3 Методы статистической обработки данных

3.1 Анализ таблиц сопряженности.

Для исследования взаимосвязи пары качественных признаков между собой применяется анализ таблиц сопряженности. Таблица сопряженности – это матричное представление частоты встречаемости объектов с той или иной комбинацией градаций (уровней значений) двух или более качественных признаков.

Важно исследовать все пары качественных признаков. Для этого строятся отдельно таблицы сопряженности содержащие ожидаемые и наблюдаемые частоты. Наблюдаемая частота n_{ij} события A_iB_j показывает количество объектов выборки обладающих комбинацией уровней A_i и B_j признаков А и В. Величина n'_{ij} называется ожидаемой частотой (ожидаемой при выполнении гипотезы H_0 об отсутствии взаимосвязи между признаками А и В).

$$n'_{ij} = P_i * P_j * n_{..} = \frac{n_{i.} * n_{.j}}{n_{..}}$$

где:

P_i – вероятность попадания объекта в i-строку,
P_j – вероятность попадания объекта в j-столбец,

$n_{i.}$ – сумма частот в i-строке по j-столбцу,

$n_{j.}$ – сумма частот по i-строке в j-столбце,

$n_{..}$ – общее число наблюдений.

Необходимо выявить различия между наблюдаемыми частотами и частотами, рассчитанными на основании гипотезы о независимости признаков. Мера расхождения между фактической частотой и ожидаемой по всем клеткам таблицы сопряженности может нести информацию относительно истинности H_0-гипотезы (об отсутствии взаимосвязи между двумя признаками).

Для проверки статистической гипотезы в анализе используется статистика Пирсона X^2:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n'_{ij})^2}{n'_{ij}};$$

Число степеней свободы *df = (r-1)(c-1).*

Если достигнутый уровень значимости «p» для вычисляемого значения статистики Пирсона более критического (5%), то нулевая гипотеза о независимости двух признаков принимается.

3.2 Однофакторный дисперсионный анализ (непараметрический)

Дисперсионным анализом называют статистический метод анализа результатов, зависящих от действия качественных факторов. Главной задачей дисперсионного анализа является расщепление общей вариации результирующего показателя на части, соответствие раздельному и совместному влиянию различных качественных факторов и остаточную вариацию, аккумулирующую влияние всех неучтенных факторов.

Для классического дисперсионного анализа существует условие наличия нормальности распределения количественного признака, сравниваемого в подгруппах уровней факторов.

В случае если отсутствует нормальность распределения изучаемых количественных признаков, применяется непараметрический дисперсионный анализ Краскела-Валлиса. Для этого заменим наблюдения X_{ij} их рангами r_{ij}. Ранг – это номер наблюдения в упорядоченном по возрастанию вариационном ряду. Необходимо упорядочить всю совокупность N наблюдений в порядке возрастания. Затем для каждого столбца таблицы рангов необходимо вычислить:

$$R_j = \sum_{j=1}^{n_j} r_{ij} \quad \text{и} \quad R_{.j} = \frac{1}{n_j}\sum_{j=1}^{n_j} r_{ij},$$

R_j – сумма рангов j-го столбца;

$R_{.j}$ – средний ранг рассчитанный по столбцу j.

Выясним различия между столбцами. Составляя общую характеристику необходимо учесть различия в числе наблюдений для разных столбцов и взять в качестве меры отступления от чистой случайности величину

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_i (R_{.k} - \frac{N+1}{2})^2 .$$

Эта величина называется статистикой Краскела-Валлиса. Множитель *12[N(N-1)]* Присутствует в этом выражении в качестве нормировочного для обеспечения асимптотической сходимости распределения H к распределению хи-квадрат с числом степеней свободы (к-1). Другая формула для вычисления H:

$$H = \frac{12}{N(N-1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1).$$

3.3 Корреляционный анализ

Корреляция в математической статистике означает зависимость, соотношение, связь между явлениями или признаками одного (нескольких) процессов.

Корреляционный анализ позволяет установить связь между различными признаками, измерить силу этой связи. Оценка значимости результатов анализа определяет достоверность выборочных показателей корреляции.

Рассмотрим основные виды корреляции.

1. По характеру проявления связи:

а) положительная корреляция: увеличение или уменьшение одной переменной приводит соответственно к увеличению или уменьшению другой переменной;

б) отрицательная корреляция: увеличение одной переменной приводит к уменьшению другой переменной, и наоборот.

2. По форме связи:

а) линейная корреляция: между взаимосвязанными переменными существуют линейные соотношения;

б) нелинейная корреляция: связь между переменными выражается нелинейными (относительно самих переменных) соотношениями.

3. По числу взаимосвязанных переменных:

а) Парная (простая) корреляция: имеется связь между двумя переменными;

б) Множественная корреляция: имеется связь между переменными, число которых более двух;

в) Частная корреляция: связь между двумя переменными при постоянных, зафиксированных значениях, других взаимосвязанных с этими двумя переменными.

В данной работе используется парная (простая) корреляция между различными признаками объекта. Таким образом, устанавливаются наиболее значимые корреляционные связи между количественными признаками в отдельных группах по градациям высоты произрастания и типу почв.

Основные задачи корреляционного анализа:

1. Количественное измерение силы, интенсивности связи двух и более явлений, переменных.

2. Отбор и ранжирование факторов по силе связи с исследуемым выходным параметром качества.

3. Обнаружение ранее не известных причинно-следственных связей между исследуемыми переменными.

Корреляционный анализ включает также построение графика имеющейся выборки наблюдений в виде двухмерной диаграммы. Совокупность всех точек этого графика в системе координат X,Y называется корреляционным полем. Для определения показателя силы, интенсивности линейной стохастической связи вычисляется коэффициент корреляции. Он был введен в практику статистического анализа Ф. Гальтоном и К. Пирсоном. Когда одна или обе переменные не подчиняются нормальному закону распределения, применяется ранговый коэффициент корреляции Спирмэна:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)},$$

где D_i – разность значений рангов, расположенных в двух рядах у одного и того же объекта.

Ранг наблюдения – это номер расположения данного наблюдения в упорядоченном ряду значений данной переменной. В курсовой работе был использован ранговый коэффициент корреляции Спирмэна (r_s), что связано с отсутствием нормальности распределения признаков. Значение этого коэффициента находится в пределах от –1 до +1. Если связь между признаками отсутствует, r_s=0. Проверка статистической гипотезы о равенстве генерального коэффициента нулю проводится с помощью t – статистики:

$$t_{расч} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}},$$

где tрасч – t критерий Стьюдента. Число степеней свободы его f=n-2. Если tрасч больше tтабл и p<0.05 , то нулевая гипотеза отвергается, т.е. существует корреляция между изучаемыми признаками.