Тема лекции №5

Статистический анализ результатов моделирования.

Цель лекции – изучить особенности статистического анализа результатов моделирования.

План лекции.

1. Особенности фиксации и статистической обработки результатов моделирования транспортных процессов.
2. Требования к оценкам характеристик.
3. Построение гистограммы.
4. Элементы дисперсионного анализа. Критерий Фишера.
5. Однофакторный дисперсионный анализ.
6. Выявление несущественных факторов.
7. Сущность корреляционного анализа.
8. Обработка результатов эксперимента на основе регрессии.

1. Особенности фиксации и статистической обработки результатов моделирования транспортных процессов.

При выборе методов обработки существенную роль играют три особенности компьютерного эксперимента с моделью системы *S:*

1. Возможность получать при моделировании системы *S* на компьютере большие выборки позволяет количественно оценить характеристики процесса функционирования системы, но превращает в серьезную проблему хранение промежуточных результатов моделирования. Эту проблему можно решить, используя рекуррентные алгоритмы обработки, когда оценки вычисляют по ходу моделирования.

2. Сложность исследуемой системы *S* при ее моделировании на компьютере часто приводит к тому, что априорное суждение о характеристиках процесса функционирования системы, например о типе ожидаемого распределения выходных переменных, является невозможным. Поэтому при моделировании систем широко используются непараметрические оценки и оценки моментов распределения.

3. Блочность конструкции машинной модели $M_м$ и раздельное исследование блоков связаны с программной имитацией входящих переменных для одной частичной модели по оценкам выходных переменных, полученных на другой частичной модели. Если компьютер (программа), используемый для моделирования, не позволяет воспользоваться переменными, записанными на внешние носители, то следует представить эти переменные в форме, удобной для построения алгоритма их имитации.

Современные системы имитационного моделирования предоставляют возможность выполнять автоматически стандартную обработку результатов моделирования:

- определение характеристик случайных параметров, главным образом, их матожиданий и дисперсий;
- фиксация минимальных и максимальных значений исследуемых величин;
- частотное распределение результатов измерений (построение гистограмм);
- расчет коэффициентов использования объектов модели.

Часто приходится выполнять более сложную обработку:

- определение функциональных или статистических зависимостей между исследуемыми величинами;
- выявление существенных или несущественных факторов, участвующих в эксперименте;
- сравнение случайных параметров процесса с целью определения значимости расхождения или совпадения их характеристик и др.

Характеристики случайных величин и процессов

В результате эксперимента с имитационной статистической моделью, состоящего из *N* наблюдений, мы получаем *N* значений исследуемой случайной величины *a*:

$$a_1, a_2, ..., a_i, ..., a_N.$$

По этим данным нужно дать всестороннее описание величины *a*. Определить случайную величину - это значит определить ее характеристики. В общем случае:

$$\overline{\Theta} = \overline{\Theta}(a_1, a_2, ..., a_i, ..., a_N).$$

где $\overline{\Theta}$ - оценка характеристики случайной величины (СВ).

Под характеристикой СВ понимают следующее.

Во-первых, это характеристики *величины*:

- матожидание (среднее арифметическое);
- медиана (срединное значение);
- мода (наиболее вероятное значение);
- среднее геометрическое и др.

В рамках задач, характерных для нашей специальности, наиболее актуальным является матожидание. Как известно, матожидание определяет центр рассеивания случайной величины, наиболее полно отмечающее ее положение на числовой оси. Будем обозначать матожидание случайной величины *a* так: ***M(a).***

Во-вторых, это характеристики *рассеивания*:

- дисперсия (матожидание квадрата отклонения случайной величины *a*);
- среднее квадратическое отклонение (квадратный корень из дисперсии); иногда целесообразно пользоваться этой характеристикой, так как она имеет размерность самой случайной величины;
- размах (max*ai* - min*ai*).

В-третьих, это характеристика *связи* между случайными величинами (корреляция); *степень связи* определяется величиной коэффициента корреляции *r*.

В случайном процессе *связь* между значениями случайной функции в моменты времени t_k, t_s, определяет коэффициент автокорреляции $k(t_k, t_s)$.

В-четвертых, это характеристика *закона распределения вероятностей* случайной величины в виде плотности или функции распределения:

$$f(a)$$

или

$$F(a) = \int_{-\infty}^{a} f(a)da.$$

2. Требования к оценкам характеристик

Ограниченное число реализаций модели не позволяет точно определить значения этих характеристик, а только приближенно, то есть так называемые *оценки характеристик*. Степень приближения оценок зависит от методов их вычислений (формул).

Чтобы оценка наилучшим образом представляла искомую характеристику, нужно, чтобы она обладала следующими свойствами:

- несмещенностью;
- состоятельностью;
- эффективностью.

Несмещенность. Это свойство означает, что оценка не содержит систематической ошибки. Т.е., математическое ожидание оценки совпадает с действительным значением характеристики Θ:

$$M[\overline{\Theta}] = M[\Theta].$$

Состоятельность. Это свойство означает, что оценка $\overline{\Theta}$ приближается сколь угодно близко к истинному значению характеристики Θ по мере увеличения объема выборки, т. е. увеличения числа реализаций модели. Формально это свойство записывают так:

$$P(\left|\overline{\Theta} - M[\Theta]\right| < \varepsilon \to 1$$

при $N \to \infty$ и любом $\varepsilon > 0$

Именно это свойство являлось определяющим при нахождении количественной связи между точностью, достоверностью оценок и числом реализаций модели.

Эффективность. Это свойство означает, что из всех несмещенных и состоятельных оценок следует предпочесть ту, у которой разброс значений меньше. Иначе: эффективной оценкой характеристики случайной величины называют ту, которая имеет наименьшую дисперсию:

$$D[\overline{\Theta}] = \min \overline{\Theta}_k ,$$

где *k* - число возможных оценок.

Таблица 2.1 - Характеристики случайных величин и их оценки

Характеристика	Оценка	Среднее квадратическое отклонение оценки
Матожидание $M[x] = \int_{-\infty}^{-\infty} xf(x)dx$	$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$	$\sigma_{\overline{x}} = \frac{S}{\sqrt{N}}$
Дисперсия $D[x] = M[x^2] - (M[x])^2$	$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - overlinex)^2$	$\sigma_{S^2} = \sqrt{\frac{2}{N}}S^2$
Среднее квадратическое отклонение $\sigma_x = \sqrt{D[x]}$	$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$	$\sigma_S = \frac{S}{\sqrt{2N}}$
Вероятность события P	$\overline{P} = \frac{m}{N}$	$\sigma_{\overline{P}} = \sqrt{\frac{P(1-P)}{N}}$
Коэффициент корреляции $Q_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$	$\overline{r}_{xy} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{S_x S_y}$	$\sigma_{\overline{r}} = \frac{1-\overline{r}_{xy}^2}{\sqrt{N}}$

Все оценки несмещенные, состоятельные, эффективные.

3. Построение гистограммы

Одной из задач моделирования может быть *определение* закона распределения вероятностей исследуемой случайной величины и количественных значений его характеристик.

Аналогом, моделью плотности распределения вероятности случайной величины является *гистограмма*, которую можно построить (аналитически или графически) *по* данным имитационного моделирования. *Гистограмма* (рис. 3.1) строится так.

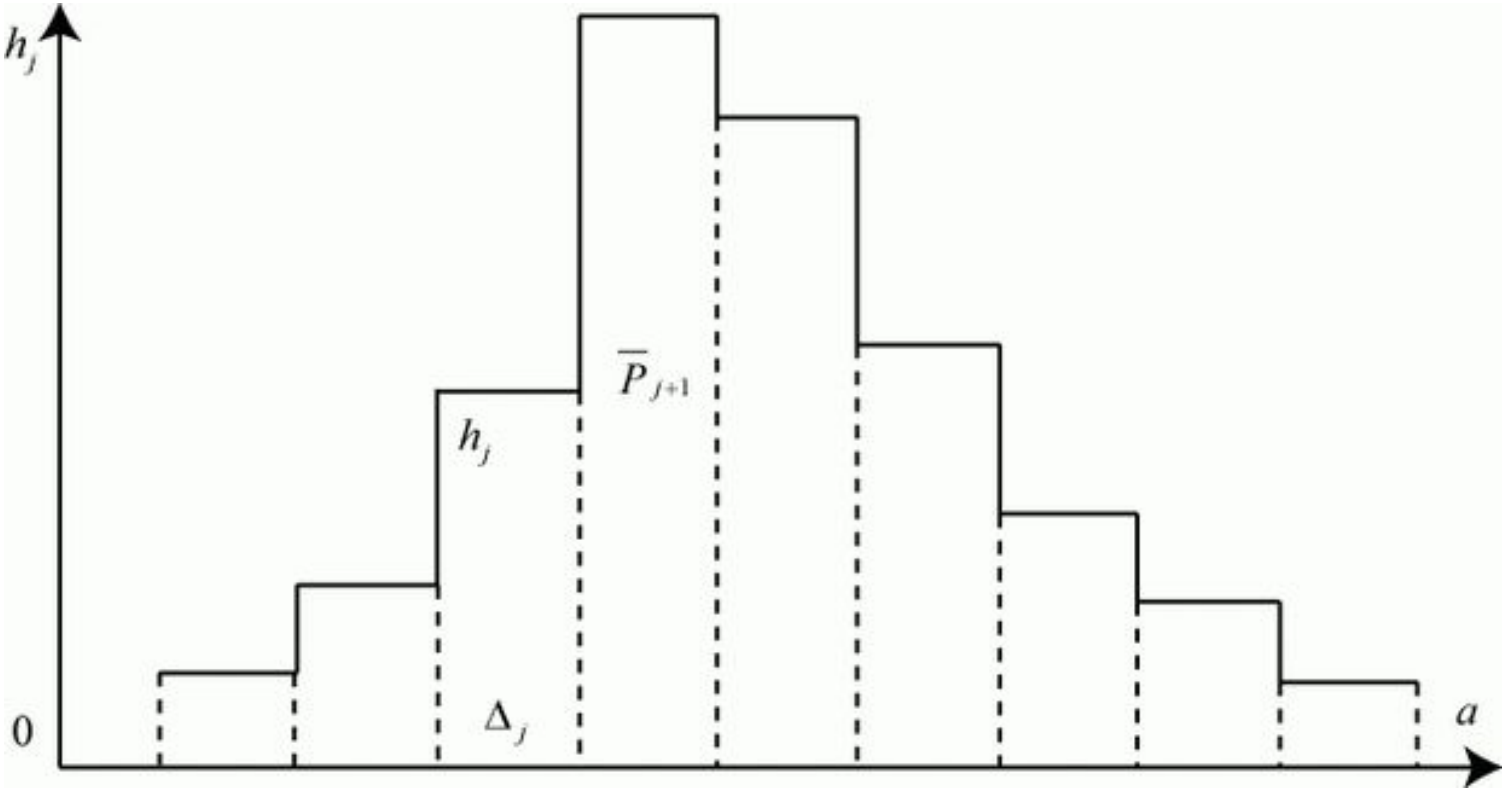


Рисунок 3.1 - Гистограмма

В результате *N* реализаций модели получен ряд случайных значений исследуемого параметра *a* : $a_1, a_2, ..., a_i, ..., a_N$.

Весь *диапазон* значений a_i разбивается на *l* интервалов (разрядов). Числовой *диапазон* каждого интервала обозначим Δ_j , *j=1,l*. Обычно все числовые диапазоны одинаковые: $\Delta_l = \Delta$.

Для каждого интервала подсчитываем число значений a_i, попавших в него - m_j.

На каждом интервале строят *прямоугольник* с высотой h_j:

$$h_j = \frac{m_j}{N \cdot \Delta} \cdot \Delta = \frac{m_j}{N}.$$

По выбору числа интервалов существуют разные эмпирические рекомендации. Чем больше *N* и *l*, а меньше Δ, тем ближе *гистограмма* совпадает с некоторым теоретическим распределением.

На основе очертания гистограммы делается предположение (выдвигается *гипотеза*) о совпадении полученного эмпирического распределения вероятностей с тем или иным теоретическим - нормальным, экспоненциальным, Вейбулла и т. д.

Затем выполняется проверка этой гипотезы с помощью *критериев согласия*. Рассматриваются некоторые (критерий Колмогорова, критерий Смирнова и др.), наиболее популярными считают критерий хи-квадрат - критерий Пирсона.

Оценки матожидания и дисперсии можно получить *по* данным гистограммы:

$$\overline{a} = \sum_{j=1}^{l} \overline{a_j} \cdot \overline{P_j}, \quad S^2 = \sum_{j=1}^{l} \overline{a_j} \cdot \overline{P_j} - \frac{\Delta}{12}.$$

где $\overline{a_j}$ - среднее значение каждого интервала;

$\overline{P_j}$ - оценка по каждому интервалу;

$\frac{\Delta}{12}$ - поправка Шеппарда.

4. Элементы дисперсионного анализа. Критерий Фишера

Гипотезой называется предположение о:

- законах распределения вероятностей случайных величин;
- значениях характеристик случайных величин;
- совпадении законов распределения двух и более случайных величин и др.

Обычно исходную гипотезу называют нулевой и обозначают H_0. Противоположное утверждение называют конкурирующей гипотезой и обозначают H_1.

Гипотеза подвергается проверке. Смысл этой проверки в том, чтобы принять или отклонить ее с допустимым минимальным риском. При этом возможны ошибки:

- забраковать проверяемую гипотезу, если она верна, что соответствует так называемой ошибке первого рода;
- принять проверяемую гипотезу, когда она не верна, значит совершить ошибку второго рода.

Правило, по которому принимается суждение об истинности или ложности основной гипотезы H_0 называют *критерием проверки или критерием согласия*.

Сущность *дисперсионного анализа* состоит в проверке гипотезы о тождественности выборочных дисперсий одной и той же генеральной дисперсии.

Также одновременно решает проблему проверки гипотезы о равенстве средних значений выборок.

Задача сравнения дисперсий сводится к проверке исходной гипотезы (нулевой гипотезы H_0) о принадлежности двух выборок одной и той же генеральной совокупности.

Для проверки гипотезы о равенстве дисперсий нужно иметь независимую функцию, вычислимую по данным эксперимента.

Такой функцией является *функция Фишера* (распределение Фишера, *F* -распределение), определяемая так:

$$F = \frac{\frac{U}{k_1}}{\frac{V}{k_2}},$$

где *U* и *V* - случайные величины, имеющие распределение χ^2;

k_1 и k_2 – соответствующие степени свободы случайных величин *U* и *V* соответственно: $k_1 = N_1 - 1$, $k_2 = N_2 - 1$;

N_1 *и* N_2 *– количество испытаний (объем выборки).*

Почему χ^2 является мерой сравнения дисперсий? А потому, что дисперсии, являясь суммой квадратов ошибок, имеют распределение χ^2.

Распределение *хи-квадрат* определяется следующим образом:

$$f(x) = \frac{1}{\Gamma(v/2) \cdot 2^{v/2}} \cdot x^{(v/2)-1} \cdot e^{-x/2} \quad x > 0 \quad v = 1, 2, \ldots$$

где v – число степеней свободы;

Г – гамма-функция.

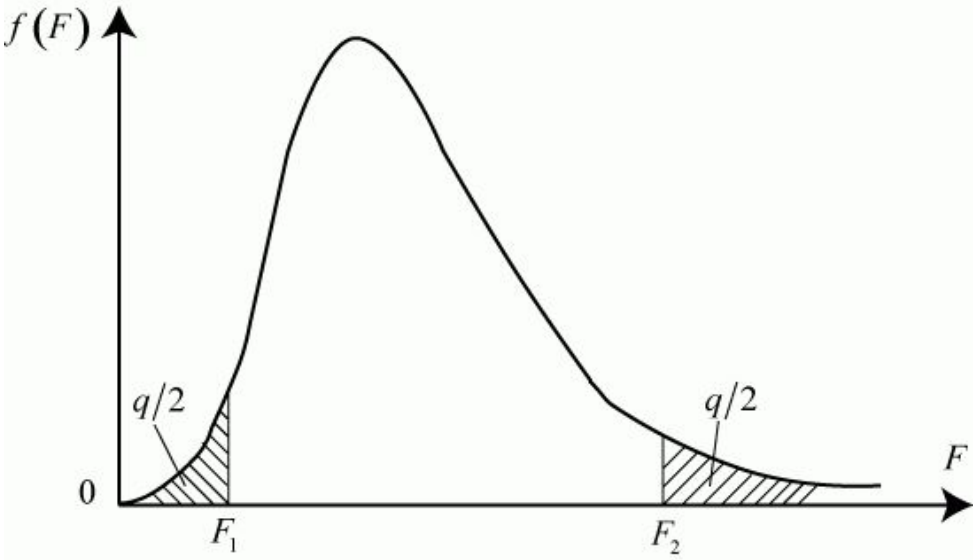
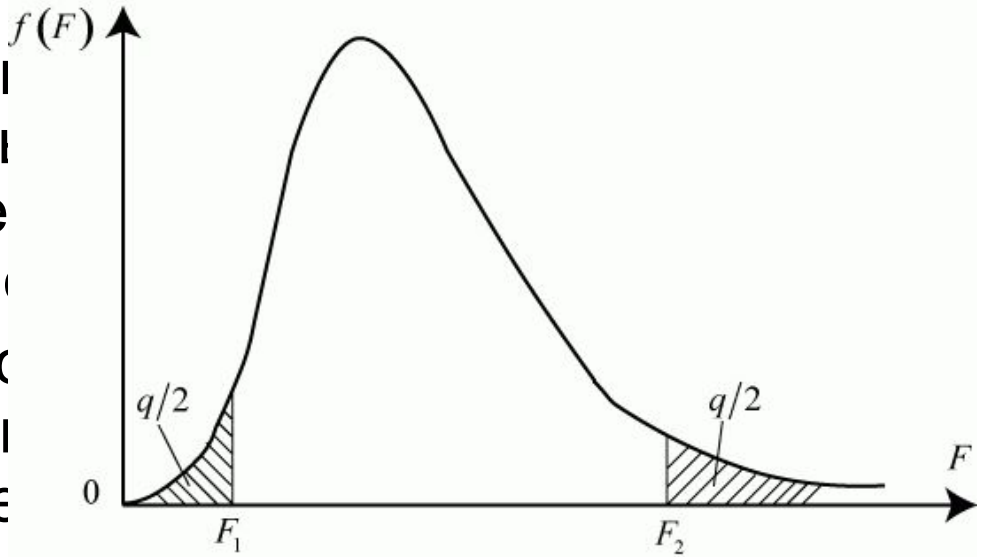


Рисунок 4.1 - График плотности F -распределения

Итак, *случайная величина*

$$F = \frac{S_1^2}{S^2},$$

где S^2_1 и S^2_2 - несмещенны
полученные из независимы
нормальных совокупносте
Фишера (*F* -распределени



Величина *F* - случайна, поэт
величине о подтверждени
об однородности исследуе

Поэтому вводится ***q*** уровень значимости, численно равный вероятности *неприемлемых* отклонений от принятой гипотезы. Области неприемлемых значений ***F*** показаны на рисунке штриховкой. Граничные точки допустимых значений ***F*** определяются точками F_1 и F_2, соответствующих вероятностям ***q/2*** .

Если вычисленное по данным эксперимента *значение* ***F*** попадает в область между точками F_1 и F_2:

$$F_1 \le F \le F_2,$$

то принятая *гипотеза* **не опровергается**.

Заметим, что *случайная величина*

$$F^* = \frac{1}{F} = \frac{S_2^2}{S_1^2}$$

также имеет *F* -распределение со степенями свободы k_1 и k_2 соответственно. Следовательно, вероятность попадания числа ***F*** в левую критическую область равна:

$$P(F < F_1) = P(\frac{1}{F} > \frac{1}{F_1}) = P(F^* > \frac{1}{F_1}).$$

Отсюда следует, что *левая критическая точка F* - распределения соответствует *правой* критической точке *F**-распределения. Т.е. правые точки распределений ***F*** и ***F**** определяют левую и правую точки F_1 и F_2. Поэтому в таблицах представлены только правые F_2 критические точки *F* -распределения.

В таблицах значения F_2 приведены в зависимости от *q/2* , числа степеней свободы k_1 и k_2 .

Обычно при вычислении ***F*** в *числитель* отношения $\frac{S_1^2}{S_2^2}$ ставят *значение* большей дисперсии.

Итак, при $F \le F_2$ принятая *гипотеза* не опровергается, при $F > F_2$ - не подтверждается.

Пример

Анализируем две хронометражны
за временем погрузки. В
измерений значений, дисперс
S_1=0,15ч., во второй – по сем
Однотипны ли измерения? Е
гипотезу об их однотип
значимости q=10.

Решение.

$$F = \frac{S_1^2}{S_2^2} =$$

$$k_1 = N_1 - 1 =$$

По таблицу *F* -распределения для
соответствующей большей
соответствующей меньшей
значимости $q/2$, находим F_2= 6,1

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Так как $F=1,5<F_2=6,16$, то для уровня значимости $q=10$ *гипотеза* об одинаковости результатов наблюдений не опровергается.

Итак: чем меньше уровень значимости q, тем меньше вероятность забраковать проверяемую гипотезу, когда она верна, т. е. совершить ошибку первого рода.

Но с уменьшением уровня значимости (увеличения F_2) расширяется область допустимых ошибок, что приводит к увеличению вероятности принятия неверного решения, т. е. совершения ошибки второго рода.

5. Однофакторный дисперсионный анализ

Эксперимент для выполнения однофакторного дисперсионного анализа (ОДА) состоит в накоплении результатов измерений *контролируемого параметра* при каждом варианте исследуемого фактора.

Введем обозначения:

- n - число вариантов фактора;
- m - число измерений при каждом варианте;
- a_{ij} - результат каждого измерения;
- $i=1,m$ - номер варианта фактора;
- $j=1,n$ - номер измерения.

Схема эксперимента заключается в следующем.

Производится n измерений *контролируемого параметра* при m вариантах фактора.

В принципе, число измерений может быть разным для каждого варианта фактора. Результаты эксперимента сводятся в табл.5.1.

Таблица 5.1 – Результаты эксперимента

№ варианта	Номер измерения						Средние значения
	1	2	...	j	...	n	
1	a_{11}	a_{12}	...	a_{1j}	...	a_{1n}	$\overline{a}_1 = \frac{1}{n}\sum_{j=1}^{N} a_{1j}$
2	a_{21}	a_{22}	...	a_{2j}	...	a_{2n}	$\overline{a}_2 = \frac{1}{n}\sum_{j=1}^{N} a_{2j}$
...							
i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{in}	$\overline{a}_i = \frac{1}{n}\sum_{j=1}^{N} a_{ij}$
...							
m	a_{m1}	a_{m2}	...	a_{mj}	...	a_{mn}	$\overline{a}_m = \frac{1}{n}\sum_{j=1}^{N} a_{mj}$

Вопрос: влияют ли варианты фактора на *точность* измерений? Или, говоря языком математической статистики, являются результаты измерений выборкой одной генеральной совокупности, или нет? Если да, то варианты фактора несущественны, если нет, то существенны.

Будем исходить из следующей нулевой гипотезы:

- наблюдения каждого варианта независимы;
- наблюдения каждого варианта имеют нормальное распределение;
- имеют одинаковую дисперсию σ^2;
- имеют одинаковые центры рассеивания.

Очевидно, если систематические ошибки вариантов не одинаковы, следует ожидать повышенного рассеивания выборочных средних $\overline{a_i}$.

Для подтверждения или отрицания выдвинутой нулевой гипотезы об идентичности вариантов фактора проведем *дисперсионный анализ*.

Общее среднее арифметическое по всем $m \cdot n$ измерениям:

$$\overline{a} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}}{m \cdot n}.$$

Сумма квадратов отклонений по всем $m \cdot n$ измерений, то есть по данным всего эксперимента:

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \overline{a})^2.$$

Эту сумму квадратов отклонений можно разложить на два независимых слагаемых:

$$Q = n\sum_{i=1}^{m}(\overline{a}_i - \overline{a})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} - \overline{a})^2.$$

Обозначим:

$$Q_1 = n\sum_{i=1}^{m}(\overline{a}_i - \overline{a})^2, \;\; Q_2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} - \overline{a})^2$$

где Q_1 - сумма квадратов отклонений между вариантами фактора, так как $\bar{a}_i$ - среднее значение измеренного параметра *i*-го варианта фактора;

Q_2 - характеризует отклонения внутри каждого варианта.

Если принятая *гипотеза* о равенстве центров рассеивания a_i и σ^2 дисперсий верна, тогда все $m \cdot n$ наблюдений значений a_{ij} можно рассматривать как выборку из одной и той же нормальной совокупности с очевидной несмещенной оценкой дисперсии:

$$S^2 = \frac{Q}{m \cdot n - 1}$$

Можно показать, что величина

$$\frac{1}{m-1} Q_1,$$

имеющая распределение χ^2 со степенями свободы $m-1$, является оценкой дисперсии S^2. И величина

$$\frac{1}{m(n-1)} Q_2$$

имеющая распределение χ^2 со степенями свободы $m(n-1)$, также является оценкой дисперсии.

Из сказанного следует, что критерий

$$F = \frac{\frac{1}{m-1}Q_1}{\frac{1}{m(n-1)}Q_2}$$

при нашей гипотезе и независимости Q_1 и Q_2 (это можно доказать) имеет F -распределение с $m-1$ и $m(n-1)$ степенями свободы.

А дальше мы уже знаем, как поступить:

- выбираем уровень значимости q;
- вычисляем число F;
- из таблицы по величине $q/2$ находим F_2.

Пример

Необходимо проверить однотипность погрузчиков с одинаковыми характеристиками трех производителей. Осуществляется контрольные погрузки одного вида груза 30 раз каждым погрузчиком и замеряли время. Отклонение от нормативного времени занесли в таблицу 5.2.

Решение.

Проверяем исходную гипотезу: погрузчики по времени выполнения операций одинаковы.

При выборе уровня значимости *q* исходим из того, что более опасна ошибка второго рода - подтвердить ошибочный выбор. Примем *q =10*.

Число вариантов фактора: *m*=3.

Число измерений: *n*=30.

Среднее отклонение времени по трем погрузчикам: $\bar{a}_1, \bar{a}_2, \bar{a}_3$.

Среднее отклонение по 90 замерам: $\bar{a}$.

Средний квадрат расхождений между вариантами факторов:

$$\frac{1}{m-1}Q_1 = \frac{1}{m-1}n\sum_{i=1}^{n}(\overline{a_i} - \overline{a})^2 =$$

Таблица 5.2 – Исходные данные

№эксперимента	Отклонение времени по погрузчикам, мин.		
	Погр.1	Погр.2	Погр.3
1	2,6	2,5	1,1
2	2,1	3,4	3,9
3	2,4	2,3	2,8
4	2,3	3,7	2,7
5	2,6	1,4	3,0
6	2,6	2,9	1,7
7	2,4	1,2	1,2
8	1,6	3,4	3,9
9	1,8	1,8	3,9
10	1,8	3,5	1,6
11	3,1	3,8	1,7
12	3,5	1,1	3,9
13	2,1	1,5	1,9
14	2,3	1,6	2,7
15	3,0	3,6	1,8
16	2,7	1,7	2,7
17	3,0	2,4	2,4
18	2,7	1,8	2,4
19	3,3	3,7	3,5
20	1,3	2,6	2,6
21	2,9	3,4	2,6
22	4,0	3,8	2,6
23	1,6	3,8	3,0
24	2,4	1,7	1,3
25	3,0	2,6	2,4
26	2,2	3,6	2,9
27	1,2	2,3	1,6
28	1,1	1,6	1,0
29	3,6	3,9	2,8
30	2,2	1,0	2,5

Число степеней свободы: *m-1=*

Средний квадрат расхождений вн

$$\frac{1}{m(n-1)}Q_2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}$$

Число степеней свободы: *m(n-1)=*

Расчет *F* -критерия: *F=*

По таблице определения критери
q/2=5 верхних пределах отклон
имеющихся степенях свободы

Далее сравниваем *F и* F_2 *и если*
вывод, что выдвинутая гипот
опровергается.

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

6. Выявление несущественных факторов

Большое количество факторов усложняет и снижает эффективность эксперимента. Среди этого *множества* могут быть и несущественные факторы. *Исключение* их упростило бы эксперимент, не снижая его информативности.

Несущественный фактор выявляется так.

Выполняются первый эксперимент из *N* наблюдений с учетом проверяемого фактора и второй эксперимент также из *N* наблюдений - без него. В обоих случаях фиксируются отклики *y*. Делается предположение, что обе выборки принадлежат одной генеральной совокупности, то есть, что проверяемый фактор - несущественный (это нулевая *гипотеза*). Дальнейшие рассуждения должны либо *не опровергнуть* эту гипотезу, либо посчитать ее *недостаточно обоснованной*.

Итак, получены две последовательности откликов, в которой y'_i и y''_i - значения откликов в *i*-м наблюдении при наличии и отсутствии проверяемого фактора соответственно:

Согласно принятой гипотезе эти последовательности имеют одинаковые матожидания *M(y)* и дисперсии σ^2_y .

Рассмотрим случайную величину *Z*, реализациями которой является последовательность случайных чисел

$$z_i = y_i^{'} - y_i^{''}, i = \overline{1, N}.$$

При независимости z_i и достаточно большом числе наблюдений *N* согласно центральной предельной теореме:

$$M[\overline{z}] = M[z] = 0, \;\; \sigma_{\overline{z}} = \frac{\sigma_z}{\sqrt{N}}.$$

Очевидно:

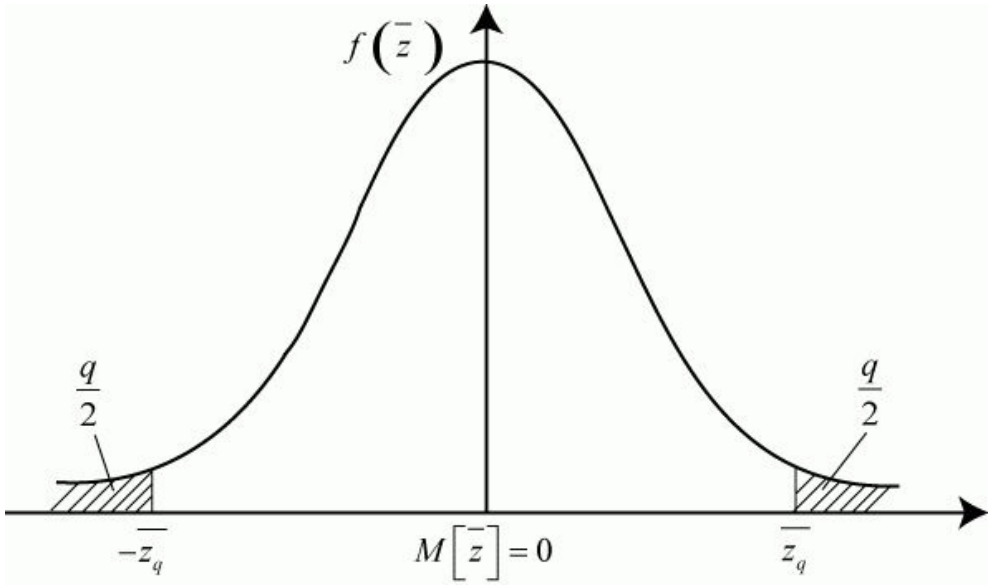
$$\sigma_z^2 = 2\sigma_y^2, \;\; \sigma_z == \sigma_y\sqrt{2}.$$

Как отделить случайные отклонения $\overline{z}$ от нуля от тех, которые мы будем считать не подтверждающими принятую гипотезу?

Такое разделение осуществляется по следующему правилу: если вычисленная величина $\overline{z}$ окажется маловероятной, в рамках нормального распределения и данном среднеквадратическом отклонении σ_z, то такое отклонение $\overline{z}$ от нуля считается не соответствующим принятой гипотезе.

Эту малую вероятность называют уровнем значимости и обозначают *q*. Обычно *q=2* - в зависимости от степени опасности совершения ошибки первого или второго рода.

На графике плотности распределения f($\overline{z}$) уровень значимости *q* показан заштрихованным участком (рис.6.1).



Для нормального закона распределения случайной величины $\overline{z}$ *вероятность* превышения $\overline{z}$ некоторого значения определяется известным выражением:

$$P\left(|\overline{z}-0|>t_\alpha\frac{\sigma_z}{\sqrt{N}}\right)=1-2\Phi(t_\alpha)$$

Следовательно:

$$\overline{z}=t_\alpha\frac{\sigma_z}{\sqrt{N}}$$

граничное значение

$$[1-2\Phi(t_\alpha)]=q$$

Аргумент функции Лапласа t_α находим из соответствующего справочника согласно $t_\alpha=\Phi^{-1}\left(\frac{1-q}{2}\right)$

и, как было указано ранее, $\sigma_z=\sigma_y\sqrt{2}$

Из изложенного следует:

- если $$\overline{z} > t_\alpha \frac{\sigma_y \sqrt{2}}{\sqrt{N}}$$

принятая гипотеза о несущественности проверяемого фактора не подтверждается;

- если $$\overline{z} \le t_\alpha \frac{\sigma_y \sqrt{2}}{\sqrt{N}}$$

принятая гипотеза не опровергается (в рамках принятого уровня значимости *q*).

Обычно величина σ_y неизвестна, поэтому следует использовать ее оценку S_y:

$$S_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\overline{y} - y_i)^2}$$

Оценку $\overline{y}$ и ряд значений y_i можно получить из данных первого эксперимента (y_i') или второго (y_i''), так как в силу рассматриваемой гипотезы они идентичны. Однако следует помнить, что если N<100 то вместо аргумента функции Лапласа t_α надо брать t'_α - *аргумент* функции Стьюдента.

Пример

Исследуется зависимость времени пребывания заявки в системе массового обслуживания от дисциплины выборки заявок из очереди: *LIFO* или *FIFO*. Проведены два эксперимента. Первый эксперимент из *N=100* наблюдений с дисциплиной *FIFO* и второй эксперимент также из *N=100* наблюдений с дисциплиной *LIFO*. Результаты измерений и вычислений: $\overline{z}$=1,8 мин., S_y=2мин. Для уровня значимости q=5, t_α=1,96.

Решение.

Принимаем, что принятая гипотеза не опровергается, тогда:

$$\overline{z}_q \le t_\alpha \frac{S_y\sqrt{2}}{\sqrt{N}} = 1.96\frac{2\cdot 1.41}{10} = 0.55$$

Так как $\overline{z} = 1.8 > \overline{z} = 0.55,$ то *гипотеза* не подтверждается. Для времени пребывания заявки в очереди в системе массового обслуживания не безразлично, какая дисциплина выборки заявок из очереди применена.

7. Сущность корреляционного анализа

Часто при исследовании объекта или его модели необходимо наблюдать за характеристиками двух и более случайных величин. Например, за двумя откликами одного эксперимента. При этом может возникнуть вопрос: есть ли *связь* между этими случайными величинами? Существенна или несущественна эта *связь*, если она есть?

Корреляционный анализ - это совокупность методов обнаружения зависимости (корреляции) между двумя или более случайными признаками или процессами.

Под *корреляцией* будем понимать статистическую зависимость между двумя случайными величинами, не имеющую, вообще говоря, строго функционального характера.

Заметим, что *корреляционный анализ* не позволяет определить вид функциональной связи между случайными величинами, а только наличие или отсутствие предполагаемой связи, например, линейной, параболической, экспоненциальной и т.д.

Название "корреляционный *анализ*" происходит от латинского слова correlatio - согласование, *связь*, соотношение, взаимосвязь.

Обычно исследуют парную корреляцию, то есть *зависимость* между двумя случайными величинами (процессами), хотя возможны и более сложные ситуации, когда необходимо обнаружить наличие или отсутствие связей между тремя или более случайными величинами.

Мы ограничимся исследованием *парной корреляции*.

Как известно, *связь* между двумя случайными величинами можно описать с помощью двумерной функции распределения. Однако такое описание часто очень сложно, а для практических целей можно удовлетвориться определением зависимостей средних значений.

Итак, целью имитационного эксперимента является *определение* характеристик двух случайных величин *a* и *b*.

Необходимо проверить: есть ли *связь* между величинами *a* и *b*?

Проверка наличия (или отсутствия) связи - корреляции - между случайными величинами выполняется так.

Проводится два эксперимента, каждый - с соответствующей моделью. В каждом эксперименте -*N* наблюдений (напоминаем, что компьютерный эксперимент состоит из наблюдений, а наблюдение - из реализаций (прогонов) модели, число которых рассчитывается с учетом требуемой точности и достоверности получаемых результатов моделирования).

В результате экспериментов получаются два *множества* значений измеряемых параметров *a* и *b*: a_i и b_i, $i = \overline{1, N}$

Из этих множеств формируются пары:

$$(a_1, b_1), (a_2, b_2), \ldots, (a_i, b_i), \ldots, (a_N, b_N).$$

Каждая пара интерпретируется как *координаты* случайной точки в системе координат *a, b*. Первичное исследование можно провести графически. Возможны следующие варианты размещения точек на графиках (рисунок 7.1).

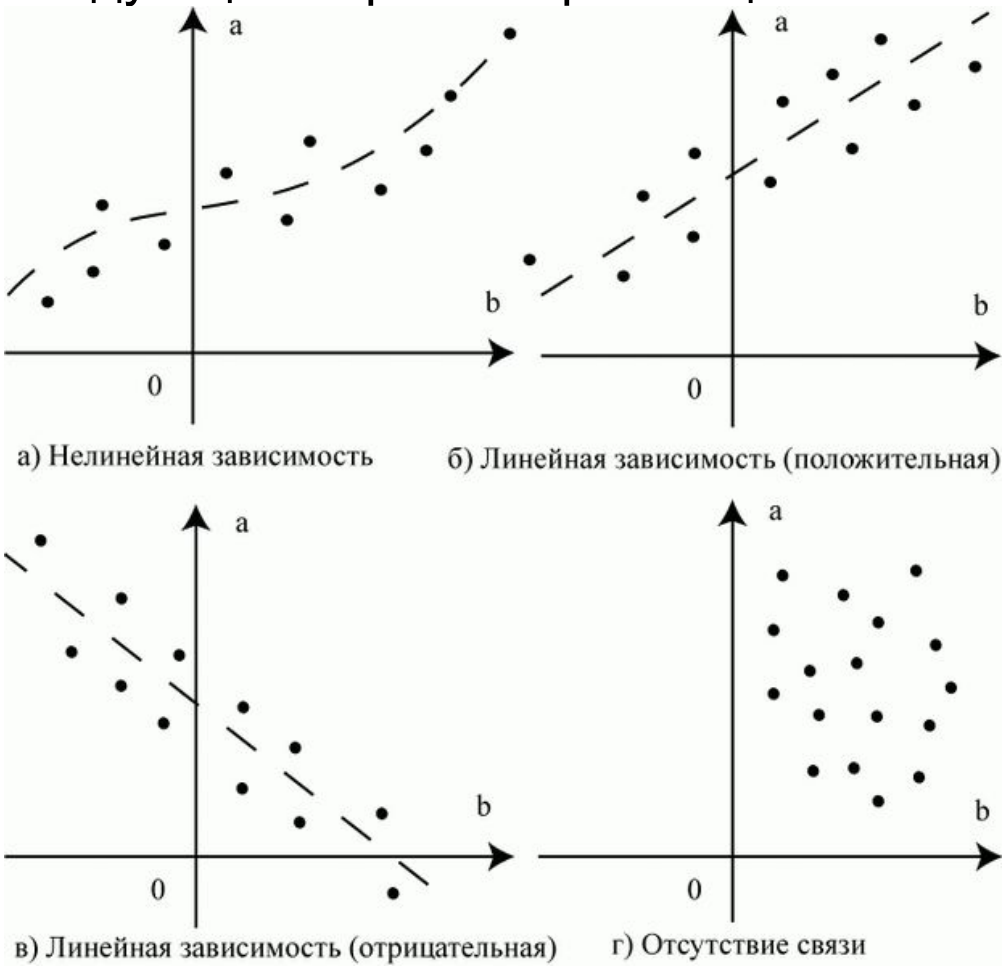


Рисунок 7.1 - Графическое исследование корреляции

Можно научиться визуально определять по расположению данных, насколько тесно они коррелированны.

Говорят, что две переменные *положительно* коррелированны, если при увеличении значений одной переменной увеличиваются значения другой переменной (рисунок 7.1-б).

Две переменные *отрицательно* коррелированны, если при увеличении одной переменной другая *переменная* уменьшается (рисунок 7.1-в).

Отсутствие корреляции - совместного поведения переменных - обнаруживается хаотическим нагромождением точек, исключающим проведение какой-либо аппроксимирующей линии (рисунок 7.1-г).

Но такое *качественное исследование* недостаточно. Необходимо иметь количественную оценку степени корреляции между величинами *a* и *b*.

Если совместное распределение вероятностей случайных величин и нормальное, то количественной характеристикой степени линейной связи между ними является *коэффициент корреляции r* (введен Пирсоном):

$$-1 \leq r \leq 1.$$

Если *r=0*, то между *a* и *b* линейная независимость.

Равенство r=±1 свидетельствует о наличии однозначной функциональной связи между *a* и *b*, то есть *b=f(a)*.

При -1<r<1 между *a* и *b* существует стохастическая *связь*, причем, чем ближе *коэффициент корреляции* |*r*| к единице, тем эта *связь* сильнее.
Стохастическая *связь* означает, что при изменении *a* имеется лишь тенденция к изменению *b*.

Коэффициент корреляции r определяется *по* данным эксперимента, следовательно, можно определить только его оценку $\overline{r}$. В качестве оценки $\overline{r}$ принят выборочный *коэффициент корреляции*:

$$\overline{r} = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - \overline{a})(b_i - \overline{b})^2}{S_a \cdot S_b}$$

где ā - оценки математических ожиданий и *M(a)* и *M(b)*;

$$\overline{a} = \frac{1}{N}\sum_{i=1}^{N} a_i$$

S_a,S_b - оценки среднеквадратических отклонений σ_a,σ_b

$$S_a = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(a_i - \overline{a})^2}, \quad S_b = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(b_i - \overline{b})^2}$$

8.Обработка результатов эксперимента на основе регрессии

Часто целью исследования является определение функциональной связи между факторами и откликом (реакцией модели) по данным, полученным при экспериментах с моделью объекта или непосредственно с объектом. Такая цель достигается регрессионным анализом значений факторов *x* и отклика *y*.

Под *регрессией* в теории вероятностей и математической статистике понимают зависимость среднего значения какой-либо величины от некоторой другой (других) величины.

Регрессионный анализ - это совокупность методов построения и исследования *регрессионной зависимости* между величинами (в нашем случае между факторами и откликом) по статистическим данным. Статистические данные накапливаются при проведении эксперимента.

Функциональную зависимость между факторами и откликом представим в виде аппроксимирующего полинома:

$$\overline{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = \sum_{i=0}^{n} \beta_i x_i, \quad x_0 = 1$$

Этот *полином* получил название уравнения регрессии, а *коэффициенты* β_i - *коэффициенты регрессии*. От точности подбора *коэффициентов регрессии*

зависит *точность* представления *f(x).*

Коэффициенты β_i определяются путем обработки полученных в ходе эксперимента варьируемых значений факторов и откликов.

Однако из-за ограниченного числа наблюдений точные значения β_i получить нельзя, будут найдены их оценки b_i:

$$\overline{\beta_i} = b_i.$$

Поэтому уравнение регрессии принимает вид:

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n = \sum_{i=0}^{n} b_i x_i, \quad x_0 = 1$$

В уравнении регрессии могут участвовать и так называемые "совместные эффекты" ($x_1 x_2$, $x_1 x_2 x_3$ и т. п.) или степени значений факторов (x^2_1, x^3_2 и т. п.). Совместные эффекты и степени факторов можно обозначать обобщенным фактором. Например, уравнение регрессии

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_2^2$$

можно представить так:

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4, \quad x_3 = x_1 x_2, \quad x_4 = x_2^2$$

Итак, для определения выражения *f(x)* надо:

- выбрать степень аппроксимирующего полинома - уравнения регрессии;
- определить *коэффициенты регрессии*.

Выбор уравнения регрессии обычно начинают с линейной модели. Например, для двухфакторного эксперимента ее вид:

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2$$

Если окажется, что такая *аппроксимация* дает неприемлемые отклонения при сравнении с экспериментальными точками отклика *y* , то модель усложняется, например, так:

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

или

$$\overline{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2$$

Коэффициенты регрессии b_i для выбранного уравнения определяются из условия минимума суммы квадратов ошибок, вычисленных по всем экспериментальным точкам. Это делается так. Введем обозначения:

x_{il} - *значение* *i*-го фактора в наблюдении номер *l*;

y_l - *значение* отклика в *l*-м наблюдении;

$\overline{y}_l$ - *значение* отклика, вычисленное *по* принятому уравнению регрессии и данным x_{il}.

Очевидно, сумма квадратов ошибок между экспериментальными значениями y_l и вычисленными по уравнению регрессии y_l для всех *N* наблюдений равна:

$$\delta = \sum_{l=1}^{N} (y_l - \overline{y}_l)^2 = \sum_{l=1}^{N} \left(y_l - \sum_{i=0}^{n} b_i x_{il} \right)^2$$

Для определения минимума ошибки возьмем частные *производные* от δ по всем неизвестным коэффициентам регрессии b_j, $j=1,n$, и приравняем их нулю:

$$\frac{\partial \delta}{\partial b_j} = -2 \sum_{l=1}^{N} \left(y_l - \sum_{i=0}^{n} b_i x_{il} \right) x_{jl} = 0$$

Это условие минимума, а не максимума. Очевидно:

$$\sum_{l=1}^{N} \left(y_l - \sum_{i=0}^{n} b_i x_{il} \right) x_{jl} = 0,$$

$$\sum_{l=1}^{N} y_l x_{jl} = \sum_{l=1}^{N} \sum_{i=0}^{n} b_i x_{il} x_{jl}.$$

Для лучшей наглядности выделим неизвестные *коэффициенты регрессии* и получим:

$$\sum_{l=1}^{N} y_l x_{jl} = \sum_{l=1}^{N} b_i \sum_{i=0}^{n} x_{il} x_{jl}$$

Данное выражение представляет собой систему из *n+1* уравнений для нахождения *n+1* неизвестных *коэффициентов регрессии* b_i, которые окончательно определят выбранное уравнение регрессии.

Нахождение *коэффициентов регрессии* справедливо при следующих допущениях:

- Случайный фактор ξ имеет нормальное распределение с матожиданием M[ξ]=0.
- Результаты наблюдений y_l - независимые нормально распределенные случайные величины. Если это не соблюдается, то следует измерять другой отклик, удовлетворяющий этому условию, но функционально связанный с исследуемым откликом *y*.
- Точность наблюдений (количество реализаций модели) не меняется от наблюдения к наблюдению.
- Точность наблюдения x_{il} должна быть выше точности y_l.