



УНИВЕРЗИТЕТ У БЕОГРАДУ
МАТЕМАТИЧКИ ФАКУЛТЕТ



Мастер рад

Тест Колмогорова и његове модификације

Универзитет у Београду
Математички факултет
297
Београд

БЕОГРАД, 2013.

САДРЖАЈ

Увод.....	4
1. Статистички тестови	5
1.1. Статистички тестови	5
1.2. Параметарски тестови	8
1.2.1. Тестирање хипотезе о математичком очекивању код нормалне расподеле када је дисперзија непозната (све три врсте алтернативне хипотезе)	8
1.2.2. Тестирање хипотезе о математичком очекивању код нормалне расподеле када је дисперзија позната	9
1.2.3. Тестирање хипотезе о дисперзији са непознатим параметром σ	10
1.2.4. Тестирање хипотезе о једнакости дисперзија код два независна обележја из нормалне расподеле	11
1.2.5. Тестирање хипотезе о једнакости средњих вредности код два независна обележја из нормалне расподеле када су дисперзије непознате и једнаке	11
1.2.6. Тестирање хипотезе о вероватноћи код биномне расподеле	13
1.2.7. Хипотеза о коефицијенту корелације	14
1.3. Непараметарски тестови	15
1.3.1. Пирсонов χ^2 -тест	15
1.3.2. χ^2 -тест независности два обележја	16
1.3.3. Тест рангова	17
1.3.4. Тест серија	18
2. Тест Колмогорова	20
2.1. Емпиријска функција расподеле	20
2.2. Централна теорема математичке статистике	21
2.3. Тест Колмогорова	24
2.4. Тест Колмогоров-Смирнов за два узорка	30
3. Модификација теста Колмогорова	32
3.1. Тест Куипера	32
3.2. Крамер-фон Мизес тест	33
3.3. Андерсон-Дарлинг тест	34

3.4. Lilliefors тест	36
Моћ тестова	37
Закључак.....	40
Литература	42

Увод

Израз статистика се у почетку односио на прикупљање података који су били од значаја за државу, као што су евиденције о становништву, поседима и приходима, а води порекло од италијанске речи **стате-држава**. Потреба за ефикаснијом државном администрацијом, као и оснивање првих осигуравајућих друштава утицали су на развој виталне статистике (праћење и анализа рађања и умирања) у империјалној Енглеској XVII века, а пионири у овој области били су *John Graunt* (1620-1674. г.) и *William Petty* (1623-1687. г.). Готово у исто време *Blaise Pascal* (1623-1662. г.) и *Pierre de Fermat* (1601-1665. г.) поставили су основе теорије вероватноће, а у сврху повећања успеха у играма на срећу, које су биле популарне у високим друштвеним круговима у Француској.

Даљи подстицај за развој статистичке методологије дала је астрономија, где је резултате многих појединачних посматрања било потребно објединити у јединствену теорију. Водеће личности у овој области били су *Pierre-Simon Laplace* (1749-1827. г.) у Француској и *Johann Carl Friedrich Gauss* (1777-1855. г.) у Немачкој. Широка примена рачунарске технологије од осамдесетих година XX века допринела је да статистика постане једна од научних области са највећим степеном развоја у последњих тридесет година.

Статистика је област примењене математике која се бави прикупљањем, организацијом, приказивањем, анализом и интерпретацијом нумеричких података, као и доношењем статистичких закључака, а њена методологија се заснива на теорији вероватноће и закону великих бројева.

Статистика нас учи како да процењујемо поузданост прикупљених података и како да неутралишемо грешке које могу настати код свих врста процена. Како се заснива на теорији вероватноће, статистика не нуди егзактне закључке, али је њена методологија тако конципирана да увек можемо израчунати снагу и ограничења добијених резултата.

Статистика има два аспекта: теоријски и примењени. Теоријска или математичка статистика бави се развојем, извођењем и доказивањем статистичких теорема, формула, правила и закона. Примењена статистика подразумева примену тих теорема, формула, правила и закона у решавању реалних проблема. Проблематика обрађена у овом раду је једним делом припада теоријској статистици, а једним делом примењеној статистици.

У првом поглављу овог мастер рада изложени су елементи математичке статистике који су потребни да бисмо могли да извршимо статистичка тестирања. Претстављени су неки од параметарских и непараметарских тестова, са урађеним примерима.

Друго поглавље садржи обраду Тест Колмогорова и Тест Колмогоров-Смирнов за два узорка, са уводном делом у коме су претстављени сви потешни елементи за примену ових тестова. Садржи назначене теореме и дефиниције са урађеним примерима, такође има примера који су урађени и у програмском језику R.

Треће поглавље садржи модификације Теста Колмогорова, са урађеним примерима у програмском језику R.

Статистички тестови

1.1. Статистички тестови

Приликом извођења статистичких тестова постоје одређени кораци којих се треба придржавати да би закључак био поуздан:

1. Поставити нулту хипотезу
2. Изабрати ниво поузданости
3. Одредити величину узорка
4. Изабрати статистички тест за тестирање хипотезе
5. Утврдити критичну вредност за одабрани статистички тест
6. Прикупити податке
7. Израчунати статистичку величину за одабрани статистички тест
8. Донети статистички закључак
9. Изразити статистички закључак.

Нека је $(X_1, X_2, \dots, X_n)$ прост случајан узорак из расподеле $F_\theta \in \{F_\theta | \theta \in \Theta\}$. Статистичка хипотеза је свака претпоставка која сужава фамилију допустивих расподела. Хипотезу коју проверавамо називамо *нулта хипотеза* и означавамо са H_0 . Проста параметарска нулта хипотеза је облика $H_0(\theta = \theta_0)$. Заједно са нултом, тестира се и *алтернативна хипотеза* у ознаци H_1 , која је обично комплементарна нултој $H_1(\theta \neq \theta_0)$, али може бити и облика $H_1(\theta < \theta_0)$ или $H_1(\theta > \theta_0)$. Избор између две хипотезе појављује се у различитим областима примене, кад год треба доказати неко тврђење или верификовати неку нову теорију. На пример, ако се појави нови лек, произвођач мора доказати да је он бољи од постојећих. Да би доказао ту хипотезу, он мора да обори супротну хипотезу.

Ако желимо да докажемо неко тврђење, онда супротно тврђењу (или неутрално или постојеће стање) узимамо за нулту хипотезу H_0 , а само тврђење за хипотезу H_1 .

Циљ поступка тестирања јесте да се испита, на основу резултата експеримента, има ли доказа против хипотезе H_0 , а у корист хипотезе H_1 .

Тест је одређен ако је дефинисана статистика S (статистика теста) и скуп вредности за S , за које одбацујемо хипотезу H_0 (област одбацивања или критична област).

Закључак теста може бити један од следећа два:

- Одбацујемо H_0 , јер смо у експерименту добили S у области одбацивања. Као објашњење нудимо хипотезу H_1 .
- Не одбацујемо H_0 јер је вредност за S у експерименту била ван области одбацивања.

Грешка прве врсте се чини ако се одбаци тачна хипотеза H_0 . Грешка друге врсте се чини ако се прихвати хипотеза H_0 у ситуацији када она није тачна. У пракси се поступа тако што се фиксира вероватноћа грешке прве врсте (означавамо је са α), а затим се међу критичним областима величине α одређује она за коју је вероватноћа грешке друге врсте (у ознаци β) најмања. Обично се узима $\alpha = 0.05$ или $\alpha = 0.01$.

Нека је $\theta \in \Theta$ права вредност параметра.

Вероватноћа да ће нулта хипотеза бити одбачена је **моћ теста**, у ознаци $\gamma(\theta), \theta \in \Theta$. Ако је C област одбацивања H_0 , а S статистика теста, онда је $\gamma(\theta) = P(S \in C)$.

Вероватноћу *грешке прве врсте* означавамо са $\alpha(\theta)$. Она је дефинисана само за $\theta \in \Theta_0$, јер се грешка прве врсте може направити само када је H_0 тачна. Очигледно је $\alpha(\theta) = \gamma(\theta), \theta \in \Theta_0$ (вероватноћа да ће H_0 бити одбачена).

Вероватноћу *грешке друге врсте* означавамо са $\beta(\theta), \theta \in \Theta_1$. Имамо да је $\beta(\theta) = 1 - \gamma(\theta)$ за $\theta \in \Theta_1$ (вероватноћа да H_0 неће бити одбачена).

Максимална вредност грешке прве врсте је ниво значајности теста и обележава се са α :

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta).$$

Ако је област одбацивања теста облика $\{S > c\}, \{S \geq c\}, \{S < c\}$ или $\{S \leq c\}$, за број c кажемо да је критична вредност теста.

За тестирање хипотезе користе се **параметарски** и **непараметарски** тестови.

Заједничка карактеристика свих параметарских тестова је да унапред знамо расподелу обележја, а тестом проверавамо претпостављене вредности појединих параметара те расподеле (нпр. μ или σ код нормалне расподеле, p код биномне расподеле, итд ...). Разликујемо две основне категорије ових тестова.:

- параметарске тестове једног узорка;
- параметарске тестове два узорка.

Код **параметарских тестова једног узорка** нулта хипотеза је облика $H_0(\theta = \theta_0)$, где је θ непознати параметар расподеле обележја, а θ_0 његова претпостављена вредност. Основни начин провере оваквих хипотеза је помоћу статистичког теста значајности, који се састоји у томе да одаберемо одговарајућу статистику $U = u(X_1, X_2, \dots, X_n)$ у којој фигурише непознати параметар θ и чија расподела нам је позната. Уз претпоставку да је $\theta = \theta_0$, добићемо конкретну вредност статистике $U = u(x_1, x_2, \dots, x_n)$. Ако је α^* , вероватноћа одступања статистике U од реализоване вредности, мања од унапред задатог прага значајности α , хипотезу одбацујемо, у противном је не одбацујемо. Алтернативни начин провере параметарских хипотеза једног узорка је помоћу интервала поверења. За задати праг значајности α направимо интервал поверења сам за тражени параметар, са нивоом поверења $\beta = 1 - \alpha$. Ако претпостављена вредност припада интервалу поверења, нулту хипотезу прихватамо са прагом значајности α . У противном је одбацујемо.

Код **параметарских тестова два узорка** циљ је да утврдимо да ли два узорка припадају истој популацији тј. да ли имају исте вредности параметара расподеле. Због тога је нулта хипотеза облика $H_0(\theta_1 = \theta_2)$.

Непараметарски статистички тестови су, са друге стране, базирани на моделима који не укључују предуслове везане за параметре популације из које узорак потиче, а претпоставке карактеристичне за већину непараметарских статистичких тестова најчешће су слабије од оних код параметарских тестова. Шта више, непараметарски тестови не захтевају тако прецизна „мерења” као параметарски тестови, а неки од њих се могу користити и у закључивањима везаним за квалитативна обележја.

Због чињенице да је често немогуће спровести мерења која омогућавају коректно коришћење параметарских тестова, непараметарским статистичким тестовима припада значајна улога у наукама у којима се користе методе статистике анализе.

1.2. Параметарски тестови

1.2.1. Тестирање хипотезе о математичком очекивању код нормалне расподеле када је дисперзија непозната (све три врсте алтернативне хипотезе)

Нека је $(X_1, X_2, \dots, X_n)$ прост случајан узорак из расподеле $N(m, \sigma^2)$. Тестирамо хипотезу $H_0(m = m_0)$ када параметар σ^2 није познат. Користимо статистику

$$\frac{\bar{X}_n - m_0}{\bar{S}_n} \sqrt{n-1} : t_{n-1}.$$

У случају опште алтернативне хипотезе $H_1(m \neq m_0)$ имамо да је критична област:

$$P\left\{\frac{|\bar{X}_n - m_0|}{\bar{S}_n} \sqrt{n-1} > c\right\} = \alpha$$
$$P\{|T| > c\} = \alpha$$

Константу c одређујемо из таблица одговарајуће Студентове расподеле. Ако је испуњен услов $\frac{|\bar{X}_n - m_0|}{\bar{S}_n} \sqrt{n-1} > c$ онда одбацујемо хипотезу H_0 и прихватамо H_1 , у супротном прихватамо H_0 .

Ако је алтернативна хипотеза облика $H_1(m < m_0)$, стављамо:

$$P\left\{\frac{\bar{X}_n - m_0}{\bar{S}_n} \sqrt{n-1} < c\right\} = \alpha$$

Опет, ако је услов у загради испуњен, хипотезу H_0 ћемо одбацити, односно прихватити ако неједнакост не важи.

Слично, за алтернативну хипотезу $H_1(m > m_0)$ имаћемо

$$P\left\{\frac{\bar{X}_n - m_0}{\bar{S}_n} \sqrt{n-1} > c\right\} = \alpha$$

Закључак о прихватању или одбацавању хипотезе се изводи на исти начин као у претходна два случаја.

пример:

Машина производи куглице пречника дебљине 0,5cm. Да бисмо проверили да ли куглице имају пречник прописане дебљине узима се узорак од 10 куглица. Ако је аритметичка

средина узорка $0,53\text{cm}$ и узорачко стандардно одступање $0,03\text{cm}$, тестирати хипотезу да машина производи куглице прописаног пречника са прагом значајности $0,05$.

решење:

$$H_0(m = 0.5\text{cm}), H_1(m > 0.5)$$

$$n = 10, \bar{X}_n = 0.53, \bar{S}_n = 0.03, T = \frac{\bar{X}_n - m_0}{\bar{S}_n} \sqrt{n-1}$$

$$T = \frac{0.53 - 0.5}{0.03} \sqrt{10-1} = 3$$

$$P\{T > c\} = 1 - P\{T < c\} = 0.05, P\{T < c\} = 0.95$$

$$P\left\{T < \frac{c - m_0}{\bar{S}_n} \sqrt{n-1}\right\} = P\left\{T < \frac{c - 0.5}{0.03} \sqrt{10-1}\right\} = P\left\{T < \frac{c - 0.5}{0.03} \sqrt{9}\right\} = 0.95$$

$$\frac{c-0.5}{0.03} \sqrt{9} = t_{9,0.95} \text{ па је}$$

$$\frac{c-0.5}{0.03} \sqrt{9} = 2.26, \frac{c-0.5}{0.03} \cdot 3 = 2.26, 3c - 1.5 = 2.26 \cdot 0.03, 3c = 0.0678 + 1.5 \quad 3c = 1.5678$$

$$c = 1.5678 : 3 = 0.5226 \quad T > c \quad \text{одбацујемо } H_0.$$

1.2.2. Тестирање хипотезе о математичком очекивању код нормалне расподеле када је дисперзија позната

Обележје X има $N(m, \sigma^2)$ расподелу са непознатим параметром μ и познатом стандардном девијацијом σ . Тестираћемо да је $H_0(\mu = \mu_0)$ против $H_1(\mu \neq \mu_0)$. Посматраћемо одступање аритметичке средине $\bar{X}_n$ узорка $(X_1, X_2, \dots, X_n)$ од очекиване вредности μ_0 . Са $\bar{x}_n$ ћемо означити аритметичку средину реализованог узорка $(x_1, x_2, \dots, x_n)$.

Користићемо чињеницу да $\bar{X}_n$ има $N(m, \frac{\sigma^2}{n})$ расподелу, одакле следи да $\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}$ има $N(0,1)$.

$$P\{|\bar{X}_n - \mu_0| \geq |\bar{x}_n - \mu_0|\} = P\left\{\left|\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n}\right| \geq \left|\frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}\right|\right\} = 1 - 2\Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}\right) = \alpha^*$$

Ако је $\alpha^* < \alpha$ хипотезу H_0 одбацујемо, а ако је $\alpha^* \geq \alpha$ хипотезу H_0 не одбацујемо.

пример:

Нека обележје X има нормалну расподелу $N(\mu, 1)$ и нека је средина узорка од 25 елемената $\bar{x}_{25} = 50$. Тестирати хипотезу $H_0(\mu = 49.5)$ против $H_1(\mu \neq 49.5)$ за праг значајности $\alpha = 0.05$.

решење:

$$\alpha^* = 1 - 2\Phi\left(\frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}\right) = 1 - 2\Phi\left(\frac{50 - 49.5}{1} \sqrt{25}\right) = 1 - \Phi(2.5) = 0.0124$$

Како је $\alpha^* < \alpha$ хипотезу H_0 одбацујемо.

1.2.3. Тестирање хипотезе о дисперзији са непознатим параметром σ

Обележје X има $N(m, \sigma^2)$ расподелу са непознатим параметром σ . Постављамо хипотезу $H_0(\sigma^2 = \sigma_0^2)$. Ако је $\bar{S}_n^2$ дисперзија узорка $(X_1, X_2, \dots, X_n)$, а $\bar{s}_n^2$ реализована вредност дисперзије узорка тада је

$$P\left\{\frac{n\bar{S}_n^2}{\sigma_0^2} \geq \frac{n\bar{s}_n^2}{\sigma_0^2}\right\} = \alpha^*$$

Знамо да $\frac{n\bar{S}_n^2}{\sigma_0^2}$ има χ_{n-1}^2 расподелу. Вероватноћу α^* упоређујемо са унапред задатим прагом значајности α и ако је $\alpha^* < \alpha$ хипотезу $H_0(\sigma^2 = \sigma_0^2)$ одбацујемо, у супротном је не одбацујемо.

пример:

Обележје има нормалну расподелу и дисперзију узорка за избрани узорак од 30 елемената. Тестираћемо хипотезу $H_0(\sigma^2 = 15)$ против $H_0(\sigma^2 \neq 15)$ за праг значајности $\alpha = 0.01$.

решење:

$$\frac{n\bar{s}_n^2}{\sigma_0^2} = \frac{30 \cdot 10}{15} = 20$$

Како је $P\left\{\frac{n\bar{S}_n^2}{\sigma_0^2} \geq 20\right\} = \alpha^* = 0.9$ прочитано из таблице

и $\alpha^* > \alpha$, хипотезу не одбацујемо.

1.2.4. Тестирање хипотезе о једнакости дисперзија код два независна обележја из нормалне расподеле

Нека су та два независна обележја $X: N(m_1, \sigma_1^2)$ и $Y: N(m_2, \sigma_2^2)$. Хипотезу $H_0(\sigma_1^2 = \sigma_2^2)$ можемо другачије записати као $H_0\left(\frac{\sigma_1^2}{\sigma_2^2} = 1\right)$.

Како узорачке дисперзије σ_1^2 и σ_2^2 оцењујемо редом са $\widetilde{S_{n_1}^2}$ и $\widetilde{S_{n_2}^2}$, где су $\widetilde{S_{n_1}^2}$ и $\widetilde{S_{n_2}^2}$ поправљене узорачке дисперзије,

$$\widetilde{S_{n_1}^2} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x}_{n_1})^2, \quad \widetilde{S_{n_2}^2} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y}_{n_2})^2$$

где је,

$$\bar{x}_{n_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \quad \bar{y}_{n_2} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i$$

уз претпоставку о једнакости дисперзија, за велике n_1 и n_2 важи:

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{\widetilde{S_{n_1}^2}}{\widetilde{S_{n_2}^2}} = \frac{\frac{\widetilde{S_{n_1}^2}}{\sigma_1^2} : \chi_{n_1-1}^2}{\frac{\widetilde{S_{n_2}^2}}{\sigma_2^2} : \chi_{n_2-1}^2} : F_{n_1-1, n_2-1}$$

Критична област ће бити облика $W = (0, c_1) \cup (c_2, \infty)$, а константе c_1 и c_2 налазимо из таблица одговарајуће Фишерове расподеле тако да важи $P\{F < c_1\} = P\{F > c_2\} = \frac{\alpha}{2}$, где је α праг значајности теста. Ако је нека од две неједнакости у заградама испуњена, хипотезу H_0 треба одбацити, у супротном ћемо је прихватити.

1.2.5. Тестирање хипотезе о једнакости средњих вредности код два независна обележја из нормалне расподеле када су дисперзије непознате и једнаке

Нека су та два независна обележја $X: N(m_1, \sigma_1^2)$ и $Y: N(m_2, \sigma_2^2)$. Претпостављајући да важи $H_0(m_1 = m_2)$ имамо:

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (m_1 - m_2)}{S_Z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{S_Z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : t_{n_1+n_2-2}$$

где је

$$S_Z^2 = \frac{(n_1 - 1)\overline{S_{n_1}^2} + (n_2 - 1)\overline{S_{n_2}^2}}{n_1 + n_2 - 2}$$

За општу алтернативну хипотезу $H_1(m_1 \neq m_2)$, из:

$$P \left\{ \frac{|\overline{X}_{n_1} - \overline{Y}_{n_2}|}{S_Z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > c_\alpha \right\} = \alpha$$

одређујемо c_α и хипотезу H_0 одбацујемо ако је неједнакост испуњена. За алтернативне хипотезе $H_1(m_1 > m_2)$ и $H_1(m_1 < m_2)$, коефицијенте одређујемо из услова

$$P \left\{ \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{S_Z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > c_\alpha \right\} = \alpha$$

односно

$$P \left\{ \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{S_Z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < c_\alpha \right\} = \alpha$$

пример:

У узорку од 14 теглица креме једног произвођача просечан садржај коензима Q10 по теглици је 23 мерне јединице, а стандардна девијација је 3 јединице. У другом узорку од 16 теглица другог произвођача просечан садржај коензима Q10 по теглици је 25, а стандардна девијација је 4 мерне јединице. Да ли се на нивоу значајности од 1% може закључити да креме ова два произвођача садрже у просеку **различит** садржај Q10? Претпоставимо да је количина Q10 по теглици креме нормално расподељена за сваког од два произвођача.

решење: m_1 је просек Q10 у првом узорку $n_1 = 14, \overline{S_{n_1}} = 3, \overline{X_{n_1}} = 23$ m_2 је просек Q10 у другом узорку $n_2 = 16, \overline{S_{n_2}} = 4, \overline{Y_{n_2}} = 25$

$$H_0(m_1 = m_2), H_1(m_1 \neq m_2)$$

Ниво значајности је $\alpha=0,01$. Знак $\neq$ у алтернативној хипотези указује да је тест двостран, са две области одбацивања нулте хипотезе. Површина на сваком крају је $\frac{\alpha}{2} = \frac{0,01}{2} = 0,005$, са 28 степени слободе, $t_{28;0,005}$, па је $t_{28;0,005} = 2.763$.

$$P\left\{T > c_{\frac{\alpha}{2}}\right\} = \frac{\alpha}{2}$$

$$S_z^2 = \frac{(n_1 - 1)\overline{S_{n_1}}^2 + (n_2 - 1)\overline{S_{n_2}}^2}{n_1 + n_2 - 2}$$

$$S_z^2 = \frac{(14 - 1) \cdot 9 + (16 - 1) \cdot 16}{14 + 16 - 2} = \frac{13 \cdot 9 + 15 \cdot 16}{28} = \frac{117 + 240}{28} = \frac{357}{28} = 12.75,$$

$$S_z = 3.57, T = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{S_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{23 - 25}{3.57 \cdot \sqrt{\frac{1}{14} + \frac{1}{16}}} = \frac{-2}{3.57 \cdot \sqrt{\frac{30}{224}}} = \frac{-2}{1.31} = -1.53, T = -1.53$$

Пошто се реализован тест-статистика $T = -1.53$ налази у области неодбацивања доносимо одлуку о неодбацивању нулте хипотезе.

1.2.6. Тестирање хипотезе о вероватноћи код биномне расподеле

Користећи исте ознаке као у претходном делу, тестирамо хипотезу $H_0(p = p_0)$. Користимо статистику $Z = \frac{\overline{X_n} - np_0}{\sqrt{p_0(1-p_0)n}} : N(0,1)$. У зависности од задатог α прага значајности теста, можемо помоћу таблица нормалне расподеле наћи константу c_α , а критична област је облика $P\{|Z| > c\} = \alpha$ за општу алтернативну хипотезу, односно $P\{Z > c\} = \alpha$ или $P\{Z < c\} = \alpha$ за неку од друге две. Као и у претходним тестирањима, H_0 одбацујемо у случају да је одговарајућа неједнакост испуњена.

пример:

У узорку од 3000 бацања новчића добијено је 1578 грбова. Вероватноћа добијања грба је 0,5 и тај податак узимамо као нулту хипотезу, а податак да ће се добити више грбова узима се као алтернативна хипотеза. Тестирати нулту хипотезу са прагом значајности од 0,01.

решење:

Нека је

$$H_0 \left(X: B \left(\frac{1}{2}, 3000 \right) \right)$$

Случајна променљива X представља број добијених грбова, са биномном расподелом, која се апроксимира нормалном расподелом.

$$p = \frac{1}{2}, n = 3000, np_0 = 1500, np_0(1 - p_0) = 750, \alpha = 0.01$$

$$P\{Z > c\} = 1 - P\{Z < c\} = \alpha \text{ па је } P\{Z < c\} = 0.99$$

$$P\left(\frac{\bar{X}_n - np_0}{\sqrt{p_0(1-p_0)n}} < c\right) = P\left(\frac{\bar{X}_n - 1500}{\sqrt{750}} < \frac{c - 1500}{\sqrt{750}}\right) =$$

$$P\left(\bar{X}_n^* < \frac{c - 1500}{\sqrt{750}}\right) = \Phi\left(\frac{c - 1500}{\sqrt{750}}\right) = 0.99$$

$$\frac{c - 1500}{\sqrt{750}} = 2.32, c - 1500 = 2.32 \cdot \sqrt{750}, c = 2.32 \cdot 27.39 + 1500 = 63.5448 + 1500$$

$$c \approx 1564$$

пошто је $c < 1578$, израчуната вредност припада критичној области па одбацујемо нулту хипотезу.

1.2.7. Хипотеза о коефицијенту корелације

Нека је (X, Y) случајни вектор и $(x_i, y_i), i = 1, 2, \dots, n$ дати подаци за дводимензионално обележје (X, Y) .

Узорачке средине и узорачке дисперзије се рачунају по формулама:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i, \overline{S_X^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2, \bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i, \overline{S_Y^2} = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_n)^2,$$

$$\sigma_x = \sqrt{\overline{S_X^2}}, \sigma_y = \sqrt{\overline{S_Y^2}}$$

Овде су парови (X_i, Y_i) независни, док случајне величине из истог пара (X_i, Y_i) имају одређену заједничку расподелу и могу бити зависне, са коефицијентом корелације ρ . Како је

$$\rho(X, Y) = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sigma_x \sigma_y}$$

методом момената добијамо оцену за ρ :

$$\hat{\rho} = \frac{\sum_{k=1}^n (X_k - \bar{\mu}_X) \cdot (Y_k - \bar{\mu}_Y)}{\sum_{k=1}^n (X_k - \bar{\mu}_X)^2 \sum_{k=1}^n (Y_k - \bar{\mu}_Y)^2}, \hat{\rho} - \text{узорачки коефицијент корелације}$$

За тестирање хипотеза у вези са ρ од користи је следећа теорема:

Теорема: Ако случајни вектор (X, Y) има дводимензионалну нормалну расподелу са $\rho = 0$, тада статистика

$$T = \frac{\hat{\rho} \sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}$$

има t_{n-2} расподелу.

Предходна теорема се користи за тестирање хипотезе $H_0(\rho = 0)$ у случају када вектор (X, Y) , има нормалну расподелу или када је обим узорка велики, па се може прихватити нормална апроксимација.

пример:

Из узорка обима $n = 27$ из дводимензионе нормалне расподеле добијено је $\hat{\rho} = 0.6$. Са нивоом значајности $\alpha = 0.05$ тестирати хипотезу $H_0(\rho = 0)$ против алтернативне хипотезе $H_1(\rho > 0)$.

решење:

$$P\{T > c\} = \alpha, 1 - P\{T < c\} = 0.05, P\{T < c\} = 0.95, c = t_{25;0.95} = 1.708$$

Применом предходне теореме добијамо тест-статистику

$$T = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} = \frac{0.6 \cdot \sqrt{27-2}}{\sqrt{1-0.6^2}} = \frac{0.6 \cdot 5}{\sqrt{1-0.36}} = \frac{3}{\sqrt{0.64}} = \frac{3}{0.8} = 3.75$$

пошто је $T > c$ нулту хипотезу одбацујемо.

1.3. Непараметарски тестови

1.3.1. Пирсонов χ^2 -тест

Пирсонов χ^2 -тест се може применити за тестирање свих расподела, ако је обим узорка бар 50. Нека је са прагом значајности α потребно тестирати хипотезу H_0 да обележје X за које има прост случајан узорак $(X_1, X_2, \dots, X_n)$ има дату функцију расподеле $F_0(x)$, $H_0(F(x) = F_0(x))$. Нека је у расподели обележја X непознато s параметара. Скуп могућих вредности обележја се разбија на r дисјунктних скупова $S_1, S_2, \dots, S_r$, тако да је број m_j елемената из узорка који су у скупу S_j најмање 5. Бројеви m_j су реализоване вредности случајних величина M_j , чије су расподеле $B(n, p_j)$, $j = 1, 2, \dots, r$. Према постављеној хипотези се налазе вероватноће: $p_j = P_{H_0}\{X \in S_j\}$.

Тест-статистика је:

$$\chi_{\bar{U}}^2 = \sum_{j=1}^r \frac{(M_j - n \cdot p_j)^2}{n \cdot p_j} = \sum_{j=1}^r \frac{M_j^2}{n \cdot p_j} - n.$$

Ако је хипотеза H_0 тачна, тада та статистика има $\chi_{r-s-1; \alpha}^2$ расподелу. На основу датог узорка израчунавамо реализовану вредност тест-статистике.

За дати ниво значајности (α) читамо вредност $\chi_{r-s-1; \alpha}^2$ из услова

$$P\{\chi_{r-s-1}^2 \geq \chi_{r-s-1; \alpha}^2\} = \alpha$$

у таблицама. Ако је у узораку регистрована вредност тест-статистике већа од табличне, хипотеза се одбацује; у супротном се, на основу датих података и за дати праг значајности; хипотеза прихвата.

пример:

Коцка је бачена 1200 пута и добијени су следећи резултати

1	2	3	4	5	6
183	211	170	220	200	216

На прагу значајности $\alpha = 0.05$ проверити да ли је коцкица хомогена.

решење:

Вероватноћа за сваки број уколико је коцкица хомогена је $\frac{1}{6}$.

$$p = P(1') = \frac{1}{6}, P(2') = \frac{1}{6}, P(3') = \frac{1}{6}, P(4') = \frac{1}{6}, P(5') = \frac{1}{6}, P(6') = \frac{1}{6}, np = 200$$

Реализована вредност тест-статистика је

$$\begin{aligned} \chi^2 &= \frac{(183 - 200)^2}{200} + \frac{(211 - 200)^2}{200} + \frac{(170 - 200)^2}{200} + \frac{(220 - 200)^2}{200} + \frac{(200 - 200)^2}{200} \\ &\quad + \frac{(216 - 200)^2}{200} \\ &= \frac{289}{200} + \frac{121}{200} + \frac{900}{200} + \frac{400}{200} + \frac{0}{200} + \frac{256}{200} = 9.83 \end{aligned}$$

Пошто нема оцењених параметара, при тачној H_0 , тест-статистика има χ^2_5 расподелу. Таблична вредност за праг значајности 0.05 је 11.070. Пошто је реализована вредност тест-статистике мања можемо прихватити хипотезу да је коцкица хомогена.

1.3.2. χ^2 -тест независности два обележја

Нека је $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ узорак из дводимензионалне расподеле. Тестирамо хипотезу H_0 да су обележја X и Y независна.

Ако су $F: \mathbb{R}^2 \rightarrow [0,1]$, $F_X: \mathbb{R} \rightarrow [0,1]$, $F_Y: \mathbb{R} \rightarrow [0,1]$ редом функције расподеле случајног вектора (X, Y) и случајних величина X и Y , тада хипотезу H_0 можемо записати на следећи начин:

$$H_0: F(x, y) = F_X(x)F_Y(y) \text{ за све } x, y \in \mathbb{R}$$

Обележја X и Y не морају бити нумеричка да бисмо могли применити хи-квадрат тест.

Нека су $x_1, x_2, \dots, x_r$ све вредности обележја X које се појављују у узорку, и слично $y_1, y_2, \dots, y_s$ вредности обележја Y . Податке приказујемо помоћу тзв. табеле контингенције (повезаности):

$X \backslash Y$	y_1	y_2	...	y_s	Σ
x_1	n_{11}	n_{12}	...	n_{1s}	$n(x_1)$
x_2	n_{21}	n_{22}	...	n_{2s}	$n(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n(x_r)$
Σ	$n(y_1)$	$n(y_2)$	...	$n(y_s)$	n

где је n_{ij} број појављивања пара (x_i, y_j) у узорку, а $n(x_i)$ и $n(y_j)$ маргинални зборови по врстама односно колонама. Како претпостављамо да су X и Y независне, важиће

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(n \cdot n_{ij} - n(x_i) \cdot n(y_j))^2}{n \cdot n(x_i) \cdot n(y_j)} = \frac{\left(n_{ij} - \frac{n(x_i) \cdot n(y_j)}{n}\right)^2}{\frac{n(x_i) \cdot n(y_j)}{n}} : \chi_{(r-1)(s-1)}^2$$

Константу c_α можемо одредити из таблица χ^2 расподеле за одговарајући степен слободe $(r-1)(s-1)$ за познати α праг значајности теста:

$$P\{\chi_{(r-1)(s-1)}^2 > c_\alpha\} = \alpha$$

Као и у претходним тестирањима хипотеза, и овде ћемо H_0 прихватити у случају да неједнакост у загради није испуњена.

1.3.3. Тест рангова (Wilcoxon–Mann–Whitneytest)

Тест се примењује за тестирање хипотеза о једнакости непрекидних расподела за обележја X и Y на основу два проста случајна узорка $(X_1, X_2, \dots, X_m)$ и $(Y_1, Y_2, \dots, Y_n)$ при чему је $m \leq n$. Нулта хипотеза је

$$H_0(F_X(x) = F_Y(x)),$$

а алтернативна је

$$H_1(F_X(x) \geq F_Y(x)).$$

Прво се формира обједињени узорак у коме су сви елементи у неоппадајућем поретку и дефинишемо:

$$h_{ij} = \begin{cases} 1, & \text{ако је } Y_j < X_i \\ 0, & \text{у супротном случају.} \end{cases}$$

Тест-статистика ће бити

$$\sum_{i=1}^m \sum_{j=1}^n h_{ij}$$

и она представља укупан број случајева у којима елемент из узорка за обележје Y предходи елементу из узорка за обележје X .

При тачној хипотези H_0 је

$$E(h_{ij}) = \frac{1}{2}, \quad E(h_{ij}^2) = \frac{1}{2}$$

па је тада

$$E(U) = \frac{1}{2}mn, \quad D(U) = \frac{mn(m+n+1)}{12}$$

ако је $m, n < 8$, постоје таблице из којих се налази граница критичне области, ако је $m, n \geq 8$, тада се расподела статистике U апроксимира нормалном расподелом

$$N(E(U), D(U)), \quad m, n \rightarrow +\infty, \frac{m}{n} \rightarrow s > 0.$$

Критична област се одређује из услова

$$P\{U \geq u_\alpha\} = \alpha.$$

пример:

У обједињеном узорку сви елементи из узорка X обележени су са x , а сви елементи из узорка Y обележене су са y . Имамо следећи обједињени узорак

$x \ y \ x \ x \ x \ y \ y$

Испитати хипотезу о једнакости расподела, та два обележја. Праг значајности је 0.05.

решење:

Укупан број случајева у којима елемент из узорка за обележје Y претходи елементу из узорка за обележје X је 5. Реализована вредност тест статистике је $u = 5$.

Из таблице за $n = m = 4$ налазимо $\alpha_0 = 0.243$, како је та вредност већа од 0.05 онда хипотезу о једнакости расподела прихватамо.

1.3.4. Тест серија

Нека је обележје X непрекидног типа и нека је $(X_1, X_2, \dots, X_n)$ прост случајан узорак обима n . Формирамо варијациони низ $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ и одређујемо узорачку медијану

$$m_e = \begin{cases} X_{(\frac{n+1}{2})}, & \text{ако је } n \text{ непарно} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}), & \text{ако је } n \text{ парно} \end{cases}$$

На основу $(X_1, X_2, \dots, X_n)$ формирамо нови низ $(Y_1, Y_2, \dots, Y_k)$ чији су елементи једнаки 0 или 1, а добијају се по следећем правилу:

$$Y_i = \begin{cases} 0, & \text{ако је } X_i < m_e \\ 1, & \text{ако је } X_i > m_e \\ \text{место остаје празно,} & \text{ако је } X_i = m_e \end{cases}$$

Тако добијамо низ од k нула и јединица и у том низу пребројавамо серије истих цифара.

пример1:

Низ 110010011100.

низ: 110010011100

110010011100 (11,00,1,00,111,00)

укупна серија је:6.

Ако је тачна хипотеза да је узорак случајан може се доказати да ће K серија имати приближно нормалну расподелу $N\left(\frac{n+2}{2}, \frac{n(n-2)}{4(n-1)}\right)$.

Границу k_α критичне области одређујемо из услова

$$P\{K \geq k_\alpha\} = \alpha$$

користећи таблице за нормалну расподелу.

пример2:

Тестирати хипотезу да је узорак

51;84;44;24;64;100;59;76;32;91;98;33;80;27;60;74;

случајан, против алтернативне да постоји периодичност у појављивању малих и великих вредности обележја. Праг значајности је $\alpha = 0.05$.

решење:

Формирамо варијациони низ:

24;27;32;33;44;51;59;**60;64**;74;76;80;84;91;98;100

Медијана је

$$m_e = \frac{60 + 64}{2} = 62$$

Формирамо низ серија:

0100110101101011(0,1,00,11,0,1,0,11,0,1,0,11)

Укупан број серија у овом низу је 12.

Асимптотска расподела за број K серија, при тачној нултој хипотези, је нормална расподела са параметрима, $n = 16$

$$K \sim N\left(\frac{16+2}{2}, \frac{16 \cdot (16-2)}{4 \cdot (16-1)}\right), K \sim (9, 3.73)$$

$$P\left\{\frac{K-9}{1.93} \geq \frac{k_\alpha-9}{1.93}\right\} = 0.05,$$

$$P\left\{K^* \geq \frac{k_\alpha-9}{1.93}\right\} = 0.05$$

$$\frac{k_\alpha-9}{1.93} = 3.1, k_\alpha = 14.98$$

Како је реализована тест-статистика ван критичне области, то се хипотеза о случајности прихвата.

Тест Колмогорова

2.1. Емпиријска функција расподеле

У првом делу смо представили статистичке тестове и дефинисали смо: популацију, обележје, прост случајни узорак.

Да би узорак добро репрезентовао генерални скуп, мора да буду испуњени следећи услови:

- 1) сваки елемент генералног скупа мора да има једнаку шансу да уђе у узорак;
- 2) узорак мора да буде довољно бројан.

деф: За дати простор узорка $(X_1, X_2, \dots, X_n)$, **емпиријска функција расподеле** се дефинише за свако $x \in R$, а са $F_n(x) = \frac{k}{n}$, где је k број елемената из узорка који нису већи од x .

Нека је $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ варијациони низ, који чине вредности случајних променљивих $X_1, X_2, \dots, X_n$ уређене по величини од најмање до највеће. Тада се емпиријска функција расподеле може одредити помоћу

$$F_n(x) = \begin{cases} 0, & \text{ако је } x < X_{(1)} \\ \frac{k}{n}, & \text{ако је } X_{(k)} \leq x < X_{(k+1)}, \quad 1 \leq k \leq n-1 \\ 1, & \text{ако је } x \geq X_{(n)} \end{cases}$$

последица: Математичко очекивање и дисперзија (варијанса) за $F_n(x)$ су:

- 1) $E(F_n(x)) = F(x)$
- 2) $D(F_n(x)) = \frac{F(x)[1-F(x)]}{n}$

-особине емпиријске функције расподеле

- 1) F_n има вредности између 0 и 1,
- 2) F_n је корак функција са скоковима на различитим вредностима $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$,
- 3) F_n је непрекидна са десне стране,
- 4) За свако фиксирано $x, -\infty < x < +\infty$, важи

$$F_n(x) \sim N\left(F(x), \frac{F(x)[1-F(x)]}{n}\right),$$

- 5) $nF_n(x)$ има биномну расподелу

$$nF_n(x) \sim B(n, F(x)).$$

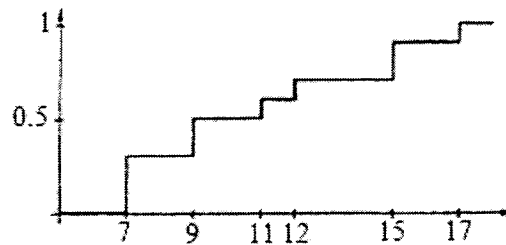
пример:

У експерименту су добијене следеће бројевне вредности, узорак је обима $n = 10$: (9; 15; 7; 11; 17; 9; 7; 12; 7; 15).

решење:

Варијациони низ је: 7; 7; 7; 9; 9; 11; 12; 15; 15; 17.

$$F_n(x) = \begin{cases} 0, & \text{ako je } x < 7 \\ \frac{3}{10}, & \text{ako je } 7 \leq x < 9 \\ \frac{5}{10}, & \text{ako je } 9 \leq x < 11 \\ \frac{6}{10}, & \text{ako je } 11 \leq x < 12 \\ \frac{7}{10}, & \text{ako je } 12 \leq x < 15 \\ \frac{9}{10}, & \text{ako je } 15 \leq x < 17 \\ 1, & \text{ako je } x \geq 17 \end{cases}$$



2.2. Централна теорема математичке статистике

Основни задатак математичке статистике јесте да помоћу узорка $(X_1, X_2, \dots, X_n)$ одреди расподелу $F(x)$ обележја X , под условом да је n врло велико. Како у применама радимо само са коначним обимом узорка, расподелу за X можемо да одредимо само приближно, утолико тачније уколико је n веће. У решавању постављеног проблема радимо са функцијама случајног узорка $(X_1, X_2, \dots, X_n)$.

Лема: Нека X има континуирану униформну расподелу, F има емпиријску функцију расподеле и $U = F(X)$. Тада $U \sim U[0,1]$.

доказ:

Нека $u \in [0,1]$. X непрекидна величина $\exists x_u \in R, F(x_u) = u$.

$$F(u) = P(U < u) = P(F(X) < F(x_u)) = P(X < x_u) = F(x_u) = u$$

Дакле $F(u) = u$ и U има униформну расподелу на $[0,1]$.

Да бисмо извршили прављење D_n треба урадити следеће:

генерисање случајних узорака величине n из стандардне униформне расподеле $U[0,1]$, са функцијом расподеле $F(u) = u$.

Пронаћи максималну апсолутну разлику између $F(u)$ и емпиријске функције расподеле $F_n(u)$ за дати узорак. ■

Теорема (Гливенко-Кантели):

$$\text{За } n \rightarrow \infty, D_n \xrightarrow{P} 0, \quad D_n = \sup_x |F_n(x) - F(x)|$$

доказ:

Пошто је конвергенција скоро сигурна тада је

$$P\left(\lim_{n \rightarrow +\infty} \sup_{x \in R} |F_n(x) - F(x)| = 0\right) = 1$$

За јаки закон великих бројева нам говори да за произвољно $x \in R$

$$P\left(\lim_{n \rightarrow +\infty} F_n(x) = F(x)\right) = 1$$

Ми ћемо доказати теорему за случај када је $F(x)$ је непрекидна. Доказ се може појачати за опште расподеле функција. Нека је $\varepsilon > 0$ фиксирани број. Пошто је F непрекидна можемо наћи m тачке које су $-\infty = x_0 < x_1 < \dots < x_m < +\infty$ и $F(x_j) - F(x_{j-1}) \leq \varepsilon$ за $j \in \{1, 2, \dots, m\}$, такве да је $x_{j-1} < x < x_j$. Пошто функција расподеле не опада закључујемо

$$\begin{aligned} F_n(x) - F(x) &\leq F_n(x_j) - F(x_{j-1}) \\ &= (F_n(x_j) - F(x_j)) + (F(x_j) - F(x_{j-1})) \\ &\leq (F_n(x_j) - F(x_j)) + \varepsilon \\ &\leq \max_{j \in \{0, 1, \dots, m\}} |F_n(x_j) - F(x_j)| \end{aligned}$$

На исти начин

$$\begin{aligned} F_n(x) - F(x) &\geq F_n(x_{j-1}) - F(x_j) \\ &= (F_n(x_{j-1}) - F(x_{j-1})) + (F(x_{j-1}) - F(x_j)) \\ &\geq (F_n(x_{j-1}) - F(x_{j-1})) - \varepsilon \\ &\geq -\max_{j \in \{0, 1, \dots, m\}} |F_n(x_{j-1}) - F(x_{j-1})| - \varepsilon \end{aligned}$$

следи

$$\sup_{x \in R} |F_n(x) - F(x)| \leq \varepsilon + \max_{j \in \{0, 1, \dots, m\}} |F_n(x_j) - F(x_j)|.$$

Сада ћемо искористити јаки закон великих бројева. Дефинисаћемо догађаје

$$A_j = \left\{ \lim_{n \rightarrow \infty} |F_n(x_j) - F(x_j)| \neq 0 \right\}$$

Ми знамо да је $P(A_j) = 0$. Дефинисаћемо такође

$$A = \left\{ \lim_{n \rightarrow \infty} \max_{j \in \{0,1,\dots,m\}} |F_n(x_j) - F(x_j)| \neq 0 \right\}$$

Јасно је $A = \bigcup_{j=0}^m A_j$ и $P(A) = P(\bigcup_{j=0}^m A_j) \leq \sum_{j=0}^m P(A_j) = 0$. Дакле, следи да је

$$\lim_{n \rightarrow \infty} \sup_{x \in R} |F_n(x) - F(x)| \leq \varepsilon$$

или другим речима

$$B_\varepsilon = \left\{ \lim_{n \rightarrow \infty} \sup_{x \in R} |F_n(x) - F(x)| \leq \varepsilon \right\}$$

Има бар један догађај за $\varepsilon > 0$, чиме смо доказали теорему. ■

Теорема (Централне математичке статистике): Ако је $F(x)$ функција расподеле случајне променљиве X и F_n емпиријска функција расподеле добијена из простог узорка $(X_1, X_2, \dots, X_n)$ обима n , тада је

$$P(\sup_{x \in R} |F_n(x) - F(x)| \rightarrow 0, \text{ када } n \rightarrow +\infty) = 1$$

Неједнакост Дворецки-Кифер-Волфовиц

Нека је дат природни број n , нека су $X_1, X_2, \dots, X_n$ независне реалне вредности и једнако расподељене случајне променљиве са функцијом расподеле $F(x)$. Нека је $F_n(x)$ емпиријска функција расподеле дефинисана:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}, \quad I\{X_i \leq x\} = \begin{cases} 1, & \text{ако је } X_i \leq x \\ 0, & \text{ако је } X_i > x \end{cases}$$

Дворецки-Кифер-Волфовиц неједнакост ограничава вероватноћу да случајна функција F_n разликује од F за дату константу $\varepsilon > 0$ било где на реалној правој. Тачније постоји процена

$$P\left(\sup_{x \in R} (F_n(x) - F(x)) > \varepsilon\right) \leq e^{-2n\varepsilon^2}, \quad \forall \varepsilon \geq \sqrt{\frac{1}{2n} \ln 2}$$

За двострану процену важи

$$P\left(\sup_{x \in R} |F_n(x) - F(x)| > \varepsilon\right) \leq e^{-2n\varepsilon^2}, \quad \forall \varepsilon \geq 0.$$

Неједнакост јача Гливенко-Кантелијеву теорему конвергенције како n тежи бесконачности, такође процењује реп вероватноћу Колмогоров-Смирнове статистике. F има униформну расподелу на $[0,1]$, где F_n има исту расподелу као G_n , G_n има емпиријску функцију расподеле, $Y_1, Y_2, \dots, Y_n$ су независне и $U[0,1]$, где важи

$$\sup_{x \in R} |F_n(x) - F(x)| = \sup_{x \in R} |G_n(F(x)) - F(x)| \leq \sup_{0 \leq t \leq 1} |G_n(t) - t|$$

Једнако је само ако и само ако је F непрекидно.

2.3. Тест Колмогорова

Колмогоров тест се користи за проверу да ли узорак $X_1, X_2, \dots, X_n$ има одређену функцију расподеле. Нулта хипотеза је

$$H_0(F(x) = F_0(x)), \forall x$$

Двострана алтернативна хипотеза је

$$H_1(F(x) \neq F_0(x)), \text{ за најмање једно } x.$$

Колмогорова тест-статистика је

$$D_n = \sup_x |F_n(x) - F_0(x)|,$$

где је $F_n(x)$ емпиријска функција расподеле дефинисана

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

где је I индикатор-функција, код које важи

$$I\{X_i \leq x\} = \begin{cases} 1, & \text{ако је } X_i \leq x \\ 0, & \text{ако је } X_i > x \end{cases}$$

која броји пропорцију тачака узорака мањих од x . У било којој фиксираној тачки $x \in R$, по закону великих бројева указује да је

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\} \rightarrow E(I\{X_i \leq x\}) = P\{X_1 \leq x\} = F(x)$$

пропорција узорка се на интервалу $(-\infty, x]$ приближава вероватноћи целог скупа.

$$P\{D_n \geq d_{n,\alpha}\} = \alpha$$

За једнострану алтернативну хипотезу $H_1(F(x) \leq F_0(x))$ тест статистика је

$$D_n^+ = \sup_x \{F_0(x) - F_n(x)\}$$

За једнострану алтернативну хипотезу $H_1(F(x) \geq F_0(x))$ тест статистика је

$$D_n^- = \sup_x \{F_n(x) - F_0(x)\}$$

Као што смо већ дефинисали Колмогорову тест-статистику

$$D_n = \sup_x |F_n(x) - F_0(x)|,$$

без улажења у детаље како дефинисати простор функција. По теореме Гливенко-Кантели $D_n \rightarrow 0$ (скоро сигурно) када $n \rightarrow \infty$, па је

$$\sqrt{n}D_n \rightarrow \sup_x |B(F_0(x))|$$

када $n \rightarrow \infty$, где је

$B(\cdot)$ -Браунов мост

Ако је $F_0(x)$ непрекидна функција, онда граница расподеле $\sqrt{n}D_n$ не зависи од $F_0(x)$, поклапа се са расподелом $\sup_{x \in [0,1]} |B(x)|$ и једнака је

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq \lambda) = K(\lambda) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 \lambda^2} = \frac{\sqrt{2\pi}}{\lambda} \sum_{i=1}^{\infty} e^{-\frac{(2i-1)^2 \pi^2}{8\lambda^2}},$$

за свако $\lambda > 0$ $K(\lambda)$ је Колмогорова расподела. Слобода из расподеле статистике важно је својство које омогућава да се добије критична вредност за било које n , без обзира на $F_0(x)$.

Средња вредност (μ) и варијанса (σ^2) за $\sqrt{n}D_n$ су приближно

$$\mu = \sqrt{\frac{\pi}{2}} \ln 2 \approx 0.87, \quad \sigma^2 = \frac{\pi^2}{12 - \mu^2} \approx 0.068, \quad \sigma \approx 0.26$$

За тест-статистику D_n^+ , када $n \rightarrow \infty$, важи

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n^+ \leq x) = 1 - e^{-x^2}$$

Дефинишимо, још једанпут, тест-статистику Колмогорова

$$D_n = \sup_{x \in R} |F_n(x) - F(x)|$$

Посебно, можемо поправити, за $\varepsilon > 0$ и ако је n велико онда са великом вероватноћом $D_n \leq \varepsilon$. Ако је тако, онда бисмо могли

$$C_n(x) = \{(x, y) : |F_n(x) - y| \leq \varepsilon\}.$$

То нам даје добру идеју за F јер је $D_n \leq \varepsilon$, ако и само ако график функције F лежи у C_n . График функције F је скуп свих парова $(x, F(x))$.

Теорема 1: Ако је $F(x)$ непрекидна

$$\sup_x |F_n(x) - F(x)|$$

не зависи од F .

доказ:

Дефинишимо инверзну функцију функције F

$$F^{-1}(y) = \min\{x : F(x) \geq y\},$$

затим чинећи промену променљивих $y = F(x)$ или $x = F^{-1}(y)$, следи

$$P\left(\sup_x |F_n(x) - F(x)| \leq t\right) = P\left(\sup_x |F_n(F^{-1}(y)) - y| \leq t\right)$$

Користећи дефиницију емпиријске функције расподеле F_n

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

слиди

$$P\left(\sup_{0 \leq y \leq 1} |F_n(F^{-1}(y)) - y| \leq t\right) = P\left(\sup_{0 \leq y \leq 1} \left|\frac{1}{n} \sum_{i=1}^n I(F(X_i) \leq y) - y\right| \leq t\right).$$

Функција $F(X_i)$ је униформна на интервалу $[0,1]$ јер $F(X_1)$ има емпиријску функцију расподеле

$$P(F(X_1) \leq t) = P(X_1 \leq F^{-1}(t)) = F(F^{-1}(t)) = t$$

дакле, случајне променљиве

$$U_i = F(X_i) \text{ за } i \leq n$$

су независне и имају униформну расподелу на $[0,1]$, тако да смо доказали

$$P\left(\sup_x |F_n(x) - F(x)| \leq t\right) = P\left(\sup_{0 \leq y \leq 1} \left|\frac{1}{n} \sum_{i=1}^n I(U_i \leq y) - y\right| \leq t\right)$$

да не зависи од F . ■

За фиксне тачке x подразумева се да

$$\sqrt{n}(F_n(x) - F(x)) \rightarrow N(0, F(x)(1 - F(x)))$$

зато што је $F(x)(1 - F(x))$ варијанса за $I\{X_1 \leq x\}$. Ако се узме у обзир да

$$\sup_{x \in R} |F_n(x) - F(x)|$$

конвергира у расподели.

Теорема2: Имамо

$$P\left(\sup_{x \in R} |F_n(x) - F(x)| \leq t\right) \rightarrow K(t) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 t}$$

где је $K(t)$ расподела Колмогоров-Смирнов.

Нулта хипотеза је

$$H_0(F(x) = F_0(x))$$

Алтернативна хипотеза је

$$H_1(F(x) \neq F_0(x))$$

Размотримо следећу статистику

$$D_n = \sqrt{n} \sup_{x \in R} |F_n(x) - F(x)|$$

Ако је нулта хипотеза тачна онда, по *теорему 1*, вредности за D_n читамо табеларно (зависиће само од n). Ако је n довољно велико тако да расподела D_n се апроксимира Колмогоров-Смирновом расподелом из *теореме 2*.

Ако претпоставимо да нулта хипотеза H_0 није испуњена, $F(x) \neq F_0(x)$. Пошто F има, по закону великих бројева емпиријска функција расподеле F_n конвергира ка F , односно за велико n имаћемо

$$\sup_x |F_n(x) - F_0(x)| > \delta$$

за довољно мало δ . Множењем са $\sqrt{n}$ добијамо

$$D_n = \sqrt{n} \sup_x |F_n(x) - F_0(x)| > \sqrt{n} \delta$$

Ако H_0 није испуњена

$$D_n > \sqrt{n} \delta \rightarrow +\infty \text{ за } n \rightarrow +\infty$$

Дакле, за тестирање H_0 размотрићемо случајеве

$$\delta = \begin{cases} H_0: D_n \leq c \\ H_1: D_n > c \end{cases}$$

Праг c зависи од нивоа значајности α и можемо наћи из услова

$$P_{H_0}(\delta \neq H_0) = \alpha$$

$$P_{H_0}(D_n \geq c) = \alpha$$

Пошто под H_0 расподела за D_n се може прочитати табеларно за свако n , можемо наћи границу $c = c_\alpha$ из таблице. У већини књига статистичке табеле за ову расподелу су за $n < 100$. Када је $n > 100$ можемо користити Колмогоров-Смирнов расподелу да пронађемо c .

$$P_{H_0}(D_n \geq c) = \alpha \approx 1 - K(c)$$

и можемо користити табелу за расподелу Колмогоров-Смирнов да бисмо нашли c .

Асимптотска дистрибуција је

$$P(D_n < \lambda) = K \cdot \left(\left[\sqrt{n} + 0.12 + \frac{0.11}{\sqrt{n}} \right] \cdot \lambda \right), n \geq 4$$

када је

$$K = 2 \sum_{i=1}^{\infty} (-1)^{i-1} \cdot e^{-2i^2 \lambda^2}$$

A	0.20	0.10	0.05	0.02	0.01
c	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Таблица када је n веће од 100.

пример 1:

На основу датог узорка мерења чије су вредности дате у табели, применом теста Колмогорова, са прагом значајности $\alpha = 0.05$ тестирати хипотезу да узорак припада нормалној расподели

$$H_0(F(x) = F_0(x)), H_1(F(x) \neq F_0(x))$$

$$P\{D_n \geq c\} = \alpha$$

N	x_i	$F_0(x)$	$F_n(x)$	$ F_n(x) - F_0(x) $
1	-1.787	0.0370	0.05	0.0130
2	-1.229	0.1095	0.1	0.0095
3	-0.525	0.2998	0.15	0.1498
4	-0.513	0.3040	0.2	0.1040
5	-0.508	0.3057	0.25	0.0557
6	-0.486	0.3135	0.3	0.0135
7	-0.482	0.3149	0.35	0.0351
8	-0.323	0.3733	0.4	0.0267
9	-0.261	0.3970	0.45	0.0530
10	-0.068	0.4729	0.5	0.0271
11	-0.057	0.4773	0.55	0.0727
12	0.137	0.5545	0.6	0.0455
13	0.464	0.6787	0.65	0.0287
14	0.595	0.7241	0.7	0.0241
15	0.881	0.8108	0.75	0.0608
16	0.906	0.8175	0.8	0.0175
17	1.046	0.8522	0.85	0.0022
18	1.237	0.8920	0.9	0.0080
19	1.678	0.9533	0.95	0.0033
20	2.455	0.9930	1	0.0070

$D_{20} = 0.149$, за $n = 20$ и $\alpha = 0.05$ из таблице за Колмогорова добија се критична област $c = 0.294$, па је $D_{20} < c$ и прихватамо нулту хипотезу.

пример2: На случајан начин, у програму R, је одабран је узорак од 50 бројева на интервалу $(-2,2)$, тестирати хипотезу да узорак припада нормалној расподели, помоћу теста Колмогоров-Смирнов.

```
x<-runif(50,-2,2)
```

```
x<-sort(x)
```

```
ks.test(x,"pnorm")
```

One-sample Kolmogorov-Smirnov test

data: x

$D = 0.2374$, $p\text{-value} = 0.00581$

alternative hypothesis: two-sided

На основу добијених података можемо рећи да случајан низ од 50 елеменета не припада нормалној расподели, p -вредност теста је мања од 0.05.

```
y<-rnorm(50,0,1)
```

```
ks.test(y,"pnorm")
```

One-sample Kolmogorov-Smirnov test

data: y

$D = 0.13$, $p\text{-value} = 0.3369$

alternative hypothesis: two-sided

Прихватамо хипотезу да узорак припада нормалној расподели, што је и било за очекивати јер смо изабрали узорак који има нормалну расподелу.

```
z<-rexp(100,1)
```

```
ks.test(z,"pnorm")
```

One-sample Kolmogorov-Smirnov test

data: z

D = 0.5043, p-value < 2.2e-16

alternative hypothesis: two-sided

Не прихватамо хипотезу да узорак има нормалну расподелу, и видимо да је р-вредност теста веома мала, што се и очекивало, јер расподела узорка има експоненцијалну расподелу.

Узећемо „мали“ случајан узорак обима четири.

```
d<-runif(4,0,2)
```

```
ks.test(d,"pnorm")
```

One-sample Kolmogorov-Smirnov test

data: x

D = 0.6027, p-value = 0.06522

alternative hypothesis: two-sided

Можемо прихватити хипотезу да узорак има нормалну расподелу, зато што је обим узорка „мали“.

пример3: Претпоставимо да је задатак истраживача у једној фабрици дечијих играчака, да открије да ли деца више воле лопте светлијих или тамнијих тонова. Тест се спроводи на узорку од 10 малишана и свакоме од њих се доноси пет лопти обележених бројевима од 1-5. Обележје чија се расподела тестира је изражена бројевима и лопте се разликују по боји, тако да је бројем 1 означена лопта на којој преовлађују најтамнији тонови, а бројем 5 лопта са најсветлијим тоновима, са нивоом значајности $\alpha = 0.01$.

	Вредности обележја				
	1	2	3	4	5
<i>f</i>	0	1	0	5	4

решење:

H_0 - Не постоји разлика у очекиваном броју избора свих пет тоналитета и свака разлика која се појави је случајна варијација у случајном узорку из униформне расподеле

$$H_0(f_1 = f_2 = f_3 = f_4 = f_5),$$

алтернативна хипотеза је

$H_0(f_1, f_2, f_3, f_4, f_5 \text{ нису све једнаке})$

$F_0(x)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$
$F_n(x)$	$\frac{0}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{6}{10}$	$\frac{10}{10}$
$ F_0(x) - F_{10}(x) $	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{2}{10}$	0

$$P\{D_{10} \geq c\} = \alpha$$

$$D_{10} = \sup |F_0(x) - F_n(x)| = \frac{5}{10} = 0.5$$

а вредност за c добијамо из табеле, $c = 0.4864$, па како је $D_{10} > c$, па је закључак који следи да хипотезу H_0 треба одбацити у корист H_1 .

2.4. Тест Колмогоров-Смирнов за два узорка

Тест Колмогоров-Смирнов за два узорка је веома сличан тесту Колмогоров за један узорак.

Предпоставимо да први узорак $X_1, X_2, \dots, X_m$ дужине m са расподелом $F(x)$ и други узорак $Y_1, Y_2, \dots, Y_n$ дужине n са расподелом $G(x)$ и желимо да тестирамо хипотезу

$$H_0(F(x) = G(x))$$

против

$$H_1(F(x) \neq G(x))$$

Ако су $F_m(x)$ и $G_n(x)$ емпиријске функције расподеле за први и други узорак тада је статистика

$$D_{mn} = \sqrt{\frac{mn}{m+n}} \sup_x |F_m(x) - G_n(x)|.$$

Праг d_n зависи од нивоа значајности α и можемо га наћи из услова

$$P_{H_0}(D_n \geq d_{m,n,\alpha}) = \alpha$$

Критичне вредности теста Колмогоров-Смирнова за $n > 40$ и $m = n$

A	0.20	0.10	0.05	0.02	0.01
$d_{n,m,\alpha}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.92}{\sqrt{n}}$	$\frac{2.15}{\sqrt{n}}$	$\frac{2.30}{\sqrt{n}}$

Критичне вредности теста Колмогоров-Смирнова за $n > 40, m > 40$ и $m = n$

A	0.20	0.10	0.05	0.02	0.01
$d_{n,m,\alpha}$	$1.07k$	$1.22k$	$1.36k$	$1.52k$	$1.63k$

тако да је $k = \sqrt{\frac{m+n}{mn}}$

пример:

На основу два узорка резултата мерења једне случајне променљиве, са нивоом поверења $\alpha = 0.05$ тестирати хипотезу да оба скупа припадају истој расподели.

39	31	39	46	54	31
54	49	46	49	39	60

$$H_0(F(x) = G(x)), H_1(F(x) \neq G(x))$$

$$P_{H_0}(D_n \geq d_{m,n,\alpha}) = \alpha$$

Прорачун је дат у таблици

x	$F_m(x)$	$G_n(x)$	$ F_m(x) - G_n(x) $
31	$\frac{2}{6}$	0	$\frac{2}{6}$
39	$\frac{4}{6}$	$\frac{1}{6}$	$\frac{3}{6}$
46	$\frac{5}{6}$	$\frac{2}{6}$	$\frac{3}{6}$
49	$\frac{5}{6}$	$\frac{4}{6}$	$\frac{1}{6}$
54	$\frac{6}{6}$	$\frac{5}{6}$	$\frac{1}{6}$
60	$\frac{6}{6}$	$\frac{6}{6}$	0

$D_n = \frac{3}{6} = \frac{1}{2} = 0.5$, из таблице Колмогоров-Смирнов одређује се критична вредност $d_{6,6,0.05} = \frac{2}{3} = 0.67$, на основу чега се закључује да се нулта хипотеза прихвата.

Модификација теста Колмогорова

3.1. Тест Куипера

Колмогоров тест мери максимално апсолутно одступање између емпиријске функције расподеле узорка и емпиријске расподеле функције теоријске функције.

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

Колмогоров је дефинисао

$$D_n^+ = \sup_x \{F_0(x) - F_n(x)\}, D_n^- = \sup_x \{F_n(x) - F_0(x)\}$$

Куиперова тест-статистика V_n ,

$$V_n = D_n^+ + D_n^- = \sup_x \{F_0(x) - F_n(x)\} + \sup_x \{F_n(x) - F_0(x)\}$$

Асимптотска дистрибуција је

$$P(V_n < \lambda) = Q \cdot \left(\left[\sqrt{n} + 0,155 + \frac{0,24}{\sqrt{n}} \right] \cdot \lambda \right)$$

када је

$$Q = 2 \sum_{i=1}^{\infty} (4i^2 \lambda^2 - 1) \cdot e^{-2i^2 \lambda^2}$$

Куиперова тест статистика важи и за упоређивање два узорка, f и g са обимом узорка n и m . У овом случају тест-статистика Куипера $V_{n,m}$ укључује одступање две емпиријске функције F и G :

$$V_{n,m} = \sup_x \{F_n(x) - G_m(x)\} + \sup_x \{G_m(x) - F_n(x)\}$$

Асимптотска дистрибуција је

$$P(\lambda > V_n) = Q \cdot \left(\left(\sqrt{\frac{nm}{n+m}} + 0,155 + \frac{0,24}{\sqrt{\frac{nm}{n+m}}} \right) \cdot \lambda \right)$$

када је

$$Q = 2 \sum_{i=1}^{\infty} (4i^2 \lambda^2 - 1) \cdot e^{-2i^2 \lambda^2}$$

У оба случаја, овај тест може да се примени ако су запажања циклична, јер тест-статистика је независна од избора порекла. Из тог разлога, овај тест је инваријантан за цикличне трансформације независних променљивих.

пример: Тестирати низове, из примера2 (2.3. Тест Колмогорова), у програму R да узорак припада нормалној расподели, помоћу теста Куипера.

x- случајан узорак од 50 елемената из интервала $(-2,2)$ (пример2).

y- случајан узорак из нормалне расподеле (пример2).

```
v.test(x, "pnorm", list(0,1), H = NA)
```

Kuiper Test

data: x

$V = 2.9498$, $p\text{-value} = 0.03$

alternative hypothesis: NA

threshold = -Inf, simulations: 100

Не прихватамо хипотезу да узорак има нормалну расподелу, p -вредност теста је мања од 0,05.

```
v.test(y, "pnorm", list(0,1), H = NA)
```

Kuiper Test

data: y

$V = 1.1212$, $p\text{-value} = 0.29$

alternative hypothesis: NA

threshold = -Inf, simulations: 100

Прихватамо нулту хипотезу, p -вредност теста је велика, а и случајан низ је из нормалне расподеле.

3.2. Крамер-фон Мизес тест

Нека је $X_1, X_2, \dots, X_n$ независне случајне променљиве из узорка X са функцијом расподелом F . Нулта хипотеза

$$H_0(F(x) = F_0(x))$$

где је $F_0(x)$ дата функција расподеле, онда је тест-статистика Крамера-вон Мизеса

$$\psi(F_n(x)) = \int_{-\infty}^{+\infty} [F_n(x) - F_0(x)]^2 dF_0(x)$$

где је F_n емпиријска функција расподеле на основу узорка $X_1, X_2, \dots, X_n$.

Нека су $x_1, x_2, \dots, x_n$ дати подаци у растућем поретку, тада је тест статистика

$$T = n\psi = \frac{1}{12n} + \sum_{i=1}^n \left[\frac{2i-1}{2n} - F(x_i) \right]$$

Ако је вредност тест статистике већа од табличне вредности, хипотеза се одбацује.

пример: Тестирати низове, из примера2 (2.3. Тест Колмогорова), у програму R да узорак припада нормалној расподели, помоћу теста Крамер-фон Мизес.

х- случајан узорак од 50 елемената из интервала $(-2,2)$ (пример2).

у- случајан узорак из нормалне расподеле (пример2).

z- случајан узорак који има експоненцијалну расподелу (пример2).

cvm.test(x)

Cramer-von Mises normality test

data: x

$W = 0.2657$, $p\text{-value} = 0.0007818$

На основу добијених вредности можемо приметити да случајан низ од 50 елемената не припада нормалној расподели, има малу p-вредност теста.

cvm.test(y)

Cramer-von Mises normality test

data: y

$W = 0.0275$, $p\text{-value} = 0.8759$

p-вредност теста је велика па прихватамо хипотезу да узорак припада нормалној расподели, а и сам узорак је на случајан начин изабран из нормалне расподеле.

cvm.test(z)

Cramer-von Mises normality test

data: z

$W = 0.9595$, $p\text{-value} = 2.111e-09$

Одбацујемо хипотезу да узорак има нормалну расподелу, p-вредност теста је мала.

3.3. Андерсон-Дарлинг тест

За дато x и хипотетичку функцију расподеле $F_0(x)$, случајна променљива nF_n има биномну расподелу $F_0(x)$. Очекивана вредност за nF_n је $F_0(x)$ и варијансом $nF_0(x)[1 - F_0(x)]$. Посебно можемо истражити репове $F_0(x)$ расподеле. Изаберимо функцију

$$\psi(x) = \frac{1}{u(1-u)}$$

тада за одређено x

$$\sqrt{n} \frac{F_n(x) - F_0(x)}{\sqrt{F_0(x)[1 - F_0(x)]}}$$

има средњу вредност 0 и варијансом 1 када је нулта хипотеза тачна.

Тест-статистика је

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{[F_n(x) - F_0(x)]^2}{F_0(x)[1 - F_0(x)]} dF_0(x)$$

или

$$A_n^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) (\log u_{(i)} + \log(1 - u_{(n-i+1)}))$$

где је $u_{(i)} = F_0(x_{(i)})$, $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

пример: Тестирати низове, из примера2 (2.3. Тест Колмогорова), у програму R да узорак припада нормалној расподели, помоћу теста Андерсон-Дарлинг.

x- случајан узорак од 50 елемената из интервала $(-2,2)$.

y- случајан узорак из нормалне расподеле.

z- случајан узорак који има експоненцијалну расподелу.

ad.test(x)

Anderson-Darling normality test

data: x

$A = 1.6233$, $p\text{-value} = 0.0003126$

На основу добијених података можемо рећи да случајан низ од 50 елемената не припада нормалној расподели, р-вредност теста је мања од 0.05.

ad.test(y)

Anderson-Darling normality test

data: y

$A = 0.18$, $p\text{-value} = 0.9113$

Прихватамо нулту хипотезу, р-вредност теста је велика, а и случајан низ је из нормалне расподеле.

ad.test(z)

Anderson-Darling normality test

data: z

$A = 5.5183$, $p\text{-value} = 1.057e-13$

Одбацујемо хипотезу да узорак има нормалну расподелу, р-вредност теста је мала, а и узорак има експоненцијалну расподелу.

3.4. Lilliefors тест

Lilliefors тест је модификација теста Колмогорова. Тест Колмогоров је одговарајући у ситуацији у којој су параметри расподеле потпуно познати. Међутим, понекад је тешко у потпуности одредити параметре, тј. расподела је непозната. У овом случају, параметри треба да се процењују на основу података из узорка и тада Колмогорова статистика може да завара. Вероватноћа грешке прве врсте може бити мања од оних датих у табlici. Lilliefors тест се процењује на основу узорака, па у ситуацији када су непознати параметри Lilliefors тест је у предности у односу на тест Колмогорова.

Тест-статистика за Lilliefors тест је дефинисана као

$$D = \max_x |F(x) - F_n(x)|$$

$F(x)$ је из нормалне расподеле са параметрима $\mu = \bar{X}$ и s^2 поправљеном узорачком дисперзијом.

Иако Lilliefors тест има исту тест-статистику као и тест Колмогорова таблица за критичну вредност је другачија, што доводи до другачијих закључака.

Ако је D веће од одговарајуће критичне вредности у табlici, онда се одбацује нулта хипотеза о нормалности узорка.

пример: Тестирати низове, из примера2 (2.3. Тест Колмогорова), у програму R да узорак припада нормалној расподели, помоћу теста Lilliefors.

x - случајан узорак од 50 елемената из интервала $(-2,2)$.

y - случајан узорак из нормалне расподеле.

lillie.test(x)

Lilliefors (Kolmogorov-Smirnov) normality test

data: x

$D = 0.163$, $p\text{-value} = 0.001949$

На основу добијених вредности за p -вредност теста можемо закључити да случајан узорак од педесет елемената нема нормалну расподелу.

lillie.test(y)

Lilliefors (Kolmogorov-Smirnov) normality test

data: y

$D = 0.0761$, $p\text{-value} = 0.6653$

Прихватимо хипотезу да низ има нормалну расподелу, што је и било за очекивати јер је низ изабран баш из нормалне расподеле, те је и зато велика p -вредност теста.

Моћ тестова

У првом поглављу смо представили шта је моћ статистичког теста. Моћ статистичког теста је вероватноћа да ће тест одбацити нулту хипотезу када је алтернативна хипотеза тачна.

Како моћ теста расте, смањује се шанса да дође до грешке друге врсте. Уколико је вероватноћа мала, било би добро изменити узорак или завршити тестирање. Фактори који могу утицати на повећање моћи теста је: повећање обима узорка, повећања нивоа значајности и смањење варијабилности у узорку.

У програмском језику *R* направићемо функцију која пореди моћ теста за поједине тестове од великог броја узорака ($N=1000$), истог обима ($n=20,30,40,50,60,70,80,90,100$) из „контаминираних“ нормалних расподела.

Прво направићемо функцију која генерише узорак произвољног обима из контаминираних нормалних расподела,

$N(m_1, 1)$ са вероватноћом p_1

$N(m_2, 1)$ са вероватноћом p_2

где важи $p_1 + p_2 = 1$, и који има мешавину две нормалне расподеле. Његова функција расподеле је $F(x) = p_1 \cdot F_1(x) + p_2 \cdot F_2(x)$.

Код којим се генерише узорак:

```
generator<-function(n,m1,m2,p1){
  x<-rep(0,n)
  y<-runif(n)
  pom<-y<=p1
  k<-length(pom[pom])
  x[pom]<-rnorm(k,mean=m1,sd=1)
  x[!pom]<-rnorm(n-k,mean=m2,sd=1)
  x
}
```

Функција генерише низ из контаминираних нормалних расподела.

t-test и Тест Колмогорова применићемо директно *t.test* и *ks.test*, док ћемо Хи-квадрат тест (Пирсонов тест) *hi0.test*, Андерсон-Дарлинг тест *ad0.test* и Крамер-фон Мизес тест *cvm0.test* модификовати у односу на стандардне тестове који се могу позвати директно.

Модификације тестова:

Хи-квадрат тест (Пирсонов тест)

```
pearson0.test<-function(x, n.classes = ceiling(2 * (n^(2/5))))
{
  x <- x[complete.cases(x)]
  n <- length(x)
  num <- floor(1 + n.classes * pnorm(x))
```

```
count <- tabulate(num, n.classes)
prob <- rep(1/n.classes, n.classes)
хрес <- n * prob
h <- ((count - хрес)^2)/хрес
P <- sum(h)
pvalue <- pchisq(P, n.classes - 1, lower.tail = F)
pvalue
}
```

Андерсон-Дарлинг тест

```
ad0.test<-function (x)
{
  x <- sort(x[complete.cases(x)])
  n <- length(x)
  if (n < 8)
    stop("sample size must be greater than 7")
  p <- pnorm(x)
  h <- (2 * seq(1:n) - 1) * (log(p) + log(1 - rev(p)))
  A <- -n - mean(h)
  AA <- (1 + 0.75/n + 2.25/n^2) * A
  if (AA < 0.2) {
    pval <- 1 - exp(-13.436 + 101.14 * AA - 223.73 * AA^2)
  }
  else if (AA < 0.34) {
    pval <- 1 - exp(-8.318 + 42.796 * AA - 59.938 * AA^2)
  }
  else if (AA < 0.6) {
    pval <- exp(0.9177 - 4.279 * AA - 1.38 * AA^2)
  }
  else {
    pval <- exp(1.2937 - 5.709 * AA + 0.0186 * AA^2)
  }
  pval
}
```

Крамер-фон Мизес тест

```
cvm0.test<-function (x)
{
  x <- sort(x[complete.cases(x)])
  n <- length(x)
  if (n < 8)
    stop("sample size must be greater than 7")
  p <- pnorm(x)
  W <- (1/(12 * n) + sum((p - (2 * seq(1:n) - 1)/(2 * n))^2))
  WW <- (1 + 0.5/n) * W
  if (WW < 0.0275) {
    pval <- 1 - exp(-13.953 + 775.5 * WW - 12542.61 * WW^2)
  }
  else if (WW < 0.051) {
    pval <- 1 - exp(-5.903 + 179.546 * WW - 1515.29 * WW^2)
  }
}
```

```
else if (WW < 0.092) {
  pval <- exp(0.886 - 31.62 * WW + 10.897 * WW^2)
}
else if (WW < 1.1) {
  pval <- exp(1.111 - 34.242 * WW + 12.832 * WW^2)
}
else {
  #p-value is smaller than 7.37e-10, cannot be computed more accurately
  pval <- 7.37e-10
}
pval
}
```

Функција која рачуна моћ сваког теста и исписује табелу на хиљаду узорака са различитим обимом, функција исписује табелу са обим од 20-100.

```
moc<-function(n=seq(20,100,by=10), m0,m1, p1, N=1000){
x<-c()
matr<-matrix(9,5)
dimnames(matr)<-list(c("n=20","n=30","n=40","n=50","n=60","n=70","n=80","n=90","n=100"),
c("t-test","Pirsonov test","Kolmogorov.test","A.Darling test","Kramer V.M test"))
ind<-rep(0,5)
for(k in 1:length(n)){
  for(i in 1:N){
    x<-generator(n[k],m1=m1,m2=m0,p1)
    niz<-x-m0
    if(t.test(niz)$p.value<0.05)
      ind[1]=ind[1]+1
    if(pearson0.test(niz)<0.05)
      ind[2]=ind[2]+1
    if(ks.test(niz,"pnorm")$p.value<0.05)
      ind[3]=ind[3]+1
    if(ad0.test(x)<0.05)
      ind[4]=ind[4]+1
    if(cvm0.test(x)<0.05)
      ind[5]=ind[5]+1
  }
  matr[k,]<-ind/N
  ind<-rep(0,5)
}
matr
}
```

$m0=2, m1=0, p1=0.9$

		тестови				
		t-test	Pirsonov test	Kolmogorov.test	A.Darling test	Kramer V.M test
Број елемената у узорку	n=20	1	1	1	0.712	0.585
	n=30	1	1	1	0.774	0.637
	n=40	1	1	1	0.808	0.681
	n=50	1	1	1	0.834	0.707
	n=60	1	1	1	0.865	0.714
	n=70	1	1	1	0.892	0.774
	n=80	1	1	1	0.905	0.789
	n=90	1	1	1	0.913	0.791
	n=100	1	1	1	0.927	0.812

На основу добијених података видимо да на овом узорку Хи-квадрат тест и Тест Колмогорова имају велику моћ, док модификације Теста Колморова, Крамер-фон Мизес тест и Андерсон-Дарлинг тест, имају мању моћ.

Закључак

У овом мастер раду су урађени параметарски и непараметарски тестови, са тиме што је посебан акценат стављен на тест Колмогорова и његовим модификацијама.

Кроз рад смо закључили да је за непараметарске тестове потребно мање претпоставки, њихова примена је много шира од параметарских тестова. Конкретно, они се могу применити у ситуацијама у којима се мање зна о узорку, па због мање претпоставки, непараметарске методе су много поузданије. Још један од разлога за употребу непараметарских тестова је њихова једноставност, лакши су за учење. Код непараметарских тестова искази добијени из вероватноће непараметарских тестова су тачне вероватноће, осим у случају великих узорака, где се користе одличне апроксимације. Постоји неколико подобних непараметарских статистичких тестова за узорке добијене из опсервација неколико различитих популација (Тест Колмогоров-Смирнов). Ниједан од параметарских тестова се не може употребити за овакве податке без увођења неких нереалних претпоставки.

Неке непараметарске методе се могу применити и у анализи квалитативних обележја чији модалитети не морају бити изражени нумерички као што смо урадили кроз пример у делу Тест Колмогорова.

Уколико узорак има мали број елемената Тест Колмогорова може са великим нивоом поузданости да нам да повратну информацију о узорку. Видели смо да се у Тесту Колмогорова сваки елемент из узорка користи без икаквог груписања, док се код Хи-квадрат теста елементи групишу, под неким условима, па се тек онда примењује тест.

Зашто уводимо модификације теста Колмогорова?

Андерсон-Дарлинг тест је модификовани тест Колмогорова и уведен је из разлога зато што је доста осетљивији на девијације претпостављене расподеле на реповима него тест Колмогорова. Колмогоров-Смирнов тест за два узорка је најкориснији непараметарски тест при поређењу два узорка јер је осетљив и на облик и на положај емпиријске функције расподеле два узорка. Куиперов тест као модификација је ефикаснији уколико је обим узорка велик.

Када су параметри непрекидне расподеле непознати, а морају бити оцењени, тада стандардне таблице критичне вредности за Колмогоров тест не одговарају правој расподели тест-статистике, тада можемо применити Lilliefors тест. Недостатак теста Колмогорова је слаба моћ детектовања разлика теоријске и емпиријске функције на реповима и зато се уводе модификације Теста Колмогорова.

Симулације, које смо урадили, показују да су модификоване верзије препоручљивије када је и обим узорка велик.

Литература

Anderson T. W. (2010.) *Anderson-Darling Tests of Goodness-of-Fit*, Stanford: Stanford University

Гилезан С., Лужанин З., Грбић Т., Михаиловић Б., Недовић Љ., Овцин З., Иветић Ј., Дорословачки К. (2009.) *Вероватноћа и статистика*, Нови Сад: TEMPUS

Јевремовић В., Малишић Ј. (2002.) *Статистичке методе у метеорологији и инжењерству*, Београд: Савезни хидрометеоролошки завод

David J. Sheskin (2000.) *PARAMETRIC and NONPARAMETRIC STATISTICAL PROCEDURES*, Danbury: Western Connecticut State University

Jean Dickinson Gibbons, Subhabrata Chakraborti (2003.) *Nonparametric Statistical Inference*, Alabama: The University of Alabama Tuscaloosa, U.S.A.

Меркле М. (2010.) *Вероватноћа и статистика*, Београд: Академска мисао

Младеновић П. (2002.) *Вероватноћа и статистика*, Београд: Математички факултет

http://www.google.rs/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&sqi=2&ved=0CDUQFjAC&url=http%3A%2F%2Focw.mit.edu%2Fcourses%2Fmathematics%2F18-443-statistics-for-applications-fall-2006%2Flecture-notes%2Flecture14.pdf&ei=DG1AUpeiKOeL4AS2woG4AQ&usg=AFQjCNEmy01Mu_boFtxBsxyas1KnG2Fx-A&sig2=8mccH8mHXo-W9yFONpptOQ&bvm=bv.52434380,d.bGE
посећено 21.07.2013. године

http://www.google.rs/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&sqi=2&ved=0CEQQFjAE&url=http%3A%2F%2Fwww.usna.edu%2FUsers%2Fmath%2Fjager%2Fcourses%2Fsm439%2Fflabs%2Fflab5.pdf&ei=821AUoafDqKn4gSihICACQ&usg=AFQjCNEDwV2Bpyt440EVP7h6_D0cwNyyZQ&sig2=TZlpZ782LS3Ouq0jawqH9Q&bvm=bv.52434380,d.Yms
посећено 19.08.2013. године

http://www.google.ru/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CDAQFjAA&url=http%3A%2F%2Fquantile.ru%2F09%2F09-IK.pdf&ei=W_5BUqHJF8iatAaBoYHICA&usg=AFQjCNGKzfweuSKWoQOa4ydETtq2k0g6rg&bvm=bv.53077864,d.bGE&cad=rjt
посећено 5.09.2013. године

www.ge.infn.it/geant4/analysis/HEPstatistics/gof/deployment/userdoc/statistics/documents/Kuiper.pdf посећено 22.04.2013. године

www.win.tue.nl/~rmcastro/2WS05/files/weak_Glivenko-Cantelli_note.pdf
посећено 15.04.2013. године

www.maths.qmul.ac.uk/~bb/CTS_Chapter3_Students.pdf
посећено 7.05.2013. године

<http://www.statmethods.net/stats/power.html>
посећено 25.09.2013. године