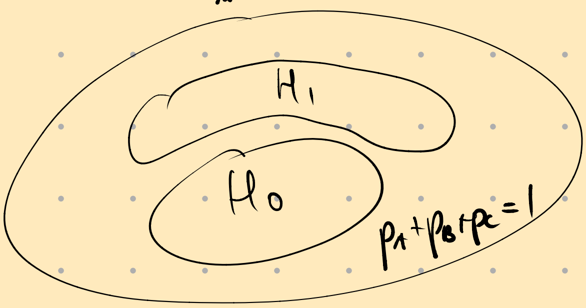
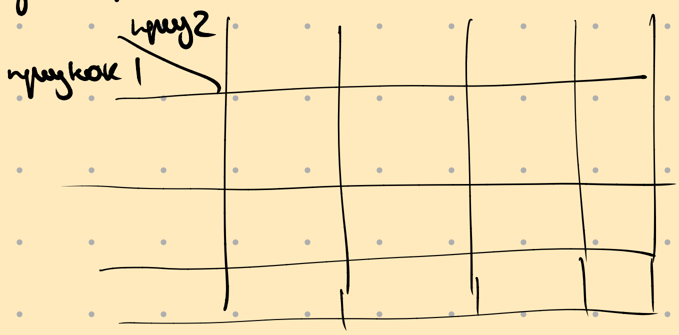
Тест χ^2 Пирсона.

$$H_0: \begin{cases} p_A = p_A(\theta) \\ p_B = p_B(\theta) \\ p_C = p_C(\theta) \end{cases} \qquad H_1: \text{[возм-ны вар-ты]}$$



Тест Пирсона на независимость признаков.

Если:
есть два дискретных (кач-х) признака



кол-во наблюдений в клетке (i,j) N_{ij}

у признака 1 возм-но k_1 значений
у признака 2 возм-но k_2 значений.

$\forall$ наблюдение попадает в одну из $k_1 \times k_2$ „клеток".
наблюдения незав-мы и один-во распр-ны.

H_0: признаки независимы
H_1: признаки зависимы

!рек! если пар-р θ — скалярный, то вместо тестирования гипотезы $H_0: \theta = 42$ всегда наглядней строить CI

N_{ij} – фактич. кол-во набл-ий, попавших в клетку (i,j)

$\hat{N}_{ij} = E(N_{ij} \mid \text{незав}, N_{i\cdot}, N_{\cdot j})$

O_{ij} E_{ij}

observed expected

$N_{i\cdot} = \sum_j N_{ij}$ – кол-во набл-ий в i-ой строке

$N_{\cdot j} = \sum_i N_{ij}$ – кол-во набл-ий в j-м столбце.

$N = \sum_{ij} N_{ij}$

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - N\hat{p}_i)^2}{N\hat{p}_i} = \sum \frac{N_i^2}{N\hat{p}_i} - N$$

$$T = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(N_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}} = \sum_{ij} \frac{N_{ij}^2}{\hat{N}_{ij}} - N$$

ТО:

при $N \to \infty$ и верной H_0

$$T \xrightarrow{dist} \chi^2_d$$

общий рецепт: $d = df_{UR} - df_R$

df_R = число свободных параметров в ограниченной модели (R=restricted) (при верной H_0)

df_{UR} = число своб-х пар-ров в неогранич. ой модели (UR = unrestricted) (при верной H_0 или H_1)

N_{ij}

Пример.

пол \ кофе	латте	кап	экспр.	Σ
M	$10_{N_{11}}$	10	10	30 $N_{1\bullet}$
F	20	30	20	70
Σ	$30_{N_{\bullet 1}}$	40	30	100 N

$N = 100$

$\hat{N}_{ij}$

	л	к	э
M	9	12	9
F	21	28	21

H_0: пол и предп. кофе незав.

H_1: пол и предп. кофе зав.

- протестировать H_0 на $\alpha = 5\%$

H_0: $P(A \cap B) = P(A) \cdot P(B)$

$$E(N_{ij} \mid H_0, N_{i\bullet}, N_{\bullet j})$$

$$E(N_{11} \mid H_0, \underline{N_{1\bullet}}, \underline{N_{\bullet 1}}) = N \cdot \hat{p}_{1\bullet} \cdot \hat{p}_{\bullet 1} = N \cdot \frac{N_{1\bullet}}{N} \cdot \frac{N_{\bullet 1}}{N}$$

$$\hat{N}_{ij} = \frac{N_{i\bullet} \cdot N_{\bullet j}}{N}$$

$$T = \sum_{i,j} \frac{(N_{ij} - \hat{N}_{ij})^2}{\hat{N}_{ij}} = \sum_{i,j} \frac{N_{ij}^2}{\hat{N}_{ij}} - N = \ldots = 0.8$$

$$\chi^2_{obs} = 0.8$$

Теорема: если $\boxed{\ldots}$ то $T \xrightarrow{dist} \chi^2_d$

$$d = df_{UR} - df_R$$

UR - модель (неогр.) : признаки м. быть зависимы / не зависимы. H_0 или H_1

p_{11}	p_{12}	p_{13}
p_{21}	p_{22}	p_{23}

любые!

ур-ие: $p_{11}+p_{12}+p_{13}+p_{21}+p_{22}+p_{23}=1$

$$d_{UR} = \underbrace{6}_{\text{пар-в}} - \underbrace{1}_{\text{ур-ие}} = 5 \text{ своб-х пар-в.}$$

R-модель (ограничение) H_0 верна (верю в незав)

Сколько своб-х пар-в?

	β_1	β_2	β_3
α_1			
α_2			

$p_{ij} = \alpha_i \cdot \beta_j$

$\alpha_1 + \alpha_2 = 1$

$\beta_1 + \beta_2 + \beta_3 = 1$

$$d_R = \underbrace{2+3}_{\text{число пар-в}} - \underbrace{2}_{\text{число ур-ий}} = 3$$

число своб-х пар-в, если верить в H_0

$$d = df_{UR} - d_R = 5 - 3 = 2$$

общая формула:

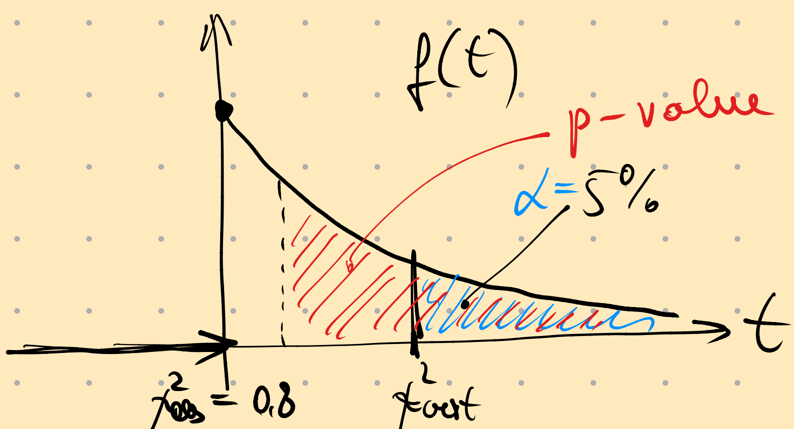
ур-ие $\sum p_{ij} = 1$

$$df_{UR} = k_1 \cdot k_2 - 1$$

$$df_R = k_1 - 1 + k_2 - 1 = k_1 + k_2 - 2$$

$$d = k_1 \cdot k_2 - 1 - (k_1 + k_2 - 2) =$$
$$= k_1 \cdot k_2 - k_1 - k_2 + 1 = (k_1 - 1)\cdot(k_2 - 1)$$

$$T \xrightarrow[H_0]{dist} \chi^2_2$$



$\chi^2_{crit} = 5.99$

[illegible] статистики [illegible])

Способ 1 (по крит. области для ст...)

$$\chi^2_{obs} = 0.8$$

если $\chi^2_{obs} > \chi^2_{crit}$, то H_0 отвергается
[иначе – H_0 не отвергается]

вывод: H_0 не отвергается.
[данные не противоречат гипотезе о независимости признаков]

Способ 2 (через p-value)

$$p\text{-value} = P(T_{new} > T \mid H_0, T)$$

$$= P(\chi^2_2 > 0{,}8) = 0.67$$

$\boxed{\text{Стандарт}}$

если p-value $< \alpha$, то H_0 отвергается
если p-value $\geq \alpha$, то H_0 не отвергается

p-value $= 0.67 > \alpha = 0.05 \Rightarrow H_0$ не отвергается.

$$P(\text{отвергнем } H_0 \mid H_0 \text{ верна}) \xrightarrow{n\to\infty} 0.05$$

0.01

$\boxed{\text{общая ситуац.}}$ c $P(\text{отверг } H_0 / H_0 \text{ верна}) \downarrow$
$P(\text{не отверг } H_0 / H_1 \text{ верна}) \uparrow$

Переход от наблюдений к рангам

X_1	Y_1
X_2	Y_2
$\vdots$	$\vdots$
X_n	Y_n

(X_i, Y_i) – незав. независимы, одинак. расп.

Теор. корреляция

$$\rho = \frac{\operatorname{cov}(X_1, Y_1)}{\sqrt{\operatorname{Var} X_1 \operatorname{Var} Y_1}} = \frac{C_{XY}}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}} \in [-1;1]$$

$$\hat{\sigma}^2_X = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$\hat{\sigma}^2_Y = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

$$\hat{C}_{XY} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n-1}$$

$E(\hat{\sigma}^2_X) = \sigma^2_X$

$E(\hat{C}_{XY}) = C_{XY}$

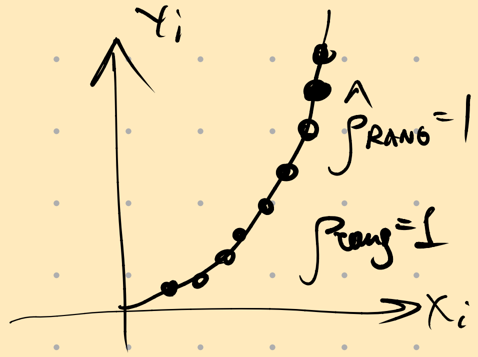
оценка корреляции

$$\hat{\rho} = \frac{\hat{C}_{XY}}{\sqrt{\hat{\sigma}^2_X \cdot \hat{\sigma}^2_Y}} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \in [-1; 1]$$

$\hat{\rho}$ — оценивает „силу" линейной зависимости.

Что делать для оценки „силы" нелинейной связи?

если $X_i > 0$ $\quad Y_i = X_i^2$



$\hat{\rho} < 1$

$\rho < 1$

$\boxed{\text{идея:}}$ перейти от (X_i, Y_i) к рангам.

ранг наблюдения — номер в порядке возрастания [по убыванию]

$$\underset{N3}{5.3}, \ \underset{N1}{2.8}, \ \underset{N2}{4.7} \longrightarrow 3, 1, 2$$

$$\begin{array}{|c|c|} X_1 & Y_1 \\ \vdots & \vdots \\ X_n & Y_n \end{array} \longrightarrow \begin{array}{|c|c|} R_1^X & R_1^Y \\ R_2^X & R_2^Y \\ \vdots & \vdots \\ R_n^X & R_n^Y \end{array}$$

оценка ранговой корр-ции

$$\hat{\rho} = \frac{\sum (R^X - \bar{R}^X) \cdot (R^Y - \bar{R}^Y)}{\ldots}$$

$$\hat{\rho}_{RANG} = \frac{\sum (R_i^x - \bar{R}^x)(R_i^y - \bar{R}^y)}{\sqrt{\sum (R_i^x - \bar{R}^x)^2 \cdot \sum (R_i^y - \bar{R}^y)^2}}$$

$$\rho_{RANG} = \operatorname{Corr}(R_i^x, R_i^y)$$

$\rho_{RANG} = 1 \iff$ идеал-ая линейная завис-сь между рангами. $\iff$

$\iff$ монотонно возрастающая зависимость между X и Y