ДИСПЕРСИОННЫЙ АНАЛИЗ; Критерий Фишера; **КОРРЕЛЯЦИЯ;** Преобразование Фишера; **ПРОСТАЯ АЛЬТЕРНАТИВА(отношение правдоподобия); ПЕРЕСТАНОВОЧНЫЕ ТЕСТЫ; МНОЖЕСТВЕННОЕ ТЕСТИРОВАНИЕ**

ANOVA = Analysis Of VAriations

ДИСПЕРСИОННЫЙ АНАЛИЗ - метод, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий в средних значениях. Дисперсионный анализ позволяет сравнивать средние значения двух и более групп.

Какие бывают задачи на зависимость

1. Оба признака категориальные

a. Наличие/отсутствие редкого аллеля в геноме, диабетики и здоровые

Это таблицы сопряжённости 2 ? 2

b. Пять городов, цвет глаз

Таблицы сопряжённости n ? m

2. Оба признака количественные

a. Вес и концентрация сахара в крови

Корреляция (Пирсона и Спирмена)

3. Один признак количественный, другой категориальный

a. Наличие/отсутствие аллеля, концентрация сахара в крови

Это сравнение двух выборок: критерии Уилкоксона (Манна – Уитни) или Стьюдента

b. Пять городов, рост взрослых мужчин

Нужно сравнить пять выборок → ANOVA

Критерий Фишера

Имеются две выборки: $X_1, \ldots, X_k$ и $Y_1, \ldots, Y_l$ (не обязательно одного размера)

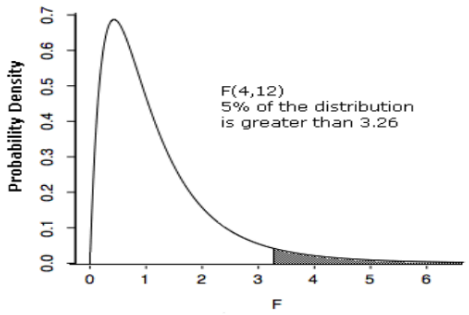
H_0: они происходят из нормальных распределений с равными дисперсиями.
H_1: они происходят из нормальных распределений с разными дисперсиями

Статистика: $\kappa = s_X^2 / s_Y^2$, где
$s_X^2 = \sum_i (X_i - \bar{X})^2 / (k-1)$,
$s_Y^2 = \sum_j (Y_j - \bar{Y})^2 / (l-1)$
— две выборочные дисперсии

Величина κ при нулевой гипотезе подчиняется так называемому F-распределению (распределению Фишера) с параметрами $k-1$ и $l-1$.

Поскольку $\sum_i (X_i - \bar{X})^2$ распределена по закону хи-квадрат, распределение Фишера определяется как распределение отношения двух независимых с.в., распределённых по хи-квадрат, умноженного на $(l-1)/(k-1)$

Fisher distribution



F - Distribution ($\alpha = 0.01$ in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659

Дисперсионный анализ (ANOVA)

Имеется несколько выборок:

X11, ..., X1n(1)

X21,...,X2n(2)

...

Xk1,...,Xkn(k)

H0: все эти выборки происходят из одного и того же нормального распределения

H1: эти ***выборки происходят из нормальных распределений с одинаковыми дисперсиями, но разными средними***

(ср. с двухвыборочным критерием Стьюдента)

Дисперсионный анализ (ANOVA)

Находим k выборочных средних и общее среднее:

$m_i = \sum_j X_{ij}/n(i) \;\; i = 1, \ldots, k$

$m = \sum_{ij} X_{ij} / \sum_i n(i)$

Оцениваем (одинаковую, по предположению) дисперсию выборок:

$s_0^2 = \sum_{ij}(X_{ij} - m_i)^2 / \sum_i (n(i) - 1)$

Находим т.н. межгрупповую дисперсию:

$s_1^2 = \sum_i n(i)\,(m_i - m)^2/(k-1)$

Теперь $\kappa = s_1^2/s_0^2$ при нулевой гипотезе имеет F-распределение с параметрами $k - 1$ и $\sum(n(i) - 1)$

Дисперсионный анализ, случай k = 2

Две выборки: $X_1, \ldots, X_g$ и $Y_1,\ldots,Y_h$

H_0: обе выборки происходят из одного и того же нормального распределения.

H_1: выборки происходят из нормальных распределений с одинаковыми дисперсиями, но разными средними.

То есть задача в точности та же, что для двустороннего двухвыборочного критерия Стьюдента.

Межгрупповая дисперсия $s_1^2 = \sum_i n(i)(m_i - m)^2/(k-1)$ при k = 2 превращается в $s_1^2 = g\,(m_X - m)^2 + h\,(m_Y - m)^2$

Статистика $\kappa = s_1^2/s_0^2 = (g\,(m_X - m)^2 + h\,(m_Y - m)^2)/s_0^2$

Статистика Стьюдента $t = (m_X - m_Y)/(s_0\,(1/g + 1/h)^{1/2})$

Нетрудно доказать, что $\kappa = t^2$. Поэтому критические множества $\kappa > a^2$ и $|t| > a$ одни и те же для любого a.

Иными словами, ANOVA для двух выборок эквивалентна двустороннему тесту Стьюдента.

Непараметрический вариант дисперсионного анализа

*Постановка задачи — как у ANOVA, **но без требования нормальности**, то есть*

H0: все выборки из одного распределения,
*H1: **распределения разные.***

*Критерий **Краскела – Уоллиса** (Kruskal — Wallis): посчитаем ранги всех наблюдений в объединённой выборке, рассчитаем средний ранг по каждой выборке и в качестве статистики возьмём среднее квадратичное отклонение этих средних рангов от ожидаемого среднего, равного (N+1)/2.*
(Упражнение: для k = 2 сравните с критерием Уилкоксона)

*Если альтернатива состоит в том, что выборки упорядочены по возрастанию действия некоторого фактора, то надо взять сумму статистик Уилкоксона (или Манна – Уитни) по всем парам выборок таким, что во второй выборке пары фактор (предположительно) действует сильнее, чем в первой. Получится статистика для критерия **Джонкхиера** (Jonckheere).*

*Замечание: оба критерия предназначены прежде всего для случая, когда **выборки отличаются сдвигом,** в противном случае их мощность невелика (но они остаются корректными, в отличие от обычной ANOVA).*

*2. **КОРРЕЛЯЦИЯ***

Примеры задач
1. Измеряем у нескольких десятков молодых людей длину волос и уровень оптимизма. Связаны ли эти параметры?
2. Измеряем у нескольких десятков больных определённым типом рака экспрессию некоторого гена в опухоли, затем собираем сведения о времени, которое они прожили после этого обследования. Влияет ли экспрессия данного гена на прогноз?
3. Собираем сведения об урожайности некоторой культуры на нескольких участках земли, одновременно определяем долю некоторого таксона бактерий в микробиоме почвы на этих участках. Есть ли связь?
4. И т.д.

В принципе можно поделить каждый из измеряемых параметров на классы по каким-то порогам и свести задачу к таблице сопряжённости. Интуитивно понятно, что это не очень хорошо:

во-первых, данные «загрубляются» (теряем часть информации),
во-вторых, трудно контролировать влияние произвола в выборе порогов.

Коэффициентом корреляции двух случайных величин ξ и η называется число:

$$R = \mathrm{E}((\xi - \mathrm{E}\xi)(\eta - \mathrm{E}\eta))/(\mathrm{D}\xi\mathrm{D}\eta)^{1/2}$$

R принимает значения от –1 до 1 (включительно).

Если величины ξ и η независимы, то их коэффициент корреляции равен 0.

Обратное, вообще говоря, неверно.

Пусть $X_1, \ldots, X_n; Y_1, \ldots, Y_n$ — две выборки чисел одинаковой длины.

Выборочной корреляцией (также корреляцией Пирсона, Pearson correlation) этих выборок называется число:

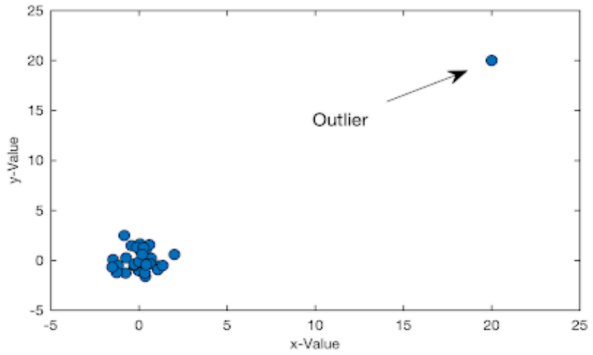
$$r = \sum(X_i - \bar{X})(Y_i - \bar{Y})/(\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2)^{1/2}$$

(являющееся оценкой коэффициента корреляции соответствующих случайных величин).

Преобразование Фишера

В предположении, что выборки X1, ..., Xn и Y1, ..., Yn происходят из двух нормально распределённых и независимых случайных величин, величина (преобразование Фишера выборочной корреляции):
z = ln((1+r)/(1–r))/2
имеет при n > 20 приблизительно нормальное распределение со средним 0 и дисперсией 1/(n – 3).
Это позволяет проверить нулевую гипотезу: «корреляция нулевая» против альтернативной гипотезы: «корреляция не нулевая».
*На **практике часто (но не всегда!) нулевая корреляция означает независимость**. **Ненулевая корреляция (генеральных совокупностей) всегда означает зависимость***

Если распределение генеральной совокупности заметно отличается от нормального, то использование такого критерия может привести к ошибке (корреляция между генеральными совокупностями может быть практически нулевой, а мы примем её за ненулевую)



http://mres.uni-potsdam.de/index.php/2017/02/14/outliers-and-correlation-coefficients/
Корреляция будет достаточно большой, чтобы казаться достоверно отличающейся от 0, но вся «зависимость» происходит из одного-единственного наблюдения

Ранговая корреляция

Корреляция Спирмена (*Spearman*) — это корреляция между **рангами** измерений
(ранг измерения – это число измерений в выборке, не превосходящих по величине данное измерение).

Если среди измерений есть одинаковые, то их ранги надо усреднить.
Например, для выборки (3; 7; 1; 3; 9) надо взять в качестве рангов числа (2,5; 4; 1; 2,5; 5)

Если все измерения различные, то корреляция Спирмена легко вычисляется по формуле:

$$\rho = 1 - 6\sum_i (\mathrm{rank}_X X_i - \mathrm{rank}_Y Y_i)^2 / n(n^2 - 1)$$

(но при наличии усреднённых рангов она не годится).

Существует ещё ранговая корреляция Кендалла (Kendall), но никаких преимуществ по сравнению с корреляцией Спирмена она не имеет, используют её редко.

Важное преимущество ранговой корреляции: независимость критерия независимости от исходных распределений.
Не уверены в нормальности распределений — используйте ранги!

Ранговая корреляция: значимость отличия от 0

Если ρ — корреляция Спирмена для n пар измерений, то величина

$$t = \rho\,((n-2)/(1-\rho^2))^{1/2}$$

распределена при нулевой гипотезе о независимости измерений приблизительно по Стьюденту с $n - 2$ степенями свободы.

При больших n можно использовать и преобразование Фишера

$$z = \ln\big((1+\rho)/(1-\rho)\big)/2$$

для случая корреляции Спирмена оно распределено нормально со средним 0 и дисперсией $1{,}06/(n-3)$

Отношение правдоподобия

3. ПРОСТАЯ АЛЬТЕРНАТИВА

Предположим, что нам известно распределение вероятностей **и для H_0 , и для H_1**.

Пусть сначала наши наблюдения X_1, X_2, …, X_n происходят из дискретного распределения.

Это значит, что каждое наблюдение может принимать лишь значения из некоторого дискретного множества, и каждое такое значение X при H_0 имеет ненулевую вероятность $P_0(X)$, а при H_1 — ненулевую вероятность $P_1(X)$.

Тогда по **теореме Неймана – Пирсона** оптимальной статистикой (т.е., статистикой, порождающей самый мощный критерий при любом заданном уровне значимости) является **отношение правдоподобия** (likelihood ratio):

$L = P_1(X_1)\cdot P_1(X_2)\cdot \ \dots \cdot P_1(X_n)/P_0(X_1)\cdot P_0(X_2)\cdot \ \dots \cdot P_0(X_n)$

Часто, чтобы иметь дело с суммами вместо произведений, используют log *L* вместо самого *L*

непрерывный случай

Пусть теперь наблюдения X_1, X_2, …, X_n происходят из непрерывного распределения.

Это значит, что каждое X может принимать любое действительное значение (или любое значение из какого-то интервала: $(0,\infty)$ или (a, b)),
вероятность каждого конкретного значения равно нулю
и для любого числа определена плотность вероятности в окрестности этого числа:

$p(x) = \lim_{\varepsilon\to 0} P(x-\varepsilon < X < x+\varepsilon)/2\varepsilon$

Тогда в качестве статистики используют отношение совместных плотностей вероятности при двух гипотезах:

$L = p_1(X_1)\cdot p_1(X_2)\cdot \ \dots \cdot p_1(X_n)/p_0(X_1)\cdot p_0(X_2)\cdot \ \dots \cdot p_0(X_n)$

или логарифм этой величины.

QEVGGALYA
QKLGEAIYS

Нулевая гипотеза: эти участки белков не гомологичны
Альтернативная гипотеза: участки гомологичны

Посчитаем вероятность такой пары последовательностей при нулевой гипотезе, затем при альтернативной и поделим второе на первое, получим статистику Неймана – Пирсона.

Вероятность пары **букв** x, y в одной колонке при нулевой гипотезе равно произведению частот этих букв: p(x)p(y), а при альтернативной гипотезе — частоте q(x,y) этой пары в *эталонных выравниваниях*.

Статистика: $\prod_k q(a_k, b_k)/p(a_k)p(b_k)$. Удобнее использовать логарифм этой величины:

$\log \prod_k q(a_k, b_k)/p(a_k)p(b_k) = \sum_k S(a_k, b_k)$, где $S(a_k, b_k) = \log q(a_k, b_k)/p(a_k)p(b_k)$

Числа $S(a_k, b_k)$ образуют **матрицу аминокислотных замен**
(зависит от эталонных выравниваний, самая известная называется BLOSUM62)

Что делать, если никакой известный критерий не подходит?

4. ПЕРЕСТАНОВОЧНЫЕ ТЕСТЫ

Идея: достаточно придумать хорошую статистику, а её распределение нам знать не надо!
Чтобы узнать p-value, достаточно придумать, как перемешать данные, чтобы при **нулевой гипотезе ничего не изменилось, а при альтернативной — изменилось**.
Тогда p-value (или как минимум его надёжная верхняя оценка) получается вычислительным экспериментом с многократным **перемешиванием**.
Вместо перемешивания можно моделировать выборку, исходя из нулевой гипотезы.

Поправки на множественность

5. МНОЖЕСТВЕННОЕ ТЕСТИРОВАНИЕ

Проблема множественного тестирования

• Тестируется новое лекарство. Группой лечения будем называть группу больных, которым выдают новое лекарство, а группой контроля — группу больных, которым его не выдают. Будем считать, что эффективность лекарства заключается в ослаблении симптомов заболевания (понижении температуры, нормализации давления, ослабления боли и т.д.). Чем больше симптомов рассматривается, тем более вероятно, что найдётся хотя бы один симптом, который в силу случайных причин окажется достоверно слабее у «группы лечения».
• Рассмотрим аналогичную ситуацию, но теперь будем считать лекарство безвредным, если оно не вызывает побочных эффектов. Чем больше возможных побочных эффектов рассматривается, тем более вероятно, что найдётся хотя бы один, который будет больше проявляться у «группы лечения» в конкретном исследовании .
• Опять аналогичная ситуация, но пусть теперь лекарство должно ослаблять симптом (например, понижать давление). Лекарство тестируется на пациентах разного возраста и пола, в разных географических зонах — пациенты разбиты на категории. **Чем больше категорий, тем более вероятно, что в одной из категорий ситуация будет выглядеть так, как если бы лекарство помогало (даже если оно реально не действует).**

Показатели, вычисляемые при множественном тестировании

FWER — family-wise error rate, вероятность ошибки первого рода хотя бы в одном варианте
FDR — false discovery rate, средняя доля ложных отклонений нулевой гипотезы (среди всех отклонений)
FCR — false coverage rate, средняя доля ложных покрытий, то есть не покрытие верных параметров в пределах выбранных интервалов.

Family-wise error rate (FWER)

Это фактически P-value, но вычисляемое с учётом множественного тестирования. На практике вычисляются P-value для каждого варианта (показателя, категории) отдельно, а затем к полученным числам применяется поправка на множественное тестирование
Варианты поправки:
• Поправка Бонферрони
• Поправка Шидака
• Поправка Холма – Бонферрони

Поправка Бонферрони

Наименьшее из полученных p-value умножается на число вариантов (показателей, категорий, ...):
P = mini Pi· N
Равносильная формулировка: при заданном уровне достоверности α отклоняем нулевую гипотезу для варианта с наименьшим P, если это наименьшее P меньше, чем α/N.
Разумно применять, если главное, что нам нужно: узнать есть ли эффект хоть в одном из вариантов.
Если нам нужны все варианты, в которых есть эффект, то такая поправка слишком «строгая» (= мощность критерия мала)

Поправка Шидака (Šidák correction)

Пусть есть N вариантов и задан уровень достоверности α .
Отклоняем нулевую гипотезу для варианта с наименьшим P, если это наименьшее P меньше, чем $1-(1-\alpha)^{1/N}$,
что равносильно $1-(1-\mathrm{P})^N<\alpha$

Чуть-чуть мощнее, чем Бонферрони.

Раскладывая по степеням P, имеем: $(1-\mathrm{P})^N=1-N\mathrm{P}+N(N-1)\,\mathrm{P}^2/2+o(\mathrm{P}^2)$
Поэтому при маленьких P условие практически превращается в
$N\mathrm{P}-N(N-1)\mathrm{P}^2/2<\alpha$ (ср. с Бонферрони: $N\mathrm{P}<\alpha$)

Поправка Холма – Бонферрони (Holm–Bonferroni method)

Нумеруем варианты так, чтобы полученные в них p-value шли по возрастанию:
$P1 < P2 < P3 < ... < PN$.

Если есть такое m, что для всех $k \le m$ имеем $Pk < \alpha/(N + 1 - k)$, то отвергаем нулевую гипотезу для всех вариантов 1, ... , m
В частности, если такое верно для всех k = 1, ... , N, то отвергаем нулевую гипотезу для всех вариантов.
Если же $P1 > \alpha/N$, то не отвергаем ни для какого варианта.
Тем самым, для «лучшего» варианта поправка равносильна Бонферрони.
Данный метод лучше простой поправки Бонферрони тем, что позволяет выявить больше вариантов, для которых эффект достоверен.

False discovery rate

FDR — это средняя доля ложных отклонений нулевой гипотезы
(какая доля наших «открытий» имеет случайные причины?)
Метод Бенджамини — Хохберга (Benjamini–Hochberg procedure)
Пусть мы готовы «сделать ложное открытие» в 5% вариантов.
Положим $\alpha = 0{,}05$.
Опять упорядочим Pi по возрастанию:
$P1 < P2 < P3 < ... < PN$.

Найдём наибольшее k такое, что $Pk < k\alpha/ N$
После этого считаем, что в вариантах 1, ..., k эффект есть
(имея в виду, что среди них можно ожидать 5% вариантов без эффекта)

Пермутации по Westfall-Young

Хочу другие p-values, и пусть они уже знают про множественность гипотез,
а независимость испытаний их вообще не волнует!
Westfall, P. H. and S. S. Young (1993). Resampling-based multiple testing. John Wiley and Sons.

$$0 \le p_1 \le p_2 \le p_3 \le p_4 \le \ldots\ldots\ldots \le p_N$$ – наши p-values

Перемешиваем исходные данные M раз так, чтобы они стали как можно менее осмысленными, но выглядели как исходные.

$0 \le p^1_1 \le p^1_2 \le p^1_3 \le p^1_4 \le \ldots\ldots\ldots \le p^1_N$
$0 \le p^2_1 \le p^2_2 \le p^2_3 \le p^2_4 \le \ldots\ldots\ldots \le p^2_N$
$0 \le p^3_1 \le p^3_2 \le p^3_3 \le p^3_4 \le \ldots\ldots\ldots \le p^3_N$
$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$
$0 \le p^M_1 \le p^M_2 \le p^M_3 \le p^M_4 \le \ldots \le p^M_N$

$$p_i^{wy} = \frac{\#k : \exists\, p_i^k \le p_i}{M}$$

ОЧЕНЬ МЕДЛЕННО!