Изследвания върху полските култури, 2009, Том V - 1
Field Crops Studies, 2009, Vol. V - 1

СТАТИСТИЧЕСКИ МЕТОДИ И СОФТУЕР В СЕЛСКОСТОПНАСКИТЕ ИЗСЛЕДВАНИЯ

Емил Пенчев
Добруджански земеделски институт - . Генерал Тошево

Резюме

Пенчев, Е., 2009. Статистически методи и софтуер в селскостопанските изследвания.

Статистическите методи са в основата на селскостопанските изследвания. Двете основни направления са : планиране и анализ на експеримента . Основна задача която трябва да реши всеки изследовател е правилното планиране на експеримента. Това действие е в основата на изследователската работа, защото предопределя точността на извеждане на опита и на получените експериментални данни, както и възможностите за пълноценния им и коректен статистически анализ. Много важен проблем при изследванията на биологичните обекти е оценяването влиянието на различни фактори. Върху развитието на растенията влияят много природни фактори, повечето, от които са неконтролируеми. Тяхното изследване е сложна за разрешаване задача, при която се прилагат различни модели на дисперсионен анализ, който се базира върху законите на нормалното разпределение и неговите производни. Важно значение има и принципния компонентен анализ, при който въз основа разлагането на дисперсията, могат да се определят структурите на факторния експеримент. Важна задача, която твърде често интересува изследователите на биологични обекти е изясняването на взаимовръзките между различните му компоненти. Основните методи които се прилагат в това отношение са корелационния анализ на фенотипно и генотипно ниво за оценка на директните ефекти . На базата на този анализ е възможно да бъдат оценени и индиректните ефекти със средствата на path- анализа и получените path – коефициенти. Изясняването конкретния вид на изследваните взаимовръзки се осъществява с методите на регресиония анализ, който дава възможност за прогнози и симулиране на биологичния обект. Голямо приложение в селскостопанските изследвания имат и методите на генетическата статистика, които изясняват сложните закони на наследяване на признаците и по този начин спомагат за разрешаването на селекционни задачи от най различен характер. Изчислителната техника и софтуера станаха един от най важните инструменти в ръцете на изследователите и интензифицираха в голяма степен изследователската дейност.

Ключови думи: Статистически методи - Принципен компонентен анализ – Корелационен анализ – Path-анализ.

Abstract

Penchev, E., 2009. Statistical methods and software in agricultural investigations

The statistical methods are at the basis of agricultural investigations. The two main directions of research are planning and analysis of the trial. The main task each researcher has to solve is the correct planning of the trial. This action lays the foundations of research

work because it predetermines the precision of the trail performance and of the experimental data obtained, as well as the potential for their thorough and correct statistical analysis. An essential problem in the investigations of biological entities is the assessment of the effect of various factors. Numerous natural factors affect plant development, most of them uncontrollable. Their investigation is a complex task which involves different models of dispersion analysis based on the laws of normal distribution and its derivatives. Principal componential analysis is also important for it allows determining the structures of the factor experiment on the basis of dispersion. An important task often intriguing researchers of biological entities is the clarification of the correlations between their individual components. The major methods used in this respect are correlation analysis at genotype and phenotype level to evaluate direct effects. On the basis of this analysis it is possible to assess the indirect effects as well through path analysis and the calculated path-coefficients. The specific correlations are clarified using the methods of regression analysis which allows making prognoses and simulations of the biological entity. The methods of genetic statistics are also widely used in agricultural researches; they clarify the complex laws of character heritability thus contributing to the solving of various breeding tasks. Computers and software have become a major tool in the hands of researchers and to a large extend intensified their research work.

Key words: Statistical methods – Principal componential analysis – Correlation analysis – Path-analysis

УВОД

Развитието на изчислителната техника и по специално създаването на персоналните компютри през 80 те години на миналия век създадоха чудесни условия за интензифициране на изследователската работа. Създаден бе софтуер във всички направления на експерименталната и аналитичната дейност който улесни и направи значително по прецизен труда на научните работници .

В областта на селскостопанската наука бе създаден и внедрен софтуер касаещ планирането на експеримента , лабораторните анализи както и различни методи за пълен статистически анализ . Като резултат са постигнати много успехи и значително по високо ниво на селскостопанските изследвания .

Основни статистически параметри

Първите анализи които са правени на експерименталните данни са били свързани с нивото на изчислителната техника . С калкулатори е възможно да бъдат изчислени основните параметри на разпределението на данните а именно средна стойност която характеризира центъра ; доверителния интервал в който тя варира и разсейването на експерименталните данни около този център с помощта на вариационен коефициент . Тези параметри дават възможност да бъде определено теоретичното разпределени на данните и доколко се различават изследваните извадки . Въпреки съществената информация която носят в себе си тези параметри не дават възможност за задълбочени статистически анализи .

Факторен експеримент

Основна задача която трябва да се разрешава при селскостопанските изследвания е влиянието на различните фактори върху полските култури . Най – често прилагани в това отношения са дизайни с балансирани блокове където с помощта на подходяща рендомизация могат да бъдат съчетани няколко фактора . Техния брой обаче е лимитиран поради големия обем който заемат подобни опити . Рендомизацията е задължителна за да бъде избегнато различното почвено плодородие на опитните парцелки и да бъде минимизирана опитната грешка . Много често прилагани са латинските квадрати и правоъгълници . Изследване влиянието

на факторите е възможно с различни модели на дисперсионен анализ (2) като основните изисквания са следните :

1. нормално разпределение на експерименталните данни
2. опитна грешка под 5 %

Разлагането на дисперсията по фактори трябва да се извършва по подходящ модел и дава възможност за оценка на степента на тяхното влияние, взаимодействието им както и интервалите на доверие , въз основа на които се определят различията между отделните нива на факторите . Такава задача често се налага да бъде разрешавана при селскостопанските изследвания за бъдат избрани най ефективните нива на факторите или тяхна комбинация .

Проучване на факторния експеримент в детайли позволява принципния компонентен анализ . Принципния компонентен анализ е метод базиращ се на множествената ковариация на изучаваните статистически променливи и даващ възможност за оценка на ролята им в изследваните взаимовръзки . Метода има основна роля в изследването на факторния експеримент както дисперсионния , корелационния и регресионния анализи . PCA е аналитична процедура за трансфирмиране на множество от променливи в друго множество от компонентни променливи имащи следните свойства :

1. те са линейна функция на оригиналните променливи
2. те са ортогонални т.е. независими една за друга
3. тоталната вариация сред тях е равна на тоталното вариране в оригиналните променливи , следователно информацията съдържаща се в наблюдаваните променливи не е загубена при трансформацията
4. първата компонента изчислява най-голямата възможна пропорция на тоталната вариация а втората най-голямата пропорция на остатъка

Геометрически имаме изобразени данни от **n** променливи изобразени в **p** мерното пространство . PCA представлява ротация на координатните оси по такъв начин , че тоталната дисперсия на проекциите на точките по първата ос е максимална , като това е първата принципна компонента . Втората ос (втората принципна компонента) е избрана ортогонално на първата и се изчислява като възможната остатъчна дисперсия . Приложението на този метод дава количествена оценка на влиянието на факторите и техните взаимодействия .

Блоковите експерименти са обемисти и скъпоструващи и поради тази причина все повече се търсят такива дизайни които да премахват тези недостатъци . Ето защо широко приложение намират в експерименталната работа също така не балансираните решеткови методи при които върху малко площи могат да бъдат изследвани много фактори . Друг подходящ подход също така е метода на дробните парцелки . Подходите при планирането на многофакторни експерименти стават все по многообразни и ефикасни но съществен проблем при тях е модела на дисперсионен анализ който е адекватен . Ако се подходи по стандартни процедури за разлагане на дисперсията то резултатите от анализа ще доведат до неверни изводи. Дисперсионния анализ намира приложение и при изследване на важното за селскостопанските изследвания взаимодействие генотип х климатична среда . Твърде важно за практиката е да бъде оценена екологическата пластичност и стабилност на сортовете . Основният параметър оценящ стабилността на сорта е дисперсията σ_i^2 . Колкото повече дисперсията на стабилността σ_i^2 клони към нула , толкова по-малко се отличават емпиричните стойности на признака от теоретически разположените на линията на регресията .Статистическите параметри YS , изчислени въз основа на тези дисперсии са комплексна оценка , като интегрират в себе си добив и стабилност. Те представляват удобен комплексен селекционен критерий .

Изследване на взаимовръзки и зависимости

Основен подход при изследването на зависимостите между различни ценни

стопански признаци и корелационния анализ . Линейния корелационен коефициент се изчислява по формулата :

$$R(x,y) = \text{cov}\,(x,y)/(D_x D_y)^{1/2}$$

Корелационния коефициент се изменя в интервала [-1,1] . Отрицателните стойности при доказаност на корелациите водят до извода за обратно пропорционални зависимости , а положителните стойности за право пропорционални зависимости . Корелационният анализ позволява да се изследват възможните взаимовръзки между показателите , които не е задължително да бъдат причинно-следствени (каузални) , тъй като корелационния коефициент представлява само количествена характеристика на степента на зависимост между две случайни величини . Твърде често при корелационен коефициент, клонящ към нула , се прави погрешен извод за независимост на изследваните случайни величини . В биологията взаимодействието генотип х среда е значително в повечето случаи и това води до проблема за репрезентативността на изследваните извадки от случайни величини към определени таксономични групи данни . Ето защо при определени екологични условия корелационния коефициент може да доказва изследвана взаимовръзка , а при други не . Корелационният анализ дава възможност за формиране на конкретни хипотези , които са валидни само при определени екологични условия .

В този смисъл path - анализа дава възможност да се установи степента на независимост между случайните величини X и Y . Path - коефициентите (2) се получават като решение на системата :

$$\sum P_i R_{ij} = R_{i,k} \quad , \quad i = 1 , k\text{-}1$$

където Rij е корелационният коефициент между случайните величини X_i и X_j , i;j= 1,k . Ясно е , че path-коефициентите носят недостатъците на корелационния анализ . Те са количествена оценка на индиректните взаимовръзки между изследваните показатели и могат само да доуточнят направените хипотези ..

Един от най - популярните методи за апроксимиране на изследваните зависимости в селското стопанство е регресионния анализ по метода на най-малките квадрати . Прилагането на този метод е при условие , че случайната грешка удовлетворява закона за нормално разпределение . Ако функциите f(x) и g(x) са зададени в интервала (a,b) в краен брой точки X_k , k=1,m , то се въвежда разстояние между тях :

$$\mathbf{r}(f,g) = \sum \boldsymbol{\lambda}_i \, (f(x_i) - g(x_i))^2 \quad .$$

Числата $\boldsymbol{\lambda}_i$ се наричат тегла . тегловните множители се поставят за да уеднаквят грешките в различните точки x_i . Методът на най-малките квадрати позволява да се реши следната задача : да се намери обобщен полином $\sum \mathbf{a_i R_i(x)}$, който най-малко да се отклонява от f(x) в смисъла на разстоянието **r**(f,g). Известно е , че задачата има единствено решение ако функциите $R_i(x)$ са линейно независими в множеството от точки x_k , k=1,m . Този метод позволява предимно полиномни приближения на реалните връзки . При изследване на климатичните фактори се прилага апроксимиране с тригонометрични полиноми от вида :

$$\boldsymbol{\tau}_m\,(x) = a_o/2 + \sum (a_k \text{cos}kx + b_k \text{sin}kx)$$

Понеже биологичните обекти се изменят във времето интерес представлява техните изменения във времето и да се апроксимира тяхната динамика на развитие .Динамичните модели се описват с помощта на диференциалните уравнения . Нека системата се определя от q променливи $X_1, X_2, \ldots , X_q$ в момента t . По принцип динамичния детерминиран модел се състои от диференциални уравнения от първи ред , които описват поведението на променливите във времето :

$$dX_1/dt = f_1(X_1,X_2, \ldots , X_q ; P; E)$$
$$dX_2/dt = f_2(X_1,X_2, \ldots , X_q ; P; E)$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$
$$dXq/dt = f_q(X_1,X_2, \ldots , X_q ; P; E) ,$$

където $f_1 , f_2 , \ldots , f_я$ са функциите на $X_1,X_2, \ldots , X_q$, параметрите P и околната среда E . Записа $f_i(X_1,X_2, \ldots , X_q ; P; E)$ не означава че функцията трябва да съдържа всички променливи и параметри .

При изследването на сортовете пшеница относно тяхната екологическа стабилност и пластичност е приложена методиката на Eberhart , Russel (1966) . Под екологическа пластичност се разбира средната реакция на сорта при изменение условията на средата , а под стабилност - отклонението на емпиричните данни при всяко условие на средата от тази средна реакция . На базата на тези тълкувания е разработена методиката . Статистическият модел на дисперсионния анализ с оценка варирането на екологическите условия е :

$$X_{ijk} = X \ldots + g_i + l_j + (gl)_{ij} + b_k + e_{ijk}$$

където X... е средната стойност на опита , g_i е ефекта на i-тия генотип , l_j е ефекта на j - тото условие , b_k е ефекта на к - тото повторение на условията ,а e_{ijk} е ефекта на случайната грешка . Анализът и оценката се извършват само при доказано взаимодействие на факторите “генотип х околна среда “ . Оценяват се отклоненията на всички сортове при дадени условия от средната стойност на опита , като :

$$l_j = X_{j.} - X ..$$

Оценката на регресионните коефициенти на показателите при промяната на екологическите условия е :

$$bi = (\sum X_{ij.} l_j) / (\sum l_i^2)$$

Оценката на дисперсията на стабилността S_i^2 дава количествен израз на колебанията на изследвания показател . Модела на теоретическото значение на признака е :

$$Yij = X_{i.} + b_i l_j$$

откъдето следва , че ;

$$S_i^2 = \sum (X_{ij.} - Y_{ij})^2/(k-2)$$

Прилага се F критерия на Фишер (5) за сравняване теоретическите и емпирически резултати .

Генетическа статистика

Статистическите методи намират широко приложение и в генетическите изследвания . Мощно приложение намира критерия χ^2 при оценка съответствието между експерименталните данни и теоретично очакваните резултати . Този критерий се прилага и при преценяване хомогенността на експерименталните данни както и в случаите когато трябва да се преценява съвпадението в няколко групи с няколко класа .

Изчисляването на основните статистически параметри за център на разпределението и разсейване на данните около този център е основен подход в количествената генетика . Техните стойности дават възможност за определяне типа на наследяване на признаците а разлагането на фенотипната дисперсия на генотипна и грешка носи информация за важни генетически статистически параметри като коефициент на наследяване в широк и тесен смисъл , селекционни индекси и корелационен коефициент на генетично ниво .

Дисперсионния анализ е в основата на подробен генетически анализ при селекционни задачи при опити заложени по диалелна или топкросна схеми . При диалелния анализ са разработени 4 модела (3) за разлагане на фенотипната дисперсията. Прилага се върху данни от P_1 , P_2 , F_1 , BCP_1 , BCP_2 и F_2 . За прилагането на този метод трябва да бъдат изпълнени следните условия :

1.) диплоидно разпадане
2.) хомозиготни родители
3.) отсъствие на генетични различия между реципрочните кръстоски
4.) независимо действие на неалелните гени , т.е. липса на епистазис
5.) отсъствие на множествен алелизъм
6.) независимо рекомбиниране на гените у родителите

При този метод още в ранните фази на селецията могат да бъдат оценени родителите , комбинационната им способност , да се получи информация за адитивността или доминантността и за степените на доминантност или свръх доминатност , броя на гените контролиращи признака и да бъдат избрани перспективните за селекцията кръстоски . Разработени са също така модели за генетически статистически анализ при топ кросни схеми , които са по лесни за реализирани в сравнение с диалелните .

Статистически софтуер

Развитието на изчислителната техника , навлизането на персоналните компютри в научно – изследователската работа , развитието на операционните системи и на обектното програмиране са условията които дадоха възможност за интензифициране на изследователската работа , както и за нейното прецизиране . Научните работници могат персонално да прилагат в своите изследвания постиженията на статистическата наука . Създадени са множество статистчески пакети като водещи в това отношение са SAS на националния статистически институт на САЩ и SPSS на университета в Кеймбридж .

Статистическия пакет БИОСТАТ е разработен в ДЗИ гр. Генерал Тошево и е ориентиран към статически анализи в биологични и селскостопански изследвания . В него са включени специфични статистически методи за оценка екологическата пластичност и стабилност , за анализ на много факторни опити , оценка на генетически параметри и генетически анализ по диалелни и топкросни схеми .

ЛИТЕРАТУРА

Дрейпер , Н., Г.Смит (1966) .Прикладной регрессионний анализ. Москва. Статистика.

Снедекор ,У.(1961) Статистические методи в применении к исследованиям в сельском хозяйстве и биологии .Москва.

Griffing , L. (1956) Concept of general and specific combining ability in relation to diallel crossing systems . Genetics .

France ,J. J.Thornley .(1987) . Mathematical models in agriculture . Boston

Little , T. , F.Hills (2002) . Agricultural Experimentation .New York

Mead , R.,R.Curnow,AHasted (2005) . Statistical methods in agriculture and experimental biology . London .

Schabenberger O.,C.Gotway (2005) . Statistical methods for spatial data analysis. New York.