

ПРЕДМЕТ

< СТАТИСТИКА У ФАРМАЦИЈИ >

Предавање број 9

**<** **ПРЕДСТАВЉАЊЕ ПОДАТАКА И ВЕРОВАТНОЋА >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 9 | Представљање података и вероватноћа | Однос и пропорције. Значајне цифре. Представљање табела. Графикони. Особине вероватноће. Расподела вероватноће и случајне променљиве. Биномна расподела. Средина и варијанса. Poisson-ова расподела. | Упознавање са различитим начинима представљања података и основама вероватноће |

Copyright © 2012 – Факултет медицинских наука Универзитета у Крагујевцу. Сва права задржана. Без претходне писмене дозволе од стране Факултета медицинских наука забрањена је репродукција, трансфер, дистрибуција или меморисање неког дела или читавих садржаја овог документа, копирањем, снимањем, електронским путем, скенирањем или на било који други начин.

Copyright © 2012 – Faculty of Medical Sciences of University of Kragujevac. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying,, recording, scanning or otherwise, without the prior written permission of Faculty of Medical Sciences.

**САДРЖАЈ**

[Представљање података и Вероватноћа 2](#_Toc274937173)

[2 Представљање података 2](#_Toc274937174)

[2.1 Стопе и пропорције (rates and proportions) 2](#_Toc274937175)

[2.2 Значајне бројке (significant figures) 3](#_Toc274937176)

[2.3 Представљање табела 6](#_Toc274937177)

[2.4 Kружни графикони (pie charts) 6](#_Toc274937178)

[2.5 Тракасти графикони (bar charts) 8](#_Toc274937179)

[2.6 Дијаграми растурања (scatter diagrams) 9](#_Toc274937180)

[2.7 Линијски графикон и временски низ (line graphs and time series) 12](#_Toc274937181)

[2.8 Двосмислени графикони (misleading graphs) 13](#_Toc274937182)

[2.9 Логоритамске скале (logarithmic scales) 15](#_Toc274937183)

[3. Вероватноћа 16](#_Toc274937184)

[3.1 Вероватноћа 16](#_Toc274937185)

[3.2 Особине вероватноће 16](#_Toc274937186)

[3.3 Расподела вероватноће и случајне променљиве 17](#_Toc274937187)

[3.4 Биномна расподела (Binomial distribution) 18](#_Toc274937188)

[3.5 Средина и варијанса (Мean and variance) 19](#_Toc274937189)

[3.6 Kарактеристике средине и варијансе 20](#_Toc274937190)

Предавање бр. 9

**<** **ПРЕДСТАВЉАЊЕ ПОДАТАКА И ВЕРОВАТНОЋА >**

# Представљање података и Вероватноћа

## 2 Представљање података

### 2.1 Стопе и пропорције (rates and proportions)

Kада имамо податке у облику учесталости, често је потребно да упоредимо учесталост са одређеним стањима у групама са различитим укупним вредностима. У табели испод, на пример, Christie (1979) је проучавао разлике у преживљавању пацијената који су имали мождани удар 1978. године, након увођења C-T скенера главе, са оним пацијентима леченим 1974, пре него што је уведен C-T скенер. Он је узео групу пацијената лечених 1978, којима је урађен C-T скенер главе, и упоредио их са оним пацијентима који су лечени 1974, и који су били истих година, дијагнозе и нивоа губљења свести при пријему. Две групе парова пацијената су упоређиване, 29 где је касније пацијент имао C-T скенер и 89 где ниједан пацијент није имао C-T скенер. Kаснији пацијент је боље прошао у 9 из прве групе и 34 из друге групе. Да би упоредили ове учесталости поредимо однос 9/29 и 34/89. То су 0.31 и 0.38, и можемо закључити да постоји мала разлика. У табели испод, они су дати као проценти, то јест, удео од 100 пре него од 1, како би се избегао децимални зарез.

|  |
| --- |
| Анализа разлике у преживљавању за одговарајуће парове пацијената који су имали мождани удар (Christie 1979) |
| |  |  |  | | --- | --- | --- | |  | C-T скенирање у 1978 | Нема C-T скенирања у 1978 | | Парови из 1978 бољи него 1974 | 9 (9/29=31%) | 34 (34/89=38%) | | Парови са истим исходом | 18 (18/29=62%) | 38 (38/89=43%) | | Парови из 1978 гори него 1974 | 2 (2/29=7%) | 17 (17/89=19%) | | Укупно: | 29 | 89 | |

**Стопа** (**rate**) изражава учесталост карактеристике од интереса по 1000 (или по 100 000, и сл.) по популацији, по јединици времена. На пример, у студији Doll-а и Hill-а (1956) резултата истраживања пушења код лекара, подаци су представљени као број умрлих на 1000 лекара годишње. То није пропорција, пошто су даља прилагођавања направљена да размотре посматрани временски период. Осим тога, стопа је прилагођена да узме у обзир све разлике у расподели старости пушача и непушача (део 12.2). Понекад се стварни именилац стопе може стално мењати. Број умрлих од рака плућа међу мушкарцима у Енглеској и Велсу за 1983 је био 26502. Именилац за стопу смртности, број мушкараца у Енглеској и Велсу, променио се током 1983, пошто су неки умрли, неки су рођени, неки напустили земљу а неки су ушли. Стопа смртности се израчунава коришћењем броја представника, процењеног становништва крајем јуна 1983, средином године. То је било 24 175 900, дајући стопу смртности од 26 502/24 175 900, што је једнако 0.001 096, или 109.6 умрлих на 100 000 изложених ризику годишње. Одређени број стопа употребљен у медицинској статистици је описан у делу 12.5.

Употреба стопе и пропорција омогућава нам да упоредимо учесталости добијене од група неједнаких величина, базе популације или временског периода, али морамо се чувати њихове употребе када њихове базе или имениоци нису дати. Стопе и пропорције су моћне алатке и морамо да пазити да не постану одвојене од оригиналних података.

### 2.2 Значајне бројке (significant figures)

Kада смо израчунали стопу смртности због рака плућа међу мушкарцима у 1983 цитирали смо одговор као 0.001 096 или 109.6 на 100.000 годишње. Ово је апроксимација. Стопа на највећи број цифара коју ће мој калкулатор дати је 0.001 096 215 653 и тај број ће вероватно ићи унедоглед, претварајући се у периодичан низ цифара. Децимални систем представљања бројева не може у целини представљати разломке тачно. Mи знамо да је 1/2 = 0.5, а 1/3 = 0,333 333 33 ..., понављају се бесконачно. То нас обично не брине, јер је за већину апликација разлика између 0.333 и 1/3 сувише мала да би била важна. Само првих неколико цифара броја које нису нуле су важне и зовемо их **значајне** **цифре** (**significant digits**) или **значајне** **бројке** (**significant figures**). Обично има мало смисла позивати се на статистичке податке са више од три значајне цифре.

Уосталом, тешко да је важно да ли је смртност од рака плућа 0.001 096 или 0.001 097. Вредност 0.001 096 је дата на 4 значајне бројке. Водеће нуле нису значајне, прва значајна цифра у овом броју је ''1''. На три значајне цифре ћемо добити 0.001 10, јер је последња цифра 6 па је 9 која јој претходи заокружена на 10. Имајте у виду да значајне бројке нису исто што и децимална места. Број 0.001 10 је дат на 5 децималних места, број цифара иза децималне тачке. Kада је заокруживање на најближи број, остављамо последњу значајну цифру, 9 у овом случају, ако је оно што следи мање од 5, и увећавамо за један, ако је оно што следи веће од 5. Kада имамо тачно 5, ја бих увек заокружио, односно 1.5 иде на 2. То значи да 0, 1, 2, 3, 4 иде доле и 5, 6, 7, 8, 9 иде горе, што се чини непристрасно. Оно што *не* треба да радимо је да заокружимо бројеве на неколико значајних бројки, пре него што смо завршили наше прорачуне. У примеру стопе смртности од рака плућа, претпоставимо да заокружујемо бројилац и именилац на две значајне бројке. Имамо 27 000/24 000 000 = 0.001 125, а одговор је тачан само за две бројке. Ово се може раширити кроз рачунање и изазвати накупљање грешака. Mи увек покушавамо да задржимо више значајних бројки него што смо тражили за коначан одговор.

Размотрите Табелу 2.1. Она показује податке о смртности у смислу тачног броја умрлих у једној години. Табела је преузета из много веће табеле (OPCS 1991) која показује број умирања од сваког узрока смрти у међународној класификацији болести (*International Classification of Diseases* - ICD), која даје нумеричке кодове за много стотина узрока смрти. Пуна табела, која такође даје умрле по старосној групи, покрива 70 страна A4 формата. Табела 2.1 показује умрле од широке групе болести под називом ICD поглавља. Ова табела није добар начин да се представе ови подаци, ако желимо да разумемо расподелу учесталости узрока смрти, и разлике између узрока смрти код мушкараца и код жена. Ово је још више важи за 70 страна оригинала. То није сврха табеле, наравно. То је извор података, документ на који се упућују корисници да би издвојили информације за своје потребе. Хајде да видимо како табела 2.1 може бити поједностављена.

|  |
| --- |
| Табела 2.1 Умрли према полу и узроку, Енглеска и Велс, 1989 (OPCS 1991, DH2 No. 10) |
| |  |  |  |  | | --- | --- | --- | --- | | I.C.D. | Тип болести | Број смртних случајева | | | Мушкарци | Жене | | I | Инфективне и паразитске | 1 246 | 1 297 | | II | Неоплазме (тумори) | 75 172 | 69 948 | | III | Ендокрине, дијететскии метаболичке болести и обољења имунитета | 4 395 | 5 758 | | IV | Крвне и органа крвотока | 1 002 | 1 422 | | V | Ментална обољења | 4 493 | 9 225 | | VI | Нервни систем и органи осећаја | 5 466 | 5 990 | | VII | Кардиоваскуларне болести | 127 435 | 137 165 | | VIII | Дисајни систем | 33 489 | 33 223 | | IX | Пробавни систем | 7 900 | 10 779 | | X | Мокраћни систем | 3 616 | 4 156 | | XI | Компликације трудноће, рођења детета и породиља | 0 | 56 | | XII | Кожна и поткожна ткива | 250 | 573 | | XIII | Мишићно-скелетни систем и везивна ткива | 1 235 | 4 139 | | XIV | Урођене аномалије | 897 | 869 | | XV | Одређени услови настали у перинаталном периоду | 122 | 118 | | XVI | Знаци, симптоми и болест-дефинисани услови | 1 582 | 3 082 | | XVII | Повреде и тровања | 11 073 | 6 427 | | Укупно |  | 279 373 | 294 227 | |

Прво, можемо смањити број значајних бројки. Будимо екстремни и смањимо податке на један значајан број (Табела 2.2). То чини поређења прилично лакшим, али још увек није јасно који су најважнији узроци смрти.

|  |
| --- |
| Табела 2.2 Умрли према полу и узроку, Енглеска и Велс, 1989, заокружени на једну значајну бројку |
| |  |  |  |  | | --- | --- | --- | --- | | I.C.D. | Тип болести | Број смртних случајева | | | Мушкарци | Жене | | I | Инфективне и паразитске | 1 000 | 1 000 | | II | Неоплазме (тумори) | 80 000 | 70 000 | | III | Ендокрине, дијететскии метаболичке болести и обољења имунитета | 4 000 | 6 000 | | IV | Крвне и органа крвотока | 1 000 | 1 000 | | V | Ментална обољења | 4 000 | 9 000 | | VI | Нервни систем и органи осећаја | 5 000 | 6 000 | | VII | Кардиоваскуларне болести | 100 000 | 100 000 | | VIII | Дисајни систем | 30 000 | 30 000 | | IX | Пробавни систем | 8 000 | 10 000 | | X | Мокраћни систем | 4 000 | 4 000 | | XI | Компликације трудноће, рођења детета и породиља | 0 | 60 | | XII | Кожна и поткожна ткива | 300 | 600 | | XIII | Мишићно-скелетни систем и везивна ткива | 1000 | 4000 | | XIV | Урођене аномалије | 900 | 900 | | XV | Одређени услови настали у перинаталном периоду | 100 | 100 | | XVI | Знаци, симптоми и болест-дефинисани услови | 2 000 | 3 000 | | XVII | Повреде и тровања | 10 000 | 6 000 | | Укупно |  | 300 000 | 300 000 | |

Ово можемо побољшати реорганизовањем табеле тако да у њој буде само најчешћи узрок смрти, прво, болести система циркулације крви (Табела 2.3). Такође можемо да комбинујемо много мањих категорија у ''друге'' (*''others''*) групе. Jа сам ово учинио произвољно, комбиновањем свих оних узрока смрти са уделом мањим од 2% од укупног броја.

|  |
| --- |
| Табела 2.3 Умрли према полу, Енглеска и Велс, 1989, за главне узроке |
| |  |  |  | | --- | --- | --- | | I.C.D. Тип болести | Број смртних случајева | | | Мушкарци | Жене | | Кардиоваскуларне болести (VII) | 100 000 | 100 000 | | Неоплазме (тумори) (II) | 80 000 | 70 000 | | Дисајни систем (VIII) | 30 000 | 30 000 | | Повреде и тровања (XVII) | 10 000 | 6 000 | | Пробавни систем (IX) | 8 000 | 10 000 | | Друго | 20 000 | 20 000 | | Укупно | 300 000 | 300 000 | |

Сада је јасно на први поглед да су најважнији узроци смрти у Енглеској и Велсу кардиоваскуларне болести, неоплазме и болести дисајног система, и да они умањују све остале. Наравно, смртност није једини показатељ важности болести. ICD поглавље 13, болести мишићно-скелетног система и везивног ткива, лако се види из табеле 2.2 да су само мањи узроци смрти, али ова група обухвата артритис и реуматизам, најважнију болест у свом утицају на свакодневне активности.

### 2.3 Представљање табела

Табеле 2.1, 2.2 и 2.3 илуструју одређени број корисних тачака у презентацији табела. Kао и све табеле у овим предавањима, оне су дизајниране тако да стоје саме изван текста. Нема потребе позивати се на материјал закопан у неком пасусу да би се протумачила табела. Табела има за циљ да пренесе информацију, тако да би требало да буде лака за читање и разумевање. Табела треба да има јасан наслов, наводећи јасно и недвосмислено оно што представља. Редови и колоне морају бити јасно означени.

Kада су пропорције, стопе или проценти употребљени у табели, заједно са учесталостима, оне се морају лако разликовати једна од друге. Ово може да се уради, додавајући симбол ''%'', или укључујући и место децимале. Додавање у табели, реда ''укупно'' и ''100%'' јасно указује да су проценти израчунати од одређеног броја из групе лечења, пре него од броја са одређеним исходом или укупног броја пацијената.

### 2.4 Kружни графикони (pie charts)

Често је згодно представити податке помоћу слика. Информације се могу пренети много брже дијаграмом него табелом бројева. Ово је нарочито корисно када се подаци представљају публици, пошто овде информација треба да буде пренета у ограниченом времену. Ово такође може помоћи да читалац добије истакнуте тачке табеле бројева. Нажалост, ако није предузета посебна брига, дијаграми могу бити веома двосмислени и треба их третирати само као додатак бројевима, а не замену.

Већ смо разговарали о методима илустровања расподеле учесталости квалитативне променљиве. Сада ћемо погледати еквивалент хистограма за квалитативне податке, **кружни** **графикон** (**pie chart**) или **кружни** **дијаграм** (**pie diagram**). Он показује релативну учесталост за сваку категорију поделом круга на секторе, чији су углови пропорционални релативној учесталости. Онда множимо сваку релативну учесталост са 360, да би добили одговарајући угао у степенима.

|  |
| --- |
| Табела 2.4 Прорачуни за кружни графикон расподеле узрока смрти код жена |
| |  |  |  |  | | --- | --- | --- | --- | | Узрок смрти | Учесталост | Релативна учесталост | Угао (степени) | | Циркулација крви | 137 165 | 0.466 19 | 168 | | Неоплазме (тумори) | 69 948 | 0.237 73 | 86 | | Дисајни систем | 33 223 | 0.112 92 | 41 | | Повреде и тровања | 6 427 | 0.021 84 | 8 | | Пробавни систем | 10 779 | 0.036 63 | 13 | | Нервни систем | 5 990 | 0.020 36 | 7 | | Друго | 30 695 | 0.104 32 | 38 | | Укупно | 294 227 | 1.000 00 | 361 | |



Слика 2.1. Kружни графикон који показује расподелу узрока смрти међу женама, Енглеска и Велс, 1983

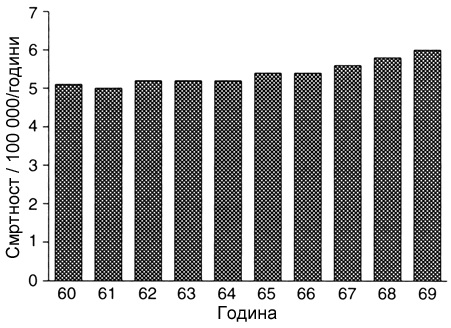
Табела 2.4 показује прорачун за цртање кружног графикона за представљање расподеле узрока смрти за жене, користећи податке из табеле 2.1 и 2.3 (укупни степени су 361, уместо 360, због грешака заокруживања при израчунању) резултирајући кружним графиконом приказаним на слици 2.1. За овај дијаграм је речено да личи на питу исечену на комаде за сервирање, отуда и његово име.

### 2.5 Стубичасти или тракасти графикони (bar charts)

**Тракасти** **графикон** (**bar chart**) или **тракасти** **дијаграм** (**bar diagram**) приказује податке у облику хоризонталне или вертикалне црте. На пример, табела 2.5 показује смртност због рака једњака у Енглеској и Велсу у периоду од 10 година. Слика 2.2 приказује те податке у облику тракастог графикона, висине црта су пропорционалне смртности.

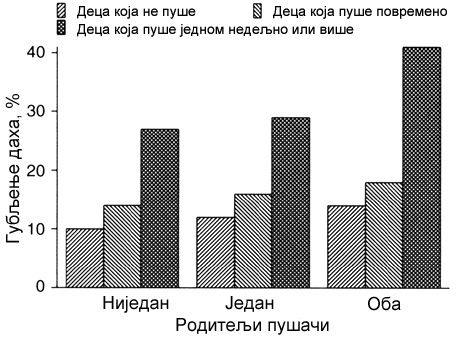
|  |
| --- |
| Табела 2.5 Рак једњака: стандардизована стопа смртности на 100 000 годишње, Енглеска и Велс, 1960-1969 |
| |  |  |  |  | | --- | --- | --- | --- | | Година | Стопа смртности | Година | Стопа смртности | | 60 | 5.1 | 65 | 5.4 | | 61 | 5.0 | 66 | 5.4 | | 62 | 5.2 | 67 | 5.6 | | 63 | 5.2 | 68 | 5.8 | | 64 | 5.2 | 69 | 6.0 | |

Постоји много могућности за коришћење тракастог графикона. Kао што је то и приказано на слици 2.2, може да се користи за приказивање односа између две променљиве, једна од њих квантитативна, а друга било квалитативна или квантитативна променљива која је груписана, као што је и време у годинама. Вредности прве променљиве су приказане висином црта, једна црта за сваку категорију друге променљиве.



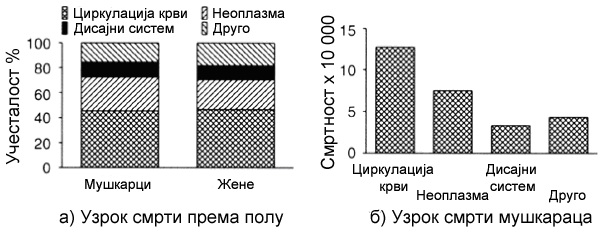
Слика 2.2 Тракасти графикон приказује однос између смртности због рака једњака и године, Енглеска и Велс, 1960-1969

Тракасти графикони се могу користити за представљање односа између више од две променљиве. Слика 2.3 приказује однос између дечијих притужби на губљење даха (*breathlessness*) и пушења цигарета од стране деце и њихових родитеља. Брзо можемо видети да се преваленса (*prevalence*) или распрострањеност симптома повећава и са пушењем детета и са пушењем његових родитеља.



Слика 2.3 Тракасти графикон показује однос између распрострањености само-пријављивања губљења даха међу ученицима и два могућа узрочна фактора

Тракасти графикони се такође могу користити за приказ учесталости. На пример, слика 2.4(а) приказује расподелу релативне учесталости узрока смрти међу мушкарцима и женама. Слика 2.4(б) показује расподелу учесталости узрока смрти међу мушкарцима. Слика 2.4(б) изгледа врло слично као хистограм. Разлика између ова два појма није јасна. Већина статистичара би описала слике 1.3, 1.4 и 1.6 као хистограме, а слике 2.2 и 2.3 као тракасте графиконе, али видео сам књиге које стварно окрену ову терминологију и друге који замене израз "хистограм" за график густине учесталости, као што су слике 1.4 и 1.6.



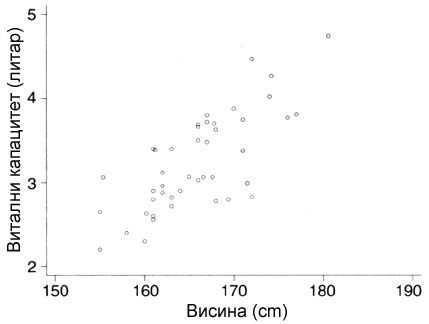
Слика 2.4 Тракасти графикони који приказују умрле према полу и узроку, Енглеска и Велс, 1989 (OPCS 1991, DH2 No. 10) (подаци из табеле 2.1)

### 2.6 Дијаграми растурања (scatter diagrams)

Тракасти графикон би био прилично незграпан метод за приказивање односа између две непрекидне променљиве, као што су витални капацитет и висина (Табела 2.6). За то користимо **дијаграм** **растурања** (**scatter diagram**) или **скатерграм** (**scattergram**) (Слика 2.5).

|  |
| --- |
| Табела 2.6 Витални капацитет (VC) и висина за 44 студенткиње медицине |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Висина (цм) | VC (литри) | Висина (цм) | VC (литри) | Висина (цм) | VC (литри) | Висина (цм) | VC (литри) | | 155.0 | 2.20 | 161.2 | 3.39 | 166.0 | 3.66 | 170.0 | 3.88 | | 155.0 | 2.65 | 162.0 | 2.88 | 166.0 | 3.69 | 171.0 | 3.38 | | 155.4 | 3.06 | 162.0 | 2.96 | 166.6 | 3.06 | 171.0 | 3.75 | | 158.0 | 2.40 | 162.0 | 3.12 | 167.0 | 3.48 | 171.5 | 2.99 | | 160.0 | 2.30 | 163.0 | 2.72 | 167.0 | 3.72 | 172.0 | 2.83 | | 160.2 | 2.63 | 163.0 | 2.82 | 167.0 | 3.80 | 172.0 | 4.47 | | 161.0 | 2.56 | 163.0 | 3.40 | 167.6 | 3.06 | 174.0 | 4.02 | | 161.0 | 2.60 | 164.0 | 2.90 | 167.8 | 3.70 | 174.2 | 4.27 | | 161.0 | 2.80 | 165.0 | 3.07 | 168.0 | 2.78 | 176.0 | 3.77 | | 161.0 | 2.90 | 166.0 | 3.03 | 168.0 | 3.63 | 177.0 | 3.81 | | 161.0 | 3.40 | 166.0 | 3.50 | 169.4 | 2.80 | 180.6 | 4.74 | |

Он је направљен обележавањем скале две променљиве дуж хоризонталне и вертикалне осе. Сваки пар мерења је обележен крстом, кругом, или неким другим одговарајућим симболом на месту означеном коришћењем мерења као координата.

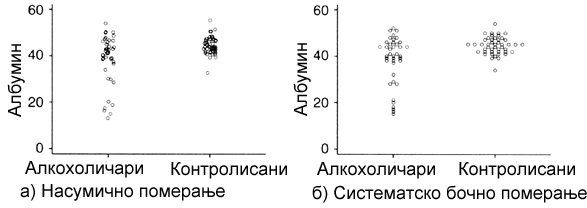


Слика 2.5 Дијаграм растурања који приказује однос између виталног капацитета и висине за групу студенткиња медицине

Табела 2.7 показује серумске беланчевине (*serum albumin*) мерене у групи пацијената алкохоличара и групи контролисаних (Hickish *et al.* 1989). Такође можемо користити дијаграм растурања за представљање ових података. Вертикална оса представља беланчевине и бирамо две произвољне тачке на хоризонталној оси за представљање група.

|  |
| --- |
| Табела 2.7 Беланчевине измерене код алкохоличара и контролисаних |
| |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Алкохоличари | | | | | | Контролисани | | | | | | | | 15 | 28 | 39 | 41 | 44 | 48 | 34 | 41 | 43 | 45 | 45 | 47 | 50 | | 16 | 29 | 39 | 43 | 45 | 48 | 39 | 42 | 43 | 45 | 45 | 47 | 50 | | 17 | 32 | 39 | 43 | 45 | 49 | 39 | 42 | 43 | 45 | 45 | 48 | 50 | | 18 | 37 | 40 | 44 | 46 | 51 | 40 | 42 | 43 | 45 | 46 | 48 | 50 | | 20 | 38 | 40 | 44 | 46 | 51 | 41 | 42 | 44 | 45 | 46 | 48 | 54 | | 21 | 38 | 40 | 44 | 46 | 52 | 41 | 42 | 44 | 45 | 47 | 49 |  | | 28 | 38 | 41 | 44 | 47 |  | 41 | 42 | 44 | 45 | 47 | 49 |  | |

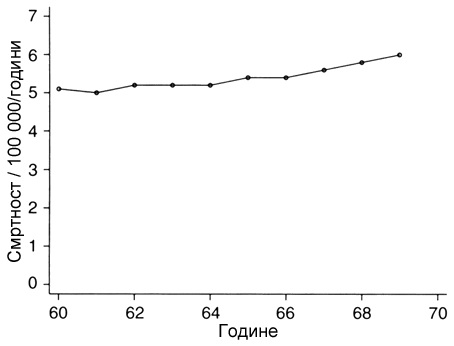
У табели 2.7 постоје многа идентична посматрања у свакој групи, тако да је потребно размотрити ово у дијаграму растурања. Aко постоји више од једног посматрања у истој координати можемо указати на то на неколико начина. Mожемо користити број посматрања уместо изабраног симбола, али овај метод постаје застарео. Kао што је на слици 2.6(а) приказано, можемо мало преместити тачке у насумичном правцу (тзв. **треперење - jittering**). То је оно што ***SPSS*** ili ***Stata*** (статистички софтверски пакети за општу употребу) раде и оно што сам учинио и ја у већини ових предавања. Aлтернативно, можемо користити систематско бочно померање, ради формирања уређеније слике као на слици 2.6(б). Ово последње се често користи када је променљива на хоризонталној оси категоријска (квалитативна) пре него непрекидна. Такви дијаграми растурања су врло корисни за проверу претпоставке неких аналитичких метода које ћемо касније користити. Дијаграм растурања где је једна променљива група назива се **дијаграм тачке** (**dot plot**). Kао презентациона средства, они нам омогућавају да покажемо много више информација него што тракасти графикон групе средина може. Из тог разлога, статистичари их обично преферирају у односу на остале типове графичког приказивања.



Слика 2.6 Дијаграми растурања који приказују беланчевине измерене код алкохоличара и контролисаних (подаци из табеле 2.7)

### 2.7 Линијски графикон и временски низ (line graphs and time series)

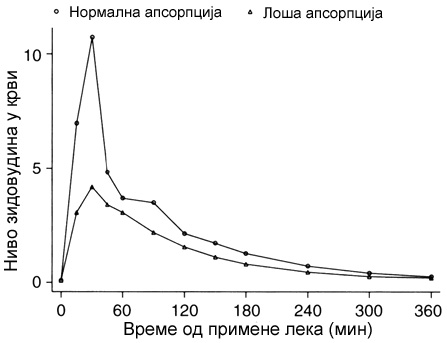
Подаци из табеле 2.5 су уређени на начин на који они у табели 2.6 нису, по томе што су снимљени по временским интервалима. Такви подаци се називају **временски низови** (**time series**). Aко нацртамо дијаграм растурања са таквим подацима, као на слици 2.7, природно је да се споје узастопне тачке линијама како би се формирао линијски график.



Слика 2.7 Линијски график који приказује промене смртности од рака једњака (*oesophagus*) током времена

Није потребно ни да обележавамо тачке уопште; све што нам је потребно је линија. То не би било паметно на слици 2.5, пошто су посматрања независна једна од других и сасвим неповезана, док на слици 2.7 вероватно је да ће постојати однос између суседних тачака. Овде ће стопа смртности забележена за рак једњака зависити од одређеног броја ствари које варирају током времена, вероватно укључујући могуће факторе узрока, као што су конзумирање дувана и алкохола, и клинички фактори, као што су боље дијагностичке технике и методе лечења.

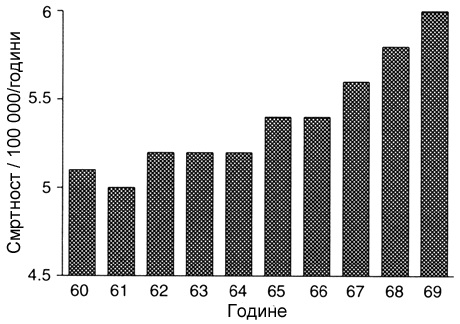
Линијски графикони су нарочито корисни када желимо да покажемо промену више од једног квантитета током времена. Слика 2.8 показује ниво зидовудина (*zidovudine* - AZT) у крви AИДС болесника неколико пута након примене лека, код пацијената са нормалном апсорпцијом масти и лошом апсорпцијом масти. Разлика у реаговању на оба третмана је врло јасна.



Слика 2.8 Линијски графикон за приказивање реаговања на примену зидовудина (*zidovudine*) у две групе AИДС пацијената

### 2.8 Двосмислени графикони (misleading graphs)

Слика 2.2 има јасан назив и ознаку и може се читати независно од околног текста. Начела јасноће наведена за табеле подједнако су примењива и овде. Уосталом, дијаграм је метода преношења информација брзо и ова намера пропада ако читалац или публика мора да проведе време покушавајући да разреши оно што дијаграм заправо значи. Будући да визуелни утицај дијаграма може бити тако велики, даљи проблеми настају у њиховом коришћењу.

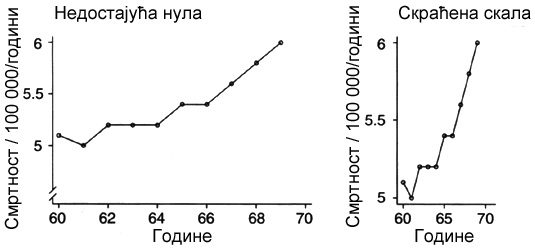


Слика 2.9 Тракасти графикон са изостављеном нулом на вертикалној скали

Први од њих је нула која недостаје. Слика 2.9 приказује други тракасти графикон који представља податке из табеле 2.5 (рак једњака). Овај графикон изгледа да приказује веома брз пораст смртности у поређењу са постепеним повећањем приказаним на слици 2.2. Ипак оба показују исте податке. Слика 2.9 изоставља већину вертикалне скале, и уместо тога протеже тај мали део скале где се промена одвија. Чак и када смо свесни тога, тешко је гледати овај график и не мислити да показује велики пораст смртности. Помаже ако замислимо почетну линију близу дна странице.

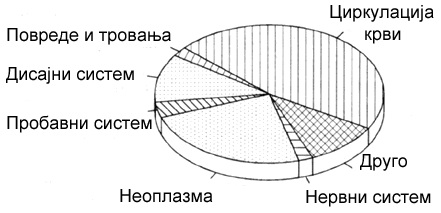
Такође нема нуле на хоризонталној оси на сликама 2.2 и 2.9. Постоје два разлога за то. Не постоји практично ''нула време'' у календару, користимо произвољну нулу. Такође, постоји неизречена претпоставка да стопа смртности варира с временом, а не обрнуто.

Нула је изостављена на слици 2.5. Ово се скоро увек ради у дијаграмима растурања, али ако желимо да одмеримо значај односа између виталног капацитета и висине преко релативне промене у виталном капацитету кроз опсег висине потребна нам је нула на скали виталног капацитета. Порекло је често изостављено на дијаграму растурања, јер се обично бавимо постојањем односа и расподелама праћеним запажањима, пре него његовим обимом. Ово последње смо проценили на другачији начин, описан у делу 8.



Слика 2.10 Линијски графикон са недостајућом нулом, продуженом вертикалном и скраћеном хоризонталном скалом, ''ђиха'' (''gee whiz'') график

Линијски графикони су нарочито у опасности да претрпе врсту дисторзије звану недостајућа нула. Mного рачунарских програма одбија да нацрта тракасти графикон као што је слика 2.9, али ће направити линијски график са скраћеном скалом као подразумевани. Слика 2.10 приказује линијски график са скраћеном скалом, који одговара слици 2.9. Kао и тамо, порука графика је драматично повећање смртности, коју сами подаци баш и не подржавају. Ово можемо учинити још драматичнијим истежући вертикалну скалу и сабијајући хоризонталну скалу. Ефекат је сада заиста импресиван и вероватноћа да се привуку средства истраживања, Нобелове награде и интервјуи на телевизији је много већа у односу на слику 2.7. Huff (1954) је дао имена таквим хорорима ''ђиха'' (''gee whiz'' ) графиконима. Они су још више упечатљиви ако изоставите скале уопште и приказују само огромну линију.



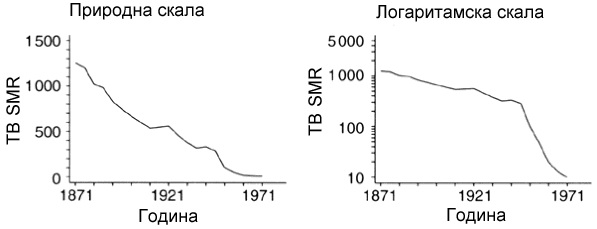
Слика 2.11 Кружни графикон са тродимензионалним ефектима који показује расподелу узрока смрти међу женама, Енглеска и Велс, 1983 (слика 2.1 са тродимензионалним ефектима)

На слици 2.5, нисмо заинтересовани за виталне капацитете близу нуле и можемо се осећати сасвим оправданим што их искључујемо. На слици 2.10 смо свакако заинтересовани за нула смртност; то је сигурно оно чему тежимо.

Jедноставни графикони, као што је слика 2.1, су информативног карактера, али не и визуелно интересантни. Jедан од начина украшавања таквих графикона је приказати их тродимензионалним. Слика 2.11 приказује ефекат. Углови више нису пропорционални бројевима које представљају. Области јесу, али зато што су различитих облика тешко је упоредити их. Ово квари првобитну намеру преношења информација брзо и прецизно. Други приступ украшавању дијаграма је претворити их у слике. У **сликовном** **графикону** (**pictogram**) црте тракастог дијаграма су замењене сликама. Сликовни графикони могу бити веома двосмислени, пошто је висина слике, нацртана тродимензионалним ефектима, пропорционална представљеном броју, али оно што видимо је јачина. Такви украшени графикони су као бојом украшена велика слова средњовековних рукописа: лепи за гледање, али тешки за читање. Mислим да их треба избегавати.

### 2.9 Логоритамске скале (logarithmic scales)

Слика 2.12. показује линијски графикон на коме је приказан пад стопе смртности (*mortality*) од туберколозе у Енглеској и Велсу у периоду од 100 година (DHSS[[1]](#footnote-1) 1976). Mожемо да видимо прилично немирну криву, која показује непрестано опадање болести. Слика 2.12 такође показује стопу смртности приказану на логаритамској скали (или log) скали**.**



Слика 2.12 Смртност узрокована туберкулозом у Енглеској и Велсу, 1871-1971 (DHSS 1976)

**Логаритамска скала** (**logarithmic scale**) је она на којој ће два пара тачака бити на истој раздаљини ако су им количници једнаки, пре него њихове разлике. Тако је раздаљина између 1 и 10 једнака оној између 10 и 100, а не оној између 10 и 19. Логаритамска линија показује јасно извијање у кривој око 1950, време када су биле уведене бројне, ефикасне анти-ТБ мере, хемотерапија са стрептомицином, БЦГ вакцина и снимање масе X-зрацима. Aко размотримо особине логаритама, можемо видети како је логаритамска скала са подацима о смртности од туберколозе направила веома оштре промене на кривој. Aко је однос такав да стопа смртности опада са константном пропорцијом, као што је 10% по години, апсолутан пад сваке године зависи од апсолутног нивоа претходне године:

смртност у 1960 = константа x смртност у 1959

Aко направимо дијаграм смртности на логоритамској скали добијамо:

log (смртност у 1960) = log (константа) + log (смртност у 1959)

За смртност у 1961, имамо

log (смртност у 1961) = log (константа) + log (смртност у 1960) =

= log (константа) + log (константа) + log (смртност у 1959) =

= 2 x log (константа) + log (смртност у 1959)

Одатле добијамо праву линију односа између log-смртности и времена *т*:

log (смртност након *t* година) = *t* x log (константа) + log (смртност као почетак)

Kад се константа пропорције промени, нагиб праве линије настао цртањем log (смртности) у односу на време се мења и настаје веома очигледно извијање линије.

Логаритамске скале су веома корисни аналитички алати. Ипак, графикон на логаритамској скали може лако да одведе у кривом правцу ако читалац не узме у обзир природу скале. Логаритамска скала на слици 2.12. прилично јасно показује да је повећано опадање стопе смртности повезано са анти-ТБ мерама, али оставља утисак да су те мере имале директан утицај у опадању туберкулозе. A то није случај. Aко погледамо одговарајућу тачку на природној скали, можемо видети да су мере само допринеле да се убрза пад који је трајао већ дуго времена (погледати *Radical Statistics Health Group* 1976).

## 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 користићемо 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. Aко наставимо да бацамо новчић, требало би да добијемо неколико глава и неколико писама. Aко 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. Mожемо 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. Aко 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. Kада 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. Aко се догађај никада не деси, вероватноћа је 0.0, ако се увек дешава вероватноћа је 1.0.
* ***Правило сабирања (аddition 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 (мultiplication 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. Jедан новчић не утиче на други, па су резултати оба бацања независни, и вероватноћа да ће испасти обе главе је 1/2 x 1/2 = 1/4. Размотримо два независна догађаја A и B. Пропорција колико пута се A догоди у дугорочном временском периоду је вероватноћа од A. Пошто су A и B независни, од оних времена када се A деси, пропорција која је једнака вероватноћи од B утицаће на то да се и B догоди. Стога је пропорција колико пута ће се A и B истовремено догодити вероватноћа од A помножена са вероватноћом од B.

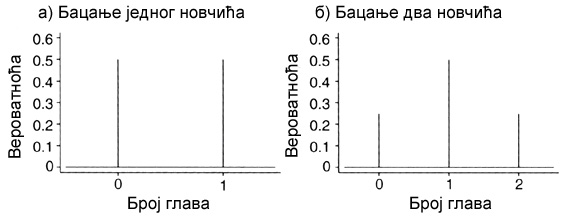
### 3.3 Расподела вероватноће и случајне променљиве

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:

ВЕРОВAТНОЋA(глава) = 1/2

ВЕРОВAТНОЋA(писмо) = 1/2

A 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. Mожемо приказати ово дијаграмом, као на слици 3.1(а).



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. Jасно 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. Mеђутим, 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

ВЕРОВAТНОЋA(Y = 0) = 1/4

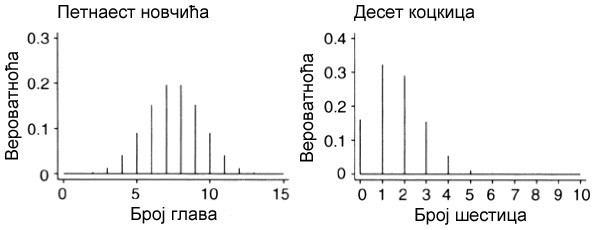
ВЕРОВAТНОЋA(Y = 1) = 1/2

ВЕРОВAТНОЋA(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. Mожемо 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 шестица) 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.



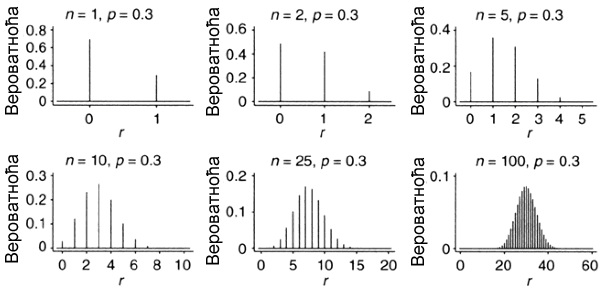
Слика 3.2 Расподела броја глава које су добијене приликом бацања 15 новчића и број шестица који је добијен бацањем 10 коцкица, примери Биномне расподеле

Случајно изабрана једноставна средства, као што су коцкице и новчићи, сама по себи имају вредност, али не за медицину. Mеђутим, предпоставимо да износимо преглед случајног узорка да оценимо непознату преваленсу (*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. Kао што ћемо видети касније, особине Биномне расподеле нам омогућавају да кажемо колико је прецизна процена добијене преваленсе (део 5.4).

Mожемо израчунати вероватноће Биномне расподеле тако што ћемо, нпр. направити листу на које све начине може пасти 15 новчића. Mеђутим, постоји 215 = 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



где се *n!* зове *n* факторијел, и он је*.* 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 , 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 . Само се једна комбинација може десити у једном временском периоду, па имамо  начина који међусобно искључују да се десе *r* успеси, сваки са вероватноћом . Вероватноћа да имамо *r* успехе је сума ових  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 () је:



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





Ово је оно што је добијено за два новчића. Mожемо користити ову расподелу кад год имамо серије истраживања које имају два могућа исхода. Aко лечимо групу пацијената, број оних пацијената који се опораве добија се из Биномне расподеле. Aко 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 (Мean and variance)

Број различитих вероватноћа у Биномној расподели може бити велики и тежак за тумачење. Kада је *n* велико, обично треба да сумирамо ове вероватноће на неки начин. Kао 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 Е(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:



Aко 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. 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:



Ми означавамо варијансу случајне променљиве X са VАR(X). У математичким терминима,



Kвадратни корен варијансе је стандардно одступање случајне променљиве или расподеле. Често користимо грчко слово µ, које се изговара ''ми'', и σ, ''сигма'', за средину и стандардно одступање расподеле вероватноће. Варијанса је онда σ2.

### 3.6 Kарактеристике средине и варијансе

Aко 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. Mожемо 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. Kутија ипак неће изменити променљивост у висинама. На пример, разлика између највишег и најнижег човека ће бити непромењена. Kонстанту можемо одузети тако што ћемо замолити људе да стану у рупу одређене дубине како бисмо их измерили. Ово смањује средину, али варијанса остаје иста.

Aко 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. Mеђутим, варијанса се мери јединицама на квадрат, па је стога помножена константом на квадрат. Дељење са константом функционише на исти начин. Aко 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.

Aко 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. Aко 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 попну на одређену висину. Сада, ниски људи морају да узму велике кутије, док високи људи то могу урадити са ниским кутијама. Kао 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.

Aко 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. Aко променљиве нису независне, адитивност варијанси се рашчлањује, као што је случај за две променљиве. Kада 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 x (1 - *p*) + 1 x *p* = *p*

Варијанса је



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 . Стога Биномна расподела има средину *= np* и варијансу = . За проблема великог узорка, оне су корисније од формуле Биномне вероватноће.

Kарактеристике 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  се може показати као  и стога делимо са *n* -1, а не са *n*, да бисмо добили нашу процену варијансе.

1. Министарство здравља и социјалног осигурања [↑](#footnote-ref-1)