. Для этого воспользуемся первыми четырьмя строками предыдущих вычислений и поместим вычисления точных границ на том же самом листе в строках 12-14.

| | A | B | C |
|---|--|--------|---|
| 1 | Норма $C_{иди}$ | 119,67 | |
| 2 | Общее число данных n | 101 | |
| 3 | Число кондиционных t | 87 | |
| 4 | Оценка вероятности выпуска хорошего изделия \tilde{p} | 0,861 | |

| | | | |
|----|------------------------|-------|-------|
| 12 | Доверительный интервал | 0,778 | 0,922 |
| 13 | Верхняя граница | | 0,914 |
| 14 | Нижняя граница | 0,792 | |

Сначала все эти ячейки
заполняются значением
 \tilde{p} ($=0,861$)

☞ Порядок вычислений точных доверительных границ.

- 1) Ячейки B12, B14, C12 и C13 заполнить значениями (не ссылками) оценки \tilde{p} .

Найти решение уравнения для нижней границы доверительного интервала:

- 2) во вспомогательную ячейку (например, B15) поместить формулу биномиального распределения
➤ =БИНОМРАСП (B3-1 ; B2 ; B12 ; 1)
- 3) вызвать процедуру поиска решений
– «Сервис» – «Подбор параметра»;
- 4) ввести параметры процедуры в таблицу запроса:

Здесь в первой строке указывается ячейка, содержащая формулу вычисления функции распределения;
во второй строке – подбираемое значение (в данном случае величина надежности $1 - \alpha / 2$);
в третьей строке – ячейка с искомым значением параметра;

5) OK

– в результате в ячейке В12 будет выведено искомое значение нижней границы, а в ячейке В15 значение 0,975.

6) Для вычисления верхней границы (в ячейке С12) необходимо заменить формулу пункта 2) на

➤ $=1-\text{БИНОМРАСП}(В3; В2; С12; 1)$

и поместить в ту же вспомогательную ячейку В15, после чего, снова вызвать метод поиска решений, заменив в третьей строке запроса пункта 4) ссылку В12 на С12.

7) Односторонние границы в ячейках С13, В14 строятся по той же самой схеме с заменой значения 0,975 в запросе пункта 4) на 0,95.

Замечание. Как видно из результатов приведенного примера, при объеме выборки порядка 100 асимптотические методы дают ошибку во втором-третьем знаке после запятой.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Приведите формулы асимптотических границ для вероятности успеха.

Ответ: см. [4] стр. 57.

3. Приведите формулы уточненных асимптотических границ.

Ответ: см. [4] стр. 57.

4. Как можно построить точную границу для вероятности успеха?

Ответ: см. [4] стр. 53.

5. Построив предварительно соответствующую доверительную границу, проверьте гипотезу о том, что вероятность рождения девочки меньше 0,5, если в 50 случаях наблюдалось 28 рождений мальчиков.

Ответ: см. [4] стр. 52, 57.

Задание 14.

**Проверить независимость двух характеристик
по критерию сопряженности хи-квадрат**

Постановка задачи.

По выборке $(x_1, y_1), \dots, (x_n, y_n)$ из двумерного распределения (не обязательно нормального) проверить гипотезу независимости компонентов наблюдаемого случайного вектора (X, Y) .

Теоретические основы.

См. стр. 59-62 и стр. 63-65 пособия [4].

Вычисления.

При построении критерия сопряженности для данных непрерывного типа основная сложность состоит в получении таблицы сопряженности. Если объем выборки не слишком велик (< 200), то наиболее простой способ – воспользоваться помощью одного из друзей и, перебирая все пары данных от первой до последней, установить принадлежность их той или иной ячейке. Процесс можно организовать следующим образом. Один из студентов произносит вслух пару чисел (x, y) , а второй студент ставит точку в ту ячейку, куда попала эта пара. По окончании перебора всех чисел останется только подсчитать количество точек в каждой из ячеек.

Пример.

Рассмотрим сначала задачу исследования зависимости между характеристиками по данным, представленным в виде таблицы сопряженности. В качестве примера взяты данные медицинского обследования абитуриентов одного из вузов г. Казани на предмет зависимости артериального давления (первый признак) от уровня употребления табачных изделий (второй признак). Данные обследования занесены в ячейки B2 : E3 приведенной ниже таблицы.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---------------------|-------|---------------|---|-------------|-----|---|----------------------------------|------|------------------------|------|
| 1 | Курение Давление | Много | Уме- ренно | Мало | Не курят | Σ | | Таблица квадратов расхождений | | | |
| 2 | Высокое | 81 | 69 | 116 | 76 | 342 | | 2,15 | 0,25 | 0,41 | 0,97 |
| 3 | Норма | 80 | 83 | 172 | 123 | 458 | | 1,61 | 0,19 | 0,31 | 0,72 |
| 4 | Σ | 161 | 152 | 288 | 199 | 800 | | | | $X^2 =$ | 6,6 |
| 5 | | | | | | | | | | $\alpha_{\text{крит}}$ | 0,08 |
| 6 | | | Вывод | Прослеживается слабо зна- чимая тенденция к увеличе- нию давления | | | | | | | |

☞ Порядок вычислений.

- 1) Подсчитать общее количество случаев по каждой группе каж-
дого из признаков и общее количество случаев
 - выделить блок ячеек B2 : F4
– исходная таблица данных плюс пустая строка ниже и
пустой столбец справа;
 - нажать кнопку Σ на панели инструментов.

Составить таблицу квадратов расхождений:

- 2) в ячейку H2 ввести формулу
 - $= (\$F\$4 * B2 - B\$4 * \$F2) ^ 2 / (\$F\$4 * B\$4 * \$F2)$
 $\{ (n \cdot n_{ij} - n_{\bullet j} \cdot n_{i\bullet})^2 / (n \cdot n_{\bullet j} \cdot n_{i\bullet}) - \text{это просто пояснение} \}$
– следите за правильностью расположения знаков \$;
- 3) скопировать ячейку H2 во все ячейки блока H2 : K3;
- 4) вычислить статистику хи-квадрат в ячейке K4
 $= \text{СУММ} (H2 : K3)$
- 5) вычислить критический уровень значимости в ячейке K5
 $= \text{ХИ2РАСП} (K4 ; (2-1) * (4-1))$
– здесь цифра “2” – это число групп по признаку «Давле-
ние», а цифра “4” – число групп по признаку «Курение»;
- 6) сделать вывод о значимости или отсутствии значимости согла-
сия данных с гипотезой независимости признаков.

Замечание. Для представленных данных критический уровень значимости близок к своему пороговому значению в 5%. Поэтому сделан такой расплывчатый вывод о возможной зависимости между уровнем курения и величиной кровяного давления.

Пример.

Рассмотрим пример применения критерия сопряженности хи-квадрат к данным, которые в следующих двух заданиях будут использованы для проверки гипотезы независимости по критерию Стьюдента и для построения линий регрессии.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|--------------------------------------|--------|-------|--------|--------|-------------------|---|-------------------------------|-------|------------------------|-------|
| 1 | $\begin{matrix} X \\ Y \end{matrix}$ | <119,7 | 121,3 | 122,9 | >122,9 | Σ | | Таблица квадратов расхождений | | | |
| 2 | <52,55 | 1 | 5 | 10 | 5 | 21 | | 0,003 | 0,025 | 0,154 | 0,132 |
| 3 | 56,05 | 4 | 28 | 17 | 4 | 53 | | 0,022 | 0,315 | 0,176 | 0,034 |
| 4 | 59,55 | 7 | 12 | 4 | 0 | 23 | | 0,152 | 0,133 | 0,022 | 0,000 |
| 5 | >59,55 | 2 | 2 | 0 | 0 | 4 | | 0,071 | 0,021 | 0,000 | 0,000 |
| 6 | Σ | 14 | 47 | 31 | 9 | 101 | | | | $X^2 =$ | 26,35 |
| 7 | | | | | | | | | | $\alpha_{\text{крит}}$ | 0,002 |
| 8 | | | | Вывод: | | Признаки зависимы | | | | | |

Замечание. Кроме вывода о значимом отклонении гипотезы независимости по этой таблице также можно сказать, что при увеличении признака X следует ожидать уменьшения признака Y – данные как бы концентрируются возле второй главной диагонали.

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Что такое независимость случайных величин?
Ответ: см. [4] стр. 62.
3. Выпишите формулу для вычисления статистики критерия сопряженности хи-квадрат.
Ответ: см. [4] стр. 64.
4. Почему эта статистика может служить мерой близости данных к гипотезе независимости?
Ответ: см. [4] стр. 64.
5. Чему равен критический уровень значимости критерия сопряженности признаков?
Ответ: см. [4] стр. 64.
6. Каким еще критерием (и в каком случае) можно проверить гипотезу независимости двух наблюдаемых характеристик?
Ответ: см. [4] стр. 66.

Задания 15-16.

**Проверить независимость двух характеристик
по критерию Стьюдента.**

Построить линии регрессии.

Постановка задачи.

По выборке $(x_1, y_1), \dots, (x_n, y_n)$ из двумерного нормального распределения проверить гипотезу независимости компонентов наблюдаемого случайного вектора (X, Y) . Построить линии регрессии одного из признаков по другому признаку. Найти наилучший прогноз признака Y при фиксированном значении признака $X = 120$.

Теоретические основы.

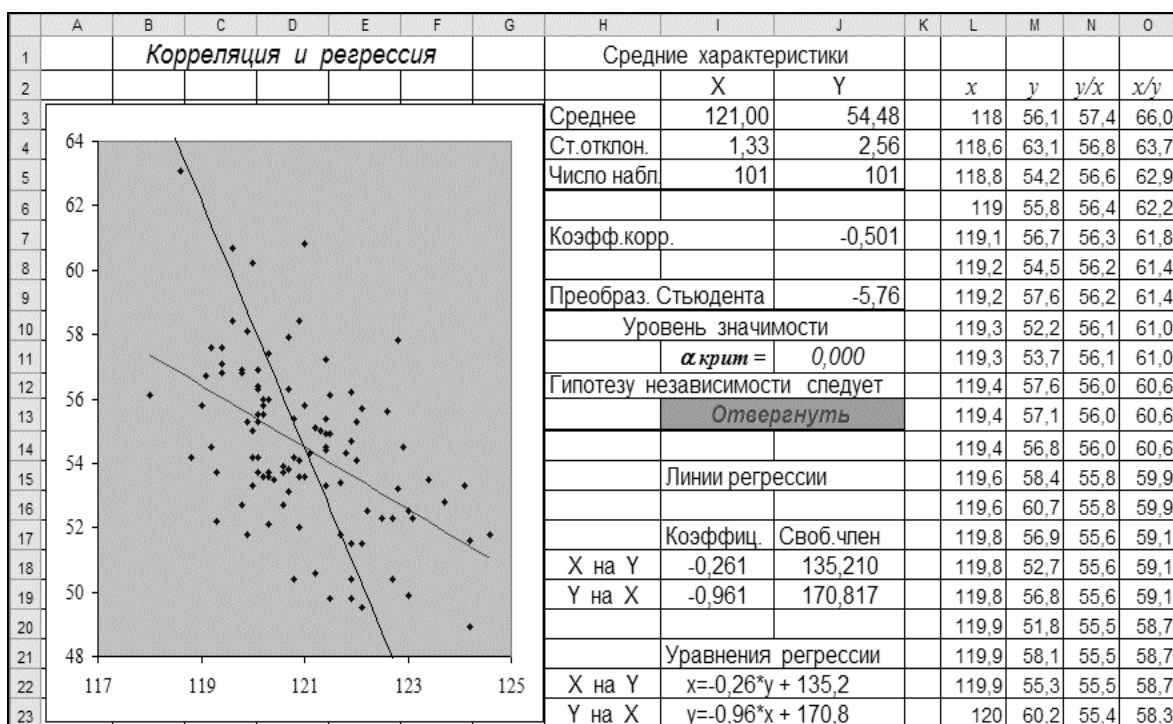
См. стр. 59-62 и стр. 65-67 пособия [4].

Вычисления.

Для вычисления выборочного коэффициента корреляции пакет Excel располагает встроенной функцией КОРРЕЛ(массив1; массив2), где массив1 и массив2 – ссылки на ячейки с наблюдениями над x и y . Количество x -ов должно совпадать с количеством y -ов.

Пример.

Ниже приведен образ листа Excel с необходимыми вычислениями. Данные по первому признаку те же, что использовались при первичной статистической обработке. Кроме того, здесь приведены результаты построения линий среднеквадратической регрессии. Гипотеза независимости проверяется при двухсторонней альтернативе.



При $x = 120$ наилучший прогноз второй характеристики $y = 55,4$.

☞ Порядок вычислений.

В целях удобства, лучше всего сначала скопировать данные на рабочий лист. Итак, пусть наши данные располагаются в ячейках L3:M103 (101 значение по каждому из признаков).


- 1) Подсчитать средние значения, стандартные отклонения и число наблюдений для обоих признаков (ячейки I3:J5);
- 2) вычислить коэффициент корреляции (ячейка J7)
 - =КОРРЕЛ(L3:L103; M3:M103)
- 3) вычислить преобразование Стьюдента для r (ячейка J9)
 - =КОРЕНЬ(I5-2) * J7 / КОРЕНЬ(1-J7^2)
 - { $\sqrt{n-2} * r / \sqrt{1-r^2}$ – это просто пояснение };
- 4) вычислить критический уровень значимости для двухсторонней альтернативы (ячейка J11)
 - =СТЮДРАСП(ABS(J9); I5-2; 2)
 - последний аргумент – число хвостов – для двухсторонней альтернативы равен 2;

- 5) сделать вывод о значимости гипотезы независимости
– ячейки I13:J13.

Перейти к построению линий регрессии.

- 6) Вычислить коэффициенты регрессии X на Y
 - $=J7*I4/J4$ – ячейка I18 – коэффициент регрессии;
 $=I3-I18*J3$ – ячейка J18 – свободный член;
- 7) вычислить коэффициенты регрессии Y на X
 - $=J7*J4/I4$ – ячейка I19 – коэффициент регрессии;
➤ $=J3-I19*I3$ – ячейка J19 – свободный член;
- 8) вычислить прогноз признака Y по значению признака $X=120$
 - $=120*I19+J19$.

Построить графики линий регрессии.

- 9) упорядочить данные по признаку X
 - выделить все ячейки данных (как x -ы, так и y -и), начиная со столбца L, и нажать кнопку  меню Excel (если начать выделение со столбца M, то упорядочение произойдет по признаку Y);
- 10) в ячейках N3, O3 вычислить значения функций регрессии (значения переменной y по значению переменной x)
 - $=L3*I\$19+J\19 – регрессия Y на X ;
 - $=(L3-J\$18)/I\18 – регрессия X на Y ;
- 11) скопировать ячейки N3, O3 параллельно данным столбца L.

Теперь все готово для построения графиков линий регрессии.

- 12) Выделить данные в четырех столбцах L, M, N, O;
- 13) вызвать “Мастера построения диаграмм”;
- 14) выбрать “Точечную диаграмму со значениями, соединенными сглаживающими линиями”;

- 15) после двух нажатий кнопки выбрать закладку \\Легенда// и удалить легенду из графика, сняв галочку на переключателе “Добавить легенду”;
- 16) ;
- 17) привести вид полученного графика в соответствие с приведенным выше стандартом
 - убрать маркеры с линий регрессии;
 - убрать линию, соединяющую исходные данные;
 - изменить границы шкал по оси ОХ и по оси ОУ
– общий принцип – щелкнуть правой кнопкой мыши в области изменяемого элемента и выбрать изменение «Формата».

Замечание 1. Отрицательное значение коэффициента корреляции говорит о том, что с ростом одного из признаков следует ожидать уменьшение другого признака – сравните с аналогичным выводом, сделанным при проверке независимости по критерию сопряженности хи-квадрат.

Замечание 2. Гипотеза независимости отвергается с очень высоким уровнем значимости (на листе приведено значение 0,000, означающее, что действительное значение меньше 0,001). Однако, величина коэффициента корреляции -0,501 имеет низкую практическую значимость – см. замечание 2 на стр. 66 пособия [4].

Контрольные вопросы.

1. Сформулируйте статистическую задачу.
2. Запишите формулу для выборочного коэффициента корреляции и найдите (вручную) его значение по следующим данным: (1;2), (2;5), (3;8).
 Ответ: см. [4] стр. 61.
3. Не вычисляя, скажите, чему равен коэффициент корреляции для следующих данных: (1;2), (2;4), (3;6), (7;14)?
 Ответ: см. [4] стр. 60.
4. Является ли выборочный коэффициент корреляции несмещенной (состоятельной) оценкой для истинного коэффициента корреляции?
 Ответ: см. [4] стр. 61.
5. Какими свойствами обладает коэффициент корреляции?
 Ответ: см. [4] стр. 60.
6. Как изменится значение коэффициента корреляции между ростом и весом человека, если значение веса измерять сначала в килограммах, а затем в граммах?
 Ответ: см. [4] стр. 60.
7. Почему (и когда) гипотезу независимости можно проверять, основываясь на значениях выборочного коэффициента корреляции?
 Ответ: см. [4] стр. 65.
8. Выпишите преобразование Стьюдента для выборочного коэффициента корреляции.
 Ответ: см. [4] стр. 65.
9. Чему равен критический уровень значимости при односторонней альтернативе?
 Ответ: см. [4] стр. 66.
10. В чем отличие статистической значимости от практической значимости?
 Ответ: см. [4] стр. 66-67.
11. Что такое линейная регрессия?
 Ответ: см. [4] стр. 59-60.

12. Выпишите уравнение линейной регрессии Y на X . Можно ли по этому уравнению вычислить приближенное значение признака X , если задано значение признака Y ?

Ответ: см. [4] стр. 60.

13. В целях упрощения вычислений очень часто значения одного (или обоих) признаков уменьшают на одну и ту же константу. Например, если все данные располагаются в пределах от 0 до 20, то, вычтя из всех данных число 10, мы сможем проводить вычисления даже без калькулятора. Как при этом изменятся коэффициенты линии регрессии?

Ответ: см. [4] стр. 60.

14. Исходя из персональных данных, найдите наилучший прогноз признака X , если значение признака $Y = 50$.

15. Известно, что высота h , с которой падает предмет, и время его падения t удовлетворяют соотношению $h = gt^2/2$, где g – ускорение свободного падения. Как с помощью методов регрессионного анализа оценить величину g по ряду связанных замеров h и t ?

16. В каком случае обе линии регрессии совпадут?

Ответ: см. [4] стр. 60.

17. Как будут располагаться линии регрессии, если коэффициент корреляции близок к 0?

Ответ: см. [4] стр. 60.

18. Как будут располагаться линии регрессии, если коэффициент корреляции близок к 1?

Ответ: см. [4] стр. 60.

Встроенные функции Excel.

Здесь мы опишем возможности Excel при вычислении статистических функций и дадим пояснения к способу их вызова.

♦ Нормальное распределение.

См. стр. 16-17 пособия [4].

Практически любой справочник по математической статистике содержит таблицы функции $\Phi(x)$ и её квантилей. Так как стандартное нормальное распределение симметрично, то эти таблицы составляют для значений $x \geq 0$ и $p < 1/2$. Приведем фрагмент таблицы из сборника [1].

Таблица 1.1. Функция нормального распределения $\Phi(x)$

| x | 0 | 1 | 2 | 3 | 4 | ... | 8 | 9 |
|------|-----------|------|------|------|------|-----|------|------|
| 2,05 | 0,97 9818 | 9867 | 9915 | 9964 | 0012 | ... | 0205 | 0253 |
| 06 | 0,98 0301 | 0349 | 0396 | 0444 | 0491 | ... | 0680 | 0727 |
| 07 | 0774 | 0821 | 0867 | 0914 | 0960 | ... | 1145 | 1191 |
| | | | | ... | | | | |

Слева в таблице представлено входное значение x с точностью до второго знака после запятой. Третий знак указан в самой верхней строке таблицы. С целью представления на одном листе по возможности большей информации, таблица разбита на блоки (выделенные чертой), в которых числа имеют несколько одинаковых первых цифр. Эти совпадающие части приведены только для одного значения (в столбце под верхней первой ячейкой с цифрой 0). Так, например,

$$\Phi(2,052) = 0,97\ 9915, \quad \Phi(2,058) = 0,98\ 0205, \quad \Phi(2,074) = 0,98\ 0960.$$

Для нахождения квантилей можно использовать таблицу исходной функции распределения, отыскивая значение вероятности внутри таблицы и находя соответствующее входное значение. Например, при $p = 0,02$ верхняя p -квантиль (то есть решение уравнения $\Phi(x) = 1 - p = 0,98$) будет

находиться где-то между 2,053 и 2,054, так как $\Phi(2,053) < 0,98 < \Phi(2,054)$. Простая линейная аппроксимация дает $t^{0,02} \approx 2,0537$.

В этом же сборнике [1] имеется таблица значений обратной функции нормального распределения, иными словами – таблица p -квантилей (не верхних).

Таблица 1.3. Функция, обратная функции нормального распределения

| p | 0 | 1 | 2 | 3 | 4 | ... | 8 | 9 |
|-------|----------|------|------|------|------|------|------|-----------|
| 977 | 1,9 9539 | 9723 | 9908 | ... | 0093 | 0279 | ... | 1030 1219 |
| 978 | 2,0 1409 | 1600 | 1792 | 1984 | 2177 | ... | 2957 | 3154 |
| 979 | 3352 | 3551 | 3750 | 3950 | 4151 | ... | 4964 | 5169 |
| 0,980 | 5375 | 5582 | 5790 | 6000 | 6208 | ... | 7056 | 7270 |
| | | | | ... | | | | |

Таким образом, $t^{0,02} = \Phi^{-1}(0,98) = 2,05375$, что весьма близко к полученному выше приближенному значению.

Пакет Excel располагает четырьмя функциями, связанными с нормальным распределением. Для вызова этих функций необходимо

- 1) вызвать «Мастера Функций»
 - нажать кнопку f_x панели инструментов или перейти в подраздел “Функция” раздела “Вставка” главного меню Excel;
- 2) в категории “Статистические” найти соответствующую функцию;
- 3) заполнить таблицу аргументов функции.

Альтернативный способ вызова состоит в непосредственном обращении к соответствующей функции из ячейки листа Excel. Общий вид такого обращения можно представить следующим образом:

=ИМЯФУНКЦИИ (аргумент1; аргумент2; ...)

Аргументами функции могут быть либо числа, либо ссылки на ячейки, их хранящие.

Рассмотрим каждую из этих функций по отдельности.

1. НОРМСТРАСП – функция стандартного нормального распределения $\Phi(x)$. Аргумент – значение x (любое число).

2. НОРМСТОБР – обратная функция стандартного нормального распределения $t^{1-p} = \Phi^{-1}(p)$. Аргумент – p (число от 0 до 1).
3. НОРМРАСП – функция распределения $F(x) = \Phi(\frac{x-\mu}{\sigma})$ или функция плотности $f(x) = \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ нормального (μ, σ^2) закона. Имеет 4 аргумента:

| | |
|------------------|--|
| х | не требует пояснений |
| Среднее | среднее значение μ |
| Стандартное_откл | корень из дисперсии $\sigma = \sqrt{\sigma^2}$ |
| Интегральная | 0 (или FALSE) – вычисляется плотность, 1 (или TRUE) – функция распределения |

4. НОРМОБР – функция, обратная функции нормального распределения $\mu + \sigma \cdot \Phi^{-1}(p)$. Имеет 3 аргумента, аналогичные первым трем аргументам предыдущей функции.

Приведем несколько примеров применения этих функций.

| Функция Excel | Значение в ячейке | Характеристика распределения |
|------------------------|-------------------|---|
| =НОРМСТРАСП(2,058) | 0,98020500 | $\Phi(2,058)$ – функция распределения |
| =НОРМСТОБР(0,05) | -1,64485348 | $t^{0,95}$ – 5%-квантиль |
| =НОРМРАСП(-1; 0; 1; 1) | 0,15865526 | $\Phi((2,058-0)/1)$ – функция распределения |
| =НОРМРАСП(-1; 0; 1; 0) | 0,24197073 | $\phi((2,058-0)/1)$ – функция плотности |
| =НОРМОБР(0,95; 0; 1) | 1,64485348 | $0+1 \cdot t^{0,05}$ – верхняя 5%-квантиль |

Задание. Объясните совпадение значений (с точностью до знака) во второй и пятой строках этой таблицы.

♦ Хи-квадрат распределение.

См. стр. 18-19 пособия [4].

Сборник таблиц [1] содержит значения так называемого интеграла вероятностей хи-квадрат – в нашей терминологии это просто функция

надежности $\bar{K}_m(x) = 1 - K_m(x)$. Таблица имеет два входа – по числу степеней свободы (верхняя строка) и по аргументу функции (левый столбец).

Таблица 2.1а. Интеграл вероятностей χ^2

| x | $m=16$ | | ... | $m=20$ | |
|------|---------|-----------|-----|---------|-----------|
| | P | $-\Delta$ | | P | $-\Delta$ |
| ... | ... | | | ... | |
| 15,0 | 0,52464 | 3627 | ... | 0,77641 | 2929 |
| 5 | 48837 | 3541 | ... | 74712 | 3050 |
| ... | ... | | | ... | |

Здесь, кроме значения функции распределения (столбец P), приведены также первые разности этой функции (столбец $-\Delta$), точнее, только 5 значащих цифр после запятой без первых нулей. Таким образом, $\bar{K}_{16}(15,5) - \bar{K}_{16}(15,0) = -0,03627$ (после запятой поставлен один ноль, чтобы получилось пять цифр). Если x_0 и x_1 – два рядом стоящие значения аргумента, то для нахождения значения функции в промежуточной точке $x \in [x_0; x_1]$ можно применить аппроксимацию $\bar{K}_m(x) \approx \bar{K}_m(x_0) - \Delta \cdot \frac{x-x_0}{x_1-x_0}$. В приведенном нами фрагменте $x_1 - x_0 = 0,5$. Поэтому $\bar{K}_{16}(15,2) \approx \bar{K}_{16}(15,0) - 0,03627 \cdot 2 \cdot 0,2 = 0,510132$.

Значения верхних p -квантилей $t^p = t^p(m) = \bar{K}_m^{-1}(p)$ распределения хи-квадрат содержатся в следующей таблице на стр.166 сборника [1].

Таблица 2.2а. Процентные точки распределения χ^2

| $m \backslash Q$ | ... | 97,5% | 95% | ... | 5% | 2,5% | ... |
|------------------|-----|-------|--------|-----|--------|--------|-----|
| ... | | | | ... | | | |
| 19 | ... | 8,907 | 10,117 | ... | 30,144 | 32,852 | ... |
| 20 | ... | 9,591 | 10,851 | ... | 31,410 | 34,170 | ... |
| ... | | | | ... | | | |

Вход в таблицу осуществляется по числу степеней свободы (m в левом столбце) и по вероятности, выраженной в процентах (Q в верхней строке). Таким образом, $t^{0,05}(19) = 30,144$, $t^{0,025}(20) = 34,170$.

Пакет Excel предоставляет возможность вычисления как значений функции надежности $\bar{K}_m(x) = 1 - K_m(x)$, так и значений p -квантилей хи-

квадрат распределения. Эти функции называются ХИ2РАСП и ХИ2ОБР. Рассмотрим несколько примеров применения этих функций.

| Функция Excel | Значение в ячейке | Характеристика распределения |
|---------------------|-------------------|---|
| =ХИ2РАСП(15, 2; 16) | 0,510041 | $\bar{K}_{16}(15,2)$ – функция надежности |
| =ХИ2РАСП(15; 20) | 0,776408 | $\bar{K}_{20}(15)$ – функция надежности |
| =ХИ2ОБР(0, 05; 19) | 30,14351 | $t^{0,05}(19)$ – верхняя 5%-квантиль |
| =ХИ2ОБР(0, 025; 20) | 34,16958 | $t^{0,025}(20)$ – верхняя 2,5%-квантиль |

♦ Распределение Стьюдента.

См. стр. 19-20 пособия [4].

Таблицы распределения Стьюдента также имеются в любом справочнике по математической статистике. Приведем здесь фрагмент соответствующей таблицы из сборника [1].

Таблица 3.1а. Функция распределения Стьюдента

| $t \backslash k$ | 11 | 12 | ... | 19 | 20 |
|------------------|--------|--------|-----|--------|--------|
| ... | | | ... | | |
| 2,0 | 0,9646 | 0,9657 | ... | 0,9700 | 0,9704 |
| 1 | 9702 | 9712 | ... | 9753 | 9757 |
| ... | | | ... | | |

Эта таблица имеет два входа – число степеней свободы k (верхняя строка) и аргумент функции t (левый столбец). Из этой таблицы находим, что $S_{12}(2) = 0,9657$, $S_{19}(2,1) = 0,9753$. При степенях свободы больше 20 можно воспользоваться нормальным приближением: $S_k(t) \approx \Phi(t), k > 20$.

Следующая таблица указанного сборника [1] содержит значения верхних p -квантилей $t^p = t^p(k) = \bar{S}_k^{-1}(p)$. Эта таблица также имеет два входа – число степеней свободы (левый столбец) и вероятность в процентах $Q = p \cdot 100\%$ (верхняя строка). Для наглядности целые части вместе с запятой приведены только для верхних чисел в блоке из пяти чисел.

Таблица 3.2. Процентные точки распределения Стьюдента

| $\begin{matrix} Q \\ k \end{matrix}$ | ... | 10% | 5% | 2,5% | ... | 0,05% |
|--------------------------------------|-----|--------|--------|--------|-----|--------|
| ... | | | | | | |
| 19 | ... | 1,3277 | 1,7291 | 2,0930 | ... | 3,8834 |
| 20 | ... | 3253 | 7247 | 0860 | ... | 8495 |
| ... | | | | | | |

Таким образом, $t^{0,1}(16) = 1,3368$, $t^{0,025}(17) = 2,1098$.

В пакете Excel имеются встроенные функции

СТЮДРАСП, вычисляющая функцию надежности, и

СТЮДРАСПОБР, вычисляющая верхние квантили.

Функция СТЮДРАСП имеет три аргумента. Кроме двух естественных (аргумента t и числа степеней свободы k), при обращении к этой функции требуется указать количество хвостов распределения, которые нужно учитывать (1 или 2). Под “хвостом” распределения понимается любой интервал с одним конечным и одним бесконечным концом. Например, при вычислении функции надежности ищется вероятность попадания в область $[x; \infty)$. Поэтому $\bar{S}_k(t)$ - это функция СТЮДРАСП с одним “хвостом”. Очень часто в статистической практике требуется найти вероятность попадания в область $(-\infty; -x] \cup [x; \infty)$ при $x \geq 0$, то есть в область с двумя “хвостами”. Легко видеть, что функция, вычисляющая вероятности таких симметричных интервалов, представляет собой не что иное, как функцию надежности распределения модуля $|T_k|$.

Обращение к функции СТЮДРАСПОБР вполне тривиально и полностью аналогично обращению к функции ХИ2ОБР (см. выше).

♦ Показательное (экспоненциальное) распределение.

См. стр. 21 пособия [4].

Как функция плотности, так и функция распределения показательного закона могут быть вычислены посредством калькулятора. В Excel обе эти функции можно вычислить, воспользовавшись функцией EXP (...).

♦ **Биномиальное распределение.**

См. стр. 22 пособия [4].

В пакете Excel имеется возможность вычисления функции

БИНОМРАСП ($x; n; p; b$),

которая при $b=1$ или $b=TRUE$ есть не что иное, как функция распределения $Bin(x+1, n, p)$ (обратите внимание на различие в первом аргументе этих функций). Четвертый параметр функции БИНОМРАСП, если он не равен нулю, указывает на необходимость вычисления именно функции распределения, то есть суммы всех биномиальных вероятностей до x включительно (в отличие от функции $Bin(x|n, p)$, которая вычисляет вероятности до $x-1$). При $b=0$ или $b=FALSE$ функция БИНОМРАСП вычисляет индивидуальную вероятность $P\{X = x\}$. Приведем несколько примеров.

| Функция Excel | n | p | Вероятность | Результат |
|----------------------------|-----|-----|------------------|-----------|
| БИНОМРАСП (3;10;0,5;1) | 10 | 0,5 | $P\{X \leq 3\}$ | 0,171865 |
| БИНОМРАСП (3;10;0,5;0) | 10 | 0,5 | $P\{X = 3\}$ | 0,117188 |
| 1-БИНОМРАСП (29;100;0,5;1) | 100 | 0,5 | $P\{X \geq 30\}$ | 0,999984 |
| 1-БИНОМРАСП (30;100;0,5;0) | 100 | 0,5 | $P\{X \neq 30\}$ | 0,999977 |