Статистическо моделиране

1. Същност на статистическо моделиране

Статистическото моделиране (СМ) е моделиране при което чрез математически зависимости се представят емперични данни, получени от проведени експерименти. Това моделиране се прилага, когато е необходимо да се представят причино-следствените връзки в изследвания обект. Получените математически модели при статистическото моделиране се наричат статистически модели. Те отразяват връзките между изследваните параметри на обектите.

Експериментът е съвкупност от целенасочени действия, чрез които се разкрива същността на състоянието и функционирането на обекта на моделиране и изследване. Той се състои от опити и наблюдения. Опитът е част от експеримента, реализирана при при определен набор от условия. В резултат на реализиране на опита протича случайно събитие. Всеки опит се изразява с точка във факторното пространство. Многократното повтаряне на опита в една точка и събирането на данни за нея се нарича наблюдение. Всеки експеримент съдържа много опити, а опитът – едно или няколко наблюдения.

Статистическото моделиране включва обикновено два етапа:

1) Планиране и провеждане на експерименти за събиране на емперични данни;
2) Прилагане на подходящ метод за математически анализ на тези данни.

СМ се основа на метода на Монте Карло, който позволява намиране на приближени решения за различни зависимости, чрез формиране на статистически извадки.

Основните стъпки за статическото моделиране са:

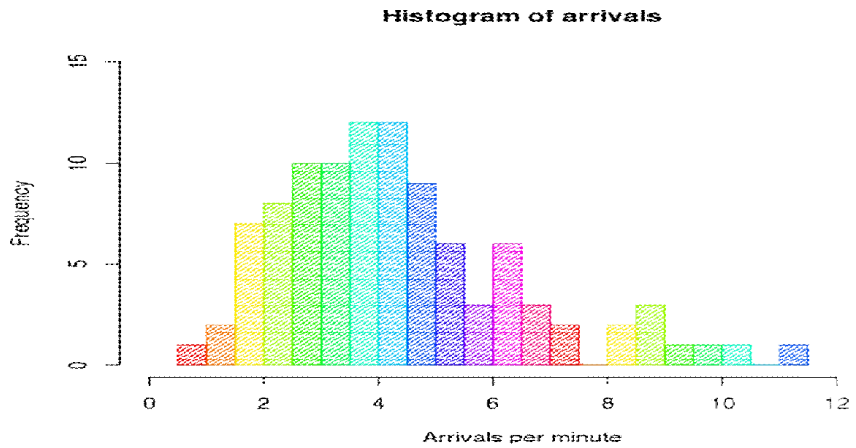
- ✓ Планиране на статистическия експеримент;
- ✓ Формиране на извадка;
- ✓ Избор на математически метод за статистическа обработка на извадката и получаване на статистически оценки за изследваните параметри.

Планирането на статистическия експеримент е свързано с определяне на стойностите на отделните променливи и броя на провежданите експерименти (опити) за натрупване на достатъчно количество данни.

За формиране на извадката се провеждат N на брой експеримента при еднакви условия и се снемат (регистрират) стойности за всяка от следените променливи. Броят на наблюденията N в дадена извадка зависи от броя на следените променливи. Той може да се определи по два начина, известни като:

- Класически, при който N е предварително определен;
- Динамично, при който N се определя в процеса на формиране на самата извадка чрез анализ на вече получените данни.

Всяка извадка се характеризира с разпределние. То може да се предтави чрез **хистограма** (фиг.1). Тя представя разпределението на стойностите на извадката по групи в зависимост от попаденията в отделните интервали от диапазона на изменение на стойностите на изследвания параметър. Хистограмата визуализира плътността на извадката.



Фиг. 1 Хистограма

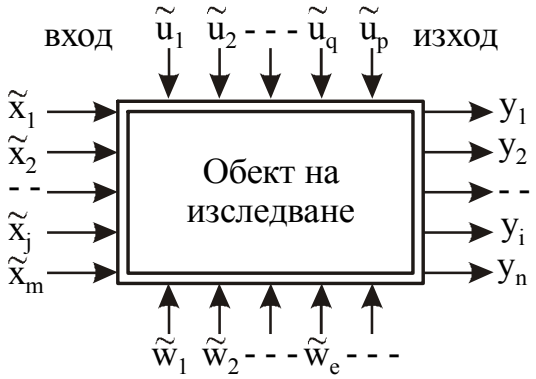
Хистограмата е форма на графично представяне на данните. Данните от хистограмата са изобразени като правоъгълници, които представляват отделните части, без да се припокриват. Изобразената област от всеки правоъгълник представлява съответната относителна честота. Обикновено по оста X представят категории данни. По оста Y се представя разпространението. Височината на правоъгълника изразява честотата или плътността на случаите.

Статистическото моделиране е свързано с приложение на методи и средства от математическата статистика за изчисляване на оценки за случайни величини и случайни процеси. Такива методи са: най-малките квадрати (МНК), стохастична апроксимация; параметрични функции и др. С тях се определят числовите характеристики (математическо очакване, дисперсия, начален момент и др).

2. Обекти и характеристики

Като обекти за статическо моделиране могат да се разглеждат компютърни системи и техните компоненти, компютърните мрежи, ресурсите и процесите, комуникационните канали и др.

Всеки обект разгледан като „черна кутия” се характеризира със входни въздействия X, изходни реакции У (фиг.2).



Фиг. 2. Обект на моделиране и изследване

Входните въздействия се представят с независими променливи. Те се наричат **фактори** и определят поведението на обектите.

Изходните реакции се представят с изходни величини. Те се наричат **параметри** и изразяват изменението на обектите под действието на факторите.

Всеки обект може да бъде определен чрез факторите и параметрите му.

Входните въздействия могат да бъдат:

- управляеми $\widetilde{x}_j$, $j \in \{1, m\}$- имат точно определена стойност или се изменят по определен начин;
- неуправляеми (смущаващи) $\widetilde{w}_l$, $l \in \{1, \infty\}$– с неизвестен характер и физическа същност;
- контролируеми $\widetilde{u}_q$, $q \in \{1, p\}$- стойностите им се контролират и поддържат постоянни през време на опита (тези фактори не участват пряко в изследването);

Обектите могат да се класифицират по различни критерии (признаци):

- по броя на факторите - еднофакторни и многофакторни;
- по броя на входните и изходните величини - еднопараметрични (един вход и един изход) и многопараметрични (всички останали);
- по вида на зависимостта между входните и изходните величини - линейни и нелинейни;
- по това дали се изменят характеристиките на обекта във времето – статични и динамични.

За всяка от изходните величини на обекта може да се разработи самостоятелен математичен модел за които:

- изходната величина зависи само от входните фактори:

$$y_i = \varphi_i(\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_m); \quad (1)$$

- изходна величина зависи от входните и изходните величини в различни моменти от време *t*:

$$y_i = \varphi_i[\widetilde{x}_1(t), \widetilde{x}_2(t), ..., \widetilde{x}_m(t), y_1(t), y_2(t), ..., y_l(t)]. \quad (2)$$

Факторите биват качествени и количествени. Качествените фактори се трансформират в количествени. Ако броят на факторите е много голям е желателно да се отразят само съществените (определящите за поведението и състоянието на обекта).

Факторите трябва да отговарят на следните изисквания: да са управляеми; да са еднозначни; да са съвместими; да са независими (един от друг); да са измерими (да допускат количествена оценка); да въздействат непосредствено върху обекта.

Множеството от всички възможни и допустими стойности на факторите се нарича областта на определение на факторите. Тя може да бъде ограничена или неограничена, непрекъсната или дискретна.

3. Методи за статистическа обработка на данните

Най-често използваните методи за анализ на резултатите от моделното изследване са на базата на статистическа обработка на данните. Някои от тези методи са:

- ✓ Честотно разпределение;
- ✓ Базови статистики;
- ✓ Корелационен анализ;
- ✓ Регресионен анализ;
- ✓ Дисперсионен анализ;
- ✓ Факторен анализ и статистически оценки.

При честотното разпределение за изследваната извадка от N регистрирани стойности за вектора на променливите се определят к непресичащи се класа от регистрации. Мощността m_i на всеки клас A_i се изчислява от относителните честоти на попадение $p_i = m_i/N$, за които е в сила равенството $\sum p_i = 1$, i=1÷k

Основните статистически оценки, използвани при метода на базовите статистики са:

- размер на извадката;
- размах на извадката;
- средно – аритметична стойност на извадката;
- средно отклонение;
- средно-квадратично отклонение;
- коефицент на отклонение (коефициент на вариация).

Корелационният анализ изследва взаимовръзката между две променливи X и Y , като изчислява коефицентите на корелация и ковариация за всяка двойка моделни параметри. Неговата основна задача е установяване на значимостта (силата) на връзката между значенията на различни случайни величини. При него се определят оценките на корелационните коефициенти и корелационната зависимост. Тя може да бъде:

- по форма: линейна или нелинейна;
- по характер: права (положителна) или обратна (отрицателна);
- според броя на едновременно действащите фактори: единична или множествена.

Единичната корелационна зависимост изразява зависимостта между два признака със съответните групи статистически случайни величини, без да се отчита влиянието на други фактори. Множествената корелационна зависимост отразява зависимостта между повече от два признака. При частната корелационна зависимост се отчита влиянието и на други фактори.

Основни величини в корелационния анализ са коефициентът на корелация и корелационното отношение.

Коефициентът на корелация е число, с което количествено се изразява съществуващата зависимост между две случайни величини X и Y. Той приема стойности в интервала [-1, +1] . В зависимост от знака на корелационния коефициент се различава положителна (при положително число) и отрицателна (при отрицателно число) корелация. От стойностите на корелационния коефициент се определя и степента на корелираност. Колкото абсолютната стойност на коефициента на корелация е по-близка до единица, толкова по-силна е зависимостта между компонентите X и Y.

Корелационното отношение е количествена характеристика на корелационна зависимост. То приема стойности в затворения интервал [0,1]. Ако стойността на корелационното отношение е 0, то няма корелационна зависимост между величините Y и X.

Регресионният анализ изследва причинната взаимовръзка при количествени фактори, представени като една зависима променлива У и една независима променлива X (проста регресия) или няколко независими променливи X1, X2, ...Xк (множествена регресия). За представяне на изследваната зависимост се прилагат линеен или нелинеен модел. За линейния модел се използва уравнение на права линия, а за нелинейния – нелинейни функции: експненциални, логаритмични и др.

Корелационният анализ показва взаимозависимост, а **регресионният** – причинно – следствени връзки между две или повече от две променливи. Може да приемат повече променливи като независими и да се търси тяхното влияние върху една зависима.

Изследователят определя кои променливи ще бъдат зависими и кои независими. Коефициент на регресия обикновено се бележи се с β.
При интерпретация, когато β>0, повишаването на значението на едната променлива води до повишаване значението на другата. Определяща е силата на тази връзка. Когато β<0, повишаването на стойността на независимата променлива води до понижаване на стойността на зависимата. При няколко независими променливи се използва допълнителен коефициент ΔR^2 (Adjusted R Square). Той показва какъв процент от случаите ще доведат до промени в зависимата променлива.
Най-често използван регресионен модел е така наречения стъпков регресионен модел. При него резултатите се извеждат на няколко стъпки. На първа стъпка се разглежда една от независимите променливи, които оказват влияние върху зависимата. След това на всяка стъпка се увеличава броят им. Колкото е по-голяма стойността, толкова по-голяма е възможността промените в независимите променливи да водят до промени в зависимата.

Дисперсионният анализ се прилага установяване на относителното влияние на различни фактори върху стойностите на входните характеристики. Обикновенно се използва за сравнение на средните стойности на отделни слоеве на една извадка (еднофакторен анализ) или при изследване на диференциалните ефекти при два фактора (двуфакторен анализ).
При еднофакторния анализ анализ се използва дисперсионен модел като математическо съотношение, при което всяка променлива се представя като сума от средна стойност и грешка.
При двуфакторния дисперсионен анализ се изледват два типа отношения:

- пресичане, когато двата фактора са представени чрез всички възможни комбинации за отделните нива;
- групиране, когато всяко отделно ниво на единия фактор участва в комбинации само с едно ниво на другия фактор.

Основни задачи на дисперсионния анализ са:

- ✓ установяване на количествена зависимост между изходната величина и входните фактори при моделиране на многофакторни обекти;
- ✓ качествена оценка (наличие или отсъствие) на влиянието на един или група фактори върху изходната величина.

Дисперсионният анализ е статистически метод за проверка на хипотези. Чрез тази проверка може да се прецени доколко влиянието на даден фактор или на група фактори е статистически значимо или не. Изследването на влиянието на факторите с помощта на оценките на дисперсиите е неговата същност. Ако оценките не се различават съществено се счита, че влиянието на фактора не е съществено. В противен случай се приема обратното твърдение (нулевата хипотеза H_0 се отхвърля).
Чрез дисперсионния анализ може само да се установи дали между изследваните параметри съществува зависимост.
За да се приложи дисперсионният анализ, е необходимо да бъдат изпълнени следните условия:

- ✓ да се направи анализ с цел да се установи кои признаци са взаимно свързани, кои са фактори и кои са резултати;
- ✓ разпределението на единиците в генералната съвкупност трябва да е нормално или близко до нормалното;

- ✓ извадките трябва да имат равни (еднакви) дисперсии;
- ✓ данните, които се използват да са от независими случайни извадки;

При **факторният анализ** него се решават две основни задачи:

- ✓ определяне на главните фактори;
- ✓ завъртането на главните фактори.

Завъртането на факторите е свързано със замяна на главните фактори с линейните им комбинации, които са взаимно некорелиарани и имат еденични дисперсии.
Статистическите оценки и проверките на хипотези се използват за формулиране на статистическите изводи. **Статистическата хипотеза** е твърдение относно валидността и адекватността на статистическия модел. **Проверката на хипотеза** е процедурата за приемане или отхвърляне на твърдение.
Статистическата оценка е изчисляване на дадена функция на променлива на базата на нейните стойности от извадката. Най-често за изчисляване на оценките се използва методът на най-малките квадрати или различни критерии за подобие (критерий на Пирсън, критерий на Колмогоров и др.).

4. Графичната интерпретация на статистически данни, резултати и оценки

Статистически данни са събраната, организирана и анализирана информация, необходима за изследване на даден обект. За представяне на измерените/снетите данни от наблюденията, както и на се използват резултатите от статистическата им обработка **статистически таблици**. Те се състоят от статистически редове.
Статистическите таблици се онагледяват графично чрез диаграми.
Диаграмите са геометрични фигури, рисунки или чертежи, които изобразяват графично данни, факти или информация в дву- или триизмерно пространство. На тях с определен мащаб се нанасят данните на признаците върху осите на подходящо избрана координатна система. Диаграмите се класифицитат на:

- Линейни диаграми – графичният образ е линия, свързваща отделните точки, представящи данните.
- Плоскостни диаграми – графичните образи са правоъгълници, триъгълници, кръгове и други.

Размерът на изучаваните явления се изразява с честотата. Тя може да бъде:

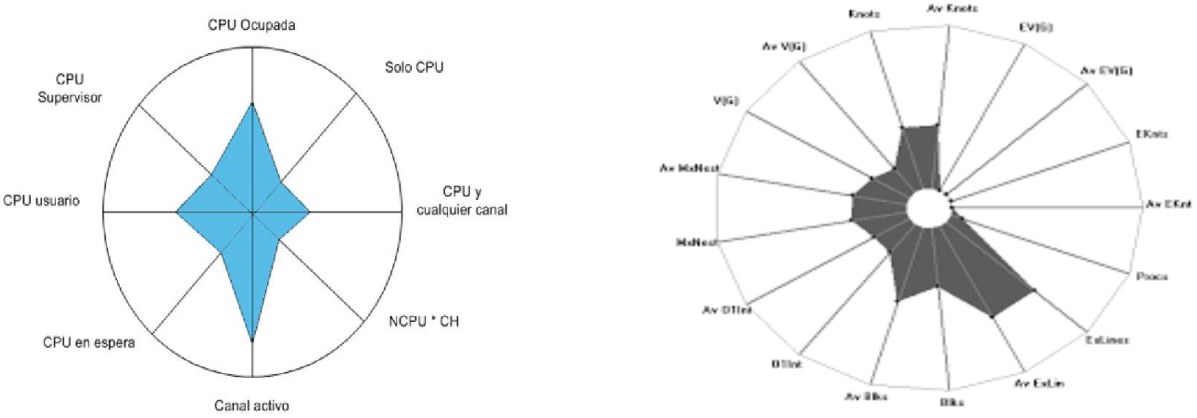
- Абсолютна честота – изразява броя на единиците от статистическата съвкупност, които се отличават по някакъв признак.
- Относителна честота (статистическа вероятност). Според закона за големите числа, колкото е по-голям броят на изследваните единици на генералната съвкупност, толкова по-малко наблюдаваните признаци се влияят от случайни причини и относителната честота се доближава до съответната вероятност.

Статистическа съвкупност е съвкупност от голям брой единици (случаи), които характеризират обекта на изследване. Тя може да бъде:

- Генерална съвкупност – обхваща всички случаи (всички данни);
- Представителна съвкупност (извадка) – обхваща част от случаите (част от данните) на генералната съвкупност.

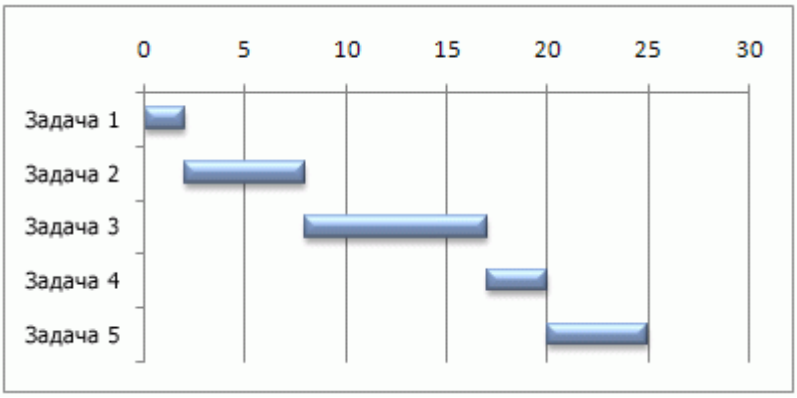
Графичната интерпретация на резултатите и оценките при статистическо моделиране може да се реализират чрез диаграма на Кивиат, диаграми на Гант, кръгови диаграми и др.

Диаграмата на Кивиат (фиг. 3) представлява многолъчева графика. В нея лъчите са равномерно разпределени радиуси на една окръжност. Радиусите са координатните оси, по които се изобразяват стонойностите на отделните фактори. Разполагането на величините по осите е произволно. Препоръчва се редуване на съседни оси на "добри" и "лоши" характеристики.



Фиг. 4. Диаграми на Кивиат

Диаграмата на Гант (фиг.4) се прилага за съвкупно представяне на характеристиките, свързани с използването на отделните компоненти на системата. За построяването и са необходими точните моменти време (регистрации на моделния таймер) за реализация на събитията. Тази диаграма предоставя графична информация за адекватността на компонентите на системите и припокриването им във времето. Диаграмата на Гант е ефективно средство за визуализация на времевите съотношения между отделните компоненти.



Фиг. 4. Диаграма на Ганд

Кръговата диаграма обикновенно представя дадена относителна характеристика, чрез площа на кръг, като отделните компоненти се представят чрез сектори. Всеки сектор изобразява част или процент, отнасяща се към дадена променлива в рамките на общосистемната характеристика. Тази диаграма позволява да се получи визуална интепретация за теглото на отделните компоненти или за значимостта на променливите на базата на относителните оценки, получени от анализите.

Кръговата диаграма представя графично количествена информация с помощта на окръжност, разделена на сектори, чиито относителни размери съответстват на пропорциите на количествата. По същество тази диаграма показва процентното отношение между частите в сравнение с цялото.