Лекция 8.

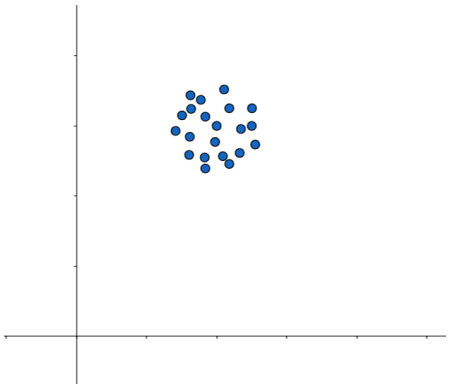
Статистическая зависимость

Def. Зависимость называется статистической, если изменение одной случайной величины вызывает изменение распределения другой. Если при этом изменяется среднее значение другой случайной величины, то такая зависимость называется коррелеционной. Если при увеличении одной случайной величины среднее значение другой также увеличивается, то говорят, что имеет место прямая корреляция. Аналогично, если уменьшается, то – обратная

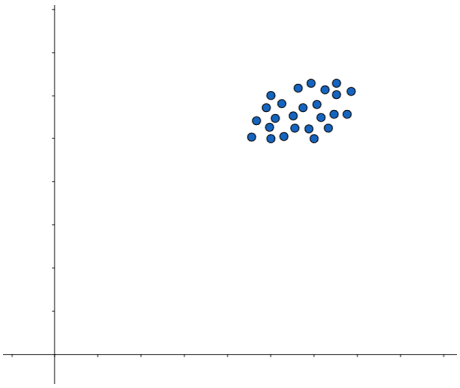
Корреляционное облако

Пусть в ходе n экспериментов появились значения двух случайных величин X и Y:
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Нанеся точки на координатную плоскость, получаем корреляционное облако, о виде которого можно делать предположения о наличии/отсутствии связи

Ex.



Здесь, возможно, нет зависимости



Здесь можно предположить прямую корреляцию

Корреляционная таблица

Пусть даны данные X и Y при n экспериментов. Эти данные удобно представить в виде коррелеционной таблицы: по вертикали отмечают различные значения x, а по горизонтали - y, в клетках таблицы отмечаются частота появления n_{xy}

Ex. $n = 50$

X\Y	10	20	30	40	n_x	$\overline{y}_x$
2	7	3	0	0	10	13
4	3	10	10	2	25	4.4
6	0	2	10	3	15	30.67
n_y	10	15	20	5	Σ 50	

По диагонали таблицы можно предположить, что корреляция есть
Имеет смысл вычислить условное среднее по формуле $\overline{y}(x) = \frac{1}{n_x}\sum n_{xy}y_i$. Так как в нашем примере условные средние растут с ростом x, то имеет место прямая корреляция

Nota. Если данных много или X и Y – непрерывные случайные величины, то лучше составить интервальную корреляционную таблицу: разбить случайные величины на интервалы, по вертикали отметить интервалы $[a_{i-1}, a_i)$ случайной величины X, по горизонтали – $[b_{j-1}, b_j)$ случайной величины Y, в клетках отметить частоты $n_{ij} : [a_{i-1}, a_i) \times [b_{j-1}, b_j)$. В дальнейшем интервалы можно заменить их серединами

Критерий «хи-квадрат» для проверки независимости

Пусть выборка $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ представлена в виде интервальной корреляционной таблицы. Случайная величина X разбита на k интервалов, а Y – на m интервалов
Обозначим $v_{i\cdot}$ – частота i-ого интервала $[a_{i-1}, a_i)$ случайной величины X, $v_{\cdot j}$ – частота j-ого интервала $[b_{j-1}, b_j)$ случайной величины Y, v_{ij} – число точек в $[a_{i-1}, a_i) \times [b_{j-1}, b_j)$

X\Y	$[b_0, b_1)$	$[b_1, b_2)$	$\ldots$	$[b_{m-1}, b_m)$	$v_{i\cdot}$
$[a_0, a_1)$	v_{11}	v_{12}	$\ldots$	v_{1m}	$v_{1\cdot}$
			$\ldots$		
$[a_{k-1}, a_k)$	v_{k1}	v_{k2}	$\ldots$	v_{km}	$v_{k\cdot}$
$v_{\cdot j}$	$v_{\cdot 1}$	$v_{\cdot 2}$	$\ldots$	$v_{\cdot m}$	Σn

Проверяется основная гипотеза $H_0 : X$ и Y независимы против $H_1 = \overline{H_0} : X$ и Y зависимы
Если H_0 верна, то $p_{ij} = P(X \in [a_{i-1}, a_i), Y \in [b_{j-1}, b_j)) = P(X \in [a_{i-1}, a_i)) \cdot P(Y \in [b_{j-1}, b_j))$
Тогда по закону больших чисел $\frac{v_{i\cdot}}{n} \xrightarrow{p} p_{i\cdot}, \frac{v_{\cdot j}}{n} \xrightarrow{p} p_{\cdot j}$
Поэтому основанием для отклонения основной гипотезы будет заметная разница между величинами $\frac{v_{i\cdot}}{n}\frac{v_{\cdot j}}{n}$ и $\frac{v_{ij}}{n}$ или v_{ij} и $\frac{1}{n}v_{i\cdot}v_{\cdot j}$

В качестве статистики берется $K = n\sum\limits_{i,j}\frac{\left(v_{ij} - \frac{1}{n}v_{i\cdot}v_{\cdot j}\right)^2}{v_{i\cdot}v_{\cdot j}}$

Th. Если H_0 верна, то $K \rightrightarrows H_{(k-1)(m-1)}$

Пусть t_α – квантиль $H_{(k-1)(m-1)}$ уровня α, тогда

$$\begin{cases} H_0 : X \text{ и } Y \text{ независимы, если } K < t_\alpha \\ H_0 : X \text{ и } Y \text{ зависимы, если } K \geq t_\alpha \end{cases}$$

Nota. Для работы критерия необходимо, что бы частота в каждой клетке была больше 5, а объем выборки был достаточно большой

Однофакторный дисперсионный анализ

Предположим, что на случайную величину X (результат) может влиять фактор Z (необязательно, что Z – случайная величина, эксперимент может быть управляемым)
Пусть при различных «k уровней» фактора Z получено k независимых выборок случайной величины X: $X^{(1)} = (X_1^{(1)}, \dots, X_{n_1}^{(1)}), \dots, X^{(k)} = (X_1^{(k)}, \dots, X_{n_k}^{(k)})$
Всего было получено $n = \sum\limits_{i=1}^{k} n_i$ значений

Nota. В общем говоря, распределение этих выборок отличается, поэтому эти выборки разных случайных величин

Общая, внутригрупповая и межгрупповая дисперсии

Для каждой выборки вычислим выборочное среднее и дисперсию: $\overline{x}^{(j)} = \dfrac{1}{n_j} \sum\limits_{i=1}^{n_j} X_i^{(j)}$, $D^{(j)} = \dfrac{1}{n_j} \sum\limits_{i=1}^{n_j} (X_i^{(j)} - \overline{x}^{(j)})^2$
Объединим все выборки в общую и также вычислим выборочнее среднее и дисперсию:
$\overline{x} = \dfrac{1}{n} \sum\limits_{i,j} x_i^{(j)} = \dfrac{1}{n} \sum\limits_{j=1}^{k} n_j \cdot \overline{x}^{(j)}$ - общее среднее
$D_{\text{О}} = \dfrac{1}{n} \sum\limits_{i,j} (X_i^{(j)} - \overline{x})^2$ - общая дисперсия

Def. Внутригрупповой (или остаточной) дисперсией называется среднее групповых дисперсий: $D_{\text{В}} = \dfrac{1}{n} \sum\limits_{j=1}^{k} n_j D^{(j)}$

Def. Межгрупповой (или факторной) дисперсией называется величина $D_{\text{М}} = \dfrac{1}{n} \sum\limits_{j=1}^{k} n_j (\overline{x} - \overline{x}^{(j)})^2$

Th. О разложении дисперсии. Общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсией: $D_{\text{О}} = D_{\text{В}} + D_{\text{М}}$

Смысл: внутригрупповая дисперсия показывает средний (случайный) разброс внутри выборок, межгрупповая - насколько отличаются среднее при различных уровнях фактора, то есть именно эта величина отражает влияния фактора

Вывод по наличии корреляции можно сделать, если доля $D_{\text{М}}$ достаточно велика

Проверка гипотезы о влиянии фактора

Предполагаем, что X имеет нормальное распределение и фактор Z может влиять только на ее математическое ожидание, но не на дисперсию и тип распределения, поэтому можно считать, что данные независимых k выборок при различных уровнях фактора Z также имеют нормальное распределение с одинаковой дисперсией: $X^{(j)} \in N(a_j, \sigma^2)$

Проверяется основная гипотеза $H_0 : a_1 = a_2 = \cdots = a_k$ (фактор не оказывает влияния) против $H_1 = \overline{H_0}$: есть влияние

По пункту 3 основной теоремы $\sum\limits_{i=1}^{n} \left(\dfrac{x_i - \overline{x}}{\sigma} \right)^2 = \dfrac{nD^*}{\sigma^2} \in H_{n-1}$

Из этого $\dfrac{n_j D^{(j)}}{\sigma^2} \in H_{n_j - 1} \quad \forall\ 1 \le j \le k$

Так как распределение «хи-квадрат» устойчиво относительно суммирования, то $\sum\limits_{j=1}^{k} \dfrac{n_j D^{(j)}}{\sigma^2} \in H_{n-k}$, так как $(n_1 - 1) + \cdots + (n_k - 1) = n - k$

Пусть основная гипотеза верна, тогда все данные можно считать выборкой одной случайной величины и по пункту 3 $\dfrac{nD_{\text{О}}}{\sigma^2} \in H_{n-1}$

Согласно теореме о разложении дисперсии $D_{\text{О}} = D_{\text{В}} + D_{\text{М}}$, тогда $\dfrac{nD_{\text{О}}}{\sigma^2} = \dfrac{nD_{\text{В}}}{\sigma^2} + \dfrac{nD_{\text{М}}}{\sigma^2}$

Так как $\dfrac{nD_{\text{О}}}{\sigma^2} \in H_{n-1}$, $\dfrac{nD_{\text{В}}}{\sigma^2} \in H_{n-k}$, то $\dfrac{nD_{\text{М}}}{\sigma^2} \in H_{k-1}$

Тогда при верной основной гипотезе получим, что $\dfrac{nD_{\text{М}}}{\sigma^2(k-1)} \dfrac{\sigma^2(n-k)}{nD_{\text{В}}} = \dfrac{n-k}{k-1} \dfrac{D_{\text{М}}}{D_{\text{В}}} \in F(k-1, n-k)$ – распределение Фишера-Снедекера со степенями $k-1$ и $n-k$

В качестве статистики берется $K = \dfrac{n-k}{k-1} \dfrac{D_{\text{М}}}{D_{\text{В}}}$, в качестве критической точки t_α – квантиль $F(k-1, n-k)$ уровня α

$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_k \text{ (фактор не оказывает влияние), если } K < t_\alpha \\ H_1 : \text{фактор влияние оказывает, если } K \ge t_\alpha \end{cases}$