

ПРЕДМЕТ

< МЕДИЦИНСКА СТАТИСТИКА И ИНФОРМАТИКА >

Предавање број 9

**<** **СУМИРАЊЕ >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 9 | Сумирање | Облици расподеле учесталости. Медијане и квантили. Средина. Варијанса, опсег и опсег међуквартила. Стандардно одступање. | Упознавање са различитим врстама података и њиховим сумирањем. |

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**

[Сумирање 2](#_Toc284064840)

[1.4 Облици расподеле учесталости 2](#_Toc284064841)

[1.5 Медијане и квантили 5](#_Toc284064842)

[1.6 Средина (mean) 6](#_Toc284064843)

[1.7 Варијанса, опсег и опсег међуквартила 7](#_Toc284064844)

[1.8 Стандардно одступање (standard deviation) 9](#_Toc284064845)

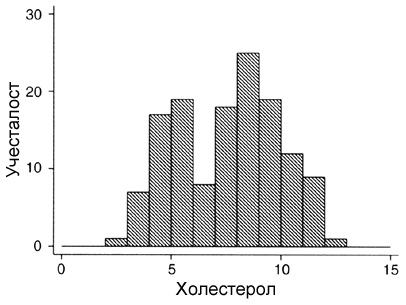
Предавање бр. 9

**<** **СУМИРАЊЕ >**

# Сумирање

### 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.

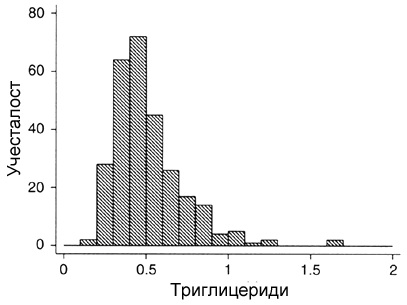


Слика 1.9 Серумски холестерол (*serum cholesterol*) код деце која су у сродству са породичним увећаним холестеролом (*hypercholesterolaemia*) (Leonard *et al* 1977)

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

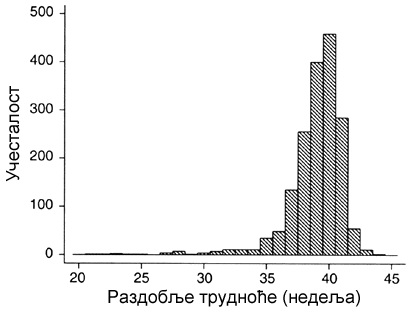
|  |
| --- |
| Табела 1.8 Mерења серумских триглицерида у крви из пупчане врпце 282 бебе |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 0.15 | 0.29 | 0.32 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.56 | 0.60 | 0.70 | 0.86 | | 0.16 | 0.29 | 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.

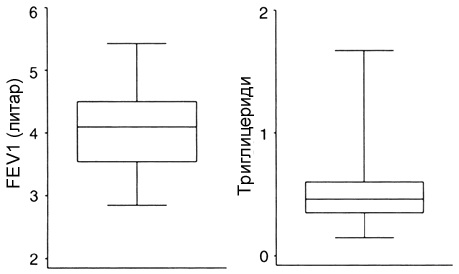
У принципу, ми процењујемо q квантил, такву вредност да ће пропорција q бити испод ње, као што следи. Имамо *n* уређених посматрања која деле скалу на *n* + 1 делова: испод најнижег посматрања, изнад највишег и између сваког суседног пара. Пропорција расподеле која лежи испод *i*-тог посматрања процењује 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



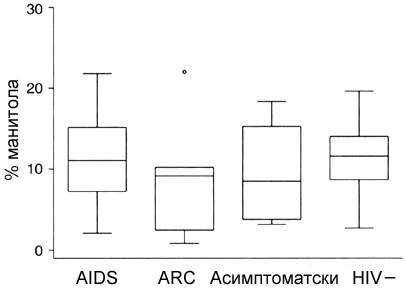
За медијану, на пример, 0.5 квантил, *i* = *q* (*n*+1) = 0.5 x (57 + 1) = 29, 29-то посматрање као и раније.

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 |

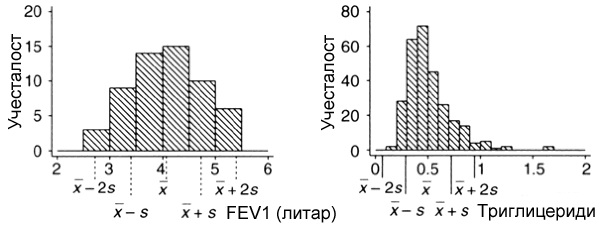
Из табеле 1.10 имамо да је *n* = 57, , ∑*xi*2 = 965.69:





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