



ПРЕДМЕТ  
< МЕДИЦИНСКА СТАТИСТИКА >

Област број 2  
< ВЕРОВАТНОЋА, НОРМАЛНА РАСПОДЕЛА И ПРЕДВИЋАЊЕ >

Област	Наставна јединица	Тематске јединице	Резултат – знања или вештине које студент треба да добије
2	Вероватноћа. Нормална расподела. Предвиђање.	Особине вероватноће. Расподела вероватноће и случајне променљиве. Биномна расподела. Средина и варијанса. Poisson-ова расподела. Нормална расподела. Променљиве које прате Нормалну расподелу. Нормални графикон. Расподеле узорака. Стандардна грешка средине узорка. Интервали поверења. Стандардна грешка и интервал поверења за пропорцију. Разлика између две средине. Поређење две пропорције.	Упознавање са основама вероватноће, Нормалне расподеле и предвиђањем.

## САДРЖАЈ

Вероватноћа, Нормална расподела и Предвиђање.....	2
3. Вероватноћа .....	2
3.1 Вероватноћа .....	2
3.2 Особине вероватноће .....	3
3.3 Расподела вероватноће и случајне променљиве .....	3
3.4 Биномна расподела (Binomial distribution).....	4
3.5 Средина и варијанса (Mean and variance).....	6
3.6 Карактеристике средине и варијансе .....	6
3.7 Poisson-ова расподела .....	7
4. Нормална расподела (Normal distribution) .....	9
4.1 Вероватноћа непрекидних променљивих .....	9
4.2 Нормална расподела (The Normal distribution) .....	11
4.3 Особине Нормалне расподеле .....	13
4.4 Променљиве које прате Нормалну расподелу .....	16
4.5 Нормални графикон (The Normal plot) .....	18
5 Предвиђање.....	21
5.1 Расподеле узорка (Sampling distributions) .....	21
5.2 Стандардна грешка средине узорка (Standard error of a sample mean) .....	23
5.3 Интервали поверења (Confidence intervals).....	25
5.4 Стандардна грешка и интервал поверења за пропорцију .....	27
5.5 Разлика између две средине.....	27
5.6 Поређење две пропорције .....	28
5.7 Који је тачан интервал поверења? .....	30

Област бр. 2

## < Вероватноћа, Нормална расподела и Предвиђање >

### Вероватноћа, Нормална расподела и Предвиђање

#### 3. Вероватноћа

##### 3.1 Вероватноћа

Податке из узорка користимо да донесемо закључке о популацији из које је узет узорак. На пример, на клиничком истраживању можемо запазити да одређени број пацијената који су добили нови третман, боље реагује од пацијената који су добили стари третман. Желимо да знамо да ли ће напредак бити уочљив код целе групе пацијената, и ако је тако, колико велики би он могао бити. Теорија вероватноће нам омогућава да повежемо узорке и популацију, и да донесемо закључке о популацији на основу узорака. Почећемо дискусију о вероватноћи са неким једноставним средствима, као што су новчићи и коцкице, али повезаност са медицинским проблемима ће убрзо постати јасна.

Прво поставимо питање шта тачно значи "вероватноћа". У овом делу користимо учесталу дефиницију: *вероватноћа да ће се неки догађај десити под одређеним условима може се дефинисати као однос понављања оних услова у којима би се догађај дешавао у дугорочном периоду*. На пример, ако бацимо новчић он падне на главу или на писмо. Пре него што га бацимо, немамо никаква сазнања на коју ће страну пасти, али знамо да ће то бити писмо или глава. Наравно, након што смо га бацили, знамо који је исход. Ако наставимо да бацамо новчић, требало би да добијемо неколико глава и неколико писама. Ако наставимо са истим извесно време, онда ћемо очекивати да добијемо исто онолико глава колико и писама. Вероватноћа да добијемо главу је половична, јер у дугом низу бацања, глава би требало да падне у половини од укупног броја бацања. Број глава који би могао да настане у неколико бацања новчића зове се случајна променљива (*random variable*), то јест, променљива која може да добије више од једне вредности у датим вероватноћама. На исти начин, бачена коцкица може да покаже шест лица, бројеве од један до шест, са једнаком вероватноћом. Можемо да испитујемо случајне променљиве као што су број шестица у датом броју бацања, број бацања пре прве шестице, и тако даље.

Учестала дефиниција вероватноће такође се односи на континуално мерење, као што је људска висина. На пример, претпоставимо да је просечна висина женске популације 168 цм. Онда је половина женске популације виша од 168 цм. Ако изаберемо жене случајно (тј. без особина жена које утичу на избор) у дугом низу изабраних, жене ће бити више од 168 цм. Вероватноћа да је жена виша од 168 цм је једна половина. Слично томе, ако 1/10 жена има висину изнад 180 цм, жене изабране случајно ће бити више од 180 цм са вероватноћом од 1/10. На исти начин можемо наћи да се вероватноћа висине налази између било којих датих вредности. Када меримо непрекидни квантитет увек смо ограничени методом мерења, и онда када кажемо да је жена висока 170 цм мислимо да је висина између, рецимо 169.5цм и 170.5цм, у зависности од прецизности са којом меримо. Значи оно што нас интересује је вероватноћа случајне променљиве која узима вредности између извесних граница, пре него између одређених вредности.

### 3.2 Особине вероватноће

Следеће једноставне особине произилазе из дефиниције о вероватноћи.

- Вероватноћа лежи између 0.0 и 1.0. Ако се догађај никада не деси, вероватноћа је 0.0, ако се увек дешава вероватноћа је 1.0.
- **Правило сабирања (addition rule).** Претпоставимо да се два догађаја од интереса међусобно искључују, тј. кад се један деси немогуће је да ће се и други десити. Вероватноћа да ће се десити један или други, је збир њихових појединачних вероватноћа. На пример, бачена коцкица може да се окрене на један или два, али никако на оба броја. Вероватноћа да се добије један или два =  $1/6 + 1/6 = 2/6$ .
- **Правило множења (multiplication rule).** Претпоставимо да су два догађаја од интереса независна, тј. то што знамо да се један десио нам не говори ништа о томе да ли се други дешава. Тада вероватноћа да се оба догађаја догоде је производ њихових вероватноћа. На пример, претпоставимо да бацимо два новчића. Један новчић не утиче на други, па су резултати оба бацања независни, и вероватноћа да ће испасти обе главе је  $1/2 \times 1/2 = 1/4$ . Размотримо два независна догађаја А и В. Пропорција колико пута се А догоди у дугорочном временском периоду је вероватноћа од А. Пошто су А и В независни, од оних времена када се А деси, пропорција која је једнака вероватноћи од В утицаће на то да се и В догоди. Стога је пропорција колико пута ће се А и В истовремено догодити вероватноћа од А помножена са вероватноћом од В.

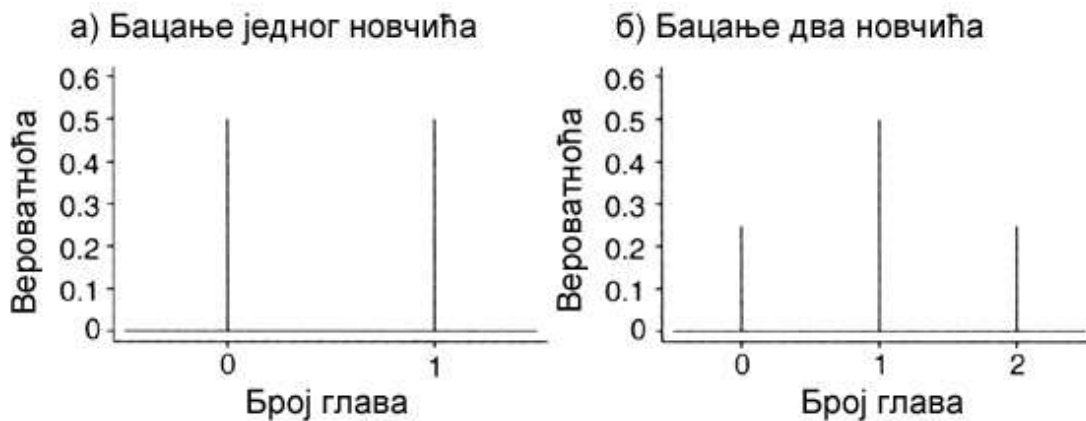
### 3.3 Расподела вероватноће и случајне променљиве

Претпоставимо да имамо скуп догађаја који се међусобно искључују и који укључује све догађаје који се могу десити. Сума њихових вероватноћа је 1.0. Скуп ових вероватноћа представља расподелу вероватноће. На пример, ако бацимо новчић, две могућности, глава или писмо, се међусобно искључују, а то су једини догађаји који се могу десити. Расподела вероватноће је:

$$\text{ВЕРОВАТНОЋА(глава)} = 1/2$$

$$\text{ВЕРОВАТНОЋА(писмо)} = 1/2$$

А сада дефинишимо променљиву, коју ћемо означити симболом  $X$ , тако да је  $X = 0$  ако новчић падне на писмо и  $X = 1$  ако новчић падне на главу.  $X$  је број глава које се покажу приликом једног бацања, а то мора да буде 0 или 1. Пре бацања не знамо шта ће  $X$  бити, али знамо да ће вероватноћа имати неку могућу вредност.  $X$  је случајна променљива (део 3.1) и расподела вероватноће је такође расподела од  $X$ . Можемо приказати ово дијаграмом, као на слици 3.1(a).



Слика 3.1 Расподела вероватноће за број глава које су приказане у бацању једног новчића и у бацању два новчића

Шта се дешава ако бацимо два новчића одједном? Сада имамо четири могућа исхода: глава и глава, глава и писмо, писмо и глава, писмо и писмо. Јасно је да су сви подједнако могући и сваки има вероватноћу  $1/4$ . Рецимо да је  $Y$  број глава.  $Y$  има три могуће вредности: 0, 1, и 2.  $Y = 0$  само онда кад добијемо писмо и писмо, и има вероватноћу  $1/4$ . Слично томе,  $Y = 2$

само кад добијемо главу и главу, па стога има вероватноћу  $1/4$ . Међутим,  $Y = 1$  или када добијемо главу и писмо, или када добијемо писмо и главу, и стога има вероватноћу  $1/4 + 1/4 = 1/2$ . Ову расподелу вероватноће можемо записати као

$$\text{ВЕРОВАТНОЋА}(Y = 0) = 1/4$$

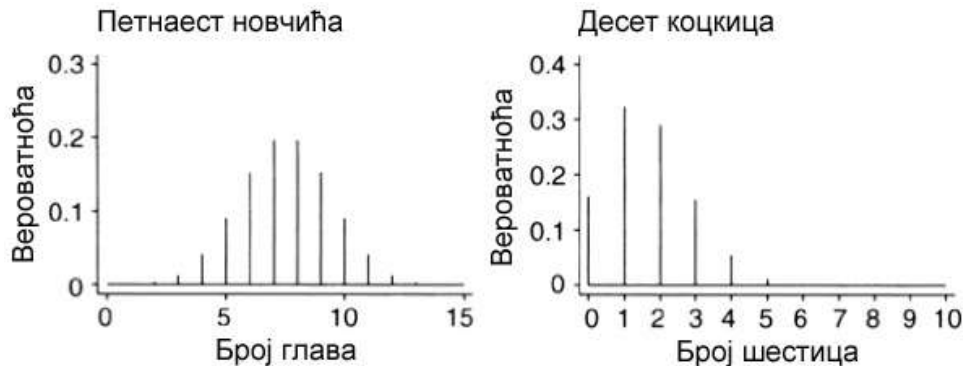
$$\text{ВЕРОВАТНОЋА}(Y = 1) = 1/2$$

$$\text{ВЕРОВАТНОЋА}(Y = 2) = 1/4$$

Расподела вероватноће од  $Y$  приказана је на слици 3.1(б).

### 3.4 Биномна расподела (Binomial distribution)

Размотрили смо расподелу вероватноће две случајне променљиве:  $X$ , број глава у једном бацању новчића, која има вредности 0 и 1, и  $Y$ , број глава у бацању два новчића, која има вредности 0, 1 или 2. Можемо повећати број новчића. Слика 3.2 приказује расподелу броја глава који је постигнут када је бачено 15 новчића. Нама не треба да вероватноћа "главе" буде 0.5; такође можемо да избројимо шестице када се баце коцкице. Слика 3.2 такође показује расподелу броја шестица која је добијена бацањем 10 коцкица. Уопштено можемо мислити о новчићу или о коцкици као о истраживањима, чији резултат може бити успех (глава или шестлица) или неуспех (писмо или један до пет). Расподеле на сликама 3.1 и 3.2 су примери Биномне расподеле, која се често појављује у медицинским апликацијама. Биномна расподела је расподела праћена бројем успеха у  $n$  независних истраживања када је вероватноћа било ког успешног појединачног истраживања  $p$ . Биномна расподела је у ствари породица расподела, а њени чланови су дефинисани вредностима  $n$  и  $p$ . Вредности које одређују ког члана из породице расподеле имамо, зову се параметри расподеле.



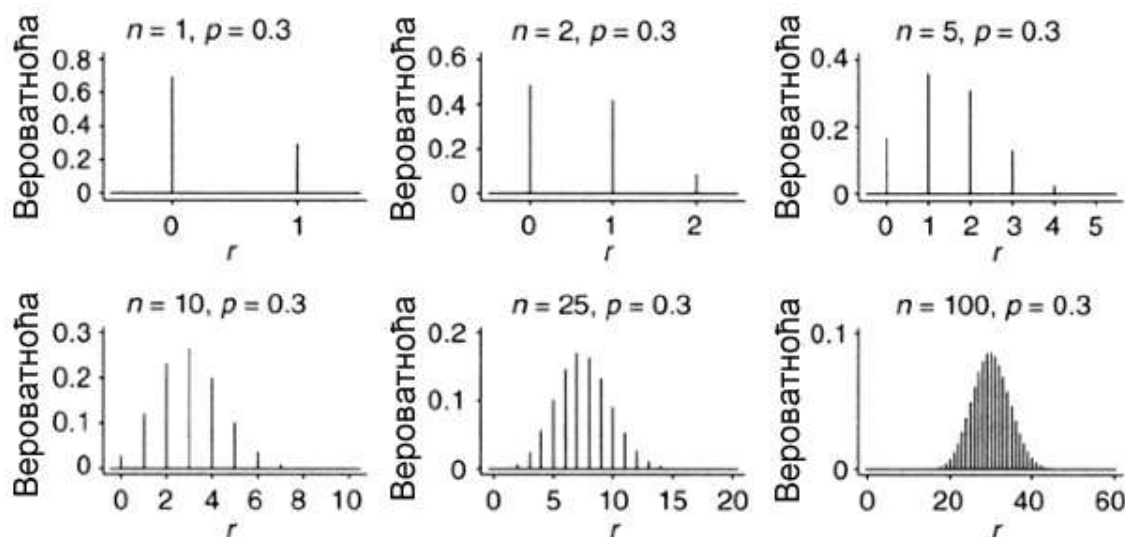
Слика 3.2 Расподела броја глава које су добијене приликом бацања 15 новчића и број шестица који је добијен бацањем 10 коцкица, примери Биномне расподеле

Случајно изабрана једноставна средства, као што су коцкице и новчићи, сама по себи имају вредност, али не за медицину. Међутим, предпоставимо да износимо преглед случајног узорка да оценимо непознату преваленцу (*prevalence*),  $p$ , болести. С обзиром да су чланови узорка одабрани случајно и независно од популације, вероватноћа да било који од одабраних субјеката има болест је  $p$ . Стога имамо серије независних истраживања, свако са вероватноћом успеха  $p$ , и бројем успеха, тј. чланови узорка, људи који имају болест ће пратити Биномну расподелу. Као што ћемо видети касније, особине Биномне расподеле нам омогућавају да кажемо колико је прецизна предвиђање добијене преваленце (део 5.4).

Можемо израчунати вероватноће Биномне расподеле тако што ћемо, нпр. направити листу на које све начине може пасти 15 новчића. Међутим, постоји  $2^{15} = 32\,768$  комбинација за 15 новчића, тако да ово и није баш практично. Уместо тога, постоји образац за вероватноћу у смислу броја бацања и вероватноће да ће глава бити та која ће пасти. Ово нам омогућава да разрешимо ове вероватноће за сваку вероватноћу успеха и било који број истраживања. У принципу, имамо  $n$  независних истраживања са вероватноћом да је резултат истраживања успешан  $p$ . Вероватноћа  $r$  успеха је

$$\text{PROB}(r \text{ uspeha}) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)}$$

где се  $n!$  зове  $n$  факторијел, и он је  $n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$ . Овај непрактичан образац настаје на овај начин. За било коју одређену серију  $r$  успеха, свака са вероватноћом  $p$ , и  $n-r$  неуспеха, сваки са вероватноћом  $1-p$ , вероватноћа да ће се серије десити је  $p^r(1-p)^{(n-r)}$ , с обзиром на то да су истраживања независна и да се примењује правило множења. Број начина у којима  $r$  ствари могу бити изабране из  $n$  ствари је  $\frac{n!}{r!(n-r)!}$ . Само се једна комбинација може десити у једном временском периоду, па имамо  $\frac{n!}{r!(n-r)!}$  начина који међусобно искључују да се десе  $r$  успеси, сваки са вероватноћом  $p^r(1-p)^{(n-r)}$ . Вероватноћа да имамо  $r$  успеха је сума ових  $\frac{n!}{r!(n-r)!}$  вероватноћа, што нам даје формула изнад. Они који се сећају развоја бинома у математици ће видети да је ово један од термина, одатле и име Биномна расподела.



Слика 3.3 Биномна расподела са различитим  $n$ ,  $p = 0.3$

Ово можемо применити код броја глава које добијемо приликом бацања два новчића. Број глава ће бити из Биномне расподеле, са  $p = 0.5$  и  $n = 2$ . Стога вероватноћа за две главе ( $r = 2$ ) је:

$$\text{PROB}(r = 2) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} = \frac{2!}{2!0!} \times 0.5^2 \times 0.5^0 = \frac{2}{2 \times 1} \times 0.25 \times 1 = 0.25$$

Запамтите да је  $0! = 1$ , и било шта степеновано на 0 је 1. Слично томе за  $r = 1$  и  $r = 0$ :

$$\text{PROB}(r = 1) = \frac{2!}{1!1!} 0.5^1 \times 0.5^1 = \frac{2}{1 \times 1} \times 0.5 \times 0.5 = 0.5$$

$$\text{PROB}(r = 0) = \frac{2!}{0!2!} 0.5^0 \times 0.5^2 = \frac{2}{1 \times 2} \times 1 \times 0.25 = 0.25$$

Ово је оно што је добијено за два новчића. Можемо користити ову расподелу кад год имамо серије истраживања које имају два могућа исхода. Ако лечимо групу пацијената, број оних пацијената који се опораве добија се из Биномне расподеле. Ако меримо крвни притисак једне групе људи, онај број који је означен као хипертензија добија се из Биномне расподеле. Слика 3.3 показује Биномну расподелу за  $p = 0.3$  и растуће вредности за  $n$ . Расподела постаје више симетрична како  $n$  расте. То конвергира (тежи) Нормалној расподели која је описана у наредном делу.

### 3.5 Средина и варијанса (Mean and variance)

Број различитих вероватноћа у Биномној расподели може бити велики и тежак за тумачење. Када је  $n$  велико, обично треба да сумирамо ове вероватноће на неки начин. Као што се расподела учесталости може описати помоћу своје *средине* (*mean*) и *варијансе* (*variance*), тако се може описати и расподела вероватноће и случајне променљиве која је у вези са њом.

Средина је просечна вредност случајне променљиве у дужем временском периоду. Она се такође зове очекивана вредност (*expected value*) или очекивање (*expectation*), и очекивање случајне променљиве  $X$  се обично означава са  $E(X)$ . На пример, размотримо број глава у бацању два новчића. Добијемо 0 глава у  $1/4$  парова новчића, тј. са вероватноћом  $1/4$ . Добијамо 1 главу у  $1/2$  парова новчића, и 2 главе у  $1/4$  парова. Просечна вредност коју треба да добијемо у дужем временском периоду добија се тако што помножимо сваку вредност са пропорцијом парова у којој се јавља и саберемо их:

$$0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 0 + \frac{1}{2} + \frac{1}{2} = 1$$

Ако наставимо да бацамо новчиће, просечан број глава по пару би био 1. Стога за сваку случајну променљиву која узима дискретне вредности средине, очекивање или очекивана вредност се налази у сумирању сваке појединачне вредности која је помножена својом вероватноћом.

Запамтите да очекивана вредност случајне променљиве не мора да буде вредност коју та променљива може да добије. На пример, за средину броја глава приликом бацања једног новчића, не добијамо ниједну главу или добијамо једну главу, свака од опција има вероватноћу пола, а очекивана вредност је  $0 \times 1/2 + 1 \times 1/2 = 1/2$ . Број глава мора бити 0 или 1, али очекивана вредност је пола, просек, који би добили у дужем временском периоду.

Варијанса случајне променљиве је просечна разлика на квадрат од средине. За број глава у бацању два новчића, 0 је 1 јединица од средине и јавља се код  $1/4$  парова новчића, 1 је 1 јединица од средине и јавља се код  $1/2$  парова и 2 је 1 јединица од средине и јавља се код  $1/4$  парова, тј. са вероватноћом  $1/4$ . Варијанса се тада налази у квадрирању тих разлика, множећи их пропорцијом броја пута у којима ће се десити разлика (вероватноћом) и сабирањем:

$$\text{Варијанса} = (0-1)^2 \times \frac{1}{4} + (1-1)^2 \times \frac{1}{2} + (2-1)^2 \times \frac{1}{4} = (-1)^2 \times \frac{1}{4} + 0^2 \times \frac{1}{2} + 1^2 \times \frac{1}{4} = \frac{1}{2}$$

Ми означавамо варијансу случајне променљиве  $X$  са  $\text{VAR}(X)$ . У математичким терминима,

$$\text{VAR}(X) = E(X^2 - E(X)^2)$$

Квадратни корен варијансе је стандардно одступање случајне променљиве или расподеле. Често користимо грчко слово  $\mu$ , које се изговара "ми", и  $\sigma$ , "сигма", за средину и стандардно одступање расподеле вероватноће. Варијанса је онда  $\sigma^2$ .

### 3.6 Карактеристике средине и варијансе

Ако додамо константу случајној променљивој, нова променљива настала на тај начин има средину једнаку оној вредности коју има оригинална променљива плус константа. Варијанса и стандардно одступање ће бити непромењени. Претпоставимо да је људска висина наша случајна променљива. Можемо додати константу висини тако што ћемо мерити висину људи који стоје на кутији. Средина висине људи плус кутија ће сада бити средина висине људи плус константна висина кутије. Кутија ипак неће изменити променљивост у висинама. На пример, разлика између највишег и најнижег човека ће бити непромењена. Константу можемо одузети тако што ћемо замолити људе да стану у рупу одређене дубине како бисмо их измерили. Ово смањује средину, али варијанса остаје иста.

Ако помножимо случајну променљиву са позитивном константом, средина и стандардно одступање су помножени константом, варијанса је помножена константом на квадрат. На пример, ако променимо јединице мерења, рецимо из инча пређемо на центиметре, сваку јединицу мерења помножимо са 2.54. Ово је исто као да помножимо средину са константом, 2.54, и стандардно одступање са константом с обзиром на то да су то, као посматрања, исте јединице. Међутим, варијанса се мери јединицама на квадрат, па је стога помножена константом на квадрат. Делјење са константом функционише на исти начин. Ако је константа

негативна, средина се множи константом и због тога мења знак. Варијанса је помножена квадратом константе, који је позитиван, стога варијанса остаје позитивна. Стандардно одступање, које је квадратни корен варијансе, је увек позитивно. Помножено је апсолутном вредношћу константе, тј. константе без негативног знака.

Ако саберемо две случајне променљиве средина суме је сума средина, и ако су две променљиве независне, варијанса суме је сума њихових варијанси. Ово можемо урадити тако што ћемо измерити висину људи који стоје на кутијама које су неједнаке висине. Средина висине људи на кутијама је средина висине људи + средина висине кутија. Варијабилност висина је такође повећана. Ово је због тога што ће неки људи који су ниски стајати на ниским кутијама, а неки високи људи ће стајати на високим кутијама. Ако две променљиве нису независне, дешава се нешто друго. Средина суме остаје сума средина, али варијанса суме није сума варијанси. Претпоставимо да су људи одлучили да стоје на кутијама, не само због статистике, већ због неке одређене сврхе. Они желе да промене сијалицу, и због тога морају да се погну на одређену висину. Сада, ниски људи морају да узму велике кутије, док високи људи то могу урадити са ниским кутијама. Као резултат тога добијамо скоро потпуно смањење у варијабилности. Са друге стране, ако кажемо највишим људима да нађу највише кутије и најнижим људима да нађу најниже кутије, варијабилност би порасла. Независност је веома битан услов.

Ако одузмемо једну случајну променљиву од друге, средина разлике је разлика између средина, и ако су две променљиве независне, варијанса разлике је сума њихових варијанси. Претпоставимо да меримо висине људи који стоје у рупама случајне дубине, а висина коју меримо је она изнад нивоа земље. Средина висина изнад земље је средина висина људи минус средина дубине рупе. Варијабилност је порасла, јер неки ниски људи стоје у дубоким рупама, а неки високи људи стоје у плитким рупама. Ако променљиве нису независне, адитивност варијанси се рашчлањује, као што је случај за две променљиве. Када људи покушају да се сакрију у рупама, они морају да пронађу рупу која је довољно дубока да би то могли да ураде, варијабилност је опет смањена.

Ефекти множења две случајне променљиве и ефекти дељења једне уз помоћ друге су много више компликовани. На срећу то ретко треба да радимо.

Сада можемо пронаћи средину и варијансу Биномне расподеле уз помоћ параметара  $n$  и  $p$ . Размотримо прво да је  $n = 1$ . Онда је расподела вероватноће:

вредност	вероватноћа
0	$1-p$
1	$p$

Стога је средина

$$0 \times (1 - p) + 1 \times p = p$$

Варијанса је

$$(0 - p)^2 \times (1 - p) + (1 - p)^2 \times p = p^2(1 - p) + p(1 - p)^2 = p(1 - p)(p + 1 - p) = p(1 - p)$$

Сада, променљива из Биномне расподеле са параметрима  $n$  и  $p$  је сума  $n$  независних променљивих из Биномне расподеле са параметрима 1 и  $p$ . Стога је њена средина сума  $n$  средина које су једнаке  $p$ , а њена варијанса је сума  $n$  варијанси једнаких  $p(1 - p)$ . Стога Биномна расподела има средину  $= np$  и варијансу  $= np(1 - p)$ . За проблема великог узорка, оне су корисније од формуле Биномне вероватноће.

Карактеристике средине и варијансе случајних променљивих омогућавају нам да пронађемо формално решење проблема степена слободе за варијансу узорка о чему смо дискутовали у делу 1 који је обрађивао сумирање. Желимо да проценимо варијансу чија очекивана вредност је варијанса популације. Очекивана вредност од  $\sum (x_i - \bar{x})^2$  се може показати као  $(n - 1) \text{VAR}(x)$  и стога делимо са  $n - 1$ , а не са  $n$ , да бисмо добили нашу процену варијансе.

### 3.7 Poisson-ова расподела

Биномна расподела је једна од многих расподела вероватноће које се користе у статистици. То је дискретна расподела, која може узети само коначан скуп могућих вредности, и то је дискретна расподела која се најчешће среће у медицинским апликацијама. Једна друга

дискретна расподела је вредна расправе у овом тренутку, Poisson-ова расподела. Иако, као и Биномна, Poisson-ова расподела настаје из једноставног модела вероватноће, укључена математика је много компликованија и биће изостављена.

Претпостављамо догађаје који се дешавају случајно и независно у времену константном брзином. **Poisson-ова расподела** је расподела праћена бројем догађаја који се дешавају у фиксном временском интервалу. Ако се догађаји дешавају са стопом  $\mu$  догађаја у јединици времена, вероватноћа  $r$  догађаја који се дешавају у јединици времена је

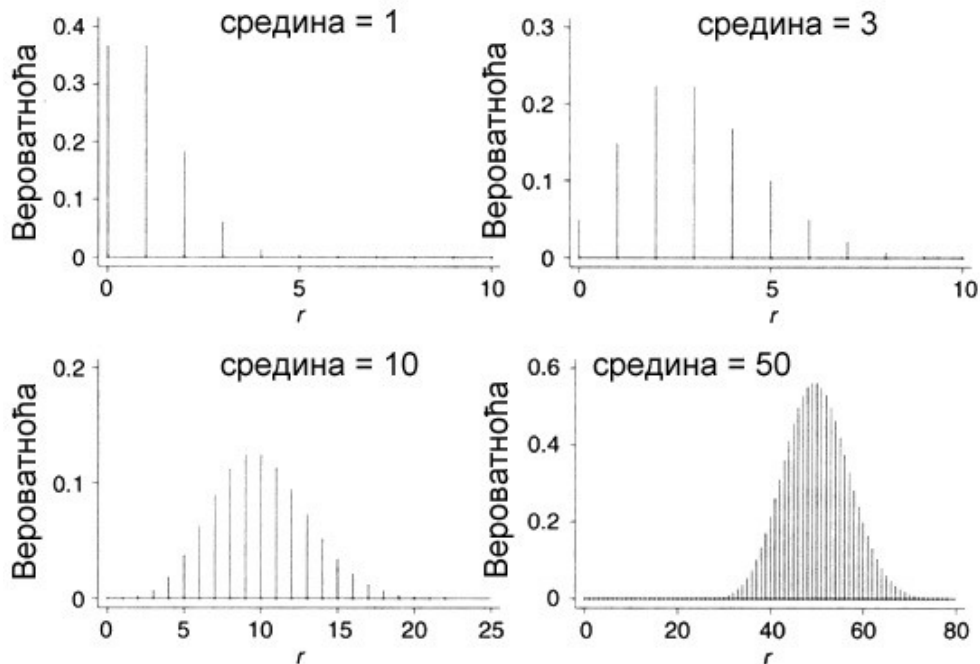
$$\frac{e^{-\mu} \mu^r}{r!}$$

где је  $e = 2.718...$ , математичка константа. Ако се догађаји дешавају случајно и независно у простору, Poisson-ова расподела даје вероватноћу за број догађаја у јединици запремине или површине.

Ретко има потребе да се користе појединачне вероватноће ове расподеле, већ је довољна њена средина и варијанса. Средина Poisson-ове расподеле за број догађаја у јединици времена је једноставно стопа,  $\mu$ . Варијанса Poisson-ове расподеле је такође једнака  $\mu$ . Тако Poisson је фамилија расподела, као што је Биномна, али само са једним параметром,  $\mu$ . Ова расподела је важна, јер се смртност од многих болести може третирати као да се дешава случајно и независно у популацији. Тако, на пример, број смртних случајева од рака плућа у једној години међу људима у радној групи, као што су рудари, биће посматрања из Poisson-ове расподеле, и ми можемо користити ово да извршимо поређења између стопа смртности (12.3).

Слика 3.4 приказује Poisson-ову расподелу за четири различите средине. Видећете да како се средина повећава Poisson-ова расподела изгледа прилично слично као Биномна расподела на слици 3.3.

#### • Poisson-ова расподела



Слика 3.4 Poisson-ове расподеле са четири различите средине

## 4. Нормална расподела (Normal distribution)

### 4.1 Вероватноћа непрекидних променљивих

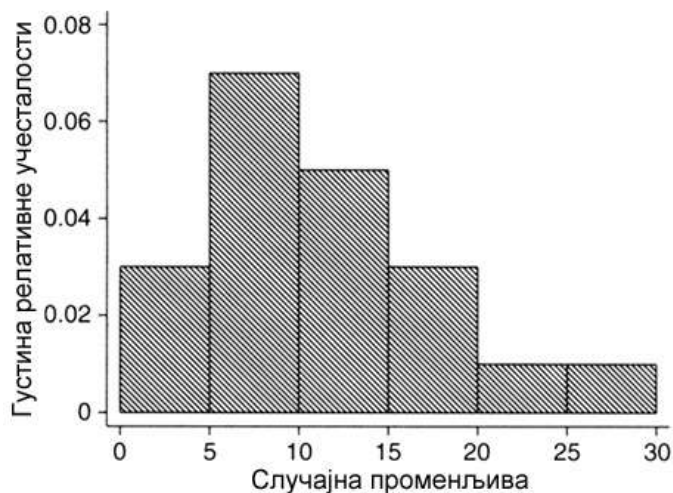
Када смо извели теорију вероватноће у дискретном случају, били смо у могућности да кажемо која је вероватноћа случајне променљиве која узима одређену вредност. Док број могућих вредности расте, вероватноћа одређене вредности се смањује. На пример, у Биномној расподели са  $p = 0.5$  и  $n = 2$ , најчешћа вредност 1, има вероватноћу 0.5. У Биномној расподели где је  $p = 0.5$  и  $n = 100$  најчешћа вредност 50, има вероватноћу 0.08. У таквим случајевима смо обично више заинтересовани за спектар вредности него за једну одређену вредност.

За непрекидну променљиву, као што је висина, скуп могућих вредности је неограничен и вероватноћа било које одређене вредности је нула (део 3.1). Заинтересовани смо за вероватноћу случајне променљиве која узима вредности између одређених граница, а не неку одређену вредност. Ако је  $p$  пропорција појединаца у популацији чије су вредности између датих граница, а ми изаберемо неку особу случајно, вероватноћа да ћемо изабрати особу која се налази у оквиру ових граница је једнака  $p$ . Ово произилази из наше дефиниције вероватноће, да је избор сваког појединца подједнако могућ. Проблем је налажење и давање вредности овој вероватноћи.

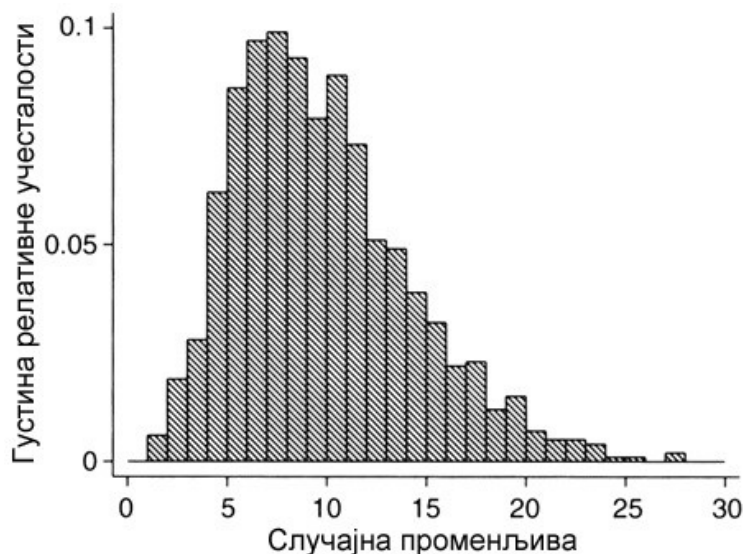
Када узмемо расподелу учесталости за узорак посматрања, рачунамо број вредности које се налазе у оквиру одређених граница (део 1.2). Ово можемо представити хистограмом као што је приказано на слици 4.1 (део 1.3). Један од начина да представимо хистограм је као густину релативне учесталости, пропорцију посматрања у интервалу по јединици  $X$  (део 1.3). Када је величина интервала 5, густина релативне учесталости је релативна учесталост подељена са 5 (Слика 4.1). Релативна учесталост у интервалу је представљена ширином интервала који је помножен са густином, чиме добијамо површину правоугаоника. Тако се релативна учесталост између било које две тачке може наћи између тачака на површини испод хистограма. На пример, да проценимо релативну учесталост између 10 и 20 на слици 4.1 имамо густину од 10 до 15 као 0.05 и између 15 и 20 као 0.03. Тако је је релативна учесталост

$$0.05 \times (15 - 10) + 0.03 \times (20 - 15) = 0.25 + 0.15 = 0.40$$

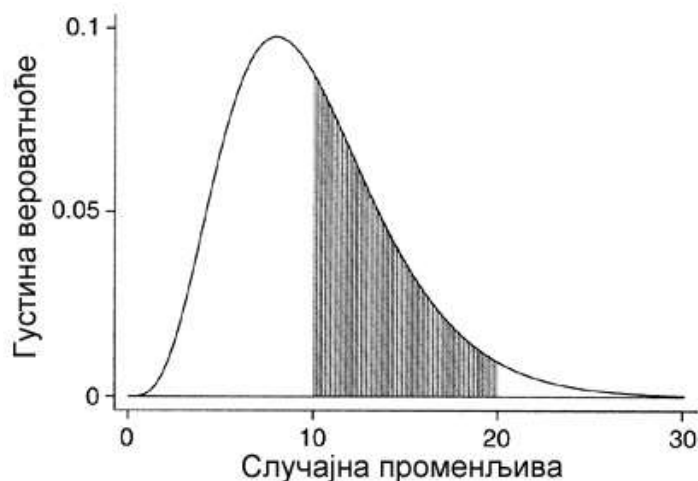
Ако узмемо већи узорак можемо користити мање интервале. Добијамо хистограм који мирније изгледа, као што је онај на слици 4.2, и како користимо све веће узорке, и све мање интервале, добијамо облик глатке криве (Слика 4.3). Како се величина узорка приближава величини популације, за коју можемо претпоставити да је доста велика, крива постаје густина релативне учесталости целе популације. Због тога можемо наћи пропорцију посматрања између било које две границе тако што ћемо наћи површину испод криве, као што је приказано на слици 4.3.



Слика 4.1 Хистограм који показује густину релативне учесталости



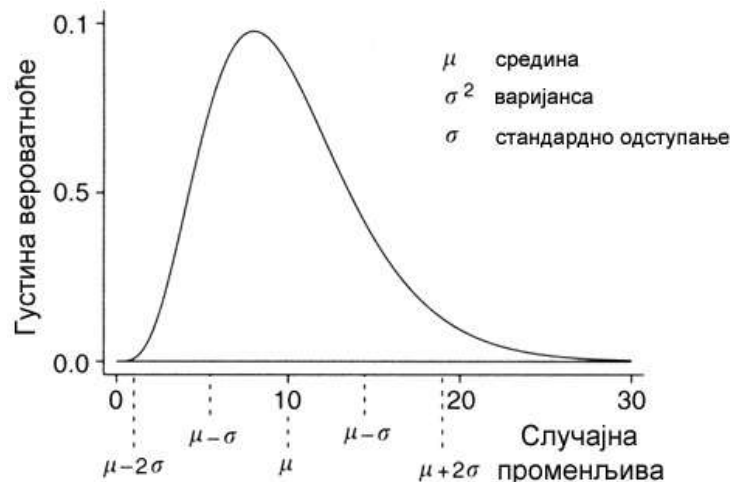
Слика 4.2 Утицај на расподелу учесталости узорка чија величина расте



Слика 4.3 Густина релативне учесталости или функција густине вероватноће, која показује вероватноћу посматрања између 10 и 20

Ако знамо једначину ове криве, можемо наћи површину испод ње (математички то радимо рачунањем интеграла, али не морамо да знамо да рачунамо интеграле како бисмо користили или разумели практичну статистику; сви интеграл који нам требају су већ израчунати и стављени у табеле). Сада, ако изаберемо неку особу случајно, вероватноћа да се  $X$  налази између било којих датих вредности је једнака пропорцији појединаца који се уклапају унутар тих граница. Тако нам расподела релативне учесталости за целу популацију даје расподелу вероватноће променљиве. Ову криву зовемо **функција густине вероватноће** (**probability density function**).

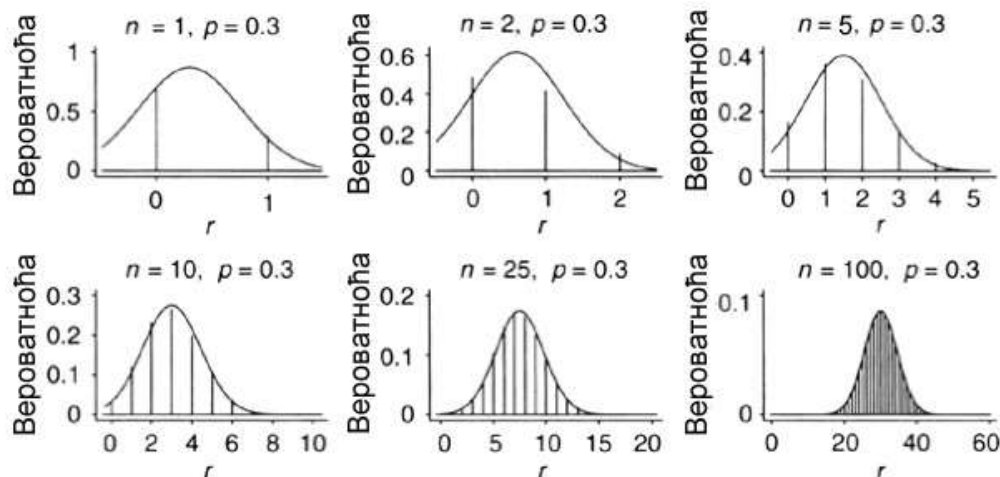
Функције густине вероватноће имају доста општих особина. На пример, цела површина испод криве мора да буде једна целина, пошто је ово укупна вероватноћа за све могуће догађаје. Непрекидне случајне променљиве имају средине, варијансе и стандардна одступања које су дефинисане на сличан начин као оне код дискретних случајних променљивих, и имају исте особине (део 3.5). Средина ће бити негде у средини криве и највећим делом површина испод криве ће бити између средине минус два стандардна одступања и средине плус два стандардна одступања (Слика 4.4).

Слика 4.4 Средина  $\mu$ , стандардно одступање  $\sigma$ , и функција густине вероватноће

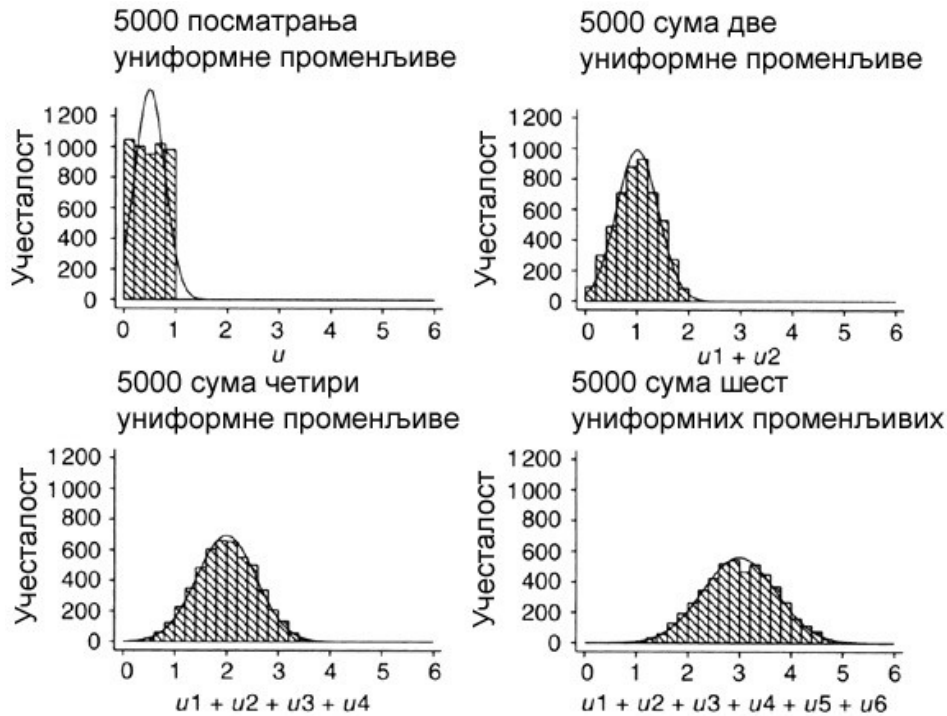
Прецизан облик криве је много теже одредити. Постоје многе вероватне функције густине вероватноће и за неке од њих можемо показати да настају из једноставних ситуација вероватноће, као што то раде Биномна и Poisson-ова расподела. Међутим већина непрекидних променљивих са којима се сусрећемо као што су висина, крвни притисак, ниво холестерола, итд., не произилазе из једноставних ситуација вероватноће. Као резултат тога имамо чињеницу да не знамо расподелу вероватноће тих мерења на теоријским основама. Као што ћемо видети, често можемо наћи стандардну расподелу чије су математичке особине познате, што одговара прикупљеним подацима и што нам омогућава да извучемо закључке о њима. Даље, како се величина узорка повећава, расподела одређених статистичких података који су израчунати из података, као што је средина, постаје независна од расподеле самих посматрања и прати једну конкретну форму расподеле, Нормалну расподелу.

## 4.2 Нормална расподела (The Normal distribution)

Нормална расподела, такође позната као Гаусова (*Gaussian*) расподела, може се оценити као основна расподела вероватноће у статистици. Реч "нормална" се овде не користи у свом основном значењу "обично или често", или у њеном значењу у медицини "без болести". Употреба ове речи се односи на њено старо значење "у складу са правилом или шаблоном", и као што ћемо видети, Нормална расподела је форма којој тежи Биномна расподела док јој параметар  $n$  расте. Не постоји импликација да већина променљивих прати Нормалну расподелу.

Слика 4.5 Биномна расподела за  $p = 0.3$  и шест различитих вредности  $n$ , са одговарајућим кривама Нормалне расподеле

Почећемо тако што ћемо посматрати Биномну расподелу док  $n$  расте. Видели смо у делу 3.4 "Биномна расподела", да док  $n$  расте, облик расподеле се мења. Највеће могуће вредности постају мање очигледне, а расподела постаје више симетрична. Ово се дешава без обзира на то колико је  $p$ . Позиција расподеле по хоризонталној оси, и њено ширење, су још увек одређени са  $p$ , али облик није. Може се нацртати глатка крива која пролази близу ових тачака. Ово је крива Нормалне расподеле, крива константне расподеле којој Биномна расподела прилази како  $n$  расте. Било која Биномна расподела може бити приближна Нормалној расподели исте средине и варијансе под условом да је  $n$  довољно велико. Слика 4.5 показује Биомну расподелу са слике 3.3 са одговарајућим кривама Нормалне расподеле. Од  $n = 10$  наредне две расподеле су веома близу. Генерално, ако обе и  $np$  и  $n(1-p)$  прелазе 5, апроксимација Биомне расподеле до Нормалне расподеле је доста добра за већину употреба у пракси.



Слика 4.6 Суме посматрања из Униформне расподеле

Биномна променљива се може гледати као сума  $n$  независних идентично распоређених случајних променљивих, које су свака настале као исход једног испитивања где су добиле вредност 1 са вероватноћом  $p$ . Уопштено, ако имамо било коју серију независних, идентично распоређених случајних променљивих, њихова сума тежи ка Нормалној расподели како број случајних променљивих расте. Ово је познато као **централна гранична теорема (central limit theorem)**. Већина мерних скупова су посматрања истих серија случајних променљивих, што је њихова веома важна особина. Из ње можемо закључити да сума или средина било које велике серије независних посматрања следи Нормалну расподелу.

На пример, размотримо **Униформу (Uniform)** или **Правоугаону расподелу (Rectangular distribution)**. Ово је расподела где су све вредности између две границе, рецимо 0 и 1, подједнако вероватне, и друге вредности нису могуће. Опажања из овога настају ако узмемо случајне цифре из табеле случајних бројева. Свако посматрање Униформне променљиве се образује помоћу оних цифри које су постављене иза децималне тачке. На дигитрону, ово је обично расподела која се добије након што се притисне RND(X) функција у BASIC језику. Слика 4.6 показује хистограм расподеле учесталости 5000 посматрања из Униформне расподеле између 0 и 1. То се доста разликује од Нормалне расподеле.

Претпоставимо да правимо нову променљиву тако што ћемо узети две Униформне променљиве и сабрати их (Слика 4.6). Облик расподеле сума две Униформне променљиве се прилично разликује од облика Униформне расподеле. Сума неће бити близу ниједне крајње вредности, овде 0 или 2, и посматрања су концентрисана у средини близу очекиване вредности. Како би одржале ниску суму, обе Униформне променљиве морају да буду ниске; а

да би направиле високу суму, обе морају да буду високе. Али добијамо суму близу средине, ако је прва променљива висока, а друга ниска, или је прва ниска а друга висока, или да су обе, и прва и друга осредње. Расподела суме две променљиве је много ближа Нормалној него Униформној расподели. Међутим, нагли прекид код 0 и код 2 не изгледа као онај који одговара Нормалној расподели. Слика 4.6 такође показује резултат који се добија сабирања четири Униформне променљиве и шест Униформних променљивих. Сличност са Нормалном расподелом се повећава како расте број који се додаје и за суму од шест Униформних променљивих је толико близу да се расподеле тешко могу разликовати.

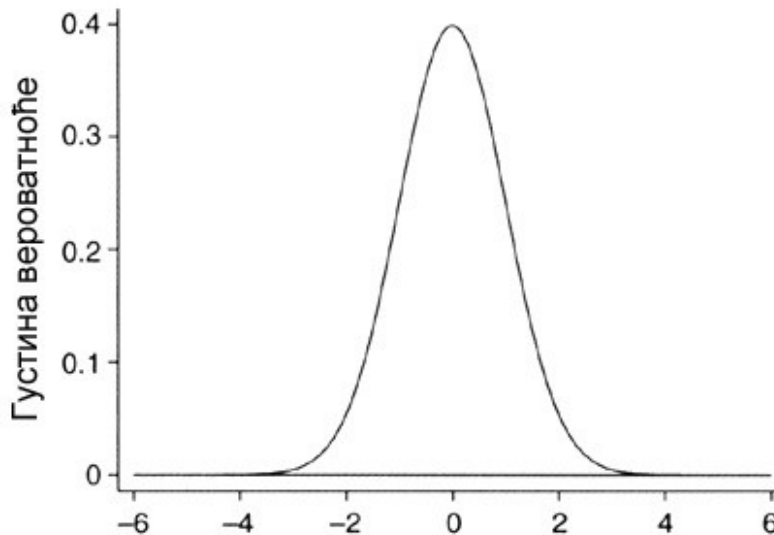
Апроксимација Биномне до Нормалне расподеле је посебан случај централне граничне теореме. Poisson-ова расподела је нешто друго. Ако узмемо скуп Poisson-ових променљивих са истом стопом и саберемо их, добићемо променљиву која је број случајних догађаја у дужем временском периоду (сума интервала за појединачне променљиве) и која је стога Poisson-ова расподела са повећаном средином. Пошто је то сума скупа независних, идентично распоређених случајних променљивих она ће тежити Нормалној расподели како средина буде расла. Зато како расте средина Poisson-ова расподела постаје приближна Нормалној расподели. За реалне потребе, ово је случај када средина пређе 10.

### 4.3 Особине Нормалне расподеле

У својој најједноставнијој форми једначина криве Нормалне расподеле, названа **Стандардизована Нормална расподела (Standard Normal distribution)**, обично се означава са  $\phi(z)$ , где је  $\phi$  грчко слово "фи":

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

где је  $\pi$  честа математичка константа која има вредност 3.14159. Онај читалац који се бави медицином може се поново уверити како не треба да користимо ову "забрањену" формулу у пракси. Стандардизована Нормална расподела има средину 0 и стандардно одступање 1, и облик као онај приказан на слици 4.7. Крива је симетрична око средине и обично описана како има облик звона (иако морам да кажем да никад нисам видео такво звоно). Можемо навести да је већина површине, тј. вероватноћа, између -1 и +1, а велика већина између -2 и +2, и скоро све су између -3 и +3.



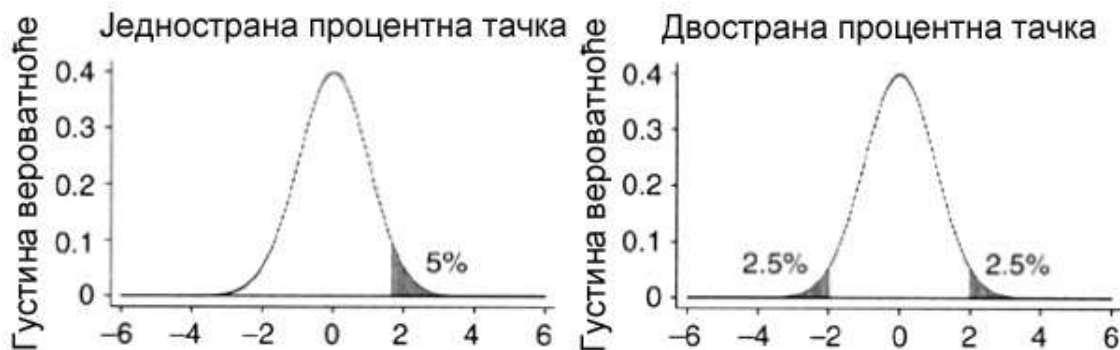
Слика 4.7 Стандардизована Нормална расподела

Иако крива Нормалне расподеле има доста значајних особина, има и једну чудну, не може да се интегрални. Другим речима, не постоји једноставна формула за вероватноћу случајне променљиве са Нормалном расподелом која лежи између датих граница. Површине испод криве могу се пронаћи нумерички, а оне су биле израчунате и убачене у табеле. Табела 4.1. показује површину испод криве густине за различите вредности Нормалне расподеле. Да

будем још прецизнији, за вредност  $z$  табела показује површину испод криве која се налази лево од  $z$ , тј. од минус бесконачно до  $z$  (Слика 4.7).

$z$	$\phi(z)$	$z$	$\phi(z)$	$z$	$\phi(z)$	$z$	$\phi(z)$	$z$	$\phi(z)$	$z$	$\phi(z)$
-3.0	0.001	-2.0	0.023	-1.0	0.159	0.0	0.500	1.0	0.841	2.0	0.977
-2.9	0.002	-1.9	0.029	-0.9	0.184	0.1	0.540	1.1	0.864	2.1	0.982
-2.8	0.003	-1.8	0.036	-0.8	0.212	0.2	0.579	1.2	0.885	2.2	0.986
-2.7	0.003	-1.7	0.045	-0.7	0.242	0.3	0.618	1.3	0.903	2.3	0.989
-2.6	0.005	-1.6	0.055	-0.6	0.274	0.4	0.655	1.4	0.919	2.4	0.992
-2.5	0.006	-1.5	0.067	-0.5	0.309	0.5	0.691	1.5	0.933	2.5	0.994
-2.4	0.008	-1.4	0.081	-0.4	0.345	0.6	0.726	1.6	0.945	2.6	0.995
-2.3	0.011	-1.3	0.097	-0.3	0.382	0.7	0.758	1.7	0.955	2.7	0.997
-2.2	0.014	-1.2	0.115	-0.2	0.421	0.8	0.788	1.8	0.964	2.8	0.997
-2.1	0.018	-1.1	0.136	-0.1	0.460	0.9	0.816	1.9	0.971	2.9	0.998
-2.0	0.023	-1.0	0.159	0.0	0.500	1.0	0.841	2.0	0.977	3.0	0.999

Тако,  $\phi$  је вероватноћа да ће случајно изабрана вредност из Стандардизоване Нормалне расподеле бити мања од  $z$ .  $\phi$  је грчко велико слово "фи". Запамтите да половина ове табеле и није потребна. Треба нам само половина за позитивно  $z$  како је  $\phi(-z) + \phi(z) = 1$ . Ово произилази из симетрије расподеле. Како бисмо нашли вероватноћу  $z$  које лежи између две вредности  $a$  и  $b$ , где је  $b > a$ , налазимо  $\phi(b) - \phi(a)$ . Како бисмо нашли вероватноћу да је  $z$  веће од  $a$ , налазимо  $1 - \phi(a)$ . Ове формуле су све примери адитивности закона вероватноће. Табела 4.1 даје само неколико вредности  $z$ , и више оних широк су доступне у литератури (Lindley and Miller 1955, Pearson and Hartley 1970). Добри статистички рачунарски програми ће израчунати ове вредности када је то потребно.



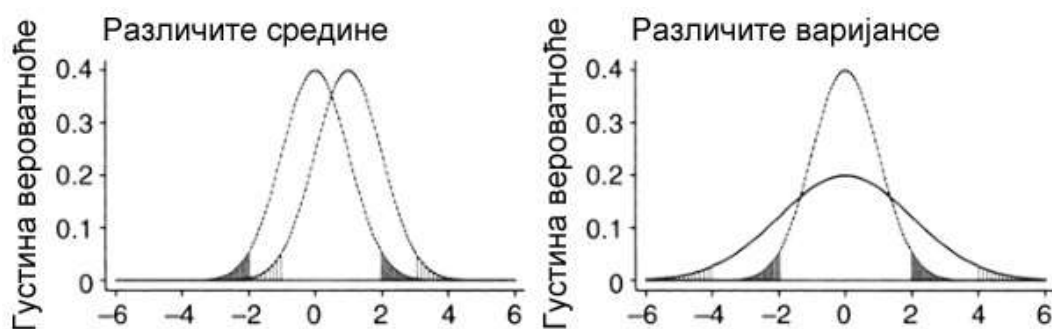
Слика 4.8 Једностране и двостране процентне тачке (5%) Стандардизоване Нормалне расподеле

Постоји други начин стављања расподеле у табелу, који користи оно што зовемо процентне тачке. **Једнострана  $P$  процентна тачка (one-sided  $P$  percentage point)** расподеле је вредност  $z$  таква да постоји вероватноћа  $P\%$  посматрања из те расподеле која је већа или једнака  $z$ , (Слика 4.8). **Двострана  $P$  процентна тачка (two-sided  $P$  percentage point)** је вредност  $z$  таква да постоји вероватноћа  $P\%$  посматрања која је већа од  $z$  или једнака  $z$ , или мања од  $z$  или једнака  $-z$  (Слика 4.8). Табела 4.2 показује обе, једностране и двостране процентне тачке Нормалне расподеле. Вероватноћа се наводи као предвиђањет јер када користимо процентне тачке обично се бавимо малим вероватноћама, као што је 0.05 или 0.01, и коришћењем обрасца предвиђањета, претварамо их у 5% и 1%, што избацује нулу испред.

Табела 4.2 Процентне тачке Нормалне расподеле

Једно-стране		Дво-стране	
$P_1$	(z)	$P_2$	(z)
50	0.00		
25	0.67	50	0.67
10	1.28	20	1.28
5	1.64	10	1.64
2.5	1.96	5	1.96
1	2.33	2	2.33
0.5	2.58	1	2.58
0.1	3.09	0.2	3.09
0.05	3.29	0.1	3.29

Табела показује вероватноћу  $P_1(z)$  да је Нормална променљива која има средину 0 и варијансу 1 већа од  $z$ , и вероватноћу  $P_2(z)$  да је Нормална променљива која има средину 0 и варијансу 1 мања од  $-z$  или већа од  $z$ .



Слика 4.9 Нормална расподела са различитим срединама и са различитим варијансама, која приказује двостране 5% тачке

До сада смо проучили Нормалну расподелу која има средину 0 и стандардно одступање 1. Ако додамо константу  $\mu$  Стандардној Нормалној променљивој, добијамо нову променљиву која има средину  $\mu$  (видети део 3.6). Слика 4.9. показује Нормалну расподелу која има средину 0 и расподелу која је добијена додавањем 1 на њу заједно са њиховим двостраним 5% тачкама. Криве су идентичне осим помака дуж осе.

Код криве која има средину 0 скоро цела вероватноћа је између  $-3$  и  $+3$ . За криву која има средину 1 вероватноћа је између  $-2$  и  $+4$ , односно између средине  $-3$  и средине  $+3$ .

Вероватноћа да постоји дати број јединица изведен из средина је исти за обе расподеле, што је такође показано 5% тачкама.

Ако узмемо Стандардну Нормалну променљиву, са стандардним одступањем 1, и помножимо је са константом  $\sigma$  добијамо нову променљиву која има стандардно одступање  $\sigma$ . Слика 4.9 показује Нормалну расподелу која има средину 0 и стандардно одступање 1 и расподелу коју смо добили множењем са 2. Криве које смо добили нису идентичне. За расподелу са стандардним одступањем 2, скоро свака вероватноћа је између -6 и +6, што је много шири интервал него онај од -3 до +3 за стандардну расподелу. Вредности -6 и +6 су -3 и +3 стандардна одступања. Можемо видети да је вероватноћа да је постојећи број онај из средине стандардног одступања исти за обе расподеле. Ово се такође види из 5% тачака, које представљају средину плус или минус 1.96 стандардних одступања у сваком случају.

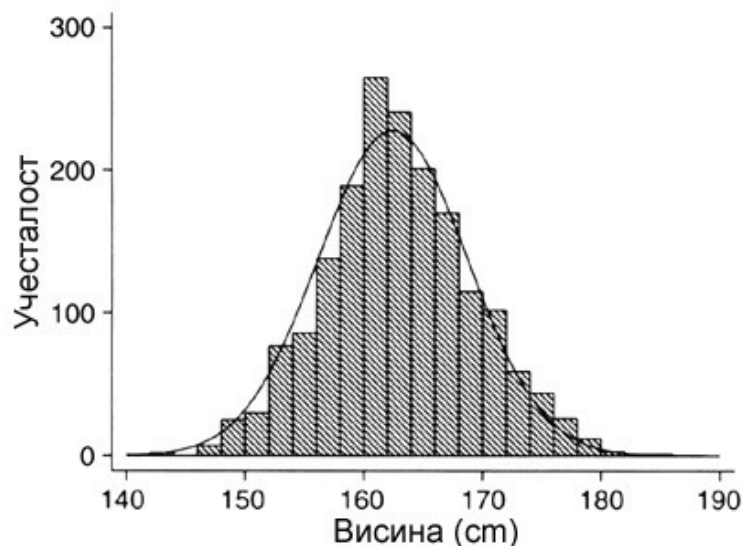
У ствари ако додамо  $\mu$  Стандардној Нормалној променљивој и помножимо је са  $\sigma$ , добијамо Нормалну расподелу средине  $\mu$ , и стандардног одступања  $\sigma$ . Табеле 4.1 и 4.2 директно одговарају томе, ако означимо са  $z$  број стандардних одступања изнад средине, а не нумеричку вредност променљиве. Тако, на пример, двостране 5% тачке Нормалне расподеле која има средину 10 и стандардно одступање 5, образују се помоћу  $10 - 1.96 \times 5 = 0.2$  и  $10 + 1.96 \times 5 = 19.8$ , вредност 1.96 се добија из табеле 4.2.

Ова особина Нормалне расподеле, да множење или додавање константи још увек даје Нормалну расподелу, није толико очигледна како се можда чини. На пример Биномна расподела то нема. Узмемо Биномну променљиву са  $n = 3$ , могућих вредности 0, 1, 2 и 3, и помножимо их са 2. Могуће вредности су сада 0, 2, 4 и 6. Биномна расподела са  $n = 6$  такође има могуће вредности 1, 3 и 5, па су расподеле различите и она коју смо добили није члан Нормалне фамилије.

Видели смо да додавањем константе променљивој са Нормалном расподелом добијамо другу променљиву која прати Нормалну расподелу. Ако саберемо заједно две променљиве са Нормалном расподелом, чак и са различитим срединама и варијансама, сума прати Нормалну расподелу. Разлика између две променљиве са Нормалном расподелом такође прати Нормалну расподелу.

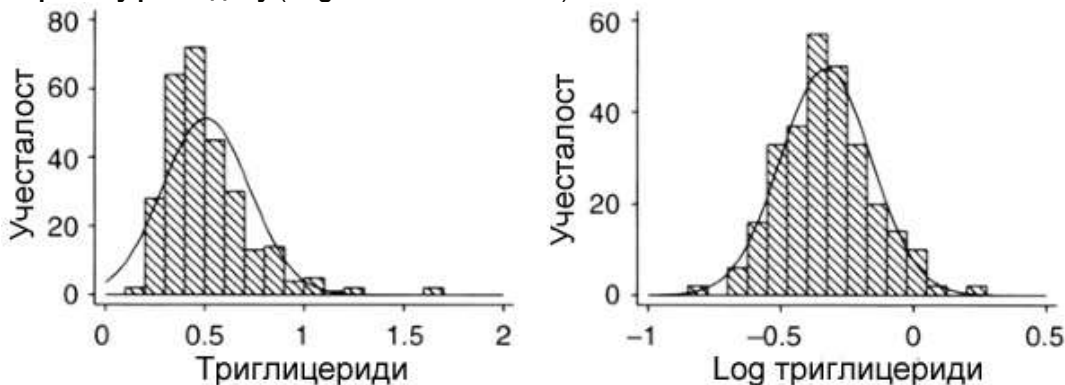
#### 4.4 Променљиве које прате Нормалну расподелу

До сада смо расправљали о томе како Нормална расподела настаје из узорак као сума или граница других расподела. Међутим, многе променљиве које постоје као такве, као што је људска висина, чини се да верно прате Нормалну расподелу. Могли би да очекујемо да ће се то десити када би променљива била резултат сабирања варијација из одређеног броја других извора. Процес који је показан помоћу централне граничне теореме може добро произвести резултат сличан ономе из Нормалне расподеле. Слика 4.10 показује расподелу висине у узорку трудних жена, и одговарајућу криву Нормалне расподеле. Поклапање са Нормалном расподелом је веома добро.



Слика 4.10 Расподела висине у узорку од 1749 трудних жена (подаци из Brooke и други 1989)

Ако је променљива коју ми меримо резултат множења неколико различитих извора варијације нећемо очекивати да резултат буде Нормалан из особина које су описане у делу 4.2, а које су све базиране на сабирању променљивих. Међутим, ако узмемо логаритамску ( $\log$ ) трансформацију такве променљиве добићемо нову променљиву која је сума неколико различитих извора варијације и која може да има Нормалну расподелу. Овај се процес често дешава са квантитетима који су део метаболичких процеса, стопа код које реакција може да се деси у зависности од концентрације других једињења. На пример, многа мерења крвних састојака то показују. Слика 4.11 показује расподелу серума триглицерида који је измерен у крви из постељице 282 бебе (Табела 1.8). Расподела је потпуно искривљена и не личи на криву Нормалне расподеле. Међутим, када узмемо логаритам концентрације триглицерида, имамо задивљујуће добро поклапање са Нормалном расподелом (Слика 4.11). Ако логаритам случајне променљиве прати Нормалну расподелу, сама случајна променљива прати **Логнормалну расподелу (Lognormal distribution)**.



Слика 4.11 Расподела серума триглицерида (табела 1.8) и  $\log_{10}$  триглицерида у крви из постељице 282 беба, која одговара кривама Нормалне расподеле

Често желимо да променимо скалу на којој анализирамо податке како бисмо добили Нормалну расподелу. Овај процес анализе зовемо математичка функција података, пре него **трансформација (transformation)** података. Логаритам је трансформација која се најчешће користи, квадратни корен и реципрочна вредност су друге трансформације (видети такође део 7.4). За један узорак, трансформација нам омогућава да користимо Нормалну расподелу да оценимо центиле (део 1.5). На пример, често желимо да одредимо 2.5-ти и 97.5-ти центил, који заједно чине 95% посматрања. За Нормалну расподелу, ово се може израчунати преко  $\bar{x} \pm 1.96s$ . Можемо да трансформишемо податке тако да расподела буде Нормална, израчунамо центиле, и онда трансформишемо назад у оригиналну скалу.

Размотримо податке о триглицеридима са слике 4.11 и табеле 1.8. Средина је 0.51, а стандардно одступање је 0.22. Средина за  $\log_{10}$  трансформисаних података је -0.33, и стандардно одступање је 0.17. Шта се дешава ако извршимо трансформацију уназад помоћу антилогаритма (*antilog*)? Антилогаритам се користи да означи функцију инверзну логаритму (експоненцијална функција, односно степеновање). Пише се као  $\text{antilog}(n)$  и значи исто што и  $b^n$ . Према томе за средину, добијамо  $10^{-0.33} = 0.47$ . Ово је мање од средине сировог (непретвореног) податка. Антилогаритам средине логаритма није исто што и нетрансформисана аритметичка средина. У ствари, ово је **геометријска средина (geometric mean)**, која је  $n$ -ти корен производа посматрања.

Ако саберемо логаритме посматрања добијамо логаритам њихових производа. Ако означимо са  $a$  и  $b$  посматрања, имамо да је  $\log(a) + \log(b) = \log(axb)$ .

Ако одузмемо логаритме посматрања добијамо логаритам њихових количника. Ако су са  $a$  и  $b$  означена посматрања, имамо да је  $\log(a) - \log(b) = \log(a/b)$ .

Ако помножимо логаритам броја са другим бројем, добијамо логаритам првог броја подигнут на степен другог броја. Ако означимо са  $a$  и  $b$  први и други број, имамо да је  $\log(a) \times b = \log(a^b)$ .

Па ако поделимо логаритам са  $n$ , добијамо логаритам  $n$ -тог корена. Тако имамо да је  $\log(a)/n = \log(\sqrt[n]{a})$

Зато је средина логаритама, логаритам геометријске средине. У трансформацији уназад, реципрочна трансформација такође производи средину посебног имена, **хармоничну средину (harmonic mean)**, реципрочну вредност средине реципрочних вредности.

Геометријска средина је у оригиналним јединицама. Ако се триглицерид мери у ммол/литру, логаритам једног посматрања је логаритам мерења у ммол/литру. Сума  $n$  логаритама је логаритам производа  $n$  мерења у ммол/литру и то је логаритам мерења у ммол/литру до  $n$ -тог. Стога је  $n$ -ти корен логаритам броја у ммол/литру, и антилогаритам је враћен назад у оригиналне јединице, у ммол/литру.

Међутим, антилогаритам стандардног одступања није мерен у оригиналним јединицама. Да израчунамо стандардно одступање узимамо разлике између сваког логаритма посматрања и одузимамо логаритам геометријске средине, користећи обично формулу  $\sum (x_i - \bar{x})^2 / n - 1$  (део 1.8). Тако имамо разлику између логаритама два броја од којих је сваки измерен у ммол/литру, дајући логаритам њиховог количника што је логаритам чистог броја који нема димензију. Било би потпуно исто да ли су триглицериди измерени у ммол/литру или у mg/100ml. Не можемо да трансформишемо стандардно одступање назад у оригиналну скалу.

Ако желимо да користимо стандардно одступање, најлакше је да сва мерења урадимо на трансформисаној скали и трансформишемо назад, ако је то потребно, на крају. На пример, 2.5-ти центил на логаритамској скали је  $-0.33 - 1.96 \times 0.17 = -0.66$  и 97.5-ти центил је  $-0.33 + 1.96 \times 0.17 = 0.00$ . Да би добили ово, узимамо логаритам нечега у ммол/литру и додајемо или одузимамо логаритам чистог броја (тј. помножен на природној скали), тако да још имамо логаритам нечега у ммол/литру. Да бисмо се вратили назад на оригиналну скалу, изводимо антилогаритам да добијемо да је 2.5-ти центил = 0.22 и 97.5-ти центил = 1.00 ммол/литру.

#### 4.5 Нормални графикон (The Normal plot)

Многе статистичке методе се могу користити само ако посматрања прате Нормалну расподелу (видети делове 7 и 8 који обрађују "Значење средине малих вредности" и "Регресију и корелацију"). Постоји неколико начина да се истражи да ли посматрања прате Нормалну расподелу. Са великим узорком можемо истражити хистограм да видимо да ли он изгледа као крива Нормалне расподеле. Ово не функционише добро са малим узорком, и бољи метод је **Нормални графикон (Normal plot)**. То је графички метод који се примењује тако што узмемо обичан графички папир и табелу Нормалне расподеле, са специјално одштампаним папиром Нормалне вероватноће, или још лакше, ако користимо рачунар. Било који добар статистички стандардни програм ће дати Нормалне графиконе; ако неће, онда програм није добар. Метода Нормалног графикона може се користити да се испита Нормална претпоставка у узорцима било које величине, и веома је корисна да се користи као провера када се користе методе као што су  $t$  методе расподеле, које су описане у делу 7.

Нормални графикон је графикон расподеле кумулативне учесталости података према расподели кумулативне учесталости за Нормалну расподелу. Прво поређамо податке од најнижих до највиших. За сваку уређено посматрање налазимо очекивану вредност посматрања ако су подаци пратили Стандардизовану Нормалну расподелу. Постоји неколико апроксимативних формула за ово. Ми ћемо пратити Armitage-ову и Berry-јеву (1994) и употребити за  $i$ -ту опсервацију  $z$  где је  $\phi(z) = (i - 0.5) / n$ . Неке књиге и програми користе  $\phi(z) = i / (n + 1)$  и постоје неке друге сложеније формуле. Нема неке велике разлике која се користи. Из табеле Нормалне расподеле налазимо вредности од  $z$  које одговарају  $\phi(z) = 0.5/n, 1.5/n$ , итд. (Табели 4.1 недостају детаљи да би се користила у пракси, али ће послужити за илустрацију.) За 5 тачака, на пример, имамо  $\phi(z) = 0.1, 0.3, 0.5, 0.7, 0.9$  и  $z = -1.3, -0.5, 0, 0.5, 1.3$ . Ово су тачке Стандардизоване Нормалне расподеле које одговарају прикупљеним подацима. Ако су прикупљени подаци из Нормалне расподеле где је средина  $\mu$  и варијанса  $\sigma^2$ , добијена тачка треба да је једнака  $\sigma(z) + \mu$ , где је  $z$  одговарајућа тачка Стандардизоване Нормалне расподеле. Ако ставимо у графикон Стандардне Нормалне тачке према добијеним вредностима, треба да добијемо нешто слично правој линији. Једначину ове линије можемо написати као  $\sigma(z) + \mu = x$ , где је  $x$  посматрана променљива и  $z$  одговарајући квантил Стандардизоване Нормалне расподеле. Ово можемо написати као

$$z = \frac{x - \mu}{\sigma}$$

које пролази кроз тачку која је означена као  $(\mu, 0)$  и има нагиб  $1/\sigma$  (погледати део 8.1). Ако подаци нису из Нормалне расподеле нећемо добити равну линију, него неку врсту криве. Зато

што у графикону цртамо квантиле расподеле учесталости коју посматрамо наспрам одговарајућих квантила теоријске (овде Нормалне) расподеле, ово се још зове **квантил-квантил графикон (quantile-quantile plot)**, или **q-q графикон (q-q plot)**.

Табела 4.3 Витамин D измерен у крви 26 здравих мушкараца, подаци из Hickish et al. (1989)

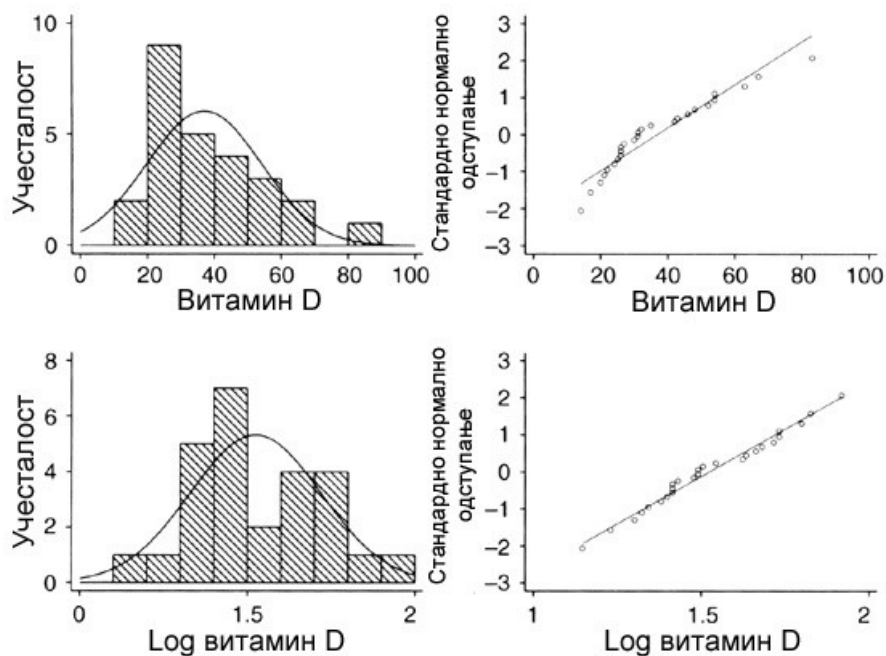
14	25	30	42	54
17	26	31	43	54
20	26	31	46	63
21	26	32	48	67
22	27	35	52	83
24				

Табела 4.4 Прорачун Нормалног графикона за податке о витамину D

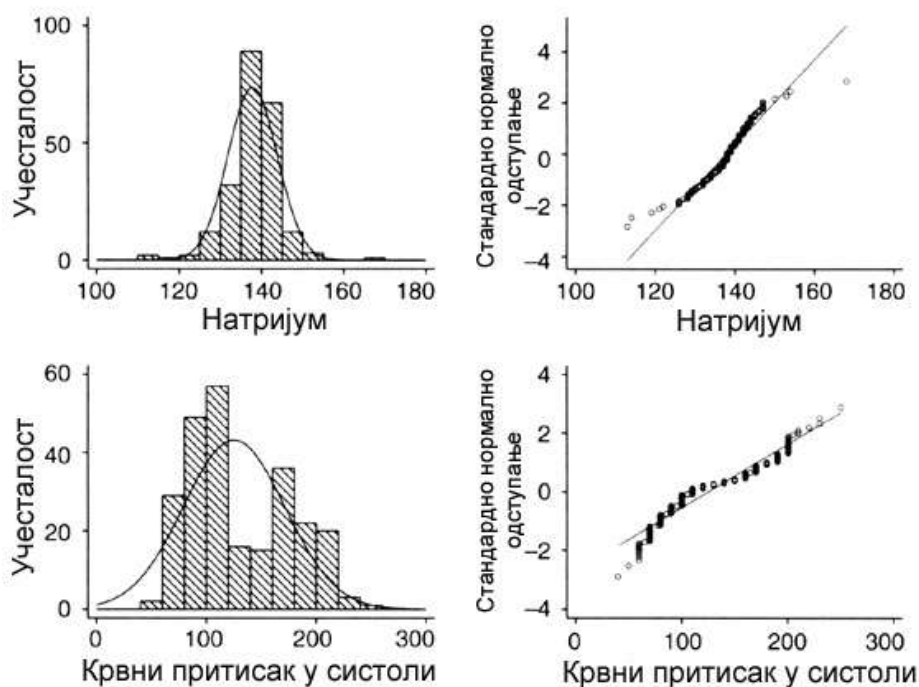
$i$	Vit D	$\phi(z)$	$z$	$i$	Vit D	$\phi(z)$	$z$
1	14	0.019	-2.07	14	31	0.519	0.05
2	17	0.058	-1.57	15	32	0.558	0.15
3	20	0.096	-1.30	16	35	0.596	0.24
4	21	0.135	-1.10	17	42	0.635	0.34
5	22	0.173	-0.94	18	43	0.673	0.45
6	24	0.212	-0.80	19	46	0.712	0.56
7	25	0.250	-0.67	20	48	0.750	0.67
8	26	0.288	-0.56	21	52	0.788	0.80
9	26	0.327	-0.45	22	54	0.827	0.94
10	26	0.365	-0.34	23	54	0.865	1.10
11	27	0.404	-0.24	24	63	0.904	1.30
12	30	0.442	-0.15	25	67	0.942	1.57
13	31	0.481	-0.05	26	83	0.981	2.07

$$\phi(z) = (i - 0.5) / 26$$

Табела 4.3 показује нивое витамина D измерене у крви 26 здравих мушкараца. Прорачун Нормалног графикона је приказан у табели 4.4. Запамтите да су  $\phi(z) = (i - 0.5) / 26$  и  $z$  симетрични, друга половина је прва половина са супротним знаком. Вредност Стандардног Нормалног одступања,  $z$ , може бити пронађена интерполацијом у табели 4.1, коришћењем пуније табеле, или преко рачунара.

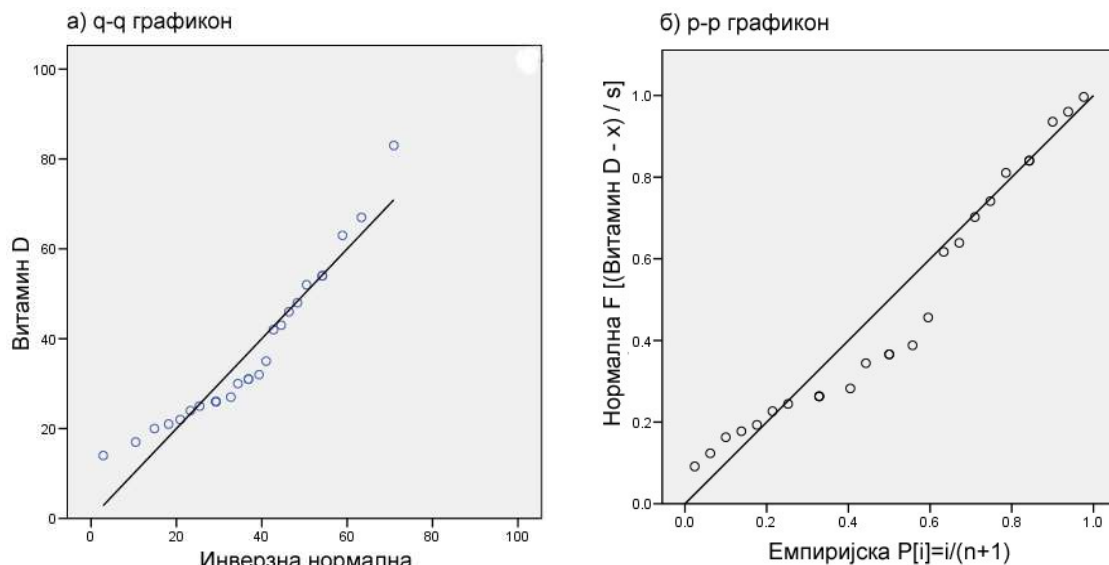


Слика 4.12 Ниво витамина D у крви и  $\log_{10}$  витамина D за 26 нормалних мушкараца, са Нормалним графиконима



Слика 4.13 Натријум у крви и крвни притисак у систоли измерен код 250 пацијената у Јединици Интезивне Терапије болнице St. George, са Нормалним графиконима (Freidland и други 1996)

Слика 4.12 показује хистограм и Нормални графикон ових података. Расподела је закошена и Нормални графикон показује јасну криву. Слика 4.12 такође показује податке о витамину D након логаритамске трансформације. Прилично је лако направити Нормални графикон пошто је одговарајуће Стандардно Нормално одступање,  $z$ , непромењено. Треба само да израчунамо логаритам посматрања и да нацртамо графикон поново. Нормални графикон трансформисаних података се добро прилагођава теоријској линији, и указује да је расподела логаритма за ниво витамина D близу Нормалне. Један лук у Нормалном графикону указује на асиметричност. Дупла крива указује да су оба краја расподеле различита у односу на Нормалну расподелу, обично су предугачки, и многе криве указују на то да је расподела бимодална (Слика 4.13). Наравно, када је узорак мали појавиће се неке случајне флуктуације.



Слика 4.14 Варијације Нормалног графикона за податке о витамину D

Постоји неколико различитих начина да се прикаже Нормални графикон. Неки програми стављају расподелу података на вертикалну осу, а теоријску Нормалну расподелу на хоризонталну осу, што утиче на правац кретања криве. Неки праве графикон за теоријску Нормалну расподелу са средином  $\bar{x}$ , средином узорка, и стандардним одступањем  $s$ , стандардним одступањем узорка. Ово је урађено израчунавањем  $\bar{x} + sz$ . Слика 4.14 (а) показује ове две особине. Нормални графикон урађен у програму SPSS помоћу команде "Q-Q Plots". Права линија је линија једнакости. Графикон је идентичан другом графикону са слике 4.12, осим промена у скали и замени оса. Незнатна варијација је **графикон стандардизоване Нормалне вероватноће (standardized Normal probability plot)**, или **p-p графикон (p-p plot)**, где стандардизујемо посматрања до средине нула и стандардног одступања један,  $y = (x - \bar{x})/s$ , и стављамо у графикон кумулативе Нормалне вероватноће,  $\Phi(y)$ , насупрот  $(i - 0.5) / n$  или  $i / (n + 1)$ , (Слика 4.14(б)), која је добијена програмом SPSS помоћу команде "P-P Plots". Постоји мала разлика између слика 4.14(а) и 4.14(б) и верзије квантила и вероватноће Нормалног графикона треба да се тумаче на исти начин.

## 5 Предвиђање

### 5.1 Расподеле узорака (Sampling distributions)

У овом поглављу ћемо видети како нам теорија вероватноће омогућава да предвидимо (проценимо) квантитете у популацији, и одредимо прецизност ових предвиђање. Прво ћемо размотрити шта се дешава када употребимо поновљене узорке из једне популације.

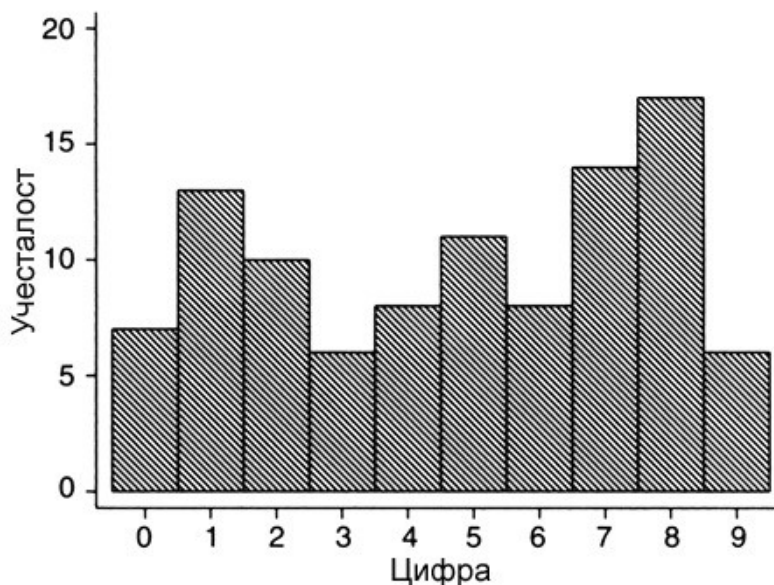
Табела 5.1 показује скуп 100 случајних цифара које можемо да користимо као популацију за експеримент узорка (узорковања) (*sampling experiment*). Расподела бројева у овој популацији је приказана на слици 5.1. Средина популације је 4.7, а стандардно одступање је 2.9.

Табела 5.1 Популација од 100 случајних цифара за експеримент узорка (узорковања)

9	1	0	7	5	6	9	5	8	8	1	0	5	7	6	5	0	2	1	2
1	8	8	8	5	2	4	8	3	1	6	5	5	7	4	1	7	3	3	3
2	8	1	8	5	8	4	0	1	9	2	1	6	9	4	4	7	6	1	7
1	9	7	9	7	2	7	7	0	8	1	6	3	8	0	5	7	4	8	6
7	0	2	8	8	7	2	5	4	1	8	6	8	3	5	8	2	7	2	4

Експеримент узорка (*sampling experiment*) се ради тако што се користи одговарајућа метода случајног узорка како би се искористили поновљени узорци популације. У овом случају, децималне коцкице су послужиле као одговарајући метод. Изабран је узорак величине четири: 6, 4, 6 и 1. Средина је израчуната као:  $17/4 = 4.25$ . Ово је поновљено како би се користио други узорак од четири броја: 7, 8, 1 и 8. Његова средина је 6.00. Ова процедура узорка је рађена свеукупно 20 пута, како би се добили узорци и њихове средине приказане у табели 5.2.

Ове средине узорака нису све исте. Оне показују случајну променљиву. Када бисмо могли да искористимо свих 3 921 225 могућих узорака за величину 4 и израчунамо њихове средине, ове средине саме би формирале расподелу. Наших 20 средина узорака су саме по себи узорци из ове расподеле. Расподела свих могућих средина узорака се зове **расподела узорка (sampling distribution)** средине. Уопштено говорећи, расподела узорка било које статистике је расподела вредности статистике која би се развила из свих могућих узорака.



Слика 5.1 Расподела популације из табеле 5.1

Табела 5.2 Случајни узорци коришћени у експерименту узорка

Узорак	6	7	7	1	5	5	4	7	2	8
	4	8	9	8	2	5	2	4	8	1
	6	1	2	8	9	7	7	0	7	2
	1	8	7	4	5	8	6	1	7	0
Средина	4.25	6.00	6.25	5.25	5.25	6.25	4.75	3.00	6.00	2.75
Узорак	7	7	2	8	3	4	5	4	4	7
	8	3	5	0	7	8	5	3	5	4
	7	8	0	7	4	7	8	1	8	6
	2	7	8	7	8	7	3	6	2	3
Средина	6.00	6.25	3.75	5.50	5.50	6.50	5.25	3.50	4.75	5.00

## 5.2 Стандардна грешка средине узорка (Standard error of a sample mean)

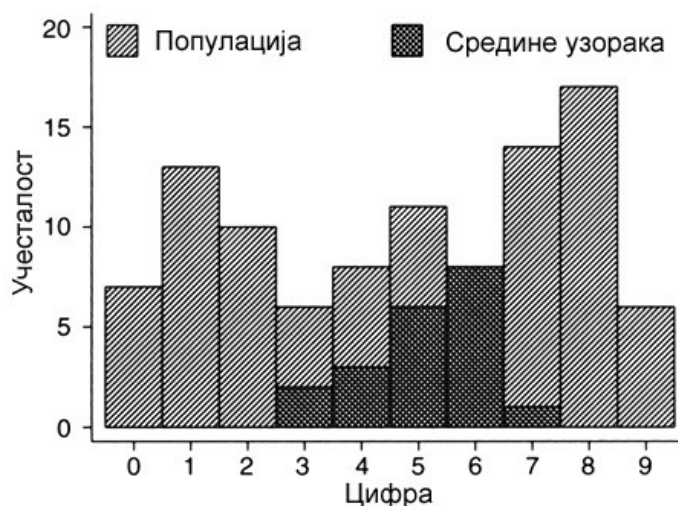
За кратко ћемо размотрити расподелу узорка само за средину. Пошто је наш узорак од 20 средина случајан узорак из средине, можемо ово користити да предвидимо неке параметре расподеле. Двадесет средина имају своју средину и стандардно одступање. Средина је 5.1 и стандардно одступање је 1.1. Сада средина целе популације је 4.7, што је близу средине узорка. Али стандардно одступање целе популације је 2.9, што је знатно веће од стандардног одступања узорка.

4,25	6,00	6,25	5,25	5,25	6,25	4,75	3,00	6,00	2,75	49,75
6,00	6,25	3,75	5,50	5,50	6,50	5,25	3,50	4,75	5,00	52,00
Средина узорка:										5,09

Ако цртамо хистограм за средине узорка (Слика 5.2), видимо да су центар расподеле узорка и расподела популације родитеља исти, али је растурање расподеле узорка доста мање.

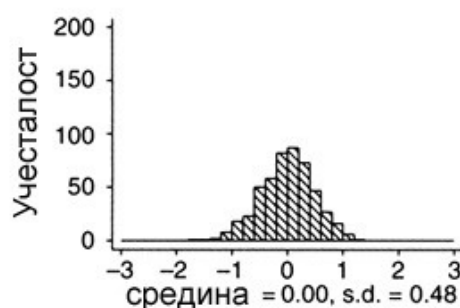
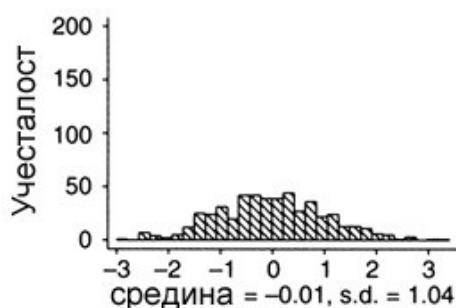
Још један експеримент узорка, на већој скали ће илустровати ово у наставку. Овог пута наша родитељска расподела ће бити Нормална расподела са средином 0 и стандардним одступањем 1. Слика 5.3 показује расподелу случајног узорка 500 посматрања из ове расподеле. Слика 5.3 такође показује расподелу средина из 500 случајних узорка величине 4 из ове популације, исте величине узорка као на слици 5.2. Слика 5.3 такође показује расподелу 500 средина величине 9 и величине 16. У све четири расподеле средине су близу 0, средине родитељске расподеле. Али стандардна одступања нису иста. Она су у ствари апроксимативно 1 (родитељска расподела); 1/2 (средина од 4), 1/3 (средина од 9) и 1/4 (средина од 16). У ствари расподела средине узорка има стандардно одступање  $\sigma/\sqrt{n}$  или  $\sqrt{\sigma^2/n}$ , где је  $\sigma$  стандардно одступање родитељске расподеле, а  $n$  је величина узорка. Средина расподеле узорка је

једнака средини родитељске расподеле. Стварна, као супротна од симулиране, расподела средине четири посматрања из Нормалне расподеле је приказана на слици 5.4.



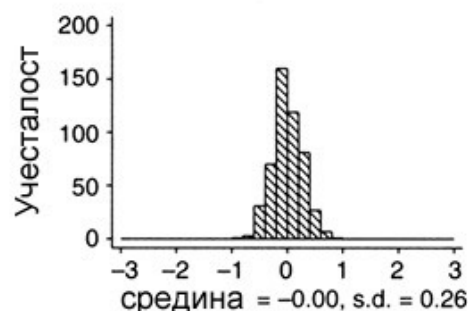
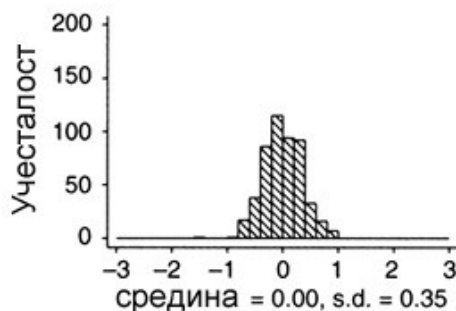
Слика 5.2 Расподела популације из табеле 5.1 и средине узорка из табеле 5.2

500 Стандардних Нормалних посматрања    500 средина 4 Нормална посматрања



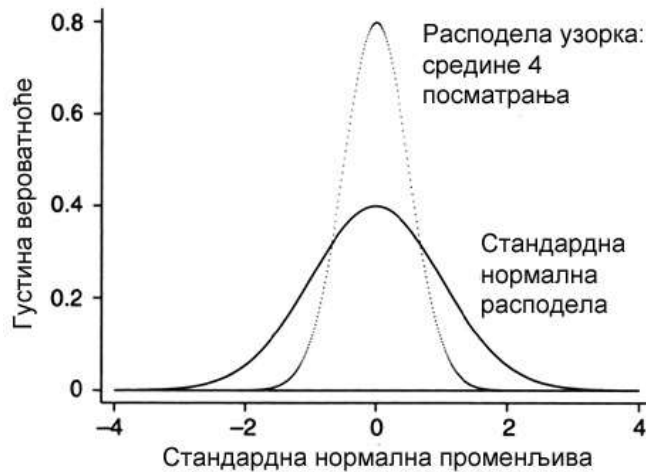
500 средина 9 Нормалних посматрања

500 средина 16 Нормалних посматрања



Слика 5.3 Узорци средина из Стандардне Нормалне променљиве

Средина узорка је предвиђање средине популације. Стандардно одступање њене расподеле узорка зове се **стандардна грешка (standard error - se)** предвиђања. Она обезбеђује меру колико далеко је предвиђање од праве вредности. У већини предвиђања, вероватно је да ће предвиђање бити у оквиру једне стандардне грешке праве средине и вероватно неће бити удаљена од ње више од две стандардне грешке. Погледајмо ово прецизније у наредном делу.



Слика 5.4 Расподела средине узорка 4 посматрања из Стандардне Нормалне расподеле

У скоро свим практичним ситуацијама не знамо праву вредност варијансе популације  $\sigma^2$  већ само предвиђање  $s^2$  (део 1.7). Ово можемо да искористимо да предвидимо стандардну грешку помоћу  $s/\sqrt{n}$ . Ова предвиђање се такође узима као стандардна грешка средине. Обично се јасно види из контекста да ли је стандардна грешка права вредност или она која је предвиђена из података.

Када је величина узорка  $n$  велика, расподела узорка од  $\bar{x}$  тежи ка Нормалној расподели. Такође можемо претпоставити да је  $s^2$  добра предвиђање од  $\sigma^2$ . Тако за велико  $n$ ,  $\bar{x}$  је у ствари, посматрање из Нормалне расподеле са средином  $\mu$  и стандардним одступањем предвиђеним преко  $s/\sqrt{n}$ . Тако са вероватноћом 0.95,  $\bar{x}$  је између два, или да будемо прецизнији унутар 1.96 стандардних грешака  $\mu$ . Са малим узорцима не можемо претпоставити да ли је Нормална расподела добра или још важније да ли је  $s^2$  добра предвиђање од  $\sigma^2$ . О овоме ћемо расправљати у делу 7 који обрађује "Значење средине малих вредности".

На пример, размотримо поново 57 FEV1 мерења из табеле 1.4. Имамо да збир 57 FEV1 је 231.51, а из тога је средина је  $\bar{x} = 231.51/57 = 4.062$  литара,  $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = 0.449174$  и  $s = \sqrt{s^2} = 0.67$  литара. Тада стандардна грешка од  $\bar{x}$  је:

$$\sqrt{s^2/n} = \sqrt{0.449174/57} = 0.08877 = 0.089 \text{ литара}$$

Најбоља предвиђање средине FEV1 у популацији је стога 4.062 литара са стандардном грешком 0.089 литара. Средина и стандардна грешка се често пишу као  $4.062 \pm 0.089$ . Ово лако може да буде варљиво, пошто тачна вредност може бити до две стандардне грешке од средине са могућом вероватноћом. Ова пракса није препоручљива.

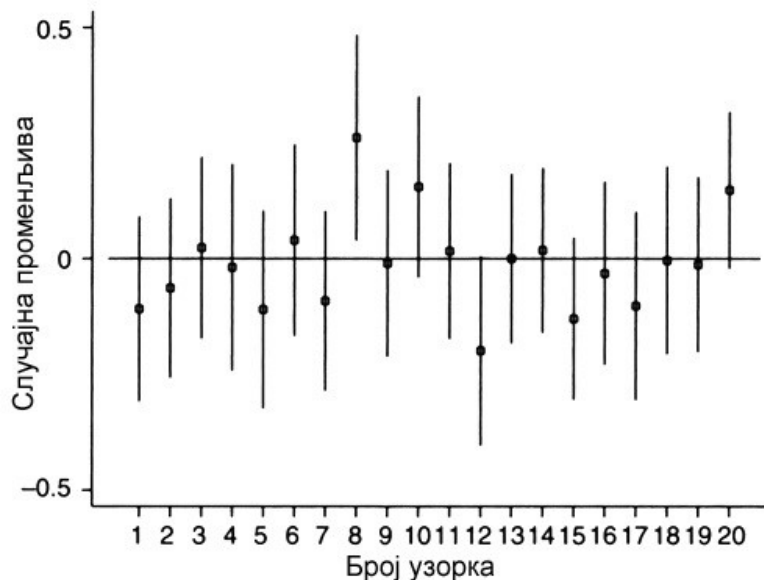
Често настају забуне између појмова "стандардна грешка" и "стандардно одступање". Ово је разумљиво, пошто стандардна грешка је стандардно одступање (расподеле узорка) и појмови се често замене у овом контексту. Конвенција је ова: користимо термин "стандардна грешка" када меримо прецизност предвиђања, и термин "стандардно одступање" када водимо рачуна о варијабилности узорака, популације или расподеле. Ако желимо да кажемо колико нам је добра предвиђање средине FEV1 мерења, наводимо стандардну грешку средине. Ако желимо да кажемо колико је широко растурање FEV1 мерења, наводимо стандардно одступање,  $s$ .

### 5.3 Интервали поверења (Confidence intervals)

Предвиђање средине FEV1 је једна вредност и зато се зове **тачка предвиђања (point estimate)**. Не постоји разлог да претпоставимо да ће средина популације бити потпуно једнака тачки предвиђања, средини узорка. Ипак, постоји могућност да ће можда бити близу ње, и износ за који ће вероватно да се разликује од предвиђања може се пронаћи из стандардне грешке. Оно што ми радимо, јесте да пронађемо границе које ће вероватно да укључе средину популације, и рецимо да предвидимо да средина популације лежи негде у интервалу (скуп свих

могућих вредности) између ових граница. Ово се зове **интервал предвиђања (interval estimate)**.

На пример, ако посматрамо 57 FEV1 мерења као велики узорак можемо да претпоставимо да је расподела средине узорка Нормална, и да је стандардна грешка добра предвиђање стандардног одступања. Због тога очекујемо да око 95% таквих средина буде унутар 1.96 стандардних грешака средине популације,  $\mu$ . Зато, за скоро 95% свих могућих узорака, средина популације мора да буде већа од средине узорка минус 1.96 стандардних грешака и мања од средине узорка плус 1.96 стандардних грешака. Ако смо израчунали  $\bar{x} - 1.96se$  и  $\bar{x} + 1.96se$  за све могуће узорке, 95% таквих интервала ће садржати средину популације. У овом случају, границе су  $4.062 - 1.96 \times 0.089$  до  $4.062 + 1.96 \times 0.089$  што даје 3.89 до 4.24 или 3.9 до 4.2 литра, заокружене на две значајне цифре; 3.9 и 4.2 се зову **95% границе поверења (95% confidence limits)** за предвиђање, и скуп вредности између 3.9 и 4.2 се зове **95% интервал поверења (95% confidence interval)**. Границе поверења су вредности на крају интервала поверења.



Слика 5.5 Средина и 95% интервал поверења за 20 случајних узорака 100 посматрања из Стандардизоване Нормалне расподеле

Прецизно говорећи, нетачно је рећи да постоји вероватноћа од 0.95 да средина популације лежи између 3.9 и 4.2, иако се често каже тако. Средина популације је број, а не случајна променљива, и нема вероватноћу. То је вероватноћа да ће границе израчунате из случајног узорка укључити вредност популације која износи 95%. Слика 5.5 показује интервале поверења за средину 20 случајних узорака од 100 посматрања из Стандардизоване Нормалне расподеле. Средина популације је наравно 0.0 и приказана је хоризонталном линијом. Неки узорци средине су близу 0.0, а неки су далеко, неки су изнад, а неки испод. Средина популације је садржана у 19 од 20 интервала поверења. У основи, за 95% интервале поверења тачно је рећи да вредност популације лежи унутар интервала. Ми само не знамо којих 95%. Ово изражавамо тако што кажемо да смо 95% сигурни да средина лежи између ових граница.

У FEV1 примеру, расподела средине узорка је Нормална и њено стандардно одступање је добро предвиђено јер је узорак велик. Ово није увек тачно и мада је обично могуће израчунати интервале поверења за неко предвиђање, они нису сви сасвим једноставни као они за средину предвиђену из великог узорка. Погледаћемо средину предвиђену из малог узорка у делу који обрађује један-узорак  $t$  метод.

Нема потребе да интервал поверења има вероватноћу 95%. На пример, можемо да израчунамо 99% границе поверења. Горња 0.5% тачка Стандардизоване Нормалне расподеле је 2.58 (табела 4.2), тако да вероватноћа да је Стандардно Нормално одступање изнад 2.58 или испод -2.58 је 1% и вероватноћа да ће бити између ових граница је 99%. 99% границе поверења за средину FEV1 су стога  $4.062 - 2.58 \times 0.089$  и  $4.062 + 2.58 \times 0.089$ , тј. 3.8 и 4.3 литра. Ово даје шири интервал него 95% граница, као што бисмо очекивали пошто смо сигурнији да ће средина бити укључена. Вероватноћа коју бирамо за интервал поверења је зато компромис између

жеље да се укључи вредност предвиђене популације и жеље да се избегну делови скале где постоји мала вероватноћа да ће средина бити пронађена. У већини случајева, 95% интервал поверења сматра се задовољавајућим.

#### 5.4 Стандардна грешка и интервал поверења за пропорцију

Стандардна грешка предвиђања пропорције се може израчунати на исти начин. Претпоставимо да је пропорција појединаца који имају одређени услов у датој популацији  $p$ , и да узмемо случајни узорак величине  $n$ , где је број посматраних са условом  $r$ . Тада предвиђена пропорција је  $r/n$ . Видели смо из дела који је обрађивао Биномну расподелу, да  $r$  долази из Биномне расподеле са средином  $np$  и варијансом  $np(1-p)$ . Под условом да је  $n$  велико, ова расподела је приближно Нормална. Тако  $r/n$ , предвиђена пропорција је Нормално распоређена са средином која се добија из  $np/n = p$ , и варијанса се добија као

$$\text{VAR}\left(\frac{r}{n}\right) = \frac{1}{n^2} \text{VAR}(r) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

пошто је  $n$  константа, и стандардна грешка је

$$\sqrt{\frac{p(1-p)}{n}}$$

Можемо предвидети ово тако што ћемо заменити  $p$  са  $r/n$ .

Стандардна грешка пропорције је од користи само ако је узорак довољно велики да се примени за Нормалну апроксимацију. Кратко објашњење овога је да  $np$  и  $n(1-p)$  оба треба да премаше 5. Ово је обично случај онда када узмемо у обзир тачно предвиђање. Ако покушамо да користимо метод за мање узорке, можемо добити апсурдне резултате. На пример, у студији преваленсе ХИВ-а код бивших затвореника (Turnbull и други 1992), од 29 жена које нису узимале дрогу једна је била ХИВ позитивна:

$$\text{Средина је } p = \frac{r}{n} = \frac{1}{29} = 0.034 = 3.4\%$$

$$\text{Стандардна грешка је } \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.034 \times 0.096}{29^2}} = 0.03365 = 3.4\%$$

$$\text{Интервал поверења од } 3.4\% - 1.96 \times 3.4\% = -3.1\% \text{ до } 3.4\% + 1.96 \times 3.4\% = 9.9\%.$$

Аутори су саопштили да је ово 3.4%, са 95% интервала поверења -3.1% до 9.9%. Нижа граница од -3.1% добијена из посматране пропорције минус 1.96 стандардних грешака, је немогућа.

Као што је Newcombe (1992) истакао, тачан 95% интервал поверења може се добити из тачних вероватноћа Биномне расподеле и он је 0.1% до 17.8%.

#### 5.5 Разлика између две средине

Претпоставимо да желимо да упоредимо средине  $\bar{x}_1$  и  $\bar{x}_2$ , два велика узорка, величина  $n_1$  и  $n_2$ . Очекивана разлика између средина узорака једнака је разлици између средина популације, то јест,  $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$ . Која је стандардна грешка разлике? Варијанса разлике између две независне случајне променљиве је збир њихових варијанси (део 3.6). Стога, стандардна грешка разлике између два независна предвиђања је квадратни корен збира квадрата њихових стандардних грешака. Стандардна грешка средине је  $\sqrt{s^2/n}$ , тако да стандардна грешка разлике између две независне средине је

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

На пример, у студији респираторних симптома код школске деце (Bland и други, 1974), желели смо да знамо да ли су деца за коју су родитељи пријавили да имају респираторне симптоме имала горе функције плућа него деца која нису пријављена да имају симптоме. Пријављено је 92 деце која су кашљала током дана или ноћи и њихова средина PEFR-а тј. максималног експираторног протока (Peak Expiratory Flow Rate) је била 294.8 литра/мин са

стандардним одступањем 57.1 литра/мин и 1643 деце која нису пријављена да имају ове симптоме имала су средину PEFR-а од 313.6 литра/мин са стандардним одступањем 55.2 литра/мин. На овај начин имамо два велика узорка, и можемо да применимо Нормалну расподелу. Имамо

$$n_1 = 92, \bar{x}_1 = 294.8, s_1 = 57.1, \text{ и } n_2 = 1643, \bar{x}_2 = 313.6, s_2 = 55.2$$

Разлика између две групе је  $\bar{x}_1 - \bar{x}_2 = 294.8 - 313.6 = -18.8$ . Стандардна грешка разлике је

$$\sqrt{se_1^2 + se_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{57.1^2}{92} + \frac{55.2^2}{1643}} = 6.11$$

Узорак ћемо третирати као јако велики, тако да се може претпоставити да разлика између средина потиче од Нормалне расподеле, и да је предвиђена стандардна грешка добра предвиђање стандардног одступања ове расподеле. 95% границе поверења за разлику су стога  $-18.8 - 1.96 \times 6.11$  и  $-18.8 + 1.96 \times 6.11$ , то јест,  $-6.8$  и  $-30.8$  литра/мин. Интервал поверења не укључује нулу, па имамо добар доказ да, у овој популацији, деца која су пријављена са дневним или ноћним кашљем имају мању средину PEFR него друга деца. Разлика је предвиђена да буде између 7 и 31 литра/мин нижа код деце са симптомом, тако да може да буде прилично мала.

Када имамо упарене податке, као што су крос-овер (*cross-over*) испитивања или одговарајућу студију контроле случаја, метод два-узорка не ради. Уместо тога, израчунамо разлике између упарених посматрања за сваки субјект, затим налазимо средину разлике, њену стандардну грешку и интервал поверења, као у делу 5.3.

## 5.6 Поређење две пропорције

Можемо да применимо метод из претходног дела на две пропорције. Стандардна грешка пропорције  $p$  је  $\sqrt{p(1-p)/n}$ . За две одвојене пропорције  $p_1$  и  $p_2$ , стандардна грешка разлике између њих је

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Ако се услови Нормалне апроксимације испуне (видети део 5.4) можемо да одредимо интервал поверења за разлику на уобичајен начин. На пример, посматрајмо табелу 5.3.

Табела 5.3 Кашаљ током дана или ноћи код деце која имају 14 година и бронхитис пре 5-те године живота (Holland и други 1978)

	Кашаљ код 14		Бронхитис пре 5	
	Да	Не	Укупно	
Да	26	44	70	
Не	247	1002	1249	
Укупно	273	1046	1319	

Истраживачи су желели да знају у којој мери деца са бронхитисом у раном детињству имају више респираторних симптома касније у животу него друга деца. Можемо да израчунамо разлику између пропорција пријављених за кашаљ током дана или ноћи међу децом са

историјом бронхитиса и код деце без историје бронхитиса пре 5-те године. Имамо предвиђање две пропорције  $p_1 = 26/273 = 0.09524$  и  $p_2 = 44/1046 = 0.04207$ . Разлика између њих је  $p_1 - p_2 = 0.09524 - 0.04207 = 0.05317$ . Стандардна грешка разлике је

$$\sqrt{\frac{p_1(1-p_1)}{n_1}} + \sqrt{\frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.09524 \times (1-0.09524)}{273} + \frac{0.04207 \times (1-0.04207)}{1046}} = \sqrt{0.000315639 + 0.000038528} = \sqrt{0.000354167} = 0.0188$$

95% интервал поверења за разлику је  $0.05317 - 1.96 \times 0.0188$  до  $0.05317 + 1.96 \times 0.0188$ , тј. од 0.016 до 0.090. Иако разлика није потпуно прецизно предвиђена, интервал поверења не укључује нулу и даје нам јасан доказ да постоји већа вероватноћа да ће деца са пријављеним бронхитисом у периоду раног детињства имати респираторне симптоме касније у животу него друга деца. Подаци о функционисању плућа у делу 5.5 дају нам разлог да претпоставимо да ово није у потпуности последица пристрасности одговора. Као и у делу 5.4, интервал поверења мора бити израчунат другачије за мале узорке.

Ова разлика у пропорцијама није лака за тумачење. Однос (*ratio*) између две пропорције је често кориснији. Однос пропорције деце са кашљем у 14-тој години и бронхитисом пре пете године према пропорцији деце са кашљем у 14-тој и оних без бронхитиса у петој години је  $p_1/p_2 = 0.09524/0.04207 = 2.26$ . Деца са бронхитисом пре пете године више него два пута су склонија кашљу у току дана или ноћи у 14-тој години од деце без такве историје.

Стандардна грешка овог односа је комплексна и, пошто је то однос, пре него разлика, он се не апроксимира добро са Нормалном расподелом. Међутим, ако узмемо логаритам односа, добијамо разлику између два логаритма, јер је  $\log(p_1/p_2) = \log(p_1) - \log(p_2)$ . За логаритам односа можемо да пронађемо стандардну грешку прилично лако. Користимо резултат да, за било коју случајну променљиву  $X$  са средином  $\mu$  и варијансом  $\sigma^2$ , приближна варијанса од  $\log(X)$  дата је изразом  $VAR(\log_e(X)) = \sigma^2 / \mu^2$  (погледати Kendall and Stuart 1969). Стога варијанса од  $\log(p)$  је:

$$VAR(\log(p)) = \frac{p(1-p)/n}{p^2} = \frac{1-p}{np}$$

За разлику између два логаритма добијамо :

$$VAR(\log_e(p_1/p_2)) = VAR(\log_e(p_1)) + VAR(\log_e(p_2)) = \frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$$

Стандардна грешка је квадратни корен овога (ова формула је често написана преко учесталости, али ја мислим да је ова верзија јаснија). На пример логаритам односа је  $\log_e(2.26385) = 0.81707$  и стандардна грешка је:

$$\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}} = \sqrt{\frac{1-0.09524}{273 \times 0.09524} + \frac{1-0.04207}{1046 \times 0.04207}} = \sqrt{\frac{0.90476}{26} + \frac{0.95793}{44}} = \sqrt{0.05657} = 0.23784$$

95% интервал поверења за логаритам односа је стога  $0.81707 - 1.96 \times 0.2378$  до  $0.81707 + 1.96 \times 0.2378$  тј. од 0.35089 до 1.28324. 95% интервал поверења за однос пропорција је антилогаритам овога:  $e^{0.35089}$  до  $e^{1.28324}$  и то је од 1.42 до 3.61. На овај начин предвиђамо да је пропорција деце пријављене са кашљем током дана или ноћи са историјом бронхитиса између 1.4 и 3.6 пута пропорције код деце без историје бронхитиса.

Пропорција појединаца у популацији код којих се обољење развија или се јавио симптом је једнака вероватноћи да ће се код било ког појединца развити ово обољење, што се назива **ризик (risk)** од индивидуалног развијања обољења. Одавде је у табели 5.3 ризик да ће дете са бронхитисом пре пете године кашљати у четрнаестој години  $26/273 = 0.09524$ , а ризик за децу која нису имала бронхитис пре пете године је  $44/1046 = 0.04207$ . Да бисмо упоредили ризике код људи са и без одређених фактора ризика, гледамо однос између ризика са фактором и ризика без фактора, **релативни ризик (relative risk)**. Тако релативни ризик за кашаљ код деце са четрнаест година која су имала бронхитис пре пете године је 2.26. Да би предвидели

директно релативни ризик, потребна је студија кохорте (*cohort study*), као у табели 5.3. Релативни ризик за студију контроле случаја израчунавамо на други начин (део 10.6).

### 5.7 Који је тачан интервал поверења?

Интервал поверења оцењује једино грешке изазване узимањем узорка. Оне не дозвољавају одступања у узорку и дају нам процене за популацију за коју наши подаци могу да се сматрају случајним узорком.

На пример, Salvesen и други (1992) је саопштио праћење два случајно одабрана процеса рутинског ултразвучног прегледа током трудноће. Током осме и девете године, деца жена које су биле посматране су била праћена. Подгрупа деце подвргнута је специфичним тестовима за дислексију. Резултати теста класификовали су 21 од 309 посматране деце (7%, 95% интервала поверења 3%-10%) и 26 од 294 контролних (9%, 95% интервала поверења 4%-12%) као дислексичне. Много кориснији био би интервал поверења за разлику између преваленси (-6.3 до 2.2 процентне тачке) или њихов однос (0.44 до 1.34), јер бисмо онда могли да упоредимо групе директно.