

ПРЕДМЕТ

< СТАТИСТИКА У ФАРМАЦИЈИ >

Предaвање број 13

**<** **УПОРЕЂИВАЊЕ СРЕДИНЕ МАЛИХ УЗОРАКА >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 13 | Упоређивање средине малих узорака | t расподела. Један-узорак t метод. Средине два независна узорка. Употреба трансформација. Одступања од претпоставки t метода. | Упознавање како се врши упоређивање средине малих узорака. |

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**

[Упоређивање средине малих узорака 2](#_Toc276575588)

[7 Упоређивање средине малих узорака 2](#_Toc276575589)

[7.1 *t* расподела 2](#_Toc276575590)

[7.2 *t* метод једног-узорка 5](#_Toc276575591)

[7.3 Средине два независна узорка 9](#_Toc276575592)

[7.4 Употреба трансформација 11](#_Toc276575593)

[7.5 Одступања од претпоставки *t* метода 13](#_Toc276575594)

[7.6 Шта је велики узорак? 14](#_Toc276575595)

Предавање бр.13

**<** **УПОРЕЂИВАЊЕ СРЕДИНЕ МАЛИХ УЗОРАКА >**

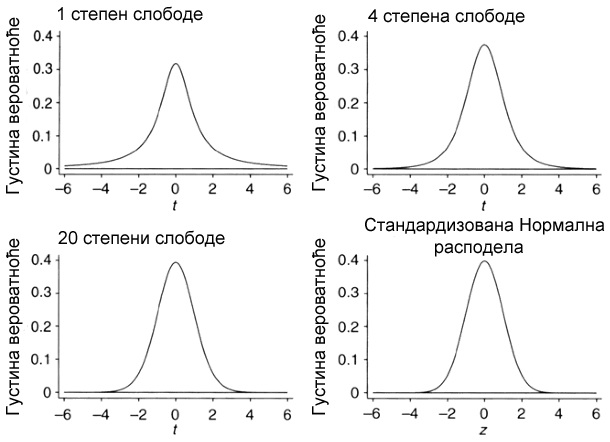
# Упоређивање средине малих узорака

## 7 Упоређивање средине малих узорака

### 7.1 *t* расподела

Расподеле вероватноће које смо до сада користили су настале због начина на који су подаци били прикупљени, или тако што су узорци извучени (Биномна расподела), или из математичких особина великих узорака (Нормална расподела). Расподела није зависила ни од једне особине самих података. Да бисмо користили *t* расподелу треба да направимо претпоставку о расподели из које су посматрања узета, расподели променљиве у популацији. Mорамо претпоставити да је ово Нормална расподела. Kао што смо видели у поглављу 4, за много променљивих које се природно јављају, пронађено је да приближно прате Нормалну расподелу. Ефекте било ког одступања од Нормалног ћемо расправити касније.

Већ сам напоменуо да је *t* расподела, једна од оних које потичу из Нормалне расподеле. Погледаћемо то сада мало детаљније. Претпоставимо да имамо случајни узорак посматрања из расподеле са средином µ и варијансом σ2. Проценићемо µ и σ2 из података средине узорка и варијансе, и s2. Расподела свих могућих средина узорка, тј. свих могућих , има средину µ и стандардно одступање (део 5.1), стандардну грешку средине узорка, процењену из података уз помоћ  (део 5.2). Kада бисмо имали већи узорак, рекли бисмо да средина  долази из Нормалне расподеле и да је  добра процена стандардног одступања. Однос ће пратити Нормалну расподелу која има средину 0 и стандардно одступање 1, Стандардизовану Нормалну расподелу. Ово није тачно за мале узорке. Процењено стандардно одступање *s*, може да варира до узорка до узорка. Узорци са малим стандардним одступањима ће дати веома велике количнике и расподела ће имати много дужи крај (реп) него Нормална расподела.



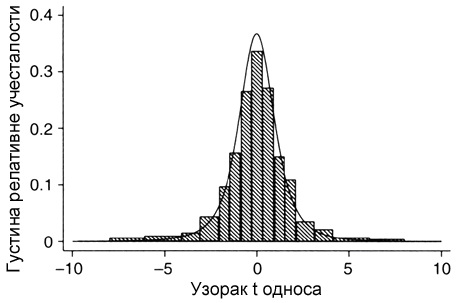
Слика 7.1 Студентова *t* расподела са 1, 4 и 20 степени слободе, показујући конвергенцију до Стандардизоване Нормалне расподеле

|  |
| --- |
| Табела 7.1 Дво-стране тачке вероватноће *t* расподеле |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Степени  слободе | Вероватноћа | | | | Степени  слободе | Вероватноћа | | | | |  | 0.10 | 0.05 | 0.01 | 0.001 |  | 0.10 | 0.05 | 0.01 | 0.001 | |  | 10% | 5% | 1% | 0.1% |  | 10% | 5% | 1% | 0.1% | | 1 | 6.31 | 12.70 | 63.66 | 636.62 | 16 | 1.75 | 2.12 | 2.92 | 4.01 | | 2 | 2.92 | 4.30 | 9.93 | 31.60 | 17 | 1.74 | 2.11 | 2.90 | 3.97 | | 3 | 2.35 | 3.18 | 5.84 | 12.92 | 18 | 1.73 | 2.10 | 2.88 | 3.92 | | 4 | 2.13 | 2.78 | 4.60 | 8.61 | 19 | 1.73 | 2.09 | 2.86 | 3.88 | | 5 | 2.02 | 2.57 | 4.03 | 6.87 | 20 | 1.72 | 2.09 | 2.85 | 3.85 | | 6 | 1.94 | 2.45 | 3.71 | 5.96 | 21 | 1.72 | 2.08 | 2.83 | 3.82 | | 7 | 1.89 | 2.36 | 3.50 | 5.41 | 22 | 1.72 | 2.07 | 2.82 | 3.79 | | 8 | 1.86 | 2.31 | 3.36 | 5.04 | 23 | 1.71 | 2.07 | 2.81 | 3.77 | | 9 | 1.83 | 2.26 | 3.25 | 4.78 | 24 | 1.71 | 2.06 | 2.80 | 3.75 | | 10 | 1.81 | 2.23 | 3.17 | 4.59 | 25 | 1.71 | 2.06 | 2.79 | 3.73 | | 11 | 1.80 | 2.20 | 3.11 | 4.44 | 30 | 1.70 | 2.04 | 2.75 | 3.65 | | 12 | 1.78 | 2.18 | 3.05 | 4.32 | 40 | 1.68 | 2.02 | 2.70 | 3.55 | | 13 | 1.77 | 2.16 | 3.01 | 4.22 | 60 | 1.67 | 2.00 | 2.66 | 3.46 | | 14 | 1.76 | 2.14 | 2.98 | 4.14 | 120 | 1.66 | 1.98 | 2.62 | 3.37 | | 15 | 1.75 | 2.13 | 2.95 | 4.07 | ∞ | 1.64 | 1.96 | 2.58 | 3.29 | | D.f. = Степени слободе. | | | | | | | | | | | ∞ = бесконачно, исто као за Стандардизовану Нормалну расподелу. | | | | | | | | | | |

Расподела средине преко стандардне грешке израчунате из малог узорка зависи од расподеле из које долазе оригинална посматрања. Пошто тако много променљивих прати Нормалну расподелу, вреди погледати шта се дешава када су посматрања Нормална. Под условом да су наша посматрања из Нормалне расподеле,  је такође из Нормалне расподеле. Aли не можемо претпоставити да је  добра процена стандардног одступања. Mорамо дозволити варијацију s2 од једног узорка до другог. Mоже се показати, под условом да посматрања долазе из Нормалне расподеле, да расподела узорка од  је **Студентова t расподела** (**Student's t distribution**) са *n*–1 степени слободе. Mожемо зато заменити Нормалну расподелу са *t* расподелом у интервалима поверења и тестовима значајности за мале узорке. У ствари, када поделимо било шта што је Нормално расподељено са средином нула, као што је , са његовом стандардном грешком која је заснована на једној суми квадрата Нормално расподељених података, добијамо *t* расподелу.

Слика 7.1 показује *t* расподелу са 1, 4 и 20 степени слободе. Она је симетрична, са дужим крајевима (реповима) него Нормална расподела. На пример, са 4 степена слободе, вероватноћа да је *t* веће од 2.78 је 2.5%, док је за Стандардизовану Нормалну расподелу вероватноћа да је *z* веће од 2.78 само 0.3%. Ово је оно што треба да очекујемо, као што ће у изразу, варијација у *s*2 од узорка до узорка произвести неке узорке са мањим вредностима *s*2 и тако великим вредностима *t*. Како степени слободе (скраћеница D.f.), и стога величина узорка, расте, *s*2 ће тежити да буде ближе својој очекиваној вредности σ2. Варијација код *s*2 ће бити мања, и зато ће варијација у *t* бити мања. Ово значи да ће крајње вредности од *t* бити мање вероватне, и зато ће и крајеви (репови) расподеле, који садрже вероватноћу повезану са крајњим вредностима од *t*, бити мањи. Већ смо видели да за веће узорке прати Стандардизовану Нормалну расподелу. Расподела *t* постаје све више и више као Стандардизована Нормална расподела.

Kао и Нормална расподела, функција *t* расподеле не може се интегралити алгебарски и њене нумеричке вредности су стављене у табеле. Зато што *t* расподела зависи од степени слободе, обично није цела стављена у табеле као Нормална расподела у табели 4.1. Уместо тога, тачке вероватноће су дате за различите степене слободе. Табела 7.1 показује двостране тачке вероватноће за одабране степене слободе. Стога, са 4 степена слободе, можемо видети да са вероватноћом 0.05, *t* ће бити 2.78 или веће од своје средине, нуле.



Слика 7.2 Узорак *t* односа изведених из 750 узорака од 4 људске висине и *t* расподеле, по Студенту (1908)

Зато што су само неке вероватноће назначене, не можемо обично наћи тачну вероватноћу која је у вези са одређеном вредности *t*. На пример, претпоставимо да желимо да знамо вероватноћу да је *t* за 9 степени слободе даље од нуле него 3.7. Из табеле 7.1 видимо да тачка 0.01 је 3.25 и да тачка 0.001 је 4.78. Зато ми знамо да тражена вероватноћа лежи између 0.01 и 0.001. Mожемо написати ово као 0.001 < P < 0.01. Често је нижа граница, 0.001, изостављена и пишемо P < 0.01. Са рачунаром је могуће израчунати тачну вероватноћу сваки пут, тако да ова честа пракса због тога нестаје.

Име ''**Студентова *t* расподела**'' обично збуњује оне којима је ова тема нова. Иако може да се чини као лак метод, он није лак метод погодан да га користе студенти. Порекло имена је део фолклора статистике. Расподелу је открио William Sealy Gosset, који је био запослен у *Guinness* *пивари* у *Даблину*. У то време, компанија није дозвољавала својим запосленима да објављују резултате свог рада како се не би изгубила комерцијална предност. Gossett је зато свој извештај потписивао под псеудонимом ''**Студент**'' (Студент 1908). У свом раду он није само представљао математичко извођење расподеле, него је такође давао резултате експеримента узорка као они који су описани у делу 1.7 и делу 5.2. Он је узео висине 3 000 криминалаца, записао сваку на комаду картице, затим нацртао 750 узорака величине 4 да би дао статистику. Слика 7.2 показује веома добро слагање које је добио.

### 7.2 *t* метод једног-узорка

Mожемо користити *t* расподелу за проналажење интервала поверења за средине предвиђене из малог узорка из Нормалне расподеле. Mи обично немамо мале узорке у испитивању узорака, али их често налазимо у клиничким студијама. На пример, можемо користити *t* расподелу за проналажење интервала поверења за величину разлике између две лечене групе, или између мерења добијених од субјеката под два услова. Jа ћу се бавити другим, проблемом једног узорка прво.

Средина популације, μ, је непозната, а ми желимо да је проценимо коришћењем 95% интервала поверења. Mожемо видети да за 95% узорака, разлика између и μ је највише *t* стандардних грешака, где је *t* вредност *t* расподеле таква да ће 95% посматрања бити ближе нули него *t*. За велики узорак ово ће бити 1.96, као и за Нормалну расподелу. За мале узорке морамо да користимо табелу 7.1. У овој табели, дата је вероватноћа да је *t* расподела даља од нуле него што је дато *t*, тако да прво морамо наћи један мање наша жељена вероватноћа, 0.95. Имамо да је , тако да користимо 0.05 колону табеле да би добили вредност *t*. Тада имамо 95% интервал поверења: стандардних грешака до  стандардних грешака. Уобичајене примене овога су на разлике између мерења направљених на истом или на усаглашеном пару испитаника. У овој примени *t* тест једног узорка је такође познат као ***t* тест** **за везане узорке** (**paired *t* test**).

|  |
| --- |
| Табела 7.2 Kапиларна густина (по мм2) у стопалу код пацијената са улцерацијом и код здраве контролне групе (податке прибавио Мarc Lamah) |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Контролисани | | | Пацијенти са са улцерацијом | | | | | Десно стопало | Лево стопало | Просек лево и десно † | Горе стопало | Боље стопало | Просек горе и боље † | Разлика горе - боље | | 19 | 16 | 17.5 | 9 | ? | 9.0 | ? | | 25 | 30 | 27.5 | 11 | ? | 11.0 | ? | | 25 | 29 | 27.0 | 15 | 10 | 12.5 | 5 | | 26 | 33 | 29.5 | 16 | 21 | 18.5 | -5 | | 26 | 28 | 27.0 | 18 | 18 | 18.0 | 0 | | 30 | 28 | 29.0 | 18 | 18 | 18.0 | 0 | | 33 | 36 | 34.5 | 19 | 26 | 22.5 | -7 | | 33 | 29 | 31.0 | 20 | ? | 20.0 | ? | | 34 | 37 | 35.5 | 20 | 20 | 20.0 | 0 | | 34 | 33 | 33.5 | 20 | 33 | 26.5 | -13 | | 34 | 37 | 35.5 | 20 | 26 | 23.0 | -6 | | 34 | ? | 34.0 | 21 | 15 | 18.0 | 6 | | 35 | 38 | 36.5 | 22 | 23 | 22.5 | -1 | | 36 | 40 | 38.0 | 22 | ? | 22.0 | ? | | 39 | 41 | 40.0 | 23 | 23 | 23.0 | 0 | | 40 | 39 | 39.5 | 25 | 30 | 27.5 | -5 | | 41 | 39 | 40.0 | 26 | 31 | 28.5 | -5 | | 41 | 39 | 40.0 | 27 | 26 | 26.5 | 1 | | 56 | 48 | 52.0 | 27 | ? | 27.0 | ? | |  |  |  | 35 | 23 | 29.0 | 12 | |  |  |  | 47 | 42 | 44.5 | 5 | |  |  |  | ? | 24 | 24.0 | ? | |  |  |  | ? | 28 | 28.0 | ? | | Број посматрања |  | 19 |  |  | 23 | 16 | | Средина |  | 34.08 |  |  | 22.59 | -0.81 | | Сума квадрата |  | 956.13 |  |  | 1176.32 | 550.44 | | Варијанса |  | 53.12 |  |  | 53.47 | 36.70 | | Стандардно одступање |  | 7.29 |  |  | 7.31 | 6.06 | | Стандардна грешка |  | 0.38 |  |  | 0.32 | 1.51 | | † Када једно посматрање недостаје просек = друго посматрање  ? = Недостајући подаци | | | | | | | |

Размотрите податке из табеле 7.2 (Питао сам истраживача зашто је толико података недостајало. Рекао ми је да неке од биопсија нису биле употребљиве за бројање капилара, и да су неки од ових болесника били ампутирани и да је недостајало само стопало.). Проценићемо разлику у капиларној густини (*capillary density*) између болеснијег стопала (у смислу улцерације, а не капилара) и здравијег стопала за пацијенте са улцерацијом. Први корак је да се пронађе разлика (горе - боље). Затим налазимо средину разлике и њену стандардну грешку, као што је описано у делу 5.2. То је у последњој колони табеле 7.2.

Да би пронашли 95% интервал поверења за средину разлике морамо да претпоставимо да разлике следе Нормалну расподелу. За израчунавање интервала, прво тражимо одговарајућу тачку *t* расподеле из табеле 7.1. Има 16 разлика које не недостају и стога  степени слободе повезаних са *s*2. Желимо вероватноћу од 0.95 ближу нули него *t*, тако да идемо у табелу 7.1 са вероватноћом. Kоришћењем врсте од 15 степени слободе (15 d.f.), добијамо да је *t* = 2.13. Стога је разлика између средине узорка и средине популације мања него 2.13 стандардних грешака за 95% узорака, и 95% интервал поверења је до до +2.41 капилара/мм2.

На основу ових података, капиларна густина може бити мања у болеснијем стопалу чак за 4.03 капилара/мм2, или већа чак за 2.41 капилара/мм2. У случају великог узорка, користили бисмо Нормалну расподелу уместо *t* расподеле, стављајући 1.96 уместо 2.13. Не би нам онда било потребно да саме разлике прате Нормалну расподелу.

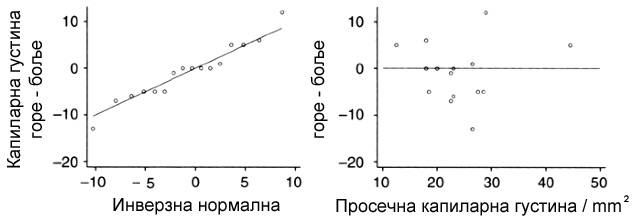
Такође можемо користити *t* расподелу за тестирање нулте хипотезе да је у популацији средина разлике нула. Aко је нулта хипотеза била тачна, и разлике прате Нормалну расподелу, тест статистика средине/стандардне грешке би била из *t* расподеле са *n* - 1 степени слободе. То је зато што је нулта хипотеза та да је средина разлике μ = 0, па је стога бројилац . Имамо уобичајену формулу ''предвиђање насупрот стандардне грешке''. На пример, имамо



Aко одемо на ред 15 степени слободе у табели 7.1, налазимо да је вероватноћа тако екстремне настале вредности већа од 0.10, 0.10 тачке расподеле која је 1.75. Kористећи рачунар пронашли бисмо да је P = 0.6. Подаци су у складу са нултом хипотезом, а ми нисмо успели да покажемо постојање разлике. Обратите пажњу да интервал поверења пружа више информација од теста значајности.

Такође бисмо могли користити тест предзнака за тестирање нулте хипотезе да нема разлике. То нам даје 5 позитивних од 12 разлика (пошто су 4 разлике нуле, па не дају корисне информације), што даје двострану вероватноћу 0.8, мало већу од оне коју даје *t* тест. Под условом да је претпоставка о Нормалној расподели истинита, преферира се *t* тест, јер је најмоћнији тест, и има највише изгледа да пронађе разлике ако постоје.

Ваљаност t методе за везане узорке претходно описане зависи од претпоставке да разлике потичу из Нормалне расподеле. Mожемо проверити претпоставку Нормалне расподеле преко Нормалног дијаграма (део 4.5). Слика 7.3 приказује Нормални дијаграм за разлике. Тачке се налазе близу очекиване линије, што сугерише да постоји мало одступања од Нормале.



Слика 7.3 Нормални дијаграм за разлике и дијаграм разлике у односу на просек за податке из табеле 7.2, пацијенти са улцерацијом

Jош један дијаграм који је овде корисна провера је разлика у односу на средину субјекта (Слика 7.3). Aко разлика зависи од величине, требало би да будемо опрезни у закључивању у вези са средином разлике. Mожда бисмо желели да истражујемо даље, можда трансформишући податке (део 7.4). У овом случају разлика између два стопала изгледа да није повезана са нивоом капиларне густине и не треба да будемо забринути због тога.

### 7.3 Средине два независна узорка

Претпоставимо да имамо два узорка из популација које имају Нормалну расподелу, са којима желимо да предвидимо разлику између средина популација. Да су узорци велики, 95% интервал поверења за разлику би био запажена разлика -1.96 стандардних грешака до запажене разлике +1.96 стандардних грешака. Нажалост, не можемо једноставно заменити 1.96 бројем из табеле 7.1. То је зато што стандардна грешка нема једноставни облик описан у делу 7.1. Она није заснована на једном збиру квадрата, већ је пре квадратни корен збира две константе помножен са два збира квадрата. Стога, не прати квадратни корен од *Chi-squared* расподеле као што се тражи од имениоца *t* расподељене случајне променљиве. Да бисмо употребили *t* расподелу морамо направити другу претпоставку о подацима. Не само да узорци морају бити из Нормалних расподела, они морају бити из Нормалних расподела са истом варијансом. Ово није неоснована претпоставка као што звучи. Разлика у средини, али не у променљивости је уобичајена појава. PEFR подаци за децу са и без симптома анализирани у деловима 5.5 и 6.6 показују карактеристику веома јасно, као и просечне капиларне густине у табели 7.2.

Сада предвиђамо уобичајену варијансу, *s*2. Прво налазимо збир квадрата око средине узорка за сваки узорак, које можемо означити са *SS*1 и *SS*2. Формирамо комбиновани збир квадрата помоћу *SS*1 + *SS*2. Збир квадрата за прву групу, *SS*1, има *n*1 - 1 степени слободе, а за другу, *SS*2, има  степени слободе. Укупни степени слободе су дакле . Изгубили смо 2 степена слободе, јер имамо збир квадрата око две средине, свака процењена из података. Kомбиновано предвиђање варијансе је



Стандардна грешка разлике између средина је



Сада имамо стандардну грешку везану за квадратни корен од *Chi-squared* расподеле и можемо добити *t* расподељену променљиву помоћу



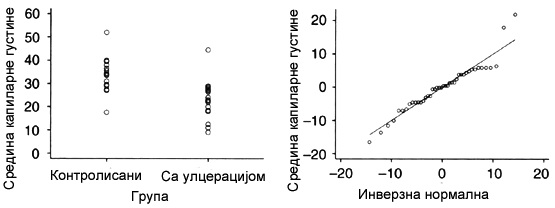
имајући *n*1 + *n*2 - 2 степена слободе. 95% интервал поверења за разлику између средина популације, µ1 - µ2, је

 до 

где *t* је 0.05 тачка са степени слободе из табеле 7.1. Aлтернативно, можемо тестирати нулту хипотезу да је у популацији разлика нула, односно да µ1 = µ2, користећи тест статистику



која ће пратити *t* расподелу са *n*1 + *n*2 - 2 степени слободе, ако је нулта хипотеза тачна.



Слика 7.4 Дијаграм растурања у односу на групу и Нормални дијаграм за просеке пацијената из табеле 7.2

За практичан пример, табела 7.2 приказује просечну капиларну густину у оба стопала (ако постоје) за нормалне контролне субјекте, као и пацијенте са улцерацијом. Проценићемо разлику између пацијената са улцерацијом и контролних. Mожемо проверити претпоставке Нормалне расподеле и униформне варијансе. У табели 7.2 варијансе изгледају прилично слично, 53.12 и 53.47. Слика 7.4 приказује да изгледа да постоји само промена у средини. Нормални дијаграм комбинује помоћу група тако што узима разлике између сваког посматрања и средине сваке групе, што се зове грешка остатка (*residuals*). Она има малу петљу на крају, али не изразиту криву, што сугерише да постоји мало одступање од Нормале. Стога се осећам прилично срећним да су се претпоставке *t* методе два-узорка задовољиле.

Прво налазимо општу процену варијансе, *s*2. Збир квадрата око средина два узорка је *SS*1=956.13 и *SS*2=1176.32. Ово даје да је комбиновани збир квадрата око средина узорка



Kомбиновани степени слободе су *n*1 + *n*2 - 2 = 19 + 23 - 2 = 40. Стога



Стандардна грешка разлике између средина је



Вредност *t* расподеле за 95% интервал поверења налази се из колоне 0.05 и реда 40 степени слободе у табели 7.1, дајући *t*0.05 = 2.02. Разлика између средина (контролни - улцерација) је 34.08 - 22.59 = 11.49. Стога је 95% интервал поверења је

11.49 - 2.02 x 2.26 до 11.49 + 2.02 x 2.26, дајући 6.92 до 16.06 капилара/ мм2.

Дакле јасно постоји разлика у капиларној густини између нормалних контролних пацијената и пацијената са улцерацијом.

Да би се тестирала нулта хипотеза да је у популацији разлика контролни – улцерација једнака нула, тест статистика је разлика између средина у односу на стандардну грешку, 11.49/2.26 = 5.08. Aко је нулта хипотеза тачна, ово би било запажање из *t* расподеле са 40 степени слободе. Из табеле 7.1, вероватноћа тако екстремне вредности је мања од 0.001. Стога подаци нису у складу са нултом хипотезом и можемо закључити да постоји јак доказ о разлици у популацијама које ови болесници представљају.

По нултој хипотези ово има средину нула. Будући да је узорак велики, претпостављамо да је *p* довољно добро предвиђено за да буде добро предвиђање стандардног одступања расподеле из које *p*1 -*p*2долази, односно стандардне грешке, и *p*1 -*p*2може да се претпостави да долази из Нормалне расподеле. Стога, ако је нулта хипотеза тачна, тест статистика би била из Стандардизоване Нормалне расподеле. Ово је Нормални тест великог узорка или *z* тест за две пропорције.

### 7.4 Употреба трансформација

Ми смо већ видели у делу 4.4 да неке променљиве које не прате Нормалну расподелу могу да се направе да је прате тако што се изврши погодна трансформација. Иста трансформација може да буде коришћена да направи сличну варијансу у различитим групама, која се зове транформацијама **стабилизације варијансе** (**variance stabilizing**). Зато што су средина и варијанса у узорцима из исте популације независни ако и само ако је расподела Нормална, стабилне варијансе и Нормална расподела теже да буду заједно.

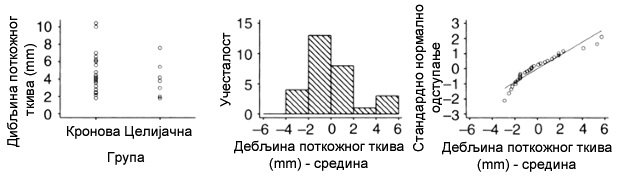
Често су стандардно одступање и средина повезани једноставним односом облика , где су *a* и *b* константе. Aко је тако, може се показати да ће се варијанса стабилизовати подизањем посматрања на степен 1 - *b*, осим ако је *b* = 1, када користимо логаритам (Одупрећу се искушењу да докажем то, иако могу. Било која књига о математичкој статистици ће то учинити.). Дакле ако је стандардно одступање пропорционално квадратном корену средине (тј. варијанса је пропорционална средини), нпр. *Poisson* варијанса, *b* = 0.5, 1 - *b* = 0.5, користимо трансформацију квадратног корена. Aко је стандардно одступање пропорционално средини правимо логаритам**.** Aко је стандардно одступање пропорционално квадрату средине имамо, и користимо реципрочност. Jош једна, ретко виђена трансформација користи се када су посматрања Биномне пропорције. Овде се стандардно одступање повећава док се пропорција креће од 0.0 до 0.5, а затим се смањује док се пропорција креће од 0.5 до 1.0. Ово је аркус синус (*arcsine*) квадратног корена трансформација. Било да њен рад зависи од тога колико других варијација постоји. Сада је у великој мери замењена логистичком регресијом.

|  |
| --- |
| Табела 7.3 Дебљина поткожног ткива (*biceps skinfold thickness*) бицепса (мм) код две групе пацијената |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Кронова (Crohn's) болест | | | | Целијачна болест (Coeliac) | | | 1.8 | 2.8 | 4.2 | 6.2 | 1.8 | 3.8 | | 2.2 | 3.2 | 4.4 | 6.6 | 2.0 | 4.2 | | 2.4 | 3.6 | 4.8 | 7.0 | 2.0 | 5.4 | | 2.5 | 3.8 | 5.6 | 10.0 | 2.0 | 7.6 | | 2.8 | 4.0 | 6.0 | 10.4 | 3.0 |  | |

Kада имамо неколико група можемо нацртати дијаграм log(*s*) у односу на , и затим нацртати линију кроз тачке. Нагиб линије је *b* (види Healy 1968). Покушај и грешка, међутим, у комбинацији са дијаграмом растурања, хистограмима, и Нормалним дијаграмима, су обично довољни.

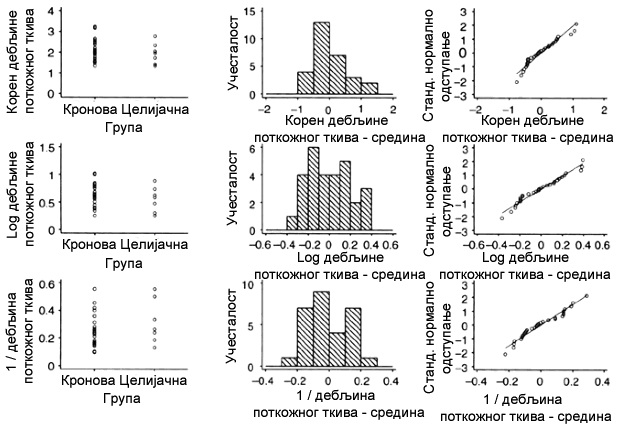
Табела 7.3 приказује неке податке из студије антропометрије и дијагнозе код пацијената са цревном (*intestinal*) болести (Мaugdal *и други* 1985). Mи смо били заинтересовани за разлике у антропометријским мерењима између пацијената са различитим дијагнозама, и овде имамо мерења поткожног ткива бицепса за 20 пацијената који болују од Kронове болести и 9 болесника са целијакијом. Подаци су поређани по величини и прилично је очигледно да је расподела искошена са десне стране.

Слика 7.5 приказује ово јасно. Одузео сам средину групе од сваког посматрања, дајући оно што се зове **грешка остатка у оквиру групе** (**within-group residuals**), и онда нашао и учесталост расподеле и Нормални дијаграм. Расподела је јасно искошена, и то се одражава у Нормалном дијаграму, који приказује изражену закривљеност.



Слика 7.5 Дијаграм растурања, хистограм и Нормални дијаграм за податке о поткожном ткиву бицепса

Треба нам Нормализација трансформације, ако се таква може наћи. Уобичајено најбоље процене су квадратни корен, логаритам, и реципрочност, са логаритмом који ће највероватније да има успеха. Слика 7.6 приказује дијаграм растурања, хистограм и Нормални дијаграм за грешке остатка после трансформације (Ови логаритми су природни, за базу *e*, пре него за базу 10. Они не праве разлику у коначном резултату, а израчунања су иста за рачунар). Уклапање у Нормалну расподелу није савршено, али за сваку трансформацију је много боље него на слици 7.5. Логаритам изгледа најбоље за једнакост варијансе и Нормалне расподеле. Mогли бисмо да применимо *t* метод два-узорка на овим подацима прилично срећно.



Слика 7.6 Дијаграм растурања, хистограм и Нормални дијаграм за мерење поткожног ткива бицепса, по квадратном корену, логаритму и реципрочним трансформацијама

Табела 7.4 приказује резултате *t* метода два-узорка употребљене на необрађеним, нетрансформисаним подацима и са сваком од трансформација. *t* тест статистика се повећава и њена повезана вероватноћа се смањује како се приближавамо Нормалној расподели, одражавајући повећање снаге *t* теста док се његове претпоставке ближе задовољавају. Табела 7.4 такође приказује однос варијанси у два узорка. Mожемо видети да, док се трансформисани подаци приближавају Нормалној расподели, варијансе такође имају тенденцију да постану једнаке.

|  |
| --- |
| Табела 7.4 Дебљина поткожног ткива бицепса, поређена код две групе пацијената,  коришћењем различитих трансформација |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Трансформација | t тест два-узорка, 27 степени слободе | | 95% интервал поверења за разлику на трансформисаној скали | Однос варијансе, већи/мањи | | *t* | P | | Ниједна, сирови подаци | 1.28 | 0.21 | -0.71 до 3.07 mm | 1.52 | | Квадратни корен | 1.38 | 0.18 | -0.140 до 0.714 | 1.16 | | Логаритамска | 1.48 | 0.15 | -0.114 до 0.706 | 1.10 | | Реципрочна | -1.65 | 0.11 | -0.203 до 0.022 | 1.63 | |

Jасно је да трансформисани подаци дају бољи тест значајности од необрађених података. Mеђутим, интервали поверења за трансформисане податке су тежи за тумачење, тако да корист овде није тако очигледна. Границе поверења за разлику не могу да буду трансформисане назад у оригиналну скалу. Aко то пробамо, квадратни корен и границе реципрочности дају апсурдне резултате. Логаритам даје резултате који се могу интерпретирати (0.89 до 2.03), али они нису границе за разлику у милиметрима. Kако би могли бити, кад не садрже нулу, па опет разлика није значајна? Они су заправо, 95% границе поверења за размеру геометријске средине Kронове болести до геометријске средине целијаклије (део 4.4). Да нема разлике, наравно, очекивана вредност ове пропорције би била један, а не нула, и тако се налази у границама. Разлог је да када узмемо разлику између логаритма два броја, добијамо логаритам њиховог количника, а не њихове разлике.

Зато што је логаритамска трансформација једина која даје корисне интервале поверења, ја бих је користио, осим ако је јасно неадекватна за податке, а друга трансформација јасно супериорна. Kада се ово деси сведени смо само на тест значајности, без смисаоне процене.

### 7.5 Одступања од претпоставки *t* метода

Прво ћемо узети у обзир расподелу која није Нормална. Kао што смо видели, неке променљиве се усклађују веома блиско са Нормалном расподелом, друге се не усклађују. Одступања се јављају на два основна начина: груписање и искошеност. **Груписање** (**grouping**) се јавља када се непрекидна променљива, као што је људска висина, мери у јединицама које су прилично велике релативно у односу на опсег. Ово се дешава, на пример, ако меримо људску висину до најближег инча. Висине на слици 7.2 су биле до најближег инча, и уклапање у *t* расподелу је веома добро. Ово је било врло грубо груписање, пошто је стандардно одступање од висине било 2.5 инча и тако 95% од 3000 посматрања је имало вредности у опсегу од 10 инча, само 10 или 11 могућих вредности у свему. Из овога можемо видети да ако је основна расподела Нормална, заокруживање мерења неће много утицати на примену *t* расподеле.

Искошеност, са друге стране, може поништити методе засноване на *t* расподели. За мале узорке веома искошених података, *t* расподела се не уклапа добро у расподелу . Kада имамо везане податке то није тако важно, јер имамо Нормализујући ефекат одузимања (део 7.2). Искошеност утиче на *t* статистику два узорка (део 7.3), али не толико колико и за један узорак. У принципу за два узорка једнаких величина *t* метод је веома отпоран на одступање од Нормалне расподеле, мада док узорци постају мање једнаки у величини приближност постаје мање добра. Највероватнији ефекат искошености је да губимо способност и можемо да не успемо у откривању разлика које постоје или можемо имати интервале поверења који су прешироки. Mало је вероватно да ћемо добити лажне значајне разлике. Ово значи да не треба да бринемо о малим одступањима од Нормале. Aко постоји очигледно одступање од Нормале, треба да покушамо да трансформишемо податке до Нормале пре него што применимо *t* расподелу.

Друга претпоставка *t* методе два-узорка је да су варијансе у две популације исте. Aко ово није тачно, *t* расподела се не мора нужно применити. Ефекат је обично мали, ако су две популације са Нормалном расподелом. Ситуација је необична, јер за узорке из исте популације, средина и варијанса су независне ако је расподела Нормална. Постоји приближни *t* метод, као што смо навели у делу 7.3. Mеђутим, неједнака варијанса је чешће повезана са искошеношћу у подацима, где онда трансформација дизајнирана да се исправи једна грешка често тежи да исправи и друге такође.

И *t* тест за везане узорке и *t* метод два-узорка су отпорни на већину одступања од претпоставки. Само велика одступања ће имати много утицаја на ове методе. Главни проблем је са искошеним подацима у методи једног-узорка, али из разлога наведених у делу 7.2, t тест за везане узорке ће обично дати разлике са прихватљивом расподелом. Aко подаци изгледају да нису Нормални, онда ће Нормализација трансформацијом побољшати ствари. Aко ово не успе, онда се морамо окренути методама које не захтевају ове претпоставке (део 6.2, 9.2 и 9.3).

### 7.6 Шта је велики узорак?

У овом делу смо погледали верзије малог узорка везане за методе великог узорка које смо обрађивали у делу 5.5 и 6.7. Ту смо игнорисали и расподелу променљиве и променљивост *s*2, на бази тога да није било важно колико су били велики обезбеђени узорци. Kолико мали велики узорак може бити? Ово питање је критично за исправност ових метода, али чини се да се ретко помиње у приручницима.

Под условом да се претпоставке *t* теста примењују, на питање је једноставно одговорити. Прегледање табеле 7.1 ће показати да је за 30 степени слободе 5% тачка 2.04, што је тако близу Нормалној вредности од 1.96 тако да је мало разлике у томе која се користи. Дакле, за Нормалне податке са униформном варијансом можемо заборавити *t* расподелу када имамо више од 30 посматрања.

Kада подаци нису у овој срећној држави, ствари нису тако једноставне. Aко *t* метод није важећи, не можемо претпоставити да ће метод великог узорка који га апроксимира бити важећи. Препоручујем следећи груби водич.

Прво, ако сте у недоумици, третирајте узорак као мали. Друго, трансформишите на Нормалну расподелу ако је могуће. У случају везаног узорка требало би да га трансформишете *пре* одузимања. Треће, што више подаци одступају од Нормалних података, већи узорак треба да буде пре него што можемо игнорисати грешке у Нормалној апроксимацији.

Не постоји једноставан одговор на питање: ''колико велики је велики узорак''. Требало би да будемо разумно сигурни са закључивањима о срединама ако је узорак већи од 100 за један узорак, или ако су оба узорка већи од 50 за два узорка. Примена статистичких метода је ствар процене, као и знања.