ЛЕКЦИЯ 6. Факторный дисперсионный анализ. Выявление несущественных факторов. Обработка результатов эксперимента на основе регрессии

Цель лекции

Сформировать у обучающихся теоретические знания об основных принципах и методах факторного дисперсионного и корреляционного анализа результатов экспериментов. Изучить методы оценки влияния исследуемых факторов на выходные характеристики системы, выявления несущественных факторов, проверки статистических гипотез и определения степени взаимосвязи между случайными величинами. Развить навыки интерпретации экспериментальных данных и применения статистических методов для обоснования результатов математического и имитационного моделирования.

План лекции

1. Понятие и назначение факторного дисперсионного анализа.
2. Однофакторный дисперсионный анализ и организация эксперимента.
3. Основные статистические характеристики и разложение общей дисперсии.
4. Проверка гипотез с использованием F-критерия Фишера.
5. Практическое применение однофакторного дисперсионного анализа.
6. Методы выявления несущественных факторов.
7. Статистическая проверка значимости влияния факторов на результаты эксперимента.
8. Понятие корреляционного анализа и его основные задачи.
9. Графическое исследование взаимосвязи случайных величин.
10. Коэффициент корреляции Пирсона и его интерпретация.
11. Оценка точности и достоверности коэффициента корреляции.
12. Практическое применение корреляционного анализа при обработке результатов модельных экспериментов.

Однофакторный дисперсионный анализ

В современной жизни - военной и гражданской часто возникают проблемы, решение которых требует научного обоснования.

Однотипны ли патроны для конкретного образца стрелкового вооружения, выпускаемые на разных заводах?

Однородны ли, например, автоматы, выпускаемые на разных заводах?

Здесь в качестве исследуемого фактора выступают заводы. Разные, но однотипные по назначению - варианты фактора, которые можно трактовать как уровни факторов.

Аналогичная задача возникает при сравнении однотипных изделий, вырабатываемых с применением различных технологий. Здесь подлежит анализу фактор - технология производства.

Эти и подобные задачи являются задачами однофакторного дисперсионного анализа (ОДА).

Иногда возможны задачи одновременного исследования влияния двух и более факторов. Например, чем объяснить рассеивание попаданий в цель: конструкторскими особенностями стрелкового вооружения, выпущенного на разных заводах, или различиями в подготовке стрелков?

Исследованием влияния факторов и занимается *факторный дисперсионный анализ*.

Мы рассмотрим ОДА, наиболее актуальный анализ на практике. Теория рассматривает и многофакторный дисперсионный анализ. В нем процедуры подобны тем, которые мы рассмотрим в ОДА. Усложняются только расчеты и при необходимости, зная ОДА, овладеть методикой многофакторного дисперсионного анализа, не составит труда.

Эксперимент для выполнения ОДА состоит в накоплении результатов измерений контролируемого параметра (угла, расстояния, наработки на отказ некоторого изделия и т. д.) при каждом варианте исследуемого фактора.

Введем обозначения:

m - число вариантов фактора;

n - число измерений при каждом варианте;

a_{ij} - результат каждого измерения;

$i = \overline{1,m}$ - номер варианта фактора;

$j = \overline{1,n}$- 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. Ход дальнейших рассуждений от этого не меняется.

Результаты эксперимента сводятся в таблицу (табл. 1).

Вопрос: влияют ли варианты фактора на точность измерений? Или, говоря языком математической статистики, являются результаты $m \cdot n$ 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.

Таблица 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}	$\overline{a_m} = \frac{1}{n}\sum_{j=1}^{N} a_{mj}$

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:

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

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

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

Эту сумму квадратов отклонений можно разложить на два независимых слагаемых:

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

Обозначим:

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

Что такое Q_1 и Q_2 ?

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 дисперсий σ^2 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 *S*.

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/2находим F_2 .

Если окажется $F > F_2$, то есть мы попали в область маловероятных значений *F*, то выдвинутая гипотеза *не подтверждается*. А это значит, что варианты фактора не однотипны. Но если $F \le F_2$, то гипотеза об однородности вариантов фактора подтверждается, конечно, в рамках допустимого риска.

Пример. Необходимо проверить однотипность патронов к автомату Калашникова, изготовленных на трех заводах.

Для получения необходимых для дисперсионного анализа данных автомат закрепили в специальном станке и сделали из него по 50 выстрелов патронами каждого завода. По результатам стрельбы измерялись радиальные отклонения пробоин от точки прицеливания.

Результаты измерений приведены в табл. 2.

Таблица 2. Результаты стрельбы

Заводы	Эксперименты и отклонения, см								
	1	2	3	…	26	27	…	49	50
№ 1	3	2	1	…	4	3	…	1	3
№ 2	2	0	4	…	3	2	…	2	3
№ 3	2	3	3	…	1	0	…	1	5

Решение

Проверяем исходную гипотезу: патроны, выпускаемые на трех разных заводах, баллистически однотипны.

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=50.

При вычислениях опустим очевидные элементарные арифметические детали.

Средние отклонения пробоин при стрельбе патронами заводов № 1, № 2, № 3 равны соответственно:

$$\overline{a_1} = \frac{3+2+1+\cdots+4+3+\cdots+1+3}{50} = 2{,}5 \text{ см},$$

$$\overline{a_2} = \frac{2+0+4+\cdots+3+2+\cdots+2+3}{50} = 2{,}8 \text{ см},$$

$$\overline{a_3} = \frac{2+3+4+\cdots+1+0+\cdots+1+5}{50} = 3{,}1 \text{ см},$$

Среднее отклонение по 150 выстрелам:

$$\bar{a} = \frac{2{,}5+2{,}8+3{,}1}{3} = 2{,}8 \text{ см}.$$

Средний квадрат расхождений между вариантами факторов:

$$\frac{1}{m-1}Q_1 = \frac{1}{m-1}n\sum_{i=1}^{n}(\overline{a_i}-\bar{a})^2 = \frac{50}{2}[(2{,}5-2{,}8)^2+(2{,}8-2{,}8)^2+(3{,}1-2{,}8)^2]$$
$$= 4{,}5\text{см}^2$$

Число степеней свободы: m – 1 =3 – 1=2.

Средний квадрат расхождений внутри вариантов:

$$\frac{1}{m(n-1)}Q_2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(a_{ij}-\overline{a_i}\right)^2$$
$$= \frac{1}{3.49}[(3-2{,}5)^2+(1-2{,}5)^2+\cdots+(4-2{,}5)^2+(3-2{,}5)^2+(2-2{,}8)^2$$
$$+\cdots+(5-3{,}1)^2] = 2{,}2\text{см}^2$$

Число степеней свободы: $m(n-1)$=3 · 49 = 147

Расчет F -критерия: $F = \frac{1.5}{2.2} \approx 2.1$

По табл. П.2 при $\frac{q}{2} = 5$ верхних пределов уклонения величины F и имеющихся степенях свободы 2 и 147 находим $F_2 = 3.06$. Величина F_2 определена при $R_2 = 150$, так как табличные значения при $k_2 = 147$ не определены. Нетрудно убедиться, что такое приближение вполне допустимо.

Поскольку 2.1 < 3.06, делаем вывод о том, что выдвинутая гипотеза об однотипности партий патронов, выпускаемых тремя разными заводами, не опровергается (в пределах принятого уровня значимости).

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:

$$y_1', y_2', y_3', \dots, y_i', \dots, y_N'; y_1'', y_2'', y_3'', \dots, y_i'', \dots, y_N''.$$

Согласно принятой гипотезе эти последовательности имеют одинаковые матожидания *M[y]* и дисперсии σ_y^2.

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:

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

Очевидно:

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

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

Такое разделение осуществляется по следующему правилу: если вычисленная величина $\bar{z}$ окажется маловероятной, в рамках нормального распределения и данном среднеквадратическом отклонении σ_z, то такое отклонение $\bar{z}$ 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 $f(\bar{z})$ уровень значимости q показан заштрихованным участком (рис. 1).

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

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

где $\Phi(t_\alpha)$ - функция Лапласа.

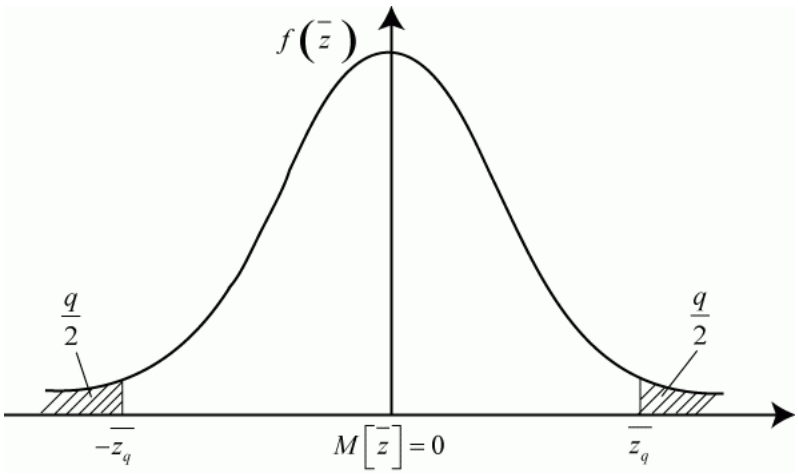

Рис.1. Плотность распределения функции *f(z)*

Следовательно:

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

• 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 $\bar{z} > t_\alpha \frac{\sigma_y\sqrt{2}}{\sqrt{N}}$, принятая гипотеза о несущественности проверяемого фактора не подтверждается;

• если $\bar{z} \le t_\alpha \frac{\sigma_y\sqrt{2}}{\sqrt{N}}$, принятая гипотеза не опровергается (в рамках принятого уровня значимости q).

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

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

Оценку $\bar{y}$ 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.

Решение

Выдвигается гипотеза о несущественности влияния на время пребывания заявки в системе массового обслуживания изменения дисциплины FIFO на LIFO.

Результаты измерений и вычислений:

• $\bar{z} = 1{,}8\text{мин}, S_y = 2$ мин;

• для уровня значимости $q = 5, t_\alpha = 1.96$

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

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

В заключение отметим, что рассмотренную проблему можно решать и методом однофакторного анализа. Однако если $m = 2$ (при сравнении двух выборок) целесообразно использовать изложенный метод.

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. В рамках этого учебного пособия мы ограничимся рассмотрением гипотезы о наличии линейной корреляции.

Название "корреляционный анализ" происходит от латинского слова correlatio - согласование, связь, соотношение, взаимосвязь. Термин впервые введен Гальтоном (Galton) в 1888 г.

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*. Например:

a	*b*
Средний балл успеваемости учебной группы по математике	Средний балл выполнения упражнения по стрельбе
Рассеивание точки падения заряда по дальности	Рассеивание точки падения заряда по боковому отклонению
Вес курсантов (студентов).	Успеваемость по физподготовке.

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:

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

Каждая пара интерпретируется как координаты случайной точки в системе координат *a*, *b*.

Первичное исследование можно провести графически. Возможны следующие варианты размещения точек на графиках (рис. 2).

Корреляция - важное понятие. Научитесь 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 (рис. 5б).

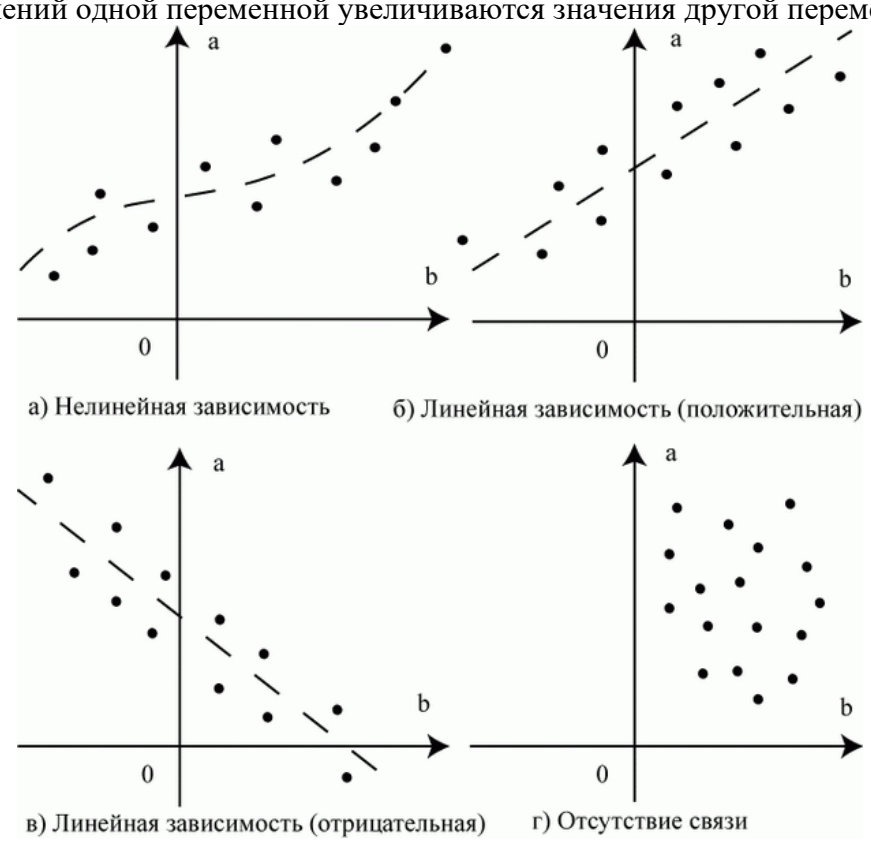

Рис. 2. Графическое исследование корреляции

Две переменные *отрицательно* коррелированны, если при увеличении одной переменной другая переменная уменьшается (рис. 5в).

Отсутствие корреляции - совместного поведения переменных - обнаруживается хаотическим нагромождением точек, исключающим проведение какой-либо аппроксимирующей линии (см. рис. 5г).

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

Если совместное распределение вероятностей случайных величин *a* и *b* нормальное, то количественной характеристикой степени линейной связи между ними является коэффициент корреляции *r* (введен Пирсоном (Pearson), 1896 г.):

$$-1 \le r \le 1.$$

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

Равенство $r = \pm 1$ 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 $\bar{r}$. В качестве оценки $\bar{r}$ принят выборочный коэффициент корреляции:

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

где $\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i$ оценки математических ожиданий и *M[a]* и *M[b];*

$S_a = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(a_i - \bar{a})^2}$, $S_b = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(b_i - \bar{b})^2}$ - оценки среднеквадратических отклонений σ_a и σ_b

Выборочный коэффициент корреляции $\bar{r}$, так же как и теоретический, принимает значения: $-1 \le \bar{r} \le 1$.

Если $\bar{r} > 0$, то наблюдается положительная корреляция (см. рис. 5б). Если $\bar{r} < 0$ - отрицательная корреляция (см. рис. 5в). Если $\bar{r} = 0$ - линейная корреляция отсутствует (но не исключена нелинейная). Если $\bar{r} = 1$, то между случайными величинами существует жесткая функциональная связь.

Заметим, что рассматриваемый коэффициент корреляции $\bar{r}$ определяет степень *линейной* связи между случайными величинами *a* и *b*. Эта корреляция наиболее популярна, поэтому часто, когда говорят о корреляции, имеют в виду именно корреляцию Пирсона.

Однако этот линейный коэффициент корреляции не является пригодным для оценки нелинейной связи, если таковая присутствует. При нелинейной зависимости степень связи между случайными величинами устанавливается более сложными характеристиками, например, *корреляционным отношением* (К. Пирсон).

Если случайные величины *a* и *b* независимы, они и не коррелированны ($\bar{r} = 0$). Но некоррелированность *a* и *b* не всегда свидетельствует об их независимости. Но если *a* и *b* имеют нормальное распределение, то условие $\bar{r} = 0$ является необходимым и достаточным условием независимости этих величин.

И еще. Наличие корреляции между случайными величинами *a* и *b* не всегда свидетельствует об их взаимосвязи. Дело в том, что при независимости *a* и *b* каждая из них в отдельности зависит от некоторого случайного фактора ξ, но эта зависимость нами не замечена.

Поэтому хорошим тоном после вычисления корреляций является построение диаграмм рассеяния, которые позволяют понять, действительно ли между двумя исследуемыми переменными имеется связь.

Оценка коэффициента корреляции должна быть определена с требуемыми точностью и достоверностью, которые зависят от числа реализаций модели. Найдем эту связь.

В предположении нормальности распределения $\bar{r}$ можно написать:

$P(|\bar{r} - r| < t_\alpha \sigma_{\bar{r}}) = 2\Phi(\mathrm{t}_\alpha)$ (1)

С выражение (1) мы уже знакомы. Здесь:

r - точное значение коэффициента корреляции;

$\sigma_{\bar{r}}$ - среднеквадратическое отклонение случайной величины $\bar{r}$;

t_α - аргумент функции Лапласа $\Phi(t_\alpha)$.

Обычно среднеквадратическое отклонение $\sigma_{\bar{r}}$ неизвестно, поэтому нужно брать ее оценку.

При больших выборках *N* оценка среднеквадратического отклонения S_r:

$$\sigma_{\bar{r}} \approx S_r = \frac{1 - \bar{r}^2}{\sqrt{N}}$$

Из (1) следует:

$$\varepsilon = t_\alpha \frac{1 - \bar{r}^2}{\sqrt{N}}, N = t_\alpha^2 \frac{(1 - \bar{r}^2)^2}{\varepsilon^2}, t_\alpha = \Phi^{-1}\left(\frac{\alpha}{2}\right)$$

ε - абсолютная величина ошибки.

Предварительное определение $\bar{r}$ осуществляется по данным пробного эксперимента в количестве $N^* = 500 \ldots 1000$ реализаций модели.

На основании изложенного и в силу случайного характера исследуемых величин *a* и *b* мы можем утверждать лишь следующее: истинное значение коэффициента корреляции *r* лежит в пределах

$$\bar{r} - \varepsilon \le r \le \bar{r} + \varepsilon$$

с заданной достоверностью α.

В заключение отметим, что если совместное распределение случайных величин *a* и *b* не является нормальным, то оценка $\bar{r}$ коэффициента корреляции может выступать в качестве ориентировочной оценки степени тесноты связи *a* и *b*.

Пример. Для оценки конструкции нового крупнокалиберного пулемета было произведено 96 выстрелов по щиту, отстоявшему на расстоянии 300 метров.

Результаты отклонений попаданий от точки прицеливания (боковые x_i, по высоте y_i) объединены в десятисантиметровые диапазоны и сведены в таблицу (табл. 3).

Для оценки конструктивных особенностей пулемета необходимо узнать: есть ли какая-то связь между боковыми отклонениями и отклонениями по высоте.

Решение

Ответ на поставленный вопрос может дать коэффициент корреляции. Предварительно заметим, что группировка измерений в десятисантиметровые диапазоны вносит некоторую ошибку в дальнейшие расчеты, однако можно показать, что при данной группировке ошибка несущественна.

В табл. 3 указаны не реальные отклонения, а центры диапазонов (-25…-15, -15…-5, -5…5 и т. д.).

Таблица 3. Отклонения от точки прицеливания

y_i	Боковые отклонения x_i							Всего n_{y_j}
	-20	-10	0	10	20	30	40	
-50	0	0	1	0	2	0	0	3
-40	0	1	1	1	2	0	0	5
-30	1	1	3	5	2	1	0	13
-20	1	3	7	3	2	2	0	18
-10	0	2	6	10	3	0	0	21
0	0	1	6	6	6	1	1	21
10	0	0	3	3	3	1	0	10
20	0	1	1	2	1	0	0	5
Всего n_{x_j}	2	9	28	30	21	5	1	96

Для определения коэффициента корреляции понадобятся следующие характеристики: $\bar{x}, S_x, \bar{y}, S_y$, ковариация $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$.

Все эти характеристики вычисляются по данным измеренных отклонений боковых *x* и по высоте *y*.

Для примера, расчет $\bar{x}$:

$$\bar{x} = \frac{(-20 \cdot 2) + (-10 \cdot 9) + (0 \cdot 28) + \cdots + (40 \cdot 1)}{96} = \frac{780}{96} \approx 8.1\text{см}.$$

Результаты расчета остальных характеристик:

$$S_x = 11.6\text{см}, y = -11.6\text{см}, S_y 16.8\text{см}, = cov(x, y) = 1518\text{см}^2$$

Теперь оценка коэффициента корреляции:

$$\bar{r} = \frac{1518}{96 \cdot 11.6 \cdot 16.8} = \frac{1518}{18525} = 0.082$$

Среднеквадратическое отклонение этой оценки:

$$S_{\bar{r}} = \frac{1 - \bar{r}^2}{\sqrt{N}} = \frac{1 - 0.082^2}{\sqrt{96}} \approx 0.1\text{см}$$

Из-за малого количества выстрелов оценка $\bar{r}$ определена с ошибкой, которая в предположении о нормальном распределении случайной величины $\bar{r}$ и достоверности, например, $\alpha = 0.8$ ($t_\alpha = 1.28$) равна:

$$\varepsilon = t_\alpha \sigma_{\bar{r}} \approx t_\alpha S_{\bar{r}} = 1.28 - 0.1 = 0.128\text{см}$$

Отсюда следует, что истинное значение коэффициента корреляции r лежит в пределах:

$$0.082 - 0.128 \le r \le 0.082 + 0.128, -0.046 \le r < 0.21.$$

Обнаружена небольшая линейная зависимость отклонений боковых и по высоте. Баллистики, отвергая непосредственную корреляцию между отклонениями x и y, объясняют значение $\bar{r} > 0$ влиянием конструктивных особенностей пулемета. Обнаружена также систематическая ошибка в прицеле: $\bar{x} \neq 0, \bar{y} \neq 0$.

Контрольные вопросы

1. Что понимается под факторным дисперсионным анализом?
2. Для каких целей применяется однофакторный дисперсионный анализ?
3. Что такое фактор и уровень фактора в экспериментальном исследовании?
4. Как организуется эксперимент для проведения однофакторного дисперсионного анализа?
5. Какие основные предположения используются при однофакторном дисперсионном анализе?
6. Что представляет собой нулевая гипотеза при исследовании влияния фактора?
7. Каким образом общая сумма квадратов отклонений разделяется на межгрупповую и внутригрупповую составляющие?
8. Что характеризует межгрупповая дисперсия?
9. Что характеризует внутригрупповая дисперсия?
10. Как используется F-критерий Фишера для проверки значимости влияния фактора?
11. Что такое уровень значимости и как он влияет на принятие статистического решения?
12. В чем заключается метод выявления несущественных факторов?
13. Почему исключение несущественных факторов может повысить эффективность модельного эксперимента?
14. Как осуществляется сравнение результатов двух экспериментов при наличии и отсутствии исследуемого фактора?
15. Как используются нормальное распределение и распределение Стьюдента при проверке статистических гипотез?
16. Что понимается под корреляционным анализом?
17. Какие основные задачи решаются с помощью корреляционного анализа?
18. Что такое парная корреляция?
19. Как с помощью диаграммы рассеяния определить характер взаимосвязи между случайными величинами?
20. Чем отличаются положительная и отрицательная корреляции?
21. Что характеризует коэффициент корреляции Пирсона?
22. Как интерпретируются значения коэффициента корреляции, близкие к −1, 0 и 1?

Список литературы

1. Гмурман В. Е. Теория вероятностей и математическая статистика: учебник для прикладного бакалавриата. — 12-е изд. — М.: Юрайт, 2014. — 479 с.
2. Montgomery D. C. Design and Analysis of Experiments. — 10th ed. — Wiley, 2019. — 688 p.
3. Kutner M. H., Nachtsheim C. J., Neter J., Li W. Applied Linear Statistical Models. — 5th ed. — McGraw-Hill/Irwin, 2005. — 1396 p.
4. Draper N. R., Smith H. Applied Regression Analysis. — 3rd ed. — John Wiley & Sons, 1998. — 736 p.