Дисперсионный анализ (ДА) = ANOVA (analysis of variance)

Примеры задач, которые решает ANOVA:

- Оценка средней успеваемости студентов после использования различных методов обучения (например, онлайн, очное, смешанное);
- Оценку влияния различных удобрений на урожайность сельскохозяйственных культур;
- Анализ эффективности рекламных кампаний, проведенных в разных регионах;
- Изучение влияния стресса на когнитивные способности у разных групп людей;
- Оценка эффективности нового лекарства по сравнению с плацебо и стандартным препаратом.

Что общего у этих задач ?

В этих экспериментах и исследованиях возникает необходимость сравнить несколько групп между собой. Например:

Группы студентов, обучающихся онлайн, очно и смешано;
Растения, выращенные с помощью азотных, калийных и комплексных удобрений;
Сильный, средний и сильный стресс и т.д

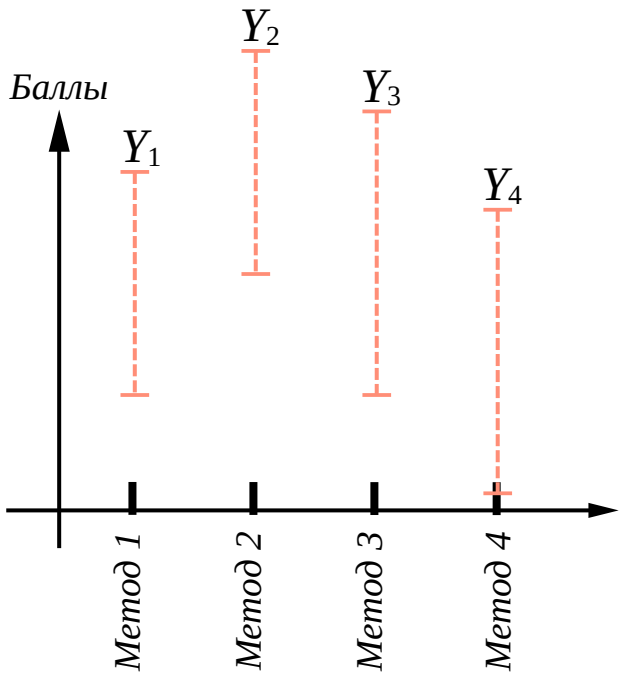
Переменная, которая будет разделять наших испытуемых или наблюдения на группы (номинативная переменная с нескольким градациями) называется **независимой переменной (фактор). Значением этой переменной мы можем управлять (задавать, контролировать).**
Пример факторов — тип обучения, вид удобрения, уровень стресса.

Количественная переменная, по степени выраженности, которой мы сравниваем группы, называется **зависимая переменная (отклик)**. Например, баллы тестирования студентов, количество плодов, количество решенных примеров и т.д.

Выделяют **однофакторный** и **многофакторный** дисперсионный анализ.

1.1 Однофакторный дисперсионный анализ

№ экс-та Уровни фактора	Один фактор **X**	Отклики объекта исследования при фиксированном уровне фактора				СВ	Групп. средняя	Групп. дисперсия
1	*Метод 1*	*50*	*85*	...	*30*	Y_1	$\bar{y}_1$	s_1^2
2	*Метод 2*	*90*	*75*	...	*70*	Y_2	$\bar{y}_2$	s_2^2
3	*Метод 3*	80	45		65	Y_3		
4	*Метод 4*	*40*	*60*	...	*55*	Y_4	$\bar{y}_k$	s_k^2



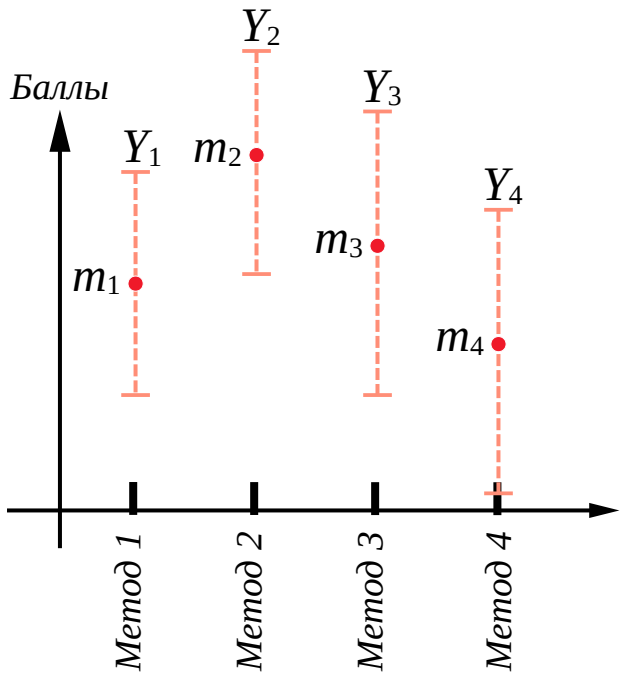
Рассмотрим для примера задачу.

Чтобы выяснить, как лучше преподавать статистику, преподаватели разбили всех студентов на 4 группы по 15 человек в каждой. В каждой группе использовалась различная методика преподавания. Т.е. фактор — метод преподавания. В конце года каждая группа написала итоговую работу. Отклик — балл за работу.

Влияет ли методика преподавания на усвоение материала?

Надо придумать какое-то математическое правила, чтобы проверить это утверждение (гипотезу)

1.1 Идея дисперсионного анализа



Влияет ли методика преподавания понимание материала?

Разброс (дисперсия) баллов возникает по двум причинам:
из-за разброса внутри группы и
из-за разброса между группами.

Почему возникает разброс внутри группы Y_i, обучающейся по одной методике? Точно не из-за методики, а по каким-то другим случайным факторам.

Почему возникает разброс между группами? Из-за разной методики преподавания.

Идея дисперсного анализа состоит в сравнении межгрупповой дисперсии («факторной дисперсии») S_X, порождаемой воздействием фактора, и внутригрупповой дисперсии («остаточной дисперсии») S_O, обусловленной случайными причинами.

⇒ Если различие ***незначимо***, то влияние несущественно.
!!! Также можно утверждать, что в этом случае средние различных групп значимо не различаются, т. е. $m_1 = m_2 = m_3 = m_4$

⇒ Если различие ***значимо***, то влияние существенно.
!!! В этом случае хотя бы одна пара средних различается значимо.

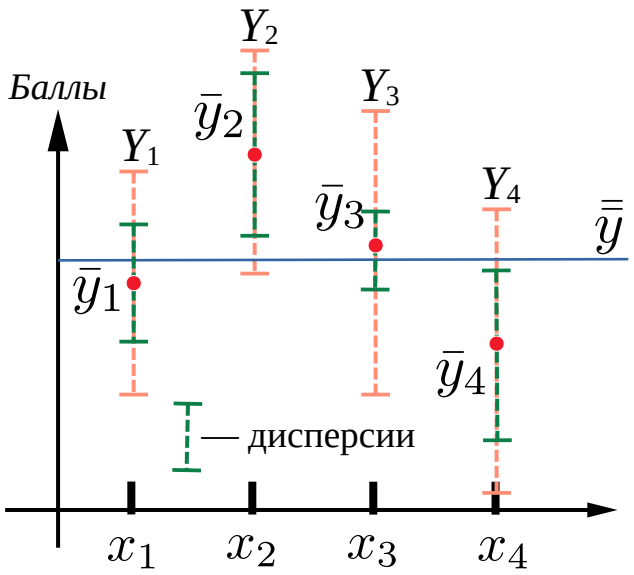
Классические методы дисперсионного анализа основываются на следующих предпосылках:
- распределение исходных случайных величин нормально;
- дисперсии экспериментальных данных одинаковы для всех условий эксперимента.

1.1 Как определяются факторная и остаточные дисперсии

№ экс-та Уровни фактора	Один фактор	Отклики объекта исследования при фиксированном уровне фактора				СВ	Групп. средняя	Групп. дисперсия
	X							
1	x_1	y_{11}	y_{12}	...	y_{1n}	Y_1	$\bar{y}_1$	s_1^2
2	x_2	y_{21}	y_{22}	...	y_{2n}	Y_2	$\bar{y}_2$	s_2^2
...								
k	x_k	y_{k1}	y_{k2}	...	y_{kn}	Y_k	$\bar{y}_k$	s_k^2

$$\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

$$\bar{\bar{y}} = \frac{1}{k}\sum_{i=1}^{k} \bar{y}_i$$



$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

$$S^2 = \frac{1}{kn-1}\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{\bar{y}})^2}_{Q}$$

Общая выборочная дисперсия

$$S_X^2 = \frac{n}{k-1}\underbrace{\sum_{i=1}^{k}(\bar{y}_i - \bar{\bar{y}})^2}_{Q_X}$$

Факторная дисперсия, характеризует изменение средних, связанное с влиянием фактора

оценка дисперсии групповых средних

$$Q = Q_O + Q_X$$

$$S^2(nk-1) = S_O^2(n-1)k + S_X^2(k-1)$$

$$S_O^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}_{Q_0}$$

Остаточная дисперсия, характеризует рассеяние вне влияния фактора

среднеарифметическое для оценок дисперсий

1.1 Алгоритм проведения однофакторнаго дисперсионного анализа

1 Проверка гипотезы о нормальном распределение исходных случайных величин Y_i, $i = 1...n$
Критерии Колмогорова и Хи-квадрат.

2 Дисперсия случайной величины $D[Y]$ является характеристикой объекта исследования и определяется только его природой. Поэтому значение величины $D[\underline{Y}]$ одинаково для всех случайных величин во всех строках таблицы данных (**однородности дисперсий или воспроизводимость дисперсий).**

Т.е внутри групп дисперсия возникает за счет каких-то внешних факторов. Например, количество часов сна перед контрольной. Требуется, чтобы эти факторы были постоянными в случае перехода от одной группы к другой. Т.е. фактор сна действует на всех студентов одинаково, не зависимо от группу. В этом случае внутригрупповые дисперсии должны быть равны (не отличаться значимо).

H_0: $\sigma_1 = \sigma_2 = \ldots. = \sigma_k$ Тесть Левена. Изучить самостоятельно!

3 Влияние фактора проверяется проверкой гипотезы о значимости статистического критерия:
H_0: $M[Y_1] = M[Y_2] = ... = M[Y_k]$ с уровнем значимости α

$$Z = \frac{S_X^2}{S_O^2} \in F(k-1, k(n-1))$$

Если $Z > Z_{KP} = F_\alpha^{-1}(k-1, k(n-1))$, то влияние фактора незначимо

Пример 1: Произведено по четыре испытания на каждом из трех уровней фактора *X*. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице:

№ экс-та Уровни фактора	Один фактор ***X***	Отклики объекта исследования при фиксированном уровне фактора				СВ	Групп. средняя $\overline{x}_i$	Групп. дисперсия s_i^2
1	x_1	*38*	*36*	*35*	*31*	Y_1	35	26/3
2	x_2	*20*	*24*	*26*	*30*	Y_2	25	52/3
3	x_3	*21*	*22*	*31*	*34*	Y_3	27	42

Методом дисперсионного анализа при уровне значимости 0.05 проверить воспроизводимость наблюдений и значимость влияния фактора

Решение: 1) Вычислим критерий Кохрена $G_p = \dfrac{42}{26/3 + 52/3 + 42} \approx 0.62 \quad < \quad G_T(f = 4-1, k = 3, \alpha = 0.05) \approx 0.8$

Гипотеза о воспроизводимости принимается, наблюдаемое различие связано со случайными факторами

2) общее среднее: $\bar{\bar{y}} = \dfrac{1}{k}\sum_{i=1}^{k} \bar{y}_i = 29$

3) найдем факторную дисперсию $S_X^2 = \dfrac{n}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{\bar{y}})^2 = \dfrac{4}{2}((35-29)^2 + 4^2 + 2^2) = 112$

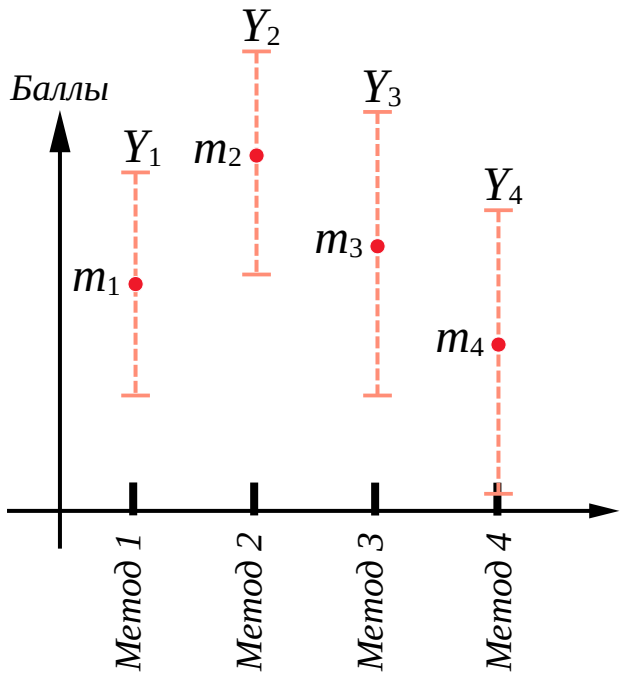
4) найдем остаточную дисперсию $S_O^2 = \dfrac{1}{k}\sum_{i=1}^{k} s_i^2 = \dfrac{26/3 + 52/3 + 42}{3} = \dfrac{68}{3}$

5) найдем значение критерия Фишера $Z = \dfrac{S_A^2}{S_O^2} = \dfrac{112 \cdot 3}{68} = 4.94$

6) найдем критическое значение критерия Фишера (используем таблицу) $F_\alpha^{-1}(k-1, k(n-1)) = F_{0.05}^{-1}(2, 9) = 4.26 = Z_{KP}$

$Z > Z_{KP}$ *Гипотеза о незначимости критерия Фишера отклоняется. Критерий значим, следовательно фактор значимо влияет на исследуемый отклик.*

1.1 Множественное сравнение в однофакторном дисперсионном анализе



Если H_0: $M[Y_1] = M[Y_2] = ... = M[Y_k]$ отклоняется с уровнем значимости α, то есть хотя бы одна пара групп, в которой имеется статистически значимое отклонение.

Какая? Для этого можно сравнивать каждую пару
H_0: $M[Y_i] = M[Y_j]$, используя тест Стьюдента.

Но нельзя при этом использовать тот же уровень значимости α, т. к. получить значимое различие увеличивается пропорционально количеству групп (*лабораторная*).
Аналог. Если бросать монетку один раз, то выпадание «орла» — 0.5, если подбросили монетку 1000 раз, то вероятность выпадание хотя бы одного «орла» стремиться к 1.

Самый простой способ — **поправка Бонферрони** $\alpha_{new} = \alpha / N$, где N — количество пар гипотез.
т. е. если групп 4, то пар $N = 4\ (4\text{-}1)/2 = 6$

Но поправка Бонферрони сильно занижает уровень, поэтому снижается вероятность выявить статистически значимые различия.

Альтернативный метод — **критерий Тьюки** (разобраться самостоятельно)

Отклик является случайной величиной, т. к. на него влияют неучтенные случайные факторы

№ экс-та	Факторы				Отклики объекта исследования			
	X_1	X_2	...	X_m				
1	x_{11}	x_{12}		x_{1m}	y_{11}	y_{12}	...	y_{1n}
2	x_{21}	x_{22}		x_{2m}	y_{21}	y_{22}	...	y_{2n}
...								
k	x_{k1}	x_{k2}		x_{km}	y_{k1}	y_{k2}	...	y_{kn}

Один уровень факторов

Эксперименты при различных уровнях факторов

Наблюдения при фиксированном уровне факторов