
Лекция 3.

Дисперсионный анализ

Фундаментальная концепция дисперсионного анализа предложена Фишером в 1920 году.

Цель *дисперсионного анализа* (**ANOVA** - **AN**alysis **Of** **VA**riance) - проверка значимости различия между средними с помощью сравнения (т.е. анализа) дисперсий.

Основа метода - **разложение общей дисперсии** статистического комплекса **на** составляющие ее **компоненты**, которые сравниваются друг с другом посредством **F-критерия** □

какая доля общей вариации учитываемого *результативного признака* (*зависимой переменной*) обусловлена действием регулируемых и не регулируемых в опыте факторов.

MANOVA – **M**ultivariate **AN**alysis **Of** **VA**riance

Если сравнивать средние в двух выборках,

дисперсионный анализ =

= обычный t -критерий для независимых выборок (если сравниваются две независимые группы объектов или наблюдений)

или

= t -критерий для зависимых выборок (если сравниваются две переменные на одном и том же множестве объектов или наблюдений).

Основная причина, по которой использование дисперсионного анализа предпочтительнее повторного сравнения двух выборок при разных уровнях факторов с помощью серий t -критерия:

дисперсионный анализ существенно более *эффективен*

и

более информативен, особенно для малых выборок

Зависимые и независимые переменные

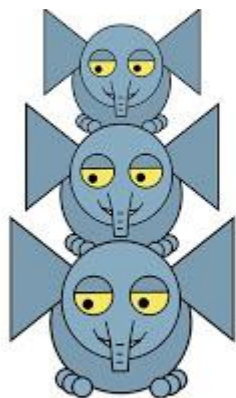
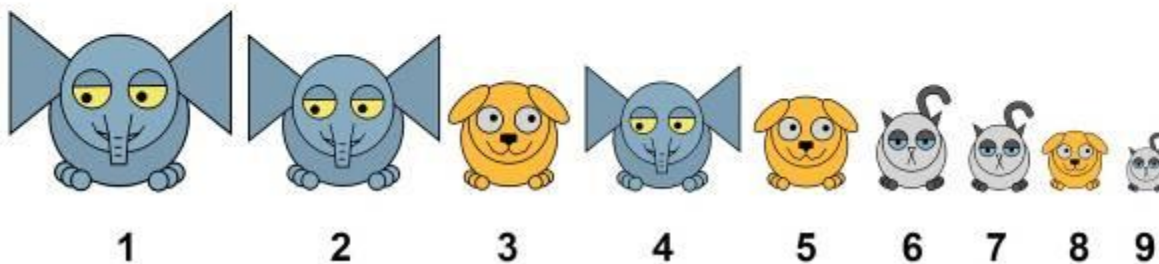
Зависимые переменные - те, значения которых определяется с помощью измерений в ходе исследования.

Шкалы отношений и интервальные

Независимые переменные или **факторы** - переменные, которыми можно управлять при проведении эксперимента (например, методы обучения) или другие критерии, позволяющие разделить наблюдения на группы или классифицировать. **Номинативные шкалы**

Как быть, если зависимая переменная задана порядковой шкалой?

Критерий Краскела-Уоллеса



$$1+2+4=7$$



$$3+5+8=16$$

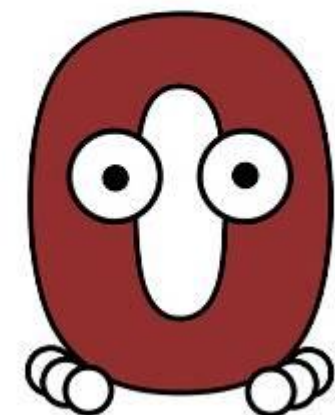
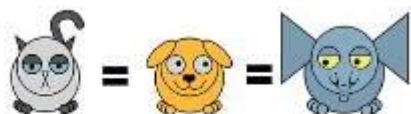


$$6+7+9=22$$

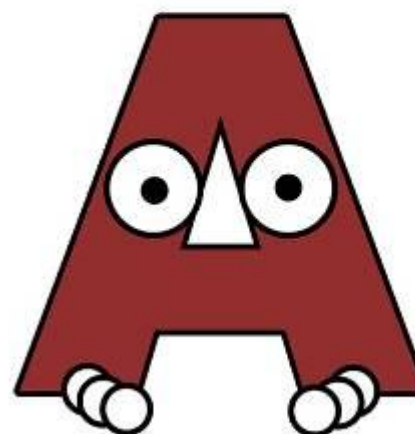
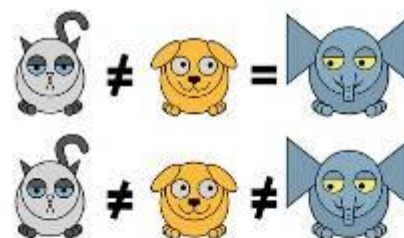
Дисперсионный анализ

Разделение общей дисперсии на несколько источников позволяет
сравнить дисперсию,
вызванную различием между группами,
с дисперсией,
вызванной внутригрупповой изменчивостью.

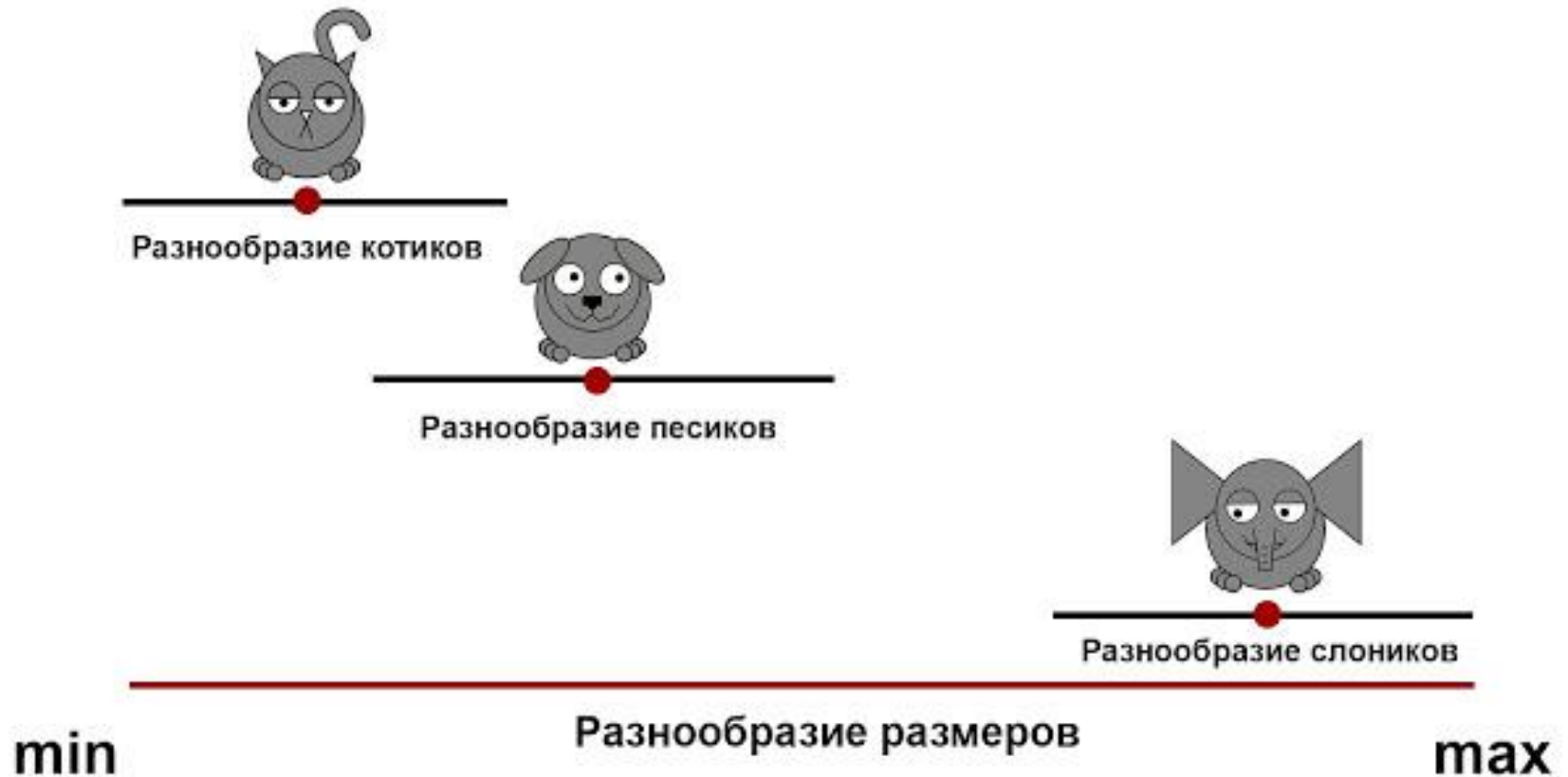
При истинности **нулевой гипотезы** (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии,
связанной с **внутригрупповой** изменчивостью,
должна быть
близкой к оценке **межгрупповой** дисперсии.



Нулевая гипотеза



Альтернативная
гипотеза



Внутригрупповая и межгрупповая (в данном случае – между биологическими видами) **изменчивости**

Внутри каждой группы, входящей в статистический (дисперсионный) комплекс, - варьирование, вызванное влиянием на признак не регулируемых в опыте факторов.

Зависимость между этими источниками варьирования выразится следующим равенством:

$$D_y = D_x + D_e$$

D_x – межгрупповая девиата - сумма квадратов отклонений групповых средних от общей средней комплекса, *взвешенная на n вариант в группе* ($N = \sum n$)

D_e – внутригрупповая девиата - сумма из сумм квадратов отклонений вариант от их групповых средних

D_y – общая девиата - сумма квадратов отклонений от общей средней комплекса в целом.

$$D_x = \sum_{i=1}^a \frac{n(\bar{x}_i - \bar{x})^2}{N}$$

$$D_e = \sum_{i=1}^a \left[\sum_{j=1}^n (x_j - \bar{x}_i)^2 \right]$$

$$D_y = \sum_{j=1}^n (x_j - \bar{x})^2$$

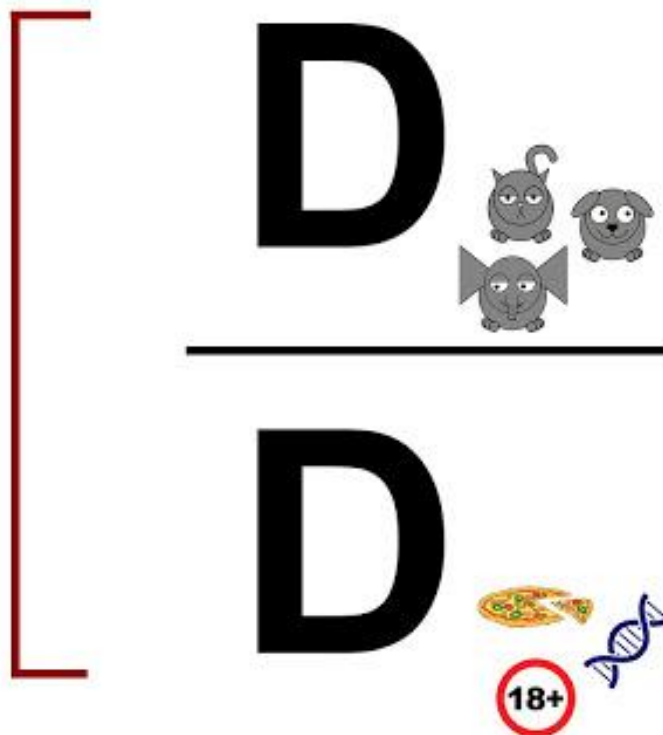
Деление сумм квадратов отклонений (**девиат**) на числа степеней свободы k дает выборочные **дисперсии** $s_y^2 = D_y/k_y$; $s_x^2 = D_x/k_x$; $s_e^2 = D_e/k_e$, которые служат оценками соответствующих генеральных параметров:

s_y^2 - оценка общей дисперсии комплекса,

s_x^2 - оценка межгрупповой дисперсии,

s_e^2 - оценка внутригрупповой или остаточной дисперсии.

F
Фишера

$$\left[\frac{D}{D} \right]$$


Основа метода - **разложение общей дисперсии** статистического комплекса **на** составляющие ее **компоненты**, которые сравниваются друг с другом посредством **F-критерия** □
какая доля общей вариации *зависимой переменной* обусловлена действием регулируемых и не регулируемых в опыте факторов.

Отношение межгрупповой дисперсии (*называется также факториальной, т.к. зависит от действия регулируемых факторов*) к внутригрупповой (остаточной) дисперсии – критерий оценки влияния регулируемых в исследовании факторов на результативный признак:

$$F = s_x^2 / s_e^2$$

Нулевая гипотеза: генеральные межгрупповые средние и дисперсии равны между собой и различия, наблюдаемые между выборочными показателями, вызваны случайными причинами, а не влиянием на признак регулируемых факторов.

Нулевую гипотезу отвергают, если $F_{\phi} \geq F_{st}$ для принятого уровня значимости α и чисел степеней свободы k_x и k_e ,

принимают, если $F_{\phi} < F_{st}$; при этом различия, наблюдаемые между групповыми средними комплекса, признают *статистически недостоверными*.

- После того как действие регулируемого фактора, нескольких факторов или их совместного действия на признак будет доказано, т.е. окажется статистически достоверным, переходят к сравнительной оценке групповых средних.
- Заключительный этап дисперсионного анализа - оценка силы влияния отдельных факторов или их совместного действия на признак:
- **Оценка post hoc и метод априорных контрастов**
 - метод наименьших значимых различий (**LSD**);
 - тест Шеффе (**Schejfe**)
 - тест Тьюки (**Tukey**)
 - тест Дункана
 - тест Бонферрони (критерий Стьюдента для множественных сравнений)

Дисперсионный анализ, как метод одновременных сравнений выборочных средних, предъявляет требования к группировке выборочных данных и к планированию наблюдений. Результаты наблюдений, подлежащие дисперсионному анализу, группируют с учетом градации каждого регулируемого фактора, воздействующего на признак.

-
- Особенность *post-hoc*-тестов - использование внутригруппового среднего квадрата для оценки любых пар средних.
 - Тесты по методам Бонферрони и Шеффе являются наиболее консервативными, так как они используют наименьшую критическую область при заданном уровне значимости .
-

- Если испытывают действие на признак одного регулируемого фактора, дисперсионный комплекс будет *однофакторным*, если одновременно исследуют действие на признак двух, трех или большего числа регулируемых факторов, комплекс называется *двух-, трех- и многофакторным*.

Числовые значения (даты) результативного признака могут распределяться по градациям комплекса равномерно, пропорционально и неравномерно. Поэтому дисперсионные комплексы называют равномерными, пропорциональными и неравномерными.

- *Равномерные и пропорциональные комплексы носят общее название ортогональные, а неравномерные комплексы называют неортогональными.*

Правильное применение дисперсионного анализа предполагает нормальное или близкое к нормальному распределению совокупности, из которой взяты выборки, объединяемые в дисперсионный комплекс.

!!! Важно, чтобы дисперсии выборочных групп были одинаковыми или не очень сильно отличались друг от друга (тесты на гомогенность дисперсий: Hartley F-max statistic, Cochran C statistic, the Bartlett *Chi*-square test; Levene's test)

Дисперсионный анализ:

- Однофакторный
- Многофакторный
 - Многомерный

Дисперсионный анализ характеризуется строгой логичностью и последовательностью вычислительных операций.

Ценность этого метода: позволяет выявить

- суммарное действие факторов,
- действие каждого регулируемого в опыте фактора в отдельности
- действие различных сочетаний факторов друг с другом на результативный признак.

Дисперсионный анализ позволяет выражать учитываемые признаки не только в абсолютных единицах измерения и счета, но и в баллах, индексах и других относительных и условных единицах.

Статистические, или дисперсионные, комплексы могут формироваться как в планах намечаемых исследований, так и на основании уже собранных данных, подвергаемых дисперсионному анализу.

При образовании дисперсионных комплексов необходимо соблюдать **два важных условия**, гарантирующих правильное применение дисперсионного анализа:

1. Действующие на признак регулируемые **факторы** должны быть **независимыми** друг от друга.
 2. **Выборки**, группируемые в статистический комплекс, должны производиться по принципу **рандомизации**, т.е. способом случайного отбора из нормально распределяющейся совокупности.
-

Видеолекция НОУ ИНТУИТ (к.физ-мат.н. Бояршинов Б.С., 1час 12 мин): https://www.youtube.com/watch?v=Wt1wdYWs_i0
