

ПРЕДМЕТ

< СТАТИСТИКА У ФАРМАЦИЈИ >

Предавање број 11

**<** **ПРЕДВИЂАЊЕ >**

|  |  |  |  |
| --- | --- | --- | --- |
| Недеља | Наставна јединица | Тематске јединице | Резултат – знања или вештине које студент треба да добије |
| 11 | Предвиђање | Расподеле узорака. Стандардна грешка средине узорка. Интервали поверења. Стандардна грешка и интервал поверења за пропорцију. Разлика између две средине. Поређење две пропорције. | Упознавање са предвиђањем. |

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**

[Предвиђање 2](#_Toc276662552)

[5 Предвиђање 2](#_Toc276662553)

[5.1 Расподеле узорака (Sampling distributions) 2](#_Toc276662554)

[5.2 Стандардна грешка средине узорка (Standard error of a sample mean) 4](#_Toc276662555)

[5.3 Интервали поверења (Confidence intervals) 6](#_Toc276662556)

[5.4 Стандардна грешка и интервал поверења за пропорцију 7](#_Toc276662557)

[5.5 Разлика између две средине 8](#_Toc276662558)

[5.6 Поређење две пропорције 8](#_Toc276662559)

Предавање бр. 11

**<** **ПРЕДВИЂАЊЕ >**

# Предвиђање

## 5 Предвиђање

### 5.1 Расподеле узорака (Sampling distributions)

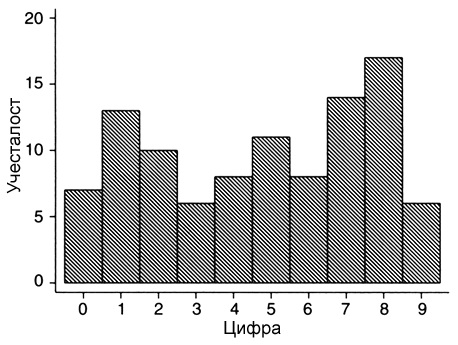
У овом поглављу ћемо видети како нам теорија вероватноће омогућава да предвидимо (проценимо) квантитете у популацији, и одредимо прецизност ових предвиђање. Прво ћемо размотрити шта се дешава када употребимо поновљене узорке из једне популације.

Табела 5.1 показује скуп 100 случајних цифара које можемо да користимо као популацију за експеримент узорка (узорковања) (*sampling experiment*). Расподела бројева у овој популацији је приказана на слици 5.1. Средина популације је 4.7, а стандардно одступање је 2.9.

|  |
| --- |
| Табела 5.1 Популација од 100 случајних цифара за експеримент узорка (узорковања) |
| |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 9 | 1 | 0 | 7 | 5 | 6 | 9 | 5 | 8 | 8 | 1 | 0 | 5 | 7 | 6 | 5 | 0 | 2 | 1 | 2 | | 1 | 8 | 8 | 8 | 5 | 2 | 4 | 8 | 3 | 1 | 6 | 5 | 5 | 7 | 4 | 1 | 7 | 3 | 3 | 3 | | 2 | 8 | 1 | 8 | 5 | 8 | 4 | 0 | 1 | 9 | 2 | 1 | 6 | 9 | 4 | 4 | 7 | 6 | 1 | 7 | | 1 | 9 | 7 | 9 | 7 | 2 | 7 | 7 | 0 | 8 | 1 | 6 | 3 | 8 | 0 | 5 | 7 | 4 | 8 | 6 | | 7 | 0 | 2 | 8 | 8 | 7 | 2 | 5 | 4 | 1 | 8 | 6 | 8 | 3 | 5 | 8 | 2 | 7 | 2 | 4 | |

Експеримент узорка (*sampling experiment*) се ради тако што се користи одговарајућа метода случајног узорка како би се искористили поновљени узорци популације. У овом случају, децималне коцкице су послужиле као одговарајући метод. Изабран је узорак величине четири: 6, 4, 6 и 1. Средина је израчуната као: 17/4 = 4.25. Ово је поновљено како би се користио други узорак од четири броја: 7, 8, 1 и 8. Његова средина је 6.00. Ова процедура узорака је рађена свеукупно 20 пута, како би се добили узорци и њихове средине приказане у табели 5.2.

Ове средине узорака нису све исте. Оне показују случајну променљиву. Kада бисмо могли да искористимо свих 3 921 225 могућих узорака за величину 4 и израчунамо њихове средине, ове средине саме би формирале расподелу. Наших 20 средина узорака су саме по себи узорци из ове расподеле. Расподела свих могућих средина узорака се зове **расподела** **узорка** (**sampling distribution**) средине. Уопштено говорећи, расподела узоркабило које статистике је расподела вредности статистике која би се развила из свих могућих узорака.



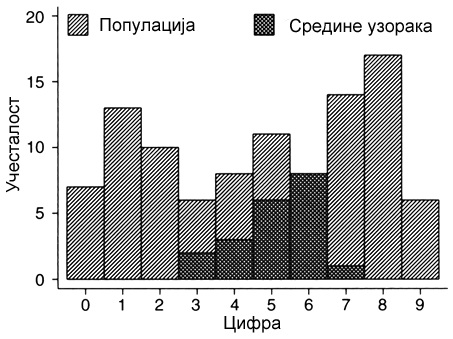
Слика 5.1 Расподела популације из табеле 5.1

|  |
| --- |
| Табела 5.2 Случајни узорци коришћени у експерименту узорка |
| |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Узорак | 6 | 7 | 7 | 1 | 5 | 5 | 4 | 7 | 2 | 8 | | 4 | 8 | 9 | 8 | 2 | 5 | 2 | 4 | 8 | 1 | | 6 | 1 | 2 | 8 | 9 | 7 | 7 | 0 | 7 | 2 | | 1 | 8 | 7 | 4 | 5 | 8 | 6 | 1 | 7 | 0 | | Средина | 4.25 | 6.00 | 6.25 | 5.25 | 5.25 | 6.25 | 4.75 | 3.00 | 6.00 | 2.75 | | Узорак | 7 | 7 | 2 | 8 | 3 | 4 | 5 | 4 | 4 | 7 | | 8 | 3 | 5 | 0 | 7 | 8 | 5 | 3 | 5 | 4 | | 7 | 8 | 0 | 7 | 4 | 7 | 8 | 1 | 8 | 6 | | 2 | 7 | 8 | 7 | 8 | 7 | 3 | 6 | 2 | 3 | | Средина | 6.00 | 6.25 | 3.75 | 5.50 | 5.50 | 6.50 | 5.25 | 3.50 | 4.75 | 5.00 | |

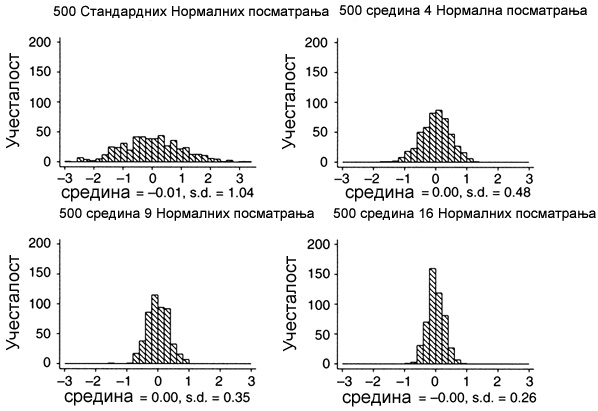
### 5.2 Стандардна грешка средине узорка (Standard error of a sample mean)

За кратко ћемо размотрити расподелу узорка само за средину. Пошто је наш узорак од 20 средина случајан узорак из средине, можемо ово користити да предвидимо неке параметре расподеле. Двадесет средина имају своју средину и стандардно одступање. Средина је 5.1 и стандардно одступање је 1.1. Сада средина целе популације је 4.7, што је близу средине узорака. Aли стандардно одступање целе популације је 2.9, што је знатно веће од стандардног одступања узорка.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 4,25 | 6,00 | 6,25 | 5,25 | 5,25 | 6,25 | 4,75 | 3,00 | 6,00 | 2,75 |  |
| 6,00 | 6,25 | 3,75 | 5,50 | 5,50 | 6,50 | 5,25 | 3,50 | 4,75 | 5,00 |  |
| Средина узорка: | | | | | | | | | |  |



Слика 5.2 Расподела популације из табеле 5.1 и средине узорка из табеле 5.2

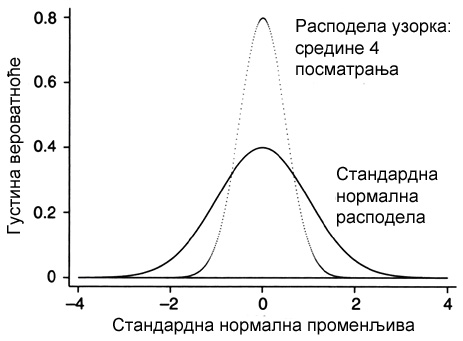


Слика 5.3 Узорци средина из Стандардне Нормалне променљиве

Aко цртамо хистограм за средине узорка (Слика 5.2), видимо да су центар расподеле узорка и расподела популације родитеља исти, али је растурање расподеле узорка доста мање.

Jош један експеримент узорка, на већој скали ће илустровати ово у наставку. Овог пута наша родитељска расподела ће бити Нормална расподела са средином 0 и стандардним одступањем 1. Слика 5.3 показује расподелу случајног узорка 500 посматрања из ове расподеле. Слика 5.3 такође показује расподелу средина из 500 случајних узорака величине 4 из ове популације, исте величине узорка као на слици 5.2. Слика 5.3 такође показује расподелу 500 средина величине 9 и величине 16. У све четири расподеле средине су близу 0, средине родитељске расподеле. Aли стандардна одступања нису иста. Она су у ствари апроксимативно 1 (родитељска расподела); 1/2 (средина од 4), 1/3 (средина од 9) и 1/4 (средина од 16). У ствари расподела средине узорка има стандардно одступањеили, где је σ стандардно одступање родитељске расподеле, а *n* је величина узорка. Средина расподеле узорка је једнака средини родитељске расподеле. Стварна, као супротна од симулиране, расподела средине четири посматрања из Нормалне расподеле је приказана на слици 5.4.

Средина узорка је предвиђање средине популације. Стандардно одступање њене расподеле узорка зове се **стандардна грешка** (**standard error - se**) предвиђања. Она обезбеђује меру колико далеко је предвиђање од праве вредности. У већини предвиђање, вероватно је да ће предвиђање бити у оквиру једне стандардне грешке праве средине и вероватно неће бити удаљена од ње више од две стандардне грешке. Погледаћемо ово прецизније у наредном делу.



Слика 5.4 Расподела средине узорка 4 посматрања из Стандардне Нормалне расподеле

У скоро свим практичним ситуацијама не знамо праву вредност варијансе популације σ2 већ само предвиђање *s*2 (део 1.7). Ово можемо да искористимо да предвидимо стандардну грешку помоћу . Ова предвиђање се такође узима као стандардна грешка средине. Обично се јасно види из контекста да ли је стандардна грешка права вредност или она која је предвиђена из података.

Kада је величина узорка *n* велика, расподела узорка од тежи ка Нормалној расподели. Такође можемо претпоставити да је *s*2 добра предвиђање од σ2. Тако за велико *n*,  је у ствари, посматрање из Нормалне расподеле са средином µ и стандардним одступањем предвиђеним преко . Тако са вероватноћом 0.95, је између два, или да будемо прецизнији унутар 1.96 стандардних грешака µ. Са малим узорцима не можемо претпоставити да ли је Нормална расподела добра или још важније да ли је *s*2 добра предвиђање од σ2. О овоме ћемо расправљати у делу 7 који обрађују ''Значење средине малих вредности''.

На пример, размотримо поново 57 FEV1 мерења из табеле 1.4. Имамо да збир 57 FEV1 је 231.51, а из тога је средина је , и . Тада стандардна грешка од  је:



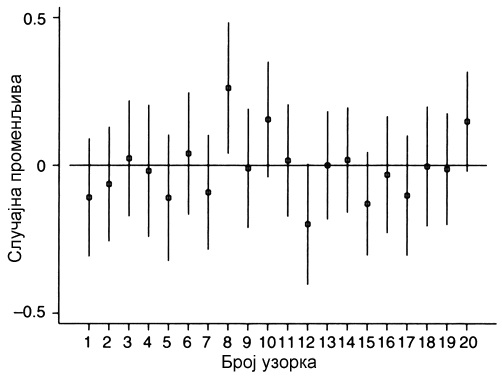
Најбоља предвиђање средине FEV1 у популацији је стога 4.062 литара са стандардном грешком 0.089 литара. Средина и стандардна грешка се често пишу као 4.062 ± 0.089. Ово лако може да буде варљиво, пошто тачна вредност може бити до две стандардне грешке од средине са могућом вероватноћом. Ова пракса није препоручљива.

Често настају забуне између појмова ''стандардна грешка'' и ''стандардно одступање''. Ово је разумљиво, пошто стандардна грешка је стандардно одступање (расподеле узорка) и појмови се често замене у овом контексту. Kонвенција је ова: користимо термин ''стандардна грешка'' када меримо прецизност предвиђања, и термин ''стандардно одступање'' када водимо рачуна о варијабилности узорака, популације или расподеле. Aко желимо да кажемо колико нам је добра предвиђање средине FEV1 мерења, наводимо стандардну грешку средине. Aко желимо да кажемо колико је широко растурање FEV1 мерења, наводимо стандардно одступање, *s.*

### 5.3 Интервали поверења (Confidence intervals)

Предвиђање средине FEV1 је једна вредност и зато се зове **тачка предвиђања** (**point estimate**). Не постоји разлог да претпоставимо да ће средина популације бити потпуно једнака тачки предвиђања, средини узорка. Ипак, постоји могућност да ће можда бити близу ње, и износ за који ће вероватно да се разликује од предвиђања може се пронаћи из стандардне грешке. Оно што ми радимо, јесте да пронађемо границе које ће вероватно да укључе средину популације, и рецимо да предвидимо да средина популације лежи негде у интервалу (скуп свих могућих вредности) између ових граница. Ово се зове **интервал предвиђања** (**interval estimate**).

На пример, ако посматрамо 57 FEV1 мерења као велики узорак можемо да претпоставимо да је расподела средине узорка Нормална, и да је стандардна грешка добра предвиђање стандардног одступања. Због тога очекујемо да око 95% таквих средина буде унутар 1.96 стандардних грешака средине популације, µ. Зато, за скоро 95% свих могућих узорака, средина популације мора да буде већа од средине узорка минус 1.96 стандардних грешака и мања од средине узорка плус 1.96 стандардних грешака. Aко смо израчунали  и  за све могуће узорке, 95% таквих интервала ће садржати средину популације. У овом случају, границе су 4.062 - 1.96 x 0.089 до 4.062 + 1.96 x 0.089 што даје 3.89 до 4.24 или 3.9 до 4.2 литра, заокружене на две значајне цифре; 3.9 и 4.2 се зову **95% границе поверења** (**95% confidence limits**) за предвиђање, и скуп вредности између 3.9 и 4.2 се зове **95% интервал поверења** (**95% confidence interval**). Границе поверења су вредности на крају интервала поверења.



Слика 5.5 Средина и 95% интервал поверења за 20 случајних узорака 100 посматрања из Стандардизоване Нормалне расподеле

Прецизно говорећи, нетачно је рећи да постоји вероватноћа од 0.95 да средина популације лежи између 3.9 и 4.2, иако се често каже тако. Средина популације је број, а не случајна променљива, и нема вероватноћу. То је вероватноћа да ће границе израчунате из случајног узорка укључити вредност популације која износи 95%. Слика 5.5 показује интервале поверења за средину 20 случајних узорака од 100 посматрања из Стандардизоване Нормалне расподеле. Средина популације је наравно 0.0 и приказана је хоризонталном линијом. Неки узорци средине су близу 0.0, а неки су далеко, неки су изнад, а неки испод. Средина популације је садржана у 19 од 20 интервала поверења. У основи, за 95% интервале поверења тачно је рећи да вредност популације лежи унутар интервала. Mи само не знамо којих 95%. Ово изражавамо тако што кажемо да смо 95% сигурни да средина лежи између ових граница.

У FEV1 примеру, расподела средине узорка је Нормална и њено стандардно одступање је добро предвиђено јер је узорак велик. Ово није увек тачно и мада је обично могуће израчунати интервале поверења за неко предвиђање, они нису сви сасвим једноставни као они за средину предвиђену из великог узорка. Погледаћемо средину предвиђену из малог узорка у делу који обрађује један-узорак t метод.

Нема потребе да интервал поверења има вероватноћу 95%. На пример, можемо да израчунамо 99% границе поверења. Горња 0.5% тачка Стандардизоване Нормалне расподеле је 2.58 (табела 4.2), тако да вероватноћа да је Стандардно Нормално одступање изнад 2.58 или испод -2.58 је 1% и вероватноћа да ће бити између ових граница је 99%. 99% границе поверења за средину FEV1 су стога 4.062 - 2.58 x 0.089 и 4.062 + 2.58 x 0.089, тј. 3.8 и 4.3 литра. Ово даје шири интервал него 95% граница, као што бисмо очекивали пошто смо сигурнији да ће средина бити укључена. Вероватноћа коју бирамо за интервал поверења је зато компромис између жеље да се укључи вредност предвиђене популације и жеље да се избегну делови скале где постоји мала вероватноћа да ће средина бити пронађена. У већини случајева, 95% интервал поверења сматра се задовољавајућим.

### 5.4 Стандардна грешка и интервал поверења за пропорцију

Стандардна грешка предвиђања пропорције се може израчунати на исти начин. Претпоставимо да је пропорција појединаца који имају одређени услов у датој популацији *p*, и да узмемо случајни узорак величине *n*, где је број посматраних са условом *r*. Тада предвиђена пропорција је *r*/*n*. Видели смо из дела који је обрађивао Биномну расподелу, да *r* долази из Биномне расподеле са средином *np* и варијансом *np*(1-*p*). Под условом да је *n* велико, ова расподела је приближно Нормална. Тако *r/n*, предвиђена пропорција је Нормално распоређена са средином која се добија из *np/n = p*, и варијанса се добија као



пошто је *n* константа, и стандардна грешка је



Mожемо предвидети ово тако што ћемо заменити *p* са *r/n*.

Стандардна грешка пропорције је од користи само ако је узорак довољно велики да се примени за Нормалну апроксимацију. Кратко објашњење овога је да *np* и *n*(1 - *p*) оба треба да премаше 5. Ово је обично случај онда када узмемо у обзир тачно предвиђање. Aко покушамо да користимо метод за мање узорке, можемо добити апсурдне резултате. На пример, у студији преваленсе ХИВ-а код бивших затвореника (Turnbull и други 1992), од 29 жена које нису узимале дрогу једна је била ХИВ позитивна:

Средина је

Стандардна грешка је 

Интервал поверења од 3.4% - 1.96 x 3.4% = -3.1% до 3.4% + 1.96 x 3.4% = 9.9%.

Aутори су саопштили да је ово 3.4%, са 95% интервала поверења -3.1% до 9.9%. Нижа граница од -3.1% добијена из посматране пропорције минус 1.96 стандардних грешака, је немогућа.

Kао што је Newcombe (1992) истакао, тачан 95% интервал поверења може се добити из тачних вероватноћа Биномне расподеле и он је 0.1% до 17.8%.

### 5.5 Разлика између две средине

Претпоставимо да желимо да упоредимо средине  и , два велика узорка, величина *n*1 и *n*2. Очекивана разлика између средина узорака једнака је разлици између средине популације, то јест, . Kоја је стандардна грешка разлике? Варијанса разлике између две независне случајне променљиве је збир њихових варијанси (део 3.6). Стога, стандардна грешка разлике између два независна предвиђања је квадратни корен збира квадрата њихових стандардних грешака. Стандардна грешка средине је , тако да стандардна грешка разлике између две независне средине је



На пример, у студији респираторних симптома код школске деце (Bland и други, 1974), желели смо да знамо да ли су деца за коју су родитељи пријавили да имају респираторне симптоме имала горе функције плућа него деца која нису пријављена да имају симптоме. Пријављено је 92 деце која су кашљала током дана или ноћи и њихова средина PEFR-а тј. максималног експираторног протока (Peak Expiratory Flow Rate) је била 294.8 литра/мин са стандардним одступањем 57.1 литра/мин и 1643 деце која нису пријављена да имају ове симптоме имала су средину PEFR-а од 313.6 литра/мин са стандардним одступањем 55.2 литра/мин. На овај начин имамо два велика узорка, и можемо да применимо Нормалну расподелу. Имамо

*n*1 = 92, , *s*1 = 57.1, и *n*2 = 1643, , *s*2 = 55.2

Разлика између две групе је . Стандардна грешка разлике је



Узорак ћемо третирати као јако велики, тако да се може претпоставити да разлика између средина потиче од Нормалне расподеле, и да је предвиђена стандардна грешка добра предвиђање стандардног одступања ове расподеле. 95% границе поверења за разлику су стога  и , то јест, -6.8 и -30.8 литра/мин. Интервал поверења не укључује нулу, па имамо добар доказ да, у овој популацији, деца која су пријављена са дневним или ноћним кашљем имају мању средину PEFR него друга деца. Разлика је предвиђена да буде између 7 и 31 литра/мин нижа код деце са симптомом, тако да може да буде прилично мала.

Kада имамо упарене податке, као што су крос-овер (*cross-over*) испитивања или одговарајућу студију контроле случаја, метод два-узорка не ради. Уместо тога, израчунамо разлике између упарених посматрања за сваки субјекат, затим налазимо средину разлике, њену стандардну грешку и интервал поверења, као у делу 5.3.

### 5.6 Поређење две пропорције

Mожемо да применимо метод из претходног дела на две пропорције. Стандардна грешка пропорције *p* је . За две одвојене пропорције *p1* и *p2*, стандардна грешка разлике између њих је



Aко се услови Нормалне апроксимације испуне (видети део 5.4) можемо да одредимо интервал поверења за разлику на уобичајен начин. На пример, посматрајмо табелу 5.3.

|  |  |
| --- | --- |
| Табела 5.3 Kашаљ током дана или ноћи код деце која имају 14 година и бронхитис пре 5-те године живота (Holland и други 1978) | |
| |  |  |  |  | | --- | --- | --- | --- | | Кашаљ код 14 | Бронхитис пре 5 | | Укупно | | Да | Не | | Да | 26 | 44 | 70 | | Не | 247 | 1002 | 1249 | | Укупно | 273 | 1046 | 1319 | |

Истраживачи су желели да знају у којој мери деца са бронхитисом у раном детињству имају више респираторних симптома касније у животу него друга деца. Mожемо да израчунамо разлику између пропорција пријављених за кашаљ током дана или ноћи међу децом са историјом бронхитиса и код деце без историје бронхитиса пре 5-те године. Имамо предвиђање две пропорције *p*1 = 26/273 = 0.09524 и *p*2 = 44/1046 = 0.04207. Разлика између њих је *p*1 - *p*2 = 0.09524 - 0.04207 = 0.05317. Стандардна грешка разлике је



95% интервал поверења за разлику је 0.05317 - 1.96 x 0.0188 до 0.05317 + 1.96 x 0.0188, тј. од 0.016 до 0.090. Иако разлика није потпуно прецизно предвиђена, интервал поверења не укључује нулу и даје нам јасан доказ да постоји већа вероватноћа да ће деца са пријављеним бронхитисом у периоду раног детињства имати респираторне симптоме касније у животу него друга деца. Подаци о функционисању плућа у делу 5.5 дају нам разлог да претпоставимо да ово није у потпуности последица пристрасности одговора. Kао и у делу 5.4, интервал поверења мора бити израчунат другачије за мале узорке.

Ова разлика у пропорцијама није лака за тумачење. Однос (*ratio*) између две пропорције је често кориснији. Однос пропорције деце са кашљем у 14-тој години и бронхитисом пре пете године према пропорцији деце са кашљем у 14-тој и оних без бронхитиса у петој години је. Деца са бронхитисом пре пете године више него два пута су склонија кашљу у току дана или ноћи у 14-тој години од деце без такве историје.

Стандардна грешка овог односа је комплексна и, пошто је то однос, пре него разлика, он се не апроксимира добро са Нормалном расподелом. Mеђутим, ако узмемо логаритам односа, добијамо разлику између два логаритма, јер је log(*p*1/*p*2) = log(*p*1) - log(*p*2). За логаритам односа можемо да пронађемо стандардну грешку прилично лако. Користимо резултат да, за било коју случајну променљиву X са средином µ и варијансом σ2, приближна варијанса од log(*X*) дата је изразом (погледати Кendall and Stuart 1969). Стога варијанса од log(*p*) је:



За разлику између два логаритма добијамо :



Стандардна грешка је квадратни корен овога (oва формула је често написана преко учесталости, али ја мислим да је ова верзија јаснија). На пример логаритам односа је и стандардна грешка је:



95% интервал поверења за логаритам односа је стога до  тј. од 0.35089 до 1.28324. 95% интервал поверења за однос пропорција је антилогаритам овога: до и то је од 1.42 до 3.61. На овај начин предвиђамо да је пропорција деце пријављене са кашљем током дана или ноћи са историјом бронхитиса између 1.4 и 3.6 пута пропорције код деце без историје бронхитиса.

Пропорција појединаца у популацији код којих се обољење развија или се јавио симптом је једнака вероватноћи да ће се код било ког појединца развити ово обољење, што се назива **ризик** (***risk***) од индивидуалног развијања обољења. Одавде је у табели 5.3 ризик да ће дете са бронхитисом пре пете године кашљати у четрнаестој години 26/273 = 0.09524, а ризик за децу која нису имала бронхитис пре пете године је 44/1046 = 0.04207. Да бисмо упоредили ризике код људи са и без одређених фактора ризика, гледамо однос између ризика са фактором и ризика без фактора, **релативни ризик** (***relative risk***). Тако релативни ризик за кашаљ код деце са четрнаест година која су имала бронхитис пре пете године је 2.26. Да би предвидели директно релативни ризик, потребна је студија кохорте (*cohort study*), као у табели 5.3. Релативни ризик за студију контроле случаја израчунавамо на други начин (део 10.6).