χ^2 тест сагласности са расподелом

У даље тестирамо $H_0: F = F_0$ vs $H_1: F \neq F_0$, при чему F не мора бити апс. непр.

Идеја: 1. поделимо домен обележја у k дисјунктних категорија $B_1, \dots, B_k$

↓

скуп свих могућих вредности које обележје може узети ако важи или расподелу F_0 (тј. ако важи A_0)

2. пребројимо колико је елемената из узорка пало у j-ту категорију и тај број означимо са m_j

(4) 3. Ако је H_0 тачна сл. вел M_j која представља број X_i-ева који су узели вредности из j-те категорије има $Bin(n, p_j)$, $p_j = P\{X \in B_j \mid H_0 \text{ тачна}\}$

$M_j = \sum_{k=1}^{n} I_k$, $I_k = I\{X_k \in B_j\}$, $I_1, \dots, I_n$ су независне!

Зато ћемо посматрати колико се свако m_j разликује од np_j, што је очекивање од M_j ако важи H_0

Ако је H_0 тачна, m_j је блиско $\mathbb{E} M_j = np_j$

Сада смо на корак од формирања теста статистике. Поменимо прво једну теорему:

(T) Пирсонова теорема

Нека је $p_j = P\{X \in B_j\}$, $j \in \{1, \dots, k\}$, при чему $B_i \cap B_j = \emptyset$ за $i \neq j$, $\sum_{j=1}^{k} p_j = 1$

Нека је $(X_1, \dots, X_n)$ ПСУ, где $X_i \overset{d}{=} X$, n велико

!

Нека је M_j бр. X_i-ева који ће узети вредности из B_j. Тада $M_j \sim Bin(n, p_j)$, $EM_j = np_j$ и:

$$\sum_{j=1}^{k} \frac{(M_j - np_j)^2}{np_j} \sim \chi^2_{k-1}$$

Дакле, за тестирање

$H_0: F = F_0$ vs $H_1: F \neq F_0$

када је обим узорка велики можемо користити тест статистику

$$T = \sum_{j=1}^{k} \frac{(M_j - np_j)^2}{np_j}, \quad T \underset{\text{при } H_0}{\sim} \chi^2_{k-1}$$

где су M_j и p_j дефинисани у кораку 3. (4)

У случају H_1 иду велике вредности $(M_j - np_j)^2$, па и велике вредности T, па је критична област:

$$W = \{T \geq c\}$$

- категорије можемо бирати произвољно, с тим што се саветује да j-ту категорију спојимо са неком другом ако је $np_j < 5$

- Ако F_0 зависи од непознатих параметара, обично* их оценимо методом максималне веродостојности. Оцене тих параметара искористимо за рачунање тест статистике T, која је сада таква да:

$$T \underset{\text{при } H_0}{\sim} \chi^2_{k-1-\text{бр. оцењених параметара}}$$

* није теоријски исправно, али се ради у пракси

51.

Видимо $P\{X=k\} = 2^{-k} = \frac{1}{2^{k-1}} \cdot \frac{1}{2^k}$, $k \in \mathbb{N}$, одговара закону геометријске расподеле са параметром $\frac{1}{2}$

Дакле, потребно је да тестирамо следеће хипотезе:

$$H_0: F = F_0 \quad \text{vs} \quad H_1: F \neq F_0$$

Где је F_0 ф-ја $G(\frac{1}{2})$ расподеле. (у овом задатку нам F_0 не треба, наводимо само како гласе H_0 и H_1)

- Шта читамо из дате табеле у задатку?
Видимо да се једнашца у узорку појавила 45 пута, двојка 30 пута, тројка 15 пута и четворка 10 пута.

Ова табела нам може послужити да одредимо категорије у које ћемо разврстати шансе из узорка.
Природно би било да за категорије узмемо скупове $\{1\}, \{2\}, \{3\}$ и $\{4\}$, али као што смо рекли, унија свих категорија мора покрити цео домен обележја X ако оно задовољава $G(\frac{1}{2})$ расподелу.
Како сл. величина из $G(\frac{1}{2})$ расподеле узима вредности из скупа $\mathbb{N}$, можемо поделу на категорије извршити на следећи начин:

$$B_1 = \{1\}, \quad B_2 = \{2\}, \quad B_3 = \{3\}, \quad B_4 = \{4, 5, \ldots\}$$

(није грешка писати B_j као интервале у збиру, али обоје мало природније..)

- Дакле наше категорије - B_j и бројеви тачака из узорка које су се нашле у њима - m_j су:

B_j	$\{1\}$	$\{2\}$	$\{3\}$	$\{4,5,\dots\}$
m_j	45	30	15	10

- Да бисмо одредили тежу статистику, потребно је да израчунамо p_1, p_2, p_3 и p_4.

$$p_1 = P\{X \in B_1 \mid H_0 \text{ тачна}\} = P\{X=1 \mid X \sim G(\tfrac{1}{2})\} = \frac{1}{2^1}$$

слично, $p_2 = \frac{1}{2^2}$, $p_3 = \frac{1}{2^3}$.

p_4 можемо добити као $1 - \frac{1}{2} - \frac{1}{2^2} - \frac{1}{2^3} = 1 - \frac{7}{8}$

- Требало би да се уверимо да ли је захтев $np_j \geq 5$ (ако није, спојићемо неке категорије):

$$n = m_1 + m_2 + m_3 + m_4 = 100$$

$$\Rightarrow np_1 = 50,\ np_2 = 25,\ np_3 = 12{,}5,\ np_4 = 12{,}5$$

- Дакле, категорије не треба спајати.
- Израчунајмо сада тежу статистику:

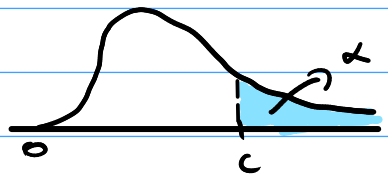
$$T = \sum_{j=1}^{4} \frac{(M_j - np_j)^2}{np_j} \Rightarrow \text{реализована вредност } gT \text{ је:}$$

$$\frac{(45-50)^2}{50} + \frac{(30-25)^2}{25} + \frac{(15-12{,}5)^2}{12{,}5} + \frac{(10-12{,}5)^2}{12{,}5}$$

$$= 2.5$$

Одредимо сада критичну област. Рекли смо да је она облика $W=\{T>c\}$.

c је т.ј. $P\{T>c \mid H_0 \text{ тачна}\}=\alpha$, а како при H_0 $T \sim \chi^2_{k-1} = \chi^2_{4-1}$, за c важи:



$$F_{\chi^2_3}(c) = 1-\alpha$$

$$\Rightarrow c = F^{-1}_{\chi^2_3}(1-\alpha)$$

У R-у: $c =$ qchisq(1- 0,05, 3)

↓ α код нас

---

52. Ако са X означимо обележје које представља одштампане грешке особине и са F означимо ф-ју расподеле обележја, потребно је тестирати:

$$H_0: F = F_0 \quad \text{vs} \quad H_1: F \neq F_0$$

где је F_0 ф-ја $N(m, \sigma^2)$ расподеле.

Видимо да у овом задатку F_0 зависи од непознатих параметара, па је потребно да их оценимо. Рекли смо да, када непознате параметре оцењујемо методом макс. веродостојности, ће су тражене оцене:

$$\hat{m}_{mmv} = \overline{X}_n \quad \text{и} \quad \hat{\sigma}^2_{mmv} = \overline{S}_n^2$$

Међутим, ове статистике не можемо израчунати на основу података које имамо, па ћемо користити модификоване оцене $\bar{X}_n^*$ и $\bar{S}_n^{2*}$, које се добијају као узорачка средина и узорачка дисперзија узорка сачињеног од средина датих интервала.

Тај узорак код нас изгледа овако:

$$(\underbrace{2.5, 2.5, \dots, 2.5}_{15\times}, \underbrace{7.5, \dots, 7.5}_{75\times}, \dots, \underbrace{22.5, \dots, 22.5}_{20\times})$$

$$\Rightarrow \bar{x}_n^* = \frac{15 \cdot 2.5 + 75 \cdot 7.5 + \dots + 20 \cdot 22.5}{15 + 75 + \dots + 20} = 12{,}2$$

$$\bar{S}_n^{2*} = 25{,}4$$

Напоменимо да овакав поступак нема никакво теоријско оправдање, користимо га јер је једноставан и зато што немамо много других опција. (За радознале које занима наложене ММВ оцена на основу броја тачака у интервалима: видети ЕМ алгоритам)

Сада када имамо оцене параметара m и σ^2, можемо наставити задатак. Следећи корак је одређивање категорија.

Било би природно за категорије узети дате интервале одступања, али претходно продужити $(0,5)$ на $(-\infty,5)$ и $(20,25)$ на $(20,+\infty)$, како бисмо покрили цео $\mathbb{R}$, што је домен $\mathcal{N}(m,\sigma^2)$ расподеле.

Скицирајмо ради једноставности табелу са B_j и m_j:

B_j	$(-\infty,5)$	$[5,10)$	$[10,15)$	$[15,20)$	$[20,+\infty)$
m_j	15	75	100	50	20

Када F_0 зависи од непознатих параметара, не можемо да одредимо p_j. Ево и зашто:

$$p_1 = P\{X \in B_1 \mid H_0 \text{ тачна}\} = P\{X < 5 \mid X \sim \mathcal{N}(m,\sigma^2)\}$$

$$= P\{\frac{X-m}{\sigma} < \frac{5-m}{\sigma} \mid X \sim \mathcal{N}(m,\sigma^2)\} = \phi\left(\frac{5-m}{\sigma}\right)$$

Слично, p_2, p_3, p_4 и p_5 ће зависити од m и σ. Дакле, сада морамо оценити p_j са $\hat{p}_j$, користећи добијене оцене $\hat{m}_{mmv}$ и $\hat{\sigma}^2_{mmv}$.

Лако добијамо да је:

$$\hat{p}_1 = \phi\left(\frac{5-\hat{m}_{mmv}}{\hat{\sigma}_{mmv}}\right) = 0{,}08$$

$$\hat{p}_2 = \phi\left(\frac{10-\hat{m}_{mmv}}{\hat{\sigma}_{mmv}}\right) - \phi\left(\frac{5-\hat{m}_{mmv}}{\hat{\sigma}_{mmv}}\right) = 0{,}25$$

$$\vdots$$

$$\hat{p}_5 = 1 - \phi\left(\frac{20-\hat{m}_{mmv}}{\hat{\sigma}_{mmv}}\right) = 0{,}06$$

Можемо проверити да је $n\hat{p}_j \geq 5$, $j \in \{1,\dots,5\}$ и да није потребно групсати категорије

Оцене $\hat{p}_j$ ћемо сад искористити за рачунање тест статистике:

$$T = \sum_{j=1}^{5} \frac{(M_j - n\hat{p}_j)^2}{n\hat{p}_j},$$ која сад при H_0 има χ^2_{5-1-2} расподелу

(оценили смо 2 параметра)

Знамо да је $n = m_1 + \ldots + M_5 = 260$, па знамо све да одредимо реализовану вредност од T:

$$\frac{(15 - 260 \cdot 0{,}07)^2}{260 \cdot 0{,}07} + \ldots + \frac{(20 - 260 \cdot 0{,}06)^2}{260 \cdot 0{,}06} = 6{,}017$$

Критична област је $W = \{T \geq c\}$, где је c т.д.

$$P\{T \geq c \mid H_0 \text{ тачна}\} = \alpha, \text{ тј } P\{T \geq c \mid T \sim \chi^2_2\} = \alpha$$

$$\Rightarrow F_{\chi^2_2}(c) = 1 - \alpha \Rightarrow c = F^{-1}_{\chi^2_2}(1-\alpha)$$

у R-у: c = qchisq(1-0,05, 2)

c = 5,991 ↓ наше α