Statistiek voor tandartsen (VIII)

Student t-toets

A. H. B. Schuurs, tandarts[1]
H. J. Duivenvoorden, methodoloog[2]
M. A. van 't Hof, statisticus[3]

Uit de [1] vakgroep Cariologie en Endodontologie van het Academisch Centrum Tandheelkunde Amsterdam (ACTA), de [2] vakgroep Medische Psychologie en Psychotherapie van de Faculteit der Geneeskunde, Erasmus Universiteit te Rotterdam en de [3] Medisch Statistische Afdeling van de Faculteit der Geneeskunde en Tandheelkunde, Katholieke Universiteit te Nijmegen.

Trefwoorden: Statistiek

Datum van acceptatie: 29 november 1992.

Adres: Dr. A. H. B. Schuurs, ACTA, Louwesweg 1, 1066 EA Amsterdam.

Samenvatting. Er bestaan twee soorten t-toetsen. Met de t-toets voor 'gepaarde waarnemingen' wordt nagegaan of de scores van één persoon, bijvoorbeeld voor en na een ingreep, van elkaar afwijken. Daarbij wordt gebruik gemaakt van de standaarddeviatie (sd) van de verschilscores. Met de gewone t-toets wordt bezien of de scores van twee steekproeven van elkaar verschillen. De standaarddeviaties van beide steekproeven worden daarbij omgerekend tot één gemeenschappelijke deviatie.

Alle t-toetsen berusten op de t-waarde, dat wil zeggen op het verschil gedeeld door de fout in het verschil. Op grond van de t-waarde en het bijbehorende aantal vrijheidsgraden kan de p-waarde worden berekend.

SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Statistiek voor tandartsen (VIII). Student-t toets. Ned Tijdschr Tandheelkd 1993; 100: 495-8.

1 Inleiding

Op veel terreinen speelt de vraag of het gemiddelde van een onderzoeksgroep afwijkt van dat van een andere onderzoeksgroep. Zo wil men bijvoorbeeld weten of kinderen die wel en die geen fluoridetabletten gebruiken, gemiddeld evenveel caviteiten hebben. Fabrikanten willen weten of restauraties van hun nieuwe composietmateriaal gemiddeld een langere levensduur hebben dan die van hun vorige materiaal. En gezondheidsvoorlichters willen nagaan of een nieuwe voorlichtingsmethode tot beter poetsgedrag leidt.

Als het gemiddeld aantal caviteiten in een steekproef uit de ene onderzoeksgroep afwijkt van het steekproefgemiddelde uit een andere onderzoeksgroep, kan men zich afvragen of ook beide groepen van elkaar verschillen, of dat het verschil in de gemiddelden van de twee steekproeven aan toeval is te wijten. Om dat te kunnen nagaan, wordt de t-toets gebruikt. Een voorwaarde voor het toepassen van de t-toets is dat er sprake moet zijn van 'nette' gegevens, dat wil zeggen dat er geen 'uitschieters' (extreme waarden) in de verzamelde gegevens mogen voorkomen.

De t-toets wordt ook wel aangeduid als 'de toets van Student', naar de schuilnaam van de Ierse bierbrouwer Gosset, die de toets in 1903 ontwierp.

2 Twee t-toetsen

Met de t-toets worden gemiddelde waarden met elkaar vergeleken, rekening houdend met de nauwkeurigheid van de gemiddelden (Standard Error of the Mean, SEM). Er bestaan twee soorten t-toetsen. De eerste is de zogenoemde 'paired t-test', de tweede is de t-toets voor twee (onafhankelijke) steekproeven.

3 Paired t-test

Vaak worden in een klinisch onderzoek patiënten met twee verschillende tandheelkundige materialen behandeld om na te gaan of er verschil bestaat tussen beide materialen, bijvoorbeeld wat betreft de toestand van de gingiva of de levensduur van de restauratie. Omdat voor ieder individu een paar waarnemingen ter beschikking staat, ligt het voor de hand om de verschillen per paar te analyseren (gepaarde t-toets). Dus in plaats van met twee reeksen waarnemingen te werken, wordt er uiteindelijk gewerkt met slechts één reeks getallen, namelijk met één steekproef van verschilscores.[1]

Een voorbeeld: een tandarts wil weten of poetsinstructie de mondhygiëne van zijn patiënten (N = 1500) verbetert. Daarom bepaalt hij bij iedere patiënt die zich voor halfjaarlijkse controle meldt, linguaal respectievelijk palatinaal van alle eerste en tweede onder- en bovenmolaren de bloedingsneiging op een driepuntsschaal: 0 = geen bloeding, 1 = matige bloeding, 2 = sterke bloeding. Per patiënt berekent hij de gemiddelde bloedingsneiging. Aan de patiënten met een gemiddelde bloedingsneiging groter dan 0,8 – en dat blijken er 65 te zijn – geeft hij vervolgens poetsinstructie.

Om na te gaan of het met deze patiënten beter gaat, laat hij hen een maand later terugkomen en bepaalt op precies dezelfde manier opnieuw de bloedingsneiging. Met de beschikbare gegevens kan de tandarts nu tabel I samenstellen. Het is interessant om hierin naar de verschillen tussen de eerste en de tweede meting, de gemiddelde bloedingsneiging en de standaarddeviatie te kijken.[2]

3.1 De gegevens

Uit tabel I (Range) blijkt dat de eerste meting varieert van minimaal 0,88 tot maximaal 1,63. Geen patiënt had een sterke bloeding op alle acht meetplaatsen, want de waarde 2 komt niet voor. De tweede meting varieert van 0,25 tot 1,50. Gezien de range van de eerste meting en die van de tweede is het begrijpelijk dat het gemiddelde van de eerste meting hoger is dan dat van de tweede. Hoewel bij de tweede meting lagere waarden dan 0,8 worden gevonden, blijkt een aantal patiënten nog steeds een hoge bloedingsneiging te tonen.

Een verandering van de conditie van de gingiva komt tot uiting in de verschilscores (1ste min 2de metingen). Komen die op nul uit, dan is de conditie van de gingiva niet verbeterd, maar ook niet slechter geworden. Een positief getal duidt op een verbetering, een negatief op een verslechtering. Het gemiddelde van de verschilscores (0,37) is natuurlijk gelijk aan het verschil tussen het eerste en tweede gemiddelde (dus X1 - X2 = 1,21 - 0,84 = 0,37). Maar de standaardafwijking voor de verschilscores kan niet worden berekend uit het verschil van de standaarddeviaties (sd) (immers 0,18 - 0,32 ≠ 0,21). De waarde van de standaarddeviatie voor de verschilscores hangt nauw samen met de correlatie tussen de eerste en tweede meting.

3.2 Het 95% betrouwbaarheidsinterval

De verbetering van de gingivale conditie komt tot uiting in de positieve waarde van het gemiddelde 0,37 van de verschilscores van de steekproef. Zoals gezegd, het gemiddelde van de steekproef (N = 65) zal niet precies gelijk zijn aan het gemiddelde van de populatie (namelijk van alle patiënten met een hoge bloedingsneiging). Het mogelijke verschil tussen steekproefgemiddelde en populatiegemiddelde komt tot uiting in de 'Standard Error of the Mean' (SEM) die, zoals bekend,[3] gelijk is aan de standaardeviatie van de populatie σ, gedeeld door $\sqrt{n}$.

Nu is, zoals gewoonlijk, de waarde van σ niet bekend, want de tandarts heeft niet de gehele populatie onderzocht. Wel mag voor een redelijk grote steekproef, zoals hier het geval is, σ worden vervangen door sd, dus door de standaarddeviatie van de verschilscores, die gelijk is aan 0,21. De Standard Error of the Mean is als volgt te berekenen:[4]

$$SEM = s/\sqrt{n} = 0{,}21/\sqrt{65} = 0{,}026$$

Nu kan op eenvoudige wijze het 95%-betrouwbaarheidsinterval (95%-BI) voor de verbetering in de populatie worden bepaald; dat is gelijk aan de gemiddelde verschilscore plus en min tweemaal SEM, dus:

$$95\%\ BI = 0{,}27 \pm 2 \times 0{,}026 \rightarrow 0{,}22 \text{ en } 0{,}32$$

Derhalve geldt met een kans van 95% dat dit interval het populatiegemiddelde bevat.

Als er geen verbetering zou zijn opgetreden, zou het gemiddelde van de verschilscores op ongeveer nul uitkomen. We zien echter dat de waarde nul ver buiten het interval ligt en daarom behoort 'geen verbetering' normaal gesproken niet tot de reële mogelijkheden. Anders gezegd, de gevonden verbetering is significant, zelfs sterk significant.

3.3 t-waarde

Als het gemiddelde van de verschilscores gedeeld wordt door de waarde van SEM, wordt de zogenoemde t-waarde gevonden;

$$t = \frac{\text{gemiddelde verschilscores}}{\text{SEM}}$$

De t-waarde houdt dus in: het aantal SEM's dat het (verschil-)gemiddelde van de waarde nul afligt. Als t = 2, dan ligt de waarde nul precies op de rand van het 95%-betrouwbaarheidsinterval (gemiddelde $\pm$ 2 x SEM). Is de waarde van t kleiner dan 2, dan ligt nul binnen het interval en is het verschil in bloedingsneiging niet significant. Is de waarde van t daarentegen groter dan 2, dan ligt het gemiddelde ver van nul weg en is het verschil wel significant.

Bij het resultaat van het voorbeeld moeten twee aantekeningen worden gemaakt.

1. De gingivale verbetering van 0,27 is niet zo groot. Hoewel het toetsingsresultaat sterk significant is, lijkt het klinisch echter weinig relevant. Significantieniveau en p-waarde geven alleen aan hoe zeker het is dat er een verschil bestaat, maar niet hoe groot of hoe belangrijk het verschil is.

2. In dit onderzoek is niet bewezen dat de poetsinstructie de oorzaak is van de gevonden verbetering van de gingiva. Sterker nog, de poetsinstructie heeft waarschijnlijk geen enkel effect gehad, omdat hier een extreme groep is onderzocht, namelijk bestaande uit alleen patiënten met hoge bloedingsneiging. Nu vertoont iedere extreme groep de tendens om na verloop van tijd minder extreem te worden. Dat verschijnsel heet 'regression to the mean' of 'teruggang naar het gemiddelde'. Om het gevonden effect toe te kunnen schrijven aan de poetsinstructie, zou gecorrigeerd moeten worden op het effect van 'regression to the mean'. Dat kan door een controlegroep op te voeren van patiënten met een hoge bloedingsneiging, die géén poetsinstructie krijgen. Zodoende ontstaat een gerandomiseerd interventie-onderzoek. Maar men moet zich dan wel realiseren dat alleen al het feit dat een patiënt weet dat hij 'als controle' meedoet aan het onderzoek, effect op zijn poetsgedrag kan hebben; dus 'blindheid' is geboden.

3.4 Kleine steekproeven en vrijheidsgraden

Bovenstaande berekening werd uitgevoerd voor een redelijk grote steekproef (N = 65). De steekproef leverde een goede schatting voor σ op: sd (berekend voor de steekproef) mocht in plaats van σ worden gebruikt. Maar voor kleine steekproeven is sd slechts een onnauwkeurige benadering voor σ en dat betekent dus onzekerheid. Daarom moet bij kleinere steekproeven voor de waarde van t een ruimere grens dan 2 worden gekozen. De grens van de t-waarde hangt dus samen met de nauwkeurigheid van sd ten opzichte van σ. Deze nauwkeurigheid wordt op zijn beurt weer bepaald door het zogenaamde aantal vrijheidsgraden (degrees of freedom = df).

Het begrip 'df' houdt het volgende in. De standaarddeviatie wordt, zoals bekend, berekend als de wortel uit de variantie. En de variantie wordt berekend uit de som van de kwadraten van de verschilscores ten opzichte van het gemiddelde, dus uit:

$$\Sigma = (X_i - \bar{X})^2.$$

Het lijkt er nu op of n afzonderlijke kwadraten worden opgeteld, maar dat is slechts schijn. De afwijkingen $(X_i - \bar{X})$ zijn positief of negatief en vallen geheel tegen elkaar weg zodat hun som gelijk is aan nul. Dus in werkelijkheid zijn er slechts n-1 verschillen 'vrij te kiezen'; het laatste verschil moet ervoor zorgen dat het totaal, dat wil zeggen $\Sigma(X_i - \bar{X})$, op nul uitkomt en bezit dus niet meer de vrijheid om te variëren. Daarom berust sd op (n-1)-vrijheidsgraden: df = n-1. Voor de reeks van 65 waarnemingen in het voorbeeld geldt dat er 64 vrijheidsgraden zijn.

In een vorig artikel werden van de grenzen voor een aantal steekproefgrootten de t-waarden gepresenteerd (tab. I in artikel V).[3] Het is echter doeltreffender de grenzen te geven in relatie tot het bijbehorende aantal vrijheidsgraden, zoals in tabel II is gedaan (zie ook par. 4.3). In deze tabel is te zien dat voor df > 30 zonder nadelig effect de grenswaarde t = 2 gekozen mag worden. In de praktijk worden de tabellen weinig gebruikt, omdat de meeste computerprogramma's het significantieniveau en de bijbehorende p-waarde opgeven, rekening houdend met het aantal vrijheidsgraden.

Het begrip 'vrijheidsgraden' zal bij de nog te bespreken toetsen variantie-analyse en chi-kwadraattoets, eveneens een belangrijke rol spelen.

3.5 Eenzijdige of tweezijdige toetsing

Toetsing van verschillen kan op twee manieren geschieden. Bijvoorbeeld, als twee nieuwe behandelmethoden A en B worden vergeleken, kan tweezijdig worden getoets of deze gelijkwaardig zijn. Op grond van de toetsing zijn drie conclusies mogelijk: A is gelijk aan B òf A is beter dan B òf B is beter dan A

Als een nieuwe behandeling (B) vergeleken wordt met de oude standaardbehandeling (A), is men vaak uitsluitend geïnteresseerd in een eventuele verbetering en kan eenzijdig worden getoetst. De conclusie is dan: B is beter dan A òf B is niet beter dan A.

Bij tweezijdige toetsing wil men dus meer alternatieven onderzoeken, met als consequentie dat de grenswaarden voor t hoger liggen (zie tab. II).

4 Twee onafhankelijke steekproeven

De tweede soort t-toets (de 'gewone' t-toets) wordt gebruikt als twee afzonderlijke populaties met elkaar worden vergeleken.[1] De werkwijze om de gemiddelden van twee onafhankelijke steekproeven met elkaar te vergelijken, is vrijwel hetzelfde als voor de paired t-test. Maar het probleem is, dat er hier sprake is van twee onbekende standaarddeviaties van twee populaties. Om de 'Standard Error' (SE) van het verschil te berekenen kan worden uitgegaan van de standaarddeviaties van beide steekproeven. Dit kan worden geïllustreerd met het volgende voorbeeld.

Het vóórkomen van wortelcariës werd onderzocht via twee aselecte steekproeven uit zwarten en blanken. Voor een steekproef van 257 blanken werd gevonden dat $\bar{X}$ = 10,2 en sd = 10,0. Voor 176 zwarten bleek het gemiddeld aantal elementen met wortelcariës $\bar{X}$ = 6,1 te zijn met sd = 7,7.

Verschillen de gemiddelde scores voor wortelcariës voor blanken en zwarten? Als de nulhypothese is dat er geen verschil tussen beide rassen bestaat, kan de vraag ook als volgt worden omschreven: zijn beide steekproeven getrokken uit populaties met identieke gemiddelden μ? Is het verschil tussen de gemiddelde score van de blanken en zwarten 10,2 - 6,1 = 4,1 louter te wijten aan toeval of aan een werkelijk bestaand verschil tussen beide populaties?

4.1 Procedure

Voor de oplossing van het probleem mag de t-toets worden gebruikt, mits bekend is dat het twee aselecte steekproeven met 'nette' data betreft.

Van zowel de blanke als de zwarte populatie is σ onbekend, maar aangenomen mag worden dat σ voor beide dezelfde is. We gebruiken ook nu weer sd als geschatte waarde voor σ, dus beide standaarddeviaties van zowel de blanken als de zwarten zijn een schatting voor de onbekende waarde van de gemeenschappelijke σ. Maar beide standaarddeviaties bleken te verschillen. Om nu σ te bepalen, berekenen we uit beide varianties samen een nieuwe schatting van σ. Deze gemeenschappelijke sd wordt 'gewogen' berekend, dat wil zeggen met inachtneming van het aantal vrijheidsgraden (personen - 1) per steekproef.[1] Het zal duidelijk zijn dat de standaarddeviatie van de 257 blanken zwaarder weegt dan die van de 176 zwarten. De gemeenschappelijke standaardafwijking in dit voorbeeld zal daarom dichter bij 10 liggen dan bij 7,7. De precieze waarde, te vinden met een formule (hier niet gepresenteerd), blijkt 9,14 te zijn, de zogenoemde 'pooled sd'.

In deze t-test gaat het, evenals in de 'paired' situatie, om het verschil tussen de gemiddelden én de fout daarin. Het verschil, bij twee steekproeven, wordt berekend als het verschil van de twee gemiddelden. De fout in het verschil wordt bepaald door de fouten (SEM) in de twee gemiddelden. Deze SEM's (gelijk aan $s/\sqrt{n}$) kunnen op twee manieren worden berekend, namelijk op basis van de gemeenschappelijke standaarddeviatie (pooled sd), zoals die zojuist is berekend, maar ook op basis van de afzonderlijke standaarddeviaties.

De fout in het verschil, de Standard Error (SE), is natuurlijk groter dan de fout (SEM) in de afzonderlijke gemiddelden en wordt verkregen door de varianties op te tellen (want toeval telt altijd op en wordt niet gecompenseerd door ander toeval). Met gebruik van de afzonderlijke standaarddeviaties geldt (zie tab. III):

$$SE = \sqrt{(0{,}62^2 + 0{,}58^2)} = 0{,}85.$$

Gebruiken we de pooled sd dan geldt:

$$SE = \sqrt{(0{,}57^2 + 0{,}69^2)} = 0{,}89.$$

De bijbehorende t-waarden zijn nu eenvoudig te berekenen als:
t = verschil/fout in het verschil en blijkt dan voor afzonderlijke sd's 4,61 en voor pooled sd 4,82 te zijn. De berekende waarden zijn samengevat in tabel III.

4.2 Hebben blanken en zwarten evenveel wortelcariës?

De voor de 'pooled sd' berekende t-waarde

Tabel I. Gemiddelde bloedingsneiging van 65 patiënten vóór en na poetsinstructie.

Patiënt	1ste meting	2de meting	Verschil 1ste-2de meting
1	1,25	1,13	0,12
2	0,88	1,00	−0,12
3	1,50	1,13	0,37
4	0,88	0,25	0,63
5	1,00	1,25	−0,25
6	1,13	1,13	0,00
.			.
.			.
65	1,38	1,50	−0,13
Gemiddelde	1,21	0,84	0,37
Standaarddeviatie	0,88	0,31	0,21
Range	0,88-1,63	0,25-1,50	−0,25-0,88

Tabel II. t-verdeling. Grenzen van de t-verdeling bij tweezijdige en eenzijdige toetsing met $\alpha = 0{,}05$, gegeven het aantal vrijheidsgraden (df). De uitkomst van de toetsing is significant als t > tabelwaarde of als t < negatieve tabelwaarde.

df	Tweezijdig	Eenzijdig
1	12,71	6,31
2	4,30	2,92
3	3,18	2,35
4	2,78	2,13
5	2,57	2,02
6	2,45	1,94
7	2,37	1,90
8	2,31	1,86
9	2,26	1,83
10	2,23	1,81
15	2,13	1,75
20	2,09	1,73
25	2,06	1,71
30	2,04	1,70
50	2,01	1,68
100	1,98	1,66
−	1,96	1,65

Tabel III. Rekenschema t-toets voor het verschil bij twee steekproeven.

Ras	Gemidd.	sd(I)	n	sd(II)	SEM(I)	SEM(II)
Blank	10,2	10,0	257	9,14	0,62	0,57
Zwart	6,1	7,7	176	9,14	0,58	0,69
Verschil	4,1				0,85	0,89

I: op basis van afzonderlijke standaarddeviaties geldt: t = 4,1/0,85 = 4,82.
II: op basis van gemeenschappelijke standaarddeviatie: t = 4,1/0,89 = 4,61.

4,61 kan nu, in tabel II, worden vergeleken met de grenswaarde bij df = (257 - 1) + (176 - 1) = 431 bij tweezijdige toetsing. De waarde df = 431 staat niet vermeld, maar deze ligt tussen 100 en ∞. Dus ligt de t-waarde-grens tussen 1,98 en 1,96. De berekende waarde 4,61 is veel groter en H_0 moet dus worden verworpen. Volgens dit onderzoek hebben blanken en zwarten een significant verschillend aantal carieuze cervicale laesies.

De t-waarde bij de afzonderlijke standaarddeviaties = 4,82 is iets hoger dan die voor de pooled sd. Het aantal vrijheidsgraden is echter wat lager en kan niet worden vastgesteld via een eenvoudige optelling (de formule wordt hier niet gegeven). Beide t-test-methoden ontlopen elkaar in de praktijk niet veel. Voorzichtigheid is geboden als de kleinste steekproef van de twee de grootste standaarddeviatie heeft.

Uit bovenstaande berekening blijkt dat het aantal vrijheidsgraden nu niet zonder meer gelijk is aan n-1. Bij twee steekproeven geldt :

$$df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 = N - 2.$$

Dit verklaart waarom t-tabellen (zie tab. II) naar df en niet naar n worden opgegeven.

4.3 Opmerkingen

Bij het voorbeeld van de wortelcariës moet een kanttekening worden geplaatst. De verdelingen van de scores zijn positief scheef, dat wil zeggen, er zijn uitschieters naar rechts. Uitschieters naar links kunnen niet bestaan, want dan zou men op negatieve aantallen wortels uitkomen. Dat uitschieters naar rechts in het materiaal voorkomen, is ook te zien aan het feit dat de standaarddeviaties ongeveer gelijk zijn aan de gemiddelden. Aan de voorwaarde van 'nette verdelingen' is dus niet echt voldaan. Een transformatie van de gegevens, bijvoorbeeld door worteltrekken, zou hier nuttig kunnen zijn. Aan de andere kant is bij grote steekproeven de 'scheefheid' van minder groot belang, reden waarom de t-toets hier toch redelijk acceptabel is. Hier stuiten we op het grootste probleem bij het toepassen van statistiek, namelijk, de praktische beoordeling van de situatie in relatie tot een strikte theorie.

5 Slot

In de tandheelkunde komt het vaak voor dat twee groepen met elkaar vergeleken moeten worden. Te denken valt aan tegengestelde (natuurlijke) paren zoals: behandeld/onbehandeld, man/vrouw, links/rechts, ziek/gezond, voor/na behandeling, succes/mislukking. De t-toets is een machtig instrument om na te gaan of er dan verschillen bestaan. Maar daarbij moet men wel steeds de opzet van het onderzoek in het oog houden, om te beslissen of al dan niet gepaard gewerkt moet worden. Zo zullen links/rechts-verschillen vaak binnen één mond (dus gepaard) bestudeerd worden. En geslachtsverschillen aan de hand van twee steekproeven. Echter, in onderzoek aan gemengde tweelingen kunnen geslachtsverschillen ook gepaard bestudeerd worden, hetgeen zeer efficiënt kan zijn.

Het behoeft nauwelijks meer betoog dat toepassing van de t-toets een beoordeling van de netheid van de verdelingen vraagt. Bij grote steekproefomvangen zijn uitschieters niet zo belangrijk, bij kleine wel. Gelukkig bestaat er een goed alternatief voor de t-toets bij uitschieters, namelijk de rangnummertoets van Wilcoxon of de Mann-Whitney U-test die later zullen worden behandeld.

Summary

STUDENT T-TEST

Key word: Statistics

In order to test whether the mean scores of either two paired or independent samples deviate from one another, the Student t-test could be applied. In the paired samples case, the mean score of the differences between the matched scores is analyzed. In case of two independent samples, the difference between the two means is analyzed. In both cases the t-value is the difference divided by the error in difference, which, in combination with the degrees of freedom, results in the p-value.

Literatuur

[1] DE JONGE H. Medische statistiek. Deel II. Groningen: WoltersNoordhoff, 1964: 476-7, 482-3 en 486.
[2] SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Statistiek voor tandartsen. III. Ned Tijdschr Tandheelkd 1990; 97: 505-8.
[3] SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Statistiek voor tandartsen. V. Ned Tijdschr Tandheelkd 1991; 98: 474-6.
[4] SNEDECOR GW, COCHRAN WG. Statistical methods. Ames: The Iowa State University Press, 1976: 59-61.