

Анализа података у биологији

Марија Цупарић

Шести час

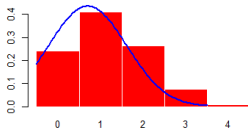


Апроксимација биномне расподеле нормалном

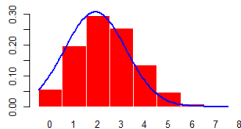
Размотримо биномну расподелу са параметрима n и $p = 0.3$.

$$P\{X = x\} = \binom{n}{x} 0.3^x \cdot 0.7^{n-x}$$

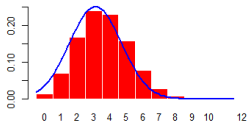
$n = 4, np = 1.2$



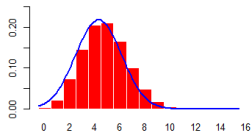
$n = 8, np = 2.4$



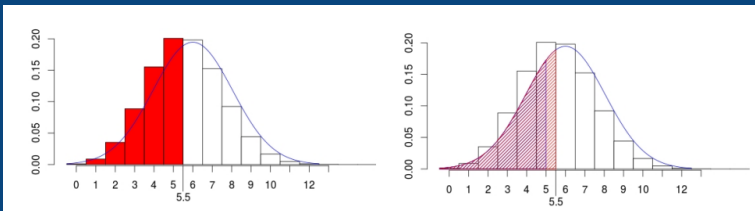
$n = 12, np = 3.6$



$n = 16, np = 4.8$



Пример Нека X има биномну расподелу са параметрима $n = 20$ и $p = 0.3$. Одредимо $P\{X \leq 5\}$



$$\mathcal{B}(20, 0.3)$$

$$\mathcal{N}(np, np(1-p)) = \mathcal{N}(6, 4.2)$$

$$\begin{aligned} P\{X \leq 5\} &= \sum_{x=0}^5 \binom{20}{x} 0.3^x \cdot 0.7^{20-x} \\ &= 0.4163708 \end{aligned}$$

$$\begin{aligned} P\{X \leq 5\} &= P\left\{ \frac{X - 6}{\sqrt{4.2}} \leq \frac{5 - 6}{\sqrt{4.2}} \right\} \\ &= P\{Z \leq -0.24\} \\ &= 0.4052 \end{aligned}$$

Нека X има биномну расподелу са параметрима n и p . Уколико је $p \leq 0.5$ и $np > 5$ или $p \geq 0.5$ и $n(1-p) > 5$, тада, за природне бројеве a и b важи

$$P\{a \leq X \leq b\} \approx P\left\{ \frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}} \right\},$$

где Z има стандардну нормалну расподелу.



Пример Вероватноћа да једно семе успешно проклија износи 0.1. Ако се посади 70 семена, колика је вероватноћа да тачно 5 семена проклија? Колика је вероватноћа да их бар 5 проклија?

Нека је X број проклијалог семења. Тада X има биномну расподелу са параметрима $n = 70$ и $p = 0.1$. Како је $p < 0.5$ и $np = 7 > 5$, па се може користити апроксимација нормалном расподелом.

$$\begin{aligned} P\{X = 5\} &= P\{5 \leq X \leq 5\} \approx P\left\{\frac{4.5 - 7}{\sqrt{6.3}} \leq Z \leq \frac{5.5 - 7}{\sqrt{6.3}}\right\} \\ &= P\{-0.996 \leq Z \leq -0.60\} = 0.274 - 0.160 = 0.114. \end{aligned}$$

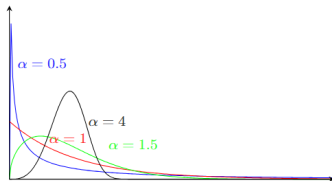
$$\begin{aligned} P\{X \geq 5\} &= 1 - P\{X < 5\} = 1 - P\{X \leq 4\} \approx 1 - P\left\{Z \leq \frac{4.5 - 7}{\sqrt{6.3}}\right\} \\ &= 1 - P\{Z \leq -0.996\} = 1 - 0.160 = 0.84. \end{aligned}$$



```
> dbinom(5,70,0.1)
0.1284271
> 1-pbinom(4,70,0.1)
0.8412055
```

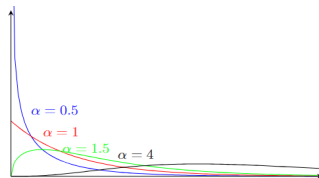
Вејбулова расподела

$$f(x) = \alpha \lambda^\alpha x^{\alpha-1} e^{-\lambda x^\alpha}, \quad x > 0$$



Гама расподела

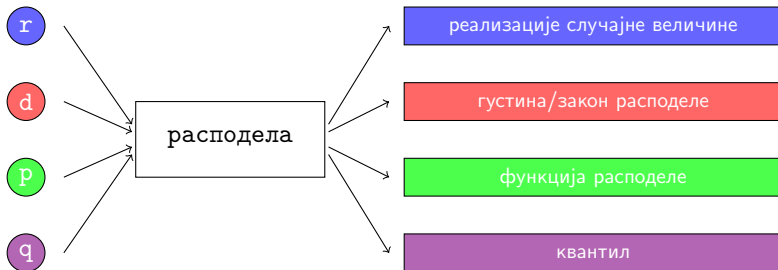
$$f(x) = \frac{x^{\alpha-1} \lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0$$



R `rweibull(n, shape, scale)`

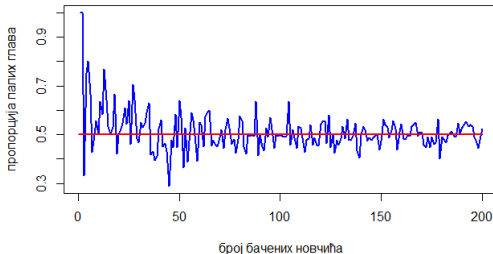
`rgamma(n, shape, rate)`

- ★ Обе расподеле имају параметре облика $\alpha > 0$ и размере $\lambda > 0$ (или скалирања $\frac{1}{\lambda}$)
- ★ За $\alpha = 1$ свде се на експоненцијалну с параметром стопе отказа λ
- ★ Померене су удесно, што је израженије за мање вредности α
- ★ Ако је $\alpha > 1$ стопа отказа расте током времена, тј. половна компонента има већу стопу отказа од нове, ако је $\alpha < 1$, онда је обрнуто



Закон великих бројева

На слици је приказана пропорција палих глава у симулацији бацања правилног новчића n пута за $n \in \{1, 2, \dots, 200\}$. Црвеном линијом је представљена теоријска вредност вероватноће да падне глава приликом бацања новчића. Можемо приметити да како се повећава n , симулиране вредности се све више групишу око тачна вредности.



Закон великих бројева објашњава да при n мерења неке неслучајне величине m , која су заправо везана за случајне резултате $X_1, \dots, X_n$, ако број мерења расте, $\frac{1}{n} \sum_{k=1}^n X_k$ је све „ближа” вредности m .



Примери

- ★ процена висине свих људи на основу узорка појединаца,
- ★ предвиђање дугорочних перформанси играча у казину,

Можете победити кућу са неколико срећних руку, али на дуге стазе, кућа увек побеђује!

- ★ предвиђање будућих одштета осигурања како би се одредиле премије,
- ★ предвиђања у вези постојаности система,...

Погрешна примена ЗВБ на „кратак рок” може довести до „коцкарске заблуде”:
Ако наставим да се коцкам, моја срећа ће се променити и повратићу новац који сам изгубио.

ЗВБ даје предвиђања само о понашању при бесконачно много експеримената
– не може се користити за предвиђање исхода одређеног експеримента.

Зато поступак треба да буде следећи

- симуирати **велики број** независних реализација посматране случајне величине или догађаја;
- израчунати одговарајућу **просечну вредност** (аритметичку средину) над симулацијама, чиме се **апроксимира** тражено очекивање, интеграл или вероватноћа.

Наведени поступак је заправо Монте Карло метод. ЗВБ гарантује да просечно симулирано понашање конвергира ка правој вредности.

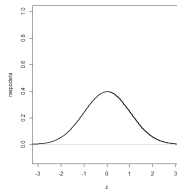
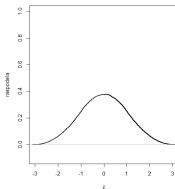
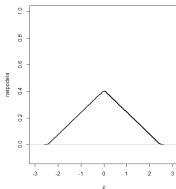
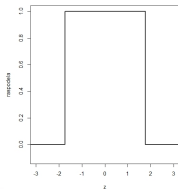
Понављање експеримента велики број пута нам омогућава да о неким величинама закључујемо емпиријски. **Монте Карло методе** користе стохастичке симулације (велики број пута) за апроксимацију решења проблема који је веома тешко аналитички решити.



Када се неки резултат формира као **збир великог броја малих, независних и случајних утицаја**, његова расподела ће **увек бити приближно нормална**, без обзира како изгледа појединачни утицај.

Пример Дужина тела јединки у популацији зависи од великог броја фактора (генетика, исхрана, услови средине). Иако је сваки од тих утицаја појединачно непредвидљив → **укупна варијација дужине прати приближно нормалну расподелу**.

Нека су $X_1, X_2, \dots$ независне случајне величине са униформном $\mathcal{U}(0, 1)$ расподелом и нека је $\tilde{Z}_n = \frac{S_n - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$



Нека је $X_1, \dots, X_n$ низ независних случајних величина са истом расподелом, очекивањем m и дисперзијом $\sigma^2 < \infty$. Тада је њихова просечна вредност $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ приближно нормално расподељена за n велико, тј.

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

Колико n је довољно велико? Чести кажемо „ако је $n \geq 30$ “, међутим, то зависи од облика расподеле полазне променљиве. Што се расподела више разликује од нормалне, величина узорка мора бити већа.

ЦТГ нам омогућава да у реалности не морамо да понављамо студије много пута да бисмо оценили расподелу средње вредности, већ да то можемо оценити из једног пута.



Основна сврха статистике је доношење закључака о популацији на основу узорка узетог из популације.

Постоје два основна начина статистичког закључивања:

оцењивање параметара примењује се кад немамо претходна знања о параметру, на основу података процењујемо вредност параметра

тестирање хипотеза примењује када имамо претпоставку од правој вредности непознатог параметра, подаци иду у прилог или не претпостављене вредности

Пример Приликом проучавања криминала популацију чине све особе старије од 16 година које су осуђене због неког кривичног дела. Занима нас:

- (1) Колики је средњи број година образовања у тој популацији?
- (2) Да ли је већина чланова популације ухапшена бар једном пре него што је први пут осуђена?

Пре самог статистичког закључивања потребно је:

- (1) одредити популацију
- (2) одредити случајну променљиву коју проучавамо
- (3) одређивање параметра који нас интересује
- (4) извући узорак из популације

Пример (наставак) Са популацијом је повезана случајна променљива X , број година формалног образовања појединца. Желимо да изаберемо подскуп од n појединаца из популације случајно. То значи да желимо да изаберемо n особа за наше истраживање на такав начин да избор једног појединца ни не искључује нити гарантује избор било ког другог. На овај начин, избор једног појединца је независан од избора било ког другог. Ово подразумева случајан узорак.

Прост случајан узорак обима n чине случајне величине $X_1, \dots, X_n$ које су међусобно независне и све имају исту расподелу као обележје.

Случајан избор се постиже или уз помоћ таблица случајних бројева или коришћењем рачунара.

Које оцене параметара су до сада рађене?

Када оцена може да се одреди без икаквог предзнања о расподели обележја онда говоримо о непараметарском оцењивању.

Сада претпоставимо да је статистички модел одређен до на непознате параметре расподеле, односно

- ★ имамо претпоставке о расподели обележја $F(\theta)$
- ★ желимо, на основу доступног узорка, да оценимо параметре те расподеле или неке функције од њих

Овакав приступ припада домену параметарске статистике.

Пример На пример, знамо да $X \in \mathcal{N}(m, \sigma^2)$ али не знамо вредности од m и σ^2 . Циљ: на основу доступног узорка оценити их.

- ★ тачкасте оцене параметра
- ★ интервалне оцене параметра

- ★ Тачкаста оцена непознатог параметра θ је нека статистика T чије вредности дају добру процену о вредности тог параметра
- ★ Оцена T је случајна величина јер за различите узорке узима различите вредности, она никад неће бити баш једнака θ , али се надамо да даје добру процену
- ★ Квалитетна оцена непознатог параметра је она која је
 - ▷ у просеку блиска стварној вредности параметра; математичко очекивање оцене једнако је параметру - непристрасност
 - ▷ њена дисперзија је мала када n расте

Најчешће коришћени методи за добијање оцена

принцип замене: тражи шта представља параметар у популацији и оцењује га одговарајућом статистиком из узорка

метод максималне веродостојности: рачуна вероватноћу добијања конкретног узорка као функцију непознатог параметра и оцењује параметар вредношћу за коју функција достиже максимум



Тачкаста оцена параметра средње вредности

Претпоставимо да имамо популацију и на њој дефинисану случајну величину X чији су средња вредност m и дисперзија σ^2 непознати.

Тачкаста оцена параметра m је $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.

Оцена $\bar{X}$ има пожељне особине, тј. $E\bar{X} = m$ и $D\bar{X} = \frac{\sigma^2}{n}$ што је мало када је n велико.

Пример Претпоставимо да студент мери температуру кључања (y °C) одређене течности и добио је резултате 102.5, 101.7, 103.1, 100.9, 100.5 и 102.2 на 6 различитих узорака течности. На основу овог узорка, добија се да је оцена за средњу температуру кључања:

$$\bar{x} = \frac{1}{6}(102.5 + 101.7 + 103.1 + 100.9 + 100.5 + 102.2) = 101.8167.$$

Дефиниција: Нека је θ непознати параметар. Нека су L_n и U_n статистике за које је $P\{L_n < U_n\} = 1$ и $P\{L_n < \theta < U_n\} = \beta$. Интервал (L_n, U_n) назива се $100\beta\%$ двострани интервал поверења за параметар θ , а β је ниво поверења.

- ★ Једностранни доњи и горњи интервал поверења за θ су интервали облика (L_n, ∞) и $(-\infty, U_n)$, где су L_n и U_n статистике одређене тако да је ниво поверења β .

У зависности од параметарског простора, $-\infty$, односно ∞ , треба заменити коначним вредностима.

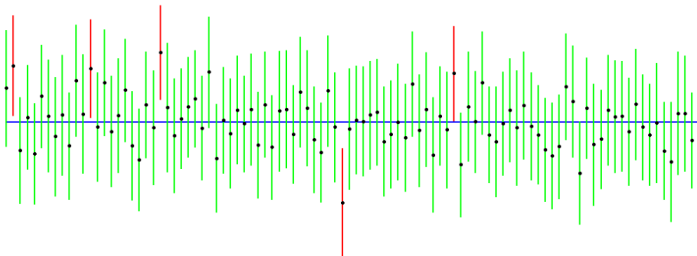
За одређивање граница интервала (из одговарајућих вероватноћа) треба нам расподела неке случајне величине.

Генерално, да би се конструисао $100\beta\%$ интервал поверења за параметар θ , потребно је пронаћи случајну променљиву чији израз садржи θ и чија је расподела вероватноће позната, макар приближно.

Интерпретација реализованог интервала:

погрешно: вероватноћа да се непознати параметар нађе у интервалу $(\hat{L}_n, \hat{U}_n)$, где су $\hat{L}_n$ и $\hat{U}_n$ реализоване вредности статистика, је β

тачно: у просеку ће се у $100\beta\%$ случајева стварна вредност параметра налазити у реализованом интервалу поверења $(\hat{L}_n, \hat{U}_n)$



Шта значи да имамо поверење од 90%? То значи да ће 90% узорака „ухватити” вредност θ , а 10% ће га промашити. Ми „верујемо” да је наш узорак онај који „хвата” праву вредност непознатог параметра.

Интервал поверења за параметар средње вредности

Видели смо да је узорачка средина $\bar{X}$ логична оцена средње вредности m .
Која је расподела од $\bar{X}$?

Теорема: Нека је $X_1, \dots, X_n$ узорак обима n из нормалне расподеле са параметрима m и σ^2 . Тада $\bar{X}$ има нормалну расподелу чија је средња вредност m и дисперзија $\frac{\sigma^2}{n}$.

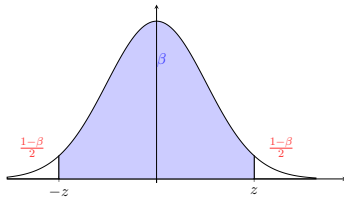
На основу стандардизације, $Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ има стандардну нормалну расподелу.

Уколико узорак није из нормалне, него из неке друге расподеле чија је средња вредност m , а дисперзија σ^2 , онда Z добијено горњом формулом нема нормалну расподелу, али за велико n ($n > 30$) има приближно нормалну расподелу (што следи из ЦТГ).

Интервал поверења за m када је σ^2 познато

Нека је $X_1, \dots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима m и познатом вредношћу σ^2 .

- ★ случајна променљива: $Z = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0, 1)$
- ★ треба одредити константу z тако да $P\{-z < Z < z\} = \beta$



- ★ z је такво да важи $P\{Z > z\} = \frac{1-\beta}{2}$ или $z = F^{-1}(\frac{1+\beta}{2})$
- ★ $-z < Z < z \iff -z < \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z \iff \bar{X}_n - z \frac{\sigma}{\sqrt{n}} < m < \bar{X}_n + z \frac{\sigma}{\sqrt{n}}$

Интервал поверења је $(\bar{X}_n - z \frac{\sigma}{\sqrt{n}}, \bar{X}_n + z \frac{\sigma}{\sqrt{n}})$, а његова дужина је $d = 2z \frac{\sigma}{\sqrt{n}}$.

Интерпретација интервала поверења: Генерисаћемо узорке обима 10 из нормалне $\mathcal{N}(0, 1)$ и формирати 95% интервале поверења (за параметар m када је σ^2 познато). Када се ова процедура понови велики број пута у приближно 95% случајева стварна вредност параметра m (у нашем случају 0) треба да се нађе унутар реализованог интервала.

```
intervalPoverenjaM <- function(n,m,sigma,beta){
  xsr <- mean(rnorm(n,m,sigma))
  L <- xsr - qnorm((1+beta)/2)*sigma/sqrt(n)
  U <- xsr + qnorm((1+beta)/2)*sigma/sqrt(n)
  return(c(L,U))
}
brojac<-0
N<-10000
for( i in 1:N){
  intpov<-intervalPoverenjaM(10,0,1,0.95)
  if((intpov[1]<0)&&(intpov[2]>0))
    brojac<-brojac+1
}
brojac/N
```

Резултат је: 0.947

