# КУБЛАНОВ
# Михаил Семенович

**103А, 207А**

Моя цель – передать свой опыт,

а не отчитать часы!

“Говорят, что если умыть кошку, то она потом уже никогда не будет умываться сама. Не знаю, правда это или нет, но несомненно одно: если человека чему-нибудь учить, он этому никогда не выучится.”

Б. Шоу

***Назначение курса:***

– усвоение основных **методологических требований** и приемов научных исследований с помощью ММ;

– прививка **математической строгости** и получение на этой основе **иммунитета** от ошибок при формулировке и решении производственных и исследовательских задач.

***Особенности курса:***

– первая дисциплина в процессе Вашего обучения, которая учит правильно организовывать и проводить научную и инженерную **исследовательскую работу**;

– **математическая строгость** к постановке задач и методам их решения;

– наличие **философской составляющей** в теории математического моделирования;

– упор на **самостоятельную работу** студента;

– возможность **сэкономить** до 50 % времени при работе в процессе семестра, а не только в сессию.

## Состав курса (160901, 160900)

- 22 лекции
- 3 лабораторные работы на компьютерах

  (3-я - репетиция зачета)
- КДЗ - РГР (3 задачи)
- зачет на компьютерах

ОБ ЭКЗАМЕНЕ (ЗАЧЕТЕ) по циклу моделирования

1. Экзамен (зачет) проводится на компьютерах в тестовой форме.
2. Разрешается пользоваться всем, чем угодно, **кроме соседа**.
3. Сдавать можно много раз.

4. Это – не игральный автомат, получить случайно удовлетворительную оценку здесь невозможно!
5. Это – **объективная** система оценки уровня знаний, так как вопросы составлены не на выбор, а на **понимание**!

6. Время ограничено таким образом, что необходимое количество правильных ответов можно успеть набрать, только имея определенный **уровень** знаний, позволяющий **свободно** (без использования учебных материалов) обращаться с достаточным количеством вопросов.

7. Вопросы сформулированы в тестовой форме – точное их содержание в учебном пособии не содержится. Для ответа необходимо задействовать свои **умственные способности**!

8. Высшее образование – не просто сумма запомненных фактов! Образование – это **понимание**, из каких фактов какими логическими путями, что следует. Такое понимание позволяет не только решать задачи школьного уровня (по предоставленному шаблону), но и ставить и решать новые задачи.

9. Поэтому все возможные вопросы на контрольных мероприятиях в вузе можно условно разбить на 3 группы:

– определения и классификации (необходимо правильно запомнить);

– фундаментальные понятия и их логические связи (надо понять);

– методы (надо "попробовать").

10.Оценка недостаточного уровня знаний по статистике дает точную и **объективную** характеристику неподготовленности (количество правильных ответов при требуемых 15 из 27):

0 – 2 – термины понимаются неверно,

3 – 4 – случайный набор ответов,

5 – 6 – что-то где-то слышал,

7 – 9 – отдельные фрагменты читал бессистемно,

10 – 14 – занимался, но бессистемно, есть существенные пробелы.

11.При условии максимальной сосредоточенности экзаменующегося разброс оценок в 90 % случаев составляет ± 1 вопрос (из 27).

12.О распространяемых среди студентов за деньги (которых они не сто́ят) списков вопросов с ответами:

– они далеко не полные, содержат ошибки, а база вопросов постоянно корректируется;

– бездумное отыскивание по алфавитному признаку нужного ответа приводит к уровню, недостаточному для получения удовлетворительной оценки (~ 9 вопросов из 27);

– если работать за компьютером с этими списками **головой**, то можно за множество попыток достичь необходимого уровня оценки; однако такой способ **обучения** требует втрое больше времени, чем традиционный – с помощью чтения учебного пособия.

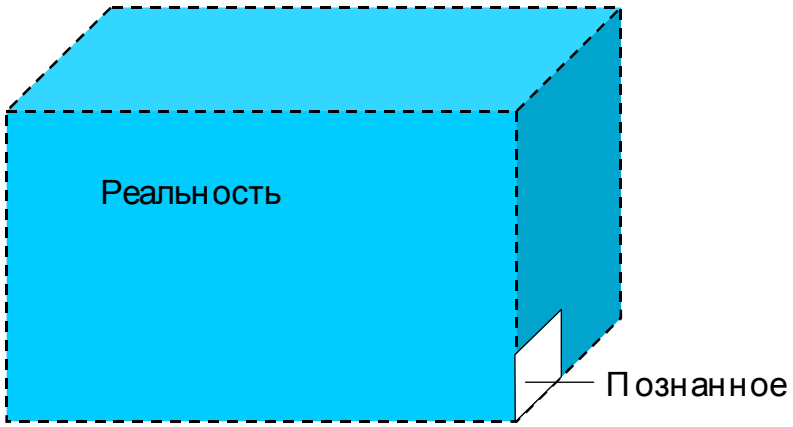
# ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА И ОБРАБОТКА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

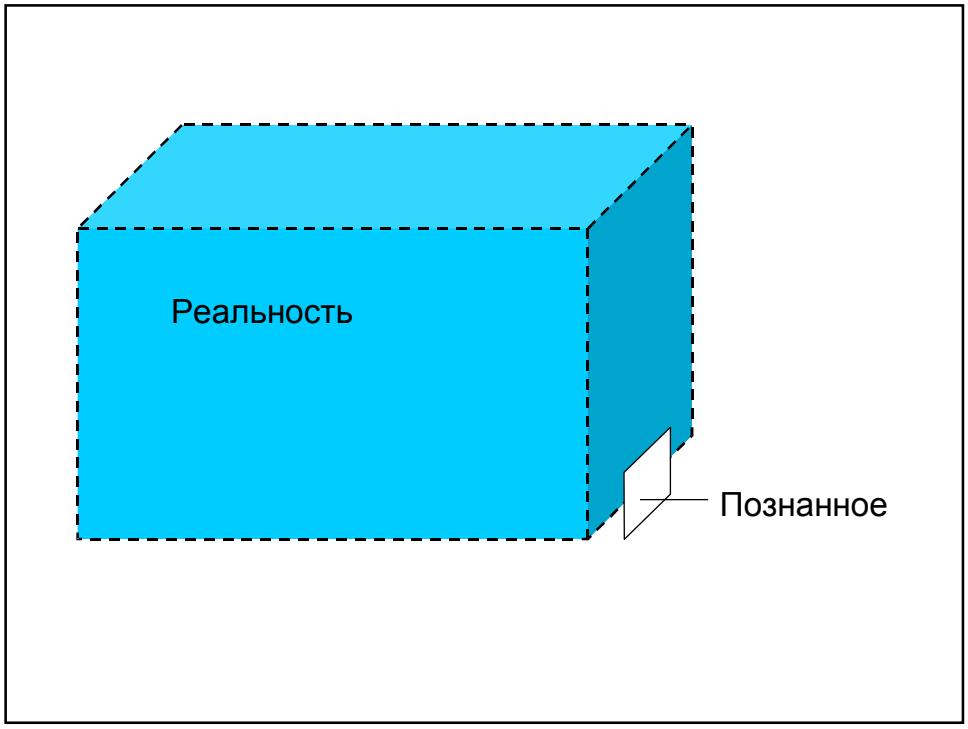
## Литература

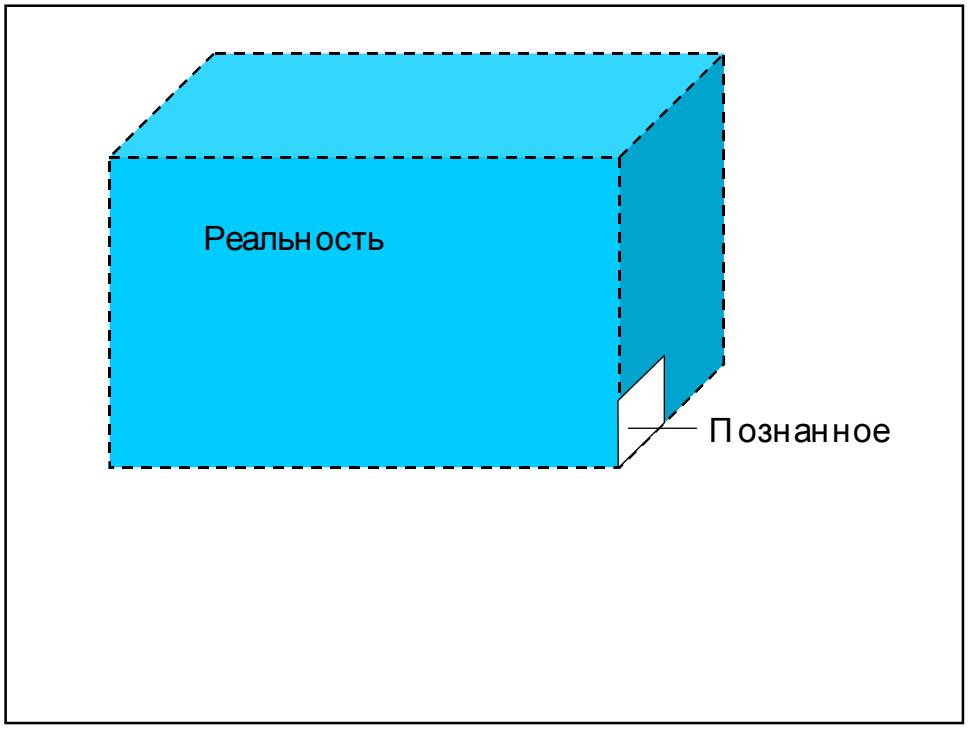
| | |
|---|---|
| Математическое моделирование:<br>Учебное пособие. Ч I, II. 2004. | 517.8<br>К88 |
| | |
| методичка по изучению дисциплины (2005 г.) | № 429 |

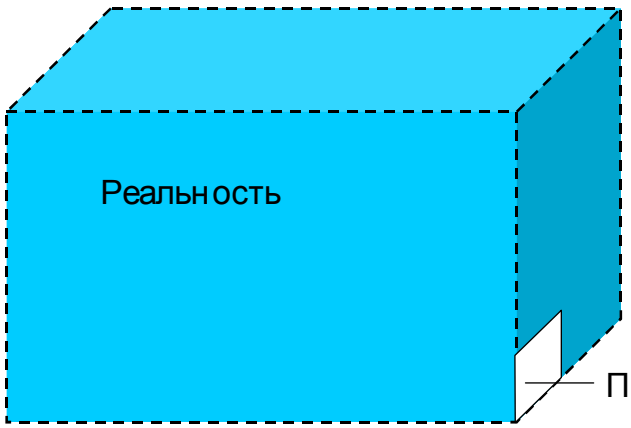
# Введение

[Часть I, стр. 6 - 8]



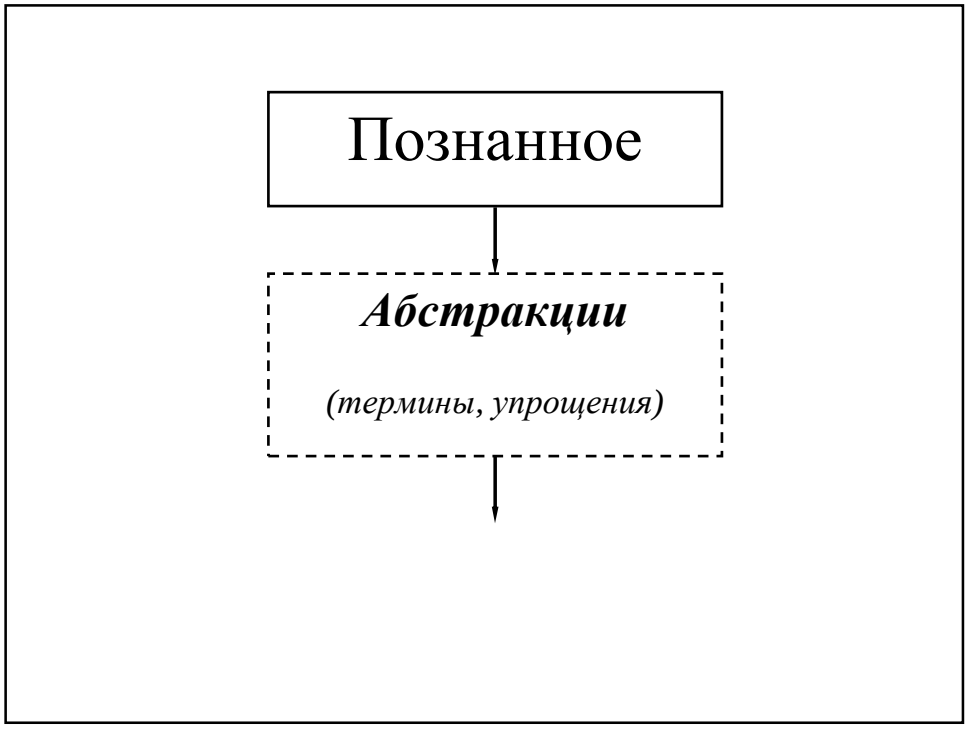

Реальность
Познанное

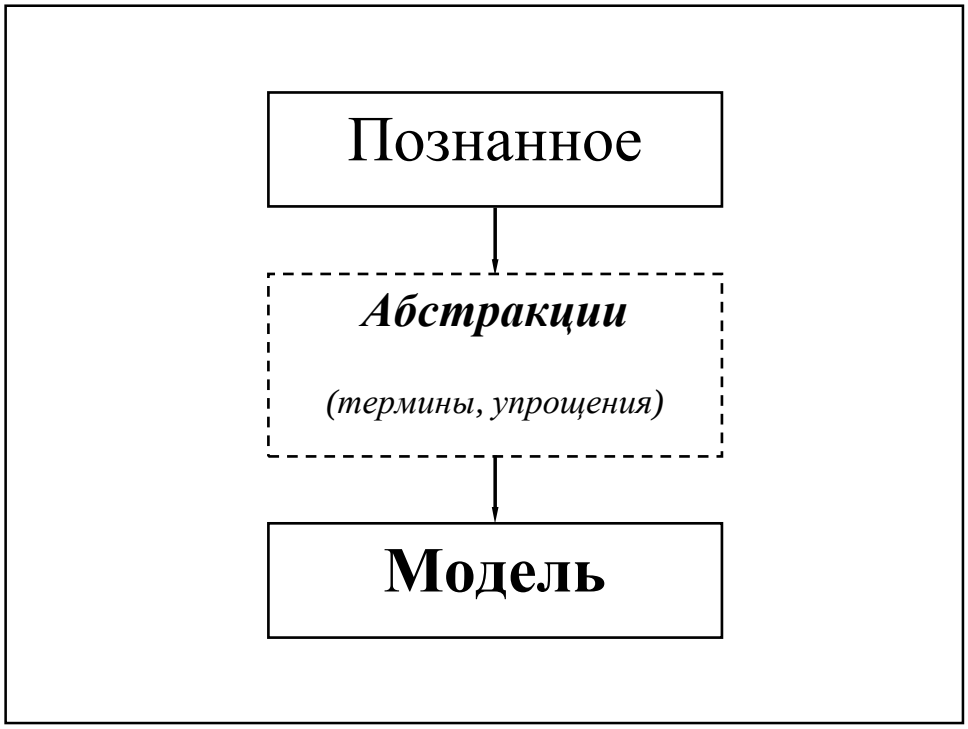

Реальность
Познанное

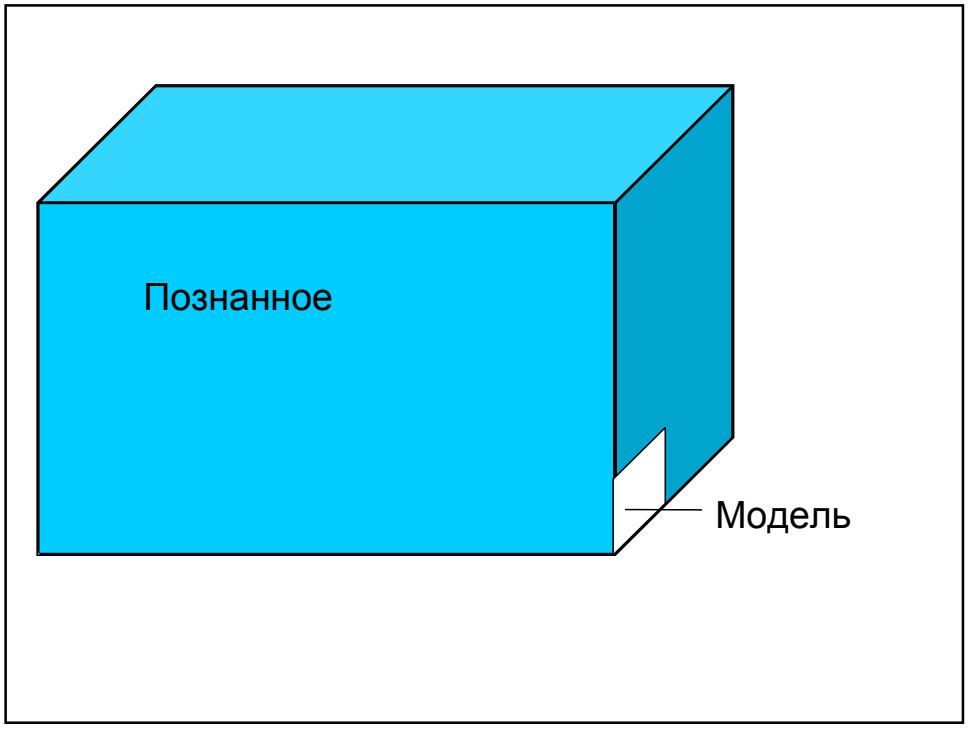


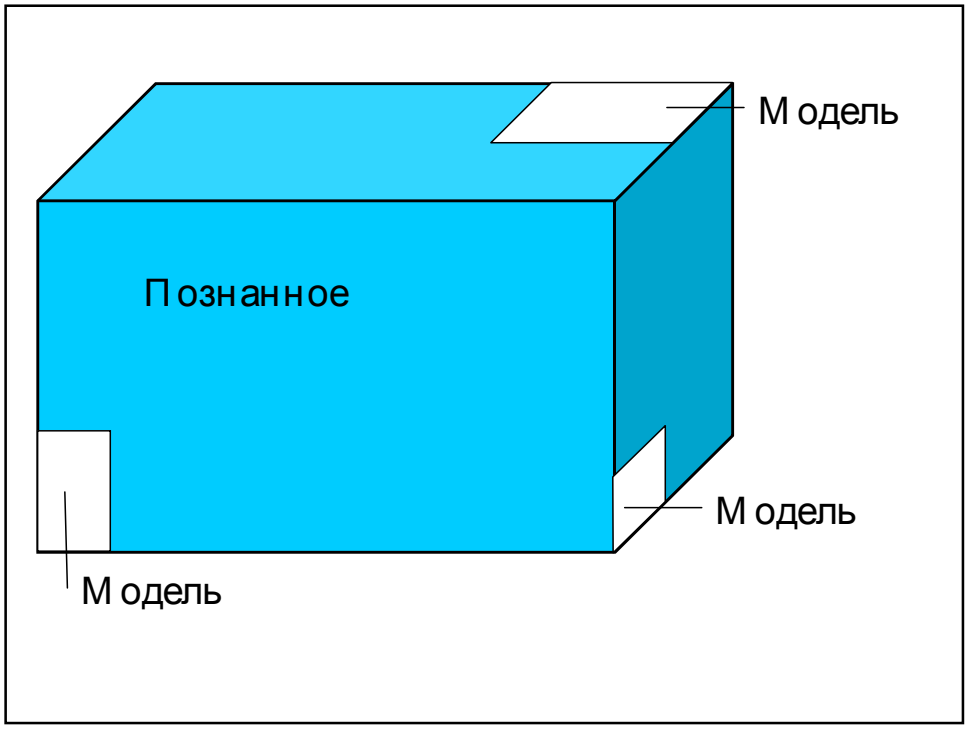

Реальность:
- истина,
- генеральная совокупность,
- оригинал.

Познанное:
- эксперимент,
- выборка,
- система.


Познанное
Абстракции
(термины, упрощения)


Познанное
Абстракции
(термины, упрощения)
Модель


Познанное
Модель


Модель
Познанное
Модель
Модель

«Хорошо организованные системы» - явления и объекты, достаточно точно и однозначно описываемые **небольшим** количеством факторов

«Плохо организованные системы» - сложные системы, в которых нельзя разделить отдельные явления

| ХОРОШО ОРГАНИЗОВАННЫЕ СИСТЕМЫ | ПЛОХО ОРГАНИЗОВАННЫЕ СИСТЕМЫ |
| --- | --- |
| ЗАКОНЫ | ЗАКОНОМЕРНОСТИ |
| • классической механики,<br>• генетики,<br>• химии... | • поляра ЛА,<br>• инфляционные ожидания,<br>• надежность... |

Целью научных исследований является познание **законов** природы.

Целью инженерных исследований следует считать познание **закономерностей**, свойственных продуктам человеческой деятельности.

Выявление закономерностей,
т.е. эксперимент нуждается
не только в четкой формулировке
**цели исследования**,
но и в знании основных свойств
оригинала,
что невозможно без его
**модели**

**Цель исследования**

↓

**Сбор и обработка информации**

↓

**Модель**

↓

**Планирование эксперимента**

↓

**Эксперимент**

↓

**Сбор и обработка информации**

# МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

## Основы теории вероятностей и математической статистики

Отбор информации

[Часть II, стр. 12 - 14]

Математическая статистика - естественный и единственный аппарат сбора и обработки информации

Но!
Математическая статистика может обработать только **собранную** информацию и дать результаты с определенной **вероятностью**

Нет и не может быть математической обработки информации **вообще**,
есть только аппарат для **целевых исследований**,
невозможных без предварительных предположений о **модели**

## Проблемы
## *сбора и обработки информации*:

– выбор существенных факторов;
– выбор процедуры отбора информации;
– обеспечение достоверности выводов по результатам анализа.

Социологический опрос.
Результаты ответов 50 человек на некоторый вопрос представлены рядом знаков «+» – «да» и «–» – «нет».

++–++++–+–+––––++–+–+––+–+–++++––––+–+++–++++–+–+–++

«+» встречается 29 раз, «–» – 21 раз.

Т.е. ответов «да» на 38 % больше, чем «нет».

Вывод очевиден?

Социологический опрос.
Результаты ответов 50 человек на некоторый вопрос представлены рядом знаков «+» – «да» и «–» – «нет».

++-++++-+-+---++-+-+--+-+++++---+-+++-+++-+-+-++

«+» встречается 29 раз, «–» – 21 раз.

Т.е. ответов «да» на 38 % больше, чем «нет».

Вывод очевиден?

Но баланс знаков, стоящих на четных местах: «+» – 18, «–» – 7; а на нечетных: «+» – 11, «–» – 14.

Насколько можно доверять общим итогам опроса в таком случае?

Отбор информации - важнейшая операция, от правильного проведения которой зависит результат статистического анализа, а, следовательно, и выводы исследования

Отбор информации не объективен!

1. Результаты наблюдений - это лишь ограниченная выборка.
2. Информация собирается для определенных целей.
3. Результаты наблюдений имеют погрешность.

Естественный отбор предполагает получение информации в виде **констатации** определенных событий, процесс которой (констатации) не зависит от исследователя.
Искусственный отбор всегда подчинен воле исследователя.

Виды искусственного отбора:

- пристрастный -

осуществляется по заранее намеченному признаку;

- случайный -

производится с помощью случайных чисел по любой методике;

- механический -

отбор данных из всей совокупности по какому-либо правилу

(например, каждый пятый);

- типический -

отбор из слоев (частей) всей имеющейся совокупности;

- аритмический -

частный случай *типического* и *механического*, когда отбор производится из равных групп по определенному правилу;

- пропорциональный -

частный случай *типического* отбора, когда из каждого слоя отбирается часть, пропорциональная его объему;

- репрезентативный -

(идеальный случай), при котором получается представительная выборка, достаточно **полно** характеризующая **всю** совокупность;

- расслоенный случайный - комбинация *типического* и *случайного* отбора, при которой из отдельных слоев (групп, частей) отбор осуществляется случайным образом...

# Основные термины теории вероятностей и математической статистики

[Часть II, стр. 5 - 12]

Теория вероятностей – наука, изучающая закономерности в случайных явлениях

Случайное явление – явление, которое при неоднократном воспроизведении одного и того же опыта протекает по-разному

Событие – всякий факт, который в результате опыта может произойти или не произойти

Если некоторое событие заведомо не может произойти, то такое событие называется невозможным

Если некоторое событие обязательно происходит, то такое событие называется достоверным

Два события называются несовместными, если их одновременное (совместное) наступление невозможно

Несколько событий образуют полную группу событий, если обязательно происходит хотя бы одно из них, т.е. никаких других, неучтенных, событий быть не может

Результаты повторения одного и того же опыта называются *исходами* (или *элементарными событиями*)

Идеализированная система исходов:
– число исходов конечно;
– все исходы образуют полную группу событий;
– все исходы попарно несовместны;
– все исходы равновозможны.

Классическое определение вероятности:
вероятность P(A) случайного события A
– это **числовая характеристика**
возможности этого события
$P(A) = m/n$,
где n – число **всех** *равновозможных*,
*несовместных* исходов, образующих
*полную группу событий*,
m – число исходов, **благоприятных**
появлению события A

Основные свойства вероятности:
$0 \le p \le 1$;
для невозможного события $p = 0$;
для достоверного события $p = 1$.

Случайной величиной называют величину, которая в результате опыта может принять только **одно из множества** возможных значений, заранее не известно какое

Дискретные *случайные величины* принимают отдельные, изолированные перечисляемые значения.

Непрерывные *случайные величины* заполняют своими возможными значениями некоторый промежуток числовой оси.

Закон распределения случайной величины ставит в соответствие каждому значению случайной величины вероятность именно его появления

Для **дискретной** случайной величины (например, число очков, выпавшее на игральной кости) *закон распределения* задается таблицей соответствия возможных значений и вероятностей их появления

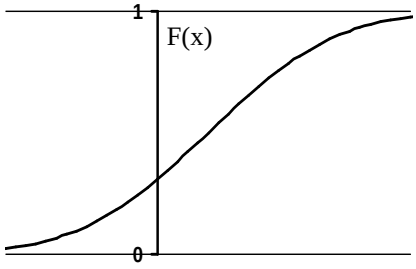
| Закон распределения | | | | |
|---|---|---|---|---|
| $\xi_i$ | $\xi_1$ | $\xi_2$ | ... | $\xi_M$ |
| $P(\xi_i)$ | $p_1$ | $p_2$ | ... | $p_M$ |

| Для игральной кости | | | | | | |
|---|---|---|---|---|---|---|
| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $p_i$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

## Для **непрерывной** случайной величины

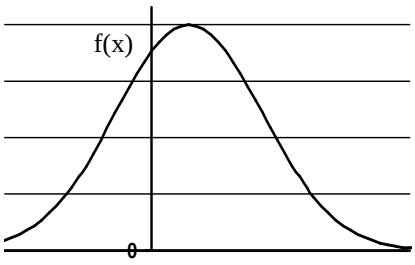
Интегральная функция распределения вероятностей

**F(x) = P(ξ<x)**

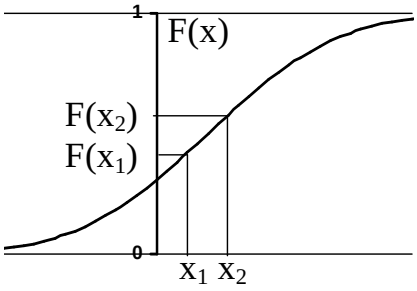


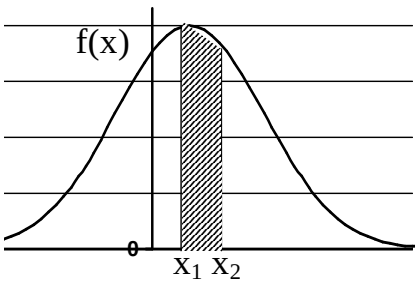

Дифференциальная функция распределения вероятностей

**f(x) = F ' (x)**




Для любой **непрерывной** случайной величины вероятность попадания ее значения в **интервал** от $x_1$ до $x_2$: $P(x_1 < \xi < x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$



Числовые характеристики
законов распределения
случайных величин:

Математическое ожидание

для дискретной случайной величины:

$$a \equiv E(\xi) = \sum_{i=1}^{M} \xi_i p_i, \quad \text{или для } j-\text{го слоя: } a_j = \sum_{i=1}^{M_j} \xi_{ji} p_{ji}$$

для непрерывной случайной величины:

$$a \equiv E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$$

Дисперсия $\sigma^2$
и среднее квадратическое отклонение $\sigma$

для дискретной случайной величины:

$$\sigma^2 \equiv D(\xi) \equiv E(\xi - a)^2 = \sum_{i=1}^{M} (\xi_i - a)^2 p_i, \qquad \sigma_j^2 = \sum_{i=1}^{M_j} (\xi_{ji} - a_j)^2 p_{ji}$$

для непрерывной случайной величины:

$$\sigma^2 \equiv D(\xi) = \int_{-\infty}^{\infty} (x - a)^2 \cdot f(x) \cdot dx$$

Медиана $\tilde{x}$: $P(\xi < \tilde{x}) = P(\xi > \tilde{x})$ – такое значение, для которого одинаково вероятно, окажется ли случайная величина меньше или больше $\tilde{x}$

Мода $x_M$: $P(x_M) = \max\{P(x)\}$ – наиболее вероятное значение случайной величины

Размах $R = x_{max} - x_{min}$ – разность между наибольшим и наименьшим из возможных значений случайной величины

СИСТЕМА ОБОЗНАЧЕНИЙ

Простой случайный отбор

| элементы генеральной совокупности | генеральные | | | элементы выборки | выборочные оценки | | |
|---|---|---|---|---|---|---|---|
| | объем | средняя | дисперсия | | объем | средняя | дисперсия |
| $\xi_1,...,\xi_M$ | M | a | D, $\sigma^2$ | $x_1,...,x_N$ | N | $\bar{x}$ | $D_B$, $s^2$ |

Расслоенный отбор

| № слоя | элементы генеральной совокупности | генеральные | | | элементы выборки | выборочные оценки | | |
|---|---|---|---|---|---|---|---|---|
| | | объем | средняя | дисперсия | | объем | средняя | дисперсия |
| 1 | $\xi_{11},...,\xi_{1M_1}$ | $M_1$ | $a_1$ | $D_1, \sigma_1^2$ | $x_{11},...,x_{1M_1}$ | $N_1$ | $\bar{x}_1$ | $D_{в1}$, $s_1^2$ |
| 2 | $\xi_{21},...,\xi_{2M_2}$ | $M_2$ | $a_2$ | $D_2, \sigma_2^2$ | $x_{21},...,x_{2M_2}$ | $N_2$ | $\bar{x}_2$ | $D_{в2}$, $s_2^2$ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| k | $\xi_{k1},...,\xi_{kM_k}$ | $M_k$ | $a_k$ | $D_k, \sigma_k^2$ | $x_{k1},...,x_{kM_k}$ | $N_k$ | $\bar{x}_k$ | $D_{вk}$, $s_k^2$ |

Здесь объемы связаны следующим образом: $M = \sum_{j=1}^{k} M_j$, $N = \sum_{j=1}^{k} N_j$, а индексация значений случайной величины имеет вид: $\xi_{ji}$ (или $x_{ji}$), где первый индекс означает номер слоя (группы) j = 1, 2,..., k, а второй – порядковый номер в слое i = 1, 2,..., $M_k$ ($N_k$).

Генеральная средняя ≡ Математическое ожидание.

Далее будем рассматривать такие дискретные случайные величины, которые принимают M различных значений с одинаковой вероятностью: $p_i = \frac{1}{M}$ или $p_{ji} = \frac{1}{M_j}$ (так себя ведут результаты наблюдений и измерений), тогда:

$$a \equiv E(\xi) = \frac{1}{M}\sum_{i=1}^{M}\xi_i, \qquad a_j = \frac{1}{M_j}\sum_{i=1}^{M_j}\xi_{ji},$$

$$\sigma^2 \equiv D(\xi) \equiv E(\xi - a)^2 = \frac{1}{M}\sum_{i=1}^{M}(\xi_i - a)^2, \qquad \sigma_j^2 = \frac{1}{M_j}\sum_{i=1}^{M_j}(\xi_{ji} - a_j)^2$$

Для **системы** двух случайных величин $\xi$, $\eta$ (встречающихся пáрами):
ковариация (*корреляционный момент*):

$$\operatorname{cov}(\xi,\eta) \equiv E(\xi - a, \eta - b) = \frac{1}{M}\sum_{i=1}^{M}(\xi_i - a)(\eta_i - b),$$

где a, b – математические ожидания случайных величин $\xi$, $\eta$;
$\xi_i$, $\eta_i$ – все возможные пары значений.

Коэффициент корреляции:

$$\rho_{\xi\eta} = \frac{\operatorname{cov}(\xi,\eta)}{\sigma_\xi \sigma_\eta}.$$

К сожалению, всеми этими формулами мы воспользоваться не можем, так как не можем знать истинных значений M, $M_j$, p, $p_j$ и законов распределения

$$P(A)=\lim_{\nu\to\infty}\frac{\mu}{\nu}$$

– статистическое определение вероятности,

где $\mu$ – число появлений события A в $\nu$ опытах

Статистическое определение вероятности основывается на априорном свойстве **состоятельности** любого массового повторения опытов. Это значит, что при бесконечном увеличении числа повторений опытов относительная частота появления интересующего нас события стремится к вероятности:

$$\frac{\mu}{\nu}\xrightarrow[\nu\to\infty]{}P(A)=\frac{m}{n}.$$

Иначе говорят, что $\frac{\mu}{\nu}$ **сходится по вероятности** к величине P(A): $P\left(\left|\frac{\mu}{\nu}-P(A)\right|<\varepsilon\right)\xrightarrow[\nu\to\infty]{}1.$

Математическая статистика – наука для разработки методов регистрации, описания и анализа экспериментальных данных наблюдения массовых случайных явлений

Центральное место в математической статистике занимают *теория оценок* и *теория проверки гипотез*

Основное правило математической статистики гласит: каждое выдвинутое предложение должно быть *оценено* и *проверено* на правдоподобие. Для обеспечения этого правила и служит аппарат математической статистики.

Математическая статистика позволяет по результатам наблюдения **частного** (*выборки*) сделать некоторые обоснованные выводы о характеристиках **общег**о (*генеральной совокупности*)

Общая последовательность применения методов математической статистики, предложенная Р. Фишером (в скобках дается комментарий в современных терминах):

**1.** Планирование исследований (*планирование эксперимента*, определение способа отбора информации)

**2.** Конкретизация математико-статистического описания (выбор *дисперсионной* или *регрессионной модели*)

**3.** Оценка параметров модели (получение *точечных и интервальных оценок*) и составление их выборочных (эмпирических) распределений

**4.** Изучение согласия между моделью и наблюдениями (*адекватность модели* оригиналу и проверка критериев согласия в обоснование модели)

**5.** Реальное решение задачи посредством оценок параметров и критериев значимости (*статистический анализ* результатов и разработка выводов)

1-й и 3-й этапы составляют процедуру первичной обработки информации (отбор информации, построение гистограмм и полигонов частот, расчет точечных и интервальных **оценок**)

Эти результаты служат исходным материалом для 4-го и 5-го этапов – статистического анализа, целью которого является установление статистических **закономерностей**

# Точечные оценки

[Часть II, стр. 14 - 19]

Определение значения некоторого (в общем случае ненаблюдаемого) параметра наблюдаемого объекта по экспериментальным данным носит название статистической точечной оценки

**Метод моментов** (предложен К. Пирсоном): приравнивание начальных $\nu_r$ или центральных $\mu_r$ моментов порядка r генеральной совокупности соответствующим моментам выборки.

Математическое ожидание — выборочная средняя

$$a \equiv \nu_1(\xi) \equiv \frac{1}{M}\sum_{i=1}^{M}\xi_i \quad \rightarrow \quad \bar{x} \equiv \nu_1(x) \equiv \frac{1}{N}\sum_{i=1}^{N}x_i$$

Дисперсия — выборочная оценка дисперсии

$$\sigma^2 \equiv \mu_2(\xi) \equiv \frac{1}{M}\sum_{i=1}^{M}(\xi_i - a)^2 \quad \rightarrow \quad D_в \equiv \mu_2(x) \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Для расслоенных выборок:

$$\bar{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}x_{ji}, \qquad D_{вj} = \frac{1}{N_j}\sum_{i=1}^{N_j}(x_{ji} - \bar{x}_j)^2,$$

$$\bar{\bar{x}} = \frac{1}{k}\sum_{j=1}^{k}\bar{x}_j = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}x_{ji}, \qquad D_в = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ji} - \bar{\bar{x}})^2.$$

Основные очевидные преимущества метода моментов в простоте вычислений и в **независимости** от законов распределения - их не нужно знать

Вопросы:

– какую из нескольких выборочных средних $\bar{x}_j$ (нескольких исследований) принять в качестве оценки математического ожидания?

– как разрешить противоречие в формуле выборочной оценки дисперсии $D_в = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ по результатам одного измерения?

Свойства точечных оценок $\lambda^*$

Несмещенность: $E(\lambda^*) = \lambda$ **(без систематической ошибки)**

Состоятельность: $\lim_{N\to\infty} P(|\lambda^*-\lambda|<\varepsilon)=1$ **($\lambda^*\to\lambda$)**

Эффективность: $D(\lambda^*) = \min$ **(наименьшая неопределенность)**

Достаточная (исчерпывающая) точечная оценка **не может** быть существенно изменена из-за получения какой-либо **дополнительной** информации
*(Эффективная* оценка обязательно является *достаточной)*

Несмещенные точечные оценки

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

$$s_j^2 = \frac{1}{N_j - 1}\sum_{i=1}^{N_j}(x_{ji} - \bar{x}_j)^2, \qquad s^2 = \frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ji} - \bar{\bar{x}})^2$$

$s^2$ – исправленная выборочная оценка дисперсии

| Оцениваемый параметр | Вычисляемая характеристика | Свойства оценок | | |
|---|---|---|---|---|
| | | несмещенность | состоятельность | эффективность |
| $a$ | $\bar{x}$ | + | + | + |
| $a$ | $\tilde{x}$ | – | + | – |
| $a$ | $x_M$ | – | + | – |
| $\sigma^2$ | $D_{\hat{a}}$ | – | + | + |
| $\sigma^2$ | $s^2$ | + | + | + |
| $\sigma$ | $\sqrt{D_{\hat{a}}}$ | – | + | + |
| $\sigma$ | $s$ | – | + | + |
| $\sigma$ | $R$ | – | + | + при $N < 10$ |

**<u>Метод наибольшего правдоподобия</u>**

(Р. Фишер )

<u>Функция правдоподобия</u>

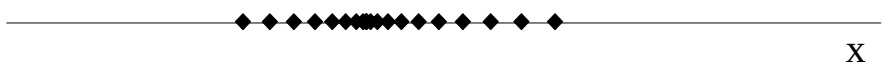
(вероятность появления наблюдаемой выборки, вычисленная при значении искомой оценки параметра $\lambda^*$):

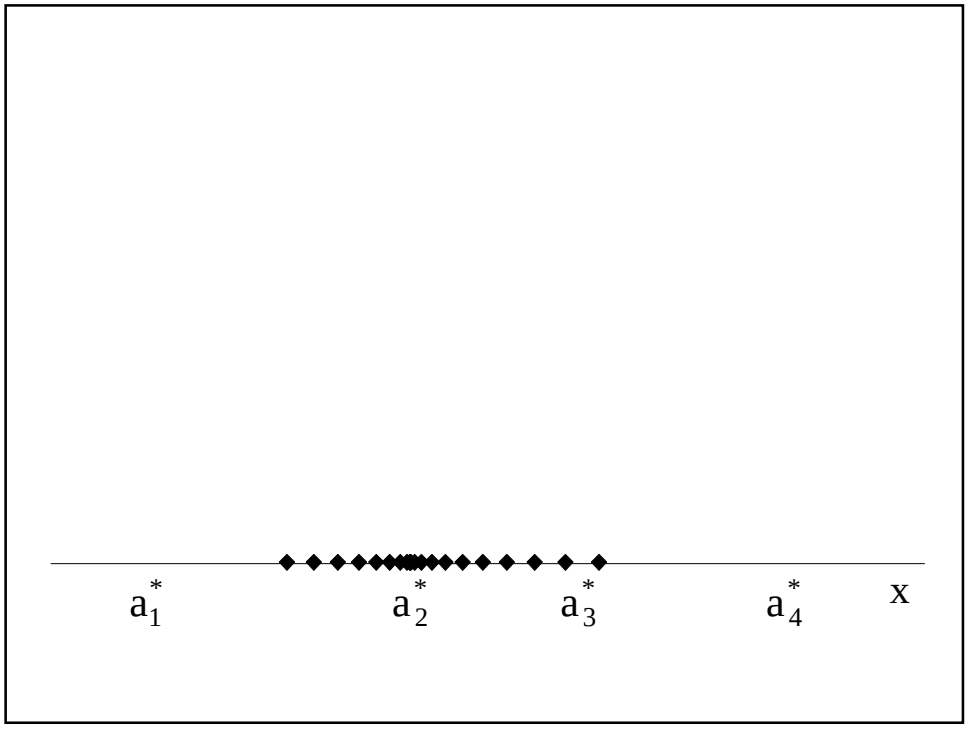
$$L = P(x_1, x_2, ..., x_N, \lambda^*),$$

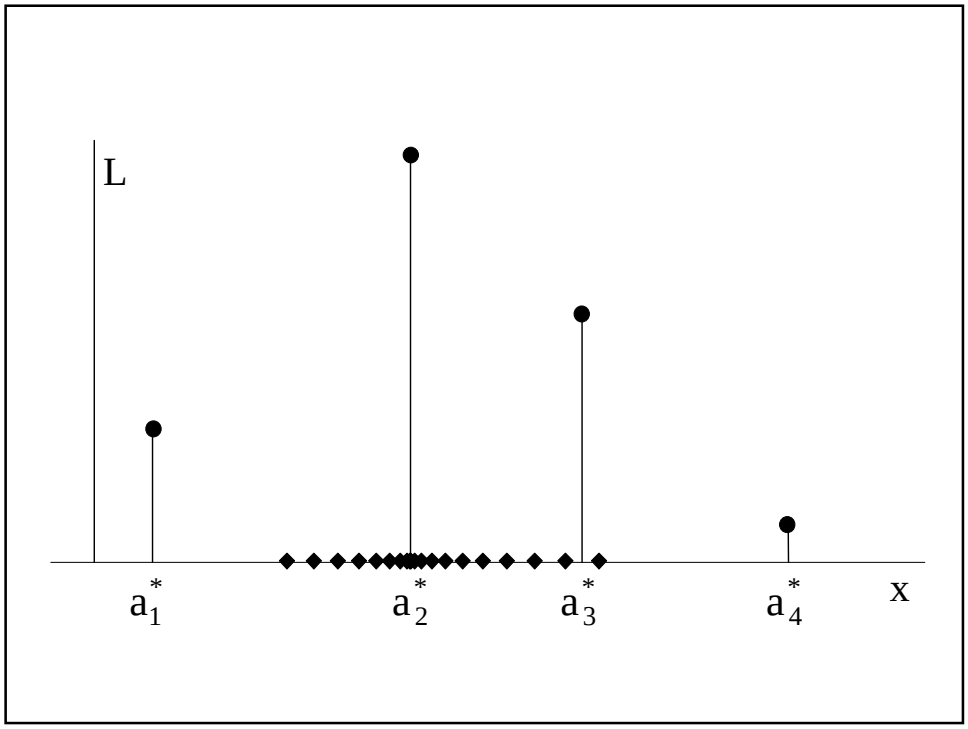
где $x_1, x_2, ..., x_N$ – выборка; результаты эксперимента, наблюдения

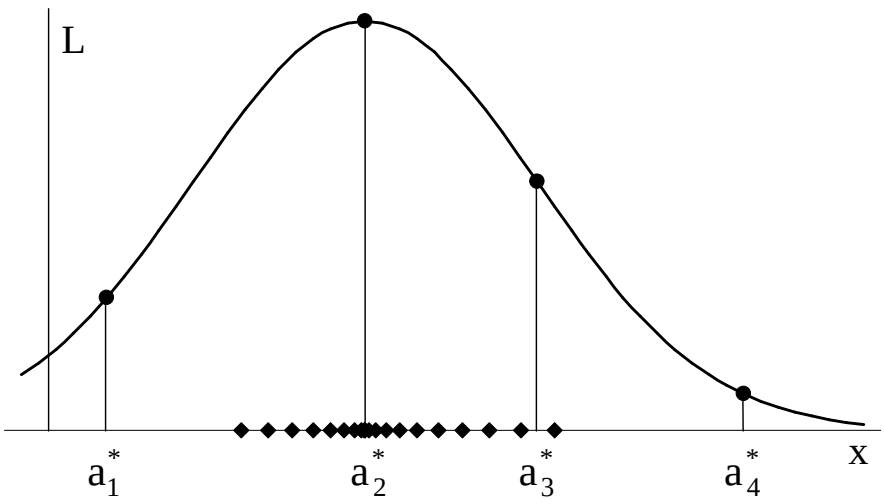
Для получения несмещенных, состоятельных и эффективных точечных оценок Р. Фишер предложил искать такие $\lambda^*$, которые максимизируют функцию правдоподобия:

$$L = P(x_1, x_2, ..., x_N, \lambda^*) \Rightarrow \max.$$




$a_1^*$
$a_2^*$
$a_3^*$
$a_4^*$
x


L
$a_1^*$
$a_2^*$
$a_3^*$
$a_4^*$
x




Обычно на практике используются законы распределения, описываемые с помощью экспонент: L ~ exp, поэтому для максимизации функции правдоподобия легче решать уравнение наибольшего правдоподобия:

$$\frac{\partial \ln L}{\partial \lambda^*} = 0$$

Для **независимо** полученных значений случайной величины
(а именно так и стремятся поставить эксперимент):

$$L = P(x_1, x_2, ..., x_N, \lambda^*) =$$

$$= P(x_1, \lambda^*) \cdot P(x_2, \lambda^*) \cdot ... \cdot P(x_N, \lambda^*) = \prod_{i=1}^{N} P(x_i, \lambda^*)\text{'}$$

откуда:

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^{N} \frac{1}{P(x_i, \lambda^*)} \cdot \frac{\partial P(x_i, \lambda^*)}{\partial \lambda^*} = 0.$$

Число степеней свободы для системы n случайных величин - это число n этих величин **минус** число линейных **связей** между ними

Число степеней свободы f:
- при определении выборочного среднего f = N (весь объем выборки);
- при определении несмещенной оценки дисперсии f = N-1 (весь объем выборки за исключением 1 найденного выборочного среднего).

# Законы распределения случайных величин

[Часть II, стр. 20 - 23]

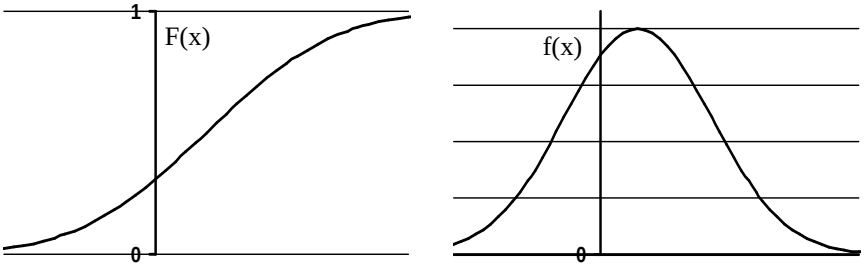
Центральная предельная теорема:
сумма **произвольно** распределенных
**независимых** случайных величин
при условии **одинакового** их влияния
подчиняется *нормальному закону* (*распределению ошибки, распределению Гаусса*):

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-a)^2}{2\sigma^2}}=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$$

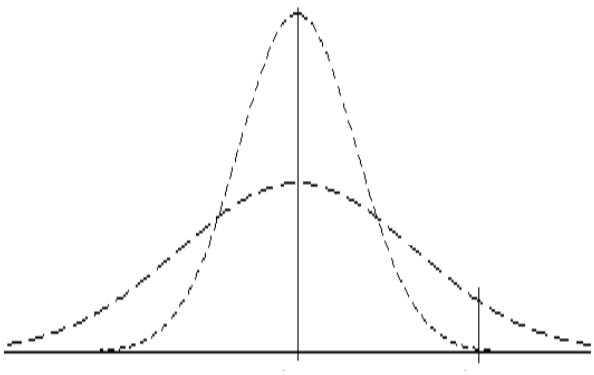
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{1}{2}\left(\frac{t-a}{\sigma}\right)^2}dt$$

с математическим ожиданием a и дисперсией $\sigma^2$.

Нормальный закон распределения с математическим ожиданием a=? и дисперсией $\sigma^2$=?




Плотность распределения вероятности нормального закона для разных дисперсий



## ФУНКЦИЯ ЛАПЛАСА

Функция распределения для **стандартизованного** нормального закона при a = 0, σ = 1 называют функцией Лапласа. Существует несколько разновидностей функции Лапласа:

$$\Phi^*(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt,\quad \Phi_1(x)=\frac{2}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{t^2}{2}}dt,$$

$$\boxed{\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{t^2}{2}}dt}$$

Через Φ(x) можно выразить интегральную функцию распределения для любых значений математического ожидания a и среднего квадратического отклонения σ:

$$F(x)=\Phi^*\left(\frac{x-a}{\sigma}\right)=0{,}5+\frac{1}{2}\Phi_1\left(\frac{x-a}{\sigma}\right)=0{,}5+\Phi\left(\frac{x-a}{\sigma}\right)$$

Множество выборочных средних (по слоям, по группам, по экспериментам) $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ претендует на оценку математического ожидания.

Каждое из них – случайная величина, распределенная согласно центральной предельной теореме по нормальному закону.

А знание закона распределения позволяет с помощью метода наибольшего правдоподобия получить несмещенную, состоятельную и эффективную оценку математического ожидания.

Так же возможно получение несмещенных, состоятельных и эффективных оценок не только математического ожидания, но и некоторых **функций** от него.

Выборочные функции – функции от выборочных значений таких величин, как среднее выборочное, выборочная оценка дисперсии, выборочные ковариация или коэффициент корреляции и т.д.

Для многих таких выборочных функций с помощью теорем математической статистики выявлены законы распределения, которым они подчиняются.

В следующей таблице предполагается, что выборка объема $N$ (или в слое $N_j$) сделана из нормально распределенной генеральной совокупности с математическим ожиданием $a$ (или $a_j$) и дисперсией $\sigma^2$ (или $\sigma_i^2$).

Обозначения: для характеристик расслоенных выборок:

$$s_A^2 = \frac{1}{k-1} \cdot \sum_{j=1}^{k} N_j (\bar{x}_j - \bar{\bar{x}})^2, \qquad s_0^2 = \frac{1}{N-k} \cdot \sum_{j=1}^{k} (N_j - 1) s_j^2$$

– *межгрупповая дисперсия* между слоями (рассеяние из-за влияния исследуемого фактора) и *остаточная* внутри слоев (*внутренняя дисперсия*, рассеяние результатов из-за влияния неучтенных факторов)

Для системы случайных величин – гипотетический (генеральный) и выборочный *коэффициенты регрессии*:

$$\beta_{\eta\xi} = \rho_{\xi\eta}\frac{\sigma_\eta}{\sigma_\xi}, \qquad b_{yx} = r_{xy}\frac{s_y}{s_x},$$

где $r_{xy} = r_{yx} = \frac{l_{xy}}{s_x s_y}$, а $l_{xy} = l_{yx} = \frac{1}{N}\cdot\sum_{j=1}^{N}(x_j - \bar{x})(y_j - \bar{y})$

– выборочные *коэффициент корреляции* и *ковариация*

В таблице обозначены следующие затабулированные в специальной литературе законы распределения:

u – стандартизованное нормальное распределение с a=0, $\sigma$=1 (u-распределение);

t – распределение Стьюдента (t-распределение);

r – r-распределение;

$\chi^2$ – $\chi^2$-распределение Пирсона;

F – распределение Фишера ($v^2$-распределение);

z – z-распределение.

# Законы распределения выборочных функций

| № | Выборочная функция | Закон распределения | Число степеней свободы закона |
|---|---|---|---|
| 1 | $\frac{x_i - a}{\sigma}$ | u | – |
| 2 | $\frac{x_i - a}{s} = \frac{x_i - a}{\sqrt{D_в}}\sqrt{\frac{N-1}{N}}$ | t | N–1 |
| 3 | $\frac{x_i - \bar{x}}{\sqrt{D_в}} = \frac{x_i - \bar{x}}{s}\sqrt{\frac{N}{N-1}}$ | r | N–1 |
| 4 | $\frac{\bar{x} - a}{\sigma}\sqrt{N}$ | u | – |
| 5 | $\frac{\bar{x} - a}{s}\sqrt{N} = \frac{\bar{x} - a}{\sqrt{D_в}}\sqrt{N-1}$ | t | N–1 |

| № | Выборочная функция | Закон распределения | Число степеней свободы закона |
|---|---|---|---|
| 6 | $\frac{ND_в}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \bar{x})^2$ | $\chi^2$ | N–1 |
| 7 | $\frac{N\hat{s}^2}{\sigma^2} \equiv \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - a)^2$ | $\chi^2$ | N |
| 8 | $(N-k)\frac{s_0^2}{\sigma^2}$ | $\chi^2$ | N–k |
| 9 | $(k-1)\frac{s_A^2}{\sigma^2}$ | $\chi^2$ | k–1 |
| 10 | $\frac{(\bar{x}_i - \bar{x}_j)(a_i - a_j)}{\sigma}\sqrt{\frac{N_i N_j}{N_i + N_j}}$ | u | – |
| 11 | $\frac{(\bar{x}_i - \bar{x}_j)(a_i - a_j)}{s_0}\sqrt{\frac{N_i N_j}{N_i + N_j}}$ | t | $N_i + N_j - 2$ |

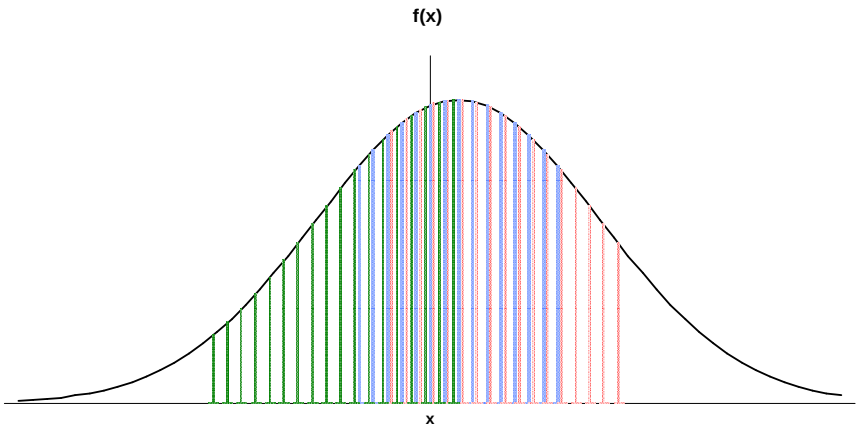
| № | Выборочная функция | Закон распределения | Число степеней свободы закона |
|---|---|---|---|
| 12 | $\frac{s_i^2}{s_j^2}$ | F | $N_i - 1, N_j - 1$ |
| 13 | $\frac{s_A^2}{s_0^2}$ | F | k–1, N–k |
| 14 | $\sqrt{N-2}\frac{r_{xy}}{\sqrt{1-r_{xy}^2}}$ | t | N–2 |
| 15 | $r_{xy}\sqrt{N-1}$ | r | N–2 |
| 16 | $\frac{s_x\sqrt{N-2}}{s_y\sqrt{1-r_{xy}^2}}(b_{yx}-\beta_{\eta\xi})$ | t | N–2 |

# Интервальные оценки

[Часть II, стр. 23 - 25]

Доверительный интервал (Ю. Нейман): интервал $(\lambda_l^*, \lambda_r^*)$ ($\lambda_l^*$ – левая граница, $\lambda_r^*$ – правая граница), в котором с заданной *доверительной вероятностью* $\gamma$ следует ожидать истинное (но не известное) значение оцениваемого параметра $\lambda$, т.е.:

$$P(\lambda_l^* < \lambda < \lambda_r^*) = \gamma$$



Наиболее естественным является выбор границ с опорой на точечную **оценку** $\lambda^*$ искомого параметра, найденную предварительно:

$$\lambda_l^* = \lambda^* - \delta_l, \quad \lambda_r^* = \lambda^* + \delta_r,$$

где $\delta_l, \delta_r$ – *погрешности* (допуски) характеризуют **точность** оценки влево и вправо от $\lambda^*$.

В простейшем случае принимают $\delta = \delta_l = \delta_r$, т.е. строят **симметричный** доверительный интервал относительно точечной оценки параметра.

**ПРИМЕР.** Найти симметричный доверительный интервал с вероятностью $\gamma$ для математического ожидания $a$ нормально распределенной случайной величины $\xi$ в случае известного среднего квадратического отклонения $\sigma$.

---

В таблице выборочных функций искомый параметр $a$ и известный параметр $\sigma$ есть только в 1-й и в 4-й строках. Из **1-й** можно оценить $a$, исходя из единственного замера искомого нормально распределенного параметра $x_i$:

$$P(x_i-\delta<a<x_i+\delta)=P(-\delta<a-x_i<\delta)=$$
$$=P\left(-\frac{\delta}{\sigma}<\frac{x_i-a}{\sigma}<\frac{\delta}{\sigma}\right)$$

Вероятность $\gamma$ попадания случайной величины, распределенной по нормальному закону, в заданный интервал выразится с помощью функции Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$:

$$P\left(-\frac{\delta}{\sigma}<\frac{x_i-a}{\sigma}<\frac{\delta}{\sigma}\right)=\gamma=F\left(\frac{\delta}{\sigma}\right)-F\left(-\frac{\delta}{\sigma}\right)=$$
$$=0{,}5+\Phi\left(\frac{\delta}{\sigma}\right)-0{,}5-\Phi\left(-\frac{\delta}{\sigma}\right)=2\Phi\left(\frac{\delta}{\sigma}\right).$$

При заданном $\gamma$ по таблице функции Лапласа можно определить ее аргумент $u_{0,5\gamma} = \delta/\sigma$ (т.е. $\Phi(u_{0,5\gamma}) = 0,5\gamma$). Таким образом численное значение погрешности $\delta$ для точечной оценки математического ожидания a по данным единственного замера $x_i$ определится: $\delta = u_{0,5\gamma}\sigma$.

В итоге доверительный интервал по единственному замеру:

$$x_i - u_{0,5\gamma}\sigma < a < x_i + u_{0,5\gamma}\sigma.$$

Аналогично из **4-й** строки, исходя из выборочного среднего $\bar{x}$ по выборке объемом $N$:

$$\bar{x} - u_{0,5\gamma}\frac{\sigma}{\sqrt{N}} < a < \bar{x} + u_{0,5\gamma}\frac{\sigma}{\sqrt{N}},$$

где погрешность $\delta$ в $\sqrt{N}$ меньше (увеличение точности) раз по сравнению с единственным замером.

Этот факт давно известен человечеству: "семь раз отмерь – один отрежь".

# Основы статистического контроля качества технологических процессов

[Часть II, стр. 39 - 41, 44 - 45]

Организуя специальным образом сбор статистического материала о параметрах производства, можно делать научно обоснованные (с оценкой ошибок) выводы о качестве процесса, тенденциях его изменения и о качестве продукции.

**Текущий контроль** качества технологических процессов с помощью *контрольных карт*.

Доверительный интервал для математического ожидания по выборочному среднему, и для выборочного среднего по известному математическому ожиданию:

$$P(\overline{x}-\delta<a<\overline{x}+\delta)=P(-\delta<a-\overline{x}<\delta)=$$
$$=P(\delta>\overline{x}-a>-\delta)=P(a+\delta>\overline{x}>a-\delta)=\gamma$$

$$P(a+\delta>\overline{x}>a-\delta)=\gamma$$

Математическое ожидание a – **нормативное значение** контролируемого параметра технологического процесса (на которое он **должен** быть настроен).

δ – допуски (допустимые погрешности) контролируемого параметра.


Распространенные <u>виды контрольных карт</u>:

1) "$\bar{x}$" – средних значений;
2) "$x_i$" – индивидуальных значений;
3) "$\tilde{x}$" – медиан;
4) "$\bar{x}$/s";
5) "$\bar{x}$/R";
6) "$\tilde{x}$/R";
7) "p" – среднего процента брака;
8) "Np" – модификация карты p;
9) "c" – дефектов, отказов;
10) "u" – обобщение карты "c" на процент дефектов, отказов.

Карты 1 – 6 называются *контрольными картами* по <u>измеримым признакам</u>, карты 7 – 10 – по <u>неизмеримым</u>.

## **Приемочный контроль с помощью выборочного метода**

Однократные выборки

Необходимо определить объем проверяемой выборки N и приемочное число с – норму браковки: для принятия всей партии из M изделий необходимо, чтобы в выборке N число бракованных изделий было меньше с.

Возможны два вида ошибок (когда выборка неверно отражает качество всей партии):

– ошибка I рода – по результатам выборочного контроля **бракуется годная** партия, ее вероятность обозначим $\alpha$,

– ошибка II рода – по результатам выборочного контроля **негодная** партия **принимается**, ее вероятность обозначим $\beta$.

Вместо одного критического значения p **доли брака** приходится назначать два: $p_\alpha$ – *допускаемый уровень качества*, и $p_\beta$ – *недопустимый уровень качества*. Между ними ($p_\alpha < p_\beta$) находится область неопределенности, которую необходимо исключить при организации приемочного контроля однократными выборками. Именно для замены этих двух границ определяются N и c.

Из теории вероятностей известно, что если доля брака определяется вероятностью p, то вероятность обнаружения среди N изделий, отобранных из M, ровно k бракованных вычисляется по формуле:

$$P_N(k)=\frac{C_{pM}^{k}C_{M-pM}^{N-k}}{C_M^N}.$$

Тогда вероятность принятия партии с **допускаемым** уровнем качества $p_\alpha$, т.е. случая, когда в выборке качества $p_\alpha$ число бракованных изделий не превосходит приемочного числа c, определится суммой:

$$\sum_{k=0}^{c}P_N(k,\alpha)=\sum_{k=0}^{c}\frac{C_{p_\alpha M}^{k}C_{M-p_\alpha M}^{N-k}}{C_M^N}=1-\alpha$$

А вероятность принятия партии с **недопустимым** уровнем качества $p_\beta$, т.е. случая, когда в выборке качества $p_\beta$ число бракованных изделий не превосходит приемочного числа c, определится суммой:

$$\sum_{k=0}^{c} P_N(k,\beta) = \sum_{k=0}^{c} \frac{C_{p_\beta M}^{k} C_{M-p_\beta M}^{N-k}}{C_M^N} = \beta$$

Из этих двух уравнений определяются два неизвестных N и c. Таким образом, строится *однократная выборка* для *приемочного контроля качества*.

## Многократные выборки

По результатам очередной i-й выборки можно сделать три вывода:

а) в случае числа бракованных изделий $k < \overline{c}_i$ принять партию,

б) в случае числа бракованных изделий $k > \overline{\overline{c}}_i$ забраковать партию,

в) в случае $\overline{c}_i < k < \overline{\overline{c}}_i$ продолжить контроль.

Таким образом, в результате процедуры *последовательного анализа* многократных выборок неизбежно принимается определенное решение.

# Проверка статистических гипотез

[Часть II, стр. 25 - 33]

Правило математической статистики: любое предположение, основанное на **выборочных** данных, должно быть проверено.

Статистический анализ не может доказать истинность, но может указать с некоторой долей уверенности на **наличие или отсутствие признаков опровержения** данного суждения.

Проверка статистических гипотез – оценка **соответствия** выдвинутой гипотезы полученному статистическому материалу, т.е. **выборке**.

Виды статистических гипотез

| | | | |
|---|---|---|---|
| Гипотезы о значениях **параметров** $\lambda$ закона распределения $F(x, \lambda)$ | $\Rightarrow$ | параметрические критерии | $H_0$: $\lambda = \lambda_0$ |
| Гипотезы о **непараметрических** свойствах распределения | $\Rightarrow$ | непараметрические критерии | $H_0$: $F(x) = F_0(x, \lambda_0)$ |

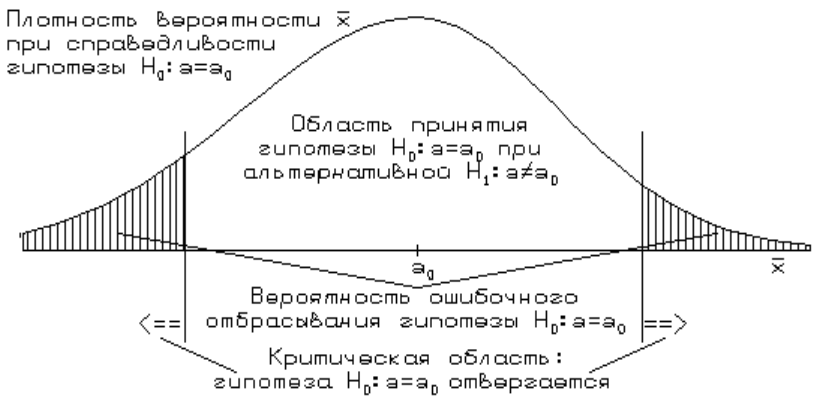
**Критерий** соответствия гипотезы статистическому материалу – достижение определенного значения ***функции правдоподобия***:

| | | | | |
|---|---|---|---|---|
| выборка попадает в область малого правдоподобия | $\Rightarrow$ | присутствуют признаки опровержения гипотезы | $\Rightarrow$ | **есть основания отвергнуть** гипотезу |
| выборка попадает в область большого правдоподобия | $\Rightarrow$ | отсутствуют признаки опровержения гипотезы | $\Rightarrow$ | **нет оснований отвергнуть** гипотезу |

Нельзя забывать о возможной **ошибке** в наших выводах, поэтому здесь возможны не два, а четыре исхода:

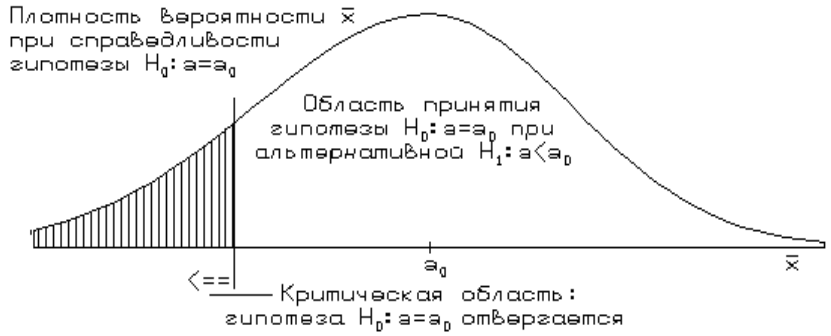
| | | |
|---|---|---|
| гипотеза **верна** и **не отвергается** согласно критерию | правильный вывод | $1 - \alpha$ |
| гипотеза **неверна** и **отвергается** согласно критерию | правильный вывод | $1 - \beta$ мощность критерия |
| гипотеза **верна**, но **отвергается** согласно критерию | ошибка I рода | $\alpha$ уровень значимости |
| гипотеза **неверна**, но **не отвергается** согласно критерию | ошибка II рода | $\beta$ |

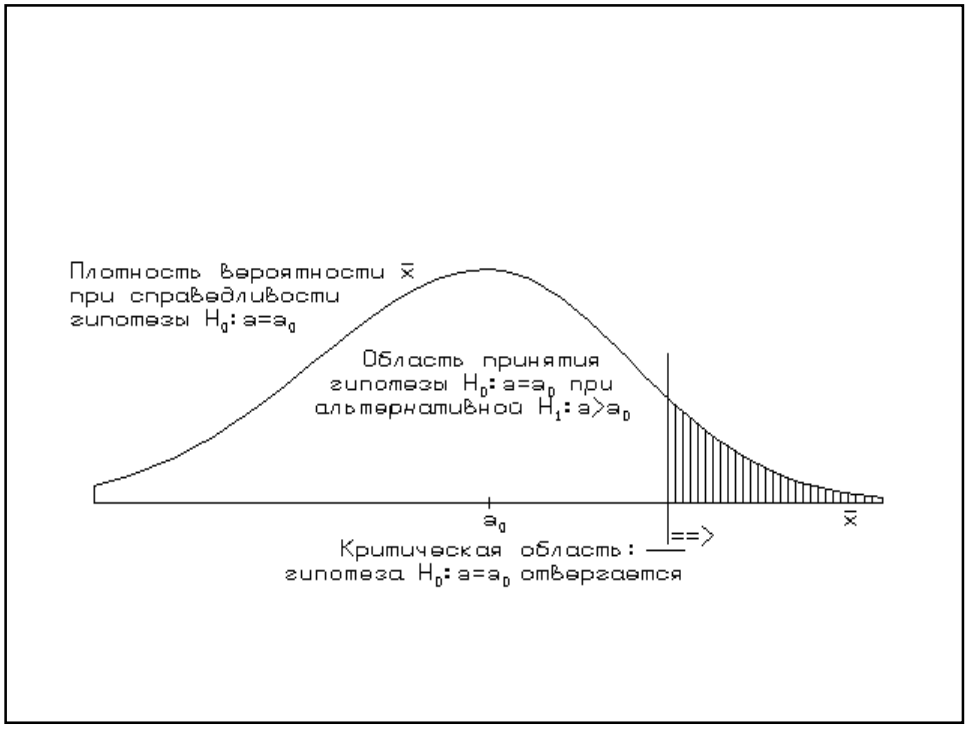
Для оценки гипотезы необходимо **назначать уровень значимости** – максимальное значение вероятности, которое принимается за **практическую невозможность** получения конкретной выборки с гипотетическими свойствами.

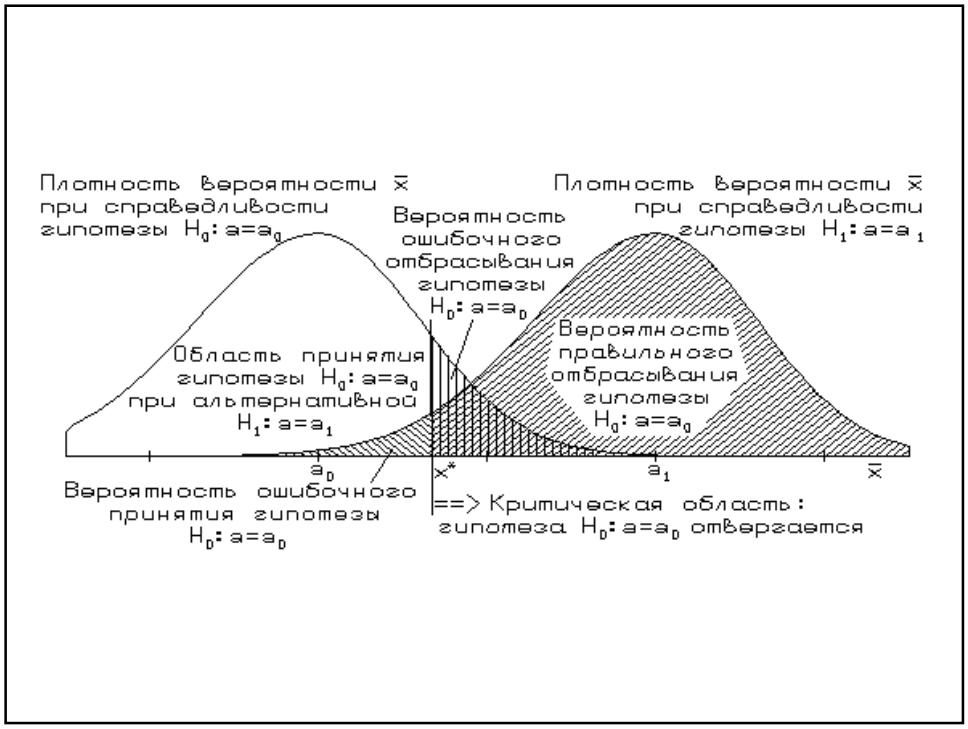



Если $\bar{x}$ попадет в критическую область, то **различие** между $\bar{x}$ и $a_0$ **значимо** (существенно, неслучайно, проистекает от каких-то неучтенных факторов, им нельзя пренебречь) и **есть основания отвергнуть** гипотезу $H_0$: $a = a_0$ – она **не может быть принята**.

Если $\bar{x}$ не попадет в критическую область, то **различие** между $\bar{x}$ и $a_0$ **незначимо** (несущественно, случайно, им можно пренебречь) и **нет оснований отвергнуть** гипотезу $H_0$: $a = a_0$ – она **может быть принята**.




Плотность вероятности $\bar{x}$
при справедливости
гипотезы $H_0: a=a_0$
Область принятия
гипотезы $H_0: a=a_0$ при
альтернативной $H_1: a>a_0$
$a_0$
$\bar{x}$
==>
Критическая область:
гипотеза $H_0: a=a_0$ отвергается


Плотность вероятности $\bar{x}$
при справедливости
гипотезы $H_0: a=a_0$
Плотность вероятности $\bar{x}$
при справедливости
гипотезы $H_1: a=a_1$
Вероятность
ошибочного
отбрасывания
гипотезы
$H_0: a=a_0$
Область принятия
гипотезы $H_0: a=a_0$
при альтернативной
$H_1: a=a_1$
Вероятность
правильного
отбрасывания
гипотезы
$H_0: a=a_0$
$a_0$
$x^*$
$a_1$
$\bar{x}$
Вероятность ошибочного
принятия гипотезы
$H_0: a=a_0$
==> Критическая область:
гипотеза $H_0: a=a_0$ отвергается

Алгоритм проверки статистических гипотез

с помощью параметрических критериев

**1.** выдвижение оцениваемой гипотезы H0;

**2.** выдвижение альтернативной гипотезы H1;

**3.** установление подходящего уровня значимости $\alpha$;

**4.** выбор подходящей выборочной функции по следующим признакам:

– подчиненность известному закону распределения (хотя бы с контролируемым **приближением**),

– простота вычислений,

– обеспечение наилучшего критерия (наиболее мощного: $1 - \beta \to 1$);

**5.** определение (вычисление или построение) распределения используемой выборочной функции в предположении гипотезы $H_0$;

**6.** определение критической области для проверки гипотезы $H_0$ с учетом альтернативной $H_1$;

**7.** получение выборки и вычисление значения выборочной функции (вычисление "статистики");

**8.** принятие решения: если вычисленное в предыдущем пункте значение ("статистика") попало в критическую область, то гипотезу следует отвергнуть, в противоположном случае нет оснований отвергнуть гипотезу.

Несогласованность $\alpha$, $\beta$ и $a_1 - a_0$ может приводить к невозможности сделать определенный вывод.

С другой стороны, некоторые параметры, например, $\sigma$, могут быть неизвестны, что требует проведения отдельного эксперимента.

Эти сложности разрешаются с помощью *последовательного анализа* и применения *секвенциальных (последовательных) критериев*.

Такие критерии позволяют сделать один из **трех** выводов: принять проверяемую гипотезу, принять альтернативную гипотезу, продолжить эксперимент.

*Секвенциальный* (последовательный) *критерий* можно построить при условиях:

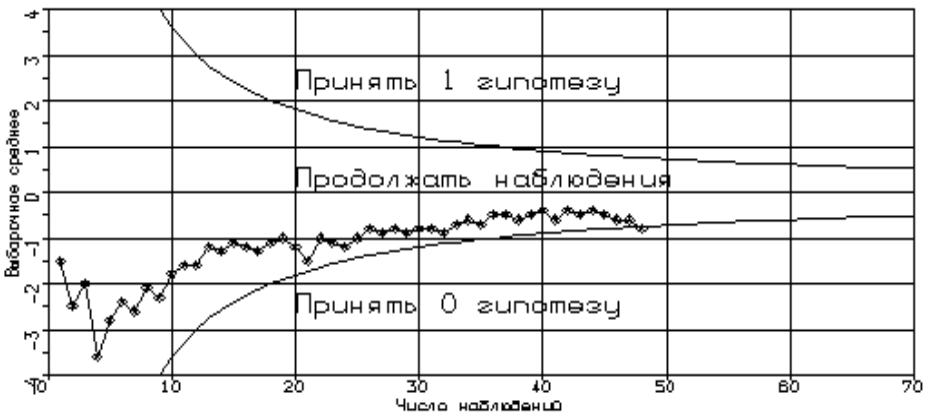
– гипотезы $H_0$ и $H_1$ фиксированы,

– вид функции распределения известен,

– $\alpha$ и $\beta$ (вероятности ошибочного отвергания гипотез) выбраны.

Секвенциальный (последовательный) критерий А. Вальда

$$\frac{\beta}{1-\alpha} < \frac{p_{1m}}{p_{0m}} < \frac{1-\beta}{\alpha},$$



Непараметрические критерии (ранговые) применяются для проверки свойств закона распределения

Ранги – числовые характеристики **упорядоченных** результатов эксперимента.

Критерий знаков – использует только два ранга: да – нет или больше – меньше.

*Задача*: сопоставление двух **непрерывных** случайных величин $\xi$ и $\eta$ по парным выборкам одного объема N, в которых значения случайных величин $\xi$ и $\eta$ встречаются па́рами: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$.

*Исходная гипотеза*: величины $\xi$ и $\eta$ **распределены одинаково**, тогда должны совпадать вероятности:

$$P(x_i > y_i) = P(x_i < y_i) = 0{,}5.$$

Вероятность того, что среди этих N пар более m имеют положительные разности $x_i - y_i > 0$:

$$p_N(m) = \frac{1}{2^N} \sum_{j=m+1}^{N} C_N^j .$$

По заданному (выбранному) уровню значимости $\alpha$ определяется $m(\alpha)$ – **наименьшее** значение m, при котором $p_N(m) \le \alpha$.

1) *Альтернативная гипотеза*: случайная величина $\xi > \eta$, тогда для опровержения исходной гипотезы необходимо, чтобы **число** положительных разностей $x_i - y_i > 0$ было больше $m(\alpha)$ – т.е. в этом случае превосходство значений $x_i$ над $y_i$ неслучайно – **зна́чимо**.

2) *Альтернативная гипотеза*: случайная величина $\xi$ существенно отличается от случайной величины $\eta$ в любую сторону ($\xi \neq \eta$), тогда исходная гипотеза опровергается, если **число** положительных **или число** отрицательных разностей окажется больше $m(\alpha)$ – при заданном уровне значимости $2\alpha$.

Критерий согласия К. Пирсона $\chi^2$ для сравнения законов распределения (эмпирического и/или теоретического) двух случайных величин.

Весь диапазон N данных эксперимента разбивается на $r \geq 6$ интервалов таким образом, чтобы в каждом i-м интервале оказалось наблюдаемых значений $N_i \geq 5$.

Вычисляется

$$\chi^2_{\text{наблюдаемое}} = \sum_{i=1}^{r} \frac{(N_i - Np_i)^2}{Np_i}$$

где $p_i$ – вероятность попадания в i-й интервал случайной величины, вычисленная по проверяемому теоретическому закону распределения.

$\chi^2_{\text{крит.}}(\alpha, n)$ определяется по таблице распределения $\chi^2$ при назначенном уровне значимости $\alpha$ с $n = r - 2$ степенями свободы.

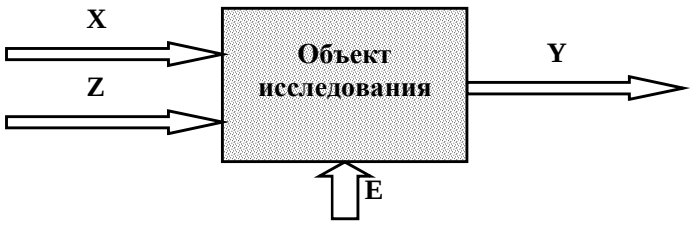
Если $\chi^2_{\text{наблюдаемое}} < \chi^2_{\text{крит.}}(\alpha, n)$, то различие статистического и гипотетического законов распределения **незна́чимо** и нет оснований отвергнуть гипотезу о совпадении законов распределения.

Если $\chi^2_{\text{наблюдаемое}} > \chi^2_{\text{крит.}}(\alpha, n)$ расхождение **зна́чимо** (не может считаться случайным) и гипотезу о совпадении законов распределения следует отвергнуть.

# Основы многомерного статистического анализа

## Задачи многомерного статистического анализа

[Часть II, стр. 46 - 47, 49 - 50]




**X** = ($x_1$, $x_2$, ..., $x_k$) – вектор **входных контролируемых** переменных, которыми **можно управлять** в исследовании;

**Z** = ($z_1$, $z_2$, ..., $z_p$) – вектор **входных контролируемых** переменных, которыми **невозможно управлять** в исследовании;

**E** = ($e_1$, $e_2$, ..., $e_f$) – вектор **входных неконтролируемых** и **неуправляемых** переменных (*шум*);

**Y** = ($y_1$, $y_2$, ..., $y_g$) – вектор **выходных** переменных.

Переменные **X, Z** и **Y** называются факторами.

Если фактор принимает фиксированные, детерминированные значения, то они называются уровнями *фактора*.

## Задачи многомерного статистического анализа

| | | |
|---|---|---|
| 1 | Существует ли **связь** между отдельными факторами (любыми из: $y_i$, $y_k$, $x_s$, $z_t$, $z_v$) | корреляционный анализ |
| 2 | Если между какими-то факторами есть связь, то насколько она **тесная** | |
| 3 | Если между какими-то факторами есть связь, то какой функцией ее можно **представить** | регрессионный анализ |
| 4 | Какие входные факторы оказывают на определенные выходные наибольшее **влияние** | дисперсионный анализ |
| 5 | Какие входные факторы можно **отбросить** из процесса изучения на основании их слабого, сравнимого с шумом, влияния | |
| 6 | Существуют ли **неучтенные** факторы, которые необходимо рассматривать ввиду их существенного влияния на выходные | |
| 7 | Существуют ли **обобщенные** факторы, которыми можно заменить несколько рассматриваемых | факторный анализ |
| 8 | Как связаны между собой **зашумленные** факторы | конфлюэнтный анализ |
| 9 | Каковы **характеристики шума** | |
| 10 | Как **выделить** "полезную" информацию из зашумленной | теория фильтрации |

Статистический анализ
предназначен
для обоснования и построения
*статистических*
*математических моделей*:
корреляционной,
регрессионной, дисперсионной

Задачи статистического анализа

| | |
|---|---|
| Корреляционный анализа | построение совместного **закона** распределения системы случайных величин |
| Регрессионный анализ | исследование вида и формы регрессии |
| Конфлюэнтный анализ | изучение **структуры** случайных величин, находящихся во взаимодействии |
| Дисперсионный анализ | **сравнение дисперсий** разных случайных величин или различных способов вычисления дисперсий |
| Факторный анализ | поиска минимального числа обобщенных факторов, заменяющих исходные |
| Теория фильтрации | выделение исходного сигнала из искаженной или неполной информации |

Разбиение методов статистического анализа условно.
Приступая к решению каждой конкретной задачи, необходимо из всей палитры методов выбрать приемлемые и решающие задачу.

# Понятие о корреляционном анализе

[Часть II, стр. 48 - 49, 50 - 53]

**<u>Корреляционный анализ</u> -** группа статистических методов установления формы и тесноты связи между факторами

Система **двух случайных величин** $\xi$ и $\eta$, принимающих возможные дискретные значения $x_i$ и $y_j$ с математическими ожиданиями a и b, соответственно.

**Дисперсии** этих величин:

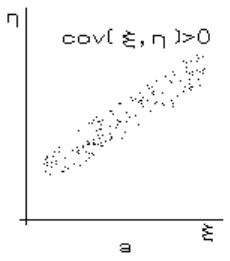
$$\sigma_\xi^2 = \sum_i (x_i - a)^2 p_i, \qquad \sigma_\eta^2 = \sum_j (y_j - b)^2$$

характеризуют **рассеяние** возможных значений,

а **ковариация**, являющаяся тоже моментом второго порядка,

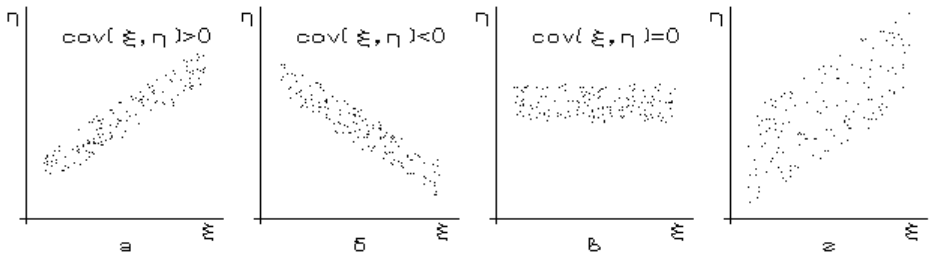
$$\operatorname{cov}(\xi,\eta)=\sum_{i,j}(x_i-a)(y_j-b)p_{ij}$$

характеризует **тенденцию** возрастания или убывания.



cov( ξ, η )>0
η
ξ
а
cov( ξ, η )<0
η
ξ
б


cov( ξ, η )>0
η
ξ
а
cov( ξ, η )<0
η
ξ
б
cov( ξ, η )=0
η
ξ
в




Ковариация показывает, насколько **связь** между случайными величинами близка к линейной. Она отражает и слишком большую **случайность**, и слишком большую **нелинейность** этой *связи*.

Если $\xi$ и $\eta$ **независимы**, то $p_{ij} = p_i p_j$ и ковариация обращается в **нуль**:

$$\operatorname{cov}(\xi,\eta)=\sum_{i,j}(x_i-a)(y_j-b)p_{ij}=$$
$$=\sum_i(x_i-a)p_i\sum_j(y_j-b)p_j=E(\xi-a)E(\eta-b)=0.$$

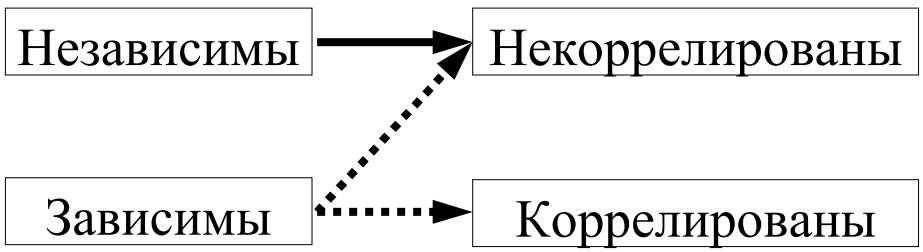
Дисперсия линейной комбинации двух случайных величин:

$$\sigma^2_{\alpha\xi+\beta\eta}=\sum_{i,j}[(\alpha x_i+\beta y_j)-(\alpha a+\beta b)]^2 p_{ij}=$$
$$=\sum_{i,j}[\alpha(x_i-a)+\beta(y_j-b)]^2 p_{ij}=$$
$$=\alpha^2\sum_i(x_i-a)^2p_i+\beta^2\sum_j(y_j-b)^2p_j+2\alpha\beta\sum_{i,j}(x_i-a)(y_j-b)p_{ij}=$$
$$=\alpha^2\sigma^2_\xi+\beta^2\sigma^2_\eta+2\alpha\beta\cdot\operatorname{cov}(\xi,\eta)$$

Такие случайные величины, ковариация которых равна нулю, называются **некоррелированными**.

В противном случае случайные величины называются **коррелированными**, т.е. между ними наблюдается **связь**.

Случайные величины $\xi$ и $\eta$



# Свойства системы двух случайных величин $\xi$ и $\eta$

Коэффициент корреляции:

$$\rho_{\xi\eta} = \frac{\operatorname{cov}(\xi,\eta)}{\sigma_\xi \sigma_\eta} = \frac{1}{\sigma_\xi \sigma_\eta} \sum_{i,j} (x_i - a)(y_j - b) p_i$$

| | | |
|---|---|---|
| независимость | $\Rightarrow$ | некоррелированность |
| $\rho_{\xi\eta} = 0$ | $\Leftrightarrow$ | некоррелированность |
| $\lvert\rho_{\xi\eta}\rvert = 1$ | $\Leftrightarrow$ | линейность |
| $\rho_{\xi\eta} = -1$ | $\Leftrightarrow$ | убывание |
| $\rho_{\xi\eta} = 1$ | $\Leftrightarrow$ | возрастание |

Связь **двух** факторов можно выявить по величине коэффициента корреляции $\rho_{\xi\eta}$, для которого надо получить статистическую оценку $r_{xy}$.

При $r_{xy} \approx 0$ факторы *некоррелированы*, при $|r_{xy}| \approx 1$ – *коррелированы* и связь почти **линейная**.

Корреляционная матрица связи

**нескольких** факторов

| | a | T | V | Q | N |
|---|---|---|---|---|---|
| a | 1,000 | -0,016 | -0,135 | -0,019 | 0,010 |
| T | -0,016 | 1,000 | -0,370 | -0,275 | -0,456 |
| V | -0,135 | -0,370 | 1,000 | 0,700 | 0,442 |
| Q | -0,019 | -0,275 | 0,700 | 1,000 | 0,352 |
| N | 0,010 | -0,456 | 0,442 | 0,352 | 1,000 |

Корреляционная модель – запись гипотетического закона распределения.

**Нормальный** закон распределения **системы** двух случайных величин:

$$f(\xi,\eta)=\frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1-\rho_{\xi\eta}^2}}e^{-\frac{1}{2(1-\rho_{\xi\eta}^2)}\left[\frac{(\xi-a)^2}{\sigma_\xi^2}-\frac{2\rho_{\xi\eta}(\xi-a)(\eta-b)}{\sigma_\xi\sigma_\eta}+\frac{(\eta-b)^2}{\sigma_\eta^2}\right]}$$

| Первичная обработка информации | | |
|---|---|---|
| $\bar{x}$ | $\Rightarrow$ | a |
| $\bar{y}$ | $\Rightarrow$ | b |
| $s_x$ | $\Rightarrow$ | $\sigma_\xi$ |
| $s_y$ | $\Rightarrow$ | $\sigma_\eta$ |
| Корреляционный анализ | | |
| $r_{xy}$ | $\Rightarrow$ | $\rho_{\xi\eta}$ |

Оценка тесноты связи факторов

**1) По доверительному интервалу** для коэффициента корреляции $\rho_{\xi\eta}$ (сложности вычислений функции распределения коэффициента корреляции).

**2) По корреляционному отношению.**

Полная дисперсия случайной величины $\eta$ как функции от $\xi$:

$$\sigma_\eta^2 \equiv D(\eta) \equiv E(\eta - b)^2 = E(\eta - \bar{y}_x + \bar{y}_x - b)^2 =$$
$$= E(\eta - \bar{y}_x)^2 + E(\bar{y}_x - b)^2 + 2E(\eta - \bar{y}_x)(\bar{y}_x - b).$$

Последнее слагаемое равно нулю, так как:

$$E(\eta - \bar{y}_x)(\bar{y}_x - b) = (\bar{y}_x - b)E(\eta - \bar{y}_x) = 0,$$

поэтому можно ввести следующие обозначения:

$$\sigma_\eta^2 = E(\eta - \bar{y}_x)^2 + E(\bar{y}_x - b)^2 \equiv \sigma_{\eta/x}^2 + \delta_{\eta/x}^2.$$

$\delta_{\eta/x}^2$ показывает **дисперсию линии регрессии** $\bar{y}_x = f(x)$ (широту размаха) **относительно математического ожидания b**, т.е. измеряет степень **влияния фактора** $\xi$ на фактор $\eta$.

$\sigma_{\eta/x}^2$ дает **дисперсию случайной величины** $\eta$ (разброс точек) **относительно линии регрессии** $\bar{y}_x$ и измеряет **влияние неучтенных факторов** на $\eta$.

Корреляционное отношение:

$$\Theta^2_{\eta/x} \equiv \frac{\delta^2_{\eta/x}}{\sigma^2_\eta} = 1 - \frac{\sigma^2_{\eta/x}}{\sigma^2_\eta}$$

| | | | |
|---|---|---|---|
| $\Theta^2_{\eta/x} = 1$ | при $\sigma^2_{\eta/x} = 0$ | влияние неучтенных факторов **отсутствует** | однозначная функциональная связь |
| $\Theta^2_{\eta/x} = 0$ | при $\delta^2_{\eta/x} = 0$ | влияние фактора $\xi$ на фактор $\eta$ не прослеживается | $\bar{y}_x = b + \text{const}$ |

## Алгоритм проведения корреляционного анализа

**1.** Получение статистических оценок $\bar{x}$, $\bar{y}$, $s_x$, $s_y$, $r_{xy}$ точечных характеристик a, b, $\sigma_\xi$, $\sigma_\eta$, $\rho_{\xi\eta}$ случайных величин.

**2.** Проверка гипотез о законах распределения изолированных случайных величин $\xi$ и $\eta$.

**3.** Проверка гипотезы о коррелированности случайных величин $\xi$ и $\eta$.

**4.** Проверка гипотезы о равенстве коэффициента корреляции определенному числу $\rho_{\xi\eta} = \rho_0$.

**6.** Построение доверительного интервала для коэффициента корреляции $\rho_{\xi\eta}$.

**7.** Построение совместного закона распределения системы случайных величин $\xi$ и $\eta$.

# Дисперсионный анализ

[Часть II, стр. 53 - 59]

Дисперсионный анализ (Р. Фишер, 1920 г.) – группа методов математической статистики для анализа результатов наблюдений, зависящих от **нескольких** одновременно действующих факторов.

Идея дисперсионного анализа заключается в **разбиении** общей дисперсии изучаемой случайной величины на независимые составляющие. Каждая из них характеризует влияние того или иного фактора или их взаимодействие, а их **сравнение** позволяет оценить **зна́чимость влияния** факторов на исследуемую величину.

Предположения дисперсионного анализа:

**1)** Исследуемые факторы стохастически **независимы**. С точки зрения способов отбора информации это означает независимость выборочных результатов наблюдения (отдельных выборок или слоев – они не преобразуются друг в друга с помощью какого-либо алгоритма).

**2)** Исследуемые факторы, каждый по отдельности, подчиняются **нормальным** законам распределения.

**3)** **Дисперсии** $\sigma_i^2$ исследуемых факторов **однородны** (априори приблизительно одного порядка).

Идею дисперсионного анализа о разбиении дисперсии изучим на примере однофакторного эксперимента по установлению связи выходного фактора системы ($\eta$) с одним входным фактором ($\xi$).

Входной фактор $\xi$ задается своими k уровнями, значения которых в дисперсионном анализе не существенны, важны лишь их номера: $j = 1, 2, \ldots, k$.

В однофакторном эксперименте при каждом j-ом уровне входного фактора проводится серия замеров выходного фактора. Каждый такой замер имеет номер:

$$i = 1, 2, \ldots, N_j$$

Тогда результат единичного i-го замера выходного фактора $\eta$ при **j-м уровне** входного фактора (в j-й серии наблюдений, группе, слое) можно представить в виде:

$$y_{ji} = b_j + \varepsilon_{ji},$$

где $b_j$ – математическое ожидание фактора $\eta$ при **j-м уровне** исследуемого входного фактора,

$\varepsilon_{ji}$ – погрешность наблюдения, независимые стохастические компоненты наблюдений, распределенные по единому нормальному закону с нулевым математическим ожиданием и дисперсией $\sigma^2$.

Допустим, что все предположения дисперсионного анализа выполнены:
- исследуемый (единственный входной) фактор независим;
- исследуемый фактор подчиняется нормальному закону распределения;
- единственная дисперсия входного фактора «однородна».

Гипотеза: **выходной фактор зависит от входного**, т.е. математические ожидания $b_j$ различаются значимо, тогда $b_j$ можно рассматривать как функцию от номера j уровня входного фактора):

$$b_j = \mu + T_j,$$

где $\mu$ – математическое ожидание фактора $\eta$ при **всех** уровнях исследуемого входного фактора,

$T_j$ – добавок к $\mu$ от **влияния** исследуемого входного фактора.

Таким образом, **дисперсионная модель** *однофакторного дисперсионного* анализа имеет вид:

$$y_{ji} = \mu + T_j + \varepsilon_{ji}$$

Однако ни $\mu$, ни $b_j$ известными быть не могут, вместо них можно использовать их оценки $\bar{\bar{y}}$ и $\bar{y}_j$:

$$y_{ji} = \bar{y}_j + \delta_{ji},$$

где $\delta_{ji}$ – независимые стохастические компоненты наблюдений, тоже распределенные по единому нормальному закону с нулевым математическим ожиданием и дисперсией $\sigma^2$.

Рассмотрим дисперсионную сумму квадратов отклонений в выражении несмещенной оценки общей дисперсии всего эксперимента:

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{\bar{y}})^2, \text{ где } N = \sum_{j=1}^{k} N_j .$$

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{\bar{y}})^2 = \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j + \bar{y}_j - \bar{\bar{y}})^2 =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{y}_j - \bar{\bar{y}})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)(\bar{y}_j - \bar{\bar{y}}) =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{k} N_j(\bar{y}_j - \bar{\bar{y}})^2 + 2\sum_{j=1}^{k}(\bar{y}_j - \bar{\bar{y}})\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j)$$

Но $2\sum_{j=1}^{k}(\bar{y}_j - \bar{\bar{y}})\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j) = 0$, так как $\sum_{i=1}^{N_j}(y_{ji} - \bar{y}_j) = 0$ по определению $\bar{y}_j$.

Первое слагаемое $\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2$ дает оценку рассеяния **внутри** серий наблюдений (отклонения единичных замеров от средней **внутри** серии), т.е. отражает влияние всех **неучтенных** факторов. Поэтому выражение:

$$s_0^2=\frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2$$

называется **остаточной (внутренней) дисперсией**.

Второе слагаемое $\sum_{j=1}^{k}N_j(\bar{y}_j-\bar{\bar{y}})^2$ дает оценку рассеяния **между** сериями наблюдений (отклонения средних по сериям от общего среднего), т.е. отражает **влияние** изменения входного **фактора.** Поэтому выражение:

$$s_A^2=\frac{1}{k-1}\sum_{j=1}^{k}N_j(\bar{y}_j-\bar{\bar{y}})^2$$

называется **межгрупповой дисперсией**.

**Основное уравнение дисперсионного анализа**:

$$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{\bar{y}})^2=\sum_{j=1}^{k}\sum_{i=1}^{N_j}(y_{ji}-\bar{y}_j)^2+\sum_{j=1}^{k}N_j(\bar{y}_j-\bar{\bar{y}})^2$$

или $(N-1)\cdot s^2=(N-k)\cdot s_0^2+(k-1)\cdot s_A^2$.

Если в последнем уравнении:

$$(N-1)\cdot s^2=(N-k)\cdot s_0^2+(k-1)\cdot s_A^2$$

$s_A^2 = s_0^2$, то $\underline{s=s_A^2=s_0^2}$.

Отсюда: если все выборочные данные подчиняются **одному** и тому же нормальному закону распределения (с общими математическим ожиданием и дисперсией), то различие между $s_A^2$ и $s_0^2$ должно быть **незна́чимо**.

Для подтверждения выдвинутой гипотезы о зависимости выходного фактора от единственного входного необходимо **значимое** превосходство межгрупповой дисперсии $s_A^2$ над остаточной $s_0^2$ .

**Критерий Р. Фишера**

Гипотеза: все выборочные данные по всем слоям подчиняются **одному** и тому же нормальному закону распределения (с общими математическим ожиданием и дисперсией), т.е. различие между $s_A^2$ и $s_0^2$ должно быть **незна́чимо**.

Из 13-й строки таблицы *выборочных функций* используется закон распределения $\frac{s_A^2}{s_0^2}$ Фишера: $F_{1-\alpha}(f_1,f_2)$ при вероятности $1 - \alpha$ и двух числах степеней свободы: $f_1$ для **большей** дисперсии и $f_2$ для **меньшей**.

Три исхода *критерия Р. Фишера*:

– если межгрупповая дисперсия **ЗНАЧИМО БОЛЬШЕ** остаточной:

$$\frac{s_A^2}{s_0^2} > F_{1-\alpha}(k-1, N-k),$$

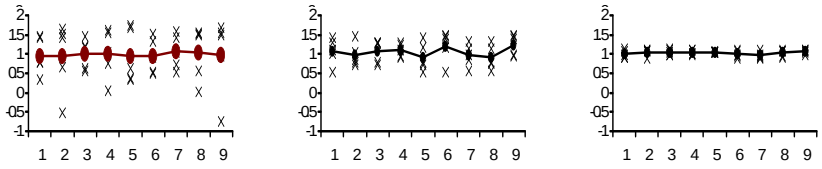
то **влияние фактора существенно** и его необходимо учитывать;

– если остаточная дисперсия **ЗНАЧИМО БОЛЬШЕ** межгрупповой:

$$\frac{s_0^2}{s_A^2} > F_{1-\alpha}(N-k, k-1),$$

то **влияние фактора несущественно** и им можно пренебречь;

– в противном случае влияние исследуемого фактора сравнимо с погрешностью эксперимента или влиянием неучтенных факторов, поэтому конкретный вывод невозможен.



а б в

а) бóльшая дисперсия – остаточная: $\frac{s_0^2}{s_A^2} = 8{,}07 > F_{1-\alpha}(N-k, k-1) = 5{,}15$ – влияние **неучтенных** факторов значительно, они "забивают" возможную зависимость от исследуемого входного фактора, признать которую нельзя.

б) бóльшая дисперсия – межгрупповая, но отношение дисперсий не достигает критического значения: $\frac{s_A^2}{s_0^2} = 1{,}21 < F_{1-\alpha}(k-1, N-k) = 3{,}04$ – уверенный вывод о влиянии или невлиянии исследуемого входного фактора сделать нельзя.

в) межгрупповая дисперсия **значимо** больше остаточной: $\frac{s_A^2}{s_0^2} = 9{,}02 > F_{1-\alpha}(k-1, N-k) = 3{,}04$ – влияние исследуемого входного фактора существенно.

Многофакторная дисперсионная модель:

$$y_{ij...m} = \mu + T_i + S_j + ... + Q_m + \varepsilon_{ij...m},$$

где $y_{ij...m}$ – результат эксперимента, в котором фактор T принял i-ый уровень, фактор S – j-ый, фактор Q – m-ый уровень.

Обеспечение предположений дисперсионного анализа

| Предположение | Меры обеспечения | Опасность |
|---|---|---|
| **Независимость** исследуемых факторов | замена факторов, корреляционный анализ, метод главных компонент, факторный анализ | бессмысленные или неверные выводы |
| **Нормальный закон** распределения факторов | перегруппировка слоев, метод главных компонент, факторный анализ | грубые, недостаточно обоснованные выводы |
| **Однородность дисперсий** в слоях | переход к новому фактору: $g(x) = q \cdot \int \frac{dx}{h(x)}$, где $h(x)$ выбирается из: $\sigma = h(a)$, а коэффициент $q = h(a) \cdot g'(a)$ | бессмысленные или неверные выводы |

# Алгоритм дисперсионного анализа

**1.** Проверка независимости (или некоррелированности) исследуемых факторов методами корреляционного анализа. Обеспечение некоррелированности.

**2.** Проверка нормального распределения исследуемых факторов по критерию согласия Пирсона. При необходимости пересмотр факторов.

**3.** Проверка однородности дисперсий по критерию Фишера. При необходимости замена факторов.

**4.** Разбиение общей дисперсии в соответствии с задачей исследований.

**5.** Вычисление необходимых межгрупповых и остаточных дисперсий и проверка гипотез о значимости их различия с помощью критерия Фишера.

**(6).** Анализ отклонений средних от общего среднего (проверка гипотезы о равенстве математических ожиданий) с помощью критерия знаков для k величин: $\frac{\overline{y}_i - \overline{\overline{y}}}{s_0}\sqrt{N_i}$, а при больших $N_i$ и k еще и проверка нормального распределения k величин (4-я или 5-я строка табл. 10 § 5.4): $\frac{\overline{y}_i - b}{\sigma}\sqrt{N_i}$ или $\frac{\overline{y}_i - b}{s}\sqrt{N_i}$.

**(7).** Если гипотеза о равенстве математических ожиданий отвергнута, то можно определить доверительные интервалы для них с помощью распределения Стьюдента с N – k степенями свободы для функции $\frac{\bar{y}_i - b_i}{s_0}\sqrt{N_i}$.

# Регрессионный анализ

[Часть II, стр. 59 - 68]

**Регрессионный анализ** предназначен для получения теоретического уравнения регрессии $\eta(\xi) = f(\xi, \lambda)$, вид которого задается, исходя **из особенностей** изучаемой системы случайных величин, а параметры $\lambda$ определяются по выборочным данным.

**Регрессия** – функциональная зависимость, аппроксимирующая (заменяющая) статистическую зависимость средних значений рассматриваемых факторов (переменных) $\overline{y}_x = f(x)$.

Регрессионная модель – регрессия для использования в математической модели исследуемого явления:

$$\mathbf{Y} = \mathbf{\Phi}(\mathbf{X}, \mathbf{Z}, \mathbf{\Lambda}) + \mathbf{E},$$

где $\mathbf{\Lambda}$ – вектор коэффициентов регрессионной модели, подлежащих определению из эксперимента.

**Ортогональными базисными функциями** $f_i(\mathbf{X}, \mathbf{Z})$ могут быть:

– тригонометрические функции $f_{(2n-1)i}(\mathbf{X}) = \cos(nx_i)$; $f_{2ni}(\mathbf{X}) = \sin(nx_i)$, если явление имеет признаки периодической ограниченной величины;

– системы ортогональных полиномов (многочленов Эрмита, Лежандра, Лагерра и т.п.);

– "полиномиальные переменные" $f_i(\mathbf{X}) = x_i;...;$ $f_{n+i}(\mathbf{X}) = x_i^2;...;$ $f_{2n+i}(\mathbf{X}) = x_{i-1}x_i;...$

Наиболее распространенная **линейная регрессионная модель** по базисным функциям **F(X, Z)** от входных факторов:

$$y = (\mathbf{\Lambda}, \mathbf{F(X, Z)}) + e = \sum_i \lambda_i \cdot f_i(\mathbf{X}, \mathbf{Z}) + e,$$

где одна выходная переменная y представлена скалярным произведением, с аддитивной (суммируемой) помехой e.

Простейшей регрессионной моделью является линейная непосредственно по входным факторам функция:

$$y = \lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k + e.$$

## Алгоритм регрессионного анализа

**1. Задание вида** линии регрессии из реальных **физических** свойств изучаемого явления: $f(\xi, \eta, \lambda)$ с неизвестными параметрами $\lambda$.

**2.** Вычисление выборочных оценок параметров предполагаемого теоретического закона распределения: $\overline{x}, \overline{y}, s_x, s_y, r_{xy}$.

**3.** Проверка гипотезы о равенстве нулю коэффициента корреляции. (Если исследуемые случайные величины $\xi$, $\eta$ некоррелированы, то следует прекратить анализ или вернуться к пункту 1.

**4.** Проверка гипотез о законах распределения исследуемых случайных величин $\xi$ и $\eta$, как изолированных. Это необходимо для следующих шагов и делается с помощью критерия согласия Пирсона $\chi^2$.

**5.** Отыскание параметров линии регрессии $\lambda$ методом наибольшего правдоподобия.

**(6).** Вычисление выборочных оценок дисперсий для разбитого на k групп (слоев) массива экспериментальных данных: остаточной $s_0^2 = \frac{1}{N-k} \cdot \sum_{j=1}^{k} (N_j - 1) s_j^2$, где $s_j^2 = \frac{1}{N_j - 1} \cdot \sum_{i=1}^{N_j} (y_{ji} - \bar{y}_j)$, и межгрупповой $s_R^2 = \frac{1}{k-\nu} \sum_{i=1}^{k} N_i [\bar{y}_i - f(x_i;\lambda)]^2$, где $\nu$ – число параметров функции регрессии, определенных из выборки.

**(7).** Проверка по критерию Фишера для $\frac{s_R^2}{s_0^2}$ с $(k - \nu, N - k)$ степенями свободы **незначимости** отличия $s_R^2$ от $s_0^2$, что характеризует такую "малость" отклонений вокруг линии регрессии, которую можно "объяснить" погрешностью самого эксперимента.

**(8).** Построение доверительных интервалов для оценок параметров распределения $\lambda$ вокруг их оценок $\lambda^*$, а также самой случайной величины $\eta$ ("коридор" вокруг эмпирической линии регрессии).

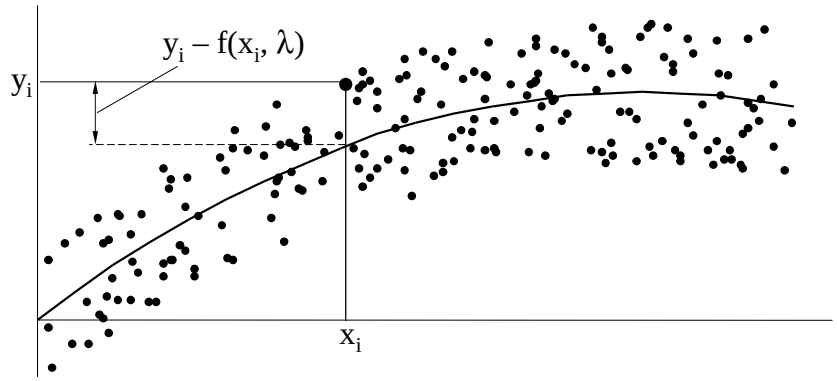
МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

- частный случай метода наибольшего правдоподобия для построения регрессии при нормальном распределении параметров

МНК имеет физический смысл и был разработан в XIX веке Лежандром и Гауссом - до метода наибольшего правдоподобия. Однако его формальное применение, без проверки нормальных законов распределения параметров, чревато систематическими ошибками в случае асимметричных распределений.

## Физический смысл МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

$$J(\boldsymbol{\lambda}) = \sum_{i=1}^{N} [y_i - f(x_i, \boldsymbol{\lambda})]^2 \Rightarrow \min$$



Параметры $\lambda_j$ регрессии $y = f(x, \boldsymbol{\lambda})$ определяются из условия минимума функции нескольких аргументов $J(\boldsymbol{\lambda})$, т.е. из системы уравнений:

$$\frac{\partial J}{\partial \lambda_j} = -2\sum_{i=1}^{N}\left[y_i - f(x_i, \boldsymbol{\lambda})\right]\cdot\frac{\partial f}{\partial \lambda_j} = 0, \qquad j = 0,1,2,...,m,$$

в которой неизвестными являются все $m + 1$ параметров $\lambda_j$ (столько же, сколько и уравнений), а все значения $x_i$ и $y_i$ известны из эксперимента.

В случае <u>полиномиальной регрессии</u> (многочленами) вида:

$$y = f(x, \boldsymbol{\lambda}) = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m = \sum_{j=0}^{m} a_j x^j$$

"система нормальных уравнений метода наименьших квадратов":

$$\sum_{j=0}^{m} a_j \cdot \sum_{i=1}^{N} x_i^{j+k} = \sum_{i=1}^{N} y_i x_i^k, \qquad k = 0,1,2,...,m,$$

которая в развернутом виде выглядит следующим образом:

$$\begin{cases} a_0 \cdot N + a_1 \cdot \sum_{i=1}^{N} x_i + a_2 \cdot \sum_{i=1}^{N} x_i^2 + ... + a_m \cdot \sum_{i=1}^{N} x_i^m = \sum_{i=1}^{N} y_i, \\ a_0 \cdot \sum_{i=1}^{N} x_i + a_1 \cdot \sum_{i=1}^{N} x_i^2 + a_2 \cdot \sum_{i=1}^{N} x_i^3 + ... + a_m \cdot \sum_{i=1}^{N} x_i^{m+1} = \sum_{i=1}^{N} y_i x_i, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_0 \cdot \sum_{i=1}^{N} x_i^m + a_1 \cdot \sum_{i=1}^{N} x_i^{m+1} + a_2 \cdot \sum_{i=1}^{N} x_i^{m+2} + ... + a_m \cdot \sum_{i=1}^{N} x_i^{2m} = \sum_{i=1}^{N} y_i x_i^m. \end{cases}$$

**ПРИМЕР.** Испытания надежности некоторой аппаратуры дали результаты, сформированные в корреляционную таблицу, в которой x задает время наладки (испытания, доработки), y – время безотказной эксплуатации конкретного прибора, а в основной части таблицы – число опытов, закончившихся с результатами x, y (время задается в %% от нормы).

| y | x | | | | |
|---|---|---|---|---|---|
| | 60 | 80 | 100 | 120 | 140 |
| 60 | 6 | 3 | - | - | - |
| 80 | 2 | 16 | 16 | 2 | 2 |
| 100 | 2 | 23 | 49 | 19 | 4 |
| 120 | - | 5 | 8 | 13 | 4 |
| 140 | - | 1 | 4 | 6 | 15 |

Предположим для начала, что исследуемая зависимость может быть представлена линейной регрессией $f(x,\mathbf{a}) = a_0 + a_1 x$. Тогда система нормальных уравнений метода наименьших квадратов приобретает вид:

$$\begin{cases} a_0 \cdot N + a_1 \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i, \\ a_0 \cdot \sum_{i=1}^{N} x_i + a_1 \cdot \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i x_i. \end{cases}$$

Из данного статистического материала:

$$N=200,\ \sum_{i=1}^{N} x_i = 20440,\ \sum_{i=1}^{N} x_i^2 = 2179200,$$

$$\sum_{i=1}^{N} y_i = 20520,\ \sum_{i=1}^{N} y_i x_i = 2148000,$$

решение системы: $a_0 = 44{,}998$, $a_1 = 0{,}56362$. Результат в этом случае годен **лишь для оценок в пределах исследованного диапазона**, ибо из всех существенных физических свойств величины времени безотказной работы отражает лишь возрастание.

Исследуем другое важное физическое свойство рассматриваемой величины – выпуклость – с помощью многочлена второй степени, имеющего это свойство: $f(x,\mathbf{a}) = a_0 + a_1 x + a_2 x^2$.

Решение системы из 3 нормальных уравнений метода наименьших квадратов для 3 коэффициентов $a_0$, $a_1$, $a_2$ в этом случае, использующем дополнительные суммы:

$\sum_{i=1}^{N} x_i^3 = 241456000,\ \sum_{i=1}^{N} x_i^4 = 27694100000,\ \sum_{i=1}^{N} y_i x_i^2 = 234176000,$

даст: $a_0 = 49{,}902$, $a_1 = 0{,}46503$, $a_2 = 0{,}00047465$, что свидетельствует о наличии в зависимости выпуклости вниз.

Замечание 1. Наличие в системе линейных алгебраических уравнений коэффициентов, различающихся на несколько порядков, существенно осложняет решение с приемлемой точностью. Поэтому следует применять масштабирование исходных параметров. В данном примере удобен один масштаб по x и y: 100, который приводит параметры к величинам, **близким** к 1.

В более общем случае для полиномиальных регрессий бывает полезно масштабирование со "сдвигом": $x = \frac{x - x_0}{m_x}$.

Замечание 2. В случае **нелинейной** регрессии бывает полезно привести исходную зависимость с помощью каких-либо замен переменных к виду, близкому к линейному или квадратичному, и уже для этой новой зависимости искать регрессию.

Сглаживание – отыскание регрессии **функциональной**, однозначной по смыслу экспериментальной зависимости (например, поляры самолета).

Аппроксимация – замена сложной функциональной зависимости более простой (например, аналитической).

*Отыскание регрессии, сглаживание и аппроксимация* – эквивалентные названия одной и той же задачи регрессионного анализа.

Погрешность аппроксимации (сглаживания) можно оценить по величине:

$$\sum_{i=1}^{N}[y_i - f(x_i,\lambda)]^2 \quad \text{или} \quad \sum_{i=1}^{N}\left|y_i - f(x_i,\lambda)\right|.$$

**ПРИМЕР 1.** Точки **поляры самолета** – зависимости $c_{xa} = f(c_{ya})$ получают из комплекса расчетов, экспериментов в аэродинамических трубах и в натурных полетах.

При полиномиальной аппроксимации чем выше степень полинома, тем он ближе к экспериментальным точкам. Но такие аппроксимации могут иметь точки перегиба и лишние экстремумы.

Поэтому необходимо учесть физические свойства поляры в интересующей нас области:

– не имеет корней,

– не имеет точек перегиба,

– вблизи $c_{уа} = 0$ имеет точку минимума.

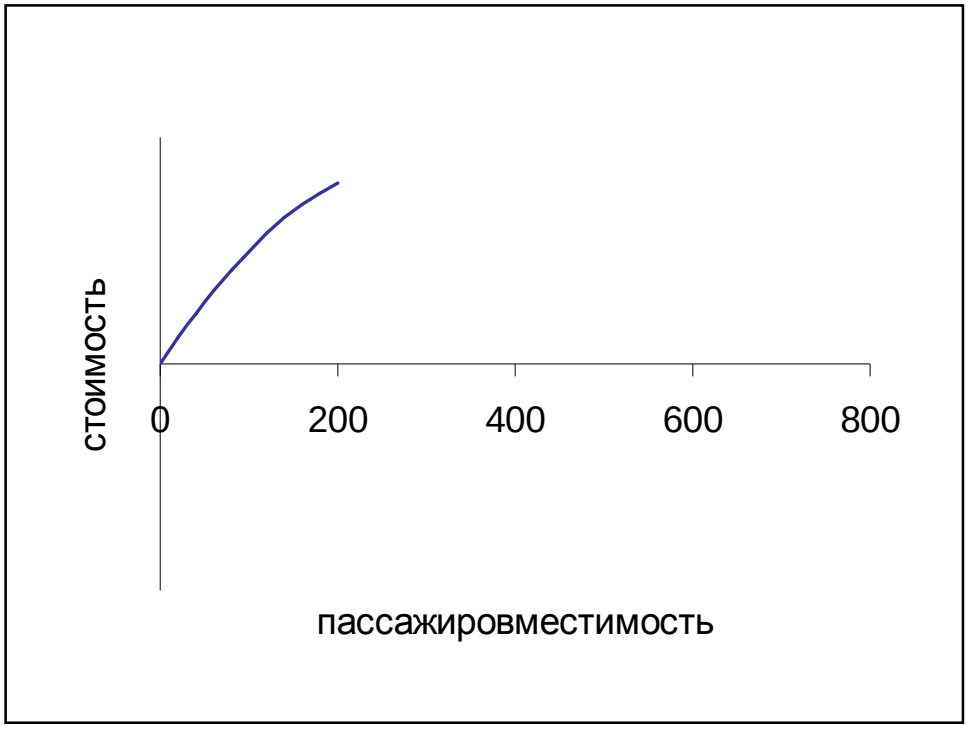
Еще более глубокие требования обнаруживаются при аппроксимации зависимости $c_{xa}$ от числа M.

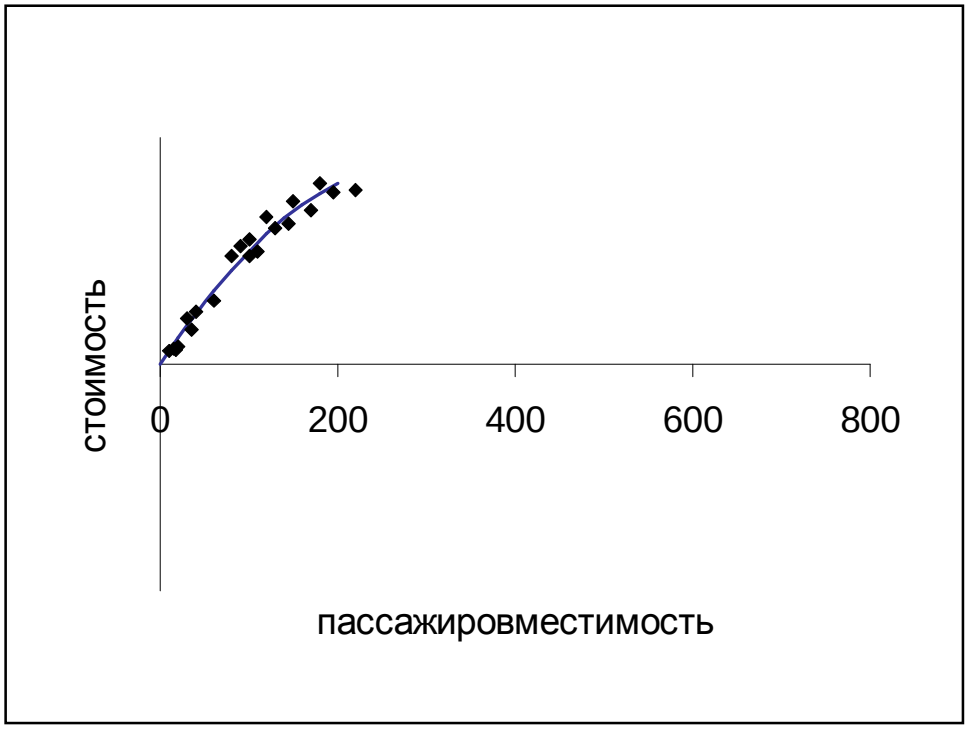
В итоге наилучший из приемлемых полиномов для поляры самолета Ту-154Б при 3 % погрешности содержит $M^0$, $M^4$ и $c_{ya}^0$, $c_{ya}^1$, $c_{ya}^2$, $c_{ya}^3$.

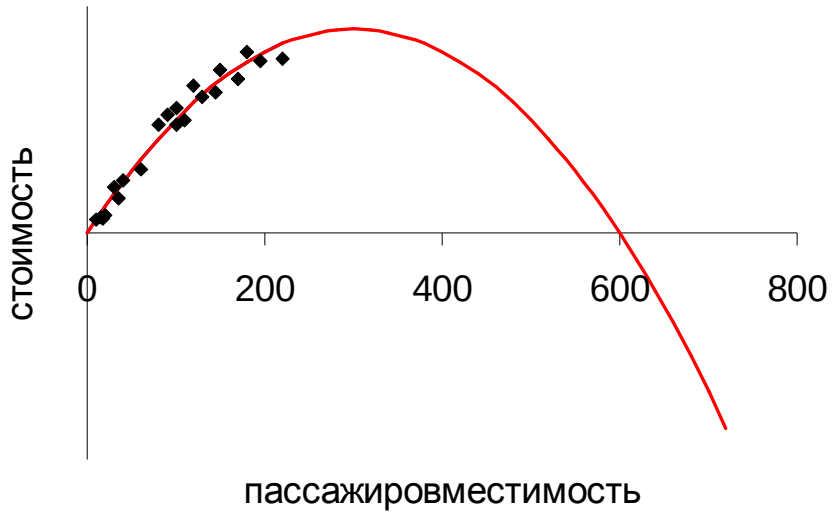
Для самолета Ту-134А – $M^0$, $M^4$, $M^8$ и $c_{ya}^0$, $c_{ya}^1$, $c_{ya}^2$, $c_{ya}^3$.

**ПРИМЕР 2.** Аппроксимация зависимости стоимости самолета C от пассажировместимости n:

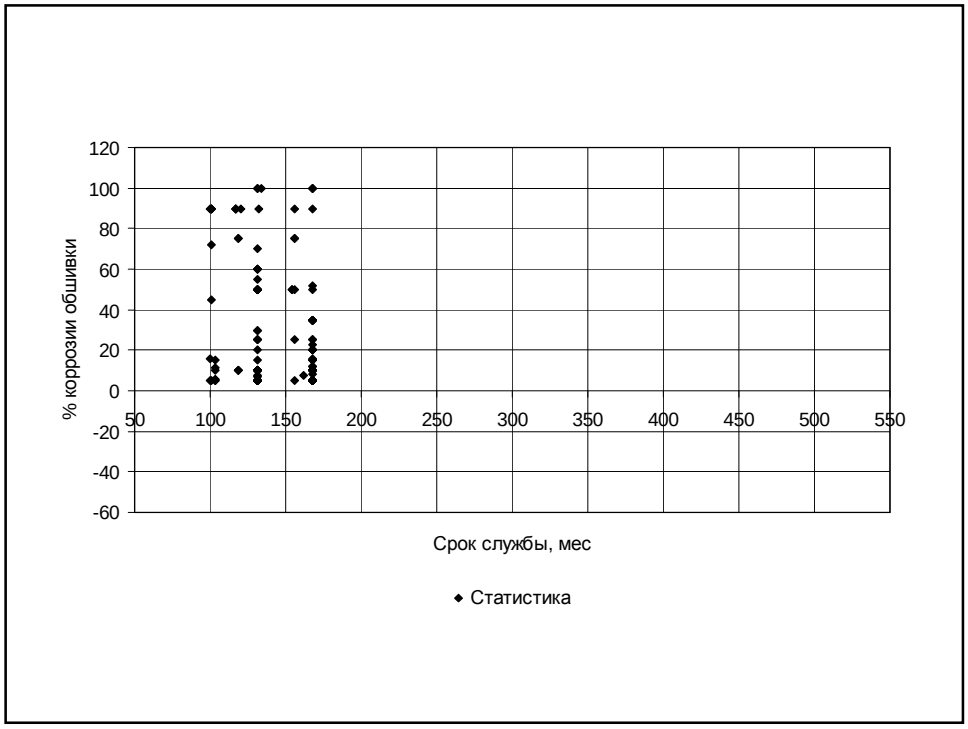
$$C = A \cdot n - B \cdot n^2.$$

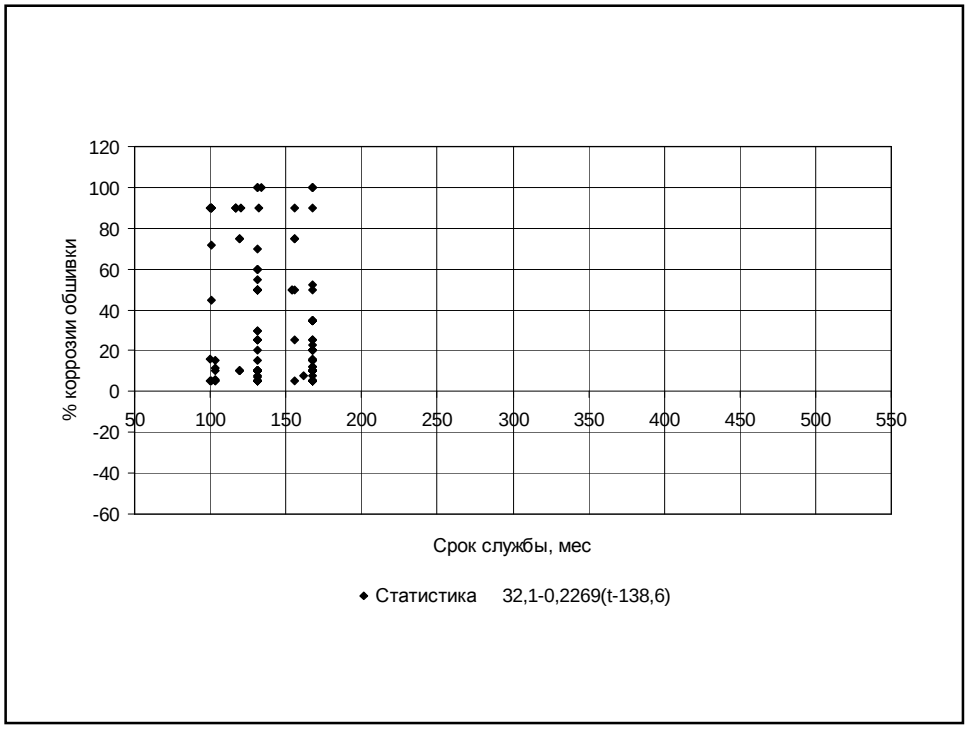

стоимость
0
200
400
600
800
пассажировместимость

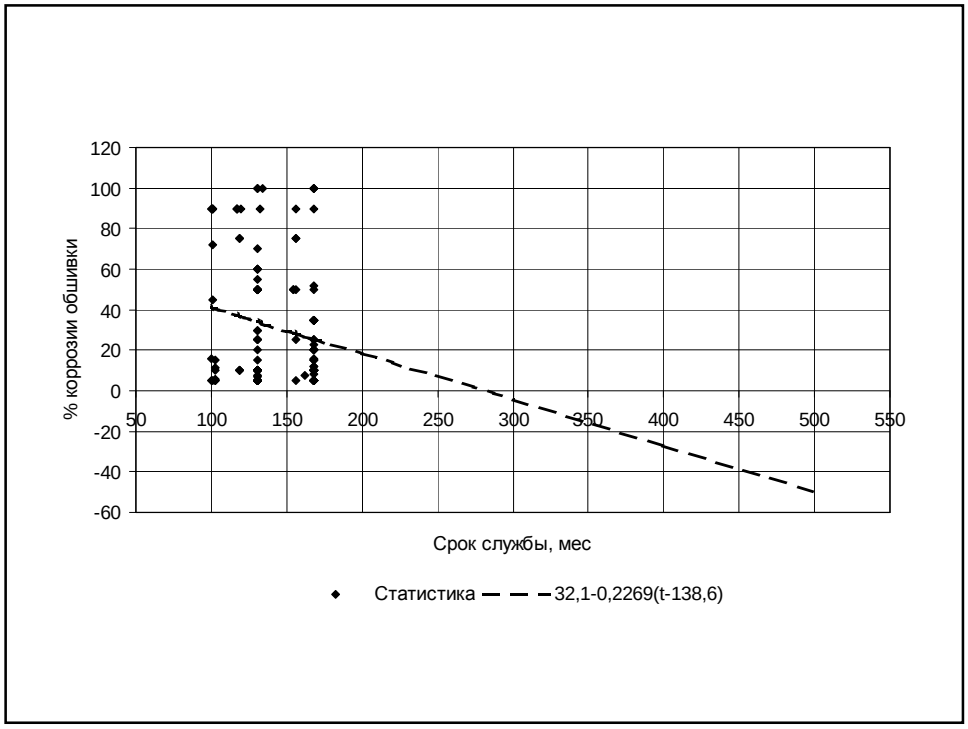

стоимость
0
200
400
600
800
пассажировместимость

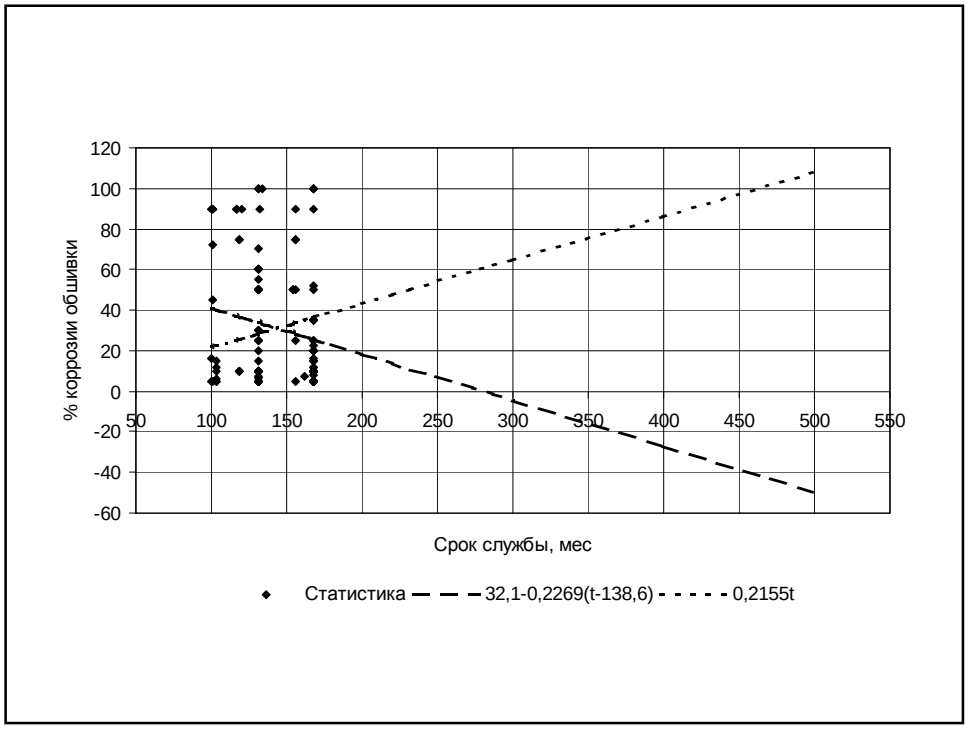



**ПРИМЕР 3.** Исследования зависимости коррозионных повреждений листов обшивки самолетов от срока службы.


120
100
80
60
40
20
0
-20
-40
-60
% коррозии обшивки
50
100
150
200
250
300
350
400
450
500
550
Срок службы, мес
Статистика

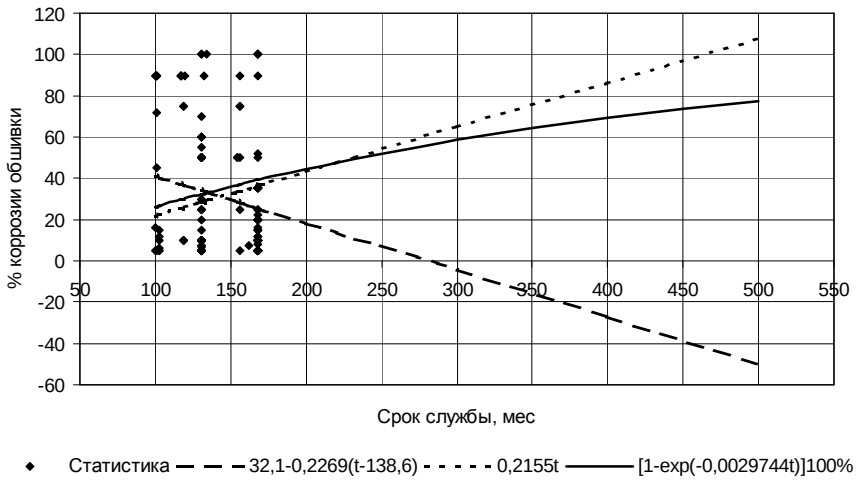

120
100
80
60
40
20
0
-20
-40
-60
% коррозии обшивки
50
100
150
200
250
300
350
400
450
500
550
Срок службы, мес
Статистика
32,1-0,2269(t-138,6)


120
100
80
60
40
20
0
-20
-40
-60
50
100
150
200
250
300
350
400
450
500
550
% коррозии обшивки
Срок службы, мес
Статистика
32,1-0,2269(t-138,6)


120
100
80
60
40
20
0
-20
-40
-60
50
100
150
200
250
300
350
400
450
500
550
% коррозии обшивки
Срок службы, мес
Статистика
32,1-0,2269(t-138,6)
0,2155t

Необходим глубокий **физичный** подход, учитывающий максимум естественных свойств зависимости:

- проходит через начало координат;
- возрастающая;
- выпукла вверх;
- асимтотически приближается к 100 % при $t \rightarrow \infty$.



# Понятие о конфлюэнтном анализе

[Часть II, стр. 68]

Конфлюэнтный анализ представляет **структуру** исследуемых случайных величин в виде двух составляющих (конфлюэнтная модель):

$$\xi = a + \xi', \qquad \eta = b + \eta',$$

где a и b математические ожидания – структурные компоненты,

$\xi'$ и $\eta'$ – стохастические компоненты (случайные) с **нулевым** математическим ожиданием.

Предполагается, что связь между факторами $\xi$ и $\eta$ определяется связью структурных компонент, а стохастические имеют характер шума.

# ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА

## Статистические методы планирования эксперимента

Проблемы построения эксперимента

[Часть II, стр. 72 - 76]

Отбор информации не объективен!

1. Результаты наблюдений - это лишь ограниченная выборка.
2. Информация собирается для определенных целей.
3. Результаты наблюдений имеют погрешность.

**Из эксперимента можно получить любые результаты!**
Поэтому для получения достоверных сведений о сложных «плохо организованных системах» необходим особый подход к организации эксперимента.

Для построения статистических математических моделей (дисперсионных, корреляционных, регрессионных) существуют математически обоснованные методы планирования эксперимента

**Планирование эксперимента** – совокупность действий, объединенных **целью** исследования и направленных на разработку **стратегии** экспериментирования от начальных до заключительных этапов изучения объекта исследований (от получения априорной информации до создания работоспособной математической модели или определения оптимальных условий)

[ГОСТ 24026–80. Исследовательские испытания. Планирование эксперимента. Термины и определения. – М.: Изд-во стандартов, 1980.].

**Эксперимент** – это система операций, воздействий и (или) наблюдений, направленных на получение информации об объекте при исследовательских испытаниях [ГОСТ 24026–80].

**Опыт** – это отдельная часть эксперимента, воспроизводящая исследуемое явление в определенных задаваемых условиях при возможности регистрации его результатов.

**Пассивный** эксперимент – эксперимент в отсутствии управляемых факторов: выходные факторы зависят только от неуправляемых входных и неконтролируемых факторов (шума).

**Активный** эксперимент – эксперимент в отсутствии неуправляемых входных факторов: выходные факторы зависят только от управляемых входных и неконтролируемых факторов (шума).

**Планом эксперимента** является некоторая **совокупность** уровней факторов **X**, построенная для определенных целей исследования.

**Матрица плана эксперимента** (строки отвечают опытам, а столбцы – факторам: элемент $x_{ij}$ матрицы плана обозначает уровень j-го фактора в i-м опыте):

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & u_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nk} \end{pmatrix}.$$

**k-факторным** называется эксперимент, в котором результат рассматривается в зависимости от изменения k контролируемых управляемых входных факторов (вектор **X** имеет размерность k).

Если все исследуемые факторы имеют в эксперименте одинаковое количество h уровней, то такой эксперимент называется **h-уровневым**.

Для наиболее экономичного получения достоверного результата эксперимента, удовлетворяющего предъявленным требованиям, необходимо решать ряд проблем

**<u>Проблемы постановки эксперимента</u>**:

*1) собственно эксперимент*:

а) формулировка целей и задач эксперимента,

б) выбор наблюдаемого выходного фактора,

в) выбор управляемых факторов,

г) выбор уровней этих факторов (количественных или качественных, фиксированных или случайных),

д) подбор сочетаний уровней факторов;

*2) планирование эксперимента:*

а) определение необходимого числа опытов,

б) определение порядка проведения отдельных опытов,

в) выбор метода рандомизации,

г) составление математической модели для описания результатов;

*3) анализ результатов эксперимента:*

а) сбор и обработка данных,

б) вычисление статистик (выборочных функций) для проверки гипотез,

в) интерпретация результатов эксперимента.

Для успешного решения 1-ой из этих проблем применяется **теория моделирования**, 3-ой - методы **обработки информации**, 2-ой и частично 1-ой - **планирование эксперимента**

## Принципы планирования эксперимента

(позволяют сделать эксперимент практически реализуемым)

**1. Принцип отказа от полного перебора всех возможных входных состояний**.

**2. Принцип последовательного усложнения математической модели (принцип последовательного планирования)**.

Е.С. Вентцель: "Основной принцип теории планирования эксперимента состоит в том, что любое принятое заранее решение должно пересматриваться с учетом полученной новой информации".

**3. Принцип сопоставления с шумом**.

Бессмысленно ставить дорогостоящий эксперимент для получения точной модели, если результаты эксперимента обладают большой погрешностью (зашумлены).

**4. Принцип рандомизации** (принцип приведения к случайности влияния факторов).

Рандомизация – это обеспечение случайности влияния действующих на систему факторов, не поддающихся или поддающихся с трудом учету и контролю.

**5. Принцип оптимальности плана** (наличие критерия оптимальности, например, D-оптимальный – минимизирует обобщенную дисперсию, A-оптимальный – минимизирует сумму дисперсий, E-оптимальный – минимизирует наибольшую дисперсию).

## Назначение плана эксперимента

[Часть II, стр. 76 - 79]

**Цель** *планирования эксперимента* – получение максимума достоверной информации при минимуме затрат.

***Достоверность*** и компактность информации можно обеспечить с помощью *дисперсионных* и *регрессионных* математических *моделей*, обладающих свойствами *эффективности, состоятельности и несмещенности*.

Эффективность статистической оценки тем выше, чем меньше ее дисперсия. А дисперсией можно **управлять** условиями постановки эксперимента, его погрешностью и его **планом**.

*Достоверная* и наиболее компактная информация

⇓

Наилучшие *статистические* математические *модели*

⇓

*Эффективные, состоятельные и несмещенные* оценки параметров моделей

⇓

Оценки параметров моделей с минимальной дисперсией

⇓

Управление общей дисперсией через дисперсии от условий постановки эксперимента, его погрешности и от его плана

# Задача о взвешивании трех арбузов

Естественный план

| № опыта | A | B | C | результат |
|---|---|---|---|---|
| 0 | **–1** | **–1** | **–1** | $y_0$ |
| 1 | **+1** | **–1** | **–1** | $y_1$ |
| 2 | **–1** | **+1** | **–1** | $y_2$ |
| 3 | **–1** | **–1** | **+1** | $y_3$ |

Естественный план

| № опыта | A | B | C | результат |
|---|---|---|---|---|
| 0 | **–1** | **–1** | **–1** | $y_0$ |
| 1 | **+1** | **–1** | **–1** | $y_1$ |
| 2 | **–1** | **+1** | **–1** | $y_2$ |
| 3 | **–1** | **–1** | **+1** | $y_3$ |

$$\sigma^2\{A\} = \sigma^2\{y_1 - y_0\} = 2\sigma^2\{y\}$$

Улучшенный план

| № опыта | A | B | C | результат |
|---|---|---|---|---|
| 0 | **+1** | **+1** | **+1** | $y_4$ |
| 1 | **+1** | **–1** | **–1** | $y_1$ |
| 2 | **–1** | **+1** | **–1** | $y_2$ |
| 3 | **–1** | **–1** | **+1** | $y_3$ |

Улучшенный план

| № опыта | A | B | C | результат |
|---|---|---|---|---|
| 0 | **+1** | **+1** | **+1** | $y_4$ |
| 1 | **+1** | **–1** | **–1** | $y_1$ |
| 2 | **–1** | **+1** | **–1** | $y_2$ |
| 3 | **–1** | **–1** | **+1** | $y_3$ |

$$\sigma^2\{A\}=\sigma^2\left\{\frac{1}{2}(y_1-y_2-y_3+y_4)\right\}=\frac{1}{4}4\sigma^2\{y\}=\sigma^2\{y\}$$

Для определения веса каждого предмета по первому плану требуется произвести только **2** опыта, зато по второму плану все **4** опыта участвуют в определении веса каждого предмета.

Таким образом, стоящее в знаменателе формулы дисперсии число степеней свободы выросло с 2 до 4.

Первый план фактически распадается на три отдельных плана 1-факторных 2-уровневых экспериментов, а второй оказывается планом 3-факторного 2-уровневого эксперимента.

Прием улучшения плана предыдущего примера является по сути приемом **рандомизации.**

*Полностью рандомизированным* планом может быть только план, в котором обеспечивается полный перебор всевозможных сочетаний всех уровней всех факторов

План, в котором обеспечивается полный перебор всевозможных сочетаний всех уровней всех факторов называется **полным планом**

Эксперимент, в котором обеспечивается полный перебор всевозможных сочетаний всех уровней всех факторов называется **полным факторным экспериментом**

**<u>Объем</u>** полного **<u>плана</u>** – *количество* необходимых *опытов*:

$$h_1 \times h_2 \times ... \times h_k = \prod_{j=1}^{k} h_j ,$$

где k – число факторов, а $h_j$ – число уровней каждого j-го фактора.

Неполный план не может быть полностью рандомизированным. Он может иметь лишь некоторые **признаки** рандомизации.

**Рандомизация** – это обеспечение случайности влияния действующих на систему факторов, не поддающихся или поддающихся с трудом учету и контролю.

В рандомизированной системе влияние действующих факторов можно считать вполне случайным, что позволяет учитывать их статистически.

Примитивнейший прием рандомизации плана эксперимента состоит в случайном переборе уровней факторов.

Если уровни исследуемых факторов распределены в плане **симметрично**, т.е. все встречаются одинаковое число раз, то такой план называется **сбалансированным**.

Полный план всегда сбалансирован.

Группы опытов в плане, объединенных каким-либо общим свойством, называются **блоками**.

Если блоки неполные, то такой *план* называется **неполноблочным**.

## Латинские квадраты

(в каждой строке и в каждом столбце размещаются только **разнотипные** элементы и только по **одному** разу)

| | | |
|---|---|---|
| A | B | C |
| B | C | A |
| C | A | B |

| | | | | |
|---|---|---|---|---|
| A | B | C | D | E |
| B | C | D | E | A |
| C | D | E | A | B |
| D | E | A | B | C |
| E | A | B | C | D |

При исследовании сложных систем некоторые неслучайные факторы невозможно учесть, что приводит к систематическим погрешностям.

В этих случаях помогает прием *рандомизации*, который **переводит** такие факторы в разряд случайных.

Полная противоположность "хорошо организованным системам", где эксперимент надо было "очистить" от "посторонних", "мешающих" факторов.

## Планирование объема эксперимента

[Часть II, стр. 79 - 84]

Рассмотрим несколько подходов к построению эксперимента по определению с заданной погрешностью значения некоторого параметра x наблюдаемого явления.

**Метод А.** Простейший подход к планированию объема эксперимента выражается известной поговоркой: "Семь раз отмерь, один – отрежь!":

из известного соотношения $\sigma\{\bar{x}\} = \frac{\sigma}{\sqrt{N}}$

необходимое число опытов для обеспечения требуемой погрешности $\delta$:

$$N = \frac{\sigma^2}{\delta^2},$$

при этом **не существенен закон** распределения x.

**Метод Б.** Симметричный доверительный интервал для математического ожидания контролируемого параметра.

Для **нормально** распределенного параметра погрешность его средней выборочной оценки (радиус доверительного интервала, половина его):

$$\delta = u_{0,5\gamma}\frac{\sigma}{\sqrt{N}}$$

при априори известном значении $\sigma$ по 4-й строке таблицы выборочных функций

$$\delta = t_{\gamma}\frac{s}{\sqrt{N}},$$

при неизвестном $\sigma$, но найденной выборочной несмещенной его оценке s по 5-й строке таблицы выборочных функций

Отсюда вытекают формулы для определения объема эксперимента, необходимого для обеспечения погрешности $\delta$ с доверительной вероятностью $\gamma$:

$$N = \frac{u_{0,5\gamma}^2\sigma^2}{\delta^2} \quad \text{или} \quad N = \frac{t_{\gamma}^2 s^2}{\delta^2}.$$

**Метод В.** "Простейший критериальный" подход с помощью проверки гипотезы $H_0$: $a = a_0$ о том, что математическое ожидание a имеет значение $a_0$.

Из 4-ой и 5-ой строк таблицы выборочных функций:

$$u = \frac{\bar{x} - a_0}{\sigma/\sqrt{N}} \quad \text{и} \quad t = \frac{\bar{x} - a_0}{s/\sqrt{N}}$$

имеют нормальное и, соответственно, распределение Стьюдента.

Объем эксперимента, необходимый для проверки гипотезы $H_0$: $a = a_0$ при возможных конкурирующих гипотезах, представлен следующей таблицей.

| Конкурирующая гипотеза | Требуемый объем эксперимента | |
|---|---|---|
| | при известном $\sigma$ | при неизвестном $\sigma$ |
| $a \neq a_0$ | $u^2_{\frac{1}{2}-\frac{\alpha}{2}} \times \frac{\sigma^2}{(\bar{x}-a_0)^2}$ | $t^2_{1-\frac{\alpha}{2}} \times \frac{s^2}{(\bar{x}-a_0)^2}$ |
| $a > a_0$ | $u^2_{\frac{1}{2}-\alpha} \times \frac{\sigma^2}{(\bar{x}-a_0)^2}$ | $t^2_{1-\alpha} \times \frac{s^2}{(\bar{x}-a_0)^2}$ |
| $a < a_0$ | $u^2_{\frac{1}{2}-\alpha} \times \frac{\sigma^2}{(\bar{x}-a_0)^2}$ | $t^2_{\alpha} \times \frac{s^2}{(\bar{x}-a_0)^2}$ |
| $a = a_1$ | $(u_{\frac{1}{2}-\beta} + u_{\frac{1}{2}-\alpha})^2 \times \frac{\sigma^2}{(a_1-a_0)^2}$ | $(t_{1-\beta} + t_{1-\alpha})^2 \times \frac{s^2}{(a_1-a_0)^2}$ |
| Проверяемая гипотеза $a = a_0$<br>$\alpha$ – вероятность ошибки I рода (отвергнуть верную гипотезу)<br>$\beta$ – вероятность ошибки II рода (принять неверную гипотезу)<br>Значения вероятностей $\gamma$ в функции $u_\gamma$ соответствуют таблице стандартизованного нормального закона распределения вероятностей из [18]. | | |

**ПРИМЕР.** Требуется с помощью летных испытаний выяснить возможность приема самолета нового типа на аэродроме с располагаемой посадочной дистанцией $L_{а/д}$ = 1500 м. Разброс посадочных дистанций на самолете-прототипе составляет 300 м. Сколько посадок необходимо произвести в летных испытаниях, чтобы обеспечить ответ на поставленный вопрос с погрешностью $\delta$ = 50 м при допустимой вероятности ошибки $\alpha$ = 0,1 %?

Оценка среднего квадратического отклонения $\sigma$ при одном опыте (посадке).

Предположим в первом приближении, что посадочная дистанция распределена по нормальному закону. По правилу $3\sigma$ (вероятность попадания нормально распределенной случайной величины в интервал радиусом $3\sigma$ вокруг математического ожидания равна 0,9973):

$$\sigma \approx \frac{300\text{м}}{2 \cdot 3} = 50 \text{ м}.$$

**1 подход** – "тривиальный" по методу **А**.

По правилу $3\sigma$ требуемая погрешность:

$$50 \text{ м} = 3\sigma(N) = \frac{3\sigma}{\sqrt{N}}, \text{ откуда } N = 3^2 = 9 \text{ посадок}.$$

Исходя из правила $3\sigma$, вероятность ошибки при этом можно оценить величиной 0,27 %, что существенно хуже требуемой заданием вероятности ошибки.

**2 подход** – "формальный доверительный" по методу **Б**.

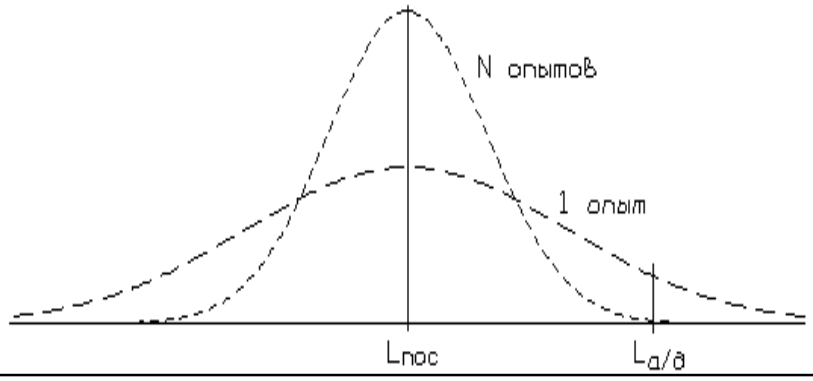
Для симметричного доверительного интервала $\gamma = 1 - \alpha$, что дает $u_{0,5\gamma} = 3,29$ и для обеспечения заданной погрешности ($\delta$ – радиус интервала) при заданной доверительной вероятности приводит к соотношению:

$$N > \frac{u_{0,5\gamma}^2 \sigma^2}{\delta^2} = \frac{3,29^2 50^2}{50^2} = 10,82.$$

Таким образом, при таком подходе требуется провести летный эксперимент из 11 посадок.

**3 подход** – "неформальный доверительный" – построение одностороннего доверительного интервала.

Область риска: $x > L_{а/д}$ с вероятностью $\alpha = 0,1$ %, т.е. $\gamma = 0,5 - \alpha$.



Предполагая, что центр распределения расположен левее $L_{а/д}$ на 50 м (для обеспечения запаса), т.е. $a_0 = L_{пос} = L_{а/д} - \delta$, среднее арифметическое значение $\overline{L}$ посадочных дистанций при N посадках должно удовлетворять условию:

$$P\left\{\overline{L} > a_0 + \delta = L_{а/д}\right\} = 0{,}001.$$

После простых преобразований дополнительная к этой вероятность:

$$P\left\{\frac{\overline{L} - a_0}{\sigma}\sqrt{N} < \frac{\delta}{\sigma}\sqrt{N}\right\} = 0{,}999 =$$

$$= P\left\{\frac{\overline{L} - a_0}{\sigma}\sqrt{N} < 0\right\} + P\left\{0 < \frac{\overline{L} - a_0}{\sigma}\sqrt{N} < \frac{\delta}{\sigma}\sqrt{N}\right\} = 0{,}5 + 0{,}499$$

с помощью таблицы функции Лапласа (раздел 8) по значению функции 0,499 дает значение аргумента:

$$\frac{\delta}{\sigma}\sqrt{N} > 3{,}09, \qquad \text{т.е.} \qquad N > \frac{3{,}09^2 50^2}{50^2} = 9{,}55.$$

Таким образом, при этом способе оценки объема эксперимента необходимо произвести 10 посадок.

**4 подход** – "формальный критериальный" по методу **В**.

Проверяемая гипотеза: $a_0 = L_{пос} = L_{а/д} - \delta$ (для обеспечения запаса). При конкурирующей гипотезе о том, что истинное значение посадочной дистанции отличается от допустимой $a_0$ в любую сторону более, чем на требуемую погрешность: $|L_{пос} - a_0| > \delta$, руководствуемся 1-й строкой таблицы объемов эксперимента. Тогда необходимое число посадок составит (при $\alpha = 0{,}001$): $N = 11 > 10{,}82$, как и при "формальном доверительном" 2-м подходе.

**5 подход** – "неформальный критериальный".

Та же проверяемая гипотеза: $a_0 = L_{пос} = L_{а/д} - \delta$. При конкурирующей гипотезе о том, что истинное значение посадочной дистанции превосходит допустимую $a_0$ более, чем на требуемую погрешность: $L_{пос} - a_0 > \delta$, руководствуемся 2-й строкой таблицы объемов эксперимента. Тогда необходимое число посадок составит (при $\alpha = 0{,}001$): $N = 10 > 9{,}55$, как и при "неформальном доверительном" 3-м подходе.

**6 подход.** По **одному** единственному испытанию с $L_{пос} = x_1 = 1410$ м.

Гипотеза $H_0$: $a = a_0 = x_1 = 1410$ м.

Пусть

– альтернативной гипотезой будет $H_1$: $a > x_1 = 1410$ м;

– вероятность ошибки $\alpha = 0{,}001$ – ошибки I рода;

– распределения близко к нормальному с $\sigma \approx 50$ м.

По таблице функции Лапласа для стандартизованной величины $\frac{x^* - a_0}{\sigma}$ находим ее значение 3,09, соответствующее вероятности $0{,}5 - \alpha = 0{,}499$.

Граница критической области:

$$x^* = 3{,}09 \cdot \sigma + x_1 = 1565 \text{ м}.$$

В этом случае расположения критической области при очевидном 1410 м < 1565 м следует вывод о приемлемости гипотезы $H_0$, т.е. о возможности эксплуатации этого типа самолета на данном аэродроме.

Однако настораживают промежуточные результаты: граница критической области находится в зоне недопустимых значений $x^* = 1565$ м > $L_{а/д} = 1500$ м!

**7 подход.** Как в 6 подходе, но с альтернативной гипотезой: $H_1$: $a + a_1$.

Остаются под вопросом лишь значения $a_1$ и $\beta$.

Заданная погрешность $\delta = 50$ м ~ величина уверенного (с некоторой вероятностью) различения двух значений $L_{пос}$. Тогда в качестве $a_1$ следует рассматривать: $a_1 = x_1 + \delta = (1410 + 50)$ м = 1460 м.

$\beta$ не задано, но эта ошибка второго рода (ошибочное принятие неверной гипотезы) много более опасна, т.е. следует назначить $\beta$ существенно меньше, чем $\alpha$, например, $\beta = 0{,}0001$.

Это 4-й случай статистических критериев и граница критической области x* должна находиться между $a_0$ и $a_1$:

$$u_{0,5-\alpha} = \frac{x^* - a_0}{\sigma} \quad \text{и} \quad u_{0,5-\beta} = \frac{a_1 - x^*}{\sigma}.$$

Однако эти соотношения невыполнимы: $u_{0,5-\alpha} = 3{,}09$, $u_{0,5-\beta} = 3{,}72$, а правые части принимают значения, безусловно меньшие 1, так как x* располагается внутри отрезка длиной 50 м, чему равен знаменатель.

**Необходимо собрать другой статистический материал**.

**8 подход.** Предыдущая постановка задачи, но с данными N посадок.

Искомое $L_{пос} = \overline{L}$. Тогда последние соотношения из предыдущего подхода примут вид:

$$u_{0,5-\alpha} = \frac{x^* - a_0}{\sigma/\sqrt{N}} \quad \text{и} \quad u_{0,5-\beta} = \frac{a_1 - x^*}{\sigma/\sqrt{N}},$$

и из них, исключая x*, можно получить выражения для необходимого объема эксперимента N:

$$N > (u_{0,5-\beta} + u_{0,5-\alpha})^2 \times \frac{\sigma^2}{(a_1 - a_0)^2},$$

а исключая $\sigma$, положение границы критической области, но только после задания $a_0$:

$$x^* = \frac{u_{0,5-\alpha} \cdot a_1 + u_{0,5-\beta} \cdot a_0}{u_{0,5-\alpha} + u_{0,5-\beta}}.$$

Расчеты для данного случая дают $N > 46{,}38$.

На практике: по результатам 47 посадок вычислить среднюю величину посадочной дистанции $\overline{L}$ и принять ее в качестве $a_0$. Если окажется $\overline{L} < x^* < L_{а/д} - \delta$, то нет оснований отвергать гипотезу $H_0$: $a = a_0$, т.е. можно разрешить эксплуатацию нового типа самолета на данном аэродроме.

Этот пример показывает, что на практике необходимо каждый раз только из условий конкретной задачи исследований **выбирать приемлемые подходы** к планированию эксперимента.

# Планирование однофакторного эксперимента

[Часть II, стр. 85 - 87]

**ПРИМЕР.** Исследование влияния срока хранения на качество ГСМ.

Подразумевается существование такого измеримого параметра y, который характеризует качество ГСМ и который может зависеть от единственного входного фактора времени T. Так как измерительная аппаратура имеет погрешность, то для исследования влияния срока хранения на параметр y необходимо производить многократные рандомизированные его замеры при различных сроках хранения, например, по n замеров в течение нескольких (m) месяцев.

Таким образом, требуется составить план однофакторного эксперимента для дисперсионного анализа

Дисперсионная математическая модель явления:

$$y_{ji} = \mu + T_j + \varepsilon_{ji},$$

План однофакторного эксперимента

| № замера ( i ) | Значения фактора – уровни ( j ) | | | |
|---|---|---|---|---|
| | 1 | 2 | ... | m |
| 1 | $y_{11}$ | $y_{21}$ | ... | $y_{m1}$ |
| 2 | $y_{12}$ | $y_{22}$ | ... | $y_{m2}$ |
| ... | ... | ... | ... | ... |
| n | $y_{1n}$ | $y_{2n}$ | ... | $y_{mn}$ |
| Средние по уровням | $\bar{y}_1$ | $\bar{y}_2$ | ... | $\bar{y}_m$ |

## Унифицированный вид плана однофакторного m-уровневого эксперимента

| № опыта | Уровни фактора | Результат замера |
|---|---|---|
| 1 | **1** | $y_{11}$ |
| 2 | **1** | $y_{12}$ |
| ... | **...** | **...** |
| n | **1** | $y_{1n}$ |
| n + 1 | **2** | $y_{21}$ |
| ... | **...** | **...** |
| $m \times n$ | **m** | $y_{mn}$ |

Этот полностью рандомизированный план позволяет провести дисперсионный анализ для ответа на вопрос о существенности влияния времени на качество ГСМ.

Статистическая оценка общей дисперсии:

$$s^2 = \frac{1}{mn-1} \sum_{j=1}^{m} \sum_{i=1}^{n} (y_{ji} - \bar{\bar{y}})^2,$$

где $\bar{\bar{y}} = \frac{1}{m} \sum_{j=1}^{m} \bar{y}_j$, а $\bar{y}_j = \frac{1}{n} \sum_{i=1}^{n} y_{ji}$.

**Остаточная дисперсия**:

$$s_0^2 = \frac{1}{m(n-1)} \cdot \sum_{j=1}^{m} \sum_{i=1}^{n} (y_{ji} - \bar{y}_j)^2 .$$

**Межгрупповая дисперсия**:

$$s_A^2 = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{y}_j - \bar{\bar{y}})^2 .$$

Критерий Фишера для сравнения двух дисперсий:

– если $\frac{s_A^2}{s_0^2} > F_{1-\alpha}[m-1, m(n-1)]$,

то существует выраженная зависимость y от T;

– если $\frac{s_0^2}{s_A^2} > F_{1-\alpha}[m(n-1), m-1]$,

то влияние фактора T на y несущественно;

– в противном случае конкретный вывод невозможен.

# Планирование двухфакторного эксперимента

[Часть II, стр. 87 - 88]

**ПРИМЕР.** Требуется спланировать эксперимент для дисперсионного анализа влияния на ходовые качества новой марки шин типа автомобиля и места установки шины на автомобиль. Выделено 12 шин и 4 типа автомобиля: I, II, III, IV с 4 местами установки: 1, 2, 3, 4.

В этом эксперименте предполагается наличие двух существенных факторов: T – автомобиль и S – место по 4 уровня. Легко видеть, что рассматриваемые факторы независимы, поэтому можно не исследовать третий фактор их совместного влияния.

Математическая дисперсионная модель:

$$y_{ijk} = \mu + T_j + S_i + \varepsilon_{ijk},$$

где $y_{ijk}$ – результат испытания шины на i-м месте j-го автомобиля;

$\mu$ – средний ожидаемый результат (математическое ожидание);

$T_j$ – отклонение от $\mu$, обусловленное влиянием типа автомобиля;

$S_i$ – отклонение от $\mu + T_j$, обусловленное влиянием места установки шины на автомобиле;

$\varepsilon_{ijk}$ – нормально распределенная погрешность оценки результата эксперимента, обусловленная влиянием неучтенных факторов.

Эксперимент получается 2-факторный 4-уровневый, полный объем которого составляет:

$$\prod_{i=1}^{k} h_i = 4 \times 4 = 16,$$

и 12 шин не позволяют построить полный план.

Эксперимент получается 2-факторный 4-уровневый, полный объем которого составляет:

$$\prod_{i=1}^{k} h_i = 4 \times 4 = 16,$$

и 12 шин не позволяют построить полный план.

Элементарные рассуждения приводят к *неполноблочному сбалансированному плану двухфакторного четырехуровневого рандомизированного эксперимента*.

# План двухфакторного эксперимента

| № места (i) | № автомобиля (j) | | | | |
|---|---|---|---|---|---|
| | I | II | III | IV | |
| 1 | **+1** | **–1** | **+1** | **+1** | $S_1$ |
| 2 | **–1** | **+1** | **+1** | **+1** | $S_2$ |
| 3 | **+1** | **+1** | **+1** | **–1** | $S_3$ |
| 4 | **+1** | **+1** | **–1** | **+1** | $S_4$ |
| | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $\mu$ |

# Унифицированный вид плана двухфакторного эксперимента

| № опыта (шины) | Факторы | |
|---|---|---|
| | № автомобиля | № места |
| 1 | **I** | **1** |
| 2 | **I** | **3** |
| 3 | **I** | **4** |
| 4 | **II** | **2** |
| 5 | **II** | **3** |
| 6 | **II** | **4** |
| 7 | **III** | **1** |
| 8 | **III** | **2** |
| 9 | **III** | **3** |
| 10 | **IV** | **1** |
| 11 | **IV** | **2** |
| 12 | **IV** | **4** |

# Планирование многофакторного эксперимента

[Часть II, стр. 89 - 92]

**ПРИМЕР.** Требуется спланировать эксперимент для дисперсионного анализа влияния на ходовые качества 4 новых марок шин: A, B, C, D (по 4 штуки каждой марки) типа автомобиля и места установки шины на автомобиль. Выделено 12 шин и 4 типа автомобиля: I, II, III, IV с 4 местами установки: 1, 2, 3, 4.

В этом эксперименте предполагается наличие трех существенных факторов: T – автомобиль, S – место установки шины и Q – марки шины по 4 уровня. Легко видеть, что рассматриваемые факторы независимы, поэтому можно не исследовать факторы их совместного влияния.

Математическая дисперсионная модель:

$$y_{ijmk} = \mu + T_j + S_i + Q_m + \varepsilon_{ijmk},$$

где $y_{ijk}$ – результат испытания шины на i-м месте j-го автомобиля;

$\mu$ – средний ожидаемый результат (математическое ожидание);

$T_j$ – отклонение от $\mu$, обусловленное влиянием типа автомобиля;

$S_i$ – отклонение от $\mu + T_j$, обусловленное влиянием места установки шины на автомобиле;

$Q_m$ – отклонение от $\mu + T_j + S_i$, обусловленное влиянием m-й марки шины;

$\varepsilon_{ijk}$ – нормально распределенная погрешность оценки результата эксперимента, обусловленная влиянием неучтенных факторов.

Эксперимент получается 3-факторный 4-уровневый, полный объем которого составляет:

$$\prod_{i=1}^{k} h_i = 4 \times 4 \times 4 = 64,$$

и 16 шин не позволяют построить полный план.

Эксперимент получается 3-факторный 4-уровневый, полный объем которого составляет:

$$\prod_{i=1}^{k} h_i = 4 \times 4 \times 4 = 64,$$

и 16 шин не позволяют построить полный план.

Но ими можно полностью оснастить все автомобили. В этом случае обеспечить *сбалансированность* и *рандомизированность* плана позволяют *латинские квадраты*.

## План трехфакторного эксперимента

| № места (i) | № автомобиля (j) | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| 1 | **A** | **B** | **C** | **D** |
| 2 | **D** | **A** | **B** | **C** |
| 3 | **C** | **D** | **A** | **B** |
| 4 | **B** | **C** | **D** | **A** |

## Унифицированный вид плана трехфакторного эксперимента

| № опыта (шины) | Факторы | | |
|---|---|---|---|
| | № автомобиля | № места | № марки шины |
| 1 | **I** | **1** | **A** |
| 2 | **I** | **2** | **D** |
| 3 | **I** | **3** | **C** |
| 4 | **I** | **4** | **B** |
| 5 | **II** | **1** | **B** |
| 6 | **II** | **2** | **A** |
| 7 | **II** | **3** | **D** |
| 8 | **II** | **4** | **C** |
| 9 | **III** | **1** | **C** |
| 10 | **III** | **2** | **B** |
| 11 | **III** | **3** | **A** |
| 12 | **III** | **4** | **D** |
| 13 | **IV** | **1** | **D** |
| 14 | **IV** | **2** | **C** |
| 15 | **IV** | **3** | **B** |
| 16 | **IV** | **4** | **A** |

## Планы трехфакторного двухуровневого эксперимента

(сбалансированный полноблочный план трехфакторного двухуровневого, полностью рандомизированного эксперимента для дисперсионного анализа)

| № места (i) | № автомобиля (j) | |
|---|---|---|
| | I | II |
| 1 | **A** | **B** |
| 2 | **B** | **A** |

| № опыта (i) | Факторы (j) | | | Результат опыта (i) |
|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | |
| 1 | **+1** | **+1** | **+1** | $y_1$ |
| 2 | **+1** | **–1** | **–1** | $y_2$ |
| 3 | **–1** | **+1** | **–1** | $y_3$ |
| 4 | **–1** | **–1** | **+1** | $y_4$ |

## Свойства плана

1) симметричность: $\sum_{i=1}^{N} x_{ij} = 0$,

(сумма элементов любого столбца равна нулю);

2) условие нормировки: $\sum_{i=1}^{N} x_{ij}^2 = k+1$,

– где k + 1 называется числом степеней свободы плана (длина столбца на единицу больше числа k факторов);

3) ортогональность:

$$\sum_{i=1}^{N} x_{im} x_{in} = 0 \qquad \text{при} \quad m \neq n,$$

(скалярное произведение любых двух различных столбцов плана равно нулю) обеспечивает **независимость** определения всех коэффициентов линейной регрессионной модели $y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3$;

4) насыщенность для выбранной **линейной регрессионной** модели: $N = k + 1$ – можно определить все коэффициенты модели;

ненасыщенность при $N > k + 1$ – кроме коэффициентов модели можно получить дополнительную информацию;

сверхнасыщенность при $N < k + 1$ – информации недостаточно для определения всех коэффициентов модели.

Оценки для всех $j \neq 0$ коэффициентов линейной регрессии даются общей формулой: $\tilde{\lambda}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} y_i$, а $\tilde{\lambda}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$ нарушает единообразие. Поэтому введем в план дополнительный фиктивный фактор $x_0$ "прочих" влияний, которые всегда присутствуют (+1), при всех опытах эксперимента.

## План трехфакторного двухуровневого эксперимента

(выделена **матрица Адамара**)

| № опыта (i) | Факторы (j) | | | | Результат опыта (i) |
|---|---|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ | $x_3$ | |
| 1 | **+1** | **+1** | **+1** | **+1** | $y_1$ |
| 2 | **+1** | **+1** | **–1** | **–1** | $y_2$ |
| 3 | **+1** | **–1** | **+1** | **–1** | $y_3$ |
| 4 | **+1** | **–1** | **–1** | **+1** | $y_4$ |
| | | план | | | |

# Неполные и неортогональные планы

[Часть II, стр. 92 - 97]

## Правило составления полного плана

| № опыта (i) | Факторы (j) | |
|---|---|---|
| | $x_0$ | $x_1$ |
| 1 | **+1** | **+1** |
| 2 | **+1** | **–1** |
| | | план |

# Правило составления полного плана

| № опыта (i) | Факторы (j) | | |
|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ |
| 1 | **+1** | **+1** | **+1** |
| 2 | **+1** | **–1** | **+1** |
| | | | план |

# Правило составления полного плана

| № опыта (i) | Факторы (j) | | |
|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ |
| 1 | **+1** | **+1** | **+1** |
| 2 | **+1** | **–1** | **+1** |
| 3 | **+1** | **+1** | **–1** |
| 4 | **+1** | **–1** | **–1** |
| | | | план |

# Правило составления полного плана

| № опыта (i) | Факторы (j) | | | |
|---|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ | $x_3$ |
| 1 | **+1** | **+1** | **+1** | **+1** |
| 2 | **+1** | **–1** | **+1** | **+1** |
| 3 | **+1** | **+1** | **–1** | **+1** |
| 4 | **+1** | **–1** | **–1** | **+1** |
| | | план | | |

# Правило составления полного плана

| № опыта (i) | Факторы (j) | | | |
|---|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ | $x_3$ |
| 1 | **+1** | **+1** | **+1** | **+1** |
| 2 | **+1** | **–1** | **+1** | **+1** |
| 3 | **+1** | **+1** | **–1** | **+1** |
| 4 | **+1** | **–1** | **–1** | **+1** |
| 5 | **+1** | **+1** | **+1** | **–1** |
| 6 | **+1** | **–1** | **+1** | **–1** |
| 7 | **+1** | **+1** | **–1** | **–1** |
| 8 | **+1** | **–1** | **–1** | **–1** |
| | | план | | |

Формула объема полного плана (при одинаковом количестве уровней всех факторов) принята и для обозначения плана:

$$h^k$$

h – общее количество уровней всех факторов;

k – количество факторов;

$h^k$ – количество строк плана.

Для предыдущего примера: $2^3$.

Полный план $2^3$ является *ненасыщенным*, так как

$$N = 8 > k + 1 = 4,$$

т.е. можно добавить 4 таких новых члена регрессионной модели, которые не изменят числа **основных** факторов.

Таким образом, в модель можно добавить всевозможные **сочетания** (произведения) **основных** факторов и рассматривать ее в нелинейном, виде:

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_{12} x_1 x_2 + \lambda_{13} x_1 x_3 + \lambda_{23} x_2 x_3 + \lambda_{123} x_1 x_2 x_3.$$

Под такую **расширенную** регрессионную модель можно построить *насыщенный* план.

Насыщенный план для расширенной регрессионной модели

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_{12} x_1 x_2 + \lambda_{13} x_1 x_3 + \lambda_{23} x_2 x_3 + \lambda_{123} x_1 x_2 x_3$$

| № опыта (i) | Факторы (j) | | | | | | | | Результат опыта |
|---|---|---|---|---|---|---|---|---|---|
| | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | $x_1x_2x_3$ | |
| 1 | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | $y_1$ |
| 2 | **+1** | **–1** | **+1** | **+1** | **–1** | **–1** | **+1** | **–1** | $y_2$ |
| 3 | **+1** | **+1** | **–1** | **+1** | **–1** | **+1** | **–1** | **–1** | $y_3$ |
| 4 | **+1** | **–1** | **–1** | **+1** | **+1** | **–1** | **–1** | **+1** | $y_4$ |
| 5 | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | **–1** | **–1** | $y_5$ |
| 6 | **+1** | **–1** | **+1** | **–1** | **–1** | **+1** | **–1** | **+1** | $y_6$ |
| 7 | **+1** | **+1** | **–1** | **–1** | **–1** | **–1** | **+1** | **+1** | $y_7$ |
| 8 | **+1** | **–1** | **–1** | **–1** | **+1** | **+1** | **+1** | **–1** | $y_8$ |
| | | план | | | | | | | |

Построенный таким образом насыщенный план обладает всеми рассмотренными ранее свойствами.

Условие **нормировки** выполняется для всех столбцов очевидным образом.

**Симметричность** столбцов парных произведений $\{x_{i1}x_{i2}\}$ следует из условия ортогональности основных столбцов.

Симметричность столбца $\{x_{i1}x_{i2}x_{i3}\}$:

$$\sum_{i=1}^{N} x_{i1}x_{i2}x_{i3} = \sum_{i=1}^{N/2} x_{i1}x_{i2}x_{i3} + \sum_{i=N/2+1}^{N} x_{i1}x_{i2}x_{i3} =$$

$$= \sum_{i=1}^{N/2} x_{i1}x_{i2} \cdot 1 - \sum_{i=N/2+1}^{N} x_{i1}x_{i2} \cdot 1 = 0,$$

так как $\sum_{i=1}^{N/2} x_{i1}x_{i2} = \sum_{i=N/2+1}^{N} x_{i1}x_{i2}$.

**Ортогональность** столбцов-произведений**:**

$$\sum_{i=1}^{N}(x_k)_i(x_l x_m)_i = \sum_{i=1}^{N} x_{ik}x_{il}x_{im} = 0,$$

$$\sum_{i=1}^{N}(x_k)_i(x_k x_m)_i = \sum_{i=1}^{N} x_{ik}x_{ik}x_{im} = \sum_{i=1}^{N} 1\cdot x_{im} = 0,$$

$$\sum_{i=1}^{N}(x_k)_i(x_k x_l x_m)_i = \sum_{i=1}^{N} x_{ik}x_{ik}x_{il}x_{im} = \sum_{i=1}^{N} 1\cdot x_{il}x_{im} = 0$$

$$\sum_{i=1}^{N}(x_k x_l)_i(x_l x_m)_i = \sum_{i=1}^{N} x_{ik}x_{il}x_{il}x_{im} = \sum_{i=1}^{N} 1\cdot x_{ik}x_{im} = 0,$$

$$\sum_{i=1}^{N}(x_k x_l)_i(x_k x_l x_m)_i = \sum_{i=1}^{N} x_{ik}x_{il}x_{ik}x_{il}x_{im} = \sum_{i=1}^{N} 1\cdot 1\cdot x_{im} = 0.$$

**Насыщенный** план 3-факторного эксперимента для **расширенной нелинейной** регрессионной модели

$$y = \lambda_0 + \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_{12} x_1 x_2 + \lambda_{13} x_1 x_3 + \lambda_{23} x_2 x_3 + \lambda_{123} x_1 x_2 x_3$$

⇓

**Насыщенный** план 7-факторного эксперимента для **линейной** регрессионной модели

$$y = \lambda_0 + \sum_{j=1}^{7} \lambda_j x_j .$$

## Дробные планы

Неполные планы при условии их **симметричности**, нормировки и **ортогональности**, позволяющие определить регрессионную модель **частного вида**, называются дробными планами или дробными репликами.

**Дробный план 1/m (1/m реплика)** k-факторного h-уровневого эксперимента $h^{k-m}$.

$h^{k-m}$ – объем плана,

k – m – число основных факторов, по которым обеспечивается полный перебор,

m – число дополнительных факторов.

Правило составления дробных планов

$2^{4-1}$

| № опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| 1 | +1 | +1 | +1 | +1 | +1 |
| 2 | +1 | –1 | +1 | +1 | –1 |
| 3 | +1 | +1 | –1 | +1 | –1 |
| 4 | +1 | –1 | –1 | +1 | +1 |
| 5 | +1 | +1 | +1 | –1 | +1 |
| 6 | +1 | –1 | +1 | –1 | –1 |
| 7 | +1 | +1 | –1 | –1 | –1 |
| 8 | +1 | –1 | –1 | –1 | +1 |

$x_4 = x_1 \cdot x_2$

Дробный план $2^{k-m}$ в $2^m$ раз меньше полного $2^k$ для того же количества факторов.

Дробный план $2^{4-1}$ составляет половину от полного $2^4$ и называется **полурепликой**.

Две полуреплики плана четырехфакторного двухуровневого эксперимента

| № опыта (i) | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|
| 1 | **+1** | **+1** | **+1** | **+1** | **+1** |
| 2 | **+1** | **–1** | **+1** | **+1** | **–1** |
| 3 | **+1** | **+1** | **–1** | **+1** | **–1** |
| 4 | **+1** | **–1** | **–1** | **+1** | **+1** |
| 5 | **+1** | **+1** | **+1** | **–1** | **+1** |
| 6 | **+1** | **–1** | **+1** | **–1** | **–1** |
| 7 | **+1** | **+1** | **–1** | **–1** | **–1** |
| 8 | **+1** | **–1** | **–1** | **–1** | **+1** |
| 9 | **+1** | **+1** | **+1** | **+1** | **–1** |
| 10 | **+1** | **–1** | **+1** | **+1** | **+1** |
| 11 | **+1** | **+1** | **–1** | **+1** | **+1** |
| 12 | **+1** | **–1** | **–1** | **+1** | **–1** |
| 13 | **+1** | **+1** | **+1** | **–1** | **–1** |
| 14 | **+1** | **–1** | **+1** | **–1** | **+1** |
| 15 | **+1** | **+1** | **–1** | **–1** | **+1** |
| 16 | **+1** | **–1** | **–1** | **–1** | **–1** |

Предыдущие расширенные регрессионные модели содержали факторы только в степени 0 или 1. Завершенная модель должна содержать степени факторов, например, квадраты:

$$y = \lambda_0 + \sum_{j=1}^{k} \lambda_j x_j + \sum_{i=1}^{k} \sum_{j=1}^{k} \lambda_{ji} x_j x_i .$$

Для такой модели с числом факторов k = 3 необходимо определить 10 различных коэффициентов (матрица $(\lambda_{ji})$ симметричная).

**Полный** план 3-факторного эксперимента дает только $2^3 = 8$ опытов, т.е. этот полный план **для такой регрессионной модели** является **сверхнасыщенным** и не позволяет определить все коэффициенты модели.

В рамках прежнего правила в план войдут одинаковые столбцы с квадратами факторов: $x_j^2 = x_i^2$, состоящие из единиц. Это приведет к **неортогональности** плана. Таким образом, все коэффициенты регрессии определить не удастся.

Необходим план специального вида.

# Неортогональный план Бокса

| № опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | | |
|---|---|---|---|---|---|---|
| **1** | **+1** | **+1** | **+1** | **+1** | | |
| **2** | **+1** | **–1** | **+1** | **+1** | | |
| **3** | **+1** | **+1** | **–1** | **+1** | | |
| **4** | **+1** | **–1** | **–1** | **+1** | | |
| **5** | **+1** | **+1** | **+1** | **–1** | | |
| **6** | **+1** | **–1** | **+1** | **–1** | | |
| **7** | **+1** | **+1** | **–1** | **–1** | | |
| **8** | **+1** | **–1** | **–1** | **–1** | | |
| | | | | | | |
| | | | | | | |

# Неортогональный план Бокса

| № опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_1^2$ | $x_2^2$ | $x_3^2$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | |
| **2** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | **я** |
| **3** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | |
| **4** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **д** |
| **5** | **+1** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | |
| **6** | **+1** | **–1** | **+1** | **–1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | **р** |
| **7** | **+1** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | |
| **8** | **+1** | **–1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **о** |
| | | | | | | | | | | | |
| | | | | | | | | | | | |

# Неортогональный план Бокса

| № опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_1^2$ | $x_2^2$ | $x_3^2$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | |
| **2** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | **я** |
| **3** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | |
| **4** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **д** |
| **5** | **+1** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | |
| **6** | **+1** | **–1** | **+1** | **–1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | **р** |
| **7** | **+1** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | |
| **8** | **+1** | **–1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **о** |
| **9** | **+1** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **0** | **з** |
| **10** | **+1** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **0** | **в** |
| **11** | **+1** | **0** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **ё** |
| **12** | **+1** | **0** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **з** |
| **13** | **+1** | **0** | **0** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **д** |
| **14** | **+1** | **0** | **0** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **ы** |
| | | | | | | | | | | | |

# Неортогональный план Бокса

| № опыта | $x_0$ | $x_1$ | $x_2$ | $x_3$ | $x_1^2$ | $x_2^2$ | $x_3^2$ | $x_1x_2$ | $x_1x_3$ | $x_2x_3$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | |
| **2** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | **я** |
| **3** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | |
| **4** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | **д** |
| **5** | **+1** | **+1** | **+1** | **–1** | **+1** | **+1** | **+1** | **+1** | **–1** | **–1** | |
| **6** | **+1** | **–1** | **+1** | **–1** | **+1** | **+1** | **+1** | **–1** | **+1** | **–1** | **р** |
| **7** | **+1** | **+1** | **–1** | **–1** | **+1** | **+1** | **+1** | **–1** | **–1** | **+1** | |
| **8** | **+1** | **–1** | **–1** | **–1** | **+1** | **+1** | **+1** | **+1** | **+1** | **+1** | **о** |
| **9** | **+1** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **0** | **з** |
| **10** | **+1** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **0** | **в** |
| **11** | **+1** | **0** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **ё** |
| **12** | **+1** | **0** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **0** | **з** |
| **13** | **+1** | **0** | **0** | **+α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **д** |
| **14** | **+1** | **0** | **0** | **–α** | **0** | **0** | **$\alpha^2$** | **0** | **0** | **0** | **ы** |
| **15** | **+1** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **0** | **центр** |

В итоге получен **ненасыщенный** неортогональный план для регрессионной модели:

$$y=\lambda_0+\lambda_1 x_1+\lambda_2 x_2+\lambda_3 x_3+$$
$$+\lambda_{11} x_1^2+\lambda_{22} x_2^2+\lambda_{33} x_3^2+$$
$$+\lambda_{12} x_1 x_2+\lambda_{13} x_1 x_3+\lambda_{23} x_2 x_3,$$

имеющий дополнительно (15 – 10) = 5 степеней свободы.

# Особые методы планирования эксперимента

## Специальные приемы планирования эксперимента

[Часть II, стр. 107 - 111]

## Выбор некоррелированных факторов – метод главных компонент

Исходная система из k наблюдаемых факторов $x_j$ (j = 1, 2,..., k) построена по итогам N наблюдений:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nk} \end{pmatrix}.$$

По этим данным можно составить корреляционную матрицу $R = (r_{ij})$ из выборочных оценок *коэффициентов корреляции*:

$$r_{ij} = \frac{1}{s_i s_j} \cdot \frac{1}{N-1} \cdot \sum_{n=1}^{N} (x_{ni} - \bar{x}_i)(x_{nj} - \bar{x}_j),$$

где $s_i$ – выборочная оценка среднего квадратического отклонения i-го фактора.

Заметим, что диагональные элементы матрицы R равны единице, а сама матрица – симметричная:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \dots & \dots & \dots & \dots \\ r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}.$$

Основная идея метода главных компонент состоит в замене переменных, характеризующих исследуемые факторы, на некоррелированные:

$$z_i = \sum_{j=1}^{k} a_{ij} x_j .$$

Для этого достаточно, чтобы матрица ковариаций новых переменных стала диагональной.

Математический метод решения такой задачи известен и реализуется с помощью стандартного программного обеспечения современных ЭВМ:

– из уравнения det(R – vE) = 0, где E – единичная матрица, находятся собственные значения матрицы R: $v_i = s^2(z_i)$;

– вычисляются собственные векторы $\mathbf{a}_i$ той же матрицы из матричного уравнения $R\mathbf{a}_i = v_i\mathbf{a}_i$ ;

– матрица искомого преобразования составляется из векторов $\mathbf{a}_i$ как столбцов: $A = (a_{mi})$.

## Факторный анализ

Выбор наименьшего числа наиболее представительных (независимых, некоррелированных) факторов с целью уменьшения объема эксперимента.

k исходных факторов заменяются на $m < k$ новых факторов $f_i$ ($i = 1, 2,..., m$):

$$x_j = \sum_{i=1}^{m} l_{ji} f_i + \varepsilon_j, \qquad j = 1,2,...,k; \quad i = 1,2,...,m,$$

где $\varepsilon_j$ – остаточная случайная поправка.

Формулируется и решается задача минимизации количества независимых (некоррелированных) факторов $f_i$.