Час 1

- Популација, обележје
- узорак, репрезентативност узорка
- типови обележја:
  - квалитативно (категоричко):
    - номинално: крвна група, пол, сешуално опредељење, вериска припадност...
    - ординално: разред, интензитет бола, статус студената (будзет, самофинансирајући)
  - (нумеричко):
    - дискретно: број деце, оцена на испиту, број искори71ених дана одмора...
    - непрекидно: тежина, висина, време чекања у реду у банци....
- дескриптивне статистике: мода,медијана,узорачка средина,распон узорка, интерквартилно растојање,узорачка дисперзија
- графичко представљање података: полигони фрквенција, кумулативни полигони фреквенција, хистограми (апсолутних, релативних фреквенција, густине), боксплот, барплот....
- идентификација аутлајера

Час 2

- емпиријска функција расподеле је дефинисана са

$$F_n(x) = \frac{\sum_{i=1}^{n} I\{X_i \leq x\}}{n}$$

  Неке њене особине су:

  - $E(F_n(x)) = F(x)$, $D(F_n(x)) = \frac{F(x)(1-F(x))}{n}$
  - $nF_n(x)$ има $\mathcal{B}(n, F(x))$ расподелу
  - уз одговарајуће скалирање, $F(x)$ се може апроксимирати нормалном расподелом

теор. м.	узор. м.	теор. цент. м.	узор. цент. м.
EX	$\bar{X}_n$	$--$	$--$
EX^2	$\frac{\sum X_i^2}{n}$	DX	$\bar{S}_n^2$
EX^3	$\frac{\sum X_i^3}{n}$	$E(X-EX)^3$	$\frac{\sum (X_i-\bar{X}_n)^3}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
EX^k	$\frac{\sum X_i^k}{n}$	$E(X-EX)^k$	$\frac{\sum (X_i-\bar{X}_n)^k}{n}$

Табела 1: Теоријски и одговарајући узорачки моменти

  - неопадајућа је функција
  - функција за одређивање у R-у је ecdf(x)

- Оцењивање параметара EX са $\bar{X}_n$ и DX са поправљеном узорачком дисперзијом $\tilde{S}_n^2$;
- $E\bar{X}_n = EX_1$ и $E\tilde{S}_n^2 = DX_1$
- оцењивање параметара расподеле, опис статистичког модела и поставка проблема
- метод момената за оцењивање θ, при чему θ може бити вишедимензионалан параметар. Оцене непознатих параметара се добијају као решење система једначина који се добије кад се изједначе теоријски моменти са одговарајућим узорачким моментима. Илустровали смо метод на оцењивању параметара униформнр $\mathcal{U}[a,b]$ расподеле, нормалне $\mathcal{N}(m,\sigma^2)$, експоненцијалне $\mathcal{E}(\lambda)$, Пуасонове $\mathcal{P}(\lambda)$ и биномне $\mathcal{B}(N,p)$ расподеле.
- особине оцена: непристрасност, постојаност, асимптотска непристрасност
- поређење оцена: оцена $\hat{\theta}_n^{(1)}$ је боља од $\hat{\theta}_n^{(2)}$ (за параметар θ) у средњеквадратном смислу ако је $E(\hat{\theta}_n^{(1)}-\theta)^2 < E(\hat{\theta}_n^{(2)}-\theta)^2$

Час 3

Метод максималне веродостојности

Основни принцип овог метода је да је оцена непознатог параметра (који може бити вишедимензионални) вредност која максимизира функцију веродостојности.

У случају дискретног обележја функција веродостојности је

$$L(\theta) = P_\theta\{X_1 = x_1, ..., X_n = x_n\}.$$

У случају простог случајног узорка

$$L(\theta) = \prod_{i=1}^{n} P_\theta\{X_i = x_i\}.$$

У случају апсолутно непрекидног обележја функција веродостојности је

$$L(\theta) = f_\theta(X_1, ..., X_n).$$

У случају простог случајног узорка

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Веома често је лакше маскимизирати неку монотону трансформацију функције веродостојности. Најчешће се максимизира $\log L(\theta)$.

Оцена добијена овомо методом не мора бити јединствена (радили смо пример $X \sim U[\theta - 1, \theta + 1]$ расподелу.)

Оцене добијене овом методом имају следеће лепо својство: Нека је g нека функција. *У* колико је $\hat{\theta}_n$ оцена методом максималне веродостојности за θ онда је $g(\hat{\theta}_n)$ оцена методом максималне веродостојности за $g(\theta)$.

На часу смо илустровали метод прмерима обележја из Пуасонове $\mathcal{P}(\lambda)$, нормалне $\mathcal{N}(m, \sigma^2)$, биномне $\mathcal{B}(N, p)$, униформне $\mathcal{U}[0, \theta]$ и униформне $\mathcal{U}[\theta - 1, \theta + 1]$ расподеле и дискретне расподеле за коју важи $P\{X = -1\} = P\{X = 1\} = \theta$ и $P\{X = 0\} = 1 - 2\theta$.

Час 4

На часу смо објаснили основне принципе Бајесовске статистике и приказали сличности и разлике са класичним (фреквенионистичким) приступом. Главна разлика је што је у Бајесовском свету непознат параметар случајна величина са неком (априорном расподелом). Та расподела се мења након сазнања о реализованом узорку. Управо та (апостериорна) расподела је основ за Бајесовско оцењивање. Најчешће се за оцену парамтера узима математичко очекивање у односу на апостериорну расподелу.

На часу смо урадили примере за оцењивање параметра p код Биномне расподеле у случају да је априорна расподела униформна, затим

Бета и нека дискретна расподела. Поред тога, Бајесовско оцењивање смо илустровали на примеру оцене параметра средње вредности Пуасонове расподеле у случају да је априорна расподела експоненцијална расподела.

Час 5

Основни састојци сваког статистичког теста су:

- Нулта хипотеза (H_0) и алтернативна хипотеза (H_1) (хипотеза која се прихвата уколико одбацујемо H_0); Одабир нулте и алтернативне хипотезе спада у дизајн експеримента. То не захтева математичке способности истраживача већ креативност и искуство.
- Тест статистика-статистика на основу чије реализоване вредности доносимо закључак.
- Критична област W (нека врста правила). Уколико реализована вредност тест статистике упадне у критичну област одбацујемо хипотезу.

Најважније је да се добро поставе хипотезе јер од тога зависе сви даљи закључци.

Пример 1. *Сматра се да студенти Математичког факултета натпросечне спадају у натпросечне грађане. С циљем да се ове тврдње оправдају насумично је одабрано 30 студената Математичког факултета и измерен им је IQ.*

У овој ситуацији је природно да нулта хипотеза буде да је просечан IQ студената Математичког факултета 100 против алтернативе да је већи од 100.

Приликом статистичког закључка могуће је направити грешке.

H_0	тачна	нетачна
прихватимо	+	−
одбацимо	−	+

Уколико одбацимо нулту хипотезу која је тачна направили смо грешку прве врсте. Уколико не одбацимо нетачну нулту хипотезу направили смо грешку друге врсте.

Вероватноћа грешке прве врсте се назива ниво значајности теста и означава са α. Вероватноћа грешке друге врсте се означава са β. $1-\beta$

представља моћ теста. Тест је моћнији уколико боље одбацује нетачне хипотезе.

Мера теста је α за које је $\sup_{H_0} P\{\text{грешка } I \text{ врсте}\} = \alpha$.

Добар пример за илустрацију врсте грешака и њиховог значаја је суђење оптуженику при чему ако се докаже да је крив, следује му смртна казна. Свако је невин док се не докаже супротно. Дакле, H_0 је да је оптужени невин, а H_1 да је крив. Грешка прве врсте би била да невин човек страда, док би грешка друге врсте била да је кривац на слободи.

Још треба напоменути да хипотезе могу бити просте и сложене. Просте су оне за које је расподела тест статистике једнозначно одређена. Ми ћемо се у овом курсу углавном сретати са простим хипотезама.

Ниво значајности теста се увек задаје пре тестирања. Најчешће вредности су $0.1, 0.05$ и 0.01. Следећи корак је да се за задати ниво значајности теста одреди критична област. Јасно је да је за то потребна расподела тест статистике под нултом хипотезом. Међутим то није увек једноставно одредити, а некада чак није ни могуће. Зато се често расподела оцењује Монте Карло методама. На часу смо приказали алгоритам за оцењивање моћи теста (чији је саставни део оцењивање расподеле тест статистике под нултом хипотезом.)

Неки тестови који се односе на параметре нормалне расподеле

Претпоставимо да обележје X има нормалну $\mathcal{N}(m, \sigma^2)$ расподелу, при чему на располагању имамо прост случајан узорак $X_1, ..., X_n$. Често се јавља потреба за тестирањем да m има баш неку одређену вредност m_0. Дакле, потребно је тестирати $H_0 : m = m_0$ против неке алтернативе. Најчешће три алтернативе су:

- $H_1 : m \neq m_0$;
- $H_1 : m < m_0$;
- $H_1 : m > m_0$;

Разликујемо два случаја, када је σ^2 познато и када није.

У првом случају користи се тест статистика

$$T_n = \frac{\bar{X}_n - m_0}{\frac{\sigma}{\sqrt{n}}}. \tag{1}$$

Уколико је H_0 тачно, из особина нормалне расподеле, закључујемо да T_n има $\mathcal{N}(0,1)$. Ово нам је јако битно за одређивање критичне области.

$\bar{X}_n$ је оцена за m тако да ако H_0 није тачно T_n неће бити "довољно блиско нули". У случају природно је да је критична област облика $W = \{|T_n| > C\}$ јер је расподела тест статистике под нултом хипотезом симетрична, а много мале и много велике вредности тест статистике упућују на алтернативну хипотезу. Константу C одређујемо из услова $P_{H_0}\{|T_n| > C\} = \alpha$. Овај услов се може записати у облику $\Phi(C) = 1 - \frac{\alpha}{2}$, па је $C = \Phi^{-1}(1 - \frac{\alpha}{2})$.

Сличним разматрањем закључујемо да је облик критичне области за $W = \{T_n < C\}$, док је за $W = \{T_n > C\}$, и добијамо вредности за C у функцији од α.

У случају да је σ^2 непознато онда у изразу (1) σ^2 замењујемо непристрасном оценом $\widetilde{S}_n^2$. Тестирање се обавља на сличан начин. Једино што је другачије у овом случају је расподела тест статистике под нултом хипотезом. Сада T_n има t_{n-1} расподелу.

Још један битан појам у статистичким тестирањима је p- вредност теста. То је најмањи ниво значајности теста за који ћемо, на основу датог узорка, одбацити H_0. Тако да ако је $p < \alpha$ онда одбацујемо хипотезу, у супротном је прихватамо. Нпр. у описаном тестирању, уколико је алтернатива p-вредност је $P\{T_n < \hat{T}_n\}$.

Час 6

Нека је X обележје које има нормалну $\mathcal{N}(m, \sigma^2)$. На прошлом часу смо видели како можемо да тестирамо нулту хипотезу да је $m = m_0$. Сада ћемо тестирати $H_0 : \sigma^2 = \sigma_0^2$ против алтератива

- $H_1 : \sigma^2 \neq \sigma_0^2$;
- $H_1 : \sigma^2 > \sigma_0^2$;
- $H_1 : \sigma^2 > \sigma_0^2$.

Како је непристрасна оцена за σ^2 поправљена узорачка дисперзија $\widetilde{S}_n^2$ природно је да управо та оцена фигурише у тест статистици. Најчешће се користи:

$$T = \frac{(n-1)\widetilde{S}_n^2}{\sigma_0^2}.$$

Уколико је H_0 тачно познато је да T има χ^2_{n-1} расподелу.

Претпоставимо да је $H_1 : \sigma^2 > \sigma_0^2$. Тада нам велике вредности тест статистике упућују на алтернативу. Зато је критична област облика $W = \{T > C\}$. Константу C одређујемо из услова $\alpha = P_{H_0}\{T > C\}$.

Тестови који се односе на два обележја из нормалне расподеле

Претпоставимо да су X и Y два обележја са $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$ расподелама. Желимо да тестирамо $H_0 : \ m_1 = m_2$. Ово можемо формулисати и као тест за $H_0 : m_1 - m_2 = 0$. Непристрасна оцена за $m_1 - m_2$ је $\bar{X}_{n_1} - \bar{Y}_{n_2}$ и ако је H_0 тачно та разлика има $\mathcal{N}(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ расподелу. Због тога је најзгодније да за тест статистику узмемо баш ову разлику адекватно нормирану да тест статистика под нултом хипотезом има баш $\mathcal{N}(0,1)$ расподелу. Тада тестирање можемо извршити на исти начин као у случају да тестирамо хипотезу о вредности параметра средње вредности једног обележја са нормалном расподелом.

Међутим, у пракси се ретко дешава да имамо познате дисперзије обележја. Слично као пре, тада те дисперзије оцењујемо на основу узорка. Како ћемо то урадити зависи од тога да ли је $\sigma_1^2 = \sigma_2^2$. Како су нам ове дисперзије непознате не можемо до закључка доћи без статистичког теста.

За тестирање о једнакости дисперзија користимо такозвани F-тест који је заснован на поређењу узорачких дисперзија из обе популације. Тест статистика је

$$T_F = \frac{\widetilde{S}^2_{n_1}}{\widetilde{S}^2_{n_2}}.$$

На предавању смо видели да ако је H_0 тачно онда T има Фишерову $\mathcal{F}_{n_1-1,n_2-1}$ расподелу.

У овој ситуацији је природно да алтернативна хипотеза буде $H_1 : \sigma_1^2 \neq \sigma_2^2$, па је критична област облика $W = \{T_F < C_1\} \cup \{T_F > C_2\}$ где се константе C_1 и C_2 одређују из услова $\frac{\alpha}{2} = P_{H_0}\{T_F < C_1\} = P_{H_0}\{T_F > C_2\}$.

Вратимо се сад тестирању једнакости очекиваних вредности.

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ Уколико смо дошли до овог закључка онда σ^2 оцењујемо на основу обједињеног узорка са

$$S^2 = \frac{(n_1 - 1)\widetilde{S}^2_{n_1} + (n_2 - 1)\widetilde{S}^2_{n_2}}{n_1 + n_2 - 2}.$$

Зато користимо тест статистику

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}.$$

Уколико је H_0 тачно T има $t_{n_1+n_2-2}$ расподелу.

$\sigma_1^2 \neq \sigma_2^2$ Сада је тест сатистика

$$T = \frac{\bar{X}_{n_1} - \bar{Y}_{n_2}}{\sqrt{\frac{\widetilde{S}_{n_1}^2}{n_1} + \frac{\widetilde{S}_{n_2}^2}{n_2}}}.$$

Ако је H_0 тачно T има t_ν где је број степени слободе

$$\nu = \frac{\left(\frac{\widetilde{S}_{n_1}^2}{n_1} + \frac{\widetilde{S}_{n_2}^2}{n_2}\right)^2}{\frac{\left(\frac{\widetilde{S}_{n_1}^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{\widetilde{S}_{n_2}^2}{n_2}\right)^2}{n_2-1}}.$$

Час 7

Нека је θ непознат параметер. Нека су L_n и U_n статистике за које је $P\{L_n \leq \theta U_n\} = \beta$. Интервал (L_n, U_n) $\beta\%$ је двострани интервал поверења за параметар θ. Аналогно се дефинишу једнострани доњи једнстрани горњи интервали поверења.

Нека су $\hat{L}_n$ и $\hat{U}_n$ реализоване вредности статистика. Тада је $(\hat{L}_n, \hat{U}_n)$ реализовани интервал поверења.

Важно је да у интерпретацији интервалних оцена не дође до забуне. **Није тачно да је $P\{\theta \in (\hat{L}_n, \hat{U}_n)\} = \beta$ јер параметар θ није случајна величина!** На основу неког другог узорка добићемо други реализовани интервал поверења, па је исправна интерпретација нивоа поверења даће се у $\beta\%$ случајева стварна вредност параметра налазити у реализованом интервалу поверења.

За налажење интервала поверења птребно је наћи неку функцију од узорка и непознатог параметра чија расподела не зависи од непознатог параметра (стожерну величину).

Пример 2. *Нека X има нормалну $\mathcal{N}(m, \sigma^2)$ расподелу при чему јсу оба параметра непозната. На располагању имамо п.с.у. $X_1, ..., X_n$.. Циљ нам је да нађемо 90% интервал поверења за m.*

Потребно је прво да нађемо помоћну функцију од узорка чију расподелу знамо, а у којој се јавља m. Једна могућност је

$$T = \frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}$$

за коју знамо да има t_{n-1} расподелу. Зато можемо одредити константу C тако да је $P\{|T| \leq C\} = \beta$. Због симетричности Студентове расподеле $X = F^{-1}_{t_{n-1}}(\frac{1+\beta}{2})$. Даље, неједнакост $|T| \leq C$ је еквивалентна са $\bar{X}_n - C\frac{\tilde{S}_n}{\sqrt{n}} \leq m \leq \bar{X}_n + C\frac{\tilde{S}_n}{\sqrt{n}}$. Одавде видимо шта су статистике L_n и U_n које смо тражили.

На часу смо видели како можемо направити интервале поверења за m када је σ^2 познато, затим интервале поверења за σ^2. Поред овога размотрили смо и случај када имамо два независна обележја X и Y са нормалним расподелама и потребно нам је да нађемо интервал поверења за разлику њихових очекивања.

Приказали смо и како се може направити интервал поверења за p код $\mathcal{B}(1, p)$ расподеле.

Час 8

Тема часа су били непараметарски тестови сагласности са расподелом односно да обележје X има расподелу F_0.

Прва група тестова које смо поменули заснована је на униформној конвергенцији емпиријске функције расподеле ка правој функцији расподеле обележја. За сада претпостављамо да је F_0 апсолутно непрекидна функција расподеле.

- Тест Колмогоров-Смирнова

$$T = \sup_x |F_n(x) - F_0(x)|$$

  За мале вредности обима узорка може се егзактно исвести расподела, док је за велико n нађена асимптотска расподеластатистике $\sqrt{n}|F_n(x) - F_0(x)|$.

  Показали смо да расподела тест статистике под нултом хипотезом не зависи од F_0 па се за одређивање критичних вредности може претпоставити да тестирамо нулту хипотезу да је узорак из униформне расподеле. Доказ је заснован на следећој особини: Ако

X има функцију расподеле F_0 онда $F_0(X)$ има унифрмну $\mathcal{U}[0,1]$ расподелу.

У случају алтернативе хипотезе $F \neq F_0$ критична област за тестирање је $W = \{T > C\}$. У R-у користимо функцију $ks.test$.

- Тест Крамер-фон Мизеса

$$T = \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 dF_0(x).$$

На сличан начин може се показати да расподела тест статистике под нултом хипотезом не зависи од F_0.

Функција у R-у коју користимо је $cvm.test$ из пакета $goftests$.

Ови тестови су предвиђени за тестирање просте нулте хипотезе, односно да су сви параметри расподеле F_0 познати. У слуйају да нису они се могу оценити. Међутим тада расподела статистике под нултом хипотезом се мења и не могу се користити критичне вредности које су добијене када параметри нису оцењени!

На часу смо видели шта радимо у тој ситуацији.

Поред ових тестова приказали смо и χ^2 тест сагласности са расподелом. У овом случају не X не мора бити апсолутно непрекидна случајна величина. Идеја је да се домен обележја X подели у k дисјунктних категорија а затим преброји број чланова из узорка у свакој од категорија и упореди са очекиваним бројем. Нека је M_j број елемената у j-тој категорији. Тада M_j има биномну $\mathcal{B}(n, p_j)$ где је $p_j = P_{H_0}(X$ је у j-тој категорији). Одавде је $EM_j = np_j$. Уколико је $np_j < 5$ спајамо категорије. Због тога формирамо тест статистику

$$T = \sum_{j=1}^{k} \frac{(M_j - np_j)^2}{np_j}.$$

Уколко је H_0 тачно Tима χ^2_{k-1} расподелу. Критична област је природно облика $W = \{T > C\}$, осим уколико не желимо и да се штитимо од "намештања података".

Ако F_0 зависи од непознатих параметара онда прво те параметре оцењујемо методом максималне веродостојности а затим вероватноће p_j одређујемо користећи управо те оцењене параметре. Тест статистика остаје иста али је сада расподела, уколико важи H_0, $\chi^2_{k-1-\text{број оцењених параметара}}$.

Час 9

На овом часу смо приказали следеће непараметарске тестове:

- Тест знакова (Sign test)

  Тестирамо H_0 $m_e = m_{e0}$, где је m_e медијана расподеле коју има обележје X. Користимо статистику

  $$T = \sum_{i=1}^{n} I\{X_i \leq m_{e0}\}.$$

  Можемо користити и "центрирану" верзију

  $$T^C = \sum_{i=1}^{n} I\{X_i \leq m_{e0}\} - \frac{n}{2}.$$

  Уколико је H_0 тачна T има $\mathcal{B}(n, \frac{p}{2})$. За велико n се може користити нормална апроксимација.

  Уколико је расподела обележја X симетрична онда се H_0 своди на H_0 $m = m_0$, где је $m = EX$.

- Вилкоксонов тест заснован на "ранговима и знаковима" (Wilkoxon sign-rank test) Тестирамо H_0 $m = m_0$. Важна претпоставка за овај тест је да је расподела обележја X симетрична. Због тога ће, ако важи H_0 $X - m_0$ имати исту расподелу као $m_0 - X$.

  Означимо са r_i ранг елемента $|X_i - m_0|$ у узорку $|X_1 - m_0|, ..., |X_n - m_0|$. Тест статистика коју је предложио Вилкоксон је

  $$T = \sum_{i=1}^{n} r_i I\{X_i - m_0 \geq 0\}$$

  Може се показати да је $ET = \frac{n(n+1)}{4}$ и да је $DT = \frac{n(n+1)(2n+1)}{24}$. Уколико је n велико, за одређивање критичне области може се корстити нормална апроксимација, тј. да $\frac{T-ET}{\sqrt{DT}}$ има стандардну нормалну расподелу.

  *Н*апомена: Радили смо и адаптације ових тестова на "спарен узорак" .

-

- Вилкоксонов тест заснован на "ранговима и знаковима" за два независна обележја.

  $H_0 : m_X = m_Y$ при чему додатно претпостављамо да расподеле X и Y су истог облика са истом дисперзијом (и симетричне су).

  Направимо обједињен узорак $X_1, ..., X_n, Y_{n+1}, ..., Y_{n+m}$. Означимо са r_i ранг сваког у обједињеном узорку.Нека је z_i индикатор да је елемент са рангом r_i из првог узорка. Тада је тест статистика

  $$T = \sum_{i=1}^{n} r_i z_i.$$

  Направимо обједињен узорак $X_1, ..., X_n, Y_{n+1}, ..., Y_{n+m}$. Означимо са r_i ранг сваког у обједињеном узорку. На следећем часу ћемо показати да је
  $ET = \frac{n(n+m+1)}{2}$ и $DT = \frac{mn(m+n+1)}{12}$.

- χ^2-тест независности

  Желимо да тестирамо H_0 да су обележја X и Y независна. Подсетимо сел то значи да је свака два скупа A и B $P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$. Зато ћемо формирати $K \times L$ категорија (K за вредности X и L за вредности Y). На располагању имамо п.с.у. $(X_1, Y_1), ..., (X_n, Y_n)$. Означимо са M_{ij} број елемената из узорка чија се X-компонента налази у i-тој категорији, и Y-компонента у j-тој категорији. Тада сличним резоновањем као у χ^2-тесту сагласности са расподелом, формирамо тест статистику

  $$T = \sum_{ij} \frac{(M_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}},$$

  при чему је $\hat{p}_{ij}$, оцењено на основу узорка под претпоставком да важи H_0, једнако $\hat{p}_{i,}\hat{p}_{.,j} = \frac{\sum_j M_{ij}}{n} \cdot \frac{\sum_i M_{ij}}{n}$. Тест статистика, уколико важи H_0, има $\chi^2_{(K-1)(L-1)}$ расподелу.

  *Н*апомена: И овде морамо извршити груписање категорија уколико је $n\hat{p}_{ij} < 5$.

- Тестови за $H_0 : \ F_X = F_Y$ засновани на емпиријским функцијама расподеле. Радили смо тестове који су аналогни тестовима сагласности заснованих на оценама функције расподеле.

– Тест Колмогоров-Смирнова

$$T = \sup_x |F_{nx}(x) - F_{my}(x)|$$

– Крамер- фон Мизесов тест

$$T = \int_{-\infty}^{\infty} (F_{nx}(x) - F_{my}(x))^2 dH_N(x),$$

где је $H_N(x)$ емпиријска функција расподеле обједињеног узорка.

Часови 10 и 11

Овај час је био посвећен простој линеарној регресији као једном од најједноставнијих регресионих модела-модека зависности једне променљиве Y (зависна променљива) од неке друге променљиве X(предиктор).

Претпоставимо да је $(X_1, Y_1), ..., (X_n, Y_n)$ прост случајан узорак.

Модел који посматрамо је

$$Y_i|X_i = aX_i + b + \varepsilon_i, \quad i = 1, 2, ..., n. \tag{2}$$

Можемо посматрати ситуацију и на следећи начин. Предиктори су нам познати и желимо да видимо колика је одговарајућа вредност зависне променљиве и писати (2)

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, 2, ..., n.$$

Низом случајних променљивих $\{\varepsilon_i\}$ се моделира "шум", односно случајна одступања од модела. Тај низ прордно треба да задовољава следеће услове:

1. $E(\varepsilon_i) = 0$;
2. $\{\varepsilon_i\}$ су међусобно некорелисане једнако расподељене случајне променљиве;
3. $E(\varepsilon_i^2) = \sigma^2$;

Ови услови се могу заменити слабијим условом да су $\{\varepsilon_i\}$ независне и једнакорасподељене случајне променљиве са нормалном $\mathcal{N}(0, \sigma^2)$ расподелом.

Приметимо да је $E(Y_i) = ax_i + b$, дакле ми овим моделом заправо средњу вредност зависне променљиве и да зависност је баш линеарна.

Поставља се питање како да оценимо непознате коефицијенте a и b. Природно је да оцене треба да буду такве да оцењена вредност што мање одступа од параве вредности. Зато је једна могућност да минимизујемо

$$S(a,b) = \sum_{i=1}^{n}(Y_i - (ax_i + b))^2.$$

Оцене ће бити решење система

$$\frac{\partial S(a,b)}{da} = 0 \qquad \frac{\partial S(a,b)}{db} = 0.$$

Добија се

$$\hat{a} = \frac{\sum Y_i x_i - n\bar{Y}\bar{x}}{\sum x_i^2 - n(\bar{x})^2}$$
$$\hat{b} = \bar{Y} - \hat{a}\bar{x}.$$

Приметимо да тачка $(\bar{x}, \bar{Y})$ припада "оцењеној правој".

На часу смо показали да, уколико важе , оцене $\hat{a}$ и $\hat{b}$ су непристрасне и постојане. Постојаност следи из израза за дисперзију оцена.

$$D(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$D(\hat{b}) = \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{nS_X^2}\Big).$$

За сада је једини проблем то што је σ^2 непознато. Означимо са $\hat{\varepsilon}_i$ резидуале модела, односно $\hat{\varepsilon}_i = Y_i - \hat{a}x_i - \hat{b}$. Даље нека је $SSE = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$. Показали смо да је $E(SSE) = \sigma^2(n-2)$. Зато ћемо σ^2 оценити са $\frac{SSE}{n-2}$.

Пре него што уведемо додатне претпоставке о моделу приметимо и следеће:

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$
$$= SSE + SSR.$$

Видимо да је укупан варијабилитет зависне променљиве представљен као сума квадрата резидуала и вариабилитета објашњеног моделом. Зато

се за меру адекватности модела може користити коефицијент детерминације $R^2 = 1 - \frac{SSE}{SST}$. Уколико је модел одговарајући R^2 је блиско јединици. Приликом евалуације модела треба имати у виду да је боље то радити на тест подацима а модел правити на тренинг скупу података.

Уколико у моделу важи додатна претпоставка о нормалности $\{\varepsilon_i\}$ важе многе лепе особине оцена. Показали смо да тада $\hat{a}$ и $\hat{b}$ имају нормалне $\mathcal{N}(a, D(\hat{a}))$ и $\mathcal{N}(b, D(\hat{b}))$. Да бисмо тестирали хипотезе у вези коефицијената модела, потребно је да оценимо дисперзије оцена. Како у оба израза дисперзију појављује σ^2, заменицемо σ^2 са непристрасном оценом $\hat{\sigma}^2 = \frac{SSE}{n-2}$. Може се показати да тада

$$\frac{\hat{a} - a}{\sqrt{\widehat{D}(\hat{a})}} \text{ и } \frac{\hat{b} - b}{\sqrt{\widehat{D}(\hat{b})}}$$

имају t_{n-2} расподелу. Сада се може извршити тестирање хипотезе $a = a_0$ и $b = b_0$. Уколико је a_0 онда се заправо тестира утицај предиктора на Y. Ако је $a = 0$ онда на Y не утиче X.

Један од главних циљева моделирања је прогнозирање. Оцена за средњу вредност Y-а кад предиктор има вредност x_0 је $\hat{Y}_0 = \hat{a}x_0 + \hat{b}$. Лако се показује да је ова оцена непристрасна, и постојана, као и да, уз додатне претпоставке о расподели ε_i, прогнозирана вредност има нормалну расподелу са дисперзиом $\sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{nS_X^2}\right)$. Уколико са $\hat{Y}_0 = \hat{a}x_0 + \hat{b}$ оценимо вредност променљиве у тацки x_0 дисперзија те оцене је значајно већа јер треба додати $D(\varepsilon)$.

Напомена 1: Оцене за a и b добијене методом најмањих квадрата су у ствари оцене добијене методом максималне веродостојности, када имамо претпоставке о нормалној расподели низа $\{\varepsilon_i\}$.

Напомена 2: У моделу може бити више предиктора, односно можемо посматрати модел

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, ...n.$$

Тада се све оцене изводе аналогно и имају слична својства као у претходно описаном случају једног предиктора.

Часови 11 и 12

На овом часу смо се бавили уопштеним линеарним моделима. У линеарном регресионом моделу смо моделирали $E(Y)$ линеаном функцијом

од предиктора. Поставља се питање да ли то можемо да урадимо у случају да Y нема нормалну расподелу. Одговор је потврдан, али уз мале модификације. Примера ради, претпоставимо да је Y индикаторнеко обележје које узима само две вредности 0 и 1. Разумно је претпоставити да $p_i = P\{Y_i = 1\}$ може зависити од предиктора. Међутим када бисмо претпоставили да је $p_i = aX_i + b$ дошли бисмо у опасност да p_i узме вредност изван свог дозвољеног опсега $(0, 1)$. Једна могуцност је да трансформишемо p_i тако да трансформисана вредност је у R а затим извршимо моделирање. Најприроднија трансформација је $F^{-1}(p_i)$, где је F функција расподеле случајне променљиве дефинисане на R. У случају да се ради о логистичкој расподели, $F(x) = \frac{1}{1+e^{-x}}$, за $x \in \mathbb{R}$, модел

$$\log \frac{p_i}{1 - p_i} = ax_i + b$$

се назива логистички регресиони модел.

Параметре a и b ооцењујемо методом максималне веродостојности. Функција веродостојности дата је са

$$L(a, b) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}.$$

Одавде је

$$l(a, b) = \sum_{i=1}^{n} \Big(Y_i \log \frac{p_i}{1 - p_i} + \log(1 - p_i) \Big) = \sum_{i=1}^{n} \Big(Y_i(ax_i + b) + \log \frac{1}{e^{ax_i+b} + 1} \Big).$$

Решавање система $\frac{\partial l(a,b)}{da} = \frac{\partial l(a,b)}{db} = 0$, се врши нумерички.

Након што одредимо $\hat{a}$ и $\hat{b}$, оцене вероватноћа су $\hat{p}_i = \frac{1}{1+e^{-(\hat{a}x_i+\hat{b})}}$. Сада на основу тога можемо вршити класификацију. Једна могућност је да ако је $\hat{p}_i > 0.5$ онда је $\hat{Y}_i = 1$, у супротном је 0. И онда можемо видети проценат добро класификованих података. Алтернативно можемо користити ROC криву којом се посматрата зависност сензитивности система од специфичности система. Ову криву можемо користити и за мерење квалитета модела, не само за одређивање најбољег прага на основу кога ћемо $\hat{Y}$ оценити са 1. Што је већа површина испод криве то је модел бољи, тако да можемо посматрати управо ту површину.

На часу смо споменули још неколико модела из класе уопштених линеарних модела.