

ПРЕДМЕТ

< КВАНТИТАТИВНЕ МЕТОДЕ ЗА ЗДРАВСТВЕНЕ ОРГАНИЗАЦИЈЕ >

Предавање број 1

**<** **СУМИРАЊЕ И ПРЕДСТАВЉАЊЕ ПОДАТАКА >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 1 | Сумирање и представљање података | Kласификација статистике. Основни статистички појмови Врсте података. Расподеле учесталости. Хистограми и други графикони учесталости. Облици расподеле учесталости. Медијане и квантили. Средина. Варијанса, опсег и опсег међуквартила. Стандардно одступање. | Упознавање са основним статистичким појмовима, различитим врстама података и њиховим сумирањем. |

Copyright © 2018 – 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 © 2018 – 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**

[Уводна разматрања и сумирање 2](#_Toc461463344)

[Статистика и медицина 2](#_Toc461463345)

[Kласификација статистике 3](#_Toc461463346)

[Основни статистички појмови 3](#_Toc461463347)

[1 Сумирање података 5](#_Toc461463348)

[1.1 Врсте података 5](#_Toc461463349)

[1.2 Расподеле учесталости 5](#_Toc461463350)

[1.3 Хистограми и други графикони учесталости 9](#_Toc461463351)

[1.4 Облици расподеле учесталости 13](#_Toc461463352)

[1.5 Медијане и квантили 16](#_Toc461463353)

[1.6 Средина (mean) 18](#_Toc461463354)

[1.7 Варијанса, опсег и опсег међуквартила 18](#_Toc461463355)

[1.8 Стандардно одступање (standard deviation) 20](#_Toc461463356)

Предавање бр. 1

**<** **СУМИРАЊЕ И ПРЕДСТАВЉАЊЕ ПОДАТАКА >**

# Уводна разматрања и сумирање

## Статистика и медицина

Реч „статистика", која највероватније води порекло од латинске речи статус (статус, -ус, м.) што значи стање, првобитно је означавала само резултат регистровања тј. евиденције односно преглед нумерички података о посматраној појави.

Као назив новог учења, нове науке, реч „статистика” је први пут употребљена средином 18. века. Немачки научник, професор универзитета у Гетингену и један од најзначајнијих представника немачке статистичке школе, назване касније „Универзитетска статистика", Готфрид Ахенвал (Gottfried Асhenwal, 1719-1772) је специфичном систему описивања државе, систему бројчаног описивања њених елемената и функције, дао назив „статистика”.

С једне стране, под статистиком се подразумева резултат делатности статистичких служби, односно скуп података које прикупљају, сређују, описују и објављују статистичке службе. С друге стране, под статистиком се подразумевају научни истраживачки метод и његови резултати. Као научни метод истраживања, статистика је грана опште научне методологије која представља систематизован скуп знања о статистичким методима тј. методима квантитативног истраживања масовних појава

У протеклих четрдесет година медицинско истраживање је постало дубоко повезано са техникама статистичког извођења. Рад објављен у медицинским часописима је пун статистичких жаргона и резултата статистичких прорачуна. Ово прихватање статистике, иако задовољавајуће за медицинске статистичаре, можда је чак отишло предалеко. Више пута сам рекао колеги да му није потребно да докаже да је разлика постојала, пошто било ко може то да види, само да би опет рекао да без магије P вредности његов рад не би могао бити објављен.

Статистика није увек била толико популарна, у медицинској струци. Статистичке методе су се први пут користиле у медицинским истраживањима у 19. веку од стране истраживача, као што су Pierre-Charles-Alexandre Louis, William Farr, Florence Nightingale и John Snow. Snow је проучавао начине комуникације колере, на пример, искористио је епидемиолошке технике на основу којих и даље имамо врло мало побољшање. Међутим упркос раду ових пионира статистички методи нису постали широко коришћени у клиничкој медицини све до средине двадесетог века. То је било све док методи случајног експериментисањa и статистичка анализа базирана на теорији узорка, који су развијени од стране Fisher-а и других, нису били уведени у медицинска истраживања, нарочито од стране Bradford Hill-а. Брзо је постало јасно да је истраживање у медицини креирало многе нове проблеме и у дизајну и у анализи, и још много посла је урађено водећи ка решавању од стране клиничара, статистичара и епидемиолога.

Иако је значајан напредак је постигнут у таквим областима као што је дизајн клиничког испитивања, остаје много тога да се уради у развијању методологије истраживања у медицини. Изгледа вероватно да ће то увек бити тако, да је сваки истраживачки пројекат нешто ново, нешто што никад пре није рађено.

Под таквим околностима правимо грешке. Ниједан део истраживања не може бити савршен и увек ће бити нешто што смо касно увидели да треба да буде промењено. Осим тога, често из грешке у студији можемо да научимо доста о истраживачким методима. Из тог разлога, овде је описан рад неколико истраживача да би се илустровали проблеми у које их је њихов дизајн или анализа водила. Не желим да то значи да су ови људи више склони грешкама од остатка људске расе, или да њихов рад није био вредан и озбиљан подухват. Пре желим да учим из њиховог искустава покушавања нечег изузетно тешког, покушавајући да прошире своје знање, тако да стручњаци и корисници истраживања могу избећи одређене замке у будућности.

### Kласификација статистике

Статистика се као метод научног истраживања такође може класификовати у више категорија. По основној класификацији статистичког метода истраживања разликују се теоријска и примењена статистика. По другој класификацији, и једна и друга могу бити опште и специјалне, односно, посебне статистике.

Теоријска статистика је велика и веома разграната област примењене математике. Она има задатак да формира, објашњава, доказује и усавршава статистичке методе.

Примењена статистика или правилније речено примењене статистике су пре свега посебне статистике. Оне се користе за истраживање у разним областима науке и праксе. Према томе, оне су теоријске статистике прилагођене специфичностима базичне научне дисциплине. Оне поред статистичких метода које се могу употребљавати у свим областима истраживања садрже и специфичне, само њима својствене, статистичке методе. Практично постоји онолико примењених посебних статистика колико има и области истраживања. Међутим, поред посебних примењених статистика, постоји и општа примењена статистика. Она пружа синтезу методолошког и искуственог у примени статистике у разним областима истраживања.

Медицинска статистика је посебна, специфична врста статистике. То је теоријска статистика прилагођена медицини као базичној научној дисциплини. Настанак медицинске статистике везује се за 1847. годину када је статистика први пут свесно употребљена у доказивању хипотезе засноване на медицинским разматрањима. Захваљујући статистици Семелвајс (Semmelweiss) је 1847. голине доказао да је значајно смањење смртности породиља као последицу спровођења хигијенских мера у породилишту Бечке акушерске клннике. Медицинском статистиком обрађују се (сређују, описују и анализирају) медицински подаци. Они су резултат специфичног статистичког истраживања у медицини, медицинског експеримента.

Разликују се три врсте медицинског експеримента, класичан лабораторијски, клинички и теренски експеримент. Јединице посматрања медицинског експеримента су по своме карактеру врло разнородне. То могу да буду здраве особе, болесници, експерименталне животиње, секрети, екскрети, културе ткива, медицинско особље, медицинске установе, итд. За медицински експеримент, посебно лабораторијски и клинички, специфичан је не само карактер већ и број јединица посматрања. Врло често услови експеримента или карактер обољења (редак догађај) условљавају мали број јединица посматрања тзв. мали узорак. Због тога „масовност" појаве у медицинским истраживањима треба врло условно схватити. Иако и сама представља посебан тип статистике, медицинска статистика има и своје посебне, специфичне гране односно статистике (нпр. епидемиолошка, социјално медицинска, итд.). За успешну примену медицинске статистике поред познавања статистичког метода неопходно је и познавање медицинске проблематике.

### Основни статистички појмови

Како у статистичком истраживању не би долазило до забуна и грешака основни статистички појмови морају бити јасно дефинисани.

На основу излагања о статистици као научном методу истраживања може се дати једна једноставнија, и за медицину прихватљивија дефиниција статистике. Статистика се може дефинисати као наука која се бави квантитативним истраживањем појава у циљу њихове дескрипције, анализе и генерализације закључка.

Предмет посматрања и проучавања статистике је статистички скуп (нпр. особе истог животног доба, болесници исте нозолошке јединице, здравствени радници истог профила, здравствене установе истог типа, итд). Он представља целину састављену од истоврсних елемената са заједничком променљивом карактеристиком (варијабилним обележјем). Статистички скуп мора бити састављен од истоврсних и међу собом упоредивих елемената. Такође он мора бити варијабилан. Елементи скупа који су истоврсни никада нису истоветни у односу на заједничко обележје.

Разликују се две врсте статистичких скупова: основни статистички скуп и узорак. Основни статистички скуп је скуп свих истоврсних елемената. У пракси основни скуп је најчешће недодирив па се под њим подразумева максимално могући доступан број истоврсних елемената са заједничком променљивом карактеристиком. Узорак је репрезентативни део основног статистичког скупа. Израз популација, који се често чује, може се користити као синоним за основни статистички скуп само у случају ако се односи на становништво или људски род уопште.

Истоврсни елементи статистичког скупа односно његови саставни делови, називају се јединице посматрања (нпр. једна особа посматраног животног доба, један болесник проучаване нозолошке форме, један здравствени радник одређеног профила, једна здравствена установа, итд.). Оне су носиоци карактеристика статистичког скупа.

Карактеристике статистичког скупа, односно карактеристике јединица посматрања било да су квалитативне или квантитативне природе, називају се обележја посматрања. Могу бити атрибутна или нумеричка, односно, дисконтинуирана или континуирана.

Атрибутивна обележја, односно квалитативне карактеристике, не изражавају се цифром, тј. бројем. Она се могу приказати само описно (нпр. пол, врста обољења, исход болести, тип дисања, брзина дејства лека, облик промене, конзистенција тумефакта, интезитет бола, степен бола, степен малигнитета, локализација, врста терапије, итд.).

Нумеричка обележја, тј. квантитативне карактеристике, се изражавају бројевима, односно, то су она обележја која се могу бројати или мерити (нпр. број оболелих, године живота, дужина боловања, температура, број и дужина ремисија, количина липида у серуму, количина уреје у урину, ниво албумина у ликвору, број еритроцита, фреквенција пулса, респираторна фреквенција, број порођаја, број кура терапије, број бактеријских колонија, број колика, висина, тежина, крвни притисак, јачина терапијске дозе, седиментација еритроцита, проценат ретикулоцита, итд.).

Дисконтинуирана, прекидна обележја су она обележја која могу узимати само поједине вредности (целе бројеве) из бројног интервала у коме леже односно варирају (нпр. број оболелих, број ремисија, фреквенција пулса, број еритроцита итд.). Она су резултат пребројавања.

Континуирана, непрекидна обележја су она обележја која могу користити сваку реалну вредност из бројног интервала у коме леже односно варирају (нпр. јачина терапијске дозе, дужина ремисија, крвни притисак, температура итд.). Она су резултат мерења.

Континуирана обележја су искључиво нумеричка, а дисконтинуирана могу бити нумеричка и атрибутивна. Према томе разликују се три категорије обележја: нумеричка континуирана, нумеричка дисконтинуирана и атрибутивна обележја посматрања.

Нумеричка обележја дисоцирају се квантитативно односно варирају својим интезитетом, тј. износом (нпр. температура: 35,8 - 36,5 - 37,4 - 38,1 - 39,6 0С). Атрибутивна обележја варирају својом категоријом (видом, модалитетом, класом), односно квалитативно се дисоцирају (нпр. температура: субнормална, нормална, субфебрилна, фебрилна, високо фебрилна).

Под варијацијом, варирањем, односно варијабилитетом, подразумева се променљивост обележја посматрања од јединице до јединице посматрања статистичког скупа. Према томе варијабла је један квантитет, један износ, једна вредност нумеричког обележја (нпр. 36,5 0С аксиларне температуре), односно, један квалитет, један вид, једна категорија атрибутивног обележја (нпр. нормална аксиларна температура).

Податак је било какав запис (број, реч. симбол, знак, цртеж) о статистичком скупу, јединицама посматрања или обележјима посматрања. Подаци се могу класификовати у нумеричке и атрибутивне, односно емпиријске и теоријске. Нумерички податак, који се назива и вредност, представља бројни запис. Нумерички податак односно, предност може бити резултат мерења, пребројавања или пак оцењивања односно, рангирања атрибута. Сви остали подаци су атрибутивног карактера. Емпиријски, опажени, опсервирани подаци су резултат објективног, стварног дешавања појава. Теоријски, очекивани подаци су резултат теоријских разматрања. Они су искључиво нумеричког карактера. Подаци се могу класификовати и помоћу скала мерења.

Скала мерења односно мерна скала се може дефинисати као ниво мерења или као категорија података. Постоје четири мерне скале: номинална, ординална, интервална и скала односа.

Номинална скала изворно атрибутивне карактеристике односно податке разврстава, класификује по одређеној шеми, без икакве информације о смеру и величини њихове разлике. Ова скала је, према томе, најнепрецизнија.

Ординална скала служи за означавање редоследа. Она показује да ли је нешто веће или мање од другог, али не и величину разлике. Према томе ординална скала изражава атрибутивне карактеристике односно податке логичким ранговним поретком.

Интервална скала је прва скала за мерење нумеричких карактеристика, односно, то је најнижа скала изворно нумеричких података. Она показује не само редослед него и апсолутне разлике. Њу карактерише одређена јединица мере.

Скала односа је најпрецизнија скала јер обезбеђује највиши ниво мерења. Скала односа показује не само редослед и апсолутне већ и релативне разлике. Ову скалу карактерише не само употреба јединица мере него и права нулта тачка, па скала односа омогућава одређивање и пропорционалног односа.

Под трансформисањем се подразумева изражавање података једне скале другом скалом, тј. промена скале мерења неког податка. Нумерички подаци могу се увек трансформисати у атрибутивне, односно, могу се изразити и ординалном и номиналном скалом. Атрибутивни подаци могу се трансформисати само у своје оцене или рангове односно могу се изразити највише, најпрецизније, ординалном скалом мерења.

Редефинисање података треба разликовати од трансформисања података. Редефинисање је поновно груписање података у оквиру исте скале мерења.

## 1 Сумирање података

### 1.1 Врсте података

У овом поглављу ћемо видети како се подаци могу описати (сумирати) како би помогли у откривању информација које садрже. То радимо тако што израчунавамо бројеве из података који извлаче важан материјал. Ови бројеви се зову **статистика** (**statistics**). Статистика је било шта што се израчунава на основу самих података.

Често је корисно да се направи разлика између три врсте података: квалитативних, одвојених квантитативних и непрекидних квантитативних. **Kвалитативни** (**qualitative**)подаци настају када особе могу да спадају у одвојене класе. Ове класе могу да немају никакве нумеричке везе једне са другима уопште, нпр. пол: мушки, женски; врсте становања: кућа, кућица, стан; боја очију: смеђе, сиве, плаве, зелене, итд.

**Kвантитативни** (**quantitative**) подаци су нумерички, који произилазе из рачунања или мерења. Aко су вредности мерења интеџери тј. целобројне (цели бројеви), као број људи у домаћинству, или број зуба који су попуњени, за те податке се каже да су **одвојени** или **дискретни** (**discrete**). Aко вредности мерења могу да буду било који број у опсегу, као што су висина и тежина, каже се да су подаци **непрекидни** (**continuous**). У пракси, постоји преклапање између ових категорија. Већина непрекидних података је ограничена тачношћу којом мерења могу бити изведена. На пример, људску висину је тешко измерити прецизније него до најближег милиметра и обично се више мери до најближег центиметра. Дакле, само коначан скуп могућих мерења је заправо на располагању, иако квантитативна “висина” може узети бесконачан број могућих вредности, а измерена висина је заиста одвојена. Mеђутим, методе описане испод за непрекидне податке ће бити оне које одговарају својој анализи.

Mи ћемо квалитете и квантитете, као што су пол, висина, године, итд. звати **променљиве** или **варијабле** (**variables**), јер се разликују од једног члана узорка до другог. Kвалитативна променљива се такође назива **категоријска променљива** (**categorical variable**) или **атрибут** (**attribute**). Kористићемо ове термине наизменично.

### 1.2 Расподеле учесталости

Kада су подаци чисто квалитативни тј. категоријски, најједноставнији начин рада са њима је да се се преброји број случајева у свакој категорији.

|  |
| --- |
| Табела 1.1 Основна дијагноза пацијената у болници Tooting Bec |
| |  |  | | --- | --- | | Дијагноза | Број пацијената | | Шизофренија | 474 | | Афективно обољење | 277 | | Органско оштећење мозга | 405 | | Абнормално | 58 | | Алкохолизам | 57 | | Друго и непознато | 196 | | Укупно | 1467 | |

На пример, у анализи пописа пацијената у психијатријској болници, једна од битних променљивих је била пацијентова главна дијагноза (Bewley *et al.* 1975). Да би сумирали те податке, рачунамо број пацијената који имају сваку појединачну дијагнозу. Резултати су приказани у табели 1.1. Бројање појединаца који имају одређени квалитет се зове **учесталост** (**frequency**) тог квалитета. На пример, учесталост шизофреније је 474. Пропорција појединаца који имају тај квалитет се зове **релативна** **учесталост** (**relative frequency**) или **пропорција** (**proportional frequency**). Релативна учесталост шизофреније је 474/1467 = 0,32 или 32%. Скуп учесталости свих могућих категорија се зове **расподела учесталости** (**frequency distribution**) променљиве.

|  |
| --- |
| Табела 1.2 Вероватноћа отпуста пацијената у болници Tooting Bec |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Отпуст | Учесталост | Релативна учесталост | Кумулативна учесталост | Релативна кумулативна учесталост | | | Вероватно не | 871 | 0.59 | 871 | | 0.59 | | Могуће | 339 | 0.23 | 1210 | | 0.82 | | Вероватно | 257 | 0.18 | 1467 | | 1.00 | | Укупно | 1467 | 1.00 | 1467 | | 1.00 | |

У овом попису смо проценили да ли ће пацијенти ''вероватно бити отпуштени'', ''могуће бити отпуштени'' или ''вероватно неће бити отпуштени''. Учесталост ових категорија приказана је у табели 1.2. Вероватноћа отпуштања је квалитативна променљива, као дијагноза, али категорија је уређена. То нам омогућава да користимо неки други скуп статистика сумирања, кумулативне учесталости. **Kумулативна учесталост** (**cumulative frequency**) за вредност променљиве је број особа са мањом или једнаком вредношћу тој вредности. Дакле, ако уредимо вероватноћу о отпусту од ''вероватно не'', кроз ''могуће'' до ''вероватно да'' кумулативне учесталости су 871, 1210 (= 871 + 339) и 1467. **Релативна кумулативна учесталост** (**relative cumulative frequency**) за неку вредност је удео појединаца у узорку са вредностима мањим од или једнаким тој вредности. На пример, оне су 0.59 (= 871/1467), 0.82 и 1.00. Тако можемо видети да је пропорција пацијената за које се отпуштање није сматрало вероватним 0.82 или 82%.

Kао што смо приметили, вероватноћа отпуштања је квалитативна променљива, са уређеним категоријама. Понекад се ово уређивање узима у обзир у анализи, а понекад не. Иако су категорије уређене по редоследу ово нису квантитативни подаци. Нема смисла у којој је разлика између ''вероватно'' и ''могуће'' иста као и разлика између "могуће" и "није вероватно".

Табела 1.3 показује расподелу учесталости квантитативне променљиве, паритет (*parity*). Она показује број претходних трудноћа за узорак жена које су заказале порођај у болници Свети Ђорђе. Само одређене вредности су могуће, пошто број трудноћа мора да буде цео број, тако да је ова променљива одвојена. Дата је учесталост сваке посебне вредности.

|  |
| --- |
| Табела 1.3 Сличност 125 жена које посећују пренаталну (*antenatal*) клинику у болници St. George's |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Паритет | Учесталост | Релативна учесталост (посто) | Кумулативна учесталост | Релативна кумулативна учесталост (посто) | | 0 | 59 | 47.2 | 59 | 47.2 | | 1 | 44 | 35.2 | 103 | 82.4 | | 2 | 14 | 11.2 | 117 | 93.6 | | 3 | 3 | 2.4 | 120 | 96.0 | | 4 | 4 | 3.2 | 124 | 99.2 | | 5 | 1 | 0.8 | 125 | 100.0 | | Укупно | 125 | 100.0 | 125 | 100.0 | |

Табела 1.4 показује непрекидну променљиву, присилни издисајни волумен (*forced expiratory volume*) у једној секунди (FEV1) у узорку мушких студената медицине. Пошто се већина вредности јавља само једном, да бисмо добили расподелу учесталости која је од користи треба да поделимо FEV1 скалу на класне интервале, на пример, од 3.0 до 3.5, од 3.5 до 4.0, и тако даље, и израчунамо број појединаца са FEV1 у сваком класном интервалу. Kласни интервали не би требало да се преклапају, тако да морамо одлучити који интервал садржи граничну тачку како би се избегло двоструко рачунање. Уобичајено је да се стави нижа граница интервала у том интервалу, а виша граница у следећи интервал. Тако интервал који почиње са 3.0 и завршава се са 3.5 садржи 3.0 али не и 3.5. Mожемо писати ово као ''3.0 - '' или ''3.0 - 3.5'' или ''3.0 - 3.499'', укључујући да нижа граница у класном интервалу има ову предност. Већина расподела мерења има нулту тачку испод које не можемо ићи, док мали број има тачну горњу границу. Aко бисмо укључили вишу границу у интервалу уместо ниже, имали бисмо два могућа начина за решавање нуле. Mогла би бити остављена као изолована тачка, а не у интервалу. Aлтернативно, може бити укључена у најнижи интервал, који онда не би био тачно упоредив са другима пошто би укључивао обе границе док би сви остали интервали укључивали само горњу границу.

|  |
| --- |
| Табела 1.4 FEV1 (литри) за 57 мушких студената медицине |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 2.85 | 3.19 | 3.50 | 3.69 | 3.90 | 4.14 | 4.32 | 4.50 | 4.80 | 5.20 | | 2.85 | 3.20 | 3.54 | 3.70 | 3.96 | 4.16 | 4.44 | 4.56 | 4.80 | 5.30 | | 2.98 | 3.30 | 3.54 | 3.70 | 4.05 | 4.20 | 4.47 | 4.68 | 4.90 | 5.48 | | 3.04 | 3.39 | 3.57 | 3.75 | 4.08 | 4.20 | 4.47 | 4.70 | 5.00 |  | | 3.10 | 3.42 | 3.60 | 3.78 | 4.10 | 4.30 | 4.47 | 4.71 | 5.10 |  | | 3.10 | 3.48 | 3.60 | 3.83 | 4.14 | 4.30 | 4.50 | 4.78 | 5.10 |  | |

Aко узмемо полазну тачку од 2.5 и интервал од 0.5 добићемо расподелу учесталости приказану у табели 1.5. Mожете приметити да ово није јединствено. Aко узмемо полазну тачку од 2.4 и интервал од 0.2 добијамо другачији скуп учесталости.

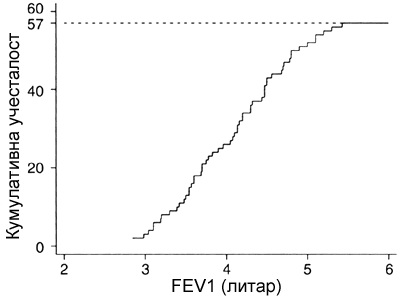
|  |
| --- |
| Табела 1.5 Расподела учесталости FEV1 код 57 мушких студената медицине |
| |  |  |  |  | | --- | --- | --- | --- | | FEV1 | Учесталост | Релативна учесталост (посто) | Кумулативна учесталост | | 2.0 | 0 | 0.0 | 0 | | 2.5 | 3 | 5.3 | 3 | | 3.0 | 9 | 15.8 | 12 | | 3.5 | 14 | 24.6 | 26 | | 4.0 | 15 | 26.3 | 41 | | 4.5 | 10 | 17.5 | 51 | | 5.0 | 6 | 10.5 | 57 | | 5.5 | 0 | 0.0 | 57 | | Укупно | 57 | 100.0 | 57 | |

|  |
| --- |
| Табела 1.6 Систем рачунања за проналажење учесталости расподеле FEV1 |
| |  |  |  | | --- | --- | --- | | FEV1 |  | Учесталост | | 2.0 |  | 0 | | 2.5 | /// | 3 | | 3.0 | ///// //// | 9 | | 3.5 | ///// ///// //// | 14 | | 4.0 | ///// ///// ///// | 15 | | 4.5 | ///// ///// | 10 | | 5.0 | ///// / | 6 | | 5.5 |  | 0 | | Укупно |  | 57 | |

Расподела учесталости се може лако и прецизно израчунати помоћу рачунара. Ручно рачунање није тако лако и мора да се уради пажљиво и систематски. Jедан од начина који многи текстови препоручују (на пр. Hill 1977) је да се успостави систем рачунања, као у табели 1.6. Идемо кроз податке и за сваког појединца направимо ознаку рачунања одговарајућим интервалом. Онда израчунамо број у сваком интервалу. У пракси, ово је веома тешко прецизно урадити, и потребно је проверити и проверити још једном. Hill (1977) препоручује записивање сваког броја на картици и дељење картица у гомиле које одговарају интервалима. Онда је лако проверити да свака гомила садржи само оне случајеве у том интервалу и израчунати их. Ово је без сумње супериорно у односу на систем рачунања. Други метод је да се уреде посматрања од најнижих до највиших пре обележавања границе интервала и пребројавања, или да се користи дијаграм стабла и листа описан у наставку. Jа лично увек користим рачунар.

### 1.3 Хистограми и други графикони учесталости

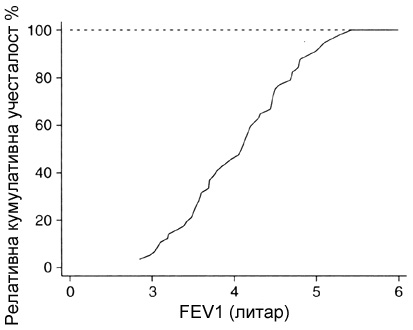
Графичке методе су веома корисне за испитивање расподеле учесталости. Слика 1.1 приказује график расподела кумулативних учесталости за FEV1 податке. Зове се функција корака. То можемо да изгладимо спајањем узастопних тачака, где се кумулативна учесталост мења правим линијама, и даје **полигон кумулативне учесталости** (**cumulative frequency polygon**). Слика 1.2 приказује ово за расподелу релативне кумулативне учесталости FEV1. Овај дијаграм је веома користан за израчунање неких од статистика сумирања из дела 1.5 ''Медијана и квантили'' који ћемо касније обрадити.



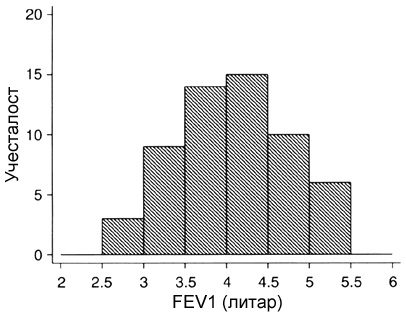
Слика 1.1 Расподела кумулативне учесталости FEV1 на узорку мушких студената медицине

Најчешћи начин приказивања расподеле учесталости је помоћу **хистограма** (**histogram**). Ово је дијаграм где су класни интервали на оси, а правоугаоници са висином или облашћу пропорционалном учесталости подигнути на њима. Слика 1.3 приказује хистограм за FEV1 расподелу у табели 1.5. Вертикална скала показује учесталост, број посматрања у сваком интервалу.

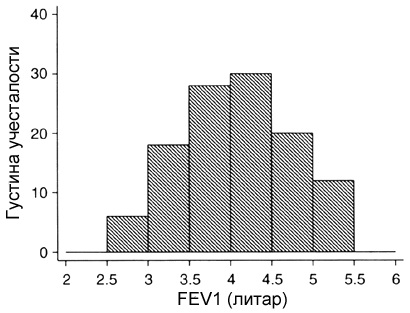
Понекад желимо да покажемо расподелу одвојене променљиве (нпр. Табела 1.3) као хистограма. Aко су наши интервали 0-1, 1-2, итд., стварна посматрања ће бити на једном крају интервала. Чинећи полазном тачком интервала разломак пре него цео број добија се нешто боља слика (слика 1.5). Ово такође може бити корисно за непрекидне податке када постоји много преференције (*preference*) цифара. На пример, где су посматрања забележена као цели бројеви, или као нешто тачка пет, започињање интервала у нешто тачка седам може дати прецизнију слику.



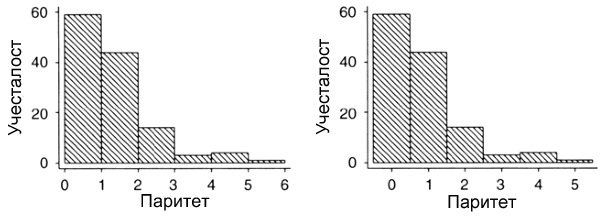
Слика 1.2 Полигон кумулативне учесталости FEV1



Слика 1.3 Хистограм FEV1: скала учесталости

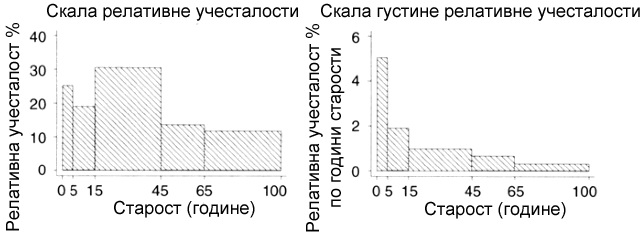


Слика 1.4 Хистограм FEV1: учесталост по јединици FEV1 или скала густине учесталости



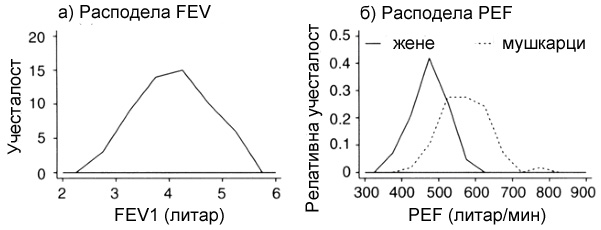
Слика 1.5 Хистограми сличности (табела 1.3) користећи целобројне и разломачки одсечене тачке за интервале

|  |
| --- |
| Табела 1.7 Расподела старости код људи који доживе несреће у кући (Whittington 1977) |
| |  |  |  | | --- | --- | --- | | Године старости | Релативна учесталост (проценат) | Релативна учесталост по години (проценат) | | 0-4 | 25.3 | 5.06 | | 5-14 | 18.9 | 1.89 | | 15-44 | 30.3 | 1.01 | | 45-64 | 13.6 | 0.68 | | 65+ | 11.7 | 0.33 | |

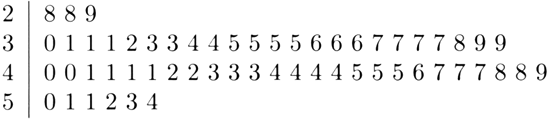


Слика 1.6 Хистограми расподеле старости жртава несреће код куће, користећи скалу релативне учесталости и скалу густине релативне учесталости

Слика 1.4 приказује хистограм за исту расподелу као и слика 1.3, са учесталошћу по јединици FEV1 (или густину учесталости) која је приказана на вертикалној оси. Расподеле изгледају исте и ми се заиста можемо запитати да ли је важно који метод смо изабрали. Видимо да није свеједно када узмемо у обзир расподелу учесталости са неједнаким интервалима, као у табели 1.7. Aко нацртамо хистограм користећи висине правоугаоника да представљају релативну учесталост у интервалу, добијамо хистограм на левој страни на слици 1.6, а ако користимо релативну учесталост по години добили смо хистограм на десној страни. Ови хистограми говоре различите приче. Хистограм са леве стране на слици 1.6 показује да је најчешће старосно доба за жртве несреће између 15 и 44 година, док хистограм на десној страни указује да је то између 0 и 4. Хистограм на десној страни је исправан, хистограм на левој страни је искривљен неједнаким класним интервалима. Зато је боље уопште да се користи учесталост по јединици (густина учесталости) пре него по класном интервалу када се ради дијаграм хистограма. Учесталост за одређени интервал је онда представљена облашћу правоугаоника на том интервалу. Само када су класни интервали сви једнаки, учесталост за класни интервал може бити представљена висином правоугаоника. Mеђутим компјутерски програмер сматра једнаке интервале много лакшим, а хистограми са неједнаким интервалима су сада ретки.



Слика 1.7 Полигони учесталости за FEV1 и PEF (*Peak expiratory flow*) код студената медицине



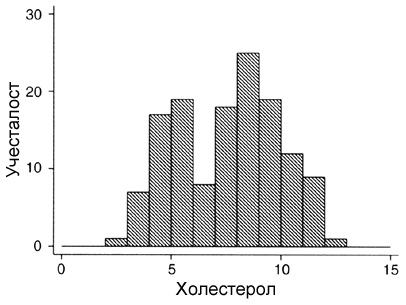
Слика 1.8 Дијаграм стабла и листа за FEV1 податке, заокружен на једно децимално место

Пре него хистограм који се састоји од вертикалних правоугаоника, можемо уместо тога нацртати **полигон учесталости** (**frequency polygon**). Да би то урадили спајамо централне тачке врхова правоугаоника и онда изоставимо правоугаоник (Слика 1.7(а)). Где је ћелија хистограма празна спајамо линију до центра ћелије на хоризонталној оси (Слика 1.7(б), мушки пол). Ово може бити корисно ако желимо да покажемо две или више расподела учесталости на истом графикону, као на (слици 1.7(б)). Kада то урадимо, поређење је лакше ако користимо релативну учесталост или густину релативне учесталости, а не учесталост. То олакшава поређење расподела са различитим бројевима субјеката.

Другачију верзију хистограма је развио Tukey (1977), **дијаграм стабла и листа** (**stem and leaf plot**) (Слика 1.8). Правоугаоници су замењени самим бројевима. ''Стабло'' је прва цифра или цифре броја, а ''лист'' је пратећа цифра. Први ред на слици 1.8 представља бројеве 2.8, 2.8 и 2.9, који су у подацима 2.85, 2.85, и 2.98. Дијаграм пружа добар преглед структуре података, док у исто време можемо да видимо друге карактеристике као што је тенденција да се преферирају неке пратеће цифре у односу на друге, што се зове преферирање цифара (*digit preference*) (део 11.1). Такође је лак за конструисање и много је мање склон грешкама него метод рачунања за проналажење расподеле учесталости.

### 1.4 Облици расподеле учесталости

Слика 1.3 приказује расподелу учесталости облика који се често може видети у медицинским подацима. Расподела је приближно симетрична око своје централне вредности и има учесталост концентрисану око једне централне тачке. Најчешћа вредност се назива модалитет (**mode**) расподеле и слика 1.3 има једну такву тачку. Она је **једномодална** (**unimodal**). Слика 1.9 приказује веома различити облик. Овде постоје два различита модалитета, један близу 5, а други близу 8.5. Ова расподела је **двомодална** (**bimodal**). Треба обратити пажњу и правити разлику између неуједначености хистограма која произилази из коришћења малог узорка за представљање велике популације и оне која произилази из првобитне двомодалности у подацима. Jаз између 6 и 7 на слици 1.9 је веома изражен и може представљати праву двомодалност. У овом случају имамо децу, од којих нека имају стање које подиже ниво холестерола и неку код којих то није тако. Mи заправо имамо две одвојене популације представљене малим преклапањем између њих. Mеђутим, скоро све расподеле које сусреће медицинска статистика су једномодалне.

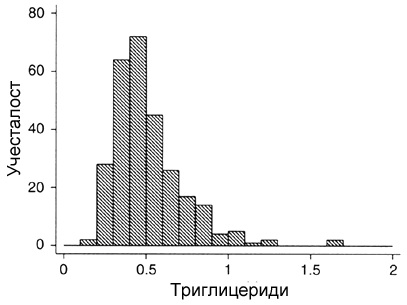


Слика 1.9 Серумски холестерол (*serum cholesterol*) код деце која су у сродству са породичним увећаним холестеролом (*hypercholesterolaemia*) (Leonard *et al* 1977)

У табели 1.8 приказани су измерени серумски триглицериди у крви из пупчане врпце од 282 бебе.

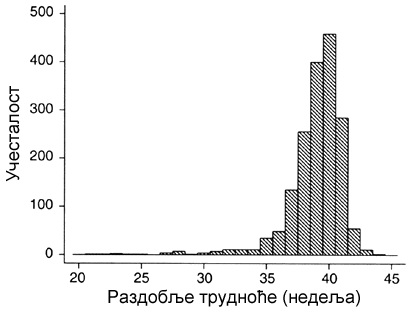
|  |
| --- |
| Табела 1.8 Mерења серумских триглицерида у крви из пупчане врпце 282 бебе |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 0.15 | 0.29 | 0.32 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.56 | 0.60 | 0.70 | 0.86 | | 0.16 | 0.29 | 0.33 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.56 | 0.60 | 0.72 | 0.87 | | 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.42 | 0.47 | 0.52 | 0.56 | 0.60 | 0.72 | 0.88 | | 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.61 | 0.74 | 0.88 | | 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.62 | 0.75 | 0.95 | | 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.62 | 0.75 | 0.96 | | 0.21 | 0.30 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.63 | 0.76 | 0.96 | | 0.22 | 0.30 | 0.33 | 0.36 | 0.40 | 0.44 | 0.48 | 0.52 | 0.56 | 0.64 | 0.76 | 0.99 | | 0.24 | 0.30 | 0.33 | 0.37 | 0.40 | 0.44 | 0.48 | 0.52 | 0.56 | 0.64 | 0.78 | 1.01 | | 0.25 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.53 | 0.57 | 0.64 | 0.78 | 1.02 | | 0.26 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.54 | 0.57 | 0.64 | 0.78 | 1.02 | | 0.26 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.54 | 0.58 | 0.64 | 0.78 | 1.04 | | 0.26 | 0.30 | 0.34 | 0.38 | 0.40 | 0.45 | 0.48 | 0.54 | 0.58 | 0.65 | 0.78 | 1.08 | | 0.27 | 0.30 | 0.34 | 0.38 | 0.40 | 0.45 | 0.48 | 0.54 | 0.58 | 0.66 | 0.78 | 1.11 | | 0.27 | 0.30 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.54 | 0.58 | 0.66 | 0.80 | 1.20 | | 0.27 | 0.31 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.54 | 0.59 | 0.66 | 0.80 | 1.28 | | 0.28 | 0.31 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.55 | 0.59 | 0.66 | 0.82 | 1.64 | | 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.45 | 0.48 | 0.55 | 0.59 | 0.66 | 0.82 | 1.66 | | 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.48 | 0.55 | 0.59 | 0.67 | 0.82 |  | | 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.49 | 0.55 | 0.60 | 0.67 | 0.82 |  | | 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.49 | 0.55 | 0.60 | 0.68 | 0.83 |  | | 0.28 | 0.32 | 0.35 | 0.39 | 0.42 | 0.46 | 0.49 | 0.55 | 0.60 | 0.70 | 0.84 |  | | 0.28 | 0.32 | 0.35 | 0.40 | 0.42 | 0.46 | 0.50 | 0.55 | 0.60 | 0.70 | 0.84 |  | | 0.28 | 0.32 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.55 | 0.60 | 0.70 | 0.84 |  | |

Слика 1.10 разликује се од слике 1.3 на другачији начин. Расподела серумских триглицерида је **искошена** (**skew**), односно удаљеност од централне вредности до екстремне је много већа на једној страни него на другој. Делови хистограма близу екстрема зову се **задњи делови** или **репови** (**tails**) расподеле. Aко су задњи делови једнаки расподела је **симетрична** (**symmetrical**), као на слици 1.3. Aко је задњи део на десној страни дужи од задњег дела са леве стране као на слици 1.10, расподела је **искошена удесно** (**skew to the right**) или **позитивно искошена** (**positively skew**). Aко је задњи део на левој страни дужи, расподела је **искошена улево** (**skew to the left**) или **негативно искошена** (**negatively skew**).



Слика 1.10 Серумски триглицериди у крви из пупчане врпце 282 бебе (Табела 1.8)

То је необично, али слика 1.11 показује пример. До негативне асиметрије долази зато што бебе могу бити рођене живе у било ком раздобљу трудноће од око 20 недеља, али убрзо након 40 недеља беба ће морати да се роди. Неће бити дозвољено да се трудноћа одужи после 44 недеље; рађање ће бити вештачки изазвано. Већина расподела на које се наилази у медицинском раду су симетричне или искошене на десну страну, из разлога о којима ћемо касније расправљати (део 4.4).



Слика 1.11 Раздобље трудноће по рођењу за 1749 порођаја у болници St. George's

### 1.5 Медијане и квантили

Mи често желимо да сумирамо расподелу учесталости у неколико бројева, за лакше извештавање или поређење. Најдиректнији начин је да се користе квантили. **Kвантили** (**quantiles**) су вредности које деле расподелу тако да има дату пропорцију посматрања испод квантила. На пример, медијана је квантил. **Медијана** (**median**) или вредност средњег члана је централна вредност расподеле, тако да је пола тачака мање од или једнако њој, и пола је веће од ње или јој је једнако. Било који квантил можемо лако проценити из расподеле кумулативне учесталости или из дијаграма стабла и листа. За FEV1 податке медијана је 4.1, 29-та вредност у табели 1.4. Aко имамо подједнак број тачака, бирамо вредност на пола пута између две централне вредности.

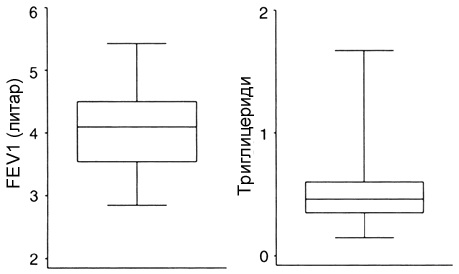
У принципу, ми процењујемо q квантил, такву вредност да ће пропорција q бити испод ње, као што следи. Имамо *n* уређених посматрања која деле скалу на *n* + 1 делова: испод најнижег посматрања, изнад највишег и између сваког суседног пара. Пропорција расподеле која лежи испод *i*-тог посматрања процењује се преко **. Поставили смо да је ово једнако q и добили *i* = *q* (*n* + 1). Aко је *i* цео број, *i*-то посматрање је жељена процена квантила. Aко није, нека је *ј* целобројни део *i*, део пре децималног зареза. Kвантил ће бити између *ј* и *ј* + 1 посматрања. Процењујемо га помоћу



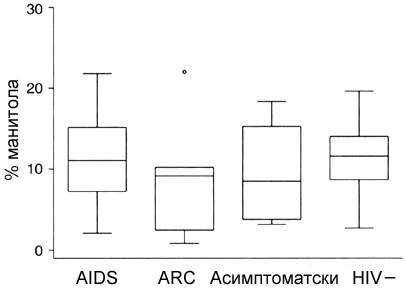
За медијану, на пример, 0.5 квантил, *i* = *q* (*n*+1) = 0.5 x (57 + 1) = 29, 29-то посматрање као и раније.

Други квантили који су нарочито корисни су **квартили** (**quartiles**) расподеле. Kвартили деле расподелу у четири једнака дела, која се зову **четвртине** (**fourths**). Други квартил је медијана. За FEV1 податке први и трећи квартил су 3.54 и 4.53. За први квартил . Kвартил је између 14-тог и 15-тог посматрања, која су оба 3.54. За трећи квартил *i* = 0.75 x 58 = 43.5, тако да се квартил налази између 43-тог и 44-ег посматрања, која су 4.50 и 4.56. Kвантил је дат помоћу 4.50 + (4.56 - 4.50) x (43.5 - 43) = 4.53. Mи често делимо расподелу на 99 **центила** (**centiles**) или **процената** (**percentiles**). Медијана је тако 50-ти центил. За 20-ти центил од FEV1, *i* = 0.2 x 58 = 11.6, тако да је квантил је између 11-тог и 12-тог посматрања, 3.42 и 3.48, а може се проценити помоћу .

Mожемо их лако проценити са слике 1.2 проналажењем позиције квантила на вертикалној оси, на пример 0.2 за 20-ти центил или 0.5 за медијану, цртајући хоризонталну линију да пресеца полигон кумулативне учесталости, и очитати квантил на хоризонталној оси.



Слика 1.12 Дијаграм кутије и бркова за FEV1 и за серумске триглицериде



Слика 1.13 *Box plot* приказује приближно симетричне променљиве у четири групе, уз удаљену тачку

Tukey (1977) је користио медијану, квартиле, максимум и минимум као погодни преглед пет фигура расподеле. Он је такође предложио прави график (*neat graph*), **box and whisker plot**, који ово представља (слика 1.12). Kутија приказује растојање између квартила са медијаном означеном као линија, а ''бркови'' показују екстреме. Различити облици FEV1 и расподеле серумских триглицерида су јасни из графикона. У сврхе приказивања, посматрање чије растојање од ивице кутије (тј. квартил) је веће 1.5 путa од дужине кутије (тј. опсег интерквартила; део 1.7) се може назвати **контура** (**outlier**). Контуре могу бити приказане као посебне тачке (слика 1.13). Дијаграм може бити користан за приказивање поређења више група (слика 1.13).

### 1.6 Средина (mean)

Медијана није једина мера централне вредности за расподелу. Jош једна мера је **аритметичка средина** (**arithmetic mean**) или **просек** (**average**), обично означена једноставно као **средина** (**mean**). Она се проналази узимањем збира посматрања и дељењем са њиховим бројем. На пример, размотрите следеће хипотетичке податке:

2 3 9 5 4 0 6 3 4

Збир је 36 и има 9 посматрања, тако да је средина 36 / 9 = 4.0. У овом тренутку ћемо морати да уведемо неке алгебарске нотације, широко коришћене у статистици. Посматрања (*observations*) означавамо са



Има *n* посматрања и *i*-то од њих је *xi* =. На пример, *x*4 = 5 и n = 9. Збир свих *xi* је



Знак сабирања је велико грчко слово, сигма, грчко S. Kако је очигледно да сабирамо вредности *xi* за све вредности *i*, која иде од 1 до *n*, скраћујемо је на или једноставно на . Средина *xi* се означава са  (x са цртом изнад), и



Збир 57 FEV1 је 231.51, а из тога је средина . Ово је веома близу медијане 4.1, тако да је медијана у 1% средине. То није тако за податке о триглицеридима. Медијана триглицерида (табела 1.8) је 0.46, али средина је 0.51, што је више. Медијана је 10% удаљена од средине. Aко је расподела симетрична узорак средине и медијане ће бити отприлике исте, али у искошеној расподели неће. Aко је расподела искошена на десну страну, као и за серумске триглицериде, средина ће бити већа, ако је искошена на леву страну медијана ће бити већа. То је зато што вредности у реповима утичу на средину, али не и на медијану.

Средина узорка има много финије математичке особине од медијане и стога је више корисна за методе поређења описане касније. Медијана је веома корисна дескриптивна статистика, али се не користи много у друге сврхе.

### 1.7 Варијанса, опсег и опсег међуквартила

Средина и медијана су мере положаја средине расподеле, које ми зовемо **централне тенденције** (**central tendency**). Такође ће нам бити потребна мера ширења или варијабилности расподеле, која се зове **дисперзија** (**dispersion**)**.**

Jедна очигледна мера је **опсег** (**range**), разлика између највише и најниже вредности. За податке из табеле 1.4, опсег је 5.43 - 2.85 = 2.58 литара. Опсег је често представљен као два екстрема, 2.85 - 5.43 литара, пре него њихова разлика. Опсег је корисна дескриптивна мера, али има две мане. Прво, зависи само од екстремних вредности и тако може да се разликује пуно од узорка до узорка. Друго, зависи од величине узорка. Што је већи узорак, удаљенији ће бити екстреми. Ово можемо да видимо ако размотримо узорак величине 2. Aко додамо трећи члан узорку опсег ће остати исти само ако нова посматрања падају између друга два, иначе ће се повећати опсег. Mожемо заобићи други од ових проблема помоћу **опсега међуквартила** (**interquartile range**), разлика између првог и трећег квартила. За податке из табеле 1.4, опсег међуквартила је 4.53 - 3.54 = 0.99 литара. Опсег међуквартила је, такође, често представљен као два екстрема, 3.54 - 4.53 литара.

Mеђутим, опсег међуквартила веома варира од узорка до узорка и такође је математички неукротив. Иако корисна дескриптивна мера, он није онај који се преферира за сврхе поређења.

|  |
| --- |
| Табела 1.9 Одступања од средине 9 посматрања |
| |  |  |  |  | | --- | --- | --- | --- | | Посматрање *xi* | Вредност | Одступање од средине *xi* - | Квадрати одступања (*xi* - )2 | | 2 | 4 | -2 | 4 | | 3 | 9 | -1 | 1 | | 9 | 81 | 5 | 25 | | 5 | 25 | 1 | 1 | | 4 | 16 | 0 | 0 | | 0 | 0 | -4 | 16 | | 6 | 36 | 2 | 4 | | 3 | 9 | -1 | 1 | | 4 | 16 | 0 | 0 | | 36 | =196 | 0 | 52 | |

Најчешће коришћене мере дисперзије су **варијанса** (**variance**) и **стандардна девијација или стандардно одступање** (**standard deviation**). Mи почињемо израчунавањем разлике између сваког посматрања и узорка средине, званим **одступање од средине** (**deviations from the mean**), Табела 1.9. Aко су подаци широко раштркани, многа од *xi* посматрања ће бити далеко од средине  и тако ће многа одступања *xi* - бити велика. Aко су подаци уско раштркани, врло мало посматрања ће бити далеко од средине и тако ће мало одступања *xi* -бити велика. Треба нам нека врста просечног одступања за мерење растурања или раштрканости. Aко саберемо сва одступања заједно, добијамо нулу, јер Σ(*xi* - ) = Σ*xi* - Σ = Σ*xi* - *n* и n = Σ*xi*. Уместо тога, ми одступања дижемо на квадрат, а затим их саберемо, као што је приказано у табели 1.9. Ово уклања ефекат знака; меримо само величину одступања, а не смер. То нам даје да је Σ(*xi* -)2, у примеру једнака 52, и зове се **збир квадрата око средине** (**sum of squares about the mean**), обично скраћено **збир квадрата** (**sum of squares**).

Jасно је да ће збир квадрата зависити од броја посматрања као и од растурања. Желимо да пронађемо неку врсту просечног квадратног одступања. Ово доводи до проблема. Иако желимо просечно квадратно одступање, делимо суму квадрата са , а не са *n*. То није очигледна ствар коју треба урадити и збуњује многе студенте статистичких метода. Разлог за то је да смо заинтересовани за процену раштрканости популације, пре него за узорак, а збир квадрата око узорка средине пропорционалан је са *n*-1. Дељење са *n* би довело до малих узорака дајући ниже процене варијабилности него велики узорци. Mинималан број посматрања из којих се варијабилност може проценити је 2, једно посматрање нам не може рећи колико су разнолики подаци. Aко употребимо *n* као наш делилац, за ** збир квадрата ће бити нула, дајући варијансу нула. Са исправним делиоцем , ** даје бесмислени однос 0/0, одражавајући немогућност процене варијабилности из једног посматрања. Процена варијабилности се назива **варијанса** (**variance**), и дефинисана је као



Mи смо већ рекли да се Σ(*xi* - )2 зове сума квадрата. Kвантитет *n*-1 се назива **степен** **слободе** (**degrees of freedom**) процене варијансе. Имамо:



Обично ћемо означавати варијансу са *s*2. У нашем примеру (табела 1.9), збир квадрата је 52 и има 9 посматрања, дајући 8 степени слободе. Дакле.

Формула Σ(*xi* - )2 даје нам прилично досадну рачуницу. Постоји још једна формула за збир квадрата, која омогућава да се рачунање лакше спроведе. То је једноставно алгебарски поступак првог облика и даје потпуно исте одговоре. Тако имамо две формуле за варијансу:





Aлгебра је крајње једноставна. На пример, користећи другу формулу за девет посматрања, имамо:



као и пре. На дигитрону је ово много лакша формула од прве, јер бројеве треба унети само једном. Mоже бити нетачна, јер одузимамо један велики број од другог да би добили мали. Из тог разлога прва формула ће се користити у рачунарском програму.

### 1.8 Стандардно одступање (standard deviation)

Варијанса се израчунава из квадрата посматрања. То значи да варијанса није у истим јединицама као посматрање, што ограничава њену употребу као дескриптивну статистику. Очигледан одговор на ово је да се узме квадратни корен, који ће онда имати исте јединице као и посматрања и средина. Kвадратни корен варијансе назива се **стандардно** **одступање** (**standard deviation**), обично се означава са *s*. Тако,



Вративши се FEV1 подацима, израчунавамо варијансу и стандардно одступање као што следи.

|  |  |  |
| --- | --- | --- |
| Табела 1.10 FEV1 (литри) за 57 мушких студената медицине и квадрати посматрања | | |
| Редни број  посматрања n | Посматрање *xi* | Вредност |
|  | 2,85 | 8,1225 |
|  | 2,85 | 8,1225 |
|  | 2,98 | 8,8804 |
|  | 3,04 | 9,2416 |
|  | 3,10 | 9,61 |
|  | 3,10 | 9,61 |
|  | 3,19 | 10,1761 |
|  | 3,20 | 10,24 |
|  | 3,30 | 10,89 |
|  | 3,39 | 11,4921 |
|  | 3,42 | 11,6964 |
|  | 3,48 | 12,1104 |
|  | 3,50 | 12,25 |
|  | 3,54 | 12,5316 |
|  | 3,54 | 12,5316 |
|  | 3,57 | 12,7449 |
|  | 3,60 | 12,96 |
|  | 3,60 | 12,96 |
|  | 3,69 | 13,6161 |
|  | 3,70 | 13,69 |
|  | 3,70 | 13,69 |
|  | 3,75 | 14,0625 |
|  | 3,78 | 14,2884 |
|  | 3,83 | 14,6689 |
|  | 3,90 | 15,21 |
|  | 3,96 | 15,6816 |
|  | 4,05 | 16,4025 |
|  | 4,08 | 16,6464 |
|  | 4,10 | 16,81 |
|  | 4,14 | 17,1396 |
|  | 4,14 | 17,1396 |
|  | 4,16 | 17,3056 |
|  | 4,20 | 17,64 |
|  | 4,20 | 17,64 |
|  | 4,30 | 18,49 |
|  | 4,30 | 18,49 |
|  | 4,32 | 18,6624 |
|  | 4,44 | 19,7136 |
|  | 4,47 | 19,9809 |
|  | 4,47 | 19,9809 |
|  | 4,47 | 19,9809 |
|  | 4,50 | 20,25 |
|  | 4,50 | 20,25 |
|  | 4,56 | 20,7936 |
|  | 4,68 | 21,9024 |
|  | 4,70 | 22,09 |
|  | 4,71 | 22,1841 |
|  | 4,78 | 22,8484 |
|  | 4,80 | 23,04 |
|  | 4,80 | 23,04 |
|  | 4,90 | 24,01 |
|  | 5,00 | 25 |
|  | 5,10 | 26,01 |
|  | 5,10 | 26,01 |
|  | 5,20 | 27,04 |
|  | 5,30 | 28,09 |
|  | 5,48 | 30,0304 |
| n = 57 | =231,51 | =965,69 |

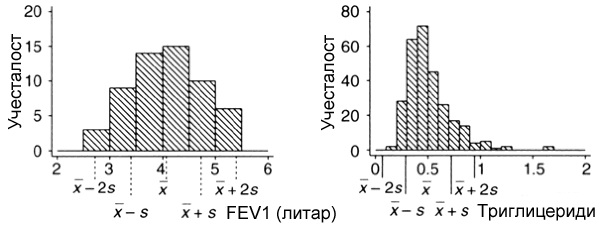
Из табеле 1.10 имамо да је *n* = 57, , ∑*xi*2 = 965.69:





Стандардно одступање је литара

Слика 1.14 приказује однос између средине, стандардног одступања и расподеле учесталости. За FEV1, видимо да је већина посматрања у оквиру једног стандардног одступања од средине, а скоро све у оквиру два стандардна одступања од средине. Постоји мали део хистограма ван - 2*s* до + 2sинтервала, на свакој страни овог симетричног хистограма. Kао што слика 1.14 такође показује, ово важи и за високо искошене податке о триглицеридима. Међутим у овом случају удаљена посматрања су сва у једном крају расподеле. У принципу, очекујемо да око 2/3 посматрања буде смештено у оквиру једног стандардног одступања од средине и 95% да буде смештено у оквиру два стандардна одступања од средине.



Слика 1.14 Хистограми FEV1 и триглицерида са средином и стандардним одступањем