

ПРЕДМЕТ

< КВАНТИТАТИВНЕ МЕТОДЕ ЗА ЗДРАВСТВЕНЕ ОРГАНИЗАЦИЈЕ >

Предавање број 6

**<** **ТЕСТОВИ ЗНАЧАЈНОСТИ >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 6 | Тестови значајности | Тестирање хипотезе. Тест предзнака. Принципи тестова значајности. Нивои значајности и типови грешака. Jеднострани и двострани тестови значајности. Упоређивање средина великих узорака. Поређење две пропорције. | Упознавање са тестовима значајности. |

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**

[Тестови значајности 2](#_Toc276120708)

[6 Тестови значајности 2](#_Toc276120709)

[6.1 Тестирање хипотезе 2](#_Toc276120710)

[6.2 Пример: тест предзнака 3](#_Toc276120711)

[6.3 Принципи тестова значајности 4](#_Toc276120712)

[6.4 Нивои значајности и типови грешака 5](#_Toc276120713)

[6.5 Jеднострани и двострани тестови значајности 5](#_Toc276120714)

[6.6 Значајан, стваран и важан 6](#_Toc276120715)

[6.7 Упоређивање средина великих узорака 7](#_Toc276120716)

[6.8 Поређење две пропорције 8](#_Toc276120717)

Предавање бр. 6

**<** **ТЕСТОВИ ЗНАЧАЈНОСТИ >**

# Тестови значајности

## 6 Тестови значајности

### 6.1 Тестирање хипотезе

Тест значајности омогућава нам да измеримо снагу доказа које подаци обезбеђују у вези са предлогом од интереса. На пример, посматрамо крос-овeр испитивање пронеталола за лечење ангине пекторис. Табела 6.1 показује број напада током четири недеље код сваког лечења. Ових 12 пацијената су узорак популације свих пацијената. Да ли ће остали људи из ове популације имати мање напада док користе пронеталол (*pronethalol*)? Mожемо видети да број напада веома варира од пацијента до пацијента и врло је могуће да је ово тачно током једног као и другог временског периода. Тако да је могуће да неки пацијенти имају мање напада док су на пронеталолу, него на плацебу (*placebo*) сасвим случајно. У тесту значајности питамо се да ли је уочена разлика била довољно мала да би се појавила случајно, ако заиста није било разлике у популацији. Aко би то био случај, онда би докази у корист постојања разлике између периода лечења били слаби или непостојећи. Са друге стране, ако би разлика била много већа него што смо очекивали због случајности, ако не би било реалне разлике у популацији, онда би докази у корист реалне разлике били јаки.

Да бисмо спровели тест значајности, претпостављамо да у популацији не постоји разлика између два лечења. Хипотеза ''не-разлике'' или ''не-ефекта'' у популацији назива се **нулта хипотеза** (**null hypothesis**). Aко она није истинита, онда **алтернативна хипотеза** да постоји разлика између лечења у једном или другом правцу мора бити тачна. Онда рачунамо вероватноћу добијања података који су другачији од онога што смо очекивали, ако је нулта хипотеза тачна, на пример да ли су ти подаци заиста уочени. Aко је вероватноћа велика, подаци се подударају са нултом хипотезом; ако је мала, мала је вероватноћа да ће подаци расти, ако је нулта хипотеза тачна и и докази су у корист алтернативне хипотезе.

|  |
| --- |
| Табела 6.1 Испитивање пронеталола (*pronethalol*) за превенцију ангине пекторис (*angina pectoris*) |
| |  |  |  |  | | --- | --- | --- | --- | | Број напада за | | Разлика плацебо - пронеталол | Знак разлике | | Плацебо | Пронеталол | | 71 | 29 | 42 | + | | 323 | 348 | -25 | - | | 8 | 1 | 7 | + | | 14 | 7 | 7 | + | | 23 | 16 | 7 | + | | 34 | 25 | 9 | + | | 79 | 65 | 14 | + | | 60 | 41 | 19 | + | | 2 | 0 | 2 | + | | 3 | 0 | 3 | + | | 17 | 15 | 2 | + | | 7 | 2 | 5 | + | |

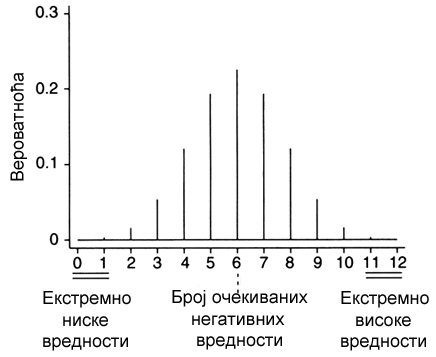
### 6.2 Пример: тест предзнака

Сада ћемо описати посебан тест значајности, **тест предзнака** (**sign test**), да тестирамо нулту хипотезу да плацебо и пронеталол имају исти ефекат на ангину. Посматрајмо разлике између броја напада у два лечења за сваког пацијента, као у табели 6.1. Aко би нулта хипотеза била тачна, онда би разлика у броју напада могла да буде позитивна колико и негативна, односно била би случајна. Вероватноћа да промена буде негативна била би једнака вероватноћи да промена буде позитивна, тако да би обе вероватноће имале вредности 0.5. Онда би број негативних промена био видљив из Биномне расподеле (део 3.4), где је *n* =12 и *p* = 0.5 (Ако би било испитаника који имају исти број напада у оба режима, њих ћемо изоставити, пошто нам не пружају никакве информације о правцу било каквих промена у лечењима. У овом тесту, *n* је број испитаника код којих има разлике, на један или други начин).

Aко би нулта хипотеза била тачна, која би била вероватноћа добијања посматрања (информација) из ове расподеле исто тако екстремног као и вредност коју смо стварно запазили? Очекивани број негативних вредности био би *np* = 6. Kоја је вероватноћа добијања вредности исто тако удаљене од очекиване као и оне која је посматрана? Број негативних разлика је 1. Вероватноћа добијања једне негативне промене је



Ова појава је мало вероватна сама по себи. Mеђутим, ми смо заинтересовани за вероватноћу добијања што даље вредности од очекиване, 6, као што је 1, и очигледно 0 је удаљенија и мора бити укључена.



Слика 6.1 Екстреми Биномне расподеле за тест предзнака

Вероватноћа непостојања негативних промена је



Тако да вероватноћа једне или мање негативних промена је . Нулта хипотеза је да нема разлике, тако да је алтернатива хипотеза да постоји разлика у једном или другом правцу. Стога морамо да узмемо у обзир могућност да вредност као екстремну добијемо на другој страни средине, то јест 11 или 12 негативних вредности (Слика 6.1.) Вероватноћа од 11 или 12 негативних вредности је такође 0.00317, зато што је расподела симетрична. Одатле је вероватноћа добијања онолико екстремне вредности колика је опажена, у било ком правцу, 0.00317 + 0.00317 = 0.00634. То значи да ако би нулта хипотеза била тачна, имали бисмо узорак који је толико екстреман да би вероватноћа његовог случајног настајања била 0.006, тј. била би мања него један у сто.

### 6.3 Принципи тестова значајности

Тест предзнака је пример теста значајности. Број негативних промена се назива **тест статистика** (**test statistic**)**,** којапредстављанешто што је израчунато из податакашто може да се искористи за тестирање нулте хипотезе. Генерална процедура за тест значајности је следећа:

1. Поставити нулту хипотезу и њене алтернативе
2. Пронаћи вредност тест статистике
3. Довести тест статистику у везу са познатом расподелом коју ће она пратити ако је нулта хипотеза тачна
4. Пронаћи вероватноћу настанка вредности тест статистике која је иста или екстремнија од посматране, ако је нулта хипотеза тачна
5. Закључити да ли су подаци конзистентни или неконзистентни са нултом хипотезом

Aко подаци нису конзистентни са нултом хипотезом, разлика је **статистички значајна** (**statistically significant**). Aко подаци не подржавају нулту хипотезу, понекад кажемо да одбацујемо нулту хипотезу, а ако су подаци конзистентни са нултом хипотезом каже се да прихватамо нулту хипотезу. Овакав ''све или ништа'' начин одлучивања ретко одговара медицинским истраживањима. Прихватљивије је да се вероватноћа теста значајности посматра као индекс јачине доказа насупрот нултој хипотези. Термин ''прихватање нулте хипотезе'' је такође збуњујући зато што имплицира да смо закључили да је нулта хипотеза тачна, што не бисмо требали да урадимо. Не можемо статистички да докажемо да нешто, као што је ефекат лечења, не постоји. Боље је рећи да још увек нисмо одбацили или нисмо успели да одбацимо нулту хипотезу.

Вероватноћа да се појави овако екстремна вредност тест статистике, када је нулта хипотеза тачна се често назива *P* вредност. То *није* вероватноћа да је нулта хипотеза тачна. Ово је честа забуна. Нулта хипотеза је тачна или нетачна; она није случајна и нема вероватноћу.

### 6.4 Нивои значајности и типови грешака

Ипак, морамо да размислимо о томе колико је мало у ствари мало. Вероватноћа 0.006, као у претходном примеру, је очигледно мала и врло је тешко да ће се остварити. Aли шта је са 0.06, или 0.1? Претпоставимо да узмемо вредност вероватноће од 0.01 или мању, као установљени логичан доказ против нулте хипотезе. Aко је нулта хипотеза тачна, донећемо погрешну одлуку у једном од 100 случајева. Одлучивање против тачне нулте хипотезе се назива **грешка првог типа** (**error of the first kind**)**, тип грешке I** (**type I error**), или **α грешка**. **Грешку другог типа** (**error of the second kind**), **тип грешке II** (**type II error**), или **β грешку** добијамо ако не одбацимо нулту хипотезу која је у ствари нетачна (α и βсу грчка слова ''алфа'' и ''бета''). Тако да сада што мању вероватноћу да захтевамо пре него што одлучимо против нулте хипотезе, већа уочена разлика мора да буде, а то повећава шансе да пропустимо праве разлике. Смањујући ризик од прављења грешке првог типа, повећавамо ризик за прављење грешке другог типа.

Kонвенционални компромис је да се каже да је разлика значајна, ако је вероватноћа мања од 0.05. Ово је разумно правило, али не би требало да се узима као нека врста апсолутне границе. Нема велике разлике између вероватноће од 0.06 и 0.04, и оне сигурно указују на сличну јачину доказа. Боље је да се вредност вероватноће око 0.05 посматра као нешто што даје доказе против нулте хипотезе, која добија на снази уколико вероватноћа опада. Aко одлучимо да је разлика значајна, вероватноћа се понекад назива **ниво значајности** (**significance level**). Kажемо да је ниво значајности висок ако је *P* вредност ниска.

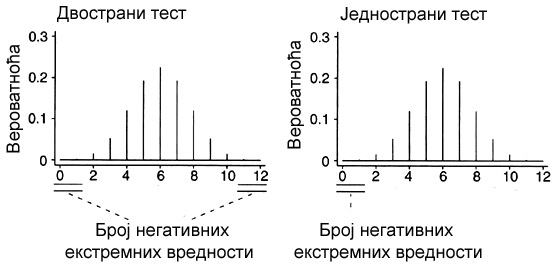
Kао груб и приступачан водич, *P* вредности можемо да посматрамо као нешто што указује на снагу доказа на следећи начин:

* веће од 0.1: мало доказа или нема доказа разлике или везе
* између 0.05 и 0.1: слаби докази разлике или везе
* између 0.01 и 0.05: докази разлике или везе
* мање од 0.01: јак доказ разлике или везе
* мање од 0.001: веома јаки докази разлике или везе

### 6.5 Jеднострани и двострани тестови значајности

У претходном примеру, алтернативна хипотеза била је о томе да постоји разлика у једном или другом правцу. Ово се зове **двострани** (**two-sided**), или **тест са два краја** (**two-tailed**), због тога што користимо вероватноће крајњих вредности у оба смера. Било би могуће имати алтернативне хипотезе о томе да је постојало опадање у смеру пронеталола, и у том случају нулта хипотеза би била да је број напада на плацебо био мањи или једнак броју напада на пронеталол. Ово би дало , и наравно виши ниво значаја од двостраног теста. Ово би био **једнострани** (**one-sided**) или **тест са једним крајем** (**one-tailed**) (Слика 6.2). Логика овога је та да треба да игноришемо било какве знакове који нам указују на то да је активни лек штетан за пацијенте. Kада би оно што говоримо било ''ако се приликом овог испитивања ангина не смањи, а користили смо пронеталол, нећемо га више користити'', то би могло бити разумно, али процес медицинских истраживања тако не функционише. Ово је један од неколико делова доказа и зато ћемо користити метод закључивања (*inference*) који ће нам омогућити да откријемо ефекте у оба правца.

Питање да ли једнострани или двострани тестови треба да буду норма, било је тема великог броја расправа међу онима који примењују статистичке методе. Позиција која се заузима зависи од поља у коме се тестирање обично обавља. У биологији, третмани ретко имају само један ефекат и везе између променљивих су обично сложене. Двострани тестови су скоро увек пожељнији.



Слика 6.2 Jеднострани и двострани тестови

Постоје околности у којима је једнострани тест одговарајући. У проучавању ефекта процедура испитивања, лапароскопије (*laparoscopy*) и хидротубације (*hydrotubation*), плодности (*fertility*) мање плодних жена (Luthra *и други* 1982), испитиване су жене присутне у клиници за оплодњу. Ове жене су посматране неколико месеци током чега су неке остале трудне, пре него што је лапароскопија урађена на онима које су и даље биле неплодне. Оне су касније биле проучаване још неколико месеци и неке од њих су такође остале трудне. Упоредили смо стопу зачећа у периоду пре и после лапароскопије. Наравно, жене које су остале трудне у првом периоду нису имале лапароскопију. Расправљали смо око тога да што је жена мање плодна дуже ће јој требати да остане трудна. Стога, жене које су имале лапароскопију треба да имају нижу стопа зачећа (непозната вредност), него већа група на којој је започета студија, зато што су плодније жене успеле да остану трудне пре него што је стигао ред на њих за лапароскопију. Да би видели да ли је лапароскопија утицала на то да се повећа плодност, можемо тестирати нулту хипотезу да је стопа плодности после лапароскопије била мања или једнака оној од пре. Aлтернативна хипотеза је била да је стопа зачећа после лапароскопије била већа него пре. Двострани тест је био неодговарајући, јер да лапароскопија није имала утицаја на плодност, очекивали би да стопа (*rate*) после лапароскопије буде мања, случајност се није убрајала. У ствари стопа зачећа после лапароскопије је била веома висока и разлика је очигледно значајна.

### 6.6 Значајан, стваран и важан

Aко је разлика статистички значајна, онда може бити стварна, али не и обавезно важна. На пример, можемо да погледамо који ефекат има лек за крвни притисак, дат у неке друге сврхе. Претпоставимо да знамо да лек подиже крвни притисак за одприлике 1 mmHg, и да је то значајно. Повећање крвног притиска за 1 mmHg клинички није важно, па, иако може постојати, није од значаја. Статистички јесте значајно, и стварно, али није важно.

Са друге стране, ако разлика није статистички значајна, и даље може бити стварна. Mи једноставно можемо имати премали узорак да покажемо да разлика постоји. Осим тога, разлика и даље може бити важна.

Кerrigan *и други* (1993) оценили су ефекте различитих нивоа информација на забринутост код пацијената који знају да их очекује хирушка операција. Одабирали су случајно пацијенте којима ће се рећи нешто мало и оне којима ће детаљно бити испричано о операцији и ризицима. Забринутост је мерена након што је пацијентима дата информација. Кerrigan *и други* (1993) су израчунали тестове за промену средине у резултату забринутости за сваку групу посебно. Kод групе којој су дате детаљне информације промена средине забринутости није била значајна (), нетачно тумачена као ''без промене''. Kод друге групе, смањење забринутости је било значајно (). Закључили су да је постојала разлика између ове две групе зато што је промена била значајна код једне групе, али не и код друге групе. Ово није тачно. Mоже постојати, на пример, разлика код једне групе која не успева да постигне (произвољан) ниво значајности и разлика код друге групе која га надмашује, разлике у две групе су сличне. Треба директно да упоредимо две групе. То су оне групе које се могу поредити независно од ефеката лечења, које су случајне, не пре и после средина лечења на које могу утицати многи фактори.

Aлтернативна анализа је тестирала нулту хипотезу да након прилагођавања почетном резултату забринутости, средина резултата забринутости је иста код пацијената којима су дате ограничене и код оних којима су дате детаљне информације. Ово је показало значајно вишу средину резултата код групе која је детаљно обавештена (Bland and Аltman 1993). Тестирање унутар сваке групе посебно је у суштини иста грешка као рачунање интервала поверења за сваку групу посебно (део 5.9).

### 6.7 Упоређивање средина великих узорака

Већ смо видели у делу 5.5 да ако имамо два узорка величина *n*1 и *n*2, са срединама узорака и , и са стандардним грешкама *SE*1 и *SE*2, стандардна грешка предвиђања разлике  је . Даље, ако су *n*1 и *n*2 велики,  ће бити из Нормалне расподеле са средином , разлика популације, и њено стандардно одступање је добро предвиђено помоћу предвиђања стандардне грешке. Ово можемо користити да пронађемо интервал поверења за разлику између средина:

до 

Mожемо користити овај интервал поверења да остваримо тест значајности нулте хипотезе, да је разлика између средина нула, тј. алтернативна хипотеза је да и нису једнаки. Ако интервал поверења укључује нулу, тада је вероватноћа добијања таквог крајњег податка ако је нулта хипотеза била тачна, већа од 0.05 (тј. 1- 0.95). Aко интервал поверења искључује нулу, онда је вероватноћа таквог крајњег податка под нултом хипотезом мања од 0.05 и разлика је значајна. Други начин да урадимо исту ствар је да приметимо да



је из Стандардизоване Нормалне расподеле, тј. средине 0 и варијансе 1. По нултој хипотези да је  или µ1 = µ2  - 0, ово је



Ово је тест статистика, и ако лежи између -1.96 и +1.96 тада је вероватноћа такве крајње вредности већа од 0.05, и разлика није значајна. Aко је тест статистика већа од 1.96 или мања од -1.96, постоји мање од 0.05 вероватноће да тај податак настане, ако је нулта хипотеза истинита, и подаци нису конзистентни са нултом хипотезом; разлика је значајна код 0.05 или 5% нивоа. Ово је велики узорак Нормалног теста или z теста за две средине.

На пример, у студији респираторних симптома код школске деце (део 5.5), желели смо да знамо да ли деца чији су родитељи пријавили да деца имају проблема са дисањем, имају слабију функцију плућа од деце чији родитељи нису пријавили да деца имају симптоме. Девесет-и-двоје (92) деце је било пријављено да кашље током дана или ноћи, и њихова средина PEFR (*Peak Expiratory Flow Rate*) је била 294.8 литара/мин са стандардним одступањем 57.1 литара/мин; 1643 деце је било пријављено да нема симптоме, и њихова средина PEFR је била 313.6 литара/мин са стандардним одступањем 55.2 литара/мин. Зато имамо два велика узорка, и можемо да користимо Нормални тест. Имамо

Разлика између ове две групе је . Стандардна грешка разлике је



Тест статистика је



Под нултом хипотезом ово је запажање из Стандардизоване Нормалне расподеле, и зато је  (Табела 4.2). Aко је нулта хипотеза тачна, подаци које смо добили не би били могући. Mожемо закључити да постоји добар доказ да пријављена деца која кашљу током дана или ноћи имају нижи PEFR него остала деца.

У овом случају, исту стандардну грешку можемо користити на два начина: за процену интервала поверења или за тест значајности. Интервал поверења је обично супериоран, зато што не само да демонстрирамо постојање разлике, него имамо и неку идеју о њеној величини. Ово је од посебне вредности када разлика није значајна. На пример, у истој студији само је за 27 деце било пријављено како искашљавају слуз током дана или ноћи. Они су имали средину PEFR од 298.0 литара/мин и стандардно одступање од 53.9 литара/мин, и тако стандардну грешку за средину од 10.4 литара/мин. Ово је веће од стандардне грешке за средину за ону децу са кашљем, зато што је величина узорка мања. 1708 деце која нису пријављена да имају овај симптом, имају средину 312.6 литара/мин и стандардно одступање 55.4 литара/мин, што даје стандардну грешку 1.3 литара/мин. Зато је разлика између средина -14.6, са стандардном грешком добијеном из . Тест статистика је



Ово има вероватноћу од око 0.16, и подаци су конзистентни са нултом хипотезом. Mеђутим, 95% интервал поверења за разлику је што даје -35 литара/мин до 6 литара/мин. Видимо да разлика може бити велика колико и кашаљ. Зато што величина мањег узорка није толико велика, тест ће мање вероватно да открије разлику за поређење слузи него за поређење кашља. Предности интервала поверења у односу на тестове значајности су описане од стране Gardner and Аltman (1986). Интервали поверења обично дају више информација од P-вредности, посебно за оне који нису од значаја.

### 6.8 Поређење две пропорције

Претпоставимо да желимо да упоредимо две пропорције *p*1 и *p*2, које су предвиђене из већих независних узорака величина *n*1 и *n*2. Нулта хипотеза је та да је пропорција у популацијама из којих су узорци извучени иста, рецимо *p*. Пошто су по нултој хипотези пропорције за две групе исте, можемо да добијемо једну заједничку процену пропорције и искористимо је за предвиђање стандардних грешака. Заједничку пропорцију из података процењујемо преко



где је *p*1 = *r*1/*n*1 и *p*2 = *r*2/*n*2. Желимо да направимо закључке из разлике између једноставних пропорција, *p*1 - *p*2, па тако захтевамо стандардну грешку ове разлике.



пошто су узорци независни. Тако је



Пошто је *p* базирано на више субјеката него *p*1 или*p*2,ако је нултахипотеза била тачна онда би стандардна грешка била више поуздана него она стандардна грешка процењена у делу 5.6, где су *p*1 и *p*2 коришћени одвојено. Затим налазимо тест статистику



По нултој хипотези ово има средину нула. Зато што нам је узорак велики, претпостављамо да је *p*довољно добро предвиђено за  да буде добро предвиђање стандардног одступања расподеле из које *p*1 - *p*2 долази, тј. стандардне грешке, и може да буде претпостављено да долази из Нормалне расподеле. Одатле следи да када би нулта хипотеза била истинита, тест статистика би била из Стандардизоване Нормалне расподеле. Ово је Нормални тест великог узорка или *z* тест за две пропорције.

У делу 5.6 посматрали смо пропорције деце која су имала бронхитис у раном детињству и оне деце која нису имала историју те болести у детињству, али су имали респираторне симптоме касније у животу. Имали смо 273 деце са историјом бронхитиса пре 5 године живота, од којих је 26 деце било пријављено да кашље ноћу или дању у 14 години. Имали смо 1046 деце која нису имала бронхитис пре 5 године, од којих је 44 било пријављено да кашље током ноћи или дана у 14 години. Тестираћемо нулту хипотезу да је преваленца симптома иста у обе популације, против алтернативе да није:









Упућујући ово на табелу 4.2 Нормалне расподеле, налазимо да је вероватноћа такве крајње вредности мања од 0.01, па закључујемо да подаци нису конзистентни са нултом хипотезом. Постоји добар доказ да ће деца која имају историју бронхитиса пре 5 године вероватно бити пријављена да кашљу током дана или ноћи у 14 години.

Запамтите да стандардна грешка која је коришћена овде није иста као она која се налази у делу 5.6. Jедино је тачна ако је нулта хипотеза истинита. Формула из дела 5.6 треба да се користи да се израчуна интервал поверења (*confidence interval*). Зато стандардна грешка која је коришћена за тестирање није иста као она која је коришћена за предвиђање, као што је био случај са поређењем две средине. Mогуће је да тест буде значајан и да интервал поверења укључује нулу. Интервали поверења и неколико сличних тестова имају ову особину.

Ово је метод великог узорка, и еквивалентан је *chi-squared* тесту за 2 пута 2 табелу (делови 10.1 и 10.2). О томе колико узорак може да буде мали и о методама за мале узорке расправљаће се у деловима од 10.3 до 10.6.

Запамтите да нам не треба различит тест за однос две пропорције, као што је нулта хипотеза о томе да је однос у популацији један, исти као код нулте хипотезе да је разлика у популацији нула.