ДИСПЕРСИОННЫЙ АНАЛИЗ

ANOVA

ANOVA - это акроним от ANalysis Of VAriance (дисперсионный анализ). Дисперсионный анализ был введён Фишером - английским учёным, сделавшим огромный вклад в развитие науки. ANOVA в статистике - это мощный инструмент для определения влияния различных групп наблюдений между собой.

ПРИМЕР

Предположим, Вы хотите эмпирическим методом провести исследование бензина на качество, для этого вы заправляете бак на одной заправке и проезжаете n километров, повторяете такой эксперимент, скажем, пять раз, затем проводите такой же эксперимент, только на другой заправке. У Вас два набора данных - заправка A и заправка B. Разумеется, цифры разбегаются, но всё же есть некоторая зависимость, так вот, что бы определить, влияет ли заправка на расход бензина (или данные не связаны между собой) Вы используете дисперсионный анализ.

Дисперсионный анализ позволяет определить какой из факторов влияет больше, внутригрупповой или межгрупповой. В примере выше Вы сможете определить, насколько влияет на расход бензина выбор заправки. В этом суть дисперсионного анализа: узнать, является ли выбранный фактор значимым для выбранных наблюдений.

В некотором смысле, дисперсионный анализ похож на регрессионный и корреляционный анализы, т.к. позволяет определить влияние переменных друг на друга.

АНАЛИЗ

В теории, для анализа дисперсии выстраивается простая модель, схожая с изучаемой в анализе временых рядов.

МОДЕЛЬ

Модель дисперсионного анализа включает в себя среднее значение, эффект эксперимента и случайную ошибку:

$y = \mu + \tau + \varepsilon$
τ - эффект эксперимента, ε - случайная ошибка

ОДНОФАКТОРНЫЙ

Однофакторный дисперсионный анализ рассматривает влияние одного критерия, делается это так: мы проводим два эксперимента, в одном из них включаем дополнительный фактор и анализируем, внёс ли этот фактор изменения. В качестве исходных данных рассмотрим результаты ряда экспериментов:

N	**E_1**	**E_2**	**E_3**	**E_4**
1	52	43	124	31
2	41	44	127	31
3	52	44	120	44
4	40	41	91	40
5	43	34	120	53
μ_i	**45.6**	**41.2**	**116.4**	**39.8**

$\mu = (45.6 + 41.2 + 116.4 + 39.8) / 4 = 60.75$
Квадрат ошибок внутри групп (Square Sum within group):
$SS_w = \Sigma_i\Sigma_j(y_{ij} - \mu_i)^2 = 1400$
Квадрат ошибок между группами (Square Sum between group):
$SS_b = \Sigma_i(\mu_i - \mu)^2 = 4147.55$
Учитывая степени свободы, ожидаемое среднее:
$MS_w = SS_w / a(n-1) = 93.33$
$MS_b = SS_b / a-1 = 1036.89$
Значение $F_{крит}$:
$F_0 = MS_b/MS_w = 11.11$

Тест Фишера: если значение F_0 окажется больше чем значение $F_{\lambda,4,15}$, значит фактор оказывает влияние.

Для n = 20 и a = 5, $F_{\lambda,n-a,a-1} = F_{\lambda,15,4} = 5{,}86$
Поскольку $F_0 = 11.11 > 5.86$, то принимаем, что введённый фактор **оказал**

влияние на результаты эксперимента.

ДВУХФАКТОРНЫЙ

При двухфакторном анализе выдвигаются три гипотезы на проверку:

- Факторы A и B не оказывают влияния на результат
- Фактор A не оказывает влияния на результат
- Фактор B не оказывает влияния на результат

Для проведения двухфакторного анализа необходимо составить группы результатов: несколько измерений для всех значения каждого из факторов, т.е.:

	$\mathbf{A_1}$	$\mathbf{A_2}$
$\mathbf{B_1}$	$X1_{a1,b1}...XN_{a1,b1}$	$X1_{a1,b2}...XN_{a1,b2}$
$\mathbf{B_2}$	$X1_{a1,b2}...XN_{a1,b2}$	$X1_{a1,b2}...XN_{a1,b2}$

Далее подсчитывается среднее значение для каждого значения факторов, т.е. среднее для A1, среднее для B1 и т.д. Затем подсчитывается общее среднее для всех результатов. Зададимся количеством критериев: k = 2 (количество критериев A) и m = 2 (количество критериев B).

$T = \Sigma\Sigma\Sigma x_{ijk}$

Сумма элементов под влиянием фактора A:

$T_{Ai} = \Sigma x_{i\cdot k}$

Сумма элементов под влиянием фактора B:

$T_{Bj} = \Sigma x_{\cdot jk}$

Сумма элементов под влиянием фактора AB:

$T_{AiBj} = \Sigma x_{ij\cdot}$

$SST = \Sigma x^2_{ijk} - T^2/N$

$SSA = \Sigma T^2_{Ai}/n\cdot m - T^2/N$

$SSB = \Sigma T^2_{Bj}/n\cdot k - T^2/N$

$SSAB = \Sigma\Sigma T^2_{AiBj}/n - SSA - SSB - T^2/N$

$SSE = \Sigma\Sigma\Sigma x^2_{ijk} - \Sigma\Sigma T^2_{AiBj}/n$

$SST = SSA + SSB + SSAB + SSE$

$MSE = SSE/(n-1)\cdot m\cdot k$

$MSA = SSA/k-1$

$MSB = SSB/m-1$

$MSAB = SSAB/(m-1)\cdot(k-1)$

Тест "Критерий A **не** оказывает влияние на результат", $\nu_1 = k-1$:

$F_A = MS_A/MS_E$

Тест "Критерий B **не** оказывает влияние на результат", $\nu_1 = m-1$:

$F_B = MS_B/MS_E$

Тест "Критерии A и B **не** оказывают влияние на результат", $\nu_1 = (k-1)(m-1)$:

$F_{int} = MS_{AB}/MS_E$

Для каждого F, если $F > F_{\alpha,\nu_1,\nu_2}$, то гипотеза отвергается. $\nu_2 = N-mk$

МНОГОФАКТОРНЫЙ

Многофакторный анализ аналогичен двухфакторному - проводятся те же операции, но критерии группируются и итеративно находится влияние каждого из факторов.

С ПОВТОРНЫМИ ИЗМЕРЕНИЯМИ

Дисперсионный анализ с повторными измерениями означает, что для каждого критерия производилось несколько замеров случайной величины для получения более точного результата (поскольку в ANOVA) используется внутригрупповая сумма квадратов.

ПРИМЕНЕНИЕ

Дисперсионный анализ применяют в самых различных отраслях науки и производства тогда, когда необходимо изучить зависимость критериев на различие средних значений, при этом сравнивается не среднее значение, а разброс результатов вокруг среднего значения, т.е. дисперсию.

РЕШЕНИЕ ЗАДАЧ

В качестве примера приведём задачу из метрологии. На заводе размещены пять станков, на которых производят валы. Необходимо определить, влияет ли выбор станка или подготовка работника на результат производства. Для анализа производят замеры для каждого станка и работника, в результате получается таблица:

Оператор 1

M1	30.319	30.33	30.406	30.339	30.305	30.333	30.319	30.392	30.48	30.404
M2	30.37	30.377	30.392	30.342	30.323	30.362	30.388	30.358	30.4	30.317
M3	30.174	30.26	30.141	30.284	30.139	30.121	30.254	30.233	30.281	30.248
M4	30.34	30.372	30.396	30.335	30.359	30.365	30.312	30.35	30.307	30.335
M5	30.312	30.313	30.385	30.357	30.346	30.348	30.377	30.326	30.357	30.358

Оператор 2

M1	30.326	30.306	30.343	30.346	30.379	30.384	30.317	30.374	30.319	30.376
M2	30.3	30.3	30.3	30.3	30.3	30.3	30.3	30.3	30.3	30.3
M3	29.903	29.956	29.947	30.062	30.018	29.735	30.011	29.676	30.168	30.18
M4	30.541	30.33	30.384	30.333	30.41	30.523	30.533	30.527	30.347	30.412
M5	30.405	30.513	30.526	30.359	30.595	30.374	30.489	30.427	30.464	30.595

Воспользуемся методом двухфакторного анализа, фактор A - оператор, фактор B - станок. Рассчитаем суммы квадратов, для этого необходимо рассчитать значение среднего для каждой из групп:

T	T_{A1}	T_{A2}	T_{B1}	T_{B2}	T_{B3}	T_{B4}	T_{B5}
3031.554	1516.341	1515.213	607.097	606.629	601.791	607.811	608.226

SSA = 0.013
SSB = 1.353
SSAB = 0.434
SSE = 0.467

MSA = 0.013
MSB = 0.338
MSAB = 0.109
MSE = 0.117

$F_A = 0.111$
$F_B = 2.889$
$F_{AB} = 0.932$

Критические значения для теста Фишера:
$F_{crit\ A} = F_{0.1,\ 1,\ 90} = 2.77$
$F_{crit\ B} = F_{0.1,\ 4,\ 90} = 2.01$
$F_{crit\ AB} = F_{0.1,\ 4,\ 90} = 2.01$

Таблица результатов:

Влияние станка на результат	Да	0.111 < 2.77
Влияние квалификации работника на результат	Нет	2.889 > 2.01
Взаимное влияние квалификации работника и выбора станка на результат	Да	0.932 < 2.01

В EXCEL/OPEN CALC

Для решения дисперсионного анализа в электронной таблице Вам потребуются следующие формулы:

sumproduct	Сумма произведений, используется для нахождения суммы квадратов
finv	Обратное значение распределения F - критерий Фишера

Таблица для скачивания в форматах ods и xls.

УДК: ГРНТИ:
Автор статьи: *Телятников З.А.*
Дата написания статьи: *30.06.2017*
Адрес статьи в интернете:
http://k-tree.ru/articles/statistika/analiz_dannyh/dispersionnii_analiz_anova

Дата формирования документа: *11.12.2017 05:12*
